

## Arthropod gene structure statistics

	<i>Daphnia</i> <sup>1</sup>	<i>Aphid</i> <sup>1</sup>	<i>Bee</i>	<i>Wasp</i> <sup>1</sup>	<i>Moth</i> <sup>1</sup>	<i>Beetle</i>	<i>Mosquito</i> <sup>1</sup>	<i>Fruitfly</i> <sup>1</sup>	<i>Tick</i> <sup>1</sup>	<i>Mouse</i>	<i>Worm</i>
Genome size Mbp	200 (175)	460 (350)	220 (200)	290 (250)	480 (426)	180 (177)	580 (400)	180 (120)	1,760 (753)	3,450 (2,600)	100 (100)
No. of genes	32,000	32,800	17,000	27,300	16,300	16,400	18,900	13,700	20,500	27,600	20,100
Gene density	0.175	0.063	0.040	0.120	0.042	0.100	0.055	0.168	0.023	0.015	0.250
Gene length	2,300	6,100	9,900	4,600	7,700	5,600	4,800	4,000	12,500	32,000	3,000
CDS size	1,360	1,340	1,690	1,620	1,460	1,420	1,400	1,650	1,070	2,140	1,300
Exons/gene	6.6	6.7	7.1	6.3	6.4	4.5	3.5	4.0	5.7	8.0	6.0
Exon size <sup>2</sup>	210	200	240	260	230	310	420	410	190	280	200
Intron size <sup>3</sup>	72	75/900	88/600	81/530	810/97	51/1700	64/1900	63/750	1730/90	1600/90	51/500
Mean	170	790	770	430	1150	1000	1400	660	2400	2800	290
Intr > Exon	10%	41%	36%	24%	86%	34%	36%	27%	87%	85%	33%
UTR size <sup>4</sup>	370	500	340	680	440	--	240	800	540	--	260
Alternate Tr.	10%	24%	--	23%	18%	--	--	36%	15%	65% <sup>5</sup>	20%
Intergenic size	4,000	7,100	21,600	--	23,300	8,500	18,200	5,400	26,700	78,000	2,400

Updated 2009/12/04. Compiled by D.Gilbert, gilbertd@indiana.edu.

**Genome size** value in parentheses is total gene-containing sequence (i.e. excluding heterochromatin, scaffolds without genes, etc.). **No. of genes** is from the gene set examined, not necessarily the official gene set for new genomes. **Gene density** is calculated as the sum of coding exon bases / total gene-containing genome bases. **Gene length** is the span including introns and UTR. **CDS size** is the coding sequence length without introns or UTRs. **Exons/gene** and **Exon size** are count and size of coding exons. Sizes are given as mean in bp except for Intron size. **Alternate Tr** is percent of genes with alternate transcripts measured from EST assemblies. **Intergenic size** is measured from distance between adjacent genes. These statistics have a standard deviation close to the mean, but Intergenic size has a much larger variance.

<sup>1</sup> Gene part sizes and exons/gene are measured with EST-validated gene models for these noted genomes. Others are measured from reference database gene feature data.

<sup>2</sup> **Exon size** distribution for *Drosophila* is strongly bimodal; one-exon genes average twice the size of multi-exon genes (830 bp versus 470 bp/exon). Other species show unimodal distribution of exon sizes.

<sup>3</sup> **Intron size** is non-normally distributed. Intron size lists the primary and secondary peaks, mean and the percent of introns larger than exons. It has a narrow, high peak frequency at the indicated (median) value. Fruitfly and nematode have a secondary peak at about 400 bp, mouse reverses this with its secondary peak at 90 bp. *Daphnia* appears to have no secondary intron size peak.

<sup>4</sup> **UTR size** is an over-estimate, as it is measured only where exons extend past coding sequence, and misses true cases of zero length UTRs.

<sup>5</sup> **Alternate Tr.** for mouse is from Le Texier et al., BMC Bioinformatics 2006, 7:16 doi:10.1186/1471-2105-7-169

Species: Aphid = *Acyrtosiphon pisum* (acyr1); Beetle = *Tribolium castenatum* (tcas3); Bee = *Apis mellifera* (ncbi1); Daphnia = *Daphnia pulex* (daphx1); Fruitfly = *Drosophila melanogaster* (fb5.5); Tick = *Ixodes scapularis* (IscaW1); Mosquito = *Culex pipens* (cpip12); Moth = *Bombyx mori* (silkworm r2); Mouse = *Mus musculus* (mgi3); Wasp = *Nasonia vitripennis* (nvit1); Worm = *Caen. elegans* (wb167);

For data of EST validated gene models, see PASA-EST-assemblies.html at <http://insects.eugenes.org/arthropods/summaries/>