FlyBase Representative Publications:
Using annotation data to identify key papers on a gene

Giulia Antonazzo[1], David B. Emmert[2], Nicholas H. Brown[1] and the FlyBase Consortium

[1] Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3DY, UK

[2] The Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA
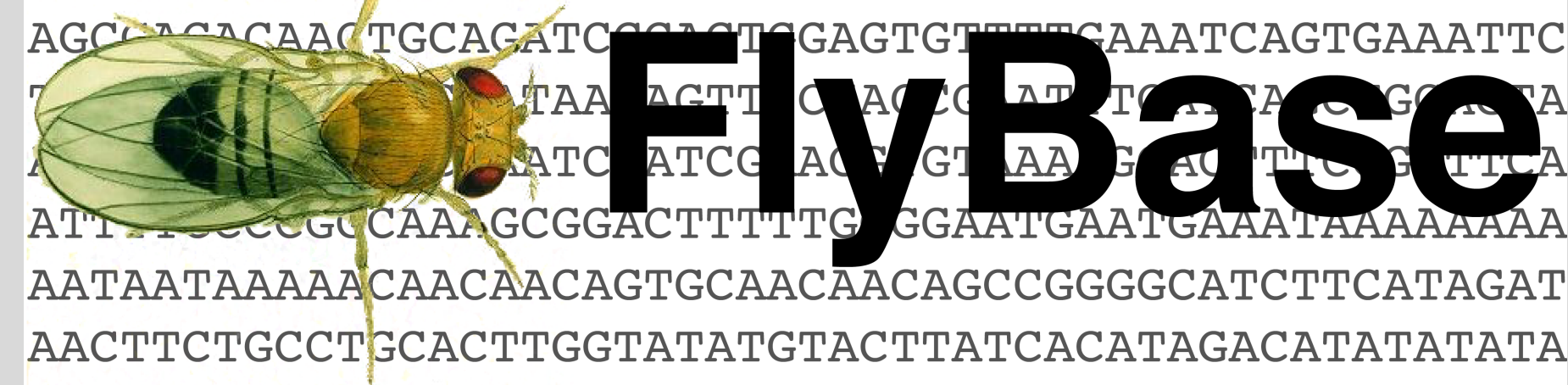
FlyBase

```
TACACAATCAGTTAGTTTCCACCGACAGTCCGCGAGAAACCATTCGACGGC
GTCGGCAATCCGTAAGATAGCCAAATATTATTATTGTTCAGATACTCACT
AGCCGCACAAGTTCCTGCCGTTCAGTGTTGAGTGTTTCAACATCAGTGAAATTC
AGCCGCACAAGTTCCTGCCGTTCAGTGTTGAGTGTTTCAACATCAGTGAAATTC
ATTCCCGCGTAAAGCGGACTTTTTCGGAATGAATGAAATAAAAAAA
AATAATAAAATCAACAACAGTGCAACAACAGCCGGGGCATCTTCATAGAT
AACTTCTGCCTGCACTTGGTATATGTACTTATCACATAGACATATATATA
```

Issue

Gene-to-publication associations are made within **FlyBase** regardless of the amount of data on a gene deriving from a given publication, resulting in comprehensive reference lists on each gene report page. However, it can be difficult for users to quickly **identify the key publications** that really characterize or focus on a gene in depth.
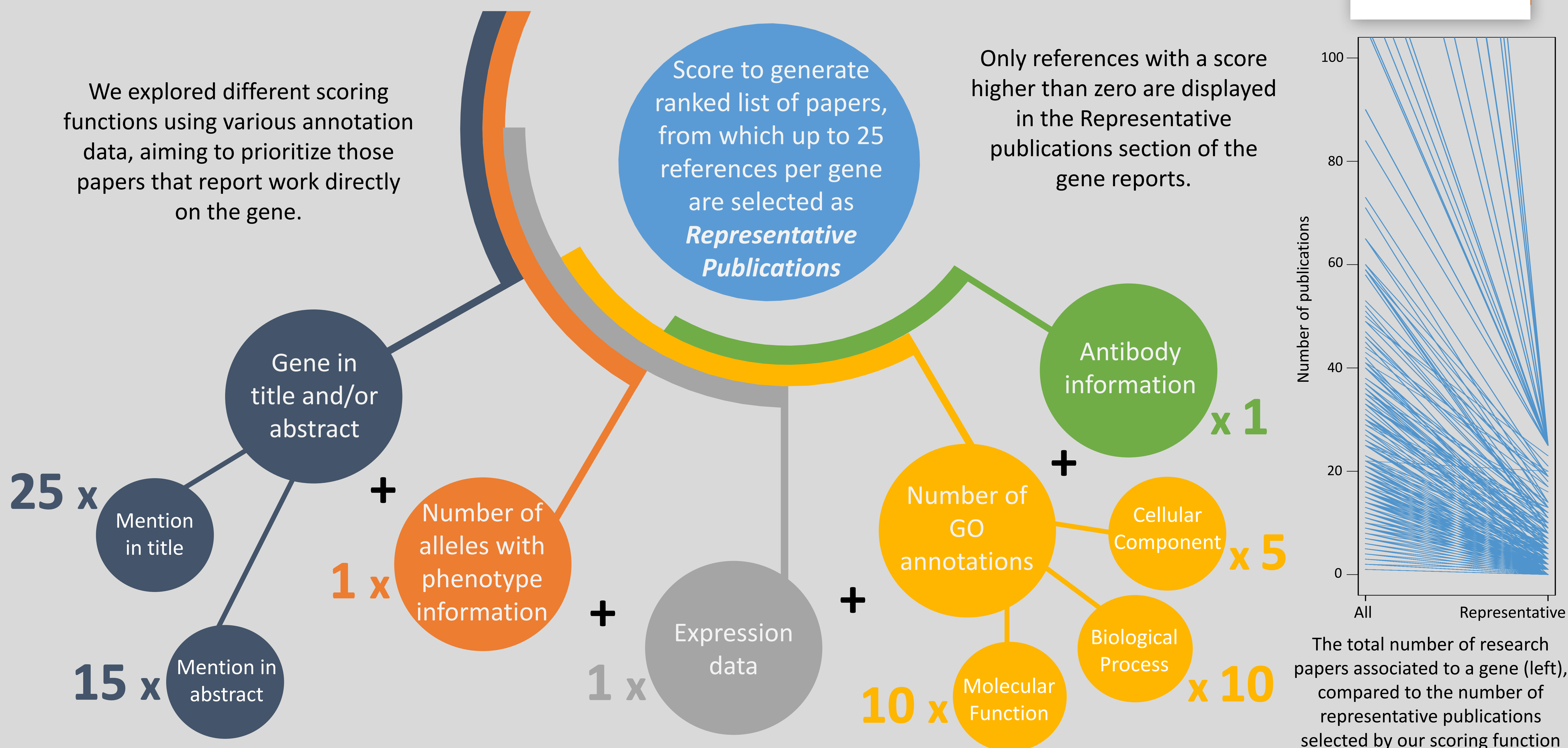
~14,000 Protein coding genes

~2,500 new research papers on *Drosophila melanogaster* per year

FlyBase solution

We have added a "*Representative Publications*" section to gene reports, which highlights **up to 25 of the most relevant research papers** for each gene. These are identified through an **in-house scoring algorithm** that assesses the amount and type of data associated to each protein coding gene, from each publication, within FlyBase.

We explored different scoring functions using various annotation data, aiming to prioritize those papers that report work directly on the gene.

Score to generate ranked list of papers, from which up to 25 references per gene are selected as *Representative Publications*

Only references with a score higher than zero are displayed in the Representative publications section of the gene reports.

Gene in title and/or abstract

25 x Mention in title

15 x Mention in abstract

+

1 x Number of alleles with phenotype information

+

1 x Expression data

+

10 x Number of GO annotations
- Cellular Component x 5
- Biological Process x 10
- Molecular Function x 10

+

Antibody information x 1

The total number of research papers associated to a gene (left), compared to the number of representative publications selected by our scoring function for the same gene (right), displayed for 500 random genes.

Contact: *ga362@cam.ac.uk*