

A Novel Method for Detecting mRNA Transcripts Using Genome-wide Tiling Arrays

Gabor Halasz¹, Marinus F. van Batenburg^{1,2}, Joelle R. Perusse³, Kevin P. White³, Harmen J. Bussemaker¹

Keywords: genomic tiling arrays, transcript detection, non-coding RNAs, probe sequence correction, normalization, cross-hybridization, *Anopheles gambiae*.

1 Introduction

DNA microarrays are invaluable tools for measuring changes in gene expression between two conditions. This technology has recently been adapted for genome-wide transcript discovery [1]. In such experiments, the question asked is not whether the expression of a certain locus changes, but whether a particular genomic region or splice isoform is transcribed at all. The corresponding quantity of interest is therefore the absolute, rather than relative, signal intensity measured by a microarray probe. Determining whether genomic regions are expressed above background requires new experimental and analytical methods. Primary considerations include selecting an appropriate negative control, accommodating dye- and array- specific effects, and controlling for the effect of probe sequence composition on signal intensity. Here we introduce PEAB-finder, our algorithm for detecting Probes Expressed Above Background, and use it to analyze data from a genome-wide tiling array assaying the *Anopheles gambiae* transcriptome.

2 Materials and Methods

Genomic tiling array data for *Anopheles gambiae* was collected essentially as described in Stolc, et al [1]. A total of 194,212 36-mer probes interrogating exonic, intronic, and intergenic regions of the *A. gambiae* genome were synthesized on five arrays. Another 1,000 probes with imperfect complementarity to the genome (at least 3 mismatches anywhere in the genome) served as negative control probes (NCP's). RNA isolated from male and female adult mosquitoes were fluorescently labeled and hybridized to each array, using an experimental design that included dye swapping.

Figure 1 shows the pipeline used to determine which exon (EP) and non-exon (NEP) probes were significantly expressed above background noise. Probe signal intensities were first corrected for sequence-specific biases by fitting a position-dependent model to the NCP probes, in which different segments of the probe are allowed to make independent contributions to binding. For each probe, we then derived a p-value reflecting the likelihood that the corresponding signal intensity belongs to the background distribution, represented by the NCP's. P-values were calculated separately for each channel (combination of array and dye), and the channel-specific values were combined into a single p-value using the method described by Bailey and Gribskov

¹Department of Biological Sciences, Columbia University. E-mail: gh74@columbia.edu

²Bioinformatics Laboratory, Academic Medical Center, University of Amsterdam, Netherlands.

³Department of Genetics, Yale University. E-mail: kevin.white@yale.edu

[3]. To address multiple hypothesis testing, the procedure of Benjamini-Hochberg [4] was then applied to these p-values, using a false discovery rate of five percent.

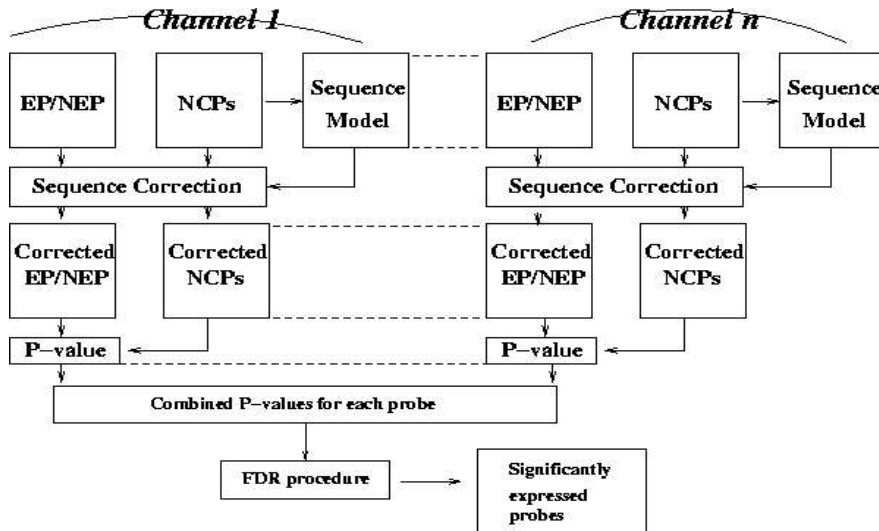


Figure 1 – pipeline for detecting probes expressed above background (PEAB)

3 Results

We created a program, PEAB-finder, that implements the pipeline detailed in figure 1. Given the probe sequences and signal intensities for a set of identically designed arrays, PEAB-finder returns a list of probes expressed above background. Our method is novel in its completely non-parametric approach to the problem of signal variability across channels- no assumptions are made about the distribution of signal intensities in each channel (in contrast to methods such as quantile normalization [5]). The only assumption is that of a monotonic relationship between RNA abundance and signal intensity.

Running PEAB-finder on the *Anopheles gambiae* data set described above revealed significant expression from 71.6 percent of exon probes and 36 percent of non-exon probes. These results are similar to those obtained in *Drosophila melanogaster* [1]. The software is available at <http://bussemaker.bio.columbia.edu/software/PEAB-finder/>.

References

- [1] Stolc, V. et al., 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* 306(5696):655-60
- [2] Bertone, P. et al, 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* 306(5705):2242-6
- [3] Bailey, TL and Gribskov, M. 1998. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14(1):48-54.
- [4] Benjamini, Y and Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, series B, 1995.
- [5] Bolstad, BM et al, 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185-93