Genomic Biomarkers for Depression: Feature-specific and Joint Biomarkers

Helena Geys, Dan Lin, Ziv Shkedy, Abel Tilahun, Luc Bijnens, Willem Talloen

hgeys@its.jnj.com

Janssen Pharmaceutical Companies

Research and Development

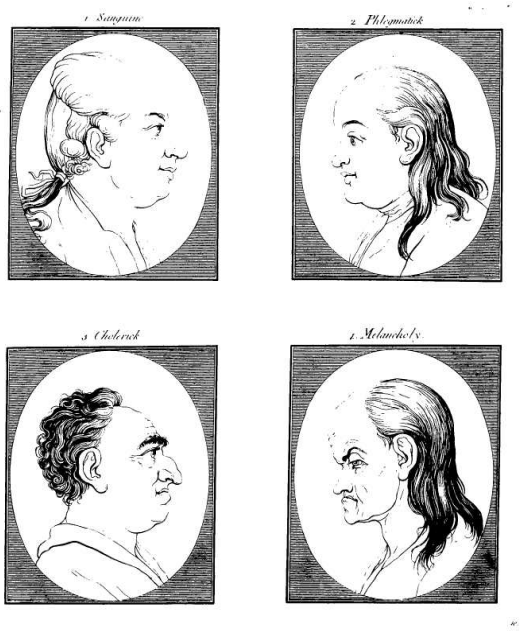
Nonclinical Statistics

Beerse, Belgium

Depression: A Historical Tour

Depression has been with mankind since the beginning of recorded history:

- Ancient Greece: disease defined by imbalance in the four body fluids or “humors": sanguine, phlegmatic, choleric, melancholic

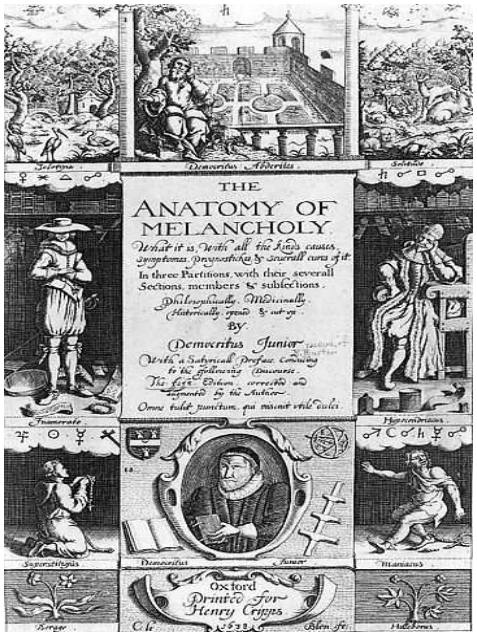


Depression: A Historical Tour

- Hippocrates referred to depression as melancholia: a disease with particular mental and physical symptoms
- Much broader than the current notion of depression
- Since Aristotle, melancholia had been associated with men of learning and intellectual brilliance, a hazard of contemplation and creativity.
- The 11th century Persian physician Avicenna described melancholia as a depressive type of mood disorder in which the person may become suspicious and develop certain types of phobias. His work, *The Canon of Medicine*, became the standard of medical thinking in Europe

Depression: A Historical Tour

- In the 17th century, Burton (*Anatomy of Melancholy*) suggested that melancholy could be combated with a healthy diet, sufficient sleep, music, and "meaningful work", along with talking about the problem with a friend.



Depression: A Historical Tour

- In the 19th century, the term “depression" (deprimere: to press down) became in use and appeared in medical dictionaries to refer to a physiological and metaphorical lowering of emotional function. The newer concept abandoned the associations of the ancient Greeks and through the 19th century, became more associated with women.

Depression Today

- Most experts agree on the following:
  - *A depressive disorder is a syndrome that reflects a sad and/or irritable mood exceeding normal sadness or grief.*
- Depression disorders today are a huge public-health problem:
  - It affects about 10% of adults, 8% of teens, 2% of preteens.
  - They involve huge amounts of direct costs (treatment) and indirect costs (loss of productivity)
  - They can increase risks for developing coronary artery disease, HIV, asthma, . . .

Causes of Depression

- Depression is not something you can just “snap out".
- It is caused by an imbalance of brain chemicals that carry signals in your brain and nerves.
- These chemicals are called neurotransmitters.

Causes of Depression

- Not a lot is known about the effects of neurotransmitters
- This is due to the fact that they are difficult to study:
- Neurotransmitters:
  - are present in very small quantities
  - in certain locations within the brain
  - disappear very quickly once used
  - and can therefore not be measured directly
- Researchers can only measure what is left over after their use in the brain:
- So-called metabolites that can be found in blood, urine, cerebrospinal fluid

HAMD: A Measure of Depression

The Hamilton Depression Scale (HAMD):

- Multiple choice questionnaire (ca. 21 questions)
  - depressed mood
  - feelings of guilt
  - insomnia
  - anxiety
  - agitation
  - weight loss
  - . . .
- Test measuring the severity of depressive symptoms in individuals
- Considered the ”Gold Standard"
- (However increased criticism that it is flawed as a test instrument)

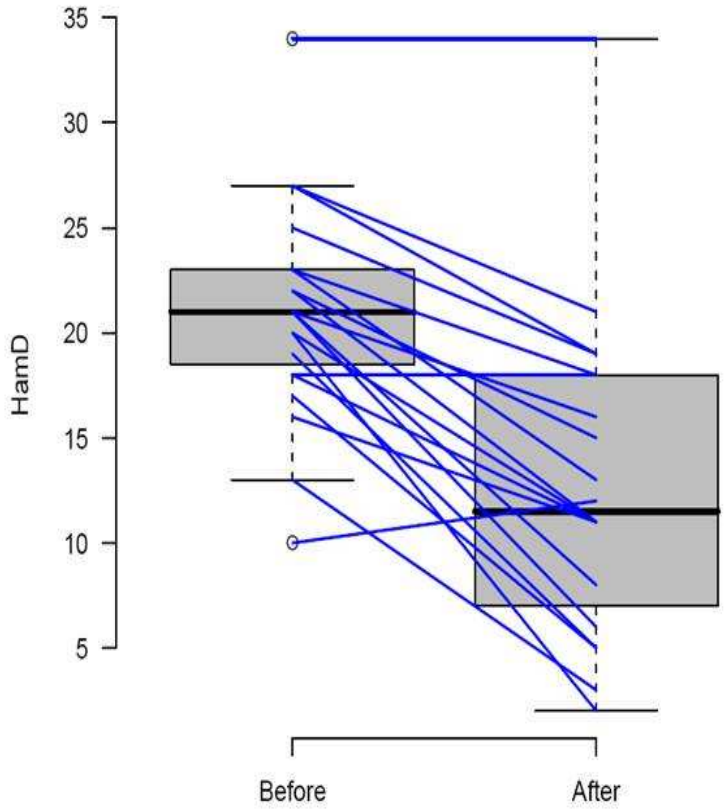
Goals and Objectives of the Current Research

- Can we identify genomic biomarkers, based on gene and metabolite expressions, for depression?
  - Such biomarkers could be applied to develop and guide more efficient drug development and testing programs.
- Objective 1: select and evaluate a set of genes and metabolites as possible biomarkers for depression
  - as measured by the HAMD score
- Objective 2: construct a joint biomarker that combines relevant information from a subset of genes or metabolites, to predict the HAMD score.
- The data: result of clinical trial in which both HAMD and gene/metabolite expressions were measured

The Data

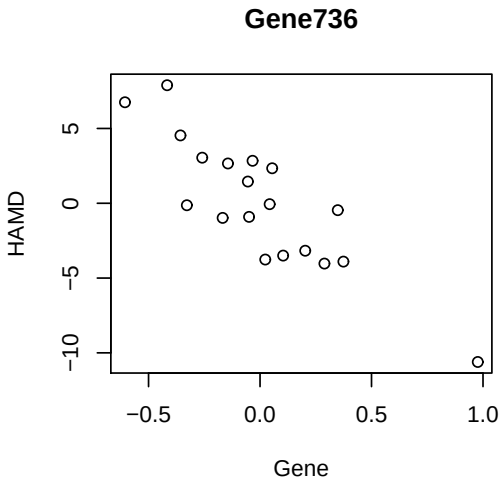
- 15 control individuals followed up 4-6 weeks after first visit
- 31 patients followed up 4-6 weeks after start of treatment with ADs
- Clinical outcome: Hamilton Depression Scale (HAMD)
- Complete information available on 14 patients for the analysis of metabolites (269 metabolites)
- Complete information available on 19 patients for the analysis of gene expression (17,502 genes)
- Other covariates: storage time of the samples, age, gender, season, fasted, . . .

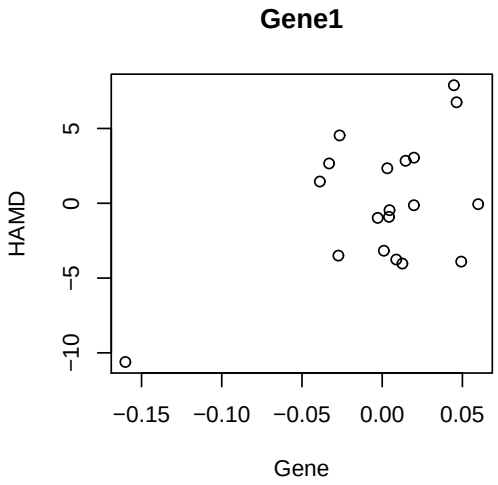
The Data

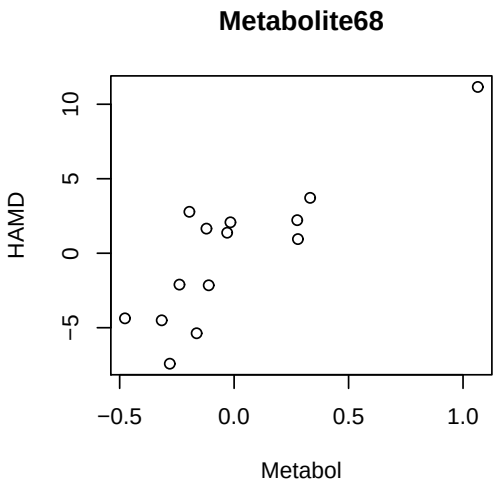


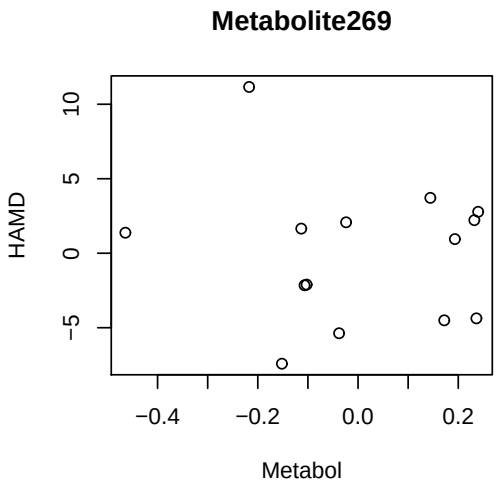
- Response of Primary Interest: Change from baseline HAMD score

The Data


Gene736
HAMD
5
0
−5
−10
−0.5
0.0
0.5
1.0
Gene


Gene1
HAMD
5
0
−5
−10
−0.15
−0.10
−0.05
0.00
0.05
Gene


Metabolite68
HAMD
10
5
0
−5
−0.5
0.0
0.5
1.0
Metabol


Metabolite269
HAMD
10
5
0
−5
−0.4
−0.2
0.0
0.2
Metabol

Joint Model for Gene Expression and HAMD

- X_{ij}: change from baseline of the jth gene expression ($j = 1, \ldots, m$) of the ith subject ($i = 1, \ldots, n$)
- Y_i: change from baseline of HAMD score
- Gene-specific Joint Model:

$$\begin{aligned} E(X_{ij}|\boldsymbol{Z}_i) &= \boldsymbol{Z}_i\boldsymbol{\alpha}_j, \quad j = 1, \ldots, m\,; \;\; i = 1, \ldots, n, \\ E(Y_i|\boldsymbol{Z}_i) &= \boldsymbol{Z}_i\boldsymbol{\beta}. \end{aligned} \tag{1}$$

  fitted for each gene separately

- $$\begin{pmatrix} X_{ij} \\ Y_i \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} \boldsymbol{Z}_i\boldsymbol{\alpha}_j \\ \boldsymbol{Z}_i\boldsymbol{\beta} \end{pmatrix}, \Sigma_j = \begin{pmatrix} \sigma_{jj} & \sigma_{jY} \\ \sigma_{jY} & \sigma_{YY} \end{pmatrix} \right). \tag{2}$$

Joint Model for Gene Expression and HAMD

- Adjusted association (Buyse and Molenberghs 1998):

$$\rho_j = \frac{\sigma_{jY}}{\sqrt{\sigma_{jj}\sigma_{YY}}}. \tag{3}$$

- ρ_j can be equal to 1 even if the gene is not differentially expressed; to select genes which can predict the response, there is no need for the gene to be differentially expressed.

Information-theory Approach

- How to quantify association if we move away from normal distribution?
- Alonso et al. (2005,2007): Likelihood Reduction Factor:

$$R^2_{hj} = 1 - \exp\left(\frac{-G^2}{n}\right), \tag{4}$$

where G^2 denotes the likelihood ratio statistics to compare models:

$$\begin{aligned} E(Y_i) &= \boldsymbol{Z}_i\boldsymbol{\beta}, \\ E(Y_i|X_{ij}) &= \boldsymbol{Z}_i\boldsymbol{\beta} + \gamma_j X_{ij}. \end{aligned} \tag{5}$$

with γ_j the gene-specific effect upon the outcome.

Four-Variate Model

- Instead of taking changes from baseline, the association can also be quantified using the pre and post measurements per subject for gene/metabolite expression and HAMD:

$$\begin{pmatrix} X_{ij0} \\ X_{ij1} \\ Y_{i0} \\ Y_{i1} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{Z}_i\boldsymbol{\alpha}_{0j} \\ \boldsymbol{Z}_i\boldsymbol{\alpha}_{1j} \\ \boldsymbol{Z}_i\boldsymbol{\beta}_{0j} \\ \boldsymbol{Z}_i\boldsymbol{\beta}_{1j} \end{pmatrix}, \Sigma_j = \begin{pmatrix} \Sigma_{XX} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{YY} \end{pmatrix} \right), \quad (6)$$

- Measure of Association (gene-specific):

$$R^2_{\Lambda j} = 1 - \frac{|\Sigma_j|}{|\Sigma_{YY}| \cdot |\Sigma_{XX}|}. \quad (7)$$

Results for top 5 genes

Gene Id	R^2	adj$_p$
736	0.7579	0.0365
2419	0.7295	0.0426
3455	0.6536	0.1553
9859	0.6507	0.1553
8427	0.5910	0.3142

- Two genes found significant
- all patients treated: no need for treatment adjustment
- treatment effect accounted for through differences from baseline
- different genes selected using R^2_Λ
- R^2_Λ quantifies the association between the vector or pre/post HAMD measurements and pre/post gene expressions

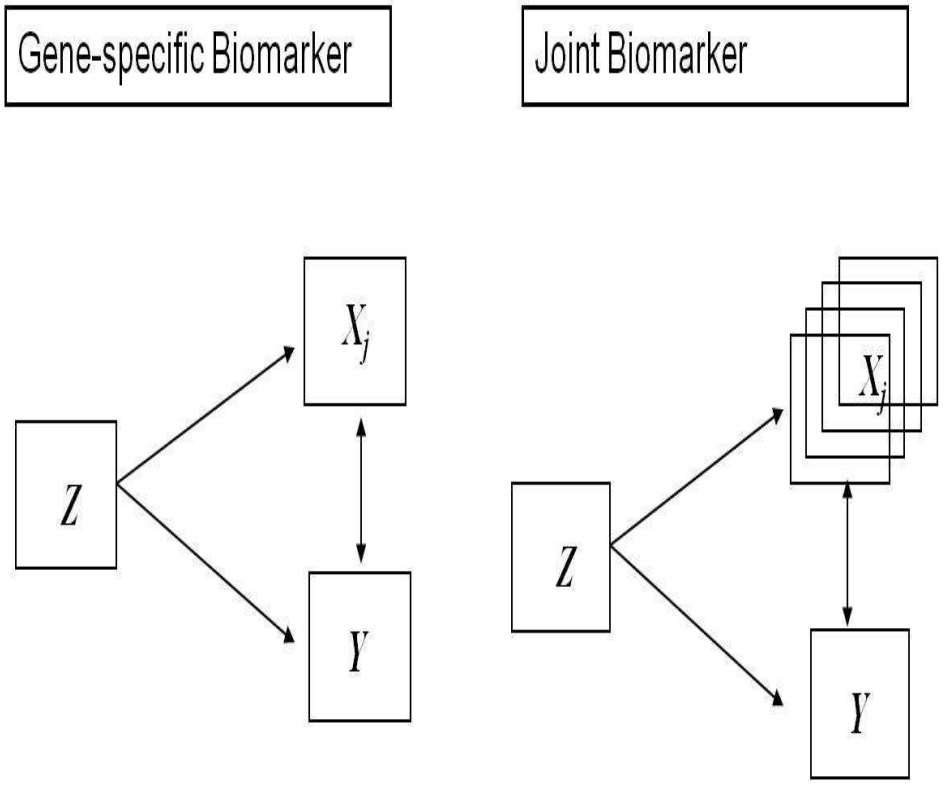
Results for top 5 metabolites

Gene Id	R^2	adj$_p$
68	0.7256	0.0317
12	0.6466	0.0489
67	0.6400	0.0489
255	0.5516	0.1302
21	0.5312	0.1302

- Three metabolites found significant

Joint Biomarker

How to combine information about expression levels from all genes/metabolites in the array into one variable?



Supervised Principal Component Analysis (SPCA)

- Basic Assumption: There exists a latent variable U(X) that is associated with the response variable Y.
- Number of genes larger than number of observations, therefore: data reduction
- Reduce dimension of X giving most weight to those genes that have the strongest relationship with the response

Supervised Principal Component Analysis (SPCA)

Steps:

1. Fit one of the gene-specific models and estimate the association measure
2. Form a reduced expression matrix consisting of only those genes whose gene-specific association measure exceeds a threshold level (or top k)
3. Let X_R be the reduced matrix.
4. Use the first principal component in a regression model to predict the response.
5. Assess the association using the methods described above

Results SPCA

	Genes		Metabolites	
Top	R^2	p-value	R^2	p-value
2	0.7791	0.2280	0.8229	0.0000
3	0.8253	0.3270	0.8029	0.0000
4	0.8301	0.2980	0.8072	0.0000
5	0.7917	0.3430	0.8342	0.0003

Updated SPCA

1. The association between the HAMD and gene profile may be heavily dependent on the threshold
2. Include a gene in the gene profile only if it gives an increase in the association between the gene profile and the HAMD

Results Updated SPCA (Genes)

- 6 genes were considered in the construction of a gene profile
- $R^2 = 0.8923$ (higher than e.g. taking top 20 at once)
- A permutation-based test did not show a significant correlation however

Results Updated SPCA (Metabolites)

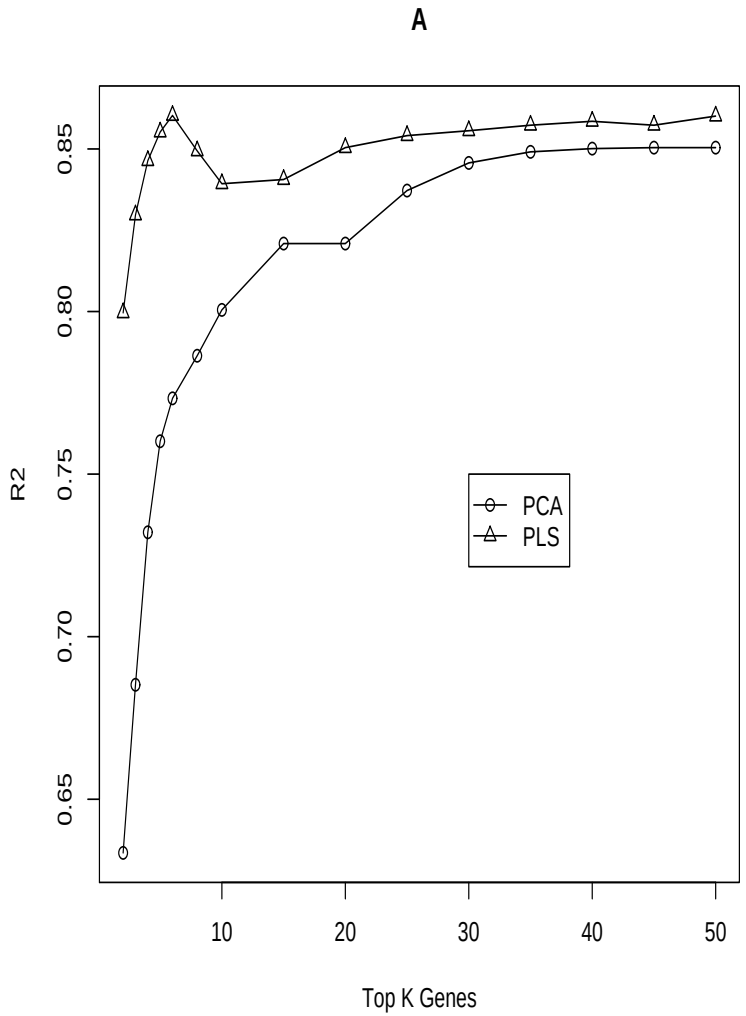
- 9 genes were considered in the construction of a metabolite profile
- $R^2 = 0.8923$ (higher than e.g. taking top 20 at once)
- Significant association between the joint metabolite biomarker and the response

Supervised Partial Least Squares (SPLS)

Steps:

1. Fit one of the gene-specific models and estimate the association measure
2. Form a reduced expression matrix consisting of only those genes whose gene-specific association measure exceeds a threshold level (or top k)
3. Let X_R be the reduced matrix.
4. Fit a partial least squares regression and take the first factor, $U(R)$ (PLS selects components of X that are also relevant to Y).
5. Select the genes with the largest influence on the resulting latent factor and use them to construct the joint biomarker
6. Assess the association using the methods described above

Results



Results

- A test of significance for the joint biomarker based on the PLS approach revealed that only the joint biomarker based on the top 2,3 or 4 genes is significant
- The joint biomarker involving any number of the top 20 metabolites was significant

Discussion

- The (weighted) PLS approach results in larger associations than the PCA approach for any number of top genes/metabolites
- Different sets of genes/metabolites were deemed important in the construction of a joint biomarker
  - PLS: genes selected such that the correlation between joint biomarker and response is maximized
  - PCA: genes selected based on their contribution to the principal component