**Supplementary material**

**果蝇（D. melanogaster）必需基因筛选方法**

**研究开发的必需基因搜索（Essential Gene Search, EGS）通过以下参数在FlyBase中筛选必需基因（EGs）：**

**等位基因类型：需为"功能缺失等位基因（loss of function allele）"、"无效等位基因（amorphic allele）"（含遗传或分子证据）、"亚效等位基因（hypomorphic allele）"（含遗传或分子证据）；**

**表型类型：必须为"致死（lethal）"。  
无效等位基因指完全丧失功能的突变，而亚效等位基因的功能较野生型降低。若等位基因表达量降低足以导致致死，则该基因应归类为必需基因。我们排除了RNAi实验数据，因其无法提供等位基因类型信息。通过FlyBase网页界面筛选符合条件的等位基因后，将其转换为基因ID。**

**非必需基因搜索（NEGS）**

**非必需基因的筛选参数为：**

1. **等位基因类型：仅包含"功能缺失"或"无效等位基因"，排除所有亚效等位基因；**
2. **表型类型：必须为"非致死（not lethal）"。**
3. **排除亚效等位基因因其仍保留部分功能活性，无法严格满足非必需基因的定义（即功能完全缺失时个体仍存活）。未描述表型的等位基因也被排除（Supplementary Table 1）。最终结果同样转换为基因ID。**

**基因分类逻辑**

1. **必需基因：只要在杂合（heterozygosity）或纯合（homozygosity）状态下观察到致死表型即归类为必需基因（显性致死效应）；**
2. **非必需基因：仅当纯合无效等位基因个体存活时才归类为非必需基因，以避免隐性致死基因的干扰（例如果蝇基因*Duox*为隐性致死经典案例）。**

**若同一基因同时出现在EGS和NEGS结果中，则优先归类为必需基因（条件必需基因）。我们通过文献综述验证了自动筛选结果的可靠性，并补充了未涵盖的基因（图1A）。最终整合FlyBase与文献数据的集合称为DMEL数据集。**

**数据来源对比分析**

**现有数据库的局限性**

1. **OGEE数据库：**

**包含两个数据集：  
（1）13,781个基因的细胞水平RNAi筛选结果（可能漏检发育必需基因）；  
（2）437个转基因RNAi沉默基因（部分结果不可重复）。**

**问题：缺乏等位基因详细信息，且非必需基因占比高达98.1%，可能仅反映持家基因。**

1. **DEG数据库：**

**339个P元件插入致死的必需基因，但86.7%的等位基因缺乏类型信息，剩余多为亚效等位基因（90/181），不适合作为严格的功能缺失标准。**

1. **FlyBase优势：**

**通过等位基因类型、表型及基因型信息的结构化描述（如无效/亚效等位基因、致死/非致死表型），可明确区分必需与非必需基因。**

**我们的筛选获得1,393个必需基因与899个非必需基因，与OGEE数据集重叠率低（>1,000个FlyBase必需基因被OGEE误判为非必需）。**

**与Campos等（2020）ES评分的差异**

**Campos团队通过富集分数（Enrichment Score, ES）将基因分为三类（必需/条件必需/非必需），其方法存在以下差异：**

1. **定义差异：**

**我们要求基因功能完全或部分缺失导致致死表型（含条件必需基因）；**

**Campos的ES评分可能将低外显率必需基因误判为非必需基因。**

1. **数据过滤：我们排除了等位基因类型或表型不明确的基因（Supplementary Figure 1），而Campos方法保留了此类数据。**

**特征工程方法**

1. **吉布斯熵（Gibbs Entropy）：**

**使用VarGibbs软件（D-CMB模型）预测DNA序列的熵与焓值，生成两个特征。**

1. **伪氨基酸组成（Pseudo-amino acid composition）：**

**参数：λ=50（最大残基距离）、ω=0.05（序列顺序权重）；**

**生成20+λ维特征（前20维为氨基酸组成，后续为序列顺序效应）；**

**长度<50氨基酸的蛋白质被排除。**

1. **蛋白质长度：**

**必需基因通常更长（Grishkevich & Yanai, 2014），直接计算序列长度作为特征。**

1. **联合三联体（Conjoint Triad）：**

**按理化性质将20种氨基酸分为7类，统计连续三残基组合频率（共343维特征），用于预测蛋白质互作（Shen et al., 2007）。**

**Supplementary References**

Batista, G.E.A.P.A., Monard, M.C. An analysis of four missing data treatment methods for supervised learning. *Appl Artif Intell.* 2003;17:14.

Boutros, M.*, et al.* Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science (New York, N.Y* 2004;303(5659):832-835.

Chen, S., Zhang, Y.E. and Long, M. New genes in Drosophila quickly become essential. *Science (New York, N.Y* 2010;330(6011):1682-1685.

Chen, W.H.*, et al.* OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic acids research* 2017;45(D1):D940-D944.

Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 2001;43(3):246-255.

Gao, F.*, et al.* Gene essentiality analysis based on DEG 10, an updated database of essential genes. *Methods Mol Biol* 2015;1279:219-233.

Grishkevich, V. and Yanai, I. Gene length and expression level shape genomic novelties. *Genome Res* 2014;24(9):1497-1503.

Kondo, S.*, et al.* New genes often acquire male-specific functions but rarely become essential in Drosophila. *Genes Dev* 2017;31(18):1841-1846.

Larkin, A.*, et al.* FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic acids research* 2021;49(D1):D899-D907.

Schmitt-Engel, C.*, et al.* The iBeetle large-scale RNAi screen reveals gene functions for insect development and physiology. *Nat Commun* 2015;6:7822.

Shen, J.*, et al.* Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences of the United States of America* 2007;104(11):4337-4341.

Spradling, A.C.*, et al.* The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. *Genetics* 1999;153(1):135-177.

Weber, G. Optimization method for obtaining nearest-neighbour DNA entropies and enthalpies directly from melting temperatures. *Bioinformatics* 2015;31(6):871-877.

Xie, X.*, et al.* NIP/DuoxA is essential for Drosophila embryonic development and regulates oxidative stress response. *Int J Biol Sci* 2010;6(3):252-267.