

Supplemental methods

Reporter construction

TRIP Reporters were constructed from pT2-HSP-GFP (Addgene plasmid #65488). A promoterless construct was obtained by digestion with EcoRI, which excised exactly the *hsp70B* promoter, followed by ligation of the backbone. Promoter II was amplified from the gDNA of Kc167 cells with primers containing EcoRI and HindIII sites (primers 1 and 2, **Supplemental Table 2**). The PCR product was digested and cloned into the linearized EcoRI/HindIII backbone plasmid. For the other constructs, we first obtained the transcription start sites (TSS) positions of *Drosophila melanogaster* from the modENCODE 5' RACE experiment (DCC ID: modENCODE_2929). We selected 4 TSS of genes expressed in Kc167 according to published expression data. The sequence of the promoter was defined as the 1000 bp sequence upstream of this position. Taking advantage of the promoterless construct, we designed Gibson compatible primers (primers 3 to 8, **Supplemental Table 2**) to extract the promoters from the Kc167 genomic DNA and clone them into the EcoRV linearized plasmid in an isothermal 1 hour reaction at 50 °C performed in the Gibson reaction mix manufactured in house (Tris-HCl 0.1 M, MgCl₂ 0.01 M, dNTPs 0.2 mM each, DTT 0.04 M, PEG 8000 0.05 g/mL, NAD 1 mM, 6.4 U T5 exo (T5E4IIK Epicentre), 40 U Phusion DNA-polymerase (F530s VITRO) and 40000 U Taq-ligase produced in house). The circularized promoter-backbone constructs were purified with QIAquick PCR Purification Kit columns in 30 uL of water, of which 2 uL were electroporated in *E. coli* TOP10. Candidate plasmids were purified by Qiagen Mini kit from liquid culture grown from single colonies. GFP expression was verified by electroporating 1 ug of the purified constructs into 50 uL medium containing 2 million Kc167 cells (90 V 500uF) using 0.2 cm

cuvettes (Biorad), and expression of the GFP was measured by flow cytometry 72h post-electroporation.

Barcoded library generation

In order to insert barcodes in the reporter constructs, the template reporter plasmids were PCR-amplified (30 cycles in 50 μ L of Phusion polymerase mix with GC Buffer following recommendations from NEB) with primers 9 and 10 (**Supplemental Table 2**), generating linear copies of the template with the addition of a random barcode of 21 nucleotides, the Illumina PE1.0 sequence and a NlaIII restriction enzyme site. Following PCR, the template was digested by adding 20 U of DpnI (NEB) for 30 min at 37 °C. The PCR products were purified with Qiagen MinElute PCR Purification Kit and eluted in 30 μ L EB buffer. Finally, to circularize the molecules, an isothermal 1 hour reaction at 50 °C was performed with Gibson reaction mix (Tris-HCl 0.1 M, MgCl₂ 0.01 M, dNTPs 0.2 mM each, DTT 0.04 M, PEG 8000 0.05 g/mL, NAD 1 mM, 6.4 U T5 exo (T5E4IIIK Epicentre), 40 U Phusion DNA-polymerase (F530s VITRO) and 40000 U Taq-ligase produced in house), 400 ng of purified PCR products were mixed with Nuclease-free water in a 1:1 (v/v) ratio with the Gibson reaction mix. Four Gibson reactions were performed per reporter plasmid, they were pooled, purified with a Qiagen MinElute PCR Purification Kit in 30 μ L of water. The concentration was adjusted to 50 ng/ μ L. 100 μ L ElectroMAX™ DH10B™ *E. coli* cells were electroporated with 2 μ L of the purified Gibson reaction products. The mix was transferred to 0.2 cm cuvettes (Biorad) and electroporated in a MicroPulser™ (Biorad) with settings Ec1 (a single pulse of 1.8 kV). Up to three electroporation per construct were performed in order to obtain a complexity higher than 10⁶ transformants. After electroporation, bacteria from the three transformations were pooled and grown overnight at 37 °C under agitation in 500

mL liquid LB. Plasmids were purified with Macherey-Nagel Endotoxin-free DNA purification kits following manufacturer's instructions. The plasmid library was resuspended in Endotoxin-free water and the concentration adjusted to 1 ug/ul. The library constructs were sequenced to verify the correct inclusion of the Barcode and the NlaIII restriction site.

Cell culture and transfection

Kc167 cells were maintained in Schneider's *Drosophila* medium (Gibco). Transfections were performed by electroporating 20 million cells with 20 ug of plasmid DNA and 10 ug pT2-promoter-GFP (see next section), 5 ug pC-sleeping beauty (Addgene plasmid #65487) and 5 ug LNGFR expression plasmid at 250 V and 1000 uF. The cells were counted, pelleted at 300 g for 5 minutes, resuspended and electroporated in 750 uL of medium in Biorad cuvettes (0.4 cm) in a GenePulser II Electroporation system (Biorad). Immediately after electroporation 700 ul of the electroporated cell suspension were transferred to 25 cm² flasks containing 5 mL of fresh medium, leaving the top layer of dead cells that forms in the cuvette. 24 hours after electroporation the expression of the Sleeping Beauty 100X transposase and of LNGFR were induced by two heat shocks at 37 °C of two hours each, with at least four hours recovery between them. The next day, two more heat shocks were given in the same conditions. At day three after electroporation, LNGFR positive cells were selected using MACSelect LNGFR micro beads (Miltenyi biotech) and pools of 10.000, 20.000 or 50.000 cells were plated in 25 cm² flasks containing 5 mL and grown for 2 weeks transferring to a 75 cm² when the culture reached a density of 10⁷ cells/mL. This pooling was done in replicate for each promoter-construct to account for the biological variability.

inverse PCR

The genomic DNA from Kc167 cells was isolated with DNeasy kit (Qiagen) and 3 ug were digested with NlaIII restriction enzyme (NEB) for 2 hours at 37 °C followed by 20 minutes heat inactivation at 65 °C. The reaction was diluted to a final volume of 1.8 mL in T4 ligase buffer (Thermoscientific) to favor self-ligation events, and ligation was carried out with 20U T4 ligase at 16 °C overnight. After ligation samples were ethanol precipitated, pellets were redissolved in water and column purified (QIAquick PCR purification kit) eluting with 100 uL EB buffer. Non circularized templates were eliminated by 2h digestion at 37 °C with Plasmid-safe DNase (Epicentre), the enzyme was inactivated by heating 30 minutes at 70 °C and the product was column purified (QIAquick PCR purification kit). The backbone of the TRIP reporters contains a I-CeuI site outside the transposable cassette, taking advantage of this, non integrated plasmids were digested by 2h digestion at 37 °C with I-CeuI restriction enzyme (NEB) in a total volume of 100 ul followed by 20 minutes heat inactivation at 65 °C. All enzymatic reactions were carried out in the recommended manufacturer's buffer. PCR was performed in two rounds. In the first round (98 °C for 1 min // 98 °C for 15 sec / 60 °C for 30 sec / 72 °C for 1 min // – 25 cycles / 72 °C for 3 min) gDNA flanking insertions and barcode was amplified using primers 11 and 12 (**Supplemental Table 1**) annealing to the transposon feet and to the Illumina PE1.0 sequence present in the integrated construct. In the second round (98 °C for 1 min // 98 °C for 15 sec / 60 °C for 30 sec / 72 °C for 1 min // – 10 cycles / 72 °C for 3 min) Illumina adaptors for pair-end sequencing and indexes needed for multiplexing (primers 12 and 13. **Supplemental Table 1**) were added.

Spikes

In order to evaluate the quality of the experiment we used single molecule spikes. Spikes consisted of 118-mer oligonucleotides (**Supplemental Table 2**) containing a constant part, that allows us to identify it after the sequencing, and a random stretch of 20 nucleotides that makes each molecule unique. We introduced approximately 5000 spike molecules in the gDNA or cDNA of each sample before PCR. The read counts obtained for each spike thus indicate the approximate number of reads per molecule of template.

DpnII assay

Kc167 cells containing integrations were electroporated in the same conditions as detailed above with one of the four Dam fusion constructs expressing Dam-Braham, Dam-HP1, Dam-Polycomb and Dam-H1 (plasmids pNDam_Myc_Brahma, pNDam_Myc_HP1, pNDam_Myc_Polycomb, pGW_NDam_Myc_H1), or control plasmids: pNDam_Myc (Dam only), pCasper (vector only). 48 hours after electroporation, the genomic DNA was isolated (DNeasy kit, Qiagen). To eliminate non methylated insertions, 2 ug of genomic DNA were treated with 40 U of DpnII enzyme overnight at 37 °C followed by 20 minutes of heat inactivation at 65 °C. Protected (methylated) barcodes were amplified by PCR (98 °C for 1 min // 98 °C for 15 sec / 68 °C for 30 sec / 72 °C for 1 min // – 25 cycles / 72 °C for 3 min) with primers 12 and 15 (**Supplemental Table 2**) containing Illumina adapters. The products were sequenced single read on a HiSeq 2000 with 50 bp single read setup. The non digested sample was used to amplify all the barcodes present in the population and normalised by barcode abundance. The binding score was calculated by dividing the normalised barcode counts in the Dam-fusion

protein sample by its counterpart in the Dam-only sample, in order to account for accessibility biases.

Circularized Chromosome Conformation Capture (4C)

Cells containing integrations of reporter pl were sorted on GFP expression and pools of 100 positive cells were grown. 10 million cells of each clone were used for the 4C experiment following the protocol previously described (Stadhouders et al. 2013). Briefly, cells at a density of 1 million per mL were crosslinked 10 minutes with 1% formaldehyde, after cell permeabilization intact nuclei were incubated at 37 °C overnight with 250 U of NlaIII (NEB) followed by overnight ligation at 16 °C with 25 U of T4 DNA Ligase (Roche). Ligated products were de-crosslinked and digested with 50 U of MluCI (NEB) at 37 °C overnight, then DNA fragments were self-ligated by incubating overnight at 16 °C with 100 U of T4 DNA Ligase (Roche). 200ng of circularized products were PCR amplified (94 °C for 2 min // 94 °C for 15 sec / 55 °C for 1 min / 68 °C for 3 min // – 27 cycles / 68 °C for 7 min) with specific primers annealing to the reporters and containing adapters and indexes for Illumina sequencing (primer 12 and primer 16. **Supplemental Table 2**). Five PCRs per clone were done and product pools were sequenced on a HiSeq 2000 in 1x50 bp single read setup.

Mapping integrated reporters

To map reporters, inverse PCR products were sequenced as 2x100 bp paired-end reads with a HiSeq2000 sequencer (Illumina) using v3 sequencing chemistry. Barcodes were extracted from forward reads by searching the first NlaIII site (CATG) between 15 and 25 bp from the 5' end of

the read. The insertion site was extracted from the reverse read through an inexact search of the pT2 transposon (TGTATGTAAACTTCCGACTTCAACTGTA) allowing up to 5 errors (mismatches, insertions and deletions) using Seeq v1.1.2 (<http://github.com/ezorita/seeq>). The portion of the read from the insertion point till the first NlaIII site or till the 3' end of the read if none was present, was extracted as the *Drosophila* sequence flanking the insertion point. Sequencing errors in the barcodes were reverted by sequence clustering using Starcode v1.0 (Zorita et al. 2015) allowing up to two errors (mismatches, insertions and deletions) and using the 'message passing' clustering algorithm. The flanking sequence was mapped on dm3/R5 with GEM v1.376 (Marco-Sola et al. 2012) with options -m3 --unique-mapping (allowing 3 mismatches and requesting mapping to be unique). For each barcode, all the candidate insertion sites were collected and those representing more than 10% of the associated mapped reads were compared to each other. If the diameter of their genomic coordinates was higher than 10 bp, the barcode was filtered out and flagged as unmapped. Otherwise, the location of the barcode was assigned to the candidate insertion site with the highest amount of reads.

Measuring the expression of the integrated reporters

Barcodes and spikes were extracted from forward reads in gDNA and mRNA through an inexact search of the sequence of GFP (CATGCTAGTTGTGGTTTGTCCAACT) or the constant part of spikes (CATGATTACCCTGTTATC) allowing up to 4 or 2 errors (mismatches, insertions and deletions), respectively. The sequence immediately upstream of the hit was considered to be the barcode or the spike identifier if its length was between 15 and 25 bp, otherwise the read was discarded. Sequencing errors in the barcodes were reverted by sequence clustering using

Starcode v1.0 (Zorita et al. 2015) allowing up to two errors (mismatches, insertions and deletions) and using the ‘message passing’ clustering algorithm.

The mean read count of the spikes indicates the approximate number of count per molecule of input to the PCR or RT-PCR. For every replicate, one half pseudo molecule was added to every barcoded (*i.e.* half of the mean spike read count for this replicate) and the numbers were sum-normalized. Replicates were averaged and the expression of a barcode was computed as the log₂ of the mRNA counts divided by the gDNA counts. The scores were mean-normalized to be comparable between promoter sets.

Contact frequencies with genomic features

Active genes were defined as those expressed higher than the median. Promoters were defined as the annotated transcription start site, and terminators were defined as the annotated transcription end site. Housekeeping and development enhancers were used as provided by (Zabidi et al. 2015). Contact frequencies between a bin and promoters were computed as the ratio between (i) the weighted number of contacts between this bin and bins containing promoters and (ii) the total number of contacts of this bin. The weight of a bin was equal to the number of promoters in this bin. This is the scaled dot product between a row of the Hi-C matrix and the number of promoters per bin. Contact frequencies with terminators and enhancers were computed similarly. To compute the contact frequencies from 4C data, we averaged replicate profiles by geometric mean, binned the values in 2 kb windows and filtered out the barcodes with contacts in less than 20 windows. From these values, the contact frequencies were

computed as above, i.e. as the dot product between the binned 4C profile and the number of promoters or enhancers per bin.

Regression analysis

Linear models were fitted to predict the expression level of the reporters measured as the mean-normalized log2 ratio between mRNA and gDNA barcode counts (named Y below). For a numeric predictor X , the three parameter model $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ was fitted to the observations with the function `lm()` from R. The term X^2 is meant to capture potential nonlinear dependencies between X and Y . For a categorical variables X such as chromatin states, the fitted model was $Y = \beta_0 + \beta_1 X$, where the number of parameters is the number of categories of X (including a term X^2 is meaningless in this case). In all cases, the predictive power was defined as the ratio between the variance of the predicted values and the variance of the observed values.

Regression with multiple predictors was carried with lasso regularization (Friedman et al. 2010). A full model with the contact frequency with active terminators plus the 112 chromatin predictors and their associated square terms (227 parameters in total) was fitted with the R package `glmnet`. The function `cv.glmnet()` was used with default parameters, i.e. we performed 10-fold cross-validation of a lasso-regularized linear model for 100 values of the tuning parameter λ . The number of non-zero parameters and the squared validation error were measured for every value of λ .

Birth-diffusion-death model

Let H be the Hi-C matrix of a single chromosome, with dimensions $N \times N$. Consider then the matrix P defined by

$$P_{ij} = H_{ij} / \sum_{k=1}^N H_{ik}$$

which represents the row-normalized matrix. The matrix P can be interpreted as the adjacency matrix of a graph, in which P_{ij} corresponds to the probability of jumping from site i to site j .

We take a vector x such that

$$\sum_{i=1}^N x_i = 1$$

so that x_i can be interpreted as the fraction of particles at site i . Since P is row-normalized, the dot product $x \cdot P$ represents the population after one step on the graph. In summary, we can describe the birth of particles as a normalized vector x , then study its diffusion properties by iterating the dot product $x \cdot P$.

We introduce a simple first-order decay, so that at each step of the random walk the overall population is damped by the factor $\exp(-n/\tau)$, where n is the step number, and τ is the particle half-life. Therefore, the equilibrium population of particles starting from x will be given by

$$x_{eq} = \sum_{n=0}^{\infty} \exp(-n/\tau) x \cdot P^n$$

Once the equilibrium population is obtained, we can calculate the logarithm of the ratio between the population of a site and the average population, and compute the correlation coefficient with the measured reporter expression.