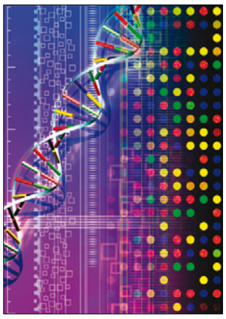
The importance of platform annotation in interpreting microarray data

Weigelt and colleagues[1] recently compared three methods for classifying breast cancers into molecular subtypes. Aside from an obvious question of whether they should have centred each gene, an issue previously discussed in the context of classifying breast-cancer subtypes,[2] there is concern that insufficient attention has been paid to the annotation of probes on the microarray platforms that were used. We review the PAM50 (pediction analysis of microarrays 50 gene set)[3] SSP (single sample predictor), as applied to the Illumina BeadArray (Illumina Inc, San Diego, CA, USA) generated Natrajan data set.[4]

Three methods of mapping microarray probes to genes were assessed by Weigelt and colleagues,[1] presenting the results from a Human Genome Organisation (HUGO) gene-symbol mapping. Gene symbols in the PAM50 list of approved HUGO symbols were updated, but not the symbols in their annotation. Consequently, probes for *NDC80* and *NUF2* could not be mapped, even though such probes exist under the former symbols of *KNTC2* and *CDCA1*. Thus Weigelt and colleagues[1] immediately restrict themselves to only 48 of the PAM50 genes.

The Illumina BeadArray differs from many microarrays; it is neither gene-centric (providing one measure of activity for each gene) nor transcript-centric (providing comprehensive interrogation of differing isoforms), but lies somewhere in between. The array contains a mixture of 3′ situated probes that target constitutive exons (which provide a good measure of overall gene expression), and additional probes that target subsets of transcripts, and often single, rare, or speculative transcripts. The best measure of gene activity is not the average of all probes mapping to a gene symbol, as used by Weigelt and colleagues,[1] but generally the single probe that has been designed to represent the overall gene.

We have previously undertaken the reannotation of this platform,[5] and have identified further problems with some probes that would be reason enough for their exclusion from the analyses by Weigelt and colleagues,[1] even if probe averaging was not considered an issue. There are probes that are too 5′ to give reliable signals, which target introns, have secondary targets, or have mismatches in their sequences. We have identified probes that cover single-nucleotide polymorphisms (SNPs), lie in repeat-masked regions, and one used by Weigelt and colleagues[1] that targets the wrong genomic strand.


Pasieka/Science Photo Library

This is not an abstract discussion, but affects results. Consider the ten PAM50 genes that discriminate the human epidermal growth factor receptor 2 (HER2; also ERBB2)-enriched class. 13 samples are noted as histologically HER2 positive, and we note that these do not cluster together by using average values for the ten genes in question, but do cluster when using an appropriate single probe for each gene. HER2 becomes dramatically more discriminatory when the appropriate probe is not diluted with noise from two others—with the measured difference in HER2 expression between the HER2 negative and HER2 positive groups changing from two-fold to four-fold. The signal-to-noise of *GRB7* expression is similarly improved, with the difference in the log-intensities from HER2 negative to HER2 positive rising from 1·85 to 2·91.

The message that inconsistencies exist between different classifiers might hold, but we would urge caution in the interpretation of microarray results without careful assessment of the microarray annotation.

*Mark J Dunning, Christina Curtis, Nuno L Barbosa-Morais, Carlos Caldas, Simon Tavaré, Andrew G Lynch**
University of Cambridge (ChC, NLB-M, CC, ST, AGL), CRUK (MJD)—Cambridge Research Institute, Department of Oncology, Robinson Way, Cambridge, CB2 0RE, UK
andy.lynch@cancer.org.uk

AGL, ChC, CC, and ST have received grants from Cancer Research UK. All other authors declared no conflicts of interest.

1 Weigelt B, Mackay A, A'hern R, et al. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol* 2010; **11:** 339–49.

2 Lusa L, McShane LM, Reid JF, et al. Challenges in projecting clustering results across gene expression-profiling datasets. *J Natl Cancer Inst* 2007; **99:** 1715–23.

3 Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 2009; **27:** 1160–67.

4 Natrajan R, Weigelt B, Mackay A, et al. An integrative genomic and transcriptomic analysis reveals molecular pathways and networks regulated by copy number aberrations in basal-like, HER2 and luminal cancers. *Breast Cancer Res Treat* 2009; **121:** 575–89.

5 Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, et al. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res* 2009; published online Nov 18. DOI:10.1093/nar/gkp942.