The genome-wide multi-layered architecture of chromosome pairing in early *Drosophila* embryos

Jelena Erceg\(^1,10\), Jumana AlHaj Abed\(^1,10\), Anton Goloborodko\(^2,10\), Bryan R. Lajoie\(^3,6\), Geoffrey Fudenberg\(^2,7\), Nezar Abdennur\(^2\), Maxim Imakaev\(^2\), Ruth B. McCole\(^1\), Son C. Nguyen\(^1,8\), Wren Saylor\(^1\), Eric F. Joyce\(^1,8\), T. Niroshini Senaratne\(^1,9\), Mohammed A. Hannan\(^1\), Guy Nir\(^1\), Job Dekker\(^3\), Leonid A. Mirny\(^2,4\) & C.-ting Wu\(^1,5\)

Genome organization involves cis and trans chromosomal interactions, both implicated in gene regulation, development, and disease. Here, we focus on trans interactions in *Drosophila*, where homologous chromosomes are paired in somatic cells from embryogenesis through adulthood. We first address long-standing questions regarding the structure of embryonic homolog pairing and, to this end, develop a haplotype-resolved Hi-C approach to minimize homolog misassignment and thus robustly distinguish trans-homolog from cis contacts. This computational approach, which we call Ohm, reveals pairing to be surprisingly structured genome-wide, with trans-homolog domains, compartments, and interaction peaks, many coinciding with analogous cis features. We also find a significant genome-wide correlation between pairing, transcription during zygotic genome activation, and binding of the pioneer factor Zelda. Our findings reveal a complex, highly structured organization underlying homolog pairing, first discovered a century ago in *Drosophila*. Finally, we demonstrate the versatility of our haplotype-resolved approach by applying it to mammalian embryos.
Although chromosomes are organized within the nucleus into distinct territories, they nevertheless come into contact in trans. In diploid organisms, including mammals, such trans contacts may involve specific interactions, such as pairing, between the homologous maternal and paternal chromosomes. Although best known in meiotic cells, homolog pairing can also occur in somatic cells and influence gene expression via phenomena such as transvection (reviewed by 1–3). In Drosophila, somatic pairing increases extensively from early embryogenesis to adulthood, making this organism an ideal system for studying trans interactions (reviewed by 1–3). What, for example, is the structure of homolog pairing in early embryos? Pairing may encompass many forms, ranging from a well-aligned juxtaposition of homologs (railroad track), to a loose, laissez-faire association of homologous regions, or even an apposition of highly disordered structures; for instance, pairing in mammalian systems can manifest as a nonrandom, approximate co-localization of homologous chromosomal regions (reviewed by 1–3).

Recently, high- and super-resolution imaging of several genomic loci has suggested that pairing can involve distinct chromosomal domains as well as more closely merged regions, while simulations of pairing via integration of the lamina-DamID genomic loci has suggested that pairing can involve distinct chromosomal domains (reviewed by 1–3). In our current study, we also explore the relationship between pairing and the developmental programs in play in the early embryo during the initial stages of pairing. We find that pairing correlates with the opening of chromatin mediated by the pioneer factor Zelda during zygotic genome activation.

### Results

**Haplotypel-resolved Hi-C for profiling embryonic pairing.** We focused our Hi-C analysis on the window of Drosophila embryogenesis extending from 2–4 h after egg laying (AEL), when zygotic gene transcription is activated and the embryo transitions from a syncytial blastoderm into a multicellular state2. It is during this window that maternal and paternal homologs come together for the first time3–13, and the rapid early cleavage cycles give way to the longer 13th and 14th cycles, thus minimizing the contribution of mitotic genome organization to our Hi-C signal33. To generate the embryos for our study, we mated two divergent Drosophila Genetic Reference Panel lines, 057 and 439 (Fig. 1a, b), as a high density of SNVs is required for assaying homolog pairing via genomic methods. Indeed, strains 057 and 439 differ with an average SNV frequency of ~5.5 per kilobase (kb), except on chromosome 4 (1.0 SNV/kb) (see Supplementary Note 1). This frequency was adjusted to ~5.1 SNVs/kb (0.01 for chr4) when we re-sequenced both lines and reconstructed an F1 diploid genome using only high-confidence homozygous SNVs (Supplementary Fig. 1 and Supplementary Table 1; see Methods); this diploid genome assembly ensured the most stringent haplotype-resolved mapping of Hi-C products. Importantly, we confirmed that hybrid embryos achieved levels of homolog pairing consistent with those observed in other studies13–17 using fluorescent in situ hybridization (FISH) to assess pairing (Fig. 1c) at heterochromatic (16.0% ± 5.7–22.0% ± 2.4) loci across different chromosomes (Fig. 1d, e). In addition, pairing increased to expected levels in 057/439 embryos as they aged (Supplementary Fig. 2a).

We performed Hi-C and recovered 513 million Hi-C products, which after mapping to the reference genome, yielded ~56% unique read pairs (Supplementary Table 2). When assembled into a map and analyzed at 1 kb resolution without regard to parental origin of the mapped fragments, these reads revealed features that are routinely seen in non-haplotype-resolved Hi-C maps33–34—a central cis diagonal representing short-distance interactions in cis (cis read pairs) as well as signatures for compartments, domains (also called contact domains33 and topologically associated domains, or TADs35,36, hereafter referred to as domains), and interaction peaks—thus confirming the quality of our Hi-C data (Supplementary Fig. 2b). Presence of these interphase hallmarks indicates that the vast majority of the cells are in interphase rather than mitosis, where such features are absent33.

**Ohm-directed distinction of trans-homolog from cis contacts.** Using our newly generated haplotype-resolved F1 diploid genome to determine the parental origin of our reads, we found that 5.8% of all mappable read pairs appeared to represent trans (trans-homolog as well as trans-heterolog) contacts and that, of these, 36.1% indicated contacts between homologous chromosomes (trans-homolog read pairs) (Supplementary Table 3). Encouraged by these numbers, we next determined the quality of the purported trans-homolog reads. For example, errors in reconstruction of the F1 diploid genome or sequencing of the Hi-C products can lead to misassignment of cis read pairs as trans-homolog read pairs (Fig. 2a) and, given the greater number of cis as versus trans-homolog read pairs, even a small rate of misassignment of cis read pairs could be confounding. Thus, to assess the potential magnitude of homolog...
misassignment, we sought a signature for it in our Hi-C dataset, focusing on read pairs where the separation along the genome (genomic separation) of the two reads of a read pair is <1 kb (Fig. 2b, shaded area). We chose to examine these read pairs because they are enriched in non-informative byproducts of the Hi-C protocol37, the most abundant being unligated pieces of DNA, because they are enriched in non-informative byproducts of the Hi-C protocol37, the most abundant being unligated pieces of DNA, as shown by dangled ends (Fig. 2b, hatched area and Supplementary Fig. 4a) and, since they are exclusively *cis* read pairs, their enrichment among *trans*-homolog read pairs can only arise via homolog misassignment (HM in figures and Methods). As such, the enrichment of short-distance (<1 kb) *trans*-homolog pairs can serve as a signature of homolog misassignment and be used to assess the overall quality of our homolog assignment (Methods).

Consistent with this hypothesis, this signature of homolog misassignment almost disappeared when we increased the stringency of our mapping by requiring at least two SNVs per read (Fig. 2c and Supplementary Fig. 4b, c, probability of homolog misassignment *P*~HM~ = 0.01%). In contrast, homolog misassignment increased when we used the less accurate DGRP SNV annotations (Supplementary Fig. 4d, *P*~HM~ = 2.4%). This approach, which we call Overview of homolog misassignment (Ohm), allowed us to estimate the probability of homolog misassignment as only ~0.17% when requiring at least 1 SNV per read (Fig. 2d, see Methods). Finally, this analysis indicated that homolog misassignment introduced less than 5% of erroneous *trans*-homolog contacts at separations above 1 kb (Fig. 2e). Having gained confidence that contamination contributes only a minor portion of purported *trans*-homolog read pairs, we observed that *trans*-homolog contacts at genomic separations of 1 kb are ~100-fold more frequent than those at genomic separations of >1 Mb, and ~500 fold more frequent than contacts between regions on different chromosomes (*trans*-heterolog, Fig. 2b). Thus, *trans*-homolog contacts constitute a major fraction of *trans*-chromosomal interactions.

**Homolog pairing in Drosophila embryos is highly structured.** We next generated a Hi-C map using ≥ 1 SNV per read and filtering our data to exclude contacts below 3 kb of genomic separation in order to substantially reduce contamination by Hi-C byproducts. The resulting map was striking in a number of ways (Supplementary Fig. 2b). In addition to displaying the expected central *cis* diagonal, it revealed prominent "*trans*-homolog" diagonals that extended across the entire length of each chromosome (Fig. 3a, black arrows). This observation demonstrated that interactions between homologs, including those within a few kilobases and up to several megabases, extend genome-wide in early embryos at the time when pairing is being initiated.

The *trans*-homolog diagonals suggest that homologs are relatively well-aligned when coming into contact, consistent with a railroad track structure (Fig. 3b). This indication, however, does not preclude less aligned forms of pairing, such as the paired domains observed at a number of genomic regions4,7. For example, Fig. 2b also shows that extensive *trans*-homolog interactions may occur between non-allelic regions corresponding to genomic separations of hundreds of kilobases or more, albeit at much reduced frequencies. We would expect such interactions to appear as signals at varying distances off the *trans*-homolog diagonals, leaving open the possibility of other structures, such as laissez-faire (Fig. 3c) and highly disordered pairing (Fig. 3d).
That pairing may assume different forms in *Drosophila* has long been considered (reviewed by1–3), with recent studies suggesting that the state of pairing is a fine balance between pairing and anti-pairing factors (reviewed by2). Here, we propose a framework in which the different aspects of pairing can be placed along the continuum of three axes (i) the precision with which homologous regions are aligned (x axis), (ii) the proximity with which homologs are held together (y axis), and (iii) the continuity or degree to which a particular state of pairing extends uninterrupted (z axis), with the continuum along any axis potentially reflecting different forms of pairing and/or the maturation (or degradation) of pairing over time (Fig. 3e).

In this context, high values along all three axes would produce railroad track pairing, intermediate values would approximate a laissez-faire mode of pairing, and minimum values would encompass interactions ranging from extreme laissez-faire to disorder. This framework can also be used to describe *trans* interactions between heterologous chromosomes, particularly prominent examples being those arising from the Rabl polarization of centromeres and telomeres, appearing in some Hi-C maps, including ours, as contacts along non-homologous arms (e.g. 2L and 2R, as well as 3L and 3R)33,34; the Rabl configuration may facilitate homolog pairing by reducing search space within the nucleus (Fig. 3f.h). Note, however, that *trans*-heterologous contacts were tighter and/or more frequent than *trans*-heterolog contacts (Fig. 3g, h), emphasizing the prevalence of homolog pairing in *Drosophila*.

We next asked whether our haplotype-resolved maps could further clarify the molecular structure of paired homologs as well as elucidate how *trans*-homolog interactions are integrated with the *cis* interactions that shape the 3D organization of the genome. Remarkably, we observed *trans*-homolog domains, domain boundaries, compartments, and interaction peaks resembling analogous features in *cis* maps (Fig. 3i–k and Supplementary Fig. 5a–e). In fact, 54% of the boundaries of *trans*-homolog domains overlapped a boundary of a *cis* domain (averaged over two homologs, Supplementary Fig. 5f). While this value is lower than the upper bound of 72% for *trans*-homolog boundaries shared between replicate haplotype-resolved Hi-C datasets, it is nevertheless significantly above the 35% overlap expected at random (P < 10−10, t-test). Note that, as we obtained 92.3% overlap of domain boundaries between replicates of non-haplotype-resolved Hi-C datasets, it is likely that the overlap values for haplotype-resolved datasets reflect the relatively low number of haplotype-resolved reads we obtained.

The similarity between *trans*-homolog and *cis* contact maps is further highlighted through comparisons with maternal- and paternal-specific Hi-C domains, which have been observed in other systems to be concordant21; in our dataset, we observed 73.9% concordance between *cis* domain boundaries of homologous chromosomes (Supplementary Fig. 5f; also Fig. 3i and Supplementary Fig. 5g, h), wherein concordance between two replicates was 73.2% (see Methods). *Trans*-homolog maps do,
**Fig. 3** Highly structured homolog pairing resembles features of cis-organization. **a** Haplotype-resolved Hi-C map of F1 hybrid embryos. Arrows, trans-homolog diagonals. **b–d** Homologs juxtaposed in a b railroad track fashion, c a loose, laissez-faire mode, or d a highly disordered structure. e Homolog pairing may encompass a range of structures defined by precision, proximity, and continuity. f Rabl positioning of chromosomes. The g zoomed-in maps of chromosomal arms (color coded boxes in a; L, left; R, right) and h respective contact frequencies as a function of genomic distance (see Methods). i trans-homolog, j cis maternal, and k cis paternal maps of matching regions on chr3L. l Ratio of cis Pat/cis Mat Hi-C maps indicates the cis contact patterns of two homologs are highly consistent. m The ratio of trans-homolog/average cis maps suggests that pairing resembles cis contacts, albeit with lower interactions in some regions (dark blue). n Homolog pairing displays highly structured trans-domains (black) and trans-boundaries (gray), reflecting cis-organization of homologs.
however, differ from somecis-defined features (Fig. 3n and Supplementary Fig. 5i, dark blue; see also31). Taken together, our findings indicate that homolog pairing goes far beyond simple genome-wide alignments of homologous pairs and includes surprisingly structured trans-homolog domains, boundaries, and specific interaction peaks that frequently correspond to cis features (Fig. 3n).

**Homolog pairing is associated with zygotic genome activation.**

Spurred on by the rich history of transvection (reviewed by1,2) we asked how broadly pairing might be correlated genome-wide with fundamental developmental events such as the binding of major transcription factors in the earlyDrosophila embryo. We focused our attention on arguably the four most prominently studied maternally-contributed factors, involved in zygotic genome activation, in particular key developmental events such as the opening of chromatin, initiation of transcription, and embryonic pattern formation: the pioneer factor Zelda (Zld), which mediates early chromatin accessibility38,39, Bicoid (Bcd)40 and Dorsal (Dl)41, which are involved in patterning the anterior-posterior (AP) and dorsal-ventral (DV) axis, respectively; and GAGA factor (GAF)42, which may facilitate transcription in later stages. We correlated the binding of these factors, as revealed by ChiP-seq datasets40,41,43, to local variations in the degree of pairing as characterized by the pairing score (PS), defined as the log2 trans-homolog contact frequency within a 28 kb window along the diagonal of Hi-C maps (Methods). Excitingly, regions associated with elevated PS values were significantly enriched for the binding of Zld and Bcd (P<10−10; Spearman; Fig. 4a) as compared to controls in which we randomized 200 kb chunks of PS values (Fig. 4b) or 200 kb chunks of binding profiles (Supplementary Fig. 6a). No correlation was observed, however, between PS values and the binding of GAF or Dl (P=0.139, P=0.029, Spearman, respectively; Fig. 4b). Then, using partial correlation analyses, we examined whether these correlations would be affected upon removal of the contribution of a third feature. For instance, how might the correlation between the binding of Bcd and PS values be affected by the binding of Zld? We found that Zld binding contributes significantly, as Spearman’s correlation coefficient dropped from 0.155 (P=2.1×10−65; Fig. 4b) to 0.044 (P=1.3×10−6; Supplementary Fig. 6b), when controlled for Zld binding. Interestingly, Spearman’s correlation coefficient between the binding of Dl and PS values decreased from −0.020 (P=0.029; Fig. 4b) to −0.157 (P=3.0×10−67; Supplementary Fig. 6b) when controlled for the contribution of Zld. These findings are consistent with a role of Zld in Bcd- as well as Dl-dependent binding and gene activation40,41. Interestingly, we observed that genes associated with both AP and DV patterning44,45 are similarly associated with higher PS values (Supplementary Fig. 6c). Moreover, loss of Bcd binding in Zld-depleted embryos40 was significantly enriched at regions otherwise associated with high PS values (P<1.21×10−68, Mood’s median test; Supplementary Fig. 6d, e). In sum, our data suggest that Zld-mediated early opening of chromatin, including the binding of Bcd, is strongly correlated with pairing in embryos.

As previous studies suggested that regions with high Zld occupancy were more dependent on Zld for establishing domain boundaries33, we asked whether Zld may play a role in domains that are established via trans-homolog interactions. Several observations supported our hypothesis. First, visual inspection of our Hi-C reveal that trans-homolog domain boundaries are associated with high Zld occupancy and, moreover, are associated with values of PS that are higher than those of the interior of domains (Fig. 4a and Supplementary Fig. 7a, P<10−10, Mood’s median test). Interestingly, 25% of the top Zld-occupied regions have high PS values (P<10−10, Mood’s median test, Fig. 4c). Second, using Hi-C data from Zld-depleted embryos33, we observed a decrease of boundary strength at regions that would otherwise be associated with high PS values and strong Zld binding (Fig. 4d and Supplementary Fig. 7b, P<10−10, Spearman). Third, we found that Zld depletion trended strongly toward reduced pairing, with the reduction being significant at four out of six loci assayed (P<2.56×10−3, Fisher’s two-tailed exact; Supplementary Fig. 7c, d). Finally, we also observed strong correlations between PS values and features known to be enriched at domain boundaries, such as RNA Pol II33, housekeeping genes35,46, and nascent zygotic gene products47 (Supplementary Fig. 6c, 7e). Thus, Zld-mediated chromatin accessibility and transcription appear correlated with the establishment of homolog pairing in the early embryo (Fig. 4e), reinforcing the notion that the establishment of homolog pairing may bear a relationship to genome function during key developmental events. For example, Zld-mediated opening of genomic regions may facilitate homologous sequences finding each other when pairing is initiated in early embryogenesis. Whether Zelda plays a direct or indirect role in initiating pairing, however, will require additional study, as will the interplay between Zld and other factors involved in pairing and other forms of chromosome organization.

**Haplotype-resolved application of Ohm to mammalian embryos.**

One of the most exciting developments in recent years has been the growing indication that homolog pairing is not special toDrosophila, that it may be a general property in many organisms, including mammals (reviewed by1,2). Thus, in order to both demonstrate the applicability of Ohm as well as assess the state of pairing in mammals, we applied our haplotype-resolved Ohm approach to mammalian embryos. As pairing in mammals is generally transient and localized, for example, during DNA repair, V(D)J recombination, imprinting, or X-inactivation (reviewed by1,2), these analyses would, at the least, also serve as a control for observations inDrosophila, where pairing is far more extensive. In particular, we considered two deeply sequenced Hi-C datasets from Du et al.25 and Ke et al.26 representing gametes and F1 hybrid mouse embryos bearing a high frequency of SNVs (1 autosomal SNV per ~130 bp or ~490 bp, respectively). Rea-soning that the haploid nature of sperm and oocytes precludes homolog pairing, we focused first on these cell types to determine the level of stringency that would be necessary to remove any apparent trans-homolog read pairs that would necessarily have resulted from homolog misassignment (Supplementary Fig. 8a). That level of stringency was ≥2 SNVs per read (Supplementary Fig. 8b and Supplementary Data 1, see Methods).

Requiring ≥2 SNVs per read, we turned to the datasets representing embryos ranging in age from the single, 2-, 4-, and 8-cell stage through the earliest days of embryogenesis and beyond. We observed no conclusive signal for trans-homolog contacts (Supplementary Figs. 9 and 10), consistent especially for the very earliest stages with previous indications that the maternal and paternal genomes are spatially segregated25,26,48–50, (Supplementary Fig. 9). More specifically, although the signal for trans-homolog contacts for some of the earlier stages seemed to hover above that expected from homolog misassignment (Supplementary Fig. 9) and thus raise the possibility of homolog pairing, this discordance between observed and expected trans-homolog contacts could also reflect artifacts arising from a non-uniform genomic density of SNVs and/or Hi-C biases (see37); we were unable to resolve this issue due to the sparsity of data (Supplementary Data 1). The data also raise the possibility of a higher order clustering of homologous as well as heterologous subtelomeric regions (Supplementary Fig. 10), reminiscent of the
Homolog pairing is related to Zld-mediated opening of chromatin. a Cis (upper panel) and trans-homolog (second panel from the top) contact maps of a 2 Mb region (2 R:16-18 Mb). Lower panels, pairing score (PS) calculated using a 28 kb window at 4 kb resolution (black), ChIP-seq profiles of Zld (dark purple), Bcd (green), GAF (blue), and DI (brown). Gray boxes, regions associated with elevated PS values in boundaries. b Correlation analyses between the PS values and the binding profiles of Zld, Bcd, GAF, and DI as determined using pairwise Spearman correlation. Control, randomized 200 kb chunks of the PS. Spearman correlation coefficients indicated in each box and by a heatmap; *P-values in parentheses. c 25% of the strongest Zld binding coincides significantly with high PS regions, compared to the remaining Zld binding (\(P < 10^{-10}\), Mood’s median test). d PS values vs loss of contact insulation upon Zld depletion, grouped by Zld binding strength quartile. Ovals represent the contours of 2D-Gaussian approximations, at a distance of 2 standard deviations from the mean (the dot). Regions with stronger Zld binding show both high PS values and increased loss of contact insulation upon Zld depletion. e Zld depletion affects insulation of domain boundaries at regions otherwise associated with high pairing. Thus, Zld may affect homologs to be further away from each other when insulation is lost.

**Discussion**

To conclude, our study addressed the fundamental nature of chromosome pairing and, to this end, developed a robust method, Ohm, for conducting haplotype-resolved Hi-C studies wherein interactions between homologous chromosomes are carefully vetted for homolog misassignment. This approach ensured the highest quality assignments of parental origin to Hi-C reads when distinguishing cis from trans contacts, and can be extended to any number of other methods.

Using Ohm, we obtained, for the first time, a genome-wide high-resolution view of homolog pairing during early Drosophila development. Our data revealed genome-wide juxtaposition of homologs along their entire lengths as well as a multi-layered organization of trans-homolog domains, domain boundaries, compartments, and interaction peaks. We also observed concordance of trans-homolog and cis- features, arguing that cis and trans interactions are structurally coordinated (see also). This striking correspondence of trans-homolog and cis domain boundaries and interaction peaks may suggest similar clustering of pericentromeric and subtelomeric regions seen previously. While, again, the sparsity of data precluded our ability to determine such clustering reflected any degree of homolog pairing, it is worth noting that telomere as well as centromere clustering has been reported in embryos across several species and may constitute a mechanism for initiating pairing.

In brief, although we found no definitive evidence of homolog pairing in early mouse embryos, our data nevertheless leave this issue open for future studies. Also, as repetitive sequences were excluded from our analyses, the potential of trans-homolog contacts at repeats remains to be explored.
mechanisms of their formation. For instance, SMC complexes implicated in loop extrusion\cite{44}, formation of cis domains, and cohesion of sister chromatids in mammals may play a role in pairing and formation of trans-homologous domains in Drosophila. We pursue this line of thought in our companion paper\cite{31}.

Moreover, by layering trans-homologous contacts on models of genome organization that have relied primarily, if not solely, on cis contacts, our findings highlight how haplotype-resolved Hi-C can shed light on paradigms of genome organization and function that would otherwise remain hidden. For example, pairing may underlie some of the structural and regulatory differences between the 3D architecture of the Drosophila genome and that of organisms lacking pairing; indeed, the close proximity of homologs may influence the manner in which loops and domains are formed in Drosophila in addition to playing a role in gene regulation\cite{33,155}.

Our observations have further revealed a correlation between pairing and Zelda-mediated chromatin accessibility during early Drosophila development, aging, and disease. Intrinsic high mobility of homologous chromosomes, which would fall in the plane of allelogous chromosomes, may be that, in some instances, only one of two alleles participates in heterologous interactions (reviewed by59).

Intriguingly, it may be that, in some instances, only one of two alleles participates in heterologous interactions (reviewed by59), and that pairing can play a potential role in gene regulation at some loci in mammals (reviewed by1,2). As such, it may be of significance that homologs are unpaired during the earliest moments of Drosophila embryogenesis\cite{36} and then initiate pairing at a time coinciding with activation of the zygotic genome. This in mind, we examined data pertaining to early mouse embryos; as in Drosophila, murine maternal and paternal genomes are segregated in the earliest cell cycles25,26,48,59, but little is known about the genome-wide status of pairing thereafter. Our analyses were further fueled by studies suggesting that pairing may be a general mechanism by which organisms detect structural heterozygosity in their genomes early in development, culling those that are deleteriously rearranged57. Although we observed some intriguing signals, ultimately, the sparsity of data did not allow us to draw strong conclusions regarding pairing in the mouse embryos, pointing to a need for additional haplotype-resolved Hi-C analyses and/or the resolution provided by imaging.

Our studies also suggest how pairing in Drosophila could shape early nuclear organization of the zygotic genome. For instance, as pairing brings similar loci together, it may enhance the compartmentalization of euchromatic and euchromatin structures during phase separation in early Drosophila embryos58. Alternatively, the segregation of heterochromatin from euchromatin could facilitate homology searches during the initiation of pairing. Indeed, how homologs identify each other remains one of the most intriguing questions. For instance, do homologs become structurally similar because of pairing, or do they have comparable conformations prior to pairing, their similarities contributing to homolog recognition? As our results do not reveal striking differences between the cis maps of homologs, they are consistent with the latter scenario. However, because our cis maps likely represent a mix of paired and unpaired homologs, potential differences between the cis maps of homologs may only become apparent in younger embryos prior to pairing.

Finally, we note that the sensitivity of Ohm has potential implications also for the analysis of trans contacts between heterochromatic chromosomes, which would fall in the plane of x = 0 in our framework of precision (x), proximity (y), and continuity (z). Intriguingly, it may be that, in some instances, only one of two alleles participates in heterologous interactions (reviewed by59), and thus, here, our haplotype-resolved approach could be applied to dissect the parental origin of the interacting alleles. Of course, when our framework is extended to encompass the dimension of time, it should then be able to capture the dynamics of trans-homolog and trans-heterolog contacts through cell division, development, aging, and disease.

**Methods**

**Selection of parental Drosophila lines.** To be able to distinguish homologs in sequencing data, we selected two parental lines with a high number of SNVs from Drosophila melanogaster Genetic Reference Panel (DGRP), which contains Drosophila lines inbred over 20 generations (see Supplementary Note 1).

**Genomic DNA isolation and library preparation.** 20 adult flies from each parental line (the Bloomington stock numbers 29652 and 29658) were homogenized with motor (Kimble-Kontes Pellet Pestle Cordless Motor) and pestle (Kimble-Kontes) on ice. Genomic DNA was isolated using DNeasy Blood & Tissue Kit (Qiagen) according to manufacturers’ recommendations. The integrity of the isolated genomic DNA (i.e. no degradation present) was verified by running a 1% agarose gel. Genomic DNA libraries were generated using Illumina TruSeq Nano DNA Library Preparation kit, and were 150 bp paired-end sequenced at the TUCF Genomics Facility using Illumina HiSeq2500.

**Collection and fixation of the F1 hybrid embryos.** The Drosophila handling was in accordance with the Harvard Medical School guidelines. The inbred DGRP-057 (maternal) and 439 (paternal) lines were mated to obtain the F1 hybrid embryos. To ensure that accurate parental genotypes were crossed, around 28,000 males and 28,000 virgin females were manually paired. The F1 hybrid embryos were collected and fixed after three pre-lays followed by aging for 2–4 h time-point at 25 °C. To validate that the embryos were of the correct stage, and did not contain contaminants from older embryos, a small aliquot (~100 embryos) per collection was set aside, devitellinised, and stored in methanol at −20 °C to verify developmental staging and cell type representation under microscopy. If even a single embryo was not aged, collection was discarded, as that embryo may contribute more nuclei than 2–4 h embryo, depending on exact developmental difference between them. The remaining embryos from collections were snap-frozen in liquid nitrogen and stored at −80 °C.

**In situ Hi-C on Drosophila F1 hybrid embryos.** Hi-C protocol from Rao et al.19 was adapted for the Drosophila embryos. 30 ng of snap-frozen fixed embryos were homogenized with pestle (Kimble-Kontes) in 500 µl of ice-cold lysing buffer, and incubated 30 min on ice. Upon cell lysis, the chromatin was digested using 500 U of DpnII restriction enzyme overnight at 37 °C. The DpnII overhangs were filled in with biotin mix during 1.5 h at 37 °C, and the chromatin was ligated for 4 h at 18 °C. RNA was degraded by adding 5 µl of 10 mg/ml RNase A (ThermoFisher Scientific) followed by incubation for 30 min at 37 °C. After degradation of proteins and crosslink reversal overnight at 68 °C, DNA was ethanol precipitated. DNA shearing was performed to obtain around 400–600 bp fragments with Qionga Instrument (Q800K) using the following parameters: 30 s on/off, 15x cycles, 70% ampl. After size-selection to around 500 bp, biotin pull-down was performed with a minor modification. Namely, the size-selected DNA was eluted in 200 µl of 10 mM Tris-HCl pH 8.0, and was mixed with equal volume of 2x Binding Buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 mM NaCl) containing pre-washed Dynabeads MyOne Streptavidin T1 beads (Life Technologies).

**Library preparation** was performed with the following modifications. The adapter ligation was in 46 µl of 1x Quick ligation reaction buffer (NEB), 2 µl of DNA Quick ligase (NEB), and 2 µl of NEXTflex DNA Barcode Adapter (25 µM, NEXTflex DNA Barcodes – 6, Bioo Scientific #514101). PCR reaction contained 23 µl of adaptor ligated DNA, 2 µl of NEXTflex Primer Mix (12.5 µM), and 25 µl of 15 Hot Start HiFi PCR Master Mix (NEB), and was run using the program: 98 °C, 30 s, (98 °C, 10 s, 65 °C, 75 s) repeated 7 times, 65 °C, 5 min, hold at 4 °C. Purification of PCR products was performed by incubation for 15 min instead of 5 min after each addition of Agencourt AMPure beads. The final libraries were eluted in 15 µl of 10 mM Tris-HCl pH 8.0 after 15 min incubation at room temperature, and 15 min incubation on a magnetic plate. The library quality was assessed using the High Sensitivity DNA assay on a 2100 Bioanalyzer system (Agilent Technologies). The libraries corresponding to two independent biological replicates were 150 bp paired-end sequenced at the TUCF Genomics Facility using Illumina HiSeq2500.

**FISH probes.** Oligo probes for heterochromatin repeats at chr2 (AACAC), chr3 (dodeca) and chrX (359) were synthesized by Integrated DNA Technologies (IDT) with the following sequences and a fluorescent dye: AACAC (Cy3-AAA CAACACAAACAAACAAACAAACAAACAAACAAACAAAC), dodeca (FAM488-ACGGG ACCAGTACGG), and 359 (Cy5-GGGATCGTTAGCACTGGTAATTAGCTGC). The libraries for euchromatin probes 89D-89E/RX-C (315 kb), 2R16 (904 kb), 3R1/C3 (744 kb), 3R1/C3 (700 kb), and 2R8 (680 kb) were designed using Oligopaint technology (Supplementary Table 4)\cite{71,62}, and were amplified using T7 amplification (see section Oligopaint probe synthesis) with forward primers containing a site for secondary oligo annealing, and reverse primers containing a T7 promoter sequence (sequences are provided in Supplementary Table 5). DNA probes were prepared together with primary oligos, using Oligopaints libraries were synthesized using T7 amplification followed by reverse transcription\cite{83}. Oligopaints libraries were first
amplified with Kapa Taq enzyme (Kapa Biosystems, 5 U/µl), and corresponding forward and reverse primers (without the site for secondary oligo annealing and the T7 promoter, without the unmapped sequences from Supplementary Table 5) using linear PCR program as follows: 95 °C, 5 min, (95 °C, 30 s, 58 °C, 30 s, 72 °C, 15 s) repeated 25 times, 72 °C, 5 min, hold at 4 °C. Upon clean up with DNA Clean & Concentrator-5 (Zymo Research), we then phased heterozygous PnM variants using bcftools isec. We picked only homozygous variants detected in the paternal DGRP-439 Drosophila line. Finally, we reconstructed the sequence of the F1 embryos with samtools consensus, using (a) the homozygous autosomal PnM SNVs, (b) the heterozygous autosomal PnM SNVs, and (c) the homozygous maternal chrX PnM SNVs as well as the homozygous SNVs detected in the paternal DGRP-439 Drosophila line.

**Mapping and parsing.** We started with sequences of Hi-C molecules and trimmed low-quality base pairs at both ends of each side using the standard mode of seqtk trimfq, v.1.2-r94 (https://github.com/lh3/seqtk). Then we mapped the trimmed sequences to the reference dm3 genome or the constructed dm3-based F1 diploid genome using bwa mem v.0.7.15 with flags -SP. We then extracted the coordinates of Hi-C contacts using the pairtools parse command line tool (https://github.com/minnlyahl/pairtools). We only kept read pairs that mapped uniquely to one of the two homologous chromosomes and removed PCR duplicates using the standard mode of the pairtools deDup command line tool.

**Contact frequency (P(s)) curves.** We used unique Hi-C pairs to calculate the functions of contact frequency P(s) to genomic separations. We grouped genomic distances into 10 bp and 100 kb bins into ranges of exponentially increasing widths, with 8 ranges per order of magnitude. For every range of separations, we found the number of observed cis- or trans-homolog interactions within this range of separations and divided it by the total number of all loci pairs separated by such distance.

**Estimating the rate of Hi-C byproducts.** Each Hi-C dataset contains non-informative byproducts, i.e. sequenced DNA molecules that originate from a single DNA fragment, and thus do not carry any information on chromatin conformation (unlike informative Hi-C molecules that form via a ligation of two spatially proximal DNA fragments and capture a spatial contact) (Supplementary Fig. 3a, b). The two well-known types of Hi-C byproducts are: (i) unligated DNA fragments (termed “dangling ends”) and (ii) self-ligated DNA fragments ("self-circles") 37. "Dangling ends" are small pieces of intact unligated DNA that passed through all selection steps of the Hi-C protocol and ended up in the library of sequenced DNA (Supplementary Fig. 3b). "Dangling ends" appear as cis read pairs at s ≪ 100–1000 bp (such molecules are left in the library after the size selection step of the Hi-C protocol), with the direction of the two reads of a read pair pointing towards each other along the reference genome. "Self-circles" are pieces of DNA whose ends were ligated to each other and the resulting circular DNA had a break in another location, producing a linear piece of DNA (Supplementary Fig. 3b). “Self-circles” also appear in cis, but at longer separations, up to a few kb (this distance depends on the frequency of the restrictions sites along the genome) and the resulting paired read orientations point away from each other along the reference genome.

**The construction of F1 diploid Drosophila genome.** We sequenced the two inbred parental Drosophila lines (DGRP-057 and DGRP-439) together with the hybrid PnM cell line31 at the average coverage of 118, 117, and 396 reads per base pair, respectively.44. We then detected the sequence variation of these three libraries using bcf-tools. We picked high-confidence variants on the maternal autosomes by selecting heterozygous PnM variants that were present among maternal DGRP-057 variants and absent among paternal DGRP-439 sequence variants (both homo- and heterozygous); we kept-confidence paternal variants phasing was selected in an opposite manner. Since PnM is a male line, for the maternal copy of chrX in F1 embryos, we considered only homozygous high-quality variants detected in the PnM cell line. To reconstruct the consensus sequence of the paternal copy of chrX in F1 embryos, we kept only homoyzygous variants detected in the paternal DGRP-439 Drosophila line. Finally, we reconstructed the sequence of the F1 embryos with samtools consensus, using (a) the homozygous autosomal PnM SNVs, (b) the heterozygous autosomal PnM SNVs, and (c) the homozygous maternal chrX PnM SNVs as well as the homozygous SNVs detected in the paternal DGRP-439 Drosophila line.
while the outer, 5′ fraction overlapped the inner fraction, but went into the opposite direction (Supplementary Fig. 3d). This non-trivial alignment pattern suggested a scenario, where, during some stage of the Hi-C protocol, a DNA molecule formed a hairpin and was extended using itself as a template (Supplementary Fig. 3e). During the work on this manuscript, this type of a Hi-C byproduct was independently reported and thoroughly characterized in66, who dubbed it as “hairpin loops”. Another, older study reported formation of similar hairpin loops in whole-genome sequencing of ancient DNA68.

The mechanism by which such “hairpin loops” form remains to be firmly established. Golosli et al.66 proposed that these “hairpin loops” were formed during the end repair stage of the Hi-C protocol via self-annealing of a pre-existing byproduct. Golloshi et al.66 suggested a scenario, where, during some stage of the Hi-C protocol, a DNA byproduct and their presence does not significantly reduce the number and identity of contacts extracted from a Hi-C library.

We produced an upper estimate of the amount of byproduct types in our sample, assuming that the true P(s) followed the same of the same-strand read pairs at s > 1 kb, and was constant at s < 1 kb. Using this assumption, we found that “dangling ends” constituted 39.5% of mapped and phased reads in our Hi-C library, “self-circles” were 0.3%, and “hairpin loops” formed from “dangling ends” made up 2.9% of the library.

Binning and balancing of Hi-C data. We aggregated unique Hi-C pairs into genomic bins of 1 kb and larger, using the cooler package (https://github.com/mirnylab/cooler). Low signal bins were excluded prior to balancing using the MADmax filter: we removed all bins, whose coverage was 7 genome-wide median deviations below the median bin coverage. Additionally, we removed all cis- and trans-homolog contact separations below 3 kb, since these read pairs were dominated by non-informative Hi-C artifacts, unligated DNA fragments and ligation sites formed via self-circularization of DNA fragments68. We then balanced the obtained contact matrices via iterative correction (IC), i.e. equalized the sum of contacts in every row/column to 1.0.

We calculated observed/expected contact frequency maps (often abbreviated as O/E CE) in cis and trans-homolog by dividing each diagonal of an IC contact matrix by its chromosome-wide average value over non-filtered genomic bins.

The definition of homolog misassignment. When we aligned a read to a diploid genome, we assigned to it three bits of information: which pair of homologous chromosomes this read originated from; the allele, i.e., which of the two homologs it originated from; and its location along the chromosome. Importantly, we could assign the allele (or, the homolog) to a read only if it overlapped some sequence that is unique to one homolog, i.e. if it overlapped a SNV; otherwise, it was impossible to tell which of the two homologs the read originated from. Even in the Drosophila and mouse models with the highest density of SNVs, the sequences of the two homologous chromosomes were still highly similar and contained only one SNV per 100–200 bp. Because of this, the majority of read pairs (150 bp) lacked allele assignment on one or both sides; and the allele assignment for the rest of pairs was based on one or a few SNVs, and thus was prone to errors.

Sources of homolog misassignment. The errors in allele assignment (below, homolog misassignment, HM) occur for two reasons: (A) due to an error in the sequencing that caused a read, or (B) due to an error in the genome. Below we discussed how such errors lead to HM. Also, note that we only discussed the case where diploid genome sequences only contained SNVs and did not analyze the effect of insertions, deletions and duplications.

(A) The type and rate of sequencing errors varies highly among different sequencing platforms69. In the case of the most popular short-read platform, Illumina, sequencing errors typically lead to substitutions at the rate around 0.1% per base pair57. Such a substitution could occasionally lead to a HM, if it occurs at a SNV site and the erroneously called base pair by chance matches the other SNV allele.

(B) Some types of errors in the sequence of a diploid genome may lead to HM. The possible cases are:

(B1) an error at a site without a SNV (e.g. the true sequence is A in both alleles, which we wrote as A-A, but in the genome database it is T-T)—in this case, reads coming from either homolog would contain a mismatch with the database. Such error does not lead to HM.

(B2) a missing SNV (e.g. true A-T, but in the database it is A-A). Such an error leads to mapping imbalance, i.e. reads from the homolog with the correct sequence were mapped, while the reads from a homolog with a sequence error will be dropped. A missing SNV does not lead to HM.

(B3) a false positive SNV (e.g. true A-A, database A-T). In this case, reads from both homologs will be assigned to the same allele (the one matching the true sequence). A false positive SNV error leads to both HM and mapping imbalance.

(B4) an erroneous SNV (e.g. true A-T, database A-C). In this case, reads from the homolog containing an error will not be mapped, which will produce mapping imbalance. An erroneous SNV does not lead to HM.

(B5) missing heterozygosity of an SNV allele. Samples derived from populations of organisms can contain heterozygous SNVs, i.e. different organisms within the population may contain different base pairs at the same location on the same homolog. The issue, however, is that genomic databases can store only one allele per homolog (below, referred to as a “known” allele), while the other alleles become ignored (referred to as “unknown” alleles) (note that this issue can be solved by using graph genomes, which can store multiple overlapping variants per position57). The effects of heterozygosity on mapping depends on the nature of “known” and “unknown” alleles on the two homologs:

(B5.1) non-overlapping “known” and “unknown” alleles, when all alleles of one homolog are different from the allele on the other homolog (e.g., true A/C-T, database A-T), this scenario leads to a partial mapping imbalance, because the reads containing the “unknown” allele do not get mapped. This does not lead to HM.

(B5.2) overlapping “known” and “unknown” alleles, when the unknown allele of one homolog is the same as the “known” allele on the other homolog (e.g. true A/T, database A-T); this scenario leads to HM, because all reads containing the “unknown” allele will be interpreted as containing the “known” allele from the other homolog. Also, this scenario leads to a mapping imbalance.

Misassigned cis pairs can contaminate trans-homolog signal. HM is particularly problematic for studying the trans-homolog contact patterns. If we would assign a wrong homolog to one side of a cis Hi-C pair, we would misinterpret this pair as a trans-homolog contact.

Given that trans-homolog contacts in our sample were much less frequent than cis ones, even a small probability of HM could lead to a significant contamination of the trans-homolog contact map. Note, that a cross-contamination between the cis maps of the two homologs was much less likely, since it required simultaneous HM on both sides of a Hi-C pair.

Estimating homolog misassignment using Hi-C byproducts. The observed trans-homolog read pairs thus was a mixture of “true” trans-homolog contacts and cis pairs with a misassigned homolog on one side. For P(s) curves, this could be expressed as:

\[ p_{\text{trans-homolog}}(s) = \frac{1 - \Phi(s)}{\Phi(s)} = p_{\text{trans-homolog}_{\text{true}}}(s) + p_{\text{PHM}} + p_{\text{cis}}(s), \]

where \( p_{\text{trans-homolog}}(s) \) was the P(s) curve for the observed trans-homolog contacts, \( p_{\text{trans-homolog}_{\text{true}}}(s) \) was for the true trans-homolog contacts (i.e. Hi-C trans-homolog pairs that represented a true trans-homolog ligation) and \( p_{\text{cis}}(s) \) was for cis contacts, and \( p_{\text{PHM}} \) was the probability of HM. Here, the probability of HM was defined per paired read (i.e. the probability per side is \( p_{\text{PHM}/2} \)); we assumed that each Hi-C pair had the same chance to get an erroneously assigned homolog, regardless of its genomic separation or orientation, i.e. that HM probability was the same for all Hi-C pairs. We also accounted for HM in trans-homolog read pairs, which were falsely interpreted as cis ones.

Estimating the probability of HM in a mapped Hi-C library was a challenging task. A primary problem was that, given trans-homolog read pair, we cannot tell if it was produced by a true trans-homolog ligation, or it resulted from HM of a cis pair. However, we could reduce the relative amount of homolog-misassigned pairs by increasing the stringency of homolog assignment. For that purpose, we selected pairs where both sides overlapped at least two or three SNVs and where read alignments had no mismatches with respect to the reference genome. We reasoned that, for such pairs, HM was less likely, since it required simultaneous errors at multiple SNVs.

For pairs with 2 + SNVs on each side, the shape of \( p_{\text{trans-homolog}}(s) \) changed at separations <1 kb: the −10-fold enrichment of trans-homolog inward pairs at s < 1 kb that was present in unfiltered data (Supplementary Fig. 4a), was greatly reduced after increasing the stringency of homolog assignment (Supplementary Fig. 4b). We tried increasing the stringency of homolog assignment further by requiring 3 + SNVs on each side, but this did not change the shape of \( p_{\text{trans-homolog}}(s) \) (Supplementary Fig. 4c). This observation suggested that the homolog-misassigned cis pairs did not significantly contribute to \( p_{\text{trans-homolog}}(s) \) for pairs with 2 + SNVs. Conversely, increasing the mapping of homolog assignments to raw data to less accurate DGRP SNV annotations (see Supplementary Note 1), increased the abundance of observed short-distance inward trans-homolog pairs (Supplementary Fig. 4d).
Fig. 4d). We concluded that the enrichment of short-distance inward trans-homolog pairs was due to homolog-misassigned inward cis pairs. Toward trans-homolog pairs, the enrichment was particularly sensitive to contamination by homolog-misassigned cis pairs, because at these separations and directionalities cis pairs were the most abundant (>104 more abundant that trans-homolog pairs of other directionalities) due to “dangling ends”. Thus, even rare HM at a probability as low as ~10−6 would generate enough false trans-homolog pairs to match the true signal, and thus change the shape of inward Ptrans-homolog(s) at < 1 kb. Thus, the ratio of Ptrans-homolog(s)/Pcis(s) for inward pairs at short separations could be used as an upper estimate for the probability of HM:

\[ \frac{P_{\text{HM}}}{P_{\text{cis}(s)}} \approx \frac{P_{\text{trans-homolog}(s)}/P_{\text{cis}(s)}}{P_{\text{cis}(s)}} \]

We tested our OHM approach to haplotype-resolved Hi-C using data from two published haplotype-resolved Hi-C studies on mice25,26. For mapping, we reconstructed two sets of diploid genomes (Blackx x DBA/2 and Blackx x PWK/Ph) using SNVs from the Mouse Genome Project71. The genome reconstructions did not require reconstruction, since it served as the basis for the reference mm10/GRCm38 genome. We reconstructed consensus genomes with bcftools consensus, using only homozygous SNVs that passed the quality control. We then mapped the publicly available sequenced Hi-C libraries to the diploid genomes (Blackx x DBA/2 for25 and Blackx x PWK/Ph) for26 and extracted the positions of Hi-C contacts using the same approach as for our Drosophila data. For contact maps, we only used Hi-C pairs overlapping 2 ± 3 SNVs on each side. We did not perform IC on the produced contacts maps due to their sparsity. Non-IcEd raw contact maps have to be interpreted with a degree of caution, since they are affected by the variation of sequencing visibility of loci due to an uneven distribution of SNVs, GC content, amount of open chromatin, etc.

In a naive interpretation, a trans of trans-homolog contact frequency Ptrans-homolog(s) with distance s is indicative of homolog pairing, since it means that pairs of loci in homologous positions make more contacts than loci in non-homologous positions. However, such decay can also be observed in samples without pairs of loci in homologous positions. To interpret the curves of trans-homolog contact frequency Ptrans-homolog(s), we compared them to a negative control, no-pairing curves Ptrans-homolog(s)-no-pairing, which described the expected amount of observed trans-homolog contacts in samples without pairing, but in presence of HM. Ps trans-homolog(s)-no-pairing could be derived from Eq. (1) assuming Ptrans-homolog(s)-no-pairing const:

\[ P_{\text{trans-homolog}(s)\text{-no-pairing}} \approx \text{avg. cross} - \text{parent} \times \text{heterolog CF} + P_{\text{HM}} \times P_{\text{cis}(s)} \]

here, we could estimate POM using Eq. (2). Thus, in order to conclude that a sample had homolog pairing, we had to observe significantly more contacts than that predicted by Ps trans-homolog(s)-no-pairing, at least in some range of separations.

Analysis of published haplotype-resolved mouse Hi-C data. To compare our data with other published data, we downloaded the structure of chromosomes in Drosophila embryos, we re-analysed the publicly available Hi-C dataset from33 using the same methods as we used for our own data.

Pairing score. To characterize the degree of pairing between homologous loci across the whole genome, we introduced a genome-wide statistics track called pairing score (PS). The PS of a genomic bin is log2 of average trans-homolog IC contact frequency between all pairs of bins within a window of ±W bins. For each genomic bin i, its pairing score with window size W was defined as:

\[ P_{\text{PS}(i)} = \log_2(CF_{\text{pair}}) \]

averaged over bins m and n between i-W-th and i+W-th genomic bins on different homologs of the same chromosome.

Importantly, under this definition, the PS quantified only contacts between homologous loci and their close neighbors and did not quantify pairing between non-homologous loci on homologous chromosomes.

The choice of the window size W was guided by the balance between specificity and sensitivity. Using a bigger window increased sensitivity, accumulating contacts across more loci pairs, while smaller windows increased specificity, allowing to see smaller-scale variances in homolog pairing; empirically, we found that, for our contact maps binned at 4 kb resolution, using a 7 x 7 bin window (W = 3) provided the optimal balance between specificity and sensitivity.

A close examination of the PS track revealed two important features: i) the Drosophila genome was divided into regions that demonstrate consistently high, relatively low, or values of PS, followed by extended regions where PS dipped into lower values, ii) switching between high- and low-PS regions seemed to occur around insulating boundaries.

Insulation scores. For every bin of a contact map binned at 4 kb resolution, we calculated the insulation score as the total number of normalized and filtered contacts formed across that bin by pairs of bins located on the either side, up to 5 bins (20 kb) away for Hi-C maps to reference dm3 maps (40 kb) away for haplotype-resolved Hi-C maps (see Supplementary Note 1). We then normalized the score by its genome-wide median. To find insulating boundaries, we detected all local minima and maxima in the log2-transformed and then characterized them by their prominence (Billauer E. peaklet: Peak detection using MATLAB, https://github.com/ENCODE-DCG-chipseq-pipeline). The detected minima insulation score corresponding to a local depletion of contacts across the genomic bin, were then called as insulating boundaries. We found empirically that the distribution of log-prominence of boundaries had a bimodal shape, and we selected all boundaries in the high-prominence mode above a prominence cutoff of 0.1 for Hi-C mapped to the reference dm3 map, and a cutoff of 0.3 for haplotype-resolved Hi-C maps. We called the insulating boundaries in trans-homolog contact maps using the same technique, requiring a minimal prominence of 0.3. Finally, we removed boundaries that were adjacent to the genomic bins that were masked out during IC.

To estimate the similarity of insulating boundaries detected in the cis and trans-homolog contact maps, we calculated the number of overlapping boundaries with each other. We allowed for a mismatch up to four genomic 4 kb bins (16 kb total) between overlapping boundaries to account for the drift caused by the stochasticity of contact maps. We reported average percentage of overlapping boundaries. For instance, the average fraction (73.9%) that overlapping boundaries between cis maternal and cis paternal boundaries occupied within cis maternal only (74.2%), and cis paternal only boundaries (73.7%). Similarly, for replicates, we calculated the average percentage of overlapping boundaries for cis maternal replicates (75.4%), and also for cis paternal replicates (71.1%), and then provided a mean of the two values (73.2%). In reference dm3 Hi-C contact maps, the percentage of overlapping boundaries between the two replicates reached 92.3%, suggesting that the lower reproducibility of haplotype-resolved boundaries is due to technical and not biological variation between the two replicates.

We estimated the significance of an overlap between two given boundary sets by comparing it to the overlap expected by chance alone, given the sizes of the two sets. We estimated an overlap using a binomial distribution, sampled random subsets of genomic bins visible in our Hi-C maps. To estimate how significantly the observed overlap deviates from the random expectation, we repeated randomization 10 times and compared the random overlaps to the observed overlap using a t-test.

ChiP-seq. We mapped the publicly available raw ChIP-seq data following the same procedure as used by the ENCODE consortium27, (https://github.com/ENCODE-DCG-chipseq-pipeline). To calculate the number of overlapping ChIP-seq peaks, we re-sized all peaks to 1 kb around their centers and used pyBedTools to calculate peak intersections27.

To generate the tracks of Zbt binding, we used ChIP datasets GSM1596215 and GSM1596219 and input datasets GSM1596216 and GSM1596220 from study GSE654411. For DI, we used ChIP datasets GSM1596223 and GSM1596227 and input datasets GSM1596224 and GSM1596228 from the same study GSE654411. For GAF, we used ChIP datasets GSM614652 and input datasets GSM614653 and GSM614654 from study GSE235374. For wild-type Bcd, we used ChIP datasets GSM1332670 and GSM1332671 and input datasets GSM1332672 and GSM1332673 from GSE552564. For Bcd in zld mutants, we used ChIP datasets GSM1332674 and GSM1332675 and input datasets GSM1332676 and GSM1332677 from GSE552564. Finally, for RNA Pol II, we used ChIP datasets GSM1596231 and GSM1596235 and input datasets GSM1596232 and GSM1596236 from GSE654411.

GRO-seq. We mapped the publicly available GRO-seq datasets (GSM1020093 and GSM1020094 from GSE416117, using STAR24 following the same procedure as used by the ENCODE consortium, (https://github.com/ENCODE-DCG-long-rna-seq-pipeline/tree/master/dnanexus).

Correlation analyses between the PS and ChIP or GRO-seq. The correlation analyses were performed by first dividing the genome into 10 kb bins. In each bin, the fraction of sequence occupied by the PA, the respective ChIP or GRO-seq signal was determined. Then, the enrichment of bin-genome-wide PS was calculated between the PA values and a ChIP or GRO-seq signal of interest. The strength of correlation was reported using Spearman correlation coefficients and corresponding P-values in two flavors, either for a pairwise comparison between two features, or as a part of analyses that examined whether the correlation between the two features was affected by co-correlation with a third feature. The Spearman correlation coefficients were also displayed using a heatmap. Control regions consisted of the randomized PS or the binding profiles for Zbt, Bcd, GAF, and DI.
where each respective feature was chunked into blocks of 200 kb size, that were randomly moved around.

**Correlation between the PS and Zld binding.** To visualize the relationship between PS and Zld binding, we plotted the distribution of PS for genomic bins from each of the four quartiles of the genome-wide distribution of Zld ChIP-seq signal.

We then tested if the effects of Zld binding on chromatin conformation were correlated with the pairing status of loci. We re-analyzed the published Hi-C datasets on Zld depletion with the same computational methods as above and obtained normalized Hi-C maps at 4 kb resolution for WT and Zld-depleted embryos at nuclear cycle 14. Then, we calculated the change of the 20-kb window insulation score upon Zld depletion, \( \Delta_{\text{insZld}} \), and compared it to PS. For bins from each quartile of Zld ChIP-seq binding, we fitted the resulting 2D-distribution of \( \Delta_{\text{insZld}} \) vs PS with a bivariate Gaussian. We then plotted the resulting Gaussians as ellipses corresponding to ±2 standard deviations from the mean along each of the independent axes of the distribution (i.e. Mahalanobis distance of 2).

**Sets of genes.** We obtained target gene lists for AP and DV factors from MacArthur et al. (1% FDR dataset) and Zeitlinger et al. We retrieved locations of housekeeping genes from the Flybase by selecting genes with at least moderate expression level (RPKM > 1) across all assessed stages and tissues.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

All raw sequencing data and extracted Hi-C contacts have been deposited in the Gene Expression Omnibus (GEO) repository under accession number "GSE121255" [www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121255]. The publically available ChIP-seq and GRO-seq datasets were obtained using accession numbers provided in the Methods section. All other relevant data is available upon request. The source data underlying Fig. 1e and Supplementary Figs. 2a and 7d are provided as a Source Data file.

**Code availability**

We performed all custom data analyses in Jupyter Notebook, using matplotlib, NumPy, and pandas. The software used in this study is available at https://github.com/mirnylab.

Received: 25 January 2019 Accepted: 27 August 2019 Published online: 03 October 2019

**References**

1. Apte, M. S. & Meller, V. H. Homologue pairing in flies and mammals: gene regulation when two are involved. Genet. Res. Int. 2012, 430587 (2012).

2. Joyce, E. F., Erceg, J. & Wu, C. T. Pairing and anti-pairing: a balancing act in chromosome pairing. Dev. Biol. 454, 1–21 (2018).

3. Fukaya, T. & Levine, M. Transcription. CELL. Biol. 27, R1047–R1049 (2017).

4. Cartoni, D. I. et al. Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. Nat. Commun. 8, 1753 (2017).

5. Szabo, Q. et al. TADs are 3D structural units of higher-order chromosome organization in Drosophila. Sci. Adv. 4, eaar8082 (2018).

6. Cardozo Gizzi, A. M. et al. Microscopy-based chromosome conformation capture enables simultaneous visualization of genome organization and transcription in intact organisms. Mol. Cell. 74, 212–222 e5 (2019).

7. Mateo, L. J. et al. Visualizing DNA folding and RNA in embryos at single-cell resolution. Nature 568, 49–54 (2019).

8. Li, Q. et al. The three-dimensional genome organization of Drosophila melanogaster through data integration. Nature 585, 233–235 (2017).

9. Ke, Y. et al. 3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis. Cell 170, 367–381 e20 (2017).

10. Barutzu, A. R., Maass, P. G., Lewandowski, J. P., Weiner, C. L. & Rinn, J. L. A TAD boundary is preserved upon deletion of the CTFC-rich Firre locus. Nat. Commun. 9, 1444 (2018).

11. Bonora, G. et al. Orientation-dependent Ddx4 contacts shape the 3D structure of the inactive X chromosome. Nat. Commun. 9, 1445 (2018).

12. Tan, L., Xing, D., Chang, C. H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. Science 361, 924–928 (2018).

13. Kim, S. et al. The dynamic three-dimensional organization of the diploid yeast genome. Elife 6, e25623 (2017).

14. AlHajj Abed, J. et al. Highly structured homolog pairing reflects functional organization of the Drosophila genome. Nat. Commun. (2019, In press).

15. Harrison, M. M. & Eisen, M. B. Transcriptional activation of the zygotic genome in Drosophila. Curr. Top. Dev. Biol. 113, 85–112 (2015).

16. Hug, C. B., Grimaldi, A. C., Kruse, K. & Vaquerizas, J. M. Chromatin architecture emerges during zygotic genome activation independent of transcription. Cell 169, 216–228 e19 (2017).

17. Stadler, M. R., Haines, J. E. & Eisen, M. Convergence of topological domain boundaries, insulators, and polytenic interbands revealed by high-resolution mapping of chromatin contacts in the early Drosophila melanogaster embryo. Elife 6, e29550 (2017).

18. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 378–380 (2012).

19. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381–385 (2012).

20. Imaiak, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods 9, 999–1003 (2012).

21. Li, E. et al. The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. Nature 456, 400–403 (2008).

22. Harrison, M. M., Li, X. Y., Kaplan, T., Botchan, M. R. & Eisen, M. B. Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-oocyte transition. PLoS Genet. 7, e1002266 (2011).

23. Xu, Z. et al. Impacts of the ubiquitous factor Zelda on Bicoid-dependent DNA binding and transcription in Drosophila. Dev. Cell 28, 608–621 (2014).

24. Delagrange, F. et al. Whole-genome ChIP-chip analysis of Drosal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes Dev. 21, 385–390 (2007).

25. Xu, Z. et al. Impacts of the ubiquitous factor Zelda on Bicoid-dependent DNA binding and transcription in Drosophila. Dev. Cell 28, 608–621 (2014).

26. Delagrange, F. et al. Whole-genome ChIP-chip analysis of Drosal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes Dev. 21, 385–390 (2007).
