



[www.csiro.au](http://www.csiro.au)

# Statistical methods for biomarker selection: method validation with CRC gene expression data

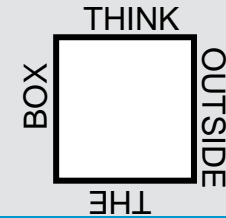
Preventative Health Flagship / CSIRO Mathematics, Informatics & Statistics

**James Doecke**  
**Statistician**  
**Wednesday 13 June 2011**

National Research  
**FLAGSHIPS**  
Preventative Health

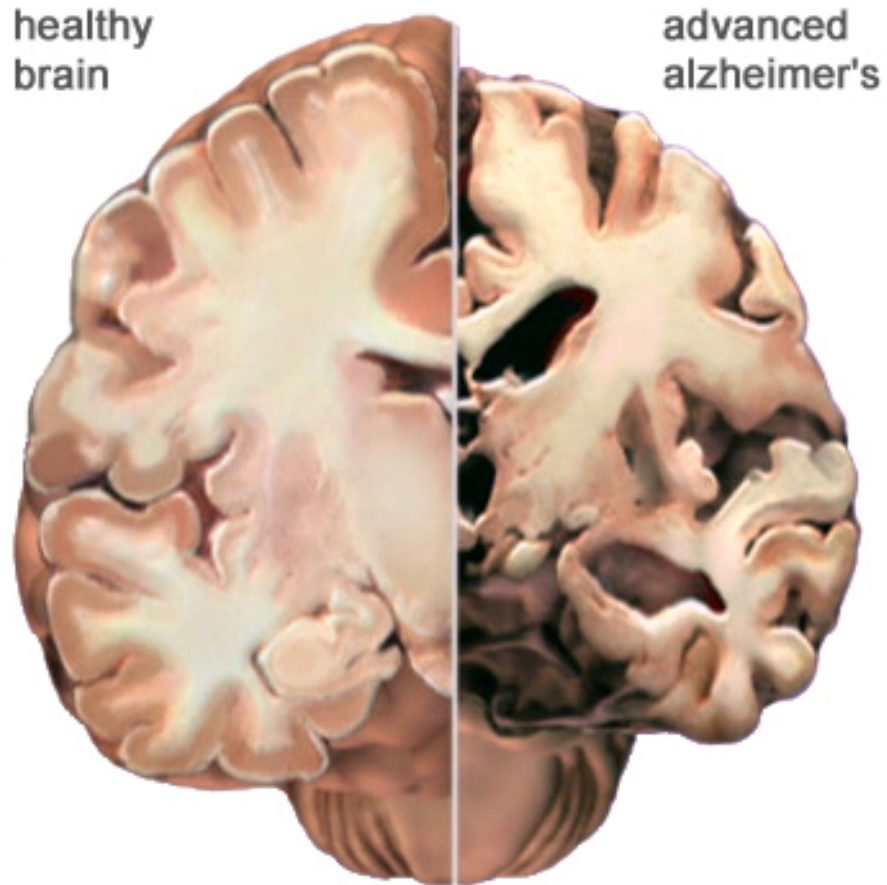
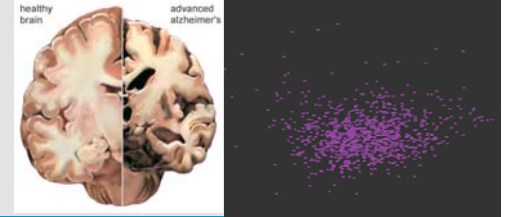


# Talk outline



- Overview of research including:
  - Types of biomarkers
    - SNP markers,
    - gene expression,
    - proteins
  - Sample size
    - Traditional gene expression  $P \gg N$
    - New protein studies where  $P \sim N$  or  $P < N$
  - Possible approaches
    - One method?
    - Two?
    - Multiple?
  - My research project
  - Sparsity based upon size of  $P$

# Background: Biomarkers



Statistical methods for biomarker selection

# Biomarker samples



SomaLogic *Unlocking Biomarker Discovery*

[Search](#) [Contact](#) [Email](#)

• Re  
pre  
se

• Ne  
of

• As  
Ru  
bio

• Th

• Cu

**RBM**  
RULES BASED MEDICINE innovative biomarker solutions

Home Products & Services Order Scientific Literature Events Company Data Quality VeriPsych Sign Up For Our Newsletter

**HumanMAPs**

Overview

- Human DiscoveryMAP® 250+
- Human DiscoveryMAP
- Human OncologyMAP
- HumanMAP
- Human PsyMAP
- Human CardiovascularMAP
- Human InflammationMAP
- Human MetabolicMAP
- Human CytokineMAP A
- Human CytokineMAP B
- Human KidneyMAP

RodentMAPs

Therapy Indications

Custom MAP

TruCulture

Strategic Biomarker Services

**HumanMAPs**

**Human DiscoveryMAP® 250+ v 1.0**

Building on RBM's DiscoveryMAP® service, the DiscoveryMAP® 250+ covers dozens of biochemical pathways with over 250 quantitative immunoassays. This service allows customers to cast the widest net in RBM's history to further bolster the discovery of new biomarker patterns across multiple therapeutic indications.

DATA QUALITY

HOW TO ORDER

HumanMAP Brochure

Sample Volume Requirements

Click or mouse over biomarkers for more information

1. 6Ckine	2. Adiponectin
3. Adrenocorticotrophic Hormone	4. Agouti-Related Protein
5. Aldose Reductase	6. Alpha-1-Antichymotrypsin
7. Alpha-1-Antitrypsin	8. Alpha-1-Microglobulin
9. Alpha-2-Macroglobulin	10. Alpha-Fetoprotein
11. Amphiregulin	12. Angiogenin
13. Angiotensin-2	14. Angiotensin-Converting Enzyme
15. Angiotensinogen	16. Annexin A1
17. Apolipoprotein A-I	18. Apolipoprotein A-II
19. Apolipoprotein A-IV	20. Apolipoprotein B
21. Apolipoprotein C-I	22. Apolipoprotein C-III
23. Apolipoprotein D	24. Apolipoprotein E
25. Apolipoprotein H	26. Apolipoprotein(a)
27. AXL Receptor Tyrosine Kinase	28. B cell-activating factor

levels of specific  
disease and/or

the measurement

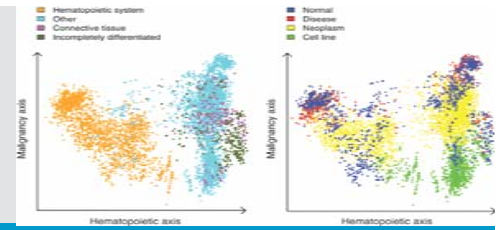
8 months ago,  
as 151

• Cu  
Statistical techniques to deal with the deluge of data.

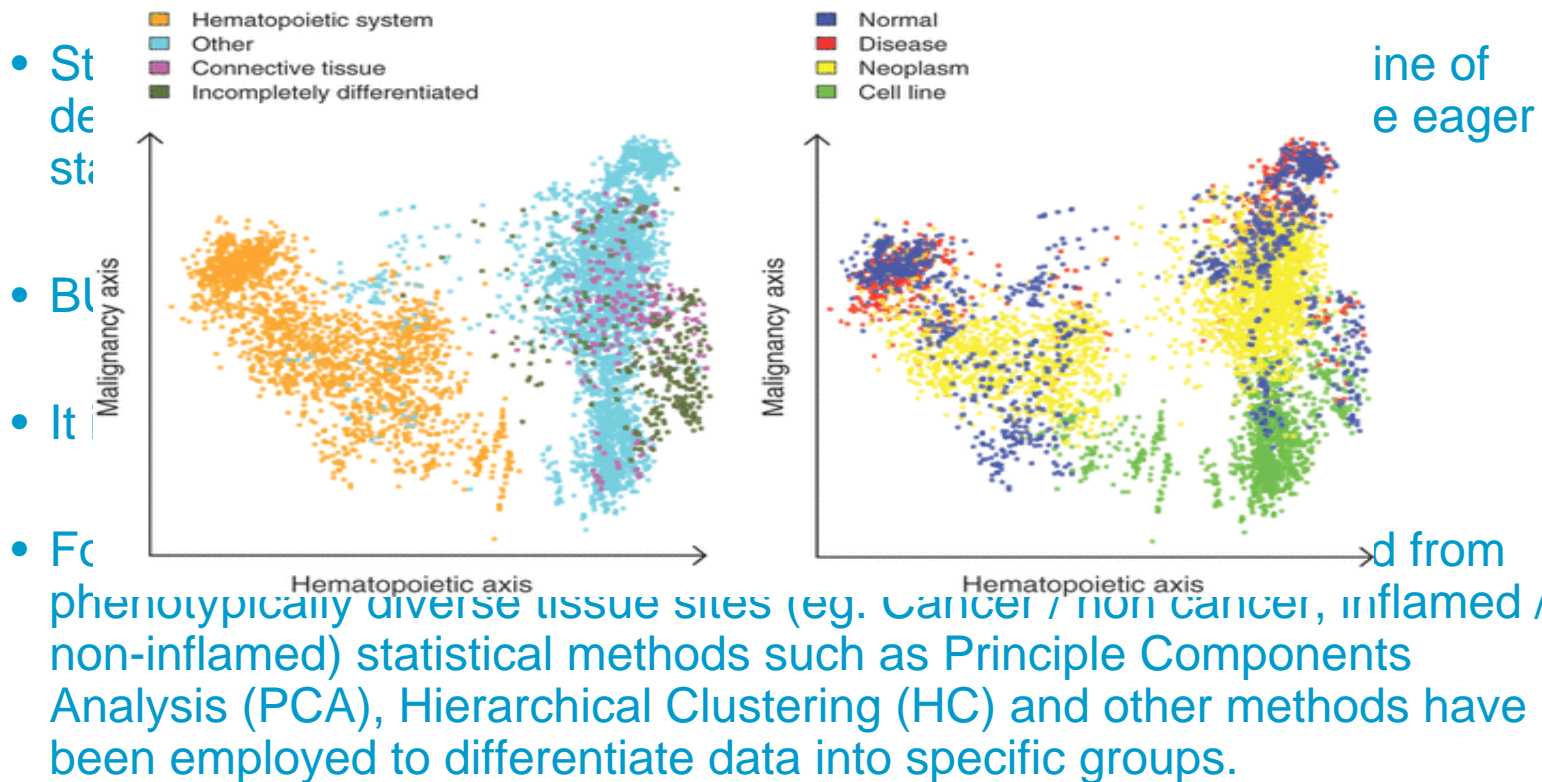
- Robust and reproducible assay
- Sample collection protocol established and

across a wide spectrum of cases and controls collected prospectively from appropriate retrospective cohorts of samples

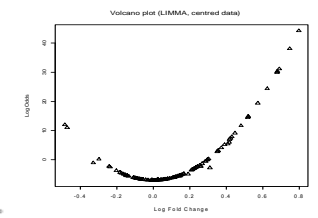
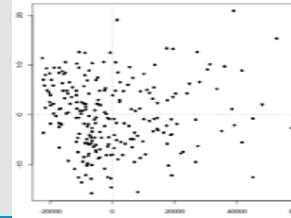
# Possible approaches I



- GWAS and genome wide expression arrays have needed very sparse methods to reduce the dimension of data.

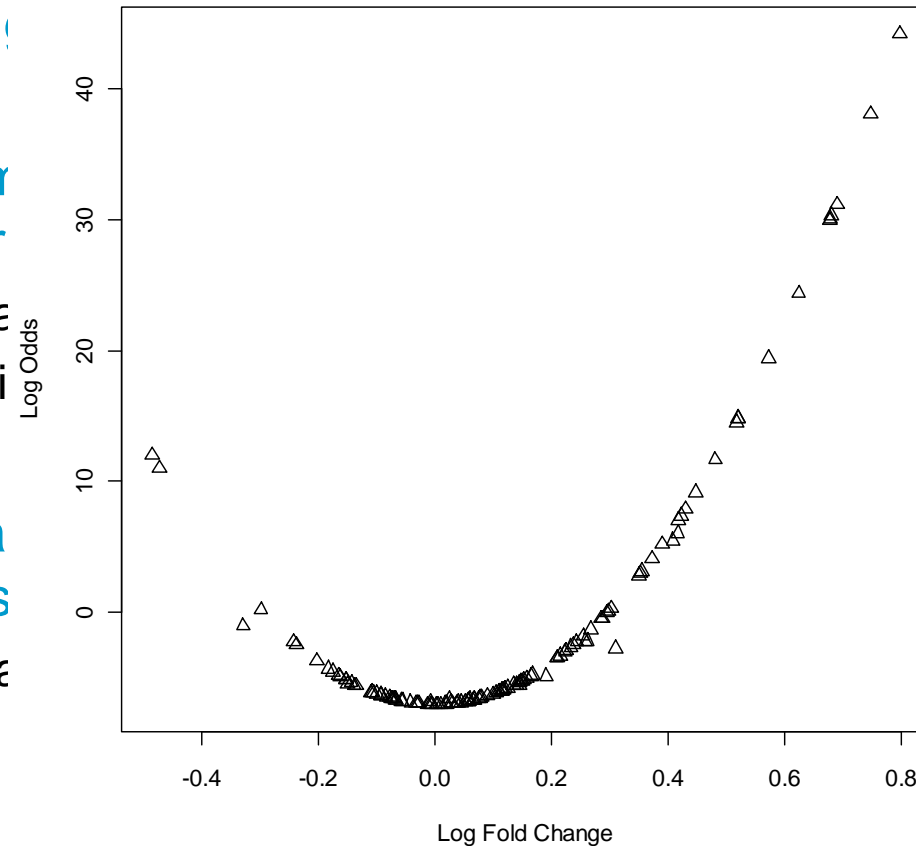


# Possible approaches II



- Protein grouping
- Some methods transfer
  - Linear
  - Significance
- As a baseline
  - Volcano

Volcano plot (LIMMA, centred data)



It's into clear

but however be

methods are

of effect sizes

# Research project: Aim



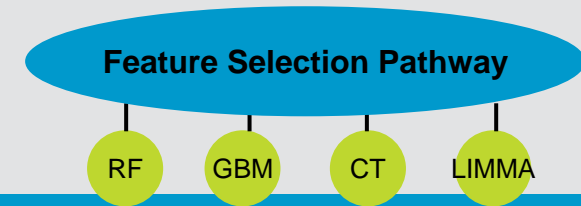
- Design an approach to identify a short list of biomarkers that could predict disease status
- Extend this approach to larger data sets
- Compare approach with large scale sparse regression using a range of penalty scores

# Research project: The data



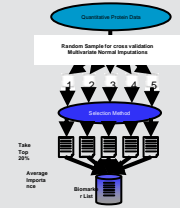
- Data from two Affymetrix Gene Chips; RNA collected from Colorectal Cancer and healthy control tissue specimens
  - ~44,000 probe sets (P)
  - ~400 tissue samples (N)
  - $P \gg N$

# Research project: Methodology



- Chose 4 different methods (interchangeable) to represent a feature selection pathway
- One of these methods was an estimator or effect size comparisons
  - LIMMA
- Other three are publically available and well referenced methods:
  - Random Forest
  - Generalized Boosted Models
  - Classification Trees

# Data analyses

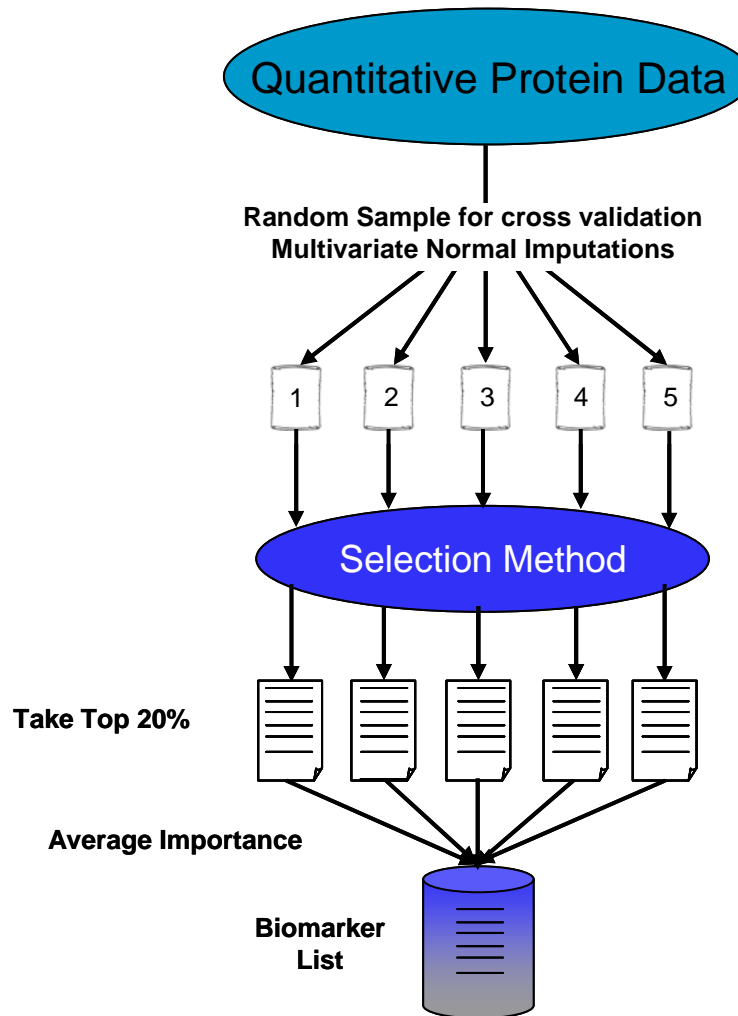


- Using 4 (deterministic) analyses

- Each method important

- I chose to

- Remember each method
- Top 20% statistic

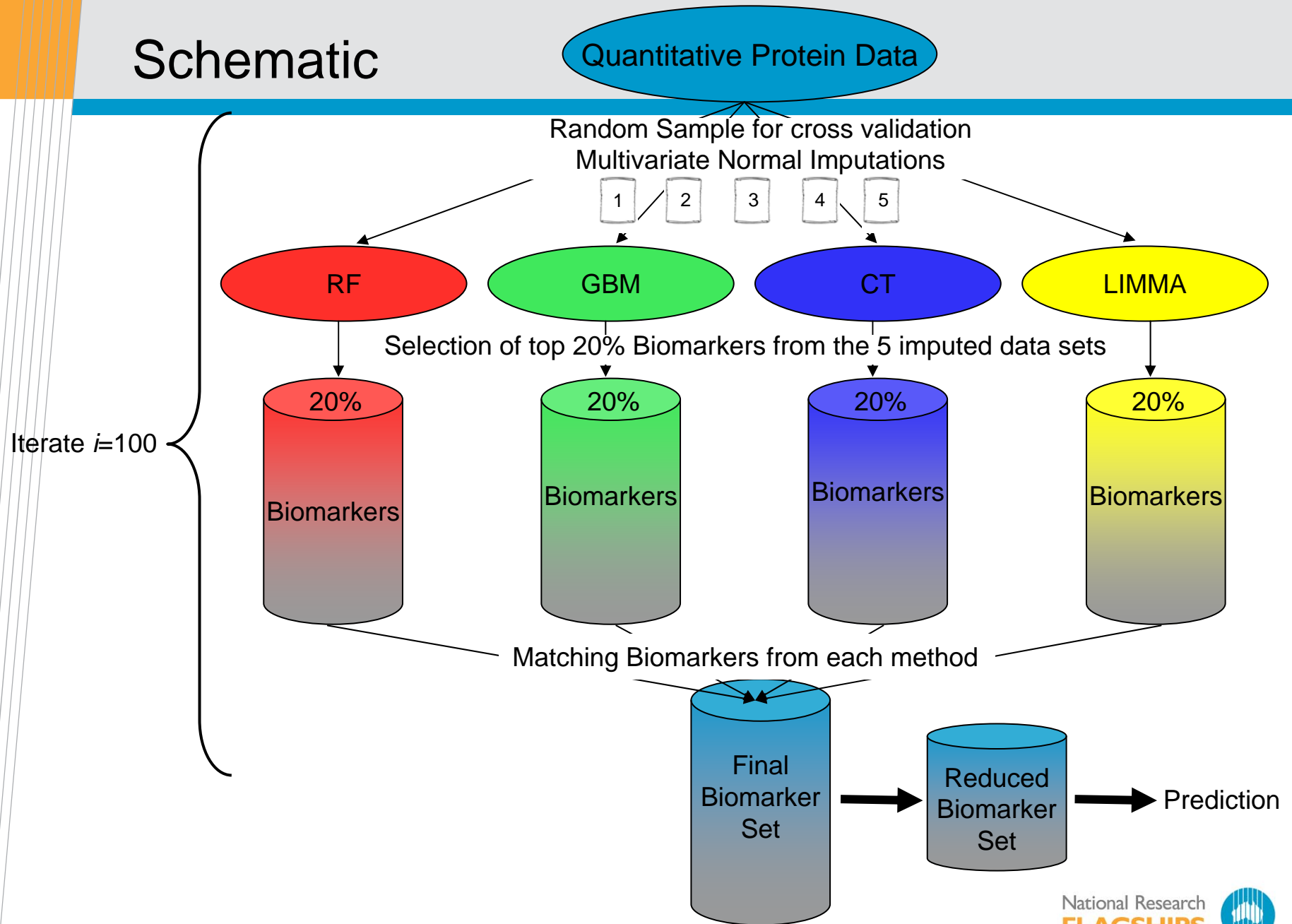


I needed a way of ranking markers from each

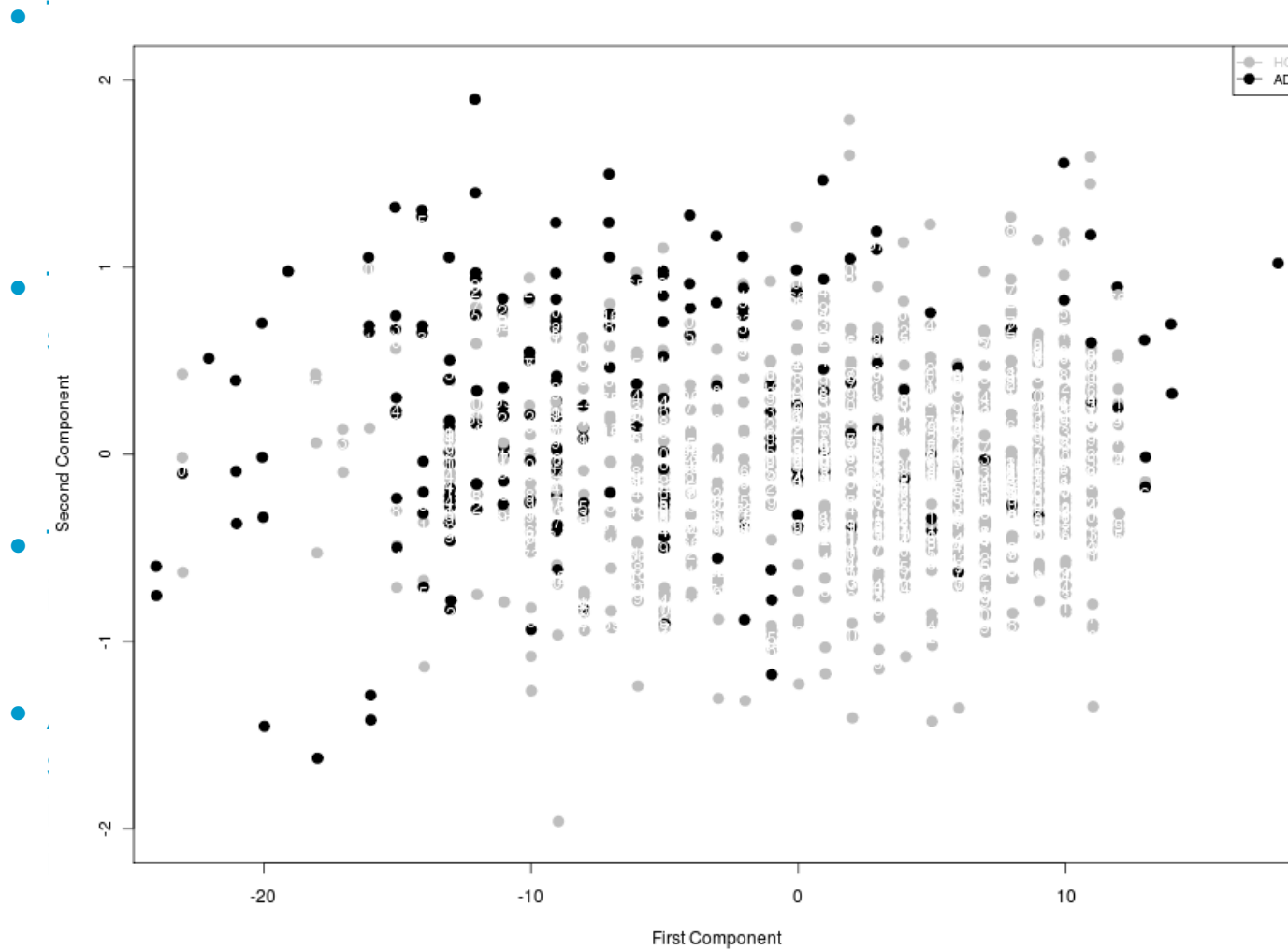
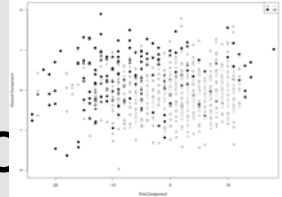
in terms of

from each method. (imputed data sets at 100 repeats) average importance

# Schematic



# Results: Potential benefits and interpretation



he

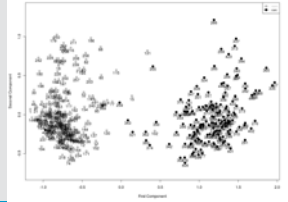
are  
or to  
100

# Suitability?

- The Feature selection pathway was designed with ~200 biomarkers in mind.
- Is it applicable for larger data sets?
- CRC Expression set:
  - ~45K markers
  - Random selection of 200 probe sets
  - Performed feature selection pathway
  - Collated results
  - Random selection of 1000 probe sets
  - Performed feature selection pathway
  - Collated results
- Use pathway for all 45K probe sets



# 45K Markers



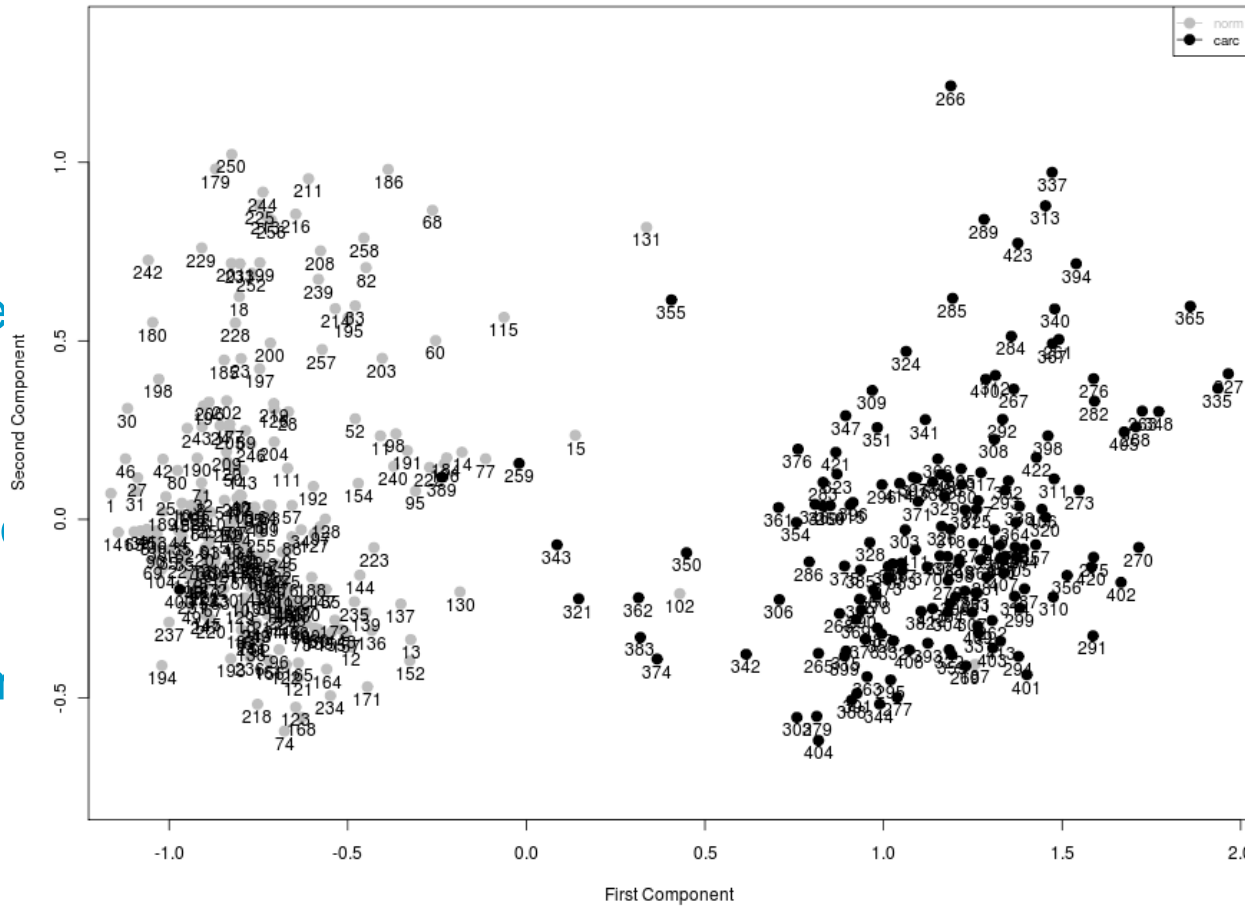
- Tri

- 
- 

- Ide an

- A t

- Tin



se

o, with

# GeneRaVE (GR)



- To validate of different ratios of variables to observations
  - 0 (7)
  - 0.5 (12)
  - 0.75 (6)
  - 1.0 (36)
  - 2 (45,0)
- The class
- Increasingly increasing
- When and samples (standard l

## BMC Bioinformatics



Open Access

Methodology article

### A general approach to simultaneous model fitting and variable elimination in response models for biological data with many more variables than observations

Harri T Kiiveri

Address: CSIRO Mathematical and Information Sciences, The Leeuwin Center, 65 Brockway Road, Floreat, 6014, Western Australia  
Email: Harri T Kiiveri - harri.kiiveri@csiro.au

Published: 15 April 2008

Received: 27 September 2007

Accepted: 15 April 2008

BMC Bioinformatics 2008, 9:195 doi:10.1186/1471-2105-9-195

This article is available from: <http://www.biomedcentral.com/1471-2105/9/195>

© 2008 Kiiveri; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

#### Abstract

**Background:** With the advent of high throughput biotechnology data acquisition platforms such as micro arrays, SNP chips and mass spectrometers, data sets with many more variables than observations are now routinely being collected. Finding relationships between response variables of interest and variables in such data sets is an important problem akin to finding needles in a haystack. Whilst methods for a number of response types have been developed a general approach has been lacking.

**Results:** The major contribution of this paper is to present a unified methodology which allows many common (statistical) response models to be fitted to such data sets. The class of models includes virtually any model with a linear predictor in it, for example (but not limited to), multiclass logistic regression (classification), generalised linear models (regression) and survival models. A fast algorithm for finding sparse well fitting models is presented. The ideas are illustrated on real data sets with numbers of variables ranging from thousands to millions. R code implementing the ideas is available for download.

**Conclusion:** The method described in this paper enables existing work on response models when there are less variables than observations to be leveraged to the situation when there are many more variables than observations. It is a powerful approach to finding parsimonious models for such datasets. The method is capable of handling problems with millions of variables and a large variety of response types within the one framework. The method compares favourably to existing methods such as support vector machines and random forests, but has the advantage of not requiring separate variable selection steps. It is also works for data types which these methods were not designed to handle. The method usually produces very sparse models which make biological interpretation simpler and more focused.

h a range  
ch run).

'sity,

:issue  
using the

# Considerations

- Larger scale feature selection methods need tuning to get the parameters for feature selection accurate.
- However, this is still only one method of selecting features.
- Comes back to the old question, is one method enough?
- Using the statistical pathway discussed here, 3 feature selection methods have been packaged with one effect size comparator.
- Thus the pathway can be tuned with four different statistical packages to either increase or decrease the sparsity.

## Considerations II

- This statistical pathway also provides FDR adjusted p-values for the group comparisons.
- Three listings of features are provided at different levels of specificity, such that there are secondary and tertiary lists of markers to follow up for hypothesis generation.
- Adding increasing numbers of markers aligns with an increase in sparsity.
- Applicable for small to large biomarker discovery data sets
- Especially interesting given the current increase in protein discovery panel molecular architecture.

**CSIRO Mathematics, Informatics and Statistics (CMIS)**

James Doecke  
Statistician

**Phone:** +61 7 3253 3697

**Email:** james.doecke@csiro.au

**Web:** <http://www.csiro.au/org/CMIS.html>

www.csiro.au

Thank you

National Research  
**FLAGSHIPS**  
Preventative Health

