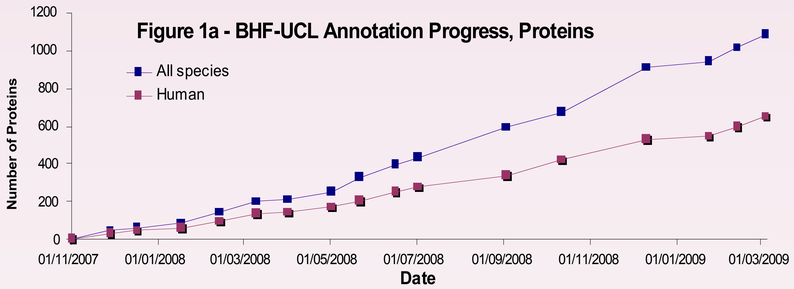
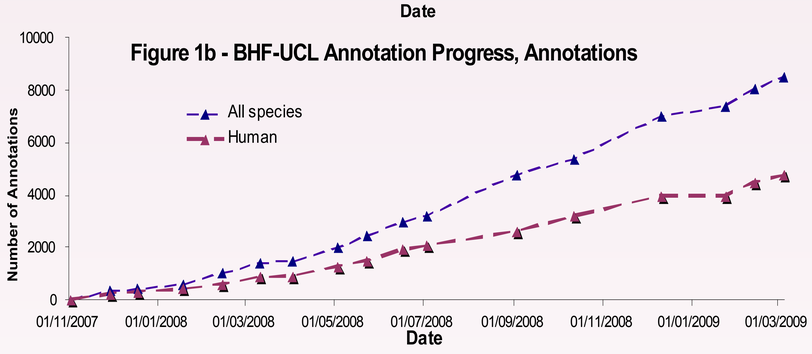
Cardiovascular Gene Ontology Annotation Initiative

Varsha Khodiyar[1], Daniel Barrell[2], Peter Scambler[3], Mike Hubank[4], Rolf Apweiler[2], Philippa Talmud[1] and Ruth Lovering[1]

[1]Centre for Cardiovascular Genetics, UCL Department of Medicine, Rayne Institute 5 University Street London WC1E 6JF.
[2]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.

Gene Ontology (GO) provides a controlled vocabulary, which is used by several groups around the world to provide functional annotation to proteins across a wide range of species (www.geneontology.org). The Cardiovascular Gene Ontology Annotation Initiative is funded the British Heath Foundation to supply GO annotation specifically for human proteins involved in cardiovascular (CV) processes. This is the first time that a physiological process-centred approach has been used for human protein GO annotation[1]. Experienced GO curators from the BHF-UCL team work alongside the bench scientists from the CV genetics group at University College London (UCL). By working in a CV dedicated environment the GO curators are developing their expertise in this field; this is leading to more detailed and accurate GO annotation of CV-relevant genes. Since the start of the project in November 2007 we have added over 4700 annotations to 649 human CV related proteins, as shown in figure 1.





Identification of CV related genes and proteins

A list of over 4000 CV-related genes was assembled by merging the following gene lists:

- the ITMAT consortium list of approximately 2,200 genes, generated for the vascular disease 50K SNP array[2].
- a list of 282 congenital heart disease candidate genes, created by Bentham and Bhattacharya following the identification of genes with a major role in mouse heart development[3].
- a list of approximately 2,500 genes identified by the BHF-UCL team as associated with CV-relevant GO terms at the start of the project.
- These gene lists were merged and supplemented by 170 more genes on the advice of our advisory board composed of expert CV researchers.

With further input from our advisory board we prioritised 250 genes from the merged list of 4054 genes, for annotation in the first year. We followed a gene by gene annotation approach for the first 12 months of the project. However whilst reading the literature for a single gene, we often annotated several other genes within the same pathway as they would be discussed within the same papers. Thus for the second year we decided to trial a process centric approach.

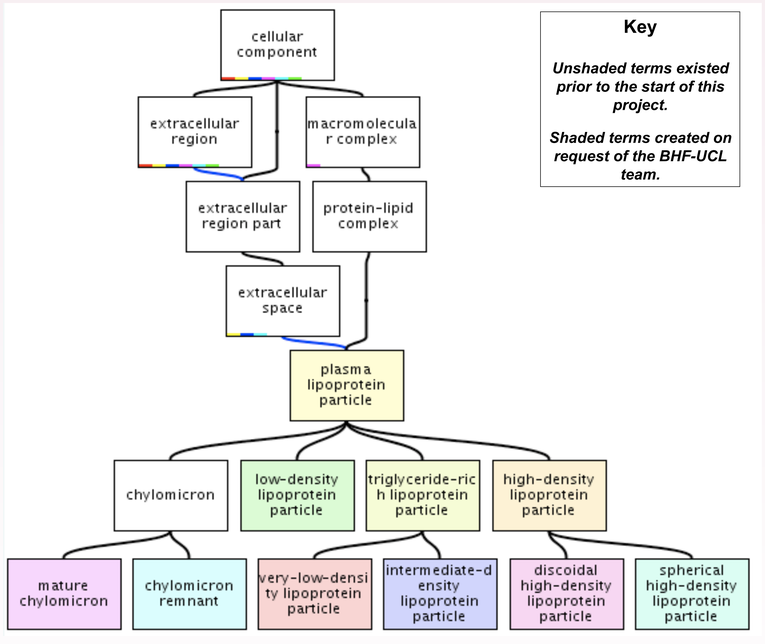
Process-centric annotation method

We have identified several pathways and processes which are relevant to the cardiovascular system. For example: TGFβ SMAD signalling, growth hormone release, foam cell differentiation and cholesterol esterification were all annotated recently.

The annotation team aims to completely annotate two pathways each month. Complete annotation means that all genes within the pathway are annotated. Some genes may also participate in other pathways and GO terms relevant to these pathways will be added when the secondary pathways are annotated. The benefit of annotating in this process-centric manner is a deeper understanding of the CV-related pathways by the BHF-UCL team. This increased understanding generally results in: i) increased granularity of GO annotation, ii) greater consistency of GO annotations across each process, and iii) BHF-UCL annotators being able to give a deeper level of input to ontology development.

For example during the annotation of the process lipoprotein particle clearance. the CV genetics group at UCL provided considerable advice on appropriate GO terms for each of the human apolipoproteins, identified relevant publications for the BHF- UCL team to annotate and contributed to the development of the lipoprotein particle cellular component ontology as shown in figure 2.

Figure 2 – How the GO was developed to reflect published knowledge on lipoprotein particle biology.



Ontology Development

The GO editorial office develops and refines the GO in small, medium and large scale projects[4]. Often GO annotators will need to request new terms in order to extract the most information from the papers they annotate and in general this results in small scale changes to the ontology. The close association of the BHF-UCL annotation team in an active bench science CV research laboratory and the intensive reading required as part of the annotation process, means that the BHF-UCL team has gained considerable expertise in various aspects of CV system, processes and development. Therefore the BHF-UCL team is able to successfully instigate both small and medium scale GO development projects in CV related areas with the GO editorial office.

Figure 2 shows one example of an ontology development project that was undertaken by the BHF-UCL team. New terms were developed in consultation with lipoprotein researchers based at the UCL CV genetics laboratory. GO terms for a number of other CV related processes have been, or are in the process of being, developed by the BHF-UCL team in collaboration with expert bench researchers including; heart conduction, heart development and cholesterol import.

Improving electronic annotations

All GO curators review the electronic annotations associated with a gene record during the annotation of a particular gene and notify the appropriate source if any errors are identified.. This provides an additional opportunity to ensure the GO annotations are as accurate as possible. For example, prior to this work, SwissProt lipoprotein particle keywords such as HDL, LDL and VLDL had been associated with several apolipoproteins by SwissProt curators. As these component terms did not exist in GO these SwissProt keywords had instead been erroneously mapped to the GO term “lipid transporter activity”, using the Keywords2GO table4. Consequently, the molecular function GO term “lipid transporter activity” had been erroneously electronically propagated to all apolipoproteins.

With ready access to bench researchers working in the lipoprotein field, the BHF-UCL team was able to confirm that the majority of apoplipoproteins are not lipid transporters and requested that this erroneous electronic mapping was replaced with an accurate mapping to the newly created lipoprotein particle component terms. The Keywords2GO table now maps the SwissProt keyword HDL to the GO term high-density lipoprotein particle.

Conclusions

- A gene-by-gene annotation process promotes the complete annotation of a single gene.
- A process-centric approach allows the annotator to gain a deeper understanding of a specific pathway, and promotes the annotation of all the genes involved within that pathway.
- A major benefit of undertaking GO annotation in a process-centric approach is the development and refinement of the ontology, which occurs alongside the annotation process.

Contact us: Email: GOannotation@UCL.ac.uk
Wiki: http://wiki.geneontology.org/index.php/Cardiovascular
Website: http://www.cardiovasculargeneontology/feedback

Grant: SP/07/007/23671

www.cardiovasculargeneontology.com

References

1. Lovering RC, Dimmer E, Khodiyar VK, Barrell DG, Scambler P, Hubank M, Apweiler R, Talud PJ. Cardiovascular GO annotation initiative year 1 report: why cardiovascular GO? Proteomics 2008, 8:1950-1953
2. Vascular Disease 50k SNP Array Consortia http://bmic.upenn.edu/cvdsnp/index.php
3. Bentham J, Bhattacharya S. Genetic mechanisms controlling cardiovascular development. Ann N Y Acad Sci. 2008, 1123:10-19
4. Harris M.A. Developing an ontology. Methods Mol Biol. 2008, 452: 111-124
5. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R. The GOA database in 2009--an integrated Gene Ontology Annotation resource. Nucleic Acids Res. 2009, 37: D396-D403