Benchmarking Next-Generation Transcriptome Sequencing for Functional and Evolutionary Genomics

John G. Gibbons, Eric M. Janson,* Chris Todd Hittinger,†‡ Mark Johnston,†‡ Patrick Abbot,* and Antonis Rokas**

*Department of Biological Sciences, Vanderbilt University, Nashville, TN; †Department of Biochemistry and Molecular Genetics, University of Colorado Denver Health Sciences Center; and ‡Department of Genetics, Center for Genome Sciences, Washington University in St Louis School of Medicine

Next-generation sequencing has opened the door to genomic analysis of nonmodel organisms. Technologies generating long-sequence reads (200–400 bp) are increasingly used in evolutionary studies of nonmodel organisms, but the short-sequence reads (30–50 bp) that can be produced at lower cost are thought to be of limited utility for *de novo* sequencing applications. Here, we tested this assumption by short-read sequencing the transcriptomes of the tropical disease vectors *Aedes aegypti* and *Anopheles gambiae*, for which complete genome sequences are available. Comparison of our results to the reference genomes allowed us to accurately evaluate the quantity, quality, and functional and evolutionary information content of our "test" data. We produced more than 0.7 billion nucleotides of sequenced data per species that assembled into more than 21,000 test contigs larger than 100 bp per species and covered ~27% of the *Aedes* reference transcriptome. Remarkably, the substitution error rate in the test contigs was ~0.25% per site, with very few indels or assembly errors. Test contigs of both species were enriched for genes involved in energy production and protein synthesis and underrepresented in genes involved in transcription and differentiation. Ortholog prediction using the test contigs was accurate across hundreds of millions of years of evolution. Our results demonstrate the considerable utility of short-read transcriptome sequencing for genomic studies of nonmodel organisms and suggest an approach for assessing the information content of next-generation data for evolutionary studies.

Introduction

Model systems have so far been the source of most genomic data, even though they are not always optimal for evolutionary studies (Abzhanov et al. 2008). Therefore, a gulf exists between the bulk of evolutionary genomics research and the detailed natural history and evolutionary information available from nonmodel organisms. Next-generation DNA sequencing technologies (e.g., Solexa/Illumina, SOLiD/ABI, 454/Roche, and HeliScope/Helicos, Margulies et al. 2005; Bentley et al. 2008; Hudson 2008; Mardis 2008a, 2008b) drastically reduce the cost, time, and labor required for large-scale sequencing, promising to transform molecular evolutionary studies by changing the ranges and types of questions that can be addressed (Rokas and Abbot 2009).

Next-generation DNA sequencing technologies differ in the length and number sequence reads produced (Holt and Jones 2008), giving rise to an important trade-off. Short-read technologies (~30- to 50-bp reads) usually generate more sequence data per run at a substantially lower overall cost than long-read technologies (~200- to 400-bp reads) (Mardis 2008b). But short-read technologies are also less successful in assembling large, complex genomes (a similar trade-off exists between all next-generation sequencing technologies and the standard capillary sequencing method) (Whiteford et al. 2005). Thus, short-read technologies are typically best suited for resequencing projects where a reference genome on which the reads can be mapped is already available (Whiteford et al. 2005; Schuster 2008), although recent algorithmic (Butler et al. 2008) and experimental (e.g., Solexa/Illumina's new mate-pair and short-read paired-ends libraries) advances are likely to increase their suitability for de novo genome sequencing projects, especially in combination with long-read technologies (Maher et al. 2009).

A large fraction of studies in molecular evolution and evolutionary genetics is concerned with the coding parts of genomes and less with their full sequences. For example, several evolutionary experiments require variable genetic markers, and coding sequences can provide a wealth of markers scattered across the genome. A standard way of obtaining coding sequences is through partial sequencing of mRNA transcripts, typically after they have been reverse transcribed into cDNA (Bouck and Vision 2007). Importantly, transcript sequence is usually much simpler to assemble compared with genomic sequence for two reasons: First, unlike genomic sequences, transcripts are unequally represented in an organism's mRNA pool: Transcripts of highly expressed genes are expected to be highly abundant, whereas transcripts of weakly expressed genes are present at low levels. Thus, sequencing of an mRNA pool can result in very deep coverage of highly expressed transcripts but in very low coverage of poorly expressed transcripts. Second, coding sequences typically contain fewer repetitive elements and have other properties (e.g., higher GC content) (Toth et al. 2000; Lander et al. 2001) that reduce the probability of errors in sequence assembly.

These predictions have been experimentally tested and verified in several next-generation transcriptome-sequencing studies of evolutionarily important organisms and populations lacking genomic resources (Cheung et al. 2006; Toth et al. 2007; Novaes et al. 2008; Vera et al. 2008; Rokas and Abbot 2009). However, all of those studies were performed using the long-read 454 technology (Margulies et al. 2005), perhaps contributing to the perception of the drawbacks of the more affordable short-read technologies. Furthermore, due to the unavailability of a reference genome for the organisms sequenced, absolute estimation of error rates was not always possible.

Key words: next-generation DNA sequencing, Solexa, transcriptome, orthology prediction, error rate, functional annotation.

E-mail: antonis.rokas@vanderbilt.edu.

Mol. Biol. Evol. 26(12):2731–2744. 2009
doi:10.1093/molbev/msp188
Advance Access publication August 25, 2009

We evaluated the potential of short-read next-generation DNA sequencing technology for transcriptome sequencing of nonmodel organisms. We sequenced the transcriptomes of the tropical disease vectors *Aedes aegypti* and *Anopheles gambiae* using the Solexa-sequencing technology (Bentley et al. 2008). The sequences of the genomes of both species have been determined with standard capillary sequencing, providing "reference" data, against which we could evaluate the information content of our de novo Solexa-based test data. We have extensively compared test and reference data to address four questions about our generated test sequence data: 1) What is their quantity? 2) What is their quality? 3) What is their functional information content? and 4) What is their evolutionary information?

Materials and Methods

Full transcriptomes (from *A. aegypti* and *A. gambiae*) and proteomes (from all insect species) were retrieved from the following databases: *A. aegypti* (www.vectorbase.org, Nene et al. 2007), *A. gambiae* (www.vectorbase.org, Holt et al. 2002), *Culex pipiens* (www.broad.mit.edu), *Drosophila melanogaster* (flybase.bio.indiana.edu, Adams et al. 2000), *Bombyx mori* (silkworm.genomics.org.cn, Xia et al. 2004), *Apis mellifera* (www.beebase.org, Honeybee Genome Sequencing Consortium 2006), and *Tribolium castaneum* (beetlebase.org, Richards et al. 2008).

Mosquito Rearing and Massively Parallel Sequencing Solexa cDNA Libraries

Male and female non–blood-fed *A. aegypti* (strain LVP-IB12) and *A. gambiae* (strain SUA2La+) mosquitoes were reared to adulthood using standard rearing protocols (Das et al. 2007). *Aedes aegypti* eggs were obtained through the MR4 (MRA-735, LVP-IB12, deposited by M. Q. Benedict), whereas *A. gambiae* eggs were kindly provided by Jason Pitts. The *A. aegypti* strain used in this study (LVP-IB12) is the same as the reference genome strain and was derived from 12 consecutive generations of single pair inbreeding from an already inbred substrain (LVP-SBM) (Nene et al. 2007). In contrast, the *A. gambiae* strain used (SUA2La+) is different from the reference genome strain (PEST) due to the latter strain's extinction.

Adult mosquitoes were homogenized for isolation of total RNA using Trizol (Invitrogen, Carlsbad, CA). Poly-A^+ RNA was extracted using the Qiagen Oligotex Poly A RNA extraction kit (Qiagen, Venlo, The Netherlands) according to the manufacturer's instructions. poly-A^+ RNA (0.5–2 μg) was chemically sheared in 10 mM zinc acetate for 5 min at 60 °C, quenched with an equal volume of 250 mM ethylenediaminetetraacetic acid (EDTA), and desalted on an illustra Microspin G-25 column (GE Healthcare, Chalfont St Giles, United Kingdom). Double-stranded cDNA was synthesized from the sheared RNA pool using the First-Strand cDNA Synthesis Kit (Fermentas Life Sciences, Burlington, Canada) with random hexamers, followed by treatment with *Escherichia coli* DNA Polymerase I and *E. coli* RNaseH according to the manufacturer's instructions.

Double-stranded cDNA was purified using a QIAquick polymerase chain reaction (PCR) Purification Kit (Qiagen) and prepared for Solexa-sequencing. DNA ends were repaired at 20 °C for 30 min with 0.5 U/μl T4 Polynucleotide Kinase, 0.15 U/μl T4 DNA Polymerase, and 0.05 U/μl Large (Klenow) Fragment of DNA Polymerase I in 1× T4 DNA Ligase Buffer with adenosine triphosphate (ATP) (New England Biolabs, Ipswich, MA) and 0.4 mM each deoxyribonucleoside triphosphate in a 100-μl reaction mixture and purified using QIAquick. DNA was A-tailed at 37 °C for 30 min with 0.3 U/μl Klenow Fragment (3′ → 5′ exo$^-$) in 1× NEBuffer 2 and 0.2 mM deoxyadenosine triphosphate in a 50-μl reaction mixture and purified using QIAquick. Forked Solexa adapters were annealed together at 100 μM in 10 mM Tris pH 8.0, 50 mM NaCl, 1 mM EDTA by slowly reducing the temperature from 94 °C to room temperature, desalting using an illustra Microspin G-25 column, and storing the adapters at −20 °C. Adapters were ligated to A-tailed DNA at a final concentration of 0.1 μM at 15–16 °C overnight with 0.6 U/μl T4 DNA Ligase in 1× T4 DNA Ligase Buffer with ATP in a 50 μl reaction mixture and purified using a MinElute PCR Purification Kit. 200- to 250-bp adapter–cDNA–adapter complexes were further purified by extraction from a well-resolved 3% agarose tris(hydroxymethyl)aminomethane borate EDTA gel using a QIAquick Gel Extraction Kit with isopropanol and triple the recommended volume of QG.

The Solexa cDNA library was amplified for 18 cycles according to the manufacturer's recommendations and diluted to either 2 or 4 pM. Four lanes of massively parallel sequencing by synthesis were performed and processed into SCARF files containing millions of 36-bp sequencing reads and raw quality scores using the Solexa Genome Analyzer I (*A. gambiae*) or Solexa Genome Analyzer II (*A. aegypti*) and the Solexa Pipeline Software according to the manufacturer's instructions (Illumina, San Diego, CA) (Bentley et al. 2008). All data are available on request.

Assessment of Data Quantity

We assembled the Solexa-generated sequence reads from *A. aegypti* and *A. gambiae* using the Velvet short-read assembler (Zerbino and Birney 2008). Velvet generates assemblies by searching for identical matches of a certain length (referred to as *k*-mer length) between reads. To identify the optimal *k*-mer value (sensu Zerbino and Birney 2008), we assembled our reads using *k*-mer lengths of 17, 19, 21, 23, 25, 27, and 29, without imposing any cutoffs for contig coverage or length. Test contigs $\geq$100 and $\geq$300 bp plus *k*-mer length were independently analyzed (this corrects for the 5′ overhang in each assembled contig, Zerbino and Birney 2008). Multiplying the coverage value of each test contig by its sequence length and dividing the product by read length yields the number of reads used in the construction of a given test contig. To calculate the fraction of reads used in a given de novo assembly, we summed the numbers of reads used in the construction of all test contigs in the assembly and divided it by the total number of reads.

We calculated the relative proportion of each reference transcript that was mapped by the test contigs for each of the

two species with BlastN (Altschul et al. 1997), retaining only those test contigs that were ≥100 bp long, 100% identical, and had e-values ≤ 1E−06 to a reference transcript.

Assessment of Data Quality

We calculated substitution, indel, and assembly error rates by comparing the de novo assembled *A. aegypti* test contigs with the *A. aegypti* reference transcriptome (Nene et al. 2007). Importantly, the *A. aegypti* strain used in our study (LVP-IB12) is the same as the reference genome strain: It is highly inbred and harbors extremely low levels of genetic variation (Nene et al. 2007). It is likely that the overwhelming majority of differences observed in our comparisons are sequencing errors and not genuine sequence polymorphisms. Differences between test and reference data were counted as errors in the test data, which might slightly overestimate the true error rates of our test data.

For the error calculations, we assigned each test contig to a reference transcript according to two criteria: 1) the test contig had a BlastN e-value ≤ 1E−06 against a reference transcript and 2) the test contig was ≥100 bp. We did not impose any filter for percent similarity between test contig and reference transcript. Furthermore, we included all test contigs mapping to a single reference transcript that passed these criteria. In cases where the Blast analysis indicated there were no assembly errors (i.e., cases in which the Blast hit length was equal to either the test contig length, the reference transcript length, or was bounded by the start/end of test contig/reference transcript), the substitution and indel error rates were calculated automatically by parsing the BlastN output. For all other cases, we aligned each test contig to its reference transcript using ClustalW (Chenna et al. 2003) and manually recorded substitution, indel, and assembly errors. To evaluate the effect of *k*-mer length, number of lanes used, and test contig length cutoff, error rates were calculated for six different assemblies (one lane, test contig length ≥ 300 bp, *k*-mer = 21; two lanes, test contig length ≥ 300 bp, *k*-mer = 21; four lanes, test contig length ≥ 300 bp, *k*-mer = 21; four lanes, test contig length ≥ 300 bp, *k*-mer = 19; four lanes, test contig length ≥ 300 bp, *k*-mer = 23; four lanes, test contig length ≥ 100 bp, *k*-mer = 21).

We mapped sequence reads to the coding portions of their respective nonredundant reference transcriptomes using the RMAPQ software with a quality filter of five, allowing up to three mismatches, and retaining only unambiguously mapped reads (Smith et al. 2008). We determined the distributions of the test contig mapping positions across three equally sized segments of their corresponding reference transcripts by calculating the midpoint of each test contig and identifying the corresponding midpoint in the reference transcript. We generated the expected frequencies of test contig positions using the following equations: middle-third segment = $(L/3)/(L - l)$ and 5′ third and 3′ third segment = $((L/3) - (l/2))/(L - l)$, where L = reference transcript and l = test contig length (Huntley and Clark 2007). We then averaged the probabilities for each test contig and multiplied them by the total number of test contigs for each species. *G*-test goodness of fit was used to assess whether observed frequencies deviated from expected (Sokal and Rohlf 1995).

Assessment of Functional Content

We plotted *A. gambiae* test contig coverage values (calculated by the Velvet software, Zerbino and Birney 2008) against the microarray expression values (retrieved from the University of California Irvine *A. gambiae* Gene Expression Profile database, Dissanayake et al. 2006). Microarray expression values were calculated by averaging expression scores of male and female non–blood-fed mosquitoes. Both microarray expression and next-generation sequencing transcript coverage values were normalized via base-10 logarithm transformations. The relationship between the two sets of values was quantified using Pearson's correlation (Sokal and Rohlf 1995).

We examined whether proteins belonging to different functional categories (FunCat) in the MIPS PEDANT3 database (http://pedant.gsf.de/index.jsp) (Riley et al. 2007) were equally represented in the test contig data relative to the reference proteome. We also examined the relative proportions of transcription factors (dbd.mrc-lmb.cam.ac.uk/DBD, Wilson et al. 2008), ribosomal proteins (ribosome.miyazaki-med.ac.jp, Nakao et al. 2004), odorant receptors (Hill et al. 2002), immunity-related proteins (www.vectorbase.org, Christophides et al. 2002), proteins containing the major insecticide-resistance domains (i.e., carboxylesterases [IPR002018]; cytochrome P450s [IPR001128]; and glutathione transferases [IPR004045, IPR004046]; all obtained from Interpro: www.ebi.ac.uk/interpro) (Ranson et al. 2002; David et al. 2005; Hunter et al. 2009), as well as proteins containing putative signal peptides (Bendtsen et al. 2004), transmembrane helix domains (Krogh et al. 2001), and glycophosphatidylinositol (GPI) anchors (Eisenhaber et al. 2004) between the test contig data and the reference proteomes of the two species. Finally, we identified all microsatellites with repeat unit sizes between 2 and 6 bp in the *A. aegypti* and *A. gambiae* reference transcripts using the EMBOSS ETANDEM software (Rice et al. 2000). Microsatellites with repeat units whose sequence was ≥90% similar and whose total sequence length was ≥24 bp were considered genuine. For each microsatellite-containing reference transcript, we then identified all test contigs that mapped to it and examined whether the actual microsatellite sequence was covered by a test contig. We assessed deviations from randomness by Fisher's exact test, implementing a multiple test corrected P value of 0.0056 (Sokal and Rohlf 1995).

Assessment of Evolutionary Content

To investigate the accuracy of ortholog prediction between test contig data from one species and full proteome data from a second one, we first identified all ortholog pairs between species with full proteome data as well as between *A. aegypti*/*A. gambiae* test contigs and a number of insect species with full proteome data. The flowchart of our experimental design is shown in figure 1. Briefly, by assuming that all ortholog pairs identified from comparisons of species with full proteome data were correct, we calculated the number of true positives and false positives for all test contigs that were predicted to be ortholog pairs to insect proteins (fig. 1). Further, we identified test contigs that

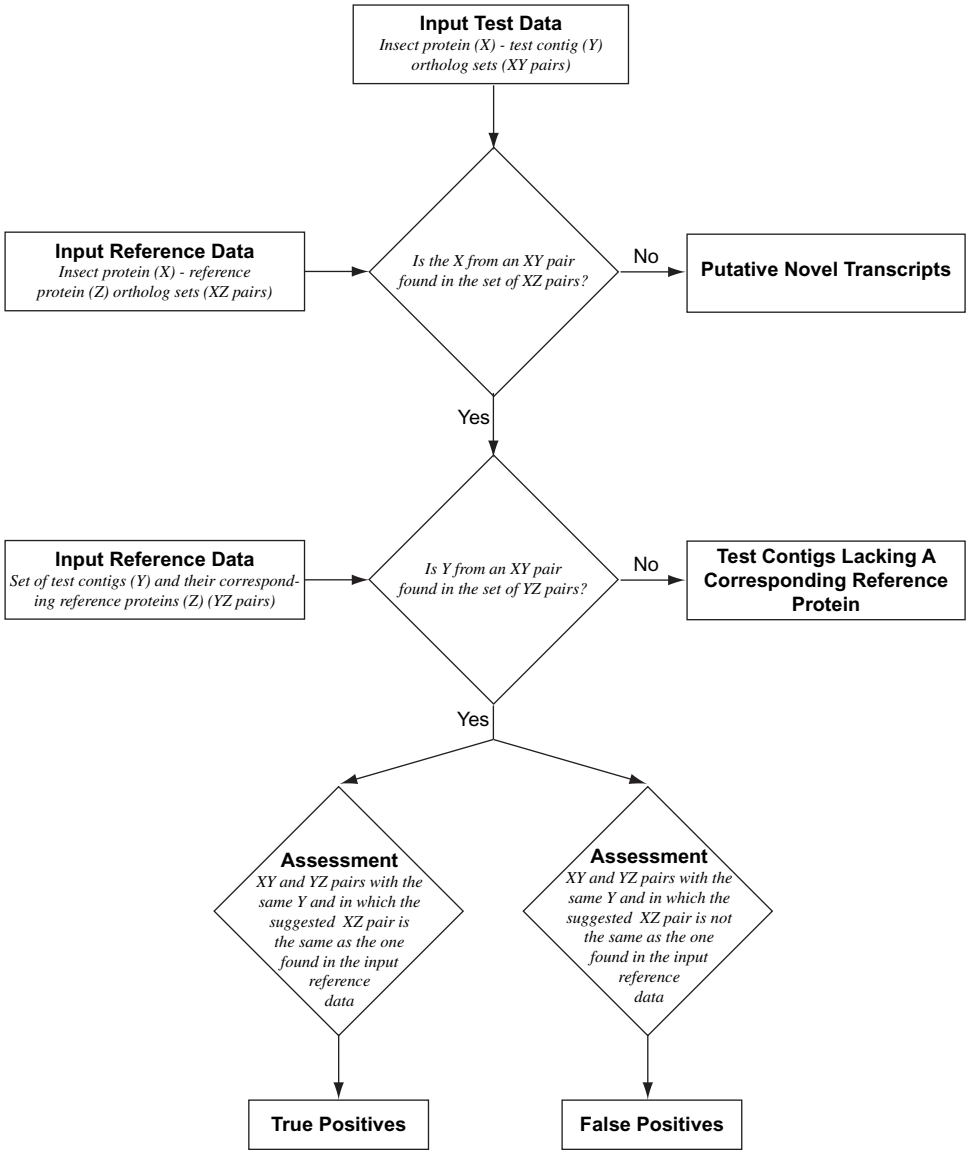

FIG. 1.—Experimental design flowchart for investigating the accuracy of ortholog prediction between test contig data from one species and full proteome data from a second one. Orthologs were calculated by comparing a number of insect proteomes (*Aedes aegypti*, *Culex pipiens*, *Anopheles gambiae*, *Drosophila melanogaster*, *Bombyx mori*, *Apis mellifera*, *and Tribolium castaneum*) against *A. aegypti/A. gambiae* test contigs or against *A. aegypti/A. gambiae* reference proteins. The sets of orthologs identified in comparisons between insect and reference proteins as well as the set of test contigs and their corresponding reference proteins were assumed to be correct and were used to evaluate the accuracy of the predicted set of orthologs between insect proteins and test contigs.

either failed to map to a reference protein or those that mapped to unpaired reference protein and therefore may represent putative novel transcripts. All analyses were conducted using two different cutoffs for minimum test contig length (test contigs $\geq$100 and $\geq$ 300 bp).

We next investigated the accuracy of ortholog prediction between the two test contig data sets from *A. aegypti* and *A. gambiae*. Assuming that all ortholog pairs of reference proteins were correct, we calculated the number of true positives, false positives, false negatives, as well as lineage-specific transcripts for all test contigs that mapped to reference proteins. The flowchart of our experimental design is shown in supplementary figure 1, Supplementary Material online.

All orthology calculations were done using the reciprocal best Blast hit algorithm (Koonin 2005), with an e-value cutoff of 1E−06, on two test contig data sets (test contigs $\geq$100 and $\geq$300 bp). Our assumption of the validity of all ortholog pairs identified in comparisons between species with full proteomes is not likely to be strictly true, so its implementation should result in conservative ortholog prediction error rate estimates for our test contigs.

Results

Assessment of Data Quantity

Four "lanes" of Solexa-sequencing data from nonnormalized cDNA libraries from *A. aegypti* and *A. gambiae*

generated 22,648,046 (average raw base quality score: 37/40) and 21,292,304 reads (average raw base quality score: 29/40), respectively (table 1, supplementary table 1, Supplementary Material online). De novo assembly of reads from all four lanes generated over 21,000 contigs ≥100 bp and a total of 4.1–4.5 Mbp of assembled sequence data per species (at 6–8× median coverage; fig. 2*A* and *B*, table 1). The fraction of total reads used in each ≥100-bp assembly was 7.6% (60.5 Mbp) and 8.1% (61.2 Mbp) for *A. aegypti* and *A. gambiae*, respectively (supplementary table 1, Supplementary Material online). The number of test contigs in the unfiltered data set was substantially larger. Specifically, 99,020 *A. aegypti* and 95,747 *A. gambiae* test contigs were recovered, yielding 10.5 and 10.1 Mbp of assembled sequence data, respectively (table 1, supplementary table 1, Supplementary Material online). The total amount of sequence reads that contributed to the assemblies was 102 and 110 Mbp for *A. aegypti* and *A. gambiae*, respectively. However, we excluded all test contigs <100 bp from further analyses because their short length and typically low coverage (54-bp median length with 3.5× median coverage for *A. aegypti*; 57-bp median length with 4.5× median coverage for *A. gambiae*) is likely to render them useless for most applications.

Numbers and total amounts of assembled test contigs were highly dependent on the starting number of reads used as input to the Velvet assembler, as evidenced by the approximate doubling of their numbers in analyses of reads from 1, 2, and 4 lanes (table 1). Importantly, the number and sequence amount of test contigs did not plateau as reads from additional lanes were used. Assuming this linear relationship between generated and assembled data holds true for larger amounts of data, one might extrapolate that the entire transcriptome would be covered with four times more data.

Numbers and total sequence amounts of assembled test contigs were also highly dependent on *k*-mer length (table 1). There was a positive relationship between *k*-mer length and median test contig coverage depth, and between *k*-mer length and average test contig length. Conversely, there was a negative relationship between *k*-mer length and test contig number. These data suggest that increasing *k*-mer length results in fewer but longer and more deeply sequenced test contigs, a result matching theoretical expectations (Zerbino and Birney 2008). Assemblies that yielded the greatest number of test contigs ≥300 bp were the *k*-mer = 21, 4-lane *A. aegypti* assembly (2,444 contigs ≥300 bp) and the *k*-mer = 23, 4-lane *A. gambiae* assembly (1,671 contigs ≥300 bp) (table 1, supplementary table 1, Supplementary Material online). These were used in all subsequent analyses unless otherwise noted. Assemblies generated using different *k*-mer lengths had very similar substitution, indel, and assembly error rates (see the "assessment of data quality" section).

The *A. aegypti* test contigs mapped to ~27% of the species reference transcriptome sequence. Additionally, ~ 29% of reference proteins were mapped by at least one test contig. However, multiple test contigs frequently mapped to the same reference transcript, albeit without any overlap. In contrast, the *A. gambiae* test contigs were mapped much less frequently to their reference transcripts,

Table 1
Summary Statistics of Test Contigs from Select Assemblies Using Varying Amounts of Solexa Sequenced Read Data and Assembly Parameters for *Aedes aegypti* and *Anopheles gambiae*, Respectively

	Assembly Statistics				All Test Contig Set				≥100-bp Test Contig Set				≥ 300-bp Test Contig Set			
Species	Lane #	Read #	ABQS	*k*-mer	Ctg #	Raw Amount	Amount	Med L	Med C	Ctg #	Raw Amount	Amount	Med L	Med C	Ctg #	Raw Amount	Amount	Med L	Med C
A. aegypti	1	5,745,510	37	21	27,930	25.99	2,744,396	77	3	4,769	15.87	988,983	167	6	530	7.11	252,490	420	14
A. aegypti	2	11,465,769	37	21	56,494	52.67	5,675,901	78	3	10,448	31.36	2,157,542	168	6	1,123	12.07	529,649	414	13
A. aegypti	4	22,648,046	37	19	137,864	108.16	13,113,033	74	4	26,534	54.11	5,228,532	165	6	2,379	15.49	1,052,360	397	11
A. aegypti	4	22,648,046	37	21	99,020	105.06	10,459,561	81	4	21,653	60.45	4,541,793	170	6	2,444	23.00	1,165,278	414	13
A. aegypti	4	22,648,046	37	23	68,157	92.79	7,803,582	87	4	16,303	61.77	3,555,687	173	7	2,016	28.87	1,010,292	422	14
A. gambiae	1	6,053,469	28	25	18,182	23.58	2,061,338	91	5	4,185	15.99	859,021	169	9	966	9.45	340,376	304	15
A. gambiae	2	12,101,924	28	23	61,359	66.98	6,426,363	85	5	12,826	39.98	2,508,624	163	8	1,008	12.14	461,086	405	17
A. gambiae	4	21,292,304	29	21	141,562	120.25	13,401,535	78	5	25,919	55.00	4,813,707	160	7	1,551	10.91	664,058	388	13
A. gambiae	4	21,292,304	29	23	95,747	110.04	10,128,107	85	5	21,193	60.63	4,157,303	165	8	1,671	18.05	753,038	400	16
A. gambiae	4	21,292,304	29	25	61,632	91.74	7,075,880	92	6	15,401	56.79	3,121,465	169	9	1,387	20.15	646,691	411	19

Summary statistics of test contigs for All, ≥100- and ≥300-bp categories are shown. Abbreviations: "Lane #": Number of Solexa-sequencing lanes used as input in the assembly; "Read #": Number of Solexa sequence reads used as input in the assembly; "ABQS": Average base quality score in a Solexa sequence read; "*k*-mer": required length of identical match between two sequence reads by the Velvet software (Zerbino and Birney 2008); "Ctg #": number of test contigs produced by the assembly; "Raw Amount": Amount of total sequence contributing to the assembly (in Mbp); "Amount": Amount of total sequence in test contigs; "Med L": median length of test contigs; and "Med C": median coverage depth of test contigs. Summary statistics for all assemblies can be found in supplementary table 1, Supplementary Material online.

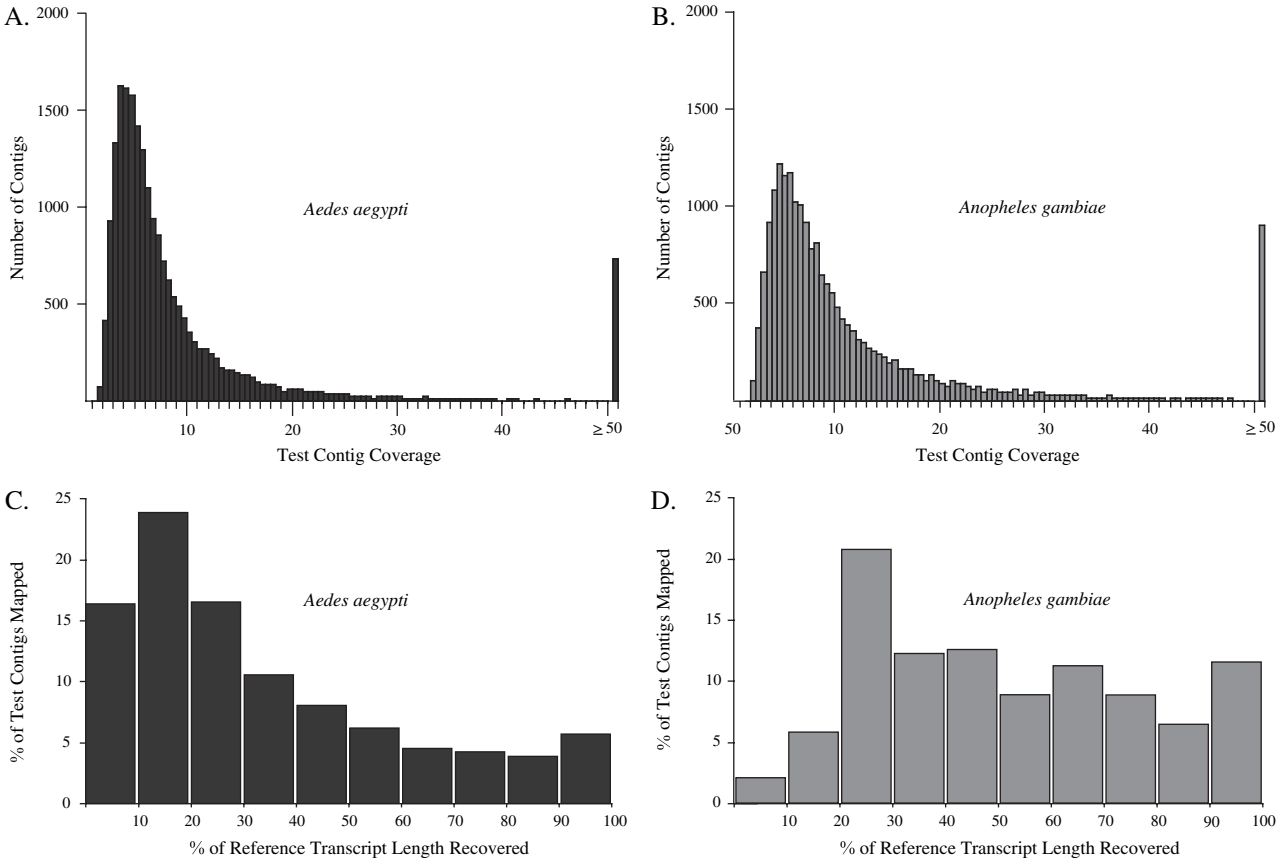

FIG. 2.—Distribution of test contig coverage values (*A*,*B*) and reference transcript length recovery (*C*,*D*) in the *Aedes aegypti* (*A*,*C*) and *Anopheles gambiae* (*B*,*D*) ≥100-bp test contig sets. In panels *A* and *B*, the *X* axis shows coverage values and the *Y* axis the total number of contigs. In panels *C* and *D*, the *X* axis shows the percentages of reference transcript lengths recovered by the test transcripts and the *Y* axis the percentage of test contigs that mapped reference transcripts. Note that the *A. gambiae* data set in panel *D* is significantly smaller than the *A. aegypti* data set in panel *C*. The difference is explained by considering the different test and reference *A. gambiae* strains used (the *A. aegypti* strain used in reference and test data generation was the same as well as highly inbred), and the lower overall average base quality score of the *A. gambiae* test data relative to the *A. aegypti* test data (table 1).

a likely byproduct of sequence variants between the strain used in this study and the extinct genome reference strain. For each reference transcript mapped by test contigs, we found that an average of 34% and 42% (with medians of 25% and 50%) of their length was covered by test contigs in *A. aegypti* and *A. gambiae* (fig. 2*C* and *D*; the transcript ids of all reference transcripts recovered by our test contigs are provided in supplementary table 2, Supplementary Material online), respectively. The mapping distributions from both species are positively skewed ($N = 4505$, $g_1 = 0.996$; $N = 385$, $g_1 = 0.659$), indicating that the majority of the reference transcripts had ≤50% of their length mapped by test contigs (fig. 2). The total number of uniquely mapped reads against the *A. aegypti* and *A. gambiae* reference transcriptomes is shown in supplementary table 3, Supplementary Material online.

Assessment of Data Quality

Substitution, indel, and assembly error rates were relatively robust to *k*-mer length (19, 21, or 23) and amount of input data (reads from 1, 2, or 4 lanes) (table 2). The only parameter that affected substitution and indel error rates was minimum test contig length. Both error rates significantly increased in the data set consisting of test contigs ≥100 bp relative to their values in data sets consisting of test contigs ≥300 bp (substitutions: $p = 1.5 \times 10^{-46}$; indels: $p = 2.6 \times 10^{-6}$). This is likely explained by median coverage depth differences (6× for the ≥100-bp data set; 11–14× for the ≥300-bp data sets), as error-free test contigs from the ≥100-bp data set have a significantly higher average coverage depth relative to all other test contigs ($P = 0.0013$). Assembly errors also did not significantly differ between assemblies (table 2).

Both species displayed a nonrandom distribution of test contigs across the length of their corresponding reference transcripts (*A. aegypti*: df = 2, $g = 245.64$, $p = 4.56e{-}54$; *A. gambiae*: df = 2, $g = 16.68$, $p = 2.39e{-}04$, respectively), with a significant overrepresentation of test contigs in the 3′ third of reference transcripts (fig. 3). These findings are in agreement with previous work showing that the 3′ ends of transcripts are more common in Solexa libraries made with poly-A+ RNA (Nagalakshmi et al. 2008).

Assessment of Functional Content

Previous studies have identified a positive correlation between transcript expression levels measured by

Table 2
Substitution, Indel, and Assembly Error Rates

Assembly				Error Rates					
Lanes	*k*-Mer	Sequence (bp)	Contigs (bp)	Substitutions	Substitution Error Rate	Indels	Indel Error Rate	Assembly Errors	Assembly Error Rate
1	21	162,798	≥300	449	2.76E−03	9	5.53E−05	38	2.33E−04
2	21	334,703	≥300	940	2.81E−03	29	8.66E−05	70	2.09E−04
	21	3,212,803	≥100	12,210	3.80E−03	297	9.24E−05	623	1.94E−04
4	19	750,858		2,163	2.88E−03	57	7.59E−05	152	2.02E−04
	21	801,979	≥300	2,216	2.76E−03	34	4.24E−05	146	1.82E−04
	23	583,658		1,587	2.72E−03	47	8.05E−05	117	2.00E−04

All error rates were calculated in six *Aedes aegypti* assemblies varying in amount of data used (number of lanes column) and assembly parameters (*k*-mer length column) by comparing test contigs to their reference transcripts.

microarrays and those measured by next-generation sequencing (‘t Hoen et al. 2008; Wilhelm et al. 2008). To test whether this was the case for our data, we plotted the $\log_{10}$ expression scores of adult male and female non–blood-fed *A. gambiae* (strain Pink-eye) from a recent microarray study (Dissanayake et al. 2006) against the $\log_{10}$ coverage values of their corresponding test contigs (fig. 4) for all 3,145 reference *A. gambiae* transcripts mapped by our ≥100-bp test contig data set. However, there were no expression values in the microarray database for 723 of the 3,145 transcripts, leaving us with a total of 2,422 transcripts for which we had both coverage and microarray expression values. We found a strong positive association between microarray expression and Solexa coverage (Pearson’s $r = 0.62$) (fig. 4). Interestingly, we found that 660 of 2,422 (27%) transcripts with both coverage and microarray expression values were included in the top 10% most highly expressed transcripts from the microarray study.

It has previously been reported that Solexa sequencing is slightly biased against sequencing AT-rich genomic regions (Hillier et al. 2008). To test whether this bias was present in our data, we independently plotted microarray expression values and test coverage values against reference transcript AT content. Both comparisons yielded very weak associations with AT content ($r = 0.04$ for microarray expression values and $r = 0.10$ for test contig coverage values, respectively).

We assessed the functional information content of our test contigs through several different analyses. We first examined the proportion of each major functional category according to the FunCat scheme (Ruepp et al. 2004) in the test contigs and reference proteins of *A. aegypti* and *A. gambiae*, respectively. Test contigs were significantly overrepresented in the Energy (*A. aegypti*, ≥100-bp test contigs, $p = 8.00e{-}05$; *A. aegypti*, ≥300-bp test contigs, $p = 2.60e{-}14$; *A. gambiae*, ≥100-bp test contigs, $p = 4.00e{-}08$; *A. gambiae*, ≥300-bp test contigs,

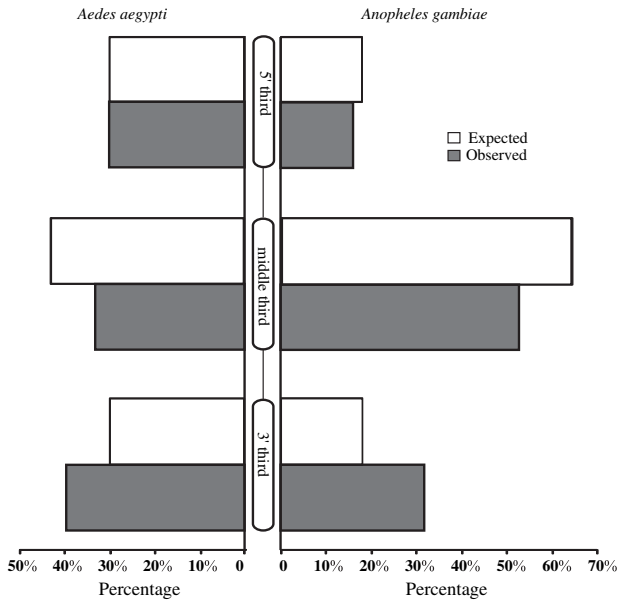

FIG. 3.—Distribution of test contig positions relative to their reference transcripts. The *X* axis represents the percentage of total test contigs whose midpoint was located in a given reference transcript region (5′ third, middle-third, 3′ third), whereas the *Y* axis represents the three reference transcript segments to which the test contigs were mapped. White and gray bars represent the expected and observed distributions of test contig positions, respectively. The left panel shows data from *Aedes aegypti* ($p = 4.56e{-}54$), whereas the right panel data from *Anopheles gambiae* ($p = 2.39e{-}04$).

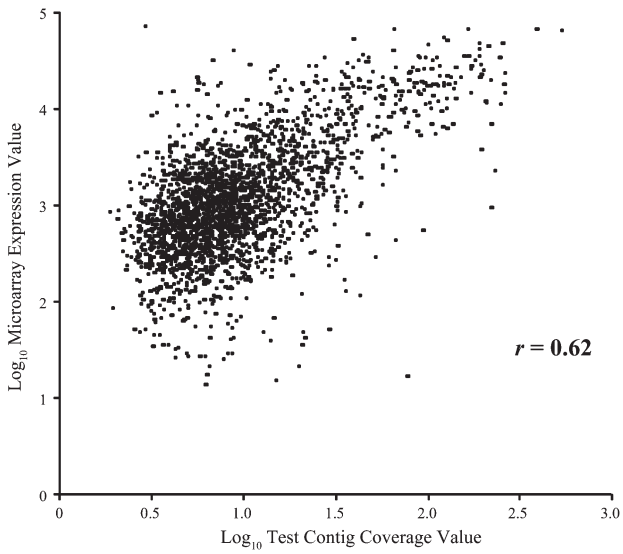

FIG. 4.—Scatter plot of the relationship between microarray transcript expression and next-generation sequencing coverage. *Anopheles gambiae* $\log_{10}$ microarray expression values (*Y* axis) are plotted against their corresponding $\log_{10}$ test contig coverage values (*X* axis). Note the strong association between the two variables (Pearson’s $r = 0.62$).

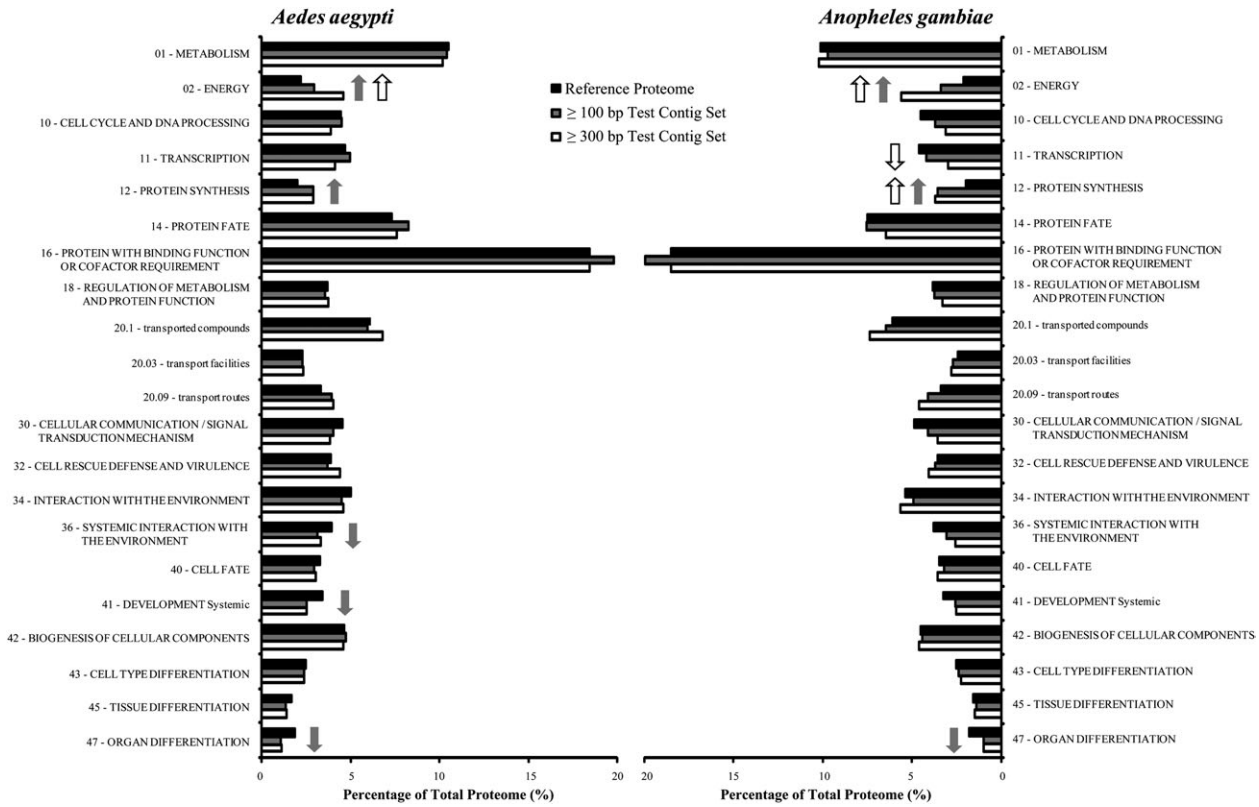

FIG. 5.—Functional classification of reference proteomes and test contigs according to the FunCat scheme. FunCat category numbers and corresponding names are shown on the *Y* axis, whereas the percentages of the contribution of each data set to each FunCat category are shown on the *X* axis. Black, gray, and white bars represent the reference proteome, ≥100-bp test contig set, and ≥300-bp test contig set, respectively. Upward and downward gray and white arrows show statistically significant overrepresentation or underrepresentation in a given FunCat category in the ≥100-bp test contig set and in the ≥300-bp test contig set, respectively. The left and right panels refer to the *Aedes aegypti* and *Anopheles gambiae* comparisons, respectively. Note that in cases where only some of the test contig data sets/species are significantly different from the reference ones, the other test contig data set(s)/species often show similar, albeit nonsignificant responses.

$p = 7.30e{-}18$), and Protein Synthesis categories (*A. aegypti*, ≥100-bp test contigs, $p = 5.20e{-}06$; *A. gambiae*, ≥100-bp test contigs, $p = 9.40e{-}12$; *A. gambiae*, ≥300 test contigs, $p = 2.23e{-}06$) (fig. 5). In contrast, test contigs were significantly underrepresented in the Organ Differentiation (*A. aegypti*, ≥100-bp test contigs, $p = 9.16e{-}08$; *A. gambiae*, ≥100-bp test contigs, $p = 2.50e{-}06$), Systemic Interaction With The Environment (*A. aegypti*, ≥100-bp test contigs, $p = 8.00e{-}04$), Systemic Development (*A. aegypti*, ≥100-bp test contigs, $p = 3.00e{-}05$), and Transcription (*A. gambiae*, ≥300-bp test contigs, $p = 4.00e{-}04$) categories (fig. 5).

We then examined whether test contigs were enriched in motifs suggestive of roles in cell secretion and cell-membrane localization functions, by comparing test contigs that contained putative signal peptides, transmembrane helices, and GPI anchors with their corresponding reference proteins. Signal peptides were significantly underrepresented only in one test contig data set (*A. aegypti*, ≥100-bp test contigs, $p = 2.00e{-}05$), whereas transmembrane helices and GPI anchors were not significantly different in any test contig data set (table 3).

We finally examined whether additional classes of proteins and protein families of evolutionary interest (transcription factors, ribosomal proteins, odorant receptors, immunity-related proteins, and proteins families involved in insecticide resistance) were present in our test contigs in proportion to their representation in the reference proteomes. Ribosomal proteins were overrepresented in all data sets examined, whereas odorant receptor and transcription factors were underrepresented in almost all data sets examined (table 3). In contrast, the proportion of proteins involved in innate immunity and insecticide resistance were not significantly different between test contig and reference proteome data sets with one exception (the test contigs associated with insecticide resistance in the *A. aegypti* ≥100-bp test contig data set were underrepresented) (table 3).

Next-generation DNA-sequencing technologies have been highly successful in detecting alternatively spliced transcripts (Carninci 2008; Vera et al. 2008; Tang et al. 2009). We also found several cases of alternatively spliced reference transcripts that were uniquely mapped by test contigs (fig. 6). Specifically, we detected 11 unambiguous cases of test contigs matching alternatively spliced transcripts in *A. aegypti* and 20 cases in *A. gambiae*.

Microsatellites contained in protein-coding regions are readily transferred between species (Ellis and Burke 2007; Gibbons and Rokas 2009). We identified 40 and 550 microsatellites in the reference transcripts of *A. aegypti* and

Table 3
The Functional Information Content of Test Contigs from *Aedes aegypti* and *Anopheles gambiae* Transcriptomes, Respectively

			Immunity-Related Transcripts			Insecticide Resistance Transcripts			Odorant Receptors		
	Data Set	Transcript #	+	Proportion	Fisher's Exact *P*	+	Proportion	Fisher's Exact *P*	+	Proportion	Fisher's Exact *P*
A. aegypti	Reference transcript set	16781				271	0.016		131	0.008	
	≥100-bp test contig set	4822	N/A			41	0.009	**5.00E−05 ↓**	0	0.000	**5.70E−15 ↓**
	≥300-bp test contig set	1325				11	0.008	0.028	0	0.000	**7.00E−05 ↓**
A. gambiae	Reference transcript set	13133	187	0.014		190	0.014		79	0.006	
	≥100-bp test contig set	3145	32	0.010	0.084	30	0.010	0.031	0	0.000	**6.77E−08 ↓**
	≥300-bp test contig set	834	8	0.010	0.359	12	0.014	1.000	0	0.000	0.015
			Ribosomal Transcripts			Transcription Factors			Microsatellite-Containing Transcripts		
			+	Proportion	Fisher's Exact *P*	+	Proportion	Fisher's Exact *P*	+	Proportion	Fisher's Exact *P*
A. aegypti	Reference transcript set	16781				928	0.055		40	0.002	
	≥100-bp test contig set	4822	N/A			131	0.027	**2.70E−17 ↓**	14	0.003	0.514
	≥300-bp test contig set	1325				18	0.014	**3.90E−14 ↓**	6	0.005	0.147
A. gambiae	Reference transcript set	13133	80	0.006		601	0.046		505	0.038	
	≥100-bp test contig set	3145	36	0.011	**0.003 ↑**	52	0.017	**2.50E−16 ↓**	140	0.045	0.127
	≥300-bp test contig set	834	36	0.043	**5.40E−17 ↑**	9	0.011	**3.74E−08 ↓**	31	0.037	0.632
			Signal Peptides			Transmembrane Helices			GPI Anchors		
			+	Proportion	Fisher's Exact *P*	+	Proportion	Fisher's Exact *P*	+	Proportion	Fisher's Exact *P*
A. aegypti	Reference transcript set	16781	3263	0.194		3626	0.216		97	0.006	
	≥100-bp test contig set	4822	808	0.168	**2.00E−05 ↓**	990	0.205	0.111	32	0.007	0.524
	≥300-bp test contig set	1325	275	0.208	0.250	267	0.202	0.224	12	0.009	0.139
A. gambiae	Reference Transcript Set	13133	2593	0.197		2971	0.226		86	0.007	
	≥100-bp test contig set	3145	571	0.182	0.045	702	0.223	0.740	27	0.009	0.231
	≥300-bp test contig set	834	180	0.216	0.195	162	0.194	0.032	7	0.008	0.506

The relative proportions of several protein families and functional domains or patterns were compared between the reference and test data sets of the two species. "+" indicates the presence of a protein, domain or pattern. "Proportion" refers to the relative proportion of proteins in each data set. Bold data represent statistically significant over (↑) or underrepresentations (↓) of proteins relative to the reference data set (at $P = 0.0056$).

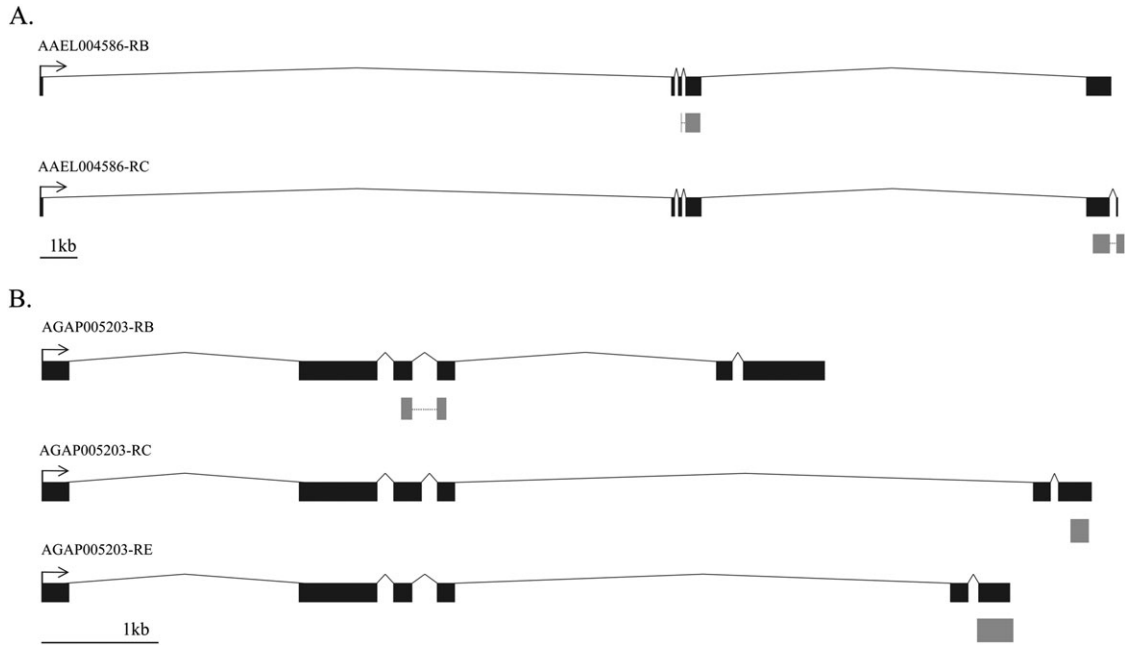

FIG. 6.—Examples of alternatively spliced transcripts identified within the *Aedes aegypti* (*A*) and *Anopheles gambiae* (*B*) test contig data sets. Arrows indicate direction of transcription, with reference transcript IDs above. Black and gray rectangles represent reference and test exons, respectively. Lines connecting exons signify introns in the reference transcripts. Dotted lines connecting test exons indicate that the sequence of test contig retrieved flanked an intron.

A. gambiae, respectively. The substantially lower number of microsatellites in *A. aegypti* is in agreement with results from several other studies (e.g., Chambers et al. 2007, and references therein). Surprisingly, we found that the proportion of test contigs mapping to microsatellite-containing reference transcripts was not significantly different from the overall proportion of reference transcripts that contain microsatellites. However, no microsatellites were completely covered by our test contigs.

Assessment of Evolutionary Content

Partial transcriptome data sets are frequently used for ortholog prediction (de la Torre et al. 2006; Hughes et al. 2006; Dunn et al. 2008), either using partial data from one species and full transcriptome/proteome data from a second one (first scenario), or using partial data sets from both species (second scenario). In the first scenario, we found that ortholog prediction accuracy was insensitive to increasing evolutionary distance, although the numbers of predicted ortholog pairs did decrease with evolutionary distance (fig. 7, supplementary table 4, Supplementary Material online). Accuracy was significantly higher for ortholog pairs constructed using the $\geq$300-bp *A. aegypti*/*A. gambiae* test contig data set than when using the corresponding $\geq$100-bp test contig data set (*A. aegypti*: $\geq$100-bp test contigs = 64%, $\geq$300 bp = 72%, $p = 1.7 \times 10^{-27}$; *A. gambiae:* $\geq$100 bp = 68%, $\geq$300 bp = 75%, $p = 7.4 \times 10^{-18}$) (fig. 7). Furthermore, the percentage of false positives was relatively low in both data sets. A number of insect proteins with predicted orthologs in the *A. aegypti*/*A. gambiae* test contigs did not have predicted orthologs to *A. aegypti*/*A. gambiae* reference proteins and may likely represent putative novel or unannotated proteins (fig. 7). Finally, a number of test contigs did not map to any reference proteins from their species' proteome (fig. 7).

In the second scenario, where ortholog prediction was performed entirely with test contigs, accuracy was significantly higher for ortholog pairs constructed using the 300-bp test contig data sets of *A. aegypti* and *A. gambiae* than their 100-bp test contig data sets. Forty-nine percent of predicted ortholog pairs were true positives in the $\geq$300-bp test contig data sets compared with $\sim$25% in the 100-bp test contig data sets. Importantly, in both analyses, the percentage of false positives was small but nontrivial (6% and 3%, respectively), whereas a very large percentage of ortholog pairs found using the reference proteome data was not predicted by the test contig data and was classified as false negatives (45% for the 300-bp comparison and 73% for the 100-bp comparison).

Discussion

We investigated the potential utility of short-read DNA sequencing for evolutionary studies of nonmodel organisms. We evaluated the quality of sequences of nonnormalized cDNA libraries of two mosquito vectors, *A. aegypti* and *A. gambiae*, by comparing them with the reference genome sequences of these species (Holt et al. 2002; Nene et al. 2007). Similarly, we compared our sequence data with proteomes of diverse insects to evaluate its utility for evolutionarily distant comparisons. We found that our short-read DNA sequence data were large in quantity and high in quality and of sufficient diversity to be of functional and evolutionary interest. We will discuss our findings in the context of the four key parameters we set out to

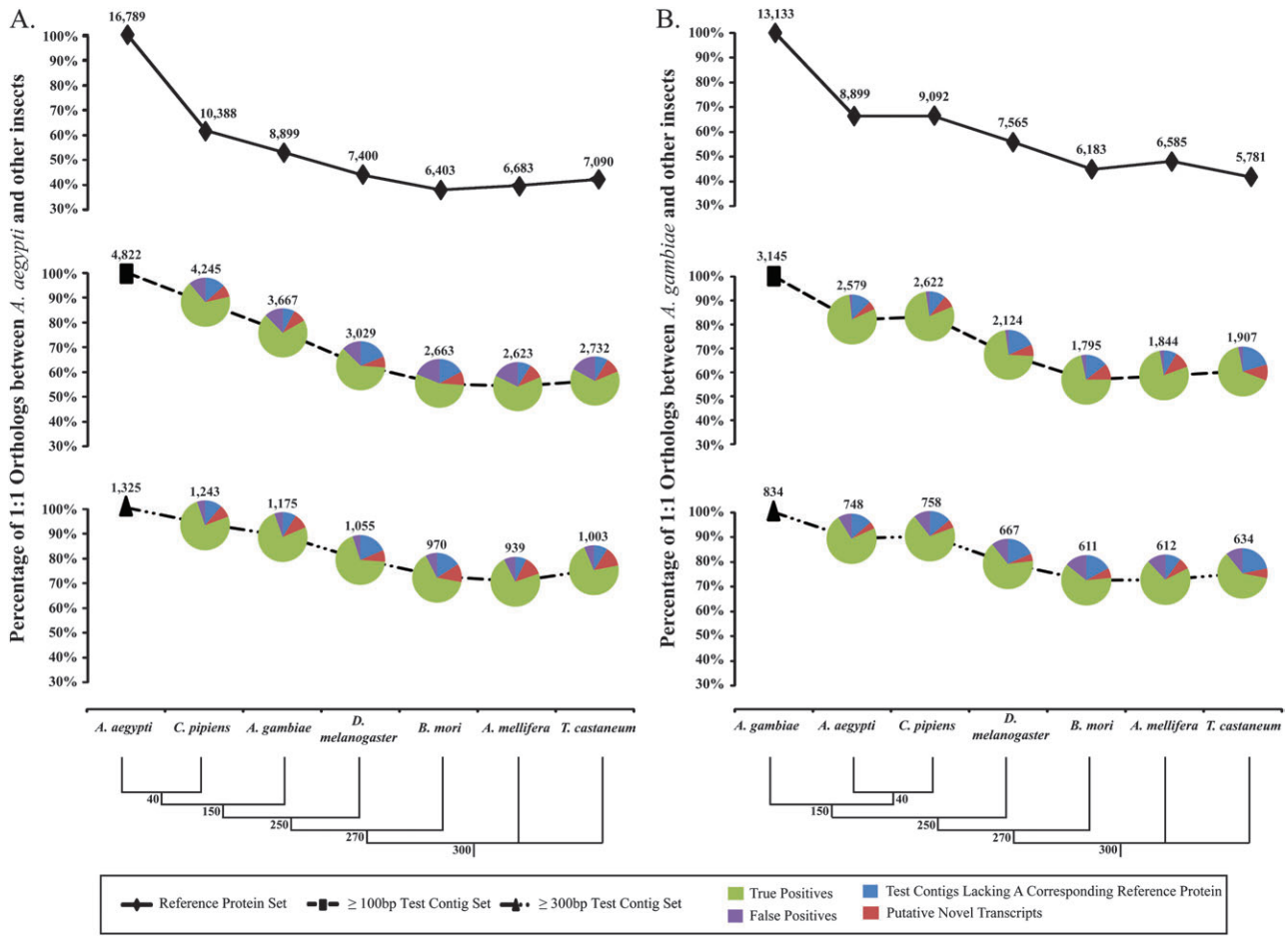

FIG. 7.—Accuracy of ortholog prediction between test contig data from *Aedes aegypti* (panel *A*)/*Anopheles gambiae* (panel *B*) and full proteome data from several insects. Percentages of 1:1 ortholog pairs recovered when *A. aegypti*/*A. gambiae* reference proteins are used are shown in the upper graphs, whereas percentages of pairs recovered when two different test contigs data sets from *A. aegypti*/*A. gambiae* are used are shown in the middle (≥100-bp test contig set) and lower (≥300-bp test contig set) graphs, respectively. For the middle and lower graphs, the percentages of true positives, false positives as well as of putative novel test contigs and test contigs lacking a reference protein are shown as pie charts (for a description of how the percentages in each category are calculated see figure 1 and for their actual numbers see supplementary table 4, Supplementary Material online). The total number of predicted *A. aegypti*/*A. gambiae* test contig—insect protein ortholog pairs are shown above each pie chart. At the bottom of each panel, a phylogenetic tree depicting the evolutionary relationships between the insects used in these calculations is shown, with numbers corresponding to divergence in million years (all dates from Grimaldi and Engel 2005).

evaluate—quantity, quality, functional, and evolutionary information content—and in comparison to similar analysis using long-read next-generation sequence data.

Data Quantity

We recovered a total of 4–4.5 Mbp (at 6–8× median coverage) of assembled sequence data per species, with *Aedes* test contigs matching ~27% of the *Aedes* reference transcriptome sequence. These values are similar to the assembled sequence data from long-read transcriptome sequencing (9.5 Mbp at 3× median coverage) (Vera et al. 2008). The relatively low fraction of the total transcriptome recovered reflects the large size of the transcriptomes we examined and the relatively small number of reads generated, because deeper short-read resequencing of the smaller yeast transcriptome accurately detected more than 90% of all transcripts (Nagalakshmi et al. 2008; Wilhelm et al. 2008). In the near future, the use of paired-end sequencing (already available for both short- and long-read next-generation technologies), the development of assembly algorithms that take advantage of base quality scores and longer read lengths promise to significantly increase the proportion of the transcriptome recovered at depths of sequence similar to what we obtained.

Data Quality

Substitution error rate in our test data (~0.25%) was robust to varying assembly parameters and amount of data used. This value is likely an overestimate of the actual error rate because in all calculations we assumed that the reference data are error-free, which is certainly not the case (Holt et al. 2002; Nene et al. 2007). Similar error rates have been estimated from de novo assembly data of Solexa-generated short reads of bacterial genomes (0.33%, Farrer et al. 2009), and from unassembled 454/Roche long reads (0.25–0.5%, Huse et al. 2007). In contrast, Wicker et al. estimated the

substitution error rate for de novo assembled long-read sequenced bacterial artificial chromosomes at approximately 0.07% (Wicker et al. 2006), but they did not include additional errors in homopolymer runs.

Functional Information Content

Transcriptome sequencing is frequently used to provide greater insight into the organism's biology (King et al. 2003; Nichols et al. 2006). What information did we retrieve from our test data about the distinctive biology of mosquitoes? It is well established that protein families involved in basic cellular functions are typically highly expressed (Holstege et al. 1998), and this is evidenced in the functional enrichment of our test contigs (fig. 5). We examined a number of other protein families and features that are relevant to mosquito biology (e.g., involved in insecticide resistance, olfaction, and immunity) or to the biology of nonmodel organisms in general (e.g., microsatellite-containing genes, transcription factors, or motifs indicative of secretion or cell-membrane localization). We found a number of proteins that contribute to immunity or have a role in insecticide resistance, indicating that the short-read, low-cost approach we took can reveal interesting components of an organism's transcriptome. On the other hand, we did not recover many weakly expressed transcripts likely to be of particular interest to some mosquito biologists, such as odorant receptors or transcription factors. It is likely that poorly expressed transcripts could be recovered by additional sequencing, although at greater cost. Moreover, assembly of transcripts that are members of gene families might be fundamentally problematic due to the inherent difficulty of assembling similar but distinct transcripts from short reads.

An important but perhaps easily overlooked point is that next-generation DNA sequencing provides quantitative information about gene expression (Nagalakshmi et al. 2008; ′t Hoen et al. 2008; Wilhelm et al. 2008; Rokas and Abbot 2009). The strong correlation between test contig coverage depth and microarray-estimated expression levels demonstrates this point (fig. 4) and suggests an underappreciated benefit of de novo transcriptome sequencing, provided one starts from a nonnormalized cDNA library.

Evolutionary Information Content

Another use of transcriptome sequencing is for generation of novel molecular markers (e.g., Rokas et al. 2005; de la Torre et al. 2006; Hughes et al. 2006; Bouck and Vision 2007; Dunn et al. 2008). Importantly, our prediction of orthologs using test contigs was relatively accurate and the number of false positives remained small across different experiments (fig. 7). The number of orthologs retrieved was remarkably robust to increasing evolutionary distance, suggesting that one can successfully annotate (for functional and evolutionary purposes) next-generation data from nonmodel organisms that are evolutionarily very distant to model ones. However, the fraction of false positives in all comparisons was not trivial (fig. 7), highlighting the need for development of experimental approaches that explicitly deal with ameliorating their influence on downstream data analyses. False positives notwithstanding, the consistent recovery of the same sets of orthologs across species that diverged hundreds of million years ago (fig. 7) suggests that this technology may prove to be very useful for the recovery of large amounts of orthologous sequence for downstream molecular evolution studies.

Supplementary Material

Supplementary figure 1 and supplementary tables 1–4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank John Tossberg for mosquito rearing and mRNA extraction, Scott Egan for insightful comments and help with statistical analyses, Paul Howell and the MR4 facility at the CDC for providing the eggs from *Aedes aegypti*, Julian Hillier, Larry Zwiebel, Jason Pitts, Jonas King, and Tania Estevez-Laofor providing access to mosquito-rearing facilities and for advice and help with mosquito rearing, Mathias Walter and Dmitrij Frishman for support in accessing FunCat data, Bob MacCallum for assistance in retrieving data from VectorBase, Osvaldo Marinotti for providing *A. gambiae* expression data, and Jason Pitts for providing information and sequence data on mosquito olfactory receptors. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University and the Genome Technology Core at Vanderbilt University Medical Center. J.G.G. is funded by the Graduate Program in Biological Sciences at Vanderbilt University. C.T.H. is the Maclyn McCarty Fellow of the Helen Hay Whitney Foundation. This work was also supported by a Washington University in St Louis Department of Genetics Fellowship (C.T.H.) and the James S. McDonnell Foundation (M.J.). E.J. and research in P.A.'s laboratory are supported by the National Science Foundation. Research in A.R.'s laboratory is supported by the Searle Scholars Program and the National Science Foundation (DEB-0844968).

Literature Cited

Abzhanov A, Extavour CG, Groover A, Hodges SA, Hoekstra HE, Kramer EM, Monteiro A. 2008. Are we there yet? Tracking the development of new model systems. Trends Genet. 24:353–360.

Adams MD, Celniker SE, Holt RA, et al. (195 co-authors). 2000. The genome sequence of *Drosophila melanogaster*. Science. 287:2185–2195.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: signalP 3.0. J Mol Biol. 340:783–795.

Bentley DR, Balasubramanian S, Swerdlow HP, et al. (194 co-authors). 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 456:53–59.
Bouck A, Vision T. 2007. The molecular ecologist's guide to expressed sequence tags. Mol Ecol. 16:907–924.
Butler J, Maccallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 18:810–820.
Carninci P. 2008. Hunting hidden transcripts. Nat Meth. 5:587–589.
Chambers EW, Meece JK, McGowan JA, Lovin DD, Hemme RR, Chadee DD, McAbee K, Brown SE, Knudson DL, Severson DW. 2007. Microsatellite isolation and linkage group identification in the yellow fever mosquito *Aedes aegypti*. J Hered. 98:202–210.
Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31:3497–3500.
Cheung F, Haas BJ, Goldberg SM, May GD, Xiao Y, Town CD. 2006. Sequencing Medicago truncatula expressed sequenced tags using 454 life sciences technology. BMC Genomics. 7:272.
Christophides GK, Zdobnov E, Barillas-Mury C, et al. (35 co-authors). 2002. Immunity-related genes and gene families in *Anopheles gambiae*. Science. 298:159–165.
Das S, Garver L, Dimopoulos G. 2007. Protocol for mosquito rearing (*A. gambiae*). J Vis Exp. 221.
David JP, Strode C, Vontas J, Nikou D, Vaughan A, Pignatelli PM, Louis C, Hemingway J, Ranson H. 2005. The *Anopheles gambiae* detoxification chip: a highly specific microarray to study metabolic-based insecticide resistance in malaria vectors. Proc Natl Acad Sci USA. 102:4080–4084.
de la Torre JE, Egan MG, Katari MS, Brenner ED, Stevenson DW, Coruzzi GM, DeSalle R. 2006. ESTimating plant phylogeny: lessons from partitioning. BMC Evol Biol. 6:48.
Dissanayake SN, Marinotti O, Ribeiro JMC, James AA. 2006. angaGEDUCI: *anopheles gambiae* gene expression database with integrated comparative algorithms for identifying conserved DNA motifs in promoter sequences. BMC Genomics. 7.
Dunn CW, Hejnol A, Matus DQ, et al. (18 co-authors). 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 452:745–749.
Eisenhaber B, Schneider G, Wildpaner M, Eisenhaber F. 2004. A sensitive predictor for potential GPI lipid modification sites in fungal protein sequences and its application to genome-wide studies for *Aspergillus nidulans*, *Candida albicans*, *Neurospora crassa*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. J Mol Biol. 337:243–253.
Ellis JR, Burke JM. 2007. EST-SSRs as a resource for population genetic analyses. Heredity. 99:125–132.
Farrer RA, Kemen E, Jones JD, Studholme DJ. 2009. De novo assembly of the *Pseudomonas syringae* pv. syringae B728a genome using Illumina/Solexa short sequence reads. FEMS Microbiol Lett. 291:103–111.
Gibbons JG, Rokas A. 2009. Comparative and functional characterization of intragenic tandem repeats in 10 *Aspergillus Genomes*. Mol Biol Evol. 26:591–602.
Grimaldi D, Engel MS. 2005. Evolution of the insects. Cambridge: Cambridge University Press.
Hill CA, Fox AN, Pitts RJ, Kent LB, Tan PL, Chrystal MA, Cravchik A, Collins FH, Robertson HM, Zwiebel LJ. 2002. G protein-coupled receptors in *Anopheles gambiae*. Science. 298:176–178.
Hillier LW, Marth GT, Quinlan AR, et al. (20 co-authors). 2008. Whole-genome sequencing and variant discovery in *C. elegans*. Nat Meth. 5:183–188.
Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. Cell. 95:717–728.
Holt RA, Jones SJ. 2008. The new paradigm of flow cell sequencing. Genome Res. 18:839–846.
Holt RA, Subramanian GM, Halpern A, et al. (123 co-authors). 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science. 298:129–149.
Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature. 443:931–949.
Hudson ME. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour. 8:3–17.
Hughes J, Longhorn SJ, Papadopoulou A, Theodorides K, de Riva A, Mejia-Chang M, Foster PG, Vogler AP. 2006. Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles). Mol Biol Evol. 23:268–278.
Hunter S, Apweiler R, Attwood TK, et al. (38 co-authors). 2009. InterPro: the integrative protein signature database. Nucleic Acids Res. 37:D211–D215.
Huntley MA, Clark AG. 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 drosophila species. Mol Biol Evol. 24:2598–2609.
Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 8:R143.
King N, Hittinger CT, Carroll SB. 2003. Evolution of key cell signaling and adhesion protein families predates animal origins. Science. 301:361–363.
Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 39:309–338.
Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305:567–580.
Lander ES, Linton LM, Birren B, et al. (256 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.
Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM. 2009. Transcriptome sequencing to detect gene fusions in cancer. Nature. 458:97–101.
Mardis ER. 2008a. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 9:387–402.
Mardis ER. 2008b. The impact of next-generation sequencing technology on genetics. Trends Genet. 24:133–141.
Margulies M, Egholm M, Altman WE, et al. (56 co-authors). 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 437:376–380.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 320:1344–1349.
Nakao A, Yoshihama M, Kenmochi N. 2004. RPG: the Ribosomal Protein Gene database. Nucleic Acids Res. 32:D168–D170.
Nene V, Wortman JR, Lawson D, et al. (95 co-authors). 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science. 316:1718–1723.
Nichols SA, Dirks W, Pearse JS, King N. 2006. Early evolution of animal cell signaling and adhesion genes. Proc Natl Acad Sci USA. 103:12451–12456.

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr., Grattapaglia D, Sederoff RR, Kirst M. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics. 9:312.

Ranson H, Claudianos C, Ortelli F, Abgrall C, Hemingway J, Sharakhova MV, Unger MF, Collins FH, Feyereisen R. 2002. Evolution of supergene families associated with insecticide resistance. Science. 298:179–181.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276–277.

Richards S, Gibbs RA, Weinstock GM, et al. (191 co-authors). 2008. The genome of the model beetle and pest *Tribolium castaneum*. Nature. 452:949–955.

Riley ML, Schmidt T, Artamonova II, Wagner C, Volz A, Heumann K, Mewes HW, Frishman D. 2007. PEDANT genome database: 10 years online. Nucleic Acids Res. 35:D354–D357.

Rokas A, Abbot P. 2009. Harnessing genomics for evolutionary insights. Trends Ecol Evol. 24:192–200.

Rokas A, Kruger D, Carroll SB. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science. 310:1933–1938.

Ruepp A, Zollner A, Maier D, et al. (11 co-authors). 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. 32:5539–5545.

Schuster SC. 2008. Next-generation sequencing transforms today's biology. Nat Meth. 5:16–18.

Smith AD, Xuan Z, Zhang MQ. 2008. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics. 9:128.

Sokal RR, Rohlf FJ. 1995. Biometry: the principles and practice of statistics in biological research. New York: W.H. Freeman.

Tang F, Barbacioru C, Wang Y, et al. (12 co-authors). 2009. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Meth. 6:377–382.

′t Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJ, den Dunnen JT. 2008. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res. 36:e141.

Toth AL, Varala K, Newman TC, et al. (11 co-authors). 2007. Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. Science. 318:441–444.

Toth G, Gaspari Z, Jurka J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10:967–981.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol. 17:1636–1647.

Whiteford N, Haslam N, Weber G, Prugel-Bennett A, Essex JW, Roach PL, Bradley M, Neylon C. 2005. An analysis of the feasibility of short read sequencing. Nucleic Acids Res. 33:e171.

Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N. 2006. 454 sequencing put to the test using the complex genome of barley. BMC Genomics. 7:275.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 453:1239–1243.

Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. 2008. DBD–taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res. 36:D88–D92.

Xia Q, Zhou Z, Lu C, et al. (80 co-authors). 2004. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science. 306:1937–1940.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.

Scott Edwards, Associate Editor

Accepted August 19, 2009