Research Plan

A. Overview

A1. Specific Aims

A1.1 Aim 1. Complete the ChIP-seq binding profiles for the remaining ~300 TFs in each organism (Section B)

A1.2 Aim 2. Perform RNA-seq profiles of TF mutant or RNAi animals in each organism (Section B)

A1.3 Aim 3. Utilize bioinformatics pipelines to assess QC and define binding sites and DE genes (Section C)

A1.4 Aim 4. Integrate data, developing temporal-spatial regulatory networks and a mini-encyclopedia (Section C)

A1.5 Aim 5. Ensure resource release and data availability (Section D)

A2. Introduction

A2.1 Rationale

Transcription factors (TFs) play key roles in diverse aspects of development and physiology, including sex determination, early pattern formation, organogenesis, and response to environmental cues. A catalog of sites where transcription factors bind (regulatory sequences) is perhaps only second in importance to a catalog of genes in understanding how a genome specifies an organism. In this proposal we seek 1) to complete an initial catalog of binding sites for all transcription factors in both the fly D. melanogaster and the worm C. elegans; 2) to validate these sites through measuring the impact of transcription factor loss on the expression of genes; 3) to integrate binding sites and gene expression profiles both with one another and with other available information to develop models of gene expression and gene regulatory networks. These catalogs will represent the first comprehensive description of the TF binding sites in any metazoan and will provide a context for understanding the catalog of TF binding sites that will emerge from ENCODE.

A2.2. Advantages of model organisms

The *C. elegans* and *Drosophila* model organisms have several advantages for global mapping of transcription factor-DNA interactions. Both organisms have extensive comparative genomics resources, have powerful tools for profiling gene expression, and are easy to manipulate in the laboratory. Their genomes are among the most thoroughly and meticulously annotated metazoan genomes (a result in part of the transcript identification and annotation efforts of the modENCODE project), providing a stable platform upon which to perform and interpret our experiments ([1](#_ENREF_1" \o "Adams, 2000 #23), [2](#_ENREF_2" \o "C. elegans Sequencing Consortium, 1998 #3)). This information is key for correct assignments of binding sites to candidate target genes. At only ~1/30th the size of the human genome, the 100MB worm genome and 130MB fly genome are compact. Identifying regulatory motifs in these genomes is relatively efficient because they are proportionately high in information content, and regulatory motifs are confined to small regions relatively close to the promoter, when compared to the human genome. Additionally, these compact genomes increase sensitivity and decrease the cost of ChIP-seq experiments, because less sequencing is required to cover the genome, permitting a high level of multi-plexing, and because the reduced complexity increases the likelihood of detecting relatively rare binding events occurring in only a few cells. Most importantly, these organisms provide the opportunity to map transcription factor binding in the living organism, something that cannot be done easily in the study of human TFs. Finally, many of the transcription factors in both worms and flies are orthologous to human, and both organisms have long been successfully used to investigate the functions of these transcription factors during development. Research on individual fly and worm orthologs has led to important insights into the function of human disease genes and human biology generally. Thus, studying key conserved factors in this project will greatly enhance analysis, interpretation and the broader relevance of data gathered in the human ENCODE project. As we transition into a period where all the “parts lists” in genomes are being defined, it will be crucial to have detailed network maps in model organisms in order to accelerate the understanding of how the cognate genes function in orthologous and analogous networks in humans.

Each organism also has unique properties that are complementary and therefore make it worthwhile to proceed with analysis of transcription factor binding in both organisms. The particular advantages of *C. elegans* include its determinate cell lineage and transparency, which permit us to precisely link transcription factor expression with cell fate at a single-cell resolution ([3](#_ENREF_3" \o "Murray, 2012 #135)). Further, deletion mutants are available for every TF ([4](#_ENREF_4" \o "Mitani, 2011 #80), [5](#_ENREF_5" \o "Moerman, 2008 #95)) as well as multiple missense alleles (Thompson et al., 2013). *Drosophila* has been used to study many aspects of behavior, including sleep, courtship aggressiveness, learning and memory, making it possible to study neural networks, and ultimately help us decode the mysteries of human behavior ([6](#_ENREF_6" \o "Buszczak, 2007 #45), [7](#_ENREF_7" \o "Villella, 2008 #56)). Each organism undergoes distinct developmental programs, yet share a large number of factors that are conserved not only between worm and fly but also with human. Thus, understanding how transcription factor usage differs or is conserved between the organisms will be enormously informative. Both organisms are increasingly used as models for studying processes involved in a variety of human diseases, including cancer, diabetes, neurodegenerative diseases, and many others (e.g. reviewed in ([8](#_ENREF_8" \o "Pandey, 2011 #21), [9](#_ENREF_9" \o "Pfleger, 2008 #22))) often with one organism providing an especially suitable model.

A2.3. Fly and worm transcription overview

The basic mechanisms controlling gene regulation in both worms and flies are comparable to mammalian systems, including the interaction of transcription factors with RNA Pol II, and the influence of chromatin status on gene expression. Thus, mapping transcription factor function in flies and worms will be informative about human transcription factor function as well, with the added advantage of being analyzed in vivo.

*Drosophila* has 708 sequence-specific predicted transcription factor genes, of which 215 still remain largely unstudied and are known only by a curated gene identifier (CG) in FlyBase (Hammonds et al., PMID:24359758). Over half of the uncharacterized predicted factors contain a zf-C2H2 DNA-binding domain and may be involved in RNA-binding. *C. elegans* has 958 predicted transcription factor genes (pseudogenes removed from ([10](#_ENREF_10" \o "Reece-Hoyes, 2005 #21)) and additional factors from Hughes (Narasimhan, 2015; PMID: 25905672)). However, several of these have now been classified as RNA-binding or chromatin remodeling factors, leaving 892 \*\*\*I have 8—could be because of the added Hughes factors.\*\*\* sequence specific transcription factors. Of these, 260 represent an expanded family of nuclear hormone receptor genes. A small fraction of these (19) have orthologs outside of nematodes, but most are nematode specific and are poorly characterized; therefore all but a representative sample of the nematode specific nuclear receptor genes are excluded from this project, leaving 691 C. elegans TFs that are candidates for analysis. For both organisms, the factors to be analyzed represent the major conserved families of transcription factors, including homeobox, GATA, ETS, winged helix, HMG, bZIP, zinc finger, zf-C2H2, and T-box DNA binding domain-containing genes. From the previous modENCODE project and the modERN project we have successfully obtained ChIP-seq data sets for 220 worm and xxx fly TFs and we expect to a completed data set on another 58 worm and xxx fly factors by the end of the current grant in August, 2017. Another 65 worm TFs had expression at such low levels in just a few cells that either ChIP-seq failed or was put aside.

Thus, for each organism, the binding sites of approximately 340 factors will remain to be determined in each species at the start of this project.

In many cases, the restricted temporal and spatial expression of a transcription factor itself plays an important role in the specificity of binding to its target sites. In other cases, the transcription factor is broadly expressed, but shows distinct binding in different tissues or in response to different cues. For instance, the *C. elegans* factor PHA-4 acts at different times and in different tissues to mediate distinct responses, including foregut development in embryos and a starvation response in first stage L1 larvae, and has distinct binding sites at the different stages ([11](#_ENREF_11" \o "Zhong, 2010 #10)). The *Drosophila* transcription factor *Ultrabithorax* (*Ubx*) has different sets of targets in the haltere and the leg disc ([12](#_ENREF_12" \o "Slattery, 2011 #139)). To maximize biologically relevant results in the ChIP-seq experiments, it is critical to define where and when each factor is expressed prior to identifying candidate binding sites. In our project, the use of GFP-tagged protein fusions allows us to track protein expression for each tagged factor. These spatial and temporal expression patterns of the transcription factors, along with existing RNAseq and anatomic data ([3](#_ENREF_3" \o "Murray, 2012 #135), [13](#_ENREF_13" \o "Graveley, 2011 #136), [14](#_ENREF_14" \o "Hillier, 2009 #137)) (Gerstein et al, 2014), are used to select the optimal developmental stage and condition during which ChIP-seq is performed. Knowledge of the expression pattern will also prove useful in validating the ChIP-seq results, and will ultimately be critical in elucidating the regulatory network of each organism. However, it is important to emphasize that the proposed project does not intend to provide all the binding sites for all the factors in all tissues and all time points. Rather, this project will create the necessary resources for the community to perform such experiments while generating an initial reference dataset that defines where in the genome each factor binds at the point of maximal expression. For the Ubx TF, for example, the initial data from the modENCODE project defined its binding sites in embryos ([15](#_ENREF_15" \o "Negre, 2011 #138)) while the reagents generated by the project led to detailed tissue-specific follow-up experiments in imaginal discs ([12](#_ENREF_12" \o "Slattery, 2011 #139)). In *Drosophila*, approximately 85% of TFs are expressed during embryogenesis ([15-17](#_ENREF_15" \o "Negre, 2011 #138)) and therefore select time points in embryogenesis will likely be examined for most of the factors. A similar fraction of TFs are expressed during embryogenesis in the worm, and two-thirds of TFs have their maximal expression in the embryo.

Finally, in our proposed genome annotation effort we recognize that the presence of a biochemically defined binding site proximal to a gene is not itself sufficient evidence of a functional regulatory role. Binding may be inconsequential or the regulatory target may not be properly inferred simply by being the closest gene to the binding site. Empirically, DNA microarray and RNA-seq experiments are commonly used to profile gene expression changes in transcription factor mutants, in order to identify affected downstream genes. In this project, we will perform RNA-seq on upon mutation or RNAi knockdown of selected TFs, in order to determine systematically the fraction of binding sites that result in local changes in gene expression *in vivo* across many different functional classes of TFs. We will also explore other methods to validate the activity of the discovered binding sites.

A2.4. Using ChIP-seq to investigate TF-DNA interactions

ChIP-seq (chromatin immunoprecipitation sequencing) is a widely-used method to map transcription factor binding sites, in which a transcription factor of interest is immunoprecipitated from a chromatin extract using an antibody specific for the transcription factor ([18-20](#_ENREF_18" \o "Horak, 2002 #16)). Deep sequencing of the DNA in the precipitate identifies the specific bound regions enriched relative to control DNA. In flies and worms, the assay provides an *in vivo* genome-wide map of binding sites for a transcription factor under endogenous regulation. We have found that the established ChIP protocol provides the best experimental strategy to balance the number of factors analyzed with the most comprehensive genome-wide data collection. We did explore methods like Exo-seq, but found it did not produce reliable results and would have slowed production. We therefore will rely primarily on the established ChIP protocol in this project. However, we will explore the recently developed alternative protocol, Cut&Run (Skene and Henikoff, 2017 PMID: 28079019), which by targeting micrococcal nuclease to just the DNA surrounding TF binding sites is reported to have a much improved signal to noise ratio and lower requirements for input material (see Resource Production (Part Bx.x) for additional details).

In both organisms, the experimental pipelines begin with the epitope-tagging of transcription factor genes with GFP through recombineering, generally as an in-frame carboxy-terminal fusion. The resulting DNA products then are used to create transgenic strains. The GFP expression pattern of each strain is determined, and an optimal developmental stage (or stages) is selected for analysis. The strain is then grown to the selected stage(s) and harvested for ChIP-seq. We developed this approach rather than raising antibodies to individual transcription factors because the requirement for very high antibody quality places a substantial time and effort burden at the very beginning of the pipeline with no certainty of success and more variable results between factors, complicating later analyses. Although cloning and transgenesis also require considerable effort, they are generally predictable. Additionally, our previous experience indicates that expression patterns of GFP-tagged factors closely match those of the corresponding endogenous factors ([21](#_ENREF_21" \o "Venken, 2010 #147)). Notably, the ENCODE project began with antibody generation, but switched to GFP-tagging because of our success and because of the relative ease it provided. Finally, transgenic lines created in this project provide the fly and worm research communities an enduring and important resource. Figure 1 shows the overall outline of the pipeline for this proposal. In Section B we describe our results to date and provide additional experimental details of the pipeline.

A2.5. Investigating the function of binding sites

As alluded to above, multiple studies have indicated that the simple presence of a binding site in a ChIP-seq assay proximal to a gene is not itself sufficient evidence of a functional regulatory role. Our experience in the modENCODE project made it abundantly clear that transcription factor binding profiles are more complex than perhaps many expected. For many TFs, thousands of statistically significant binding sites are identified, and the mechanism of binding to, as well as the function of, many of these sites is difficult to assess. Not all sites display an obvious consensus sequence, even for factors whose binding preference is known. Additionally, many sites are bound by multiple, often unrelated factors, and it is unclear whether such sites are bound in the same or different tissues, or simultaneously within a given cell. Finally, many of the genes adjacent to binding sites do not exhibit changes in expression levels in transcription factor mutants. Chronic loss of a transcription factor could lead to buffering or compensation of gene expression levels by the time expression is analyzed, and many changes could be missed. Understanding these complexities will require the efforts of many different labs.

Despite this complexity, we chose in the modERN project to define which binding sites are associated with changes in gene expression for several reasons. First, we have developed a strategy that enriches for gene expression changes that are the initial, direct consequence of the absence of transcription factor function, rather than downstream, indirect changes or missing relevant changes that are compensated for. Second, this information will help to make an important distinction between those candidate target genes whose expression relies heavily on that particular transcription factor, and those target genes whose expression is unaffected, either because other transcription factors have greater effects, or for other reasons. This distinction can be used in further analyses for comparison of other properties, such as the identification of consensus binding sites. Finally, a systematic analysis of binding sites associated with expression for many transcription factors will permit a global assessment of the fraction of binding sites that can be expected to be associated with gene expression.

A3. Resource Summary

There are three major resources for each organism that will be provided by this project. The first of these is a set of transgenic strains expressing GFP-tagged TFs. In worms, about 340 TFs still need to have tagged strains produced. For ~150 of these, we will utilize our existing pipeline, in which fosmids containing the entire endogenous TF locus with a GFP-3xFLAG tag inserted at the carboxy-terminus will be introduced into worms as integrated, low-copy transgenes. The remaining TFs do not have available fosmids, so tagged strains will be generated using CRISPR/Cas9 strategies to introduce the GFP-3xFLAG tag into the exogenous locus. In flies, about 320 TFs still require strain generation. The vast majority have BACs available for recombineering, and transgenic flies will be generated using the existing pipeline. For the 41 TFs with a BAC we will use the MIMIC transposon strategy, in which a transposon is used to insert GFP into an intron spanning two coding exons at the endogenous locus (PMID:26102525). Worm strains will be made in the Waterston lab at UW, and fly strains by the Celniker lab at Berkeley, as established with our current pipeline. All strains that pass quality control will be used for subsequent ChIP-seq analyses and will be deposited in strain repositories for each organism, the Caenorhabditis Genetics Center (CGC) and the Drosophila Stock Center, respectively.

The second resource generated will be the dataset of binding sites for these 340 worm and 320 fly TFs, at one or more developmental stages. When added to the existing binding site data from modENCODE and modERN, the catalog of TF binding sites for these two crucial model organisms will be essentially complete. To perform these experiments, we will utilize the existing pipeline that we have optimized during the modERN project, briefly described as follows. Worm strains are sent from UW to the Reinke lab at Yale, at which point TF expression is confirmed and catalogued. The strain is grown and harvested at the peak developmental stage for TF expression and function. ChIP is performed using an anti-GFP goat antibody, and the isolated, purified chromatin is sent to the White lab at University of Chicago for Illumina library preparation and sequencing on an Illumina HiSeq4000 machine. Similarly, fly strains are sent from Berkeley to UC for growth to the appropriate developmental stage, ChIP is performed using the same anti-GFP antibody, and library construction, and sequencing is the same as for worms. ChIP-seq is performed in duplicate for each TF in each organism. The resulting sequencing files are processed using the established ENCODE pipeline, which applies an IDR (irreproducible discovery rate) analysis to determine the quality of the experiment and to identify significant binding sites. With this pipeline, therefore, several of the steps are performed identically for both species, resulting in economies of scale and production, and also simplifying cross-species comparisons. In this second funding period, in addition to completing the whole-animal analysis for all encoded TFs, we also propose to test whether tissue isolation can improve binding site identification for TFs with extremely tissue-specific expression patterns by focusing on a select few cases. We will also perform ChIP-seq analysis of a select histone modification (H3K27ac), which marks active enhancers at multiple developmental stages in both organisms. These data will continue to be deposited in the ENCODE database to provide a public portal for the entire research community.

The third resource is a dataset of gene expression profiles in which the effects of TF loss-of-function is determined in each organism. During the current grant period, we established effective experimental approaches in each species to monitor the effects of loss of TF activity on gene expression. In worms, a subset of TFs whose expression begins during embryogenesis was identified, and deletion mutant strains for those TFs were ordered from either the CGC or the Mitani group. Staged embryos were collected for RNA isolation at three timepoints during embryogenesis that correspond to 1) immediately before TF expression begins, 2) the time of peak TF expression, and 3) shortly after peak TF expression. These timepoints were selected to best capture baseline gene expression, then the earliest, most direct TF-dependent gene expression changes, and finally the more consequential but possibly more indirect TF-dependent gene expression changes. In flies, a similar approach is used, but transgenic RNAi strains (PMID:26320097) are typically used to disrupt TF activity instead of genetic mutants; timepoints taken follow the same logic. RNA isolated from both worms and flies is sent to UW for library preparation and sequencing. Thus, once again, after dealing with strain-specific experimental constraints, we consolidate the sequencing and informatics pipeline to streamline the process and increase the ability to make cross-species comparisons. In this proposal, we plan to continue these efforts to generate TF-dependent gene expression profiles, at the rate of 15 factors per year per organism.

Importantly, our ChIP-seq data demonstrates that certain pairs of TFs have adjacent binding sites more frequently than expected, so we also propose to examine the functional relationship of key TF pairs on gene expression by performing RNA-seq analysis in strains in which the activity of both TFs is disrupted. Our goal is to complete 5 factor-pairs per year per organism. As with the ChIP-seq data, all of these profiles will be deposited in the ENCODE database and publically accessible via their website.

In addition to providing these resources, we will also perform a series of bioinformatic analyses on these data. We will identify binding sites and candidate gene targets from the ChIP-seq data, and determine which sites are likely to be HOT (highly occupied target) sites, and which are likely to represent specific regulatory events. We will determine differentially expressed genes from the RNA-seq experiments. Finally, by further integrating the data from these “ground level” annotations along with publicly available datasets, we will define middle and top level annotations. For instance, to identify enhancer elements, we will use chromatin accessibility, histone modifications, TF binding sites and HOT regions. Also, we also intend to construct regulatory networks by integrating the ChIP-seq data and the RNA-seq data within each organism to define meta-transcriptional regulatory networks. We will include analyses and parameters such as motif enrichment and hierarchical structure, and use spatial and temporal restrictions to constrain these networks.

The strategies to generate all of these resources are described in detail in Section B, and the bioinformatic analysis planned are described in detail in Section C.

Resource Sharing: To distribute and share those resources, we will place the strains generated in the respective organismal strain repositories (CGC and Drosophila Stock Center). The ChIP-seq and RNA-seq data will continue to be sent through the established pipeline to the ENCODE database. The plans for data dissemination and resource sharing are described in detail in Section D. \*\*\*Might amplify this section some.\*\*\*

A4. Community Support

The resources we generated during the modENCODE and modERN projects to date are of great interest to the community. We have put in effort to spread the word to the community, not only through publications but by presenting multiple workshops on these resources at international meetings over the years in order to describe our progress, explain how to access data, and to elicit community feedback. Specific resources that we supply have been heavily accessed. The strain usage is easiest to track since we deposit all our strains in the respective stock centers and they keep excellent records of the strain requests. For the fly, the Bloomington Stock Center has received 5400 requests for strains we have generated since 2012. For the worm, the *Caenorhabditis* Genetics Center has answered 1600 requests. For both Centers, our strains are among the largest single source \*\*\*Checking on this with Ann Rougvie.\*\*\* of strain requests. In addition, members of the community frequently contact our labs, requesting specific strains, ChIP-seq data, and asking for additional data collection. We take their requests into account; indeed, dozens of factors have been prioritized for GFP tagging and ChIP analysis based on community requests. In some cases, even specialized TF factors that require unusual growth conditions have been analyzed on request, including hypoxia treatment and dauer formation in worms. Usage of the ChIP-seq data itself is harder to track, but our publications on transcription factor binding sites in the fly and worm have been cited more than 1800 times (Google Scholar), with many of these citations using the data for a particular factor or set of factors to understand gene regulation. We anticipate that a complete catalog of TF binding sites in the worm and fly would be of even greater interest to the community. [[Should we list some more examples of individual contacts with potential collaborations/assistance, and/or get letters from Wormbase and Flybase?]]

A5. Impact

The primary, most immediate, impact of the resources described above is to provide the broader research community with freely accessible, substantial transcriptional regulatory data in two key model organisms. This “raw material” will facilitate a vast number of scientific inquiries into the function of individual TFs in key developmental, metabolic, defense and homeostatic regulatory pathways, as well as provide a broader perspective on how individual TFs work together in local networks and even globally across the lifespan of an organism. Buttressed with the classical experimental strengths of worms and flies, including sophisticated genetic and cell biological approaches, the genomic regulatory data we generate should be leveraged very efficiently and extensively by the larger community into a sophisticated understanding of how complex regulatory systems are integrated in the living animal.

In addition to providing a complete catalog of binding sites, we will be developing and testing new experimental strategies to improve specificity and accuracy. In particular, we are focusing on applying new tissue-specific strategies and beginning to define the regulatory relationships between pairs of TFs. Using our fluorescent tags and FACS, we will evaluate the effect of TF disruption on the expression of possible targets. We will disrupt the activity of pairs of TFs whose sites co-occur at higher than expected frequencies and investigate their effect on gene expression using RNA-seq. We attempt to assay TF binding sites in specific tissues, combining CUT&RUN with FACS of labeled cells. These new approaches will provide new insights into our data sets. [[etc this paragraph will give us a chance to highlight our new technologies tried to do this!]]

Finally, we also believe our experience will complement efforts in human, both those already underway and planned for the near future. The challenge to define regulatory networks in human is much greater than worms and flies, with ~1500 TFs needed to be assayed across multiple cell lines and where progress to date we believe has been more modest, with xxx, yyy and zzz factors assayed on the most studied cells lines, often with very little overlap between cell types. [[we should get these numbers - \*\*\*Maybe phrase in terms of the maximal number of TFs for any one cell type is only xxx. I put in some possible language, if we can get the numbers. \*\*\*]]. With our data now deposited in the ENCODE DCC, links between our results and human projects should be easier to establish. In addition to providing comparative data for individual factors and a general understanding of how the full set of TFs drive gene expression, our results may also inform the ENCODE project in other ways. For example, what is the value of the full catalog? Can the last fraction of TF profiles be imputed? What other data types are most useful in any imputation? What other data types are most effective in translating the binding site data into a more profound understanding of gene regulation? In sum, we believe our project will inform, supplement and facilitate other large-scale projects with a similar focus.

B. Research Project/Resource Production

B1. Specific Aims

B1.1 Aim 1 Generate transgenic strains and ChIP-seq profiles for remaining TFs for each organism

B1.2 Aim 2 Generate RNA-seq binding profiles upon loss-of-function for a subset of TFs for each organism

B2. Progress Report

B2.1 Summary of achievements and implemented optimizations of ChIP-seq (Aim 1)

During the the modERN project so far, we have generated XX strains (XX worm and XX fly) and XX ChIP-seq profiles (XX worm and XX fly). When combined with the data from the modENCODE project, we have completed XX (XX worm and XX fly) profiles for XX TFs (XX worm and XX fly) . This section describes our accomplishments in detail, as well as how we optimized many aspects of the pipelines. \*\*\*I hope ot get these numbers from people tomorrow.\*\*\*

B2.1.1 Streamlined ChIP-seq production pipeline for worms

For strain generation, we utilize fosmid-based transgenes that have a GFP-3xFLAG tag inserted at the C-terminus of transcription factor genes (Figure 2). Recombineering was used to first insert the tag into a 30-40kb fosmid in which the transcription factor gene is roughly centered and surrounded by extensive flanking regulatory sequence (Sarov et al, 2012 PMID: 22901814). In turn the *unc-119* selection marker was inserted into the vector backbone of this engineered fosmid also by recombineering. The insertion sites were checked by sequencing. In total, 154 of the remaining worm factors have recombineered fosmids available for transgenesis by the bombardment technique, which results in low-copy-number, randomly-integrated transgenes. The Waterston lab performs biolistic transformation to introduce transgenic DNA into worms. DNA-coated gold beads are bombarded into *unc-119* mutant worms, which are then put through a multiple-week selection process to identify animals at least rescued for the *unc-119* phenotypes of uncoordinated movement and a failure to form dauers upon starvation. Animals that have formed dauers and are non-Unc are picked and their progeny analyzed for the frequency of transmission of the transgene. If 100% of the progeny are non-Unc over multiple generations, then the transgene is considered integrated into the genome. These strains are then assessed for GFP expression of the transgene, in particular whether the factor is localized to the nucleus, and whether the expression pattern matches that of the endogenous factor, if known. With the efficiencies developed over time, one person can bombard at least twelve fosmids constructs per week and perform the initial screening for rescued animals. Strains with detectable nuclear expression are sent to the CGC and to the Reinke lab.

We have made several improvements over the past four years to increase the yield of transformed animals. Rather vortexing the fosmid DNA with the beads to keep the beads in suspension, we now use a gentle rocking procedure to minimize shearing of the fosmid DNA. The result was an increase in both the number of rescued animals and a higher fraction of expressing lines. We switched from using standard ngm plates with OP50 bacteria, to peptone rich plates with NA22 bacteria to provide a richer food source, allowing a larger brood of eggs from the bombed animals to reach the dauer stage. We switched from spermadine to protamine to bind the DNA to gold beads, because protamine is more stable and, hence, more reliable. Finally, we developed an overarching SQL Lab Information System to track the precise status of all TFs in pipeline.

Once the strains are received by the Reinke lab, the next step of the pipeline is to culture worms to the appropriate stage. For each transgenic strain, we record which cell type(s) exhibit GFP expression at each of the major stages of development (embryos, larval stages 1-4 and young adult hermaprhodites and in select cases, dauers and males), comparing our observations to the RNA-seq data sets. Almost but a very small number of strains (28) have had sufficient numbers of cells expressing at a high enough level to warrant proceeding with ChIP analysis. The failed strains often have very low expression levels in the RNA-seq data; others may have lost expression during propagation. These factors might be assayable with a method more sensitive than whole animal ChIP-seq. In a larger number (64) the fluorescent signal is detected only in the cytoplasm. This location could indicate that the co-factor responsible for transporting the TF to the nucleus is absent under laboratory conditions. More likely, our predicted transcription factor list contains RNA binding factors as well as TFs, since the list was deliberately inclusive (Reece-Hoyes and Walhout, 2005 PMID:16420670).

We also gather information for each transcription factor, including mutant or RNAi phenotypes, from WormBase and the literature. Based on these data, we select a primary developmental stage at which the factor has maximal expression and/or function for analysis. We also select secondary stages, if warranted. We assess whether the transgenic strain exhibits any features that possibly indicate an over-expression phenotype or a disruption of an important gene by the transgene insertion, and might thus preclude analysis. Very few lines examined to date have exhibited visible phenotypes. Worms are then grown on enriched growth plates in the presence of food (OP50 bacteria) and bleached to obtain eggs. The embryos are placed in liquid culture and allowed to hatch in the absence of food to synchronize to the early L1 larval stage (a two-hour window). At least 1x106 synchronized L1 worms are then grown on plates to the desired developmental stage and collected by extensive washing and centrifugation to remove the bacteria. The cleaned worm pellets are exposed to formaldehyde at room temperature for 30 minutes to crosslink proteins to chromatin, rinsed, and then placed in lysis buffer and stored at -80 until ready for ChIP. Approximately 6 preps can be grown and harvested per week per person. Note that for each factor at each stage, duplicate preps are grown, making this 3 factors per week.

To perform ChIP, preps previously fixed in formaldehyde are thawed in lysis buffer and the chromatin is sonicated. Two chromatin preps are used for a typical ChIP-seq experiment. Small amounts of each chromatin prep are removed and pooled to serve as a control “input” sample. A completed dataset for a given factor then includes 2 ChIP samples and 1 input control. A total of 2mg of sonicated chromatin is immunoprecipitated with our validated goat anti-GFP antibody. Importantly, this antibody is the same anti-GFP antibody as used for *Drosophila* ChIP, and is another advantage of combining worm and fly pipelines. Crosslinks are reversed on both samples, and the three samples describing a dataset are submitted to the White lab at the University of Chicago for library construction and sequencing.

During the modERN project, we have improved several aspects of the ChIP protocol. In particular, we often had low DNA yields that led to failed library construction about 20% of the time. We determined that this was likely due to inefficient disruption of the worms during lysis and sonication, and have subsequently added a douncing step to further break open the worms prior to sonication. We also tested several sonication parameters to establish the setting that allows maximal breakage without destroying the sample by overheating. These two improvements decreased the failures to less than 5%. Finally, we have recently begun tests in which we alter the timing of the fixation. We reasoned that fixation prior to lysis was likely to toughen the cuticle and reduce efficiency of chromatin extraction, so we tested whether fixation after douncing and before sonication resulted in improved chromatin isolation. Our preliminary results suggest that we see about 10-fold more chromatin with this method, and that this chromatin is capable of being immunoprecipitated and results in very robust library construction. Sequencing of this sample produced excellent peaks, and we are awaiting sequencing of the input sample to complete the quality control analysis. Assuming that this is a successful ChIP experiment, we will implement the same strategy for a few additional test factors before adopting it generally.

B2.1.2 Streamlined ChIP-seq production pipeline for flies

GFP-tagging of fly TF genes are performed in the White lab. Recombineering is also used to insert a GFP tag into the C-terminus of fly transcription factor genes using the P[acman] (PhiC31 artificial chromosome for manipulation) system. Two P([acman] BAC libraries have been constructed, one with an average 30 kb and one with an average 80 kb genomic fragments ([30](#_ENREF_30" \o "Venken, 2009 #152)). Carefully selected BACs ensure that tagged transcription factors retain their genomic regulatory context; we predict that the regulatory elements for most transcription factors lie between the two closest insulator elements, and we use modENCODE-produced insulator data to ensure the tagged BAC includes all the DNA between the two closest predicted insulators or extends to cover the nearest genes upstream and downstream of the TF.  BAC transgenes are, therefore, expected to express encoded genes at physiological levels similar to the endogenous gene of interest, an expectation that has been borne out by our empirical observations ([30](#_ENREF_30" \o "Venken, 2009 #152), [49](#_ENREF_49" \o "Poser, 2008 #151)).

Once a BAC for a given transcription factor gene is selected, we use recombineering to tag the C-terminus of the protein. The C-terminal recombineering cassette contains sequence encoding a version of GFP that is codon optimized for expression in *Drosophila*, followed by a short bacterial promoter (gb3) that drives expression of the neomycin-kanamycin resistance gene in *E. coli*. We insert C-terminal tagging cassettes, containing 50 nucleotides of PCR-introduced homology arms, into the BAC by recombineering in front of the stop codon of the gene. All steps of transgene production can be carried out in multi-well-plate format, making it readily possible to generate 100 constructs per year, as we propose for this project. *E. coli* cells that have successfully recombined the cassette are selected for kanamycin resistance in liquid culture. We check multiple independent clones for each reaction by PCR across the tag insertion point. Our initial experience generating over 100 recombineered BACs indicates that the necessary recombineering steps can be carried out with extremely high fidelity (see Summary of Accomplishments To Date -- Drosophila). \*\*\*Does this need to be updated since you have obviously done many more than the initial 100.\*\*\* Notably we are applying this same methodology to produce over 500 human BACs for ENCODE, where we have also extensively validated this method ([49-51](#_ENREF_49" \o "Poser, 2008 #151)) (Poser et al 2008 PMCID: PMC2871289, Hua et al. 2012 PMCID: PMC3374131, Human Encode 2012 PMID: 22955616).

The P([acman] vector has the advantage of containing a *miniwhite+* marker and an *attB* site for precise integration into the *Drosophila* genome. The PhiC31 integrase system ensures that a single copy of the BAC integrates into a well-characterized location in the genome. The P[acman] vector contains a copy number inducible origin of replication. Low copy numbers can be maintained to protect the long term stability of large BAC clones in bacteria; following induction, however, copy number can be increased, allowing us to easily recover large amounts of DNA when preparing BACs for injection.

Strain generation is then performed in the Celniker lab. For TFs on the X, 2nd, and 4th chromosomes we inject into line VK00033 (Bloomington stock number 42673), which has an AttP docking site on the third chromosome. For TFs on the 3rd chromosome we now inject into a stock from the Norbert Perrimon lab, AttP40, which has the AttP docking site on the second chromosome. For small BACs (<50kb) we inject 100-200 embryos, depending on the docking site used, using a concentration of 150 ug/ml. For large BACs (>70kb) we inject 300-600 embryos, depending on the docking site, at a concentration of 50ug/ml. Transformed lines are discovered by screening progeny for w+ eye color (yellow-orange to red). We attempt to generate at least two independent insertions for each experiment, in case one line fails PCR. Homozygous lines are constructed where possible. For lines that are lethal as homozygotes (~10%), a balanced stock is generated. The lines are PCR-verified to confirm that they contain the expected TF and that the transgene inserted in the correct attP-landing site.

Once generated, tagged fly lines are characterized for GFP expression patterns throughout embryogenesis and wandering third instar larvae. Based on information from modENCODE RNA-seq, in vivo GFP expression patterns, and the literature, we determine the time of highest expression of the factor for ChIP. Our previous observations have shown that greater than 85% of all TFs are expressed during embryogenesis (RPKM>1), and by adulthood virtually all TFs are detectable in at least one developmental stage using whole animal extracts ([13](#_ENREF_13" \o "Graveley, 2011 #136), [16](#_ENREF_16" \o "Arbeitman, 2002 #141), [55](#_ENREF_55" \o "Stolc, 2004 #48)). These fly lines are then sent to the White lab for growth, collection and ChIP.

In the White lab, samples for ChIP-seq production are prepared on a large scale from one of seven developmental stages. Embryos are collected in three developmental windows from 0-8 hr, 8-16 hr, or 16-24 hr after egg laying, aged on apple plates, dechorionated, washed, and processed for ChIP as described below. Representative samples are staged under light microscopy as previously described to ensure that the appropriate stages are present ([16](#_ENREF_16" \o "Arbeitman, 2002 #141)). Post-embryonic collections are divided into four developmental windows: wandering 3rd instar larvae (W3L), white prepupae (WPP), 3-5 day old adult males, 3-5 day old adult females. For post-embryonic collections, animals are cultured in flasks containing molasses media. As with embryo collections, animals collected from post-embryonic stages are washed to remove food and processed for ChIP as described below. We use ~80 mg of starting material per biological replicate regardless of stage, and we have successfully performed ChIP on all stages of fly development described above. Whole animal collections are made at 25°C and 50% humidity.

To perform ChIP, fresh embryos, larvae, or adult flies are simultaneously homogenized and treated with formaldehyde to crosslink transcription factors and chromatin proteins to their DNA substrates. The cross-linked chromatin is then isolated, fragmented by sonication to an average size of approximately 300-400bp, and stored in appropriate aliquots at -80°C for further experiments. Chromatin samples prepared above are used for immunoprecipitation of specific DNA/protein complexes. As with the worm ChIPs, two chromatin preps are used for IP and, prior to sonication and ChIP, a small amount of chromatin is removed from each, pooled and set aside as an “input” control sample. For the IP, chromatin is first incubated with a polyclonal antibody raised against GFP (or another epitope, if appropriate) followed by immobilization with Protein A or G beads. After a series of washes, chromatin is eluted from the beads, treated to reverse the crosslinking, and digested with proteinase K to release the DNA. DNA is then purified by phenol/chloroform extraction and ethanol precipitation to obtain the genomic fragments corresponding to the binding sites of a transcription factor. Similar to *C. elegans*, a single technician can carry out enough ChIPs per week to produce 3-4 datasets (TFs).

During the modERN project, we have improved upon the modENCODE ChIP-seq protocol by collecting chromatin at developmental stages in which the TF (transcription factor) being investigated is most highly expressed. Care was also taken to assay TFs at different developmental time points in order to create a resource that is as comprehensive as possible.

Importantly, we have successfully transitioned from using TF-specific antibodies on wild type animals to using transgenic animals with GFP-tagged TFs engineered into BACs, allowing a GFP-specific antibody to be used in a uniform ChIP experimental pipeline. Additionally, we successfully modified our workflow by collecting additional replicates of fixed chromatin for immunoprecipitation (IP) and library preparation, rather than two replicates as in the previous pipeline, due to the time and effort required grow each transgenic line (thus improving efficiency of the pipeline). Analogously, we routinely sequenced two to three IP libraries, in addition to an input library, to analyze using an automated data processing and evaluation pipeline. If one of these replicates failed, this method allowed us to quickly reprocess additional replicates, minimizing the effort required to triage experiments. Furthermore, it minimized batch effects as each additional replicate was collected and ChIPed in parallel.

B2.1.3 ChIP library construction and sequencing strategies

Both worm and fly ChIP samples are sequenced in the White lab on Illumina HiSeq 2000 sequencers. All DNA samples, control and IP, obtained from the ChIP step are converted to libraries for Illumina sequencing using the Nugen Prep Kit. DNA is end repaired to create blunt ends, 3’ adenylated, then ligated to adapters that are compatible with Illumina sequencing. Importantly, these adapters are indexed (barcoded), allowing for sequencing of multiple samples in a single lane of a flow cell. After adapter ligation, samples are size selected, amplified by PCR, and purified. DNA libraries from this step are then sequenced on an Illumina HiSeq 2000 sequencer.

With the multiplexing strategies in place, samples from worms and flies can be combined in a single lane, so we don't have to wait for sufficient samples from one group to run a lane. This has increased turnaround during the period of the modERN project. One technician can generate 48 sequencing libraries per week. We initially multiplexed 8 libraries per lane of a flow cell and obtained sufficient genome coverage for both worm and fly ChIP-seq purposes, sowith 8 lanes on a flow cell a total of 64 libraries are needed to run a full flow cell. Thus, we are able to generate enough libraries to run a full flow cell after 1.5 weeks and, with a 3 day sequencing turnaround, sequencing data are available approximately 2 weeks after library production begins, assuming 64 samples are available for library preparation.

We made several improvements to the sequencing pipeline during this granting period. After testing several NGS (Next Generation Sequencing) library methods, we found that the Nugen Ovation Ultralow Library System yields the highest quality data. We also introduced multiple additional steps of quality control into our pipeline. Intermediate libraries are run on Bioanalyzer DNA 1000 chips and final purified libraries are run on DNA high-sensitivity chips (Agilent) and quantified by qPCR, to determine the concentration of sequenceable fragments within the library. We replaced the Invitrogen E-gel size selection with SPRIselect beads (Beckman Coulter). A wider range of fragment sizes can be purified using this bead purification, maximizing the amount of purified DNA, as well as removing human bias in choosing the optimal sized fragments in the final library.

With the ongoing improvements to the NGS platforms, we continue to optimize the number of libraries we’re able to pool, decreasing per sample cost. We began the project by pooling between 8 and 10 libraries. We are currently pooling 16 libraries, and plan to increase this further soon. All these improvements have helped us increase our production efficiency while still generating high-quality data at a reduced cost. We will continue to optimize our production pipeline, as we have during the modENCODE and modERN projects, making them increasingly efficient.

B2.2 Summary of achievements and implemented optimizations of RNA-seq (Aim 2)

During this funding period of the modERN project, we established a functional system to define gene expression profiles in animals with loss-of-function of particular TFs in both worms and flies. We have successfully collected gene expression profiles for 30 worm factors, and XX fly factors, as described in the next sections.

B2.2.1 Streamlined RNA-seq data production pipeline for worms

At the beginning of the modERN project, we selected 125 factors whose expression is known to be heavily induced during a particular stage of the life cycle and about half of these have fly or human orthologs (Appendix I). RNA-seq data shows that most of these rapid onsets occur in the process of embryogenesis. These factors aren’t present in early embryos (i.e., aren’t maternally loaded), but become expressed in later embryos (i.e., are zygotically induced). Importantly, the embryonic DNA binding profiles for these factors have been and are continuing to be collected as part of Aim 1. We originally planned to use RNAi to reduce expression of these TFs in embryos, but quickly realized that the strategy of “feeding RNAi”, which is the most commonly used, produced variable knockdown of the TF. We therefore switched to using deletion mutants of the TFs instead. The Reinke lab collects deletion mutant strains that are available from either of two resources, the CGC, and the Mitani group. These strains are backcrossed to a selected wild type strain 4-6x to remove most background variants from the genome and optimize strain growth. We then isolate total RNA just before, and at two time points after, the time of normal TF expression (Figure). The selection of these timepoints represents specific responses to loss of TF function. The first, collected prior to TF expression, provides the baseline gene expression profile. The second, collected at peak TF expression, provides the early response genes whose expression is most immediately affected by loss of TF function. The third, collected after peak TF expression, should capture more genes whose changes were too subtle or whose expression levels were too low to see in the second timepoint, but might also include secondary, indirect changes. During the modERN project, we realized that matching these profiles to a wild type timecourse was difficult due to subtle changes in developmental timing that were not apparent phenotypically but were apparent at the level of gene expression. We therefore adjusted our strategy by collecting “double samples” at each timepoint with a 30 minute spacing between them (Figure), which serve as internal replicates to better match to the wild type timecourse and improve the identification of differentially expressed genes (described below). The isolated RNA is analyzed for quality on an Agilent 2100 Bioanalyzer and is then sent to the Waterston lab for library preparation and sequencing.

B2.2.2 Streamlined RNA-seq data production pipeline for flies

For the fly TF function studies we have chosen 125 TFs to study based on their role in embryogenesis, correlation with ChIP experiments and their orthology with worm and human TFs. The Celniker lab developed a pipeline to generate animals with TF loss-of-function phenotypes to study the affects on the transcriptome. For the fly TF knockdown experiments, we are taking advantage of the large collection of RNAi lines being generated by the Transgenic RNAi Project (TRiP) at Harvard (see letter of support from Norbert Perrimon). The current TRiP collection includes 465 TFs in the Valium20 vector, which allows expression of the RNAi construct in both soma and germline, and more lines are being generated at the rate of about 300 per month. Of the 320 GFP tagged TFs we have generated for ChIP studies to date, 195 are also represented in the TRiP RNAi collection. These lines express RNA silencing short hairpin structures under the control of UAS, and therefore are activated by the presence of a Gal4 driver. To generate embryos with knockdown of a specific TF, we cross females expressing a ubiquitous Gal4 driver under the control of the *daughterless* gene (da-Gal4) to males from the respective homozygous UAS RNAi line. For all experiments, three embryonic time points are collected in duplicate: 0-1.5 hours post egg laying (before zygotic expression begins), a two-hour collection at the point of peak expression (specific to each TF), and 16-18 hours post egg laying (late embryogenesis).

Parallel control embryo collections are made from a cross of the of da-Gal driver females to UAS RNAi males where the RNAi target is an mCherry gene, which is not part of the fly genome, so that any differences seen in the RNA-seq data are likely to be specific to the knockdown of the TF and not just a general effect of RNAi activation. To increase embryo collection and sequencing efficiency we group RNAi knockdown embryo collections of two to three TFs with the same peak expression period along with the parallel control (18-24 samples per set). Embryos are flash frozen in liquid nitrogen and stored at -80 degrees C until all collections from a set are complete. Following protocols established for modENCODE, total RNA is prepared from the frozen samples by extraction with Trizol followed by purification using Qiagen RNAeasy. The extracted RNA is sent to the Waterston lab for library generation and sequencing.

B2.2.3 RNA library construction and sequencing strategies

All sequencing of both worm and fly RNA-seq libraries is performed by the Waterston lab. By combining samples from both labs we can more effectively utilize the capacity of the sequencing instruments and improve turnaround time.

*C. elegans* RNA is treated with a Ribozero (Epicentre) kit to subtract rRNA, followed by cDNA synthesis and library prep. Ribozero provides two advantages: we can work with smaller amounts of RNA than required for poly A selection (<100ng vs. 10ug), which allows us to collect more temporally precise populations, and we can better assess the effects of TF loss on non-polyadenylated noncoding RNAs. After cDNA synthesis, the cDNA is sonicated to a target size of 300 bp, then undergoes end repair (NEB), A-tailing, and Y-adapter ligation. qPCR is then used to amplify the cDNA libraries and to label each library with one of 95 available multiplexing indexes. The amplified libraries are run on a PAGE gel, from which 350-550 bp fragments are selected and extracted.

Drosophila RNA is checked for quality on an Agilent 2100 Bioanalyzer. The samples are then poly A selected and prepared for sequencing using the Illumina TruSeq Stranded mRNA LT Sample Prep Kit. Ribozero rRNA depletion is not used with the fly samples because the available kits are not compatible with fly RNA. The TruSeq kits provide an additional advantage of preserving information about the strandedness of the RNA samples, which is compatible with previous datasets collected by the Berkeley group. The kits produce cDNA fragments with a median insert length of 155 bp and allow for multiplexing of up to 24 different libraries.

The libraries from both organisms undergo an initial quality control round of sequencing to check for successful library preparation and sufficient complexity. Once the libraries pass this step, they are more deeply sequenced to a target depth of 20 million 75 bp paired end reads per library, a depth is sufficient to provide reasonable estimates of expression levels for most transcripts. The quality control round was originally performed using the Illumina MiSeq system, which produced ~20 million single end 50 bp reads per run. The deep sequencing was then done on a HiSeq 2500 machine with other pooled samples. We have recently transitioned to using the NextSeq 550 system with mid- and high-output kits for improvements in both time and cost efficiency (details in Innovations section) and, with paired end data, allows more accurate quality assessment including estimates of PCR duplicate rates.

To avoid possible problems from PCR duplicates, we have begun to implement some small but important changes in our pipeline. We are increasing the starting material in the Ribozero reaction from 150 ng to 500 ng where possible, increasing the complexity of our libraries as well as decreasing the number of PCR cycles that are needed for amplification. We have also modified the TruSeq protocol to use qPCR instead of standard PCR so that libraries to monitor the amplification.

Other improvements in streamlining our sequencing pipeline include switching from MiSeq to NextSeq instruments. Previously, we used MiSeq kits for quality control because they were cheaper and quicker than HiSeq runs. However, with the MiSeq we could only get single-end data, which could not be combined with the paired end data for DE analysis. With the acquisition of several NextSeq 550 machines. we are now able to use the NextSeq mid- and high-output kits to do both quality control and deep sequencing runs. This has the advantage of producing usable data from the quality control run, as well as decreasing turnaround time, since a NextSeq high output kit takes about 18 hours for a 75 bp paired end run.

B3. Specific Aim 1. Complete binding site discovery for remaining TFs

The majority of our effort (~75%) will be concentrated on continuing to perform ChIP-seq on tagged TFs for both *C. elegans* and *Drosophila*. Steps specific to the individual species, such as strain generation, characterization, growth, and ChIP will still be performed separately by labs expert in *C. elegans* and *Drosophila* techniques, respectively. However, the same well-vetted anti-GFP antibody will be used for ChIP in both organisms, and all libraries and sequencing will be performed at a single sequencing center. Data quality assessment and initial analysis will be performed before depositing to the ENCODE database, as before (described in Section C). Here, we describe our plans to continue the current ChIP-seq pipelines outlined above in order to complete the binding site identification for the remaining ~300 TFs in each organism, along with further optimization and also new strategies to capture the more difficult and refractory factors.

B3.1. Worm-specific experimental pipeline

B3.1.1. Strain generation (Waterston lab)

For those 154 factors with a recombineered fosmid available, the Waterston lab will continue with the recombineering strategy described in the Progress Report (Bx.x), since these constructs continue to produce a reasonable rate of successful transgenic lines. The transgenic, tagged factors recapitulate known expression patterns of the endogenous factors, and rescue mutant phenotypes (6 out of 6 tested). Most importantly, the binding site profiles of the tagged factors generally match those of endogenous factors ([11](#_ENREF_11" \o "Zhong, 2010 #10), [26](#_ENREF_26" \o "Niu, 2011 #143)); Kudron et al., 2013). Having these constructs ready to go will allow us to continue producing binding site datasets for additional factors at the current rate, while we address factors lacking tagged fosmid clones.

A total of 184 factors are not captured in the fosmid library, and/or have not been successfully recombineered to include a GFP-3XFLAG tag. Additionally, we expect that a small fraction of recombineered fosmids fail to yield expressing transgenic lines. For these factors, we will utilize a CRISPR/Cas9 strategy to insert the same GFP-3XFLAG tag into the carboxy-terminus of these TF genes. For cases where carboxy-terminal tagging has repeated failed to yield an expressing line, we will try inserting GFP-3X FLAG tag at the N-terminus.

To obtain GFP tagged strains for TFs that lack fosmids, we are using the SapTrap method (Schwartz and Jorgensen, 2016; PMID:26837755) in which CRISPR is used to insert GFP at either the 5’ or 3’ end of the endogenous gene. Earlier, we had success with a similar method that employs Gibson assembly to create GFP insertion cassettes (Dickinson et al, 2015; PMID: 26044593), but the time needed to design and create the homology arm cassettes in this method is substantial. Also, the method uses PCR based techniques to generate the homology arms, which can be inefficient and requires additional design steps to eliminate the sequences that are targeted by the sgRNA.

Instead in the SapTrap method (Schwartz and Jorgensen, 2016), modular plasmid assembly is combined with simple annealed oligos specific to the gene of interest to create GFP insertion cassettes rapidly and efficiently (Figure XX). Separate plasmids carrying sequence for GFP, for a flexible linker, and for an sgRNA targeting sequence are digested with SapI, a type II restriction enzyme that cuts outside of its recognition sequence, leaving 3-base overhangs each with distinct sequences. Synthetic oligonucleotides are designed for the guide RNA (sgRNA) and the 5’ and 3’ homology arms, with each having the desired 3-base overhangs. The latter are designed to alter the sequences recognized by the sgRNA. The resulting DNA fragments can then be mixed in one-step reaction, with only the desired sequences annealing and ligating in a modular fashion to complete assembly of a single plasmid with all the desired components. Additionally, the screening process for identifying successfully CRISPR edited animals is optimized for both injection and bombardment based approaches due to the use of an *unc-119* rescue module within a synthetic intron in the GFP sequence. We will compare the ease and efficiency of generating tagged lines using injection and bombardment protocols and use the most effective protocol.

Based on our experience to date, we expect that roughly 10-15% of predicted TFs will have only cytoplasmic expression. With this knowledge we will be able to refine our list of worm TFs. Another smaller fraction (~3%) will have only very few cells expressing the GFP at low levels. If we are able successfully to implement CUT&RUN and it does indeed provide better signal to noise data, we will proceed with TF peak analysis on these strains as well as similar strains put aside previously.

B3.1.2. Growth of strains and ChIP (Reinke lab)

The Reinke lab will continue to perform the strain growth and ChIP as described in the Progress Report. As we receive each strain from the Waterston lab, we examine it for expression and record our findings, collecting fluorescent images (which are included in the ENCODE database) of the key developmental stages at which ChIP will be performed. We also freeze the strains so that we have a backup to both UW and to the CGC. We then organize how to most efficiently grow the strains (grouping strains to be collected at particular stages), and grow them on large plates so that they are synchronized by developmental stage, collecting the worms at the selected stage for ChIP. With the improvements to chromatin isolation described in the Progress Report, we expect that much less starting material might need to be collected in order to have a sufficient amount for ChIP. This will reduce the amount of effort that will have to be put into strain growth, and should speed our production efforts.

We will continue to use the anti-GFP antibody that we have been using for the last four years for chromatin immunoprecipitation, which is also used by the Drosophila ChIP group. We will also continue to test various adjustments to the ChIP protocol, including other ways in which input chromatin amounts can be reduced, such as [[look up]]. We will also investigate the use of CUT&Run, a recently published method that uses micrococcal nuclease tethered to proteinA to localize the DNase activity primarily to sites where antibody binds (Skene and Hennikof, 2017, PMID: 28079019 ). The result is increased sensitivity and lower background. The lower background may allow us to detect additional peaks even in whole animals for factors such as MEC-3, which is present in only ten cells among the 558 embryonic cells (Wray and Chalfie 1989, PMID: 2576011). Alternatively, the increased sensitivity may allow us to use FACS cells as the source material, greatly enriching for signal over noise. In addition, the ability to ChIP-seq on sorted cells would allow us to determine if factors such RNT-1 bind distinct sites in the different cell types in which it is expressed. To do this we would use a second marker to label the distinct cell types. Pete Skene and Steve Henikoff have offered to help us apply their method to *C. elegans* and *Drosophila* (see letter).

[[In addition to these efforts, we will look at histone modification H3K27Ac at all developmental stages because it will help define peaks. We know how to do it, it should be straightforward, and we can add more if feasible.]]

B3.2. Fly-specific experimental pipeline

B3.2.1. Strain generation

[[need some discussion of how we will continue to do things and what we will do for final TFs that are difficult]]

B3.2.2. Growth of strains and ChIP

For the majority of TFs for which we have generated ChIP-seq data, we are collecting chromatin at developmental stages with the highest TF expression in the modENCODE RNA-seq data and this has been satisfactory. Additional published data will be essential to determine the best approach for difficult factors, e.g. if they are expressed in only a few tissues or require induction.

Whole organisms are collected from one of four developmental stages with minor deviations based upon expression levels. Embryos are collected on apple plates, dechorinated and washed prior to processing as described below. Embryo cages are incubated overnight at 25°C prior to collection. Post-embryonic collections of wandering 3rd instar larva, white prepupa or adults are cultured in flasks with molasses media at 20°C and 50% humidity and washed prior to processing. For each stage, 100 mg of material is collected for each of four replicates.

Organisms are simultaneously homogenized and treated with formaldehyde to crosslink DNA with nearby proteins. This chromatin is then isolated and fragmented by sonnication with a Diagenode bioruptor to fragments smaller than 1kb and stored at -80°C. An equivalent portion of each chromatin is retained and combined to be used as an input sample. For the IP, chromatin is first incubated with a polyclonal antibody raised against GFP (or another epitope, if appropriate) followed by immobilization with Protein G beads. After a series of washes, chromatin is eluted from the beads, reverse crosslinked, and digested with proteinase K to release the DNA. DNA is then purified using QIAGEN MinElute columns to obtain the genomic fragments corresponding to the binding sites of a transcription factor.

Besides the strains that tagged with site-specific transcription factors, we will collect wild type fly for each of the four developmental stages following the same growth condition. These wild type samples will be used to conduct mock-IP with GFP antibody and H3K27Ac chromatin IP following the same ChIP protocol as described above. We discussed the necessity of doing these experiments in **B3.3**.

B3.3 Joint ChIP sequencing pipeline for worm and fly

Two to three IPs, along with their corresponding input are used to make libraries utilizing the Nugen Ovation Ultralow kit. In order to balance the cost and efficiency, libraries are checked on an Agilent Bioanalyzer DNA1000 chip and successful libraries are further passed to size selection for fragments from 200-1000bp using SPRIselect beads from Beckman Coulter. Libraries are again quantified on a DNA1000 chip in order to not overload the Agilent High-Sensitivity chip. Equal molar amounts from each library are pooled and sequenced on Illumina HiSeq 4000 sequencer.

We will continue to optimize our production pipeline, as we have during the modENCODE and modERN projects, making them increasingly efficient. We plan on continuing our history of technical innovation to make several improvements. (1) we will generate mock-IP data for each developmental stage we have used. Recently we observed higher numbers of peaks with a greater extent of co-localization between TFs than had been previously seen in the modENCODE project. We propose to do wild type mock-IP to better distinguish the true biological signals from technical bias. (2) We will generate H3K27Ac data for each developmental stage to improve peak calling. Most of the H3K27Ac data from the modENCODE project was generated by ChIP-chip. More importantly, these data were collected from only from a limited number of developmental stages with different sources of antibodies. We intend to perform ChIP-seq for H3K27Ac in each of the four fly developmental stages and the six worm developmental stages used in our standard modERN project pipeline. These additional experiments will only require a small amount of extra effort, but will be very useful for enhancer prediction and data interpretation. (3) We will introduce additional multiplexing barcodes to utilize the increased sequencing output of the HiSeq4000 platform. (4) We will explore more sophisticated immunoprecipitation approaches such as high resolution definition of TF binding sites (e.g. ChIP-nexus [He et al. PMID: 25751057]).

B4. Aim 2. Identify biological relevance of TF binding sites using gene expression profiles

The remainder of our effort (~25%) will be concentrated on continuing to perform RNA-seq on strains with TF loss-of-function for both *C. elegans* and *Drosophila*. Steps specific to the individual species, such as mutant backcrossing, RNAi, strain growth and RNA isolation, will still be performed separately by labs expert in *C. elegans* and *Drosophila* techniques, respectively. However, all library construction and sequencing will continue to occur at a single sequencing center, at UW. Data quality assessment and initial analysis will be performed before depositing to the ENCODE database, as before (described in Section C). Here, we describe our plans to continue the current RNA-seq pipelines outlined above in order to further understand the relationship between TF binding and gene expression, along with further optimization and also new strategies to investigate more tissue-specific profiles.

B4.1. Worm-specific experimental pipeline

We will continue to examine the effects of TF loss-of-function strains to look for associations between TF binding sites and changes of expression in likely target genes. Based on experience with RNA-seq analysis of strain with deleted TFs we will modify our strategies to increase the informativeness of our experiments. First, we will study essential TFs using RNAi, or where available ts alleles (or both), again focusing on those that are zygotically expressed and lack a maternal component. We will assay their effects on gene expression using RNAi or ts alleles starting from when they are expressed until two hours after peak expression. (should we start with balanced lethal strains when available to knock down at least half the RNA? Or will the balancer confuse things?) We will verify that the RNAi is reducing expression by visually assessing the impact of the RNAi on our GFP tagged strains, by performing qPCR on the treated strain compared to wild type and monitoring the phenotypic effects.

Secondly, we will use FACS to enrich for cell types in which the deleted TF is expressed, testing for the possibility that changes in gene expression in one cell population are masked by expression in other cell types in which the regulation is controlled by other TFs. We will mark the desired cells with the TF promoter::GFP fusion if available and expressed in the same cell types as the protein fusion. (We cannot use the protein GFP fusion for obvious reasons.) We have the expression patterns already determined for xx TF promoter fusions (Murray et al., PMID: 22508763; Waterston, unpublished results) and can readily generate additional expression patterns for a gene of interest. Alternatively, we can use existing promoter fusions whose expression patterns contain the cell types of interest, choosing the most appropriate among the more than 200 patterns we have determined. In some situations we may use doubly marked strains, one using a promoter fusion of the deleted TF and a second marking a subset of cells in which the TF is expressed. This strategy will allow us to determine if the TF is targeting different genes in different cell populations. We have extensive experience in performing RNA-seq of FAC sorted cell types from synchronized embryos, using essentially the same protocols we use on whole animals (Adam Warner and Waterston, unpublished). We have even been able to sort cells late in embryogenesis (after the cuticle is formed) by incorporating post-embryonic methods (Zhang et al, 2011, PMID: 21559335, Spencer et al., 2014 PMID: 25372608 ) with minor modifications. We can reliably recover sufficient cells for RNA-seq from markers expressed in only 2% of cells.

Thirdly, we will look at strains lacking two TFs to examine the role of redundancy in the control of gene expression. Our ChIP-seq data shows the co-association of several TFs above that expected by chance (See Cx.x). For example, UNC-120 clusters with CEH-18 more than 10-fold greater than by chance. Similarly, PQM-1 clusters with ELT-2 8-fold greater than by chance and CEH-42 clusters with CES-1 10-fold greater than by chance. We will construct double mutants of the deletion strains, where available and evaluate the phenotype. We will also perform RNA-seq to look for greater effects on gene expression than by either mutant alone. If the double mutant is dead, we will combine RNAi with a deletion mutant to achieve the double knockdown.

Lastly, we will directly assay a sample of DNA sequences identified by clusters of TF binding sites for enhancer activity. We will insert synthetic oligonucleotides upstream of Ppes-10, a minimal promoter, driving a codon optimized GFP with 2x NLS (*egl-13*/SV40) in a vector for mosSCI insertion at the ttTi5605 site, carrying the *unc-119* gene for selection (REF). The “naked” vector inserted as a single copy gives only very dim expression in the head while single copy enhancer constructs give bright expression (Frokjaer-Jensen, personal communication). We will test clusters that contain combinations of binding sites for TFs with known tissue/cellular specificity as well as clusters that contain clusters of TFs with less well characterized specificity. We will also evaluate clusters of increasing complexity. We will evaluate embryonic expression patterns by comparing them to patterns of determined by our lineaging based methodology () and in difficult cases we will take full movies and establish the precise cellular patterns with minute resolution. Because over the past two years we have installed a di-SPIM microscope () and have developed software to analyze these images, we can now follow expression up to the point of animal movement.

We expect to assay 10 essential factors in the initial year, along with 5 factors using FACS and 5 double mutant combinations. We will also expect to be able to create transgenic lines for 10 potential enhancers and evaluate their expression patterns. Depending on the results we obtain, we will either continue with this mix or alter the mix to obtain the most information about the function of the binding sites.

B4.2. Fly specific experimental pipeline

[[needs more info and set up]]

For a limited number (<10) of high priority TFs for which there is no available RNAi line, or where the RNAi line is homozygous lethal (approximately 5% of the lines), we are developing an alternate protocol using lethal, loss-of-function mutants and single embryo screening (PMID:17068956). We first establish lines of each mutant chromosome with a GFP marked balancer. We can then use PCR primers that identify the presence of the balancer chromosome marked by GFP and another set of control primers to detect genomic DNA. Individual embryos are homogenized in extraction buffer, with a portion taken for PCR and the remaining extract transferred into Trizol and frozen in a 96-well plate for RNA isolation after homozygotes are identified. Trizol extracts for approximately 80 homozygous mutant embryos are pooled to prepare sufficient RNA for sequencing (2 ug per sample).

Finally for a limited number of TFs with rare cell type expression patterns it may be necessary to isolate specific cell populations. This has been successfully done for Drosophila embryos PMID 25151164, larval imaginal disc PMID 27057746 and adult midgut intestinal cells PMID 26237570. We plan to isolate cells and interrogate the transcriptional profile from wild-type and TF mutant animals. \*\*\*Anything about FACS?\*\*\*

B4.3. Joint RNA sequencing pipeline for worm and fly

We will continue to perform RNA-seq as described above, incorporating the improvements that we have developed. \*\*\*Need to add more here.\*\*\*

[[put in here how we will continue as before, but add any changes to libraries or sequencing]]

B5. Milestones

Our plan is ambitious, aiming to complete the binding site analysis for approximately 600 transcription factors in the course of 4 years. This requires a pace of data acquisition at or slightly greater than what we achieved in the modERN project. However, we have generally increased our throughput through the course of the modERN project (barring maternity leave from one key member of the worm ChIP team). By further improvements to our pipeline as described above, we are confident we can achieve these goals. Our detailed milestones are presented below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Milestones** | **Organism** | **Year 1** | **Year 2** | **Year 3** | **Year 4** | **Year 5** |
| Genes with transgenic lines | Worm | ? | ? | ? | ? | ? |
| Fly | ? | ? | ? | ? | ? |
| Whole animal ChIP-seq | Worm | ? | ? | ? | ? | ? |
| Fly | ? | ? | ? | ? | ? |
| RNAi/RNA-seq | Worm | ? | ? | ? | ? |  |
| Fly | ? | ? | ? | ? |  |

C. Informatics Core

C1. Specific Aims

C2. Progress Report

C2.1 Summary of achievements and innovations in analyzing ChIP-seq data

C2.1.1 Analysis of hundreds of ChIP-seq datasets for quality control

We have implemented an in-house ChIP-seq processing pipeline, in order to uniformly process all our data, as well as to reprocess the ChIP-seq data generated during modENCODE. Our pipeline closely mirrors that of ENCODE, using similar software and generating common QC metrics. Raw fastq files are aligned to the reference genome (for fly: dm3 and dm6; for worm WS220 and WS245) using BWA. Aligned reads are scored for mappability and PCR duplication rate. High quality, unique reads are fed into SPP [Kharchenko PV et al, PMID: 19029915] to call peaks; with rare exception MACS2 has been used when broad peaks are expected. Peaks are scored using IDR with threshold 0.01. A validated dataset consists of ChIP-seq experimental replicates with: (1) a self-consistency ratio of less than two; (2) a rescue ratio less than two. After consulting ENCODE project members, we no longer use normalized strand coefficient minimum and relative strand correlation score. The thresholds previously used by modENCODE and ENCODE, were subjectively chosen based on Human ChIP-seq data, and do not translate to worm and fly due to their smaller genomes. Datasets that pass these initial uniform metrics and manual review are available through the UCSC Track Hub service for easy access across the modERN groups. Furthermore, successful datasets are submitted to the ENCODE DCC and released to public through the ENCODE central data depository (see **Section C5**).

A validated dataset consists of replicate ChIP-seq experiments that have met a normalized strand coefficient minimum of 0.5 and/or relative strand correlation score of >0.8, which is the standard utilized by modENCODE and ENCODE over the past several years ([77](#_ENREF_77" \o "Landt, 2012 #166)) (Landt et al., 2012). \*\*\*The earlier paragraph seems to imply that earlier filters were not appropriate for the smaller worm and fly genomes, but here it seem like we are using them. What am I missing?\*\*\*

C2.1.2 Analysis of HOT regions and clustering of TF binding sites in the genome

With more than 340 data sets for ~ 200 DNA binding proteins, several features begin to emerge. The binding sites are often tightly clustered, with almost 728,381 sites from 378 data sets involving 216 TFs and other DNA binding proteins generating just 34,684 clusters containing more than site and 41,926 single sites. (Sites were clustered by demanding that the peaks of sites within a cluster were no more than 40 bases from the nearest peak.) Further, 92.9% of those clusters spanned less than 200 bases. These clusters and single binding sites are not uniformly distributed – more than 2,000 genes have not sites between the end of the upstream gene and the start of the downstream gene. These genes without sites include many very poorly expressed genes. At the other extreme, many genes have multiple clusters as well as many single sites in the same region. The smaller clusters (< 20 sites) can sometimes contain sites for factors many of which are expressed within the same tissue. When this is the only cluster associated with the gene, the cluster can be highly predictive of tissue specific expression. For example, genes containing HLH-1 sites along with CEH-18, UNC-120, RNT-1 or UNC-62 sites are overwhelming expressed in muscle (Figure browser shot tnt-3). Using a multiple linear regression analysis, \*\*\*I need to get more here\*\*\*



Figure tnt-3 A cluster of sites near the body wall specific gene tnt-3 with HLH-1 and other sites associated with muscle expression

Another finding is that some factors, such as MEC-3 bind to relatively few sites outside of large clusters, and those sites are often among the lowest scoring sites (Figure mec-3). These results suggest that the ChIPseq performed on whole animals may have difficulty detecting sites present in only a few cells in the embryo.



Figure MEC-3 cluster size versus score Clusters containing MEC-1 sites show a paucity of high scoring sites (x-axis) in smaller clusters (y-axis).

Because of the prevalence of HOT sites, we wondered if we could detect such sites with just antibody alone.  We performed mock-IPs on lines without a tagged-TF while using the same antibody as in the transgenic IP experiments. Peaks observed in these mock-IP data significantly overlapped with the HOT regions from both projects, suggesting that non-specific interactions of the antibody with these sites could be contributing to the signal. We suspect that these peaks overlap true TF binding sites, and thus do not simply want to remove them. Instead, we are developing more sophisticated approaches to remove the mock-IP signal. \*\*\*Can you provide some specificity here about what you might do?\*\*\*

C2.2 Summary of achievements and innovations in analyzing RNA-seq data

C2.2.1 Identification of differentially expressed genes in TF mutant worms (Waterston lab)

In the worm transcription modENCODE project (ref PMID: 25164755), the Waterston and Reinke labs were among the first to apply RNA-seq to transcriptomes. Over the course of the project, we assayed more than 150 samples to characterize the protein coding and non-coding genes of *C. elegans* and those of 4 other *Caenorhabditis* species. This effort identified new genes, new first exons, new 3’ UTRs, revised exon boundaries and tens of thousands of new splice junctions, as well as determining gene expression levels for these genes. We therefore are highly experienced with analyzing RNA-seq data and have in place the data analysis tools to monitor changes in gene expression in the mutant and RNAi-depleted animals.

The informatics pipeline for generating the gene expression profiles is implemented in conjunction with an SQL database. The metadata for the samples and libraries are entered into the database. The samples are organized into datasets reflecting the time series nature of the samples. After sequencing, the raw data is archived in permanent storage with locations recorded in the database. Alignments are then performed with the STAR aligner (<https://code.google.com/p/rna-star>) and used by BitSeq (<https://code.google.com/p/bitseq>) to calculate expression profiles. The results of aligning and expression calculations are tracked and also stored in the database. After an entire time series has been processed, the metadata, alignments and expression profiles are uploaded to the ENCODE database. The process of uploading is fully automated and tracked by the informatics system.

Differential expression analysis of the worm RNA-seq data is performed using the DESeq2 (PMID: 25516281) package in R. For each mutant sample, Spearman correlations are calculated for expression of a subset of 6,407 genes, which are well-expressed and have large expression changes over the course of embryonic development, against a panel of 113 wild type RNA-seq samples. These wild type samples represent a range of 0 to 1200 minutes post-two embryo formation covering embryonic development (Boeck et al. PMID: 27531719; this grant). The correlations are used to find the best wild type match for each experimental sample. For the DESeq2 analysis, two timepoints are 30 minutes apart for each experimental time series are treated as “pseudoreplicates,” and the two most highly correlated wild type samples are treated as pseudoreplicates of a control series. rRNA reads are removed prior to providing read counts per gene used by DESeq2. For each time point (early, peak, and late TF expression), the control and experimental timepoints are compared to calculate differential expression.

We have recently incorporated a PCR duplicate estimation program into our pipeline (written by Brent Ewing in the Green Lab at UW) and have found a variable duplication rate in our data. We used this information to make improvements to the library construction and sequencing methods, as described in the Progress Report in Section B.

C2.2.2 Identification of differentially expressed genes in TF RNAi flies (Waterston lab)

[[any differences between the worm and fly RNA informatics pipeline?]] Sue, can you fill this in, since up till now you have been doing this and we will try to replicate it?\*\*\*

C3. Aim 3 Processing and analyzing sequencing data

C3.1. Assigning peaks for ChIP-seq data

We have extensive experience in peak calling, having developed PeakSeq \cite{19122651}, one of the standard peak calling programs used by the ENCODE and modENCODE consortia for various ChIP-Seq datasets. More recently, we developed another peak caller, MUSIC. It performs multiscale decomposition of ChIP signals to enable simultaneous and accurate detection of enrichment at a range of narrow and broad peak breadths. This tool is particularly applicable to studies of histone modifications and previously uncharacterized transcription factors, both of which may display both broad and punctate regions of enrichment, which will be generated in this project. We plan to use MUSIC to call peaks for the histone modification data generated in this proposal. Moreover, unlike most previous ChIP-Seq experiments, the experiments in modERN have recently produced controls from both DNA inputs and mock immunoprecipitations (IP). To take advantage of both controls, we plan to develop new statistical models to identify TF binding events. More specifically, due to the increased model complexity, we plan to use Markov chain Monte Carlo (MCMC) methods to estimate the posterior probability of TF binding. To our best knowledge, this is the only method that designed to consider multiple control types. These state-of-art ChIP-Seq data and methods should substantially improve TF binding identification. \*\*\*This sound like it belongs in what we will do.\*\*\*

C3.2 Analyzing HOT regions

Apart from promoters and enhancers, an important annotation for the ChIP data is the HOT (High Occupancy Target) regions. We have pioneered several research areas on HOT regions in the genome of human, worm and fly. Using co-binding information for more than 100 transcription factors annotated by ENCODE, we developed machine learning methods that capture the genomic features of HOT regions, among other annotations \cite{22950945}. In the modENCODE project, we annotated HOT regions in human, worm and fly \cite{25164757}. Making use of the ChIP-seq peaks identified for the large number of transcription factors generated by this project, we plan to build a pipeline for annotating HOT regions in worm and fly across different stages. We will further compare HOT regions between different developmental stages, and across the species based on orthology relationships. Making use of the publicly available Hi-C data on worm and fly, we aim to understand the role of HOT regions in relation to spatial organization of genomes.

C3.3 Mapping RNA-seq reads and determining differential gene expression

The data collected for the three different types of RNA-seq/TF loss-of-function experiments are compatible with our established pipeline. We will continue to align the reads to the genome using STAR and calculate the expression levels of transcripts using BitSeq. We will use the reads aligned to the genome to determine the number of mapped reads per gene to feed to DESeq2. In addition we will incorporate the program developed in conjunction with Brent Ewing to remove likely PCR duplicated reads. Simply removing all duplicate reads (paired end reads having identical start sites in genome alignments) can unwittingly depress expression of highly expressed genes, where the high expression levels can result in true independent duplicate reads. Instead our approach utilizes signals inherent in the data to define the duplicates. Minikel’s Poisson distribution, for example, assumes independent samples; however, drawing a volume of ligated cDNA from the library solution corresponds to dependent samples so we instead use the hypergeometric probability density function:

where

is the number of reads sequenced (or read pairs in paired end data)

is the number of read duplicates sequenced

is the hypergeometric probability density function

is the number of read duplicates of a particular insert (template?)

number of duplicates of each insert in the library ()

number of inserts before PCR amplification

number of inserts after PCR amplification

number of inserts sequences (reads for SE or read pairs for PE)

Based on this, we can estimate the duplicate read fraction dependence on fraction of inserts sequenced for a given library. Having this tool allows us to both assess library quality and to thus improve both our initial expression estimates and our ability to detect differentially expressed transcripts.

Having found differentially expressed genes, we will evaluate the TF binding sites associated with them. We will also use our increasing data sets of tissue and cell-type expression patterns to see if the genes with altered expression are consistent with the known expression pattern of the perturbed TF. \*\*\*Add control analyses here from Sue’s section.\*\*\* These data will be essential for helping us to recognize functionally active sites.

C4. Aim 4. Integrating data, developing temporal-spatial regulatory networks and a mini-encyclopedia

C4.1 Integrating ChIP- and RNA-seq data

To generate something comparable to the “Encyclopedia” produced by ENCODE, we will develop a mini-encyclopedia that will consist of different levels of annotation. The ground level annotations include peaks derived directly from our TF ChIP-Seq data, some of the existing modENCODE chromatin data, as well as the new H3K27ac data. The ground level peaks will then be integrated to predict regulatory elements like enhancers. We have developed a set of tools to identify enhancers and the genes they regulate. First, we have applied machine-learning methods that integrate multiple genomics features, for example, using ENCODE data of >100 TF binding sites, to classify regulatory regions. A particular class of regions is the gene-distal regulatory modules comprised of potential enhancers \cite{22950945}. \*\*\*What is the relevance of this to the compact genome of the worm, where almost all studies have shown that upsteam region and in some cases the first introns are sufficient to generate the native expression pattern? \*\*\* More recently, using signal processing techniques, we have developed another machine learning model to identify enhancers based on histone marks such as H3K27ac. The method has been used for enhancer prediction in human, mouse and fly. We plan build a pipeline specific for the purpose of this proposal. The enhancers predicted by our pipeline will be an important resource for the model organism community, and there will be many research avenues. \*\*\*This doesn’t say anything.\*\*\*

A higher level annotation is the linkage between enhancers and targets. We have much experience in using these information \*\*\*Beyond the grammar, what does this mean?\*\*\* for functional annotations. First, we will annotate the regulatory roles of genomic regions according to their TF binding and histone modification status, using our statistical models \cite{22950945}. We developed the Function-based Prioritization of Sequence Variants (FunSeq) tool \cite{24092746} for identification of candidate drivers in tumor genomes, and more recently, a more elaborate and flexible framework for this tool. \cite{25273974}. A possible project \*\*\*This is too tentative. What will we do here?\*\*\* is to apply similar machine learning models for understanding the potential functions of these enhancer candidates, generating the linkages between enhancers and their target genes. \*\*\*This seem overkill in the worm, where the assignment for half of the genes seems trivial (for head to tail genes, the upstream region targets the downstream gene. For the other half of genes – head-to-head – proximity is a good rule of thumb. \*\*\* Overall, the mini-encyclopedia of model organisms will be compatible with the ENCODE Encyclopedia and facilitate the comparison between human and model organisms.

C4.2 Developing temporal-spatial regulatory networks

The prediction of promoters, enhancers, and their target genes will lead to a draft of comprehensive regulatory networks. The network representation of the regulome will form the high level annotation in the mini-encyclopedia. The networks will be further refined by the transcriptome changes due to the TF knocked out/down. We will also use the transcriptome changes to understand how TFs contribute to gene expression, and thus improve our model predicting gene expression \cite{21926158}. More intriguingly, we have will have double deletions of TFs. Together with their single deletions, these data can reveal the interactions among TFs regulating gene expressions. For spatial networks, we further aim to develop a deconvolution method to resolve binding profiles at different tissues, similar to what has been done in expression data obtained from a mixture of cell lines. We will compare the refined networks in regard of their topological structures. \*\*\*What does this mean?\*\*\* We will further compare the better-studied human regulatory network with the networks we obtain in this project. Making use of the orthology relationships, we will identify the conserved components across these species and specific components that are unique to worm or fly. Information of TF-gene interactions in human could also be used to refine our networks.

C4.2 Developing temporal-spatial regulatory networks

To mimic the “Encyclopedia” produced by ENCODE, we plan to present our resource as a mini-encyclopedia. We will organize our data and analysis into different levels of annotations. The ground level annotations include peaks derived directly from our TF ChIP-Seq data and existing modENCODE histone modification ChIP-Seq data. To have a histone modification data consistent with our TF data, we seek in this proposal to perform extra ChIP-Seq experiments for a key chromatin mark, H3K27ac, for all developmental stages in worm and fly. The integration of the mark with existing chromatin accessibility and other histone modification data, as well as the TF binding sites generated by this proposal, will better predict the functional role of a genomic region, forming the middle level of the encyclopedia such as promoters, enhancers and HOT regions. Finally, based on the linkage between regulatory elements and target genes, the regulome will be presented in a network fashion. The representation will form the high level annotation in the mini-encyclopedia. Overall, the mini-encyclopedia of model organisms will be compactable with the ENCODE Encyclopedia and facilitate the comparison between human and model organisms. \*\*\*This paragraph has too much redundancy with the previous one, doesn’t it? \*\*\*

C5. Strain and Data submission and dissemination

We generate two types of resources for the community, transgenic lines of worms and flies, and datasets of ChIP-seq and RNA-seq. We have made, and will continue to make, both available to the community as quickly as possible while not compromising on integrity or reliability of either resource. Currently we have been depositing transgenic strains in the respective model organism stock centers, the Caenorhabditis Genetics Center at University of Minnesota and the Bloomington Drosophila Stock Center at Indiana University. The advantage of this option is that these stock centers are well-known and heavily used, and the per-strain fee is nominal. Therefore, we will continue to utilize this resource, and on a quarterly basis will send them the strains that have been produced and are expressing, regardless of whether ChIP-seq has been completed or not. Performing this at defined intervals simplifies strain tracking and makes the deposition process simpler for both the stock centers and ourselves, and minimizes costs.  For worms, strains are submitted to the CGC using their standard protocol with the exception that we submit frozen stocks due to the volume of strains generated by this project.  Strains are processed into frozen glycerol stocks as they are generated and shipped to the CGC. We check on these strains on multiple levels. \*\*\*Need to fill in the various qc steps – comparing patterns with known patterns, tissue and cell specific RNA-seq data sets, the presence of the integrated fosmid for the worm in the ChIP-seq data. Other things?\*\*\*

The ChIP-seq and RNA-seq data will be distributed in various formats to multiple places. Primarily, the processing pipeline at University of Chicago directly deposits the raw and processed ChIP-seq data into the ENCODE database. The ChIP-seq datasets generated at University of Chicago for both organisms are sent through an established pipeline to the ENCODE database. Through the ENCODE database, the raw reads for both data sets are deposited in GEO/SRA [[or are directly available?]] \*\*\*Who can answer this? Alec?\*\*\* , as is the processed data, which includes called peaks and statistical analysis of data quality (the pipeline described above) for ChIP-seq data and expression levels for genes and exons for RNA-seq data (FPKM).

Additionally, we plan to facilitate the deposition of the processed data in the model organism databases, Wormbase and Flybase (see attached letters of support from both databases). The original intent was to have this occur through the ENCODE DCC, but in practice, this intermediary has slowed access to the data by the respective communities. In this coming year, we will develop a simple and straightforward data file format system that allows rapid inclusion of our data in to the public community databases. Processed, validated RNA-seq datasets will also be submitted on a regular basis (e.g. the last Friday of every month) to the MODs on schedule to be worked out with the databases. Finally, we have established a UCSC track hub for both types of data, which allows users (including ourselves) to have easy access to large datasets. We expect to be able to update this as data becomes available to facilitate our own review of the data. We will work with the UCSC browser group to ensure that this track hub becomes publicly available.

In addition, we plan to host an independent website that will contain certain simple analyses of the data, for instance, lists of genes assigned to binding sites for each TF, along with basic statistical metrics that to assess the confidence of each assignment (TIP analysis). Additionally, binding sites that might be HOT sites will be indicated, as described above. Finally, for the RNAi experiments, lists of genes whose expression changes in response to RNAi of a given TF will be listed, along with statistical analysis of the change of expression. \*\*\* Who is responsible for this? How will this be sustained after the project is completed?\*\*\*

C6. Milestones

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Milestones** | **Organism** | **Year 1** | **Year 2** | **Year 3** | **Year 4** | **Year 5** |
| ChIP-seq datasets analyzed | Worm | ? | ? | ? | ? | ? |
| Fly | ? | ? | ? | ? | ? |
| RNA-seq datasets analyzed | Worm | ? | ? | ? | ? | ? |
| Fly | ? | ? | ? | ? | ? |

D. Management/Dissemination Core

D1. Specific Aims (distribution of the resources and data)

D2. Progress Report (mention here establishment of the pipeline into the ENCODE database?)

D3. Aim 5 Data release, publication and leadership plan

D3.1. Release of the Data to the Research Community

Our plans for strain distribution and data release are detailed in Sections B? and C5. Briefly,…

D3.2. Collaboration Leadership Plan

Dr. Robert Waterston will oversee the entire project. He has extensive experience managing large projects of this kind, including 12 years of experience as Director of the Genome Sequencing Center at Washington University. He has also served as department head for 24 years. The four collaborating investigators (Celniker, Gerstein, Reinke, White) will serve as the executive committee. This group along with relevant members of their laboratories will have conference calls bi-weekly to review research progress and plan strategies. The coPIs will talk afterwards to discuss any administrative issues. Based on the past several years in the modERN project, this structure works very well and we have not had any difficulties agreeing on strategies or with participation or teamwork. We therefore expect to reach consensus on most decisions, and Dr. Waterston is responsible for all final decisions related to this project.

Each individual investigator (or in some cases two investigators) will be responsible for their particular projects. In most cases of a large group, a senior laboratory member will oversee the daily activities of the project. In the event that any group falls behind, funds will be reallocated to ensure that the goals of the project are met. Because different groups have overlapping expertise one group can often readily step in to alleviate any productivity problems that may occur. However, based on this current granting period, we do not expect to have any difficulties with this aspect of the project.

D3.2 Resource Sharing

We have a long history of advocating early data release and will continue these policies here. As the data are initially validated we will work with the public databases, particularly WormBase, Flybase, and the ENCODE Data Analysis Center, to ensure that the data is released to the public. We will also adhere to data release policies developed by the ENCODE consortium.

\*\*\*This needs to be expanded. Should use this to discuss the extensive efforts with the DCC. Alec, can you add some of that here?\*\*\*