Research Statement Sachin Pundhir

Recent advances in next-generation sequencing (NGS) have posed a unique challenge in terms of the analysis and interpretation of huge amount of data generated by this technology. *Computational biology*, a research area that lies at the intersection of computer science and biology is making immense contributions in addressing this 'big data' challenge. My research focuses in this area, specifically on to develop and/or use computational and statistical techniques to interpret complex biological datasets. Central to my research are the questions as to:

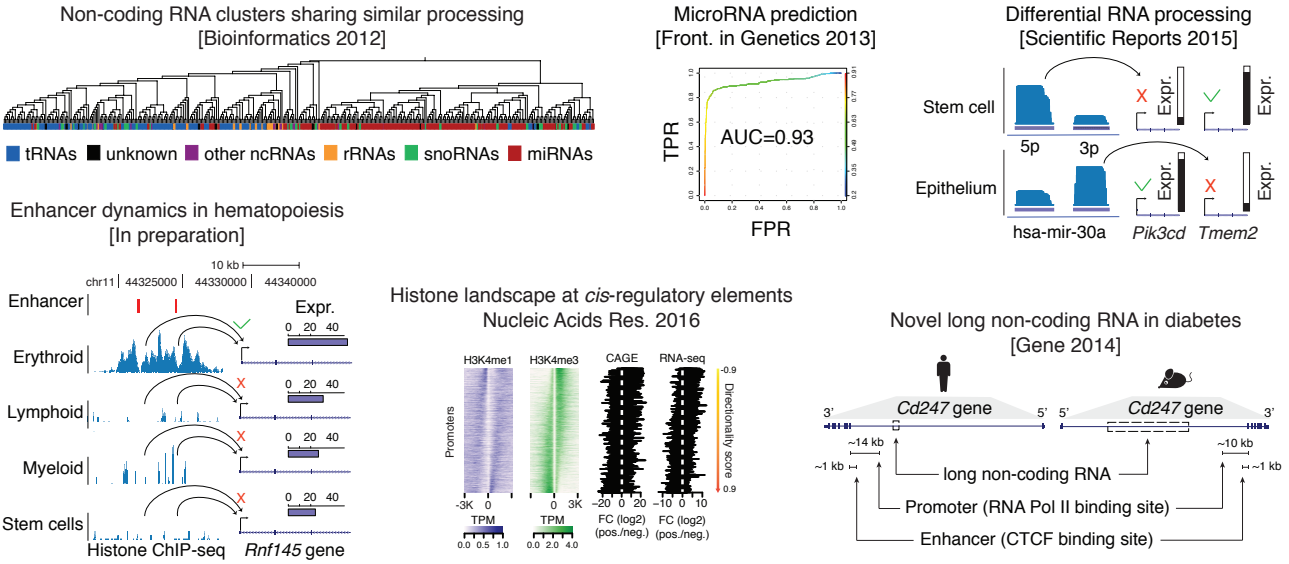
1. How despite having same genetic code a huge diversity of cell arises?
2. How transcriptome (RNA-seq) and epigenetic (ChIP-seq) landscape defines the identity of a cell?
3. How alterations in this landscape drive cellular differentiation, and if aberrant to the formation of cancer cells?

These are fundamental questions, an exploration of which can lead to better prognosis of disease and personalized medicine. However, they also raise several challenges due to non-coherent nature of the data and the inherently interdisciplinary nature of the problems.

My goal is to address these questions by proposing new computational methods that use standardized NGS data formats and robust statistics to capture patterns that are unique to a particular cell state or disease condition, and by maintaining collaborations with biologists for experimental follow-up and validation. In the following paragraphs, I summarize my previous research work and outline my future research plan.

Past and current research

1. How to robustly detect functional non-coding RNAs in eukaryotic genome? We now understand that non-protein coding fraction of an organism's genome play important role in defining its physiological complexity, for example 99% of human genome is non-coding as compared to 30% in case of yeast. However, unlike protein, non-coding regions with similar function share low nucleotide sequence conservation, and are thus undetectable using traditional



sequence comparison-based approaches. I have worked on a novel approach using post-transcriptionally processed RNAs from non-coding regions, and captured during RNA-seq technology to show that they can be used to accurately predict different functional classes of non-coding regions (miRNA, tRNA and snoRNA) (1,2). Using transcriptome and epigenome data, I detected a novel long non-coding RNA (lncRNA) transcribed from the first intron of *Cd247* gene in both human and mouse. *Cd247* gene is important due to its association with Type 1 diabetes (T1D). We experimentally validated the expression of predicted lncRNA by collaborating with a wet-lab group, and proposed it as an important player in T1D due to it also being enriched for T1D associated-SNPs (3).

2. How does the post-transcriptional processing of non-coding RNAs varies across different cell types? One good example of such a dynamics occurs during the process termed as '*arm-switching*' in which the pre-microRNA (pre-miRNA) switches the arm (5' or 3') from where the mature miRNA is processed (4). To study this dynamics on a genome-wide scale, I used small RNA-seq data performed on biological replicates of multiple tissues from the ENCODE project. A major challenge was to normalize the read count in a way that increases the signal to noise ratio. This is essential to ensure effective comparison of RNA processing between multiple tissues sequenced at variable sequencing depths. On literature search, I identified one such method that is widely used for differential gene expression analysis (5). I optimized this method by integrating it with widely known Fisher's exact test to compare RNA processing across multiple tissues. I discovered, apart from miRNA, numerous examples of snoRNA and tRNA exhibiting a phenomenon similar to 'arm-switching' (6). Besides, I also identified genomic regions in proximity to Transcription start sites

(TSS) that are processed precisely at the same nucleotide position to produce small RNAs of length <22 nt across all the analyzed cell types. I also made the computational algorithm available as downloadable package to wider research community (6).

3. How transcription factors and cis-regulatory elements drive differentiation of stem cells into mature blood cells? *Enhancers* and *promoters* are important *regulators* located within the non-coding region of genome that dynamically switch on and off the expression of target gene(s), thus conferring directly to cell type diversity. My current research is focused on to understand the role of two transcription factors (TF), PU.1 and CEBPA in differentiation of stem cells into mature blood cells (hematopoiesis). I use RNA-seq and ChIP-seq data to study how the expression of the two TFs shapes the *cis*-regulatory landscape (enhancers) across different differentiation stages.

We show that, in contrast to previous notion, CEBPA knock out alters the enhancer landscape much earlier in the differentiation process. We show distinct classes of enhancers bound by PU.1, CEBPA or both that became activated at different rate during the differentiation. A major challenge during this project was to accurately predict active enhancer and promoter elements using histone modification data. Unlike standard approaches using genetic sequence to predict these elements, I exploited the pattern by which epigenome data map at these elements (Figure 1; sky blue). This pattern as being a more direct reflection of actual biological events occurring at these elements (nucleosome rearrangement in this case) detects genomic regions involved in active regulation with high specificity (Figure 1; dark blue). This research was listed as a featured article by Danish Stem cell association (DanStem). Two manuscripts, one published (7) and another one in preparation have resulted from my post-doctoral research till now.

Future research directions

My research is focused on understanding the role of non-coding RNAs and epigenome in defining cell identify. With concerted efforts from many research groups, we have only started understanding the crucial role these elements play in regulating spatiotemporal expression of protein coding genes. However, there are many open questions, some of which I enlist below that I would like to address in my future research.

4. How histone modifications shape the transcriptomic landscape of a cell? We now know that various post-translational modifications (PTM) at histone play important role in defining regulatory regions in the genome (enhancers, promoters and repressed regions), which eventually lead to precise regulation of gene expression (both coding and non-coding). There are many PTM at histone, different combinations of which lead to distinct chromatin environment across a chromosome. I am interested in understanding the role of two such combinations, a) H3K4me1 and H3K4me3; and, b) H3K4me3/me1 and H3K27me3. While H3K4me1 and me3 distinctly defines the presence of an enhancer and promoter, respectively, a subset of regions carries both these marks. I want to understand on which regions does it happen, is it consistent across different cell types, is it proximal to any specific functional class of genes and how does it effect stable vs. unstable transcription ratio at gene promoters?

5. What role does histone modifications and long non-coding RNAs play in higher-level organization of chromatin? Similar to their role in defining regulatory environment, histone modifications have also been suggested to play a role in defining higher-level organization of chromatin into loops, domains and compartments within the nucleus. These two roles of PTM are perhaps inter-dependent and in my future research I would like to study this relationship.

Along the same lines, I am interested in understanding the role of long non-coding RNAs (lncRNAs) in regulating histone environment and expression of genes. A well-known example of such a regulation has recently been shown for *Xist* lncRNA, which is involved in silencing one of the two X chromosomes in females. *Xist* performs this function by activating a chromatin modifier, histone deacetylases enzyme (HDAC) leading to deacetylation on X chromosome, which inhibits binding of RNA polymerase II on inactive X chromosome, thus silencing gene expression (8). Intriguingly, *Xist* was also shown to play role in organization of inactive X chromosome within the nucleus (8).

There are also some challenges that I foresee during these studies:

1. Analysis of NGS data requires extensive computer infrastructure and optimization of the computational algorithms for speed and efficiency.

2. The most important criterion of success for bioinformatics algorithms is to bring out novel biology that is supported by experimental data. Therefore, it would be crucial to validate many of the computational predictions with experimental methods.

To address these challenges, I plan to collaborate with fellow researchers in the field of statistics, computer science and biology. During my research career, I have published articles with more than 30 different coauthors, both nationally and internationally. I plan to continue the collaborative work in my future research. These collaborations potentially will generate novel biological hypothesis that can direct further computational and biological experiments. New experimental data can then be iteratively combined with old data to further optimize the computational algorithms and analysis.

References

1. Langenberger D*, Pundhir S*, Ekstrøm CT, Stadler PF, Hoffmann S, Gorodkin J. DeepBlockAlign: A tool for aligning RNA-seq profiles of read block patterns. Bioinformatics. Oxford Univ Press; 2012;28(1):17–24 (* join first authors).

2. Pundhir S, Gorodkin J. MicroRNA discovery by similarity search to a database of RNA-seq profiles. Front Genet. 2013 Jan;4(July):133.

3. Pundhir S, Hannibal TD, Bang-Berthelsen CH, Wegener A-MK, Pociot F, Holmberg D, et al. Spatially conserved regulatory elements identified within human and mouse Cd247 gene using high-throughput sequencing data from the ENCODE project. Gene. Elsevier B.V.; 2014 Jul 15;545(1):80–7.

4. Griffiths-Jones S, Hui JHL, Marco A, Ronshaugen M. MicroRNA evolution by arm switching. EMBO Rep. Nature Publishing Group; 2011 Feb;12(2):172–7.

5. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010 Jan;11(10):R106.

6. Pundhir S, Gorodkin J. Differential and coherent processing patterns from small RNAs. Sci Rep. Nature Publishing Group; 2015;5:12062.

7. Pundhir S, Bagger FO, Lauridsen FB, Rapin N, Porse BT. Peak-valley-peak pattern of histone modifications delineates active regulatory elements and their directionality. Nucleic Acids Res. 2016;gkw250.

8. McHugh CA, Chen C-K, Chow A, Surka CF, Tran C, McDonel P, et al. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature. 2015;521(7551):232–6.