w

ENCODE DAC Progress Report

This grant is from 2/01/2017 to 1/31/2021

The first annual report covers 2/01/2017 to 11/1/2017.

B.1 What are the major goals of the project?

We propose an ENCODE Data Analysis Center (EDAC, DAC in short) to support, facilitate, and enhance integrative analyses of the ENCODE Consortium data on human and mouse. We will work closely with Consortium members to identify and prioritize integrative analyses that should be carried out, identify the best groups and methods to accomplish them, coordinate all necessary data transformations, and undertake these analyses with the other Consortium members. Our ultimate goal is to ensure a successful final product of high-quality annotation in human and mouse, called the ENCODE Encyclopedia, and gain new insights into the biology and genetic regulation of animal genomes.

The DAC performs activities to achieve the following four goals: First, analyzing and integrating data and metadata from a broad range of functional genomics projects. (Aim 2). Serving as an informatics resource by supporting the activities of the ENCODE AWG. (Aim 3). Creating high quality Encyclopedias of DNA elements in the human and mouse genomes. (Aim 4) Assessing quality & utility of the ENCODE data & providing feedback to the Consortium. To achieve these four aims, the proposed DAC will work closely with members of the Consortium and in particular two entities within it: firstly, the Analysis Working Group (AWG), consisting of all Principal Investigators (PIs) of the production centers, PIs of functional characterization centers, informatics PIs, and personnel from each of the groups; and, secondly, the Data Coordination Center (DCC), responsible for all data and metadata submission, data formatting and uniform processing, and data sharing with the larger scientific community.

B.2 What was accomplished under these goals?

Zhiping:

Mark:

EN-TEx

Since the start of the fourth phase of the ENCODE project we have continued co-chairing (together with Tom Gingeras, Barbara Wold and Roderic Guigo) the bi-monthly EN-TEx working group conference calls which includes the previous ENCODE3 members of the working group as well as new members of ENCODE4 that have been invited to participate. The main goal of the working group to demonstrate the value of analysing functional genome data using the personal genome of an individual rather than the reference genome as is currently common practice. Towards this effort a main focus of the working group has been assembling personal genomes for the four EN-TEx individuals using a variety of different sequencing technologies:

Illumina paired end short reads, PacBio and 10X Genomics. Tools developed in the Gerstein lab are currently being used to convert VCF files of variants (SNVs, indels and SVs) into maternal and paternal genome sequences as well as map files in order to liftover annotations and coordinates between the reference and the maternal and paternal genomes. Effort has been devoted to compare the improvement of the different versions of the EN-TEx genomes compared to the reference genome by comparing the mapping rate of functional genomic reads from RNA-Seq, ChIP-Seq and HiC.

As a way to benchmark the performance and utility of using a personal genome effort has also been devoted to analyzing the vast amount of the current ENCODE data for GM12878 on the deeply sequenced NA12878 genome. One way to evaluate the effect of a personal genome is to investigate the phenomena of allele-specific behaviour. The Gerstein lab has extensive experience in this area and using an initial versions of the personal genomes for the four EN-TEx individuals and has analyzed the allele-specific expression and binding for all available EN-TEx RNA-Seq and ChIP-Seq data. The Gerstein lab is also developing methods for performing allele-specific analyses for the the EN-TEx HiC data that is available.

The Gerstein lab has been in communication with the DCC about file formats for transferring the work product from the EN-TEx working group in order to be hosted on the main ENCODE DCC portal. We have supplied them with sample output from our analysis pipelines for allele-specific SNVs and regions as well as personal genome sequences and auxiliary files. There has been extensive discussion for the ongoing analyses by the EN-TEx group in order to finish these analyses with the intention of submitting publications in early to mid 2018.

EN-TEx Extension

Together with Lieberman-Aiden Lab at Baylor, we have started calls for a working group called the “EN-TEx-ENCODE4 Personal Genomes” in August. The aims of this group are (1) to focus on the long-term goals of EN-TEx and involve ENCODE4 PIs in the current ENTEx effort started with ENCODE3 and (2) to start a potential extension of the assembly and annotation of personal genomes to the ENCODE cell lines/types and tissues that were studied extensively. We have had two calls so far. The first call was devoted to the discussion of logistics (email group, outreach to interested PIs and potential cell lines to work with). The second call had a strong attendance from ENCODE4 participants and we discussed the potential values of personal genomes and the long-term goals in terms of the genomes for which cell lines should be assembled and what these assemblies will add to the current agenda for the ENCODE consortium.

Nuclear Architecture

The Gerstein lab is participating in the Nuclear Architecture Working Group (NAWG) which is currently focused on standardizing analysis pipelines for processing both HiC and ChIA-PET data that the Lieberman-Aiden and Ruan labs are generating. Conference calls have included members of the NIH Common Fund 4D Nucleome Consortium in order to standardize pipelines and processed output data from the ENCODE and 4D Nucleome Consortium. The Gerstein Lab has also recently published two

papers focusing on the analysis of HiC data. The first is a method called HiC-spector (Yan et al. (2017) *Bioinformatics*) which develops a methodology for comparing the reproducibility of HiC data using contact maps. The second is a method called MrTADFinder (Yan et al. (2017) *Plos Comp. Bio.*) which uses a network modularity based approach in order to identify topologically associating domains at multiple resolutions.

Enhancer Prediction

Our lab has developed a framework for enhancer identification which can be applied to different tissues and cell lines across mammalian organisms with high specificity. The model adopts the matched-filter algorithm which is a well-developed method in signal processing for supervised enhancer prediction. From massive parallel reporter assays (MPRA), we created meta-profiles from ChIP-seq signals of different histone modifications around active enhancers, The meta-profiles show peak-trough-peak pattern as reported in several previous literatures. A separate meta-profile is created from DNase-seq signals which demonstrates a single peak at enhancer regions. The model scans the genome with these meta-profiles and integrates the matched-filter scores with SVM to generate genome-wide enhancer regions prediction. We have validated our model with transgenic assays in different mouse tissues and reporter assays in human cell lines. The manuscript of this framework has been submitted to Nature Methods and is currently under review.

Disease-specific Annotations from ENCODE

It is challenging to discover associated or even causal genes or loci for diseases. This is partially because testing on the whole set of comprehensive annotations will significantly reduce statistical power. Hence, the Gerstein lab aims to refine the encyclopedia annotations to such analysis for a variety of diseases.

Specifically, we started from refining the enhancer predictions from two aspects. First, we assigned different cell-type specific confidence levels of enhancers by integrating multiple functional characterization assays, including ChIP-Seq, DNase-Seq, and STARR-Seq. Second, we tried to trim the enhancer regions down to core elements by combining the chromatin status with motif information. Users can select different levels of disease-specific annotations according to their power requirements and sample size.

Besides, we also try to link various noncoding elements to genes with high confidence to provide cell-type specific extended gene definitions, which include transcription factor/RBP/miRNA binding sites, promoters, and enhancers. In particular, we developed two enhancer-gene linkage prediction methods JEME and ENGINE, which integrate DNAse-Seq, ChIP-Seq, RNA-Seq, Hi-C and ChIA-pet data.

With the refined linkage analysis, we further extended the TF regulatory networks by merging the proximal (TF-promoter-gene) and distal (TF-enhancer-gene) networks to illuminate potential regulatory changes (e.g. key rewiring TFs) and pinpoint key regulators that reshapes the disease-specific expression profiles. We also constructed such networks in stem cells, e.g.

H1-hESC, to test whether the network changes during the normal-to-disease transition is more stem-like or not.

Manolis:

Anshul:

Roderic

Bill

Shirley:

With the increasing in the number of publically available ChIP-seq samples in GEO we have relied increasingly on an automated parsing procedure to annotate the factor and cell type associated with each sample. Unfortunately, annotation in GEO has not been done in a consistent way and partial information is often given without use of any standardized vocabulary. To improve the number and accuracy of transcription factor annotations we developed a system to extract this information from antibody ids. We identified the companies with the most ChIP-seq samples in GEO: ABcam, SantaCruz, Sigma, Millipore, Bethyl, Active Motif, and Cell Signalling, and retrieved antibody information from their webpages. In addition, we improved TF recognition by keyword, by assembling a list of known gene names. To improve cell line recognition we collected cell line information from scicrunch.org/resources and from ATCC.

We have processed 949 new histone modification samples, including 300 H3K4me3, 112 H3K4me2 and 158 H3K4me1 samples. along with 444 transcription factor ChIP-seq samples representing 88 different TFs. For these samples we have calculated quality control statistics similar to those in ENCODE, including the fraction of reads in peaks (FRiP) score and (PBC) PCR bottleneck coefficient. Our scores are calculated consistency across all Cistrome data, including previously collected samples, and are therefore not precisely the same as ENCODE scores. Nevertheless, we have calculated QC metrics for ENCODE3 data to assess the overall quality of the newly collected data. For the histone modifications the data we collected tended to be sequenced to a lower depth (30.6M median reads) than ENCODE data (42.9M median reads). The median sequencing depth for transcription factors was 28.4M, similar to ENCODE (33.2M). Median uniquely mapped ratios were very similar for collected data and ENCODE for TFs and histone marks (82%-88%). Median PCB scores were also similar (0.98-0.99). Median FRiP scores for TFs were similar 4% for collected data, 3% for ENCODE samples. Collected histone modification data median FRiP (8%) was a little higher than ENCODE3 (5%), although this probably reflects the preponderance of active marks in the collected data. Overall the quality of the collected data was very similar to that of ENCODE.

Rafa

B.4 What opportunities for training and professional development has the project involved?

Zhiping:
Professional development and training opportunities play a vital role in achieving success and attaining greater proficiency. UMass Medical School recognizes the critical importance of preparing the postdoctoral scholars for success within a broad spectrum of scientific careers. All incoming graduate students and postdocs receive training in how to create an Individual Development Plan (IDP) as part of the required onboarding course in Responsible Conduct of Research. They are constantly given feedback to help assess both their strengths and areas for growth. We encourage the attendance of conferences, workshops, and professional courses, and provide multiple options for convenience and flexibility. We also help them build professional networks and collaborations.

Mark:
Professional development and training opportunities play a vital role in achieving success and attaining greater proficiency. Yale University recognizes the critical importance of preparing the postdoctoral scholars for success within a broad spectrum of scientific careers. All incoming graduate students and postdocs receive training in how to create an Individual Development Plan (IDP) as part of the required onboarding course in Responsible Conduct of Research. They are constantly given feedback to help assess both their strengths and areas for growth. We encourage the attendance of conferences, workshops, and professional courses, and provide multiple options for convenience and flexibility. We also help them build professional networks and collaborations.

Manolis:

Anshul:

Roderic

Bill

Shirley:

Rafa

B.6 What do you plan to do during the next reporting period to accomplish the goals?

Zhiping:

Mark:

The Gerstein lab will work with the DCC in order make sure all the analysis work products from the EN-TEx working group is submitted and incorporated into the ENCODE portal at the DCC. We will work together with the Lieberman-Aiden lab in order to develop methods for determining allele-specific behaviour using HiC data. We will continue our efforts to develop methods for determining allele-specific behaviour for larger regions and analysis the significance of allelic structural variants. We will also develop an integrative method for a multi-assay based allelic state caller.

We will continue refining the encyclopedia annotations for disease genomes by extending the current work in model cell lines to a variety of other cell types. For cell types without enough functional characterization data, we will try to synthesize imputed information to construct disease-specific networks to monitor and investigate regulatory changes in different diseases.

Manolis:

Anshul:

Roderic

Bill

Shirley:

Rafa

C.1 Publications

A pubmed search using the grant number HG009446 and did not find any publications.

Zhiping:
None

Mark:
KK Yan, S Lou, M Gerstein (2017). *MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions.* **PLoS Comput Biol 13: e1005647.** PMID: 28742097 PMCID: PMC5546724.

KK Yan, GG Yardimci, C Yan, WS Noble, M Gerstein (2017). *HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps.* **Bioinformatics 33: 2199-2201.** PMID: 28369339.

Manolis:

Anshul:

Roderic

Bill

Shirley:

Rafa

C.2 Websites or other internet sites

Zhiping:
http://screen.encodeproject.org
http://factorbook.org

Mark:

Manolis:

Anshul:

Roderic

Bill

Shirley:

Rafa

C.4 Inventions, patent applications, or licenses

Zhiping:
None.

Mark:

Manolis:

Anshul:

Roderic

Bill

Shirley:

Rafa