Invited Talk:

Gene annotation in transcriptomic analysis

Andy Cuming

Centre for Plant Sciences, Leeds University.

The availability of a sequenced genome enables global gene expression analysis at a number of levels through 'omics approaches. However, interrogating the entire genome of an organism also presents challenges, if the appropriate conclusions are to be drawn from the analysis of experimental data.

In particular, newly sequenced genomes are often poorly annotated in both structural and functional terms. Whilst available computational tools enable the development of automated pipelines for identifying potential coding sequences and regions of shared homology between genes in different species, the gene models and annotations that are generated are frequently inaccurate and misleading.

The *Physcomitrella* genome is no different from the other, rapidly emerging genome sequences in this respect. Experience of analysing transcriptional responses to a highlights some basic rules of gene annotation:-

1. The "Filtered gene model" (the one with the GenBank accession number) is most likely wrong.
2. Identify and prioritise a specific subset of genes for detailed analysis – with 30,000-odd genes, you can't curate them all!
3. Correct structural annotation is often a prerequisite for correct functional annotation, so make use of EST data and BLASTX alignments
4. Use your common sense (does that gene ***really*** have a 20kilobase intron…or has the software misinterpreted a gene duplication?)
5. Use BLASTP, the "conserved domains" function, TargetP and online resources for other genomes
6. The gene showing the most striking response to your experimental treatment will only be annotatable with the legend "Hypothetical expressed gene of unknown function"

These principles will be illustrated with examples from analyses of both microarray and RNA-seq transcriptomic experiments.