Statistical analysis of RIM data (retroviral insertional mutagenesis)

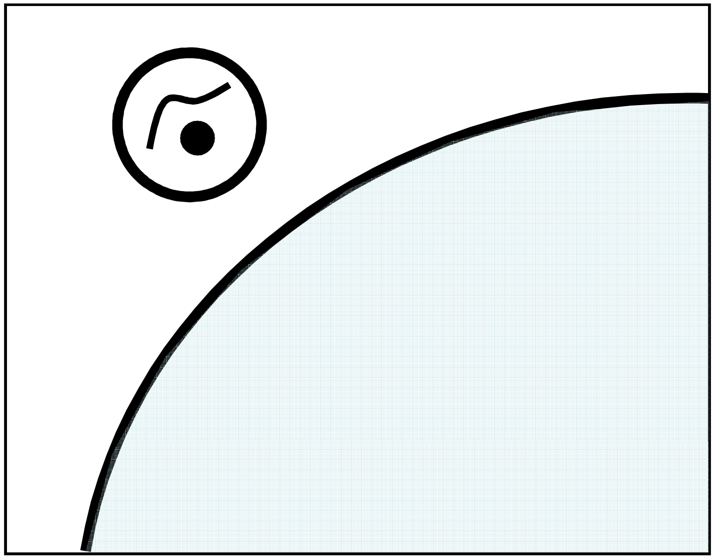
Lodewyk Wessels

Bioinformatics and Statistics

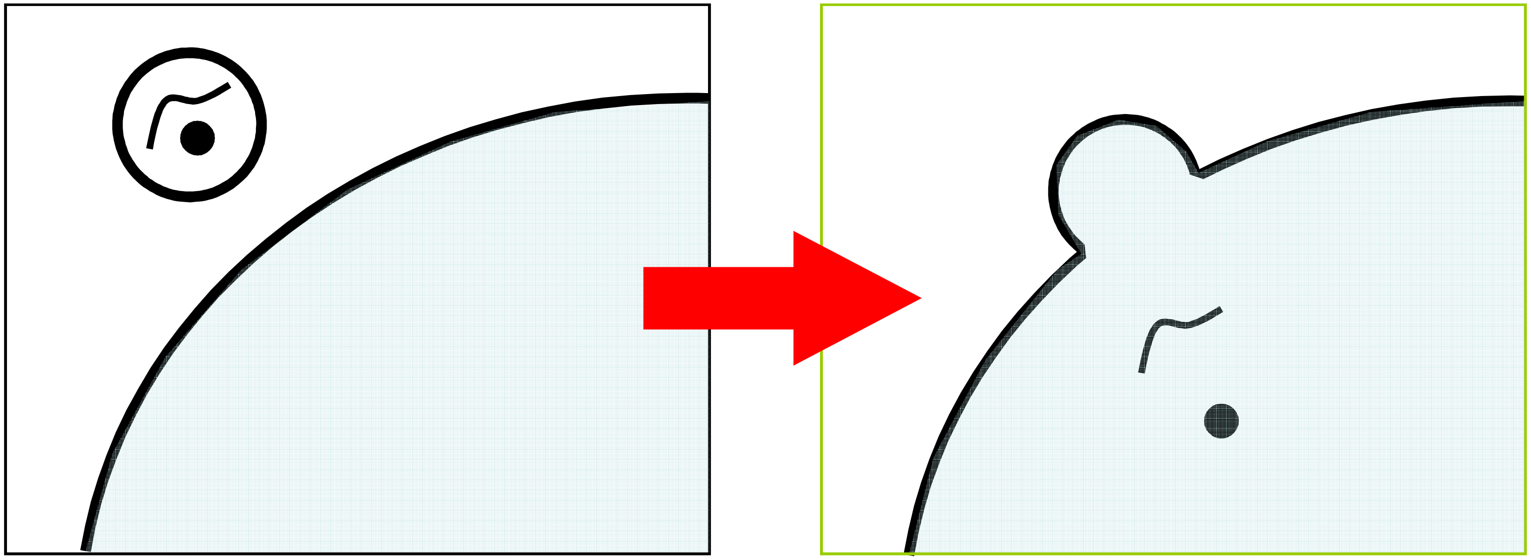
The Netherlands Cancer Institute

Amsterdam

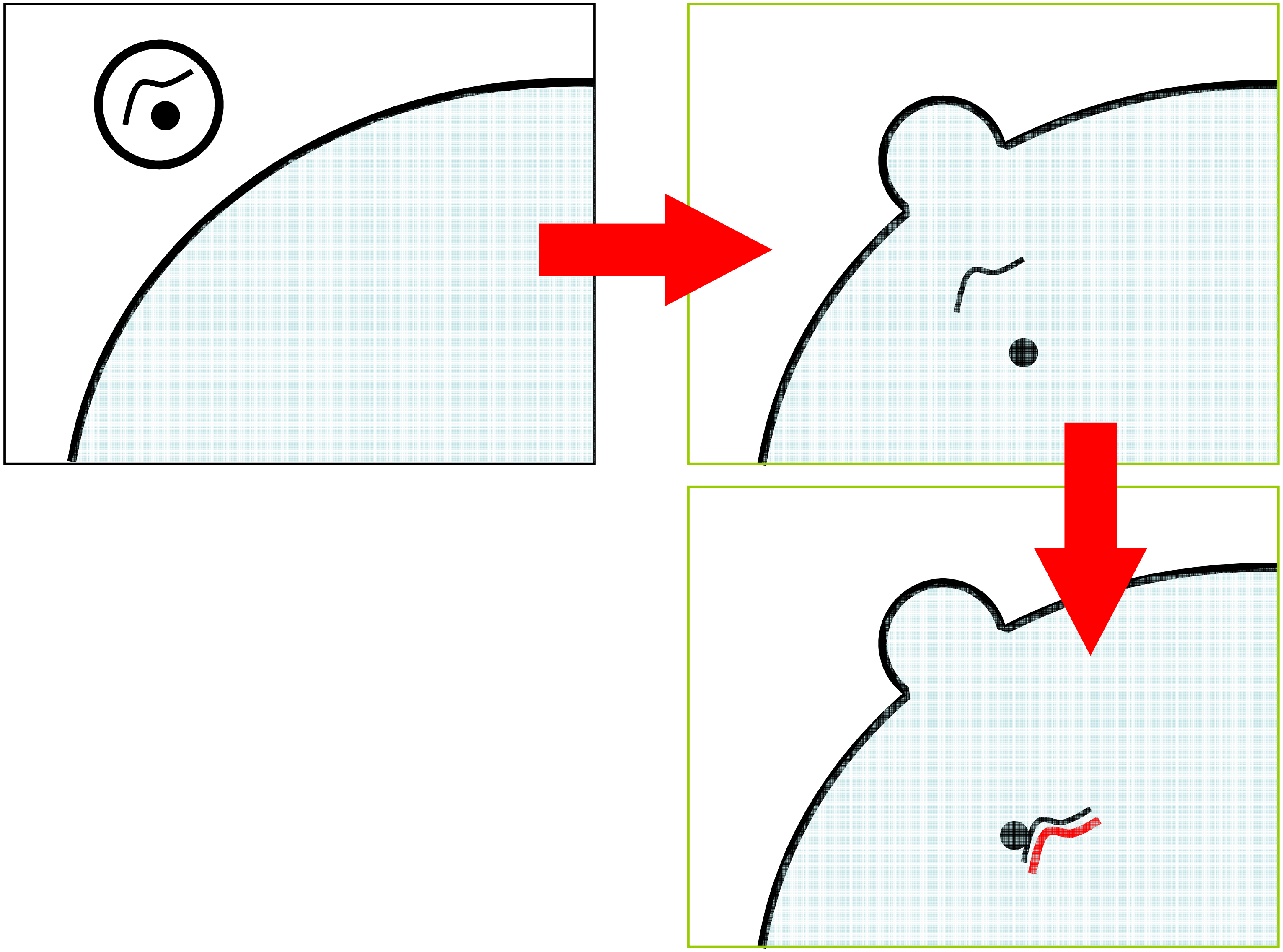
Viral integration



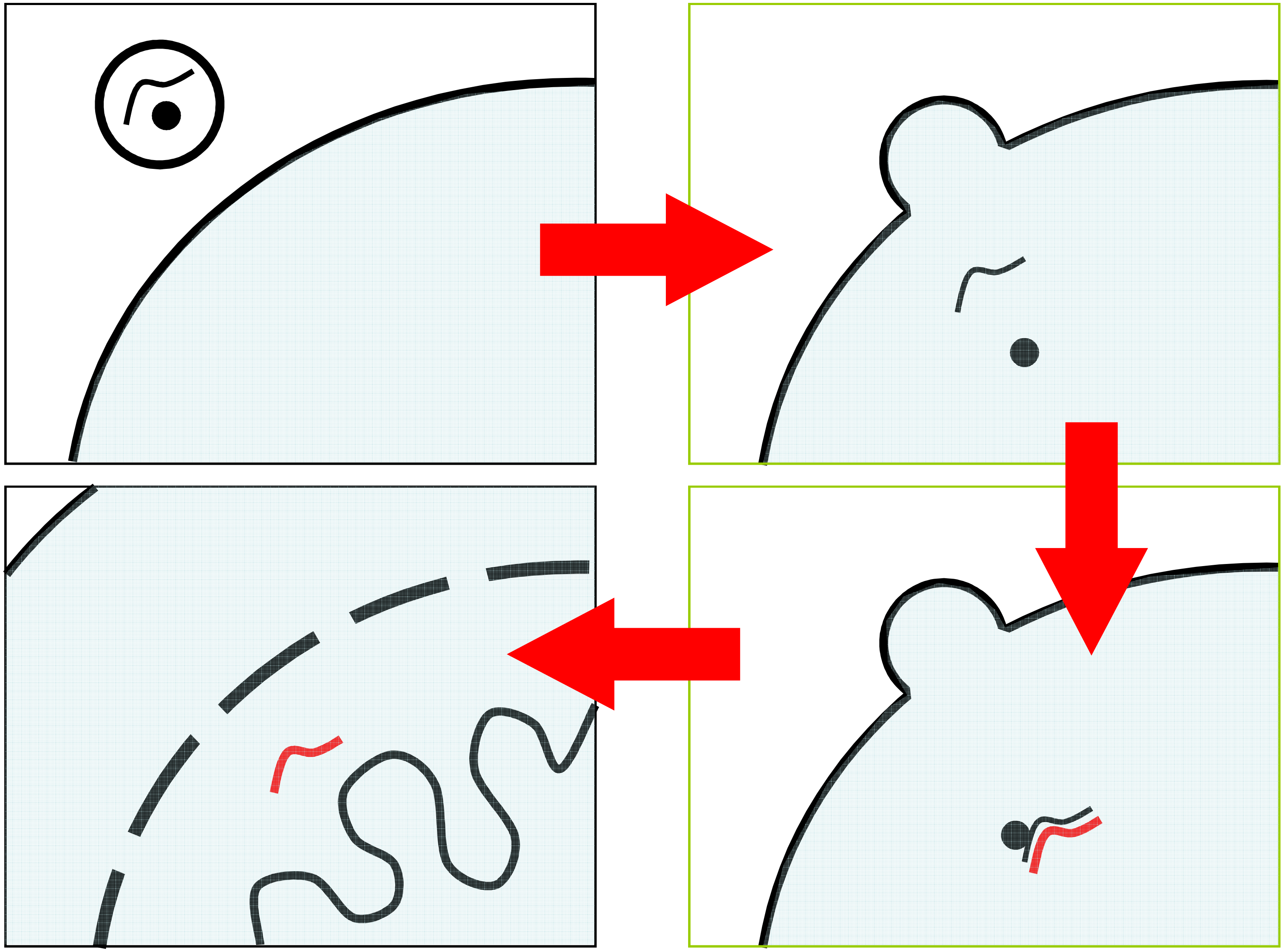
Viral integration



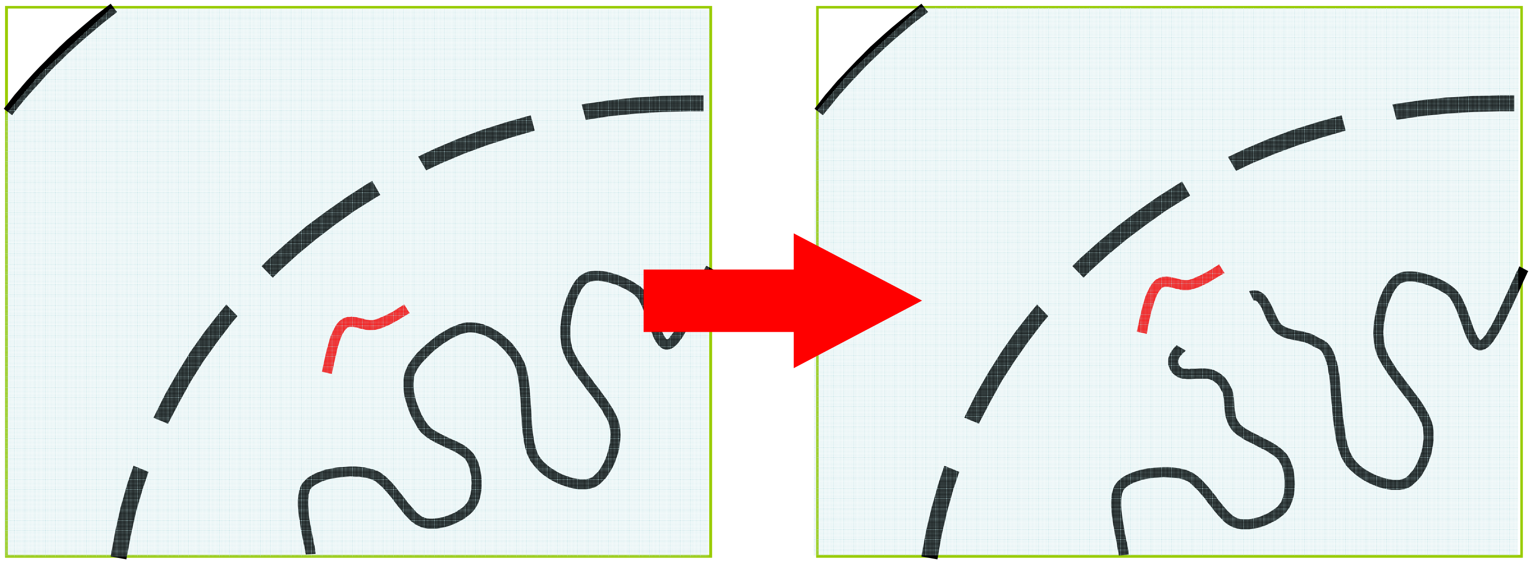
Viral integration



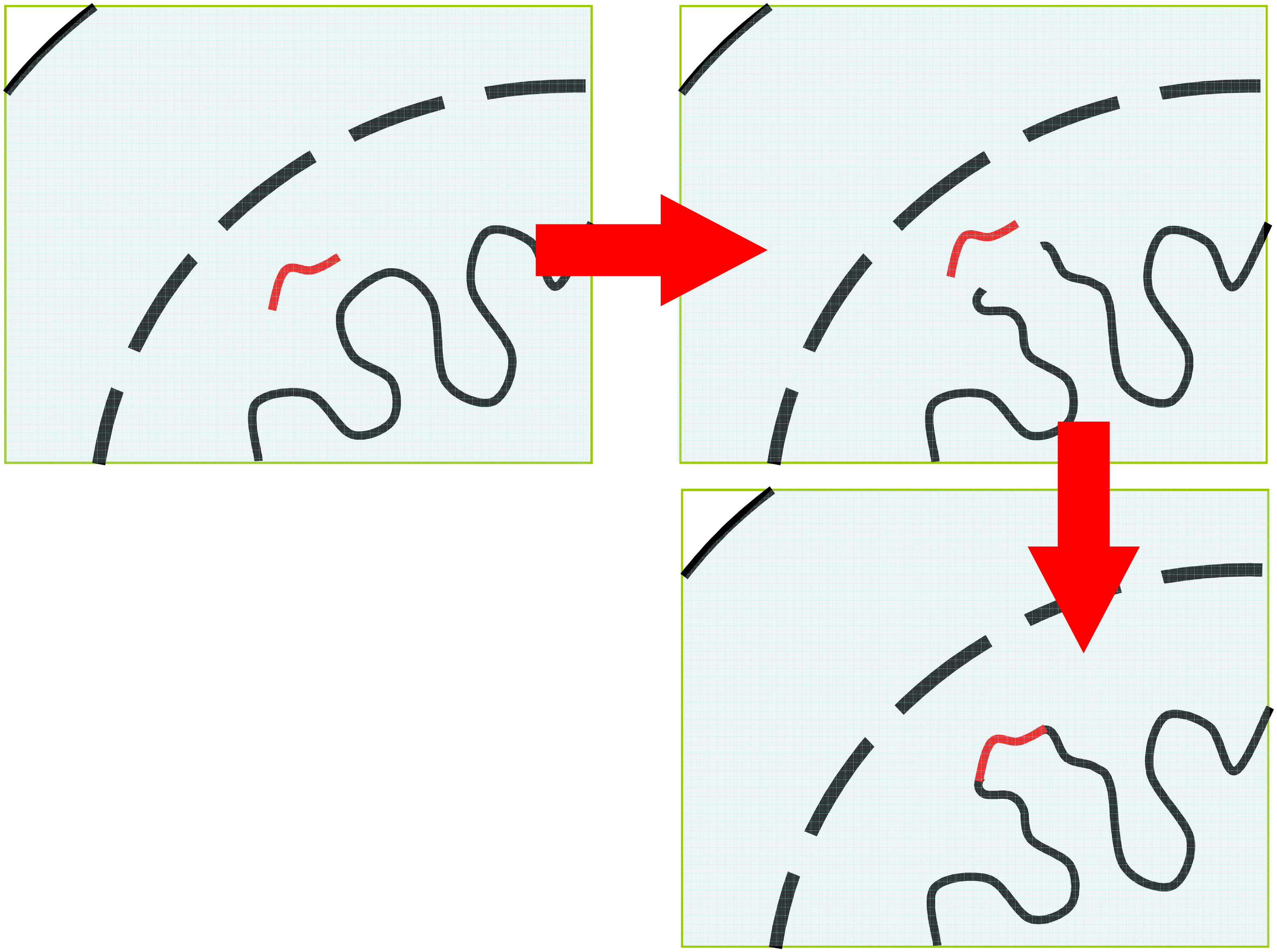
Viral integration



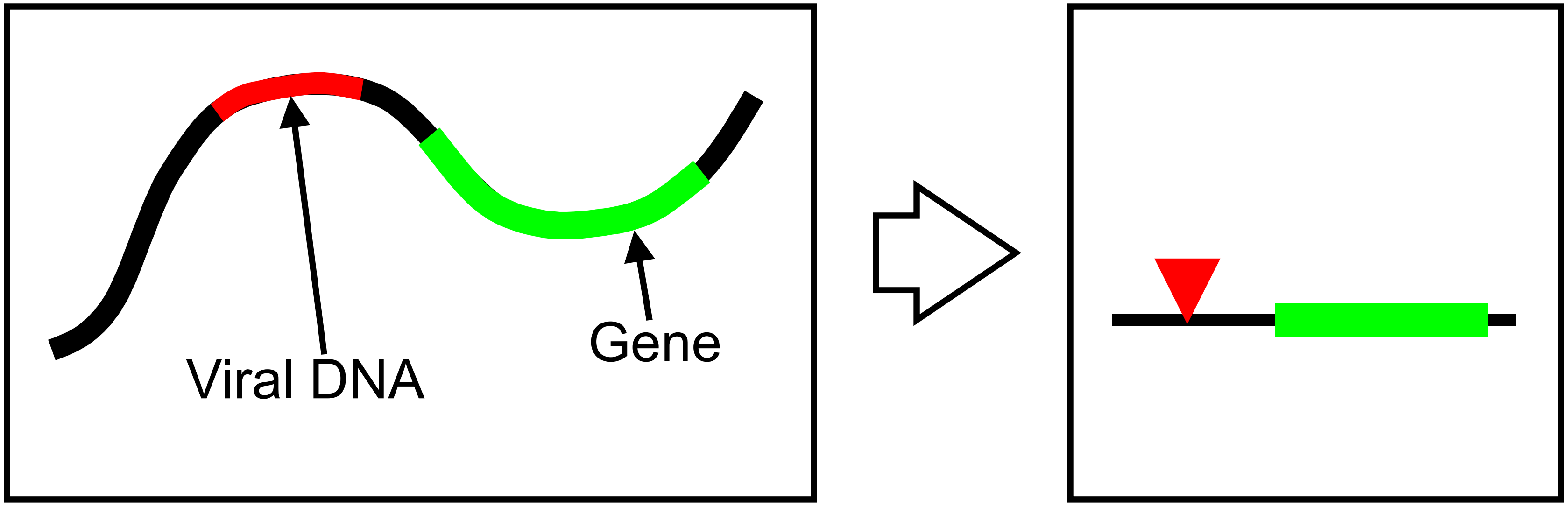
Viral integration



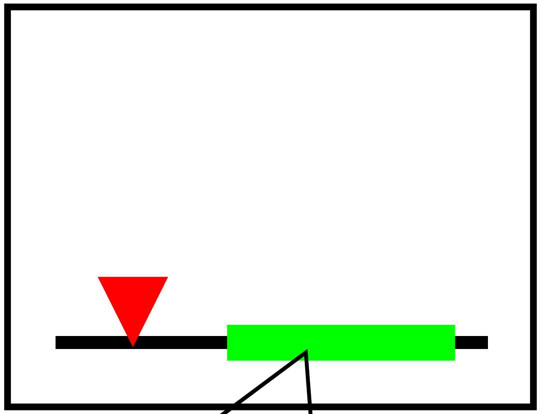
Viral integration



Effects of a viral insert

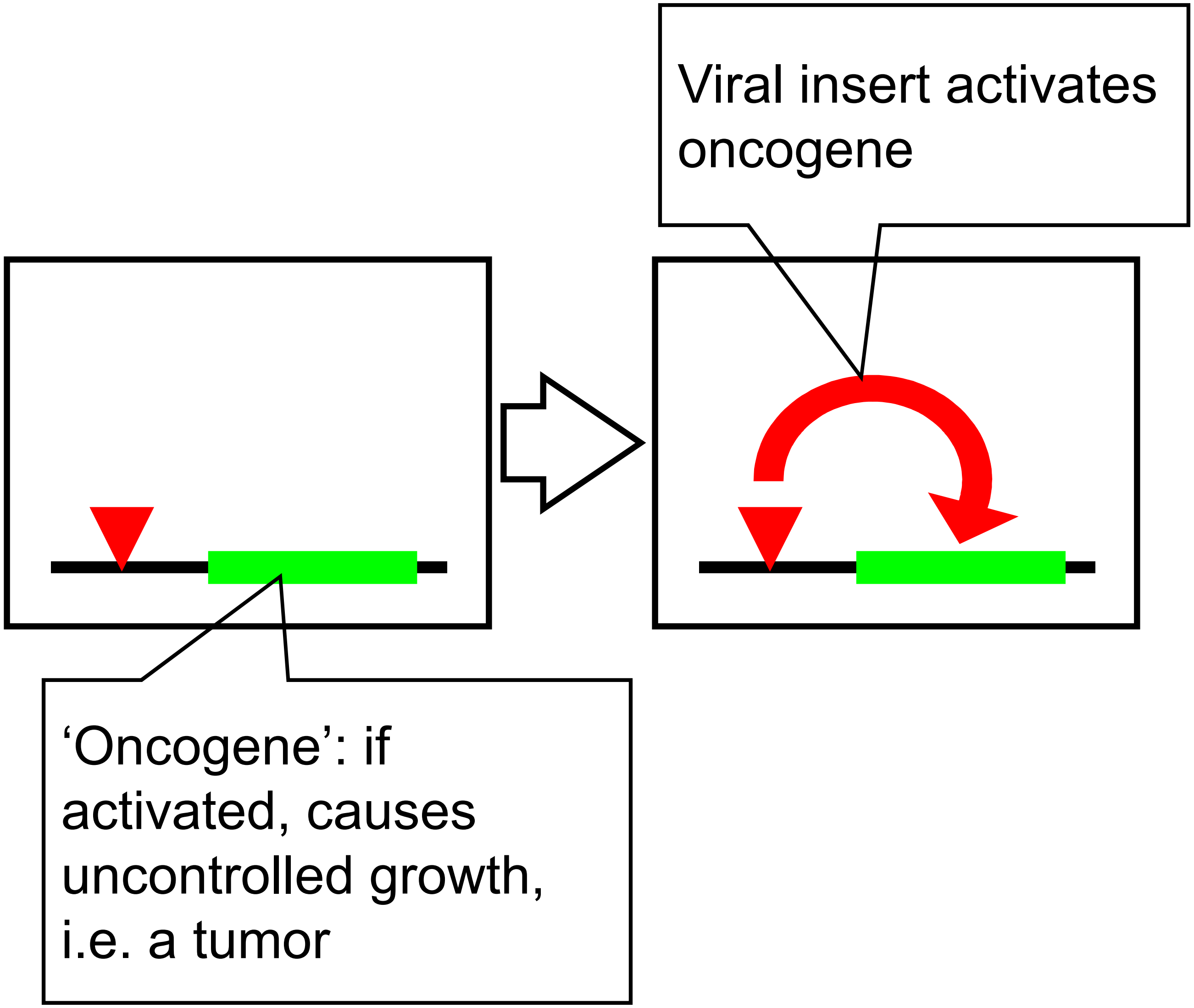


Effects of insert: oncogene activation



‘Oncogene’: if activated, causes uncontrolled growth, i.e. a tumor

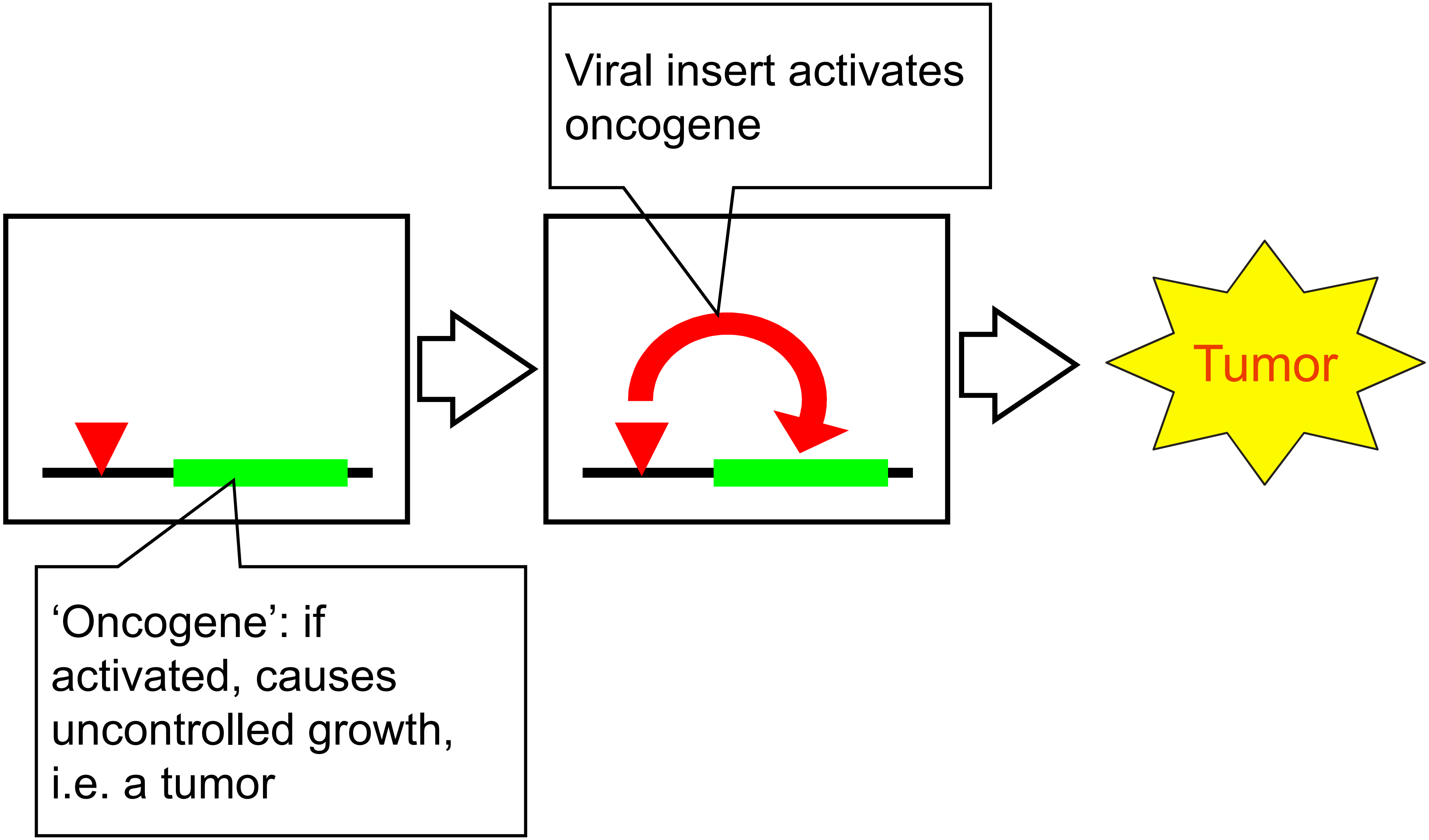
Effects of insert: oncogene activation



Viral insert activates oncogene

‘Oncogene’: if activated, causes uncontrolled growth, i.e. a tumor

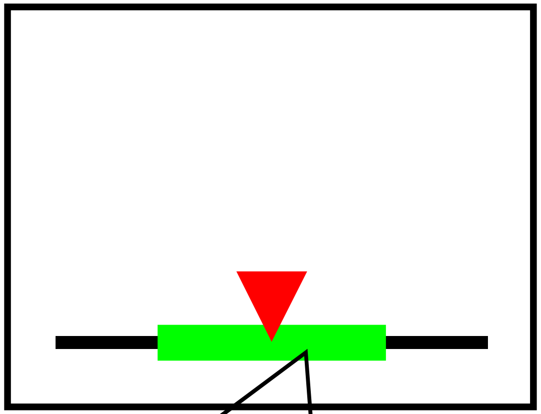
Effects of insert: oncogene activation



Viral insert activates oncogene

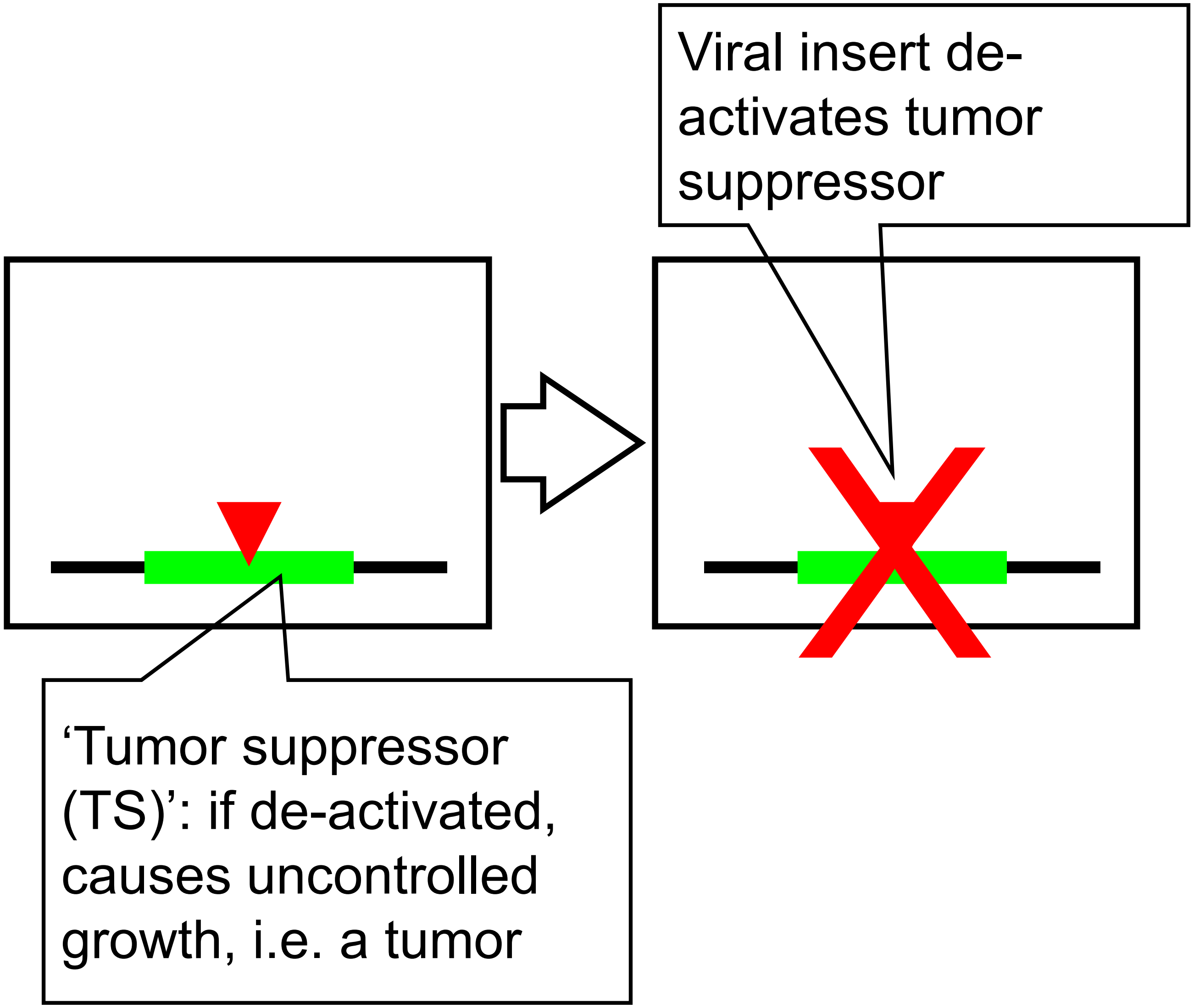
‘Oncogene’: if activated, causes uncontrolled growth, i.e. a tumor

Effects of insert: TS inactivation



‘Tumor suppressor (TS)’: if de-activated, causes uncontrolled growth, i.e. a tumor

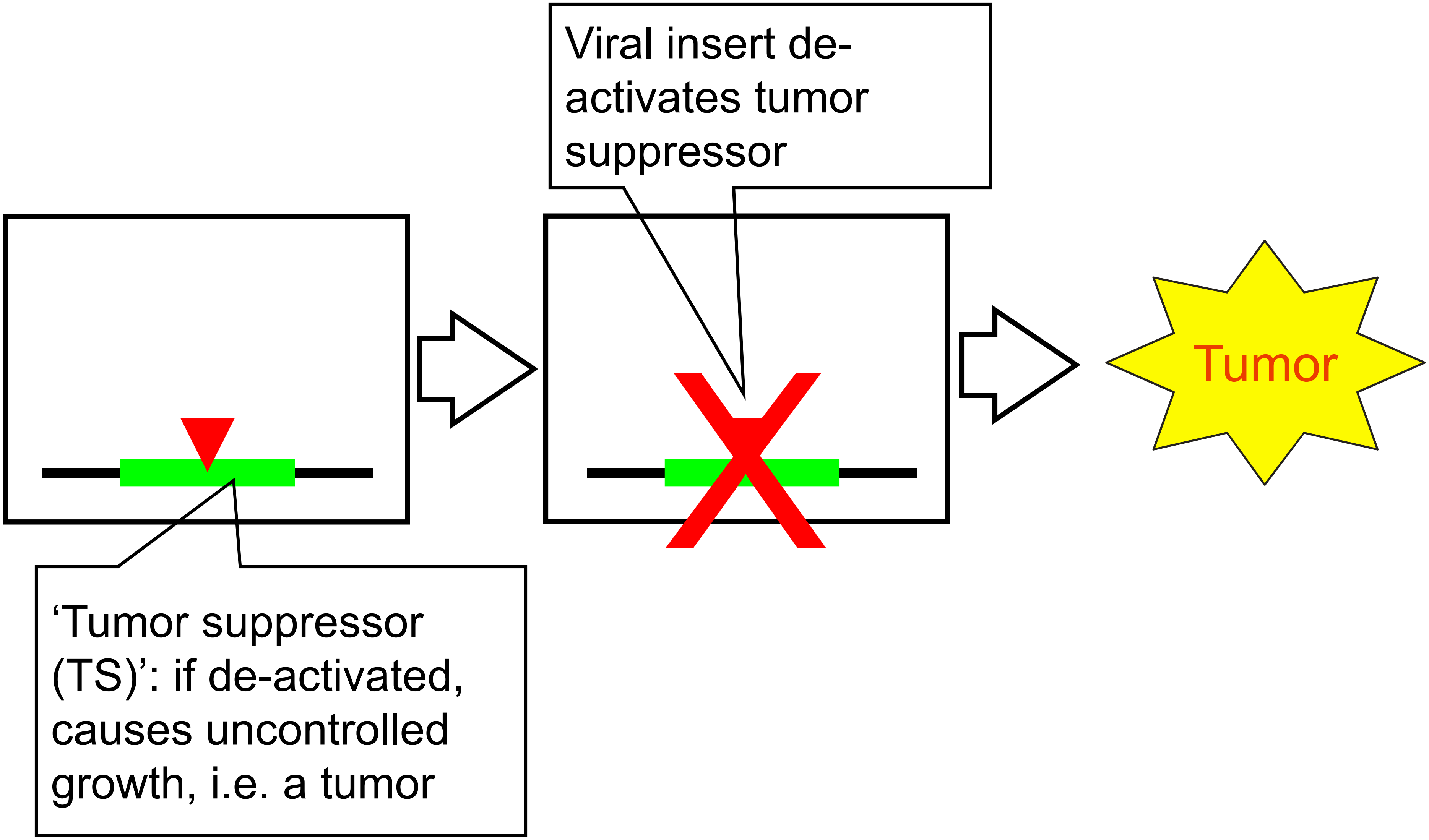
Effects of insert: TS inactivation



Viral insert de-activates tumor suppressor

‘Tumor suppressor (TS)’: if de-activated, causes uncontrolled growth, i.e. a tumor

Effects of insert: TS inactivation



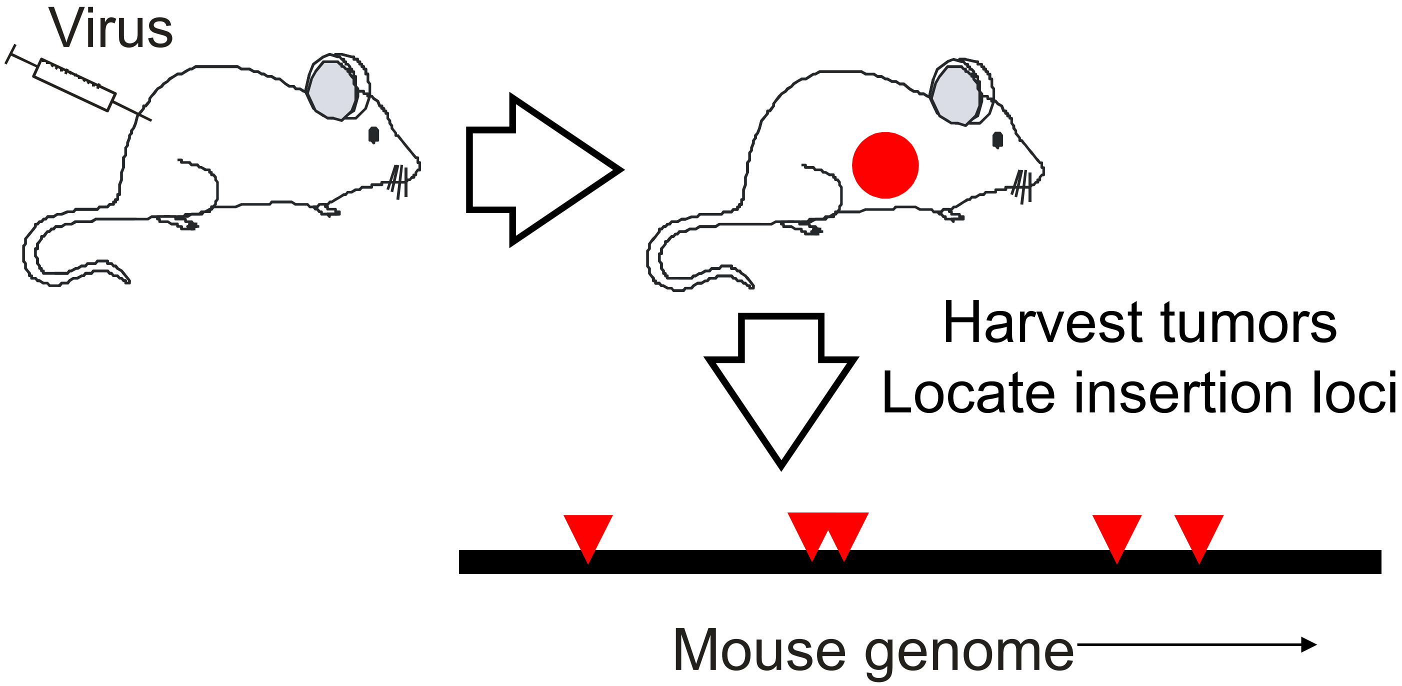
Viral insert de-activates tumor suppressor

‘Tumor suppressor (TS)’: if de-activated, causes uncontrolled growth, i.e. a tumor

Cancer research

- Find oncogenes and tumor suppressors
- If these are known we better understand mechanisms of cancer
- We can devise better targeted treatments
- Exploit viral integration in model systems to hunt for oncogenes and tumor suppressors

Experimental overview

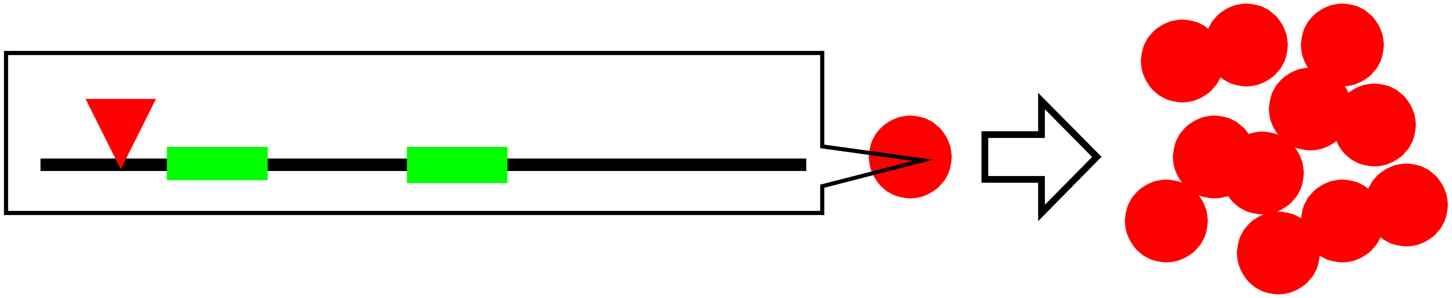


What happens in a tumor?

- Not all insertions are informative (random integration)
- Cells with oncogenic mutations have a growth advantage

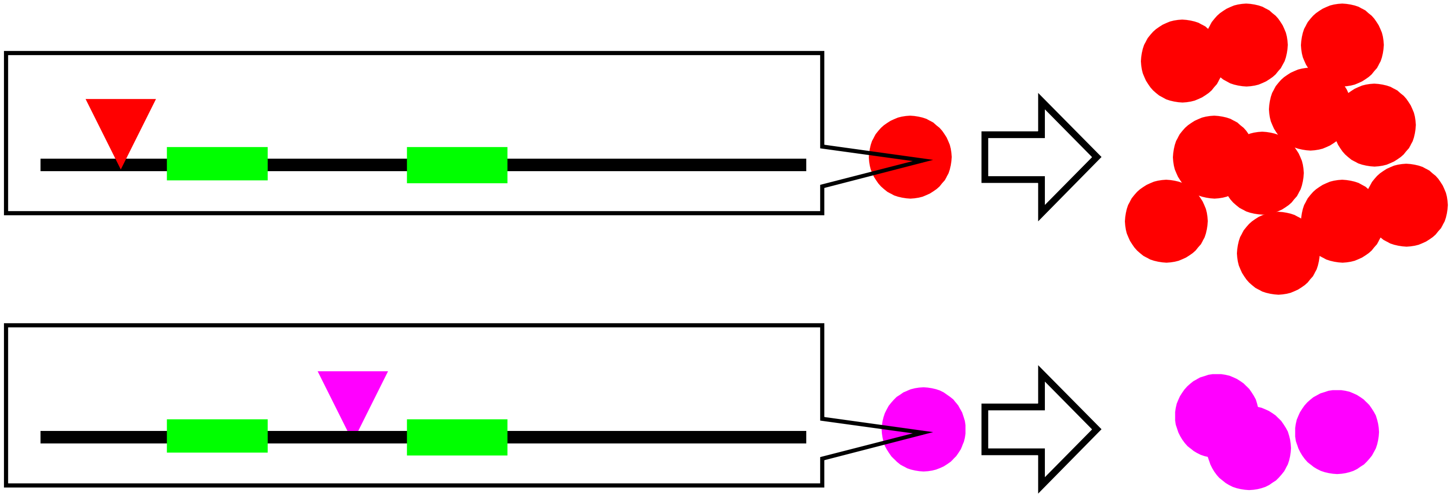
What happens in a tumor?

- Not all insertions are informative (random integration)
- Cells with oncogenic mutations have a growth advantage



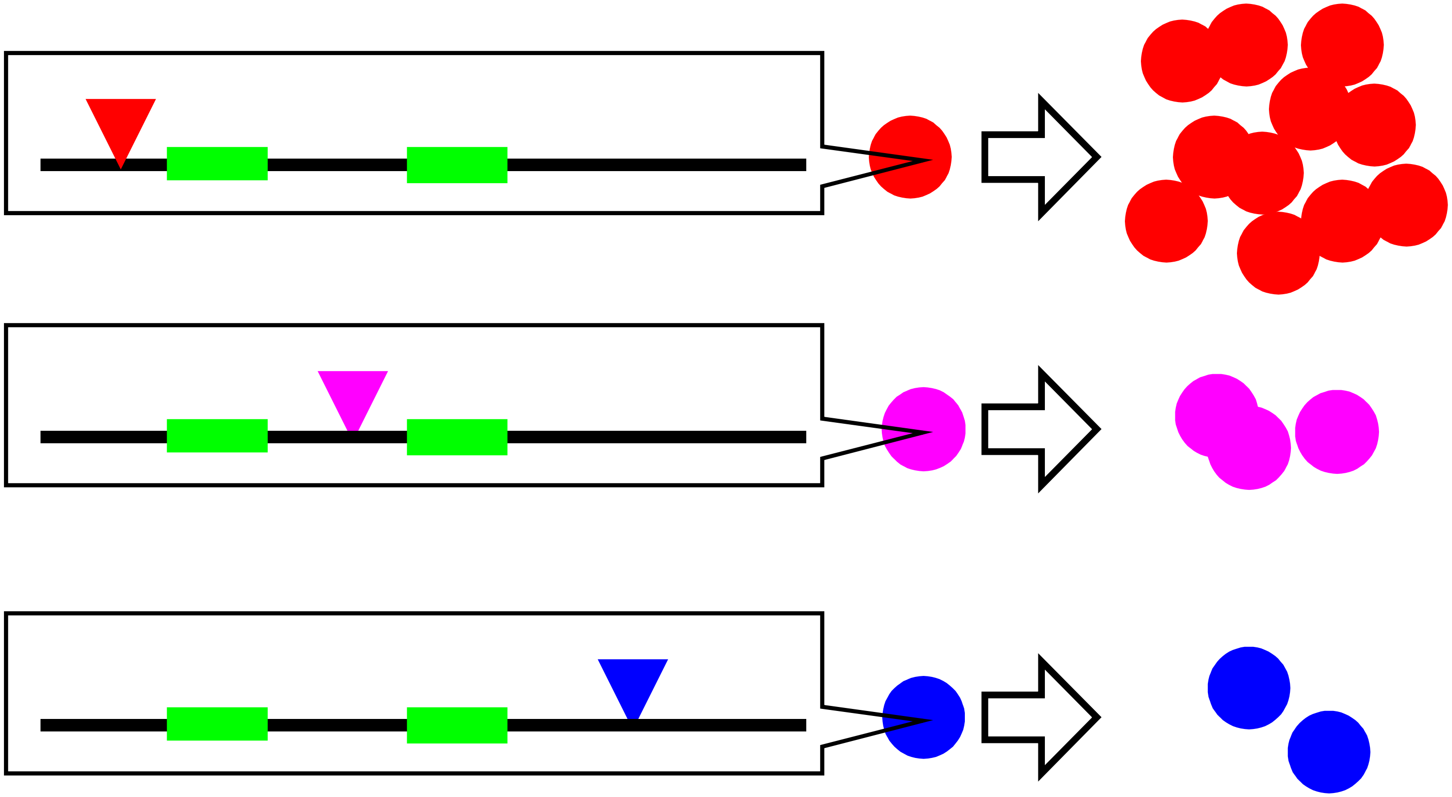
What happens in a tumor?

- Not all insertions are informative (random integration)
- Cells with oncogenic mutations have a growth advantage



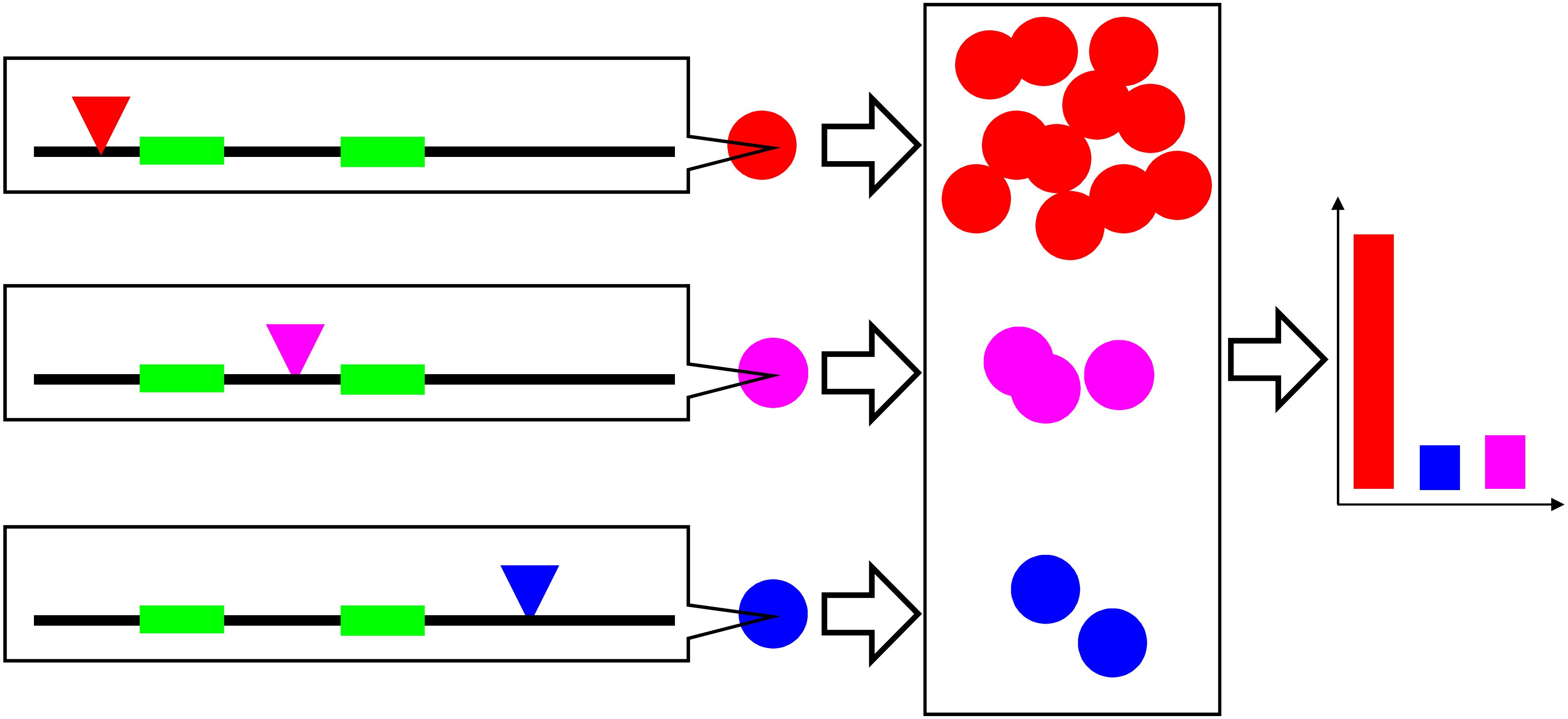
What happens in a tumor?

- Not all insertions are informative (random integration)
- Cells with oncogenic mutations have a growth advantage



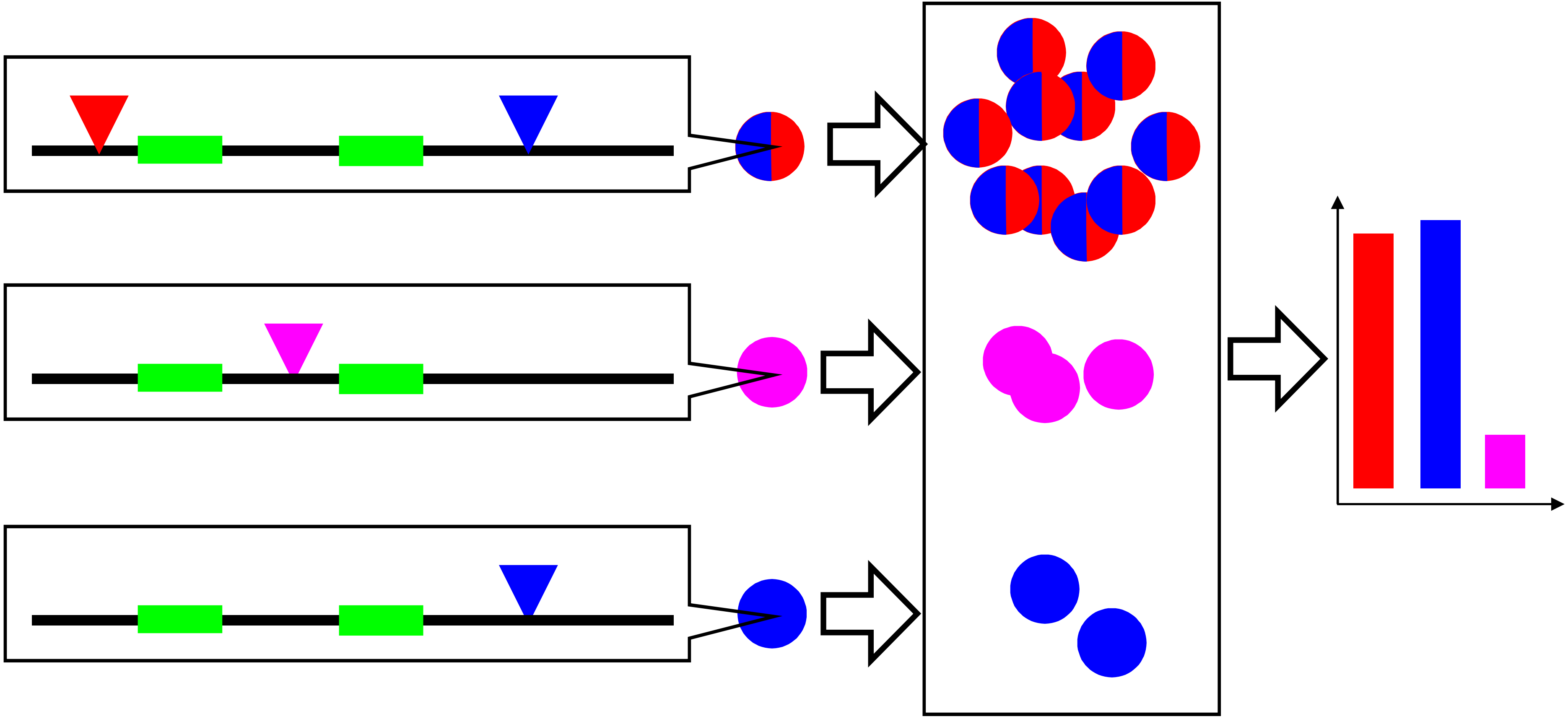
What happens in a tumor?

- Not all insertions are informative (random integration)
- Cells with oncogenic mutations have a growth advantage



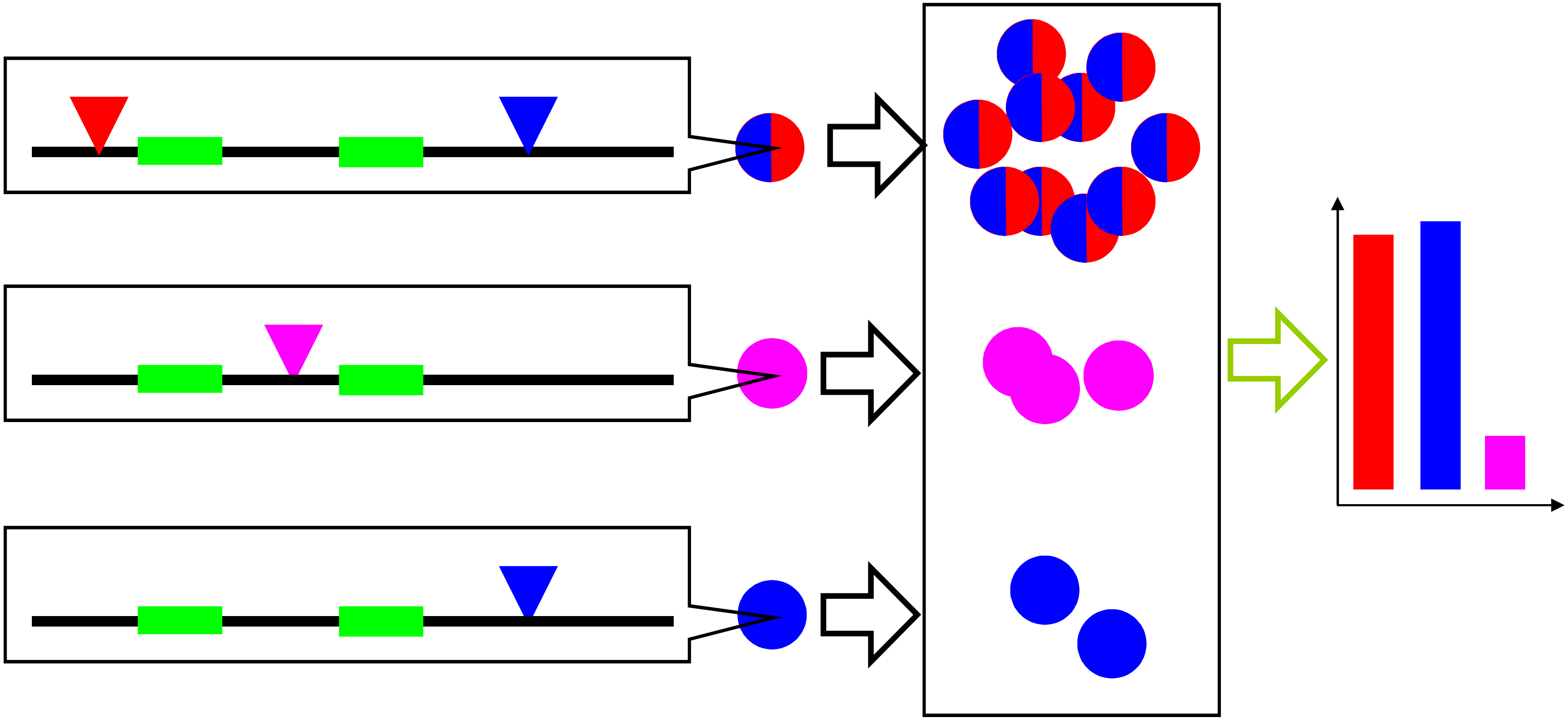
What happens in a tumor?

- But, non-causal insertions ‘piggy-back’



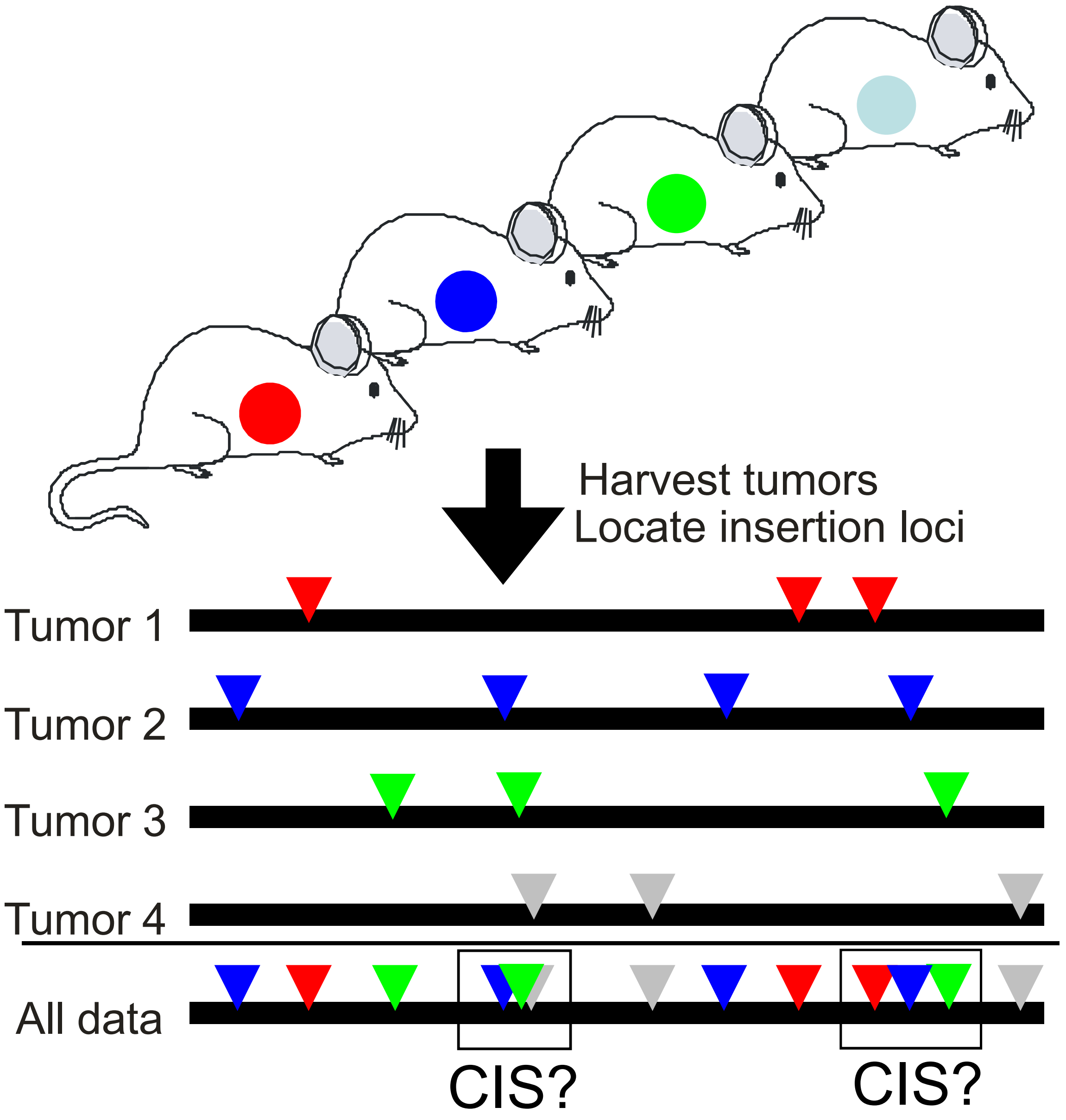
What happens in a tumor?

- But, non-causal insertions ‘piggy-back’

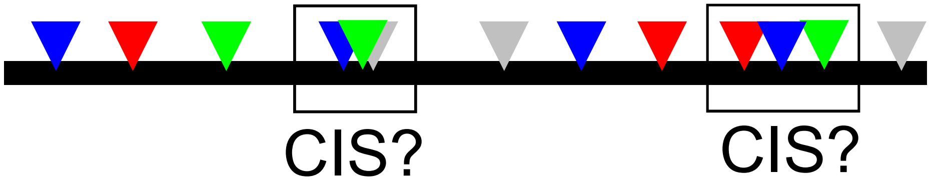


Require that insertion occurs frequently across tumors

Cancer genes: common insertion sites



Common insertion sites (2)



- Common Insertion Site (CIS):

  Region in the genome hit by viral inserts in multiple independent tumors significantly more than expected.
- CISs can be different sizes

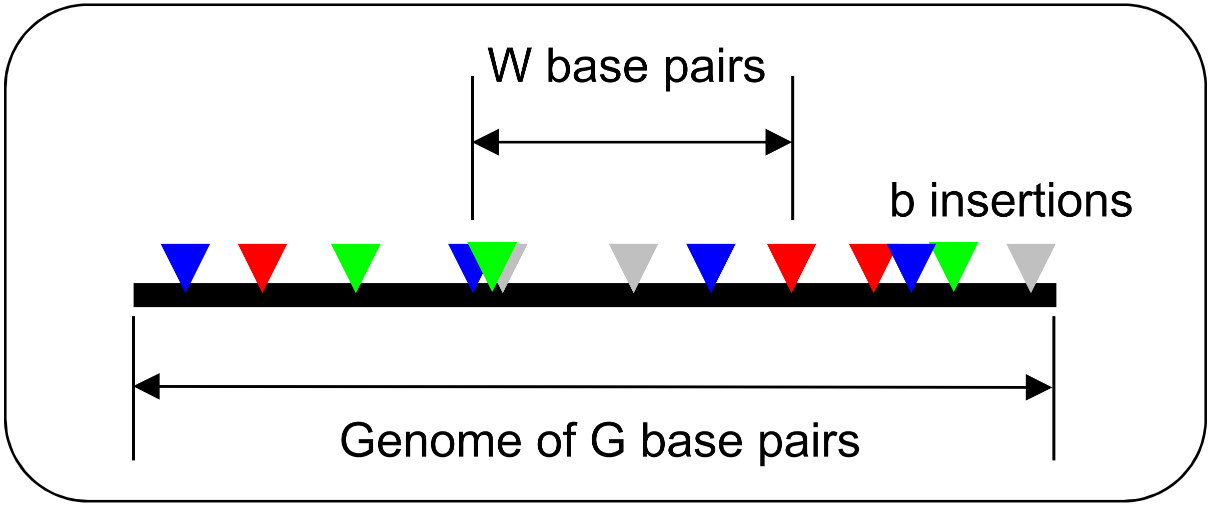
Finding cancer genes and cancer pathways

- Cancer genes:
  - genes individually frequently ‘hit’

- Cancer gene ‘pairs’:
  - pairs of genes frequently ‘hit’ in a specific pattern
  - (a gene and a family of genes frequently hit)
  - Co-operating, mutually exclusive

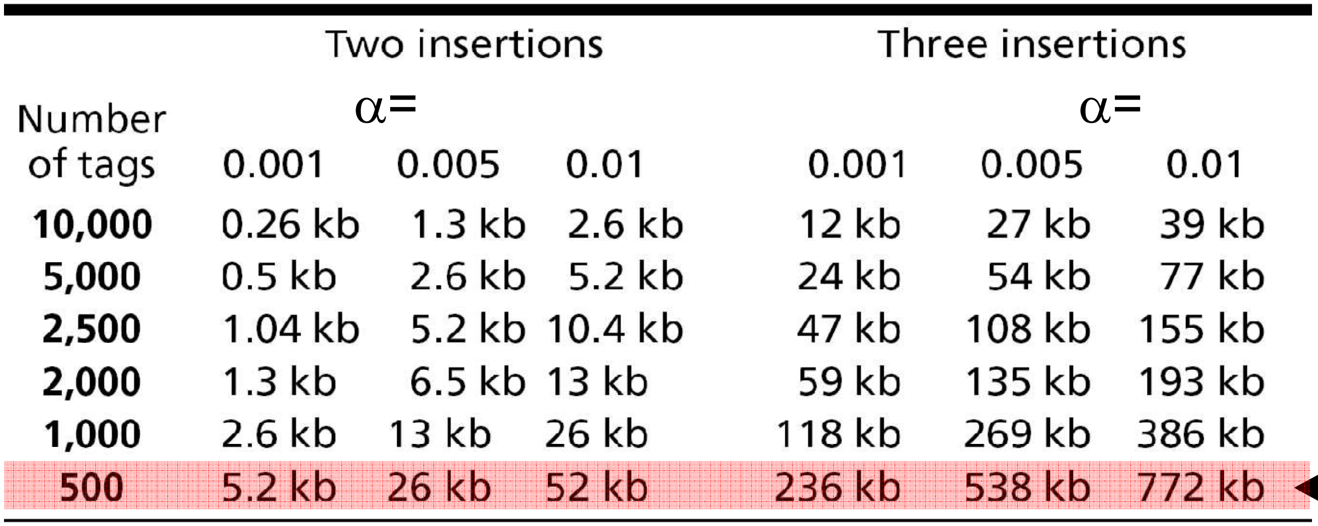
- Cancer pathways/networks
  - groups of genes frequently ‘hit’ in a specific pattern

Previous approaches: Poisson model



- Background model: insertions uniformly distributed
- k = Number of insertions in a window of W base pairs
- k ~ Poisson(W;λ)
- λ = Average number of insertions in W base pairs $\approx$ b/G
- Compute when the number of insertions exceeds the background, at a fixed α-level

Previous approaches (1)

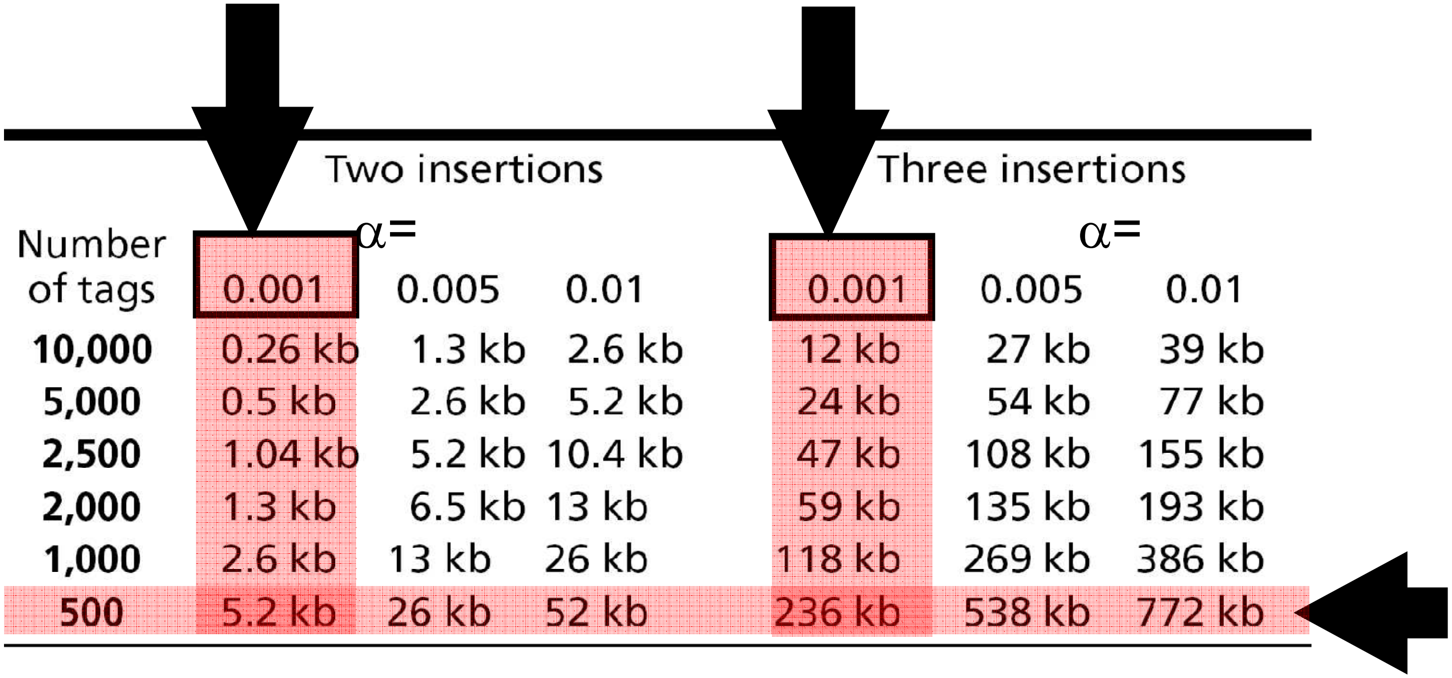


	Two insertions			Three insertions		
	α=				α=	
Number of tags	0.001	0.005	0.01	0.001	0.005	0.01
10,000	0.26 kb	1.3 kb	2.6 kb	12 kb	27 kb	39 kb
5,000	0.5 kb	2.6 kb	5.2 kb	24 kb	54 kb	77 kb
2,500	1.04 kb	5.2 kb	10.4 kb	47 kb	108 kb	155 kb
2,000	1.3 kb	6.5 kb	13 kb	59 kb	135 kb	193 kb
1,000	2.6 kb	13 kb	26 kb	118 kb	269 kb	386 kb
500	5.2 kb	26 kb	52 kb	236 kb	538 kb	772 kb

b: screen size

Previous approaches (2)

Choose significance level: α

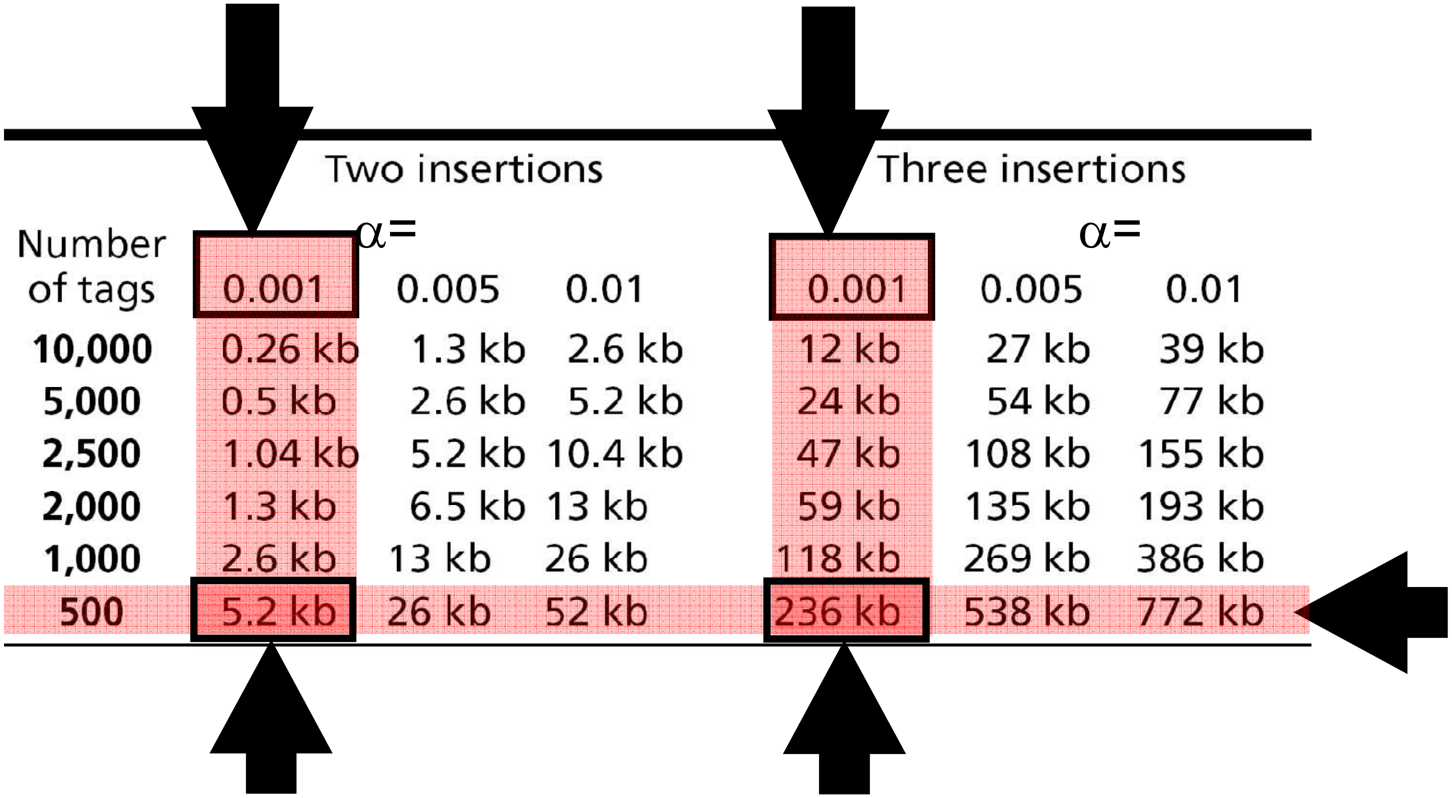


	Two insertions			Three insertions		
	α=			α=		
Number of tags	0.001	0.005	0.01	0.001	0.005	0.01
10,000	0.26 kb	1.3 kb	2.6 kb	12 kb	27 kb	39 kb
5,000	0.5 kb	2.6 kb	5.2 kb	24 kb	54 kb	77 kb
2,500	1.04 kb	5.2 kb	10.4 kb	47 kb	108 kb	155 kb
2,000	1.3 kb	6.5 kb	13 kb	59 kb	135 kb	193 kb
1,000	2.6 kb	13 kb	26 kb	118 kb	269 kb	386 kb
500	5.2 kb	26 kb	52 kb	236 kb	538 kb	772 kb

b: screen size

Previous approaches (3)

Choose significance level: α



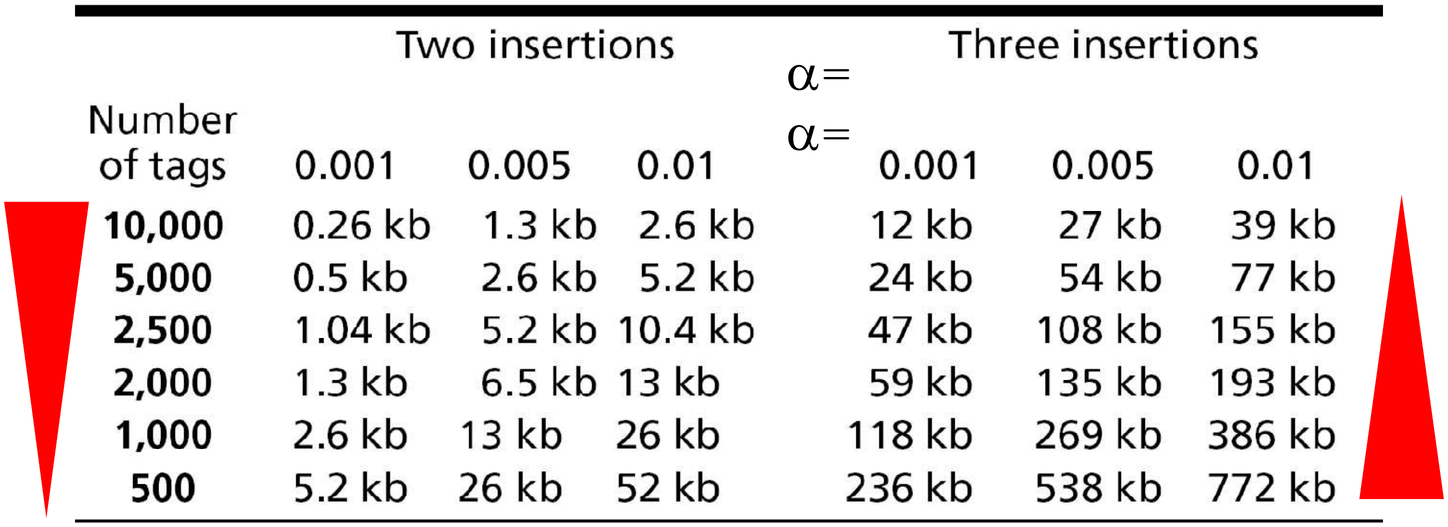
Number of tags	Two insertions $\alpha=$ 0.001	0.005	0.01	Three insertions $\alpha=$ 0.001	0.005	0.01
10,000	0.26 kb	1.3 kb	2.6 kb	12 kb	27 kb	39 kb
5,000	0.5 kb	2.6 kb	5.2 kb	24 kb	54 kb	77 kb
2,500	1.04 kb	5.2 kb	10.4 kb	47 kb	108 kb	155 kb
2,000	1.3 kb	6.5 kb	13 kb	59 kb	135 kb	193 kb
1,000	2.6 kb	13 kb	26 kb	118 kb	269 kb	386 kb
500	5.2 kb	26 kb	52 kb	236 kb	538 kb	772 kb

b: screen size

W: genomic window

Previous approaches (3)

- Large datasets (large *b*) → more FPs
- To reduce FPs, reduce window size, *W*
- Undesirable error control, window size is a *biological* parameter
- Desirable: decouple error control and scale



Number of tags	Two insertions			α=	Three insertions		
	0.001	0.005	0.01	α=	0.001	0.005	0.01
10,000	0.26 kb	1.3 kb	2.6 kb		12 kb	27 kb	39 kb
5,000	0.5 kb	2.6 kb	5.2 kb		24 kb	54 kb	77 kb
2,500	1.04 kb	5.2 kb	10.4 kb		47 kb	108 kb	155 kb
2,000	1.3 kb	6.5 kb	13 kb		59 kb	135 kb	193 kb
1,000	2.6 kb	13 kb	26 kb		118 kb	269 kb	386 kb
500	5.2 kb	26 kb	52 kb		236 kb	538 kb	772 kb

Goal

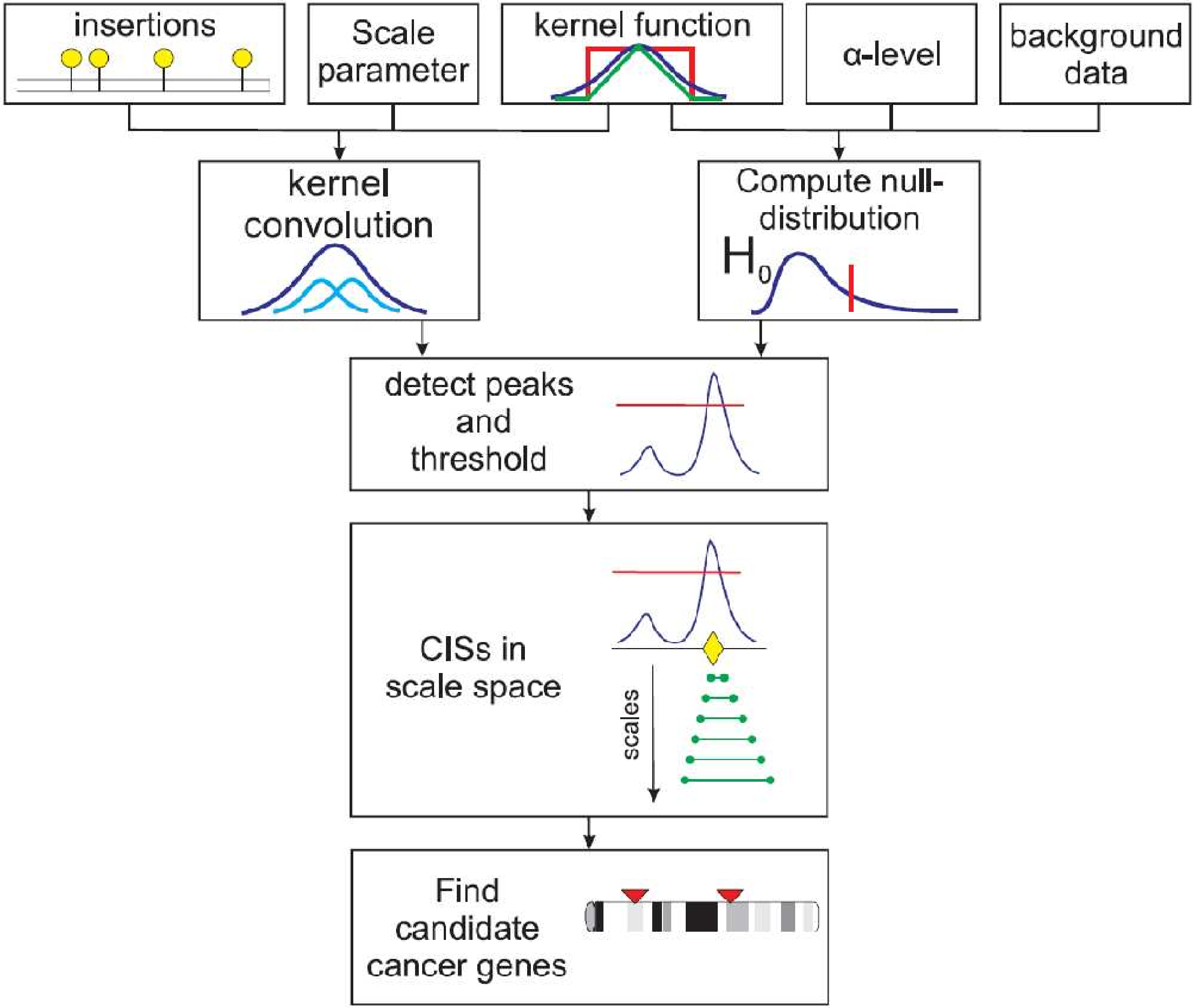
Develop a framework to analyze insertional mutagenesis data which:

1. Evaluates significance at any desired scale
2. Keeps control of the error
3. Compensates for background biases

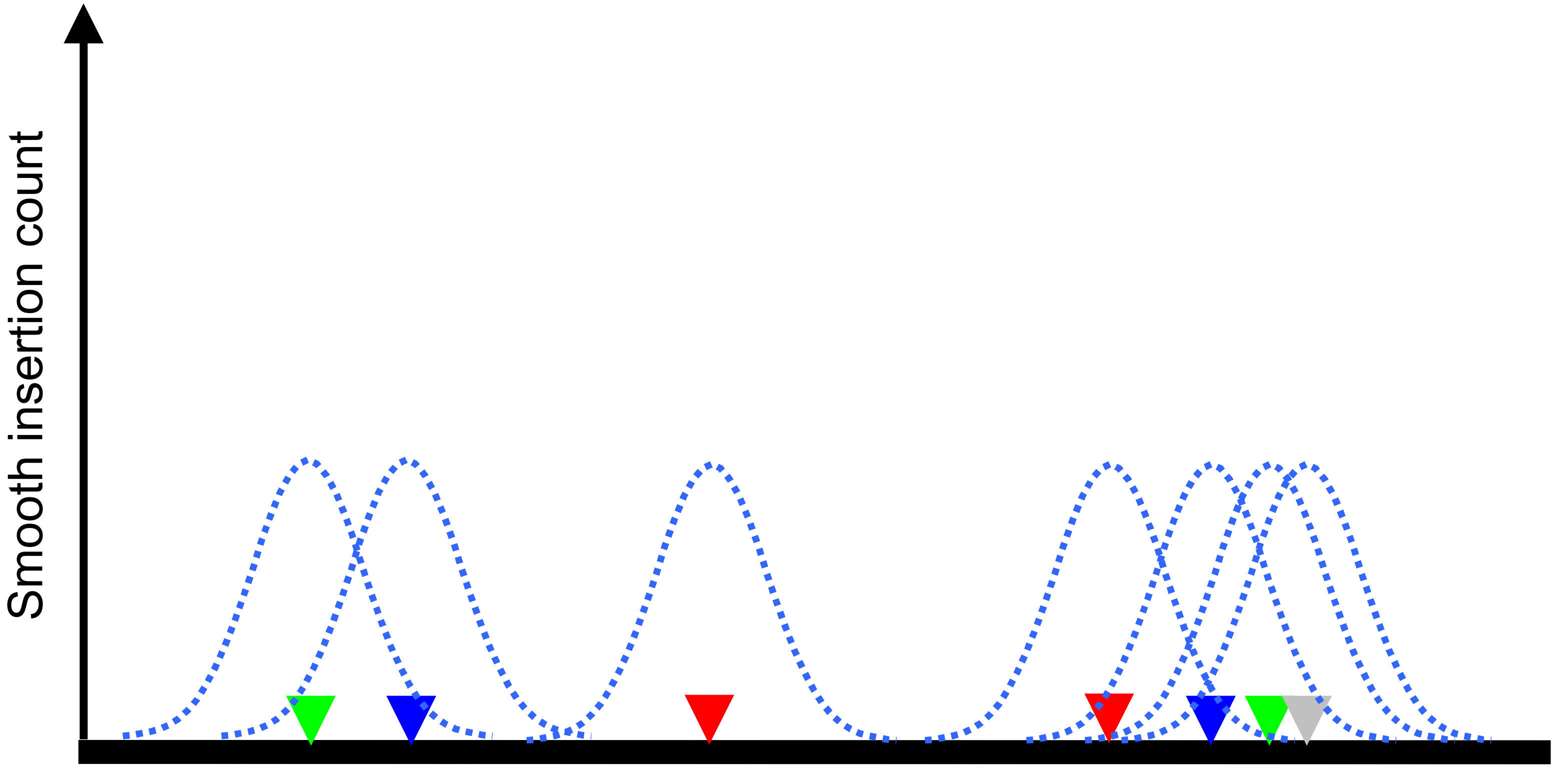
Kernel convolution framework

Main ingredients:

1. Kernel smoothing
   1. smoothed count
   2. alleviates data sparseness
   3. models effect of insertion on neighborhood
2. Permutation scheme to keep the FWE under control
3. Scale space: vary kernel width to vary smoothing
4. Background model


insertions
Scale parameter
kernel function
α-level
background data
kernel convolution
Compute null-distribution
H_0
detect peaks and threshold
CISs in scale space
scales
Find candidate cancer genes

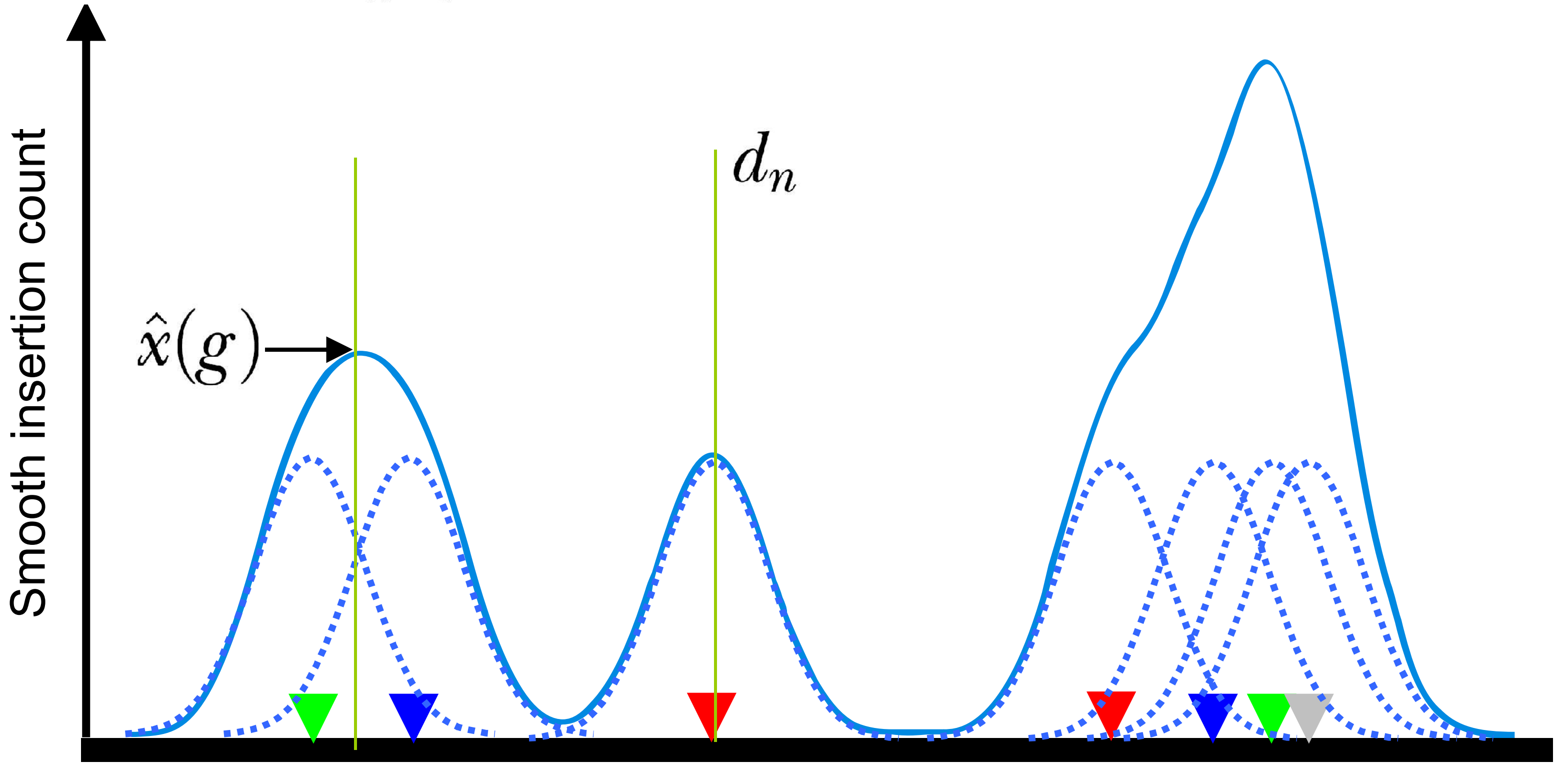
Kernel smoothing (4)



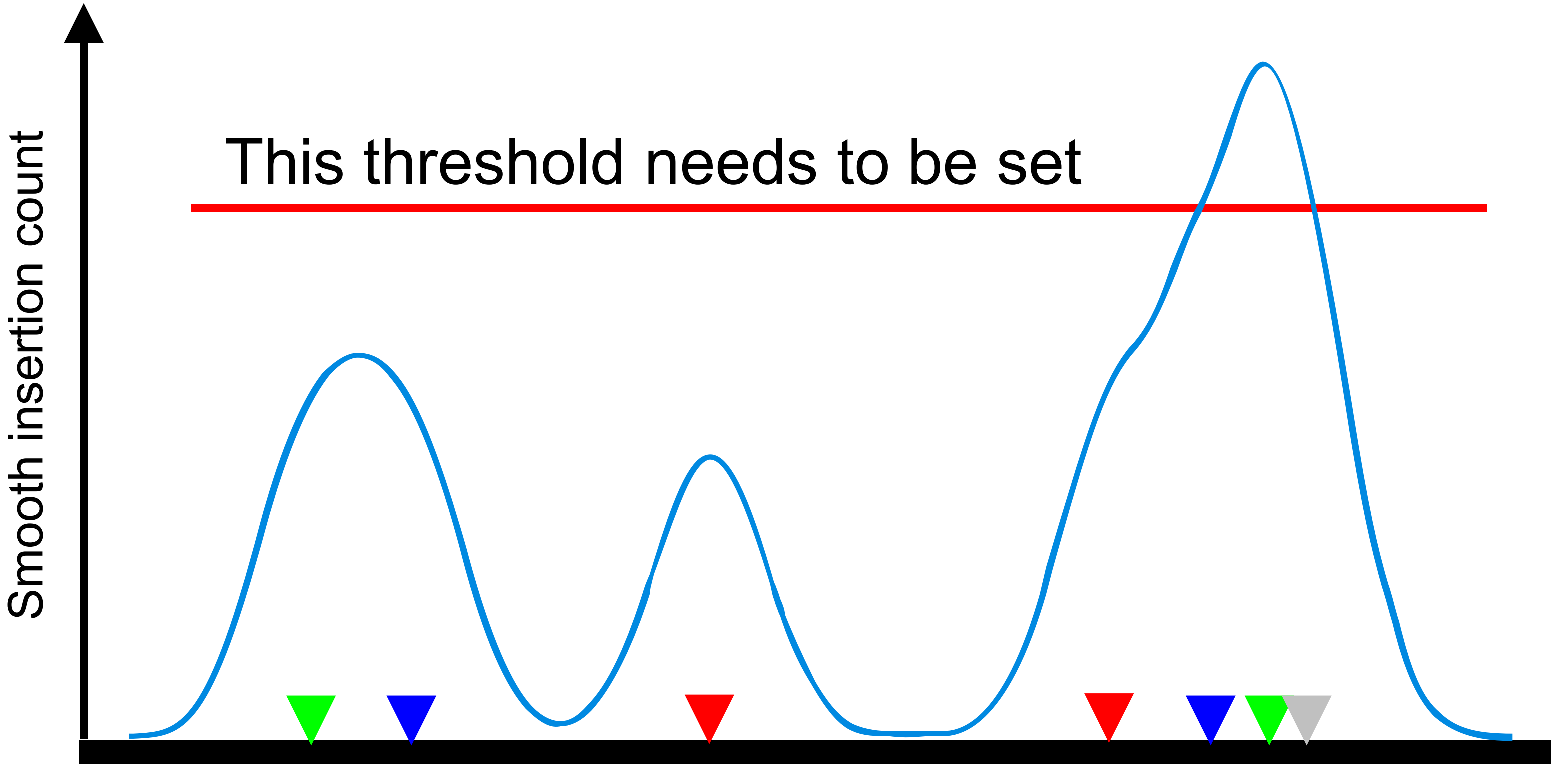
De Ridder et al. 2006, PLoS Comput Biol. 2(12): e166.

Kernel smoothing (4)

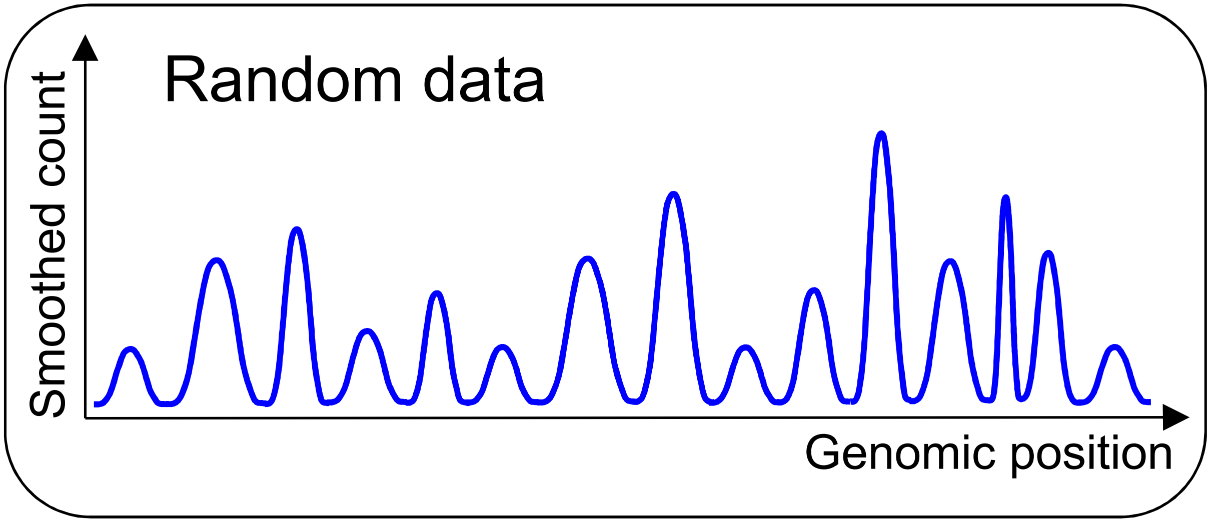
$$\hat{x}(g) = \sum_{n=0}^{N} K(g - d_n) \quad \text{with} \quad g = [0, ..., G]$$



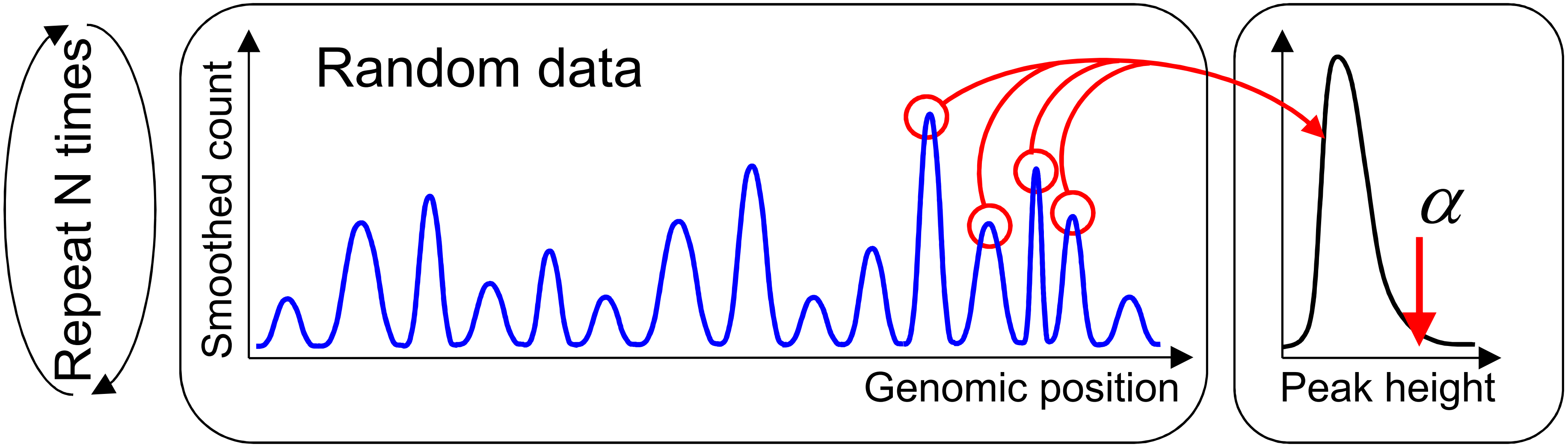
Kernel smoothing (4)



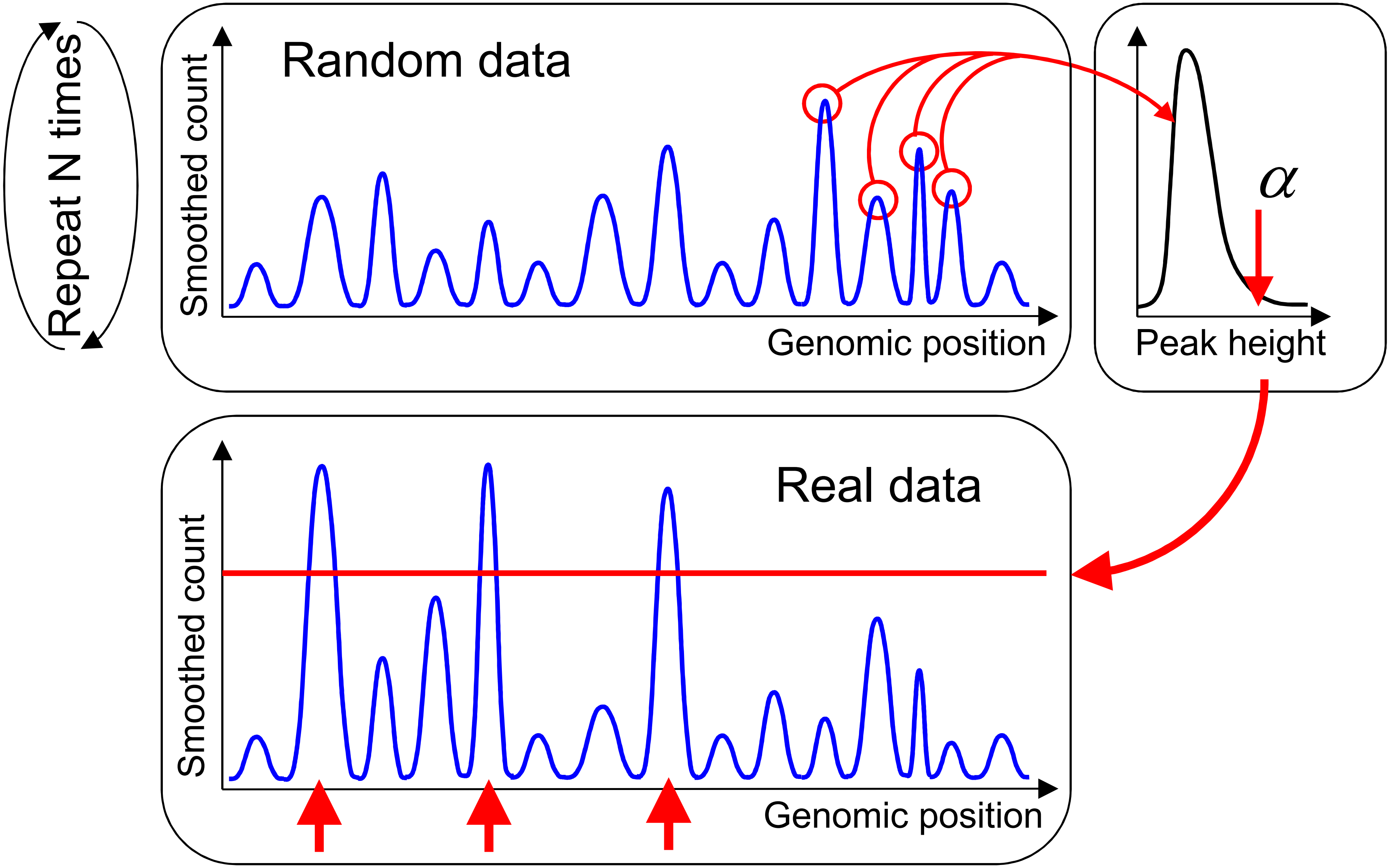
Random permutations: CIS threshold



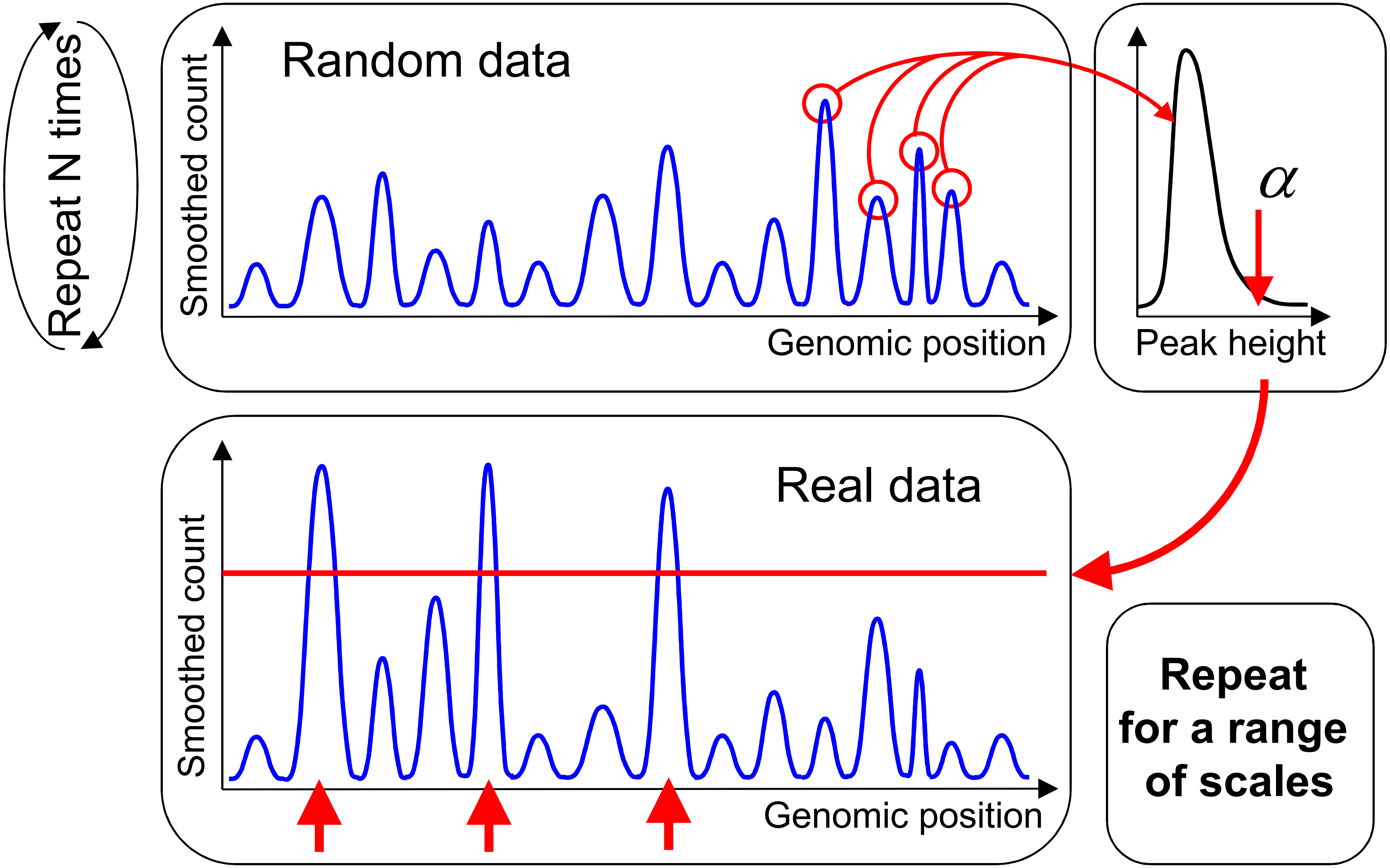
Random permutations: CIS threshold



Random permutations: CIS threshold

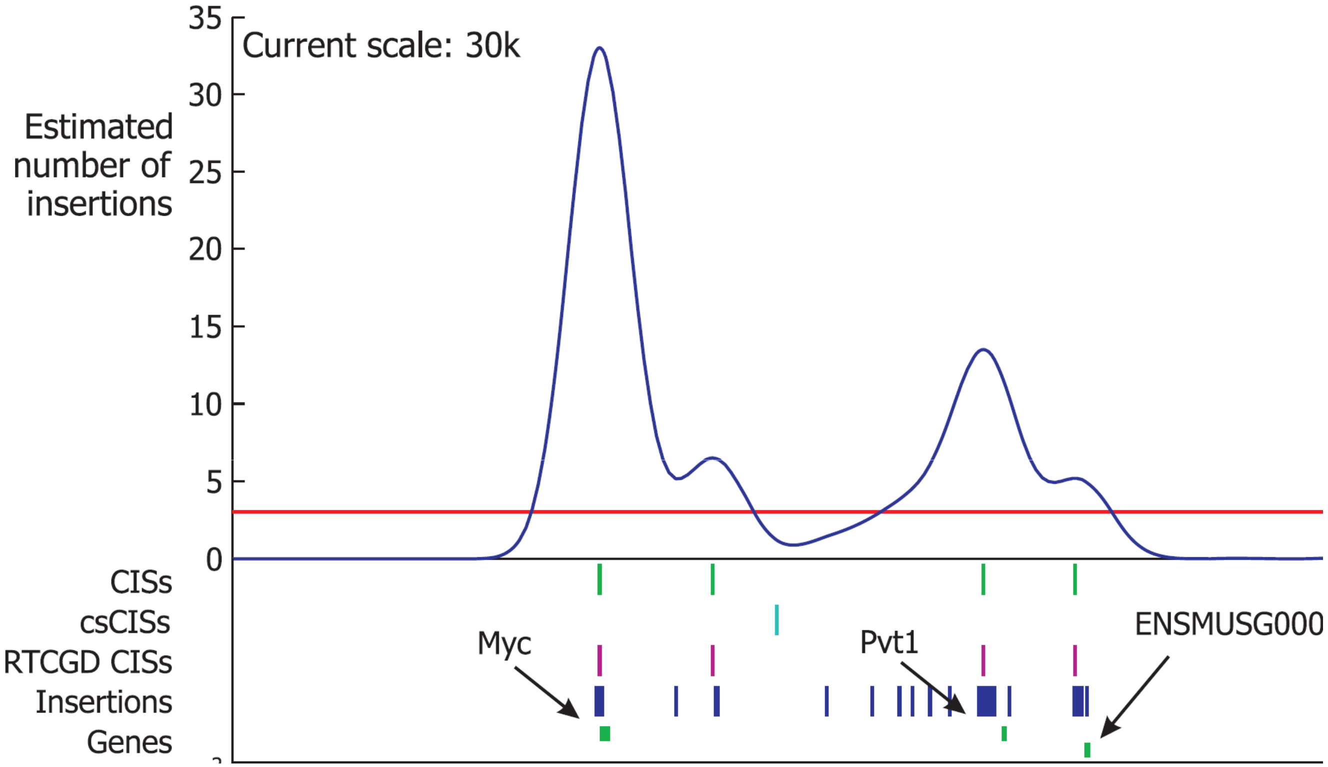


Random permutations: CIS threshold

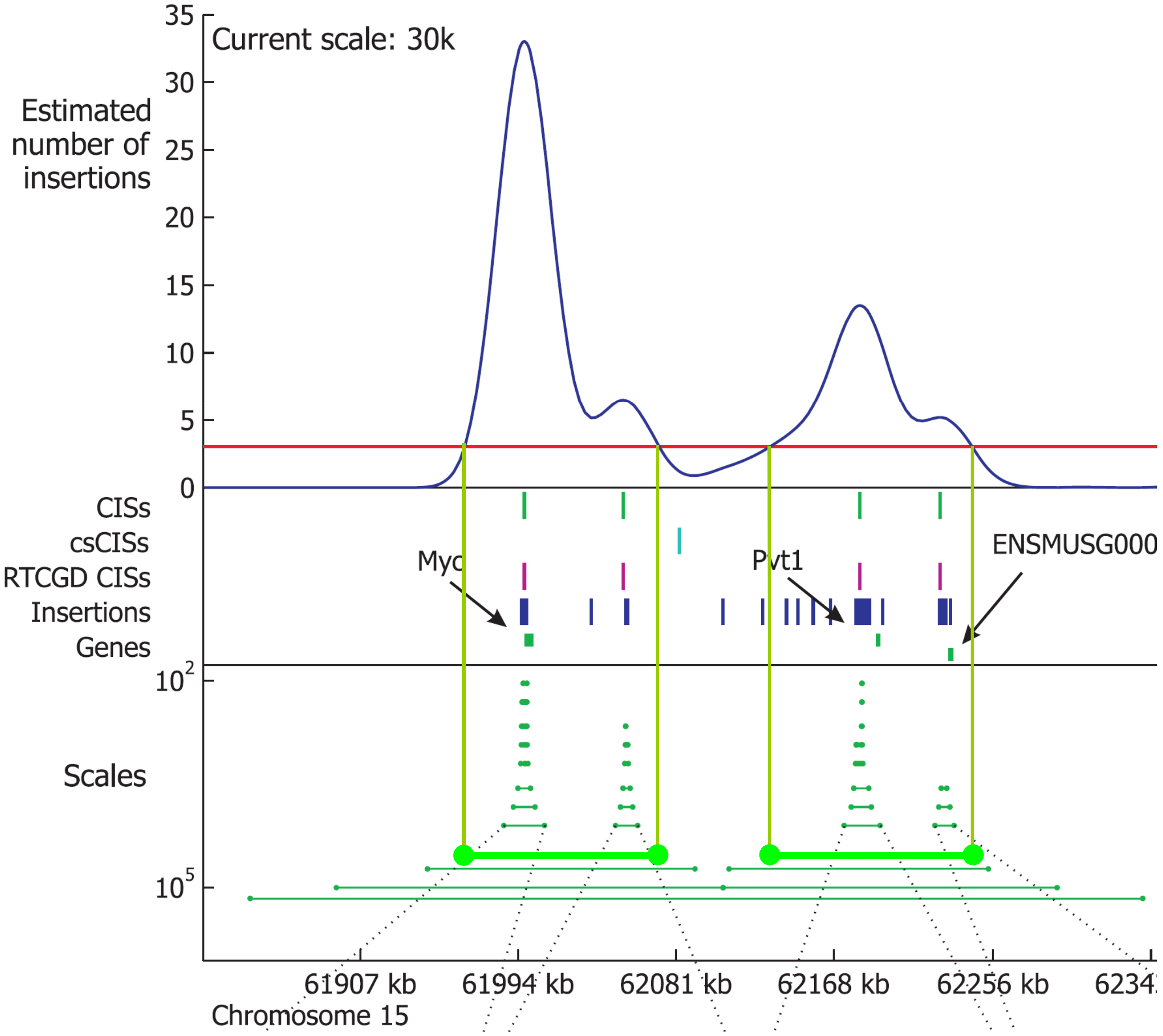


Repeat for a range of scales

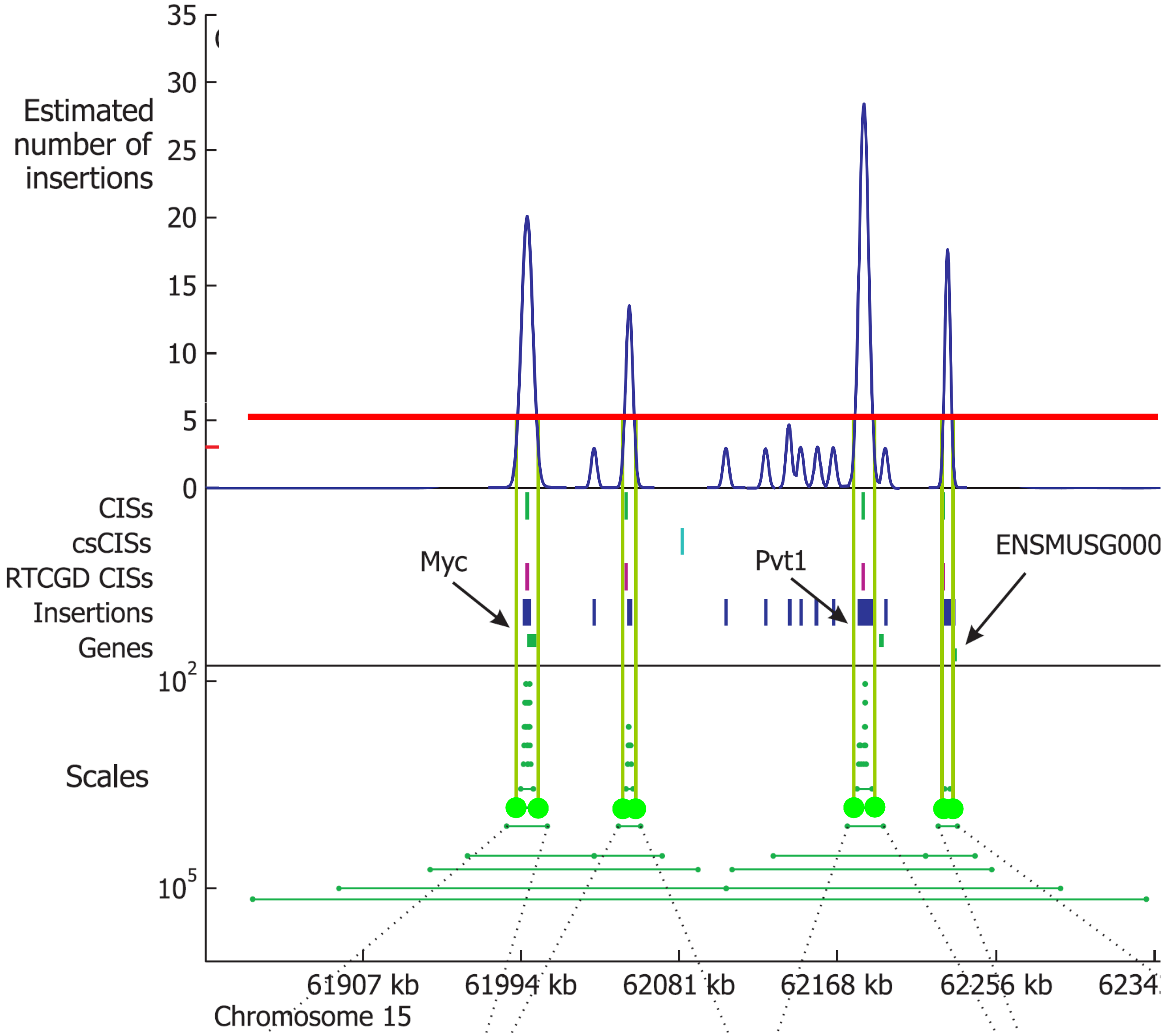
Scale space example (1)



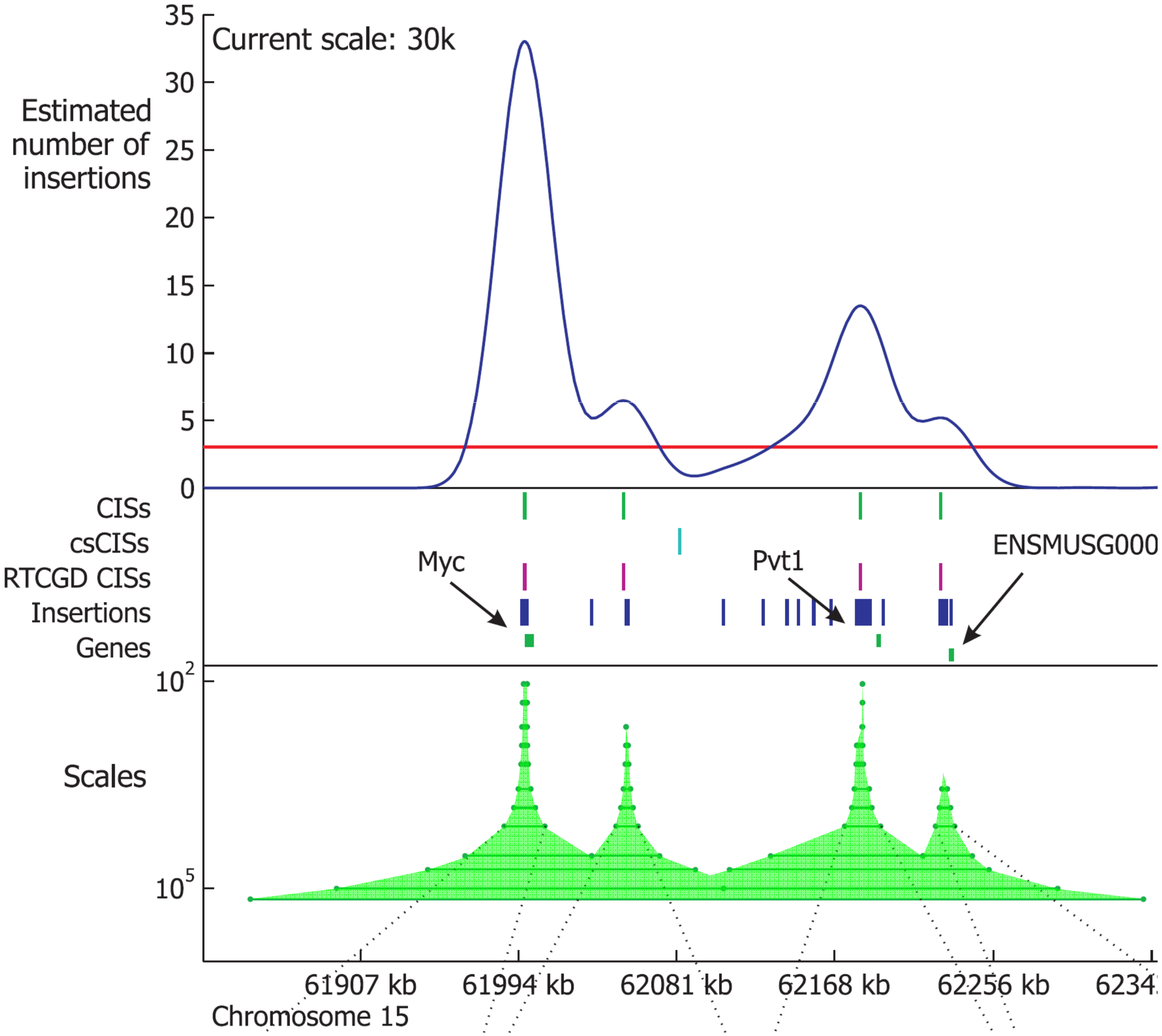
Scale space example (2)



Scale space example (3)

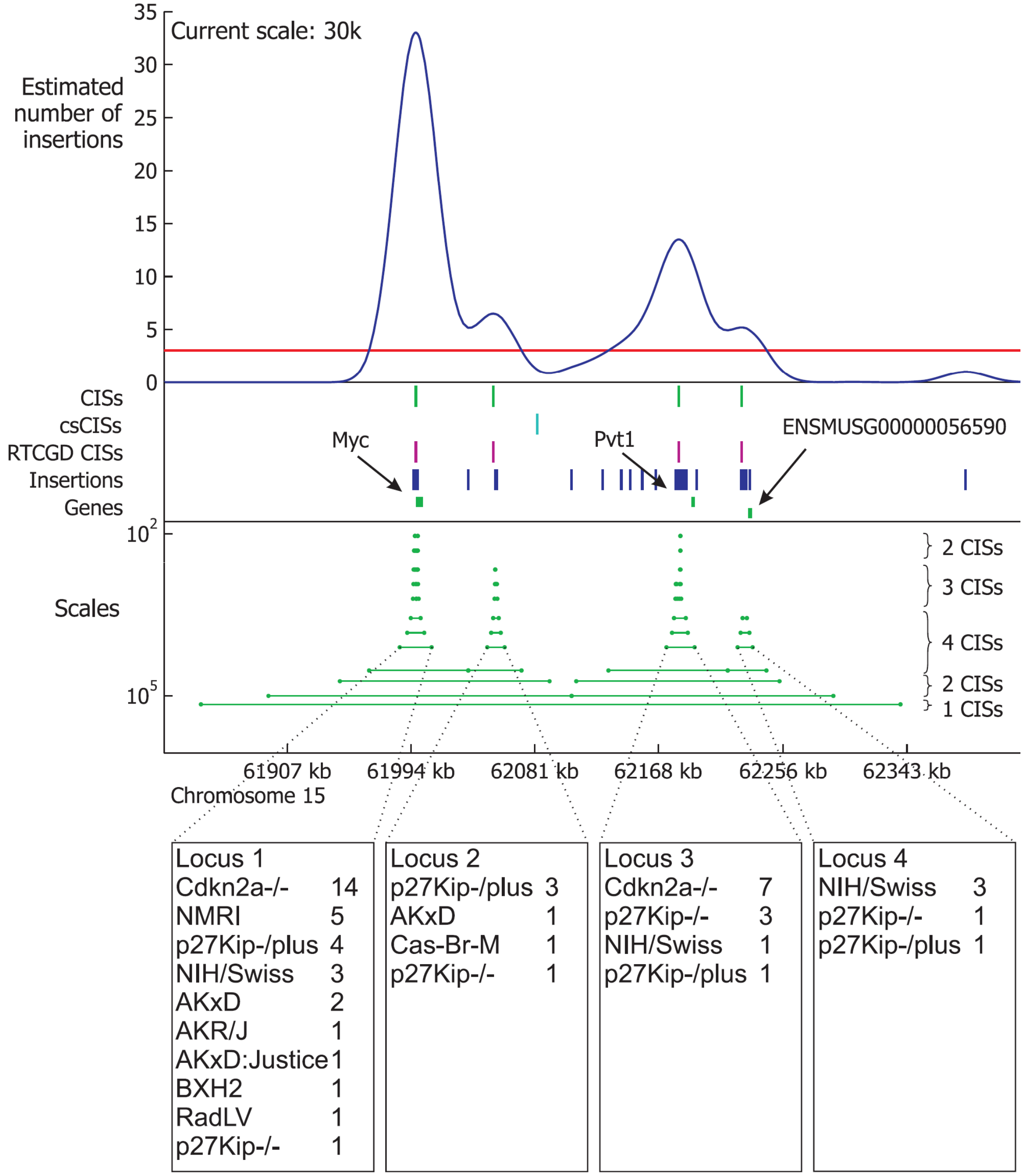


Scale space example (4)



RTCGD Results

- Cdkn2a-/- have a notable bias towards sub-CIS1 and sub-CIS3
- Functionality of sub-CISs



Locus 1	
Cdkn2a-/-	14
NMRI	5
p27Kip-/plus	4
NIH/Swiss	3
AKxD	2
AKR/J	1
AKxD:Justice	1
BXH2	1
RadLV	1
p27Kip-/-	1

Locus 2	
p27Kip-/plus	3
AKxD	1
Cas-Br-M	1
p27Kip-/-	1

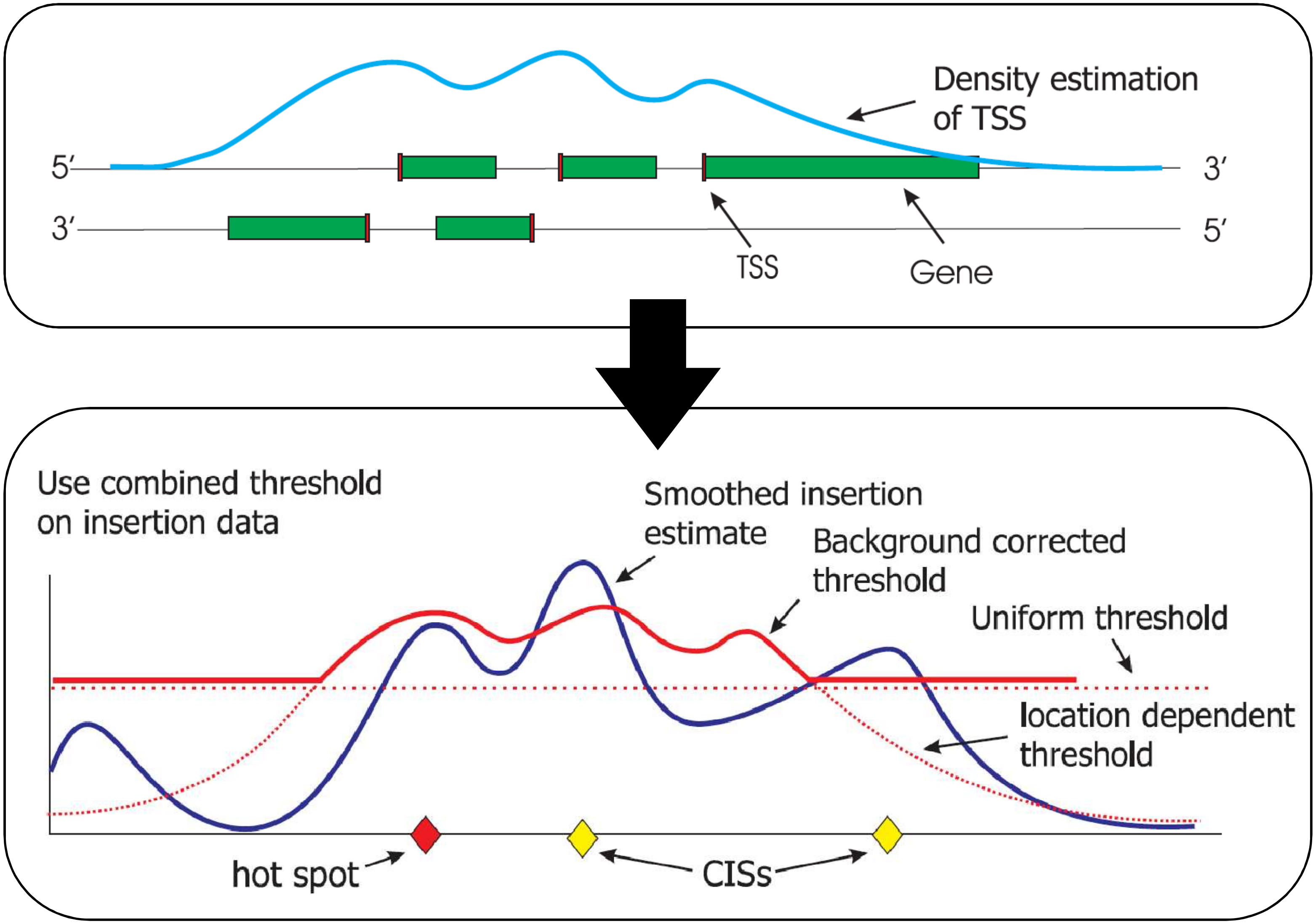
Locus 3	
Cdkn2a-/-	7
p27Kip-/-	3
NIH/Swiss	1
p27Kip-/plus	1

Locus 4	
NIH/Swiss	3
p27Kip-/-	1
p27Kip-/plus	1

Background correction

- MLV favors integration near TSS
- TSS may be a predictor for hot spots.
- Background model:
  - locations of the 5’ ends of ENSEMBLE genes (should be ‘active’ genes)
- There are more (unknown) factors influencing the selective behavior

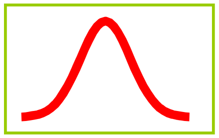
Background correction (2)



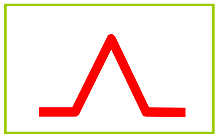
Kernel function

- Many possibilities
- We consider Gaussian, Triangular, Rectangular

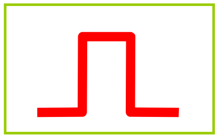
$$\text{Gaussian}: K(z) = e^{-2z^2/h^2}$$



$$\text{Triangular}: K(z) = \begin{cases} -\dfrac{|z|}{\gamma_t h} + 1 \text{ for } |z| < \gamma_t h \\ 0 \text{ otherwise} \end{cases}$$



$$\text{Rectangular}: K(z) = \begin{cases} 1 \text{ for } |z| < \gamma_r h/2 \\ 0 \text{ otherwise} \end{cases}$$



Artificial data

Goal:

- Evaluate kernel functions
- Characterize error properties

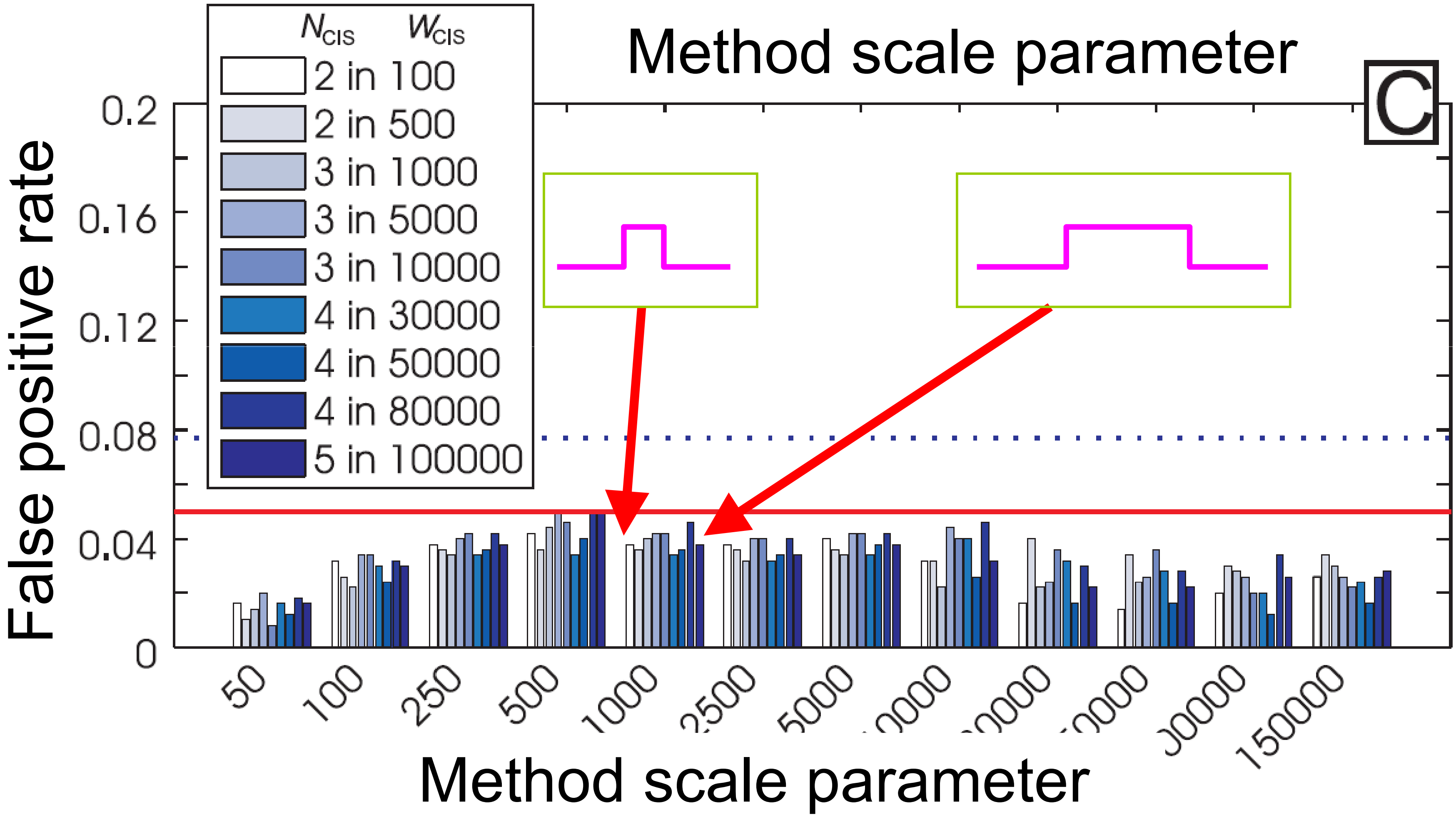
Experiment:

- Uniform background (400 insertions on 2.6 x 10^8 bp genome)
- One CIS locus:
  - uniform distribution
  - W_{CIS} [100bp – 100kbp] wide
  - N_{CIS} insertions in a window
- Insertion frequency slightly higher in CIS locus
- For each setting, 500 artificial datasets were generated.

Definitions

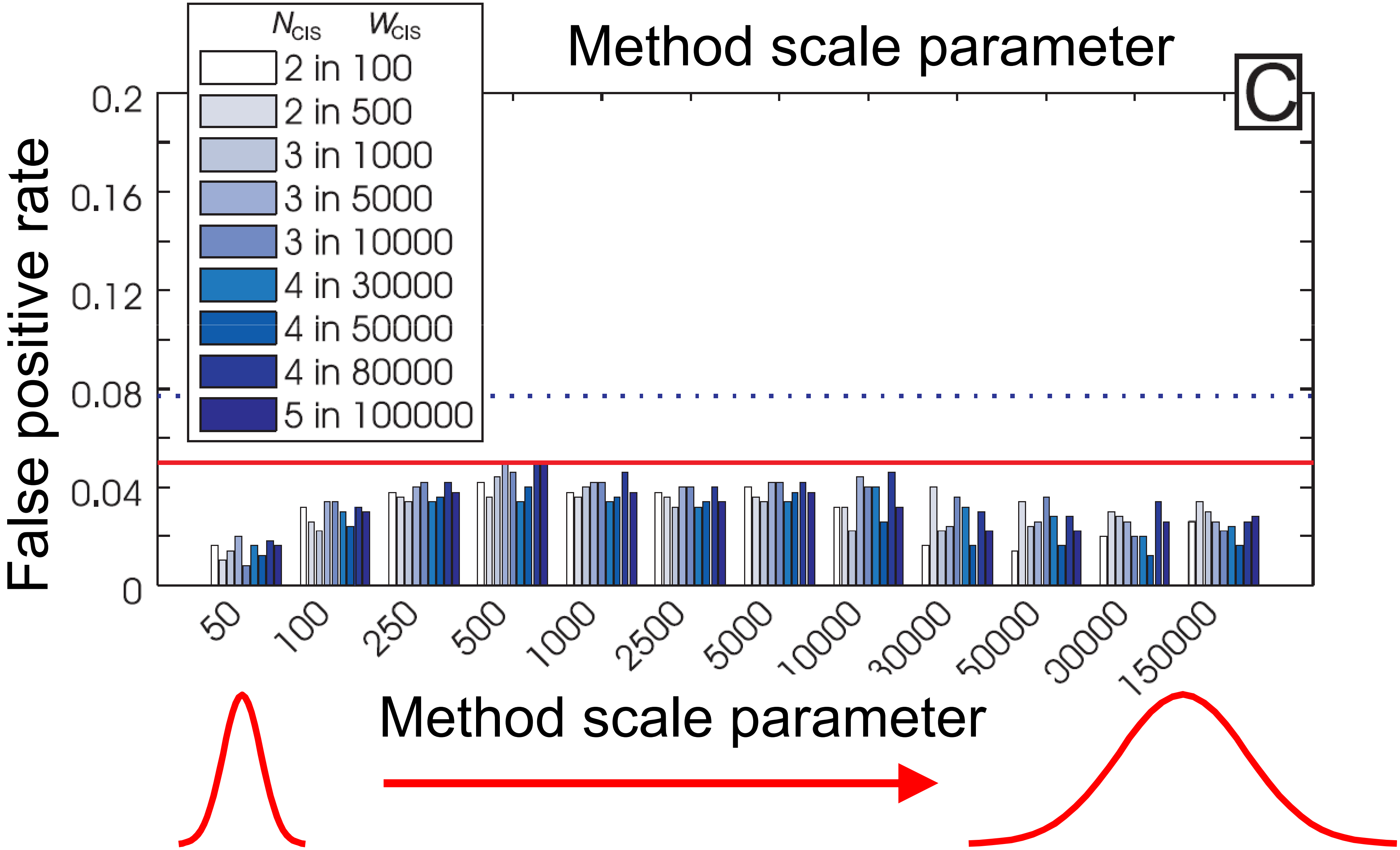
- TP: detection of artificial CIS (overlap of estimated and artificial CIS)
- FP: detection of all other CISs

Results (2)

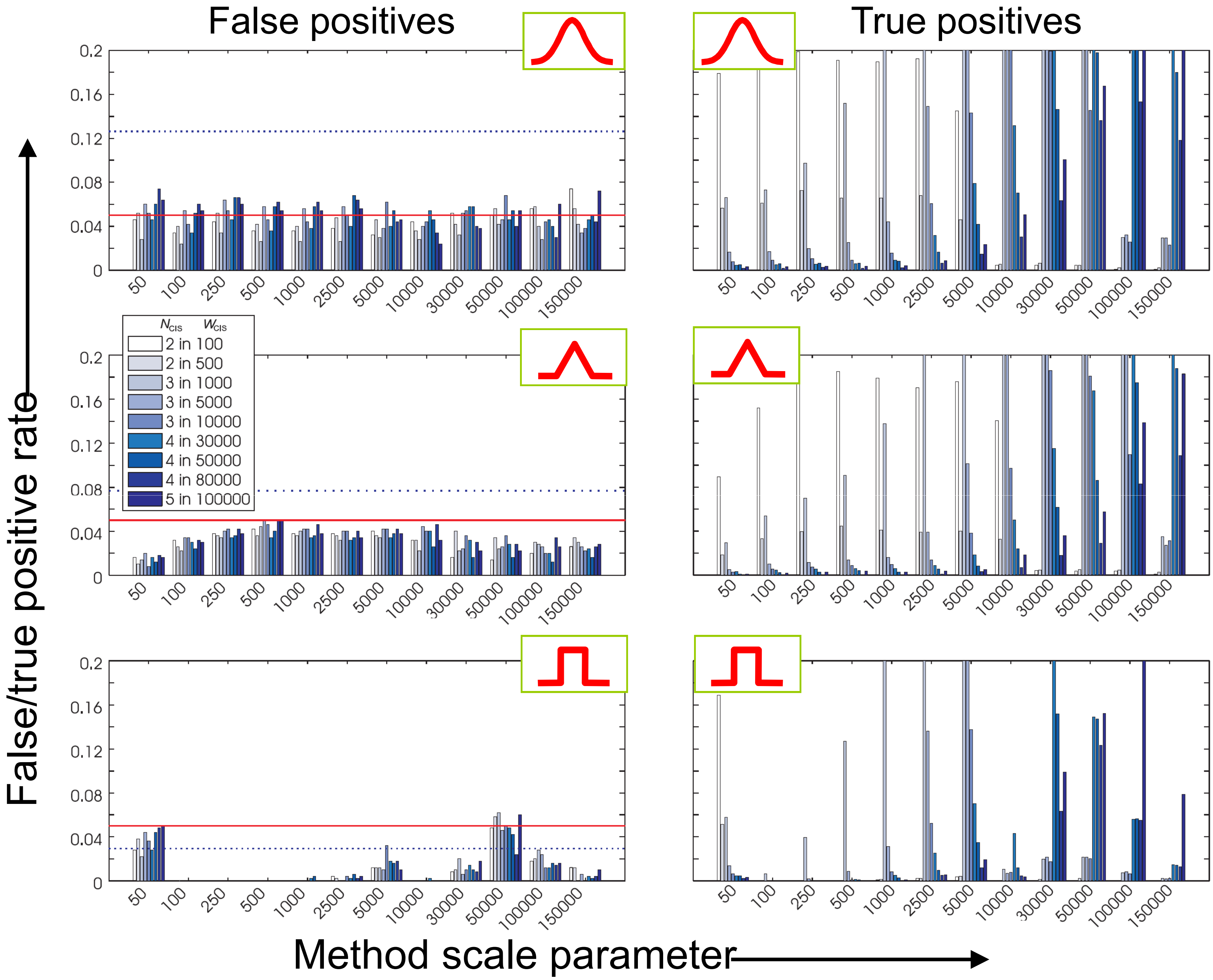


Method scale parameter

Results (1)

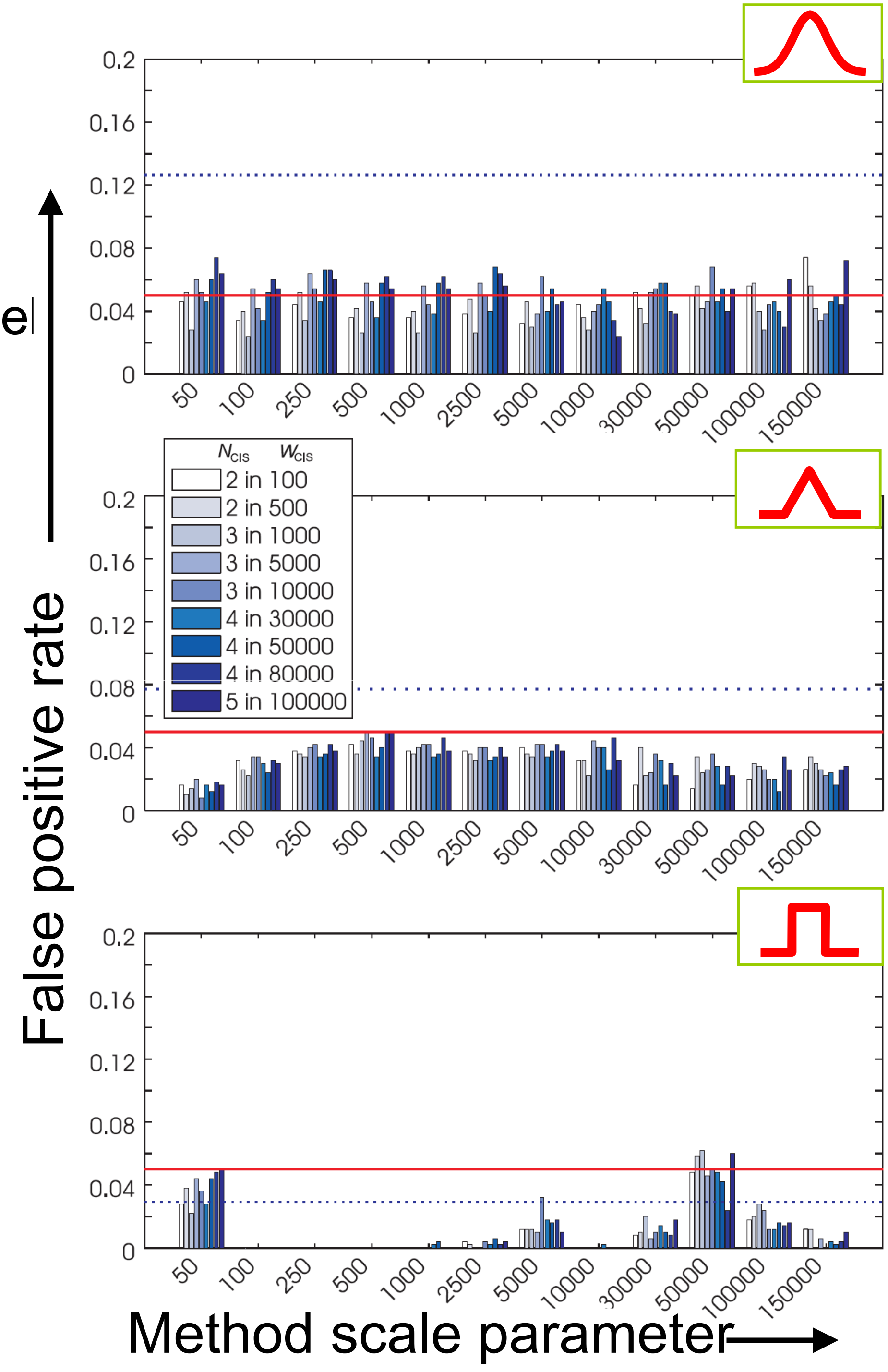


Method scale parameter
C
N_{CIS} W_{CIS}
2 in 100
2 in 500
3 in 1000
3 in 5000
3 in 10000
4 in 30000
4 in 50000
4 in 80000
5 in 100000
False positive rate
0.2
0.16
0.12
0.08
0.04
0
50
100
250
500
1000
2500
5000
10000
30000
50000
00000
150000
Method scale parameter


False positives
True positives
False/true positive rate
Method scale parameter
0.2
0.16
0.12
0.08
0.04
0
50
100
250
500
1000
2500
5000
10000
30000
50000
100000
150000
N_{CIS} W_{CIS}
2 in 100
2 in 500
3 in 1000
3 in 5000
3 in 10000
4 in 30000
4 in 50000
4 in 80000
5 in 100000

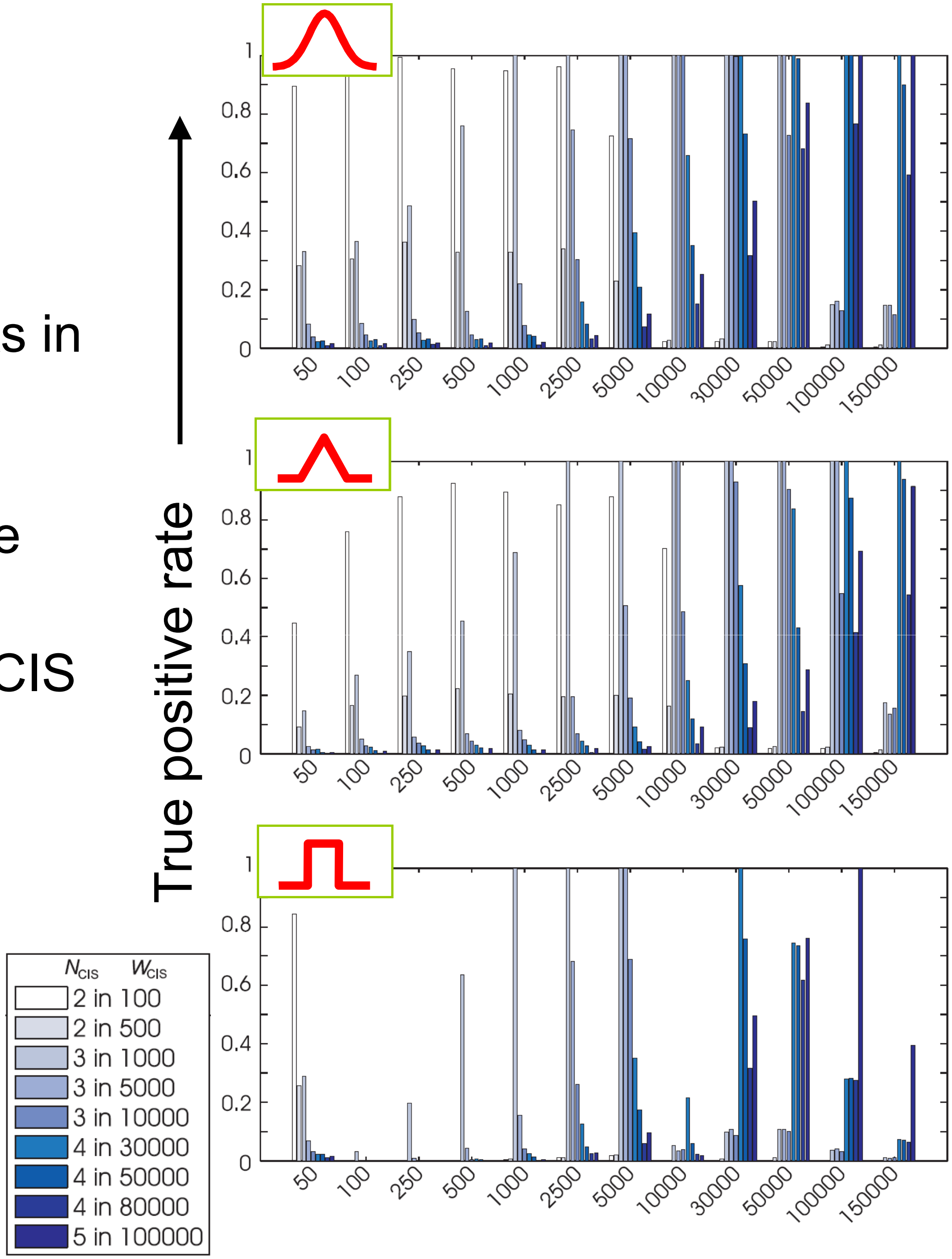
Results (4)

- All kernels control error at 5%-level for all scales
- The GKC controls at an average of 5%
- TKC and especially the RKC are more conservative
- GKC has constant control across scales
- Discrete nature of RKC causes non-uniform control



Results (5)

- Conservativeness results in lower TPs
- Max TP at scale that matches CIS width ('blue shift')
- Range of scales where CIS is detected is largest for GKC
- Measure of robustness



Summary

- GKC shows
  - Some advantage on positional accuracy
  - consistent error distribution across scales
- Therefore, use the GKC to analyze the data from the RTCGD (Retroviral Tagged Cancer Gene Database)

RTCGD

- Retroviral Tagged Cancer Gene Database
- RTCGD contains 1076 tumors, 4K inserts
- Various genetic backgrounds
- Various methods to define CISs

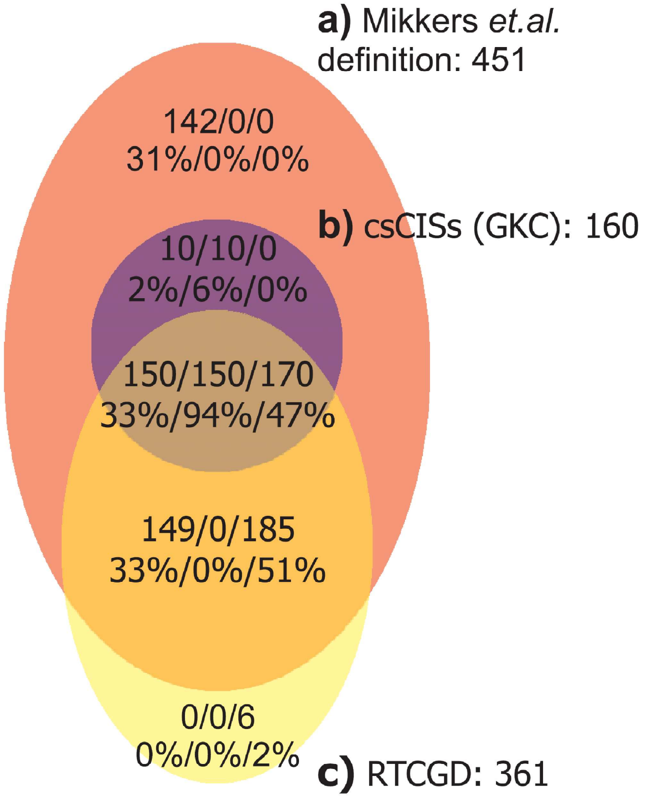
Results

Simulation results:

- Framework suitable for large datasets
- Decoupled error control and window size

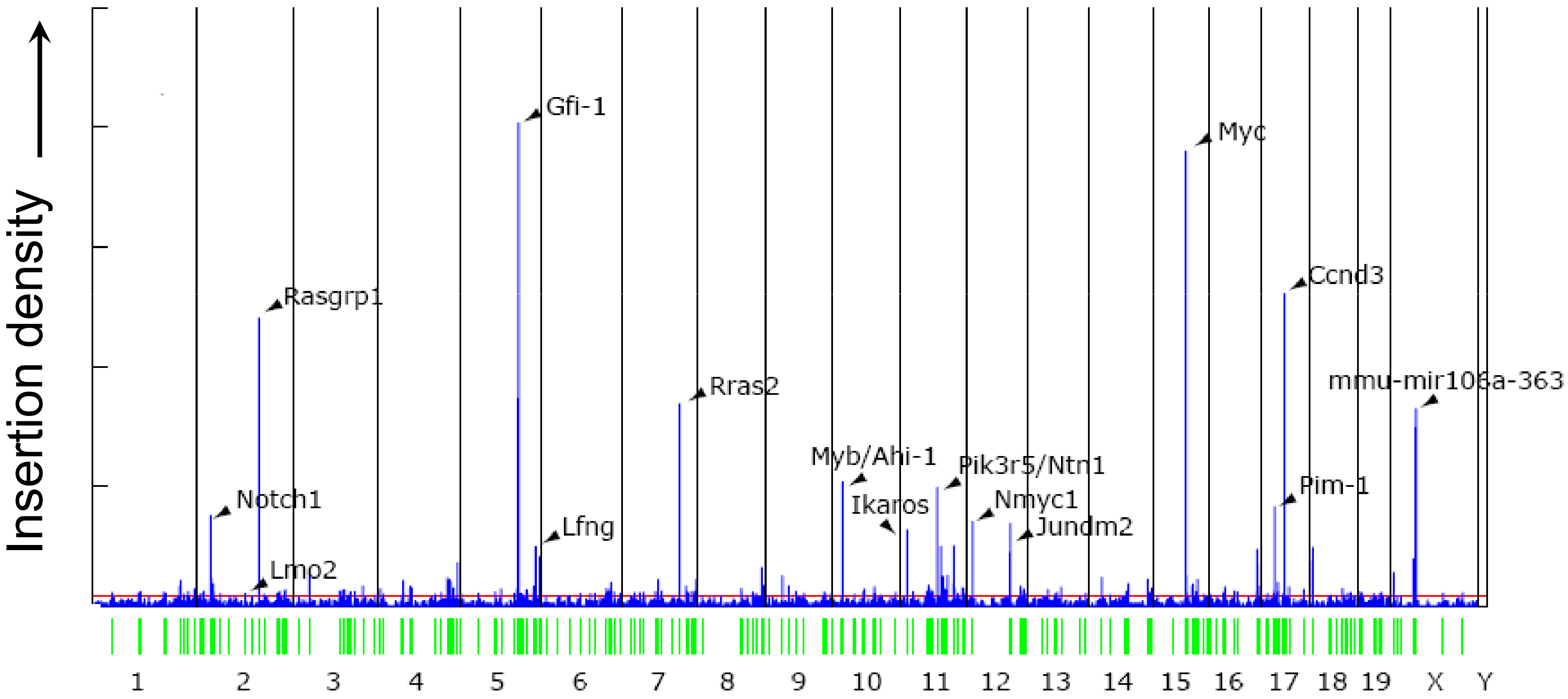
RTCGD results

- We find only 160, but at a guaranteed FWE
- 10 novel CISs over RTCGD.
- 6 of these due to integral analysis.
- Mikkers et al. CIS definition (2 inserts in 26kb) → 451 CISs
- 244 (54%) are estimated to be false detections with MC



Mutapedia results

(500 tumors, ~11K insertions, ~300 CISs, $p < 0.05$)



Uren, Kool *et al.*, Cell. 2008;133(4):727-41.

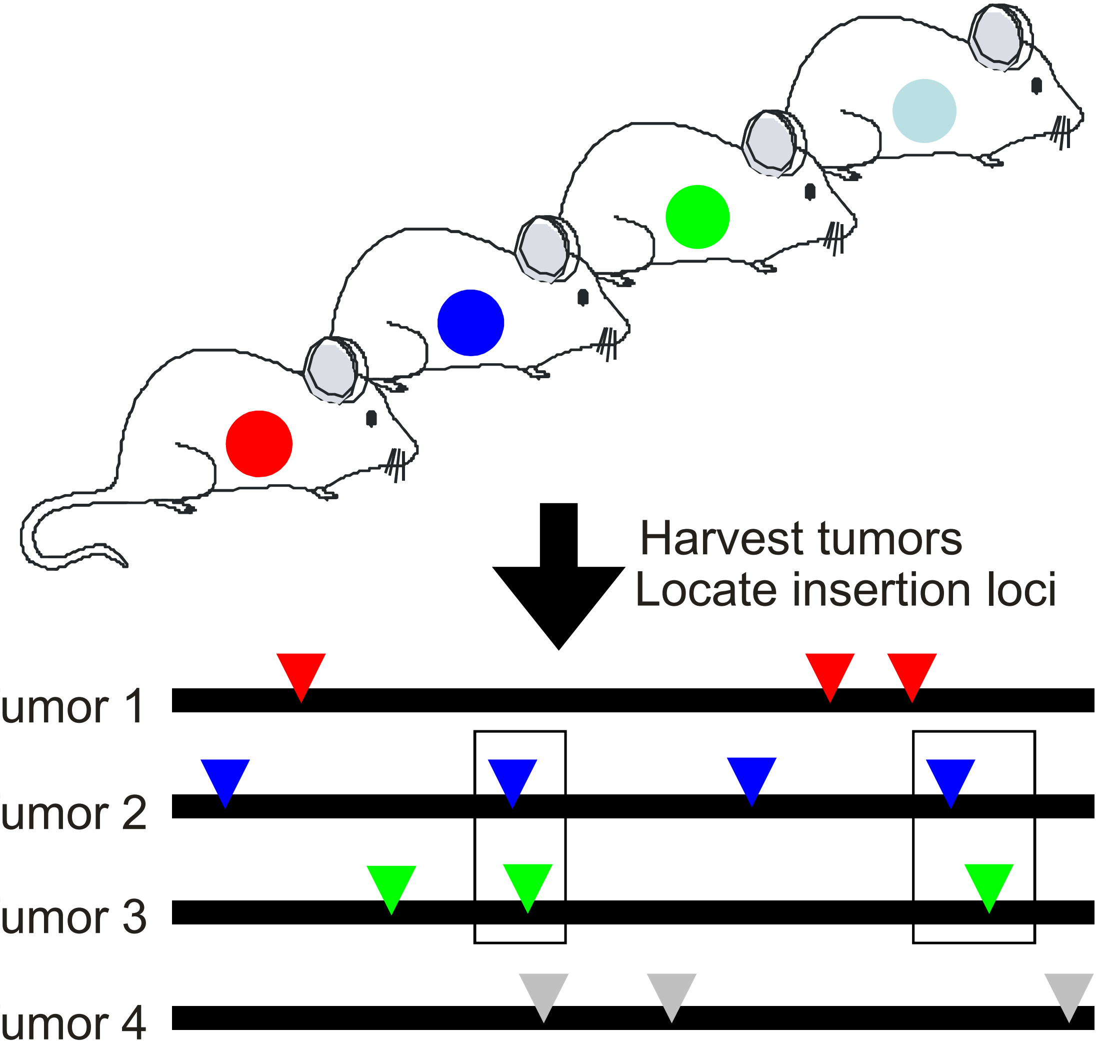
Finding cancer genes and cancer pathways

- Cancer genes:
  - genes individually frequently ‘hit’

- Cancer gene ‘pairs’:
  - pairs of genes frequently ‘hit’ in a specific pattern
  - (a gene and a family of genes frequently hit)
  - Co-operating, mutually exclusive

- Cancer pathways/networks
  - groups of genes frequently ‘hit’ in a specific pattern

Cancer pairs:
Common co-occurrence of insertions



Harvest tumors
Locate insertion loci

Tumor 1

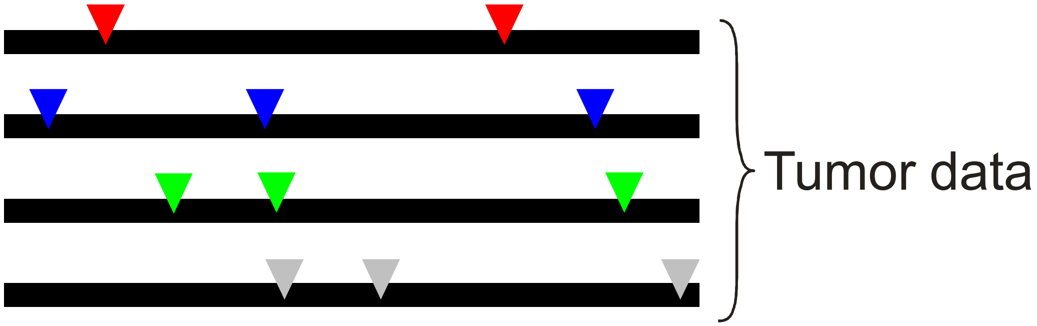
Tumor 2

Tumor 3

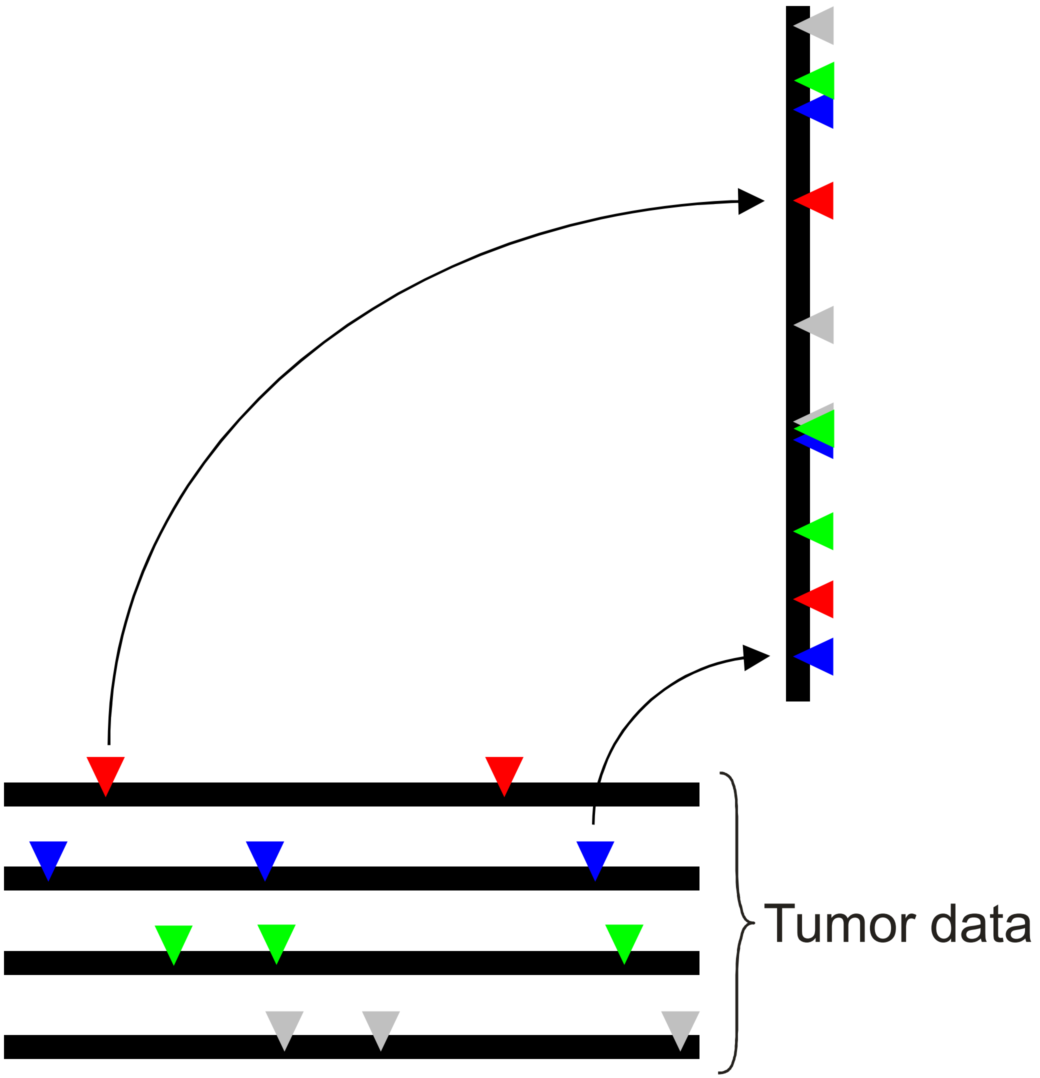
Tumor 4

Co-occurrence?

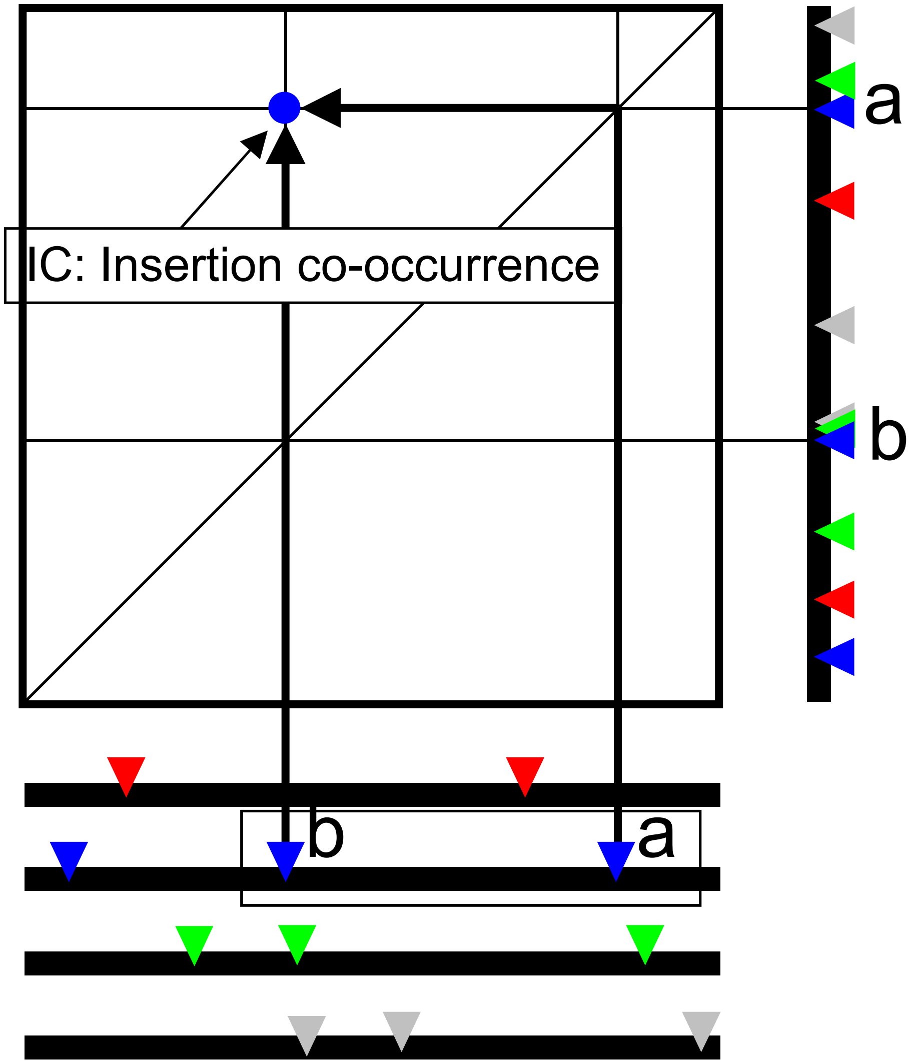
The co-occurrence space



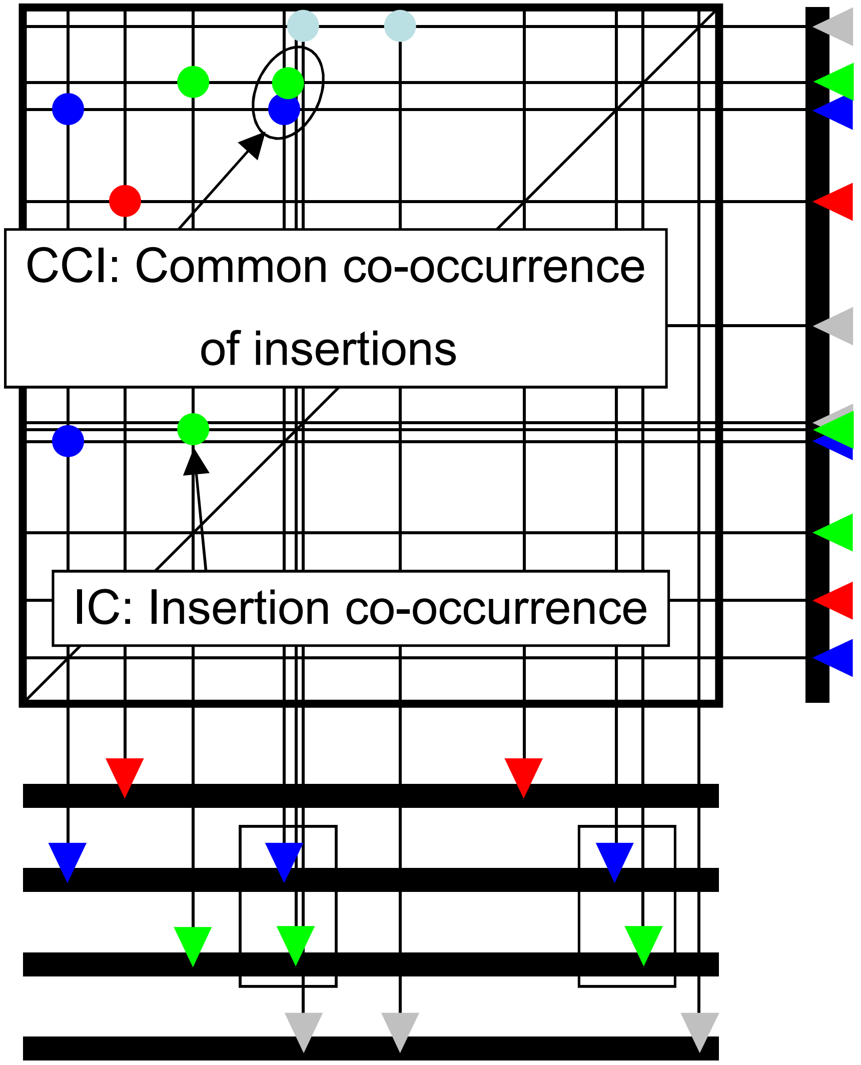
The co-occurrence space



The co-occurrence space



The co-occurrence space

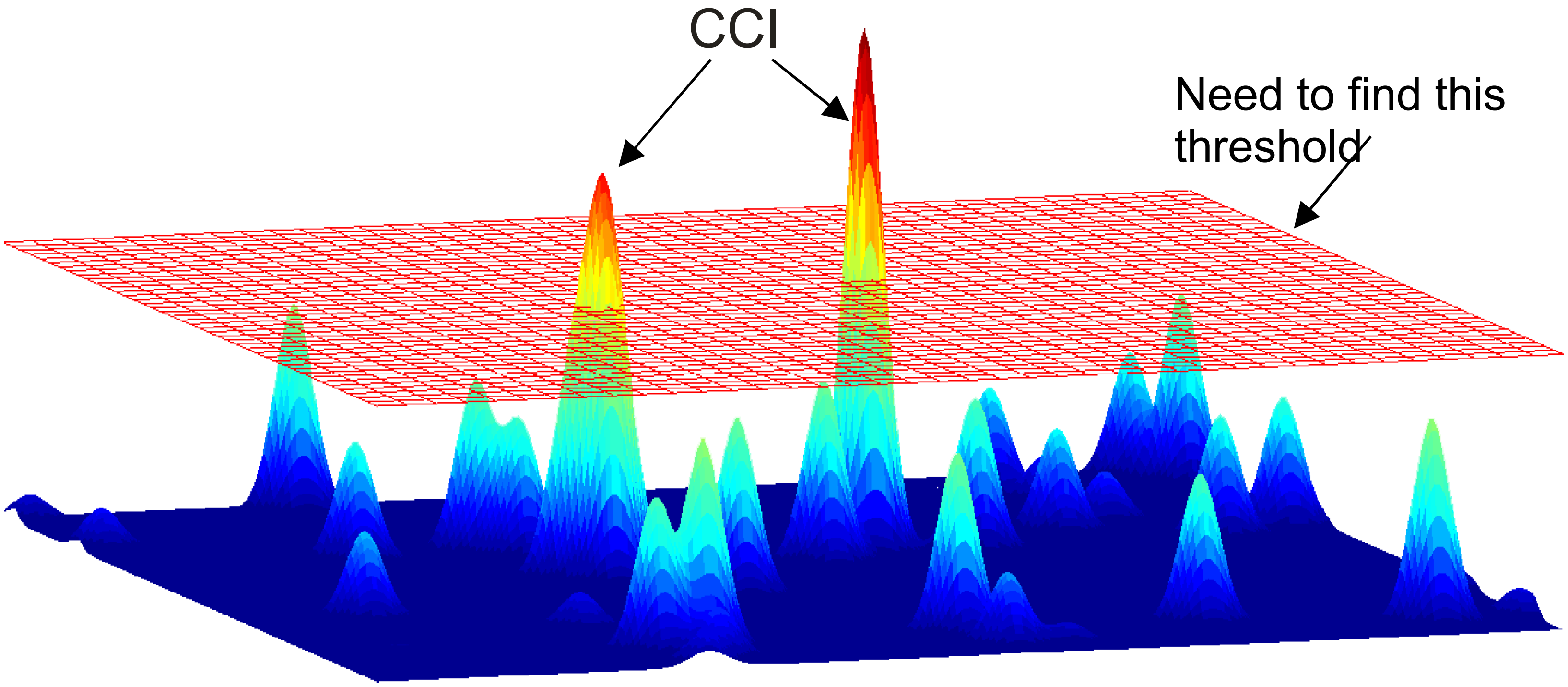


IC: insertion co-occurrence

CCI: Common Co-occurrence of Insertions

CCI: region in the co-occurrence space hit by viral inserts in multiple independent tumors significantly more than expected.

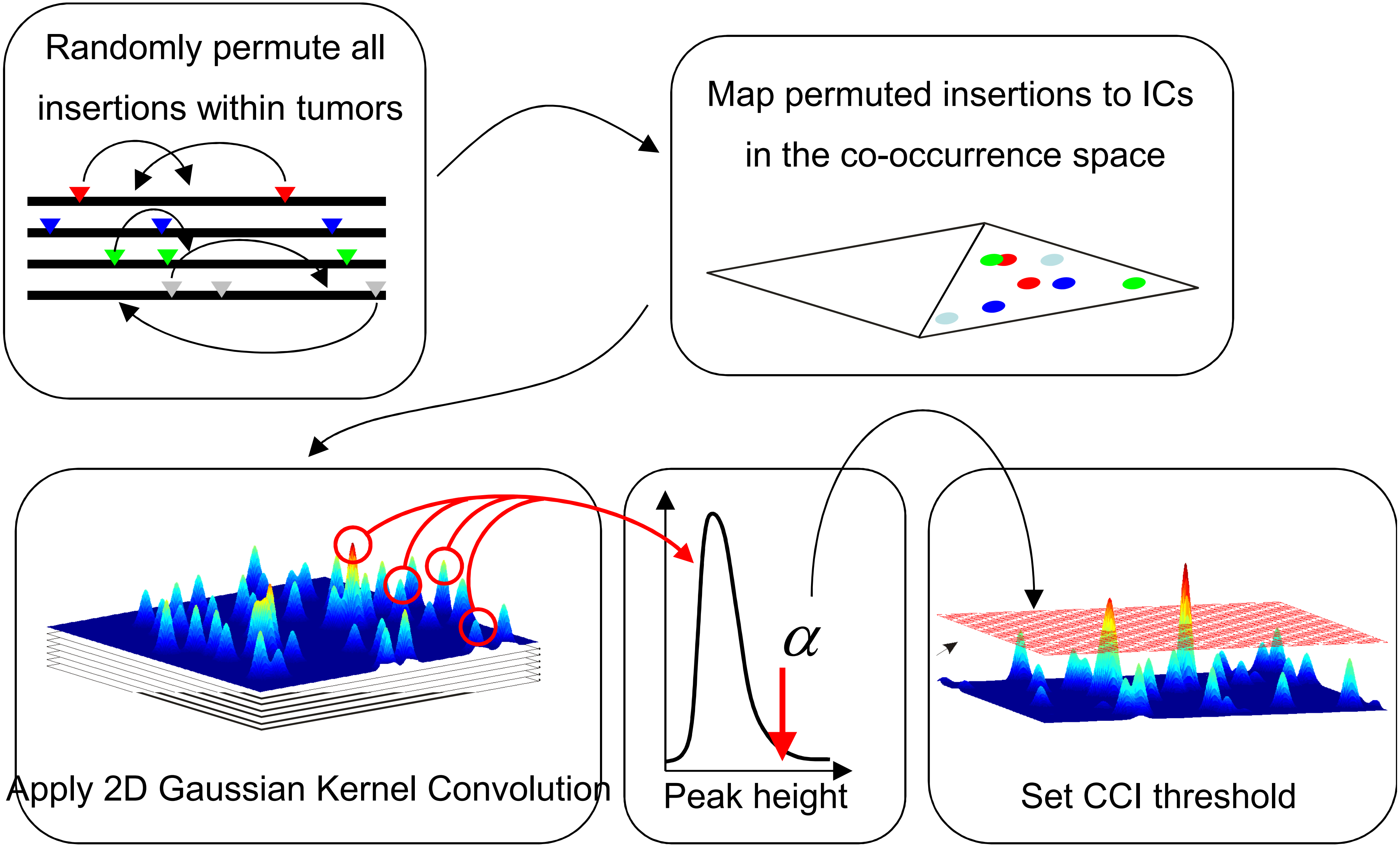
2D Gaussian Kernel Convolution



CCI

Need to find this threshold

Permutation procedure



Randomly permute all insertions within tumors

Map permuted insertions to ICs in the co-occurrence space

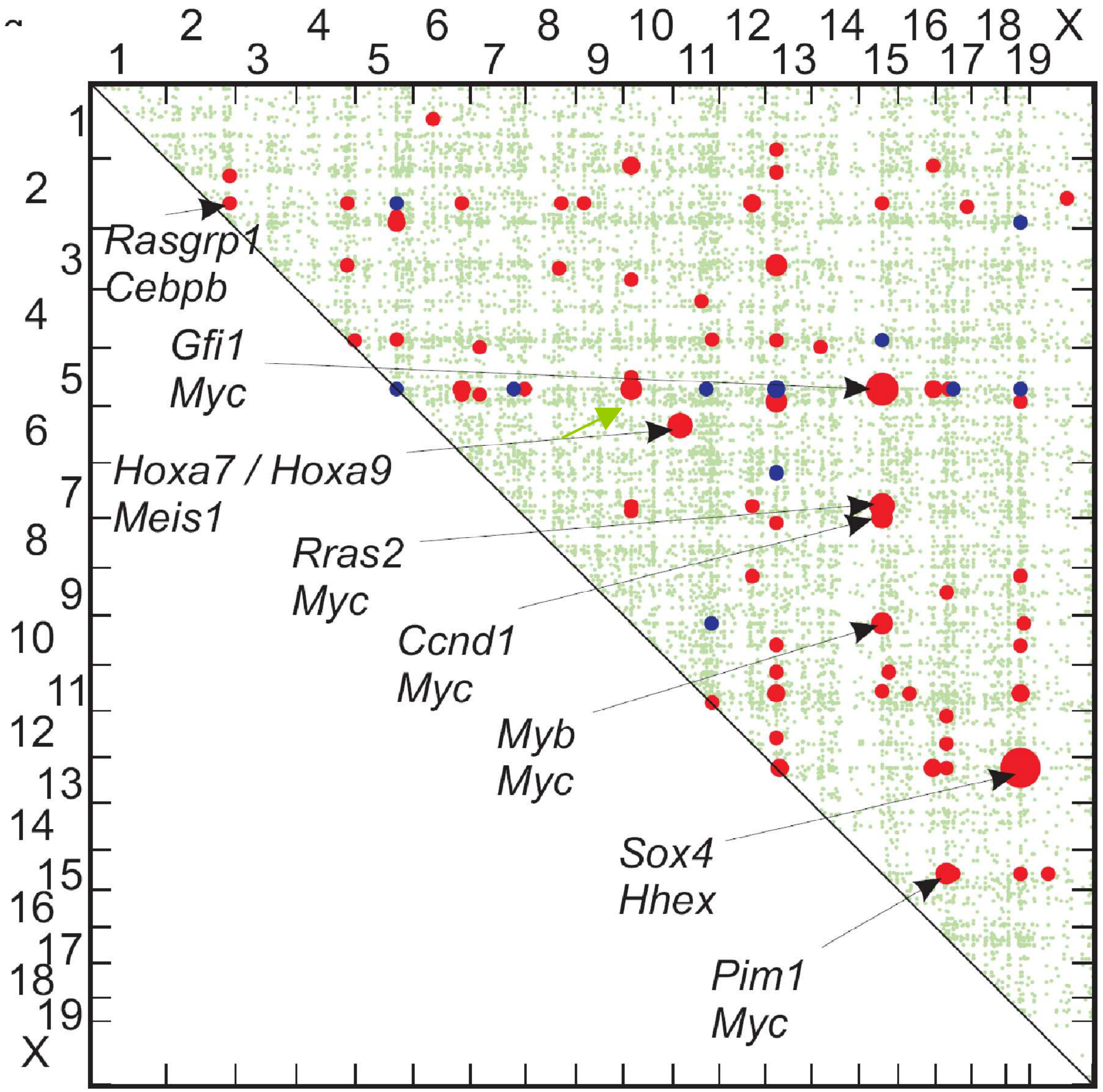
Apply 2D Gaussian Kernel Convolution

Peak height

Set CCI threshold

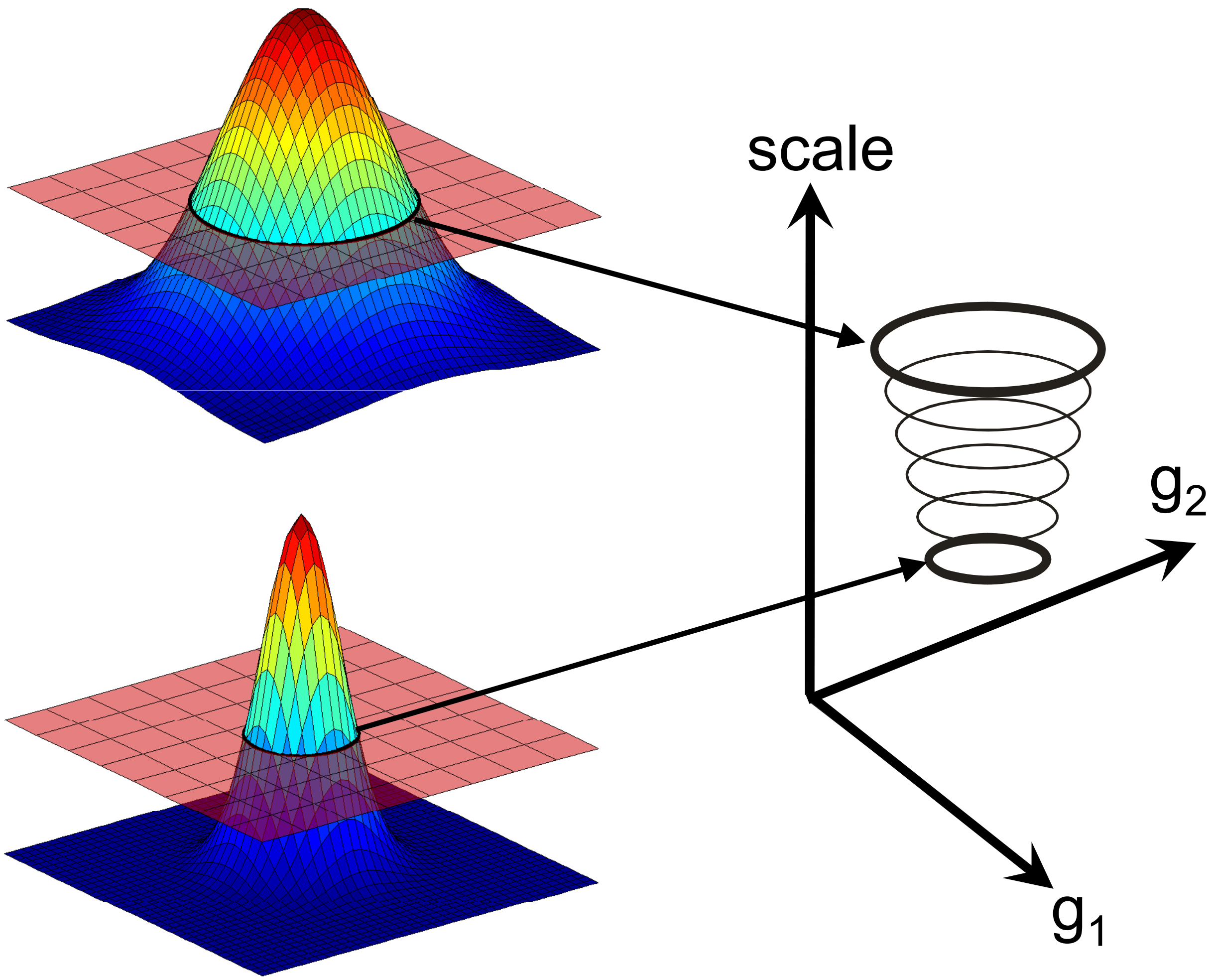
RTCGD result (1076 tumors, 4K inserts)

(Retroviral Tagged Cancer Gene Database)

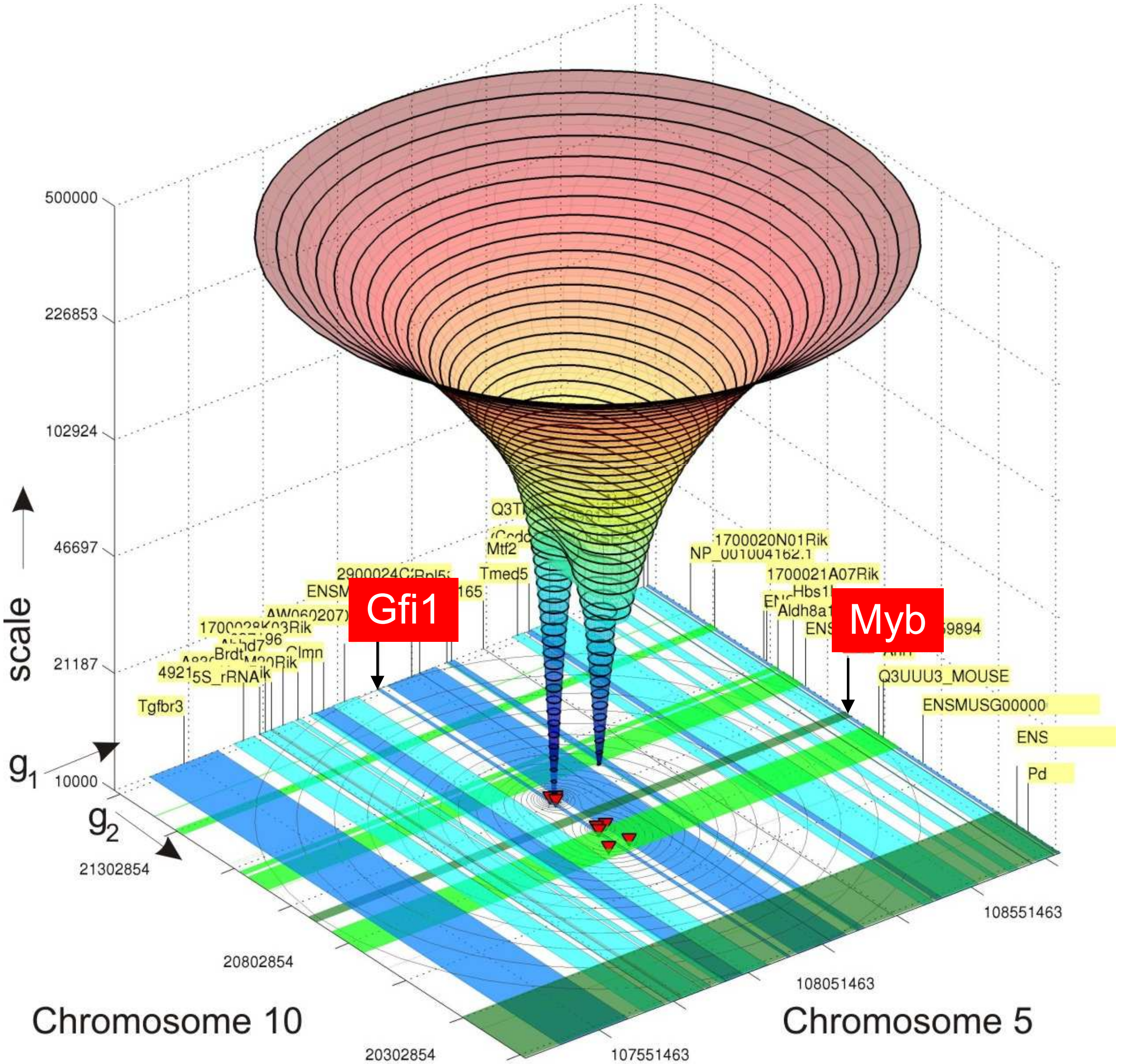


de Ridder *et al.* (2007) Bioinformatics 23; i133-i141.

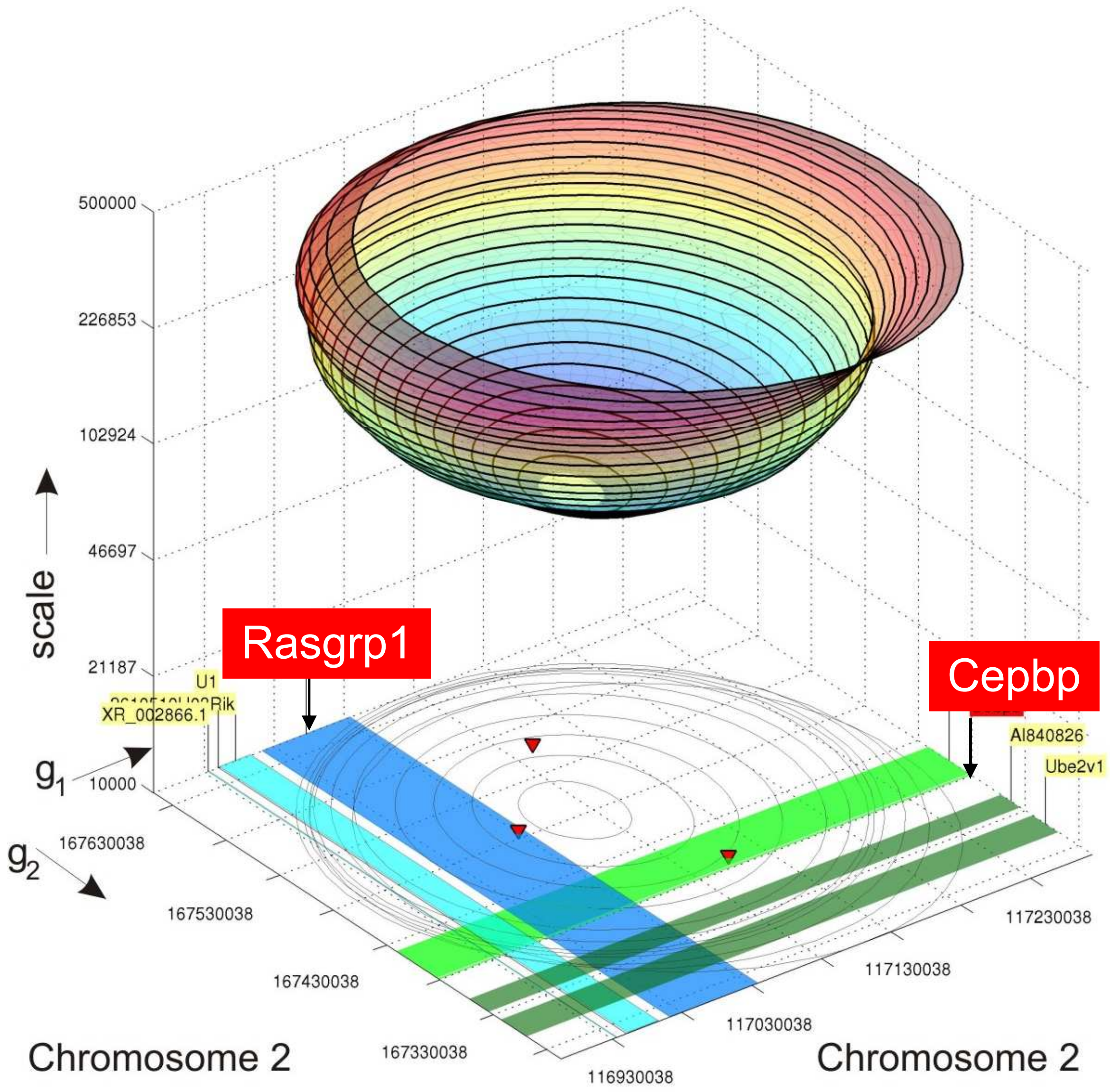
Building a scale space



Scale space for 2D GKC: *Myb-Gfi1*



Scale space for 2D GKC: *Rasgrp1-Cebpb*



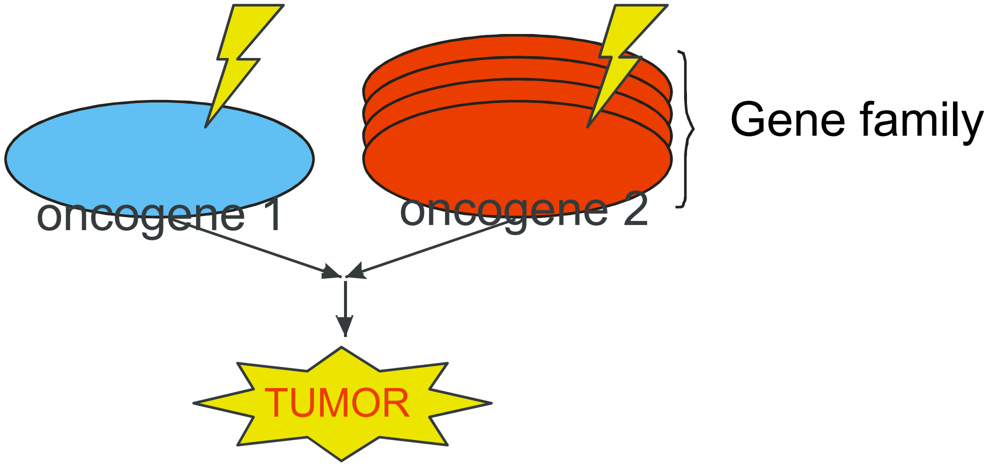
Finding cancer genes and cancer pathways

- Cancer genes:
  - genes individually frequently ‘hit’

- Cancer gene ‘pairs’:
  - pairs of genes frequently ‘hit’ in a specific pattern
  - (a gene and a family of genes frequently hit)
  - Co-operating, mutually exclusive

- Cancer pathways/networks
  - groups of genes frequently ‘hit’ in a specific pattern

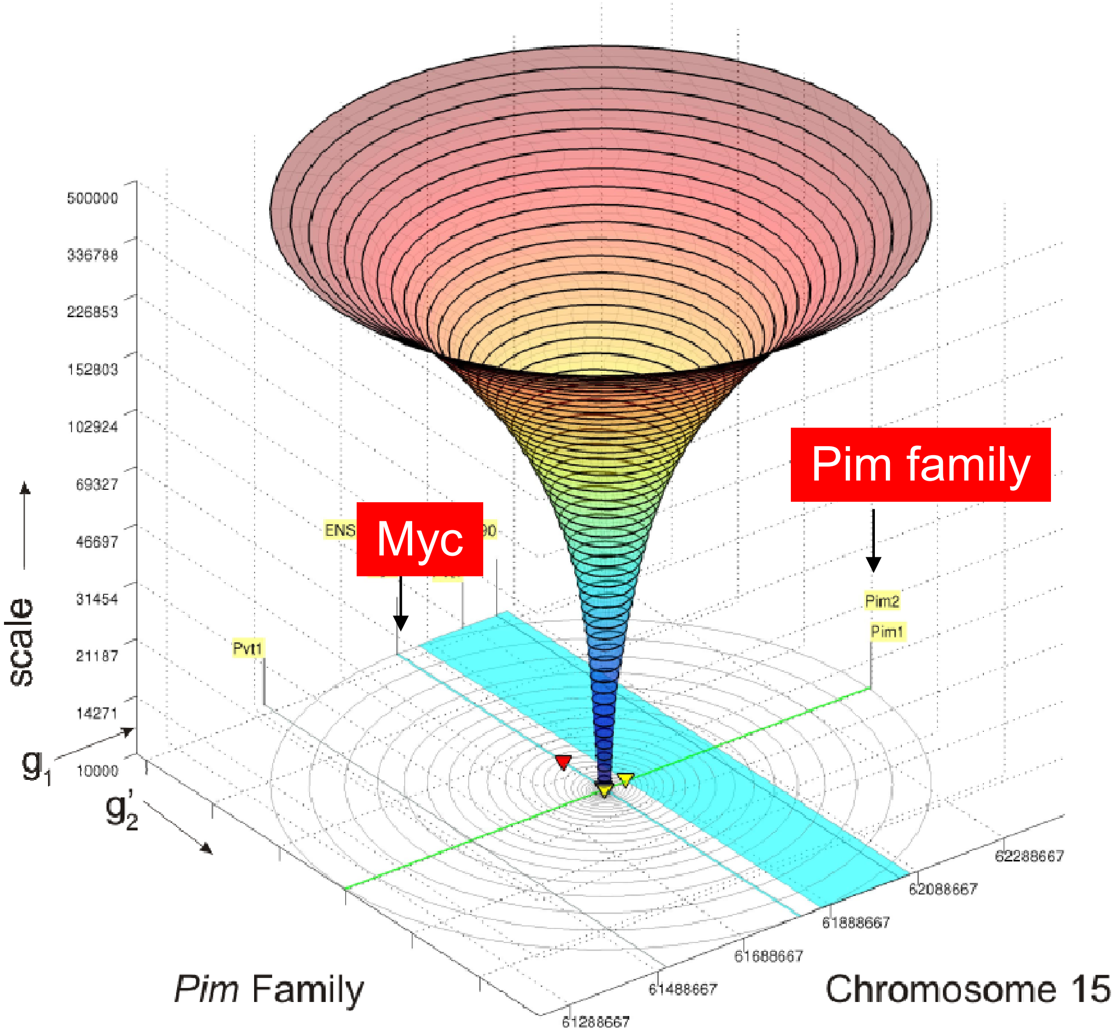
Cooperating genes and families



- Genes cooperate interchangeable
- Example: Myc and the Pim-Family
- Exploit this to detect more interactions

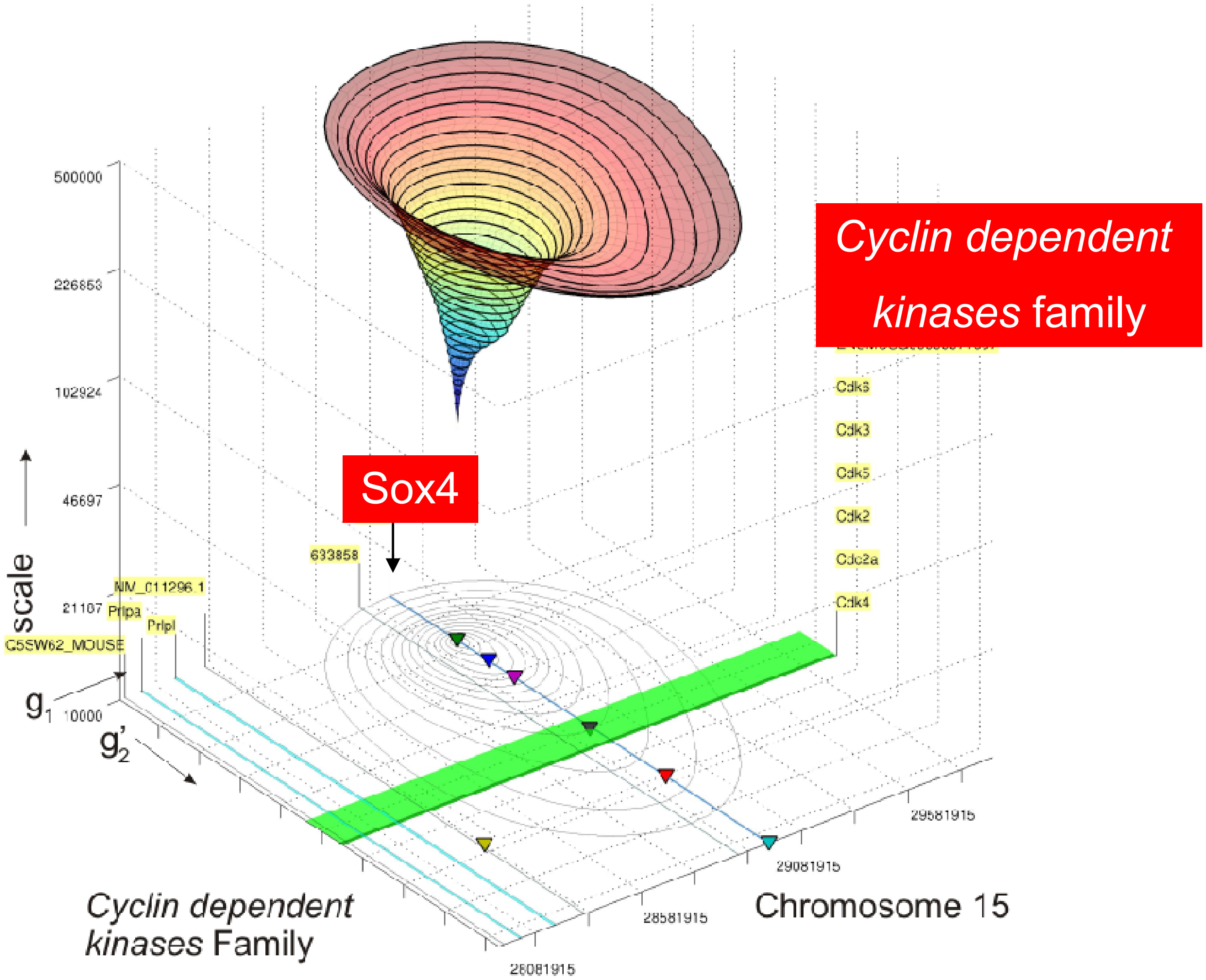
Family-mapped CCI

Myc and the *Pim* family



Family-mapped CCI

Sox4 and the *Cyclin dependent kinases* family



Acknowledgements

Jeroen de Ridder

Marcel Reinders

Jos Jonkers

Jaap Kool
Anthony Uren
Maarten van Lohuizen
Anton Berns

http://bioinformatics.nki.nl/ (vacancies)