University College London

CoMPLEX

Summer Project

Dr Daniel Jeffares, Dr Chris Illingworth &
Prof. Jürg Bähler

Statistical analysis of saturating transposon mutagenesis in fission yeast

Christoph Sadée (15084362) September 5, 2016

Abstract

In this project a transposon mutagenesis library in fission yeast, assembled in the Bahler-Lab at UCL, is investigated using a 3 state and 4 state Hidden Markov Model. An initially assembled Hidden Markov Model was adjusted to allow for the classification of regions within the yeast genome. Classifications group regions by their essentiality to a cells fitness and growth, based on transposon density within that region. This allows for a genome wide classification and the ability to find essential non-coding RNAs as laid out in the result section. Also a very interesting large region, classified as essential by the HMM, has been identified within chromosome III with almost no prior annotation from the PomBase data base. A region, one of many, which has been identified by the HMM as essential to growth and fitness of a cell but very little a prior annotation.

Acknowledgements

Dr. Daniel Jeffares

Thanks to Dr. Jeffares for his guidance and helpful attitude to all project related matters.

Leanne Greche

Thanks to Ms. Greche for the friendly and informative talks about the experimental work behind the data.

Dr. Maarten Speekenbrink

Thanks to Dr. Speekenbrink for his involvement in this project.

Prof. Jürg Bähler

Thanks to Prof. Bähler for his most interesting suggestions and useful inputs.

Contents

1 Introduction

Fission yeast or S. pombe is a popular model organism, allowing the investigation of many cellular processes of eukaryotic cells. A transposon screen has been conducted with the yeast genome, where transposons (2k basepairs) are inserted artificially at random into the yeast genome by transposase hermes. An insertion generally results in the disruption of a gene and hence it's knockout. Saturation transposon mutagenesis describes the saturation or transposon insertion in the yeast genome and accompanied growth of cells. Cells with an insertion in an essential region for growth and fitness are most likely to die, resulting in transposon depleted regions. Therefore previously non annotated regions can be discovered if such regions fall within less well a priory investigated regions. Due to the arbitrary length of depleted regions, under sampling and insertion biases a Hidden Markov Model is employed to classify regions as essential and non essential.

2 Biological Background

An integral part in biological genetics is to identify functional elements in the DNA of an organism. This is commonly performed by knockout of a gene or sequence of interest either at the translational level or by gene deletion. At the translational level mRNA knockdown is performed by RNA interference mechanisms. A short interfering RNA (siRNA) is introduced and complementary to the mRNA of interest, signalling the RNA sequence for cleavage by ribonucleases and abolishing the translational stage [2]. Although successfully used in gene identification, a major downside is the resource intensive methods required for RNAi production and possible incomplete knockdown.

An alternative approach, performed in S. pombe and S. cerevisiae, is the culture of strains with systematic deletions in predicted coding sequences. This allows for screening of cells under a variety of conditions to access the purpose of the deleted sequence and compare it with it's functioning counterpart. The possibility to produce such strains en masse, allows for high throughput data and testing of a wide range of coding sequencing but also has several problems associated with it. One complication is the remain of genes targeted for deletion in large cultures [11] and the inability to identify function of sub regions in an open reading frame (ORF) and non-coding RNA regions.

Each of the above named methodologies has advantages and disadvantages but one inherent problem of both is their approach of targeting a specific sequence at a time and the inevitability to skip functional sequences (such as non-coding elements) when selecting regions of interest. Here instead a transposon screen is described in S. pombe, a more recently developed procedure, established with the event of deep sequencing and discovery of DNA transposons/transposases that can efficiently act in a wide variety of eukaryotic species other than their host of origin [8]. A transposon in the most simple terminology is a strand of DNA (here about 2k nt in length) that is inserted into the organisms genome. For the transposon screen in S. pombe the hermes transposase from musca domestica is an efficient tool for facilitating insertion. A transposon screen makes use of the quasi random insertion mechanism, which causes the insertion to be non specific and intercalate anywhere within the genome. Transposons have in general a disruptive effect when inserted within a coding sequence (CDS), resulting in down regulation of the gene and/or knockout. For the transposon screen, cells are grown during and beyond transposon mutagenesis, causing those with insertions in essential regions to display hindered growth, no growth and/or die, whereas non essential regions are predicted to have an avid display of insertions. This allows, for dense insertion libraries, to highlight regions of importance that

have previously not been annotate and the influence of non-coding elements genome wide [7].

2.1 The model organism

Schizosaccharomyces pombe (S. pombe), also referred to as fission yeast is one of the primarily used model systems of the yeast species beside S. cerevisiae, fission yeast will from here onwards be referred to as yeast, unless otherwise indicated. It is a unicellular organism of the domain Eukaryota and has an approximate size of 3.5μm by 10μm. Three chromosomes and organelles make up the yeast genome of 14M nucleotides with about 5k protein coding genes. This is a much more reasonable size as compared to other model systems such as mice with 2.7 billion base pairs and chromosomes, increasing the complexity of analysis and experiment [9]. Due to its easy care and size it is ideal for large scale studies of cellular processes such as cell cycle, cytokinesis, ribosome biogenesis, regulation of transcription[15] to name but a few and with relevance to other eukaryotic organisms such as humans.

2.2 Hermes transposition mechanisms

Hermes as previously mentioned originates from Musca domestica and is part of class II transposable elements (TE), which use a DNA intermediate (i.e. from a donor plasmid) and "cut and paste" mechanism for insertion. This results in one insertion per donor plasmid as compared to retrotransposons (class I TE's) which first transcribe a sequence into an RNA intermediate before reintegrating it into the genome - "copy and paste" -. Hermes is part of the hAT superfamily of transposons within the class II TE's which are commonly found in animals and plants and with their common ability to transpose in species other than their host of origin. Specifically hermes is the first non-drosphillid class II transposable element which has been converted into a gene vector*, giving it great range to perform transposition within a variety of species and it's use in S. pombe [14].

The transposon screen in S. pombe is started by introducing two plasmids into the yeast cell, each of them a gene vector [8]. One labelled as donor plasmid, carrying the transposon, to be inserted into the yeast genome. There are different possible donor plasmids that have been investigated for use in S. pombe [5] but all have the imperfect terminal inverted repeats (TIR) at either end with a length of 17bp, required by the transposase for target recognition.

The second plasmid, also referred to as expression plasmid, carries the hermes transposase and a promotor, this is to control/ activate the expression of hermes. when promoted it performs the excision of the gene vector and insertion into the yeast genome. Upon artificial (or natural) promotion the Hermes gene, consisting of a single open reading frame, is transcribed and translated into a 612 amino acid long protein with an N-terminal (residues 1-150), a catalytic domain (150-265) and a large α-helical insertion domain (265-552), a more in-depth discussion of the hermes structure is omitted but can be found here [14]. The N-terminal has a DNA binding domain with a nuclear localisation signal which identifies and binds to a specific motif within the TIRs of the transposon on the donor plasmid [10]. At each TIR a synaptic complex is formed by binding between transposon and transposase, followed by allocation of the hermes catalytic domain to both ends, see Figure 2.2. Cleavage is performed and a hair pin structure is formed of the excised DNA sequence[3]. Next the hermes transposon complex or more commonly referred to as the hermes transposon is aiming for a target site for insertion on the yeast genome. Insertion takes place through the involvement of the hermes α-helical insertion domain with a

*DNA molecule to artificially carry foreign genetic material into another cell

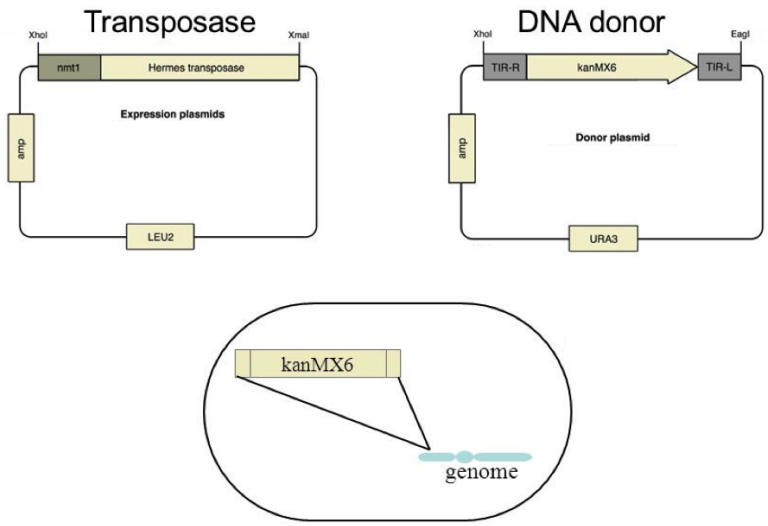


Figure 2.1: Hermes transposition in S. pombe [5]

preference for a specific 8 base pair consensus/motif (a specific sequence of nucleotides recognised by hermes), see next section.

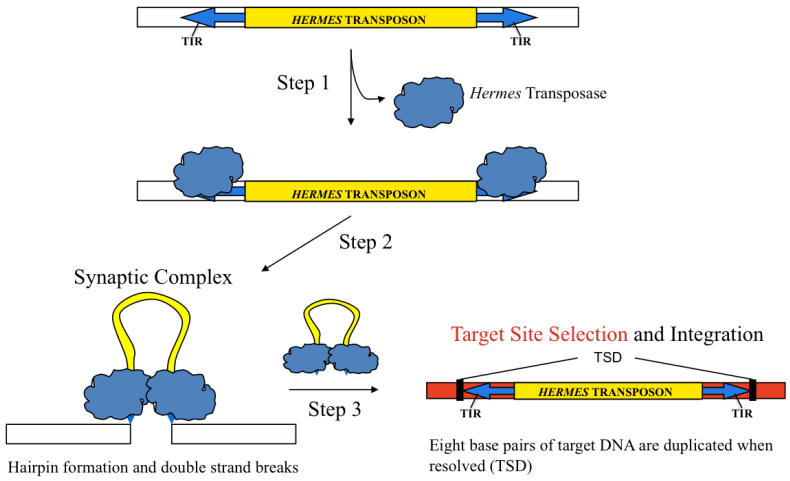


Figure 2.2: Hermes mechanism [14]

2.3 Insertion biases

Hermes is most likely to insert into an 8 bp long consensus/ motif, targeting a sequence of type nTnnnnAn, where n is any base and T and A stand for the usual nucleotides [6]. Upon insertion, target sequence duplication (TSD) takes place and the consensus is found to either side of the inserted transposon Figure 2.3.

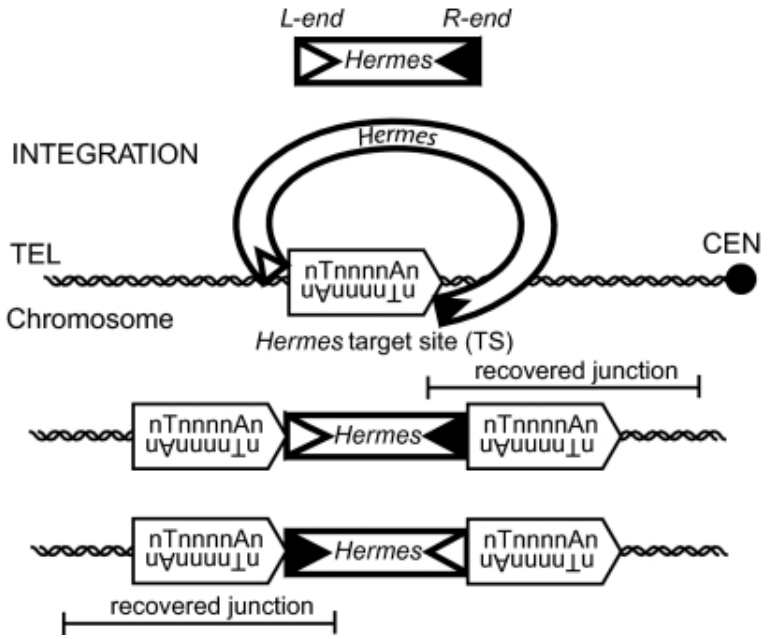


Figure 2.3: Hermes insertion bias and TSD[6]

The preference has been observed in previous screens and quantified for in vitro data Figure 2.4 and in vivo data Figure 2.5[8]. The in vitro data is highly valuable since it has been generated with naked DNA or nucleosome depleted DNA, where nucleosome occupancy is another bias, affecting Hermes insertion.

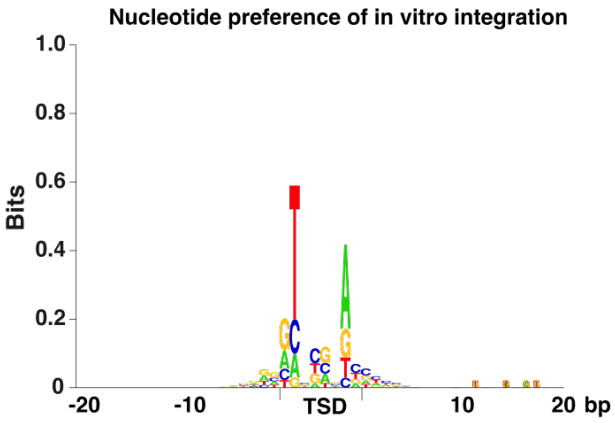


Figure 2.4: Nucleotide preference for in vitro insertion sites [8], with naked DNA (nucleosome depleted) of S.pombe

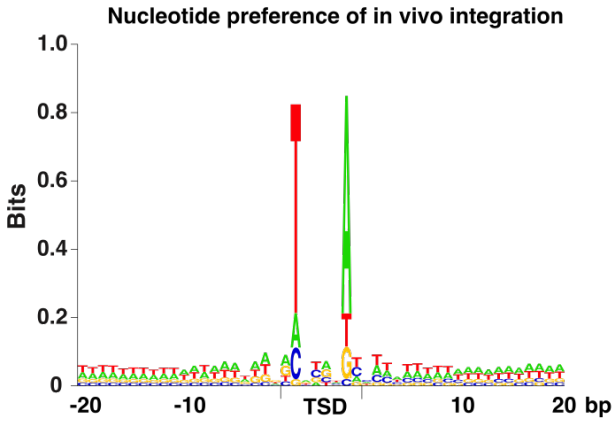


Figure 2.5: Nucleotide preference for in vivo insertion sites [8] of S.pombe

The second major bias affecting hermes insertion is the nucleosome occupancy/density. DNA in eukaryotes is wrapped around a nucleosome molecule or histone octamer, made up of eight histone subunits (H2A to H4). The length of DNA wrapped around the nucleosome can be as accurately determined as 150bp±5 which is linked to the next nucleosome by nucleosome free linker DNA Figure 2.6. During transcription, DNA will unwind from the nucleosome, making it accessible to the transcriptome [16]. Hence it is not surprising that in [8] an increase in insertions was found every 150 bp, in nucleosome free regions, also offering a possible explanation for an increase of insertions at the start of open reading frames such as transcription start sites (TSS)

Figure 2.7. Nucleosome occupancy is generally low at TSS, to allow for access of transcription factor proteins.

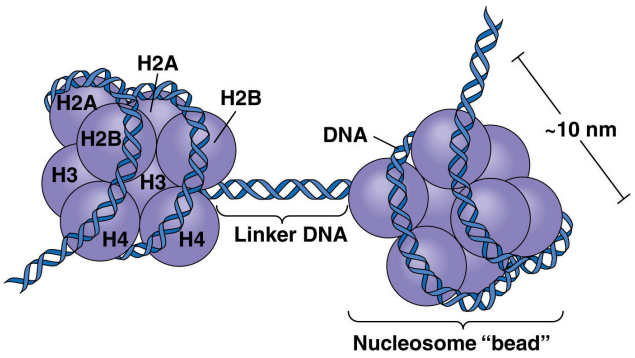


Figure 2.6: Nucleosome structure [13]

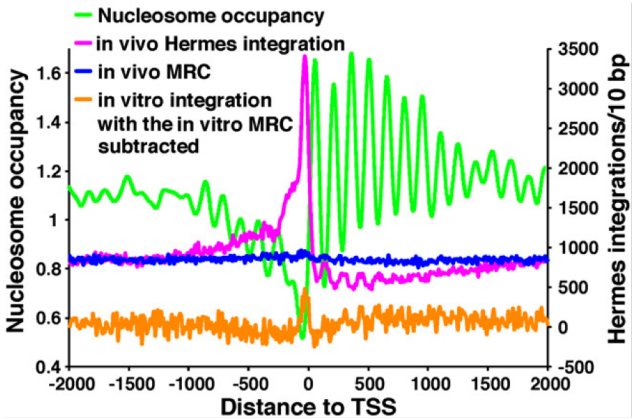


Figure 2.7: Nucleosome occupancy and hermes insertion. MRC is a random insertion signal generated in silico as comparision[8].

3 Data

The data was generated by L. Greche under D. Jeffares supervision in the Jürg Bähler laboratory [1]. The final data file contains several pooled data sets of hermes transposon insertions into the S. pombe genome. The final data file lists each nucleotide as a position with an associated read count larger than zero when insertion occurred at that position and zero when no insertion occurred. To be clear in terminology, a read count refers to how many insertions there at one position. Note this is not due to hermes having inserted several times at the same location within one cell, but the sequencing of several cells each having one insertion at that position. It is assumed that there is one insertion per cell, an appropriate assumption as previously discussed that hermes is part of the hAT transposase family which uses a "cut and paste" mechanism, hence requiring one donor plasmid per insertion and the experiment was optimised to facilitate one plasmid per cell. Then an insertion refers to the position where hermes transposition took place. Two sample regions within the first chromosome are given below.

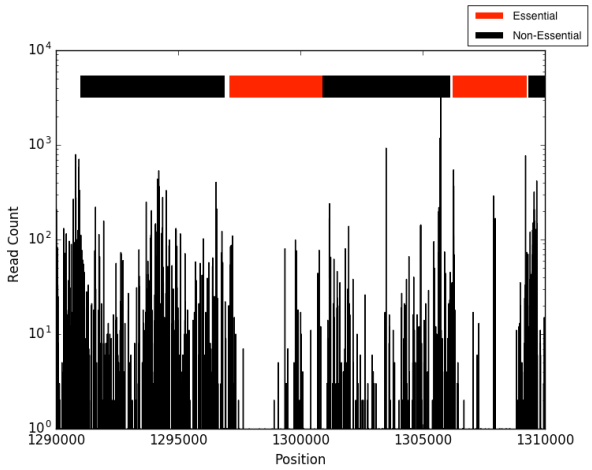


Figure 3.1: Data excerpt from chromosome I with annotations from PomBase[18]

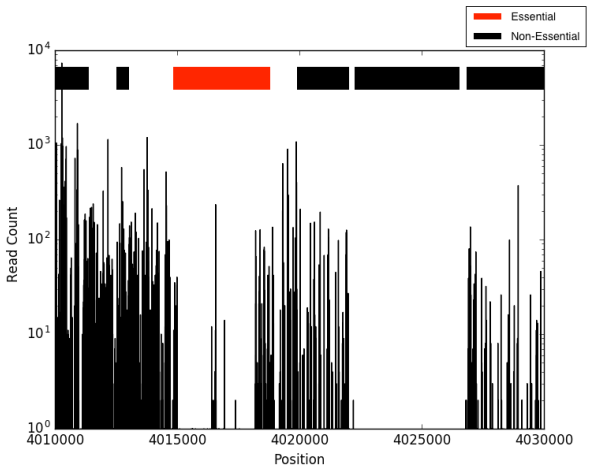


Figure 3.2: Data excerpt from chromosome I with annotations from PomBase[18]

Note that Figures 3.1 and 3.2 plot a region of 20k nucleotides each, making insertions appear next to each other in dense regions. This is generally not the case, even between dense regions there are several non-insertion positions, with an average of 17 zero insertion positions between every insertion for chromosome I. The red and black regions above the insertions are annotated regions from the PomBase [18], which have been identified by previous experiments as essential for growth and/or fitness, these regions have been determined by i.e. gene knockout. Interesting to note is that not all regions are annotated which this study aims to remedy and that a previously annotated region (red), appears to have less insertions and read counts over its span (approx. positions 1,295,500 to 1,300,000) in Figure 3.1, underlining the validity of the data. Of particular interest are regions which have not been highlighted as essential but appear to be of importance for the fitness to the cell, judging from the data, see Figure 3.2 positions 40,225,000 to 40,275,000.

Using further annotations from PomBase, one can find the average read counts and insertions per site for each type of annotation. Figure 3.3 B very nicely displays how essential coding sequences (those that led to cell death when knocked out) are very conserved in the data and that regions of long non-coding RNAs have a high average read count associated with them.

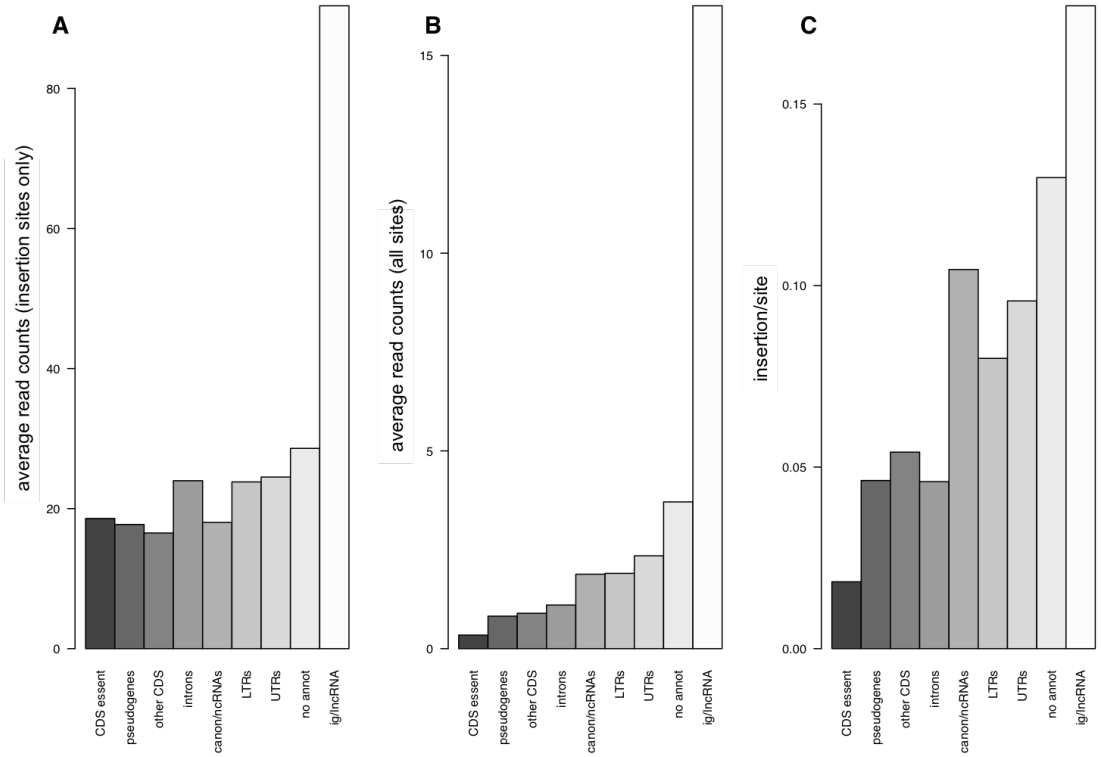


Figure 3.3: **A**: Average read counts of insertions sites that fall into the respective annotated region from PomBase **B**: Average read counts of all sites that fall into the respective annotated region from PomBase **C**: Insertion per site in the respective annotated region from PomBase, obtained by dividing insertions in a region by total of sites within region. **Annotations correspond to**: Essential coding sequences: CDS essent. and pseudogenes. Coding sequences of non-essential genes: other CDS, introns, canonical RNAs (tRNAs, sno/snRNAs and rRNAs), Tf-type retrotransposons and solo LTRs, 5' and 3' untranslated regions (UTRs). Regions without any annotation: no annot. And intergenic long ncRNAs.[12]

4 Model

A Hidden Markov Model (HMM) was first described in [4] to analyse a transposon screen for the Bacterial genome. A HMM is supposed to smooth out variations between insertion sites in close proximity but define larger regions as part of a predefined state. Such states can classify a region as essential (generally few insertions and read counts) or non-essential (insertions are more abundant with higher read counts) as an example of a 2 state model. An initial HMM with 3 states was constructed by Maarten Speekenbrink [17] based on the theory below. The code was debugged, trained on sample data to classify states, altered to incorporate log_2 transformed data and expanded to for a 4th state.

4.1 HMM

First consider a discrete Markov model, before advancing to the more complex hidden markov model. Consider that every position in the 14M base pair long yeast genome could be in one of several predefined states

$$q_t = S_j \tag{4.1}$$

where q_t = state of nucleotide t
t = position of nucleotide with $1 \leq t \leq T$
S_j = state j

This results in a state sequence $Q = q_1...q_t...q_T$, with $T = 14M$ if considering the whole genome at once. It is also easy to analyse sub-domains of the genome separately, such as individual chromosomes. S_j is part of a pre-defined set of states such as

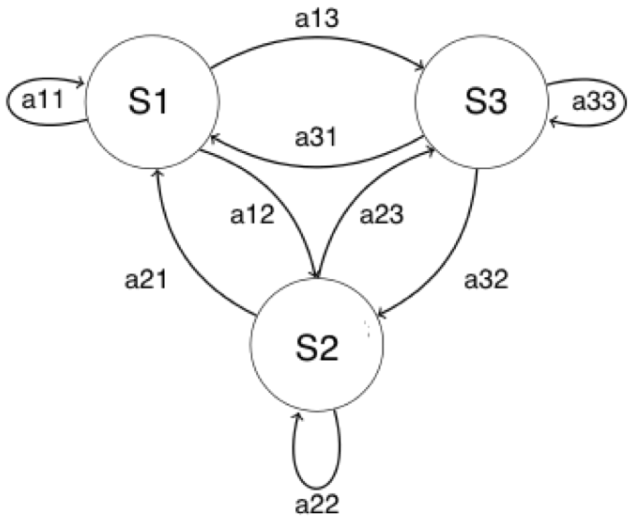
$$S = \{S_1, S_2, S_3\} \tag{4.2}$$

where S_1 = essential
S_2 = intermediate
S_3 = non-essential

The states are in relation to the importance for growth, i.e. if a nucleotide at that position is essential for growth of the cell, intermediate for growth of the cell etc.[†] It should be obvious that it is rather hard to infer if an individual base position on its own is essential, a more sensical notion is the essentiality of regions of the genome. Hence consider that the state at position q_t is predicted by some probability depending on the preceeding states $P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ...]$ or making use of the first order Markov assumption, only the most adjacent preceding state

$$a_{ij} = P[q_t = S_j | q_{t-1} = S_i] \tag{4.3}$$

The so called transition probabilities a_{ij} are part of a Markov chain Figure 4.1 with probabilites for every transition, represented as a transition matrix A, also referred to as transition probability distribution.



$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \tag{4.4}$$

Figure 4.1: Three state, ergodic Markov Chain

[†]This should not be confused with regions in the genome previously annotated as essential, found on a database

For the model an ergodic process is considered, meaning that each state can be reached from another $a_{ij} > 0$. The state of a nucleotide is not directly observable, it has to be inferred by the presence of an insertion and it's read count (increased likelihood to be intermediate or non-essential) or absence (increased likelihood to be essential), resulting in a Hidden Markov Model (HMM)[15].

The HMM is defined by M, the distinct observation symbols per state which corresponds to all possible/found values of read counts with the addition of no insertion (the zero value). Due to the biases (nucleosome and specific motif) a high read count can still correspond to an essential state, therefore all observation symbols are possible for each state. The individual observation symbols v_k are part of the set V

$$V = \{v_1, v_2, ..., v_M\} \tag{4.5}$$

The data can then be interpreted as an observation sequence $O = O_1O_2...O_t...O_T$, where the position is denoted by t and the observation at O_t is v_k.

Each state has an observation symbol probability distribution B associated to it.

$$B = \{b_j(k)\} \tag{4.6}$$
$$b_j(k) = P[v_k \text{ at } t | q_t = S_j] \tag{4.7}$$

Equation (4.9) is the probability to be in state S_j when there is the observation $O_t = v_k$ at position t. Note that B is affected by the bias and hence is not solely based on the observation sequence.

The sole left parameter required to be defined for a complete description of the HMM is the initial state distribution π.

$$\pi = \pi_j \tag{4.8}$$
$$\pi_j = P[q_1 = S_j] \tag{4.9}$$

This is required as each state depends on the previous state by some probability a_{ij} and since there is no preceding state for the first position, there must be some initial state distribution defining the state of the first element at $t = 1$. All required HMM distributions can now be grouped together for simplicity into

$$\lambda = \{A, B, \pi\} \tag{4.10}$$

The problem is now defined by 1) How to best adjust the model parameters λ, given the biases, in order to maximise the probability of finding the given observation sequence of insertions and read counts. Using the maximised model parameters and observation sequence 2) What is the best possible state sequence $Q = q_1q_2...q_T$ for the given parameters and observation sequence.

The first problem can be approached by applying the Baum-Welch algorithm or expectation maximisation (EM) algorithm whereas the second problem can be resolved using the Verterbi algorithm, resulting in the sought after state sequence.

4.2 Implementation

The depmixS4.R package [17] has a readily available EM algorithm that models the observation symbol probability B as a generalised linear response model, which allows for the addition of

the nucleosome bias and the 8nt consensus as covariates. Each position in the genome requires a measure of increased or decreased likelihood for transposition to take place due to the presence or absence of the consensus and nucleosome density.

The likelihood for a transposition to take place based on the consensus was calculated by Chris Illingworth using insertion data from a previous study[8]. A window with 41nt centered on every insertion was used to calculate the percentage of each nucleotide present at each position and compared to the percentage composition across the whole genome. A window of 20 positions was identified for which the composition differed from the genome-wide composition by at least 1% for one nucleotide. For every position t, the probability of observing a specific nucleotide is given by $p_t(\alpha) : 1 \leq t \leq 20$ where α denotes any of the 4 possible nucleotides. The genome wide probability of observing the nucleotide α is given by $p^{gw}(\alpha)$. Next the genome was segmented into 20 nucleotide windows and for each window a likelihood measure was calculated, measuring its similarity to the motif in relation to the genome wide base composition.

$$L = \underbrace{\sum_t \log p_t(\alpha_t)}_{\text{similarity to motif}} - \underbrace{\sum_t \log p^{gw}(\alpha_t)}_{\text{similarity to base comp.}} \tag{4.11}$$

A similar measure was constructed for the nucleosome bias, based on the average nucleosome density over a cell cycles.

The state sequence Q is calculated post model parameter optimisation. The Verterbi algorithm maximises $P(Q|O, \lambda)$ (same as $P(Q, O|\lambda)$), the probability of finding a particular state sequence for the given model parameters λ and observation sequence O. This is done via an iterative process, when moving along the observation sequence, by defining δ, the highest probability along a single path

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P[q_1, q_2, ..., q_t = i, O_1 O_2 ... O_t|\lambda] \tag{4.12}$$

it follows by induction

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] \cdot b_j(O_{t+1}) \tag{4.13}$$

The most likely state sequence is then the sequence of arguments which maximises Equation (4.13), which can be evaluated sequentially from the first position 1 to the final position T, resulting in an assigned state for every nucleotide position.

5 Results and Discussion

The Hidden Markov Model was altered to be applied to each of the three yeast chromosomes separately. The first runs were conducted on the raw read count data. A relation between read count data and cell growth was uncovered and implemented into the HMM with very promising results. A HMM with a fourth state was then constructed, assigning an additional state to very high read counts. A short analysis of the most highly conserved regions of state 1 in relation to their biological meaning was conducted with focus on non-coding RNAs. Beyond analysing the Growth data, a first attempt at analysing the Chronological Life Span Assay was made, a separate data set, where cells with insertions were declined nutrient to measure the effect of insertions on ageing. No further results of the Chronological Life Span Assay are discussed within this report.

5.1 Raw Data

The raw data entails the read count of every insertion and the insertion at every position of the genome. Three states were classified for the hmm, "S_1 = essential", "S_2 = intermediate" and "S_3 = non-essential". In [4] these states were defined by appropriate likelihood functions for read counts, here instead the states are defined by selecting a sub region of the data (about 10%) and associating annotated essential regions and their read counts and insertion density information with state 1. Similarly state 2 is classified by the read counts and insertion density of untranslated regions, whereas state 3 is classified by the read counts and insertion density of all non-annotated sites. Note that these sites are all within the 10% proportion of the data. The classification was chosen based on Figure 3.3, where annotated sites are related to their read count and insertion site content. One should note that this classification helps to associate certain read count numbers, insertion densities and bias effects to states and is not the same as carrying out analysis for an annotated region. Regions with no annotation should overall have the most read counts and insertions[‡] and hence serve as state 3 classifiers. An ergodic, symmetric initial transition matrix was chosen for the HMM:

$$A_{i,j} = \begin{pmatrix} 0.998 & 0.001 & 0.001 \\ 0.001 & 0.998 & 0.001 \\ 0.001 & 0.001 & 0.998 \end{pmatrix} \tag{5.1}$$

Large values on the diagonal assure smoothing of the data, and should result in continuous runs of one state. The initial distribution was chosen as seen below, with an increased likelihood for state 1 at the first position due to a long insertion free sequence at the start.

$$\pi = \{\pi_1 = 0.5, \pi_2 = 0.25, \pi_3 = 0.25\} \tag{5.2}$$

A region of the genome of Chromosome II with the hmm predicted states is plotted in Figure 5.1, a histogram based on the run length (how many nucleotides within one state defined window) for all of Chromosome II is given in Figure 5.2, bin size is based on the Freedman-Diaconis formulae.

[‡]Note: Overall this is save to assume but not true for individual regions since this work in particular is aimed at identifying essential regions previously non-annotated

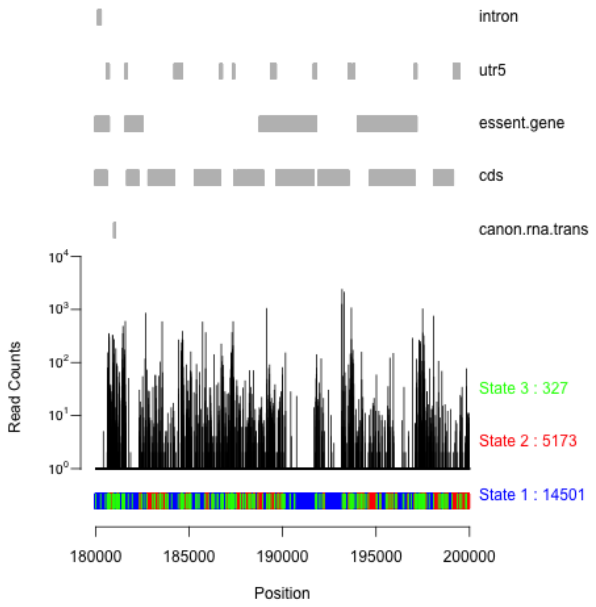


Figure 5.1: HMM classified regions w.r.t. data for chromosome II

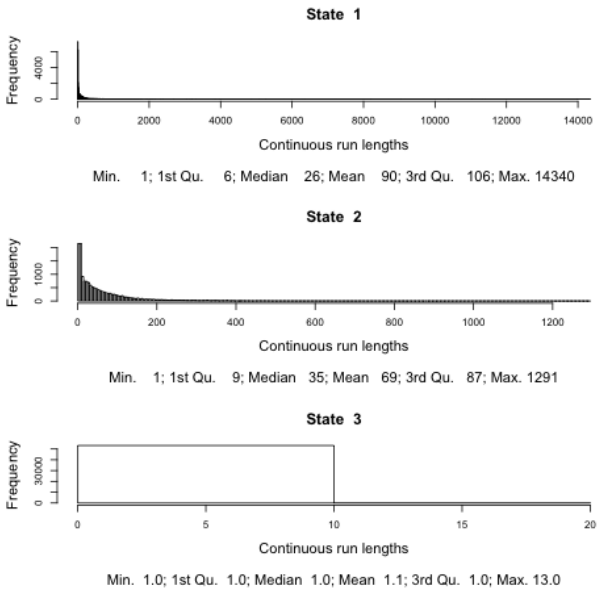


Figure 5.2: Histograms of continuous run length of each state for chromosome II

Figure 5.1 shows quite strong alternation between states with very few continuous runs. State "S_3 = non essential" is almost not present within the selected region[§] and has very discontinuous runs of $mean = 1$ and $max = 13$ as shown and summarised in Figure 5.2. $S_1 = essential$ is primarily represented and has an abundance of 14051 S_1 nucleotides in the selected window and a mean continuous run length of 90 over all of chromosome II. A qualitative analysis of Figure 5.1 shows that as expected S_1 annotated regions fall within very low insertion areas, with a clear S_1 window just after position 190,000, similarly S_2 and S_3 blocks have medium amount of insertions and read counts and high amount of insertions and read counts respectively. An issue that becomes apparent based on the two plots is that the hmm is too sensitive, changing state when there is a high read count or a short sequence of no insertions, resulting in the quick alternation between states and explaining the odd histogram of continuous runs for S_3 in Figure 5.2 as the hmm flips to S_3 when there is an insertion with a high read count and immediately flips back to S_1 when followed by no insertion. This could be an indication that the transition matrix $T_{i,j}$ does not smooth out the data enough, hence resulting in short continuous state runs, but as seen in Equation (5.1) the probability for remaining in the same state is initially set to have a probability of 99% to remain in the current state, a rather high chance, also the transition matrix is optimised by the HMM. The obvious conclusion is that the read counts are too spread out, ranging from 1 to approx. 2M, in particular note the log scale on the y-axis in Figure 5.1 showing a large spread of read counts causing rapid state flips.

5.2 Relation of read count to growth cycle

The available data was further inspected and the read counts for every individual insertion (non-zero read counts) were used to construct a histogram for each chromosome. Due to the large spread of individual values ranging from 1 to approx 2M, each of the selected read counts

[§]The green lines, corresponding to S_3 appear to be more frequent than 327, as indicated on the side of Figure 5.1. This is due to plotting a large range of a 20,000 nucleotide window. The actual presence of S_3 state nucleotides is as indicated.

was scaled by the natural logarithm. As expected small read count values are much more frequent as compared to large values and decay in an exponential manner. Surprisingly certain values at regular intervals appear to be more frequent than their neighbouring values. The process was repeated by scaling read counts with $\log_2$ as instead with the natural logarithm, detecting the same increased frequency at certain values, seen in Figure 5.3 and Figure 5.4.

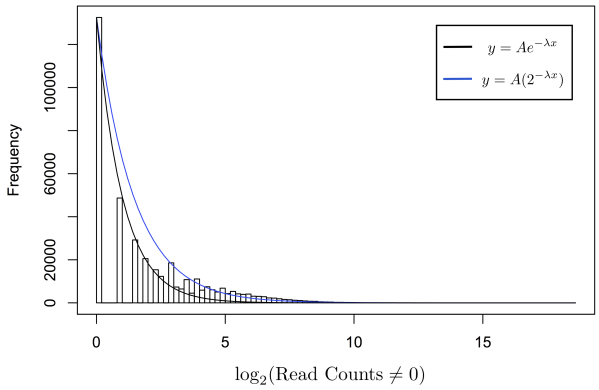


Figure 5.3: Histogram of $\log_2$ scaled non-zero read counts with 100 bins of chromosome I. Two exponential decay functions with base 2 and e were fitted to the data with $A = max(Freq.)$ and $\lambda = 1$.

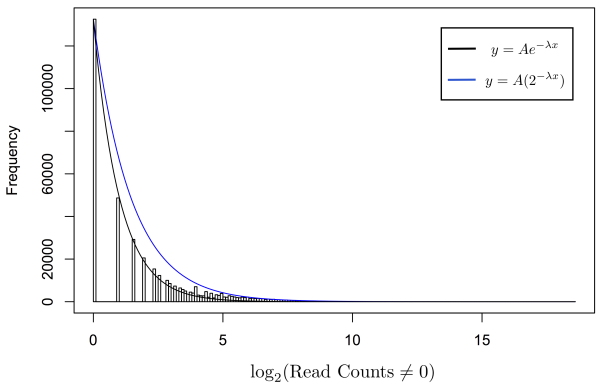


Figure 5.4: Histogram of $\log_2$ scaled non-zero read counts with bins size determined by Freedman-Diaconis of chr. I. Two exponential decay functions with base 2 and e were fitted to the data with $A = max(Freq.)$ and $\lambda = 1$.

As becomes apparent from Figures 5.3 and 5.4, the non zero read counts follow two different types of exponential decay. The same is observed when analysing chromosome II and III. Not as easy visible, on either plot, is the location of spikes, appearing in what appears to be an otherwise uniform decay. Selecting a subrange of Figure 5.3 on the x-axis from 0.5 to 5.5 yields Figure 5.5 and showing increases in frequency at integer values 3, 4 and 5.

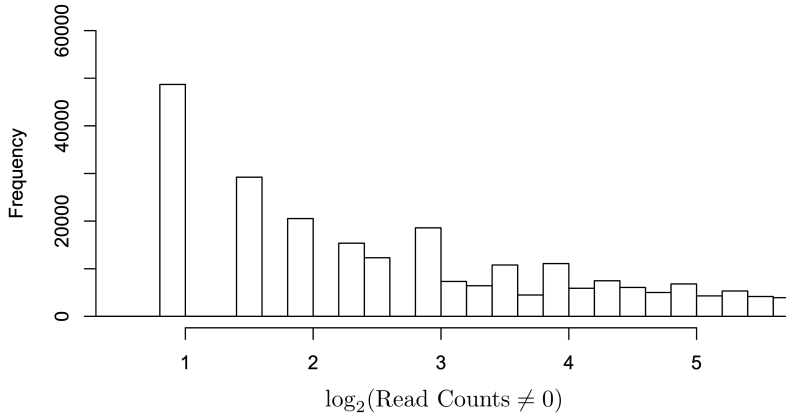


Figure 5.5: Zoomed in histogram of log2 scaled non-zero read counts with 100 bins of chromosome I Figure 5.3

Figure 5.4: Histogram of log2 scaled non-

The integer values are related to read counts around and about $2^3 = 8$, $2^4 = 16$ and $2^5 = 32$,

this can be interpreted as a signal from a geometric series: 1, 2, 4, 8, 16, 32, 64, etc and can be related to mitotic cell growth in yeast

$$x(t) = A \cdot 2^{t/\tau} \tag{5.3}$$

where x = cells at time t

t/τ = growth cycle

τ = time required for single growth cycle

$A = x(0)$ = initial population

Hence the data can be reinterpreted. It was previously assumed that the read count refers to the amount of cells in which hermes transposition had taken place. This seems rather unlikely with very high read counts such as the maximum of 378,184 in chromosome I. A different interpretation is that an insertion took place at a particular site early within the 25 carried out growth cycles. The cell undergoing mitosis (if insertion is not detrimental) duplicates in the next growth cycles, producing offsprings with the exact same transposon at the same position. Hence a high read count is possibly not due to many insertions at the same position within different cells but instead due to an insertion which duplicated many time. A cell growth tree depicted in Figure 5.6 shows how an insertion (red and blue) can take place within a cell line and it's propagation through it's offspring in case of a non majorly interfering insertion. Consultation with L. Greche and D. Jeffares [1] resulted in agreement with the interpretation and it's adaptation for the model.

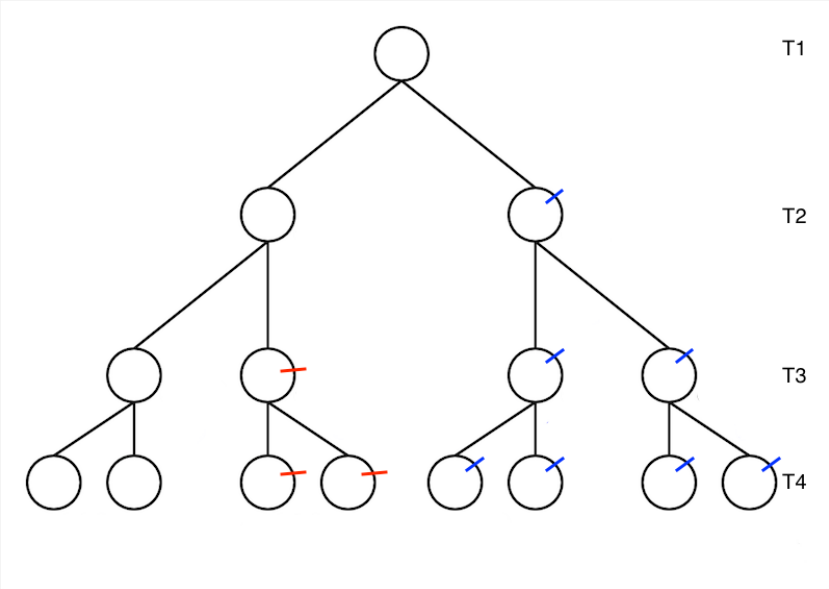


Figure 5.6: Schematic of hermes insertion (red and blue) in a cell line and the resulting offspring carrying copy of transposon at exact same position.

The interpretation of a read count as a growth cycle is then given by:

$$log_2(x) = t/\tau = \text{growth cycle} \tag{5.4}$$

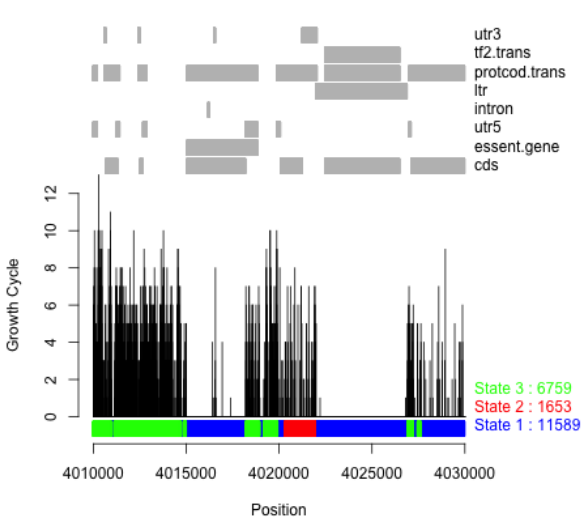
Note that the growth cycle is individual to each insertion since the exact start of transposition cannot be determined for an individual cell. It is also important to see that the growth time

cannot be inferred by "t/τ = growth cycle". τ is the time require for a single cell to duplicate, this measure does not take into account any effect due to insertion such as a decrease in growth speed or increase.

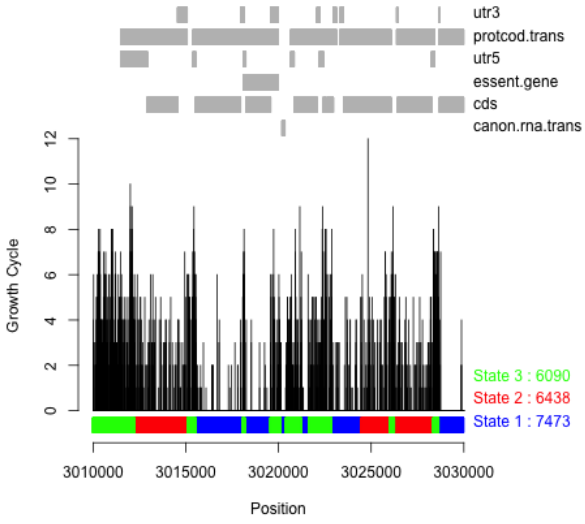
It is possible to associate ultra high read counts with a beneficial effect on growth, but care has to be taken since the maximum read count measure corresponds to 21 growth cycles but cells were grown for 25 growth cycles in total and time of hermes insertion is hard to determine.

5.3 3 state HMM for growth

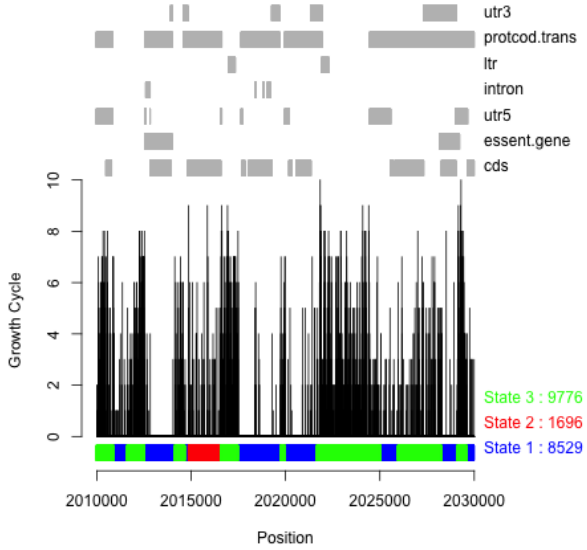
All non-zero insertion read counts are $\log_2$ transformed, resulting in a type of growth cycle measure for each insertion site instead of read counts. This results in values scaling from 0 for a read count of 1, to ≈21 for a read count of ≈2,097,152. Next the full procedure of the hidden markov model implementation is repeated as described in Section 5.1, with the same transition matrix, initial state probability and training set selection. A graphical summary of the HMM on each chromosome with growth cycle measures are given in Figure 5.7. Note how the states have continuous runs of one state and smooth out the data, which is in sharp contrast to the initial HMM run in Section 5.1 where states appear to flip with every insertion. Converting the read counts to growth cycles allows the data to have less spread out values per insertion and more easily comparable regions, resulting in less state flips.



(a) Chromosome I sub region



(b) Chromosome II sub region

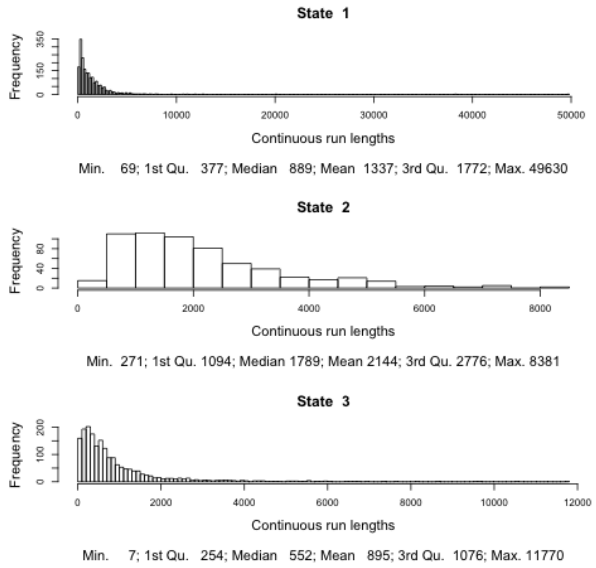


(c) Chromosome III sub region

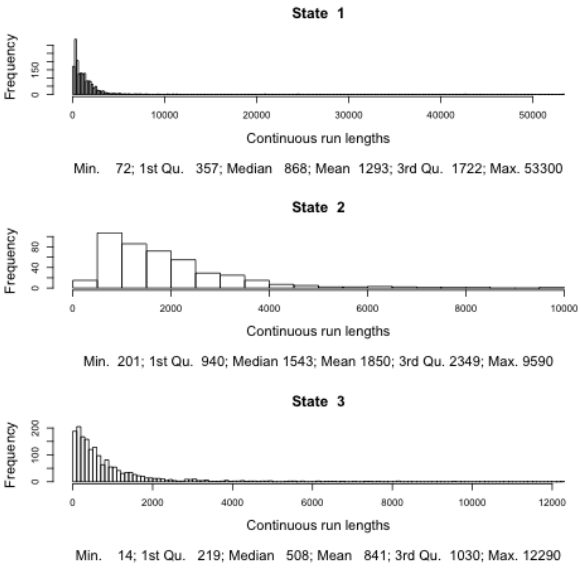
Figure 5.7: Annotations and HMM states in relation to Data: Grey segments on top show annotations from PomBase data base. Annotations stand for: coding sequence (cds), essential genes (essent. gene), introns (introns), untranslated regions (utr3, utr5), Tf-type retrtransposons (tf2.trans), solo LTRs (ltr) and protein coding sequences (protcod. tans). Coloured segments show states of region as predicted by the HMM, state 1 = essential, state 2 = intermediate and state 3 = non-essential.

Note that short continuous runs of one state should still be taken with caution since there is always the possibility that a region is undersampled, which becomes increasingly less likely with increasing length. It is therefore a great feature of the HMM analysis that most states have an average continuous run length (length sequence of nucleotides in one state) much greater than length of nucleotides occupied by a histone ($\approx$ 150bp). The average continuous run length for chromosome I for state 1is 1337, for state 2 is 2144 and for state 3 is 895 as found in Figure 5.8a. This is similar for the other chromosomes in a sense, where state 2 has on average the longest continuous runs but as every histogram in Figure 5.8 shows is also the least frequent. A summery of the continuous run length of states is given below each histogram.

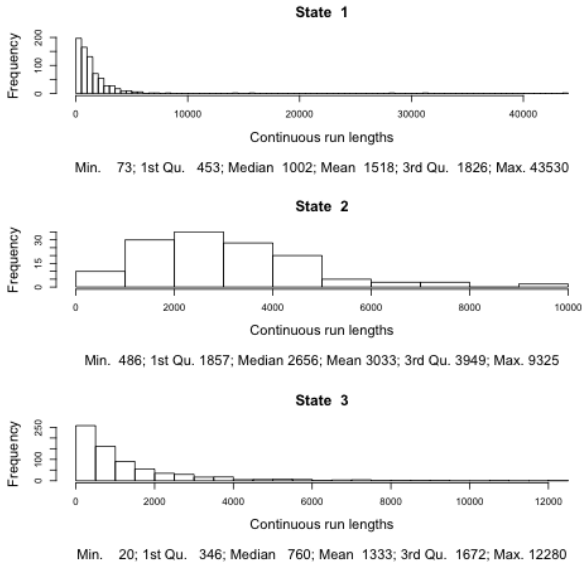
Something peculiar are the maximum run lengths for state 1, which are 49,630, 53,300 and



(a) Chromosome I



(b) Chromosome II



(c) Chromosome III

Figure 5.8: Histograms of continuous run length of each state for every chromosome individually. State 1 = essential, state 2 = intermediate and. state 3 = non-essential.

43,530 for chromosome I, II and III respectively. The regions have been manually inspected and the region for chromosome 3 is plotted in Figure 5.9.

As expected, the region is appropriately classified by the HMM as an essential region, judging on the amount of insertions, but the more astounding finding is that the region is not annotate within the PomBase data base. This indicates that the region has not yet been thoroughly investigated, yet it's function appears to be of great essentiality in growth. A most interesting finding which should be further investigated, experimentally and in other gene libraries. Another possibility is that Hermes insertion could not take place in this region due to an unknown bias, also a very likely.

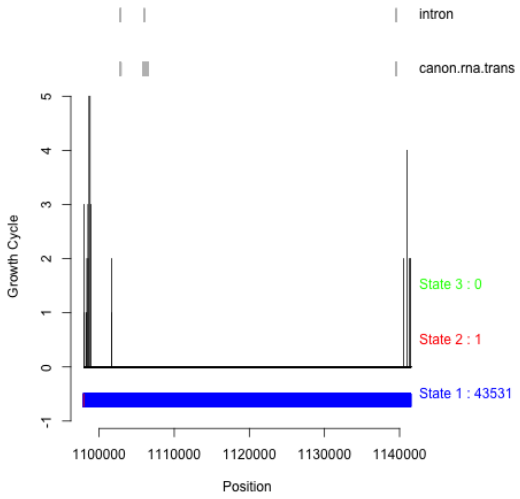
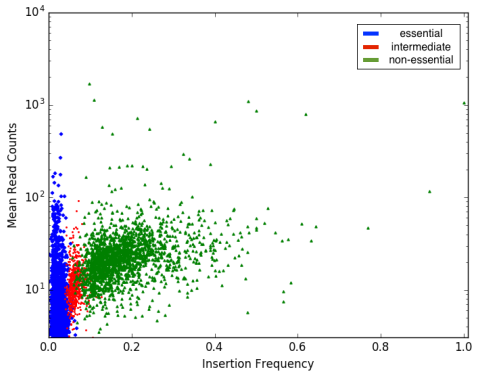
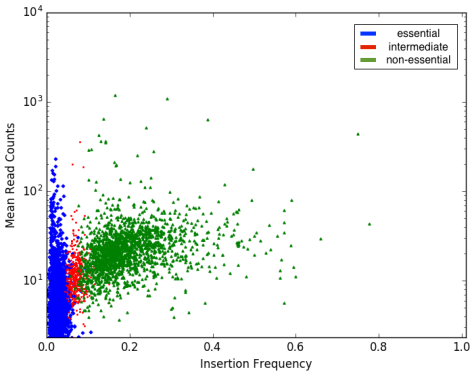


Figure 5.9: Annotations and HMM states in relation to data for the longest continuous run of state 1 in chromosome III.

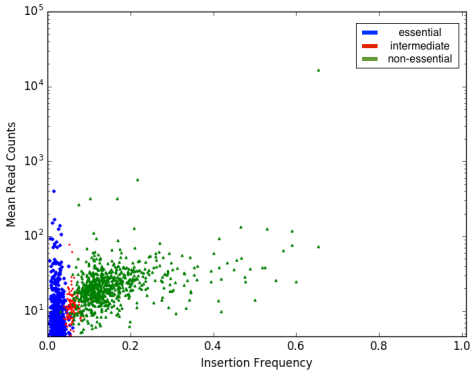
To quantitatively investigate the validity of the model, the growth cycle values were reversed to read counts and the mean read count of insertions was plotted against the insertion frequency or insertions per sight (type of density measure) for each individual window. As expected state 1 grouped for lower values of each scale, with higher values advancing to the intermediate and final non-essential state as depicted in Figure 5.10.



(a) Chromosome I



(b) Chromosome II



(c) Chromosome III

Figure 5.10: Mean read counts of insertions (only non-zero values) vs insertion frequency (insertions/sites in state classified region).

HMM States	Total % of chromosome	Overall insertion frequency	Overall mean read counts
Essential	47	0.020	305.94
Intermediate	23	0.063	1897.85
Non-essential	30	0.156	4087.38

Table 5.1: Chromosome I: Statistics for state classifications. Total % of chromosome stand for the ratio of nucleotide classified as one state to the total amount of nucleotides. Overall insertion frequency is the ratio of all insertion within one state to the total amount of nucleotides within one state across the whole chromosome. Overall mean read counts refers to the total of non zero read counts within one state, divided by the non zero insertions within the same state across the whole chromosome.

An interesting fact to notice in Figure 5.10 is that the essential regions appear to be able to have a high mean read count while maintaining a low insertion frequency. One could therefore possibly infer that a low insertion density is more important for a region to be conserved than a high read count of a particular insertion.

Table 5.1 summarises the findings on a chromosomal level for chromosome I. A surprising detail is that the essential regions make up 47% of the whole chromosome, a rather large portion. The analysis clearly indicates that essential regions are much more conserved since Overall insertion frequency is low and Overall mean read counts are low as compared for intermediate regions and non-essential regions. An even more conservative measure for essential regions, decreasing the overall mean read counts of 305.94 might be an interesting adjustment to make by making a more carful selection of the training data for the essential state (state 1). It should also be argued that the data is possibly undersampled, leading to many classified essential regions and hence more credibility should be given to regions comprising many nucleotides.

5.4 Adjusted 4 state HMM for growth

Another HMM was constructed with 4 states instead of 3 to account for really high growth cycle values. As previously mentioned these values might correspond to growth enhancing insertions, having a positive effect on fitness of the cell such as for example a higher metabolic rate and therefore a faster reproductive rate. The initial probability was changed to:

$$\pi = \{\pi_1 = 0.5, \pi_2 = 0.2, \pi_3 = 0.2, \pi_4 = 0.1\} \tag{5.5}$$

The training set for the 4th state comprised growth cycles above the 99th percentile (corresponding to ultra high read counts) and was added to the other already existing 3 training sets. The initial transition matrix was adapted as seen in Equation (5.6).

$$A_{i,j} = \begin{pmatrix} 0.998 & 0.0006 & 0.0006 & 0.0006 \\ 0.0006 & 0.998 & 0.0006 & 0.0006 \\ 0.0006 & 0.0006 & 0.998 & 0.0006 \\ 0.0006 & 0.0006 & 0.0006 & 0.998 \end{pmatrix} \tag{5.6}$$

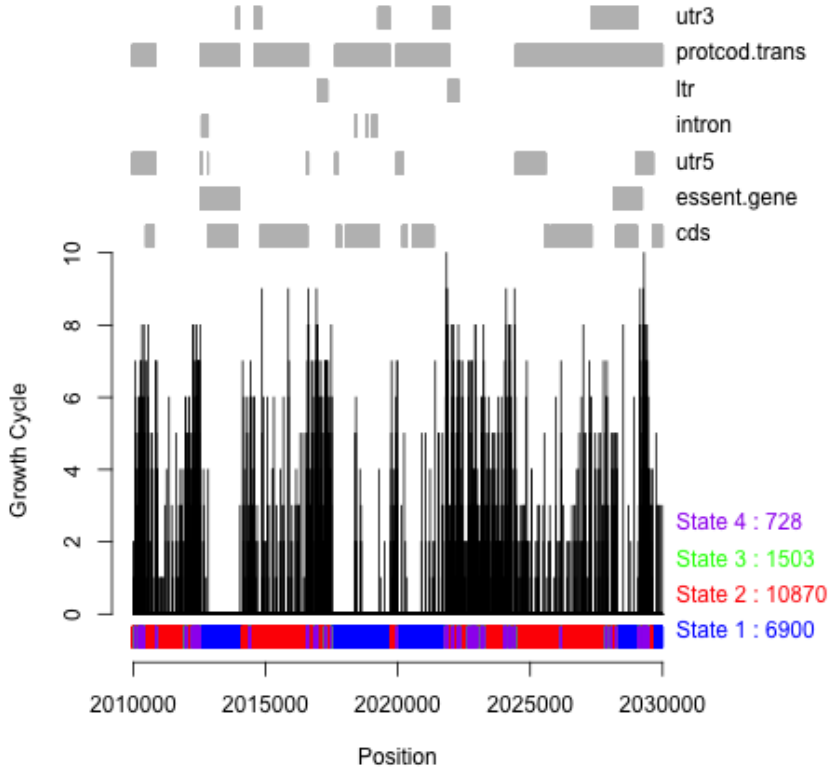


Figure 5.11: Annotations and HMM states in relation to Data, using the 4 state HMM.

Here only a position annotation plot (Figure 5.11) with corresponding states is given for chromosome III, further analysis has to be performed. Note how the 4 state HMM accurately identifies essential regions. The purple state 4 lines in Figure 5.11, not to be mixed up with the blue essential (state I) lines are located at high growth cycle values. A rather strange effect of HMM 4 is the high presence of state 2 = intermediate and should be further investigated. Care must be taken in the interpretation of the 4th state as it is not fully assured that high read counts have an enhancing effect but might instead have no effect on the cell at all.

5.5 Biological interpretation

The role of non coding RNA's is not very well understood so far, leaving several open questions about their importance in cellular processes. This study for the first time allows to associate previously annotated ncRNA's with growth and fitness of a cell and their importance to such. State 1 classified regions with a very low insertion frequency, mean read count (of insertion sites) and a long continuous run length (comprising many nucleotides) have the highest likelihood to be highly conserved and be of great importance to fitness and growth of a cell. Therefore for chromosome I, all state I regions where selected with a mean read count less than 2.6, an insertion frequency less than 0.015 and a continuous run length greater than 1772nt. This corresponds to the lower 25 percentile for mean read counts and insertion frequency and the

upper 75 percentile for continuous run length respectively. A total of 21 state 1 regions satisfied these conditions, with a full list given in the appendix. Each of these regions entails several non-coding RNAs, with a minimum of 66 in one region and a maximum of 609 in another as found on BioMart [9] and depicted in Figure 5.12. Another bar chart was made with all number of protein coding RNAs in the same 21 regions plotted in Figure 5.13. An interesting feature of both charts is that the total of protein coding RNAs and ncRNAs appears to decent the further one advances in the genome of chromosome I. As expected the protein coding RNAs are much more frequent as compared to the ncRNAs.

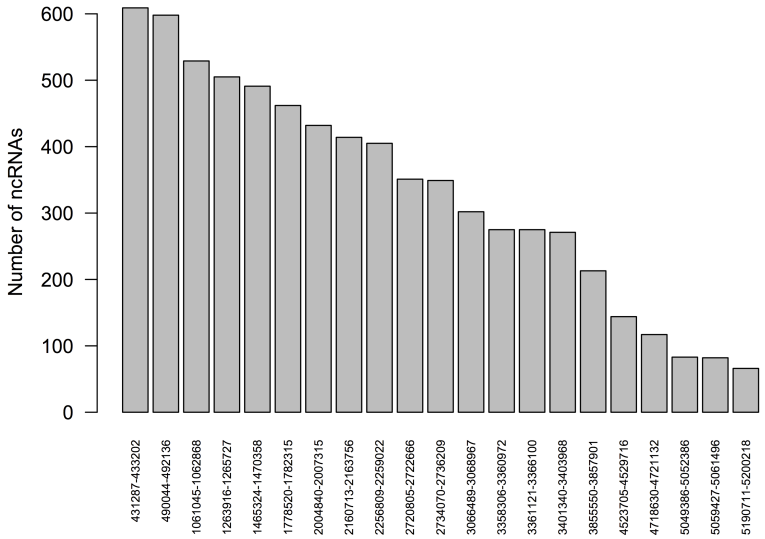


Figure 5.12: Bar chart of non coding RNAs in highly conserved state 1 regions. X-axis displays start and end of region in chromosome I, y-axis displays RNA count. Analysis done through BioMart [9].

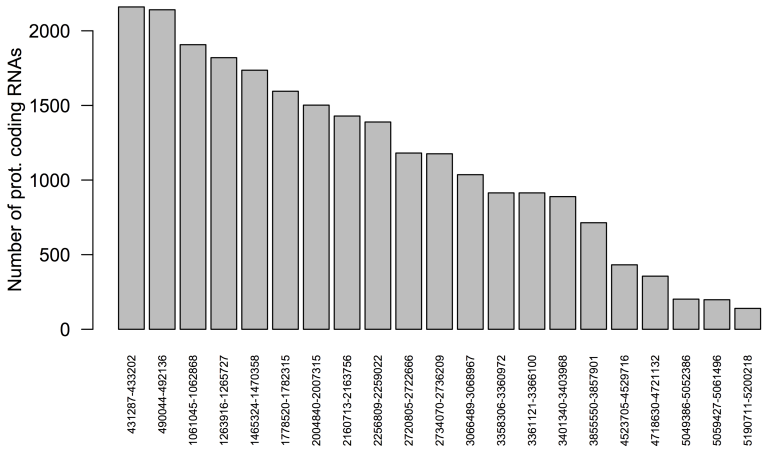


Figure 5.13: Bar chart of protein coding RNAs in highly conserved state 1 regions. X-axis displays start and end of region in chromosome I, y-axis displays RNA count. Analysis done through BioMart [9].

6 Conclusion

The developed 3 state HMM was run on the raw data, entailing insertion sites and their corresponding read counts. The classified state regions by the HMM alternated rapidly, most likely due to the huge difference in high an low read counts. Relating read counts to growth cycles by scaling with log_2 allowed for a much more uniform data set, with non large peak heights. The resulting HMM state 3 run on growth data, produced great results with very promising classifications. Especially a very large essential region, almost full void of insertions, with a length of 43530nt in the centre of the 3rd chromosome should be investigated in more depth in future experiments. The constructed 4 state HMM shows promising results in identifying essential regions but the accumulation of state 2 intermediate read counts should be closely watched and the HMM adjusted accordingly. Overall the HMM has shown very promising results for all 3 chromosomes, with data available for further analysis and interpretation.

References

[1] Jurg Bahler. Bahlerlab: Genome regulation. http://www.bahlerlab.info/home/.

[2] David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.

[3] Guillaume Carpentier, Jérome Jaillet, Aude Pflieger, Jérémy Adet, Sylvaine Renault, and Corinne Augé-Gouillou. Transposase–transposase interactions in mos1 complexes: a biochemical approach. *Journal of molecular biology*, 405(4):892–908, 2011.

[4] Michael A DeJesus and Thomas R Ioerger. A hidden markov model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data. *BMC bioinformatics*, 14(1):303, 2013.

[5] Adam G Evertts, Christopher Plymire, Nancy L Craig, and Henry L Levin. The hermes transposon of musca domestica is an efficient tool for the mutagenesis of schizosaccharomyces pombe. *Genetics*, 177(4):2519–2523, 2007.

[6] Sunil Gangadharan, Loris Mularoni, Jennifer Fain-Thornton, Sarah J Wheelan, and Nancy L Craig. Dna transposon hermes inserts into dna in nucleosome-free regions in vivo. *Proceedings of the National Academy of Sciences*, 107(51):21966–21972, 2010.

[7] Jennifer E Griffin, Jeffrey D Gawronski, Michael A DeJesus, Thomas R Ioerger, Brian J Akerley, and Christopher M Sassetti. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathog*, 7(9):e1002251, 2011.

[8] Yabin Guo, Jung Min Park, Bowen Cui, Elizabeth Humes, Sunil Gangadharan, Stevephen Hung, Peter C FitzGerald, Kwang-Lae Hoe, Shiv IS Grewal, Nancy L Craig, et al. Integration profiling of gene function with dense maps of transposon integration. *Genetics*, 195(2):599–609, 2013.

[9] Syed Haider, Benoit Ballester, Damian Smedley, Junjun Zhang, Peter Rice, and Arek Kasprzyk. Biomart central portal?unified access to biological data. *Nucleic acids research*, 37(suppl 2):W23–W27, 2009.

[10] Alison B Hickman, Hosam E Ewis, Xianghong Li, Joshua A Knapp, Thomas Laver, Anna-Louise Doss, Gökhan Tolun, Alasdair C Steven, Alexander Grishaev, Ad Bax, et al. Structural basis of hat transposon end recognition by hermes, an octameric dna transposase from musca domestica. *Cell*, 158(2):353–367, 2014.

[11] Timothy R Hughes, Christopher J Roberts, Hongyue Dai, Allan R Jones, Michael R Meyer, David Slade, Julja Burchard, Sally Dow, Teresa R Ward, Matthew J Kidd, et al. Widespread aneuploidy revealed by dna microarray expression profiling. *Nature genetics*, 25(3):333–337, 2000.

[12] Daniel Jeffares. Bahlerlab: Genome regulation. http://www.bahlerlab.info/home/.

[13] Young Zoon Kim. Altered histone modifications in gliomas. *Brain Tumor Research and Treatment*, 2(1):7–21, 2014.

[14] Joshua Allen Knapp. Dissecting the hermes transposase: Residues important for target dna binding and phosphorylation. 2011.

[15] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[16] Ignacio Soriano, Luis Quintales, and Francisco Antequera. Clustered regulatory elements at nucleosome-depleted regions punctuate a constant nucleosomal landscape in schizosaccharomyces pombe. *BMC genomics*, 14(1):1, 2013.

[17] Ingmar Visser and Maarten Speekenbrink. depmixS4: An R package for hidden markov models. *Journal of Statistical Software*, 36(7):1–21, 2010.

[18] McDowall MD Rutherford K Vaughan BW Staines DM Aslett M Lock A Bhler J Kersey PJ Oliver SG" "Wood V, Harris MA. Pombase: a comprehensive online resource for fission yeast. http://www.pombase.org/browse-curation/fission-yeast-go-slim-terms.

A HMM 3

HMM States	Total % of chromosome	Overall insertion frequency	Overall mean read counts
Essential	51	0.022	314.12
Intermediate	17	0.071	1852.73
Non-essential	32	0.176	5583.15

Table A.1: Chromosome II: Statistics for state classifications. Total % of chromosome stand for the ratio of nucleotide classified as one state to the total amount of nucleotides. Overall insertion frequency is the ratio of all insertion within one state to the total amount of nucleotides within one state across the whole chromosome. Overall mean read counts refers to the total of non zero read counts within one state, divided by the non zero insertions within the same state across the whole chromosome.

B Highly Conserved Regions in Chromosome I

label start end length
114 431287 433202 1916
134 490044 492136 2093
348 1061045 1062868 1824
419 1263916 1265727 1812
492 1465324 1470358 5035
599 1778520 1782315 3796
677 2004840 2007315 2476
748 2160713 2163756 3044
783 2256809 2259022 2214
961 2720805 2722666 1862
967 2734070 2736209 2140
1088 3066489 3068967 2479
1195 3358306 3360972 2667
1196 3361121 3366100 4980
1210 3401340 3403968 2629
1364 3855550 3857901 2352
1610 4523705 4529716 6012
1677 4718630 4721132 2503
1788 5049386 5052386 3001
1793 5059427 5061496 2070
1848 5190711 5200218 9508

HMM States	Total % of chromosome	Overall insertion frequency	Overall mean read counts
Essential	45	0.019	364.08
Intermediate	17	0.058	2289.88
Non-essential	38	0.126	6820.37

Table A.2: Chromosome III: Statistics for state classifications. Total % of chromosome stand for the ratio of nucleotide classified as one state to the total amount of nucleotides. Overall insertion frequency is the ratio of all insertion within one state to the total amount of nucleotides within one state across the whole chromosome. Overall mean read counts refers to the total of non zero read counts within one state, divided by the non zero insertions within the same state across the whole chromosome.