Sparse partitioning: a tool for analyzing regression problems with many tertiary predictors

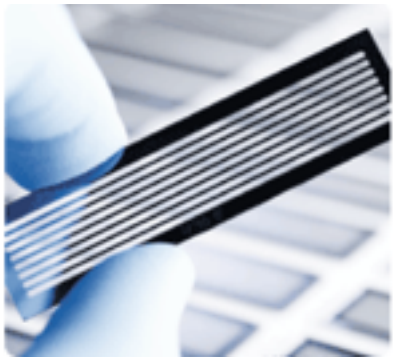
Simon Tavaré

CRI & DAMTP, University of Cambridge

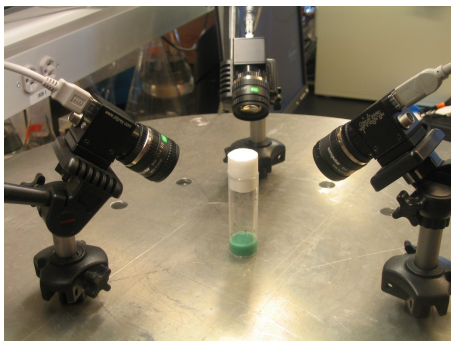
Molecular & Comp Bio, USC

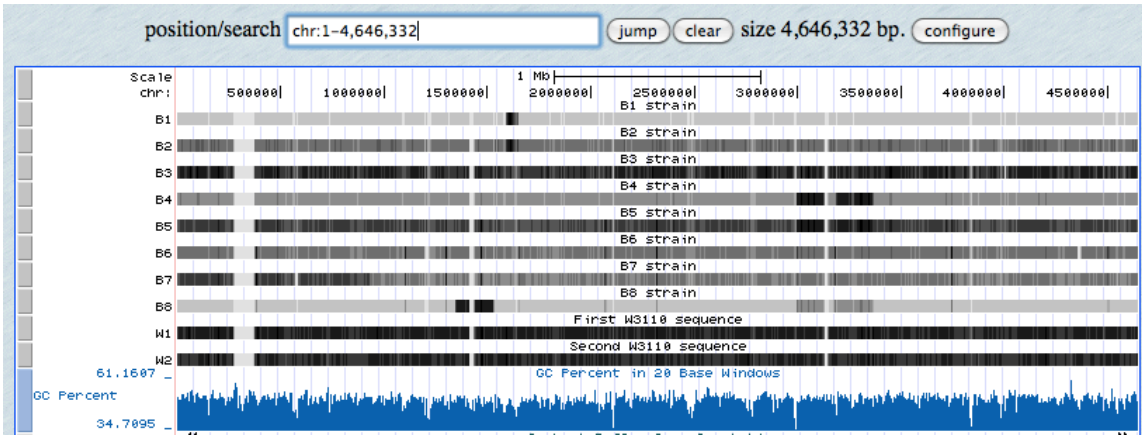
BioC 2010. Seattle, July 30 2010

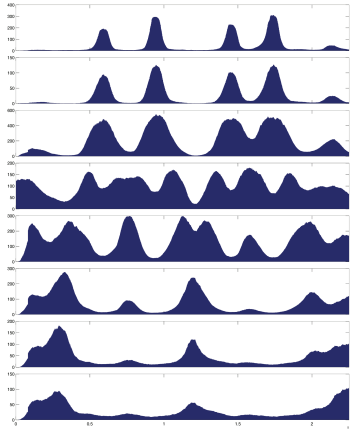
This-generation sequencing



USC CEGS: the genotype-phenotype map

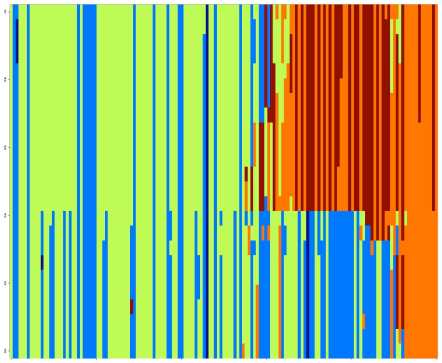




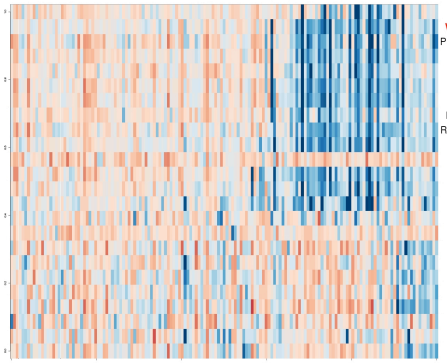


CRI: METABRIC

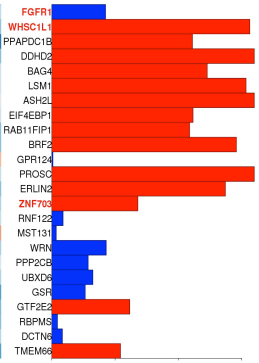
Copy number



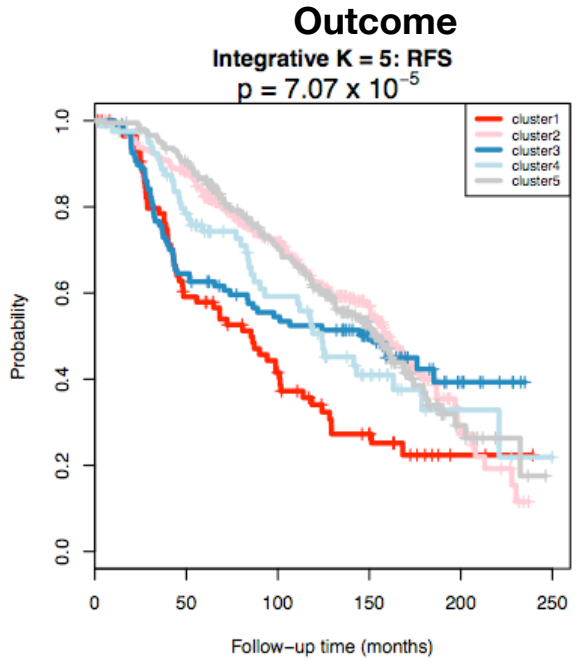
mRNA expression



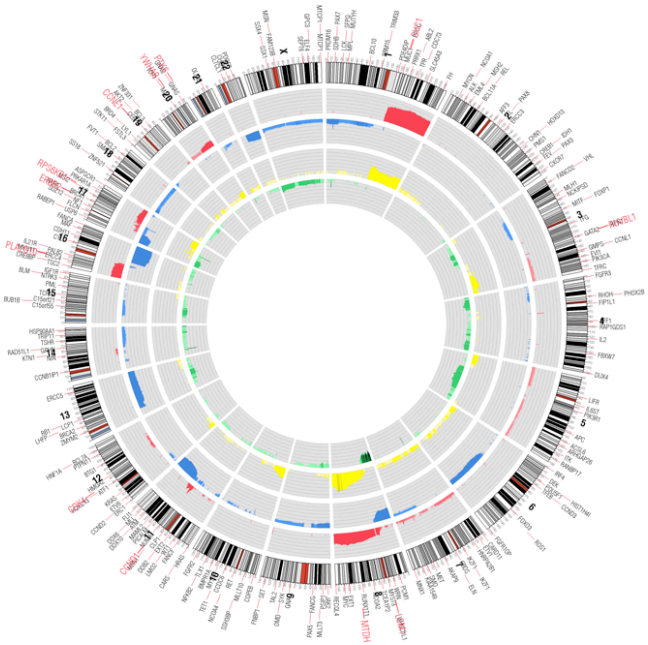
-Log10 pval



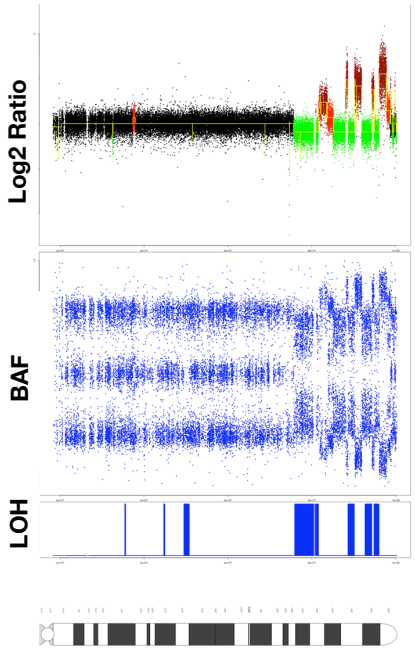
Outcome



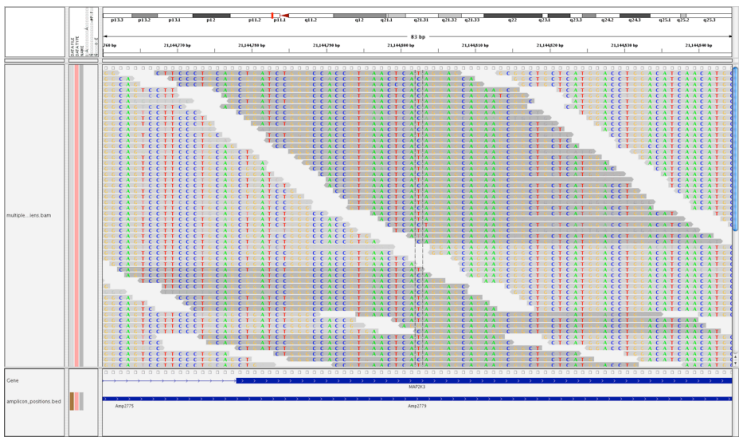
Recurrent alterations



Allele-specific copy number



Sequencing pileups



A genome-wide cancer CNV map

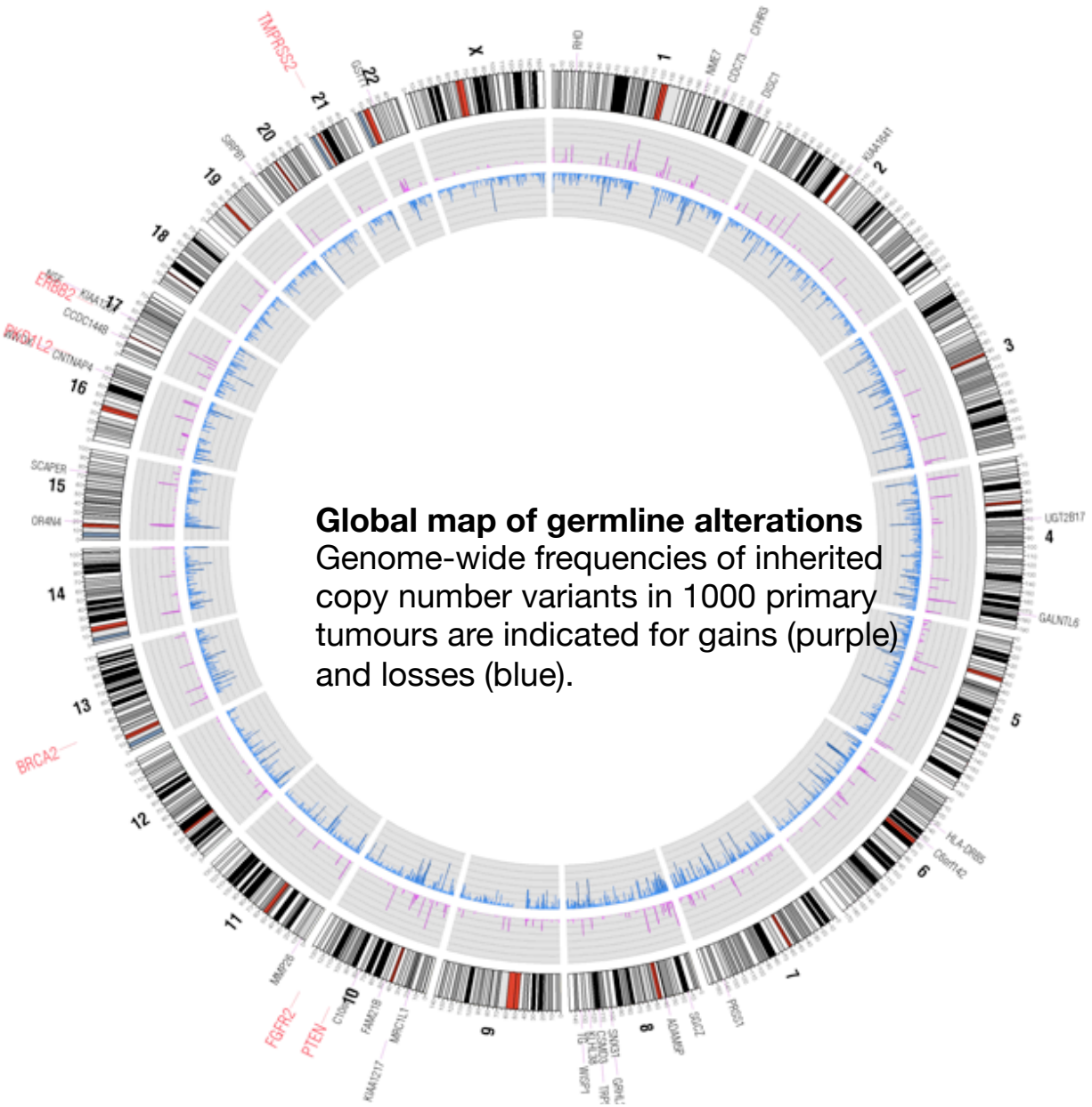
- Median of 600 CNVs per matched normal sample

-40% overlap with Conrad et al, where CNVs were discovered in 2 HapMap populations (CEPH,YRI)

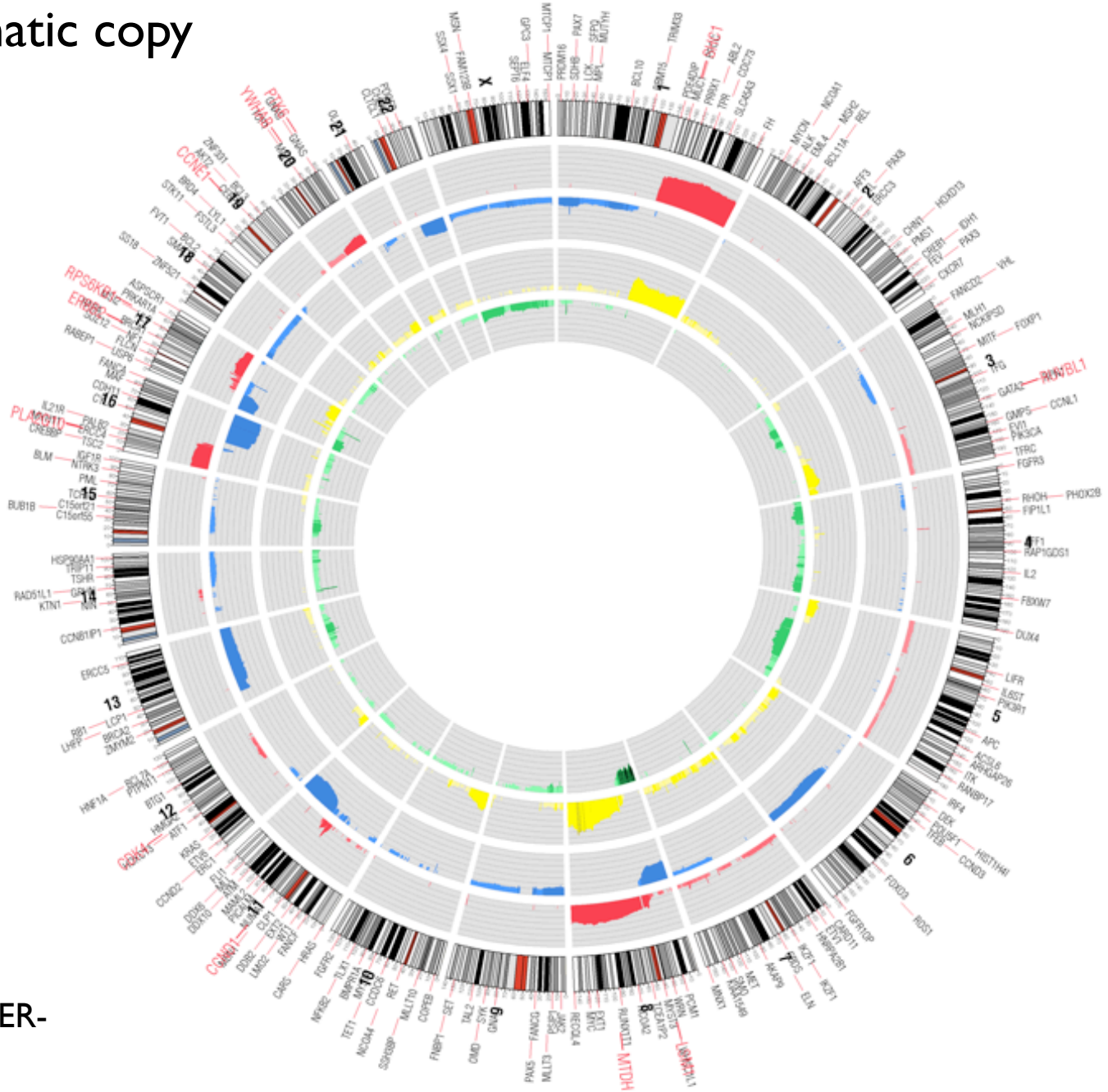
-focal CNVs include those spanned by a single probe (17%)

-Several known breast cancer genes are targeted by focal CNVs

-These variants affect a substantial proportion of expression variation



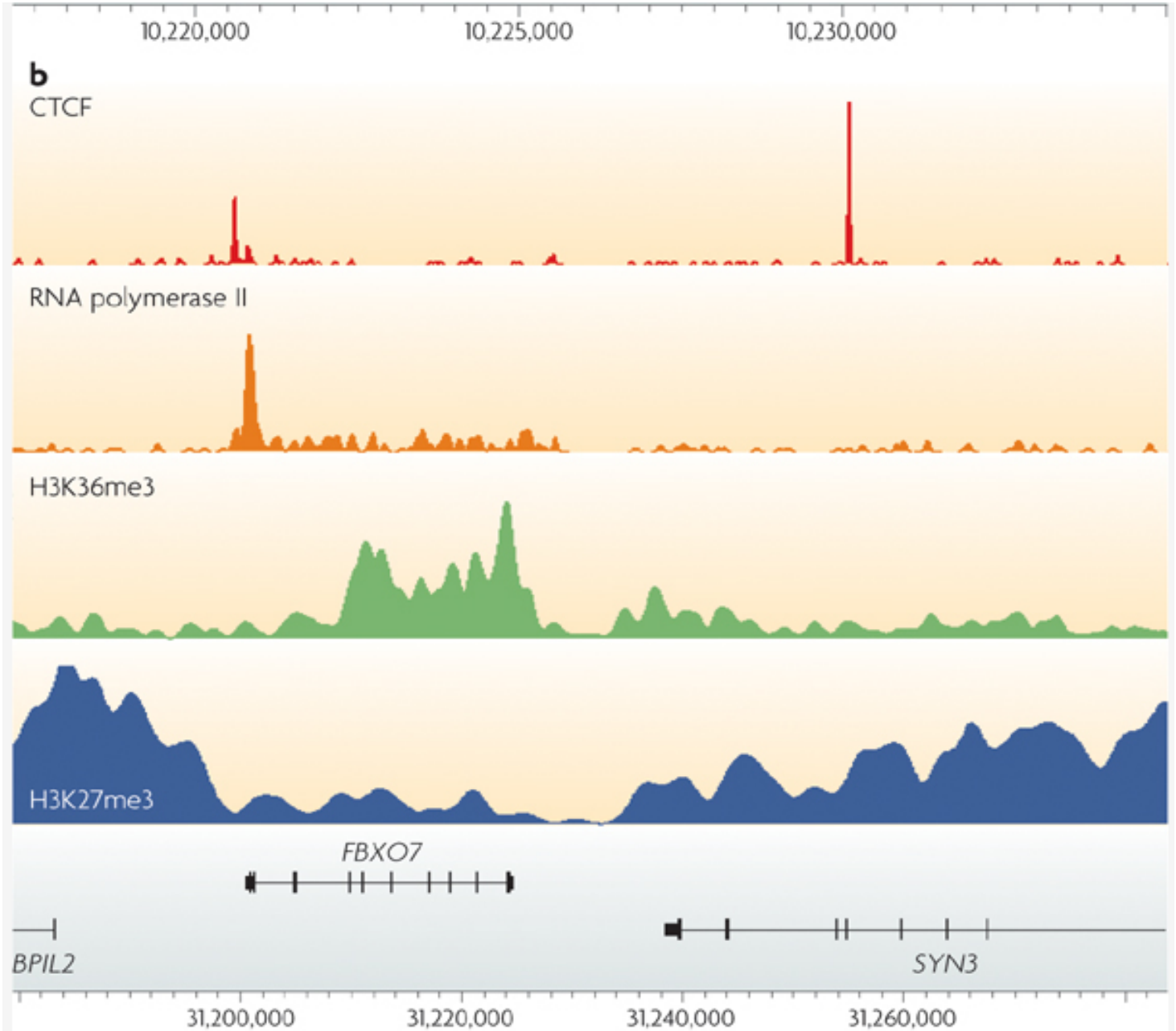
The landscape of somatic copy number alterations



Global map of somatic aberrations in ER positive versus negative disease
Genome-wide frequencies of acquired gains and losses are indicated for ER+ (red, blue) and ER- (yellow, green) cases.

beadarray

- Bioconductor package for *bead-level* analysis of Illumina bead arrays (M. Dunning)
- Pipeline for analysis of multiple samples (BADGER)
- Artefact detection: BASH, HULK, registration (A. Lynch, M. Smith, J. Cairns)
- Annotation a problem (N. Barbosa-Morais, NAR 2009) ReMOAT (S. Samarajiwa)
- *CRLMM* for Illumina SNP chips (B. Carvalho, M. Ritchie)



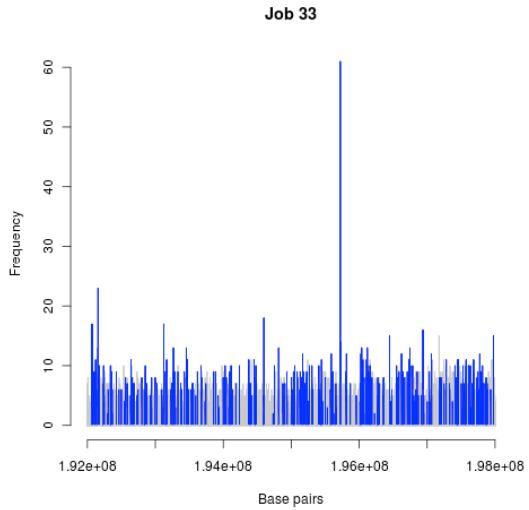
Park J, Nature Reviews Genetics, 2009

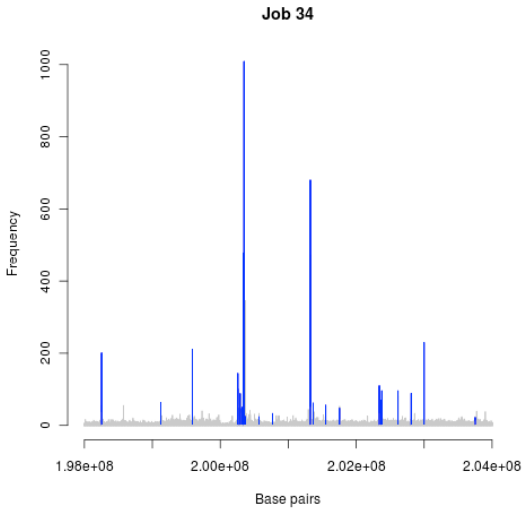
(Some) peak callers

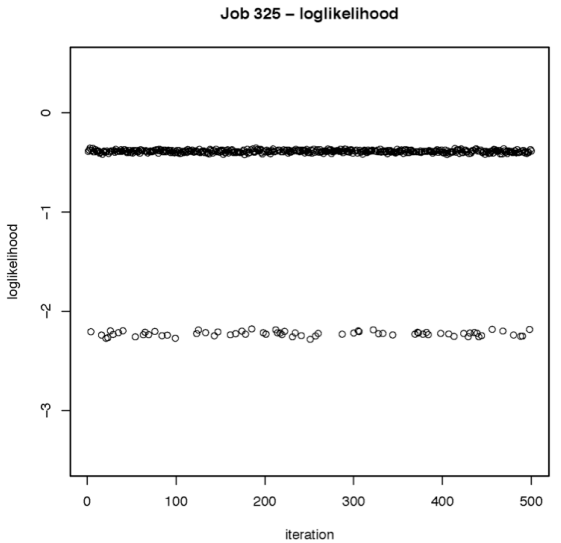
	IP only, & control, either	read features	data from different strands	masks genomic repeats	scoring criteria	confidence in results, FDR estimates sensitivity / specificity	for both TF & HM
CSPF	& / or	read length no orientation	merges strands	N	simple height criteria	empirically: ROC curve	both
XSET	& / or	mean fragment length orientation	merges strands	N	simple height criteria	FDR based on randomised sample and Poisson probabilities	both
Mikkelsen et al.	IP only	no orientation	no merge / shift	Y	p-values produced by randomising the datasets	no official FDR	both
MACS	& / or	mean fragment length orientation ignores duplicated reads	shifts reads merges strands	N	Poisson p-values	FDR = no. peaks in control : IP	both
QuEST	&	orientation	shifts reads merges strands	N	kernel density estimation	FDR based on calling peaks in 1/2 the control sample	TF
FindPeaks	IP only	mean fragment length orientation	no merge / shift	N	simple height criteria	Monte-Carlo based FDR (ie. from randomised sample)	both
SISSR	& / or	mean fragment length orientation	no merge / shift	N	compares read density on different strands	FDR comparing simulated background peaks to real data	better for TF
Kharchenko et al.	&	orientation	no merge / shift	N	Poisson probabilities	FDR based on different randomised versions of the input sample	better for TF
PeakSeq	&	mean fragment length orientation	merges strands	Y	pre-processing: normalisation Binomial p-values	FDR: q-values after multiple correction adjustment	both
BayesPeak	& / or	mean fragment length orientation	no merge / shift	N	Negative Binomial distribution Bayesian posterior probabilities	posterior probabilities of enrichment presence	both

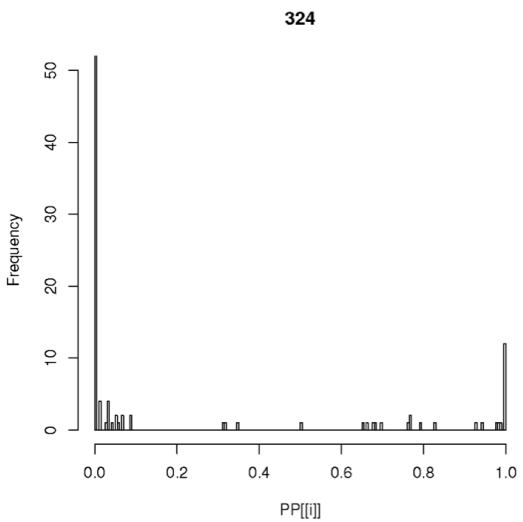
Spyrou et al., *BMC Bioinformatics*, **10**: 299, 2009

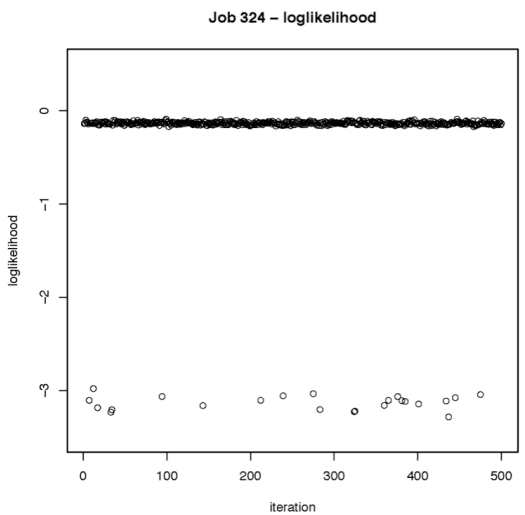
BayesPeak (genome-wide)
J. Cairns

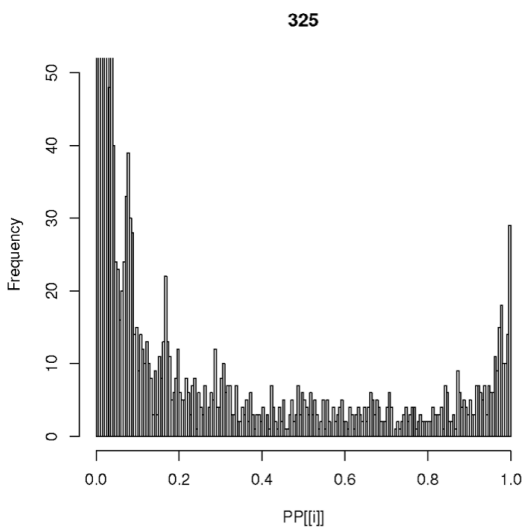






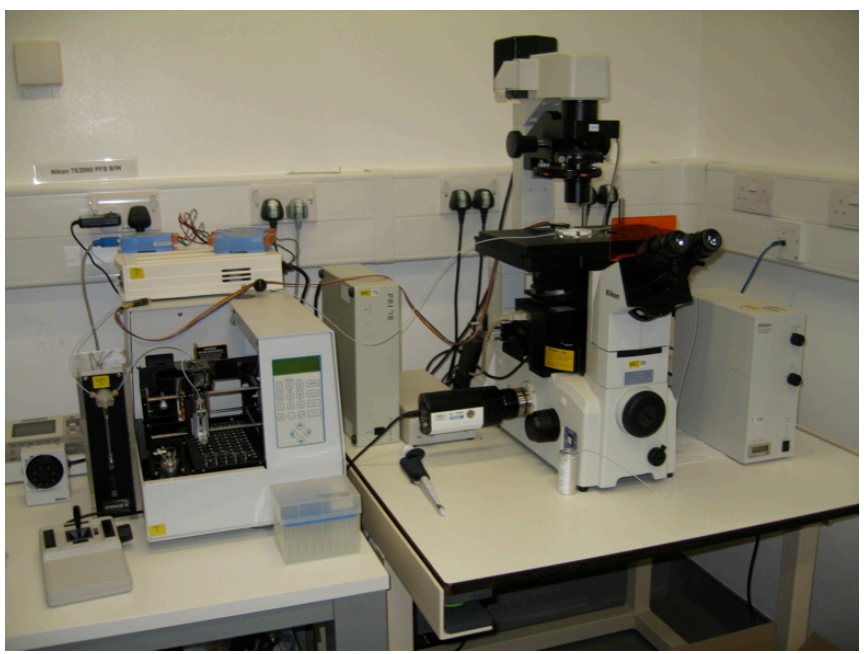






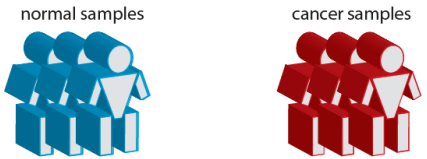
CNAnova

S. Ivakhno, *Bioinformatics*, 2009



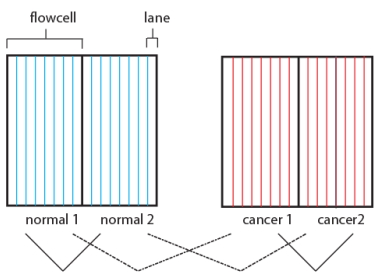
HMMseg

1. Pre-processing and normalization of the dataset

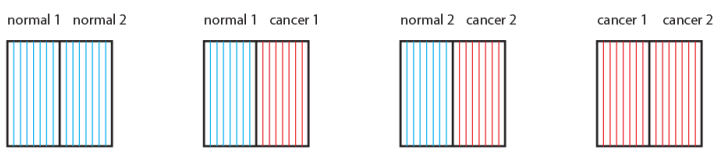


Extract read counts from BAM files.
GC-normalize the data
Perform read count smoothing using discrete wavelet transform (DWT)

2. HMM segmentation

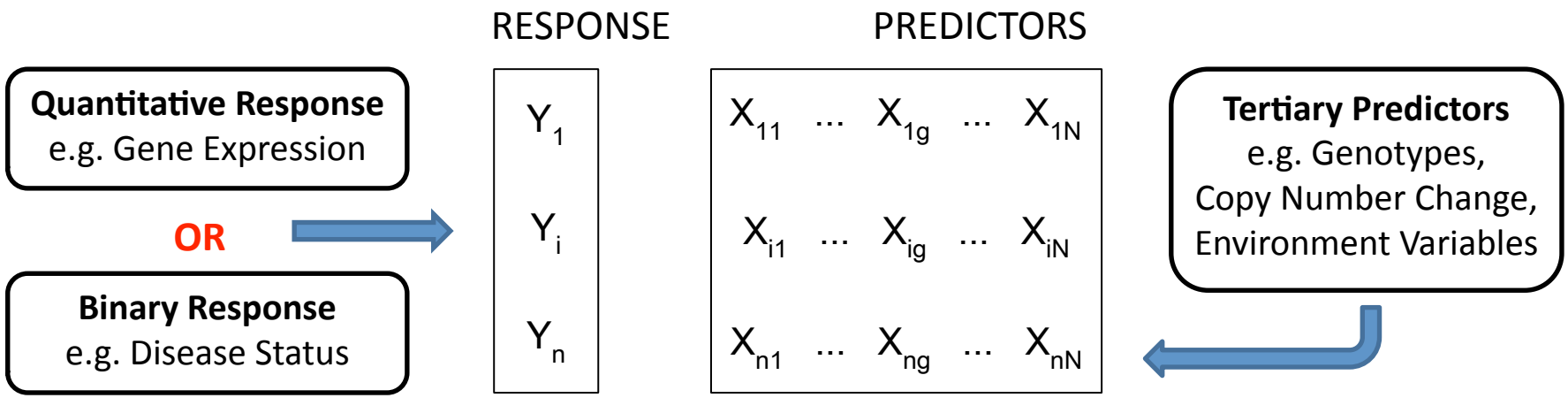


Split reads by the flowcell of origin. Contrast read counts between flowcells withing the same sample or between different samples



Perform HMM segmentation on four contrasts

NONLINEAR REGRESSION (Doug Speed)



Regression Model: $g(E[Y]) = f(\mathbf{X})$ (using an appropriate link function g).

Aim: Design regression methodology to identify which predictors contribute to f(**X**) and how they contribute.

Sparsity Assumption: Very few predictors are causal for a given response.

Interactions: Might interactions between predictors affect the response?
e.g. Gene Regulatory Networks, Oncogenes.

EXISTING METHODS

One-at-a-time methods test each predictor for association separately

Consider models of the form: $f(\mathbf{X}) = \alpha + \beta X_g$

Pros: Extremely fast and very simple to understand.

Cons: Unlikely to be a realistic model and detection based only on marginal effects.

Possible alternatives suitable for high-dimensional analysis:

- Multiple regression with a penalization term
- Pair-wise maximum likelihood tests
- Classification & Regression Trees (CART) or Random Forests
- Logic Regression
- Multivariate Adaptive Regression Splines (MARS)

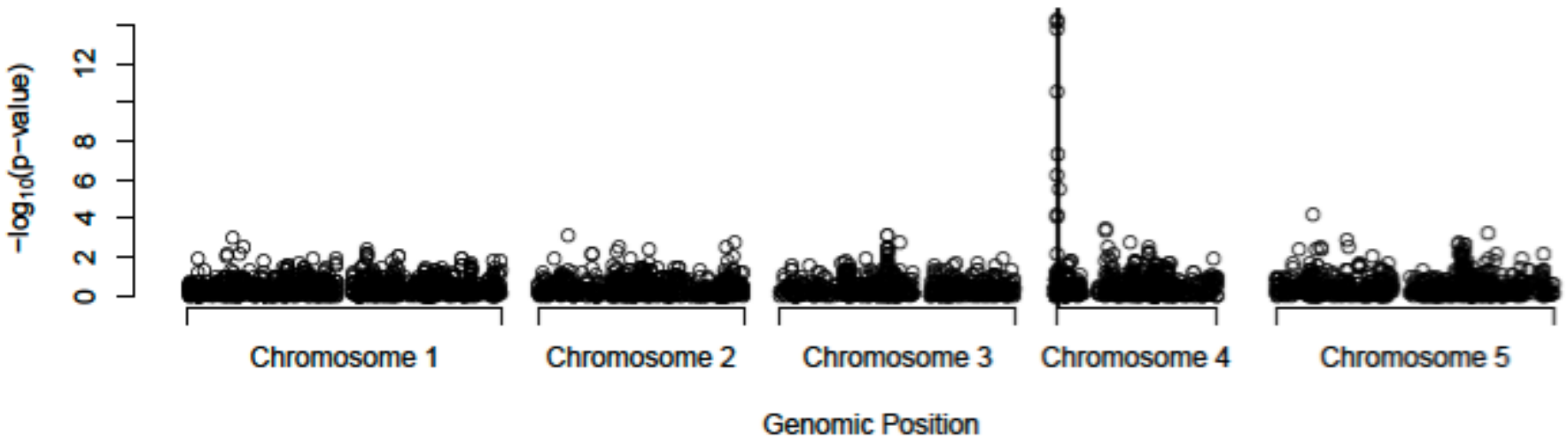
EXAMPLE ONE - *ARABIDOPSIS*

Sample: 95 accessions.

Response: Expression level of FRIGIDA Gene.

Predictors: 5419 SNPs – approximate spacing 25 kbp.

Arabidopsis thaliana

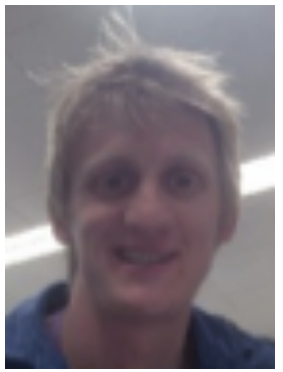


EXAMPLE TWO - HUMAN

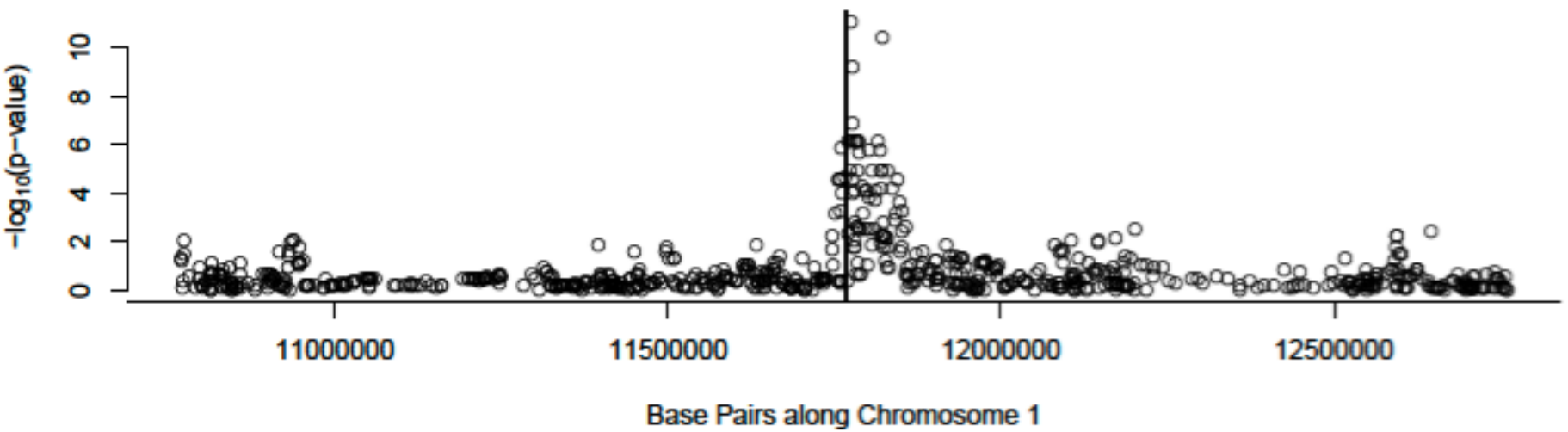
Sample: 109 CEPH HapMap individuals.

Response: Expression level of MTHFR gene.

Predictors: 763 SNPs within 1Mbp of gene locus.



Human

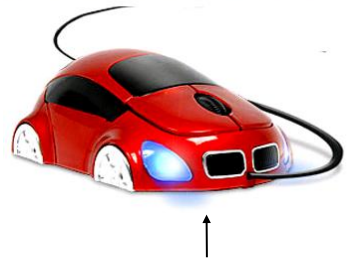


EXAMPLE THREE - MOUSE

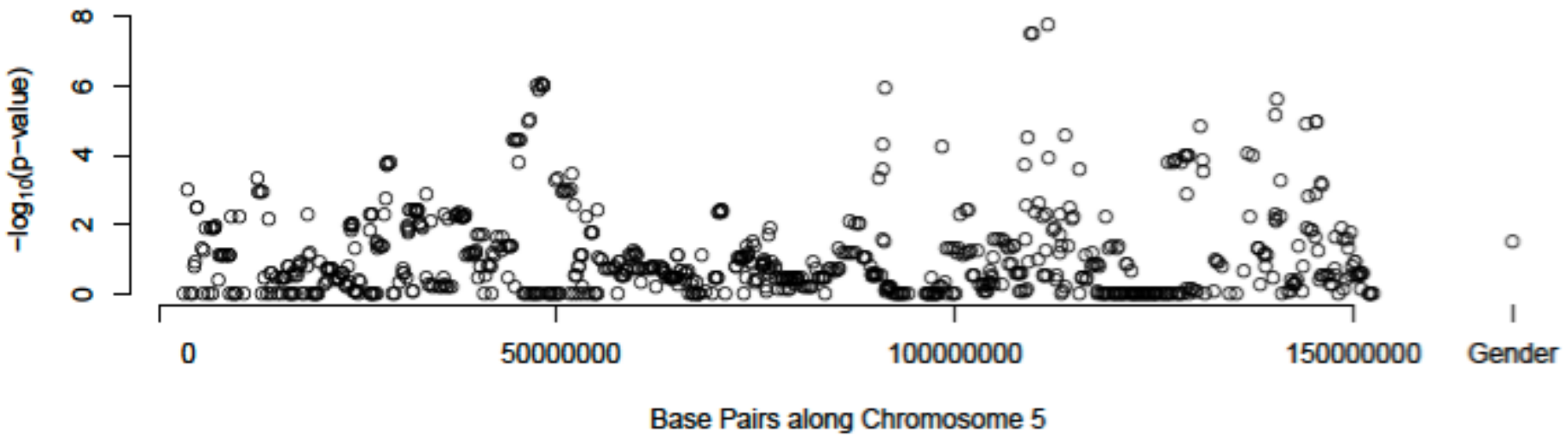
Sample: 1274 “Heterogeneous Stock” Mice.

Response: CD4 Count.

Predictors: Sex + 770 SNPs along Chr 5 – approximate spacing 200 kbp.



Mice



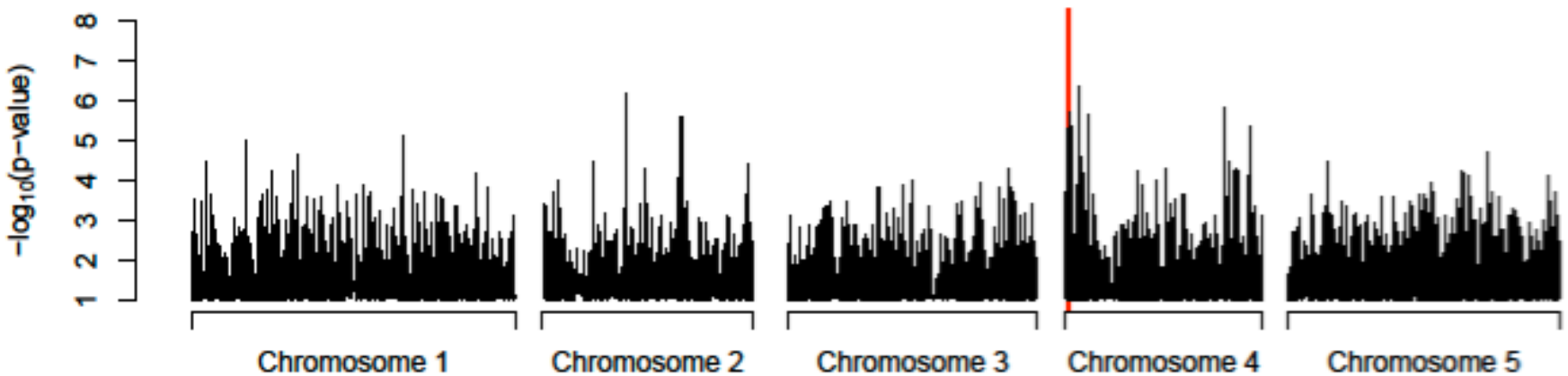
EXAMPLE FOUR - *ARABIDOPSIS*

Sample: 166 accessions

Response: Expression level of FLC Gene

Predictors: 216,130 SNPs – approximate spacing 580bp

More *Arabidopsis*

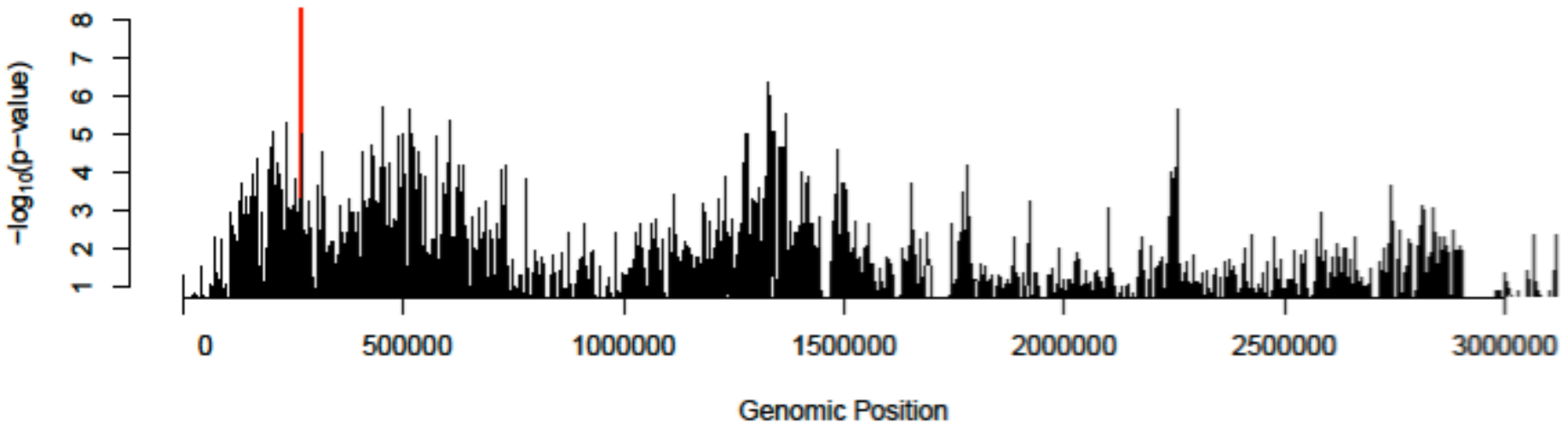


EXAMPLE FOUR - *ARABIDOPSIS*

Close-up of Chromosome 4

A true causal loci “known” to exist near FRIGIDA region (250kbp).

But one-at-a-time tests find at least two larger peaks (500kbp, 1.3Mbp).



SPARSE PARTITIONING

Bayesian nonlinear regression method allowing for interactions between predictors.
Suitable for problems where all predictors tertiary (take no more than three distinct values).

Considers how f(**X**) partitions predictor set:

Predictor set: $\{ 1 , 2 , 3 , 4 , 5 , \ldots , N \}$

Underlying relationship: e.g. $f(\mathbf{X}) = X_1 \times X_3 + X_4$

Partitions predictors as: $\{ 1 , 3 \}$ $\{ 4 \}$ $\{ 2 , 5 , \ldots , N \}$

Group 1, G_1 Group 2, G_2 Null Group, G_0

Sparse Partitioning explores space of partitions.

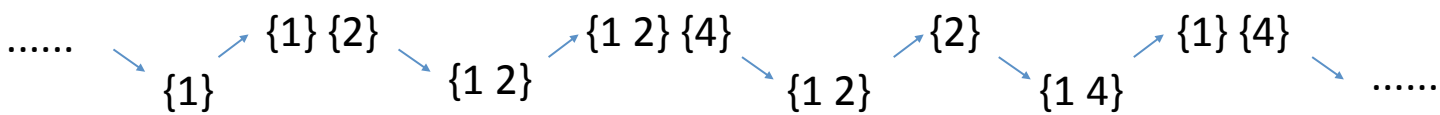
In reverse: Partition $G = \{G_0, G_1, G_2, \ldots, G_K\}$ => $f(\mathbf{X}) = f_1(X_{G1}) + f_2(X_{G2}) + \ldots + f_K(X_{GK})$

BAYESIAN METHODOLOGY

Priors:
Prior on the partitions reflects belief that each predictor is associated.
Also assign a prior on the functions of groups which prefers smoother functions.

Likelihood:
Scores each partition by integrating over all possible functions.

Posterior:
Approximated via Markov Chain Monte Carlo using a stepwise search of partitions



Sparse Partitioning estimates:

probability a predictor not in null group => predictor associated

probability a predictors in same non-null group => predictors interact

ADVANTAGES OF SPARSE PARTITIONING

Most methods only allow certain forms for f (**X**)

e.g. insist it is linear, or only consider multiplicative interactions.

Sparse Partitioning fits full degrees of freedom model for f (**X**)

e.g. suppose $G_1 = \{X_1, X_2\}$ => $f_1(X_1, X_2) = \beta_{X1,X2}$

This approach inevitably overfits the true model at times

but penalty for overfitting seems less than penalty for underfitting.

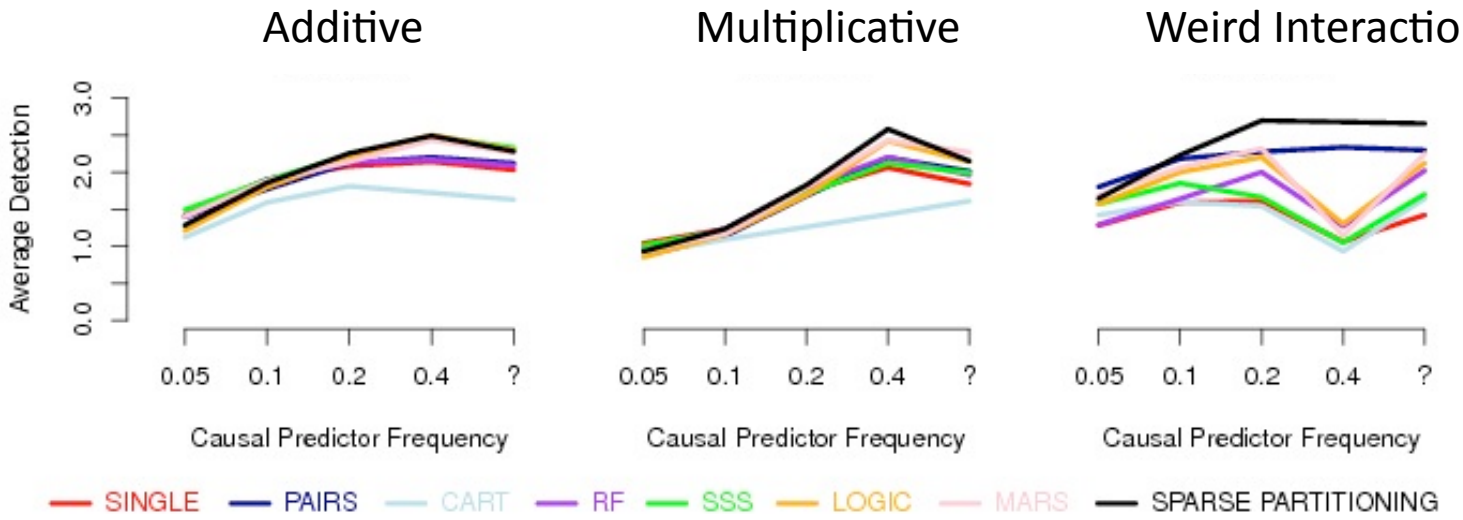
Sparse Partitioning only concerned with which predictors contributes to f (**X**)

Much easier to search space of partitions rather than space of possible f (**X**)

Exact form of f (**X**) can be investigated in a follow-up experiment.

SIMULATION STUDY RESULTS

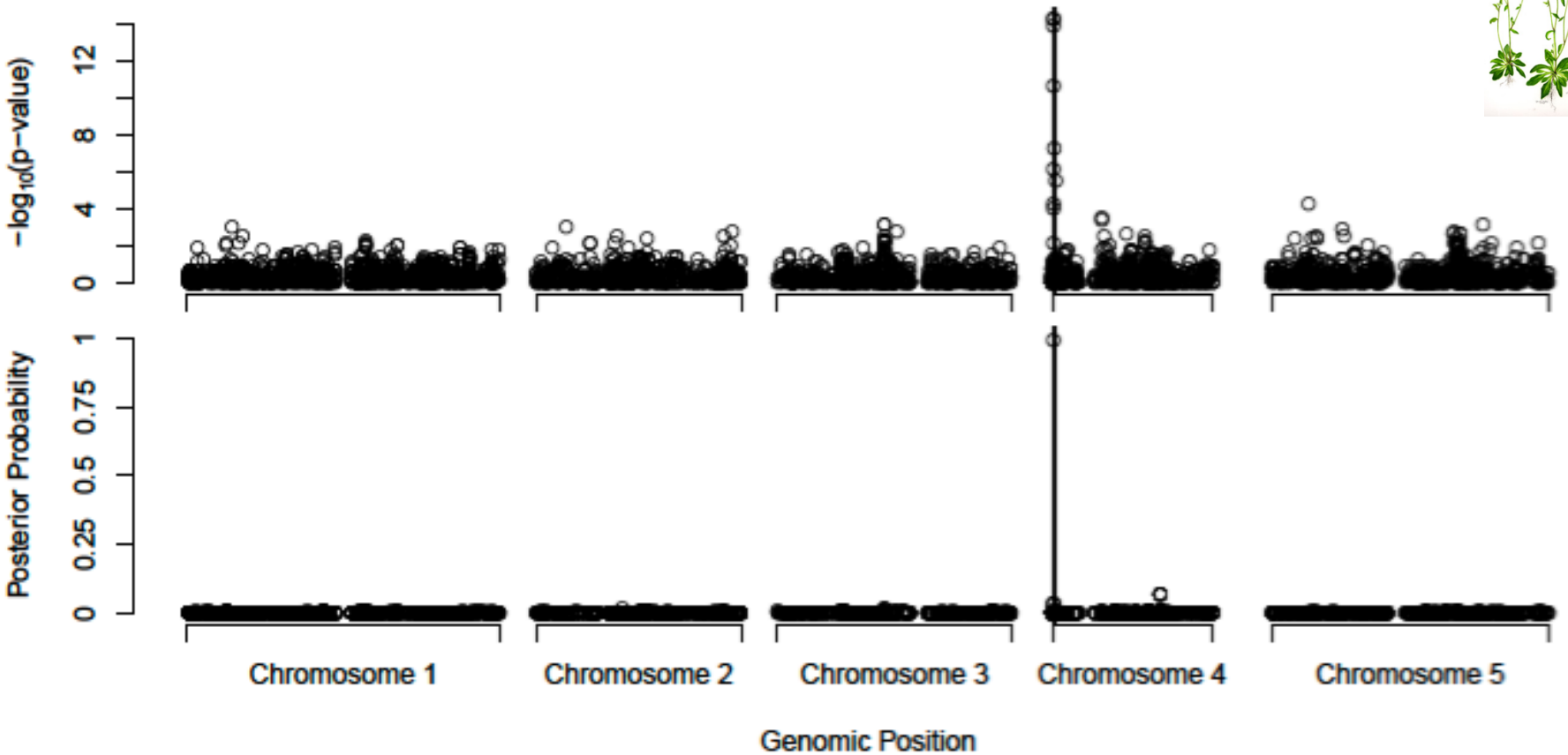
Compare methods by their power to detect causal predictors in simulated data.
(Binary predictors, continuous response)



Sparse Partitioning appears robust to different regression models.

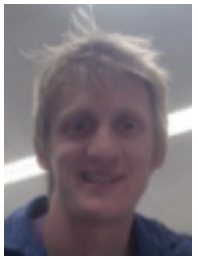
It is able to consider all possible forms for f (**X**)
so maintains power in scenarios where other methods fail.

EXAMPLE ONE - *ARABIDOPSIS*



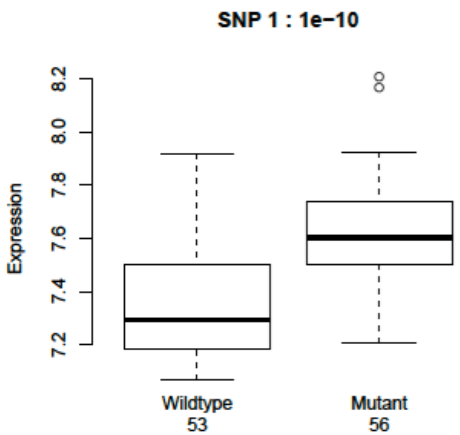
Easy example, one strong association.

EXAMPLE TWO - HUMAN

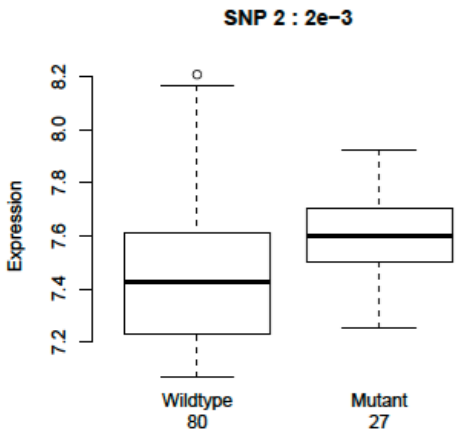


(Two predictors are almost identical, so given equal posterior scores)
Sparse Partitioning finds two associations and evidence for interaction between them.

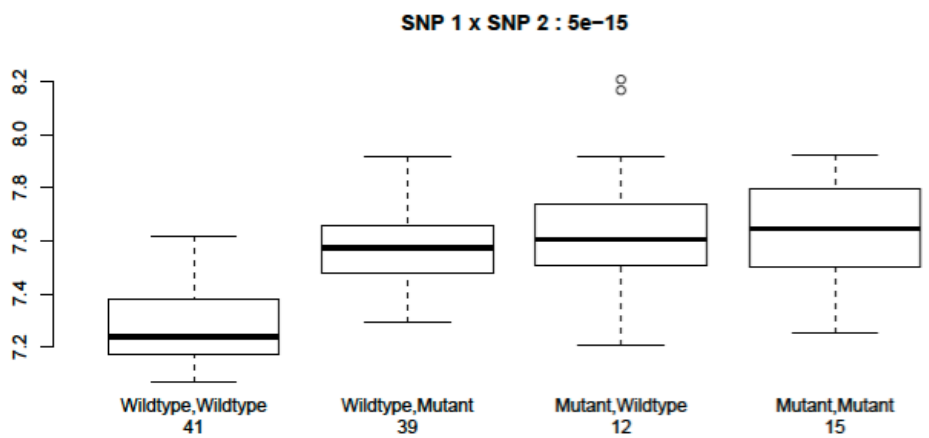
EXAMPLE TWO - HUMAN



G = { SNP 1 }



G = { SNP 2 }

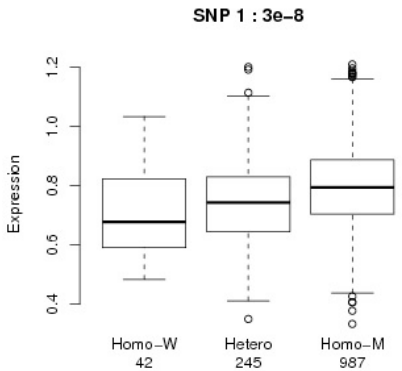


G = { SNP 1 , SNP 2 }

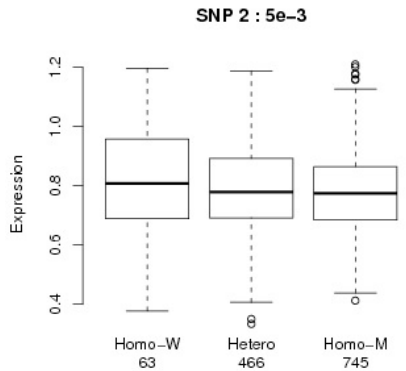
EXAMPLE THREE - MOUSE

Sparse Partitioning finds strongest association from one-at-a-time test, but also a SNP with no marginal effect. Finds evidence for interaction between both SNPs and gender.

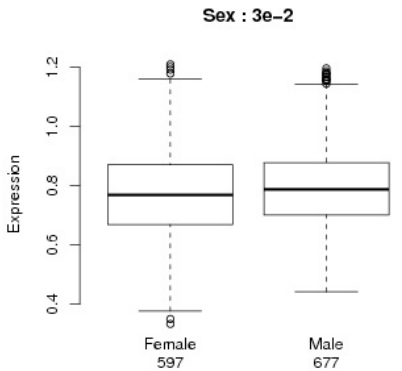
EXAMPLE THREE - MOUSE



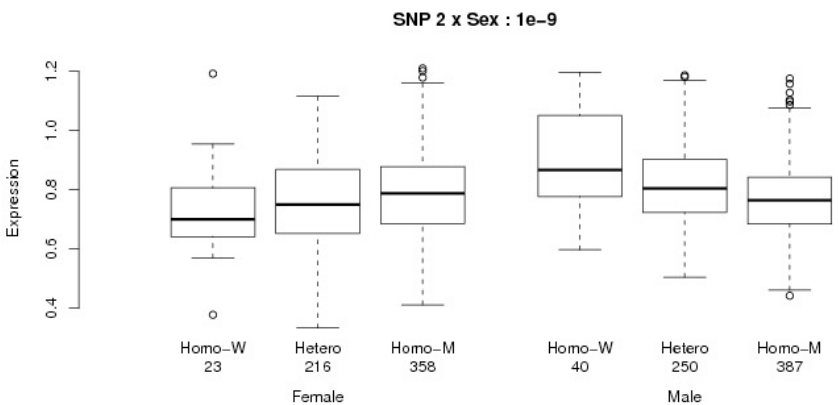
G = { SNP 1 }



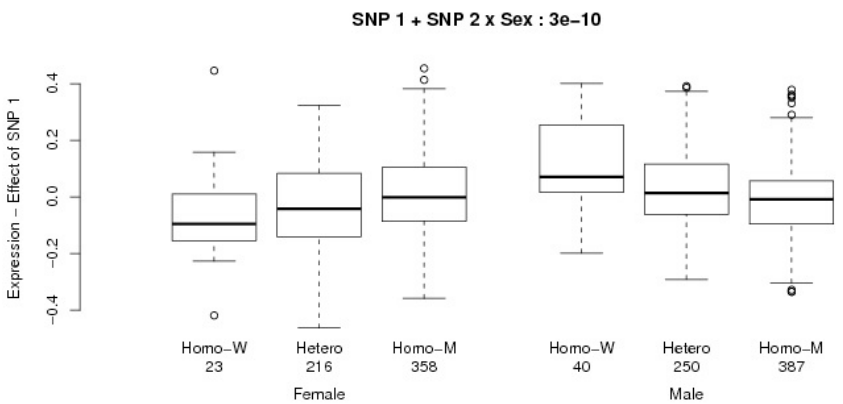
G = { SNP 2 }



G = { SEX }



G = { SNP 2 , SEX }

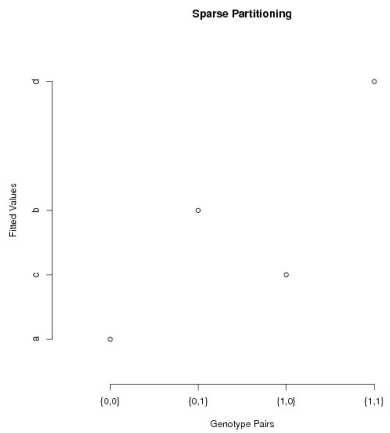


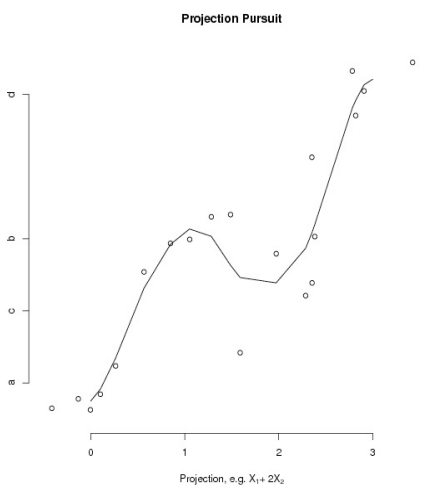
G = { SNP 1 } { SNP 2 , SEX }

LIMITATIONS

Requires tertiary predictors

But can handle quantitative predictors by introducing splines.





Relies on convergence of MCMC sampling which limits size of dataset it can handle

e.g. analysis of 5,000 predictors will take ~1 hour (but can be parallelized).

Currently limited to c.15,000 predictors, so not applicable on a genome-wide scale.

Could filter dataset first or…

DETERMINISTIC VERSION

Instead of exploring partition space, deterministic version explores set of predictors associated
Finds marginal score for sets indicating which predictors are associated:

Predictor set { 1 , 2 , 4 }

Possible Partitions:

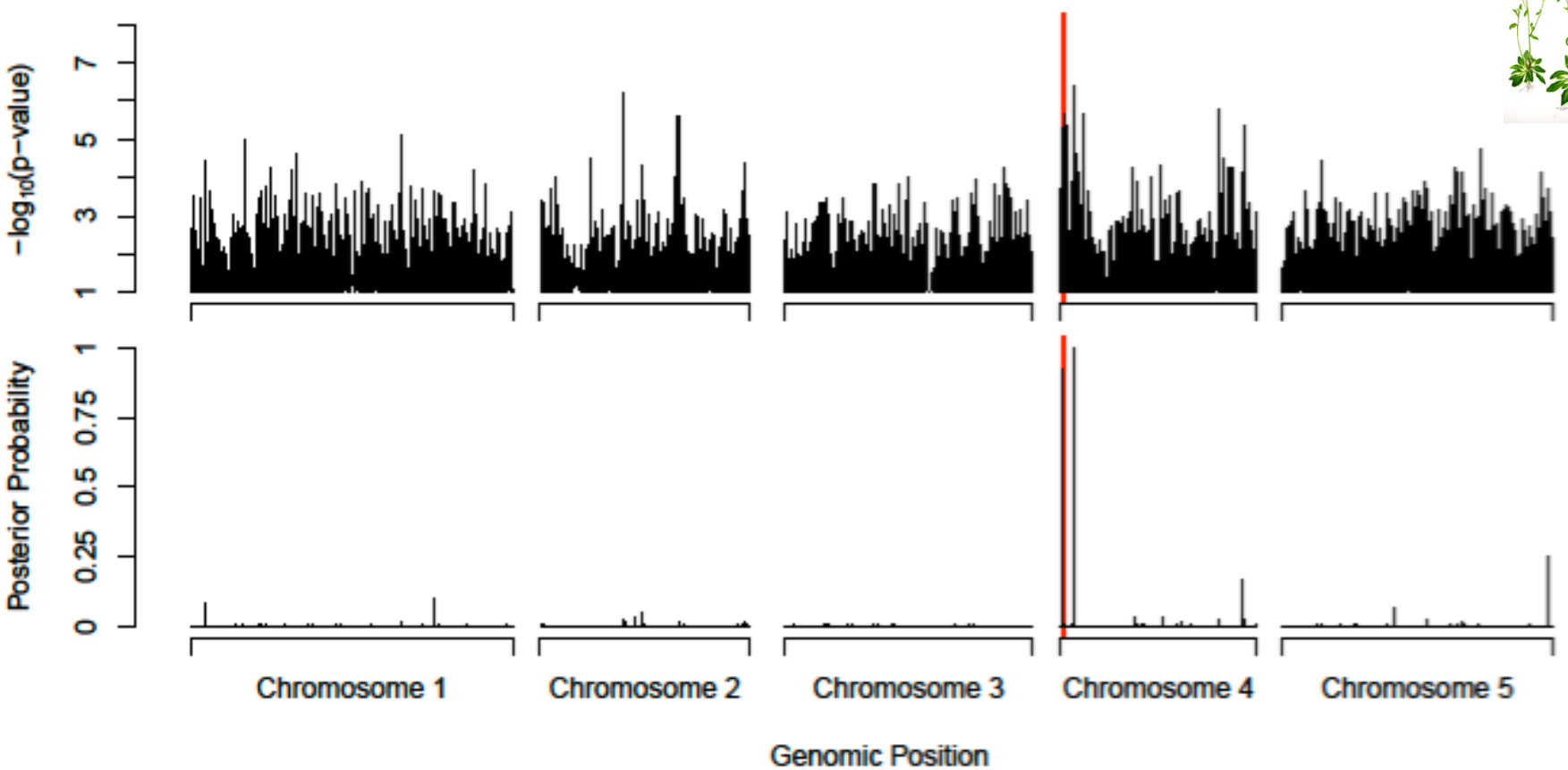
{ 1 } { 2 } { 4 }
{ 1 } { 2 , 4 } { 2 } { 1 , 4 } { 4 } { 1 , 2 }
{ 1 , 2 , 4 }

Score for predictor set is sum of posterior scores for 5 possible partitions.

Stops moving once no improvement possible.

Using this set as “null model” finds posterior probabilities of including other predictors.

EXAMPLE FOUR - *ARABIDOPSIS*

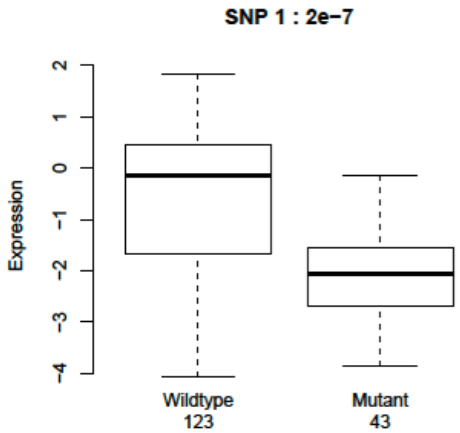


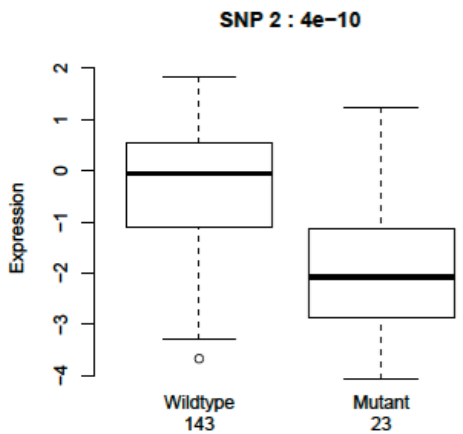
Deterministic version able to find sensible results on whole-genome data.

EXAMPLE FOUR - *ARABIDOPSIS*

Finds evidence for an association 30kbp upstream of locus, which seems far more plausible.

EXAMPLE FOUR - *ARABIDOPSIS*

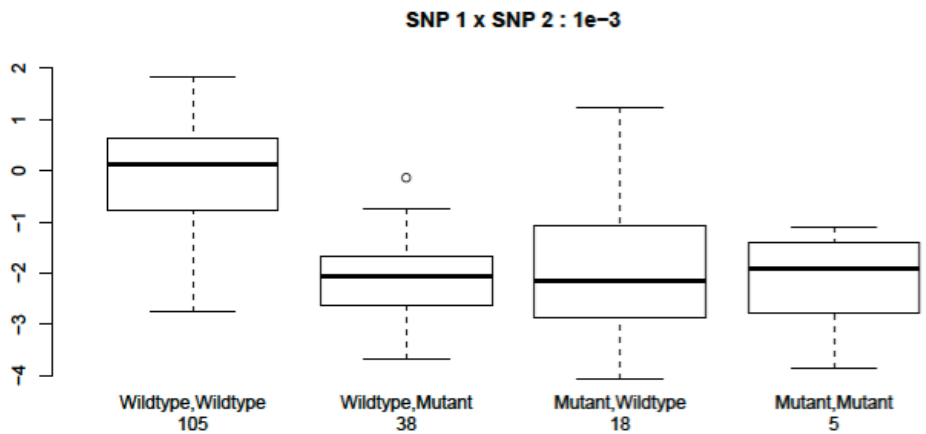




Highest scoring set of associations:
{ SNP 1 , SNP 2 }

Two possible partitions:

{ SNP 1 } { SNP 2 } 0.01
{ SNP 1 , SNP 2 } 0.99



PROS / CONS OF DETERMINISTIC VERSION

Run time takes a few minutes for 200,000+ predictors.
Slightly less powerful, as only allowed to move to higher scoring models
e.g. will struggle to find pair of interacting predictors if weaker marginal effects.

Also able to incorporate prior knowledge concerning associations.

If desired can set number of groups / maximum size of groups
e.g. maximum group size one => linear model (method becomes forward regression).

If we know a predictor is causal we can set its prior to 1:
then the method will allow for its additive effect (as in forward regression).
BUT ALSO consider the way other predictors might interact with it.

Both versions available at:
http://www.compbio.group.cam.ac.uk/software.html

Acknowledgements

METABRIC

Cambridge Breast Cancer Functional Genomics
Carlos Caldas
Suet-Feung Chin
Oscar Rueda
Stefan Gräf
Zhihao Ding
Elena Provenzano

Cambridge Ovarian Cancer Functional Genomics
James Brenton

Cambridge CompBio
Florian Markowetz
Yinyin Yuan

CRI Bioinformatics Core
Mark Dunning
Roslin Russell
Kevin Howe

Cambridge CompBio Stats
Christina Curtis
Andy Lynch
Shamith Samarajiwa
Doug Speed

Vancouver
Samuel Aparicio
Sohrab Shah
Gulisa Turashvili

Oslo
Anne-Lise Børresen-Dale
Anita Langerød

Nottingham
Ian Ellis
Sarah Watts
Andrew Green

CRI Genomics
James Hadfield
Michele Osborne

CRI Bioinformatics
Rory Stark

USC CEGS

Magnus Nordborg
Sergey Nuzhdin
Simon Knott
Reza Ardekani

+ 40 MCB faculty, postdocs, students

Comp Bio & Statistics

Mike Smith
Benilton Carvalho
Jonathan Cairns
Nuno Barbosa Morais
Sergii Ivakhno
Julie Woolford

BBS & tumor evolution

Cambridge
Irene Tiemann-Boege
Christina Curtis
Daniel Goodman
Inma Spiteri
Andrea Sottoriva

USC
Darryl Shibata
Kim Siegmund
Paul Marjoram

CENTER OF EXCELLENCE IN GENOMIC SCIENCE

NATIONAL INSTITUTES OF HEALTH

CANCER RESEARCH UK

MRC Medical Research Council

THE ROYAL SOCIETY

bbsrc biotechnology and biological sciences research council

EPSRC Engineering and Physical Sciences Research Council