Res. Microbiol. 151 (2000) 113–119

S0923250800000120-0/REV

Transcriptional profiling: is it worth the money?

Jörg D. Hoheisel[a]*, Martin Vingron[b]

[a] *Functional Genome Analysis, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 506, 69120 Heidelberg, Germany*
[b] *Theoretical Bioinformatics, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 506, 69120 Heidelberg, Germany*

Abstract — Transcriptional profiling on DNA arrays has become a synonym for the type of analyses that aim to understand cellular functioning in a comprehensive manner. In this review, the status of the technology is briefly discussed, with emphasis on some inherent weaknesses and problems. © 2000 Éditions scientifiques et médicales Elsevier SAS

transcriptional profiling / differential gene expression / DNA microarray / functional genomics

1. Introduction

For many, transcriptional profiling seems to have made its appearance 5 years ago with a publication by Patrick Brown and colleagues on initial analyses on a few *Arabidopsis* genes [23]. However, the ingredients that this success story is made of are older. Arrayed DNA in the form of clones or isolated DNA has been used for decades in the form of dot-blots. Hans Lehrach then initiated both the development and application of high-resolution robotics for the purpose of arraying large numbers at high density [15, 21], also found to be a successful tool for making library resources commonly accessible [17, 28]. The use of cDNA libraries in this format was discussed [16] and its value for actual transcriptional analyses demonstrated by several groups early on [1, 11, 12]. DNA chip technology emerged from work carried out at several places [2, 8, 10, 14, 25] and a first meeting on the subject was organised by Andrei Mirzabekov at the Engelhardt Institute in Moscow in 1991 [5]. However, those were the times of sequencing and, apart from technical issues, the focus then was on the use of this technology for high-throughput sequencing. Only later, when many technical problems of this most challenging task became more apparent (and were found difficult to solve) and other applications grew more imminent and important, did emphasis on chip-based work change, but it was still more directed toward sequence variations than functional analyses. Also, the dual fluorescence colouring scheme used by many for detecting different transcript levels had been utilised previously in microscopic analyses (e.g. [9, 19]).

As ever so often, however, the accomplishment of Patrick Brown and his colleagues was the clever combination of all this technology, throwing in a few extras for good measure, and doing so at the right time, when not sequencing but the (subsequent) functional analyses were becoming the bottleneck in genomics; in addition, they provided a perspective on what could be achieved by this technology. Ever since, transcriptional analysis has moved – and justifiably so – more and more into the centre of scientific attention, mostly because of the availability of the technology described above. Other approaches based on different technologies, such as differential display [18] or SAGE (serial analysis of gene expression) [26], for example, for all their merits, seem to become obsolete, whether for the right or wrong reasons. Only DNA arrays appear to combine the ability of quantification with the high degree of parallel-

* Correspondence and reprints
Tel.: +49 6221 424680; fax +49 6221 424682;
j.hoheisel@dkfz-heidelberg.de

ism essential for genomic approaches; they exhibit the – at first glance – advantage that all genes can be looked at simultaneously and still allow for the analysis of relatively rare transcripts. Currently, there exists a sort of gold-rush expectation that transcriptional profiling experiments on microarrays will produce many of the results necessary for understanding cellular activities, and infrastructure for such studies is being created at very many places worldwide. Here, a brief assessment is given of the current technology status.

2. Hybridisation

2.1. Support media

Since the methodology is based on hybridisation, the mechanisms that govern this process are of much concern to the procedure. Everything that influences both kinetics and thermodynamics can have substantial consequences. For the dynamic range of measurement and thus accuracy of quantification, the type of solid support is crucial. While glass exhibits an inertness to most chemical processes and acts favourably because of its good optical characteristics, it has the disadvantage that biomolecules can be attached only at a low concentration. Nylon filters, on the other hand, bind much more material but concomitantly limit both the probe accessibility and the density of arrays because of their porous structure. Alternative support media and/or linkage chemistries exist, however, that do not impede advantageous features while simultaneously getting rid of most if not all of the problems [3, 20, 22]. Nevertheless, many still use filters since their high loading capacity translates directly into good dynamic range.

2.2. Specificity

Although double-strand formation is rather sensitive to base mismatches, cross-hybridisation is a frequently occurring problem that can strongly distort results. The relatively large degree of redundancy of (especially eukaryotic) genomes, the repeated presence of certain sequence features in otherwise unrelated genes and the sequence biases observed in basically all organisms make the clear distinction of all genes difficult to achieve. With PCR products being the probes attached to the support (*figure 1*), common sequence domains or repetitive elements result in cross-hybridisation. Also, frequently, classes of genes are present that consist of similar sequence but might not even be involved in the same functional process and hence could be regulated very differently. Oligonucleotides exhibit a higher specificity, since mismatch discrimination is much improved for their shortness. However, even for semi-quantitative analyses the average of relatively numerous oligonucleotides must be determined per individual gene, because of their highly variable performance in hybridisa-

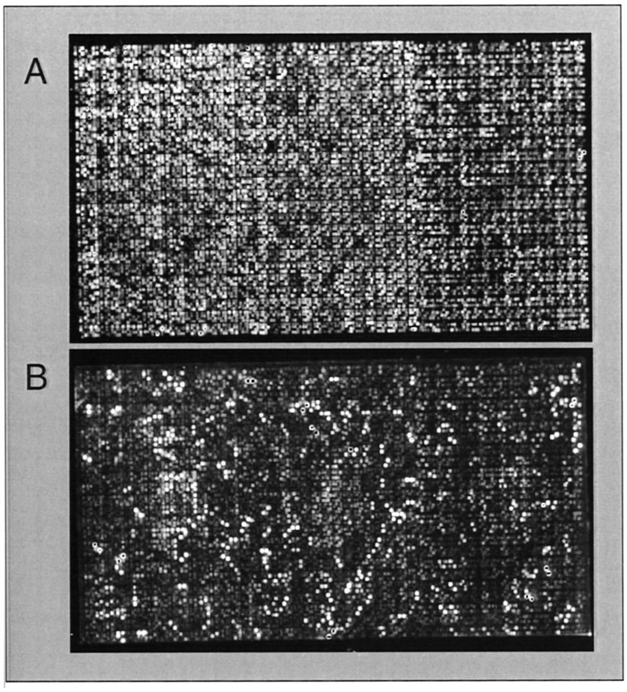

Figure 1. Hybridisations to the probe array. An array consisting of PCR products that represent the about 6 200 yeast genes was hybridised with an oligonucleotide binding to a common tag sequence present in the forward primer molecules of all PCR products (top panel) and with a complex DNA target generated by oligo-dT primed reverse transcription from total yeast RNA (bottom panel).

tion. Therefore, the identification of a set of experimentally unique oligonucleotides proves more difficult than it looks at first glance. Because of their very different duplex stabilities, differences between oligonucleotides should be larger than just one or two nucleotides and preferentially be located around the centre of the oligonucleotide. These and other restrictions considerably limit the freedom of choice and make the definition of a well-working set a formidable task.

2.3. Probe quality

Purity and status of the probe material attached to the chip is another factor that influences the analysis. PCR products are usually purified before being spotted to glass (while unpurified material is applied to nylon filters). With the target DNA being far in excess anyway, the unpurified PCR probe would result in loss of signal unless either the capacity is increased or bonding occurs selectively, both of which are possible. Direct in situ synthesis of oligonucleotide arrays is a very versatile procedure for the generation of DNA arrays. In particular, photolithographically controlled synthesis combines the power of producing oligomer arrays of extremely high density and flexible patterns with a relatively simple procedure for independently directing the sequence of the molecules synthesised at the individual array positions; in addition, it facilitates large-scale chip production. However, established chemistries produced stepwise yields of some 85% only. In consequence, the total yield of a 20-mer oligonucleotide was in the range of 4%, while the majority of molecules consisted of shorter derivatives. This has quite apparent effects not only on the dynamic range but also on the discriminative power in hybridisation. Only very recently, photolithographic synthesis with quantitative yields was established [4].

3. Sample preparation

Sample preparation is another critical, often underestimated or sometimes even ignored, issue in transcriptional analysis. Anything done to the cells prior to or during RNA preparation is mirrored in the eventual analysis. This period is therefore prone to the introduction of (often unnoticed) artificial variations due to experimental action. Therefore, considerable care has to go into the design of the study as well as the actual procedures involved, and the potential risk of ‘contaminating’ effects should be assessed very carefully.

Generation of the complex target that is hybridised to the arrays is yet another process by which biases are introduced. For eukaryotic organisms, priming by oligo-dT is still the most frequently used method for reverse transcribing the RNA into DNA. However, the length of the poly-A tail varies a lot and there are genes without it, which are consequently missed altogether. Random priming would be a way out, but this requires the isolation of messenger RNA, a task that is basically impossible to perform in prokaryotes. The best solution seems to be the use of a specific primer pair for each open reading frame (ORF). Thereby, reverse transcription can occur on total RNA and differences in priming efficiency between genes should be small. For most prokaryotes, the complexity of such a primer pool is in a range similar to a random hexamer mixture and should therefore behave similarly. For larger transcriptomes (the complete set of coding sequences), the kinetic component could become critical, however. In particular, very rare transcripts could be missed entirely, although strongly variable in transcription. In addition, knowledge of the ORFs is prerequisite to this approach. Oligo-dT and random priming could produce target molecules from unknown transcripts that could subsequently be identified on arrays that contain fragments resembling the entire genome rather than the transcribed sequences only, a type of array relatively easy to obtain for prokaryotes and particularly useful for the improvement of sequence annotation. The example of yeast demonstrates that the proportion of genes missed during the initial annotation can be considerable.

4. Controls

4.1. Quantification

Transcriptional analysis relies on the presence of appropriate standards for quantification. Frequently, housekeeping genes are used for this purpose, as it is assumed that their transcript level does not vary significantly. However, variations larger than tenfold have been seen. Therefore, the introduction of artificial transcripts into the system is superior, also because they can be added early during sample handling and thus, in addition, check the various processes prior to hybridisation. In our laboratory, two sets of some 20 transcripts of that kind have each been generated and are being used in analyses (Scheideler et al., in prep.). Only by use of such numbers is the inherent statistical deviation taken care of and subsequent analyses much more significant.

4.2. Data confirmation

In most studies published to date, part of the results is verified by cross-checking with northern blot analyses. However, there is no evidence that northern blots perform in any way differently from, let alone better so than arrays except for the fact that multiple bands could indicate cross-hybridisation. In all other respects, they are based on the same principles and assumptions and therefore are prone to being similarly biased. Reverse transcription PCR [27], however, rests on a different principle and is currently unquestionably more accurate than both arrays and northern blots and should therefore be used for independent confirmation.

5. Analysis

5.1. Data quality

From all data known to date, it can be concluded that the inherent experimental variation is large (e.g. *figure* 2). Hence, at least six data points seem to be the minimum for anything close to a quantitative analysis. If only patterns are compared – useful, for example, for screening large numbers of components with the aim of identifying those that act in a similar fashion – this redundancy is not required. For investigations aimed at quantitative analysis, however, the actual data quality should be assessed thoroughly. Simply stated, the statistical deviation when dealing with an entire transcriptome makes the frequently mentioned rule of 'larger-than-twofold-variation' obsolete.

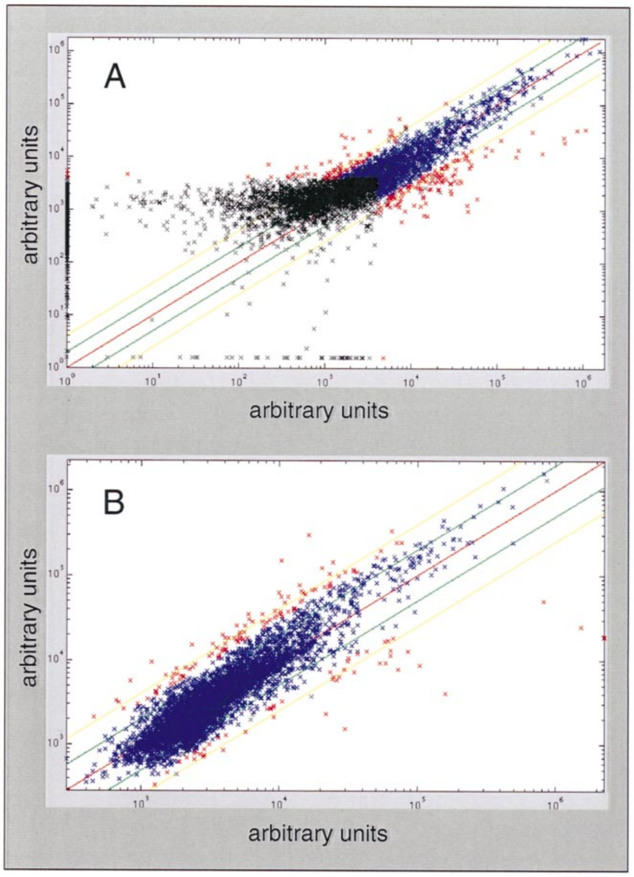

Figure 2. Correction of transcriptional analysis. Transcript levels obtained on a yeast array under two different growth conditions are compared by a scatter plot. The x:y coordinate of each gene – represented by a cross – was plotted according to its signal intensity obtained in the respective experiment. In the top panel, an irregular and artificial deviation from the diagonal distribution occurs for the weakly transcribed genes because of the lack of background correction. The bottom panel (B) shows an analysis after correction. The colour-coding indicates confidence levels (blue/black) and highlights differentially transcribed genes (red). The diagonal lines mark transcriptional changes of zero- (red), two- (green) and fourfold (yellow), respectively.

5.2. Data interpretation

While lots of raw data can be produced with array-based approaches (e.g. [6, 7]), data interpretation is an even more complex part of analysis. Even in a well-studied organism such as yeast, the amount of raw data is overwhelming – more than 1 000 conditions at Stanford University (Patrick Brown, pers. comm.) and about a tenth of this number in our database, for example – with only a very little portion of this yet integrated into the vast aggregation of knowledge about yeast biology. What is needed is the development of expert systems that automatically do at least the basic analysis procedures. As the (in relation to this problem) simple task of sequence annotation demonstrates, the establishment of such systems is not trivial and is still some time off. Without it, however, most of the results will be for database storage only.

6. Standardisation

Data comparability was until recently not addressed at all as an issue of transcriptional profiling experiments, although it is clearly of critical importance. Since many technical aspects differ between the various systems that are in use, rules and common standards are required in order to allow comparison of the various data sets. At an inaugural meeting in November 1999 in Cambridge, UK, the relevant questions were discussed (http://www.ebi.ac.uk/microarray/MGED), yielding a list of accepted general recommendations and the establishment of working groups with the task to elaborate the issues. A follow-up meeting will take place in Heidelberg in May 2000.

7. Alternative procedures

While at first glance the advantages of array-based transcriptional analysis seem to be convincing, the technology nevertheless also has disadvantages that can continue to make other approaches preferable under certain circumstances. One critical issue is sensitivity, for instance. Although detection limits will be pushed by technological developments – a very nice example being the electric field control [24] by which the target concentration is increased substantially during association and the discrimination during dissociation is controlled accurately and reproducibly – the sensitivity of PCR-based methods will most likely never be reached. Hence, technology such as representational difference analysis (RDA; *figure 3*) [13] offers a sound alternative and supplement to arrays. In addition, it has the great advantage that no prior knowledge on sequence is required for the analysis and that, in principle at

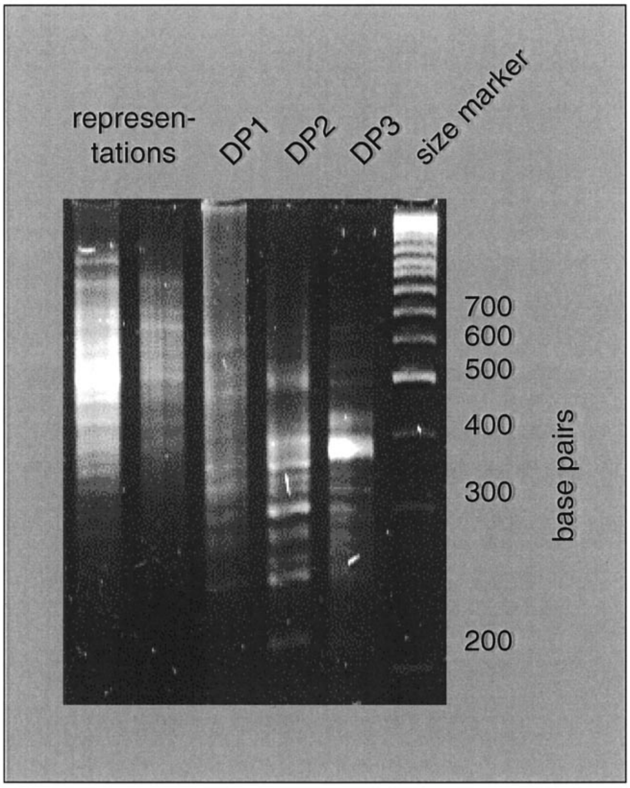

Figure 3. Representational difference analysis on kidney carcinoma tissue. A representation generated from RNA isolated from kidney carcinoma (tester) was compared to an equivalent preparation from normal kidney tissue of the same patient (driver). By three iterative cycles of difference analysis with increasing ratios of driver to tester, difference products (DP) 1 to 3 were prepared and separated on an agarose gel. The reduction in complexity is clearly visible, with eventually mainly one fragment being the dominant portion of DP3.

least, all transcribed sequences are being analysed. For current analyses in humans, for example, only part of the transcriptome is available, although a first draft sequence of the human genome will be finished soon, as will the sequence analysis of more and more organisms in the future.

Another presumed advantage, the analysis of all genes simultaneously, may be a mixed blessing. For many applications, and probably most practical uses in the future, the concentration on a well-defined set of genes could be more informative, quicker and easier to interpret. Thus, the issue of high density might not be as important as is being discussed at the moment. Again, methods such as RDA have the intrinsic advantage that they select for the differences only, hence automatically focusing on the interesting portion of an analysis.

8. Conclusions

As was outlined above, there are still many pitfalls and shortcomings in array-based transcript analyses. Therefore, data must be contemplated with care. When studied in consideration of this fact, however, and in combination with results obtained from other technical approaches, they truly are a critical and essential contribution towards the understanding of cellular biology. Nevertheless, more and very different types of information are needed for broader comprehension even at the level of nucleic acids, such as the epigenetic status and changes in conformation.

Acknowledgments

We are indebted to our collaborators for their contributions that form the basis of this manuscript, especially to Marcus Frohme, Nicole Hauser, Marcel Scheideler, Frank Diehl and Markus Beier (with respect to experimental work), and Kurt Fellenberg, Tim Beiβbart and Benedict Brors concerning bio-informatic issues. Work in our laboratories was supported by several grants from the German Federal Ministry for Education and Science (BMBF) and the European Commission.

References

[1] Augenlicht L.H., Taylor J., Anderson L., Lipkin M., Patterns of gene expression that characterise the colonic mucosa in patients at genetic risk for colonic cancer, Proc. Natl. Acad. Sci. USA 88 (1991) 3286–3289.

[2] Bains W., Smith G., A novel method for nucleic acid sequence determination, J. Theor. Biol. 135 (1988) 303–307.

[3] Beier M., Hoheisel J.D., Versatile derivatisation of solid support media for covalent bonding on DNA-microchips, Nucleic Acids Res. 27 (1999) 1970–1977.

[4] Beier M., Hoheisel J.D., Production of individually quality-checked, re-usable DNA microarrays by quantitative photolithographic synthesis, Nucleic Acids Res. 28 (2000) e11.

[5] Cantor C.R., Mirzabekov A., Southern E., Report on the sequencing by hybridisation workshop, Genomics 13 (1992) 1378–1383.

[6] Cho R.J., Campell M.J., Winzeler E.A., Steinmetz L., Conway A., Wodicka L., Wolfsberg T.G., Gabrielian A.E., Landsman D., Lockhart D.J., Davis R.W., Genome-wide transcriptional analysis of the mitotic cell cycle, Mol. Cell 2 (1998) 65–73.

[7] DeRisi J.L., Iyer V.R., Brown P.O., Exploring the metabolic, genetic control of gene expression on a genomic scale, Science 278 (1997) 680–686.

[8] Drmanac R., Labat I., Brukner I., Crkvenjakov R., Sequencing of megabase plus DNA by hybridisation: theory of the method, Genomics 4 (1989) 114–128.

[9] Du Manoir S., Speicher M.R., Joos S., Schröck E., Popp S., Döhner H., Kovacs G., Robert-Nicoud M., Lichter P., Cremer T., Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridisation, Hum. Genet. (1993) 590–610.

[10] Fodor S.P., Read J.L., Pirrung M.C., Stryer L., Lu A.T., Solas D., Light-directed spatially addressable parallel chemical synthesis, Science 251 (1991) 767–773.

[11] Gress T.M., Hoheisel J.D., Lennon G.G., Zehetner G., Lehrach H., Hybridisation fingerprinting of high density cDNA-library arrays with cDNA pools derived from whole tissues, Mamm. Genome 3 (1992) 609–619.

[12] Höög C., Isolation of large number of novel mammalian genes by a differential cDNA library screening strategy, Nucleic Acids Res. 19 (1991) 6123–6127.

[13] Hubank M., Schatz D.G., Identifying differences in mRNA expression by representational difference analysis of cDNA, Nucleic Acids Res. 22 (1994) 5640–5648.

[14] Khrapko K., Lysov Y., Khorlyn A., Shick V., Florentiev V., Mirzabekov A., An oligonucleotide hybridisation approach to DNA sequencing, FEBS Lett. 256 (1989) 118–122.

[15] Lehrach H., Drmanac R., Hoheisel J.D., Larin Z., Lennon G., Monaco A.P., Nizetic D., Zehetner G., Poustka A., Hybridisation fingerprinting in genome mapping and sequencing, in: Davies K.E., Tilghman S. (Eds.), Genome Analysis: Genetic and Physical Mapping, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1990, pp. 39–81.

[16] Lennon G.G., Lehrach H., Hybridisation analyses of arrayed cDNA libraries, Trends Genet. 7 (1991) 314–317.

[17] Lennon G., Auffray C., Polymeropoulos M., Soares M.B., The I. M. A. G. E. Consortium: an integrated molecular analysis of genomes and their expression, Genomics 33 (1996) 151–152.

[18] Liang P., Pardee A.B., Recent advances in differential display, Curr. Opin. Immunol. 7 (1995) 274–280.

[19] Lichter P., Bentz M., du Manoir S., Joos S., Comparative genomic hybridisation, in: Verma R.S., Babu A. (Eds.), Human Chromosomes: Principles and Techniques, McGraw-Hill, New York, 1995, pp. 191–210.

[20] Matson R.S., Rampal J., Pentoney S.L., Anderson P.D., Coassin P., Biopolymer synthesis on polypropylene support: oligonucleotide arrays, Anal. Biochem. 224 (1995) 110–116.

[21] Poustka A., Pohl, T., Barlow D.P., Zehetner G., Craig A., Michiels F., Ehrich E., Frischauf A.-M., Lehrach H., Molecular approaches to mammalian genetics, Cold Spring Harbor Symposia on Quant. Biol. 51 (1986) 131–139.

[22] Proudnikov D., Timofeev E., Mirzabekov A., Immobilisation of DNA in polyacrylamide gel for the manufacture of DNA and DNA-oligonucleotide microchips, Anal. Biochem. 259 (1998) 34–41.

[23] Schena M., Shalon D., Davis R.W., Brown P.O., Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.

[24] Sosnowski R.G., Tu E., Butler W.F., O'Connell J.P., Heller M.J., Rapid determination of single base mismatch mutations in DNA hybrids by direct electric field control, Proc. Natl. Acad. Sci. USA 94 (1997) 1119–1123.

[25] Southern E.M., Maskos U., Elder J.K., Analysing and comparing nucleic acid sequences by hybridisation to arrays of oligonucleotides: evaluation using experimental models, Genomics 13 (1992) 1008–1017.

[26] Velculescu V.E., Zhang L., Vogelstein B., Kinzler K., Serial analysis of gene expression, Science 270 (1995) 484–487.

[27] Veres G., Gibbs R.A., Scherer S.E., Caskey C.T., The molecular basis of the sparse fur mouse mutation, Science 237 (1987) 415–417.

[28] Zehetner G., Lehrach H., The reference library system - sharing biological material and experimental data, Nature 367 (1994) 489–491.