References: Ashburner, M., 1967, Chromosoma 21: 398-428; Fraenkel, G., and U.J. Brooker 1953, Biol. Bull. 105: 442-446; Korge, G., 1977, Dev. Bio. 58: 339-355; Kress, H., A. Jarrin, E. Thuroff, R. Saunders, C. Weise, M.S. Busch, E.W. Knapp, M. Wedde, and A. Vilcinskas 2004, Insect Biochemistry and Molecular Biology 34: 855-869; Loveriza, A.S., and H.K. Mitchell 1982, Developmental Genetics 3: 255-272; Mitchell, H.K., U.W. Tracy, and L.S. Lipps 1977, Biochemical Genetics 15: 563-573; Nirmala, S.S., and N.B. Krishnamurthy 1973, Orient. Insect. 7: 267-270; Ramesh, S.R., and W.E. Kalisch 1988, Biochemical Genetics 26: 527-540; Ramesh, S.R., and W.E. Kalisch 1989, Biochemical Genetics 27: 507-51; Shirk, P.D., P.A. Roberts, and C.H. Harn 1988, Roux's Arch. Dev. Biol. 197: 66-74; Velissariou, V., and M. Ashburner 1980, Chromosoma 77: 13-27.

Raw whole *Drosophila* genome sequence traces have contaminant sequences from bacterial symbionts.

O'Connell, Karen E., and Mohamed A.F. Noor. Biology Department, Duke University, Durham, NC USA; karen.oconnell@duke.edu.

Abstract

Many *Drosophila* genomes have recently been sequenced and assembled, and many more genome sequencing projects are in progress. *Drosophila* species have bacterial, fungal, and protozoan symbionts; therefore, DNA from these symbionts may be isolated in the course of sequencing *Drosophila* genomes. Here, we assess how much sequence is isolated from these symbionts and whether sequence contamination from symbionts affected the assembly of published *Drosophila* genomes. We find raw sequence from bacterial symbionts and humans in genome sequence traces analyzed. Surprisingly, the four most-common contaminant species were shared among the *Drosophila* genomes. However, we do not find evidence of bacterial sequences in two published *Drosophila* genome assemblies.

Introduction

The genus *Drosophila* has been a vital model system in studying adaptation, speciation, and genetics (*e.g.*, Morgan, 1910; Mustonen and Lassig, 2007; Matute *et al*., 2010;). Further investigation of such processes can now be accomplished in *Drosophila* species using bioinformatic tools and data sets, because an increasing number of *Drosophila* genomes have been assembled and annotated (Clark *et al.*, 2007). Bacterial, fungal, and protozoan symbionts, however, live in the gut and on the exterior surfaces of the *Drosophila* species (Ebbert *et al.*, 2003; Cox and Gilmore, 2007), and sequences from these symbionts may contaminate sequence from the focal organism (*e.g.*, Salzberg *et al*., 2005). If this "contamination" is not eliminated from raw genomic data sets, it has the potential to be incorporated incorrectly into a final genome assembly.

To alleviate this concern, researchers bioinformatically filter out raw sequence reads from known symbionts or rear flies in a manner that limits their bacterial content (Myers *et al*., 2000; Clark *et al.*, 2007). Filtration of all bacterial sequences may, however, be difficult when sequencing a genome *de-novo* or assembling a genome without knowledge of all potential symbionts. Whole adult

flies were used for DNA isolation and extraction, and no bioinformatic filtering was performed on raw genome sequences when assembling genomes of *D. pseudoobscura pseudoobscura* (Kulathinal *et al*., 2008), *D. pseudoobscura bogotana* (Kulathinal *et al*., 2009), *D. persimilis* (Stevison and Noor, 2010), and *D. miranda* (Kulathinal *et al*., 2009). These studies assembled genomes using two published *Drosophila* genome assemblies as a backbone. Due to the lack of filtering and the isolation of DNA from whole adult flies, contaminant sequences were likely isolated in these raw genomic data sets and could have been incorporated into the genome sequence assemblies of these flies (Richards *et al*., 2005; Clark *et al.*, 2007). To examine the extent of foreign contamination that is possible, we quantify and characterize raw sequence reads derived from symbionts using a custom bioinformatic pipeline. We then ask if the identified contaminant sequences were incorrectly incorporated into published *Drosophila* genome assemblies (Richards *et al*., 2005; Clark *et al*., 2007).

Materials and Methods

The genomes of *D. persimilis* and *D. pseudoobscura pseudoobscura* were sequenced and assembled by the *Drosophila* community (Richards *et al*., 2005; Clark *et al.*, 2007). More recently, our laboratory isolated raw genomic DNA from multiple adult females via PureGene protocol from four different species of *Drosophila* and sequenced their genomes at low coverage using high-throughput 454-FLX technology. *D. miranda* (SRA accession: SRX003254), *D. persimilis* (SRA accession: SRX015434 and SRX015435), and *D. pseudoobscura bogotana* (SRA accession: SRX003253 and SRX003254) were sequenced at Duke University's IGSP facility, while *D. pseudoobscura pseudoobscura* (SRA accession: SRX001087 and SRX003252) and more *D. pseudoobscura bogotana* (SRA accession: SRX003252) were sequenced at 454 Life Sciences (Branford, CT).

A pipeline of computer scripts was constructed using Perl (available via Dryad accession http://dx.doi.org/10.5061/dryad.8085), and sequence similarity was assessed using a Basic Local Alignment Search Tool (BLAST: Altschul *et al*., 1990). The pipeline aims to identify raw sequence reads that do not appear to come from *Drosophila*, and then identifies which species shares the most similarity with each ambiguous sequence using sequences available at the National Center for Biotechnology Information's (NCBI) nucleotide database. Raw sequence reads less than 51 base pairs were not used in these analyses, and we imposed an e-value cutoff of 1×10^{-10} in all BLAST alignments. Specifically, the pipeline proceeded through 5 steps: 1) BLAST each raw sequence read in all four fly species to the published assembled genomes of *D. persimilis* and *D. pseudoobscura pseudoobscura*, 2) designate those reads that do not successfully align to these published genomes as "ambiguous", 3) individually BLAST these ambiguous sequence reads to the non-redundant 'nr' nucleotide database on NCBI, 4) record which species' sequence best aligned to each given ambiguous sequence, and 5) informatively summarize the species to which all the ambiguous reads aligned.

To see if bacterial symbionts identified in the pipeline were incorporated into genome assemblies, we BLASTed whole genome assemblies of *Drosophila* to bacterial contaminant genomes and manually analyzed the resulting alignments. We also BLASTed genome scaffolds from *D. persimilis* to the "whole genome sequence" nucleotide database on NCBI. Any regions of *Drosophila* genome assemblies that seemed to have bacterial contamination incorporated were directly tested using PCR.

Table 1. Summary of BLAST Pipeline of Raw Genomic Sequence. Each cell contains the number of raw sequence reads (bigger than 50bp) in a given species that belong in the category described.

Fly species	Raw # reads	# not aligning to *Drosophila* genome assemblies	# aligning to something on NCBI	# aligning to *Homo sapiens*	# aligning to other *Drosophila*	# aligning to something else*
D. miranda	486,125	16,264	2,623	210	695	1,718
D. persimilis	426,051	13,237	4,488	9	381	4,102
D. p. bogotana	330,071	10,402	1,061	127	318	616
D. p. pseudoobscura	261,948	1,944	352	32	198	122

* 'Something else' means something non-human and non-Drosophila

Table 2. Summary of Non-human and Non-Drosophila Alignments. Each cell contains the number of raw sequence reads from the fly genome listed across the top row that aligned to the bacterial species listed on the left-hand column.

Bacterial Species	*D. miranda*	*D. persimilis*	*D. p. bogotana*	*D. p. pseudoobscura*
Acetobacter pasteurianus	231	334	37	4
Gluconobacter oxydans	140	1622	125	0
Gluconacetobacter diazotrophicus	417	192	59	2
Enterococcus faecalis	21	16	10	6
Lysinibacillus sphaericus	1	0	0	29
Lactobacillus brevis	40	8	0	0
Bacillus cereus	1	0	0	6
Propionibacterium acnes	1	0	1	0

Results

Most raw sequence reads successfully aligned to one of the two assembled and published genomes (Table 1). Of the sequences not aligning to the published genomes, many strongly matched human genome sequences, indicating a degree of human contamination from handling. Most ambiguous sequences did not, however, successfully align to anything on NCBI's non-redundant 'nr' nucleotide database, suggesting most ambiguous sequences come from regions of genome that are difficult to assemble (*e.g.*, heterochromatic regions) or from organisms that are not yet sequenced. We identified a common set of four bacterial species (*Acetobacter pasteurianus*, *Gluconobacter oxydans*, *Gluconacetobacter diazotrophicus*, and *Enterococcus faecalis*) found in three *Drosophila* genomes (Table 2) when analyzing the species to which ambiguous sequences align. These bacterial profiles are unlikely to result from contamination at the sequencing facilities, because the fly species sharing bacterial contaminant profiles were sequenced at different sequencing facilities. The four bacterial species, or their close relatives, have potential to be associated with *Drosophila* species, because two of them are found in soils with fruits and vegetables (*Gluconobacter oxydans* and *Acetobacter pasteurianus*), one is a symbiont of sugar cane (*Gluconacetobacter diazotrophicus*), and the last is a common probiotic found in human intestines (*Enterococcus faecalis*). To verify that sequences from these bacteria were isolated in genomic DNA from these flies, we designed PCR primers specific to these four species of bacteria and successfully amplified DNA specific to

Gluconacetobacter diazotrophicus from *D. pseudoobscura pseudoobscura* and *D. pseudoobscura bogotana* genomic DNA.

We asked if any bacterial sequence reads from symbionts had been incorporated into published *Drosophila* whole genome assemblies (Richards *et al*. 2005; Clark *et al*. 2007) after identifying bacterial sequences that were isolated in raw genomic DNA. In the first analysis, we focused on *D. persimilis'* genome given its lower coverage and quality. No alignments to bacterial genomes were observed after BLASTing each genome scaffold of *D. persimilis* to the whole genome sequence or 'wgs' database on NCBI. Second, we downloaded the whole genomes of bacteria that have been found to live within or on the surface of *D. melanogaster* (Martin *et al*., 1972; Cox *et al.*, 2007; Ren *et al*., 2007; Roh *et al*., 2008; Ryu *et al.*, 2008) and the bacterial species identified in the bioinformatic pipeline. We then asked if these bacterial whole genome sequences aligned to any portion of the whole genome assemblies of *D. persimilis* and *D. pseudoobscura pseudoobscura*. Most positive hits in these alignments of whole genome sequences had either high identity and very short length or low identity and longer length (Table 3). One whole genome alignment between *Enterobacter cloacae* and *D. pseudoobscura pseudoobscura* was the best candidate for misincorporation of symbiotic DNA, because it was 228bp long and had 98% identity. The contig identified in this alignment, however, has 1× coverage and part of the contigs shows homology to other *Drosophila* species.

Table 3. Successful Hits from Whole Genome Alignments.

Bacterial Genome Assemblies	*D. persimilis* Genome Assembly			*D. p. pseudoobscura* Genome Assembly		
	Length of hit	Identity	Gaps	Length of hit	Identity	Gaps
Acetobacter pasteurianus	33	100%	0%	33	96%	0%
Gluconacetobacter diazotrophicus	407	80%	2%	407	80%	2%
Cladosporium shaerospermum	62	87%	6%	938	83%	6%
Lactobacillus brevis	171	78%	8%	293	74%	10%
Enterobacter cloacae	No match			228	98%	0%

The similarity between *Enterobacter cloacae* and *D. pseudoobscura pseudoobscura* could exist due to chance, a horizontal transfer event between the bacteria and the fly, or it could simply be an artifact of incorrect genome assembly. To determine if a horizontal gene transfer event had occurred in the history of *Drosophila pseudoobscura pseudoobscura*, we attempted to amplify chimeric sequence from fly whole genome DNA. The PCR was unsuccessful, indicating the alignments observed may have resulted from incorrect genome assembly. We then performed another PCR and sequenced the region of genome surrounding the putative bacterial transfer. The sequence isolated did not show homology to any bacteria and indicated that the region of the *D. pseudoobscura pseudoobscura* assembly incorrectly inferred a gap of 60 ambiguous base pairs (N's). Instead, we found the region where 60 ambiguous bases were inferred is 93 bases long and showed homology to *Drosophila melanogaster* chromsome 2R (Available as a GenBank accession: HQ828988). This *D. pseudoobscura pseudoobscura* contig was most likely not assembled into the final published scaffold because of its low sequence coverage.

Discussion

Genome sequencing and annotation is useful in addressing genetic and evolutionary questions typically explored in the genus *Drosophila*. Here, we have shown the sequencing and assembly process can be complicated by sequence from symbionts living in or on a fly. Fortunately, it appears the contaminant sequences were not incorporated into two published genome assemblies. We were, however, able to find common bacterial contaminant species in our laboratory's raw *Drosophila* genome sequences. Additionally, a large portion of sequence reads did not align to the *D. persimilis* and *D. pseudoobscura pseudoobscura* assemblies nor to any sequence available on NCBI's whole genome database (Table 1). These sequences may come from difficult-to-assemble regions of the genome, but they could also suggest there is contamination in genomes (perhaps bacterial or fungal) researchers have yet to sequence and characterize. When sequencing a genome *de-novo*, there is often no closely related species to use as a reference, and filtration of potential bacterial contaminants is potentially much more important. Genome sequencing and assembly is a complicated process that should not be viewed as a "black box" raw sequences go into and genomes come out of. Knowledge of potential sequence contamination and confidence in the assembly process (*e.g.*, presence of reference genome) helps to eliminate incorrect genome assembly.

References: Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman 1990, J. Mol. Biol. 5: 403-410; Cox, C.R., and M.S. Gilmore 2007, Infect. Immun. 75: 1565-1576; Clark, A.G., M.B. Eisen, D.R. Smith, C.M. Bergman, B. Oliver, T.A. Markow, T.C. Kaufman, M. Kellis, and W. Gelbart 2007, Nature 450: 203-218; Ebbert, M.A., J.L. Marlowe, and J.J. Burkholder 2003, J. Invertebr. Pathol. 83: 37-45; Goecks, J., Nekrutenko, A., Taylor J., and The Galaxy Team 2010, Genome Biol. 11: R86; Kulathinal, R.J., S.M. Bennett, and C.L. Fitzpatrick 2008, Proc. Natl. Acad. Sci. USA 105: 10051-10056; Kulathinal, R.J., L.S. Stevison, and M.A.F. Noor 2009, PLoS Genet. 5(7): e1000550; Martin, J.D., and J.O. Mundt 1972, Amer. Soc. Microbiol. 24: 575-580; Matute, D.R., I.A. Butler, D.A. Turissini, and J.A. Coyne 2010, Science 329: 1518-1521; Morgan, T.H., 1910, Science 32: 120-122; Mustonen, V., and M. Lassig 2007, Proc. Natl. Acad. Sci. USA 104: 2277-2282; Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H.J., Remington, K.A., Anson, E.L., Bolanos, R.A., Chou, H., Jordan, C.M., Halpern, A.L., Lonardi, S., Beasley, E.M., Brandon, R.C., Chen, L., Dunn, P.J., Lai, Z., Lian, Y., Nusskern, D.R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G.M., Adams, M.D. and J.C. Venter 2000, Science 287: 2196-2204; Ren, C., P. Webster, S.E. Finkel, and J. Tower 2007, Cell Metab. 6: 144-152; Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., Couronne, O., Hua, S., Smith, M.A., Zhang, P., Liu, J., Bussemaker, H.J., van Batenburg, M.F., Howells, S.L., Scherer, S.E., Sodergren, E., Matthews, B.B., Crosby, M.A., Schroeder, A.J., Ortiz-Barrientos, D., Rives, C.M., Metzker, M.L., Muzny, D.M., Scott, G., Steffen, D., Wheeler, D.A., Worley, K.C., Havlak, P., Durbin, K.J., Egan, A., Gill, R., Hume, J., Morgan, M.B., Miner, G., Hamilton, C., Huang, Y., Waldron, L., Verduzco, D., Clerc-Blankenburg, K.P., Dubchak, I., Noor, M.A.F., Anderson, W., White, K.P., Clark, A.G., Schaeffer, S.W., Gelbart, W., Weinstock, G.M., and R.G. Gibbs 2005, Genome Res. 15: 1-18; Roh, S.W., Y. Nam, H. Chang, K. Kim, M. Kim, J. Ryu, S. Kim, W. Lee, and J. Bae 2008, Appl. and Environ. Microbiol. 74: 6171-6177; Ryu, J., S. Kim, H. Lee, J.Y. Bai, Y. Nam, J. Bae, D.G. Lee, S.C. Shin, E. Ha, and W. Lee 2008, Science 319: 777-782; Salzberg, S.L., J.C. Dunning Hotopp, A.L. Delcher, M. Pop, D.R. Smith, M.B. Eisen, and W.C. Nelson 2005, Genome Biol. 6: R23; Stevison, L.S., and M.A.F. Noor 2010, J. Mol. Evol. 71: 332-345.