Manual: Sasquatch Web-tool

A tool for predicting the impact of sequence variation on DNase I footprinting potential.

Date: April 28, 2016
Authors: Ron Schwessinger, Maria Suciu, Simon J McGowan, Jelena M Telenius, Douglas R. Higgs & Jim R. Hughes
Groups: Genome Biology and Computational Biology Research Group, WIMM, Oxford
Contact: Ron Schwessinger, ron.schwessinger@ndcls.ox.ac.uk
CBRG, managers@molbiol.ox.ac.uk
Jim Hughes, jim.hughes@imm.ox.ac.uk
License: GPL3

Contents

1 Introduction ......................................................... 2
  1.1 DNase I Average Footprints ..................................... 2
  1.2 Quantifying Average Footprints .................................. 3

2 Workflow 1: Query single k-mer .................................. 4
  2.1 Input: .......................................................... 4
  2.2 Output: ........................................................ 4

3 Workflow 2: Dissect longer sequences ............................... 5
  3.1 Input: .......................................................... 5
  3.2 Output: ........................................................ 5

4 Workflow 3: Compare sequences ..................................... 5
  4.1 Input: .......................................................... 6
  4.2 Output: ........................................................ 6

5 Workflow 4: Compare in batch mode ................................ 7
  5.1 Input: .......................................................... 7
  5.2 Output: ........................................................ 7

6 Workflow 5: In silico mutation ...................................... 7
  6.1 Input: .......................................................... 8
  6.2 Output: ........................................................ 8

7 Methods ................................................................. 8
  7.1 Calculating k-mer based Average DNase I cut profiles ........... 8
  7.2 Sequence bias & Normalisation .................................. 9
  7.3 Strand imbalance ............................................... 9

8 References .......................................................... 11
1 Introduction

1.1 DNase I Average Footprints

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. At sufficient coverage, footprints can be identified at single loci as short sequences protected from DNase I cleavage within open-chromatin, potentially pointing to transcription factor occupancy.

To bring this type of analysis to a global scale, various approaches have been proposed to pile-up the average DNase I cut profiles over TF specific motif matches using PWMs [1–3]. Although these approaches demonstrate the great potential of average footprint analysis, there are certain limitations associated with the TF specific viewpoint. PWMs have to be pre-calculated and may vary depending on the underlying tissue and methods used for identification.

To overcome these limitations, Sasquatch was designed to exploit the comprehensive nature of exhaustive, k-mer based approaches and combines them with the single base pair resolution of footprinting analysis. Sasquatch estimates the DNase I footprinting potential of short sequences (k-mers) in the open-chromatin of a tissue of interest. All DHS of that tissue are searched for every possible k-mer of size 5 - 7 bp. For every match, all DNase I cut sites in a 250 bp surrounding window are extracted and attributed to the respective, centric k-mer. Thus a piled-up cut profiles is generated and stored per k-mer. From this repository, every k-mer of interest can be queried and the relative DNase I cut probability within that 250 bp window can be calculated and visualized (Figure 1). Technical biases like sequence preferences and strand imbalance have to be taken into account when analyzing average cut profiles. For more information on how Sasquatch accounts for these biases, please read the more detailed Methods section.

Average DNase I cut profile can be interpreted much like their single locus counter-parts. A characteristic average footprint shows a cut depleted region in k-mer centered region flanked by shoulders of increased cut frequency. On that basis, one can directly infer the protein bound or unbound state of genomic DNA in a tissue specific manner, probing a wide range of transcription factors with a single experimental input. Using simple comparative analysis one can predict the impact of sequence variations on transcription factor binding.

Figure 1: Average footprints over k-mer occurrences in open-chromatin. Sasquatch calculates and analyses average DNase I cut probability profiles in a 250 bp window surrounding every possible k-mer.
1.2 Quantifying Average Footprints

For quantifying average DNase I footprints, Sasquatch first estimates the optimal footprint shoulders. Then the ratio of the average cut probability within the shoulders to the embedded footprint region is calculated as the \textbf{Shoulder-to-Footprint-Ratio}. Background noise will result in SFRs of 1.0 while average footprints caused by average protein occupancy will result in SFRs starting from 1.5 depending on the involved factor and underlying dataset.

Quantification allows to perform comparative analysis. On that basis, Sasquatch can compare sequences associated with sequence variation and estimate the resulting damage in footprinting potential. The damage is calculated as the difference between the reference and the variant SFR ($Dmg = SFR_{ref} - SFR_{var}$). Thus a positive damage is associated with a decrease of the footprinting strength, while a negative damage indicates increase footprinting strength and may point to novel introduced binding sites.

Sasquatch compares on a $k$-mer basis. To increase the potential of capturing larger motifs and the robustness of the results, we implemented and recommend to use a sliding window approach. For example, for comparing a single SNP on a 7-mer basis, extract the sequence from the 13 bp window centered on the variant position and compare this window with reference against the variant base. Sasquatch will employ the sliding window approach by calculating the damage for every pair of $k$-mers and summing up all single values as the total damage score.

Complementary, Sasquatch calculates the relative damage for the highest scoring $k$-mer window, by calculating the percentage change relative to the $k$-mer with the highest SFR.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Predicting damage in footprinting potential. The strength of an average footprint is calculated as the Shoulder-to-Footprint-Ratio. For estimating the damage in footprinting potential, associated with a single base pair change, Sasquatch is designed to compare the surrounding sequences on a sliding $k$-mer basis. The damage at each window is calculated as the difference between the $SFR_{ref}$ and $SFR_{var}$. For the highest scoring $k$-mer window, Sasquatch further calculates the percentage change relative to the highest SFR.}
\end{figure}
2 Workflow 1: Query single $k$-mer

Query a single $k$-mer of interest from a given pre-processed tissue dataset. Visualize and quantify the respective average cut profile.

2.1 Input:

After selecting the Workflow, select the organism and fragmentation type of interest to guide you to the available data for your selection:

Organism: \{human, mouse\}
Fragmentation: \{DNase, ATAC\}

Afterwards, enter the $k$-mer of interest and choose from the available tissue datasets:

$k$-mer: $k$-mer to query of length 5, 6 or 7 bp. Only capital letters of standard FASTA conventions are allowed as input, but $k$-mers may include ambivalent letters (N, S, W, ...)
Tissue: ID of pre-processed tissue dataset
Normalisation: Normalisation mode \{propensity-based (erythroid)\} is the default and the remaining options are for internal testing purposes only
Raw data plot: select if to show the average profiles smoothed or not \{smoothed, no smoothed\}

2.2 Output:

The output headed by a summary of the query made followed by 4 tabs:

Merged:
Visualizes the average cut profile merged from both strands (Figure 3). Note that merging is performed in a strand-specific fashion with respect to the fragmentation type (see also section 7.3). Estimated footprint shoulders and the number of $k$-mer occurrences within DHS are indicated in the plot. Note that for datasets merged from multiple replicates the recorded cut profiles and $k$-mer occurrences are averaged and may therefore result in odd values. The SFRs calculated for the merged and the strand-specific profiles are listed below.

![Figure 3: Example average DNase I cut profile.](image)

Relative DNase I cut probabilities within the 250 bp window are plotted relative to the centric $k$-mer, marked by black, dashed lines. Occurrences of the queried $k$-mer within DHS of the respective tissue are indicated in the title. Red (dashed) lines mark the estimated footprint shoulder regions.

Strand-specific
Visualizes the strand-specific cut profiles for the plus (reference strand) and minus strand respectively.

Background
Visualizes the strand-specific cut profiles in the background (deproteinized genome digested with the respective fragmentation enzyme). They are provided as sanity controls indicating if an observed average footprint might be caused by artifacts determined by the underlying sequence. The titles include the (mappable) genome-wide occurrences of the respective $k$-mer.
**Heavy smooth overlay**
Overlay of the merged sample and background profile, both smoothed with a forceful running Gaussian filter (Figure 4). Intended for quick identification of potential artifact-caused trends within the profiles.

![Heavy Smoothed Profiles CACGTG](image.png)

**Figure 4: Example heavy smoothed overlay plot.** Average cut profiles associated with the respective \( k \)-mer in DHS of the queried tissue and the deproteinized genome-wide background are smoothed with a forceful Gaussian filter and overlayed. This allows a quick judgment if an observed average footprint is likely to be explained by biases present in the background. Here the footprint present in the (blue) data profile is clearly distinct from the expected (red) background profile. The small, centric peak in the background profile is observed frequently and is caused from smoothing single positions of increased cut frequency caused by sequence preference.

### 3 Workflow 2: Dissect longer sequences

Query a longer sequence for analysis. The sequence is split up into \( k \)-mers of the desired length and analysed on a running window basis. Visualizes and quantifies the profiles for each \( k \)-mer step.

#### 3.1 Input:
After selecting the Workflow, select the organism and fragmentation type of interest to guide you to the available data for your selection:
- **Organism:** \{human, mouse\}
- **Fragmentation:** \{DNase, ATAC\}

Afterwards, enter the sequence and \( k \)-mer length of interest and choose from the available tissue datasets:
- **Sequence:** Input sequence. Only capital letters of standard FASTA conventions are allowed as input. Note that the runtime increases with the length of the sequence. (To keep runtime modest and the results interpretable, we limited the input to max 30 bp)
- **Split using kmer:** Select the length of \( k \)-mers to split the sequence. \{5, 6, 7\} [default 7]
- **Tissue:** ID of pre-processed tissue dataset
- **Normalisation:** Normalisation mode \{propensity-based (erythroid)\} is the default and the remaining options are for internal testing purposes only

#### 3.2 Output:
Output of workflow 2 is a \( k \)-mer table listing the \( k \)-mer windows with the respective, estimated SFR values. Links on the right-hand site offer a short-cut to workflow 1 and queries the respective \( k \)-mer in the already chosen tissue for visualization.

### 4 Workflow 3: Compare sequences

Compare the average footprinting potential of two sequences (e.g. a reference against a variant sequence). Longer sequences will be split into \( k \)-mer windows of a selected size and compared on a running window basis. The workflow is designed to estimate differences associated with single base
pair variations. For that, enter two variable \(k\)-mers or for more robustness, include the surrounding sequences. For example, for estimating the impact of one SNP on a 7-mer basis, enter the reference and variant 13 bp window centered on the variant base.

4.1 Input:

After selecting the Workflow, select the organism and fragmentation type of interest to guide you to the available data for your selection:

- **Organism**: \{human, mouse\}
- **Fragmentation**: \{DNase, ATAC\}

Afterwards, enter the sequence 1 (reference) and sequence 2 (variant) you want to compare. Select the \(k\)-mer length to split the sequences of interest and choose from the available tissue datasets:

- **Sequence1**: Input sequence 1 (treated as reference). Only capital letters of standard FASTA conventions are allowed as input. Note that the runtime increases with the length of the sequence. (To keep runtime modest and the results interpretable, we limited the input to max 20 bp)
- **Sequence2**: Input sequence 2 (treated as variant). Note that Sasquatch was designed for single base pair changes. Thus the sequences to compare must be of equal length. (max 25 bp)
- **Split using kmer**: Select the length of \(k\)-mers to split the sequence \{5, 6, 7\}. [default 7]
- **Tissue**: ID of pre-processed tissue dataset
- **Normalisation**: Normalisation mode \{propensity-based (erythroid)\} is the default and the remaining options are for internal testing purposes only

4.2 Output:

Output of workflow 3 is a comparison table listing for every \(k\)-mer window, the estimated SFRs for the respective sequence \(k\)-mers and the calculated damage in footprinting potential for the merged and the single strands respectively. The calculated damage is the difference between the SFRRef and the SFRvar. For visual inspection, an overlay plot is generated for every window step and visualized in separate tab (Figure 5). The overlay profiles plot the sequence 1 (reference) in blue and sequence 2 (variant) in red. The compared \(k\)-mers and their respective occurrence within DHS in the tissue are indicated alongside each profile.

![Figure 5: Example of variant associated profiles](image)

Relative DNase I cut probabilities within the 250 bp window are plotted for the respective reference (blue) and variant (red) \(k\)-mer. The respective \(k\)-mer occurrences are indicated alongside the cut profiles. In this example, it a striking change in footprinting potential is associated with the single base pair variant, while the \(k\)-mer occurrences are comparable.
5 Workflow 4: Compare in batch mode

Batch query multiple sequence (reference vs. variant) pairs. The workflow is designed to batch analyze and prioritize a larger number of sequence pairs. Each pair is analyzed as in Workflow 3 and summarized to give total predicted damage per variant. Variants are then sorted by their predicted damage and listed. For a quick indicator of potentially involved transcription factors, selected sequences can be matched with PWM motifs from JASPAR2014 [4].

5.1 Input:

After selecting the Workflow, select the organism and fragmentation type of interest to guide you to the available data for your selection:

Organism: {human, mouse}
Fragmentation: {DNase, ATAC}

Afterwards, enter a space or tab separated, two column list of reference and variant sequence pairs to analyse. Select the k-mer length to split the sequence and choose from the available tissue datasets:

sequence pairs: Two column (tab or space separated) table of sequence pairs with the reference in the left and the variant in the right column (e.g. For 7-mer based analysis, use the 13 bp windows centered on the variant base.)
Split using kmer: Select the length of k-mers to split the sequence {5, 6, 7}. [default 7]
Tissue: ID of pre-processed tissue dataset
Normalisation: Normalisation mode {propensity-based (erythroid)} is the default and the remaining options are for internal testing purposes only
Search mode: Select mode how the total predicted damage for a sequence pair is calculated. Mode “exhaustive” sums up all k-mer pairwise damages to a sum of damage, offering more robustness. Mode “local” takes the highest k-mer pairwise damage

5.2 Output:

kmer table
Output is a table listing the reference and variant sequence, the respective k-mer pair that yielded the highest damage, the reference and variant SFR and the total damage predicted. Entries are sorted by their total damage, with high damage scores indicating decreased footprinting potential (disrupting of a potential binding site) and negative damage scores indicating enhanced footprinting potential (e.g. potentially introducing novel binding sites). Each entry is assigned with a link to analyse and visualize this sequence pair in detail using workflow 3.

Jaspar analysis
Allows to select total damage thresholds above and below which, sequences should be matched against JASPAR2014 motif PWMs. This was implemented using the “JASPAR2014” and “TFBSTools” R-packages. For each selected sequence pair, the sequence harboring the highest SFR is selected and matched against PWMs from the same organism. Relative scores are retrieved and all matches with a score ≥ 0.8 are displayed with a link to investigate the PWM in JASPAR.

6 Workflow 5: In silico mutation

Query a genomic region of interest and perform an in silico mutation by mutating every reference base to every possible variant. This workflow allows A) to identify variants and clusters of variants that have a high footprint damaging potential; B) Scan a region for sequence parts that appear susceptible to variations, thus pointing to potentially actively bound sequences in a tissue-specific manner. This perspective yields an interesting alternative to looking at sequence conservation for finding regulatory important regions.
6.1 Input:

After selecting the Workflow, select the organism and fragmentation type of interest to guide you to the available data for your selection:

Organism: {human, mouse}
Fragmentation: {DNase, ATAC}

Afterwards, select the genome build and chromosome and enter a start and stop coordinate (1-based). Select the k-mer length as split basis and select from the available tissue datasets:

Genome build: Genome build for the reference sequence and coordinates. {hg18, hg19, mm9}. [default hg18]
Region start: Start coordinate for region of interest (1-based)
Region end position: End coordinate for region of interest (1-based)
Tissue: ID of pre-processed tissue dataset

6.2 Output:

Output are two in silico mutation plots. A full (combined) plot and a plot only showing a the positive damage scores. We switching between plot modes useful, depending on the question. If looking for the impact of variations and if potentially site introducing variants should be captured, the combined plots serves this purpose great by capturing all variation. However, plotting all negative damage score reduces the resolution of the positive damage part. If only looking for TFBS disrupting mutations or if scanning a region for parts that are in particular susceptible to mutations, plotting only the positive damage values leaves you with more resolution for that purpose.

![Figure 6: Example in silico mutation plots.](image)

The full mutation and damage prediction data table can be viewed and downloaded from the link button.

7 Methods

7.1 Calculating k-mer based Average DNase I cut profiles

Given a DNase-seq input (aligned reads and called peaks), we calculated the k-mer based, average DNase I cut profiles in the genome-wide open-chromatin. For every possible k-mer with $k \in \{5, 6, 7\}$,
the average DNase I cut profile of the 250 bp window surrounding the k-mer was calculated and only k-mer occurrences within DHS were considered. DHS (extended by 125 bp to each side) were scanned with a sliding window of length 250 bp + k-mer length. At every position the mapped DNase I cut sites within the window were recorded and attributed to the centric k-mer. The total number of k-mer occurrence and the summed up surrounding cut profile was stored in a table format. The cut profiles were recorded in a strand-specific fashion. For every processed tissue, cut profiles of every k-mer are stored in a separate file per k-mer length and reference strand. Regions blacklisted for mapability reasons were subtracted from peak sets prior to the analysis.

Each possible k-mer can than be queried from the pre-calculated database. For every retrieves profile, Sasquatch calculates the relative DNase I cut probabilities in the 250 bp window.

7.2 Sequence bias & Normalisation

Although DNase I cuts the DNA in general, the enzyme is known to exhibit certain sequence preferences leading to sequences biases, especially in piled-up profiles [2,5,6].

To correct for the DNase I associated sequence bias, each recorded DNase I cut was multiplied with a weighting factor matching to the 6 bases surrounding the expected cut site. These weighting factors were calculate according to the DNase I sequence preference. For estimating these propensities, deproteinized genomic DNA was digested with DNase I and sequenced according to the protocol described above. DNase I cuts were mapped and average cut profiles of all possible 6-mers within the mappable genome were recorded as described above. The relative sequence cut probability of each 6-mer was than calculated as the recorded number of cuts at the 4th base position of each 6-mer divided by the number of recorded 6-mer occurrences along the entire mappable genome. The relative cut probabilities were shifted to achieve a median cut probability of 1 and relative weights were then calculated to correct the relative sequence cut probability of every 6-mer to 1. A sequence context of 6 was chosen because it has been shown to capture the DNase I sequence bias sufficiently [2,5–7].

7.3 Strand imbalance

DNase I footprints are known to exhibit strand imbalance, showing more reads mapping to the plus (reference) strand in the footprint upstream region and more reads mapping to the minus strand in the footprint downstream region [8]. This pattern can be observed in the derived average cut profiles as well (Figure 7). To account for that, Sasquatch merges strand-specific profiles accordingly. DNase cuts upstream of the k-mer region are retrieved solely from the plus strand profile, while downstream cuts are derived only from the minus strand. In the centre k-mer region, the profiles are averaged.
Figure 7: DNase-seq strand imbalance in average cut profiles.
8 References

References


