Supplementary files to accompany "Genome Scale Analysis of Positionally Relocated Genes" (Arjun Bhutkar, Susan M. Russo, Temple F. Smith, William M. Gelbart[1]

List of supplementary files:

1. supplementary_all_PRG_data.xml
2. supplementary_mel_embedded_relationships_inferred_ancestral.xml
3. supplementary_mel_to_anoph_ansc_co_located_genes.xml
4. supplementary_PRGs_expr_in_testes.xml

Description of supplementary files:

1. supplementary_all_PRG_data.xml

   Genes in the set of 1383 PRG candidates (Fig. 1b) are listed. There are 6 sheets in this Excel file:

   - single_species_hi_conf_one_gene: One-gene single-species PRGs
   - single_sp_hi_conf_multi_gene: Multi-gene single-species PRGs
   - lineage_supp_hi_conf_one_gene: One-gene lineage-supported PRGs
   - lineage_supp_hi_conf_multi_gene: Multi-gene lineage-supported PRGs
   - single_species_low_conf: Single-species low confidence set
   - lineage_supp_lo_conf: Multiple-species low confidence set

   Columns in each sheet are as follows:

   - xlation_start_order_id: Internal code for gene
   - Mel_gene: *D. melanogaster* CG id for the PRG candidate
   - Classification: PRG category for Fig. 1b. *See below*.
   - Comments: Internal evidence code (2-COPY implies that there was evidence to classify the PRG as being part of a duplicative transposition)
   - SYN_ARM: Synpipe (see text) *Drosophila* Muller element assignment for the gene in each species. Species are in the order listed below. A "-" implies that the gene does not have a placement in that species. "Q" implies that the gene was inferred to lie in an assembly gap.
   - SYN_EXONS: Number of exons in the Synpipe ortholog. Exon count was determined from a GLEANR gene model in the corresponding location. Automated gene prediction sometimes results in multiple gene models where a single gene is inferred. Exon counts of all such models are listed (separated by ";") for the Synpipe extents of the ortholog. Species are in the order listed below.

- SYN_FLAGS: Confidence estimate of Synpipe ortholog assignment (numbers 0 or greater indicate higher confidence). Numbers less than 0 indicated paralogs hits or probable locations in assembly gaps.
- SYN_IN_SPECIES: Presence or absence (“-“) codes in each species (for an ortholog of the gene). Species are in the order listed below (with codes).
- GLEAN_ARM: *Drosophila* Muller element inferred for the gene in each species based on majority GLEANR orthologs assigned to the corresponding scaffold. Species are in the order listed below. A “-“ implies that the gene does not have a placement in that species.
- GLEAN_EXONS: Number of exons in the gene model of the corresponding GLEANR ortholog.
- GLEAN_IN_SPECIES: Presence or absence (“-“) codes in each species (for an ortholog of the gene). Species are in the order listed below (with codes).

Species order(code): *D. melanogaster(m),D. sechellia(S), D. simulans(s), D. yakuba(y), D. erecta(e), D. ananassae(a), D. pseudoobscura(P), D. persimilis(p), D. willistoni(w), D. virilis(v), D. mojavensis(m; no conflict with D. melanogaster due to position in order), D. grimshawi(g)*

Classification column:
3.x: Single-species one-gene PRGs
4.x: Lineage-supported one-gene PRGs
8.x: Single-species multi-gene PRGs
5.x: Lineage-supported multi-gene PRGs
10: Single-species one-gene PRGs where PRG has 2 copies (original & relocated)
11: Lineage-supported one-gene PRGs where PRG has 2 copies (original & relocated)
***Where*:**
x =1 implies ancestral state is multi-exon and derived state is 1-exon
x= 3 implies ancestral state is multi-exon and derived state is multi-exon
x= 5 implies ancestral state is 1-exon and derived state is 1-exon
x= 7 implies ancestral state is 1-exon and derived state is multi-exon

All others codes are part of the low-confidence set.

2. supplementary_mel_embedded_relationships_inferred_ancestral.xml

   Lists the 763 embedded relationships found in *D. melanogaster* annotation Release_4.3. For each relationship the “surrounding gene” and the “embedded gene is listed (FlyBase CG id). Additionally, the third column specifies whether the relationship was inferred to be ancestral i.e. (at the genus

Drosophila root; 1 = inferred ancestral). 544 such relationships were inferred (Fig. 3c). See Methods for details.

3. supplementary_mel_to_anoph_ansc_co_located_genes.xml

   Lists the 5653 *D. melanogaster* genes from the boxed areas in Fig. 4(a) (see main text). These are genes inferred to have been arm-conserved between *Drosophila* and *A. gambiae* and are used to study genome organization in *A. mellifera* (Fig. 4(c)). *A. gambiae* and *D. melanogaster* chromosome arms are listed for each gene.

4. supplementary_PRGs_expr_in_testes.xml

   Testes expression was determined for two datasets:

   The first datasheet in this file tags the lineage supported 87 PRGs with single exon derived states (94 total genes accounting for *D. melanogaster* lineage duplicative transpositions with distinct gene ids.) depending on whether or not they are expressed in *D. melanogaster* testes (based on data from FlyAtlas and FlyMine – see main text). A total of 39 of these (42%) are found to be expressed in testes.

   A second datasheet in this file shows similar information for lineage supported DNA-transposed genes. 4 out of 16 genes (25%) are found to be expressed in the testes.
