Chapter 14

Identification of Transcription Factor Binding Sites Derived from Transposable Element Sequences Using ChIP-seq

Andrew B. Conley and I. King Jordan

Abstract

Transposable elements (TEs) form a substantial fraction of the non-coding DNA of many eukaryotic genomes. There are numerous examples of TEs being exapted for regulatory function by the host, many of which were identified through their high conservation. However, given that TEs are often the youngest part of a genome and typically exhibit a high turnover, conservation-based methods will fail to identify lineage- or species-specific exaptations. ChIP-seq has become a very popular and effective method for identifying in vivo DNA–protein interactions, such as those seen at transcription factor binding sites (TFBS), and has been used to show that there are a large number of TE-derived TFBS. Many of these TE-derived TFBS show poor conservation and would go unnoticed using conservation screens. Here, we describe a simple pipeline method for using data generated through ChIP-seq to identify TE-derived TFBS.

Key words: Transposable elements, ChIP-seq, gene regulation, gene expression, transcription factors, CTCF.

1. Introduction

Transposable elements (TEs) are segments of DNA that possess the ability to 'transpose,' meaning that they can move themselves to distant locations of the host genome and replicate when they do so. TEs are present in all domains of life and are abundant in the genomes of many sequenced eukaryotes accounting for a large portion of non-coding DNA and the genomes as a whole (nearly 50%, ~1.4 Gb of the human genome) (1). Broadly speaking, there are two types of TEs. Type I TEs, or retroelements, transpose by a copy and paste mechanism via an RNA intermediate, generating a new insertion. Type II TEs, or DNA transposons,

I. Ladunga (ed.), *Computational Biology of Transcription Factor Binding*, Methods in Molecular Biology 674,
DOI 10.1007/978-1-60761-854-6_14, © Springer Science+Business Media, LLC 2010

move by a 'cut-and-paste' mechanism where the actual insertion is moved (2). Most TEs harbor their own promoters and regulatory sequences, and many active elements encode genes for their own transposition. Active elements are a small minority, however, and most TE insertions are unable to transpose.

1.1. Exaptation of Transposable Elements

TEs exist solely to continue their own existence; they do not, simply by their replication, contribute anything to the host (3, 4). It is likely that many, if not the large majority of TE insertions, have little or no functional role for the host and are effectively under neutral or nearly neutral selection. However, given the very large number of TE insertions in eukaryotic genomes and the opportunistic nature of evolution, it is only reasonable to expect that some would be 'exapted' (5) over time to take on a functional role that benefits the host, a process that could have a wide variety of results (6, 7). A key factor in TE exaptation events is their ability to promote their own transcription; without this ability, they could not replicate themselves. Given this ability, it stands to reason that TEs could be exapted to provide alternative promoters for host genes; this has been seen a number of times (8, 9). Of most importance to this chapter, however, is the ability of TEs to provide new TFBS to the host. If there existed an active TE that contained a TFBS, then each new insertion that the TE generated would also contain the TFBS. If the TE were highly active, it could quickly spread the TFBS around the genome. Even if the TE simply had a sequence that was only close to the TFBS, it could still spread this 'progenitor sequence' around the genome. Over time, point mutations in individual insertions could alter the progenitor sequence so that it would now be bound by the TF (10). Either way, the TE could spread the TFBS around the genome over timer and create a network of TFBS, and in doing so alter the expression patterns of host genes. For example, it was recently shown that a large number of human c-myc binding sites are located in TE insertions, possibly creating a sub-network for c-myc control (11). For a comprehensive review of TE-derived regulatory networks, *see* (12).

1.2. Transposable Elements Evolve Rapidly

Transposable elements are generally the most rapidly evolving part of a genome; so long as their insertions are not too deleterious to the host, TEs can quickly increase in copy number and then are generally free to accumulate point mutations. The rapid activity of TEs relative to the host genome means that lineage-specific insertions can be accumulated in a very short time frame. In the 6 million years since the human–chimpanzee divergence, for example, there have been several thousand new TE insertions in each genome (13). There also appears to be very little selective pressure on the deletion of most insertions, which can result in their chance deletion from one lineage, while they are retained in others. Between human and mouse, there is generally very little conservation of non-coding regions in the genome, including TEs. Many insertions that appear to predate the human–mouse divergence are present in one genome, but have been lost in the other (**Fig.** 14.1) (14). The rapid insertion of TEs combined with their rapid loss means that two lineages can develop distinct TE complements in a relatively short time after divergence. Given that two lineages can have very different TE complements, it could be possible for a large number of lineage or even species-specific exaptation events (**Fig.** 14.1). If the exaptation events were the creation of new TFBS or promoters, then the spread of TEs could create species-specific patterns of gene expression (15, 16).

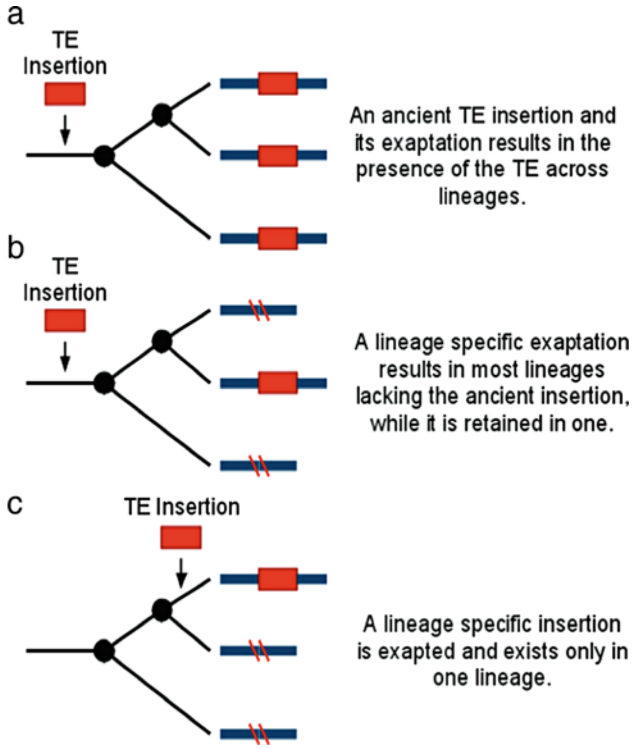
a
TE Insertion
An ancient TE insertion and its exaptation results in the presence of the TE across lineages.
b
TE Insertion
A lineage specific exaptation results in most lineages lacking the ancient insertion, while it is retained in one.
c
TE Insertion
A lineage specific insertion is exapted and exists only in one lineage.

Fig. 14.1. Evolutionary scenarios related to TE exaptation events. **a** An ancient insertion is exapted and the resulting regulatory sequences are shared across multiple derived evolutionary lineages. **b** An ancient insertion is exapted but only selectively conserved in some of the derived evolutionary lineages. This could result in regulatory divergence between lineages. **c** A recent lineage-specific insertion is exapted resulting in regulatory differences between lineages. TEs are particularly prone to this scenario given how dynamic and rapidly evolving they are.

1.3. Detection of Functional TE-Derived Non-coding Sequences

There are three widely used methods to find TFBS in genomes. It should be noted that these approaches are not mutually exclusive; indeed, the methods are often combined to more rigorously predict and locate TFBS.

1.3.1. Phylogenetic Footprinting

The first approach, phylogenetic footprinting (17), can be done solely computationally via comparative sequence analysis. A phylogenetic screen attempts to find regions of different genomes that have been conserved over time and, in the case of TFBS, looking for conserved non-coding elements (CNEs). Screens looking for conserved non-coding elements (CNEs) represent a very successful technique for identifying the oldest and, due to their conservation most likely to be essential, non-coding parts of the genome. Shortly after the sequencing of the human and mouse genomes, it was shown that a larger than expected number of mouse MIR and L2 elements had human orthologs (14). Subsequently, several thousand insertions or insertion fragments near human genes were shown to be under purifying selection, suggesting their exaptation and possible involvement in transcriptional control (18). In recent years, a number of insertions have been shown to be enhancers for human and vertebrate genes, many identified with phylogenetic screens. An insertion from the CORE-SINE family was shown to be conserved across the mammalian lineage and to be an enhancer of the POMC gene in mice (19). The amniote SINE 1, AmnSINE1, family of TEs is a very old family that spread early in the amniote lineage. However, a number of conserved AmnSINE1 insertions exist in the human genome, two of which were shown to be enhancers involved in brain development (20–22). A mammalian interspersed repeat (MIR) was shown to have enhancer 'boosting' activity, in that its presence greatly increased the action of a nearby enhancer, while the MIR could not on its own be an enhancer (23). The problem with an approach based on conservation is that, while it will find many important regions, the screen will miss other regions that are also important, but also lineage specific. Lineage-specific TFBS, such as those that could be provided by lineage-specific TE insertions, could generate lineage-specific expression, and this would be missed by CNE screens (16). Another case in which older elements may be overlooked in CNE screens is one in which an old insertion has been lost, as many are, in several lineages, but exapted in one (**Fig.** 14.1). Such an insertion may well play some role in the lineage that kept it, but it will be completely missed in CNE screens. CNE screens will miss not only new TE exaptations but also other non-coding functional elements. It has been shown previously that sequences with low conservation can play important functional roles, such as rapidly evolving, long non-coding RNAs (24).

1.3.2. Motif Search

The second of the three methods to identify TFBS is also computational and involves scanning a genome for the sequence motif that the TF in question recognizes. REST, the RE1 silencing transcription factor, is known to repress neuronal genes in non-neuronal cells. Using experimentally identified REST binding sites, which contain the RE1 motif, Johnson et al. (25) created a position-specific scoring matrix, PSSM, for the motif and used it to screen for possible REST binding sites in the human genome. Johnson et al. were able to show that there are a number of TE-derived REST binding sites that had the ability to bind REST in vitro, suggesting that TEs have helped to spread the REST network. When a PSSM is used to search for new TFBS in a genome, false positives are controlled by shuffling the sequence in the PSSM, re-scanning the genome with the shuffled sequence, and comparing the number of sites identified with the original PSSM to those found with the shuffled PSSM (26). This approach will not work, however, for TFs that recognize motifs smaller than the RE1 motif as there will likely be many false positives. In addition, the presence of a TFBS sequence motif does not guarantee that the sequence that bears it is actually bound by its corresponding TF, while sequences that lack similarity to the motif may in fact be bound by that factor. These challenges to the sequence-based computational approach necessitate an approach to identifying TFBS on a genome-wide scale that does not depend on the sequence of the TFBS, only the binding of the TF to the region.

1.3.3. ChIP-seq or ChIP-chip

The third major approach to finding TFBS is identifying in vivo protein–DNA interactions via chromatin immunoprecipitation (ChIP) followed by microarray analysis (ChIP-chip) or sequencing of the captured DNA (ChIP-seq, *see* **Chapters 9**, **10**, and **11**). Of the three approaches, this one offers the greatest sensitivity and potential specificity. ChIP is able to find genomic DNA that is bound by a transcription factor, not just those regions that are conserved or for which there exists a well-defined TFBS motif. ChIP is also distinguished from the other approaches in the sense that it identifies sequences that are experimentally characterized to be bound by transcription factors, i.e., not just computational predictions. Genome-wide ChIP assays such as ChIP-PET or ChIP-chip have been used successfully in the past; however, a newer and relatively inexpensive method, ChIP-seq, has quickly become the dominant method of experimentally identifying TFBS, and it is on ChIP-seq that we focus the rest of our discussion. The ChIP-seq method combines ChIP with massively parallel sequencing of the bound DNA (27). The sequencing is usually carried out on one of the currently available short-read sequencers: Illumina Genome Analyzer, ABI SOLiD, or Helicos HeliScope. ChIP-seq has a number of advantages over ChIP-chip and ChIP-PET. There is no cross-hybridization, as can occur in ChIP-chip, and the ChIP-seq signal is a digital count of reads mapping to the TFBS, rather than a fluorescence signal. ChIP-seq is also far less costly than ChIP-PET, which typically relied on capillary sequencing. Using several ChIP-based data sets, including one derived with ChIP-seq, Bourque et al. (28) identified a large number of TE-derived TFBS. The majority of TFBS they observed were not well conserved, with many being lineage specific. This strongly suggests that expansion of TEs within a genome can lead to the concurrent expansion of transcription regulatory networks. Below, we provide a specific example detailing how analysis of ChIP-seq data can be used to identify TE-derived TFBS.

2. Software

All the software we describe and recommend here is publicly available.

Bowtie (29) http://bowtie-bio.sourceforge.net/

MuMRescueLite (30) http://genome.gsc.riken.jp/osc/english/dataresource/

UCSC Genome Browser (31) http://genome.ucsc.edu

UCSC Table Browser (32) http://genome.ucsc.edu

3. Methods

This section describes our choice of tools for the identification of TFBS derived from TE insertions using ChIP-seq data, and we show how these tools can be assembled into an analytical pipeline. The tools presented were chosen for their speed, utility for analysis of TE-derived TFBS, ease of use, and good documentation. To illuminate the use of these tools, we first provide an overview of our analytical pipeline for the detection of TE-derived TFBS (**Fig.** 14.2) and then we give a specific example of how ChIP-seq data can be analyzed to yield genome-wide set of TE-derived TFBS.

3.1. Methods Basics

3.1.1. Mapping

The first step in finding TE-derived TFBS is to map reads generated by ChIP-seq back to the genome used. Massively parallel sequencers generate millions of reads in run of a ChIP-seq experiment. Mapping these reads in a genome as large as the human or mouse genomes with traditional techniques like BLAST (33) or BLAT (34) quickly becomes computationally overly expensive. Fortunately, a number of programs have been developed explicitly for the mapping of short-read data. The fastest of these

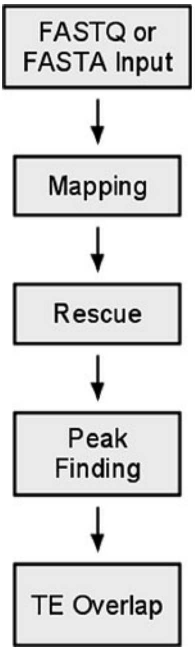

Fig. 14.2. Schematic of the analytical pipeline presented here for finding TE-derived TFBS with ChIP-seq. Each individual step is described in detail in the text along with important caveats, which are listed in 'Notes' section.

are those that employ the Burroughs–Wheeler transform (35) to build a very dense index of the genome, then map reads using the index. We recommend Bowtie for general mapping because of its speed and useful options (*see* **Note 1**). Bowtie is generally the fastest of these aligners, and it can utilize read quality information in the FASTQ format data generated from Illumina sequencing. However, it cannot currently use colorspace reads generated from SOLiD sequencing (*see* **Note 2**).

3.1.2. Read Rescue

Were genomes fully random sequences of the four bases, then almost any ChIP-seq read would be mappable to a unique region of the genome. However, due in large part to the vast number of TE insertions, this is not the case. There are numerous repeated sequences in eukaryotic genomes, and sequence tags derived from these regions may not map unambiguously back to the genome, i.e., they may map to multiple genomic regions with equal probability. The problem of such multiple-mapping ChIP-seq reads arises in part due to their short length. ChIP-seq reads must necessarily be short in order to provide good resolution protein binding locations in the genome; a 500 bp read from ChIP-seq would be easy to unequivocally map to the genome, but would give very little information about the exact location of the DNA–protein interaction. A shorter read, on the order of <50 bp, as most

ChIP-seq data sets contain, gives good resolution regarding the location of the DNA binding, but will have a much greater probability of mapping to multiple locations in the genome. If a TE insertion provides a TFBS, the insertion is very young, and there are many similar TEs in the genome, then it may not be possible to map the ChIP-seq reads from that insertion. For slightly older elements, there will be far fewer possible places to map the reads. Many studies have simply discarded multi-mapping reads for both simplicity of analysis and a desire to be conservative in their findings. However, this becomes an obvious problem when studying TEs, as this will result in the loss of many of the reads coming from TE insertions. To appropriately analyze ChIP-seq data in regard to TEs, some 'rescue' method must be used to resolve reads of the map to multiple locations.

3.1.3. Different Methods of Rescue

There are currently several different schools of thought regarding 'rescuing' reads that map to multiple genomic locations. MAQ (36) is a very commonly used mapping utility for short-read data. When it encounters reads that map to multiple locations with equal probability, it randomly chooses one of the locations to map the tag. This poses problems for TE-derived sequences, as it will dilute the signal from legitimate TFBS, potentially resulting in both false positives and false negatives. This method also ignores information on the local context of potential map positions given by uniquely mapping reads. MUMRescueLite (30, 37) takes this information into account and assumes that multi-mapping reads are more likely to come from regions which already have more uniquely mapping reads and probabilistically determines where a read most likely came from. We recommend that MuMRescueLite be used after the initial mapping to resolve multi-mapping reads.

3.1.4. Peak Calling

Quality mapping is critically important for downstream analysis, and once this has been achieved, the first step is often finding 'peaks' or, more generally speaking, regions that have a density of mapped ChIP-seq reads significantly higher than the background (*see* **Note 3**). These peaks are the regions bound by the TF that are being looked at in the ChIP assay and should contain the TFBS. Methods for peak calling, and indeed the area itself, are still new, and while there is work to be done in the area, there are several quality software choices available for identifying peaks in ChIP-seq data. Quantitative Enrichment of Sequence Tags (QuEST) is reviewed in **Chapter 10** and CisGenome in **Chapter 9**. PeakSeq (38) and SISSRs (39) are two widely used utilities, and in this review, we recommend SISSRs due to its good documentation.

3.1.5. Finding TE-Derived TFBS

SISSRs attempts, and in general is highly successful at, finding the TFBS to within a few tens of base pairs based on the strand orientations of reads forming the peak, as well as the density of reads in the region. Ideally, the TFBS would always be at the point of highest read density. In reality, it is very often co-located with the highest density or if not that then very near by, and SISSRs is correct in its predictions the large majority of the time. What this means, practically, is that finding those regions identified by SISSRs that are contained within TEs will tell us which TFBS are TE derived (*see* **Note 4**). This can be accomplished in a number of ways, the simplest being the creation of two BED-formatted custom tracks for the UCSC Genome Browser (31), one from the predicted TFBS and one from the TEs, and uploading them to the browser. Then, the table browser can be used to intersect the tracks (*see* **Note 5**). Below, we provide a specific step-by-step example of how this can be done using the software cited in **Section** 2.

3.2. Example

Here we provide an example using ChIP-seq data for the CCCTC-binding factor (CTCF) from the human ENCODE (ENCyclopedia of DNA Elements) project (40). CTCF is zinc finger binding protein with multiple regulatory functions including both transcriptional activation and repression as well as insulator and enhancer blocking activity (41). The ChiP-seq data for CTCF are available at http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeChromatinMap/. For this example, we will be using the first repetition of CTCF and the control. The majority of the steps in this procedure are done from the command line in the Unix/Linux operating system environment.

3.2.1. Mapping

The program Bowtie requires an index for the genome that the user wishes to map the tags to. This is accomplished with the 'bowtie-build' utility. It takes as input a FASTA file that contains the genome in question, the human genome in our example:

```
$bowtie-build <human genome FASTA> <index name>
```

Building the index typically takes several hours depending on the machine, though once built there is no need to build it again for different samples. Bowtie takes as input a FASTQ file and the parameters to control the mapping (*see* **Note 1**), as well as the index to use for the mapping:

```
$bowtie -q -k 10 -m 10 --best --strata <index name>
<FASTQ> <bowtie output>
```

The mapping should be done for both the CTCF ChIP-seq set and the control set. Bowtie is capable of mapping several thousand reads per second, or far more, depending how many cores it is allowed to use (*see* **Note 1**).

3.2.2. Multi-mapping Read Rescue

MuMRescueLite takes all of the information that the Bowtie output has, but the information needs to be rearranged to meet the requirements of MuMRescueLite:

```
$awk ′/./ {print $1"\t"$7 + 1"\t"$3"\t"$2"\t"$4"\t"$4 +
length($5)"\t1"}′ < <bowtie output> > <MuM Input>
```

While the above command may appear daunting, it is simply using awk to rearrange the columns of the Bowtie output and put tabs between them. MuMRescueLite is invoked with a much simpler command:

```
$MuMRescueLite.py <MuM Input> <MuM Output><Window Size>
```

Keeping the window size small will prevent distant reads from rescuing reads that do not really come from the location. We suggest keeping the window size under 100. MuMRescueLite produces output that is the same as the input, with an additional column that represents the calculated probability that the read in question is from that site. Using the desired probability cutoff for multi-mapping read, use awk to create a BED track from the MuMRescueLite output for analysis with SISSRs:

```
$awk ’$8 > <cut off> {print $3"\t"$5"\t"$6"\t"$4}’
< <MuM Output> > <Mapping BED>
```

The output should then be sorted by chromosome, then start, then stop:

```
$sort -k 1,1 -k 2n,2n -k 3n,3n -o <Mapping BED>
<Mapping BED>
```

As with the mapping, the rescue should be done for both sets.

3.2.3. Peak Calling

SISSRs takes as input the two BED files created in the previous step and creates another file with peak calls:

```
$sissrs.pl -i <CTCF File> -b <Control File> -o
<Output File>
```

Use the `-i` option to specify the ChIP set as the input and the `-b` option to specify the control set as the background. The `-o` option tells SISSRs where to write the output. Formatting the output into a BED file will allow overlap of the identified TFBS with TEs in the UCSC genome browser:

```
$awk ’/^chr/ {print 1,2,3}’ < <Output File> >
<TFBS BED>
```

3.2.4. Identification of TE-Derived TFBS

The final step is to upload the SISSRs-identified TFBS, BED-formatted track to the UCSC genome browser as a custom track. The name of the track should be changed so as not to be overwritten by later tracks. Once that is done, create another custom track that will contain only TEs using the table browser.

This can be done by filtering the RepeatMasker track for only those repeats which have a ‘repClass’ of ‘LINE,’ ‘SINE,’ ‘LTR,’ or ‘DNA.’ Intersecting the track of CTCF TFBS with this TE-only track will give those TFBS that reside in TE insertions. If everything has gone right, then there should be examples like that shown in **Fig.** 14.3. Here, two distinct CTCF binding sites are shown for a solo long terminal repeat sequence from the endogenous retrovirus family K (ERVK). Although these particular binding sites were identified solely based on ChIP-seq data, they can also be seen to possess known CTCF binding site sequence motifs at the bound genomic intervals. Thus, a computational survey of TE sequences that possess TFBS motifs may have turned up this example.

Genome wide there are 326 CTCF-bound sites located within ERVK sequences, and ERVK elements show more than an order of magnitude greater likelihood to be bound by CTCF than members of other ERV families. The number of CTCF-bound ERVK sequences suggests that these TE-derived TFBS may play some role in regulating human genes, and in fact many ERVs

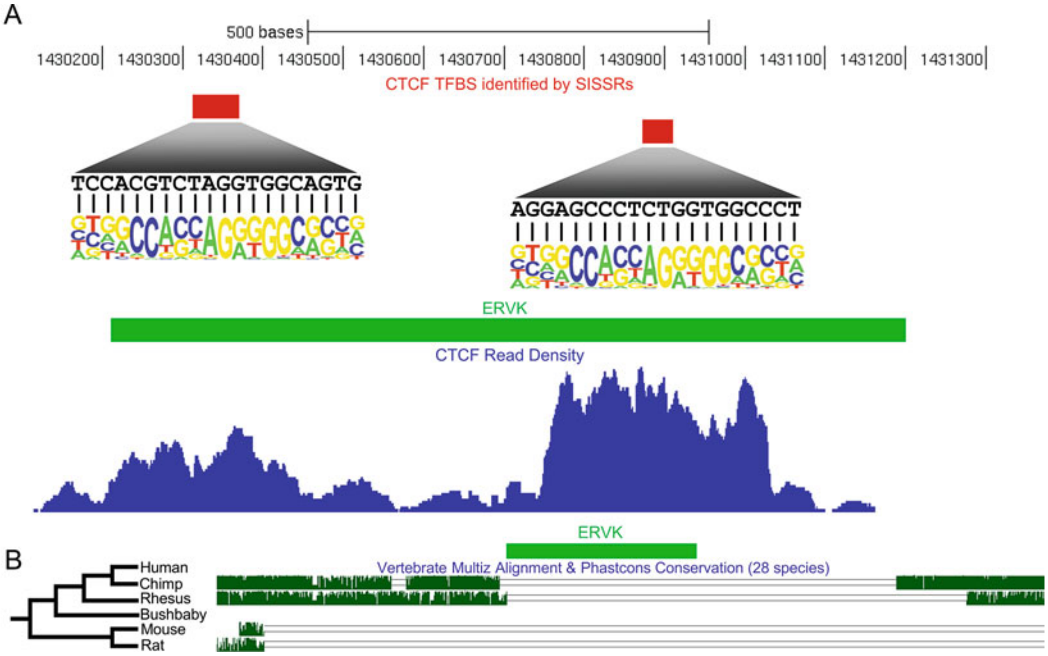

Fig. 14.3. An example of two TE-derived CTCF binding sites found using ChIP-seq data. **a** Two CTCF TFBS identified by the SISSRs program are found within the long terminal repeat sequence of an endogenous retrovirus TE (ERVK). The ChIP-seq read density shows two peaks in the ERVK that correspond to the CTCF-bound regions. Analysis of the bound regions with a CTCF position weight matrix (PWM) (45) using the program CLOVER (46) confirms the presence of two conserved CTCF binding site sequence motifs in the regions identified with the ChIP-seq data. The sequences of the binding sites are shown compared to the sequence logo representing position-specific variation in the CTCF PWM. **b** Regions orthologous to the ERVK insertion site from completely sequenced mammalian genomes were compared using the vertebrate Multiz alignment. Sequence regions conserved between species are shown. Regions flanking the ERVK element are conserved in other mammalian genomes, but the insertion itself is human specific.

are located in close proximity to genes. For instance, the CTCF-bound ERVK shown in **Fig.** 14.3 is located in the 5′ regulatory region ~6 kb upstream of the ATAD3A gene.

ERV sequences in general and members of the ERVK family in particular are young lineage-specific elements that are poorly conserved across species. Phylogenetic analyses revealed that the ERVK family invaded the primate lineage subsequent to the diversification between New World and Old World monkeys (42). Consistent with their recent evolutionary origin in the human genome, ERVK sequences have a mean PhyloP (http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=147315896&c=chr1&g=phyloPCons28way) base-wise conservation score of 0.22, while the genome as a whole has a mean score of 0.47. Therefore, phylogenetic footprinting approaches, which identify regulatory sequences in non-coding DNA by virtue of their sequence conservation, would be exceedingly unlikely to turn up any cases of ERVK-derived TFBS. Indeed, comparison of the CTCF-bound ERVK insertion shown in **Fig.** 14.3 with orthologous mammalian genomic regions indicates that this particular ERVK insertion is human specific and missing in all other mammals. Such lineage-specific TE-derived regulatory sequences may be of particular interest in the sense that they could be responsible for driving regulatory divergence between species (15, 16).

4. Notes

1. Bowtie is currently the fastest short-read aligner available and our preference for mapping short-read data, such as that generated by ChIP-seq or RNA-seq. It has many of the same advantages of MAQ, such as taking quality information into account, but also has other features useful for looking at TE-derived sequences that MAQ currently lacks. Bowtie is also quite memory efficient and it scales well with genome size. Bowtie can be run with the human genome on a computer with 4 GB of RAM, though on such a computer nothing else should be started in the meantime, as when Bowtie is forced out of memory it tends not to recover. Bowtie has a large number of options for controlling mapping and output, which can be listed by executing bowtie with no arguments. The more important options are listed and explained here:

   `-k <integer>` this option is critically important among those available. This option tells bowtie that it should report more than one mapping, as by default it reports only the first. At the current time, MAQ will not report more than one mapping. Currently, MAQ will use the quality scores to choose a location and assign the mapping a quality of 0. Output of multi-mapping reads and their possible location is essential for the rescue and analysis of TE-derived sequences.

`--best` giving this option will cause bowtie to report only those mappings which have the highest quality and is recommended if you have the FASTQ data and not just the FASTA data of base calls. This can greatly reduce the number of multi-mapping reads.

`--strata` This option is used along with the `--best` option and will cause bowtie to return only the highest quality mappings.

`-m <integer>` will eliminate reads that map more than *m* times. We suggest making it the same as `k`. This will remove reads that map to so many places in the genome that they could likely never be placed with confidence.

One major advantage of Bowtie is that it allows for the easy use of multiple cores, which every desktop shipped in the last ~3 years has. Speed will become increasingly important as the number of reads generated per run increases. On a dual-core machine, such as a machine with an Intel Core Duo, only one core is advisable. However, on a quad-core machine, it is generally advisable to use two or three cores. On an eight-core machine six cores are recommended. The number of cores (processors) is set with the `-p` option. In some unfortunate cases, FASTQ files from a ChIP-seq experiment are not available, and only the base calls are supplied. In this case, you would not supply the '`-q`' flag to indicate FASTQ format. It is in these cases that the rescue is especially important.

2. The ABI SOLiD sequencing platform does not produce base calls like the Illumina platform, but rather 'color' calls that represent transitions between two bases. Bowtie cannot currently map colorspace reads, and we suggest the SOCS program for this purpose (43). Like Bowtie, it has generally low memory requirements and is also capable of using multiple cores when available.

3. Though many peaks from ChIP-seq data will be quite large and obvious, others may be closer to the background noise. Complicating this is that the background in ChIP-seq is non-random and tends to form peaks of its own. Most peak-finding utilities will look for peaks with just the ChIP-seq data alone, but many also allow the use of both the ChIP-seq data and a control set. By comparing the control set and the experimental set, false positives that result from peaks not related to the ChIP can be removed.

4. While SISSRs and other peak finders do a very good job of finding the actual TFBS from ChIP-seq data, they may still be off on occasion. A more accurate way to find the exact TFBS is to scan the identified TFBS, along with their flanks, with a PSSM for the TFBS motif with a program such as MAST (44). This will give the exact location of the TFBS if it exists in the peak region.
5. In this chapter, we suggest using the UCSC Genome Browser and table browser for the overlap of the identified TFBS and transposable elements. This is very simple to do, but requires loading BED-formatted tracks to the browser and (relatively) lots of manual work. 'Kent Source Tree' is a large series of utilities, many of which form the back end of the browser. One such utility, 'bedOverlap,' will overlap two sets of tracks without having to upload them to the browser. Numerous other useful utilities include the 'bedItemOverlapCount' utility that can produce custom 'wiggle' tracks for the UCSC Genome Browser, which visualize the density of ChIP-seq reads, and hence protein binding intensity, along the genome. Compilation and installation of the Kent Source Tree is not always easy, but is recommended if possible.

Acknowledgments

King Jordan was supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). King Jordan and Andrew B. Conley were supported by NIH HG000783 granted to Mark Borodovsky. The authors would like to thank Jianrong Wang for help with the CTCF analysis.

References

1. Lander, E.S., Linton, L.M., Birren, B., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
2. Wicker, T., Sabot, F., Hua-Van, A., et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 8, 973–982.
3. Doolittle, W.F., and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284, 601–603.
4. Orgel, L.E., and Crick, F.H. (1980) Selfish DNA: the ultimate parasite. *Nature* 284, 604–607.
5. Gould, S.J., and Vrba, E.S. (1982) Exaptation; a missing term in the science of form *Paleobiology* 8, 4–15.
6. Jordan, I.K. (2006) Evolutionary tinkering with transposable elements. *Proc Natl Acad Sci USA* 103, 7941–7942.
7. Kidwell, M.G., and Lisch, D.R. (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* 55, 1–24.
8. Cohen, C.J., Lock, W.M., and Mager, D.L. (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* 448, 105–114.

9. Conley, A.B., Piriyapongsa, J., and Jordan, I.K. (2008) Retroviral promoters in the human genome. *Bioinformatics* 24, 1563–1567.
10. Zemojtel, T., Kielbasa, S.M., Arndt, P.F. et al. (2009) Methylation and deamination of CpGs generate p53-binding sites on a genomic scale. *Trends Genet* 25, 63–66.
11. Wang, J., Bowen, N.J., Chang, L. et al. (2009) A c-Myc regulatory subnetwork from human transposable element sequences. *Mol Biosyst* 5, 1831–1839.
12. Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9, 397–405.
13. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87.
14. Silva, J.C., Shabalina, S.A., Harris, D.G. et al. (2003) Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet Res* 82, 1–18.
15. Marino-Ramirez, L., and Jordan, I.K. (2006) Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biol Direct* 1, 20.
16. Marino-Ramirez, L., Lewis, K.C., Landsman, D. et al. (2005) Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet Genome Res* 110, 333–341.
17. Zhang, Z., and Gerstein, M. (2003) Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J Biol* 2, 11.
18. Lowe, C.B., Bejerano, G., and Haussler, D. (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci USA* 104, 8005–8010.
19. Santangelo, A.M., de Souza, F.S., Franchini, L.F. et al. (2007) Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. *PLoS Genet* 3, 1813–1826.
20. Nishihara, H., Smit, A.F., and Okada, N. (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res* 16, 864–874.
21. Sasaki, T., Nishihara, H., Hirakawa, M. et al. (2008) Possible involvement of SINEs in mammalian-specific brain formation. *Proc Natl Acad Sci USA* 105, 4220–4225.
22. Hirakawa, M., Nishihara, H., Kanehisa, M. et al. (2009) Characterization and evolutionary landscape of AmnSINE1 in Amniota genomes. *Gene* 441, 100–110.
23. Smith, A.M., Sanchez, M.J., Follows, G.A. et al. (2008) A novel mode of enhancer evolution: the Tal1 stem cell enhancer recruited a MIR element to specifically boost its activity. *Genome Res* 18, 1422–1432.
24. Pang, K.C., Frith, M.C., and Mattick, J.S. (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* 22, 1–5.
25. Johnson, R., Gamblin, R.J., Ooi, L. et al. (2006) Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res* 34, 3862–3877.
26. Thornburg, B.G., Gotea, V., and Makalowski, W. (2006) Transposable elements as a significant source of transcription regulating signals. *Gene* 365, 104–110.
27. Johnson, D.S., Mortazavi, A., Myers, R.M. et al. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316, 1497–1502.
28. Bourque, G., Leong, B., Vega, V.B. et al. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* 18, 1752–1762.
29. Langmead, B., Trapnell, C., Pop, M. et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25.
30. Hashimoto, T., de Hoon, M.J., Grimmond, S.M. et al. (2009) Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite. *Bioinformatics* 25, 2613–2614.
31. Kuhn, R.M., Karolchik, D., Zweig, A.S. et al. (2009) The UCSC genome browser database: update 2009. *Nucleic Acids Res* 37, D755–D761.
32. Karolchik, D., Hinrichs, A.S., Furey, T.S. et al. (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res* 32, D493–D496.
33. Altschul, S.F., Madden, T.L., Schaffer, A.A. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389–3402.
34. Kent, W.J. (2002) BLAT – the BLAST-like alignment tool. *Genome Res* 12, 656–664.
35. Burrows , M., and Wheeler, D.J. (1994) A block-sorting lossless data compression algorithm. *Digital Systems Research Center.*
36. Li, H., Ruan, J., and Durbin, R. (2008) Mapping short DNA sequencing reads and

calling variants using mapping quality scores. *Genome Res* 18, 1851–1858.

37. Faulkner, G.J., Forrest, A.R., Chalk, A.M. et al. (2008) A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* 91, 281–288.
38. Rozowsky, J., Euskirchen, G., Auerbach, R.K. et al. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* 27, 66–75.
39. Jothi, R., Cuddapah, S., Barski, A. et al. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* 36, 5221–5231.
40. Birney, E., Stamatoyannopoulos, J.A., Dutta, A. et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.
41. Gaszner, M., and Felsenfeld, G. (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* 7, 703–713.
42. Sverdlov, E.D. (2000) Retroviruses and primate evolution. *Bioessays* 22, 161–171.
43. Ondov, B.D., Varadarajan, A., Passalacqua, K.D. et al. (2008) Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. *Bioinformatics* 24, 2776–2777.
44. Bailey, T.L., and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 14, 48–54.
45. Kim, T.H., Abdullaev, Z.K., Smith, A.D. et al. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 128, 1231–1245.
46. Frith, M.C., Fu, Y., Yu, L. et al. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* 32, 1372–1381.