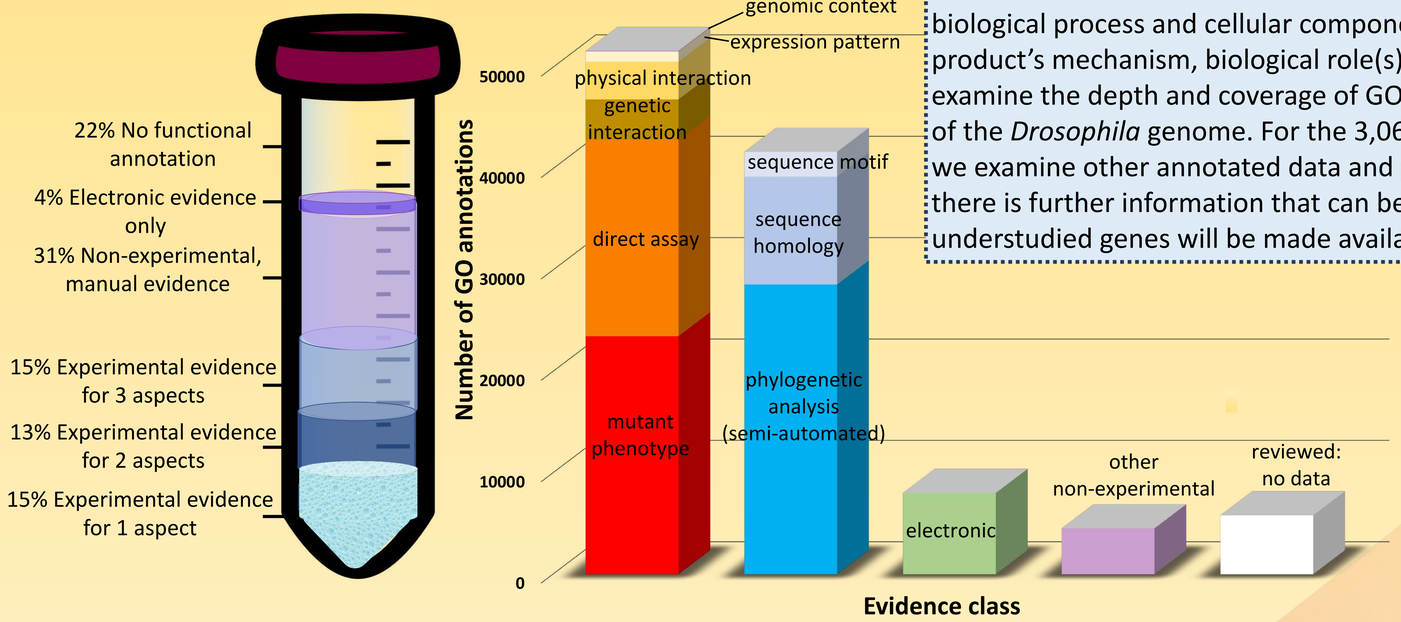
Progress towards functional understanding of the gene repertoire of *Drosophila*

Helen Attrill, Giulia Antonazzo, Phani Garapati, Nicholas H. Brown & The FlyBase Consortium

Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, CB2 3DY, UK. E-mail: hla28@cam.ac.uk

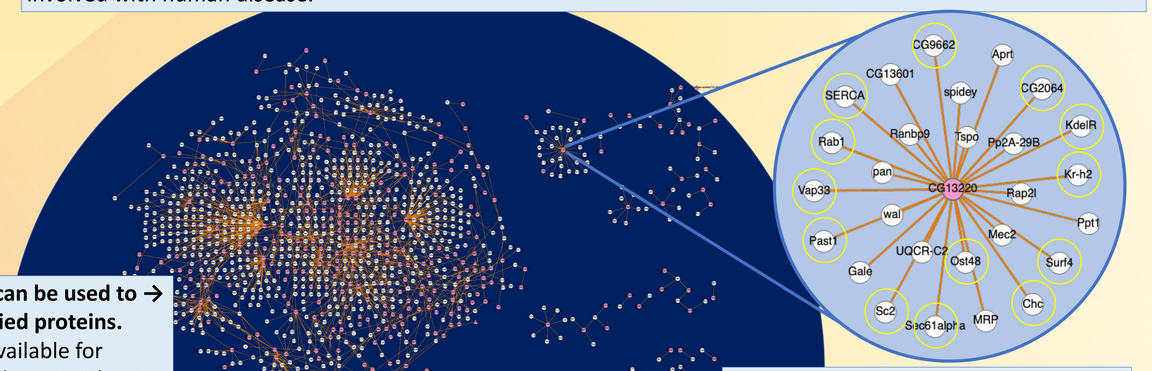
Poster #211A

Introduction The Gene Ontology (GO) is used to annotate gene products with functional information. The GO is divided into three 'aspects': molecular function, biological process and cellular component - capturing information about a gene product's mechanism, biological role(s) and where these activities occur. Here we examine the depth and coverage of GO annotation for the 13,969 protein-coding genes of the *Drosophila* genome. For the 3,064 protein-coding genes without GO annotation, we examine other annotated data and predicted sequence features, illustrating that there is further information that can be mined for functional inference. The list of understudied genes will be made available with the aim of stimulating future work.



D.mel	3,064 genes	
Other Drosophilids	2,754 orthologs	90%
Other Insects	1,107 orthologs	36%
Human	738 orthologs	24%
S.cerevisiae	238 orthologs	8%

7. 24% of fly proteins without GO annotation have putative orthologs in humans. Thus, future effort by the fly community to use the experimental advantages of *Drosophila* to elucidate the functions of these unknown genes will also aid our understanding of their function in humans. 13% of the human orthologs are predicted to be involved with human disease.

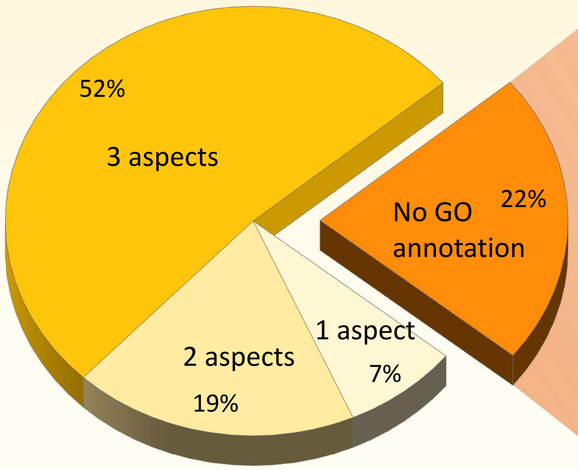


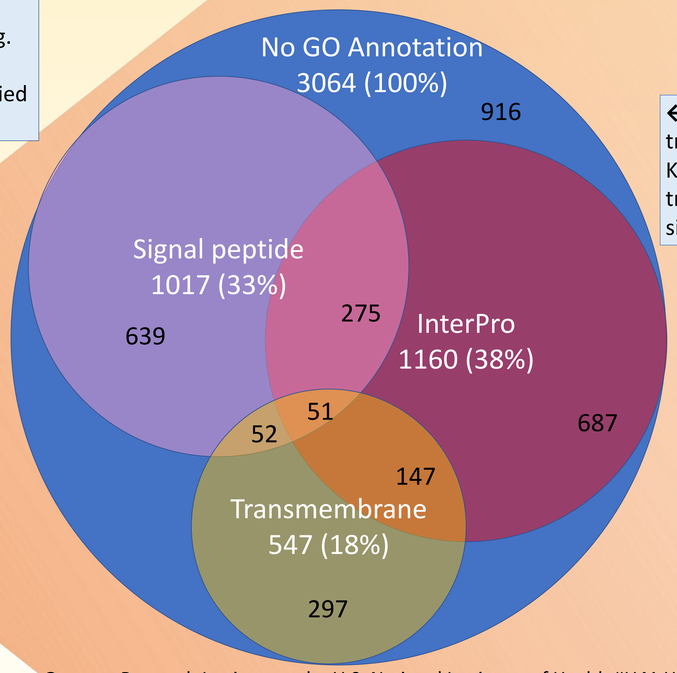
6. CG13220 is associated with very ↑ little curated data. Of its 27 curated interacting partners, 13 have GO cellular component annotations to the endomembrane system (circled in yellow), suggesting an involvement in intracellular trafficking pathways.

5. Physical interaction data can be used to → infer function for understudied proteins. Physical interaction data is available for 433/3064 of the proteins in the set without GO annotation. Of the set of unannotated proteins (pink nodes), 80% of proteins have a single curated interaction partner and 96% of proteins have <5 interacting partners.

↑ 1. GO annotation coverage of the protein-coding genome. ↑
The conical tube illustrates the depth of GO annotation evidence for the function of fly proteins – experimental, non-experimental or electronic (domain-based automated inference). 43% of protein-coding genes have some experimental information and 78% have at least one GO annotation. The bar chart shows the number of annotations supported by each evidence class. Overall, there are 105,867 annotations for protein-coding genes.

↓ 2. GO annotation coverage of the protein-coding genome. Over half of protein-coding genes have information in all three aspects. 22% have no GO annotation, this is similar to other estimates of uncharacterized proteins in other eukaryotes e.g. 20% in a cross-species study, Wood *et al*., 2019 (PMID:30938578); 27% of human proteins "never been studied by a full publication" Stoeger *et al.,* 2018 (PMID:30226837).





← 3. Domain/sequence analysis of protein-coding genes with no GO annotation .18% are predicted to possess transmembrane-spanning segments. This is very similar to the predicted complement of the protein-coding genome (20%, Korgh *et al*., 2001 (PMID:11152613). 33% are predicted by SignalP to possess a signal peptide, indicating that they are type I transmembrane proteins or enter the secretory pathway. 38% of unannotated proteins are associated with an InterPro signature.

Domain Name (top 25% of domain groups)	Number of genes	% of domains	Putative Function
Protein of unknown function DUF1091	86	7	unknown, postulated lipid binding
CHK kinase-like/Ecdysteroid kinase-like	42	4	kinase or other
Protein of unknown function DM4/12	28	2	unknown
Zinc finger, FLYWCH-type	19	2	Zn^{2+} binding, chromatin/DNA binding
Leucine-rich repeat domain superfamily	25	2	protein-protein interaction
Retinin-like protein	12	1	unknown
Protein of unknown function DUF1431 (Drosophila)	12	1	unknown
Protein of unknown function DUF4729	11	1	unknown
Protein TsetseEP	11	1	unknown
Transcription activator MBF2	10	1	transcription regulator
Protein of unknown function DUF745	10	1	unknown
MADF domain	10	1	transcription regulator
LPS-induced tumour necrosis factor alpha factor	9	1	Zn^{2+} binding, membrane-associated

← 4. Major domain groupings & putative functions. Many group into domains of unknown function. The largest grouping, DUF1091, is only found in insects. Based on similarity to MD-2-related lipid-recognition domain, it is hypothesized that this domain interacts with lipids.

Funding: British Medical Research Council #MR/N030117/1 & National Human Genome Research Institute at the U.S. National Institutes of Health #U41 HG000739