Single-Nucleotide-Specific Targeting of the Tf1 Retrotransposon Promoted by the DNA-Binding Protein Sap1 of *Schizosaccharomyces pombe*

Anthony Hickey,* Caroline Esnault,* Anasuya Majumdar,*[,1] Atreyi Ghatak Chatterjee,*[,2] James R. Iben,† Philip G. McQueen,‡ Andrew X. Yang,* Takeshi Mizuguchi,§ Shiv I. S. Grewal,§ and Henry L. Levin*[,3]
*Section on Eukaryotic Transposable Elements, Program in Cellular Regulation and Metabolism and †Program in Genomics of Differentiation, Eunice Kennedy Shriver National Institute of Child Health and Human Development, ‡Mathematical and Statistical Computing Laboratory, Division of Computational Bioscience, Center for Information Technology, and §Laboratory of Biochemistry and Molecular Biology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892

ABSTRACT Transposable elements (TEs) constitute a substantial fraction of the eukaryotic genome and, as a result, have a complex relationship with their host that is both adversarial and dependent. To minimize damage to cellular genes, TEs possess mechanisms that target integration to sequences of low importance. However, the retrotransposon Tf1 of *Schizosaccharomyces pombe* integrates with a surprising bias for promoter sequences of stress-response genes. The clustering of integration in specific promoters suggests that Tf1 possesses a targeting mechanism that is important for evolutionary adaptation to changes in environment. We report here that Sap1, an essential DNA-binding protein, plays an important role in Tf1 integration. A mutation in Sap1 resulted in a 10-fold drop in Tf1 transposition, and measures of transposon intermediates support the argument that the defect occurred in the process of integration. Published ChIP-Seq data on Sap1 binding combined with high-density maps of Tf1 integration that measure independent insertions at single-nucleotide positions show that 73.4% of all integration occurs at genomic sequences bound by Sap1. This represents high selectivity because Sap1 binds just 6.8% of the genome. A genome-wide analysis of promoter sequences revealed that Sap1 binding and amounts of integration correlate strongly. More important, an alignment of the DNA-binding motif of Sap1 revealed integration clustered on both sides of the motif and showed high levels specifically at positions +19 and −9. These data indicate that Sap1 contributes to the efficiency and position of Tf1 integration.

KEYWORDS Sap1; Tf1; integration; transposition; *Schizosaccharomyces pombe*

RETROTRANSPOSONS are pervasive among eukaryotes and, in many cases, account for a substantial portion of the host genome (Moore *et al.* 2004; Scheifele *et al.* 2009; Levin and Moran 2011). The ability of these elements to selectively integrate into specific target sequences has been paramount to their success because the employment of specific targeting mechanisms has allowed these retrotransposons to propagate within host genomes without disrupting genes and compromising the host's survival (Levin and Moran 2011). The long terminal repeat (LTR) retrotransposons Ty1, Ty3, and Ty5 of *Saccharomyces cerevisiae* avoid causing damage to the host by targeting noncoding genomic regions; Ty1 and Ty3 integrate upstream of RNA polemerase III–transcribed genes, while Ty5 integrates into heterochromatin (Chalker and Sandmeyer 1990; 1992; Ji *et al.* 1993; Kirchner *et al.* 1995; Devine and Boeke 1996; Zou *et al.* 1996; Zou and Voytas 1997; Yieh *et al.* 2000; Sandmeyer 2003; Lesage and Todeschini 2005).

In *Schizosaccharomyces pombe*, the LTR retrotransposon Tf1 has a unique targeting mechanism that directs integration to promoters of RNA polymerase II–transcribed genes with a bias for the promoters of stress-response genes (Behrens *et al.* 2000; Singleton and Levin 2002; Bowen *et al.* 2003; Leem *et al.* 2008; Guo and Levin 2010). Surprisingly, insertion of Tf1 into

doi: 10.1534/genetics.115.181602
Manuscript received August 3, 2015; accepted for publication September 3, 2015; published Early Online September 9, 2015.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.181602/-/DC1

[1] Present address: Vidyasagar College for Women, Kolkata 700 006, West Bengal, India

[2] Present address: Department of Biological Sciences, Presidency University, Kolkata 700 073, West Bengal, India

[3] Corresponding author: National Institutes of Health, 18 Library Dr., Rm. 106, Bethesda, MD 20892. E-mail: henry_levin@nih.gov

promoters rarely reduces the expression of their downstream genes, and in approximately 40% of cases, it enhances it (Feng *et al.* 2013).

The LTR elements Ty1, Ty3, and Ty5 in *S. cerevisiae* rely on host factors to tether integrase (IN) to insertion sites. Ty1 IN binds the AC40 subunit of RNA Pol III to direct integration to sites upstream of transfer RNA (tRNA) genes (Bridier-Nahmias *et al.* 2015). Ty3 IN binds transcription factors TFIIIB and TFIIIC to position insertions upstream of tRNA genes, while phosphorylated Ty5 IN interacts with the heterochromatin protein Sir4 to direct Ty5 insertion into heterochromatin (Kirchner *et al.* 1995; Gai and Voytas 1998; Zhu *et al.* 1999; Yieh *et al.* 2000, 2002; Xie *et al.* 2001; Qi and Sandmeyer 2012). Retroviruses such as human immunodeficiency virus 1 (HIV-1) and murine leukemia virus (MLV), which share significant similarities with LTR retrotransposons in their genetic structures and mechanisms of propagation (Levin and Moran 2011), depend on host factors for targeting integration. HIV-1 insertion is directed to the body of RNA polemerase II–transcribed genes by host factor LEDGF, while MLV IN interacts with BET proteins to direct its integration into enhancer sequences of RNA polemerase II–transcribed genes (Ciuffi *et al.* 2005; Llano *et al.* 2006; Shun *et al.* 2007; Gupta *et al.* 2013; Sharma *et al.* 2013). The mechanism by which Tf1 accomplishes targeting appears to be significantly different from the preceding examples except perhaps for MLV, which does integrate into promoter sequences. While the BET proteins are transcription coactivators, none of the transcription factors of stress-responce genes in *S. pombe* appear to play a role in Tf1 integration (Majumdar *et al.* 2011). As a result, there is still little understanding about how Tf1 integration is positioned.

Switch-activating protein 1 (Sap1), an essential DNA-binding protein in *S. pombe* (Arcangioli *et al.* 1994), binds sequences in the LTR of Tf1, as well as genomic regions where Tf1 insertion occurs (Zaratiegui *et al.* 2011). Sap1 has multiple reported functions, including facilitating mating type switching and causing replication fork arrest at places of genomic instability (Arcangioli and Klar 1991; Krings and Bastia 2005, 2006; Mejia-Ramirez *et al.* 2005; Noguchi and Noguchi 2007). To determine whether Sap1 plays a role in Tf1 retrotransposition, we studied *S. pombe* harboring the temperature-sensitive mutant *sap1-1* (Noguchi and Noguchi 2007). We found that Tf1 transposition is reduced 10-fold in the *sap1-1* mutant strain compared to wild-type *sap1*$^+$, and this defect was not due to decreases in levels of Tf1 proteins or complementary DNA (cDNA). Together with results of a recombination assay indicating that the *sap1-1* mutant did not inhibit transport of Tf1 cDNA to the nucleus, these data argue that Sap1 contributes to the integration of Tf1. Analysis of ChIP-Seq data reveals that ~6.85% of the *S. pombe* genome is bound by Sap1. Genome-wide profiles of Tf1 integration with measures of independent insertions at single-nucleotide positions revealed that 73.4% of Tf1 insertions occurred within these Sap1-bound sequences. A strong correlation was observed between positions with high numbers of repeated integration and locations where Sap1 binding was greatest. In addition, analysis of promoter sequences identified strong binding of Sap1 at the nucleosome-free regions (NFRs), and this binding correlated not only with peaks of integration but also with the size of the NFRs. We identified a Sap1 DNA-binding motif and found that Tf1 insertions clustered at two specific nucleotide positions adjacent to the motif, providing additional evidence that Sap1 promotes Tf1 integration.

Materials and Methods

Strains used in this work are listed in Table 1, and the plasmids are listed in Table 2. The media used in this study included yeast extract medium with supplements (YES) and essential minimal medium (EMM), which were prepared as described previously (Forsburg and Rhind 2006) with the following modifications: YES was supplemented with 2 g complete dropout powder, while EMM was supplemented with 2 g dropout powder lacking leucine and uracil. Dropout stock powder was prepared by mixing 5 g adenine SO_4 with 2 g each of the remaining 19 amino acids and uracil.

Drop assays

sap1$^+$ and *sap1-1* cells with pHL449-1 (wild-type Tf1) were collected from EMM-uracil plates and resuspended in liquid EMM at a starting OD_{600} of 0.5. From these initial resuspensions, four fivefold serial dilutions were prepared, and 10 μl of each resuspension and dilution was spotted onto EMM-uracil or YES plates and grown at either 25° or 32° for 5 days. Three independent transformants of each genotype were assessed.

Transposition assays and homologous recombination assays

Assays to determine Tf1 transposition and homologous recombination frequencies were conducted as described previously with the following modification: all assays performed with Tf1-*neoAI* in the *sap1-1* mutant *S. pombe* and the wild-type controls were incubated at 25° (Levin 1995, 1996; Teysset *et al.* 2003). Briefly, Tf1 transposition was monitored by placing a *neo*-marked Tf1 element (Tf1-*neoAI*) under the control of an inducible *nmt1* promoter in a donor plasmid. After the artificial intron (AI) is spliced out, the *neo* gene allows cells to grow in the presence of 500 μg/ml G418. Patches of *S. pombe* strains containing donor plasmids were grown on EMM-uracil dropout agar plates in the absence of thiamine to induce transcription of the *nmt1* promoter and were further incubated for 4 days. The plates then were replica printed to EMM containing 1 mg/ml 5-fluoroorotic acid (5-FOA) to counterselect against the donor plasmid (Boeke *et al.* 1987). As a final step, patches were replica printed to plates containing YES, G418, and 5-FOA and incubated 2 additional days to detect strains with integration of Tf1. Wild-type Tf1 produced confluent patches of G418 resistance, while protease frameshift (PRfs) and IN frameshift (INfs) mutations in the Tf1 element reduced cellular growth on the G418-containing plates.

Table 1 Yeast strains

Strain number	Genotype	Proteins expressed	Source
YHL912	h^{-} *ura4-294 leu1-32*		Boeke *et al.* 1987, 21X5A
YHL9752	h^{+} *leu1-32 ura4 D-18 Sap1-1ts -3X*::*NAT*		This study
YHL5661	Diploid *ura4 D-18/ura4 D-18 ade6-m210/ade6-m216 leu1-32/leu1-32*::*nmt1-lacZ-leu1*		Singleton and Levin 2002
YHL9716	*CTY10-5d MAT***a** *ade2 trp1-901 leu2-3,112 his3-200 gal4 gal80 ura3*::*lexAop-lacZ ura3-52*		Studamire and Goff 2008
YHL9777	YHL9716/pHL2781, pHL2793	LexA-Tf1-IN, Gal4-Tf1-IN	This study
YHL9774	YHL9716/pHL2778, pHL2783	LexA, Gal4	This study
YHL9775	YHL9716/pHL2778, pHL2793	LexA, Gal4-Tf1-IN	This study
YHL9776	YHL9716/pHL2781, pHL2783	LexA-Tf1-IN, Gal4	This study
YHL10822	YHL9716/pHL2936, pHL2780	LexA-Sap1, Gal4	This study
YHL10798-10800	YHL9716/pHL2936	LexA-Sap1	This study
YHL10804-10815	YHL9716/pHL2936, pHL2793	LexA-Sap1, Gal4-Tf1-IN	This study
YHL10823	YHL9716/pHL2936, pHL2938	LexA-Sap1, Gal4-Sap1	This study
YHL10801-10803	YHL9716/pHL2938	Gal4-Sap1	This study
YHL10820-10821	YHL9716/pHL2938, pHL2778	LexA, Gal4-Sap1	This study
YHL10816-10819	YHL9716/pHL2938, pHL2781	LexA-Tf1-IN, Gal4-Sap1	This study
YHL10014	YHL912/pHL449-1	Tf1-NeoAI	This study
YHL10015	YHL912/pHL490-80	Tf1-NeoAI (PRfs)	This study
YHL10016	YHL912/pHL472-3	Tf1-NeoAI (Infs)	This study
YHL10017	YHL9752/pHL449-1	Tf1-NeoAI	This study
YHL10018	YHL9752/pHL490-80	Tf1-NeoAI (PRfs)	This study
YHL10019	YHL9752/pHL476-3	Tf1-NeoAI (INfs)	This study

Homologous recombination between cDNA and plasmid sequences was assayed using a protocol similar to the transposition assay with the following modification (Atwood *et al.* 1996): strains harboring Tf1-*neoAI* donor plasmids were first grown as patches on agar plates that contained EMM-uracil (plus 10 μM thiamine) and then replica printed to EMM-uracil plates that lacked thiamine. After 4 days of incubation, the patches were replica printed directly to YES containing 500 μg/ml G418.

Transposition assays conducted to identify the location and frequency per position of Tf1 integration were performed using the Tf1 serial number library, as described previously (Chatterjee *et al.* 2014). *S. pombe* strains YHL9752 (*sap1-1*) and YHL5661 (wild type) were transformed with the serial number Tf1-*neo* plasmid library, and serial number insertion libraries were constructed for each individual strain by pooling approximately 55,000 and 60,000 independent wild-type and *sap1-1* transformants, respectively, from EMM plates lacking uracil. To repress expression of *neo*-marked Tf1 from the *nmt1* promoter, thiamine was added to plated medium at a concentration of 10 μM, which was removed prior to induction by mixing the pooled cells for 1 hr at 25° and washing them four times with 225 ml EMM lacking both uracil and thiamine. Transposition was induced by growing cells at 25° in EMM in the absence of uracil and thiamine. The wild-type and *sap1-1* mutant cultures were passaged with repeated dilutions to an OD$_{600}$ of 0.05 until they reached 53 generations; then the cultures were diluted to an OD$_{600}$ of 0.25 in EMM containing 5-FOA and regrown to an OD$_{600}$ of 5.0. This selected against cells retaining the Tf1-containing plasmids. The cultures were diluted 10-fold to an OD$_{600}$ of 0.5 in YES containing both 5-FOA and G418 and grown to an OD$_{600}$ of 5.0 to isolate both wild-type and *sap1-1* mutant cells containing integrated copies of Tf1$_s$-*neo*.

Quantitative transposition assay

Quantitative transposition assays were performed as described previously (Majumdar *et al.* 2011) to measure the frequencies of transposition of wild-type and *sap1-1* mutant *S. pombe* strains. Briefly, strains were grown on solid EMM without thiamine and after 4 days were resuspended in liquid medium to an OD$_{600}$ of 5.0. A series of five 10-fold dilutions were generated from each resuspension starting with 10^8 cells/ml and ending with 10^4 cells/ml, and 100 μl of cells from the three lowest dilutions was spread onto YES plates containing 5-FOA, while 100 μl of the three highest dilutions were spread onto YES plates containing both 5-FOA and G418. The transposition frequency is reported as the percentage of 5-FOA/G418-doubly-resistant colonies relative to the total number of 5-FOA-resistant colonies. The data presented in this work were compiled from three independent transformants.

Quantitative recombination assay

Cells were resuspended in liquid EMM without thiamine to an OD$_{600}$ of 0.05 and were grown for 6 days at 25°. A series of five 10-fold dilutions was generated from each culture starting with 2×10^7 cells/ml and ending with 2×10^3 cells/ml, and 100 μl of cells from the three lowest dilutions was spread onto YES plates, while 100 μl of the three highest dilutions were spread onto YES plates containing G418. The transposition frequency is reported as the percentage of G418-resistant colonies relative to the total number of colonies on nonselective YES plates. The data presented in this work were generated from four to five independent transformants.

Table 2 Plasmids used in this study

Plasmid number	Plasmid	Description	Source
pHL2778	pSH2-1	Expresses LexA DBD for yeast two-hybrid analysis	Studamire and Goff 2008
pHL2780	pACT2	Expresses Gal4 AD for yeast two-hybrid analysis	Studamire and Goff 2008
pHL2783	pACT1	Expresses Gal4 AD for yeast two-hybrid analysis	Durfee *et al.* 1993
pHL2781	pSH2-1-Tf1-IN	Expresses Tf1-intergrase fused to the LexA DBD	This study
pHL2793	pACT1-Tf1-IN	Expresses Tf1-intergrase fused to the Gal4 AD	This study
pHL2938	pACT2-Sap1	Expresses Sap1 fused to the Gal4 AD	This study
pHL2936	pSH-Sap1	Expresses Sap1 fused to the LexA-DBD	This study
pHL2944	Serial number library	Introduces serial number into wild-type Tf1-LTR	Chatterjee *et al.* 2014
pHL449-1	Tf-*neo*AI	Expresses Tf1 containing the *neo*-tagged artificial intron	Levin 1995
pHL490-80	Tf1-*neo*AI (PRfs)	Expresses Tf1 PRfs containing the *neo*-tagged artificial intron	Atwood *et al.* 1996
pHL472-3	Tf1-*neo*AI (INfs)	Expresses Tf1 INfs containing the *neo*-tagged artificial intron	Levin 1995

Two-hybrid analysis

The yeast two-hybrid system was used as described previously (Studamire and Goff 2008) with the following modifications: in brief, DNA segments encoding IN and Sap1 were amplified via PCR using *Eco*RI- and *Bam*HI-tailed primers in the case of Sap1 and *Eco*RI and *Sal*I in the case of Tf1 IN, which were designed to create the respective restriction sites on the 5′ and 3′ ends of the amplified products. These fragments were subsequently both cloned into the LexA DNA-binding domain (DBD) expression vector pSH2-1 and into the GAL4 activation domain (AD) expression vector pACT. The plasmids were transformed into *S. cerevisiae* strain YHL9716, which contains a copy of the *LacZ* gene downstream of a LexA operator. Transformants were patched onto synthetic complete (SC) plates lacking histidine and leucine. After 3 days of growth, the patches were transferred to a nitrocellulose membrane and frozen overnight at −80°. Potential interactions were identified by incubating the nitrocellulose in Z-buffer (Miller 1972) and identifying blue patches producing β-galactosidase.

Comparison of Sap1$^+$ and Sap1-1 integration profiles

Methods for sequencing insertion sites are included in the Supplemental Methods (Supporting Information) and the oligos used are listed in Figure S1. The integration site data are provided in File S1, File S2, File S3, File S4, File S5, File S6, File S7, File S8, and File S9. PERL scripts were used to identify the number of insertions per intergenic region that were sequenced from six independent cultures: three independent cultures of wild-type *S. pombe* and three independent cultures of the *Sap1-1* mutant (Table S1). Additional PERL scripts were used to generate matrices reflecting the change in the number of insertions per intergenic region and per nucleotide position for the three combined integration profiles from each genotype (Table S1). Positions were discarded if both strains had fewer than three independent insertions. Because 512,312 insertions were mapped in the *sap1-1* strain and 1,086,402 insertions were mapped in the *sap1*$^+$ strain, the data from the *sap1-1* strain were normalized by multiplying each value in the *sap1-1* data set by 2.12. Graphs were generated of the normalized data using the software R, and linear regression and R^2 analyses were performed using Graphpad Prism. PERL scripts were used to identify insertion positions exhibiting a greater than twofold difference between the *sap1*$^+$ and *sap1-1* strains (Table S1).

Detection of Tf1 glycosaminoglycan, integrase, and reverse transcriptase by immunoblot analysis

Total proteins were extracted from cells grown under Tf1-inducing conditions using a previously published protocol (Atwood *et al.* 1996). Briefly, cultures were grown to an OD_{600} of 10 and were harvested by resuspension in 400 μl extraction buffer (15 mM NaCl, 10 mM HEPES-KOH, pH 7.8, and 5 mM EDTA). Cells were then vortexed in the presence of 0.4-mm acid-washed glass beads (Sigma), and the resulting crude protein extracts were recovered and mixed with an equal volume of 2× sample buffer and then boiled for 10 min. Then 10 μg of total protein from each sample was collected and loaded on SDS–10% polyacrylamide gels for immunoblot analysis. Standard electrophoresis and transfer techniques were used with Immobilon-P membranes (Millipore) in conjunction with the ECL System for detection as described by the manufacturer (Amersham). To visualize that equal amounts of protein were loaded in each lane and to determine that the transfer occurred evenly, we stained the membranes with 0.1% Coomassie Brilliant Blue G250 in 50% methanol. After 5 min, the filters were rinsed with water and destained in a solution of 50% methanol and 10% acetic acid. Horseradish peroxidase–conjugated donkey anti-rabbit immunoglobulin was used at a 1:10,000 dilution. The polyclonal antisera used to detect Gag, integrase (IN), and reverse transcriptase (RT) have been described previously (Levin *et al.* 1993; Hoff *et al.* 1998).

Preparation and analysis of nucleic acid

cDNA preparations were performed as described previously (Atwood-Moore *et al.* 2006). In brief, total DNA was extracted from cells grown for 36 hr using glass beads and phenol. The DNA was digested with *Bst*XI prior to analysis by DNA blotting. The probe consisted of a 1-kb *Bam*HI fragment with the sequence of *neo* (Atwood *et al.* 1996). The radiolabeled probe was generated using ^{32}P random primer labeling (Roche).

Data availability

All strains, plasmids, and computer programs/scripts are available upon request. Sequence reads were submitted to the Short Read Archive (SRA) at National Center for

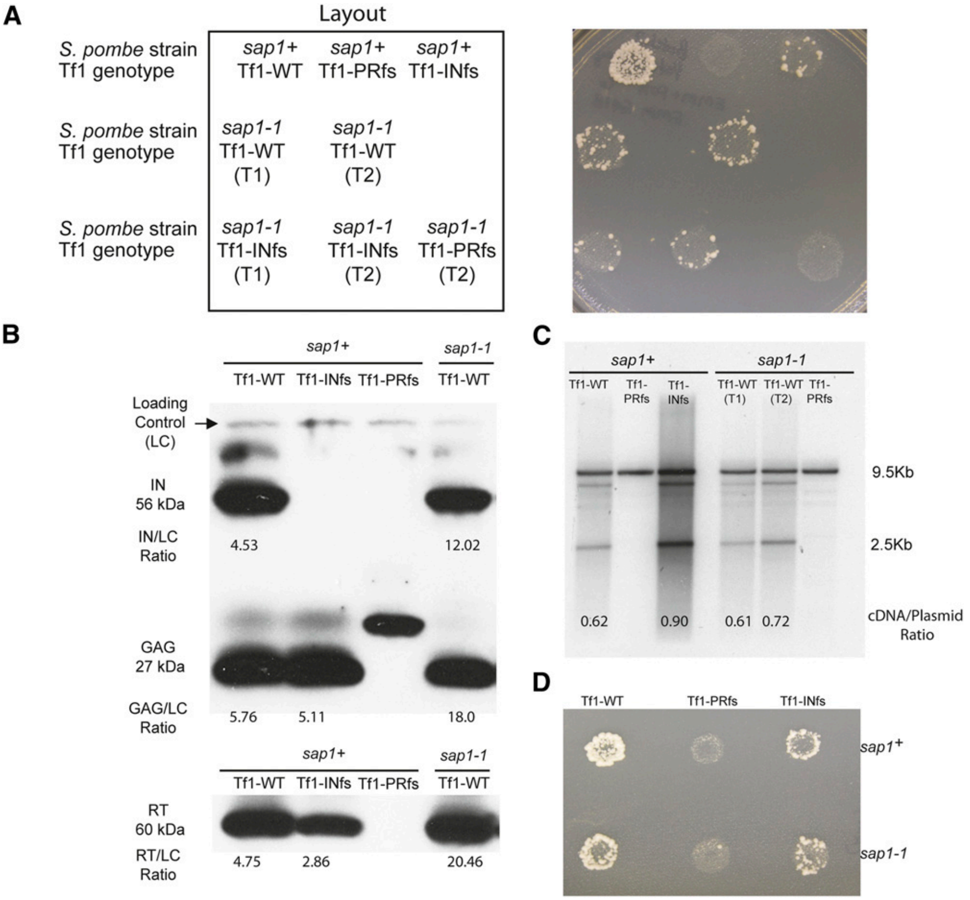

Figure 1 (A) Tf1 retrotransposition is reduced in *S. pombe* with a temperature-sensitive *sap1-1* allele. The results of transposition patch assays of wild-type *S. pombe* (*sap1*$^+$) and of *S. pombe* harboring a temperature sensitive allele of *sap1* (*sap1-1*). Tf1-PRfs, protease frameshift Tf1 mutant; Tf1-INfs, IN frameshift Tf1 mutant. Patches in which the expression of Tf1 was induced were first replica printed to plates containing 5-FOA and then printed to plates with 5-FOA and G418. T1 and T2 indicate two individual transformants. (B) Immunoblot analysis of Tf1 Gag, IN, and RT in wild-type (*sap1*$^+$) and *sap1-1* mutant *S. pombe* extracts. Polyclonal antibodies specific for IN, Gag, and RT detected the mature 56-kDa IN, 27-kDa Gag, and 60-kDa RT species. "Loading control" indicates a nonspecific band used for the normalization of IN, Gag, and RT with ImageJ. The numbers represent the ratio of the intensity of the indicated band to the loading control. The frameshift transposons used in this experiment are the same as those listed in A. (C) DNA blot analysis of total DNA extracted from wild-type (*sap1*$^+$) and *sap1-1* mutant *S. pombe* cells and digested with *Bst*XI. Blots were probed with a randomly labeled 1-kb *Bam*HI fragment containing the *neo* sequence. The large 9.5-kb fragment resulting from the digestion of the Tf1 donor plasmid served as the loading control for comparing levels of the 2.5-kb cDNA. Numbers indicate the ratio of the intensities of the indicated cDNA band to the loading control. (D) To measure recombination, cells were replica printed to plates with 5-FOA and then printed to plates with 5-FOA and G418.

Biotechnology Information (NCBI) under the accession number PRJNA279274.

Results

Sap1 promotes Tf1 integration

To test whether Sap1 is a host factor that plays a role in Tf1 transposition, assays were conducted in mutant *S. pombe* harboring the temperature-sensitive allele *sap1-1*. The *sap1-1* mutant showed no apparent growth defects on EMM or YES at 25° compared to the *sap1*$^+$ control; however, as reported previously, it was unable to grow at 37° (Figure S2). Transposition was measured using an assay that detects the insertion of *neo*-marked Tf1 (Tf1-*neoAI*) elements into the *S. pombe* genome (Levin 1996; Atwood *et al.* 1998; Teysset *et al.* 2003). This assay relies on the expression of a plasmid-encoded Tf1-*neoAI* to generate integration events, which are detected by the ability of cells to grow in the presence of 5-FOA to select against cells containing the original *URA3*-tagged donor plasmid and G418 to select for insertions. At the permissive temperature of 25°, cells containing the *sap1-1* mutant allele had substantially less transposition activity than the wild-type *sap1*$^+$ cells (Figure 1A, *sap1*$^+$, Tf1-WT *vs.* *sap1-1*, Tf1-WT). The transposition frequency of the *sap1-1* strain expressing wild-type Tf1-*neoAI* was only slightly higher than that of the *sap1*$^+$ strain expressing Tf1-*neoAI* with the INfs, indicating that the *sap1-1* mutant exhibited low-level residual Tf1 activity. INfs is used as a baseline for transposition assays because no IN is expressed, and the low level of G418R is due to homologous recombination between cDNA and LTR sequences in the genome (Levin 1996). Quantitative transposition assays (see *Materials and Methods*) revealed that Tf1 transposition was reduced by 10-fold in the *sap1-1* mutant (Table 3).

To determine whether the *sap1-1* mutation reduced Tf1 transposition by lowering expression, immunoblot analyses were performed using lysates of *sap1*$^+$ and *sap1-1* cells expressing wild-type and various mutant versions of Tf1. To quantitate the levels of Tf1 Gag, IN, and RT, ImageJ software was used to normalize each of the bands representing these proteins to a nonspecific band on the blot that was present in each lane. None of the Tf1 proteins analyzed were reduced in the *sap1-1* strain relative to *sap1*$^+$ but actually appeared to be increased by three- to fourfold (Figure 1B). While this apparent increase could simply be the result of a weak signal given by the nonspecific species in the *sap1-1* lane, these data argue that Tf1 Gag, IN, or RT expression was

Table 3 Results of quantitative Tf1 transposition assay with wild-type and *Sap1-1* mutant *S. pombe*

S. pombe genotype	Tf1 genotype	Average[a]	SD
sap1$^+$ (25°)	Wild type	0.17	3.5×10^{-3}
sap1$^+$ (25°)	INfs	0.010	3.5×10^{-3}
sap1$^+$ (25°)	PRFs	0	1×10^{-4}
sap1-1 (25°)	Wild type	0.019	9.5×10^{-3}

[a] Percentage of cells with Tf1 integrations from three independent transformants.

not reduced by the *sap1-1* mutation. To determine whether Tf1 cDNA production was reduced by the *sap1-1* mutation, DNA blot analysis was performed using a probe specific for the *neo* gene. To differentiate between the Tf1 cDNA and the original donor plasmid, the samples were digested with *Bst*XI, which results in a 2.1-kb band from the Tf1 cDNA and a 9.5-kb band from the donor plasmid (Atwood *et al.* 1996). We used ImageJ to quantitate the amount of Tf1 cDNA present in the lysates relative to the plasmid, and we found that the *sap1-1* mutation did not result in a defect in Tf1 cDNA production (Figure 1C).

Having determined that levels of Tf1 proteins and cDNA were not reduced in *sap1-1* cells, we next considered whether their reduced transposition was due to a defect in the nuclear import of Tf1 cDNA. This was tested using the homologous recombination assay, as described previously (Atwood-Moore *et al.* 2006), which measures the amount of Tf1 cDNA in the nucleus by detecting homologous recombination between Tf1 cDNA and the Tf1 donor plasmid. In this assay, the *neo* gene within the Tf1 transposon is disrupted by an AI that renders it inactive until the intron is removed by splicing during transcription. An active copy of the *neo* gene is generated during reverse transcription that is then able to convey resistance to G418 either by integration into the genome or, in this case, by homologous recombination with the donor plasmid. Unlike the transposition assay, the donor plasmid remains in the cells throughout the assay, allowing efficient homologous recombination to occur between the cDNA and the plasmid. The levels of G418 resistance produced by the INfs provided the measure of cDNA present in the nucleus. The *sap1-1* mutation did not reduce the level of homologous recombination in patch assays (Figure 1D; compare top-right to bottom-right patches), indicating that the nuclear import of the Tf1 cDNA is not significantly inhibited in *S. pombe* harboring the *sap1-1* mutant allele. To measure more precisely whether Sap1 makes a contribution to homologous recombination, we performed quantitative recombination assays in liquid cultures. The results of these assays revealed that homologous recombination was reduced in the *sap1-1* mutant by approximately 2.5-fold when comparing the INfs strains (Table 4). The contribution of Sap1 to recombination could be the result of reduced nuclear import. It is also possible that the Sap1-binding LTR sequence could stimulate homologous recombination. This would not affect transposition measures. Regardless, the 2.5-fold contribution of Sap1 to recombination is substantially less than the 10-fold contribution Sap1 makes to transposition, indicating that the bulk of the defect is in integration. Together with the observations that the levels of Tf1 protein and cDNA were not reduced in the *sap1-1* mutant, these data indicate that the *sap1-1* mutant significantly affected the process of integration.

If Sap1 contributes directly to Tf1 integration, the *sap1-1* mutation would have the potential to alter the positions of integration. We tested this possibility by generating dense profiles of integration sites by inducing plasmid-encoded copies of Tf1$_s$-*neo* carrying serial number tags of eight random base pairs that allowed us to measure independent insertions at single-nucleotide positions (Chatterjee *et al.* 2014). As explained previously (Chatterjee *et al.* 2014), we used rate-distortion analysis to compensate for errors in serial number measures that result from sequence misreads (Supplementary Methods, File S10 and Supplementary Figure S5, Figure S6, Figure S7, and Figure S8). We isolated genomic DNA from cells with Tf1 insertions, and by ligation-mediated PCR and high-throughput sequencing, we determined dense profiles of integration that included independent insertions at single-nucleotide positions. Three independent cultures of wild-type (*sap1*$^+$) cells produced integration profiles with an average of 92.3% of events located in intergenic sequences, a level similar to previous profiles and that results from the targeting of specific promoter sequences (Guo and Levin 2010; Chatterjee *et al.* 2014).

Inspection of integration in individual regions of the genome suggested that the numbers of insertions between replicas were highly reproducible (Figure 2A). Also, in these examples, the *sap1-1* mutation did not alter the integration pattern. To evaluate the integration patterns genome-wide, the numbers of insertions per intergenic region were tabulated. The amount of integration within intergenic regions was highly reproducible between the three independent replicas for both the *sap1*$^+$ and *sap1-1* experiments (Table 5, $R^2 > 0.8$). More important, the levels of integration in the intergenic regions correlated strongly between the *sap1*$^+$ and *sap1-1* experiments, indicating that the *sap1-1* mutation did not significantly change the integration pattern (Table 5, R^2 between 0.75 and 0.91). The combined integration in intergenic sequences of all three wild-type cultures when compared by linear regression with the combined integration of all three *sap-1-1* cultures showed a high level of correlation, with an R^2 of 0.92 (Figure 2B).

To test whether the *sap1-1* mutation altered the integration patterns within intergenic sequences, we conducted a more thorough examination of integrations at single-nucleotide positions. The tabulation of serial number measures of integration from the *sap1*$^+$ and *sap1-1* strains revealed a total of 153,848 independent insertion sites. Most of these sites had fewer than three insertions in both strains and thus were eliminated from the subsequent analysis. The analysis was conducted with the remaining 34,418 sites. The number of insertions mapped in the *sap1-1* strain was normalized to account for differences in the total number of insertions mapped in each strain (512,312 *vs.* 1,086,402 insertions in the *sap1-1* and *sap1*$^+$ strains, respectively). Analysis of the

Table 4 Quantitative Tf1 homologous recombination assay with *sap1*$^+$ and *sap1-1* strains

Tf1 genotype	*S. pombe* Genotype	Average[a]	SD	*sap1*$^+$/*sap1-1* ratio
Wild type	*sap1*$^+$ (25°)	1.34	2.90×10^{-1}	2.48
Wild type	*sap1-1* (25°)	5.4×10^{-1}	6.56×10^{-2}	
INfs	*sap1*$^+$ (25°)	6.9×10^{-1}	2.09×10^{-1}	2.48[b]
INfs	*sap1-1* (25°)	2.8×10^{-1}	1.17×10^{-1}	
PRfs	*sap1*$^+$ (25°)	0.0	1.34×10^{-4}	0.0
PRfs	*sap1-1* (25°)	1×10^{-3}	8.00×10^{-4}	

[a] Percentage of cells with G418 resistance from four or five independent transformants.
[b] Student's *t*-test *P* = 0.0083.

34,418 insertion sites revealed that the integration patterns of *sap1*$^+$ cells correlated highly with those of *sap1-1* cells (Figure 2C, $R^2 = 0.91$), indicating that although the *sap1-1* mutation lowered integration frequencies substantially, it did not cause a substantial change in the targeting of integration. On closer examination, we identified a total of 12,172 positions (35.37% of the 34,418 positions analyzed) that exhibited a greater than twofold change in the number of integrations between the two strains (Table 6). Of these positions, 8755 exhibited greater than twofold decreased numbers of Tf1 integrations in the *sap1-1* strain. These contained 8.09% of all integration events in the *sap1*$^+$ strain and only 2.14% of all events in the *sap1-1* mutant. Thus, at the positions that decreased more than twofold in the *sap1-1* mutation, there were 5.96% fewer integration events. Overall, comparing the 5.96% decrease in integration at positions with reduced integration to the 9.60% increase in integration that occurred at positions with greater than a twofold increase in integration, the *sap1-1* mutation caused a total increase of 3.64% at these positions.

To gain a better understanding of how the *sap1-1* mutation influenced integration at individual sites, all 12,172 of the positions with more than a twofold change were sorted into groups based on the number of integrations each position holds within the *sap1*$^+$ strain. This analysis revealed that while these 12,172 positions accounted for 12.18 and 15.83% of all insertions in the *sap1*$^+$ and *sap1-1* strains, respectively, the vast majority of these positions (7260 + 2878 = 10,138, or ~82%) contained fewer than 11 integration events within the *sap1*$^+$ strain, showing that the *sap1-1* mutation has the most impact on integration in positions that are weak targets for Tf1 integration. While a few of these altered positions were found to be integration hotspots (positions with over 100 integrations) in the reference *sap1*$^+$ strain, most of these hotspots were located in sites whose Tf1 activity increased in the *sap1-1* mutant rather than decreased (99 *vs.* 37 positions, respectively), further suggesting that the few strong integration targets that are affected by the *sap1-1* mutation generally become stronger targets in the *sap1-1* strain.

Despite identification of the preceding positions where Tf1 activity is altered by more than twofold by the *sap1-1* mutation, most of the Tf1 integrations in each strain (78.60 and 74.56% in the *sap1*$^+$ and *sap1-1* strains, respectively) were at positions that did not change more than twofold in Tf1 activity. This is consistent with our initial conclusion from our linear regression analysis: Tf1 integration-site preference was not grossly altered by the *sap1-1* mutation. While the *sap1-1* mutation does appear to result in minor changes in integration patterns, it is also possible that these differences resulted from selection because the *sap1-1* strain was haploid and the *sap1*$^+$ cells were diploid.

The strong contribution Sap1 makes to integration frequency suggests that it may play a role in integration. Such a role could be mediated through direct interaction. To test for direct interaction between Sap1 and Tf1 IN, we performed a series of pull-down experiments. We tested Sap1 and IN as purified recombinant proteins as well as proteins expressed in *S. pombe*. Despite testing many configurations of cell extracts and purified proteins, we were unable to obtain evidence that Sap1 interacts with IN (data not shown). We also tested for *in vivo* interaction between Sap1 and IN using the two-hybrid assay of *S. cerevisiae*. Fusion of IN to the C termini of the LexA DBD and the Gal4 AD detected IN:IN interaction, as indicated by high expression of a LexA operator–lacZ reporter (Figure 3, A and B, bottom-right panel). Similarly, LexA DBD and Gal4 AD fusions to Sap1 detected Sap1:Sap1 interactions (Figure 3B, bottom-right panel). Importantly, when Sap1 was fused to the C terminus of the LexA DBD and IN was fused to the C terminus of the Gal4 AD, a strong interaction was detected (Figure 3B, top-left panel). Interactions were not observed if the LexA DBD lacked Sap1 or if the Gal4 AD lacked IN. Also, interaction was not observed when IN was fused to the LexA DBD and Sap1 was fused to the Gal4 AD (Figure 3B, top-right panel).

Tf1 insertion sites align with genomic regions of enriched Sap1 binding

Tf1 insertion data lacking serial number measures of independent insertion at single-nucleotide positions (Guo and Levin 2010) were aligned with positions of Sap1 binding determined by ChIP-Seq (Zaratiegui *et al.* 2011). The alignment of the insertion sites revealed a peak of Sap1 binding approximately 60 bp from the integration sites (Zaratiegui *et al.* 2011). The correspondence between insertion sites and Sap1 binding in the genome suggested that Sap1 participates in integration. Here we performed a similar alignment using the previously reported positions of Sap1 binding (Zaratiegui *et al.* 2011) with Tf1 integration sites recently determined with the serial number system (Chatterjee *et al.* 2014). Sap1 enrichment throughout the *S. pombe* genome was determined by calculating the $\log_2$ ratio of Sap1 binding signal to that of the whole-cell-extract (WCE) control and was then tabulated relative to aligned

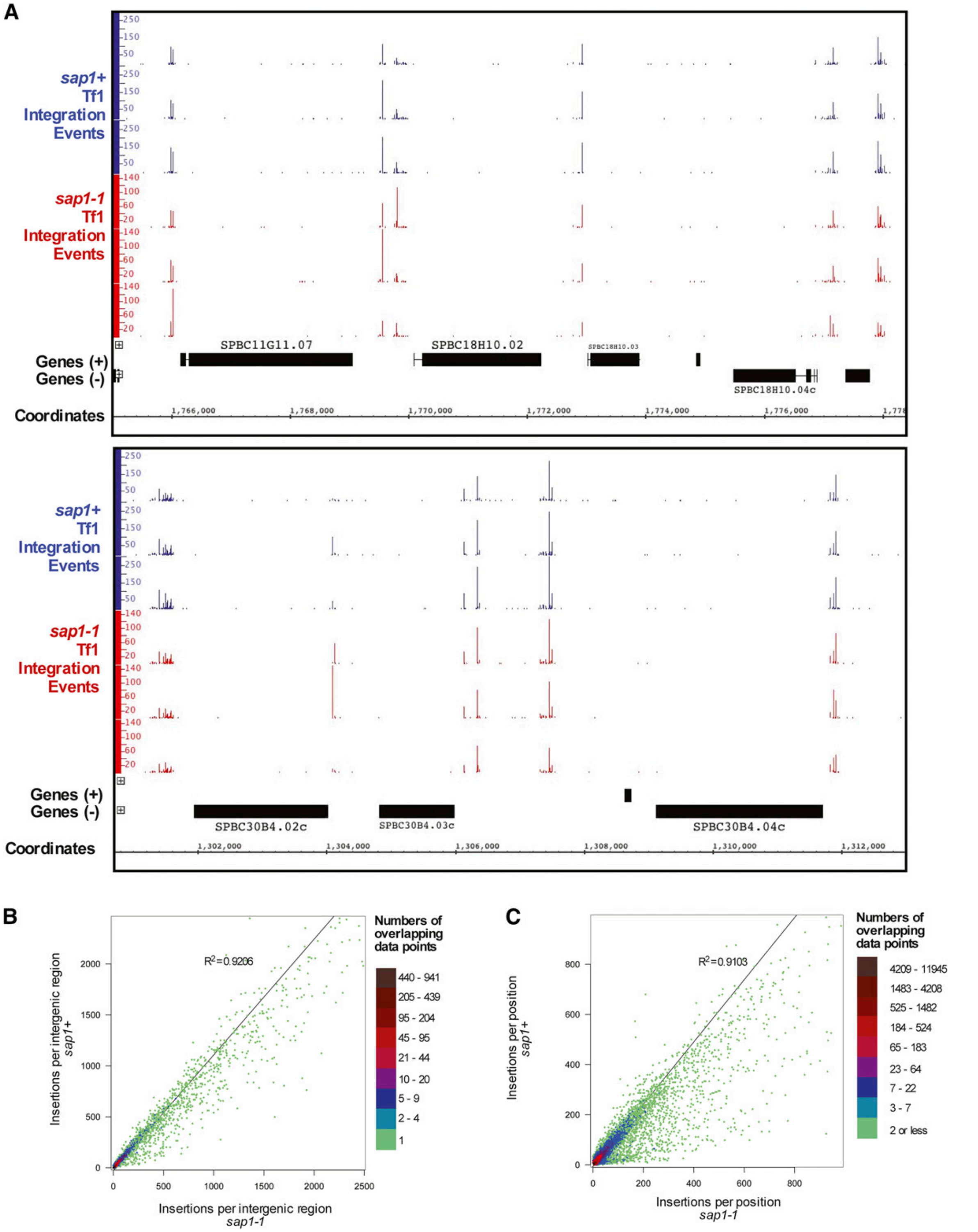

Figure 2 Integration positions are not substantially altered in *sap1-1* mutant cells. (A) Integration levels from *sap1*$^+$ and *sap1-1* cells are shown in two sample regions of the genome. The data shown are from three independent experiments. (B) Density scatter plot and linear regression analysis comparing the number of Tf1 insertions occurring within the intergenic regions of wild-type (*sap1*+) (*y*-axis) and mutant (*sap1-1*) cells (*x*-axis). The boxes to the left indicate ranges of the number of data points that overlap at any given plot in the graph (*i.e.*, the number of intergenic regions having the same *x*- and *y*-axis values). The data shown are combined from three independent experiments. (C) Density scatter plot and linear regression analysis of the number of Tf1 insertions occurring at specific nucleotide positions within wild-type (*sap1*$^+$) and mutant (*sap1-1*) cells. The boxes to the left indicate ranges of the number of data points that overlap at any given plot in the graph (*i.e.*, the number of positions with the same *x*- and *y*-axis values). The data shown are combined from three independent experiments. For B and C, density scatter plots were created using the function hexbin of the program R. Linear regression analyses were performed, and R^2 values were calculated using Graphpad Prism.

Table 5 Correlation coefficients (R^2) of the integration levels within the intergenic regions of wild-type and *Sap1-1* mutant *S. pombe* collected from three independent experiments

	sap1$^+$ (25°) #1	*sap1*$^+$ (25°) #2	*sap1*$^+$ (25°) #3	*sap1-1* (25°) #1	*sap1-1* (25°) #2	*sap1-1* (25°) #3
sap1$^+$ (25°) #1		0.96	0.91	0.77	0.81	0.76
sap1$^+$ (25°) #2			0.97	0.87	0.91	0.86
sap1$^+$ (25°) #3				0.92	0.90	0.91
sap1-1 (25°) #1					0.84	0.86
sap1-1 (25°) #2						0.84

Tf1 insertion positions. This analysis confirmed that Tf1 has a strong preference for integration into genomic positions of increased Sap1 binding (Figure 4A). No 60-bp offset between the insertion sites and Sap1 binding was observed.

The relationship between Sap1 binding and integration at individual genomic sequences revealed strong correlation between positions with high numbers of insertions and peaks of Sap1 enrichment (Figure 4B). To determine what fraction of integration events occurred at areas of Sap1 binding genome-wide, we identified all the regions of the *S. pombe* genome that were enriched twofold or greater for Sap1 binding (receiving a $\log_2$ score of 1) and then calculated the number of Tf1 insertions present within and outside these regions. While only 6.85% of the *S. pombe* genome was enriched above this twofold threshold for Sap1 binding, the vast majority of all integration events (73.1%) were found to lie within this fraction of the genome (Table 7).

In a detailed analysis that relied specifically on the serial number integration data, all Tf1 integration positions were sorted and grouped based on the number of independent

Table 6 Change in the distribution of Tf1 integrations resulting from *sap1-1* mutation

	No. of positions	Percent of total positions	No. of integrations in *sap1*$^+$ strain	Percent of integrations in *sap1*$^+$ strain	No. of integrations in *sap1-1* strain	Percent of integrations in *sap1-1* strain	Percent of integrations that change positions
Positions with a greater than twofold change in *sap1-1* relative to *sap1*$^+$	12,172	35.37	132,379	12.18	81,095	15.83	3.64
Positions with a greater than twofold decrease in integrations in *sap1-1* relative to *sap1*$^+$	8,755	25.44	87,944	8.09	10,947	2.14	−5.96
Positions with fewer than 11 integrations in *sap1*$^+$	7,260	21.09	34,429	3.17	2,438	0.48	−2.69
Positions with between 11 and 100 integrations in *sap1*$^+$	1,458	4.24	31,897	2.94	5,435	1.06	−1.88
Positions with between 101 and 500 integrations in *sap1*$^+$	33	0.10	5,449	0.50	1,102	0.22	−0.29
Positions with between 501 and 1000 integrations in *sap1*$^+$	3	0.01	1,761	0.16	256	0.05	−0.11
Positions with over 1000 integrations in *sap1*$^+$	1	0.00	14,408	1.33	1,716	0.33	−0.99
Positions with a greater than twofold increase in Tf1 integrations in *sap1-1* relative to *sap1*$^+$	3,417	9.93	44,435	4.09	70,148	13.69	9.60
Positions with fewer than 11 integrations in *sap1*$^+$	2,878	8.36	8,947	0.82	13,793	2.69	1.87
Positions with between 11 and 100 integrations in *sap1*$^+$	440	1.28	14,910	1.37	27,712	5.41	4.04
Positions with between 101 and 500 integrations in *sap1*$^+$	94	0.27	17,249	1.59	23,583	4.60	3.02
Positions with between 501 and 1000 integrations in *sap1*$^+$	5	0.01	3,329	0.31	5,060	0.99	0.68
Positions with over 1000 integrations in *sap1*$^+$	0	0.00	0	0.00	0	0.00	0.00
Positions with a less than twofold change in *sap1-1* relative to *sap1*$^+$	22,246	64.63	853,947	78.60	382,001	74.56	−4.04

Note: Numbers are nonnormalized values, and percentages are derived from nonnormalized values.

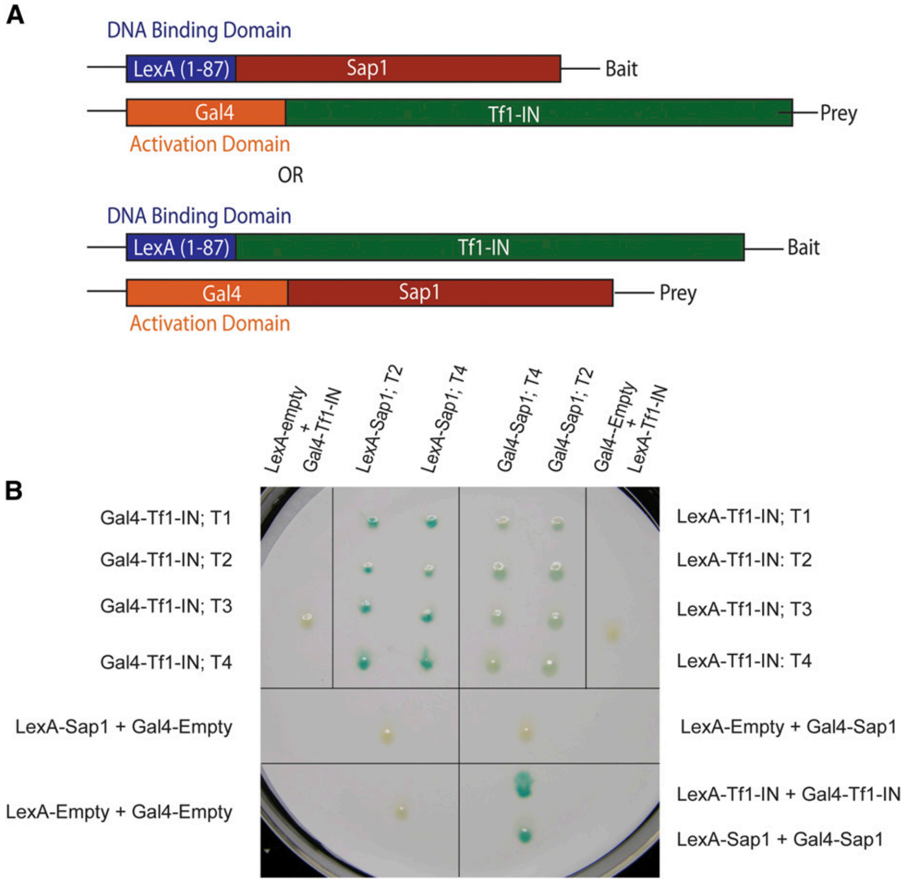

Figure 3 (A) Sap1 interacts with Tf1 IN in yeast two-hybrid assays. Diagram of the fusion proteins used in the two-hybrid assay. (B) Results of yeast two-hybrid assay. The fusion proteins expressed by each *S. cerevisiae* patch are indicated. Independent transformants are indicated by T followed by a number. "Gal4-Empty" indicates patches that express the Gal4 portion of the fusion protein only. "LexA-Empty" indicates patches that express the LexA portion of the fusion protein only. An interaction between the two fusion proteins is indicated by the color blue.

insertions that occurred per position. Integration sites inside regions of Sap1 enrichment were analyzed separately from those outside Sap1-enriched regions. The number of insertions within each group and their location relative to regions of Sap1 enrichment were assessed (Figure 4C). Positions with single insertions are the most evenly distributed between regions with and without Sap1 binding, with most of them lying outside areas of Sap1 enrichment. However, the sites with single insertions constitute just 4.0% of all integration events. Positions with 2–10 independent insertions had more integration in regions of Sap1 binding than outside. This bias becomes significantly more pronounced as the number of insertions per positions increases, with almost no positions containing over 500 insertions found outside Sap1-enriched regions. These analyses argue that there is a very strong correlation between sites with high levels of integration and positions where Sap1 binding is enriched.

Sap1 binding has been demonstrated previously to align with NFRs upstream of RNA polemerase II transcription start sites (Tsankov *et al.* 2011). The association of regions of Sap1 binding with integration sites led us to ask whether high levels of Tf1 integration were associated with large NFRs. To address this question, we sorted all the intergenic sequences within the *S. pombe* genome based on the number of Tf1 insertions within them and binned these sequences into groups of 500. The transcription start sites (TSSs) in each bin were aligned. Relying on previously published data, the average nucleosome occupancy (de Castro *et al.* 2012), Sap1-binding (Zaratiegui *et al.* 2011), and the normalized average number of Tf1 insertions (Chatterjee *et al.* 2014) at each nucleotide position within 1 kb of these TSSs were tabulated. Most Tf1 integration events lie upstream of the TSSs and form a peak that closely matches the binding of Sap1 (Figure 5A). This pattern is most apparent in bins 1–5, which had the greatest number of integration events. Importantly, bin 1 with the highest number of integration events contained the greatest enrichment for Sap1 binding, which peaked upstream of the TSS and aligned well with the Tf1 integration positions containing the greatest number of insertions. Bin 1 also had the largest average NFR. Although similar patterns were observed in bins 2–4, the peaks of Sap1 binding were significantly diminished in each subsequent bin, corresponding with decreases in insertion events relative to the number of TSSs and decreases in the average NFR. Bin 5 had almost no enrichment for Sap1 binding and very little integration. Linear regression analysis performed on the sums for the nucleotides lying within 1000 bp upstream of the TSS for each data set in each bin demonstrated that there were particularly strong correlations between Tf1 integration, Sap1 binding ($R^2 = 0.9748$), and the area of the NRFs ($R^2 = 0.9510$) (Figure 5B).

Tf1 insertions cluster around the Ter1 sequence in ribosomal DNA (rDNA) repeats but not around switch-activating site 1 (SAS1) in the mating locus

Sap1 has been previously reported to bind specific sequences in both the mating locus and the rDNA repeats in *S. pombe*.

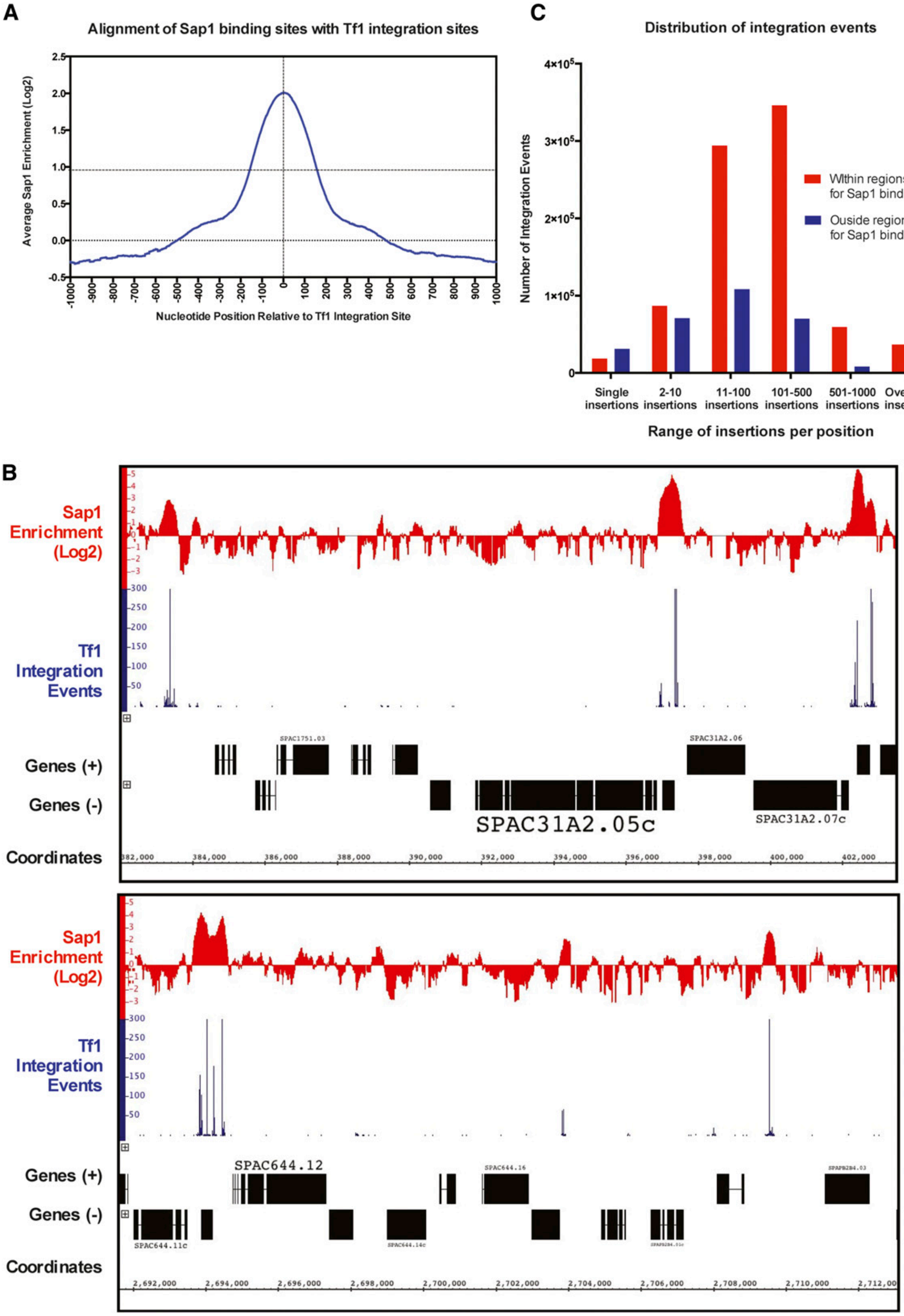

Figure 4 Tf1 prefers to integrate in Sap1-bound regions of the genome. (A) Graph showing the alignment of all genomic Tf1 insertion sites with the tabulated average of Sap1 enrichment ($\log_2$ ratio of Sap1 binding to WCE signal) at single-nucleotide positions within 1000 bp of the aligned insertion sites. (B) Representative regions of the genome showing that high integration sites align with regions of the genome enriched for Sap1 binding. (C) 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-enriched regions in the genome.

Table 7 Tf1 integration positions are predominantly in regions of Sap1 binding

	No. of nucleotide positions	Percentge of genome	No. of Tf1 integrations	Percentage of Tf1 integrations
Within Sap1-enriched regions	861,300	6.84	841,476	73.11
Outside Sap1-enriched regions	11,729,955	93.16	309,516	26.89
Total	12,591,255	100	1,150,992	100

In the rDNA repeats, Sap1 binds a direct repeat known as Ter1, where it initiates replication fork arrest and acts as a fork barrier (Krings and Bastia 2005, 2006; Mejia-Ramirez *et al.* 2005). In the mating locus, Sap1 binds SAS1, an inverted repeat, where it facilitates mating type switching (Arcangioli and Klar 1991, Krings and Bastia 2006). Examination of the Ter1 loci was challenging because these sites are within repetitive DNA located inside the rDNA genes (Mejia-Ramirez *et al.* 2005). While this makes it impossible to differentiate integration positions within specific Ter1 loci, the combined integration of all Ter1 sequences can be analyzed. As expected, both the SAS1 sequence in the mating loci and the Ter1 sequences in the rDNA repeats, as well as the surrounding regions, are enriched for Sap1 binding. However, very few Tf1 insertions were found in or near the SAS1 site, with the greatest number of insertions per position amounting to only 5, which is small considering that single-nucleotide positions elsewhere can have well over 1000 insertions (Figure 6A). By contrast, regions surrounding the Ter1 sequence contained high numbers of Tf1 insertions (Figure 6B). While relatively few insertions were within the Ter1 sequence itself, a single-nucleotide position containing over 1300 insertion events was found to be located 17 bp downstream of the Ter1 sequence 3′ end (Figure 6).

Tf1 insertion sites cluster at the Sap1-binding motif

Most Tf1 integration is located within genomic regions enriched for Sap1 binding. If Sap1 were directly responsible for positioning Tf1 integration, we would expect that integration would take place at specific nucleotide positions relative to the nucleotides bound by Sap1. The resolution of Sap1-binding sites provided by ChIP-Seq is not sufficient to determine the specific nucleotides bound by Sap1. To identify precise sites of Sap1 binding, we determined a motif sequence by processing Sap1 ChIP-Seq reads with the MEME Suite (Bailey *et al.* 2009). The resulting motif presented in Logo form was 21 bp and shared similarity with the Ter1 sequence (Figure 7A) (Krings and Bastia 2005, 2006) and also shares strong similarity with previously published Sap1-binding motifs (Zaratiegui *et al.* 2011). To identify positions within the *S. pombe* genome containing the Sap1-binding motif, the FIMO program of the MEME Suite (Grant *et al.* 2011) was used to perform genomic searches, and these identified 5013 locations that matched this motif. The alignment of all these motifs revealed that 82% of all integration events cluster within 1 kb of this motif (Figure 7B). Importantly, a large fraction of the integration events mapped to four positions, all of which occurred within 20 bp of the motif (Figure 7C). The two largest single-nucleotide hotspots of integration occur either 9 bp upstream or 19 bp downstream of the motif, demonstrating an asymmetrical pattern of integration around the motif (Figure 7D). In addition, the sites with fewer insertions form a sine wave pattern, which is more clearly defined downstream of the motif, with single-nucleotide peaks appearing at approximate 10-bp intervals. Of the insertions that do occur within the motif, most are located asymmetrically toward the 5′ end of the motif.

The serial number measures of integration analyzed earlier were assayed at 32° (Chatterjee *et al.* 2014). We also mapped Tf1 integration relative to the Sap1 motif using serial number data measured at 25° in the profiles reported here. In integration relative to the Sap1 motif, we saw no changes in the pattern from assays conducted at 32° with *sap1*$^+$ cells, at 25° with *sap1*$^+$ cells, or at 25° with *sap1-1* cells (Figure S4).

Sap1 binding to chromosomal DNA is not sufficient to mediate Tf1 integration

The finding that 73.1% of insertion events occurred at sites where Sap1 binding is enriched together with the substantial reduction in integration caused by the *sap1-1* mutation indicates that Sap1 promotes integration. We also evaluated whether Sap1 binding to DNA was sufficient to mediate integration. Inspection of the ChIP-Seq data of Sap1 identified some genomic sequences bound by high levels of Sap1 but that had little Tf1 integration (Figure 8). To quantitate what fraction of Sap1-enriched sequences contained Tf1 insertions, all peaks of Sap1 binding that were enriched two-fold or more were grouped based on the numbers of inserts they contained. Surprisingly, of a total of 7819 peaks of Sap1 enrichment, 5153 (66.0%) contained no integration events (Table 8), and 812 (10.4%) Sap1 peaks contained between 1 and 10 inserts, 483 (6.2%) peaks contained between 11 and 100 inserts, 1152 (14.7%) Sap1 peaks contained between 101 and 1000 inserts, and 219 (2.8%) Sap1 peaks contained 1001 or more Tf1 integration events. Combining the Sap1 peaks containing between 0 and 10 inserts constitutes a full 76.3% of all peaks, but these include just 0.2% of the integration events. The remaining 23.2% of the Sap1 peaks with 11 or more insertions contained 99.7% of all inserts present in Sap1-enriched sequences.

To determine what features distinguish the Sap1 peaks with high numbers of insertion events, we analyzed the size of the peaks. The Sap1-binding peaks with zero inserts averaged 42 nt in length and 55.4 (arbitrary units) in area, measures that are approximately two-fold smaller than the averages for peaks containing 1 to 10 insertions, suggesting that a threshold level of Sap1 binding is required to induce integration (Table 8). Similarly, the Sap1 peaks containing 11 to 100 inserts averaged 45.8 inserts per peak, a level 16-fold

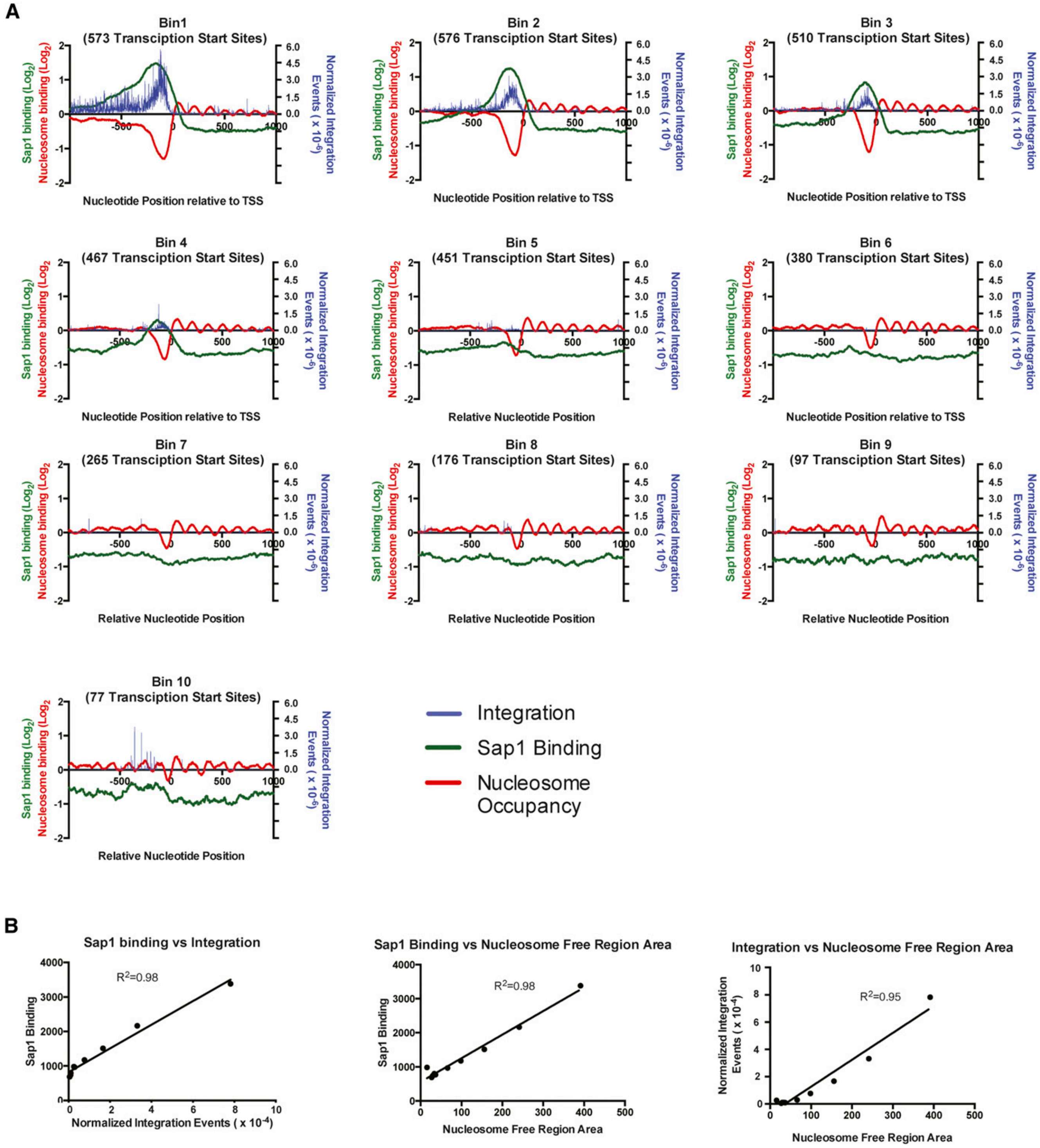

Figure 5 Tf1-insertion and Sap1-binding regions both align with NFRs at TSSs. (A) Graphs showing the alignment of all TSSs within 10 bins of intergenic regions from the *S. pombe* genome. The bins are sorted based on the number of Tf1 insertions occurring within the intergenic regions and are arranged in descending order. For each bin, the average number of Tf1 insertions and average value of Sap1 enrichment ($\log_2$ ratio of Sap1 binding to WCE signal) and nucleosome occupancy are plotted on the *y*-axis at each nucleotide position within 1000 bp of the aligned TSSs. For these analyses, the number of Tf1 insertions was normalized by dividing the amount of integration of each position by the total number of Tf1 insertions. (B) Graphs showing the pairwise comparisons and linear regression analysis of the sum of Tf1 integration, Sap1 enrichment, and area of NFRs that occur 1000 bp upstream of all TSSs within each bin. For these analyses, Sap1 enrichment values were back-transformed into their original non-$\log_2$ number, and only the negative values of nucleosome occupancy data were used to sum the total area of NFRs.

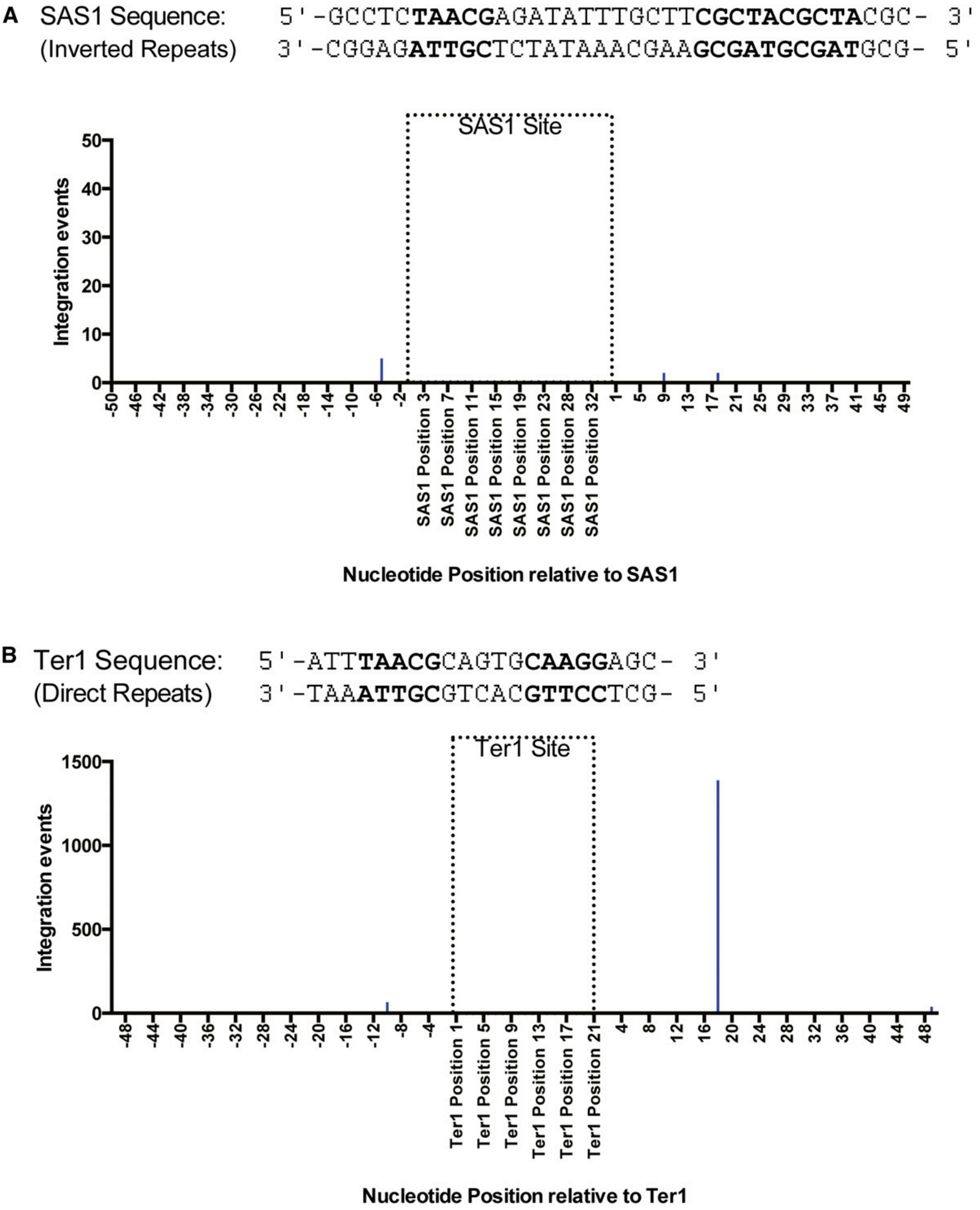

Figure 6 Tf1 insertions cluster around the Ter1 motif in rDNA repeats but not the SAS1 site in the mating locus. (A) Top: Sequence of both strands of the SAS1 element. Bottom: Graph showing the number of Tf1 insertions at single-nucleotide positions within 1000 pb of the SAS1 sequence. (B) Top: Sequence of both strands of the Ter1 element. Bottom: Graph showing the alignment of the three published Ter1 motifs (Pombase.com) and the tabulated number of Tf1 insertions present at single-nucleotide positions within 1000 bp of the aligned Ter1 motifs. The nucleotides in bold indicate nucleotides within core motifs of the Ter1 and Sas1 sequence, as identified by Krings and Bastia (2006).

greater than Sap1 peaks with 1 to 10 insertions. Importantly, the Sap1 peaks with 11 to 100 inserts had average lengths and areas approximately two-fold greater than the peaks with 1 to 10 insertions. This trend of increasing integration events with larger Sap1 peaks continued through the Sap1 peaks containing 1001 or more insertions, which averaged 1672 inserts per peak and had average lengths of 545 nt. These numbers indicate that Tf1 integration was associated with Sap1 peaks that had greater than a threshold length and area.

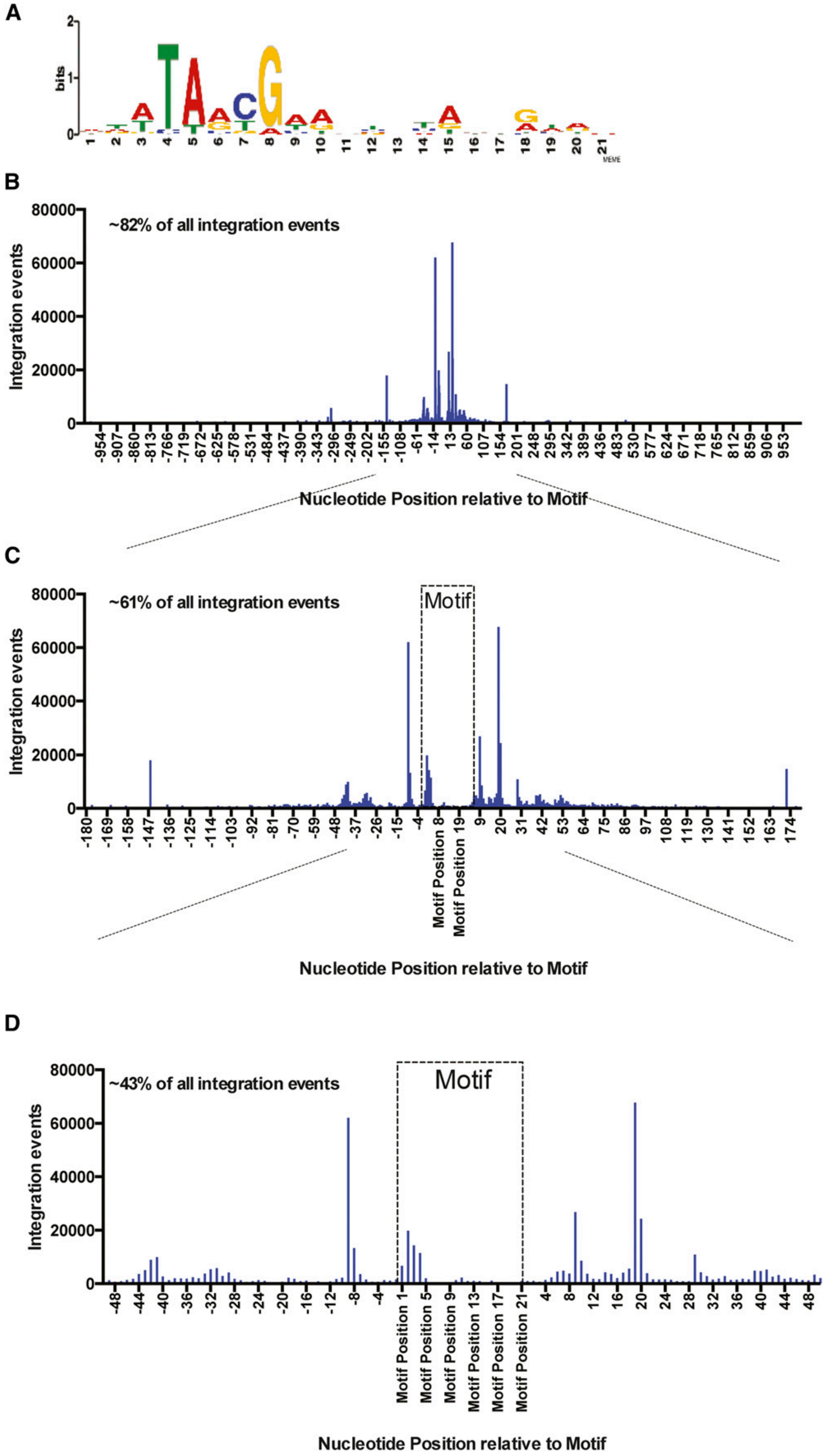

Figure 7 Tf1 insertions cluster at specific nucleotide positions around Sap1-binding motifs. (A) Logo of a Sap1 DNA-binding motif identified using the MEME Suite. (B) Graph showing the alignment of ∼5000 genomic Sap1 motifs that were identified using FIMO of the MEME Suite. The tabulated numbers of Tf1 insertions that occur at single-nucleotide positions within 1000 bp of the aligned motifs are plotted on the *y*-axis. Approximately 82% of all Tf1 insertion events occur within 1000 bp of a Sap1 motif. (C and D) Zoomed-in regions of the graph displayed in B showing the tabulated number of Tf1 insertions that occur at single-nucleotide positions within 180 (C) and 50 bp (D) of the aligned Sap1 motifs. Approximately 61% of all Tf1 insertions occur with 180 bp of a Sap1 motif, and approximately 43% of insertions occur within 50 bp of a Sap1 motif.

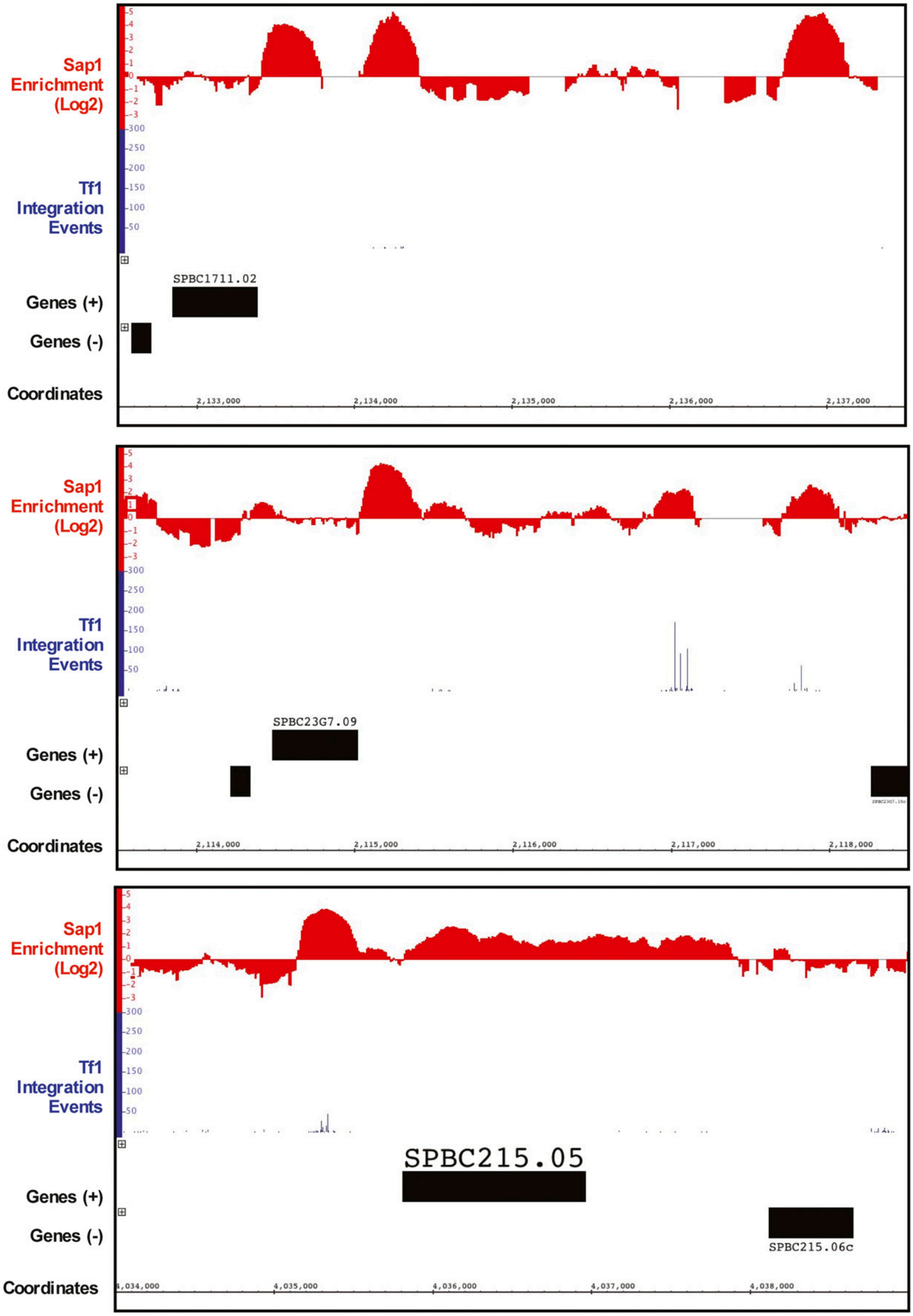
Sap1 Enrichment (Log2)
5
4
3
2
1
0
-1
-2
-3
-300
250
200
150
100
50
Tf1 Integration Events
SPBC1711.02
Genes (+)
Genes (-)
Coordinates
2,133,000
2,134,000
2,135,000
2,136,000
2,137,000
Sap1 Enrichment (Log2)
Tf1 Integration Events
SPBC23G7.09
Genes (+)
Genes (-)
SPBC23G7.10c
Coordinates
2,114,000
2,115,000
2,116,000
2,117,000
2,118,000
Sap1 Enrichment (Log2)
Tf1 Integration Events
SPBC215.05
Genes (+)
Genes (-)
SPBC215.06c
Coordinates
4,034,000
4,035,000
4,036,000
4,037,000
4,038,000

Table 8 Sorting of Sap1 peaks based on the number of Tf1 integrations per peak

Part A

Integrations per Sap1 peak	No. of Sap1 peaks	Percentage of Sap1 peaks	No. of integrations	Percent of integrations in Sap1-enriched regions	Percent of total integrations in the genome	Average number of integrations per peak
0	5153	65.90	0	0.00	0.00	0.00
1–10	812	10.38	2,311	0.27	0.20	2.85
11–100	483	6.18	22,048	2.62	1.92	45.79
101–1000	1152	14.73	450,977	53.59	39.18	391.47
1001 and up	219	2.80	366,140	43.51	31.81	1671.87
Total	7819	100.00	841476	100.00	73.11	NA

Part B

Integrations per Sap1 peak	Average peak length (nt)	Total peak length (nt)	Percentage of peak length	Average peak area	Total peak area	Percentage of peak area
0	41.97	216,270	25.11	55.44	285,685.52	16.36
1–10	93.12	75,370	8.75	136.56	110,887.00	6.35
11–100	178.43	86,180	10.01	332.44	160,568.53	9.20
101–1000	315.58	361,940	42.02	733.69	840,509.96	48.15
1001 and up	554.98	121,540	14.11	1589.67	348,137.61	19.94
Total	NA	861,300	100	NA	1,745,788.616	100

Discussion

Our data demonstrate that Sap1 is important for Tf1 transposition and indicate that Sap1 plays a role in the integration of the element. This conclusion is based on the substantial decrease in Tf1 transposition in *sap1-1* mutant cells and the presence of transposition intermediates in these cells (Figure 1).

The alignment of Sap1-binding sites with Tf1 inserts is consistent with a role of Sap1 in integration (Zaratiegui *et al.* 2011) (Figure 4A). Further, the highly quantitative measures of integration provided by the serial number system reveal that Sap1-enriched DNA sequences accounted for 73.4% of all integration events. Importantly, serial number data showed that integration positions with high numbers of independent insertions were much more likely to coincide with Sap1-binding sites than positions with low numbers of inserts, indicating that Sap1 contributes to the efficiency of integration (Figure 4C). Our finding that the *sap1-1* mutation greatly reduced integration frequency without causing significant changes in the location of integration also argues that Sap1 promotes the efficiency of integration. One possible contribution Sap1 may make to integration efficiency is nucleosome occlusion. The striking correlation between Sap1 binding, NFR size, and integration at promoters is consistent with a function of Sap1 that excludes nucleosomes and allows the integration complex access to target sites. This model is supported by the finding that a mutation in *sap1* causes increased nucleosome occupancies at the promoters that Sap1 binds (Tsankov *et al.* 2011).

Sap1 also may be playing a direct role in integration by positioning insertion sites. In this case, the *sap1-1* mutation would be expected to alter the profile of integration sites. However, the *sap1-1* mutation did not cause substantial alterations in the profile. The *sap1-1* mutation is in domain IV, a region important for dimerization (Noguchi and Noguchi 2007; Ghazvini *et al.* 1995). Therefore, the mutation may reduce levels of Sap1 dimer bound to DNA and not the positions where it binds. This would explain why the mutation did not cause large changes in the integration profile. The highly specific clustering of integration events at positions +19 and −9 bp relative to the Sap1-binding motif provides strong support for the model that Sap1 is responsible for positioning integration events (Figure 7C).

The clustering of inserts at single-nucleotide sites flanking the Sap1 motif would be expected to occur if Sap1 covers its binding site on the DNA and directs integration to sites on either side of the protein. Sap1 binds as a head-to-tail dimer to two direct repeats in the Ter1 sequence. On binding this sequence, the Sap1 dimer bends the DNA and distorts the DNA helix, a process that is required for Sap1-mediated replication fork arrest (Bada *et al.* 2000; Krings and Bastia 2005; 2006). If Sap1 molecules were to bind the motif sequence near insertion sites in the same manner, IN interaction with equivalent regions of each head-to-tail dimer could explain the asymmetry in the single-nucleotide clusters flanking the Sap1 motif. Thus far we have been unable to demonstrate an interaction between Sap1 and Tf1 IN with pull-down assays. This could be because post-translational modifications of IN or Sap1 are required. But the *in vivo* two-hybrid assay provided support for a Sap1:IN interaction that could directly position integration by a tethering mechanism similar to those of HIV-1, Ty1,

Figure 8 Sap1 binding is necessary but not sufficient for Tf1 integration. These regions of the *S. pombe* genome show large Sap1 peaks containing few to no Tf1 insertions. Note that the top two panels shows tall Sap1-binding peaks that have no insertion events, while the bottom panels shows a broad Sap1-binding peak with no insertion events.

Ty3, and Ty5. Finally, a third possibility that must be considered is that an additional host factor interacts with both Sap1 and Tf1 IN/cDNA and forms a bridge in the complex that mediates integration.

The sine wave pattern of integration near the Sap1 motif could be indicative of integration occurring as a result of helical distortion induced by Sap1 binding. The peaks of integration within these sine waves occur in ~10-bp intervals, the approximate length of a DNA helical turn (Wang *et al.* 1979). Perhaps the binding of a Sap1 dimer to the DNA results in conformational changes at regular intervals along the DNA helix, which makes certain regions of each helical turn more amenable to Tf1 insertion.

Not all Sap1-binding sites appear to be equal in terms of efficiency of Tf1 insertion, which suggests that Sap1 binding to genomic target sites is important but not sufficient for integration. Fully 25.1% of the genomic sequences enriched for Sap1 binding (total peak length) have no integration events (Table 8). These Sap1-enriched peaks lacking integration indicate that Sap1 binding is not the only feature needed to mediate integration. Conversely, not all integration occurs in Sap1-enriched peaks. Twenty-six percent of the Tf1 insertions occur outside the sequences enriched twofold for Sap1 binding. Even if all sequences with any level of detectable Sap1 enrichment are considered, 13% of the insertions fall outside Sap1-bound peaks (Figure S3 and Table S2). While these data argue for a Sap1-independent mechanism of integration, it is formally possible that the differences in the culture conditions used for the ChIP-Seq experiments compared to the transposition cultures could account for the integration that occurred outside Sap1-binding sites. Additionally, insertion sites with single integration events were not enriched in Sap1-bound sequences, suggesting that these single insertion sites may represent random events that also use a Sap1-independent mechanism.

The integration of Tf1 is known to recognize the promoters of stress-response genes (Guo and Levin 2010; Chatterjee *et al.* 2014). The binding of Sap1 to targets of Tf1 integration raises the possibility that Sap1 may play a role in responding to environmental stress. Sap1 is known to be able to interrupt the migration of replication forks (Arcangioli and Klar 1991; Krings and Bastia 2005, 2006; Mejia-Ramirez *et al.* 2005; Noguchi and Noguchi 2007). Future studies of Sap1 will be needed to determine whether its ability to block replication forks functions in response to stress and whether its role in Tf1 integration is regulated by stress.

Acknowledgments

We thank Amnon Hizi for purifying Sap1 protein, conducting pull-down experiments, and providing helpful discussions. We thank Yabin Guo for the use of his computer programs and scripts in the bioinformatic analysis of our data. We thank Chanan Rubin for his helpful contributions. This research was supported by the Intramural Research Programs of the National Institutes of Health, the Eunice Kennedy Shriver National Institute of Child Health and Human Development (H.L.L.), and the National Cancer Institute (S.I.S.G.). P.G.M. was supported by the Intramural Research Program of the National Institutes of Health, Center for Information Technology. This study used the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov).

Literature Cited

Arcangioli, B., T. D. Copeland, and A. J. Klar, 1994 Sap1, a protein that binds to sequences required for mating-type switching, is essential for viability in Schizosaccharomyces pombe. Mol. Cell. Biol. 14: 2058–2065.

Arcangioli, B., and A. J. Klar, 1991 A novel switch-activating site (SAS1) and its cognate binding factor (SAP1) required for efficient mat1 switching in Schizosaccharomyces pombe. EMBO J. 10: 3025–3032.

Atwood, A., J. Choi, and H. L. Levin, 1998 The application of a homologous recombination assay revealed amino acid residues in an LTR-retrotransposon that were critical for integration. J. Virol. 72: 1324–1333.

Atwood, A., J. H. Lin, and H. L. Levin, 1996 The retrotransposon Tf1 assembles virus-like particles that contain excess Gag relative to integrase because of a regulated degradation process. Mol. Cell. Biol. 16: 338–346.

Atwood-Moore, A., K. Yan, R. L. Judson, and H. L. Levin, 2006 The self primer of the long terminal repeat retrotransposon Tf1 is not removed during reverse transcription. J. Virol. 80: 8267–8270.

Bada, M., D. Walther, B. Arcangioli, S. Doniach, and M. Delarue, 2000 Solution structural studies and low-resolution model of the Schizosaccharomyces pombe Sap1 protein. J. Mol. Biol. 300: 563–574.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant *et al.*, 2009 MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37: W202–208.

Behrens, R., J. Hayles, and P. Nurse, 2000 Fission yeast retrotransposon Tf1 integration is targeted to 5' ends of open reading frames. Nucleic Acids Res. 28: 4709–4716.

Boeke, J. D., J. Trueheart, G. Natsoulis, and G. R. Fink, 1987 5-Fluoro-orotic acid as a selective agent in yeast molecular genetics. Methods Enzymol. 154: 164–175.

Bowen, N. J., I. Jordan, J. Epstein, V. Wood, and H. L. Levin, 2003 Retrotransposons and their recognition of pol II promoters: a comprehensive survey of the transposable elements derived from the complete genome sequence of Schizosaccharomyces pombe. Genome Res. 13: 1984–1997.

Bridier-Nahmias, A., A. Tchalikian-Cosson, J. A. Baller, R. Menouni, H. Fayol *et al.*, 2015 Science 348(6234): 585–8.

Chalker, D. L., and S. B. Sandmeyer, 1990 Transfer RNA genes are genomic targets for *de novo* transposition of Ty3. Genetics 126: 837–850.

Chalker, D. L., and S. B. Sandmeyer, 1992 Ty3 integrates within the region of RNA polymerase III transcription initiation. Genes Dev. 6: 117–128.

Chatterjee, A. G., C. Esnault, Y. Guo, S. Hung, P. G. McQueen *et al.*, 2014 Serial number tagging reveals a prominent sequence preference of retrotransposon integration. Nucleic Acids Res. 42: 8449–8460.

Ciuffi, A., M. Llano, E. Poeschla, C. Hoffmann, and J. Leipzig *et al.*, 2005 A role for LEDGF/p75 in targeting HIV DNA integration. Nat Med. 11: 1287–1289.

de Castro, E., I. Soriano, L. Marin, R. Serrano, L. Quintales *et al.*, 2012 Nucleosomal organization of replication origins and meiotic recombination hotspots in fission yeast. EMBO J. 31: 124–137.

Devine, S. E., and J. D. Boeke, 1996 Integration of the yeast retrotransposon Ty1 is targeted to regions upstream of genes transcribed by RNA polymerase III. Genes Dev. 10: 620–633.

Durfee, T., K. Becherer, P. L. Chen, S. H. Yeh, Y. Yang *et al.*, 1993 The retinoblastoma protein associates with the protein phosphatase type 1 catalytic subunit. Genes Dev. 7(4): 555–69.

Feng, G., Y. E. Leem, and H. L. Levin, 2013 Transposon integration enhances expression of stress response genes. Nucleic Acids Res. 41: 775–789.

Forsburg, S. L., and N. Rhind, 2006 Basic methods for fission yeast. Yeast 23: 173–183.

Gai, X., and D. F. Voytas, 1998 A single amino acid change in the yeast retrotransposon Ty5 abolishes targeting to silent chromatin. Mol. Cell 1: 1051–1055.

Ghazvini, M., V. Ribes, and B. Arcangioli, 1995 The essential DNA-binding protein sap1 of Schizosccharomyces pombe contains two independent oligomerization interfaces that dictate the relative orientation of the DNA-binding domain. Mol. Cell. Biol. 15: 4939–4946.

Grant, C. E., T. L. Bailey, and W. S. Noble, 2011 FIMO: scanning for occurrences of a given motif. Bioinformatics 27: 1017–1018.

Guo, Y., and H. L. Levin, 2010 High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in Schizosaccharomyces pombe. Genome Res. 20: 239–248.

Gupta, S. S., T. Maetzig, G. N. Maertens, A. Sharif, M. Rothe *et al.*, 2013 Bromo and ET domain (BET) chromatin regulators serve as co-factors for murine leukemia virus integration. J. Virol. 87: 12721–12736.

Hoff, E. F., H. L. Levin, and J. D. Boeke, 1998 Schizosaccharomyces pombe retrotransposon Tf2 mobilizes primarily through homologous cDNA recombination. Mol. Cell. Biol. 18: 6839–6852.

Ji, H., D. P. Moore, M. A. Blomberg, L. T. Braiterman, D. F. Voytas *et al.*, 1993 Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. Cell 73: 1007–1018.

Kirchner, J., C. M. Connolly, and S. B. Sandmeyer, 1995 Requirement of RNA polymerase III transcription factors for in vitro position-specific integration of a retroviruslike element [see comments]. Science 267: 1488–1491.

Krings, G., and D. Bastia, 2005 Sap1p binds to Ter1 at the ribosomal DNA of Schizosaccharomyces pombe and causes polar replication fork arrest. J. Biol. Chem. 280: 39135–39142.

Krings, G., and D. Bastia, 2006 Molecular architecture of a eukaryotic DNA replication terminus-terminator protein complex. Mol. Cell. Biol. 26: 8061–8074.

Leem, Y. E., T. L. Ripmaster, F. D. Kelly, H. Ebina, M. E. Heincelman *et al.*, 2008 Retrotransposon Tf1 is targeted to pol II promoters by transcription activators. Mol. Cell 30: 98–107.

Lesage, P., and A. L. Todeschini, 2005 Happy together: the life and times of Ty retrotransposons and their hosts. Cytogenet. Genome Res. 110: 70–90.

Levin, H. L., 1995 A novel mechanism of self-primed reverse transcription defines a new family of retroelements. Mol. Cell. Biol. 15: 3310–3317.

Levin, H. L., 1996 An unusual mechanism of self-primed reverse transcription requires the RNase H domain of reverse transcriptase to cleave an RNA duplex. Mol. Cell. Biol. 16: 5645–5654.

Levin, H. L., and J. V. Moran, 2011 Dynamic interactions between transposable elements and their hosts. Nat. Rev. Genet. 12: 615–627.

Levin, H. L., D. C. Weaver, and J. D. Boeke, 1993 Novel gene expression mechanism in a fission yeast retroelement: Tf1 proteins are derived from a single primary translation product. EMBO J. 12: 4885–4895 [erratum: EMBO J 13:1494 (1994)].

Llano, M., M. Vanegas, N. Hutchins, D. Thompson, S. Delgado *et al.*, 2006 Identification and characterization of the chromatin-binding domains of the HIV-1 integrase interactor LEDGF/p75. J. Mol. Biol. 360: 760–773.

Majumdar, A., A. G. Chatterjee, T. L. Ripmaster, and H. L. Levin, 2011 The determinants that specify the integration pattern of retrotransposon Tf1 in the fbp1 promoter of Schizosaccharomyces pombe. J. Virol. 85: 519–529.

Mejia-Ramirez, E., A. Sanchez-Gorostiaga, D. B. Krimer, J. B. Schvartzman, and P. Hernandez, 2005 The mating type switch-activating protein Sap1 Is required for replication fork arrest at the rRNA genes of fission yeast. Mol. Cell. Biol. 25: 8755–8761.

Miller, J., 1972 *Experiments in Molecular Genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Moore, S. P., G. Liti, K. M. Stefanisko, K. M. Nyswaner, C. Chang *et al.*, 2004 Analysis of a Ty1-less variant of Saccharomyces paradoxus: the gain and loss of Ty1 elements. Yeast 21: 649–660.

Noguchi, C., and E. Noguchi, 2007 Sap1 promotes the association of the replication fork protection complex with chromatin and is involved in the replication checkpoint in Schizosaccharomyces pombe. Genetics 175: 553–566.

Qi, X., and S. Sandmeyer, 2012 In vitro targeting of strand transfer by the Ty3 retroelement integrase. J. Biol. Chem. 287: 18589–18595.

Sandmeyer, S., 2003 Integration by design. Proc. Natl. Acad. Sci. USA 100: 5586–5588.

Scheifele, L. Z., G. J. Cost, M. L. Zupancic, E. M. Caputo, and J. D. Boeke, 2009 Retrotransposon overdose and genome integrity. Proc. Natl. Acad. Sci. USA 106: 13927–13932.

Sharma, A., R. C. Larue, M. R. Plumb, N. Malani, F. Male *et al.*, 2013 BET proteins promote efficient murine leukemia virus integration at transcription start sites. Proc. Natl. Acad. Sci. USA 110: 12036–12041.

Shun, M. C., N. K. Raghavendra, N. Vandegraaff, J. E. Daigle, S. Hughes *et al.*, 2007 LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. Genes Dev. 21: 1767–1778.

Singleton, T. L., and H. L. Levin, 2002 A long terminal repeat retrotransposon of fission yeast has strong preferences for specific sites of insertion. Eukaryot. Cell 1: 44–55.

Studamire, B., and S. P. Goff, 2008 Host proteins interacting with the Moloney murine leukemia virus integrase: multiple transcriptional regulators and chromatin binding factors. Retrovirology 5: 48.

Teysset, L., V. D. Dang, M. K. Kim, and H. L. Levin, 2003 A long terminal repeat-containing retrotransposon of Schizosaccharomyces pombe expresses a Gag-like protein that assembles into virus-like particles which mediate reverse transcription. J. Virol. 77: 5451–5463.

Tsankov, A., Y. Yanagisawa, N. Rhind, A. Regev, and O. J. Rando, 2011 Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. Genome Res. 21: 1851–1862.

Wang, J. C., 1979 Helical repeat of DNA in solution. Proc. Natl. Acad. Sci. USA. 76: 200–203.

Xie, W., X. Gai, Y. Zhu, D. C. Zappulla, R. Sternglanz *et al.*, 2001 Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. Mol. Cell. Biol. 21: 6606–6614.

Yieh, L., H. Hatzis, G. Kassavetis, and S. B. Sandmeyer, 2002 Mutational analysis of the transcription factor IIIB-DNA target of Ty3 retroelement integration. J. Biol. Chem. 277: 25920–25928.

Yieh, L., G. Kassavetis, E. P. Geiduschek, and S. B. Sandmeyer, 2000 The Brf and TATA-binding protein subunits of the RNA polymerase III transcription factor IIIB mediate position-specific integration of the gypsy-like element, Ty3. J. Biol. Chem. 275: 29800–29807.

Zaratiegui, M., M. W. Vaughn, D. V. Irvine, D. Goto, S. Watt *et al.*, 2011 CENP-B preserves genome integrity at replication forks paused by retrotransposon LTR. Nature 469: 112–115.

Zhu, Y. X., S. G. Zou, D. A. Wright, and D. F. Voytas, 1999 Tagging chromatin with retrotransposons: target specificity of the Saccharomyces Ty5 retrotransposon changes with the chromosomal localization of Sir3p and Sir4p. Genes Dev. 13: 2738–2749.

Zou, S., N. Ke, J. M. Kim, and D. F. Voytas, 1996 The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. Genes Dev. 10: 634–645.

Zou, S., and D. F. Voytas, 1997 Silent chromatin determines target preference of the Saccharomyces retrotransposon Ty5. Proc. Natl. Acad. Sci. USA 94: 7412–7416.

Communicating editor: D. Voytas

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.181602/-/DC1

Single-Nucleotide-Specific Targeting of the Tf1 Retrotransposon Promoted by the DNA-Binding Protein Sap1 of *Schizosaccharomyces pombe*

Anthony Hickey, Caroline Esnault, Anasuya Majumdar, Atreyi Ghatak Chatterjee, James R. Iben, Philip G. McQueen, Andrew X. Yang, Takeshi Mizuguchi, Shiv I. S. Grewal, and Henry L. Levin

DOI: 10.1534/genetics.115.181602

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	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACGTCTCACCGCAGTTGATGCATAGGAAGC	Tf1-LTR primers with barcodes (*sap1*$^+$ (25°C) #1)
HL3512	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCACTCACCGCAGTTGATGCATAGGAAGC	Tf1-LTR primers with barcodes (*sap1*$^+$ (25°C) #2)
HL3513	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTACCTCACCGCAGTTGATGCATAGGAAGC	Tf1-LTR primers with barcodes (*sap1*$^+$ (25°C) #3)
HL3514	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCATGCTCACCGCAGTTGATGCATAGGAAGC	Tf1-LTR primers with barcodes (*sap1-1* (25°C) #1)
HL3520	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTCCTCACCGCAGTTGATGCATAGGAAGC	Tf1-LTR primers with barcodes (*sap1-1* (25°C) #2)
HL3521	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGACTCTCACCGCAGTTGATGCATAGGAAGC	Tf1-LTR primers with barcodes (*sap1-1* (25°C) #3)

C

Sequence of PCR Primer		
Number	Sequence	Description
HL2216	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGTAATACGACTCACTATAGGGC	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°C) #1: AGCT, *sap1*$^+$ (25°C) #2: TGCA, *sap1*$^+$ (25°C) #3: GTAC, *sap1-1* (25°C) #1: CATG, *sap1-1* (25°C) #2: AGTC, and *sap1-1* (25°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).

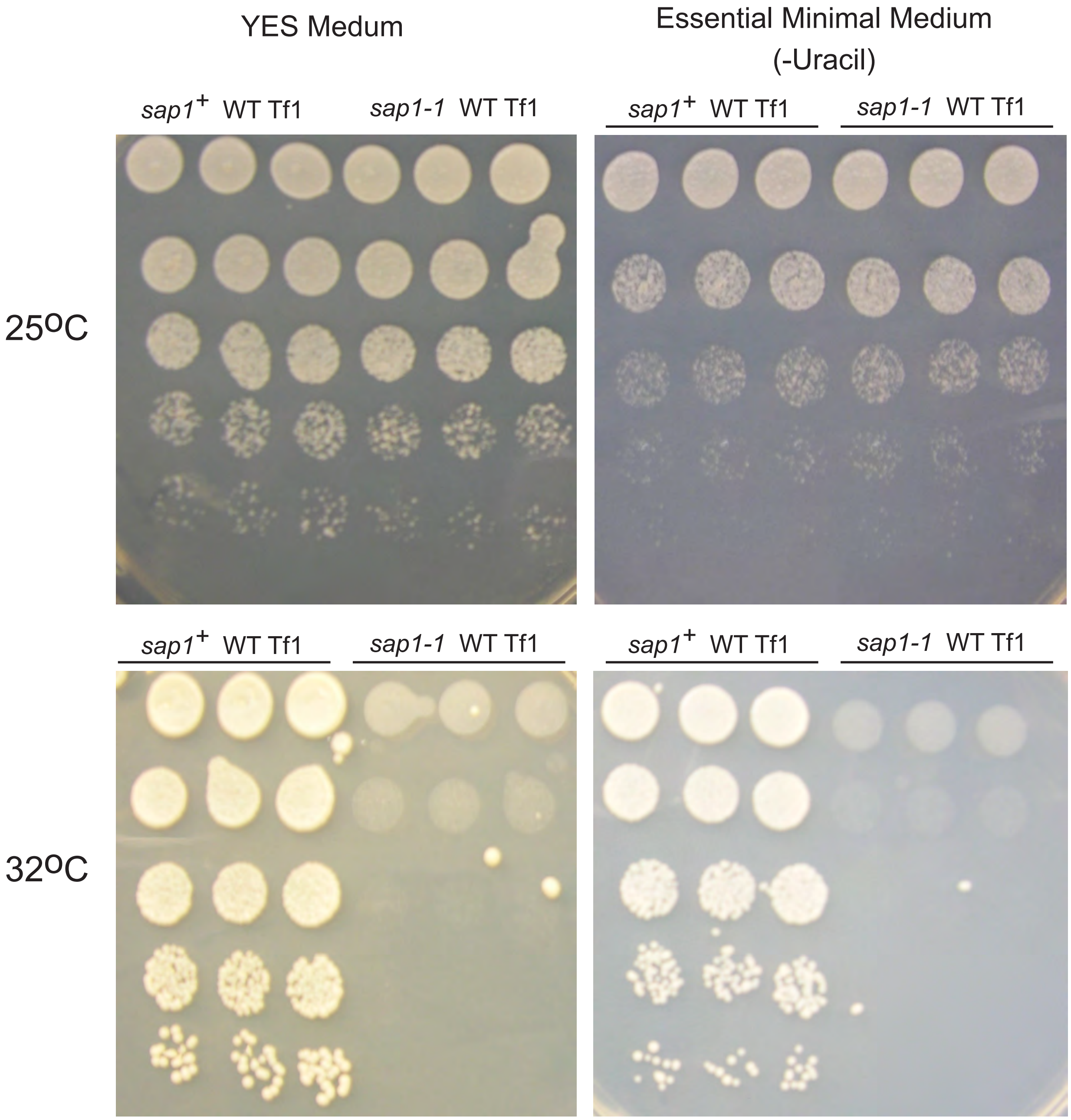

Figure S2. The *sap1-1* mutation causes no apparent growth defect at 25^{O}C. Five fold serial dilutions of *sap*$^{+}$ and *sap1-1 S. pombe* transformed with plasmids expressing WT Tf1 were spotted onto both solid YES and EMM –uracil media, and were grown at the indicated temperatures for 5 days. Three 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.

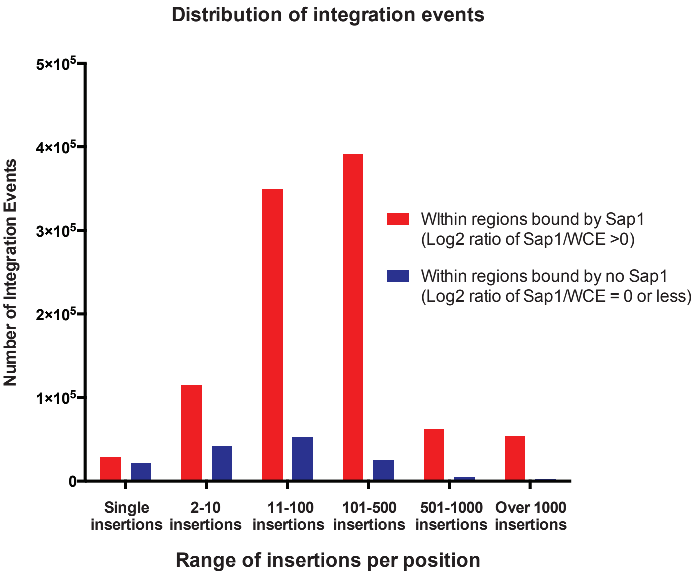

Figure S3. The majority of Tf1 integration occur in regions of the genome that are bound by Sap1. Graph showing the number of Tf1 insertions that occur within the indicated ranges of insertions per positions, as well as their occurrence in- and outside Sap1-binding regions in the genome.

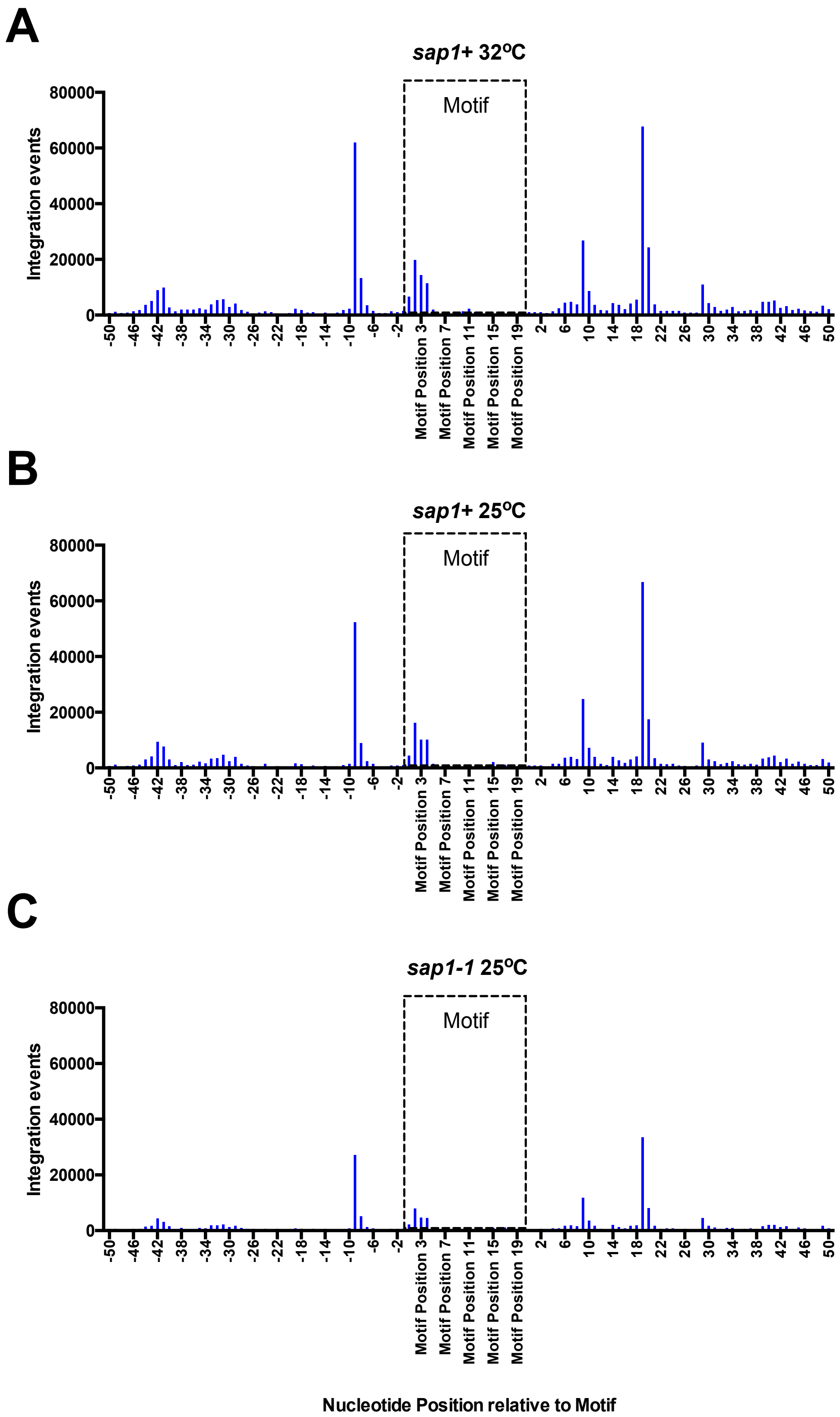
A
sap1+ 32°C
Motif
Integration events
80000
60000
40000
20000
0
-50
-46
-42
-38
-34
-30
-26
-22
-18
-14
-10
-6
-2
Motif Position 3
Motif Position 7
Motif Position 11
Motif Position 15
Motif Position 19
2
6
10
14
18
22
26
30
34
38
42
46
50
B
sap1+ 25°C
Motif
Integration events
80000
60000
40000
20000
0
-50
-46
-42
-38
-34
-30
-26
-22
-18
-14
-10
-6
-2
Motif Position 3
Motif Position 7
Motif Position 11
Motif Position 15
Motif Position 19
2
6
10
14
18
22
26
30
34
38
42
46
50
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
Motif Position 3
Motif Position 7
Motif Position 11
Motif Position 15
Motif Position 19
2
6
10
14
18
22
26
30
34
38
42
46
50
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) *sap1*$^{+}$ cells grown at 25^{O}C and C) *sap1-1* cells grown at 25^{O}C.

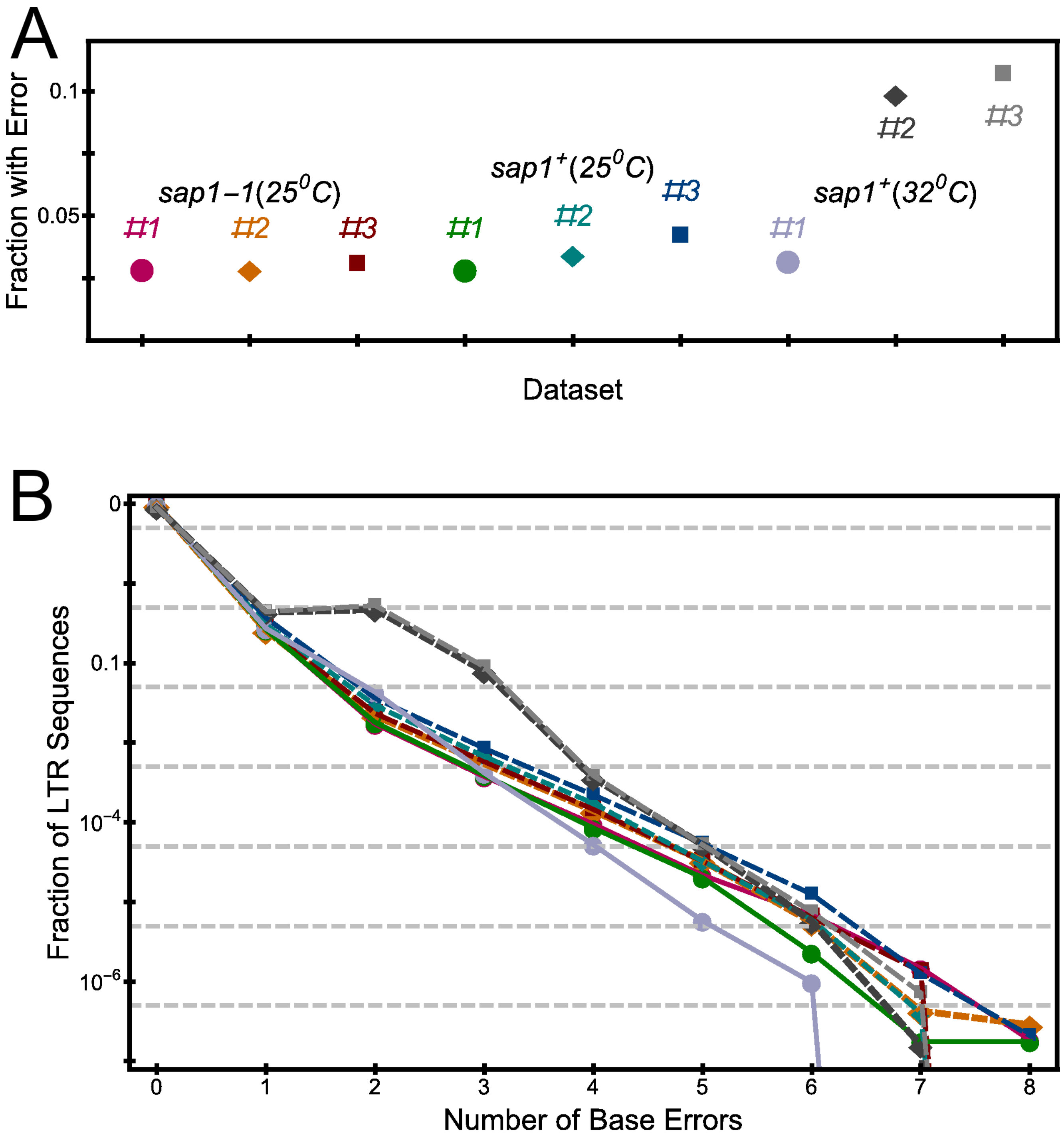

Figure S5. Error rate in reading the eight base pair serial number sequence. (A) Fraction of eight base pair sequences mis-read for each of the nine datasets. (B). Proportion of serial sequence reads with a given number of base mis-matches for each of the nine datasets. The symbols and colors for each database match those in (A). The dashed gray lines show fraction = 5x10^{-n}, n = 1, 2, 3, 4, 5, 6.

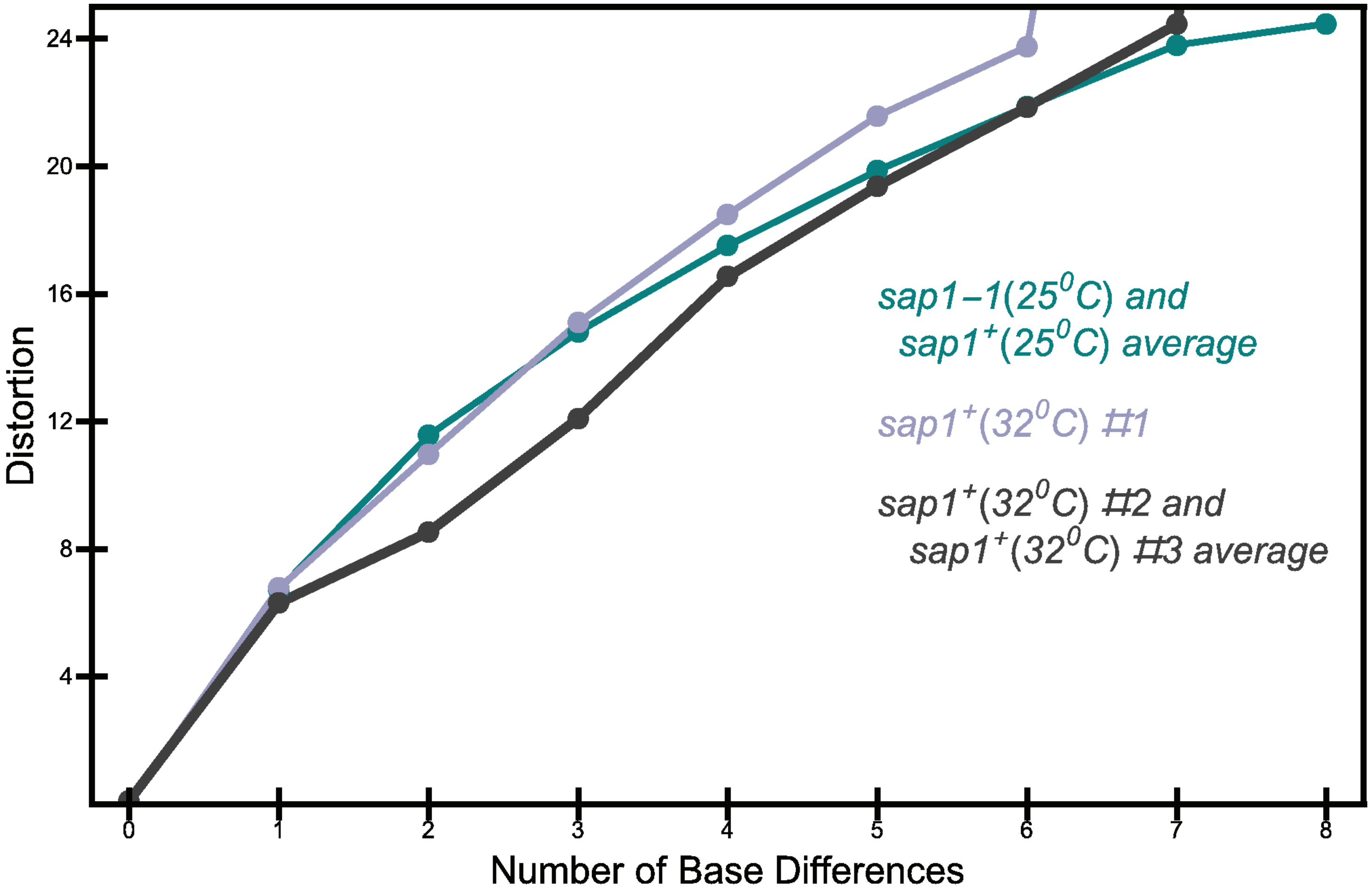

Figure S6. The distortion function between two serial numbers as a function of the number of base differences between them.

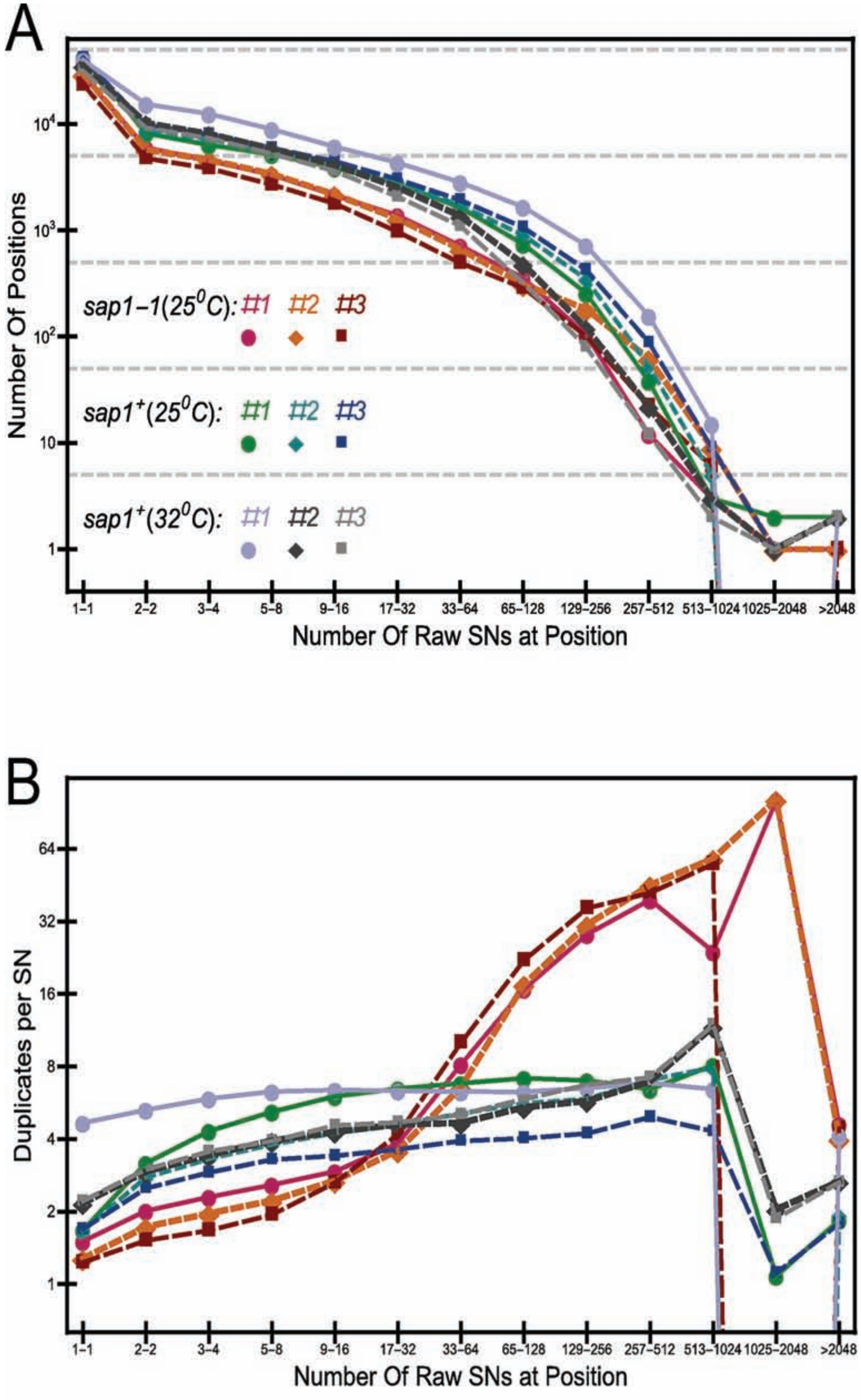

Figure S7. (A) Number of positions in each of the datasets versus the number of raw serial number at the position. The dashed gray lines show fraction = $5x10^n$, n = 1, 2, 3, 4. (B) Number of average duplicates per serial number at a position versus the number of raw serial numbers at the position.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

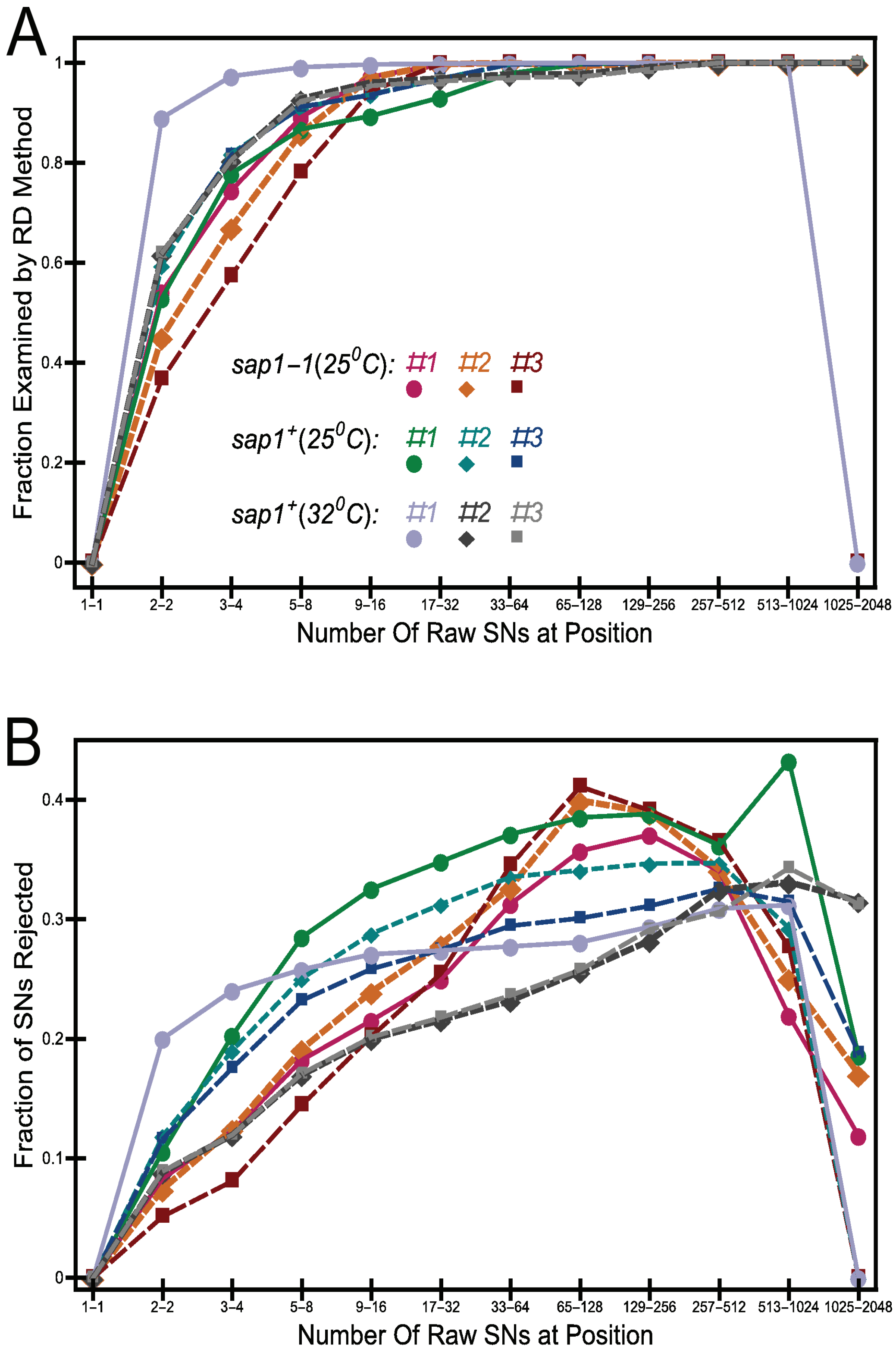

Figure S8. (A) Proportion of positions that could be analyzed by the rate distortion method versus the number of raw serial numbers at the position. (B) Average proportion of serial numbers judged as not real at a position by the rate distortion method versus the number of raw serial numbers at the position. 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 *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

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

The rate distortion procedure is to chose probabilities P(*Y*j|*X*i) 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

$$Z(X_\mathrm{i},T) = \Sigma_{\ \mathrm{j}=1,N_\mathrm{Y}}\ \mathrm{Pc}(Y_\mathrm{j})\ \exp(-\log(2)\ d(X_\mathrm{i},Y_\mathrm{j})\ /T\ ) \qquad \text{(Eq. 5)}$$

Since the $\mathrm{Pc}(Y_\mathrm{j})$ are dependent on the $P(Y_\mathrm{j}|X_\mathrm{i})$ variables, this solution for the $P(Y_\mathrm{j}|X_\mathrm{i})$ 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, $\mathbf{S} = \{S_1, S_2, \ldots S_\mathrm{N}\}$. One would like to estimate the probability $P(S_\mathrm{j}|S_\mathrm{i})$ 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 $P(S_\mathrm{j}|S_\mathrm{i})$'s are a stochastic mapping of **S** back onto itself, the rate distortion algorithm can be applied to find the $P(S_\mathrm{j}|S_\mathrm{i})$ 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

$$\mathrm{Pr}(SN_\mathrm{i}) = dup_\mathrm{i}/N_\mathrm{dup} \qquad \text{(Eq. 6)}$$

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

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

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

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

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 $\{P(S_j|S_i)\}$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 $P(S_x|S_j)$ 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

$$<dup_i> = \Sigma_j \, Pmr(m_{ij}) \, dup_j' \qquad \text{(Eq. 10)}$$

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:

$$G = \Sigma_i \, dup_i \ln(dup_i/<dup_i>) \qquad \text{(Eq. 11)}$$

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 $Pmr(0)*N_{\mathrm{dup}}$, 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 $Pmr(0)*N_{\mathrm{dup}}$. 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 $Pmr(0)*N_{\text{dup}}$.
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^{0}C) 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.