ADAPT: a database of affymetrix probesets and transcripts

Hui Sun Leong, Tim Yates, Claire Wilson and Crispin J. Miller*

Paterson Institute for Cancer Research, Wilmslow Road, Manchester M20 4BX, UK

Received on November 24, 2004; revised on January 31, 2005; accepted on February 24, 2005

Advance Access publication March 3, 2005

1 INTRODUCTION

Affymetrix microarrays record the presence of a transcript in a solution by measuring the level of hybridization between the transcript and a set of short (typically 25mer) oligonucleotide probes anchored to the array surface. Each ‘probeset’ consists of a series of ‘perfect match’ (PM) probes, designed to match exactly the transcript, and a series of ‘mismatch probes’ (MM), identical to the PM probes except that the middle residue has been changed. Hybridization conditions are controlled with the aim of maximizing the binding between a transcript and its PM probes, while minimizing the binding to its MM probes. The intention is that the PM probes record the presence of the transcript, while MM probes measure background and non-specific hybridization (Affymetrix, 2004, http://www.affymetrix.com/support/technical/manual/expression_manual.affx). One advantage of this approach is that the combination of short oligos and strict hybridization conditions makes it possible to use in silico searches to predict which probes are likely to bind to which transcripts—information that is important because many transcripts have similar sequences (e.g. alternate splicing can lead to a set of transcripts being encoded by a single gene, due to homology or due to repetitive or low complexity regions).

Each probeset is typically designed to match the (more variable) 3’ non-coding region of its target transcript; however, it is not always possible to identify a set of probes that reliably and uniquely identify a particular transcript, and relax the design criteria accordingly. This is indicated in the naming convention used for the probesets. The situation is complex, but generally:

- Those ending with ‘_at’ are designed to recognize transcripts uniquely.
- Those ending with ‘_s_at’ or ‘_a_at’ are designed to recognize multiple transcripts from the same gene family.

- Those ending with ‘_x_at’ may cross-hybridize into completely unrelated sequences.

Other suffixes exist, and the exact meaning can be dependent on the array type. More information can be found at Affymetrix’s website: http://www.affymetrix.com

Not only do some probesets target multiple transcripts, the reverse is also true—there are multiple probesets that target a single transcript. This can occur, for example, with probesets designed to identify different splice-variants of the same gene, or where one probeset is designed to identify a gene family, while another probeset targets a particular family member.

Identifying these situations is useful when considering experimental data in which evidence from a particular probeset is weak. If all the other probesets targeting the same transcript behave similarly, this can provide supporting evidence; if they behave differently it may be possible to discount the probeset from further analysis.

Another confounding factor occurs because arrays are designed against sequence databases that are in a state of continual growth. Each array therefore represents a snapshot based on the knowledge available at the time it was created. A number of resources provide more recent annotations [e.g. NetAffx (Liu et al., 2003), Resourcerer (Tsai et al., 2001), DAVID (Dennis et al., 2003), ARROGANT (Kulkarni et al., 2002), GeneAnnot (Chalifa-Caspi et al., 2004) and Ensembl (Birney et al., 2004)].

While NetAffx, GeneAnnot and Ensembl all offer information describing potential cross-hybridizations, they do so as a part of a comprehensive collection of functional annotation. ADAPT, by focusing specifically on cross-hybridization, is able to offer greater detail on a larger range of arrays (including HGU95, HGU133 and recent murine and human ‘plus 2’ arrays). In particular, ADAPT makes it possible to investigate not only which transcripts a probeset might be expected to hit, but also which probesets hit multiple transcripts. This is achieved through clickable images that show the location of all probes and probesets that match each individual transcript (Fig. 1).

The database is queried by providing one or more probesets, transcript or gene ids, accompanied by an optional filter to restrict the results to a specific array type. If a single identifier is specified, the search returns a series of images such as Figure 1; if multiple identifiers are used, an overview page provides a summary of the matches. Clicking on a probeset within an image results in all matching transcripts to that probeset being shown (along with all other probesets that hit each transcript). In this way the database can be browsed.
ADAPT is populated by searching all probe sequences for exact matches to RefSeq and Ensembl transcripts (for more details see the database FAQ on the website). For RefSeq both ‘known’ and ‘model’ mRNA sequences are used; for Ensembl, ADAPT uses cDNA sequences assigned ‘known’, ‘novel’ or ‘pseudo’ status. Both databases were used because they employ different methods to predict transcript/gene sequences. In particular, it was found that the RefSeq entries sometimes contained more complete 3′-UTRs (e.g. the ADAPT entry for 219550_at). This is crucial, because this is where the majority of Affymetrix probesets are designed to hybridize.

A conscious decision was made not to extract large amounts of annotation data from the primary sources, but instead to provide hyperlinks to existing resources (via the ‘find’ icons in Fig. 1). It was felt that this made it easier to offer the full functionality offered by sites such as GeneCards, Ensembl and the NCBI, rather than providing access to a subset of the data.

At the time of writing, ADAPT contained mappings for over 250,000 probesets, hitting \( \sim 180,000 \) RefSeq transcripts represented on 23 different array types. On the HGU133 plus 2.0 array, 7124 (23.12%) probesets match more than one RefSeq transcript, while 2489 (8.08%) hit sequences emanating from different genes. Given that a significant number of probesets are likely to cross-hybridize, it is important to be aware of these issues when interpreting microarray data.

ACKNOWLEDGEMENT

This work was funded by Cancer Research, UK.

REFERENCES

Tsai,J. et al. (2001) RESOURCERER: a database for annotating and linking microarray resources within and across spec. Genome Biology, 2, software0002.1–0002.4.