Ab initio construction and evolutionary analysis of protein-coding gene families with partial homologous relationships:
The closely related Drosophila genomes as a case study

Supplementary Data
Table S1. Reconstructed architecture scenarios (RAS) of ACC.

|                      | Number of ACC | Number of Wu's ACC |
|----------------------|---------------|--------------------|
| Total                | 4,173         | 4,107              |
| With RAS             | 4,018         | 3,882              |
| Without RAS          | 155           | 225                |
| Consistent with Wu's | 3,882         | -                  |
| Consistent with Wu's & with RAS | 3,708 | -                  |
| Need to be reconstructed | 465   | -                  |
| Successfully reconstructed | 310    | -                  |

With RAS: an ACC with reconstructed architecture scenario.
Without RAS: an ACC without reconstructed architecture scenario.
Consistent with Wu's: an ACC consistent with the ACC constructed by Wu.
Consistent with Wu's & with RAS: the ACC consistent with Wu's ACC and also with a reconstructed architecture scenario (inferred by Wu).
Need to be reconstructed: the remaining ACC we need to reconstruct its architecture scenario.
Successfully reconstructed: the remaining ACC we had reconstructed its architecture scenario.
Notions

$T$: the rooted binary species tree, with nodes $N(T)$ and leave nodes $L(T)$.

$g \in G$: an architecture connected component (ACC) belong to the set of ACCs which have reconstructed architecture scenarios.

$A$: architectures with single-module $A^{sm}$ or multi-modules $A^{mm}$.

$m \in A$: the modules of an architecture.

$M^*$: a set of novel modules with generation events.

Pseudocode

**Supplementary algorithm: RASfam pseudocode.**

**Input:** $T$, $g \in G$, set of $A$ at $n \in N(T)$, $M^*$

**Output:** A set of gene families $F$.

/* Initialization

$O$ = empty ancestors/origins

$F$ = empty families

/* Recursion

for $g \in G$ do

    // Determine the ancestors of modules

    flag = 0

    for $n \in N(T)$ (pre-order traversal) and $n \notin L(T)$ do

        // Determine the most ancient architecture/architectures

        if flag == 0 \&\& A(n) exist then

            flag = 1

            for $A_i \in A(n)$ do

                for $m_i \in m(A_i)$ do

                    $O[m_i] \leftarrow A_i$


                end

            end

        end

    end

    // Determine the novel architecture/architectures generated in internal branch

    if flag == 1 \&\& $m_i(n) \in M^*$ then

        if $m_i \in A_i^{mm}$ then

            $O[m_i] \leftarrow m_i$

        else

            if $m_i \in A_i^{mm}$ \&\& all other modules of $A_i^{mm}$ are in $M^*$ then

                $O[m_i] = A_i^{mm}$

                $O[each\ of\ other\ modules] = A_i^{mm}$

            else if $m_i \in A_i^{mm}$ \&\& there is a $m_j \in A_i^{mm}$ not in $M^*$ then

                $O[m_i] = O[m_j]$

        end

    end

    // Assignment extant architectures

    for $A_i \in A(g)$ do

        for $m_i \in m(A_i)$ do

            $F[O[m_i]] \leftarrow A_i$

        end

    end

end
Table S2. Constructed homologous gene family.

|                         | Number of gene families | Number of ACC |
|-------------------------|-------------------------|---------------|
| Total                   | 15222                   | -             |
| Derived from AST        | 10145                   | -             |
| Derived from ACC        | 5077                    | 4035          |
| Derived from ACC which is without RAS | 17                      | 17            |
| Derived from ACC which is with RAS | 5060                    | 4018          |
| Share partial homologs with other families | 1832                    | 790           |

Derived from ACC which is without RAS: A gene family derived from an ACC which is strongly connected although without RAS.

Table S3. The assignment of MMA within comparable ACCs.

| Method        | Comparable ACCs | MMA solely divided | MMA partially divided | MMA partially and solely divided |
|---------------|-----------------|--------------------|-----------------------|----------------------------------|
| OrthoFinder   | 468             | 464                | 1                     | 3                                |
| CompositeSearch | 201           | 161                | 7                     | 33                               |

Comparable ACCs: ACCs which derive families that can share common proteins with others, and their corresponding protein set is consistent with the protein set of one or several orthogroups or compositefamilies.

MMA solely divided: proteins with an MMA are assigned into a unique family.

MMA partially divided: proteins with an MMA are assigned into different families.

MMA partially and solely divided: some MMAs in an ACC are assigned into a unique family and some MMAs are partially divided into different families.
Table S5. The expanded families with annotated functions identical to previous study.

| Expanded Species | Family annotation                                      | Family ID                  |
|------------------|--------------------------------------------------------|---------------------------|
| dmel             | structural constituent of cytoskeleton                 | AST.5643                  |
| dmel             | Sdic/Cdic                                              | ACC-derived.3540          |
| dyak             | centrosome                                             | ACC-derived.30            |
| dyak             | chromosome segregation                                 | ACC-derived.615           |
| dyak             | defense response                                       | ACC-derived.1942          |
| dyak             | proteolysis                                            | ACC-derived.2443          |
| dere             | nucleus                                                | ACC-derived.1341;         |
| dere             | oxidoreductase activity                                | ACC-derived.1422          |
| dana             | oxidoreductase activity                                | ACC-derived.1438          |
| dana             | proteolysis                                            | ACC-derived.249           |
| dana             | serine-type endopeptidase activity                     | ACC-derived.3799          |
| dpse             | protein serine/threonine kinase activity               | ACC-derived.966           |
| dwil             | glucuronosyltransferase activity                       | ACC-derived.790           |
| dwil             | juvenile-hormone esterase activity                     | ACC-derived.845           |
| dwil             | membrane                                               | AST.8728                  |
| dwil             | proteolysis                                            | ACC-derived.126           |
| dmoj             | oxidoreductase activity                                | ACC-derived.904           |
| dgri             | hydrolase activity, acting on ester bonds              | ACC-derived.1810          |
| dgri             | odorant binding                                        | ACC-derived.112; ACC-derived.4179; ACC-derived.4260; AST.4133 |
| dgri             | olfactory receptor activity                            | ACC-derived.112; ACC-derived.4260 |
Table S6. Branch numbers of architecture change patterns in 38,491 branches.

| foldchange | novel | duplicated | lost | number of branches | n1 / 38,491 * 100(%) | n1 / n2 * 100(%) | n1 / n3 * 100(%) |
|------------|-------|------------|------|--------------------|----------------------|-------------------|-------------------|
| <1         | -     | ✓          | ✓    | 35                 | 0.090                | 2.728             | 0.541             |
|            | ✓     | -          | ✓    | 1,244              | 3.195                | 96.960            | 19.221            |
|            | ✓     | ✓          | ✓    | 4                  | 0.010                | 0.312             | 0.062             |
|            | -     | -          | ✓    | 5,189              | 13.325               | 100.000           | 80.176            |
| 1          | -     | ✓          | ✓    | 117                | 0.300                | 2.802             | 0.388             |
|            | ✓     | -          | ✓    | 4,053              | 10.408               | 97.078            | 13.433            |
|            | ✓     | ✓          | ✓    | 5                  | 0.013                | 0.120             | 0.017             |
|            | -     | -          | -    | 25,998             | 66.763               | 100.000           | 86.163            |
| >1         | -     | ✓          | ✓    | 25                 | 0.064                | 4.039             | 1.094             |
|            | ✓     | -          | ✓    | 565                | 1.451                | 91.276            | 24.716            |
|            | ✓     | ✓          | ✓    | 29                 | 0.074                | 4.685             | 1.269             |
|            | -     | ✓          | -    | 443                | 1.138                | 26.575            | 19.379            |
|            | ✓     | -          | -    | 1,150              | 2.953                | 68.986            | 50.306            |
|            | ✓     | ✓          | -    | 74                 | 0.190                | 4.439             | 3.237             |
|            | -     | ✓          | -    | 10                 | 0.026                | 100.000           | 100.000           |
Figure S1. Ratio of expanded ACC-derived families in species-specific expanded families.
Figure S2. The reconstructed architecture scenario of Sdic family. The expansion lineage is marked with red frame. Different modules are indicated by different colored blocks with module ID. Protein ID and number are show next to the respective module architectures.

Figure S3. The reconstructed architecture scenario of Cyp6a21-PA/Cyp6a9-PA family.
Figure S4. The reconstructed architecture scenario of Ugt37c1-PA/Ugt36Ba-PA family.

Figure S5. The reconstructed architecture scenario of Cyp6a4-PA family.
Figure S6. The reconstructed architecture scenario of Obp58b-PA family and Or42a-PA.
Figure S7. Distribution of foldchanges in all ACC-derived families. Color denotes the respective species.
Figure S8. The summarized enriched GO terms of gene families unchanged only in (A) *Drosophila* genus, (B) *Drosophila* subgenus and (C) *Sophophora* subgenus. GO enrichment analysis was done by R package clusterprofile. The enriched GO terms and the corresponding Benjamini & Hochberg (B & H) corrected *p* values were used by REVIGO method to summarize by removing redundant GO terms, and visualize semantic clusters of the top scoring GO terms.
Figure S9. Distribution of gene families with different evolutionary details in each species. ‘duplicated’ and ‘novel’ denote the duplicated and the novel genes respectively. Families unchanged or only with loss event in corresponding species are not shown.

Figure S10. Distribution of evolutionary events along each branch.
Table S7. Distribution of evolutionary events along each branch.

| Branch                  | Gene level | Subgene (module) level |
|-------------------------|------------|------------------------|
|                         | Duplication | Loss | Duplication | Loss | Merge | Split | Generation |
| Dmel                    | 67          | 2,991 | 155         | 1,047 | 435   | 237   | 0          |
| The MRCA of dmel and dyak | 2,777      | 479  | 421         | 230   | 497   | 327   | 452        |
| Dere                    | 61          | 2,861 | 134         | 875   | 224   | 366   | 0          |
| The MRCA of dere and dyak | 10         | 107  | 27          | 352   | 249   | 169   | 149        |
| Dyak                    | 348         | 2,839 | 576         | 914   | 256   | 674   | 0          |
| The MRCA of dmel and dana | 115        | 865  | 95          | 187   | 315   | 166   | 287        |
| Dana                    | 104         | 686  | 347         | 1,131 | 311   | 544   | 0          |
| The MRCA of dmel and dpse | 269        | 402  | 88          | 195   | 287   | 143   | 359        |
| Dpse                    | 237         | 1,426 | 485         | 1,150 | 354   | 554   | 0          |
| The MRCA of dmel and dwil | 884        | 163  | 65          | 2     | 196   | 85    | 291        |
| Dwil                    | 245         | 1,514 | 500         | 1,177 | 303   | 556   | 0          |
| Dmoj                    | 130         | 1,969 | 247         | 999   | 286   | 571   | 0          |
| The MRCA of dmoj and dvir | 27         | 246  | 86          | 264   | 274   | 161   | 184        |
| Dvir                    | 96          | 1,898 | 208         | 834   | 292   | 429   | 0          |
| The MRCA of dmoj and dgri | 1,713      | 218  | 297         | 5     | 285   | 207   | 351        |
| Dgri                    | 360         | 2,195 | 842         | 1,054 | 267   | 774   | 0          |
| Terminal branch         | 1,648       | 18,379 | 3,494       | 9,181 | 2,728 | 4,705 | 0          |
|                         | 20,027      | 20,108 |             |       |       |       |            |
| Internal branch         | 5,795       | 2,480 | 1,079       | 1,235 | 2,103 | 1,258 | 2,073      |
|                         | 8,275       | 7,748  |             |       |       |       |            |
| Total                   | 7,443       | 20,859 | 4,573       | 10,416 | 4,831 | 5,963 | 2,073      |
|                         | 28,302      | 27,856 |             |       |       |       |            |
Table S8. Distribution of module rearrangement events for gene families expanded along the corresponding branch.

| species | S-L | S-L-D | D-M | S-D | S-L-M | S-D-M | S-M | S-D-L-M | S | D | total |
|---------|-----|-------|-----|-----|-------|-------|-----|---------|---|---|-------|
| **Multi-modules architectures formation** |
| dmel    | 0   | 0     | 1   | 0   | 1     | 0     | 0   | 6       | 8 |   | 281   |
| dere    | 0   | 0     | 1   | 0   | 3     | 0     | 0   | 1       | 0 | 6 | 11    |
| dyak    | 1   | 1     | 2   | 9   | 0     | 1     | 0   | 0       | 0 | 40 | 54    |
| dana    | 0   | 1     | 1   | 4   | 4     | 1     | 0   | 1       | 16| 29|       |
| dpse    | 0   | 1     | 2   | 4   | 0     | 1     | 0   | 0       | 25| 33|       |
| dwil    | 0   | 2     | 2   | 10  | 0     | 2     | 0   | 1       | 27| 45|       |
| dmoj    | 1   | 0     | 0   | 2   | 0     | 0     | 0   | 3       | 10| 16|       |
| dvir    | 0   | 0     | 2   | 0   | 1     | 3     | 0   | 0       | 16| 22|       |
| dgri    | 0   | 1     | 0   | 20  | 0     | 2     | 0   | 1       | 39| 63|       |
| **Both of single-module and multi-modules architectures formation** |
| dmel    | 1   | 0     | 1   | 3   | 0     | 3     | 0   | 0       | 0 | 8 | 237   |
| dere    | 2   | 0     | 0   | 4   | 1     | 1     | 0   | 0       | 2 | 0 | 10    |
| dyak    | 2   | 4     | 0   | 12  | 1     | 4     | 0   | 1       | 12| 28|       |
| dana    | 4   | 1     | 0   | 9   | 0     | 4     | 3   | 0       | 4 | 25|       |
| dpse    | 7   | 2     | 0   | 9   | 1     | 1     | 1   | 0       | 8 | 1 | 30    |
| dwil    | 2   | 6     | 1   | 13  | 0     | 1     | 1   | 0       | 3 | 28|       |
| dmoj    | 2   | 2     | 0   | 9   | 1     | 0     | 0   | 13      | 0 | 28|       |
| dvir    | 8   | 3     | 0   | 5   | 0     | 1     | 3   | 0       | 7 | 28|       |
| dgri    | 2   | 8     | 0   | 21  | 0     | 1     | 3   | 1       | 6 | 42|       |
| **Single-module architectures formation** |
| dmel    | 0   | 0     | 0   | 27  | 0     | 0     | 0   | 8       | 3 | 38| 1180  |
| dere    | 7   | 0     | 0   | 27  | 0     | 0     | 0   | 10      | 2 | 46|       |
| dyak    | 16  | 13    | 0   | 146 | 0     | 0     | 0   | 15      | 11| 201|      |
| dana    | 12  | 4     | 0   | 67  | 0     | 0     | 0   | 13      | 9 | 105|      |
| dpse    | 14  | 4     | 0   | 101 | 0     | 0     | 0   | 22      | 22| 163|      |
| dwil    | 8   | 5     | 0   | 82  | 0     | 0     | 0   | 10      | 15| 120|      |
| dmoj    | 17  | 3     | 0   | 55  | 0     | 0     | 0   | 27      | 8 | 110|      |
| dvir    | 17  | 0     | 0   | 39  | 0     | 0     | 0   | 17      | 5 | 78 |      |
| dgri    | 7   | 9     | 0   | 258 | 0     | 0     | 0   | 19      | 26| 319|      |

Gene family expansions are identified when foldchanges larger than 1 along the corresponding branch. Based on whether the formed architecture is single-module or multi-modules, the expansions are classified into three categories, including single-module architectures formed only (1,180, 69.49%), multi-modules architectures formed only (281, 16.55%), and both of them formed (237, 13.96%). Module rearrangement types which did not occur among all three categories are not shown.