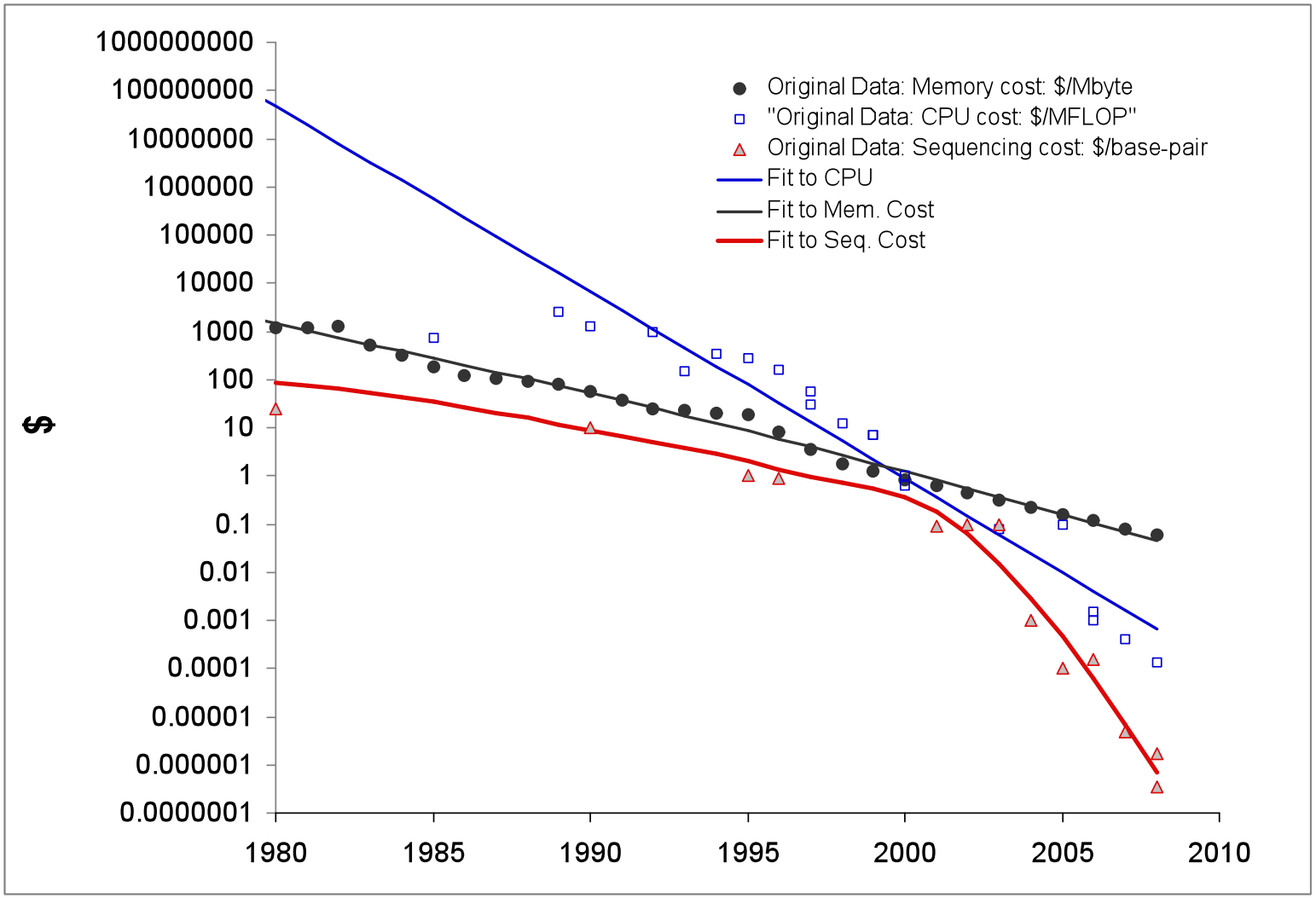
Detection and analysis of structural variants in personal genomes

Mark B Gerstein
Yale

Slides at
Lectures.GersteinLab.org
(See Last Slide for References & More Info.)

Plummeting Cost of Sequencing...



[Greenbaum et al., Am. J. Bioethics ('08)]

... has led to the era of Personal Genomics

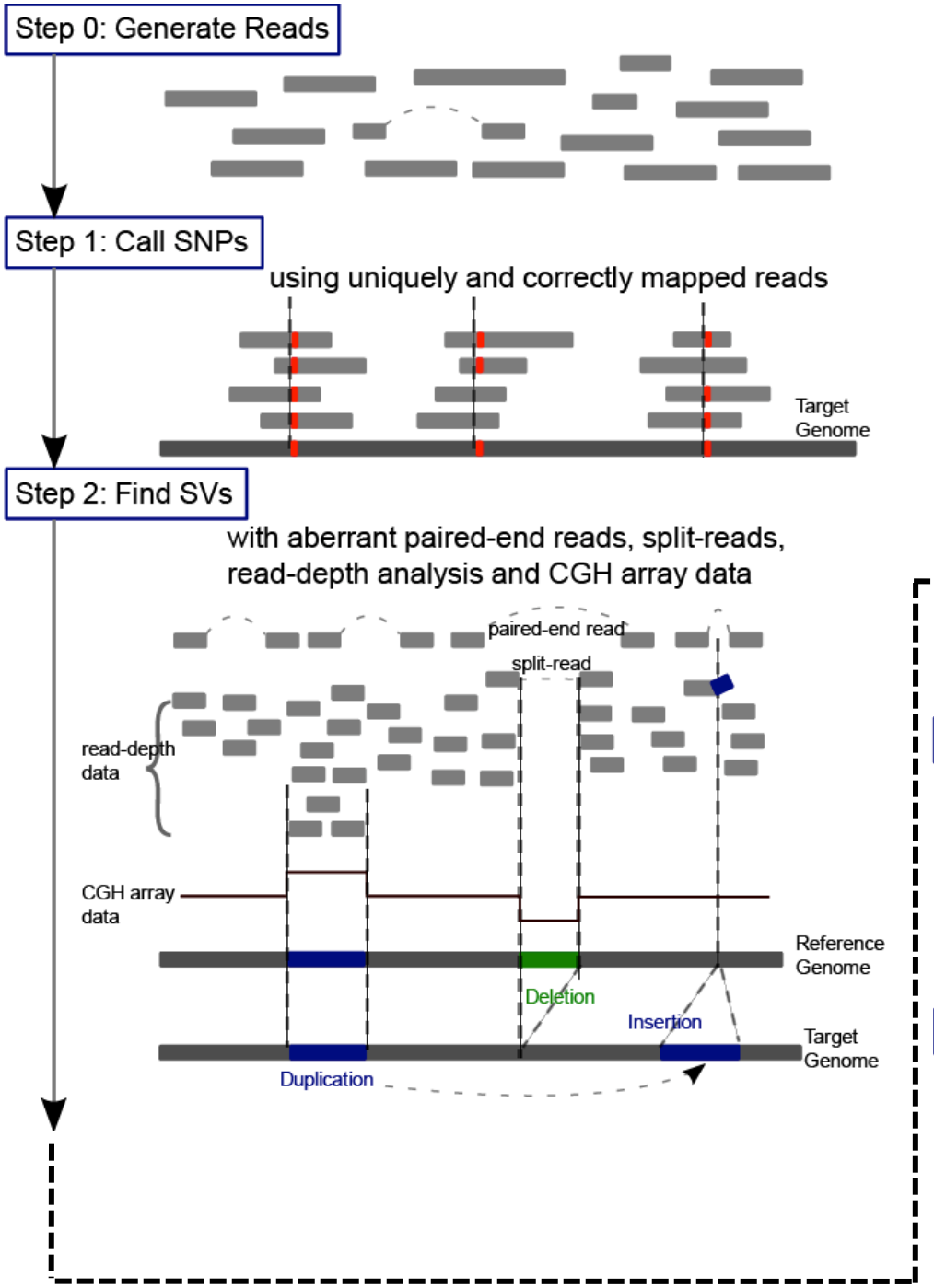
- Resequencing of individuals' genomes
- Now for a few, eventually for many

- Easy (!) except for the genome's complex,

REPEAT-CON**TA**INING **STRUCTURE**

Main Steps in Genome Resequencing

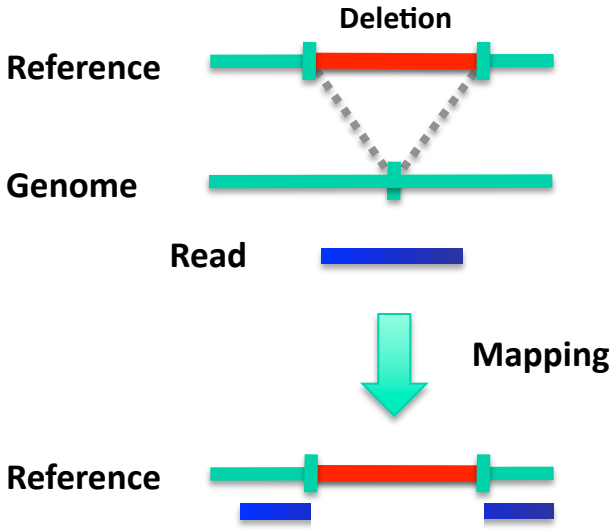
[Snyder et al. Genes & Dev. ('09), submitted]



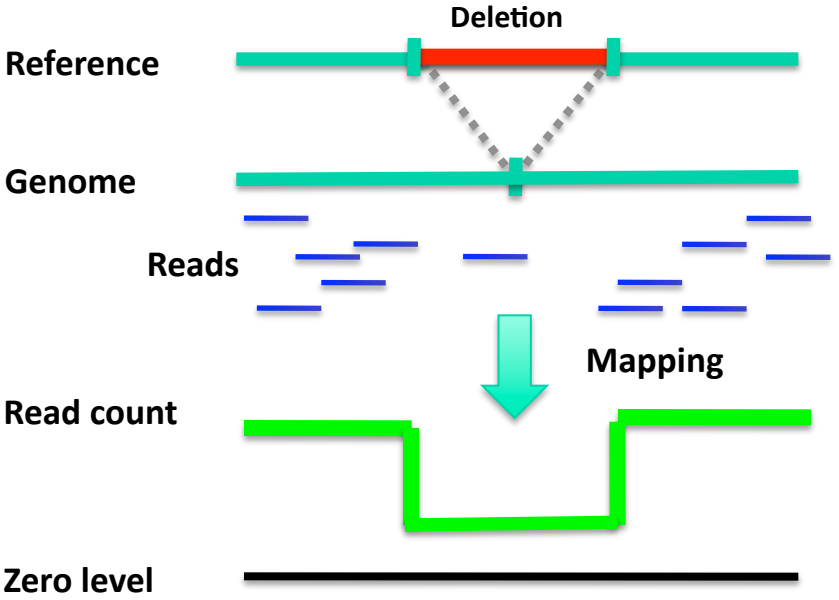
Methods to **Find SVs**

1. Paired ends

2. Split read



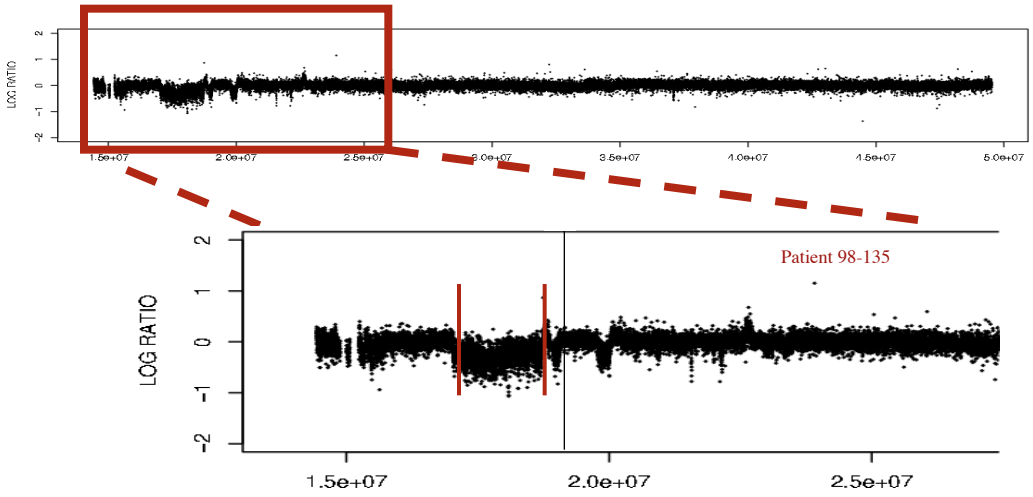
3. Read depth (or aCGH)



4. Local Reassembly

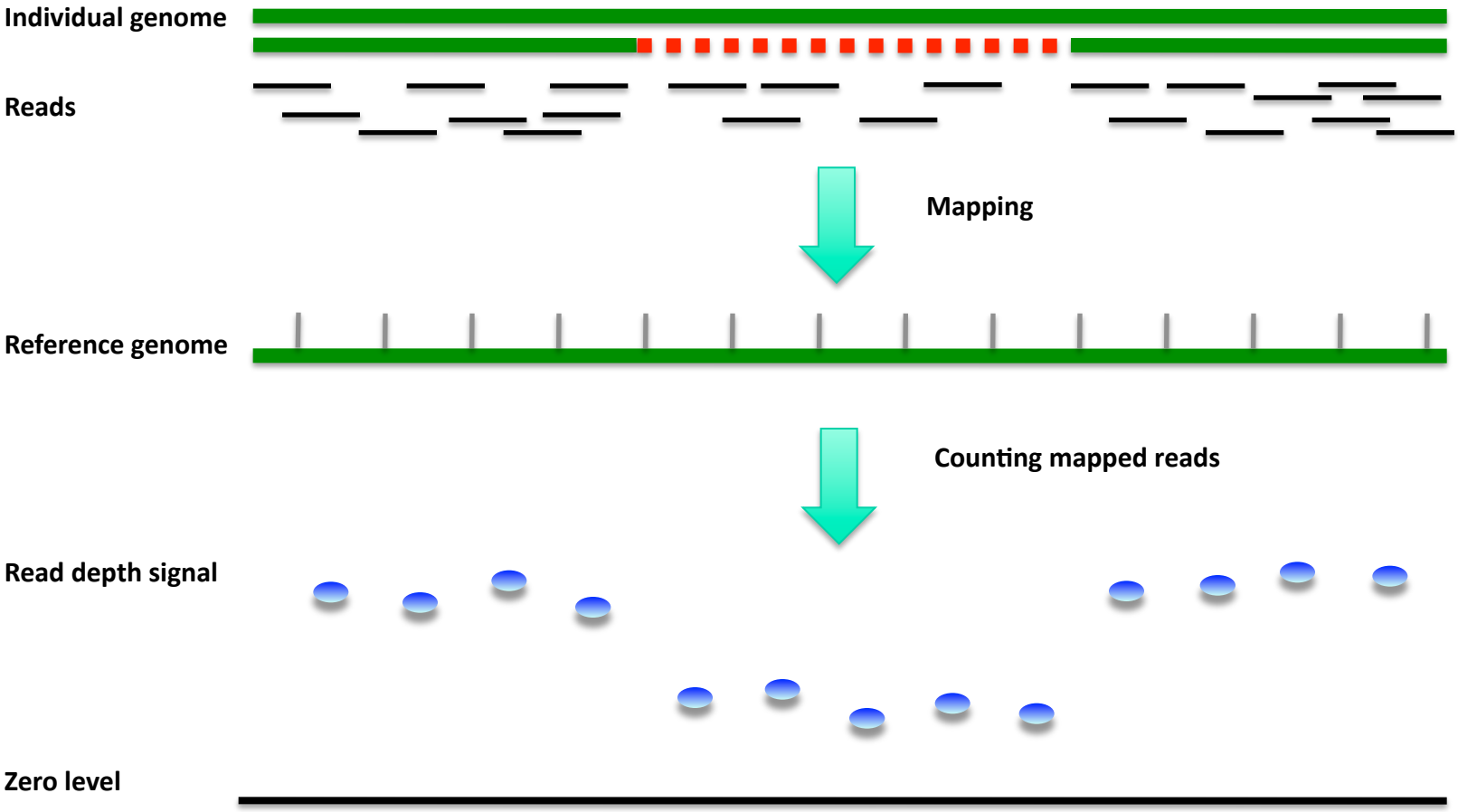
[Snyder et al. Genes & Dev. ('09), submitted]

Breakpointer: Segmentation of Array Signal as precursor to Read Depth



Array Signal

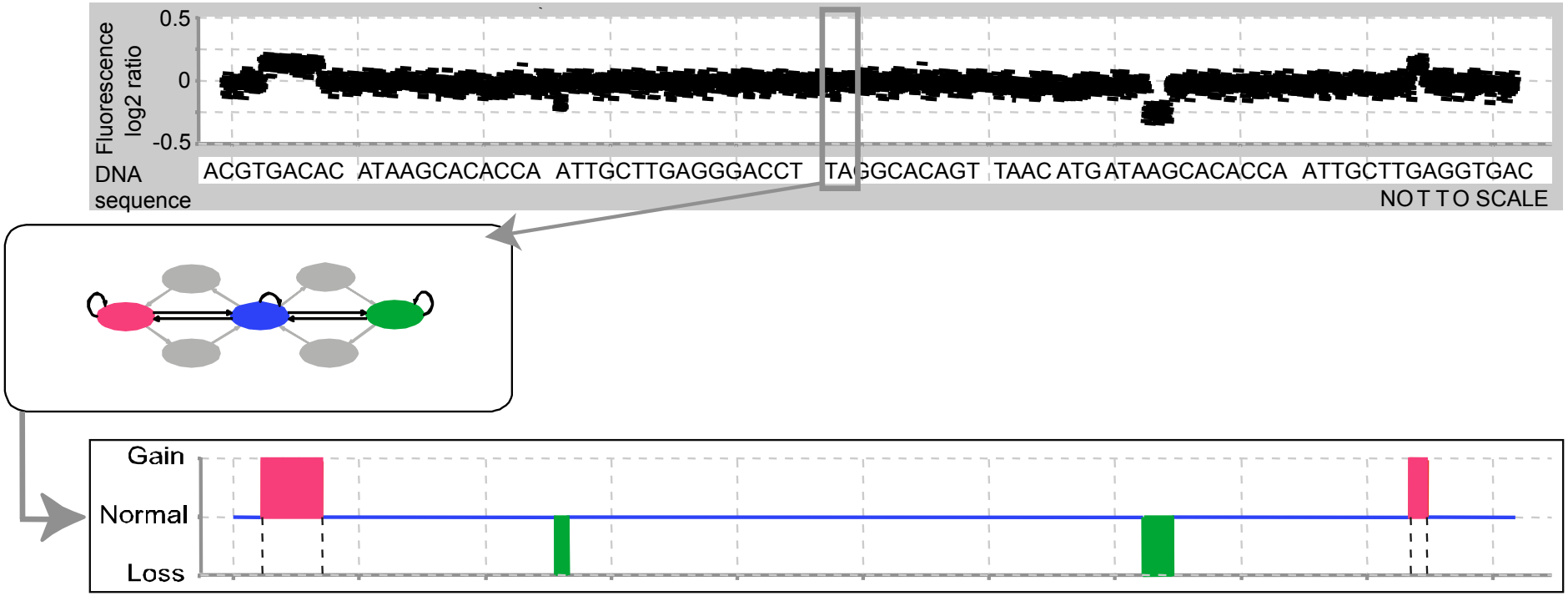
Read depth



[Urban et al. ('06) PNAS; Wang et al. Gen. Res ('09)]

BreakPtr HMM

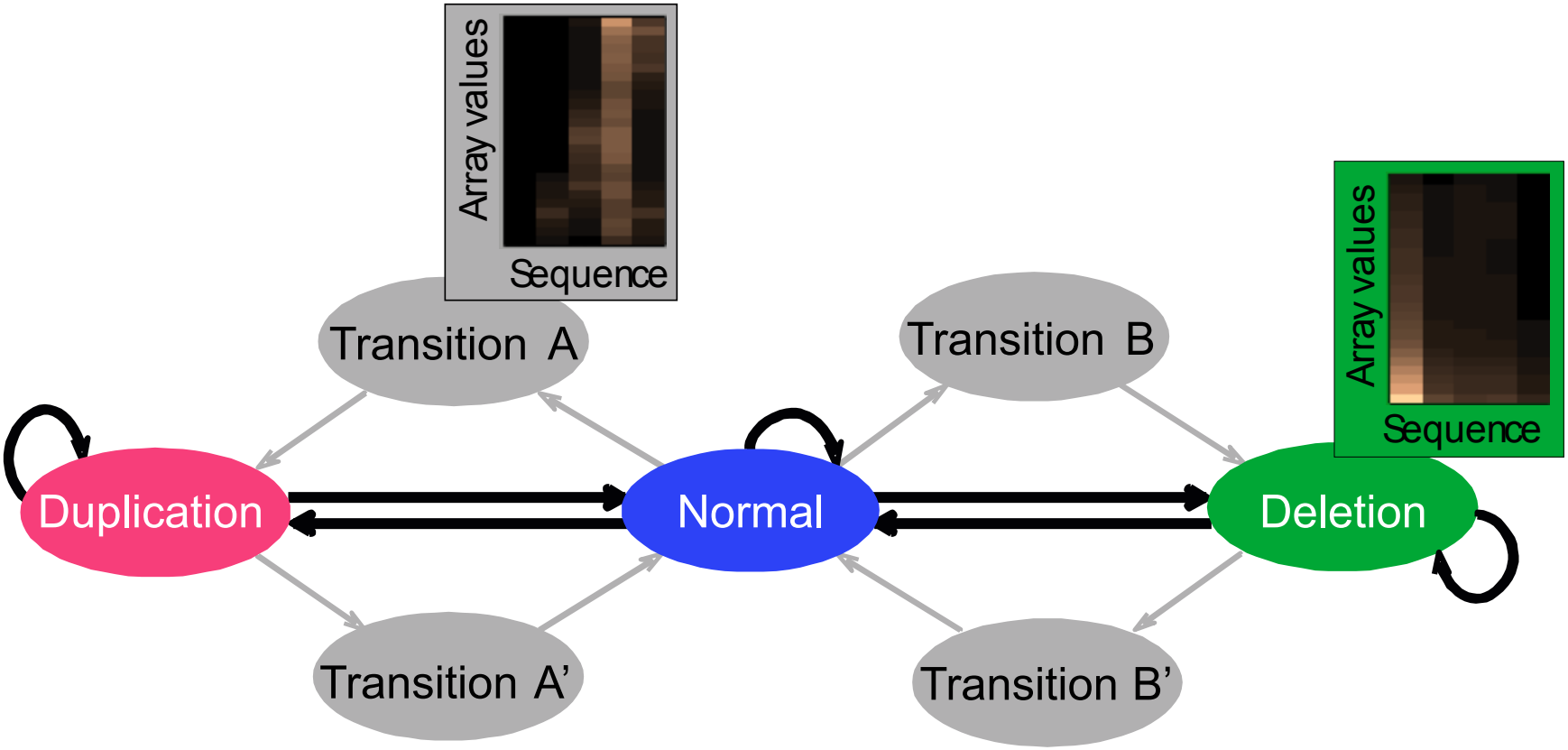
- To get highest resolution on breakpoints need to smooth & segment the signal
- BreakPtr: prediction of breakpoints, dosage and cross-hybridization using a system based on Hidden Markov Models



Korbel*, Urban* *et al.*, PNAS (2007)

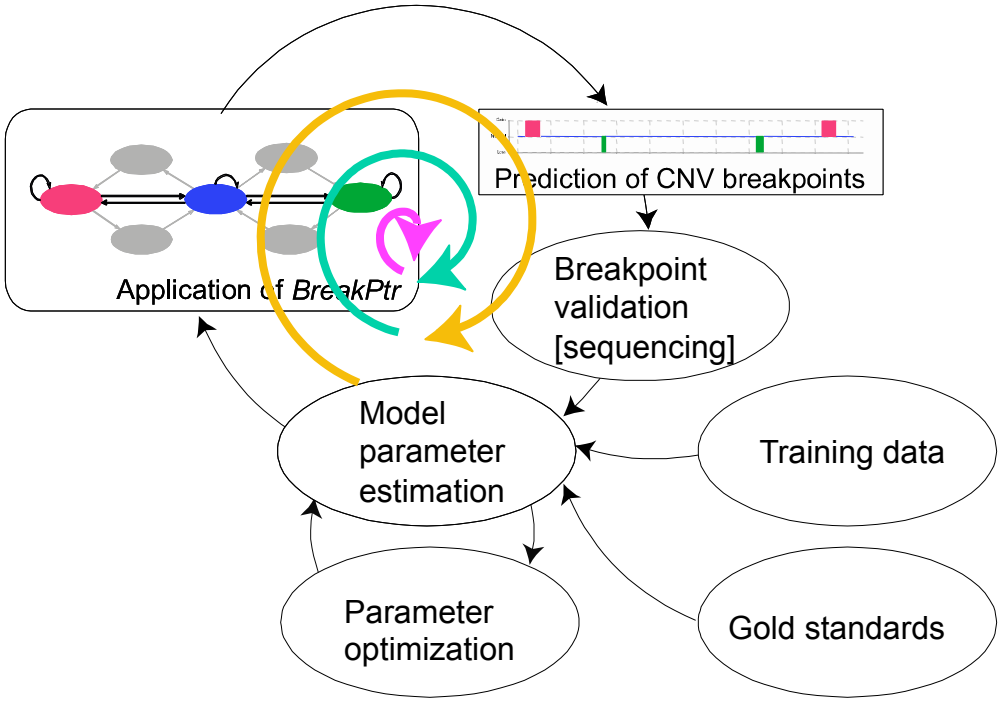
http://breakptr.gersteinlab.org

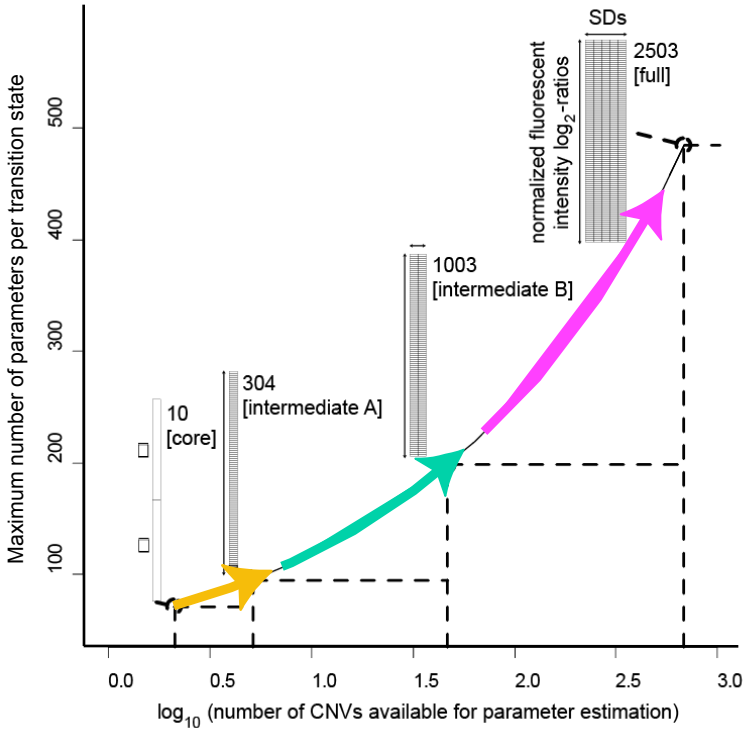
BreakPtr statistically integrates array signal and DNA sequence signatures (using a discrete-valued bivariate HMM)



Korbel*, Urban* *et al.*, PNAS (2007)

‘Active’ approach for breakpoint identification: initial scoring with preliminary model, targeted validation (with sequencing), retraining, and rescoring





CNV breakpoints sequenced in ~10 cases following BreakPtr analysis;

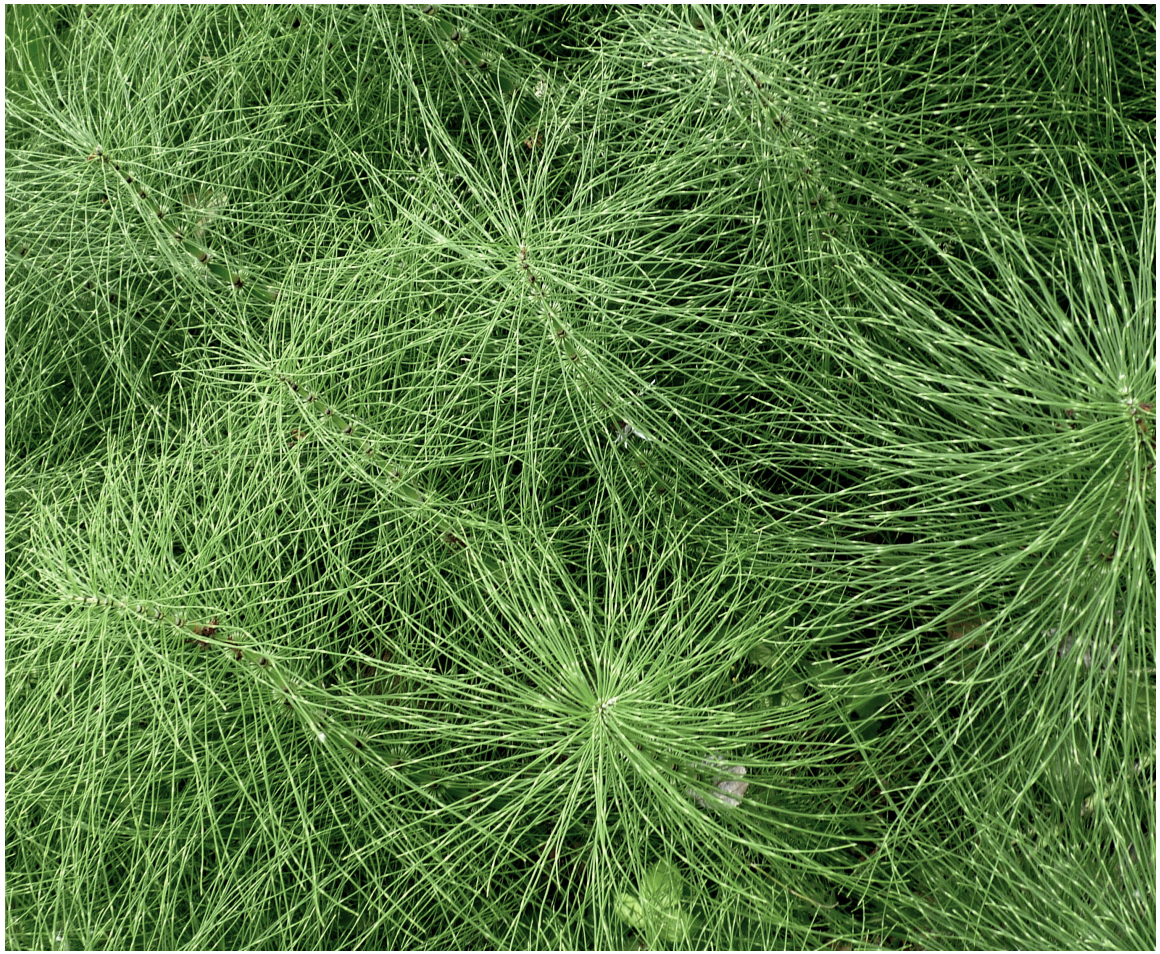
Median resolution <300 bp

No improvement in accuracy with higher resolution (9nt tiling)

HMM optimized iteratively (using Expectation Maximization, EM)

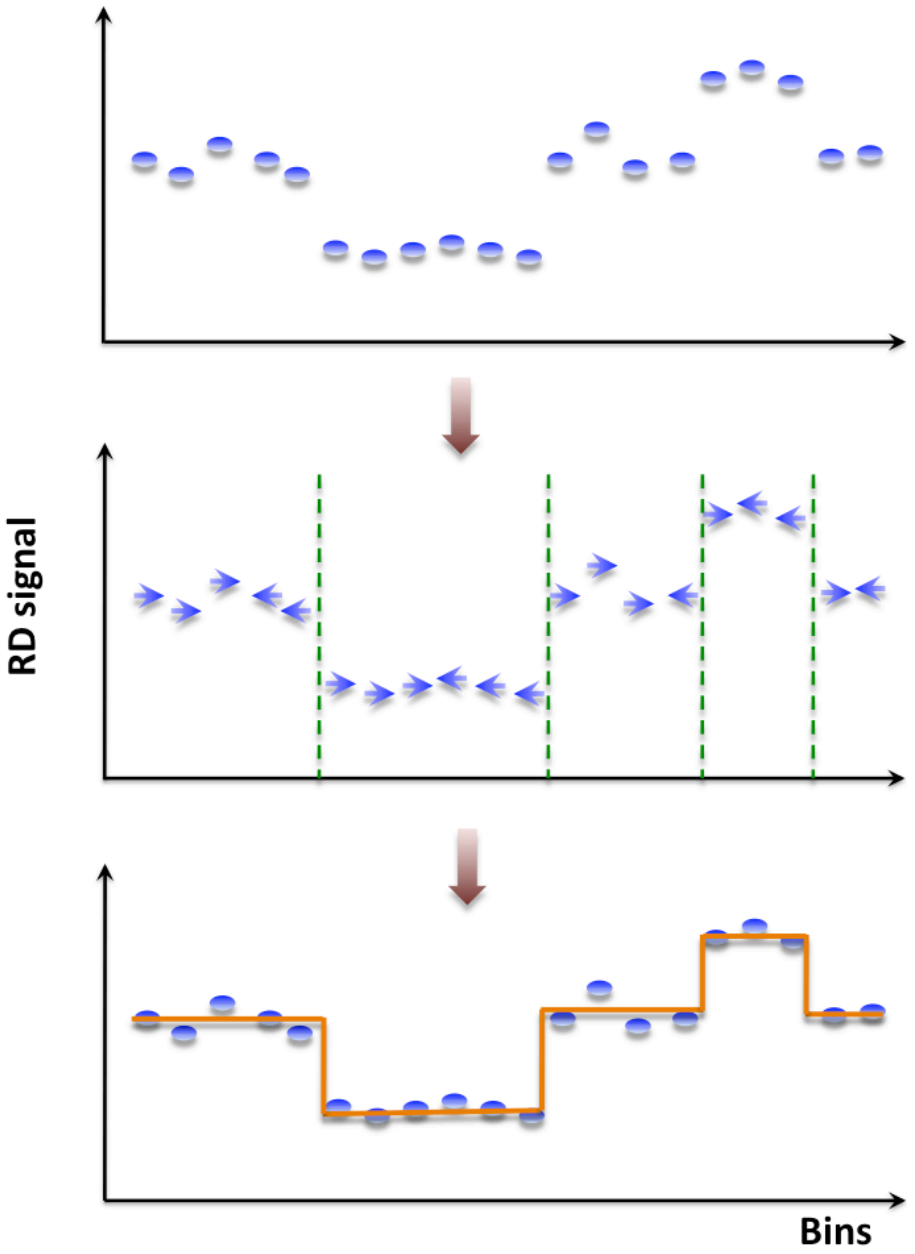
Korbel*, Urban* *et al.,* PNAS (2007)

MSB: Read-Depth Segmentation



Mean-shift-based (MSB) Segmentation: no explicit model

- For each bin attraction (mean-shift) vector points in the direction of bins with most similar RD signal
- No prior assumptions about number, sizes, haplotype, frequency and density of CNV regions
- Not Model-based (e.g. like HMM) with global optimization, distr. assumption & parms. (e.g. num. of segments).
- Achieves discontinuity-preserving smoothing
- Derived from image-processing applications

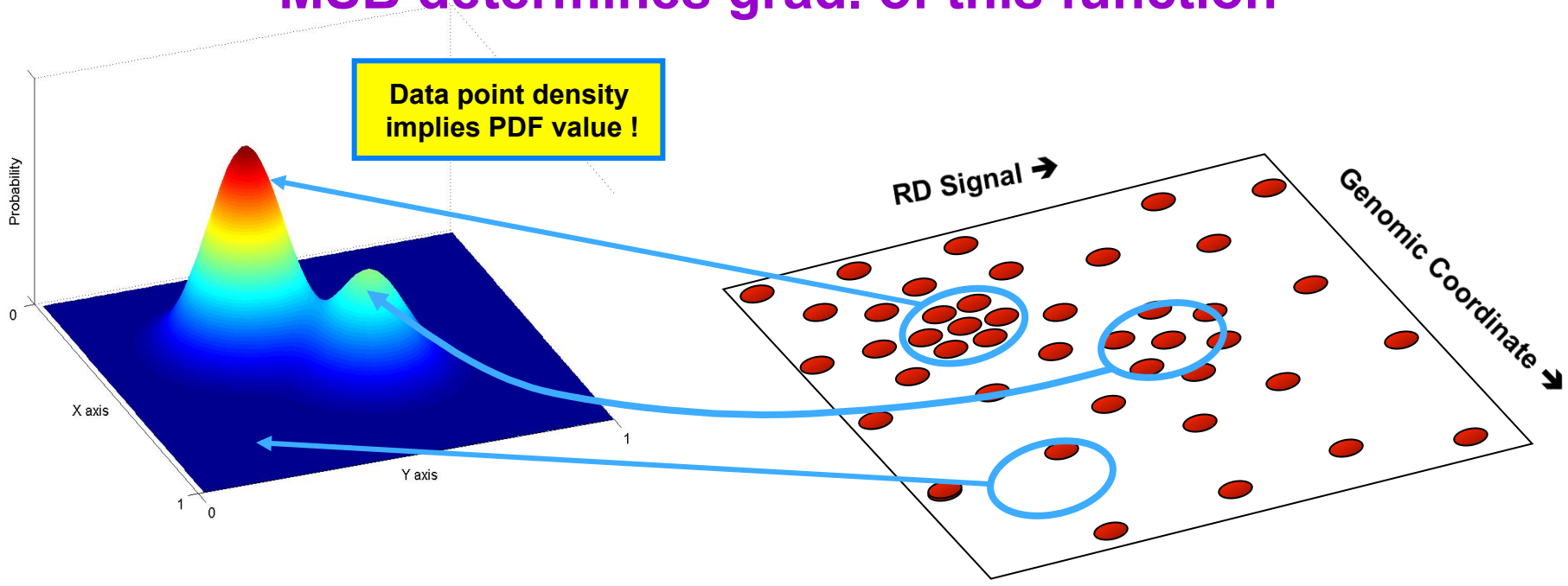


[Wang et al. Gen. Res ('09) 19:106]

Some Intuition on how MSB works: Non-Parametric Density Estimation

Assumption : The data points are sampled from an underlying PDF

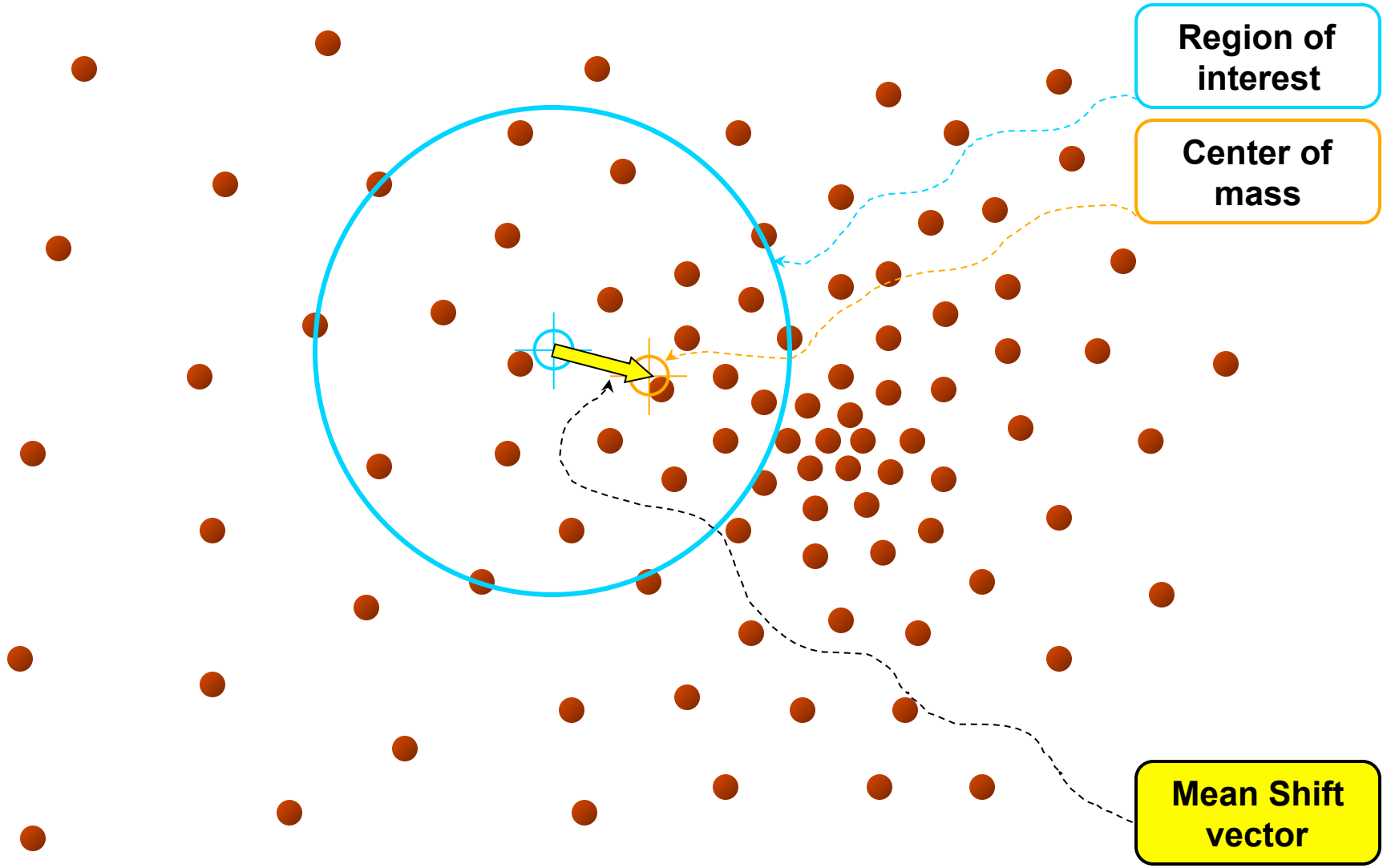
MSB determines grad. of this function



Assumed Underlying PDF **Real Data Samples**

[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

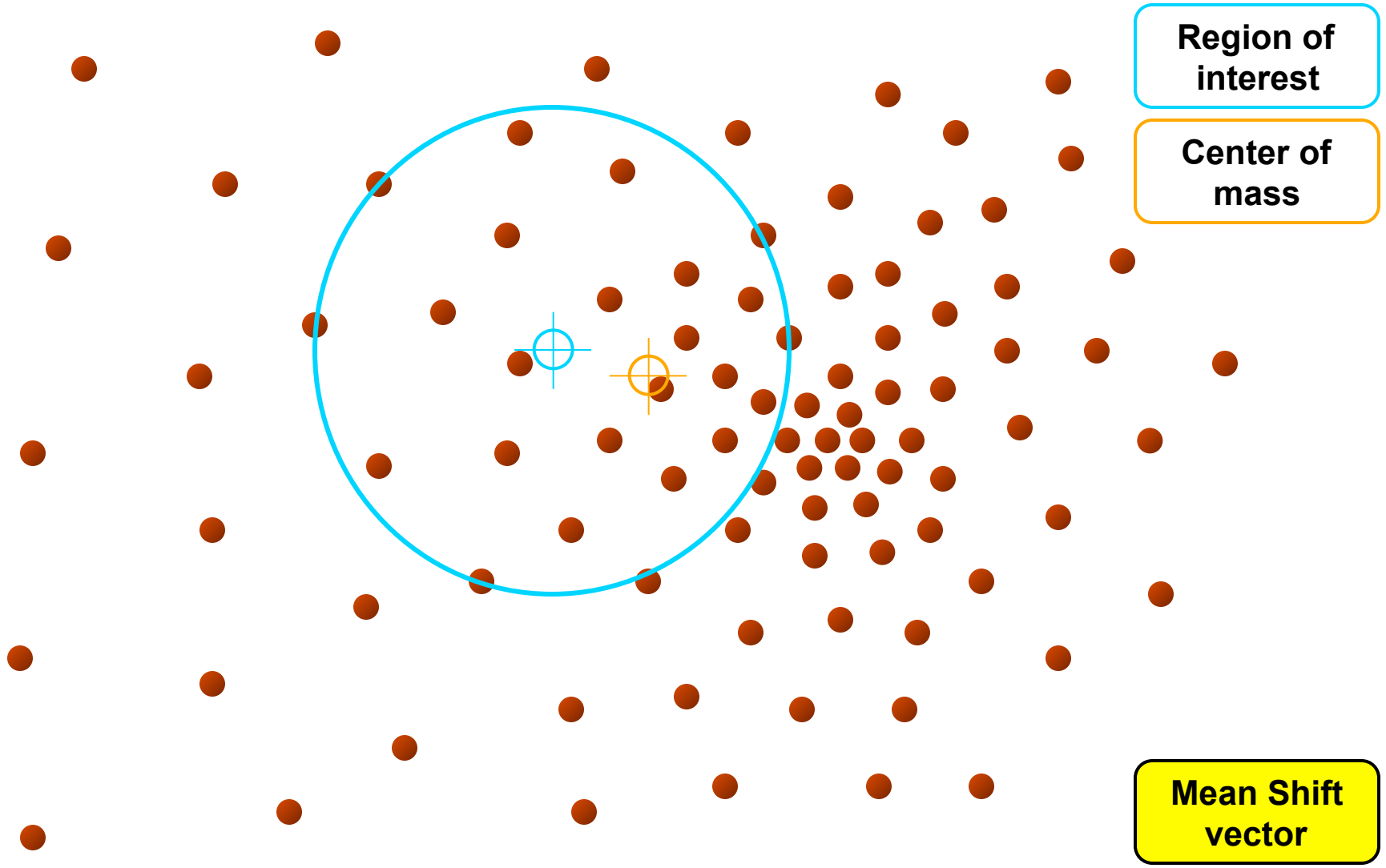
Intuitive Description



Objective : Find the densest region
Distribution of identical billiard balls

[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

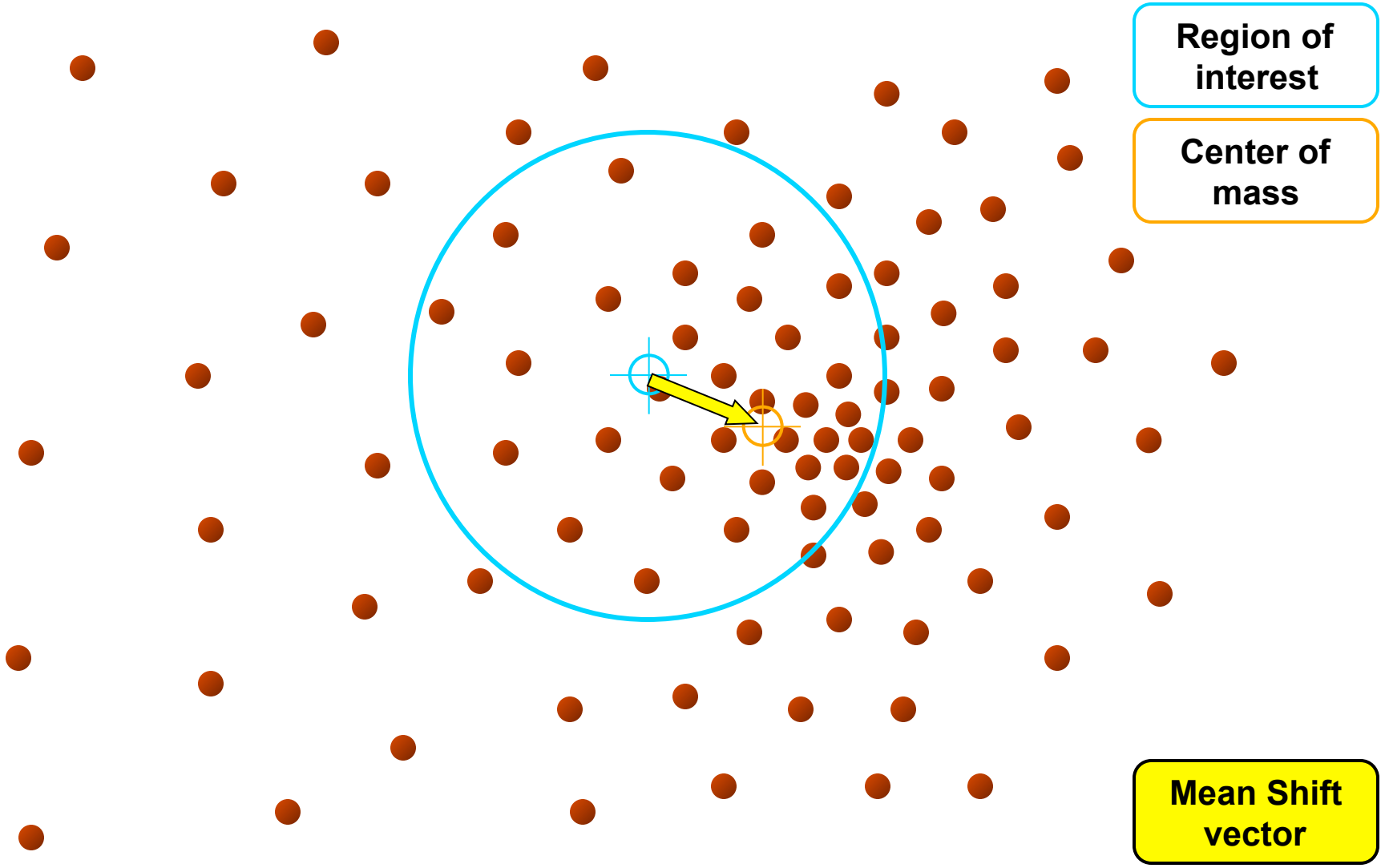
Intuitive Description



Objective : Find the densest region
Distribution of identical billiard balls

[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

Intuitive Description

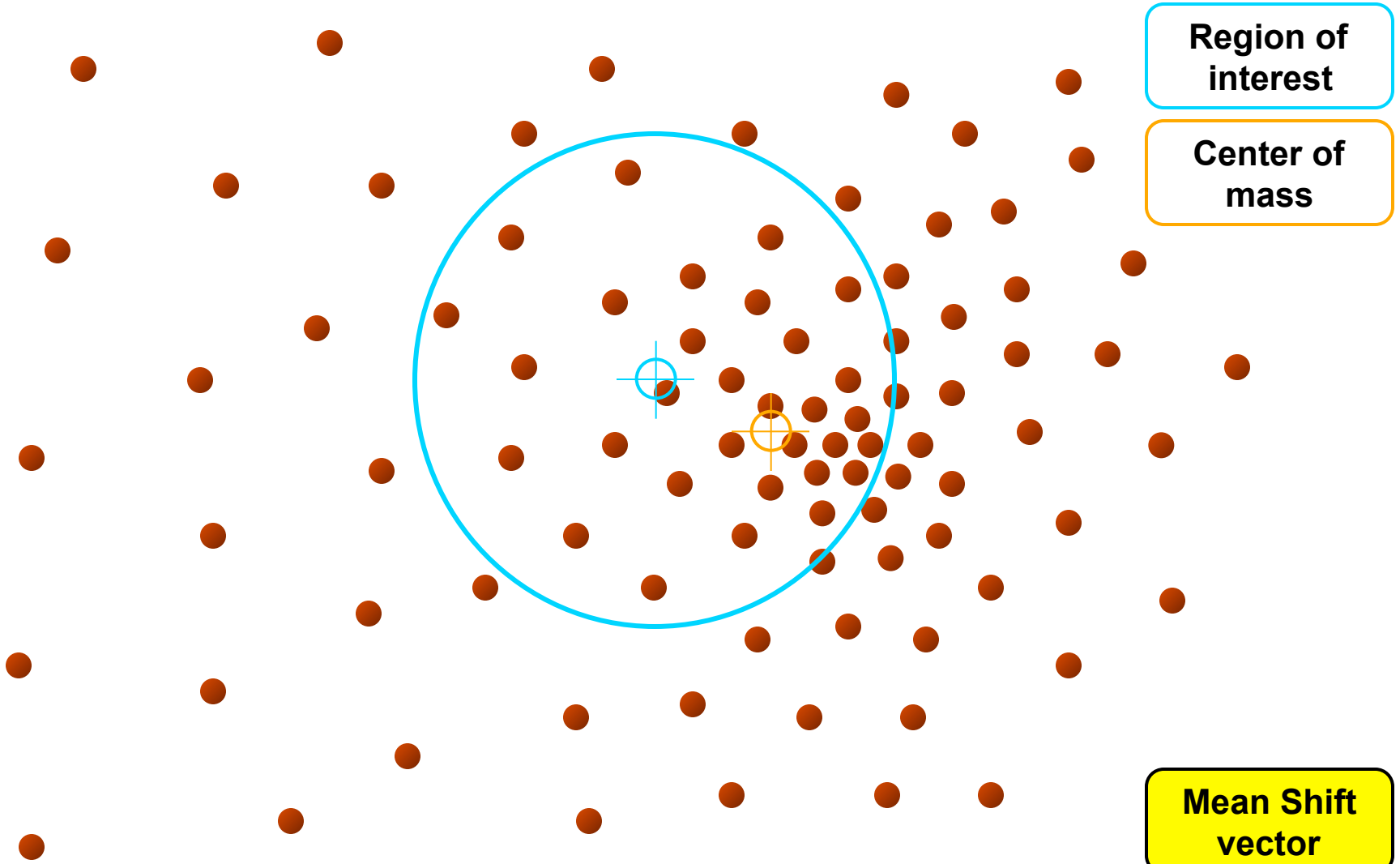


Objective : Find the densest region

Distribution of identical billiard balls

[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

Intuitive Description

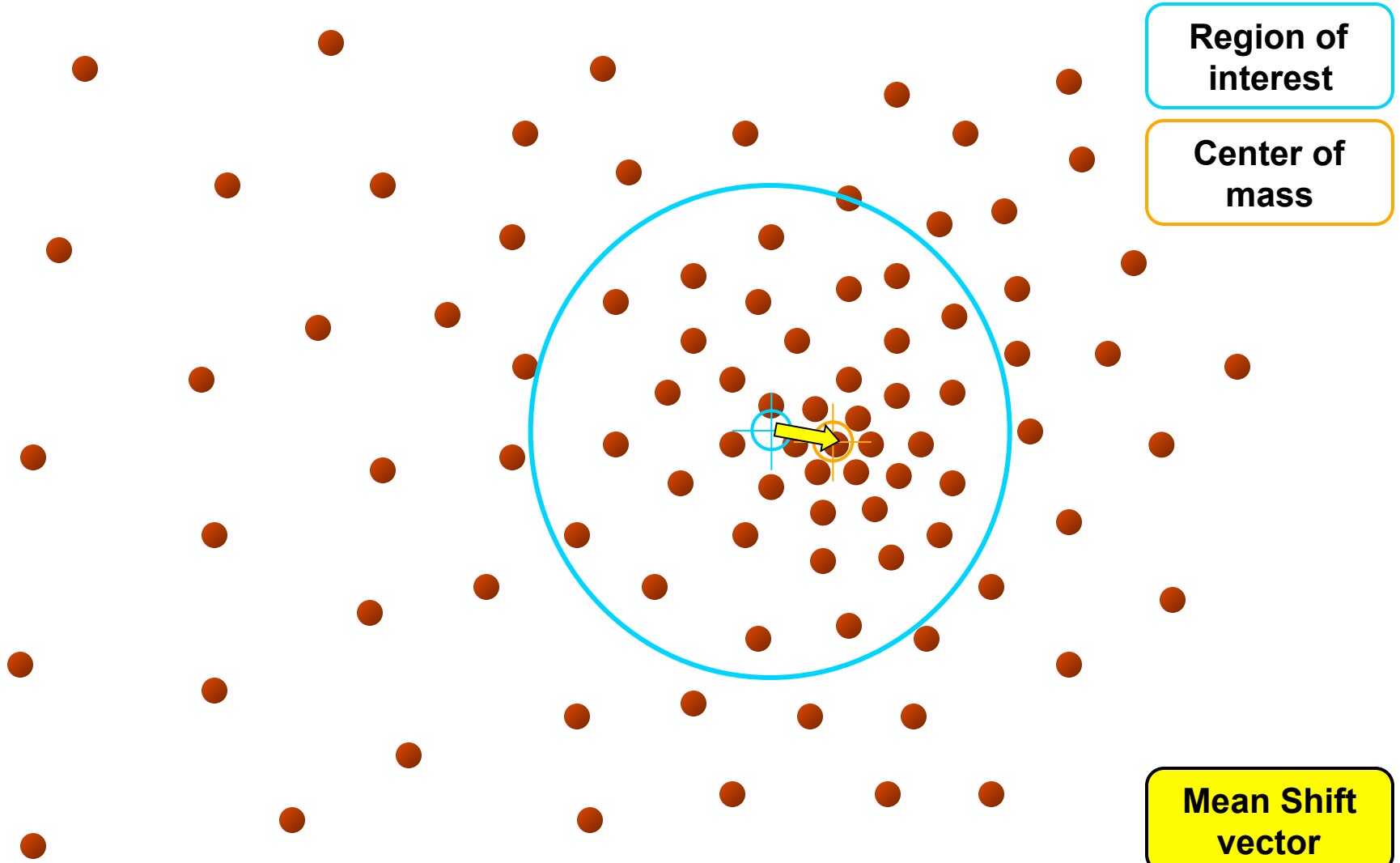


Objective : Find the densest region

Distribution of identical billiard balls

[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

Intuitive Description

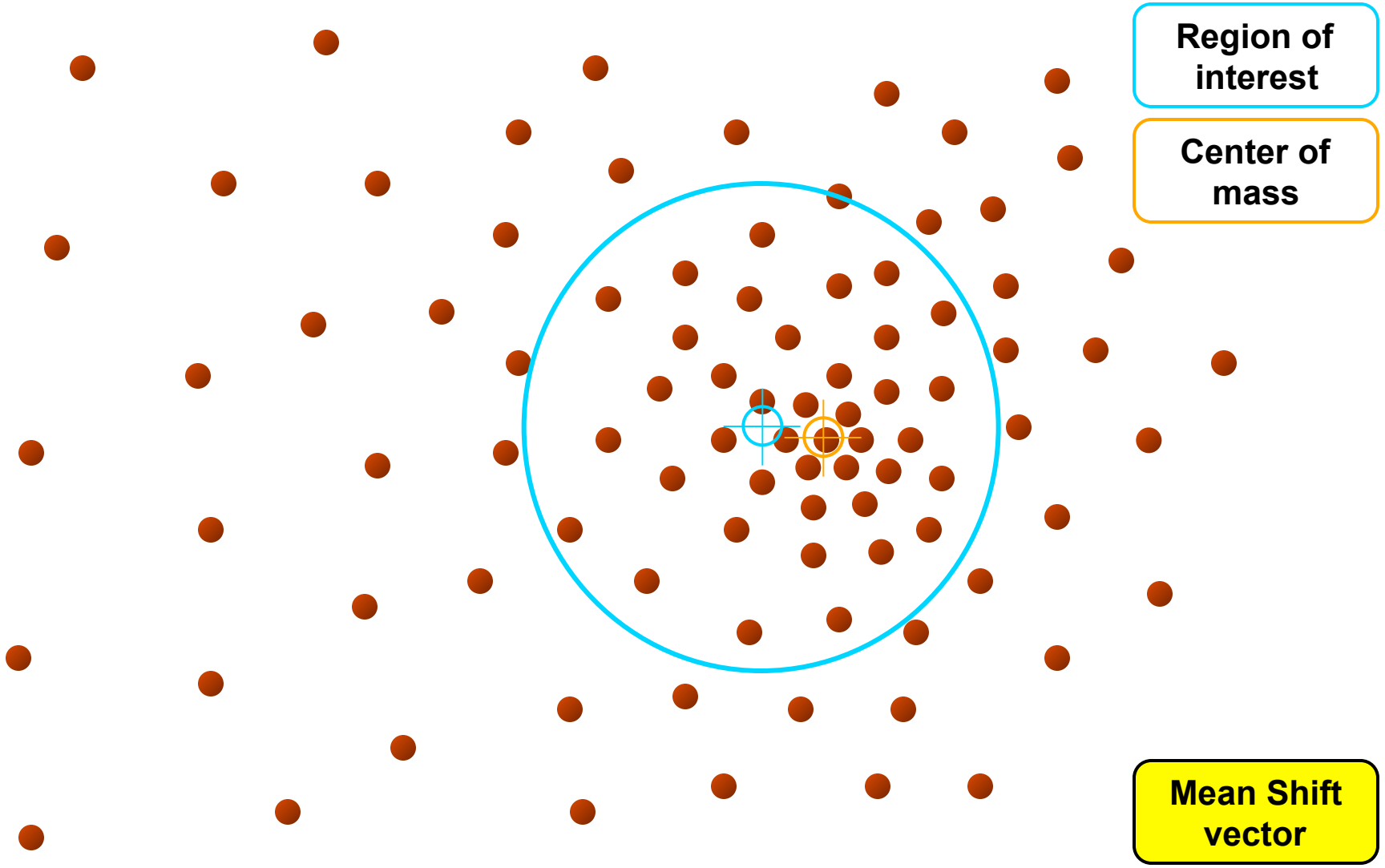


Objective : Find the densest region

Distribution of identical billiard balls

[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

Intuitive Description

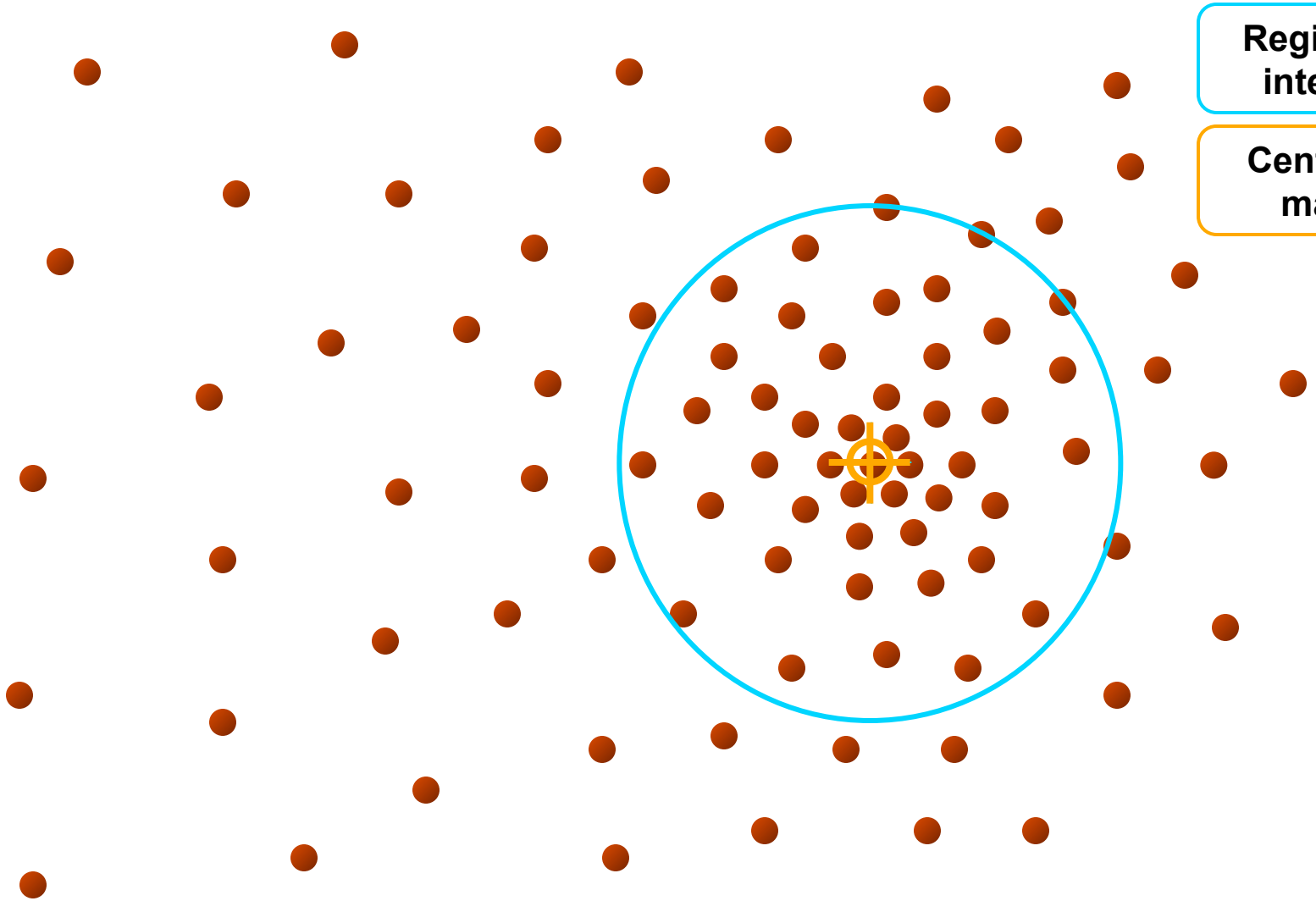


Objective : Find the densest region

Distribution of identical billiard balls

[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

Intuitive Description

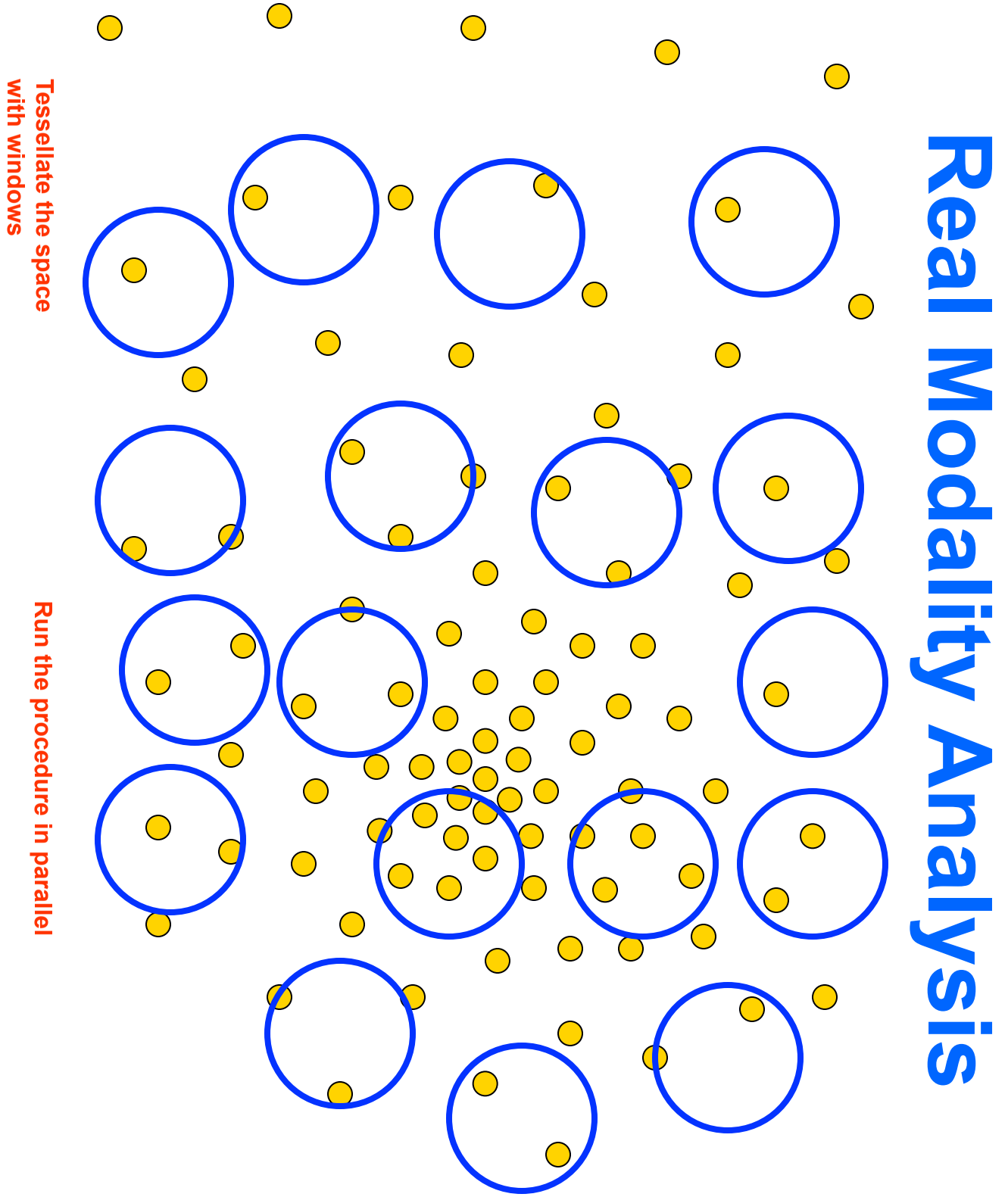


Objective : Find the densest region
Distribution of identical billiard balls

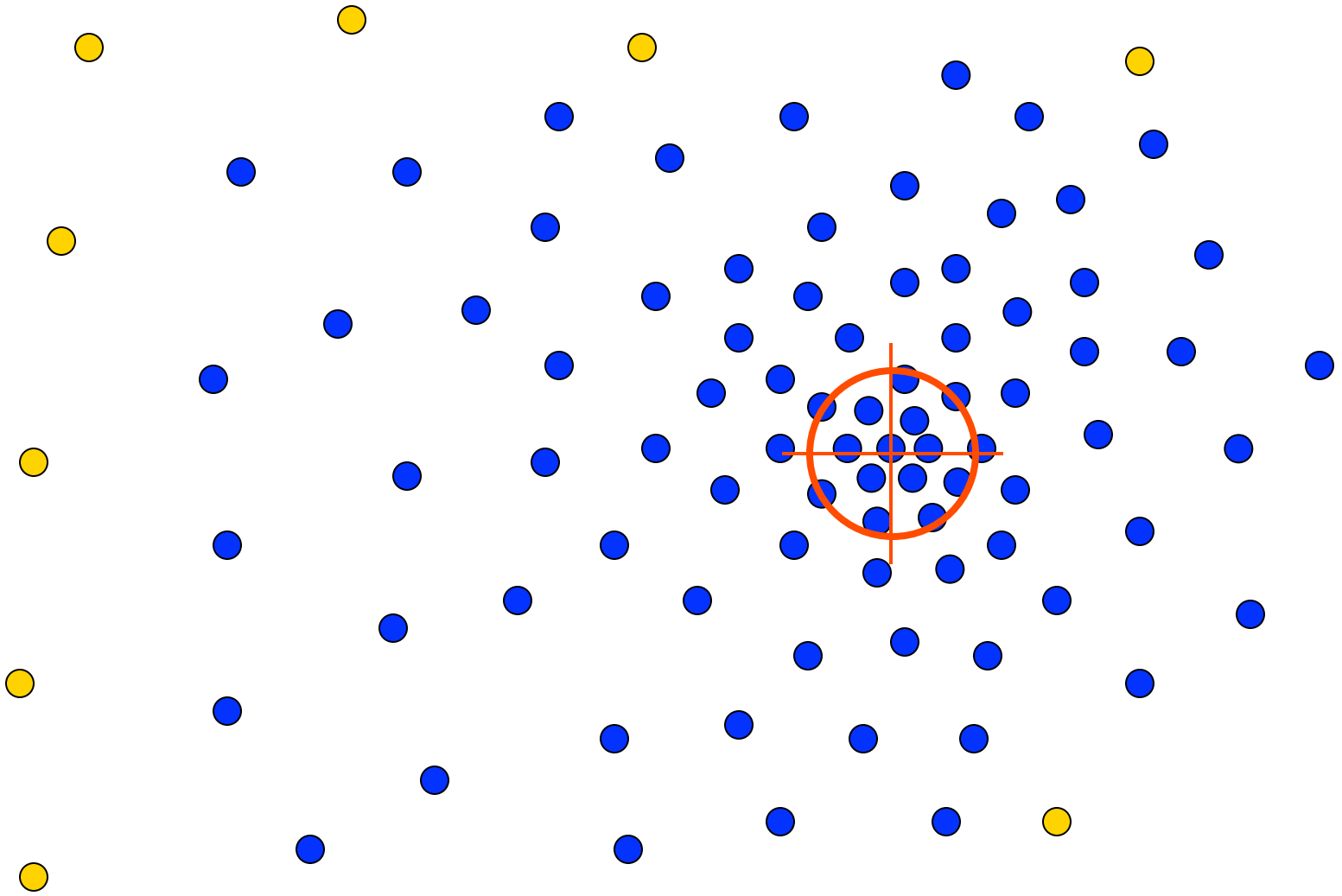
[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

Real Modality Analysis



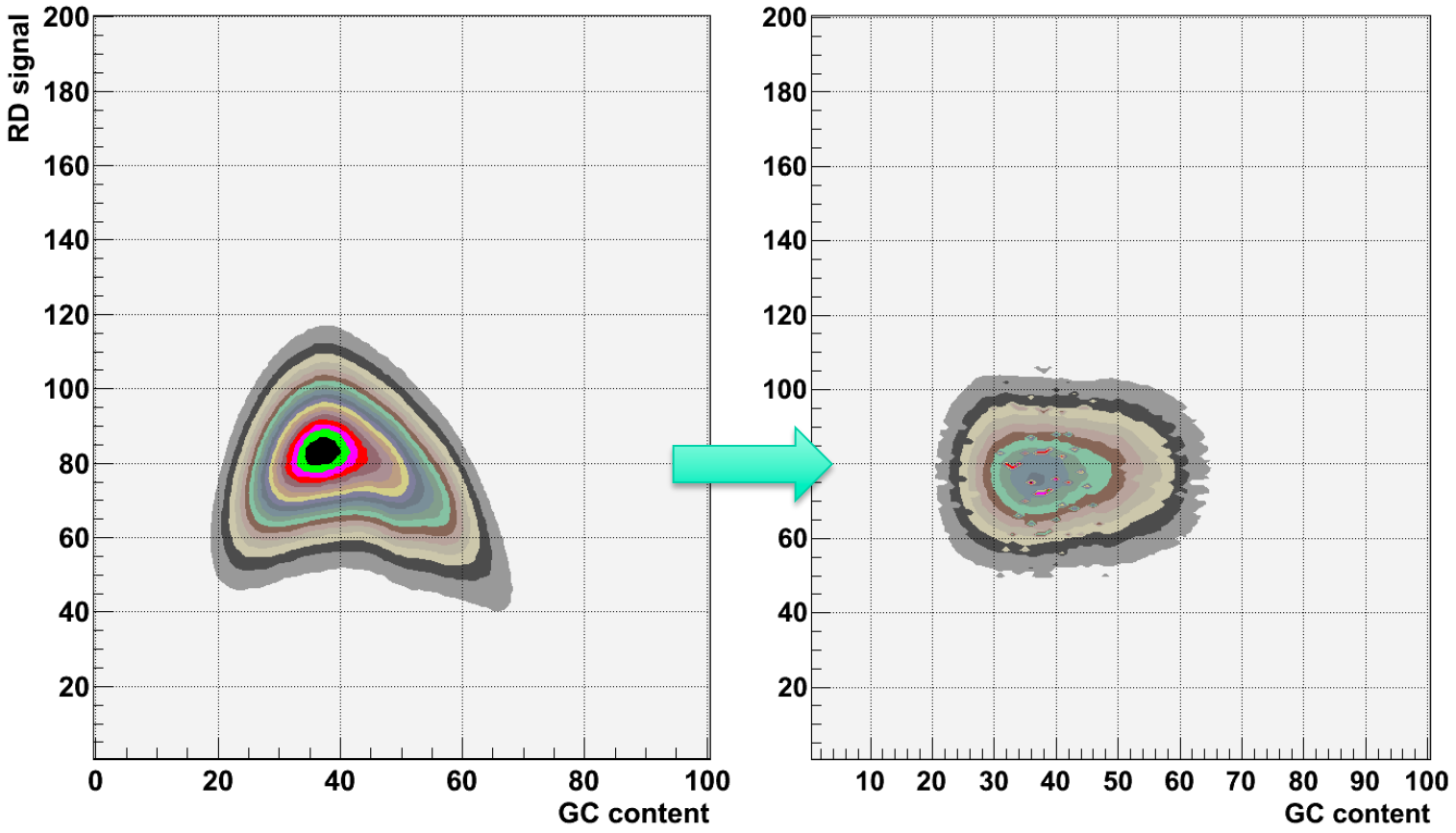
Real Modality Analysis



The blue data points were traversed by the windows towards the mode

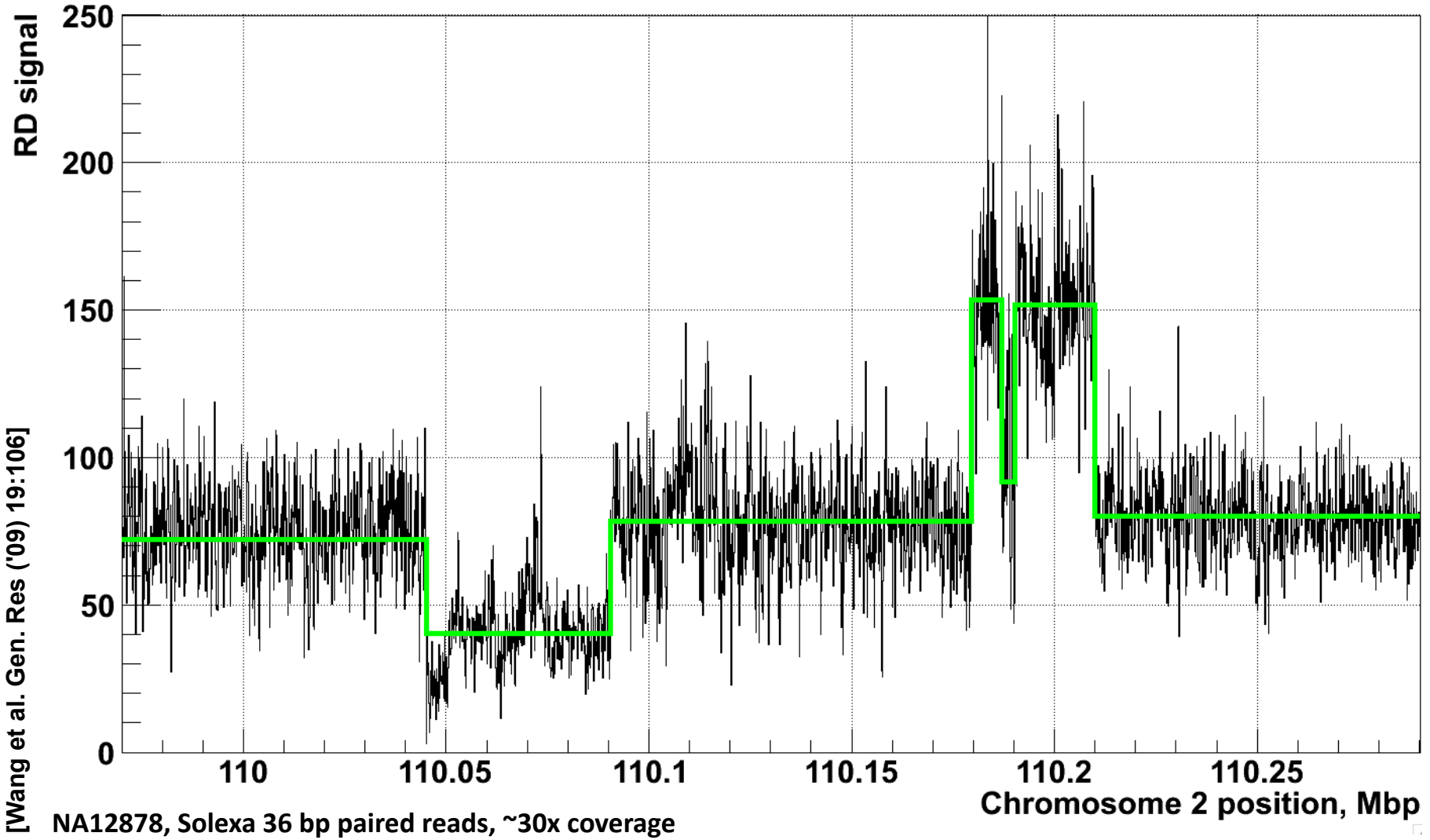
[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

GC bias correction

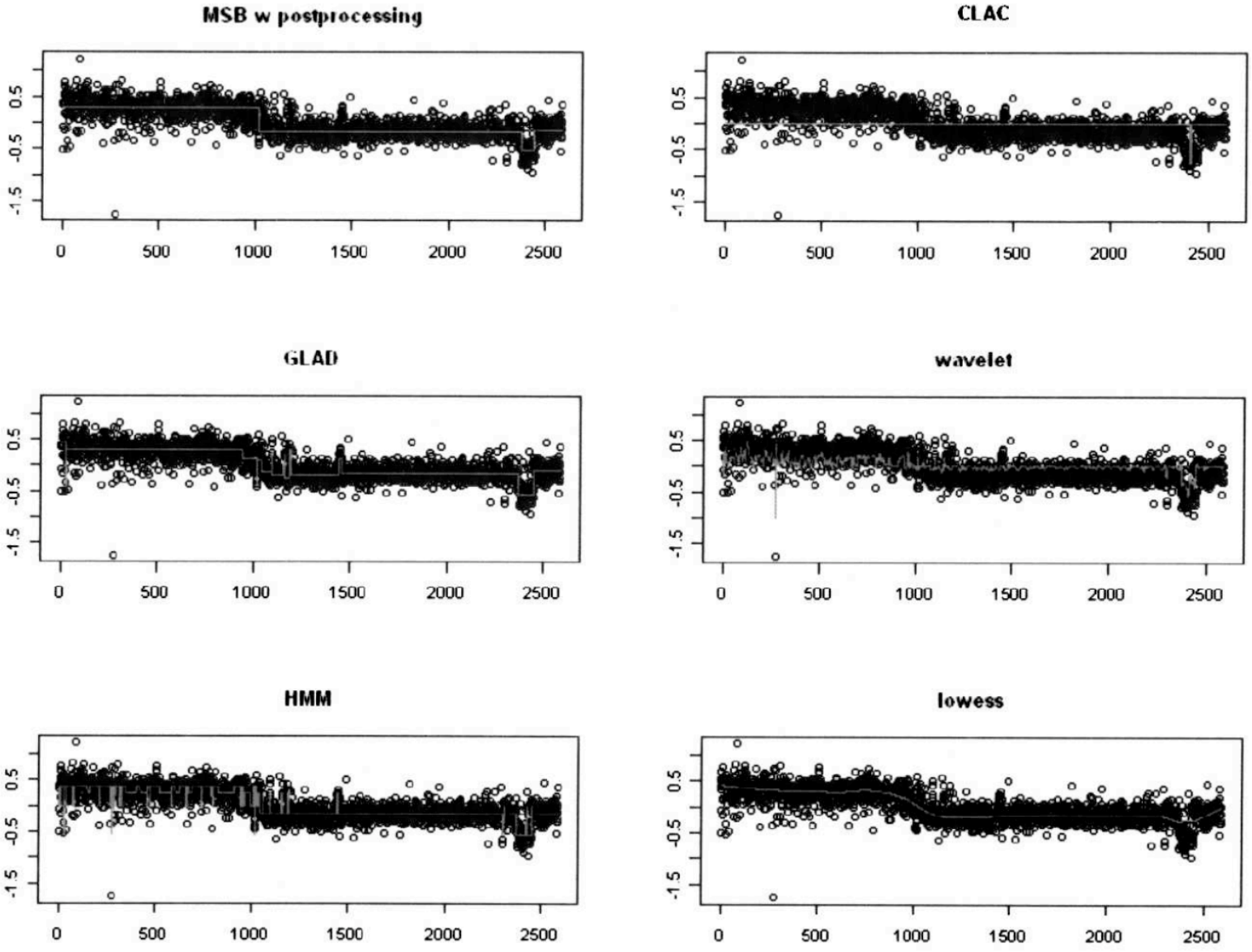


$$RD_{corrected} = \overline{RD}_{global}\ RD/\overline{RD}_{GC}$$

Example of Application of MSB to RD data

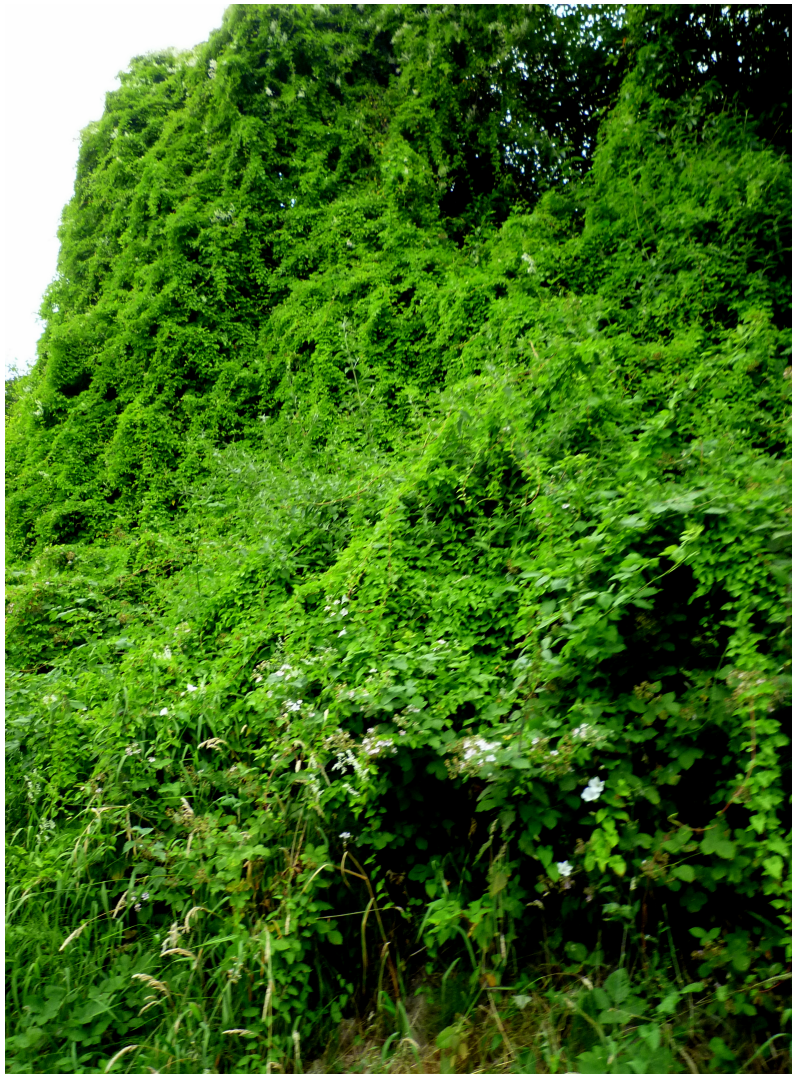


MSB works well on array data too

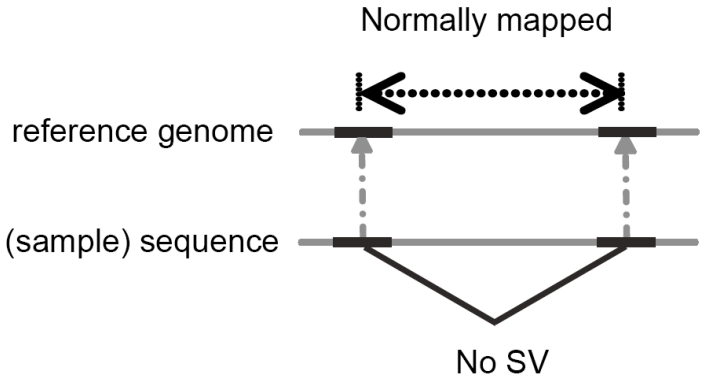


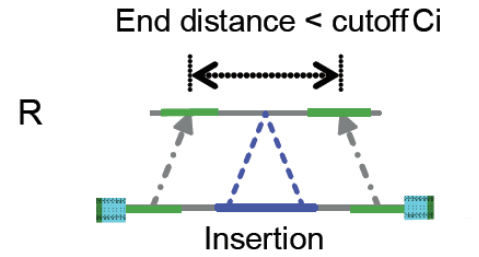
[Wang et al. Gen. Res ('09) 19:106]

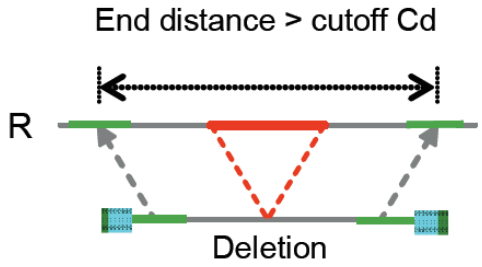
Looking for Aberrantly Placed Paired Ends



PEMer: Detecting Structural Variants from Discordant Paired Ends in Massive Sequencing

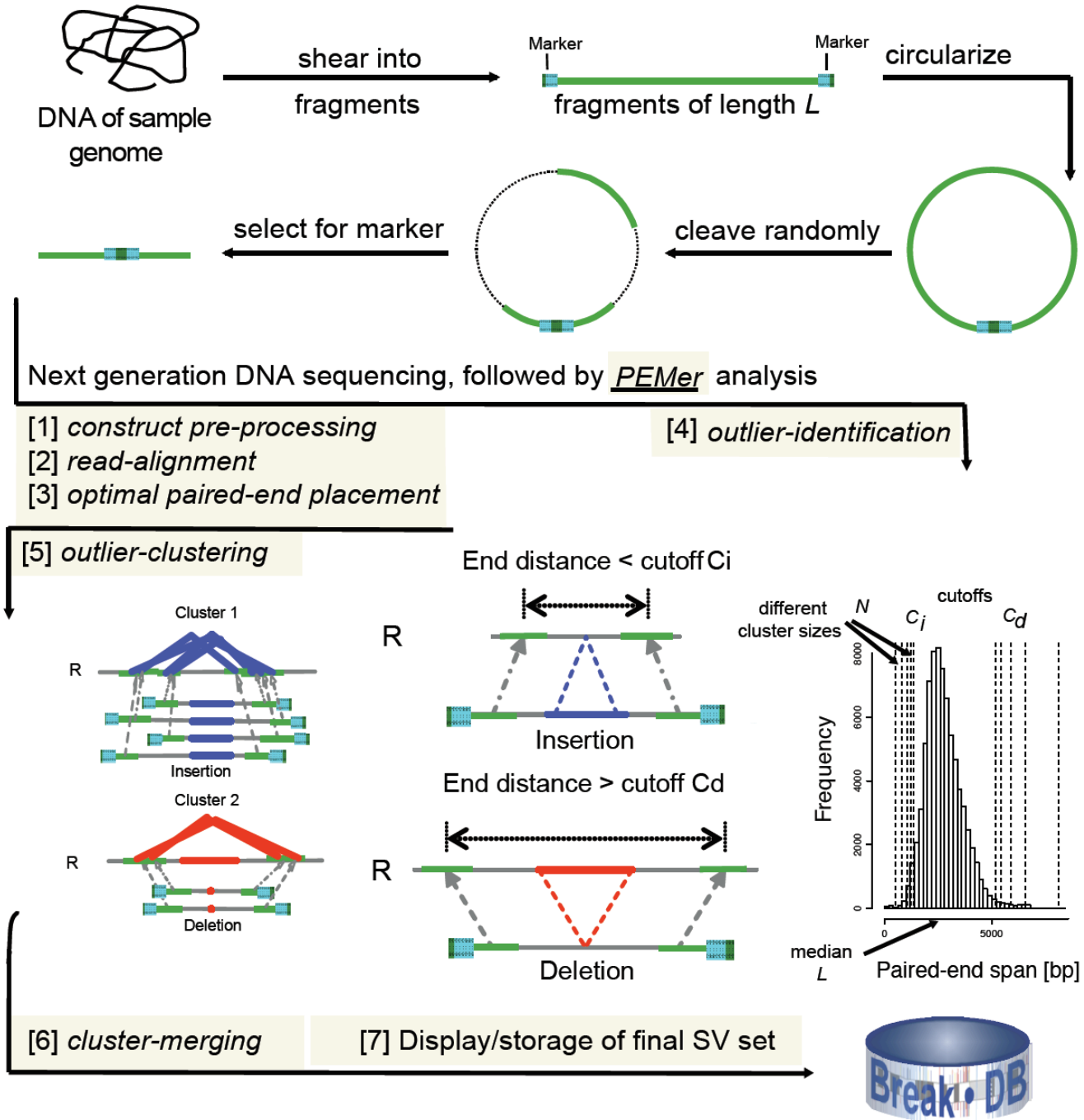






[Korbel et al., Science ('07); Korbel et al., GenomeBiol. ('09)]

Overall Strategy for Analysis of NextGen Seq. Data to Detect Structural Variants



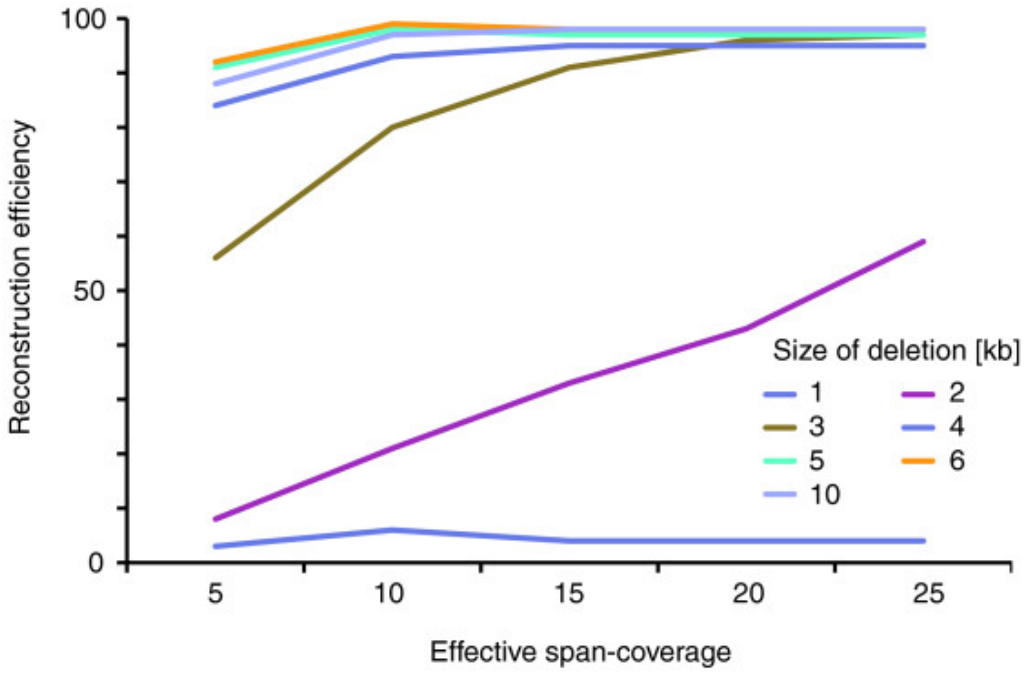
[Korbel et al., Science ('07); Korbel et al., GenomeBiol. ('09)]

Parameterize Error Models through Simulation

Reconstruction efficiency at different coverage

[Korbel et al., GenomeBiol. ('09)]

Deletion size	Reconstruction efficiency at 5x coverage by 2.5 kb inserts
1000	3
2000	11
3000	49
4000	80
5000	91
6000	92
10000	88
Total	414
False positives	5



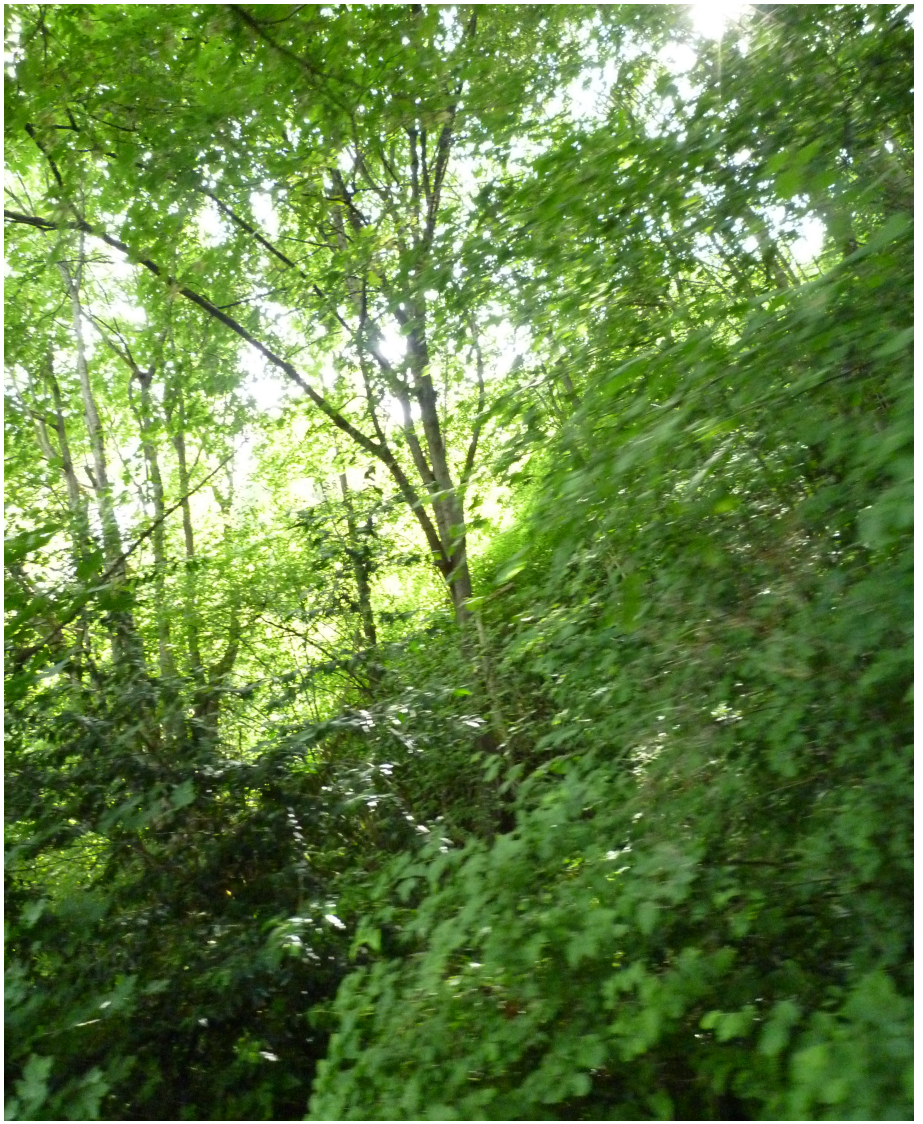
Reconstruction of heterozygous insertions

5x coverage by 2.5 kb inserts		5x coverage by 10 kb inserts	
Insertion size	Reconstruction efficiency	Insertion size	Reconstruction efficiency
250	0	1000	8
500	1	2000	42
750	2	3000	72
1000	1	4000	69
1250	8	5000	61
1500	3	6000	55
1750	3	7000	37
2000	1	8000	23
2250	1	9000	4
2500	0	10000	1
2750	0		
3000	0		
False positives	4		4

Better coverage and fewer reads allow to relax cutoff on outlier lengths and reconstruct more insertions

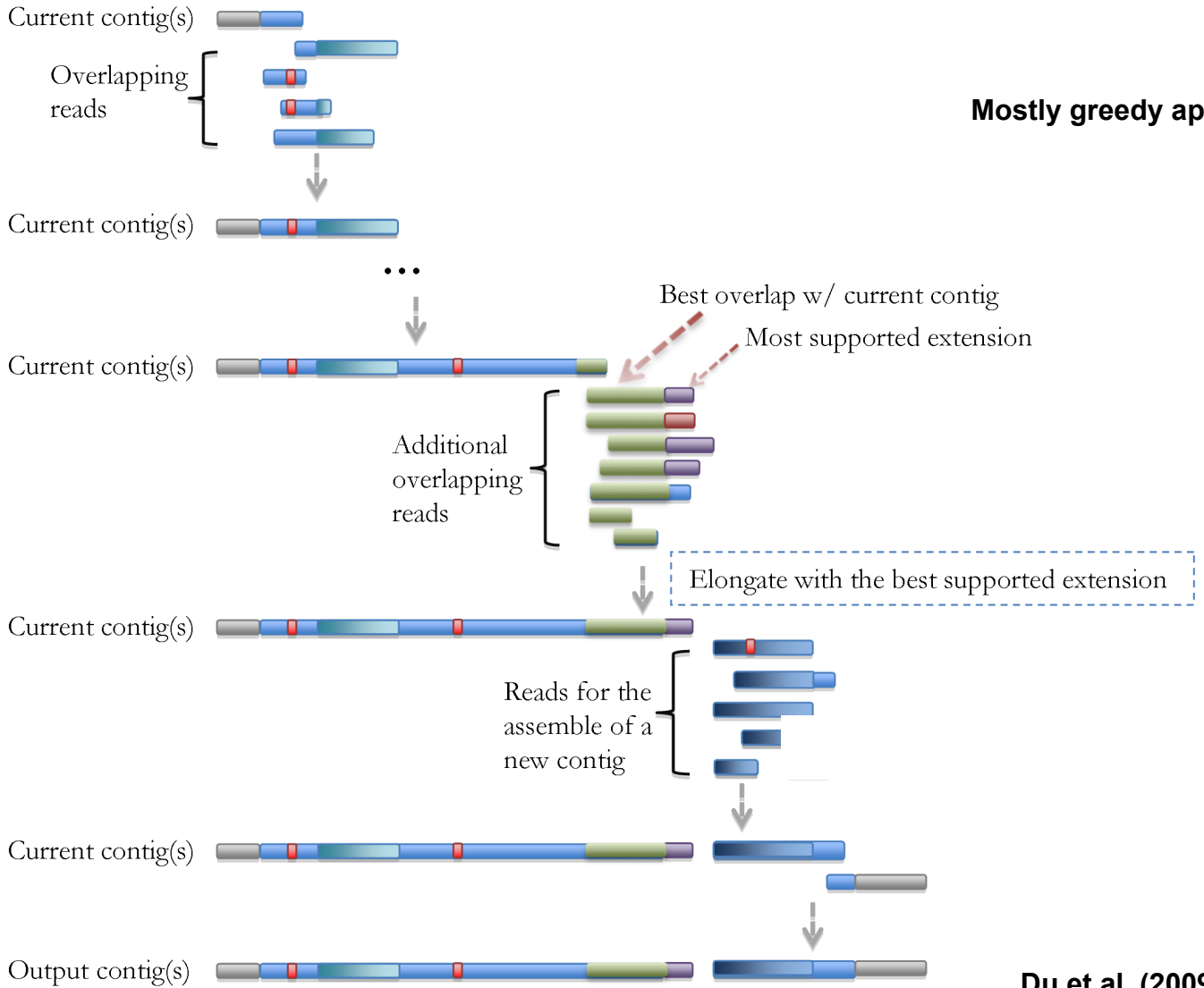
[Korbel et al., GenomeBiol. ('09)]

Local Reassembly



Simple Local Assembly: iterative contig extension

G Iterative contig elongation with the best supported extension

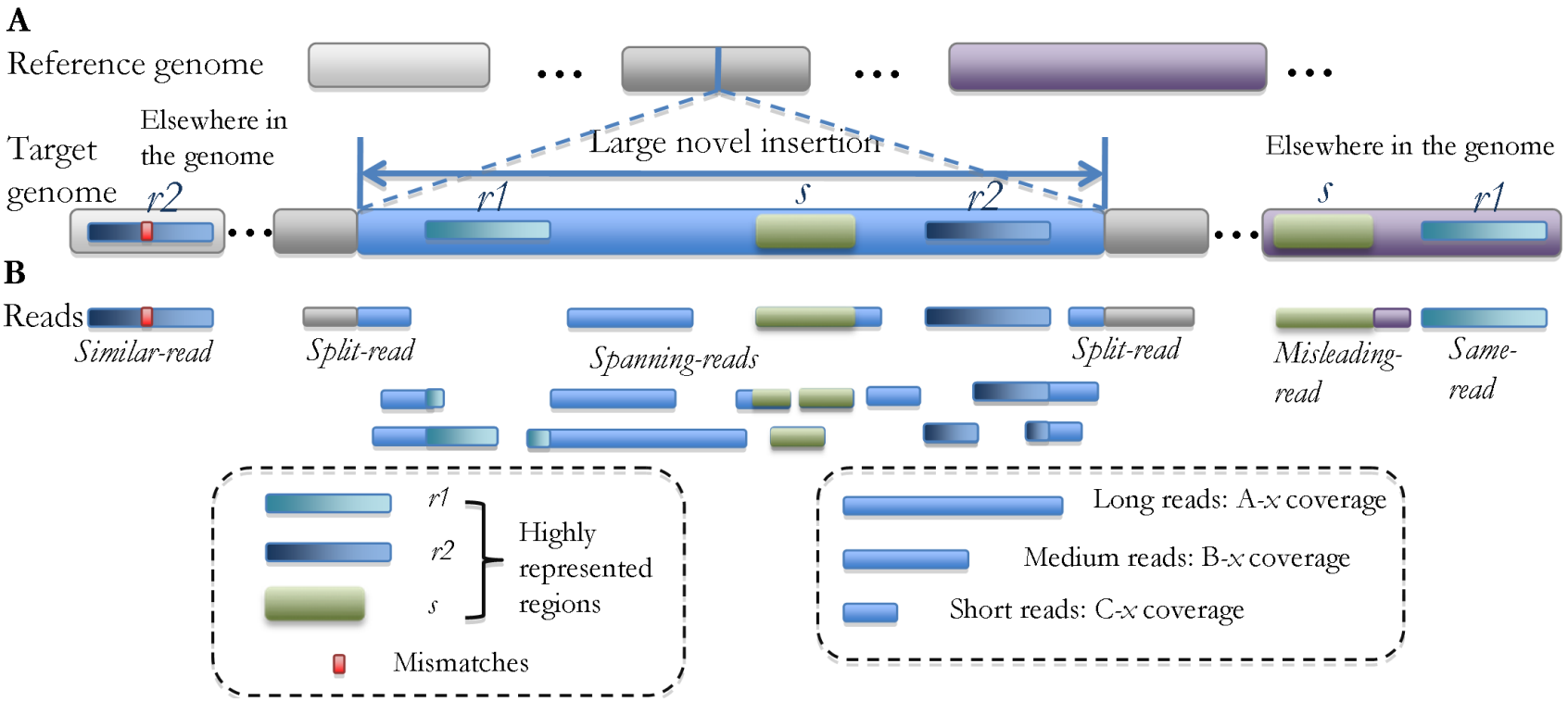


Mostly greedy approach....

Du et al. (2009), PLoS Comp Biol.

Optimal integration of sequencing technologies:
Local Reassembly of large novel insertions

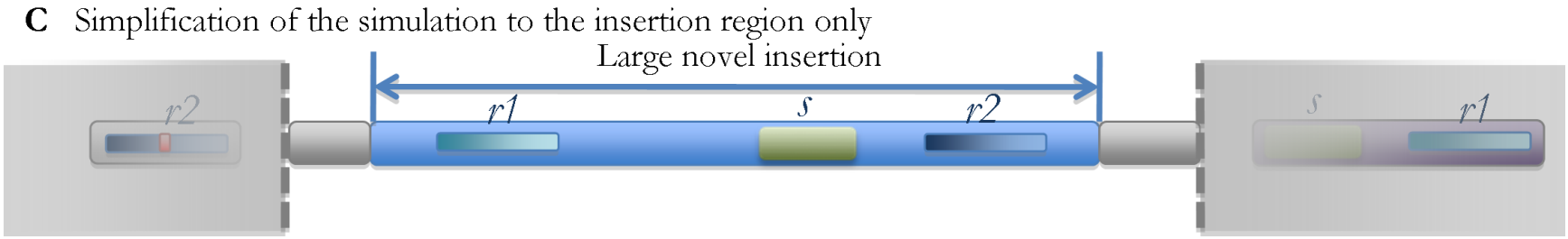
Given a fixed budget, what are the sequencing coverage A, B and C that can achieve the maximum reconstruction rate (on average/worst-case)? Maybe a few long reads can bootstrap reconstruction process.



Du et al. (2009), PLoS Comp Biol, in press

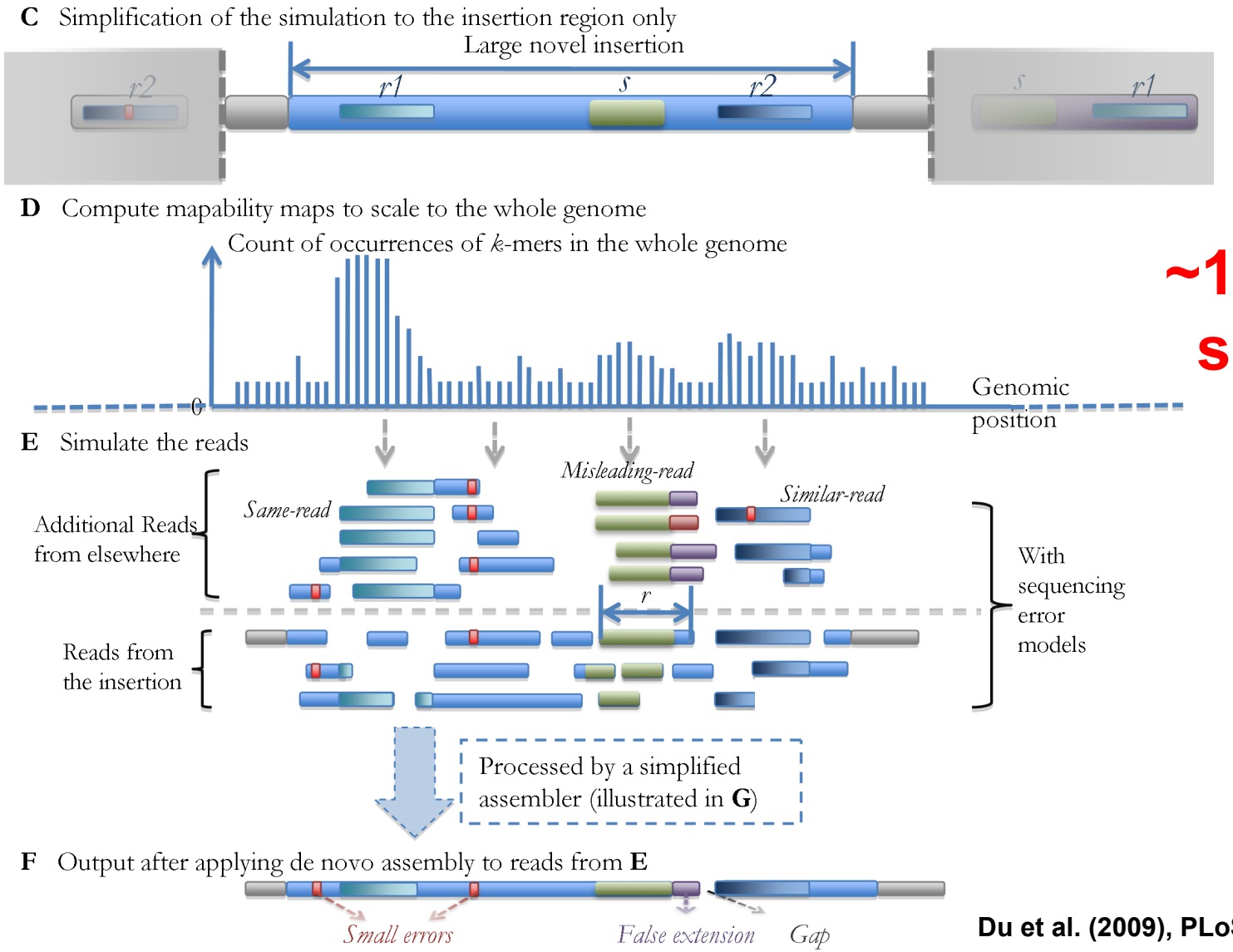
Optimal integration of sequencing technologies:
Need Efficient Simulation

Different combinations of technologies (i.e. read lenghs) very expensive to actually test.
Also computationally expensive to simulate.
(Each round of whole-genome assembly takes >100 CPU hrs; thus, simulation exploring 1K possibilities takes 100K CPU hr)



Du et al. (2009), PLoS Comp Biol, in press

Optimal integration of sequencing technologies:
Efficient Simulation Toolbox using Mappability Maps



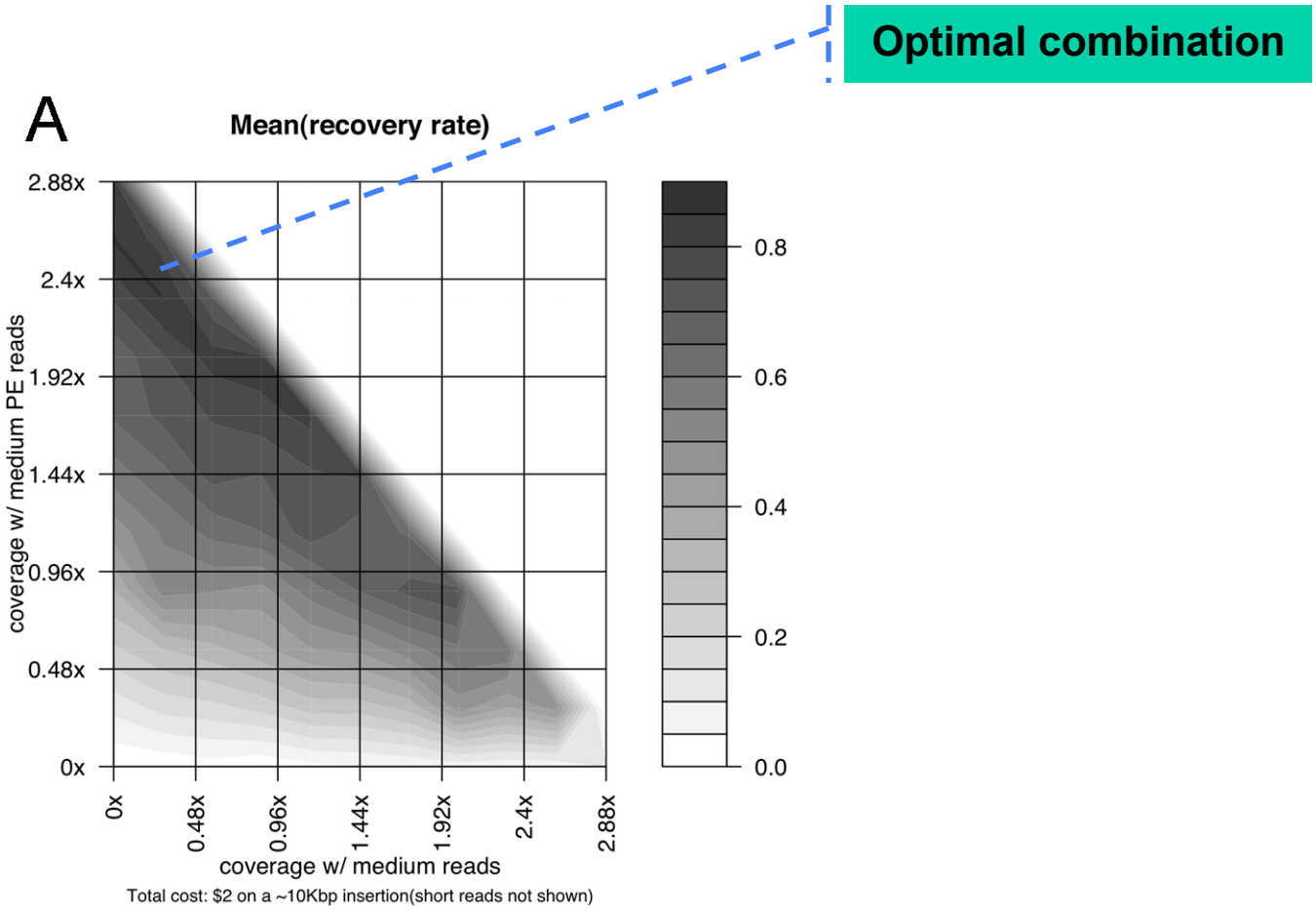
~100,000 X speedup

Du et al. (2009), PLoS Comp Biol, in press

Optimal integration of sequencing technologies:

Simulation shows power of PEs

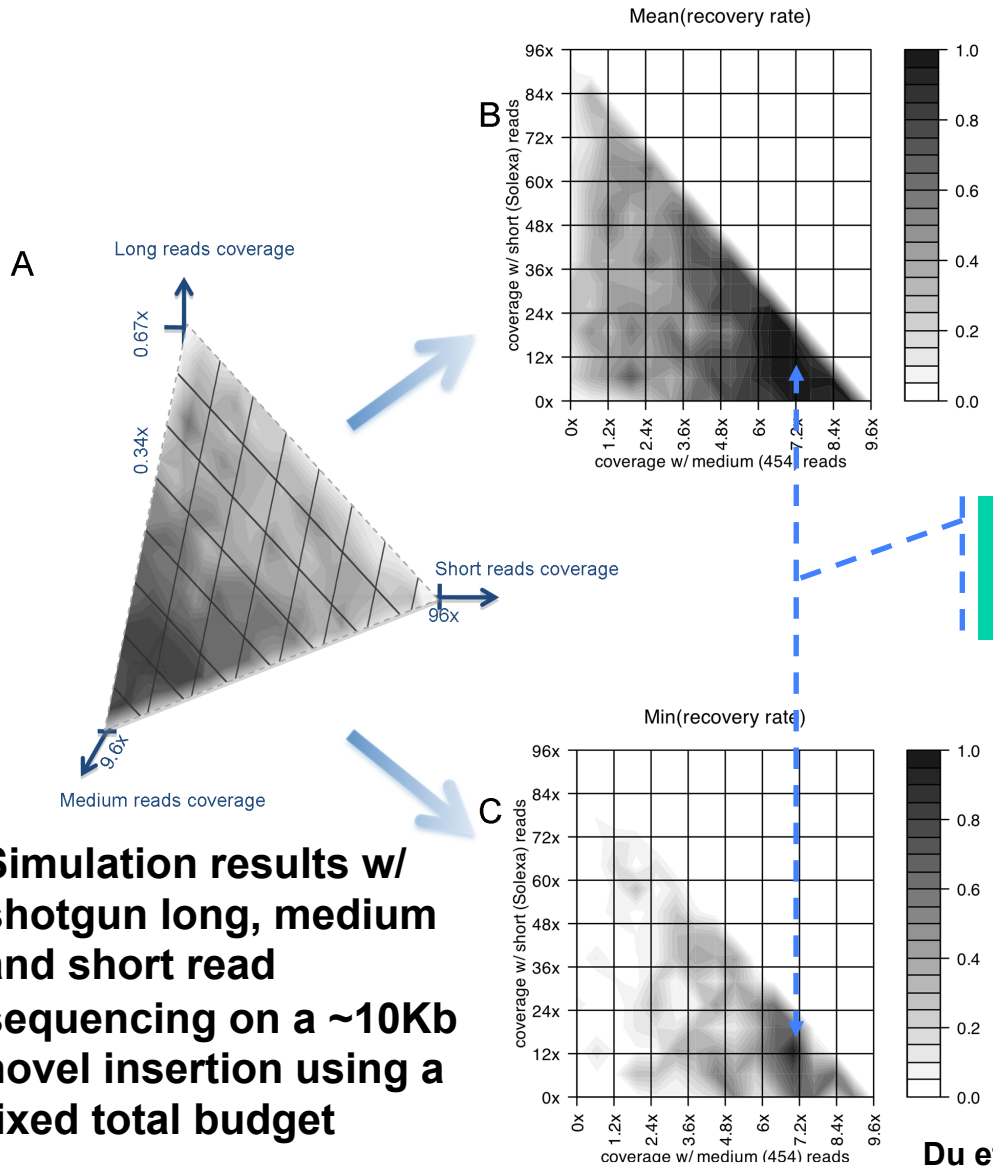
Simulation results w/ shotgun & paired-end reads on the same ~10Kb insertion



Source: Du et al. (2009), PLOS Comp Biol, in press

Optimal integration of sequencing technologies:

Simulation shows combination better than single technology



Simulation results w/ shotgun long, medium and short read sequencing on a ~10Kb novel insertion using a fixed total budget

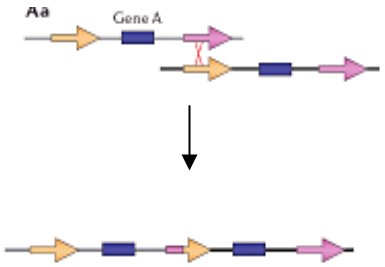
Optimal combination of different technologies

Result dependent on specific parameter setting of different sequencing technologies

Du et al. (2009), PLOS Comp Biol, in press

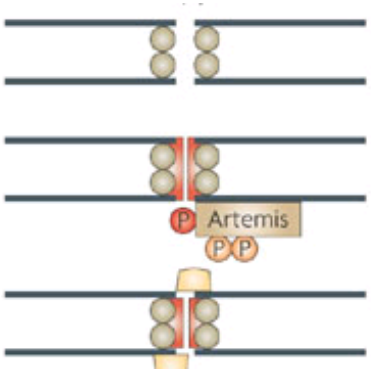
Analyzing Repeated Blocks in the Genome (SDs & CNVs)

SEGMENTAL DUPLCATIONS AND COPY NUMBER VARIANTS ARE RELATED PHENOMENA AND HAVE BEEN CREATED BY SEVERAL DIFFERENT MECHANISMS



NAHR
(Non-allelic homologous recombination)

Flanking repeat
(e.g. Alu, LINE…)

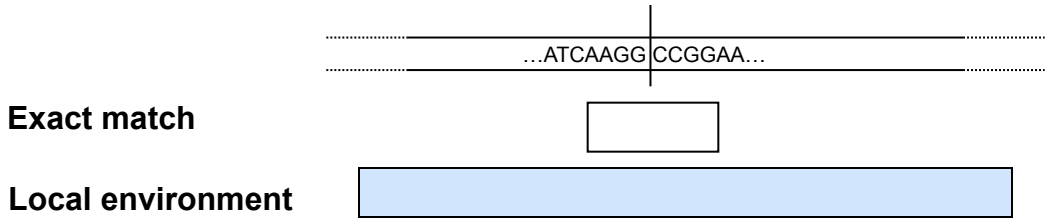


NHEJ
(Non-homologous-end-joining)

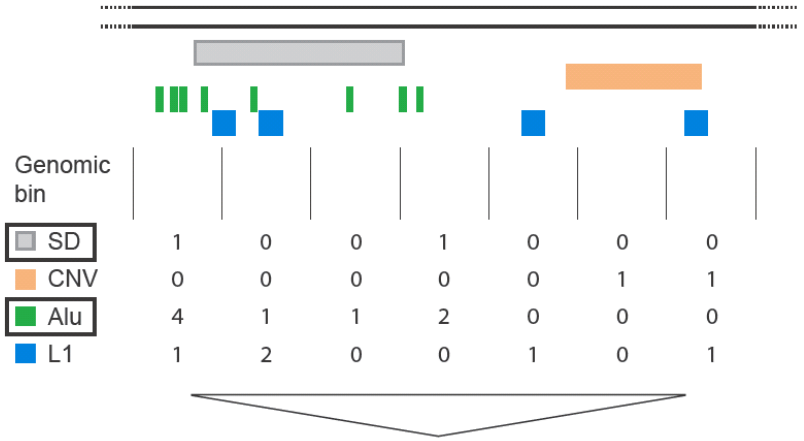
No (flanking) repeats.
In some cases <4bp microhomologies

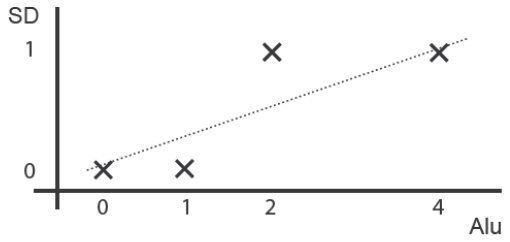
PERFORM LARGE SCALE CORRELATION ANALYSIS TO DETECT REPEAT SIGNATURES OF SDs AND CNVs

If exact CNV breakpoints are known, we can calculate the enrichment of repeat elements relative to the genome or relative to the local environment



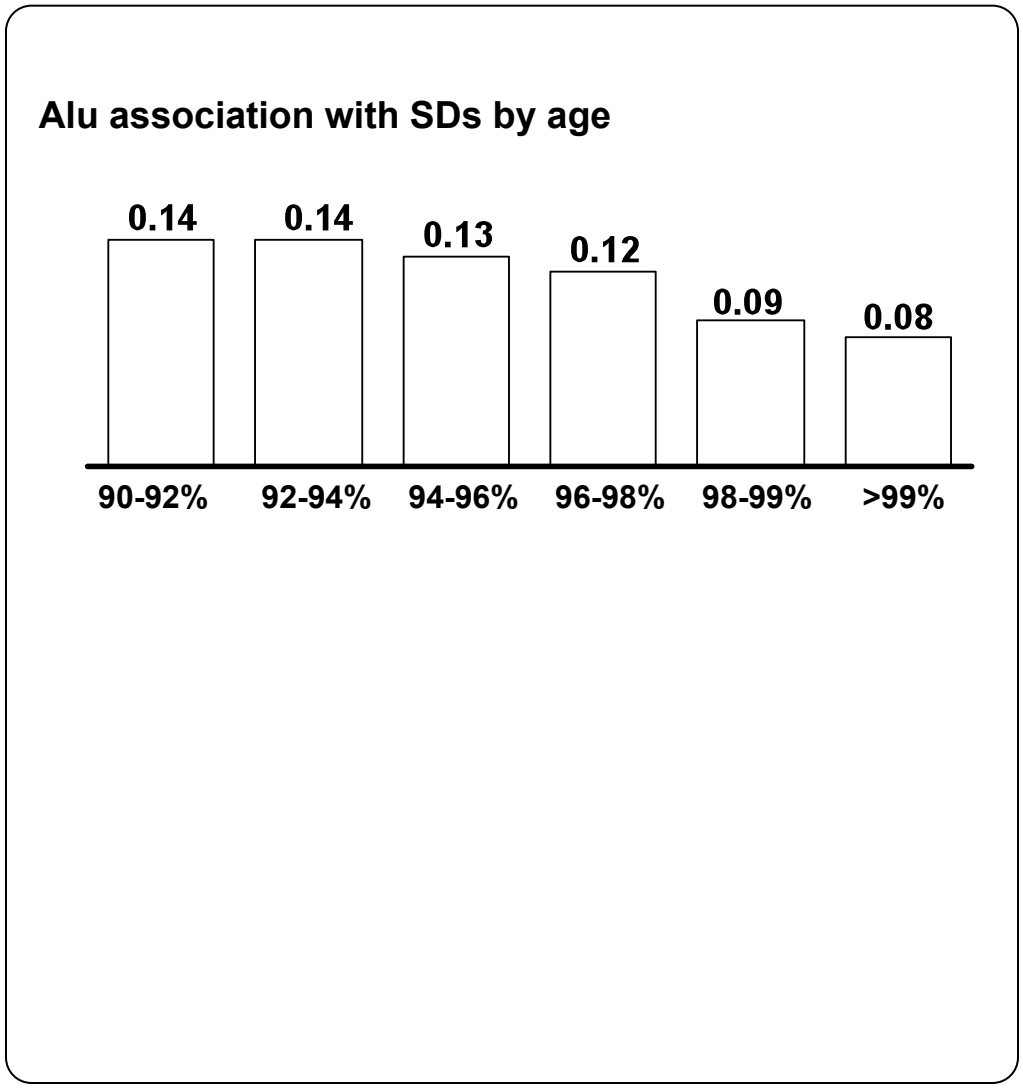
1. Survey a range of genomic features
2. Count the number of features in each genomic bin (100kb)
3. Calculate correlations / enrichments using robust stats





[Kim et al. Gen. Res. ('08), arxiv.org/abs/0709.4200v1]

OLDER SDs ARE MUCH MORE LIKELY TO BE FORMED BY ALU ELEMENTS

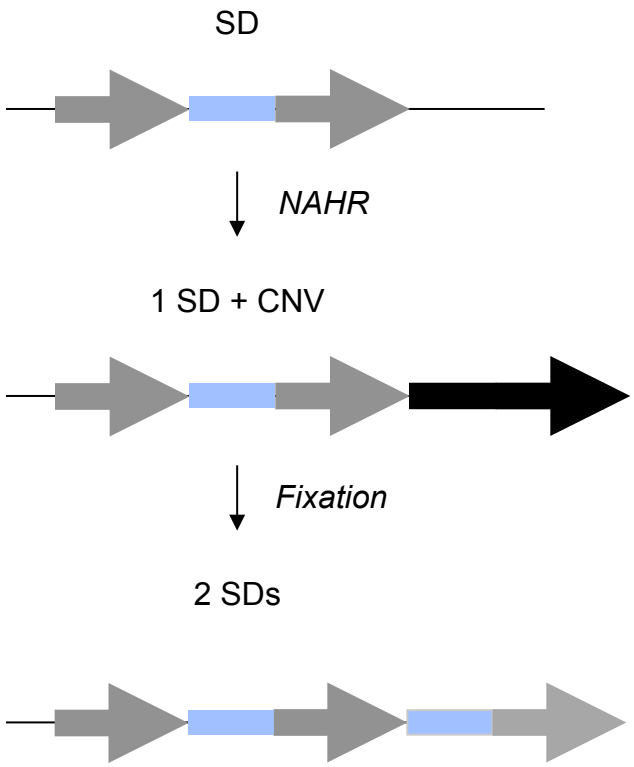


- The co-localization of Alu elements with SDs is highly significant.
- Older SDs have a much higher association with Alus than younger SDs.
- Hence it is likely, that Alu elements were more active in mediating NAHR in the past (consistent with the Alu burst)

[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1]

FOCUSSING ON SDS: SDS CAN PROPAGATE THEMSELVES, WHICH LEADS TO A POWER-LAW DISTRIBUTION

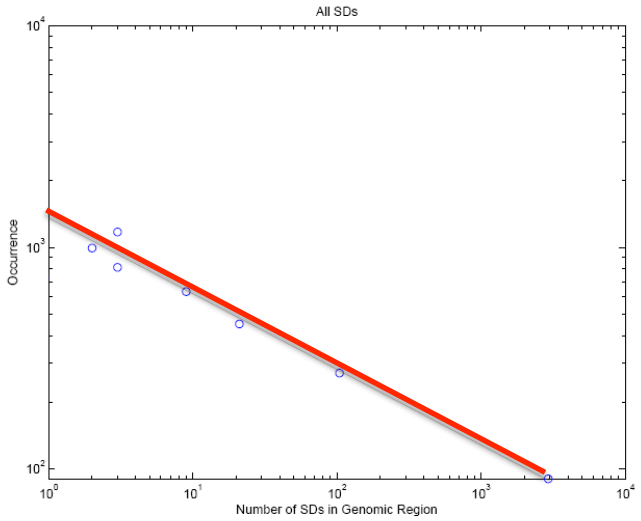
Hypothesis



"SD selfpropagation"

Corollary

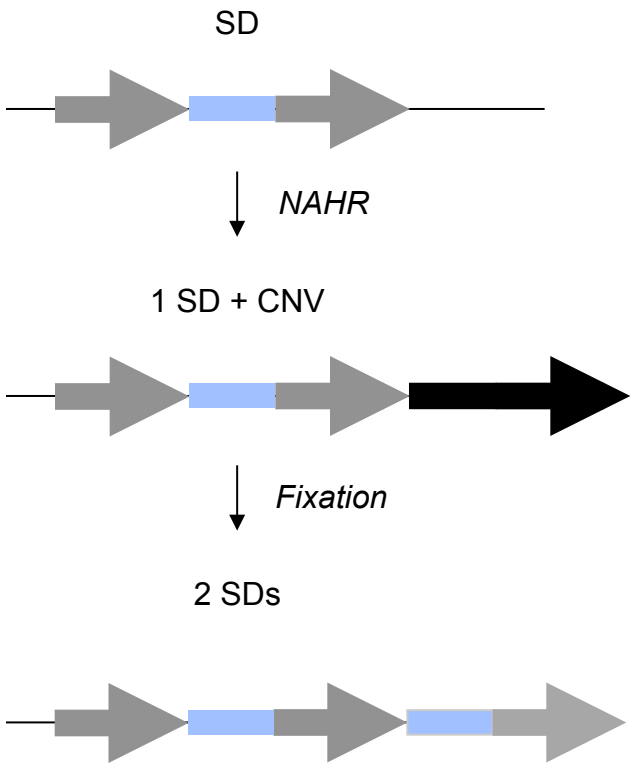
- SDs can mediate NAHR and lead to the formation of CNVs
- CNVs can become fixed and then be SDs
- Such mechanisms ("preferential attachment") are well studied in physics and should leads a very skewed ("power-law") distribution of SDs.



[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1]

FOCUSSING ON SDS: SDs COLOCALIZE WITH EACH OTHER

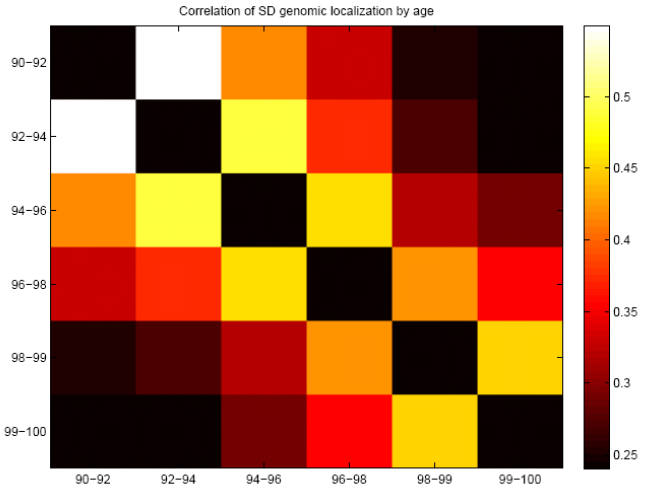
Hypothesis



"SD self-propagation"

Corollary

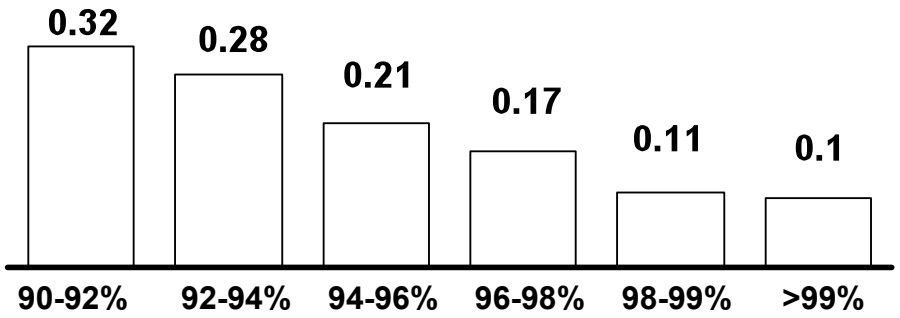
- SDs can mediate NAHR and lead to the formation of CNVs
- CNVs can become fixed and then be SDs
- SDs of similar age should co-localize better with each other:



[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1]

Pseudogenes & CNV/SDs (whole genome, not just encode pilot)

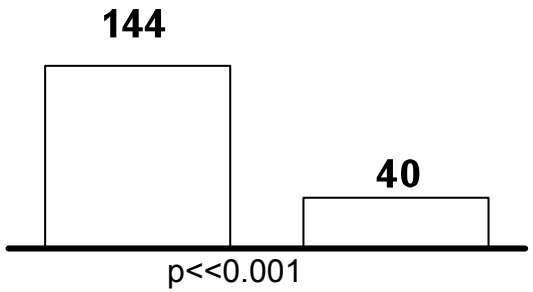
Pseudogene association with SDs by age



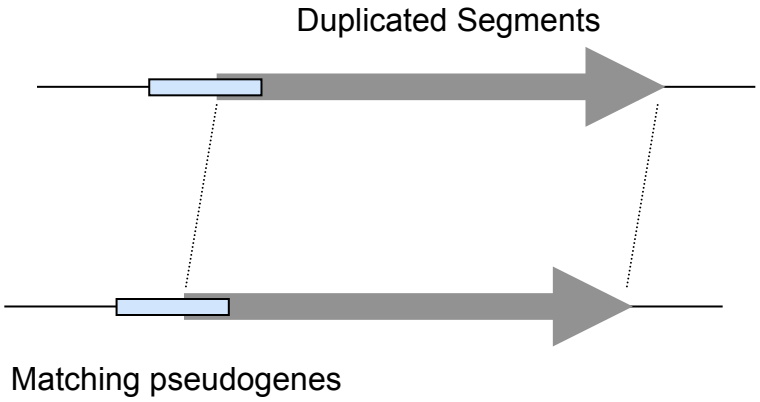
Duplicated pseudogenes associated with SDs, particularly older ones

SDs comprise ~5% of the human genome but contain ~18% genes, 46% duplicated pgenes and 22% processed pgenes

Processed pseudogenes at SD junctions



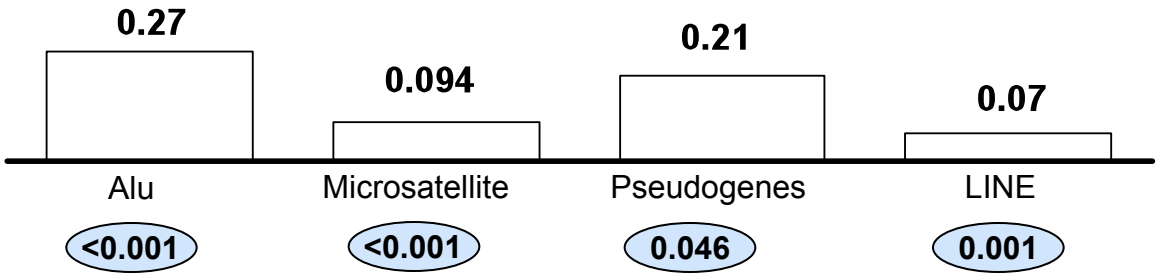
Processed Pseudogenes: serving as repeats for mediating NAHR



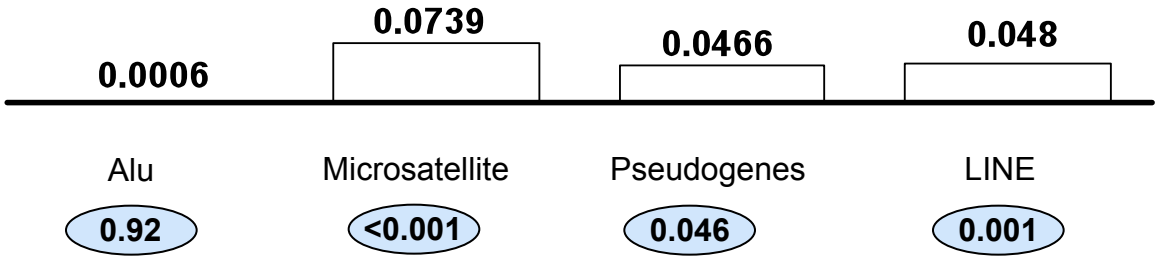
[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1]

ASSOCIATIONS ARE DIFFERENT FOR SDs AND CNVs

SD association with repeats

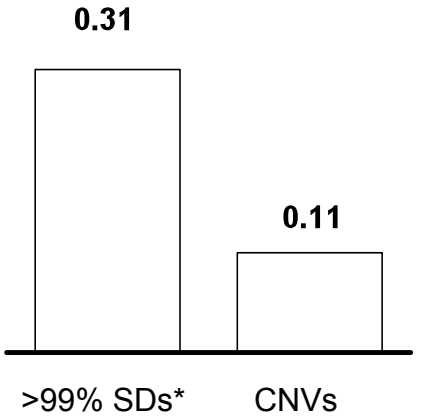


CNV association with repeats

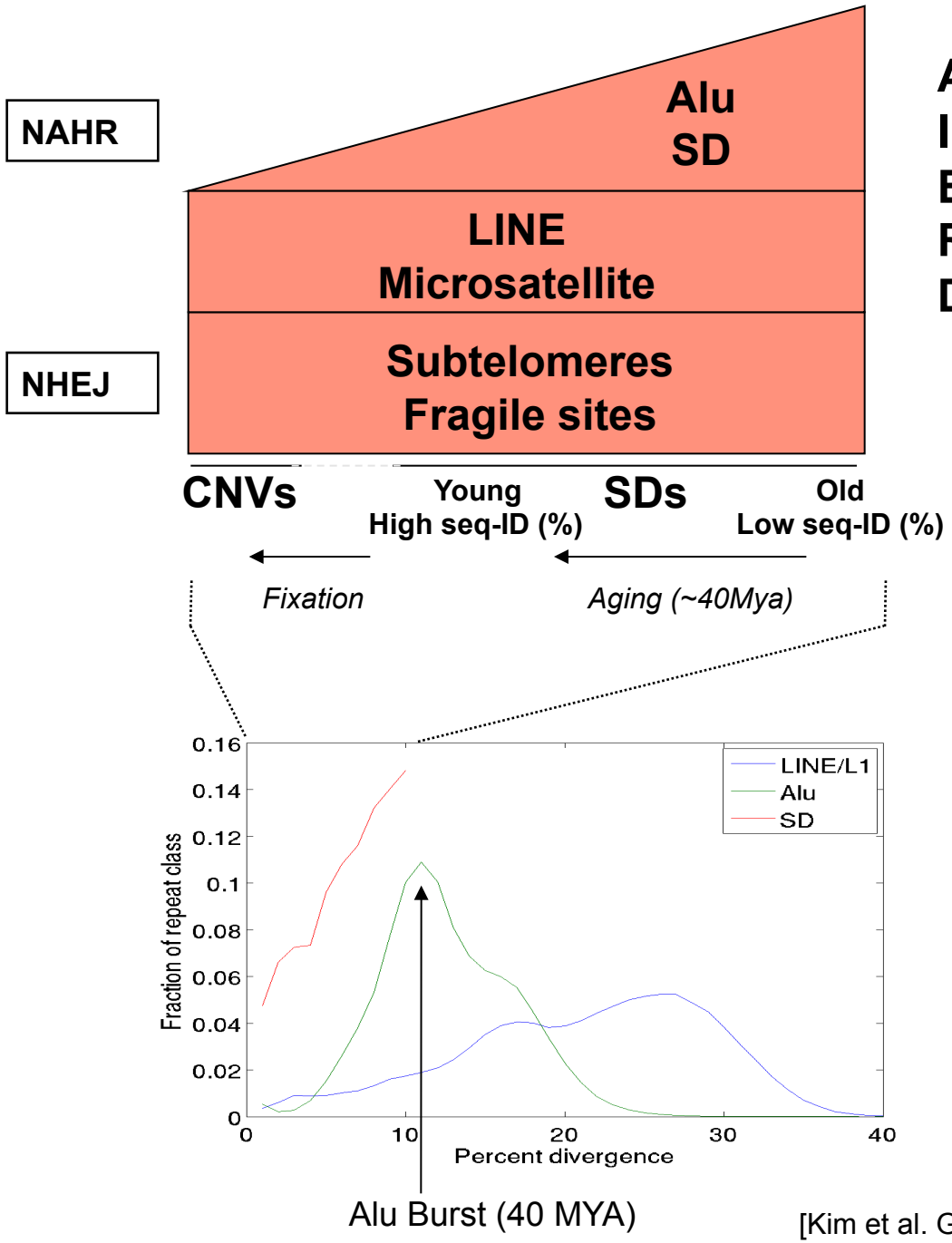


CNVs ARE LESS ASSOCIATED WITH SDs THAN THE GENERAL SD TREND

CNV Association with SDs



[Kim et al. Gen. Res. ('08), arxiv.org/abs/0709.4200v1]



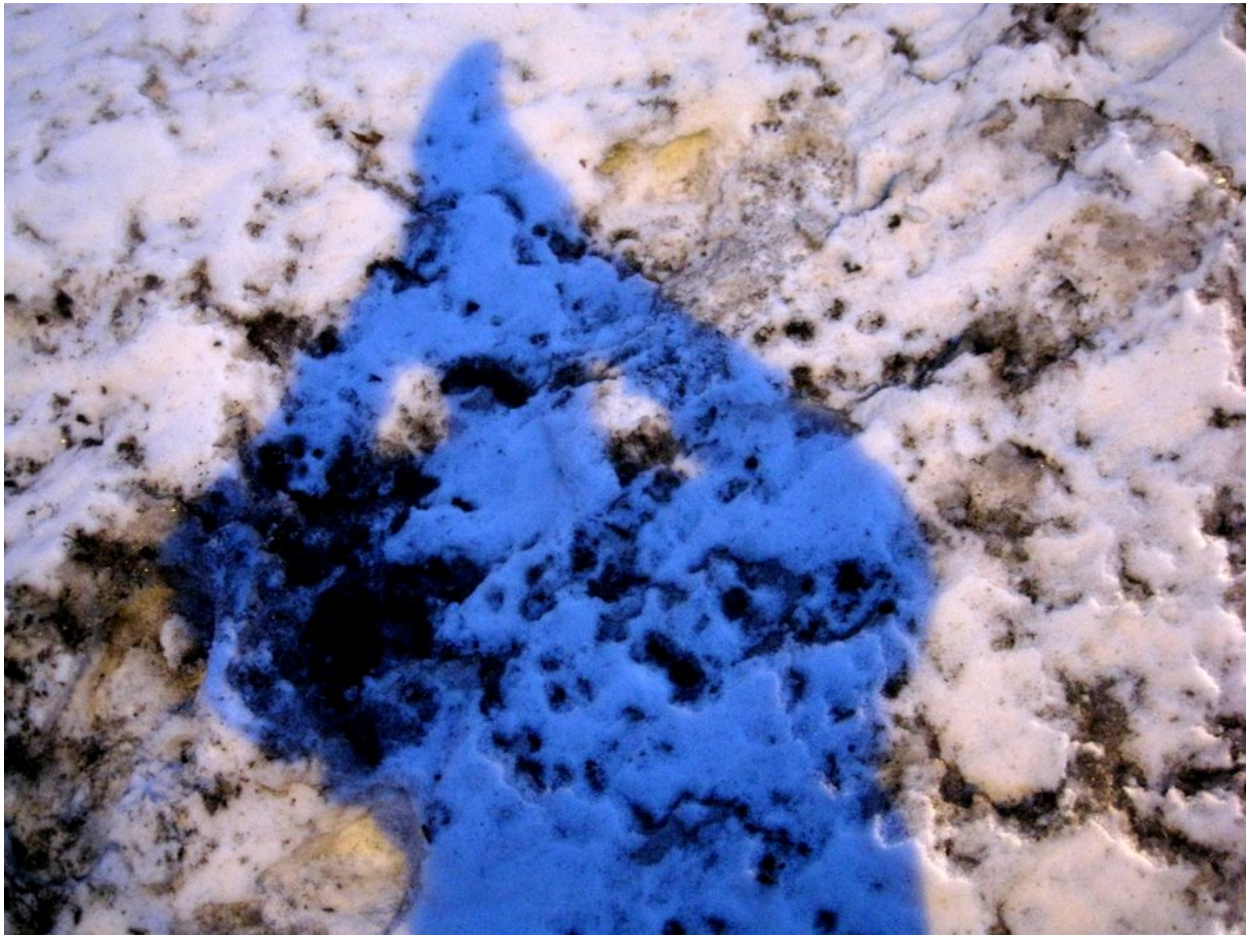
AFTER THE ALU BURST, THE IMPORTANCE OF ALU ELEMENTS FOR GENOME REARRANGEMENT DECLINED RAPIDLY

- About 40 million years ago there was a burst in retrotransposon activity
- The majority of Alu elements stem from that time
- This, in turn, led to rapid genome rearrangement via NAHR
- The resulting SDs, could create more SDs, but with Alu activity decaying, their creation slowed

[Kim et al. Gen. Res. ('08), arxiv.org/abs/0709.4200v1]

Summary:

Looking Back Over the Talk



Identifying Structural Variants in Human Population

- BreakPtr
  - ◊ Model-based segmentation using bivariate HMM
- MSB
  - ◊ Mean-shift segmentation approach following grad. of PDF
  - ◊ Equally applied to aCGH and depth of coverage of short reads
- PEMer
  - ◊ Detecting Variants from discordantly placed paired-ends
  - ◊ Simulation to paramaterize statistical model
- ReSeqSim
  - ◊ Efficiently simulating assembly of a representative variant
  - ◊ Shows that best reconstruction has a combination of long, med. and short reads

Analysis of Duplication in the Genome: SVs and SDs

- Large-scale analysis of existing CNVs & SDs in human genome
- SDs assoc. with Alu, pseudogenes and older SDs
- CNVs assoc. other repeats (microsat.) and not as much with SDs
- Suggestion: Alu burst 40 MYA triggered much NAHR rearrangement, then dupl. feed on itself in hotspots but now dying down and NAHR assoc. with other repeats and CNVs also from NHEJ

YK Lam
J Du
J Korbel
L Wang
P Kim
A Abyzov
M Snyder

X Mu,
D Greenbaum,
A Urban,
P Cayting,
J Rozowsky,
R Bjornson,
S Weissman,
Z Zhang,
S Balasubramanian

GenomeTECH.gersteinlab.org

More Information on this Talk

SUBJECT: GenomeAssembly

DESCRIPTION:
Banbury meeting on Structural Variation in the Human Genome, Lloyd Harbor, NY 2009.11.17, 9:00-9:20; [I:**BANBURYCNV**] (Adaption of GenomeAssembly talk, building on [I:**BIBM**].)

(Works equally well on mac or PC. Paper references in the talk were mostly from Papers.GersteinLab.org. The above topic list can be easily cross-referenced against this website. Each topic abbrev. which is starred is actually a papers "ID" on the site. For instance,
the topic **pubnet*** can be looked up at
http://papers.gersteinlab.org/papers/pubnet)

.

PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as **kwpotppt** , that can be easily queried from flickr, viz: **http://www.flickr.com/photos/mbgmbg/tags/kwpotppt** .