Supplemental Figure Legends

Suppl. Figure S1. The primers used to amplify the integration products of ligation mediated PCR. A. The primers that comprised the ligation linker (HL1870 and HL1871) and the amplification primer (HL2216) are shown. B. The amplification primer/s that annealed to the LTR upstream of the serial numbers are shown. The z’s represent six different barcodes. WT1: CATG, WT2: AGTC, WT3: GACT, CHD1: ACGT, CHD2: TGCA, and CHD3: GTAC. HL2747 is the specialized sequencing primer used with the Illumina GAIIX. C. The primers used to create and amplify the serial numbers were HL2827 and HL2828. The resulting product was ligated into the Tf1 expression plasmids pHL2763 and pHL2770 to generate the expression libraries of WT and CHD versions of Tf1, respectively.

Suppl. Figure S2. Design of the serial number system. A. The wild-type sequence of the 5’ LTR of Tf1 is shown. The red sequence is the first 11 nucleotides of the Tf1 mRNA that forms the self-primer of reverse transcription (1). In orange are 5 nucleotides that anneal in the self-priming structure. Nucleotides in grey are present in the 5’ end of the transcript. TTAA is underlined to indicate the position of an MseI restriction site. Nucleotides in yellow participate in another 11 bp duplex structure of the self-primer structure. The green nucleotides anneal to form the U5-IR stem necessary for self-priming (2). The sequence of the U5-IR loop, previously shown to be unimportant for transposition, is underlined in blue. B. The sequence of the 5’ LTR was modified in the expression vector in order to introduce the serial
number. The nucleotides in green anneal to form the U5-IR stem. Nucleotides of the U5 loop were replaced with a unique sequence tag and 8 nucleotides of random sequence that constituted the serial numbers. The nucleotides in orange are a component of the self-priming structure.

Suppl. Figure S3. Over 1,000 intergenic regions had higher numbers of insertions than if integration were random. All intergenic regions of *S. pombe* were ordered on the X axis based on numbers of integration events. The amount of integration in each intergenic region is plotted on the Y axis. The integration produced by WT2 (Tf1s-neo) and WT3 are displayed along with an equal number of insertions positioned randomly (MRC) in A. and B. respectively. The integration of CHD2 and CHD3 (Tf1s-CHD-neo) are compared to an equal number of insertions positioned randomly (MRC) in C. and D. respectively. The centromeres and telomeres were excluded from the intergenic sequences because they are significantly longer than other intergenic regions.

Suppl. Figure S4. Sequence preference was observed in a substantial proportion of integration events of Tf1 lacking the chromodomain (Tf1s-CHD-neo). A. The 260,346 insertion events produced by CHD1, CHD2, and CHD3 experiments were ranked based on the number of independent insertion events that occurred at their position. The events that occurred at the sites with the greatest number of insertions had the highest ranking while the events that occurred at sites with single insertions had low rankings. These ranked events were put into 20 bins; each
included 13,017 events or 5% of the total insertions. Then for the insertion events in each bin we determined the average sequence similarity to the sequence preferences of the top 150 sites shown in Fig. 8A. B. Insertion events (MRC) positioned in silico at random sites were ranked and organized into 20 bins as in A.

Suppl. Figure S5. Integration frequencies at the 5,000 positions that had the highest match to the logo of the strongest integration sites (Fig. 8A). The positions were sorted on the X axis starting with the strongest match at position 1 and the lowest at position 5,000. Total integration levels (WT1+WT2+WT3) at each position is indicated on the Y axis.

Suppl. Figure S6. From the 5,000 positions that best matched the DNA logo of the strongest integration sites, the positions without integration were mapped relative to the closest ORF. The ORFs of the S. pombe genome were subdivided into 15 equal segments (red) and the positions within these bins were tabulated. Positions outside of ORFs were mapped upstream of the ORF if they were closest to the 5’ end of an ORF and positions were mapped downstream of the ORF if they were closest to the 3’ end of an ORF.
Bioinformatic Analysis

Sequence reads were trimmed and screened using HTtools, a visual interface using XCode for Macintosh computers. HTtools interfaces with our laboratory software which is written in Perl. This software screened for reads containing the end of the LTR. Then the LTR and any linker sequences were trimmed. The trimmed sequences were positioned in the genome using the NCBI BLAST software on a local computer. The *S. pombe* genome database used in BLAST was the Feb. 2007 version of the chromosome contigs from the Wellcome Trust Sanger Institute. The BLAST results were filtered to collect matches with genomic sequence that started from the first nucleotide after the LTR with expect (e) values less than or equal to 0.05. Of the matches that met these criteria, the one with the highest bit score was used to obtain the coordinates for the unique insertion sites. If there is no single largest bit score such that the difference between the two highest is less than 0.0001, the insertion was assumed to be in a repeat region and was omitted. If sequence reads match orientation and unique chromosome position but have different serial number sequences they are considered to be the result of independent insertion events. The total number of unique serial number sequences for each orientation and unique chromosome position was determined.

The logo analyses were created by a program written in Visual Basic that was based on an existing algorithm (Schneider and Stephens, 1990). The program, logo-cal-3.2.pl calculates the bit scores by compensating for the CG content of the *S. pombe* genome. The program DNAlogo then generates the graphs with the bit scores corresponding to the heights of the nucleotides.
The CDS coordinates for the *S. pombe* genome were from Wellcome Trust Sanger Institute, the January 2010 version.

Schneider, T. D., and Stephens, R. M., 1990 Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18:6097-6100.
Applying Rate Distortion Theory to the Mis-Read Problem

Introduction to Information Entropy

The idea of an entropy-like quantity as a measure of the information contained in a probability distribution was introduced by Shannon (3) who showed that given a probability distribution \( \{p_1, p_2, p_3, \ldots, p_N\} \), then a quantity \( H \) which behaves as an entropy is uniquely given (to within a proportionally constant) by

\[
H = -\sum_n p_n \log(p_n) \tag{Eq. 1}
\]

(Note: it is assumed that the probabilities sum to one.) \( H \) has a minimum value of 0 which occurs if some \( p_n = 1 \) (and all other \( p \)'s equal to zero). Its maximum value is \( \log(N) \), occurring when \( p_n = 1/N \) for all \( n \), which makes sense since in this situation the probability distribution has no structure.

Shannon went on to consider the entropy of the output of a communication when there is knowledge of the input (3). Suppose there are two random variables \( x \) and \( y \), and \( p(x_A, y_B) \) is the probability of the co-occurrence given value \( x_A \) for \( x \) and \( y_B \) for \( y \). Let \( \{x_i\} \) be a set of values observed for variable \( x \), and \( \{y_j\} \) be a set of values observed for variable \( y \). Then define the conditional probability distribution \( p_i(j) \) as

\[
p_i(j) = p(x_i, y_j) \left( \sum_j p(x_i, y_j) \right)^{-1} \tag{Eq. 2}
\]

The conditional entropy \( H_x(y) \) is defined as

\[
H_x(y) = -\sum_i \sum_j p(x_i, y_j) \log(p_i(j))
\]

The idea of this quantity is to define the uncertainty in \( y \) given knowledge of \( x \). If \( x \) and \( y \) are independent, then \( H_x(y) \) reduces to the entropy of the distribution of \( y \). If \( x \) and \( y \) are 1-to-1, then \( H_x(y) \) is zero.

Mutual Information and Rate Distortion Theory

Rate distortion theory is concerned with the change in information content of data during compression. The discussion of the theory here follows closely the introductory sections of key articles (5-8). Consider a stochastic mapping of data points \( \{X_i, 1 < i < N_X\} \) to another set of the data, \( \{Y_j, 1 < j < N_Y\} \), where \( N_Y \) are \( N_X \) are not necessarily equal. (As examples consider the mapping of phonemes from one language to another, or the compressing of a large data file into a much smaller one.) Let \( P(Y_j|X_i) \) be the probability that value \( X_i \) implies value \( Y_j \) (the mapping), and 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. If every \( X \) data point is unique and is to be mapped, then \( \Pr(X_i) \) is
simply $N_X^{-1}$. Based on the conditional entropy developed by Shannon (3), the mutual information $M(Y;X)$ of the mapping of $X$ to $Y$ is defined as

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

where

$$
Pc(Y_j) = \sum_{i=1}^{N_X} P(Y_j|X_i) \Pr(X_i)
$$

(Eq. 3)

($Pc(Y_j)$ can be thought as the probability that at least one $X$ will be mapped to $Y_j$). $M$ is defined so that it is maximized if there is no loss of data in the mapping, and minimized if the mapping completely destroys the data: 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 $M$ is $\log_2(N_X)$ bits. 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.

For a well-posed problem, every $X_i$ must be mapped into something, even if it is just to itself. Thus, there is the constraint

$$
\sum_j P(Y_j|X_i) = 1
$$

(Eq. 4)

The main issue in choosing the probabilities $P(Y_j|X_i)$ is to preserve relevant features in the data while keeping the amount of information transformed as small as possible. Suppose $d(X_i,Y_j)$ is a measure of how different data points $X_i$ and $Y_j$ are from each other. Then the goal of the rate distortion procedure is to minimize $M(Y;X)$ subject to minimizing the distortion in the mapping, $D(Y;X)$, defined as

$$
D(Y;X) = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} d(X_i,Y_j) \Pr(Y_j|X_i) \Pr(X_i)
$$

(Eq. 5)

The idea is that $D(Y;X)$ is to be kept small if $Y$ is to preserve some important property of $X$. For example, the dataset $X$ could be the pixel information in a gigantic image file, and $d$ chosen so that a much smaller dataset $Y$ preserves some feature in the image such as location of the edges. In formal terms, the problem is chose probabilities $P(Y_j|X_i)$ to minimize the functional

$$
F = D(Y;X) + TM(Y;X)
$$

(Eq. 6)

The parameter $T$ is called the information temperature, chosen to balance $D$ against $M$: the larger $T$ is, the more information that is lost. $T$ is not fixed intrinsically by Eq 6: it must be chosen so that features considered relevant are preserved. The $P(Y_j|X_i)$ that minimize $F$ are the solutions to the implicit equations.
\[ P(Y_j|X_i) = \frac{P_c(Y_j)}{Z(X_i,T)} \exp\left(-\frac{\log(2) \ d(X_i,Y_j)}{T}\right) \quad \text{(Eq. 7)} \]

where

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

Remember that the \( P_c(Y_j) \) are dependent on the \( P(Y_j|X_i) \) variables, so this solution for the \( P(Y_j|X_i) \) variables must be solved by iteration.

One can see the analogy with the statistical mechanics of a gas of particles, with \( d \) defining an energy landscape of particle interaction that yields internal energy \( D \), \( -M \) is the entropy, and \( F \) a free energy.

**Rate Distortion Algorithm to eliminate the false Serial Numbers generated by Illumina sequence errors**

Consider a set of serial number sequences read at a chromosome position, \( S = \{S_1, S_2, \ldots, S_N\} \). Because the sequencing process has an error rate, some of the serial numbers in \( S \) may actually be mis-reads. 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 \( S_i \). (In the language of the above section, finding the probabilities \( P(S_j|S_i) \) is equivalent to finding a mapping of \( S \) back to itself that indicates that some serial numbers are likely mis-reads.) Since the mis-read information in the fixed LTR sequence gives an estimate of the sequencing error rate, we make use of the mis-read information in the fixed LTR sequence to help determine a distortion function for a rate distortion calculation to estimate the \( P(S_j|S_i) \). Although examination of the read errors in the LTR sequence shows that the mis-read probability is position and base dependent, it also indicates that the probability that a sequence is a mis-read of the LTR is inversely proportional to the number of base differences between that sequence and the LTR. In order to proceed, two assumptions are made: (1) that the probability that one serial number is a mis-read of another decreases with difference in base pairs, and (2) that the frequency that the serial number is mis-read is the overall frequency of mis-reads for the LTR sequence. Based on assumption (1), given sequences \( S_i \) and \( S_j \), \( d(S_i,S_j) \) for use in determining the distortion \( D \) is defined as the number of positions for which bases are different between \( S_i \) and \( S_j \).

We now indicate how assumption (2) is applied. At a given chromosome position let \( ndp_n \) be number of duplicate reads of serial number \( n \). Then the total number of duplicates is \( N_{\text{dup}} = \sum_n ndp_n \). (The sum is over all \( N \) serial numbers at the position.) If one were to choose some random duplicate from the set of all duplications for the positions, the probability that it is a duplicate for serial number \( S_i \) is \( dup_i/N_{\text{dup}} \). Thus, in the language of the above section, the Bayes prior probability for serial number \( S_i \) that would be used for a rate distortion calculation is \( Pr(S_i) = dup_i/N_{\text{dup}} \). Given an estimate for all the \( P(S_j|S_i) \) probabilities, the predicted true number of duplicates for a serial number \( S_i \), \( nprd(S_i) \) would be
\[ nprd(S_i) = \sum_j ndp_j P(S_j \mid S_i) \]  
(Eq. 8)

(Note that the sum \( \Sigma_i nprd(S_i) \) is \( N_{dup} \).) Since \( P(S_j \mid S_i) \) is the probability that \( S_i \) is really its own read, then \( nprd(S_i) - ndp_i P(S_j \mid S_i) \) is the expected number of duplicates from mis-reads of \( S_i \). Thus, the predicted number of duplicates at that chromosomal position that are due to mis-reads is

\[ nprMS = N_{dup} - \Sigma_i ndp_i P(S_j \mid S_i) \]  
(Eq. 9)

Using assumption (2), we assume that the probability that a serial number will be mis-read is the same as the frequency of misreads of an identical sized sequence of LTR that was adjacent to the serial number. This in turn gives a way to set the information temperature. For example, for the WT1 sequence dataset, the proportion of LTR reads that were mis-read within an 8 bp window was \(~0.032\). Thus, under assumption (2), we assume that at a chromosomal position for the WT1 dataset, the expected number of mis-reads, \(<MS>\), is \(~0.032\ \( N_{dup} \). The information temperature \( T \) is set by requiring \( nprMS = <MS> \). The predicted number of true serial numbers for a position is taken as the number of serial numbers with \( nprd(S_i) > 0.5 \). (The idea is that if the number of predicted duplications for a serial number does not round up one or more, then the serial number is considered a mis-read of another serial number.) The mis-read frequencies used for the datasets are shown in the table below.

**Details of Implementation**

For the yeast datasets, the algorithm was implemented as follows: for each chromosomal position in the dataset do the following actions:

1) If the number of serial numbers is just one, regardless of \( N_{dup} \), no calculation is done. The predicted number of mis-reads \( nprMS \) for the position is set to zero. Go on to next position.
2) If the number of serial numbers is greater than one, calculate the expected number of mis-reads for the position, \(<MS>\).
3) Start with initial \( T = 0.000001 \). Call this temperature \( T_L \). Calculate \( nprMS \) for \( T_L \) and call it \( nprMS_L \).
4) Consider temperature \( T = 10 \), and call it \( T_H \). Calculate the \( nprMS \) and call it \( nprMS_H \). If \( nprMS_H < <MS> \), or if \( nprMS_H < nprMS_L \), stop the temperature iteration go to step 7.
5) Compute a new trial temperature, \( T_{Tr} = \exp( \log(T_H) / 2 + \log(T_L) / 2 \) ), and compute \( nprMS \). Call it \( nprMS_{Tr} \). If \( |nprMS_{Tr} - <MS>| \) is within 0.1, then stop temperature iteration and go to step 7. If not, go to step 6.
6) If \( nprMS_{Tr} > <MS> \) take \( T_{Tr} \) as the new value of \( T_H \), else take \( T_{Tr} \) to be the new value of \( T_L \). Go to step 5.
7) Find the number of serial number sequences for which \( nprd(S_i) > 0.5 \). That is the prediction of the number of true serial numbers.

8) Go on to the next chromosome position in the database until the end is reached.

9) If for a given temperature, the rate distortion algorithm cannot converge to a set of \( P(S_i | S_i) \), then abandon the calculation for the position. However, for the choice of distortion used above, this never had to be done for the tens of thousands chromosomal positions in the six yeast datasets examined.

### Table Frequency of Mis-reads of the LTR Sequence for the Yeast Datasets.

| Yeast Dataset | Frequency of Mis-read of LTR |
|--------------|-------------------------------|
| CHD1         | 0.039199281                   |
| CHD2         | 0.041220426                   |
| CHD3         | 0.045630872                   |
| WT1          | 0.032158218                   |
| WT2          | 0.098894804                   |
| WT3          | 0.106690631                   |

### References

1. Lin, J.H. and Levin, H.L. (1997) A complex structure in the mRNA of Tf1 is recognized and cleaved to generate the primer of reverse transcription. *Genes Dev*, 11, 270-285.
2. Lin, J.H. and Levin, H.L. (1998) Reverse transcription of a self-primed retrotransposon requires an RNA structure similar to the U5-IR stem-loop of retroviruses. *Mol Cell Biol*, 18, 6859-6869.
3. Shannon, C.E. (1948) A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27, 379-423, 623-656.
4. Le Grice, S.F., Panin, M., Kalayjian, R.C., Richter, N.J., Keith, G., Darlix, J.L. and Payne, S.L. (1991) Purification and characterization of recombinant equine infectious anemia virus reverse transcriptase. *J Virol*, 65, 7004-7007.
5. Tishby, N., Pereira, F.C. and Bialek, W. (1999) In Hejek, B. and Sreenivas, R. S. (eds.), *Thirty-seventh Annual Allerton Conference on Communication, Control, and Computing*. University of Illinois, Allerton House, Monticello, Illinois.
6. Still, S. and Bialek, W. (2004) How many clusters? An information-theoretic perspective. *Neural Computation*, 16, 2483-2506.
7. Still, S., Bialek, W. and Bottou, L. (2004) In Thrun, S., Saul, L. and Schoelkopf, B. (eds.). MIT Press, Cambridge, MA, Vol. 16, pp. 1165-1172.

8. Slonim, N., Atwal, G.S., Tkacik, G. and Bialek, W. (2005) Information-based clustering. Proc Natl Acad Sci U S A, 102, 18297-18302.
Serial number system

A.

WT_Tf1..................................................LTR→TGTCAGCAATACACTACACTACGTCTATGATACACTACGTTACACTACGTTGCGTATGATATGTACACTACGTTGCTTGGGATATAGCTAAGATCCGTTATTGATAATTACGGGTAACTATCCGTCTATATGCATAGATACCCGTTGATAAACAACTACTTATAATATTATAATAGTGTACCAACTGAACCTCGTTCTCTAGTTCAGTTATGAGCTATATAATGGATAGGTAACATTATAACCCTGTTAATACAATACTATACTCAGTGTCTACTCATACAACCTGTGTATTGTAATATAATTAGATCGCAAGGAAAAACTCACCGCAGTTCTACGTATCCTTAAATCATAATACCAAACTGCGTAGCTAACA

U5 loop

B.

Seq. modifications with serial numbers

Suppl. Figure S2
Suppl. Figure S3.
**A.**
Average percent similarity of target sequence to logo of top 150 CHD sites

**B.**
Average percent similarity of MRC to logo of top 50 CHD sites

**Suppl. Figure S4**
Suppl. Figure S5.
Suppl. Figure S6.