Research Statement Sachin Pundhir

Recent progress in next-generation sequencing (NGS) technology is driving rapid accumulation of biological data, robust analysis and interpretation of which is of crucial importance. *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 is in this area. Specifically, I develop and use computational and statistical methods to interpret complex biological datasets. Central to my research are the questions as to:

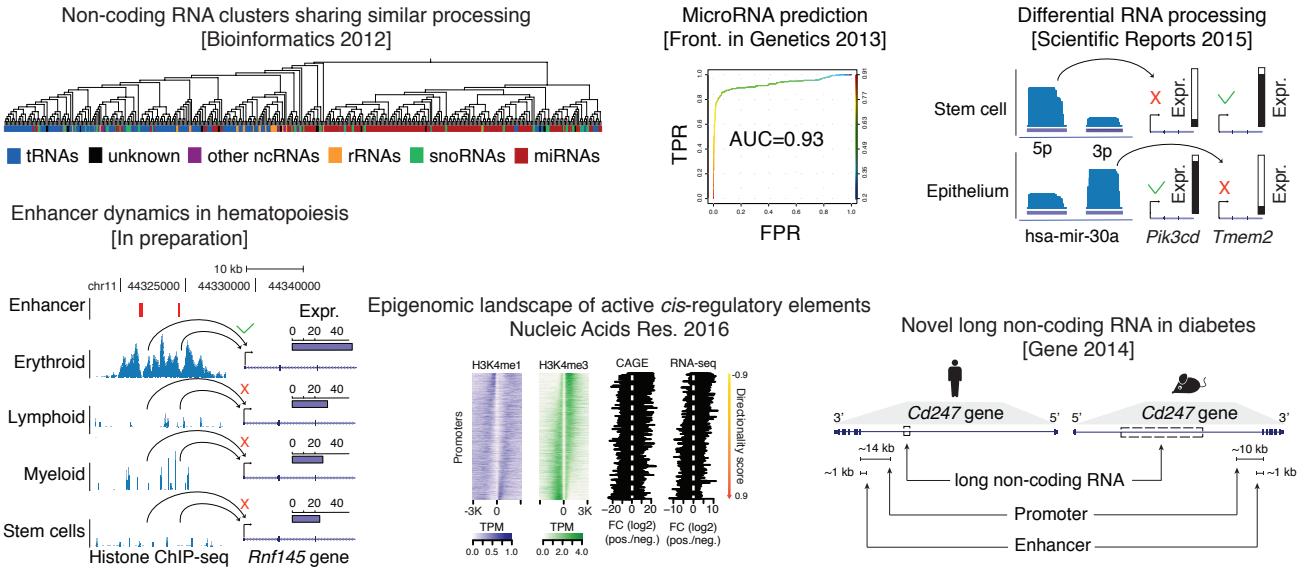
1. How despite having the same genetic code does the huge diversity of cell arise?
2. How do the *non-coding RNA* (RNA-seq) and *epigenetic* (ChIP-seq) landscape define the *identity* of a cell?
3. How do alterations in this landscape drive cellular differentiation, and if aberrant to the formation of 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 unique for a particular cell state or disease condition, and by maintaining collaborations with biologists for experimental follow-up and validation.

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. I developed a novel approach to predict different functional classes of non-coding RNAs (miRNA, tRNA and snoRNA), and showed it as more accurate than available methods. The approach uses post-transcriptionally processed RNAs from non-coding regions, captured using RNA-seq technology, for its predictions (Langenberger et al. 2012; Pundhir and Gorodkin 2013). I used



transcriptome and epigenome data to detect a novel long non-coding RNA (lncRNA) transcribed from the first intron of *Cd247* gene in both human and mouse. 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 (Pundhir et al. 2014).

2. How does the post-transcriptional processing of non-coding RNAs varies across different human cell types? One good example of such a dynamics occurs when pre-microRNA (pre-miRNA) switches the arm (5' or 3') from where the mature miRNA is processed ('*arm-switching*') (Griffiths-Jones et al. 2011). I studied this phenomenon on transcriptome-wide scale using RNA-seq data from nine human cell types and showed that many non-coding RNAs (snoRNA and tRNA) are processed and show dynamic activity similar to miRNAs (Pundhir and Gorodkin 2015). In contrast, short RNAs (<22 nt) proximal to transcription start site (TSS) are processed precisely at the same nucleotide position across all the analyzed cell types (Pundhir and Gorodkin 2015). I ensured robust comparison of RNA processing between samples sequenced at variable sequencing depths by using a novel normalization strategy based on integration of estimated expression levels with fisher's exact test (Pundhir and Gorodkin 2015).

3. How transcription factors and cis-regulatory elements drive differentiation of stem cells into mature blood cells? *Enhancers* and *promoters* are non-coding regions in genome that confers directly to cell type diversity by dynamically switching on and off the expression of target gene(s). I used RNA-seq and ChIP-seq data to study how the expression of the two TFs (PU.1 and CEBPA) shapes the *cis*-regulatory landscape (enhancers) across different differentiation stages of hematopoiesis (blood formation). I show that, in contrast to previous notion, CEBPA knockout

alters the enhancer landscape much earlier in the differentiation process. I show distinct kinetics in the activity of enhancers bound by the two TFs, where late activity of enhancers is driven almost exclusively by CEBPA. 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. This pattern as being a more direct reflection of actual biological events occurring at these elements (nucleosome rearrangement in this case) helped us in detecting genomic regions involved in active regulation with high specificity (Pundhir et al. 2016). This research was listed as a featured article by Danish Stem cell association (DanStem).

Future research directions

My research aims to understand the role *non-coding RNAs* and *epigenome* play in cell identity, its differentiation, and if aberrant to its malignant state. I plan to use my skills in machine learning, statistics and programming to develop computational algorithms to identify patterns (transcriptomic and epigenetic) unique to various cell types or their differentiation stages, and study how these patterns changes during malignant state. Most of the NGS data required for these analysis, I plan to obtain from public repositories (GEO, ENCODE, FANTOM, Roadmap Epigenomics) and from my present collaborators. In the following paragraphs, I list some of the specific questions that I aim to address in the next five years.

4. How do histone modifications shape transcriptomic landscape during hematopoiesis? We now know that different combinations of histone modifications are necessary for the precise regulation of gene expression by *cis*-regulatory elements (enhancers and promoters) and protein complexes (PRC2). I am interested in understanding the role of two such combinations: a) H3K4me1 and H3K4me3; and, b) H3K4me3 and H3K27me3 during formation of mature blood from stem cells in mouse (hematopoiesis). While H3K4me1 and me3 individually defines the location of an enhancer and promoter, respectively, a subset of regions carries both these marks or includes H3K27me3. I want to understand on which regions does it happen, is it consistent during differentiation, is it proximal to any specific functional class of genes or enriched for any specific group of transcription factors? I plan to use ATAC-seq, ChIP-seq (epigenetic) and RNA-seq data corresponding to: a) eleven differentiation stages of hematopoiesis; and, b) knockout of protein responsible for H3K27me3 modification, available from GEO (GSE60103) and my collaborators in Denmark, respectively.

5. How H2A.Z modification at transcription start sites (TSS) effect hippocampal transcription landscape? In this work, I along with my collaborators from Canada seek to understand how age-dependent variations in histone variant H2A.Z deposition in the rat genome associates with transcriptional changes that mediate hippocampal development. Some of the most striking changes observed occur at the transcription start site. I plan to use our computational method for the analysis of bimodal peak pattern (Pundhir et al. 2016) to study how H2A.Z pattern flanking TSS correlates with the extent of hippocampal transcription.

6. 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 been suggested to play a role in higher-level organization of chromatin into loops, domains and compartments within the nucleus. These two roles of histone modifications are perhaps inter-dependent and in my future research I would like to study this relationship. Similarly, I am interested in understanding the role of long non-coding RNAs (lncRNAs) in regulating histone environment and expression of genes. Such a regulation has recently been shown for *Xist* lncRNA, which is involved in silencing one of the two X chromosomes in females (McHugh et al. 2015). I plan to train machine-learning methods (deep neural networks) on histone modification and lncRNA expression data to understand their relationship with chromatin organization.

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 and optimization of previous computational algorithms.

References

Griffiths-Jones S, Hui JHL, Marco A, Ronshaugen M. 2011. MicroRNA evolution by arm switching. *EMBO Rep* **12**: 172–7.

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

McHugh CA, Chen C-K, Chow A, Surka CF, Tran C, McDonel P, Pandya-Jones A, Blanco M, Burghard C, Moradian A, et al. 2015. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**: 232–236.

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

Pundhir S, Gorodkin J. 2015. Differential and coherent processing patterns from small RNAs. *Sci Rep* **5**: 12062.

Pundhir S, Gorodkin J. 2013. MicroRNA discovery by similarity search to a database of RNA-seq profiles. *Front Genet* **4**: 133.

Pundhir S, Hannibal TD, Bang-Berthelsen CH, Wegener A-MK, Pociot F, Holmberg D, Gorodkin J. 2014. Spatially conserved regulatory elements identified within human and mouse Cd247 gene using high-throughput sequencing data from the ENCODE project. *Gene* **545**: 80–7.