Good Footprints, Bad Footprints

An introduction to average DNase I footprint profiles and their interpretation.

July 7, 2017

Contents

1 DNase I Footprinting 2

2 Average DNase I Footprint Profiles 2

3 Determinants of Footprint Shape 3

4 Good Footprints 4
  4.1 Visualisation in Sasquatch 4
  4.2 Good Examples of Diverse Factors 4
  4.3 GATA across Protocols and Tissues 5
  4.4 ATAC Footprints 6

5 Bad Footprints 7

6 Estimating Differences 8

7 References 11
1 DNase I Footprinting

Digestion enzymes used for open-chromatin assays, like DNase I, cut the genomic DNA. This cutting is obstructed by proteins binding to the DNA like transcription factors and nucleosomes. Therefore, the cutting is more frequent in accessible (or open) regions of the chromatin. Since these regions are associated with regulatory activity, open-chromatin assays like DNase-seq or ATAC-seq are frequently used to map these regions of open-chromatin. Bringing this type of analysis to a higher resolution, footprinting techniques can be used to map the cut sites at a base pair resolution (Figure 1). For this, only the very ends of every sequencing read is extracted and mapped back to the reference genome (instead of the whole read or fragment for the accumulated DNase-seq signal). At sufficient coverage, footprints can then be identified at single loci as short sequences protected from DNase I cleavage within open-chromatin, potentially pointing to transcription factor occupancy.

![Figure 1: From DNase I digestion to DNase I footprinting.](image)

DNase I cutting is obstructed by proteins occupying the DNA. Sequencing of the digestion fragments and mapping of the reads or entire fragments results in accumulated signals indicating open chromatin sites as peaks. By only mapping the 5’ ends of every read to its respective strand, the individual DNase I cut sites can be resolved at a single bp resolution. At sufficient sequencing depth, this can identify footprints, sequences protected from cleavage, potentially pointing to protein factor occupancy.

2 Average DNase I Footprint Profiles

Accumulating and averaging the footprinting signal over multiple loci creates average profiles of the DNase I cut frequency. This can be calculated over motif instances from matching position weight matrices (PWM) [1–3] or directly over instances of short sequences (k-mers) as Sasquatch works. Each way of deriving these average profiles is associated with specific advantages and limitations, but the profiles share the same core features. If a significant portion of the sequence or PWM matched instances exhibit a footprint at their single locus, the average profile will recapitulate the average footprint over that instance. Average footprint profiles exhibit the clear footprint characteristics, a central, cut depleted, “footprint” region flanked by “shoulders” of increased cut frequency. A straightforward, yet eager interpretation is that: ”A significant portion of the analysed loci exhibits a DNase I footprint and is therefore likely to be bound by (protein) factors that prevent the DNase I cleavage”.

A visualization of an exemplary average profile linked to a sketch of DNase I cutting of a protein bound DNA is represented in figure 2.
Figure 2: Cartoon of DNase I cleavage surrounding a DNA binding Transcription Factor (TF) and the average DNase I cut frequency profile underneath. DNase I cleaves accessible DNA frequently. If the cleavage is blocked, e.g., by proteins binding the DNA, the cut sites accumulate in the regions directly flanking the bound region. Thus, when mapping the cut sites and averaging over multiple instances like E-boxes CACGTG, the average pattern exhibits a typical, central, cut depleted region, flanked by "shoulders" of increased cut frequency.

Average footprint profiles are a useful tool for visualizing, quantifying and interpreting transcription factor binding characteristics. They can be used interpret binding properties of DNA binding proteins [1,3,4] or to infer if binding happens in that context at all.

3 Determinants of Footprint Shape

A strong average footprint profile is characterized by a deep, central footprint and high shoulders. Multiple factors influence the strength and shape of a particular average footprint. This includes:

1. The binding protein: Footprints are highly factor dependent, determined by:
   (a) Size: a large transcription factor or complex will create a broader footprint.
   (b) Affinity and DNA residing time. [4]
   (c) Binding mode: The way a factor binds to DNA may also alter the footprint shape.
2. Tissue: as binding affinity and actively bound instances of a transcription factor vary across tissue types.
3. Number of analysed instances and the fraction of instances actually bound by a protein. Averaging over too few instances results in frequency noise, especially in the flanking regions. In contrast, averaging over more positive, "bound" instance reduces noise and clarifies the footprint shape.
4. DNase-seq protocol: Technical specificities of different DNase-seq protocols, especially fragment size selection. [5]
5. Sequencing depth: In general, we found that more reads enhance the shoulders surrounding a footprint.
6. Biases: Especially DNase I is known to suffer from sequence and methylation biases. [6]

Note: The list above has no claim for being comprehensive.
4 Good Footprints

Since the shape of footprint profiles are determined by the above mentioned factors and more, it is crucial to get a basic feeling for footprint shapes. Being able to recall examples of strong footprints helps assessing the characteristics of unfamiliar profiles. Therefore, this section is designed to walk through a few characteristic examples of strong footprint profiles.

4.1 Visualisation in Sasquatch

In Sasquatch the average footprints are calculated as relative cut probability profiles for a 250 bp window surrounding the searched $k$-mer. Sasquatch averages over all $k$-mer instances found in DNase hypersensitive sites (DHS). The relative DNase I cut probability is than calculated as the relative cut frequency in that window. Hence, a relative cut frequency of 0.01 refers to 1 % of all DNase I cuts in that window being located at that particular position and consequentl a 1 % chance of cutting at this position when introducing a single cut. For visualisation and further downstream quantification, the profiles are smoothed. Each average footprint is visualised as profile along the 250 bp window surrounding the queried $k$-mer. The $k$-mer is located in the center of the profile, indicated by black dashed lines (Figure 3).

4.2 Good Examples of Diverse Factors

The housekeeping transcription factor NRF1 produces one of the strongest footprints across various tissues. We use it as quality control for all newly added DNase-seq data querying the core of the NRF1 motif GCATGCG see figure 3A fr an example from primary erythroid tissue DNase-seq. The actual shape of a footprint is highly factor and protocol dependent. The shoulders may not be as distinct as in the NRF1 example. Some factor profiles primarily exhibit the cut depleted region, flanked by broad shoulders. This is the case when querying the core sequence of the promoter binding co-activator SP1 GGCAGGG (Figure 3B) in a dataset derived from a DNase-seq assay on MCF-7 cells [5]. The Sasquatch repository benefits from a broad range of DNase-seq assays performed as part of the ENCODE consortium [7]. Due to differences in DNase-seq protocols and data analysis the shapes of the footprints may appear different compared to more recent assays. For example, the CACGTG E-box motif, queried and merged from multiple replicates of DNase-seq in the K562 cell line (source: ENCODE University of Washington, UW), produces a weaker average footprint see Figure 3C. However, the shape of the footprint still allows to interpret it with some confidence. Differences regarding protocol, data analysis and tissue specificity should be considered when comparing footprints across different data sets.
Figure 3: Good Examples of average DNase-seq footprints. Average footprints are shown as visualized in Sasquatch as the relative cut frequency within the 250 bp + queried k-mer window. The location of the k-mer is indicated with black dashed lines. The number of occurrences within DHS that were analysed by Sasquatch is indicated as number in the respective header of the plot. 

(A) NRF1 core motif GCATGCG in primary erythroid cells. This k-mer of the housekeeping transcription factor core produces one of the strongest average footprints across various tissues. Therefore, we use it as a control to help getting an overall expression about quality and other shape influencing factors of a newly analysed dataset. 

(B) The core of the SP1 binding motif GGCGGG queried in a DNase-seq dataset on MCF-7 cells [5]. SP1 is a transcription factor that frequently binds to GC-rich promoter regions, as indicated by its high number of occurrence. High occurrences tend to smooth the shoulders. But we observe broader shoulders like these frequently in different datasets. In combination with the clearly cut depleted region, such shapes can still be assigned to average footprints with some confidence. 

(C) E-box motif CACGTG queried in a dataset derived from ENCODE DNase-seq data on K562 cells [3, 7]. The footprint appears weaker compared to other profiles. The actual shoulders are almost vanished due to smoothing of the weaker profile. However, the central region is clearly cut depleted over a large number of k-mer occurrence. Especially, the steep drop at the footprint borders are characteristic for a true average footprint. Note that the number of occurrence is no integer because this dataset was merged from two replicates and the occurrence was determined as the mean occurrence across replicates.

4.3 GATA across Protocols and Tissues

As the shape of footprints is determined by the binding factor and potentially by frequently associated (DNA-binding) proteins, average footprints can exhibit various shapes. A prominent example in erythroid cells are GATA-factors binding to a consensus motif of WGATAA (W = [A/T]). Querying WGATAA in our primary erythroid cell DNase-seq dataset exhibits a strong and broad footprint, where the cut depleted region expands upstream of the k-mer, indicating the frequently binding of the TAL1:GATA complex (rough consensus motif CTGNGWGATAA) (Figure 4 A). Querying the same k-mer in K562 ENCODE datasets generates weaker footprints (Figure 4 B/C). Note that ENCODE derived data ship with more liberal peak calls and increase the number of k-mer abundance strikingly. More recent optimizations of the DNase-seq protocol not fully utilized in the ENCODE data, for example optimized fragment size selection, increase the resolution of the average footprints. However, the general shape can still be recognized.
Figure 4: Average footprints over WGATAA across different tissue datasets. WGATAA was queried in our (A) primary erythroid DNase-seq and in ENCODE K562 DNase-seq datasets derived from (B) UW and (C) Duke University (Duke) DNase-seq data. The ENCODE derived average footprints appear weaker. As can be seen in the respective header, the number of analysed occurrences of the WGATAA \(k\)-mer is significantly higher in the ENCODE data, due the more liberal DHS calling. The core binding is visible across the data sets, but the resolution of the frequent upstream extension differs. Note, the 3' shoulders are slightly smoothed into the \(k\)-mer location. This does not indicate frequent cuts of the motif. When looking at unsmoothed profiles, this effect vanishes.

4.4 ATAC Footprints

ATAC-seq is a versatile and simplifying alternative approach to genome-wide open-chromatin analysis [8]. In general, footprinting and calculation of average footprint profiles using ATAC-seq data is possible, but is associated with certain ATAC specific characteristics and limitations. However, this section is not designed to discuss ATAC-seq footprinting in general. It should rather demonstrate that ATAC-seq derived average footprints appear different in shape and we give a few examples to develop a basic feeling for them.

In general, ATAC derived average footprints have comparably lower shoulders but the cut depleted regions appear broader. This may point to additional steric hindrance caused by the structure of the transposase. From our analysis so far, ATAC footprints exhibit less fluctuations within the cut depleted regions may pointing to more robustness against sequence or methylation biases. More importantly, we have found several \(k\)-mers where ATAC-seq does not reveal average footprints where DNase shows strikingly clear footprints. We have to this date not quantified that proportion fully but do warn that we expect that ATAC-seq suffers from significant sensitiviy loss since certain footprints are not detectable due to limitations of the assay. Overall, given the significant advantages of the ATAC-seq protocol, we expect an increasing amount ATAC-seq data being published. The principle of average footprint profiles is applicable to ATAC-seq data and Sasquatch is able to deal with them, though the biology of ATAC footprints is not yet understood at the same level as DNase-seq footprints. ATAC-seq related analysis should thus be made with appropriate caution and we do only recommended it for investigating the properties of ATAC-seq footprinting in general rather then extensive SNP screens.
Figure 5: Examples for average footprint profiles derived from ATAC-seq data. The profiles presented are derived from one of the replicates of an ATAC-seq assay applied on the lymphoblastoid cell line GM12878 [8]. (A) Querying the NRF1 core results in a strong average footprint. Compared to its DNase-seq counterpart, the shoulders appear lower but the cut depleted site is broader. (B) The E-box motif exhibits a strong footprint flanked by lower but clear shoulders. (C) Querying the SP1 core results in significantly more analysed instances and exhibits a weaker footprint, that is hardly identifiable on its own, at least in this dataset.

5 Bad Footprints

Multiple sources of bias and noise can influence the shape of average footprints. DNase I is known to exhibit a sequence and methylation bias. This bias is at least partly dictated by alterations in the minor groove architecture of the DNA [6]. The underlying sequence influence the overall probability of being cut by DNase I. This causes each k-mer to exhibits a distinct pattern of cleavage in the average footprint profiles. Relative cleavage rates k-mer were determined via DNase-seq assays of de-proteinized DNA [5, 6]. In addition, if the average surrounding sequence of a queried k-mer has a biased base composition (e.g. repeat regions) the sequence bias leads to variations in the flanking regions. Sasquatch filters for repetitive regions but can not guarantee to not suffer from that source of bias. So far, it was proposed that the composition of 6-mers [6] or just 4-mers [4] effectively captures the DNase I sequence bias. We conducted a genome-wide DNase I (and ATAC) digestion of de-proteinized DNA of primary erythroid cells and sequenced the fragments with high coverage. From that we derived linear normalisation factors per 6-mer. Sasquatch uses these factors to correct the DNase I cut sites prior to calculating the profiles. Thus we account for sequence bias as suggested in [4–6, 9]. Furthermore Sasquatch visualises the average cut profile of each queried k-mer in the background model, which allows to quickly compare if a footprint like signature is present in the background and thus likely caused by bias.

Generally, when averaging over a sufficient number of instances, the flanking regions outside footprint and shoulders flatten. Profiles derived from averaging over fewer instances exhibit more fluctuations in the flanking regions that are often still present after smoothing. However, such profiles are most
often still interpretable. But one has to be aware of this to avoid assigning footprints falsely to noise.

The way Sasquatch calculates and displays the average footprint profiles can lead to artefacts falsely assigned with relative high scores. Sasquatch calculates the average cut frequency within a window. Hence, observing sparse cut events in the first place increases fluctuations dictated by all kinds of noise and biases. For example, when querying an unspecified k-mer that is not functional and evenly distributed across the DHS analysed, the profile will be rather flat as cuts are evenly distributed at all positions relative to the k-mer. Consequently, biases at the k-mer itself accumulate relative to the flat profile and therefore disturb the averaging profile vigorously. In general, such biases are present as distinct (single base pair) peaks in the profile. Smoothing deals with this issue but after smoothing may they still appear as relatively sharp peaks and are located at the center over the queried k-mer.

We find most of the k-mers that do not exhibit a footprint to accumulate such central "bias peak" to different strengths. However, as this pattern is frequent and distinct to certain extend, one can easily distinguish bias artefacts from actual footprints by visual inspection of the profiles. Some examples for artefacts are assembled in Figure 6. In general, "true" average footprints exhibit the characteristic "shoulder-footprint-shoulder" pattern. In contrast, bias artefacts show a single peak located over the central k-mer region.

![A] Overlay cut profiles GACCTT unsmoothed vs. smoothed

![B] Overlay cut profiles GACCTT unsmoothed vs. smoothed

**Figure 6: Examples for average footprint artefacts** (A) Un-smoothed and smoothed version of TATGAG, queried in a dataset merged from replicates of K562 cell line DNase-seq assays (ENCODE UW). The k-mer does not exhibit any clear footprint shape. However, bias caused, single base pair fluctuations are smoothed and appear as peaks. Such bias peaks are frequently observed in uninformative (unbound) k-mers. The rather central positioning of the peak and especially the lack of a depleted region and a second shoulder indicate a clear bias artefact. (B) GACCTT queried in a dataset merged from DNase-seq replicates of HepG2 cells (ENCODE Duke). This k-mer exhibits a more complex, bias dictated profile. The highest peak is strongly shifted into the central k-mer region. Because of additional surrounding fluctuations, such pattern may be falsely interpreted as weak average footprints, but the positioning and appearance of the peaks points to bias rather than footprints.

### 6 Estimating Differences

Sasquatch is designed to overlay average footprint profiles and allows for their comparison. One straightforward question is, "if different k-mers exhibit an average footprint in the same dataset (tissue)". This allows to predict if a particular k-mer is on average bound when present in accessible chromatin. From that, one can predict if changes in a certain k-mer, influence it’s potential of "being bound". Sasquatch’s workflows 3 and 4 are designed for predicting the impact of such changes. Examples of overlay profiles, exhibiting clear differences are shown in Figure 7.
Figure 7: Examples of profile overlays for comparison across k-mers. All profiles are derived from a primary erythroid DNase-seq dataset. (A) Overlay of the SP1 core motif against a single base pair variant. The two k-mers exhibit strikingly different profiles. The original motif creates a strong footprint indicating that it is on average bound by a protein factor. The variant lacks the characteristic footprint shape. Furthermore, the occurrences differ significantly as well. While the true SP1 motif occurs frequently within DHS and is thus frequently analysed, the variant seems to lack that enrichment. (B) A qualitative differences in a more complex footprint appears when querying the GATA-core against a single base pair variant at the highly conserved “G” position. Note that the variant occurs less frequent within DHS in that tissue. Therefore, the profile of variant appears more noisy. In general, we found that profiles from the same tissue and protocol are comparable in a qualitative and quantitative manner, as long as a certain number of instances is reached for each k-mer. With decreasing instances and increasing noise, comparisons, especially in a quantitative fashion should be performed with caution. (C) Possible quantitative difference between two variants of the E-box motif. Both k-mers exhibit a clear footprint and the similar shape suggests that both are on average bound by the same or at least similar factors. The less frequent k-mer exhibits a weaker footprint. However, as indicated by the flanking regions, it does not suffer from noise caused by its occurrence number. As both k-mers were queried in the same tissue, this difference may be interpreted as quantitative difference. The less frequent k-mer may have a reduced affinity to the E-box binding factors. (D) Overlay of two different variants of E-box motifs. The k-mers occur comparably often. The footprint shapes show similarities but are different in their range. Thus, they seem to be caused by binding of different factors or complexes. Therefore, these k-mers should rather be compared in a qualitative fashion.
Estimating differences between average footprints should be done between comparable environments, favourably. Ideally, this should be done using data from the same DNase-seq experiment. Beyond that, the footprint profiles derived from Sasquatch may also be used for comparing across tissues in a qualitative fashion. For example, one may be interested in predicting if a certain motif is active in particular tissue opposed to another tissue. When comparing across experiments, one has to consider sequencing depth, DNase-seq protocol and tissue context, especially when addressing quantitative questions. Furthermore, the number of analysed instances should breach a minimal level of \(k\)-mer occurrences (100). In general, we found that fewer instances are associated with more noise, especially in the flanking regions. Above a certain threshold, having more instances reduces the noise slightly but does not contribute significantly to the strength of the footprint itself. Observing a strongly fluctuating cut frequency in the flanking regions also points to low occurrence effects and comparing such profiles should be done with caution.
7 References

References


