transactions
Biochemical Society Transactions www.biochemsoctrans.org

The use of Affymetrix GeneChips as a tool for studying alternative forms of RNA

Andrew P. Harrison[1], Joanna Rowsell, Renata da Silva Camargo, William B. Langdon, Maria Stalteri, Graham J.G. Upton and Jose M. Arteaga-Salas

Departments of Biological Sciences and Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, U.K.

Abstract

We are developing a computational pipeline to use surveys of Affymetrix GeneChips as a discovery tool for unravelling some of the biology associated with post-transcriptional processing of RNA. This work involves the integration of a number of bioinformatics resources, from comparing annotations to processing images to determining the structure of transcripts. The rapidly growing datasets of GeneChips available to the community puts us in a strong position to discover novel biology about post-transcriptional processing, and should enable us to determine the mechanisms by which some groups of genes make co-ordinated changes in their production of isoforms.

Introduction

Microarrays are pervasive technology, widely used in the life sciences to measure genome-wide transcriptional output in many organisms, phenotypes and tissues. A high-density oligonucleotide microarray, such as an Affymetrix GeneChip, contains hundreds of thousands of different probes, measuring the transcriptional concentration of tens of thousands of distinctive transcripts. GeneChip technology utilizes multiple independent probe hybridization events to measure the expression level for each gene investigated. Each probe is a 25 nt oligomer (25-mer) and each probeset, designed to represent a different gene transcript, typically consists of 11 perfect match probes as well as corresponding mismatch probes. A typical microarray experiment focuses on finding a set of genes associated with a given biological process. However, owing to the high costs associated with running a microarray experiment, a typical study will only use a relatively small number of microarrays, ranging from a handful to a few hundred. Moreover, because follow-up studies are laborious and expensive, each experiment leads to detailed studies of typically less than 100 genes associated with the process of interest. Consequently, the transcriptional information from tens of thousands of genes is being ignored by the experimenter, because it appears to have no relevance to the biological topic of interest. However, the data are available for analysis by others because many of the leading journals require the raw data from microarray studies to be deposited in public repositories. The popularity of microarrays is leading to a rapid growth of public repositories, such as GEO (Gene Expression Omnibus) [1], typically doubling in size every year. This provides a glut of data available for analysis.

Even with the successful use of GeneChips, it is clear that we can learn much more from these data. Many of the improvements in the use of GeneChips will derive from computational methods to either extract biological signals from the data or remove systematic errors introduced by the technology. However, because the repositories contain data from experiments that cover a range of organisms, phenotypes and disease states, care needs to be taken in order to bring subsections of the data together in order to ask meaningful biological questions. The strategy we are adopting is to focus on aspects of post-transcriptional processing of RNA which we have discovered acts to modify the interpretation of Affymetrix GeneChips [2], and to look for these signals in GEO.

Alternative splicing can be detected using GeneChips

GeneChips can detect alternative splicing signatures, as we have recently illustrated for the *Calca* (calcitonin/calcitonin-related polypeptide α) gene [2]. The mammalian *Calca* gene has six exons and is considered to be a model gene for the study of alternative splicing [3]. Splicing together the first four exons produces the mRNA for calcitonin, a hormone produced by the thyroid gland. Splicing together exons 1–3, 5 and 6 produces the mRNA for CGRP (calcitonin gene-related peptide), which is expressed in the nervous system. The *Calca* gene is represented by three probesets on the rat RAE230A chip (Figure 1): 1369116_a_at maps to exons 2 and 3 (calcitonin and CGRP); 1369117_at maps to exons 1–4 (mainly calcitonin); 1370775_a_at maps to exons 3, 5 and 6 (mainly CGRP). GeneChips from an experiment studying the effects of an operation on the spinal cord indicated that probeset 1370775_a_at was significantly down-regulated, whereas probeset 1369117_at showed little variation [2]. These observations make sense only because we have a detailed knowledge of the splicing and expression patterns for calcitonin/CGRP. However, without this knowledge, a confusing picture of expression arises from

Key words: Affymetrix GeneChip, alternative polyadenylation, alternative splicing, microarray, post-transcriptional processing, RNA.

Abbreviations used: *Calca*, calcitonin/calcitonin-related polypeptide α; CGRP, calcitonin gene-related peptide; GEO, Gene Expression Omnibus.

[1]To whom correspondence should be addressed (email harry@essex.ac.uk).

Figure 1 | **Mapping of individual probes in probe sets 1369117_at (black lines; top), 1370775_a_at (light-grey lines; middle) and 1369116_a_at (dark-grey lines; bottom) to genomic sequence of the rat *Calca* gene on chromosome 1 (UCSC June 2003 assembly, http://genome.ucsc.edu)**

The introns are represented by broken lines and are not drawn to scale.

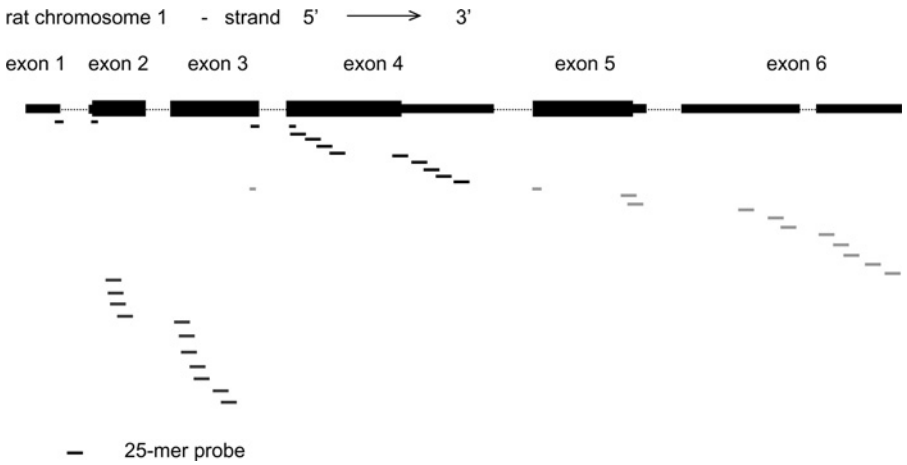
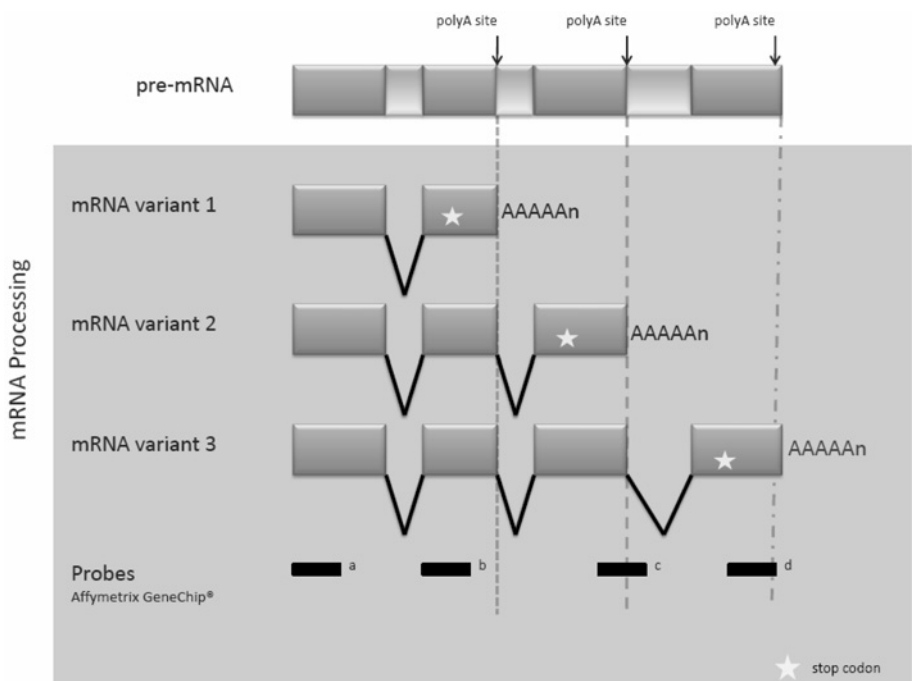


Figure 2 | **Effect of alternative polyadenylation on Affymetrix probe detection**

Based on data from [4].



combining multiple probesets for the same 'gene'. The moral of this is that, through correcting for splicing, we identify the differences in differential expression for two isoforms with low noise, whereas not including knowledge of splicing leads to a more confusing and noisy estimate of differential expression. Moreover, from the observed differences between differential expression estimates for groups of probesets taken from the same genes, we have evidence for alternative splicing in several other genes, including the α1I subunit of a calcium channel (*Cacna1i*) and A-kinase-anchoring protein 1 (*Akap1*). This suggests that splicing modifies the interpretation of a significant number of probesets in different GeneChip experiments, as expected for a phenomenon which affects up to 50 % of genes in higher eukaryotes.

Alternative polyadenylation can be detected using GeneChips

We have identified alternative polyadenylation as a reason some groups of probesets appeared to show a gene being up-regulated and down-regulated together [2]. As Figure 2 shows, the probes affected by alternative polyadenylation are those which map to regions downstream of the most upstream poly(A) site and regions with a poly(A) site. D'mello et al. [4] studied human, mouse and rat Affymetrix microarray probes and found that a large number could be affected by alternative polyadenylation. This may have consequences for interpreting the microarray data because alternative polyadenylation may be the cause of changes in expression values. It was found that the Affymetrix probes are biased toward the 3′-end of the transcripts, leading to possible detection bias [4]. Moreover, the probes that detect sequences before and after a polyadenylation site could be used to discover more about alternative polyadenylation, such as finding groups of genes using the same polyadenylation signals.

Summary

We are developing a computational pipeline to use GeneChips as a tool for identifying co-ordinated events such as alternative polyadenylation and chimaeric transcripts. This work involves the integration of a number of bioinformatics resources, from comparing annotations to processing images to determining the structure of transcripts. The rapidly growing datasets of GeneChips available to the community puts us in a strong position to discover novel biology about post-transcriptional processing, and should enable us to determine the mechanisms by which groups of genes make co-ordinated decisions in their choice of exons, UTRs (untranslated regions) and transcript terminations.

R. da S.C. and W.B.L. are supported by a grant from the BBSRC (Biotechnology and Biological Sciences Research Council) (BB/E001742/1), and J.R. is supported by a Strategic Studentship from the BBSRC (BBS/S/H/2005/11996A). J.M.A.-S. is supported by a scholarship from CONACyT (Consejo Nacional de Ciencia y Tecnología).

References

1 Barrett, T., Suzek, T., Troup, D., Wilhite, S., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A., Fujibuchi, W. and Edgar, R. (2005) NCBI GEO: mining millions of expression profiles-database and tools. Nucleic Acids Res. **33**, D562–D566

2 Stalteri, M.A. and Harrison, A.P. (2007) Interpretation of multiple probesets mapping to the same gene in Affymetrix GeneChips. BMC Bioinformatics **8**, 13

3 Coleman, T., Tran, Q. and Roesser, J. (2003) Binding of a candidate splice regulator to a calcitonin-specific splice enhancer regulates calcitonin/CGRP pre-mRNA splicing. Biochem. Biophys. Acta **1625**, 153–164

4 D'mello, V., Lee, J.Y., MacDonald, C.C. and Tian, B. (2006) Alternative mRNA polyadenylation can potentially affect detection of gene expression by Affymetrix GeneChip® arrays. Appl. Bioinformatics **5**, 249–253

Received 14 January 2008
doi:10.1042/BST0360511