BMC Genomics

BioMed Central

Methodology article

Digital analysis of cDNA abundance; expression profiling by means of restriction fragment fingerprinting

Peter Hof, Claudia Ortmeier, Kirstin Pape, Birgit Reitmaier, Johannes Regenbogen, Andreas Goppelt and Joern-Peter Halle*

Address: Switch Biotech AG, Fraunhoferstr. 10, 82152 Martinsried, Germany

E-mail: Peter Hof - hof@switch-biotech.com; Claudia Ortmeier - ortmeier@switch-biotech.com; Kirstin Pape - pape@switch-biotech.com; Birgit Reitmaier - reitmaier@switch-biotech.com; Johannes Regenbogen - regenbogen@switch-biotech.com; Andreas Goppelt - goppelt@switch-biotech.com; Joern-Peter Halle* - halle@switch-biotech.com

*Corresponding author

Published: 6 March 2002

BMC Genomics 2002, **3**:7

Received: 11 October 2001
Accepted: 6 March 2002

This article is available from: http://www.biomedcentral.com/1471-2164/3/7

Abstract

Background: Gene expression profiling among different tissues is of paramount interest in various areas of biomedical research. We have developed a novel method (DADA, **D**igital **A**nalysis of c**D**NA **A**bundance), that calculates the relative abundance of genes in cDNA libraries.

Results: DADA is based upon multiple restriction fragment length analysis of pools of clones from cDNA libraries and the identification of gene-specific restriction fingerprints in the resulting complex fragment mixtures. A specific cDNA cloning vector had to be constructed that governed missing or incomplete cDNA inserts which would generate misleading fingerprints in standard cloning vectors. Double stranded cDNA was synthesized using an anchored oligo dT primer, unidirectionally inserted into the DADA vector and cDNA libraries were constructed in E. coli. The cDNA fingerprints were generated in a PCR-free procedure that allows for parallel plasmid preparation, labeling, restriction digest and fragment separation of pools of 96 colonies each. This multiplexing significantly enhanced the throughput in comparison to sequence-based methods (e.g. EST approach). The data of the fragment mixtures were integrated into a relational database system and queried with fingerprints experimentally produced by analyzing single colonies. Due to limited predictability of the position of DNA fragments on the polyacrylamid gels of a given size, fingerprints derived solely from cDNA sequences were not accurate enough to be used for the analysis. We applied DADA to the analysis of gene expression profiles in a model for impaired wound healing (treatment of mice with dexamethasone).

Conclusions: The method proved to be capable of identifying pharmacologically relevant target genes that had not been identified by other standard methods routinely used to find differentially expressed genes. Due to the above mentioned limited predictability of the fingerprints, the method was yet tested only with a limited number of experimentally determined fingerprints and was able to detect differences in gene expression of transcripts representing 0.05% of the total mRNA population (e.g. medium abundant gene transcripts).

Background

Knowing the differences in gene expression levels among different tissues is of paramount interest in various areas of biomedical research. These include, but are by no

means limited to, (i) comparisons of healthy and diseased tissue in order to understand malfunctions in regulation and identify genes essential for control, thus identifying target molecules for the development of novel therapeutics and (ii) the observation of changes in expression over time after addition of a drug to elucidate the mechanism of action of pharmaceuticals and predict their toxicology. Experiments in that realm furthermore help to understand the molecular basis of diseases, provide means for early diagnostics, and facilitate monitoring of therapy.

Examples for analog techniques representing mRNA expression patterns are subtractive cDNA libraries [1–8] and the Differential Display method and its derivatives [9–12]. Macro- and microarrays [13] are rapidly becoming the analog method of choice, mainly due to their high throughput in data production which allows complex biological questions to be addressed.

Digital methods for counting differences in gene expression levels provide specific advantages. With the expressed sequence tag (EST) approach [14,15] expression patterns are analyzed by sequencing many clones from cDNA libraries. Even limited sequence information on the cDNA 3'-end (tag sequences) permits unambiguous identification of the cDNA and the corresponding gene. The different frequencies of cDNAs in libraries derived from different sources give evidence of possible changes in gene expression. This approach provides accurate quantitative information and has a flexible degree of sensitivity which depends solely on the number of analyzed clones. However, this is very labor-intensive. In order to analyze the expression of low abundant cDNAs that represent mRNA in the range of one copy per mammalian cell, more than 100.000 colonies have to be analyzed per library. Improvements in sequencing automation and analysis such as capillary electrophoresis have speeded up the process considerably and recent developments such as a sequencing method based on real-time pyrophosphate [16], sequencing on microchips [17], and massive parallel signature sequencing on microbead arrays [18] also contribute to the speed and depth of EST gene expression analysis. Oligonucleotide fingerprinting [19] characterizes expressed genes via the hybridization of hundreds of synthetic oligonucleotides to cDNA that produces unique fingerprints of matching and non-matching oligonucleotides. Another approach to increase throughput is the serial analysis of gene expression (SAGE). Short defined cDNA sequences are initially prepared from mRNA which are then dimerized, multimerized, cloned, and sequenced [20]. SAGE accelerated the process of gene expression analysis more than an order of magnitude compared to the conventional EST analysis and still is compatible with most of the improvements of DNA sequencing mentioned above. However, after the identification of differentially expressed cDNA the full length gene has to be cloned starting from minor sequence information (12 bp tag) which hampers further functional analysis of the corresponding protein, a complex task that requires the complete coding sequence.

Each of these methods clearly has its specific advantages, and often different subsets of differentially expressed genes are identified employing a certain method. For example, cDNAs that are not easily amplified by means of PCR are inadequately represented in PCR-based methods. DADA was designed in order to overcome specific shortcomings of established differential expression analysis technologies used today, e.g. the need to sequence all of the examined genes or the involvement of PCR steps. In addition, further cloning of the complete coding sequence of the identified genes of interest is facilitated by ending up with rather long (e.g. in comparison to the SAGE method) and multiple corresponding cDNAs comprising at least part of the coding sequence. DADA is a digital method which identifies and counts the abundance of genes by means of restriction fragment fingerprinting.

Results and Discussion

I – principle of the method

DADA distinguishes and identifies genes by their specific patterns of cDNA fragments derived from digests with 6 different restriction enzymes recognizing sequences of 4 specific nucleotides. In order to avoid detection of other fragments than the desired cDNA fragments spanning from the 3', poly-A end towards the nearest upstream restriction site, a special cloning vector has been devised. It harbors a particular combination of cleavage sites that permits cloning of a nucleic acid in defined orientation and controlled labeling of only the desired fragments with fluorescent dyes (Figs. 1,2,3).

Double stranded cDNA was synthesized by a reverse transcription reaction using isolated mRNA and an anchored oligo dT primer (see material and methods). The resulting cDNA pool containing a defined poly A stretch of 16 nucleotides was cloned uni-directionally in the vector, a cDNA library was constructed in E. coli and plasmid DNA was isolated from the resulting colonies.

The part of the vector containing the cDNA fragment was labelled by means of a simultaneous restriction/ligation reaction [21] with double stranded, fluorescently labeled oligonucleotides using the Bgl I site (Fig. 1, Step 1) located adjacent to the cloning site of the cDNA representing the 3'/poly-A end of the corresponding mRNA. The labelled oligonucleotide was only ligated to the cDNA end of the BglI restriction site due to the non-palindrome nature of the overhang (see Fig. 2). In addition, the other BglI site located within the vector backbone (Bgl I*, Fig. 2) and vir-

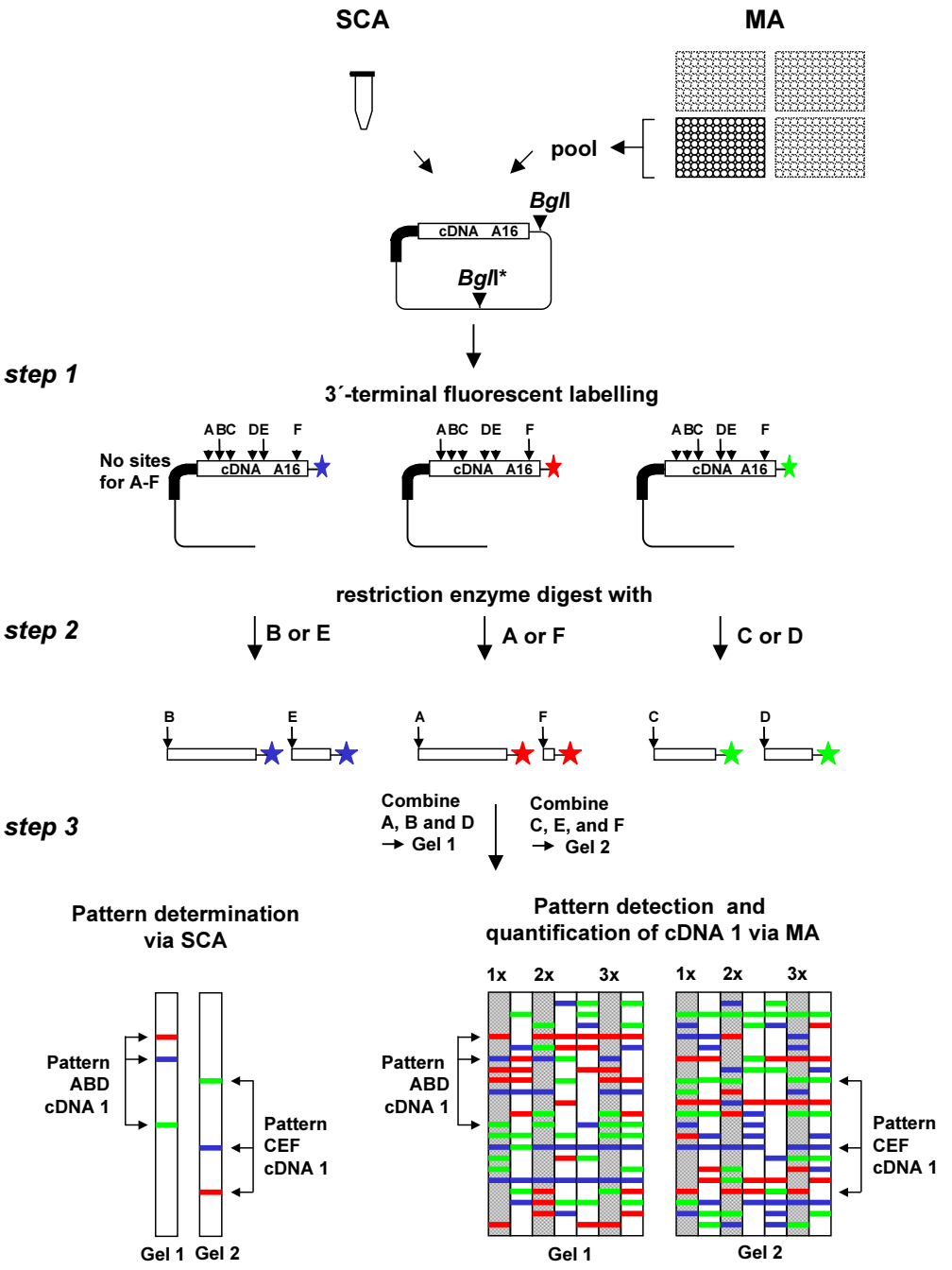

Figure 1
Experimental flow. The procedure can be performed with either a single colony from a cDNA library at a time (single clone analysis, SCA) or several colonies in parallel (96 in mixed analysis, MA). After splitting up the plasmid preparation into 6 aliquots, two at a time are labeled at a Bgl I site with one of the fluorescent dyes FAM (blue), JOE (green), or NED (yellow) – step 1. Subsequently, the 6 fractions are digested individually employing 6 different restriction enzymes recognizing sites of 4 base pairs – step 2. 3 digests are mixed together and resolved on a gel in the presence of an internal size marker labelled with the dye ROX (red) – step 3. For the digital analysis, the existence of patterns derived from a SCA is probed in the MA. In the example given, the fingerprint is identified in lanes 1,3, and 6, thus 3 out of the 6 time 96 colonies analyzed contained the cDNA of the specific gene.

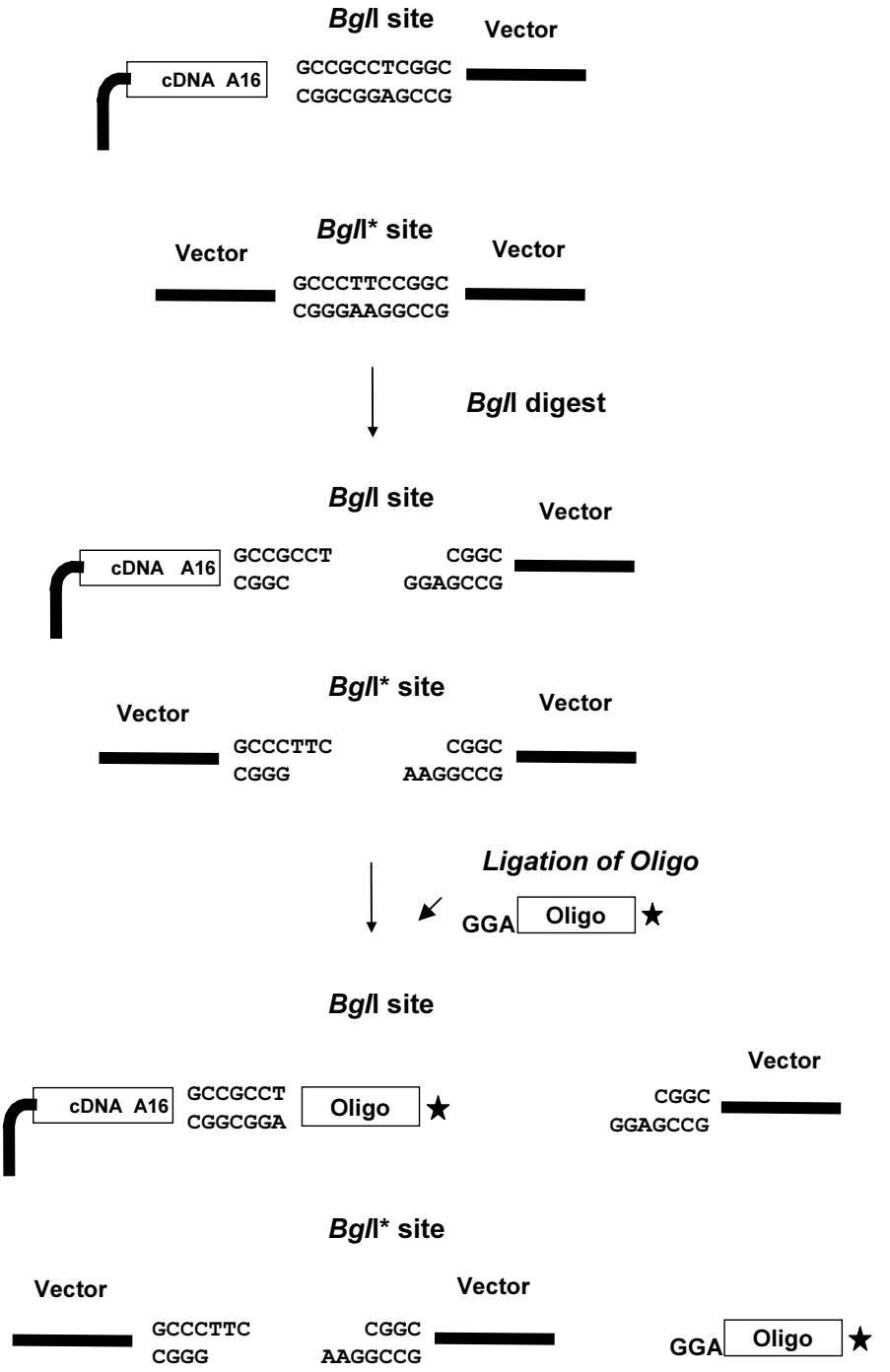

Figure 2
Close up of the one strand labelling procedure. Only those Bgl I sites that incidentally contain CCT in the arbitrary 5 N stretch inside GCC-NNNNN-GGC ligate to the fluorescently labelled oligo. Moreover, due to the in general non-palindrome nature of the 5 N stretch, only the side of the construct is labelled that contains the cDNA.

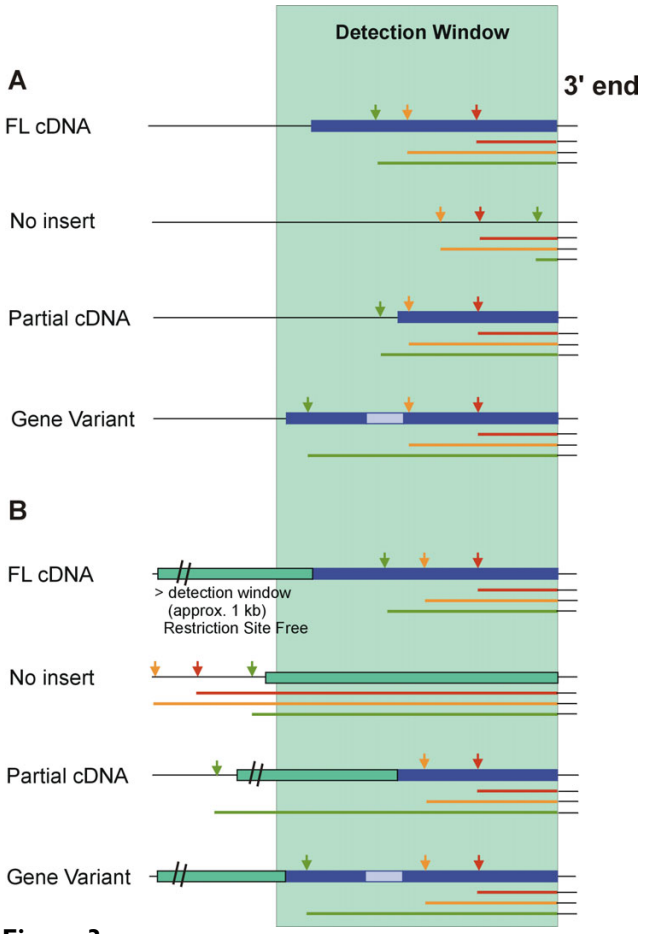

Figure 3
DADA vector. Panel A depicts the situation for conventional fragment length analysis in a standard vector. It can yield either the desired cDNA fragments or any number of wrong fragments stemming from the vector that cannot be distinguished. In Panel B the situation in the DADA vector is shown. The insertion of an approx. 1 kb long vector segment, containing none of the 4 bp cutter sites that are used for the analysis, pushes all of the undesired fragments out of the detection window. Additionally only the cDNA from the 3' end towards the gene is marked with fluorescent dye to avoid signals from vector on the 3' side. Gene variants (e.g. splice variants) can only be discriminated if the variation is located within the most 3' 900 bps of the cDNA sequence.

tually all Bgl I sites (with the exception of sites with CCT overhang) occasionally found in the inserted cDNA sequences were not labelled, as the overhangs created by the restriction digest were not compatible with the overhang of the used oligonucleotides. In comparison, the labelling of the Bgl I was less affected by artifacts resulting from labelling of internal restriction sites within the cDNA as compared to labeling of a NotI site in a different vector construct, although Bgl I sites are more often found in cDNAs compared to Not I sites (data not shown).

In a second step, the plasmids were digested with restriction enzymes recognizing 4 bp sequences (Alu I, Bfa I, Dde I Dpn I, Hinf I, or Rsa I) in six separate reactions (Fig. 1, step 2). This created labelled cDNA fragments comprising the most 3' located site of each restriction enzyme, respectively, towards a defined stretch of the poly-A tail of the cDNA, a defined part of the cloning vector and the labelled oligonucleotide. The resulting fragments were separated by polyacrylamide gel electrophoresis (ABI377 DNA analyzer) and the length of the labelled fragments were analyzed by comparison to a specially designed internal size marker (Fig. 1, step 3; see material and methods for details). The pattern of restriction fragment sizes generated by the procedure was characteristic for each individual gene transcript and represented an unique fingerprint of the corresponding gene. Partial cDNA inserts and empty vectors represent a substantial part of cDNA libraries that would hamper the analysis of restriction fingerprints in standard cloning vectors; e.g. the empty cloning vector itself would generate a characteristic restriction fingerprint that would overlay and disturb every analysis. To distinguish fragments derived from restriction digests within the cDNA inserts from fragments derived from restriction digests within the vector resulting in mixed cDNA/vector or vector only sequences, the DADA cloning vector contains an engineered segment that is free from the recognition sites used in the fragment analysis (Fig. 1) and that is longer than the detection window of 45 to approximately 900 bp (Figs. 3 and 4). This extended the restriction fingerprint of the empty vector to fragments over 900 bp which were not detectable in the analysis. In addition, the "restriction site-free" segment of the vector guaranteed that all detected fragments within the detection window stemmed from restriction sites within the cDNAs. If a specific restriction site is not contained within 900 bp of the 3' end of a cDNA, no detectable fragment was generated.

There are 2 variations of the experimental procedure for a given cDNA library (see Fig. 1 and the detailed description in Material and Methods). In (i) a single clone analysis (SCA) the procedure was applied to only one clone at a time, and yielded accurate fingerprints for certain genes experimentally. This did not speed up the process of expression analysis compared to sequencing of clones from cDNA libraries (EST approach), but generated the fingerprints used for the mixed analysis. In (ii) the mixed analyses (MA) the procedure was applied to pools of 96 cDNA clones multiplexing the plasmid preparation, labelling and analysis, thereby speeded up the process by nearly two orders of magnitude. In spite of the multiplexing, already established fingerprints of genes could be unambiguously identified in the restriction fragment mixture derived from the pools (Fig. 1 and 4). If the characteristic fingerprint of a certain gene was identified in the fragment

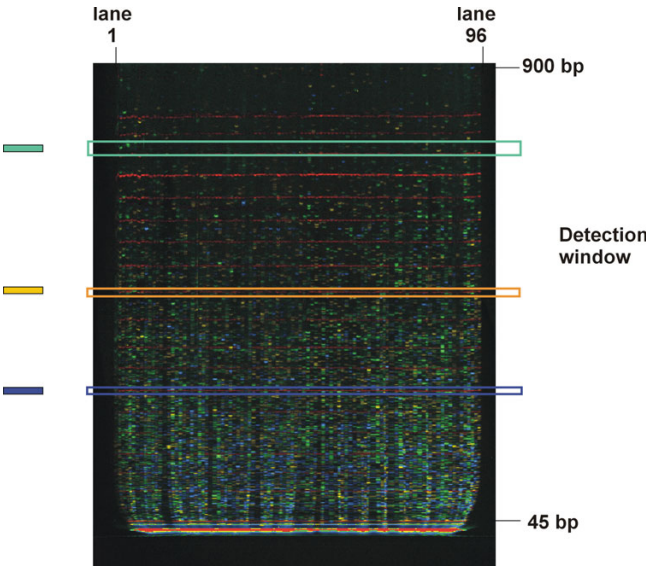

Figure 4
Visualizing the digital counting of patterns (and thus genes) in cDNA libraries. After a pattern for a given gene is determined (eg. via SCA, described in the method section), every lane from the mixed analysis to be examined receives a count that contains all 3 fragments of a fingerprint at the same time (actually all 6 from 2 lanes of the 2 corresponding gels). The width of the search windows indicate that the search algorithm tolerates small deviations in fragment lengths. Due to the higher variations in the analysis of long fragments, the width of the search window is adjusted accordingly.

mixture, at least one of the 96 colonies of the pool included a vector with the corresponding cDNA. Therefore, the frequency of occurrence of a fingerprint enabled us to count the abundance of genes and to calculate expression levels from cDNA libraries derived from different statuses at a high throughput. The high throughput only applied to the analysis of the abundance of fingerprints, as in contrast to the SCA, no new fingerprints could be generated by the mixed analysis.

2 – obtaining patterns from a single clone analysis

We randomly picked single clones of 96 colonies from a murine wounded skin library and compared the analyzed restriction fingerprints with the corresponding, experimentally determined cDNA sequence of the vectors. Vectors containing no cDNA inserts produced no signals in the detection window of 45 to 900 base pairs and incomplete or short cDNA inserts yielded only signals of fragments derived from restrictions within the cDNA sequence. This demonstrated that the restriction free region next to the cDNA cloning site fulfills its function in repressing misleading fragments and the corresponding signals.

A thorough comparison of the analyzed fragment lengths with the corresponding sequences of the clones resulted in a good agreement of theoretical and experimental fragment length for the 6 restriction enzymes. In general, only those fragments were detected, that were derived from the nearest restriction site of the corresponding enzyme as judged by sequencing. However, in several cases the first Alu I site was not recognized by the enzyme despite the use of an at least 4-fold excess of enzyme units compared to the amount of cDNA, indicating site preferences of this restriction enzyme that may depend on the context of the recognition sequence [22]. This did not affect the analysis of fingerprints in mixed clone analyses reported below as the same site preference was also observed under these conditions.

Another inherent discrepancy is the integer character of the predicted values versus the non-integer values stemming from the comparison to the size marker. The average of the difference of the experimentally analyzed length to the calculated length was 0.258 bp (median 0.115 bp) which means that the analyzed values were generally a little longer than the predicted ones. About 95% of the analyzed lengths differed less than 1 bp and about 50% differed less than 0.5 bp from the predicted lengths which by far exceeded the experimental deviations in repeated experiments (standard deviation of less than 0.2 bp, see below) and is in accordance to results from other authors [23,24]. These data demonstrated, that there is a reproducible but not yet predictable shift of the fragments within the polyacrylamide gel run. Together with the above mentioned site preference of AluI and additional problems like the not well defined polyadenylation-site in public database entries, this hinders the construction of fingerprints solely based on sequence data. Attempts to predict these shifts based on the sequence composition of the fragments were not successful in the first trials.

Experimentally determined fingerprints from single clone analysis could be used for the analysis without any of the above mentioned hindrances. The fragment lengths from single clone analyses that by chance contained cDNAs from the same gene reproduced very accurately. The standard deviation from the average value was below 0.2 bp for fragments below 500 bp and increased with larger fragment size. The range in which a band was accepted as being the same as the search value was determined empirically. Colonies of specific genes were mixed into different pools of defined clone compositions and the detection width was adjusted to be the best compromise between not missing any real hits and counting false ones. As assumed from the increasing peak width of the analyzed fragments and the internal size marker, the detection width varied with the length of the fragments (see material and methods). The detection range was increased above

Table 1: Results of digital analyses and independent verification of DADA quantification. For two genes in each scenario the DADA counts are given and compared to quantitative RT-PCR (DEX-PBS), respectively. The values from the RT-PCR are normalized to GAPDH.

	MCP-2		Cystatin C	
	DEX	**PBS**	**DEX**	**PBS**
DADA	**3**	**0**	**4**	**0**
Q-RT-PCR	**0,48**	**0,09**	**2,18**	**0,76**

the determined 0.2 bp standard deviation especially for the longer fragments. The broader detection width that could in principle lead to false positive identification of a fingerprint was counterbalanced by the fact that the peak density is much lower towards the longer fragments (Fig. 4).

3 – proof of concept

The differences in gene expression following treatment of a tissue with an active compound was monitored using two libraries of wound tissue derived from control (phosphate buffer saline, PBS, normally healing) and dexamethasone (DEX, badly healing) treated mice. Steroid treatment is widely used as a model for impaired wound healing [25]. From each library about 37.000 colonies were analyzed in pools of 96 colonies each. The complete analysis of 74.000 colonies required only 16 runs on an ABI377 DNA analyzer (96 lanes with 96 colonies each, 2 gels of 3 enzymes each for the analysis of 6 different enzymes). All calculated fragment sizes were analyzed and stored in a relational database, representing nearly 500.000 data points from this analysis alone.

Applying the restriction fingerprints of the above mentioned single clone analysis, we found several genes that were up-regulated in wounds of dexamethasone treated animals. Two genes are shown that were confirmed with quantitative Real Time RT-PCR and represent intriguing examples (Table 1). MCP-2 and Cystatin C cDNAs contain all six restriction sites within 500 bp from the poly-A tail (experimental fragment lengths including poly-A tail and labelling oligonucleotides: MCP-2/Cystatin C: Dde I: 196.23/276.11 bp; Alu I: 67.51/118.08 bp; Hinf I: 349.88/274.51 bp; Bfa I: 531.02/270.83 bp; Dpn I: 245.02/265.17 bp; Rsa I: 254.23/89.18 bp, respectively). These fingerprints are unambiguously related to their cDNA as a search in the complete GenBank database (see material and methods) showed only the corresponding murine cDNA hits even though the search parameters were loosened to account for the above mentioned differences in experimentally derived fragment lengths to the real sequence.

Finally, these two fingerprints were used for a digital analysis of the DEX and PBS libraries. The fingerprints of all six enzymes were used for the search of the database from the 74.000 colonies. Each of the 768 subsets of the data representing a pool of 96 colonies was analyzed by a computer program for the concurrent appearance of the 6 fragments of a specific fingerprint. The search for the appearance of the fingerprints from MCP-2 and Cystatin C resulted in 3 and 4 counts, respectively. Interestingly, the fingerprints were only found in datasets/pools from the cDNA library derived from DEX treated animals and not in the library from control animals. The concluded induction of both transcripts in wounds from badly healing animals compared well to quantitative RT-PCR data (Table 1).

In order to explore the limitation of the method we also computed all possible restricted fingerprints of MCP-2 and Cystatin C containing only 5 different restriction enzymes. Here we obtained the same results when we used the five shortest fragments of the fingerprints, but misleading results, if we left out for example the restriction enzyme that produced the shortest fragment. This observation can be easily explained, as partial cDNA sequences derived from oligo dT synthesis can lack the most 5' restriction site and include the most 3' restriction site, but never vice versa. If the numbers of fragments were reduced to the 4 shortest ones, the results were perturbed by unspecific fingerprints that were not derived from the specific cDNAs but were erratically composed of fragments derived from different cDNAs within the pool of 96 colonies. More of those comparisons started to reveal significance limits such as the number of fragments that should be contained in a fingerprint. From the limited amount of comparisons it would seem that having 6 different fragments in a pattern and finding all of them works reliably even with low abundant genes like MCP-2 and Cystatin C that were induced to medium abundance by the treatment (3 to 4 transcripts within the 72.000 analyzed transcripts or about 0.05%).. Finding only 3 to 4 fragments is sufficient only for high abundant genes with a significant differential expression. E.g., serum albumin and SPR1a with abundances of about 0.5% in liver and skin libraries, respectively [26,27], could be identified and quantified by using only 3 restriction enzymes in the corresponding cDNA libraries [28]

The clones of MCP-2 and Cystatin C were examined more closely regarding their length. They contained the complete cDNA sequence. This is not the case for all other genes, but the clones often contain substantial parts of the coding regions and they almost always prolong the public

database cDNA entries toward the 3' end. This supports the notion that DADA will also facilitate a rapid full length cloning effort after identification of interesting genes.

Conclusion

In conclusion, DADA is a digital method for analyzing expression levels based on the counting of restriction fingerprints of cDNA clones. It is built around a specifically designed cloning vector that suppresses misleading fingerprints derived from partial cDNAs or empty vectors. The method yields quantitative data and absolute figures that do not depend on amplification by polymerase chain reaction. Compared to the sequencing of cDNA libraries (conventional EST approach [14]), it is substantially faster and more economical. In comparison to the SAGE method [20] it results in physical cDNA clones of substantial length which speed up further analysis.

Within the limited number of examples, there is a good correlation between the clear results from DADA and other gene expression analysis methods such as quantitative real time RT-PCR, RNase Protection Assays, and various hybridization procedures (data not shown). A number of additional genes with differential regulation discovered by DADA and confirmed by independent methods are now under investigations for the use as therapeutical targets (unpublished observations). Some of the genes identified by DADA including MCP-2 and Cystatin C were neither discovered by means of subtractive hybridization [3], nor by differential display [9], although these methods were successfully used in comparable settings in our laboratory and lead to a high number of differentially expressed genes. Vice versa, DADA could not detect certain differentially expressed genes, that were discovered by standard methods. As an example, the injury-induced differential expression of S100A9 could be detected by subtractive hybridization [33]. The DADA fingerprint of S100A9 was detected in both analyzed cDNA libraries. However, as it occurs only once in each library, no statistical significant differential expression could be identified by the DADA method. Only a subset of genes was identified by all methods. In general, the different approaches can be seen as complementary rather than competitive. The specific findings will certainly vary with changes in the completeness and individual execution of the respective screens. Additionally, DADA offers advantages with regard to absolute and quantitative data at the level of screening.

As of yet we can only perform digital analyses with patterns experimentally analyzed from single clone analysis. These fragment lengths reproduce accurately enough to search and count them in mixed analyses applying rigid queries. The usage of restriction fingerprints computationally derived from sequence database entries is hampered by the lack of predictability of gel positions from sequences and the integer versus non-integer issue. The experimental fingerprints, due to the narrower search space, allowed for the identification of fingerprints in mixtures of 96 colonies as demonstrated here. The broader search space of sequence-based predictions allow only for the identification of fingerprints in mixtures of about 10 colonies (data not shown). The one order of magnitude lower throughput in the case of predicted fingerprints would render the method non-superior to others.

The generation of experimental fingerprints is the bottleneck of the procedure as described here. Once the fingerprints are generated for a certain species they can be rapidly applied in all possible settings. Therefore, we are developing improvements of this step. In addition, the use of a now available fifth fluorescent dye as a size marker in sequencing lanes should allow for a fast correlation between restriction fragment sizes (derived from the sequence) and gel run behaviors of DNA fragments (relative to the size marker). The latter improvement could not only lead to a faster generation of experimental fingerprints, but could also supply a large dataset to improve the prediction of gel run behaviors of fragments based on the sequence composition.

Another improvement could result from alternative DNA separation and detection methods such as capillary electrophoresis or mass spectrometry. This could further improve the reproducibility of the fragment analysis and lead to a reduction of the search space for every fragment. As discussed above, this parameter is directly linked to the throughput of the method. In addition, DADA would profit from the higher throughput of capillary electrophoresis compared to the gel electrophoresis used in this study. On the other hand, values created by mass spectrometry are precisely predictable from the sequence of DNA fragments [29]. This would allow for the generation of fingerprints from sequence data in silico and speed up this limiting step of the procedure.

Materials and Methods

Cloning vectors

The vector pUC19 (Yanisch-Perron et al., 1985) was cut with Hind III and Aat II and the 2170 bp fragment containing the β-lactamase gene and the ColE1 replication origin was isolated, and two different synthetic double stranded oligonucleotides were inserted (order of restriction sites Hind III-Asc I-EcoR I-Xho I-Sfi I/Bgl I-T7promoter-Aat II). To generate a restriction-free region at the 5' side of the cDNA cloning site (EcoR I-Xho I) a 860 bp long PCR fragment of human genomic DNA (primers: CCCCAAGCTTGAGTATGAACAAATTTACTTTCTTCTTTC and CCGGCGCGCCTCCTAAAGTGCTGGATTATAG) de-

void of Alu I, Dpn I, Dde I, Hinf I and Rsa I was inserted between the Hind III and Asc I site of the vectors. The BfaI site originally located in the PCR fragment from genomic DNA was deleted by site directed mutagenesis (details available upon request).

cDNA library construction

mRNA was extracted from murine tissues (BALB/c mice: Id wound of controls, 1d wound of mice treated with 0.5 mg Dexamethason per kg bodyweight twice a day for 5 days before wounding) according to standard procedures [30]. Methylated cDNA was synthesized from an anchored Xho I-oligo-dT primer ($(GA)_{10}$ACTAGTCTCGAGT$_{16}$VN) that secured a defined start of the cDNA synthesis at the cDNA-polyA border using the Stratagene cDNA synthesis kit according to the instructions of the supplier. EcoR I adapter was added, the cDNA was digested by Xho I and the cDNA was inserted between the EcoR I and Xho I sites of the vectors in a directional fashion. cDNA libraries were constructed by electroporation into E. coli SURE and single colonies were picked either manually or by robots (Q-Pix, GeneScreen) into 96 or 384 microtiter plates, respectively.

DNA preparation

For the single clone analysis, colonies were grown in 96 deep well plates. For the mixed clone analysis, 9216 single colonies were inoculated in 24 384- microtiterplates in 50 µl of TYGPN-medium [31] and grown for 48 hours at 37°C. The cultured medium of 96 colonies each was collected by means of a BIOMEK pipetting robot (Beckman) and pooled in one well of a 96 deep well plate. DNA was prepared according to the REAL DNA preparation kit (Qiagen) and dissolved at a concentration of about 100 ng/µl in H_2O.

Labelling of cDNA

Double stranded, fluorescently labelled (FAM, JOE, or NED) oligonucleotides (5' labelled oligonucleotide: CAGGAGATGCTGTTCGTAGG, unlabelled oligonucleotide: ACGAACAGCATCTCCT, supplier: Applied Biosystems) were annealed in 10 mM Tris pH 8.0, 10 mM NaCl, and 1 mM EDTA (3 min. at 94°C, cooling to 20°C within 15 minutes). Two times three labelling reactions were prepared in 6 different 96-microtiterplates (2×FAM, 2×JOE, and 2×NED, respectively). 500 ng (FAM labelling) or 1 µg (JOE and NED labelling) of the cDNA plasmid preparation was mixed with 0.5 pmol labelled oligonucleotides, 0.25 (FAM) or 0.5 (JOE and NED) Units Bgl I, and 10 (FAM) or 20 (JOE and NED) Units T4-DNA-ligase (New England Biolabs), respectively, in 20 mM Tris-acetate (pH 7.9), 10 mM magnesium acetate, 50 mM potassium acetate, 1 mM DTT, 100 µg/ml BSA, 1 mM ATP. The reactions were incubated over night at 37°C and stopped by heating to 65°C for 10 minutes. The restriction enzymes for the second digest (0.5 U Bfa I, 1.0 U Dde I, 1.5 U Dpn I, 2.0 U Alu I, 1.0 U Rsa I, or 2.0 U Hinf I) were diluted in 10 µl buffer (20 mM Tris-acetate (pH 7.9), 10 mM magnesium acetate, 50 mM potassium acetate, 1 mM DTT, 100 µg/ml BSA) and added separately to the six rections (FAM: Bfa I and Dde I, JOE: Dpn I and Alu I, NED: Rsa I and Hinf I). After 2 hours at 37°C, the reactions were stopped by heating to 80°C for 20 minutes and the reactions Bfa I, Dpn I and Rsa I as well as Dde I, Alu I and Hinf I were pooled, respectively. The reaction products were purified by means of three repeated gel chromatographies using water saturated Sephadex G-50 in Millipore Multiscreen® filtration plates according to the instructions provided by the supplier and dried under vacuum.

Fragment analysis

The dried DNA was dissolved in 2 µl of loading buffer (7 M Urea, 5 mM EDTA, 0.5% Blue Dextrane, pH 8.0) containing a ROX-labelled size marker derived from specifically designed λ-phage PCR fragments (45 bp, 80 bp, 120 bp, 160 bp, 200 bp, 240 bp, 280 bp, 320 bp, 360 bp, 400 bp, 450 bp, 500 bp, 550 bp, 600 bp, 650 bp, 700 bp, 750 bp, 800 bp, 850 bp, 900 bp). The fragments were denaturated at 90°C for 6 minutes, loaded onto a 36 cm long, 5% (29:1) polyacrylamid, 7 M urea gel, and separated at 2000 V, 50 mA, and 51°C in a 96 lane ABI377 DNA analyzer. The fragment sizes of the FAM, JOE, and NED labeled fragments between 45 and 900 bp were calculated from the raw data relative to the ROX size marker by the GeneScan program (Applied Biosystems) and exported as text files. The data pertaining to the peaks (lane number, height, area, colour, calculated length of fragments) were imported into a relational database scheme (Oracle 8.01 relational database management system) and integrated with the information on the cDNA library, the location of the colonies on the microtiterplates, the colony composition of the DNA preparation, the labelling reaction, and the gel conditions.

Analysis of restriction fingerprints derived from single clones – single clone analysis (SCA) A computer program (written with the Microsoft Visual Studio 6 software suite) was designed that extracted fingerprints from the raw data of single clone analysis. In one lane, the peak for each restriction enzyme was chosen, that had the maximum height (NewHeight) after applying the following empirically determined formula in order to account for the broadening in peak shape and the resulting lower peak height that goes along with increasing numbers of base pairs: NewHeight = MeasuredHeight / ((CalculatedBasePairs / - Denominator) + Addend), with Denominator = 3.0, Addend = 500. The corresponding calculated lengths of the fragments were stored as fingerprints in the database.

Calculation of the abundance of fingerprints in mixed clone analysis (MA)

A second computer program was designed that analyzed the datasets from the mixed analysis for the frequency of occurrence of the stored fingerprints. Each dataset derived from a pool of 96 colonies (6 complex fragment mixtures in 2 corresponding lanes on two corresponding gels) was queried for the concurrent occurrence of the peaks / fragments of each fingerprint, respectively (see Fig. 4). As an example, for a certain cDNA fingerprint the the peak from SCA for restriction enzyme DdeI was determined at 326,5 bp. In one lane in a MA there are multiple peaks stemming from DdeI digestion (50.3, 67.4, 98.5,.....326.7....500.1...). The computer program checked if the peak for Dde I was contained in the complex peak mixture derived from the Dde I digest, considering a certain bandwidth. In the example above, 326.7 bp in the MA would be considered as a hit. The computer program did the same for the other 5 restriction enzymes. When all peaks are simultaneously present in one dataset from a pool of colonies the counter for this fingerprint/gene goes up by one. It is possible to adjust the program to query the data with only a subset of the 6 restriction enzymes of the fingerprints.

The empirical determined allowed bandwidth for a peak to be counted was set to the following parameters fragments from 40 to 100 bp: bandwidth +-0.5 bp, 100–300 bp: +-0.3, 300–500 bp: +-0.7, 500–700 bp:+-2.0, 700–900 bp:+-5.0).

Pattern search in Database Sequences

Search strings of the form (eg. fingerprint MCP-2, Bfa I, Hinf I, Rsa I, Dpn I, Dde I, Alu I) $CTAGN_{177-181}GANTCN_{90-94}GTACN_{3-7}GATCN_{41-45}CTNAGN_{123-127}AGCT$ were derived from the restriction fingerprints. The distances between the restriction sites were set to be flexible (+/-2 bp of the empirical length), because the experimentally analyzed fragment lengths have non-integer numbers and do not correspond to sequence length in base pairs better than +/-1 bp. The strings served as an input to the Perl script prosite_scan (author: Kay Hofmann [http://www.isrec.isb-sib.ch/ftp-server/prosite_scan/pattern_find]) searching the nucleotide section of GenBank (downloaded April 12th 2000).

RNA Preparation and reverse transcription

Total RNA was extracted from the frozen mouse skin biopsies of 30 animals as described by Chomczynski and Sacchi (1987) [32]. The total RNA was digested by RNase free DNaseI and phenol extracted in order to remove genomic DNA contaminations. First strand cDNA was synthesized from 1 μg of total RNA in a 100 μl reaction volume using random hexamers as primer and Multiscribe reverse transcriptase (TaqMan Reverse Transcription Reagents Kit, PE Biosystems).

Real-time polymerase chain reaction (PCR)

Specific primers were designed using the Primer-Express program (PE Biosystems) and synthesized by Interactiva (Ulm, Germany). The sequences and necessary concentrations in the PCR reaction are outlined in Table 2. The PCR mixture included 0.625 U AmpliTaq Gold DNA polymerase in the 2× SYBR Green Master Mix (PE Biosystems), the required concentration of specific forward and reverse primers, 10 ng of cDNA template and 0.25 U AmpErase UNG (PE Biosystems) in a 25 μl reaction volume. The quantification relative to the house keeping gene GAPDH was carried out in MicroAmp Optical 96-well reaction plates (PE Biosystems). On each plate a standard curve was generated for both the GAPDH and target PCR reactions by amplifying 5 different known amounts of cDNA derived from total RNA. For each cDNA sample under investigation triplicate reaction wells were set up for both GAPDH and target amplification. The amplification was carried out and analysed in the GeneAmp 5700 Sequence Detection System (PE Biosystems). The efficiency of each PCR was calculated from the slope of the standard curve ($E = 10^{(-1/s)} - 1$). The abundance of the target relative to GAPDH was calculated as: $X_n = (1 + E_{GAPDH})^{Ct,gapdh}/(1 + E_{target})^{Ct,target}$, where C_t is the threshold cycle determined from the amplification curves, and the relative abundances from one quantification were set into relation with one another.

Table 2: The primer sequences and necessary concentrations for the quantification via real-time PCR.

product	forward primer	conc.	reverse primer	conc.
GAPDH	ATCAACGGGAAGCCCATCA	100 nM	GACATACTCAGCACCGGCCT	100 nM
MCP-2	CTTCTCTGGGCTGACAGGGA	300 nM	TCTACGCAGTGCTTCTTTGCC	300 nM
cystatin C	CAAGAAGAGTGGAGCCAGGG	50 nM	GCAGGCAGGTTCTGCACAT	50 nM

List of abbreviations

cDNA – DNA complementary to ribonucleic acid (RNA), DNA – deoxyribonucleic acid, PBS – phosphate buffer saline, DEX – dexamethasone, EST – expressed sequence tag, SAGE – serial analysis of gene expression, SCA – single clone analysis, MA – mixed analysis, bp – base pairs, StDev – Standard Deviation, RT-PCR – real-time polymerase chain reaction. poly-A – polyadenylation, GAPDH – Glyceraldehyde 3-phosphate dehydrogenase

Acknowledgments

The authors thank Gabriele Stumpf, Jens Herold, Eckhard Wolf, Sabine Werner and Horst Domdey for discussions, and all people at SWITCH Biotech for their support. This work was supported by a grant from the Bayerisches Staatsministerium fur Wirtschaft, Verkehr und Technologie (BayTOU Nr. 07 03/683 64/870/99/733/2000/734/2001/735/2002).

References

1. Akopian AN, Wood JN: **Peripheral nervous system-specific genes identified by subtractive cDNA cloning.** *J Biol Chem* 1995, **270:**21264-70
2. Deleersnijder W, Hong G, Cortvrindt R, Poirier C, Tylzanowski P, Pittois K, Van Marck E, Merregaert J: **Isolation of markers for chondro-osteogenic differentiation using cDNA library subtraction. Molecular cloning and characterization of a gene belonging to a novel multigene family of integral membrane proteins.** *J Biol Chem* 1996, **271:**19475-82
3. Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, Siebert PD: **Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries.** *Proc Natl Acad Sci USA* 1996, **93:**6025-30
4. Gurskaya NG, Diatchenko L, Chenchik A, Siebert PD, Khaspekov GL, Lukyanov KA, Vagner LL, Ermolaeva OD, Lukyanov SA, Sverdlov ED: **Equalizing cDNA subtraction based on selective suppression of polymerase chain reaction: cloning of Jurkat cell transcripts induced by phytohemaglutinin and phorbol 12-myristate 13-acetate.** *Anal Biochem* 1996, **240:**90-7
5. Hubank M, Schatz DG: **Identifying differences in mRNA expression by representational difference analysis of cDNA.** *Nucleic Acids Res* 1994, **22:**5640-8
6. Lisitsyn N, Wigler M: **Cloning the differences between two complex genomes.** *Science* 1993, **259:**946-51
7. Yang M, Sytkowski AJ: **Cloning differentially expressed genes by linker capture subtraction.** *Anal Biochem* 1996, **237:**109-14
8. Zeng J, Gorski RA, Hamer D: **Differential cDNA cloning by enzymatic degrading subtraction (EDS).** *Nucleic Acids Res* 1994, **22:**4381-5
9. Liang P, Averboukh L, Keyomarsi K, Sager R, Pardee AB: **Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelial cells.** *Cancer Res* 1992, **52:**6966-8
10. Liang P, Pardee AB: **Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction [see comments].** *Science* 1992, **257:**967-71
11. Prashar Y, Weissman SM: **Analysis of differential gene expression by display of 3' end restriction fragments of cDNAs.** *Proc Natl Acad Sci USA* 1996, **93:**659-63
12. Sutcliffe JG, Foye PE, Erlander MG, Hilbush BS, Bodzin LJ, Durham JT, Hasel KW: **TOGA: An automated parsing technology for analyzing expression of nearly all genes.** *Proc Natl Acad Sci USA* 2000, **97:**1976-1981
13. Cohen B, ed: *Nature Genetics 21, Supplement* 1999
14. Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C, Venter JC: **Sequence identification of 2,375 human brain genes [see comments].** *Nature* 1992, **355:**632-4
15. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al*: **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 1991, **252:**1651-6
16. Ronaghi M, Uhlen M, Nyren P: **A sequencing method based on real-time pyrophosphate.** *Science* 1998, **281:**363-365
17. Liu S, Ren H, Gao Q, Roach DJ, Loder RT Jr, Armstrong TM, Mao Q, Blaga I, Barker DL, Jovanovich SB: **Automated parallel DNA sequencing on multiple channel microchips.** *Proc Natl Acad Sci USA* 2000, **97:**5369-74
18. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, T B, Pallas M, RB D, Kirchner J, Fearon K, Mao J, Corcoran K: **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.** *Nature Biotechnology* 2000, **18:**630-40
19. Meier-Ewert S, Lange J, Gerst H, Herwig R, Schmitt A, Freund J, Mott R, Herrmann B, Lehrach H: **Comparative gene expression profiling by oligonucleotide fingerprinting.** *Nucleic Acids Res* 1998, **26:**2216-23
20. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270:**484-7
21. Carrano AV, Lamerdin J, Ashworth LK, Watkins B, Branscomb E, Slezak T, Raff M, de Jong PJ, Keith D, McBride L, *et al*: **A high-resolution, fluorescence-based, semiautomated method for DNA fingerprinting.** *Genomics* 1989, **4:**129-36
22. Gingeras TR, Brooks JE: **Cloned restriction/modification system from Pseudomonas aeruginosa.** *Proc Natl Acad Sci USA* 1983, **80:**402-6
23. Bowling JM, Bruner KL, Cmarik JL, Tibbetts C: **Neighboring nucleotide interactions during DNA sequencing gel electrophoresis.** *Nucleic Acids Res* 1991, **19:**3089-97
24. Frank R, Koster H: **DNA chain length markers and the influence of base composition on electrophoretic mobility of oligodeoxyribonucleotides in polyacrylamide gels.** *Nucleic Acids Res* 1979, **6:**2069-87
25. Beer HD, Longaker MT, Werner S: **Reduced expression of PDGF and PDGF receptors during impaired wound healing.** *J Invest Dermatol* 1997, **109:**132-8
26. Kartasova T, Darwiche N, Kohno Y, Koizumi H, Osada S, Huh N, Lichti U, Steinert PM, Kuroki T: **Sequence and expression patterns of mouse SPR1: Correlation of expression with epithelial function.** *J Invest Dermatol* 1996, **106:**294-304
27. Sellem CH, Frain M, Erdos T, Sala-Trepat JM: **Differential expression of albumin and alpha-fetoprotein genes in fetal tissues of mouse and rat.** *Dev Biol* 1984, **102:**51-60
28. Halle J-P, Regenbogen J, Goppelt A: **Cloning vector, its production and use for the analysis of mRNA expression pattern.** *European Patent Application EP0965642;* 1999
29. Koster H, Tang K, Fu DJ, Braun A, van den Boom D, Smith CL, Cotter RJ, Cantor CR: **A strategy for rapid and efficient DNA sequencing by mass spectrometry [see comments].** *Nature Biotechnology* 1996, **14:**1123-8
30. Chomczynski P, Mackey K: **Substitution of chloroform by bromo-chloropropane in the single-step method of RNA isolation.** *Anal Biochem* 1995, **225:**163-4
31. Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K, eds: **Current Protocols in Molecular Biology.** *Massachusetts General Hospital, Harvard Medical School, Boston, MA: John Wiley & Sons, Inc.;* 1999
32. Chomczynski P, Sacchi N: **Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform *extraction.*** *Anal Biochem* 1987, **162:**156-9
33. Thorey IS, Roth J, Regenbogen J, Halle JP, Bittner M, Vogi T, Kaesler S, Bugnon P, Reitmaier B, Durka S, Graf A, Wockner M, Rieger N, Konstantinow A, Wolf E, Goppelt A, Werner S: **The Ca2+-binding proteins S100A8 and S100A9 are encoded by novel injury-resulated senes.** *J Biol Chem* 2001, **38:**35818-25