

Letter to the Editor

Data storage: bringing us a step closer to data sharing?

The work by Saito *et al.* (2005) presented in a recent issue of this journal is laudable and timely for two reasons. First of all, it raises awareness about the importance and the challenges of data storage and management, an issue the field of nutrition cannot ignore. Second, the article reminds us that, beyond data storage, data sharing is fundamental to establishing such resources. As pointed out by the authors, out of 250 publications currently stored in their tool, only a handful are actually available for comprehensive review (i.e. include raw data). Surveying PubMed with keywords such as 'adipose tissue', 'microarray', 'obesity' and 'diabetes type 2' identifies over thirty array-based publications (Cao *et al.* 2001; Atzmon *et al.* 2002; Barta *et al.* 2002; Gregoire *et al.* 2002; Napoli *et al.* 2002; Roche *et al.* 2002; Roy *et al.* 2002; Sreekumar *et al.* 2002; Castro-Chavez *et al.* 2003; Fujiwara *et al.* 2003; Gabrielsson *et al.* 2003; Lan *et al.* 2003; Lopez *et al.* 2003; Moraes *et al.* 2003; Sartipy & Loskutoff, 2003; Tanaka *et al.* 2003; Almind & Kahn, 2004; Becker *et al.* 2004; Crott *et al.* 2004; Deng *et al.* 2004; Dhahbi *et al.* 2004; Mutch *et al.* 2004; Recinos *et al.* 2004; Vohl *et al.* 2004; von Eyben *et al.* 2004; Eletto *et al.* 2005; Guan *et al.* 2005; Oana *et al.* 2005; Rota *et al.* 2005; Tseng *et al.* 2005; Tsuda *et al.* 2005; van Breda *et al.* 2005; van Schothorst *et al.* 2005; Xiao *et al.* 2005; Yagil *et al.* 2005), accounting for roughly 600 hybridisations worth of data. Assuming 8500 genes per array and twenty hybridisations on average per experiment, this corresponds to over 5 million data points. With little doubt, these datasets could be used as seeding material for establishing a nutrient–gene interaction knowledge base. Sadly for the nutrition community, only one out of these thirty datasets has been deposited in a public repository (GSE1392) (Mutch *et al.* 2004) and, for that one, the raw data are not available.

So is the nutrition community really ready for data sharing? Is it that effective data sharing is just too complex to be feasible yet? Well, one can arguably say no to the latter, considering the vast amount of effort spent in the field of microarrays to enhance data exchange and access. It is almost 4 years since the publication of the 'Minimum Information About a Microarray Experiment' (MIAME) paper by the Microarray Gene Expression Society (MGED) (Brazma *et al.* 2001). The same group of individuals has come up with an object model for database implementation (MAGE-OM) that would enable data persistence and an XML format (MAGE-ML) to enhance data exchange between institutions. Last, MGED has made an attempt to provide the community with a common set of descriptors (controlled vocabularies) arranged in an ontology referred to as 'MO', standing for MGED Ontology. Why generate such an ontology? If MIAME defines the amount of information to describe, it does not formulate any recommendation about which terminology should be used to provide annotation. Hence, one can be MIAME

compliant with simple free text. This is where the sting is. Making sense of free text is computationally expensive and currently available text-mining techniques clearly under-perform. A way around this hurdle is to promote the use of community-vetted annotation standards, such as controlled vocabularies and ontologies. Their use eventually ensures that terms and descriptors are employed consistently throughout a community, which is made resource aware.

This is one drawback of the work by Saito *et al.* (2005), as the resource they have set up makes little use of controlled vocabularies. Still, their work should be viewed as an essential reminder that more attention should be paid to consistent data annotation and at establishing curated resources.

But why make so much fuss about annotation and does missing annotation really matter? Again, let us take a simple example. How does the presence or absence of information impact the understanding of a study? What if one does not report the sex of the animal or the strain used or the developmental stage or the target organ? Can the information be reconstructed simply by reading the articles? The latter task is certainly manageable when dealing with, say, four or ten datasets, but what about 1000 datasets?

This is where journals and funding agencies can play a central role by insisting on having raw data and metadata (i.e. all the necessary descriptive ancillary data enabling use of raw data files) deposited in public databases (for example, Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) (Barrett *et al.* 2005) or ArrayExpress at the European Bioinformatics Institute (EBI) (Parkinson *et al.* 2005)). Furthermore, making sure that data can be peer-reviewed for publications or grant appraisal is a critical step of scientific assessment (Ball *et al.* 2002). Therefore, data sharing is vital for ensuring that knowledge is not lost and work is not unnecessarily duplicated.

Last, depositing data is a long-term investment: facilitated data access allows bioinformaticists to develop new algorithms and tools while enabling students and trainees to get acquainted with complex datasets combining data from omics technologies with classical phenotypic anchoring.

Nor should we stop at microarray data. Mass spectrometers and NMR instruments are becoming more broadly available for protein characterisation and metabolic studies. These techniques are adding new dimensions to the space of metrics nutritionists can use to explore biological systems.

Structured management of these data is simply becoming a necessity. Reassuringly, both the proteomics and metabolomics communities have organised themselves and are striving to develop data exchange standards in their own fields, namely the Human Proteome Organisation Proteomics Standards Institute (HUPO-PSI) and the Standard Metabolic Reporting Structure (SMRS) group (http://smrsgroup.sourceforge.net/mm_report.html) respectively (Taylor *et al.* 2003; Lindon *et al.*

2005). HUPO-PSI has delivered the mzData format for describing MS data, which has been accepted by major hardware companies and a data repository, 'PRoteomics IDentifications database' (PRIDE), relying on the format that has been set up at EBI (<http://www.ebi.ac.uk/pride/>) (Martens *et al.* 2005). The good thing is that these efforts are not being undertaken in isolation. Rather, synergies and integration capabilities are highly encouraged. To this end, all these groups are working collaboratively alongside the Reporting Structure for Biological Investigation Working Group (RSBI-WG), an offshoot from the MGED society, which brings together representatives from Nutritional Genomics, Toxicogenomics and Environmental Genomics communities (<http://www.mged.org/Workgroups/rsbi/rsbi.html>). The RSBI working group has a key liaising role and the group reports about its activities to the Functional Genomics (FuGE) and Functional Genomics Ontology (FuGO) groups in order to bring forward use cases and specific need from the nutrition research arena, a body working to provide tools, exchange formats and annotation standard for describing the complexity of functional genomics experiments (<http://fuge.sourceforge.net> and <http://fuge.sf.net/fugo>).

All this will be ultimately beneficial to the field of nutritional science but on one condition, that of data sharing.

The authors are members of The European Nutrigenomics Organisation (NuGO). The European Nutrigenomics Organisation, linking genomics, nutrition and health research (NuGO, CT-2004-505944), is a Network of Excellence funded by the European Commission's Research Directorate General under Priority Thematic Area 5 Food Quality and Safety Priority of the Sixth Framework Programme for Research and Technological Development.

Philippe Rocca-Serra¹ and Ruan M. Elliott²

¹EMBL-EBI

Wellcome Trust Genome Campus

Hinxton

Cambridge CB10 1SD

UK

²Institute of Food Research

Norwich NR4 7UA

UK

Fax: 44 1603 507723

Email: ruan.elliott@bbsrc.ac.uk

References

- Almind K & Kahn CR (2004) Genetic determinants of energy expenditure and insulin resistance in diet-induced obesity in mice. *Diabetes* **53**, 3274–3285.
- Atzmon G, Yang XM, Muzumdar R, Ma XH, Gabriely I & Barzilai N (2002) Differential gene expression between visceral and subcutaneous fat depots. *Horm Metab Res* **34**, 622–628.
- Ball CA, Sherlock G, Parkinson H, *et al.* (2002) The underlying principles of scientific publication. *Bioinformatics* **18**, 1409.
- Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W & Edgar R (2005) NCBI GEO: mining millions of expression profiles – database and tools. *Nucleic Acids Res* **33**, D562–D566.
- Barta P, Monti J, Maass PG, Gorzelnik K, Muller DN, Dechend R, Luft FC, Hubner N & Sharma AM (2002) A gene expression analysis in rat kidney following high and low salt intake. *J Hypertens* **20**, 1115–1120.
- Becker W, Kluge R, Kantner T, Linnartz K, Korn M, Tschank G, Plum L, Giesen K & Joost HG (2004) Differential hepatic gene expression in a polygenic mouse model with insulin resistance and hyperglycemia: evidence for a combined transcriptional dysregulation of gluconeogenesis and fatty acid synthesis. *J Mol Endocrinol* **32**, 195–208.
- Brazma A, Hingamp P, Quackenbush J, *et al.* (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat Genet* **29**, 365–371.
- Cao SX, Dhahbi JM, Mote PL & Spindler SR (2001) Genomic profiling of short- and long-term caloric restriction effects in the liver of aging mice. *Proc Natl Acad Sci USA* **98**, 10630–10635.
- Castro-Chavez F, Yechoor VK, Saha PK, Martinez-Botas J, Wooten EC, Sharma S, O'Connell P, Taegtmeier H & Chan L (2003) Coordinated upregulation of oxidative pathways and downregulation of lipid biosynthesis underlie obesity resistance in perilipin knockout mice: a microarray gene expression profile. *Diabetes* **52**, 2666–2674.
- Crott JW, Choi SW, Ordovas JM, Ditelberg JS & Mason JB (2004) Effects of dietary folate and aging on gene expression in the colonic mucosa of rats: implications for carcinogenesis. *Carcinogenesis* **25**, 69–76.
- Deng X, Elam MB, Wilcox HG, Cagen LM, Park EA, Raghov R, Patel D, Kumar P, Sheybani A & Russell JC (2004) Dietary olive oil and menhaden oil mitigate induction of lipogenesis in hyperinsulinemic corpulent JCR:LA-cp rats: microarray analysis of lipid-related gene expression. *Endocrinology* **145**, 5847–5861.
- Dhahbi JM, Kim HJ, Mote PL, Beaver RJ & Spindler SR (2004) Temporal linkage between the phenotypic and genomic responses to caloric restriction. *Proc Natl Acad Sci USA* **101**, 5524–5529.
- Eletto D, Leone A, Bifulco M & Tecce MF (2005) Effect of unsaturated fat intake from Mediterranean diet on rat liver mRNA expression profile: selective modulation of genes involved in lipid metabolism. *Nutr Metab Cardiovasc Dis* **15**, 13–23.
- Fujiwara K, Ochiai M, Ubagai T, Ohki M, Ohta T, Nagao M, Sugimura T & Nakagama H (2003) Differential gene expression profiles in colon epithelium of two rat strains with distinct susceptibility to colon carcinogenesis after exposure to PhIP in combination with dietary high fat. *Cancer Sci* **94**, 672–678.
- Gabrielsson BG, Johansson JM, Lonn M, Jernas M, Olbers T, Peltonen M, Larsson I, Lonn L, Sjostrom L, Carlsson B & Carlsson LM (2003) High expression of complement components in omental adipose tissue in obese men. *Obes Res* **11**, 699–708.
- Gregoire FM, Zhang Q, Smith SJ, Tong C, Ross D, Lopez H & West DB (2002) Diet-induced obesity and hepatic gene expression alterations in C57BL/6J and ICAM-1-deficient mice. *Am J Physiol* **282**, E703–E713.
- Guan H, Arany E, van Beek JP, Chamson-Reig A, Thyssen S, Hill DJ & Yang K (2005) Adipose tissue gene expression profiling reveals distinct molecular pathways that define visceral adiposity in offspring of maternal protein-restricted rats. *Am J Physiol* **288**, E663–E673.
- Lan H, Rabaglia ME, Stoehr JP, Nadler ST, Schueler KL, Zou F, Yandell BS & Attie AD (2003) Gene expression profiles of nondiabetic and diabetic obese mice suggest a role of hepatic lipogenic capacity in diabetes susceptibility. *Diabetes* **52**, 688–700.
- Lindon JC, Nicholson JK, Holmes E, Keun HC, *et al.* (2005) Summary recommendations for standardization and reporting of metabolomic analyses. *Nat Biotechnol* **23**, 833–838.
- Lopez IP, Marti A, Milagro FI, Zulet Md Mde L, Moreno-Aliaga MJ, Martinez JA & De Miguel C (2003) DNA microarray analysis of genes differentially expressed in diet-induced (cafeteria) obese rats. *Obes Res* **11**, 188–194.
- Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J & Apweiler R (2005) PRIDE: the proteomics identifications database. *Proteomics* **5**, 3537–3545.

- Moraes RC, Blondet A, Birkenkamp-Demtroeder K, Tirard J, Orntoft TF, Gertler A, Durand P, Naville D & Begeot M (2003) Study of the alteration of gene expression in adipose tissue of diet-induced obese mice by microarray and reverse transcription-polymerase chain reaction analyses. *Endocrinology* **144**, 4773–4782.
- Mutch DM, Simmering R, Donnicola D, Fotopoulos G, Holzwarth JA, Williamson G & Corthesy-Theulaz I (2004) Impact of commensal microbiota on murine gastrointestinal tract gene ontologies. *Physiol Genomics* **19**, 22–31.
- Napoli C, de Nigris F, Welch JS, Calara FB, Stuart RO, Glass CK & Palinski W (2002) Maternal hypercholesterolemia during pregnancy promotes early atherogenesis in LDL receptor-deficient mice and alters aortic gene expression determined by microarray. *Circulation* **105**, 1360–1367.
- Oana F, Homma T, Takeda H, Matsuzawa A, Akahane S, Isaji M & Akahane M (2005) DNA microarray analysis of white adipose tissue from obese (fa/fa) Zucker rats treated with a beta3-adrenoceptor agonist, KTO-7924. *Pharmacol Res* **52**, 395–400.
- Parkinson H, Sarkans U, Shojatalab M, *et al.* (2005) ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* **33**, D553–D555.
- Recinos A III, Carr BK, Bartos DB, Boldogh I, Carmical JR, Belalcazar LM & Brasier AR (2004) Liver gene expression associated with diet and lesion development in atherosclerosis-prone mice: induction of components of alternative complement pathway. *Physiol Genomics* **19**, 131–142.
- Roche HM, Noone E, Sewter C, Mc Bennett S, Savage D, Gibney MJ, O’Rahilly S & Vidal-Puig AJ (2002) Isomer-dependent metabolic effects of conjugated linoleic acid: insights from molecular markers sterol regulatory element-binding protein-1c and LXRalpha. *Diabetes* **51**, 2037–2044.
- Rota C, Rimbach G, Minihane AM, Stoecklin E & Barella L (2005) Dietary vitamin E modulates differential gene expression in the rat hippocampus: potential implications for its neuroprotective properties. *Nutr Neurosci* **8**, 21–29.
- Roy S, Lado BH, Khanna S & Sen CK (2002) Vitamin E sensitive genes in the developing rat fetal brain: a high-density oligonucleotide microarray analysis. *FEBS Lett* **530**, 17–23.
- Saito K, Arai S & Kato H (2005) A nutrigenomics database - integrated repository for publications and associated microarray data in nutrigenomics research. *Br J Nutr* **94**, 493–495.
- Sartipy P & Loskutoff DJ (2003) Expression profiling identifies genes that continue to respond to insulin in adipocytes made insulin-resistant by treatment with tumor necrosis factor-alpha. *J Biol Chem* **278**, 52298–52306.
- Sreekumar R, Halvatsiotis P, Schimke JC & Nair KS (2002) Gene expression profile in skeletal muscle of type 2 diabetes and the effect of insulin treatment. *Diabetes* **51**, 1913–1920.
- Tanaka T, Yamamoto J, Iwasaki S, *et al.* (2003) Activation of peroxisome proliferator-activated receptor delta induces fatty acid beta-oxidation in skeletal muscle and attenuates metabolic syndrome. *Proc Natl Acad Sci USA* **100**, 15924–15929.
- Taylor CF, Paton NW, Garwood KL, *et al.* (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat Biotechnol* **21**, 247–254.
- Tseng YH, Butte AJ, Kokkotou E, *et al.* (2005) Prediction of pre-adipocyte differentiation by gene expression reveals role of insulin receptor substrates and necdin. *Nat Cell Biol* **7**, 601–611.
- Tsuda T, Ueno Y, Kojo H, Yoshikawa T & Osawa T (2005) Gene expression profile of isolated rat adipocytes treated with anthocyanins. *Biochim Biophys Acta* **1733**, 137–147.
- van Breda SG, van Agen E, van Sanden S, Burzykowski T, Kienhuis AS, Kleinjans JC & van Delft JH (2005) Vegetables affect the expression of genes involved in anticarcinogenic processes in the colonic mucosa of C57BL/6 female mice. *J Nutr* **135**, 1879–1888.
- van Schothorst EM, Franssen-van Hal N, Schaap MM, Pennings J, Hoebee B & Keijer J (2005) Adipose gene expression patterns of weight gain suggest counteracting steroid hormone synthesis. *Obes Res* **13**, 1031–1041.
- Vohl MC, Sladek R, Robitaille J, Gurd S, Marceau P, Richard D, Hudson TJ & Tchernof A (2004) A survey of genes differentially expressed in subcutaneous and visceral adipose tissue in men. *Obes Res* **12**, 1217–1222.
- von Eyben FE, Kroustrup JP, Larsen JF & Celis J (2004) Comparison of gene expression in intra-abdominal and subcutaneous fat: a study of men with morbid obesity and nonobese men using microarray and proteomics. *Ann N Y Acad Sci* **1030**, 508–536.
- Xiao R, Badger TM & Simmen FA (2005) Dietary exposure to soy or whey proteins alters colonic global gene expression profiles during rat colon tumorigenesis. *Mol Cancer* **4**, 1.
- Yagil C, Hubner N, Monti J, Schulz H, Sapojnikov M, Luft FC, Ganten D & Yagil Y (2005) Identification of hypertension-related genes through an integrated genomic-transcriptomic approach. *Circ Res* **96**, 617–625.