Sequencing of Insertion Sites

Genomic DNA was isolated from the final YES cultures containing 5-FOA and G418, and samples were prepared for Illumina sequencing as described previously (1). In brief, the genomic DNA was purified from 200 O.D. units of cells by zymolyase 100T (Sigma) treatment and spheroblast extraction. The restriction enzyme Msel was used to fragment the DNA because previous data indicated this enzyme did not introduce a bias in detection of insertion sites and because, in our lab, restriction enzyme-cleaved ends are ligated to linkers more efficiently than sonicated DNA fragments. For each library, six-2 µg samples of genomic DNA were digested in 100 µl volumes with Msel for 16 hours and were then purified using the Qiagen PCR purification kit. The digested DNA for each library was eluted in a volume of 50µl and used in 10 duplicate linker ligations with Invitrogen T4 DNA ligase for 1 h at 25°C (See Suppl. Fig. S1A for linker oligos). After heat inactivation at 65°C for 10 min, 10 units of SpeI was added to separate the 5′ LTR from the 3′ LTR which is used in the amplification of the insertion sites. All the SpeI cut DNA was used directly as template in 80 PCR reactions, 20 µl per well, with Titanium Taq (Clontech). The primer that recognizes the linker end is HL2216 and the LTR amplification primers with barcodes are described in Suppl. Fig. S1B. The PCR program used was:

1. 94°C 4 min
2. 94°C 15 s
3. 65°C 30 s
4. 72°C 45 s
5. go to step 2 for a total of six cycles.
6. 94°C 15 s
7. 60°C 30 s
8. 72°C 45 s
9. go to step 6 for a total of 24 cycles.
10. 68°C 10 min
11. 4°C until sample is retrieved.

All PCR reactions were pooled and then divided into 6 samples, each of which were purified on a separate Qiagen PCR purification column. Each set of 80 PCR reactions was purified on a single 10 cm 2% TBE agarose gel, which was run at 70 volts until the dye reached half the length of the gel. The DNA of size 150–500 bp was extracted from the gel and purified with Qiagen gel extraction kits. The concentration of the purified DNA was determined using qPCR (KAPA SYBR FAST kit, Kapa Biosystems) and a fluorimeter using picogreen. All six libraries were combined and loaded onto one lane of an Illumina Genome Analyzer IIx (GAIIx) and primer HL2747 was used to sequence 100 nt single end reads. The sequencing was performed by the University of California, Irvine Genomics High-throughput Facility. The sequence reads were submitted to the Short Read Archive (SRA) at National Center for Biotechnology Information (NCBI) under the accession number PRJNA279274.

The computational methods of data analysis and the use of Rate Distortion Theory to remove erroneous serial numbers generated by Illumina misreads were performed as previously described (1) and are explained below. However, in this set of data two genome nucleotide positions had over 14,000 independent serial number coded insertions. This high number of serial numbers could not be analyzed with the rate
distortion software to correct Illumina misreads and were left uncorrected in our integration analyses. The two positions were Chromosome 1: 98781 (1.4% of insertions) and Chromosome 2: 4414490 (1.2% of insertions). The corrected data sets are Supplemental Data files #1 through #9.

**Bioinformatic Analysis**

The CDS coordinates for the *S. pombe* genome were from the Feb. 2007 version of the chromosome contigs from the Wellcome Trust Sanger Institute (ftp://ftp.sanger.ac.uk/pub/yeast/pombe/Chromosome_contigs/OLD/20070206/). The coordinates within intergenic regions or ORFs and the distance to the start or end of the nearest ORF of each integration site were calculated with scripts written with PERL or Python.

To generate a genomic Sap1 binding profile, paired-end alignments of previously published CHIP-seq DNA sequence reads from cross-linked Sap1-antibody and whole cell extract (2) aligned to the 2007 build of the *S. pombe* genome were generated using the bioinformatics tool Burrow’s Wheeler Alignment (BWA) (3). The resulting BAM files from each data set were converted into WIG format using MACS (version 14), and the Log2 ratio of Sap1 signal to the WCE control was calculated using Python scripts. The Sap1 binding data output was aligned with previously published Tf1 serial number integration positions (1) using custom PERL scripts (Suppl. Table S1).

For the analysis of the Sap1 binding and Tf1 integration site preferences near transcription start sites (TSS) all intergenic sequences were ranked by the number of integration events and grouped into 10 bins of 500 sequences (bin 11 was disregarded...
since it had only 61 intergenic regions and 11 TSSs). All the reported TSSs (Wellcome Trust Sanger Inst. 2011-02-04 genome build) within each bin were aligned and the number of independent Tf1 insertion events, the amount of Sap1 binding enrichment, and the size of the nucleosome free regions were normalized to the number of TSSs in each bin of intergenic sequences (nucleosome data available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM715390) (4). The integration values used for the above analysis were normalized by dividing the number of integration events at each position by the total number of independent integration events in each dataset. Pairwise correlations of Sap1 binding, NFR size, and Tf1 integration numbers were attained by summing modified values of each data set within 1000bp upstream of the TSS in each bin. For these analyses the following data sets were modified as follows: Sap1 data values were back-transformed from their Log 2 values into their non-logarithmic ratio values, and only negative nucleosome occupancy data points were summed to reflect areas of nucleosome free regions (NRFs). Linear regression analysis was performed with Graphpad Prism to derive correlation coefficients.

The Sap1 binding motif was identified and its Logo was constructed from the analysis of previously published CHIP-Seq data (2) using MEME suite. Briefly, paired-end alignments of the CHIP-seq data to the 2007 build of the S. pombe were generated as described above, and the resulting BAM files were further analyzed using the HOMER suite to identify peak sequences of Sap1 binding. The resulting output was converted into FASTA format using a custom made PERL script which was then applied to MEME (5). To find occurrences of the identified motif in the S. pombe genome, the resulting motif was applied using the FIMO tool of MEME suite (6). The location of Tf1 insertion positions relative to these genomic Sap1 binding motifs was assessed using
custom made PERL Scripts that plotted data derived from our previous published Tf1 serial number library (1) relative to the positions of the motifs.
Compensation for Sequence Errors Using Rate Distortion Theory

Introduction

Illumina sequencing produces sequence errors in our data of approximately 0.5% per base pair. While these mis-reads can be readily identified when they occur in genomic sequence, they cannot be directly corrected when they occur in the eight base pair randomized serial number sequence. Because mis-reads in the serial number sequences would artificially increase the measures of independent insertion, we developed a method to correct for this distortion. The correction analyzes all sequences that map to a single genomic position and considers the number of duplicate sequence reads with the same serial number sequence. Since the error rate is the same for each sequence read we can estimate the probability that any serial number sequence results from errors derived from sequencing high numbers of duplicate reads. In other words, our method is based on the number of duplicate sequence reads and the sequence divergence of the eight serial number base pairs. As an example, using a sequence error rate of 0.5% per bp (4% per eight bp serial tag), a set of 1,000 duplicate serial number sequences mapping to a single genomic site would be expected to generate 40 erroneous serial number sequences that differed from the original sequence by a single nucleotide. With this information about the sequence distribution we find the probability is high that these 40 single sequences resulted from mis-reads. If a high number of duplicate serial sequences, say 300, differed from the serial sequence of the 1,000 duplicates by two or three nucleotides the probability that the 300 reads resulted from mis-reads of the 1,000 duplicates is very low.

We view the problem of compensating for sequence errors of serial numbers resulting from the Illumina sequencing process as a form of data compression, and we use rate distortion theory to guide this compression. Claude Shannon introduced rate distortion theory in his seminal 1948 paper on information theory [Shannon 1948]. The idea is to balance the amount that the information is compressed against the distortion of the information generated by the compression. The algorithm described here to implement the rate distortion method is similar to the one discussed in the Supplement Data of a previous publication [Chatterjee et al. 2014], but with some differences in how it is implemented. The first section of this supplemental discussion gives a brief overview of rate distortion theory for those unfamiliar with it, and the other sections discuss how to adapt the method to compensate for serial number mis-reads.

Mutual Information and Expected Distortion in Data Compression

Consider a set of data \( \{X_i, 1 < i < N_X\} \) that are mapped stochastically into another set of the data, \( \{Y_j, 1 < j < N_Y\} \), where \( N_Y \) are \( N_X \) are not necessarily equal. The mapping is specified by \( P(Y_j|X_i) \), the probability that value \( X_i \) implies value \( Y_j \). For a well-posed problem, every \( X_i \) must be mapped into something, even if it is just to itself, so the following constraint is enforced:
$$\sum_j P(Y_j|X_i) = 1$$  \hspace{1cm} \text{(Eq. 1)}$$

Let Pr(X_i) be the Bayesian prior probability for data point X_i; this is usually set to be the probability distribution of the X data points itself. Based on the conditional entropy developed by Shannon (7), the mutual information $M(Y;X)$ of the mapping of X to Y is defined as

$$M(Y;X) = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} P(Y_j|X_i) \Pr(X_i) \log_2(P(Y_j|X_i) \Pr(Y_j)^{-1})$$

where

$$\Pr(Y_j) = \sum_{i=1}^{N_X} P(Y_j|X_i) \Pr(X_i)$$  \hspace{1cm} \text{(Eq. 2)}$$

($\Pr(Y_j)$ can be thought as the probability that at least one X will be mapped to $Y_j$).

To understand what this quantity means, consider two extreme examples: (1) if the mapping from X to Y were deterministic and 1-to-1 so that there is no loss of data in the mapping, and $\Pr(X_i) = N_X^{-1}$, then $M$ reduces to $\log_2(N_X)$ bits, the negative of the entropy of the distribution of X_i’s. This value of $M$ indicates that all information contained by the X distribution is preserved. (2) On the other hand, if $P(Y_j|X_i)$ has the same value for every i and j, then $M$ is 0, reflecting complete data loss by the mapping.

For data compression one would like a clever choice of the probabilities $P(Y_j|X_i)$ that minimize mutual information $M(Y;X)$, yet preserve relevant features in the data. Let $d(X_i,Y_j)$ be a measure of how different data points $X_i$ and $Y_j$ are from each other, the distortion. The expected distortion in the mapping, $D(Y;X)$, is

$$D(Y;X) = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} d(X_i,Y_j) P(Y_j|X_i) \Pr(X_i)$$  \hspace{1cm} \text{(Eq. 3)}$$

The rate distortion procedure is to chose probabilities $P(Y_j|X_i)$ that minimize the functional

$$F = D(Y;X) + T M(Y;X)$$  \hspace{1cm} \text{(Eq. 4)}$$

The parameter $T$ is called the information temperature, chosen to balance $D$ against $M$: the larger $T$ is, the more information that is lost. The analogy with statistical mechanics is apparent: $d$ defines an energy landscape of gas particle interaction that yields internal energy $D$, $-M$ is the entropy, and $F$ a free energy.

The $P(Y_j|X_i)$ that minimize $F$ are the solutions to the implicit equations [Rose et al. 1990]

$$P(Y_j|X_i) = \Pr(Y_j) Z(X_i,T)^{-1} \exp(-\log(2) d(X_i,Y_j)/T)$$

A. Hickey et al.
where

\[ Z(X_i, T) = \sum_{j=1}^{N_Y} \text{Pc}(Y_j) \exp(-\log(2) \frac{d(X_i, Y_j)}{T}) \]  

(Eq. 5)

Since the Pc(Y_j) are dependent on the P(Y_j|X_i) variables, this solution for the P(Y_j|X_i) variables must be solved by iteration. Now T is not fixed intrinsically by Eq 4: it must be chosen so that features considered relevant are preserved.

**Serial Number Mis-read Problem as Data Compression**

Consider a set of serial number sequences read at a chromosome position, \( S = \{S_1, S_2, \ldots, S_N\} \). One would like to estimate the probability \( P(S_j|S_i) \) that serial number \( S_j \) is really a mis-read of serial number \( S_i \) for all pairing \((i,j)\) of serial numbers read at the position. Since the the \( P(S_j|S_i) \)'s are a stochastic mapping of \( S \) back onto itself, the rate distortion algorithm can be applied to find the \( P(S_j|S_i) \) that minimizes the mutual information with some suitable choice of distortion function \( d \) and information temperature \( T \). Since the mis-read information in the fixed 8-base LTR sequence was carefully recorded, we used that information to determine \( d \) and \( T \). When applying the rate distortion calculation at a given position on a chromosome, we took the prior probability for serial number \( SN_i \) as

\[ \text{Pr}(SN_i) = \frac{\text{dup}_i}{N_{\text{dup}}} \]  

(Eq. 6)

where \( \text{dup}_i \) is the number of duplicate reads for \( SN_i \), and, and \( N_{\text{dup}} \) is the sum of the number of duplicates over all the independent serial number reads at the position.

Overall, the fraction of the eight base pair serial sequences that were misread was \(< 5\%\) for the sap1-1(25°C) and sap1+(25°C) datasets and dataset sap1+(32°C) #1, but \(~10\%\) for the sap1+(32°C) #2 and sap1+(32°C) #3 datasets. (See Suppl. Fig. S5A.) The rate of misreading \( m \) of the eight bases, \( r(m) \), tends to decrease with increasing \( m \), (although for the datasets sap1+(32°C) #2 and #3, the decline was not monotonic). (See Suppl. Fig. S5B.) Given two serial numbers \( SN_1 \) and \( SN_2 \) that differ by \( m \) bases, we estimate the probability \( Pmr(m) \) that a serial number \( SN_1 \) will be mis-read as serial number \( SN_2 \) to be

\[ Pmr(m) = \frac{r(m)}{8m!} \cdot 3^{-m} \]  

(Eq. 7)

This expression assumes that (1) the serial number mis-read rate is the same as for the LTR sequence, and (2) the probability of mis-reading is independent of position and base pair of the mis-read. For the sap1-1(25°C) and sap1+(25°C) datasets \( Pmr(m) \) was calculated using the average of \( r(m) \) over the six datasets, since their \( r(m) \) functions are so similar. (See Suppl. Fig. S5B.) For sap1+(32°C) #2 and sap1+(32°C) #3, \( Pmr(m) \) was calculated from the average of \( r(m) \) over those two datasets since they are essentially identical for \( m < 6 \). Because the \( r(m) \) function for sap1+(32°C) #1 is some what smaller for \( m > 3 \) than for the other datasets, \( Pmr(m) \) for sap1+(32°C) #1 data was calculated...
directly from its $r(m)$ function. For a given dataset, if two serial numbers $SN_1$ and $SN_2$ have a base difference of $m$, we defined the distortion between them to be

$$d(SN_1, SN_2) = -\ln(P_{mr}(m))$$  \hspace{1cm} (Eq. 8)

(If $P_{mr}(m) = 0$, we take $d(SN_1, SN_2)$ as $\ln(10^{22})$, which is $\sim 50.7$.) The distortion as a function of base differences for the datasets is plotted in Suppl. Fig. S6.

In our previous report [Chatterjee et al. 2014] we took $d(SN_1, SN_2)$ to simply be the number of base differences between them. As can be seen in Suppl. Fig. S6, the choice defined by Eq. 8 grows less than linearly in $m$, so contributions to $F$ from serial numbers pairs that have a greater number of base differences will more important than for the distortion used in the earlier report. Furthermore, the choice specified in Eq. 8 allows the $P_{mr}(m)$ function to enter into the rate distortion calculation in a very intuitive manner: at a given information temperature $T$, then ratio of the probability that serial number $SN_1$ is a mis-read of $SN_2$ to the probability that it is a mis-read of $SN_3$ is

$$P(SN_2|SN_1)/P(SN_3|SN_1) = P_{c}(SN_2)*P_{c}(SN_3)^{-1}*(P_{mr}(m_{12})/P_{mr}(m_{13}))^{ln2/T}$$  \hspace{1cm} (Eq. 9)

Here $m_{1x}$ is the number of base differences between $SN_1$ and $SN_x$.

To use the $P_{mr}(m)$ function to fix the information temperature, we proceed as follows:

1) Consider a set of serial numbers \{$S_1, S_2, \ldots, S_N$\} sequenced at a chromosome position, and let \{\text{dup}_1, \text{dup}_2, \ldots, \text{dup}_N\} be the corresponding number of duplicate reads observed for each sequence. A rate distortion calculation is done at information temperature $T$ that gives the probability mappings \{P(S_j|S_i)\} for that $T$.

2) Define a new set of duplicate assignments, \{\text{dup}_1’, \text{dup}_2’, \ldots, \text{dup}_N’\}, such that $\text{dup}_i’ = \text{sum over all } \text{dup}_j \text{ such that } P(S_x|S_i) \text{ is at a maximum for } x = i$. Call this the maximal assignment for $T$ of the duplicate reads.

3) If the maximal assignment were the true distribution of duplicate reads among the serial numbers at the chromosome position, then the expected distribution of duplicates from the sequencing process would be

$$<\text{dup}_i> = \sum_i P_{mr}(m_{ij}) \text{dup}_j’$$  \hspace{1cm} (Eq. 10)

Here, the sum is over all serial numbers read at the position, and $m_{ij}$ is number of base differences between $SN_i$ and $SN_j$.

4) The “fit” of the maximal assignment distribution to the actual observed distribution is assessed using a $G$-statistic:

$$G = \sum_i \text{dup}_i \ln(\text{dup}_i/ <\text{dup}_i>)$$  \hspace{1cm} (Eq. 11)
The strategy of the rate distortion calculation is to find the value of $T$ that has the maximal distribution of duplicates that gives a value of $G$ closest to $3N$. Usually, two distributions are considered statistically undistinguishable with the $G$ test if $G < \text{twice the degrees of freedom}$, which here would be $2(N-1)$. [Sokal and Rohlf 1995] The target of $3N$ was chosen to allow for the fact that thousands of positions were examined within each dataset, (see Suppl. Fig. S7A), so the criterion for significance was set higher. Serial numbers with zero duplicates in the maximal distribution with $G$ closest to $3N$ are judged as likely mis-reads by the sequencing process.

There are other criteria that could be used to fixed $T$ and assess which serial numbers are likely mis-reads. In our previous report [Chatterjee et al. 2014], $T$ was adjusted so that the number of mis-read duplicates matched $Pmr(0)\times N_{\text{dup}}$, the expected number of mis-read duplicates. The criterion used in the current analysis allows for more use of the details in the $Pmr(m)$ function. Spot checks of the results suggest that the “$G$ closest to $3N$” rule allowed for about twice the number of mis-reads predicted by $Pmr(0)\times N_{\text{dup}}$. Whatever criterion one uses, the point is to be conservative in predicting what are the true reads.

The number of genome positions in each of the nine integration data sets versus the number of raw serial number sequence reads at a position (the quantity $N$ above) is shown in Suppl. Fig. S7A. The average number of duplicates per serial number versus $N$ is shown in Suppl. Fig. S7B. As mentioned in the main text, positions with $N > 2048$ were not evaluated. There were two reasons for this: (1) the amount of time and memory needed to do the calculation in the current implementation grows with $N^2$, and (2) as indicated in Suppl. Fig. S7B, the average number of duplicates per serial number declines sharply for $N > 1024$. This means that positions with $N > 1024$ tend to be “low information” positions for which most of the serial numbers are likely to be true reads. Preliminary calculations at positions with $N > 2048$ indicated that the time to come to an answer for those positions would be many hours.

**Details of Implementation of the Rate Distortion Algorithm**

For an integration dataset, the algorithm was implemented as follows:

1) If there was only one serial number indicated at a position, or if there was more than one serial number, but the number of duplicate reads for each serial number is the same, then no rate distortion calculation was performed. The prediction number of true reads for the position is set to the reported number of reads. The computation advances next position on the chromosome. If the number of serial numbers is two or more, and not all serial numbers have the same number of duplicates, proceed to step 2.

2) Start with initial $T = 1$.

2a) If $G$ for the maximal assignment is $> 3N$, reduce the $T$ by a factor of $e$. Continue until $G < 3N$. If temperate collapses to $T < 0.01$, stop the calculation, set the prediction number of true reads for the position is set to the reported number of reads. Proceed to next position on chromosome. We found that at positions where the $T$ collapsed, there was
usually one serial number with a large N, but the combined number of duplicates for the other serial numbers was less than $Pmr(0)N_{dup}$.

2b) If for initial $T = 1$, $G$ for the maximal assignment is $< 3N$, define a variable $T_L = 1$. Increase $T$ by a factor $e - 1$ until $G$ for the maximal assignment $> 3N$. Call the higher temperature $T_H$. If $T$ increases to $>110$, and $G$ for the maximal assignment is still $< 3N$, stop the calculation, and use the maximal assignment of duplicates $T$ for which $G$ was closest to $3N$ (usually $T = 110$) to set the predicted number of true reads. Proceed to next position on chromosome.

If both $T_L$ and $T_H$ can be established, proceed to step 3.

3) Use the interval halving algorithm to adjust $T_L$ (with $G < 3N$) and $T_H$ (with $G > 3N$) until $G$ converges to $3N$ or $G$ oscillates between two values. If the latter happens, use the maximal assignment for the $T$ with $G$ closest to $3N$ to set the predicted number of true reads. Go onto next position on the chromosome.

We found that for all nine datasets, almost no positions with $N > 256$ had all serial numbers with the same number of duplicate reads, so that rate distortion method could analyze all positions with $N > 256$. (See Suppl. Fig. S8A.) A plot of the average proportion of serial numbers at a position judged as not real versus $N$ is shown in Suppl. Fig. S8B. In line with positions with $N > 1024$ being less informative, the rate of reduction declines sharply for $N > 1024$, except for datasets $sap{1^7}(32^0C)$ 2 and 3.

References


