Supporting Information

Hittinger et al. 10.1073/pnas.0910449107

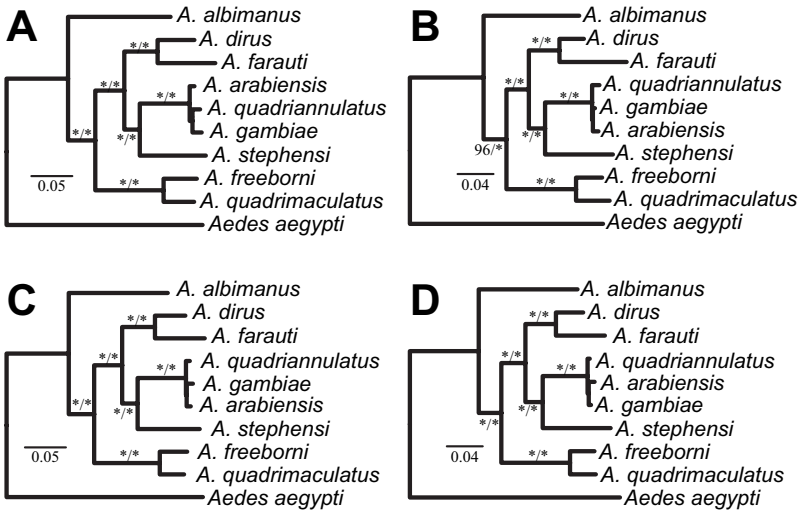

Fig. S1. Removal of loci of ambiguous orthology has no topological effect on phylogenetic inference from short-read next-generation DNA sequencing. (*A*) Maximum likelihood (ML) phylogeny produced from the data matrix constructed by considering all contigs ≥100 bp assembled from ~13 million sequence reads per species using *A. aegypti* full transcripts as references under the single-contig strategy after exclusion of all loci in which *A. gambiae* contigs were inaccurately assigned to their *A. aegypti* reference transcript orthologs. (*B*) Same analysis as in *A* but on the data matrix constructed by considering all contigs ≥300 bp. (*C*) ML phylogeny produced from the data matrix constructed by considering all contigs ≥100 bp assembled from ~13 million sequence reads per species using *A. aegypti* full transcripts as references under the single-contig strategy after exclusion of all loci that contained paralogs using a phylogeny-based assessment of orthology assignment. (*D*) Same analysis as in *B* but on the data matrix constructed by considering all contigs ≥300 bp. Clade support near internodes represents bootstrap support (ML) and posterior probability (Bayesian inference), respectively. Asterisks denote absolute support. Branch lengths represent estimated substitutions per site.

Table S1. Species taxonomy and collection information

Species	Strain	Stock no.	Collection location
Anopheles albimanus Wiedemann (Nyssorhynchus)	STECLA	MRA-126	Santa Tecla, El Salvador
Anopheles arabiensis Patton (Cellia)	KGB	MRA-339	Kanyemba, Zimbabwe
Anopheles dirus Payton and Harrison (Cellia)	WRAIR2	MRA-700	Thailand
Anopheles farauti Laveran (Cellia)	FAR1	MRA-489	Rabaul Colony, Papua New Guinea
Anopheles freeborni Aitken (Anopheles)	F1	MRA-130	Marysville, CA
Anopheles gambiae Giles (Cellia)	SUA2LA	MRA-765	Suakoko, Liberia
Anopheles quadriannulatus Theobald (Cellia)	SKUQUA	MRA-761	Skukuze, South Africa
Anopheles quadrimaculatus Say (Anopheles)	ORLANDO	MRA-139	United States
Anopheles stephensi Liston (Cellia)	STE2	MRA-128	Delhi, India
Aedes (Stegomyia) aegypti (Linnaeus)	LVP-IB12	MRA-735	West Africa

Table S2. Summary statistics of assembled test contigs from 13 million Solexa/Illumina 36-bp sequence reads from all mosquito species

	Assembly statistics						≥100-bp test contig set				≥300-bp test contig set			
Species	Read no.	ABQS	*k*-mer	Node no.	N50	Maximum length	Contig no.	Amount	Median length	Median coverage	Contig no.	Amount	Median length	Median coverage
A. albimanus	13,741,955	32	23	89,735	87	2,041	10,738	1,983,006	142	7	1,143	528,330	403	16
A. arabiensis	12,180,498	36	21	161,248	74	1,987	19,172	3,139,670	133	6	1,241	534,257	379	14
A. dirus	14,659,921	34	23	101,182	89	1,952	14,162	2,575,380	141	7	1,444	661,028	399	16
A. farauti	12,114,242	34	23	97,831	82	1,152	15,457	2,587,562	136	6	1,090	467,309	389	14
A. freeborni	14,107,744	33	21	173,715	81	1,863	21,828	3,857,785	140	6	2,011	873,651	384	12
A. gambiae	12,101,924	28	23	160,346	81	1,529	20,364	3,331,879	134	6	1,246	534,880	377	16
A. quadriannulatus	13,079,694	36	23	100,959	78	2,590	14,207	2,377,237	131	6	1,006	473,416	399	16
A. quadrimaculatus	13,803,823	34	21	169,661	81	1,635	24,152	4,085,079	137	6	1,729	759,914	383	12
A. stephensi	13,547,996	36	21	160,555	79	1,903	17,660	3,092,149	138	6	1,504	677,568	390	12
A. aegypti	11,465,769	37	21	91,591	81	1,430	15,712	2,727,917	137	5	1,327	592,760	385	12

"Read no.," no. 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 (1); "Node," number of raw contigs produced by the VELVET software; "N50," the length-weighted average of contig length, such that the average base in the assembly will appear in a contig of N50 length or greater; "Maximum length," length of longest contig in the assembly; "Amount," amount of sequence found in contigs ≥100/300 bp; "Median length," median length of contigs ≥100/300 bp; "Median coverage," median coverage depth of contigs ≥100/300 bp. The data for *A. gambiae* and *A. aegypti* are from Gibbons et al. (2).

1. Zerbino DR (2008) Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829.
2. Gibbons JG, et al. (2009) Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. *Mol Biol Evol* 26:2731–2744.

Table S3. Quantity, putative ortholog detection, and accuracy summary statistics for data matrices constructed from ≥100 and ≥300 bp contigs using the single-contig strategy

			Data matrices constructed from ≥100-bp contigs			
Data set	Total alignment length	Overlapping alignment length	Missing data, %	No. of orthologs	No. of true orthologs	Accuracy, %
100,000	NA	NA	NA	0	NA	NA
250,000	342	1	41	1	0	0
500,000	3,588	100	47	11	10	91
1,000,000	30,462	1,744	44	72	69	96
2,000,000	57,864	6,004	38	120	113	94
3,000,000	90,705	10,718	40	173	161	93
4,000,000	119,316	12,556	42	212	197	93
5,000,000	148,653	13,873	44	252	235	93
~6,500,000	214,905	15,030	46	333	311	93
~13,000,000	389,364	15,239	51	553	526	95
			Data matrices constructed from ≥300-bp contigs			
100,000	NA	NA	NA	0	NA	NA
250,000	NA	NA	NA	0	NA	NA
500,000	NA	NA	NA	0	NA	NA
1,000,000	NA	NA	NA	0	NA	NA
2,000,000	4,896	792	37	8	6	75
3,000,000	17,037	3,576	28	29	27	93
4,000,000	22,926	4,493	26	36	33	92
5,000,000	27,966	5,295	28	40	38	95
~6,500,000	33,465	4,091	34	37	36	97
~13,000,000	72,564	5,518	44	69	68	99

"Data set," no. of 36-bp sequence reads used as input in the assembly; "Total alignment length," total length of alignment in data matrix; "Overlapping alignment length," total length of alignment after excluding all alignment columns with data missing or gaps; "Missing data, %," percentage of missing data in the alignment; "No. of orthologs," no. of putative orthologs identified across all 10 species; "No. of true orthologs," no. of true orthologs in alignment; "Accuracy, %," percentage of orthologs detected accurately in alignment. Note that placements whose accuracy could not be confirmed include both real errors and possible reference transcriptome annotation errors, which makes our accuracy assessment conservative.

Table S4. Clade support values for phylogenetic analyses of phylogenomic data matrices constructed from varying amounts of starting sequence data using the single-contig strategy

Clade										
	Clade support values for ML analysis of data matrices constructed from ≥100-bp contigs									
(Agam, Aara, Aqan)	NA	99	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste)	NA	4	18	0	100	100	100	100	100	100
(Adir, Afar)	NA	66	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar)	NA	28	100	100	100	100	100	100	100	100
(Afre, Aqma)	NA	71	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar, Afre, Aqma)	NA	19	0	100	100	100	100	100	100	100
	Clade support values for BI analysis of data matrices constructed from ≥100-bp contigs									
(Agam, Aara, Aqan)	NA	69	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste)	NA	7	6	0	100	100	100	100	100	100
(Adir, Afar)	NA	88	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar)	NA	0	100	100	100	100	100	100	100	100
(Afre, Aqma)	NA	79	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar, Afre, Aqma)	NA	0	10	100	100	100	100	100	100	100
	Clade support values for ML analysis of data matrices constructed from ≥300-bp contigs									
(Agam, Aara, Aqan)	NA	NA	NA	NA	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste)	NA	NA	NA	NA	0	100	100	100	100	100
(Adir, Afar)	NA	NA	NA	NA	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar)	NA	NA	NA	NA	0	100	100	100	100	100
(Afre, Aqma)	NA	NA	NA	NA	0	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar,Afre,Aqma)	NA	NA	NA	NA	13	68	86	100	100	100
	Clade support values for BI analysis of data matrices constructed from ≥300-bp contigs									
(Agam, Aara, Aqan)	NA	NA	NA	NA	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste)	NA	NA	NA	NA	0	100	100	100	100	100
(Adir, Afar)	NA	NA	NA	NA	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar)	NA	NA	NA	NA	0	100	100	100	100	100
(Afre, Aqma)	NA	NA	NA	NA	0	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar, Afre, Aqma)	NA	NA	NA	NA	11	11	100	100	100	100
No. of 36-bp sequence reads used in assembly	1×10^5	2.5×10^5	5×10^5	1×10^6	2×10^6	3×10^6	4×10^6	5×10^6	~6.5×10^6	~13×10^6
Length of data matrix constructed from ≥100-bp contigs	NA	342	3,588	30,462	57,864	90,705	119,316	148,653	214,905	389,364
Length of data matrix constructed from ≥300-bp contigs	NA	NA	NA	NA	4,896	17,037	22,926	27,966	33,465	72,564

Table S5. Quantity, putative ortholog contig detection, and accuracy summary statistics for data matrices constructed from ≥100- and ≥300-bp contigs using the supercontig strategy

				Data matrices constructed from ≥100-bp contigs			
Data set	No. of orthologs	No. of contigs	Total alignment length	Overlapping alignment length	Missing data, %	No. of true contigs	Accuracy, %
100,000	4	3	459	0	49	1	33
250,000	50	34	6,141	0	52	27	79
500,000	124	128	26,625	285	45	110	86
1,000,000	226	287	60,135	4,188	36	245	85
2,000,000	430	550	122,358	16,563	34	460	84
3,000,000	630	825	188,784	28,548	35	680	82
4,000,000	850	1,152	260,844	36,684	35	945	82
5,000,000	1,054	1,449	332,871	43,164	36	1,202	83
~6,500,000	1,591	1,872	521,352	57,441	37	1,569	84
~13,000,000	2,661	4,118	970,746	82,650	38	3,496	85
				Data matrices constructed from ≥300-bp contigs			
100,000	0	0	NA	NA	NA	0	NA
250,000	2	2	630	0	49	1	50
500,000	10	6	2,433	0	47	4	67
1,000,000	64	36	17,850	0	48	31	86
2,000,000	148	86	53,169	876	40	74	86
3,000,000	198	146	76,398	3,957	37	128	88
4,000,000	255	183	100,476	5,832	37	159	87
5,000,000	302	222	124,512	6,867	37	190	86
~6,500,000	445	290	190,155	7,413	39	251	87
~13,000,000	725	523	345,312	10,227	42	451	86

"Data set," no. of 36-bp sequence reads used as input in the assembly; "No. of orthologs," no. of putative ortholog supercontigs identified; "No. of contigs," no. of putative ortholog contigs identified; "Total alignment length," total length of alignment in data matrix; "Overlapping alignment length," total length of alignment after excluding all alignment columns with data missing or gaps; "Missing data, %," percentage of missing data in the alignment; "No. of true contigs," no. of true ortholog contigs in alignment; "Accuracy, %," percentage of ortholog contigs detected accurately in alignment. Note that placements whose accuracy could not be confirmed include both real errors and possible reference transcriptome annotation errors, which makes our accuracy assessment conservative.

Table S6. Clade support values for phylogenetic analyses of phylogenomic data matrices constructed from varying amounts of starting sequence data using the supercontig strategy

Clade	Clade support values for ML analysis of data matrices constructed from ≥100-bp contigs									
(Agam, Aara, Aqan)	67	100	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste)	57	46	97	100	100	100	100	100	100	100
(Adir, Afar)	0	29	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar)	0	26	100	100	100	100	100	100	100	100
(Afre, Aqma)	2	100	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar, Afre, Aqma)	32	11	100	100	100	100	100	100	100	100
	Clade support values for BI analysis of data matrices constructed from ≥100-bp contigs									
(Agam, Aara, Aqan)	90	100	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste)	0	97	100	100	100	100	100	100	100	100
(Adir, Afar)	0	100	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar)	0	100	100	100	100	100	100	100	100	100
(Afre,Aqma)	0	100	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar, Afre,Aqma)	0	12	100	100	100	100	100	100	100	100
	Clade support values for ML analysis of data matrices constructed from ≥300-bp contigs									
(Agam, Aara, Aqan)	NA	13	82	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste)	NA	6	45	77	100	100	100	100	100	100
(Adir, Afar)	NA	0	22	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar)	NA	3	0	100	100	100	100	100	100	100
(Afre, Aqma)	NA	0	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar, Afre, Aqma)	NA	7	0	41	100	100	100	100	100	100
	Clade support values for BI analysis of data matrices constructed from ≥300-bp contigs									
(Agam, Aara, Aqan)	NA	47	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste)	NA	11	0	5	100	100	100	100	100	100
(Adir, Afar)	NA	0	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar)	NA	93	100	100	100	100	100	100	100	100
(Afre, Aqma)	NA	16	100	100	100	100	100	100	100	100
(Agam, Aara, Aqan, Aste, Adir, Afar, Afre,Aqma)	NA	25	25	33	100	100	100	100	100	100
No. of 36-bp sequence reads used in assembly	1×10^5	2.5×10^5	5×10^5	1×10^6	2×10^6	3×10^6	4×10^6	5×10^6	~6.5×10^6	~13×10^6
Length of data matrix constructed from ≥ 100 bp contigs	459	6,141	26,625	60,135	122,358	188,784	260,844	332,871	521,352	970,746
Length of data matrix constructed from ≥300-bp contigs	NA	630	2,433	17,850	53,169	76,398	100,476	124,512	190,155	345,312