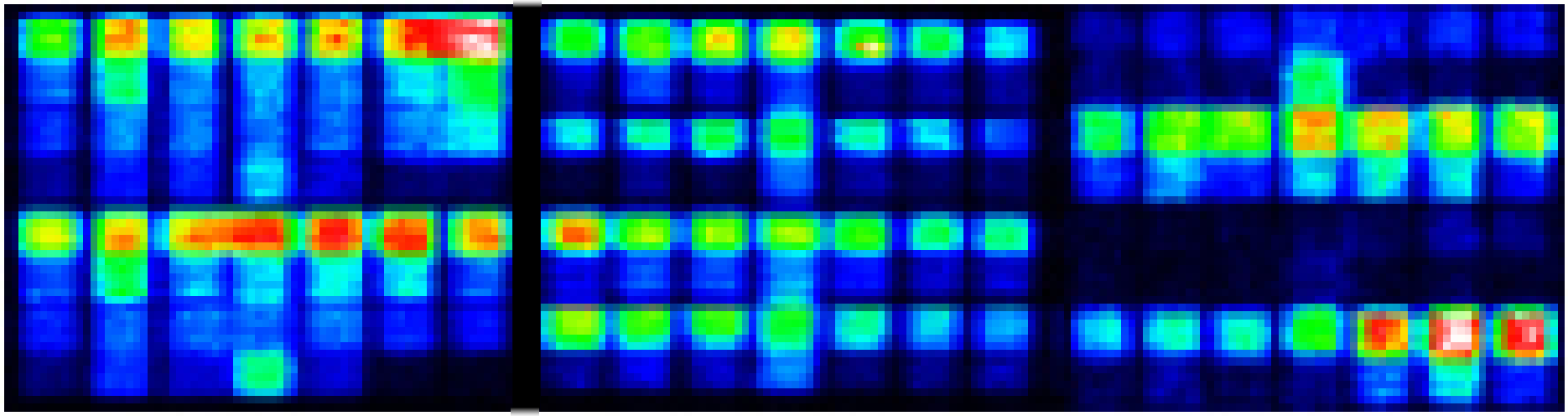


Building Models from High Throughput Biotechnology Data

Harri Kiiveri

CMIS Bioinformatics for Human Health

Feb 2006

Talk outline

1. **Background on high throughput biological data**

2. **Response modelling**

3. **Local gene network construction**

4. **Network simulation**

1. High Throughput Biological Data

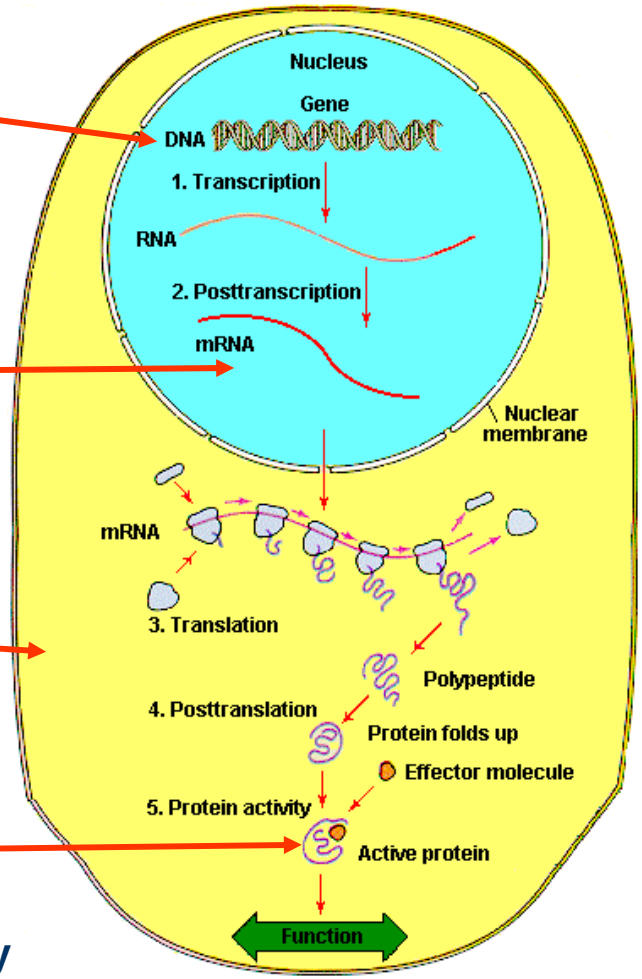
Probing the cell

Common DNA mutations- SNPs
- SNP chips

Gene expression
- microarrays

metabolites

Protein expression
- mass spectrometry
- gas chromatography



Features of the data

DNA sequence data – SNP chips

(measures millions of variables)

Gene Expression - microarrays

(measures 30,000 – 500,000 variables)

Protein expression – mass spectrometry

measures 100,000+ variables

Metabolites

measures 200,000+ variables for humans

The number of samples will typically be of the order of 100's

Many more variables than observations!

2. Response Modelling

Each sample has a characteristic or response that we would like to predict from our measurements “inside” the cell

y (n by 1) X (n by p)

Say n=100 and p=30000

Possible responses (y) of interest

- Binary – cancer vs healthy
- categorical – sub types of a disease
- ordered categorical – benign, cancer, metastasized (disease stages)
- continuous – survival time, obesity, seed size….
- gene expression itself

Algorithm for solving the problem

1. **Model the effect of each variable on the response as a variable specific weight times its value**

2. **Sum the effects over all variables**

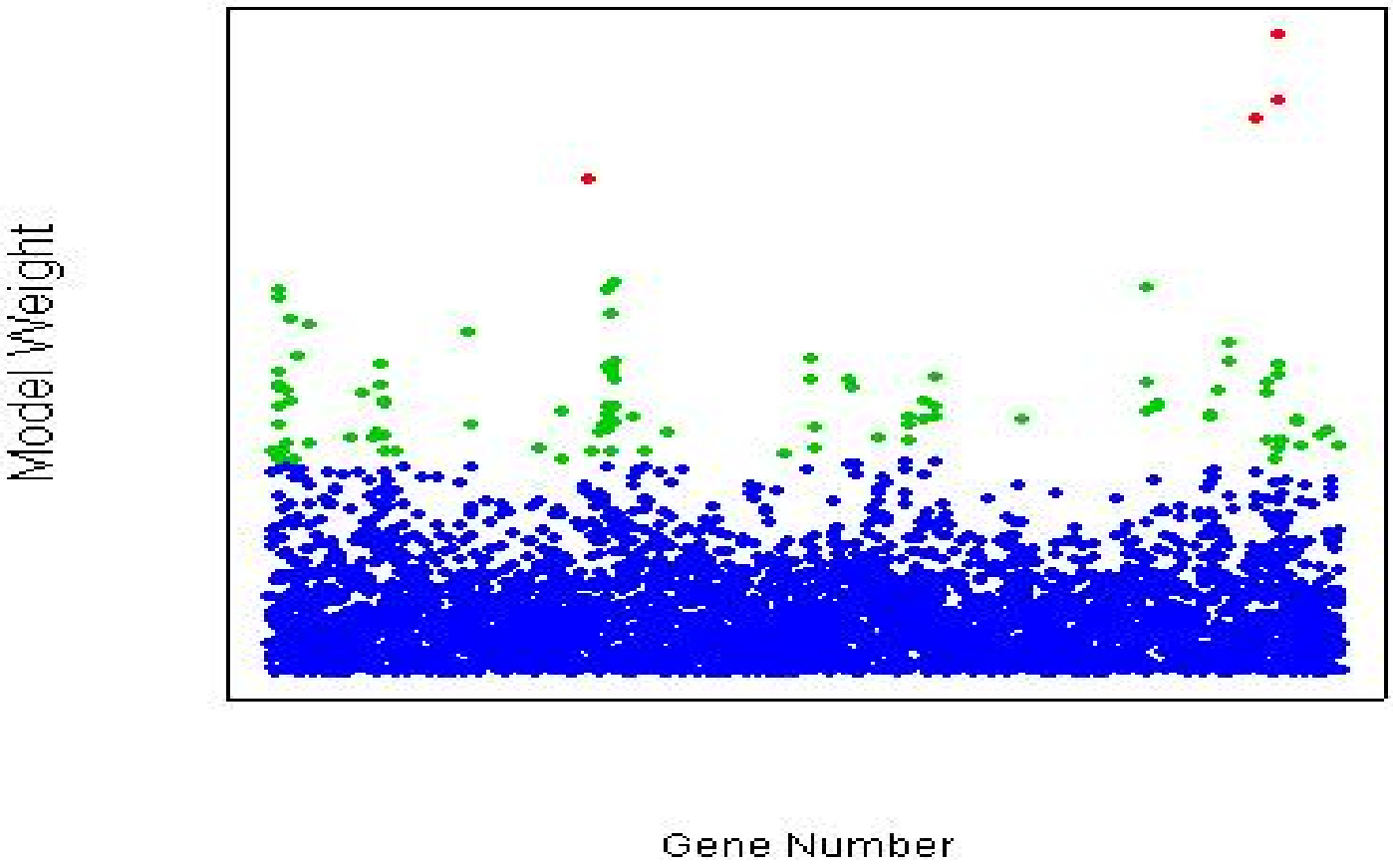
3. **Define a model which converts the total effects into a predicted response value**

4. **Assume that it is highly likely that a variable effect is zero**

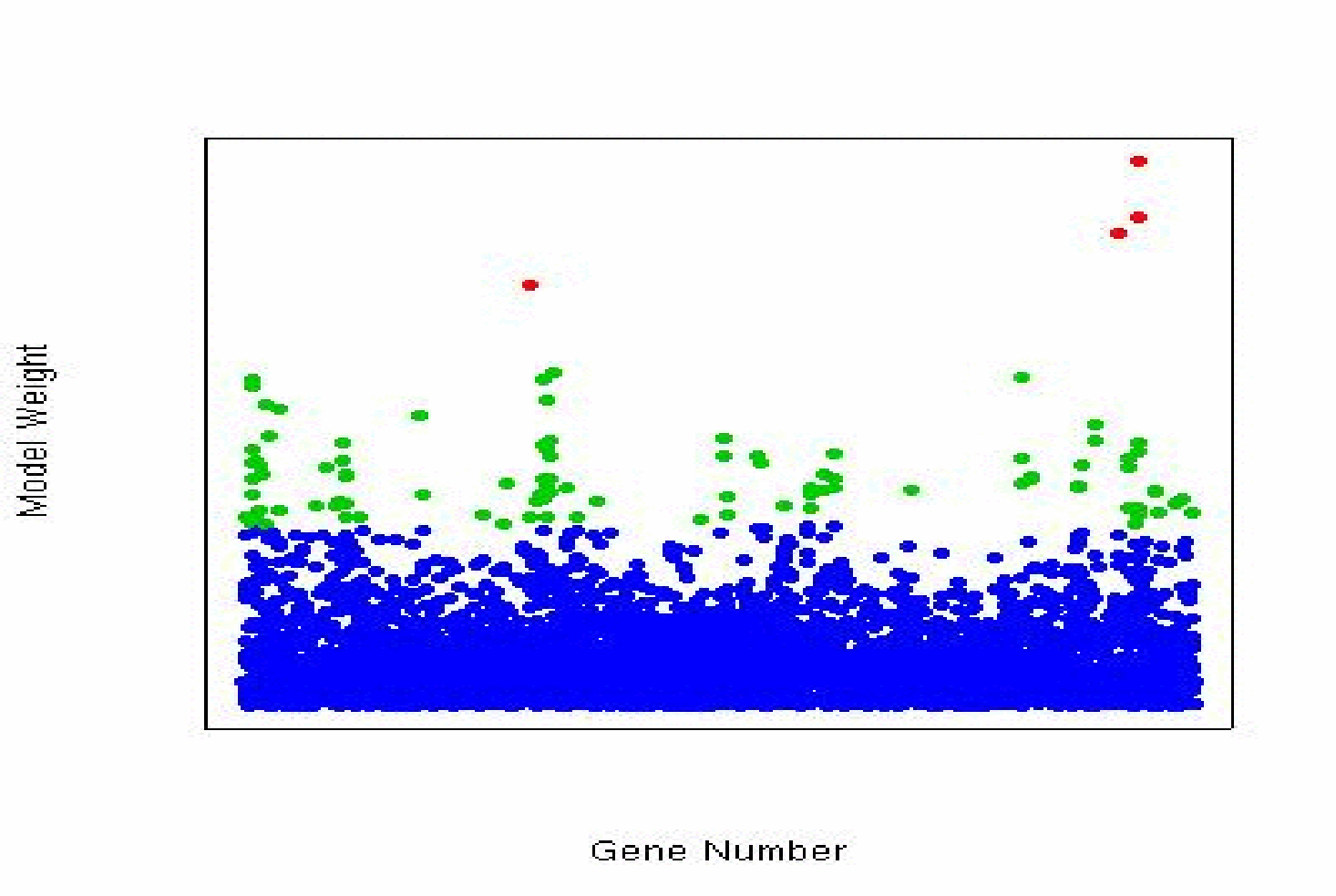
5. **Define a criterion for any set of weights which measures goodness of fit and model simplicity or sparseness**

6. **Search for the best set of weights to give to each variable (variable selection and parameter estimation are simultaneous)**

GeneRave in Action



GeneRave in Action



Examples

St Jude's leukemia data (6 classes)

n=104 p=44,000+ "genes"

predicting leukemia subtype

Perlegen SNP data

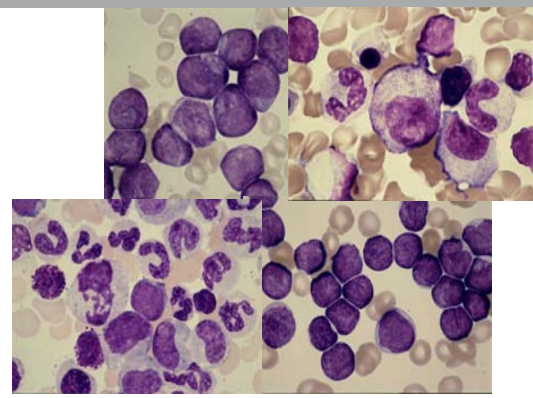
n=71 p=1,500,000 - SNP's

(3 million variables)

predicting sex and race

Examples run in

Example 1: St Jude's Leukaemia data

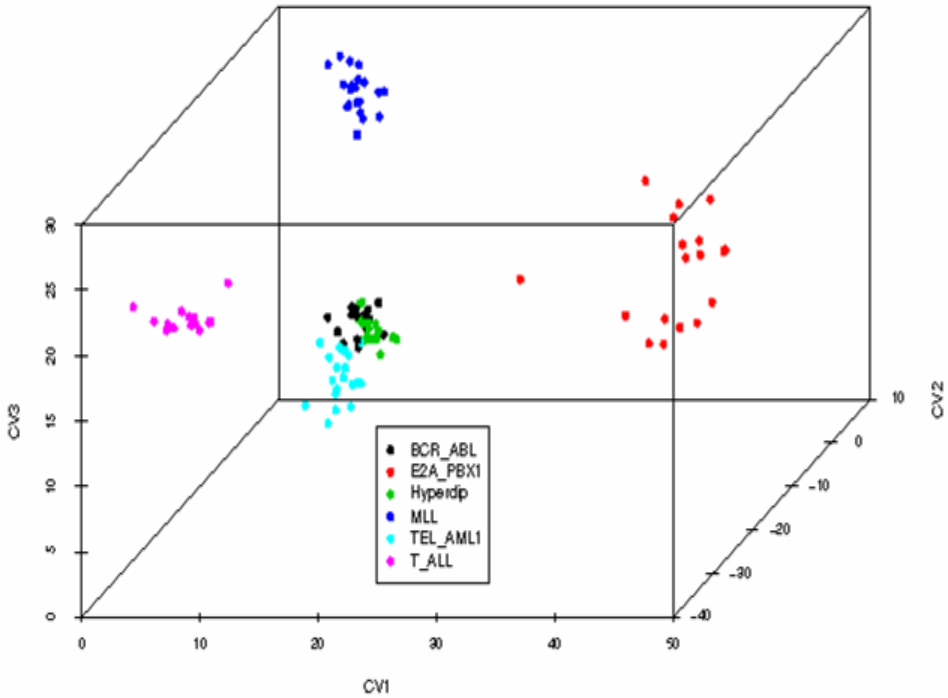


p = 44,000 "genes" or >500,000 probes (Affymetrix U133A/B)

- n = 104 samples
- 6 leukaemia subtypes

Results

- 6-gene classification model
- Cross-validated error < 5%



- Validated with PCR data
- Explore genes related to the 6 predictors…

Example 2: Perlegen SNP data

Reference:

Whole-Genome Patterns of Common DNA Variation in Three Human Populations.(2005) Hinds et al, Nature (2005).

http://genome.perlegen.com/browser/download.html

71 individuals **~1.5 million SNPS**

33 males

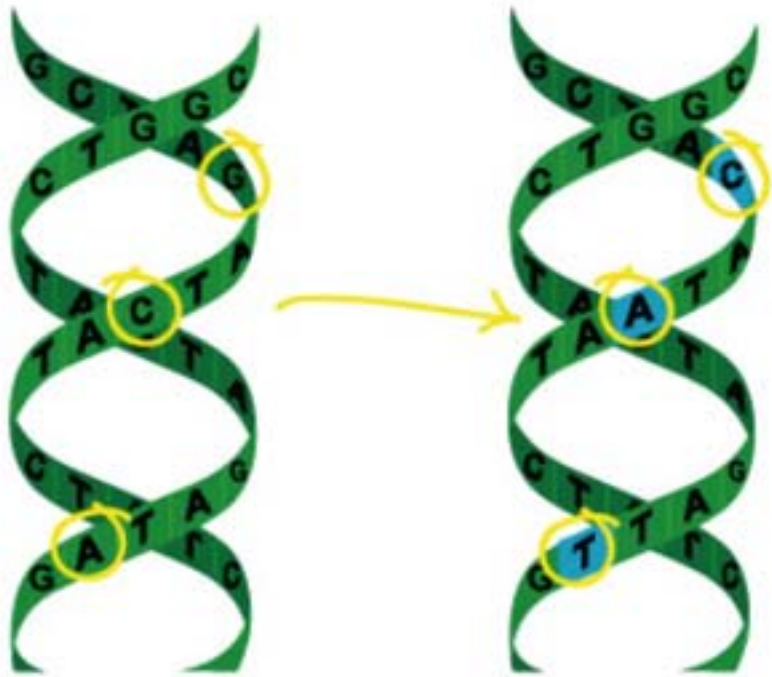
38 females

23 African Americans

24 European Americans

24 Han Chinese

Single Nucleotide Polymorphisms

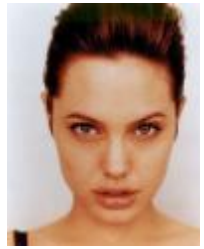


SNP Statistics

- Estimated SNPs in human genome: 10 million
- Number that have been seen twice: about two million

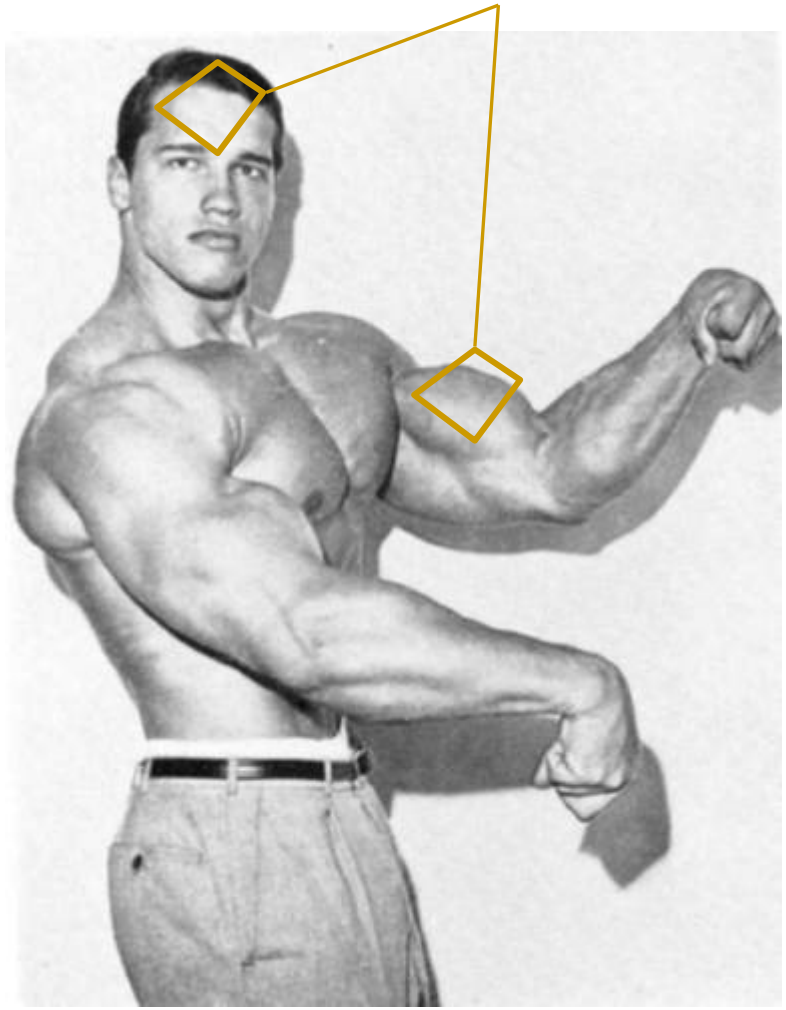
SNP

AGCTCCTAAGCTTAAGCTACT
AGCTCCTAACCTTAAGCTACT
AGCTCCTAAGCTTAAGCTACT
AGCTCCTAAGCTTAAGCTACT

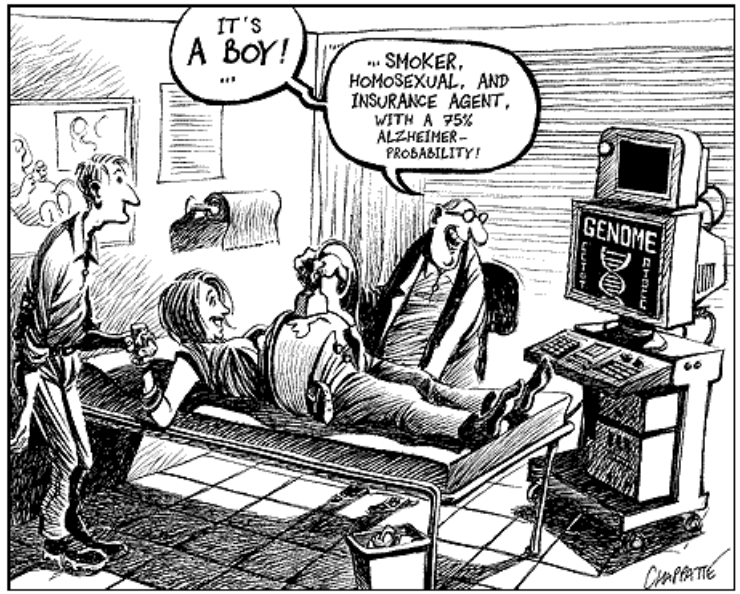


SNP's are a major determinant of phenotype

quantitative traits



- **Strength**
- **Intelligence**
- **Response to drugs**



Data and model

We fit a sparse “main effects” model to the data using the GeneRave algorithm

On an appropriate scale each SNP genotype has an additive effect on the probability of race or sex.

Most effects are expected to be zero and the effects of a small number of SNP genotypes will dominate

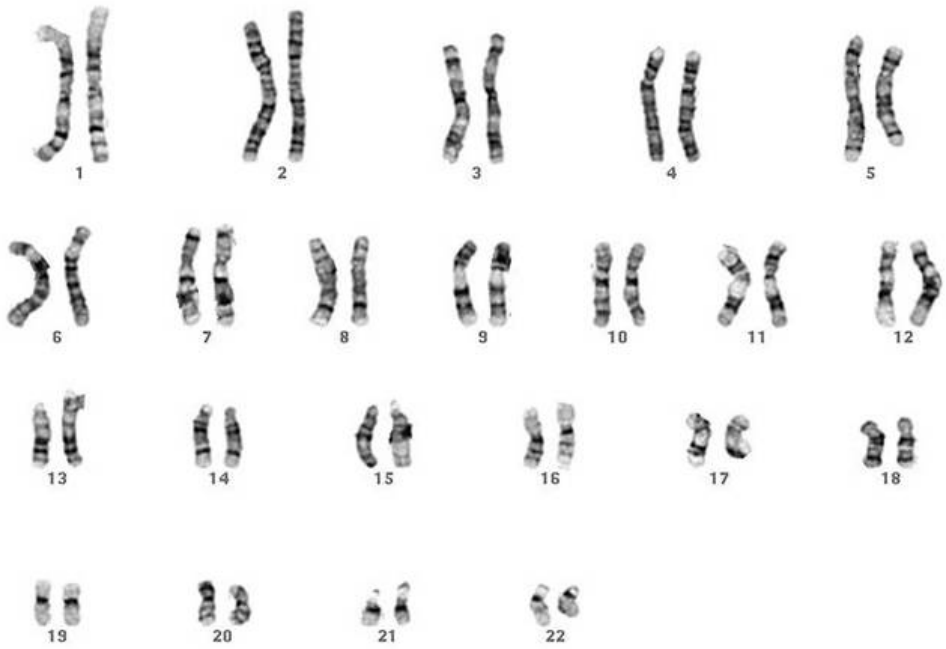
For the Perlegen SNP data there are 71 samples and 3,096,617 variables !!

GeneRave – Perlegen SNP Data

1,548,308 SNPS on chromosomes 1 to 22

Race data
- 23 african americans,
- 24 european americans
- 24 han chinese

Sex data
- 33 males
- 38 females

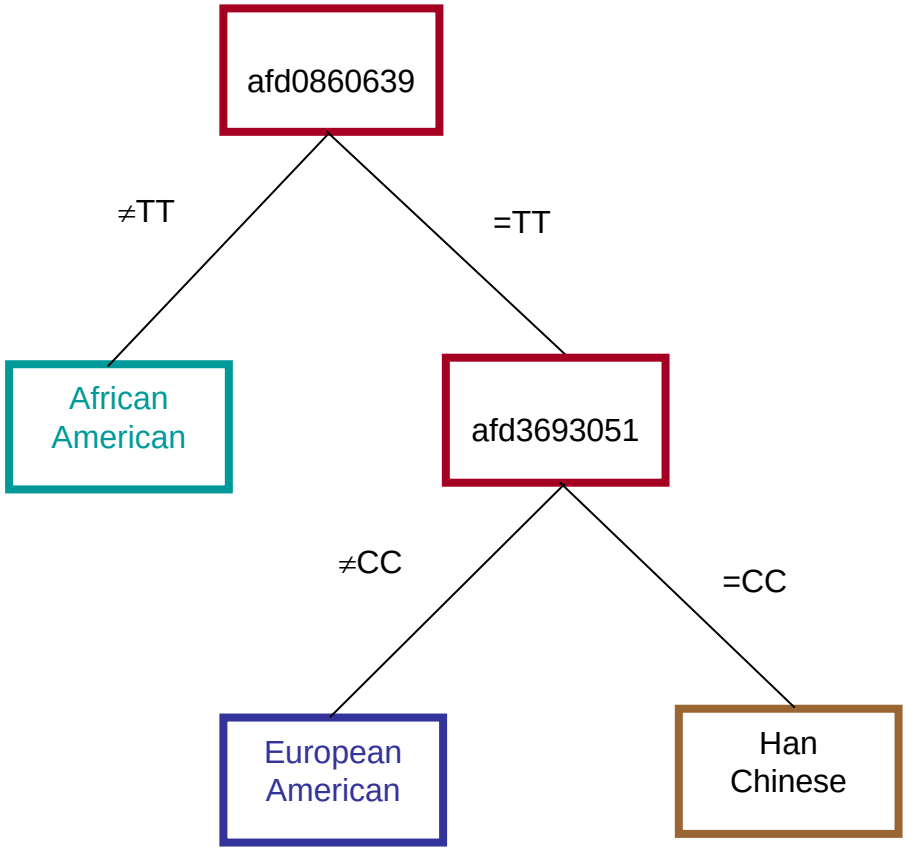


Results

Race
- 3 SNPs (0.082)

Sex
- 2 SNPs (0.00)

SNP race classifier



Validation data - Hapmap data set

http://www.hapmap.org

270 individuals

142 males

128 females

~5 million SNPS

90 Utah residents (European Americans)

45 Han Chinese

45 Japanese

90 Yoruba in Ibadan Nigeria

Independent validation of results

The SNPS picked up in the GeneRave analysis have been genotyped in the Hapmap project

The SNP on chromosome 1 classifies males and females in the Hapmap data set with zero error

The SNP on Chromosome 15 doesn't

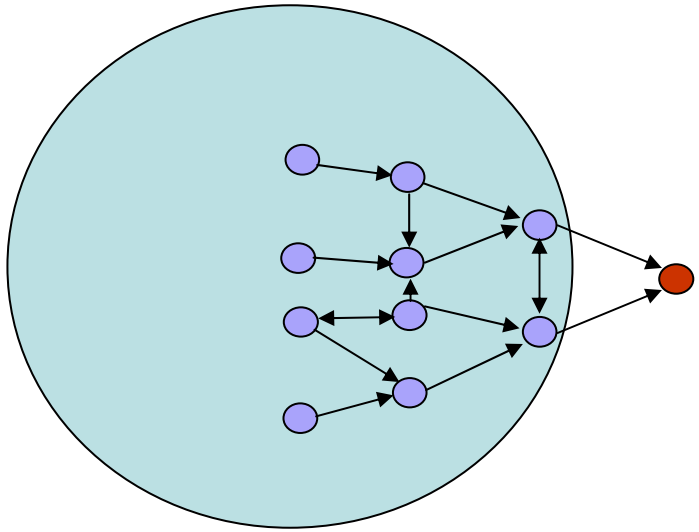
The SNP from the Perlegen Analysis which classifies Han chines and European Americans works in the validation data with zero error

SNP Analysis Conclusion

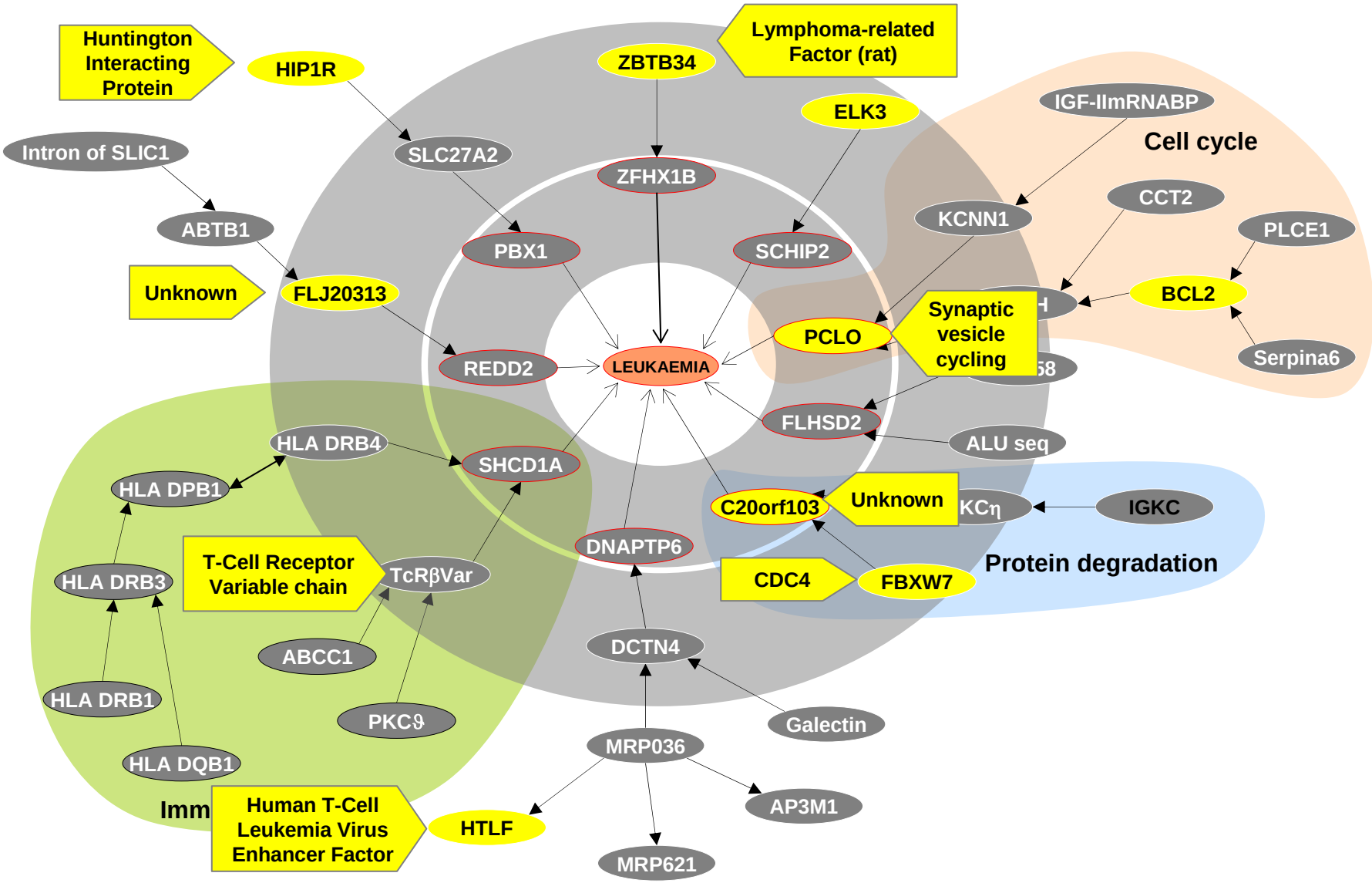
- **The sex SNP on chromosome 1 is highly likely to be a cross hybridisation problem with the SNP Chips**

- **The Race SNP is associated with a gene which codes for skin colour**

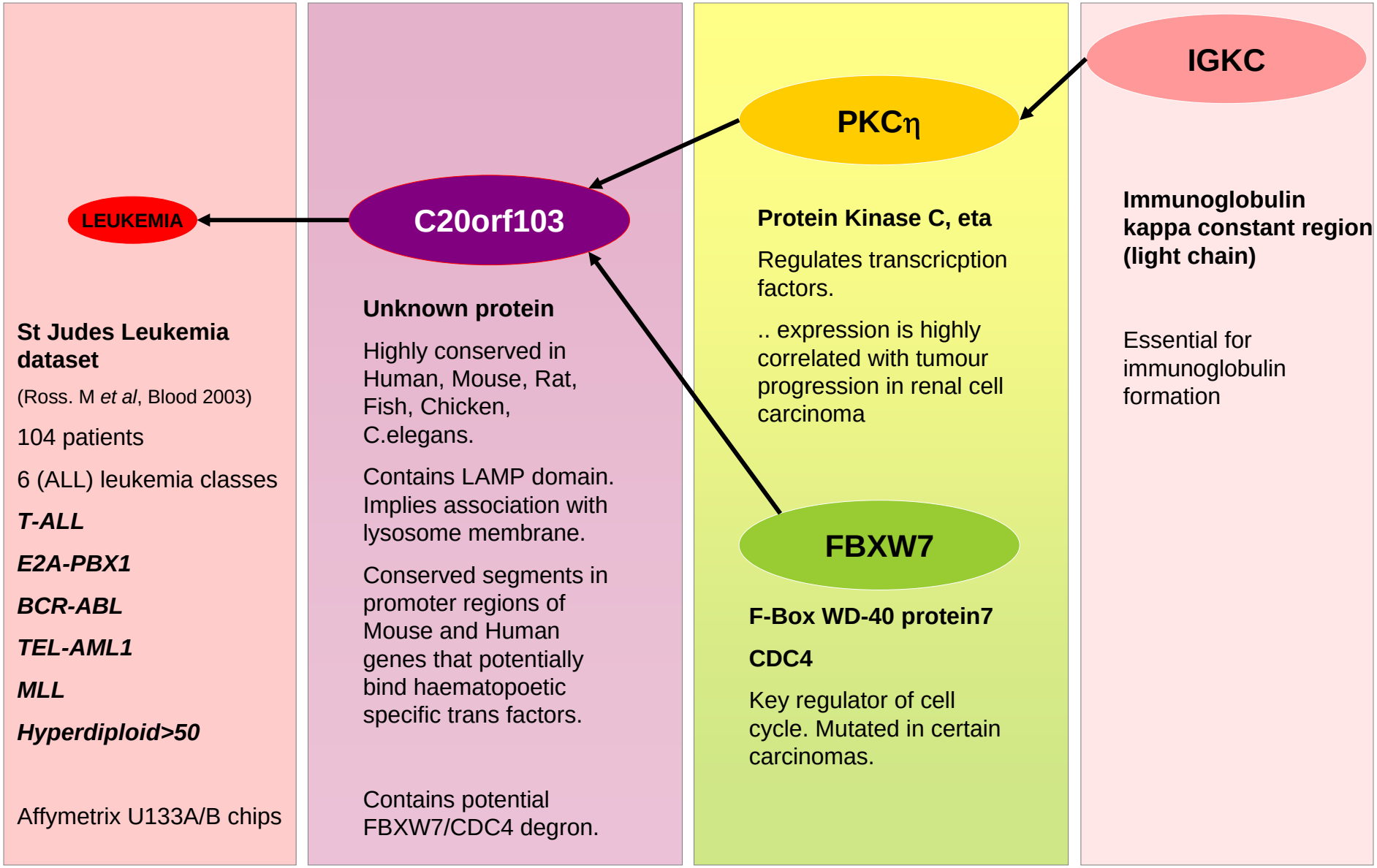
3. Local gene network construction



GeneRave - Sparse Networks



Hypothesis Testing



Networks - An Exploratory tool

Should consider these networks as exploratory data analysis

Hopefully suggestive of Hypotheses and further LAB experiments

Building Gene Networks
using additional information

The algorithms can use other data sets to improve the network construction algorithms

For example

- **Protein-protein interactions**
- **Sequence information**
  - **Transcription factor binding sites in a genes promoter region**

4. Network simulation

Luo et al prostate cancer data

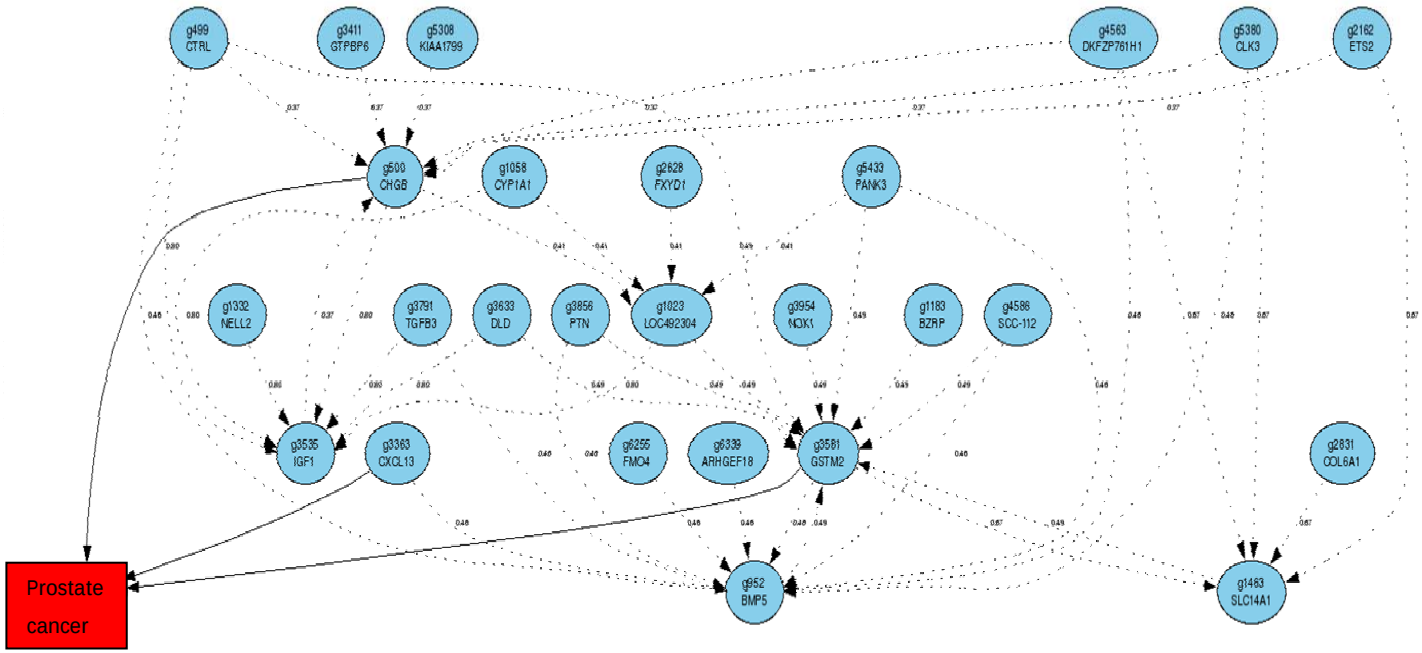
25 subjects

16 malignant

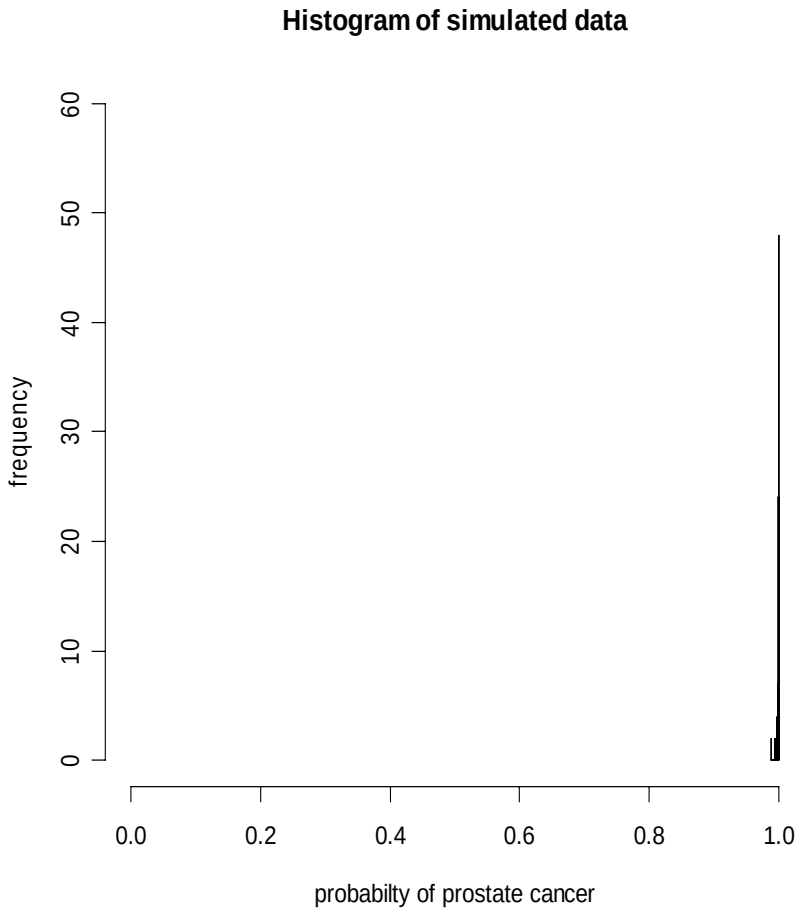
9 benign

Expression measurements for 6500 genes

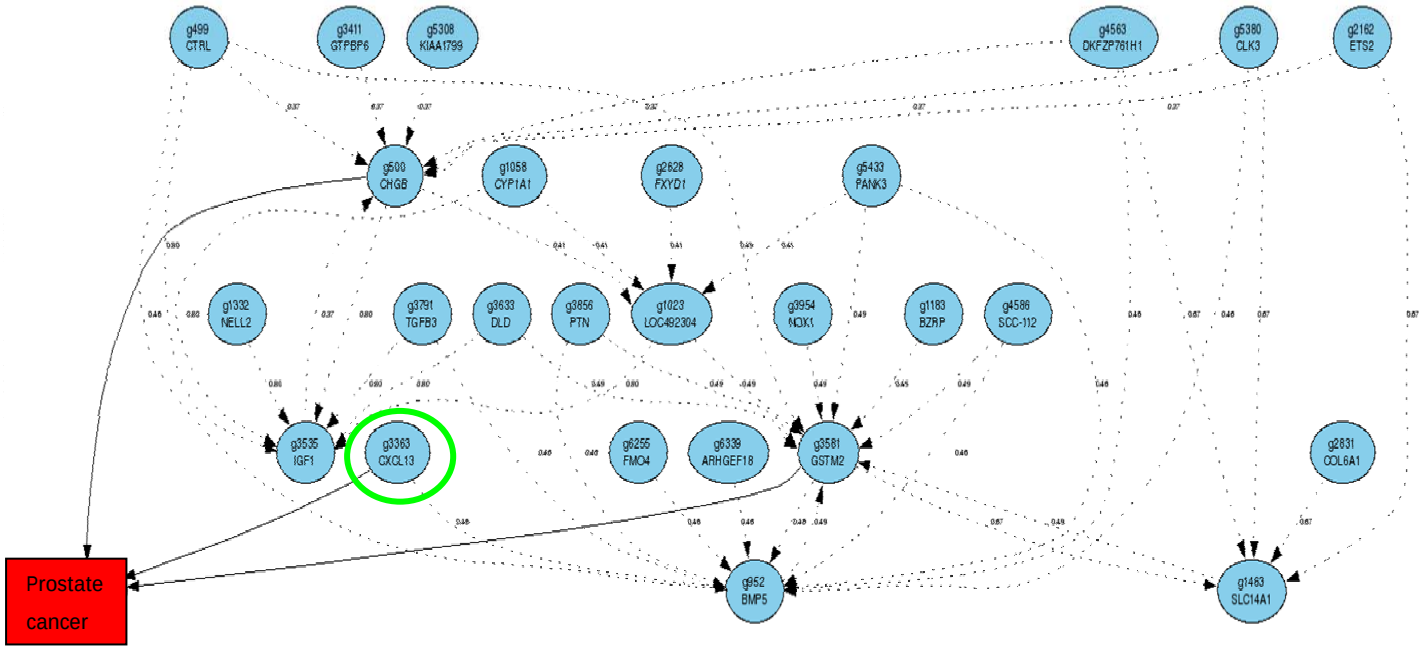
Prostate cancer network



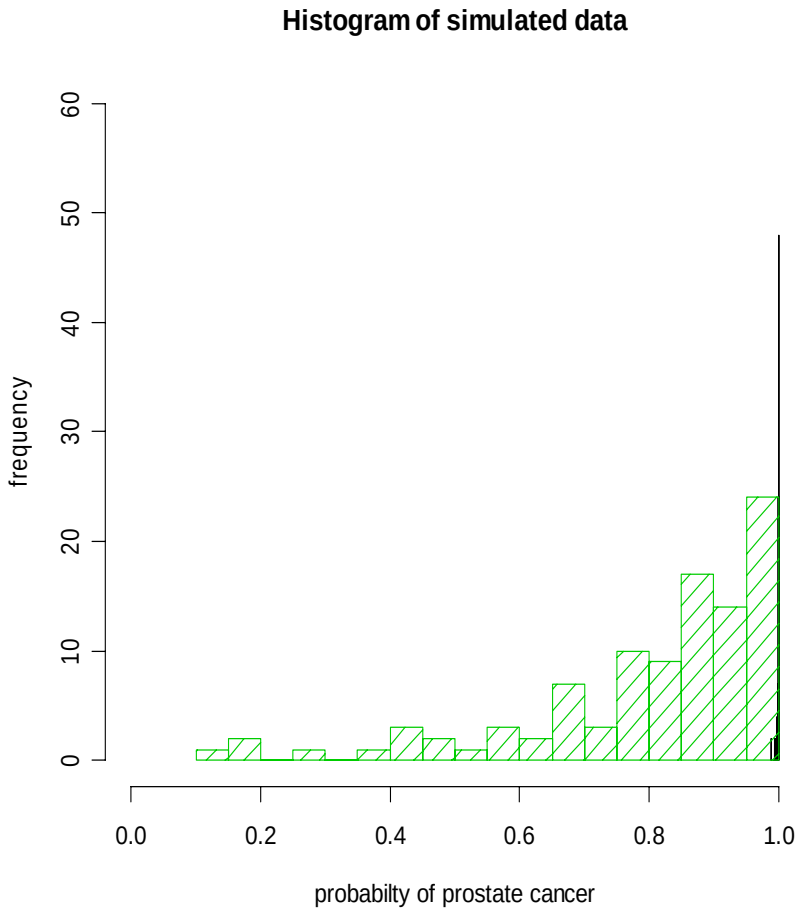
Simulation of 100 observations



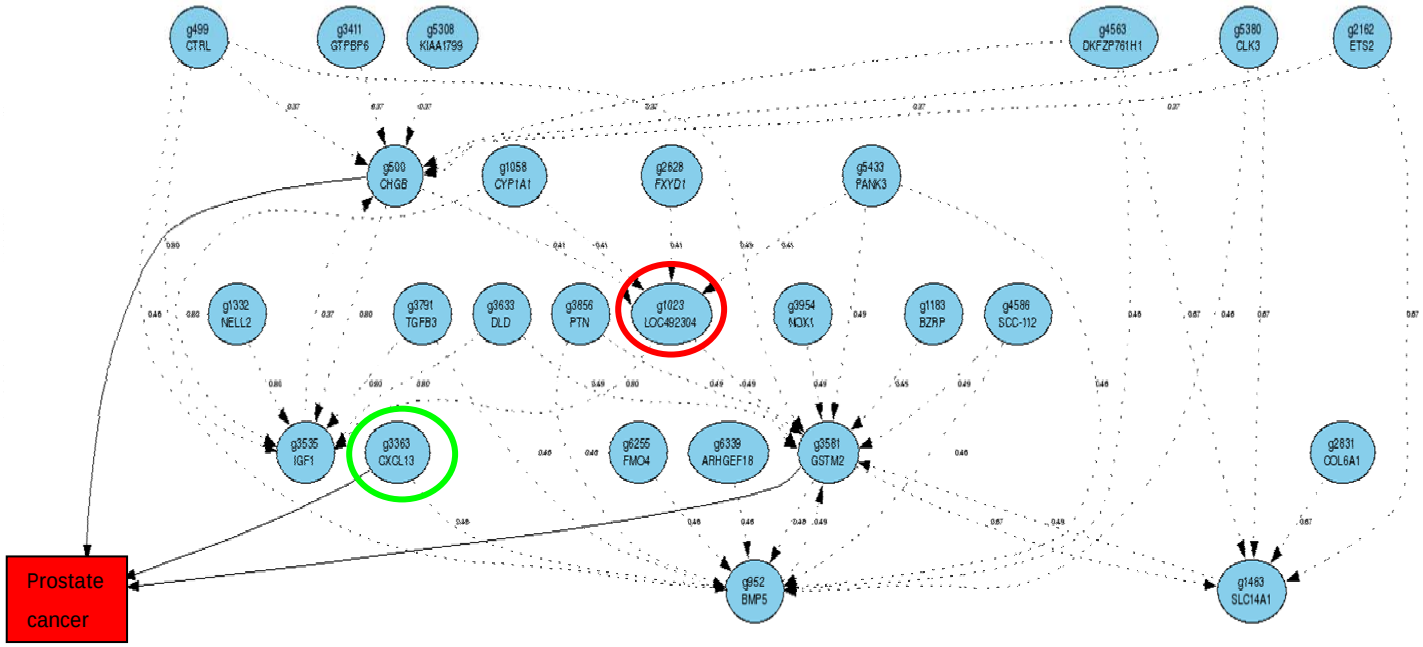
Prostate cancer network



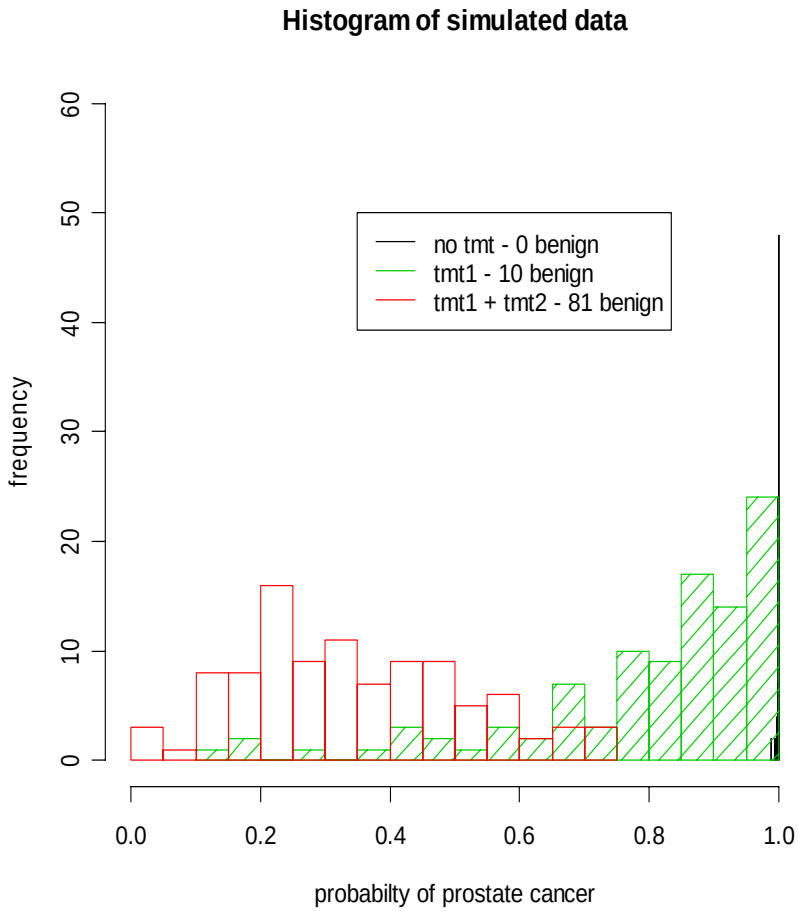
Effect of controlling gene expression



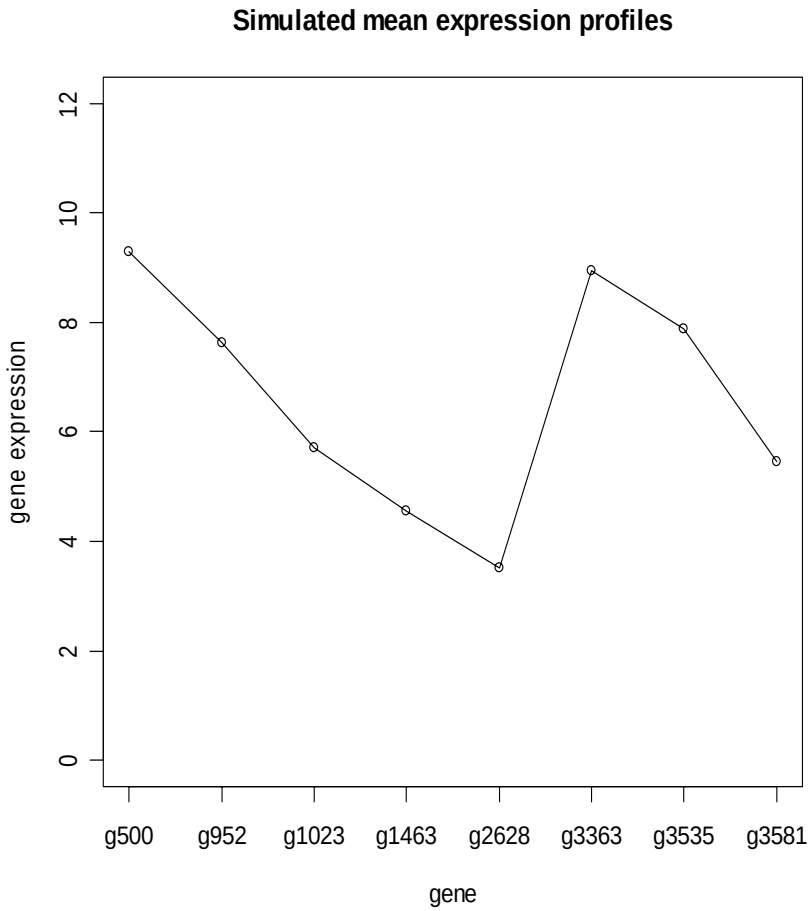
Prostate cancer network



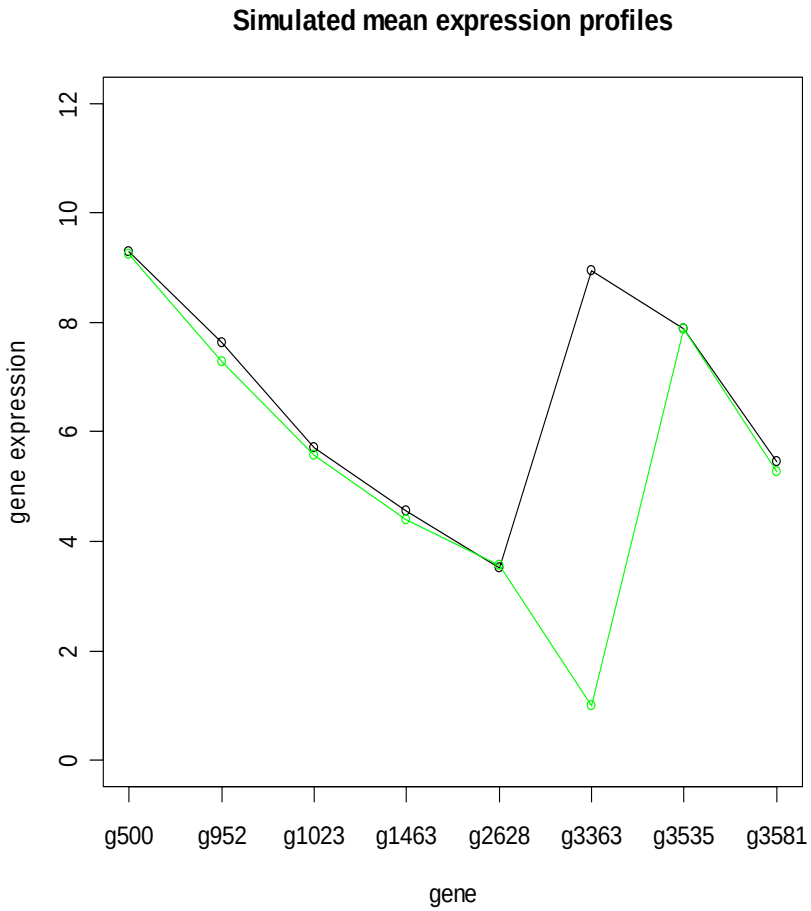
Effect of controlling gene expression



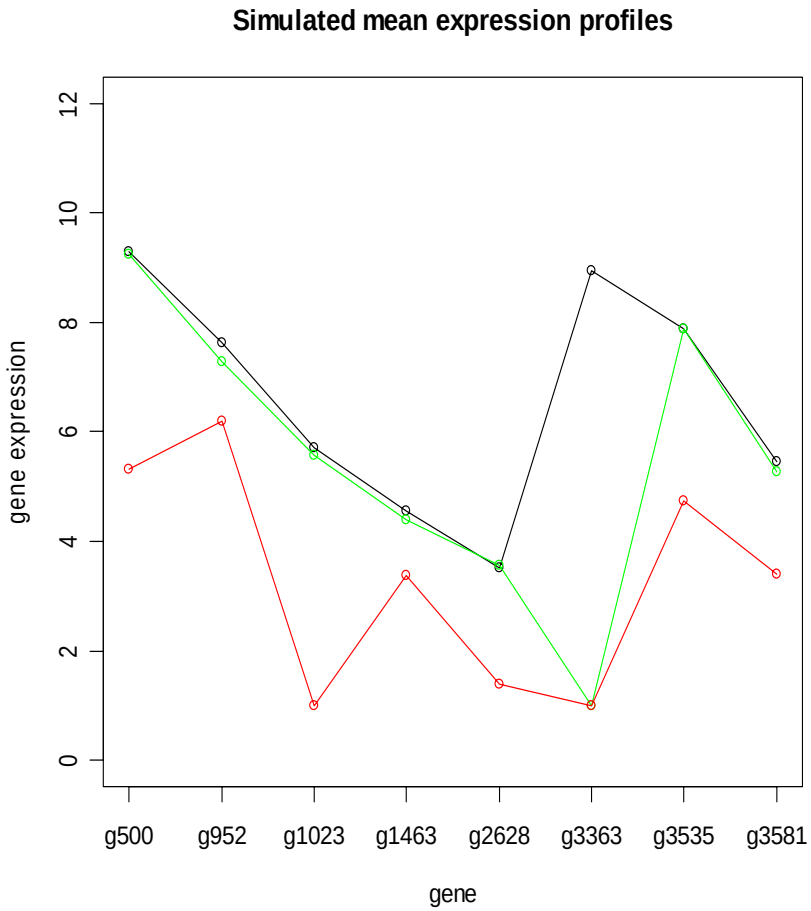
Side effects ?



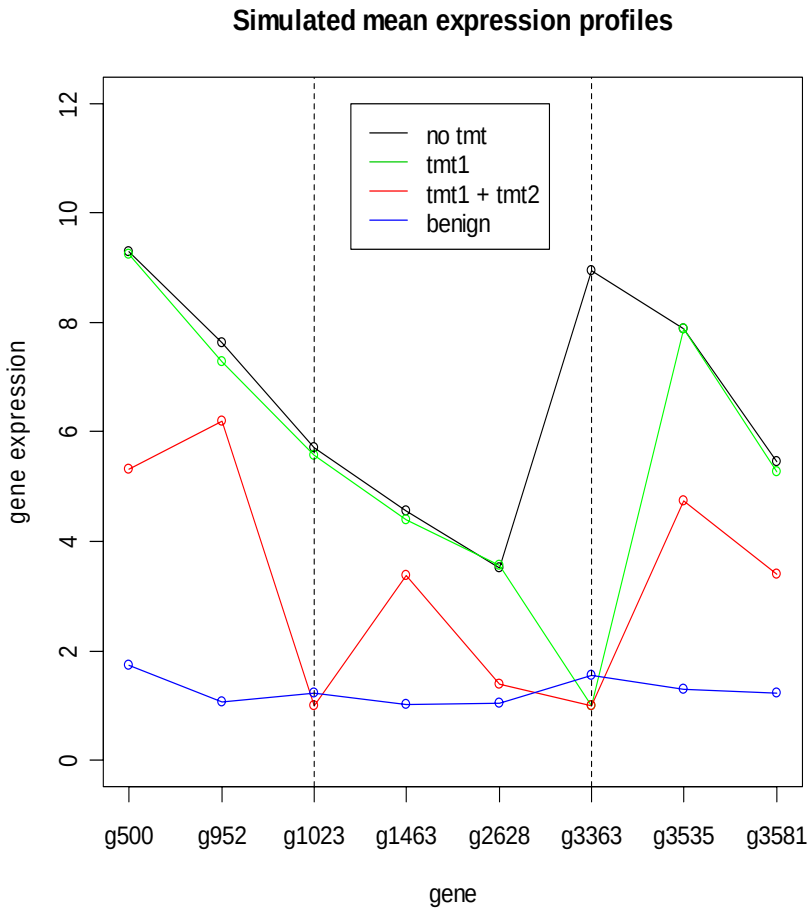
Side effects



Side effects



Side effects



Contact

Name Harri Kiiveri

Title Research Scientist

Phone 61 8 9332 3317

Email Harri.Kiiveri@csiro.au

Web www.cmis.csiro.au/BHI

Contact

Name Rob Dunne

Title Stream Leader

Phone 61 2 9325 3263

Email Rob.Dunne@csiro.au

Web www.cmis.csiro.au/BHI

Thank You

Contact CSIRO

Phone 1300 363 400

+61 3 9545 2176

Email enquiries@csiro.au

Web www.csiro.au