Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Biotechnology

Bioinformatics applications for pathway analysis of microarray data

Thomas Werner

Changes in transcript levels are assessed by microarray analysis on an individual basis, essentially resulting in long lists of genes that were found to have significantly changed transcript levels. However, in biology these changes do not occur as independent events as such lists suggest, but in a highly coordinated and interdependent manner. Understanding the biological meaning of the observed changes requires elucidating such biological interdependencies. The most common way to achieve this is to project the gene lists onto distinct biological processes often represented in the form of gene-ontology (GO) categories or metabolic and regulatory pathways as derived from literature analysis. This review focuses on different approaches and tools employed for this task, starting form GO-ranking methods, covering pathway mappings, finally converging on biological network analysis. A brief outlook of the application of such approaches to the newest microarray-based technologies (Chromatin-ImmunoPrecipitation, ChIP-on-chip) concludes the review.

Addresses
Genomatix Software GmbH, Bayerstr. 85A, D-80335 München, Germany

Corresponding author: Werner, Thomas (Werner@genomatix.de)

Current Opinion in Biotechnology 2008, **19**:50–54

This review comes from a themed issue on
Analytical biotechnology
Edited by Thomas Joos and Paul E. Kroeger

Available online 22nd January 2008

0958-1669/$ – see front matter

DOI 10.1016/j.copbio.2007.11.005

Microarrays made it possible to survey changes in the mRNA levels of genes on a genome-wide scale in a single experiment, promising an unbiased overview over changes in the transcriptome. Quickly new or adapted methods for mastering the statistical part of microarray analysis were introduced including the still very popular significance analysis of microarrays (SAMs) [1]. However, after the initial enthusiasm had subsided, it became quite clear that even the statistically best-supported lists of up-regulated and down-regulated genes were most of the time as cryptic as the primary nucleotide sequence of the genome. There are two reasons for this: first of all many (if not almost all) genes serve multiple context-dependent functions. On the contrary, not all changes in mRNA levels are directly connected to the experiment conducted. Therefore, it is no surprise that soon after 2001 the necessity to go beyond simple clustering and statistics has been recognized [2].

Gene ontology

The first thing that comes to mind when trying to bring genes from a list into some biologically meaningful context is gene ontology (GO), an expert-curated database assigning genes to various functional categories. Although this currently covers only 17 348 of the 38 675 'genes' annotated in Entrez Gene, it is a great tool to see which of the genes in a list belong together in terms of one of the GO branches: biological process, molecular function, and cellular component. *p*-Values can be used to rank the GO categories in relationship to the genes found as significantly regulated on the micorarray.

The DAVID program is a widely used web-based application [3] focusing on GO classification. The same year another web-based application called Onto-Express was published by a group from Wayne State University in Detroit [4], which gradually developed into a host of applications in the meantime (see Table 1). The idea is taken one step further by Hvisten *et al.* going beyond GO classification and employing expression data themselves into what they call 'minimal decision rules' [5]. The basic idea is that genes with a similar expression profile are more likely to be part of one active biological process than genes with distinct expression profiles. These are only the early implementations of GO analysis. One of the latest review-style publications by Ochs *et al.* [6] is cited as a place holder for the more than 70 papers on GO-based microarray analysis methods, not cited here.

Pathways

There are two ways to carry the analysis beyond GO classification deeper into biology: going to the molecular level, which is promoter and regulatory network analysis, or employing the vast-accumulated knowledge from the literature to carry out pathway analysis (see Figure 1 for overview). Although both are intrinsically linked, let us start out with the more popular and widespread pathway analysis. Pathways focus on physical and functional interactions between genes rather than taking the gene-centered view of GO-based analyses [7]. Therefore, it is intriguing to map the list of significantly regulated genes onto usually precompiled pathways in order to elucidate whole chains of events observed in a microarray experiment. Not all pathways are equally suitable for micro-

Table 1

Web-accessible tools for microarray pathway analysis (academic)

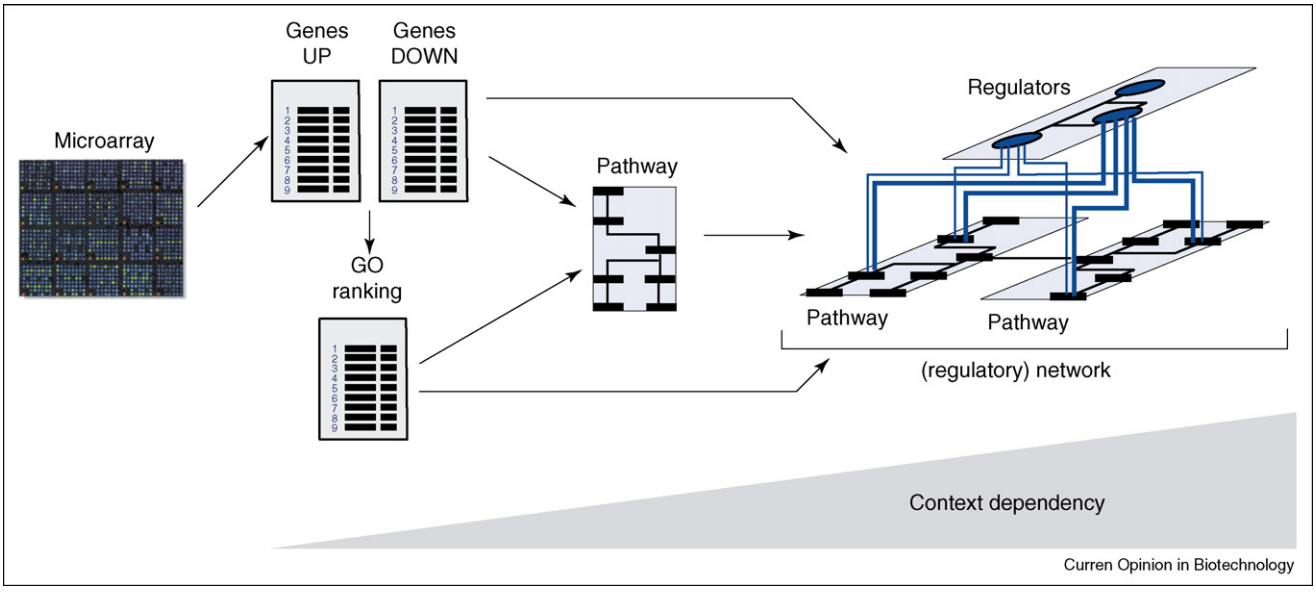
Tool	URL
ArrayXPath	http://www.snubi.org/software/ArrayXPath/
CRSD	http://biochip.nchu.edu.tw/crsd1/
DGEM	http://dgem.dhcp.iupui.edu/
Eu.Gene	http://www.ducciocavalieri.org/bio/Eugene.htm
GenMAPP 2	http://www.genmapp.org
KEGG	http://www.g-language.org/data/marray/
KOBAS	http://kobas.cbi.pku.edu.cn
Onto-Express	http://vortex.cs.wayne.edu/Projects.html#Onto-Express
PathExpress	http://bioinfoserver.rsbs.anu.edu.au/utils/PathExpress/
Pathway Miner	http://www.biorag.org/index.php
PAP	http://bioinformatics.wustl.edu/webTools/portalModule/PromoterSearch.do
VisANT3.0	http://visant.bu.edu

array analysis. Metabolic pathways are controlled to a large extent by protein-based events, not observable on microarrays as only steady-state levels of mRNAs are monitored. Kinase-based signaling cascades also do not necessarily involve changes in mRNA levels. The best case for microarray-based pathway analysis is transcriptional-signaling pathways that are directly coupled to *de novo* transcription. Although they also contain post-transcriptional steps usually there is enough transcriptional feedback regulation of pathway genes to allow identification of the pathways via mRNA level changes.

Most pathway analysis tools rely on precompiled databases of pathways derived from large-scale literature analysis requiring constant updating because of the continuous growth of the literature. Especially the maintenance efforts result in a strain on the resources academic groups can hardly accommodate, which is why academic efforts almost exclusively rely on the KEGG database (Table 1). Pathway tools backed by the more elaborative literature research and databases are entirely commercial efforts. There is a summary of such methods on the Affymetrix gene-chip-compatible pages on which Table 2 is based. As I will not discuss commercial applications, Table 2 serves as overview and further reference for such methods.

As microarray data can be mapped onto most of all pathways it is important to rank those association to highlight the most tightly associated, that is, the most relevant, pathway. One of the first attempts published for pathway analysis is the GenMAPP tool [8], a Windows-based application that maps genes to pathways graphically allowing the user to modify and construct pathways as well. GenMAPP leaves the ranking to the user or other statistical analyses of the initial data but has been upgraded and extended recently [9], adding phylogenetic information on pathways as well as protein–protein interaction data. The web-based pathway miner provides ranking of the gene/pathway groups via a Fisher's exact test on top of the gene–pathway association analysis [10]. Another tool that publishes the same year takes a very

Figure 1



From microarray to biological interpretation of results. Genes Up, Genes Down, and GO ranking symbolize lists, which are essentially one-dimensional representations of the data, though GO ranking already contains groups of genes. The pathway symbol indicates the two-dimensional nature of pathways that can branch, while networks are essentially multidimensional interlinking several pathways. The arrows indicate that there are many ways to develop the more complex structures from the initial data, for example, GO ranking is an optional but not mandatory step in pathway and network mapping.

Table 2

Gene-Chip® compatible tools for DNA sequence and expression analysis from Affymetrix web site (commercial)

Company	Tool
DNA sequence analysis	
Biotique Systems	Local integration system (BLIS)
GenomeQuest	GenomeQuest
Genomatix	ElDorado
	GEMS Launcher
Exon expression	
Biotique Systems	XRAY
Genomatix	ChipInspector
Partek	Genomics Suite
SAS	JMP Microarray for Affymetrix Exon Analysis
Gene expression	
Applied Math	GeneMaths XT
Biodiscovery	GeneSight
Biotique Systems	XRAY
DNASTAR	ArrayStar v2.0
Gene Data	Expressionist
Genomatix	ChipInspector
Ocium Biosolutions	Genowitz
Partek	Genomics Suite
Rosetta Genomics	Rosetta Resolver System
SAS	JMP Microarray for Affymetrix Exon Analysis
Spotfire	DecisionSite for Functional Genomics
VizX Labs	GeneSifter
Regulation analysis	
Genomatix	ChipInspector
Partek	Genomics Suite
Pathway network analysis	
Ariadne	Pathway Studio
BioBase	ExPlain
GeneGo	MetaCore
Genomatix	BiblioSphere PathwayEdition
Ingenuity	Ingenuity Pathway Analysis (IPA)

similar approach adding Scalable Vector Grafics (SVGs) to the picture and relying on gene expression clusters as input [11]. The program got updated a year later but still relies on SVG [12], support of which was discontinued by Adobe. Tomita *et al.* are responsible for generation and maintenance of the KEGG database of metabolic and other pathways [13], also joined the ranks of microarray-based pathway analysis with a web-accessible tool [14]. Wu *et al.* published another KEGG-related web-accessible tool called KOBAS, using the KEGG orthology (KO) as a controlled vocabulary to map gene sequences or identifiers back to KEGG pathways. They added four statistical tests (binominal, Chi-square, Fisher's exact, and hypergeometric distribution test) to assess the relevance of the found pathways [15]. Taking a slightly different approach, the DGEM system of Xia *et al.* focuses on statistical association of microarray data with clinical disease parameters utilizing pathway analysis (again KEGG based) as a subsequent information [16]. The VisANT 3.0 web-based system developed by Hu *et al.* also utilizes KEGG as its backbone for pathway analysis but adds the ability to predict pathways based on physical interaction data and coexpression profiles, also using SVG [16]. Another microarray pathway analysis tool focusing entirely on the metabolic pathways in KEGG is PathExpress by Goffard and Weiller [17]. Finally, a stand alone JAVA-based application named Eu.Gene utilizes information from KEGG as well as from GenMapp and the Reactome database to analyze the correlation of genes to pathways. Two statistical tests, Fisher's exact and Gene Set Enrichment Analysis (GSEA) are used to determine the significance of these correlations [18].

From pathway to network

Pathway analysis is useful for the interpretation of microarray data. However, the major difference to GO categories is also their most crucial shortcoming: biological processes usually involve more than one pathway, more precisely pathways interconnect in a context-specific manner. The result is a network, in case of microarray data mainly a regulatory network (Figure 1). Such networks cannot be derived from literature or precompiled pathways easily as network structures are not fixed and change with context. Regulatory networks also target and change usually only a few genes per pathway involved but do so across several pathways connected via the particular regulatory network as became evident especially in development [19]. This explains why a few genes usually only hit pathways and complete pathways never show up as coordinately up-regulated or down-regulated on the level of individual genes. Tackling regulatory networks requires to tap into new resources, genomic regulatory sequences (enhancers and promoters) as well as to differentiate transcripts rather than genes, as alternative transcript often originate also from alternative promoters and transcript identification is required in order to select the correct promoter for analysis [20]. New exon junction microarrays might aid this process [21]. On top of that, tools and databases for the analysis of regulatory sequences are required such as databases of transcription factor binding site (TFBS) descriptions (usually weight matrices) and corresponding programs to analyze the sequences. However, as has been firmly established, statistical enrichment of particular TFBSs in sets of promoters is not sufficient to explain specific regulation as the arrangement of the TFBSs relative to each other is also an important factor [22–24].

Till date, several attempts have been published to approach a molecular analysis of regulatory networks either as a stand alone or combined with GO and pathway analysis. Initially, sequence-based approaches were confined to yeast and similar relatively simple systems [25,26]. However, in the meantime several groups also approached the more complex mammalian systems.

Bluthgen *et al.* combined TFBS analysis of promoters of coregulated genes with GO annotation of the respective TFs in order to link coregulation and biological functional context [27]. The CRDS system aims at providing a more complete microarray analysis incorporating statistical and cluster analysis with some initial promoter analysis for TFBSs along with GO and pathway analysis [28]. Veerla *et al.* used a purely microarray and promoter analysis-based approach to identify relevant TFBSs by clustering of promoters or regulated genes according to high-scoring TFBS motifs in such promoters [29]. Chang *et al.* have more recently proposed a system called PAP that attempts to analyze promoters of coexpressed genes from miroarrays for common TFBSs including phylogenetic conservation of such sites in other genomes in order to identify potential regulators of the genes [30].
There are also at least two published studies including specific organization of TFBSs within promoters of significantly regulated genes in order to elucidate relevant regulatory networks.

Di Cara *et al.* introduced PromoterPlot, a web-based system, that utilizes an initial TFBSs analysis of potentially coregulated promoters and attempts to find TFBSs triplets with conserved internal spacing and consistent binding strands. Although the system requires extensive preanalysis of the sequences and is not directly applicable to microarray data, it used the concept of transcriptional modules for comparative promoter analysis [31]. Complete analysis of microarray data from an experiment of PDFG stimulation of fibroblasts in another study using a set of commercial tools revealed the complete regulatory network linking PDFG to several key transcriptional changes [32].

The latest challenge for pathway/network analysis: ChIP-on-chip

The genome-wide application of chromatin-immunoprecipitation (ChIP) in connection with promoter/genome tiling arrays poses a new challenge for pathway and network analyses as in this case there is no obvious way to subgroup results. Some of the resulting challenges have been recently reviewed [33]. A genomic sequence is either bound by a TF or not, but there are no obvious groups such as coexpression profiles unless additional experiments are carried out.

There were already some initial tools/approaches for ChIP-on-chip analysis published, such as Chipper [34], RINGO/Bioconductor [35], original [36], the multiscale analysis by Lerman *et al.* [37], and the MAT system from a Harvard group [38]. However, none of these tools ventures beyond identification and annotation of the bound genomic regions. A recent review by Goutsisas and Lee lists the required resources for regulatory network analysis including ChIP data: gene expression profiling, *cis*-regulatory element identification, TF target gene identification, and gene silencing by RNA interference [39]. In one word: the analysis of regulatory networks, which in this case is the only means to directly group gene as required for pathway analysis in contrast to expression array experiments. This will be even more important as ChIP-on-chip with nuclear receptors targets many more unknown enhancers than known promoters [40,41].

Whatever approach is used, the link between ChIP-on-chip and pathways goes definitely via molecular regulatory networks. As already indicated by Goutsisas and Lee, the key to understanding biological networks is to combine as many data and analyses as possible, which basically is a hallmark of systems biology. As noted earlier, consequent regulatory network analysis of microarrays links this technology directly into a variety of systems-oriented approaches from different fields of research [23].

As a final note, I would like to add that new high-throughput sequencing technologies such as 454 and Solexa yield data that are intrinsically similar to microarray data [42] and can actually be processed by the same strategies.

References

1. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci U S A* 2001, **98**:5116-5121.
2. Altman RB, Raychaudhuri S: **Whole-genome expression analysis: challenges beyond clustering**. *Curr Opin Struct Biol* 2001, **11**:340-347.
3. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: database for annotation, visualization, and integrated discovery**. *Genome Biol* 2003, **4**:P3.
4. Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA: **Global functional profiling of gene expression**. *Genomics* 2003, **81**:98-104.
5. Hvidsten TR, Laegreid A, Komorowski J: **Learning rule-based models of biological process from gene expression time profiles using gene ontology**. *Bioinformatics* 2003, **19**:1116-1123.
6. Ochs MF, Peterson AJ, Kossenkov A, Bidaut G: **Incorporation of gene ontology annotations to enhance microarray data analysis**. *Methods Mol Biol* 2007, **377**:243-254.
7. Thomas PD, Mi H, Lewis S: **Ontology annotation: mapping genomic regions to biological function**. *Curr Opin Chem Biol* 2007, **11**:4-11.
8. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR: **GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways**. *Nat Genet* 2002, **31**:19-20.
9. Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, Dahlquist KD, Doniger SW, Stuart J, Conklin BR, Pico AR: **GenMAPP 2: new features and resources for pathway analysis**. *BMC Bioinformatics* 2007, **8**:217.
10. Pandey R, Guru RK, Mount DW: **Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data**. *Bioinformatics* 2004, **20**:2156-2158.
11. Chung HJ, Kim M, Park CH, Kim J, Kim JH: **ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics**. *Nucleic Acids Res* 2004, **32**: W460-W464.

12. Chung HJ, Park CH, Han MR, Lee S, Ohn JH, Kim J, Kim J, Kim JH: **ArrayXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using Scalable Vector Graphics**. *Nucleic Acids Res* 2005, **33**:W621-W626.

13. Kanehisa M, Goto S, Kawashima S, Nakaya A: **The KEGG databases at GenomeNet**. *Nucleic Acids Res* 2002, **30**:42-46.

14. Arakawa K, Kono N, Yamada Y, Mori H, Tomita M: **KEGG-based pathway visualization tool for complex omics data**. *In Silico Biol* 2005, **5**:419-423.

15. Wu J, Mao X, Cai T, Luo J, Wei L: **KOBAS server: a web-based platform for automated annotation and pathway identification**. *Nucleic Acids Res* 2006, **34**:W720-W724.

16. Xia Y, Campen A, Rigsby D, Guo Y, Feng X, Su EW, Palakal M, Li S: **DGEM — a microarray gene expression database for primary human disease tissues**. *Mol Diagn Ther* 2007, **11**:145-149.

17. Goffard N, Weiller G: **PathExpress: a web-based tool to identify relevant pathways in gene expression data**. *Nucleic Acids Res* 2007, **35**:W176-W181.

18. Cavalieri D, Castagnini C, Toti S, Maciag K, Kelder T, Gambineri L, Angioli S, Dolara P: **Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases**. *Bioinformatics* 2007.

19. Boyle S, de Caestecker M: **Role of transcriptional networks in coordinating early events during kidney development**. *Am J Physiol Renal Physiol* 2006, **291**:F1-F8.

20. Tzvetkov MV, Meineke C, Oetjen E, Hirsch-Ernst K, Brockmoller J: **Tissue-specific alternative promoters of the serotonin receptor gene HTR3B in human brain and intestine**. *Gene* 2007, **386**:52-62.

21. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays**. *Science* 2003, **302**:2141-2144.

22. Werner T, Fessele S, Maier H, Nelson PJ: **Computer modeling of promoter organization as a tool to study transcriptional coregulation**. *FASEB J* 2003, **17**:1228-1237.

23. Werner T: **Regulatory networks: linking microarray data to systems biology**. *Mech Ageing Dev* 2007, **128**:168-172.

24. Howard ML, Davidson EH: ***cis*-Regulatory control circuits in development**. *Dev Biol* 2004, **271**:109-118.

25. Pilpel Y, Sudarsanam P, Church GM: **Identifying regulatory networks by combinatorial analysis of promoter elements**. *Nat Genet* 2001, **29**:153-159.

26. Sudarsanam P, Pilpel Y, Church GM: **Genome-wide co-occurrence of promoter elements reveals a *cis*-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae***. *Genome Res* 2002, **12**:1723-1731.

27. Bluthgen N, Kielbasa SM, Herzel H: **Inferring combinatorial regulation of transcription in silico**. *Nucleic Acids Res* 2005, **33**:272-279.

28. Liu CC, Lin CC, Chen WS, Chen HY, Chang PC, Chen JJ, Yang PC: **CRSD: a comprehensive web server for composite regulatory signature discovery**. *Nucleic Acids Res* 2006, **34**:W571-W577.

29. Veerla S, Hoglund M: **Analysis of promoter regions of co-expressed genes identified by microarray analysis**. *BMC Bioinformatics* 2006, **7**:384.

30. Chang LW, Fontaine BR, Stormo GD, Nagarajan R: **PAP: a comprehensive workbench for mammalian transcriptional regulatory sequence analysis**. *Nucleic Acids Res* 2007, **35**:W238-W244.

31. Di Cara A, Schmidt K, Hemmings BA, Oakeley EJ: **PromoterPlot: a graphical display of promoter similarities by pattern recognition**. *Nucleic Acids Res* 2005, **33**:W423-W426.

32. Seifert M, Scherf M, Epple A, Werner T: **Multievidence microarray mining**. *Trends Genet* 2005, **21**:553-558.

33. Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, Weissman S, Snyder M, Gerstein M: **Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping**. *Trends Genet* 2005, **21**:466-475.

34. Gibbons FD, Proft M, Struhl K, Roth FP: **Chipper: discovering transcription-factor targets from chromatin immunoprecipitation microarrays using variance stabilization**. *Genome Biol* 2005, **6**:R96.

35. Toedling J, Sklyar O, Huber W: **Ringo — an R/Bioconductor package for analyzing ChIP-chip readouts**. *BMC Bioinformatics* 2007, **8**:221.

36. Reimers M, Carey VJ: **Bioconductor: an open source framework for bioinformatics and computational biology**. *Methods Enzymol* 2006, **411**:119-134.

37. Lerman G, McQuown J, Blais A, Dynlacht BD, Chen G, Mishra B: **Functional genomics via multiscale analysis: application to gene expression and ChIP-on-chip data**. *Bioinformatics* 2007, **23**:314-320.

38. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS: **Model-based analysis of tiling-arrays for ChIP-chip**. *Proc Natl Acad Sci U S A* 2006, **103**:12457-12462.

39. Goutsias J, Lee NH: **Computational and experimental approaches for modeling gene regulatory networks**. *Curr Pharm Des* 2007, **13**:1415-1436.

40. Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I: **Predicting tissue-specific enhancers in the human genome**. *Genome Res* 2007, **17**:201-211.

41. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF *et al.*: **Genome-wide analysis of estrogen receptor binding sites**. *Nat Genet* 2006, **38**:1289-1297.

42. Kaller M, Lundeberg J, Ahmadian A: **Arrayed identification of DNA signatures**. *Expert Rev Mol Diagn* 2007, **7**:65-76.