# The impact of focused Gene Ontology annotation efforts on high-throughput data analysis

*Ruth C Lovering[1], Varsha K Khodiyar[1], Rachael P Huntley[2], Yasmin Alam-Faruque[2], Emily C Dimmer[2], Tony Sawford[2], Claire O'Donovan[2], Peter Scambler[3], Mike Hubank[4], Rolf Apweiler[2], Philippa J Talmud[1].*

[1]Centre for Cardiovascular Genetics, Institute of Cardiovascular Science, University College London, Rayne Building, 5 University Street, London WC1E 6JF.
[2]UniProt Gene Ontology Annotation Project, European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD.
[3]Molecular Medicine Unit, Institute of Child Health, 30 Guilford Street, London WC1N 1EH.
[4]Molecular Hematology and Cancer Biology Unit, Institute of Child Health, 30 Guilford Street, London WC1N 1EH.
Correspondence: r.lovering@ucl.ac.uk

## Introduction

Gene Ontology (GO) was one of the first biological ontologies to be created[1,2] and is a key resource for researchers wishing to understand the biological role of a gene product. Whilst GO is widely used by the high-throughput scientific community it is also used by many other scientists as a way of quickly identifying the function of any protein, the processes it is involved in and its location within the cell.

Over the last 4 years, the Cardiovascular Gene Ontology Annotation Initiative, funded by the British Heart Foundation (BHF) and based at UCL, has supplied GO annotation specifically to human proteins involved in cardiovascular processes (see Figure 1). The GO uses structured controlled vocabulary terms, to describe three aspects of a gene product's attributes: the *molecular function(s)*, or activities that a gene product can directly perform; the *biological process(es)* it contributes to; and the subcellular locations (*cellular component*) in which it is present[1]. Around 37,000 GO terms describe a wide range of concepts to differing levels of specificity and are organised as directed acyclic graphs using descriptive relationship type.

## Results

This is the first annotation effort to focus on a specific area of biology[3] and the first time that GO annotators have been placed in a laboratory research environment (rather than in a bioinformatics environment). Working at UCL has enabled the BHF-funded GO curators to establish collaborations with local cardiovascular researchers and build up their own expertise in this area. This has fed back into the further development of the Gene Ontology itself, with the UCL curation team being responsible for the creation of ~1,500 GO terms. Some of these terms have been created as the result of concerted ontology development efforts, such as terms to describe heart development[4], others have been created on an ad-hoc basis as needed. In addition, this initiative has increased the number of GO annotations associated with human proteins, providing ~10% of the 170,000 manual GO annotations currently available.

The interpretation of high-throughput datasets is often limited by the quality and quantity of the annotations available. We have demonstrated that the ability to identify discriminatory groupings within a cardiovascular high-throughput dataset can be vastly improved by combining the creation of more specific GO terms with the use of these terms to provide more descriptive gene annotations[5]. A microarray dataset was chosen for re-analysis that had identified differentially regulated genes in peripheral blood mononuclear cells from patients with systemic scleroderma-related pulmonary arterial hypertension (PAH-SSc) compared to healthy controls[6]. The re-analysis of this dataset using more recent GO annotation data identified the significant enrichment of GO terms relevant to the disease phenotype, which were not originally reported, such as '*cytokine-mediated signaling pathway*' and '*positive regulation of nitric oxide biosynthetic process'* (Table 1). Removing the BHF-funded annotations from the analysis decreased the significance of the majority of enriched GO terms and several GO terms were no longer identified (Table 1)[5].

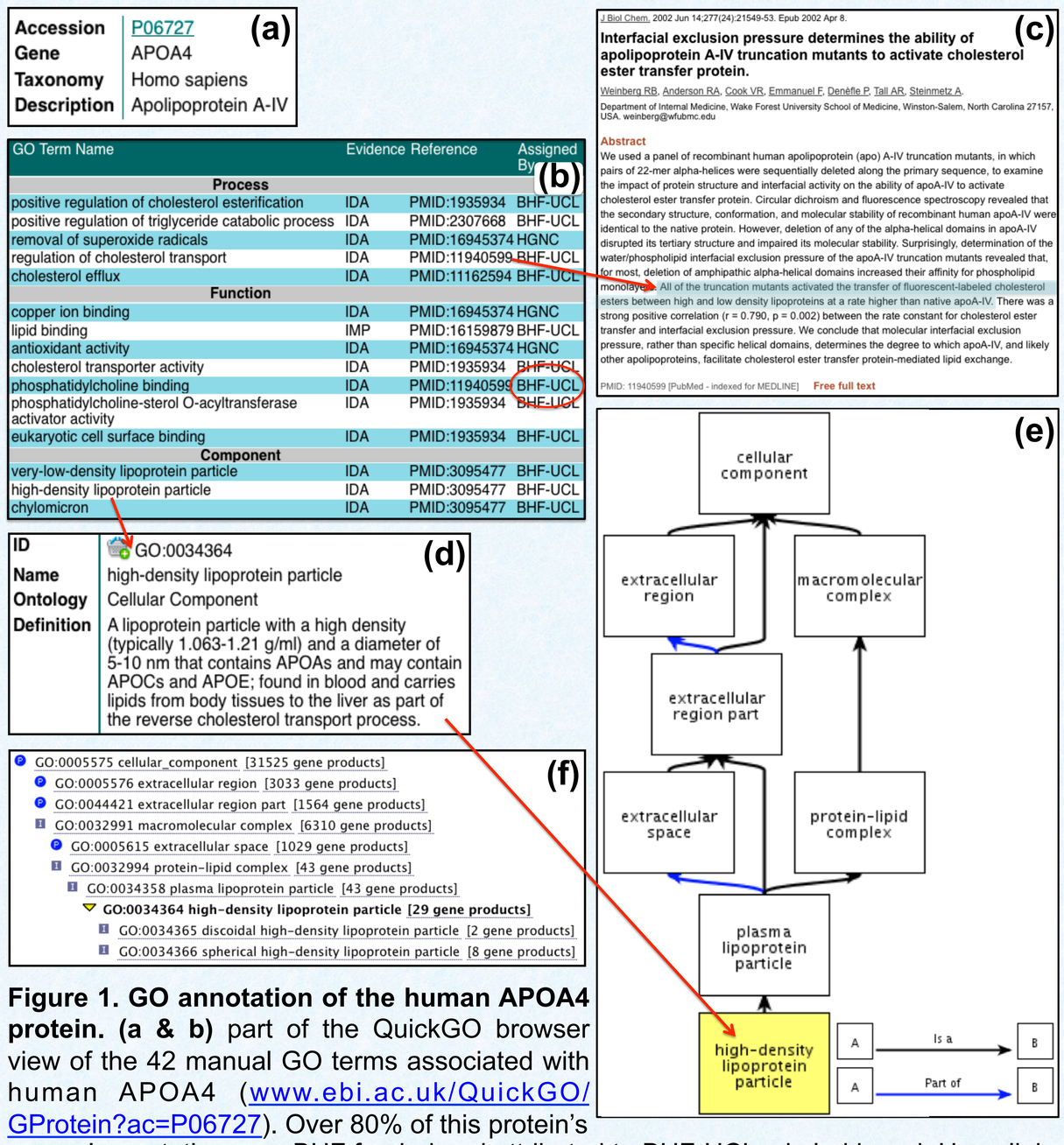


**Figure 1. GO annotation of the human APOA4 protein. (a & b)** part of the QuickGO browser view of the 42 manual GO terms associated with human APOA4 (www.ebi.ac.uk/QuickGO/GProtein?ac=P06727). Over 80% of this protein's manual annotations are BHF funded and attributed to BHF-UCL, circled in red. Hyperlinks (red arrows) to **(c)** the listed abstract and **(d & e)** the QuickGO term record (www.ebi.ac.uk/QuickGO/GTerm?id=GO:0034364#term=info). **(f)** AmiGO browser view of the ontology structure of the GO term 'high-density lipoprotein particle' including the child terms associated with this GO term (http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0034364).

## Conclusion

GO term enrichment analysis of this PAH-SS dataset confirms that GO annotations created through three years of annotation focused on cardiovascular-relevant proteins, rather than specific annotation of just a few genes within a study dataset, can lead to significantly improved data interpretation. This demonstrates the need for comprehensive, information-rich annotation datasets and a more knowledgeable use of existing public data to aid in pathway identification and to fully harness bioresources and biomodelling. Hence the continued improvements in both protein GO annotation and ontology development can enable researchers to gain improved biological insights into their proteins of interest and hence guide their future research towards alleviating various human diseases.

## Future Directions

Many biological domains remain under-represented in GO. To address this concern we are looking for funding opportunities to enable us to focus our annotation on specific areas and help improve the interpretation of specific high-throughput datasets.

**Table 1. Comparison of hypertension microarray data analysis using GO annotation dataset with and without the human BHF-funded annotations.**

| GO term | GO dataset including BHF annotations p-value | GO dataset without BHF annotations p-value |
|---|---|---|
| response to lipopolysaccharide | 5.45E-05 | 1.22E-03 |
| inflammatory response | 2.03E-04 | 1.59E-02 |
| positive regulation of anti-apoptosis | 5.28E-04 | 2.06E-03 |
| positive regulation of nitric oxide biosynthetic process | 1.75E-03 | #N/A |
| immune response | 1.82E-03 | 1.30E-03 |
| positive regulation of smooth muscle cell proliferation | 2.22E-03 | 6.23E-03 |
| response to organic cyclic substance | 3.02E-03 | 2.82E-03 |
| leukocyte migration | 3.60E-03 | 1.94E-02 |
| response to corticosterone stimulus | 6.69E-03 | 6.23E-03 |
| cytokine-mediated signaling pathway | 9.66E-03 | 1.32E-01 |

GO processes with p-values<0.01 identified are considered as significantly enriched. Shading indicates p-values that are not significant.

## References

1. Gene Ontology Consortium (2001) Creating the gene ontology resource: Design and implementation. Genome Res 11(8): 1425-1433.
2. Gene Ontology Consortium 2012. The Gene Ontology: enhancements for 2011. *Nucleic acids research,* 40, D559-64.
3. Lovering, R. C., Dimmer, E., Khodiyar, V. K., Barrell, D. G., Scambler, P., Hubank, M., Apweiler, R. & Talmud, P. J. 2008. Cardiovascular GO annotation initiative year 1 report: why cardiovascular GO? *Proteomics,* 8, 1950-3.
4. Khodiyar, V. K., Hill, D. P., Howe, D., Berardini, T. Z., Tweedie, S., Talmud, P. J., Breckenridge, R., Bhattarcharya, S., Riley, P., Scambler, P. & Lovering, R. C. 2011. The representation of heart development in the gene ontology. *Developmental biology,* 354, 9-17.
5. Alam-Faruque, Y., Huntley, R. P., Khodiyar, V. K., Camon, E. B., Dimmer, E. C., Sawford, T., Martin, M. J., O'Donovan, C., Talmud, P. J., Scambler, P., Apweiler, R. & Lovering, R. C. 2011. The impact of focused Gene Ontology curation of specific mammalian systems. *Plos one,* 6, e27541.
6. Grigoryev DN, Mathai SC, Fisher MR, Girgis RE, Zaiman AL, et al. (2008) Identification of candidate genes in scleroderma-related pulmonary arterial hypertension. Transl Res 151(4): 197-207.

Grant: SP/07/007/23671

goannotation@ucl.ac.uk
www.ucl.ac.uk/cardiovasculargeneontology