Published in final edited form as:
Science. 2007 June 15; 316(5831): 1586. doi:10.1126/science.1139815.

The Release 5.1 Annotation of *Drosophila melanogaster* Heterochromatin

Christopher D. Smith[1,2], **ShengQiang Shu**[3], **Christopher J. Mungall**[3], and **Gary H. Karpen**[2,4,*]

[1]Department of Biology, San Francisco State University, San Francisco, CA 94132, USA

[2]Drosophila Heterochromatin Genome Project, Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[3]National Center for Biomedical Ontology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[4]Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, CA 94720, USA

Abstract

The repetitive DNA that constitutes most of the heterochromatic regions of metazoan genomes has hindered the comprehensive analysis of gene content and other functions. We have generated a detailed computational and manual annotation of 24 megabases of heterochromatic sequence in the Release 5 *Drosophila melanogaster* genome sequence. The heterochromatin contains a minimum of 230 to 254 protein-coding genes, which are conserved in other Drosophilids and more diverged species, as well as 32 pseudogenes and 13 noncoding RNAs. Improved methods revealed that more than 77% of this heterochromatin sequence, including introns and intergenic regions, is composed of fragmented and nested transposable elements and other repeated DNAs. *Drosophila* heterochromatin contains "islands" of highly conserved genes embedded in these "oceans" of complex repeats, which may require special expression and splicing mechanisms.

The goal of genome annotation is to identify sequence features that have a biological role in the organism, but a telomere-to-telomere DNA sequence is not yet available for complex metazoans, including humans. The missing genomic "dark matter" is the heterochromatin, which is generally defined as repeat-rich regions concentrated in the centric and telomeric regions of chromosomes. Centric heterochromatin makes up at least 20% of human and 30% of fly genomes, respectively; thus, even for well-studied organisms such as *Drosophila melanogaster*, fundamental questions about gene number and global genome structure remain unanswered.

Once considered "junk" DNA, it is now clear that heterochromatin contains essential genes and also contributes to genome stability and chromosome segregation (1–3). In addition, heterochromatin participates in RNA interference mechanisms that epigenetically repress gene and transposable element (TE) expression, which may "immunize" genomes against the invasion and expansion of selfish DNA elements (4,5). Despite differences in heterochromatin

*To whom correspondence should be addressed. karpen@fruitfly.org.

Supporting Online Material www.sciencemag.org/cgi/content/full/316/5831/1586/DC1 Materials and Methods SOM Text SOM Tables (D to F, I, and J) References

sequence composition between genomes, commonalities in the structures, chromatin modifications, and presence of genes suggest that *D. melanogaster* heterochromatin is an excellent model for studying repeat-rich genomic DNA in other species, including the 40% repetitive human euchromatin (6,7).

Annotation overview

The Drosophila Heterochromatin Genome Project has generated 16 megabases (Mb) of finished or near-finished heterochromatin sequence from *D. melanogaster*, as well as 8 Mb of draft whole genome shotgun (WGS) heterochromatic assemblies (8,9). We performed computational and manual curation to produce the Release 5.1 annotation of this 24 Mb of heterochromatin sequence, which excludes degenerate sequence reads not incorporated into the final sequence assembly (ArmUextra) (10,11). Several repeat-finding programs were implemented, including RepeatRunner (12) and Tandem Repeats Finder (TRF) (13). New data from the research community were incorporated, including GenBank third-party annotations and Heidelberg gene predictions (14). Lastly, the conservation of *D. melanogaster* heterochromatin genes was assessed by identifying putative gene orthologs in more than 16 species [supporting online material (SOM) text and data].

The annotations include protein-coding genes, non–protein-coding RNAs (ncRNAs), repetitive sequences, and other functional elements. The majority of nonrepeat annotations (64%) mapped to a chromosome arm, including regions contiguous with the euchromatic arms (chromosome arm h; e.g., 2Rh) and internal scaffolds that have been cytologically localized to an arm (chromosome arm Het; e.g., 2RHet) (8,9). The remaining annotations are not mapped to a chromosome arm (36%) and reside on Arm U (Unmapped). Although the highly repetitive simple sequence component of *D. melanogaster* heterochromatin remains unassembled (8,9), few unique genes have been identified in these regions, and thus the current annotation is likely to include the majority of heterochromatic genes.

Protein-coding genes and comparative analysis

Previous studies indicated that at least 32 essential genetic loci are present in the heterochromatin (1). Using clone-based evidence [expressed sequence tag (EST) and cDNA] and predicted orthology, we annotated 613 protein-coding genes and gene fragments in the 24 Mb of Release 5 heterochromatin (Table 1,SOM text, and data) (11). Currently, 41% of these genes have supporting clone-based evidence, including 137 annotations with full-length cDNAs generated by the Berkeley Drosophila Genome Project (BDGP) (15). The incorporation of new EST sequences from Exelixis (16), which were generated with random primers, was instrumental for finding new internal splice sites and missed exons and was used to refine at least 43 gene models (Fig. 1) (11). New evidence resulted in 16 cases of Release 3.2b genes that were merged into seven larger Release 5.1 genes (e.g., CG41520 in Fig. 1) and two genes from Release 3.2b that were split into four smaller genes in Release 5.1.

A subset of the annotations represents single-exon fragments of coding sequences present in small scaffolds, and these fragments should not yet be considered as complete genes. We classified gene annotations as either single-exon genes (186) or multi-exon genes (427) in order to more conservatively estimate higher-quality complete gene models without losing information about potentially incomplete genes (Fig. 2). Heterochromatin single-exon genes are currently less supported by clone-based evidence; roughly 50% of heterochromatic multi-exon genes and 61% of euchromatic single-exon genes had EST or cDNA support, compared to only 20% of single-exon genes.

Because many genes lacked EST or cDNA support, the translated basic local alignment sequence tool (TBLASTN) was used to identify putative orthologs for heterochromatin genes

in 16 other insect genomes and more distantly related vertebrate species (17) (Fig. 3A). We defined orthologs as unique, top-scoring TBLASTN-identified sequences, although no evidence currently exists to support conserved function. Overall, more than 99% of the annotations had an ortholog identified in at least one species (Fig. 3A). Orthologs were identified for 86 to 98% of heterochromatin protein-coding genes in the four species (*D. simulans*, *D. erecta*, *D. sechellia*, and *D. yakuba*) most closely related to *D. melanogaster*, and 55 to 70% of genes are conserved in the more distantly related Drosophilids. In addition, 22 to 46% of genes are conserved in distantly related insects (such as the silkworm, mosquito, honeybee, wasp, and beetle), and 13% of all protein-coding genes had significant alignments to proteins in even more diverged species (Fig. 3A). We conclude that the majority of *D. melanogaster* heterochromatic genes are highly conserved in insect lineages that span 300 million years of evolution and that a surprising number share significant similarity with proteins from vertebrate species that diverged more than 900 million years ago (18).

The conservation patterns were bimodal, such that nearly 20% of protein-coding genes were conserved in fewer than 4 species, whereas more than 30% were conserved in all 16 insect species (Fig. 3B). This trend suggests the existence of distinct groups of *Drosophila* lineage-specific genes and another subset of ultraconserved insect genes. Closer examination of the 163 genes with orthologs in four or fewer species revealed that 93% had orthologs only in the four species most closely related to *D. melanogaster* (Fig. 3B, SOM text, and data). Conservation in only the Drosophilid lineage or melanogaster subgroup may identify more recently evolved genes (SOM text). Indeed, some open reading frames (ORFs), such as odorant-binding proteins or receptors, tend to evolve faster than do other genes and are often found in only a single species (19,20). More in-depth study will be required to determine whether the lineage-specific Drosophilid genes are conserved because of function or because there has been insufficient time for ORFs to diverge.

Analysis of orthology data, gene prediction, and cDNA and EST clone evidence shows that 96% of heterochromatin multi-exon genes and 89% of single-exon genes are supported by two or more types of evidence (Fig. 2, A and B). These results demonstrate that single-exon genes probably represent parts of bona fide, full-length genes that will be merged into more complete genes as the sequence assemblies and cDNA resources are improved, as observed for fragmented annotations in prior releases (e.g., CG41520 in Fig. 1). Comparative genomic support for small single-exon genes also suggests that there may be many more conserved exons in other parts of less well assembled Drosophilid genomes (21).

We inferred 1030 conserved introns between well-conserved exons of the orthologs identified in the 16 insect genomes. Heterochromatin gene introns are, on average, five times longer than introns in euchromatic genes [4949 versus 1149 base pairs (bp)], and may be as long as 1 Mb (22). The longest contiguous intron we have identified is 224,977 bp in the *Snap25* gene (11). Intron lengths are highly conserved among euchromatin gene orthologs (Fig. 3C), whereas heterochromatin intron lengths were correlated in species closely related to *D. melanogaster* but were poorly correlated for more distantly related species. For example, intron lengths for *D. erecta* orthologs were correlated in both euchromatin [correlation coefficient (r) = 0.48] and heterochromatin (r = 0.58), whereas *D. virilis* gene intron lengths were correlated in euchromatin (r = 0.37) but not in heterochromatin (r = 0.14) (Fig. 3C). A list of conserved introns, their average lengths, and the percent conservation of flanking coding exons is provided (SOM text and data).

One of the features of heterochromatic genes is that introns and intergenic regions are composed almost entirely of repeated sequences, predominantly fragmented TEs (Fig. 1). There was no appreciable difference in the average repeat density or composition of intronic (56%) versus intergenic sequences (63%) (11). Changes in intron length are most often due to TE insertions

and excisions and simple repeat expansions. The high repeat content of introns and regulatory regions suggests that regulation of heterochromatic gene expression may differ from euchromatic genes. We identified 16 recursive splice sites (RSSs) shown to aid in the splicing of long introns (23), including 8 RSSs located in long heterochromatin gene introns ranging from 11 to 166 kilo-bases (kb). Of particular interest are three RSSs predicted in the 23.6-kb intron of CG40120, only one of which was previously predicted because of gaps in the sequence assembly (23). Fifty-six percent of the RSS motifs were embedded within retrotransposons, suggesting that cis TE sequences may be used to splice out TEs that invade heterochromatin genes.

Non–protein-coding genes

We identified 13 putative non–protein-coding genes in the heterochromatin (Table 1), defined as single-copy genes with EST or cDNA support that contained protein-coding ORFs that were substantially shorter than the length of the transcript. Spliced ESTs and cDNAs were identified for 11 of the ncRNA annotations, excluding the possibility of false positives generated by the priming of polymerase chain reaction products from adenine (A)–rich regions. Analysis of the two unspliced ncRNAs suggested that the clones were not primed from genomic A-rich regions. 5 out of 13 ncRNAs contain ~100 bp of the 600-bp *INE-1* TE consensus in the 3' end of the transcript, which is insufficient for autonomous transposition but may represent a conserved motif. The ncRNA gene CR40375 is nested in the intron of the protein-coding gene CG41520 (Fig. 1). Recent reports (24) suggest that thousands of transcribed regions are coregulated with genes and may represent missing coregulated genes or novel 5' untranslated region (UTR) exons. We found the converse case; new cDNA evidence suggests that the 5' UTR exon of the Release 3.2b gene CG40084 represents a ncRNA (CR41594) (11). Further analysis will be required to determine whether these ncRNA annotations represent functional genes.

Pseudogenes

Protein-coding genes that are near-perfect but truncated copies of genes found elsewhere in the genome were annotated as pseudogenes. The *D. melanogaster* genome has far fewer identified pseudogenes than do other metazoan genomes (25). For example, the three-gigabase human genome is estimated to contain ~20,000 pseudogenes of various types (26), whereas the 120-Mb *D. melanogaster* R4.3 euchromatin has only 51 annotated pseudogenes (27). We identified 32 putative pseudogenes in the *D. melanogaster* heterochromatin sequence, representing a threefold increase in pseudogene density in heterochromatin versus euchromatin (1.3 versus 0.425 pseudogenes per Mb) (Table 1 and SOM text). The enrichment of repetitive sequences in both fly heterochromatin and human euchromatin, relative to *D. melanogaster* euchromatin, may facilitate the formation of pseudogenes by increasing the probability of large- or small-scale duplications.

Repeats and transposable elements

Repeats were defined by significant alignment to known RepBase repetitive sequences identified by RepeatMasker (28), BLASTX homology to TE proteins, or TRF (29) results (10). The application of these improved methods demonstrated that 18 Mb (77%) of the annotated heterochromatin could be classified as repetitive or transposable elements, a 4-Mb (20%) increase compared with the previous annotation (1). About 50 previously identified protein-coding genes have been reannotated as repetitive features. For example, CG40388 was annotated as a protein-coding gene in Release 3.2b but annotated as a *1731* repeat in Release 5.1 (Fig. 1). The euchromatin sequence was previously reported to consist of 3.86 to 6% of repetitive sequence (30,31). Our methods provide better recognition of TE fragments, small tandem repeats, and short ORFs from degenerate TEs, and they identified 7% of the Release 5 euchromatin as repetitive (Fig. 4), similar to other estimates (32). The sequenced *D.*

melanogaster heterochromatin has an overall repeat and TE content more than 10 times that of fly euchromatin and is more similar to the repeat density in human euchromatin (40%) (6, 7).

We measured the repeat content and gene distribution across the heterochromatin sequence in 100-kb sections and calculated the average density for each region (Fig. 4). Lower overall repeat content was observed for regions more distal from the centromere, especially for the heterochromatic regions on chromosome arm 3Rh (Fig. 4). In addition, the unmapped scaffolds had a higher repeat content than the heterochromatin regions that have been mapped to the chromosome arms (Fig. 4). In general, there was a strong inverse correlation between the repeat and gene content of a region ($r = -0.89$) (Fig. 4). Euchromatin had an average gene density of 12.6 genes per 100 kb, whereas heterochromatin contained 1.8 to 4.4 genes per 100 kb (2.9 genes per 100 kb overall). Exceptions include Xh and XHet, with 9.2 and 6.4 genes per 100 kb. Chromosome arm 3Rh had an average gene content even higher than that of typical euchromatin (19 genes per 100 kb) and a 20% average repeat content, which is significantly lower than the rest of the heterochromatin (Fig. 4).

We further categorized the average percentage of retrotransposons, DNA transposons, and other repeats (Fig. 4). Roughly two-thirds (16 Mb, 66%) of the heterochromatin is composed of retrotransposon sequences [33% long terminal repeats (LTRs) and 33% long interspersed nuclear elements (LINEs)]. DNA transposons are overrepresented on the relatively repeat-rich euchromatin regions of the fourth chromosome of *D. melanogaster* (33) but constitute only 15% of the Release 5 heterochromatin, which does not include the fourth chromosome heterochromatin (8,9). TRF-identified tandem repeats and satellite repeats make up ~10% of the available heterochromatin sequence, which is significantly higher than in the euchromatin sequence (~3%), especially in the proximal centric regions of chromosomes 2 and 3 and the Y chromosome (Fig. 4). Calculations of tandem-repeat content are likely to be an underestimate, because WGS3 sequence underrepresents the difficult-to-clone satellite DNA and tandem-repeat regions. Unlike the Y chromosome and autosomal heterochromatin, the available X chromosome sequence is not enriched for tandemly repeated sequences. Our repeat analysis indicates that nearly all of the ArmU and ArmUExtra (10) sequence is composed of repetitive sequences, further suggesting that we have identified most of the unique sequence available in the *D. melanogaster* heterochromatin.

The majority of repetitive TE-like sequences in heterochromatin is not intact. We found 202 full-length TEs in the heterochromatin (2% of heterochromatic TEs), compared with 361 full-length TEs reported for the nonpericentromeric euchromatin (20.6% of euchromatic TEs). The most recent annotation of the Release 4 euchromatin identified nests of TEs that were fragmented, interdigitated, and transposed into one another (32). Our manual curation identified 846 repeat nests in newly sequenced regions of the Release 5 heterochromatin, compared with 112 nested TEs in the euchromatin. We annotated 117 instances where there were two nested TEs (i.e., a TE jumped into a TE that itself had jumped into a TE) and 17 instances where four or more TEs were nested (Fig. 1).

Conclusions

The assembly of a more complete genome sequence (9) and an integrated annotation set for *D. melanogaster* provides a reference set of genes and other features that will be useful for investigating the biological functions of heterochromatin in flies and other organisms. These results are now fully integrated, with information about the euchromatin, in FlyBase (34) and GenBank (35).

These results demonstrate that repetitive and TE sequences constitute at least 77% of the 24 Mb of heterochromatin sequence. Although there are more full-length TEs and TE nests in the centric heterochromatin relative to the euchromatin, it appears that most heterochromatic TEs are fragmented and not capable of autonomous transposition. As with the euchromatin, the repetitive sequences in heterochromatin are dominated by LTR- and LINE-like retrotransposons. We have identified a substantial amount of tandemly repeated sequences in the most proximal centric heterochromatin of the second, third, Y, and unmapped chromosomes, but not in the currently sequenced X chromosome. These carefully annotated repetitive regions provide opportunities for more in-depth analysis of their functions and evolution. For example, a recent study showed that at least 80% of *piwi*-associated small RNAs, which regulate transposon activity, map to the release 5 heterochromatin (4).

We found that more than 99% of the protein-coding genes and gene fragments are conserved in other insect species. A subset of the protein-coding genes (35%) appears to be present in only the *melanogaster* group of Drosophilids, suggesting recent evolution. Of the 613 protein-coding genes identified, 137 are considered complete by the criterion of having a full-length cDNA. Another 115 protein-coding genes are only partially supported by clone evidence, with ~360 genes lacking EST or cDNA evidence. Thus, there are 475 annotated protein-coding genes that are likely to represent fragments from larger genes. Based on an average of ~four to five exons observed for complete euchromatin and heterochromatin genes, we estimate that these 475 fragments represent 95 to 119 full-length genes, and thus we approximate that there are 230 to 256 protein-coding genes in the currently sequenced heterochromatin.

The gene density in heterochromatin is substantially lower than it is in euchromatin and is inversely correlated with repeat content. Based on our RepeatRunner analysis, only 9% (2.2 out of 24 Mb) of the Release 5 heterochromatin is a unique sequence, of which 60% (1.3 Mb) is annotated as exons; thus, only 5.4% of the sequenced heterochromatin is exonic, compared with 25% of the euchromatin. The average protein-coding and ncRNA gene density for the annotated heterochromatin is 10 to 11 genes per Mb, compared with 127 genes per Mb in the euchromatin. We have identified 32 pseudogenes in the heterochromatin, including 8 in the poorly represented Y chromosome sequences, representing a density of pseudogenes that is at least three times that of euchromatin. The high repeat content of heterochromatin may provide recombination substrates that increase the frequency of tandem and segmental duplications.

Despite differences in gene density, there are many similarities between the basic structures and putative functions of euchromatic and heterochromatic genes. Based on cDNA-supported genes, it appears that euchromatin and heterochromatin genes have, on average, a similar number of exons and transcript variants per gene. In general, heterochromatic and euchromatic genes appear to encode a similar spectrum of functions, based on gene ontology (GO) analysis (Fig. 5). Some classes of genes are overrepresented in the heterochromatin, relative to the euchromatin. For example, heterochromatin genes are enriched 35-fold for putative membrane cation transporters domains (4 out of 308 heterochromatin domains versus 5 out of 13,500 euchromatin domains). Heterochromatic genes are also enriched for domains involved in DNA (53 domains) or protein binding (122 domains) that may regulate chromatin structure or function, including histone variants and proteins (Fig. 5, SOM text, and data) (11). This raises the intriguing possibility that heterochromatin may encode genes involved in its own establishment or maintenance.

Heterochromatin genes can reside in regions that approach 90% repeat content. Heterochromatin gene introns are usually composed of fragmented TE sequences (Fig. 1), are on average five times longer than euchromatin gene introns, and display less length conservation in inter-species comparisons. We found nine recursive splice site motifs nested in the long introns of heterochromatin genes, which may regulate splicing in repeat-rich

regions. The underlying mechanisms that allow essential genes to be expressed and regulated in otherwise silent chromatin remain unknown. Studying heterochromatin in other species promises to shed light on whether there are cis sequences that define or regulate boundaries between euchromatin and heterochromatin and if there are genic and nongenic regions of heterochromatin in other repeat-rich regions, including human euchromatin.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank E. Frise for maintaining the hardware and software used in these studies; M. Yandell for providing the specialized comparative genomics library–based software used in our analyses; A. Dernburg, D. Acevedo, J. Carlson, S. Celniker, R. Hoskins, and C. Kennedy for their helpful comments on the manuscript and input on annotations; and the members of the BDGP for cDNA sequencing. This work was supported by the National Human Genome Research Institute grant R01-HG000747 to C.D.S. and G.H.K. and NIH grant U54 HG004028-01 to S.S. and C.J.M.

References and Notes

1. Hoskins RA, et al. Genome Biol 2002;3:RESEARCH0085. [PubMed: 12537574]
2. The Arabidopsis Genome Initiative. Nature 2000;408:796. [PubMed: 11130711]
3. Sullivan BA, Blower MD, Karpen GH. Nat. Rev. Genet 2001;2:584. [PubMed: 11483983]
4. Brennecke J, et al. Cell 2007;128:1089. [PubMed: 17346786]
5. Elgin SC, Grewal SI. Curr. Biol 2003;13:R895. [PubMed: 14654010]
6. Lander ES, et al. Nature 2001;409:860. [PubMed: 11237011]
7. Venter JC, et al. Science 2001;291:1304. [PubMed: 11181995]
8. Carlson, HR., Jr., et al. 2006. (www.fruitfly.org/sequence/release5genomic.shtml)
9. Hoskins R, et al. Science 2007;316:1625. [PubMed: 17569867]
10. Materials and methods are available as supporting material on Science Online.
11. Supplemental data can be downloaded from ftp://ftp.dhgp.org/pub/DHGP/Science_2007_Supplemental_Data. Future updates will be released through www.flybase.net.
12. Smith CD, et al. Gene 2007;389:1. [PubMed: 17137733]
13. Benson G. Nucleic Acids Res 1999;27:573. [PubMed: 9862982]
14. Hild M, et al. Genome Biol 2003;5:R3. [PubMed: 14709175]
15. Berkeley Drosophila Genome Project. www.fruitfly.org/
16. Kopczynski, C., et al. www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=59874298
17. O'Brien KP, Remm M, Sonnhammer EL. Nucleic Acids Res 2005;33:D476. [PubMed: 15608241]
18. Peterson KJ, et al. Proc. Natl. Acad. Sci. U.S.A 2004;101:6536. [PubMed: 15084738]
19. Foret S, Maleszka R. Genome Res 2006;16:1404. [PubMed: 17065610]
20. Robertson HM, Wanner KW. Genome Res 2006;16:1395. [PubMed: 17065611]
21. Assembly, Alignment, and Annotation of Drosophilid Genomes. http://rana.lbl.gov/drosophila/wiki/index.php/Main_Page
22. Reugels AM, Kurek R, Lammermann U, Bunemann H. Genetics 2000;154:759. [PubMed: 10655227]
23. Burnette JM, Miyamoto-Sato E, Schaub MA, Conklin J, Lopez AJ. Genetics 2005;170:661. [PubMed: 15802507]
24. Manak JR, et al. Nat. Genet 2006;38:1151. [PubMed: 16951679]
25. Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M. Nucleic Acids Res 2003;31:1033. [PubMed: 12560500]
26. Torrents D, Suyama M, Zdobnov E, Bork P. Genome Res 2003;13:2559. [PubMed: 14656963]
27. Drysdale RA, Crosby MA. Nucleic Acids Res 2005;33:D390. [PubMed: 15608223]

28. Smit, AFA.; Hubley, R.; Green, P. RepeatMasker Open-3.0. www.repeatmasker.org/
29. Benson G. Nucleic Acids Res 1999;27:573. [PubMed: 9862982]
30. Kaminker JS, et al. Genome Biol 2002;3:RESEARCH0084. [PubMed: 12537573]
31. Quesneville H, et al. PLoS Comput. Biol 2005;1:e22.
32. Bergman CM, Quesneville H, Anxolabehere D, Ashburner M. Genome Biol 2006;7:R112. [PubMed: 17134480]
33. Slawson EE, et al. Genome Biol 2006;7:R15. [PubMed: 16507169]
34. FlyBase. www.flybase.net/
35. GenBank. www.ncbi.nlm.nih.gov/
36. Lewis SE, et al. Genome Biol 2002;3:RESEARCH0082. [PubMed: 12537571]

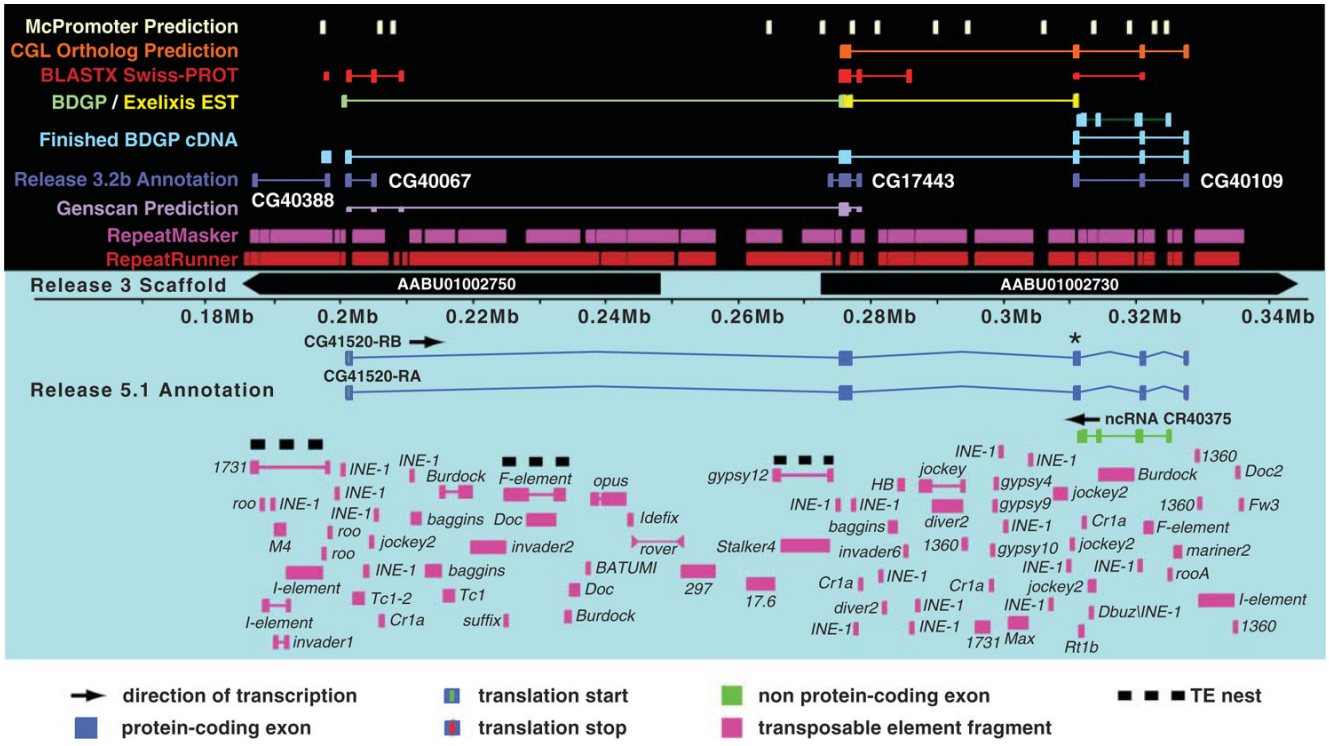

Fig. 1.

Computational pipeline results used for the Release 5.1 annotation. An *Apollo* (36) screenshot of the evidence (black region) and Release 5.1 annotations (light blue region) for scaffold CP000218, which was produced by merging, extending, and finishing the Release 3 WGS scaffolds (AABU01002750 and AABU010027230) (9). New cDNA evidence was used to merge the Release 3.2b annotations CG40067, CG17443, and CG40109 into one Release 5.1 gene (CG41520) and to identify an alternative exon for CG41250-RB (asterisk). ncRNA CG40375 is shown on the opposite strand to illustrate that it is nested within CG41520. CG40388 represents a Release 3.2b gene that is now annotated as a TE fragment. Complete annotation and evidence are shown in (11).

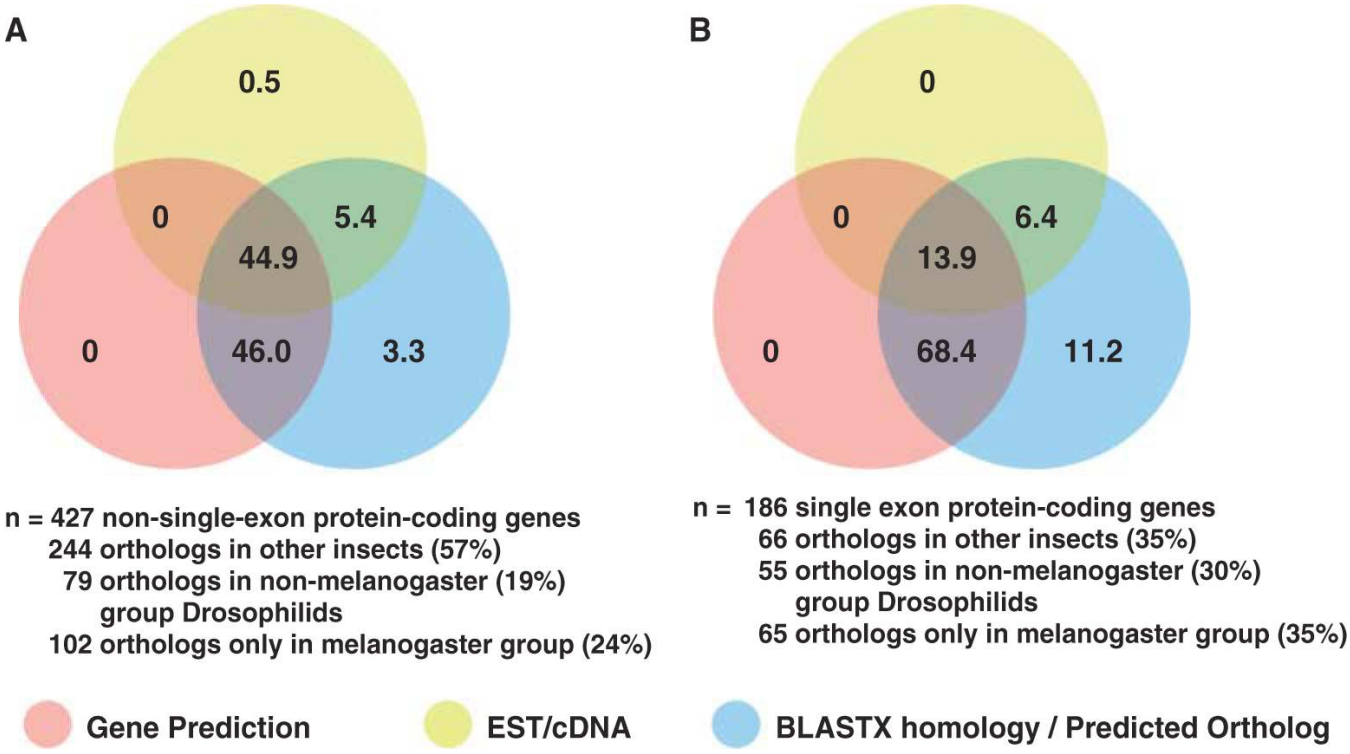

Fig. 2.

Evidence of *D. melanogaster* heterochromatin protein-coding gene annotations. Venn diagrams show the percentage of protein-coding genes supported by gene prediction (pink), EST or cDNA (yellow), and/or BLASTX/TBLASTN comparative genomic evidence (blue). (**A**) Multi-exon genes are likely to be complete, whereas (**B**) single-exon genes are likely to represent genes that are fragmented across multiple scaffolds. The number of genes measured for each class are indicated, as well as the number and percent of genes with putative orthologs in melanogaster group species, nonmelanogaster group Drosophilids, or other insect species.

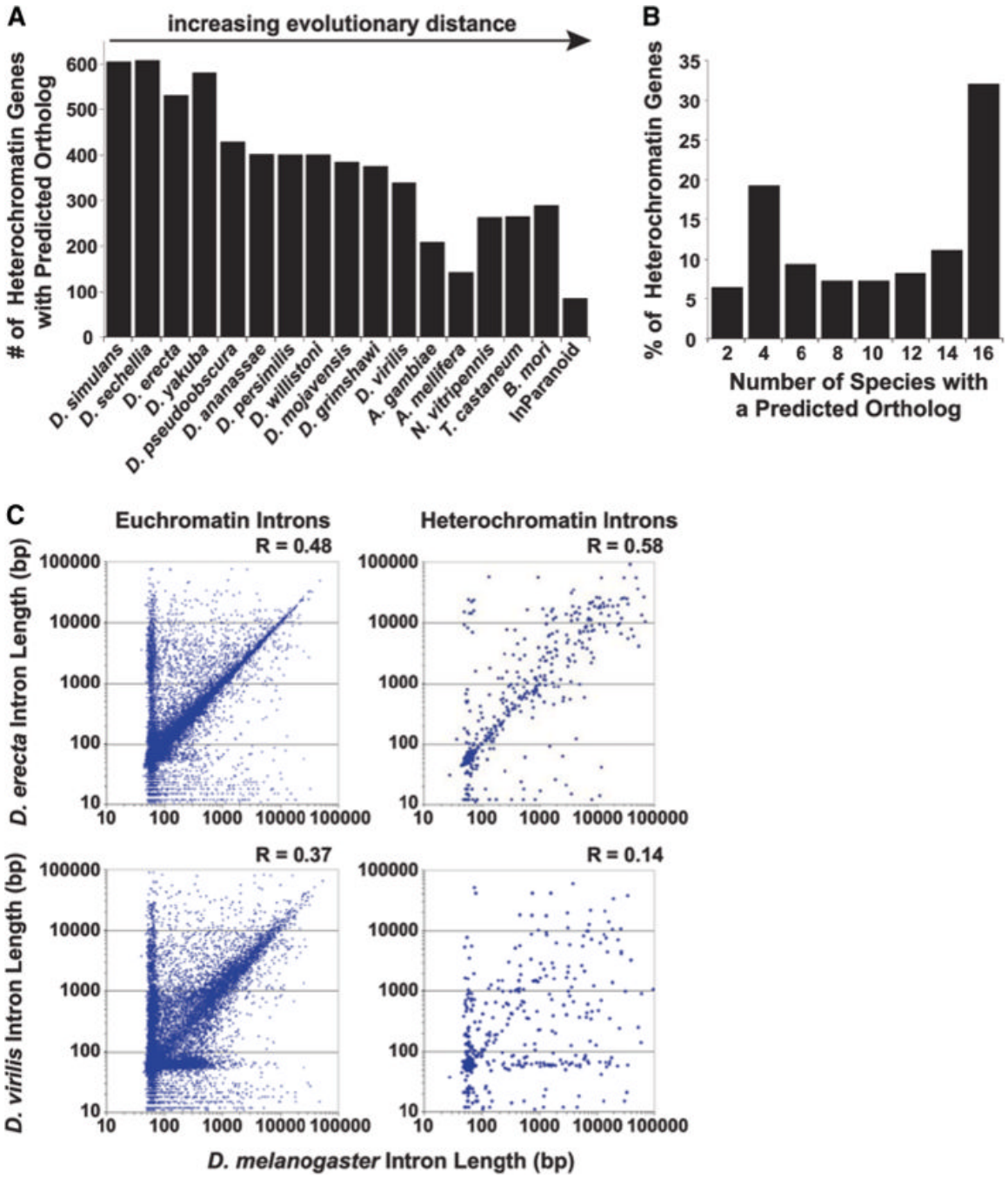

Fig. 3.

Comparative analysis summary for *D. melanogaster* heterochromatin genes. (**A**) Number of heterochromatin protein-coding genes with a predicted ortholog in a given species, ordered (left to right) by increasing evolutionary distance from *D. melanogaster*.(**B**) Frequency histograms showing the percentage of heterochromatin protein-coding genes with a predicted ortholog in the 16 insect species tested. (**C**)Scatter plots of intron lengths(bp) for euchromatin and heterochromatin protein-coding gene introns conserved in either *D. erecta* or *D. virilis*. Each data point refers to a single conserved intron. Correlation coefficients (*r*) for intron lengths are indicated.

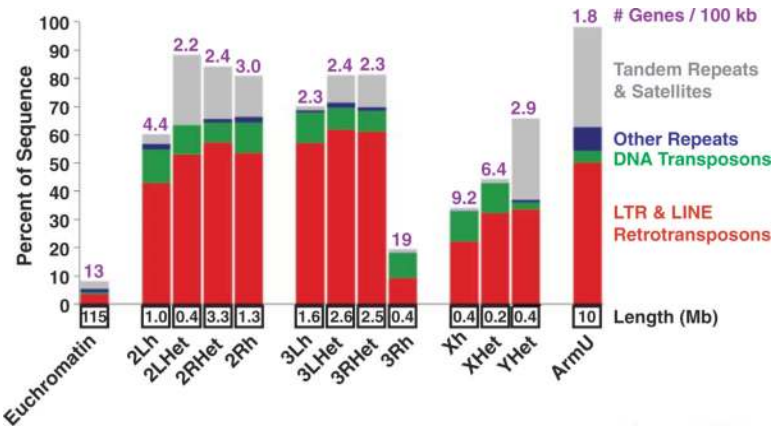

Fig. 4.

Density of repeat and gene features across the heterochromatin. The average percentages of indicated annotation types are shown for each chromosome region (total length in boxes below the *x* axis). Euchromatin is an average of the noncentric regions of arms 2, 3, and X only. 2Lh, 2Rh, 3Lh, 3Rh, and Xh describe heterochromatic regions that are contiguous with the chromosome arms, whereas the Het regions are mapped to arms and ordered, but not necessarily in the correct orientation (8,9). The average percentages of sequences for LTR- and LINE-like retrotransposons (red), DNA transposons (green), other and unknown repeats (blue), and TRF tandem repeats and satellite sequences (gray) are indicated. The average number of genes per 100 kb is shown above each histogram.

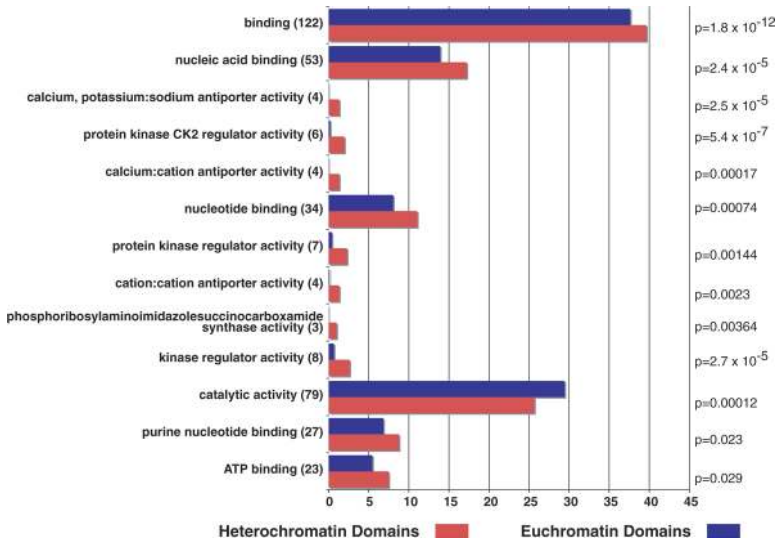

Fig. 5.

Overrepresented GO terms in heterochromatin versus euchromatin genes. The percentages of GO molecular-function domains for genes in heterochomatin (red) and euchromatin (blue) are shown. Numbers in parentheses indicate the actual number of domains in heterochromatin. *P*-value significance scores are shown to the right. Complete GO analyses are presented in the SOM text.

Table 1

Annotation summary. ND, not done.

Data type	Release 2 (2000)	Release 3.1 (2002)	Release 3.2b (2004)	Release 5.1 (2006)
Annotated sequence (Mb)	3.8	12.1	14.2	24
Sequence length of repeats (Mb/%)	ND	6.3/52	6.3/75	18/77
Sequence length of exons (Mb/%)	0.15/4	0.33/2.7	0.43/3.0	1.33/5.5
Repeat nest fragments (number/Mb)	ND	ND	ND	10084/10
Full-length TEs	ND	ND	ND	202
Total annotations	130	447	556	11038
Protein-coding genes	130	297	472	613
Single-exon genes	43	58	195	187
Genes with finished cDNAs	48	58	92	137
Protein-coding genes with any EST/cDNA clone evidence	ND	80	142	250
Pseudogenes	0	1	7	32
ncRNAs	0	3	14	13
Recursive splice sites	ND	ND	ND	16
Miscellaneous annotations	ND	ND	ND	9
Unassembled ribsomal DNA fragments	0	6	52	67