Towards the Construction of a Transcriptional Landscape of the Human Genome: Data Analysis and Data Compression

by

Yuefeng Lin

B.E. Harbin Institute of Technology, 2006

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Master of Computational Biology

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This thesis was presented

by

Yuefeng Lin

It was defended on

August 26, 2014

and approved by

Dr. Carlos J Camacho, Assoc. Professor, Dept. of Computational and Systems Biology

Dr. Daniel M. Zuckerman, Assoc. Professor, Dept. of Computational and Systems Biology

Dr.Nathan L. Clark, Asst. Professor, Dept. of Computational and Systems Biology

Thesis Advisor: Dr. Carlos J Camacho, Assoc. Professor

Towards the Construction of a Transcriptional Landscape of the Human Genome: Data Analysis and Data Compression

Yuefeng Lin

University of Pittsburgh, 2014

In the thesis, we built a genome-wide polyadenylation map with sequencing data sets from various human tissues and cell lines. With the map, we analyzed the pattern and distribution of polyadenylation sites in human genome. And we explored the differential polyadenylation patterns of non-coding and novel genes. Meanwhile, we have created the Expression and Polyadenylation Database (xPAD) as a web portal for the polyadenylation map. Moreover, we revealed the regulatory marks that might correlated with polyadenylation sites we have found.

Besides, we unveiled a novel group of small YB-1 associated RNAs and investigate their possible regulation mechanism where we found multiple transcription factors and histone modification may mark the location of YB-1 associated RNA.

We also implemented an Assembly-based Sequencing data Encoding Tool, AbSEnT. With this tool, we exhibited the feasibility and efficiency of the novel assembly-based compression algorithm by achieving a higher compression ratio than general-purpose compression tools. Meanwhile, we investigated the distribution of word frequency in sequencing data and found it shares a similarity with natural languages we used. If the connection could be proved, we may borrow the knowledge and experience from what we have learned in the research of natural language into the analysis of sequencing data.

TABLE OF CONTENTS

1.0 INTRODUCTION 1

1.1 OVERALL GOALS AND SPECIFIC AIMS 1

1.1.1 Compressing sequencing data with assembly-based algorithm 1

1.1.2 Building comprehensive polyadenylation map of human genome 2

1.1.3 Characterizing a novel group of short RNAs that associates with YB-1 3

1.2 BACKGROUND AND SIGNIFICANCE 3

1.2.1 Gene expression analysis using next-generation sequencing 3

1.2.2 Non-coding RNAs 4

1.2.3 Technology challenges in next-generation sequencing 5

1.2.4 Basic algorithm of data compression 6

2.0 COMPRESSING SEQUENCING DATA WITH ASSEMBLY-BASED ALGORITHM 8

2.1 INTRODUCTION 8

2.1.1 Sequencing Data Accumulation 8

2.1.2 Possible Solutions 10

2.1.3 Redundancy in the sequencing data set 10

2.1.4 Available Frameworks and Tools 12

2.1.5 Advantages and disadvantages 14

2.2 METHODS 16

2.2.1 Framework of the genome-independent compression method AbSEnT 16
2.2.2 de-Brujin graph build-up 18
2.2.3 Simplification and pruning of de-Brujin graph 19
2.2.3.1 Branch zipper and Bubble zipper 20
2.2.3.2 Low importantance nodes filter 21
2.2.4 Mapping and encoding of the Sequences 22
2.2.5 Decoding of the sequences 23
2.3 RESULTS 23
2.3.1 Redundancy of information in sequencing dataset 23
2.3.2 de-Brujin Graph pruning and simplification 26
2.3.3 Compression efficiency of AbSEnT 26
2.4 DISCUSSION 27
2.4.1 Evaluation of compression algorithm 27
2.4.2 Lossless compression and lossy compression 28
2.5 SUMMARY 30
3.0 BUILDING COMPREHENSIVE POLYADENYLATION MAP OF HUMAN GENOME 31
3.1 INTRODUCTION 31
3.1.1 Alternative polyadenylation 32
3.1.2 Genome-wide Polyadenylation map 33
3.2 MATERIALS AND METHODS 34
3.2.1 DRS sequencing and genome mapping 34
3.2.2 Sequencing data preprocessing with in-house tool 36
3.2.3 Determination of usage of short and long 3’ UTR isoforms 38

3.2.4 Motif-enrichment analysis 39
3.2.5 Statistical tests 40
3.3 RESULTS 40
3.3.1 Genomic feature of polyadenylation sites 41
3.3.2 Polyadenylation patterns of non-coding and novel genes 44
3.3.3 Polyadenylation sites contain isoform-dependent regulatory marks 47
3.4 DISCUSSION 48
3.5 SUMMARY 51
4.0 CHARACTERIZING A NOVEL GROUP OF SHORT RNAS THAT ASSOCIATES WITH YB-1 52
4.1 INTRODUCTION 52
4.2 METHODS 55
4.2.1 Genomic analysis 55
4.2.2 Statistical analysis 55
4.3 RESULTS 58
4.3.1 Identified novel class of small RNAs associated with YB1 protein 58
4.3.2 Regulatory marks around shyRNA locations 59
4.3.3 Transcription factors and histone modifications mark shyRNA locations 61
4.4 DISCUSSION 65
4.5 SUMMARY 66
5.0 CONCLUSIONS 68
APPENDIX A 70
APPENDIX B 76
BIBLIOGRAPHY 114

LIST OF TABLES

Table 1. Bzip2 and AbSEnT compression statistics on *Pseudomonas Syringae* whole genome sequencing data. 27

LIST OF FIGURES

Figure 1. Sequencing Data Accumulation. 9

Figure 2: Compression and decompression schema for read sequences data..... 17

Figure 3. The distribution of word frequency in sequencing data sets.. 24

Figure 4. Galley of dense graphs after pruning and simplication.. 25

Figure 5: The histogram of KL-divergence of a set of random generated sequences.. 38

Figure 6: Characteristics of polyadenylation sites.. 41

Figure 7: Overall properties of polyadenylation sites are consistent across tissues.. 44

Figure 8: Genomic view of polyadenylated non-coding RNAs and novel gene locations that are aberrantly expressed in cancer using xPAD.. 45

Figure 9: Polyadenylation maps enable the identification of isoform-dependent regulatory marks.47

Figure 10: Illustration of xPAD.. 49

Figure 11: Characteristics of YB-1 associated small and short RNAs. 57

Figure 12: Specific transcription factor and histone modifications mark shyRNA loci in a cell-type dependent manner.. 60

Figure 13. Specific transcription factor and histone modifications mark shyRNA loci.. 64

1.0 INTRODUCTION

1.1 OVERALL GOALS AND SPECIFIC AIMS

Cells make various types of primary and processed RNAs in the course of expressing the encoded genetic information in genomes. The produced RNAs carry a significant portion of the responsibility of directing the synthesis of proteins and regulating cell behavior and gene expression. Thus, the synthesis, processing, transport, modification and translation of RNAs reveal us how the genetic information is utilized and the cell functions. However, the complete landscape of these RNAs is not readily available and the different characteristics of them are still poorly understood. Therefore, we aim to construct a complete transcriptional landscape of the human genome and further investigate patterns and characteristics of previously unannotated RNAs. In the meantime, due to the fact that massive sequencing data are generated in the field, the problem of data storage and share hinders the fast-pace development of the field. To accommodate it, we expect to develop a novel computational method to efficiently compress sequencing data.

1.1.1 Compressing sequencing data with assembly-based algorithm

We compressed large next-generation sequencing (NGS) data to efficiently store and transfer data sets. Specifically, we aim to construct an assembly-based compression method. Unlike the existing method that using available curated genome, we used a genome assembled directly from each

dataset by de-novo assembler, which would be suitable when a high-quality reference is not available or the sequencing is made on a mixture of samples from multiple and might be unknown genomes. We compressed not only the sequence data but also quality and read identifier data all together. The method was superior to general-purpose compressor, e.g. gzip or bzip2, in term of compression ratio on our test data set. We also targeted to implement the method as a publicly available software package.

- Assemble a virtual genome as reference assembly from sequencing data with de-novo assembler;
- Develop assembly-based method to compress read identifiers, sequences and quality data separately;
- Implement assembly-based method as a publicly available software package.

1.1.2 Building comprehensive polyadenylation map of human genome

To facilitate investigation of polyadenylation patterns in normal and tumor human cells, we built an open-access community resource using Direct RNA sequencing (DRS)[1] with multiple experiments of different tissues and cell lines. The resource showed cell type-dependent polyadenylation map, helped the comparison of polyadenylation patterns among various tissues and cell lines, and facilitated the discovery of novel genes and gene isoforms that might potentially important to tumorigenesis. The resource is publicly accessible through a web portal, which not only provides direct comparison of the pattern of polyadenylation sites across various tissue/cell types, but also enables users to search, download and analyze the data of their interest.

- Build polyadenylation map from sequencing data of various human tissues and cell lines;
- Develop a web portal of polyadenylation map for easy access and direct comparison;
- Analyze and discover differential polyadenylation patterns of non-coding and novel genes.

1.1.3 Characterizing a novel group of short RNAs that associates with YB-1

We found the multifunctional nucleic acid binding protein, Y-box binding protein 1 (YB-1), is associated with two distinct but related classes of RNAs in a pilot study. Our goal is to characterize the novel group of RNAs and explore their biogenesis and regulation mechanisms. We investigated the relationship between these RNAs with Illumina and Helicos sequencing techniques. We also researched the relationship of these RNAs with annotated genomic marks to identify the mechanisms that control their biogenesis.

- Analyze the relationship between two novel classes of YB-1 associated RNAs;
- Investigate the relationship between the YB-1 associated RNAs and various annotated genomic marks, including histone marks, transcription factor binding sites and small RNAs.

1.2 BACKGROUND AND SIGNIFICANCE

1.2.1 Gene expression analysis using next-generation sequencing

Gene expression is the fundamental process of cell, by which information flow from genome to corresponding functional product. Genetic information stored in genome is expressed through multiple steps, including transcription and RNA processing; and if the product is protein, more steps like translation and post-translational modification are required. All these products work as structural and functional components of cell, or serve as regulation factors to give control over different machineries of cell. Therefore, measuring gene expression level genome-wide is of great value to understand the various mechanisms in cells, including gene regulation and cancer development.

Before next-generation sequencing techniques, gene expression profiling is first practically done by microarrays. As the most commonly used high-throughput screening method a decade ago, microarrays enable the monitoring of thousands of genes at a time and have provided a huge amount of information about gene expression in human that are unknown before. However, this technique has two disadvantages. One is that the range of detected signal is limited to several orders from the minimum to the maximum; the other is that the locations on the genome that it can study are restricted to known genes. The next-generation sequencing technique overcomes these shortcomings in that it can detect the gene expression signal by hit counts that are ranged from 0 to the theoretically unlimited; and it can detect transcripts anywhere from the whole genome. Therefore, the next-generation sequencing techniques offer the opportunity of revealing comprehensive landscape of transcripts, including splicing variants, identifying novel transcripts as well as quantifying their expression levels in a given sample. Although it increases the challenge of data analysis and rebuilding transcript models, the next-generation sequencing techniques have been able to expand gene expression profiling to genome-wide expression analysis, from annotated gene regions to the whole genome, including previously underestimated non-coding RNAs.

1.2.2 Non-coding RNAs

It is clear that a large portion of the human genome is transcribed. However, we have not yet addressed a key question. Are the transcribed sequences functional or are they merely transcriptional noise such as random products and degradation-resistant RNAs?

When we look back the history, it must be noted that when the first miRNA, *lin-4*, was discovered more than a decade ago, it was generally believed that such small RNAs are not wide-spread. It took approximately another ten years to realize that miRNAs are a large class of

functionally important non-coding RNAs (ncRNAs) and several of them are strongly involved in cancer pathways[2,3,4]. Several reports unequivocally demonstrate that a very large number of ncRNAs have important roles in mammals. Also, a large fraction of the reported transcripts generally have polyA tails, have experimentally verified transcription factor binding sites near their genomic loci, and appear to be spliced[5,6]. Several of the reported transcripts were verified by northern-blot analysis or real time RT-PCR, thereby revealing their relatively high abundance. Although some of these transcripts may have protein-coding regions that were not previously annotated, many do not appear to have any significant coding potential[7,8,9]. Moreover, the emerging role of large numbers of miRNAs in genome-wide regulation of protein-coding genes, their tissue specific expression patterns, and their roles in diverse pathways are excellent examples for the wide-spread expression and influence of ncRNAs.

In light of realization that many functional ncRNAs have been identified in a variety of species, it is reasonable to postulate that the human genome encodes a large number of ncRNAs involved in numerous processes. The continuing discoveries of ncRNAs using NGS strategies are beginning to unravel many surprises in molecular biology. Based on multiple independent observations that indicate that the extent of the genome transcription is much greater than previously thought, we must no longer ignore the possibility that ncRNAs and other genomic products that are yet uncharacterized may help unravel the central events that lead to cancer.

1.2.3 Technology challenges in next-generation sequencing

With the fast evolving in recent years and many advantages it bears, NGS becomes the method of choice nowadays in genome and transcriptome analysis. Meanwhile, many big projects, for example The 1000 Genomes Project[10] and Genome 10K[11], have been launched to study a

comprehensive overview of genomes and aim to understand various mechanisms of cells and life complexity. These large-scale projects produce terabytes of sequencing data daily for storage and distributing to clients and collaborators. Moreover, the cost of NGS dropped in recent years at a pace far outstripping Moore's law[12], which describes the pace of the growth of computing power. These trends boost the data accumulation speed dramatically and introduce a huge pressure and challenge on how to efficiently store and manage these data with limited budget and share them with scientists all over the world within a reasonable time.

1.2.4 Basic algorithm of data compression

Given these trends and challenges, compression of NGS data seems to be a natural direction to pursue. Currently, the most commonly used compression strategy of these raw sequencing datasets is applying the general-purpose compressors, for example, gzip and bzip2. These methods have the advantages of ease of use and broad applicability. The compression ratio, however, is far from optimal as the specific properties of sequencing data are not fully utilized. Thus, novel compression methods that are optimized for sequencing data need to be explored and applied to solve the data storage and sharing challenge in the field.

There are two main approaches in data compression. One approach, including dictionary coders and entropy compressors, is to build a statistical model to represent data efficiently by exploiting redundancy. For example, Lempel-Ziv (LZ) compression methods, one category of dictionary coders, compress data by replacing repeat occurrences of data with references to a dictionary that is built on based on input data stream. These coders form the basis of many ubiquitous compression schemes, including the algorithms used in gzip, GIF and PNG. The entropy compressors, including Huffman and arithmetic coders, compress data by encoding higher

frequent patterns with shorter bit strings and low frequent patterns with longer bit strings. The utilization of these coders is universal, including the back-end algorithm for JPEG, MP3 and many others.

The other data compression approach is data differentiation, which produces a difference given a source and a target as compression, and then represents the target by patching the source with the difference produced as decompression. Delta encoding, which records the relative change between sequential data instead of the complete files, roots in this idea. These methods are widely used in data synchronization and version control. Despite the conceptual difference between these approaches, applications of compression algorithms generally mix up multiple methods, hoping for fully utilizing their advantages under different circumstances.

2.0 COMPRESSING SEQUENCING DATA WITH ASSEMBLY-BASED ALGORITHM

2.1 INTRODUCTION

2.1.1 Sequencing Data Accumulation

In recent years, the fast development next-generation sequencing technology (NGS) has made this technology vastly and successfully used in multiple biological research fields, such as genome and transcriptome analysis, Copy-number Variantion (CNV), Single-nucleotide polymorphism (SNP), alternative splicing detection and many more. With the progress in the sequencing techniques, more researchers migrate from traditional microarray technology, which generates megabytes of data per experiment, to NGS technology, which produces gigabytes, if not more, of data per experiment, e.g. there would be ~100 GB result data in a Illumina HiSeq 2000 instrument run. When researchers enjoy the advantage of much more detail information of genome and cells, they have to suffer the vast increase of data. Currently, the accumulation speed of NGS data is much faster than the increase of computing power, which empirically obeys Moore's law[12]. Given the situation that researchers are optimistic and yearn to personalize medicine and have put a huge effort to reduce the cost of sequencing to $1000 per genome, they may first face the obstacle of how to effectively handling these huge and even fast increasing data.

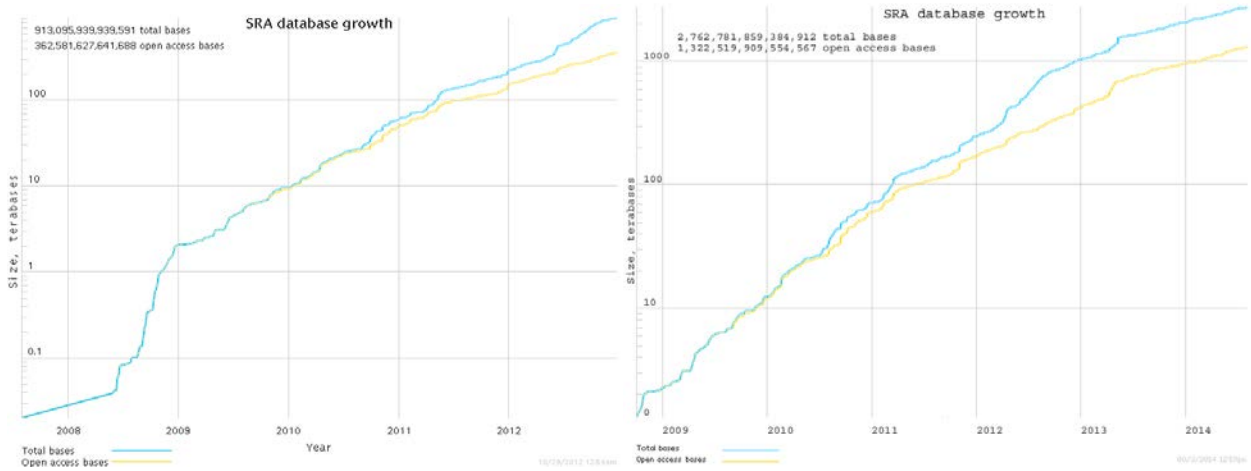


Figure 1. Sequencing Data Accumulation. The increase of data accumulation in the Short Read Archive (SRA) shows an exponential growth over the past decade. The capacity of SRA database has grown 100k times in the past 5-6 years, from 0.02 terabases in the middle of 2008 to now more than 2.7k terabases in July 2014 .

The first realistic problem is data storage. With data accumulating, extra computing infrastructure and man power has to be added and maintained. Further data analysis may require much more capacity to store the intermediate and final results. Another problem with huge volume of data is data sharing and transferring. Researchers tend to utilize data from colleagues to support, supplement or verify their own experiments. However, facing the huge NGS data, one may find it is a nightmare to transfer data back and forth through network. The rapid growth of the data gives the whole community great pressure, both for scientists and engineers, on store, analyze and interpret these huge data. For a centralized database which provides NGS experiment data storage, like SRA, the demand to solve these two problems seems more urgent. In 2011, NCBI announced a date for terminate SRA service due to the shortage of funding, which remind the community of a potential problem for maintaining the public data effectively and efficiently. It cannot be imagined that scientists cannot find the raw and processed sequencing data of peers from the public sector, which is essential resource for them to peer review or cross-validate other's results. Although, NCBI cancelled the decision later on, and has continued providing the archive for public access in

the following years, the situation still needs enough attention and effort from all related staffs from biological science, bioinformatics and computer science, until it can be solved.

2.1.2 Possible Solutions

There are three ways to solve these problems. One is discarding the sequencing data but keep the original sample of the experiment. It is, however, unrealistic in most cases. Since NGS is still not cheap enough to redo multiple times now, it is a waste of funding to discard the result. Also, data sharing among researchers is impossible after data discarding. Keeping the sample consistent over time is also a challenge. Another way to deal with huge NGS data is reducing the size or format of the original data, for example, decreasing the level of quality data. However, it may affect the signal to noise ratio of experiment and need further research to verify the feasibility. The third way is compressing the raw NGS data. The reduction of data set size would allow more data being stored and less data transfer lapse. This approach is currently the most feasible method to solve data storage and transfer problem of NGS data sets.

2.1.3 Redundancy in the sequencing data set

The compression of nucleotides sequences has been studied in the past decades. The compression of NGS data sets, however, is a relevant but different topic. The relevance between these topics roots in the fact that they both target compressing nucleotides sequences, which are comprised of four kinds of nucleotides, which are A, C, G and T, and probably an extra wildcard, N, which stands for missing or unidentified nucleotide. The difference comes from the distribution and characteristic of data. Compression of nucleotide sequences targets a small number of long

sequences, while NGS data compression targets a large amount of, usually millions of, short reads. Also, the NGS data compression involves handling accessory information of each read, such as read identifier and quality data.

To compress a dataset, the vital thing for considerration is to explore the special properties of the subject of the research and choose effective and efficient algorithms that are designed to pinpoint and utilize that property to perform for a high compression ratio. For example, in the compression of images, a straight forward presentation of image is matrix of pixels, in which each pixel stores various properties of that spot, such as grayscale, color channels and so on. Theoretically, images can be any random matrix of pixels; but in reality, there are always some patterns for a meaningful picture: e.g. pixels are locally correlated in most cases, meaning that the adjacent pixels share similar properties, including but not limit to color and grayscale, which is an important property that can be utilized by image compression algorithms. The correlationship indicates information redundancy in the straightforward storage format, which gives the space for compression to come in to store information in another format to gain high efficiency in storage by sacrisfing native storage format.

In sequencing datasets, the native format for gene sequences is a set of strings of A/C/G/T characters, reperesenting corresponding nucleotides. These sequences of nucleotides in cell do not come from a purely uniform distribution. And when they store, deliver and express the information within genome, the expression level could be vary in several magnitudes, which makes the presents of each nucleotides and short reads quite different. Thus, some sequences which have just more occurrences or are easier to be transcribed than the others in the genome would results in a large amount of repeat in the data set, the fact of which leads to an imbalance of these occurrence numbers in experiment result and could be a vital feature that can be utilized in data compression. Other than that, in next-generation sequencing datasets, all short reads come from a single or

several genomes with only subtle variation here and there. And these millions of reads overlap with other reads by part or whole to increase the coverage and fidelity of the sequencing experiment, which inevitably lead to the fact that a large of fraction of the information stored is redundant.

2.1.4 Available Frameworks and Tools

To handle short reads in NGS data sets, several methods have been put forward in recent years. The most straight forward and widely-used method is taking advantage of existing general purpose compressors, e.g. gzip or bzip2, directly. This approach bears several advantages, including ease of use and wide adaptability, since any data set can be compressed with these tools regardless of the format or completeness. The main disadvantage, however, is the compression efficiency, since these methods do not make use of any specific NGS data property, such as the short reads are derived from longer nucleotide sequences. Thus, many NGS data oriented tools have been designed in recent several years to answer the quest.

Reference-based framework, which based on the assumption that most short reads generated in the NGS data set comes from a set of longer sequences, which can be called as reference sequences, is the most developed method to handle exponential growing NGS data sets nowadays. In reference-based framework, all short reads from NGS data set are first aligned to a specific reference, e.g. GRCh37/hg19 when the NGS data set is generated from human sample, and only the locations and differences of hit on the reference are recorded, while the original read sequences are discarded; then the information of aligned reads and unalignable read sequences are further encoded with statistical compression method. In the aligning part, the sequences information is converted into position information, which is supposed to be more concise and easier to compress under statistical method. In the process, the reference genome works as a

dictionary in Lempel-Ziv compression method, and save the space to record the shared information between data set and reference.

A bunch of tools have been developed under this framework in the last several years. One of these tools is CRAMtool[13], which is a set of Java tools and APIs for efficient compression of sequence read data sets that developed by EMBL-EBI. It follows the principle and framework that described in [13], and is developed with the objective of achieving better loseless compression than BAM, which is a compressed binary file format of SAM to store aligned nucleotides sequences. While the main target is to keep archive lossless, CRAMtool also provide the options of lossy compression strategies of data set to enable users to make choice of whch data should be preserved in the archives. Currently, CRAMtool is utilized by SRA as a comparative compression method to reduce the size of some sequencing data sets compatible with the method.

Besides, there are several other compression tools that are designed for a close related problem that is, archiving genome, one of which is Genome Differential Compressor (GDC)[14]. The algorithm behind GDC is to compress multiple genomic sequences that comes from same or similar species by choosing one of the sequences as a reference and encode the variance of the other sequences to the reference sequence with a carefully designed LZ-77 scheme. There are several other tools, such as Genome Re-Sequencing(GRS)[15], Genome Re-sequencing Encoding (GReEN)[16], , GenomeZip[17], that adopt similar philosophy to squeeze data sets. While bearing the advantage of high compression ratio and fast compression speed, these tool has a limited usage scope for modern large scale sequencing data sets, as the reads in these sequencing data are short ones, not as long as genomes used by the tool making choose one data set as reference problematic; and the difference among short reads data sets is much higher than long genomic sequences making a great loss of the compression ratio and efficiency.

Recently, other than reference-based framework, several statistical compression methods that tailor to NGS data have also been proposed, such as DSRC[18] and G-SQZ[19]. These tools divide the FASTQ files and compress them independently with LZ-based or Huffman coding approaches. Although they can only handle FASTQ files, these methods show a better compression ratio (×1.1-1.5)[18,19] than the general-purpose compressors. Another tool, Quip[20], which is based on statistical model using arithmetic coding, can handle both FASTQ and SAM/BAM format and is able to achieve an even better compression ratio (×1.5-2)[20].

2.1.5 Advantages and disadvantages

The advantage of reference framework is extremely obvious in re-sequencing data set, since this type of data set comes from a defined set of reference sequences, e.g. human or *Arabidopsis thaliana* genome, and the reference genomes are widely available and verified. The high quality of reference genome increase the chance of hit when we try to align short reads back to these references. The disadvantage of the framework lies in the fact that it is highly reliance on reference quality and availability. For species that do not have high quality annotated genome, the method may fail because of a low percentage of reads can be aligned to the reference genome, which heavily affect the compression ratio of the scheme. Moreover, many NGS data sets are sequenced with a biological sample consisted of different species, which may not be known or identified beforehand, e.g. the data sets of Human Microbiome Project[21]. Under such circumstance, no reference is available, which make the reference-based compression scheme not applicable. Another issue to be considered is that the versions of references being used in various archives need careful management. With more researches and experiments performed, the annotated genomes that are used as references may be updated so frequently that either older versions of

references need to be maintained in publicly database or the compressed archives would be updated for each new version of references, which would be a huge hassle or waste of computing power.

Moreover, with many curated genome sequences available in public databases, the methods of differential encoding become feasible, including the algorithm[13] on which the developing CRAM tool is based. These methods are generally reference-based and implemented with two stages. First, reads are mapped to various locations on publicly available genome sequences. Afterwards, these locations, as well as the mismatches between reads and reference genome, are recorded and further compressed. The philosophy behind this method is the same as delta encoding. An extreme example of this application is DNAzip[22], with which James Watson's genome was compressed to a file of size ~4 MB, with the aid of curated human genome[22]. The strength of the reference-based method is that the reference genome can be stored in publicly accessible database, instead of each compressed file, to save space. The weak points, however, are that the compressed files are not self-contained, making decompression complicate; and the method suffers when a close related reference is not readily available for the data set.

Besides, with complete genome sequencing has been available and affordable in recent years, genomic data increase rapidly as well, which leads a close related problem as sequencing data compression. The difference between genomic data archiving and sequencing data compression lies in the property of data, that is, in sequencing data compression, data is always short reads generated by shot-gun sequencing strategy, while the targets in genomic data archiving are different samples sequenced from same genome, e.g. different individuals or different experimental conditions. Thus, the facts that the variation among genomes from multiple samples could be limited and the genomic data are long, assembled genome make reference-based framework a natural choice to approach the problem.

To summarize, these methods bear one or more advantages of the following: 1) compressing subfield of FASTQ file separately; 2) utilizing reference; 3) independence of available genomes and self-certainness. The idea of utilizing all these properties leads to our study of assembly-based compression tool.

In this dissertation, we put forward a novel compression method for NGS data. This method utilizes an assembly-based framework, which is based on non-parametric philosophy, not only to avoid the limitation of heavy reliance on reference availability and quality, but also get rid of the hassle of version control of references.

2.2 METHODS

Our ultimate aim is to solve the challenge of sequencing data explosion in recent years. To start with, we proposed to develop a computational method, Assembly-based Sequencing Encoding Tool (AbSEnT), to compress large NGS data sets, which are stored as FASTQ format, into files that are smaller than the result of compression tools which are commonly used in the field nowadays.

2.2.1 Framework of the genome-independent compression method AbSEnT

To fully utilize the different statistical properties of the three fields, including read identifiers, read sequences and read quality in the FASTQ format, we encode them separately with different strategies. First, the read identifiers are encoded with dictionary coders. The read identifier field is usually composed of several subfields, storing experiment information like instrument names, run ids, flowcell ids and so on. These subfields are identical for each read entry in most case. Thus, to

reduce the required space for storage of read identifiers, we compile the identifiers into patterns that capture the structure of identifiers, and encode the variable subfields with arithmetic and delta coding algorithm.

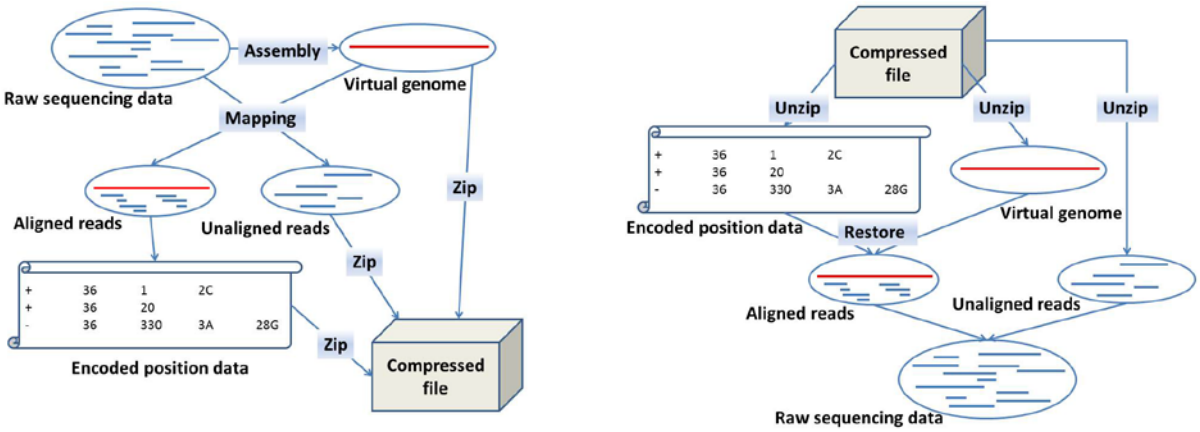


Figure 2: Compression and decompression schema for read sequences data. A) The raw sequences are first assembled to a virtual genome serving as reference using de Bruijn graph. All raw reads are mapped to the virtual genome. The aligned reads are further encoded as relative information to the reference, such as strand, read length, position on reference and difference between read and reference. The virtual genome, unaligned reads and relative information to the reference are then encoded with entropy compression models into a single file. B) The decompression schema is basically the reverse of compression.

Then, the read sequences are encoded by assembly-based method (**Figure 2A**). The assembly-based method first assembles short reads into a reference assembly with de Bruijn assembly method. And then data differencing strategy is used, that is all reads will be mapped to the reference assembly. For the aligned reads, only positions on the reference and mismatches between reads and reference are recorded. These differences, the reference assembly as well as the unaligned reads are passed over to arithmetic coder separately for further compression. In this assembly-based method, the raw reads are expected to be closer to the reference assembly than to the curated genomes from public database, given the fact that the assembly is directly generated

from the specific data set. We found the assembly-based compression method is both effective and efficient on out test NGS datasets.

The compression of read quality data is notorious since the pattern of the quality data are more random and hence harder to find than that of read sequences or identifiers. Thus, many tools are trying to use lossy compression on quality data. However, to keep compressed file faithfully to the raw data, we chose to compress quality data lossless. We used run-length algorithm and delta encoding to take advantage of local similarity property of quality data. Meanwhile, we investigated the positional entropy of quality data for possible improvement of compression ratio. Afterwards, arithmetic coder is used to achieve better efficiency.

As the encoder is a lossless compressor and the encoded file is self-contained, the decompressor is coded to extract the encoded file and restore the original data file faithfully. The read labels are recovered by substitute words back into patterns, which are constructed in compression, and form original identifier entries. The read sequences are decoded by the stored positions on the reference assembly and the difference to the assembly using reverse strategy (**Figure 2B**). The read quality values can also be restored straightforwardly by the information and models built in the compression process.

2.2.2 de-Brujin graph build-up

First of all, to convert sequences of nucleotides into a de-brujin graph representation, the sequence data was interpreted as overlapped k-mer with k = 32 nts and the adjacent 32-mer would be connected with a directed edge from former 32-mer to the latter one. Thus, take a sequence with length of 100 nts for example, it would be converted to a chain of 69 32-mer, and millions of such sequences would be converted to a de-brujin graph with millions of nodes and even more edges.

Hash map was used to represent the de-brujin graph in the lower level data structure in which each node stores the information of one unique 32-mer. In the node, the count of each unique 32-mer was recorded by incrementing 1 when each time it occurred, and the number would be used as the weight for each node in the following network pruning step. Besides, the weights of neigbour nodes were also stored for fast access. Each node is bound with the unique 32-mer where the sequence of 32-mer serves as key and the node struct serves as value in the key-value paradigm of hash map data structure. The connections between nodes are implicitly represented by shifting keys, which means trimming the first nucleotide of current key and appending any one nucleotide of A, C, G or T would make a key for the next node.

2.2.3 Simplification and pruning of de-Brujin graph

When a de-Brujin graph is built from a genome scale data set, the network would always be huge and hard to process and interpret, since the vast number of nodes, each of which stands for a unique fix-length word that appears in the data set, and edges, which stands for the relationship between these unique words. With such complex network in hand, how can we get useful information from there, say a reasonable assembly of short reads? Thus, it is crucial to develop a method to simplify the network and rule out unnecessary information from it. And the simpler the network, the easier and longer the genome can be read out, which increases the confidence of the assembled genome.

A large set of simplification techniques can be applied to the de-brujin graph. The goals of them are always pruning out the irrelated reads and keep the network as simple as possible. Here in our research, there is an addition requirement, which is mapping as many reads as possible to the assembly, which is read directly from the network, so that more reads can be compressed in the encoded way. Thus we could not just prune nodes and edges out just by the structure of the

network, but also consider the weight of each node and edge, making the highly weighted nodes and edges survive in the pruning.

The overall process is a cycle consisted of two steps, the pruning step and the contraction step. In the pruning step, different methods can be applied to the network to prune out unnecessary or less confindent information given the current network structure or weights of nodes, while in contraction step, the pruned network has to be simplified and restructured for the next step. The whole process stops until there is no node or edge can be pruned out or changed from the network. It is pretty like the classic process in EM algorithm.

To complete the above process, another hash map was introduced in our data structure, which stores the contracted network of de-Brujin graph. In this data structure, each node would not represent one unique 32-mer and store the related information from it, but a long string that is contracted from a chain of nodes in de-Brujin graph and its information. The first 32 nucleotides would be used as the key of this node, and the value includes the string it stores, the total number of occurrences of all 32-mers it includes, pointers to the neighbor nodes and their counts of occurrences, the average coverage of this string from nodes in the previous de-Brujin graph, and additional information for data analysis and graph traversal. Again, the information of edge was stored implicitly within the two nodes it connected. We call the contracted network as dense graph.

2.2.3.1 Branch zipper and Bubble zipper

Under some cases, there would be some similar sequences with minor difference either from seqeuncing noise or biological mutation in the cell. These sequences would be stored in differnet nodes in the network. To make the assembled genome concise, we would like to combine these nodes together to remove the local similarity among sequences and save a version with more weight. Foutunately, because of the way de-Brujin graph and the dense map is built, these nodes

would be close in the network structure: either share the same ancestor or same decedant, which could be simplified with zipper algorithms.

For branches in the network, we would first find the shared ancestor or decedant of the two branches and set it as an end-point. Then, we would compared the two branches to check whether they are similar by a simple alignment algorithm. If they are similar, we would merge the branch with lower coverage to the other, keep the merged branch and remove the other afterwards. The whole process is metaphored as zipper, because of the similarity.

For bubbles in the network, there would be two end-points to detect and merge the pathes between these two end-points if they are similar with each other. The whole process is implemented with breadth-first search and basic seqeunce alignment algorithm. Similar to branch zipper, bubble zipper is, however, more complicate, since the bubbles could be tangled together as knit, forming a complicated thread to zip.

2.2.3.2 Low importantance nodes filter

Besides structure simplification, we pruned out low importance nodes each cycle when we got a simplified dense graph. Which are low importance nodes? Here we judged by the averge coverage index, which is calculated by dividing the weights of the nodes by difference of the length of the string that the nodes represent minus 31. We could imagine a node with low coverage could have a high possibility been generated from background noise. And even the sequence was with real meaning, e.g. from low frequent RNA products, it would not benefit compression, since the repeatence is so low. We could see that lots of sequence located in this part in the result section. The cutoff could be set to arbitrary value, as each data set could be different, but the sweet point is where the balance between over-pruning and under-pruning. Over-pruning would break the structure of network, making the assembled genome short and fragmented, while under-pruning

would make the read-out of network hard, as the structure of the network was not reasonably simplified.

After the whole process converges, the whole dense graph would be read out by breaking down in the ambiguous nodes, which is defined as the crossring of two paths in the graph, so that the graph can be converted to a set of chains of nodes. The strings stored in each chain would be concatenated together and saved as one assembled chromesome in the assembled genome.

2.2.4 Mapping and encoding of the Sequences

After the assembly is created, we mapped all reads from the original sequencing data set to the assembly with bowtie. The mapping result file is processed by a python script to reorganize the mapping information. For example, if a read can be fully mapped to the genome, then the mapped chromosome, the position on the chromosome would be recorded. If the length of sequence in the data set is not same for all, the length of the sequence would also be saved after python processing, otherwise, a global value of sequence length would be set in the result. When a read can be mapped to the assembly genome with minor difference, the difference would also be recorded, such as the position of mismatch, the kind of mismatch, like mismatch, insertion and deletion, and the portion of subsequence of the mismatch, so that the read could recovered in the decoding process. When a read cannot be mapped to the assembly genome, they would be saved to a separate file. The three files, one that saves difference information of mappable reads, one that records the unmappable reads, and one that includes the assembled genome, would be processed by statistical encoding tool to a binary compressed package in the following step, which can be stored, shared and delivered with efficiency.

2.2.5 Decoding of the sequences

The decoding process is basically the reverse step of encoding. Since the whole process is lossless, all information is contained in the compressed binary package. The package is first decoded by statistical tool, and then the three files it contains would be utilized to recover all raw reads from the original sequencing data set.

2.3 RESULTS

2.3.1 Redundancy of information in sequencing dataset

To illustrate the redundancy in sequencing datasets, we investigated the distribution of frequency of unique subsequence (word) with 32-base long. We divided long sequence of reads into overlapped 32-base words, that is, two adjacent words overlap by 31 bases; and further, we calculated the frequency of each unique 32-base word. Then we binned all words by the occurrences in the dataset and counted the total occurrences of each bin. We further plotted the data in Figure 3, in which the bins are plotted as x-axis and the total occurrences of each bin as y-axis in log-log scale.

In the log-log plot, we can find that the distribution asymptotically converges to a straight line, which may indicates the distribution follows the power law distribution. The graphical illustration of the power law distribution can be straightforward by plotting a log-log plot, but a strict validation of whether a distribution follows power law is difficult and we did not perform this test right now. Although we canno not validate the power law property, this graph reminds us Zipf's law, which is a widely-used empirical law in linguistics analysis, stating that given a large

and structured set of uninterrupted text, the frequency of any word is inversely proportional to its rank in the frequency table. It is also closely related to discrete power law probability distribution. In the case of unique fix-length words in sequencing data, the relationship between the occurrence of each unique word in a dataset and the number of unique words with same occurrence also follows an inversely proportion relationship and might follows the power law.

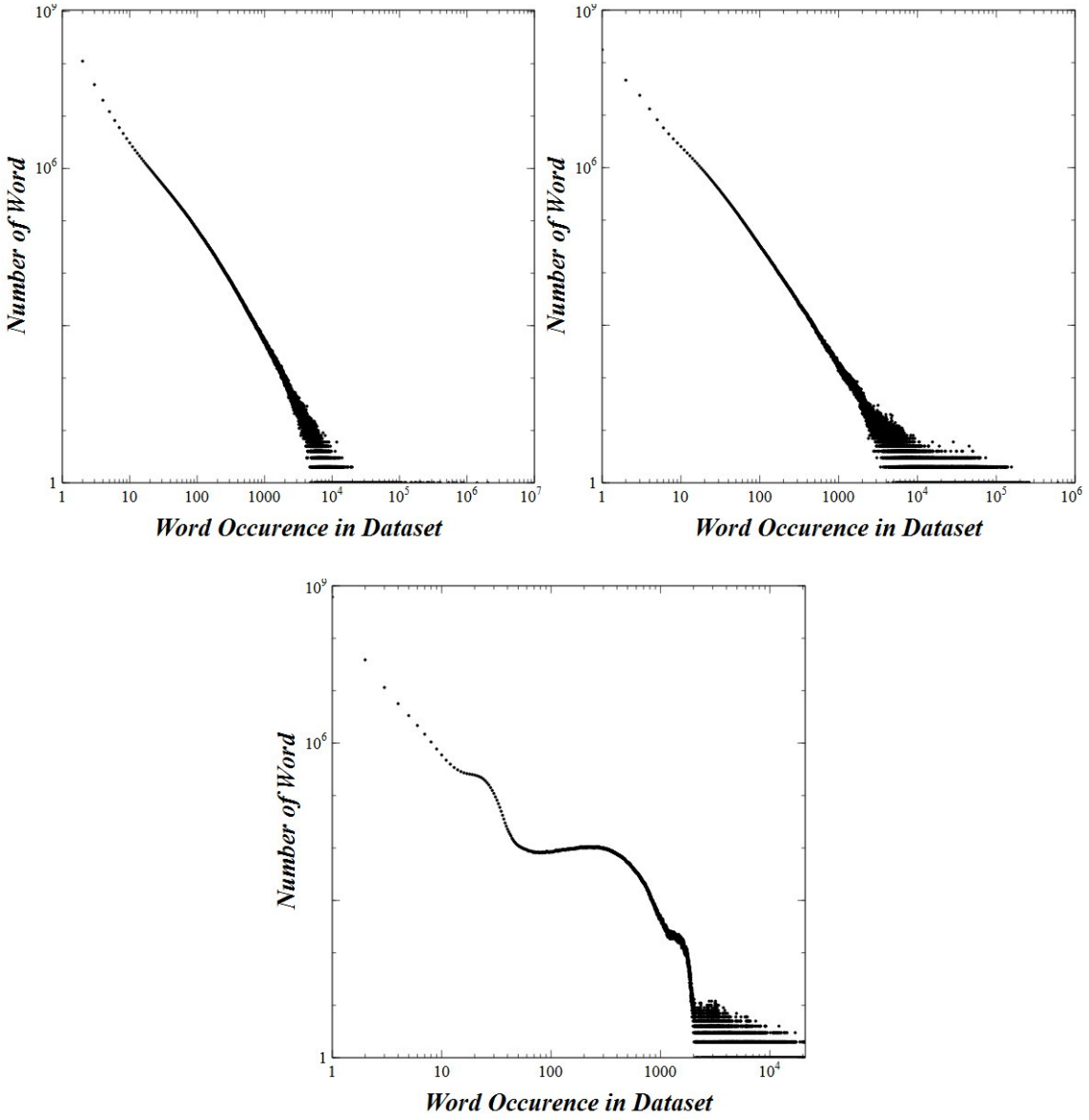


Figure 3. The distribution of word frequency in sequencing data sets. A-B) Dataset ERR030867 and Dataset ERR030894, respectively. In the log-log plot, a straight line declining pattern might indicate a power law distribution of word frequency. C) Dataset SRR359032. A wavy curve in the log-log plot does not indicate a power law

distribution. But if it varies around the presumed straight line, the answer might be different.

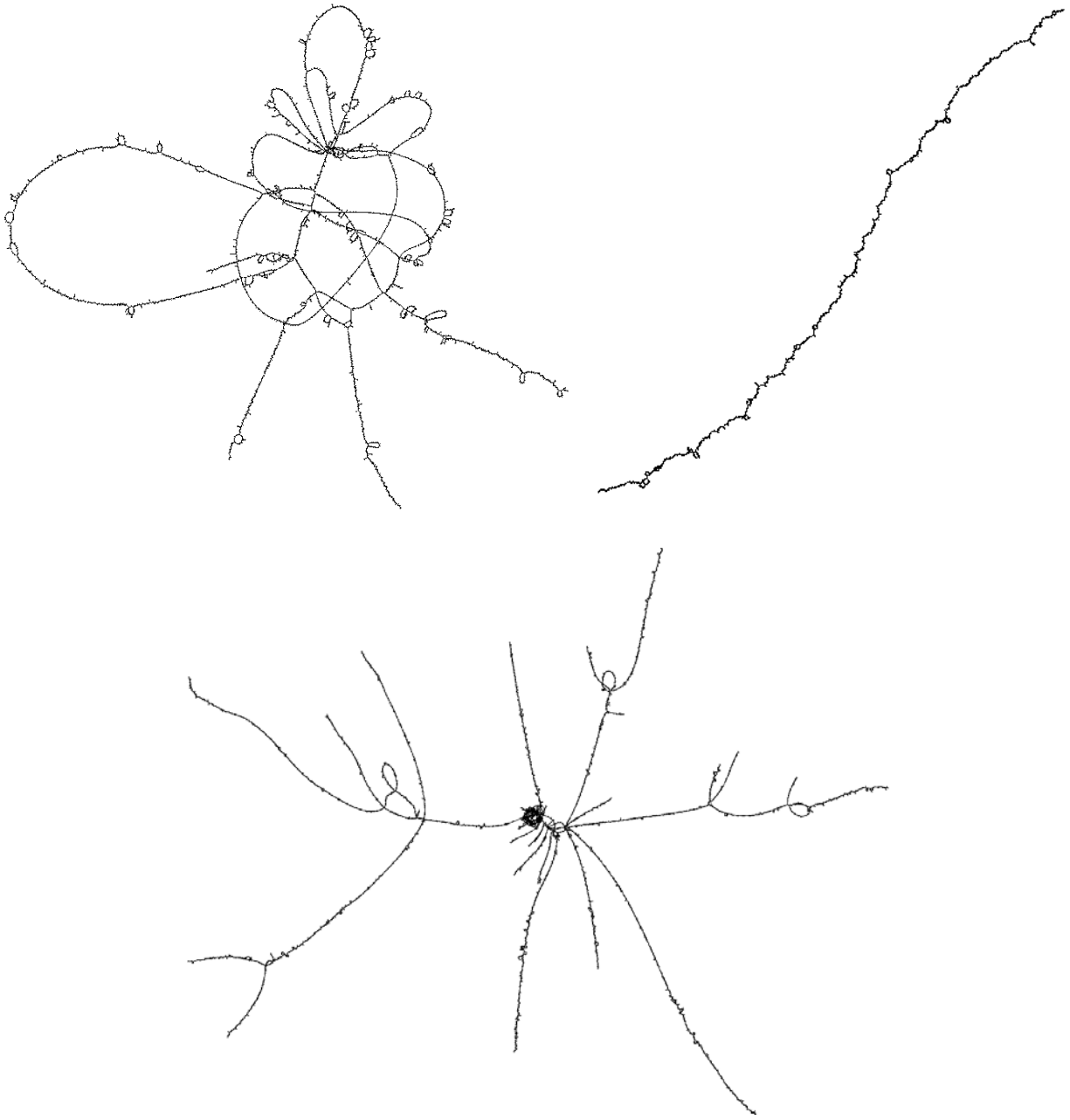


Figure 4. Galley of dense graphs after pruning and simplication. Top-left graph is one sample subgraph after graph simplication. There are some loops and bubbles in the subgraph which could not be removed in the pruning step. The subgraph could also be as simple as a linear or circular chromosome, like the top right subgraph; or it could still be hard to read out, e.g. the center part of the bottom subgraph.

2.3.2 de-Brujin Graph pruning and simplification

After several rounds of pruning and simplification of de-Brujin graph and dense map through various filters, such as lower importance filter and different zippers, the network can be simplified from unreadable complex network to hundres of realative simpler and smaller subgraphs (**Figure 4**). Each subgraph would be read out by breaking the graph into paths and each path with enough coverage and length is treated as a chromosome in the assembled genome.

2.3.3 Compression efficiency of AbSEnT

To show the efficiency of assembly-based compression method, we used a whole genome sequencing dataset of *Pseudomonas Syringae* as an example. To build a reference, we first assembled a genome from all raw reads in the data set with de Bruijn graph method that mentioned in previous section. The assembly of all reads is of length 14,924,049 nucleotides. Out of all 3,551,133 raw reads, ~93.5% of them can be mapped back to the reference assembly, indicating the high quality of the assembly. To test the compression efficiency of our method, we extracted all read sequences, which is of size 131,391,921 bytes, from the original data. With general-purpose compressor bzip2, the compressed file is of size 37,616,023 bytes, the compression ratio of which is 2.290 bits/nt. With our assembly-based method, the compressed file size is 10,143,007 bytes, which is less than one third of bzip2 file and of compression ratio of 0.618 bits/nt. More in detail, the compressed file is comprised of three parts, which are encoded assembled genome, mapped reads and unmapped reads. In this particular dataset, the encoded reference assembly is of size 3,993,063 bytes; the size of encoded mapped reads is 3,081,001 bytes; the remaining ~6.5% unmapped reads takes 3,068,943 bytes which is about one third of total size of the compressed file.

These statistics reveal us that mapped reads can be encoded much more efficiently than unmapped reads.

Table 1. Bzip2 and AbSEnT compression statistics on *Pseudomonas Syringae* whole genome sequencing data.

	Bzip2	AbSEnT
Raw data	131,391,921 bytes	131,391,921 bytes
Virtual genome	-	14,924,049 bytes
Compressed virtual genome	-	3,993,063 bytes
Percentage of mapped reads	-	93.53%
Compressed mapped reads	-	3,081,001 bytes
Compressed unmapped reads	-	3,068,943 bytes
Compressed data	37,616,023 bytes	10,143,007 bytes
Compression ratio	28.6%	7.7%

2.4 DISCUSSION

2.4.1 Evaluation of compression algorithm

To evaluate a compression algorithm, several aspects should be considered. The first and the obvious one is the compression ratio, which is the purpose and goal of compression algorithm. However, the compression ratio is not the only criteria that matters. To make a compression algorithm feasible, one important factor to be considered is the speed of both compression and decompression. The scalability is also one merit that a good compression algorithm should have. Both of them might have a direct impact on the utilization of the algorithms.

The bottleneck of assembly-based compression method is the efficiency of de-novo assembly. As we do not need a curated-level genome, the efficiency can be largely increased by slightly losing the accuracy of the reference assembly, as long as it accommodates a high portion of all raw reads. Moreover, a statistical de-Bruijin method can also be incorporated into the tool to improve the efficiency and scalability of the assembly algorithm.

The evaluation data set should be comprised with multiple next-generation sequencing data sets from different application of NGS, including whole genome sequencing, exome sequencing, ChIP-seq, RNA-seq, and metagenomic DNA sequencing. The performance of AbSEnT as well as other available tools, including gzip, bzip2, 7z, DSRC, G-SQZ and Quip, should be evaluated on all categories in terms of compression ratio and speed to illustrate the feasibility and efficiency of AbSEnT and make comparison with other tools.

2.4.2 Lossless compression and lossy compression

In addition to the lossless compression we mentioned above, we might also perform data compression in a lossy way. Again, let us take image compression for example. People has a detect limitation when an image is shown as our eyes can hardly discriminize subtle difference of small pixels. Therefore, a small variation of the color or grayscale of a small number of pixels might just merely change our interpretation of the image at most, but would benefit a lot for compression algorithms, since the outliers always take a large of mount of information in the compressed dataset. Also, for each image, we concentrate on the subject we are interested in or the theme that was designed to deliver by the image but overlook the background information there. Thus, in some lossy image compression algorithm, the valuable space in bits would be allocated more to the main

subject of the image or the region of interest and less for less informative region, so that a higher compression ratio can be achieved without losing too much information and quality of the image.

Similar to the image compression, bit allocation concept could also be utilized in the compression of sequencing data. One of the advantages of sequencing data compression comparing to image compression lies in the fact that there is a quality field in the sequencing data, which naturally confines the region of interest. In image compression, algorithms designers have to design a smart method that the algorithm can find the region of interest by it self, which is not always handy. In contrast, although one might argue that the regions that constrained by sequence quality data might not be the regions of research interest, the quality field gives researchers the region where the sequencing data can be trusted or not, which provides researchers an easy option to allocate more bits to the trusted region instead of a uniform allocation. In this way, the information of the sequencing experiment would mostly be reserved and it can be squeezed with high efficiency, as the data of untrusted region are more likely to be unrelated or unstructured noise and thus harder to be compressed, which seems to be a balanced tradeoff between storage space and scientific truth.

It is interesting to discuss the tradeoff between the lost of information and the gain of space saving in lossy compression. Is it worth to loss the valuable information that we get from the expensive experiment but gain some more, e.g. 20%, space saving in the compressed data? No concensus answer can be achieved among scientists easily until lots of contrast experiments are performed and many arguments are considered. However, insprired by the image compression, we found that not all information of the raw data, no matter it is images or sequencing data, would be utilized in the downstream analysis. Specifically speaking of sequencing data, the sequences with low quality would probably be filtered out by preprocessing procedure for quality control, making these data not necessary to be stored in the compressed data set. Then, a new question comes up: what is the best cutoff of quality to store the meaningful data? The answer is still unsettled and

need time for scientists to determine. But in fact, the cutoff really depends on the property of the data set and what the data set is designed for. For example, if the data set is designed to discover some rare sequences in the cell, the cutoff might be set to a lower one so that to keep these rare seuqnces from filtering out because of random noise. In terms of compression, we might set multiple cutoffs for the quality and compress the data set separately for different purposes.

2.5 SUMMARY

We release an assembly-based sequencing compression tool, AbSEnT, that contains compressor and the corresponding decompressor to process FASTQ files. The compressed files generated by this tool are not only lossless and self-contained to free the storage and manpower needed to maintain a set of reference genomes, but also with a higher compression ratio than general-purpose compression tools. The tool efficiently and effectively lowers the demand of storage of massive sequencing datasets and reduces the time for sharing the datasets among collaborating labs and groups. With these merits, we expect this tool would be broadly useful for the biological scientists, especially who work on next-generation sequencing intensively. Moreover, we discover the distribution of words frequency in sequencing data set shares a similar pattern with books and social network. We expect to improve our understanding of these statistical properties of different sequencing data sets in the course of pursuing higher compression ratio. We have also discussed the evaluation of compression method from various aspects. We expect to improve the efficiency of AbSEnT and optimize the speed of encoding and decoding.

3.0 BUILDING COMPREHENSIVE POLYADENYLATION MAP OF HUMAN GENOME

3.1 INTRODUCTION

We present a comprehensive map of over 1 million polyadenylation sites and quantify their usage in major cancers and tumor cell lines using direct RNA sequencing. We built the Expression and Polyadenylation Database to enable the visualization of the polyadenylation maps in various cancers and to facilitate the discovery of novel genes and gene isoforms that are potentially important to tumorigenesis. Analyses of polyadenylation sites indicate that a large fraction (~30%) of mRNAs contain alternative polyadenylation sites in their 3′ untranslated regions, independent of the cell type. The shortest 3′ untranslated region isoforms are preferentially upregulated in cancer tissues, genome-wide. Candidate targets of alternative polyadenylation-mediated upregulation of short isoforms include POLR2K, and signaling cascades of cell–cell and cell–extracellular matrix contact, particularly involving regulators of Rho GTPases. Polyadenylation maps also helped to improve 3′ untranslated region annotations and identify candidate regulatory marks such as sequence motifs, H3K36Me3 and Pabpc1 that are isoform dependent and occur in a position-specific manner. Moreover, thousands of novel polyadenylated sites were found both in 3' UTR and intergenic regions, out of which a large number (~1000) of sites show a consistent up/down-regulation in majority of tissues tested. In summary, these results highlight the need to go beyond

monitoring only the cumulative transcript levels for a gene, to separately analysing the expression of its RNA isoforms.

3.1.1 Alternative polyadenylation

Alterations in 3′ untranslated regions (UTRs) that drive post-transcriptional control of gene expression is a common theme in multiple diseases including cancer[23,24,25]. Recent reports also provide surprising and compelling evidence that in proliferating/cancer cells, genes often switch their expression toward the short 3′ UTR isoforms that correspond to truncated versions of the canonical long isoforms[26,27]. However, 3′ UTR isoforms are poorly annotated in databases[28] because common high-throughput technologies such as microarrays, RNA-Seq and quantitative reverse transcriptase-polymerase chain reaction cannot readily distinguish between 3′ UTR isoforms. This limitation arises because the long 3′ UTR isoforms encompass the short isoforms, and therefore the short isoform-specific probes/primers hybridize to both short and long isoforms, leading to inseparable, mixed signals[27,29]. Such limitations in precisely identifying and quantifying 3′ UTR variants can lead to erroneous conclusions in not only gene expression studies but also in investigations of posttranscriptional events[26,27]. New approaches such as Direct RNA sequencing or DRS[30], 3P-Seq[28], PAS-Seq[31] and others[32,33] are beginning to be adopted to identify 3′ UTR variants, allowing the construction of a near-complete map of 3′ UTR variants of the human genome[31,34].

Polyadenylation can generate many alternative transcripts of a given gene, with important consequences. Alternative polyadenylation (APA) can lead to truncated protein isoforms, abrogate protein-coding capacity, affect transcript stability, alter translation efficiency and affect transcript export[35]. Important oncogenes such as p53 and CCND1 are known to have altered polyadenylation

that results from simple 3′ UTR mutations[36,37]. A related emerging theme in cancer biology is that APA within the same 3′ UTR can enhance production of oncogenes[27] because shorter 3′ UTR isoforms have higher translational efficiency than their respective long isoforms. Although APA within the same 3′ UTR simply results in mRNA isoforms that code for identical proteins, it can alter miRNA targeting regions[27], subcellular localizations and stability[38], and protein production rate[35]. Thus, tandem APA seems to provide an alternative mechanism to complement more subtle posttranscriptional regulatory modes such as miRNAs that can cause either translational arrest[39] or destabilize target RNAs[40,41] in human cells[42]. The existence of a mechanism that preferentially alters the relative expression ratio of long and short isoforms is also observed in embryonic cells[31,43]. The diversity of transcripts produced via polyadenylation and its consequences underscore the need to accurately catalog and study polyadenylation in both normal and diseased cells.

3.1.2 Genome-wide Polyadenylation map

We sought to build a comprehensive polyadenylation map of the human genome across major cancers and their cognate normal tissues, which could also facilitate studies on 3′ UTR isoforms, and enable discoveries of important novel gene isoforms. To reduce artifacts associated with common deep sequencing strategies, we used DRS to construct the polyadenylation landscape and measure gene expression directly without manipulating purified RNA (308). We built a public resource for tissue-specific polyadenylation sites termed Expression and PolyAdenylation Database (xPAD), which represents a comprehensive analysis and discovery platform for the investigation of more than 1 million polyadenylation sites, 3′ UTRs and their cellular usage, across five major organs, their cognate tumor samples, and six different cell types including commonly used

ENCODE[44] cell types (Hela-S3, HepG2, K562 and MCF7). Approximately 9000 polyadenylation sites are well expressed (≥10 reads) and are at least 1000 nucleotides (nts) away from annotated gene regions, indicating the presence of many novel polyadenylated gene isoforms and genes. DRS reads also reveal potential polyadenylation in many non-coding RNAs including the lncRNA[45], GAS5 that encodes 10 snoRNAs, which are all unexpectedly polyadenylated. The simultaneous determination of gene isoforms and their levels also opens up additional opportunities in exploring the data, such as using network and motif analysis for studying gene isoforms. Indeed, network analysis of genes that manifest APA suggest that cell–cell and cell–extracellular matrix (ECM) contact mediated signaling cascades are potentially key targets of APA in cancer, particularly involving regulators of Rho family small GTPases that are consistently targeted by these mechanisms across the samples tested. The nucleotide level resolution of the polyadenylation maps also enabled the discovery of novel sequence motifs and regulatory marks (e.g. H3K36Me3) that are highly position specific with respect to polyadenylation sites and are isoform dependent. In summary, the polyadenylation site maps and its usage reported here reveal an extensive landscape of both known and novel polyadenylated genes, their isoforms, and their regulation.

3.2 MATERIALS AND METHODS

3.2.1 DRS sequencing and genome mapping

Matched pair (normal/tumor) total RNA from tissues was used for DRS[30,34]; all paired RNA samples for sequencing were purchased from Biochain (Hayward, CA), except for Breast (Asterand, MI). Raw DRS reads were quality filtered using HeliSphere

(http://open.helicosbio.com) and in-house tools to remove low-quality sequences and reads shorter than 25 nts. The filtered reads were mapped to the human genome (GRCh37) with maximum allowed error of 10% (mismatches and indels) by MOSAIK[46] with conservative parameters (-mmp 0.1 -mhp 100 -act 20 -hs 15 -p 8 -bw 13). Although internal priming in DRS data is negligible, we implemented filters for quality control, following a more stringent criterion than the recommended removal of sequences with a stretch of eight consecutive adenosines[47]. For quality control, uniquely mapped reads were further filtered to remove all sequences that genomically contained either six consecutive Adenosines or at least seven Adenosines within 10 nts downstream of the end of the mapped DRS reads.

The mapped locations were annotated using the UCSC genome browser tables[48]. When a locus could be attributed to multiple possible annotations, the locus was assigned with a single annotation in the following priority order: 3′ UTRs (sense), coding sequences (CDS, sense), 5′ UTRs (sense), intron (sense), non-coding RNAs (ncRNAs, sense), 5′ UTR antisense, CDS antisense, 3′ UTR antisense, intron antisense, promoter antisense, ncRNA antisense and intergenic. Unlike other regions, intergenic regions are not separated into sense and antisense strands. Promoter regions were defined as regions 1000 nts upstream of transcription start sites provided by UCSC genome browser tables. DRS reads mapping to regions that are at least 5000 nts away from the closest gene region (sense/antisense) were annotated as intergenic.

Because polyadenylation sites vary by a few nucleotides in a given isoform, we used the snow-ball method to define polyadenylation sites[49]. The snow-ball method iteratively clusters genomic locations that are located within 24 nts, using the genomically mapped 5′ positions of the DRS reads as a reference. For analysis of genes containing tandem polyadenylation sites, polyadenylation sites with less than 10 total reads in the combined set of normal and tumor reads, were removed. The proximal polyadenylation site was defined as the first polyadenylation site after

the end of the CDS on the 3′ UTR. The distal polyadenylation site was defined as the last polyadenylation site after the end of CDS on the 3′ UTR. If a gene contains more than two tandem polyadenylation sites on the 3′ UTR, only the first and last sites are used for analysis[27]. To account for variations in sample loading, the number of reads from each polyadenylation site was standardized to contain an equal number of reads across all tissues, using the normal breast tissue reads as a reference. To enable more accurate comparisons, the resulting expression levels were further quantile-normalized (R package) between normal and tumor within each tissue. To analyse whether quantile normalizations of the samples are justified, Quantile–quantile plots of raw read numbers (log2 scale) between tumor and normal samples were analysed, which yielded high correlation ($R = 0.98-0.99$) along the diagonal. The distance between adjacent polyadenylation sites were defined as the distance between their 5′ most cleavage sites. To determine variances in polyadenylation site locations, the mean location of each polyadenylation cluster (≥10 reads) was determined. For a given cluster, polyadenylation sites corresponding to every read was used to determine the mean location of the polyadenylation site. For each polyadenylation site cluster, the distance between the polyadenylation site of each read and the mean location was used to calculate variance.

3.2.2 Sequencing data preprocessing with in-house tool

In sequencing data analysis, data preprocessing is important step in the whole pipeline, since in large sequencing data sets, there are a lot kinds of noises coming from various sources, which include but not limited to sequencing machine's system errors, noise from samples which may come from different labs with a more or less different experiment conditions and imperfections of various sequencing techniques. These noises could lead to serious results in the statistical analysis

and might further impact on the interpretation of the data, leading. Also date sets of next generation sequencing are high-throughput which is prone to be sensitive to background noise, resulting in a relative low ratio of signal and noise (SNR). Thus, a careful noise filtering before the analysis routine is crucial for the result of each successful experiment.

In the analysis of polyadenylation sites in various cell lines and tissues, we utilized Helicos next generation sequencing techniques for sites detection. Due to the fact that Helicos's technology has a relative higher error rate compared to other mainstream NGS techniques, we would put more attention on data preprocessing to filter out reads from background noise. To help retain only high quality reads, in addition to Helisphere software, we used in-house tools to remove additional reads that occur due to imperfections in surface that results in reads that resemble the sequence of sequencing reactions, resulting in perfect or imperfect repeats of AGCT. Although Helisphere software already eliminates such reads, we have observed using other data sets that the use of the method outlined below can be helpful. For each read, we determined the extent of divergence from the repeating sequence of AGCT; Kullback–Leibler (KL) divergence measure (MATLAB) was used based on the two different frequency distributions (***P***,***Q***) of tri-nucleotides of the read (***P***) and that of the repeat sequence of AGCT [$\boldsymbol{Q}(\boldsymbol{i})$ =0.25 for $\boldsymbol{i}$ = AGC/GCT/CTA/TAG; $\boldsymbol{Q}(\boldsymbol{i})$ =0 otherwise] KL score was defined as $\sum_i \boldsymbol{P}(\boldsymbol{i}).\, log\, \frac{\boldsymbol{P}(\boldsymbol{i})}{\boldsymbol{Q}(\boldsymbol{i})}$, where $\boldsymbol{i}$ represents one of the 64 possible tri-nucleotides. In this setting, a biased read with high AGCT repeat would have a lower KL score, while an unbiased read from polyadenylation site would generally have higher KL score. Only reads above a KL divergence threshold of 40 was considered for further filtering; threshold was determined by tests on randomly generated sequence datasets, which on average retained ~90% of reads at the threshold of 40 (**Figure 5**). The reads were then further filtered using genome mapping, as described afterwards.

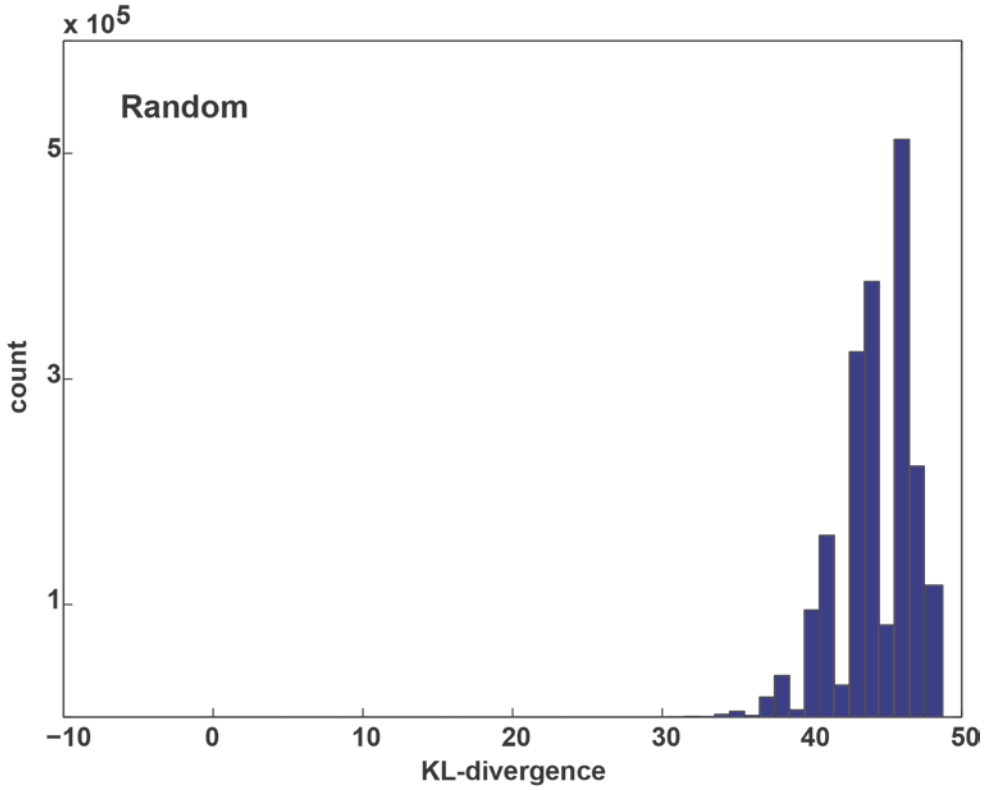


Figure 5: The histogram of KL-divergence of a set of random generated sequences. The KL-divergence between each random sequence and the repeat sequence of AGCT is calculated individually and then the distribution of this statistic is plotted as a histogram. We can see from the histogram, ~90% of the sequences have a KL-divergence higher than 40, which is used as a threshold for the high quality reads filter.

3.2.3 Determination of usage of short and long 3' UTR isoforms

For genes with more than one tandem polyadenylation site, Short and Long represents the total number of reads (quantile normalized) from the proximal and distal polyadenylation sites, respectively. To compare the relative usage of the distal polyadenylation site between normal and tumor tissues, we used the previously reported RUD indices (RUD = log2[Long/Short]; ΔRUD= $RUD_{tumor} - RUD_{normal}$). RUD is used in the strict sense of its definition; manipulations to emulate mixed signals from short isoforms in parent studies that used RUD were avoided for sake of simplicity, particularly since our conclusions were consistent with previous studies. For example,

$\Delta RUD \geq 1$, corresponds to at least two-fold increase in Long/Short ratio in tumor, and hence represents those genes that favor the expression of their long isoforms in tumor. To determine differences in expression levels of either short (ΔShort) or long (ΔLong) isoform levels between normal and tumor tissues, we computed the log2-transformed fold-change ratio ($\Delta LogShort=log[Short_{tumor}/Short_{normal}]$; $\Delta LogLong=log[Long_{tumor}/Long_{normal}]$).

3.2.4 Motif-enrichment analysis

The total number of genes used for motif-enrichment analysis is identical for both short and long isoforms. Because some of the polyadenylation sites are tissue-dependent, motif analysis was constrained to short and long isoforms detected in a single tissue-type (normal breast). To ensure that the polyadenylation sites of short and long isoforms are well separated, we further constrained the analysis to those sites that are at least 100 nts apart, resulting in a total of 3270 genes. Motifs in regions flanking (±500 nts) the polyadenylation sites of the short and the long isoforms were detected based on the DREME[50] motif discovery algorithm, followed by a statistical assessment ($E < 0.05$) of the positional preference of the motifs, using CentriMo[51], a position-specific motif-enrichment analysis method. The analysis resulted in a total of 27 motifs, six of which were eliminated because they occur in a limited number of sequences, corresponding to less than 5% of the sequences at those distinguishing positions that are most enriched for the motifs. To focus on isoform-dependent motifs, the observed preference of each motif was further tested by the bootstrapping approach to determine statistical significance, and a final non-redundant set of eight motifs were selected using default parameters of TOMTOM[52].

3.2.5 Statistical tests

Both the Ansari–Bradley and F tests were used to test whether the variances in locations of polyadenylation-sites in short and long isoforms are statistically different (MATLAB). As F-tests yielded more statistically significant numbers, we report only the P values obtained using the most conservative test (one-tailed Ansari–Bradley). For testing the observed increase/decrease in tumor cells for the absolute or relative expression of short and long isoforms, the non-parametric one tailed Wilcoxon signed rank test was used. Bootstrapping analysis for motif detection was performed by randomly sampling (with replacement, $n = 3270$), the complete set of either short or long isoforms to yield 10 different datasets for each isoform. The resulting distribution of occurrences for each motif at their characteristic location in each of the 10 datasets were calculated for both short and long isoforms, and compared using two-tailed student t test.

3.3 RESULTS

Polyadenylation can generate many alternative transcripts with important consequences. Alternative polyadenylation (APA) can lead to truncated protein isoforms, annul protein-coding capacity, affect transcript stability, alter translation efficiency, and transcript export[35]. An emerging theme in cancer biology is that APA within the same 3' untranslated region (UTR) can enhance production of oncogenes[27] because shorter 3' UTR isoforms have higher translational efficiency than their respective long isoforms. Although APA within the same 3' UTR simply results in mRNA isoforms that code for identical proteins, it can alter miRNA targeting regions[27], subcellular localizations and stability[38], and protein production rate[35]. Thus, 3' UTR APA seem to provide an

alternative mechanism to complement more subtle post-transcriptional regulatory modes such as miRNAs that can cause either translational arrest[39] or destabilize target RNAs[40,41] in human cells[42]. The existence of a mechanism that preferentially alters the relative expression ratio of long and short isoforms is also observed in embryonic cells[31,43]. The diversity of transcripts produced via polyadenylation underscore the need to accurately catalog and study polyadenylation in both normal and diseased cells.

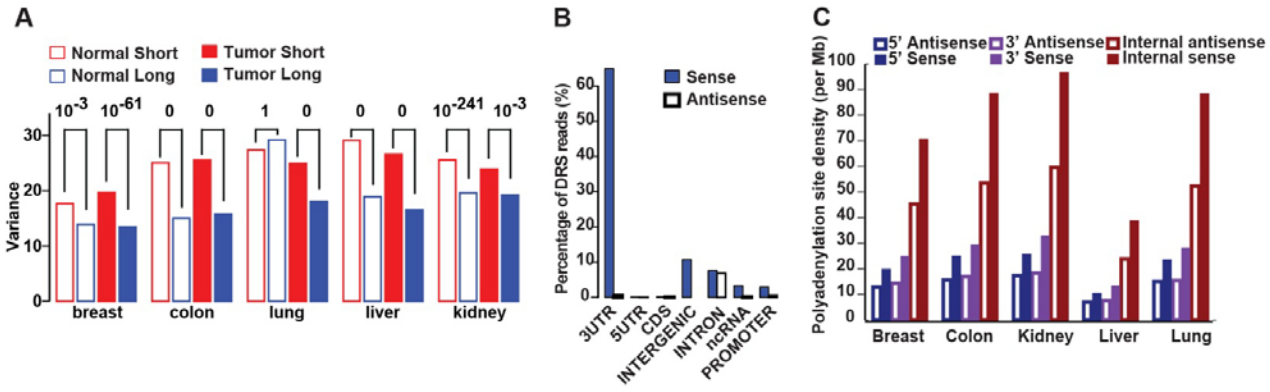


Figure 6: Characteristics of polyadenylation sites. DRS reads from normal breast tissue are used for this illustration. A) Cleavage sites of short isoforms in either normal (Normal Short) or tumor (Tumor Short), are generally (9/10) more variant than that of the corresponding long isoforms (p-values on top). B) In breast, the majority (~90%) of DRS reads match to sense strands of transcriptionally active regions and the remaining reads mainly map to intergenic regions and introns. For illustration purposes intergenic polyadenylation sites are assigned to the sense strand because categorizing them into sense and antisense strands separately can be ambiguous. C) Polyadenylation density in internal introns is higher than that of terminal (5' and 3') introns for both sense and antisense intronic transcripts.

3.3.1 Genomic feature of polyadenylation sites

Five matched pairs of tumor and normal tissues from human breast, colon, kidney, liver and lung and six commonly used cell lines, including Hela-s3, Hepg2, K562, LNCaP, MCF7 and PC3, were

used to obtain DRS[30,34] reads using the Helicos Sequencer. Both normal and tumor tissue samples yielded comparable number of reads. Moreover, the total number of genes that were expressed in a given cancer type (≥ 10 reads) and contained tandem polyadenylation sites were approximately the same across organs. Consistent with previous reports, we find that although polyadenylation sites frequently vary[53,54] the average absolute deviation based on all distinct locations within a cleavage site cluster is small (±2.7 nts), indicating that cleavage is highly precise. Intriguingly, perhaps due to differences in the factors that govern the cleavage of the short and long isoforms, short isoforms manifest a statistically significant higher variance for cleavage sites than long isoforms, across all tissues (**Figure 6A**).

In normal breast tissue, the majority (73-90%) of the reads mapped to the sense strands (**Figure 6B**) of functionally important regions (3'/5' UTR, coding sequence i.e. CDS, intron, non-coding RNA i.e. ncRNA, and promoter). The observation that sense transcripts constitute the major fraction (~70%) of polyadenylated total RNA is fully in agreement with other studies[55,56]. Notably, anti-sense transcripts are enriched in intronic regions across all 10 tissue types. Although most of these antisense transcripts are not abundant, ~17% of these transcripts occur in clusters separated by less than 2500 nts, suggestive of transcription and subsequent polyadenylation of antisense transcripts in intronic regions. Further analysis revealed that in comparison with terminal (5' or 3') introns, internal introns are significantly more enriched in both sense and antisense polyadenylated transcripts (**Figure 6C**). The enrichment is consistently stronger for intronic sense transcripts than antisense transcripts. Although the polyadenylation site density in antisense internal intronic regions is lower than that of the sense strand, it is always higher than that of terminal introns. The observed biases in intronic polyadenylation might be related to nucleosome depletion at internal introns[57] that may be more susceptible to aberrant intragenic transcription[58]. The database xPAD is

built on these data and can be accessed through the web portal (**Figure 10**).

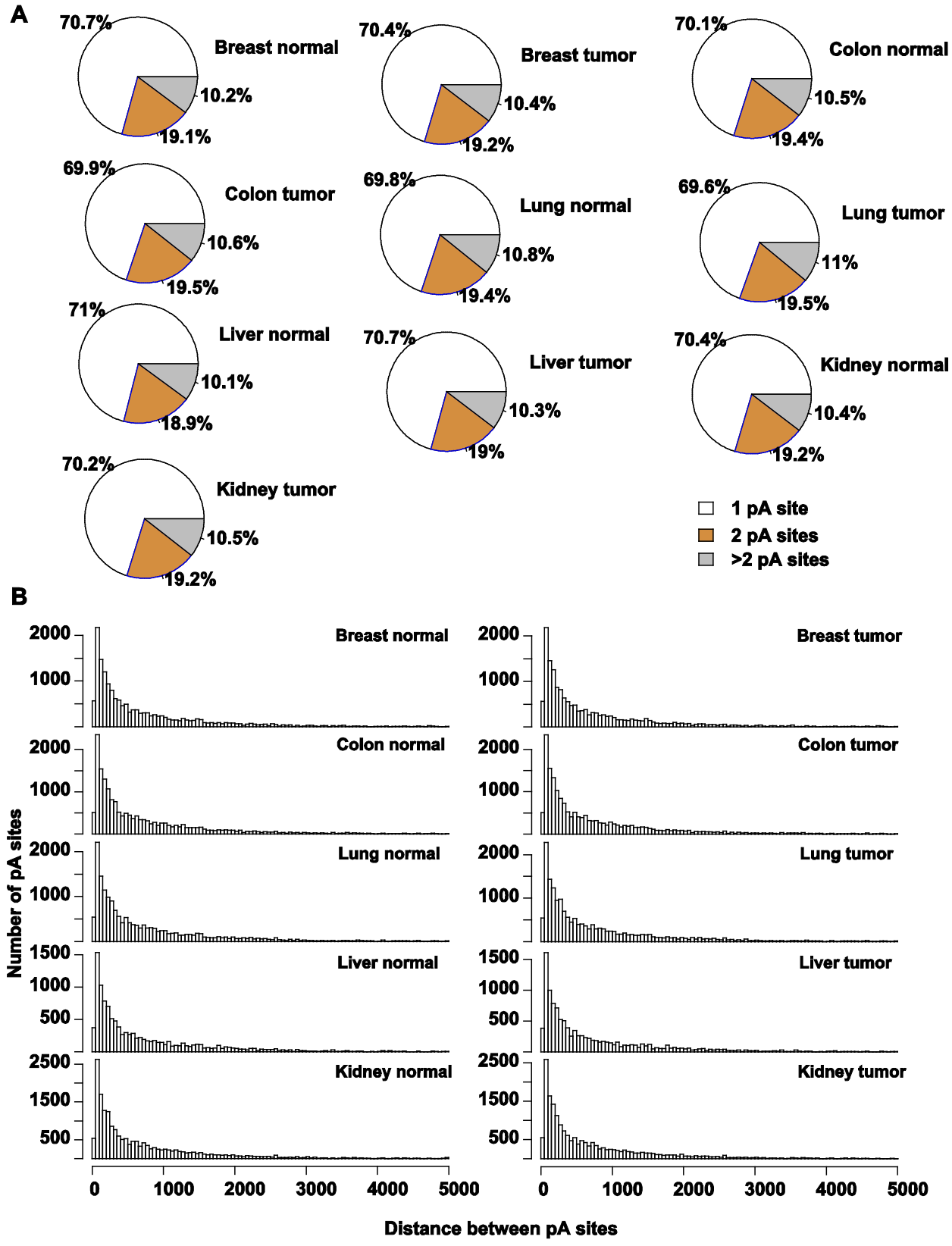


Figure 7: Overall properties of polyadenylation sites are consistent across tissues. A) Approximately 30% of genes contain tandem APA sites. B) Distribution of distances between adjacent tandem polyadenylation sites in the 3' UTR of genes expressed (bin size=50 nts). Adjacent pA sites that are separated by at least 5000 nts correspond to a small fraction (0.9-1.1%).

Because tandem APA-mediated gene regulation that occurs without a change in the cognate protein sequence 59) is an emerging theme in cancer progression[26,27,60,61], we investigated whether polyadenylated 3′ UTR isoforms are commonly produced. Approximately 30% of genes have multiple tandem polyadenylation sites in their 3′ UTRs in any given tissue (**Figure 7A**). Most of these genes have precisely two polyadenylation sites. The observed proportion is nearly identical to the previously reported percentage of ∼34%, determined using EST mapping[49]. Similarly, in accordance with the EST-mapping study, for those genes with multiple tandem polyadenylation sites, the adjacent polyadenylation sites peak ∼100 nts apart (**Figure 7B**) and have a median of 368 nts. Thus, the overall genomic characteristics of polyadenylation events captured by both DRS and ESTs are nearly identical.

3.3.2 Polyadenylation patterns of non-coding and novel genes

Non-coding RNAs such as miRNAs and lncRNAs[45] are frequently implicated in cancer progression[62] and in some cases are known to undergo polyadenylation[63,64]. We analyzed whether miRNA and lncRNA regions are polyadenylated and if so whether those polyadenylated forms are aberrantly expressed in cancer. A modest proportion (12.3%) of the 14,403 lncRNAs locations is polyadenylated within 100 nts of their annotated 3' ends. However, most of these polyadenylated locations are not abundant, as only a small fraction (1.7%; 244 loci) of the lncRNAs generates many (≥ 10) reads in at least one of the 10 tissue samples tested. Abundantly polyadenylated lncRNAs loci include the well-known breast cancer metastasis-associated HOTAIR[63] that is highly

(>12 fold) upregulated in breast tumor (49 versus. 4 reads). Although polyadenylated lncRNAs do not manifest a consistent pattern of differential expression across the majority of the tumors tested, the GAS5 and TMEM191A are two good examples of polyadenylated lncRNA regions that are aberrantly regulated in the majority of tumors tested (**Figure 8A and B**). The 4 kb long GAS5 that hosts 10 snoRNAs is frequently down-regulated gene in breast cancer[65]. In DRS results, GAS5 is also 3-fold down-regulated in breast tumor. Notably, although only 36% of snoRNAs are polyadenylated in at least one of the tissues, all GAS5 snoRNA loci are usually polyadenylated in multiple tissues, generally within ~5 nts of their annotated 3' ends.

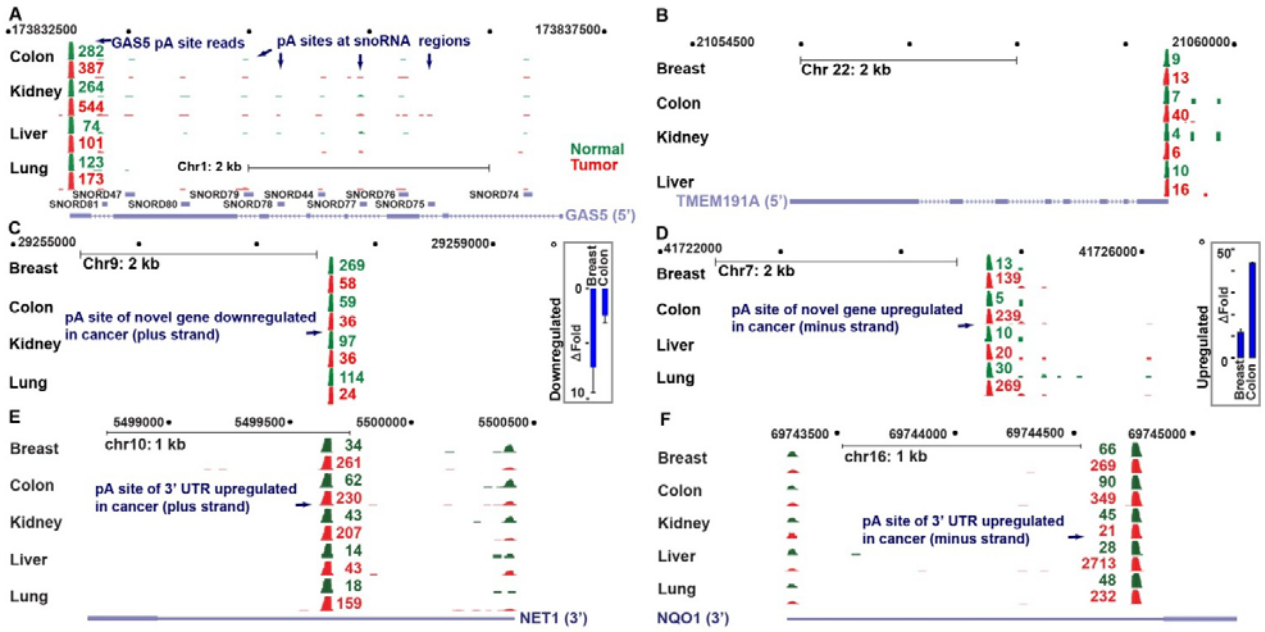


Figure 8: Genomic view of polyadenylated non-coding RNAs and novel gene locations that are aberrantly expressed in cancer using xPAD. A-D) All polyadenylation sites detected by DRS reads (green: normal; red: tumor) are indicated for all four gene regions. A-B) GAS5 (A) and TMEM191A (B) represent lncRNAs that are upregulated in the majority of tumor samples, as indicated. In contrast to the polycistronic GAS5 which hosts multiple polyadenylated snoRNAs, polyadenylation of TMEM191A is limited to its 3' UTR. C-D) End locations and the expression levels of

two potentially differentially regulated novel genes that are distantly located from known genes. Real-time PCR results also reveal similar expression patterns (fold change) in two additional, matched-pair patient samples ($p<0.001$); error bars represent standard deviation (n=3). E-F) End locations and the expression levels of two potentially differentially regulated novel polyadenylation sites that are located in 3' UTR.

We next investigated the presence of novel polyadenylated genes that may be important for cancer development or progression. To identify such novel RNA transcripts, we first probed for novel polyadenylation locations that are within annotated 3' UTR regions, are abundant ($\geq$ 10 reads) and are consistently either upregulated or downregulated in the majority of tissues. A total of 19275 polyadenylated sites were identified, of which 842 were upregulated and 94 were downregulated in the majority ($\geq$ 3) of the tumor samples at more than 2-fold levels, two of which are shown (**Figure 8E and F**). Further, we probed for novel polyadenylation locations that are at least 1 kb away from any known gene with the other criteria the same as the previous. A total of 9612 potentially novel polyadenylated genes were identified, of which 77 were upregulated and 41 were downregulated in the majority ($\geq$ 3) of the tumor samples at more than 2-fold levels. The DRS results on the differential expression of two of these locations were also confirmed by real-time reverse transcriptase-polymerase chain reaction (**Figure 8C and D**) which is experimented by Dr. Teresa Liu. These results underscore the notion that many additional RNAs and RNA-isoforms important for cancer and possibly for other diseases exist and that visualization tools for deep sequencing data could be useful in identifying such genes.

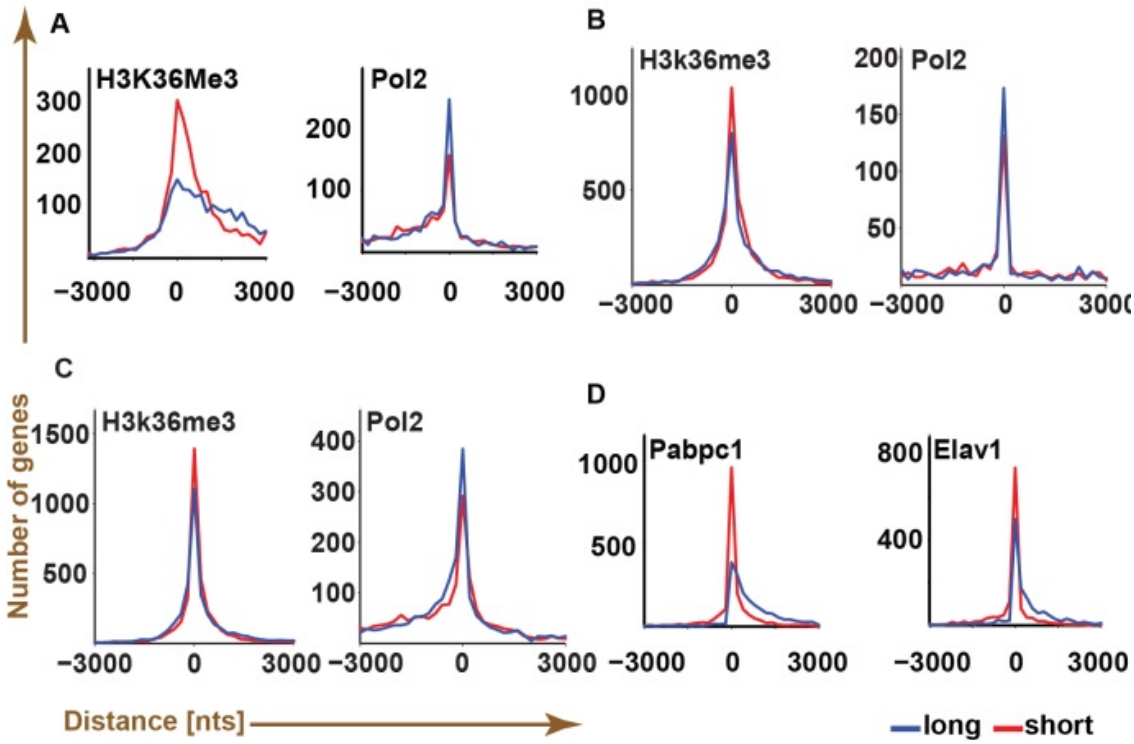


Figure 9: Polyadenylation maps enable the identification of isoform-dependent regulatory marks. A-D) A total of 3,270 genes containing both long and short forms that are genomically separated by at least 100 nts at their 3' ends are used for the analysis. Chip-Seq/RIP-Seq data comparisons of H3K36Me3 and Pabpc1 to their functional analogs (Pol2 and Elav1) in identical cell lines (B: HepG2, C: HelaS3, D, E: GM12878) suggest preferential marking of polyadenylation sites of short isoforms by H3K36Me3 and Pabpc1. The curves corresponds to the distance distribution between the location of a given polyadenylation site and the nearest regulatory mark, as inferred using Chip-Seq/RIP-Seq.

3.3.3 Polyadenylation sites contain isoform-dependent regulatory marks

We sought to identify proteins that may differentially regulate short and long isoforms. Based on a large number (708) of publically available ENCODE ChIP-Seq or RIP-Seq data sets of transcription factors, histones and RNA binding proteins, we found that the histone mark, H3K36Me3 is consistently more enriched at short isoform polyadenylation site locations than that

of long isoforms in all 42 datasets tested. In contrast, Pol2 occupancies at polyadenylation sites of long isoforms in the corresponding cell lines are higher than that of short isoforms (**Figure 9A, B, C**). ENCODE RIP-Seq data analysis revealed that polyA-binding protein 1, Pabpc1 (**Figure 9D**), could be another candidate that preferentially associates with polyadenylation sites of short-isoforms. To ensure that the observed preference is not due to the mRNA binding ability of Pabpc1, the Pabpc1 profile was compared with that of a related mRNA binding protein Elav1[80], performed using identical conditions. The comparisons reveal that the Pabpc1 is more preferentially enriched at short isoform locations, suggesting that Pabpc1 likely has a role in regulating the polyadenylation sites of short isoforms (**Figure 9D**). In conclusion, multiple regulatory marks that correspond to both transcriptional and post-transcriptional regulatory complexes are present at locations proximal to polyadenylation sites in an isoform-dependent manner, indicating the interplay between the proteins and APA.

3.4 DISCUSSION

The biogenesis of eukaryotic RNA transcripts is made possible by a myriad of protein complexes that act in concert, orchestrating key tasks such as transcription, splicing, capping, 3′ end cleavage and polyadenylation. Although gene regulation by transcription is a widely studied process, the effects of other terminal processes such as polyadenylation that regulate transcript stability, transport and expression of RNA transcripts are poorly understood. Therefore, to help with the investigations of polyadenylation locations, 3′ UTRs, their different isoforms and usage, we built comprehensive maps of polyadenylation sites and quantified the usage of each polyadenylated gene isoform. The complete polyadenylation landscape is made publicly available through our web-

portal, xPAD (**Figure 10**), which integrates the widely used UCSC genome browser[48] to enable detailed investigation of any genomic region by intuitive queries involving gene name, keywords or gene position.

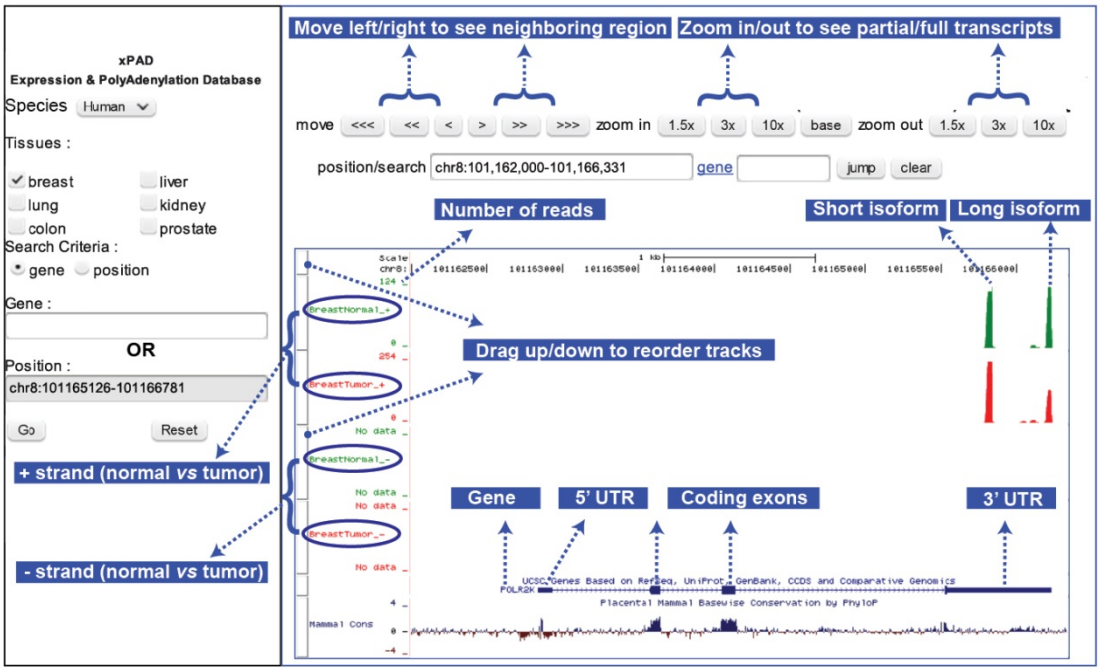


Figure 10: Illustration of xPAD. xPAD integrates the UCSC genome browser to provide a web-interface to visualize both the precise polyadenylation locations of different isoforms, as well as their expression levels across tissues of interest. For brevity, many additional features such as evolutionary conservation (bottom track) and methylation marks, which are available via the UCSC browser panel (right) are not illustrated.

xPAD integrates the UCSC genome browser to provide a web-interface to visualize both the precise polyadenylation locations of different isoforms, as well as their expression levels across tissues of interest. The complete gene structure of POLR2K highlights the utility of DRS; in both normal and tumor tissues, all polyadenylation sites exclusively occur in the 3′ UTR of the gene and within the sense strand, and the 3′ end of the reads (green/red bars) mapping to the long isoform matches within 2 nts of the 3′ UTR polyadenylation site. Normal breast contains 120 reads of long, and 109 reads of short isoforms, whereas breast tumor contains 257 reads of short isoform, which is upregulated, and 134 reads of long isoform that is almost unchanged. For brevity, many

additional features such as evolutionary conservation (bottom track) and methylation marks (not shown), which are available via the UCSC browser panel (right) are not illustrated.

The different polyadenylated transcript isoforms could be regulated the transcriptional[86,87] as well as posttranscriptional[82] levels, and the role of these two processes in polyadenylation may not be readily separable[88]. As most regulatory mechanisms that act in 3′ UTRs are thought to be repressive and not activating[27], it is likely that if the causative factor is a canonical UTR-regulation mechanism such as that of miRNAs, then it must be downregulated in cancer. An alternative explanation is that non-canonical, isoform-dependent, UTR regulatory factors that do not repress, but activate (e.g. polyadenylation/cleavage proteins) short isoform levels are upregulated in cancer. Identification of potential isoform-dependent regulatory roles for Pabpc1, H3K36Me3, and the ATATAT motif that is linked to the yeast heterogeneous nuclear ribonucleoproteins (hnRNP)-like protein Hrp1 highlights the possibility of such isoform-dependent factors. Hrp1 contains two tandem RNA Recognition Motifs domain, an arrangement shared by various human hnRNPs. Although the protein sequence of Hrp1 is similar to all human hnRNPs, the Hrp1 mRNA sequence is most similar to hnRNPA3. Several hnRNP proteins seem upregulated in tumor, including one of the hnRNPA3 isoforms that is upregulated in four of the five tumor samples. Similarly, across all five samples, Pabpc1 reads in tumor samples are higher than that of normal tissues; manifesting greater than 2-fold up-regulation in lung and kidney samples. Given the emerging theme that polyadenylation is interlinked to other key processes such as transcription elongation and posttranscriptional regulation[86,89], some of these regulatory marks such as H3K36Me3 may play a dual role in controlling both transcription and polyadenylation of the isoforms.

3.5 SUMMARY

In summary, we built a comprehensive polyadenylation and APA map, which is made publically available as a webserver (johnlab.org/xpad) and as a UCSC track hub (http://www.johnlab.org/xpad/Hub/UCSC.txt). We have highlighted potential uses of the resource in studying polyadenylation-mediated gene regulation and generated new insights into the regulation of polyadenylation with an emphasis on gene isoform-dependent signatures in cancer. Although these observations will need to be followed up with more detailed studies, xPAD serves as a unique tool to investigate various questions in biology, relating to polyadenylation, APA-mediated gene regulation, gene expression analysis, and discovery of new genes and gene isoforms. We aim to further expand xPAD with additional samples (tissue-types, cancer-subtypes, and cell lines), with the goal of defining gene-isoform signatures that may be useful for diagnosis or prognosis of specific human diseases, particularly in cancer[61]. These studies which provide cell-type specific annotation and usage of 3′ UTRs are expected to also help other genome-wide studies of UTRs such as miRNA target analysis[90,91], which could benefit by incorporating the expression levels of individual 3′ UTR isoforms.

4.0 CHARACTERIZING A NOVEL GROUP OF SHORT RNAS THAT ASSOCIATES WITH YB-1

The highly conserved, multifunctional YB-1 is a powerful breast cancer prognostic indicator. We report on a pervasive role of YB-1 in which it associates with thousands of novel non-polyadenylated, primarily intragenic, short RNAs termed shyRNAs that are further processed at their 5'/3' termini to novel small RNAs. The shyRNA genomic locations are preferentially marked within 500 bases by chromatin regulatory factors. shyRNA genomic locations manifest a strong preference for histone modifications associated with active transcription and open chromatin configurations and are bound by multiple transcription factors that precisely span the locations. The shyRNAs are also enriched in known regulatory RNAs such as snoRNAs. One 3'-modified, YB-1 associated RNA Shad1 (short RNA antisense to Dicer1) co-localizes with YB-1 under stress conditions and whose expression is closely linked with that of YB-1. We propose that the majority of shyRNA loci represent alternative transcripts that are regulated by multiple factors in a cell-type dependent manner and have evolved diverse cellular roles.

4.1 INTRODUCTION

Human Y-box binding protein 1 (YB-1) is a multifunctional protein that belongs to the highly conserved superfamily of cold shock domain (CSD) proteins known to bind both DNA and RNA[89].

It is involved in many cellular functions including transcription/translation[90,91], alternative splicing [92,93], and mRNA degradation/processing in P-bodies[91,93,94]. Inverse CAAT motifs, typically localized within promoter regions, serve as binding sites for YB-1. Occupancy of YB-1 at these sites can trigger either transcriptional activation through the recruitment of RNA polymerase II[95] or repression via displacement of bound activator proteins[96]. YB-1 has also been implicated as a protein partner in the splicing and transport of pre-mRNAs[97,98]. In the cytoplasm, YB-1 is also a component of the messenger ribonucleoprotein (mRNP) complexes formed by human Argonautes (AGO) which are central to small RNA-mediated silencing[99,100,101,102]. In addition to the link between YB-1 and mature microRNAs (miRNAs), YB-1's relatedness to lin-28, a pluripotency factor and a processor of the let-7 miRNA precursor[103,104,105] suggests that it may have a broader role in small RNA processing and function than previously recognized.

The role of YB-1 in human biology and disease has been most thoroughly examined in cancer. YB-1 expression levels have been shown to correlate with drug resistance and poor patient outcome in various cancers106[18]. Furthermore, YB-1 may be a more powerful prognostic marker for relapse and survival in breast cancer patients than the commonly used markers Her-2 and estrogen receptor[107]. Finally, YB-1 expression levels have also been shown to correlate with malignant transformation and castrate resistant prostate cancer[108]. Although the mechanistic basis for YB-1's role in cancer has not been revealed, it is a partner of the p53 tumor suppressor protein and may therefore influence its many functions in tumor progression[109].

A pilot project to predict RNA binding proteins important to small RNA pathways led us to test whether YB-1 associates with small RNAs in the androgen independent PC-3 prostate cancer cell line. In recent years, various classes of small and long non-coding RNAs have been identified in mammals[110,111,112]. In the most generalized form, genome-wide profiling of transcription factor binding sites and RNA polymerase II (RNAPII) occupancy have revealed a large number of

unexpected loci that are far away from annotated transcription start sites[113]. Studies in yeast have revealed that RNAPII can generate transcripts from sites quite far removed from gene-associated promoters with different degrees of stability; these are categorized as cryptic unstable transcripts (CUTs)[114] or stable uncharacterized transcripts[115,116]. The vast majority of CUTs and SUTs (95%) are predominately transcribed in the antisense orientation from bidirectional promoters and are frequently capped and polyadenylated. Similar RNAs that arise primarily from promoter/enhancer regions, termed TSSa-RNAs[117], eRNAs[118], and the polyadenylated miRNAs[119], exist in mammals. While the functions of these RNAs are largely unclear, other less characterized small/short RNAs, such as Y RNAs and vault RNAs are now recognized to have regulatory functions[120,121]. For example, Y RNAs (~100 nts) are highly conserved from bacteria to eukaryotes[120,122] and all four human Y RNAs are more highly expressed in tumors[123]. The bulk of these non-coding RNAs differ from mi/pi/siRNAs because they are thought to comprise diverse RNA pathways rather than conforming to a single molecular network.

We found that YB-1 associates with two distinct but related classes of RNAs in PC-3 cells. These RNAs correspond to thousands of distinct short YB1-associated RNAs (shyRNAs) and their cognate processed small RNAs that primarily match to the 5' or 3' terminal regions of shyRNAs. Many of the shyRNAs are also identical to known regulatory RNAs such as Y RNAs, vault RNAs and snoRNAs. The genomic locations of shyRNAs are bound by multiple transcription factors and polymerases similar to that of super-enhancers. And the very large number (~230,000) of shyRNA genomic locations are primarily intragenic and surpasses the number of protein-coding genes by an order of magnitude suggesting that their purpose likely extends beyond the generation of mRNA transcripts. One of the shyRNAs derive from the anti-sense strand of the Dicer1 gene, termed shad1 (short anti-sense to Dicer1), and found to regulate the proliferation of PC-3 cells.

4.2 METHODS

4.2.1 Genomic analysis

All sequence mapping and filtering were performed in house. Illumina FASTQ files were processed to retrieve sequences with Phred quality score to ensure a base call accuracy of 99%. The 3' adapter sequences and reads shorter than 15 nucleotides were removed using cutadapt. The filtered RNA reads were mapped to the human genome (GRCh37) using Bowtie. The mapped locations were then further filtered to only retain those reads that matched to a single unique location in the genome and were >1 RPKM. Neighboring (5' to 5' distance < 10 nts) reads from identical strands were considered as part of a single genomic cluster for annotations using UCSC Genome Browser and Bioconductor[154,155]. For the paired-end sequencing reads, the libraries were normalized to the sequenced input RNA. Input RNAs that were > 2 fold enriched over YB-1 RNAs was considered to be the control RNA sequences; YB-1 RNAs that were > 2 fold enriched over input RNAs was considered to be the YB-1 associated RNA sequences.

4.2.2 Statistical analysis

Publicly available (embargo free) ChIP-seq data (ENCODE) of histone modifications(72 cell lines, 682 samples) and transcription factors (88 cell lines, 218 transcription factors, 1284 samples) from the UCSC genome browser portal were used to establish the preference of various genomic features to shyRNA locations. As a control group, the input RNA locations with at least 2-fold enrichment over YB-1 and >1 RPKM (96,335 locations) were defined. An equivalent number based on RPKM of YB-1 associated RNAs were selected. Bootstrapping analysis for YB-1 to histone

modification/TF binding site enrichment was performed by randomly sampling (with replacement, n=96,335) the described sets of both YB-1 and control samples to yield 10 different datasets for each sample (YB-1 or control). The distances to the closest histone modification or TF binding sites for each 5' end of the YB-1 associated RNAs or its equivalent control 5' locations were calculated and binned (size = 200 nts). P-values were computed using the two tailed t-test based on the probability distribution of the YB-1 and control bootstrapped data sets at position 0 (+/- 100 nts).

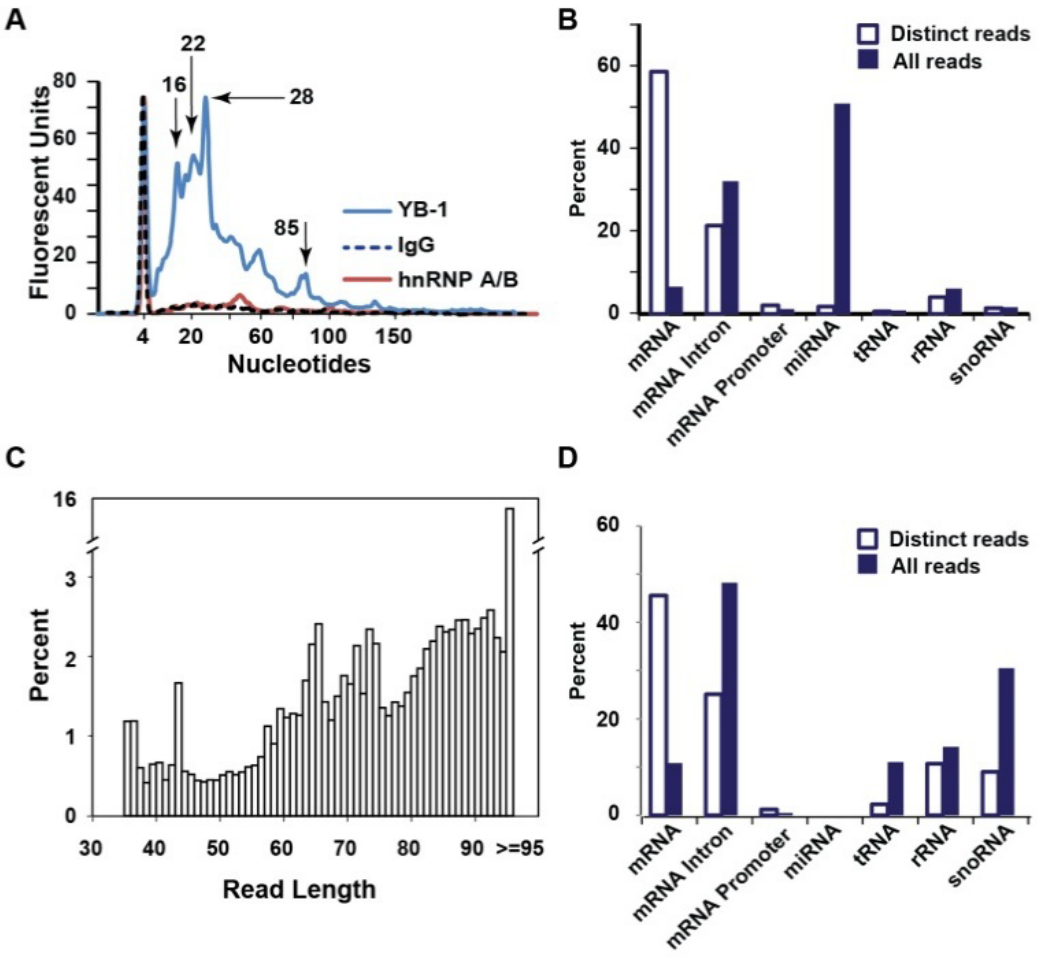


Figure 11: Characteristics of YB-1 associated small and short RNAs. A) Length distribution of small RNAs immunoprecipitated with YB-1, hnRNP A/B (control), and rabbit IgG (control), based on Agilent Bioanalyzer (marker peak at 4 nts). (Experimented by Dr. Teresa Liu) B) Analysis of distinct sequences reveal that although mRNAs account for the majority of distinct reads they correspond to small fraction of all reads; such low abundance reads may correspond to mRNA breakdown products. In contrast, although distinct reads from introns is smaller than those from mRNAs, they correspond to the second most abundant group of small RNAs, next to miRNAs. C) Length distribution of reads from YB-1 immunoprecipitated and size (~80-100 nts) fractionated RNAs. D) The majority of shyRNAs reads derive from intragenic regions, followed by snoRNAs.

4.3 RESULTS

4.3.1 Identified novel class of small RNAs associated with YB1 protein

The small YB1-associated RNAs are sequenced with Illumina Hi-seq sequencer. The Illumina reads represent a total of 436,671 distinct genomic loci (18,871,897 reads) among which 226,412 loci were cloned more than once. Among the genomically matching Illumina reads, 1.3 million reads (7%) maps to rRNAs and tRNAs, which may be due to the non-specific association of these abundant RNAs to YB-1. In other categories, 9.6 million reads (50.1%) map to miRNAs, 1.1 million reads (6.2%) map to sense mRNA exons and 6.6 million reads (32%) to mRNA introns (**Figure 11B**)**.** Among the 1,238 annotated mature miRNAs (miRBase), 553 (44%) were identified among the YB-1 associated small RNA fraction. A much higher proportion (69.8%) of the 381 annotated snoRNAs match to YB-1 small RNAs, suggesting a substantial enrichment of sdRNAs124[36]. To investigate the possible presence of short RNAs indicated in in the YB-1 RNA Bioanalyzer profile (**Figure 11A**), the 80 to 100 nts fraction of RNAs immunoprecipitated with YB-1 was extracted and sequenced using the Illumina GAII sequencing approach. The transcripts were sequenced that ranged from 35 to 100 nts with a broad ~84-92 nts peak (**Figure 11C**). In total, we estimate to have sequenced 257,907 distinct high-quality shyRNA loci. Despite the differences in the RNA sizes, the proportions of shyRNAs that map to mRNA and intronic regions resembled that of YB-1 associated small RNAs (**Figure 11D**). The potential contamination by tRNAs represent a low percentage of all reads (11%), corresponding to 19.8% (124/625) of all annotated tRNA locations. In contrast, 73% (279/381) of snoRNAs are present among the shyRNA fraction. Recurrent enrichment of snoRNAs in both the short and small RNA fractions, over

similarly sized RNAs, adds support to the notion that YB-1 preferentially associates with both snoRNAs and their processed, smaller sdRNAs.

4.3.2 Regulatory marks around shyRNA locations

To gain insights into shyRNA biogenesis, we probed whether known RNAPII binding sites preferentially overlap with well-transcribed shyRNA genomic locations (>5 reads; 51,510 locations). As controls, we used 10 different datasets of random genomic locations (random control), as well as 10 independent datasets of randomly selected exon locations (exon control). In general, while RNAPII binding sites are preferentially enriched at shyRNA locations relative to the random control, the exon controls are more enriched over the shyRNA locations. However, interrogation of GRO-seq[125] data of transcriptionally engaged RNA polymerases (RNAP I-III) reveal that shyRNA locations across all five datasets are several fold more enriched than both random and exon control (**Figure 12A**). While the discrepancy between ChIP-seq data and GRO-seq could be due to exonic RNAPII binding sites that are not used in either transcription initiation or productive elongation, it raised the possibility that shyRNA locations might preferentially associate with RNAPI/III genomic binding sites. However, since RNAPI activity is limited primarily to rRNAs[126] and RNAPIII sites account for a small fraction (<2%) of shyRNA locations, it suggests that the majority of shyRNA transcripts are generated by RNAPII.

To test for TF binding, we downloaded ENCODE[127] ChIP-seq binding sites for 148 TFs across 433 experimental conditions and analyzed TF binding patterns at shyRNA locations. Several TFs emerged as significantly ($p<10^{-10}$) overrepresented at shyRNA locations. Notably, CTCF, a well-known partner of YB-1[128][40] is overrepresented at shyRNA locations in the T47D cell line at levels higher than that of both random and exon controls and accounts for ~6% of the shyRNA

locations (± 500 nts). In contrast, the pattern is reversed in other cell lines where CTCF occupancy pattern at shyRNA locations resemble that of random control locations (**Figure 12A)**. In addition to known master regulators such as CTCF and p300, binding patterns of a few additional TFs are noteworthy, particularly Znf274, KAP1, ZBTB33, Bcl3, and Bcl11a.

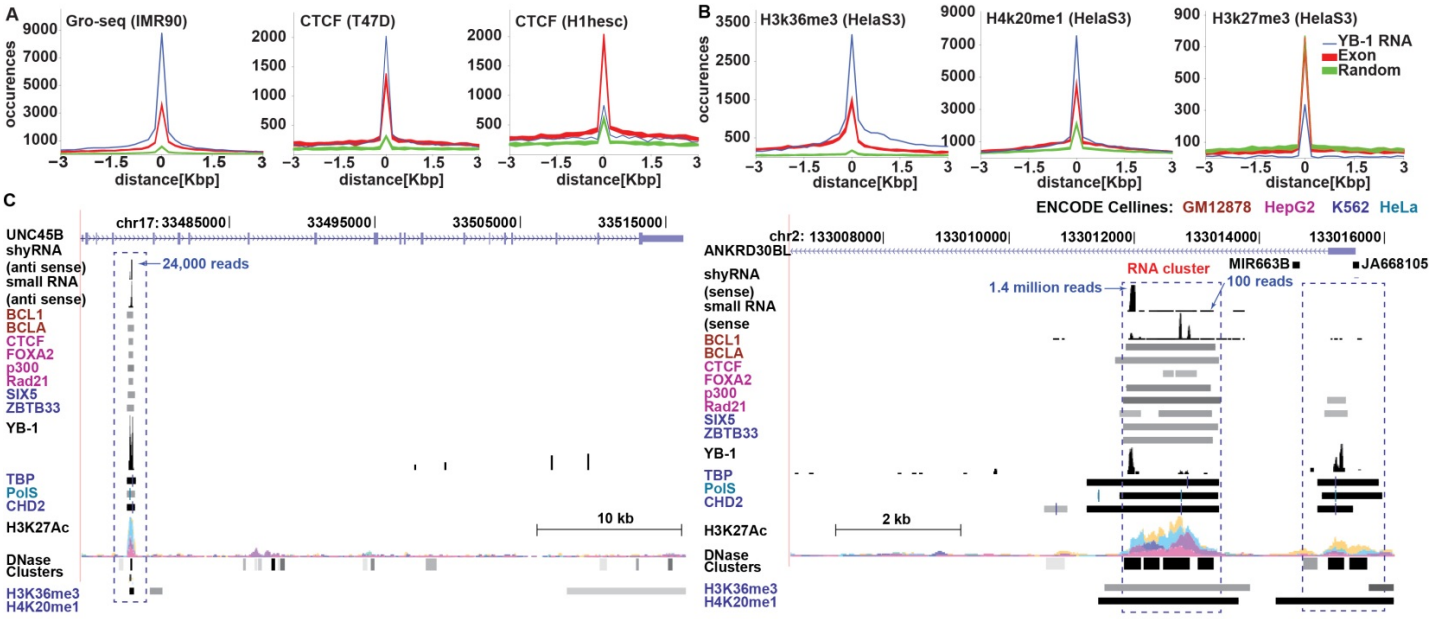


Figure 12: Specific transcription factor and histone modifications mark shyRNA loci in a cell-type dependent manner. A) shyRNA 5' locations are frequently located proximally (position 0) to protein binding sites in a context-dependent manner. The curves correspond to the distance distribution between the location of each feature and the nearest histone mark. Transcriptionally engaged polymerase locations (GRO-seq, left), are consistently more enriched towards shyRNAs in all cell lines analyzed. In contrast, regulatory factors such as CTCF manifest a cell-type dependent enrichment profile. B) Histone marks H3K36me3, and H4k20me1 are consistently overrepresented at shyRNA 5' locations, while the repressive mark H3k27me3 is under-represented at shyRNA locations. C) Examples of well-expressed shyRNA locations that precisely overlap with known binding sites of multiple TFs. Both locations reside within very low abundance genes (ANKRD30BL, UNC45B) as revealed by the lack of reads from both DRS and RNA-Seq of PC-3 cells.

In comparison to both random and exon controls, shyRNA genomic locations manifest a strong preference ($p < 10^{-10}$) towards H3K36Me3, consistently across all 42 datasets. Analysis of an additional 227 datasets spanning 12 histone marks, indicate that H4K20Me1 mark is also

preferentially enriched towards shyRNA genomic locations, at levels that surpass ($p < 10^{-10}$) those of both random ($p < 10^{-20}$) and exon controls ($p < 10^{-10}$) across all 11 cell lines. A similar pattern was also found for H3K9Me1 in several cell lines. Consistent with the notion that shyRNA locations have evolved as platforms of active transcription, the repressive histone mark, H3K27Me3, associated with gene silencing or heterochromatin formation[129,130] is significantly depleted at shyRNA genomic locations, occurring at levels that are generally lower than both random and exon controls across all experimental conditions (**Figure 12B**). A closer examination of several abundantly sequenced shyRNAs reveals an extraordinary convergence of several transcription factor binding sites including YB-1, and specific histone marks (H3K27Ac, H3K36Me3, and H4K20me1). The TF binding sites and histone marks frequently superpose precisely with shyRNA locations (**Figure 12C**). In summary, these observations indicate that shyRNA locations represent regions of the genome that are coordinately targeted in a context-dependent manner by several transcription factors and histone marks.

4.3.3 Transcription factors and histone modifications mark shyRNA locations

To gain insights into shyRNA biogenesis, we probed for transcriptional regulators and chromatin remodeling at shyRNA locations as compared to input controls (**Figure B1, B2**). RNA polymerase II (RNAPII) ChIP-seq data from several cell lines were obtained from Gene Expression Omnibus (GEO) and binding site locations overlapped with the shyRNA genomic locations. In the 88 cell lines examined, RNAPII was slightly enriched in the shyRNAs over the input controls even though an equal number of total reads were used from both fractions (**Figure 13A**). Multiple cell lines were used for comparison because data was unavailable in PC-3 cell lines. An in-depth interrogation using GRO-seq data of all transcriptionally engaged RNA polymerases (RNAP I-III),

the YB-1 specific RNAs were more highly enriched (**Figure 13B**). While the differences between the ChIP-seq and GRO-seq data could be due to exonic RNAPII binding sites that are not used in transcription initiation or productive elongation, the observed preference cannot be attributed to shyRNA locations that might preferentially associate with RNAPI/III genomic binding sites. This is due to the fact that RNAPI activity is limited primarily to a small number of rRNAs[133] and RNAPIII binding sites[134] only account for a very small fraction (<2%) of shy RNA locations (**Figure 13C**). This suggests that while RNAPIII may contribute to some shyRNA synthesis, the majority of shyRNA transcripts are generated by RNAPII.

Association of histone modifications with shyRNA locations provides an alternative method to gain insights into the underlying mechanisms that generate them. shyRNA genomic locations manifest a strong preference ($p < 10^{-20}$-10^{-35}) towards H3K36me3, consistently across all 86 datasets (**Figure 13D**). Trimethylation at H3K36 results in a more open chromatin configuration and is generally found within the body of active genes specifically at active exons[135]. H3K27ac and H4K20me1 are two other histone modifications that mark transcriptional activation that are preferentially enriched in shyRNA genomic locations over input controls[136,137](**Figure 13E**). H3K4me3, another activation modification normally marking the promoter regions of active genes, can be found overlapping shyRNA genomic locations, but the enrichment while statistically significant is not as striking over the control. Consistent with the hypothesis that shyRNA locations have evolved as platforms of active transcription, the repressive histone modification, H3K27me3, associated with gene silencing or heterochromatin formation is significantly depleted at shyRNA genomic locations over control[138,139] (**Figure 13F**).

The association of shyRNA locations with active transcription suggests that transcriptional regulators may also bind preferentially to these locations. ChIP-seq binding sites for 148 transcription factors (TF) across 433 experimental conditions were downloaded from ENCODE[140].

Several TFs emerged as significantly overrepresented at shyRNA locations enriched over input controls. Notably, two master regulators of transcription, CTCF and p300, are known to interact with YB-1 and are overrepresented at shyRNA locations in all cell lines examined[96,141] (**Figure 13G**). In conjunction with the association of shyRNAs with p300, shyRNA locations are also enriched at YY1 binding sites. YY1 is a zinc-finger transcription factor that has the ability to recruit coactivators such as p300[142]. Other noteworthy TFs that bind to shyRNA locations include KAP1/TRIM28, ETS1 and MYC/MAX. KAP1 is a multifunctional protein implicated in the assembly of epigenetic machinery[143,144]. It can interact with other transcription factors to promote transcription or bind to promoters and repress gene expression[144,145]. ETS-1 is a transcription factor commonly fused with androgen regulated genes in prostate cancer[146]. In PC-3 cells, ETS-1 promotes cell proliferation and migration but not invasion[147,148]. ETS-1 expression also correlates with poor prognosis in breast, ovarian and cervical cancers[149,150,151]. Interestingly, the TF binding sites enriched in shyRNAs are not evenly distributed among the shyRNA locations but rather precisely span shyRNA locations. The abundance of TFs at shyRNA locations combined with GRO-seq suggest that shyRNA transcription start sites are likely proximal to their 5' locations. These observations indicate that shyRNA locations represent regions of the genome that are targeted by several transcription factors and histone modifications, likely in a combinatorial manner.

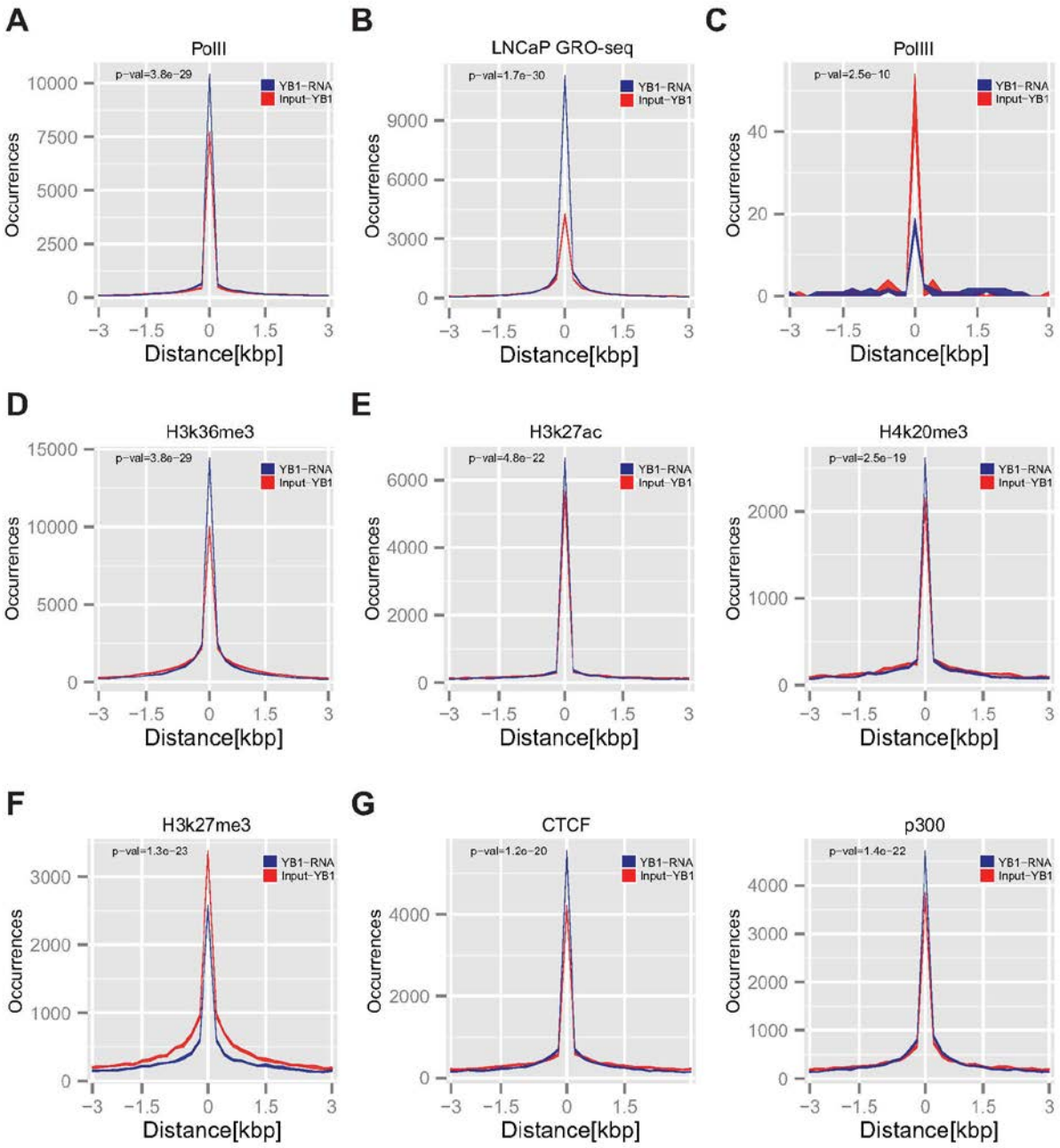


Figure 13. Specific transcription factor and histone modifications mark shyRNA loci. A) shyRNA correlation with PolII over input control in K562 cells. B) shyRNA correlation with GRO-seq data over input control in LNCaP cells. C) Input control correlation with PolIII over shyRNAs in K562. D) shyRNA correlation with H3K36me3 over input control. E) shyRNA correlation with H3K27ac and H4K20me3 over input control. F) Input control correlation with H3K27me3 over shyRNAs in Huvec cells. G. shyRNA correlation with CTCF and p300 in K562 cells over input control.

4.4 DISCUSSION

The YB-1 protein utilizes its DNA and RNA binding activity to influence various aspects of nucleic acid metabolism including DNA replication, transcription, mRNA splicing and export. We have identified a unique regulatory mechanism for YB-1 that is mediated by a novel class of short RNAs termed shyRNAs and their processed small RNAs that specifically associate with YB-1, perhaps as shown in one case by a direct interaction. All analysis that we have performed suggests that only a minor proportion of shyRNAs represent degradation products of longer non-coding RNAs or mRNAs since the overlap between shy RNAs and RNA-seq data from endo- and exonuclease knockdown cells are low. The biosynthesis of the YB-1 associated RNAs we identified is unknown, but these RNAs correlate with not only PolII binding sites but also are in locations of active transcription, as determined by GRO-seq and histone modifications. Additionally, many of shyRNA locations are binding sites for multiple different transcriptional regulators. The presence of multiple transcription factor binding sites along with fully assembled and initiated transcriptional complexes at regions where the chromatin status allows for active transcription supports the notion that shyRNAs may be derived by direct RNA PolII transcription. The possibility exists that, like TSSaRNAs, shyRNAs are byproducts related to RNAPII pausing [117,132]. This is unlikely because the shyRNAs are not found in promoter regions, do not manifest divergent transcription at the locations, and are found within intronic regions. It is also possible that shyRNAs are a function of abortive transcription and YB-1 associates with these RNAs as a clearance mechanism. This is also unlikely due to the conservation of the chromatin landscape at the shyRNA locations across all of the different cell lines we computationally examined. This is supported by the lack of a 5' modification of Shad1 and the minimal overlap between these locations and RNAs degraded by endo- and exo-nucleases.

The intersection between the transcription binding sites, histone modifications and the shyRNA locations also point to an interesting regulatory model with physiological implications. Recent studies have shown the presence of locations termed super-enhancers where large numbers of transcription factor binding sites cluster along with activating histone modifications at promoter regions[152,153]. These super-enhancer regions, found upstream of tumor associated genes in glioblastoma, multiple myeloma, and small cell lung cancer, often are more sensitive to chromatin regulators than regular enhancer regions[152]. The association of all of these factors at the genomic location of shyRNAs could be a mechanism by which these RNAs are transcribed and regulated. We speculate that shyRNAs are products of active genomic regions that are "polymerase friendly" (**Figure 7**). Taken together with the observation that shyRNAs include many known regulatory short RNAs such as Y RNAs, pre-miRNAs, and snoRNAs, these observations support the notion that shyRNAs represent a group of RNAs that share a common pathway but have evolutionarily acquired diverse cellular roles.

4.5 SUMMARY

In the study, we analyzed the property and relationship between two novel classes of YB-1 associated RNAs. And we found that shyRNAs are primarily intragenic and further processed at their 5'/3' termini to novel small RNAs. We also noticed that the shyRNAs are derived from genomic locations highly enriched in binding sites of RNA polymerases, transcription factors such as CTCF, as well as activating histone marks H3K36Me3 and H4K20Me1. These enrichments frequently occur at levels much higher than control exon locations in a cell-type-specific manner. Furthermore, many shyRNA locations precisely overlap with reported binding sites of several

transcription factors and co-regulators, suggesting a context-dependent convergence of multiple transcriptional regulators at these locations. These results point to a mechanism that might control their biogenesis.

5.0 CONCLUSIONS

The goal of the research is to investigate, annotate and understand the diverse roles of multiple RNAs in the genesis and regulation mechanism of human cancer with transcriptome analysis. To accomplish this goal, we first built a comprehensive polyadenylation map of human genome across major cancers, their cognate normal tissues and multiple cell lines, which could facilitate studies on 3' UTR isoforms and enable discoveries of important novel gene isoforms. Meanwhile, we have created the Expression and Polyadenylation Database (xPAD) as a web portal for the polyadenylation map. With these in place, we revealed polyadenylation patterns of non-coding and novel genes across major cancers and their cognate normal tissues and demonstrated that polyadenylation sites contain isoform-dependent regulatory marks

Moreover, we discovered and characterized two novel classes of YB-1 associated RNAs and investigated possible biogenesis and regulation mechanism of them where we discovered that various transcription factors and histone modifications may mark the YB-1 associated RNA locations.

Further, we designed and implemented an assembly-based compression tool, AbSEnT, with a de-Bruijn graph method, to squeeze the sequencing data,. We have demonstrated the feasibility and efficiency of the assembly-based algorithm and the tool. Also, when analyzing the distribution of word frequencies in sequencing data set, we have found a property of sequencing data set that is probably shared with books and social media data set, which might indicate that we could migrate

knowledge and experience in the natural language processing field to boost and facilitate the analysis of sequencing data.

APPENDIX A

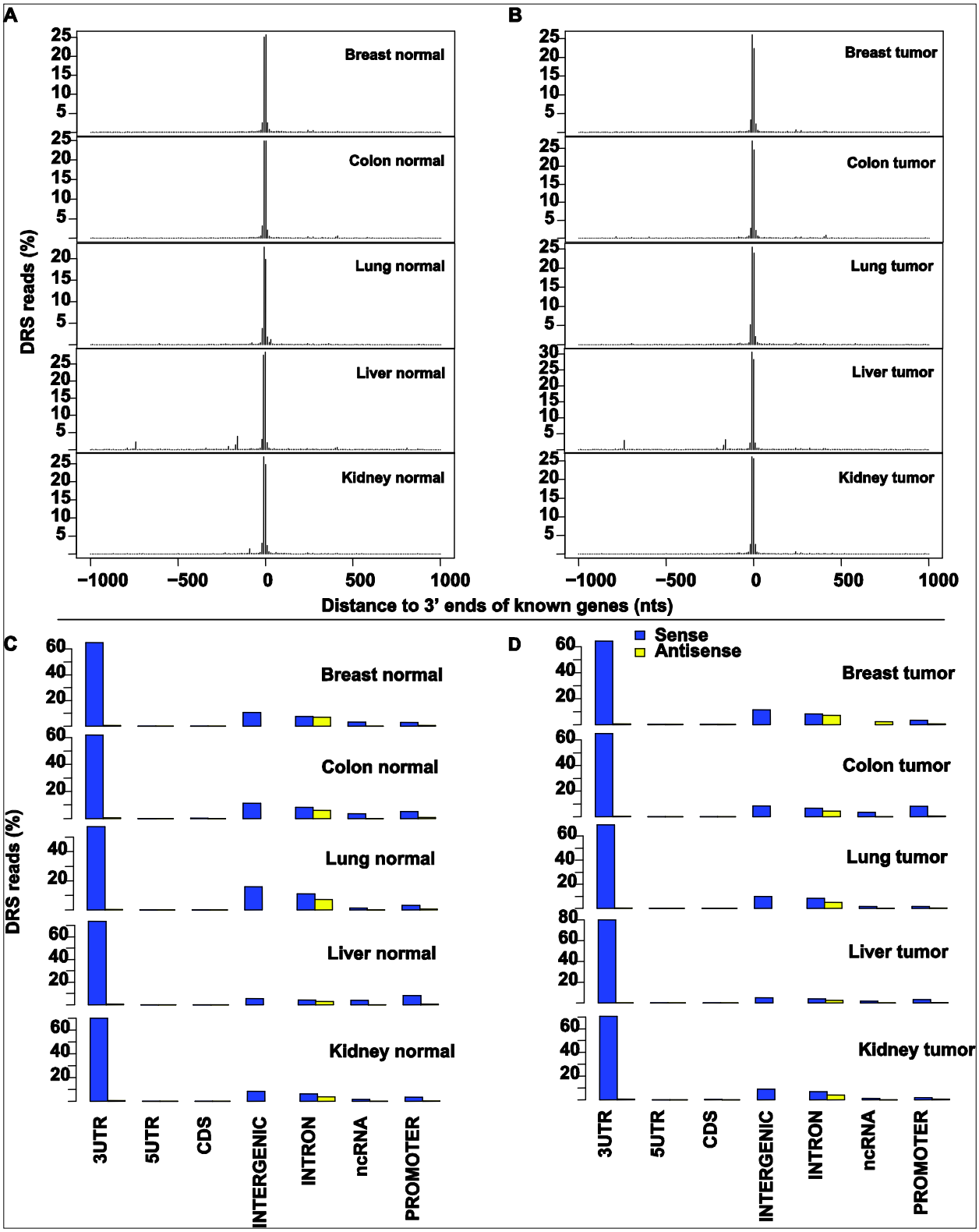

A
B
DRS reads (%)
Breast normal
Colon normal
Lung normal
Liver normal
Kidney normal
Breast tumor
Colon tumor
Lung tumor
Liver tumor
Kidney tumor
−1000
−500
0
500
1000
Distance to 3' ends of known genes (nts)
C
D
Sense
Antisense
3UTR
5UTR
CDS
INTERGENIC
INTRON
ncRNA
PROMOTER

Figure A1. Polyadenylation sites manifest similar overall characteristics across all tissues A-B) DRS reads predominantly match (bin size=10 nts) to annotated 3' ends of known genes across all tissue types in both normal (A) and tumor (B). C-D) Majority of DRS reads match to sense strand of transcriptionally active regions.

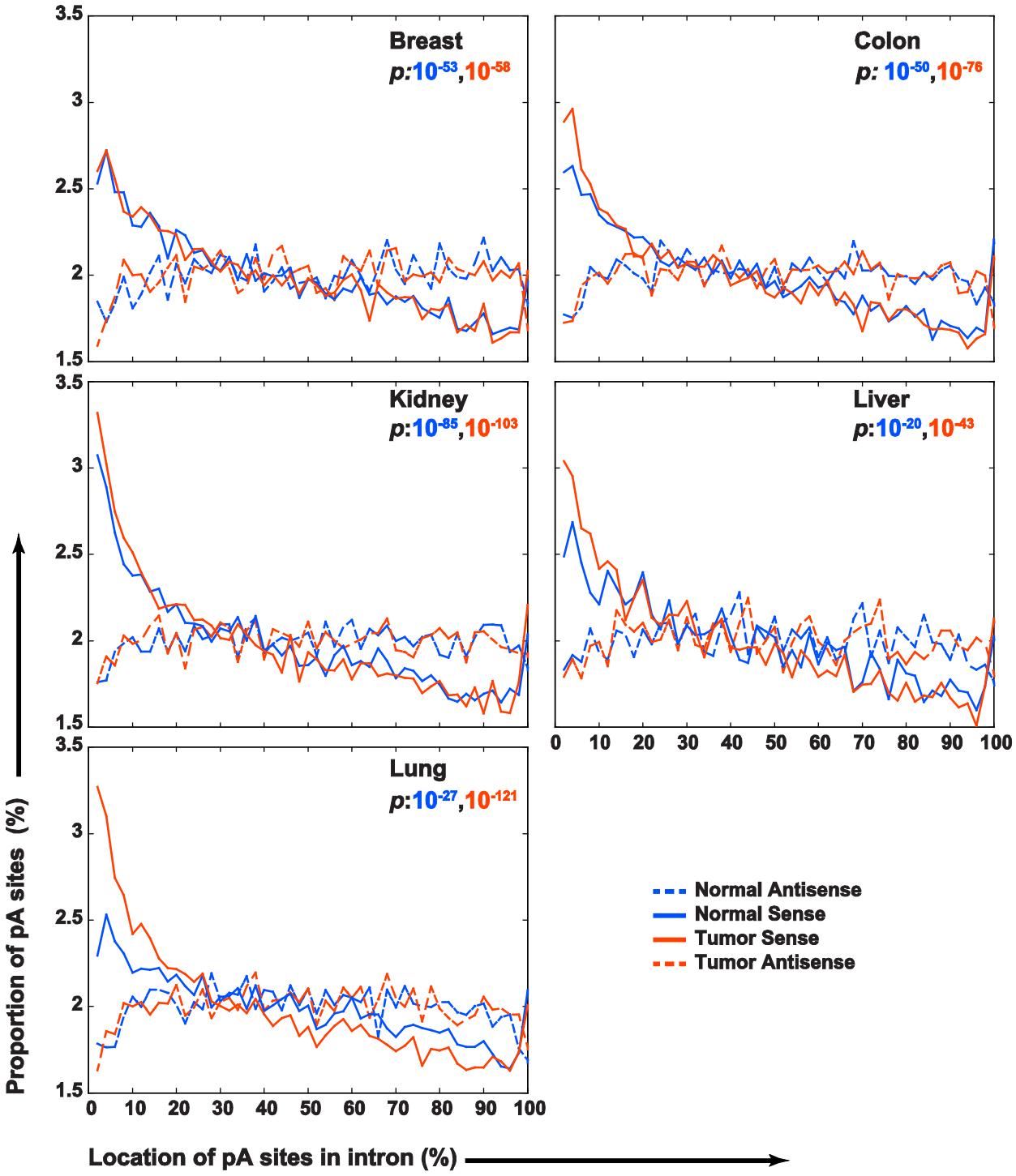


Figure A2. Distribution of the locations of polyadenylation sites in sense and antisense intronic transcripts. The full-range of each intron is standardized to 0-100%, and the polyadenylation site location is with respect to the start of

the intron (sense strand; bin size=2%). For example, in normal breast, ~3% of pA sites that occur in sense introns are located within the first 2% of the full-length of intron. For each pair of sense and antisense distributions in either normal or tumor tissue, one-sided two-sample Kolmogorov-Smirnov test was performed to analyze if the observed preference of sense intronic transcripts to occur towards intron start is statistically significant in comparison to that of antisense intronic pA sites.

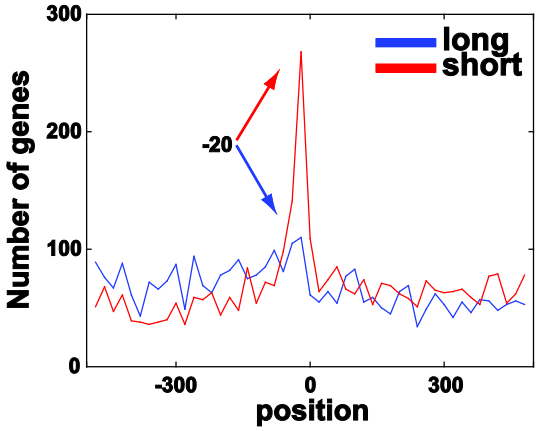


Figure A3. Strong positional preference of ATATAT motif. The motif occurs 20 nts upstream of the polyadenylation site, exclusive to short isoforms. We note that a considerable proportion (~12% or ~400) of genes containing 3' UTR isoforms used for the analysis seem affected by this motif.

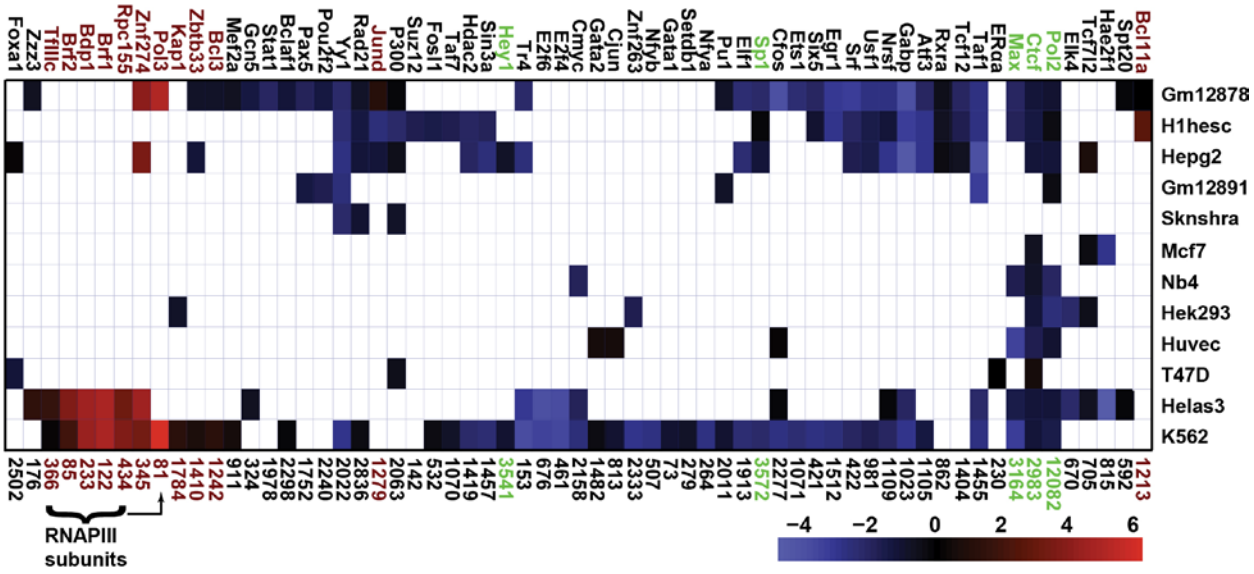


Figure A4. Summary of transcriptional regulators bind to genomic locations.

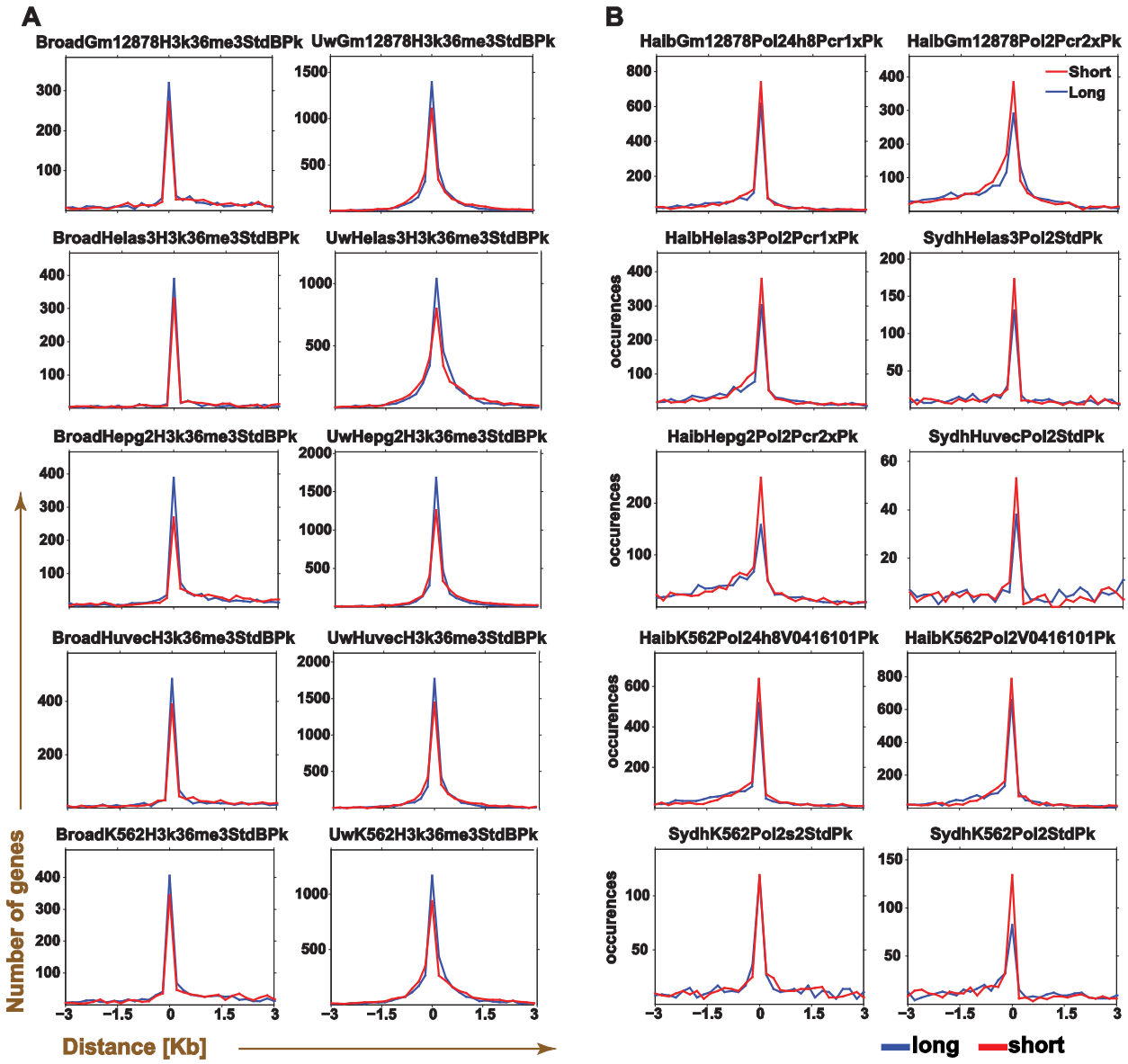


Figure A5. Distance distribution of the H3K36Me3 and Pol2 marks next to short and long isoform polyadenylation sites in multiple cell lines.

APPENDIX B

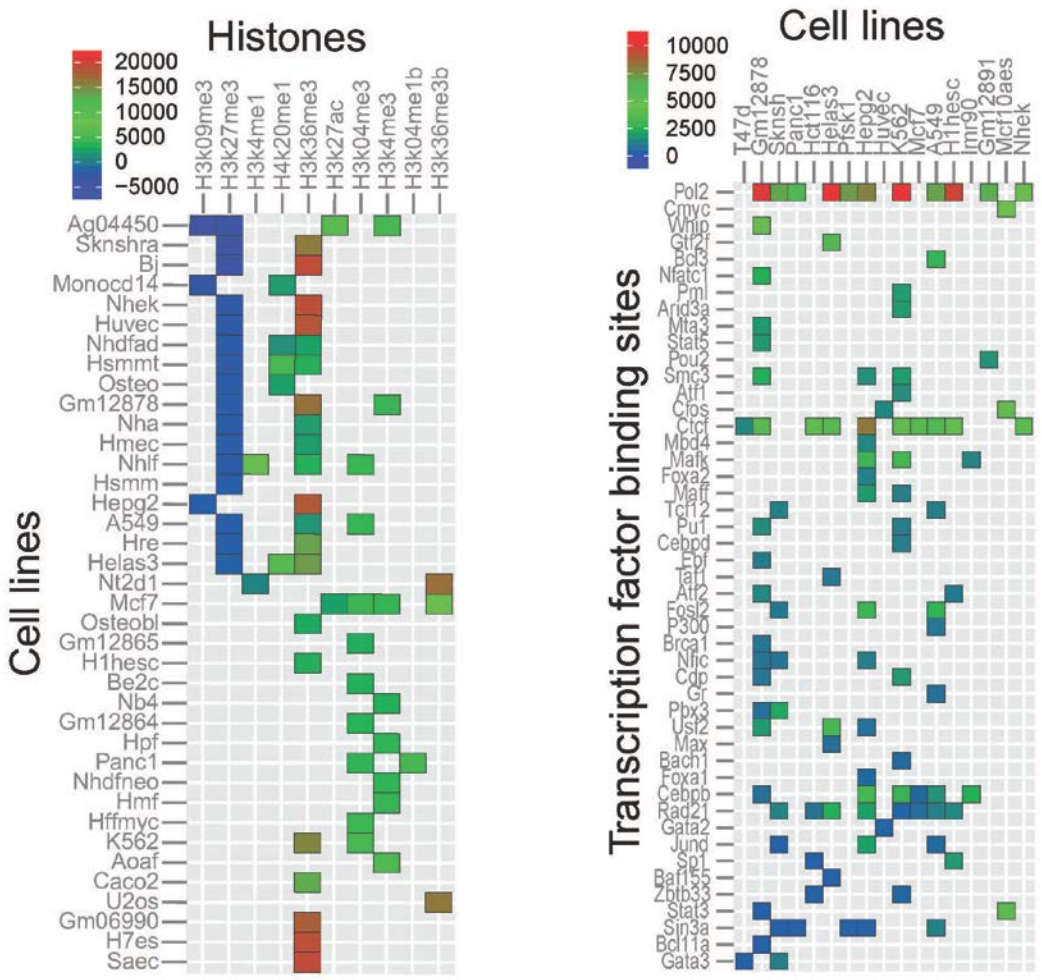
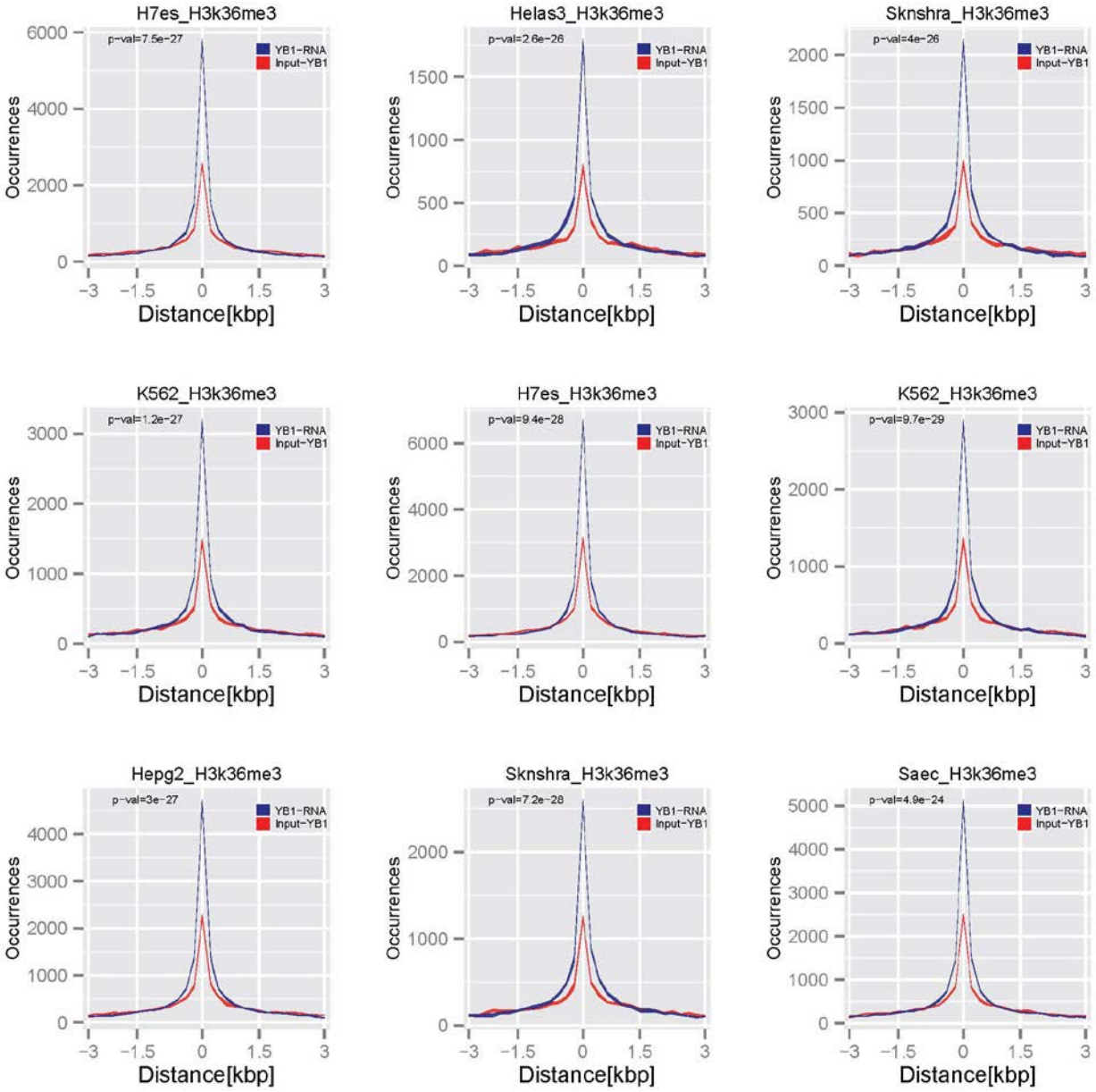
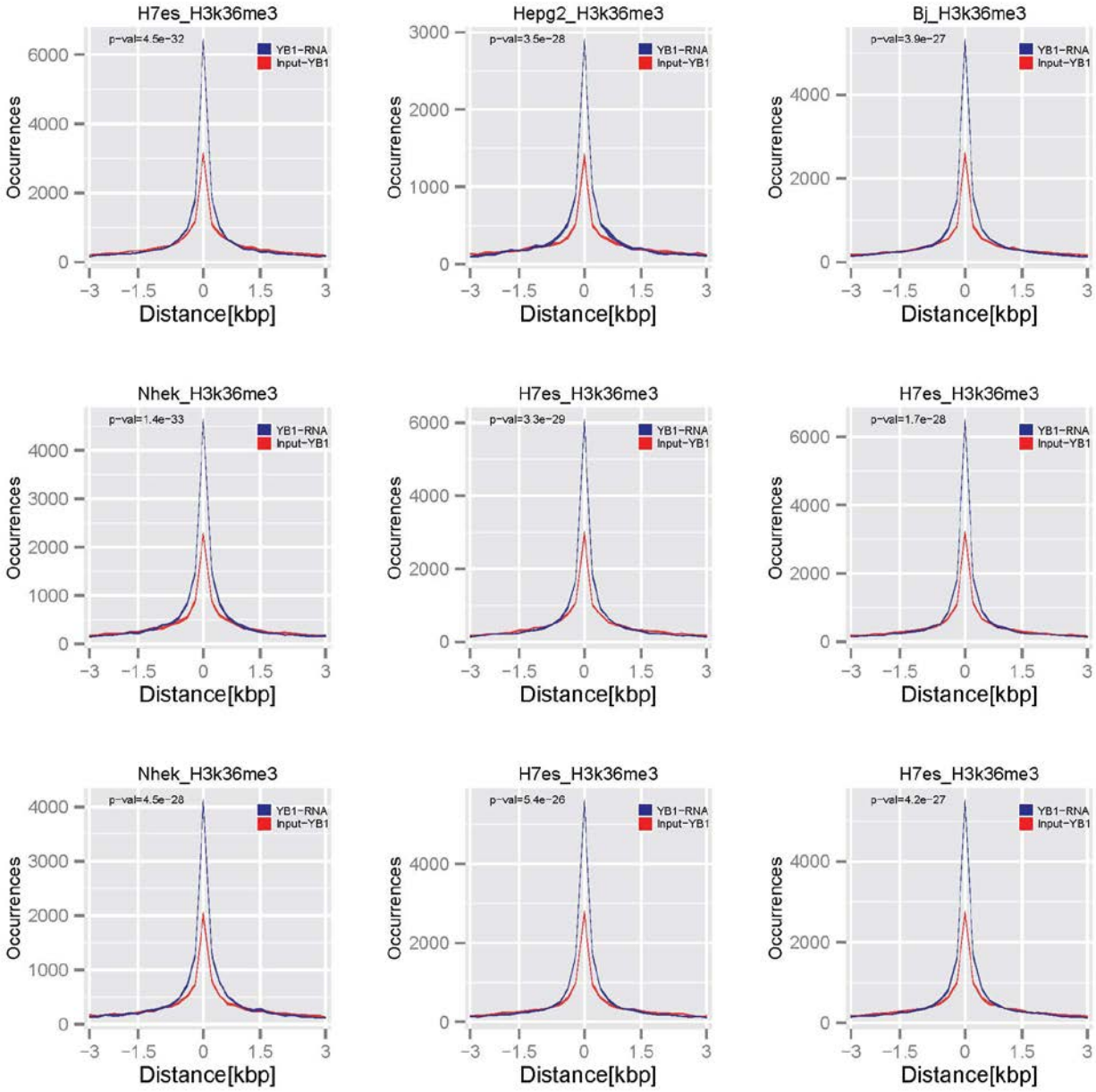
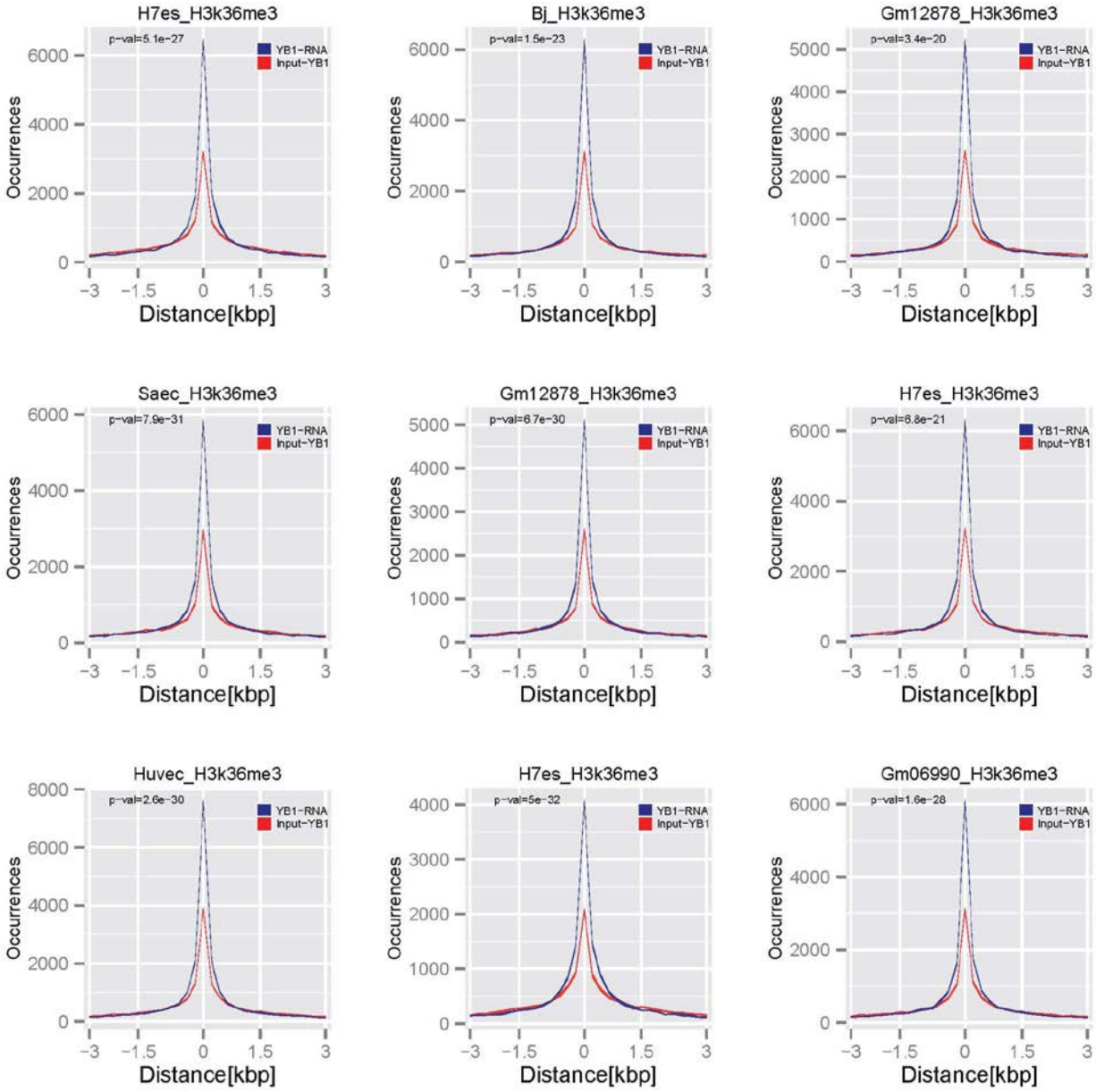
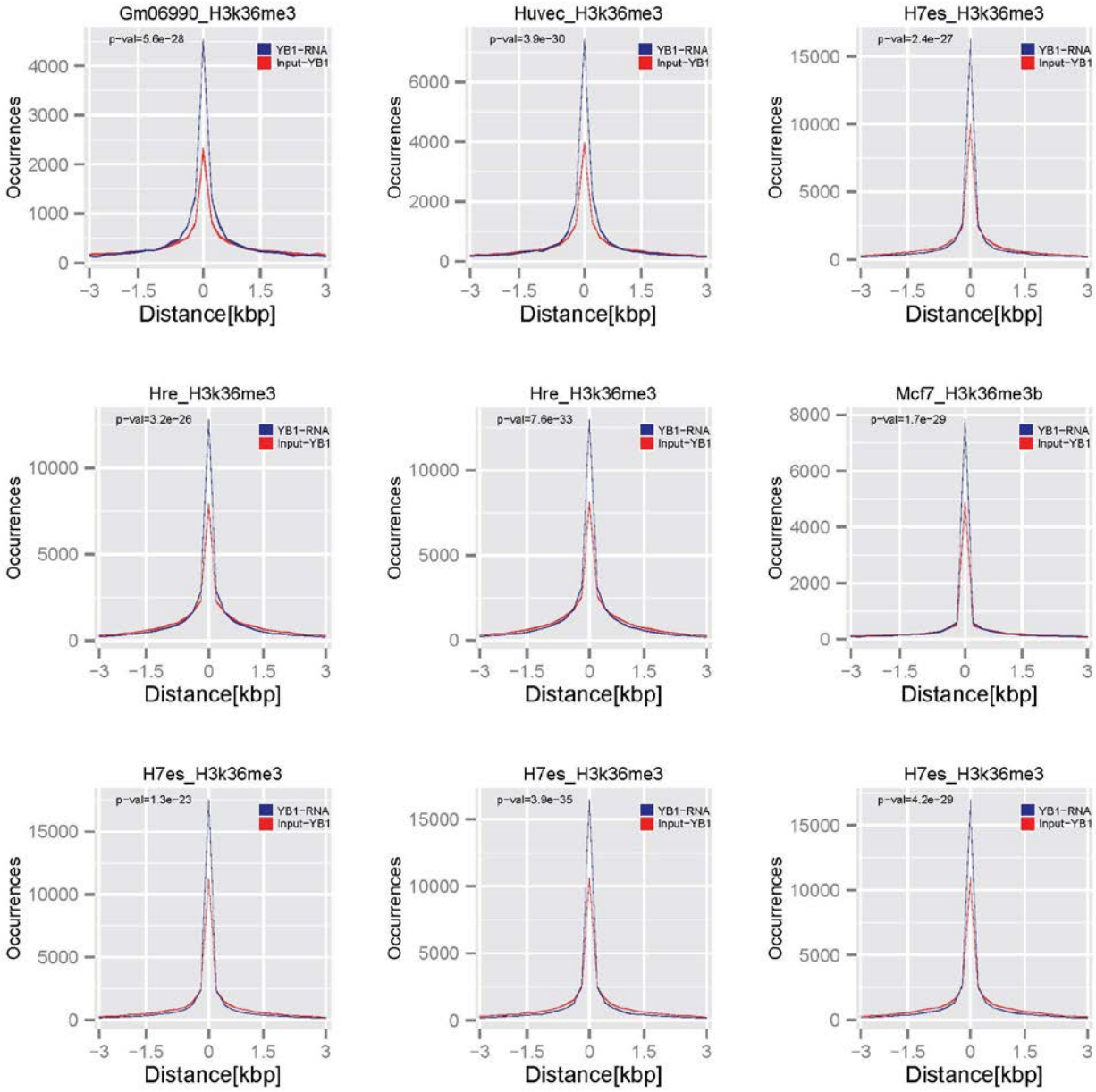


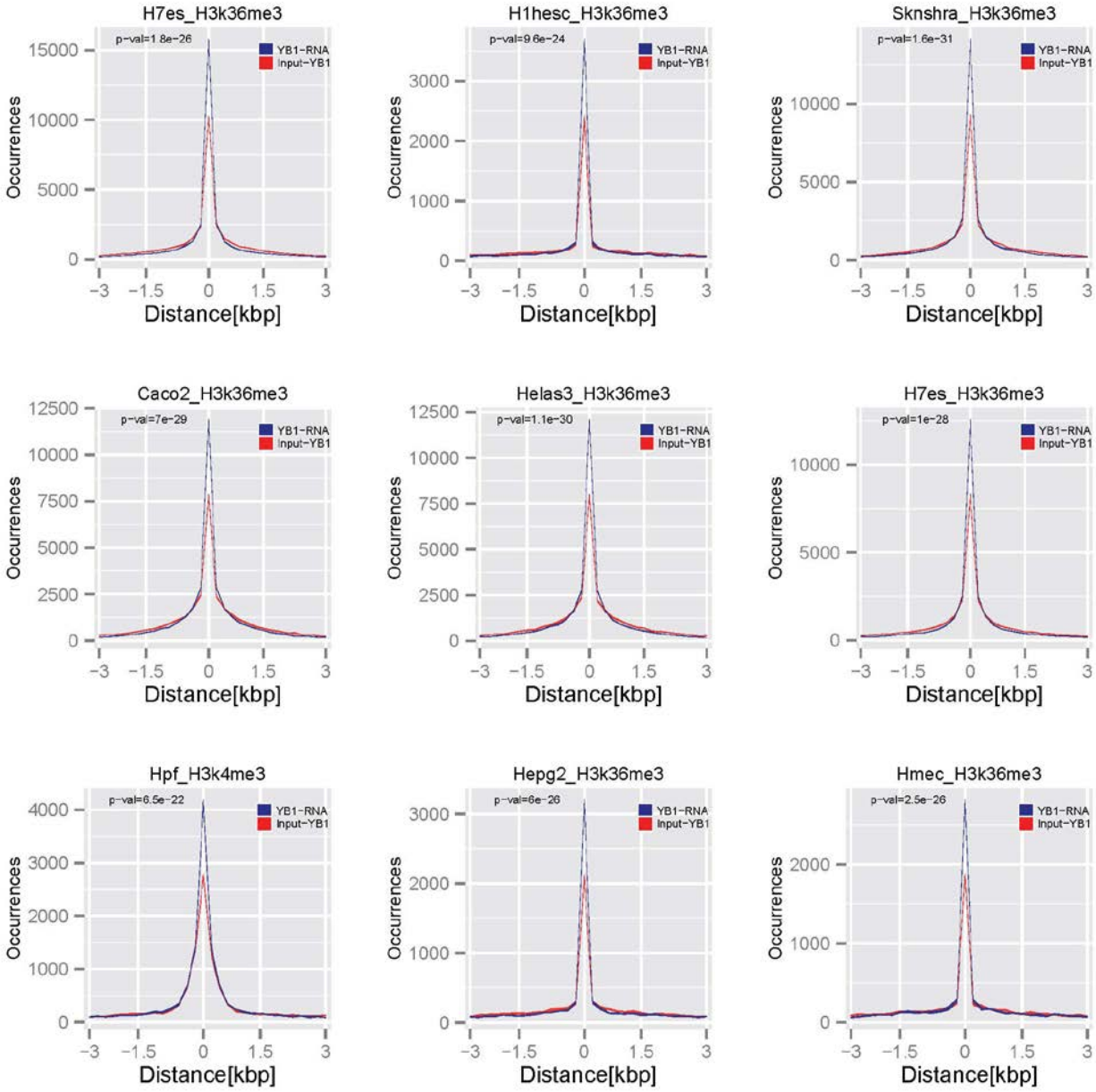
Figure B1. Summary of chromatin remodeling factors and transcriptional regulators bind to shyRNA genomic locations.

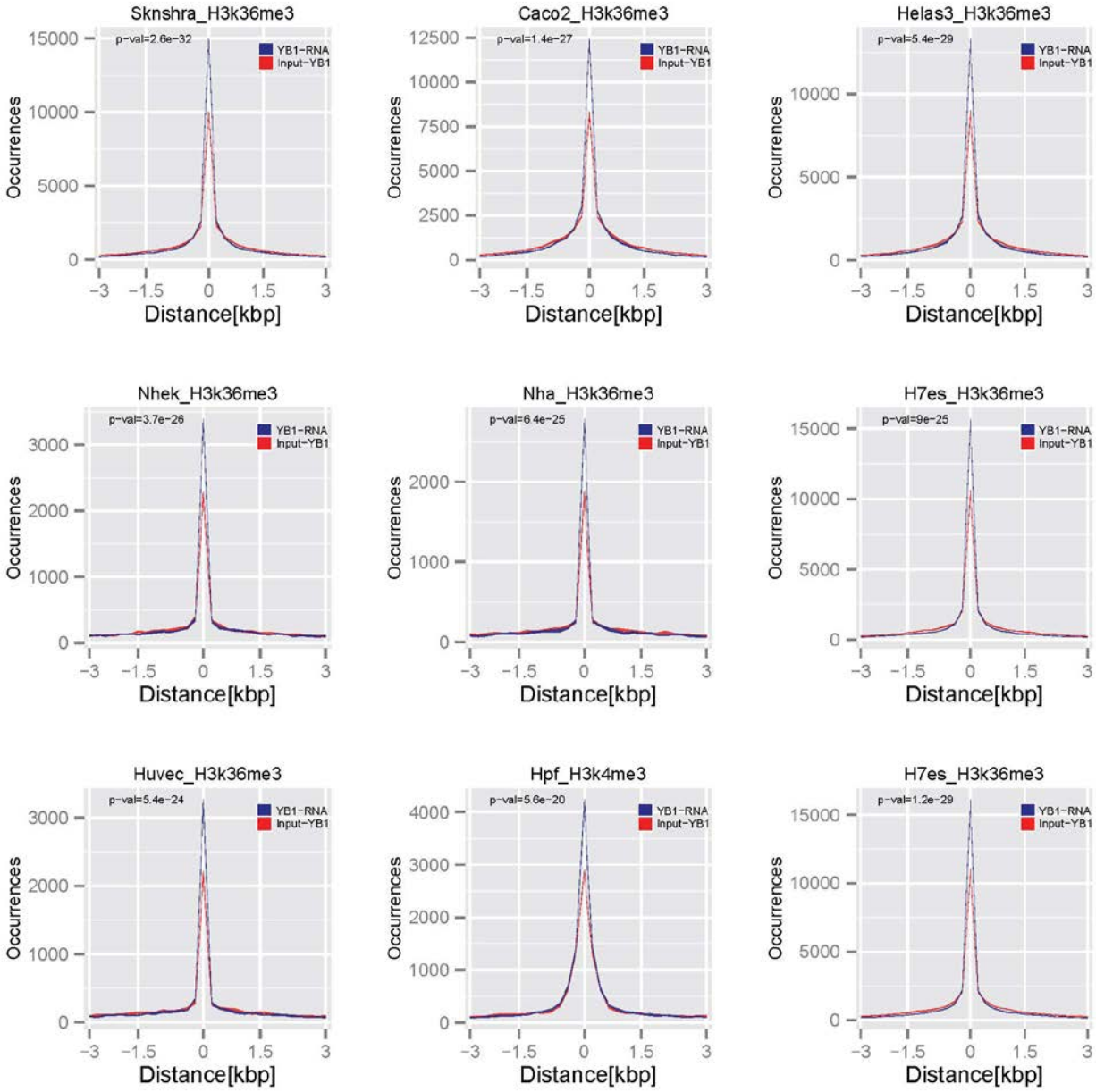

H7es_H3k36me3
p-val=7.5e-27
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Helas3_H3k36me3
p-val=2.6e-26
Sknshra_H3k36me3
p-val=4e-26
K562_H3k36me3
p-val=1.2e-27
H7es_H3k36me3
p-val=9.4e-28
K562_H3k36me3
p-val=9.7e-29
Hepg2_H3k36me3
p-val=3e-27
Sknshra_H3k36me3
p-val=7.2e-28
Saec_H3k36me3
p-val=4.9e-24

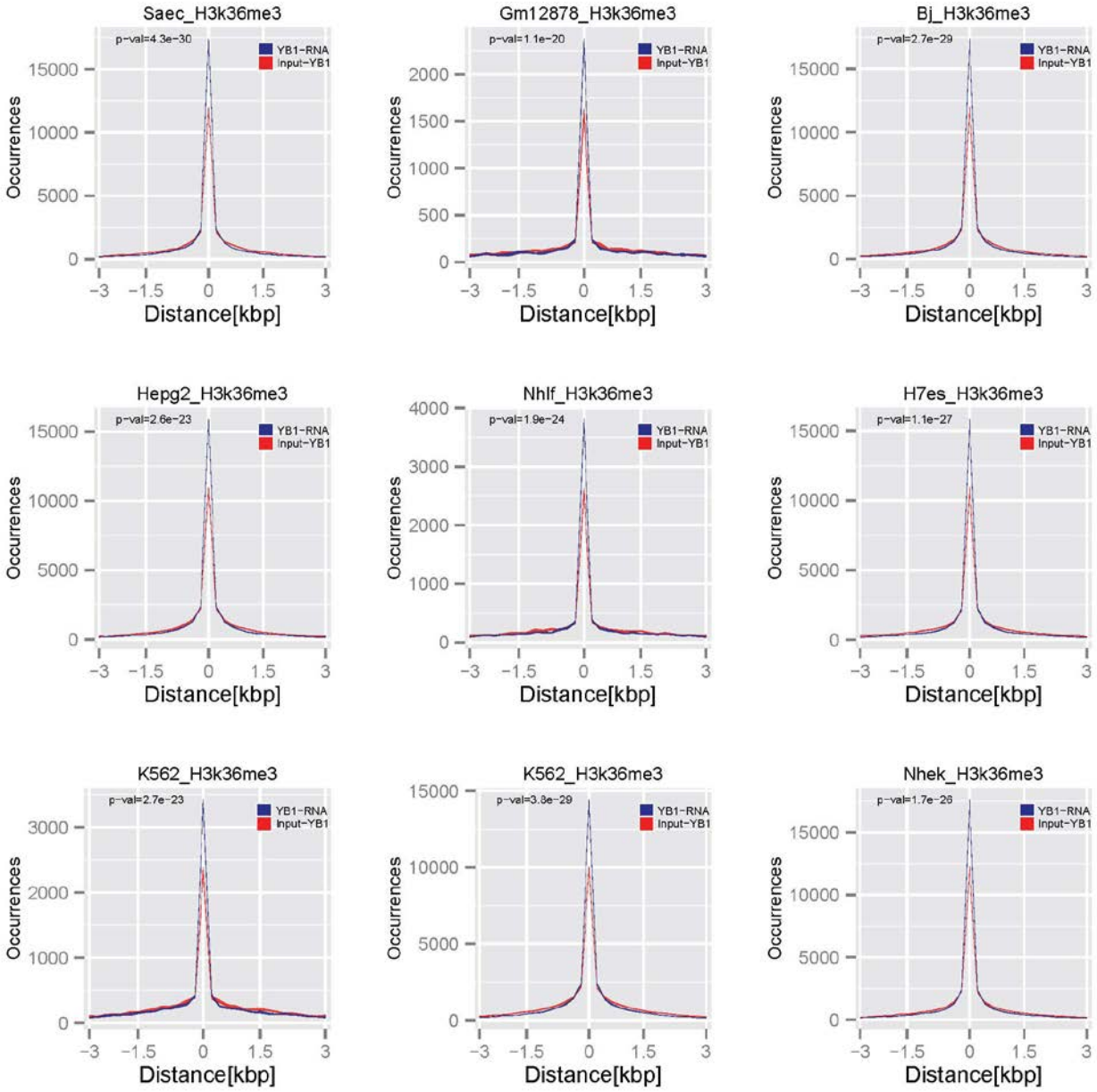

H7es_H3k36me3
p-val=4.5e-32
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_H3k36me3
p-val=3.5e-28
Bj_H3k36me3
p-val=3.9e-27
Nhek_H3k36me3
p-val=1.4e-33
H7es_H3k36me3
p-val=3.3e-29
H7es_H3k36me3
p-val=1.7e-28
Nhek_H3k36me3
p-val=4.5e-28
H7es_H3k36me3
p-val=5.4e-26
H7es_H3k36me3
p-val=4.2e-27

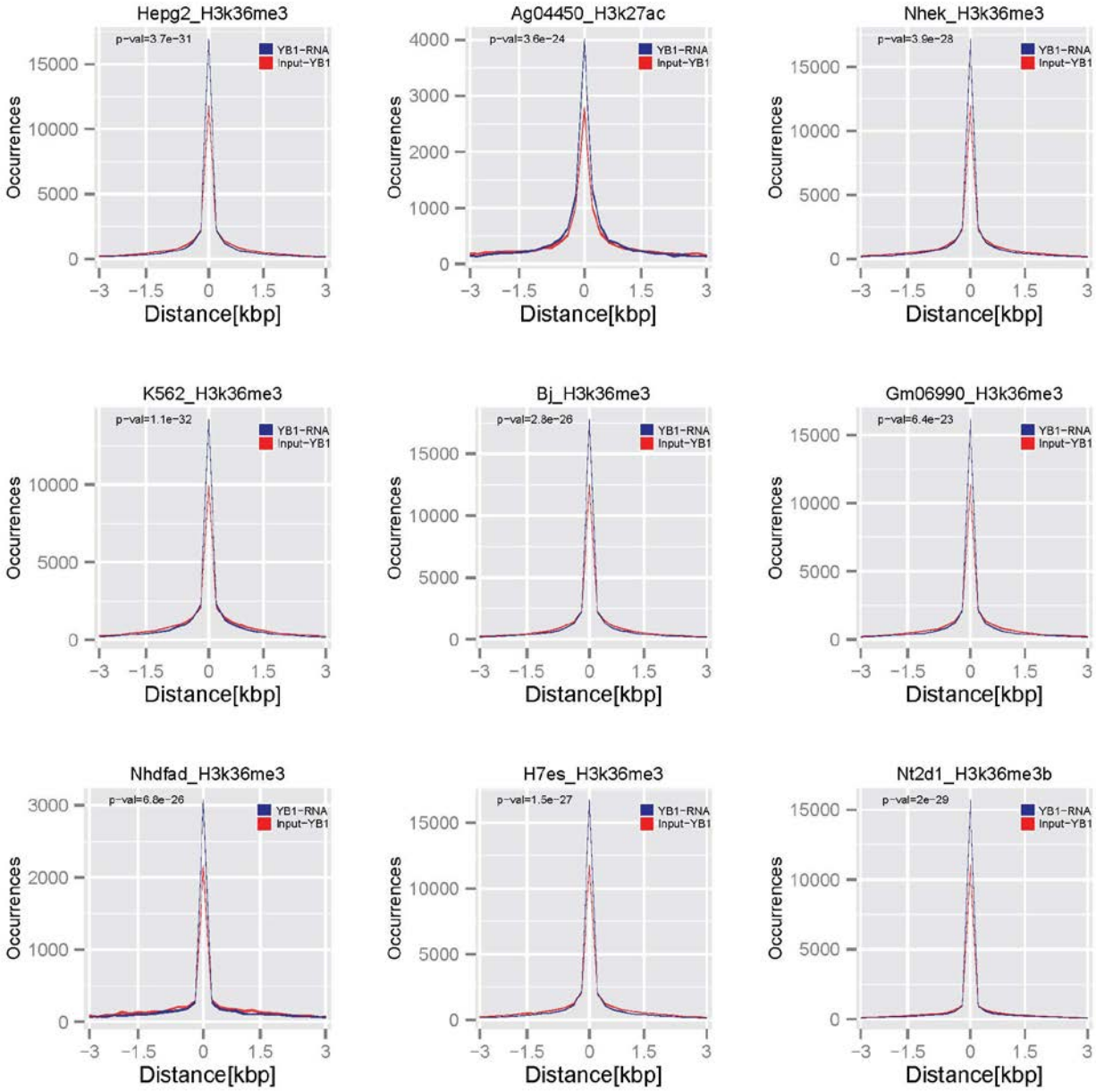

H7es_H3k36me3
p-val=5.1e-27
Bj_H3k36me3
p-val=1.5e-23
Gm12878_H3k36me3
p-val=3.4e-20
Saec_H3k36me3
p-val=7.9e-31
Gm12878_H3k36me3
p-val=6.7e-30
H7es_H3k36me3
p-val=6.8e-21
Huvec_H3k36me3
p-val=2.6e-30
H7es_H3k36me3
p-val=5e-32
Gm06990_H3k36me3
p-val=1.6e-28
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]

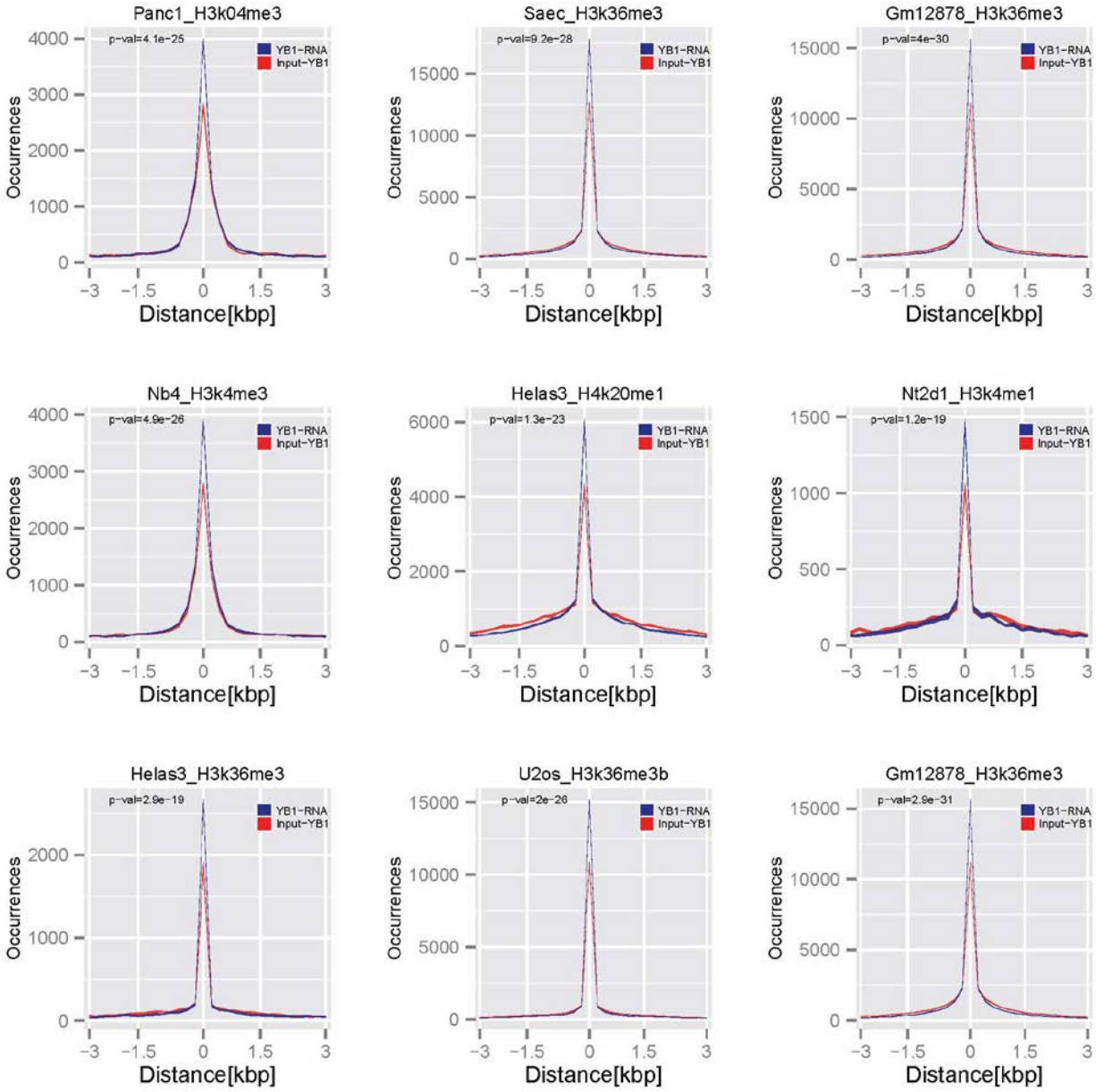

Gm06990_H3k36me3
p-val=5.6e-28
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Huvec_H3k36me3
p-val=3.9e-30
H7es_H3k36me3
p-val=2.4e-27
Hre_H3k36me3
p-val=3.2e-26
Hre_H3k36me3
p-val=7.6e-33
Mcf7_H3k36me3b
p-val=1.7e-29
H7es_H3k36me3
p-val=1.3e-23
H7es_H3k36me3
p-val=3.9e-35
H7es_H3k36me3
p-val=4.2e-29

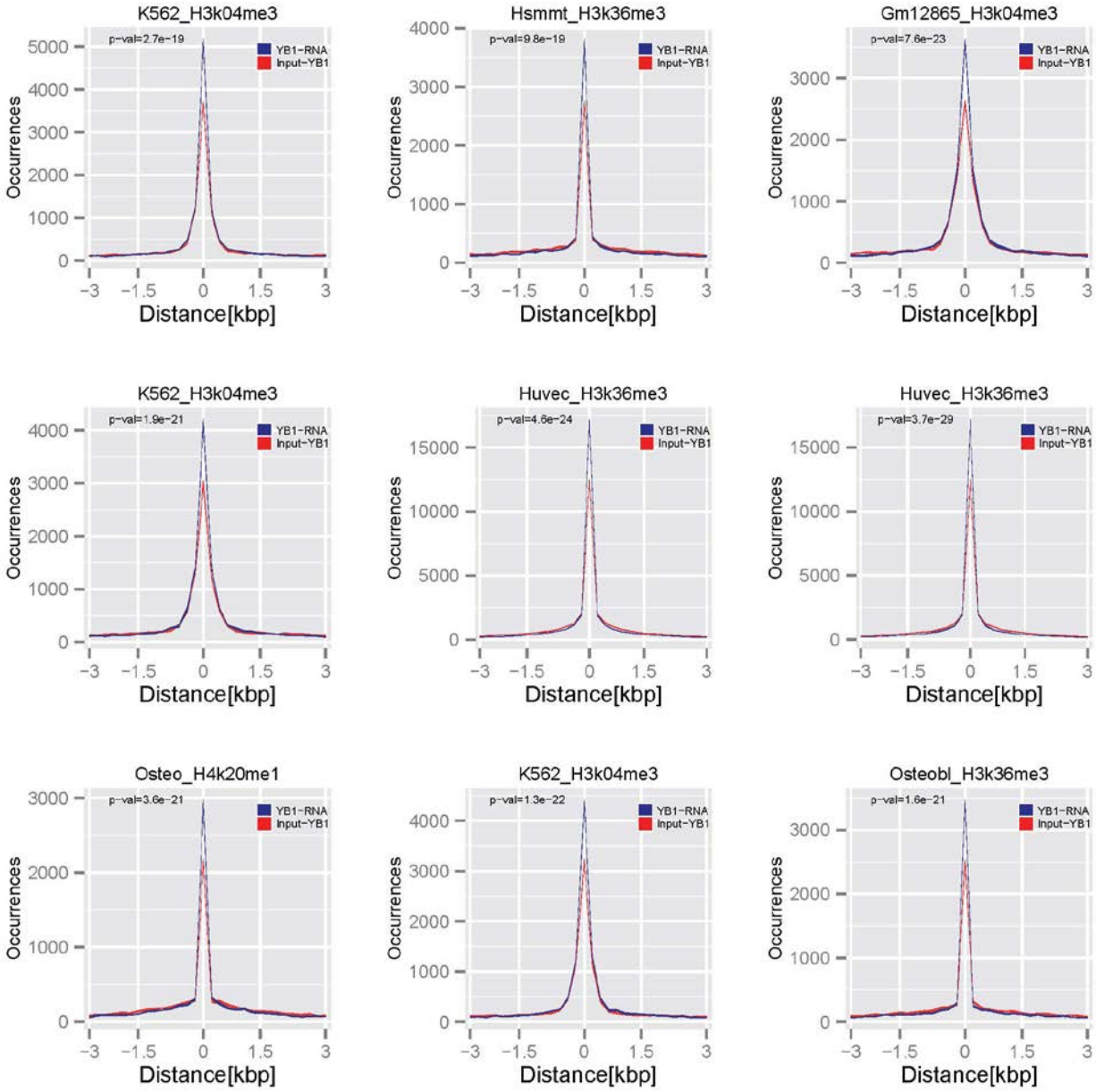

H7es_H3k36me3
p-val=1.8e-26
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
H1hesc_H3k36me3
p-val=9.6e-24
Sknshra_H3k36me3
p-val=1.6e-31
Caco2_H3k36me3
p-val=7e-29
Helas3_H3k36me3
p-val=1.1e-30
H7es_H3k36me3
p-val=1e-28
Hpf_H3k4me3
p-val=6.5e-22
Hepg2_H3k36me3
p-val=6e-26
Hmec_H3k36me3
p-val=2.5e-26

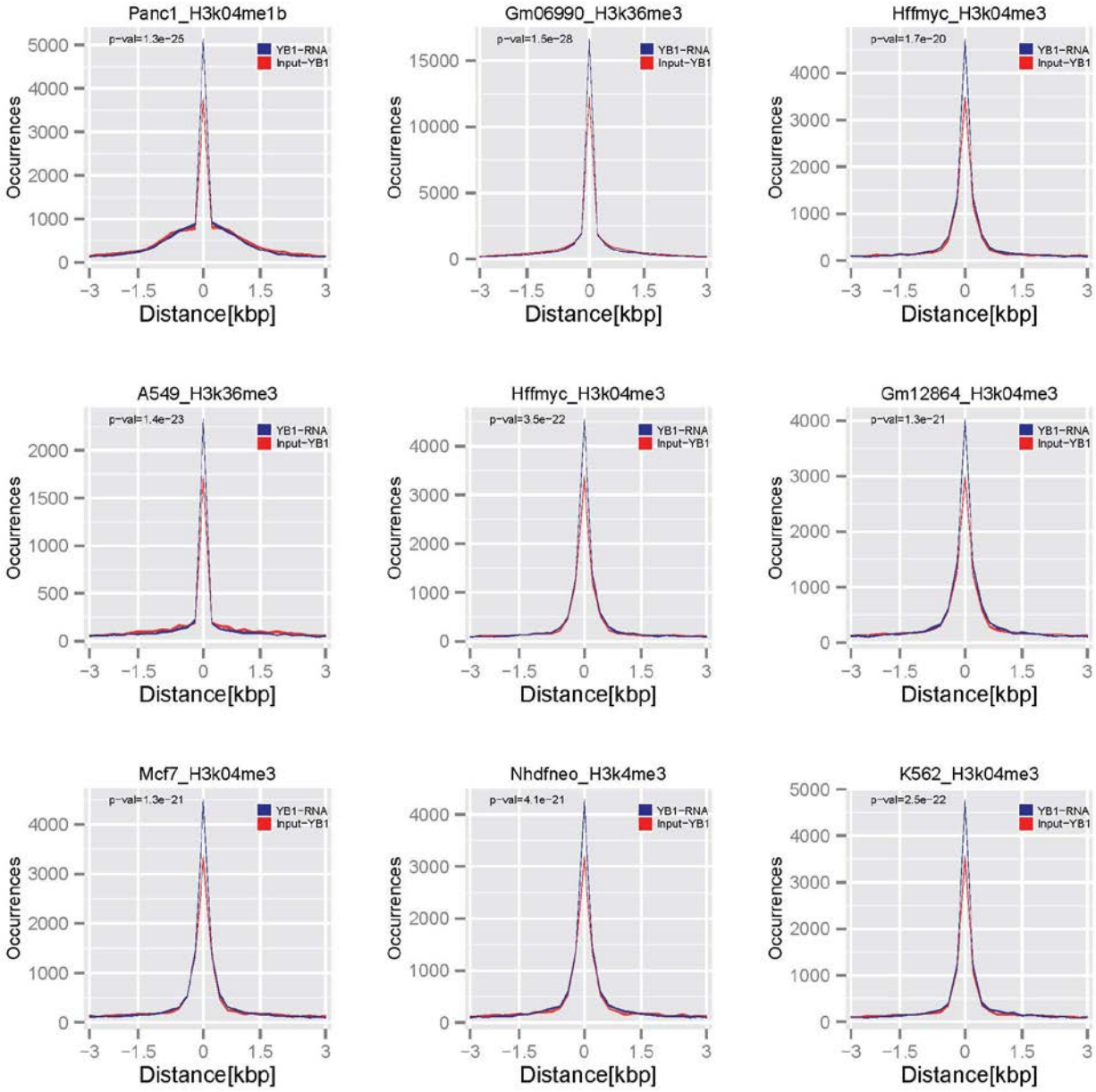

Sknshra_H3k36me3
p-val=2.6e-32
Caco2_H3k36me3
p-val=1.4e-27
Helas3_H3k36me3
p-val=5.4e-29
Nhek_H3k36me3
p-val=3.7e-26
Nha_H3k36me3
p-val=6.4e-25
H7es_H3k36me3
p-val=9e-25
Huvec_H3k36me3
p-val=5.4e-24
Hpf_H3k4me3
p-val=5.6e-20
H7es_H3k36me3
p-val=1.2e-29
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]

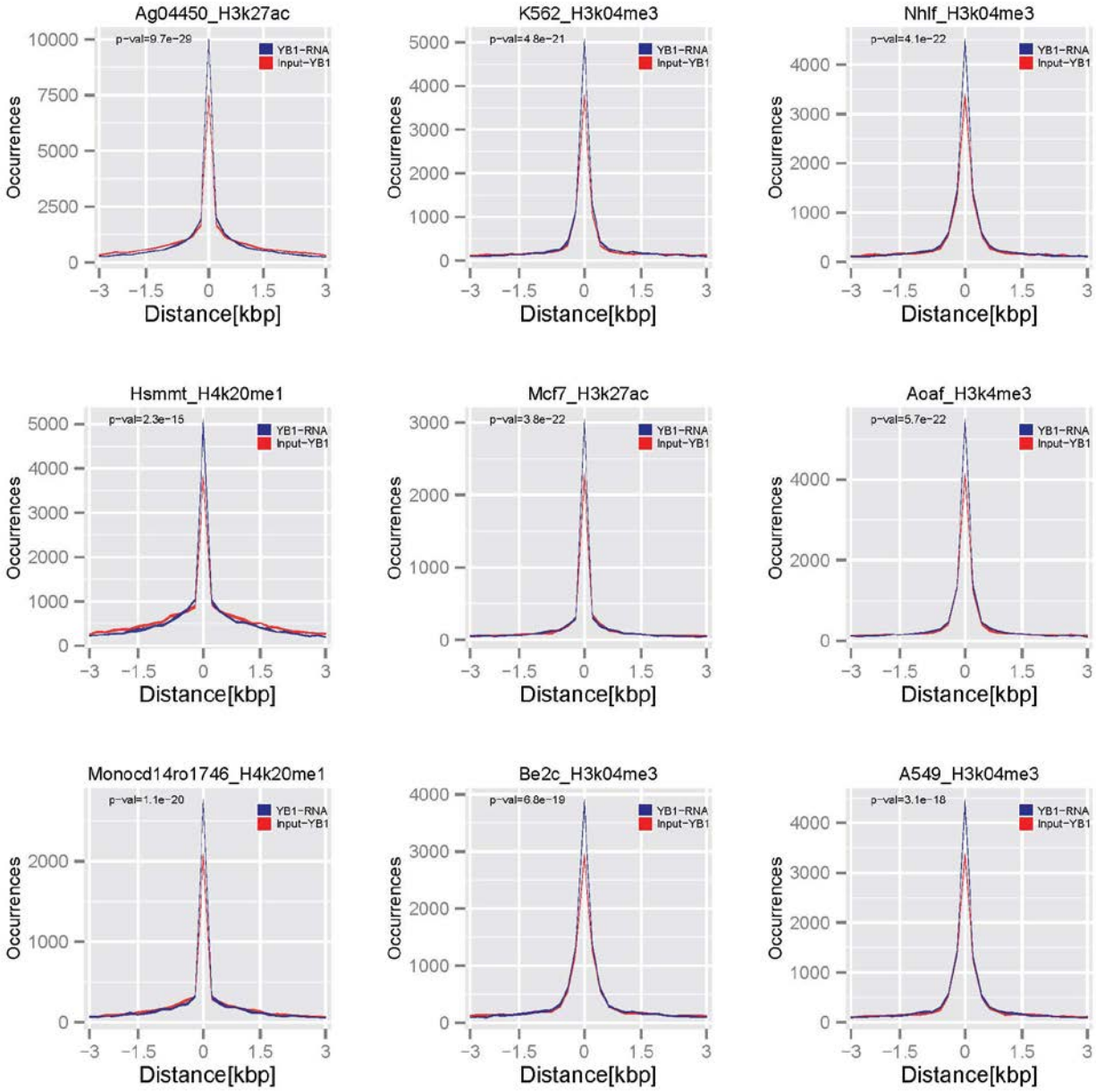

Saec_H3k36me3
p-val=4.3e-30
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Gm12878_H3k36me3
p-val=1.1e-20
Bj_H3k36me3
p-val=2.7e-29
Hepg2_H3k36me3
p-val=2.6e-23
Nhlf_H3k36me3
p-val=1.9e-24
H7es_H3k36me3
p-val=1.1e-27
K562_H3k36me3
p-val=2.7e-23
K562_H3k36me3
p-val=3.8e-29
Nhek_H3k36me3
p-val=1.7e-26

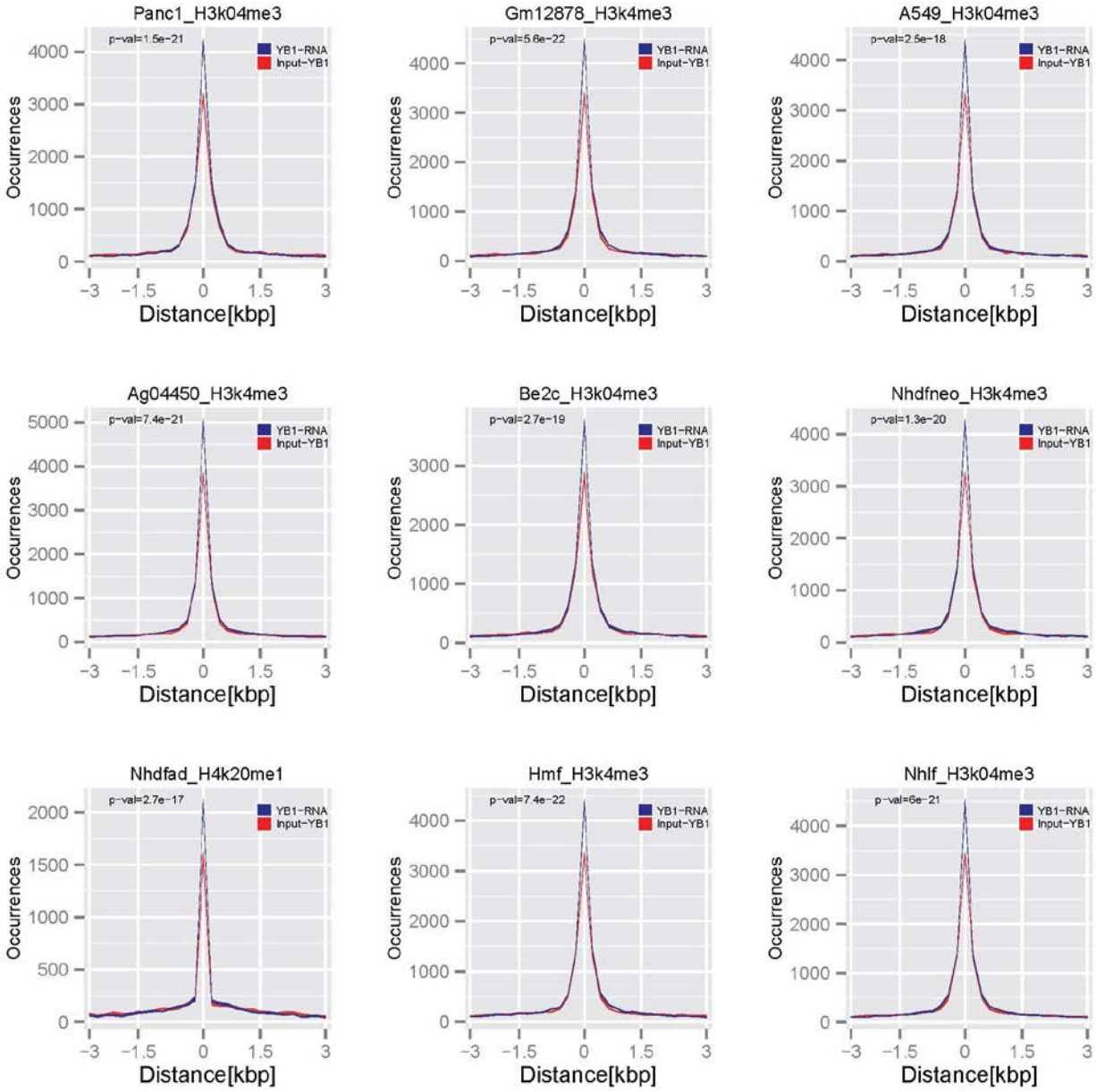

Hepg2_H3k36me3
p-val=3.7e-31
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Ag04450_H3k27ac
p-val=3.6e-24
Nhek_H3k36me3
p-val=3.9e-28
K562_H3k36me3
p-val=1.1e-32
Bj_H3k36me3
p-val=2.8e-26
Gm06990_H3k36me3
p-val=6.4e-23
Nhdfad_H3k36me3
p-val=6.8e-26
H7es_H3k36me3
p-val=1.5e-27
Nt2d1_H3k36me3b
p-val=2e-29

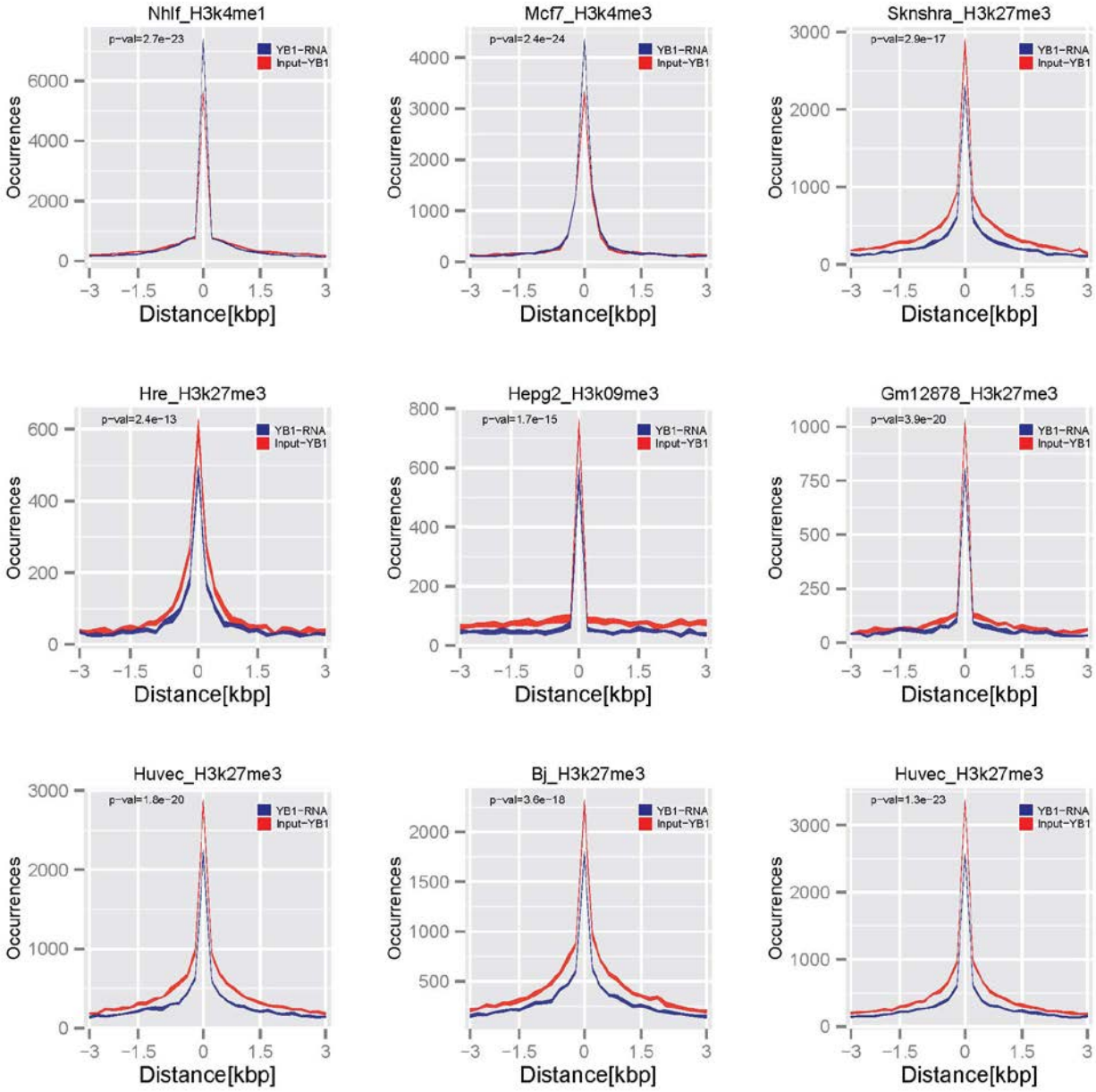

Panc1_H3k04me3
p-val=4.1e-25
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Saec_H3k36me3
p-val=9.2e-28
Gm12878_H3k36me3
p-val=4e-30
Nb4_H3k4me3
p-val=4.9e-26
Helas3_H4k20me1
p-val=1.3e-23
Nt2d1_H3k4me1
p-val=1.2e-19
Helas3_H3k36me3
p-val=2.9e-19
U2os_H3k36me3b
p-val=2e-26
Gm12878_H3k36me3
p-val=2.9e-31

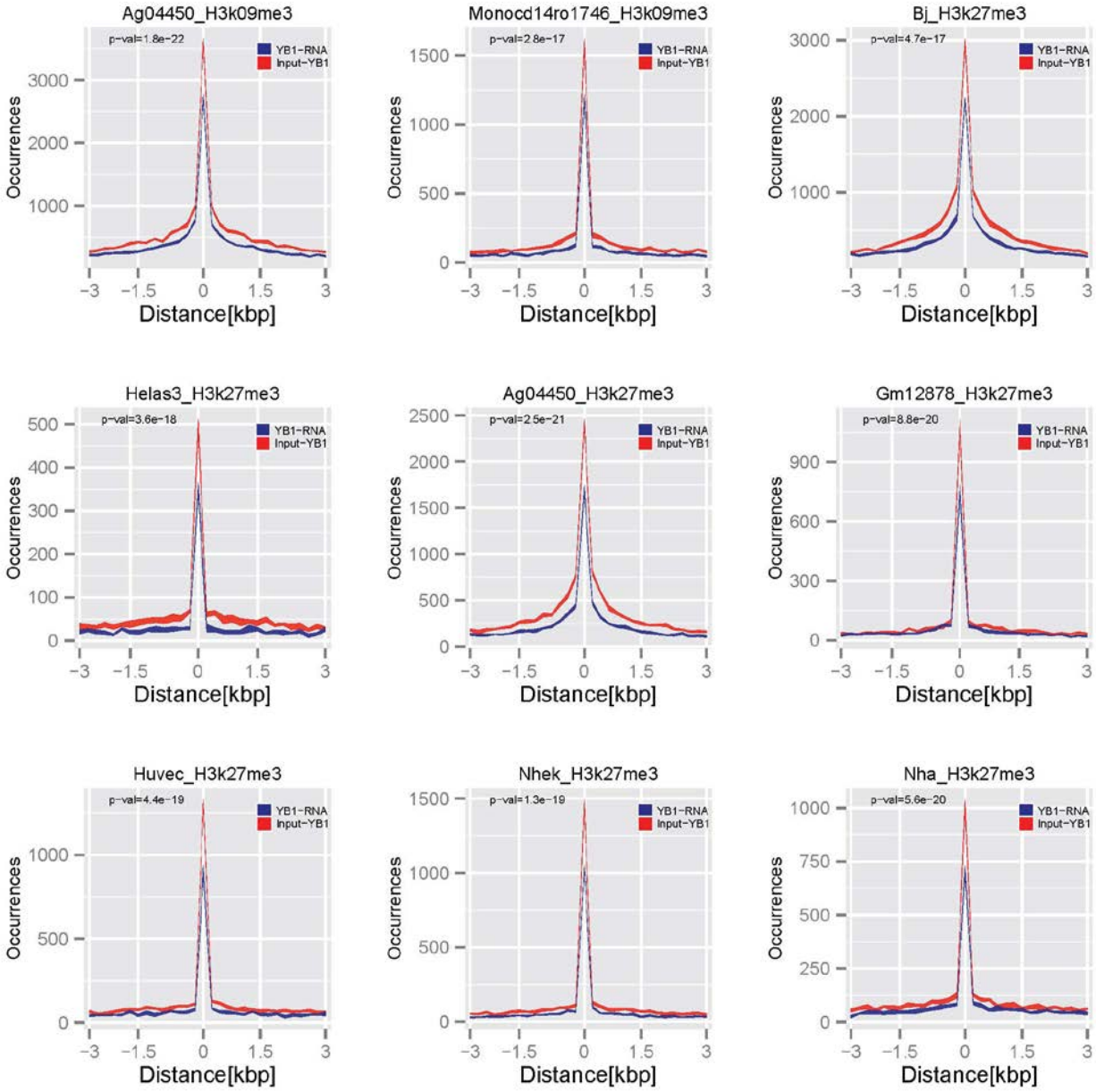

K562_H3k04me3
p-val=2.7e-19
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hsmmt_H3k36me3
p-val=9.8e-19
Gm12865_H3k04me3
p-val=7.6e-23
K562_H3k04me3
p-val=1.9e-21
Huvec_H3k36me3
p-val=4.6e-24
Huvec_H3k36me3
p-val=3.7e-29
Osteo_H4k20me1
p-val=3.6e-21
K562_H3k04me3
p-val=1.3e-22
Osteobl_H3k36me3
p-val=1.6e-21

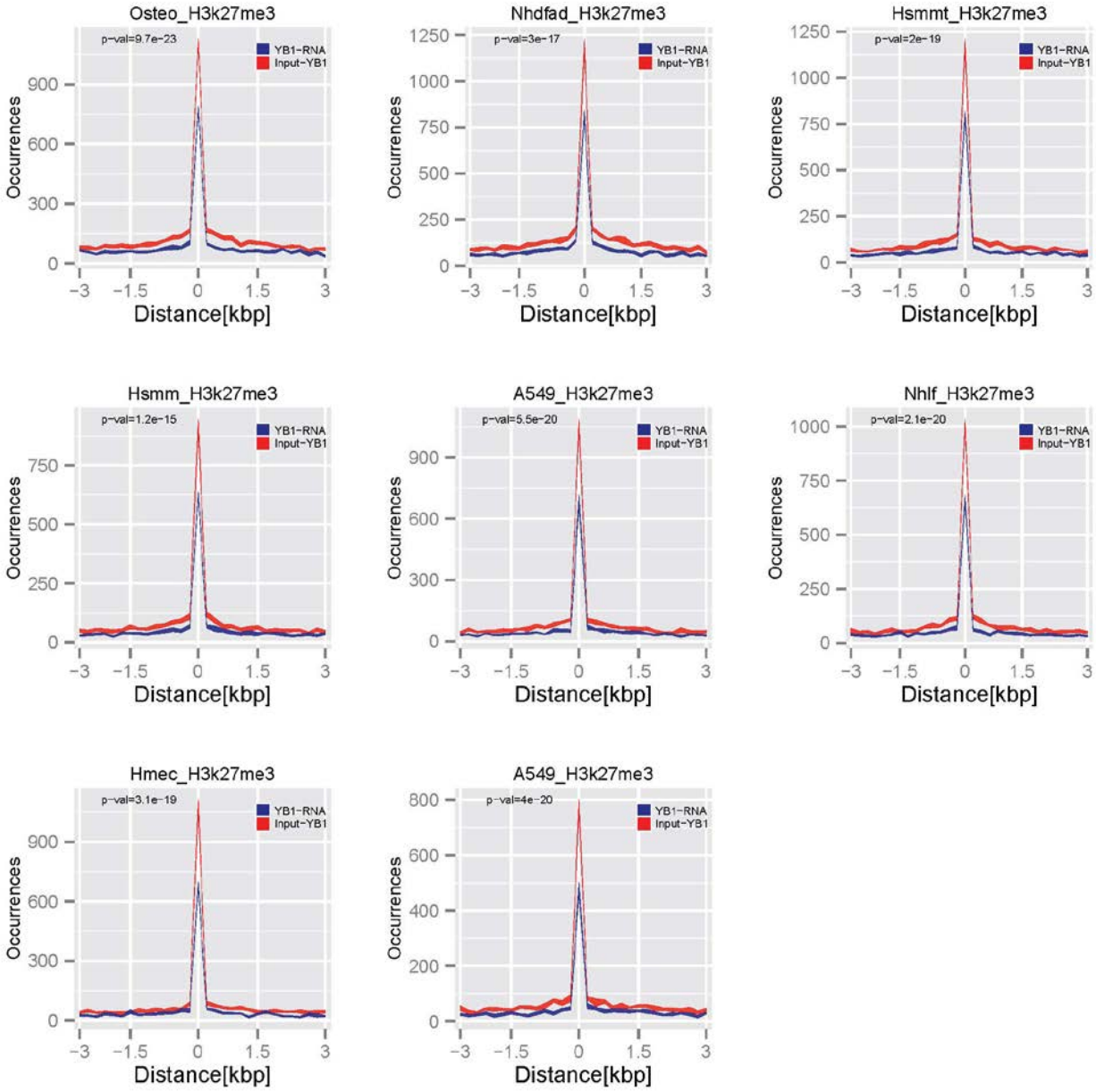

Panc1_H3k04me1b
p-val=1.3e-25
Gm06990_H3k36me3
p-val=1.5e-28
Hffmyc_H3k04me3
p-val=1.7e-20
A549_H3k36me3
p-val=1.4e-23
Hffmyc_H3k04me3
p-val=3.5e-22
Gm12864_H3k04me3
p-val=1.3e-21
Mcf7_H3k04me3
p-val=1.3e-21
Nhdfneo_H3k4me3
p-val=4.1e-21
K562_H3k04me3
p-val=2.5e-22
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]

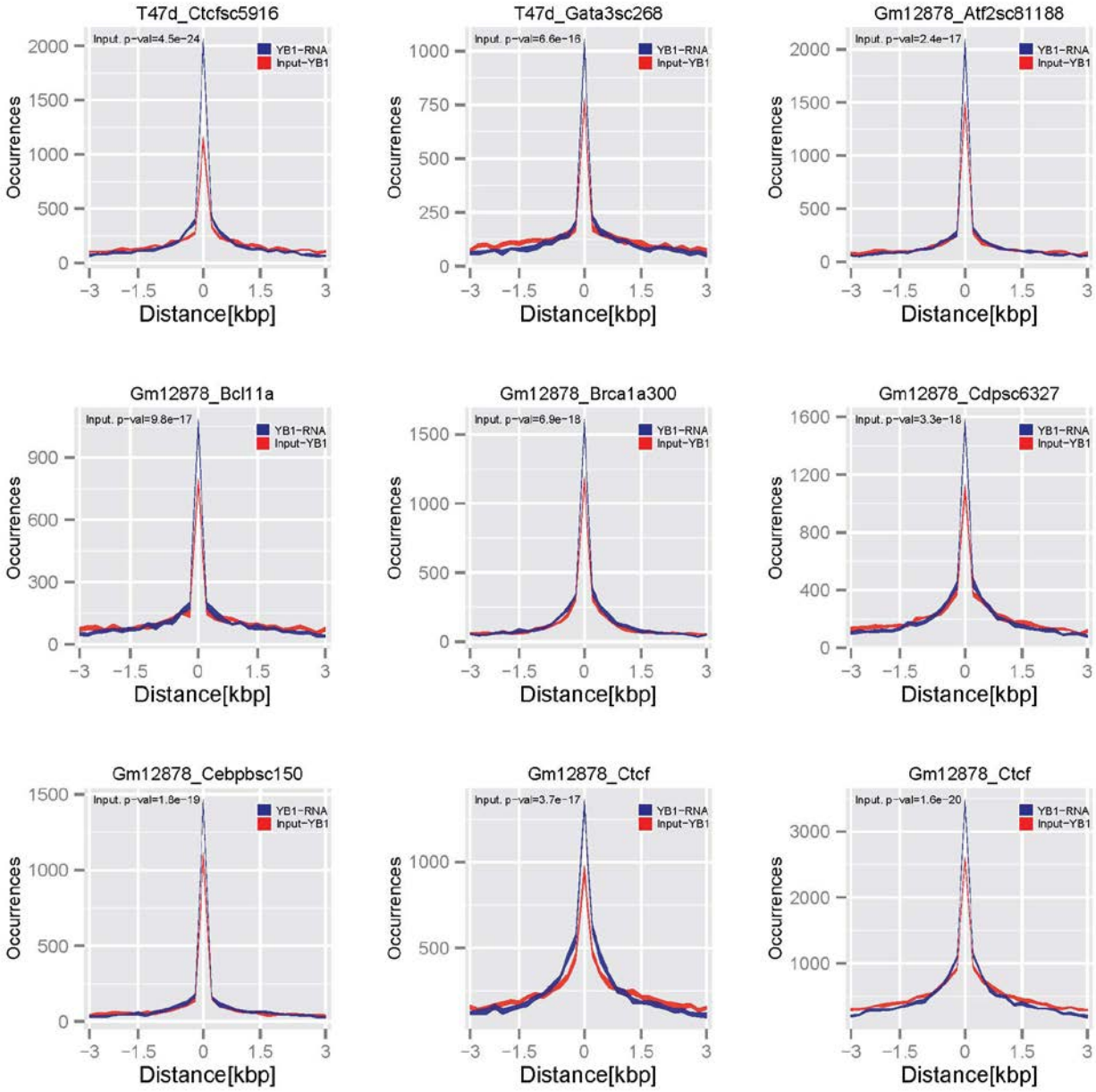

Ag04450_H3k27ac
p-val=9.7e-29
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
K562_H3k04me3
p-val=4.8e-21
Nhlf_H3k04me3
p-val=4.1e-22
Hsmmt_H4k20me1
p-val=2.3e-15
Mcf7_H3k27ac
p-val=3.8e-22
Aoaf_H3k4me3
p-val=5.7e-22
Monocd14ro1746_H4k20me1
p-val=1.1e-20
Be2c_H3k04me3
p-val=6.8e-19
A549_H3k04me3
p-val=3.1e-18

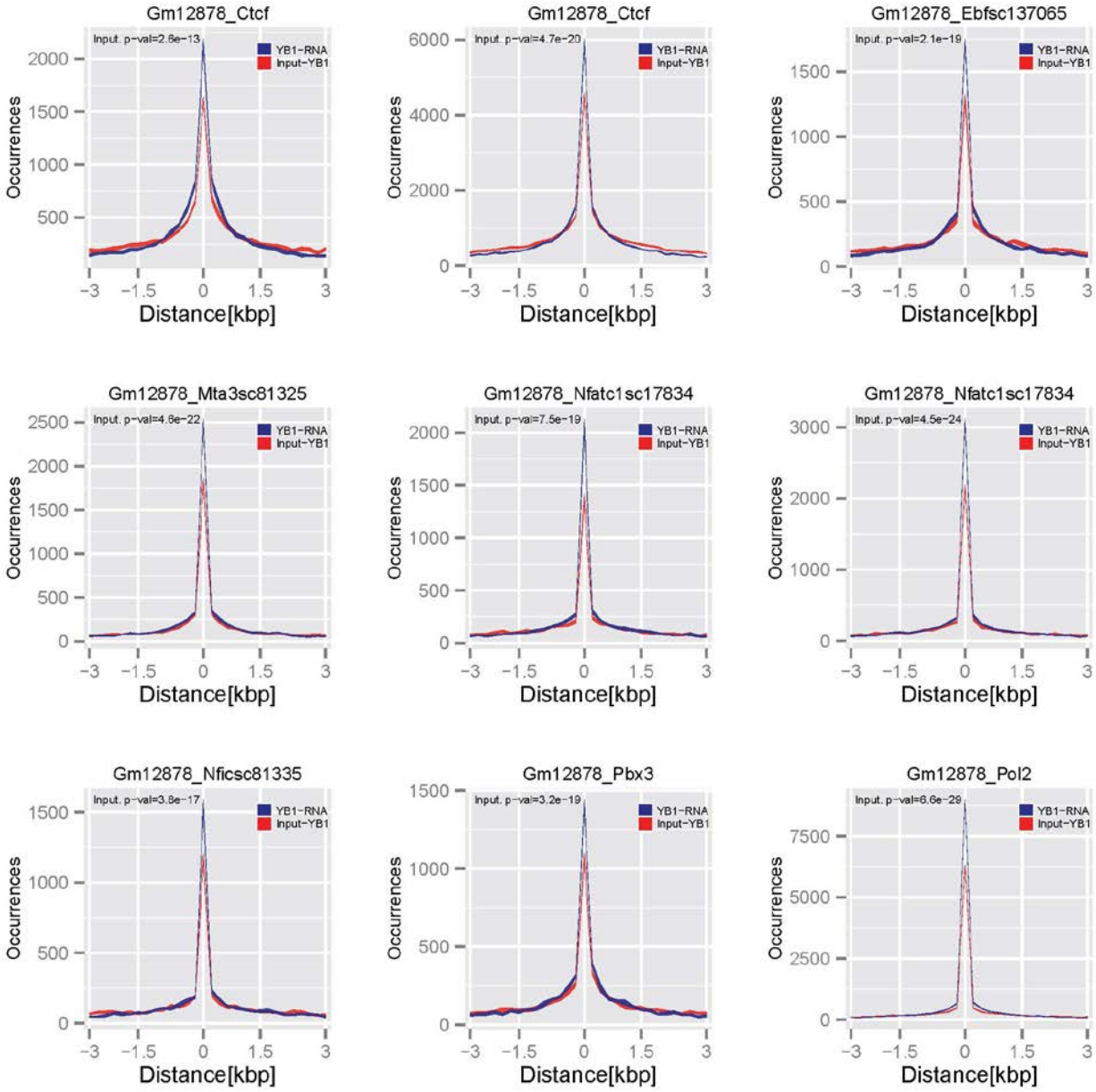

Panc1_H3k04me3
p-val=1.5e-21
Gm12878_H3k4me3
p-val=5.6e-22
A549_H3k04me3
p-val=2.5e-18
Ag04450_H3k4me3
p-val=7.4e-21
Be2c_H3k04me3
p-val=2.7e-19
Nhdfneo_H3k4me3
p-val=1.3e-20
Nhdfad_H4k20me1
p-val=2.7e-17
Hmf_H3k4me3
p-val=7.4e-22
Nhlf_H3k04me3
p-val=6e-21
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]

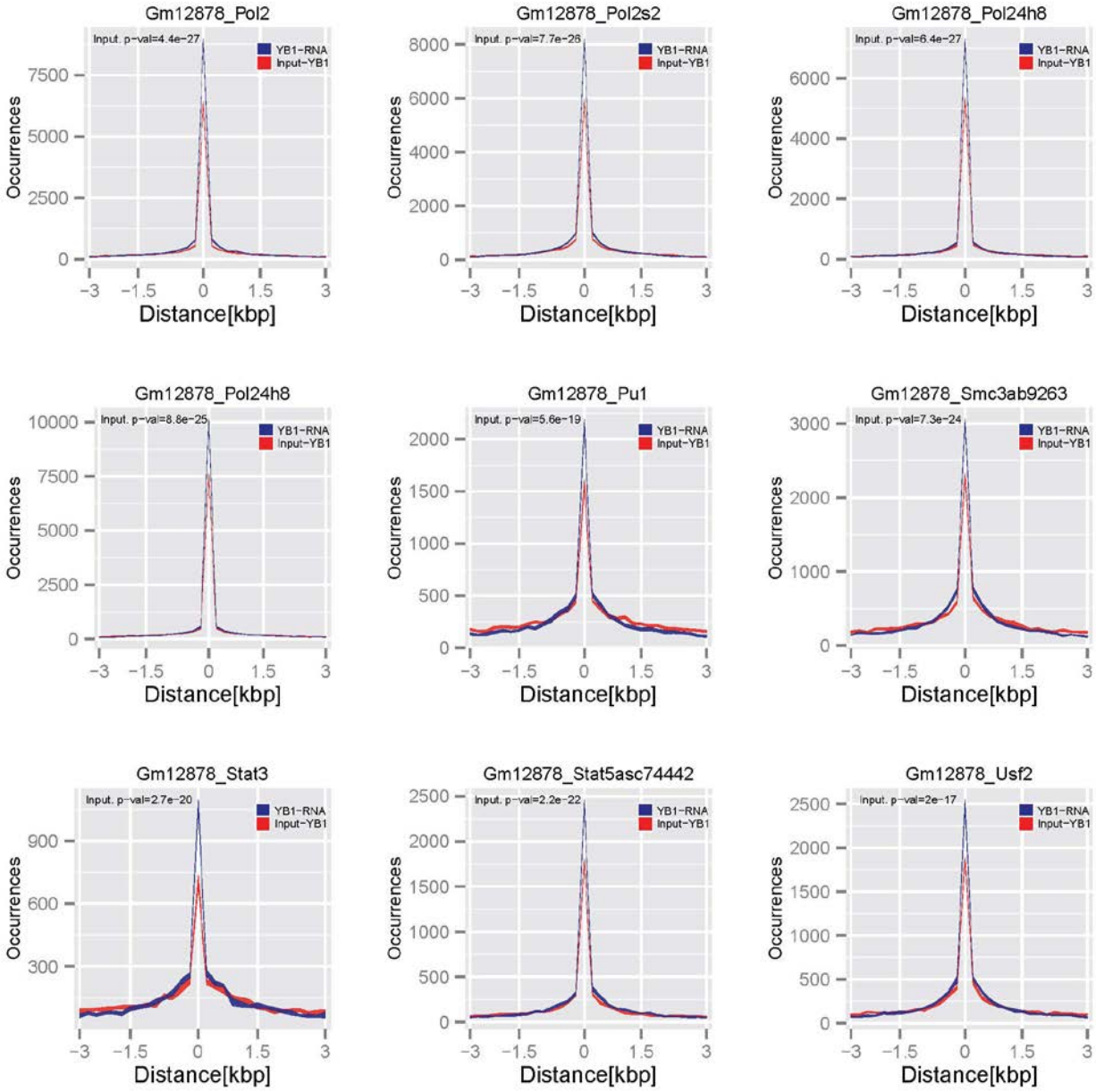

Nhlf_H3k4me1
p-val=2.7e-23
Mcf7_H3k4me3
p-val=2.4e-24
Sknshra_H3k27me3
p-val=2.9e-17
Hre_H3k27me3
p-val=2.4e-13
Hepg2_H3k09me3
p-val=1.7e-15
Gm12878_H3k27me3
p-val=3.9e-20
Huvec_H3k27me3
p-val=1.8e-20
Bj_H3k27me3
p-val=3.6e-18
Huvec_H3k27me3
p-val=1.3e-23
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]

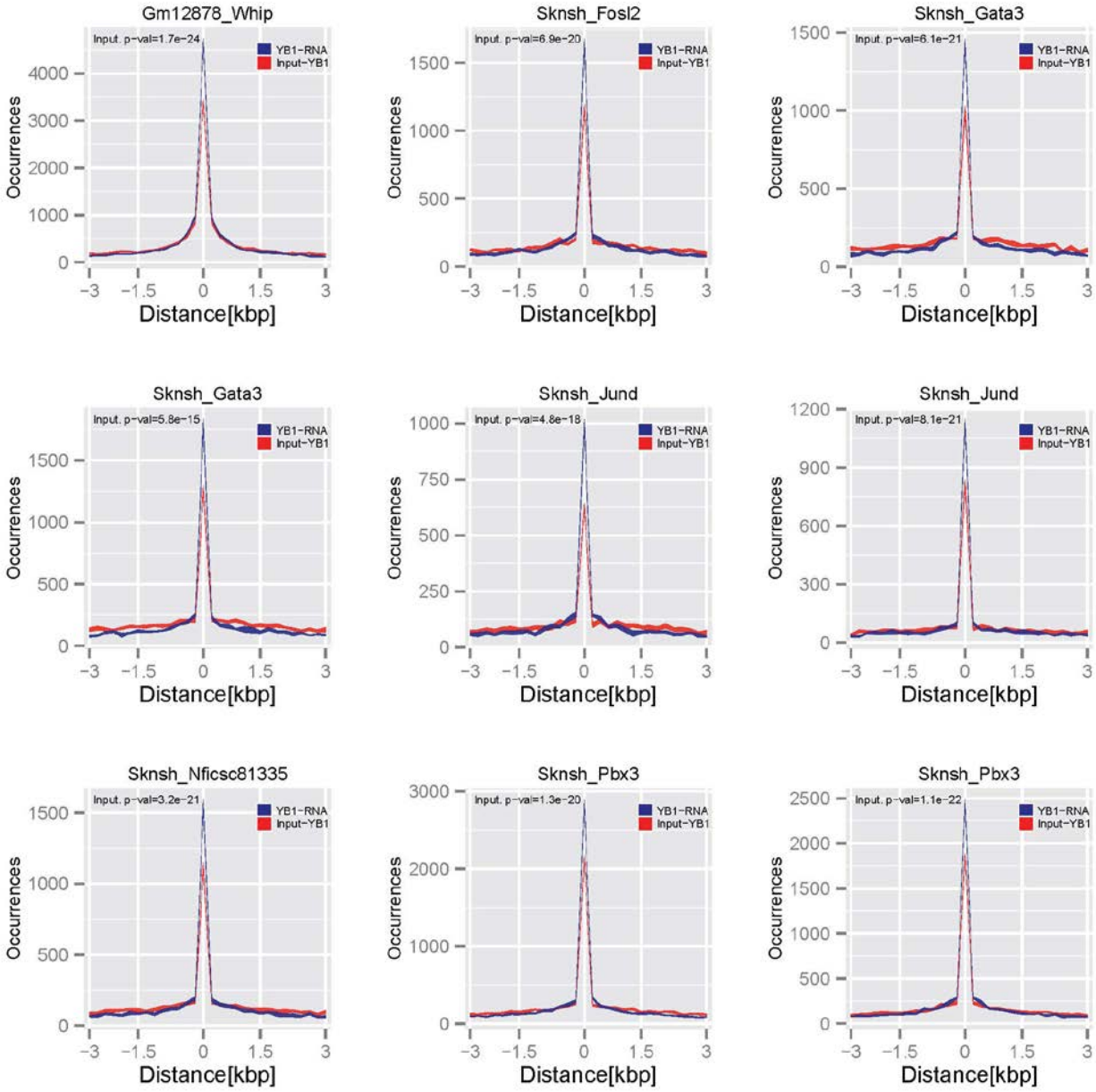

Ag04450_H3k09me3
p-val=1.8e-22
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Monocd14ro1746_H3k09me3
p-val=2.8e-17
Bj_H3k27me3
p-val=4.7e-17
Helas3_H3k27me3
p-val=3.6e-18
Ag04450_H3k27me3
p-val=2.5e-21
Gm12878_H3k27me3
p-val=8.8e-20
Huvec_H3k27me3
p-val=4.4e-19
Nhek_H3k27me3
p-val=1.3e-19
Nha_H3k27me3
p-val=5.6e-20

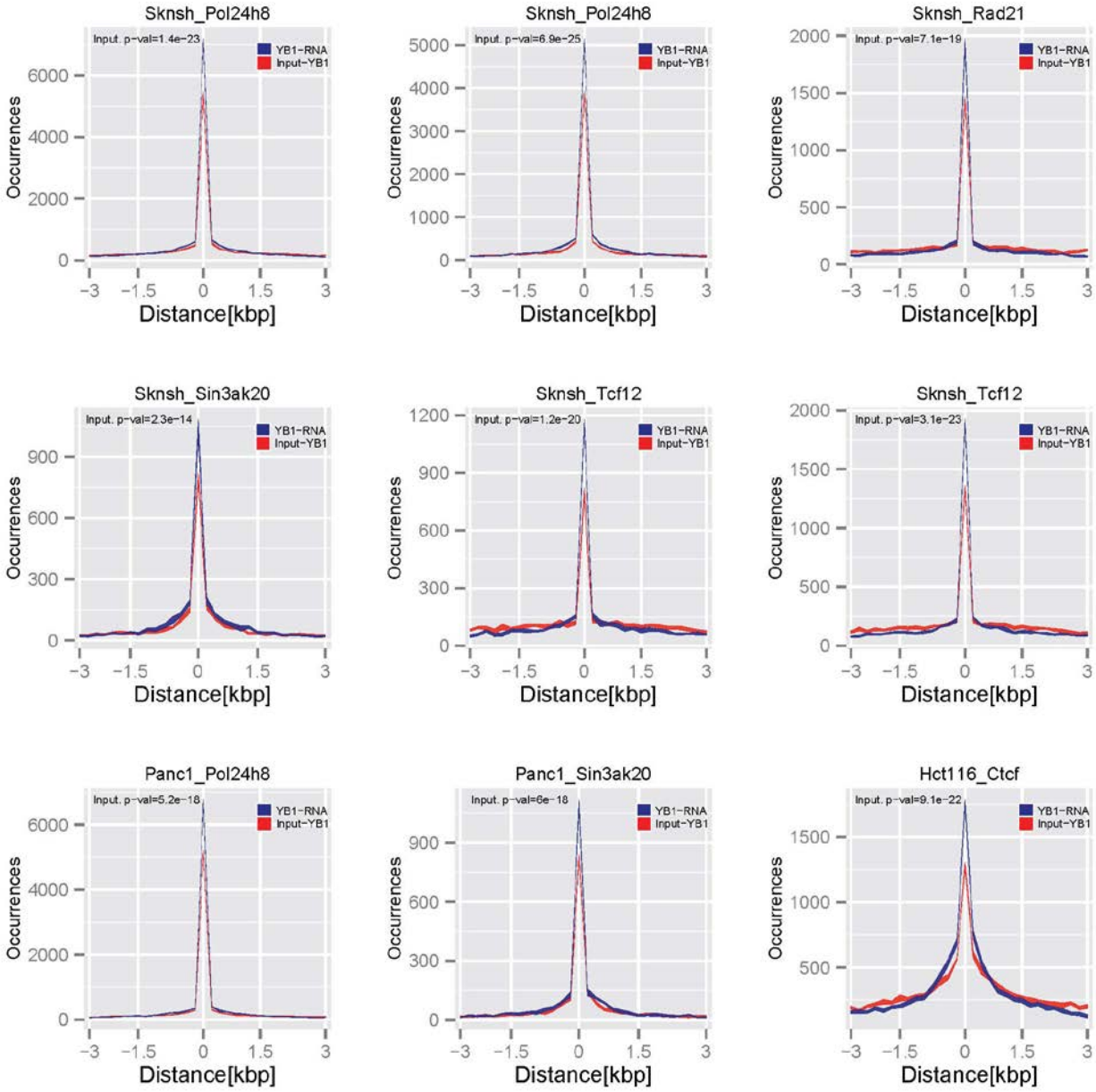

Osteo_H3k27me3
p-val=9.7e-23
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Nhdfad_H3k27me3
p-val=3e-17
Hsmmt_H3k27me3
p-val=2e-19
Hsmm_H3k27me3
p-val=1.2e-15
A549_H3k27me3
p-val=5.5e-20
Nhlf_H3k27me3
p-val=2.1e-20
Hmec_H3k27me3
p-val=3.1e-19
A549_H3k27me3
p-val=4e-20

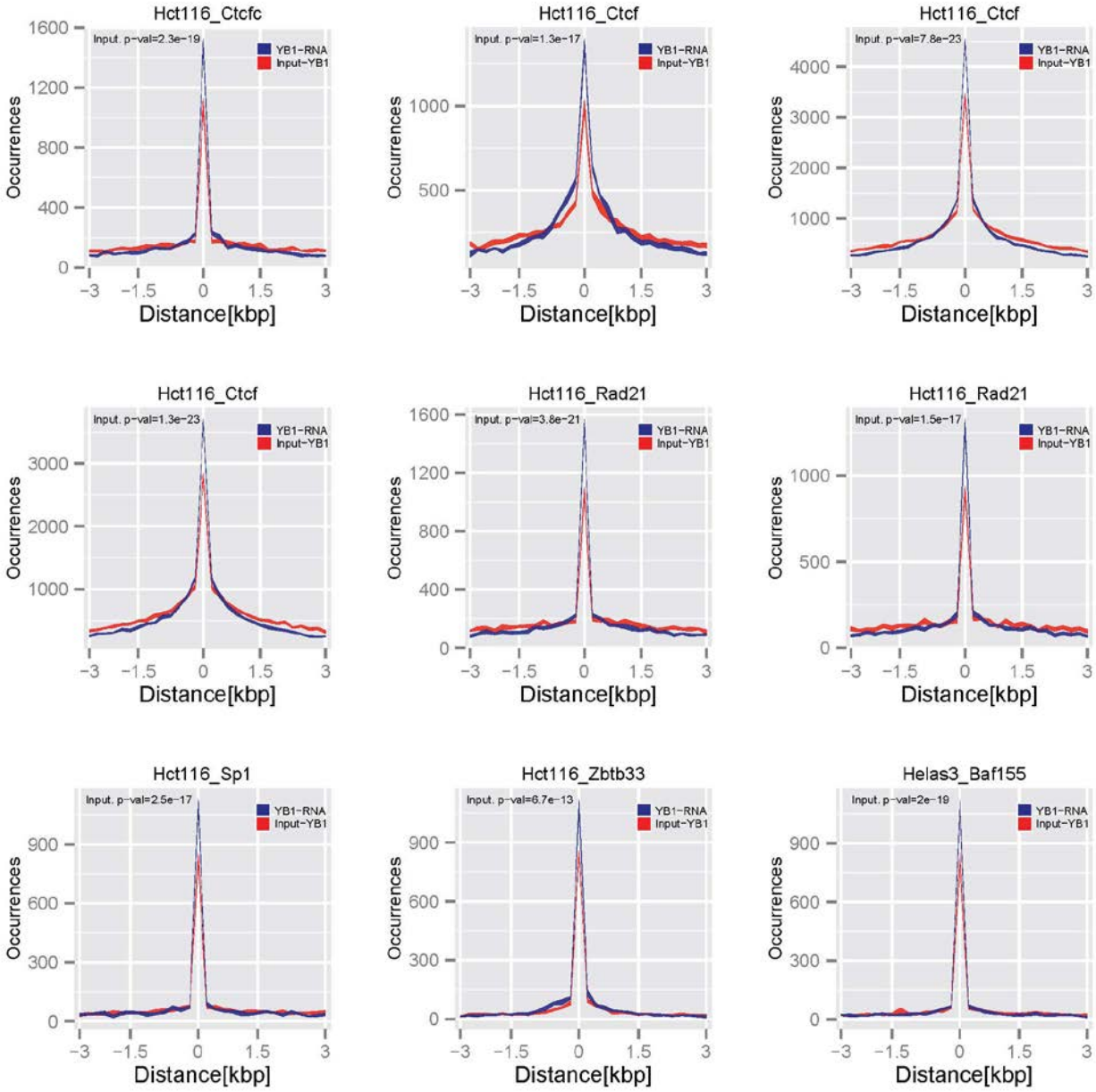

T47d_Ctcfsc5916
Input. p-val=4.5e-24
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
T47d_Gata3sc268
Input. p-val=6.6e-16
Gm12878_Atf2sc81188
Input. p-val=2.4e-17
Gm12878_Bcl11a
Input. p-val=9.8e-17
Gm12878_Brca1a300
Input. p-val=6.9e-18
Gm12878_Cdpsc6327
Input. p-val=3.3e-18
Gm12878_Cebpbsc150
Input. p-val=1.8e-19
Gm12878_Ctcf
Input. p-val=3.7e-17
Gm12878_Ctcf
Input. p-val=1.6e-20

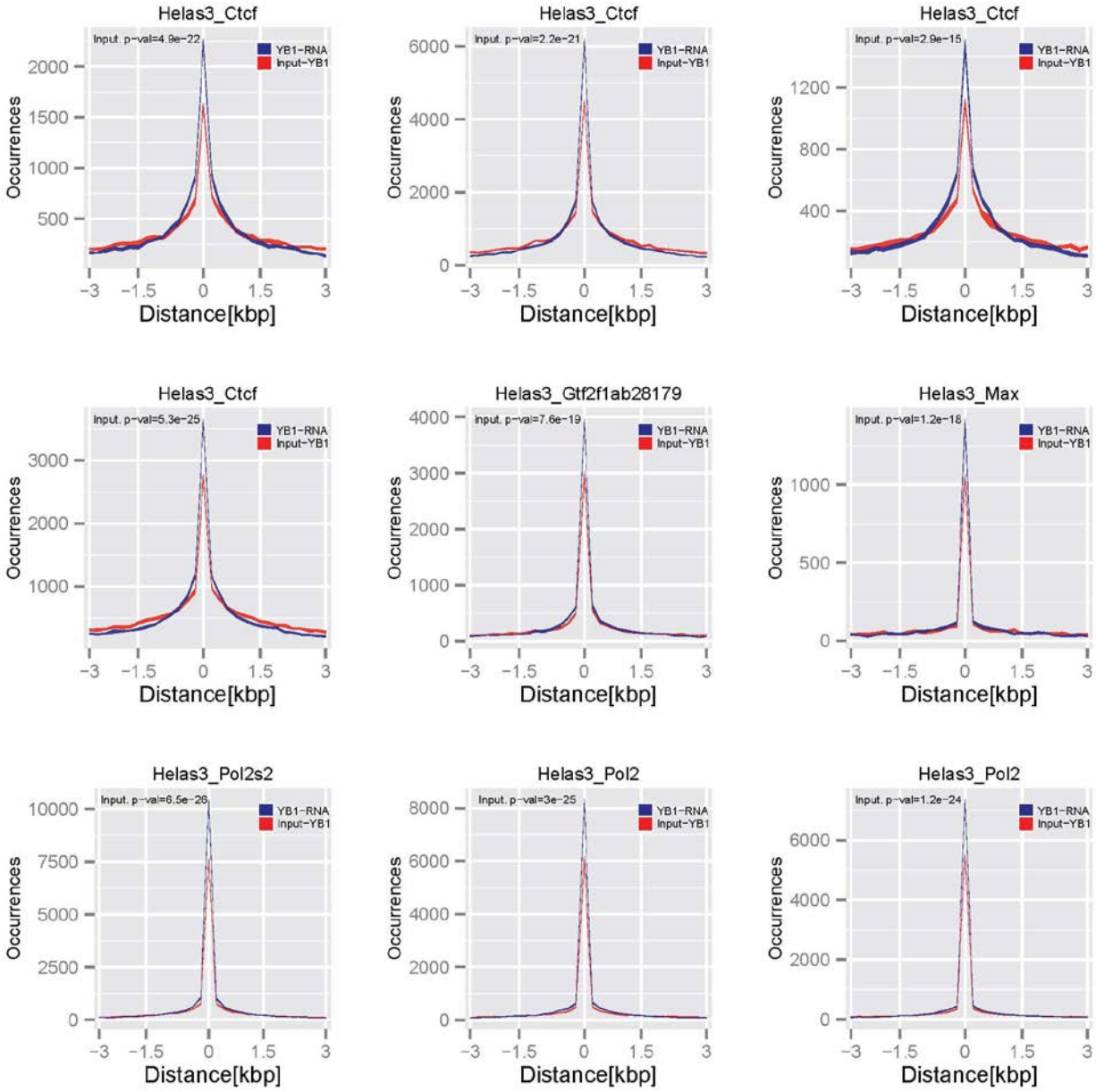

Gm12878_Ctcf
Input. p-val=2.6e-13
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Gm12878_Ctcf
Input. p-val=4.7e-20
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Gm12878_Ebfsc137065
Input. p-val=2.1e-19
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Gm12878_Mta3sc81325
Input. p-val=4.6e-22
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Gm12878_Nfatc1sc17834
Input. p-val=7.5e-19
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Gm12878_Nfatc1sc17834
Input. p-val=4.5e-24
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Gm12878_Nficsc81335
Input. p-val=3.8e-17
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Gm12878_Pbx3
Input. p-val=3.2e-19
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Gm12878_Pol2
Input. p-val=6.6e-29
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]

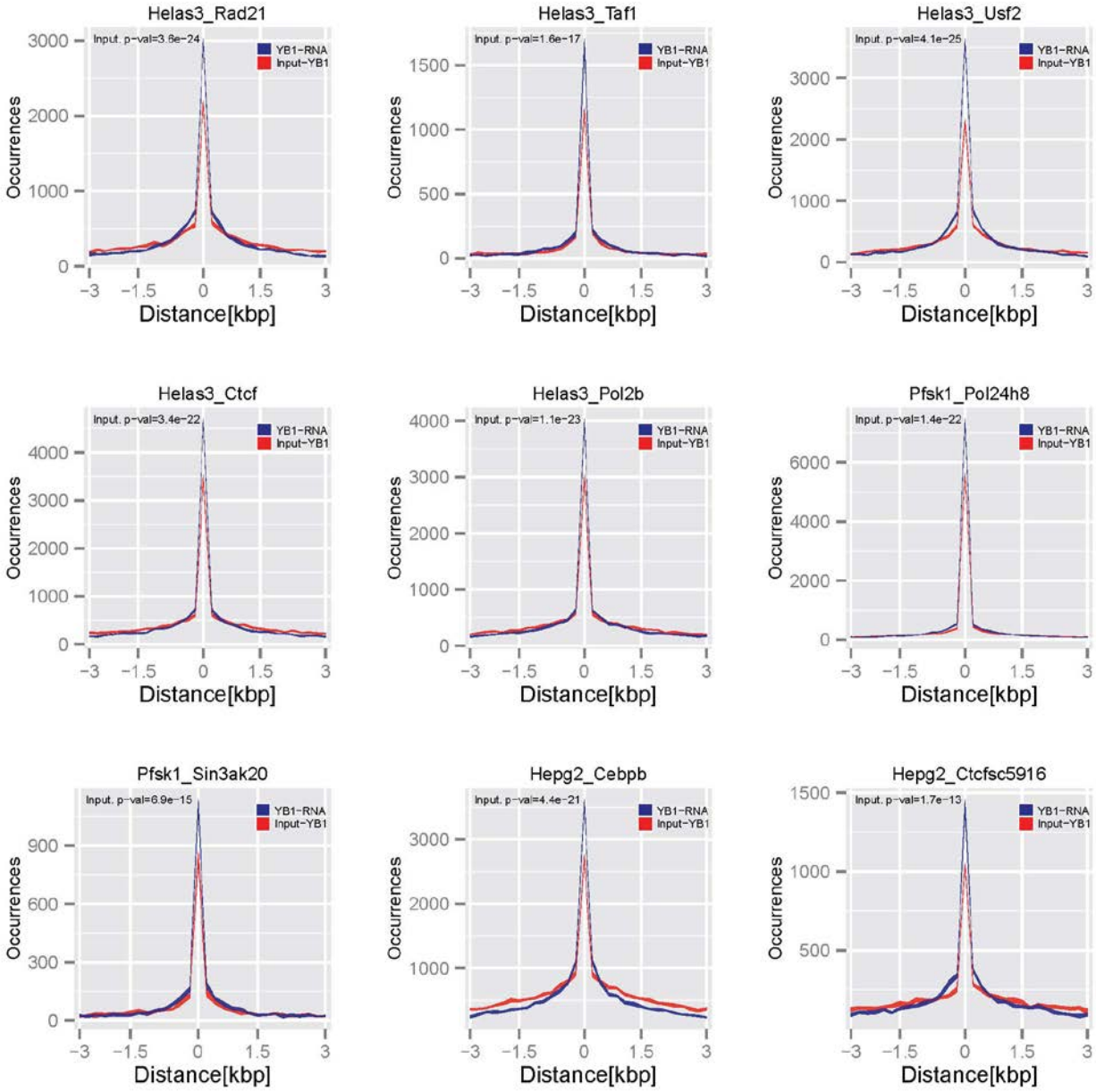

Gm12878_Pol2
Input. p-val=4.4e-27
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Gm12878_Pol2s2
Input. p-val=7.7e-26
Gm12878_Pol24h8
Input. p-val=6.4e-27
Gm12878_Pol24h8
Input. p-val=8.8e-25
Gm12878_Pu1
Input. p-val=5.6e-19
Gm12878_Smc3ab9263
Input. p-val=7.3e-24
Gm12878_Stat3
Input. p-val=2.7e-20
Gm12878_Stat5asc74442
Input. p-val=2.2e-22
Gm12878_Usf2
Input. p-val=2e-17

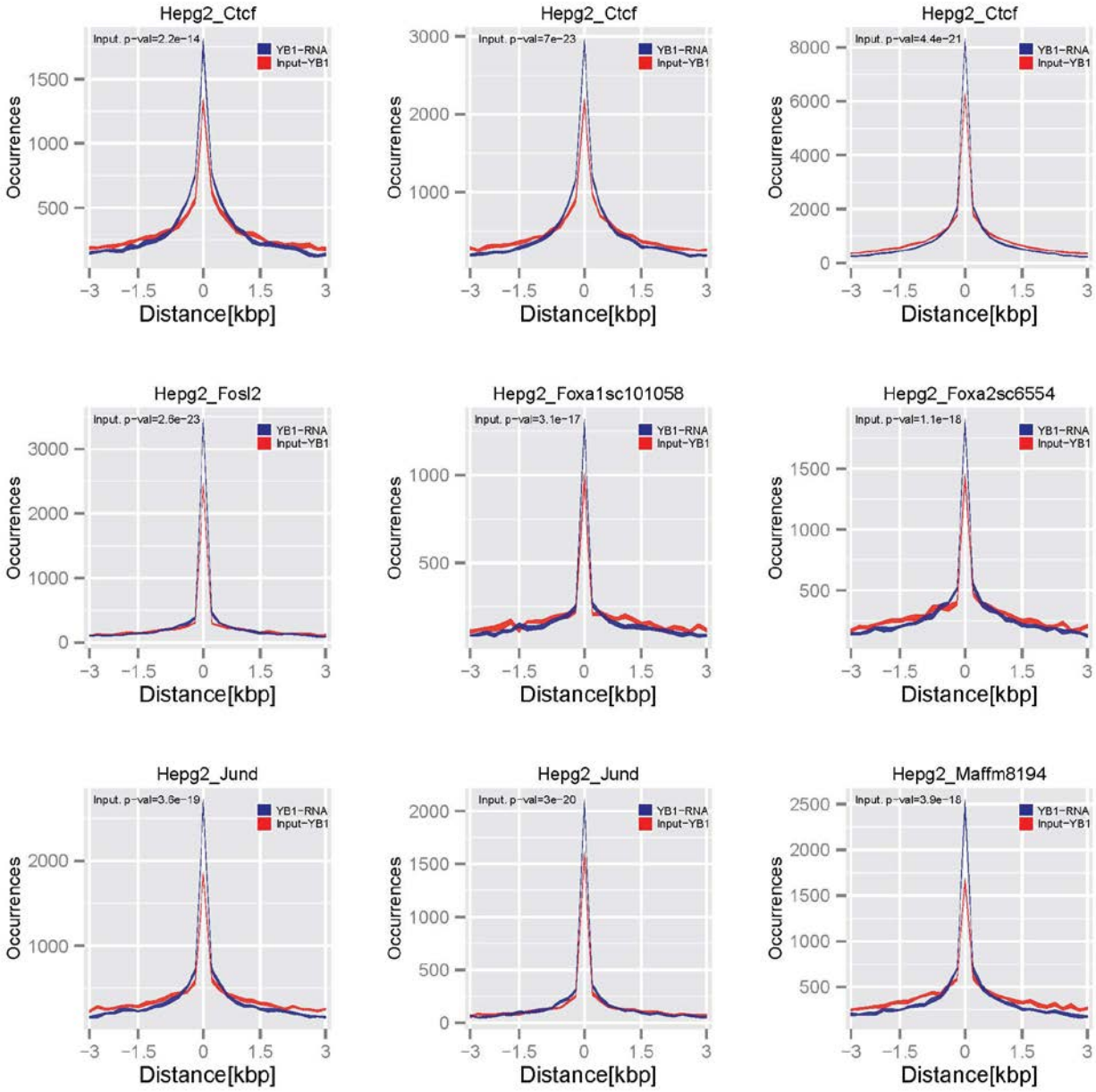

Gm12878_Whip
Input. p-val=1.7e-24
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Sknsh_Fosl2
Input. p-val=6.9e-20
Sknsh_Gata3
Input. p-val=6.1e-21
Sknsh_Gata3
Input. p-val=5.8e-15
Sknsh_Jund
Input. p-val=4.8e-18
Sknsh_Jund
Input. p-val=8.1e-21
Sknsh_Nficsc81335
Input. p-val=3.2e-21
Sknsh_Pbx3
Input. p-val=1.3e-20
Sknsh_Pbx3
Input. p-val=1.1e-22

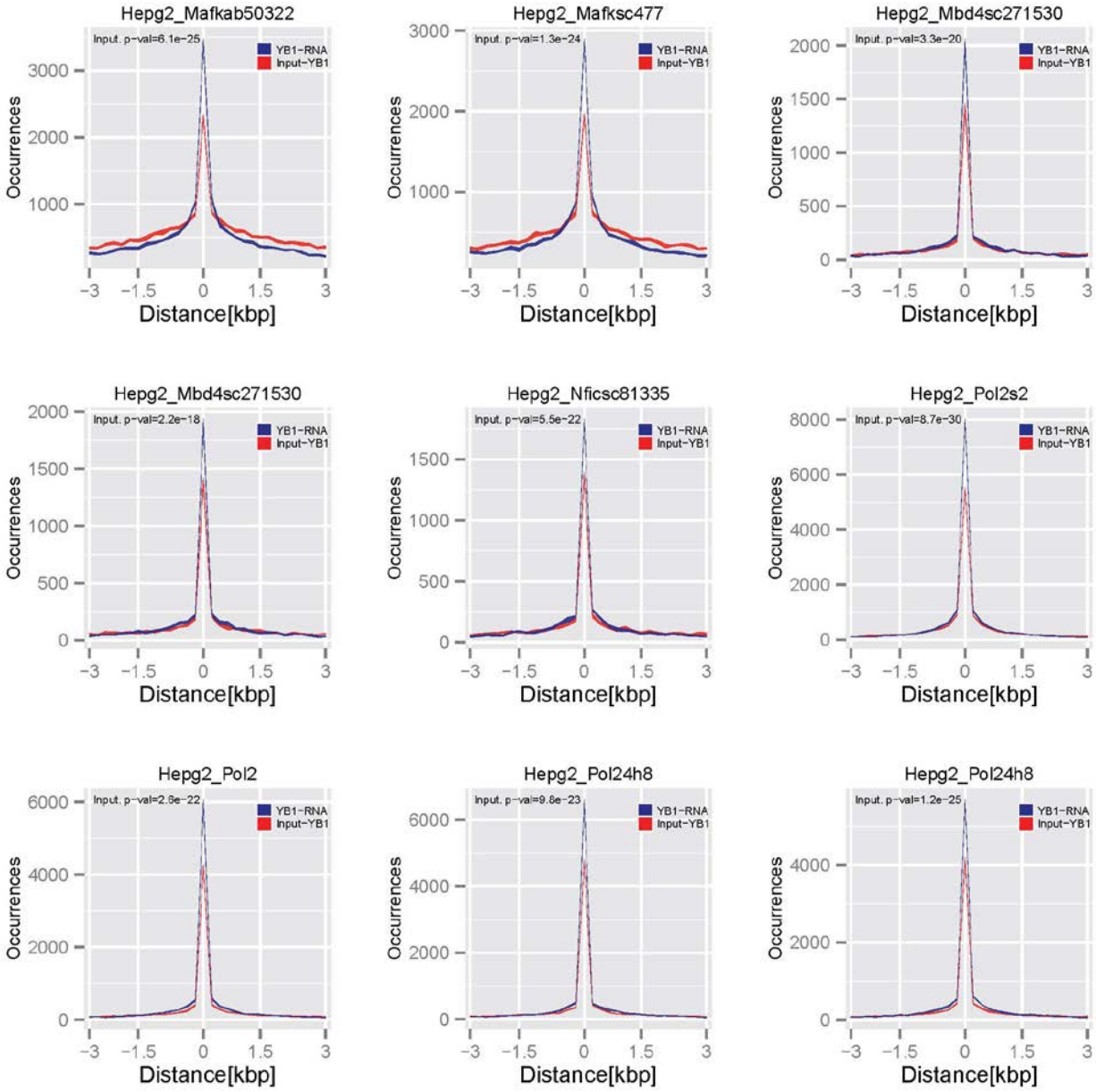

Sknsh_Pol24h8
Input. p-val=1.4e-23
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Sknsh_Pol24h8
Input. p-val=6.9e-25
Sknsh_Rad21
Input. p-val=7.1e-19
Sknsh_Sin3ak20
Input. p-val=2.3e-14
Sknsh_Tcf12
Input. p-val=1.2e-20
Sknsh_Tcf12
Input. p-val=3.1e-23
Panc1_Pol24h8
Input. p-val=5.2e-18
Panc1_Sin3ak20
Input. p-val=6e-18
Hct116_Ctcf
Input. p-val=9.1e-22

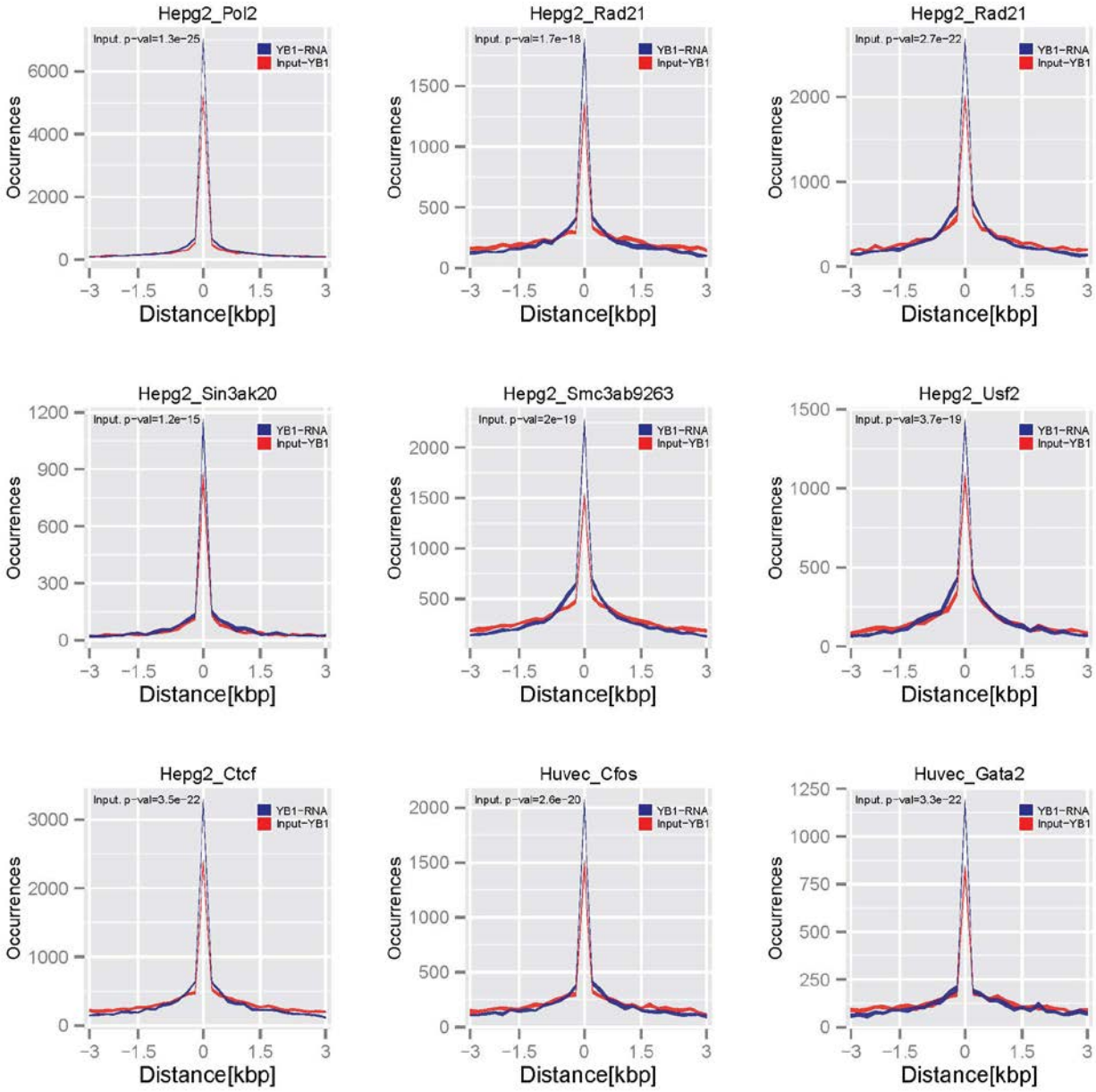

Hct116_Ctcfc
Input. p-val=2.3e-19
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hct116_Ctcf
Input. p-val=1.3e-17
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hct116_Ctcf
Input. p-val=7.8e-23
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hct116_Ctcf
Input. p-val=1.3e-23
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hct116_Rad21
Input. p-val=3.8e-21
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hct116_Rad21
Input. p-val=1.5e-17
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hct116_Sp1
Input. p-val=2.5e-17
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hct116_Zbtb33
Input. p-val=6.7e-13
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Helas3_Baf155
Input. p-val=2e-19
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]

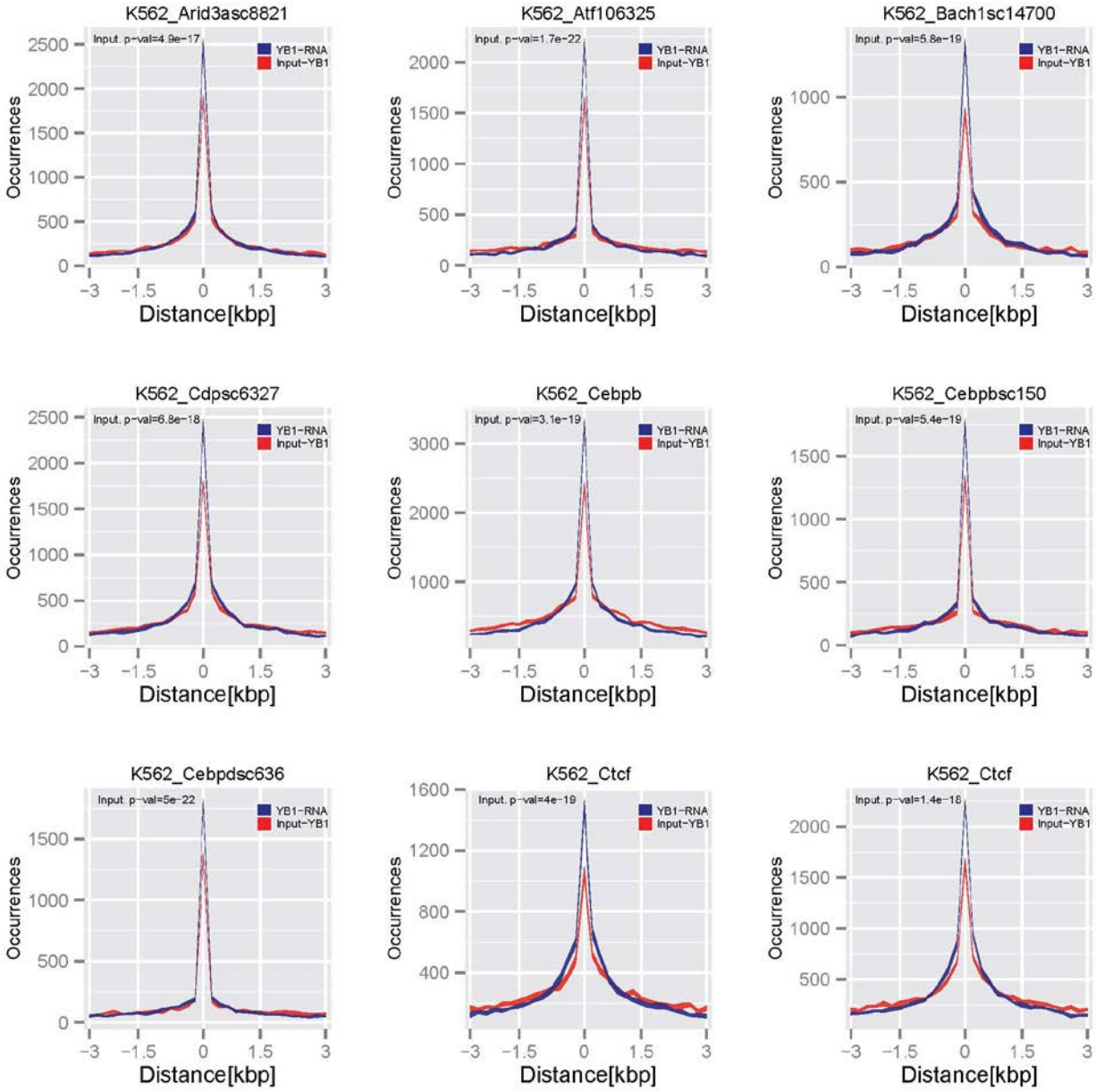

Helas3_Ctcf
Input. p-val=4.9e-22
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Helas3_Ctcf
Input. p-val=2.2e-21
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Helas3_Ctcf
Input. p-val=2.9e-15
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Helas3_Ctcf
Input. p-val=5.3e-25
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Helas3_Gtf2f1ab28179
Input. p-val=7.6e-19
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Helas3_Max
Input. p-val=1.2e-18
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Helas3_Pol2s2
Input. p-val=6.5e-26
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Helas3_Pol2
Input. p-val=3e-25
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Helas3_Pol2
Input. p-val=1.2e-24
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]

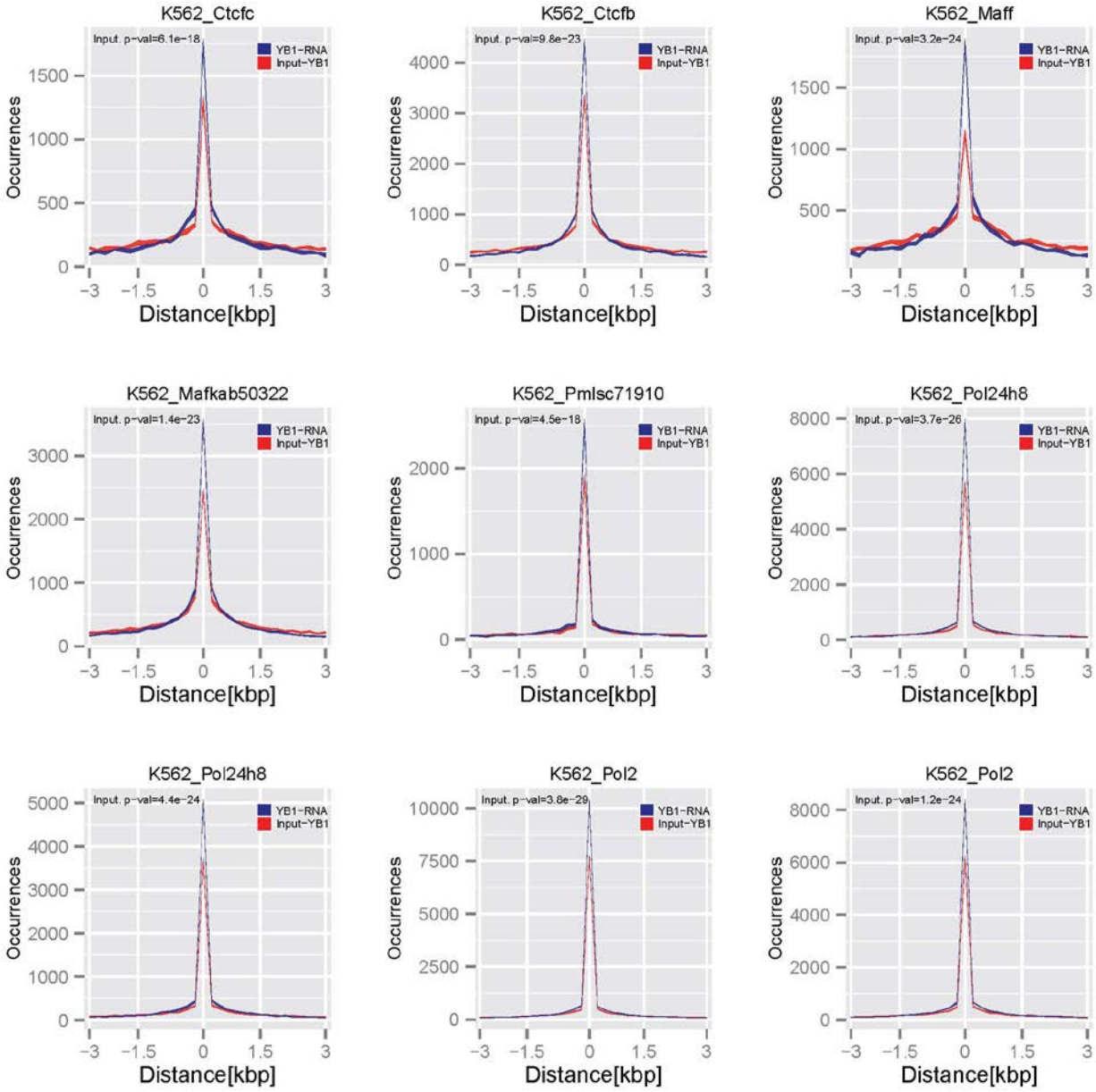

Helas3_Rad21
Input. p-val=3.6e-24
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Helas3_Taf1
Input. p-val=1.6e-17
Helas3_Usf2
Input. p-val=4.1e-25
Helas3_Ctcf
Input. p-val=3.4e-22
Helas3_Pol2b
Input. p-val=1.1e-23
Pfsk1_Pol24h8
Input. p-val=1.4e-22
Pfsk1_Sin3ak20
Input. p-val=6.9e-15
Hepg2_Cebpb
Input. p-val=4.4e-21
Hepg2_Ctcfsc5916
Input. p-val=1.7e-13

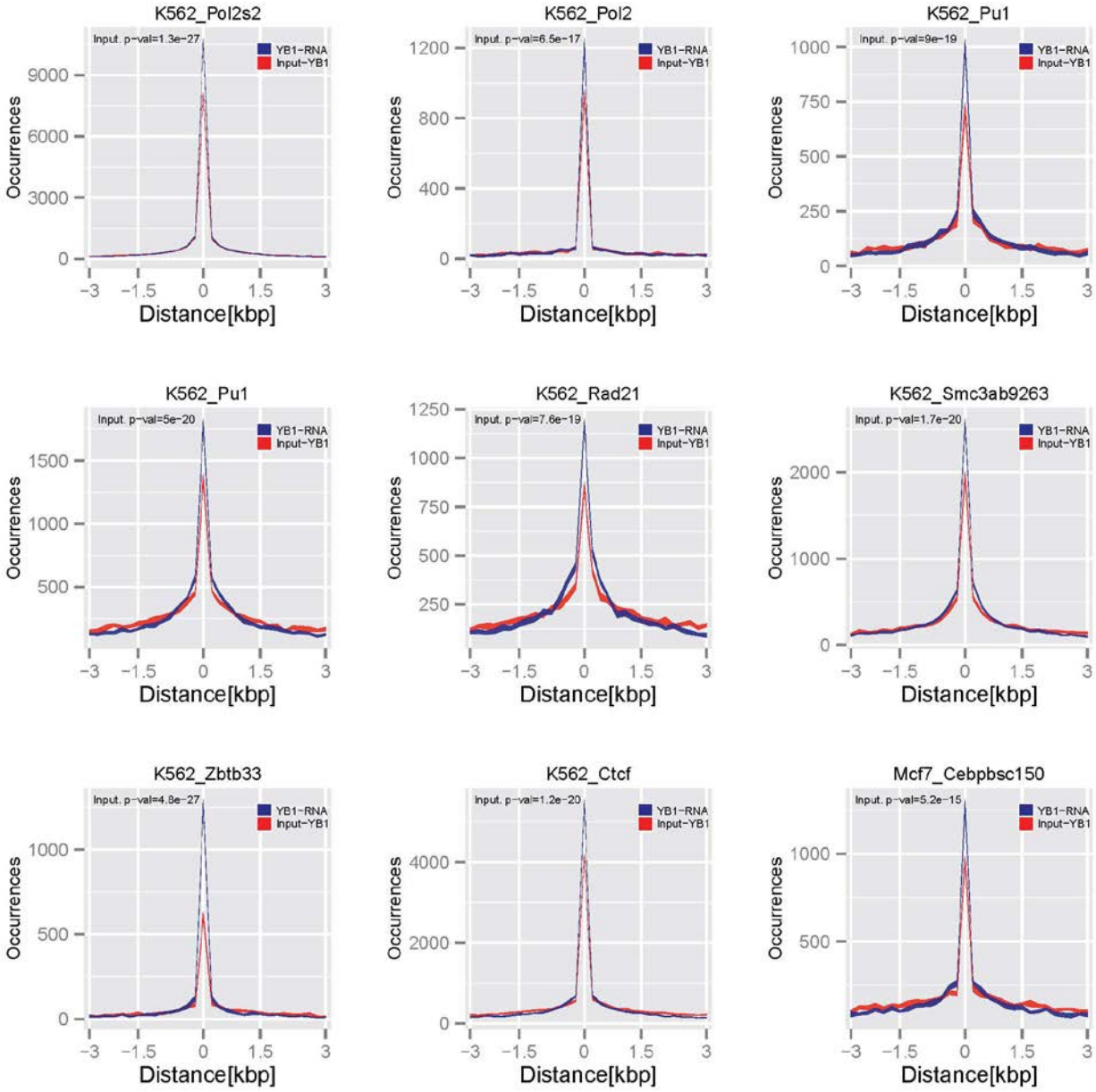

Hepg2_Ctcf
Input. p-val=2.2e-14
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Ctcf
Input. p-val=7e-23
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Ctcf
Input. p-val=4.4e-21
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Fosl2
Input. p-val=2.6e-23
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Foxa1sc101058
Input. p-val=3.1e-17
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Foxa2sc6554
Input. p-val=1.1e-18
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Jund
Input. p-val=3.6e-19
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Jund
Input. p-val=3e-20
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Maffm8194
Input. p-val=3.9e-18
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]

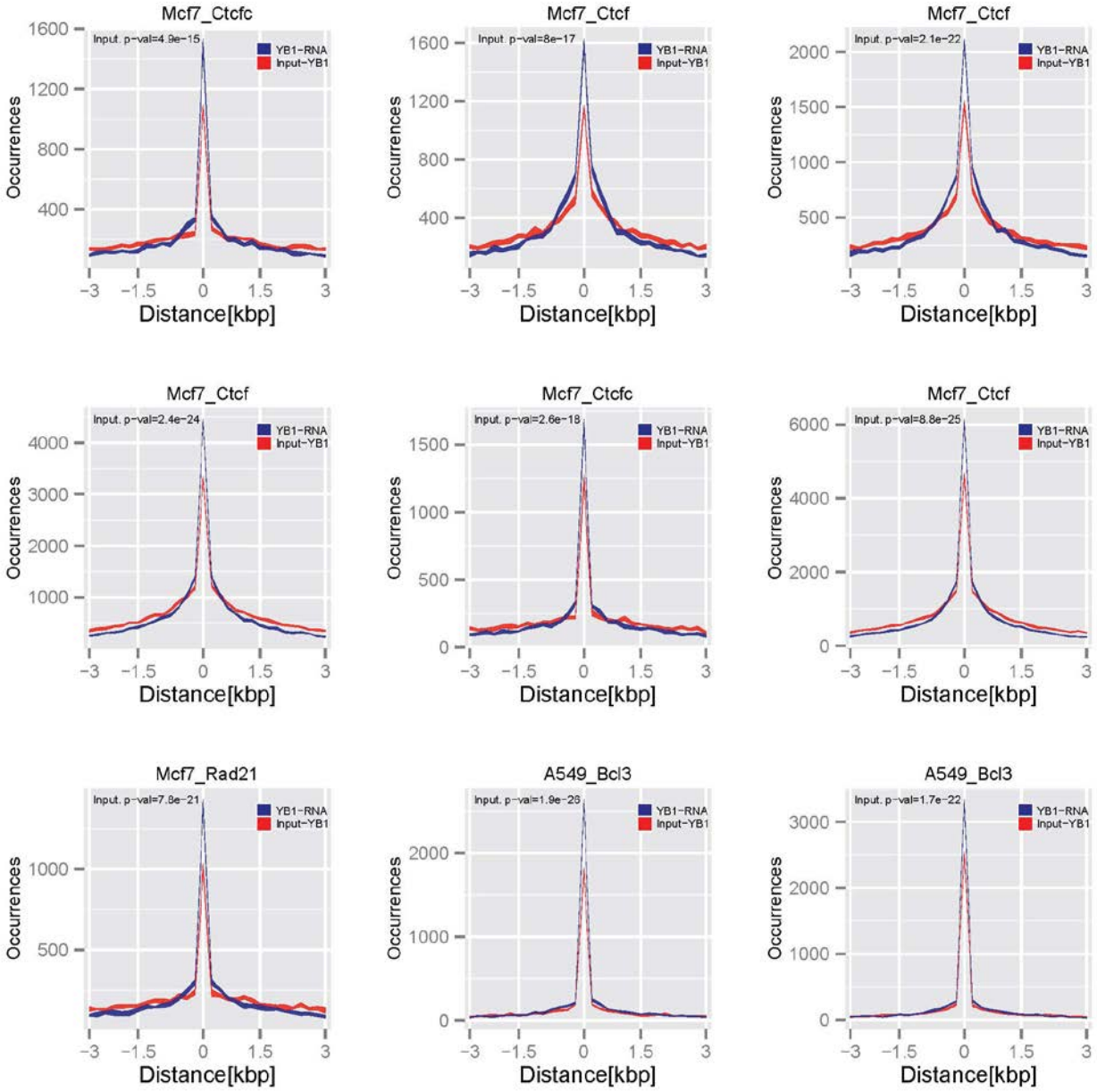

Hepg2_Mafkab50322
Input. p-val=6.1e-25
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Mafksc477
Input. p-val=1.3e-24
Hepg2_Mbd4sc271530
Input. p-val=3.3e-20
Hepg2_Mbd4sc271530
Input. p-val=2.2e-18
Hepg2_Nficsc81335
Input. p-val=5.5e-22
Hepg2_Pol2s2
Input. p-val=8.7e-30
Hepg2_Pol2
Input. p-val=2.6e-22
Hepg2_Pol24h8
Input. p-val=9.8e-23
Hepg2_Pol24h8
Input. p-val=1.2e-25

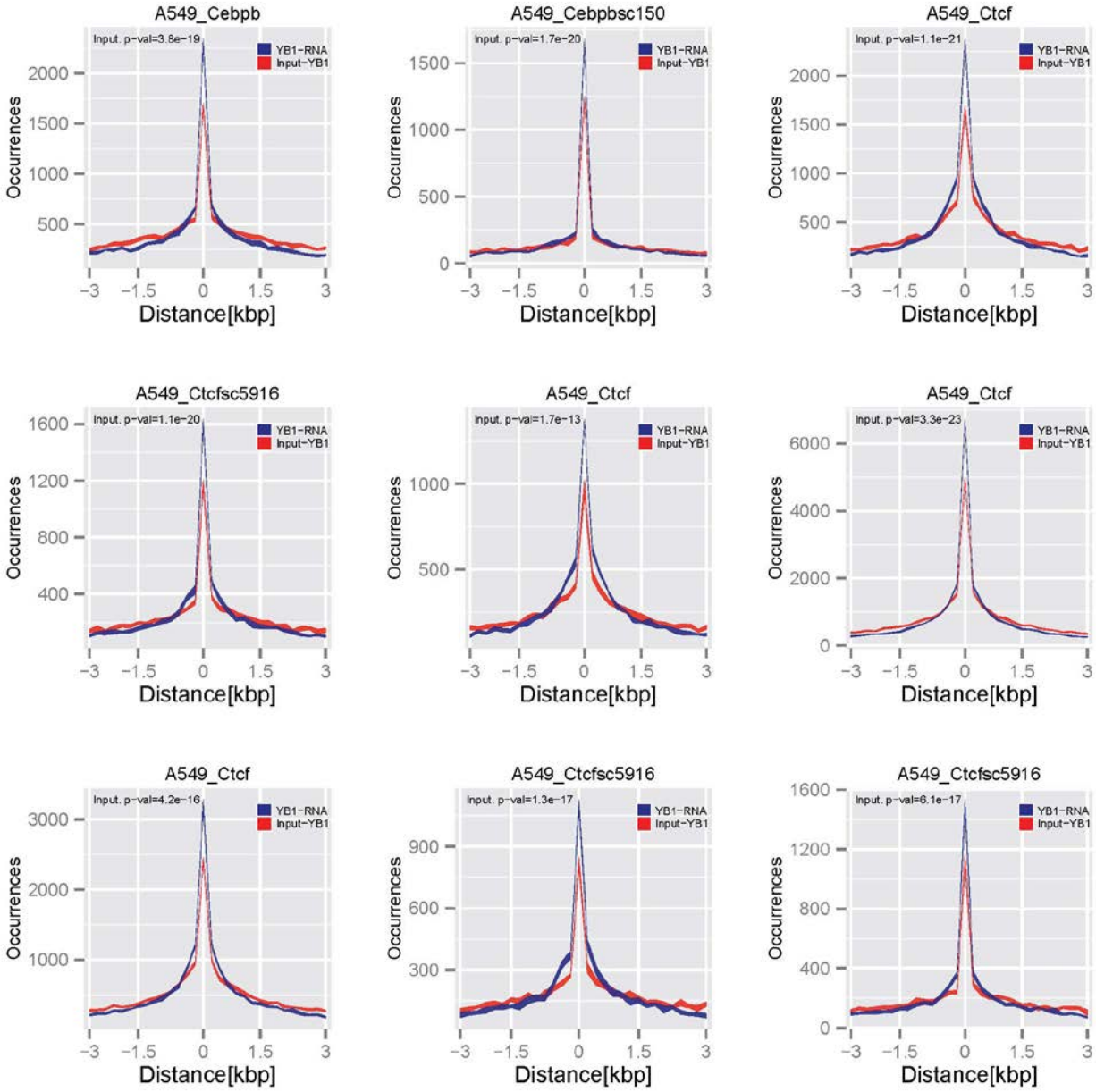

Hepg2_Pol2
Input. p-val=1.3e-25
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Rad21
Input. p-val=1.7e-18
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Rad21
Input. p-val=2.7e-22
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Sin3ak20
Input. p-val=1.2e-15
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Smc3ab9263
Input. p-val=2e-19
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Usf2
Input. p-val=3.7e-19
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Hepg2_Ctcf
Input. p-val=3.5e-22
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Huvec_Cfos
Input. p-val=2.6e-20
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Huvec_Gata2
Input. p-val=3.3e-22
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]

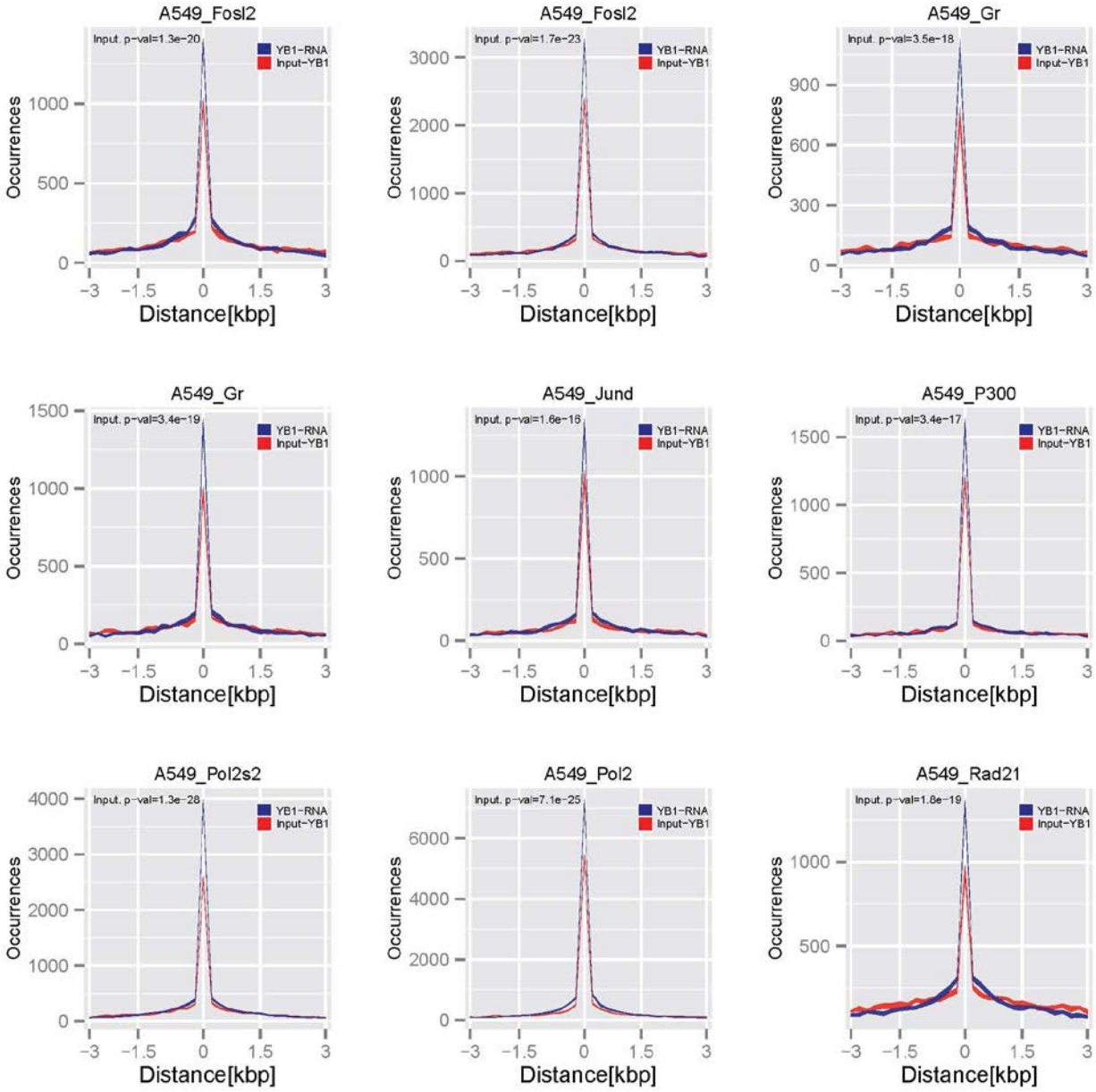

K562_Arid3asc8821
Input. p-val=4.9e-17
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
K562_Atf106325
Input. p-val=1.7e-22
K562_Bach1sc14700
Input. p-val=5.8e-19
K562_Cdpsc6327
Input. p-val=6.8e-18
K562_Cebpb
Input. p-val=3.1e-19
K562_Cebpbsc150
Input. p-val=5.4e-19
K562_Cebpdsc636
Input. p-val=5e-22
K562_Ctcf
Input. p-val=4e-19
K562_Ctcf
Input. p-val=1.4e-18

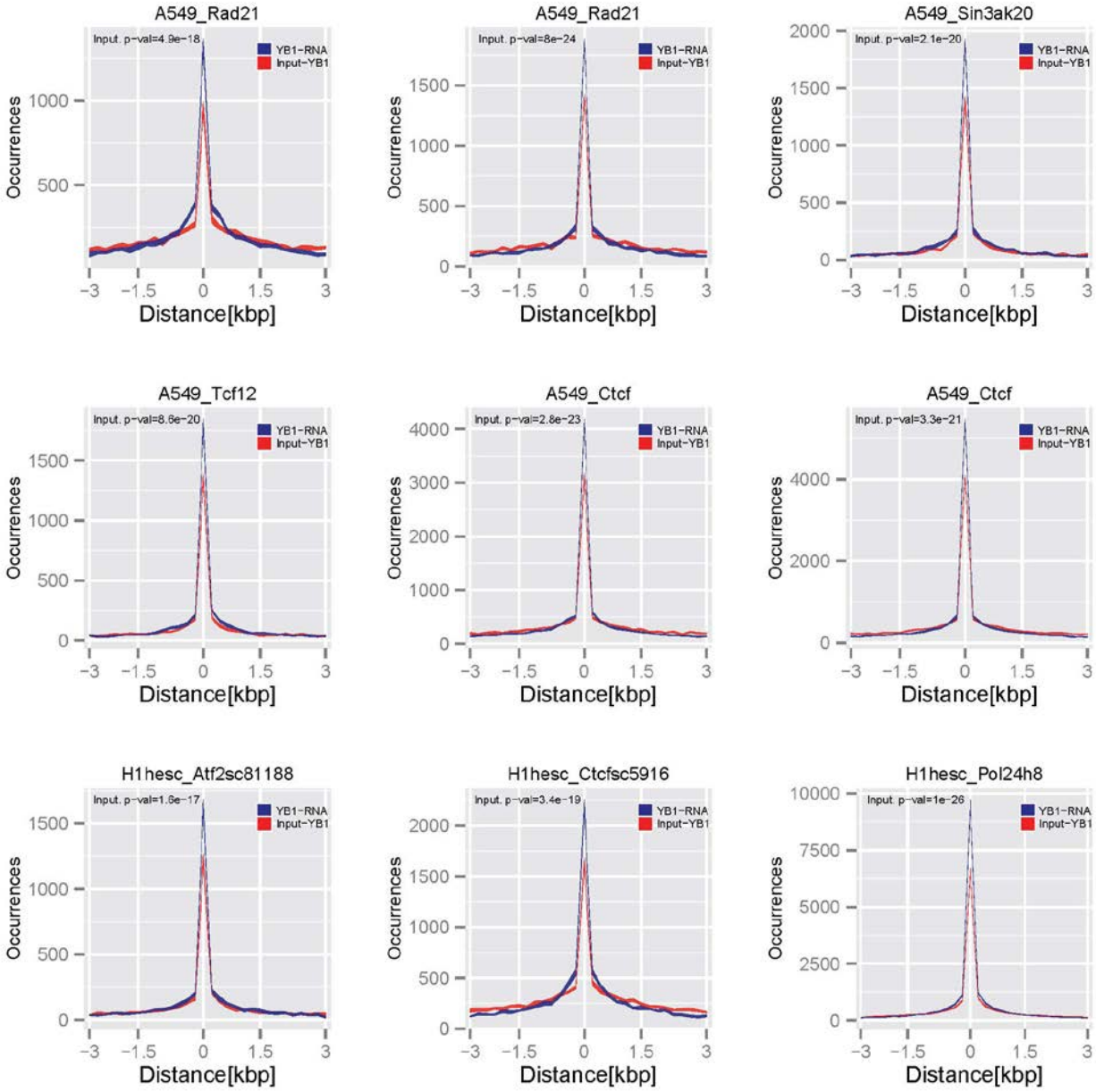

K562_Ctcfc
Input. p-val=6.1e-18
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
K562_Ctcfb
Input. p-val=9.8e-23
K562_Maff
Input. p-val=3.2e-24
K562_Mafkab50322
Input. p-val=1.4e-23
K562_Pmlsc71910
Input. p-val=4.5e-18
K562_Pol24h8
Input. p-val=3.7e-26
K562_Pol24h8
Input. p-val=4.4e-24
K562_Pol2
Input. p-val=3.8e-29
K562_Pol2
Input. p-val=1.2e-24

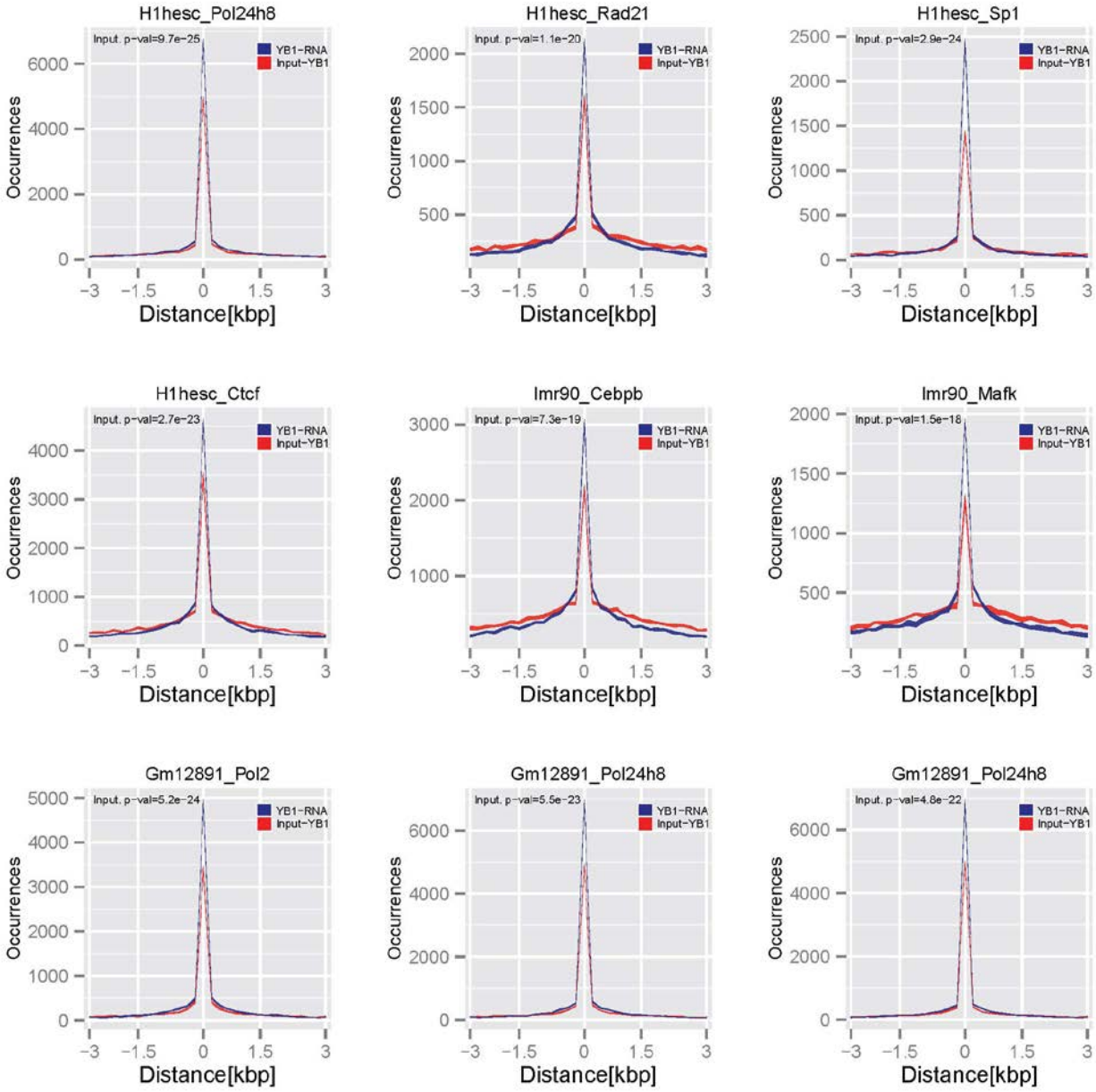

K562_Pol2s2
Input. p-val=1.3e-27
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
K562_Pol2
Input. p-val=6.5e-17
K562_Pu1
Input. p-val=9e-19
K562_Pu1
Input. p-val=5e-20
K562_Rad21
Input. p-val=7.6e-19
K562_Smc3ab9263
Input. p-val=1.7e-20
K562_Zbtb33
Input. p-val=4.8e-27
K562_Ctcf
Input. p-val=1.2e-20
Mcf7_Cebpbsc150
Input. p-val=5.2e-15

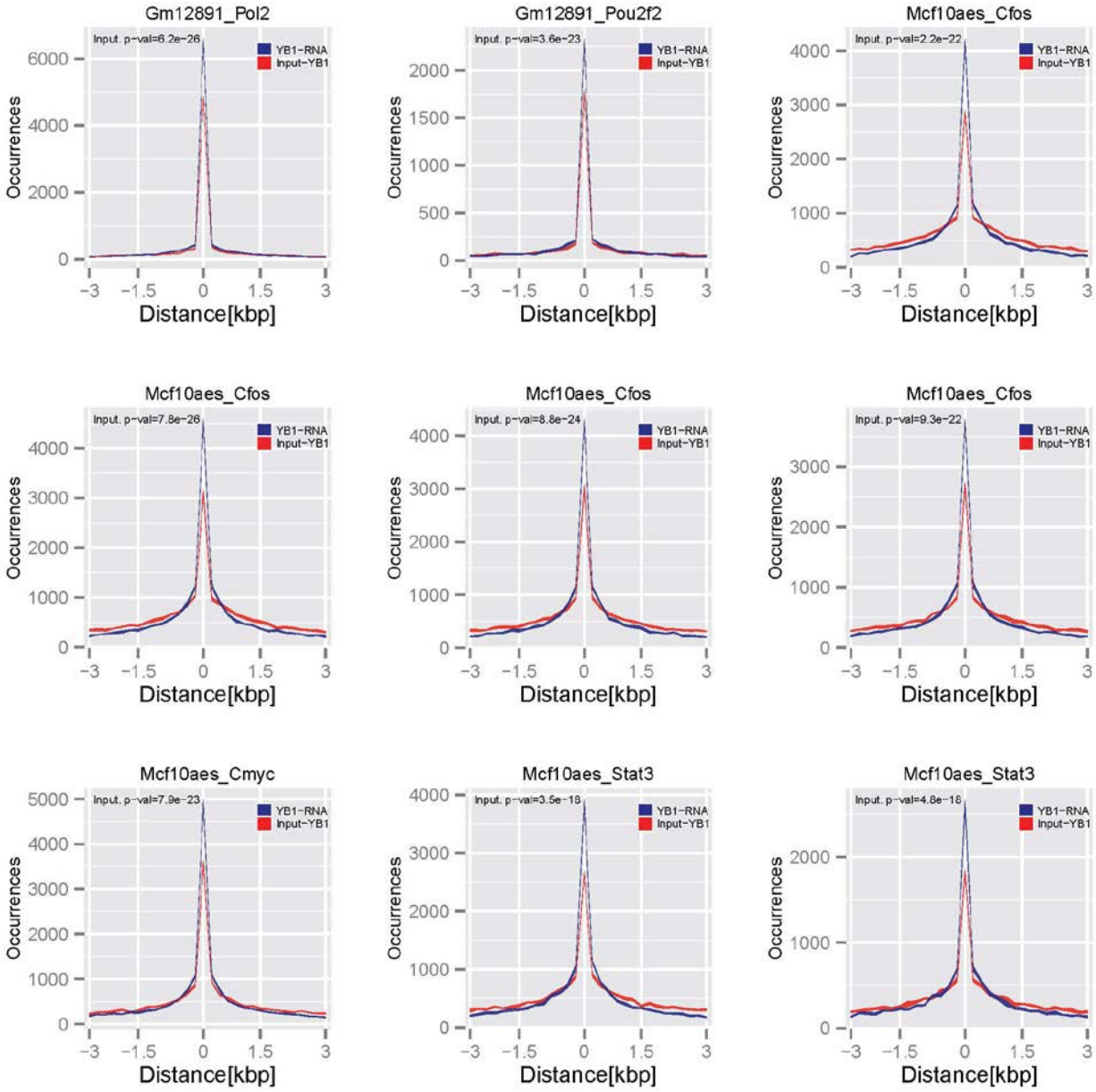

Mcf7_Ctcfc
Input. p-val=4.9e-15
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Mcf7_Ctcf
Input. p-val=8e-17
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Mcf7_Ctcf
Input. p-val=2.1e-22
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Mcf7_Ctcf
Input. p-val=2.4e-24
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Mcf7_Ctcfc
Input. p-val=2.6e-18
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Mcf7_Ctcf
Input. p-val=8.8e-25
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Mcf7_Rad21
Input. p-val=7.8e-21
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
A549_Bcl3
Input. p-val=1.9e-26
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
A549_Bcl3
Input. p-val=1.7e-22
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]


A549_Cebpb
Input. p-val=3.8e-19
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
A549_Cebpbsc150
Input. p-val=1.7e-20
A549_Ctcf
Input. p-val=1.1e-21
A549_Ctcfsc5916
Input. p-val=1.1e-20
A549_Ctcf
Input. p-val=1.7e-13
A549_Ctcf
Input. p-val=3.3e-23
A549_Ctcf
Input. p-val=4.2e-16
A549_Ctcfsc5916
Input. p-val=1.3e-17
A549_Ctcfsc5916
Input. p-val=6.1e-17


A549_Fosl2
Input. p-val=1.3e-20
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
A549_Fosl2
Input. p-val=1.7e-23
A549_Gr
Input. p-val=3.5e-18
A549_Gr
Input. p-val=3.4e-19
A549_Jund
Input. p-val=1.6e-16
A549_P300
Input. p-val=3.4e-17
A549_Pol2s2
Input. p-val=1.3e-28
A549_Pol2
Input. p-val=7.1e-25
A549_Rad21
Input. p-val=1.8e-19


A549_Rad21
Input. p-val=4.9e-18
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
A549_Rad21
Input. p-val=8e-24
A549_Sin3ak20
Input. p-val=2.1e-20
A549_Tcf12
Input. p-val=8.6e-20
A549_Ctcf
Input. p-val=2.8e-23
A549_Ctcf
Input. p-val=3.3e-21
H1hesc_Atf2sc81188
Input. p-val=1.6e-17
H1hesc_Ctcfsc5916
Input. p-val=3.4e-19
H1hesc_Pol24h8
Input. p-val=1e-26


H1hesc_Pol24h8
Input. p-val=9.7e-25
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
H1hesc_Rad21
Input. p-val=1.1e-20
H1hesc_Sp1
Input. p-val=2.9e-24
H1hesc_Ctcf
Input. p-val=2.7e-23
Imr90_Cebpb
Input. p-val=7.3e-19
Imr90_Mafk
Input. p-val=1.5e-18
Gm12891_Pol2
Input. p-val=5.2e-24
Gm12891_Pol24h8
Input. p-val=5.5e-23
Gm12891_Pol24h8
Input. p-val=4.8e-22


Gm12891_Pol2
Input. p-val=6.2e-26
YB1-RNA
Input-YB1
Occurrences
Distance[kbp]
Gm12891_Pou2f2
Input. p-val=3.6e-23
Mcf10aes_Cfos
Input. p-val=2.2e-22
Mcf10aes_Cfos
Input. p-val=7.8e-26
Mcf10aes_Cfos
Input. p-val=8.8e-24
Mcf10aes_Cfos
Input. p-val=9.3e-22
Mcf10aes_Cmyc
Input. p-val=7.9e-23
Mcf10aes_Stat3
Input. p-val=3.5e-18
Mcf10aes_Stat3
Input. p-val=4.8e-16

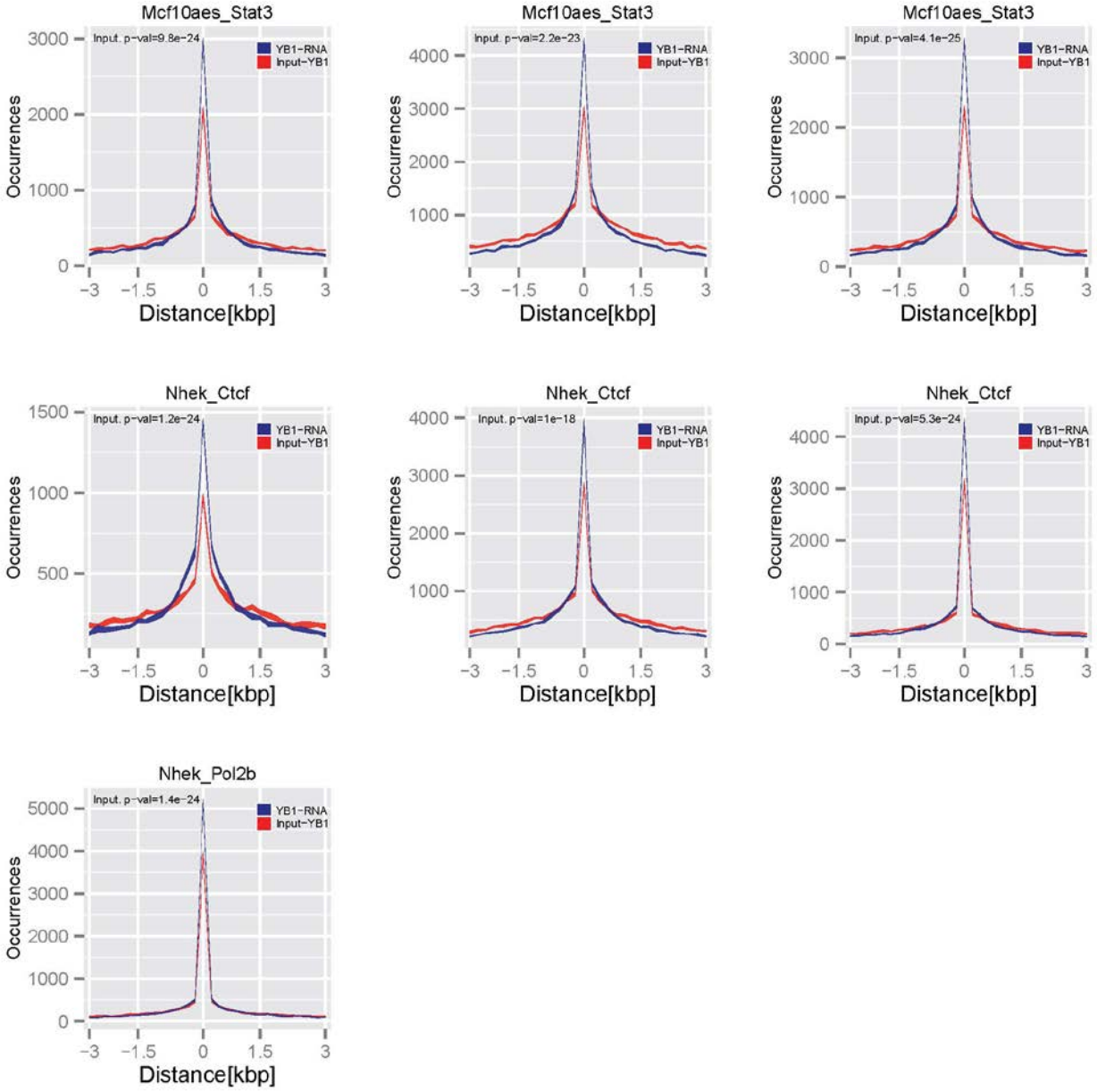


Figure B2. Detail information of chromatin remodeling factors and transcriptional regulators bind to shyRNA genomic locations.

BIBLIOGRAPHY

1. Ozsolak, F. et al. Direct RNA sequencing. *Nature* **461**, 814-818 (2009).
2. Hayashita, Y. et al. A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. *Cancer Res* **65**, 9628-9632 (2005).
3. Eis, P.S. et al. Accumulation of miR-155 and BIC RNA in human B cell lymphomas. *Proc Natl Acad Sci U S A* **102**, 3627-3632 (2005).
4. Hossain, A., Kuo, M.T. & Saunders, G.F. Mir-17-5p regulates breast cancer cell proliferation by inhibiting translation of AIB1 mRNA. *Mol Cell Biol* **26**, 8191-8201 (2006).
5. Cawley, S. et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499-509 (2004).
6. Kampa, D. et al. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* **14**, 331-342 (2004).
7. Kapranov, P. et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484-1488 (2007).
8. Willingham, A.T. & Gingeras, T.R. TUF love for "junk" DNA. *Cell* **125**, 1215-1220 (2006).
9. Cheng, J. et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149-1154 (2005).
10. Genomes Project, C. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
11. Genome, K.C.o.S. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of heredity* **100**, 659-674 (2009).
12. Sboner, A., Mu, X.J., Greenbaum, D., Auerbach, R.K. & Gerstein, M.B. The real cost of sequencing: higher than you think! *Genome Biol* **12**, 125 (2011).
13. Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G. & Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res* **21**, 734-740 (2011).
14. S. Deorowicz, Sz. Grabowski, Robust relative compression of genomes with random access. *Bioinformatics*, **27**(21):2979–2986 (2011).
15. Wang C, Zhang D, A novel compression tool for efficient storage of genome resequencing data. *Nucleic Acids Res.* 39:e45 (2011).
16. Pinho A, Pratas D, Garcia S. GReEn: a tool for efficient compression of genome resequencing data. *Nucleic Acids Res.* 40:e27 (2011).
17. Dmitri S. Pavlichin, Tsachy Weissman, Golan Yona. The human genome contracts again *Bioinformatics* 29 (17): 2199-2202 (2013)
18. Deorowicz, S. & Grabowski, S. Compression of DNA sequence reads in FASTQ format. *Bioinformatics* **27**, 860-862 (2011).

19. Tembe, W., Lowey, J. & Suh, E. G-SQZ: compact encoding of genomic sequence and quality data. *Bioinformatics* **26**, 2192-2194 (2010).
20. Jones, D.C., Ruzzo, W.L., Peng, X. & Katze, M.G. Compression of next-generation sequencing reads aided by highly efficient de novo assembly. *Nucleic Acids Res* (2012).
21. The NIH HMP Working Group et, al. The NIH Human Microbiome Project Genome Res. 19(12): 2317–2323 (2009).
22. Christley, S., Lu, Y., Li, C. & Xie, X. Human genomes as email attachments. *Bioinformatics* **25**, 274-275 (2009).
23. Chatterjee S, Pal JK. Role of 5′- and 3′-untranslated regions of mRNAs in human diseases. *Biol. Cell.* 101*:*251-262 (2009).
24. Pickering BM, Willis AE. The implications of structured 5′ untranslated regions on translation and disease. *Semin. Cell. Dev. Biol.* 16*:*39-47 (2005).
25. Hesketh J. 3′-Untranslated regions are important in mRNA localization and translation: lessons from selenium and metallothionein. *Biochem. Soc. Trans.* 32:990-993(2004).
26. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* 320*:*1643-1647 (2008).
27. Mayr C, Bartel DP. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 138*:*673-684 (2009*).*
28. Jan CH, Friedman RC, Ruby JG, Bartel DP. Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. *Nature* 469*:*97-101(2011*).*
29. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456*:*470-476 (2008).
30. Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 143*:*1018-1029 (2010).
31. Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* 17*:*761-772 (2011).
32. Fu Y, Sun Y, Li Y, Li J, Rao X, Chen C, Xu A. Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.* 21*:*741-747(2011).
33. Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, *Gutwein* MR, Khivansara V, et al. The landscape of *C. elegans 3′UTRs. Science* 329*:*432-435(2010).
34. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM. Direct RNA sequencing. *Nature* 461*:*814-818(2009).
35. Lutz CS. Alternative polyadenylation: a twist on mRNA 3′ end formation. *ACS Chem. Biol.* 3*:*609-617(2008).
36. Stacey SN, Sulem P, Jonasdottir A, Masson G, Gudmundsson J, Gudbjartsson DF, Magnusson OT, Gudjonsson SA, Sigurgeirsson B, Thorisdottir K, et al. A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat. Genet.* 43*:*1098-1103(2011).
37. Wiestner A, Tehrani M, Chiorazzi M, Wright G, Gibellini F, Nakayama K, Liu H, Rosenwa ld A, Muller-Hermelink HK, Ott G, et al. Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. *Blood* 109*:*4599-4606 (2007).

38. Moore MJ. From birth to death: the complex lives of eukaryotic mRNAs. *Science* 309:1514-1518(2005).
39. Pillai RS, Bhattacharyya SN, Artus CG, Zoller T, Cougot N, Basyuk E, Bertrand E, Filipow icz W. Inhibition of translational initiation by Let-7 MicroRNA in human cells. *Science* 309:1573-1576 (2005).
40. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466*:*835-840 (2010).
41. Hendrickson DG, Hogan DJ, McCullough HL, Myers JW, Herschlag D, Ferrell JE, Brown PO. Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS Biol.* 7:e1000238 (2009).
42. *Neilson JR, Sandberg R.* Heterogeneity in mammalian RNA 3′ end formation. *Exp. Cell Res.* 316*:*1357-1364 (2010).
43. Ji Z, Lee JY, Pan Z, Jiang B, Tian B. Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. USA* 106*:*7028-7033 (2009).
44. Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, et al. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9:e1001046 (2011).
45. Wang KC, Chang *HY*. Molecular mechanisms of long noncoding RNAs. *Mol. Cell;*43*:*904-914 (2011).
46. Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, Shen L, Donahue WF, Tusneem N, Stromberg MP, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.* 18*:*1638-1642 (2008).
47. Nam DK, Lee S, Zhou G, Cao X, Wang C, Clark T, Chen J, Rowley JD, Wang SM. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc. Natl. Acad. Sci. USA* 99*:*6152-6156 (2002).
48. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, et al. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* 39*:*D876-D882 (2011).
49. Tian B, Hu J, Zhang H, Lutz CS. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* 33*:*201-212 (2005).
50. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 27*:*1653-1659 (2011).
51. Bailey TL, Machanick P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res* May 18 (2012).
52. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34*:*W369-W373 (2006).
53. Beaudoing E, Gautheret D. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res.* 11*:*1520-1526 (2001).
54. Pauws E, van Kampen AH, van de Graaf SA, de Vijlder JJ, Ris-Stalpers C. Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. *Nucleic Acids Res.* 29*:*1690-1694 (2001).
55. Kapranov P, St. Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, Reaman G, Milos P, Arceci RJ, Thompson JF, et al. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol.* 8*:*149 (2010).

56. van Bakel H, Nislow C, Blencowe BJ, Hughes TR. Most “dark matter” transcripts are associated with known genes. *PLoS Biol.* 8*:*e1000371 (2010).
57. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* 21*:*1757-1767 (2011).
58. Kowalczyk MS, Hughes JR, Garrick D, Lynch MD, Sharpe JA, Sloane-Stanley JA, McGowan SJ, De Gobbi M, Hosseini M, Vernimmen D, et al. Intragenic enhancers act as alternative promoters. *Mol. Cell* 45*:*447-458 (2012).
59. Edwalds-Gilbert G, Veraldi KL, Milcarek C. Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res.* 25*:*2547-2561 (1997).
60. Singh P, Alley TL, Wright SM, Kamdar S, Schott W, Wilpan RY, Mills KD, Graber JH. Gl obal changes in processing of mRNA 3′ untranslated regions characterize clinically distinct cancer subtypes. *Cancer Res.* 69*:*9422-9430 (2009).
61. Lembo A, Di Cunto F, Provero P. Shortening of 3′UTRs correlates with poor prognosis in breast and lung cancer. *PLoS One* 7*:*e31129 (2012).
62. Esteller M. Non-coding RNAs in human disease. *Nat. Rev. Genet.* 12*:*861-874 (2011).
63. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129*:*1311-1323 (2007).
64. Bracht J, Hunter S, Eachus R, Weeks P, Pasquinelli AE. Trans-splicing and polyadenylation of let-7 microRNA primary transcripts. *RNA* 10*:*1586-1594 (2004).
65. Mourtada-Maarabouni M, Pickard MR, Hedge VL, Farzaneh F, Williams GT. GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene* 28*:*195-208 (2009).
66. Lee K, Kunkeaw N, Jeon SH, Lee I, Johnson BH, Kang GY, Bang JY, Park HS, Leelayuwa t C, Lee YS. Precursor miR-886, a novel noncoding RNA repressed in cancer, associates with PKR and modulates its activity. *RNA;*17*:*1076-1089 (2011).
67. Meiri E, Levy A, Benjamin H, Ben-David M, Cohen L, Dov A, Dromi N, Elyakim E, Yerushalmi N, Zion O, et al. Discovery of microRNAs and other small RNAs in solid tumors. *Nucleic Acids Res.*38*:*6234-6246 (2010).
68. Pleasance ED, Stephens PJ, O’Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 463*:*184-190 (2010).
69. Takai Y, Miyoshi J, Ikeda W, Ogita H. Nectins and nectin-like molecules: roles in contact inhibition of cell movement and proliferation. *Nat. Rev. Mol. Cell. Biol.* 9*:*603-615 (2008).
70. Kim SH, Turnbull J, Guimond S. Extracellular matrix and cell signalling: the dynamic cooperation of integrin, proteoglycan and growth factor receptor. *J. Endocrinol.* 209*:*139-151 (2011).
71. Yamazaki D, Kurisu S, Takenawa T. Regulation of cancer cell motility through actin reorganization. *Cancer Sci.* 96*:*379-386 (2005).
72. White RJ. RNA polymerase III transcription-a battleground for tumour suppressors and oncogenes. *Eur. J. Cancer.* 40*:*21-27 (2004).
73. Rubbi L, Labarre-Mariotte S, Chedin S, Thuriaux P. Functional characterization of ABC10alpha, an essential polypeptide shared by all three forms of eukaryotic DNA-dependent RNA polymerases. *J. Biol. Chem.* 274*:*31485-31492 (1999).

74. Alberts AS, Treisman R. Activation of RhoA and SAPK/JNK signalling pathways by the RhoA-specific exchange factor mNET1. *EMBO J.* 17:4075-4085 (1998).
75. Murray D, Horgan G, Macmathuna P, Doran P. NET1-mediated RhoA activation facilitates lysophosphatidic acid-induced cell migration and invasion in gastric cancer. *Br. J. Cancer* 99*:*1322-1329 (2008).
76. van Helden J, del Olmo M, Perez-Ortin JE. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.* 28*:*1000-1010 (2000).
77. Kessler MM, Henry MF, Shen E, Zhao J, Gross S, Silver PA, Moore CL. Hrp1, a sequence-specific RNA-binding protein that shuttles between the nucleus and the cytoplasm, is required for mRNA 3′-end formation in yeast. *Genes Dev.* 11*:*2545-2556 (1997).
78. Perez-Canadillas JM. Grabbing the message: structural basis of mRNA 3′UTR recognition by Hrp1. *EMBO J.* 25*:*3167-3178 (2006).
79. Ji Z, Tian B. Reprogramming of 3′ untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS One* 4*:*e8419 (2009).
80. Jayaseelan S, Doyle F, Currenti S, Tenenbaum SA. RIP: an mRNA localization technique. *Methods Mol. Biol.* 714*:*407-422 (2011).
81. Yao P, Potdar AA, Arif A, Ray PS, Mukhopadhyay R, Willard B, Xu Y, Yan J, Saidel GM, Fox PL. Coding region polyadenylation generates a truncated tRNA synthetase that counters translation repression. *Cell* 149*:*88-100 (2012).
82. Yang Z, Kaye DM. Mechanistic insights into the link between a polymorphism of the 3′UTR of the SLC7A1 gene and hypertension. *Hum. Mutat.* 30*:*328-333 (2009).
83. Nagaike T, Logan C, Hotta I, Rozenblatt-Rosen O, Meyerson M, Manley JL. Transcriptional activators enhance polyadenylation of mRNA precursors. *Mol. Cell* 41*:*409-418 (2011).
84. Ji Z, Luo W, Li W, Hoque M, Pan Z, Zhao Y, Tian B. Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol. Syst. Biol.* 7*:*534 (2011).
85. Moore MJ, Proudfoot NJ. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* 136*:*688-700 (2009).
86. Jing Q, Huang S, Guth S, Zarubin T, Motoyama A, Chen J, Di Padova F, Lin SC, Gram H, Han J. Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell* 120*:*623-634 (2005).
87. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. *Nature* 455*:*58-63 (2008).
88. Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. *Nature* 455*:*64-71 (2008).
89. Kohno, K, Izumi, H, Uchiumi, T, Ashizuka, M and Kuwano, M. "The pleiotropic functions of the y-box-binding protein, YB-1." *Bioessays* 25(7): 691-8 (2003).
90. Didier, DK, Schiffenbauer, J, Woulfe, SL, Zacheis, M and Schwartz, BD. "Characterization of the cdna encoding a protein binding to the major histocompatibility complex class ii y box." *Proc Natl Acad Sci USA* 85(19): 7322-6 (1988).
91. Evdokimova, V, Ruzanov, P, Anglesio, MS, Sorokin, AV, Ovchinnikov, LP, Buckley, J, et al. Akt-mediated YB-1 phosphorylation activates translation of silent mRNA species. *Mol Cell Biol* 26(1):277-92. (2006).

92. Dutertre, M, Sanchez, G, De Cian, MC, Barbier, J, Dardenne, E, Gratadou, L, et al. Cotranscriptional exon skipping in the genotoxic stress response. *Nat Struct Mol Biol* 17(11):1358-66. (2010).
93. Stickeler, E, Fraser, SD, Honig, A, Chen, AL, Berget, SM and Cooper, TA. The RNA binding protein YB-1 binds a/c-rich exon enhancers and stimulates splicing of the cd44 alternative exon v4. *EMBO J* 20(14): 3821-30 (2001).
94. Evdokimova, V, Ruzanov, P, Imataka, H, Raught, B, Svitkin, Y, Ovchinnikov, LP, et al. The major mRNA-associated protein YB-1 is a potent 5' cap-dependent mRNA stabilizer. *EMBO J* 20(19): 5491-502 (2001).
95. Spitkovsky, DD, Royer-Pokora, B, Delius, H, Kisseljov, F, Jenkins, NA, Gilbert, DJ, et al. Tissue restricted expression and chromosomal localization of the YB-1 gene encoding a 42 kd nuclear ccaat binding protein. *Nucleic Acids Res* 20(4): 797-803 (1992)..
96. Van de Putte, T, Maruhashi, M, Francis, A, Nelles, L, Kondoh, H, Huylebroeck, D, et al. Mice lacking zfhx1b, the gene that codes for smad-interacting protein-1, reveal a role for multiple neural crest cell defects in the etiology of hirschsprung disease-mental retardation syndrome. *Am J Hum Genet* 72(2): 465-70 (2003).
97. Agafonov, DE, Deckert, J, Wolf, E, Odenwalder, P, Bessonov, S, Will, CL, et al. Semiquantitative proteomic analysis of the human spliceosome via a novel two dimensional gel electrophoresis method. *Mol Cell Biol* 31(13): 2667-82 (2011).
98. Deckert, J, Hartmuth, K, Boehringer, D, Behzadnia, N, Will, CL, Kastner, B, et al. Protein composition and electron microscopy structure of affinity-purified human spliceosomal b complexes isolated under physiological conditions. *Mol Cell Biol* 26(14): 5528-43 (2006).
99. Hock, J, Weinmann, L, Ender, C, Rudel, S, Kremmer, E, Raabe, M, et al. Proteomic and functional analysis of argonaute-containing mRNA-protein complexes in human cells. *EMBO Rep* 8(11): 1052-60 (2007).
100. Girard, A, Sachidanandam, R, Hannon, GJ and Carmell, MA A germline-specific class of small RNAs binds mammalian piwi proteins. *Nature* 442(7099): 199-202 (2006).
101. Hammond, SM, Boettcher, S, Caudy, AA, Kobayashi, R and Hannon, GJ. Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science* 293(5532): 1146-50 (2001).
102. Meister, G, Landthaler, M, Dorsett, Y and Tuschl, T Sequence-specific inhibition of microRNA- and siRNA-induced RNA silencing. *RNA* 10(3): 544-50 (2004).
103. Moss, EG and Tang, L Conservation of the heterochronic regulator lin-28, its developmental expression and microRNA complementary sites. *Dev Biol* 258(2): 432-42 (2003).
104. Yu, J, Vodyanik, MA, Smuga-Otto, K, Antosiewicz-Bourget, J, Frane, JL, Tian, S, et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318(5858): 1917-20 (2007).
105. Viswanathan, SR, Powers, JT, Einhorn, W, Hoshida, Y, Ng, TL, Toffanin, S, et al. Lin28 promotes transformation and is associated with advanced human malignancies. *Nat Genet* 41(7): 843-8 (2009).
106. Bargou, RC, Jurchott, K, Wagener, C, Bergmann, S, Metzner, S, Bommert, K, et al. Nuclear localization and increased levels of transcription factor YB-1 in primary human breast cancers are associated with intrinsic mdr1 gene expression. *Nat Med* 3(4): 447-50 (1997).
107. Habibi, G, Leung, S, Law, JH, Gelmon, K, Masoudi, H, Turbin, D, et al."Redefining prognostic factors for breast cancer: YB-1 is a stronger predictor of relapse and disease-

specific survival than estrogen receptor or her-2 across all tumor subtypes. *Breast Cancer Res* 10(5): R86 (2008).
108. Gimenez-Bonafe, P, Fedoruk, MN, Whitmore, TG, Akbari, M, Ralph, JL, Ettinger, S, et al. YB-1 is upregulated during prostate cancer tumor progression and increases p-glycoprotein activity. *Prostate* 59(3): 337-49 (2004).
109. Zhang, YF, Homer, C, Edwards, SJ, Hananeia, L, Lasham, A, Royds, J, et al. Nuclear localization of y-box factor YB1 requires wild-type p53. *Oncogene* 22(18): 2782-94 (2003)..
110. Gupta, RA, Shah, N, Wang, KC, Kim, J, Horlings, HM, Wong, DJ, et al. Long non-coding RNA hotair reprograms chromatin state to promote cancer metastasis." *Nature* 464(7291): 1071-6 (2010).
111. Lai, F, Orom, UA, Cesaroni, M, Beringer, M, Taatjes, DJ, Blobel, GA, et al. Activating RNAs associate with mediator to enhance chromatin architecture and transcription. *Nature* 494(7438): 497-501 (2013).
112. Meiri, E, Levy, A, Benjamin, H, Ben-David, M, Cohen, L, Dov, A, et al. Discovery of microRNAs and other small RNAs in solid tumors. *Nucleic Acids Res* 38(18): 6234-46 (2010).
113. Layer, JH and Weil, PA. Ubiquitous antisense transcription in eukaryotes: Novel regulatory mechanism or byproduct of opportunistic RNA polymerase? *F1000 Biol Rep* 1: 33 (2009).
114. Thompson, DM and Parker, R Cytoplasmic decay of intergenic transcripts in saccharomyces cerevisiae. *Mol Cell Biol* 27(1): 92-101 (2007).
115. Ito, K, Tsutsumi, K, Kuzumaki, T, Gomez, PF, Otsu, K and Ishikawa, K. A novel growth-inducible gene that encodes a protein with a conserved cold-shock domain. *Nucleic Acids Res* 22(11): 2036-41 (1994).
116. Neil, H, Malabat, C, d'Aubenton-Carafa, Y, Xu, Z, Steinmetz, LM and Jacquier, A Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* 457(7232):1038-42 (2009).
117. Seila, AC, Calabrese, JM, Levine, SS, Yeo, GW, Rahl, PB, Flynn, RA, et al. Divergent transcription from active promoters. *Science* 322(5909): 1849-51 (2008).
118. Kim, TK, Hemberg, M, Gray, JM, Costa, AM, Bear, DM, Wu, J, et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465(7295): 182-7 (2010).
119. Kowalczyk, MS, Hughes, JR, Garrick, D, Lynch, MD, Sharpe, JA, Sloane-Stanley, JA, et al. Intragenic enhancers act as alternative promoters. *Mol Cell* 45(4): 447-58 (2012).
120. Christov, CP, Gardiner, TJ, Szuts, D and Krude, T. Functional requirement of noncoding y RNAs for human chromosomal DNA replication. *Mol Cell Biol 26*(18): 6993-7004 (2006).
121. Persson, H, Kvist, A, Vallon-Christersson, J, Medstrand, P, Borg, Å and Rovira, C. The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. *Nat Cell Biol* 11(10): 1268-71 (2009).
122. Krude, T, Christov, CP, Hyrien, O and Marheineke, K. Y RNA functions at the initiation step of mammalian chromosomal DNA replication. *J Cell Sci* 122(Pt 16): 2836-45 (2009).
123. Christov, CP, Trivier, E and Krude, T Noncoding human Y RNAs are overexpressed in tumours and required for cell proliferation. *Br J Cancer* 98(5): 981-6 (2008).
124. Mortazavi, A, Williams, BA, McCue, K, Schaeffer, L and Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* 5(7): 621-8 (2008).
125. Taft, RJ, Glazov, EA, Lassmann, T, Hayashizaki, Y, Carninci, P and Mattick, JS Small RNAs derived from snoRNAs. *RNA* 15(7): 1233-40 (2009).

126. Preker, P, Nielsen, J, Kammler, S, Lykke-Andersen, S, Christensen, MS, Mapendano, CK, et al. RNA exosome depletion reveals transcription upstream of active human promoters. *Science* 322(5909): 1851-4 (2008).
127. Ozsolak, F, Kapranov, P, Foissac, S, Kim, SW, Fishilevich, E, Monaghan, AP, et al. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 143(6): 1018-29 (2010).
128. Kapranov, P, Cheng, J, Dike, S, Nix, DA, Duttagupta, R, Willingham, AT, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316(5830):1484-8 (2007).
129. Lee, YS, Shibata, Y, Malhotra, A and Dutta, A. A novel class of small RNAs: tRNA-derived RNA fragments (trfs). *Genes Dev* 23(22): 2639-49 (2009).
130. Lin, Y, Li, Z, Ozsolak, F, Kim, SW, Arango-Argoty, G, Liu, TT, et al. "An in-depth map of polyadenylation sites in cancer." *Nucleic Acids Res* 40(17): 8460-71 (2012).
131. Sibley, CR, Seow, Y, Saayman, S, Dijkstra, KK, El Andaloussi, S, Weinberg, MS, et al. The biogenesis and characterization of mammalian microRNAs of mirtron origin. *Nucleic Acids Res* (2011).
132. Valen, E, Preker, P, Andersen, PR, Zhao, X, Chen, Y, Ender, C, et al. Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nat Struct Mol Biol* 18(9): 1075-82 (2011).
133. Zentner, GE, Saiakhova, A, Manaenkov, P, Adams, MD and Scacheri, PC Integrative genomic analysis of human ribosomal DNA. *Nucleic Acids Res* 39(12): 4949-60 (2011).
134. Raha, D, Wang, Z, Moqtaderi, Z, Wu, L, Zhong, G, Gerstein, M, et al. Close association of RNA polymerase ii and many transcription factors with pol iii genes. *Proc Natl Acad Sci USA* 107(8):3639-44 (2010).
135. Sims, RJ, 3rd and Reinberg, D Processing the h3k36me3 signature. *Nat Genet* 41(3): 270-1 (2009).
136. Rada-Iglesias, A, Bajpai, R, Swigut, T, Brugmann, SA, Flynn, RA and Wysocka, J. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470(7333):279-83 (2011).
137. Vavouri, T and Lehner, B Human genes with cpg island promoters have a distinct transcription-associated chromatin organization. *Genome Biol* 13(11): R110 (2012)..
138. Boyer, LA, Plath, K, Zeitlinger, J, Brambrink, T, Medeiros, LA, Lee, TI, et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* 441(7091): 349-53 (2006).
139. Martens, JH, O'Sullivan, RJ, Braunschweig, U, Opravil, S, Radolf, M, Steinlein, P, et al. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J* 24(4): 800-12 (2005).
140. Myers, RM, Stamatoyannopoulos, J, Snyder, M, Dunham, I, Hardison, RC, Bernstein, BE, et al. A user's guide to the encyclopedia of DNA elements (encode). *PLoS Biol* 9(4): e1001046 (2011).
141. Chernukhin, IV, Shamsuddin, S, Robinson, AF, Carne, AF, Paul, A, El-Kady, AI, et al. Physical and functional interaction between two pluripotent proteins, the y-box DNA/RNA-binding factor, YB-1, and the multivalent zinc finger factor, ctcf. *J Biol Chem* 275(38): 29915-21 (2000).
142. Lee, CW, Sorensen, TS, Shikama, N and La Thangue, NB Functional interplay between p53 and e2f through co-activator p300. *Oncogene* 16(21): 2695-709 (1998).

143. Ayyanathan, K, Lechner, MS, Bell, P, Maul, GG, Schultz, DC, Yamada, Y, et al. Regulated recruitment of hp1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: A mammalian cell culture model of gene variegation. *Genes Dev* 17(15): 1855 -69 (2003).
144. Sripathy, SP, Stevens, J and Schultz, DC The kap1 corepressor functions to coordinate the assembly of de novo hp1-demarcated microenvironments of heterochromatin required for krab zinc finger protein-mediated transcriptional repression. *Mol Cell Biol* 26(22): 8623 -38 (2006).
145. Rooney, JW and Calame, KL. Tif1beta functions as a coactivator for c/ebpbeta and is required for induced differentiation in the myelomonocytic cell line u937. *Genes Dev* 15(22):3023 -38 (2001).
146. Tomlins, SA, Rhodes, DR, Perner, S, Dhanasekaran, SM, Mehra, R, Sun, XW, et al. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *Science* 310(5748): 644 -8 (2005).
147. Shaikhibrahim, Z, Langer, B, Lindstrot, A, Florin, A, Bosserhoff, A, Buettner, R, et al. Ets-1 is implicated in the regulation of androgen co-regulator fhl2 and reveals specificity for migration, but not invasion, of pc3 prostate cancer cells. *Oncol Rep* 25(4): 1125 -9 (2011).
148. Holterman, CE, Franovic, A, Payette, J and Lee, S Ets-1 oncogenic activity mediated by transforming growth factor alpha. *Cancer Res* 70(2): 730-40 (2010).
149. Fujimoto, J, Aoki, I, Toyoki, H, Khatun, S and Tamaya, T Clinical implications of expression of ets [illegible] and ets-related genes in uterine cervical cancers. *Ann Oncol* 13(10): 1598-604 (2002)..
150. Span, PN, Manders, P, Heuvel, JJ, Thomas, CM, Bosch, RR, Beex, LV, et al. "Expression of the transcription factor ets-1 is an independent prognostic marker for relapse -free survival in breast cancer." *Oncogene* 21(55): 8506-9 (2002).
151. Davidson, B, Reich, R, Goldberg, I, Gotlieb, WH, Kopolovic, J, Berner, A, et al. "Ets-1 messenger RNA expression is a novel marker of poor survival in ovarian carcinoma." *Clin Cancer Res* 7(3): 551-7 (2001).
152. Loven, J, Hoke, HA, Lin, CY, Lau, A, Orlando, DA, Vakoc, CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* 153(2): 320-34 (2013).
153. Whyte, WA, Orlando, DA, Hnisz, D, Abraham, BJ, Lin, CY, Kagey, MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 153(2): 307-19 (2013).
154. Gentleman, RC, Carey, VJ, Bates, DM, Bolstad, B, Dettling, M, Dudoit, S, et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol* 5(10): R80 (2004).
155. Karolchik, D, Hinrichs, AS and Kent, WJ The ucsc genome browser. *Curr Protoc Bioinformatics* Chapter 1: Unit1 4 (2009).