

IMBA Workshop

Berlin, May 27, 2008

Validity and reliability of current omics studies for cancer risk
assessment
(EMF studies)

PD Dr. Ludger Klein-Hitpass
Institute of Cell Biology (Tumor Research)
Biochip-Lab
Medical Center Essen, Germany

„Transcript-omics“

Most widely used „omics“ tool: microarray (gene chip, genome chip, DNA chip, biochip, gene array, GeneChip®)

Possible applications: genotyping (SNP), copy numbers variation, LOH, resequencing, splicing analysis, steady-state mRNA level of nearly all genes

Aims of transcriptomic studies using microarrays: Identification of EMF specific gene signatures or pattern indicating exposure, initiation of biological events and mechanisms involved.

Approach has been proved to be successful in many areas of research, particularly in tumor research.

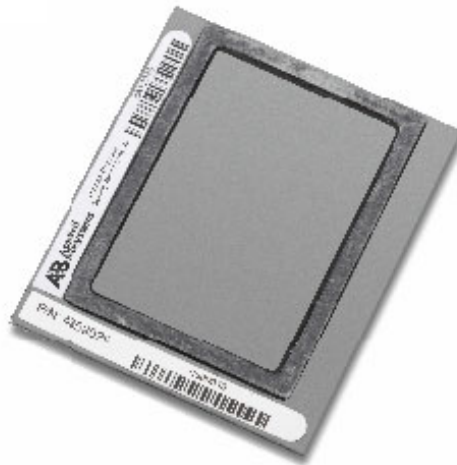
Key components of microarray analyses

- 1. Microarray:** a collection of different nucleic acid sequences, which are immobilized onto a matrix in an ordered fashion.
- 2. Labeled target,** representing the mRNA population of the sample of interest (cells, tissue, organ), which is hybridized to the array (labeling often includes a linear amplification step).
- 3. Detection system,** that measures the amount of labeled mRNAs bound to complementary probes on the array, which is a measure of the relative abundance of the transcript in the sample.

Microarray system component 1: the microarray



Applied Biosystems Array
(60mer oligonucleotides,
spotted)



Affymetrix GeneChip®
(up to 1.3×10^6 25mer oligonucleo-
tides, in situ synthesis)

Genome-wide expression analysis strategies

1. Use genome-wide microarray covering ~30.000 genes/transcript variants

Affymetrix HG-U133Plus_2.0

Applied Biosystems Human Genome Survey Array v2.0

Agilent Whole Human Genome

Illumina Humanwg-6 beadchip

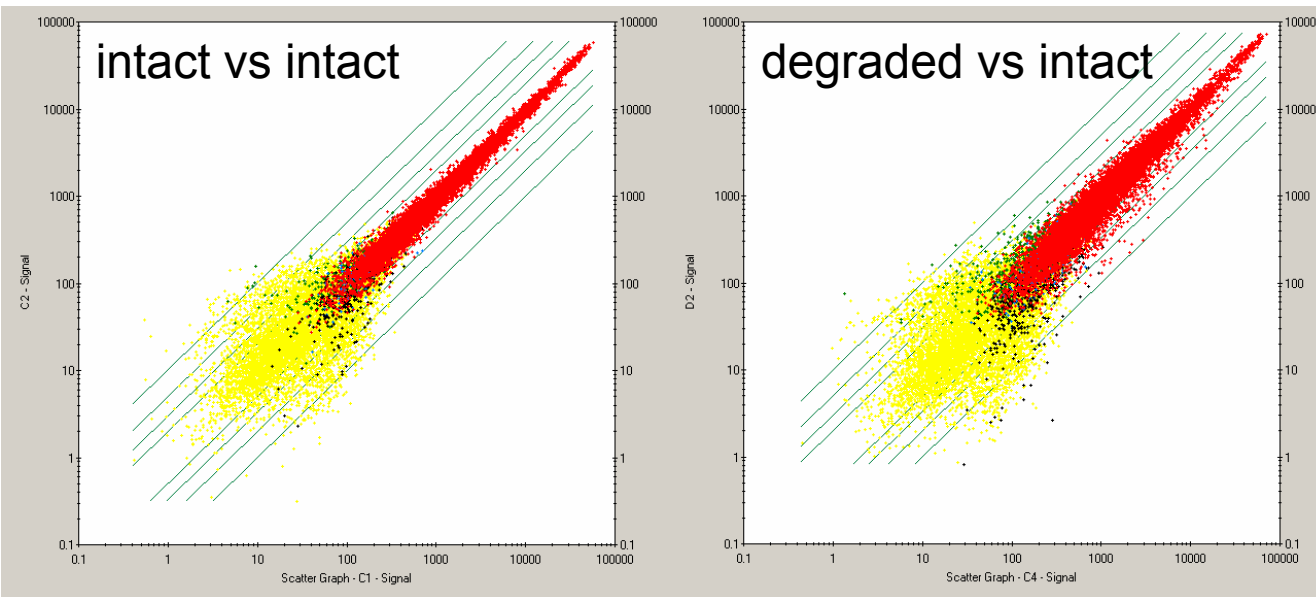
Homemade array containing cDNA or oligonucleotide probes

2. Serial analysis of gene expression (SAGE)

3. High-throughput sequencing of clonally amplified cDNA tags

Key challenges in microarray studies: RNA quality

Comparable RNA purity and quality is a prerequisite for an excellent microarray analysis



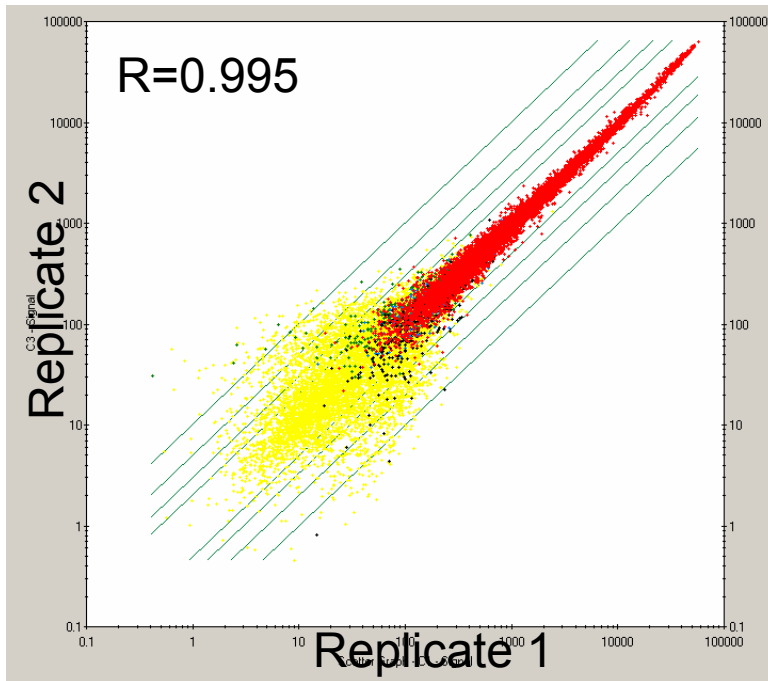
Comparison of partially degraded samples vs high quality RNAs leads to greatly increased numbers of apparently regulated genes (type I errors).

Quality of RNA samples and labeled targets has to be checked for signs of degradation and incomplete reverse transcription.
RNA quality (Affymetrix 3'/5'-ratio) should be documented.

Key challenges in microarray studies: experimental variation

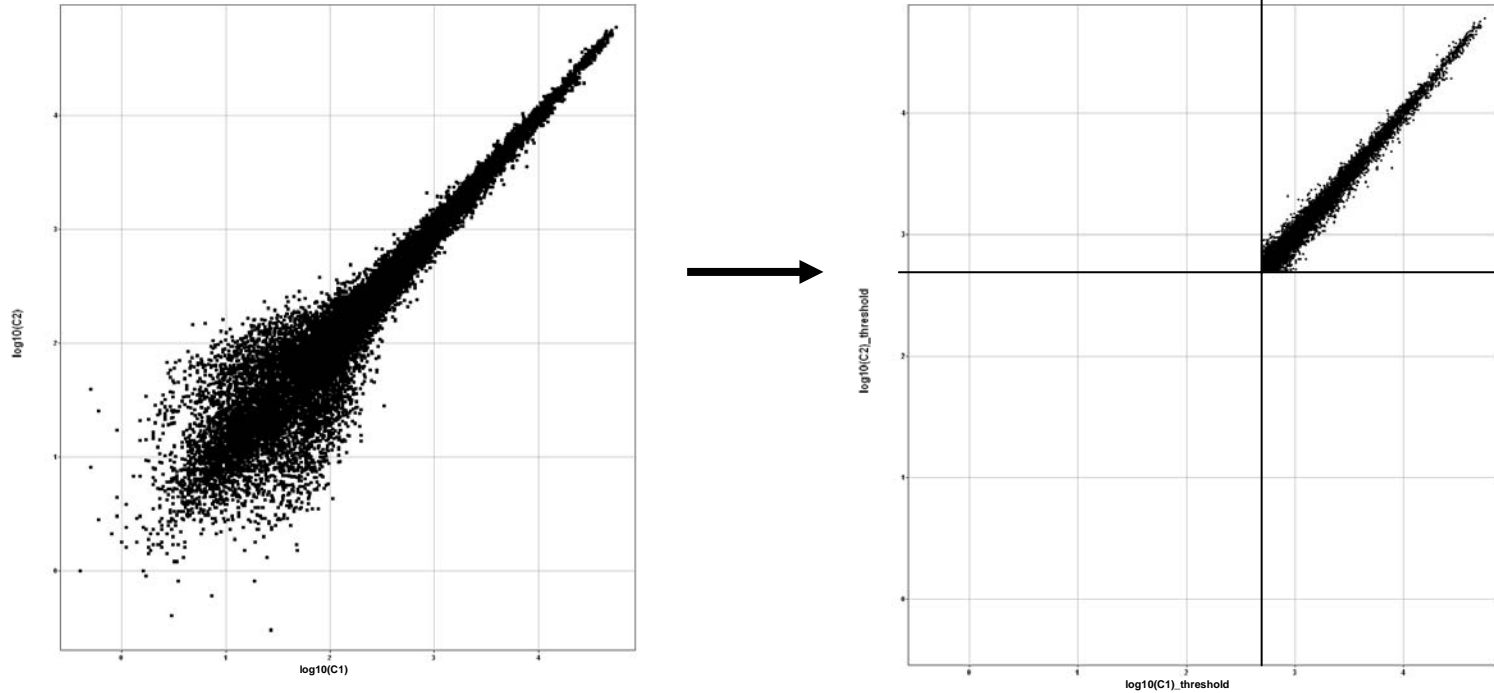
Noise or experimental variation causes unpredictable fluctuations of the measurements resulting in statistical errors.

Sources of variation: each step and reagent used from the sample over treatment to image acquisition.



Sufficient replicates are recommended, to allow estimation of experimental variance and to reduce impact of variability in statistical analyses.
(Intra- and inter-group correlation values should be reported).

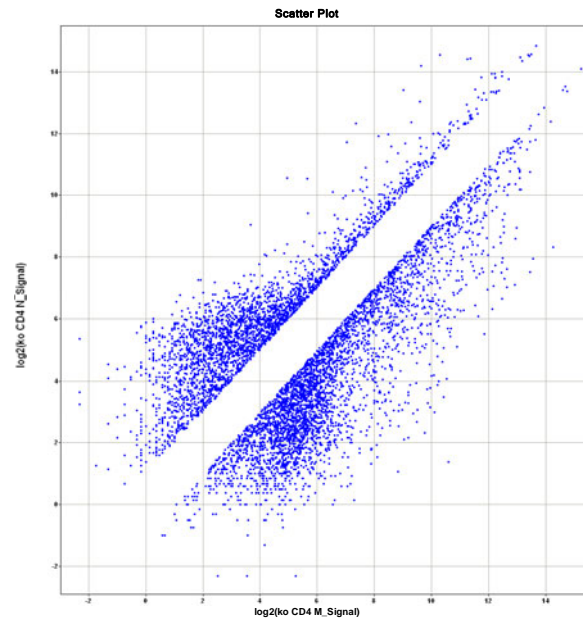
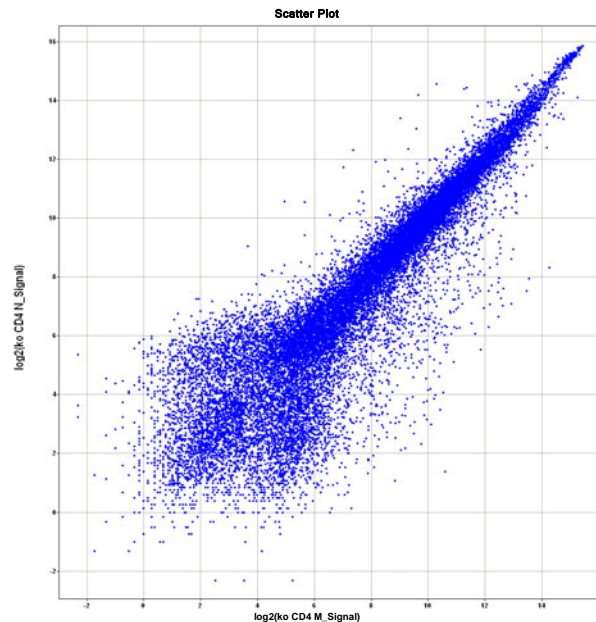
Key challenges in microarray studies: sensitivity (application of signal thresholds)



Adjustment of below threshold signals to an arbitrarily chosen signal value should be considered carefully to avoid a great loss of sensitivity (\uparrow type II error rate).

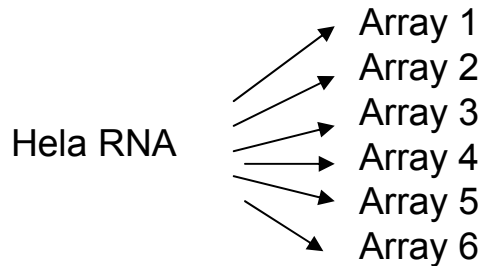
Key challenges in microarray studies: sensitivity (fold change thresholds)

Pair-wise comparison of two different cell populations. Application of a fold change threshold of 2



Application of fold change thresholds has to be avoided, because it results in loss of sensitivity (\uparrow type II error rates) and possible loss of „biology“.

Key challenges in microarray studies: a statistical issue (multiple testing problem)



$$R_{\text{avg}} = 0.995 \pm 0.0015$$

HG-U133A (22283 probesets)

t-test	Class 1	Class 2	# Probesets $p \leq 0.05$	# Probesets $p \leq 0.05$ corrected*	# Probesets $p \leq 0.05$ Avg FC > 2.0 $\uparrow\downarrow$
Test 1	1, 2, 3	4, 5, 6	966 (4.3%)	0	204
Test 2	2, 3, 6	1, 4, 5	472 (2.1%)	0	215
Test 3	3, 4, 5	1, 2, 6	1122 (5.0%)	0	278
Test 4	2, 5, 6	1, 3, 4	410 (1.8%)	0	216
Test 5	4, 6, 1	3, 5, 2	519 (2.3%)	0	231
Test 6	1, 3, 6	2, 4, 5	3203 (14.4%)	1	374

*Benjamini-Hochberg

Application of multiple testing correction adjustment methods (Bonferroni etc.) to classical test results effectively removes false positives, but has to be carefully considered, because they often eliminate truly significant genes (too conservative > increased type II error rate).

Permutation based FDR approaches, such as SAM, appear to be more adequate, but require a larger number of replicates ($n \geq 6$).

Key challenges in microarray data analysis: biological interpretation of the data

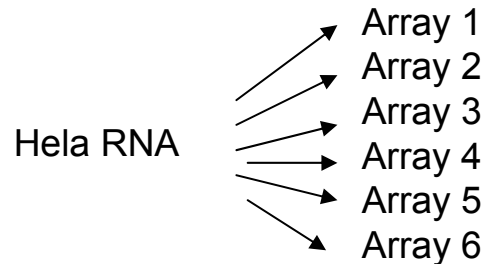
Candidate gene lists may contain a great number of genes with diverse function and many false positives.

Method1: Sort genes according to „relevant“ functions or associated biological processes to identify most frequently observed biological processes (less meaningful; bias)

Method 2: GO over-representation, GSEA and pathway analyses (statistical approaches) by web-based tools, such as GOstat, GoMiner, GeneTrail, GSEA

Target gene lists should be analyzed by statistical approaches to obtain hints for processes induced by treatment.

GO overrepresentation analysis of target list consisting of type I errors only:
biological interpretation of the data



$$R_{avg} = 0.995 \pm 0.0015$$

HG-U133A (22283 probesets)

t-test	Class 1	Class 2	# Probesets $p \leq 0.05$	GO Biological processes $p \leq 0.001$, corrected*
Test 1	1, 2, 3	4, 5, 6	966 (4.3%)	0
Test 2	2, 3, 6	1, 4, 5	472 (2.1%)	1 n.i.
Test 3	3, 4, 5	1, 2, 6	1122 (5.0%)	0
Test 4	2, 5, 6	1, 3, 4	410 (1.8%)	0
Test 5	4, 6, 1	3, 5, 2	519 (2.3%)	0
Test 6	1, 3, 6	2, 4, 5	3203 (14.4%)	0

n.i.: not informative

*Bonferroni correction

Target lists mostly consisting of type I errors rarely show highly significant overrepresentation of GO categories.

GO overrepresentation analysis of a target list containing all increased and decreased genes identified by comparison analysis:

Gene list with ~2000 probesets identified by pairwise comparison analysis as Increased or Decreased: 26 GO categories with $p < 0.001$

Best GOs	Biological Process	Count target list	Count chip	P-Value
GO:0006333	chromatin assembly or disassembly	107	170	8.04E-49
GO:0006325	establishment and/or maintenance of chromatin architecture	120	223	4.17E-41
GO:0006323	DNA packaging	120	223	4.17E-41
GO:0051276	chromosome organization and biogenesis	144	296	1.94E-40
GO:0006259	DNA metabolic process	161	367	5.58E-36
GO:0044249	cellular biosynthetic process	173	452	2.89E-27
GO:0006412	translation	118	264	5.01E-27
GO:0044237	cellular metabolic process	635	2524	1.67E-25
GO:0009058	biosynthetic process	186	516	6.44E-25
GO:0043170	macromolecule metabolic process	543	2094	2.99E-24

GOstat analysis indicates highly significant overrepresentation of genes associated chromatin assembly processes (suspected regulator of chromatin remodeling had been knocked out by siRNA).

Experimental design and identification of regulated genes: the need for validation experiments

Depending on the number of replicates and the data mining strategy (single or cross-comparison analysis, TTEST, SAM, etc.) performed, any candidate gene lists obtained from microarray studies may contain a great number of false positives or even consist of false positives only.

Thus candidate genes have to be validated by an independent method.

Method of choice: quantitative RT-PCR (qPCR, qRT-PCR, real-time PCR)

- high dynamic range
- superior sensitivity
- cheap, medium through-put method

Validation should involve a larger number of independent samples (increased statistical power) and a greater number of candidate genes (enables estimate of type I error rate).

Key challenges in microarray data analysis: data accessibility

Microarray studies produce easily 100,000 to millions of data points (genes x samples). Entire dataset cannot be published along with paper.

Solution: submission to public microarray data repositories

GEO (NCBI, NIH)

ArrayExpress (EBI, Heidelberg)

CIBEX (Japan)

Table with normalized signal intensities, raw images, sample information, experimental design, data processing steps, array platform, annotation (MIAME standard). Data viewing, browsing, query, and retrieval functions ensure that data is fully and permanently accessible to the public.

Deposition of microarray data (raw and processed) to a public data repository has to become an obligatory part of the publication process and condition of acceptance to allow that conclusions can be re-evaluated independently

Recommended approach in microarray studies:

Step 1: Perform a sensitive genome-wide array analysis

- Whole genome array combined with a sensitive detection method
 - **Six or more** replicates of independent samples
 - Avoid signal or FC thresholds, which compromise sensitivity
 - Analyze data by classical test with/without correction or by permutation based FDR statistic (SAM)
-
- > Sensitive study exhibiting low rates of both error types
 - > Candidate list capturing the „biology“ completely
 - > Subject candidate list to statistically based biological interpretation

Step 2: Validation of candidate genes

- Validate candidate genes by qPCR in a larger set of independent samples to obtain additional evidence for gene regulation

Step 3 : Establish data transparency

- Submission of processed and raw array data into public data repository
- Provide all information on data processing steps

An alternative approach in microarray studies:

Step 1: Array analysis

- Whole genome array combined with a sensitive detection method
 - **Three (or less)** array replicates of independent samples
 - Avoid signal or FC thresholds, which compromise sensitivity
 - Classical statistic without correction or comparison analysis (ratio)
- > Analysis methods will create a relatively high rate of type I errors
- > High sensitivity, relatively low rate of type II errors
- > „Biology“ may be diluted out
- > Subject candidate list to statistically based biological interpretation

Step 2: Validation of candidate genes

- Validate candidate genes by qPCR in a **much larger set** of independent samples to obtain additional evidence for gene regulation

Step 3 : Establish data transparency

- Submission of processed and raw array data into public data repository
- Provide all information on data processing steps

Evaluation of studies using microarrays in EMF research

Study retrieval: Pubmed, IMBA database

Time interval: 2003-2008

No. of studies retrieved/accessible: 15

Most frequent journals:

Proteomics (n=4, IP=5.732)

Bioelectromagnetics (n=4, IP=1.514)

Radiation Research (n=3, IP=2.602)

Evaluation of studies using microarrays in EMF research: experimental parameters

Sample types: cell lines (11/15), primary cells (4/14), brain tissue (2/15)

EMF exposure devices: many different types, one study uses mobile phone

Temperature monitoring: all but one study exclude thermal effects

SAR: up to 10 W/kg

Field frequencies: ~900 or ~1800 MHz, various modulation forms

Time of exposure: variable, 1-72 h

„Recovery“ periods: mostly none; 2 or 6 h

Published microarray studies use large variety of unique combinations of experimental variables and, thus, can hardly be compared.

Evaluation of studies using microarrays in EMF research: array systems used

Fluorescence based detection:	10 (8x Affymetrix, 2x Agilent)
Radioisotope detection:	3 (Clontech Atlas, RZPD-2)
Chemiluminescence:	1 (SuperArray Biosc. Apoptosis)
Sequencing (SAGE):	1

Genes on arrays:
between 96 (Apoptosis array) and 75,000 probes (RZPD-2)

Published microarray studies use a variety of array systems, making comparisons between studies difficult.

Microarray studies in EMF research: „negative“ studies according to author's conclusion

Authors	Port et al., 2003	Whitehead et al., 2006	Gurisek et al., 2006	Zeng et al., 2006	Qutob et al., 2006	Chauhan et al., 2007	Paparini et al., 2008
Array type	Clontech Atlas	Affymetrix U74Av2	Affymetrix HF	Affymetrix U133A	Agilent 1Av1	Agilent 1Av1	Affymetrix MOE430A
Detection system	X-ray film	Laser scanner	Laser scanner	Laser scanner	Laser scanner	Laser scanner	Laser scanner
No. of genes on array	1178	~9200	~8400	~14500	~18000	~18000	~14500
No. of genes detected	n.i.	n.i.	n.i.	~45%	n.i.	n.i.	n.i.
No. of replicate experiments	2	3	1	2	5	5	3 (pool of 5 samples)
SAR (W/kg)	2	5	0,2	2 and 3.5	0.1, 1, 10 (4h)	0.1, 1, 10 (24h)	0,2
Additional experimental variables	4 time points	2 modulation forms				2 cell lines	in vivo exposure
Analysis method	Ratio method	ttest	Affymetrix comparison?	Affymetrix cross-comparison	MAANOVA SAM	MAANOVA with correction	ttest
Signal threshold	yes	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.
Present call filtering	n.a.	yes; no bias	n.i.	n.i.	n.a.	n.a.	yes; bias?
FC threshold	yes, 2-fold	no	no	no	no	yes, 1.35-fold	yes, 1.5-fold
Consistency threshold	100%	n.a.	no	100%	n.a.	n.a.	n.a.
No. of candidates	1 to 8	~200 (\leq no. of exp. type I errors)	8	5 at 3.5 W/kg	0	0	75
Controls		Positive x-ray control			Positive heatshock control	Positive heatshock control	
Genome coverage	low	moderate	moderate	moderate	moderate	moderate	moderate
Data mining strategy	very conservative	ok	ok	conservative	ok	conservative	not overly conservative
Validation	no	no	partial	yes	n.a.	n.a.	yes
Validation (candidates; independent samples)	no	no	Failed (2; 3)	Failed (5; unclear)	n.a.	n.a.	Failed ? (30; 15): FC shown, but statistic not given!
GO enrichment/pathway analysis	n.a.	no	n.a.	n.a.	n.a.	n.a.	no (negative)
Raw data accessibility and data transparency	no	no	no	no	no	no	no
Compliance w. standards	no	no	no	no	no	no	no

Microarray studies in EMF research:
„positive“ studies according to author's conclusion

Authors	Lee et al., 2005	Belyaev et al., 2006	Nylund et al., 2006	Remondini et al., 2006	Zhao R. et al., 2007	Zhao TY et al., 2007
Array type	SAGE	Affymetrix U34A	Clontech Atlas	RZPD-2	Affymetrix RN U34A	SuperArray apoptosis
Detection system	Sequencer	Laser scanner	X-ray films	PhosphorImager	Laser scanner	CCD camera
No. of genes on array	n.a.	~8800	1176	75000 cDNAs	~1200	96
No. of genes detected	~10000	n.i.	~100	n.i.	n.i.	n.i.
No. of replicate experiments	1	3	3	1 to 2 (2 technical replicates)	1?	2
SAR (W/kg)	10	0,4	2,8	1 to 2.5	2	unclear
Additional exp. Variables	2 time points	in vivo exposure	2 cell lines	6 cell types, 2 modes	primary neuronal cell cultures	2 types of primary neuronal cells
Analysis method	Tag sampling statistic	Affymetrix cross-comparison (3x3)	ttest (p≤0.05)	>3 σ and SAM Δ >1.2	Affymetrix comparison?	Ratio method
Signal threshold	n.a.	no	yes	n.i.	n.i.	n.i.
Present call filtering	n.a.	n.i.	n.a.	n.a.	n.i.	n.a.
FC threshold	yes, 4-fold	no	yes, 2-fold	no	no	yes, 1.35-fold
Consistency threshold	n.a.	100%	n.a.	n.a.	n.a.	100%
No. of candidates	221, 759	12	2, 12	12, 32, 34	34	9
Controls						
Genome coverage	moderate	moderate	low	high	low	poor, biased
Data mining strategy	conservative	conservative	ok	conservative	ok	conservative
Validation (candidates; independent samples)	no	no	no	no	yes (25; 0) 3 repeats	yes (5; unclear)
Validation results	no	no	no	no	successful (19/25)	successful (3/5 and 4/5)
GO enrichment/pathway analysis	no	n.a.	n.a.	no	n.a.	n.a.
Raw data accessibility and data transparency	NCBI GEO	no	no	no	no	no
Compliance w. standards	no	no	no	no	no	no

Mostly due to the lack of positive validation, these „positive“ publications do not provide substantial evidence for EMF induced gene expression changes.

Microarray studies in EMF research: „negative“ studies according to author's conclusion

Authors	Hirose et al., 2006	Hirose et al., 2007
Array type	Affymetrix U133Plus_2.0	Affymetrix U133Plus_2.0
Detection system	Laser scanner	Laser scanner
No. of genes on array	~39000	~39000
No. of genes detected	~15000	~15000
No. of replicate experiments	6 (2 sets of 3 arrays)	6 (2 sets of 3 arrays)
SAR (W/kg)	0.08, 0.8	0.08, 0.25, 0.8
Additional experimental variables		2 cell lines, 3 time points (2, 24 and 48h), CW or W-CDMA; 12 experiments
Analysis method	Welch's ttest on each set ($p < 0.05$)	Welch's ttest on each set of 3 array with correction (BH) ($p < 0.05$)
Signal threshold	n.i	n.i.
Present call filtering	unclear	probably; bias?
FC threshold	no	no
Consistency threshold	n.a.	n.a.
No. of candidates	0 (?); only negative results of 21 p53-related genes reported	0 (?); only negative results of 4 hsp genes explicitly reported
Controls		Positive heatshock control
Genome coverage	high	high
Data mining strategy	ok	too conservative
Validation	yes	no
Validation (candidates; independent samples)	successful (4 negative p53-related genes; unclear)	n.a.
GO enrichment/pathway analysis	n.a.	n.a.
Raw data accessibility and data transparency	no	no
Compliance w. standards	no	no

Largest studies available up to date.
(Funded by Japanese mobile communication company NTT DoCoMo)

Obligatory submission of raw and processed data to public data repositories would greatly increase data transparency, allow re-analysis and that conclusions could be re-evaluated independently.

Evaluation summary

Two alternative guidelines for the design and analysis of array experiments were defined.

Based on these guidelines, 15 published studies using whole genome expression analysis methods were evaluated. Neither positive nor negative studies proved to be fully compliant with guidelines.

Only in very few studies genome coverage and sensitivity were satisfactory. In general, data analysis allowing a better balancing of type I and II errors, and the inclusion of substantial and more meaningful validation efforts, is required in future EMF studies.

All „positive“ EMF studies failed to provide supporting evidence by validation experiments involving a larger set of independent samples. Not a single positive finding has been independently reproduced in a different laboratory. Positive studies do not reveal a consistent EMF gene signature, processes induced in the exposed cells, or hints to mechanisms involved.

Study transparency has to be improved substantially by submission of raw and processed data to public data repositories (condition of grant approval and manuscript submission).