The Armitage Lecture

Some statistics in bioinformatics

Patty Solomon

The University of Adelaide

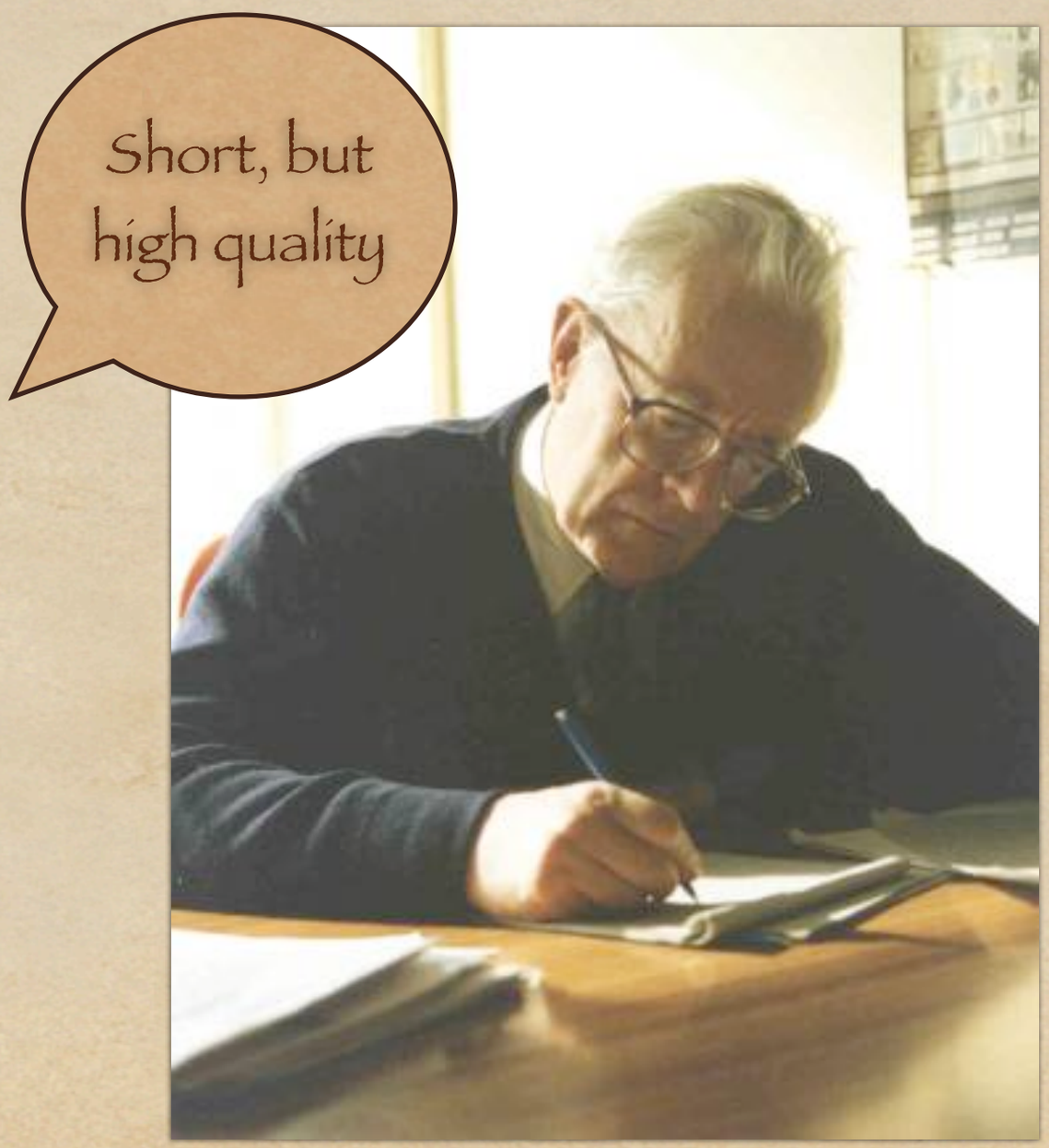
Cambridge 15 November 2007

PROFESSOR PETER ARMITAGE

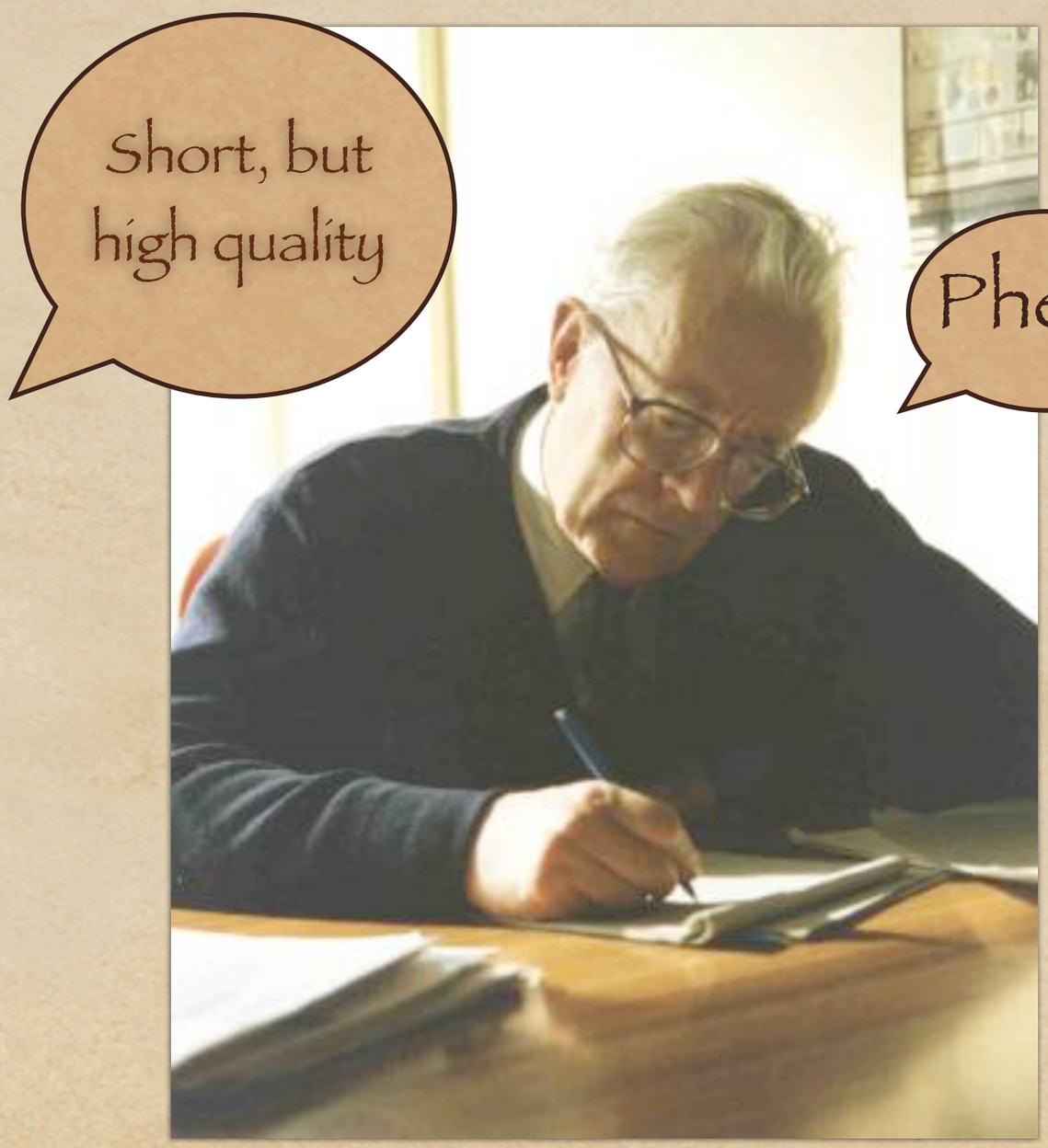
Photo courtesy of Ted Colton

Peter Armitage was the External Examiner for my PhD, and Sir David Cox was my Supervisor

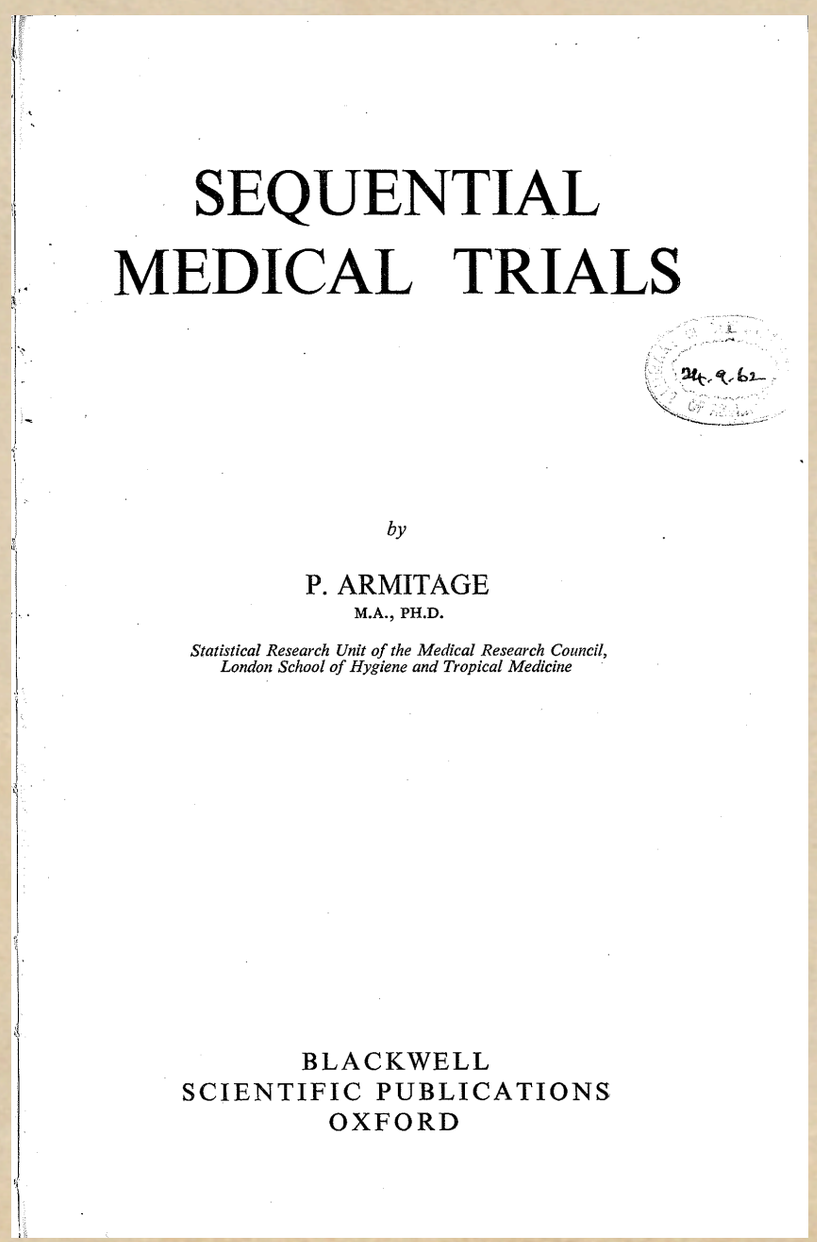
Peter Armitage was the External Examiner for my PhD, and Sir David Cox was my Supervisor



Peter Armitage was the External Examiner for my PhD, and Sir David Cox was my Supervisor



Peter's pioneering 1960 book '*Sequential Medical Trials*' was many, many years ahead of its time



SEQUENTIAL
MEDICAL TRIALS

by

P. ARMITAGE
M.A., PH.D.

Statistical Research Unit of the Medical Research Council,
London School of Hygiene and Tropical Medicine

BLACKWELL
SCIENTIFIC PUBLICATIONS
OXFORD

Motivation:

Post- WWII era of "controlled medical trials to compare the effectiveness of different therapeutic or prophylatic treatments".

PREFACE

It is now widely accepted that the most reliable way to compare the effectiveness of alternative medical treatments is to carry out a controlled trial, in which the treatments are allocated at random. In recent years a number of these clinical trials have been conducted by sequential methods, and there appears to be a growing interest in this development. In a sequential analysis the observations are examined as they become available, and the total number of subjects to enter the trial is not predetermined, but depends on the accumulating results.

Sequential analysis has an immediate appeal in clinical research.

Peter's book (and its 2nd edition in 1975) is a goldmine of information and good advice on experimental design (see next slide) ...

1.2 Comparative experimentation

- "If the same medical treatment is given to a number of patients with the same illness, their apparent responses will inevitably differ. Similar variation in response will appear if the treatment is given repeatedly to the same patient. Any attempt to compare the relative effectiveness of two or more treatments must therefore take into account the unpredictable variation in response from one occasion to another."
- "The problem presents itself in almost every field of biological experimentation, for it is in the nature of biological material to show some degree of unpredictable variability."
- *"It has been generally recognized, since the work of R.A. (now Sir Ronald) Fisher in agricultural research in the 1920's, that a valid comparison can be achieved only by some form of randomization ..."*

Peter's early work on the comparison of survival curves stimulated much later research

Vol. 122.] [Part III

Journal of the Royal Statistical Society

SERIES A (GENERAL)

PART III, 1959.

THE COMPARISON OF SURVIVAL CURVES

By P. ARMITAGE

Statistical Research Unit of the Medical Research Council,
London School of Hygiene and Tropical Medicine

[Read before the ROYAL STATISTICAL SOCIETY, February 18th, 1959, the PRESIDENT, SIR HARRY CAMPION, C.B., C.B.E., in the Chair.]

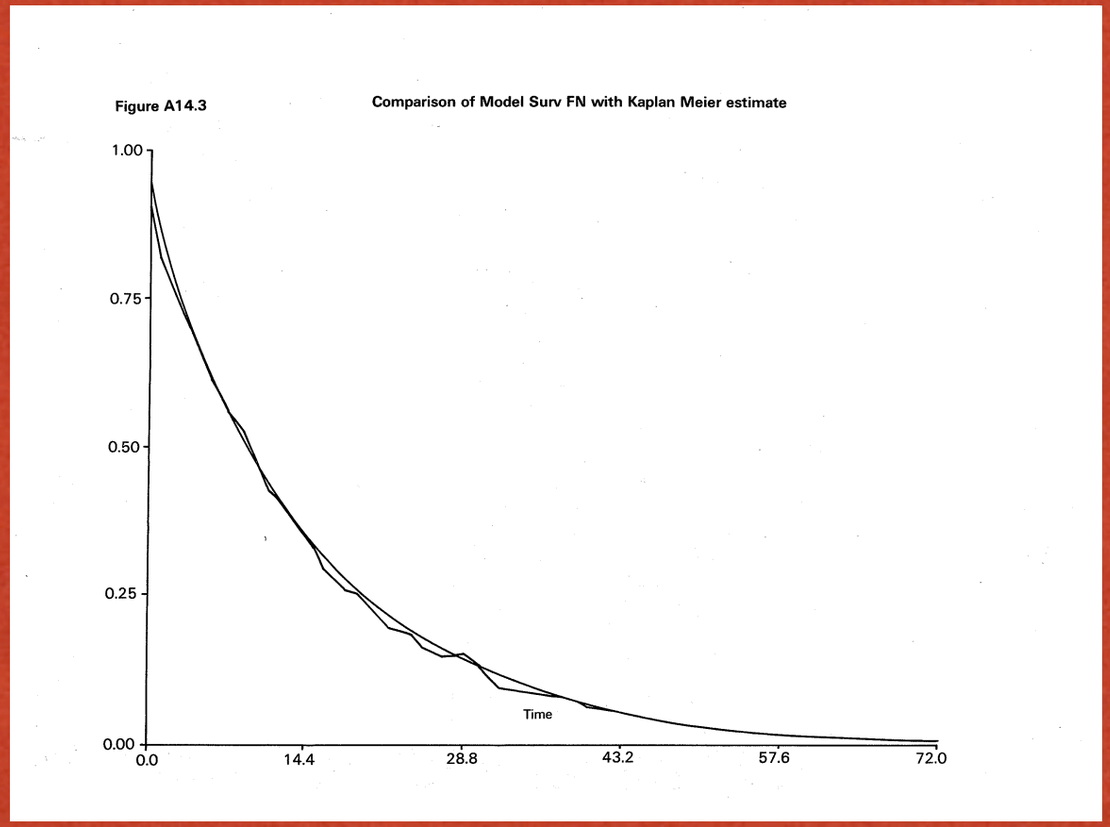
'The comparison of survival curves' compares the asymptotic relative efficiencies of four methods for comparing survival time distributions, when survival times are *exponential*:

- maximum likelihood
- the sign method*, in which individuals entering are paired and their times compared
- a comparison of proportions
- Kaplan-Meier.

THE COX REPORT

Short-term prediction of HIV infection and AIDS in England and Wales 1988

$$S(t) = \begin{cases} (1-\theta) & t = 0 \\ (1-\theta)S(t) & t > 0 \end{cases}$$



The overall distribution of survival time for UK AIDS patients, G.R. Reeves

The comparison of survival curves

The comparison of survival curves

- Dr D.R. Cox, Dr Boag and others were discussants.

The comparison of survival curves

- Dr D.R. Cox, Dr Boag and others were discussants.
- *Whence followed some of the most important work in modern statistics:*

The comparison of survival curves

- Dr D.R. Cox, Dr Boag and others were discussants.
- *Whence followed some of the most important work in modern statistics:*
- *Cox's proportional-hazards model.*

The comparison of survival curves

- Dr D.R. Cox, Dr Boag and others were discussants.
- *Whence followed some of the most important work in modern statistics:*
- *Cox's proportional-hazards model.*
- *If there was a Nobel prize for Statistics, this early work and the Cox model would surely have won it!*

The comparison of survival curves

- Dr D.R. Cox, Dr Boag and others were discussants.
- *Whence followed some of the most important work in modern statistics:*
- *Cox's proportional-hazards model.*
- *If there was a Nobel prize for Statistics, this early work and the Cox model would surely have won it!*
- *Why? Because applications in medicine have significantly improved the human condition and saved many hundreds of thousands of lives.*

Motivation

1. Introduction

The investigation reported in this paper arose out of a consideration of the possibility of using sequential methods in the design and analysis of clinical trials for treatments of chronic diseases in which the main criterion of success is length of survival after treatment.

In his discussion of Peter's paper, Professor A. Bradford Hill writes:

"I am glad that Dr Armitage has turned his attention so helpfully to a problem that can be very troublesome in clinical medicine."

Some Truths

1. Innovation in statistical thinking and methods is best driven by substantive applications.

Some Truths

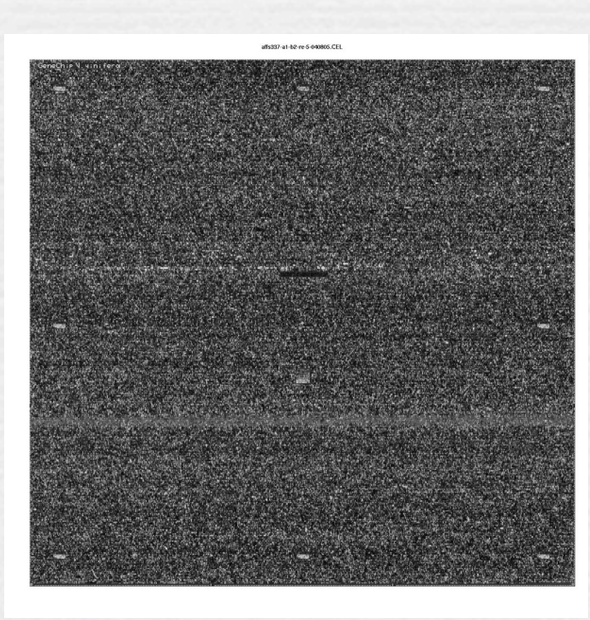
1. Innovation in statistical thinking and methods is best driven by substantive applications.

- Statistical analysis of noisy space probe images in 1950's.
- Demand for financial services.

Some Truths

1. Innovation in statistical thinking and methods is best driven by substantive applications.

- Statistical analysis of noisy space probe images in 1950's.
- Demand for financial services.



A processed Affymetrix chip

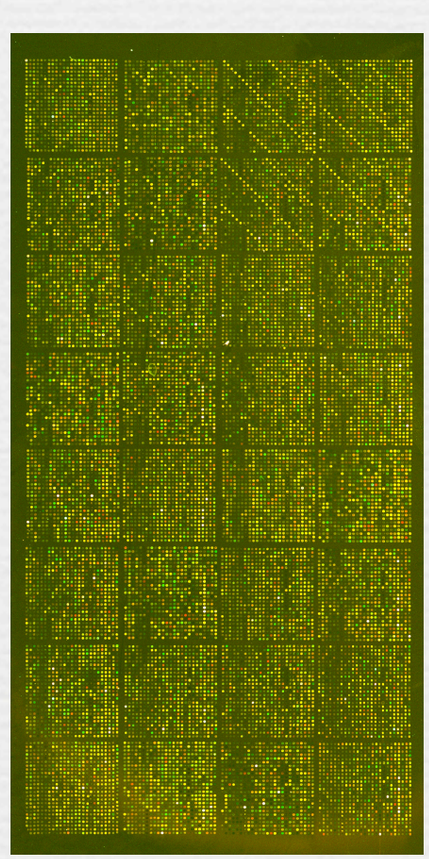


Image of a human long oligonucleotide microarray

Innovation in statistics can be driven by disasters

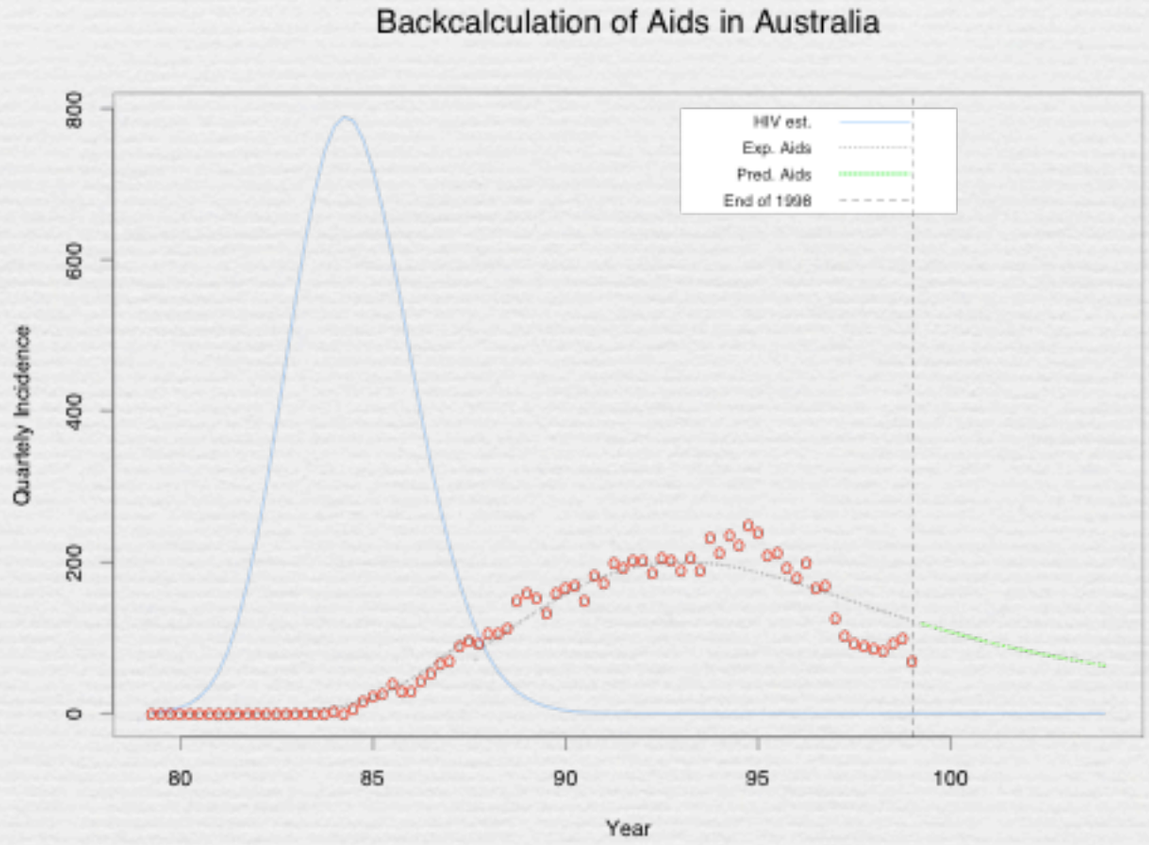
- HIV/AIDS pandemic
- challenges in early AZT trials
- method of *backcalculation* to reconstruct time-varying HIV infection incidence
- As *SIR* model:

$$a(t) = \int_0^t h(u) f(t-u)\, du$$

Innovation in statistics can be driven by disasters

- HIV/AIDS pandemic
- challenges in early AZT trials
- method of *backcalculation* to reconstruct time-varying HIV infection incidence
- As *SIR* model:

$$a(t) = \int_0^t h(u) f(t-u)\, du$$



Innovation in statistics can be driven by disasters

- HIV/AIDS pandemic
- challenges in early AZT trials
- method of *backcalculation* to reconstruct time-varying HIV infection incidence
- As *SIR* model:

$$a(t) = \int_0^t h(u) f(t-u)\, du$$

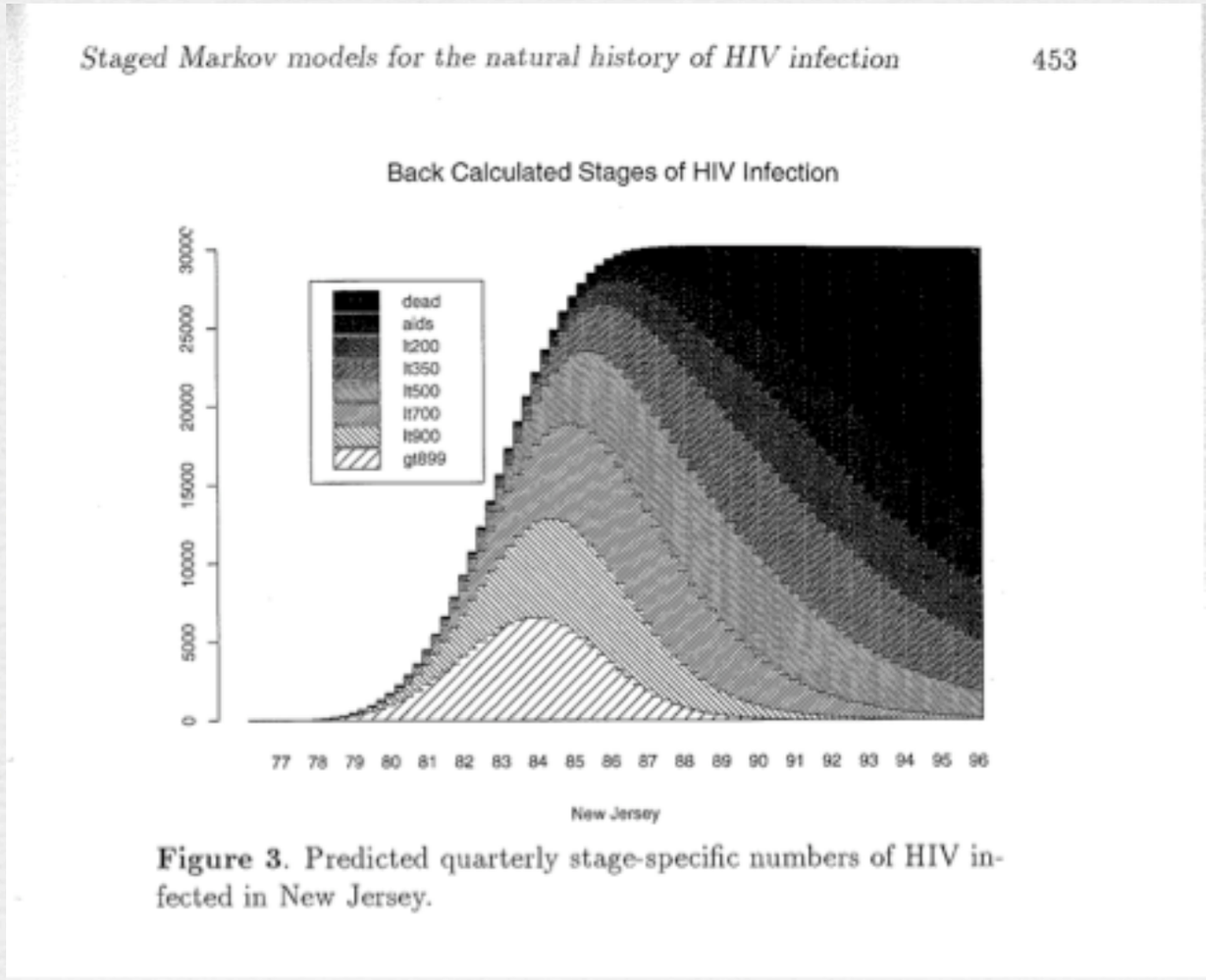
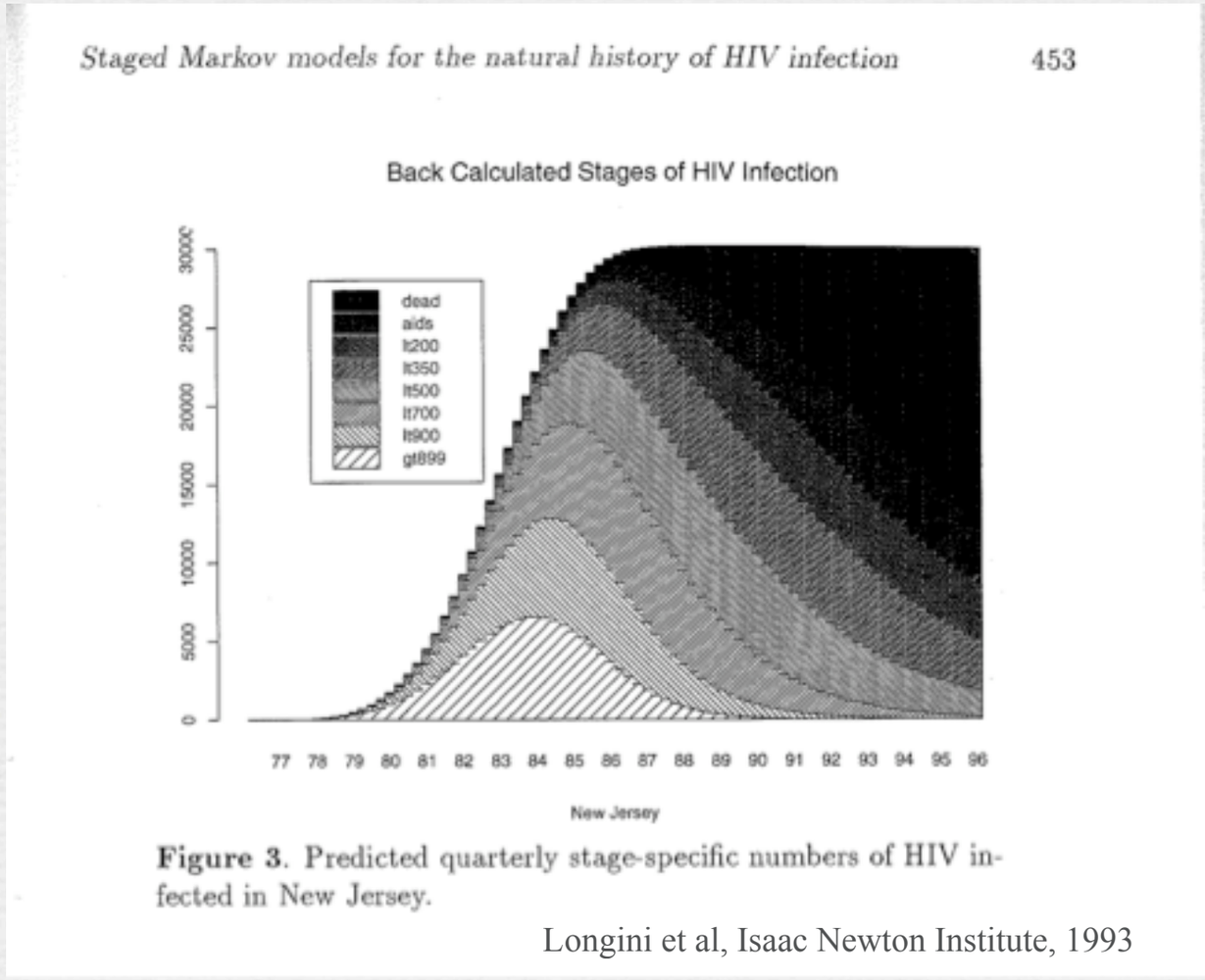


Figure 3. Predicted quarterly stage-specific numbers of HIV infected in New Jersey.

Innovation in statistics can be driven by disasters

- HIV/AIDS pandemic
- challenges in early AZT trials
- method of *backcalculation* to reconstruct time-varying HIV infection incidence
- As *SIR* model:

$$a(t) = \int_0^t h(u) f(t-u)\, du$$



Longini et al, Isaac Newton Institute, 1993

The sequencing of the human genome, together with increasingly accurate high-throughput technologies, has led to the mathematisation of biology.

The sequencing of the human genome, together with increasingly accurate high-throughput technologies, has led to the mathematisation of biology.

- In the beginning of **microarray data analysis**, we clustered or looked for differentially expressed genes using a statistic, e.*g.* *t*, and produced lists of ranked genes based on suitably chosen cut-offs.

The sequencing of the human genome, together with increasingly accurate high-throughput technologies, has led to the mathematisation of biology.

- In the beginning of **microarray data analysis**, we clustered or looked for differentially expressed genes using a statistic, e.*g.* *t*, and produced lists of ranked genes based on suitably chosen cut-offs.

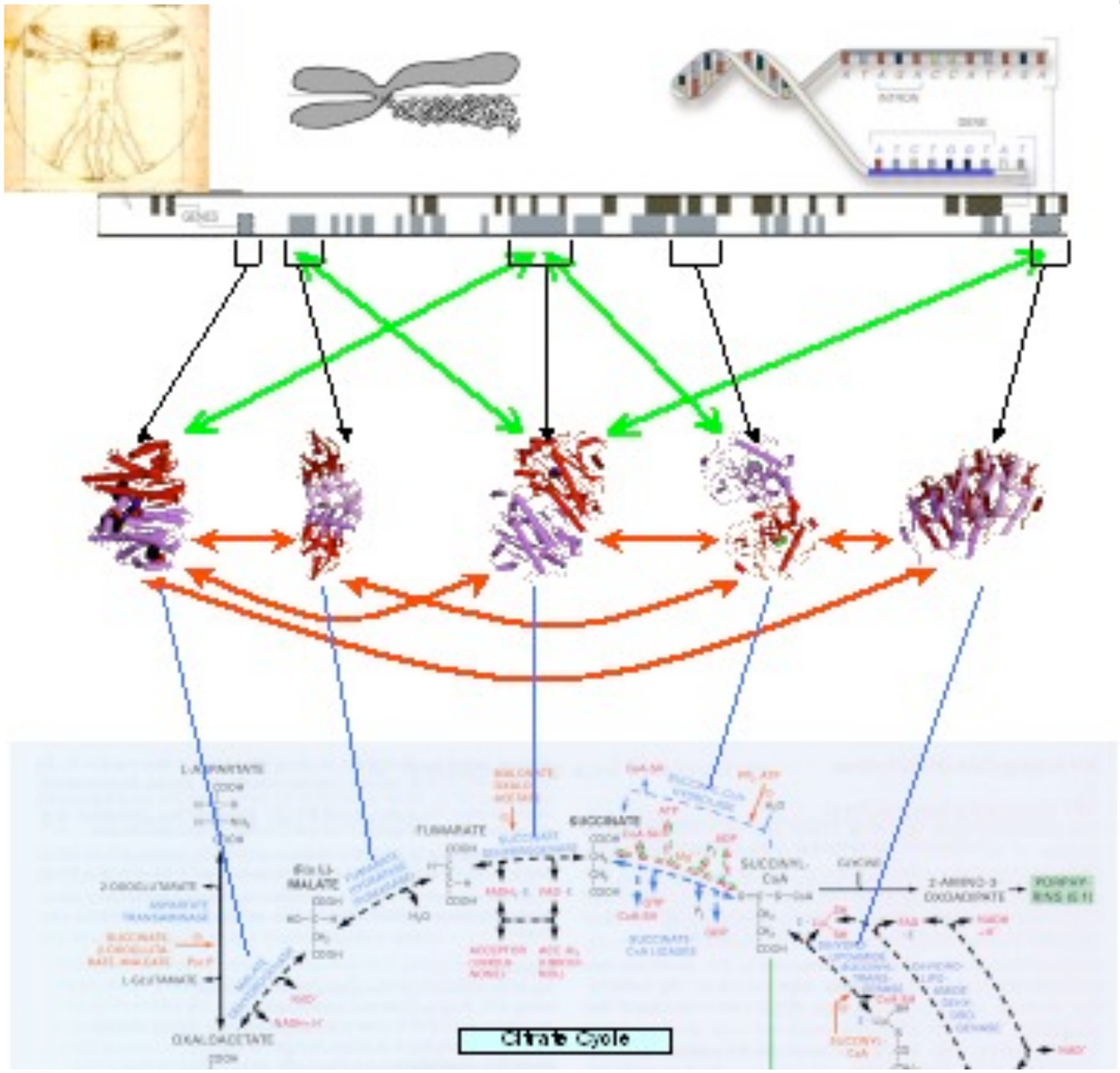
- Later, we examined these clusters of lists of d.e. genes for enrichment with various pre-defined sets of genes such as the Gene Ontology categories. Later still, we did other things ...

The sequencing of the human genome, together with increasingly accurate high-throughput technologies, has led to the mathematisation of biology.

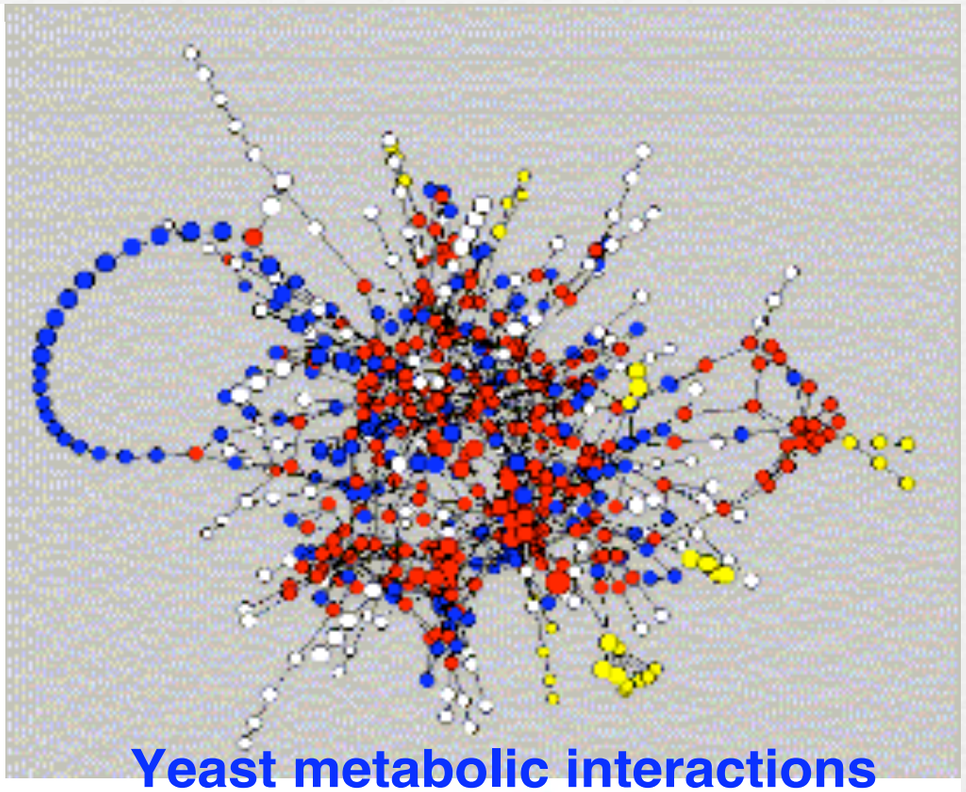
- In the beginning of **microarray data analysis**, we clustered or looked for differentially expressed genes using a statistic, e.*g.* *t*, and produced lists of ranked genes based on suitably chosen cut-offs.

- Later, we examined these clusters of lists of d.e. genes for enrichment with various pre-defined sets of genes such as the Gene Ontology categories. Later still, we did other things ...

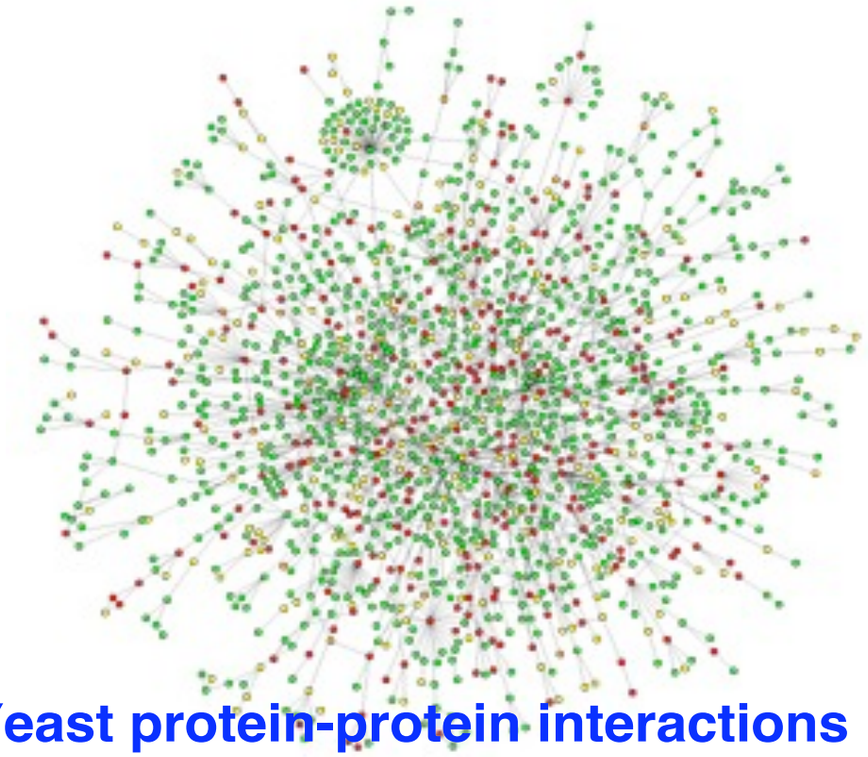
- Gene Networks took us one step further in the evolving sequence of methods for the analysis of gene expression microarray data.



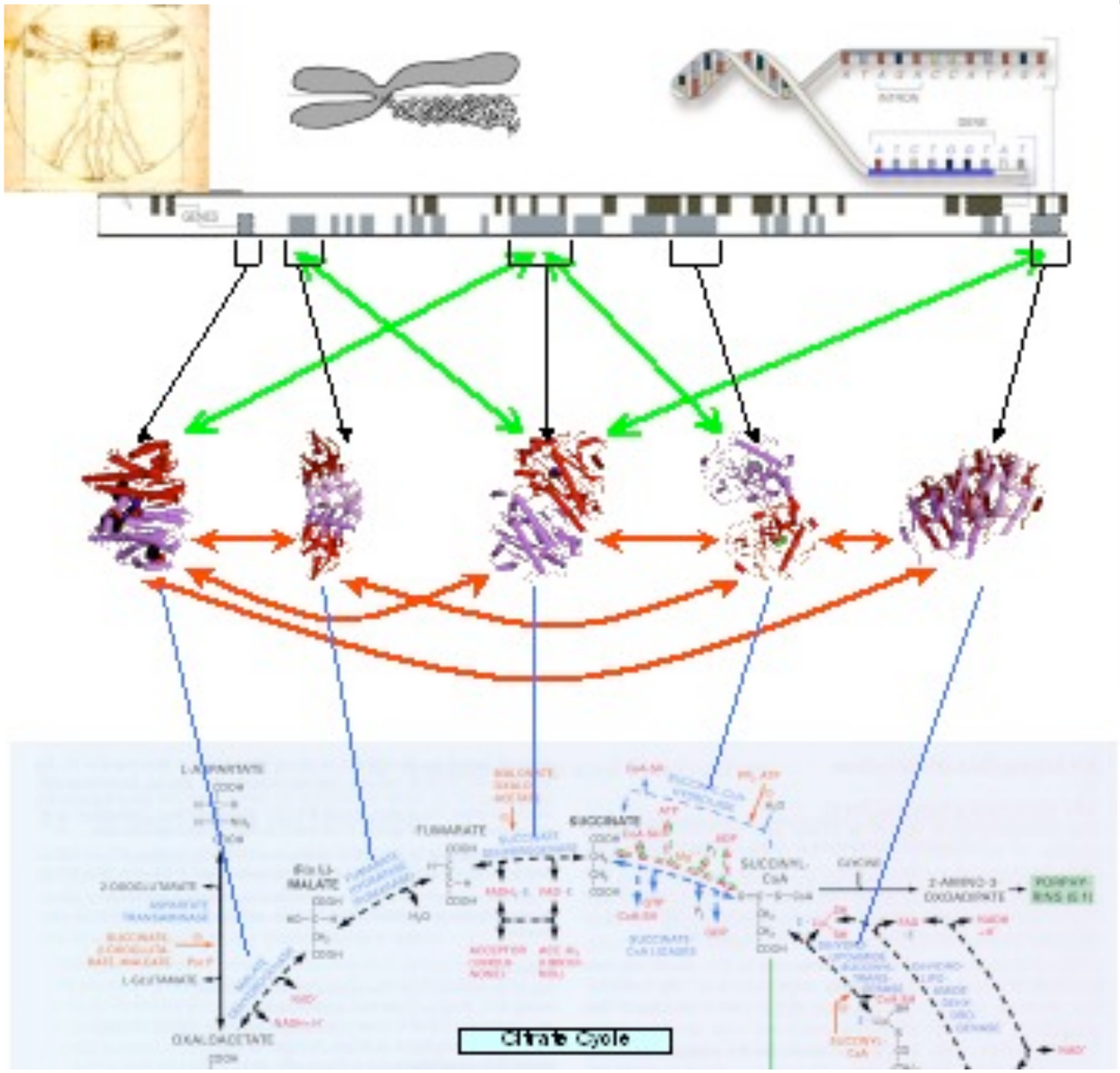
Some molecular interactions in the genome



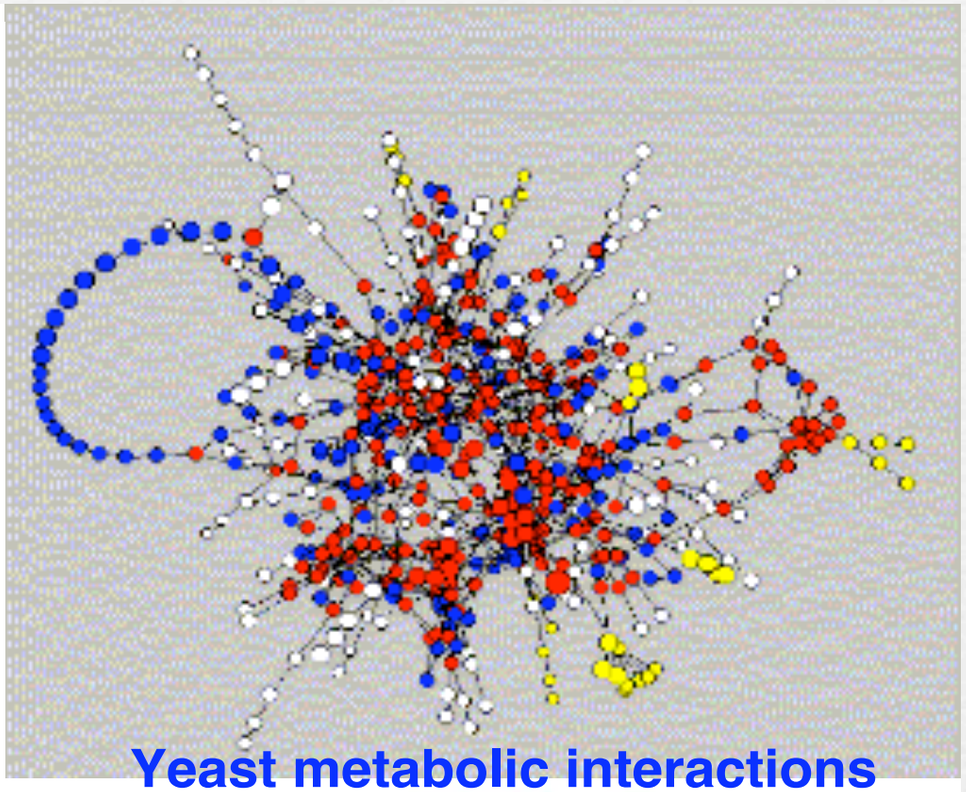
Yeast metabolic interactions



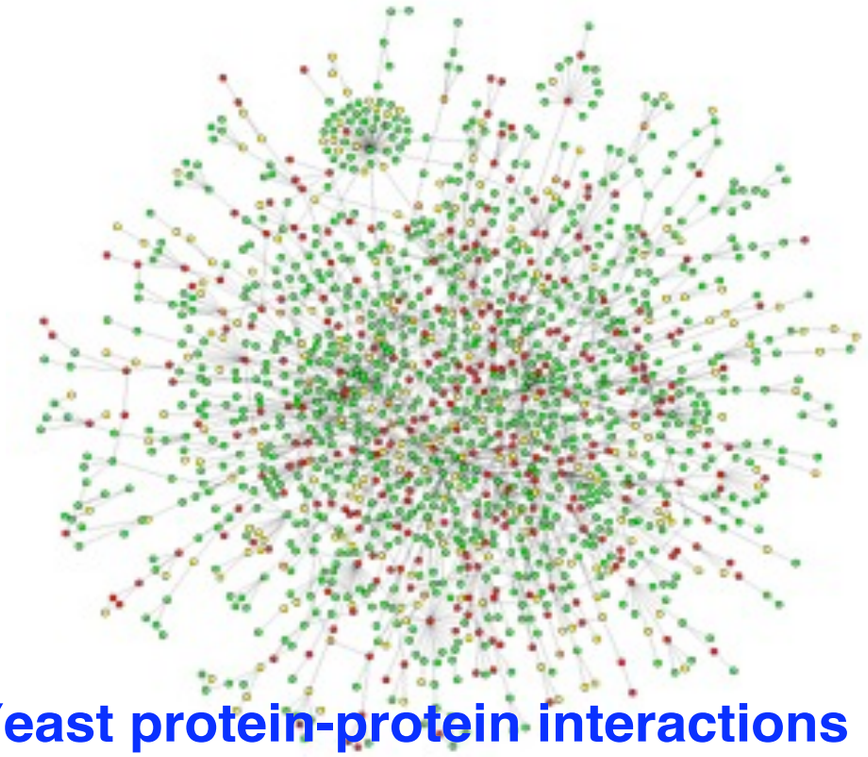
Yeast protein-protein interactions



Some molecular interactions in the genome



Yeast metabolic interactions



Yeast protein-protein interactions

- We are now in the post-genome era of *bioinformatics* and *systems biology*.

Some Truths

II. Biology is dominating statistics at the beginning of this century, just as it did at the beginning of the last one.

Some Truths

II. Biology is dominating statistics at the beginning of this century, just as it did at the beginning of the last one.

Why?

Some Truths

II. Biology is dominating statistics at the beginning of this century, just as it did at the beginning of the last one.

Why?

III. Statistics is an *enabling discipline*.

Some Truths

II. Biology is dominating statistics at the beginning of this century, just as it did at the beginning of the last one.

Why?

III. Statistics is an *enabling discipline*.

It has its own internal dynamics and coherence, but good statistical analysis is the key to getting the best out of the new technologies.

Some Truths

II. Biology is dominating statistics at the beginning of this century, just as it did at the beginning of the last one.

Why?

III. Statistics is an *enabling discipline*.

It has its own internal dynamics and coherence, but good statistical analysis is the key to getting the best out of the new technologies.
We have by training the skills of experimental design, data analysis, *synthesis* and reasoning which are essential to bioinformatics and systems biology.

Some Truths

IV. The statistical sciences must themselves be strong to enable high-level collaborations with scientists from cognate disciplines.

V. *"When we entered the era of high technology, we entered the era of mathematical technology"*

Ad Hoc Committee on Resources for the Mathematical Sciences,
US National Research Council, 1981

STATISTICAL BIOINFORMATICS MEETS EPIDEMIOLOGY

- ☐ DNA MAKES RNA MAKES PROTEIN.

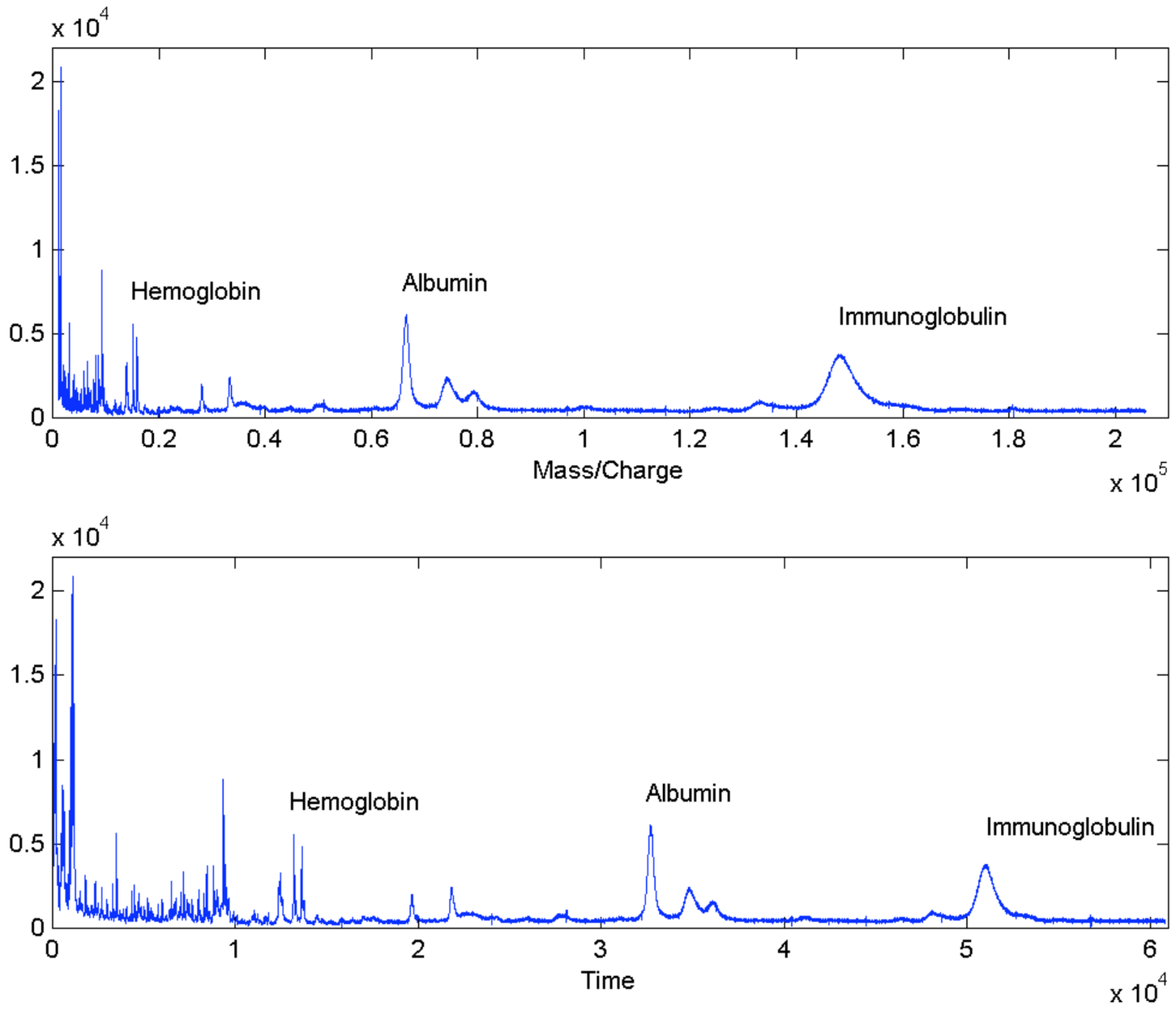
- ☐ MASS SPECTROMETRY ALLOWS US TO MEASURE THE PROTEIN COMPLEMENT (OR SUBSET THEREOF) OF A SET OF CELLS.

- ☐ THERE IS A GREAT DEAL OF INTEREST IN DISCOVERING PROTEIN BIOMARKERS IN BLOOD TO IDENTIFY CANCER PATIENTS EARLY ON.

WHAT'S THE EXCITEMENT ABOUT?

- ☐ PROFILES ARE BEING ASSESSED USING SERUM AND URINE, NOT TISSUE BIOPSIES.
- ☐ PROTEOMIC SPECTRA ARE CHEAPER TO RUN ON A PER UNIT BASIS THAN MICROARRAYS.
- ☐ CAN RUN SAMPLES ON LARGE NUMBERS OF PATIENTS.

A MASS SPECTRUM OF HUMAN SERUM



Mechanisms of disease

Lancet, 359, 2002:572-7

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

HTTP://HOME.CCR.CANCER.GOV.NCIFDAPROTEOMICS

Mechanisms of disease

Lancet, 359, 2002:572-7

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

- 100 OVARIAN CANCER PATIENTS; 100 NORMAL CONTROLS; 16 PATIENTS WITH 'BENIGN DISEASE'

HTTP://HOME.CCR.CANCER.GOV.NCIFDAPROTEOMICS

Mechanisms of disease

Lancet, 359, 2002:572-7

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

- ☐ 100 OVARIAN CANCER PATIENTS; 100 NORMAL CONTROLS; 16 PATIENTS WITH 'BENIGN DISEASE'

- ☐ USED 50 CANCER AND 50 NORMAL SPECTRA TO TRAIN A CLASSIFIER AND TESTED THE ALGORITHM ON THE REMAINING SAMPLES.

HTTP://HOME.CCR.CANCER.GOV.NCIFDAPROTEOMICS

Mechanisms of disease

Lancet, 359, 2002:572-7

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

- ☐ 100 OVARIAN CANCER PATIENTS; 100 NORMAL CONTROLS; 16 PATIENTS WITH 'BENIGN DISEASE'

- ☐ USED 50 CANCER AND 50 NORMAL SPECTRA TO TRAIN A CLASSIFIER AND TESTED THE ALGORITHM ON THE REMAINING SAMPLES.

- ☐ CORRECTLY CLASSIFIED 50/50 OF THE OVARIAN CANCER CASES.

HTTP://HOME.CCR.CANCER.GOV.NCIFDAPROTEOMICS

Mechanisms of disease

Lancet, 359, 2002:572-7

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

- 100 OVARIAN CANCER PATIENTS; 100 NORMAL CONTROLS; 16 PATIENTS WITH 'BENIGN DISEASE'
- USED 50 CANCER AND 50 NORMAL SPECTRA TO TRAIN A CLASSIFIER AND TESTED THE ALGORITHM ON THE REMAINING SAMPLES.
- CORRECTLY CLASSIFIED 50/50 OF THE OVARIAN CANCER CASES.
- CORRECTLY CLASSIFIED 46/50 OF THE NORMAL CASES.

HTTP://HOME.CCR.CANCER.GOV.NCIFDAPROTEOMICS

Mechanisms of disease

Lancet, 359, 2002:572-7

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

- 100 OVARIAN CANCER PATIENTS; 100 NORMAL CONTROLS; 16 PATIENTS WITH 'BENIGN DISEASE'
- USED 50 CANCER AND 50 NORMAL SPECTRA TO TRAIN A CLASSIFIER AND TESTED THE ALGORITHM ON THE REMAINING SAMPLES.
- CORRECTLY CLASSIFIED 50/50 OF THE OVARIAN CANCER CASES.
- CORRECTLY CLASSIFIED 46/50 OF THE NORMAL CASES.
- CORRECTLY CLASSIFIED 16/16 OF THE BENIGN DISEASE AS 'OTHER'.

HTTP://HOME.CCR.CANCER.GOV.NCIFDAPROTEOMICS

MUCH EXCITEMENT ...

*SORACE & ZHAN BMC BIOINFORMATICS 2003; ZHANG ET AL CANCER RES 2004

MUCH EXCITEMENT ...

- ☐ GROUPS AROUND THE WORLD START ASKING HOW TO DO THIS WITH THEIR TYPE OF CANCER ...

*SORACE & ZHAN BMC BIOINFORMATICS 2003; ZHANG ET AL CANCER RES 2004

MUCH EXCITEMENT ...

- ☐ GROUPS AROUND THE WORLD START ASKING HOW TO DO THIS WITH THEIR TYPE OF CANCER ...

- ☐ BUT IN 2002-3, E. DIAMANDIS (AND OTHERS) RAISED OBJECTIONS ABOUT THE APPROACH: IT SHOULDN'T WORK, OWING TO LIMITATIONS OF THE TECHNOLOGY.

*SORACE & ZHAN BMC BIOINFORMATICS 2003; ZHANG ET AL CANCER RES 2004

MUCH EXCITEMENT ...

- ☐ GROUPS AROUND THE WORLD START ASKING HOW TO DO THIS WITH THEIR TYPE OF CANCER ...

- ☐ BUT IN 2002-3, E. DIAMANDIS (AND OTHERS) RAISED OBJECTIONS ABOUT THE APPROACH: IT SHOULDN'T WORK, OWING TO LIMITATIONS OF THE TECHNOLOGY.

- ☐ VARIOUS QUESTIONS ABOUT ODDITIES IN THE DATA BEGIN TO CROP UP (KEITH BAGGERLY AND OTHERS.)

*SORACE & ZHAN BMC BIOINFORMATICS 2003; ZHANG ET AL CANCER RES 2004

MUCH EXCITEMENT ...

- ☐ GROUPS AROUND THE WORLD START ASKING HOW TO DO THIS WITH THEIR TYPE OF CANCER ...

- ☐ BUT IN 2002-3, E. DIAMANDIS (AND OTHERS) RAISED OBJECTIONS ABOUT THE APPROACH: IT SHOULDN'T WORK, OWING TO LIMITATIONS OF THE TECHNOLOGY.

- ☐ VARIOUS QUESTIONS ABOUT ODDITIES IN THE DATA BEGIN TO CROP UP (KEITH BAGGERLY AND OTHERS.)

- ☐ THE RESULTS ARE NOT REPRODUCIBLE FROM THE 'SAME' DATA.

*SORACE & ZHAN BMC BIOINFORMATICS 2003; ZHANG ET AL CANCER RES 2004

MUCH EXCITEMENT ...

- ☐ GROUPS AROUND THE WORLD START ASKING HOW TO DO THIS WITH THEIR TYPE OF CANCER ...

- ☐ BUT IN 2002-3, E. DIAMANDIS (AND OTHERS) RAISED OBJECTIONS ABOUT THE APPROACH: IT SHOULDN'T WORK, OWING TO LIMITATIONS OF THE TECHNOLOGY.

- ☐ VARIOUS QUESTIONS ABOUT ODDITIES IN THE DATA BEGIN TO CROP UP (KEITH BAGGERLY AND OTHERS.)

- ☐ THE RESULTS ARE NOT REPRODUCIBLE FROM THE 'SAME' DATA.

- ☐ PERFECT CLASSIFICATION OF PEAKS ACHIEVED WITH NOISE ...*

*SORACE & ZHAN BMC BIOINFORMATICS 2003; ZHANG ET AL CANCER RES 2004

MUCH EXCITEMENT ...

- ☐ GROUPS AROUND THE WORLD START ASKING HOW TO DO THIS WITH THEIR TYPE OF CANCER ...

- ☐ BUT IN 2002-3, E. DIAMANDIS (AND OTHERS) RAISED OBJECTIONS ABOUT THE APPROACH: IT SHOULDN'T WORK, OWING TO LIMITATIONS OF THE TECHNOLOGY.

- ☐ VARIOUS QUESTIONS ABOUT ODDITIES IN THE DATA BEGIN TO CROP UP (KEITH BAGGERLY AND OTHERS.)

- ☐ THE RESULTS ARE NOT REPRODUCIBLE FROM THE 'SAME' DATA.

- ☐ PERFECT CLASSIFICATION OF PEAKS ACHIEVED WITH NOISE ...*

*SORACE & ZHAN BMC BIOINFORMATICS 2003; ZHANG ET AL CANCER RES 2004

- [ ] ALL THIS (AND MORE) STRONGLY SUGGESTED A QUALITATIVE DIFFERENCE IN HOW THE SAMPLES WERE PROCESSED, AND POSSIBLY NOT JUST A DIFFERENCE IN THE BIOLOGY.

- [ ] IN JANUARY 2004, CORRELOGIC, QUESTDIAGNOSTICS AND LABCORP ANNOUNCED PLANS TO OFFER A 'HOME BREW' TEST CALLED OVACHECK: SAMPLES WOULD BE SENT BY CLINICIANS FOR DIAGNOSIS.

- [ ] ESTIMATED MARKET: 8 TO 10 MILLION WOMEN

- [ ] ESTIMATED COST: US$100-200 PER TEST.

IN AN ABORTIVE SECOND PAPER, CONDRADS ET AL*

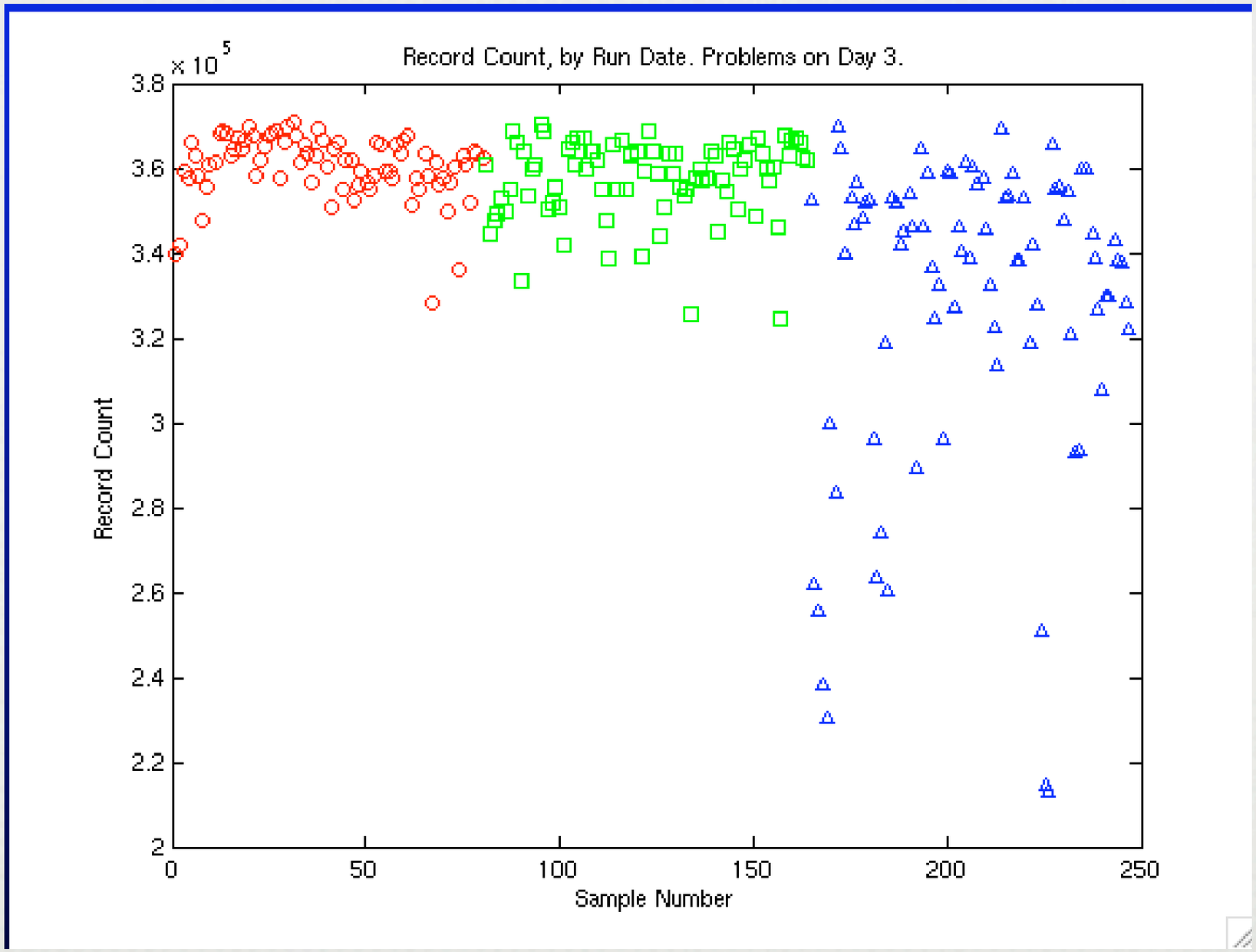
- PROCESSED SAMPLES WITH THEIR ORIGINAL SELDI TECHNOLOGY AND ALSO WITH A HIGHER RESOLUTION INSTRUMENT (QSTAR). THEY ADDED SOME QA/QC STEPS TO REMOVE BAD SPECTRA.

- DEMONSTRATED 100% SENSITIVITY AND 100% SPECIFICITY FOR IDENTIFYING CANCER FROM NORMAL, AND STATED THAT THIS "EMERGING PARADIGM" IS READY TO GO TO A LARGE CLINICAL STUDY.

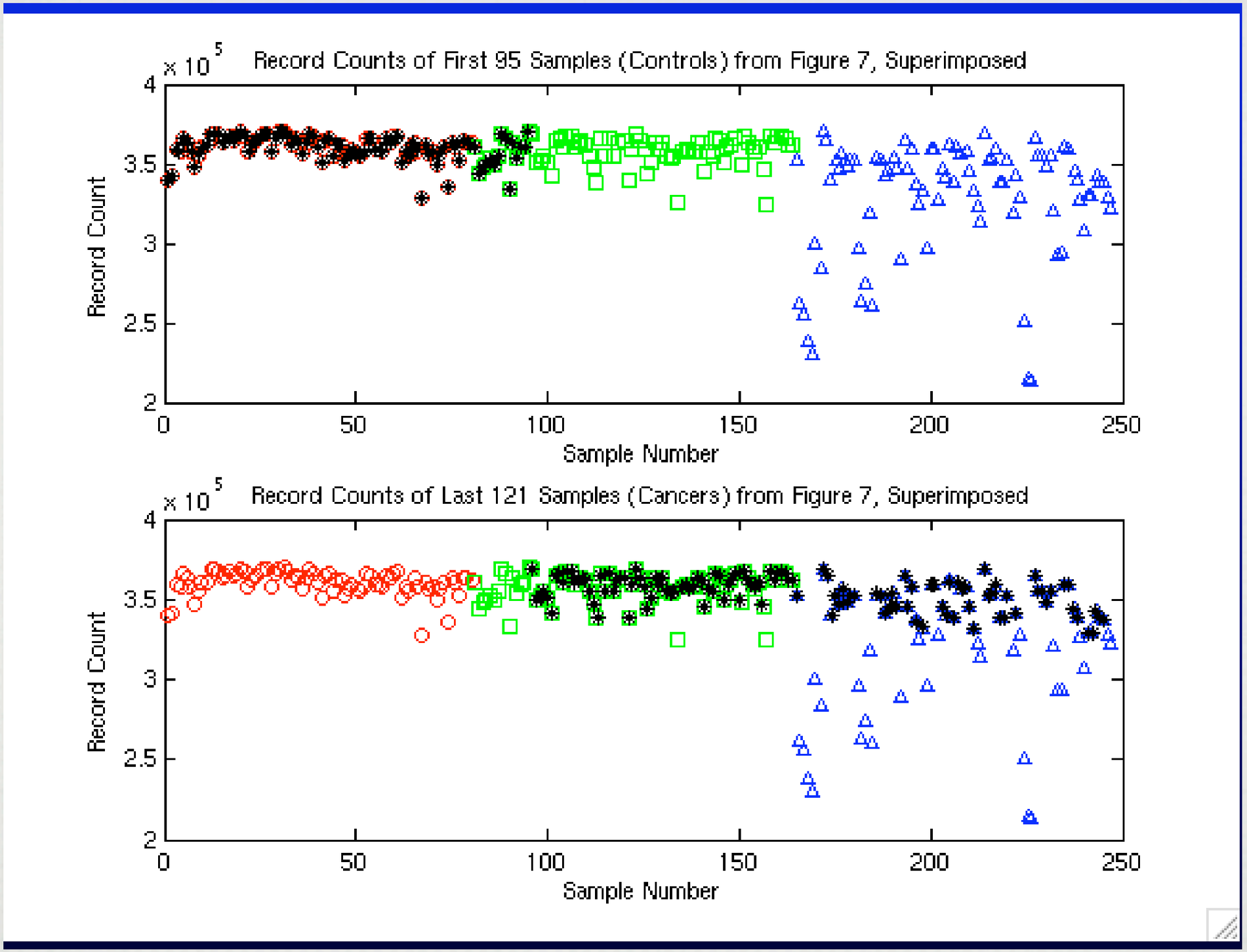
SO WHAT WAS GOING ON?

- * ENDOCRINE RELATED CANCER 11, 163-178, 2004

HERE ARE THEIR FIGURES 6A AND 7



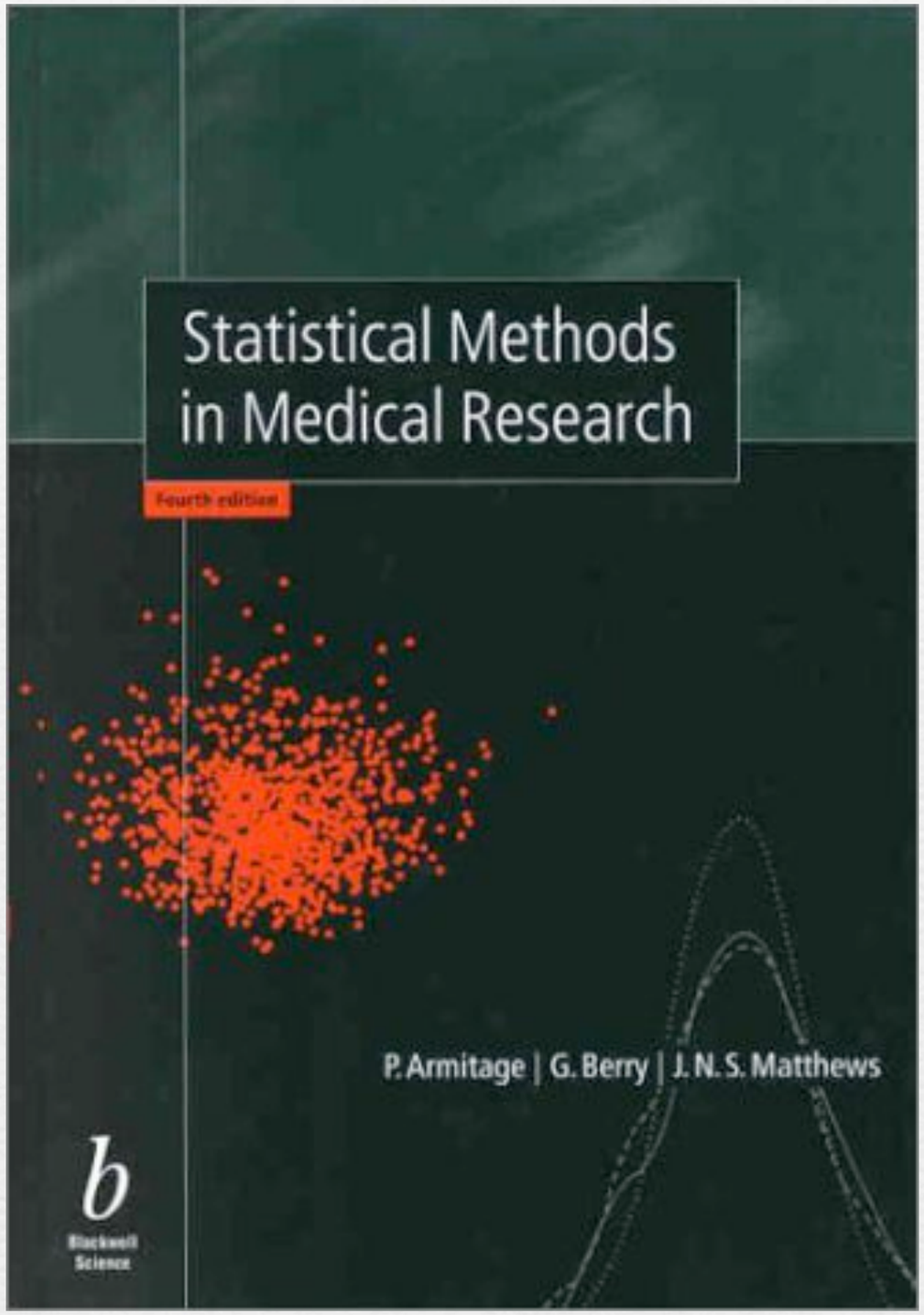
COLOUR = DAY 1, 2, 3



ALL OF THE CONTROLS WERE RUN BEFORE *ALL* OF THE CANCERS.

THE MORAL OF THE STORY

- ☐ THERE IS NO WAY A WOMAN SHOULD BE TOLD SHE NEEDS AN OOPHORECTOMY BASED ON THESE TESTS!
- ☐ IN JUNE 2004, THE FDA RULED THAT OVACHECK COULD NOT BE MADE AVAILABLE UNDER THE "HOME BREW" EXEMPTION, AS THE SOFTWARE PROGRAM WAS A 'DEVICE' THAT NEEDED TO BE MORE TIGHTLY REGULATED.
- ☐ IN SEPTEMBER 2006, THE FDA RELEASED DRAFT GUIDANCE ON 'IN VITRO DIAGNOSTIC MULTIVARIATE INDEX ASSAYS' (IVDMIAS).
- ☐ THESE RULES ARE BEING DEBATED EVEN NOW.


Statistical Methods
in Medical Research
Fourth edition
P. Armitage | G. Berry | J. N. S. Matthews
b
Blackwell
Science

A story about how statistics is helping to discover genes involved in pluripotency

Gene-profiling for a time course microarray experiment in stem cells

Pluripotency: Refers to the potential of an early stem cell to develop into any type of mature cell, depending on environment.

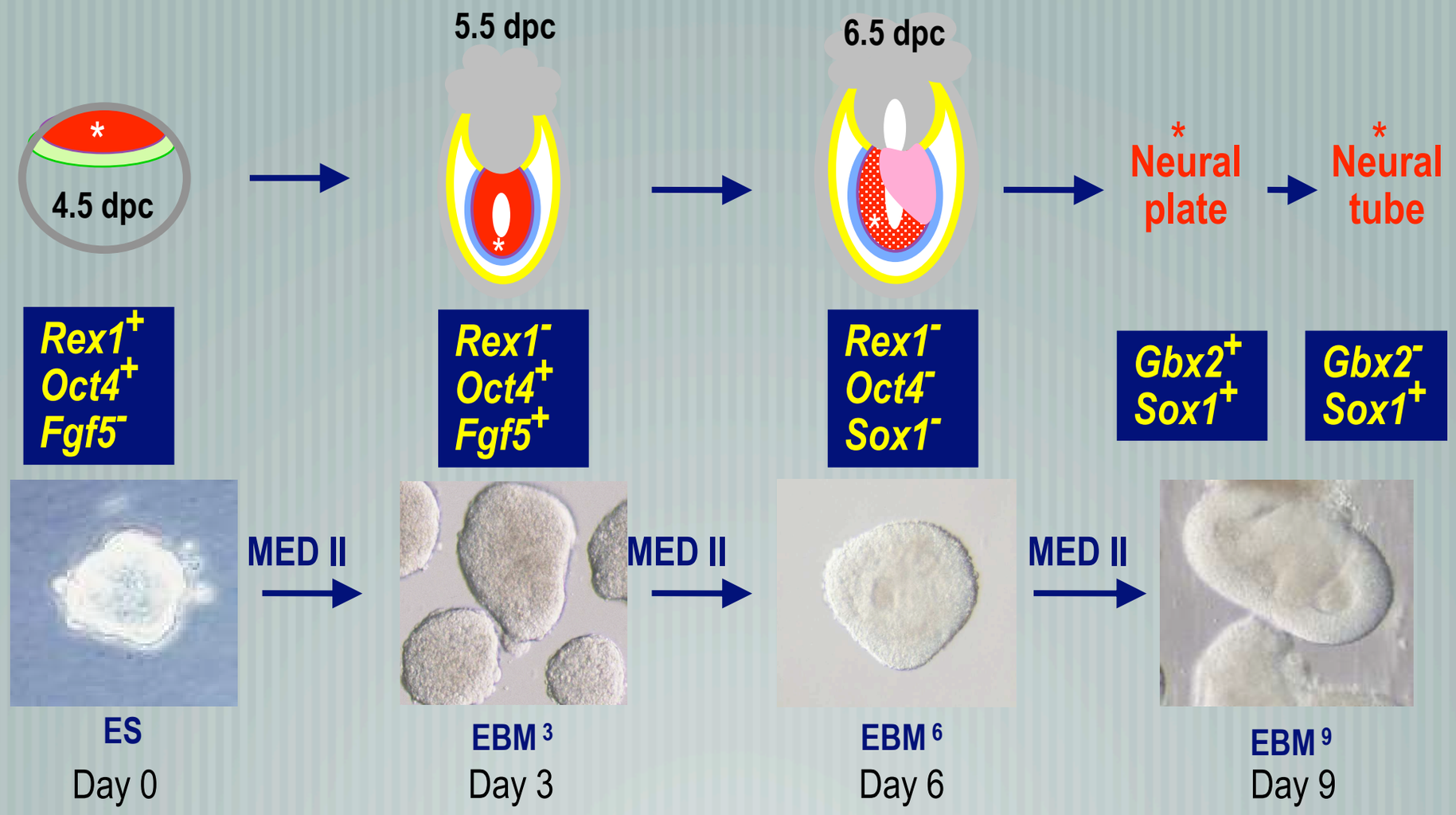
Important in research ranging from organ transplants, the treatment of diabetes, to the treatment of spinal injuries.

Motivation for our research

- Collaboration with Rathjen Lab, formerly of University of Adelaide.*
- They have developed a *mouse embryonic stem (ES) cell line* for studying lineage specific differentiation.
- *Aim:* to use microarrays to study changes in gene expression as a population of pluripotent ES cells are directed down the neuronal lineage via replacement of ES supporting media (LIF) with MEDII media.
- "**Omnibus**" experiment.

* Australian Stem Cell Centre,
ARC Special Research Centre for the Molecular Genetics of Development

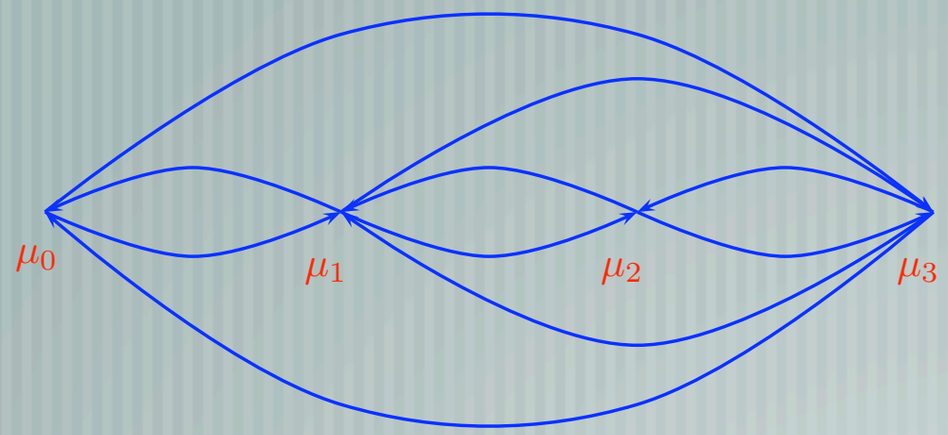
Cell line model for studying directed differentiation of ES cells down neuronal lineage



Aims, within resource constraints

- study loss of pluripotence over time
- identification of genes specific to each state
- rank genes according to association with pluripotency.
- 20 hybridisations.

Design of the ES cell experiment



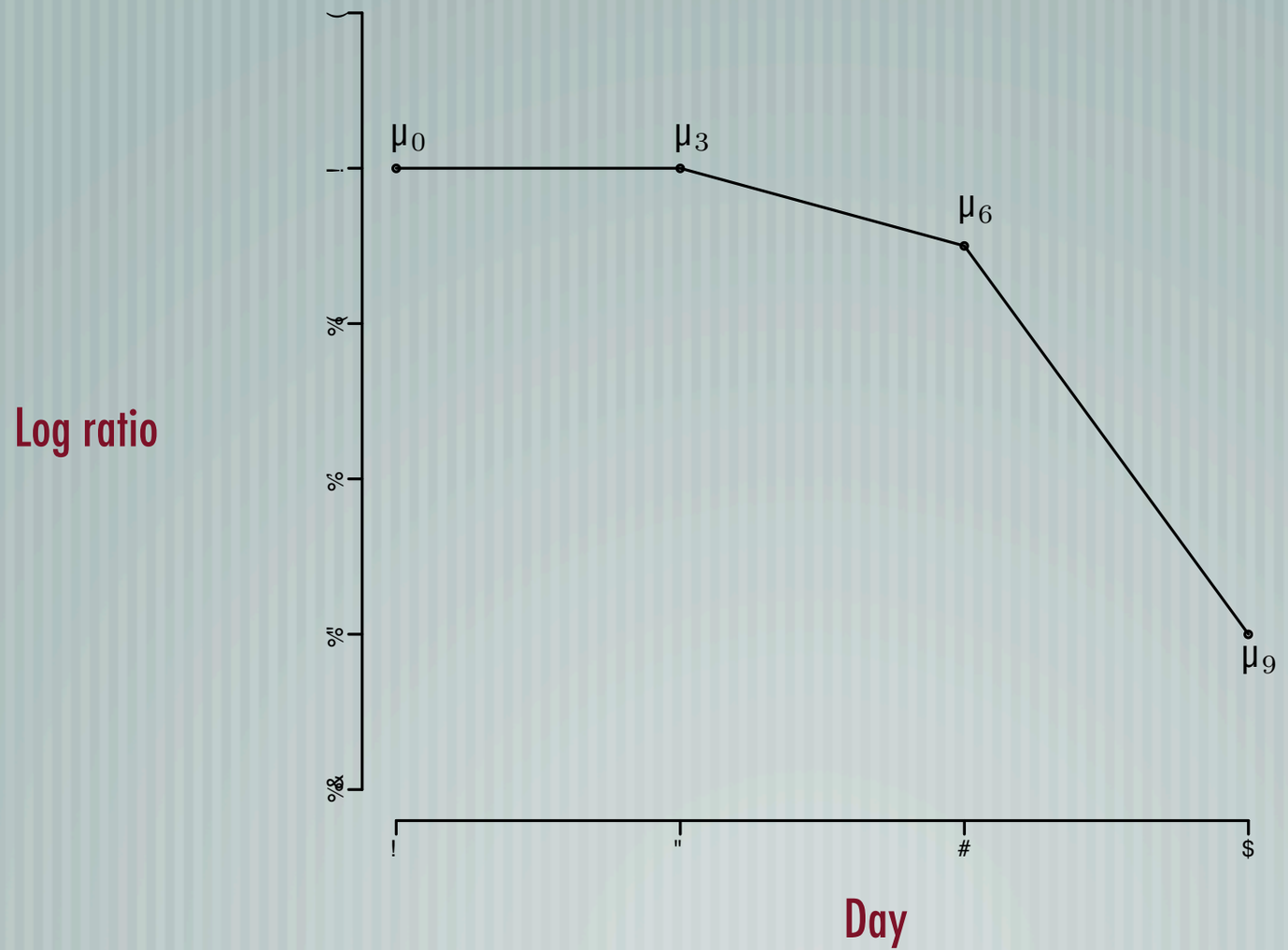
Day 0 3 6 9

CompuGen Mouse 22K long Oligo library

16 mRNA samples harvested on days 0,3,6,9

Passages p21 - p24 treated as biological replicates

The hypothetical profile for pluripotency



Log ratio with respect to Day 0 is plotted.

- We want to identify genes with the pre-specified time-course profile.

- Available methods typically not stringent enough - get part of profile, and spurious other parts. For example, Pareto optimization*; inner product of observed log ratios and a pre-specified profile**.

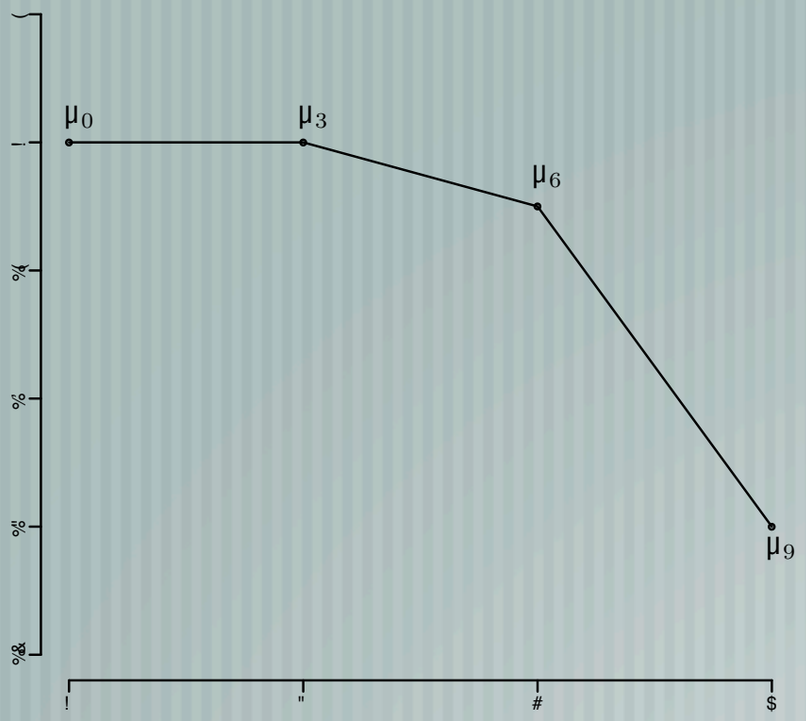
- Require inferential procedures which accommodate testing hypotheses of equivalence of gene expression and hypotheses of differential gene expression simultaneously.

*Fleury, Hero et al 2002, 2004; **Lonnstedt et al 2003.

Gene profiling

- We identify genes matching a pre-specified gene expression profile.
- We treat the vector of true gene expression levels (for each gene) as a linear combination of linearly independent vectors, chosen to represent the pre-specified time profile.
- The gene-profile model is fitted to the data, and the genes are ranked according to a suitable test statistic, incorporating equivalence and differential gene expression hypotheses.

Parameterisation for stem cell experiment



$$\mu = 0\begin{pmatrix}1\\1\\1\\1\end{pmatrix} + 1\begin{pmatrix}1\\1\\0\\0\end{pmatrix} + 2\begin{pmatrix}1\\1\\1\\0\end{pmatrix} + 3\begin{pmatrix}1/2\\-1/2\\0\\0\end{pmatrix}$$

Pluripotent profile

With this choice of model, it follows that

$$
\begin{aligned}
\gamma_0 &= \mu_9 \\
\gamma_1 &= \frac{\mu_0 + \mu_3}{2} - \mu_6 > 0 \\
\gamma_2 &= \mu_6 - \mu_9 > 0 \\
\gamma_3 &= \mu_0 - \mu_3 = 0
\end{aligned}
$$

We want to actively demonstrate equivalence of gene expression (on a gene-by-gene basis) on days 0 and 3, not simply fail to find a significant difference in the expression levels.

Equivalence testing

Null and alternative hypotheses:

$$H_0 \quad : \quad |y| \geq \epsilon, \quad \epsilon > 0$$

$$H_A \quad : \quad |y| < \epsilon$$

Confidence interval inclusion: create $(L_\alpha(\boldsymbol{X}), U_\alpha(\boldsymbol{X}))$

s.t. $P(y \in (L_\alpha(\boldsymbol{X}), \infty)) = P(y \in (-\infty, U_\alpha(\boldsymbol{X}))) = 1 - \alpha$

Reject null hypothesis in favour of equivalence iff

$$(L_\alpha(\boldsymbol{X}), U_\alpha(\boldsymbol{X})) \subset (-\epsilon, \epsilon)$$

Wellek *Testing Statistical Hypotheses of Equivalence* 2002

A snag ...

What do we mean by 'equivalence' of gene expression?

- This is analogous to the "minimum clinically significant difference" of interest in a superiority trial
- or "tolerance limit" in a non-inferiority trial.

But, we have relatively little information on what it means to say genes are 'equivalently expressed', let alone their gene-specific variation, or their interactions with other genes. And we have **lots** of genes ...

Linear model for pluripotency

- M is the vector of observed log ratios

$$\boldsymbol{M} = X^* \boldsymbol{\mu} + E$$

- we are measuring relative gene expression;
- we estimate γ via least squares;
- obtain an empirical Bayes estimate of σ^2.

Intersection-union test*

I-UT:

$$H_0 \quad : \quad (\gamma_1 \leq 0) \cup (\gamma_2 \leq 0) \cup (|\gamma_3| \geq \epsilon), \quad \epsilon > 0$$
$$H_A \quad : \quad (\gamma_1 > 0) \cap (\gamma_2 > 0) \cap (|\gamma_3| < \epsilon)$$

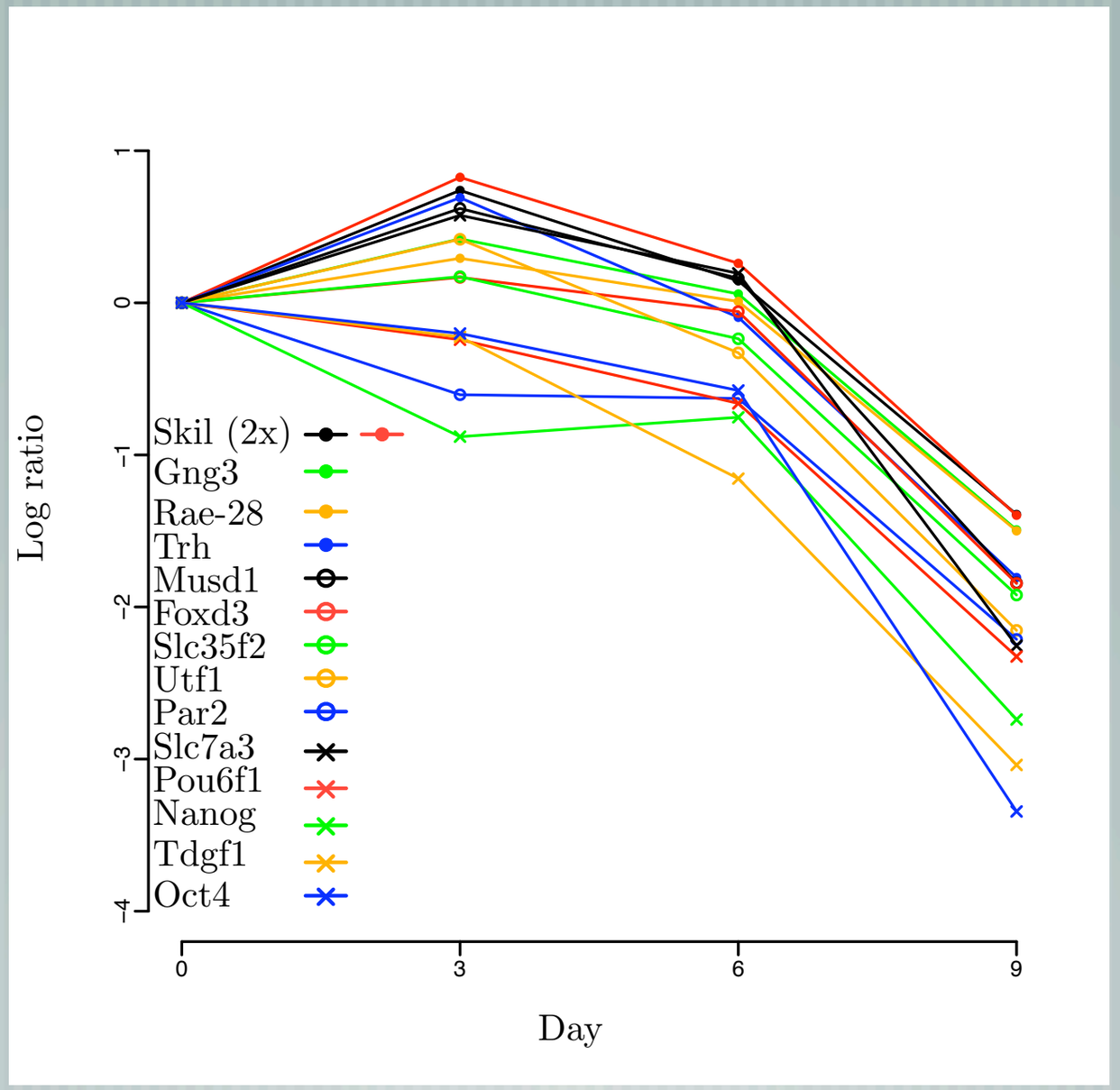
Rejection region (RR): intersection of separate rejection regions: $\alpha = \sup \alpha_\gamma$

Our aim is to rank the genes according to their match with the pre-specified profile: base on width of the largest c.i. contained within the RR.

Distance to the nearest boundary of the RR is calculated in standard errors of the estimate, larger values indicative of pluripotency.

*Berger, 1982

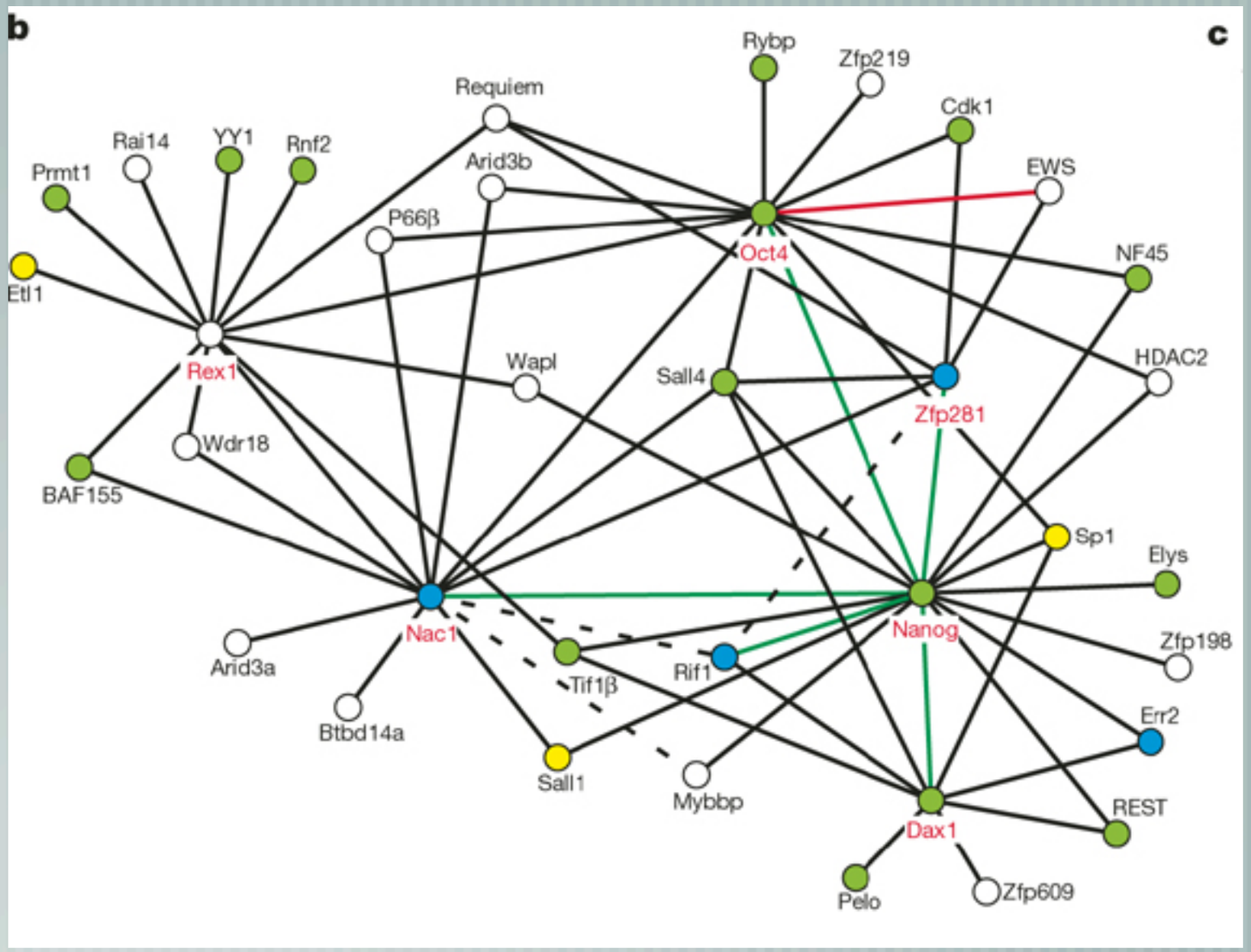
Pluripotency profile: fitted log ratios with respect to Day 0 for the ranked genes



Rodda et al, 2005; Loh et al, 2006; Nishimoto et al, 2005; Wang et al, 2006

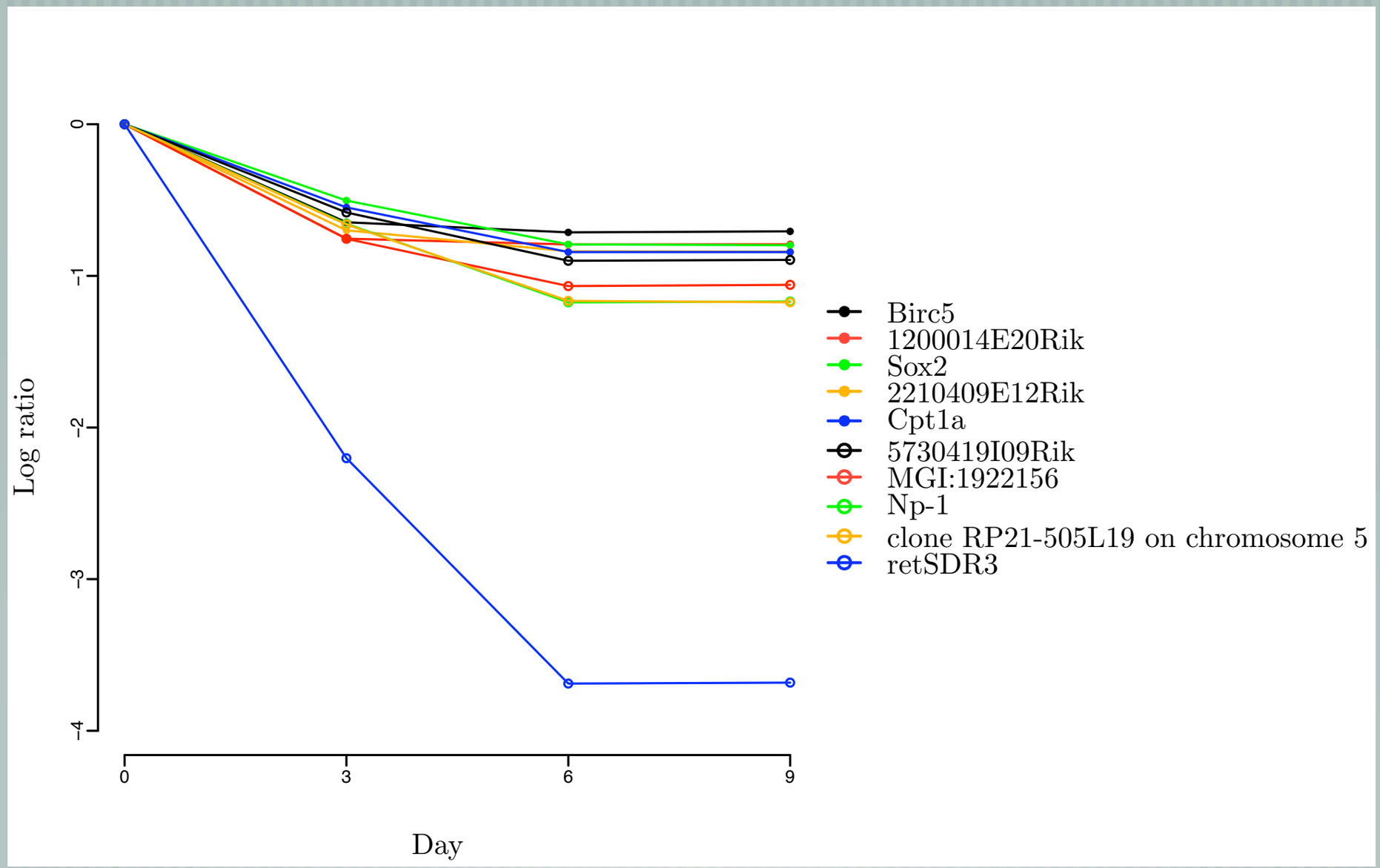
Ranked pluripotent-profile genes

Gene Names	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	U
Oct4	0.48	2.77	0.20	2.53
Utf1	0.54	1.82	−0.42	2.46
Tdgf1	1.04	1.88	0.22	1.80
Slc35f2	0.32	1.69	−0.17	1.53
Trh	0.44	1.71	−0.69	1.50
Foxd3	0.14	1.79	−0.17	1.33
Musd1	0.15	2.00	−0.62	1.17
Skil	0.15	1.66	−0.83	1.16
Pou6f1	0.54	1.66	0.24	1.13
Par2	0.33	1.58	0.60	0.75
Nanog	0.31	1.99	0.88	0.69
Slc7a3	0.09	2.45	−0.58	0.67
Gng3	0.15	1.55	−0.42	0.33
Skil	0.23	1.54	−0.74	0.28
Rae-28	0.14	1.51	−0.29	0.08



Protein interaction network, Wang et al, Nature 2006

‘Sox2’ profile ranked genes



There's more ...

We note in passing that equivalence is not transitive, i.e., not invariant to re-parameterisation (but then nor are Wald tests, for example).

Gene profiling can be used for a variety of data types, e.g., microRNAs, single-channel gene expression data, ...

Straightforward to fit models in Limma, R and C.

Is proving valuable as a tool for exploring and building networks, finding patterns based on differences in treatment and control on a genome-wide scale.

Projects: role of transcription factor MYB in microRNAs; responses of Th1 and Th2 cells to IL4 and IL12.

Acknowledgements

Peter Armitage

David Cox

Vern Farewell

The University of Adelaide

Jonathan Tuke

Gary Glonek

For extraordinary photo-searching capability:

John Bithell

M.D. Anderson Cancer Centre

Keith Baggerly

PROFESSOR PETER ARMITAGE

Photo courtesy of Ted Colton