Supplemental Materials

Supplemental materials include full methods, supplemental figures and tables.

Full methods

Genetics and fly stocks

Oregon R flies were used as the wild type. Embryos with no X chromosome were obtained by crossing attached-X/Y females to X/Y males. The stock used was C(1)*DX*, y f (Wieschaus and Sweeton, 1988). Embryos with no X and Y chromosomes were obtained by crossing attached-X/Y females (C(1)RM, $y^2su^{wa}w^a$) to attached-XY males (YSX YL, In(1)EN, y B). The compound II chromosomes RM(2L); RM(2R)=C(2)*v* and the compound III chromosomes RM(3L); RM(3R)=C(3)*se*, in which the two left arms or the two right arms segregate together, were used to generate 2L- and 2R-, and 3L- and 3R- embryos, respectively(Merrill *et al*., 1988). The compound II C(2)*EN* and compound III C(3)*EN* st^1 cu^1e^s, stocks (Bloomington 2974 and 1117) were used to generate embryos deficient for the entire second and third chromosome, respectively. The compound IV C(4)RM, $ci^1ey^R/0$ (Bloomington 1785) were used to generate embryos deficient for the fourth chromosome. Embryos deficient for chromosome 4 were identified by their defects in denticle belt patterning during late embryogenesis, whereas embryos deficient for other chromosome/chromosome arm were recognized based on their specific phenotypic defects during early embryonic development (Wieschaus and Sweeton, 1988; Merrill *et al*, 1988) . For example, embryos lacking the left arm of the second chromosome can be recognized by their three-layered appearance ("halo") during cycle 14 due to an altered distribution of the lipid droplets prior to cellularization, while embryos lacking the right arm of the second chromosome can be recognized by their failure in making ventral furrows at early gastrulation (Merrill *et al*, 1988). The early embryonic development phenotypes can be easily recognized by covering embryos in oil to make them transparent, and then viewing them in transmitted light under a stereomicroscope. All embryos were collected at room temperature.

To generate embryos deficient for smaller regions of 2R heterochromatin, T(2;Y) or T(2;3) translocation males were crossed to C(2)*EN* females. One eighth of the embryos

that lack 2R were identified according to their failure in forming the ventral furrow during early gastrulation. To generate embryos lacking different portions of the X heterochromatin, two types of stocks were used: (1) males carrying a deficient X chromosome missing part of the heterochromatin (generally *fog*$^{-}$) and a duplication on the Y chromosome that complements the deficiency, and (2) males carrying a duplication of X on the Y chromosome which covers part of the X heterochromatin. In both cases males were crossed to attached X females. In the first case, one quarter of embryos lack most of the X chromosome except the duplication of X on Y. These embryos were identified according to their defects during cellularization. Another quarter of the embryos carry the deficient X chromosome as their sole X chromosome. These embryos were identified according to their defects in posterior midgut formation during early gastrulation. In the second case, one quarter of embryos that lack most of the X chromosome except the duplication of X on Y were identified according to their defects in cellularization or posterior midgut formation. To obtain embryos lacking the euchromatin portion of the chromosome arms, translocation males bearing breakpoint at the euchromatin/heterochromatin boundary of 2L, 3L and X (Lindslet *et al*, 1972) were crossed to C(2)*EN*, C(3)*EN* or attached X females, respectively. Translocations and other deficiencies that were used to produce embryos with smaller genomic deficiencies are listed in Table S14.

Deep sequencing and identification of H-probes

In order to acquire potential *Drosophila* heterochromatin sequences for microarray construction, DNA libraries were prepared from approximately 250 2L- and 2R- embryos collected at the cellular blastoderm or early gastrulation stages, before the heterochromatin is underreplicated. The embryos were dechorionated and digested with proteinase K prior to phenol/chloroform extraction and sodium acetate/ethanol precipitation. DNA was sheared on a Covaris S2 instrument to a mean length of 200 bp, end repaired and ligated to adaptors for Illumina sequencing. The resulting short sequence reads were mapped to the Release 5 reference *Drosophila* genome using BWA with default settings (Li and Durbin, 2010). Sequences that mapped to the euchromatic arms and the heterochromatin sequences contiguous with the euchromatic arms (“h”)

were removed. The remaining sequences, which were considered probable unmapped heterochromatic sequence, were assembled into contigs using SSAKE (Warren *et al*. 2007). The reads were first quality trimmed using SSAKE provided scripts to require at least 20 consecutive bases above a threshold value of 10. The SSAKE assembly was run using default parameters. Custom microarrays were designed for manufacture by Agilent. Assembled contigs of 60 nt length or greater following masking by DUST were passed to ArrayOligoSelector (http://arrayoligosel.sourceforge.net/) for probe design, using a database of all assembled contigs plus the *Drosophila* genome as the background set used to determine the maximally unique probe for each contig sequence (Bozdech *et al*., 2003). Probe sequences were supplied to Agilent for manufacture of custom microarrays, available as Agilent design ID 024708. About 80,000 60-mer nucleotide probes were selected corresponding to the assembled contigs, including both forward and reverse complements in order to allow for use in subsequent expression analysis (~ 20,000 reverse probes that were not expressed were later removed from the microarray design). The microarray also contains ~23,000 probes designed by Agilent for expression sequences (annotated genes) covering the entire reference genome as reference probes.

BLAST analysis

BLAST analyses were performed using blast-2.2.23 (NCBI; ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/ncbi-blast-2.2.23). BLAST analysis of candidate H-probes against the entire genome and TEs was performed with seed length 15. BLAST analysis of candidate H-probes against the satellite DNAs was performed with seed length 5 with the low complexity filter (DUST) turned off. The release 5 reference *D. melanogaster* genome and its annotation (release 5.29) were downloaded from Flybase (www.flybase.org). To generate pseudo-probes for arm U that we used in the BLAST analysis, 60mer nucleotide probes spanning the entire arm U with each adjacent probe shifted by 30 nucleotides (30nt overlaps) were generated *in silico*.

CGH analysis and data normalization

For each experiment 100-150 embryos of the appropriate genotype were collected from apple juice-agar plates, dechorionated for 2 min in sodium hypochlorite (bleach), washed

in water, and then frozen in 1 ml of heptane (Sigma) using a dry ice chamber. Embryos were digested in 0.4 ml SNET buffer (20 mM Tris pH 8, 5 mM EDTA pH 8.0, 400 mM NaCl, 1% SDS, 0.5% triton) supplied with 500 ug/mL proteinase K at 50-55 ˚C over night, followed by phenol/chloroform extraction. DNA was sheared by sonication to a mean length of 500 bp, precipitated and redissolved in water. DNA purified from 0-8 hr wild type Oregon R embryos was used as a reference. Purified DNA was labeled using the BioPrime kit (Life Technologies). 500 ng of Cy5-labeled test DNA and Cy3-labeled reference DNA were mixed together and hybridized with the microarray for 17 h in a 65 ˚C oven rotating at 20 rpm. Arrays were washed according to the Agilent CGH protocol (including a 37˚C wash) and subsequently scanned using Agilent scanner. Feature extraction was performed by Agilent feature extraction software using the CGH protocol. For each test DNA, hybridization data were normalized by setting the median $\log_2$-ratio to 0 for chromosomes present as diploids in a given sample (i.e. no deletion). DNA used for the underreplication analysis was isolated from whole salivary glands dissected from roaming third-instar larvae or from ovaries dissected from adult females.

Mapping assignment by support vector machine

Samples with large chromosomal deletions result in arrays containing a large number of features with very extreme red to green ratios. Standard normalization approaches assume that on average all features will center at 0, and so are not applicable to data from samples with large deletions. Array data were normalized using the limma package in R (Smyth and Speed, 2003). Arrays were first normalized by variance stabilization between arrays using NormalizeBetweenArrays (Huber *et al*., 2002). A control array of cycle 14 wild type Oregon R embryos hybridized to the usual 0-8 hour wild type Oregon R embryo reference was the training model for variance stabilization. Following normalization between arrays, each array was normalized using NormalizeWithinArrays with all mapped probes present at normal diploid levels as control probes. The result of this normalization was that normal diploid probes were centered to a $\log_2$ ratio equal to 0, and the $\log_2$ ratios of all other probes were normalized accordingly. Replicate probes present on the array were collapsed with the avedups function in the limma package. These normalized ratiometric data were then classified using a support vector machine,

SVM-multiclass from the SVM Light package (Joachims *et al*., 1999) trained using the locations of euchromatic probes with known positions that have single BLAST hits to the genome. Classification using solely $\log_2$ ratios was less accurate than classification using both ratiometric data and raw intensities of the array data corresponding to the deletion data, so both were used. The SVM scores of the classifier on known probes were used to determine a cutoff score for which 95% of positive calls are accurate. This cutoff was then used to group H-probe position calls into "high confidence" and "low confidence" groups.

Microarray analysis of gene expression

Oregon R embryos visually staged under a stereomicroscope were collected from apple juice-agar plates, dechorionated for 2 min in sodium hypochlorite (bleach), washed in water, and then frozen in 1 ml of heptane (Sigma) using a dry ice/ethanol chamber. Six developmental stages were examined: (1) 0-1 hr; (2) cycle13 to early cycle 14 (~2-2.5 hr); (3) Mid- to late cycle 14 (~2.5-3 hr); (4) 3-4 hr; (5) 4-5 hr; and (6) 19-22 hr. Total RNA was extracted with TRIzol (Invitrogen), and 325 ng of RNA (approximately 5-10 embryos) was used to synthesize complementary RNA (cRNA) according to the Agilent protocol. cRNA prepared from an Oregon R embryo sample with broad developmental stages (0-16 hr) was used as the reference for hybridization. The same amount of Cy5-labeled sample cRNA and Cy3-labeled reference cRNA were mixed together for hybridization such that there was at least 5 pmol of dye in each channel with equal quantity of cDNA (approximately 1000 ng of cRNA probes for each channel). Each array was hybridized with probes for 17 hr in a 65 ˚C oven rotating at 20 rpm. Arrays were washed according to the Agilent CGH protocol (including a 37˚C wash) and subsequently scanned using Agilent scanner. Feature extraction was performed by Agilent feature extraction software using the gene expression protocol. Raw intensities of the array data were normalized using the limma package in R. Arrays were first loess normalized using NormalizeWithinArrays, and then all arrays were quantile normalized according to the reference channel using NormalizeBetweenArrays. The expression detection threshold is determined according to the hybridization intensities of a set of non-expressed control probes such that 99% of the control probes have normalized hybridization intensity below

the threshold. Probes were clustered by Pearson correlation according to the ratiometric expression data at different developmental stages.

Analysis of genomic sequence datasets

The FASTX-toolkit package was used for data preprocessing such as removal of adapters in small RNA data, trimming for length, and conversion of quality scores. Sequence alignments of Illumina sequence data were done using Bowtie (Langmead *et al*., 2009). For the analysis of DNA copy number, Illumina raw sequence data of 100 read length was trimmed to the first 35 bases to permit mapping to the 60mer array probes, and we counted the number of reads matching each probe. Reads were permitted to map to all probes with a perfect or single mismatch (-a option in bowtie).

For comparison of our data with patterns of small RNA expression, the raw data from Chung *et al*. 2008 were downloaded from the Short Read Database (Langmead *et al*., 2009) and trimmed to remove 3' adapter sequence. These reads were mapped using Bowtie to the H-probes, and the number of small RNA reads per probe was counted. Since these libraries are constructed in a strand-specific fashion, strand information from the mapping was retained in order to determine strand bias.

For comparison of our probes with chromatin modification patterns during development, ModENCODE ChipSeq data were downloaded from the Short Read Archive, mapped to the H-probes, and the number of reads matching each probe was counted. The ChipSeq method used by the ModENCODE consortium produces sequence data on both strands, so reads from both strands were summed.

Supplemental Figures

Supplemental Figure 1. Summary of heterochromatin mapping strategy. (A) Flowchart for Identification of heterochromatic sequence-specific probes (the "H-probes") for *D. melanogaster* heterochromatin. These probes were used to establish custom microarrays for array CGH analysis. **(B)** Flowchart for mapping H-probes by array CGH analysis of chromosome deletions.

Supplemental Figure 2. Generation of embryos deficient for large regions of a chromosome (here: autosomes). (A) Compound autosome stocks. Instead of the normal diploid arrangement in which the left and right arms of a given chromosome are attached to the same centromere, compound autosomes have the two left or two right arms attached to the same centromere. Such compound chromosomes had been generated in the 1960s (Rasmussen, 1960; Scriba, 1967, 1969) by inducing breaks at the centromeric regions of the left arm of one chromosome and the right arm of the other, followed by joining the two left arms to one centromere, and the two right arms to the other. The surviving adults of the stock generate offspring that are either normal wild type or that totally lack the right or left arm of a given chromosome (Merrill *et al*., 1988). **(B)** Compound entire chromosome stocks. Compound (2 or 3) entire chromosomes have two copies of each arm of the autosome joint to a single centromere, such that a quarter of the embryos produced by these flies will lack the entire autosome (2 or 3En-). **(C)** Generation of embryos deficient for smaller regions of a chromosome (here: autosomes) by translocations. To obtain smaller deletions, compound autosome females were crossed to males bearing Y-autosome translocations with breakpoints at various points along the chromosome. One eighth of the embryos from the cross will lack the sequence more distal to the translocation breakpoint.

Supplemental Figure 3. BLAST analysis of candidate H-probes in category 1. Histogram showing the distribution of H-probes in category 1 (those can be aligned to the reference genome) according to the number of chromosome arms on which the H-probe can be aligned to. For this analysis arm U, arm Uextra, individual arm Het, and euchromatic arms are considered as different chromosome arms. The pie graphs show the

distribution of the best BLAST hits for H-probes that can be aligned to one or two arms, respectively. 55.5% of the H-probes can only be aligned to a single arm. Another 40.5% can be aligned to two arms, but most were shared between arm U and Uextra, or beween armU/Uextra and a single armHet. Therefore, for most H-probes, there is no evidence indicating that they can be aligned to different chromosome arms other than arm U or Uextra, which is consistent with the observation that most of the H-probes were mapped to a defined chromosome region in our mapping analysis.

Supplemental Figure 4. Selection of H-probes is robust to variations in filtering stringency. We selected our H-probes according to two criteria: (1) H-probes that can be aligned to arm Het, arm U or arm Uextra (allowing up to 1 mismatch), or (2) H-probes that cannot be aligned to the reference genome (two or more mismatches) but were mapped to the heterochromatic region. To test whether altering the stringency of filtering by allowing different numbers of mismatches in the alignment would alter the selection of H-probes, candidate H-probes were divided into two categories as described in Figure 1A but allowing 0, 2, 3, 4, 5 or 6 mismatches. **(A)** The distribution of the best BLAST hits for H-probes in each category remains largely unchanged up to allowing 3 mismatches. Allowing 4 or more mismatches leads to an increase in the fraction of candidate H-probes in category 1 whose best match is located in euchromatin. As demonstrated in (D), most of these candidate H-probes would be recognized as polymorphic euchromatic probes and excluded. **(B-C)** Mapping location of H-probes in category 1 when 4 mismatches were allowed. (B): H-probes aligned to arm Het. (C): H-probes aligned to arm U or Uextra. "H", "L" and "U" indicate that the probes were mapped with high confidence, low confidence, or unmappable, respectively. The distribution of H-probes to different mapping categories is very similar to Figure 2B-C when only 1 mismatch was allowed in the alignment. **(D)** Comparing to 1 mismatch, allowing 6 mismatches leads to a shift of 7840 probes from category 2 to category 1, of which 2/3 can be mapped with high confidence. Among these probes, 10% have best alignment within heterochromatin, including arm Het, U and Uextra (group 1), whereas 90% have best alignment on euchromatic arms (group 2). According to our selection criteria, the former will be included as H-probes while the latter will be excluded. If we

filter the data by allowing 1 mismatch, both groups will be classified as category 2 and subject to test as described in Figure 2D. As shown in the bar graph, more than 80% of the group 1 probes were recognized as "Novel H-probes" and included, whereas more than 95% of the group 2 probes were recognized as "Polymorphic euchromatic probes" and excluded from H-probes. Therefore, although the exact fractions vary depending on the criteria used in their initial classification, our grouping of candidate H-probes into heterochromatic and polymorphic euchromatic, as well as our mapping conclusions are remarkably robust to the filtering step.

Supplemental Figure 5. Comparison of the control annotated euchromatic probes that are present or absent in specific chromosome/chromosome arm deletions. CGH analysis of embryos lacking specific chromosomes or chromosome arms. The $\log_2$ ratios of test versus reference (x-axis) and $\log_{10}$ intensities of hybridization (y-axis) were plotted. For each genotype, genes separate into two distinct classes. Genes present in deficiency DNA (red) show $\log_2$ ratios (x-axis) close to 0 and higher $\log_{10}$ intensities of hybridization (y-axis), whereas genes not represented in the deficiency DNA (blue) show low $\log_2$ ratios and low intensities.

Supplemental Figure 6. Correlation of hybridization intensities between test samples and the wild type control. (A) Scatter plot comparing normalized hybridization intensities of mapped H-probes present at normal diploid levels between test deficiency samples and the control Oregon R samples. The raw hybridization intensities were normalized as described in Materials and Methods. The result of normalization was that normal diploid control annotated euchromatic probes were centered to a $\log_2$-ratio equal to 0, and the $\log_2$-ratios of all other probes were normalized accordingly. **(B)** Bar graph showing the r-squares and slopes of the linear fits in (A). The high correlation of hybridization intensities suggests that the repetitiveness of the H-probes were comparable between test deficiencies and the Oregon R.

Supplemental Figure 7. Mapping analysis of the polymorphic euchromatic probes. CGH analysis of embryos lacking specific chromosome or chromosome arm. DNA

extracted from randomly staged 0-8 hr wild type Oregon R embryos was used as reference for all experiments. Heat map shows probes that are similar to the euchromatic sequences (with BLAST e-value between 10^{-7} and 10^{-22}) and mapped with high confidence. Probes are clustered according to their predicted chromosome locations on the basis of sequence similarity (labeled on the right side of the heat map). Green: fold decrease. Red: fold increase. Table shows the correct recognition rate for the high confident assignments.

Supplemental Figure 8. Sequence element families of the novel H-probes. (A) Distribution of the best BLAST alignment (euchromatin v.s. heterochromatin) for the novel H-probes in each BLAST e-value category. H-probes with e-values lower than 1E-7 (dashed line) were considered to be similar to the reference genome, while those with e-values higher than 1E-7 were considered as dissimilar. **(B-C)** H-probes were categorized into two groups according to their sequence similarity to the reference genome (B) and the sequence composition of each group is shown (C). Group 1 is similar to the known heterochromatic sequences and mainly composed of Satellite-like and TE-like sequences, while group 2 does not show clear sequence similarity to known heterochromatin and is mainly composed of non-satellite, non-TE sequences.

Supplemental Figure 9. CGH analysis of translocations and deficiencies used to map 2R-het and X-het. (A) CGH analysis of translocations bearing breakpoints in 2R heterochromatin. Heat map showing $\log_2$-ratios for annotated euchromatic genes. Genes were clustered according to their chromosomal locations. Euchromatic genes from 2R are deleted from the deficiency embryos. Green: fold decrease. Red: fold increase. **(B)** CGH analysis of the X chromosome rearrangements. Two different types of X chromosome rearrangements were analyzed: (1) Y duplicated for a piece of proximal X, and (2) X deficiencies encompassing part or all of the X heterochromatin. Showing clustering of annotated euchromatic genes on the X chromosome according to their locations. Upper panel is a model showing deleted regions on X for each X chromosome rearrangement. The euchromatin portion is determined by the chromosomal location of the deleted annotated genes on X. The heterochromatin (Het) portion is determined by

clustering of H-probes mapped to X and justified according to the history of each chromosome rearrangement (data not shown).

Supplemental Figure 10. Sequence composition of pseudo-probes derived from arm U. 60mer pseudo-probes were generated spanning the entire arm U with each adjacent probe shifted by 30 nucleotides (30nt overlaps). Map positions were assigned to these probes according to the homology between our mapped H-probes and arm U contigs. The sequence similarity between these probes and known satellite DNAs or TEs was analyzed by BLAST analysis. **(A)** Sequence composition at different heterochromatic regions. **(B, C)** Bar graphs summarizing the TE (B) or satellite DNA (C) populations in different heterochromatic regions.

Supplemental Figure 11. H-probe coverage of the arm U contigs. Distribution of arm U contigs according to the fraction of the contig sequence that is covered by H-probes. The U-contigs were categorized according to the mapping position and sequence composition of the H-probes aligned to them. U-contigs were included in the analysis only when more than 95% of the aligned H-probes were mapped to same region and shared similar sequence property (Satellite-like: “Sat”, or non-Satellite-like: “non-Sat”).

Supplemental Figure 12. CGH analysis of polytene chromosomes in ovaries. Polytene DNA was purified from ovaries and compared with DNA from blastoderm stage (cycle 14) embryos. The data have been normalized such that the mean $\log_2$-ratios of ovary polytene DNA and embryonic DNA was zero for single copy euchromatic genes (Eu). Comparing to the euchromatic genes, the H-probes were underrepresented in the polytene DNA, but to a lesser extent comparing to the salivary gland polytene chromosomes (Figure 4A).

Supplemental Figure 13. Characterization of the polytenized H-probes. (A) CGH analysis of polytene DNA purified from larval salivary glands and DNA from blastoderm stage embryos in which tissue differentiation and chromosome polytenization have not started. Data were normalized as described in Figure 4A. Histogram showing the

distribution of $\log_2$(salivary gland/embryos) for annotated euchromatic probes (Eu) and H-probes (Het). Note that a subset of H-probes (right to the dashed line) were not underreplicated in the polytene chromosomes, even though they mapped to heterochromatic regions. This subset of H-probes lacks any satellite-like probes, while TE-like probes are also much underrepresented. **(B)** Distribution of polytenized and underreplicated non-satellite H-probes according to their transcription levels as determined by hybridization intensities of the cRNAs on the microarray (Figure 6A). Insert showing the enlarged view of the boxed region which demarcates highly transcribed H-probes. **(C)** Distribution of polytenized and underreplicated non-satellite H-probes according to their enrichment for H3K9Me3 during 0-12 hrs (left) or 12-24 hrs (right).

Supplemental Figure 14. Comparison of H-probes and annotated heterochromatic genes for their repetitiveness and degree of underreplication in polytene chromosomes. Bivariate scatter plots comparing sequence repetitiveness with level of underreplication in polytene chromosomes. The degree of underreplication and repetitiveness was measured as in Figure 4. **(A)** Comparing annotated heterochromatic genes with the non-satellite-like H-probes on the same chromosome arm. Annotated genes are grouped according to their relative proximity to the centromere as shown in (B). The overall distribution of the non-satellite-like H-probes is mostly reminiscent of the annotated heterochromatic genes located at the more proximal region of the β-heterochromatin ("arm Het"), but not those located at the most distal heterochromatic regions ("h") which appear to be euchromatin-like. **(B)** The cytogenetic position of the annotated heterochromatic genes. **(C)** Comparing H-probes mapped to different subregions of 2R-het (see Figure 2E). The cytogenetic position of each group is shown on the top.

Supplemental Figure 15. Comparing the repetitiveness of H-probes between two different laboratory stocks. In order to test whether the degree of repetitiveness would change between different *D. melanogaster* strains, we analyzed an additional genomic library from ModENCODE which was generated from the Oregon R strain (this data set

was served as input control of a ChIP-seq experiment). The repetitiveness of H-probes was quantified as reads/probe and compared to the reads from the original genomic library derived from cn^1 bw^1 (see Figure 4C). **(A)** Scatter plot comparing the two libraries for the number of reads mapped to each H-probe. The reads/probe from two libraries showed a strong correlation. **(B-D)** Histogram showing the distribution of probes according to their repetitiveness as measured by the number of genomic reads mapped to them. (B): distribution of the control annotated euchromatic probes. (C): distribution of H-probes according to reads from the cn^1 bw^1 library. (D): distribution of H-probes according to reads from the Oregon R library. The reads from the Oregon R library has been normalized according to the linear regression in (A). The overall distribution of H-probes in (C) and (D) look very similar, except that a higher fraction of H-probes showed "0 reads" in D, likely because the total number of reads in the Oregon R library is much lower than the cn^1 bw^1 library. Note that some "0 reads" bars in (D) were truncated, with the length labeled above the bar.

Supplemental Figure 16. Strand bias of small RNA expression. Bar graph showing the distribution of ratios of small RNA reads matching to the forward or reverse strand of H-probes. The plus strand is defined as the strand to which more small RNA reads are mapped. H-probes were categorized into three groups according to the number of small RNA reads mapped to them ($n < 10$; n between 10 and 100; and $n > 100$). Degree of strand bias was categorized as (1) low strand bias: ratio < 2; (2) moderate strand bias: ratio between 2 and 10; and (3) high strand bias: ratio > 10. Highly abundant small RNAs often showed strong strand bias.

Supplemental Figure 17. Comparison of gene expression levels measured by microarray (this study) with published ModENCODE RNA sequencing data. RNA seq reads were downloaded from ModENCODE (Graveley *et al*., 2011) and mapped to our H-probes. The number of reads mapped to each H-probe was used as an estimation of expression level and compared to the normalized array hybridization intensities for samples collected at comparable developmental stages. No clear correlation was observed for individual H-probes. It is worth noting that only a small fraction (5%) of the control

annotated euchromatic probes or H-probes have more than 2 reads mapped to them (data not shown). Nevertheless, as we categorized the probes according to the number of reads mapped to them, the average hybridization intensity within each category correlated with the average number of reads, suggesting that the microarray and RNA seq approaches were overall consistent.

Supplemental Figure 18. Pericentric heterochromatin is enriched for H3K9Me3. Enrichment for H3K9Me3 in 0-12 hr embryos are shown for the proximal 3Mb of chromosomes 2, 3, and X, as well as the distal 1.35 Mb of chromosome 4. The arm Het scaffolds for chromosomes 2, 3, and X are also shown. ChIP-seq reads for H3K9Me3 (modENCODE) were aligned to H-probes and the number of reads for each H-probe was normalized to the input. Log_2(normalized reads) are plotted relative to the chromosomal position. Blue and grey boxes underneath the bar graph demonstrate the genomic scaffolds and gaps, respectively. Gaps within each arm Het are not shown. Red boxes demarcate heterochromatic regions determined in Riddle *et al*., 2011. C: centromeres.

Supplemental Figure 19. Histone modifications associated with H-probes. ChIP-seq reads for H3K9Me3 and H3K27Me3 (modENCODE) were aligned to H-probes and the number of reads for each H-probe was normalized to the input. Showing bivariate scatter plots comparing H3K9Me3 or H3K27Me3 enrichment with level of transcription. The x-axis is the average of Log_2(normalized reads for H3K9Me3 or H3K27Me3 ChIP-seq) at designated stages of the embryonic development. The y-axis is the Log_2 scale of the transcription level at the corresponding developmental stages detected by microarray. Left panels: average of H3K9Me3 enrichment at 12-16 hrs, 16-20 hrs and 20-24 hrs versus transcription during 19-22 hrs. Middle panels: average of H3K27Me3 enrichment at 0-4 hrs, 4-8 hrs and 8-12 hrs versus maximal transcription during 2.5-5 hrs. Right panels: average of H3K27Me3 enrichment at 12-16 hrs, 16-20 hrs and 20-24 hrs versus transcription during 19-22 hrs.

Supplemental Figure 20. Comparison of the H3K27Me3 pattern with transcription profiles for H-probes mapped to Xd-het. ChIP-seq reads for H3K27Me3

(modENCODE) were aligned to H-probes and the number of reads for each H-probe was normalized to the input. Heat map showing Log_2(normalized reads). H-probes mapped to each subregion of Xd-het (X2-X4) were clustered by hierarchical clustering. H-probes mapped to the same subdivision share similar pattern of H3K27Me3 enrichment and transcription. Region X2, which is not expressed shows high levels of H3K27Me3 modification, whereas the adjacent X3 and X4 regions, which show new transcription during blastoderm stages are depleted of H3K27Me3 during early development.

Table 1 (Continued): Amount of sequence added to each heterochromatic region from arm U and arm Uextra ("Perfectly mapped").

arm U			Basepairs					Number of Contigs			
Category		n=1	2≤n≤9	n≥10	Total	Category		n=1	2≤n≤9	n≥10	Total
2L		0	3289	19498	**22787**	2L		0	2	1	**3**
2CEN		14634	0	0	**14634**	2CEN		7	0	0	**7**
2R		52826	201603	285720	**540149**	2R		8	32	9	**49**
	2R1	0	12100	0	**12100**		2R1	0	4	0	**4**
	2R2	30268	51003	0	**81271**		2R2	4	9	0	**13**
	2R3	0	0	0	**0**		2R3	0	0	0	**0**
	2R4	15556	70850	200113	**286519**		2R4	1	2	2	**5**
	2R5	4146	25406	61952	**91504**		2R5	2	8	5	**15**
	2R6	2856	21233	4484	**28573**		2R6	1	6	1	**8**
	SUM	52826	180592	266549	**499967**		SUM	8	29	8	**45**
3L		8702	5061	0	**13763**	3L		3	3	0	**6**
3CEN		185240	66599	525514	**777353**	3CEN		41	16	145	**202**
3R		51997	26398	19473	**97868**	3R		18	12	2	**32**
4		0	5345	0	**5345**	4En		0	1	0	**1**
X		31621	166171	792303	**990095**	X		15	84	395	**494**
	X1	0	1956	520224	**522180**		X1	0	2	300	**302**
	X2	8336	25020	25743	**59099**		X2	4	10	9	**23**
	X3	5927	24191	0	**30118**		X3	4	15	0	**19**
	X4	17577	63478	14661	**95716**		X4	9	23	9	**41**
	X3/X4	0	46413	202082	**248495**		X3/X4	0	23	70	**93**
	SUM	31840	161058	762710	**955608**		SUM	17	73	388	**478**
XY		73995	58516	18087	**150598**	XY		32	36	15	**83**
Y		105287	494285	128955	**728527**	Y		58	223	59	**340**
Total		**524302**	**1027267**	**1789550**	**3341119**	Total		**182**	**409**	**626**	**1217**

arm Uextra			Basepairs					Number of Contigs			
Category		n=1	2≤n≤9	n≥10	Total	Category		n=1	2≤n≤9	n≥10	Total
2L		1496	49046	4237	**54779**	2L		1	62	1	**64**
2CEN		33174	0	0	**33174**	2CEN		43	0	0	**43**
2R		159838	114646	0	**274484**	2R		195	129	0	**324**
	2R1	50037	12382	0	**62419**		2R1	68	16	0	**84**
	2R2	28962	14349	0	**43311**		2R2	33	14	0	**47**
	2R3	0	0	0	**0**		2R3	0	0	0	**0**
	2R4	8300	1634	0	**9934**		2R4	10	2	0	**12**
	2R5	40240	51506	0	**91746**		2R5	46	59	0	**105**
	2R6	32299	24934	0	**57233**		2R6	38	28	0	**66**
	SUM	159838	104805	0	**264643**		SUM	195	119	0	**314**
3L		53977	52552	0	**106529**	3L		67	59	0	**126**
3CEN		298346	754156	353561	**1406063**	3CEN		351	1068	468	**1887**
3R		168759	232043	4443	**405245**	3R		167	206	1	**374**
4		0	0	0	**0**	4		0	0	0	**0**
X		718377	1929714	2635488	**5283579**	X		963	2615	3611	**7189**
	X1	44449	938134	2518856	**3501439**		X1	64	1365	3462	**4891**
	X2	106240	145323	47458	**299021**		X2	136	192	56	**384**
	X3	230593	174718	0	**405311**		X3	309	226	0	**535**
	X4	295289	361092	2852	**659233**		X4	389	468	3	**860**
	X3/X4	0	262005	12396	**274401**		X3/X4	0	316	13	**329**
	SUM	676571	1881272	2581562	**5139405**		SUM	898	2567	3534	**6999**
XY		431831	770127	1692	**1203650**	XY		610	1103	2	**1715**
Y		395489	698902	6437	**1100828**	Y		515	894	8	**1417**
Total		**2261287**	**4601186**	**3005858**	**9868331**	Total		**2912**	**6136**	**4091**	**13139**

Note: "n" is the frequency of BLAST hits of H-probes for a given U_contig. A U- or Uextra-contig is "perfectly mapped" to a heterochromatic region if ≥ 95% of the hits (H-probes) are mapped to that region.

Supplemental Tables (Table S1-S13 are attached as additional files)

Supplemental Table 1. List of all candidate probes, mapping assignments, confident scores, and their similarity to the reference genome, satellite DNAs and TEs.

Supplemental Table 2. List of candidate probes in category 1 and the number of BLAST hits in the reference genome.

Supplemental Table 3. List of mapped H-probes and their sequence element families.

Supplemental Table 4. Mapping assignments for the arm U contigs.

Supplemental Table 5. Mapping assignments for the arm Uextra contigs.

Supplemental Table 6. List of H-probes with their sequence repetitiveness and degree of underreplication in differentiated tissues.

Supplemental Table 7. List of polytenized H-probes.

Supplemental Table 8. CGH normalized red intensities.

Supplemental Table 9. List of H-probes with number of small RNAs mapped to them.

Supplemental Table 10. Positioning piRNA clusters mapped to arm U.

Supplemental Table 11. List of H-probes with their transcription levels.

Supplemental Table 12. List of highly transcribed H-probes.

Supplemental Table 13. List of H-probes and their association with histone modifications and transcription machineries.

Supplemental Table 14. List of translocations, duplications and deficiencies used in this study.

Name	Cytology	Reference	Early Phenotype
T(2;3)ftz^{Rpl}	41	Duncan (1986)	Lack ventral furrow
T(Y;2)B238	41	Lindsley *et al*. (1972)	Lack ventral furrow
T(Y;2)B63	41	Lindsley *et al*. (1972)	Lack ventral furrow
T(Y;2)B^{SV5}	41	Craymer (1984)	Lack ventral furrow
T(2;3)bxd^{68}	41	Bender *et al*. (1985)	Lack ventral furrow
T(Y;2)G10	41	Lindsley *et al*. (1972)	Lack ventral furrow
T(2;3)E(da)	41	Nicholls *et al*. (1998)	Lack ventral furrow
Dp(1;Y)y^{2}67g	2B; 20A	Lefevre (1981)	Cellularization defects
Dp(1;Y)y^{2}sc	1F	Lefevre	Cellularization defects
Dp(1;Y)y^{+}mal^{126}	1A;1B;18F;20A; 20E-F	Schalet and Lefevre (1973)	Cellularization defects
Df(1)fog^{114}	20E-F	Perrimon (1989)	Gastrulation defects
Dp(1;Y)y^{+}mal	1A;1B;18F;20F	Schalet and Lefevre (1973)	Cellularization defects
Df(1)mal12	19A;h26-h32	Schalet and Lefevre (1973)	Gastrulation defects
Dp(1;Y)ct^{+} y^{+}	1A;1B;6E;7C	Johnson and Judd (1979)	Gastrulation defects
Dp(1;Y) ct y^{+}	1A;1B	This study	Cellularization defects
T(1;Y)B91	20A-B	Merriam *et al*. (1998)	Cellularization defects
T(Y;2)R146	40	Lindsley *et al*. (1972)	“halo” phenotype
T(2;3)H31	79F	Hilliker *et al*. (1987)	Gastrulation defects

Supplemental References:

Bender, W., B. Weiffenbach, F. Karch, and M. Peifer. 1985. Domains of cis-interaction in the bithorax complex. Cold Spring Harb Symp Quant Biol 50: 173-180.

Bozdech, Z., J. Zhu, M.P. Joachimiak, F.E. Cohen, B. Pulliam, and J.L. DeRisi. 2003. Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol 4: R9.

Craymer, L. 1984. Techniques for manipulating chromosomal rearrangements and their application to *Drosophila* melanogaster. II. Translocations. Genetics 108: 573-587.

Duncan, I. 1986. Control of bithorax complex functions by the segmentation gene fushi tarazu of D. melanogaster. Cell 47: 297-309.

Hilliker, A.J. and S.N. Trusis-Coulter. 1987. Analysis of the functional significance of linkage group conservation in *Drosophila*. Genetics 117: 233-244.

Huber, W., A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18 Suppl 1: S96-104.

Joachims, T. 1999. Making large-Scale SVM Learning Practical. In *Advances in Kernel Methods -Support Vector Learning* (ed. B. Schölkopf, C. Burges and A. Smola). MIT-Press. Cambridge, MA.

Johnson, T.K. and B.H. Judd. 1979. Analysis of the Cut Locus of *DROSOPHILA* MELANOGASTER. Genetics 92: 485-502.

Langmead, B., C. Trapnell, M. Pop, and S.L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

Lefevre, G. 1981. The distribution of randomly recovered X-ray-induced sex-linked genetic effects in *Drosophila* melanogaster. Genetics 99: 461-480.

Li, H. and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.

Lindsley, D.L., L. Sandler, B.S. Baker, A.T. Carpenter, R.E. Denell, J.C. Hall, P.A. Jacobs, G.L. Miklos, B.K. Davis, R.C. Gethmann, *et al*. 1972. Segmental

aneuploidy and the genetic gross structure of the Drosophila genome. Genetics 71: 157-184.

Merriam, J., M.T. Yamamoto, B. Stewart, M.R. Rahman, and T. Nicolau. 1998. X chromosome segmental aneuploidy and response to gene dosage in Drosophiola melanogaster. Personal communication to FlyBase.

Merrill, P.T., D. Sweeton, and E. Wieschaus. 1988. Requirements for autosomal gene activity during precellular stages of *Drosophila* melanogaster. Development 104: 495-509.

Nicholls, R.E. and W.M. Gelbart. 1998. Identification of chromosomal regions involved in decapentaplegic function in *Drosophila*. Genetics 149: 203-215.

Perrimon, N., D. Smouse, and G.L. Miklos. 1989. Developmental genetics of loci at the base of the X chromosome of *Drosophila* melanogaster. Genetics 121: 313-331.

Rasmussen, I.E. 1960. Attached 2R;2L. Dros. Information Serv. 34.

Riddle, N.C., A. Minoda, P.V. Kharchenko, A.A. Alekseyenko, Y.B. Schwartz, M.Y. Tolstorukov, A.A. Gorchakov, J.D. Jaffe, C. Kennedy, D. Linder-Basso, *et al*. Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. Genome Res 21: 147-163.

Schalet, A. and G. Lefevre, Jr. 1973. The localization of "ordinary" sex-linked genes in section 20 of the polytene X chromosome of *Drosophila* melanogaster. Chromosoma 44: 183-202.

Scriba, M.E. 1967. Embryonale Entwicklungsstorungen bei Defizienz und Tetraploidie des 2. Chromosoms von *Drosophila* melanogaster. Wilhelm Roux' Arch. EntwMech Org. 159.

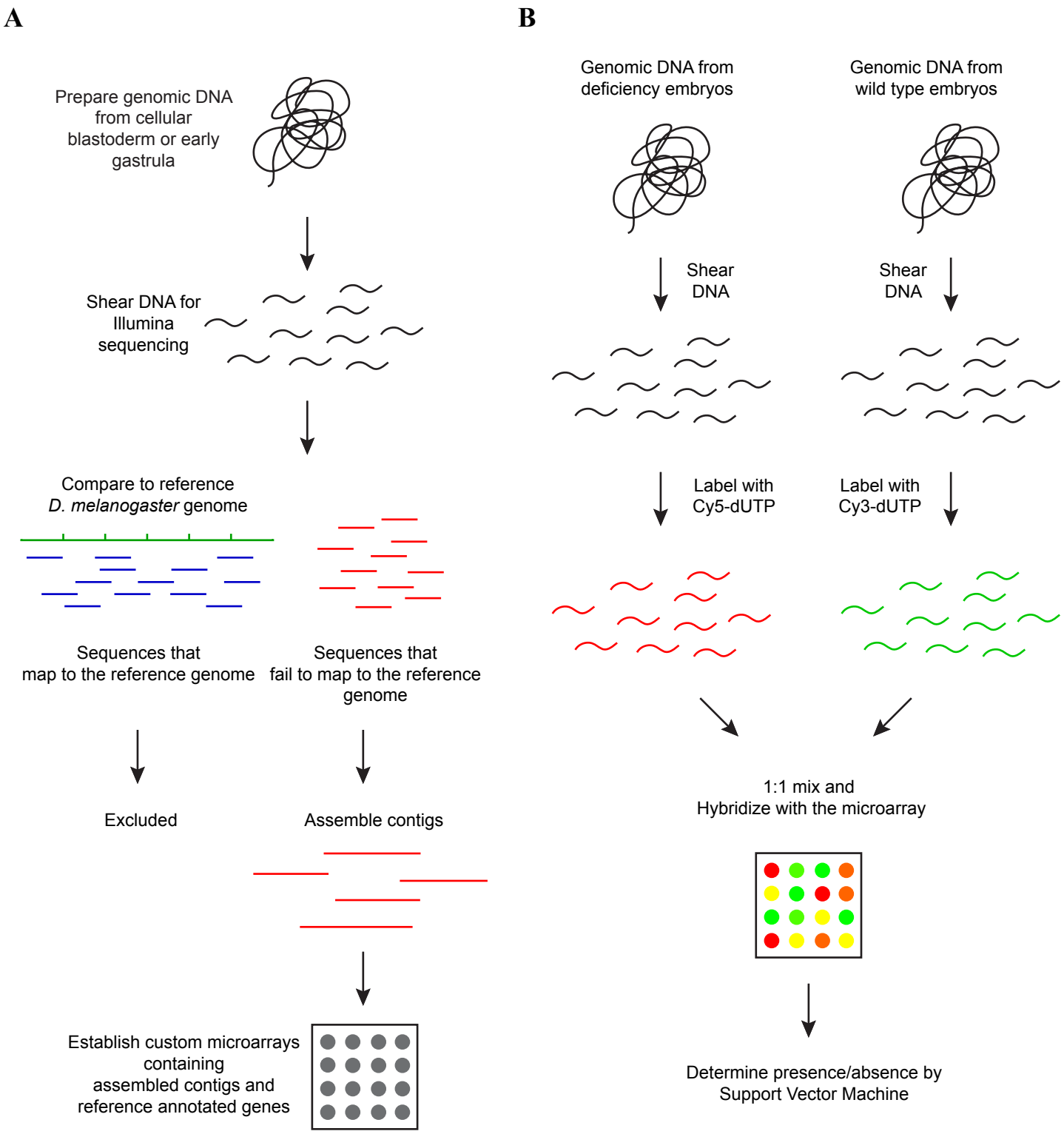
Scriba, M.E. 1969. [Embryonale hypogenesis at nullosomy and tetrasomy of *Drosophila* melanogaster third chromosome]. Dev Biol 19: 160-177.

Smyth, G.K. and T. Speed. 2003. Normalization of cDNA microarray data. Methods 31: 265-273.

Warren, R.L., G.G. Sutton, S.J. Jones, and R.A. Holt. 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23: 500-501.

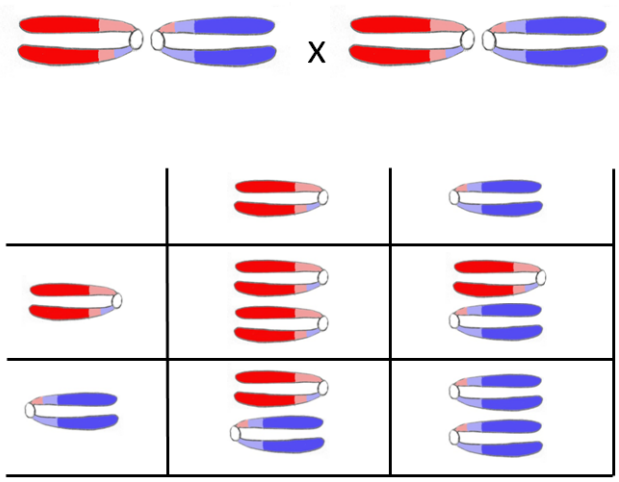
Wieschaus, E. and D. Sweeton. 1988. Requirements for X-linked zygotic gene activity during cellularization of early *Drosophila* embryos. Development 104: 483-493.

Supplemental Figure 1

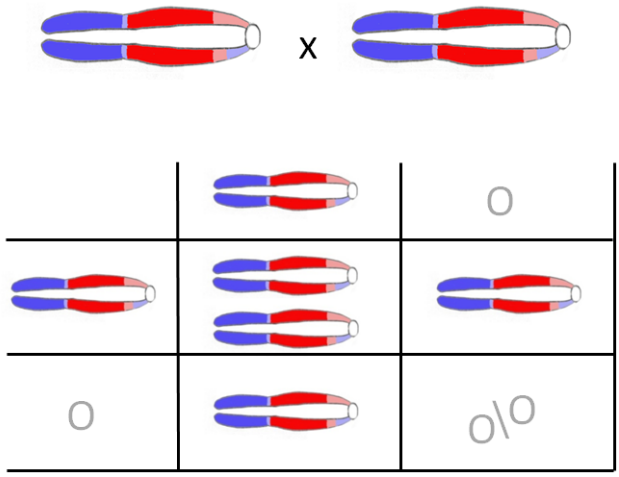


Supplemental Figure 2

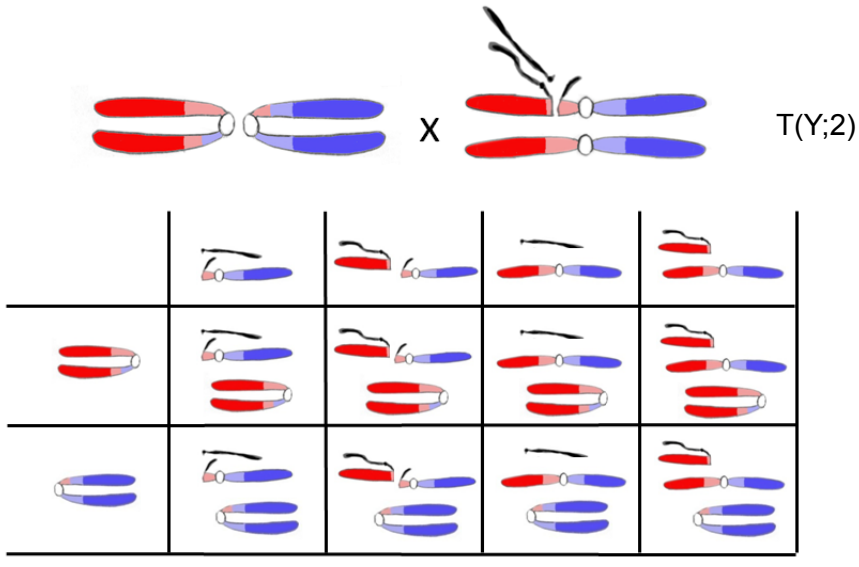
A



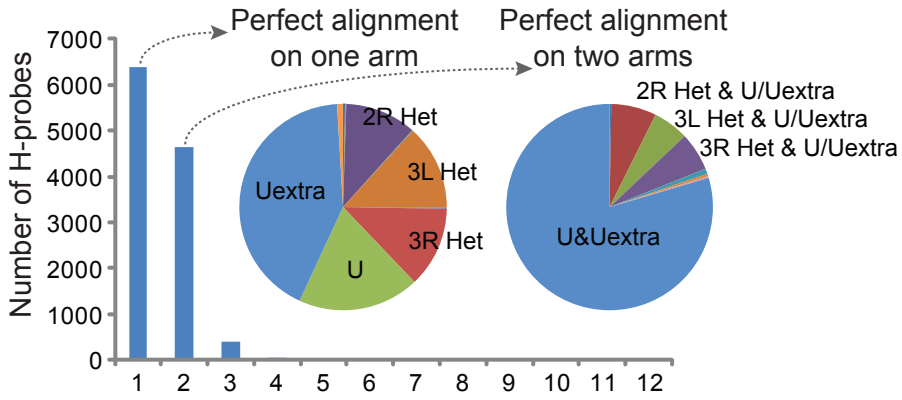
B



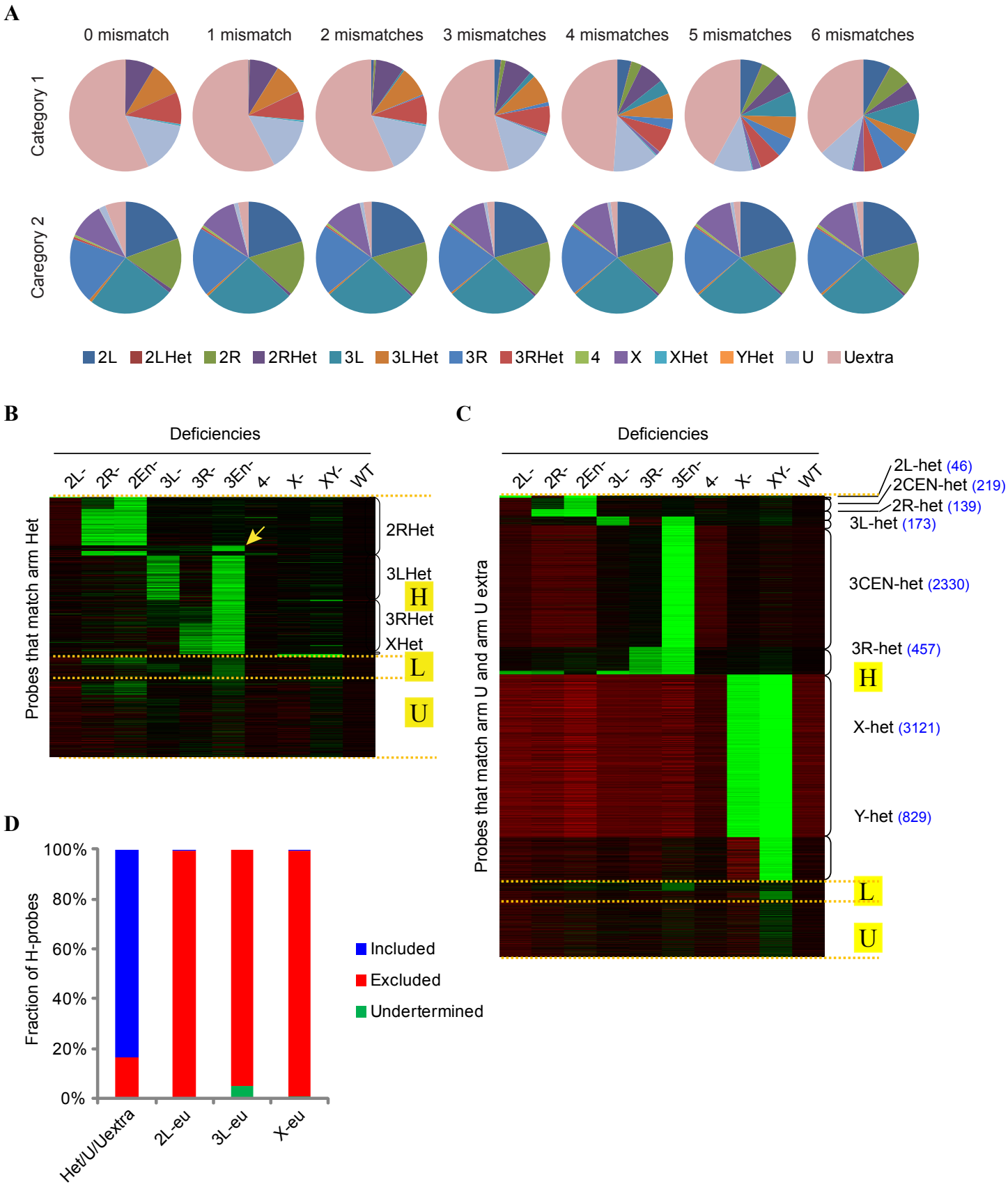
C



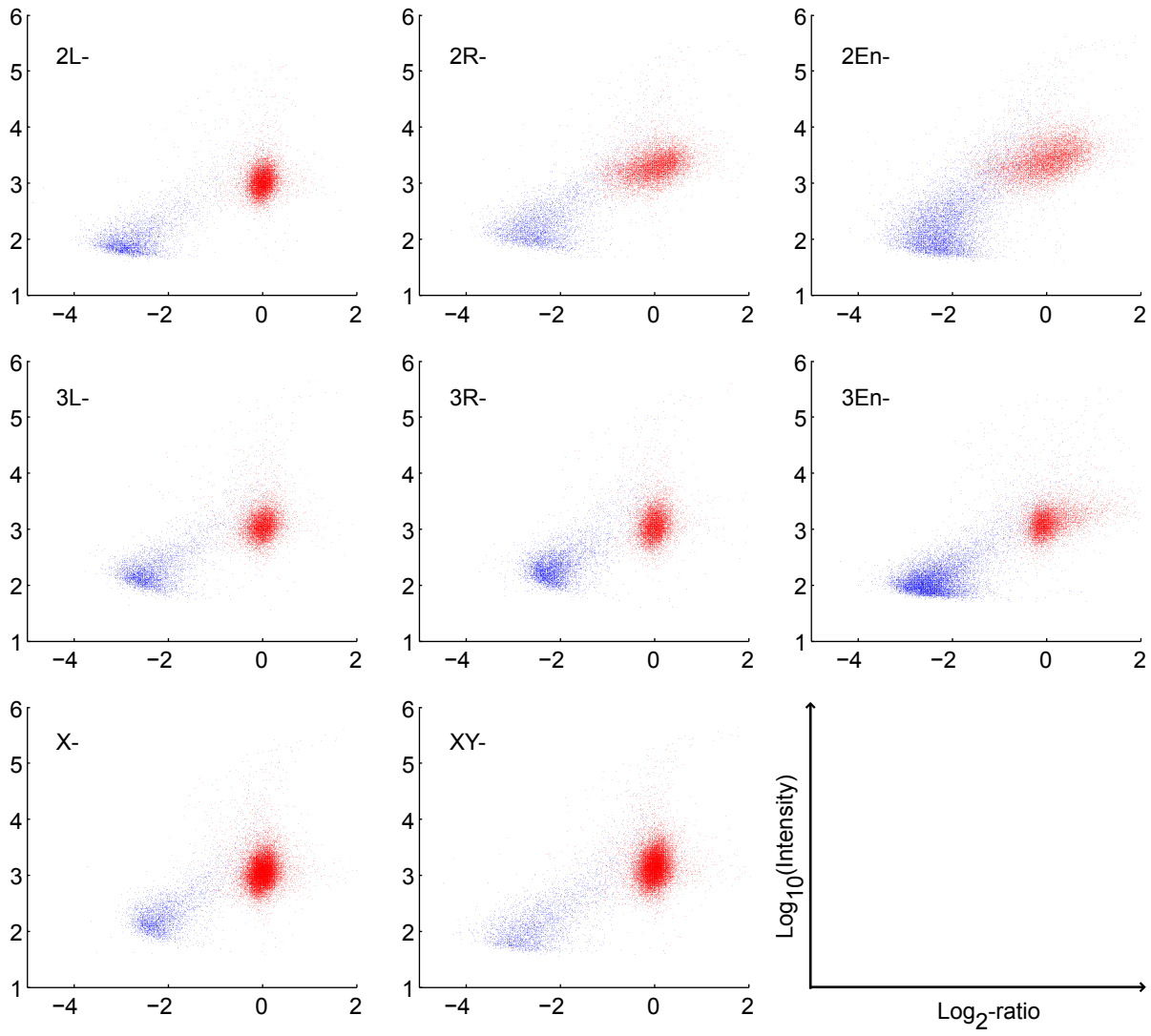
Supplemental Figure 3



Supplemental Figure 4

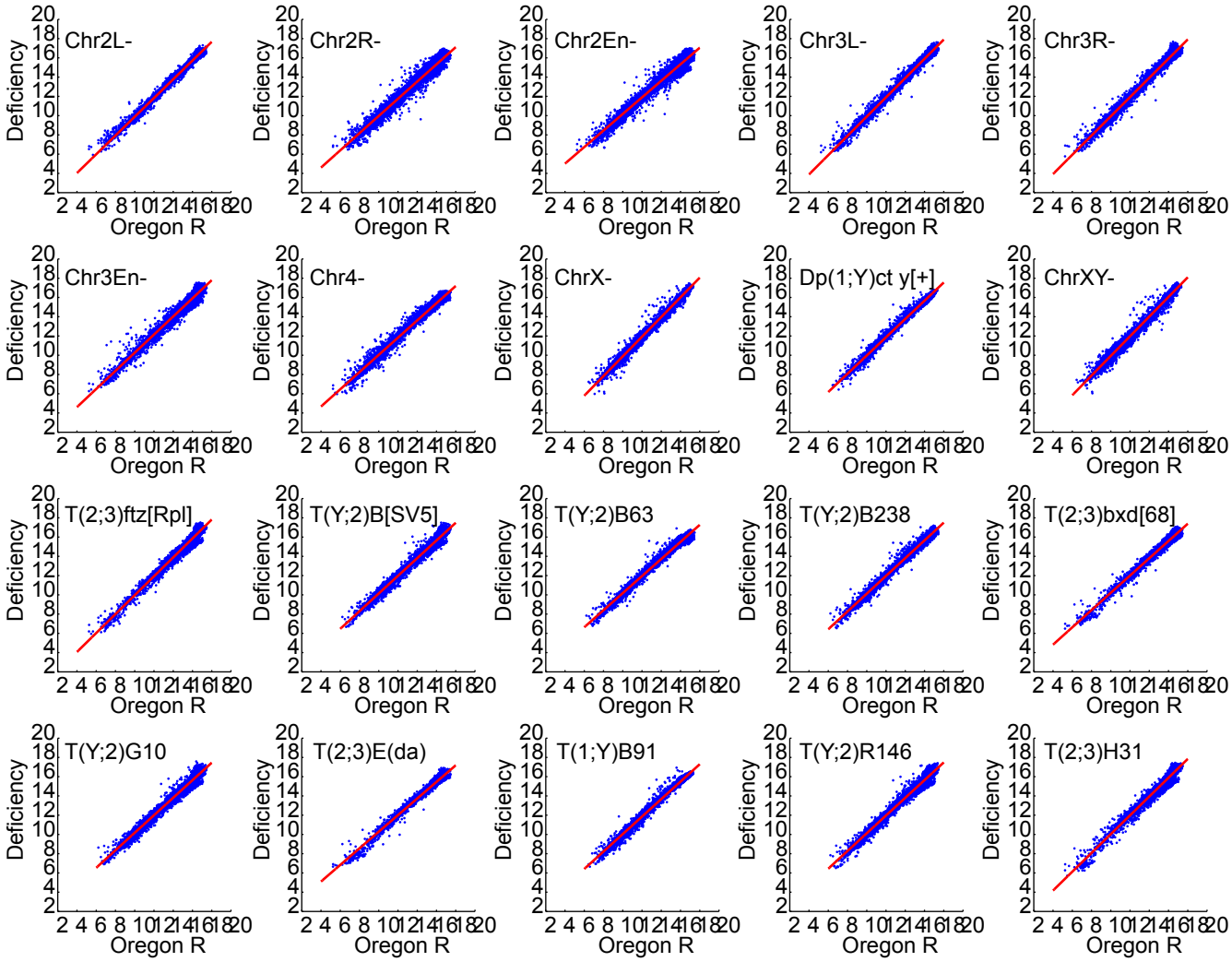


Supplemental Figure 5

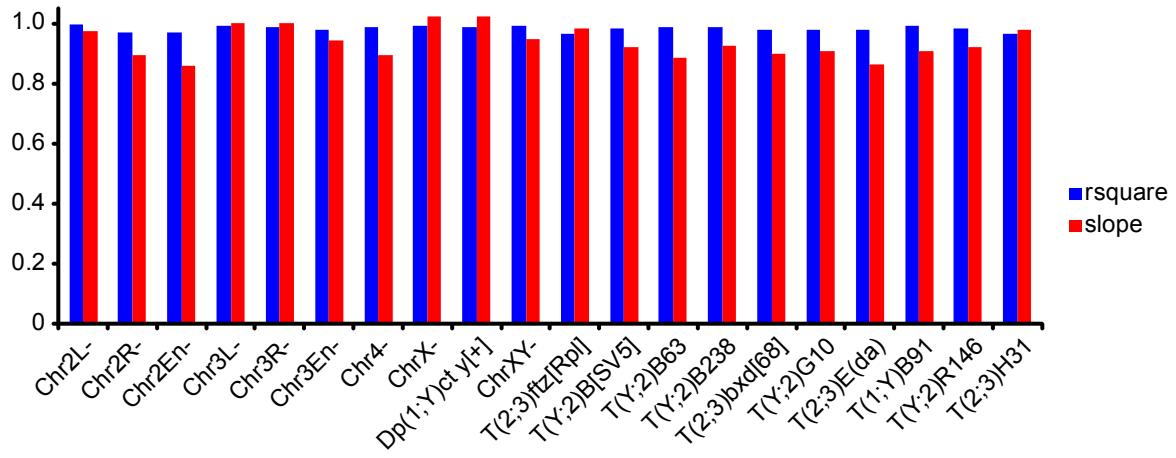


Supplemental Figure 6

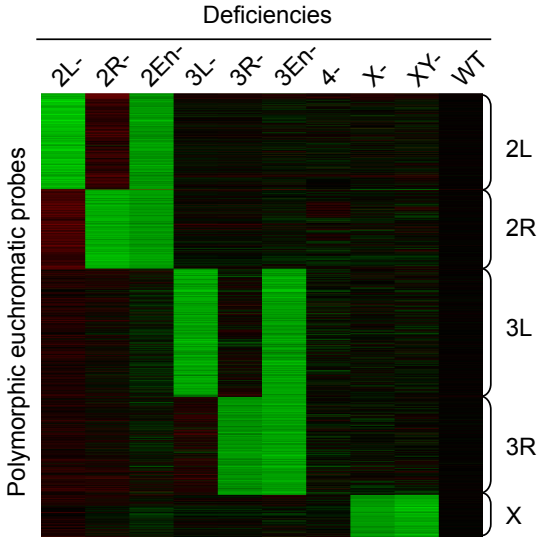
A



B

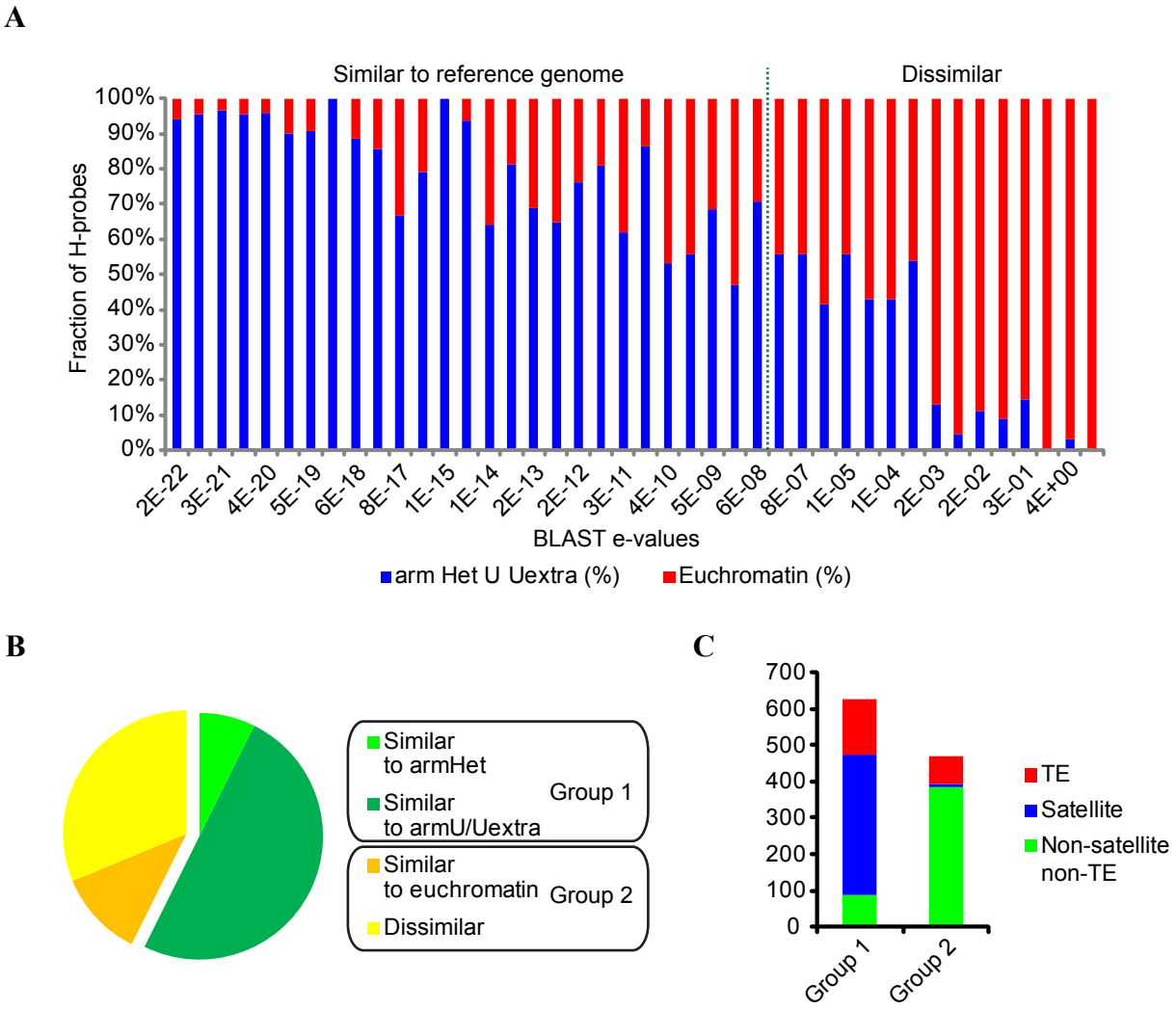


Supplemental Figure 7



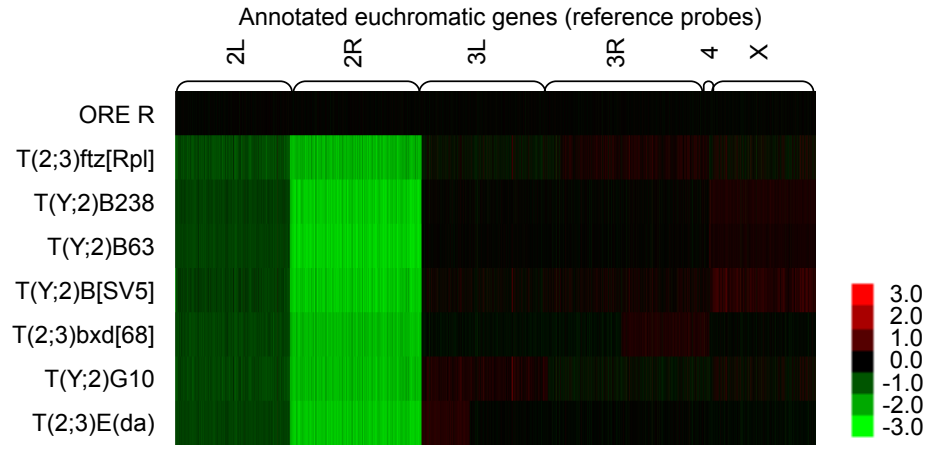
Chr Arm	Total	Correct	%Correct
2L	1794	1745	97.3%
2R	1490	1428	95.8%
3L	2404	2075	86.3%
3R	1834	1655	90.2%
X	788	774	98.2%

Supplemental Figure 8

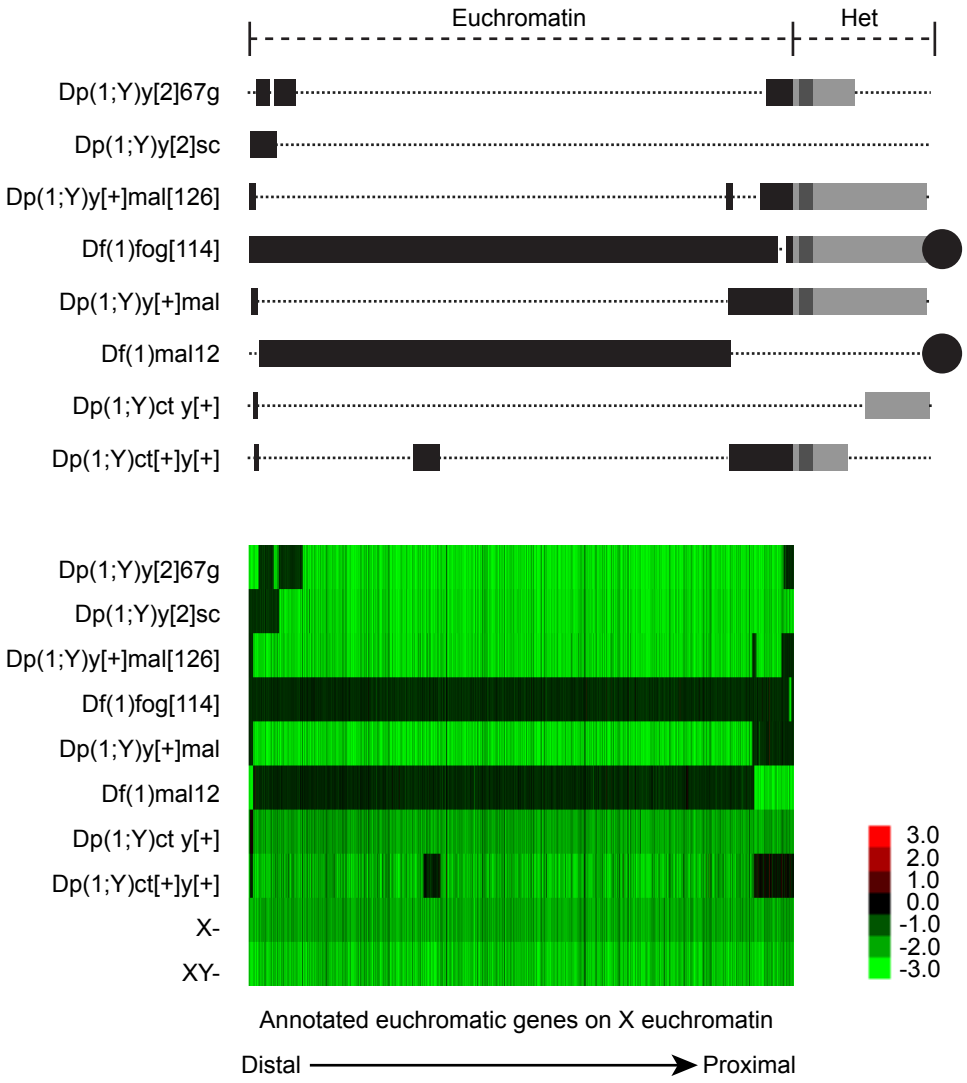


Supplemental Figure 9

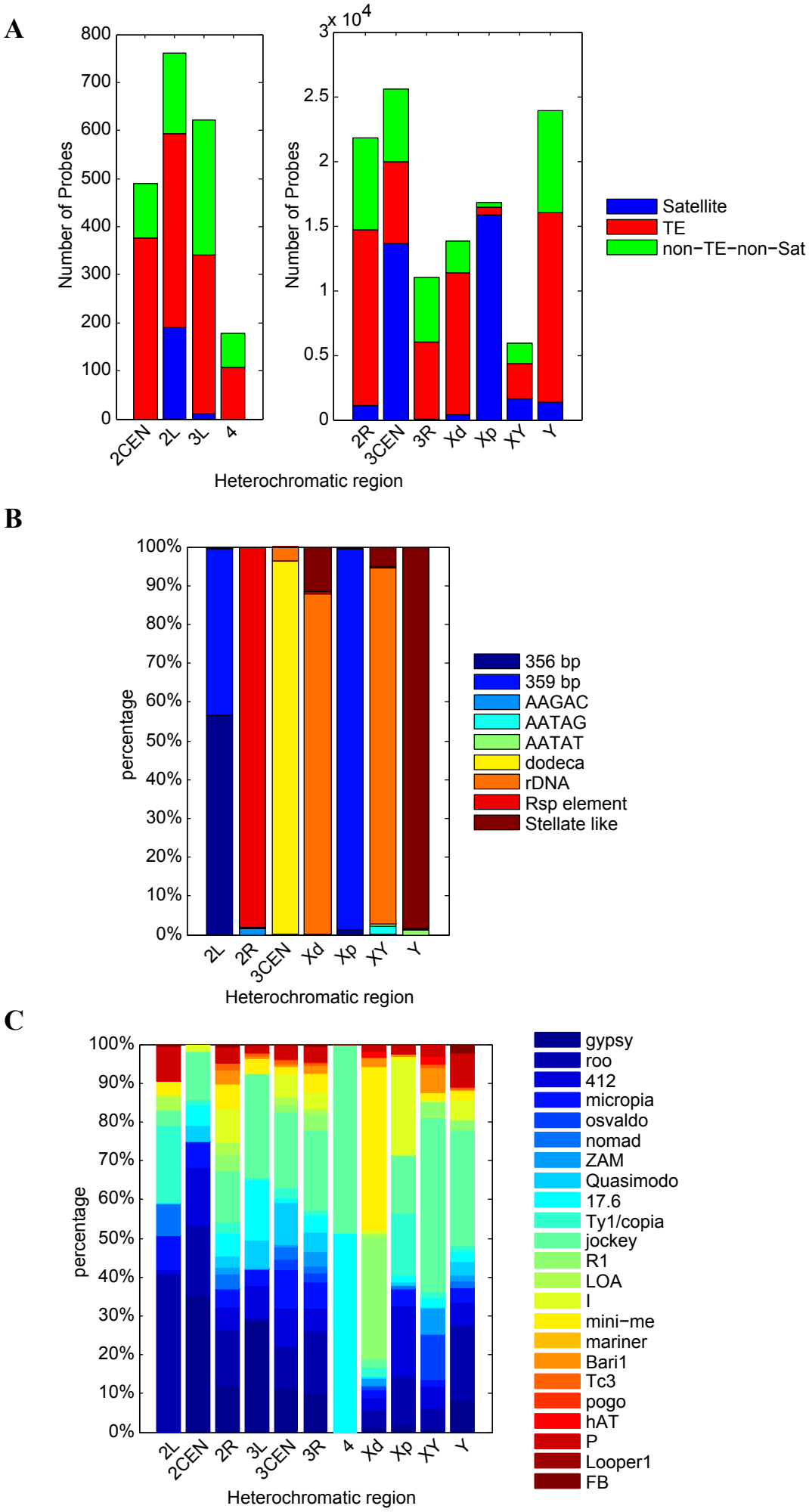
A



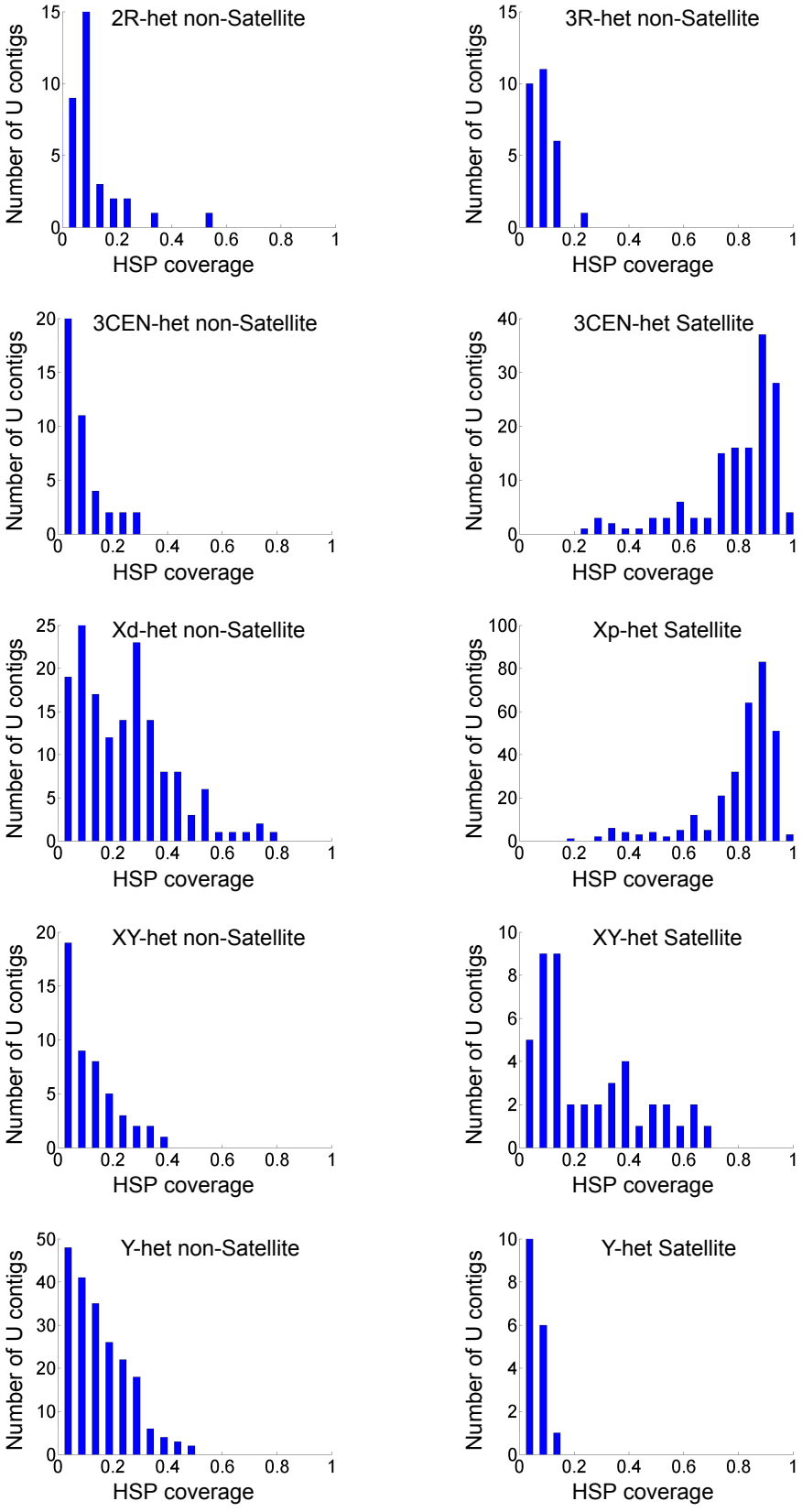
B



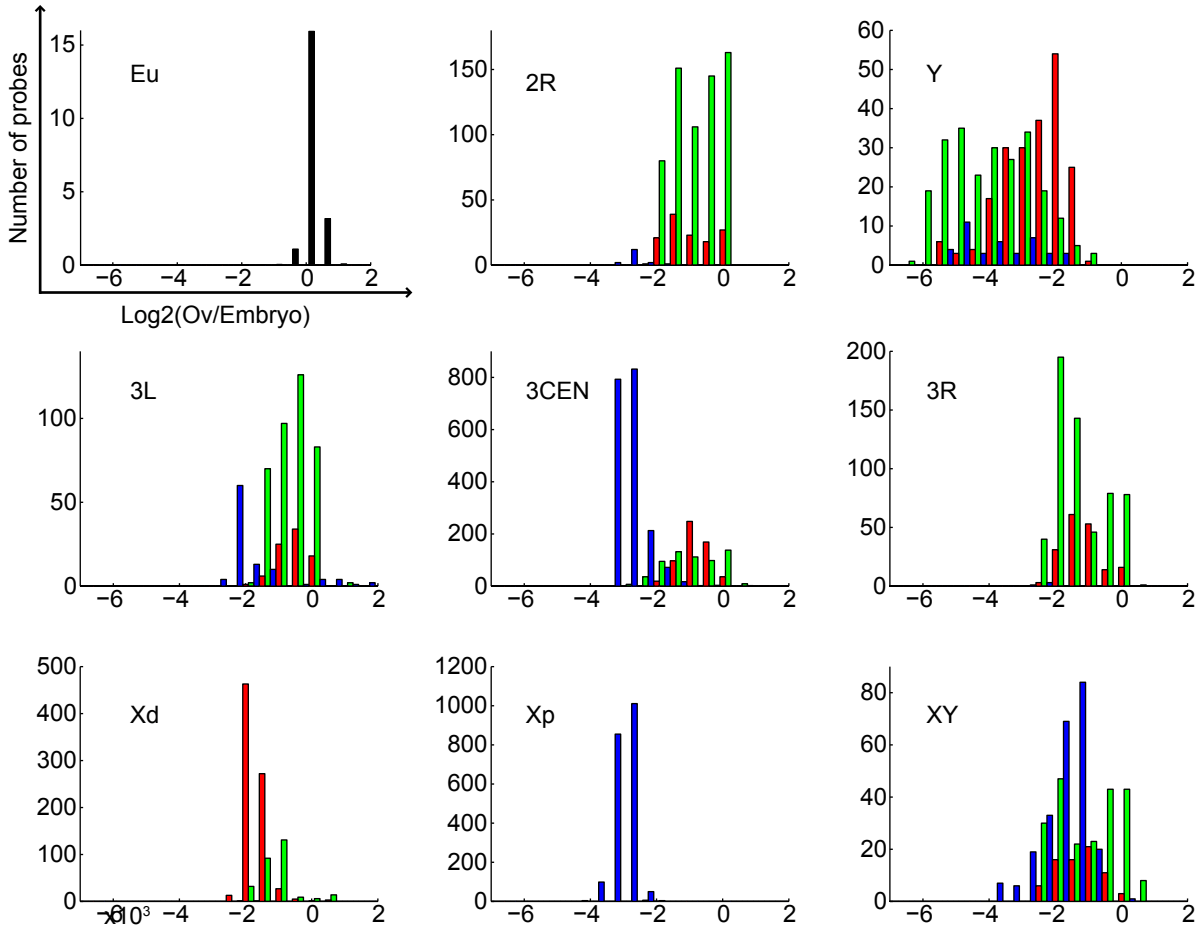
Supplemental Figure 10



Supplemental Figure 11

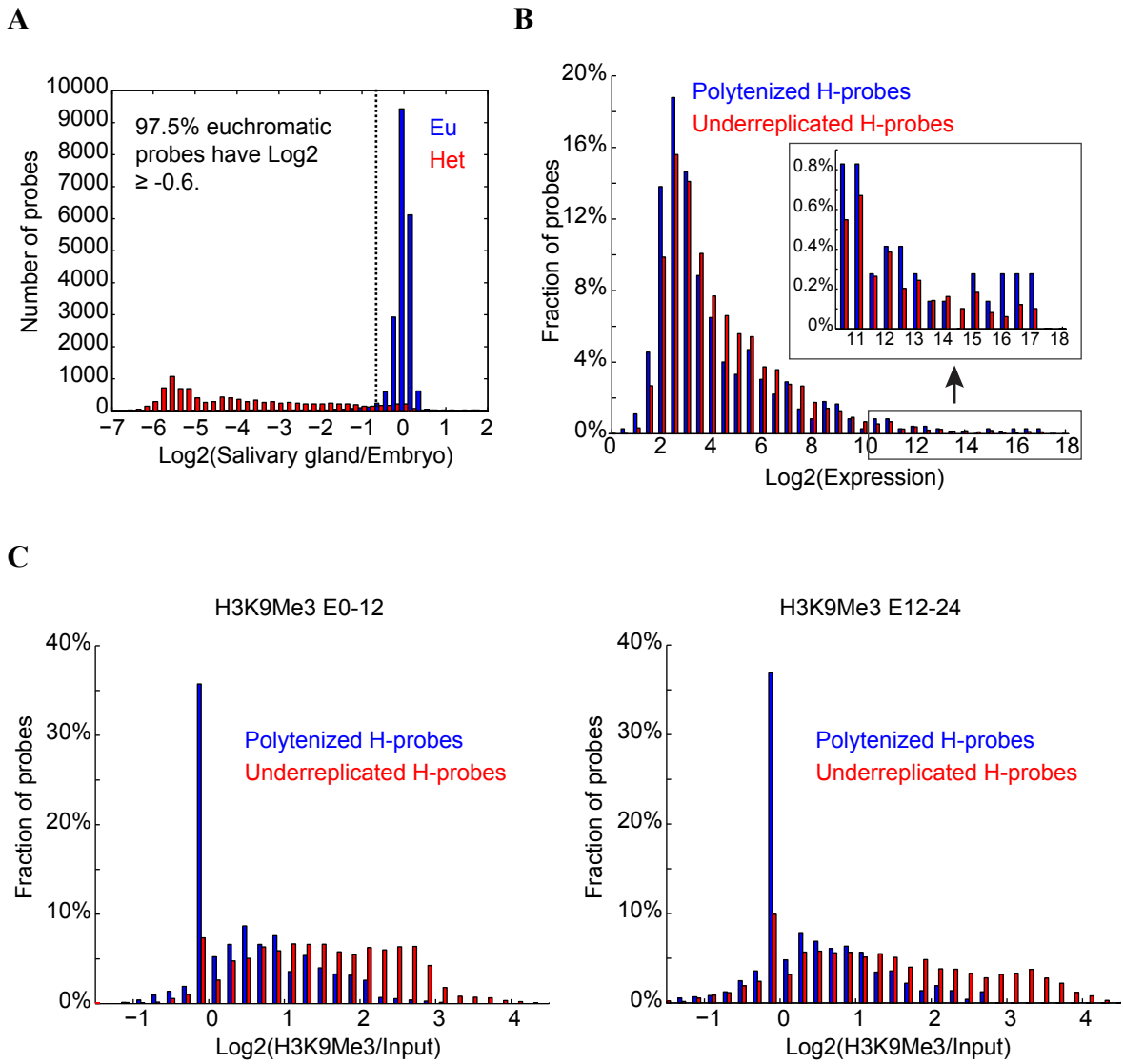


Supplemental Figure 12

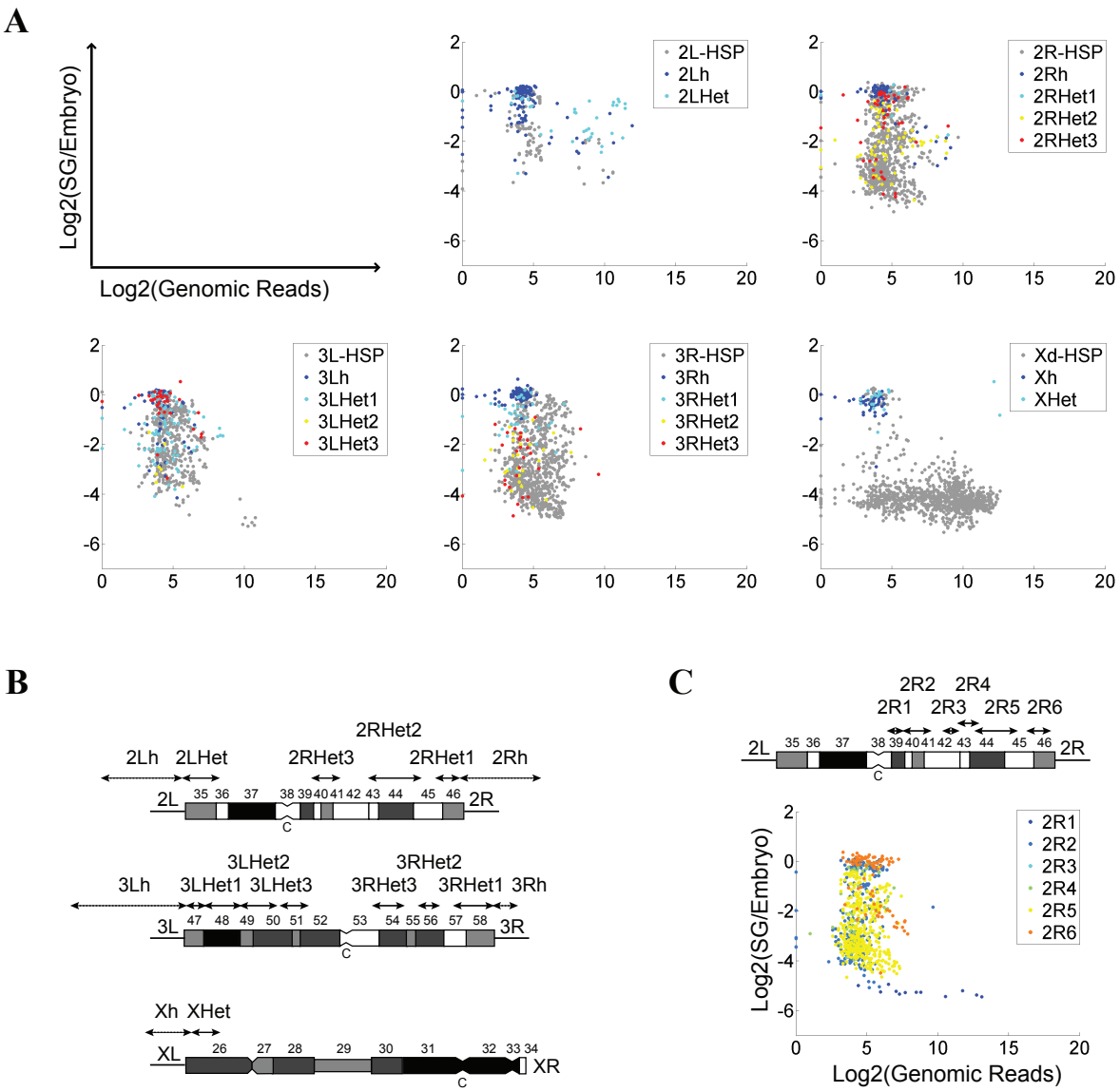


Red: TE Blue: Satellite + rDNA Green: non-Satellite non-TE

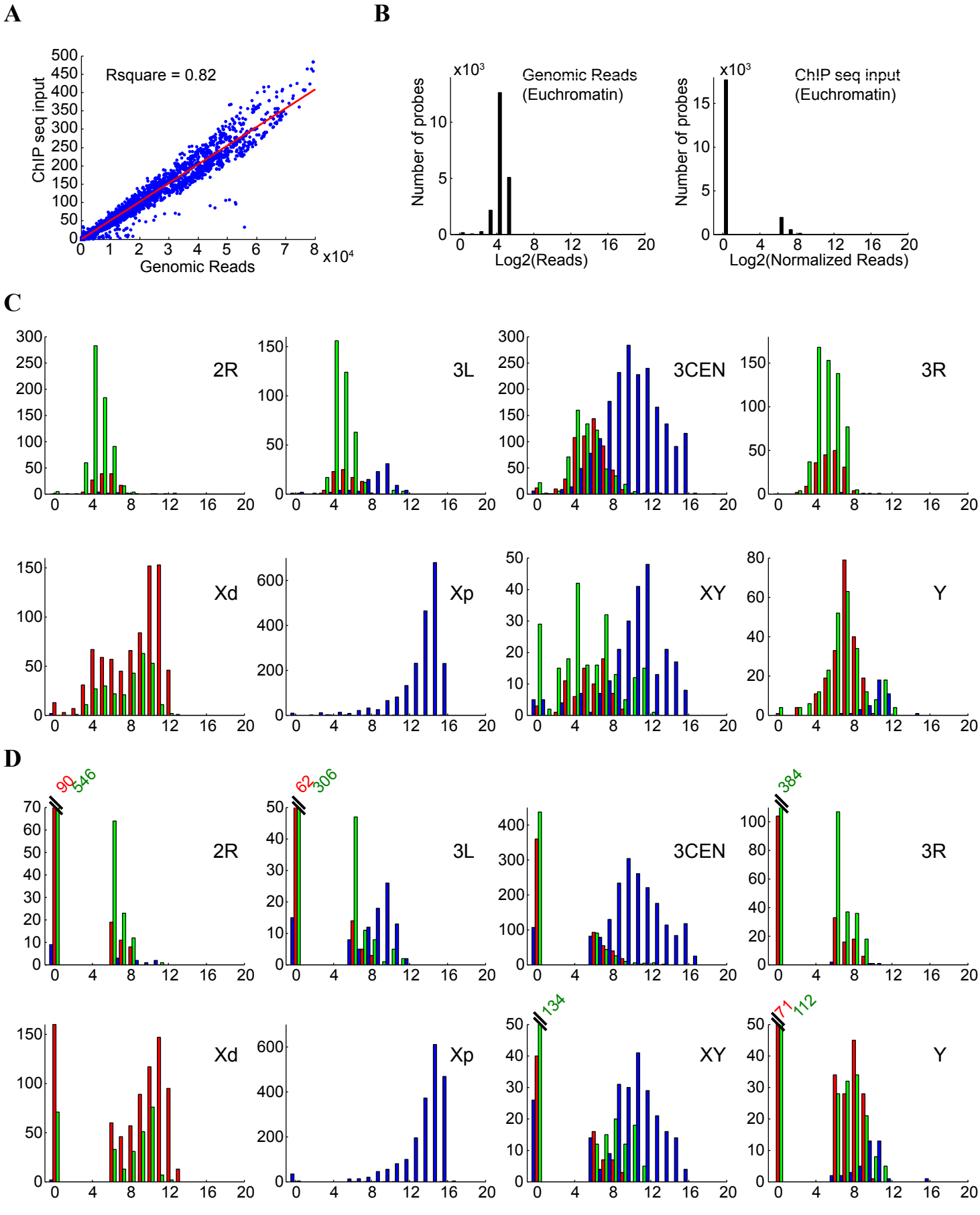
Supplemental Figure 13



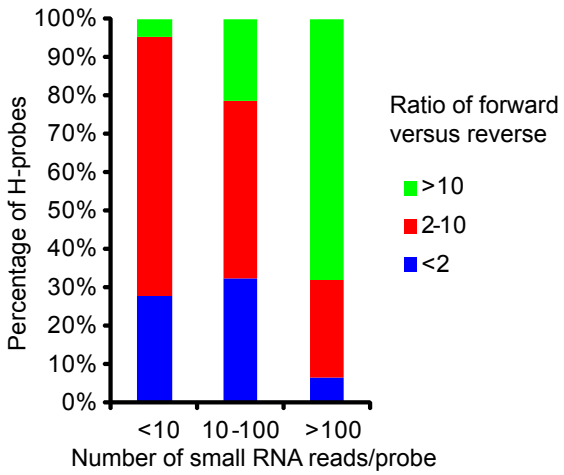
Supplemental Figure 14



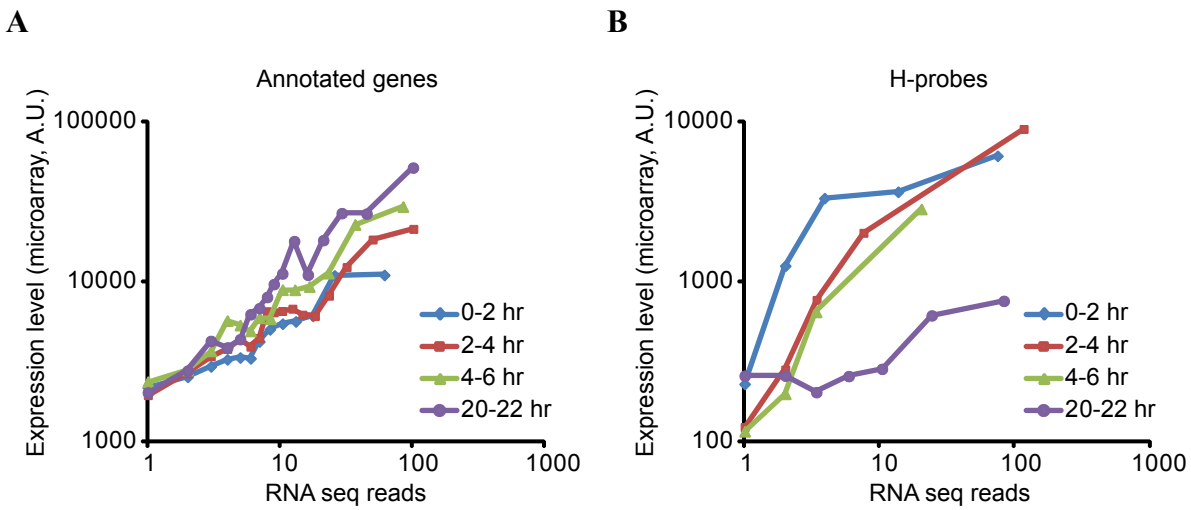
Supplemental Figure 15



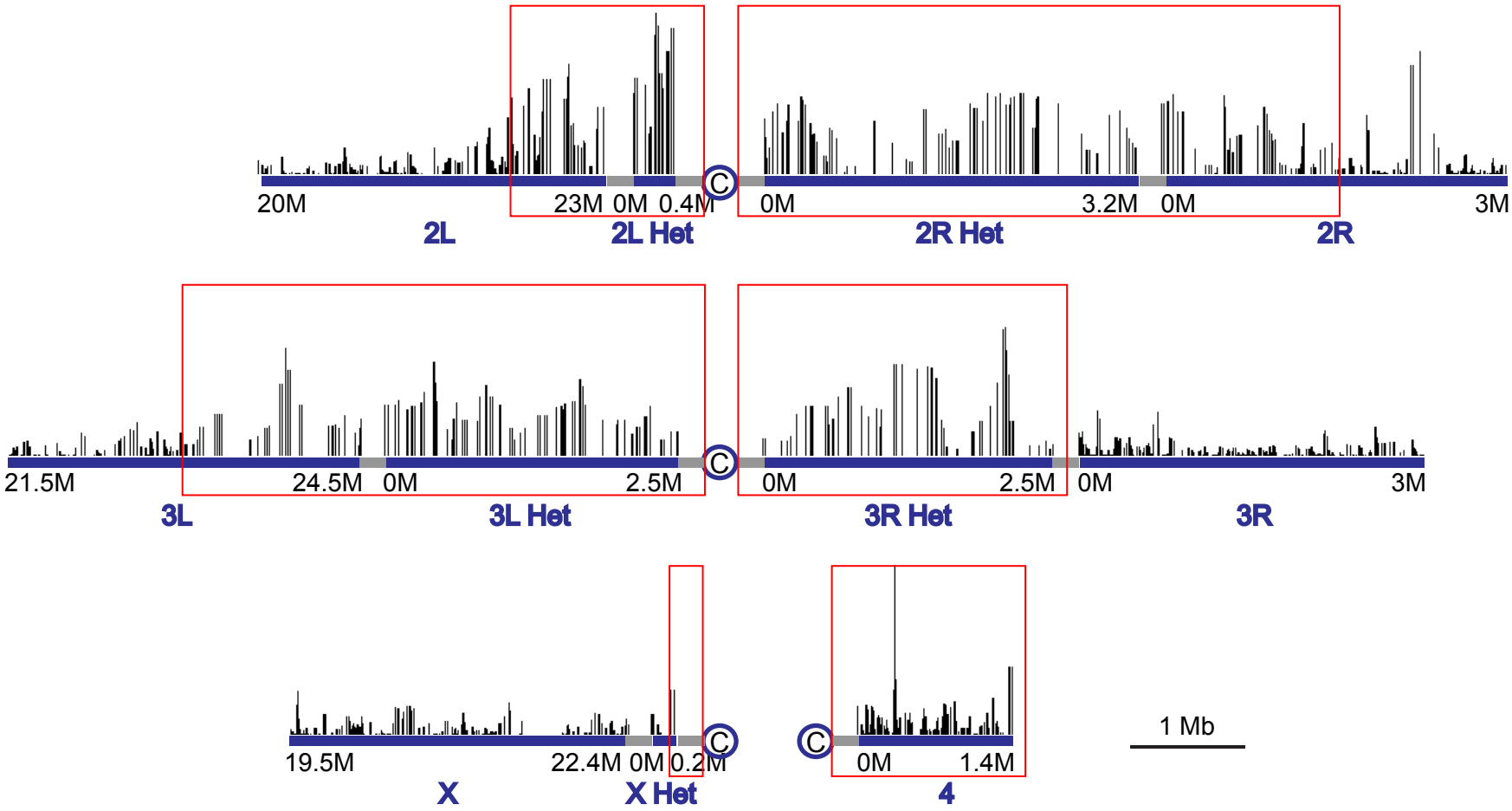
Supplemental Figure 16



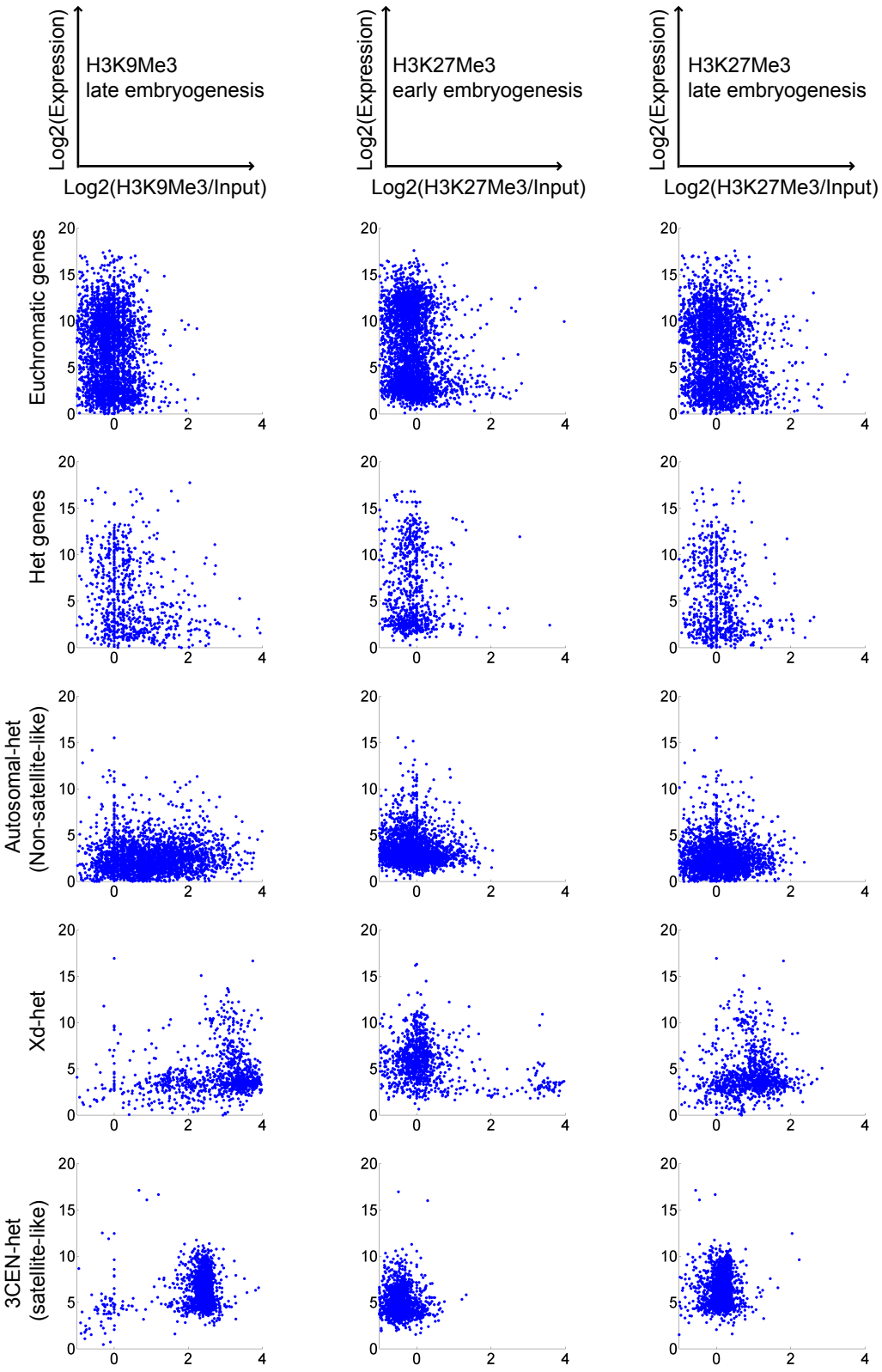
Supplemental Figure 17



Supplemental Figure 18



Supplemental Figure 19



Supplemental Figure 20