Genetica **117:** 227–237, 2003.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Y chromosome and other heterochromatic sequences of the *Drosophila melanogaster* genome: how far can we go?

A. Bernardo Carvalho[1], Maria D. Vibranovski[1], Joseph W. Carlson[3], Susan E. Celniker[3], Roger A. Hoskins[3], Gerald M. Rubin[4], Granger G. Sutton[5], Mark D. Adams[5], Eugene W. Myers[5] & Andrew G. Clark[2,5]

[1]*Departamento de Genética, Universidade Federal do Rio de Janeiro, Caixa Postal 68011 CEP 21944-970, Rio de Janeiro, Brazil (Phone: +55-21-2562-6387; Fax: +55-21-2562-6397; E-mail: bernardo@biologia.ufrj.br);* [2]*Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University, University Park, PA 16802, USA;* [3]*Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94708, USA;* [4]*Howard Hughes Medical Institute, BDGP, University of California, Berkeley, CA 94720;* [5]*Celera Genomics, Inc., 45 West Gude Drive, Rockville, MD 20850, USA*

Key words: Drosophila melanogaster, genomics, heterochromatin, whole genome shotgun, Y chromosome

Abstract

Whole genome shotgun assemblies have proven remarkably successful in reconstructing the bulk of euchromatic genes, with the only limit appearing to be determined by the sequencing depth. For genes imbedded in heterochromatin, however, the low cloning efficiency of repetitive sequences, combined with the computational challenges, demand that additional clues be used to annotate the sequences. One approach that has proven very successful in identifying protein coding genes in Y-linked heterochromatin of *Drosophila melanogaster* has been to make a BLASTable database of the small, unmapped contigs and fragments leftover at the end of a shotgun assembly, and to attempt to capture these by blasting with an appropriate query sequence. This approach often yields a staggered alignment of contigs from the unmapped set to the query sequence, as though the disjoint contigs represent small portions of the gene. Further inspection frequently shows that the contigs are broken by very large, heterochromatic introns. Methods of this sort are being expanded to make best use of all available clues to determine which unmapped contigs are associated with genes. These include use of EST libraries, and, in the case of the Y chromosome, testing of male specific genes and reduced shotgun depth of relevant contigs. It appears much more hopeful than anyone would have imagined that whole genome shotgun assemblies can recover the great bulk of even heterochromatic genes.

Introduction

The study of the heterochromatic regions of a genome is a major challenge, because their properties confound the tools used to study the euchromatin, even in favorable model organisms such as *Drosophila melanogaster*. Heterochromatic regions have very low levels of meiotic recombination, which hampers formal genetic mapping. Much of the heterochromatin undergoes little or no polytenization, making cytogenetic studies more difficult. *P*-element mutagenesis is also challenging, because marker genes are often silenced when inserted into heterochromatin. Finally, heterochromatin is packed with repetitive DNA, which creates problems at virtually every step in genomic sequencing projects, including cloning, assembly and annotation. Progress in the analysis of heterochromatin relies heavily on the development of special tools to circumvent these problems. Deletion analysis, particularly with the use of translocation stocks (e.g., Kennison, 1981), has allowed gene mapping, and special banding techniques have allowed the division of the *Drosophila* heterochromatin into 61 bands (Gatti & Pimpinelli, 1983). Improved *P*-constructs can escape silencing (Zhang & Stankiewicz, 1998). Genome sequencing projects have primarily targeted euchromatic regions (Myers et al., 2000; Lander et al., 2001), and clearly there is room for improvements, with development of methods tailored for assembly and annotation of the heterochromatic portions of genomes. In this paper we review these methods and propose new ones. We concentrate on the whole genome shotgun (WGS) strategy, used to produce the bulk of the *D. melanogaster* genome sequence (Adams et al., 2000; Myers et al., 2000).

The technical difficulties posed by heterochromatin and its relatively low gene density have led to a focusing of genome sequencing projects on euchromatin. As Myers et al. (2000) put it, "...the implicit goal of all sequencing efforts (...) is to determine the sequence of the euchromatic segments of the genome." As sensible as this choice is, heterochromatic sequences are directly involved in essential biological functions, including centromeres, telomeres, protein-coding loci, and rDNA loci in many eukaryotes. Y chromosomes are heterochromatic as well, in *Drosophila* and in mammals. Interesting evolutionary phenomena occur in these regions: the evolution of centromeres and Y chromosomes is unusual (Lahn & Page, 1997; Carvalho et al., 2001; Henikoff, Ahmad & Malik, 2001), apparently being driven by meiotic drive and accumulation of male-related genes, respectively, and rDNA loci exhibit concerted evolution (Tartof, 1979; Coen & Dover, 1983). Furthermore, proximity to heterochromatin influences gene regulation (reviewed in Weiler & Wakimoto, 1995). Finally, repetitive DNA plays a key role in assessing the relative merits of alternative strategies for genome sequencing, including 'clone-by-clone' (CBC) and WGS. For repeat-poor genomes, such as those of most microorganisms, WGS is the standard method, and the skepticism toward its application to the human genome was in part driven by the fear that WGS would fail for the repeat-rich genomes of complex eukaryotes (Green, 1997). Thus, sequence assembly of heterochromatic regions may be used as a benchmark, particularly for software improvements in WGS. We argue that the *Drosophila* genome is a remarkably informative experimental model for testing methods for assembling heterochromatic sequences from WGS data.

The fate of heterochromatic sequences in genome projects

Current sequencing technology produces sequence reads of 1000 bases, at most. The average sequence read length in the *Drosophila* Genome Project (DGP) was 550 bases (Myers et al., 2000). Hence, the sequence of a genome can only be obtained by assembly of a large collection of these short reads. Two main strategies are used to accomplish this. In the CBC approach, a library of large-insert clones (e.g., BAC clones with ∼200 kb inserts) is first used to construct a genome physical map. A subset of large-insert clones covering the genome in a 'tiling path' is then selected and sequenced. Thus, assembly is divided into two phases, assembly of BAC clones into the full genome, and assembly of reads into BAC clones. In WGS projects, genomic DNA is directly cloned into sequencing vectors with small inserts. In the DGP, paired end sequences, called 'mate-pairs', were collected from genomic clones with 2 and 10 kb inserts. The 'mate-pairs' convey three pieces of information – two sequence reads and their physical separation – and are essential for the method because they 'bridge' most repetitive elements (Weber & Myers, 1997), which in *Drosophila* are usually shorter than 10 kb. The DGP also used mate-pairs from large insert BACs, which provide long-range linking information. Note that WGS reads are essentially random; full coverage of the genome is achieved by sequencing enough clones to ensure that each base is on average sequenced multiple times (5 × or 8 × in the human genome, and ∼13 × in DGP). Multiple reads are also necessary to produce high quality sequence. Intensive computation is then used to assemble the large collection of reads into contigs and scaffolds. A sequence scaffold is a set of contigs that are ordered and oriented with respect to each other by 'mate-pair' links, and for which the gaps are of known size. Finally, the sequence scaffolds are mapped on the chromosomes using genetic and/or cytological data.

Both methods face similar difficulties with heterochromatic regions. First, many heterochromatic sequences are not stably maintained in cloning vectors. For example, although they comprise 70–80% of the heterochromatin (and ∼25% of the genome), heterochromatic simple sequence repeats account for only

2% of the WGS sequence reads in the DGP (Lohe, Hilliker & Roberts, 1993; Adams et al., 2000). This type of heterochromatic sequence seems to represent the worst case. Regions of single-copy sequences intermixed with transposable element sequences are much more stable in cloning vectors. However, the limits of 'clonability' are not clear: recent results show that human peri-centromeric sequences containing up to 98% repetitive DNA (and including a 50 kbp block of gamma satellite and the 171 bp alpha satellite arrays) can be stably propagated in BAC clones, which were supposed to be quite restrictive due to their size (Schueler et al., 2001).

The second problem caused by heterochromatic sequences occurs in assembly. Like a puzzle with many identical or nearly identical pieces, the clones (BAC clones in CBC; individual reads in WGS) are hard to place into the correct order. BAC clones with a high content of repetitive DNA are difficult to place correctly in the physical map (but see Tilford et al., 2001), and thus were not sequenced in the CBC-based Human Genome Project (Lander et al., 2001). On the other hand, WGS is random in its coverage, and cannot avoid heterochromatin-derived sequences. Depending on their nature, WGS assembly places these reads in two different bins. Islands of unique sequence are correctly assembled into short contigs and scaffolds, but cannot be connected to other contigs to form large scaffolds because they are isolated from the euchromatin by 'unbridgeable' sequence gaps (say, a huge block of satellite DNA). This set of assembled sequences was called 'armU' [arm unmapped] in the Celera/BDGP CD-ROM release of the *Drosophila* genome, and was submitted to GenBank as 704 individual scaffolds. More difficult sequences (satellite DNA, low quality reads, contaminations, etc.) that fail to assemble into contigs are placed in the 'chaff', which is a collection of unassembled WGS reads. In both the human and *Drosophila* WGS projects, the chaff accounts for 25% of the reads (Venter et al., 2001; unpublished) and, as we will see later, contains fragments of heterochromatic genes.

WGS randomly samples the entire genome. This is perhaps its main advantage for those with a 'heterochromatin-centric' view of the world: a 10 kb gene embedded in the middle of, say, 3 Mbp of ATATAT satellite sequence is bound to be detected by WGS. This simple observation propelled us to search in the armU sequence for fragments that belong to our favorite heterochromatic region, the *Drosophila* Y chromosome.

Detecting Y-chromosome genes

In this section, we briefly review the detection of Y-linked genes using WGS sequences (Carvalho, Lazzaro & Clark, 2000; Carvalho et al., 2001). More details and a short historical account can be found in the original papers. Formal genetic analysis has established that the *Drosophila* Y chromosome contains six genes essential for male-fertility (*kl-1*, *kl-2*, *kl-3*, *kl-5*, *ks-1* and *ks-2*; Kennison, 1981). Another key discovery is that some of the Y-linked fertility genes, including *kl-5*, are unusually large (∼3 Mbp), and, at least in *D. hydei*, contain gigantic heterochromatic introns (>1 Mbp; Gatti & Pimpinelli, 1983; Reugels et al., 2000). Gepner and Hays (1993) showed that *kl-5* encodes a dynein heavy chain that is expressed in the testis. Nothing was known about the molecular nature of the remaining genes, although there was strong evidence that *kl-2* and *kl-3* also encode dynein heavy chains (Hardy, Tokuyasu & Lindsley, 1981; Goldstein, Hardy & Lindsley, 1982).

The *Drosophila* WGS project assembled essentially all 120 Mbp of the euchromatic portion of the genome, but only 15 kb (a small portion of the *kl-5* gene) of the Y chromosome was identified. Given the randomness of WGS, the remaining single-copy sequences of the Y (including the fertility genes) should have been at least partly recovered. A BLASTN search using *kl-5* cDNA as the query confirmed this: in addition to the fragment already identified, it retrieved most of the *kl-5* gene, scattered across five scaffolds in armU (Figure 1). These scaffolds contain complete exons, intact 3′ and 5′ splice junctions, and a variable portion of upstream and downstream intronic sequences. The explanation is in Figure 2: exons and normally sized introns are unique sequences that are cloned and assembled by WGS to produce isolated scaffolds, whereas large heterochromatic introns fail to be assembled. Genes like *kl-5* are chopped into several small, unlinked scaffolds, separated by 'difficult' sequences such as satellite DNA. Such genes will often be missed by gene annotation procedures, because these methods work poorly with individual exons. However, the sequence scaffolds are present in the *Drosophila* WGS, and can be detected using the appropriate query sequences and tools such as BLAST. For known genes, we can use cDNA as a query in a BLASTN search, and for unknown genes, a related protein and TBLASTN search. In both cases, the BLAST hits have a characteristic staggered pattern, resulting from the fragmentation of the gene

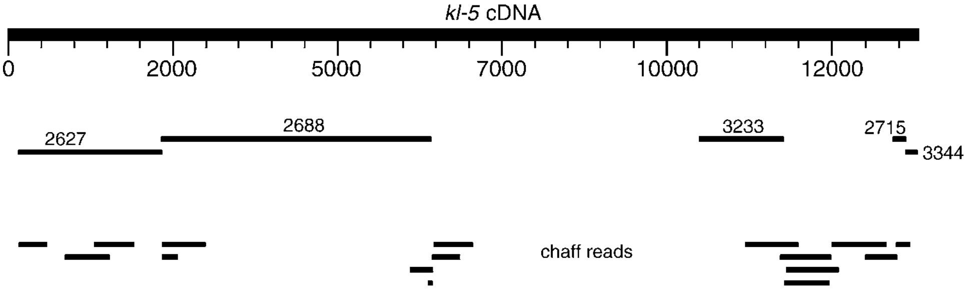

Figure 1. Results of a BLASTN search using the *D. melanogaster kl-5* cDNA sequence (AF210453) as a query against armU (upper) or against a database of chaff reads (bottom; see the *Chaff* section). Note the staggered appearance of the matches to armU. Numbers above the bars are the abridged accession numbers of the armU scaffolds (AE002627 was abridged to '2627' and so on). Unlabeled bars belong to the chaff. Note that some of the gaps in armU were covered by chaff reads.

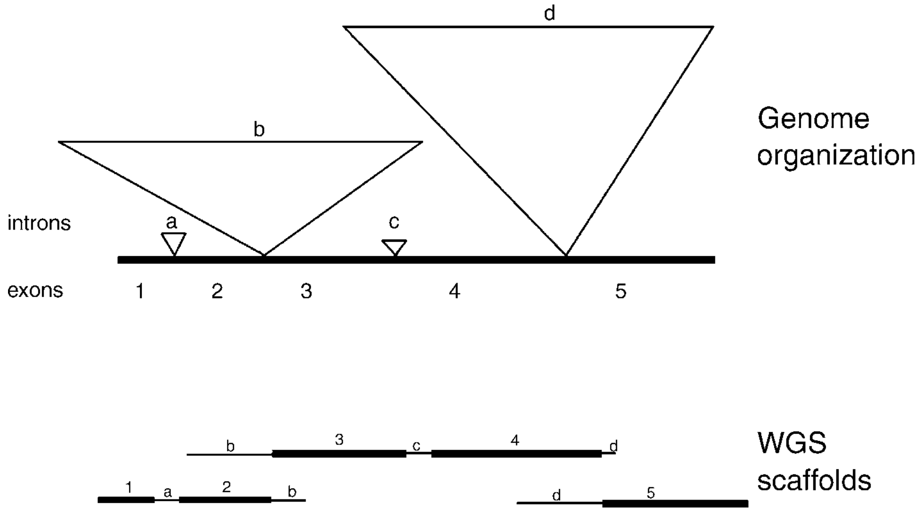

Figure 2. Staggered patterns of matches. Upper: genomic organization of a heterochromatic gene. Exons are represented by numbers and introns by letters. Bottom: the same region, as it appears in a BLAST search against a WGS database. Large introns are only partially assembled and cause a staggered pattern of matches. Modified from Charlesworth (2001), with permission.

into isolated sequence scaffolds (Figures 1 and 2). A TBLASTN search of each of the roughly 500,000 protein sequences in the NCBI nr database against a database of the armU *Drosophila* scaffolds yielded some 20 candidate genes with a staggered pattern of matches, of which seven proved to be Y-linked. One additional Y-linked gene was detected using testis ESTs as the query. Thus, WGS allowed the molecular identification of eight Y-linked genes (only *kl-5* was known previously). Two of the genes encode dynein heavy chains and correspond to the long sought fertility factors *kl-2* and *kl-3* (Hardy, Tokuyasu & Lindsley, 1981; Goldstein, Hardy & Lindsley, 1982), whereas some of them do not correspond to any of the fertility factors, and are not essential for male fertility, although they may encode necessary but redundant functions.

The results summarized above are interesting for several reasons. The closest paralogs of nearly all identified Y-linked genes are autosomal and not X-linked. This pattern strongly suggests that Y-linked genes were acquired from the autosomes, rather than being present on the hypothetical chromosome pair that gave rise to the X–Y pair. This supports the hypothesis that much of the *Drosophila* Y is a new construct rather than a degenerate X (Lohe, Hilliker & Roberts, 1993; Hackstein et al., 1996). The *Drosophila* Y chromosome has a functional specialization unheard of in eukaryotic chromosomes: among the nine known single copy genes, three are sperm-specific dynein heavy chains (*kl-2*, *kl-3* and *kl-5*), and three are protein phosphatases (*Pp1-Y1*, *Pp1-Y2* and *PPr-Y*). Most of these genes have male-related functions (e.g., beating of flagella). A possible explanation for these patterns was proposed by Fisher (1931), who noted that the Y chromosome is expected to accumulate male-related genes because male–female antagonistic effects of genes may hamper the evolution of male-related traits, unless they are located in a male-specific region of the genome. Thus, the identification of Y-linked genes directly addressed questions such as the origin and evolution of the Y, and opens the possibility of large-scale polymorphism studies on the *Drosophila* Y chromosome (Zurovcova & Eanes, 2000; Charlesworth, 2001; A.G. Clark & A.B. Carvalho, unpublished). Furthermore, these studies also provide methods for detection of heterochromatic genes on the other chromosomes of *Drosophila* (see below), and in all WGS projects. For example, the WGS assembly of the human genome produced an 'armU' comprising 11,542 scaffolds and 53 Mbp of sequence (Venter et al., 2001). There is no doubt that these sequences contain heterochromatic genes.

Tools and resources

During the search for Y-linked genes, we found many putative heterochromatic genes that are autosomal or X-linked. Some of them have perfect EST matches, and hence may be considered 'bona fide' genes. In this section, we summarize these non-Y genes. We also review and propose some tools and methods to study the *Drosophila* heterochromatin. We concentrate on the bio-informatics tools that use as raw data the WGS assembly (including armU and chaff) and publicly available EST data. What can be gleaned from the already available WGS and EST sequences?

ArmU

'ArmU' is the set of 704 WGS sequence scaffolds that were not mapped to any chromosome arm. As Adams et al.(2000) suggested, most of these scaffolds represent pieces of heterochromatic genes, including Y-linked ones. Subsequently, some 84 scaffolds were identified as contaminants (mostly human and *E. coli* DNA) and were removed from GenBank. Currently, armU contains 620 scaffolds representing ∼4 Mbp of sequence. Some 35 scaffolds have already proved to be Y-linked. ArmU probably contains most of the single-copy sequence in the heterochromatin. Though improved assemblies of the *Drosophila* genome have been generated (see the *armU2* section), armU will remain the reference for some time to come, because it is deposited in GenBank, it has been curated (whereas the improved assemblies have not yet been annotated and are raw output files from the assembly algorithm), and it has already been used for gene prediction and other purposes. The map locations of all *Drosophila* sequence scaffolds, which are useful for downloading armU from NCBI, can be found at ftp://ncbi.nlm.nih.gov/genbank/genomes/D_melanogaster/Chromosomes/scaffold_map.txt.

Putative non-Y heterochromatic genes detected with the TBLASTN method

The TBLASTN method detected 12 non-Y putative genes with a staggered pattern of matches. Six of these were confirmed by EST matches. Thus, as with the Y-linked scaffolds, the candidates detected with the TBLASTN method often correspond to transcribed genes. The paucity of pseudo-genes in *Drosophila* helps this method, but pseudo-genes are problematic for any method of gene detection. The finding that non-Y heterochromatic genes also show staggered patterns of matches in BLAST searches shows that their introns also contain 'difficult' sequences (but see the *armU2* section). On the other hand, some heterochromatic genes are contained in single armU scaffolds (e.g., CCY; Carvalho et al., 2001). We are now testing these 'single-hit' genes for Y-linkage.

Table 1 summarizes the number of putative heterochromatic genes detected with TBLASTN and other methods. These data will be more useful when each putative heterochromatic gene is mapped to a cytological region (2L centric heterochromatin, band h39, etc.). Many heterochromatic genes have been identified by deficiency analysis, but are not known at the molecular level (reviewed in Gatti & Pimpinelli, 1992). For example, there are 10 vital loci in the 3L heterochromatin. Mapping of heterochromatic scaffolds and ESTs will suggest candidates for these genes, which may then be confirmed by other methods. An easy way to map them is to perform PCR with a panel of deficiency strains. In the case of Y-linked genes we used six deletions, each covering a region of the chromosome (Carvalho, Lazzaro & Clark, 2000). Chromosome 4 and other heterochromatic regions may be similarly studied, by extracting DNA from appropriate deficiency-bearing embryos. (Homozygous deletions will be lethal in most of the cases.) Experiments to test this approach are being planned in collaboration with B. Wakimoto.

Table 1. Candidate heterochromatic genes

Method	Identified candidates	Comments
TBLASTN (protein × armU)	19	staggered: 7 Y-linked, 12 non-Y. Single-hits: not yet analyzed
ESTs × armU	43	9 staggered, 34 single-hits. Not yet fully analyzed
Gap ESTs	∼5	2900 ESTs (out 170,000) do not match the published genome. 169 clusters (but many contaminants). Not yet analyzed

Customized databases

For many purposes, it is desirable to run BLAST searches against a customized database. The staggered pattern of 'hits' is much harder to visualize if one uses the whole *Drosophila* genome (as available in NCBI web site) instead of just armU, as the database for searches. Those interested in the distribution of transposable elements (TEs) in heterochromatic sequences will benefit from a small, well-curated database of these elements, instead of being drowned by the thousands of hits produced by the usual BLAST searches. An essential tool for these searches is StandAloneBlast (which does not have a graphical interface, but can process thousands of searches in batch mode) and WWWStandAloneBlast, which requires a Linux operating system to run in PC computers and has a graphical interface (e.g., Figure 1). Both programs can be downloaded at NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/).

Expressed sequence tags (ESTs)

ESTs are partial cDNA sequences representing single reads from randomly selected cDNA clones (Adams et al., 1991). A match between a genomic region and an EST is strong evidence of a gene. Application of the method to heterochromatic regions is straightforward (e.g., the CCY gene was discovered in this way; Carvalho et al., 2001). In addition to gene identification, ESTs provide linking information that can be used to order and orient genomic sequence scaffolds. For example, the five armU scaffolds shown in Figure 1 are separated by Mbp-sized blocks of satellite DNA, and the *kl-5* cDNA gives their order and orientation in the genome (Carvalho, Lazzaro & Clark, 2000; this method of joining unmapped sequence scaffolds using cDNA sequences was independently proposed by G. M. Rubin, A. Spradling and G. Karpen). The BDGP has deposited more than 250,000 EST sequences in GenBank from various sources (larva-pupae, embryo, Schneider cell line, head, testis and ovaries), and Andrews et al. (2000) have deposited some 7300 from testis. A caveat of ESTs is that rare transcripts are missed quite frequently, that is, many true genes are not represented by ESTs. For example, only one of the Y-linked genes, CCY, has a matching EST, but the mRNAs of the remaining eight were all readily detected by RT-PCR. Another nuisance is that many ESTs represent genes in TEs, and these match the genome sequence at multiple locations, especially in the armU sequences. Removing these ESTs with RepeatMasker (A.F.A. Smit & P. Green, unpublished; available at http: //repeatmasker. genome. washington.edu/cgi-bin/ RepeatMasker) or similar procedures, and building a customized database for StandAloneBlast, solves the problem. We found 43 ESTs with matches to armU, nine of them with a staggered pattern of matches, and 34 with a single match. These figures do not include TE-derived ESTs, which are much more numerous. Most of the nine ESTs with staggered matches overlap with TBLASTN-predicted genes (Table 1).

Heterochromatic genes may also be identified by ESTs that *do not* match the published *Drosophila* genome sequence ('gap ESTs'). The rational is that heterochromatic sequences have a much higher chance of being lost during assembly (e.g., the numerous gaps in Y-linked genes; Carvalho, Lazzaro & Clark, 2000; Carvalho et al., 2001). We found 2900 such ESTs; after clustering the ones encoded by the same gene we obtained 169 contigs, which we are still analyzing. As expected, many represent *Drosophila* mitochondrial genes, histone genes (the later filtered out prior to WGS assembly; Adams et al., 2000), and contaminants from various sources (*E. coli*, *Saccharomyces*,

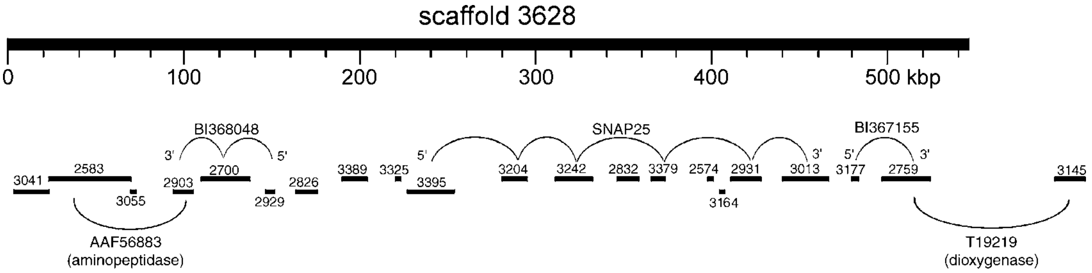

Figure 3. An unmapped scaffold from the new WGS assembly. The 546 kbp sequence spans 20 armU scaffolds, and contains at least three genes: SNAP25, an aminopeptidase similar to AAF56883, and a dioxygenase similar to T19219. The last two genes were detected with the TBLASTN method and confirmed with ESTs BI368048 and BI367155, respectively. Complete sequencing of the corresponding cDNAs may link the 546 kbp scaffold to other scaffolds (the sequence of both ESTs do not include the 3′ end). TBLASTN suggests that AE002759 is linked to AE003145. The latter is included in another armU2 scaffold. Numbers above the bars are the abridged accession numbers of the armU scaffolds (AE003041 is abridged to '3041' and so on).

human). However, the method works: we found some known heterochromatic genes that were not included in the Release 2.0 Drosophila WGS sequence (e.g., EST BG638729 is part of the plexin B gene, located on chromosome 4), and several new ones. Euchromatic genes localized in sequence gaps of the genome were also detected. For example, EST BF485746 and others encode a gluthatione transferase that was absent from the published genome sequence. PCR readily amplifies this gene, which confirms that it is present in the *Drosophila* genome. The sequence is present in a BAC clone that maps to 16F (accession AC012097), and in a new WGS assembly (see below) it also maps to the same region. Thus, this gluthatione transferase falls into a euchromatic gap in the Release 2.0 sequence that has not yet been filled. A curious gene that most likely *does not* belong to *Drosophila* is represented by 437 ESTs (e.g., BG639625), all from Schneider cells (a permanent cell line). The encoded protein is viral in origin, and is almost certainly derived from the transforming virus used to establish this immortalized cell line.

A full list of heterochromatic ESTs is being prepared in collaboration with the BDGP and will be available at http://www.fruitfly.org/. The ESTs can be downloaded from the same site.

The unmapped sequence scaffolds of a new WGS assembly (armU2)

Drosophila was the first complex eukaryote sequenced by WGS, and since then Celera has continued to improve its assembly algorithms. The first assembly was from February 2000, and is the basis of Releases 1 and 2 of the Drosophila Genome (Adams et al., 2000). Two new assemblies have been done, in March 2001 (referred to in this section) and in November 2001 (not examined yet). These assemblies have not yet been annotated or curated. In the March 2001 version, the assembly of heterochromatic regions was dramatically enhanced due to improvements in the algorithms. Preliminary analysis shows that the new armU ('armU2') represents ~17 Mbp of genomic sequence, distributed in 2183 scaffolds. Thus, many reads that were not assembled in Release 2.0 are now incorporated into scaffolds (see the *Chaff* section). Many ESTs, cDNAs and proteins that produce a staggered pattern of matches to armU of Release 2.0 are contained in single armU2 scaffolds. Thus, the improved algorithms assembled large introns composed of repetitive sequences. Figure 3 shows an example: the armU2 scaffold 3628 spans 546 kbp of sequence and joins 20 armU scaffolds! Scaffold 3628 is colinear with the three genes present in this region (i.e., no errors in the order and orientation of exons, and no missing exons), which strongly suggests that it is an essentially correct assembly of the region. Many things can be learned from armU2. Genes located in low recombination regions of the *Drosophila* genome tend to have larger introns (Carvalho & Clark, 1999; Comeron & Kreitman, 2000), and full sequence of these regions may shed light on this controversial subject. The SNAP25 gene was broken into six scaffolds in the first release of the Drosophila genome, and now we know that this is due to its unusually large introns, which are packed with TEs, and small exons, a combination that leads to very small islands of unique sequences in a sea of repetitive elements (Table 2). The SNAP25 gene spans over 220 kb. Its reported location is euchromatic (polytene bands 64E11 to 65B5), but probably is wrong, since the gene is removed by a deletion that maps to region 80, in the 3L

Table 2. Genomic structure of the SNAP25 gene. Adjacent exons separated by smaller introns (exons 1–2 and 6–7) were assembled in the same armU scaffold, whereas larger introns (>30 kbp) disrupted the assembly

Exon #	Scaffold	Size (bp)	Size of downstream intron (kbp)
1	AE003395	81	5
2	AE003395	93	52
3	AE003204	42	30
4	AE003242	48	45
5	AE003379	118	42
6	AE002931	129	4
7	AE002931	145	39
8	AE003013	1426	–

centromeric region (deletion *DE(3L)1–16*; Rao et al., 2001). Another interesting case is scaffold 3285 of the new assembly (100 kbp). It spans five armU scaffolds and the complete *rolled* gene, which is very close to the centromere of chromosome 2. This 'mapping' of armU Release 2.0 scaffolds into scaffolds of the new assembly is useful because it reduces the number of scaffolds that need to be localized experimentally, with PCR on deleted lines or FISH to metaphase chromosomes. For example, we tested two armU scaffolds shown in Figure 3 for Y-linkage; armU2 would have spared us this effort, since SNAP25 is known to be autosomal. Ironically the assembly of Y-linked sequences was worse in armU2: the scaffolds are the same size or shorter than in armU. This must be related to the low WGS depth of the Y chromosome (see the *Y-depth* section). Most of the autosomal or X-linked introns that caused a staggered pattern of matches in armU were in single scaffolds in armU2. Although these introns are large (some are 60 kbp long), they still do not seem to contain satellite DNA, and could be assembled with improved WGS computational methods. Thus, it is possible that gigantic, intractable introns are restricted to Y-linked genes. We are presently mapping all armU scaffolds on the new WGS assemblies and investigating the assembly of the Y chromosome.

Chaff

There are 808,042 unassembled sequence reads in the *Drosophila* WGS assembly used to produce the Release 2.0 genome sequence, totaling 650 Mbp of redundant shotgun sequence data. These data are an invaluable resource for studying repetitive sequences such as satellite DNA, which end up there. Figure 1 shows that the chaff also contains pieces of heterochromatic protein-coding genes not covered by armU, which make it even more useful. For example, the Y-linked ORY gene is derived from the autosomal gene CG6059 by duplication, but its N-terminal exon seems to be missing in armU (Carvalho et al., 2001). Alternatively, ORY may be shorter, or its N-terminus too divergent to be detected by similarity with CG6059. A TBLASTN search using the N-terminus of CG6059 as the query and the chaff as the database quickly identified five candidate reads. Thus, it appears that the N-terminus of ORY is not too divergent from CG6059, and was missing in armU. Although 5′ RACE might have produced the same information, it will be much easier to do a simple RT-PCR experiment to confirm that the five reads encode the N-terminus of ORY. Many other missing exons of Y-linked genes that we have sequenced *de novo* are also represented in the chaff; it would have spared us a lot of effort. Another use of the chaff is exemplified by the gluthatione transferase gene mentioned above, which lies entirely within this data set (not shown). Thus, the chaff may be used to help distinguish *Drosophila* ESTs from contaminants (the later will be absent from the chaff). However, note that pieces of Y-linked genes are missing even from the chaff (see the next section). It is also informative to examine the chaff before concluding that a given sequence does not exist in the *Drosophila* genome. Like the new WGS assembly (which includes armU2), the chaff is not publicly available data.

Y-depth

The sequence data in the *Drosophila* Genome Sequencing Project has been generated from genomic DNA isolated from unsexed embryos. Thus, the sequence reads sample the Y, X and autosomes in a 1:3:4 ratio. The genome was sequenced to an average shotgun depth of 12.8 ×, so Y chromosome sequences should be represented by ~3.2 × shotgun sequence. A symptom of this low depth is the relatively high frequency of sequencing errors (frame-shifts) we have identified in Y-linked scaffolds (Carvalho, Lazzaro & Clark, 2000). Also, low depth (besides gigantic introns) probably contributes to the scattering of exons of Y-linked genes in different scaffolds. In principle, the Y-linked scaffolds should be identifiable by their lower depth. Figure 4 shows that this simple idea works: the distributions of shotgun sequence depth for Y and non-Y

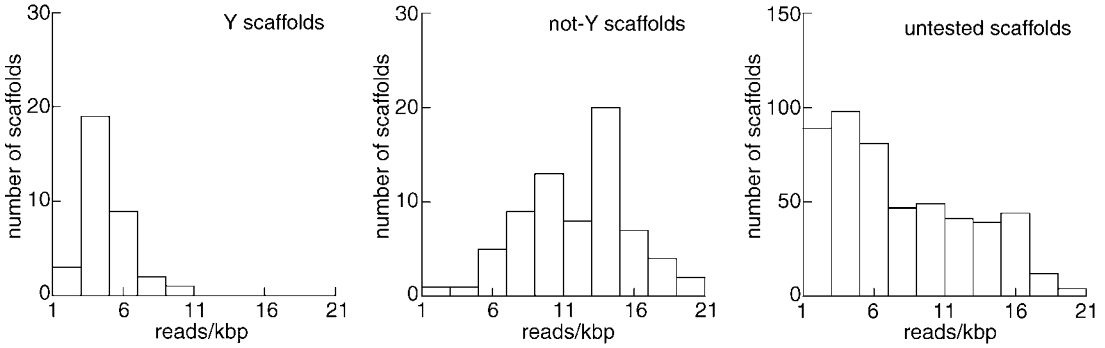

Figure 4. Distribution of shotgun depth (reads/kbp of sequence) for Y, 'not-Y' (i.e., X or autosomal), and untested scaffolds. The averages are 4.7, 12.0 and 7.9 reads/kbp, respectively. The number of reads used to build each scaffold was obtained from Celera, and the length of scaffolds was obtained from the GenBank FASTA files.

(i.e., autosomal or X-linked) sequence scaffolds are indeed very different. The small overlap in the distributions will probably be reduced when we take into account additional data such as the length of each individual sequence read. It may even be possible to identify candidate X-linked scaffolds. This beautiful aspect of WGS is very useful. In our previous experience, one in four armU scaffolds tested by genomic PCR proved to be Y-linked; the success rate is expected to be higher when we take shotgun sequence depth into account. This approach may allow the identification of essentially all genes from the *Drosophila* Y chromosome. The average shotgun depth of armU scaffolds that we have not yet tested by genomic PCR (7.9 reads/kbp) is a weighted average of Y (4.7) and non-Y scaffolds (12.0). A simple calculation suggests that 56% of the untested scaffolds should be Y-linked! The utility of this approach is not restricted to *Drosophila*, since humans and many other organisms of interest have Y chromosomes. Finally, examining the distribution of shotgun depth across scaffolds may help to identify misassemblies; correctly assembled scaffolds should have a relatively constant level of shotgun depth.

Perspectives

A wealth of sequence data is available to study the *Drosophila* heterochromatin. This is particularly true for gene identification, because the WGS approach appears to retrieve nearly all non-repetitive sequences, and the *Drosophila* genome was sequenced to very high depth (~13×). WGS allowed rapid progress in the identification of Y-linked genes, and the same prospect holds for other heterochromatic regions as well. Our main goal is to identify 'all' genes from the *Drosophila* Y, using the sequence data and methods described in this paper. This is particularly exciting because the gene content of the Y is functionally coherent, and gene identification reveals the forces at play in the evolution of the Y chromosome (Carvalho, Lazzaro & Clark, 2000; Carvalho et al., 2001).

There are important questions about heterochromatic sequences beyond gene identification, and in this regard it is not clear yet how far WGS assembly can go. Another new assembly was very recently produced (in November 2001), and it will be exciting to compare it with the previous ones. For the most challenging genomic regions, WGS assemblies and the chaff may be used at least as starting points for more detailed studies using BAC clones, FISH, *P*-element insertion, and other methods. Regions composed of very large blocks of satellite DNA are likely to remain intractable, and may never be fully sequenced.

At present, there are 620 armU scaffolds in GenBank (Drosophila Genome Release 2.0), and we have mapped 35 of them to the Y chromosome. Long-range linking information from genes (staggered hits in BLAST searches using cDNA, ESTs, and proteins as queries) as well as from new WGS assemblies effectively links these scattered scaffolds, putting them in the correct order and orientation (e.g., Figures 1 and 3). Preliminary analysis of this linking information consolidated the 620 armU scaffolds into ~330 larger ones, which can be mapped into heterochromatic regions with PCR and/or FISH. BDGP is doing further analysis that will lead to more complete linkage information and will display annotated versions to the public.

Repetitive DNA poses the most daunting challenge to genome sequencing methods and strategies, but the limits are not yet clear. Schueler et al. (2001) assembled with BACs 450 kbp from the end of human Xp euchromatin to the edge of the centromeric satellite repeats, a region containing 98% repetitive DNA and satellite arrays. Thus BACs can accommodate heterochromatic sequences, and can be used to map, sequence and assemble heterochromatic sequences. It is equally surprising that 265 kbp of this region of the human genome has been correctly assembled by Celera into a single scaffold. Most likely this was done with 5 × WGS depth because the public human sequence (which was used to increase Celera's 5 × depth to 8 ×) does not cover this region (Venter et al., 2001; Schueler et al., 2001). Thus, it is clear that both methods (CBC and WGS) are more powerful than initially imagined, and that greater effort is bound to yield further advances. While it is true that the repeat structure of every genome is different with its own challenges, the principles of WGS are the same, and the algorithm that was improved for the human genome also produced an improved assembly of *Drosophila* (see the *armU2* section). No other complex eukaryote has been shotgun sequenced to the high depth employed for *Drosophila*, so the strengths and limitations of WGS may be evaluated by how deep it can go inside the *Drosophila* heterochromatin. Given this, the large store of genetic information on the species, and its reasonable genome size, *Drosophila* is an ideal benchmark for WGS software improvements. It will not be the last time that our favorite experimental model will be called to tackle the hardest questions.

Acknowledgments

We thank S. Shavirin and D. Vakatov (NCBI) for help during the set-up of WWWStandAloneBlast and compilation of StandAloneBlast. We also thank M. Alvim and B. Ribeiro for the computer programs used to estimate depth and to map armU into armU2, and B. Charlesworth for permission of use of Figure 2. This work was supported by fellowships from the Pew Latin American Fellows Program and FAPERJ (A.B.C.), CNPq (M.D.V.), and by grants from the US National Institutes of Health and National Science Foundation (to A.G.C.).

References

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno et al., 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252: 1651–1656.

Adams, M.D., S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle et al., 2000. The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.

Andrews, J., G.G. Bouffard, C. Cheadle, J. Lu, K.G. Becker & B. Oliver, 2000. Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. Genome Res. 10: 2030–2043.

Carvalho, A.B. & A.G. Clark, 1999. Intron size and natural selection. Nature 401: 344.

Carvalho, A.B., B.P. Lazzaro & A.G. Clark, 2000. Y chromosomal fertility factors *kl-2* and *kl-3* of *Drosophila melanogaster* encode dynein heavy chain polypeptides. Proc. Natl. Acad. Sci. USA 97: 13239–13244.

Carvalho, A.B., B.A. Dobo, M.D. Vibranovski & A.G. Clark, 2001. Identification of five new genes on the Y chromosome of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 98: 13225–13230.

Charlesworth, B., 2001. Genome analysis: more *Drosophila* Y chromosome genes. Curr. Biol. 11: R182–R184.

Coen, E.S. & G.A. Dover, 1983. Unequal exchanges and the coevolution of X and Y rDNA arrays in *Drosophila melanogaster*. Cell 33: 849–855.

Comeron, J.M. & M. Kreitman, 2000. The correlation between intron length and recombination in *Drosophila*: Dynamic equilibrium between mutational and selective forces. Genetics 156: 1175–1190.

Fisher, R.A., 1931. The evolution of dominance. Biol. Rev. 6: 345–368.

Gatti, M. & S. Pimpinelli, 1983. Cytological and genetic-analysis of the Y-chromosome of *Drosophila melanogaster*. 1. Organization of the fertility factors. Chromosoma 88: 349–373.

Gatti, M. & S. Pimpinelli, 1992. Functional elements in *Drosophila melanogaster* heterochromation. Annu. Rev. Genet. 26: 239–275.

Gepner, J. & T.S. Hays, 1993. A fertility region on the Y chromosome of *Drosophila melanogaster* encodes a dynein microtubule motor. Proc. Natl. Acad. Sci. USA 90: 11132–11136.

Goldstein, L.S., R.W. Hardy & D.L. Lindsley, 1982. Structural genes on the Y chromosome of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 79: 7405–7409.

Green, P., 1997. Against a whole-genome shotgun. Genome Res. 7: 410–417.

Hackstein, J.H., R. Hochstenbach, E. Hauschteck-Jungen & L.W. Beukeboom, 1996. Is the Y chromosome of Drosophila an evolved supernumerary chromosome? Bioessays 18: 317–323.

Hardy, R.W., K.T. Tokuyasu & D.L. Lindsley, 1981. Analysis of spermatogenesis in *Drosophila melanogaster* bearing deletions for Y-chromosome fertility genes. Chromosoma 83: 593–617.

Henikoff, S., K. Ahmad & H.S. Malik, 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293: 1098–1102.

Kennison, J.A., 1981. The genetic and cytological organization of the Y chromosome of *Drosophila melanogaster*. Genetics 98: 529–548.

Lahn, B.T. & D.C. Page, 1997. Functional coherence of the human Y chromosome. Science 278: 675–680.

Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh et al., 2001. Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Lohe, A.R., A.J. Hilliker & P.A. Roberts, 1993. Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. Genetics 134: 1149–1174.

Myers, E.W., G.G. Sutton, A.L. Delcher, I.M. Dew, D.P. Fasulo, M.J. Flanigan, S.A. Kravitz, C.M. Mobarry, K.H. Reinert, K.A. Remington et al., 2000. A whole-genome assembly of Drosophila. Science 287: 2196–2204.

Rao, S.S., B.A. Stewart, P.K. Rivlin, I. Vilinsky, B.O. Watson, C. Lang, G. Boulianne, M.M. Salpeter & D.L. Deitcher, 2001. Two distinct effects on neurotransmission in a temperature-sensitive SNAP-25 mutant. EMBO J. 20: 6761–6771.

Reugels, A.M., R. Kurek, U. Lammermann & H. Bunemann, 2000. Mega-introns in the dynein gene *DhDhc7*(Y) on the heterochromatic Y chromosome give rise to the giant threads loops in primary spermatocytes of *Drosophila hydei.* Genetics 154: 759–769.

Schueler, M.G., A.W. Higgins, M.K. Rudd, K. Gustashaw & H.F. Willard, 2001. Genomic and genetic definition of a functional human centromere. Science 294: 109–115.

Tartof, K.D., 1979. Evolution of transcribed and spacer sequences in the ribosomal RNA genes of *Drosophila*. Cell 17: 607–614.

Tilford, C.A., T. Kuroda-Kawaguchi, H. Skaletsky, S. Rozen, L.G. Brown, M. Rosenberg, J.D. McPherson, K. Wylie, M. Sekhon, T.A. Kucaba et al., 2001. A physical map of the human Y chromosome. Nature 409: 943–945.

Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt et al., 2001. The sequence of the human genome. Science 291: 1304–1351.

Weber, J.L. & E.W. Myers, 1997. Human whole-genome shotgun sequencing. Genome Res. 7: 401–409.

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

Zhang, P. & R.L. Stankiewicz, 1998. Y-linked male sterile mutations induced by P element in *Drosophila melanogaster*. Genetics 150: 735–744.

Zurovcova, M. & W.F. Eanes, 1999. Lack of nucleotide polymorphism in the Y-linked sperm flagellar dynein gene *Dhc-Yh3* of *Drosophila melanogaster* and *D. simulans*. Genetics 153: 1709–1715.