Supporting Information

Gangadharan et al. 10.1073/pnas.1016382107

SI Materials and Methods

Experimental Methods. ***Genetic selection system for isolating Hermes transposon integrants.*** To recover in vivo integrations, we first constructed a yeast *ARS CEN* donor plasmid, pSG36, containing a *URA3* marker, a *Hermes-NatMX* transposon and a *GALs* promoter-regulated *Hermes* transposase gene expressing a hyperactive G366W/M286T mutant form of the *Hermes* transposase protein. The *NatMX* cassette that confers resistance to antibiotic clonNAT is flanked by *Hermes* transposon terminal inverted repeats (TIRs), 711 bp of the left L-end TIR and 512 bp of the right R-end TIR. Because of the inherent low in vivo activity of the WT enzyme (transposition frequency ≈ 2.5×10^{-5}) it proved technically difficult to obtain a large set of insertions. The number of insertions recovered improved considerably with the use of a hyperactive form of the enzyme G366W/M284T (transposition frequency ≈ 6×10^{-2}). We sequenced >80 terminal junctions of insertions testing both the wild-type and hyperactive versions of the *Hermes* transposase by 454 sequencing and they have the same nTnnnnAn target site sequence.

For measuring transposition frequency, yeast strain BY4727 (*MATα his3Δ200, leu2Δ0, lys2Δ0, met15Δ0 trp1Δ63 ura3Δ0*) was transformed with the transposon donor plasmids by using PEG/Lithium acetate method. Colonies were streaked onto Synthetic Complete media lacking Uracil (SC-Ura) plates containing 2% galactose for transposase induction and incubated for 5 d at 30° C. Single colonies were then resuspended in water, serially diluted, and plated on SC+5-FOA (1 mg/mL) to determine the total number of plasmid-free cells and on SC+5-FOA+ClonNAT (100 μg/mL) to determine the number of integrants. The frequency of transposition is the ratio of the number of colonies on SC+5-FOA+ClonNAT plates to the total number of cells on SC+5-FOA.

Cell culture and selection of integrants for transposition profiling. We recovered insertions by growing liquid cultures of yeast cells in galactose-containing media in log phase by serially passaging cells over a period of 5 d. Briefly, three independent transformants were picked to inoculate 5 mL of liquid SC-Ura containing 2% glucose and grown at 30 °C 16 h overnight. Each of these cultures were induced for transposition in liquid SC-Ura medium containing 2% galactose, and cell density measurements were taken before and after growth to calculate the number of generations. First, the overnight starter culture in liquid SC-Ura medium containing 2% glucose was washed and added to a 50-mL culture in SC-Ura+2% galactose so that the initial OD_{600} of the culture was 0.05. After cultures reached an OD_{600} of 5, they were diluted and new cultures were started. This protocol prevented the cells from entering stationary phase. In all, a series of five sequential cultures were used for each strain. The final 50-mL cultures were resuspended in 500 mL of Sc+5-FOA to a final OD of 0.25 to select against the donor plasmid. These cells were grown for 20–24 h. Next, the SC+5-FOA cultures were diluted down to OD_{600} of 0.5 in 500 mL of SC+5-FOA+ClonNAT and incubated to isolate the cells with insertions. The cultures were grown for 24 h to an OD_{600} of no higher than 5.0. Cells containing the donor plasmid are already resistant to ClonNAT because of the plasmid. Thus, only cells lacking the *URA3*-containing donor plasmid with *NatMX* (ClonNAT resistance gene cassette) are selected for the resistance to ClonNAT caused by *Hermes* integration. Plating on SC-Ura plates confirmed that the majority of cells that remain lacked the donor plasmid and also contained a transposon integration.

Genomic DNA preparation, LM-PCR, and high-throughput sequencing (HTS) on the Solexa platform. Integration sites were recovered by the ligation mediated PCR approach described in ref. 1, modified for the Solexa HTS platform. Briefly, 25 OD_{600} equivalent yeast cells were used to prepare genomic DNA by using the Epicentre yeast MasterPure yeast DNA purification kit (catalog no. MPY80010) followed by RNase digestion and phenol-chloroform extraction to remove nucleases. Two micrograms of genomic DNA were digested overnight with *MseI*, ligated to linkers overnight at 16 °C, and digested a second time with *SpeI*. Nested PCR was carried out under stringent conditions by using *Hermes* R-end TIR-specific primers. Oligonucleotides used in this study are listed in Table S3. Solexa HTS-platform specific tags were introduced in the primers, SUN221 and SUN222, used in the final nested PCR. Amplification products were gel-purified size selected (100–500 bp) and sequenced by the massively parallel cluster formation-based Solexa method from Illumina using the custom sequencing primer, SUN254, corresponding to the *Hermes* R-end.

Yeast strains. The haploid yeast strain BY4727 (*MATα his3Δ200, leu2Δ0, lys2Δ0, met15Δ0 trp1Δ63 ura3Δ0*) was used in this study. The systematic sequence of the yeast genome in SGD (*Saccharomyces* Genome Database), which was used for genome feature annotations and analyzing *Hermes* insertions, contains *MATα* sequences.

For analysis of insertions in a diploid genome we used the yeast strain BY4743 (*MATα/MATa his3Δ200, leu2Δ0,LYS2/lys2Δ0, MET15/met15Δ0 ura3Δ0*).

Plasmid construction. The *NatMX* cassette (*NAT1* gene with *TEF* promoter and *TEF* terminator) was amplified by using plasmid pAG25 as template and primers SUN130 and SUN131, digested with *PstI* and *HindIII* sites and ligated into the *PstI* and *HindIII* sites of plasmid pBS*Hermes* (2) to give pSG16.

Hermes Left and Right TIRs flanking the *NatMX* cassette were amplified from pSG16 by using primers SUN132 and 133, digested with, and ligated into *BamHI* site of pRS416 to give pSG17.

Site-directed mutagenesis of pSG17 using SUN142 and SUN143 was done to introduce a single mutation at position 8, a G to C base change in the *CDE1* element of *CEN6* in the pRS416 plasmid vector backbone [to increase the plasmid loss rate after removal of selection (3)] to give pSG30.

The plasmid pHH1.9 (2) was used as a template to amplify *Hermes* transposase ORF using oligos NLC1496 (*SmaI*) and NLC1497 (*XhoI*). This sequence was ligated into p414G*als* (ATCC87344) (4) to give plasmid pSG2 (p*GalsHermes* Transposase).

The plasmid pSG3 is the same as pSG2 except that the *Hermes* transposase ORF harbors mutations that encode a G366W/M286T mutant transposase hyperactive for transposition activity, which was recovered from a screen looking for excision hyperactive transposases.

G366W/M286T mutant version of the *Hermes* transposase gene under the control of *GALs* promoter was amplified by using pSG3, as template in a PCR using oligos SUN154 and SUN113, ligated to plasmid pSG30 cut with *Not1* and *SacII* to generate pSG36.

In vitro Hermes insertions in deproteinized yeast DNA. Yeast genomic DNA was isolated as described above except that the DNA was treated with Proteinase K for 2 h at 42 °C after RNase treatment but before phenol-chloroform extraction and ethanol precipition to get highly pure genomic DNA with an A_{260}/A_{280} optical density ratio of >2. The genomic DNA was then sonicated on ice with 3 pulses for 1 s each at amplitude of 20% on a Branson sonifier. This setting on the sonifier gave an average fragment size of ≈3–5 kb in length. This DNA was ethanol precipitated, concentrated, and used as target DNA in the strand transfer reaction.

Strand transfer reaction was carried out by using 4 μg of purified full-length WT *Hermes* transposase expressed in *Escherichia coli*, 10 μM biotinylated double-stranded *Hermes* L-end oligo made by annealing oligos SUN360 and 361 and 2 μg of the fragmented genomic DNA in a strand transfer buffer containing 250 mM Mops, 1.0 M NaCl, 100 mM $MgCl_2$, 50% Glycerol, 100 mM DTT, 10 mg/mL BSA, 100% DMSO for 3 h at 37 °C. The reaction was stopped by adding stop mix to get a final concentration of 40 mM EDTA and 0.1% SDS and heating at 65 °C for 20 min.

The transferred strand was bound to Streptavidin beads (Invitrogen) in binding and washing buffer (B & W buffer): 100 mM Tris·HCl at pH 8.0, 1 mM EDTA, and 1 M NaCl. After the beads were washed with B &W buffer, the beads were resuspended in *MseI* restriction digestion mix, and the bound DNA was digested at 37 °C overnight. The beads are washed and linkers prepared by annealing oligos SUN360 and SUN205 were ligated on to the *Mse1*-digested ends of the *Hermes* L-end attached DNA. After the beads were washed to remove unligated linkers, PCR amplification of the *Hermes* L-end insertion site junction was carried out using SUN359 [primer with a Solexa (Illumina) platform specific sequence fused to *Hermes* L-end complementary sequence] and the linker primer SUN222 with the DNA attached to beads as a template. The PCR mix was separated from the beads, the amplicons size selected on an agarose gel and purified by gel extraction. Massively parallel sequencing was then carried out on the Solexa HTS platform by using the *Hermes* L-end specific custom sequencing primer, SUN 358.

Oligos used in this study are listed in Table S3.

Bioinformatic Methods. ***Filtering of the sequences.*** We considered only those Solexa reads that started with the *Hermes* end "AAGTTCTCTG", allowing one "N" mismatch. The resulting sequences were aligned to the yeast genome by using Bowtie (5). We trimmed the first 10 bp of each reads (the *Hermes* ends) to align only the yeast genomic sequences, and we used a seed of 20 bp instead of the default 28. We considered only those reads that aligned without mismatches to the yeast genome and among these, only those that mapped uniquely in a single location. Insertion events can occur multiple times in the same location because of biological and/or technical reasons. Because we could not distinguish between these two cases, we only considered nonredundant genomic positions.

Distribution of *Hermes* Around Genomic Features. Gene coordinates were downloaded from *Saccharomyces* Genome Database (www.yeastgenome.org) on February 5, 2008. TSS positions have been taken from a large-scale study by Jiang and Pugh (6). The distances between *Hermes* insertions and other genomic features are expressed in base pairs except in the analysis of *Hermes* insertion patterns within ORFs, where each ORF was rescaled to the same length and relative positions (percentages) were used.

Nucleosome Occupancy. Nucleosome occupancy data obtained from cultures grown in glucose is from Lee et al (7). We defined three genomic regions: (*i*) nucleosome free regions (NFR) are regions of DNA of at least 100 bp in length where the nucleosome occupancy is always lower than the second quartile (−0.361), (*ii*) nucleosome occupied regions (NOR) are regions of DNA of at least 100 bp in length where the nucleosome occupancy is always greater than the third quartile (0.137), and (*iii*) intermediate occupancy regions (IOR) are regions of DNA of at least 100 bp in length where the nucleosome occupancy is in the remaining intermediate range. The areas of the genome the nucleosome occupancy level does not fall into only one class over the 100-bp region are marked as "other."

Clusters. Clusters of insertions are defined as regions within a chromosome where the density of insertions is higher than what is expected by chance; here, we used kernel smoothing to determine the boundaries and significance of clusters (8). To define statistically significant clusters (those that are denser than expected by random chance), we performed 100 randomizations of the target site positions along the chromosome (including only sites that have the correct target motif and that are unique in the genome) and reclustered after every simulation. We define the cutoff as the highest value above the third quartile excluding outliers of a set of maximum peaks from 100 randomizations of the positions of the insertions within a chromosome. Any observed cluster higher than this cutoff was considered statistically significant.

Simulation. We created a simulated dataset of *Hermes* insertions, taking into account the limitations of the experimental procedure. The yeast genome was virtually fragmented by *MseI*. Fragments were randomly chosen, and if a fragment could have an insertion (contained NTNNNNAN) and was experimentally recoverable (between 50 and 3,000 bp, with the target site >10 bp from the end of the fragment, and unique in the yeast genome), that target site was included in the simulated dataset. Target sites not meeting these constraints, and non-NTNNNNAN sites, were not included in the simulated dataset.

Computing *P* Value for the Distribution of Target Site Recovery Frequencies. We used an extremely conservative simulation to determine whether it is expected that many experiments recover the same target site. Taking the 175,600 total unique target sites as the entire universe of possible target sites, we chose x_i numbers between 0 and 175,600, for i in {1,2,3,4,5,6}, with x_i being the number of unique target sites recovered in experiment i. For each of these trials we noted how many times each target site was chosen. The simulation was done 100,000 times, and in only three of these trials was a single target site chosen by all six experiments.

Programming Language and Statistical Analysis. All analysis was done by using the Python programming language (http://www.python.org/) and the R statistical package (R Development Core Team, 2009).

1. Ciuffi A, et al. (2009) Methods for integration site distribution analyses in animal cell genomes. *Methods* 47:261–268.
2. Warren WD, Atkinson PW, O'Brochta DA (1994) The *Hermes* transposable element from the house fly, Musca domestica, is a short inverted repeat-type element of the hobo, Ac, and Tam3 (hAT) element family. *Genet Res* 64:87–97.
3. Hegemann JH, Shero JH, Cottarel G, Philippsen P, Hieter P (1988) Mutational analysis of centromere DNA from chromosome VI of Saccharomyces cerevisiae. *Mol Cell Biol* 8: 2523–2535.
4. Mumberg D, Müller R, Funk M (1995) Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene* 156:119–122.
5. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
6. Jiang C, Pugh BF (2009) A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genome. *Genome Biol* 10:R109.
7. Lee W, et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* 39:1235–1244.
8. Tukey JW (1977) *Exploratory Data Analysis* (Addison–Wesley, Reading, MA).

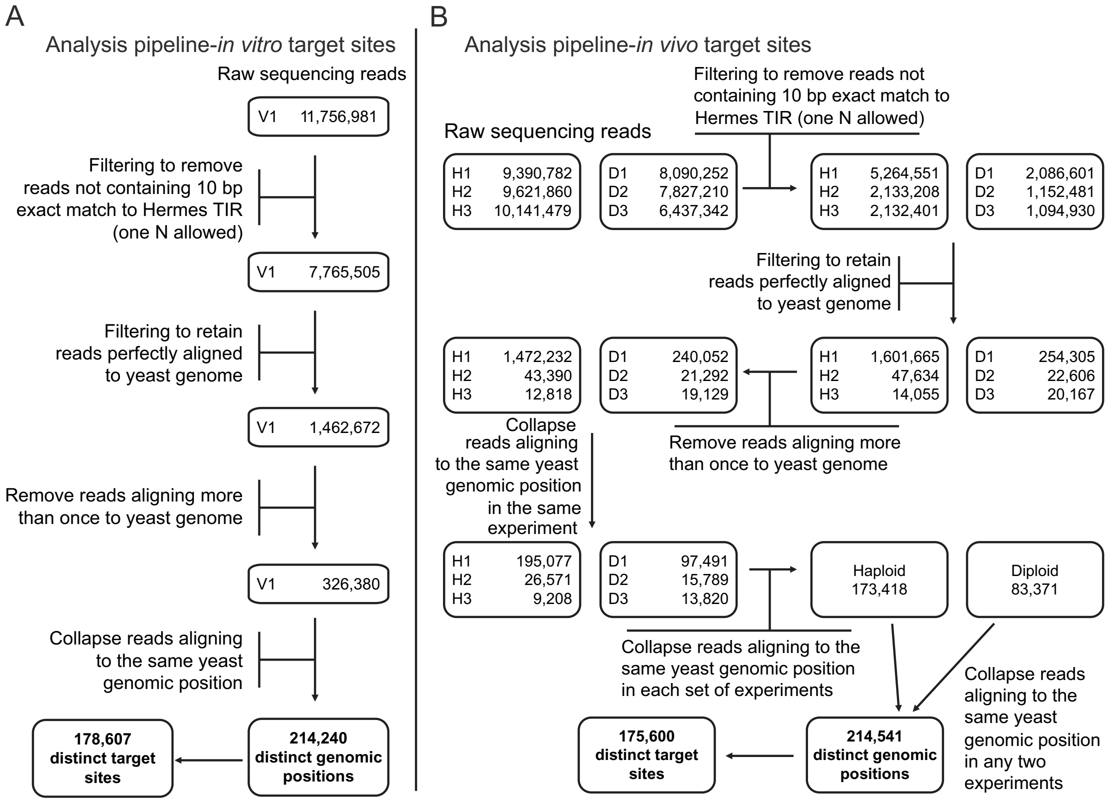
A
Analysis pipeline-*in vitro* target sites
Raw sequencing reads
V1 11,756,981
Filtering to remove reads not containing 10 bp exact match to Hermes TIR (one N allowed)
V1 7,765,505
Filtering to retain reads perfectly aligned to yeast genome
V1 1,462,672
Remove reads aligning more than once to yeast genome
V1 326,380
Collapse reads aligning to the same yeast genomic position
214,240 distinct genomic positions
178,607 distinct target sites
B
Analysis pipeline-*in vivo* target sites
Raw sequencing reads
H1 9,390,782
H2 9,621,860
H3 10,141,479
D1 8,090,252
D2 7,827,210
D3 6,437,342
Filtering to remove reads not containing 10 bp exact match to Hermes TIR (one N allowed)
H1 5,264,551
H2 2,133,208
H3 2,132,401
D1 2,086,601
D2 1,152,481
D3 1,094,930
Filtering to retain reads perfectly aligned to yeast genome
H1 1,601,665
H2 47,634
H3 14,055
D1 254,305
D2 22,606
D3 20,167
Remove reads aligning more than once to yeast genome
H1 1,472,232
H2 43,390
H3 12,818
D1 240,052
D2 21,292
D3 19,129
Collapse reads aligning to the same yeast genomic position in the same experiment
H1 195,077
H2 26,571
H3 9,208
D1 97,491
D2 15,789
D3 13,820
Collapse reads aligning to the same yeast genomic position in each set of experiments
Haploid 173,418
Diploid 83,371
Collapse reads aligning to the same yeast genomic position in any two experiments
214,541 distinct genomic positions
175,600 distinct target sites

Fig. S1. Analysis pipelines. Raw sequencing reads that did not contain the expected 10 bp of the substrate *Hermes* transposon R end and 25-bp alignments to the yeast genome with more than one mismatch were excluded, and the remaining reads were processed as shown for the in vitro insertion data (*A*) and in vivo insertion data (*B*). The final number of distinct target sites for the in vivo data are derived from all genomic positions mapped from all six experiments.

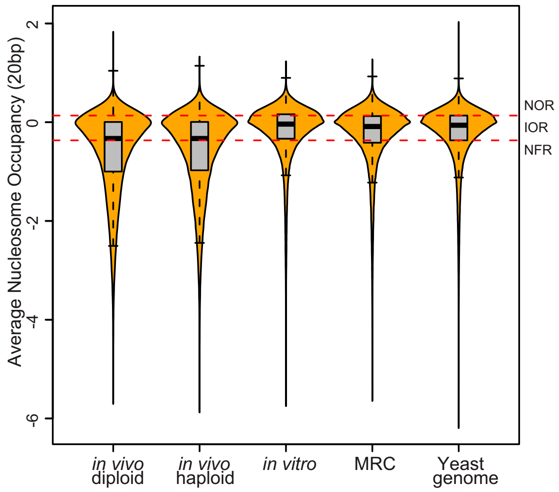

Fig. S2. Nucleosome occupancy distributions. Nucleosome-free regions (NFR), nucleosome occupied regions (NOR), and regions of intermediate nucleosome occupancy (IOR) were defined based on the dataset of Lee et al., 2007 (7). Regions at least 100 bp long with nucleosome occupancy in the highest quartile are defined as NOR, regions at least 100 bp long with nucleosome occupancy in the lowest quartile are defined as NFR, and regions at least 100 bp long with nucleosome occupancy in the middle two quartiles were considered to be intermediate in nucleosome occupancy. Note that if a 100-bp region contains areas of varying nucleosome occupancy, it is not placed into any class.

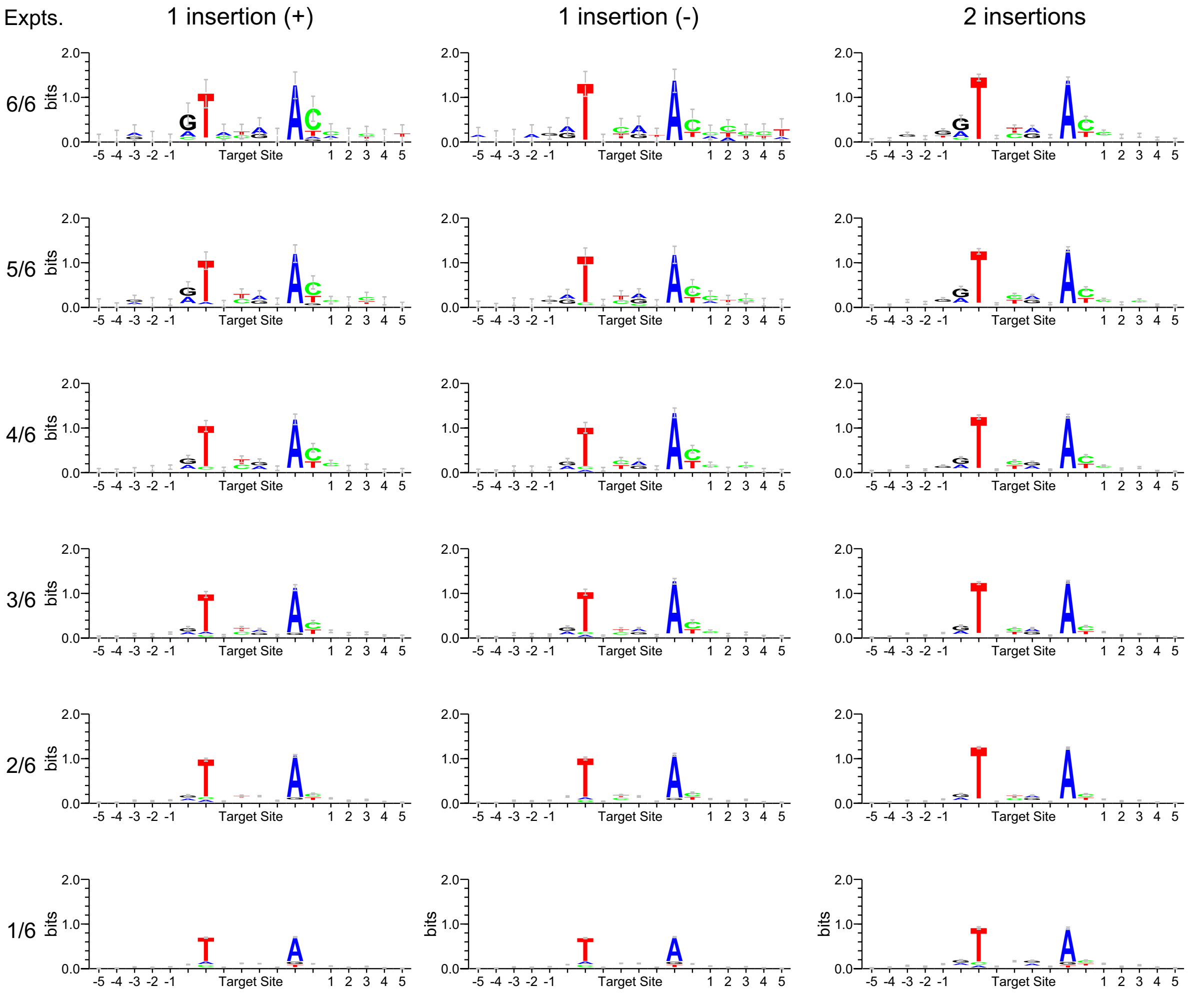

Fig. S3. Sequence logos of target sites from all in vivo experiments. Target sites from in vivo experiments are shown (and further divided into those harboring one or two insertions).

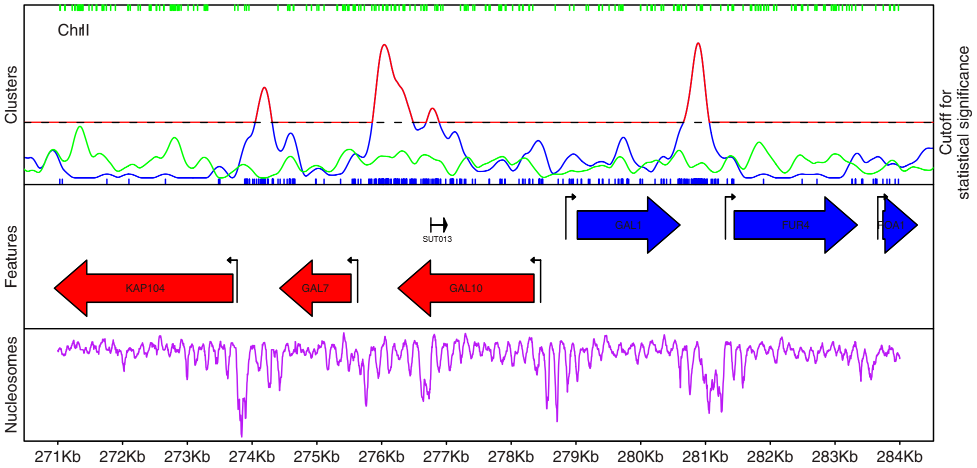

Fig. S4. Insertion clusters near the *GAL10* ORF. Clusters (red peaks achieve statistical significance above a simulated null distribution) of in vivo target sites (blue bars) are found near the promoters and/or terminators of many genes at this locus, including that for a stable untranslated transcript in *GAL10*. In vitro clusters (green line) of target sites (green bars) are also indicated in *Upper* with respect to to nucleosome occupancy (purple line).

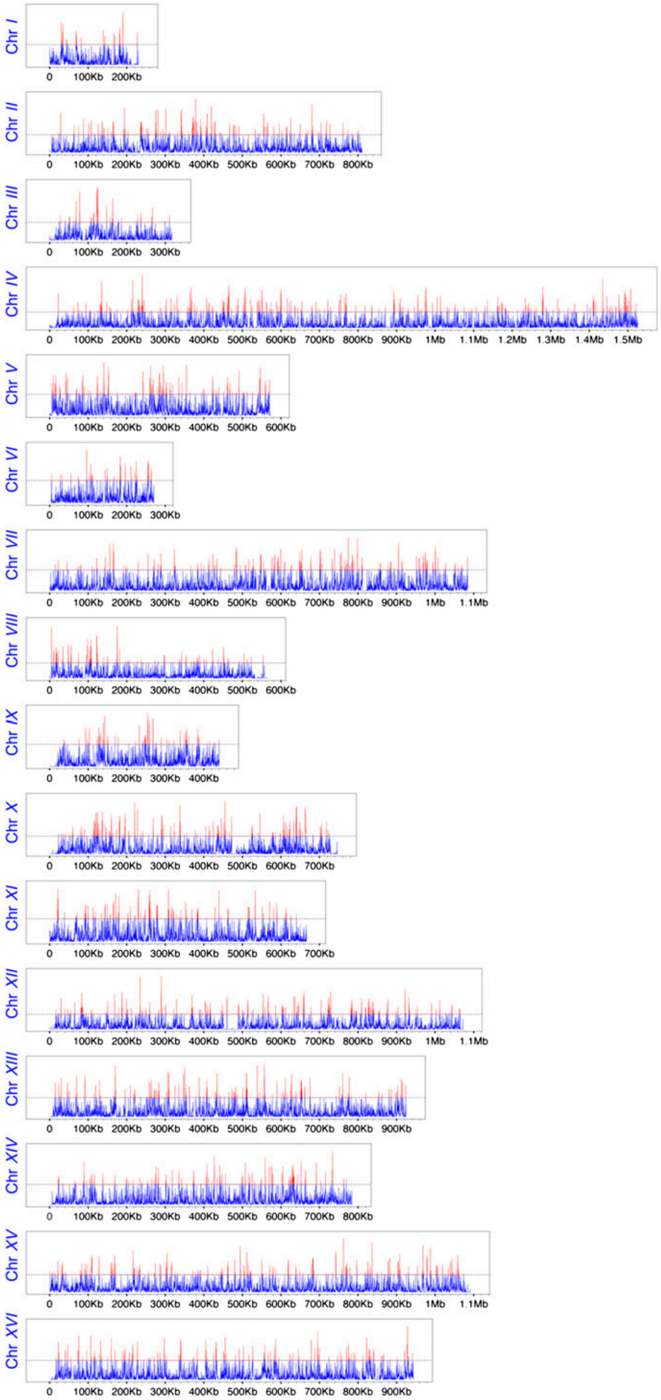

Fig. S5. A genome-wide view of *Hermes* insertions in vivo. The entire yeast genome is shown. Blue lines trace insertion density computed by kernel smoothing; red peaks are those that attain significance.

Table S1. Status of ORFs targeted by *Hermes* insertions

Datasets	Essential ORFs	Nonessential ORFs
Haploid	758	4,789
Diploid	1,070	5,345
Haploid and diploid	561	4,565
Yeast genome	1,211	5,773

Table S2. Genome-wide distribution of target site nucleosome occupancy levels

Experiments	No. of TSD	In NFR (%)	In IOR (%)	In NOR (%)	Unclassified (%)
Present in all 6	977	588 (60.2)	43 (4.4)	5 (0.5)	341 (34.9)
5 or more	3,062	1,686 (55.1)	159 (5.2)	22 (0.7)	1,195 (39.0)
4 or more	7,524	3,744 (49.8)	464 (6.2)	62 (0.8)	3,254 (43.2)
3 or more	18,310	8,172 (44.6)	1,235 (6.7)	171 (0.9)	8,732 (47.7)
2 or more	48,440	18,554 (38.3)	3,812 (7.9)	543 (1.1)	25,531 (52.7)
1 or more	175,600	52,770 (30.1)	17,139 (9.8)	2,463 (1.4)	103,228 (58.8)
In vitro	178,607	16,213 (9.1)	25,626 (14.3)	4,470 (2.5)	132,298 (74.1)
MRC	175,600	19,812 (11.3)	25,989 (14.8)	2,809 (1.6)	126,990 (72.3)
Yeast genome, bp	12,156,679	1,282,708 (10.6)	1,909,878 (15.7)	270,995 (2.2)	8,693,098 (71.5)

Table S3. Oligos

No.	Oligo	Sequence and modifications	Description
1	SUN204	/5Phos/TAGTCCCTTAAGCGGAG/3AmM/-NH2	Top strand of linker (*Mse1*)
2	SUN205	GTAATACGACTCACTATAGGGCTCCGCTTAAGGGAC	Bottom strand of linker
3	SUN243	CTTGCACTCAAAAGGCTTGACAC	5′ *Hermes* R-end specific primer (PCR1)
4	SUN206	GTAATACGACTCACTATAGGGCTC	3′ Linker specific primer
5	SUN221	AATGATACGGCGACCACCGAGATCTCTATGTGGC TTACGTTTGCCTG	5′ Solexa tag fused to *Hermes* R-end specific sequence (nested PCR2)
6	SUN222	CAAGCAGAAGACGGCATACGAGCTCTTCC GATCTGTAATACGACTCACTATAGGGC	3′ Solexa tag-linker sp primer (nested PCR2)
7	SUN254	CTATGTGGCTTACGTTTGCCTGTGGCTTGTTG	Custom solexa sequencing primer for *Hermes* transposon R-end
8	SUN360	TTGACACCCAAAACACTTGTGCTTATCTATGTGGCTTAC GTTTGCCTGTGGCTTGTTGAAGTTCTCTG	Bottom strand of *Hermes* TIR L-end (in vitro)
9	SUN361	5′P-CCAGAGAACTTCAACAAGCCACAGGCAAACGTAAG CCACATAGATAAGCACAAGTGTTTTGGGTGTCAA	Top strand of *Hermes* TIR L-end (in vitro)
10	SUN362	TAGTCCCTTAAGCGGAGCCCTATAGTGAGTCGTATTAC	Annealed with SUN205 for making linker (in vitro)
11	SUN358	GCAAGTGGCGCATAAGTATCAAAATAAGCCACTTGTTG	Custom Solexa sequencing primer for *Hermes* L-end (in vitro)
12	SUN359	AATGATACGGCGACCACCGAGATCTTAGCAAGT GGCGCATAAGTATCA	5′ Solexa tag fused to *Hermes* L-end specific sequence (in vitro) used with SUN222

Other Supporting Information Files

Dataset S1 (XLS)