(sfi)^2 Statistics for Innovation

Statistical variable screening in high dimensions: the example of genomics

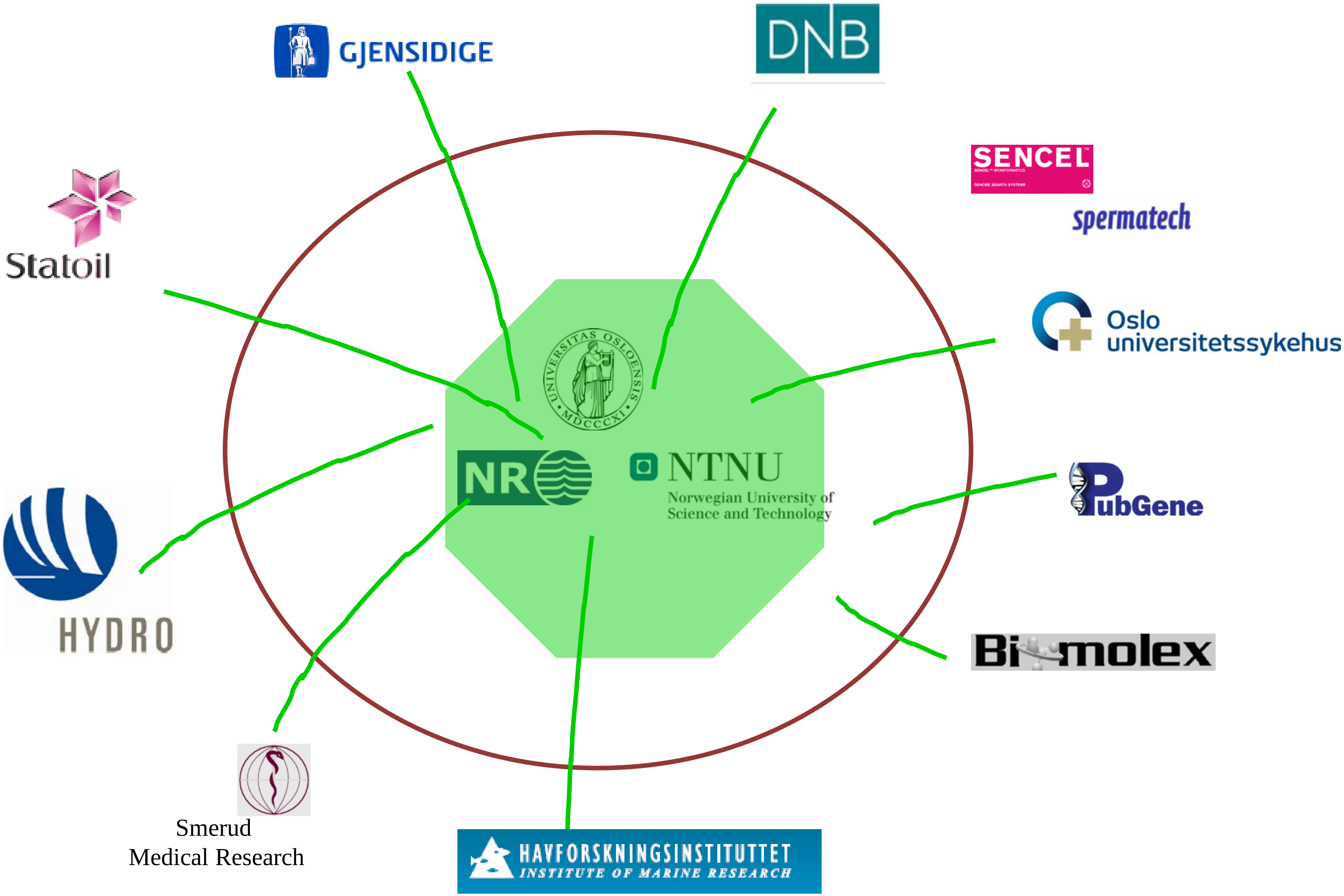
Arnoldo Frigessi,
University of Oslo
frigessi@medisin.uio.no

(sfi)^2 Statistics for Innovation

Mission:

(sfi)^2 develops core statistical methodologies, strategically necessary to achieve innovation goals in four key sectors:

- petroleum
- marine resources
- health
- finance and insurance



Smerud
Medical Research

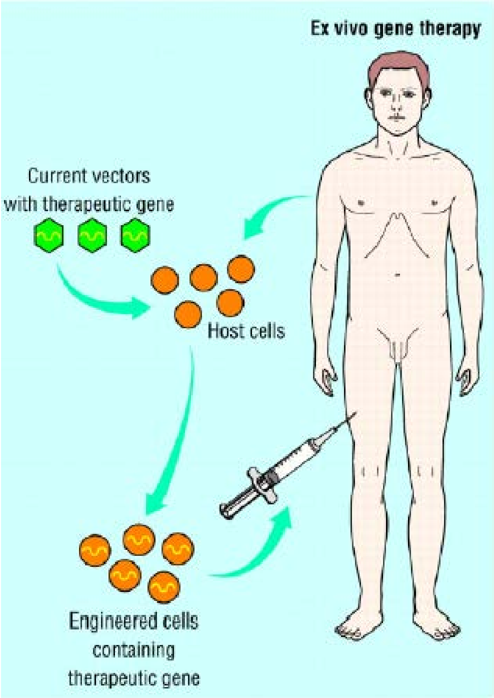
BioInfStat -	Statistics for bioinformatics
ClimateInsure -	Climate Change and Insurance Industry
ComplexClin -	Statistics for complex design of clinical studies
ComplexDepend -	Statistics for complex stochastic dependence
CustomerLife -	Statistics for modelling customer life in an insurance company
Elprice -	Electricity price sensitivity
FindOil -	Statistics for oil and gas exploration
Genestat -	Statistics for genomic research
Infect -	Modelling spread of infectious diseases in fish farming
StatMarine -	Statistics for management of Norwegian marine resources
TotalRisk -	Statistics for modelling the risk of financial institutions

PLAN

- Genomic Hyperbrowser
- p>>n regression → LASSO

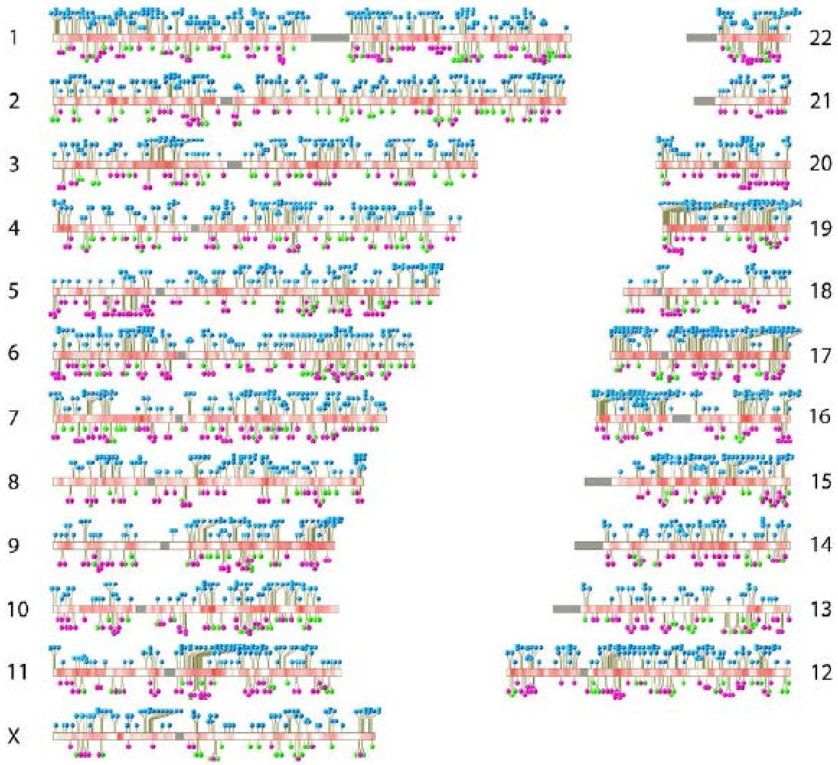
Gene therapy

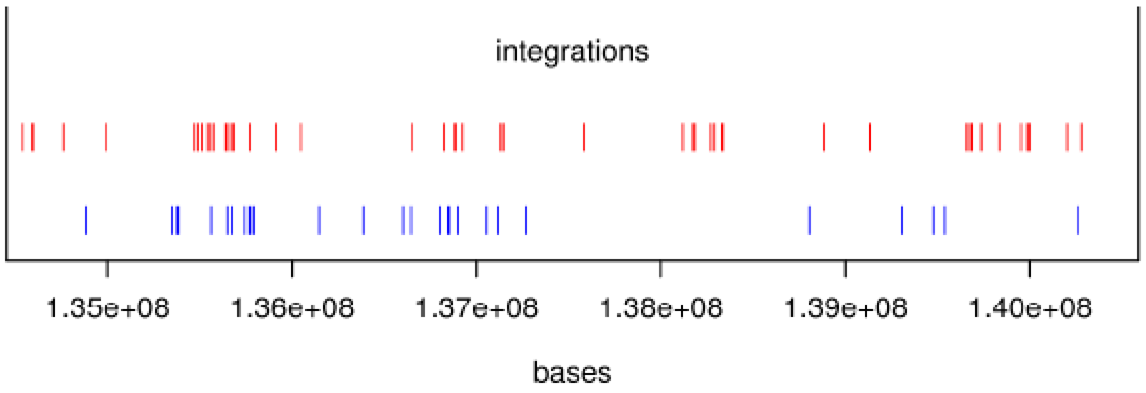
- Treats genetic diseases by replacing the defective gene with a functional one
- The working gene is introduced via a virus – the vector, which integrates into the DNA
- **Where in the DNA?**
- Different viruses prefer different integration regions.
- In trials, gene therapy caused too often leukemia.
- Integration behaviour is a major safety issue in gene therapy.


Ex vivo gene therapy
Current vectors with therapeutic gene
Host cells
Engineered cells containing therapeutic gene

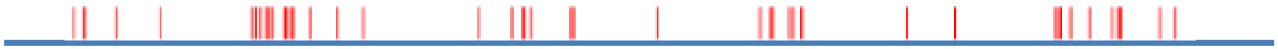
Moloney murine leukaemia virus (MLV)
Human immunodeficiency virus (HIV)

- Do HIV and MLV have different preferred areas of integrations?

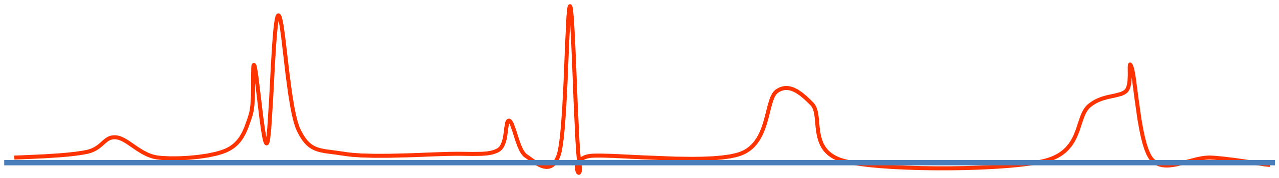




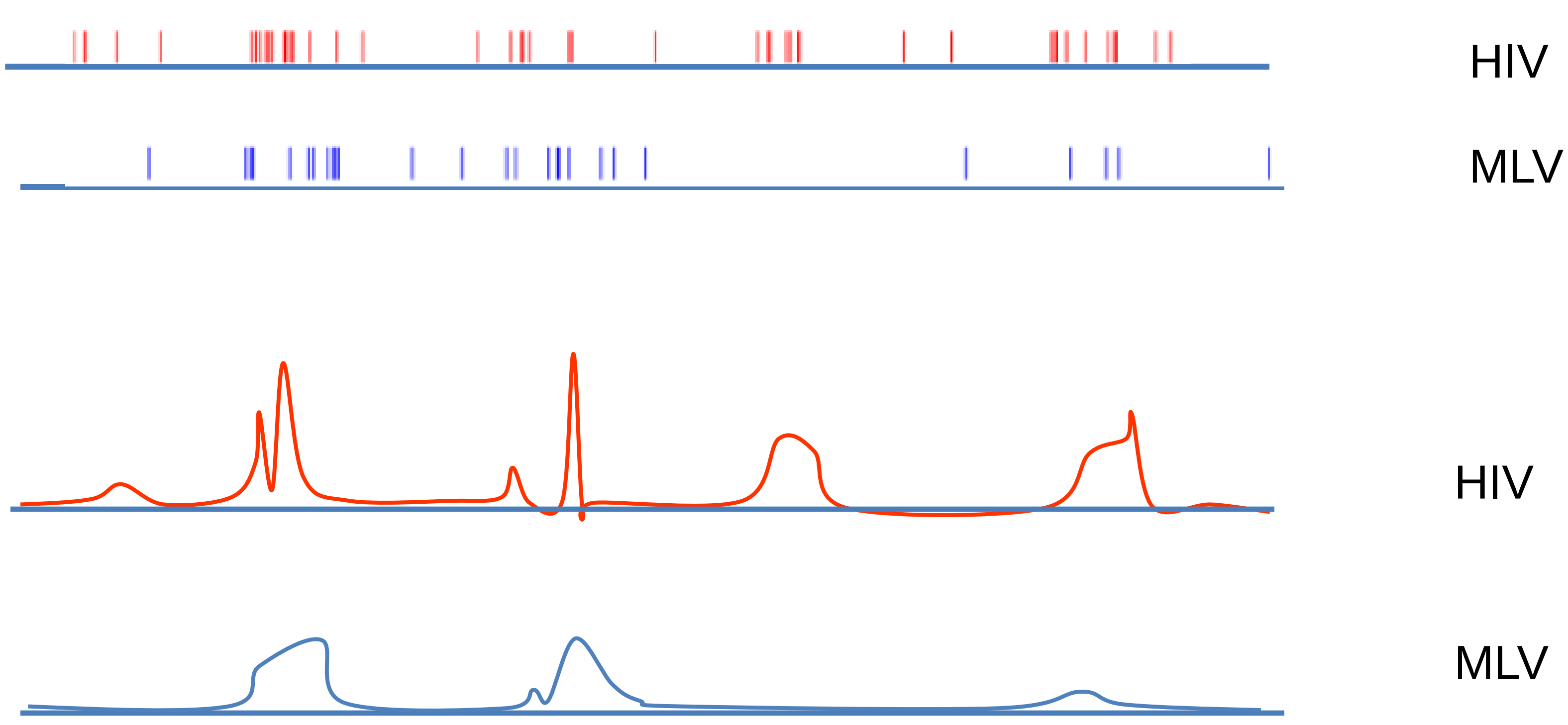
integration sites



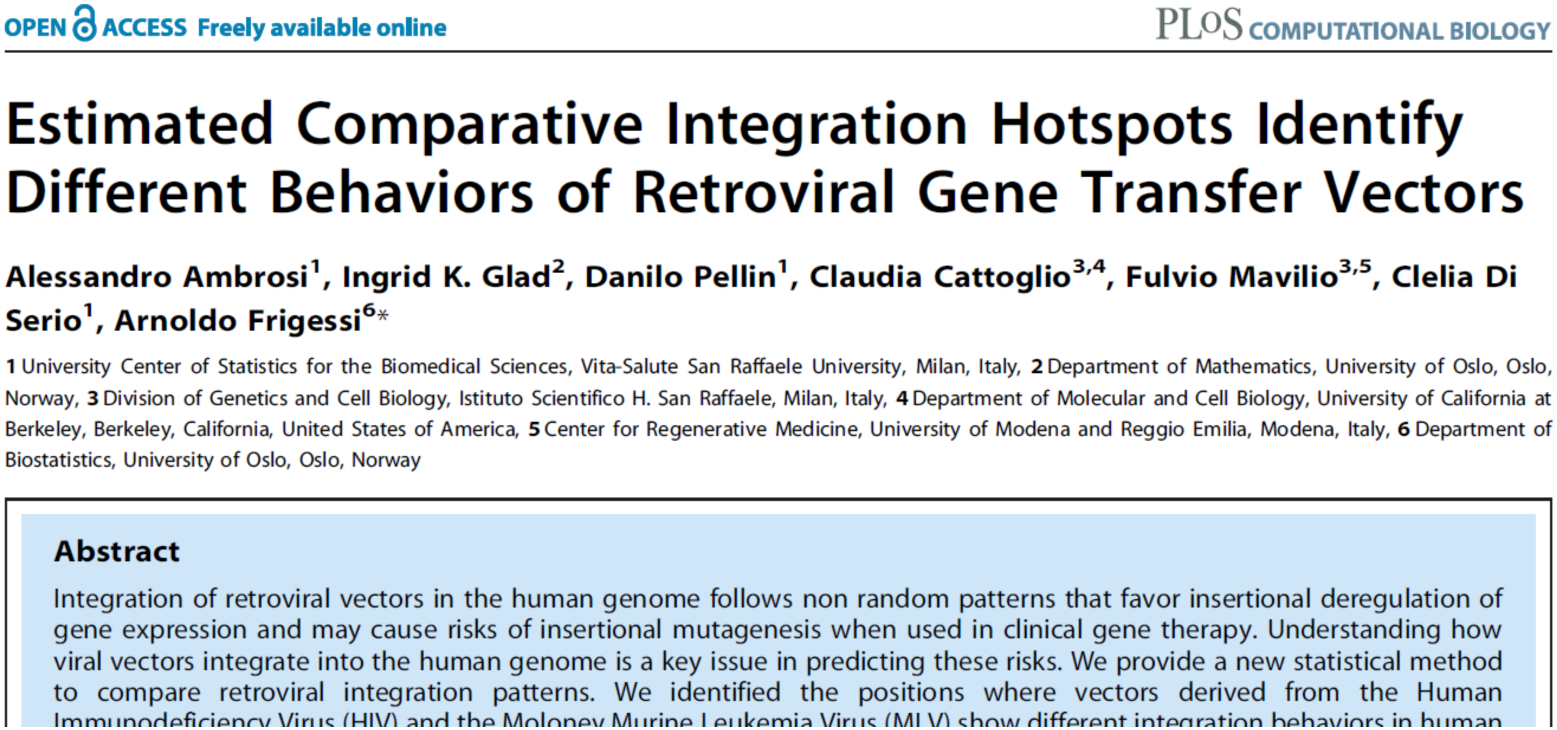
points on a line



Integration density



- Comparing two densities, estimated non-parametrically
- Different sample sizes: uncertainty is different!
- Where are the densities **significantly** different?



OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Estimated Comparative Integration Hotspots Identify Different Behaviors of Retroviral Gene Transfer Vectors

Alessandro Ambrosi[1], Ingrid K. Glad[2], Danilo Pellin[1], Claudia Cattoglio[3,4], Fulvio Mavilio[3,5], Clelia Di Serio[1], Arnoldo Frigessi[6]*

1 University Center of Statistics for the Biomedical Sciences, Vita-Salute San Raffaele University, Milan, Italy, **2** Department of Mathematics, University of Oslo, Oslo, Norway, **3** Division of Genetics and Cell Biology, Istituto Scientifico H. San Raffaele, Milan, Italy, **4** Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California, United States of America, **5** Center for Regenerative Medicine, University of Modena and Reggio Emilia, Modena, Italy, **6** Department of Biostatistics, University of Oslo, Oslo, Norway

Abstract

Integration of retroviral vectors in the human genome follows non random patterns that favor insertional deregulation of gene expression and may cause risks of insertional mutagenesis when used in clinical gene therapy. Understanding how viral vectors integrate into the human genome is a key issue in predicting these risks. We provide a new statistical method to compare retroviral integration patterns. We identified the positions where vectors derived from the Human Immunodeficiency Virus (HIV) and the Moloney Murine Leukemia Virus (MLV) show different integration behaviors in human

- Confidence bands for non-parametrically estimated densities are difficult!

- $$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

- The 0.99 variability band for this density was computed starting with the Taylor expansion

$$Var\left(\sqrt{\hat{f}}\right) \sim \frac{1}{4}\frac{1}{nh}R(K) \quad \text{where} \quad R(K) = \int K^2(x)dx$$

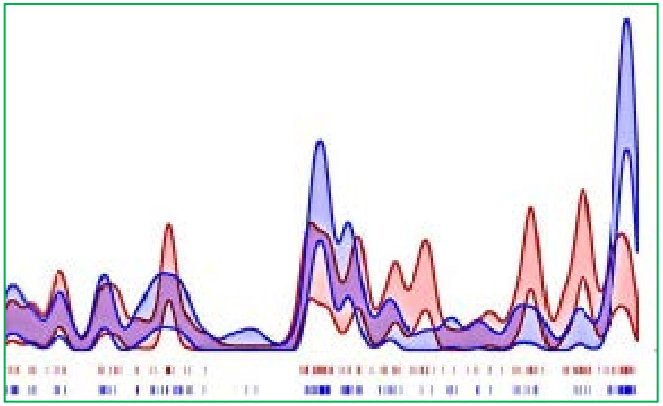
- The root transform allows to obtain an approximation of the variance which is independent from the unknown density. Therefore, on the square root scale, a level error band can be computed, using the half width

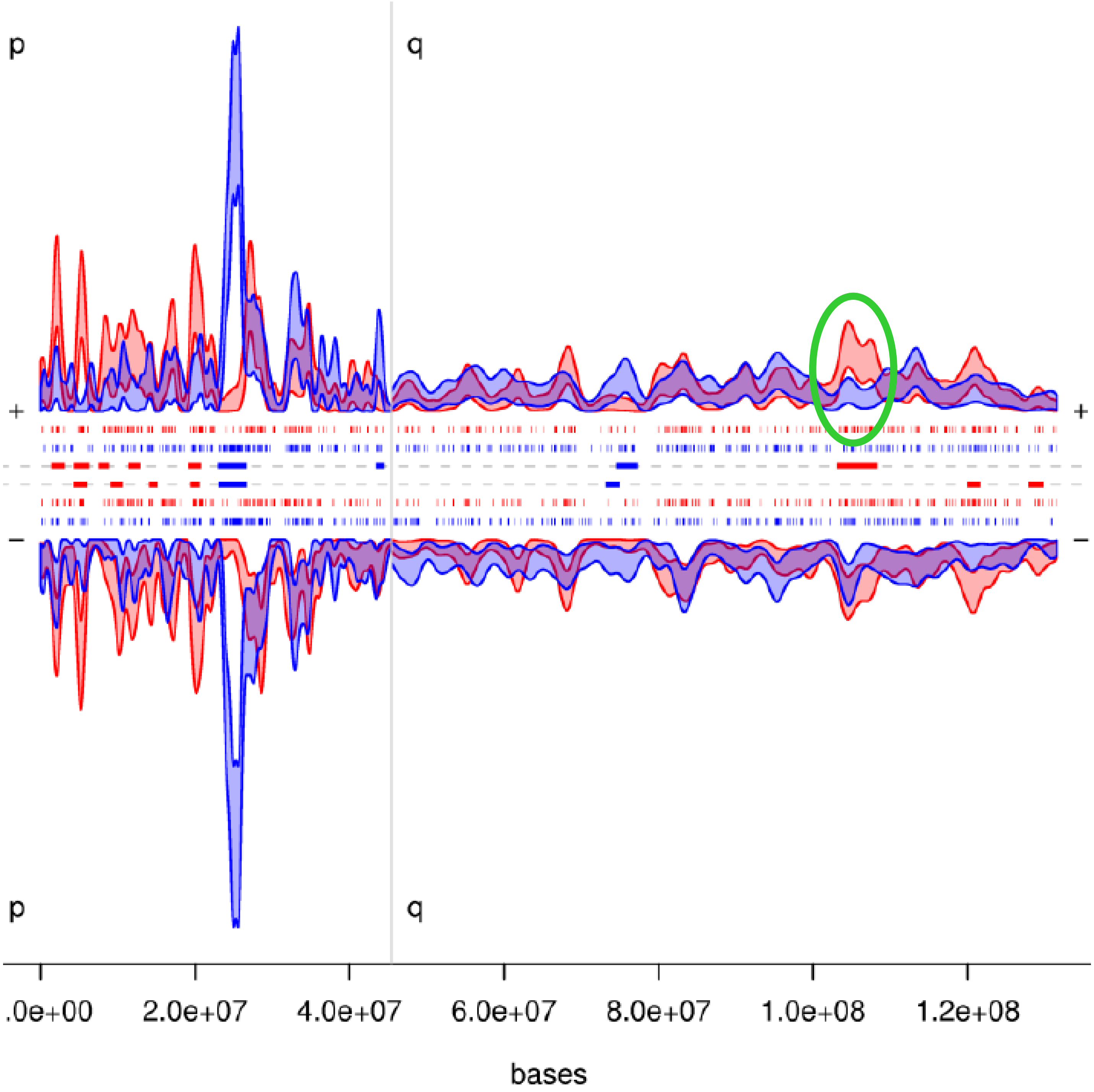
$$Z_{\alpha/2}\sqrt{\frac{R(K)}{4nh}}$$

where $Z_{\alpha/2}$ is the quantile of the normal standard distribution.

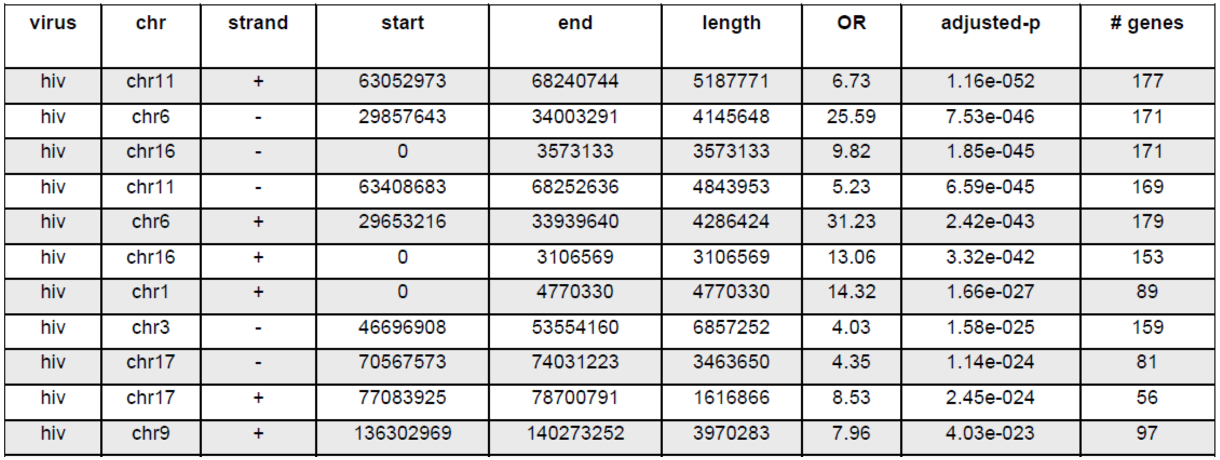
- The variability band is transformed back to the original scale as

$$\left(\sqrt{\hat{f}} \pm Z_{\alpha/2}\sqrt{\frac{R(K)}{4nh}}\right)^2$$

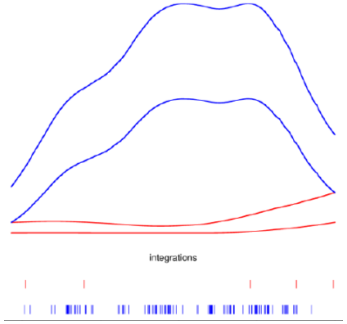



p
q
+
+
−
−
p
q
.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
bases

- 94 comparative hotspots are found significant.



virus	chr	strand	start	end	length	OR	adjusted-p	# genes
hiv	chr11	+	63052973	68240744	5187771	6.73	1.16e-052	177
hiv	chr6	-	29857643	34003291	4145648	25.59	7.53e-046	171
hiv	chr16	-	0	3573133	3573133	9.82	1.85e-045	171
hiv	chr11	-	63408683	68252636	4843953	5.23	6.59e-045	169
hiv	chr6	+	29653216	33939640	4286424	31.23	2.42e-043	179
hiv	chr16	+	0	3106569	3106569	13.06	3.32e-042	153
hiv	chr1	+	0	4770330	4770330	14.32	1.66e-027	89
hiv	chr3	-	46696908	53554160	6857252	4.03	1.58e-025	159
hiv	chr17	-	70567573	74031223	3463650	4.35	1.14e-024	81
hiv	chr17	+	77083925	78700791	1616866	8.53	2.45e-024	56
hiv	chr9	+	136302969	140273252	3970283	7.96	4.03e-023	97



- HIV hotspots contained more genes than MLV hotspots.

- HIV hotspots showed enrichment of genes involved in antigen processing.

- ….

- Important indications for drug design.

MOLECULAR BIOLOGY

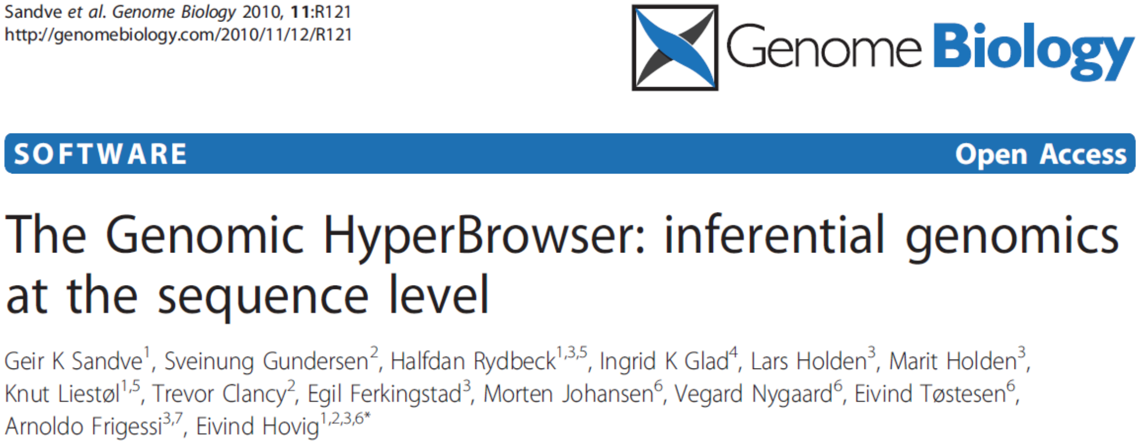
Gene therapy

Genome Biology

The Genomic HyperBrowser

The Genomic HyperBrowser: inferential genomics at the sequence level

Sandve et al



Sandve et al. Genome Biology 2010, 11:R121
http://genomebiology.com/2010/11/12/R121

Genome Biology

SOFTWARE Open Access

The Genomic HyperBrowser: inferential genomics at the sequence level

Geir K Sandve[1], Sveinung Gundersen[2], Halfdan Rydbeck[1,3,5], Ingrid K Glad[4], Lars Holden[3], Marit Holden[3], Knut Liestøl[1,5], Trevor Clancy[2], Egil Ferkingstad[3], Morten Johansen[6], Vegard Nygaard[6], Eivind Tøstesen[6], Arnoldo Frigessi[3,7], Eivind Hovig[1,2,3,6*]

http://hyperbrowser.uio.no

Comparing two tracks

1. Representation of generic genomic elements as mathematical objects on the line.

2. Hypotheses of interest are translated into mathematical relations between these objects.

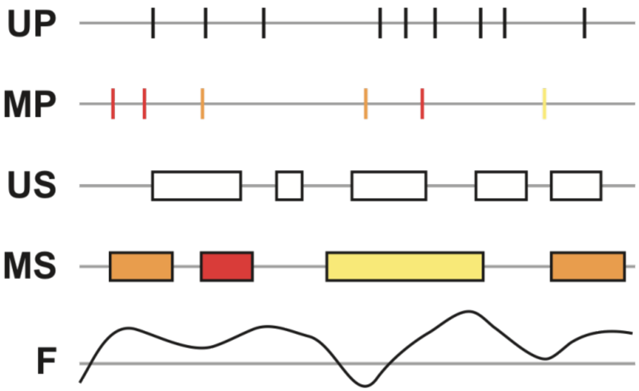
3. Concepts of randomization and track structure preservation are used to build problem-specific null models of the relation between two tracks.

4. Formal inference is performed at a global or local scale, taking confounder tracks into account when necessary.

1. Representation of genomic elements on the real line.

Five genomic types:

- unmarked points (UP),
- marked points (MP),
- unmarked segments (US),
- marked segments (MS)
- functions (F).



These five types completely represent every one-dimensional geometry with marks.

2. Catalogue of investigations

We translate biological hypotheses of interest on the relation between the two tracks, into a study of statistical relations between the geometric objects.

This leads to a large collection of generic investigations.

Example: Relation between histone modifications and gene expression

- Biology:

Does the number of nucleosomes with a given histone modification correlate with the expression of that gene?

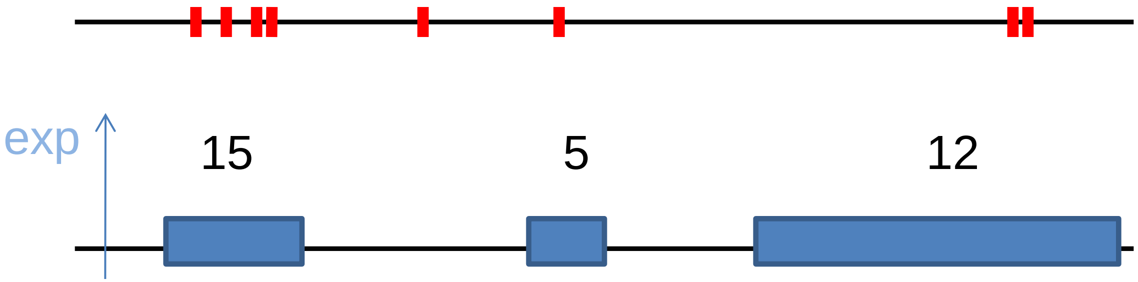
- Representation:

histone modifications: points

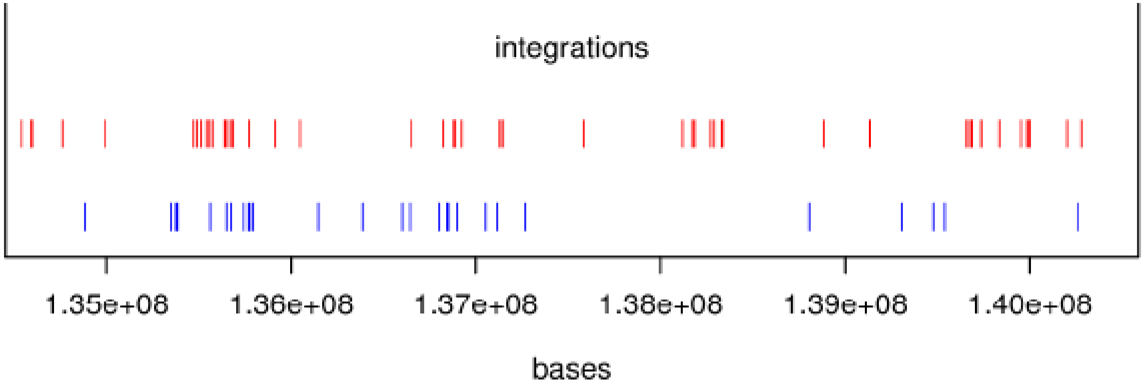
gene expressions: marked segments

- Generic investigation between a pair of tracks (T1=UP, T2=MS).

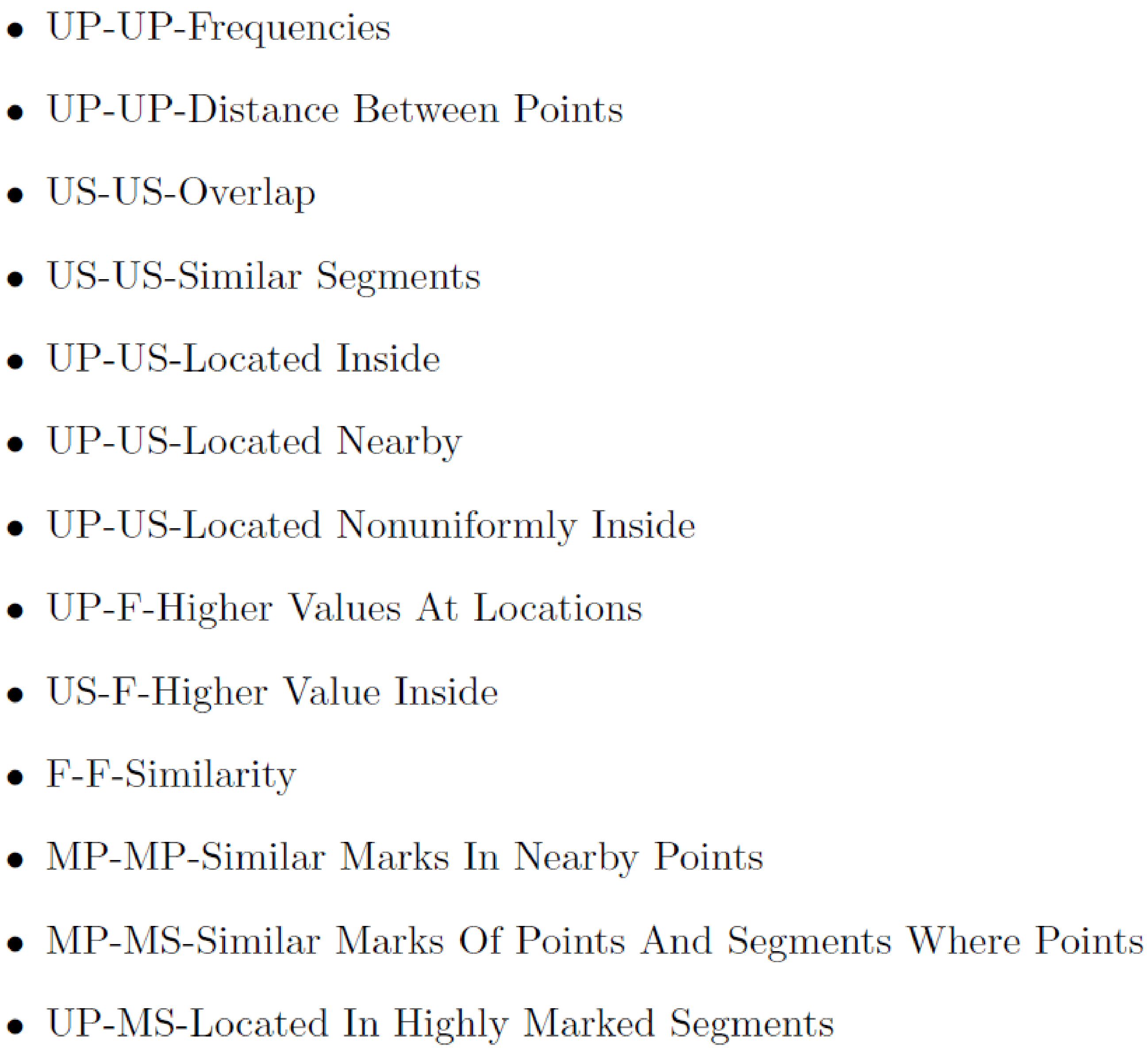
  Is the number of T1 points inside T2 segments correlated with T2 marks?



Example: Are T1=UP and T2=UP differentially located, more than expected by chance?



We have currently implemented about 20 different analyses, including:



- UP-UP-Frequencies
- UP-UP-Distance Between Points
- US-US-Overlap
- US-US-Similar Segments
- UP-US-Located Inside
- UP-US-Located Nearby
- UP-US-Located Nonuniformly Inside
- UP-F-Higher Values At Locations
- US-F-Higher Value Inside
- F-F-Similarity
- MP-MP-Similar Marks In Nearby Points
- MP-MS-Similar Marks Of Points And Segments Where Points
- UP-MS-Located In Highly Marked Segments

3. Global and local inference

- A global analysis investigates if a certain relation between two tracks is found in a domain (typically a chromosome) as a whole.

- A local analysis is based on partitioning the domain into smaller units – **bins**, and performing the analysis in each bin separately.

- Local analysis is used to investigate if and where two tracks display significant discordant behaviour, generating hypotheses on the existence of biological mechanisms explaining such perturbations.

Inference is then based on the computation of p-values, locally in each bin, or globally, under the null model.

Bins = scale

- Not too large (but not too empty either)
- Freedman and Diaconis automatic rule for histograms
- Self defined bins
- Adaptive binning (no test where there are no chances)
- Scan statistics, moving window

Example: Viral integration.

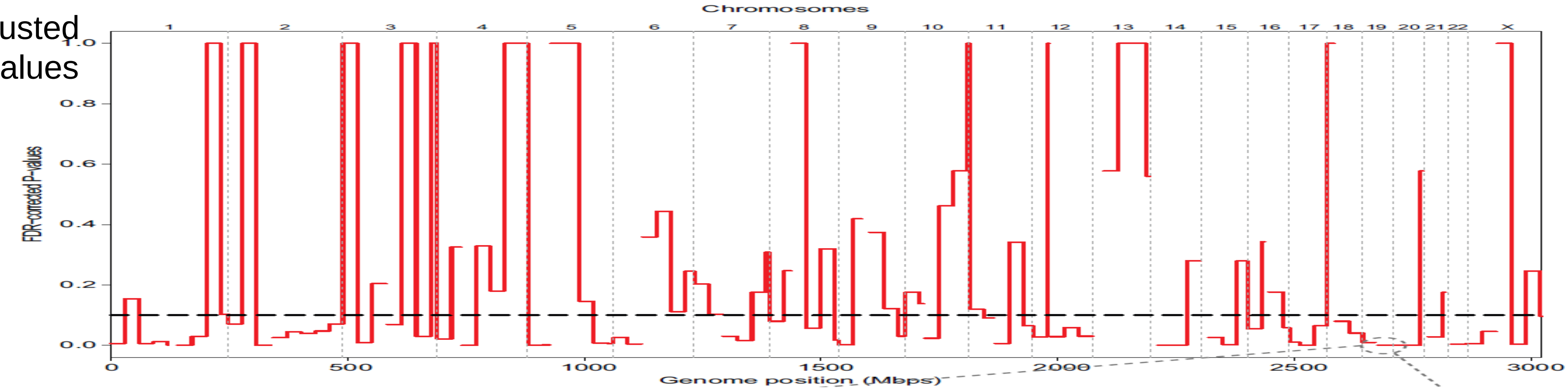
Track 1: integration sites for a specific retrovirus (UP).

Track 2: 2 kb flanking regions of predicted promoters (US)

Question: *Where in the genome*, are the points falling inside the segments more than expected by chance?

P-values in **bins** across the genome.

FDR adjusted p-values



Derse et al *J Virol 2007, 81:6731-6741.*

4. Null models

- “uniform” -- unrealistically simple null models may lead to false positives.

- The Genomic HyperBrowser allows the user to define an appropriate null model by specifying

  (a) a **preservation rule** for each track, and

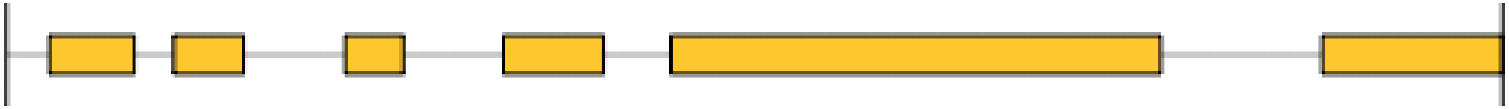
  (b) a **stochastic process**, describing how the non-preserved elements should be randomized.

- Preservation fixes (some) elements or characteristics of a track as present in the data.

- For each genomic type, we developed a hierarchy of less and less strict preservation rules, starting from preserving the entire track exactly

- (a) **preservation rule** for each track

For example, if one track is US:



(i) preserve all, as in data;
(ii) preserve segments and intervals between segments, in number and length, but not their ordering;
(iii) preserve only the segments, in number and length, but not their position;
(iv) preserve only the number of bps in segments, not segment position or number.

- (b) a rule on how the non-preserved elements should be **randomized**:

(ii) permute segments and intersegments
(iii) permute segments and give a law that says the length of the intersegments
(iv) give a law to generate segments and intersegments

- Depending on the test statistic T, the level of preservation and the chosen randomization, p-values are computed exactly, asymptotically or by standard or sequential Monte Carlo.

- Preservation leads to conditional p-values, given preservation and randomisation rules. P-values are not ordered even if the preservations rules are so. This is in analogy to tests for two-by-two contingency tables, where row or column totals can be preserved - or not -, though p-values are not decreasing.

- Choices of the null should reflect biological knowledge. Very hard. Should in principle model 3-4 billion years of the random processes that contributed to evolution.

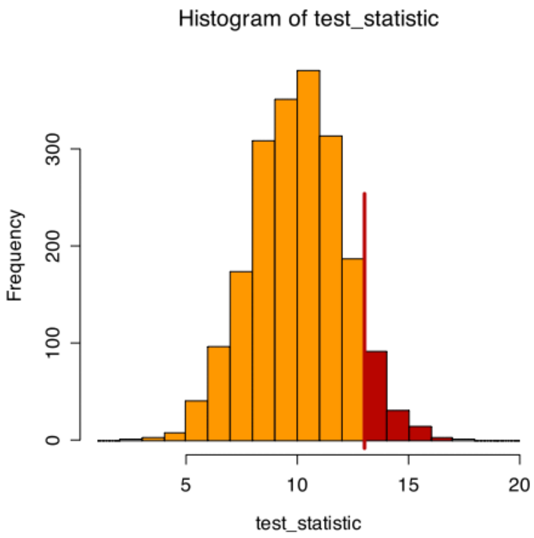
Monte Carlo Test

- When the exact calculation of the p-value is not possible, nor asymptotic arguments can be applied, standard or sequential Monte Carlo testing is performed.

  Assume T is test statistics with observed value To.

  p-value = P (T>To | Ho)

- 1. Sample new data according to Ho
  2. Compute test statistics T
  3. Repeat B (many) times
  4. Check where To falls.
  5. Estimate prob. > To.



Sequential Monte Carlo testing

(Besag J, Clifford P: Sequential Monte Carlo p-values. *Biometrika 1991*)

- Continue to sample until the sampled test statistics T is **w** times larger (or smaller, depending on side of the test) than the observed value To, or if a max number of samples N_0 has been drawn.

- p-value is then = **w**/number of samples needed

- Typically **w** = 20.

- More samples needed if p-value is small, few is p-values is large.

- Large p-values are not well estimated, but it does not matter.

- Sequential MC produces p-values that can be adjusted by FDR in the usual way

Sequential Monte Carlo multiple testing

(Sandve, Ferkingstad, Frigessi, Nygård, 2011**)**

- A= ∅

- Do all tests first: sample $H_{\mathbf{0i}}$ but stop as soon as the sampled test statistics $T_{\mathbf{i}}$ is **w** times larger than the observed value $T_{\mathbf{0i}}$: put **i** into A. For all other tests, N_0 has been reached.

- **Compute estimated p_i-values, and FDR-adjusted q_i-values for all tests.**

- **If q_i<α, put i into A.**

- For the test not in A (yet), draw new $N_{\mathbf{1}}$ samples and iterate.

- Stop when all tests are in A, or when max total number of samples achieved.

- Theorem: FDR(α) controlled.

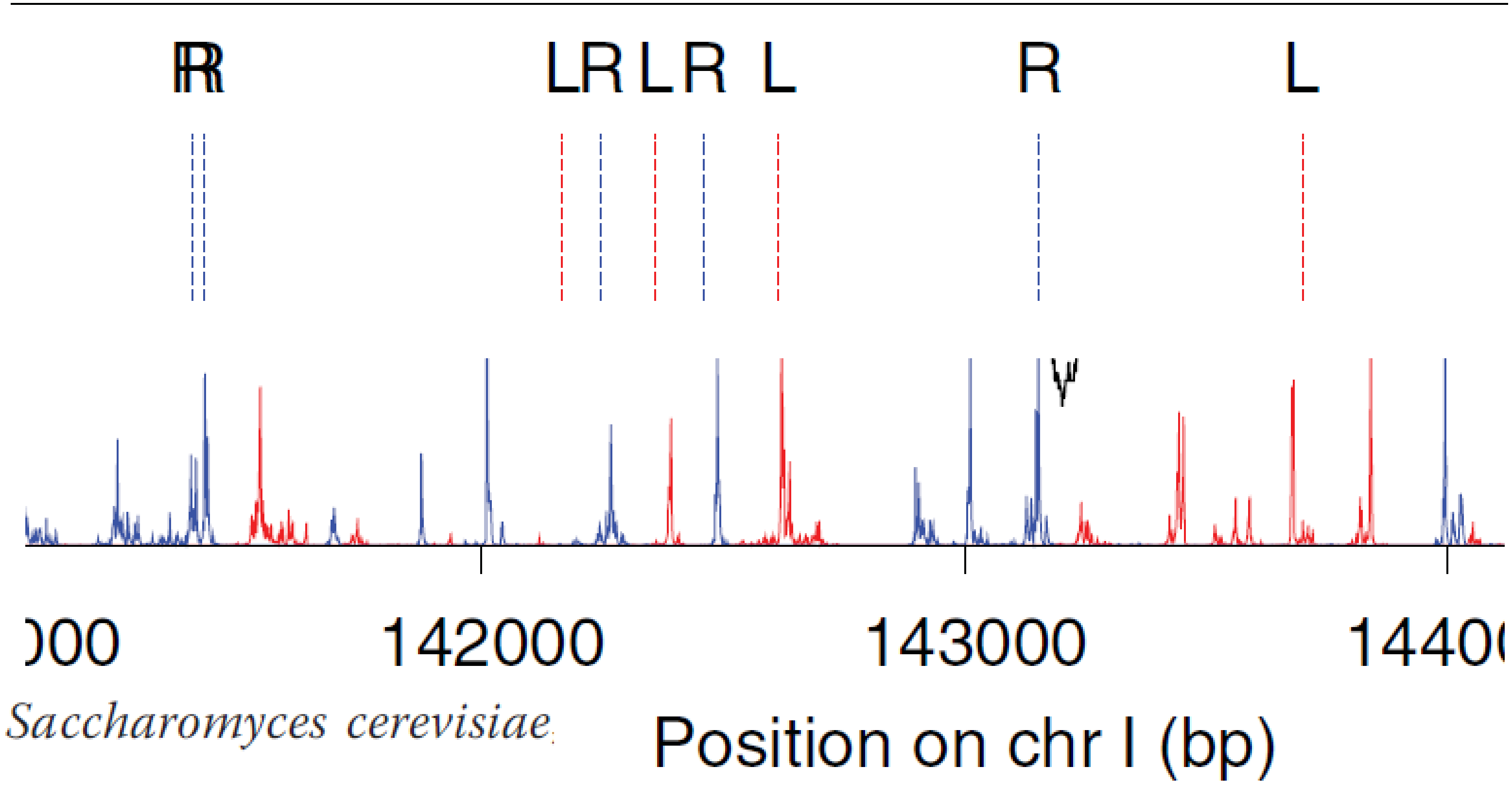
Null models with confounder track

- The relation between two tracks can be modulated by a third track.

- Such a third track acts as a confounder: if ignored it leads to wrong conclusions on the relation between the two tracks of interest.

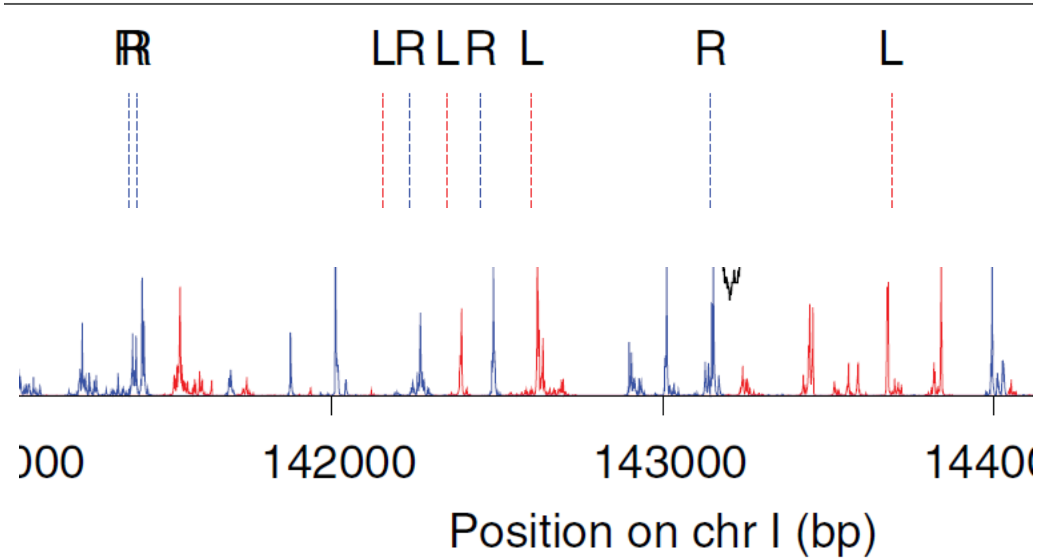
Example: the relation of coding regions with the melting stability of the DNA double helix

T1= UP Locations of exon boundaries (left and right)



T2=F Probability of melting (left and right)

⟶ **Melting peaks appear to coincide with exon boundaries!**



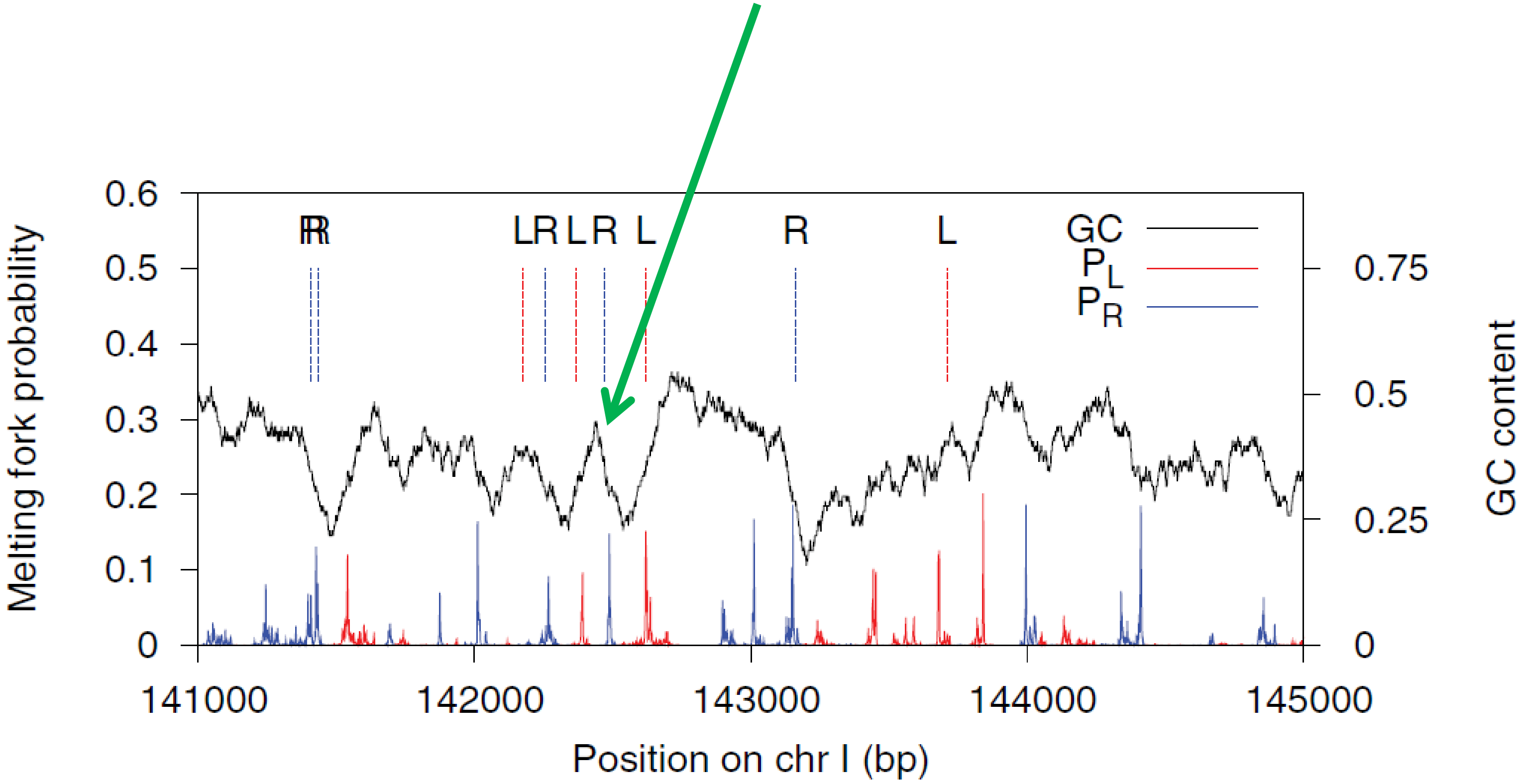
- Are melting fork probabilities higher at the exon boundaries than elsewhere? Higher than expected by chance?

- Null model: the function was conserved, while points were uniformly randomized in each chromosome.

- Monte Carlo testing was carried out on each chromosome separately, giving p-values <0.0005.

- There is an interesting relation between DNA melting and coding regions!

The black curve shows the GC content (%) in a 100-bp sliding window



- An alternative view is that the GC content is governing the relation between exons and melting probability. This because GC% is higher inside exons than outside, and higher when melting probability is high.

Confounder tracks

- Non-preserved elements of a null model can be randomized according to a non-homogeneous Poisson process with a bp-varying intensity, which can depend on a third (or several) modulating genomic tracks .

- Algebra for the construction of intensities: tracks are combined in various ways (with a biological meaning), to allow rich and flexible constructions of randomness and to modulate the effect of other tracks on the comparison.

Question "Do the elements of the two tracks show positive association, more than what expected by the fact that both are associated to the third track?“

- Non-homogeneous Poisson process on the line.

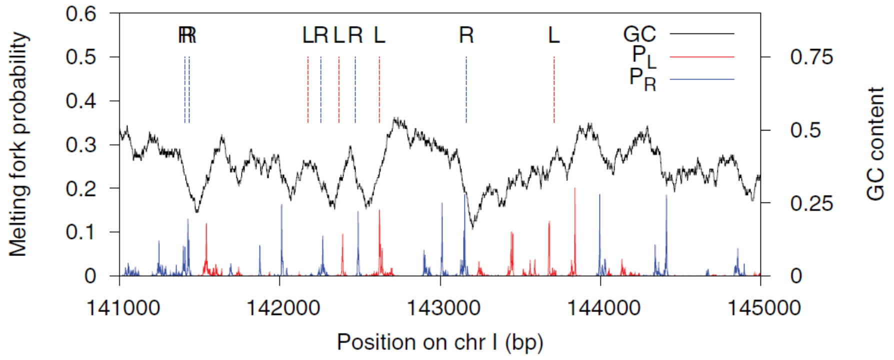
- The intensity $\lambda_3(b)$ describes how "nature" has randomised the elements, now observed as in the present genomic track.

- A third track can be used as intensity curve in randomisation: in this way, all simulated Monte Carlo configurations would adhere to the third track.

- Small p-values would indicate that the association between the two tracks is significantly higher than what expected by their joint dependency on third track.

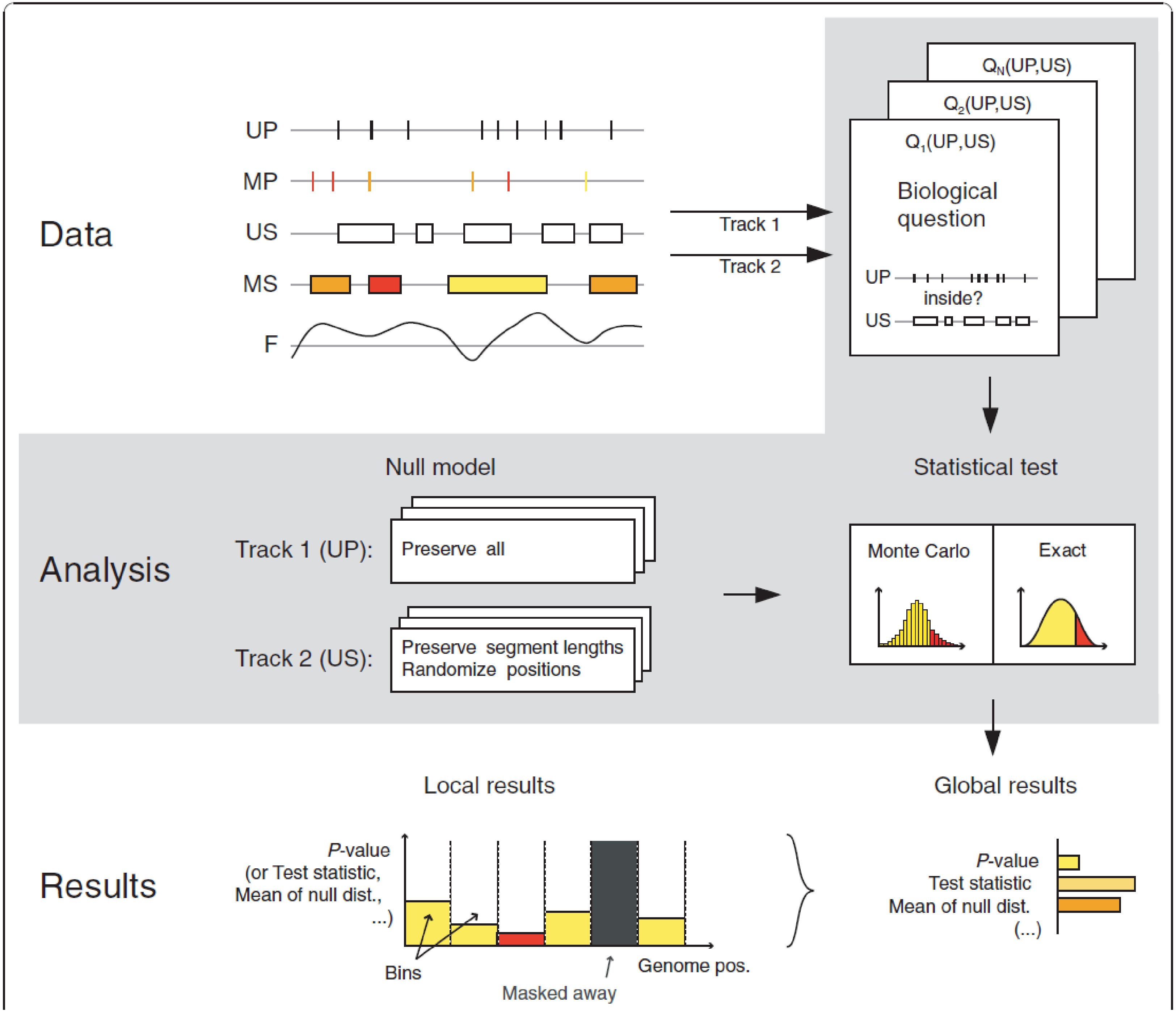
- When still significant concordance, there must be other phenomena that act on the association, in addition to the third track – a further mechanism of interference.

- The GC-content function is used as intensity $\lambda_3(b)$ when randomizing exons.



- When performing the same analysis as before, but now using the null model based on the GC intensity curve, a significant relationship was found only in **one** yeast chromosome.

- There is a melting-exon relationship in yeast, but it may simply be a consequence of differences in GC content at the exon boundaries, which may exist for biological reasons not involving melting fork locations. However, there might be some additional local mechanism disturbing the association in **one** chromosome.


Data
UP
MP
US
MS
F
Track 1
Track 2
Q_N(UP,US)
Q_2(UP,US)
Q_1(UP,US)
Biological question
UP
inside?
US
Statistical test
Analysis
Null model
Track 1 (UP):
Preserve all
Track 2 (US):
Preserve segment lengths
Randomize positions
Monte Carlo
Exact
Results
Local results
P-value
(or Test statistic,
Mean of null dist.,
...)
Bins
Masked away
Genome pos.
Global results
P-value
Test statistic
Mean of null dist.
(...)

PLAN

- Genomic Hyperbrowser
- p>>n regression ➡ LASSO

$p >> n$ regressions in genomics

$$y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \epsilon_i, \qquad i = 1, \cdots, n$$

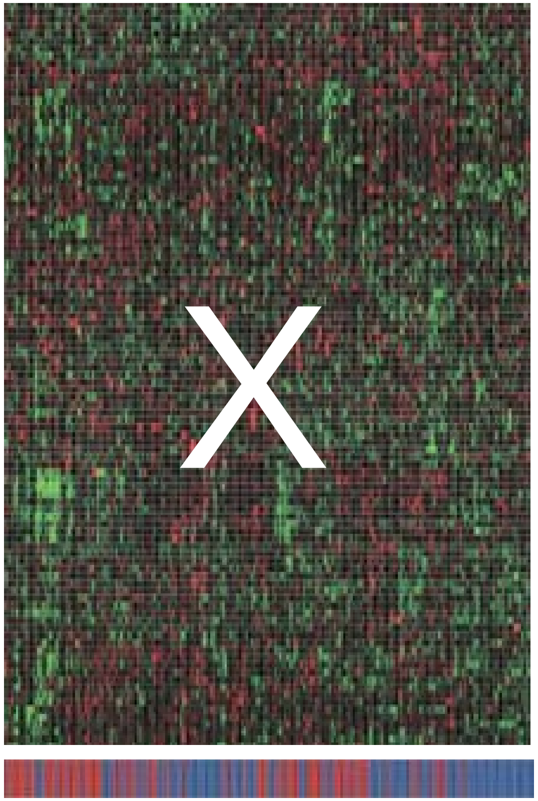
Bone biopsy data (Reppe et al., 2010)

- $n = 84$ women, y bone density, $p = 22815$ gene expressions

GWAS, Parkinson's disease data (Hamza et al., 2010)

- $n = 3986$ cases and controls, $p = 811917$ SNPs, logistic regression n=198 breast cancer patients

p = 19800 genes



survival after surgery Y

Which genes (if any) can help make a prognosis and predict survival?

Variable selection.

The linear regression model

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i, \; i = 1, \ldots, n$$

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

responses

$$\boxed{p \leq n} \qquad \boxed{p > n}$$

The linear regression model

$\boxed{p \leq n}$

More patients than unknowns!

$\boxed{p > n}$

Less patients than unknowns!
Non-identifiable. Infinitely many equivalent solutions.

p=20 000 genes
n=100 patients

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \qquad \text{with} \;\; p >> n$$

Goals:

- Prediction of new response, w.r.t. squared prediction error
- Estimation of $\boldsymbol{\beta}$, w.r.t. $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2$
- Variable selection, estimating active set (variables with coefficient $\neq 0$)

$p >> n$: Fewer equations than unknowns - infinitely many solutions - need to impose additional assumptions.

Many approaches:

- Bayesian variable selection
- Forward selection
- Preliminary dimension reduction (pre-selection)
- Penalization and shrinking

l_1 penalized regression

Regularize with l_1-penalty: $||\beta||_1 = \sum_{j=1}^{P} |\beta_j|$

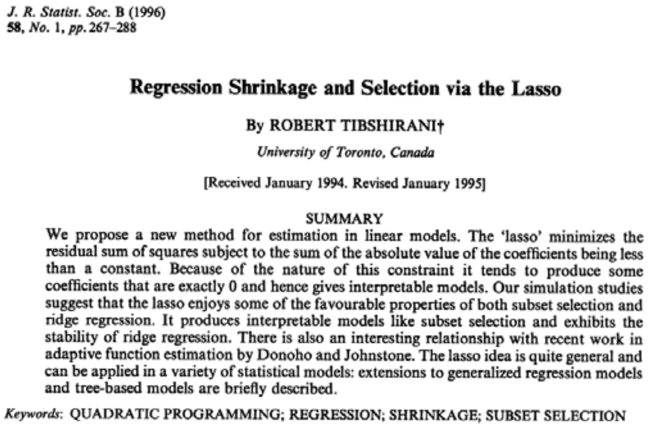
$$\hat{\beta} = \text{argmin}_{\beta}(-n^{-1} l(\beta) + \lambda ||\beta||_1)$$

$l(\beta)$ log-likelihood function, $\lambda \geq 0$ penalization parameter.

Much theory:

- Assume that the true solution is sparse, |Active set| $= s_0$ is small $< n << p$.
- If sparsity is actually true, you will recover it (theorems and algorithms).
- If truth not sparse, then no method can do well (Bühlmann, Van de Geer).

The LASSO



J. R. Statist. Soc. B (1996)
58, *No.* 1, *pp.* 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]

SUMMARY

We propose a new method for estimation in linear models. The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ridge regression. There is also an interesting relationship with recent work in adaptive function estimation by Donoho and Johnstone. The lasso idea is quite general and can be applied in a variety of statistical models: extensions to generalized regression models and tree-based models are briefly described.

Keywords: QUADRATIC PROGRAMMING; REGRESSION; SHRINKAGE; SUBSET SELECTION

$$\hat{\beta}(\lambda) = \arg\min_{\beta}\left(\underbrace{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n}_{\text{fit}} + \underbrace{\lambda\|\beta\|_1}_{\text{penalisation}} \right)$$

- Selects the few variables which are really useful!
- Estimates the parameters

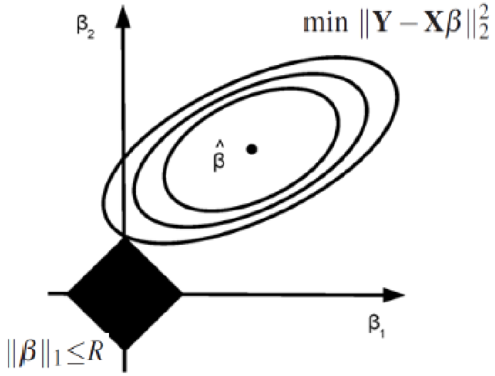
The lasso (Tibshirani, 1996)

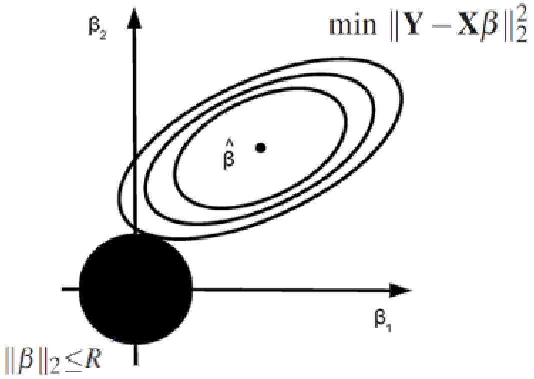
$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\{||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda \sum_{j=1}^{P} |\beta_j|\},$$

NB: MAP estimate with doubly-exponential prior on β.

Lasso does variable selection:(many) parameters are estimated to be exactly zero. To see this, use the equivalent form

$$\hat{\beta} = \operatorname{argmin}_{\beta:||\beta||_1 \leq R} ||\mathbf{y} - \mathbf{X}\beta||_2^2/n$$





$$\hat{\beta}_{\text{Ridge}}(\lambda) = \underset{\beta}{\arg\min}\left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_2^2\right)$$

Choosing the penalization parameter λ

K-fold cross-validation

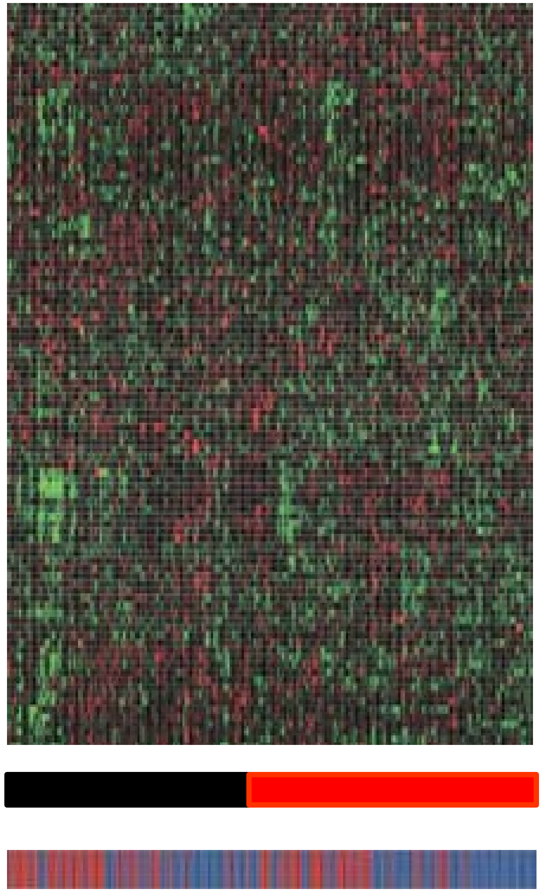
- Make a grid of $\lambda_{min} < \lambda < \lambda_{max} = \max_j |\mathbf{x}_j^T \mathbf{y}|$ (no variable selected).
- Minimize estimated prediction error $CV(\lambda)$ over λ-gridpoints.
- For the linear model:

$$CV(\lambda) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in f_k} (y_i - \hat{y}_i^{-k}(\lambda))^2.$$

f_k is the set of indices for samples in fold k, $\hat{y}_i^{-k}(\lambda)$ is the fitted predicted value for observation i when fold k involving observation i is left out of the estimation.

n=198 breast cancer patients

p = 19800 genes

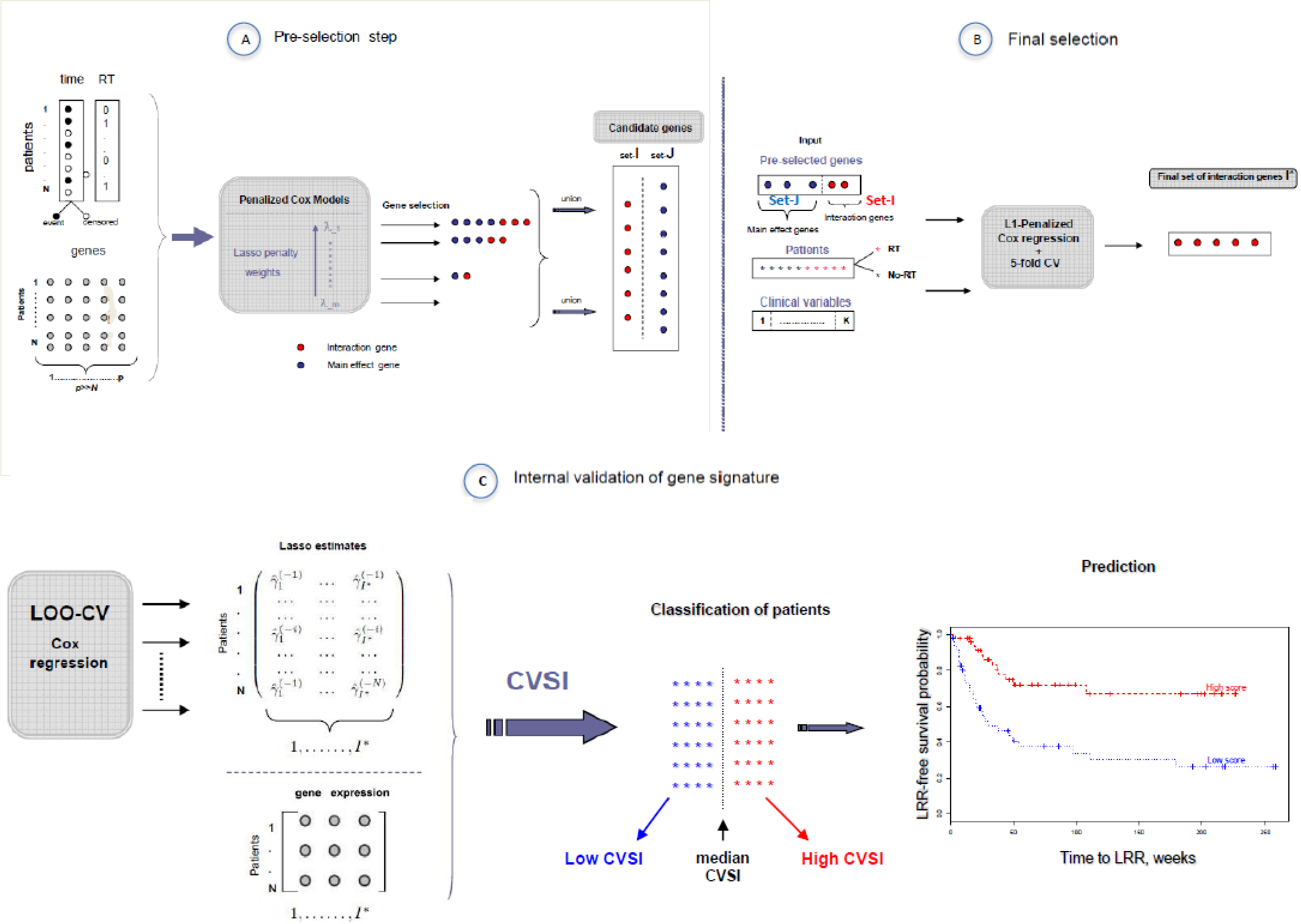


Radiotherapy or not

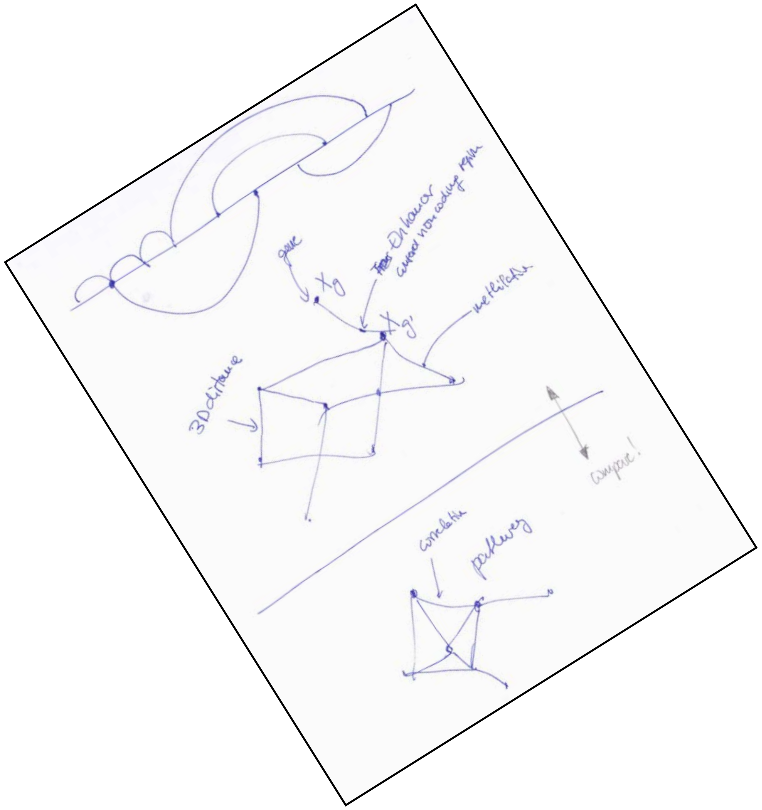
survival after surgery

Which genes (if any) can help to decide when radiotherapy prolongs survival?

Variable selection of the genes that **interact** with radiotherapy.


A
Pre-selection step
time
RT
patients
1
N
event
censored
genes
Patients
1
N
1.....................p
p>>N
Penalized Cox Models
Lasso penalty weights
λ_1
λ_m
Gene selection
union
union
Candidate genes
set-I
set-J
Interaction gene
Main effect gene
B
Final selection
Input
Pre-selected genes
Set-J
Set-I
Main effect genes
Interaction genes
Patients
RT
No-RT
Clinical variables
1
K
L1-Penalized Cox regression + 5-fold CV
Final set of interaction genes I*
C
Internal validation of gene signature
LOO-CV
Cox regression
Lasso estimates
Patients
1
N
1,......., I*
gene expression
Patients
1
N
1,......., I*
CVSI
Classification of patients
Low CVSI
median CVSI
High CVSI
Prediction
LRR-free survival probability
Time to LRR, weeks
High score
Low score
0
50
100
150
200
250

Conclusions



- Statistics for Innovation is an exciting and successful experiment on the international arena of statistics
- A new sfi application is in progress.
- Exiting years to come for statistics!
- There should be many exciting occasions of collaborations, across disciplines.

Thank You