Aus dem Institut für Medizinische Bioinformatik und Biostatistik

Geschäftsführender Direktor: Prof. Dr. Ho Ryun Chung

des Fachbereichs Medizin der Philipps-Universität Marburg

Identifying genome-wide transcription units from histone modifications using EPIGENE

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften dem Fachbereich Medizin

der Philipps-Universität Marburg

vorgelegt von

Anshupa Sahu

aus Jeypore, India

Marburg, 2020

Angenommen vom Fachbereich Medizin der Philipps-Universität Marburg am: 28.10.2020

Gedruckt mit Genehmigung des FachbereichsMedizin.

Dekan: Prof Dr. Rolf Müller

Referent: Prof Dr. Ho Ryun Chung

1. Korreferent: Prof Dr. Dominik Heider

Acknowledgements

First and foremost, I would like to thank my supervisor Prof.Dr. Ho Ryun Chung, Institute of Medical Bioinformatics and Biostatistics (IMBB) for providing me the opportunity to work on such an interesting problem. Our discussions contributed greatly to my knowledge of transcription regulation and computational epigenetics. I would like to thank him for the freedom to explore my ideas and pursue my scientific interests, which I believe, made me a better scientist. His guidance and suggestions helped me to grow both professionally and personally.

I would like to thank my former colleagues in the Max Planck Institute for Molecular Genetics, Berlin (MOLGEN): Sarah Kinkley, Alisa Fuchs, Anna Ramisch, Giuseppe Gallone and Tobias Zehnder for the inspiring discussions on transcription regulation and epigenetics data analysis that helped me to develop new ideas and analysis strategies for my projects. I am also thankful to Donald Buczek, Thomas Kreitler, and Paul Menzel (IT Services, MOLGEN), who were extremely helpful with the technical problems I encountered during my Ph.D.

Special thanks to my current colleagues at IMBB Clemens Thoelken and Till Adhikary, who influenced considerably my knowledge on transcriptomic data analysis and tumor biology. Especially Clemens, I greatly appreciate his patience and readiness to help me with any problems I encountered during my Ph.D. I am also thankful to Petra Fischer (Secretary, IMBB) for helping me with the administrative formalities during my Ph.D.

I owe my deepest gratitude to my family for believing in me and for always being supportive. Above all and everything, I want to thank my friend Manish Goel, for his constant support and encouragement that always kept me motivated during my Ph.D.

Contents

List of Figures

List of Tables

List of Abbreviations

AUC	Area Under Curve
bp	base pair
cDNA	Complementary DNA
ChIP-seq	Chromatin Immunoprecipitation Sequencing
CTD	C-terminal domain
DBN	Dynamic Bayesian Network
DNA	Deoxyribonucleic acid
EGA	European Genome-phenome Archive
ENA	European Nucleotide Archive
GEO	Gene Expression Omnibus
GRO-cap	Capped Global Run-On Sequencing
GRO-seq	Global Run-On Sequencing
HMM	Hidden Markov Model
lncRNA	long non-coding RNA
mRNA	messenger RNA
NET-seq	Native Elongating Transcript Sequencing
PIC	Pre-Initiation Complex
Pol II	RNA Polymerase II
PRC	Precision-Recall Curve
PRO-cap	Capped Precision nuclear Run-On sequencing
RNA	Ribonucleic acid
ROC	Receiver Operating characteristics Curve
snoRNA	small nucleolar RNA

snRNA	small non-coding RNA
TF	Transcription Factor
TSS	Transcription Start Site
TTS	Transcription Termination Site
TT-seq	Transient Transcriptome Sequencing
TU	Transcription Unit

Publication

Sahu, A., Li N., Dunkel I *et al.* EPIGENE: Genome-wide transcription unit annotation using a multivariate probabilistic model of histone modifications, *Epigenetics and Chromatin* **13**, 20 (2020). https://doi.org/10.1186/s13072-020-00341-z

Authors list: Anshupa Sahu[a, b], Na Li[b, c], Ilona Dunkel[b], Ho-Ryun Chung[a, b]

Author affiliations: [a] Institute of Medical Bioinformatics and Biostatistics, Philipps University of Marburg, 35037 Marburg, Germany; [b] Otto-Warburg-Laboratory, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany; [c] Guangzhou Institute of Pediatrics, Guangzhou Women and Children's Medical Center, Guangzhou, 510623, China.

Authors' contributions: The project was conceived by HC. AS implemented EPIGENE and performed all the analyses. NL performed the ChIP-seq experiment for histone modifications in K562. ID performed the ChIP-seq experiment for RNA Polymerase II in K562.

Abstract

With the successful completion of the human genome project and the rapid development of sequencing technologies, transcriptome annotation across multiple human cell types and tissues is now available. Accurate transcriptome annotation is critical for understanding the functional as well as the regulatory roles of genomic regions. Current methods for identifying genome-wide active transcription units (TUs) use RNA sequencing (RNA-seq). However, this approach requires large quantities of mRNAs making the identification of highly unstable regulatory RNAs (like microRNA precursors) difficult. As a result of this complexity in identifying inherently unstable TUs, the transcriptome landscape across all cells and tissues remains incomplete. This problem can be alleviated by chromatin-based approaches due to a well-established correlation between transcription and histone modification.

Here, I present EPIGENE, a novel chromatin segmentation method for identifying genome-wide active TUs using transcription-associated histone modifications. Unlike existing chromatin segmentation approaches, EPIGENE uses a constrained, semi-supervised multivariate Hidden Markov Model (HMM) that models the observed combination of histone modifications using a product of independent Bernoulli random variables to identify the chromatin state sequence underlying an active TU.

Using EPIGENE, I successfully predicted genome-wide TUs across multiple human cell lines. EPIGENE predicted TUs were enriched for RNA Polymerase II (Pol II) at the transcription start site (TSS) and in gene body indicating that they are indeed transcribed. Comprehensive validation using existing annotations revealed that 93% of EPIGENE TUs can be explained by existing gene annotations and 5% of EPIGENE TUs in HepG2 can be explained by microRNA annotations. EPIGENE predicted TUs more precisely compared to existing chromatin segmentation and RNA-seq based approaches

across multiple human cell lines. Using EPIGENE, I also identified 232 novel TUs in K562 and 43 novel cell-specific TUs in K562, HepG2, and IMR90, all of which were supported by Pol II ChIP-seq and nascent RNA-seq evidence.

Zusammenfassung

Mit dem erfolgreichen Abschluss des Humangenomprojekts und der raschen Entwicklung von Sequenzierungstechnologien ist nun die Annotation von Transkriptomen über mehrere menschliche Zelltypen und Gewebe hinweg verfügbar. Eine genaue Annotation des Transkriptoms ist entscheidend für das Verständnis der funktionellen und regulatorischen Rolle genomischer Regionen. Aktuelle Methoden zur Identifizierung genomweiter aktiver Transkriptionseinheiten (TUs) verwenden die RNA-Sequenzierung (RNA-seq). Dieser Ansatz erfordert jedoch große Mengen an mRNA, was die Identifizierung von hochinstabilen regulatorischen RNAs (wie microRNA-Vorläufern) schwierig macht. Aufgrund dieser Komplexität bei der Identifizierung von inhärent instabilen TUs bleibt die Transkriptomlandschaft über alle Zellen und Gewebe hinweg unvollständig. Dieses Problem kann durch Chromatin-basierte Ansätze aufgrund einer gut etablierten Korrelation zwischen Transkription und Histonmodifikation reduziert werden.

Hier präsentiere ich EPIGENE, eine neuartige Chromatinsegmentierungsmethode zur Identifizierung genomweiter aktiver TUs unter Verwendung transkriptionsassoziierter Histonmodifikationen. Im Gegensatz zu bestehenden Ansätzen zur Chromatinsegmentierung verwendet EPIGENE ein eingeschränktes, halbüberwachtes multivariates Hidden Markov-Modell (HMM), das die beobachtete Kombination von Histonmodifikationen unter Verwendung eines Produkts unabhängiger Bernoulli-Zufallsvariablen modelliert, um die einer aktiven TU zugrunde liegende Chromatin-Zustandssequenz zu identifizieren.

Mit EPIGENE konnte ich erfolgreich genomweite TUs über mehrere menschliche Zelllinien hinweg vorhersagen. Von EPIGENE vorhergesagte TUs wurden an der Transkriptionsstartstelle (TSS) und im Genkörper auf RNA-Polymerase II (Pol II) angereichert, was darauf hinweist, dass sie tatsächlich transkribiert sind. Eine umfassende

Validierung unter Verwendung vorhandener Annotationen ergab, dass 93% der EPIGENE-TUs durch vorhandene Genannotationen und 5% der EPIGENE-TUs in HepG2 durch microRNA-Annotationen erklärt werden können. EPIGENE prognostizierte TUs genauer im Vergleich zu bestehenden Ansätzen zur Chromatinsegmentierung und RNA-Sequenz über mehrere menschliche Zelllinien hinweg. Unter Verwendung von EPIGENE identifizierte ich auch 232 neue TUs in K562 und 43 neue zellspezifische TUs in K562, HepG2 und IMR90, die alle durch Pol II ChIP-seq- und entstehende RNA-seq-Beweise gestützt wurden.

1 Introduction

Transcription is one of the fundamental processes of life and is necessary for the development of living organisms. It involves the formation of single-stranded mRNA from a double-helical DNA template. Transcription is carried out by RNA Polymerase II (Pol II) in the cell nucleus. The synthesized mRNA is then transported to the cytoplasm, where it is translated into proteins by a multi-protein complex called the ribosome. In addition to mRNA, Pol II transcribes several other kinds of non-coding RNAs like long non-coding RNA (lncRNA), small non-coding RNA (microRNA, snRNA, snoRNA) and other stable and unstable RNAs such as stable unannotated transcripts (SUT), cryptic unstable transcripts (CUT) and enhancer RNAs (Jacquier, 2009; Xu *et al.*, 2009; Kim *et al.*, 2010). Besides Pol II, there exist two other kinds of RNA Polymerases, RNA Polymerase I (Pol I) and RNA Polymerase III (Pol III). While, Pol I transcribes ribosomal RNAs, which are part of the ribosome, Pol III transcribes transferRNAs (tRNA), which are involved in the transportation of amino acids to the ribosome, where they are further incorporated to proteins.

1.1 Transcription units and their function

The transcribed DNA template is called a transcription unit (TU) which can either be a protein-coding gene or a precursor for regulatory RNA. A TU has three components: promoter, RNA-coding sequence also referred to as the actively transcribed region and terminator. The promoter is the region of DNA where RNA Polymerase II binds and initiates transcription. The promoter region is followed by the DNA sequence that primarily codes for a protein or a regulatory RNA. Transcription ends at the terminator region which contains the Poly-A signal that instructs Pol II to terminate transcription (**Figure 1**).

Transcription of TUs can produce mRNA as well as other regulatory RNA such as lncRNA, microRNA, exogenous small interfering RNA (siRNA), and piwi-interacting RNA (piRNA). The polyadenylated mRNA is spliced and translated to proteins that carry out vital biological processes necessary for the survival of an organism, whereas, regulatory RNAs act as sequence-specific transcriptional and post-transcriptional regulators of gene expression (Bartel, 2004; Meister and Tuschl, 2004; Zamore and Haley, 2005; Kim and Nam, 2006).

Figure 1: Structure of a transcription unit

Transcription is regulated by distal and proximal elements called "enhancers" and "promoters" respectively. These regulatory elements contain the binding sites for transcription factors (TFs) that decide when a TU is active and how abundantly it is transcribed. Therefore, most TUs that are active in certain conditions can be inactive in another. Further complicating the picture, is the presence of TUs with a low steady-state abundance where the transcribed RNA degradation rate exceeds the TU transcription rate (as in the case of microRNA precursors), hence rendering the detection of these TUs difficult. Indeed, recent studies have shown the presence of a large number of such TUs (Preker *et al.*, 2008), some of which have been associated with diseases like HIV, Alzheimer's disease and Cancer (Sethi and Lukiw, 2009; Bail *et al.*, 2010; Shah *et al.*, 2016; Wang *et al.*, 2018; Wang, Qin and Tang, 2019; Zhang *et al.*, 2019). For example, the microRNA precursor, pri-miRNA-223 is a functional lncRNA in Acute Myeloid Leukemia that is rapidly processed out to miR-223 and lncRNA-223. Both these RNAs are expressed at different levels and have distinct functions in the myeloid lineage (Mangiavacchi *et al.*, 2016; Wallace and O'Connell, 2017). This suggests that accurate and efficient identification

of TUs would help in improving our understanding of the transcriptomic landscape and its regulation across cell types and tissues.

1.2 Transcription cycle

Eukaryotic transcription is carried out by Pol II in the cell nucleus. Transcription involves a series of stages, often referred to as "the transcription cycle", in which each stage involves specific proteins and protein modifications. The transcription cycle can be roughly divided into three stages: initiation, elongation, and termination.

1.2.1 Initiation

Transcription begins at the transcription start site (TSS) in which the region surrounding the TSS also called the "promoter" directs accurate transcription. According to several *in vitro* transcription studies, six "general transcription factors" (GTFs) (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH) assemble at the promoter into the pre-initiation complex (PIC) (Nikolov and Burley, 1997) before transcription initiation. It is important to note that PIC assembly can be more variable *in vivo* with the involvement of several other TFs, as different promoters suggest different paths to its recognition (Nikolov and Burley, 1997; Sikorski and Buratowski, 2009).

After the formation of the PIC, transcription initiation begins with the formation of a jaw-like open complex, where TFIIB aids the insertion of double-stranded DNA into the jaw and downstream cleft of Pol II (Sainsbury, Niesser and Cramer, 2013). The DNA-helicase TFIIH separates both the strands and inserts the single-stranded DNA to the active site of Pol II (Sainsbury, Niesser and Cramer, 2013). With the help of the TFIIB-reader domain, Pol II then scans the downstream nucleotides for the TSS and initiates the transcription resulting in the formation of the "initial transcribing complex" (Kostrewa *et al.*, 2009; Sainsbury, Niesser and Cramer, 2013).

1.2.2 Elongation

Pol II transits from initiation to elongation stage at approximately 150 base-pairs (bp) downstream of TSS (Mayer *et al.*, 2010). This involves the exchange of initiation factors with elongation factors and the phosphorylation of Serine 5 and Serine 7 residues in the C-terminal domain (CTD) of Pol II (Mayer *et al.*, 2010; Lidschreiber, Leike and Cramer, 2013). After the exchange of initiation factors with elongation factors, Pol II undergoes Serine 2 phosphorylation at the CTD (Ahn, Kim and Buratowski, 2004), followed by the recruitment of c-Abl (in humans) which triggers binding of the elongation factor Spt6 and suppresses transcription termination by blocking the recruitment of termination factors (Mayer *et al.*, 2015; Burger, Schlackow and Gullerova, 2019).

1.2.3 Termination

Transcription termination occurs at polyadenylation (poly-A) sites, that are marked by the presence of a highly conserved consensus sequence motif AATAAA located 10-30 bp upstream of the site (McLauchlan *et al.*, 1985; Proudfoot, 2011). The poly-A site is recognized by cleavage and specificity factors (CPSF), that processes the RNA by endonucleolytic cleavage and polyadenylation (Richard and Manley, 2009). Transcription continues for several thousand bps after the poly-A site (in humans) and the mechanisms that result in the release of Pol II and transcribed RNA from DNA are still unknown. It is, however, believed that the speed of transcription elongation and stability of RNA-DNA hybrid contributes to the destabilization and release of Pol II from DNA (Skourti-Stathaki, Proudfoot and Gromak, 2011; Fong *et al.*, 2015).

1.3 The histone code of transcription

Eukaryotic DNA is tightly packed into a macromolecular complex of histone proteins and DNA called chromatin. Chromatin can be classified as heterochromatin and euchromatin. Heterochromatin is highly compacted and hence genes located in these regions are repressed. On the contrary, euchromatin has a low degree of compaction, and genes within them can be transcribed.

Chromatin comprises of repeating units of 147 DNA base pairs (bp) wrapped around an octamer of four histones (2 copies each) H2A, H2B, H3, and H4 called the nucleosome. Post-translational chemical modifications to histones in the form of acetylation, methylation, phosphorylation, ubiquitination, and sumoylation play a significant role in transcription. These modifications are added, removed, and recognized by other proteins. Hence, nucleosomes serve as signaling platforms (Turner, 2012) that control the regulatory mechanisms in chromatin by enabling the localized activity of chromatin signaling networks partaking in transcription and other chromatin-related processes (Perner and Chung, 2013). In fact, it has been established that histone modifications correlate with transcription (**Table 1**). For instance, the histone modification H3 lysine 4 trimethylation (H3K4me3) and H3 lysine 27 acetylation (H3K27ac) are positively correlated to transcription initiation and are enriched at promoters (Bernstein *et al.*, 2002; Barski *et al.*, 2007; Creyghton *et al.*, 2010; Karlic *et al.*, 2010) whereas H3 lysine 9 trimethylation (H3K9me3) is common in heterochromatin. H3K4me3 is involved in the recruitment of chromatin remodeling factors like CHD1 and BPTF (Flanagan *et al.*, 2005; Li *et al.*, 2006), which result in the opening of chromatin and also prevents the binding of repressive NuRD and INHAT complexes (Nishioka *et al.*, 2002; Schneider *et al.*, 2004).

Table 1: Core histone modifications and their correlation with transcription

Histone	Correlation to transcription	Location
H3K4me3	positive	promoters
H3K4me1	positive	enhancer
H3K36me3	positive	gene bodies
H3K27ac	positive	promoters, enhancers
H3K27me3	negative	repressed genes
H3K9me3	negative	heterochromatin

This makes the DNA accessible to TFs, hence allowing transcription. Whereas, H3K9me3 is involved in recruiting heterochromatin protein HP1 to genomic regions, hence regulating gene expression and heterochromatin formation.

1.4 Experimental approaches for studying transcription

Currently, multiple sequencing techniques exist for studying various aspects of transcription. These can be (a) RNA-based, that measure the abundance of processed or nascent RNA and characterize the expression levels of TUs in cells, or (b) chromatin-based, that detect DNA-protein interactions and allow *in vivo* genome-wide identification of binding sites for TFs, histone modifications, and other proteins.

1.4.1 RNA-seq

Principle

RNA sequencing is a widely used technique for measuring the transcriptome across cell types and tissues. It identifies the complete set of processed RNAs and their isoforms, as well as measures their abundance for a developmental stage or specific condition (Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008).

The protocol starts with isolating the RNA of interest from a given cell type (**Table 2**). The isolated RNA molecules are either first fragmented and then reverse transcribed to cDNA or vice versa. The cDNA fragments are amplified and sequenced using high-throughput sequencing technology.

Table 2: RNA-sequencing methods and their benefits

RNA-seq method	Benefits
mRNA seq	polyA selection to sequence mRNA for gene expression analysis
small RNA seq	evaluation of small RNAs (< 200bp) and discovery of novel small RNA
total RNA seq	enables analysis of coding and non-coding RNAs (> 200bp)
targeted RNA seq	sequence and analyze specific TUs of interest

Limitation

RNA-seq approaches require relatively high quantities of target RNAs, hence, limiting the accurate identification of highly unstable TUs (like precursors of microRNA, enhancer RNA), for which the transcribed RNA concentration is quite less. These can be identified by total RNA-seq but this requires an ultra-deep sequencing or a targeted RNA-seq or small RNA-seq for identifying small RNAs < 200 bp. Additionally, RNA-seq data is prone to sequencing errors that negatively impact downstream bioinformatics analysis and can lead to incorrect conclusions regarding the set of active TUs (Le *et al.*, 2016; Tong *et al.*, 2016).

1.4.2 Nascent RNA-seq

Principle

Recently, nascent RNA-seq approaches have also emerged as a valuable method to study Pol II-mediated transcription. These approaches detect nascent RNA or primary RNA from the entire pool of cellular RNA, either by chemically inducing point mutations or by biochemical enrichment. The isolated RNAs are then reverse transcribed, ligated with adapter, amplified, and deep sequenced.

Most nascent RNA-seq approaches differ considerably in their ability to detect or enrich for nascent RNA (**Table 3**). For instance, chromatin isolation based methods such as NET-seq (Mayer *et al.*, 2015), enrich for Pol II-associated RNA and can reliably detect transcription termination site (TTS) (Nojima *et al.*, 2015), while run-on techniques like GRO-cap (Core *et al.*, 2008) and PRO-cap (Kwak *et al.*, 2013), enrich for capped RNA and can reliably detect TSS (Core *et al.*, 2014). Therefore, the choice of the method determines the stage of transcription cycle that can be analyzed and additionally influences the resolution and stringency of data generated (Wissink *et al.*, 2019).

Table 3: Overview of nascent RNA-seq methods

Method	Technique	Protocols
chromatin isolation	Isolation of chromatin-bound RNA by antibodies or high salt washes	Start-seq (Williams *et al.*, 2015), mNET-seq (Nojima *et al.*, 2015)
run-on	Labeling nascent RNAs with 5-bromouridine 5'-triphosphate (BrUTP) in presence of anionic detergent sarkosyl, followed by isolating the labeled RNAs with antibodies	GRO-cap, PRO-cap, GRO-seq
Metabolic labeling	Labeling the living cells by modified ribonucleotides (such as 4-thiouridine (4sU)), followed by affinity purification of labeled RNA	TT-seq (Schwalb *et al.*, 2016)
imaging	Detection of nascent RNAs by fluorescence in situ hybridization of labeled oligos in fixed cells or engineering transcripts to encode hairpin-like structure and *in vivo* recognition by tagged cognate binding proteins	FISH (Bauman *et al.*, 1980), MS2-GFP (Yunger *et al.*, 2010)

Limitations

Although nascent RNA-seq techniques have proven to be a valuable method to study Pol II-mediated transcription, most of these approaches require a high amount of input material and are limited to cell cultures and artificial systems (Gardini, 2017). Besides, each nascent RNA approach was initially designed to answer very specific questions about transcription regulation and hence, identify different stages of transcription such as initial transcribing complex (section 1.2.1), CTD modification (section 1.2.2), TTS, etc. It is important to note that integration and comparison of results from multiple nascent RNA-seq approaches could potentially provide a comprehensive overview of transcription and identify genome-wide active TUs. However, methods performing such an integrative analysis are yet to be developed.

1.4.3 ChIP-seq

Principle

ChIP-seq is a technique to analyze DNA-protein interactions. It involves chromatin immunoprecipitation followed by massively parallel DNA sequencing (Barski *et al.*, 2007; Johnson *et al.*, 2007; Mikkelsen *et al.*, 2007) and is commonly used to detect genome-wide binding sites for histone, transcription factors, and other proteins *in vivo*. Currently, with decreasing sequencing costs, ChIP-seq has become an imperative method to study transcription regulation and epigenetic mechanisms and can be reliably applied for a low number of cells (Gustafsson *et al.*, 2019).

ChIP-seq protocol starts with crosslinking DNA-bound protein to chromatin by exposing the cells to formaldehyde, resulting in covalent bond formation between them. The DNA is fragmented using sonication or enzyme digestion and fragments linked to the protein of interest are isolated using an antibody that recognizes it. The filtered DNA-protein complexes are then reverse cross-linked and the resulting DNA fragments are amplified and sequenced.

Limitation

One of the major limitations of the ChIP-seq technique is its dependence on the quality of the antibody (Park, 2009). Various commercially available antibodies widely differ in quality not just across suppliers but also across batches. Antibody quality can be evaluated but such methods are time consuming and laborious. Additionally, ChIP-seq techniques are comparatively more expensive than RNA-seq and nascent RNA-Seq techniques.

1.5 Computational approaches for identifying transcription units

Several computational methods have been developed for identifying genome-wide active TUs. These methods are either (a) RNA-seq based, that assemble the transcriptome and identify active TUs using RNA-seq data, or (b) ChIP-seq based, that

identify genomic elements by modeling the observed combination of transcription-associated histone modifications.

1.5.1 RNA-seq based

With the decrease in sequencing costs, RNA-seq has emerged as a valuable technique for genome-wide TU detection. As a result, most, TU detection approaches like idba-Tran, Oases, Cufflinks, Trinity, StringTie, etc (Trapnell *et al.*, 2010; Grabherr *et al.*, 2011; Schulz *et al.*, 2012; Peng *et al.*, 2013; Pertea *et al.*, 2015) use RNA-seq data. These approaches assemble RNA-seq reads either *de novo* using graph models or use a reference genome-guided graph model (**Figure 2**).

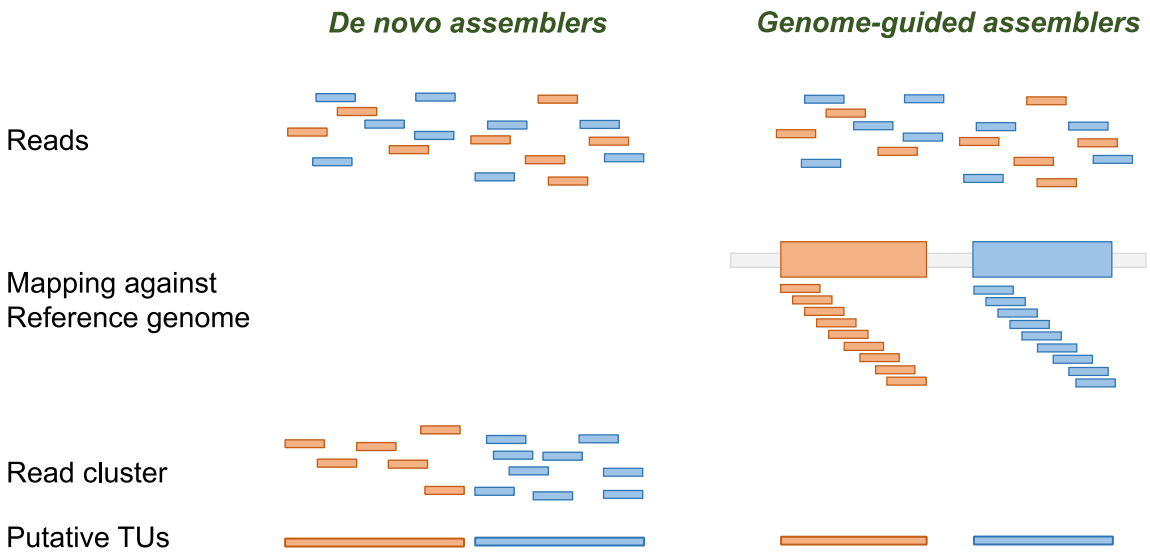


Figure 2: Principle of *de novo* and genome-guided transcriptome assemblers

De-novo assemblers

De-novo assemblers generate contigs based on the input RNA-seq data. Most present-day *de novo* assemblers like Trinity, Oases, idba-Tran rely on de Bruijn graphs generated from the k-mer decomposition of RNA-seq reads. *De novo* transcriptome assembly begins with dividing RNA-seq reads into shorter sequences of length *k* called k-

mers and reconstructing the original sequence by overlapping these k-mer sequences. A major limitation of de Bruijn graphs is its requirement for a k-mer to start at every position along the original sequence so that the graph covers the complete sequence (Chevreux *et al.*, 2004). Additionally, this limitation creates a tradeoff regarding the length of the k-mers. Short k-mers are more likely to fully cover the original sequence but are ambiguous with a single k-mer mapping to multiple reads from multiple TUs. Long k-mers avoid such ambiguity but may not cover the complete sequence of some TUs. As a result, each TU with its unique combination of sequence and abundance levels has a different k-mer length for its optimal assembly. Hence, even when using the same *de novo* assembly algorithm, multiple transcriptome assemblies with varying k-mer lengths generate different sets of contigs with a different set of correctly assembled contigs.

Genome-guided assemblers

The limitations of k-mer decomposition used in de Bruijn graphs are alleviated by genome-guided assemblers like Cufflinks and StringTie (Trapnell *et al.*, 2010; Pertea *et al.*, 2015) that aligns the RNA-seq reads to the reference genome. These assemblers account for introns, by allowing the read mapping for genome-guided assembly to split, such that the first half of the read maps to the exon and the second half maps to the subsequent downstream exon. This read mapping can be performed by split-read mappers like STAR, Top-Hat, HPG-Aligner, HISAT, etc. (Trapnell, Pachter and Salzberg, 2009; Dobin *et al.*, 2013; Kim, Langmead and Salzberg, 2015; Medina *et al.*, 2016). Each of these read mappers uses a different strategy and result in a slightly different read mapping which can influence the quality of the subsequent transcriptome assembly.

Although both *de novo* and genome-guided assemblers reliably detect TUs with high steady-state abundance, detecting unstable TUs such as microRNA precursors remains problematic. This is due to the inherent experimental limitation of RNA-seq that requires relatively high quantities of target RNA.

1.5.2 ChIP-seq based

Since the successful completion of the human genome project and release of the human genome sequence in 2001, several large-scale projects and consortia such as ENCODE, NIH Roadmap Epigenomics, DEEP, Blueprint, and IHEC (Feingold *et al.*, 2004; Bernstein *et al.*, 2010; Adams *et al.*, 2012; DEEP, 2012; Stunnenberg *et al.*, 2016) have been initiated to identify the functional elements of DNA and understand their effects on diseases and human development. To accomplish this, large scale genome-wide transcriptome and epigenome maps were generated, that allowed the analysis of histone modifications and their role in transcription. As a result, in recent years it has been established that essential genomic features such as promoter, enhancer, transcribed regions, and heterochromatin domain exhibit a characteristic and recurrent histone modification pattern commonly referred to as "chromatin state" (Ernst and Kellis, 2012). For example, a combination of H3K27ac and H3K4me1 occurs at active enhancers and a combination of H3K27ac and H3K4me3 is associated with an active promoter region.

The deluge of ChIP-seq data for histone modifications (**Figure 3**) and the association of these modifications to transcription allows for an integrative analysis of histone modifications also referred to as "chromatin segmentation". This allows robust identification of genomic regions like enhancers, promoters, and insulators as well as annotates heterochromatin domain and transcribed regions. Currently, several chromatin segmentation approaches such as ChromHMM, EpiCSeg, chroModule, GENOSTAN, etc. (Ernst and Kellis, 2012; Won *et al.*, 2013; Mammana and Chung, 2015; Zacher *et al.*, 2017) exist that use genome-wide ChIP-seq profiles of histone modifications as input to provide genome-wide chromatin state annotation. These approaches use a variety of machine learning algorithms with the most prominent one being Hidden Markov Models (HMM). HMM is a probabilistic framework used to model a sequence of observations. It assumes that the sequence of observations is generated by the underlying hidden states, that emit observations according to a particular probability distribution.

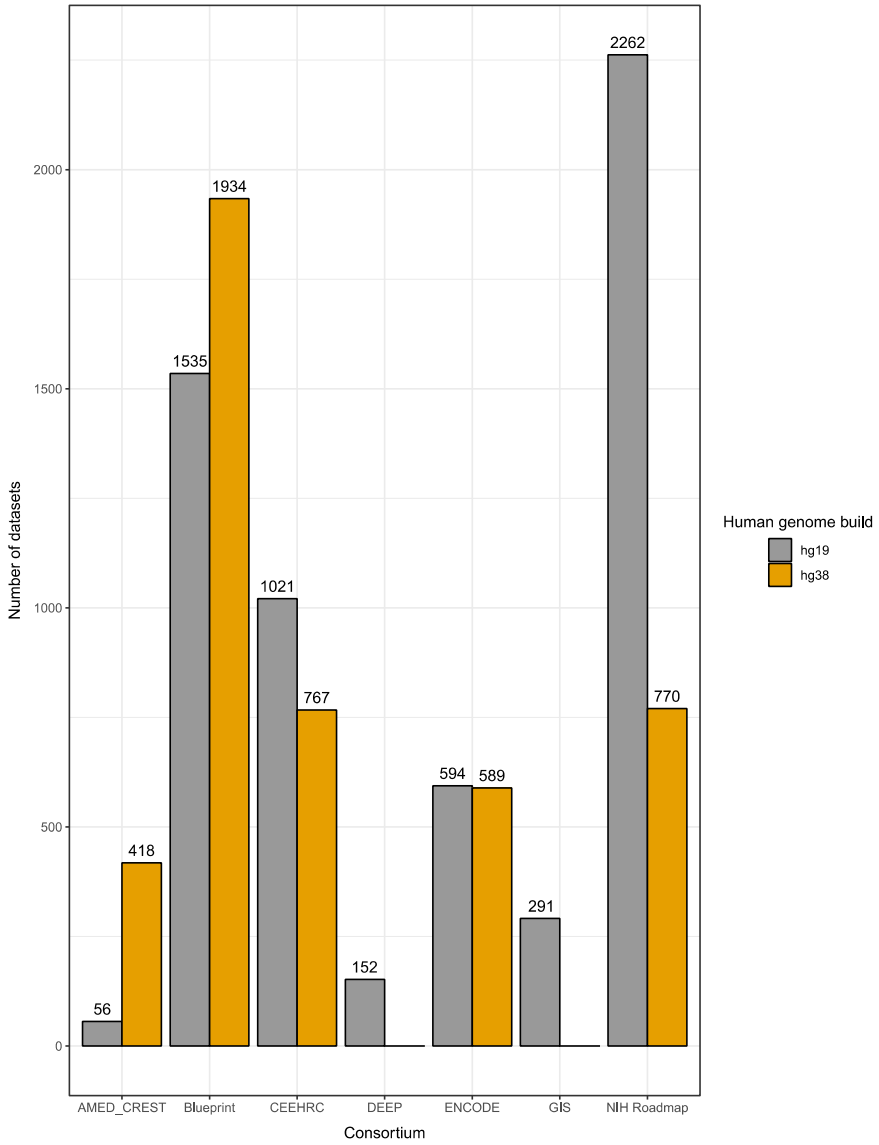


Figure 3: Distribution of IHEC class 1 histone modifications across multiple consortiums

Hence, it is are ideal for identifying chromatin states based on the observed combination of histone modifications. By formal definition, an HMM comprises of:

$Q = q_1, q_2, \dots q_N$	a set of N hidden states
$O = o_1, o_2, \dots o_K$	a sequence of K observations

$T = t_{11}, \dots t_{ij} \dots t_{NN}$	a transition matrix T is a $N \times N$ matrix, where each element t_{ij} represents the probability of moving from hidden state i to hidden state j such that $\sum_{j=1}^{N} t_{ij} = 1 \; \forall i$
$E = e_i(o_t)$	an emission matrix, where each element represents the probability of an observation o_t being generated from state i
$\pi = \pi_1, \pi_2, \dots \pi_N$	the initial probability distribution vector, where each element π_i is the probability of beginning with state i such that $0 < \pi_i < 1, \forall i \, \epsilon \, N$, and $\sum_{i \, \epsilon N} \pi_i = 1$

In the context of modeling the observed combinations of histone modifications, Q represents the set of hidden chromatin states of an HMM. These hidden states are linked by transition probabilities that represent the spatial constraints of how the combination of histone modifications occur relative to each other. The emission probability vector of each hidden chromatin state represents the probability with which a histone modification is observed in that chromatin state. The transition and emission probabilities can either be trained using highly confident histone modification data sets (supervised) or can be learned *de novo* from the input chromatin data (unsupervised) (**Figure 4**). The following sections will briefly present the state-of-art approaches used for genome-wide chromatin state annotation.

Unsupervised HMMs

These HMMs do not rely on prior biological information and therefore require the user to interpret and annotate the learned states based on existing knowledge of functional genomics. Some of the widely used unsupervised HMMs are ChromHMM, EpiCSeg, and GENOSTAN (Mammana and Chung, 2015; Ernst and Kellis, 2017; Zacher *et al.*, 2017). All of these approaches use different variants of unsupervised HMM, that operates on different kinds of inputs. e.g. ChromHMM models the presence or absence of histone modifications in 200 bp bins using a product of independent Bernoulli random

variable and hence the input needs to be binarized for this model, whereas, EpiCSeg models the raw read counts in 200 bp bins using a negative multinomial distribution and hence except for read mapping, it does not rely on preprocessing of data. Both ChromHMM and EpiCSeg provide chromatin state annotation without the strand information. GENOSTAN addresses this issue by using bidirectional HMMs that integrate strand-specific data (e.g. RNA expression) with non-strand specific data (e.g. ChIP-seq) to infer directed chromatin states from genomic data *de novo*. Additionally, all of these approaches use the Baum-Welch algorithm (Baum *et al.*, 1970) to fit the model parameters and infer the hidden chromatin states using Viterbi or posterior decoding (Viterbi, 1967; Fariselli, Martelli and Casadio, 2005).

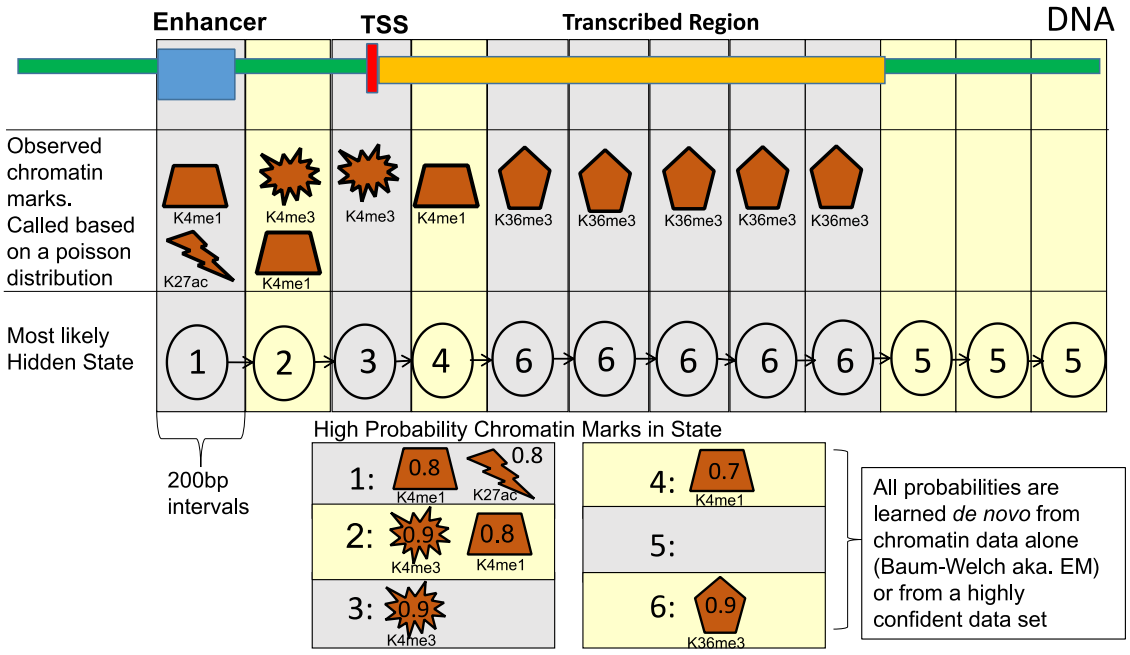


Figure 4: Core principle underlying existing chromatin segmentation HMMs. Chromatin segmentation HMMs model the observed combination of histone modifications that are emitted by a sequence of hidden chromatin states.

Supervised HMMs

These HMMs use high confidence labeled data for training the transition and emission probabilities and hence do not require the user to explicitly interpret the results. An example of a supervised HMM is chroModule (Won *et al.*, 2013) that uses a left-right structured HMM to identify genomic modules such as an enhancer, promoter, transcribed, repressed and background, and incorporate all these modules into one model. It integrates existing biological knowledge into the model by learning the transition and emission probabilities on preselected training sets. The individual models are trained separately using the Baum-Welch algorithm (Baum *et al.*, 1970) and all the modules are then linked to construct the final model. Chromodule operates at 100 bp resolution and each 100 bp bin is assigned to the hidden chromatin state using the Viterbi algorithm (Viterbi, 1967).

Dynamic Bayesian Networks

In addition to HMMs, dynamic bayesian networks (DBN) has also been used to provide genome-wide chromatin state annotations. DBNs are similar to HMM with several hidden chromatin states and multiple observation tracks. A well-known application of DBN for genome-wide chromatin state annotation is Segway (Hoffman *et al.*, 2012). Unlike other approaches, Segway operates at a 1-bp resolution and is comparatively slower than HMM-based approaches (Mammana and Chung, 2015).

Although the above-mentioned chromatin segmentation approaches identify important genome regions such as enhancers, promoters, transcribed regions, etc., they fail to identify genome-wide active TUs as the underlying model does not constrain the chromatin state sequence to begin with a TSS and end with a TTS.

1.6 Aim of the thesis

As described earlier, accurate identification of TUs is essential to better understand the transcriptomic landscape of a cell. Most of the existing approaches either

fail to identify short-lived TUs or are unable to identify TUs due to the usage of a flexible model.

In this thesis, I address these shortcomings by developing a semi-supervised HMM referred to as EPIGENE. EPIGENE models the observed combination of histone modifications to predict genome-wide active TUs. In contrast to existing chromatin segmentation approaches, EPIGENE assigns a direction (forward or reverse) to the TUs which is essential to characterize transcription.

This thesis is divided into three parts. Chapter 2 introduces, validates, and compares EPIGENE predictions using existing gene annotations, Pol II ChIP-seq, nascent-RNA, RNA-seq data, and chromatin segmentation approaches. I compare EPIGENE with existing RNA-seq based and chromatin segmentation approaches and also demonstrate its applicability across cell lines and tissues. Chapter 3 presents the methods and analysis strategies that have been used in this thesis. Finally, I critically discuss EPIGENE and also suggest future improvements in Chapter 4.

2 Results

The results presented in this chapter were published as a peer-reviewed research article in Epigenetics and Chromatin (Sahu et al. 2020). Please refer to Page xi for author contributions.

A longstanding challenge in molecular biology is to elucidate the transcriptomic landscape across cells and tissues. This task becomes even more challenging due to the presence of TUs that gives rise to unstable RNAs like microRNA precursors. Analysis of genome-wide ChIP-seq profiles of histone modifications in humans has established the presence of characteristic histone modification patterns in different parts of a TU. Thus, an integrative analysis of transcription-associated histone modifications can be used to identify genome-wide active TUs.

In this thesis, I validate this hypothesis across multiple cell lines by developing a novel computational method "EPIGENE" that models the observed combination of transcription-associated histone modifications to predict genome-wide active TUs. EPIGENE is the first method that uses histone modifications for identifying genome-wide TUs.

In this chapter, I present the EPIGENE approach and validate EPIGENE predictions with existing gene annotations, RNA-seq, and ChIP-seq datasets. I also discuss the performance of EPIGENE in gold standard datasets and its comparison with existing RNA-seq and chromatin segmentation approaches. Additionally, I demonstrate the applicability of EPIGENE across multiple cell lines and present multiple examples of cell-type specific unannotated TUs and cell-type specific microRNA precursors that could not be identified by existing approaches due to the absence of RNA-seq evidence.

2.1 Schematic overview of EPIGENE

EPIGENE learns the TU state signatures using a multivariate HMM, which probabilistically models the combinatorial presence and absence of IHEC class 1 histone modifications.

2.1.1 EPIGENE input

EPIGENE requires a class matrix as input where each row corresponds to a 200 bp non-overlapping genomic interval called bin and each column corresponds to a histone modification. The values in the matrix represent the presence or absence of the histone modification in the 200 bp bin. The class matrix was computed by dividing the mappable regions of the genome into non-overlapping 200 bp bins and computing the ChIP and control read counts in each bin for each histone modification. The ChIP and control read counts were subsequently converted to the presence and absence calls using normR (Kinkley *et al.*, 2016) (**Figure 5**; see section 3.2).

2.1.2 The EPIGENE model

EPIGENE models the class matrix using a multivariate HMM. The multivariate HMM has 14 TU states and 3 background states (**Figure 6A**), where each TU state represents individual components of a gene such as TSS, exon, intron, etc., and each background state represents genomic regions other than TU such as an enhancer, heterochromatin, etc. The TU states were duplicated, running from TSS to TTS and from TTS to TSS, hence allowing the identification of TUs on both forward and reverse strands.

In contrast to existing chromatin segmentation approaches, the transition and emission probabilities of EPIGENE were trained in a semi-supervised manner, to obtain a probabilistic model for the chromatin state sequence underlying active TUs. The model constrains this chromatin state sequence to always begin with a TSS state, proceed through exon and intron states, and end with a TTS state. Hence, naturally recovering the genomic region that is spanned by a putative TU - a task that is much harder using unconstrained chromatin state models, such as ChromHMM or EpiCSeg.

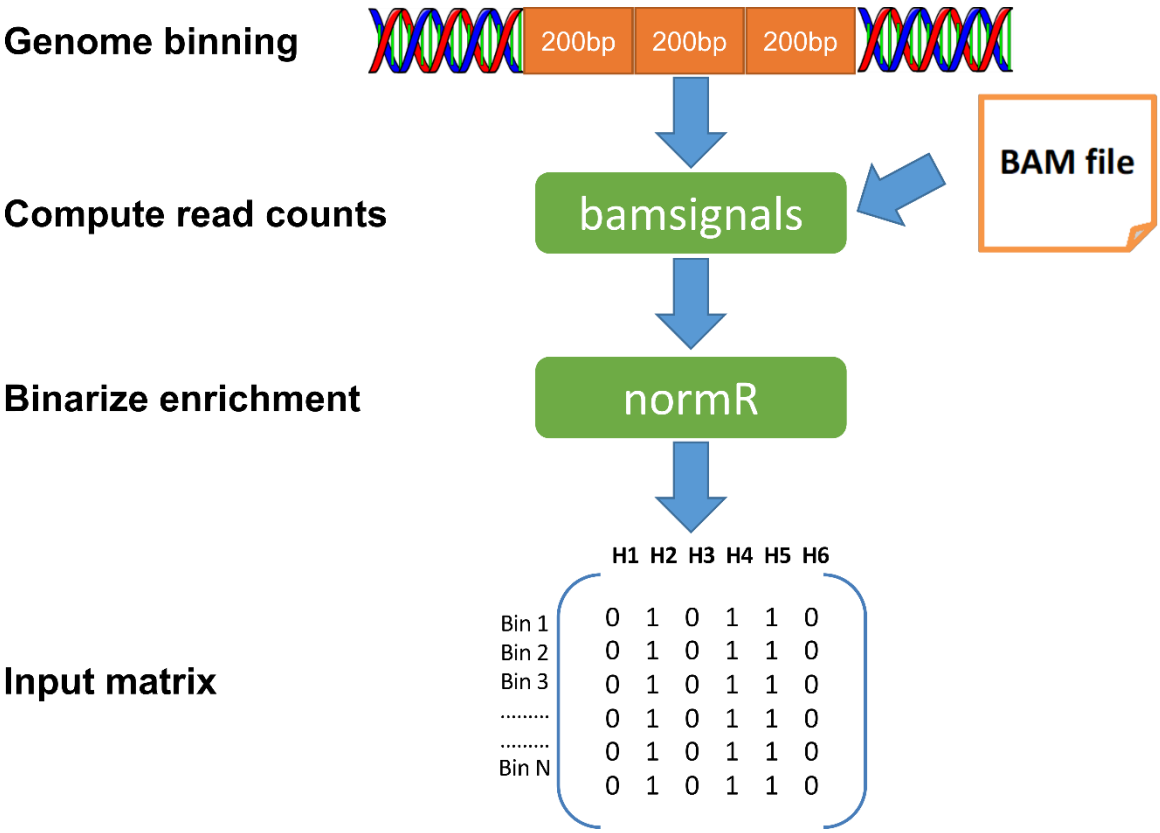


Figure 5: Preparing input data for EPIGENE model

Given the class matrix and the transition and emission probabilities, EPIGENE outputs a bin-state vector where each bin is assigned to TU or background state. The vector is further filtered to obtain active TUs that begin with a TSS state and end with a TTS state (**Figure 6B**) (method details in section 3.3).

2.1.3 EPIGENE model parameters

The transition probabilities of the HMM represents the spatial constraints of how the combination of histone modifications occur relative to each other, and, the emission probabilities represent the probability with which a histone modification occurs in a TU state.

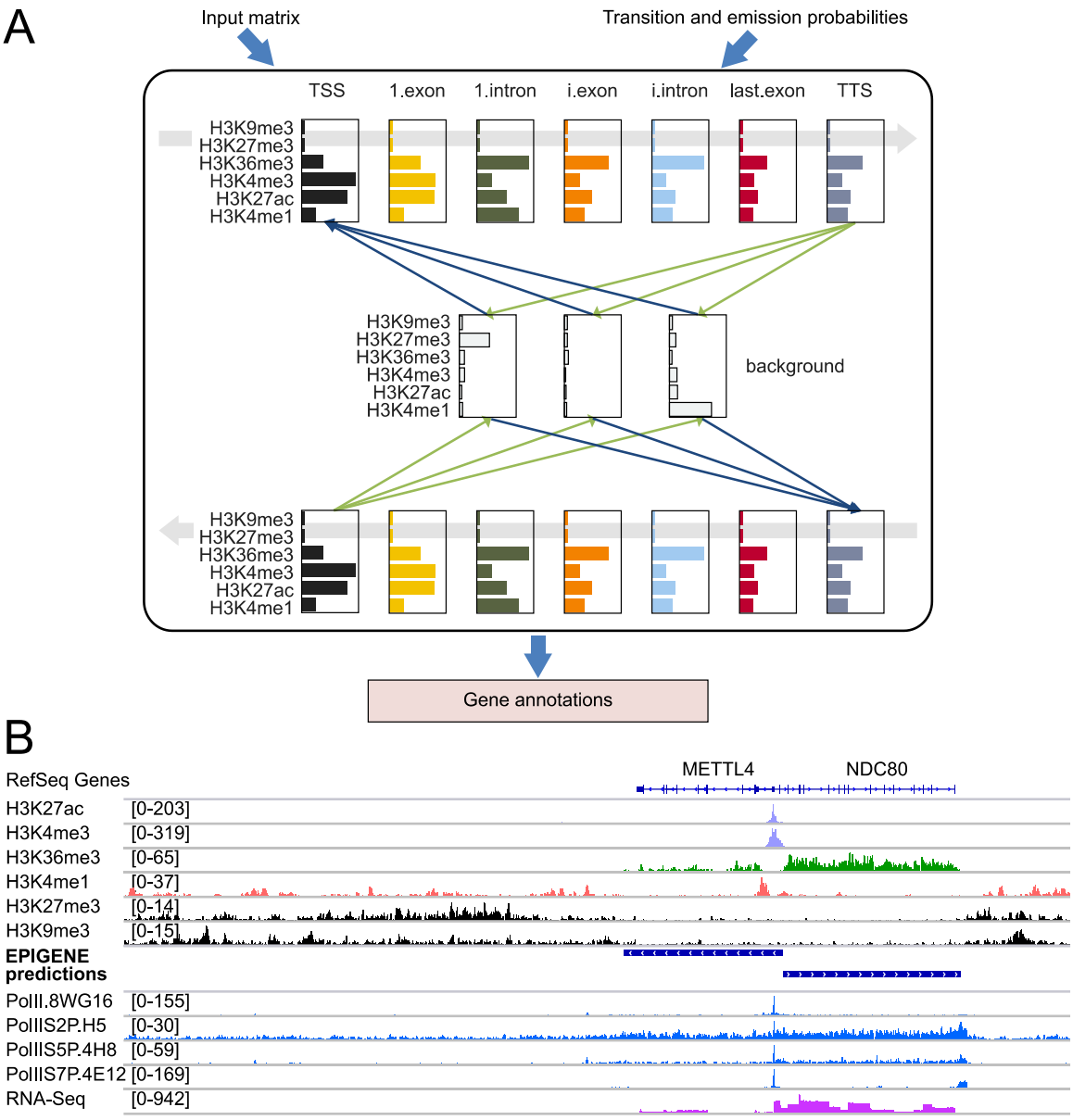


Figure 6: EPIGENE workflow and example of EPIGENE TU (a) Schematic overview of EPIGENE framework. (b) An example of EPIGENE prediction. EPIGENE predictions of METTL4 and NC80 gene show an enrichment of H3K27ac and H3K4me3 at TSS (tracks shown in light violet), H3K36me3 in gene body (tracks shown in green), enhancer mark H3K4me1 few bps upstream or downstream of TSS (tracks shown in pink), RNA Polymerase II in TSS and gene body (tracks shown in blue). The predictions also show an absence of repression marks H3K27me3 and H3K9me3 (tracks shown in black). The corresponding RNA-seq evidence in this genomic region can be seen in the lower-most track (track shown in dark pink)

To probabilistically model the combinatorial absence/presence of different histone modifications and the topology of TU and background states, the transition and emission probabilities were trained in a semi-supervised manner (refer section 3.4). Except for the transition between TSS to TTS and vice versa, the transition probabilities between the TU states were trained using GENCODE (Frankish *et al.*, 2019) annotations. The emission probabilities of the TU states were trained on the GENCODE TUs that show enrichment of Pol II in the reference epigenome, which was generated as per IHEC guidelines (IHEC, 2012) (see section 3.1 and **Table S1**). The emission and transition probabilities between the background states were trained in an unsupervised manner. Additionally, the transition probabilities from or to either the TSS and TTS states and the transition probabilities between TSS and TTS states were trained in an unsupervised manner. This semi-supervised training of model parameters ensures the applicability of the model across multiple cell-types and tissues without the need for additional training or tuning of parameters.

2.2 Validation of EPIGENE predicted TUs

To access the quality of EPIGENE predicted TUs, I validated the predicted TUs with existing gene annotations, RNA-seq, and ChIP-seq evidence in the K562 cell line. The validation with RNA-seq evidence was performed using existing RNA-seq based TU identification approaches, while the validation with ChIP-seq evidence was performed using ChIP-seq profiles of Pol II and histone modifications.

2.2.1 Validation with existing gene annotations and RNA-seq

I created a consensus TU set to investigate the presence of unannotated EPIGENE TUs (24,571 TUs; **Table S2**) and also to estimate the proportion of EPIGENE TUs that are also supported by RNA-seq and existing gene annotations. The RNA-seq TUs were obtained from StringTie (101,656 TUs; **Table S3**) and Cufflinks (32,079 TUs; **Table S4**). A *union* operation between EPIGENE, StringTie, and Cufflinks predictions was

performed to obtain the consensus TU set (refer section 3.7). The consensus TU set containing 24,874 TUs was overlapped with GENCODE and CHESS annotations (Pertea *et al.*, 2018; Frankish *et al.*, 2019) to compute the fraction of annotated and unannotated EPIGENE TUs that are: (a) supported RNA-seq evidence, and (b) exclusive to EPIGENE (**Figure 7**).

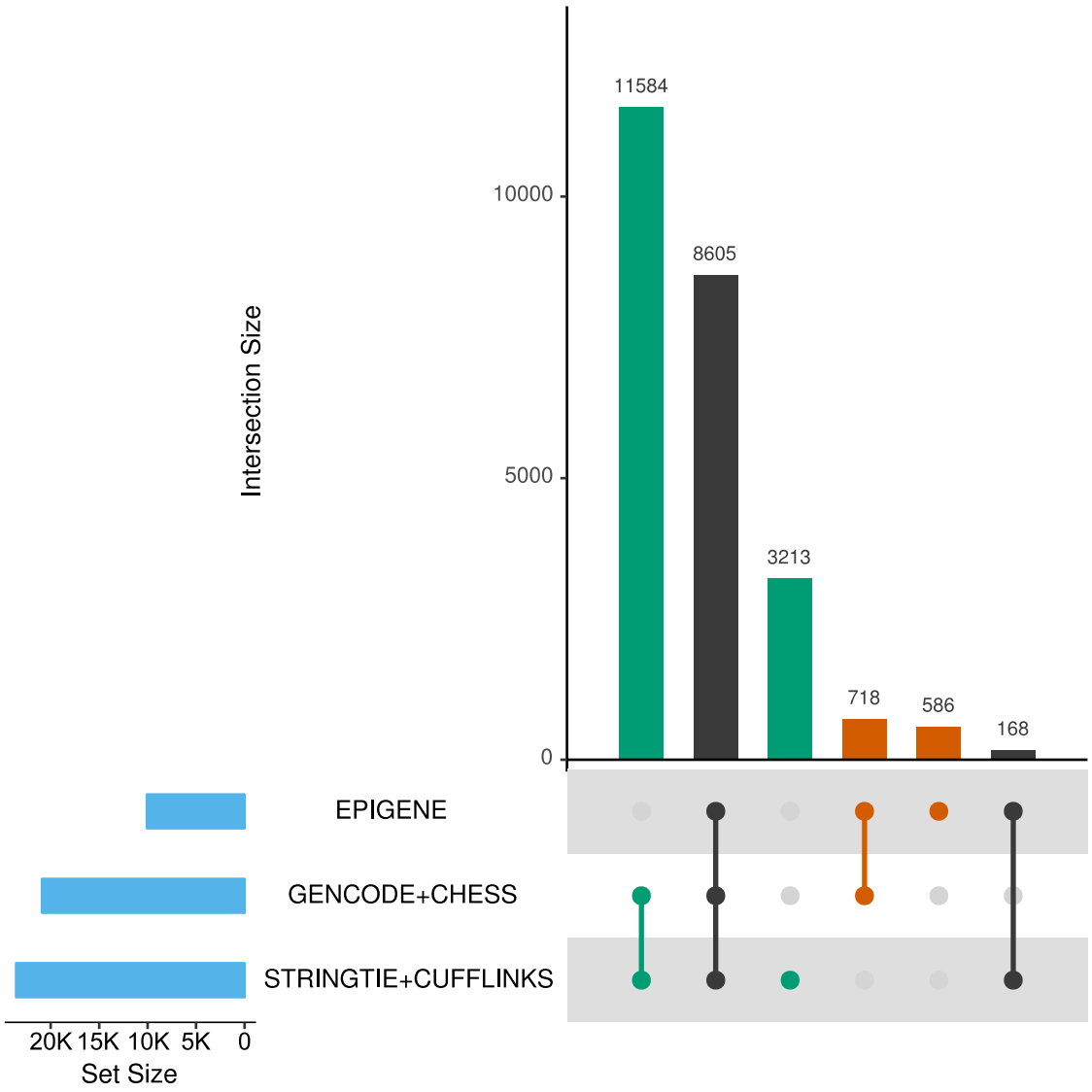


Figure 7: EPIGENE TUs overlapping gene annotations and RNA-seq TUs

I found that 10,077 (8605 + 718 + 586) out of 24,874 TUs in the consensus TU set were predicted by EPIGENE. 92.5% (9,323 out of 10,077) of these EPIGENE TUs overlapped with existing gene annotations irrespective of the TU strand. I identified 1304 (718: annotated, 586: unannotated) EPIGENE-exclusive and 14,797 (11,584: annotated, 3213: unannotated) RNA-seq exclusive TUs.

Further integration of Pol II and nascent RNA-seq (TT-seq and GRO-seq) data revealed that 88.4% (518 out of 586) of EPIGENE unannotated TUs were supported by either nascent RNA-seq or Pol II ChIP-seq evidence and 40% (232 out of 586) of EPIGENE unannotated TUs were supported by both nascent RNA-seq and Pol II ChIP-seq data (please refer Sahu *et al.*, 2020, Supplementary file S2).

2.2.2 Validation with Pol II and histone modifications

As mentioned in section 1.2, transcription in eukaryotes is regulated by Pol II phosphorylation in CTD at serine 2,5 and 7. The promoter regions are characterized by a strong phosphorylation signal for serine 5 and 7, whereas the transcription elongation regions show a strong phosphorylation signal for serine 2 and 5. Here, I used genome-wide ChIP-seq profiles of Pol II, that were generated using four antibodies (PolII8WG16, PolIIS2PH5, PolIIS5P4H8, and PolIIS7P4E12), to compute Pol II occupancy at TSS and gene body respectively.

To estimate the correctness of EPIGENE predictions, I computed the enrichment of histone modifications and Pol II in predicted TUs using normR. I found that majority of EPIGENE TUs showed typical TU characteristics with high enrichment of H3K4me3 and H3K27ac in TSS and H3K36me3 in the gene body (**Figure 8A; Figure S1**). A significant proportion of EPIGENE TUs (78%) showed enrichment of Pol II in TSS and gene body (**Figure S1**).

I further integrated RNA-seq data in this analysis. The RNA-seq RPKM values were computed for EPIGENE TUs and the TUs were classified as high RPKM TUs and

low RPKM TUs based on RNA-seq evidence (threshold = upper quartile of RPKM distribution). I visualized the distribution of Pol II enrichment score for both these TU classes (**Figure 8B**) and found the presence of 622 EPIGENE TUs that are enriched for Pol II but had very low or no RNA-seq evidence. Additional overlap with GENCODE and CHESS gene annotations revealed that 3.8% (24 of 622) of these TUs are unannotated and 96% (598 of 622) of these TUs are annotated.

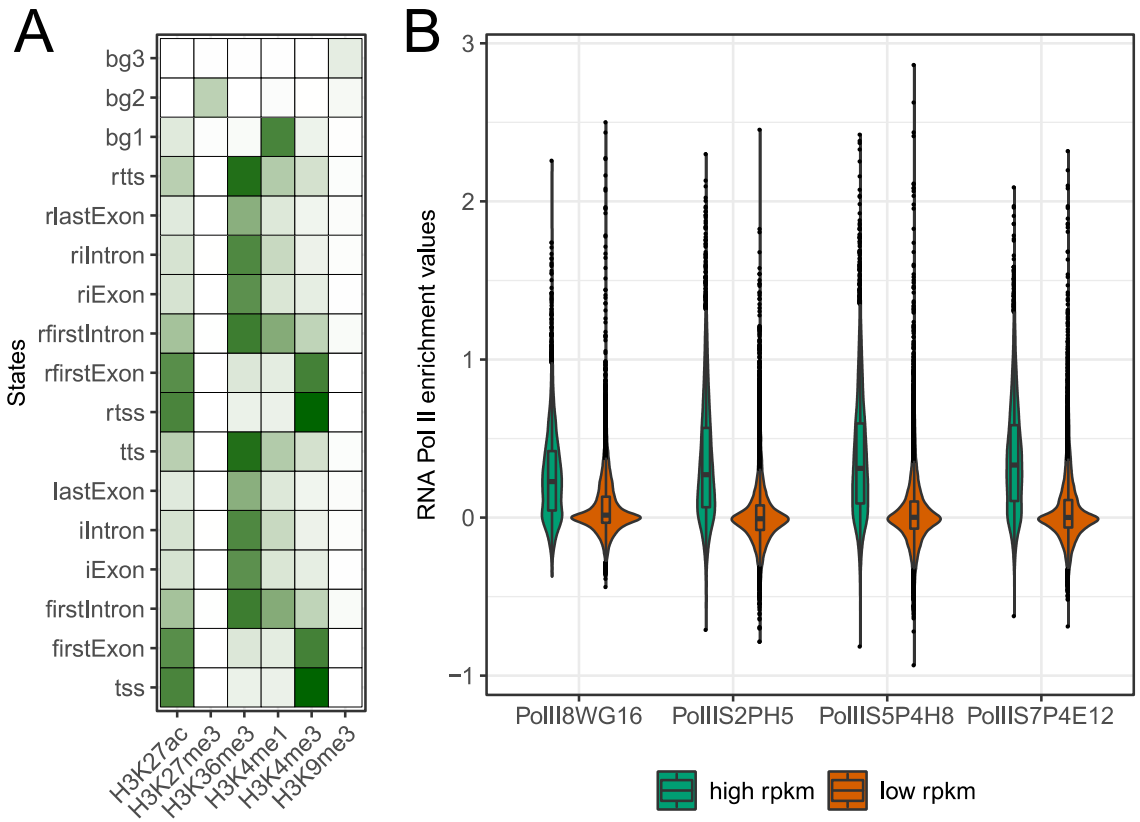


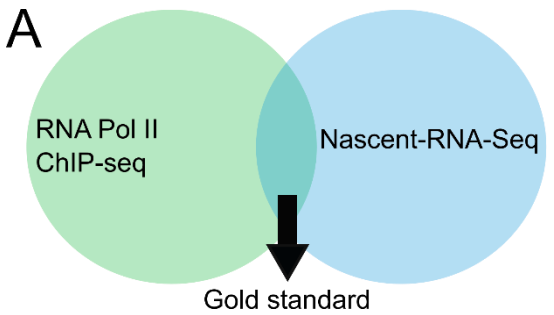
Figure 8: Correctness of EPIGENE TUs (a) EPIGENE-estimated parameters for K562 using 17 chromatin states, ranging from 0 (white) to 1 (dark green). (b) Distribution of Pol II enrichment score in EPIGENE predictions. The EPIGENE predictions are classified as: high RPKM (RPKM ≥ upper quartile) and low RPKM (RPKM < upper quartile) based on RNA-seq evidence in predicted transcripts

2.3 Method comparison

Currently, there is no gold standard set for TUs across cell types and tissues. However, as mentioned in section 1.4, several experimental approaches exist to study Pol II-mediated transcription. Hence, to perform a comprehensive and fair comparison, for each cell line, a cell-specific gold standard TU set was defined based on the enrichment profiles of Pol II ChIP-seq and nascent RNA-seq (**Figure 9A**). These gold standard cell-specific TU sets were then used to quantitatively compare EPIGENE with existing chromatin segmentation and RNA-seq based TU prediction methods (analysis details in section 3.8). The method comparison was performed in two stages: within cell line and cross cell line comparison using cell-specific gold standard TU set as a performance indicator (**Figure 9B**).

2.3.1 Comparison with RNA-seq based approaches

EPIGENE was compared with two widely used RNA-seq based methods StringTie and Cufflinks, across multiple human cell lines. The choice of RNA-seq based methods was due to the superior performance of genome-guided assemblers and also due to the availability of high-quality reference genome in humans (Liu *et al.*, 2016; Venturini *et al.*, 2018).



B

		Actual class: Gold-standard	Actual class: Not Gold-standard
Predicted class	Transcript	TP	FP
	Not transcript	FN	TN

Figure 9: Defining the gold standard for method comparison (a) Set of gold standard regions obtained by combining Pol II ChIP-seq and nascent RNA-seq. (b) Contingency matrix used for method comparison.

Within cell line comparison

This comparison was done on the K562 cell line, the ChIP-seq profile of Pol II in the K562, and the nascent RNA TUs supported by TT-seq and GRO-seq evidence in Schwalb et al (Schwalb *et al.*, 2016) were used as a performance indicator. For this comparison, I used the ChIP-seq profiles of Pol II obtained using the PolIIS5P4H8 antibody because it identifies Pol II occupancy both at TSS and gene body.

I performed the method comparison at 200 bp resolution and found that EPIGENE reports higher AUC (PRC: 0.83, ROC: 0.85; Figure 10A, 10B) than StringTie (PRC: 0.77, ROC: 0.82) and Cufflinks (PRC: 0.60, ROC: 0.63) in both Precision-Recall (PRC) and Receiver-Operating Characteristic (ROC) curve.

This analysis was repeated for 3 different resolutions (50 bp, 100 bp, and 500 bp) and EPIGENE consistently reported a superior performance than StringTie and Cufflinks for varying resolutions (**Figure 10C**). Overall, Cufflinks reported a lower AUC than EPIGENE and StringTie, which is likely due to the usage of RABT assembler resulting in a large number of false positives (Janes *et al.*, 2015).

StringTie also reported a lower AUC than EPIGENE for varying resolutions. Further examining the sensitivity, precision, and specificity values for EPIGENE, StringTie and Cufflinks revealed that lower AUC for RNA-seq based methods was due to spurious read mappings of RNA-seq that results in higher false positives in StringTie and Cufflinks. **Figure S2** shows an example of StringTie and Cufflinks TU that was identified due to read mapping. The TU exactly overlaps with a repetitive sequence occurring in chromosomes 1, 5, 6, X.

Cross cell line comparison

In this comparison, I used datasets from 2 cell lines provided by ENCODE (Feingold *et al.*, 2004) and DEEP (DEEP, 2012) consortium:

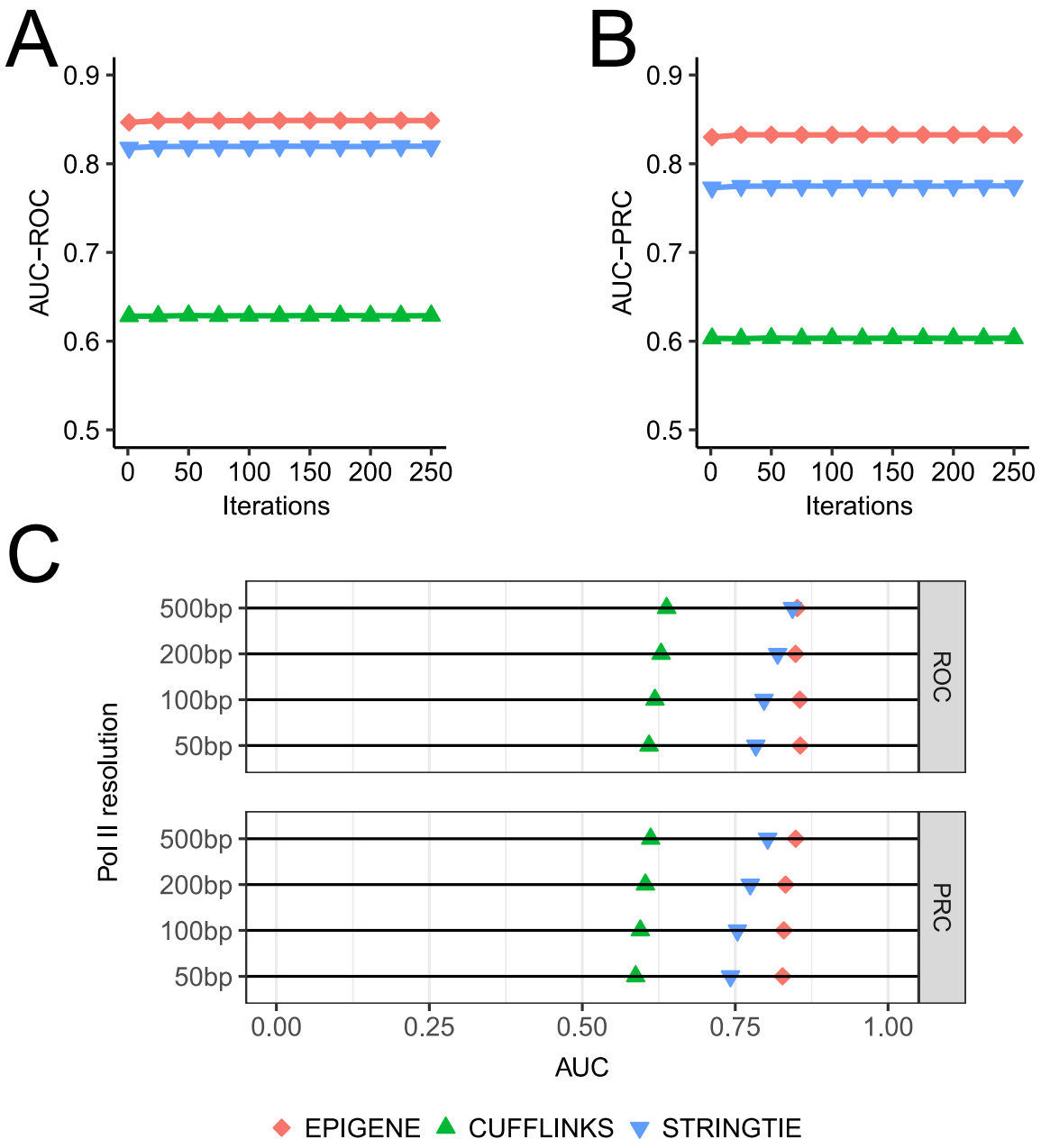


Figure 10: Comparing EPIGENE, STRINGTIE, and CUFFLINKS in K562 (A) Receiver-operating characteristic curve. (B) Precision–recall curve. (C) Area under ROC and PRC curve for varying Pol II resolution for EPIGENE, Cufflinks and StringTie

- IMR90: ChIP-seq profiles of core histone modifications in lung fibroblast cells obtained from Lister et al (Lister *et al.*, 2009), ChIP-seq profile of Pol II obtained from Dunham et al (Dunham *et al.*, 2012), two control experiments (one each for Pol II (Dunham *et al.*, 2012) and histone modifications (Lister *et al.*, 2009)), RNA-seq profile obtained from Dunham et al (Dunham *et al.*, 2012) and GRO-seq profile obtained from Jin *et al* (Jin *et al.*, 2013).
- HepG2, 2 replicates: ChIP-seq profiles of core histone modifications and control in hepatocellular carcinoma and matched RNA-seq profiles obtained from Salhab *et al* (Salhab *et al.*, 2018), where two replicates were available for each histone modification and RNA-seq. ChIP-seq profiles of Pol II and control were obtained from Dunham *et al* (Dunham *et al.*, 2012) and the GRO-seq profile was obtained from Bouvy-Liivrard *et al* (Bouvy-Liivrand *et al.*, 2017).

I applied the K562-trained EPIGENE model to predict active TUs in HepG2 and IMR90 and compared its predictions with that of StringTie and Cufflinks. The GRO-seq profiles and ChIP-seq profiles of Pol II were used to define the cell-specific gold standard TU set in HepG2 and IMR90. The method comparison was performed with a similar strategy used in the previous section.

I found that the K562-trained EPIGENE model consistently reported higher AUC than StringTie and Cufflinks in both PRC and ROC curves (**Figure 11, Figure S3, and Table 4-6**), hence suggesting that EPIGENE predicts TUs with superior precision than RNA-seq based methods across multiple cell lines.

Table 4: AUC-ROC and AUC-PRC values for EPIGENE, CUFFLINKS, and STRINGTIE in IMR90

Method	AUC-ROC	AUC-PRC
EPIGENE	0.77	0.78
STRINGTIE	0.72	0.68
CUFFLINKS	0.54	0.54

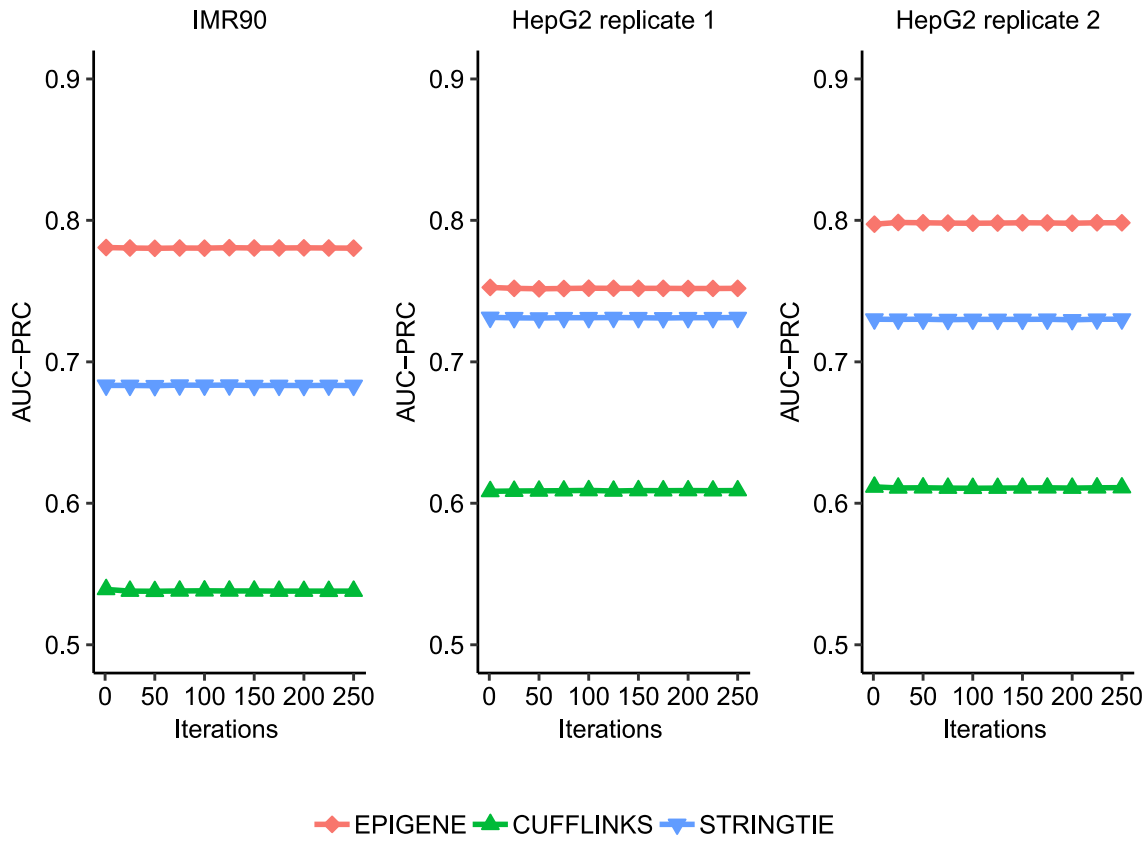


Figure 11: Comparing K562-trained EPIGENE models, STRNGTIE and CUFFLINKS across cell lines EPIGENE reports a superior AUC compared to STRINGTIE and CUFFLINKS.

Table 5: AUC-ROC and AUC-PRC values for EPIGENE, CUFFLINKS, and STRINGTIE in HepG2 replicate 1

Method	AUC-ROC	AUC-PRC
EPIGENE	0.77	0.75
STRINGTIE	0.77	0.73
CUFFLINKS	0.64	0.61

Table 6: AUC-ROC and AUC-PRC values for EPIGENE, CUFFLINKS, and STRINGTIE in HepG2 replicate 2

Method	AUC-ROC	AUC-PRC
EPIGENE	0.80	0.80
STRINGTIE	0.78	0.73
CUFFLINKS	0.64	0.61

2.3.2 Comparison with chromatin segmentation methods

As mentioned in section 1.5.2, currently several chromatin segmentation approaches exist that provide genome-wide chromatin state annotation using histone modifications. It is important to note that these approaches were initially developed for chromatin state annotation and therefore, the model parameters do not represent the topology of a TU. Here, I evaluate the performance of these approaches in identifying genome-wide active TUs. I compared EPIGENE predictions with the predictions of a widely used chromatin segmentation approach, ChromHMM which also uses a binning strategy. Segway was not included in this comparison because it is comparatively slower. Additionally, it operates at a 1-bp resolution and, hence, restricts a fair comparison across multiple cell types and tissues.

It is important to note that chromatin annotations obtained from ChromHMM do not contain the strand information, hence, to perform a fair comparison, ChromHMM was evaluated for TU prediction in a strand-specific and unstranded manner. Chromatin state annotations across multiple cell lines were obtained with ChromHMM using the same set of histone modifications that were used as features in the EPIGENE model. These chromatin state annotations were then filtered to obtain strand-specific and unstranded TUs. Strand-specific TUs were obtained by linking active TSS (state 9 in Figure 6B) and transcription elongation states (states 4, 5, and 8 in **Figure 12A**). In this case, an active TU

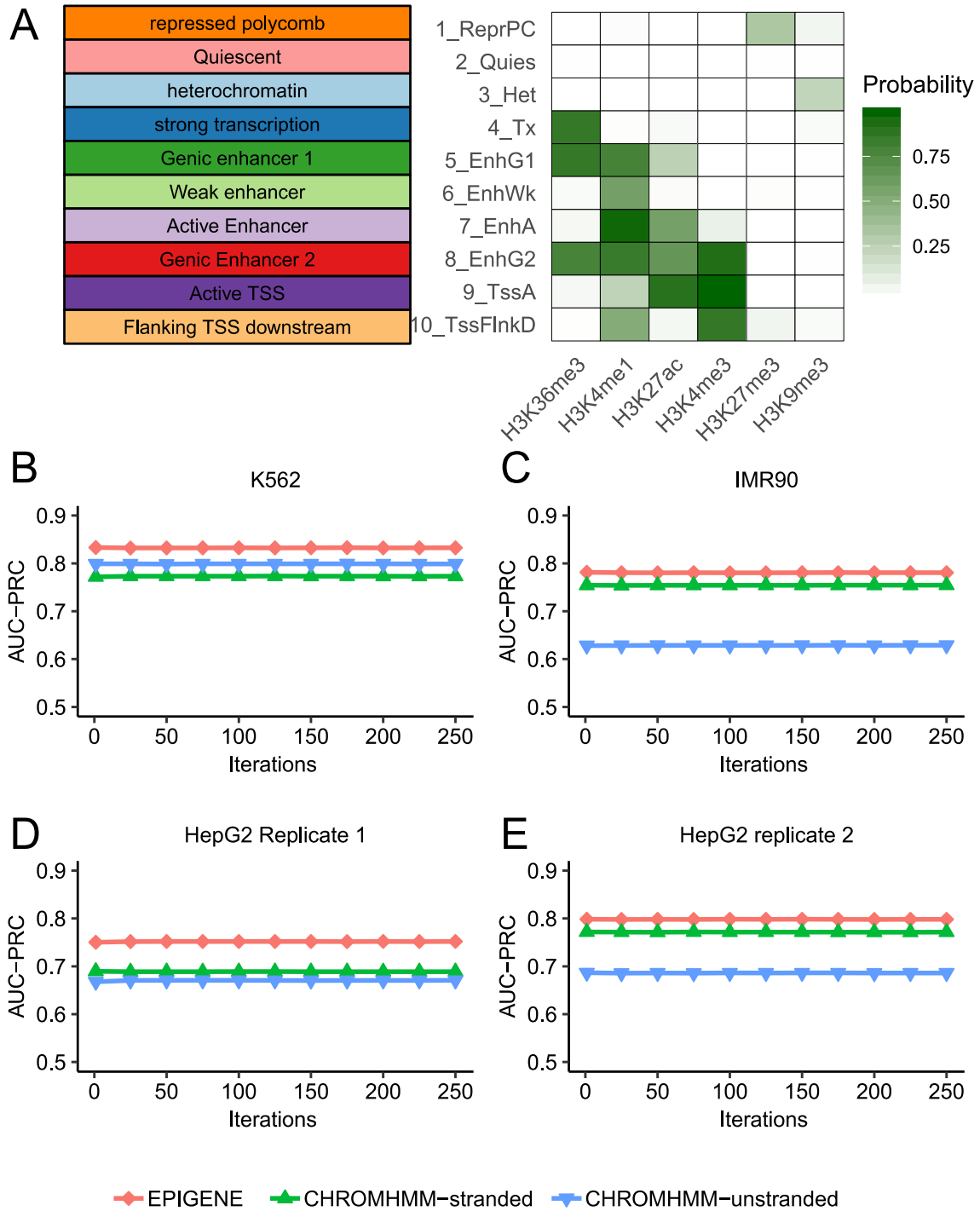


Figure 12: Comparing K562-trained EPIGENE with ChromHMM across cell lines (A) Emission probabilities of ChromHMM model trained in K562 cell line. (B-E) Performance of K562-trained EPIGENE model and K562-trained ChromHMM model in K562, IMR90 and HepG2

is defined as a genomic region that begins with an active TSS state and proceeds through transcription elongation states in forward or reverse direction. An active TSS state was defined by an enrichment of H3K27ac and H3K4me3 and a transcription elongation state is defined by an enrichment of H3K36me3. For unstranded TUs, it was assumed that a TU is a genomic region that is enriched for H3K36me3, hence, unstranded TUs were obtained by filtering the ChromHMM chromatin state annotations for transcription elongation states (states 4, 5 and 8 in **Figure 12A**). The performance of EPIGENE was compared with ChromHMM using the cell-specific gold standard regions defined in section 2.3.1. As evident from **Figure 12B-E, Figure S4 and Table 7-10** EPIGENE consistently report a superior performance than ChromHMM strand-specific and ChromHMM unstranded TUs.

Table 7: AUC-ROC and AUC-PRC values for EPIGENE, chromHMM strand-specific, and chromHMM unstranded in K562

Method	AUC-ROC	AUC-PRC
EPIGENE	0.85	0.83
ChromHMM stranded	0.73	0.77
ChromHMM unstranded	0.79	0.80

Table 8: AUC-ROC and AUC-PRC values for EPIGENE, chromHMM strand-specific, and chromHMM unstranded in IMR90

Method	AUC-ROC	AUC-PRC
EPIGENE	0.77	0.78
ChromHMM stranded	0.69	0.75
ChromHMM unstranded	0.60	0.63

Table 9: AUC-ROC and AUC-PRC values for EPIGENE, chromHMM strand-specific, and chromHMM unstranded in HepG2 replicate 1

Method	AUC-ROC	AUC-PRC
EPIGENE	0.75	0.77
ChromHMM stranded	0.63	0.69
ChromHMM unstranded	0.67	0.67

Table 10: AUC-ROC and AUC-PRC values for EPIGENE, chromHMM strand-specific, and chromHMM unstranded in Hepg2 replicate 2

Method	AUC-ROC	AUC-PRC
EPIGENE	0.80	0.80
ChromHMM stranded	0.71	0.77
ChromHMM unstranded	0.66	0.69

The lower AUC of unstranded and strand-specific ChromHMM TUs was due to the presence of intermediate low coverage states (state 2 in **Figure 12A**) and intronic enhancers that resulted in shorter strand-specific and unstranded ChromHMM TUs (**Figure 13**) and fewer strand-specific ChromHMM TUs.

2.4 EPIGENE TUs with negligible RNA-seq evidence

Previous validations and comparisons of EPIGENE with existing gene annotations, Pol II ChIP-seq profiles, and existing RNA-seq approaches (section 2.2 and 2.3) revealed the presence of EPIGENE TUs that were supported by nascent RNA and Pol II evidence but with negligible RNA-seq evidence. In this section, I analyze these TUs across multiple cell lines by: (a) identifying cell-specific TUs that showed TU characteristics but lacked RNA-seq evidence, and (b) investigating the presence of microRNA precursors that were not identified by RNA-seq.

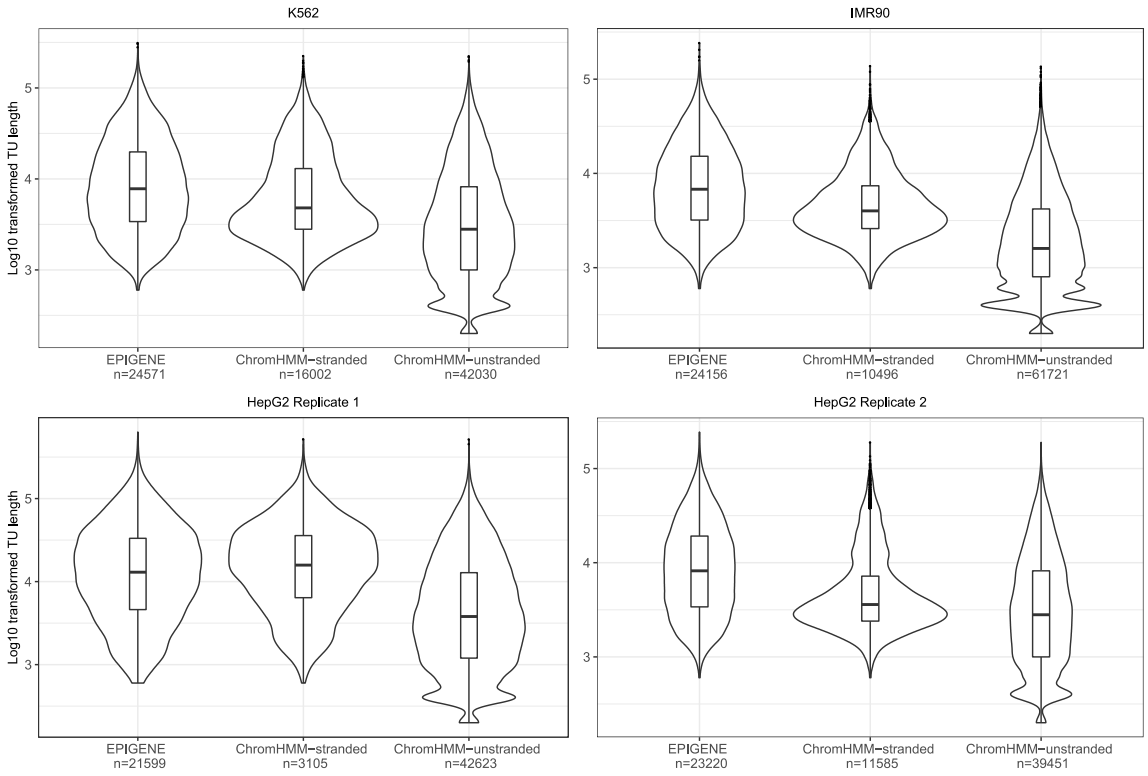


Figure 13: Length distribution of EPIGENE and ChromHMM TUs across cell lines

2.4.1 EPIGENE predicts cell-specific TUs

EPIGENE predicted TUs in K562, IMR90, and HepG2 cell lines were combined using *union* operation to create a consensus EPIGENE TU set (refer section 3.9 for analysis details). This consensus TU set constituted 18,248 TUs, of which ~78% TUs were enriched for Pol II. I identified 10,233 differentially enriched TUs, of which 8047 TUs were exclusive to cell lines (HepG2: 1255; IMR90: 2545; K562: 4247; **Figure S5**). Additional integration of RNA-seq evidence revealed the presence of 43 highly confident cell-specific TUs that lacked RNA-seq evidence but showed typical TU characteristics, with enrichment of Pol II, GRO-seq, H3K27ac and H3K4me3 at TSS and H3K36me3 and Pol II in the gene body. **Figure 14** shows one such K562 exclusive EPIGENE TU located between lncRNA RP5-

952N6.1 and CASP3P1. This TU was additionally supported by Pol II and nascent RNA evidence but lacks RNA-seq evidence.

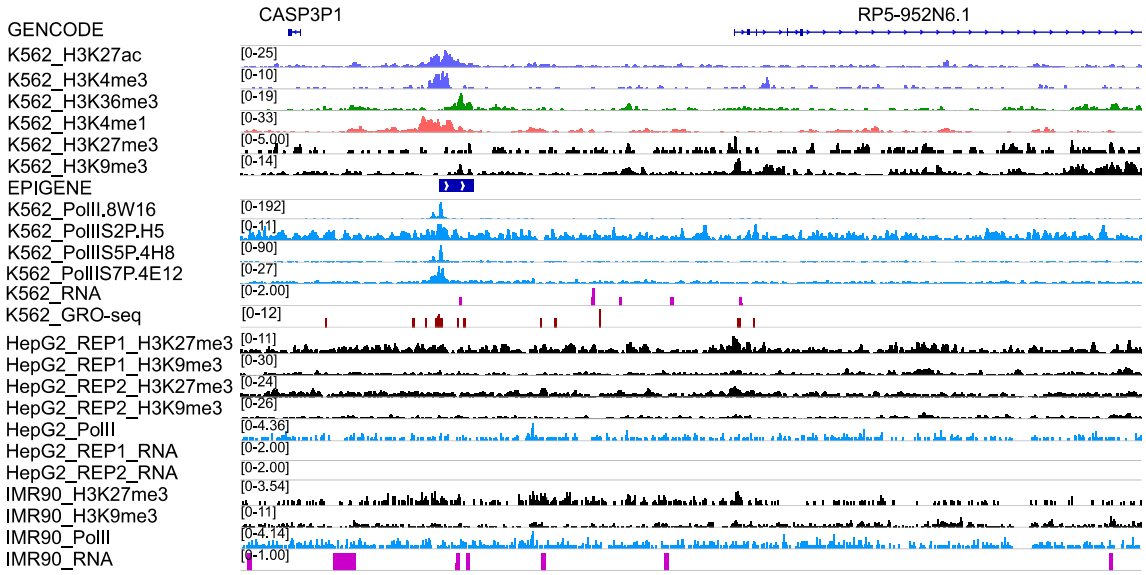


Figure 14: Example of EPIGENE-predicted TU that lacks RNA-seq evidence. The TU was predicted to be active in K562 but not in HepG2 and IMR90, and is located between pseudogene CASP3P1 and lncRNA RP5-952N6.1. The TU (shown in dark blue in EPIGENE-K562 track) shows an enrichment of H3K27ac and H3K4me3 at TSS (tracks shown in light violet), H3K36me3 in gene body (tracks shown in green), enhancer mark H3K4me1 few bps upstream of TSS (tracks shown in pink), GRO-seq in TSS (tracks shown in brown), K562 Pol II in TSS and gene body (tracks shown in blue). The TU also shows an absence of repression marks H3K27me3 and H3K9me3 in K562 (tracks shown in black). We additionally observe the enrichment of repression mark in H3K27me3 in HepG2 and IMR90 indicating that the region is repressed in both these cell lines. There is a negligible RNA-seq evidence (shown in dark pink in K562-RNA-seq track) for this predicted TU.

2.4.2 EPIGENE predicts microRNAs precursors

MicroRNA precursors are a class of TUs that gives rise to ~70 bp pre miRNA which is further processed by endonucleases (Bartel, 2004; He and Hannon, 2004) to small, evolutionally conserved, non-coding and mature microRNAs (Lagos-Quintana *et al.*, 2001; Lee and Ambros, 2001).

These RNAs have been shown to regulate several fundamental biological processes such as differentiation, development, and apoptosis by post-translational regulation of target genes via gene silencing (Plasterk, 2006; Carleton, Cleary and Linsley, 2007) and are involved in disease pathogenesis (Calin and Croce, 2006). Due to the short lifetime of microRNA precursors, identifying them with existing RNA-seq based TU detection approaches becomes challenging.

I created a consensus TU set for individual cell lines (K562, IMR90, and HepG2) by combining the predictions of EPIGENE, StringTie, and Cufflinks. I integrated miRbase annotation (Griffiths-Jones *et al.*, 2006) to the consensus TU set and found that 655 EPIGENE TUs in the HepG2 cell line is supported by microRNA annotations.

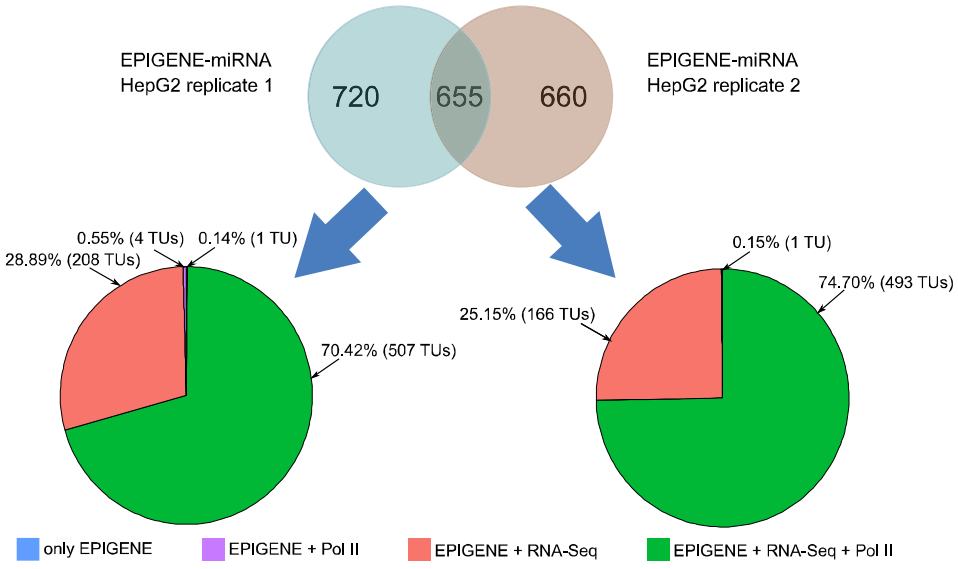


Figure 15: EPIGENE TUs overlapping miRbase annotations. Overview of potential primary miRNAs predicted by EPIGENE in HepG2

These potential microRNA TUs constitutes 5% of total EPIGENE TUs that are common in both HepG2 replicates. Further integration of Pol II and RNA-seq data, revealed that majority of these TUs were supported by Pol II and RNA-seq evidence (**Figure 15 and S6**). Additionally, I found 2 microRNA TUs in HepG2, that were enriched

for H3K4me3, H3K27ac, GRO-seq and Pol II at TSS, and H3K36me3 in their gene body but lacked RNA-seq evidence. **Figure 16** shows an example of a potential microRNA TU in HepG2, which overlaps with a microRNA cluster, located between lncRNA RP11-738B7.1 and NRF1 gene. This microRNA cluster has been shown to arise from the same microRNA precursor and is associated with cell proliferation in the HepG2 cell line.

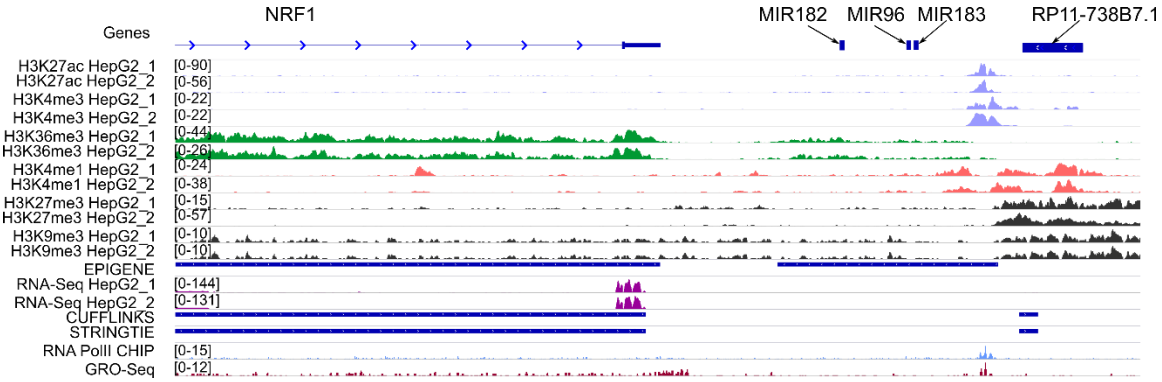


Figure 16: EPIGENE-predicted TU overlapping a microRNA cluster in HepG2 cell line. This region is located between lincRNA RP11-738B7.1 and gene NRF1. The TU shows an enrichment of H3K27ac and H3K4me3 at TSS (tracks shown in light violet), H3K36me3 in gene body (tracks shown in green), enhancer mark H3K4me1 few bps upstream and downstream of TSS (tracks shown in pink), GRO-seq in TSS (tracks shown in brown) and RNA Polymerase II ChIP-seq in TSS (tracks shown in blue). The predictions also show an absence of repression marks H3K27me3 and H3K9me3 (tracks shown in black) and RNA-seq evidence (tracks shown in dark pink)

3 Methods

Parts of this chapter were published as a peer-reviewed research article in Epigenetics and Chromatin (Sahu et al. 2020). For detailed author contributions, please refer to Page xi.

In this chapter, I provide additional details about the steps and methods used in data pre-processing and analyzing EPIGENE predictions. Additionally, I also describe the strategies used for training and evaluating EPIGENE.

3.1 Data pre-processing

As mentioned in section 2.1, EPIGENE was trained on the reference epigenome. The reference epigenome of K562 cell line was generated as per IHEC standards. The ChIP-seq libraries for IHEC class 1 histone modifications and Pol II were prepared by my colleagues (mentioned in Author Contributions in Page xi) as per the instructions mentioned in my publication (Sahu *et al.*, 2020). Other datasets used in performance evaluation of EPIGENE were download from GEO (Clough and Barrett, 2016), ENCODE, European Genome-phenome Archive (EGA), and European Nucleotide Archive (ENA) (Feingold *et al.*, 2004; Leinonen *et al.*, 2011; Lappalainen *et al.*, 2015; Clough and Barrett, 2016) (refer **Table S1**).

3.1.1 Sequencing and processing ChIP-seq data

The ChIP-seq libraries for histone modifications and Pol II were sequenced on Illumina Highseq 2500 resulting in 50 bp reads. The reads were aligned to genome assembly "hs37d5" using STAR aligner (Dobin *et al.*, 2013) with the parameter setting: *intronMax = 1* and the duplicate reads were located and marked using Picard tools (Wysoker, A., Tibbetts, K., and Fennell, 2013). The quality of ChIP-seq data was then evaluated using the *plotFingerprint* method of deepTools (Ramírez *et al.*, 2014).

3.1.2 Processing of RNA-seq data

The reads from RNA-seq experiments were download from ENA (SRR 315336, 315337 for K562), EGA (EGAD00001002527 for HepG2), and ENCODE (ENCSR00CTQ for IMR90). The quality of reads was first evaluated with FastQC (Andrews, 2010) and the reads were then aligned to genome assembly ”hs37d5” with STAR aligner.

3.1.3 Processing of Nascent RNA-seq data

The genome-wide TU set for K562 that was identified based on TT-seq was downloaded from GEO (GSE 75792). As these TUs were identified using hg38, the genome-coordinates of TUs were lifted over to hg19 for valid and efficient comparison. The raw reads from the GRO-seq experiment for HepG2 were downloaded from GEO (GSM2428726). These raw reads were aligned to hg19 and the processing was done based on the steps described in Bouvy-Liivrard *et al*. For IMR90, I used the TU annotation described in Jin *et al* which was generated using GRO-seq. This annotation was obtained using hg18 genome build, hence, to perform a fair and valid comparison, the TUs were lifted over to hg19.

3.2 Binarization of ChIP-seq profiles

As mentioned in section 2.1.1, EPIGENE requires the enrichment scores of IHEC class 1 histone modifications in binarized form i.e as presence or absence calls referred to as “class matrix”. This was done by partitioning the mappable regions of the genome to contiguous non-overlapping genomic intervals of the same size called bins. Currently, EPIGENE performs the binarized (presence or absence call) enrichment calling at 200 bp resolution as this roughly corresponds to the size of a nucleosome and spacer regions.

3.2.1 Obtaining read counts

The read counts for all genomic bins were computed from ChIP and input alignment files using the *bamCount* function of the *bamsignals* R package (Mammana and Helmuth, 2015), with the following parameter settings: paired.end = "midpoint", mapqual = 255, and filteredFlag = 1024. mapqual was set to 30, for publicly available datasets that were aligned using bwa or bowtie (Langmead *et al.*, 2009; Li and Durbin, 2009).

3.2.2 Binarized enrichment calling

After obtaining the ChIP and input read counts for individual histone modifications, in each 200 bp bin, the binarized enrichment score (1: present or 0: absent) of histone modifications $E(bin, HM_i)$ (where $i \in$ IHEC class I histone modifications) and Pol II $E(bin, Pol\ II_j)$ (where j ∈ (Pol II.8WG16, Pol IIS2P.H5, Pol IIS5.4H8, Pol IIS7P.4E12)) across all bins, were computed using *enrichR* and *getClasses* functions from normR package. For accurate binarized enrichment calling, I prefilter the genomic bins such that bins with input and ChIP read counts > 0 should be considered for background estimation. This was done by setting, binFilter = "zero" in *enrichR*. The binary enrichment scores for individual bins were computed using *getClasses*. This step results in the "class matrix" which was used as an input to the multivariate HMM.

3.3 The EPIGENE model

EPIGENE models the class matrix of histone modifications using a semi-supervised multivariate HMM to predict genome-wide active TUs. The class matrix C is a $A \times B$ matrix, where $A =$ number of 200 bp bins, and $B =$ number of histone modifications (in this case, $B = 6$) and each matrix entry C_{ij} represent the presence or absence of the jth histone modification in the ith bin. As mentioned earlier in section 2.1.2, the model consists of 17 hidden states (14 TU states and 3 background states) and each row of the class matrix corresponds to one of these hidden states. The transition probabilities between the hidden states capture the position biases of TU states relative to each other. The emission probabilities of each hidden state represent the probability with which each

histone modification occurs in the hidden state. The transition and emission probabilities were trained in a semi-supervised manner which is discussed in the forthcoming section. Given this multivariate HMM and the model parameters, the algorithm:

1. assigns the initial probabilities
2. fits the transition and emission probabilities using the Baum-Welch algorithm (Baum *et al.*, 1970).
3. infers the final segmentation. The final segmentation i.e. the sequence of hidden states can be inferred using posterior decoding or Viterbi algorithm. As I was concerned about the most probable sequence of active TUs rather than the most probable hidden state for each bin, the final segmentation was obtained using the Viterbi algorithm.
4. filters the output vector i.e. the sequence of hidden states. The vector of hidden states was filtered to obtain genomic regions beginning with a TSS state and terminating with a TTS state.

3.4 Training the transition and emission probabilities

The transition and emission probabilities of the multivariate HMM were trained in a semi-supervised manner using GENCODE annotations in the following steps:

1. The genomic bins obtained from section 3.2 were overlapped with GENCODE transcripts to identify "gencode bins".
2. For individual transcript IDs, the gencode bins were classified as TSS, exon, intron, and TTS bins based on their overlap with gencode transcript components. The rank of each exon and intron bins was obtained from GENCODE.
3. For each gencode transcript, the coverage (in bp) of individual TU elements (e.g. TSS, 1st exon, 1st intron, 2nd exon, 2nd intron, etc) was computed i.e for each 200 bp bin B, the number of bps overlapping with individual TU components was computed. Finally, a coverage

table $T_{m\times n}$ was created, where, *m = number of bins overlapping the transcrip*t and *n = number of hidden states* (in this case n=number of TU states). Each value of the coverage matrix represents the coverage (in bps) of the 200bp bin for the TU hidden state. The coverage tables for individual transcript IDs were combined to a coverage list, where each entry in the list contains the coverage table for a transcript ID.

4. The transition probabilities between TU states were computed from the coverage list generated in Step 3. The missing transition probabilities from and to the background states and also between the background states were generated in an unsupervised manner.
5. The coverage list and class matrix were filtered to identify transcripts and bins that were enriched for Pol II. This was done by performing a k-means clustering for all the TSS and TTS bins of the class matrix and identifying the bin cluster that reported a high cluster mean for Pol II. The emission probabilities for each TU state was computed from the filtered class matrix and coverage list. The emission probabilities of background states were trained in an unsupervised manner.

3.5 Binarization of Nascent RNA-seq profiles

The nascent RNA TUs for K562 were obtained from Schwalb *et al* (Schwalb *et al*., 2016), while, the TUs for HepG2 and IMR90 were obtained from GRO-seq profiles using groHMM (Chae, Danko and Kraus, 2015). The TUs for HepG2 were obtained using groHMM with default parameters, whereas, for IMR90, the TUs were obtained with parameter values specified in Chae *et al* (Chae, Danko and Kraus, 2015). Hence, for a cell line C, the binary enrichment i.e. presence or absence of nascent RNA-seq signal across the 200 bp bins $E_C(bin, nascent\ RNA)$ is given by:

$$E_C(bin, NascentRNA) = \begin{cases} 1 & if O(bin, TU_{nascent}) \geq 1 \\ 0 & otherwise \end{cases}$$

where $O(bin, TU_{nascent})$ represents the degree of overlap between the 200 bp bin and the nascent RNA TUs in cell line C.

3.6 Binarization of RNA-seq profiles

The aligned reads from RNA-seq experiments were assembled using genome-guided assemblers StringTie and Cufflinks to identify genome-wide TUs. For both StringTie and Cufflinks, the genome-wide TUs were obtained from aligned reads using GRCh37 gencode gene annotation (Harrow *et al.*, 2012) with the parameter setting –G and --rf (for StringTie) and –g and --rf (for Cufflinks). The binary enrichment of RNA-seq signal in cell line C across the 200 bp bins $E_C(bin, RNA)$ is given by:

$$E_C(bin, RNA) = \begin{cases} 1 \; if \; O(bin, TU_{RNA}) \geq 1 \\ 0 \qquad otherwise \end{cases}$$

where $O(bin, TU_{RNA})$ represents the degree of overlap between the 200 bp bin and the RNA-seq TUs in cell line C that was obtained from StringTie and Cufflinks.

3.7 Validation with gene annotations and RNA-seq

To efficiently validate the fraction of EPIGENE predictions explained by existing annotations and RNA-seq approaches, a consensus TU set was created by combining the gene annotations (obtained from GENCODE and CHESS) with the predictions of EPIGENE, StringTie and, Cufflinks in K562 cell line. This was done using the *union* method of the GenomicRanges package (Lawrence *et al.*, 2013). After obtaining the consensus TU set, a validation matrix was created. The validation matrix V is a $m \times n$, where, m corresponds to the total number of genomic regions of the consensus TU set, and n corresponds to the types of datasets combined to form the consensus TU set. In this case, $n = 3$, as we combined three different kinds of datasets namely, EPIGENE predictions, gene annotations (GENCODE+CHESS) and, predictions from RNA-seq based approaches (StringTie + Cufflinks). Each entry of the validation matrix represents if the genomic region of the consensus TU set overlaps with the corresponding predictions or

annotations. The validation matrix was then used to calculate the summary statistics shown in **Figure 7**, and the results were plotted using Upset (Lex *et al.*, 2014).

3.8 Performance evaluation

In section 2.3, I compared the performance of EPIGENE with existing RNA-seq and chromatin segmentation approaches. The performance of all of these methods was evaluated at 200 bp resolution using Pol II and nascent RNA-seq signal as a performance indicator. The actual transcription status of each bin $T_{Actual}(bin)$ was given by:

$$T_{Actual}(bin) = \begin{cases} 1 \; if \; E_C(bin, Pol\ II) \; \wedge \; E_C(bin, nascent\ RNA) = 1 \\ 0 \qquad\qquad otherwise \end{cases}$$

where, $E_C(bin, Pol\ II)$ represents the presence or absence of Pol II ChIP-seq signal that was obtained in section 3.2.2 and $E_C(bin, nascent\ RNA)$ represents the binary enrichment of the nascent RNA-seq signal in the bin that was obtained in section 3.5.

The predicted transcription status of the bin for method m, $T_{predicted}(bin)$ was given by:

$$T_{predicted}(bin) = \begin{cases} 1 \; if \; O(bin, TU_m) \geq 1 \\ 0 \qquad otherwise \end{cases}$$

where, $O(bin, TU_m)$ represents the degree of overlap between the 200 bp bin and the TUs predicted by method m.

The predictions of EPIGENE, RNA-seq, and chromatin segmentation approaches were evaluated by computing the AUC for ROC and PRC curve. Due to a very high class imbalance i.e. $bin_{Pol\ II^+} \ll bin_{Pol\ II^-}$, the AUC-ROC and AUC-PRC were computed by random sampling as:

$$AUC = mean(L_{AUC}) - \left(\frac{stdDev(L_{AUC})}{\sqrt{n_i}}\right)$$

where, n_i is the number of iterations and L_{AUC} is the list of AUCs obtained for n_i iterations.

3.9 Identifying cell-specific TUs with negligible RNA-seq evidence

For identifying cell-specific EPIGENE TUs, a consensus TU set was created by combining the predictions of K562, HepG2, and IMR90 using the *union* method from the GenomicRanges package. This consensus set was divided into individual subsets based on the presence of a TU for a given cell line (**Figure S5**). The Pol II enrichment for each subset of TUs was obtained using the same strategy described in section 3.2. The RNA-seq evidence for a TU is given by a binary value representing the overlap of the TU with StringTie or Cufflinks predictions. Lastly, each of the subsets was filtered for RNA-seq evidence to obtain valid cell-specific TU with negligible RNA-seq evidence.

4 Discussion and conclusion

Parts of this chapter were published as a peer-reviewed research article in Epigenetics and Chromatin (Sahu et al. 2020). For detailed author contributions, please refer to Page xi.

In this chapter, I critically discuss the existing TU prediction approaches and their limitations and the need for a chromatin-based TU prediction method. Additionally, I discuss the benefits and shortcomings of our novel chromatin-based TU prediction method and also suggest possible solutions to resolve the shortcomings.

4.1 Genome-wide TU identification

Accurate identification of genome-wide TU is essential for understanding the transcriptomic landscape of a cell and analyzing its differences across multiple cell types and conditions. Currently, due to efforts of several epigenomic consortia like ENCODE, Roadmap epigenomics, Blueprint, DEEP, CEEHRC, and IHEC the ChIP-seq profiles of transcription-associated histone modifications are now available across many cells types and tissues. This vast amount of data requires novel and efficient methods to analyze and integrate these data. Hence, efficient methods to accomplish this task are crucial for current and future transcriptomic research and can potentially identify genome-wide active TUs.

4.1.1 Modifying the strategy for genome-wide TU identification

Recent advances in transcriptome sequencing have made it possible to identify and quantify genome-wide TUs in a cost-effective manner. As a result, most computational approaches for identifying genome-wide TUs are based on RNA-seq data. Although these approaches correctly identify stable mRNAs, they, however, fail to identify TUs that are transcribed to unstable regulatory RNAs such as microRNAs

precursors. Several recent studies have demonstrated the presence of many unstable TUs that are rapidly degraded (Preker *et al.*, 2008; Tani *et al.*, 2012; Li *et al.*, 2013), some of which have been reported to be associated with diseases (Sethi and Lukiw, 2009; Bail *et al.*, 2010; Shah *et al.*, 2016; Wang, Qin and Tang, 2019; Zhang *et al.*, 2019).

Recently, nascent RNA-seq has also been developed for studying Pol II-mediated transcription. However, majority of these approaches are limited to cell cultures and cannot be implemented *in vivo*. These shortcomings can be alleviated with a chromatin-based TU prediction method due to the availability of a vast amount of ChIP-seq profiles of transcription-associated histone modifications. In the past, chromatin segmentation methods have been developed that identify genomic features like promoters, enhancers, transcribed regions, etc., using histone modifications. These methods, however, do not identify TUs as their model parameters do not capture the topology of a TU.

4.1.2 Predicting TUs with histone modifications

In this thesis, I developed a multivariate HMM called EPIGENE, which predicts genome-wide TUs using histone modifications. EPIGENE consists of two types of hidden states: TU states and background states. The TU states were trained in a supervised manner whereas the background states were trained in an unsupervised manner. This semi-supervised training captures the probability of occurrence of histone modifications in different components of an active TU (such as TSS, exon, intron, TTS, etc.) as well as the topology of the combination of histone modifications in active TUs. Additionally, duplicating the TU state sequence to run from TSS to TTS and vice versa also enables EPIGENE to capture the directionality of active TUs.

4.2 Unbiased and accurate TU prediction by EPIGENE

I validated the EPIGENE predictions with existing gene annotations, ChIP-seq, and nascent RNA-seq profiles and show that majority of EPIGENE predictions can

be explained by GENCODE, CHESS, Pol II, and nascent RNA evidence. Additionally, a quantitative comparison with Pol II ChIP-seq and RNA-seq profiles revealed the presence of active TUs with negligible RNA-seq evidence.

I compared the performance of EPIGENE with existing RNA-seq (StringTie and Cufflinks) and chromatin segmentation (chromHMM) methods by defining a cell-specific gold standard with Pol II and nascent RNA-seq as true transcription indicator. Based on AUC-PRC and AUC-ROC as performance metrics, EPIGENE reports a superior performance compared to RNA-seq and chromatin segmentation approaches. Additional performance evaluation across multiple cell lines showed that EPIGENE consistently achieves superior performance than RNA-seq and chromatin segmentation approaches and, therefore, can be reliably applied across different cell types and tissues without the need to retrain the model parameters.

I examined other performance metrics like specificity, sensitivity, and precision and found that the low AUC of RNA-seq methods was due to RNA-seq mapping artifacts that result in higher false positives. Additionally, the extremely low AUC of Cufflinks is due to the usage of RABT assembler that further increases the number of false positives (Janes *et al.*, 2015). Further evaluation of EPIGENE TUs across K562, IMR90, and HepG2 revealed the presence of cell-specific TUs with negligible RNA-seq evidence. Additionally, EPIGENE also predicts microRNA precursors that lack RNA-seq signal supposedly due to their unstable nature.

4.3 Limitations of EPIGENE

One of the major current shortcomings of EPIGENE is its inability to differentiate between the functional and non-functional elements (exons and introns) of a TU, as the association between alternative splicing and histone modifications is yet to be elucidated.

Additionally, it is important to note that EPIGENE requires all the core histone modifications to efficiently predict genome-wide active TUs and the accuracy of TUs

decreases in the absence of a core histone modification. However, all the histone modifications used as features in EPIGENE are available for many cell types and tissues and are consistently being generated for several cell types by many consortia like ENCODE and Roadmap Epigenome. In the absence of a core histone modification, imputation techniques such as PREDICTED and ChromImpute (Ernst and Kellis, 2015; Durham *et al.*, 2018) can be used to impute the missing histone modification and then use the imputed histone modification along with the available histone modifications to identify genome-wide active TUs.

EPIGENE also does not accurately identify complex TUs such as TUs with multiple active promoters. For such cases, EPIGENE predicts a new TU each time it encounters an active promoter and hence predicting multiple small TUs rather than predicting a single large TU. This problem can be resolved with a multimodal HMM by incorporating nascent RNA-seq data as features, in addition to the existing chromatin state features. Nascent RNA-seq data from TT-seq or mNET-seq can be used for this purpose, as both these techniques are known to reliably detect transcription termination (Wissink *et al.*, 2019).

4.4 Conclusion

With recent advances in ChIP-seq and increasing efforts in the direction of epigenomics, several consortia continue to provide high-quality ChIP-seq profiles of transcription-associated histone modifications. However, determining the transcriptomic landscape and identifying genome-wide TUs with this data remains unexplored. During my Ph.D. research, I address this shortcoming by developing a novel chromatin-based TU prediction method, EPIGENE for predicting genome-wide active TUs. EPIGENE uses a semi-supervised strategy to train the HMM parameters which enables the efficient prediction of genome-wide TUs in both forward and reverse strands. Extensive validations and evaluation of EPIGENE predictions demonstrate its superior performance

over existing RNA-seq and chromatin segmentation approaches and its broader applicability across cell types and tissues.

EPIGENE is user-friendly and can be executed by providing the aligned ChIP-seq and control reads without the need to re-train the model. Its predictions strongly agree with Pol II and nascent RNA-seq data indicating their superior accuracy. Taken together, the superior performance, broader applicability, and ability to predict highly unstable TUs makes EPIGENE a valuable method to provide genome-wide TU annotations. EPIGENE annotations will improve current TU annotations as more data becomes available and additionally provide valuable insights about transcriptomic landscape across cell types and tissues.

5 Bibliography

Adams, D. *et al.* (2012) 'BLUEPRINT to decode the epigenetic signature written in blood', *Nature Biotechnology*. Nature Publishing Group, 30(3), pp. 224–226. doi: 10.1038/nbt.2153.

Ahn, S. H., Kim, M. and Buratowski, S. (2004) 'Phosphorylation of Serine 2 within the RNA Polymerase II C-Terminal Domain Couples Transcription and 3′ End Processing', *Molecular Cell*. Cell Press, 13(1), pp. 67–76. doi: 10.1016/S1097-2765(03)00492-1.

Andrews, S. (2010) 'FastQC: A Quality Control tool for High Throughput Sequence Data', *Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bail, S. *et al.* (2010) 'Differential regulation of microRNA stability.', *RNA (New York, N.Y.)*. Cold Spring Harbor Laboratory Press, 16(5), pp. 1032–9. doi: 10.1261/rna.1851510.

Barski, A. *et al.* (2007) 'High-Resolution Profiling of Histone Methylations in the Human Genome', *Cell*, 129(4), pp. 823–837. doi: 10.1016/j.cell.2007.05.009.

Bartel, D. P. (2004) 'MicroRNAs', *Cell*. Elsevier, 116(2), pp. 281–297. doi: 10.1016/S0092-8674(04)00045-5.

Baum, L. E. *et al.* (1970) 'A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains', *The Annals of Mathematical Statistics*. Institute of Mathematical Statistics, 41(1), pp. 164–171. doi: 10.1214/aoms/1177697196.

Bauman, J. G. J. *et al.* (1980) 'A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA', *Experimental Cell Research*, 128(2), pp. 485–490. doi: 10.1016/0014-4827(80)90087-7.

Bernstein, B. E. *et al.* (2002) 'Methylation of histone H3 Lys 4 in coding regions

of active genes', *Proceedings of the National Academy of Sciences*, 99(13), pp. 8695–8700. doi: 10.1073/pnas.082249499.

Bernstein, B. E. *et al.* (2010) 'The NIH Roadmap Epigenomics Mapping Consortium', *Nature Biotechnology*, 28(10), pp. 1045–1048. doi: 10.1038/nbt1010-1045.

Bouvy-Liivrand, M. *et al.* (2017) 'Analysis of primary microRNA loci from nascent transcriptomes reveals regulatory domains governed by chromatin architecture', *Nucleic Acids Research*. Oxford Academic, 45(17), pp. 9837–9849. doi: 10.1093/nar/gkx680.

Bujold, D. *et al.* (2016) 'The International Human Epigenome Consortium Data Portal.', *Cell systems*. Elsevier, 3(5), pp. 496-499.e2. doi: 10.1016/j.cels.2016.10.019.

Burger, K., Schlackow, M. and Gullerova, M. (2019) 'Tyrosine kinase c-Abl couples RNA polymerase II transcription to DNA double-strand breaks', *Nucleic Acids Research*, 47(7), pp. 3467–3484. Available at: https://academic.oup.com/nar/article/47/7/3467/5298634 (Accessed: 2 June 2020).

Calin, G. A. and Croce, C. M. (2006) 'MicroRNA signatures in human cancers', *Nature Reviews Cancer*. Nature Publishing Group, 6(11), pp. 857–866. doi: 10.1038/nrc1997.

Carleton, M., Cleary, M. A. and Linsley, P. S. (2007) 'MicroRNAs and Cell Cycle Regulation', *Cell Cycle*. Taylor & Francis, 6(17), pp. 2127–2132. doi: 10.4161/cc.6.17.4641.

CEEHRC (2013) 'Canadian Epigenetics, Environment and Health Research Consortium', *http://www.epigenomes.ca/*, (Accessed on: 16.03.2020). Available at: http://www.epigenomes.ca/ (Accessed: 31 January 2019).

Chae, M., Danko, C. G. and Kraus, W. L. (2015) 'groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data', *BMC Bioinformatics*. BioMed Central, 16(1), p. 222. doi: 10.1186/s12859-015-0656-3.

Chevreux, B. *et al.* (2004) 'Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs', *Genome*

Research. Cold Spring Harbor Laboratory Press, 14(6), pp. 1147–1159. doi: 10.1101/gr.1917404.

Clough, E. and Barrett, T. (2016) 'The Gene Expression Omnibus Database', in. Humana Press, New York, NY, pp. 93–110. doi: 10.1007/978-1-4939-3578-9_5.

Core, L. J. *et al.* (2008) 'Transcription regulation through promoter-proximal pausing of RNA polymerase II.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 319(5871), pp. 1791–2. doi: 10.1126/science.1150843.

Core, L. J. *et al.* (2014) 'Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers', *Nature Genetics*. Nature Publishing Group, 46(12), pp. 1311–1320. doi: 10.1038/ng.3142.

Creyghton, M. P. *et al.* (2010) 'Histone H3K27ac separates active from poised enhancers and predicts developmental state', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 107(50), pp. 21931–21936. doi: 10.1073/PNAS.1016071107.

DEEP (2012) 'The German epigenome programme', *http://www.deutsches-epigenom-programm.de/*, (Accessed on: 16.03.2020).

Dobin, A. *et al.* (2013) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*. Oxford University Press, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Dunham, I. *et al.* (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*. Nature Publishing Group, 489(7414), pp. 57–74. doi: 10.1038/nature11247.

Durham, T. J. *et al.* (2018) 'PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition', *Nature Communications*. Nature Publishing Group, 9(1), p. 1402. doi: 10.1038/s41467-018-03635-9.

Ernst, J. and Kellis, M. (2012) 'ChromHMM: automating chromatin-state discovery and characterization', *Nature Methods*. Nature Publishing Group, 9(3), pp. 215–

216. doi: 10.1038/nmeth.1906.

Ernst, J. and Kellis, M. (2015) 'Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues', *Nature Biotechnology*. Nature Publishing Group, 33(4), pp. 364–376. doi: 10.1038/nbt.3157.

Ernst, J. and Kellis, M. (2017) 'Chromatin-state discovery and genome annotation with ChromHMM', *Nature Protocols*. Nature Publishing Group, 12(12), pp. 2478–2492. doi: 10.1038/nprot.2017.124.

Fariselli, P., Martelli, P. L. and Casadio, R. (2005) 'A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins', *BMC Bioinformatics*. BioMed Central, 6(SUPPL.4), p. S12. doi: 10.1186/1471-2105-6-S4-S12.

Feingold, E. A. *et al.* (2004) 'The ENCODE (ENCyclopedia Of DNA Elements) Project.', *Science*. American Association for the Advancement of Science, 306(5696), pp. 636–40. doi: 10.1126/science.1105136.

Flanagan, J. F. *et al.* (2005) 'Double chromodomains cooperate to recognize the methylated histone H3 tail', *Nature*. Nature Publishing Group, 438(7071), pp. 1181–1185. doi: 10.1038/nature04290.

Fong, N. *et al.* (2015) 'Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition', *Molecular Cell*. Cell Press, 60(2), pp. 256–267. doi: 10.1016/j.molcel.2015.09.026.

Frankish, A. *et al.* (2019) 'GENCODE reference annotation for the human and mouse genomes', *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D766–D773. doi: 10.1093/nar/gky955.

Gardini, A. (2017) 'Global Run-On Sequencing (GRO-Seq).', *Methods in molecular biology (Clifton, N.J.)*. NIH Public Access, 1468, pp. 111–20. doi: 10.1007/978-1-4939-4035-6_9.

Grabherr, M. G. *et al.* (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nature Biotechnology*. Nature Publishing Group, 29(7), pp. 644–652. doi: 10.1038/nbt.1883.

Griffiths-Jones, S. *et al.* (2006) 'miRBase: microRNA sequences, targets and gene nomenclature', *Nucleic Acids Research*. Oxford University Press, 34(90001), pp. D140–D144. doi: 10.1093/nar/gkj112.

Gustafsson, C. *et al.* (2019) 'High-throughput ChIPmentation: Freely scalable, single day ChIPseq data generation from very low cell-numbers', *BMC Genomics*. BioMed Central Ltd., 20(1), p. 59. doi: 10.1186/s12864-018-5299-0.

Harrow, J. *et al.* (2012) 'GENCODE: the reference human genome annotation for The ENCODE Project.', *Genome research*, 22(9), pp. 1760–74. doi: 10.1101/gr.135350.111.

He, L. and Hannon, G. J. (2004) 'MicroRNAs: small RNAs with a big role in gene regulation', *Nature Reviews Genetics*. Nature Publishing Group, 5(7), pp. 522–531. doi: 10.1038/nrg1379.

Hoffman, M. M. *et al.* (2012) 'Unsupervised pattern discovery in human chromatin structure through genomic segmentation', *Nature Methods*. Nature Publishing Group, 9(5), pp. 473–476. doi: 10.1038/nmeth.1937.

IHEC (2012) *Reference Epigenome Standards, http://ihec-epigenomes.org/research/reference-epigenome-standards/*. Available at: http://ihec-epigenomes.org/research/reference-epigenome-standards/ (Accessed: 20 July 2020).

Jacquier, A. (2009) 'The complex eukaryotic transcriptome: Unexpected pervasive transcription and novel small RNAs', *Nature Reviews Genetics*. Nature Publishing Group, pp. 833–844. doi: 10.1038/nrg2683.

Janes, J. *et al.* (2015) 'A comparative study of RNA-seq analysis strategies', *Briefings in Bioinformatics*. Oxford University Press, 16(6), pp. 932–940. doi: 10.1093/bib/bbv007.

Jin, F. *et al.* (2013) 'A high-resolution map of the three-dimensional chromatin

interactome in human cells', *Nature*. Nature Publishing Group, 503(7475), pp. 290–294. doi: 10.1038/nature12644.

Johnson, D. S. *et al.* (2007) 'Genome-wide mapping of in vivo protein-DNA interactions', *Science*. American Association for the Advancement of Science, 316(5830), pp. 1497–1502. doi: 10.1126/science.1141319.

Karlic, R. *et al.* (2010) 'Histone modification levels are predictive for gene expression', *Proceedings of the National Academy of Sciences*, 107(7), pp. 2926–2931. doi: 10.1073/pnas.0909344107.

Kim, D., Langmead, B. and Salzberg, S. L. (2015) 'HISAT: A fast spliced aligner with low memory requirements', *Nature Methods*. Nature Publishing Group, 12(4), pp. 357–360. doi: 10.1038/nmeth.3317.

Kim, T. K. *et al.* (2010) 'Widespread transcription at neuronal activity-regulated enhancers', *Nature*. Nature Publishing Group, 465(7295), pp. 182–187. doi: 10.1038/nature09033.

Kim, V. N. and Nam, J.-W. (2006) 'Genomics of microRNA.', *Trends in genetics : TIG*. Elsevier, 22(3), pp. 165–73. doi: 10.1016/j.tig.2006.01.003.

Kinkley, S. *et al.* (2016) 'reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in CD4+ memory T cells', *Nature Communications*. Nature Publishing Group, 7(1), p. 12514. doi: 10.1038/ncomms12514.

Kostrewa, D. *et al.* (2009) 'RNA polymerase II-TFIIB structure and mechanism of transcription initiation', *Nature*. Nature Publishing Group, 462(7271), pp. 323–330. doi: 10.1038/nature08548.

Kwak, H. *et al.* (2013) 'Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing', *Science (New York, N.Y.)*, 339(6122), pp. 950–953. doi: 10.1126/science.1229386.

Lagos-Quintana, M. *et al.* (2001) 'Identification of Novel Genes Coding for Small Expressed RNAs', *Science (New York, N.Y.)*, 294(5543), pp. 853–858. doi:

10.1126/science.1064921.

Langmead, B. *et al.* (2009) 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome', *Genome Biology*. BioMed Central, 10(3), p. R25. doi: 10.1186/gb-2009-10-3-r25.

Lappalainen, I. *et al.* (2015) 'The European Genome-phenome Archive of human data consented for biomedical research', *Nature Genetics*. Nature Publishing Group, pp. 692–695. doi: 10.1038/ng.3312.

Lawrence, M. *et al.* (2013) 'Software for Computing and Annotating Genomic Ranges', *PLoS Computational Biology*. Edited by A. Prlic. Public Library of Science, 9(8), p. e1003118. doi: 10.1371/journal.pcbi.1003118.

Le, H.-S. *et al.* (2016) 'Probabilistic error correction for RNA sequencing', *Nucleic Acids Research*, 41(10), p. e109. doi: https://doi.org/10.1093/nar/gkt215.

Lee, R. C. and Ambros, V. (2001) 'An Extensive Class of Small RNAs in Caenorhabditis elegans', *Science (New York, N.Y.)*, 294(5543), pp. 862–864. doi: 10.1126/science.1065329.

Leinonen, R. *et al.* (2011) 'The European Nucleotide Archive', *Nucleic Acids Research*, 39, pp. D28–D31. doi: 10.1093/nar/gkq967.

Lex, A. *et al.* (2014) 'UpSet: Visualization of intersecting sets', *IEEE Transactions on Visualization and Computer Graphics*. IEEE Computer Society, 20(12), pp. 1983–1992. doi: 10.1109/TVCG.2014.2346248.

Li, H. *et al.* (2006) 'Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF', *Nature*. Nature Publishing Group, 442(7098), pp. 91–95. doi: 10.1038/nature04802.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform.', *Bioinformatics (Oxford, England)*, 25(14), pp. 1754–60. doi: 10.1093/bioinformatics/btp324.

Li, Y. *et al.* (2013) 'Genome-Wide Analysis of Human MicroRNA Stability',

BioMed Research International, 2013, pp. 1–12. doi: 10.1155/2013/368975.

Lidschreiber, M., Leike, K. and Cramer, P. (2013) 'Cap Completion and C-Terminal Repeat Domain Kinase Recruitment Underlie the Initiation-Elongation Transition of RNA Polymerase II', *Molecular and Cellular Biology*. American Society for Microbiology, 33(19), pp. 3805–3816. doi: 10.1128/mcb.00361-13.

Lister, R. *et al.* (2009) 'Human DNA methylomes at base resolution show widespread epigenomic differences', *Nature*. Nature Publishing Group, 462(7271), pp. 315–322. doi: 10.1038/nature08514.

Liu, J. *et al.* (2016) 'TransComb: Genome-guided transcriptome assembly via combing junctions in splicing graphs', *Genome Biology*. BioMed Central Ltd., 17(1), p. 213. doi: 10.1186/s13059-016-1074-1.

Mammana, A. and Chung, H.-R. (2015) 'Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome', *Genome Biology*. BioMed Central, 16(1), p. 151. doi: 10.1186/s13059-015-0708-z.

Mammana, A. and Helmuth, J. (2015) 'Introduction to the bamsignals package'. Available at: http://bioconductor.org/packages/release/bioc/html/bamsignals.html (Accessed: 31 January 2019).

Mangiavacchi, A. *et al.* (2016) 'The miR-223 host non-coding transcript linc-223 induces IRF4 expression in acute myeloid leukemia by acting as a competing endogenous RNA', *Oncotarget*. Impact Journals LLC, 7(37), pp. 60155–60168. doi: 10.18632/oncotarget.11165.

Mayer, A. *et al.* (2010) 'Uniform transitions of the general RNA polymerase II transcription complex', *Nature Structural and Molecular Biology*. Nature Publishing Group, 17(10), pp. 1272–1278. doi: 10.1038/nsmb.1903.

Mayer, A. *et al.* (2015) 'Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution', *Cell*. Cell Press, 161(3), pp. 541–554. doi:

10.1016/j.cell.2015.03.010.

McLauchlan, J. *et al.* (1985) 'The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini.', *Nucleic Acids Research*, 13(4), pp. 1347–1368. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC341077/ (Accessed: 2 June 2020).

Medina, I. *et al.* (2016) 'Highly sensitive and ultrafast read mapping for RNA-seq analysis', *DNA Research*, 23(2), pp. 93–100. Available at: https://academic.oup.com/dnaresearch/article/23/2/93/1745299 (Accessed: 22 May 2020).

Meister, G. and Tuschl, T. (2004) 'Mechanisms of gene silencing by double-stranded RNA', *Nature*. Nature Publishing Group, 431(7006), pp. 343–349. doi: 10.1038/nature02873.

Mikkelsen, T. S. *et al.* (2007) 'Genome-wide maps of chromatin state in pluripotent and lineage-committed cells', *Nature*. Nature Publishing Group, 448(7153), pp. 553–560. doi: 10.1038/nature06008.

Mortazavi, A. *et al.* (2008) 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nature Methods*. Nature Publishing Group, 5(7), pp. 621–628. doi: 10.1038/nmeth.1226.

Nagalakshmi, U. *et al.* (2008) 'The transcriptional landscape of the yeast genome defined by RNA sequencing', *Science*. NIH Public Access, 320(5881), pp. 1344–1349. doi: 10.1126/science.1158441.

Nikolov, D. B. and Burley, S. K. (1997) 'RNA polymerase II transcription initiation: A structural view', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 94(1), pp. 15–22. doi: 10.1073/pnas.94.1.15.

Nishioka, K. *et al.* (2002) 'Set9, a novel histone H3 methyltransferase that facilitates transcription by precluding histone tail modifications required for heterochromatin formation', *Genes and Development*. Cold Spring Harbor Laboratory Press, 16(4), pp. 479–489. doi: 10.1101/gad.967202.

Nojima, T. *et al.* (2015) 'Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing.', *Cell*. Elsevier, 161(3), pp. 526–540. doi: 10.1016/j.cell.2015.03.027.

Park, P. J. (2009) 'ChIP-seq: Advantages and challenges of a maturing technology', *Nature Reviews Genetics*. Nature Publishing Group, pp. 669–680. doi: 10.1038/nrg2641.

Peng, Y. *et al.* (2013) 'IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels', *Bioinformatics*, 29(13), pp. i326–i334. doi: https://doi.org/10.1093/bioinformatics/btt219.

Perner, J. and Chung, H.-R. (2013) 'Chromatin signaling and transcription initiation', *Frontiers in Life Science*. Taylor & Francis, 7(1–2), pp. 22–30. doi: 10.1080/21553769.2013.856038.

Pertea, M. *et al.* (2015) 'StringTie enables improved reconstruction of a transcriptome from RNA-seq reads', *Nature Biotechnology*. Nature Publishing Group, 33(3), pp. 290–295. doi: 10.1038/nbt.3122.

Pertea, M. *et al.* (2018) 'CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise', *Genome Biology*. BioMed Central, 19(1), p. 208. doi: 10.1186/s13059-018-1590-2.

Plasterk, R. H. A. (2006) 'Micro RNAs in Animal Development', *Cell*. Cell Press, 124(5), pp. 877–881. doi: 10.1016/J.CELL.2006.02.030.

Preker, P. *et al.* (2008) 'RNA exosome depletion reveals transcription upstream of active human promoters.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 322(5909), pp. 1851–4. doi: 10.1126/science.1164096.

Proudfoot, N. J. (2011) 'Ending the message: Poly(A) signals then and now', *Genes and Development*. Cold Spring Harbor Laboratory Press, pp. 1770–1782. doi: 10.1101/gad.17268411.

Ramírez, F. *et al.* (2014) 'deepTools: a flexible platform for exploring deep-

sequencing data', *Nucleic Acids Research*. Narnia, 42(W1), pp. W187–W191. doi: 10.1093/nar/gku365.

Richard, P. and Manley, J. L. (2009) 'Transcription termination by nuclear RNA polymerases', *Genes and Development*. Cold Spring Harbor Laboratory Press, pp. 1247–1269. doi: 10.1101/gad.1792809.

Sahu, A. *et al.* (2020) 'EPIGENE: Genome-wide transcription unit annotation using a multivariate probabilistic model of histone modifications', *Epigenetics and Chromatin*. BioMed Central Ltd., 13(1), p. 20. doi: 10.1186/s13072-020-00341-z.

Sainsbury, S., Niesser, J. and Cramer, P. (2013) 'Structure and function of the initially transcribing RNA polymerase II-TFIIB complex', *Nature*. Nature Publishing Group, 493(7432), pp. 437–440. doi: 10.1038/nature11715.

Salhab, A. *et al.* (2018) 'A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains', *Genome Biology*. BioMed Central, 19(1), p. 150. doi: 10.1186/s13059-018-1510-5.

Schneider, R. *et al.* (2004) 'Direct binding of INHAT to H3 tails disrupted by modifications', *Journal of Biological Chemistry*. JBC Papers in Press, 279(23), pp. 23859–23862. doi: 10.1074/jbc.C400151200.

Schulz, M. H. *et al.* (2012) 'Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels', *Bioinformatics*. Oxford University Press, 28(8), pp. 1086–1092. doi: 10.1093/bioinformatics/bts094.

Schwalb, B. *et al.* (2016) 'TT-seq maps the human transient transcriptome', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 352(6290), pp. 1225–1228. doi: 10.1126/science.aad9841.

Sethi, P. and Lukiw, W. J. (2009) 'Micro-RNA abundance and stability in human brain: Specific alterations in Alzheimer's disease temporal lobe neocortex', *Neuroscience Letters*. Elsevier, 459(2), pp. 100–104. doi: 10.1016/j.neulet.2009.04.052.

Shah, M. Y. *et al.* (2016) 'microRNA Therapeutics in Cancer — An Emerging

Concept', *EBioMedicine*. Elsevier B.V., pp. 34–42. doi: 10.1016/j.ebiom.2016.09.017.

Sikorski, T. W. and Buratowski, S. (2009) 'The basal initiation machinery: beyond the general transcription factors', *Current Opinion in Cell Biology*. NIH Public Access, pp. 344–351. doi: 10.1016/j.ceb.2009.03.006.

Skourti-Stathaki, K., Proudfoot, N. J. and Gromak, N. (2011) 'Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination', *Molecular Cell*. Elsevier, 42(6), pp. 794–805. doi: 10.1016/j.molcel.2011.04.026.

Stunnenberg, H. G. *et al.* (2016) 'The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery', *Cell*. Cell Press, pp. 1145–1149. doi: 10.1016/j.cell.2016.11.007.

Tani, H. *et al.* (2012) 'Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals', *Genome Research*, 22(5), pp. 947–956. doi: 10.1101/gr.130559.111.

Tong, L. *et al.* (2016) 'Evaluating the impact of sequencing error correction for RNA-seq data with ERCC RNA spike-in controls', in *3rd IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2016*. Institute of Electrical and Electronics Engineers Inc., pp. 74–77. doi: 10.1109/BHI.2016.7455838.

Trapnell, C. *et al.* (2010) 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nature Biotechnology*. Nature Publishing Group, 28(5), pp. 511–515. doi: 10.1038/nbt.1621.

Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) 'TopHat: discovering splice junctions with RNA-Seq', *Bioinformatics*, 25(9), pp. 1105–1111. Available at: https://academic.oup.com/bioinformatics/article/25/9/1105/203994 (Accessed: 22 May 2020).

Turner, B. M. (2012) 'The adjustable nucleosome: an epigenetic signaling module', *Trends in Genetics*. Elsevier Current Trends, 28(9), pp. 436–444. doi:

10.1016/J.TIG.2012.04.003.

Venturini, L. *et al.* (2018) 'Leveraging multiple transcriptome assembly methods for improved gene structure annotation', *GigaScience*. Oxford University Press, 7(8). doi: 10.1093/gigascience/giy093.

Viterbi, A. (1967) 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm', *IEEE Transactions on Information Theory*, 13(2), pp. 260–269. doi: 10.1109/TIT.1967.1054010.

Wallace, J. A. and O'Connell, R. M. (2017) 'MicroRNAs and acute myeloid leukemia: Therapeutic implications and emerging concepts', *Blood*. American Society of Hematology, pp. 1290–1301. doi: 10.1182/blood-2016-10-697698.

Wang, J. *et al.* (2018) 'Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation', *BMC Genomics*. BioMed Central, 19(1), p. 633. doi: 10.1186/s12864-018-5016-z.

Wang, M., Qin, L. and Tang, B. (2019) 'MicroRNAs in Alzheimer's disease', *Frontiers in Genetics*. Frontiers Media S.A. doi: 10.3389/fgene.2019.00153.

Williams, L. H. *et al.* (2015) 'Pausing of RNA Polymerase II Regulates Mammalian Developmental Potential through Control of Signaling Networks', *Molecular Cell*. Cell Press, 58(2), pp. 311–322. doi: 10.1016/j.molcel.2015.02.003.

Wissink, E. M. *et al.* (2019) 'Nascent RNA analyses: tracking transcription and its regulation', *Nature Reviews Genetics*. Nature Publishing Group, 20(12), pp. 705–723. doi: 10.1038/s41576-019-0159-6.

Won, K.-J. *et al.* (2013) 'Comparative annotation of functional regions in the human genome using epigenomic data', *Nucleic Acids Research*. Narnia, 41(8), pp. 4423–4432. doi: 10.1093/nar/gkt143.

Wysoker, A., Tibbetts, K., and Fennell, T. (2013) *Picard Tools*. Available at: http://broadinstitute.github.io/picard/ (Accessed: 12 February 2019).

Xu, Z. *et al.* (2009) 'Bidirectional promoters generate pervasive transcription in

yeast', *Nature*. Nature Publishing Group, 457(7232), pp. 1033–1037. doi: 10.1038/nature07728.

Yunger, S. *et al.* (2010) 'Single-allele analysis of transcription kinetics in living mammalian cells', *Nature Methods*. Nature Publishing Group, 7(8), pp. 631–633. doi: 10.1038/nmeth.1482.

Zacher, B. *et al.* (2017) 'Accurate Promoter and Enhancer Identification in 127 ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN', *PLOS ONE*. Edited by R. Mantovani. Public Library of Science, 12(1), p. e0169249. doi: 10.1371/journal.pone.0169249.

Zamore, P. D. and Haley, B. (2005) 'Ribo-gnome: the big world of small RNAs.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 309(5740), pp. 1519–24. doi: 10.1126/science.1111444.

Zhang, Z. *et al.* (2019) 'Transcriptional landscape and clinical utility of enhancer RNAs for eRNA-targeted therapy in cancer', *Nature Communications*. Nature Publishing Group, 10(1), p. 4562. doi: 10.1038/s41467-019-12543-5.

6 Appendix

6.1 List of datasets used

Table S1: List of datasets used in this study

Cell line	Residue	Sequencing	Accession number	Source
IMR90	H3K27ac	ChIP-seq	ENCSR002YRE	ENCODE
IMR90	H3K4me3	ChIP-seq	ENCSR087PFU	ENCODE
IMR90	H3K4me1	ChIP-seq	ENCSR831JSP	ENCODE
IMR90	H3K36me3	ChIP-seq	ENCSR437ORF	ENCODE
IMR90	H3K27me3	ChIP-seq	ENCSR431UUY	ENCODE
IMR90	H3K9me3	ChIP-seq	ENCSR055ZZY	ENCODE
IMR90	input	ChIP-seq	ENCSR001BSB, ENCSR704GTT	ENCODE
IMR90	Pol II-Input	ChIP-seq	ENCSR000EFL	ENCODE
IMR90	Pol II	ChIP-seq	ENCSR000EFK	ENCODE
IMR90	Poly A selected RNA	RNA-seq	ENCSR000CTQ	ENCODE
HepG2	Pol II-Input	ChIP-seq	ENCSR000EEM	ENCODE
HepG2	Pol II	ChIP-seq	ENCSR000EEN	ENCODE
HepG2	H3K27ac, H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3, input, poly A selected RNA	ChIP-seq	EGAD00001002527	DEEP
K562	Poly A selected RNA	RNA-seq	SRR315336, SRR315337	ENA

K562	Nascent RNA	TT-Seq	GSE75792	GEO
K562	Nascent RNA	GRO-Seq	GSM1480325	GEO
HepG2	Nascent RNA	GRO-Seq	GSM2428726	GEO
IMR90	Nascent RNA	GRO-Seq	GSM1055806	GEO

6.2 Summary statistics of EPIGENE, StringTie, and Cufflinks TUs

Table S2: Summary statistics of EPIGENE predicted TUs in K562

	genes	+ strand	- strand	median length
all	24,571	13,410	11,161	7,800
gencode V19 + chess 2.1 same strand overlap	18,184	9,774	8, 410	9,800
gencode V19 + chess 2.1 any overlap	23,542	12,921	10,621	8,400
no match	1,029	489	540	2,000

Table S3: Summary statistics of StringTie predicted TUs in K562

	genes	+ strand	- strand	median length
all	101,656	50,636	51,020	5,481
gencode V19 + chess 2.1 same strand overlap	93,006	46,448	46,558	6,719
gencode V19 + chess 2.1 any overlap	97,300	48,531	48,769	6,110
no match	4,356	2,105	2,251	613

Table S4: Summary statistics of Cufflinks predicted TUs in K562

	genes	+ strand	- strand	median length
all	32,079	15,262	15,095	8,851

gencode V19 + chess 2.1 same strand overlap	26,452	12,986	12,671	16,486
gencode V19 + chess 2.1 any overlap	27,157	13,320	13,042	15,392
no match	4,992	1,942	2,053	962

6.3 Additional Figures

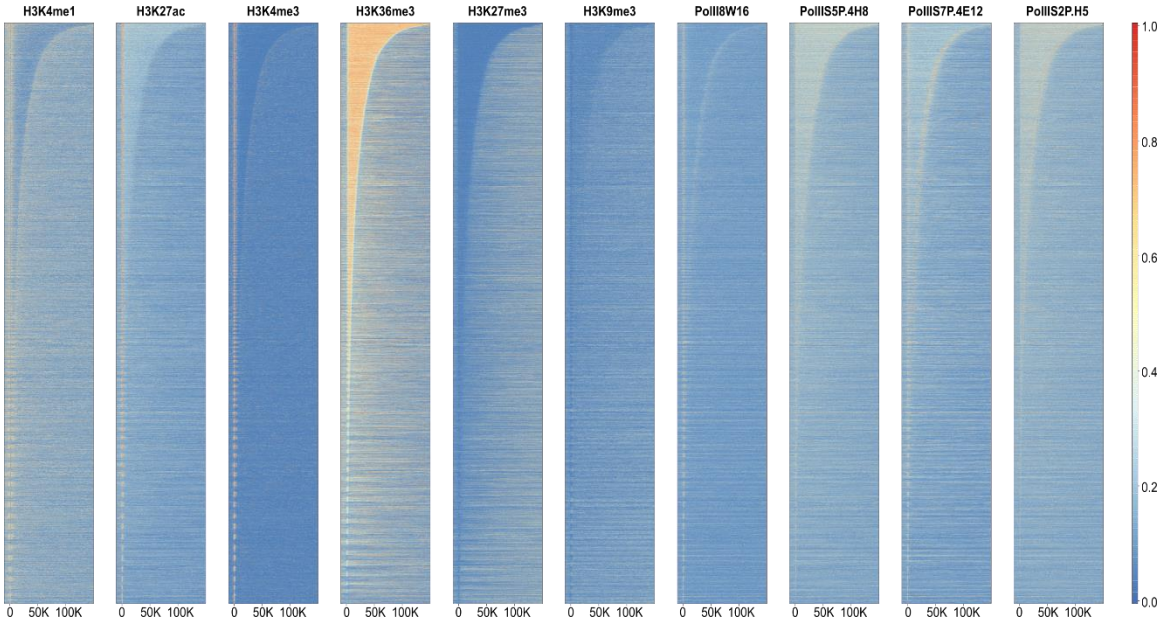


Figure S1: ChIP-seq profiles of IHEC class 1 histone modifications and Pol II

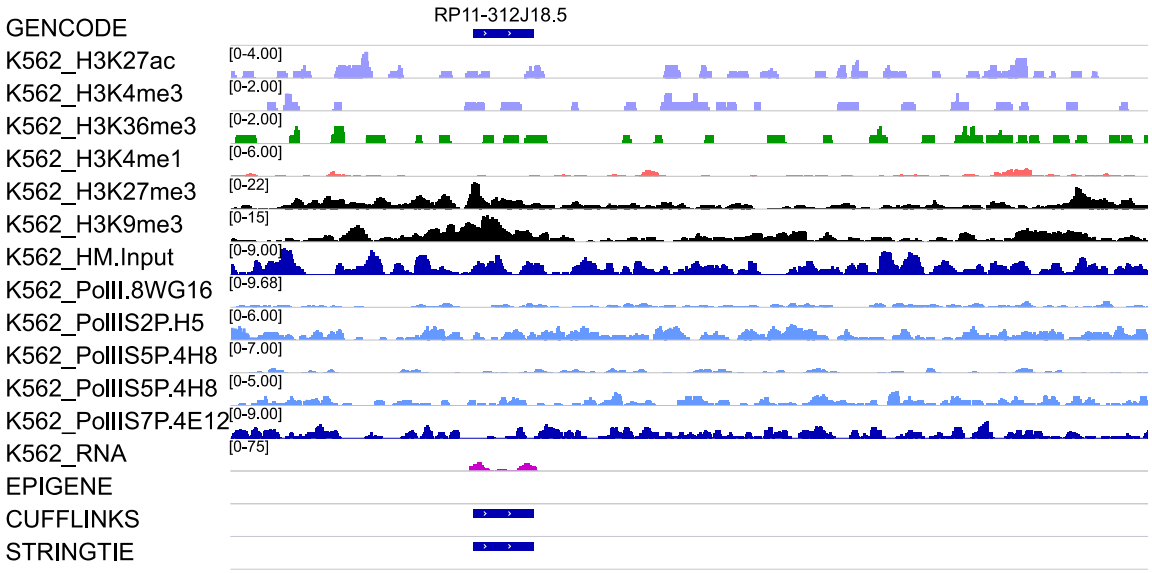


Figure S2: STRINGTIE and CUFFLINKS TU identified due to spurious read mapping

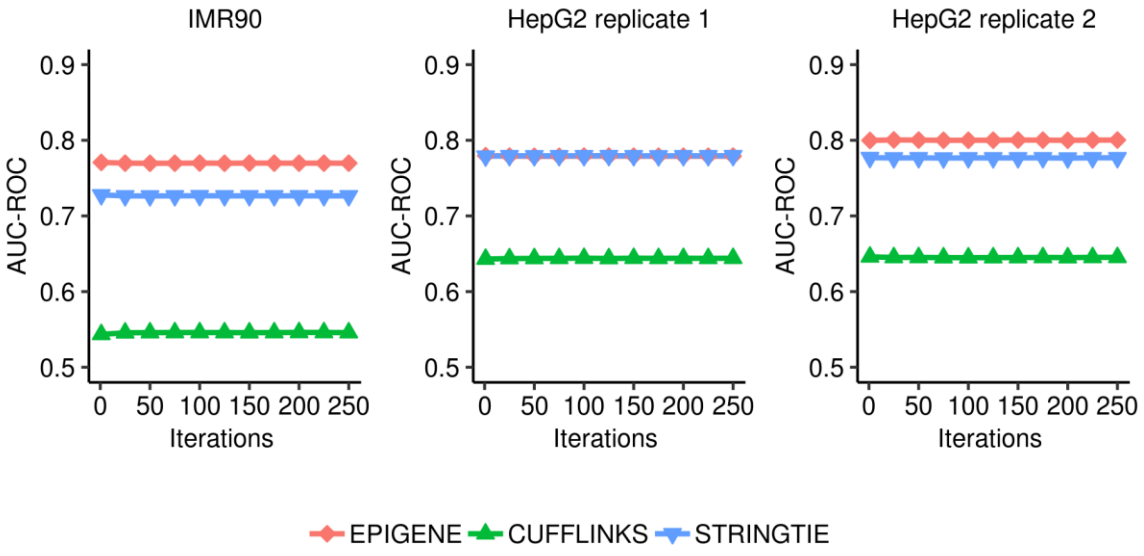


Figure S3: Comparing K562-trained EPIGENE model, STRINGTIE and CUFFLINKS across cell lines

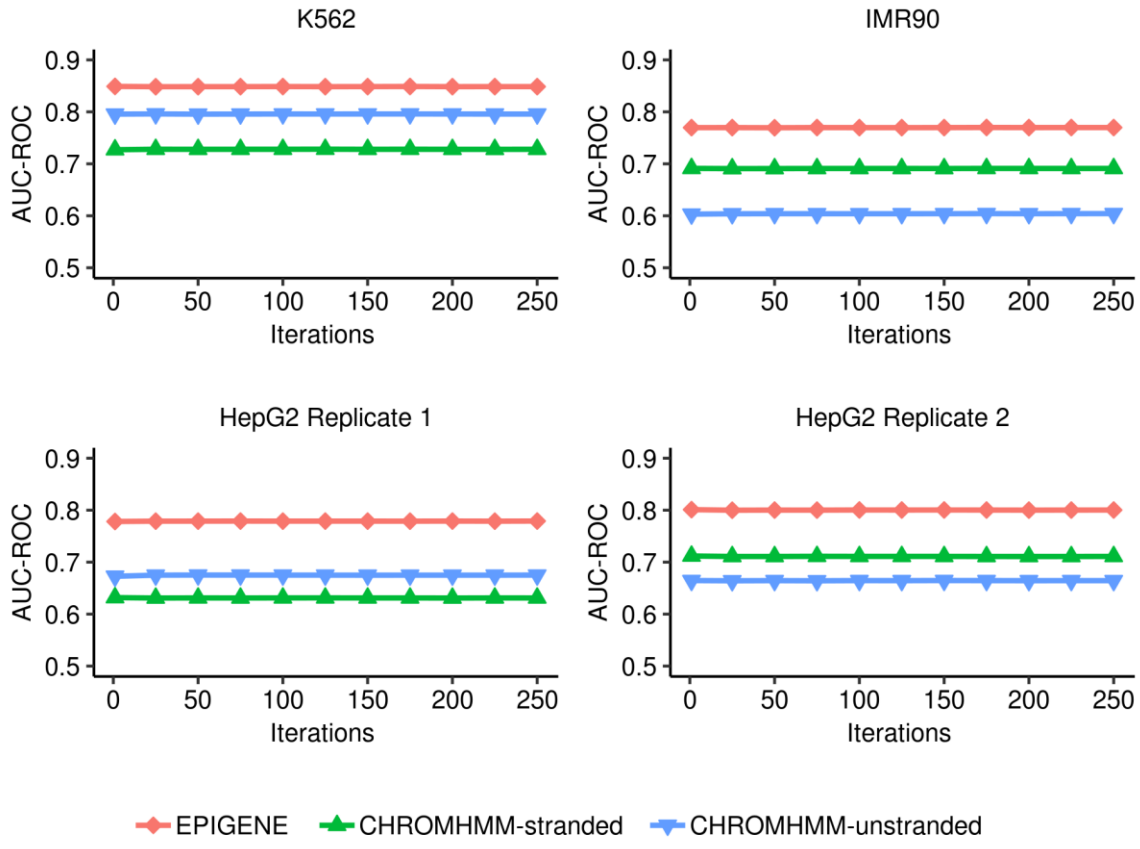


Figure S4: Comparing K562-trained EPIGENE and ChromHMM models across cell lines

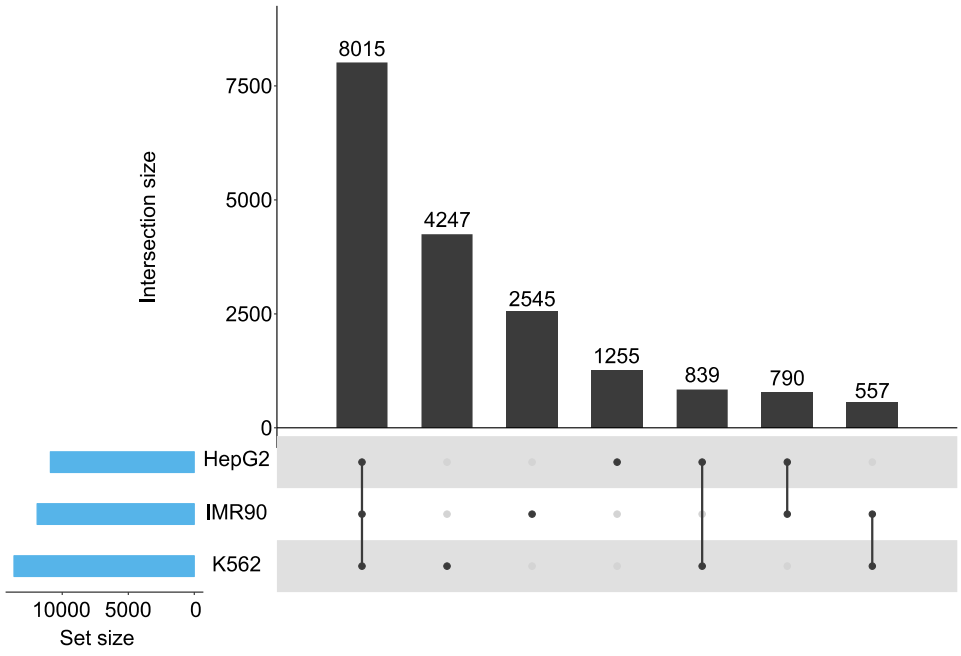


Figure S5: Distribution of EPIGENE TUs across cell lines

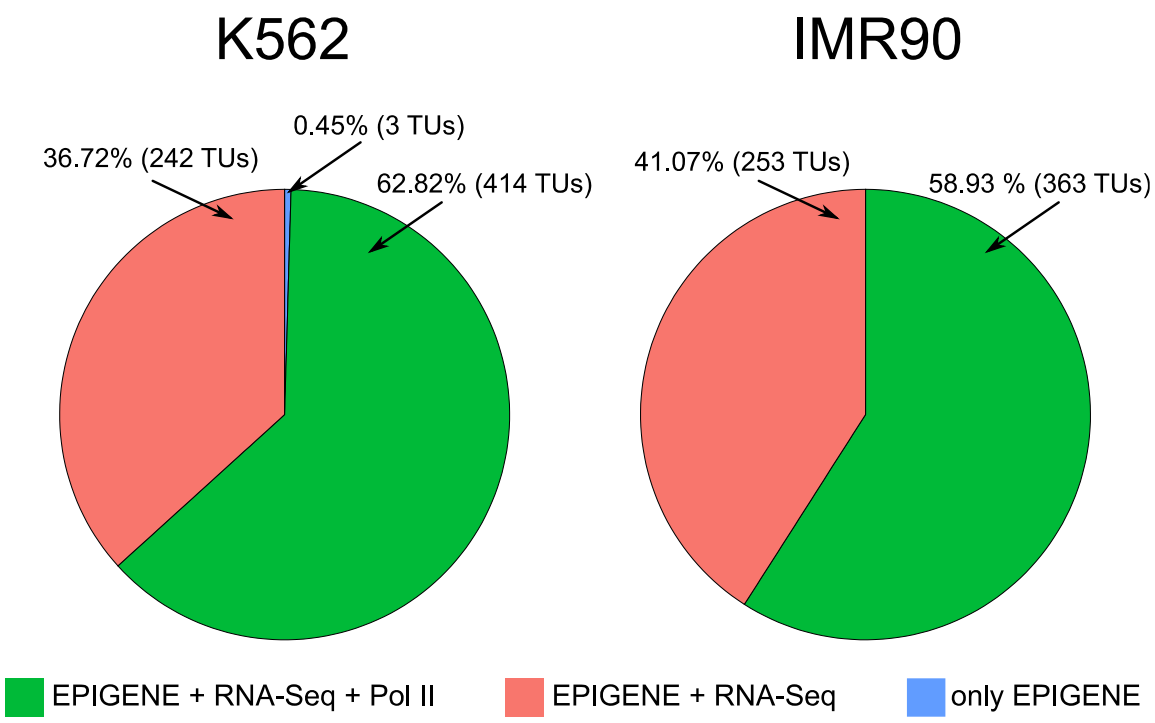


Figure S6: EPIGENE predicted TUs overlapping miRbase annotations across K562 and IMR90 cell line

Verzeichnis der akademischen Lehrer/-innen

Meine akademischen Lehrenden an der Philipps-Universität in Marburg waren:

Fachbereich Medizin:		
Adhikary	Brehm	Chung
Fachbereich Informatik:		
Heider		

Meine akademischen Lehrenden an der Rheinische Friedrich-Wilhelms-Universität Bonn waren:

Faculty of Life Science Informatics:		
Bajorath	Berlage	Fluck
Froehlich	Hoffmann-Apitius	Reitalmann
Schulz	Schoof	Senger
Weber	Zimmermann	

Meine akademischen Lehrenden im Deutsches Krebsforschungszentrum waren:

Department of Pediatric Neurooncology:		
Chavez	Pfister	Zapatka