An Introduction to Analysis of Multiple Gene Expression Datasets

Pratyaksha Wirapati

Statistical Analysis Applied to Genome and Proteome Analysis
EMBnet Course, Lausanne, 5 February 2008

Outline

- Why should we analyze multiple datasets?
- How to get the datasets?
- How to analyze them together?

Microarray Datasets

A dataset is

- a set of gene expression arrays (from cell lines, tumors, etc.)
- collected under a certain study design (either experimental or observational)
- typically done in one specific microarray technology platform

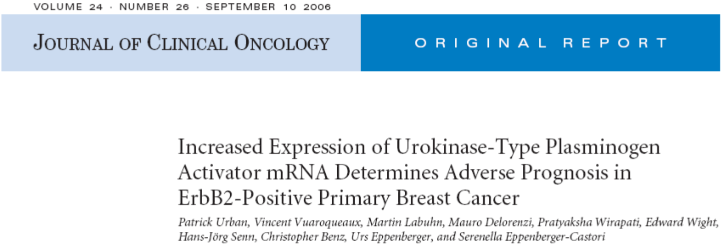
Usually, a new dataset is introduced and reported by a journal article.

Publication of the raw data of a microarray study is required by many journals, mainly to allow verification of the results by others.

However, there are other benefits for the research community

Example I

Urban et al. (2006) J Clin Oncol 24:4245



VOLUME 24 · NUMBER 26 · SEPTEMBER 10 2006

JOURNAL OF CLINICAL ONCOLOGY

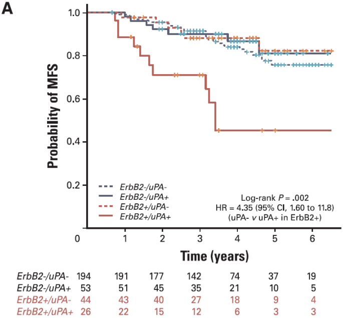
ORIGINAL REPORT

Increased Expression of Urokinase-Type Plasminogen Activator mRNA Determines Adverse Prognosis in ErbB2-Positive Primary Breast Cancer

Patrick Urban, Vincent Vuaroqueaux, Martin Labuhn, Mauro Delorenzi, Pratyaksha Wirapati, Edward Wight, Hans-Jörg Senn, Christopher Benz, Urs Eppenberger, and Serenella Eppenberger-Castori

- A study at Stiftung Tumorbank Basel and Oncoscore AG
- A collection of 317 breast cancer patients
- Expression of 60 genes assayed by quantitative RT-PCR
- Interested in subset of tumors with ERBB2+ (*her2/neu*) amplification

Example I: The Problem



ErbB2-/uPA-	194	191	177	142	74	37	19
ErbB2-/uPA+	53	51	45	35	21	10	5
ErbB2+/uPA-	44	43	40	27	18	9	4
ErbB2+/uPA+	26	22	15	12	6	3	3

Initial finding based on their data:

- The expression gene uPA is can predict metastasis within ERBB2+, but not in ERBB2– tumors.
- This interaction is not known before, and has potentially important clinical implications
- Is this “real”?

How do we know this is not an artefact of “data dredging” (overfitting)? There are 59 possible interacting genes.

Is it generalizable to all breast cancer population? Or only applies to this particular patient collection?

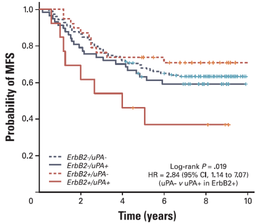
Example I: The Solution

- Check if there were publicly accessible datasets containing:
  - expression of the two genes (ERBB2 and uPA)
  - survival data (time-to-metastasis) for each patient
- Test if the two genes interact in the same way in these datasets as they do in the Basel dataset
  - There is no need to analyze all genes in these datasets, because we already have a very specific question about ERBB2, uPA and survival

Example I: Results

Rotterdam (EMC) dataset (n=286)
Affymetrix U133A chip

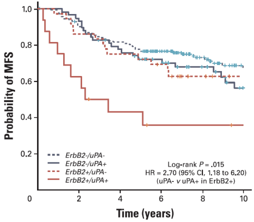
B



	0	2	4	6	8	10
ErbB2-/uPA-	177	153	129	112	78	24
ErbB2-/uPA+	58	46	40	30	20	13
ErbB2+/uPA-	38	33	28	24	15	8
ErbB2+/uPA+	13	8	7	4	4	0

Amsterdam (NKI) dataset (n=295)
Agilent custom chip

C

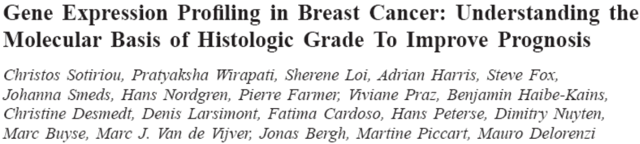


	0	2	4	6	8	10
ErbB2-/uPA-	185	165	148	116	73	47
ErbB2-/uPA+	58	55	44	35	26	16
ErbB2+/uPA-	36	31	27	22	13	10
ErbB2+/uPA+	16	10	6	4	4	3

- Similar results (ERBB2+/uPA+ ⇒ bad survival) were reproduced!
- Strong independent evidence (different patients, different platforms)
- External validation can be done very efficiently
- Existing data is useful beyond the purpose of original studies

Example II

Sotiriou et al. (2006) J Natl Cancer Inst 98:262

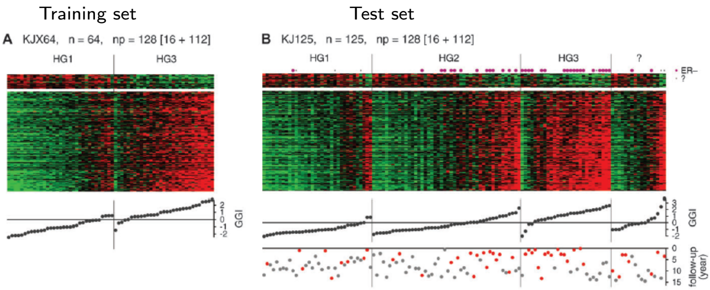


Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis

Christos Sotiriou, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, Pierre Farmer, Viviane Praz, Benjamin Haibe-Kains, Christine Desmedt, Denis Larsimont, Fatima Cardoso, Hans Peterse, Dimitry Nuyten, Marc Buyse, Marc J. Van de Vijver, Jonas Bergh, Martine Piccart, Mauro Delorenzi

- Histologic tumor grade is produced by pathologists based on conventional techniques (microscopy of stained tumor specimens)
- In breast cancer, it is a strong prognostic factor (high grade means bad survival), but intermediate grade is ambiguous
- What are the genes related to tumor grade?

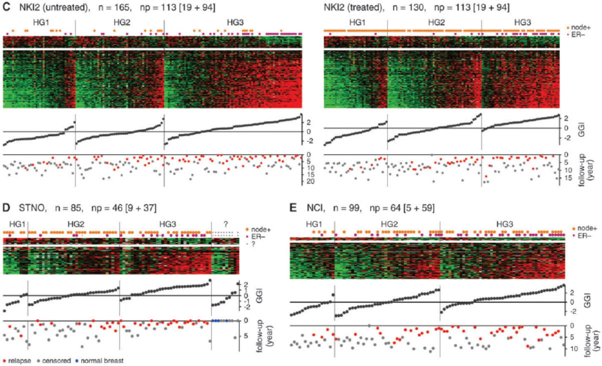
Example II: Initial Finding



- The training set was scanned for genes distinguishing low histologic grade (HG1) and high grade (HG3); 128 probesets were found
- The patterns were confirmed in the test set
- Additionally, intermediate grade tumors (HG2) were found to have patterns like HG1 or HG2
- Also potential association with survival

Example II: External Validation

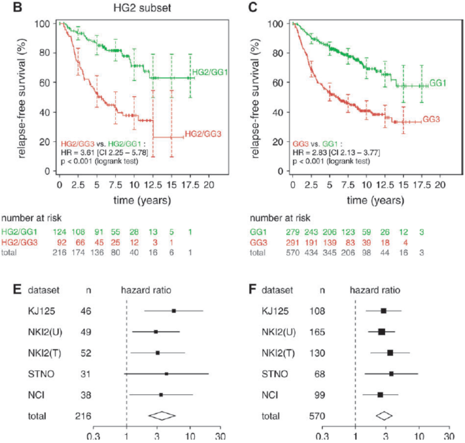
Independent datasets confirmed the findings



Only the genes in the signature (gene list) need to be checked in the external datasets (along with relevant clinical data)

Some genes can not be mapped across platforms

Example II: Summarizing Survival Analysis



number at risk

HG2/GG1	124	108	91	55	28	13	5	1
HG2/GG3	92	66	45	25	12	3	1	
total	216	174	136	80	40	16	6	1

number at risk

GG1	279	243	206	123	59	26	12	3
GG3	291	191	139	83	39	18	4	
total	570	434	345	206	98	44	16	3

dataset	n	hazard ratio
KJ125	46	
NKI2(U)	49	
NKI2(T)	52	
STNO	31	
NCI	38	
total	216	

dataset	n	hazard ratio
KJ125	108	
NKI2(U)	165	
NKI2(T)	130	
STNO	68	
NCI	99	
total	570	

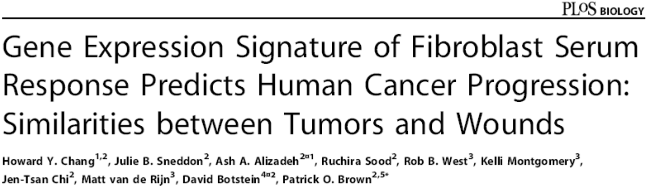
⇐ Kaplan-Meier plots of pooled data

⇐ Forest plot: summaries of individual datasets

A standard technique in meta-analysis: displays *consistency* across datasets

Example III

Chang et al (2004) PLoS Biol 2:0206



PLoS BIOLOGY

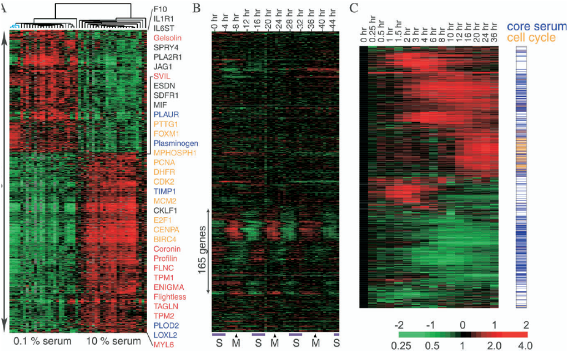
Gene Expression Signature of Fibroblast Serum Response Predicts Human Cancer Progression: Similarities between Tumors and Wounds

Howard Y. Chang[1,2], Julie B. Sneddon[2], Ash A. Alizadeh[2¤1], Ruchira Sood[2], Rob B. West[3], Kelli Montgomery[3], Jen-Tsan Chi[2], Matt van de Rijn[3], David Botstein[4¤2], Patrick O. Brown[2,5*]

Previous examples: multiple datasets of the same type of studies (prognosis in breast cancer patients)

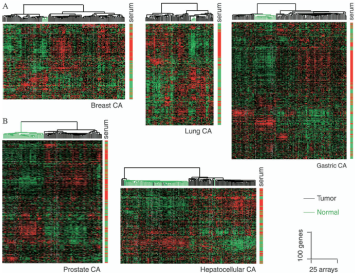
This example: results from cell-line experiment (fibroblast challenged by serum) were compared against array data from various tumors (breast, lung, liver, prostate)

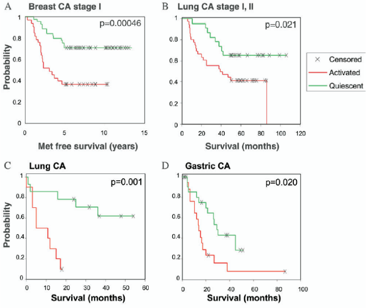
Example III: Cell-line Experiment



- Identify differential expression in fibroblast (0.1% vs 10% serum)
- Subtract non-interesting cell-cycle genes; use dataset from Whitfield (2002) *Mol Biol Cell* **13**:1977 to define "cell cycle genes"
- Compare with data from temporal study

Example III: Connection with Cancer





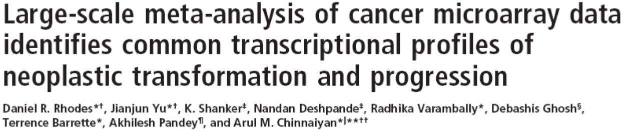
Again, cancer datasets are from already existing studies

The cancer data “annotate” the experimental results

The connection with cancer data adds *clinical importance* to the experimental cell-line study

Example IV

Rhodes (2004) PNAS 101:9309



Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression

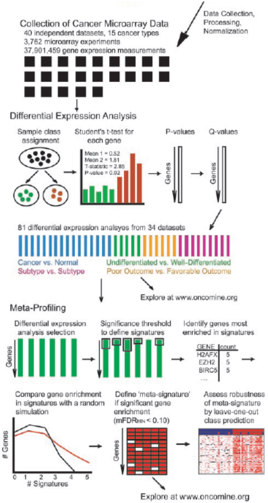
Daniel R. Rhodes[*†], Jianjun Yu[*†], K. Shanker[‡], Nandan Deshpande[‡], Radhika Varambally[*], Debashis Ghosh[§], Terrence Barrette[*], Akhilesh Pandey[¶], and Arul M. Chinnaiyan[*‖**††]

Previous examples: use external datasets to confirm or to relate to the results of an expression study

This example: use all datasets in a large collection to identify a new gene signature

Problem: How to analyze incompatible platforms?
Meta-analysis ⇒ combine summary results

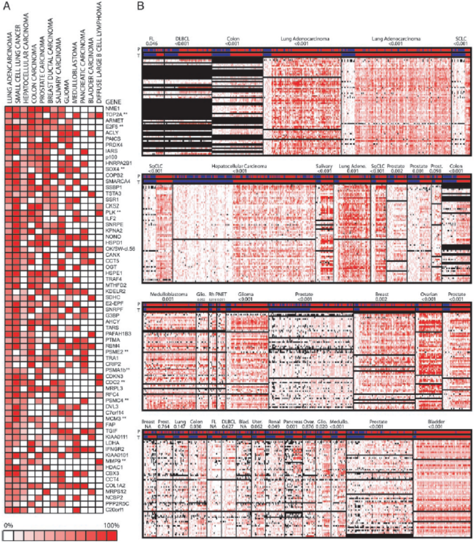
Example IV: Analysis Pipeline


Collection of Cancer Microarray Data
40 independent datasets, 15 cancer types
3,762 microarray experiments
37,901,459 gene expression measurements
Data Collection,
Processing,
Normalization
Differential Expression Analysis
Sample class assignment
Student's t-test for each gene
Mean 1 = 0.52
Mean 2 = 1.81
T-statistic = 2.85
P-value = 0.02
P-values
Q-values
Genes
Genes
81 differential expression analsyes from 34 datasets
Cancer vs. Normal
Subtype vs. Subtype
Undifferentiated vs. Well-Differentiated
Poor Outcome vs. Favorable Outcome
Explore at www.oncomine.org
Meta-Profiling
Differential expression analysis selection
Significance threshold to define signatures
Identify genes most enriched in signatures
Genes
GENE
count
H2AFX
5
EZH2
5
BIRC5
5
Compare gene enrichment in signatures with a random simulation
Define 'meta-signature' if significant gene enrichment (mFDR_{MN} < 0.10)
Assess robustness of meta-signature by leave-one-out class prediction
# Genes
# Signatures
0
1
2
3
4
5
Genes
Explore at www.oncomine.org

- Collect large number of publicly available cancer datasets (15 cancer types, 40 datasets, 3762 arrays)
- Different types of comparisons (cancer vs normal, good vs poor outcome, etc.), but all are related to cancer progression
- For each dataset, find genes differentially expressed (use t-test, find Q-value)
- Choose genes passing significance threshold of Q-value (< 0.10)
- For each gene, count the number of times it is significant
- Rank the genes based on this count
- Use permutation test to assess significance of the ranking

⇒ This is a big, industrial-scale operation

Example IV: Meta Signature



panel A:
How often a gene is significant across datasets

panel B:
How well the meta signature predict the classes

Why analyzing multiple datasets?

- Increase statistical power

  Datasets from similar studies ⇒ hypothesis tests with a *larger sample size*

- Validate results independently

  Note that cross-validation and multiple-testing correction are *internal checks* to control analysis procedure. They can not control *intrinsic biases* (e.g. due to study design or confounding variables).

- Highlight consistent relationships

  Select genes behaving the same way in many datasets

- Extend biological insights

  Comparison of different types of studies (e.g. cell line experiment and tumors from patients)

Multi-dataset analyses are not trivial tasks; many people still have not fully exploited the potentials

Using Published Microarray Datasets

- What is the question?
  - Add interpretation and/or support to our own data
  - Meta-analysis: don't have data, just want to re-analyze existing ones (such as in example IV)
- How do we know relevant datasets exist?
  - Pointed by journal articles
  - Survey of major microarray data repositories
- Data preparation
  - Download
  - Data clean-up and reformatting (still largely manual!)
  - Mapping probes to genes
- Statistical analysis

Components of a Microarray Dataset

Primary information:

- Description about the general study design
- Experimental conditions or clinical data (subject annotation)
- Description about the microarray platforms (probe annotation)
- The expression data themselves (raw or normalized data)

Derived information:

- Analysis results of original authors, such as clustering dendograms, gene list (signatures), subject classifications, ...

Where the datasets may be found?

- Original author's website (URL in journal article)
- Journal article's supplementary materials
- Public repositories:
  - GEO (Gene Expression Omnibus) `[www.ncbi.nlm.nih.gov/geo]`
  - ArrayExpress `[www.ebi.ac.uk/arrayexpress]`
  - Stanford Microarray Database `[genome-www5.stanford.edu]`
- Third-party curators, e.g.
  - Oncomine `[www.oncomine.org]`
  - CleanEx `[www.cleanex.isb-sib.ch]`

Issues in Data Collection and Preparation

- Comprehensive survey of what might be relevant and available

  ⇒ Traditional literature reviews, scanning table of contents of GEO, ArrayExpress, Oncomine, . . . (and of course, Google)

- Parts of the same dataset may be in different places

  e.g., clinical tables are in supplementary materials of several related articles (but only expression data is in GEO).

- Not all parts of a dataset are available

  e.g. Expression data are in GEO or ArrayExpress, in order to be MIAME compliant, but clinical data are not available anywhere or incomplete

- Manual data clean-up, reformatting, standardizing names and values, etc. are required (and tedious!)

- Our knowledge of the transcriptome is still evolving ⇒ probe mapping to genes needs to be regularly updated

Some Solutions

- Ongoing efforts by the community (e.g. MGED Society) to streamline the process by making the data structure more "explicitly semantic" by controlled vocabularies, data schema, and stricter requirements for publishing

  ⇒ This a complex problem! We are not there yet...

- Do-it-yourself

  Needs specialist (particularly for larger tasks). Projects often focus on studies of immediate relevance (e.g. breast cancer data only)

- Third-party projects to provide curated data

  Oncomine, but public access are limited: no complete access to data matrices; only results of pre-specified analysis types are available. (Still useful! e.g. the table of contents, simple questions)

  CleanEx provides updated probe mapping via Unigene

Statistical Analyses of Multiple Datasets

What to do on the external datasets depends on the problem:

- Confirming the action of a few genes

  No need to analyze the whole dataset; just get the relevant genes and conditions/phenotypes. In particular, multiple testing of the whole array is not needed (but of course need to done over multiple genes in question)

- Validating signatures

  Take only relevant genes; cross-validation needed if parameters is re-estimated. More powerful than fresh signature identification, because no need for feature selection.

- New discoveries (e.g. meta-analysis)

  All genes in all datasets need to be considered; multiple testing or cross-validation over *combined* results

Approaches to Combined Analysis

- Combine raw data (after some kind of normalization)

  Advantages: treat as a single dataset; existing software directly applicable

  Disadvantages: statistical validity: samples are intrinsically stratified (correlated within studies) instead of independent, identically distributed, may results in strange effect (Simpson's paradox) or loss of power

- Combine final results (e.g. Venn diagram)

  Advantages: no new analysis, just put together existing results

  Disadvantages: suboptimal conclusions (loss of power), due to premature thresholding

- Combine intermediate summary statistics (meta-analysis)

  Advantages: stratified analysis (statistically sound); good literature for hypothesis testing (i.e., for differential expression)

  Disadvantages: access to raw data needed; maybe less powerful when the first approach is applicable; not clear what to do in PCA/clustering or classification/prediction

Meta-Analysis

Several types of summary statistics to combine:

- Model parameters (regression coefficients); e.g. fold-change
- Effect size measures; e.g. signal-to-noise ratio, correlation
- Significance test statistics; e.g. Z-scores, χ^2, p-values

These are combined for each gene; ranking genes is done *afterward*, using the combined scores

Further reading:
Hedges and Olkin (1985) Statistical Methods for Meta-Analyis [the classic!]

Choi (2003) Bioinformatics 19sup1:84
Ghosh (2003) Funct Integr Genomics 3:180
Gentleman (2005) J Société Française de Statistique 146:1

Combining p-values

- The inverse normal method (Z-score):

  $$\bar{Z} = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}}$$

  Note $Z_i = \Phi^{-1}(p_i)$ (or `zi = qnorm(pi)` in R)
  Z_i is standard normal under the null hypothesis

- The inverse chi-squared method (Fisher's method)

  $$X^2 = -2 \sum_{i=1}^{k} \log p_i$$

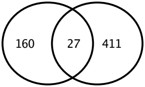
  X^2 is χ^2_{2k} under the null hypothesis

Advantage of the first one: the sign of effect (up- or down-regulation) is taken care of (strong Z_i's with opposite sign will cancel).

In Fisher's method, one-side tests has to be used. The method is applied separately for up- and down-regulation.

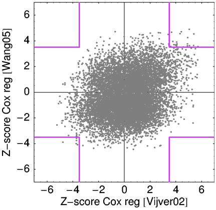
An example: relapse in breast cancer

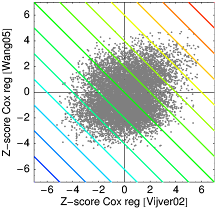
The Z-score of Cox regression summarizes association between expression and cancer metastasis (correspond to p-value); analogous to t-statistics in normal model



Venn diagram

Meta-analysis using "inverse normal" $\sum_{i=1}^{K} Z_i/\sqrt{K}$





Conclusions

- Integrative genomic data analysis is still evolving (not just for expression microarray; more genomic data types are coming)
- In principles, it is a straightforward idea and with examples of useful results

  In practice, there are still statistical issues and technical problems
- The aim here is to give a "taste" of this field

Afternoon practicals:

- Exploring GEO database
- Meta-analysis of differential expression