**Lecture 14 - Epigenomics/Chromatin States**

**Introduction**

Within the human body there are approximately 210 different cell types all of which share the same genomic sequence. Cells not only develop into distinct types from this same sequence, they also maintain the same cell type over time and across divisions (unless perturbed). This information about the cell type and the state of the cell is called *epigenomic* information. Most of this information is contained in the way DNA is packaged. In order to fit two meters of DNA into a 5-20-diameter cell nucleus and arrange the DNA for easy access to transcriptional machinery, DNA is packaged into chromatin. Nucleosome is the main player of this packaging; it is composed of eight histone proteins (named histone H2A, H2B, H3, H4) and occasionally a linker histone H1 or H5. Each unique histone appears twice in the octamer) with DNA wrapped around the outside of this ball of histones in two revolutions. Each nucleosome consumes approximately 150-200 bp of DNA. While the structure and importance of higher-level packaging of nucleosomes is less known, the lower-level arrangement and modification of nucleosomes is very important to transcriptional regulation and the development of different cell types. Histone proteins H3 and H4 are the most highly conserved proteins in the eukaryotic domain of life. If DNA contains the blueprints of life, nucleosomes contain the blueprints of multi-cellular life.

Nucleosomes have several features that contain epigenetic information. First, their positions on the DNA are important. Nucleosomes are often bound to the promoters of inactive genes. To initiate transcription of a gene, transcription factors (TFs) and the General Factors have to bind to its promoter. Therefore, when a gene becomes active, the nucleosomes located on its promoter are often pushed aside or removed. The promoter will remain exposed until further modifications are made. Hence, nucleosome positioning on the DNA is stable, yet mutable. This property of stability and mutability is a prerequisite for any form of epigenetic information because cells need to maintain its identity of a particular cell type, yet still be able to change their epigenetic state to respond to environmental circumstances.

Second, nucleosomes contain tails of amino acid residues protruding from the ends of their histones. These tails can undergo modification such as methylation, acetylation and phosphorylation. For instance, the fourth residue from the N-terminus of histone H3, lysine, is often methylated at the promoters of active genes. Modifications of this nature are so common that a shorthand notation has been developed. The above modification would be described as H3K4me3 if it were methylated three times. H3 = histone H3, K4 = lysine, 4th residue from end, me3 = 3 methyl groups added.

Lastly, epigenetic information can also be stored by DNA methylation. DNA is often methylated at CpG islands. Methylated cytosine is more likely to mutate into thiamine, which is why transition matrices weight C \( \Rightarrow \) T transitions so heavily. Adenosine, in rare cases, may also be methylated.
Technologies for measurement of epigenomic signals

Given the importance of epigenomic information in biology, great efforts have been made to study these signals on DNA. One common method for epigenomic mark measurement is chromatin immunoprecipitation (ChIP). The procedures of ChIP are described as follows:

1) Cells are exposed to a cross-linking agent such as formaldehyde, which causes covalent bonds to form between DNA and its bound proteins (e.g. histones with specific modifications);
2) Genomic DNA is isolated from the cell nucleus;
3) Isolated DNA is sheared by sonication or enzymes;
4) Antibodies against a specific epigenetic mark are then added to pull out its associated DNA. These antibodies are generated by exposing protein of interest to mammals (e.g. goats or rats). The resulting immune response will cause the production of specific antibodies.
5) The cross-linking between the protein and DNA is reversed and the DNA fragments specific to the epigenetic marks are purified.

![Figure 1: Chromatin immunoprecipitation](Nature_Methods_-_4,_613-614_(2007))

To identify these DNA fragments, one can hybridize them to known DNA segments on an array or gene chip and visualize them with fluorescent marks (called ChIP-chip). Alternatively, one can do massive parallel next-generation sequencing of these fragments (called ChIP-seq). Each sequence tag is 30 base pairs long. These tags are mapped to unique positions in the reference genome of 3 billion bases. The number of reads depending on sequencing depth, but typically
there are on the order of 10 million mapped reads for each ChIP-seq experiment. ChIP-seq is preferred over ChIP-chip nowadays because it has wider dynamic range of detection and can avoid problems such as cross-hybridization in ChIP-chip.

Given the vast amount of reads and the size of human genome, one common problem encountered in all high-throughput sequencing is the algorithmic efficiency of assigning reads to the best matching location. There are two major approaches in matching short reads to the genome. The first one is a hashtable-based approach such as MAQ, which basically trying to put all the reads in a hashtable and use it to search for matches in the genome. The second approach, **Burrows-Wheeler transformation**, is preferred because it is much faster than MAQ. The details of Burrows-Wheeler transformation are described as follows and illustrated in figure 2:

1. For a given reference genome, add a special character (e.g. $) and then generate all the short reads.
2. Sort strings lexicographically, i.e. in alphabetical order based on the first column of each string, with special character comes first.
3. The last column of the sorted list of strings contains the transformed genome. An example is given in figure 2a.

**Key properties of such transformation:**
1. All occurrences of the same suffix are effectively next to each other as opposed to scatter throughout the genome, as illustrated in figure 2c.
2. The \(i\)th occurrence of a character in the first column corresponds to the \(i\)th occurrence in the last column, as illustrated in figure 2b.

![Figure 2: Burrows-Wheeler transform. (a) The Burrows-Wheeler matrix and transformation for 'aacaag'. (b) Repeatedly apply the last first (LF) mapping to recover the original text (in red on the top line) from the Burrows-Wheeler transform (in black in the rightmost column). (c) Steps taken by to identify the range of rows, and thus the set of reference suffixes, prefixed by 'aac'. Langmead et al. Genome Biology 2009 10:R25](image)

Langmead et al. Genome Biology 2009 10:R25
After reads are aligned, signal tracks can be computed. This data can be ordered into a long histogram spanning the length of the genome and indicating the number of reads (or degree of fluorescence in the case of ChIP-chip) found at each position in the genome. More reads (or florescence) means the epigenomic marker of interest is most often present at this particular location. For histone modifications, peak/interal calling methods are based on univariate Hidden Markov Model, or scan statistics (count the reads within bins of certain size and apply statistical analysis). Problems tend to arise with wider domains due to the ambiguity of calling them one large peak or multiple smaller peaks.

There are many ways to analyze epigenomic marks. For instance, one can aggregate signal (such as H3K4me3) on known feature types (e.g. promoters of genes with high or low expression levels). Additionally, one can incorporate machine learning methods to derive epigenomic features that are predictive of different types of genomics elements such as promoters, enhancers or large intergenic non-coding RNAs.

**Annotate the genome using chromatin signature**

**Formulation of the problem**: Epigenetic factors play an important role biologically. They can act to recruit protein elements or conversely to inhibit DNA access. Over 100 distinct histone modifications have been described. Different combinations of these chromatin modifications, when taken together, can help determine how a region of DNA is interpreted by the cell (i.e. as a transcription factor binding domain, a splice site, an enhancer region, an actively expressed gene, a repressed gene, or a non functional region). Stated another way, DNA can take on a series of (hidden) states (coding, noncoding, etc). Each of these states emits a specific combination of epigenetic modifications (H3K4me3, H3K36me3, etc) that the cell recognizes. We want to be able to predict these hidden, biologically relevant states from observed epigenetic modifications.

**The Solution**: Use a multivariate Hidden Markov Model (HMM) for the *de novo* identification of ‘chromatin states’, or biologically meaningful and spatially coherent combinations of chromatin marks, in a given cell population (e.g. CD4+ T-cells). As stated above, ‘chromatin states’ are considered our hidden states. States can switch to other states with certain ‘transition’ probabilities, and each state emits a certain combination of epigenetic markers. This model can capture both the functional ordering of different states (e.g from promoter to transcribed regions) and the spreading of certain chromatin domains across the genomes. Since we do not know what biological chromatin states exist (and how many there are), Baum-Welch training will be used. There are a few differences between this model and the HMM that we have studied in the past:

**Emission of a vector**: In HMM models from previous lectures, each state would emit either a single nucleotide or a single string of nucleotides at a time. In this HMM model, each state emits a combination of epigenetic marks with a specific probability. Each combination can be represented as an n-dimensional vector where n is the number of epigenetic modification being analyzed. For example, assuming you have four possible epigenetic modifications: H3K4me3, H2BK5ac, Methyl-C, and Methyl-A, a sequence containing H3K4me3 and Methyl-C could be presented as the vector (1, 0, 1, 0). This
data binarization can generate biologically interpretable models that can be robustly learned.

However, a present call for each epigenetic mark is not as exact as discrete nucleotides A,G,C or T. Here, let \( C_{ij} \) be the number of reads for mark \( i \) detected by ChIP-seq, mapping to bin \( j \). Let \( \lambda_i \) be the average number of reads mapping to a bin for mark \( i \). The input for feature \( i \) becomes ‘1’ if \( P(X>C_{ij}) < 10^{-4} \) where \( X \) is a Poisson random variable with mean \( \lambda_i \). In order words, the read enrichment for a specific mark has to be significantly greater than a random process of putting reads into bins.

Additionally, for \( M \) input marks each state \( k \) has a vector of \((p_{k1}, \ldots, p_{kM})\) of parameters for independent Bernoulli random variables to determine the emission probability for an observed combination of marks.

**State Lengths:** Since the emissions are no longer nucleotide sequences, but epigenetic marks, it becomes necessary to set a specific length for each state so that the presence or absence of an epigenetic mark can be measured in this region. For the HMM discussed in class a length of 200 nucleotides was established.

**Choosing the number of states to model:**

As with most machine learning algorithms, increasing the number of categories (i.e. hidden states) during training will increase the variability captured by a model. However, part of this variability is due to limited sampling of the true population in our training set. Extra categories can only reduce the variability in a particular sample, but not the true population. This is often called over-fitting of the data. Bayesian Information Criterion (BIC) is a common technique for optimizing the number of states permitted in a model given finite training data. Using BIC, one can visualize the increasing power of a model as a function of the number of states. Generally, one will choose a value for \( k \) (the number of states) such that the addition of more states has relatively little benefit in terms of predictive power gain. However, even with BIC, the resulting model is likely to have more states than an ideal model with most useful states for biological interpretation. The reason is that the human genome is so big that statistically significant differences are easy to find, yet many of these differences are not biologically significant.

To solve this problem, a model with 79 states was initially generated based on the best BIC score after being trained on a population of CD4+ T cells using Baum-Welch algorithm. Of note, in Baum-Welch training, different initializations can turn into very different resulting models. In other words, different initialization can confound the result when doing comparisons among models. Hence, nested initialization rather than randomized initialization was used in generating a model. A single state was greedily removed from the 79-state model. When removing a state the emission probabilities would be removed entirely, and any state that transitioned to it would have that transition probability uniformly redistributed to all the remaining states. The remaining 78 states were used to initialize another round of Baum-Welch training. The correlation between this model and the initial 78 state model was calculated and this process was repeated for all 1-78 state models. This procedure was repeated three times and the model with the maximum
correlation was saved. The number of states for a model to analyze can then be selected by choosing the model trained from such nested initialization with the smallest number of states that sufficiently captures all states offering distinct biological interpretations. The resulting final model had 51 states and it was further verified by showing the independence of individual modifications in each class, demonstrating that the majority of information contained in the epigenetic modifications was captured.

Results: This multivariate HMM model resulted in a set of 51 biologically relevant ‘chromatin states’. However, there were no one-to-one relationship between each state and known classes of genomic elements (e.g. introns, exons, promoters, enhancers, etc) Instead, multiple chromatin states were often associated with one genomic element. Each chromatin state encoded specific biological relevant information about its associated genomic element. For instance, three different chromatin states were associated with transcription start site (TSS), but one was associated with TSS of highly expressed genes, while the other two were associated with TSS of medium and lowly expressed genes respectively. Such use of epigenetic markers greatly improved genome annotation, particularly when combined with evolutionary signals discussed in previous lectures. The 51 chromatin states can be divided in five large groups. The properties of these groups are described as follows and further illustrated in figure 3:

1) **Promoter-associated states (1-11):**

   These chromatins states all had high enrichment for promoter regions. 40-89% of each state was within 2 kb of a RefSeq TSS. These states all had a high frequency of H3K4me3, significant enrichments for DNase I hypersensitive sites, CpG islands, evolutionarily conserved motifs and bound transcription factors.

   These states differed in their functional enrichment based on Gene Ontology (GO). For instance, genes associated with T cell activation were enriched in state 8 while genes associated with embryonic development were enriched in state 4. Additionally, among these promoter states there were distinct positional enrichments. States 1-3 peaked both upstream and downstream of TSS; states 4-7 were concentrated right over TSS whereas states 8-11 peaked between 400 bp and 1200 bp downstream of TSS.

2) **Transcription-associated states (12-28):**

   These states were not predominantly associated with a single mark but rather they were defined by a combination of several marks. Some of these states were associated with spliced exons, transcription start sites or end sites. Of interest, state 28, which was characterized by high frequency for H3K9me3, H4K20me3, and H3K36me3, showed a high enrichment in zinc-finger genes. This specific combination of marks was previously reported as marking regions of KAP1 binding, a zinc-finger specific co-repressor.

3) **Active intergenic states (29-39):**

   These states were associated with several classes of candidate enhancer regions and insulator regions. These regions were usually away from promoters and were outside of transcribed genes. Interestingly, several active intergenic states showed a significant enrichment for disease SNPs, or single nucleotide polymorphism in genome-wide association study (GWAS). For instance, a SNP (rs12619285) associated with plasma
eosinophil count levels in inflammatory diseases was found to be located in the chromatin state 33, which was enriched for GWAS hits. In contrast, the surrounding region of this SNP was assigned to other chromatin states with no significant GWAS association. This can shed light on the possible functional significance of disease SNPs based on its distinct chromatin states.

4) **Large-scale repressed states (40–45):**
These states marked large-scale repressed and heterochromatic regions, representing 64% of the genome. H3K27me3 and H3K9me3 were two most frequently detected marks in this group.

5) **Repetitive states (46-51):**
These states showed strong and distinct enrichments for specific repetitive elements. For instance, state 46 had a strong sequence signature of low-complexity repeats such as (CA)n, (TG)n, and (CATG)n.

Figure 3: Chromatin state definition and functional interpretation. Ernst and Kellis. *Nature Biotechnology*. 2010.
a) Chromatin mark combinations associated with each state. Each row shows the specific combination of marks associated with each chromatin state and the frequencies between 0 and 1 with which they occur in color scale. These correspond to the emission probability parameters of the HMM learned across the genome during model training.

(b) Genomic and functional enrichments of chromatin states, including fold enrichment in different part of the genome (e.g. transcribed regions, TSS, RefSeq 5’ end or 3’ end of the gene etc), in addition to fold enrichment for evolutionarily conserved elements, DNasel hypersensitive sites, CpG islands, etc. All enrichments are based on the posterior probability assignments.

(c) Brief description of biological state function and interpretation (chr, chromatin; enh, enhancer).

The predictive power of chromatin states for discovery of functional elements consistently outperformed predictions based on individual marks. Such unsupervised model using epigenomic mark combination and spatial genomic information performed as well as many supervised models in genome annotation. It was shown that this HMM model based on chromatin states was able to reveal previously unannotated promoters and transcribed regions that were supported by independent experimental evidence. Hence, such genome-wide annotation based on chromatin states can help better interpret biological data and potentially discover new classes of functional elements in the genome.

**Encode:** All of the above work was done in a single cell type (CD4+ T cells). Since epigenomic markers vary over time, across cell types, and environmental circumstances, it is important to consider the dynamics of the chromatin states across different cell types and experimental conditions. The ENCODE project (ENCyclopedia Of Dna Elements, http://www.genome.gov/10005107) has measured 11 different chromatin marks in eight different cell types. Each ChIP-seq experiment in the project produced ~10^7 reads, so the total quantity of data is immense and requires computational analysis. A similar multivariate HMM model was developed for each of these cell types resulting in 15 states. Each cell type was analyzed for class enrichment. It was shown that some chromatin states, such as those encoding active promoters were highly stable across all cell types. Other states, such as those encoding strong enhancers, were highly enriched in a cell-type specific manner, suggesting their roles in tissue specific gene expression. Finally, it was shown that there was significant correlation between the epigenetic marks on enhancers and the epigenetic marks on the genes they regulate, even though these can be thousands of base pairs away. Such chromatin state model has proven useful in matching enhancers to their respective genes, a problem that has been largely unsolved in modern biology.