In the second year, we continued to make progress on the aims of the project. Overall, we continued to develop our computational models and do validations.We continued to have calls. Also, On Feb 20, 2017, the Yu and Rubin labs visited Yale University for a two-day meeting and discussed the details and made plans for the following year. On June 6， 2017, we have a one-day meeting in New York City to present the progress of each group and have a face-to-face discussion on the problems we have on both the experimental and computational sides.

Specifically, in relation to the aims:

**In related to Aim1:** The Gerstein Lab further expand the model by adding new features, which will consider gene expression level of transcription factor. The gene expression level are re-ordered by the waiting time of TF binding, which is thought as a indicator of directly binding to indirectly binding. By adding this cell specific feature, our model can be generalized to different cell lines. We also collect more data from TCGA cancer mutation dataset, including Prostate cancer, Breast Cancer and other cancer types.

Some coding variants have been mis-annotated as non-coding variants because it is in upstream open reading frame. Gerstein Lab built a simple Bayesian classifier using 89 attributes of uORFs labeled as active in ribosome profiling experiments. This allowed us to extrapolate to a comprehensive catalog of likely functional uORFs and generate a true non-coding variants set.

The estimation of non-coding variants effect reply on its association with coding genes, especially for transcription factor binding and gene regulatory network. To aggregate the effect of non-coding variants, we also incorporate the loss-of-function effect of coding variants on that target genes. Gerstein Lab presents ALoFT (annotation of loss-of-function transcripts) tool to help annotate the variants on the target genes of a non-coding SNV.

Identifying genomic regions with higher than expected mutation count is useful for the prioritization of variants effect. Previous parametric approaches require numerous cell-type-matched covariates for accurate background mutation rate (BMR) estimation, which is not practical for many situations. Non-parametric, permutation-based approaches avoid this issue but usually suffer from considerable compute-time cost. Hence, Gerstein Lab introduce Mutations Overburdening Annotations Tool (MOAT), a non-parametric scheme that makes no assumptions about mutation process except requiring that the BMR changes smoothly with genomic features. MOAT randomly permutes single-nucleotide variants, or target regions, on a relatively large scale to provide robust burden analysis.

**In related to Aim2:** To further refine our model, Gerstein Lab continue to use the online learning framework to tune the parameters incorporating the data from public domain and our experimental result**.**

**In related to Aim3**: Site-directed mutagenesis was performed by the Yu Lab on the pDONR223 entry clones of each of the previously tested 14 elements in order to introduce two independent SNVs into each (28 total). The Rubin Lab then verified correct mutagenesis via Sanger sequencing by and then cloned into pDEST-hSCP1-luc using Gateway LR clonase. These mutant luciferase reporters were then transfected into K562 cells and luciferase activity was measured in comparison to the reporters containing the WT element. 8 of the 14 elements demonstrated a significant change (up or down) in reporter activity with the mutations, however for 5 of these elements the mutation predicted to have no effect of the enhancer activity did alter it.

The Rubin Lab also began scaling up the cloning pipeline, this time amplifying 100 individual elements. Elements were amplified via PCR with primers containing attB1 and attB2 sequences and cloned into pDONR223 using Gateway BP clonase in a 96-well format. Four colonies for each element were picked and an aliquot of each were pooled, creating four pools each containing plasmids from 96 colonies. These four pools were separately barcoded and sequenced on a MiSeq. Of the 96 elements 86 were successfully sequenced. 31 (36%) of the elements contained a common SNP at all 4 of the clones sequenced, however in no occasion did they occur in the sequences of interest for site-directed mutagenesis.

Based on the prediction, Yu lab used a webtool (primer.yulab.org) to design site-directed mutagenesis primers to introduce the target SNVs into the 14 elements. Two SNVs will be introduced to each element respectively to generate two single-SNV-containing alleles, one of which would result in a significant change in the enhancer activity while the other would not. After sequencing-confirmation of the mutant clones, the enhancer activity of the SNV-containing elements and their corresponding WT elements were compared via the dual luciferase assay.

In the meantime, due to the progress of the field, Yu lab are working to establish three new experimental methods that were not described in the original application:

1. We are working to establish a new sequencing-based enhancer reporter assay to be used in place of luciferase assays (luciferase assay will still be used, but will be used only to validate eSTARR-seq results). STARR-seq (self-transcribing active regulatory region-sequencing) is a recently-established method that can identify enhancer elements genome-wide(Arnold, Gerlach et al. 2013). Importantly, STARR-seq does not quantify the enhancer activity of individual candidate fragments, but instead requires creation of a complex library of unique but overlapping fragments for each candidate region to be tested. Thus the original STARR-seq protocol ***cannot*** be directly used to measure enhancer activities from a clonal library of WT and mutant enhancer elements, where each element has one and only one clone with defined boundaries, as is the case for our proposed research. To circumvent these difficulties, we developed the element-clone-compatible STARR-seq (**eSTARR-seq**) protocol to incorporate a unique molecular barcode to the cDNA of each mRNA molecular produced at the reverse transcription step, allowing direct quantification of enhancer activity for each individual enhancer by counting only reads with unique molecular barcodes.
2. In Conjuncation with eSTARR-seq, we are working on establishing a massively multiplexed cloning technique, using Long-Adaptor Single-Strand Oligonucleotide (LASSO) probes (Tosi *et al*., 2017). This cloning technique will enable us to clone several thousands of gDNA fragments, ranging from hundreds to thousands of base pairs, in one microtube of reaction. In brief, it starts with the array-based synthesis of “pre-LASSO” probes. After a series of modification steps, the single-strand LASSO probes are generated. When mixed with gDNA templates, the LASSO probes can target the gDNA regions of interest via capture by circularization. While the overall procedure resembles MIP (molecular inversion probe)-based cloning, LASSO-based cloning is more powerful in terms of its higher cloning capacity and lower length bias. We are working on incorporating LASSO-based cloning in capturing the WT gDNA fragments of interest before introducing variants in those fragments.
3. In terms of mutagenesis, we are also working on a new *en masse* directed mutagenesis pipeline, MassClone-seq. Compatible with the massively multiplexed format of cloning, the new mutagenesis technique will also allow us to introduce hundreds to thousands of pre-determined mutations in a one-pot fashion.

B.6 WHAT DO YOU PLAN TO DO DURING THE NEXT REPORTING PERIOD TO ACCOMPLISH THE GOALS?

* Perform the last run parameter tuning and finalize the model
* Perform large-scale traditional PCR and/or LASSO-based cloning to capture gDNA regions of interest.
* Establish eSTARR-seq and corresponding computational analysis pipeline to quantify enhancer activities from a LASSO clone library.
* Perform large-scale target mutagenesis to introduce two SNVs per element using our Clone-seq pipeline and/or MassClone-seq pipeline.
* Determine the enhancer activity of the WT and corresponding variant-containing gDNA fragments with eSTARR-seq and luciferase assays.
* Final assessment of the model and evaluation of the impact on gene expression and functional consequences.

G.10 ESTIMATED UNOBLIGATED BALANCE
G.10.a Is it anticipated that an estimated unobligated balance (including prior year carryover) will be greater than 25% of the current year's total approved budget?

Yes, We anticipate a large carry forward on the grant. The carry forward on this grant reflects a number of things. First and most importantly, one of the co-investigators has moved from New York to Switzerland. This has involved a lot of administrative paperwork and it took a very long time to get approvals from the NIH plus not getting approvals until just a month or so ago. This, of course, has slowed the project in Switzerland overall quite dramatically and has led to considerable carry forward.

Secondly, the field as a whole is moving quickly and new technologies are being developed. Therefore, in particular in Ithaca, we've decided to slow the production process a little bit and invest time in focusing on what technologies might be good, which we plan to ramp up immediately in the end phase of the project.

## Report of Last year (2016):

We initiated the project in June 2016 and started to collect data and information. During this period, the Gerstein, Rubin and Yu labs discussed the collaboration details and made a plan for a regular bi-weekly working group call. The first working group call was held on Sept 12, with the Gerstein lab reporting an update in statistical modeling. This was followed by regular calls. On Feb 20, 2017, the Yu and Rubin labs visited Yale University for a two-day meeting.

**In related to Aim1:** The Gerstein Lab built a statistical model to predict and prioritize germline variants that can significantly affect enhancer regulatory activity, which employs conservation score, histone modification, transcription factor binding profile, CAGE data information. The model can do both classification and regression: the classification model can predict a probability whether variants can significantly change the regulatory activity while the regression model can predict the regulatory activity change between mutant and wild type (fold change). To iteratively tune parameter using experimental results, we developed an online learning framework to refine model for **Aim 2** and then select about 200 regions and predicted variants with different level of regulatory changes to further tune parameters for **Aim 3**.

Besides model learning, we have conducted a case study to identify and prioritize impactful noncoding variants on Renal cell carcinoma, which accounts for more than 90% of kidney cancers. Papillary renal cell carcinoma (pRCC) is the second most common subtype of renal cell carcinoma. We carry out the first whole-genome study of pRCC to discover triggering DNA changes explaining these cases. We find two impactful noncoding variants which are associated with prognosis and can aid clinical decisions. The results have been published PLOS Genetics on Mar 30, 2017.

**In related to Aim 3:** The Rubin lab assessed experimental conditions and developed a high- throughput pipeline for WT element cloning into a pDEST-hSCP1-luc luciferase reporter vector on a small selection (14) of predicted enhancer elements. Elements were amplified via PCR with primers containing attB1 and attB2 sequences and cloned into pDONR223 using Gateway BP clonase. Four colonies for each element were picked and sequenced via Sanger sequencing. 54 out of the 56 colonies contained the precisely cloned WT element with no nucleotide changes, indicating a PCR-induced error rate of <4%.

One clone for each element with the correct sequence was then cloned into pDEST-hSCP1-luc using Gateway LR clonase, and luciferase reporters containing the elements were then transfected into K562 cells. 13 of the 14 predicted enhancer elements demonstrated positive enhancer activity indicating a high success rate for our predictions.

To construct a positive control set for the enhancer activity assays, the Yu Lab cloned several widely used promoters that have been implied to possess enhancer activities. These include CMV (cytomegalovirus) promoter, PGK (phosphoglycerate kinase) promoter, SV40 (Simian virus 40) promoter, and RSV (Rous sarcoma virus) promoter. They delivered the luciferase assay vector containing the promoters into K562 cells via electroporation and detected up to 9-fold higher enhancer activities in these elements compared to the CDS of EGFP by dual luciferase assay.

Based on the prediction, they used a webtool to design site-directed mutagenesis primers to introduce the target SNVs into the 14 elements. Two SNVs will be introduced to each element respectively to generate two single-SNV-containing alleles. We predicted one of the SNVs would result in a significant change in the enhancer activity while the other would not. The experiments for mutagenesis and the following enhancer activity comparison between the WT and mutant elements are currently in progress.