**ENCODE DAC Progress Report**

**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 Aiden Lab at Baylor, we have started calls for a working group called “EN-TEx-ENCODE4 Personal Genomes” at the end of 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.

**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**

Our lab aim to perform large-scale data integrations to tailor the ENCODE annotations specifically for diseases. We started from data-rich cell types to deeply investigate disease-specific genomes and extended our framework to other cancers will less functional characterization data.

Specifically, we first paired ENCODE top-tier cell lines (K562), breast (MCF-7), liver (HepG2), lung (A549), brain (SK-N-SH), and cervix (HeLa-S3) with various normal cell types to deeply annotate cancer genomes and investigate regulatory tumor-to-normal regulatory changes. By integrating DNAseq, ChIP-seq, eCLIP, STARR-Seq, Hi-C and ChIA-pet data, we built up accurate extended gene annotations for each cancer by linking all the noncoding annotations to genes with high confidence, which include transcription factor/RBP/miRNA binding sites, promoters, and enhancers. We also built cancer-specific gene regulatory networks to illuminate potential regulatory changes (e.g. key rewiring TFs) and pinpoint key regulators that reshapes the disease-specific expression profiles. Then, we extended the above extended gene annotation and regulatory networks on other diseases by data reconciliation and imputation in extensive cell types. For example, we set up tissue-specific TF networks by integrating DNAse-seq and context based motif data for 20 different cell types.

Finally, we leverage the the above cancer-specific annotation resource to provide a prioritization scheme to pinpoint key regulatory elements and SNVs for small-scale follow-up. Our approach underscores the value of large-scale data integration and we anticipate that our annotation resource will be more accurate with increasing ENCODE assays on specific tissues and tumor samples.