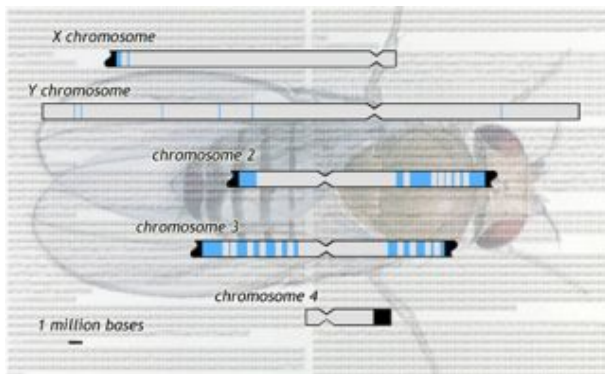


# Exploring the Dark Matter of the Genome



A diagram of *Drosophila*'s centromeric heterochromatin, which extends toward the center of the chromosomes from the gene-rich euchromatin regions (black). Sequenced regions of the heterochromatin are shown in blue. The gray regions are unsequenced "seas" of short repeats. Credit: Berkeley Lab

**Not so long ago, the difficult-to-sequence, highly repetitive, gene-poor DNA found in regions of chromosomes known as heterochromatin was called "junk." Like dark matter in the universe, the true nature of heterochromatin was unknown.**

Now members of the *Drosophila* Heterochromatin Genome Project (DHGP), headed by Gary Karpen of the Department of Energy's Lawrence Berkeley National Laboratory, are approaching a complete assembly, mapping, and functional analysis of those portions (other than simple repeats) of the heterochromatic DNA of *Drosophila melanogaster*, the fruit fly. The results confirm that heterochromatin is far from mere junk.

"Most researchers thought heterochromatin had little or no function, because it appeared to lack the protein-coding genes that occur so richly in the chromosomes' more accessible and better-studied euchromatin," says Karpen, a senior scientist in Berkeley Lab's Life Sciences Division and an adjunct professor of cell and molecular biology at the University of California at Berkeley. "In recent years it has become apparent that heterochromatin is critical for many essential functions."

Advances in sequencing the *Drosophila* heterochromatin have overcome previous technical limitations, extended understanding of heterochromatin's organization and constitution, and led to new insight into how it helps cells and organisms survive. The latest results from the DHGP are reported in a pair of papers in the June 15, 2007 issue of *Science*.

The annotated heterochromatic sequences reveal over 200 protein-coding genes. The heterochromatin also includes other features of biological importance, including sequences that code for non-protein-coding RNAs and other functional elements, such as small RNAs that neutralize transposable elements, or transposons — DNAs similar to viruses that hop around the genome and are capable of disrupting gene function.

"*Drosophila* is ideal for studying genomics for many reasons, and especially for studying heterochromatin," says Susan Celniker, a scientist in Berkeley Lab's Life Sciences Division and a longtime member of the Berkeley *Drosophila* Genome Project (BDGP). "Over a third of the total DNA in a fruit fly is heterochromatin. The female fly has an estimated 60 megabases of heterochromatin" — a megabase (Mb) is a million bases, the nucleotide building blocks of DNA — "and the male has an estimated 100 Mb, because the Y chromosome, estimated at 40 Mb, is all heterochromatin."

Heterochromatin is concentrated in the chromosome's centromeres and telomeres. In a typically bow-tie-shaped chromosome, the centromere is the knot. Centromeres play a crucial role in controlling chromosome duplication during cell division. Telomeres are a chromosome's end-caps; they help prevent the accumulation of genomic damage.

Both heterochromatin and euchromatin are sequenced using a method called whole-genome shotgun sequencing. Celniker says, "We grind up whole flies and produce libraries of DNA fragments of two sizes, some that are 2 kilobases long" — a kilobase (Kb) is 1,000 bases — "and some that are 10 kilobases long. Contiguous lengths of sequence are assembled by matching overlaps of these fragments. With euchromatin, large, single-copy fragments practically assemble themselves, but it's harder to know how shorter pieces with many repeating sequences, typical of heterochromatin, fit together."

Thus the first "substantially complete" genome sequence of *Drosophila*, published in *Science* in March, 2000 by the Berkeley *Drosophila* Genome Project and Celera Genomics, was actually far from complete. It left out a third or more of the genome by covering only the fly's euchromatin and almost none of its heterochromatin.

Says Celniker, "First we devoted our efforts towards finishing the euchromatin, leaving the part of the sequence rich in repeats, the centromeric and telomeric regions, until later. The present work extends the sequence into those regions."

Repeating sequences are the hallmark of heterochromatin, and there are several distinct kinds. Simple, short repeats are called satellite DNAs, which tend to become more abundant near the centromeres, adding up to hundreds of thousands or even millions of bases in length. In these "seas" of satellite DNA there are "islands" of moderate-length repeats totaling only tens or hundreds of kilobases, made up of transposons or fragments of transposons.

In other regions of heterochromatin, the transposons constitute the sea. Here the islands are single-copy genes, or lengths of DNA that code for RNAs other than the messenger RNA needed to make proteins, and other functional elements.

Moderately repeating fragments like transposons and single-copy genes were assembled by comparing numerous copies with unique or sufficiently distinctive sequences. The assembly was checked by matching it to clones of longer sequences. With the painstaking manual assembly taken as far as practical, the researchers mapped the sequences to their physical locations on the chromosomes. The sequence and maps prepared the ground for the next stage in the process, the functional analysis of the fly's heterochromatin.

"Historically it was called junk. We set out to see if there was any information in that junk," says Chris Smith, formerly in Berkeley Lab's Life Sciences Division and now an assistant professor of bioinformatics at San Francisco State University. "We used a pipeline of computer programs to analyze the raw sequence data, in search of genes. We identified patterns of codons that might indicate a gene splice-site or a promoter, for example. We mapped experimentally derived evidence of messenger RNAs back to the matching heterochromatin sequence. And we looked for sequences similar to ones already known from protein databases." These standard approaches to finding genes, Smith says, are relatively easy to perform in euchromatin but harder in heterochromatin, "because it is so rich in repeats."

Smith and his colleagues found evidence for 230 to 254 protein-coding genes in the heterochromatin, previously thought to contain a mere 30 to 40 (the fly's total is 14,000 or more). Many of these genes were organized quite differently from genes in euchromatin, with much longer gaps (introns) between the coding sections of the gene (exons); unlike the introns in euchromatic genes, these long gaps consisted almost entirely of repeating sequences derived from disabled transposons. The evidence suggests that heterochromatic genes are regulated differently from euchromatic ones.

Besides protein-coding genes, the annotators found other significant elements in the heterochromatin, including 13 single-copy genes that do not code for proteins but for small RNA structures called noncoding RNAs.

They also found pseudogenes — truncated copies of genes which have become inactive, most likely because they have been duplicated or code for traits no longer needed by the organism. *Drosophila* has very few pseudogenes compared to most complex organisms (humans have some 20,000), but the annotators found 32 new ones in the heterochromatin, more than twice as many as previously described.

Finally, the annotators calculated the kind and distribution of repeating elements. "The heterochromatin is incredibly repeat-rich," says Smith, "and most of it is transposable elements that have been chewed up and fragmented."

Susan Celniker emphasizes that transposable elements, of which 96 families have been found in *Drosophila* so far, "can be very dangerous. They are viruses — literally." She points out that during subsequent sequencing efforts, DNA was made three times from the same strain of flies — in 1990, 1998, and 1999 — and that new patterns of transposons were found in each. "We can see intact transposons moving from one position to the next."

Says Smith, "The heterochromatin is where transposons may be actively regulated — they go in, get stuck, and get chopped up. Other researchers have shown that small interfering RNAs are made in the heterochromatin, which seek out transposons and inactivate them. That's the kind of knowledge that comes from sequencing heterochromatin."

Smith uses the metaphor of dark matter to suggest the significance of heterochromatin. "We don't know what holds the galaxies together, and the same is true of the genome. We're pretty good at understanding how individual genes work, but we don't understand, for example, how the large-scale structure of genomes affects cellular processes. We hear too much about the 'post-genomic era' — it's underappreciated that we don't understand the genome yet."

Celniker agrees. "What we would like to see is a complete sequence of the fruit fly from telomere to telomere" — that is, from one end of the chromosomes to the other — including the centromere in the middle." Obtaining a more complete picture of the sequence of heterochromatin is clearly a crucial step toward a better understanding of genome functions.

Karpen notes that one intensively studied and puzzling problem in chromosome biology will benefit enormously from the new findings: epigenetics — the inheritance of traits and genetic information as controlled by proteins associated with the chromosomes, rather than by DNA sequence. Starting in the 1920s, epigenetics was discovered through biological studies of heterochromatin, and is now known to regulate essential functions such as those of the centromeres. The heterochromatin sequence of *Drosophila* will provide an essential foundation for identifying the proteins and components involved in epigenetic inheritance, as well as other mysteries surrounding the genome's "dark matter."

Citations: "Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin," by Roger A. Hoskins, Joseph W. Carlson, Cameron Kennedy, David Acevedo, Martha Evans-Holm, Erwin Frise, Kenneth H. Wan, Soo Park, Maria Mendez-Lago, Fabrizio Rossi, Alfredo Villasante, Patrizio Dimitri, Gary H. Karpen, and Susan Celniker appears in the June 15, 2007 issue of *Science*.

"The release 5.1 annotation of *Drosophila melanogaster* heterochromatin," by Christopher D. Smith, ShenQiang Shu, Christopher J. Mungall, and Gary H. Karpen appears in the June 15, 2007 issue of *Science*.

Source: Berkeley Lab

*This document is subject to copyright. Apart from any fair dealing for the purpose of private study, research, no part*

*may be reproduced without the written permission. The content is provided for information purposes only.*