Data Set Descriptions

Big Data Biology (BIO00047I) | January 2019
Workshop material:
http://www-users.york.ac.uk/~dj757/BIO00047I/BIO00047I_index.html

Fission yeast data

This is gene-centric data from the fission yeast *Schizosaccharomyces pombe*. The data has been gathered from multiple different high-throughput studies, including some recent unpublished data (*1*). Much of it was downloaded from the Angeli website (http://bahlerweb.cs.ucl.ac.uk/cgi-bin/GLA/GLA_input).

Each row contains information about one gene. Both protein-coding genes and non-coding RNAs (ncRNAs) are included. Each column contains some categorical or quantitative information about each gene. The data columns are described in the table on the next page.

Oilseed rape data

This is gene-centric data from plant *Brassica napus* that produces canola oil.
Each row contains information about one gene.

The data in OSR101_sample.txt shows measurements of gene expression in a random sample of 101 oilseed rape cultivars. The gene expression measure used here is RPKM, or reads per kilobase per million aligned reads. The data in the file Glucosinolates.txt contains the quantitative measure of glucosinolates in the seed oil from each cultivar.

These data can be used to conduct an association study (comparing gene expression to glucosinolate levels). This is called associative transcriptomics (for example, see (*2*)).

Fission yeast data table

Column name	Data	Type* (Q/C)	Notes	Reference
NumberIntrons	Number of introns in the gene	Q		(*3*)
NumResidues	Number of amino acid residues in the protein.	Q	Will be NA for ncRNAs.	(*3*)
protein_coding	Is the gene protein-coding?	C	1 = protein coding 0 = ncRNA	(*3*)
ncRNA	Is the gene a non-canonical ncRNA?	C	1 = ncRNA 0 = either protein-coding or a canonical ncRNA (rRNA, tRNA, snoRNA)	(*3*)
Rel_telomere	The relative distance to the telomere.	Q	0 = at the telomere, 1 = in the middle of the chromosome	(*3*)
mRNA_copies_per_cell	The number of mRNA copies per cell, from proliferating (growing) cells.	Q		(*4*)
protein_copies_per_cell	The number of protein copies per cell, from proliferating (growing) cells.	Q		(*4*)
mRNA.stabilities	The half-life of the mRNA in minutes.	Q		(*5*)
GeneticDiversity	The level of genetic diversity within *S. pombe* strains.	Q	The average pairwise similarity (π)	(*6*)
ProteinHalfLife	The half-life of the protein in minutes.	Q		(*7*)
Golgi, Mitochondrion, Nuclear_dots, Nuclear_envelope, Nucleolus, Nucleus, Vacuole	Where the protein is located.	C	1 = in this location 0 = not in this location	(*8*)
essential	Is this an 'essential' gene (required for cell survival)?	C		(*9*)
chromosome	Which chromosome the gene is on.	C	*S. pombe* has three chromosomes (I, II, III) and a mitochondria (MT).	(*3*)
start, end	The position of the gene in the chromosome	Q	Gene length = end - start +1	(*3*)
solid.media.KO.fitness	The colony size of a strain with this gene knocked out (on agar medium).	Q	The colony size is a proxy for knockout 'fitness'.	(*10*)
gene.expression.RPKM	The RNA expression level from RNA-seq, from proliferating (growing) cells.	Q		(*11*)
conservation.phyloP	The level of conservation in this gene.	Q	How fast the gene has changed over time.	(*1*)

* Quantitative or Categorical.

References

1. L. Grech *et al.*, Fitness Landscape of the Fission Yeast Genome. *bioRxiv*, 398024 (2018).
2. A. L. Harper *et al.*, Molecular markers for tolerance of European ash (Fraxinus excelsior) to dieback disease identified using Associative Transcriptomics. *Sci Rep*. **6**, 19335 (2016).
3. V. Wood *et al.*, The genome sequence of Schizosaccharomyces pombe. *Nature*. **415**, 871–880 (2002).
4. S. Marguerat *et al.*, Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell*. **151**, 671–683 (2012).
5. A. Hasan, C. Cotobal, C. D. S. Duncan, J. Mata, Systematic Analysis of the Role of RNA-Binding Proteins in the Regulation of RNA Stability. *PLoS Genet*. **10**, e1004684

(2014).
6. D. C. Jeffares *et al.*, The genomic and phenotypic diversity of Schizosaccharomyces pombe. *Nature Genetics*. **47**, 235–241 (2015).
7. R. Christiano, N. Nagaraj, F. Fröhlich, T. C. Walther, Global proteome turnover analyses of the Yeasts S. cerevisiae and S. pombe. *Cell Rep*. **9**, 1959–1965 (2014).
8. A. Matsuyama *et al.*, ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. *Nat Biotechnol*. **24**, 841–847 (2006).
9. D.-U. Kim *et al.*, Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nat Biotechnol*. **28**, 617–623 (2010).
10. M. Malecki *et al.*, Functional and regulatory profiling of energy metabolism in fission yeast. *Genome Biol*. **17**, 240 (2016).
11. S. R. Atkinson *et al.*, Long noncoding RNA repertoire and targeting by nuclear exosome, cytoplasmic exonuclease, and RNAi in fission yeast. *RNA*. **24**, 1195–1213 (2018).