EXTENDED EXPERIMENTAL PROCEDURES

Data Analysis

Microarray Data

Microarray data normalization and analysis were performed with R (R Development Core Team, 2009). All DamID data were subjected to loess normalization. The median correlation between independent replicate experiments was 0.72.

The DamID procedure relies on laying a methylation ''footprint'' on the DNA at GATC sites, the motif that is recognized by Dam. The methylation is subsequently assayed by the enzyme DpnI that specifically cuts DNA on methylated GATC sites. For those reasons, the target material in DamID experiments consists of DNA fragments flanked by GATC sites. Those fragments are typically 200-300 bp in size, but can be longer and hybridize to several probes of the array. To avoid this complication, the main text refers to them as ''probed loci.'' For the sake of clarity, they will be referred to as 'GATC fragments' in the present document.

The array platform covers 183,258 GATC fragments (FlyBase release 5) on six chromosome arms. When several probes mapped to the same GATC fragment, we averaged their normalized $\log_2$ ratio to obtain a single score per GATC fragment.

Alignment Plots

Custom R scripts were used to calculate average binding profiles around 5' and 3' ends of genes. Genomic locations were converted to coordinates relative to the nearest 5' (respectively 3') end of a gene, before applying a running median with a window covering 2% of the plotted data. To ensure that points are aligned only once, windows around the end of a gene range from the midpoint of the gene to the mid distance to the next gene. For 5' (respectively 3') alignments, genes that had an upstream (downstream) neighbor in tandem orientation closer than 500 bp were excluded. This was done to avoid that features in 3' (5') of a gene would influence the alignment plot.

Highly Conserved Noncoding Elements

Highly conserved noncoding elements (HCNEs) were defined as sequences of minimal length 50 bp and having at least 98% identity with another species (Engström et al., 2007). Accordingly, we mapped HCNEs by aligning the introns and intergenic regions of *Drosophila melanogaster* (FlyBase release 5.17) to the genome of *D. mojavensis* (FlyBase release 1) by exonerate (Slater and Birney, 2005).

Mapping Digital Gene Expression Tags

Mapping of Digital Gene Expression tags was carried out by BLAST. We prepared a database of tags consisting of the 23 nucleotides downstream of the last GATC site of every annotated transcript. Using the BLAST default parameters to map the Solexa reads against this database, we could map 70.3% and 69.4% of the total number of reads from both experiments, respectively. All hits had at most two mismatches. Counts per gene were computed by adding up the counts of all the transcripts of a given gene. Counts were then normalized to the total number of reads per experiment and replicates were averaged. Gene mapping informations were taken from *D. melanogaster* FlyBase release 5.8.

GO Analysis

The GO Slim terms for *D. melanogaster* were downloaded from http://www.geneontology.org/GO_slims/archived_GO_slims/goslim_Drosophila.0200. The obsolete terms were updated according to the data obtained from http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology_ext.obo (date: 09:11:2009 14:57). Gene associations were downloaded from http://www.geneontology.org/cgi-bin/downloadGOGA.pl/gene_association.fb.gz (CVS version 1.159). Genes were further associated with all the terms higher in the GO hierarchy as indicated by the ''is a'' and ''part of'' keywords. Enrichment or depletion for GO terms were tested against the hypergeometric distribution, at alpha level 0.01 (correcting for multiple testing by the Bonferroni method).

FlyAtlas Data

We took the $\log_{10}$ of the intensity reads from the FlyAtlas database (Chintapalli et al., 2007) and mean normalized them. 4086 genes from the database were annotated as BLACK in our study. The genes were clustered by using the hclust and dist functions from R with default parameters.

Analysis of DNA-Binding Factor Targeting

First we obtained a position-specific affinity matrix (PSAM) for each of the DBFs in our compendium. For each separate chromatin type, we subtracted the type-specific mean DamID $\log_2$ ratio from all GATC fragments assigned to it. This was done to prevent the inferred PSAMs from converging to a motif with vastly different abundance among the chromatin types. For each DBF, we then used the OptimizePSAM tool from the REDUCE suite (http://bussemakerlab.org/software/REDUCE/) to fit a PSAM to the 10,000 GATC fragments with the highest DamID signal, expected to cover an adequate range of binding affinities. A 2 kb window around the center of the GATC fragment was associated with each DamID $\log_2$ ratio. To seed the fit, we used a consensus motif based on the known binding preferences of each factor (AGGTGGCGC for CTCF; GAGAG for GAF; CGGN(11)CCG for GAL4; TGMSWMA for JRA; CACGTG for MNT; and GCATAYYY for SU(HW)). Optimal relative affinity parameters were determined for each nucleotide position with this consensus, as well as 3–4 flanking nucleotide positions on each side.

We used the inferred PSAM corresponding to each DBF to predict its relative *in vitro* binding affinity to the 2 kb region using the AffinityProfile tool. To analyze the trend between predicted affinity and DamID $\log_2$ ratio, affinities were first converted to ranks, and the loess function from the R language was used (using span = 1/10) to fit a smooth curve for each chromatin type. Statistical

significance of the association between predicted affinity and DamID $\log_2$ ratios was determined using a rank correlation (implemented in the function cor.test(method="spearman") in R).

Hidden Markov Models

DamID Data Structure

DamID $\log_2$ ratios (Dam fusion protein over Dam only) of biological replicates were loess-normalized and averaged to give a probe-wise estimated $\log_2$ ratio. Since several probes of the array can map to the same GATC fragment of index k, we averaged the normalized $\log_2$ ratios of those probes into a single score x_k. In our experience, after loess normalization, the Student's t represents an excellent approximation of the distribution of x_k, whereas the Gaussian does not (see Figure S5A).

The DamID profile of protein FBpp0074431 used in Figure S5A showed no specific target. As a result, the distribution of the scores x_k is unimodal. Note that when a protein binds specific targets, the distribution of scores is actually a mixture of two Student's t distributions. Another peak centered around a higher average score corresponds to the score distribution of the bound loci.

Target Identification

To identify the binding sites of the protein of interest, we assume the existence of two states: 'unbound' and 'bound.' Intuitively, a GATC fragment will be more likely in state 'bound,' say, if its DamID score is close to the mean score of all GATC fragments in the state 'bound,' and if its immediate neighbors are also in the state 'bound.' Hidden Markov Models (HMMs) provide a statistical framework to integrate the information of probes along with their neighbors and provide tractable algorithms to find the most likely segregation into 'unbound' and 'bound,' *i.e.* the optimal set of targets (Cappé et al., 2005).

The fundamental concept of HMMs is to use the linear ordering of the observations to decompose their probability into a *transition* term and an *emission* term. The transition term is the probability of going from one state to another at any single jump, for example the probability of going from 'unbound' to 'bound' between two consecutive GATC fragments. Those probabilities are defined and can be estimated. The emission term describes the probability of the observation given a state, for example the probability that the DamID score would equal x_k given that the GATC fragment is in state 'bound'. Parameter estimation in the HMM framework is typically carried out through the Baum-Welch algorithm, which is a particular case of EM (Estimation-Maximization) algorithm (Dempster et al., 1976). Each iteration of the algorithm consists in computing the expected likelihood of the model and then finding the values of the parameters for which this quantity is maximal. The form of the likelihood function allows to carry out the computation independently for the transition parameters, through the Forward-Backward algorithm, and the emission parameters. Computing and optimizing the expected likelihood of the emission parameters depends on the distribution.

In the case of the Student's t distribution, this can be done by an EM algorithm. Because the Baum-Welch algorithm is itself an EM algorithm, their iterations can be combined in a single EM algorithm optimizing both transition and emission parameters. However, the convergence of the EM algorithm is very slow in the case of the Student's t distribution, motivating the use of an alternative algorithm termed MCECM (Liu and Rubin, 1995). This algorithm, specifically designed for the Student's t distribution, features a 4-step iteration. First, the degree of freedom ν is fixed and the expected likelihood is computed and maximized as a function of the other parameters (μ and σ^2). Then μ and σ^2 are fixed and the expected likelihood is computed and maximized as a function of ν. We combined the MCECM, and the Baum-Welch algorithms into an algorithm allowing efficient parameter estimation of HMMs with Student's t distribution.

Detailed Derivation

We model the series of DamID scores x_k $(k = 1, \ldots, N)$ by a two-state HMM with emissions following a Student's t distribution. The states 'unbound' and 'bound' are denoted S^0 and S^1, respectively. Target identification amounts to inferring the most likely sequence of states $(S_1, \ldots, S_N)$. Given $S_k = S^i (i = 0, 1)$, x_k follows a Student's t distribution with mean μ_i, variance σ_i^2 and ν degrees of freedom. Note that ν is the same for both states. The transition between the states are described by the 2×2 matrix Q. We will refer to the probability of transition from state S^i to state S^j by $Q(S^i, S^j)$.

We will denote $\theta^{(t)} = (\mu_0^{(t)}, \mu_1^{(t)}, \sigma_0^{2(t)}, \sigma_1^{2(t)}, \nu^{(t)}, Q^{(t)})$ the vector of parameters at iteration t of the algorithm. Following the Baum-Welch procedure, we will write the full log-likelihood of the observations (*i.e.*, the likelihood of the observed and unobserved variables) and iteratively maximize its expected value conditional on the observed data and the current values of the parameters.

The likelihood of the Student's t distribution is best described as that of a Gaussian random variable with random variance. More specifically, $X \sim t(\mu, \sigma^2, \nu)$ if and only if $X \sim N(\mu, \sigma^2/\tau)$ and $\tau \sim \chi^2(\nu)$, where X and τ are independent. Importantly, τ is not observed. Note that $\tilde{\sigma}^2$, the variance of X, is not σ^2, but rather a function of (σ^2, ν). The complete log-likelihood of X is decomposed as a sum of two terms: the log-likelihood of the Gaussian variable, ℓ_N, and that of the χ^2 variable, ℓ_{χ^2}.

$$\ell(\mu, \sigma^2, \nu | x, \tau) = \ell_N(\mu, \sigma^2 | x, \tau) + \ell\chi^2(\nu | \tau) \tag{1}$$

$$= -\frac{1}{2}\log 2\pi - \log\sigma + \frac{(x-\mu)^2\tau}{2\sigma^2} + \\ -\nu/2\log 2 - \log\Gamma(\nu/2) - \tau/2 + (\nu/2 - 1)\log\tau$$

From this, we can write the complete log-likelihood of the data as:

$$\ell\left(\mu_0,\mu_1,\sigma_0^2,\sigma_1^2,\nu,Q\middle|(S_1,x_1,\tau_1),\ldots,(S_N,x_N,\tau_N)\right)$$
$$=\log(\mathrm{Prob}(S_1))+\sum_{k=2}^{N}\log Q(S_{k-1},S_k)+\sum_{k=1}^{N}\ell_N(\mu_0,\mu_1,\sigma_0^2,\sigma_1^2|S_k,x_k,\tau_k)+\sum_{k=1}^{N}\ell\chi^2(\nu|\tau_k). \quad (2)$$

The conditional expectation of the full log-likelihood can be computed separately for all the terms of the sum. The term

$$E_{\theta^{(t)}}\left\{\log(Prob(S_1))+\sum_{k=2}^{N}\log Q(S_{k-1},S_k)\middle|x_1,\ldots,x_N\right\}$$

corresponds to the transition parameters and can be computed by the Forward-Backward algorithm. Using results and notations from Cappé et al. (2005), this yields the estimate of $Q^{(t+1)}$ as:

$$Q^{(t+1)}\left(S^i,S^j\right)=\frac{\sum_{k=1}^{N}\phi_{k-1:k|n}(i,j)}{\sum_{k=1}^{N}\sum_{l}\phi_{k-1:k|n}(i,l)}, \quad (3)$$

where $\phi k(i)=\mathrm{P}\theta^{(t)}(S_k=S^i|x_1,\ldots,x_N)$, *i.e.* the probability that the GATC fragment of index k is in state S^i $(i=0,1)$ given the whole sequence of observations provided by the Forward-Backward algorithm.

The second term can be expressed as a function of moments, which makes subsequent computations easier.

$$E\theta^{(t)}\left\{\sum_{k=1}^{N}\ell_N\left(\mu_0,\mu_1,\sigma_0^2,\sigma_1^2|S_k,x_k,\tau_k\right)\middle|x_1,\ldots,x_N\right\}=Constant-E_{\theta^{(t)}}\left\{\sum_i\sum_{k=1}^{N}1_{\{S_k=S^i\}}\left(\log\sigma_i+\frac{(x_k-\mu_i)^2\tau_k}{2\sigma_i^2}\right)\middle|x_1,\ldots,x_N\right\}=Constant$$
$$-\sum_i\log\sigma_i E_{\theta^{(t)}}\left\{\sum_{k=1}^{N}1_{\{S_k=S^i\}}\middle|x_1,\ldots,x_N\right\}+\sum_i\frac{1}{2\sigma_i^2}E_{\theta^{(t)}}\left\{\sum_{k=1}^{N}1_{\{S_k=S^i\}}\tau_k x_k^2\middle|x_1,\ldots,x_N\right\}+\sum_i\frac{\mu_i}{\sigma_i^2}E_{\theta^{(t)}}\left\{\sum_{k=1}^{N}1_{\{S_k=S^i\}}\tau_k x_k\middle|x_1,\ldots,x_N\right\}$$
$$-\sum_i\frac{\mu_i^2}{2\sigma_i^2}E_{\theta^{(t)}}\left\{\sum_{k=1}^{N}1_{\{S_k=S^i\}}\tau_k\middle|x_1,\ldots,x_N\right\}. \quad (4)$$

Here, $1\{S_k=S^i\}(\cdot)$ is the indicator function of the set $\{S_k=S^i\}$. It can be shown by direct computation that the conditional distribution of τ_k given $(x_1,\ldots,x_n)$, $\{S_k=S^i\}$ and $\theta^{(t)}$ is $\Gamma((\nu^{(t)}+1)/2,\ (\nu^{(t)}+(x_k-\mu_i^{(t)})^2/\sigma^{2(t)})/2)$, so that

$$E_{\theta^{(t)}}\left\{1_{\{S_k=S^i\}}\tau_k\middle|x_1,\ldots,x_n\right\}=\phi_k(i)w_k(i),where \quad (5)$$

$$w_k(i)=\frac{\nu^{(t)}+1}{\nu^{(t)}+\left(x_k-\mu_i^{(t)}\right)^2/\sigma^{2(t)}}. \quad (6)$$

In other terms, we define weights $w_k(i)$ at index k and for state S^i that are the expected values of the unobserved τ_k conditional on the observations and the current values of the parameters. Note that the weights are small for outliers (*i.e.* for $(x_k-\mu_i^{(t)})^2/\sigma^{2(t)}$ high), so that those observations contribute little to the final estimate of the mean and variance. By substituting (5) into (4), one obtains the following values for the expected sufficient statistics.

$$E_{\theta^{(t)}}\left\{\sum_{k=1}^{N}1_{\{S_k=S^i\}}\tau_k\middle|x_1,\ldots,x_N\right\}=\sum_{k=1}^{N}\phi_k(i)w_k(i),$$

$$E_{\theta^{(t)}}\left\{\sum_{k=1}^{N}1_{\{S_k=S^i\}}\tau_k x_k\middle|x_1,\ldots,x_N\right\}=\sum_{k=1}^{N}\phi_k(i)w_k(i)x_k,$$

$$E_{\theta^{(t)}}\left\{\sum_{k=1}^{N}1_{\{S_k=S^i\}}\tau_k x_k^2\middle|x_1,\ldots,x_N\right\}=\sum_{k=1}^{N}\phi_k(i)w_k(i)x_k^2.$$

The expected sufficient statistics turn out to be weighted moments of the observations. The combined weights $\phi_k(i)w_k(i)$ show that observations unlikely to be in state S^i (*i.e.* $\phi_k(i)$small) or being an outlier for state S^i (*i.e.* $w_k(i)$small) contribute little to the estimation of

mean and variance for state S^i. The updated values of μ_0, μ_1, σ_0^2 and σ_1^2 are the ones that maximize the expected log-likelihood. They are obtained by differentiating (4) with respect to the parameters:

$$\mu_i^{(t+1)} = \frac{\sum_{k=1}^{N} \phi_k(i) w_k(i) x_k}{\sum_{k=1}^{N} \phi_k(i) w_k(i)}, \tag{7}$$

$$\sigma_i^{2(t+1)} = \frac{\sum_{k=1}^{N} \phi_k(i) w_k(i) \left(x_k - \mu_i^{(t+1)}\right)^2}{\sum_{k=1}^{N} \phi_k(i)}. \tag{8}$$

Following the EM procedure, one could update $\nu^{(t)}$ by differentiating the last term of (2). However this procedure proves to give slow convergence. Instead, we used the idea of the MCECM algorithm (Liu and Rubin, 1995), which is substantially faster than the EM in terms of number of cycles. This involves performing another Forward-Backward pass with parameters $\tilde{\theta}^{(t)} = (\mu_0^{(t+1)}, \mu_1^{(t+1)}, \sigma_0^{2(t+1)}, \sigma_1^{2(t+1)}, \nu^{(t)}, Q^{(t+1)},)$ and computing the weights $w_k(i)$ again through formula (6) before differentiating the last term of (2). Let

$$\tilde{w}_k = \sum_i \phi_k(i) w_k(i). \tag{9}$$

The weights $\tilde{w}_k$ are the weighted averages of the weights $w_k(i)$ and are designed to be small for observations that are outliers in every state. The updated value $\nu^{(t+1)}$ is then the value ν such that

$$1 - \psi\left(\frac{\nu}{2}\right) + \log\left(\frac{\nu}{2}\right) + \frac{1}{N}\sum_{k=1}^{N} \log(\tilde{w}_k) - \tilde{w}_k + \frac{1}{N}\sum_{k=1}^{N} \psi\left(\frac{\nu+1}{2}\right) - \log\left(\frac{\nu+1}{2}\right) = 0, \tag{10}$$

where $\psi(z)$ is the digamma function *i.e.* $\psi(z) = d\log\Gamma(z)/dz$. The solution of this equation is computed numerically (by dichotomy).

Here is a summary of the different steps carried out at each cycle of the algorithm, where CM stands for 'constrained' or 'conditional' maximization (Liu and Rubin, 1995).

E-step 1: Compute the Forward-Backward smoothing estimates using $\theta^{(t)}$ and compute the weights $w_k(i)$ through (6).

CM-step 1: Update the transition probabilities through (3), the estimated means through (7), and the estimated variances through (8).

E-step 2: Compute the Forward-Backward smoothing estimates using $\tilde{\theta}^{(t)}$ and compute the weights again through (6). Compute the averaged weights $\tilde{w}_k$ through (9).

CM-step 2: Compute $\nu^{(t+1)}$ by solving (13).

After multiple cycles, the estimates usually converge to stable values, at which point the algorithm is stopped. The most likely sequence of states $(S_1, \ldots, S_N)$ is then computed through the Viterbi algorithm.

Virtual Bridging

Gaps in the array design are filled by virtual positions for which emissions are not available (NA). Gaps longer than $2D$ are bridged by $L/(D-1)$ virtual positions, where D is the mean distance between two consecutive GATC fragment 'start' positions and L is the gap size.

The emission probability of those virtual positions is set to 1 for every state. As a consequence, virtual bridging directly influences the estimation of the transition probabilities Q in the modified Baum-Welch algorithm presented above because it modifies the spacing between reads.

Multidimensional Inference

The modified Baum-Welch algorithm presented above can easily be applied to the case of multidimensional measurements. These are typically different binding profiles, or different linear combinations of binding profiles.

Assume that measurements x_k $(k = 1, \ldots, N)$ are now p -dimensional vectors, such that $x_{k,p}$ denotes the value of the k th read in the p th dimension. The distribution of x_k given $S_k = S^i$ is a multidimensional Student's t variable with mean vector μ_i, and mean variance matrix $\tilde{\Sigma}_i$ and ν degrees of freedom.

Again, the likelihood of the multidimensional Student's t distribution is described as that of a multidimensional Gaussian random variable with random variance. $X \sim T(\mu, \Sigma, \nu)$if and only if $X \sim N(\mu, \Sigma/\tau)$ and $\tau \sim \chi^2(\nu)$, where X and τ are independent. Again $\tilde{\Sigma}$, the variance matrix of X, is not Σ, but rather a function of (Σ, ν).

Formula (6) giving the weights is replaced by

$$w_k(i) = \frac{\nu^{(t)} + p}{\nu^{(t)} + \left(x_k - \mu_i^{(t)}\right)^2 / \sigma^{2(t)}}. \tag{11}$$

The sufficient statistics now include the cross-products of x_k, and equation (8) should be replaced by

$$\Sigma_i^{(t+1)}(p,q) = \frac{\sum_{k=1}^{N} \phi_k(i) w_k(i) \left(x_{k,p} - \mu_{i,p}^{2(t+1)}\right)\left(x_{k,q} - \mu_{i,q}^{2(t+1)}\right)}{\sum_{k=1}^{N} \phi_k(i)}. \quad (12)$$

Equation (10) giving the updated value of ν is replaced by

$$1 - \psi\left(\frac{\nu}{2}\right) + \log\left(\frac{\nu}{2}\right) + \frac{1}{N}\sum_{k=1}^{N} \log(\tilde{w}_k) - \tilde{w}_k + \frac{1}{N}\sum_{k=1}^{N} \psi\left(\frac{\nu+p}{2}\right) - \log\left(\frac{\nu+p}{2}\right) = 0. \quad (13)$$

The MCECM procedure outlined at the end of "Detailed Derivation" can be applied to identify targets in a multidimensional setting. Also, the procedure simply extends to the case of more than two states. If m is the number of states, the matrix Q is an $m \times m$ matrix and equation (3) is still valid. The bounds of the sum in equation (4) and (9) have been left out on purpose so that all the formulas hold irrespective of the number of states.

Definition of the Chromatin Types

Determining the Number of States

By compiling the DamID scores of the 53 profiles, the probed genome can be viewed as a cloud of points in a 53-dimensional space. In this space, two GATC fragments are close to each other if they are bound by the same proteins (irrespective of their genomic coordinates). Principal Component Analysis (PCA) gives the approximation of the cloud in lower dimensions that best preserves the overall spread (*i.e.* the variance) of the cloud. Each recurrent protein binding signature should appear as an individual lobe of the projected cloud. Importantly, one expects the projected coordinates to be distributed as a Student's *t* variable because they are weighted averages of the DamID binding scores, meaning that the projected cloud can be described by the multi-state multi-dimensional HMM presented above.

We performed scaled PCA on the 53 profiles and examined the projection of the cloud on the first principal components. It immediately appeared that at least three components are required to give a qualitative representation of the data, because a distinct subcloud separates in the third component. The projection on the first 3 principal components reveals 5 distinct lobes. Whereas the plot shown in Supplementary Figure 5B suggests that the fourth component might be included in the analysis, none of the 5 clouds forks into multiple sub-clouds in the fourth or higher order principal components. Therefore, we reduced the data to its first 3 principal components.

A substantial part of the variance in tiling array experiments is noise. For example, discretizing a DamID binding profile by a two-state HMM accounts for approximately 40% of the total variance (the median score out of 53 profiles is 39.4%). In comparison, the total variance captured by the first 3 principal components is 57.7%. This figure suggests that the procedure efficiently removes the technical variation from the dataset without causing a strong loss of information.

Mapping the Chromatin Types

Fitting an HMM to the original dataset would involve estimating 7,441 parameters (5 × 4 transition parameters, 5 × 53 means, 5 × 53 variances, 5 × 1374 covariances and 1 degree of freedom). In contrast, if the dataset is replaced by its projections on the first 3 principal components, the number of parameters drops to 66 (5 × 4 transition parameters, 5 × 3 means, 5 × 3 variances, 5 × 3 covariances and 1 degree of freedom), representing a substantial gain of robustness.

To map the chromatin types in the fly genome, we used the method described in "Multidimensional Inference" with 5 states. The initial degree of freedom was set to 6. The initial means of the modified Baum-Welch algorithm were determined visually. The initial covariance matrices were set to the covariance matrix of the 3-dimensional cloud. Note that within each lobe, the projections on the principal components may be correlated, even if their correlation is null on the whole dataset. Thus, after the first iteration the terms in equation (12) are not expected to be 0 for $p \neq q$.

The output of the modified Baum-Welch algorithm showed some sensitivity to initial parameter values, mainly to the means. For some initial values, the modified Baum-Welch algorithm failed to identify 5 clearly separated states. However, there was only one fitted model where the states were clearly separated, showing that the segmentation is robust to initial conditions.

The initial DamID binding values and final state calls are provided as supplementary files. An implementation of the procedure as an R package (R Development Core Team, 2009) is available upon request.

SUPPLEMENTAL REFERENCES

Bushey, A.M., Ramos, E., and Corces, V.G. (2009). Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions. Genes Dev. 23, 1338–1350.

Cappé, O., Moulines, E., and Rydén, T. (2005). Inference in Hidden Markov Models (Springer), pp. 369.

Chintapalli, V.R., Wang, J., and Dow, J.A. (2007). Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat. Genet. 39, 715–720.

Dempster, A.P., Laird, M., and Rubin, D.B. (1976). Maximum likelihood from Incomplete Data via the EM Algorithm. J.R. Stat. Soc., B 39, 1–38.

Engström, P.G., Ho Sui, S.J., Drivenes, O., Becker, T.S., and Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res. 17, 1898–1908.

Liu, C., and Rubin, D.B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. Statist. Sinica 5, 19–39.

R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. http://www.R-project.org

Slater, G.S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31.

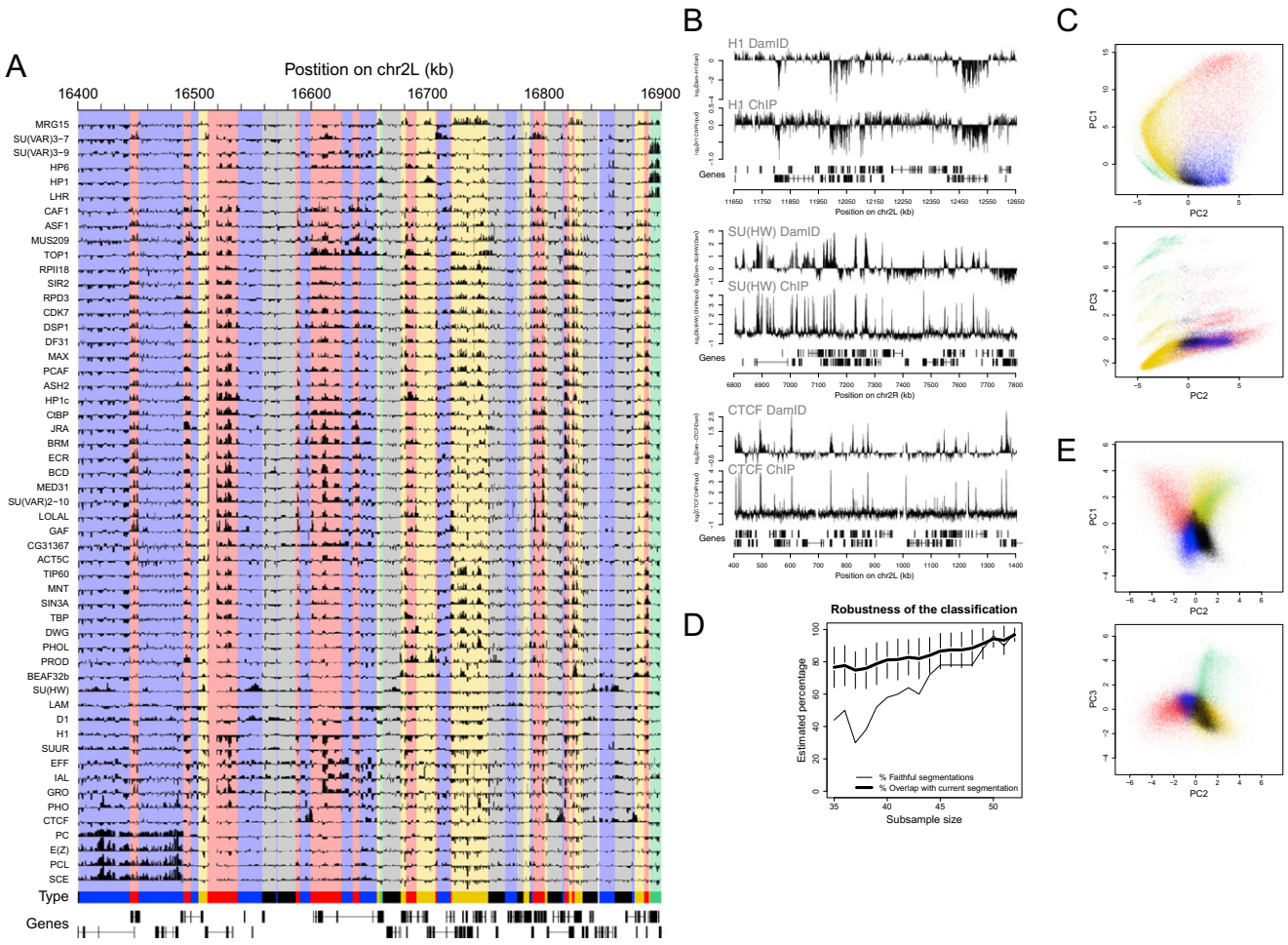

Figure S1. Validation of DamID and Robustness Analysis of the Five-Type segmentation, Related to Figure 1

(A) Magnification of the 53 DamID profiles as depicted in Figure 1. Traces show $\log_2$ DamID ratios, and width of black bars represents the length of DpnI restriction fragments.

(B) Comparisons of DamID (Braunschweig et al., 2009) and ChIP for histone H1 (this study), SU(HW), and CTCF (Bushey et al., 2009). All data sets were subjected to running mean smoothing with a window of three data points. Genes on the top and bottom strand are depicted as lines with blocks indicating exons.

(C) The five-state classification is robust to the quantification method. DamID profiles for each protein were binarized before projection onto the first three principal components. The shape of the cloud is different from the one shown in main Figure 1B, but the five types still form separated clouds in the first three principal components.

(D) Sensitivity analysis of the segmentation. The principal component analysis and HMM procedure was applied as in Figure 1 to data sets where one or more proteins were left out. It is expected that smaller sets will not always exhibit all five states; for example, a subset that lacks the four PcG proteins will not identify the BLUE state. In that case, the loci formerly assigned to BLUE will be assigned to another color by the HMM. We call such a segmentation "unfaithful." On the contrary, "faithful" segmentations show no replacement of a color by another. The percentage of overlap between the current segmentation (based on the complete data set) and unfaithful segmentations is irrelevant: it mostly reflects the size of the type that has been replaced. For example if BLUE has been replaced by another color at least 20% of the calls will differ (i.e., all the former BLUE calls). For subsets of 35–51 proteins, 50 samples were drawn at random without replacement. For subsets of 52 proteins, all possible 53 subsets were tested. For each subsample size, the percentage of faithful segmentations (thin lines) and their mean percentage of overlap with the current segmentation (thick lines) were determined. Vertical bars represent ± standard deviations. The plot shows that larger sets of proteins give an increasingly reliable discovery of the five states. Faithful segmentations show substantial agreement with the current definition, even for smaller sample sizes. Thus, identification of the states is sensitive to protein choice, but the classification itself is robust.

(E) Minimal set of proteins defining the five types. A carefully selected set of five proteins (histone H1, PC, HP1, MRG15, and BRM) summarizes the segmentation in 5 types. The data points were projected on their first three principal components, showing that the types are clearly visible with this minimal set of proteins. The coloring was obtained by applying the HMM to the first three principal components, exactly as was done for the 53 proteins. The agreement with the 53 profile segmentation is 85.5%.

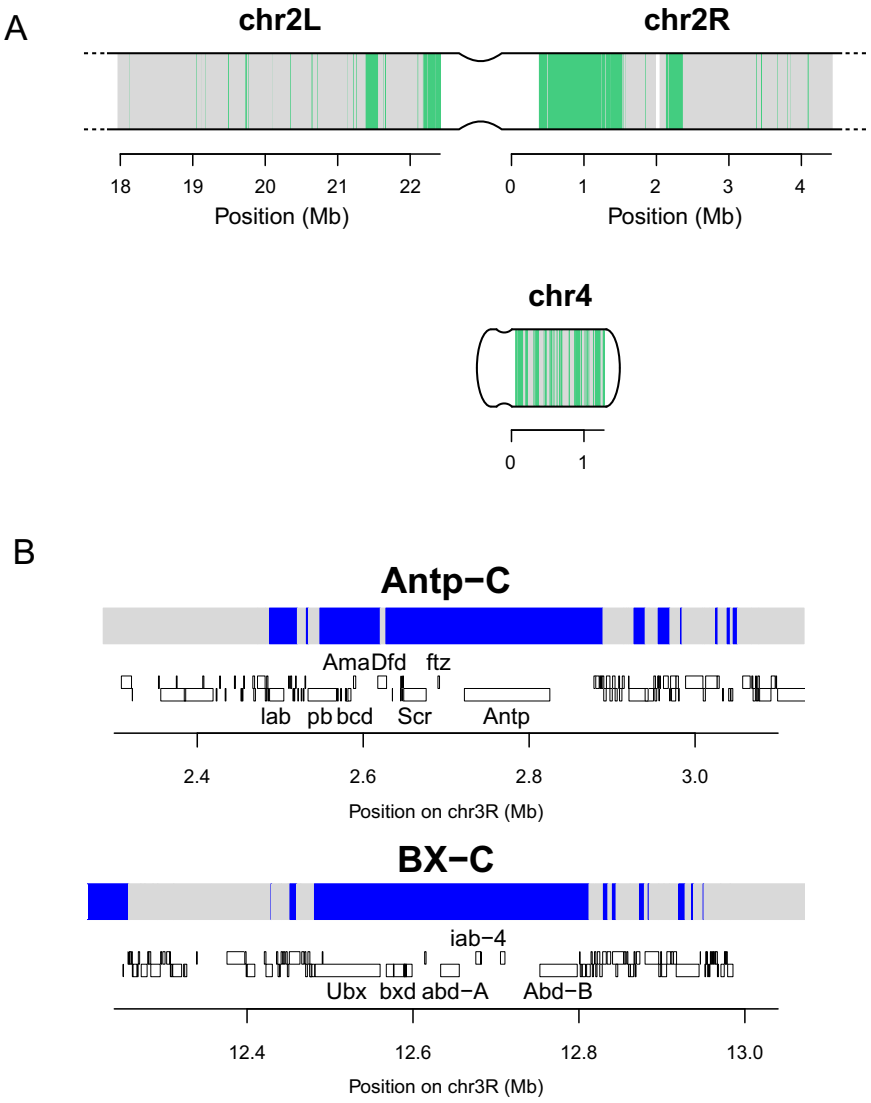

Figure S2. Localization of GREEN and BLUE Chromatin Supports Their Identity with Known Heterochromatin Types, Related to Figure 3

(A) Chromosomal maps of the pericentric region of chromosome 2 and of the entire chromosome 4. GREEN chromatin domains are shown, and other types are collectively represented in gray. Constrictions symbolize centromeres.

(B) Close-up view of the HOX gene clusters. The HOX genes are known to be Pc targets in Kc167 cells. Genes on the top and bottom strands are shown as boxes.

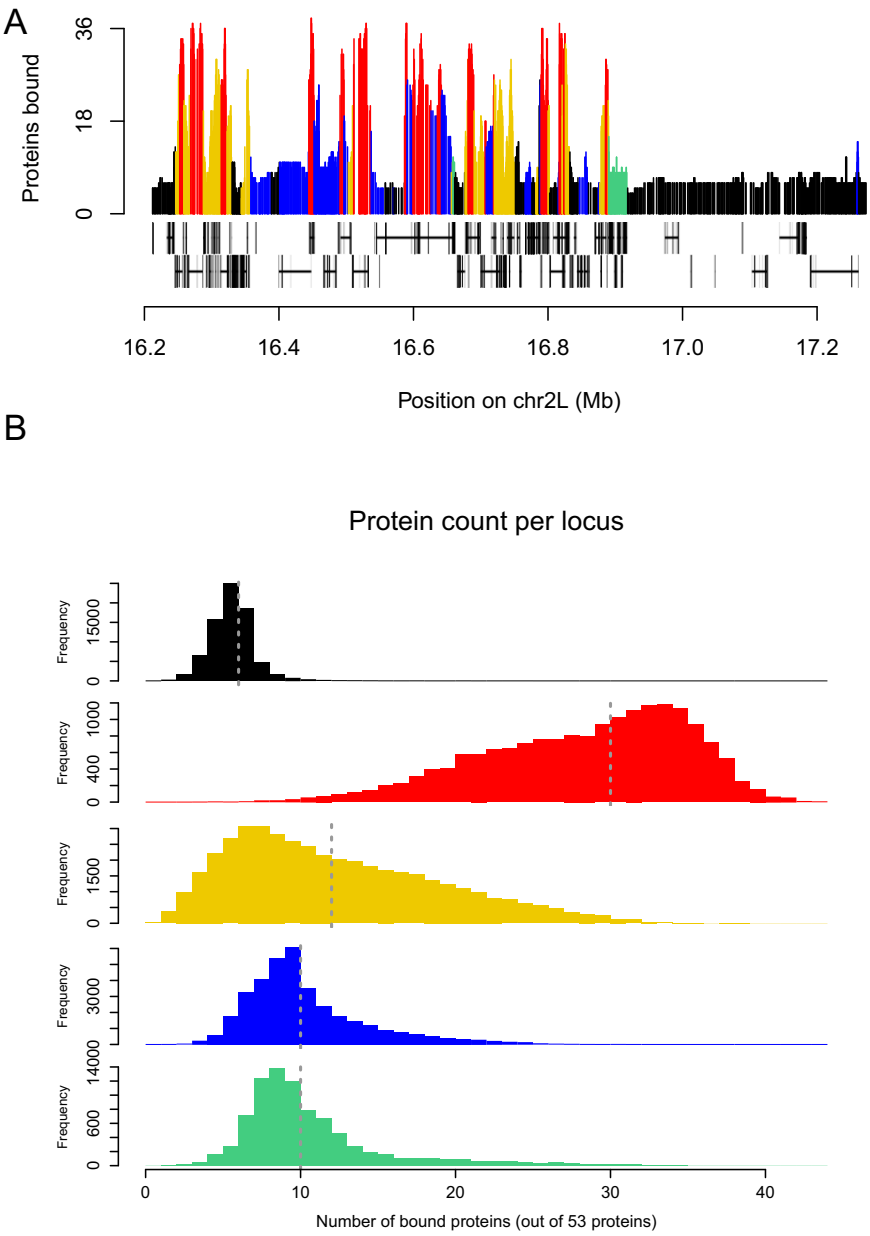

Figure S3. Chromatin Types Differ Widely in Their Total Protein Occupancy, Related to Figure 6

(A) Sample plot of the total occupancy (out of 53 mapped proteins) on a 1 Mb segment of chromosome 2L. The height of each vertical line indicates the number of proteins bound to a locus. The color of the line indicates the local chromatin type.

(B) Histograms of total occupancy distributions per chromatin type. Gray vertical dashed lines indicate median values.

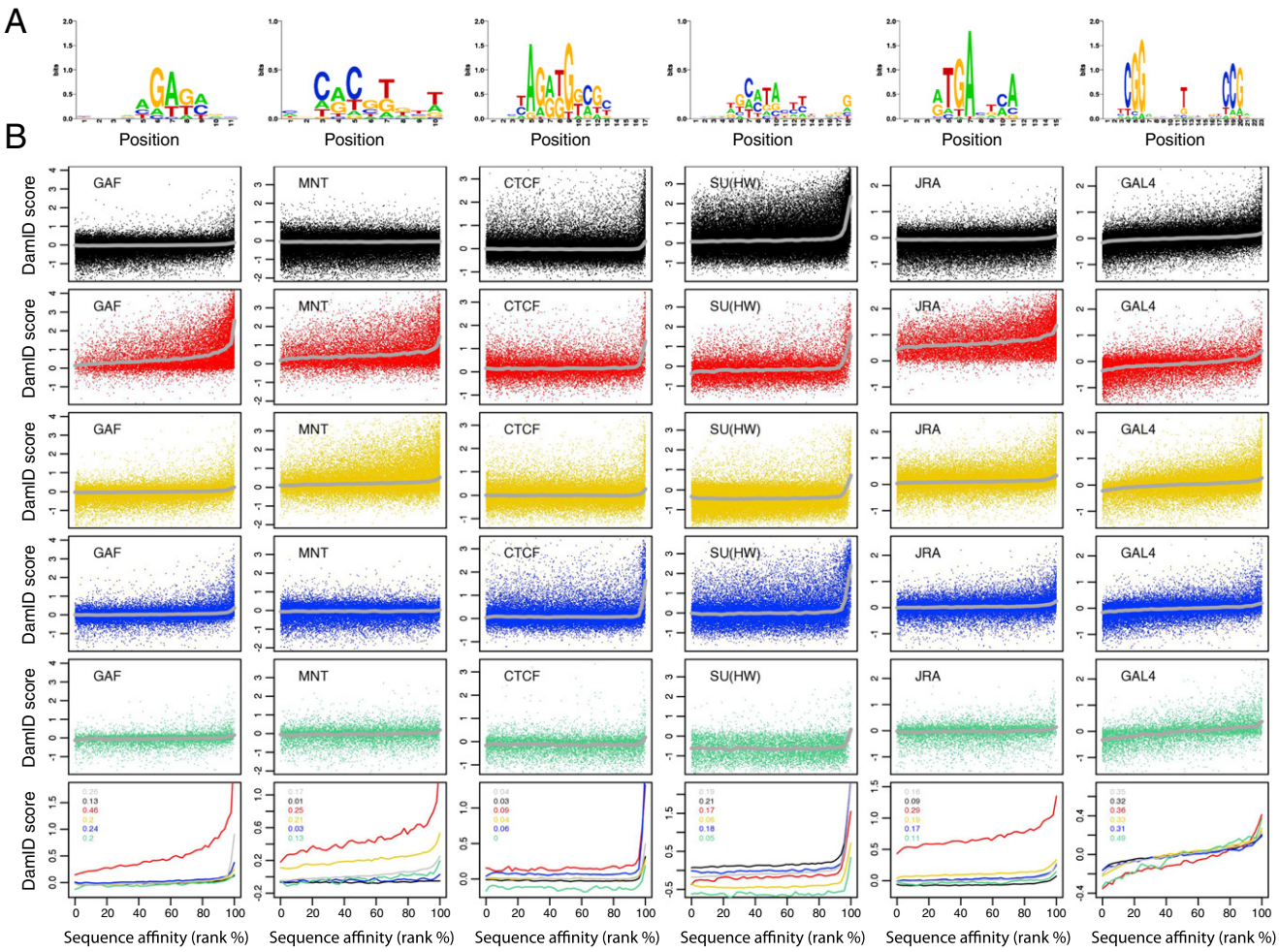

Figure S4. Chromatin Types Guide DBF Binding to Their Motifs, Related to Figure 7

(A) Optimized position-specific affinity matrices for 5 *Drosophila* DBFs and Gal4-DBD (see Extended Experimental Procedures for details).

(B) Distributions of relative affinites of GATC fragments mapping within each of the five types. For each chromatin type, DamID scores of GATC fragments ranked by the predicted affinity of a 2 kb window around its center. Gray lines represent the loess fit. Last row: superimposed loess fits for the five chromatin types. Colored numbers indicated the Spearman's rank correlation coefficients of DamID value with predicted affinities.

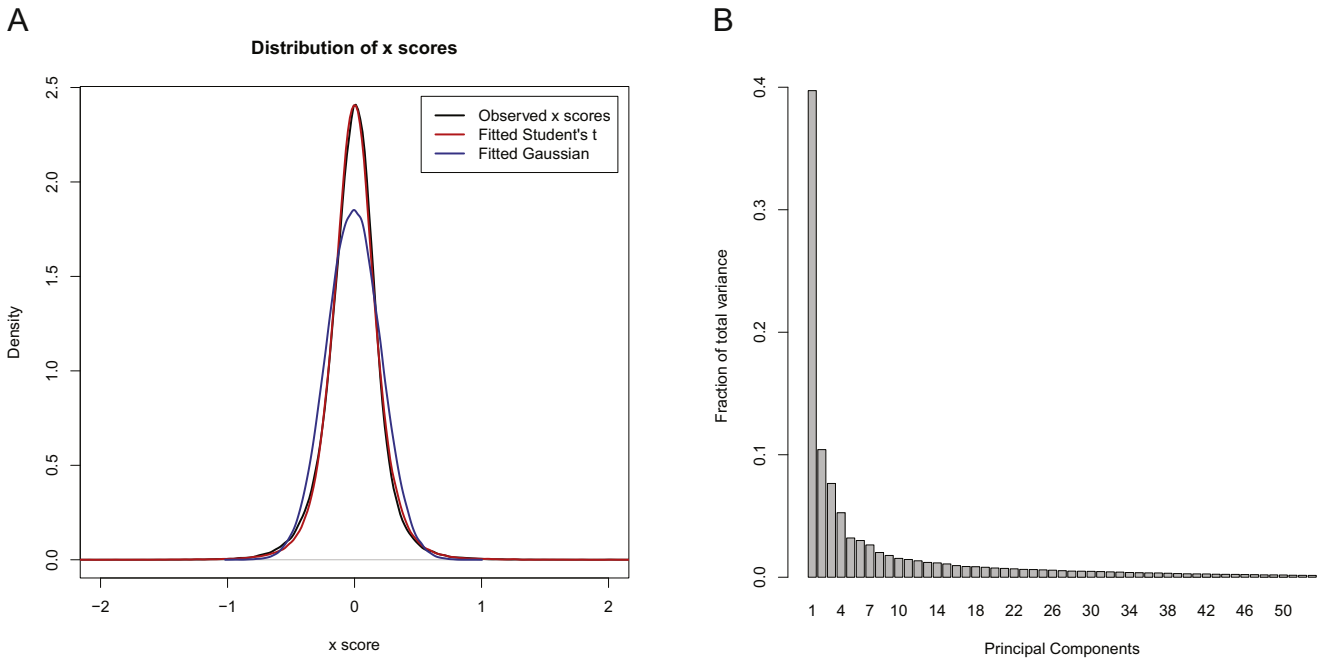

Figure S5. Addendum to the Analytical Methods, Related to Extended Experimental Procedures

(A) The scores x follow a Student's t distribution. The plot shows the density of loess-normalized and averaged per GATC fragments $\log_2$ ratios of the DamID profile of protein FBpp0074431 (black line). DamID for this protein shows no specific signal, and no targets were detected. The red line is the density of Student's t distribution fitted by maximum likelihood, and the blue line is the density of a fitted Gaussian distribution.

(B) Variance accounted for by individual principal components.