SACHIN PUNDHIR RESEARCH STATEMENT

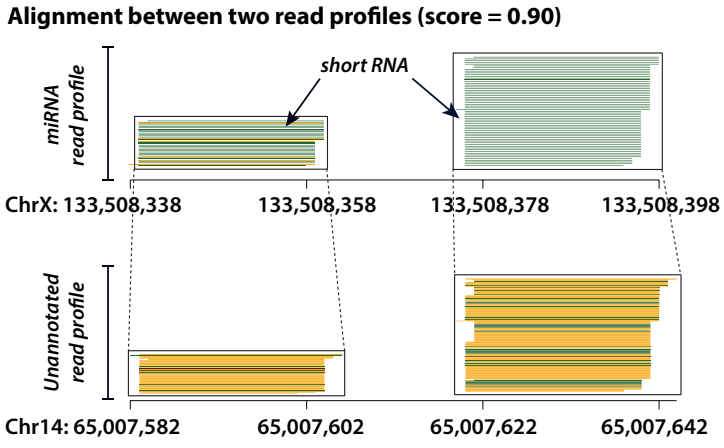
My research interest is computational biology, a rapidly growing research area where computational and statistical methods are applied to solve the biological problems. Completion of human genome project in 2001 and recent advancement in high-throughput sequencing (HTS) technology has revolutionized the field of molecular biology. The genome and transcriptome of many prokaryotic and eukaryotic organisms is publicly available now and onus lies on the scientific community to analyze this data to make meaningful biological inferences. An essential step towards this direction is to develop innovative and efficient computational algorithms for large-scale data analysis.

Thanks to my interdisciplinary background, my pre-doctoral and doctoral research covers a diversity of topics spanning from developing algorithms for the analysis of prokaryotic (virulence factor identification) and eukaryotic genome (non-coding RNA prediction) to analyzing HTS data for the identification of functional (post-transcriptional processing) and regulatory (*cis*- and *trans*-) regions in the human genome. In the following paragraphs, I summarize my previous research work, subdivided into three categories, and outline my future research plans.

Doctoral research: efficient prediction of non-coding RNAs using high-throughput sequencing (HTS) data

One of the significant findings from the last decade has been that the inherent complexity of higher organisms, such as human in comparison to much simpler organisms, such as worm lies in the non-protein coding part of its genome. In human, although, ~75% of the genome is transcribed (1), only ~1.5% encodes for protein (2). What fraction of the non-protein coding transcriptome is actually functional is thus far not understood and is subject to much debate within the scientific community (3, 4). Therefore, it is important to determine characteristic features in the transcriptomic data that can aid to identify functional transcripts from the non-functional ones, and also to understand their finely tuned regulatory mechanism.

During my Ph.D., I have worked on one such feature, which arises upon post-transcriptional processing of many transcripts in the form of short RNA fragments, inside the cell. When captured during HTS experiments and mapped back to the host genome, these fragments often leads to a unique signature of the transcript in the form of its '**read profile**' (5). Essentially, a read profile represents the result of an underlying processing mechanism that a transcript has gone through. Comparison of read profiles can provide valuable insights into the diversity of functional non-protein coding RNAs and potentially the regulation in their processing mechanisms.

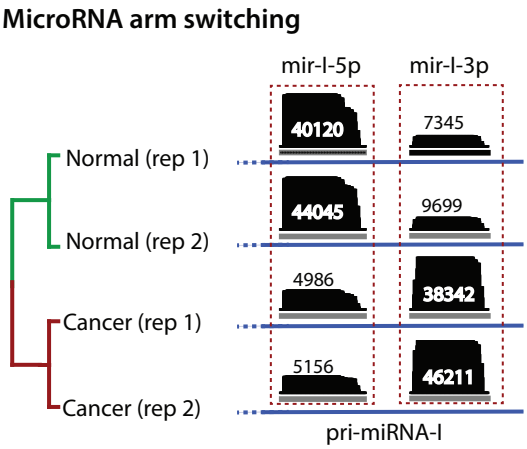


Since, there was no method available for comparing the similarity between two read profiles, I along with my co-author developed a computational method for the optimal alignment of two read profiles (5). We showed its application in the accurate classification of three major classes of non-coding RNAs i.e. miRNA, snoRNA and tRNA. Later, I utilized this

method for the efficient prediction of microRNAs (miRNA), solely based on HTS data (6). Although this method showed performance similar to previously published methods for miRNA prediction, I showed its advantage in its ability to predict putative miRNAs in genomic regions that are devoid of RNA secondary structure information. This information can be missing either due to low sequence conservation across multiple organisms that many tools like RNAz (7) require or due to inherent limitation of tools based on single sequence to predict a RNA secondary structure.

Doctoral research: using high-throughput sequencing data to identify functional and regulatory regions in the eukaryotic genome

In the later part of my Ph.D., I explored the possibility of using the concept of 'read profiles' to understand the regulation in post-transcriptional processing of small RNAs. One good example of such a regulation 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 (8). To address this, I first needed a small RNA-seq data performed on biological replicates of multiple tissues. Currently ongoing ENCODE project has generated a wealth of HTS data performed on multiple human and mouse cell lines (9). I used one such small RNA-seq dataset fulfilling our requirements. Another 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 read profiles between multiple tissues sequenced at variable sequencing depths. On literature search, I identified one such method that is widely used in differential gene expression analysis software such as DESeq (10). I tuned this method along with integrating it with widely known Fisher's exact test for the comparison of read profiles. On analyzing the small RNA-seq data from the ENCODE project, I identified, apart from miRNA, numerous examples of snoRNA and tRNA exhibiting a phenomenon similar to 'arm-switching' (11). Besides, I also observed genomic regions in proximity to Transcription start sites (TSS) that are processed precisely to produce small RNAs of length <22 nt across all the tissues analyzed. In a separate study, I again utilized the HTS data from the ENCODE project to identify two *cis*-regulatory regions (enhancer and promoter) and a long non-coding RNA within the characteristically long (~70 kb) first intron of the *Cd247* gene (12). I studied this gene due to its crucial role in diabetes.



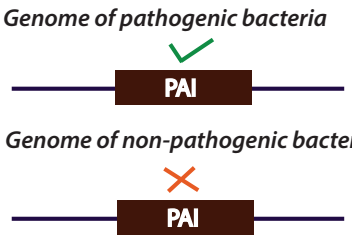
Pre-doctoral research: prediction of pathogenicity islands in prokaryotic genome

Pathogenicity islands (PAIs) are distinct chromosomal regions of pathogenic bacteria that contain genes encoding for virulence factors, for example adhesins and toxins (13). The virulence factors (proteins) help a pathogen in gaining access inside a host cell and evading the host immune system. Due to increase in bacterial resistance to conventional antibiotics, identification of novel drug targets is crucial, and being vital for the bacterial pathogenesis, PAI-encoded virulence factors can be ideal drug targets.

Common genetic features associated with pathogenicity islands are the presence of one or more virulence genes, significant composition bias from core genome (%GC bias, dinucleotide bias and codon bias), presence in pathogenic species, while being absent in benign relatives, tRNA gene acting as insertion site and proximity with mobile genetic elements like integrase,

transposase and insertion sequence (IS) elements. Most of the tools developed for *in-silico* detection of PAIs were based on measuring nucleotide composition bias, which limits their prediction to genomic regions that are acquired by horizontal gene transfer, which may or may not be PAIs. Applications like PAIDB overcome this shortcoming by integrating composition-based search with similarity search against published PAIs (14). This approach, however, has a limitation that the detected PAIs are limited by the dataset of known PAIs.

To address these limitations, I worked on developing a method (15) to identify known and predict novel PAIs. To increase the efficiency, I devised two novel features: a) A 'compare genome feature' using which I can determine if a PAI predicted in a pathogenic bacteria (say *E. coli* K-12) is present or absent in its benign relative (say *E. coli* strain 536), and; b) A 'virulence factor database' (VFDB) consisting of protein sequence of known virulence factors. Using these two features, I strengthened the PAI predictions by ensuring that they are absent in the benign relative (compare genome feature) and one or more proteins encoded from the potential PAIs show significant similarity to VFDB. On performance evaluation, I observed good agreement in the predictions with the already known PAIs from the literature. The developed method was later reviewed as one of selected method for the prediction of genomic and pathogenicity islands in a nature review article (16).



***PredictBias [15]* and *SSPred [17]* are computational methods to predict the pathogenicity islands (PAIs) in bacterial genome.**

As an extension, I later developed a prediction server based on machine learning approach, Support Vector Machine (SVM) for the efficient prediction of virulence factor proteins in bacteria (17). One of the major challenges I faced during this project was how to address the imbalance in positive and negative training dataset, which greatly compromised the prediction efficiency. To solve this, I created an ensemble of SVM modules trained on a balanced subset of training dataset. On 5-fold cross-validation, the trained models showed an accuracy of 89.73%.

Post-doctoral and future research directions

In my post-doctoral research, I am working on the computational analysis of another form of high-throughput sequencing (HTS) data, named ChIP-seq (18). From this analysis, I aim to determine genome-wide distribution of regions where a Transcription factor is bound, thus being involved in the regulation of gene expression or where the DNA exists in uncoiled formation as it happen during transcription (19).

In my immediate future, I plan to employ my skills, learnt during the analysis of RNA-seq and ChIP-seq data, to determine characteristic features that are present in the actively transcribed region of a eukaryotic genome but are absent in the dormant region. Based on these characteristics, I aim to develop computational algorithms to predict a genomic region as actively transcribed or otherwise. An immediate application of such a method would be in studying the temporal and spatial regulation of gene expression between different physiological conditions, such as normal and cancer. Some of the points that encourage me to work in this direction are:

1. Recent studies have identified several shorts RNAs, produced during gene transcription (20) or regulation (21), which form a unique signature when sequenced and mapped back to the host genome in the form of their read profile. These signatures, however, have not yet been analyzed, for example their comparison between different physiological conditions, such as normal and cancer. During my

doctoral studies, I have intensively studied and developed computational algorithms to analyze the read profiles associated with non-coding RNAs. Therefore, I would like to extend their analysis for understanding how the expression of various genes within the genome is regulated between different physiological conditions.

2. Much of the high-throughput sequencing data required for the proposed analysis is publicly available and therefore would not be a limiting factor in initiating the study.

There are also some challenges that I foresee during such a study:

1. Statistical modeling of randomly mapped reads to the genome would be required to filter authentic signals from the background noise.
2. Analysis of high-throughput sequencing data requires extensive amount of computer infrastructure and optimization of the computational algorithms for speed and efficiency.
3. The most important criterion of success for bioinformatics algorithms is being able to solve real biological problems. Therefore, it would be crucial to validate some 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 15 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 prediction efficiency of computational algorithms.

References

1. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. Nature Publishing Group; 2012;489(7414):101–8.

2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. Nature Publishing Group; 2001;409(6822):860–921.

3. Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. On the immortality of television sets:function in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. Oxford University Press; 2013;5(3):578–90.

4. Doolittle WF. Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci. National Acad Sciences; 2013;110(14):5294–300.

5. 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. 2012 Jan 1;28(1):17–24 (*equal contribution).

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

7. Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF. RNAZ 2.0: Improved noncoding RNA detection. Pac Symp Biocomput. 2010;15:69–79.

8. 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.

9. Encode T, Consortium P. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011 Apr;9(4):e1001046.

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

11. **Pundhir S**, Gorodkin J. From microRNA arm switching to differential and coherent post-transcriptional processing of small RNAs in human. Submitted. 2014;

12. **Pundhir S**, Dahlbæk T, Bang-Berthelsen CH, Wegener AMK, 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. 2014.

13. Schmidt H, Hensel M. Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev. Am Soc Microbiol; 2004;17(1):14–56.

14. Yoon SH, Park Y-K, Lee S, Choi D, Oh TK, Hur C-G, et al. Towards pathogenomics: a web-based resource for pathogenicity islands. Nucleic Acids Res. Oxford Univ Press; 2007;35(suppl 1):D395–D400.

15. **Pundhir S**, Vijayvargiya H, Kumar A. PredictBias: a server for the identification of genomic and pathogenicity islands in prokaryotes. In Silico Biol. 2008;8(3-4):223–34.

16. Langille MGI, Hsiao WWL, Brinkman FSL. Detecting genomic islands using bioinformatics approaches. Nat Rev Microbiol. Nature Publishing Group; 2010 May;8(5):373–82.

17. **Pundhir S**, Kumar A. SSPred: A prediction server based on SVM for the identification and classification of proteins involved in bacterial secretion systems. Bioinformation. 2011 Jan;6(10):380–2.

18. Schmidt D, Wilson MD, Spyrou C, Brown GD, Hadfield J, Odom DT. ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. Methods. Elsevier Inc.; 2009 Jul;48(3):240–8.

19. Euskirchen GM, Rozowsky JS, Wei C, Lee WH, Zhang ZD, Hartman S, et al. Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. Genome Res. Cold Spring Harbor Lab; 2007;17(6):898–909.

20. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014 Mar 26;507(7493):455–61.

21. Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, et al. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. Genome Res. 2012 Sep;22(9):1735–47.