Minireview

Does massively parallel transcriptome analysis signify the end of cancer histopathology as we know it?

Samuel AJR Aparicio, Carlos Caldas and Bruce Ponder

Address: Department of Oncology, University of Cambridge, Wellcome Trust/MRC Building, Addenbrookes Hospital, Cambridge CB2 2XY, UK.

Correspondence: Samuel AJR Aparicio. E-mail: saparici@hgmp.mrc.ac.uk

Published: 15 September 2000

Genome ***Biology*** 2000, **1(3)**:reviews1021.1–1021.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/3/reviews/1021

Abstract

Transcriptional analysis of all the genes expressed by breast tumors has provided the first steps towards defining a molecular signature for the disease, and might ultimately make conventional diagnostic techniques obsolete.

The traditional way of classifying tumors is by histopathology, the staining and analysis of tissue samples. Now, the ability to analyse changes in the levels of the transcripts and/or protein products for literally thousands of genes promises interesting possibilities as a research tool - for understanding the underlying molecular mechanisms, but also for automated tissue diagnosis. Such approaches to biology and medicine have been termed 'massively parallel analysis'. Although the technologies which permit massively parallel analysis of the transcriptome (the transcribed fraction of genes in a genome) or the proteome (the expressed fraction of genes in a genome) are still in a phase of rapid evolution, the first studies applying these techniques and addressing the most obvious initial questions are now being published.

A key question in assessing the utility of these techniques is whether sufficiently dense and accurate sampling of gene expression in any given tissue would allow objective molecular classification of that tissue. If this were to prove possible, then objective and automated diagnosis within an intact tissue would become a realistic possibility. A potentially formidable obstacle to reaching this goal is that tissues are multicellular by definition, and they therefore contain cells in different states and in varying quantities. It is widely assumed that in order to obtain meaningful data, it would be necessary physically to separate different cell populations in a given tissue sample, before undertaking expression analysis. Another potential concern, specific to studies of cancer, is that genetic heterogeneity between tumor cells with unstable genomes would lead to heterogeneous and uninterpretable expression data. In a recent article published in *Nature* [1], the groups of Botstein and Brown show that, at least in the case of advanced breast tumors, not only does each tumor have a unique transcriptome signature but subclassification of tumor types is possible by computationally extracting cell-type-specific signatures out of the expression data for the whole tumor (which would include admixed non-neoplastic cells).

In this study, the Stanford group used a cDNA microarray to assess changes in the expression level of some 8,102 genes or expressed sequence tags (ESTs) in 40 patients with advanced (T3/T4 clinical stage) breast carcinoma. As a 'baseline', the study used a pooled set of RNA samples from disparate cell lines; these might reflect non-tumor cell types present in the 'tumor' sample. Crucially, in 20 cases the authors were able to obtain paired samples of tumor, a 'Before' sample, taken at the time of diagnosis, and an 'After' sample taken from same the tumor 16 weeks after preoperative doxorubicin therapy. In two cases the paired samples were tumor material and corresponding lymph node metastasis. No attempt was made to microdissect the tumors, but only to ensure that the samples contained a high proportion

of tumor. Perou *et al.* chose to analyse only the 1,753 (22%) genes/ESTs that showed a greater than four-fold difference in expression between samples; the basis for this analysis is the clustering of expression signatures.

Molecular signatures

The clustering process allows the most similar expression profiles to be grouped together, and the result can be represented as a dendrogram, or tree, of the type used to show evolutionary relationships. Crucially, it is a consequence of clustering, which places the most related samples in proximity, that the expression values of key gene sets can be seen to group together (as a consequence of the clustering). Since this occurs independent of any knowledge of the type of gene involved, anonymous genes/ESTs may become implicated in certain categories by association. For example, by grouping together genes known to encode targets for present chemotherapeutic agents, it can become apparent that some anonymous ESTs also cluster in these groups; and assuming that the ESTs represent genes, then the anonymous genes become potential targets for new therapeutic agents.

The initial clustering by Perou *et al.* [1] generates a surprising conclusion - namely that individual signatures can be recognized for each tumor and that these signatures are stable. That is, in all but five cases, the clustering of genes expressed in the Before and After samples for any individual patient's tumor was closer than that between different patients. Although the paper [1] does not try explicitly to address correlations with clinical outcomes or response to therapy, it appears that at least three of the After samples that clustered closer to the normal or benign fibroadenoma branches of the cluster tree were clinically 'doxorubicin responders' - the tumor cells were killed by doxorubicin. In a further analysis, the authors reclustered the samples, but using different subsets of genes. The fact that paired tumor samples from the same patient were available in 20 cases allowed the authors to derive from the expression data for the 1753 genes a subset that shows the greatest informative difference between tumors (as opposed to differing between the Before and After samples of the same tumor). A second subset, of 'epithelial' genes, was generated from previous work on epithelial cell lines and dissected tissues; and this subset is particularly important because breast tumors arise from, and are surrounded by, epithelial tissue. Re-clustering of the samples on the basis of these two subsets produces patterns of clustering nearly identical to each other (although only 25% of the genes in overlap between the two subsets), suggesting that the classification of tumors by gene expression profile is robust.

Furthermore, the dendrograms reveal sub-groupings of tumors which, when gated for genes known to be expressed in certain cell subtypes - for example myoepithelial versus luminal epithelial - reveals that apparently different molecular subtypes of tumor can be robustly distinguished and classified. This form of 'signature extraction' - in effect re-analyzing the relationships within the tumor sample in the light of known gene expression patterns from microarray experiments on different cell types that might be present - is likely to prove crucial not only to the molecular diagnosis and classification of tumors but also as a tool for attacking mechanistic questions. For example, it allows 'stromal' expression components to be separated from 'epithelial' by extracting these signatures from the general expression 'noise' in the data. Analysis of the resulting data subset suggests that the tumor cell is driving the expression of stromal genes in a clone-dependent fashion. So, not only does a particular cancer cell have an intrinsic pattern of gene expression but it also induces a stereotyped pattern of gene expression in the surrounding stroma. It will prove important to the future progress of such studies to ensure that different microarray experiments, done in different labs, can be compared for these purposes.

Towards molecular diagnosis

The data and analysis of the study by Perou *et al.* [1] show that molecular classification of whole tumor samples is possible. Since the need for physical separation of cell types was avoided in this case, in principle it suggests a basis for automated diagnosis of tumors. But certain caveats apply to this study. First, the tumors sampled were large, and it is not known whether it will prove to be the case that earlier-stage tumors will also show such canalized expression patterns. For example, the molecular evolution of ductal carcinoma *in situ*/lobular carcinoma *in situ* as precursor lesions into invasive carcinomas has not yet been studied using massively parallel arrays. It may be that tissue samples which show mixed 'precursor plus invasive' lesions would prove unclassifiable by this method. The general applicability of massively parallel analysis to diagnosis will therefore not be known until more tumor types have been examined, to test whether expression analysis can reliably separate tumor types and grades in different situations. It may be that, in some cases, better 'control' RNA samples will be needed or that a combination of whole-genome analysis and expression analysis will be required to separate tumor types.

Further challenges

The challenges presented by the huge quantities of data produced by microarray expression analysis have analogies in the fields of signal processing and astronomy, where signatures can be extracted from noisy data and where changes in tens of thousands of data points have to be monitored and analysed. Furthermore, it is not at all clear that the cluster analysis presented by Perou *et al.* is in fact the optimal basis for analysis of these data. The future should see some interesting cross-fertilization from areas such as physics,

astronomy and signal processing, offering potentially more powerful analytical methods.

These studies are of course only the beginning. The same group has already shown [2] that molecular sub-classification of certain lymphomas can be achieved, and that this can correlate with prognosis. Another group has also shown that expression profiling can 'blindly' sub-classify acute leukemias [3]. In addition to obtaining mechanistic insights, the molecular classification of solid tumors may have an impact in at least two key areas. First, objective and reproducible classification of tumors will greatly facilitate future clinical studies of treatment outcomes. Second, it may eventually offer much more accuracy and precision in forecasting outcome and treatment modalities in a way that will allow individualization of therapy. To achieve these goals, it will first be necessary to correlate the molecular classification with treatment outcomes. In this respect, RNA expression analysis will certainly be only one aspect. Informative data and increased correlative power may be derived from proteomics or from detailed analysis of the profile of genomic copy numbers for individual tumors, using array-type comparative genomic hybridization methods, for example. Future studies that establish correlations between the analytical capabilities of these methods on the one hand, and parameters such as disease progression and response to therapy on the other, will reveal the extent to which each of these techniques can contribute. For the near-term future, however, it is unlikely that clinical or diagnostic practice will be changed by massively parallel analysis; but in the medium to long term some aspects of current practice may be completely replaced by these methods and their derivatives.

References

1. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al.*: **Molecular portraits of human breast tumors.** Nature 2000, **406:**747-752.
2. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI *et al.*: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** Nature 2000, **403:**503-511.
3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al.*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** Science 1999, **286:**531-537.