Conclusion and future directions

The study of RNA binding proteins, like the study of transcription factors before it has opened up a brand new field in genomics. We have used this power to understand the effects of single RBPs on human health and disease, and are just now understanding how to integrate multiple RBPs to gain a better picture of that is going on in the cell.

\section{TAF15}

Single RBP studies can help move the field, even if they don’t provide conclusive evidence towards the mechanism of disease. With the TAF15 study we characterized the RBPs binding in the most relevant cell types we could easily gain access to, but conclusions we drew from the results were weak at best. TAF15 appears to at least be loosely associated with ALS, so it may be worthwhile to keep searching for TAF15s true mechanism of action in the disease. It could an unobserved effect on RNA processing, or it may have to do with its aggregation into stress granules, which we were not equipped to study.

\section{Other RBPs}

The value of single RBP studies have been shown in the additional collaborative analysis I performed in the process of writing my dissertation. Thanks CLIP-seq experiments, and integration of results we now better understand the mechanism by which UPF1 functions to degrade NMD targets. Additionally our studies of Musahi 2 have elucidated the mechanism of action drugs used in the process of growing hematopoietic stem cells. I expect in the future this research will directly lead to improved efficiency of bone marrow transplants.

Looking at single RBPs can be hugely informative, and I expect that the majority of studies will continue to be performed utilizing CLIPs on single RBPs.

\section {Quality Control}

My work on quality control in eCLIP data is the first published suggestions detailing considerations for experimental and computational analysis. I provided pre-sequencing measures of quality control to save money and time for scientists before spending thousands of dollars finishing an experiment. I provided suggestions as to exactly how deeply to sequence to reach an optimal cost to discovery ratio, and most importantly I provided computational quality control metrics post-sequencing so people don’t waste time analyzing bad data.

Studies of quality control will continue to be performed. We are just now learning how to effectively scale CLIP-seq assays to understand the effects of hundreds to thousands of experiments at once. In the end quality control for each RBP is not a one size fits all approach. My methods were mostly developed for point source binders, and fail to appreciate the diversity of RBP binding locations and effects. In the future we will need to create more fine grained quality control metrics based on an understanding of RBP function. From the over 2000 datasets that we have in our lab we have just started scratching the surface of binding diversity. At a very basic level we should mark the difference between narrow and broad binders. Pre-mRNA and mRNA binders should also be looked at differently. Finally RBPs that primary map to repetitive elements need to be looked at in a structured way. We still have trouble knowing exactly what it means to bind to a repetitive element, let alone what level of binding is significant. Repetitive element binders pass QC right now likely due to technical artifacts, but this should be fixed in the future.

In addition to methods refinement to QC researchers could, and should continue to improve the quality of the experimental assay. This could subtlety change the metrics that are used to show quality of data.

Likely the observation with most staying power in my research is the estimate of library complexity. Often times people think that a sequencing library contains infinite molecules. CLIP and iCLIP proved this is not the case. Being able to communicate that sequencing deeper won’t actually yield additional information and quantifying that is an obvious, but critical observation for the field.

In the end automated quality control metrics, like automated methods to understand the effect of data can only detect signals you know you’re searching for. This is great most of the time, but biology is about the discovery of the unknown, the thing that makes you say “hmm, that’s unexpected”. The best quality control, especially for new RBPs will always be looking at each dataset by eye, and seeing what is unexpected.

/section{CLIP-seq field}

I see the CLIP-seq field as a whole as just starting, only within the last two years have truly reproducible, scalable methods to understand RBP binding been developed. This will enable a whole new area of basic biology research. With this comes a host of challenges and opportunities.

First and foremost the question we must know who is going to analyze the data. Currently expert experimentalists hand off data to expert computational analysts. This approach worked when the rate and scale of data generation was small. As high throughput methods get cheaper and faster this will be an inefficient way of doing things. The handoff of data and experimental question between two individuals takes time, not to mention buy in from both parties. There is a need for the data generators to take over primary data analysis, so they can ask the right question, and ask it more quickly than a handoff would allow for. The field isn’t there yet, but it is moving in that direction.

On the experimental side understanding the function of an RBP is still an open question. RBPs regulate many more individual pathways than transcription factors, which at the end of the day turn genes on or off, and can be assayed with a simple RNA-seq experiment. The mechanism of action for TFs is still often an open question, but the readout is well known.

For RBPs the readout can be any one of tens of different high throughput experiments. RBPs can effect stability, splicing, translation, polyadenylation, capping, one of the 100s of RNA modifications. Going in blind there are too many hypothesis to test to understand the function of a single RBP. We are addressing this somewhat using high throughput screens for RBP function including stability, splicing and translation, but these don’t address RBP function in its native context.

In the future we should develop computational approaches, based on RBP binding profiles to automatically provide suggestions as to the function of an RBP. This would help save time and money when trying to assay its function.

Finally we are just now scratching the surface of true integrative analysis between many RBPs binding locations, and other assays to understand their function. We need better methods to understand how RBPs behave cooperatively to regulate cellular state or disease.

To achieve this dream of integrative analysis though more near term issues must be more effectively addressed. It will be difficult or impossible to integrate RBP data without high quality peak calls, and to a lesser extent a better understanding of how to quantify differential binding. In the TF field, integrative finding common co-bound states enabled analysis, based off of high quality peak calls. Truly high quality peak identification algorithms must be developed for RBPs as well.

The largest current drawback to CLIP-seq peak calling algorithms today is they can’t differentiate peak size effectively. Some RBPs bind a specific motif, some bind entire transcripts. Current peak callers use the same model to call point-source peaks for both types of RBPs. A true peak caller would accurately identify both the region of binding for a specific RBP, and provide an estimate of the point-source binding location (if possible) as well.

Important to small scale studies is an understanding of differential binding. Currently peak calling methods have trouble comparing between two conditions. This is primarily because there isn’t a solid understanding of the correlation between binding strength and read density. This understanding must be solidified to better understand RBP binding between different conditions.

The future of the RBP field is in basic research. Once RBP binding patterns are better understood machine learning methods can be applied to understand how mutations or changes in gene expression will effect binding and downstream gene regulation. This kind of research and machine learning approach will need to be enabled by vast scale. Currently CLIP-seq is labor intensive. To truly understand effects of all RBPs at scale CLIP-seq will need to be automated, and the price per experiment will need to drop drastically.

These predictions will not only allow the field to predict effects of mutations, but also ideally allow for the prediction of customized RNA (or otherwise) drugs that can treat an individuals specific disease.

In the end through biomedical research we are attempting to cure diseases. Analysis of RNA binding proteins will play their role through basic research, allowing scientists to understand the cause of disease more easily, and providing new therapeutic avenues for treatment.