

Intraspecific DNA variation in nuclear genes of the mosquito *Aedes aegypti*

I. Morlais and D. W. Severson

Center for Tropical Disease Research and Training,
Department of Biological Sciences, University of Notre
Dame, Notre Dame, IN 46556, USA

Abstract

Single nucleotide polymorphisms (SNPs) are an abundant source of genetic variation among individual organisms. To assess the usefulness of SNPs for genome analysis in the yellow fever mosquito, *Aedes aegypti*, we sequenced 25 nuclear genes in each of three strains and analysed nucleotide diversity. The average frequency of nucleotide variation was 12 SNPs per kilobase, indicating that nucleotide variation in *Ae. aegypti* is similar to that in other organisms, including *Drosophila* and the malaria vector *Anopheles gambiae*. Transition polymorphisms outnumbered transversion polymorphisms, at a ratio of about 2 : 1. We examined codon usage and confirmed that mutational bias favours G and C ending codons. Codon bias was most pronounced in highly expressed genes. Nucleotide diversity estimates indicated that substitution rates are positively correlated in coding and non-coding regions. Nucleotide diversity varied from one gene to another. The unequal distribution of SNPs among *Ae. aegypti* nuclear genes suggests that single base variations are non-neutral and are subject to selective constraints. Our analysis showed that ubiquitously expressed genes have lower polymorphism rates and are likely under strong purifying selection, whereas tissue specific genes and genes with a putative role in parasite defence exhibit higher levels of polymorphism that may be associated with diversifying selection.

Keywords: transitions/transversions, synonymous/non-synonymous substitutions, Culicidae, genomics, codon bias.

Received 26 March 2003; accepted after revision 5 August 2003. Correspondence: David W. Severson, Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA. Tel.: +1 574 631 3826; fax: +1 574 631 7413; e-mail: david.w.severson.1@nd.edu

Introduction

Single-nucleotide polymorphisms (SNPs) are frequently observed in vertebrate and invertebrate genomes (Jakubowski & Kornfeld, 1999; Sachidanandam *et al.*, 2001; Taillon-Miller *et al.*, 1998). With recent advances in high-throughput sequence analysis technology, SNPs have become the marker of choice for large-scale mapping and genotyping (Berger *et al.*, 2001; Lindblad-Toh *et al.*, 2000; Taillon-Miller *et al.*, 1999; Wang *et al.*, 1998). SNP markers therefore provide powerful tools for investigating population genetics and characterizing candidate disease genes, as well as elucidating evolutionary processes at the molecular level (see reviews in Black *et al.*, 2001; Akashi, 2001; Brookes, 1999).

Nucleotide diversity has been used to define codon usage patterns within taxa or species (Argentine & James, 1993; Besansky, 1993; Ikemura, 1985; Sharp *et al.*, 1986; 1988; Shields *et al.*, 1988). In *Drosophila* and bacteria, codon usage bias, e.g. the unequal usage of synonymous codons, is hypothesized to occur as a result of selection for efficient translation; the preferred codons in highly biased genes match the most abundant isoaccepting tRNAs (Gouy & Gautier, 1982; Moriyama & Powell, 1997a; Powell & Moriyama, 1997; Sharp & Li, 1986a). With the increase in large-scale genome sequencing projects, nucleotide polymorphisms have been more widely used in phylogenetic studies (Aquadro *et al.*, 2001; Bielawski *et al.*, 2000; Comeron & Aguade, 1998; Dunn *et al.*, 2001; McVean & Vieira, 1999).

The mosquito *Aedes aegypti* has a world-wide distribution and, because it is easily worked with in the laboratory and is the major vector of yellow fever and dengue fever viruses, is one of the most intensively studied mosquito species. Studies of population genetics as well as genetic mapping and quantitative trait loci (QTL) characterization require large numbers of polymorphic markers to genotype a target population. However, simple-sequence repeats or microsatellites are not abundant or useful as genetic marker loci in *Ae. aegypti* (Fagerberg *et al.*, 2001). Therefore, the most common genetic markers used to date in *Ae. aegypti* are the RFLP, RAPD and SSCP markers that are generally labour intensive for large populations and in some

Table 1. Nucleotide polymorphisms in *Aedes aegypti* nuclear genes

Gene	N_{all}	L (bp)	Coding region												Non-coding region						
			Polymorphic sites												Polymorphic sites						
			Transitions				Transversions				Nucleotide diversity				Polymorphic sites			Nucleotide diversity			
			1st	2nd	3rd	Σ	1st	2nd	3rd	Σ	Syn	Rep	Total	π	K_s	K_a	L (bp)	T_s	T_v	Σ	π
<i>Fer (H)</i>	4	630	1	2	8	11	0	0	1	1	10	2	12	0.0119	0.0340	0.0028	183	3	3	6	0.0200
<i>AEGL8</i>	4	1008	1	1	5	7	1	2	2	5	7	5	12	0.0066	0.0154	0.0033	83	0	1	1	0.0060
<i>Sec61</i>	4	1431	4	4	8	16	2	0	4	6	12	10	22	0.0080	0.0145	0.0052	531	2	2	4	0.0038
<i>AelMUC1</i>	12	828	12	11	16	39	4	8	10	22	25	36	61	0.0245	0.0312	0.0183	273	6	5	11	0.0271
<i>Chym</i>	8	807	5	3	9	17	3	1	5	9	15	11	26	0.0124	0.0236	0.0079	153	1	2	3	0.0131
<i>AEGL2</i>	4	624	2	3	9	14	0	2	3	5	12	7	19	0.0174	0.0427	0.0090	156	1	1	2	0.0085
<i>RpS11</i>	1	459	0	0	0	0	0	0	0	0	0	0	0	0.0000	0.0000	0.0000	66	0	0	0	0.0000
<i>RpL31</i>	1	375	0	0	0	0	0	0	0	0	0	0	0	0.0000	0.0000	0.0000	174	0	0	0	0.0000
<i>RpL17A</i>	3	423	0	0	2	2	0	0	0	0	2	0	2	0.0024	0.0069	0.0000	419	3	3	6	0.0096
<i>AEGL23</i>	5	237	2	0	2	4	1	1	2	4	4	4	8	0.0169	0.0363	0.0114	50	1	1	2	0.0251
<i>PGK</i>	6	1248	4	4	8	16	4	2	5	11	12	15	27	0.0095	0.0197	0.0059	642	10	8	18	0.0182
<i>ODC-AZ</i>	3	723	1	1	3	5	1	1	1	3	5	3	8	0.0058	0.0129	0.0027	291	1	1	2	0.0034
<i>CYP9J</i>	4	1611	9	8	19	36	7	2	8	17	26	27	53	0.0191	0.0321	0.0136	252	2	1	3	0.0074
<i>CRALBP</i>	8	873	9	1	31	41	2	0	4	6	41	6	47	0.0212	0.0687	0.0025	408	11	6	17	0.0181
<i>Ef-2</i>	4	2535	1	5	14	20	3	2	3	8	16	12	28	0.0057	0.0134	0.0032	123	1	0	1	0.0041
<i>AEGBS11</i>	4	423	3	4	7	14	0	0	4	4	10	8	18	0.0260	0.0542	0.0171	53	0	0	0	0.0000
<i>APN</i>	4	2868	22	12	39	73	3	7	16	26	57	42	99	0.0217	0.0446	0.0133	117	2	2	4	0.0199
<i>SDR</i>	4	765	2	1	2	5	0	2	2	4	4	5	9	0.0070	0.0182	0.0037	145	0	1	1	0.0046
<i>mRNABP</i>	4	1260	2	3	7	12	1	0	3	4	9	7	16	0.0063	0.0135	0.0039	270	2	2	4	0.0111
<i>CecA</i>	2	180	1	0	2	3	0	0	0	0	2	1	3	0.0167	0.0323	0.0088	75	1	0	1	0.0133
<i>TSF</i>	6	1902	6	8	49	63	5	6	13	24	61	26	87	0.0202	0.0535	0.0071	190	3	3	6	0.0168
<i>DefA</i>	2	297	0	0	4	4	0	0	0	0	4	0	4	0.0090	0.0259	0.0000	170	0	0	0	0.0000
<i>DCE</i>	6	1392	10	4	51	65	3	1	20	24	74	15	89	0.0303	0.1030	0.0053	117	6	3	9	0.0377
<i>DDC</i>	2	999	2	0	7	9	0	0	3	3	12	0	12	0.0035	0.0131	0.0000	6	0	0	0	0.0000
<i>NaK</i>	2	322	0	0	3	3	0	0	0	0	3	0	3	0.0031	0.0089	0.0000	208	6	10	16	0.0874
Total		24 220	99	75	305	479	40	37	109	186	423	242	665				5155	62	55	117	
Average														0.0122	0.0287	0.0058					0.0142

L, length in nucleotide of the coding sequence; N_{all} , number of alleles; Syn, synonymous substitutions; Rep, replacement substitutions; T_s , transitions; T_v , transversions; π , average number of nucleotide substitution per site; K_s , per synonymous site; K_a , per non-synonymous site.

instances are limited by the small amounts of extractable genomic DNA per individual (Antolin *et al.*, 1996; Fulton *et al.*, 2001; Severson *et al.*, 1999, 2002).

In this study, we analysed nucleotide polymorphism and distribution at 25 independent nuclear genes in *Ae. aegypti*. Our results document extensive intraspecific nucleotide variation within some *Ae. aegypti* genes. We confirm that codon usage is biased toward C- and G-ending codons and, as for other Diptera, the most biased genes represent ubiquitous, highly expressed molecules. Finally, our analyses suggest that the frequency of SNPs within individual genes varies with gene function, with more SNPs being observed in genes with tissue-specific expression or with a putative role in parasite defence.

Results and discussion

The distribution of SNPs among 25 *Ae. aegypti* nuclear genes is shown in Table 1. With an average of 12 SNPs per kilobase (kb) ($\pi = 0.0122$, Table 1), our *Ae. aegypti* laboratory strains contain high levels of nucleotide heterogeneity. In comparison, the typical SNP frequency observed in

human genomic DNA is about 1 every 1000 bp (Aquadro *et al.*, 2001; Brookes, 1999). Thus, because of the large nucleotide variation in *Ae. aegypti*, SNPs should prove extremely useful as genetic markers, particularly for high-throughput, fine-scale linkage and complex trait analysis. We also observed that the frequency of SNPs varies considerably from one gene to another, ranging from none (*RpS11*, *RpL31*) to 99 (*APN*). Although not directly comparable, the *Anopheles gambiae* genome sequence also shows a highly variable SNP distribution, wherein some genome regions have few SNPs and others have more than eight per kb (Holt *et al.*, 2002).

Our results suggest that insertion/deletion polymorphisms (indels) may be frequent throughout the *Ae. aegypti* genome. That is, six (24%) of the 25 genes contained indels (Table 2). These indels can easily be exploited as PCR-based markers for genetic mapping (Bhattaramakki *et al.*, 2002). Our previous analysis of the *AelMUC1* gene identified indels within the coding sequence that result in different protein isoforms which may have distinct structural/functional properties that influence *Plasmodium gallinaceum* susceptibility (Morlais & Severson, 2001). The presence of

Table 2. Indel polymorphisms

Gene	Coding region	Non-coding region		
		introns	5' end	3' end
<i>AeIMUC1</i>	1 bp, 24 bp		30 or 51 bp	
<i>RpL17A</i>		1 bp		
<i>AEGL23</i>				6 bp
<i>CRALBP</i>			60 bp	42 bp
<i>DCE</i>			11 bp	
<i>NaK</i>		2 bp		

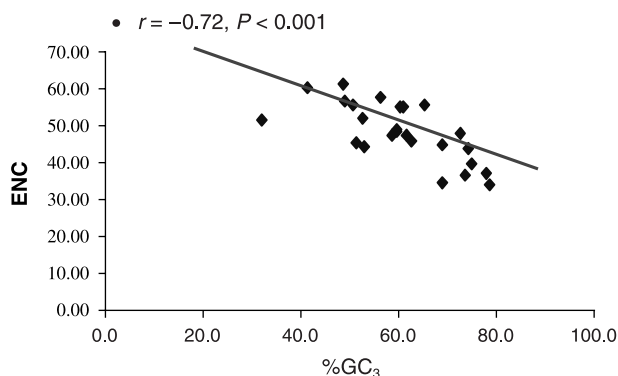
indels and a high number of non-synonymous amino acid substitutions in such genes indicates that the proteins are not under strong negative or purifying selection. However, indels located within non-coding sequences should be neutral, arising by misreading or by slippage during replication, particularly at repeat regions. For example, indels in the 3' untranslated regions of the *AeMUC1* and *CRALBP* genes have repeats flanking the inserted/deleted sequence. However, not all indels in untranslated regions are associated with repeats. The 60 bp indel found at the 5' end of *CRALBP* mRNAs has characteristic features of U2-type introns, e.g. with GT-AG splicing junctions, and is removed during RNA splicing, as seen by comparing genomic and cDNA sequences (data not shown). A G→A transition at the 3' splicing site leads to an unspliced haplotype. As in the rice β -tubulin gene (*Ostub16*), polymorphisms that influence intron splicing in the 5' untranslated region may have significance for transcriptional expression (Morello *et al.*, 2002). Such polymorphisms could also represent an intermediate state of intron gain/loss with associated selection intensities at least equal to that observed for non-synonymous substitutions (Llopart *et al.*, 2002).

For all genes, we observed that transition substitutions are more common than transversion substitutions (Table 1). The frequency of transitions for all coding and non-coding sequences is 69.2%, which is similar to the $\approx 2/3$ ratio reported for *Drosophila* and humans (Brookes, 1999; Moriyama & Powell, 1996). For coding vs. non-coding regions, the frequencies of transitions are 72.0% and 53.0%, respectively, and are significantly different ($\chi^2 = 16.9$; $P < 0.001$). SNPs occur more frequently as transitions in coding sequences than in non-coding regions and are more common at the third codon position (62.3%). These results are similar to those observed in three *Drosophila* species (Moriyama & Powell, 1996). The unequal distribution of SNPs is likely to be due to the degeneracy of the genetic code and selective constraints for gene conservation; transition polymorphisms at the wobble position are more likely to result in synonymous substitutions. Selection at fourfold degenerate codons should then be nearly neutral. Transitions and transversions were computed at these sites for all genes and the results are presented in Table 3. As pre-

Table 3. Transition (Ts) and transversion (Tv) polymorphisms for different classes of DNA

	Polymorphism			λ^2 (Fisher's exact)	
	Ts	Tv	% Tv	Cd-R	Fourfold
Non-coding regions	62	55	47.0	16.91***	1.29, ns
Coding regions (Cd-R)	479	186	28.0	–	11.5***
Fourfold degenerate sites	122	83	40.5	–	–

*** $P < 0.001$.

**Figure 1.** Correlation of codon usage with GC₃ content.

dicted, the frequency of transversions at fourfold degenerate sites (40.5%) is not significantly different than for non-coding regions ($\chi^2 = 1.29$; $P = 0.255$).

As a measure of codon usage bias, we used the 'effective number of codons' or ENC (Wright, 1990). The ENC varies from 20, the most extreme bias wherein only one codon is used per amino acid, to 61 for unbiased genes where all synonymous amino acids are used equally. The ENC values range from 33.6 for the ribosomal protein S11 to 61.0 for *Cecropin A* (Table 4) indicating that, in *Ae. aegypti*, codon usage bias differs from one gene to another, as is seen in other organisms (Besansky, 1993; Ikemura, 1985; Powell & Moriyama, 1997; Sharp *et al.*, 1988). We found the highest codon usage bias for ribosomal proteins and *Ef-2*, which agrees with the premise that highly expressed genes have a higher codon usage bias (Marais *et al.*, 2001; Moriyama & Powell, 1997a; Pal *et al.*, 2001; Shields *et al.*, 1988). The *Ae. aegypti* genes show an average ENC of 48.6, which is similar to that reported for *Drosophila melanogaster*, 46.2 (Powell & Moriyama, 1997).

We also examined the base composition for individual genes relative to the overall nucleotide sequence and for the third codon position (Table 4). The results indicate that base composition varies among genes and that the G+C content at the third codon position (GC₃) is negatively correlated ($r = -0.72$, $P < 0.001$, Fig. 1) with the ENC. The most biased genes have higher GC₃ frequencies, indicating that codon usage in *Ae. aegypti* is similar to *An. gambiae*,

Table 4. Base composition in *Aedes aegypti* nuclear genes

Gene	L (bp)	ENC	Total				Third codon position					Fourfold degenerate sites			
			T	C	A	G	T3	C3	A3	G3	GC ₃	T4	C4	A4	G4
<i>Fer (H)</i>	630	39.15	18.0	27.3	27.6	27.1	11.3	41.8	14.0	32.9	74.7	11.8	43.5	17.9	26.8
<i>AEGB8</i>	1008	59.87	28.1	19.3	26.6	26.0	36.3	19.9	22.5	21.3	41.2	33.7	19.4	25.1	21.8
<i>Sec61</i>	1431	47.15	28.3	24.9	22.0	24.8	24.4	32.7	17.2	25.7	58.4	31.5	22.2	23.5	22.8
<i>AelMUC1</i>	828	55.32	23.8	28.1	24.4	23.7	27.5	30.0	21.9	20.6	50.6	25.5	25.6	30.6	18.3
<i>Chym</i>	807	45.65	22.9	26.0	21.8	29.3	21.5	34.2	15.7	28.6	62.8	22.5	33.3	13.8	30.4
<i>AEGB12</i>	624	47.14	27.8	23.6	22.9	25.7	24.9	33.9	13.6	27.6	61.5	26.9	31.6	17.0	24.5
<i>RpS11</i>	459	33.57	18.3	28.5	25.7	27.5	13.7	45.1	7.9	33.3	78.4	20	40	8.6	31.4
<i>RpL31</i>	375	36.97	16.0	29.9	26.6	27.5	12.8	41.6	9.6	36.0	77.6	15.2	42.4	9.1	33.3
<i>RpL17A</i>	423	48.37	22.0	25.7	22.2	30.1	25.0	35.3	15.4	24.3	59.6	30.1	36.1	14.2	19.6
<i>AEGB23</i>	237	45.06	22.4	23.0	27.9	26.7	23.0	29.7	25.7	21.6	51.3	14.5	36	23.7	25.8
<i>PGK</i>	1248	51.81	24.9	22.6	24.6	27.9	30.3	29.1	17.2	23.4	52.5	38.4	28.5	20.6	12.5
<i>ODC-AZ</i>	723	55.18	24.2	28.3	24.4	23.1	20.5	33.1	14.3	32.1	65.2	20.3	33.3	16.9	29.5
<i>CYP9J</i>	1611	54.78	24.5	24.1	26.9	24.5	19.8	30.6	20.0	29.6	60.2	20.8	24.6	21.0	33.6
<i>CRALBP</i>	873	47.39	21.5	27.5	25.2	25.8	14.9	37.3	12.7	35.1	72.4	17.6	30.8	13.0	38.6
<i>Ef-2</i>	2535	34.04	22.8	27.3	22.8	27.1	22.0	40.0	9.1	28.9	68.9	30.1	39.5	9.3	21.1
<i>AEGBS11</i>	423	54.48	25.6	21.2	22.6	30.6	21.6	28.0	17.6	32.8	60.8	17.7	24.6	21.2	36.5
<i>APN</i>	2868	57.39	23.7	26.0	26.6	23.7	23.9	31.2	20.0	24.9	56.1	21.2	30.2	23.7	24.9
<i>SDR</i>	765	56.12	22.3	22.4	29.2	26.1	24.8	26.3	26.5	22.4	48.7	19.8	27.0	28.6	24.6
<i>mRNABP</i>	1260	51.08	17.9	18.4	40.5	23.2	32.6	15.7	35.6	16.1	31.8	38.5	17.6	32.1	11.8
<i>CecA</i>	180	61.00	24.7	20.8	25.6	28.9	27.5	19.2	24.1	29.2	48.4	40.3	20.8	20.8	18.1
<i>TSF</i>	1902	44.35	20.3	26.1	24.3	29.3	17.4	35.4	13.9	33.3	68.7	16.8	32.0	15.9	35.3
<i>DefA</i>	297	47.93	25.3	26.6	19.1	29.0	30.3	33.3	10.1	26.3	59.6	26.9	30.8	7.7	34.6
<i>DCE</i>	1392	43.41	21.7	30.8	21.9	25.6	16.6	46.3	9.3	27.8	74.1	16.4	45.1	9.3	29.2
<i>DDC</i>	999	43.87	25.6	23.4	25.5	25.5	26.6	28.9	20.6	23.9	52.8	36.7	20.8	22.9	19.6
<i>NaK</i>	322	36.29	25.8	25.8	19.7	28.7	19.6	37.4	7.0	36.0	73.4	14.6	41.8	0.9	42.7
Total	24 220														
Average		47.89	23.1	25.1	25.1	26.7	22.8	32.6	16.9	27.7	60.4	24.3	31.1	17.9	26.7

L, length in nucleotide of the coding sequence; ENC, effective number of codons; GC₃, G+C content at the third codon position. All nucleotide frequencies are given in per cent and are averages over all alleles examined.

Table 5. Codon usage in *Aedes aegypti* nuclear genes

aa	Codon	N	RCSU	aa	Codon	N	RCSU	aa	Codon	N	RCSU	aa	Codon	N	RCSU	aa	Codon	N	RCSU				
Twofold degenerate codons																							
Phe	UUU(F)	88	(0.45)	His	CAU(H)	61	(0.73)	Gln	CAA(Q)	93	(0.66)	Asp	GAU(D)	275	(1.12)	Cys	UGU(C)	61	(0.82)				
	UUC(F)	301	(1.55)		CAC(H)	106	(1.27)		CAG(Q)	190	(1.34)		GAC(D)	218	(0.88)		UGC(C)	88	(1.18)				
Asn	AAU(N)	133	(0.70)	Tyr	UAU(Y)	81	(0.64)	Lys	AAA(K)	183	(0.68)	Glu	GAA(E)	293	(1.21)	Ile	AUU(I)	128	(0.91)				
	AAC(N)	249	(1.30)		UAC(Y)	174	(1.36)		AAG(K)	358	(1.32)		GAG(E)	193	(0.79)		AUC(I)	262	(1.86)				
																	AUA(I)	32	(0.23)				
Fourfold degenerate codons																							
Val	GUU(V)	171	(1.09)	Ala	GCU(A)	206	(1.27)	Thr	ACU(T)	92	(0.88)	Pro	CCU(P)	51	(0.57)	Gly	GGU(G)	159	(1.17)				
	GUC(V)	212	(1.34)		GCC(A)	255	(1.58)		ACC(T)	187	(1.80)		CCC(P)	80	(0.90)		GGC(G)	136	(1.01)				
	GUA(V)	87	(0.55)		GCA(A)	94	(0.58)		ACA(T)	50	(0.48)		CCA(P)	104	(1.17)		GGA(G)	220	(1.63)				
	GUG(V)	161	(1.02)		GCG(A)	93	(0.57)		ACG(T)	87	(0.84)		CCG(P)	121	(1.36)		GGG(G)	26	(0.19)				
Sixfold degenerate codons																							
Leu	CUU(L)	71	(0.61)	Ser	UCU(S)	70	(0.81)	Arg	CGU(R)	143	(2.16)	unique codons				Met	AUG(M)	181	(1.00)	Trp	UGG(W)	95	(1.00)
	CUC(L)	93	(0.80)		UCC(S)	114	(1.32)		CGC(R)	93	(1.40)												
	CUA(L)	40	(0.34)		UCA(S)	57	(0.66)		CGA(R)	61	(0.92)												
	CUG(L)	322	(2.77)		UCG(S)	130	(1.51)		CGG(R)	45	(0.68)					stop	UGA(*)	5	(0.62)				
	UUA(L)	21	(0.18)		AGU(S)	69	(0.80)		AGA(R)	32	(0.48)						UAA(*)	11	(1.38)				
	UUG(L)	151	(1.30)		AGC(S)	78	(0.90)		AGG(R)	24	(0.36)						UAG(*)	8	(1.00)				

Total number of codons examined was 8073; RCSU, relative synonymous codon usage.

D. melanogaster, *E. coli* and *S. cerevisiae* (Besansky, 1993; Grosjean & Fiers, 1982; Sharp & Li, 1986b; Shields *et al.*, 1988). Our results suggest that, as observed for other species, in *Ae. aegypti* the codon usage pattern reflects selection for translational efficiency. Codon usage is biased

toward codons corresponding to the most abundant tRNAs (Duret, 2000; Grantham *et al.*, 1981; Ikemura, 1985; Moriyama & Powell, 1997a; Sharp *et al.*, 1988).

Codon usage for all 25 *Ae. aegypti* genes is presented in Table 5. As for *Drosophila* and *An. gambiae* (Besansky,

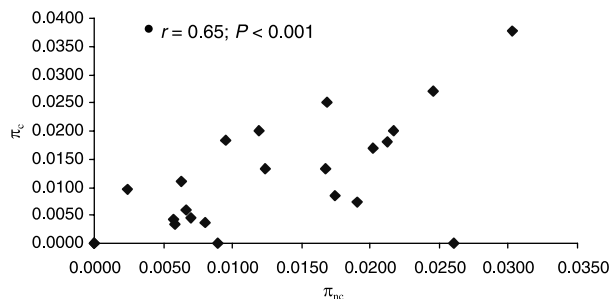


Figure 2. Correlation of nucleotide diversity in coding (π_c) and non-coding regions (π_{nc}).

1993; Powell & Moriyama, 1997), and confirming a previous study in *Ae. aegypti* (Argentine & James, 1993), we observed a preference of codon usage for C- or G-ending triplets. However, not all amino acids reflect a wobble position bias toward G or C. That is, Asp and Arg, Glu and Gly show a third codon position preference for T and A, respectively. The third position T bias in Asp was also observed in *Drosophila* species, but not in *An. gambiae* (Besansky, 1993; Powell & Moriyama, 1997). At synonymous sites where G or C can be used, we observed that C is favoured ($P < 0.01$, Table 5), and if we compile data for the six most biased genes ($ENC < 45$), the preference for C is even more evident, 40.4% vs. 29.5%, respectively. Codon preference also varies considerably from one gene to another, and individual genes can reflect specific preferences. For example, with the highly biased *Ef-2* gene we observed a mean RSCU value of 4.74 for the Arg codon CGU.

Nucleotide diversity was estimated separately for coding and non-coding sequences, and in coding regions nucleotide heterogeneity was calculated at both synonymous and replacement sites. Nucleotide diversity is similar in coding and non-coding regions (Table 1). Moreover, the level of polymorphism in coding and non-coding regions is positively correlated ($r = 0.65$, $P < 0.001$; Fig. 2). This suggests that some factor, such as recombination, may act on DNA sequences at a regional level (Moriyama & Powell, 1996; Nachman, 2001). Higher nucleotide diversity is more often found in regions of high recombination, and recombination 'hot spots' seem a general feature of the human genome (Aquadro *et al.*, 2001; Lercher & Hurst, 2002; Nachman, 2001; Reich *et al.*, 2002). However, the average nucleotide diversity was lower in non-coding regions than at synonymous sites of the coding sequences, 0.0142 vs. 0.0287 ($P = 0.005$). This implies that non-coding regions are under greater purifying selection than synonymous sites within coding regions (Moriyama & Powell, 1996). The 5'-flanking regions may contain regulatory elements that play critical roles in transcription, and single-base mutations can alter essential structures for splicing and processing (Shen *et al.*, 1999). The role, origin and adaptative

significance of introns in eukaryote genes remains controversial (Long *et al.*, 1995; Lynch, 2002). Introns can be involved in maintaining the secondary structure of pre-mRNA (Kirby *et al.*, 1995) and it has been shown that pre-mRNA splicing is an important process for regulating gene expression (Chapman & Walter, 1997; Morello *et al.*, 2002). These structural and functional constraints, even if acting on non-coding DNA, are likely to be associated with high selective pressure.

Nucleotide diversity varies considerably from one gene to another (Table 1), and is likely related to individual gene function and selective constraints. We observed that there is a trend for a positive correlation between the rates of non-synonymous and synonymous substitutions ($r = 0.38$, $P = 0.07$). This is consistent with, but less obvious than that previously reported in *Drosophila* spp. and bacteria (Comeron & Kreitman, 1998; Dunn *et al.*, 2001; Sharp & Li, 1987), and may indicate that the synonymous substitution rates in *Ae. aegypti* are also influenced by selective constraints acting at the amino acid level. Indeed, silent substitutions have been shown to alter mRNA secondary structure and subsequent processing (Shen *et al.*, 1999).

Typically, genes with the strongest selective constraints should show the lowest nucleotide polymorphism. Indeed, we observed the lowest nucleotide diversities for genes coding for highly expressed molecules that are involved in transcriptional or translational regulation (*RpS11*, *RpL31*, *RpL17A*, *AEG18* and *Ef-2*) or in signalling processes (*mRNABP* and *ODC-AZ*). This is also observed in yeast and *Drosophila*; for example, in *Drosophila* spp., substitution rates between conservative genes and fast evolving genes differ by about 10-fold (Moriyama & Powell, 1997b; Pal *et al.*, 2001; Schmid & Tautz, 1997).

With genes for which expression patterns in *Ae. aegypti* tissues are known (Table 6), we compared the substitution rates relative to their general expression patterns. Thirteen genes were expressed in a particular tissue location. Substitution rates are significantly different between tissue-specific and ubiquitously expressed genes (0.0183 and 0.0049, respectively, $P < 0.001$), with tissue-specific genes showing higher polymorphism rates. In mammals, tissue-specific genes show threefold more replacement substitutions than ubiquitous genes (Duret & Mouchiroud, 2000). However, because synonymous base changes are presumably not constrained by selection in mammals, silent substitution rates do not vary with the expression pattern (Duret & Mouchiroud, 2000). For *Ae. aegypti*, we compared the overall nucleotide diversities, although we also observed significant differences when silent and replacement sites were considered separately (data not shown).

Finally, genes involved in specific adaptations that evolve very rapidly, such as defence mechanisms against parasites, are likely to exhibit high levels of polymorphism. To investigate this assumption, we examined genes that are

Table 6. Characteristics of genes used in this study

Gene	Identity	N _{al}	Strains	GENBANK accession nos.	Tissue expression	Reference
<i>Fer (H)</i>	Ferritin, Heavy chain	4	R, M	AF326341-2, AY064105-6	gut	Morlais <i>et al.</i> (2003)
<i>AEI8</i>	Transcription factor	4	R, M	AF326339-40, AY064076-7	ubiquitous	...
<i>Sec61</i>	Sec61 isoform	4	R, M	AF326338, AF392805, AY064124-5	ubiquitous	...
<i>AelMUC1</i>	Mucin-like protein	12	R, M	AF387486, AY008350-2, AY064110-7, AY133345	gut, Pg+	Morlais & Severson (2001)
<i>Chym</i>	Chymotrypsin	8	R, M	AY038039-40, AY008348-9, AY064082-5	gut	Morlais <i>et al.</i> (2003)
<i>pG12</i>	Protein G12 precursor	4	R, M	AY009155-6, AY038041-2	gut, Pg+	...
<i>RpS11</i>	Ribosomal protein	1	R, M	AF315552, AY133344	ubiquitous	...
<i>RpL31</i>	Ribosomal protein	1	R, M	AF324863, AY009157	ubiquitous	...
<i>RpL17A</i>	Ribosomal protein	3	R, M	AF315596-7, AF399675, AY064121	ubiquitous	...
<i>AEI23</i>	unknown	5	R, M, L	AY081830-1, AY033624-5, AY064074-5	\	...
<i>PGK</i>	Phosphoglycerate kinase	6	R, M	AY043171-2, AY064542-5	ubiquitous	...
<i>ODC-AZ</i>	Ornithine decarboxylase antizyme	3	R, M	AF396870-1, AY064120, AY133346	ubiquitous	...
<i>CYP9J</i>	Cytochrome P450	4	R, M	AF329892, AF390099, AY064092-3	gut, Pg+	...
<i>CRALBP</i>	Cellular retinaldehyde-binding protein	8	R, M	AF329893, AF390101, AY064086-91	gut, Pg+	...
<i>Ef-2</i>	Elongation factor	4	R, M	AF331798, AY040342, AY064103-4	ubiquitous	...
<i>AEGBS11</i>	unknown	4	R, M	AY033622-3, AY064072-3	gut	...
<i>APN</i>	Aminopeptidase N	4	R, M	AF378117, AF390100, AY064078-9	gut, Pg+	...
<i>SDR</i>	Short-chain dehydrogenase/reductase	4	R, M	AY033621, AY033626, AY064122-3	gut, Pg+	...
<i>mRNABP</i>	mRNA-binding protein	4	R, M	AY033620, AY064107-9	ubiquitous	...
<i>CecA</i>	Cecropin A	2	R, M	AY064080-1	haemolymph	Lowenberger <i>et al.</i> (1999)
<i>TSF</i>	Transferrin	6	R, M, L	AF387489, AY064537-41	haemocoel, Fil+	Beerntsen <i>et al.</i> (1994)
<i>DefA</i>	Defensin A	2	R, M, L	AF392802-4	haemolymph, fat body	Lowenberger <i>et al.</i> (1995)
<i>DCE</i>	Dopachrome conversion enzyme	6	R, M, L	AY064094-9	haemolymph, ovaries, Fil+	Li <i>et al.</i> (1994)
<i>DDC</i>	Dopa decarboxylase	2	R, M, L	AY064100-2, U27581	haemolymph, ovaries	Ferdig <i>et al.</i> (1996)
<i>NaK</i>	Sodium/potassium channel	2	R, M	AY064118-9	\	none

N_{al}, number of alleles; R, Red-eye, M, Moyo, R, L, Liverpool; tissue expression for AEI23 and NaK genes is unknown; Pg+ and Fil+ indicate genes that are up-regulated upon *P. gallinaceum* and filaria infection, respectively.

known or suspected to play a role in the immune response of *Ae. aegypti* to metazoan parasites including *Plasmodium gallinaceum* and *Brugia malayi* (Table 6 and Fig. 2). Note that all genes induced in response to parasite infection ($n = 8$) are also tissue-specific. Interestingly, we found that the substitution rates for genes associated with parasite defence are significantly higher than the overall substitution rate (0.0205 vs. 0.0132, respectively, $P = 0.05$). In *Plasmodium*, it has been shown that parasite surface proteins are under diversifying selection to evade the host immune system (Hughes & Hughes, 1995). Therefore, genes with a putative function in host–parasite interactions reflect higher levels of nucleotide variation, consistent with the hypothesis that genes induced in response to parasites are subject to diversifying selection.

Experimental procedures

Sequences of 25 nuclear genes were obtained from the Red-eye, Moyo-R and Liverpool laboratory strains of *Ae. aegypti*. The Liverpool strain was originally obtained from the London School of Hygiene and Tropical Medicine and is permissive to filarial nematode parasites. The Red-eye strain is a mutant marker strain and is refractory to filaria but is an efficient vector of *P. gallinaceum*. The Moyo-R strain was genetically selected from the Kenyan Moyo-in-Dry strain for its refractoriness to *P. gallinaceum* (Thathy *et al.*, 1994). Genomic DNA was prepared from single mosquitoes and cDNA from pools of six females as previously described (Severson, 1997; Morlais & Severson, 2001). The loci were amplified by PCR using specific primers. The optimal primer annealing temperatures were established via gradient PCR using an Eppendorf Mastercycler (Eppendorf). PCR reactions were performed in a final volume of 25 μ l containing 50 mM Tris-HCl pH 8.3, 3 mM $MgCl_2$, 50 mM KCl, 400 μ M of each dNTP, 0.25 U of *Taq* polymerase, 10 pmol of each primer and 5 ng of template DNA. Cycling conditions were: 5 min at 95 °C, then 30 cycles of 1 min at 95 °C, 1 min at the optimal annealing temperature, 2 min at 72 °C, and final extension of 10 min at 72 °C. The PCR products were purified using the QIAquick PCR purification kit (Qiagen). Purified PCR products from individual mosquito genomic DNAs were directly sequenced on both strands using an ABI 310 with Big Dye Terminators (PE Applied BioSystems). Purified cDNAs were cloned using the pCR-2.1 TOPO TA cloning kit (Invitrogen) prior to sequencing. Two clones were sequenced for each cDNA. A list of the sequenced genes with their accession numbers is provided in Table 6. Two sequences, *DDC* and *NaK*, were partial and were not included in our correlation analyses.

The allele sequences for each gene were aligned using CLUSTAL W (Thompson *et al.*, 1994). Each alignment comprised the same number of alleles for each strain, except for the *AeIMUC1* gene. SNPs were identified as transitions or transversions for both coding and non-coding regions, and for SNPs occurring in coding sequences, nucleotide variations were also classified as synonymous or non-synonymous. Nucleotide diversity analyses were conducted using MEGA version 2.1 (Kumar *et al.*, 2001). The average number of nucleotide substitutions per site, π , was calculated for each gene, while estimates of synonymous and non-synonymous substitution rates, K_s and K_a , were computed following the method of Nei & Kumar (2000) which corrects for transition/transversion

bias and degenerate sites. Codon usage was analysed using MEGA ver. 2.1 (Kumar *et al.*, 2001). The program calculates the nucleotide composition for each codon position and the relative synonymous codon usage (RCSU). The RCSU is given as the observed frequency of a codon relative to its expected frequency under the assumption of equal codon usage (Sharp & Li, 1986b). For each gene, codon frequencies were calculated as averages over all alleles examined. The effective number of codons in a gene, ENC (Wright, 1990), was determined using the CHIPS program from the EMBOSS package <<http://bioinfo.pbi.nrc.ca:8090/cgi-bin/emboss>>. Correlation coefficients were calculated with the assumption that both variables are stochastic using the EPI INFO package.

Acknowledgements

We thank H. Hollocher and J. Romero-Severson for helpful suggestions on the manuscript. This study was supported by National Institutes of Health Grants RO1 AI33127 and RO1 AI34337.

References

- Akashi, H. (2001) Gene expression and molecular evolution. *Curr Opin Genet Dev* **11**: 660–666.
- Antolin, M.F., Bosio, C.F., Cotton, J., Sweeney, W., Strand, M.R. and Black, W.C.I.V. (1996) Intensive linkage mapping in a wasp (*Bracon hebetor*) and a mosquito (*Aedes aegypti*) with single-strand conformation polymorphism analysis of random amplified polymorphic DNA markers. *Genetics* **143**: 1727–1738.
- Aquadro, C.F., Bauer DuMont, V. and Reed, F.A. (2001) Genome-wide variation in the human and fruitfly: a comparison. *Curr Opin Genet Dev* **11**: 627–634.
- Argentine, J.A. and James, A.A. (1993) Codon preference of *Aedes aegypti* and *Aedes albopictus*. *Insect Mol Biol* **1**: 189–194.
- Berntsen, B.T., Severson, D.W., Kinkhammer, J.A., Kassner, V.A. and Christensen, B.M. (1994) *Aedes aegypti*: characterization of a hemolymph polypeptide expressed during melanotic encapsulation of filarial worms. *Exp Parasitol* **79**: 312–321.
- Berger, J., Suzuki, T., Senti, K.A., Stubbs, J., Schaffner, G. and Dickson, B.J. (2001) Genetic mapping with SNP markers in *Drosophila*. *Nat Genet* **29**: 475–481.
- Besansky, N.J. (1993) Codon usage patterns in chromosomal and retrotransposon genes of the mosquito *Anopheles gambiae*. *Insect Mol Biol* **1**: 171–178.
- Bhatramakki, D., Dolan, M., Hanafey, M., Wineland, R., Vaske, D., Register, J.C. 3rd, Tingey, S.V. and Rafalski, A. (2002) Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. *Plant Mol Biol* **48**: 539–547.
- Bielawski, J.P., Dunn, K.A. and Yang, Z. (2000) Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. *Genetics* **156**: 1299–1308.
- Black, W.C., Baer, C.F., Antolin, M.F. and DuTeau, N.M. (2001) Population genomics: genome-wide sampling of insect populations. *Annu Rev Entomol* **46**: 441–469.
- Brookes, A.J. (1999) The essence of SNPs. *Gene* **234**: 177–186.
- Chapman, R.E. and Walter, P. (1997) Translational attenuation mediated by an mRNA intron. *Curr Biol* **7**: 850–859.

- Comeron, J.M. and Aguade, M. (1998) An evaluation of measures of synonymous codon usage bias. *J Mol Evol* **47**: 268–274.
- Comeron, J.M. and Kreitman, M. (1998) The correlation between synonymous and nonsynonymous substitutions in *Drosophila*: mutation, selection or relaxed constraints? *Genetics* **150**: 767–775.
- Dunn, K.A., Bielawski, J.P. and Yang, Z. (2001) Substitution rates in *Drosophila* nuclear genes: implications for translational selection. *Genetics* **157**: 295–305.
- Duret, L. (2000) tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet* **16**: 287–289.
- Duret, L. and Mouchiroud, D. (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* **17**: 68–74.
- Fagerberg, A.J., Fulton, R.E. and Black, W.C. IV (2001) Microsatellite loci are not abundant in all arthropod genomes: analyses in the hard tick, *Ixodes scapularis* and the yellow fever mosquito, *Aedes aegypti*. *Insect Mol Biol* **10**: 225–236.
- Ferdig, M.T., Li, J., Severson, D.W. and Christensen, B.M. (1996) Mosquito dopa decarboxylase cDNA characterization and bloodmeal-induced ovarian expression. *Insect Mol Biol* **5**: 119–126.
- Fulton, R.E., Salasek, M.L., DuTeau, N.M. and Black, W.C. (2001) SSCP analysis of cDNA markers provides a dense linkage map of the *Aedes aegypti* genome. *Genetics* **158**: 715–726.
- Gouy, M. and Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* **10**: 7055–7074.
- Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res* **9**: r43–74.
- Grosjean, H. and Fiers, W. (1982) Preferential codon usage in prokaryotic genes: the optimal codon–anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* **18**: 199–209.
- Holt, R.A., Subramanian, G.M., Halpern, A., et al. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**: 129–149.
- Hughes, M.K. and Hughes, A.L. (1995) Natural selection on Plasmodium surface proteins. *Mol Biochem Parasitol* **71**: 99–113.
- Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**: 13–34.
- Jakubowski, J. and Kornfeld, K. (1999) A local, high-density, single-nucleotide polymorphism map used to clone *Caenorhabditis elegans* cdf-1. *Genetics* **153**: 743–752.
- Kirby, D.A., Muse, S.V. and Stephan, W. (1995) Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci USA* **92**: 9047–9051.
- Kumar, S., Tamura, K., Jakobsen, I.B. and Nei, M. (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**: 1244–1245.
- Lercher, M.J. and Hurst, L.D. (2002) Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18**: 337–340.
- Li, J., Zhao, X. and Christensen, B.M. (1994) Dopachrome conversion activity in *Aedes aegypti*: significance during melanotic encapsulation of parasites and cuticular tanning. *Insect Biochem Mol Biol* **24**: 1043–1049.
- Lindblad-Toh, K., Winchester, E., Daly, M.J., Wang, D.G., Hirschhorn, J.N., Laviolette, J.P., Ardlie, K., Reich, D.E., Robinson, E., Sklar, P., Shah, N., Thomas, D., Fan, J.B., Gingeras, T., Warrington, J., Patil, N., Hudson, T.J. and Lander, E.S. (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat Genet* **24**: 381–386.
- Llopert, A., Comeron, J.M., Brunet, F.G., Lachaise, D. and Long, M. (2002) Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc Natl Acad Sci USA* **99**: 8121–8126.
- Long, M., de Souza, S.J. and Gilbert, W. (1995) Evolution of the intron-exon structure of eukaryotic genes. *Curr Opin Genet Dev* **5**: 774–778.
- Lowenberger, C.A., Bulet, P., Charlet, M., Hetru, C., Hodgeman, B., Christensen, B.M. and Hoffmann, J.A. (1995) Insect immunity: isolation of three novel inducible antibacterial defensins from the vector mosquito, *Aedes aegypti*. *Insect Biochem Mol Biol* **25**: 867–873.
- Lowenberger, C., Charlet, M., Vizioli, J., Kamal, S., Richman, A., Christensen, B.M. and Bulet, P. (1999) Antimicrobial activity spectrum, cDNA cloning, and mRNA expression of a newly isolated member of the cecropin family from the mosquito vector *Aedes aegypti*. *J Biol Chem* **274**: 20092–20097.
- Lynch, M. (2002) Intron evolution as a population-genetic process. *Proc Natl Acad Sci USA* **99**: 6118–6123.
- Marais, G., Mouchiroud, D. and Duret, L. (2001) Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proc Natl Acad Sci USA* **98**: 5688–5692.
- McVean, G.A. and Vieira, J. (1999) The evolution of codon preferences in *Drosophila*: a maximum-likelihood approach to parameter estimation and hypothesis testing. *J Mol Evol* **49**: 63–75.
- Morello, L., Bardini, M., Sala, F. and Breviaro, D. (2002) A long leader intron of the Ostub16 rice beta-tubulin gene is required for high-level gene expression and can autonomously promote transcription both in vivo and in vitro. *Plant J* **29**: 33–44.
- Moriyama, E.N. and Powell, J.R. (1996) Intraspecific nuclear DNA variation in *Drosophila*. *Mol Biol Evol* **13**: 261–277.
- Moriyama, E.N. and Powell, J.R. (1997a) Codon usage bias and tRNA abundance in *Drosophila*. *J Mol Evol* **45**: 514–523.
- Moriyama, E.N. and Powell, J.R. (1997b) Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes. *J Mol Evol* **45**: 378–391.
- Morlais, I., Mori, A., Schneider, J.R. and Severson, D.W. (2003) Targeted approach toward identification of candidate genes determining *Plasmodium gallinaceum* susceptibility in *Aedes aegypti*. *Mol Genet Genomics*, in press.
- Morlais, I. and Severson, D.W. (2001) Identification of a polymorphic mucin-like gene expressed in the midgut of the mosquito, *Aedes aegypti*, using an integrated bulked segregant and differential display analysis. *Genetics* **158**: 1125–1136.
- Nachman, M.W. (2001) Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet* **17**: 481–485.
- Nei, M. and Kumar, S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
- Pal, C., Papp, B. and Hurst, L.D. (2001) Highly expressed genes in yeast evolve slowly. *Genetics* **158**: 927–931.
- Powell, J.R. and Moriyama, E.N. (1997) Evolution of codon usage bias in *Drosophila*. *Proc Natl Acad Sci USA* **94**: 7784–7790.
- Reich, D.E., Schaffner, S.F., Daly, M.J., McVean, G., Mullikin, J.C., Higgins, J.M., Richter, D.J., Lander, E.S. and Altshuler, D. (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* **5**: 5.
- Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J.,

- Willey, D.L., Hunt, S.E., Cole, C.G., Coggill, P.C., Rice, C.M., Ning, Z., Rogers, J., Bentley, D.R., Kwok, P.Y., Mardis, E.R., Yeh, R.T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R.H., McPherson, J.D., Gilman, B., Schaffner, S., Van Etten, W.J., Reich, D., Higgins, J., Daly, M.J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M.C., Linton, L., Lander, E.S. and Atshuler, D. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.
- Schmid, K.J. and Tautz, D. (1997) A screen for fast evolving genes from *Drosophila*. *Proc Natl Acad Sci USA* **94**: 9746–9750.
- Severson, D.W. (1997) RFLP analysis of insect genomes. In *The Molecular Biology of Insect Disease Vectors: a Methods Manual* (Crampton, J.M., Beard, C.B. and Louis, C., eds), pp. 309–320. Chapman & Hall, London.
- Severson, D.W., Meece, J.K., Lovin, D.D., Saha, G. and Morlais, I. (2002) Linkage map organization of expressed sequence tags and sequence tagged sites in the mosquito, *Aedes aegypti*. *Insect Mol Biol* **11**: 371–378.
- Severson, D.W., Zaitlin, D. and Kassner, V.A. (1999) Targeted identification of markers linked to malaria and filarioid nematode parasite resistance genes in the mosquito *Aedes aegypti*. *Genet Res* **73**: 217–224.
- Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H. and Wright, F. (1988) Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Res* **16**: 8207–8211.
- Sharp, P.M. and Li, W.H. (1986a) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res* **14**: 7737–7749.
- Sharp, P.M. and Li, W.H. (1986b) An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* **24**: 28–38.
- Sharp, P.M. and Li, W.H. (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* **4**: 222–230.
- Sharp, P.M., Tuohy, T.M. and Mosurski, K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14**: 5125–5143.
- Shen, L.X., Basilion, J.P. and Stanton, V.P. Jr (1999) Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc Natl Acad Sci USA* **96**: 7871–7876.
- Shields, D.C., Sharp, P.M., Higgins, D.G. and Wright, F. (1988) 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol Biol Evol* **5**: 704–716.
- Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L. and Kwok, P.Y. (1998) Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res* **8**: 748–754.
- Taillon-Miller, P., Piernot, E.E. and Kwok, P.Y. (1999) Efficient approach to unique single-nucleotide polymorphism discovery. *Genome Res* **9**: 499–505.
- Thathy, V., Severson, D.W. and Christensen, B.M. (1994) Reinterpretation of the genetics of susceptibility of *Aedes aegypti* to *Plasmodium gallinaceum*. *J Parasitol* **80**: 705–712.
- Thompson, J.D., Higgins, D., G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680.
- Wang, D.G., Fan, J.B., Siao, C.J., et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077–1082.
- Wright, F. (1990) The 'effective number of codons' used in a gene. *Gene* **87**: 23–29.