Genetic and Bioinformatic Analysis of 41C and the 2R Heterochromatin of *Drosophila melanogaster*: A Window on the Heterochromatin-Euchromatin Junction

Steven H. Myster,* Fei Wang,[†,1] Robert Cavallo,[‡,1] Whitney Christian,[†,1] Seema Bhotika,[†] Charles T. Anderson[†] and Mark Peifer*[,†,‡,2]

[†]*Department of Biology,* [‡]*Curriculum in Genetics and Molecular Biology and *Lineberger Comprehensive Cancer Center,*
University of North Carolina, Chapel Hill, North Carolina 27599-3280

Manuscript received September 2, 2003
Accepted for publication November 12, 2003

ABSTRACT

Genomic sequences provide powerful new tools in genetic analysis, making it possible to combine classical genetics with genomics to characterize the genes in a particular chromosome region. These approaches have been applied successfully to the euchromatin, but analysis of the heterochromatin has lagged somewhat behind. We describe a combined genetic and bioinformatics approach to the base of the right arm of the *Drosophila melanogaster* second chromosome, at the boundary between pericentric heterochromatin and euchromatin. We used resources provided by the genome project to derive a physical map of the region, examine gene density, and estimate the number of potential genes. We also carried out a large-scale genetic screen for lethal mutations in the region. We identified new alleles of the known essential genes and also identified mutations in 21 novel loci. Fourteen complementation groups map proximal to the assembled sequence. We used PCR to map the endpoints of several deficiencies and used the same set of deficiencies to order the essential genes, correlating the genetic and physical map. This allowed us to assign two of the complementation groups to particular "computed/curated genes" (CGs), one of which is *Nipped-A*, which our evidence suggests encodes Drosophila Tra1/TRRAP.

EUKARYOTIC chromosomes are organized into domains termed euchromatin and heterochromatin (reviewed in Henikoff 2000; Grewal and Elgin 2002). Euchromatin is composed primarily of single-copy DNA and is condensed during mitosis and decondensed during interphase. In contrast, heterochromatin is largely composed of repetitive DNA that remains condensed during interphase, is replicated late in S-phase, and is relatively gene poor. Heterochromatin is concentrated near telomeres and in the pericentric region spanning the centromere. However, despite or perhaps because of its repetitive nature, heterochromatin has several important functions. It contains the centromere in most eukaryotes and plays important roles in meiotic pairing and sister chromatid cohesion (reviewed in Henikoff 2000; Sullivan *et al.* 2001). Advances in whole-genome sequencing have provided great insights into the composition and organization of genes in euchromatin and also have provided geneticists with tools to extend genetic analysis to a new level by comprehensively characterizing a region using a combination of classical genetic, reverse genetic, and bioinformatics tools (*e.g.*, Ashburner *et al.* 1999). Our understanding of the composition, organization, and regulation of the heterochromatin has lagged behind, but recent sequencing efforts and functional studies have begun to shed new light on the structure and function of the heterochromatin.

One interesting property of heterochromatin is that it can silence euchromatic genes that are placed within it by chromosomal rearrangements such as translocations or transposable element insertions (reviewed in Grewal and Elgin 2002). This silencing property is epigenetic and is clonally inherited at a cellular level, resulting in variegated expression—a phenomenon termed position-effect variegation. This type of silencing occurs in organisms as diverse as yeasts, Drosophila, and mammals. Although heterochromatin in general is highly repetitive, many single-copy genes, which must have unique mechanisms of escaping gene silencing, are located there.

Our best understanding of the centromere and of the mechanisms of heterochromatic silencing comes from budding yeast (reviewed in Moazed 2001; Cleveland *et al.* 2003). However, its genome differs from that of multicellular eukaryotes and even from that of some other fungi in the small size of its centromere and the relatively low levels of repetitive and heterochromatic DNA. Drosophila is an excellent model for studying heterochromatin in an animal. It provided the first examples of position-effect variegation (Müller 1930)

[1]These authors contributed equally to this work.

[2]*Corresponding author:* Department of Biology, CB#3280, Coker Hall, University of North Carolina, Chapel Hill, NC 27599-3280. E-mail: peifer@unc.edu

and is where the genetic basis of this phenomenon is best understood (reviewed in Wallrath 1998). In addition, genetic experiments defined a minimal centromeric region and revealed some of the *cis* sites and transacting factors necessary for its segregation (reviewed in Sullivan *et al.* 2001).

Further, the Drosophila genome is well characterized. In Release 1 of the genome, most of the 120 Mb of the euchromatic genome were represented as complete and contiguous sequence (Adams *et al.* 2000). The heterochromatin was less completely assembled. However, the recently released new whole-genome shotgun sequence assembly (WGS3) greatly increased assembly of the pericentric heterochromatin (Celniker *et al.* 2002; Hoskins *et al.* 2002). Release 3 of the genome also provided improved gene annotation (Misra *et al.* 2002) and a more comprehensive look at transposon content (Kaminker *et al.* 2002), both of which are relevant to the heterochromatin. In addition, Gary Karpen's group defined a minimal functional centromere using genetic techniques and characterized it by molecular mapping and partial sequencing (Sun *et al.* 1997, 2003).

Together, these analyses reveal that heterochromatin is not a single entity. The ∼420-kb functional centromere is composed of large blocks of simple repeat satellite DNA (350 kb) interspersed with more complex sequence composed of transposons (Sun *et al.* 2003). In contrast, the sequence at the euchromatin-heterochromatin junction is largely composed of transposable elements (at least ∼50% of a characterized contig in the 2L heterochromatin; Hoskins *et al.* 2002), with single-copy genes interspersed at a density much lower than that found in the standard euchromatin (one gene per 50 kb, approximately sixfold lower than that in the euchromatin; Hoskins *et al.* 2002). The accumulation of transposons in the heterochromatin is an interesting and conserved phenomenon (reviewed in Dimitri and Junakovic 1999) that may reflect the low meiotic recombination rate in the region or may suggest functional roles for transposons in the structure or function of the heterochromatin.

Classical genetics has also been used to study the heterochromatin. For example, the pericentric heterochromatin of the right arm of the second chromosome (2R) of Drosophila was the target of several genetic screens that identified a number of essential loci (Hilliker 1976; Dimitri 1991; Dimitri *et al.* 1997; Rollins *et al.* 1999). However, the number of loci identified genetically does not approach the number of predicted genes in the region (Hoskins *et al.* 2002), suggesting that the screens did not reach saturation and/or that many predicted genes are either not genes at all or not essential.

One way to link genetic loci and those defined by sequence is via transposon mutagenesis. Transposons provide a molecular tag that allows one to relatively easily determine which gene is disrupted by a given mutation. The most common transposon used for this purpose in Drosophila is the *P* element. In a concerted effort, *P* elements that disrupt ∼25% of all essential loci in Drosophila were collected (Spradling *et al.* 1999). However, few were inserted in heterochromatin. Two reasons for this seem plausible (and are not mutually exclusive): *P* elements might transpose into heterochromatin at reduced frequency, or heterochromatic insertions might not be recognized due to the silencing of the selectable markers used to follow them. Dimitri *et al.* (1997) successfully used the LINE-like *I* factor to mutagenize the heterochromatin, suggesting that it can transpose into this region.

Two strategies were developed to allow recovery of heterochromatic *P*-element insertions. Roseman *et al.* (1995) generated a *P* element, the *SUPorP*, in which they flanked the $white^+$ marker by insulator elements to protect it from silencing. When this *P* element was used, heterochromatic insertions were obtained, suggesting that silencing was a major reason for the previous difficulties. Yan *et al.* (2002) utilized a *P* element carrying the $yellow^+$ selectable marker and have screened for insertions with variegated expression, allowing them to efficiently collect insertions in the centric heterochromatin. These opened up a powerful new approach for genetic analysis in the heterochromatin, and these *P* elements are now being used in ongoing systematic efforts to generate *P*-element insertions in additional genes and regions (Yan *et al.* 2002; H. Bellen, R. Hoskins, R. Levis, G. Luo, G. M. Rubin and A. C. Spradling, unpublished data; http://flypush.imgen.bcm.tmc.edu/pscreen/).

We describe below a genetic and bioinformatic analysis of the 2R euchromatin-heterochromatin junction. We built on earlier genetic work in the region, carrying out a large-scale genetic screen for essential genes, and used the genetic and bioinformatics tools developed by the Drosophila genome project to connect the genetic and physical maps, providing an example of how genetics and bioinformatics can be integrated to analyze the Drosophila heterochromatin.

MATERIALS AND METHODS

Bioinformatics: Analysis was done using tools and databases of the Berkeley *Drosophila* Genome Project (BDGP; www.fruitfly.org) and the National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov). For analysis of the "computed/curated genes" (CGs) in the region, each predicted protein was used as a query in a protein BLAST search of the nonredundant protein database at NCBI. If no significant match was found, then the predicted coding sequence was translated in all frames and used to query the same database, using a translated BLAST search. Some transposon matches were found via the latter search type. Figure 2 was created using scaffold maps of the BDGP Armview viewer (www.fruitfly.org/cgi-bin/annot/arm_view.pl), as well as complete sequences of each scaffold as annotated in GenBank. For analysis of the repetitive DNA in the vicinity of the *p120* gene, we began with the 27 kb of sequence beginning ∼3 kb upstream of the *p120* start site (this limit of the region analyzed was imposed by the presence of an unsequenced region of scaffold AE002751 beginning there) and extending through the next downstream gene, *CG17486*. One- to 2-kb segments across this region were used as queries of BLAST searches of the repeats and transposons database, using FlyBlast. We also searched the full Drosophila genome to look for repetitive DNA that is not included in the repeats database and searched the predicted genes and expressed sequence tag (EST) databases for matches to potential coding sequences.

Constructing a physical map of the region: To construct a physical map of the region, we began by assuming that the scaffolds containing *p120* and *Nipped-B*, AE002751 and AE003040, must map to the region on the basis of their genetic or physical map positions (Rollins *et al.* 1999; Myster *et al.* 2003). We then attempted to order these with respect to the scaffolds from AE30788 to the right that were part of the assembled release 1 genome and to identify additional unassigned scaffolds that might map to this region. We began with the sequence-tagged site (STS) content map of bacterial artificial chromosomes (BACs) generated by BDGP (http://www.fruitfly.org/seq_tools/displays/ArmView.html). STSs in the region were used to BLAST search the entire fly genome, using FlyBlast (http://www.fruitfly.org/blast/) to search for matches to scaffolds, BACs, or predicted genes. This identified several candidate scaffolds that mapped to the region and allowed us to tentatively order them. We then used selected regions of these scaffolds (in particular, genes that mapped onto them) as BLAST queries, confirming and extending our hypothetical map. We also used BAC end sequences that were not in the original STS content map as BLAST queries. These allowed us to make a proposed tiling path of BACs across the region. Finally, we used the sequence of BACR11B22, which was complete, to more accurately order and orient the scaffolds at the right end of our map and to identify one additional scaffold that mapped to this region (AE003064).

Fly stocks: Canton-S, *cn bw*, *vlc*^{07022}, *Bub1*^{k03113}, *rl*, *Df(2R)M41A8*, *Df(2R)M41A10*, *Df(2R)nap1*, *M(2)41A*^{2}, and *w**; *wg*$^{Sp\text{-}1}$/*CyO*; *ry*^{506} *Sb*^{1} *P{ry*$^{+t7.2}$*=Delta2-3}99B/TM6B*, *Tb*$^{+}$ were provided by the Bloomington Stock Center and mutations are described in FlyBase (flybase.bio.indiana.edu/). The *Cy Kr GFP* line is described in Casso *et al.* (2000). *l(2)41Ae*$^{34\text{-}14}$, *l(2)41Af*$^{45\text{-}72}$ (Hilliker 1976), *IR3*, *IR23* (Dimitri *et al.* 1997), and all *Nipped* alleles (Rollins *et al.* 1999) were used in complementation tests. All tests were performed at 25°.

Ethyl methanesulfonate mutagenesis: The 25 mm ethyl methanesulfonate (EMS) was fed to flies in 1% sucrose according to standard procedures (Grigliatti 1998). In seven independent rounds of mutagenesis >6000 *cn bw* males were mutagenized and crossed to *Df(2R)M41A8 al /SM1* to capture individual mutagenized and balanced second chromosomes. These males were crossed to a *p120* deficiency line that had the recessive marker *al* recombined onto the deficiency chromosome [*Df(2R)M41A8 al /SM1*]. *al* is also carried on *SM1*. This allowed us to distinguish the deficiency used in the screen from mutations generated by mutagenesis. Crosses were scored for the presence of unbalanced flies. From crosses that contained only balanced progeny, indicating the presence of a new mutation lethal over the deficiency chromosome, balanced males and females carrying the mutated chromosome were identified by the presence of aristae and stocks were established. A total of 6284 individual mutagenized males were crossed to the deficiency line and 226 lines that are lethal over the deficiency were established. Twenty-four lines complemented the deficiency on the retest and were discarded, 29 died before genetic analysis was completed, and 45 lines were unhealthy and could not be maintained. A total of 128 lines were placed on the genetic map.

***P*-element mobilization mutagenesis:** We desired to make small deletions and recover local transpositions in the *p120* region. We began with a *P* element inserted between *p120* and the neighboring gene, the *SUPorP* strain, *KG01086* (H. Bellen, R. Hoskins, R. Levis, G. Luo, G. M. Rubin and A. C. Spradling, unpublished data; http://flypush.imgen.bcm.tmc.edu/pscreen/), and backcrossed it to *y w; Pin/CyO* three times to segregate the insertion at 41C away from additional *P*-element insertions on other chromosomes, selecting by eye color for the loss of additional insertions. The retention of the *KG01086* element was confirmed by PCR amplification of the insertion junction using a *P*-element-specific primer (*P*-out, 5′-ccgcggccgcggaccaccttatgttatttc-3′) and a primer located ~7.7 kb downstream of *p120* (5′-ccgtctttaagcacgagtacacag-3′). To mobilize the element, *KG01086* was crossed to a strain carrying a source of transposase (*w; Sp/CyO; Sb*Δ*2-3/TM6 Tb*) and single males carrying both *KG01086* and the transposase were crossed to *y w; Pin/CyO*. Progeny carrying *KG01086* but not the transposase were scored for changes in eye and body color due to mobilization or deletion of the element and backcrossed to establish stable lines. Each line was crossed to the *p120* deficiency line *M(2)41A*^{2}*/SM1*, and progeny were scored for viability. DNA was isolated from heterozygotes containing both the deficiency chromosome and the mobilized *KG01086* chromosome for PCR analysis. Initial tests used three primer pairs: one spanned the *KG01086* insertion (*p120* side forward primer 5′-ccgtctttaagcacgagtacacag-3′and *CG17486* side reverse primer 5′-agcagacaactgcatgtgtgcac-3′), and the second and third pair involved use of a *P*-element primer to the inverted terminal repeats (*P*-out, see above) paired with each of the genomic primers flanking the insertion. All lines missing one or both junction fragments in the initial assay were analyzed further with primer pairs in the *p120* and *CG17486* coding regions to assess if the deletions extended into these genes. Six hundred crosses were screened for mobilization. A total of 401 independent lines were established and assayed by PCR. Mobilization events fell into the following classes: 287 lines lost both the *yellow* and *white* markers (y^{-} w^{-}), 48 lines were y^{+} w^{-}, 5 lines were y^{-} w^{+}, 2 lines were w^{+} *y* variegated, 5 lines were w^{-} *y* variegated, 18 lines had lighter eye color, and 36 lines had darker eye color. DNA was isolated from one to two flies using a scaled-down version of the BDGP protocol (http://www.fruitfly.org/about/methods/inverse.pcr.html). PCR conditions were: 3 min at 95°, followed by 35 cycles of 95° for 30 sec, 60° for 1 min, and 72° for 1 min.

Deficiency endpoint mapping and mutation identification: Deficiency lines were rebalanced over *CyO KrGFP* (Casso *et al.* 2000) and homozygous deficiency [non-green fluorescent protein (GFP)] embryos were picked for DNA isolation. DNA was isolated as in Gloor *et al.* (1993) and PCR reactions were performed as described above. Primer pairs for genes are in listed in supplemental Table 1 at www.genetics.org/supplemental/. The predicted coding regions and intron-exon boundaries of *CG2905* were sequenced from *Nipped-A* alleles *l(2)NC116*, *l(2)NC186* (both from this study), and *Nipped-A*$^{357.2}$ (Rollins *et al.* 1999). Genomic DNA isolation and PCR amplification were performed as described above, using balanced flies as starting material. PCR products were separated on agarose gels, extracted, and directly sequenced using the ABI PRISM BigDye Terminator cycle sequencing ready reaction kit with AmpliTaq DNA polymerase on a 3100 genetic analyzer (Applied Biosystems, Foster City, CA). Amplification and sequencing primer information are available upon request.

RESULTS

Rationale: Our interest in the genetics and molecular genetics of the heterochromatin-euchromatin junction of 2R was initiated by the fact that *p120*, a gene of interest to our lab, maps to this region. We thus began parallel genetic and molecular genetic analysis of this region, building on earlier genetic work in the region, and utilizing the genetic and bioinformatics tools developed by the Drosophila genome project. We carried out a screen for essential genes that map to this region and then used genetic and molecular methods to connect the genetic and physical maps. Our goal was to integrate genetics and bioinformatics and thus obtain new insights into the Drosophila heterochromatin.

The heterochromatin defined by high-resolution banding of mitotic chromosomes differs somewhat from the heterochromatin as defined on polytene chromosomes, where unamplified sequences form the chromocenter. This was clarified by parallel fluorescent *in situ* hybridization (FISH) analysis of mitotic and polytene chromosomes, using BACs from the 2R heterochromatin as probes (Corradini *et al.* 2003). These data suggest that the BACs that together span the heterochromatic region h46 on mitotic chromosomes hybridize to 41C–E as defined on the polytene chromosomes. We refer to the entire region as 41C for simplicity.

Bioinformatic analysis of the 2R euchromatin/heterochromatin junction: Previous work defined a number of lethal complementation groups in the 2R heterochromatin (Hilliker 1976; Dimitri *et al.* 1997; Rollins *et al.* 1999) and mapped these relative to several deficiencies. In addition, the BDGP in collaboration with Celera Genomics generated both a physical map (Hoskins *et al.* 2000) and sequence information (Adams *et al.* 2000; Celniker *et al.* 2002; Hoskins *et al.* 2002; http://www.fruitfly.org/). While most of 2R was assembled into a continuous sequence in the initial whole-genome assembly (Adams *et al.* 2000), the 41C region was not (the Release 3 sequence does fully assemble the region; see below). In that initial analysis, *p120* mapped to the most proximal of the scaffolds assembled (AE002751), with *p120* the most proximal sequenced gene then defined. One additional scaffold was assigned to 41C (AE002760), which was thought to lie between *p120* and the rest of 2R (which began with scaffold AE003788). The BDGP also assigned numerous BAC clones to the region (Hoskins *et al.* 2000) and mapped numerous STSs, most derived from BAC end sequences.

We used this information as a starting point to attempt to derive a physical map of the region (Figure 1). One additional scaffold, AE003040, which was at that point unassigned to a chromosome, clearly belonged in this region, as it carries *Nipped-B*, which genetically maps to 41C (Rollins *et al.* 1999). We then carried out BLAST searches of the Release 2 sequence scaffolds with STSs mapped to the region by the BDGP (using the FlyBlast server; http://www.fruitfly.org/blast/). This identified several additional scaffolds as candidates that might map to the region and allowed us to tentatively order them with respect to the scaffolds known to map to this region. We then carried out BLAST searches of partially sequenced BDGP BAC clones with both STSs and genes from these scaffolds, to further test our map. These data allowed us to derive a proposed physical map of the region (Figure 1).

In 2002 the BDGP/Celera Genomics genome project released an improved whole-genome shotgun assembly (WGS3; Celniker *et al.* 2002; Hoskins *et al.* 2002). This includes an assembled sequence of much of the 2R heterochromatin, including the entire region we analyzed: the Release 3 scaffold AE003788 is a high-quality finished sequence that encompasses the Release 2 scaffolds AE003024, AE003064, AE003056, and AE003788; while the WGS3 2R wgs3 centromere extension scaffold encompasses Release 2 scaffolds AE002751, AE003040, AE003032, and AE002760. Our proposed physical map is fully consistent with the Release 3 sequence assembly, testifying to its quality. Our proposed physical map also agrees with the mapping of BACs by FISH on both mitotic and polytene chromosomes (Corradini *et al.* 2003).

We next examined each CG assigned by the BDGP/Celera genome project (Rubin *et al.* 2000; Hoskins *et al.* 2002; Misra *et al.* 2002) to the scaffolds in the region. We performed BLAST searches of the NCBI nonredundant protein database to determine whether each CG was conserved in other organisms and whether any of its relatives had a known or predicted function. A small number appear to be transposons or transposon remnants (Table 1). Most of the remaining CGs have strong support as *bona fide* genes, as they have clear orthologs or sequence relatives in other species (Table 2; Figure 2), many with known or inferred functions.

Previous analysis demonstrated that average gene density in the heterochromatin is quite low. In the portion of the WGS3 heterochromatic sequence in scaffolds large enough to be annotated in detail, average gene density was 1 gene per 42 kb (287 genes in 12.1 Mb; Hoskins *et al.* 2002). This contrasts with the genome-wide average of 1 gene per 9 kb (Adams *et al.* 2000; Misra *et al.* 2002). In the 594-kb *light* region, which is in the 2L pericentric heterochromatin, gene density was 1 gene per 50 kb (Hoskins *et al.* 2002). To compare 41C to these, we analyzed the annotated scaffolds of the region (Misra *et al.* 2002), creating a picture of gene density across 41C (Figure 2; this extends further distal to the region covered in Figure 1). Gene density through much of the region is quite low. The region can be roughly divided into four parts on the basis of gene density. In the most proximal region (2R wgs3 centromere extension; 11 genes in ~345 kb), gene density is low—1 gene per 32 kb. Next is a region of >210 kb containing no predicted genes (all of AE003788 except its distal end). Next most distal is a long region with low gene density (1 gene per 29 kb; AE003787–AE003786; 20 genes per 585 kb). Gene density then increases fairly abruptly in the most proximal scaffold to a density similar to that of most of the euchromatic

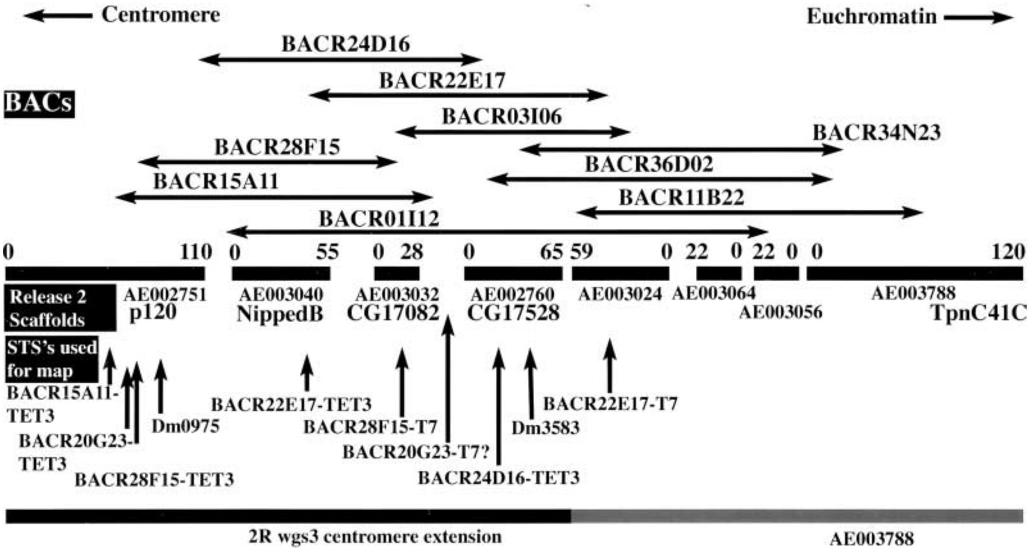

Figure 1.—Physical map of 41C. The centromere is to the left and the euchromatic region of 2R is to the right. At the center are the Release 2 scaffolds that mapped to the region when we began our analysis, which we ordered and oriented on the basis of matches to sequences identified in the overlapping BAC clones (see text for details). Gaps are indicated by spaces and scaffold size is in kilobases. The locations of selected genes on each scaffold that carries genes are indicated below the scaffold name. Above the scaffolds is an overlapping set of BAC clones covering the region. Below the scaffold are some of the STSs (Hoskins *et al.* 2000) that support the map (see text for further details). At the bottom are the WG3 sequence assemblies (Celniker *et al.* 2002; Hoskins *et al.* 2002). Our map fully supports their sequence assembly.

genome (1 gene per 7.1 kb in the first 50 kb of AE003785; 7 genes per 50 kb). A similar regional organization was previously observed in the heterochromatin-euchromatin junction of the X and 2L (Adams *et al.* 2000; Hoskins *et al.* 2002)—in each case a region devoid of genes was found intervening between regions of lowered gene density.

Earlier analyses of the X and 2L (Adams *et al.* 2000; Hoskins *et al.* 2002), along with the analysis of individual heterochromatic genes (referenced in Hoskins *et al.* 2002), suggest that the low gene density has two causes. First, many heterochromatic genes have large introns (Adams *et al.* 2000; Hoskins *et al.* 2002) relative to the genome as a whole (Mount *et al.* 1992; Adams *et al.* 2000; Misra *et al.* 2002). Second, within the regions of low gene density are stretches devoid of genes (*e.g.*, Hoskins *et al.* 2002). We observed similar features in the 41C region. Many genes on the most proximal scaffold (*e.g.*, *CG40293, p120ctn*, and *Nipped-B*) and in the more proximal region of the more distal scaffolds (*e.g.*, *d4, Ogt, CG30437*, and *CG30438)* are interrupted by large introns, a feature that is less frequent for genes on the most distal scaffold (AE003785). Second, within the regions of low gene density are several shorter stretches (40–50 kb each) devoid of genes. Our analysis thus reinforces the picture derived from the earlier analyses of the X and 2L (Adams *et al.* 2000; Hoskins *et al.* 2002), which suggested that the heterochromatin does not have a sharp boundary with the euchromatin, but rather that gene density rises and repetitive DNA content decreases gradually across several megabases.

The heterochromatin-euchromatin junctions thus far analyzed (X, 2L) are rich in repetitive DNA, as is the *rolled* region of 2R, which is deeper in the heterochromatin (*e.g.*, Miklos *et al.* 1988; Adams *et al.* 2000; Hoskins *et al.* 2002). A total of 52% of the 20.7-Mb WGS3 heterochromatic sequence is accounted for by transposable elements, and 78% of the repetitive sequence represented LTR retrotransposons (Hoskins *et al.* 2002). In contrast, only ~4% of the euchromatin is composed of transposons (Kaminker *et al.* 2002). To obtain a more detailed view of a sample of the 41C region, we analyzed 27 kb of sequence in the vicinity of *p120* (Figure 3), of which 4 kb (~15%) is composed of exons of *p120* and

TABLE 1

CGs that may be transposons or other repetitive DNA

CG40290, repetitive in the genome—7 BLAST matches with $P(n) < e - 10$.

CG40279 = CG17479, retrotransposon. BLAST match to Q9N9Z1 "ENDONUCLEASE/REVERSE TRANSCRIPTASE," $P(n) = e - 8$.

CG40280 = CG17478, possible retrotransposon. BLAST match of predicted protein to BAA95569 RNase H and integrase-like protein [*Bombyx mori*]; 53% identical over first 30 of 55 amino acids. $P(n) = 0.24$. When translated in all three frames, three additional matches to BAA95569 are found. $P(n) = 6e - 6$.

CG30442, possible retrotransposon. Repetitive in genome—8 BLAST matches with $P(n) < e - 10$. BLAST match to CAC59744 putative retrovirus-like env glycoprotein [*Drosophila virilis*]; 33% identity over 65 of 126 amino acids. $P(n) = 0.046$.

CG1294, repetitive in the genome—24 BLAST matches with $P(n) < e - 10$. Best BLAST protein match is AAB66824 Tel1, a transposable element of *D. virilis*. $P(n) = 1e - 5$.

TABLE 2

Annotation of predicted proteins in 41C

2R WGS3 centromere extension (11 genes)
CG40278 (=CG18001), 70 aa, ribosomal protein L38.
Closest human hit: NP_000990.1, ribosomal protein L38 ($P(n) = 5e - 14$).

CG40293, 333 aa, Ste-20-like protein kinase.
Has close Drosophila relatives including Frayed ($P(n) = 8e - 21$), CG5169.
Closest human hit: AAG48269.1, breast cancer antigen NY-BR-96 ($P(n) = 4e - 41$).

p120ctn (CG17484), 781 aa, p120 family, adherens junction component.
Closest human hit: AAB97957, Arm-repeat protein NPRAP ($P(n) = e - 128$).

CG17486, 564 aa, possible asparagine synthetase.
Closest human hit: NP_061921.1, hypothetical protein ($P(n) = 8e - 73$).
Closest hit with known function: NP_578800.1, asparagine synthetase [*Pyrococcus furiosus*] ($P(n) = 1e - 12$).

CG17883, 312 aa, TBC domain protein.
Closest human hit: NP_653229.1, chromosome 20 open reading frame 140 ($P(n) = 1e - 62$).

Nipped B (CG17704), 2053 aa, chromosomal adherin family member.
Closest human hit: NP_597677, IDN3 protein ($P(n) = 0.0$).

CG40282, 128 aa, and CG40287 (=CG17706), 77 aa. Both closely related to Drosophila NonA (*e.g.*, ($P(n) = 1e - 40$), but transcribed in opposite directions.
Rearranged and diverged; pseudogene?

CG17082, 629 aa, Rho GAP.
Closest human hit: NP_277050.1, MacGAP protein ($P(n) = 4e - 22$).

CG12547, 717 aa, N-terminal thioredoxin domain, C-terminal NHL repeat.
Closest human hit: XP_089702.1, hypothetical protein ($P(n) = e - 103$).

CG17528, 560 aa, doublecortin kinase-like.
Closest human hit: NP_004725.1, doublecortin and CaM kinase-like 1 ($P(n) = 2e - 66$).

CG40285 (=CG14464), 127 aa. Has human relative of unknown function.
Closest human hit: NP_689529.1, hypothetical protein ($P(n) = 9e - 14$).

AE003788 (one gene)
TpnC41C (CG2981), 154 aa, troponin C.
Closest relatives other Drosophila proteins, *e.g.*, CG7930 ($P(n) = 7e - 40$).
Closest human hit: AAH08437, calmodulin 2 ($P(n) = 8e - 25$).

AE003787 (11 genes)
CG3107, 1112 aa, metalloprotease.
Closest human hit: NP_055704.1, metalloprotease 1 (pitrilysin family) ($P(n) = 0.0$).

CG2944, 349 aa, SSB1 homolog.
Closest human hit: XP_045247.2, SPRY domain-containing SOCS box protein SSB-1 ($P(n) = e - 115$).

CG3136, 739 aa, bZIP family transcription factor.
Closest human hit: NP_004372.2, cAMP-responsive element-binding protein-like 1 ($P(n) = 3e - 13$).

CG2905, 3435 aa, Tra1/TRRAP, part of SAGA acetyltransferase/transcriptional adaptor complex.
Closest human hit: NP_003487, transformation/transcription domain-associated protein ($P(n) = 0.0$).

d4 (CG2682), 495 aa, homolog of requiem PhD finger.
Closest human hit: NP_006259.1, requiem; apoptosis response zinc-finger protein ($P(n) = 2e - 48$).

Ogt (CG10392), 1059 aa, *O*-glycosyltransferase.
Closest human hit: NP_858059, *O*-linked GlcNAc transferase isoform 2 ($P(n) = 0.0$).

CG10465, 301 aa, BTB domain protein.
Closest human hit: Q13829, TNF-α-induced protein, B12 BTB domain homolog ($P(n) = 5e - 90$).

CG10395, 281 aa, PAP-1-binding protein.
Closest human hit: NP_112578.1, PAP-1-binding protein ($P(n) = 3e - 14$).

(*continued*)

TABLE 2

(Continued)

CG30441, 126 aa intraflagellar transport protein 20.
Closest human hit: AAH02640, intraflagellar transport protein IFT20 ($P(n) = 1e - 9$).

CG10396, 162 aa, cytochrome C oxidase polypeptide IV.
Closest human hit: P13073, cytochrome C oxidase polypeptide IV ($P(n) = 5e - 23$).

CG10417, 662 aa, protein phosphatase 2C gamma.
Closest human hit: O15355, protein phosphatase 2C gamma ($P(n) = 8e - 87$).

AE003786 (8 genes)

CG30437, 733 aa, laccase.
Several others in fly genome [*e.g.*, CG7871 ($P(n) = 1e - 57$), CG5959].
Hit in another insect: CAD20461, laccase, venom protein, parasitic wasp ($P(n) = e - 117$).
Matches in fungi, plants, nematodes, but not in mammals.

CG32838, 733 aa, another laccase, 70% identical to CG30437.

CG30440, 1057 aa, Ost/trio-like rho GEF.
Closest human hit: AAA52172.1, DBL-transforming protein ($P(n) = 6e - 74$).

CG30438, 413 aa, putative UDP-glycosyltransferase.
>5 in flies (*e.g.*, CG6658, CG6644, UGT35A ($P(n) = 1e - 59$).
Closest human hit: AAA83406, UDP-Glucuronosyltransferase ($P(n) = 2e - 56$).

CG12408, 140 aa, troponin C relative.
Best matches are in Drosophila, including TpnC41C ($P(n) = 1e - 41$).

CG17510, 98 aa, related to Fis1.
Closest human hit: AF151893, human Fis1 (role in mitochondrial fission) ($P(n) = 3e - 6$).

CG17508, 321 aa, has human homolog of unknown function.
Closest relative: Drosophila CG15403 ($P(n) = 5e - 39$).
Closest human hit: NP_543011.1, chromosome 20, open reading frame 108 ($P(n) = 6e - 38$).

CG11665, 442 aa, monocarboxylate transporter.
Closest relatives are several similar proteins in Drosophila, *e.g.*, CG8271 ($P(n) = 3e - 12$).
Closest human hit: O60669, monocarboxylate transporter 2.

AE003785 (7 genes up to and including Vulcan)

CG1344, 641 aa, Arm/HEAT repeat, N-terminal kinase like domain.
Closest human hit: NP_065156.1, hypothetical protein LOC57147 ($P(n) = 2e - 51$).

CG8426, 948 aa, Cdc39/NOT3 transcriptional repressor homolog.
Closest human hit: NP_055331.1, CCR4-NOT transcription complex, subunit 3 ($P(n) = 2e - 74$).

CG8245, 343 aa, has human homolog of unknown function.
Closest human hit: NP_078863.1, hypothetical protein ($P(n) = 5e - 35$).

CG1298, 264 aa.
Related to Sinuous = CG10624 ($P(n) = 1e - 22$).
No close human hits.

CG11066, 655 aa, Serine protease, prophenoloxidase activating factor?
Closest relative is Drosophila CG40160 ($P(n) = 7e - 23$).
Insect relative with known function: CAC12665.1, prophenoloxidase activating factor ($P(n) = 6e - 20$).
Closest human hit: (likely NOT ortholog) NP_000883.1, plasma kallikrein B1 precursor ($P(n) = 9e - 16$).

CG17337, 462 aa, glutamate carboxypeptidase.
Closest human hit: CAC69883.1, glutamate carboxypeptidase ($P(n) = e - 172$).

Vulcan (CG8390), 605 aa.
Closest human hit: T03306, PSD-95/SAP90-associated protein-2 ($P(n) = 9e - 14$).

aa, amino acid.

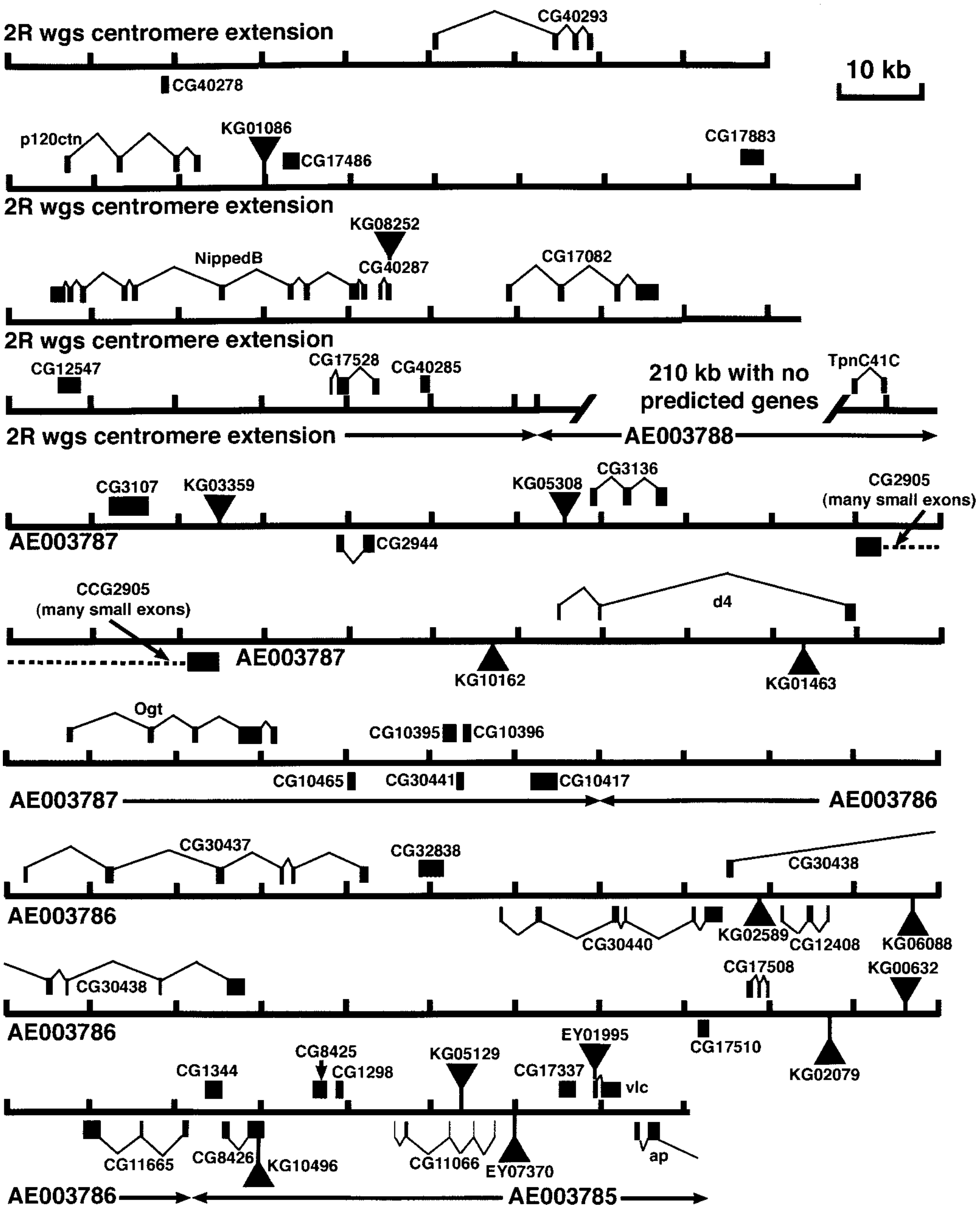

FIGURE 2.—Gene density varies across the 41C interval. Scaffolds, genes, and *P*-element insertions in 41C are displayed, using information from the BDGP and the *P*-element Gene Disruption Project. Proximal to the centromere is top left and distal is bottom right. Scaffolds are marked at 10-kb intervals. Predicted coding sequences of genes are indicated above (mRNA transcribed away from centromere) or below the contig (transcription is toward the centromere). Gene annotations are found in Table 1. Genes determined to be transposon remnants (supplemental Table 1 at http://www.genetics.org/supplemental/) were not included. *P*-element insertions from the *P*-element Gene Disruption Project (KGXXXXX and EYXXXXX) are indicated as triangles.

CG17486. We analyzed this by FlyBlast, using 1- to 2-kb segments of the nucleotide sequence as queries to search the transposon and repeat databases of the BDGP, as well as the EST, predicted gene, and genomic databases (http://www.fruitfly.org/blast/). The majority of the region was composed of repetitive DNA, largely the remnants of various transposons and retrotransposons. In most cases, only fragmentary elements appeared to be present, which were internally deleted or otherwise rearranged. Two elements, 1360/Hoppel, an element in the terminally inverted repeat class of DNA transposons (Kholodilov *et al.* 1988; Kaminker *et al.* 2002), and Narep1/Dine1, which has weak similarity to SINE retrotransposons but is structurally distinct from the major retrotransposon classes (Locke *et al.* 1999a; Kaminker *et al.* 2002), together account for 7 kb (~26%) of the 27 kb. These elements are also overrepresented in sequenced regions of the fourth chromosome (Locke *et al.* 1999b; Kaminker *et al.* 2002), which is largely heterochromatic. Various LTR-class retrotransposons make up another significant fraction of the *p120* region (5 kb; ~19%). In addition to matches to known transposable elements, other regions were clearly repetitive, although they were not closely related to any known transposon. Only a small block of simple sequence DNA was found in this region ((TA_n)), in contrast to what is observed in the centromeric region (Sun *et al.* 2003). The coding exons of the two genes in the region are very closely hemmed in by repetitive DNA, both in their 5′ and 3′ flanking regions and in their introns, and the

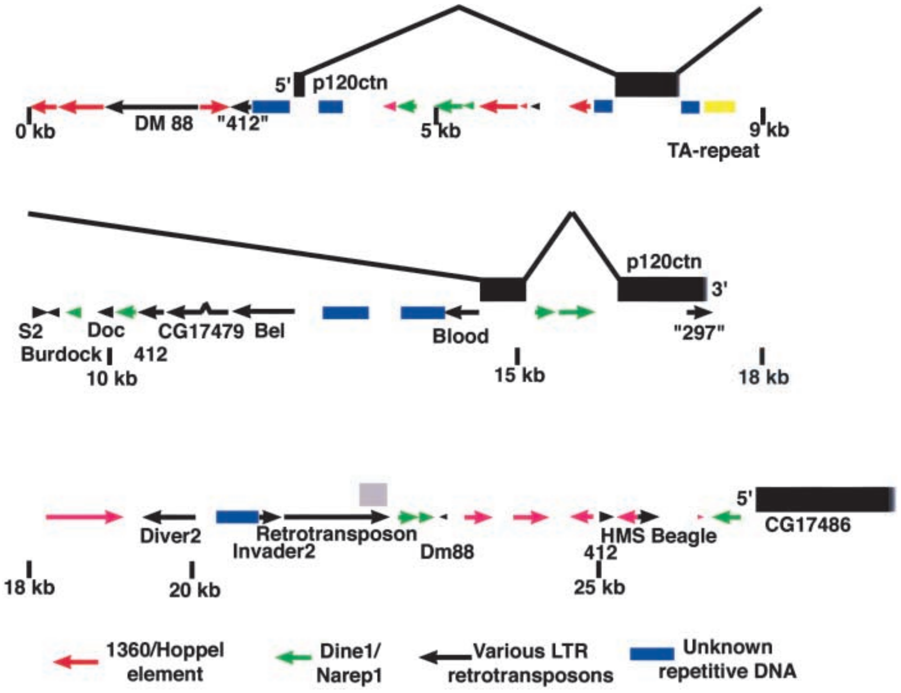

Figure 3.—The region surrounding *p120* is highly repetitive. An analysis of 23 kb of sequence spanning the *p120* gene and extending distal to *CG17486* is presented. The exons and introns of the two genes are displayed as black boxes and thin lines, respectively. The gray box indicates an incorrectly annotated "fifth exon" of p120, which was removed in WGS3. The majority of the sequence surrounding these two genes is highly repetitive. Lines with arrowheads represent blocks of sequence that are identifiable as remnants of known transposons. Colored rectangles represent other repetitive DNA, as indicated in the key. Portions of the remaining DNA may be repetitive as well.

p120 3′ untranslated region includes a retrotransposon remnant. In the *light* region of 2L, exons are also embedded in repetitive DNA (Hoskins *et al.* 2002).

A genetic screen for essential genes in the p120 region: Having this picture of predicted gene content as a foundation, we initiated genetic analysis of the essential genes in the region. Our initial goal was to obtain mutations in p120, which encodes a component of the cell-cell adherens junction and is well conserved in all animals thus far examined (reviewed in Anastasiadis and Reynolds 2000). We began with the hypothesis that p120 would be an essential gene and thus set out to collect lethal mutations in the region to which it mapped.

We first used *in situ* hybridization to polytene chromosomes to map *p120* to a region in 41C defined by the overlap between *Df(2R)M41A10* and *Df(2R)M41A8* (Myster *et al.* 2003; Figure 4A). In addition, we determined that *p120* was not deleted by *Df(2R)nap1*, but was deleted by *Df(2R)M41A4* (data not shown). Our analysis used polytene chromosomes, which have lower cytological resolution than mitotic chromosomes. Previous analysis of mitotic chromosomes suggests that *Df(2R)M-41A10* removes the entire 2R mitotic heterochromatin (h39–h46), while *Df(2R)M41A4* deletes only h46 (Dimitri 1991), suggesting that *p120* maps to h46, a conclusion supported by the recent work of Corradini *et al.* (2003). We obtained from our colleagues alleles of the known genes in the region: *Nipped-A, Nipped-B, l(2)41Ae, l(2)41Af, l(2Rh)IR3,* and *l(2Rh)IR23* (Hilliker 1976; Dimitri *et al.* 1997; Rollins *et al.* 1999). We also initiated a search for new mutations. We selected *Df(2R)M-41A8* for further genetic analysis as it was the smallest deficiency in our initial analysis that removed *p120,* and it was a relatively healthy stock.

We then carried out a screen for lethal mutations uncovered by *Df(2R)M41A8* (Figure 4B). We EMS mutagenized males carrying an isogenic second chromosome marked with the recessive visible markers *cn* and *bw* and crossed them to females carrying a second chromosome balancer (see materials and methods). Balanced F_1 males were individually mated to balanced females carrying *Df(2R)M41A8,* and crosses were screened for those in which all of the progeny carried the balancer—*i.e.*, stocks in which a new mutation that was lethal over the *al Df(2R)M41A8* chromosome had been induced. Unbalanced progeny were also scored for visible phenotypes. Candidate lethal or visible mutations were retested to verify the original result. We screened 6284 chromosomes and recovered 226 lethal mutations and 5 visible mutations. The 5 visible mutants all share the same partially penetrant phenotype when *trans*-heterozygous with *Df(2R)M41A8*: they have ectopic wing veins posterior to longitudinal vein 5. To date, these have not been analyzed further.

Placing deficiencies on the physical map and using them to map new mutations: In addition to the deficiencies we initially analyzed, we obtained from others or generated (see below) a number of other deficiencies in the region, many of which were smaller than that used for the screen (for purposes of this analysis, we hypothesize that these represent deficiencies rather than more complex rearrangements—while the latter possibility remains, the data below are consistent with most or all being simple deficiencies). We characterized existing and newly generated chromosomal deficiencies in two ways: we mapped their endpoints on the physical map by PCR, and we characterized them genetically by crossing them both to the preexisting complementation

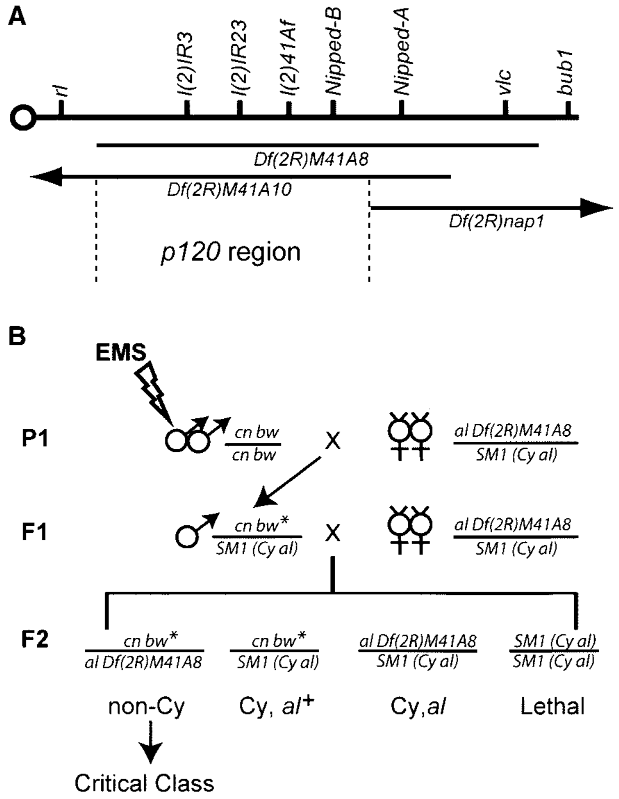

Figure 4.—Mutagenesis screen to identify lethal mutations at the 2R heterochromatin/euchromatin junction. (A) Previously identified lethal loci and overlapping deficiencies in the region. *Df(2R)M41A8* and *Df(2R)M41A10* remove *p120* whereas *Df(2R)nap1* does not (Myster *et al.* 2003). (B) Outline of the strategy for the EMS mutagenesis screen to generate recessive lethal mutations uncovered by *Df(2R)M41A8* (see materials and methods for details).

groups in the region and to our newly generated mutations. Deficiency endpoints were mapped by PCR amplification from multiple DNA preparations from single homozygous deficiency embryos (selected using a GFP-marked balancer), using primer pairs throughout the region. For each DNA preparation we used a set of primers from outside the region as a positive control for the quality of the DNA, and we used a wild-type strain as a positive control for each primer pair. Because of the repetitive nature of most of the DNA, we selected primer pairs from the coding sequence of predicted genes, with the result that our resolution is limited by the density of predicted genes in a region. This anchored the deficiency map on the physical map (Figure 5). Our mapping of *Df(2R)M41A10* is also consistent with the mapping of BAC clones by FISH onto chromosomes carrying this deficiency (Corradini *et al.* 2003).

We then characterized the lethal mutations we generated in our screen (Figure 6). We first crossed them to a subset of the deficiencies in the region, allowing us to assign them to given deficiency intervals. We then crossed them to additional deficiencies, known mutations in the region, and to one another. This allowed us to place all of the mutations into complementation groups, many of which were ordered with respect to one another (Figure 6; unordered complementation groups are joined by brackets). Interestingly, 14 of the complementation groups map more centromere proximal within the heterochromatin, in a region proximal to the contiguously assembled sequence (Hoskins *et al.* 2002). The deficiencies also allowed us to connect the genetic and physical maps by providing common points of reference. We mapped alleles of the cloned gene *Nipped-B* (Rollins *et al.* 1999), as well as mutations in *p120* and *CG17486* (from the *P*-mobilization screen described below), providing three additional anchor points between the physical and genetic maps. Interestingly, our EMS screen generated many deficiencies, in addition to the expected point mutations (Figure 6). Some are relatively small and fail to complement alleles at only two loci whereas others are quite large and fail to complement all of the mutant genes generated in the screen. Two deficiencies extend even more distally, failing to complement *bub1*, which lies outside *Df(2R) M41A8*.

Generating additional deletions in the *p120* region: None of the complementation groups from our initial analysis was a good candidate for a mutation in *p120*. Only one initially mapped to the same deficiency interval as *p120,* and sequencing of the *p120* coding region from that mutant line [*l(2)41Af*] revealed no mutations. We thus needed an alternate approach. Fortunately, by this point the *P*-element screen/Gene Disruption Project of the Bellen/Rubin/Spradling labs had begun generating and mapping new *P*-element insertions (H. Bellen, R. Hoskins, R. Levis, G. Luo, G. M. Rubin and A. C. Spradling, unpublished data; http://flypush.imgen.bcm.tmc.edu/pscreen/) and had used as one of their *P* elements the *SUPorP P* element. This carries a *white*$^+$ gene surrounded by insulator elements from the *suppressor of Hairy wing,* helping insulate the gene from chromosomal position effects (Roseman *et al.* 1995). It also carries a *yellow*$^+$ gene outside the insulators. Previous work suggested that insertions of this *P* element would be more effectively recovered from the heterochromatin (Roseman *et al.* 1995), and this has been borne out in our region of interest. The *P*-element Gene Disruption Project recovered at least 14 new *P*-element insertions in 41C, 12 of which are *SUPorP* insertions (Figure 2). Most are intergenic insertions and are viable. Even these *SUPorP* insertions are biased toward the more distal scaffolds.

One of these insertions, *KG01086,* is ~7 kb 3′ to *p120* and 2 kb 5′ of *CG17486.* This insertion is viable and fertile. We mobilized this insertion (see materials and methods), generating 401 putative mobilizations from 600 crosses. Among these was one relatively large deficiency, *Df(2R)247,* which deletes many genes (Figures 5 and 6). We used this deficiency in the mapping of complementation groups described above. The mobili-

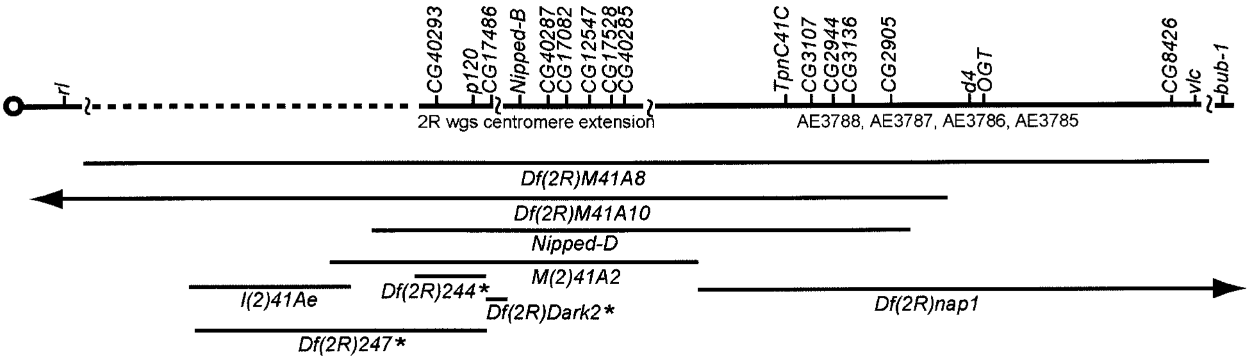

Figure 5.—Placing deficiencies on the physical map. At the top is a diagram of the physical map (this is only roughly to scale—a correctly scaled version is presented in Figure 2). Genes used in defining the endpoints of deficiencies are indicated above each scaffold. The region indicated by a dotted line has not been assembled into finished sequence and thus was not included in this analysis. The centromere is to the left. Deficiency endpoints were determined by PCR amplification of coding sequence from the indicated genes, using homozygous mutant genomic DNA as a template (see materials and methods). Asterisks indicate deficiencies generated by mobilization of the *KG1086* SUPorP *P* element.

zation of *KG01086* also generated smaller deficiencies confined to the immediate region of *p120* and its neighboring genes. We mapped these 401 lines using a standard set of PCR reactions, searching for deletions with one endpoint in the *P* element and the other in flanking DNA (the mapping of those that delete *p120* is described in detail in Myster *et al.* 2003). We began with two primer pairs: one within the *P*-element inverted repeat and one in the DNA flanking the insertion to the right or left. We scored for the presence or absence of a PCR product and used a primer pair from outside the region as a positive control for the quality of the DNA preparation. For lines in which one end of the *P* element was deleted, we then used primer pairs within the coding exons of *p120, CG40293*, and *CG17486* to amplify DNA from homozygous mutant lines to determine the extent of the deletion (for all PCR reactions, the *KG01086* strain was used as a positive control). Representative PCR data for the strains deleting *p120* can be seen in Myster *et al.* (2003). This revealed deletions in both directions of a variety of sizes. One, *Df(2R)Dark2*, deletes *CG17486*, but does not extend into *Nipped B* (as determined genetically). This deletion is viable, demonstrating that *CG17486* is not essential. Two others delete *p120* and do not extend into the next gene—the phenotype of these is described in detail elsewhere (Myster *et al.* 2003)—but they are viable and fertile, demonstrating that *p120* is not essential. Finally, *Df(2R)244* deletes both *p120* and *CG40293*. This deletion is viable and fertile, demonstrating that *CG40293* is also not essential.

Correlating the genetic and physical maps: We then used our alignment of the genetic and physical maps to identify a candidate for the *Nipped-A* gene. *Nipped-A* was originally identified as a modifier of the phenotype of the effects of certain *cut* mutations on the wing. It is the sole complementation group that fails to complement both the *Nipped-D* and *Df(2R)nap1* deficiency strains. Five predicted genes are removed by these deficiencies: *TpnC41C, CG3107, CG2944, CG3136,* and *CG2905* (Figure 6). We initially used RT-PCR to analyze transcripts from 10 different *Nipped-A* alleles, hypothesizing that one of these alleles might not produce a stable mRNA. However, a product of the predicted size was generated, using exonic primers designed to amplify *CG3107, CG2944, CG3136*, and *CG2905* (data not shown). Of the five genes in the region, *CG2905* is the largest, spanning ~35 kb, containing 15 predicted exons, and encoding a 3435-amino-acid predicted protein that is the homolog of mammalian TRRAP and yeast Tra1 (Grant *et al.* 1998; Kusch *et al.* 2003). Because *Nipped-A* was mutated the most frequently in our screen (36 alleles), we hypothesized that *Nipped-A* alleles might have mutations in the *CG2905* gene. The *CG2905* coding region was sequenced from three alleles of *Nipped-A* [two EMS-induced alleles generated in our study (*l(2)NC116* and *l(2)NC186*) and a γ-ray-induced allele (*Nipped-A*$^{357.2}$; Rollins *et al.* 1999)] . In *l(2)NC116*, a G-to-A transition was identified in the first base of the intron following exon 4. This position lies in the highly conserved GT dinucleotide in the 5′ splice site consensus sequence (Mount *et al.* 1992) and thus should disrupt proper splicing (Figure 7, top). In *l(2)NC186*, an A-to-T transversion that is predicted to result in a nonconservative valine to aspartic acid missense mutation at amino acid 885 was identified. This lies in a region conserved in the Drosophila, human, and Arabadopsis homologs: all have valine at this position (Figure 7, bottom). No mutations in the *CG2905* predicted coding region were identified in *Nipped-A*$^{357.2}$, but due to the complex genomic structure of *NippedA*, it may be that this γ-ray-induced allele results from a DNA rearrangement with a breakpoint in an intron or 5′ to the coding sequences.

We identified one additional anchor between the genetic and physical maps by examining additional *SUPorP* insertion lines in 41C whose physical location has been determined by sequence analysis of the insertion junc-

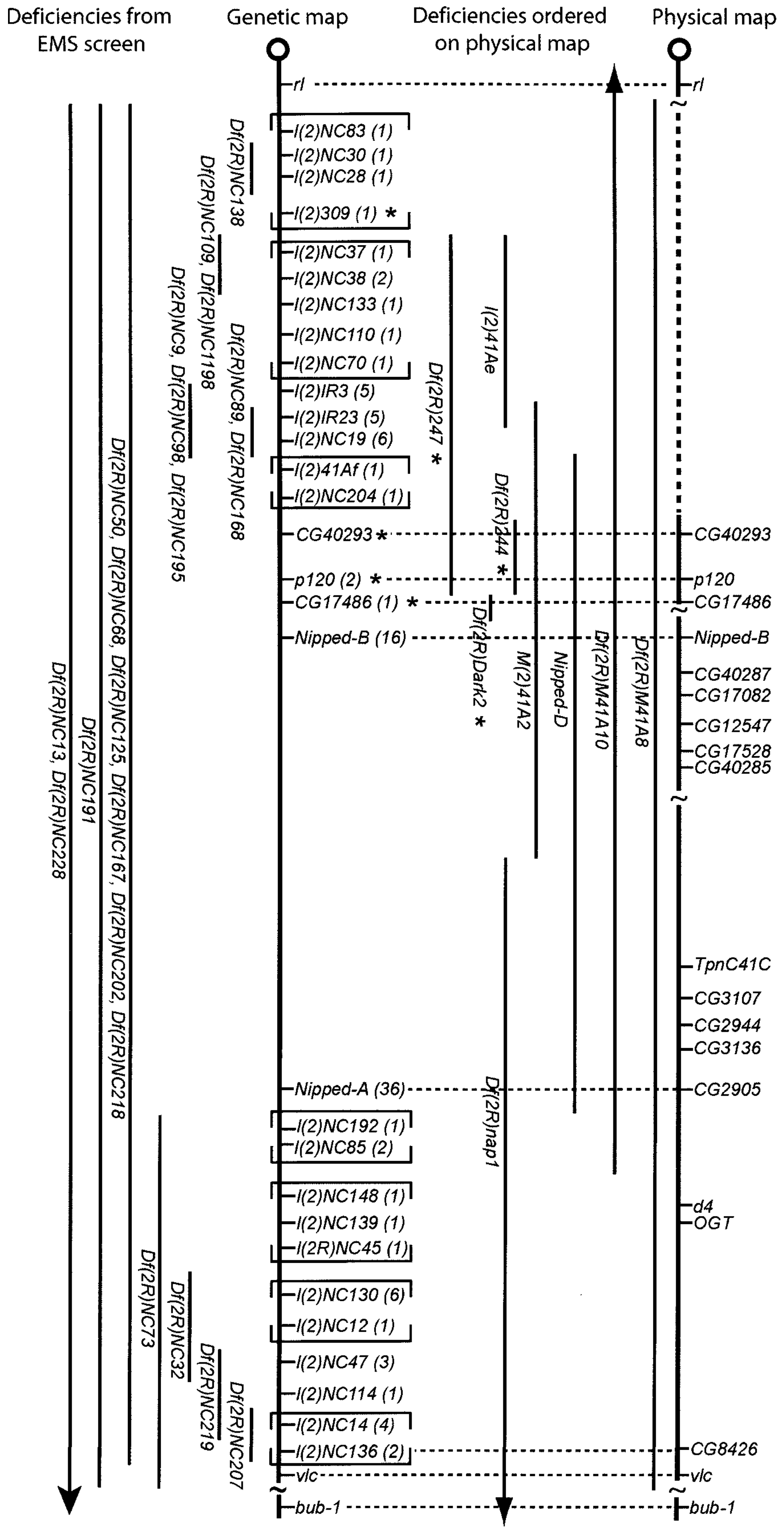

FIGURE 6.—Correlating the genetic and physical maps. At the right is the set of overlapping deficiencies correlated with the physical map, as displayed in Figure 5. To the left of them is the genetic map, with lethal complementation groups ordered on the basis of complementation data with the set of overlapping deficiencies (distances on the genetic map are arbitrary). The order of loci within deficiency intervals has not been determined, and brackets join loci unordered with respect to one another. The number of alleles for each complementation group generated in our screens is indicated in parentheses. Dashed lines represent anchor points where complementation groups (including the nonessential genes *CG40293*, *p120*, and *CG17486*) have been assigned to mutations in identified genes. At the far left are deficiencies generated in our EMS mutagenesis screen, the endpoints of which have been mapped genetically but that have not been placed on the physical map. The extent of each deletion was determined by complementation testing against the lethal loci. Multiple deficiencies that fail to complement the same loci are listed together. The distal endpoint of two deficiency lines has not been determined and is indicated by an arrowhead.

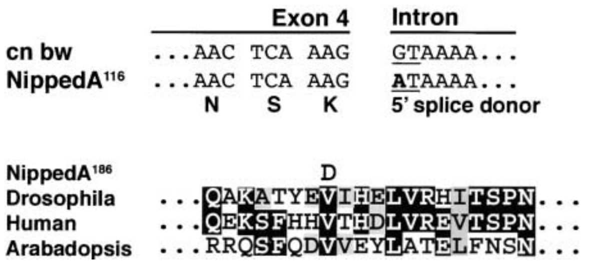
Exon 4
Intron
cn bw ...AAC TCA AAG GTAAAA...
NippedA116 ...AAC TCA AAG ATAAAA...
N S K 5' splice donor
NippedA186 D
Drosophila ...QAKATYEVIHELVRHITSPN...
Human ...QEKSFHHVTHDLVREVTSPN...
Arabadopsis ...RRQSFQDVVEYLATELFNSN...

Figure 7.—*Nipped-A* mutations affect CG2905/Tra1/TRRAP. (Top) Nucleotide sequence of the junction between exon 4 and the downstream intron of *CG2905*, from *cn bw*, the isogenic stock in which the mutations were induced, and *Nipped-A*^{116}. This mutation alters the conserved GT dinucleotide that is an essential part of the splice donor site (Mount *et al.* 1992). (Bottom) A portion of the predicted amino acid sequence of CG2905 (amino acids 878–897) and the corresponding region of its human and Arabidopsis homologs. Identical residues are indicated by white type in black boxes, while similar amino acids are indicated by black type in gray boxes. The valine residue affected by *Nipped-A*^{186} is indicated.

tion. Of 17 insertion lines, only 1, *KG10496*, appears to be lethal, as assessed by the presence or absence of homozygous flies in the stocks. This line is inserted into the coding region of *CG8426*, a predicted transcription factor. The physical location predicts that the insertion line would fail to complement *DF(2R)M41A8* and *Df(2R) nap1* and would complement *Nipped-D* and *M(2)41A2*. Our complementation tests confirmed this (data not shown). Testing of EMS lines in the region identified *l(2)NC136* as allelic to the *KG10496* insertion (Figure 6).

DISCUSSION

Our interest in the proximal region of the second chromosome was initiated by the fact that *p120*, a gene that encodes a component of adherens junctions, is located at polytene band 41C. Core components of adherens junctions are required to establish cellular adhesive contacts and mutations in many adherens junction components are embryonic lethal (for reviews see Yap *et al.* 1997; Tepass *et al.* 2001). With the goal of identifying mutations in *p120*, two screens for mutations in the region were performed. In addition to providing information about the phenotype of *p120*, which was recently published elsewhere (Myster *et al.* 2003), the screen provides an example of how classical genetic approaches can be combined with sequence information, annotation, and bioinformatics tools emerging from the genome project to create a combined genetic and bioinformatics picture of a heterochromatic region, as a resource for future work on the genes within it.

We identified lethal mutations in 26 loci, 21 of which appear to be novel. Fourteen of the complementation groups map in the heterochromatin proximal to the assembled sequence [recent analysis of mitotic chromosomes by FISH with BACs from the assembled sequence suggests that these will map in chromosome region h45 or more proximal (Corradini *et al.* 2003)]. Thus this region of the heterochromatin may contain many essential loci that remain to be molecularly analyzed (the only molecularly characterized mutation in this proximal region is *rolled*). While we clearly did not reach saturation (many complementation groups have a single allele), our data also begin to suggest that the heterochromatin may contain many nonessential although conserved loci. We directly identified three such genes, *p120, CG40293*, and *CG17486*, and the excess of identified CGs to genetically identified genes in the region between *p120* and *vulcan* suggests that a number of other loci may be nonessential.

Heterochromatin and unique coding DNA sequences: The pericentric heterochromatin is composed of the centromeric region, whose composition is largely simple sequence DNA arranged in tandem repeats (Sun *et al.* 1997, 2003), and the "boundary region," where transposons comprise much of the sequence (*e.g.*, Hoskins *et al.* 2002). Genes are absent from the centromere (with the exception of scattered retrotransposons), while in the boundary region gene density is quite low, confined to small islands of unique coding sequences interspersed throughout. Previous mutational screens identified a small number of lethal loci in the heterochromatin of 2R and in this study we report the generation of additional alleles of each (Hilliker 1976; Dimitri *et al.* 1997; Rollins *et al.* 1999). Surprisingly, our screens identified a number of new lethal loci in the heterochromatin proximal to the published sequence. This suggests that there may be many more essential loci in this region of the heterochromatin than was previously thought and fits well with recent work by the genome project that predicts ~450 genes in the heterochromatin as a whole (Hoskins *et al.* 2002; Misra *et al.* 2002).

One reason for the earlier underestimate of essential gene number in the 2R heterochromatin is that earlier screens likely did not reach saturation, which is also likely for our own screen, as described above. A second potential cause of the previous underestimate is the apparent tendency for mutagens to generate deficiencies at a high rate in this region (Figure 6, left; see below). For example, we interpret our data to suggest that one of the lethal loci described in an earlier screen, *l(2)41Ae* (Hilliker 1976), may be a deletion, as it fails to complement seven of our complementation groups (Figure 6). If this is the correct interpretation of these data, it would suggest that the number of lethals in the region has been underestimated. However, two caveats to this conclusion must be noted. First, Hilliker offered an alternate explanation for the behavior of *l(2)41Ae* in complementation tests. He suggested that it is a complex locus, with an unusual degree of intraallelic complementation, and thus suggested that all of the complementation groups in this region are alleles of a single complex locus. This is a possibility, although we favor our own interpretation. Second, as some of our complementation groups contain only a single allele, their pattern of complementation is slightly less secure than that of complementation groups with multiple alleles.

The *SUPorP* transposable element allows genetic access to heterochromatin: The repetitive nature of heterochromatin made it challenging to clone, sequence, and correctly assemble in large-scale sequencing efforts. Recent efforts have made inroads into these regions of the genome (Hoskins *et al.* 2002; Sun *et al.* 2003). It is estimated that ~60 Mb of heterochromatin are in the genome of Drosophila females and 90 Mb of heterochromatin in males (Celniker *et al.* 2002; Hoskins *et al.* 2002). A genomics-based estimate of total heterochromatic gene number will have to await the completion of the sequence in this region, but current estimates suggest that there are ~450 genes are in the heterochromatin as a whole (Hoskins *et al.* 2002; Misra *et al.* 2002).

Genetics-based approaches provide an alternative method for identifying genes in heterochromatin. *P*-element transposons can effectively insert into heterochromatin (Roseman *et al.* 1995; Yan *et al.* 2002), and thus the low number of insertions previously identified in heterochromatin is probably due to its gene silencing properties (for a review see Weiler and Wakimoto 1995), preventing the expression of the scorable markers. When we started our work no known *P*-element insertions were close to *p120*, ruling out *P*-element mobilization as a viable mutagenesis strategy. Fortunately, modified *P* elements have provided access to the heterochromatin. The *SUPorP* was designed to allow efficient insertion in silenced regions. In it, the *white*$^+$ selectable marker is flanked by insulator elements carrying Suppressor of hairy wing binding sites, effectively blocking the silencing properties of the heterochromatin (Roseman *et al.* 1995). Yan *et al.* (2002) utilized this element as well, screening for variegated expression of the *yellow*$^+$ selectable marker, which is not flanked by insulators elements. Together, these efforts have allowed the identification of many new insertions into previously untagged regions of the genome (Yan *et al.* 2002; H. Bellen, R. Hoskins, R. Levis, G. Luo, G. M. Rubin *et al.*, unpublished data; http://flypush.imgen.bcm.tmc.edu/pscreen/).

These insertions provide the ability to genetically manipulate the surrounding region, both through the direct insertional inactivation of genes and through mobilization of the *P* elements to create new insertions or deletions. Our work provides an illustration of each of these. We found that *l(2)NC136* is allelic to the *KG10496* insertion. After mobilizing the *P* element in the *p120/CG17486* intragenic region, we identified five deficiencies of variable length extending in both directions from the original insertion site among 600 mobilization events. In addition, a local hop identified an additional lethal complementation group [*l(2)309*] proximal to *p120*. An added advantage of screening at the molecular level is that nonessential mutations can be identified. Our screen revealed that mutations in *p120*, *CG40293*, and *CG17486* are viable and fertile. These illustrate how the growing bank of *P*-element insertions in the heterochromatin will be a great resource to identify or analyze both lethal and nonessential heterochromatic loci in the future.

Mutagens and repetitive DNA: EMS is generally considered to be a point mutagen, and previous mutagenesis of the euchromatin supports this (*e.g.*, Gray *et al.* 1991). We were thus surprised to find that many of our EMS-generated alleles (Figure 6, left), as well as alleles generated in earlier screens in the region (Rollins *et al.* 1999), are deficiencies that fail to complement multiple loci. We suspect that the repetitive DNA in the region may contribute to this. After induction of a new mutation, the cellular repair mechanisms are activated and use the complementary strand as a template for repair. Due to the highly repetitive nature of the region, we imagine that in the process of repairing individual base-pair mutations, misalignment could occur, resulting in a looping out of a region of DNA. This could lead to the generation of a deficiency.

Anchoring the genetic and physical maps: Our EMS screen generated additional alleles of each of the previously identified loci in the *p120* region, including 36 new mutations in *Nipped-A* and 16 new mutations in *Nipped-B*. Conversely, 16 of the newly identified complementation groups contain a single member. Taken together, these results imply that some loci are highly mutable and our screens are probably not saturating. The published Drosophila genomic sequence and its annotation provide a powerful data set that could be used to learn more about our many newly identified loci (Adams *et al.* 2000 ; Celniker *et al.* 2002; Hoskins *et al.* 2002; Misra *et al.* 2002). We used the set of overlapping deficiency strains to genetically order many of our complementation groups with respect to each other and exploited the deficiency lines as an inroad to correlate the genetic and physical maps. In addition, the *Nipped-B* gene was previously analyzed at the molecular level and thus provided an additional anchor point between the two maps (Rollins *et al.* 1999). We focused on the 14 complementation groups that are distal to *p120*. *Nipped-A* is the only locus that fails to complement both the *Nipped-D* and *Df(2R)nap1* deficiencies (see Figure 6). Five genes are predicted to be located in this interval. Interestingly, one of these genes, *CG2905*, is very large, with 15 exons encoding a 3435-amino-acid protein. Due to the high mutability of *Nipped-A* (36 alleles in our EMS screen) we suspected that *CG2905* encoded *Nipped-A*. This appears to be the case, as mutations in the *CG2905* gene were identified in two alleles of *Nipped-A* that were generated in this study (Figure 7).

Nipped-A was originally identified in a screen for genes that modified the phenotype of a regulatory allele of the *cut* gene (Rollins *et al.* 1999). *cut* has a complex regulatory region, with distant enhancers that regulate tissue-specific expression. The *cut* mutation used in the screen was caused by an insertion of the *gypsy* retrotransposon, which has the ability to block interactions of distal enhancers with promoters. Mutations in a number of different genes were isolated in this screen. They include mutations of transcriptional regulators like Chip and Mastermind, as well as mutations in Nipped-B, a member of a family of proteins involved in chromatid cohesion, chromosome condensation, and DNA repair. Our identification of *Nipped-A* as a mutation in *CG2905* fits into this picture, as *CG2905* encodes the fly homolog of Tra1/TRRAP, a component of SAGA/GNAT-type multiprotein histone acetyltransferase complexes (Grant *et al.* 1998; Kusch *et al.* 2003). Tra1/TRRAP is a distant relative of ATM, the gene mutated in the human disease ataxia-telangiectasia (reviewed in Shiloh 2000), and is thus thought by analogy to be a protein kinase or possibly a lipid kinase. Our identification of *Nipped-A* with Tra1/TRRAP opens the way for genetic analysis of the role of this protein complex in transcriptional regulation in Drosophila.

Our alignment of the genetic and physical maps provides a framework for future molecular identification studies. It is our hope that future investigators will utilize our reagents and view of the heterochromatin-euchromatin region of 2R as a starting point for examining the function of the genes in this interesting region of the genome.

We thank P. Cayirlioglu for advice on and V. Morel for assistance with *P*-element mobilization; K. Maners for helping with complementation tests; R. Hoskins, A. Hilliker, J. Gates, B. McCartney, M. Price, and the two anonymous reviewers for comments on the manuscript; and members of the Peifer lab for helpful discussions. We are also very grateful to A. Hilliker, P. Dimitri, D. Dorsett, the *P*-element Gene Disruption Project, and the Bloomington Drosophila Stock Center for fly stocks and are especially grateful to R. Rollins and D. Dorsett as well as R. Hoskins and the Berkeley *Drosophila* Genome Project for sharing unpublished data and for many helpful discussions. This work was supported by National Institutes of Health grant GM47857 to M. Peifer. S. H. Myster was supported by National Institutes of Health National Research Service Award GM19888, R. Cavallo by a Department of Defense Breast Cancer Research Program predoctoral fellowship, C. T. Anderson by the Pfizer Summer Undergraduate Research Fellowship program, C. T. Anderson and S. Bhotika by Thompson Undergraduate Research awards, and M. Peifer in part by a Department of Defense Breast Cancer Research Program Career Development Award and by the Welsh Distinguished Term Professorship.

LITERATURE CITED

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of Drosophila melanogaster. Science **287:** 2185–2195.

Anastasiadis, P. Z., and A. B. Reynolds, 2000 The p120 catenin family: complex roles in adhesion, signaling and cancer. J. Cell Sci. **113:** 1319–1334.

Ashburner, M., S. Misra, J. Roote, S. E. Lewis, R. Blazej *et al.*, 1999 An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the Adh region. Genetics **153:** 179–219.

Casso, D., F. Ramirez-Weber and T. B. Kornberg, 2000 GFP-tagged balancer chromosomes for Drosophila melanogaster. Mech. Dev. **91:** 451–454.

Celniker, S. E., D. A. Wheeler, B. Kronmiller, J. W. Carlson, A. Halpern *et al.*, 2002 Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol. **3:** RESEARCH0079.

Cleveland, D. W., Y. Mao and K. F. Sullivan, 2003 Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. Cell **112:** 407–421.

Corradini, N., F. Rossi, F. Verni and P. Dimitri, 2003 FISH analysis of Drosophila melanogaster heterochromatin using BACs and P elements. Chromosoma **112:** 26–37.

Dimitri, P., 1991 Cytogenetic analysis of the second chromosome heterochromatin of *Drosophila melanogaster*. Genetics **127:** 553–564.

Dimitri, P., and N. Junakovic, 1999 Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. Trends Genet. **15:** 123–124.

Dimitri, P., B. Arca, L. Berghella and E. Mei, 1997 High genetic instability of heterochromatin after transposition of the LINE-like I factor in Drosophila melanogaster. Proc. Natl. Acad. Sci. USA **94:** 8052–8057.

Gloor, G. B., C. R. Preston, D. M. Johnson-Schlitz, N. A. Nassif, R. W. Phillis *et al.*, 1993 Type I repressors of *P*-element mobility. Genetics **135:** 81–95.

Grant, P. A., D. Schieltz, M. G. Pray-Grant, J. R. Yates, III and J. L. Workman, 1998 The ATM-related cofactor Tra1 is a component of the purified SAGA complex. Mol. Cell **2:** 863–867.

Gray, M., A. Charpentier, K. Walsh, P. Wu and W. Bender, 1991 Mapping point mutations in the Drosophila rosy locus using denaturing gradient gel blots. Genetics **127:** 139–149.

Grewal, S. I., and S. C. Elgin, 2002 Heterochromatin: new possibilities for the inheritance of structure. Curr. Opin. Genet. Dev. **12:** 178–187.

Grigliatti, T. A., 1998 Mutagenesis, pp. 55–83 in *Drosophila: A Practical Approach*, edited by D. B. Roberts. Oxford University Press, New York.

Henikoff, S., 2000 Heterochromatin function in complex genomes. Biochim. Biophys. Acta **1470:** 01–08.

Hilliker, A. J., 1976 Genetic analysis of the centromeric heterochromatin of chromosome 2 of *Drosophila melanogaster*: deficiency mapping of EMS-induced lethal complementation groups. Genetics **83:** 765–782.

Hoskins, R. A., C. R. Nelson, B. P. Berman, T. R. Laverty, R. A. George *et al.*, 2000 A BAC-based physical map of the major autosomes of Drosophila melanogaster. Science **287:** 2271–2274.

Hoskins, R. A., C. D. Smith, J. W. Carlson, A. B. Carvalho, A. Halpern *et al.*, 2002 Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome Biol. **3:** RESEARCH0085–0085.

Kaminker, J. S., C. M. Bergman, B. Kronmiller, J. Carlson, R. Svirskas *et al.*, 2002 The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. **3:** RESEARCH0084.

Kholodilov, N. G., V. N. Bolshakov, V. M. Blinov, V. V. Solovyov and I. F. Zhimulev, 1988 Intercalary heterochromatin in Drosophila. III. Homology between DNA sequences from the Y chromosome, bases of polytene chromosome limbs, and chromosome 4 of D. melanogaster. Chromosoma **97:** 247–253.

Kusch, T., S. Guelman, S. M. Abmayr and J. L. Workman, 2003 Two Drosophila ada2 homologues function in different multiprotein complexes. Mol. Cell. Biol. **23:** 3305–3319.

Locke, J., L. T. Howard, N. Aippersbach, L. Podemski and R. B. Hodgetts, 1999a The characterization of DINE-1, a short, interspersed repetitive element present on chromosome 4 and in the centric heterochromatin of Drosophila melanogaster. Chromosoma **108:** 356–366.

Locke, J., L. Podemski, K. Roy, D. Pilgrim and R. Hodgetts, 1999b Analysis of two cosmid clones from chromosome 4 of Drosophila melanogaster reveals two new genes amid an unusual arrangement of repeated sequences. Genome Res. **9:** 137–149.

Miklos, G. L., M. T. Yamamoto, J. Davies and V. Pirrotta, 1988 Microcloning reveals a high frequency of repetitive sequences characteristic of chromosome 4 and the beta-heterochromatin of Drosophila melanogaster. Proc. Natl. Acad. Sci. USA **85:** 2051–2055.

Misra, S., M. A. Crosby, C. J. Mungall, B. B. Matthews, K. S. Campbell *et al.*, 2002 Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. Genome Biol **3:** RESEARCH0083.

Moazed, D., 2001 Common themes in mechanisms of gene silencing. Mol. Cell **8:** 489–498.

Mount, S. M., C. Burks, G. Hertz, G. D. Stormo, O. White *et al.*, 1992 Splicing signals in Drosophila: intron size, information content, and consensus sequences. Nucleic Acids Res. **20:** 4255–4262.

Müller, H. J., 1930 Types of viable variations induced by X-rays in *Drosophila.* J. Genet. **22:** 299–334.

Myster, S. H., R. Cavallo, C. T. Anderson, D. T. Fox and M. Peifer, 2003 Drosophila p120catenin plays a supporting role in cell adhesion but is not an essential adherens junction component. J. Cell Biol. **160:** 433–449.

Rollins, R. A., P. Morcillo and D. Dorsett, 1999 Nipped-B, a Drosophila homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes. Genetics **152:** 577–593.

Roseman, R. R., E. A. Johnson, C. K. Rodesch, M. Bjerke, R. N. Nagoshi *et al.*, 1995 A *P*-element containing suppressor of hairy-wing binding regions has novel properties for mutagenesis in Drosophila melanogaster. Genetics **141:** 1061–1074.

Rubin, G. M., M. D. Yandell, J. R. Wortman, G. L. Gabor Miklos, C. R. Nelson *et al.*, 2000 Comparative genomics of the eukaryotes. Science **287:** 2204–2215.

Shiloh, Y., 2000 ATM: sounding the double-strand break alarm. Cold Spring Harbor Symp. Quant. Biol. **65:** 527–533.

Spradling, A. C., D. Stern, A. Beaton, E. J. Rhem, T. Laverty *et al.*, 1999 The Berkeley Drosophila Genome Project gene disruption project: single *P*-element insertions mutating 25% of vital Drosophila genes. Genetics **153:** 135–177.

Sullivan, B. A., M. D. Blower and G. H. Karpen, 2001 Determining centromere identity: cyclical stories and forking paths. Nat. Rev. Genet. **2:** 584–596.

Sun, X., J. Wahlstrom and G. Karpen, 1997 Molecular structure of a functional Drosophila centromere. Cell **91:** 1007–1019.

Sun, X., H. D. Le, J. M. Wahlstrom and G. H. Karpen, 2003 Sequence analysis of a functional Drosophila centromere. Genome Res. **13:** 182–194.

Tepass, U., G. Tanentzapf, R. Ward and R. Fehon, 2001 Epithelial cell polarity and cell junctions in *Drosophila.* Annu. Rev. Genet. **35:** 747–784.

Wallrath, L. L., 1998 Unfolding the mysteries of heterochromatin. Curr. Opin. Genet. Dev. **8:** 147–153.

Weiler, K. S., and B. T. Wakimoto, 1995 Heterochromatin and gene expression in Drosophila. Annu. Rev. Genet. **29:** 577–605.

Yan, C. M., K. W. Dobie, H. D. Le, A. Y. Konev and G. H. Karpen, 2002 Efficient recovery of centric heterochromatin *P*-element insertions in *Drosophila melanogaster.* Genetics **161:** 217–229.

Yap, A. S., W. M. Brieher and B. M. Gumbiner, 1997 Molecular and functional analysis of cadherin-based adherens junctions. Annu. Rev. Cell Dev. Biol. **13:** 119–146.

Communicating editor: J. Tamkun