Data re-use walk-through

In this supplement we provide an example of a specific data re-analysis, aimed to demonstrate some of the ideas described in the main text. Screen shots taken during the analysis are available as a supplement and referred to by numbers (#) in the following text. URLs for all the resources mentioned are provided in the main article text. There are many areas and tools we don't cover (e.g., quality control and meta-analysis), and the analysis is rather cursory, but we hope it provides some useful guidance.

We start with a gene selected by hand from the recent literature, *doublecortex* (DCX), an X-linked gene which is mutated in some forms of lissencephaly. Lissencephaly is a human neurodevelopmental disorder characterized by errors in neuronal migration; mouse Dcx is expressed in migrating cortical neurons (Friocourt et al., 2007) and is thought to be involved in process growth (Friocourt et al., 2003). We can use existing expression data to help investigate the function of DCX as it relates to other genes and experimental paradigms.

We first used Gemma and GeneNetwork to examine how DCX is related to other genes. In Gemma we used the coexpression analysis tool to examine the mouse Dcx gene. The top hit is Hn1, coexpressed with Dcx in 8 of 40 data sets with relevant data (2). Interestingly, Hn1 also showed up in a search for genes coexpressed with the human DCX gene, though not at the top of the list. A quick literature search indicates that not much is known about Hn1, but it is striking that Hn1 was found to be upregulated after nerve injury, and was hypothesized to play a role in neuronal development (Zujovic et al., 2005).

In GeneNetwork, we examined mouse Dcx expression in BXD recombinant inbred mice (3). Dcx knockout mice have mild hippocampal defects, so we thought looking at expression QTL in the hippocampus expression data would be a good place to start, but we also looked at a whole brain expression data set. Given that the known role of Dcx is most prominent in development, one might not think that looking in adult brain would yield the most interesting results, but if the gene is expressed in the adult brain (and it is,

as a quick look at the Allen brain atlas confirms (4,5)), we still might expect some useful insight.

Dcx has four probes in each of these data sets (6). We used the QTL cluster map function to produce a quick diagram showing loci whose variants are correlated with the expression of Dcx. While there are a number of sites that received significant LODs, the patterns are not all that consistent across the probes (which may target different transcripts) and the data sets. However, there is a region on chromosome 2 that has high LODs for three of the data points (7). For our exercise we focused on one probe in one data set, 1418141_at, in the hippocampus data.

Viewing the details for this probe, we first chose the "interval mapping" function, choosing to produce just a map for chromosome 2 (8). There is indeed a major apparent QTL for Dcx, spanning a region of at least 15 Mb (9). This means that genetic variants in the region are highly correlated with the expression level of Dcx; this type of "trans-acting" QTL might be hypothesized to involve genes that affect the activity of upstream regulators of Dcx expression. The region contains about 30 genes, and further work would involve investigating each gene's function, and, if sufficiently encouraged, looking for further evidence of a link to Dcx.

The last part of our investigation is to look for some potentially relevant data in GEO and find out how Dcx behaves. We searched GEO for "brain development" (10) and found many hits. One of the interesting data sets is GDS2135, a time course of postnatal forebrain development (Semeralul et al., 2006) (11). We used the GEO search facility to find expression profiles for Dcx in this data set, and the same probe we looked at in GeneNetwork appears to decline in expression after post-natal week 3, by a factor of at least 2 (12,13). Thus encouraged, we did a search for "Dcx AND brain" to look at other profiles for this gene. There is a lot to look at, but we were drawn to a profile in a human study of bipolar disorder (GDS2190, (Ryan et al., 2006)), where there looks to be somewhat lower expression of Dcx in cortical tissue from affected individuals (14). We used the t-test tool in GEO (15) to do a two-tailed t-test between the controls and affected groups (t-tests are only available for GDS* entries in GEO), yielding 985 probes meeting

a (lax) threshold of $p < 0.01$. Unfortunately GEO does not list the p-values and it is a little hard to navigate the results, so we reanalyzed the data using R (Team, 2007).

To analyze the data in R, we first downloaded the expression matrix file for the corresponding GEO series, GSE5388 (16). The details of analysis methods in R are beyond the scope of this walk-through, so we only mention that we found that the t-test p-value for Dcx is 0.03, suggesting a change, but hardly overwhelming. Furthermore, we found that another probe for Dcx shows no change (17). Next steps might be to do a more exhaustive and methodical examination of Dcx in all the data sets; unfortunately at this writing this is difficult to do.

To summarize, in this walk-through we demonstrated how to re-use expression data to examine gene coexpression to identify possible functionally-related genes; study quantitative trait loci for a gene of interest; use GEO to locate and examine expression profiles for our gene; use an atlas to confirm the expression of the gene in the brain; and finally how to download the data for an expression study in a format that is suitable for analysis in other tools.

References

Friocourt, G., Koulakoff, A., Chafey, P., Boucher, D., Fauchereau, F., Chelly, J., and Francis, F. (2003). Doublecortin functions at the extremities of growing neuronal processes. Cereb Cortex *13*, 620-626.

Friocourt, G., Liu, J.S., Antypa, M., Rakic, S., Walsh, C.A., and Parnavelas, J.G. (2007). Both doublecortin and doublecortin-like kinase play a role in cortical interneuron migration. J Neurosci *27*, 3875-3883.

Ryan, M.M., Lockstone, H.E., Huffaker, S.J., Wayland, M.T., Webster, M.J., and Bahn, S. (2006). Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. Mol Psychiatry *11*, 965-978.

Semeralul, M.O., Boutros, P.C., Likhodi, O., Okey, A.B., Van Tol, H.H., and Wong, A.H. (2006). Microarray analysis of the developing cortex. Journal of neurobiology *66*, 1646-1658.

Team, R.D.C. (2007). R: A language and environment for statistical computing. (R Foundation for Statistical Computing).

Zujovic, V., Luo, D., Baker, H.V., Lopez, M.C., Miller, K.R., Streit, W.J., and Harrison, J.K. (2005). The facial motor nucleus transcriptional program in response to peripheral nerve injury identifies Hn1 as a regeneration-associated gene. J Neurosci Res *82*, 581-591.