A

Sequence of Linker Oligonucleotides		
Number	Sequence	Description
HL1870	GTAATACGACTCACTATAGGGCTCCGCTTAAGGGAC	Linker oligonucleotide
HL1871	5' P- TAGTCCCTTAAGCGGAG-NH_3 3'	Amino tailed linker oligonucleotide

B

Sequence of Illumina Sequence Primers		
Number	Sequence	Description
HL3498	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT TCCGATCTACGTCTCACCGCAGTTGATGCATAGGAAGC	Tf1-LTR primers with barcodes (*sap1*$^+$ (25^{o}C) #1)
HL3512	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT TCCGATCTTGCACTCACCGCAGTTGATGCATAGGAAGC	Tf1-LTR primers with barcodes (*sap1*$^+$ (25^{o}C) #2)
HL3513	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT TCCGATCTGTACCTCACCGCAGTTGATGCATAGGAAGC	Tf1-LTR primers with barcodes (*sap1*$^+$ (25^{o}C) #3)
HL3514	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT TCCGATCTCATGCTCACCGCAGTTGATGCATAGGAAGC	Tf1-LTR primers with barcodes (*sap1-1* (25^{o}C) #1)
HL3520	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT TCCGATCTAGTCCTCACCGCAGTTGATGCATAGGAAGC	Tf1-LTR primers with barcodes (*sap1-1* (25^{o}C) #2)
HL3521	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT TCCGATCTGACTCTCACCGCAGTTGATGCATAGGAAGC	Tf1-LTR primers with barcodes (*sap1-1* (25^{o}C) #3)

C

Sequence of PCR Primer		
Number	Sequence	Description
HL2216	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGTAATACGACTCACTA TAGGGC	PCR primer that anneals to linker end

Figure S1. Sequences of primers/oligonucleotides used in the generation of Tf1 serial number data in *sap1*$^+$ and *sap1-1* cells. A. The sequence of amplification primers that annealed to the LTR upstream of the serial numbers are shown. The nucleotides highlighted in red signify six different barcodes: *sap1*$^+$ (25^{o}C) #1: AGCT, *sap1*$^+$ (25^{o}C) #2: TGCA, *sap1*$^+$ (25^{o}C) #3: GTAC, *sap1-1* (25^{o}C) #1: CATG, *sap1-1* (25^{o}C) #2: AGTC, and *sap1-1* (25^{o}C) #3: GACT. B and C. The sequences of (B) the oligonucleotides that comprised the ligation linker (HL1870 and HL1871) and of (C) the amplification primer (HL2216) are shown. Serial numbers were created and amplified using methods previously described (Chatterjee *et al*. 2014).

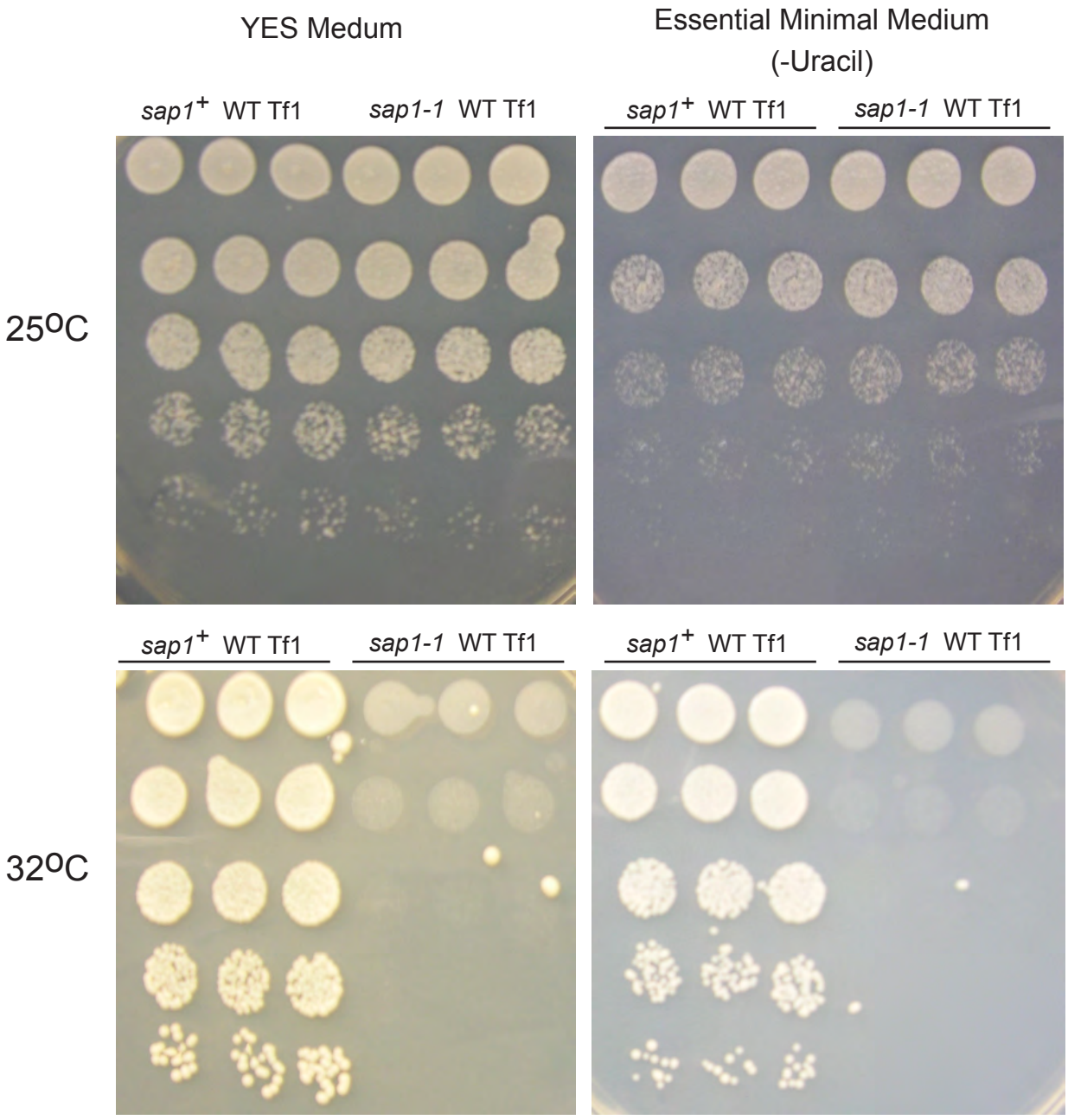

Figure S2. The *sap1-1* mutation causes no apparent growth defect at 25^{o}C. Five fold serial dilutions of *sap*$^{+}$ and *sap1-1 S. pombe* transformed with plasmids expressing WT Tf1 were spotted onto both solid YES and EMM –uracil media, and were grown at the indicated temperatures for 5 days. Three independent transformants from each genotype were assessed. The serial dilutions were prepared from resuspensions of plate grown yeast in liquid EMM-uracil at a starting O.D.600 of 0.500.

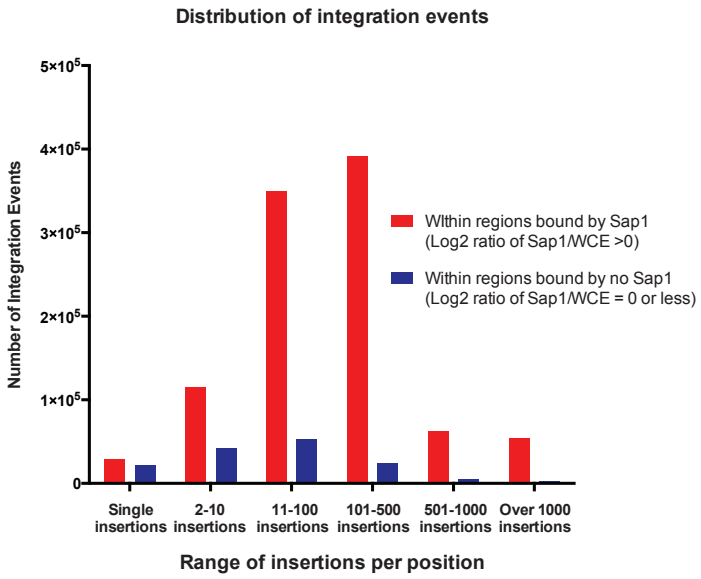

Figure S3. The majority of Tf1 integration occur in regions of the genome that are bound by Sap1. Graph showing the number of Tf1 insertions that occur within the indicated ranges of insertions per positions, as well as their occurrence in- and outside Sap1-binding regions in the genome.

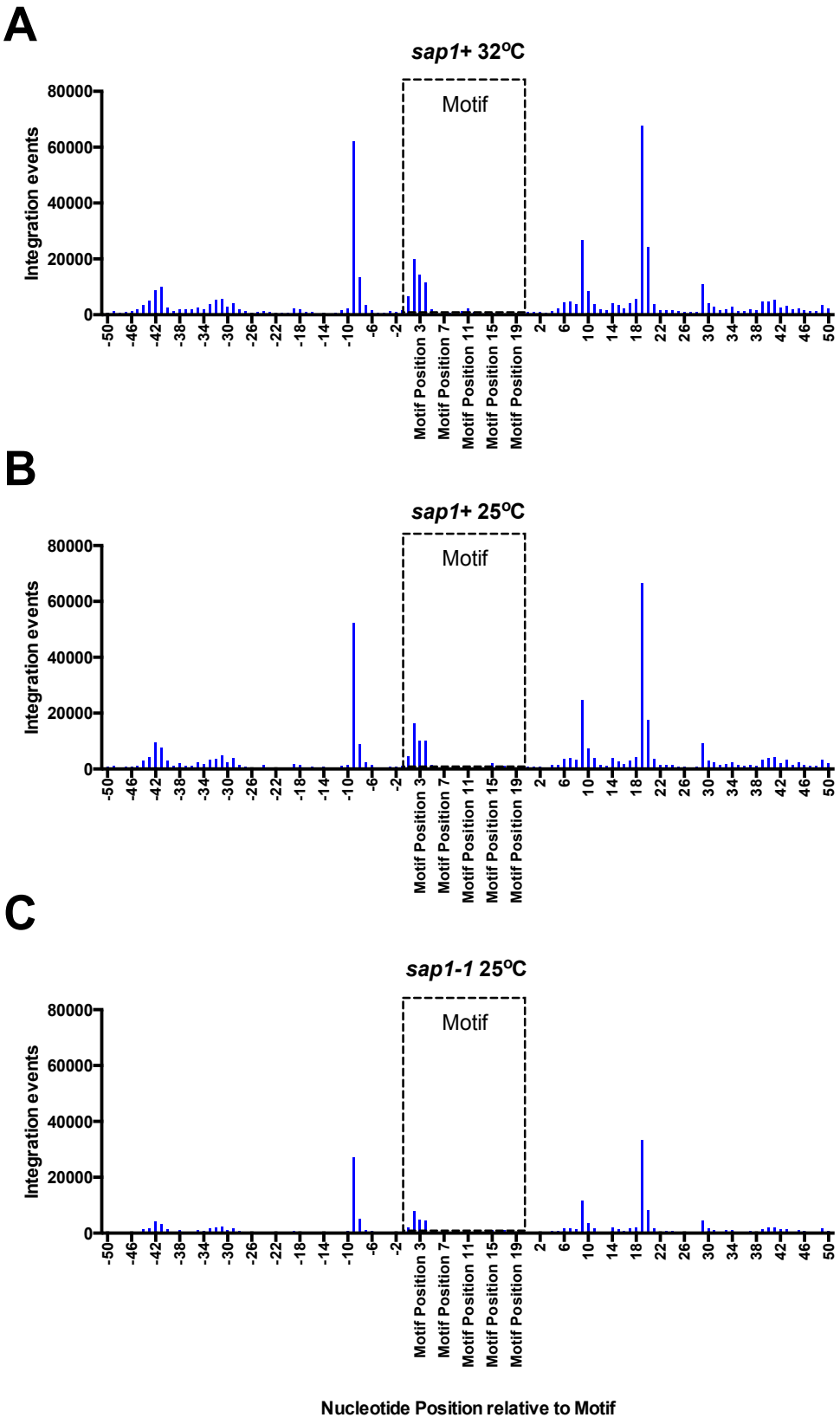
A
sap1+ 32°C
Motif
Integration events
80000
60000
40000
20000
0
-50 -46 -42 -38 -34 -30 -26 -22 -18 -14 -10 -6 -2 2 6 10 14 18 22 26 30 34 38 42 46 50
Motif Position 3
Motif Position 7
Motif Position 11
Motif Position 15
Motif Position 19
B
sap1+ 25°C
Motif
Integration events
80000
60000
40000
20000
0
-50 -46 -42 -38 -34 -30 -26 -22 -18 -14 -10 -6 -2 2 6 10 14 18 22 26 30 34 38 42 46 50
Motif Position 3
Motif Position 7
Motif Position 11
Motif Position 15
Motif Position 19
C
sap1-1 25°C
Motif
Integration events
80000
60000
40000
20000
0
-50 -46 -42 -38 -34 -30 -26 -22 -18 -14 -10 -6 -2 2 6 10 14 18 22 26 30 34 38 42 46 50
Motif Position 3
Motif Position 7
Motif Position 11
Motif Position 15
Motif Position 19
Nucleotide Position relative to Motif

Figure S4. The pattern of Tf1 insertions at specific nucleotide positions near Sap1 binding motifs is not altered in *sap1-1 yeast*. A. Graph showing the alignment of ~5000 genomic Sap1 motifs that were identified using FIMO of MEME Suite. The tabulated number of Tf1 insertions that occur at single nucleotide positions within 50bp of the aligned motifs in *sap1+* cells grown at 32^{o}C are plotted on the Y-axis. B and C. Same graph as in A generated from data collected from B) s*ap1*$^{+}$ cells grown at 25^{o}C and C) s*ap1-1* cells grown at 25^{o}C.

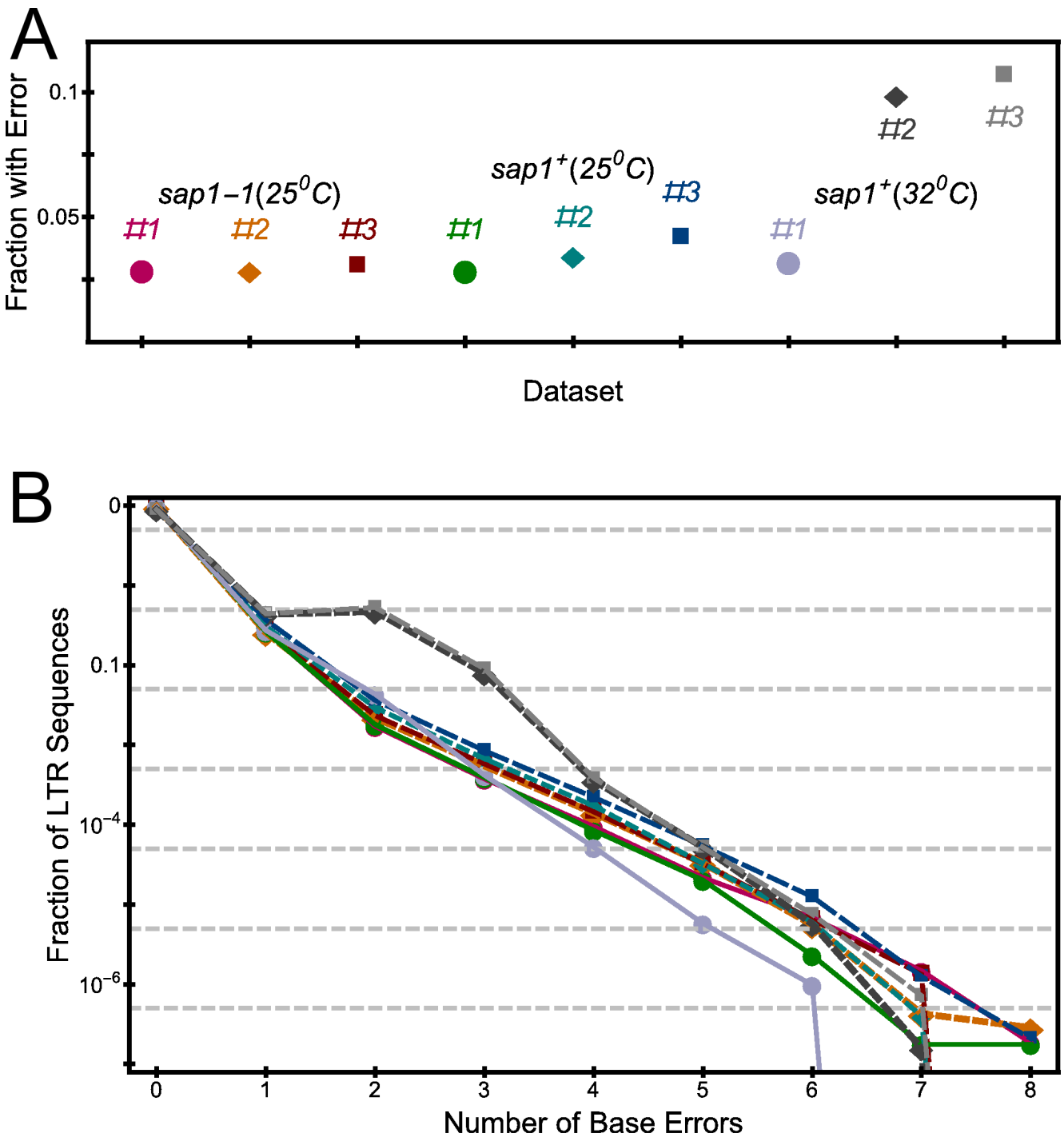

Figure S5. Error rate in reading the eight base pair serial number sequence. (A) Fraction of eight base pair sequences mis-read for each of the nine datasets. (B). Proportion of serial sequence reads with a given number of base mis-matches for each of the nine datasets. The symbols and colors for each data-base match those in (A). The dashed gray lines show fraction = 5×10^{-n}, n = 1, 2, 3, 4, 5, 6.

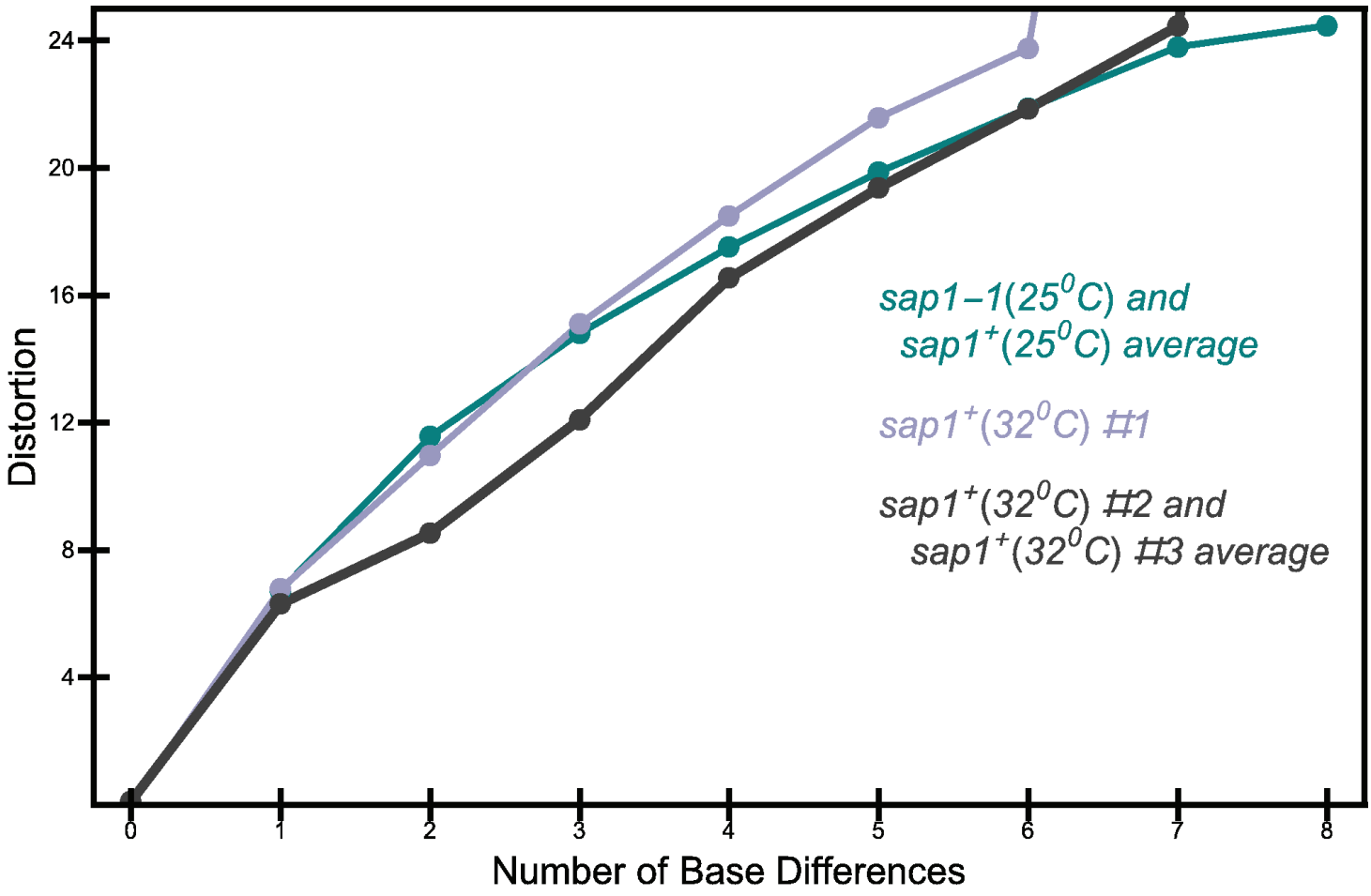

Figure S6. The distortion function between two serial numbers as a function of the number of base differences between them.

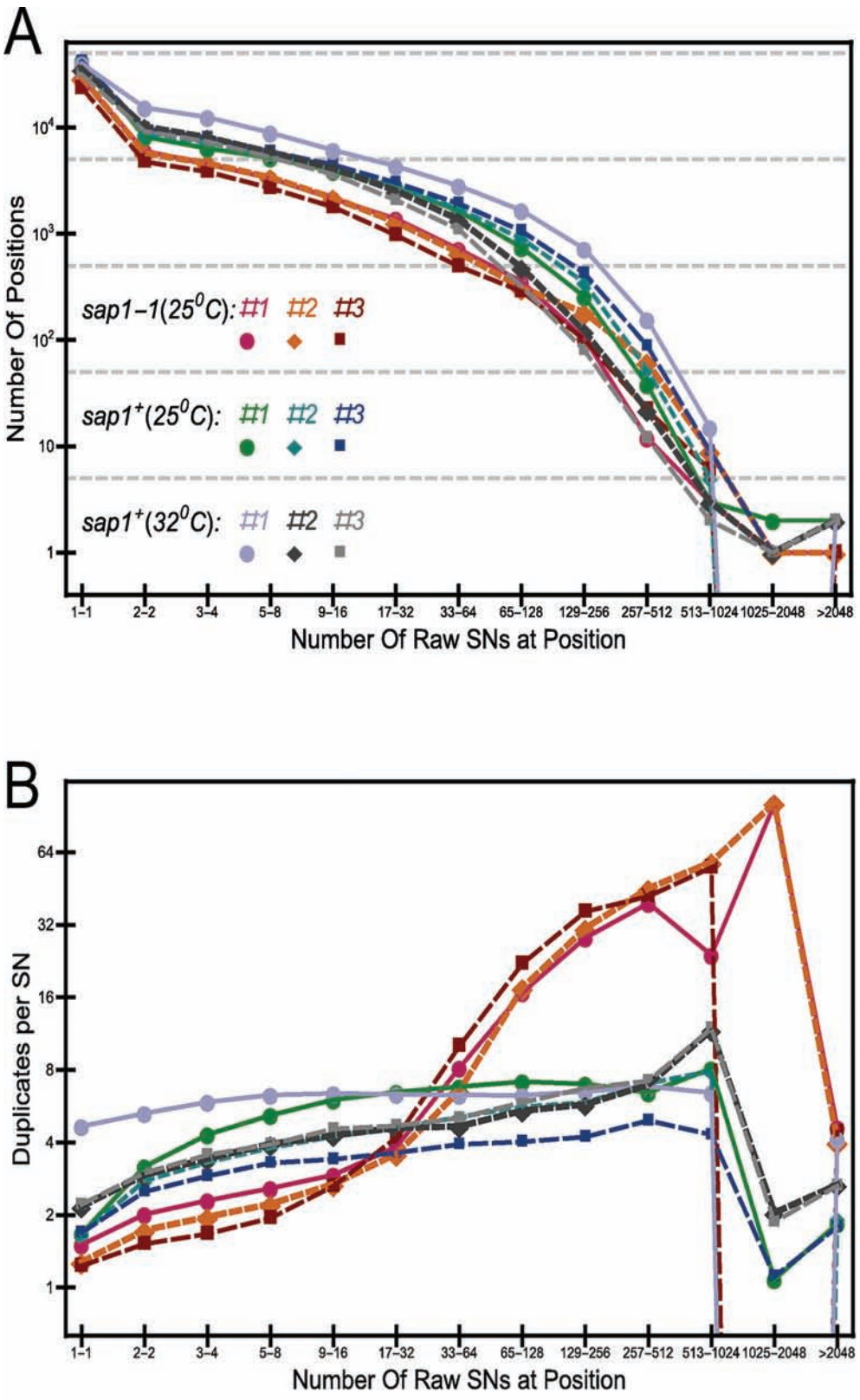

Figure S7. (A) Number of positions in each of the datasets versus the number of raw serial number at the position. The dashed gray lines show fraction = $5x10^n$, n = 1, 2, 3, 4. (B) Number of average duplicates per serial number at a position versus the number of raw serial numbers at the position. The symbols and colors for each database match those in (A).

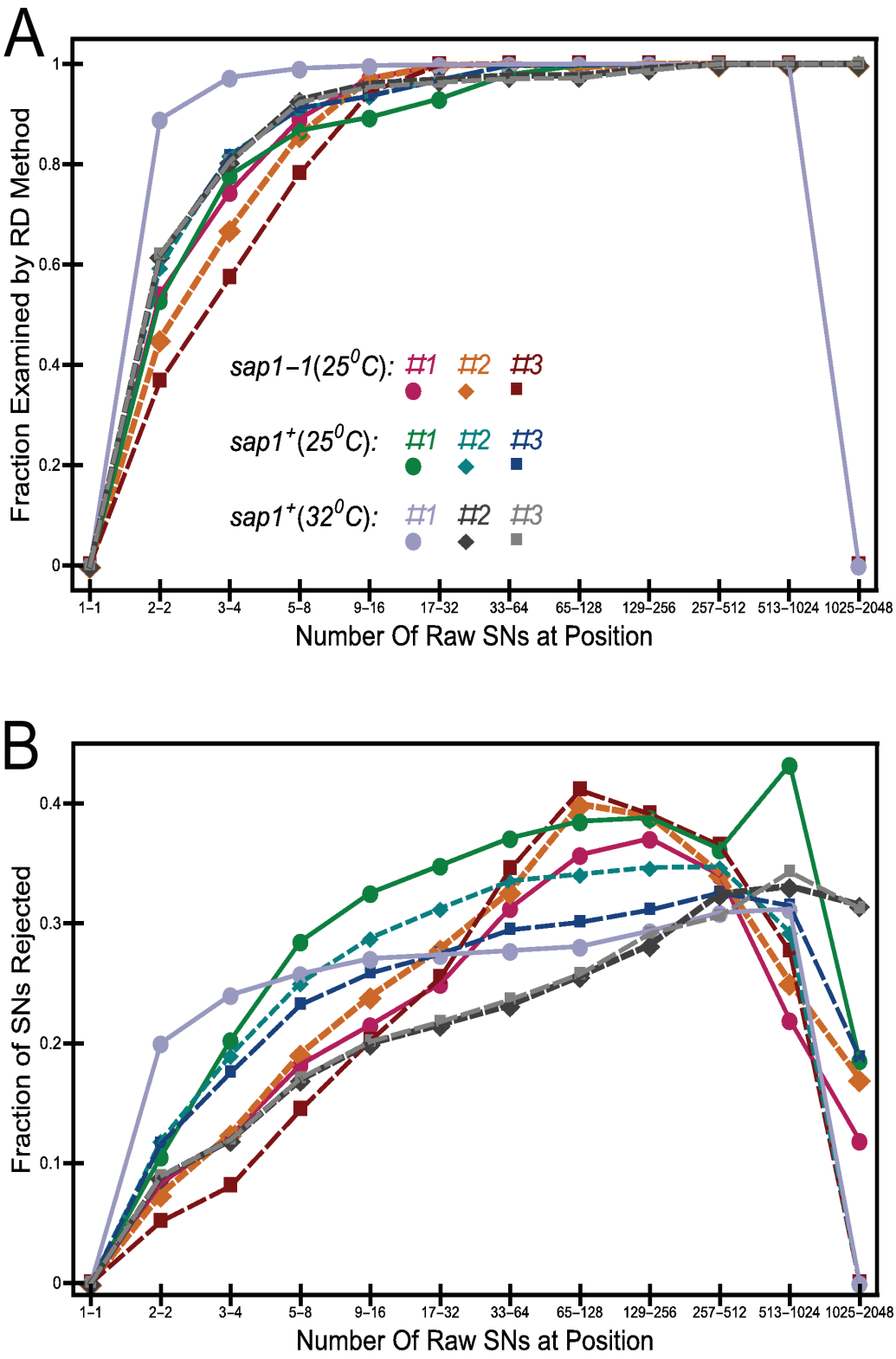

Figure S8. (A) Proportion of positions that could be analyzed by the rate distortion method versus the number of raw serial numbers at the position. (B) Average proportion of serial numbers judged as not real at a position by the rate distortion method versus the number of raw serial numbers at the position. The symbols and colors for each database match those in (A).

Table S1. List of computer programs and scripts used in this study (available upon request)

A. Perl Scripts		
Script Name	**Function/Description**	**Author**
CombineIntegrationsFiles.pl	Combines serial number Integration files from multiple experiments into 1 file with all the data.	Esnault
location_t_SN-1.0.pl	Analyzes intergenic regions, or regions between motif locations, and tabulates how many Tf1 insertion events occur in ORFs or motifs, as well as in the regions between. It will also assign intergenic/inter motif insertions to an ORF, or a motif, based on proximity	Guo
ORF_map_v5.3.pl	Takes the output from location_t_SN-1.0.pl, aligns all the ORF/Motifs, and tabulates the total integration at position flanking them (1000 positions upstream and 1000 positions downstream).	Guo, Hickey
Map_binding_profile_around_Tf1.pl	Used to align the Tf1 serial number data with Sap1-CHIPseq data.	Esnault
Sap1_Integration_Counter.pl	Counts the number of TF1 integration events that occur in regions of Sap1 enrichment, and groups insertions based on size of the peaks they are found in.	Hickey
GR_converter.pl	Converts an Integration serial number text file and generates 3 .gr files from it; one for each chromosome	Hickey
group_orientations_inGr-141027.pl	Takes an integration .gr file that has 2 sets of insertions values for the same positions, one for each orientation/strand (indicated by +/- values), and generates a new .gr file where those positions have singe positive values (the absolute values of both numbers combined).	Esnault
Master_GR_maker.pl	Takes 3 integration .gr files, one for each chromosome, and combines the data into a single master gr-like file, in the following format, chromosome # (as chr #), location, and # of insertions.	Hickey
gr_peakfinderv2.pl	Identifies Sap1 peak locations from Sap1 .gr files. Will assign peaks positions to any values above a selected threshold. This program not only gives coordinates of Sap1-peaks but also calculates the peak area by summing up all the Y-axis values of all coordinates within the peak.	Hickey
Sap1peakintcountV5.pl	Counts the number of Tf1 integration events that occurs in each Sap1 peak, and list as an output peak position, # of insertions, peak length, the percentage of peak total peak length each peak is, and peak area.	Hickey
Int_Peak_sorter.pl	Sorts the output of Sap1peakintcountV4.p, and groups peaks based on the number of Tf1 insertion evens in each peak.	Hickey
gr_fillerV2.pl	Scans a .gr file for nucleotide positions with no reported values and assigns them a value of "-1", indicating that Sap1 binding in not enriched for these positions. Such a manipulation was necessary for some future analyses.	Hickey
Master_Sap1_Gr_maker.pl	Takes 3 Sap1 .gr files, one for each chromosome, and combines the data into a single master .gr-like file, in the following format, chromosome # (as chr-#), location, and # of insertions. If the Sap1 .gr file is a "filled" .gr file (see gr_filler.pl) it replaces all values of "-1" with "0." This is necessary when tabulating Sap1 binding values around insertion sites.	Hickey
Master_Sap1_Gr_maker_chrm.pl	Similar to "Master_Sap1_Gr_maker.pl" except that for the output each chromosome is represented only as its number and does not have the "chr" prefix before it. Output format is: chromosome # (as #),	Hickey

	location, # of insertions	
Gr_to_Csv.pl	Converts the output from any of the above master .gr maker files and converts them to .csv format	Hickey
Combine_integration_into_mastermatrix.pl	Creates a comparative matrix of insertion positions and numbers between each position from 2 or more Integration serial number files.	Esnault
matrix_converter.pl	Takes a matrix output integration file that has 2 sets of insertions values for the same positions, one for each orientation/strand (indicated by +/- values), and generates a new .gr file where those positions have single positive values (the absolute values of both numbers combined).	Hickey
matrix_eliminator.pl	Allows the user to designate the minimum cutoff value required and will remove any values lower from the matrix file for further analysis. i.e. it was used to remove positions from the output file generated from matrixconverter.pl that were less than 3 insertions.	Hickey
matrix_eliminated_sorter.pl	Used to identify integration positions (based on the output from matrix-eliminator.pl) in which there is a greater than two fold difference in the number of Tf1 integrations in the *Sap1*$^{+}$ and *Sap1-1* strains. This program was also used to sort identified positions based on whether Tf1 integration is increased or decreased in the *Sap1-1* strain, as well as how many insertions were in these positions in the Sap1+ reference strain.	Hickey
LTR_PEAK_indentifier.pl	Identifies peaks that are associated with LTRs and lists them.	Hickey
duplicate_trimmer.pl	Some peaks are large and are associated with multiple LTRs, and as a result, are listed multiple times. This programs eliminated duplicate peaks from the list	Hickey
B. Python Scripts		
Script Name	**Function/Description**	**Author**
wigConverter.py	Used to calculate the Log 2 ratio of Sap1 signal to that of WCE, and generate the output as a WIG file.	Yang
aligner.py	Used to sort intergenic regions into bins of 500 TSSs and sorted them based on TF1 insertion number. It also aligned integration events in the region with Sap1 binding and nucleosome occupancy	Yang
C. R-Scripts		
Script **Name**	**Function/Description**	**Author**
Density_scatterplot-3datasets.R	Used to generate density plots comparing numbers of integration events as specific nucleotide positions in s*ap1*$^{+}$ and *sap1-1* cells.	Esnault

Table S2. Tf1 insertion numbers relative to genomic regions of Sap1 binding (Log_2 Ratio of Sap1/WCE > 0)

	Number of nucleotide positions	Percentage of genome	Number of Tf1 integrations	Percentage of Tf1 Integrations
Within regions of Sap1 binding	3,543,717	28.15	1,002,276	87.08
Within regions of No Sap1 binding	9,047,538	71.85	148,716	12.92
Total	12,591,255	100	1,150,992	100

Files S1-S9

Available for download as .txt files at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.181602/-/DC1

File S1 lists integration sites for *sap1*$^{+}$ (32°C) replica #1

File S2 lists integration sites for *sap1*$^{+}$ (32°C) replica #2

File S3 lists integration sites for *sap1*$^{+}$ (32°C) replica #3

File S4 lists integration sites for *sap1*$^{+}$ (25°C) replica #1

File S5 lists integration sites for *sap1*$^{+}$ (25°C) replica #2

File S6 lists integration sites for *sap1*$^{+}$ (25°C) replica #3

File S7 lists integration sites for *sap1-1* (25°C) replica #1

File S8 lists integration sites for *sap1-1* (25°C) replica #2

File S9 lists integration sites for *sap1-1* (25°C) replica #3

File S10

Supplemental Methods

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 MseI 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 MseI 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, $\{Yj, 1 < j < N_Y\}$, 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:

$$\Sigma_j \, P(Y_j|X_i) = 1 \qquad \text{(Eq. 1)}$$

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

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

where

$$Pc(Y_j) = \Sigma_{i=1,N_X} P(Y_j|X_i) \, Pr(X_i) \qquad \text{(Eq. 2)}$$

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

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 Xi's. This value of M indicates that all information contained by the X distribution is preserved. (2) On the other hand, if P(Yj|Xi) 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(Yj|Xi) 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

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

The rate distortion procedure is to chose probabilities P(Yj|Xi) that minimize the functional

$$F = D(Y;X) + T \, M(Y;X) \qquad \text{(Eq. 4)}$$

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]

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

where

$$Z(X_i,T) = \sum_{j=1,N_Y} Pc(Y_j) \exp(-\log(2)\, d(X_i,Y_j)\, /T\,) \qquad \text{(Eq. 5)}$$

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, $\mathbf{S} = \{S_1, S_2, \ldots S_N\}$. 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

$$Pr(SN_i) = dup_i/N_{dup} \qquad \text{(Eq. 6)}$$

where dup_i is the number of duplicate reads for SN_i, and, and N_{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*(25^0C) and *sap1*$^+$(25^0C) datasets and dataset *sap1*$^+$(32^0C) #1, but ~10% for the *sap1*$^+$(32^0C) #2 and *sap1*$^+$(32^0C) #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*$^+$(32^0C) #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

$$Pmr(m) = r(m)*(\ (\ 8!\ (8 - m\ )!\ m!\ \ )^{-1}\ 3^{-m} \qquad \text{(Eq. 7)}$$

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*(25^0C) and *sap1*$^+$(25^0C) 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*$^+$(32^0C) #2 and *sap1*$^+$(32^0C) #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*$^+$(32^0C) #1 is some what smaller for $m > 3$ than for the other datasets, $Pmr(m)$ for *sap1*$^+$(32^0C) #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

$$d(SN_1,SN_2) = -ln(\ Pmr(m)) \qquad \text{(Eq. 8)}$$

(If $Pmr(m) = 0$, we take $d(SN_1,SN_2)$ as $ln(10^{22})$, which is ~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 *Pmr(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

$$\mathrm{P}(SN_2|SN_1)/\mathrm{P}(SN_3|SN_1) = \mathrm{Pc}(SN_2)*\mathrm{Pc}(SN_3)^{-1}*(Pmr(m_{12})/Pmr(m_{13}))^{ln2/T} \qquad \text{(Eq. 9)}$$

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

To use the *Pmr(m)* function to fix the information temperature, we proceeded as follows: 1) Consider a set of serial numbers $\{S_1, S_2, \ldots S_N\}$ sequenced at a chromosome position, and let $\{dup_1, dup_2, \ldots dup_N\}$ 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 $\{\mathrm{P}(S_j|S_i)\}$for that *T*.

2) Define a new set of duplicate assignments, $\{dup_1', dup_2', \ldots dup_N'\}$, such that dup_i' = sum over all dup_j such that $\mathrm{P}(S_x|S_j)$ 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

$$\langle dup_i \rangle = \Sigma_j\ Pmr(m_{ij})\ dup_j' \qquad \text{(Eq. 10)}$$

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:

$$G = \Sigma_i\ dup_i \ln(dup_i/\langle dup_i \rangle) \qquad \text{(Eq. 11)}$$

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* < 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 $Pmr(0)*N_{dup}$, 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 $Pmr(0)*N_{dup}$. 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 bothT_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 (wth 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 *sap1*$^{+}$(32^0C) 2 and 3.

References

1. **Chatterjee AG, Esnault C, Guo Y, Hung S, McQueen PG, Levin HL.** 2014. Serial number tagging reveals a prominent sequence preference of retrotransposon integration. Nucleic Acids Res **42:**8449-8460.
2. **Zaratiegui M, Vaughn MW, Irvine DV, Goto D, Watt S, Bahler J, Arcangioli B, Martienssen RA.** 2011. CENP-B preserves genome integrity at replication forks paused by retrotransposon LTR. Nature **469:**112-115.
3. **Li H, Durbin R.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25:**1754-1760.
4. **de Castro E, Soriano I, Marin L, Serrano R, Quintales L, Antequera F.** 2012. Nucleosomal organization of replication origins and meiotic recombination hotspots in fission yeast. EMBO J **31:**124-137.
5. **Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS.** 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res **37:**W202-208.
6. **Grant CE, Bailey TL, Noble WS.** 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics **27:**1017-1018.
7. **Shannon CE.** 1948. A Mathematical Theory of Communication. The Bell System Technical Journal **27:**379-423, 623-656.