FORUM

A Case for Sequencing the Genome of *Musca domestica* (Diptera: Muscidae)

J. G. SCOTT,¹ N. LIU,² M. KRISTENSEN,³ AND A. G. CLARK⁴


ABSTRACT  House flies are carriers of >100 devastating diseases that have severe consequences for human and animal health. Despite the fact that it is a passive vector, a key bottleneck to progress in controlling the human diseases transmitted by house flies is lack of knowledge of the basic molecular biology of this species. Sequencing of the house fly genome will provide important inroads to the discovery of novel target sites for house fly control, understanding of the house fly immune response, rapid elucidation of insecticide resistance genes, and understanding of numerous aspects of the basic biology of this insect pest. The ability of the house fly to prosper in a remarkably septic environment motivates analysis of its innate immune system. Its polymorphic sex determination system, with male-determining factors on either the autosomes or the Y chromosome, is ripe for a genomic analysis. Sequencing of the house fly genome would allow the first opportunity to study the interactions between a pest insect and its parasitoid (*Nasonia vitripennis*) at the whole genome level. In addition, the house fly is well placed phylogenetically to leverage analysis of the multiple Dipteran genomes that have been sequenced (including several mosquito and *Drosophila* species). The community of researchers investigating *Musca domestica* are well prepared and highly motivated to apply genomic analyses to their widely varied research programs.

KEY WORDS  genome sequencing, insect immunity, insecticide resistance, sex determination, comparative genomics

House Flies

Biology and Importance. House flies, *Musca domestica* L. (Diptera: Muscidae), are cosmopolitan, ubiquitous, and transmit >100 human and animal diseases (Scott and Lettig 1962, Greenberg 1965, Keiding 1986), including bacterial infections such as salmonellosis, anthrax, ophthalmia, shigellosis, typhoid fever, tuberculosis, cholera, and infantile diarrhea; protozoan infections such as amebic dysentery; helminthic infections such as pinworms, roundworms, hookworms, and tapeworms; and viral and rickettsial infections. Recent house flies were shown to spread a deadly strain of *Escherichia coli* in Japan (Sasaki et al. 2000). Flies also transmit pathogens responsible for eye diseases such as trachoma and epidemic conjunctivitis and infect wounds or skin with diseases such as cutaneous diphtheria, mycoses, yaws, and leprosy (Keiding 1986). Fly-transmitted trachoma alone causes six million cases of childhood blindness each year (World Health Organization 2004). Considering that house flies are highly mobile, come into contact with excreta, carcasses, garbage, and other septic matter, and that they are intimately associated with humans, our food, and utensils, it is not surprising that they are involved in transmission of so many serious and widespread diseases (Scott and Lettig 1962, Keiding 1986). Most recently, house flies have been shown to transmit life-threatening antibiotic-resistant bacteria (Rahuma et al. 2005, Macovei and Zurek 2006), which are an ever increasing problem in hospitals and other health care facilities (Sundin 1996, Graczyk et al. 2001, Maisnier-Patin and Andersson 2004, Boulesteix et al. 2005).

House flies are always found in association with humans and human activities. In fact, house flies and humans have evolved together, with house flies following the spread of *Homo sapiens* across the planet (Mündi 1994). House flies are also one of the most serious pests at dairy, horse, hog, sheep, and poultry facilities worldwide. Exposure to debilitating disease-causing agents, public health and nuisance concerns, lowered levels of milk and egg production, and reduced feed conversion all result from house fly activity. Economic losses and the cost associated with fly suppression are difficult to quantify, but costs of pesticides for fly control at poultry facilities alone are estimated at more than $200 million annually in the United States (Geden et al. 1994). With the loss of available insecticides (caused by governmental can-
cellations, the high cost of reregistration in the absence of patent protection, and the development of insecticide resistance), there is a pressing need for the development of new insecticides and control strategies.

Many species of arthropods are the sources of potent allergens that sensitize and induce IgE-mediated allergic reactions in humans. Most of these arthropod allergens are proteins, and the allergic response mechanism is the same as those from other sources such as plant pollens, molds, and foods. Allergies to house flies are rare, but cases of respiratory allergy from occupational exposure (farmers) have been reported (Wahl and Fraedrich 1997, Focke et al. 2003). Identification of house fly allergens could lead to recombinant allergens with a potential use in diagnosis and immunotherapy.

Given the importance of house flies in the transmission of human and animal diseases, there has been substantial effort to control fly populations, primarily with insecticides. Generally, house fly control has involved DDT and methoxychlor, as well as other chlorinated hydrocarbons (e.g., lindane, and chlordane), organophosphates (e.g., malathion, diazinon, and dimethoate), carbamates (e.g., methomyl), pyrethrins (usually with piperonyl butoxide), pyrethroids (e.g., permethrin, fenvalerate, and cyfluthrin), and most recently spinosad (limited use) and neonicotinoid baits (e.g., imidacloprid). House flies have shown a remarkable ability to rapidly evolve resistance to each of the insecticides used against it (Keiding 1999). Identification of the gene involved in a specific resistance is hampered by the lack of a genome sequence. House fly control is also practiced using biological control agents such as pteromalid wasps (Axtell 1990).

House flies do not suffer from severe inbreeding depression (Reed and Bryant 2004), and many highly inbred strains (which are preferred for genome sequencing) are available globally. For identification of genes responsible for various traits, as well as for identification of polymorphisms, house fly strains are available globally for identification of genes responsible for various traits, as well as for identification of polymorphisms. House flies do not suffer from severe inbreeding depression (Reed and Bryant 2004), and many highly inbred strains (which are preferred for genome sequencing projects) are available.

In the house fly, sex is determined by a dominant factor, M, which is located on the Y chromosome in “standard” populations. Thus, males are XYM and females XX (Hiroyoshi 1964, DiBenedorfer et al. 2002). This is believed to be the ancestral state of sex determination in house flies (Bull and Charnov 1977, Denholm et al. 1983). However, there are “autosomal male” (A^M) strains in which the M factor is located on one or more of the five autosomes (I–V) (Franco et al. 1982, Inoue et al. 1983, Tomita and Wada 1989, Hamm et al. 2005, Hamm and Scott 2009, Kozielska et al. 2008) or even rarely on X (Schmidt et al. 1997). The M factor located on Y functions biologically in a way identical to the M located on any of the other autosomes (Tomita and Wada 1989, Schmidt et al. 1997). The sequence of M is unknown. In the A^M strains, females are XX and males are also XX (or XO) (Hiroyoshi 1964; Wagener 1969; Franco et al. 1982, Denholm et al. 1983, 1990). Populations are found in which males are A^M/A^M (Tomita and Wada 1989, Hamm and Scott 2009). Such populations have females with F (feminizing factor located on autosome 4), which is epistatic to M, as a means to produce female offspring. In these populations, females have become the heterogametic sex. F has recently been sequenced (D. Bopp, personal communication).

Quantitative real-time polymerase chain reaction (qRT-PCR) has been shown to be a reliable method for determination of genome size (Wilhelm et al. 2003), and the genome sizes of D. melanogaster and M. domestica were recently compared using this method. The size of the D. melanogaster genome was found to be 180–181 Mbp (in agreement with the published genome size: Adams et al. 2000), and the size of the M. domestica genome was found to be 309–312 Mbp (Gao and Scott 2006) or ~1.7-fold larger than D. melanogaster. This indicates that the size of the house fly genome makes it an excellent candidate for whole genome sequencing.
Rationale for Sequencing of the House Fly Genome

Expanding Our Understanding of Basic House Fly Biology to Develop New Control Strategies. Given the tremendous importance of house flies in the transmission of human and animal diseases, substantial effort has been made to control this pest. Availability of the house fly genome will allow for identification of important target sites and will allow for the development of selective new insect control agents. Identification of novel target sites in the house fly will also aid in the development of new insecticides for control of house flies, as well as of agricultural pests that limit the supply (and quality) of human foods. A genome sequence would also provide an opportunity to explore biological control in novel ways, including disruption of the unusual autosome-based sex determination system, sterile male release, confounding signals for mate recognition, etc. Such approaches may be safer than insecticides, given the proximity of house flies to humans, animals, and many of their important food sources, and would offer important alternative control measures for organic farmers.

The biochemistry and genetics of insecticide resistance have been well studied in the house fly, arguably more widely than in any other insect. This is because of the medical and economic importance of house flies, that they are direct targets of insecticide control, the relatively rapid rate at which they develop resistance, the availability of strains resistant to almost every class of insecticide, the well-understood biology of the house fly, and that the house fly has proven to be a useful model for understanding and predicting resistance in other insect species (Scott 1990, 1991, 1999). Availability of the house fly genome would allow for more rapid identification of the genes and regulatory sequences involved in resistance to insect control agents.

The house fly has been, and continues to be, a major insect for studies of environmental toxins. It has had a preeminent role in insect toxicology studies, especially with focus on comparative toxicity between insects and mammals, in the development of new insecticides (Casida and Quistad 2004) and insecticide resistance (see above). The house fly has been a model insect for these scientific areas of inquiry. Completion of the house fly genome will facilitate this research by the identification of novel target sites, further elucidation of differences in target sites (ion channels, neuroreceptors, hormones, etc.) between insects and mammals, and by facilitating identification of genes involved in insecticide resistance. The neonicotinoids serve as an excellent example of the payoff that comes from this comparative biochemical approach. Neonicotinoids are the fastest growing class of insecticides and were developed specifically by the process of selecting agents that interact with insect and not mammalian receptors (Matsuda et al. 2001, Nauen et al. 2003, Tomizawa and Casida 2003, Wakita et al. 2003). Studies of house fly nicotinic acetylcholine receptor subunits have recently identified novel RNA editing sites (Gao et al. 2007a, b, c), but lack of a genome sequence hampers efforts to understand the scope of RNA editing in this species.

Despite efforts by developmental biologists, the molecular identity of M has remained elusive. In Drosophila, Sex-letal (Sxl) integrates information about the dose of X and autosomes and provides the initial switch for the sex determination cascade. In M. domestica, M is not a signal for Sxl, and in fact, Sxl is not involved in sex determination (Meise et al. 1998). Most intriguingly, in Drosophila, sex determination and dosage compensation are tied to the same pathway, whereas these processes are decoupled in Musca. It has been speculated that it is this decoupling that gives Musca the impressive flexibility and polymorphism in sex determination mechanisms (Dubendorfer et al. 2002, Hediger et al. 2004). A full genome sequence of M. domestica would allow immediate identification of all homologs to the Drosophila sex determination cascade and would greatly accelerate discovery of the genes that cause the radical divergence in the fundamental processes of sex determination found in most Diptera compared with dosage compensation found in Drosophila (Dubendorfer et al. 2002, Hediger et al. 2004).

The house fly has been a model system for studies of insect olfaction (Kelling et al. 2002, 2003), and (Z)-9-tricosene plays an important role in intersex communication and mate selection in house flies. Sequencing of the house fly genome will identify receptor molecules (in antennal and palpal olfactory cells) that will aid olfact studies, and will facilitate development of attractants for house flies to baits in management systems (Darbro and Mullens 2004, Hanley et al. 2004).

Improved Understanding of Insect Immunity. The house fly thrives in a virtual sea of animal pathogens. Sequencing of the house fly genome will shed light on the immune defense systems of this important species and provide valuable information about how it is able to flourish, despite living in intimate contact with septic flora and fauna. Comparison between the innate immune systems of Musca, Drosophila, and Anopheles, which face different ecological pressures and pathogens, will be informative, just as the Drosophila–Anopheles comparison has been (Christophides et al. 2002). The relatively close relationship to Drosophila has already greatly expedited this analysis, because >30 individual innate immunity genes have been sequenced in Musca. The advantage of a genome sequence is that it will allow discovery of genes unique to Musca and regulatory systems that allow it to survive in a far more septic environment.

Completion of the Drosophila and Anopheles genomes provided unprecedented opportunities to study insect–pathogen interactions (Christophides et al. 2002, Lazzaro and Clark 2003, Schlenke and Begun 2003, Lazzaro et al. 2004, Osta et al. 2004, Srinivasan et al. 2004). The house fly will also be of great value for two reasons. First, house flies live in intimate association with vertebrate pathogens such as Helicobacter pylori (causative agent of gastric ulcer; Li and Stutzenberger 2000), Salmonella, Campylobacter jejuni
(Hald et al. 2004), *E. coli* (including toxin producing strains E105 and O157:H7 that cause food poisoning; Moriya et al. 1999, de Jesus et al. 2004), and trachoma (i.e., transmission of *Chlamydia trachomatis*, Emerson et al. 1999, 2000). However, house flies are remarkably resilient to pathogens. Understanding the basis for their refractoriness to many pathogens would offer important insights into ways to improve human health. Second, house fly populations in temperate climates are occasionally decimated by Entomophthora, an entomopathogenic fungus (*Zygomycetes*, Entomophthoraceae). A genome sequence would expedite the study of why certain populations of *Musca* are sensitive to this fungus, whereas others are refractory. Microarray studies using the house fly genome to investigate genes associated with pathogen exposure will be a cornerstone in future studies in this field (Jensen et al. 2001, Kalsbeek et al. 2001, Zurek et al. 2002) and would become a model system for biological control (entomopathogenic fungi; parasitic hymenoptera; microsporidia) of insects.

**Comparative Genomics of a Host–Parasitoid System.** Sequencing of the parasitoid wasp *N. vitripennis* has completed the genome. *Nasonia* is a parasitoid of the house fly (*Nasonia* is sold commercially for fly control). Having the genome of both the parasitoid (*Nasonia*) and the host (*M. domestica*) will allow the first opportunity to study the interactions between a pest insect and its parasitoid at the whole genome level. How do *Nasonia* eggs evade the immune response of the house fly? What is the response of the house fly to *Nasonia* venoms? What immunological factors permit some flies but not others to destroy the developing *Nasonia* embryo? Sequencing of the house fly genome is strongly supported by the *Nasonia* community (Scott et al. 2008).

**Improving Genome Annotation, Especially Within Insects and Diptera.** The Dipteran clade has radiated into >120,000 known species since its origin in the late Jurassic. *M. domestica* is well placed within the Diptera to maximize the utility of sequence data for comparison between existing Dipteran genomes. Although systematic/phylogenetic research on Diptera has been carried out for more than a century, a well-supported tree for the entire order has not been completed (Yeates and Wiegmann 1999). However, it is clear that house fly and *Drosophila* represent a different suborder than *Anopheles*, and house fly represents a different Section (Calyptrate) than *Drosophila* (Acalyptrate) (note: some classifications differ in the taxonomic level where they split calyptrate and acaleyptate flies). Multiple, deeply divergent comparisons within the order allows identification of lineage effects on rates and patterns of genomic diversity. These comparisons become more powerful in elucidating genome evolution as the phylogenetic context is broadened. Given the well-centered position between *Drosophila* and mosquitoes, the *Musca* genome would be nearly ideal for leveraging analysis and annotation of the *Anopheles* and *Aedes* genomes by bridging this gap. In addition, the house fly genome will provide a valuable outgroup for analyses of *Drosophila* genomes, given their more recent common ancestor (compared with *Drosophila* versus *Anopheles* or *Aedes*). The deepest common ancestor to the set of *Drosophila* species whose genomes were sequenced is estimated to be 60–40 million years ago (MYA), and the common ancestor between *D. melanogaster* and *M. domestica* has been estimated to be ~100 MYA (Beverley and Wilson 1984). This places it remarkably well in the gap between *Drosophila* and *Anopheles* and will allow a very broad evolutionary analysis across the Dipteran order.

The *Glossina* genome has recently been suggested as worthy for having its genome sequenced (Aksoy et al. 2005), and this effort is underway. However, this does not disqualify the house fly for a number of reasons. First, the *Glossina* genome is quite large (500–600 Mb) and contains numerous repeat sequences (Aksoy et al. 2005) that will make the sequencing effort quite difficult. Second, *Glossina* and *Musca* would both provide useful outgroups for *Drosophila* and mosquito genomes (without being redundant). At the same time, they would provide useful comparisons with each other at a level of difference similar to a comparison of *Anopheles* and *Aedes*. In addition, house flies have many unique features (including their profound impact on human and animal health) and have entirely different lifestyles, habitats, and behaviors from *Glossina*, all of which justify sequencing the house fly genome, whether or not *Glossina* is sequenced. Furthermore, house fly has far better genetics; more is known about its sex determination, physiology, biochemistry, neurobiology, and evolution. The house fly is a global pest, whereas *Glossina* is a pest only in Africa.

The completed genome sequences of *Drosophila melanogaster* and *Anopheles gambiae* have been extremely valuable for deductions about the evolutionary origins, structure, and even the function of many human genes (Kortschak et al. 2003). Nevertheless, a significant number of gene modifications and extensive gene loss has occurred in *Drosophila*. Although the genomes and proteomes of *An. gambiae* and *D. melanogaster*, which diverged ~220–240 MYA (Wiegmann et al. 2003), show considerable similarities, both lineages have experienced multiple gene acquisitions and losses, especially through expansions and contractions of gene families (Zdobnov et al. 2002). Sequences of orthologous genes in these two insect species have diverged to the point that synonymous positions are virtually randomized (Zdobnov et al. 2002). It was hoped that regulatory regions of genes would become clear by comparison of 5′ regions of genes in *Drosophila* and *Anopheles*, but this has proven to be much more difficult. For example, the 5′ and 3′ regulatory flanking regions of many genes in house flies are virtually unalignable to the orthologous sequence in *Drosophila* (Shaw et al. 2001); the genetic cascades regulating sex determination of the house fly and *D. melanogaster* seem strikingly different, and the upstream regulators of sex determination genes are different between these two insect species (Düben-dorfer et al. 2002). Furthermore, 24% of *Apis* expressed
sequence tags (ESTs) showed better matches to Chor-
data than to Drosophila genes (Whitfield et al. 2002). Some Apis ESTs showed significant matches to human sequences, but no matches to the Drosophila genome (inferred to be genes that were lost from Drosophila). Similar results have also been identified in the current house fly EST sequences (N.L., unpublished data). Although either Drosophila or An. gambiae (or both) homologs could be recognized for more than one half of the house fly EST sequences, some of these EST sequences showed better matches to other more distant species, such as Plasmodium falciparum, Carassius auratus (goldfish), and Homo sapiens, than to Dro-
sophila and/or An. gambiae homologs. Some of the Musca EST sequences showed no matches to the Dro-
sophila and/or An. gambiae genome. These results indicate that the genomic sequences from other insect species will be extremely important for linking human genes to their Drosophila or An. gambiae homologs.

Understanding the evolution of cis-regulatory se-
quences in Drosophila has proven difficult in some cases (e.g., achaete-scute genes), because the patterns of expression are not substantially different between Drosophila species but are so extremely diverged in Anopheles that analysis is difficult. Thus, the house fly genome would provide a critical resource for the analysis of cis-regulatory sequences in Drosophila.

Conclusions

A thorough understanding of the biology of com-
plex organisms requires complete sequencing infor-
mation and identification of all functional elements from the genomes of these organisms. The “whole genome” approach has vastly improved comparative and evolutionary studies, as well as physical map building. It has addressed several important scientific ques-
tions about genome evolution, such as evolutionary rate, speciation, genome reorganization, and origins of variation. The approach has also been important for identification of conserved sequences involved in gene regulation and other genomic functions, identification of specific functional sequences (i.e., those that have been substituted or modified during evolution, and which have undergone recent selection; Vandahl et al. 2004), and elucidation of sequence variation in the population of organisms (such as alter-
native splicing in the regulation of gene function; Tan et al. 2002). The whole genome approach will also be important for identification of sequences that are broadly conserved across insect genomes to provide insight into the unique features in the genome and for obtaining a broader and more complete assessment of the extent of genetic variation in the population of organisms; identification of variation in gene expression; and understanding the evolution (Yan et al. 2002). The whole genome is also necessary to under-
stand the interactions of house flies with the parasitoid wasp N. citripennis. Although some of the genes that are expected to be regulated by parasitoid venoms and egg laying could be inferred from other studies, only a whole genome will provide comprehensive insight into the genes involved in host/parasitoid interac-
tions.

The calypter flies, with M. domestica as the most prominent experimental organism, includes a large number of important vectors of human and veterinary diseases, as well as important species for forensic en-
tomology: dog dung fly (Musca sorbens), face fly (Musca autumnalis), blow flies (Lucillia, Calliphora, Chrysomya), flesh flies (Sarcophaga), screwworm (Cochliomyia), tsetse fly (Glossina), the little house fly (Fannia), warble flies (Hylopera), yellow dung flies (Scathophaga), and the root maggot fly (Antho-
myiida). By using genetic manipulations of M. domes-
tica to place function of novel genes in its genome, we anticipate that it will be easy to transfer the knowledge gained to other synanthropic flies.

Many genes, especially regulatory genes, are often expressed at a very low level, and they would be rare in EST libraries. The entire house fly genome sequence will, especially when compared with the Dro-
sophila and mosquito genome sequences, facilitate the identification of homologous genes expressed at low levels or in a specific tissue. Expression patterns can be validated with high-throughput real-time PCR sys-
tems for use in either general population or micro-
evolutionary studies (e.g., the spread and fitness of resistance genes).

Currently there are ≈40 laboratories worldwide whose primary research focus is the house fly. About one half of these are engaged in studies of molecular biology that would immediately benefit from a complete genome sequence. Most of the others are study-
ing aspects of toxicology and pest control, and imme-
diate access to design of primers for PCR analysis would open the door to simple but powerful molecular approaches to this group. Drosophila researchers would benefit from, and are strongly supportive of (Scott et al. 2008), sequencing of the house fly gene-
nome. The white papers that resulted in funding to sequence an additional 11 genomes of Drosophila spe-
cies failed to include an outgroup to the set of Dro-
sophila species, and Anopheles gambiae is just too dis-
tantly related for optimal analysis (in most cases).

Suitable tools are available to facilitate sequencing of the house fly genome, and the scientific community is solidly behind this effort. Several cDNA and geno-
ic libraries (various tissues, strains, and life stages) have been prepared, and at least two pilot EST projects are underway. More than 450 nucleotide and >750 EST sequences from M. domestica can be found in GenBank. As a part of USDA NRI and Auburn University Biogrant funded project, a house fly nor-
malized cDNA library has been constructed from the mRNA of house flies. More than 300 ESTs have been generated, resulting in 292 high-quality cDNA se-
quence reads. Thirty-nine ESTs were assembled into eight contigs. The remaining 253 ESTs are unique, suggesting a 15% redundancy in the house fly se-
quence set. This EST sequencing effort, combined with other larger EST projects, will be excellent re-
sources for the genomic library (BAC library) screen-
ing and building contig maps for comparative genomic
studies. House flies can be readily transformed with mobile elements such as piggyback, hermes, or hobo (Atkinson et al. 1993; O’Brochta et al. 1994, 1996; Warren et al. 1994; O’Brochta and Atkinson 1996, 1997; Sarkar et al. 1997; Hediger et al. 2001), and some genes can be silenced using RNAi (McGregor et al. 2001, Burghardt et al. 2005).

Letters of support written for the house fly white paper (Scott et al. 2008) eloquently showed how researchers from diverse scientific areas (genomics, proteomics, developmental biology, population genetics, evolutionary biology, etc.) that use a wide range of study animal (mosquitoes, Drosophila, Nasonia, Tribolium, Musca, etc.) would make immediate use of the M. domestica genome sequence to accelerate their research programs on fundamental aspects of genetics (sex determination, dosage compensation, olfaction, immunology, etc.), as well as practical problems of pest control. It is clear that the scientific community considers sequencing of the house fly genome to be an extremely high priority.

References Cited


Hamm, R. L., T. Shono, and J. G. Scott. 2005. A cline in frequency of autosomal males is not associated with in-


Received 14 June 2008, accepted 4 November 2008.