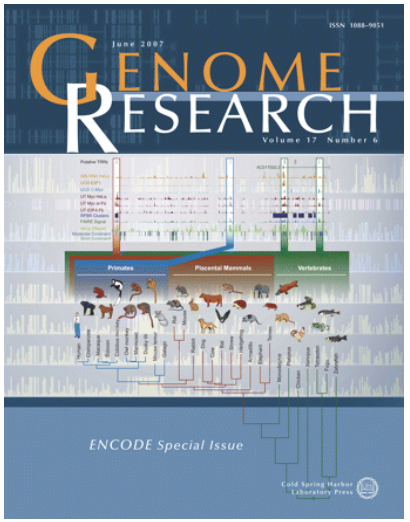
Human Genome Annotation, focusing on SVs

Mark B Gerstein
Yale

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

2001: Most of the genome is not coding (only ~1.2% exon). It consists of elements such as repeats, regulatory regions, non-coding RNAs, origins of replication, pseudogenes, segmental duplications....What do these elements do? How should they be annotated?

[IHGSC, ***Nature*** 409, 2001]
[Venter et al. ***Science*** 29, 2001]



2007 : Pilot results from ENCODE Consortium on decoding what the bases do

- 1% of Genome (30 Mb in 44 regions)
- Tiling Arrays to assay Transcription & Binding
- Multi-organism sequencing and alignment
- Careful Annotation
- Variation Data

[IHGSC, ***Nature*** 409, 2001]
[ENCODE Consortium, ***Nature*** 447, 2007]

Different Views of the Function of Junk DNA

[NY Times, 26-Jun-07]

ESSAY

Human DNA, the Ultimate Spot for Secret Messages (Are Some There Now?)

By DENNIS OVERBYE

In Douglas Adams's science fiction classic, "The Hitchhiker's Guide to the Galaxy," there is a character by the name of Slartibartfast, who designed the fjords of Norway and left his signature in a glacier.

I was reminded of Slartibartfast recently as I was trying to grasp the implications of the feat of a team of Japanese geneticists who announced that they had taught relativity to a bacterium, sort of.

Using the same code that computer keyboards use, the Japanese group, led by Masaru Tomita of Keio University, wrote four copies of Albert Einstein's famous formula, $E=mc^2$, along with "1905," the date that the young Einstein derived it, into the bacterium's genome, the 400-million-long string of A's, G's, T's and C's that determine everything the little bug is and everything it's ever going to be.

The point was not to celebrate Einstein. The feat, they said in a paper published in the journal Biotechnology Progress, was a demonstration of DNA as the ultimate information storage material, able to withstand floods, terrorism, time and the changing fashions in technology, not to mention the ability to be imprinted with little unobtrusive trademark labels — little "Made by Monsanto" tags, say.

In so doing they have accomplished at least a part of the dream that Jaron Lanier, a computer scientist and musician, and David Sulzer, a biologist at Columbia, enunciated in 1999. To create the ultimate time capsule as part of the millennium festivities at this newspaper, they proposed to encode a year's worth of the New York Times magazine into the junk DNA of a cockroach. "The archival cockroach will be a robust repository," Mr. Lanier wrote, "able to survive almost all conceivable scenarios."

If cockroaches can be archives, why not us? The human genome, for example, consists of some 2.9 billion of those letters — the equivalent of about 750 megabytes of data — but only about 3 percent of it goes into composing the 22,000 or so genes that make *us* what we are.

The remaining 97 percent, so-called junk DNA, looks like gibberish. It's the dark matter of inner space. We don't know what it is saying to or about us, but within that sea of megabytes there is plenty of room for the imagination to roam, for trademark labels and much more. The King James Bible, to pick one obvious example, only amounts to about five megabytes.

Jimmy Turrell

If a bacterium can be encoded with $E=mc^2$, if cockroaches can be archives, why not us?

Inevitably, if you are me, you begin to wonder if there is already something written in the warm wet archive, whether or not some Slartibartfast has already been here and we ourselves are walking around with little trademark tags or more wriggling and squiggling and folded inside us. Gill Bejerano, a geneticist at the University of California, Santa Cruz, who mentioned Slartibartfast to me, pointed out that the problem with raising this question is that people who look will see messages in the genome even if they aren't there — the way people have claimed in recent years to have found secret codes in the Bible.

Nevertheless, no less a personage than Francis Crick, the co-discoverer of the double helix, writing with the chemist Leslie Orgel, now at the Salk Institute in San Diego, suggested in 1973 that the primitive Earth was infected with DNA broadcast through space by an alien species.

As a result, it has been suggested that the search for extraterrestrial intelligence, or SETI, should look inward as well as outward. In an article in New Scientist, Paul Davies, a cosmologist at Arizona State University,

change, and have remained identical in humans, rats, mice, chickens and dogs for at least 300 million years.

But Dr. Bejerano, one of the discoverers of these "ultraconserved" strings of the genome, said that many of them had turned out to be playing important command and control functions.

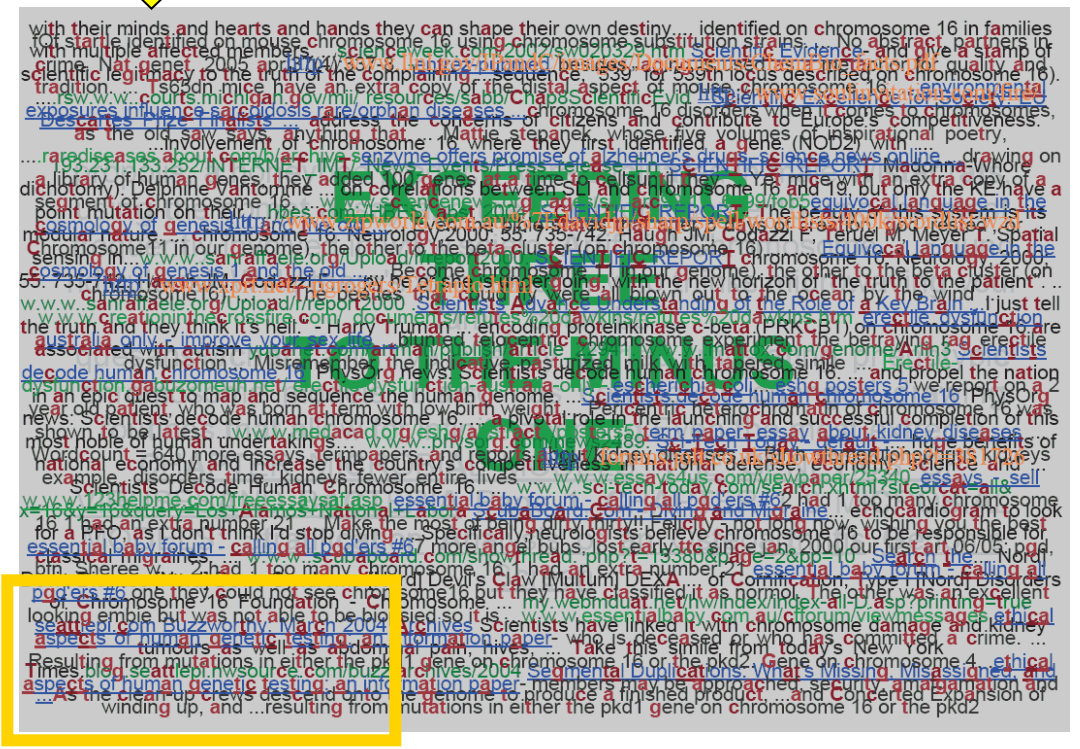
"Why they need to be so conserved remains a mystery," he said, noting that even regular genes that do something undergo more change over time. Most junk bits of DNA that neither help nor annoy an organism mutate even more rapidly.

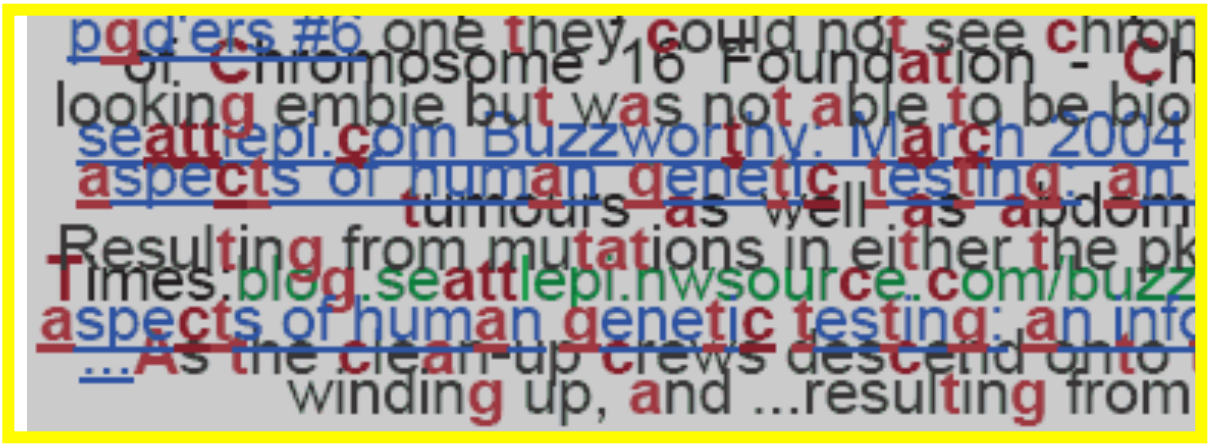
The Japanese team proposed to sidestep the mutation problem by inserting redundant copies of their message into the genome. By comparing the readouts, they said, they would be able to recover Einstein's formula even when up to 15 percent of the original letters in the string had changed, or mutated. "This is the major point

sections of junk DNA seem to be markedly resistant to

Slartibartfast.

Using the same code that computer keyboards use, the Japanese group... wrote four copies of Albert Einstein's famous formula, E=mc2... into the bacterium's genome... In so doing they have accomplished at least a part of the dream that Jaron Lanier, a computer scientist and musician, and David Sulzer, a biologist at Columbia, enunciated in 1999. To create the ultimate time capsule as part of the millennium festivities at this newspaper, they proposed to encode a year's worth of the New York Times magazine into the junk DNA of a cockroach. "The archival cockroach will be a robust repository," Mr. Lanier wrote, "able to survive almost all conceivable scenarios."





Junk DNA as Art

How might we annotate a human text?

Color is Function

Lines are Similarity

The Semicolon Wars

Brian Hayes

IF YOU WANT TO BE a thoroughgoing world traveler, you need to learn 6,912 ways to say "Where is the toilet, please?" That's the number of languages known to be spoken by the peoples of planet Earth, according to Ethnologue.com.

If you want to be the complete polyglot programmer, you also have quite a challenge ahead of you, learning all the ways to say:

```
printf("hello, world\n");
```

(This one is in C.) A catalog maintained by Bill Kinnersley of the University of Kansas lists about 2,500 programming languages. Another survey, compiled by Diarmuid Piggott, puts the total even higher, at more than 8,500. And keep in mind that whereas human languages have had millennia to evolve and diversify, all the computer languages have sprung up in just 50 years. Even by the more-conservative standards of the Kinnersley count, that means we've been inventing one language a week, on average, ever since Fortran.

For ethnologists, linguistic diversity is a cultural resource to be nurtured and preserved, much like biodiversity.

Every programmer knows there is one true programming language. A new one every week

a good-enough notation—for expressing an algorithm or defining a data structure.

There are programmers of my acquaintance who will dispute that last statement. I expect to hear from them. They will argue—zealously, ardently, vehemently—that we have indeed found the right programming language, and for me to claim otherwise is willful ignorance. The one true language may not yet be perfect, they'll concede, but it's built on a sound foundation and solves the main problems, and now we should all work together to refine and improve it. The catch, of course, is that each of these friends will

cide which end of a boiled egg to crack. This famous tempest in an egg cup was replayed 250 years later by designers of computer hardware and communications protocols. When a block of data is stored or transmitted, either the least-significant bit or the most-significant bit can go first. Which way is better? It hardly matters, although life would be easier if everyone made the same choice. But that's *not* what has happened, and so quite a lot of hardware and software is needed just to swap ends at boundaries between systems.

This modern echo of Swift's Endian wars was first pointed out by Danny Cohen of the University of Southern California in a brilliant 1980 memo, "On holy wars and a plea for peace." The memo, subsequently published in *Computer*, was widely read and admired; the plea for peace was ignored.

Another feud—largely forgotten, I think, but never settled by truce or treaty—focused on the semicolon. In Algol and Pascal, program statements have to be separated by semicolons. For example, in `x:=0; y:=x+1; z:=2` the semicolons tell the compiler where one statement ends and the next begins. C

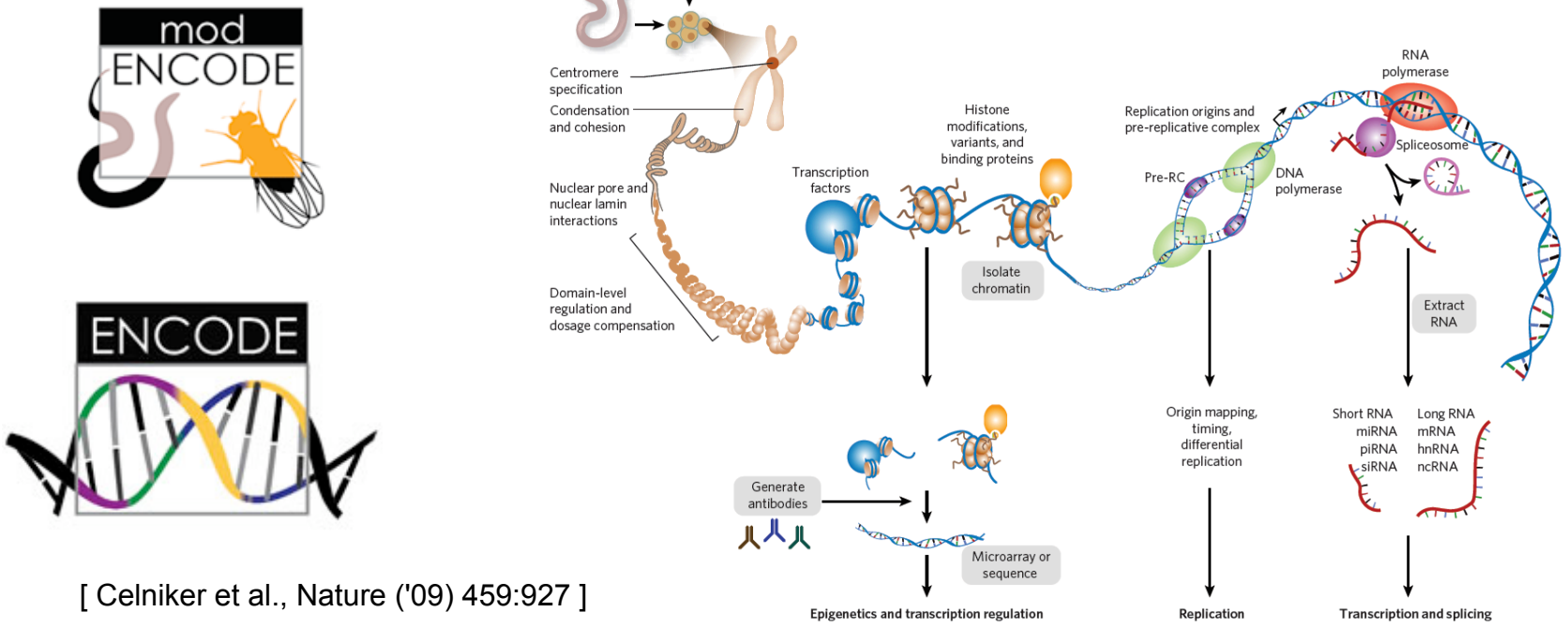
[B Hayes, Am. Sci. (Jul.- Aug. '06)]

Overview of the Process of Annotation of non-coding Regions

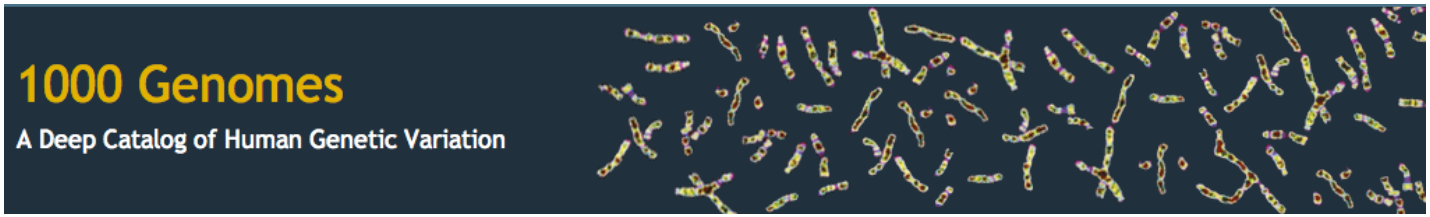
- Basic Inputs
  1. Comparative Genomics.
     Doing large-scale similarity comparison,
     looking for repeated or deleted regions
  2. Functional Genomics.
     Determining experimental signals for activity (e.g. transcription)
     across each base of genome
- **Comparative Genomics**

  Finding repeated or deleted blocks in the genome
  1. As a function of similarity (i.e. age, perhaps using explicit models)
  2. vs. other organisms, vs. human reference, or within the human population
     (synteny, SDs, and CNVs)
  3. Big and small blocks
     (duplicated regions and retrotransposed repeats)
  4. Creation of formal annotations (e.g. genes and pseudogenes)

ENCODE + modENCODE Consortia for functional annotation & 1KG Consortium for variable blocks in human population



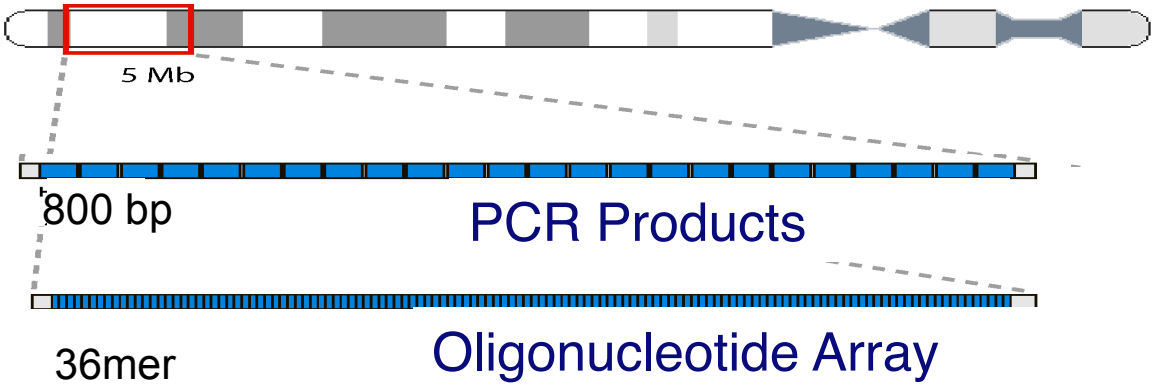
[Celniker et al., Nature ('09) 459:927]



Technologies used for Interrogating the Human Genome, over the past 6 years: Reading out "active" or "tagged" regions

Tiling Arrays

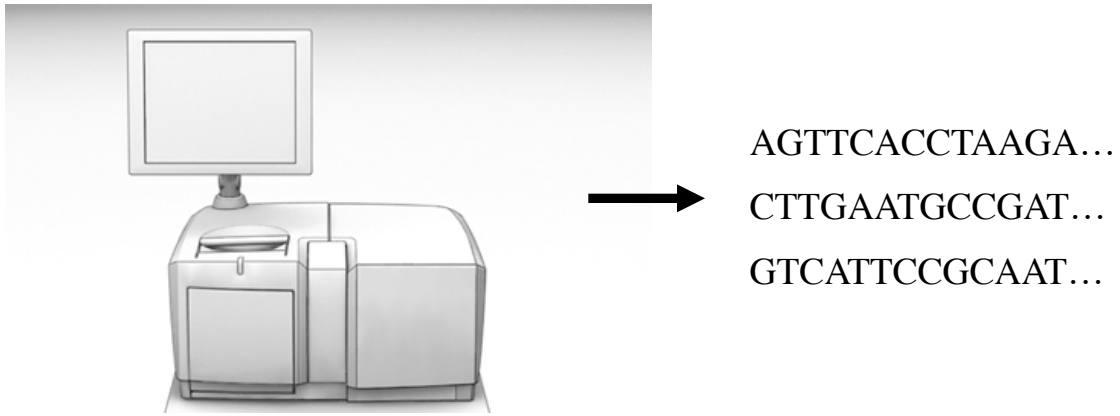
‘02



‘04

Massively Parallel Sequencing

‘06+



Application in a variety of contexts:

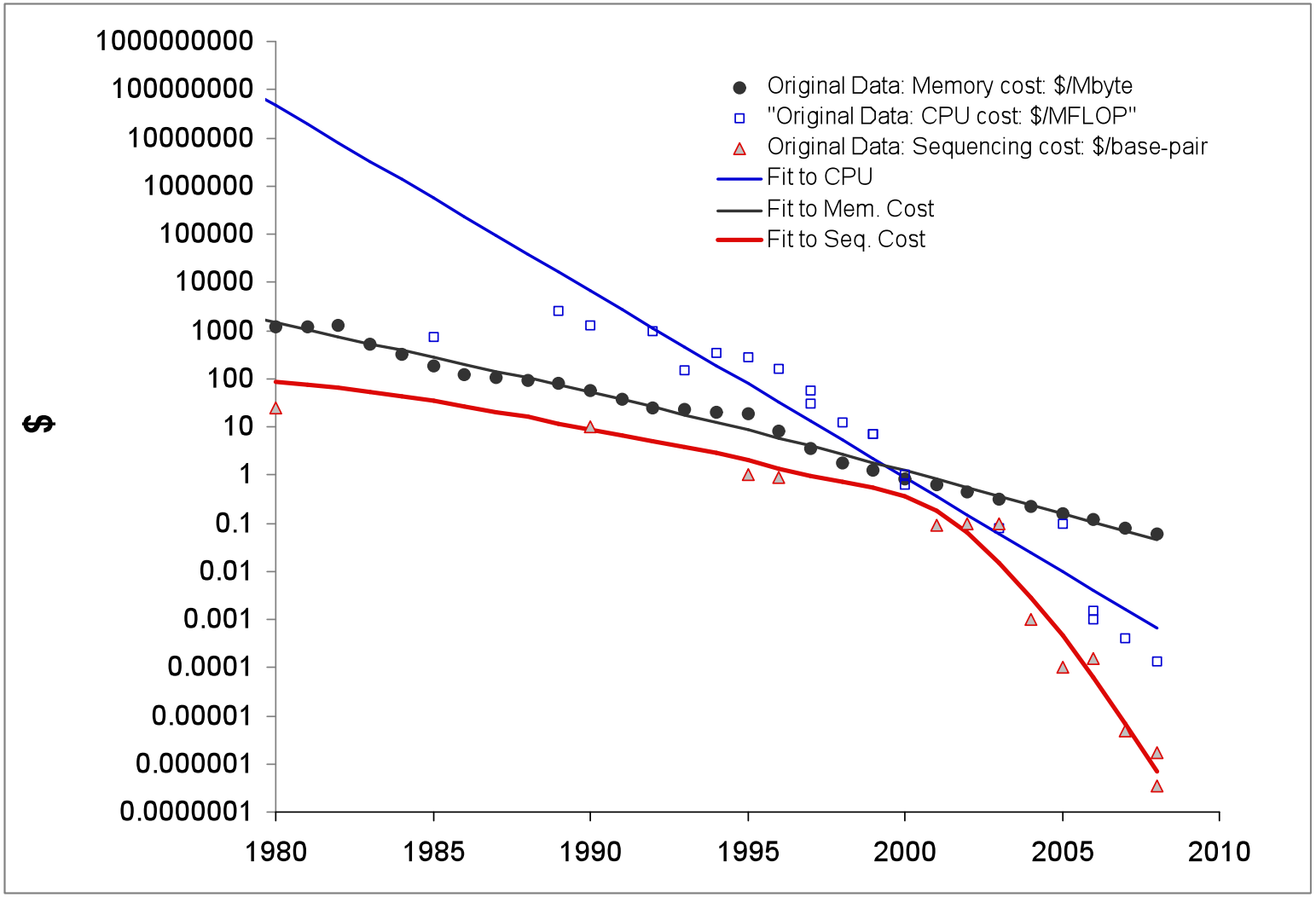
Transcription Mapping

DNA binding (inc. chromatin struc.)

Replication

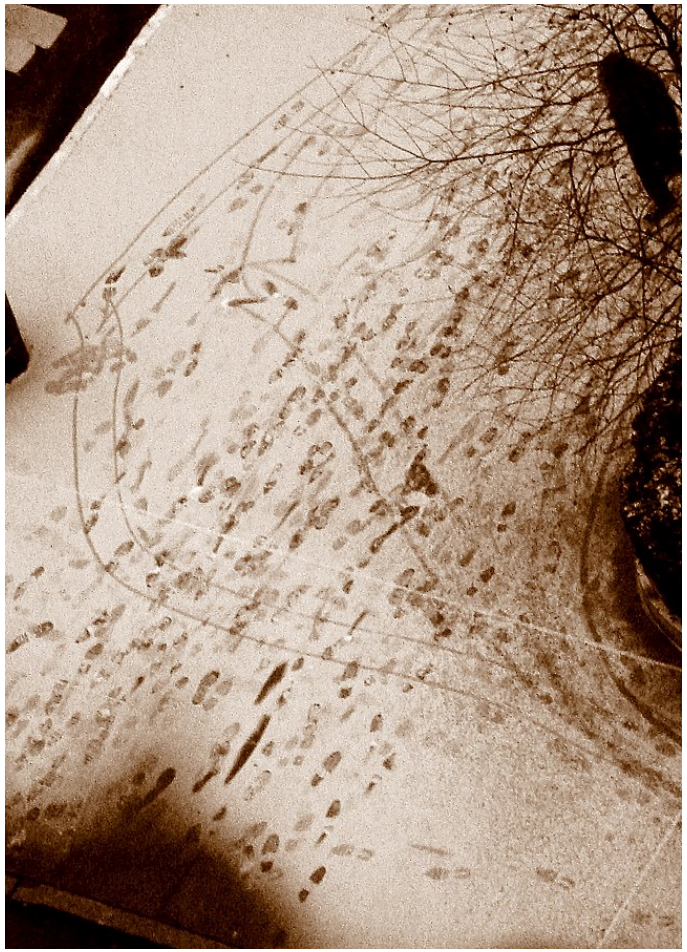
Structural Variation

Plummeting Cost of Sequencing



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

Outline



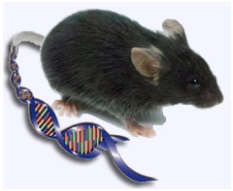
- Calling Variable Blocks in Genome (CNVs,SDs)
  - Calling them with various signal processing approaches
- Analyzing Association of Variable Blocks with repeats, in relation to formation mechanisms

Terminology for Variable Duplicated Elements in the Human Genome

Blocks in both mouse & human

CNVs (variable blocks in the human population)

Fixed SDs ("fixed" blocks in the human population)



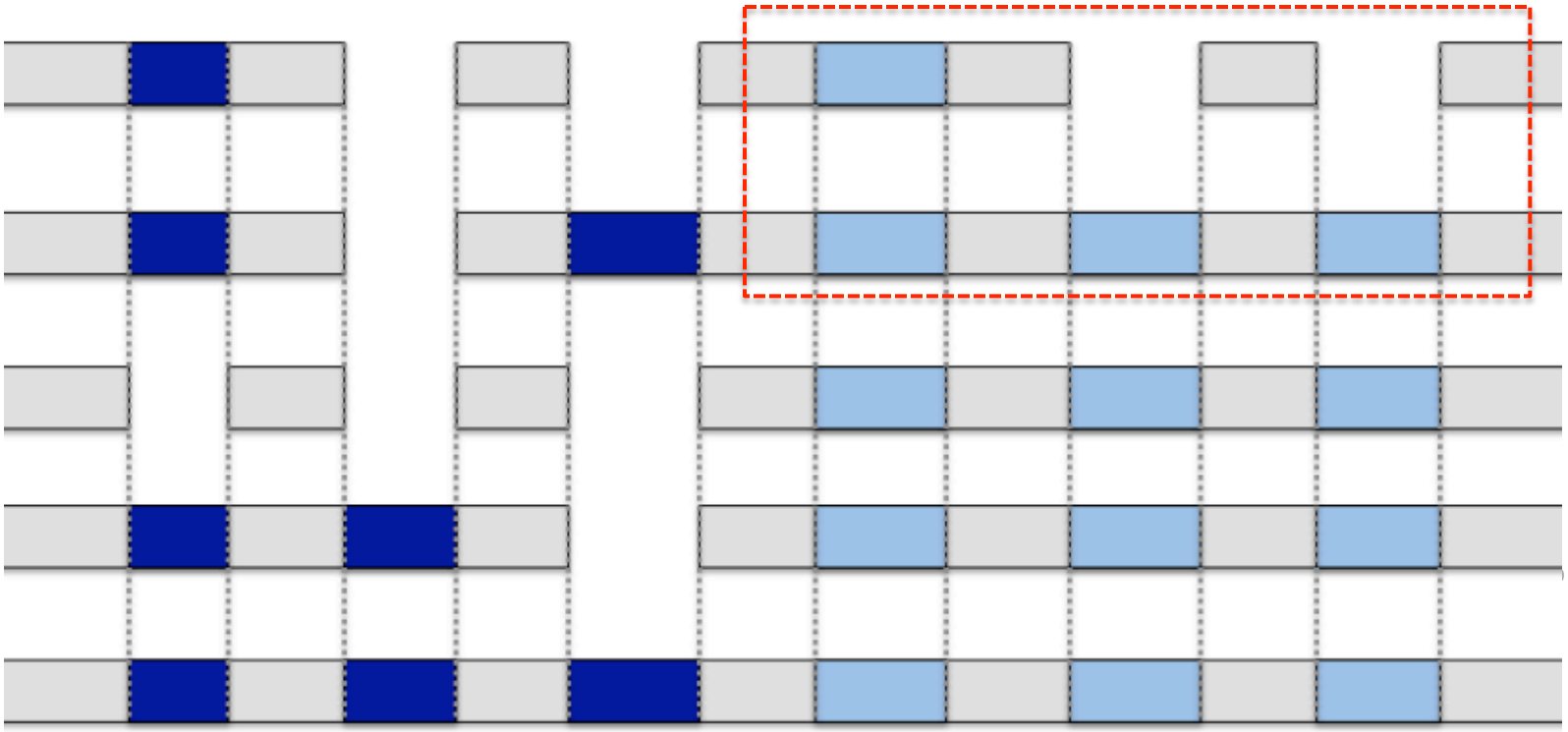
ref.

Human ref.

Venter

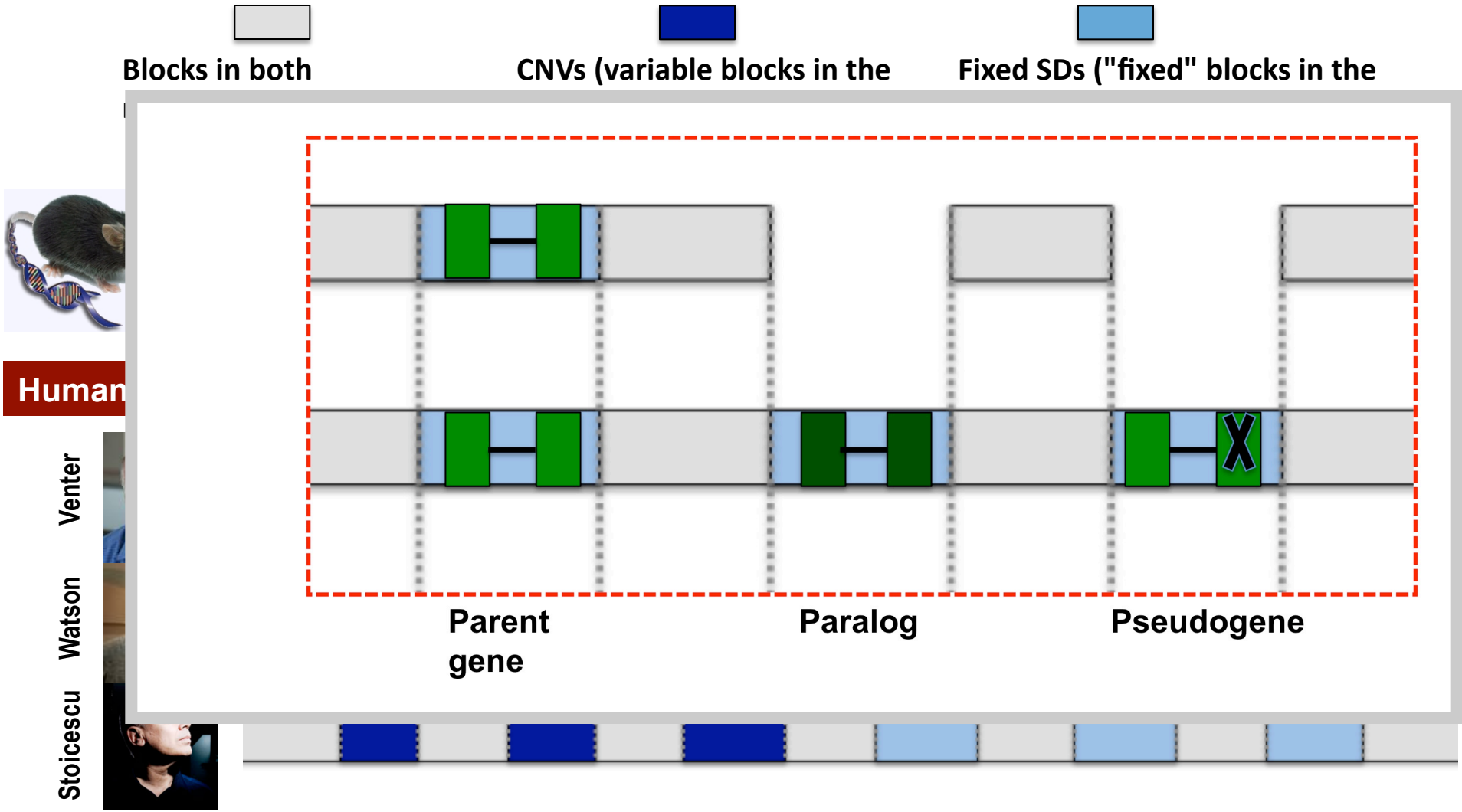
Watson

Stoicescu



Segmental duplications (SDs) - Recent duplications (~40 million years and younger)

Terminology for Variable Duplicated Elements in the Human Genome



Human

Venter

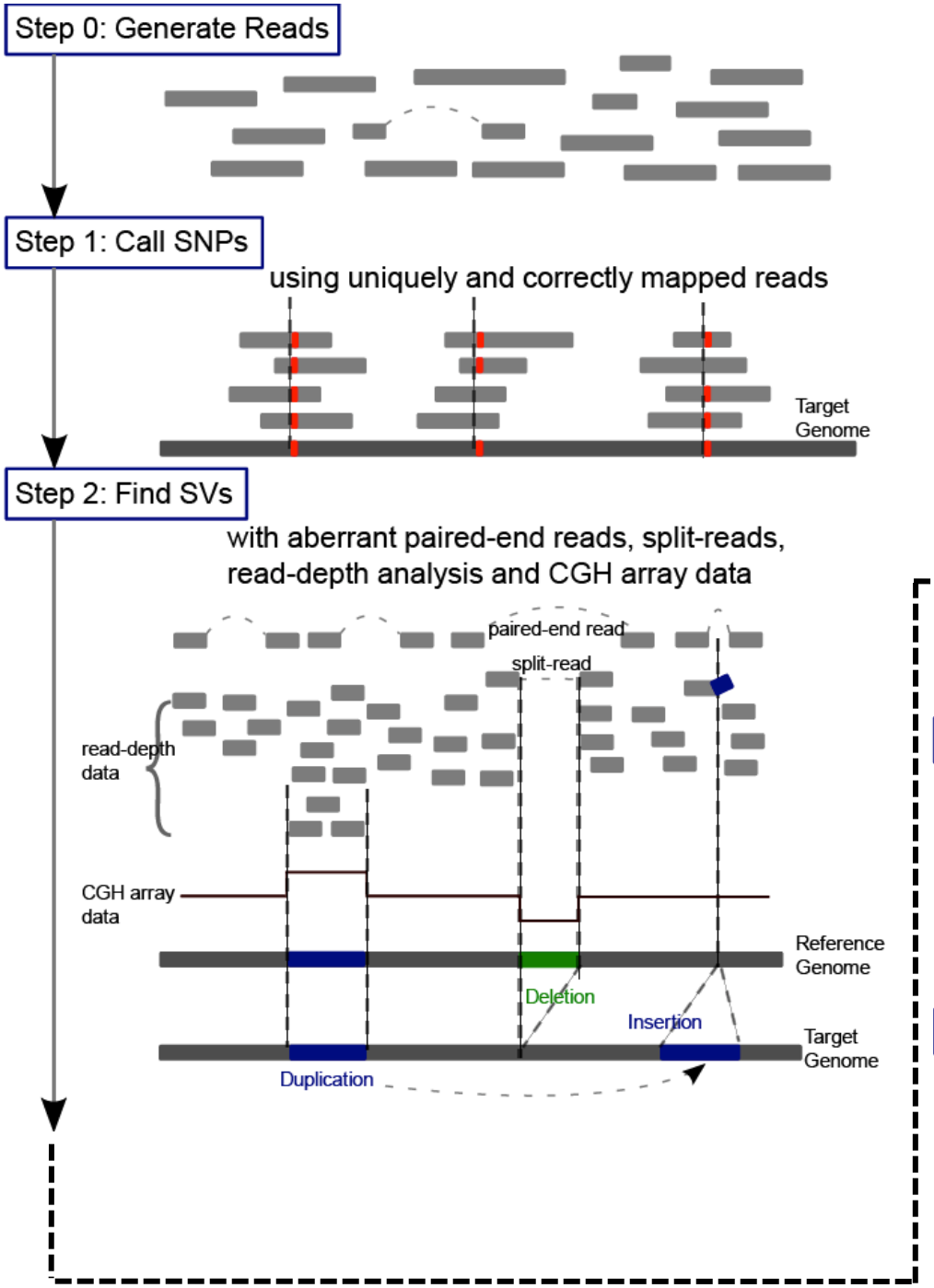
Watson

Stoicescu

Segmental duplications (SDs) - Contain Duplicated Paralogs and Duplicated Pseudogenes

Main Steps in Genome Resequencing

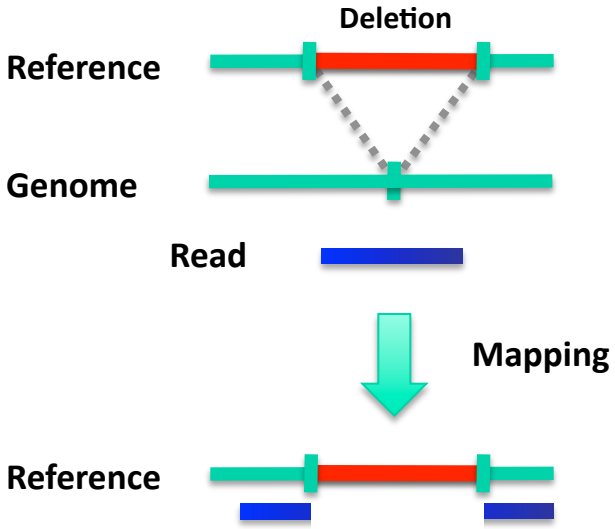
[Snyder et al. Genes & Dev. ('09), submitted]



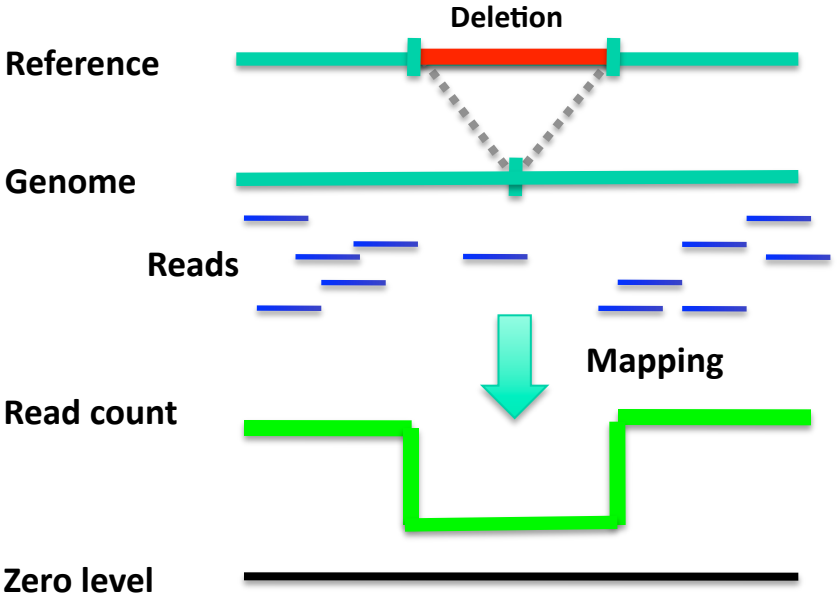
Methods to **Find SVs**

1. Paired ends

2. Split read



3. Read depth (or aCGH)

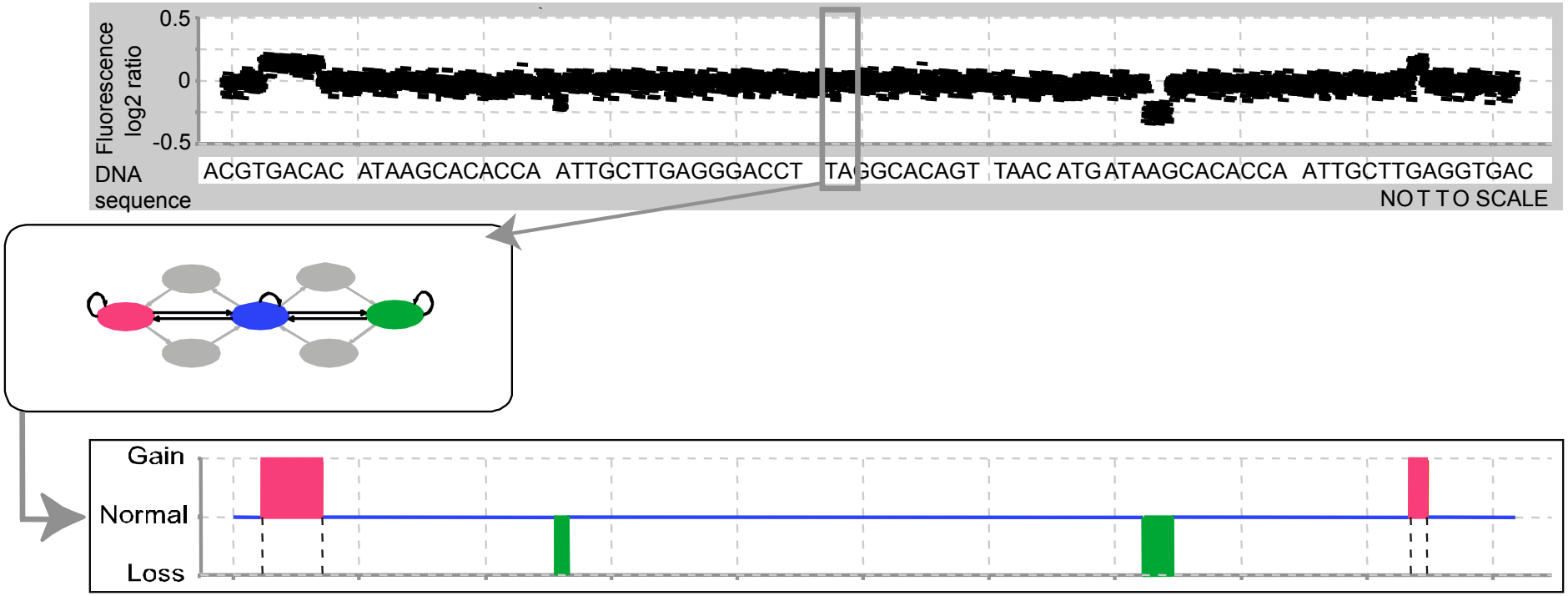


4. Local Reassembly

Segmentation of Array Signal (a precursor to Read Depth)

BreakPtr HMM

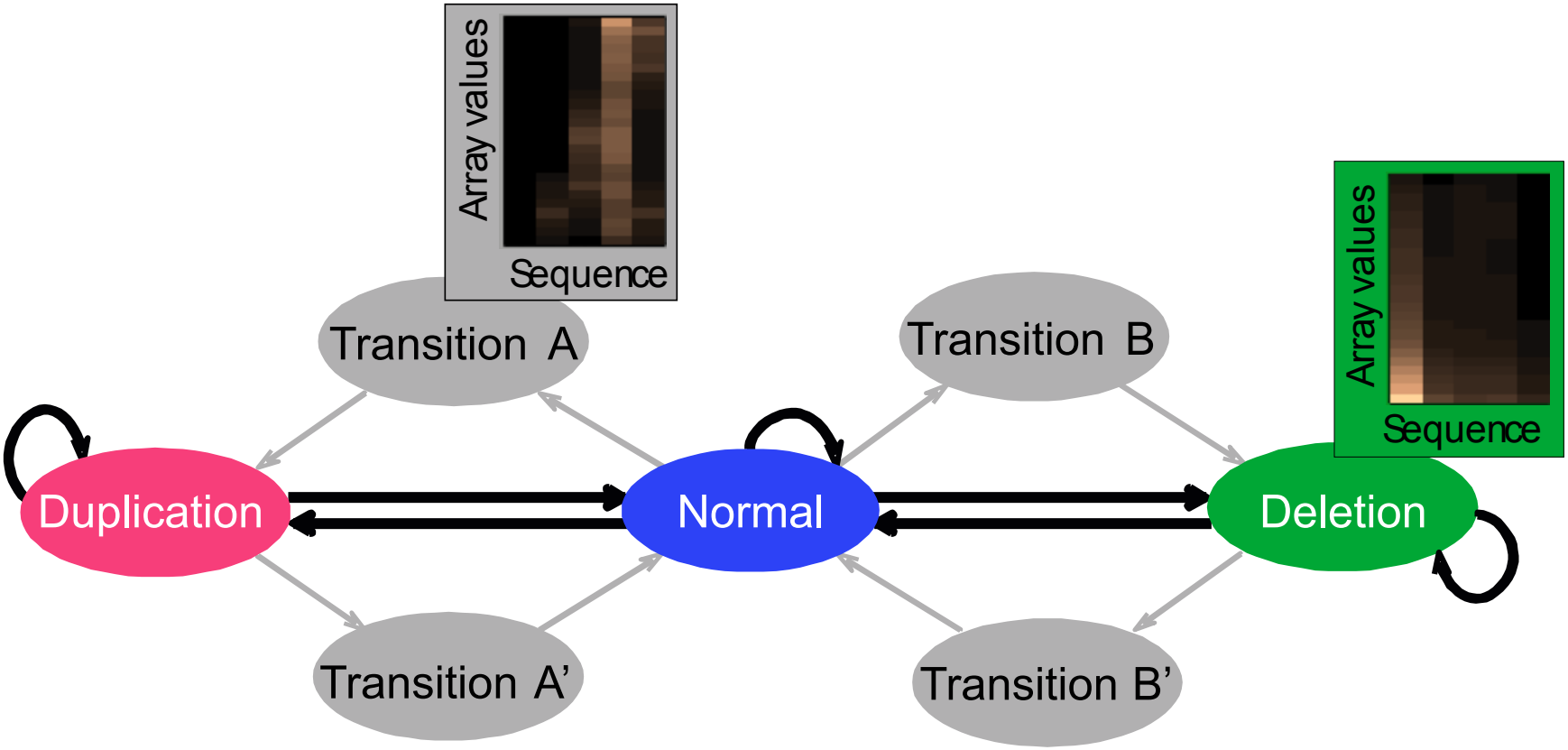
- To get highest resolution on breakpoints need to smooth & segment the signal
- BreakPtr: prediction of breakpoints, dosage and cross-hybridization using a system based on Hidden Markov Models



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

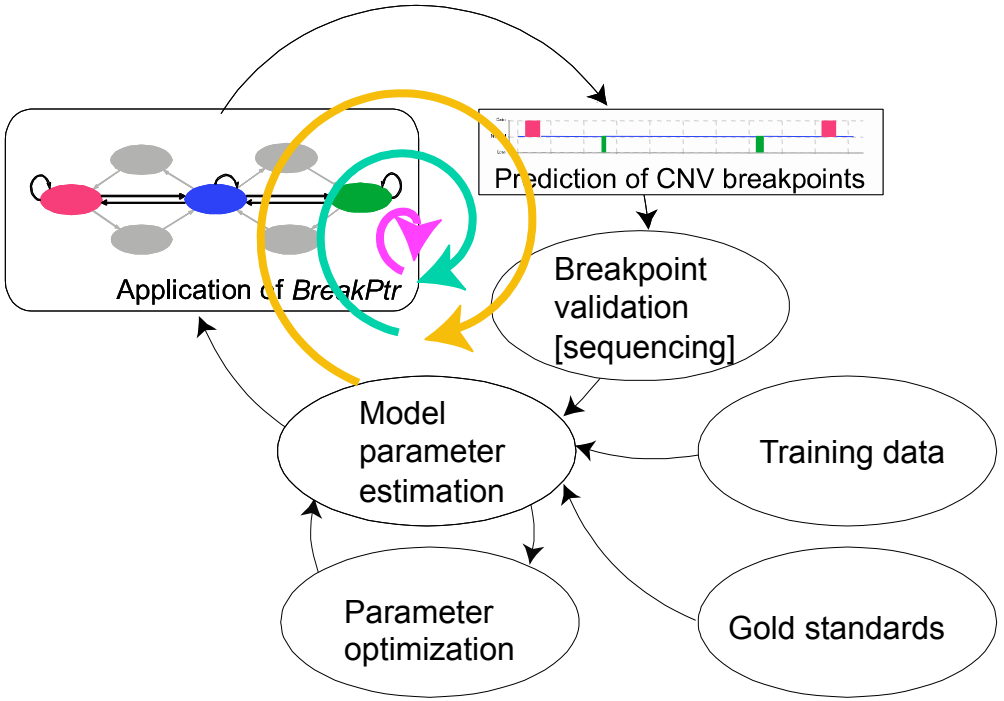
http://breakptr.gersteinlab.org

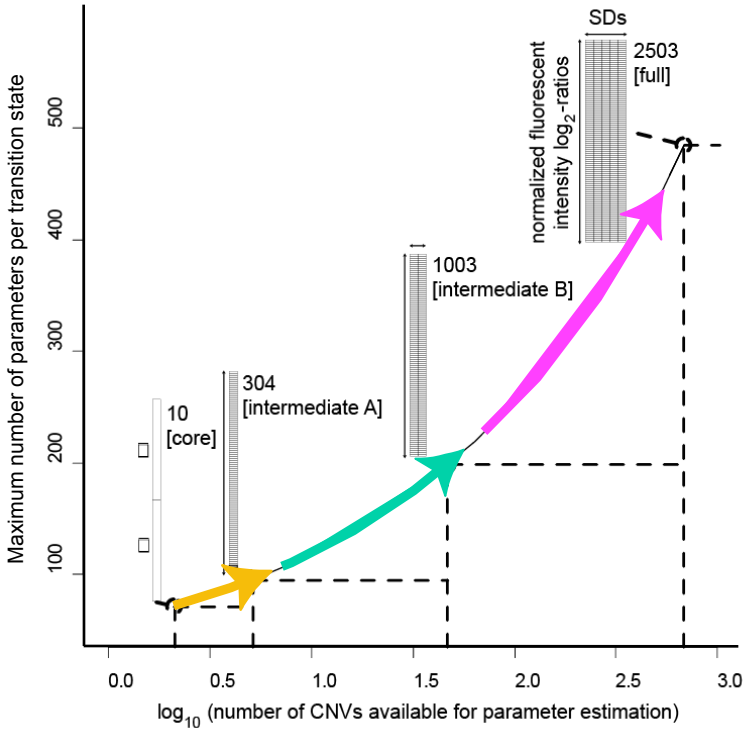
BreakPtr statistically integrates array signal and DNA sequence signatures (using a discrete-valued bivariate HMM)



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;

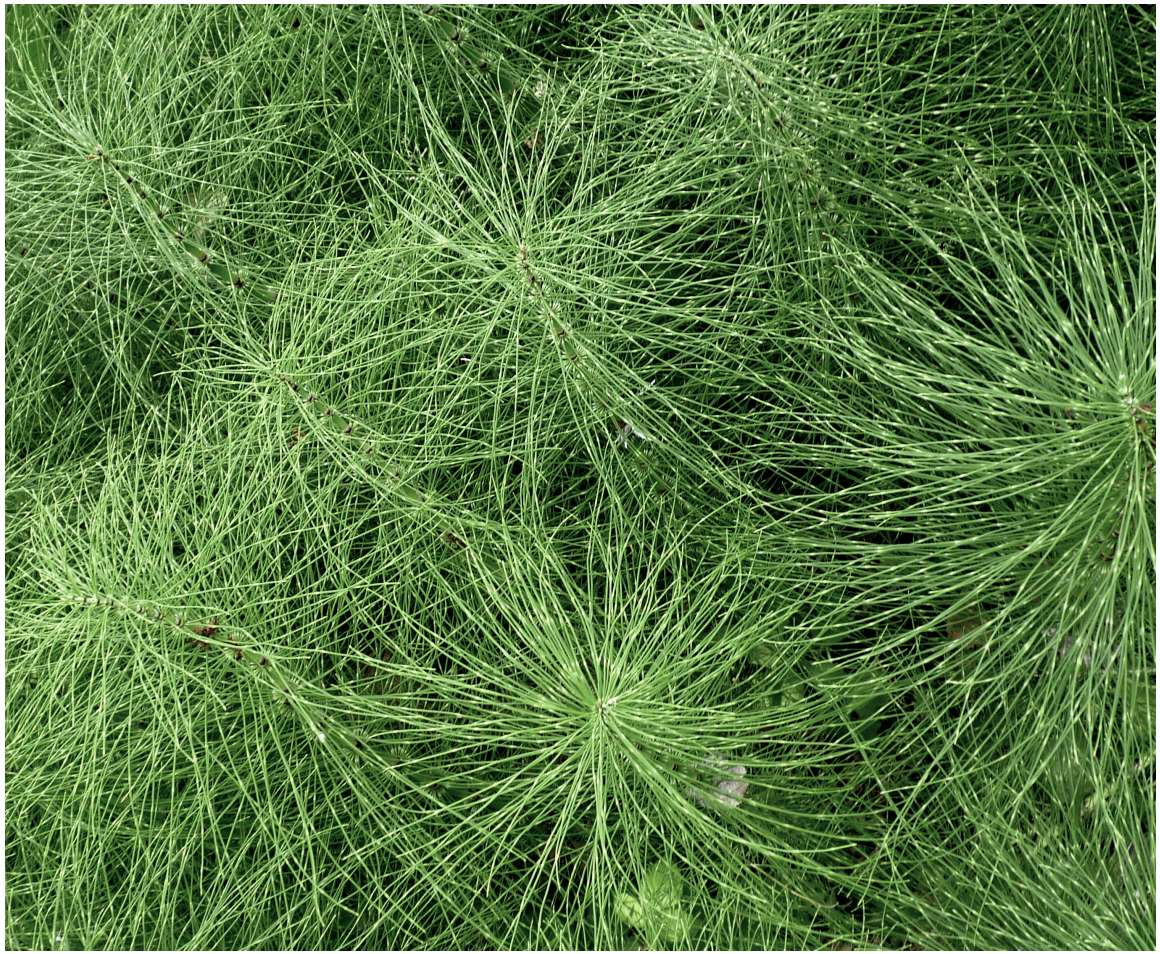
Median resolution <300 bp

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

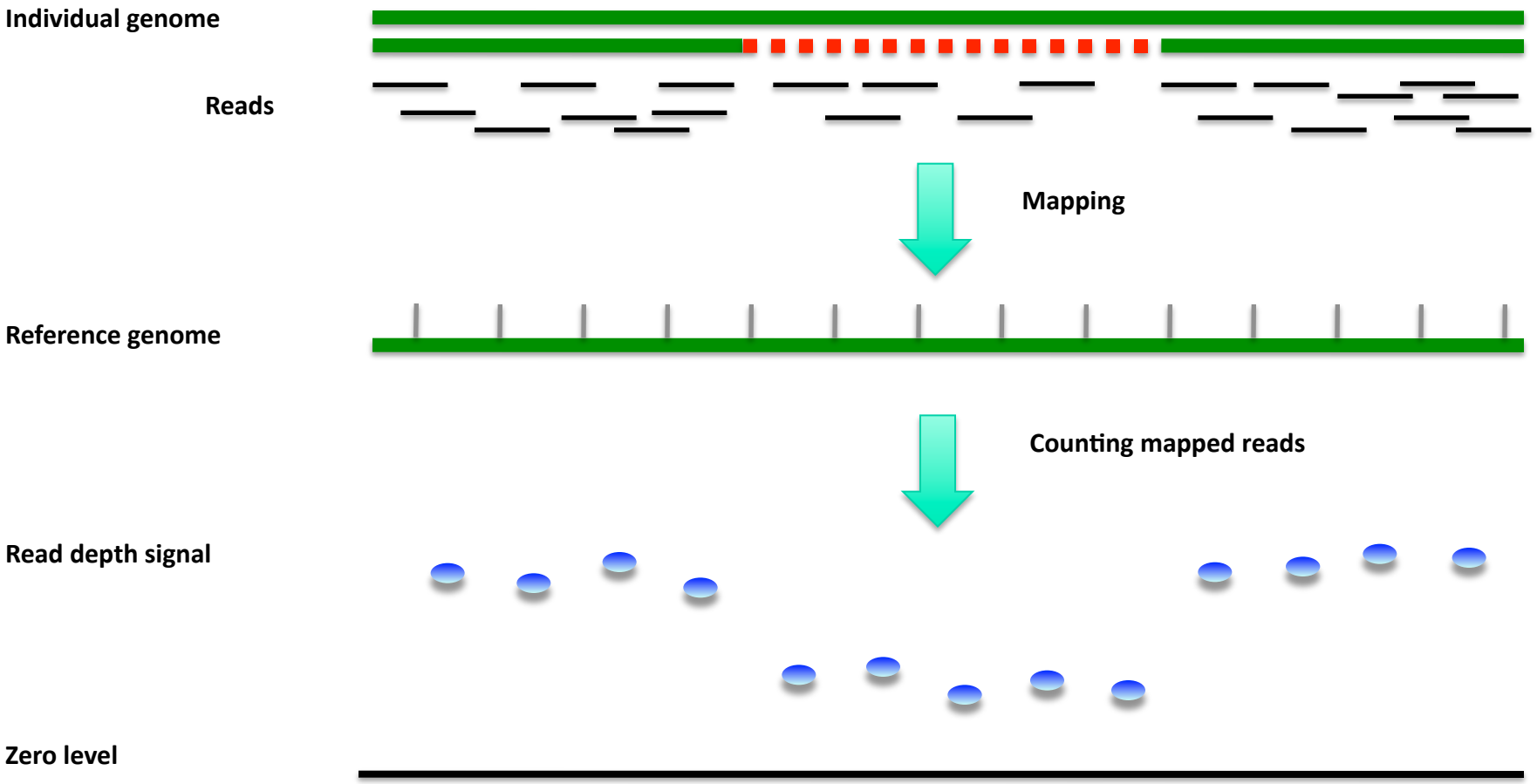
HMM optimized iteratively (using Expectation Maximization, EM)

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

Read-Depth from sequencing



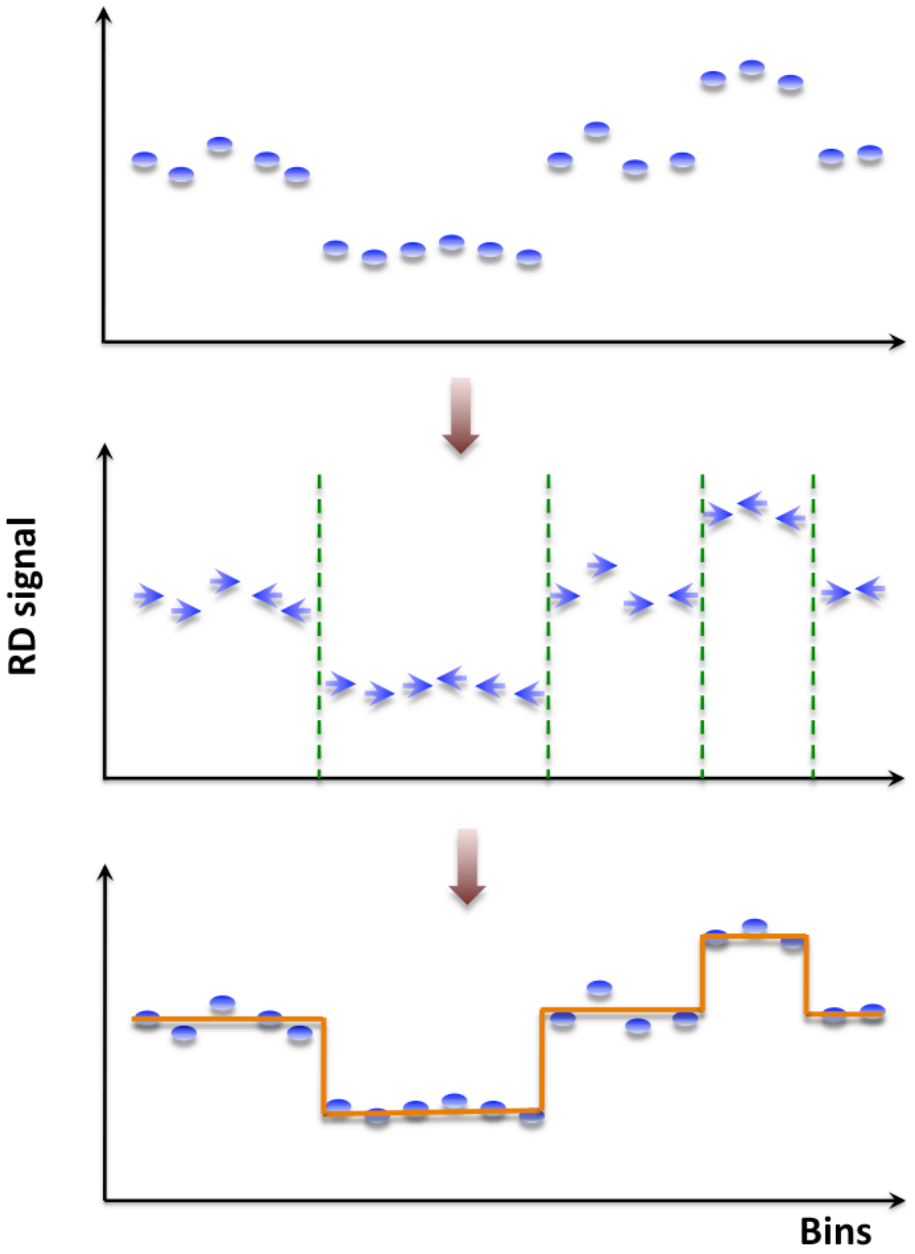
Read depth



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

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

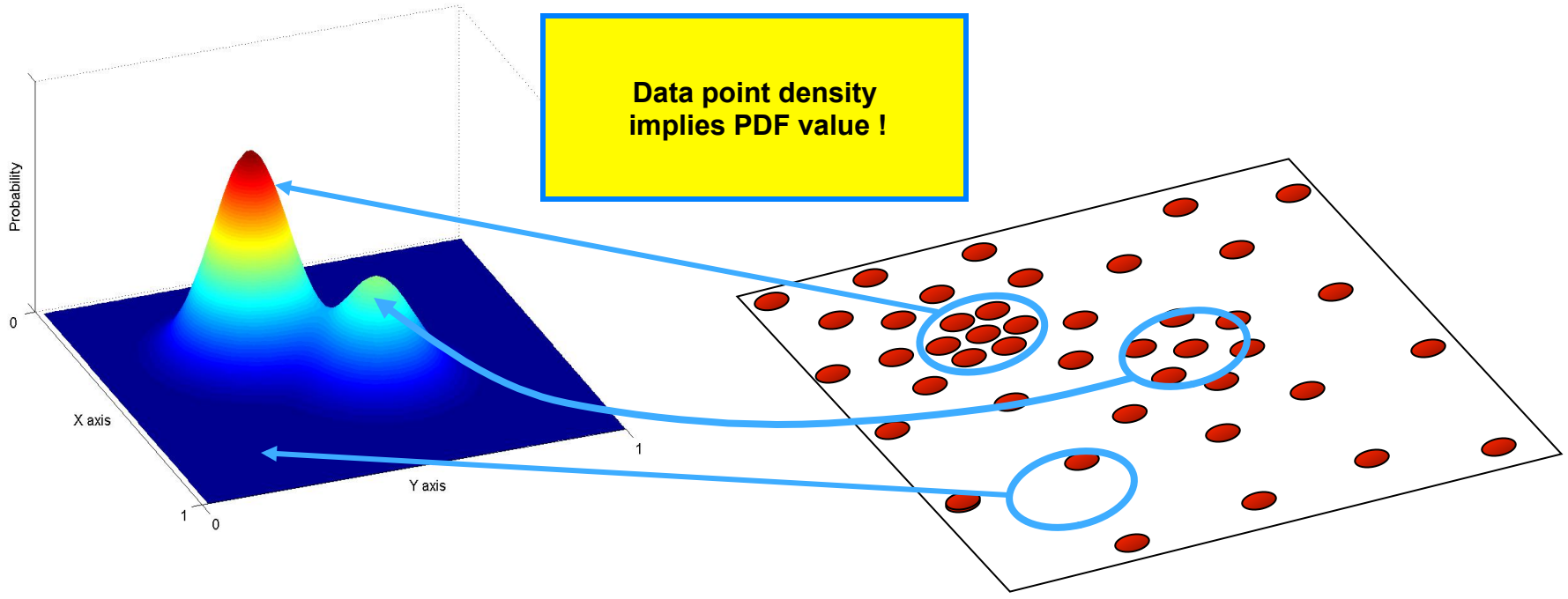
- For each bin attraction (mean-shift) vector points in the direction of bins with most similar RD signal
- No prior assumptions about number, sizes, haplotype, frequency and density of CNV regions
- Not Model-based (e.g. like HMM) with global optimization, distr. assumption & parms. (e.g. num. of segments).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

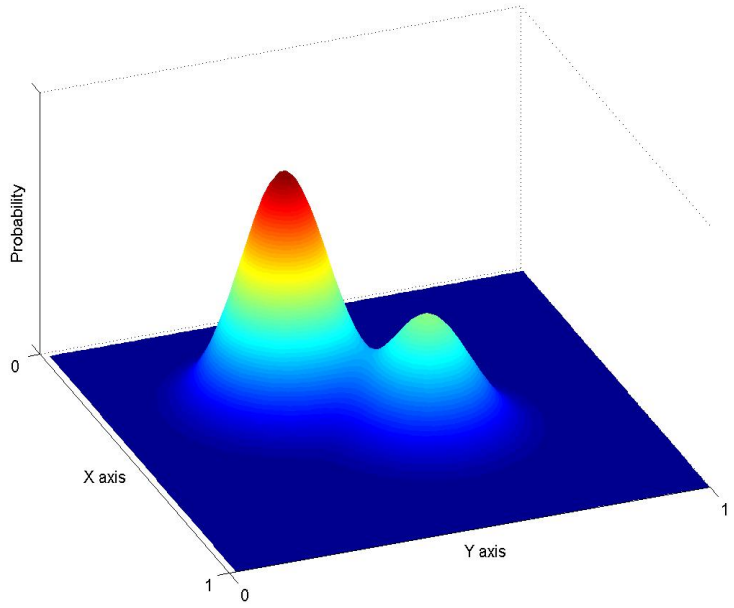
Assumption : The data points are sampled from an underlying PDF

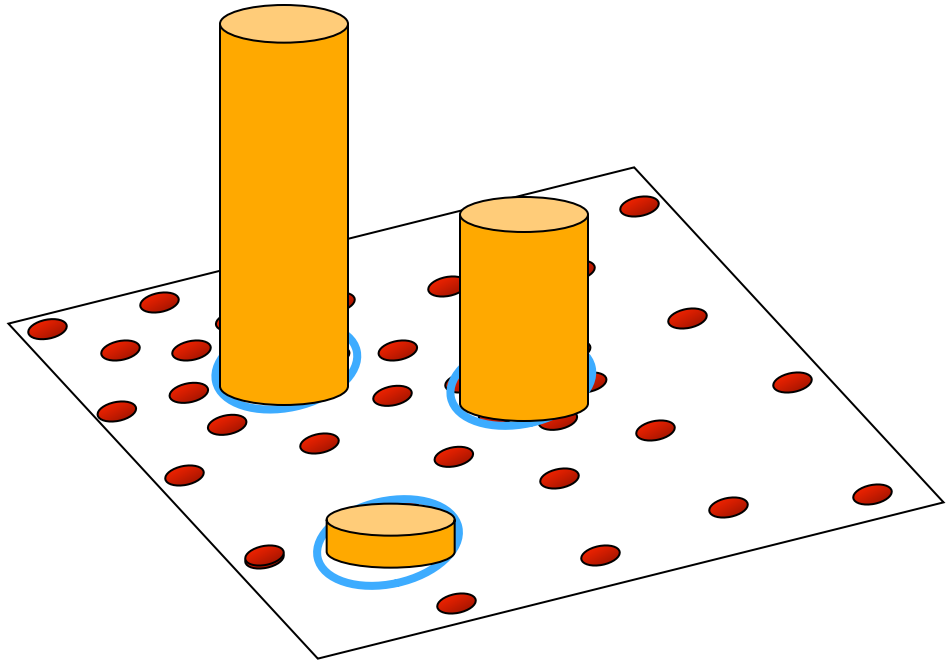


Assumed Underlying PDF **Real Data Samples**

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

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



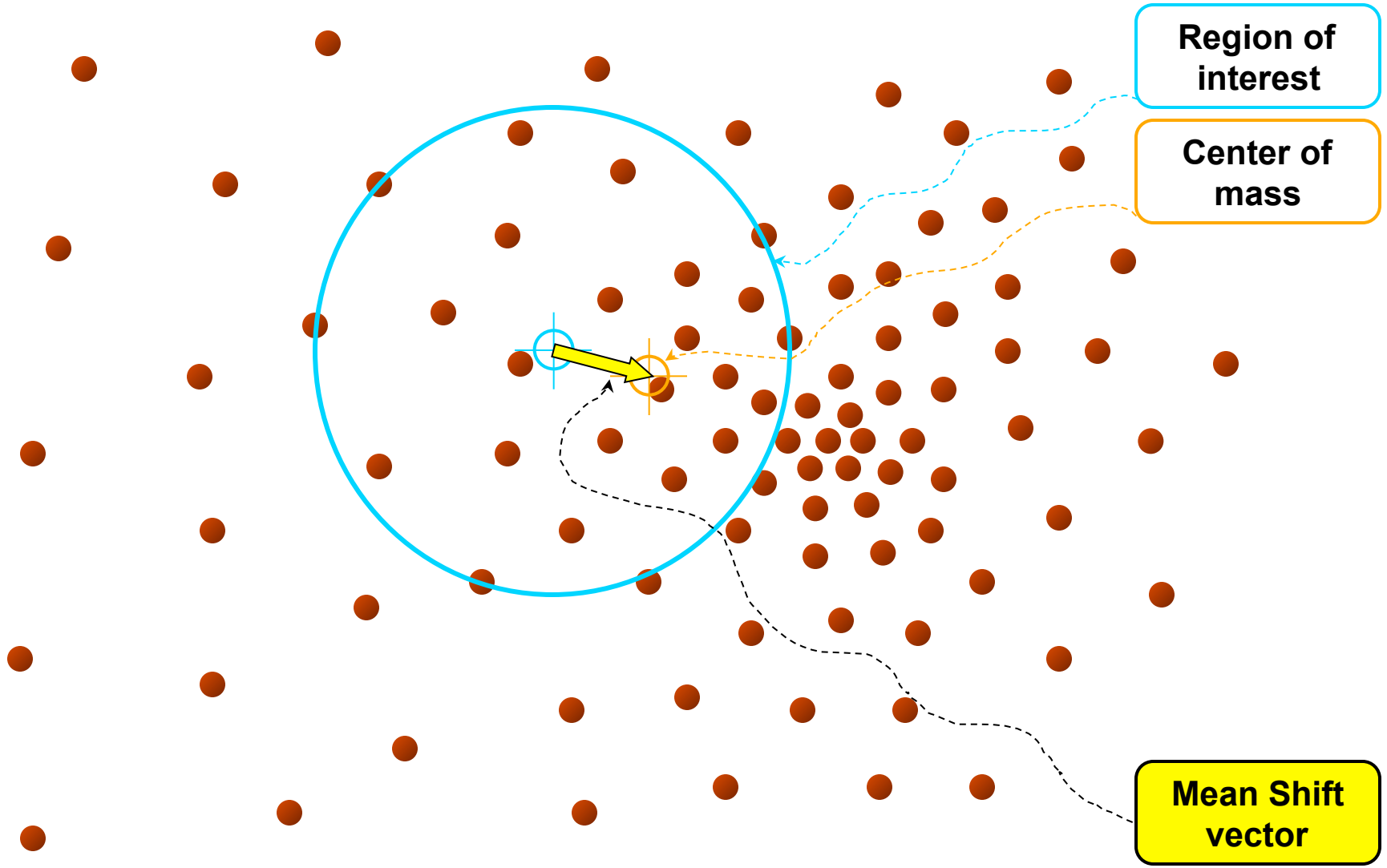


Assumed Underlying PDF

Real Data Samples

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

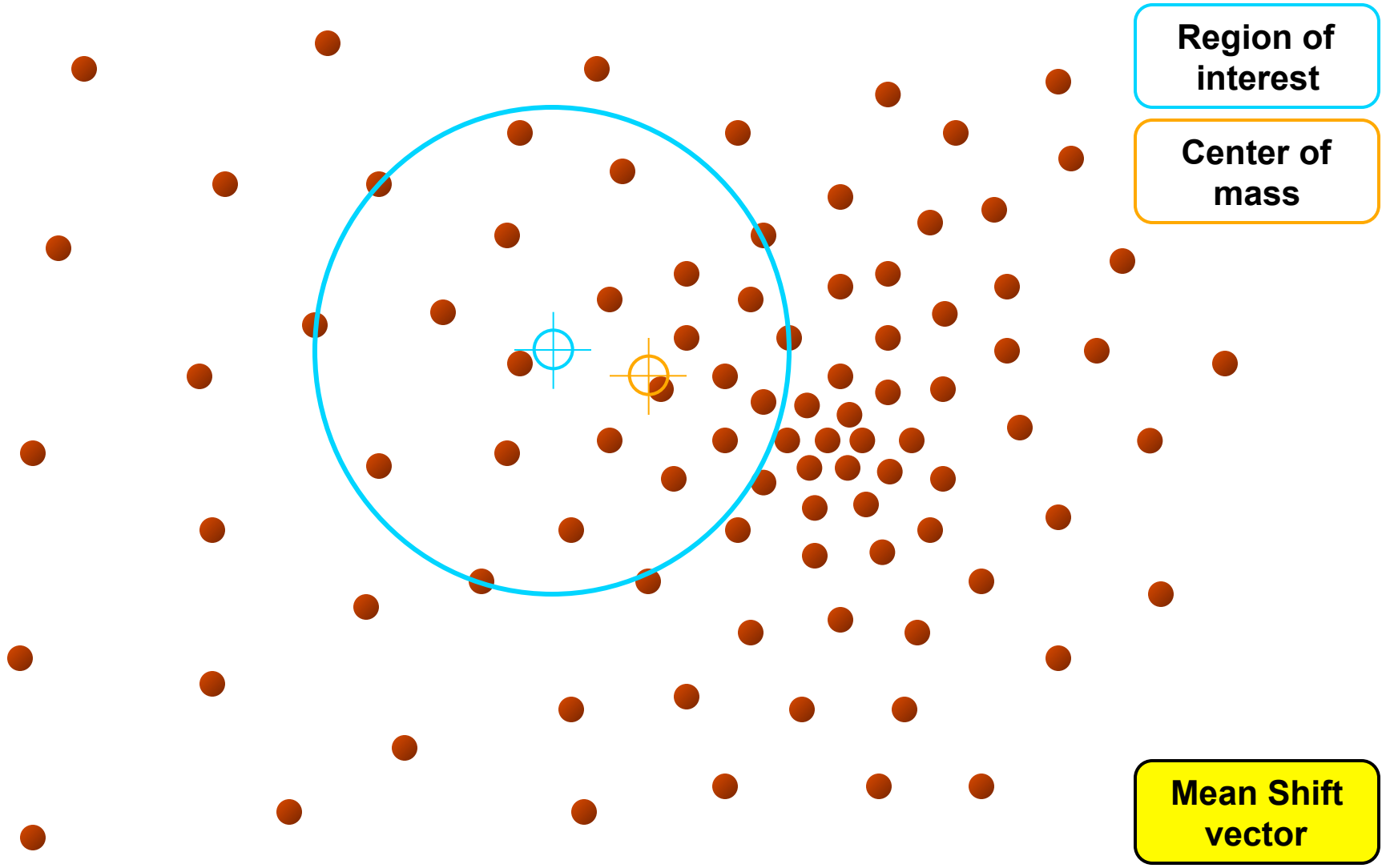
Intuitive Description



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

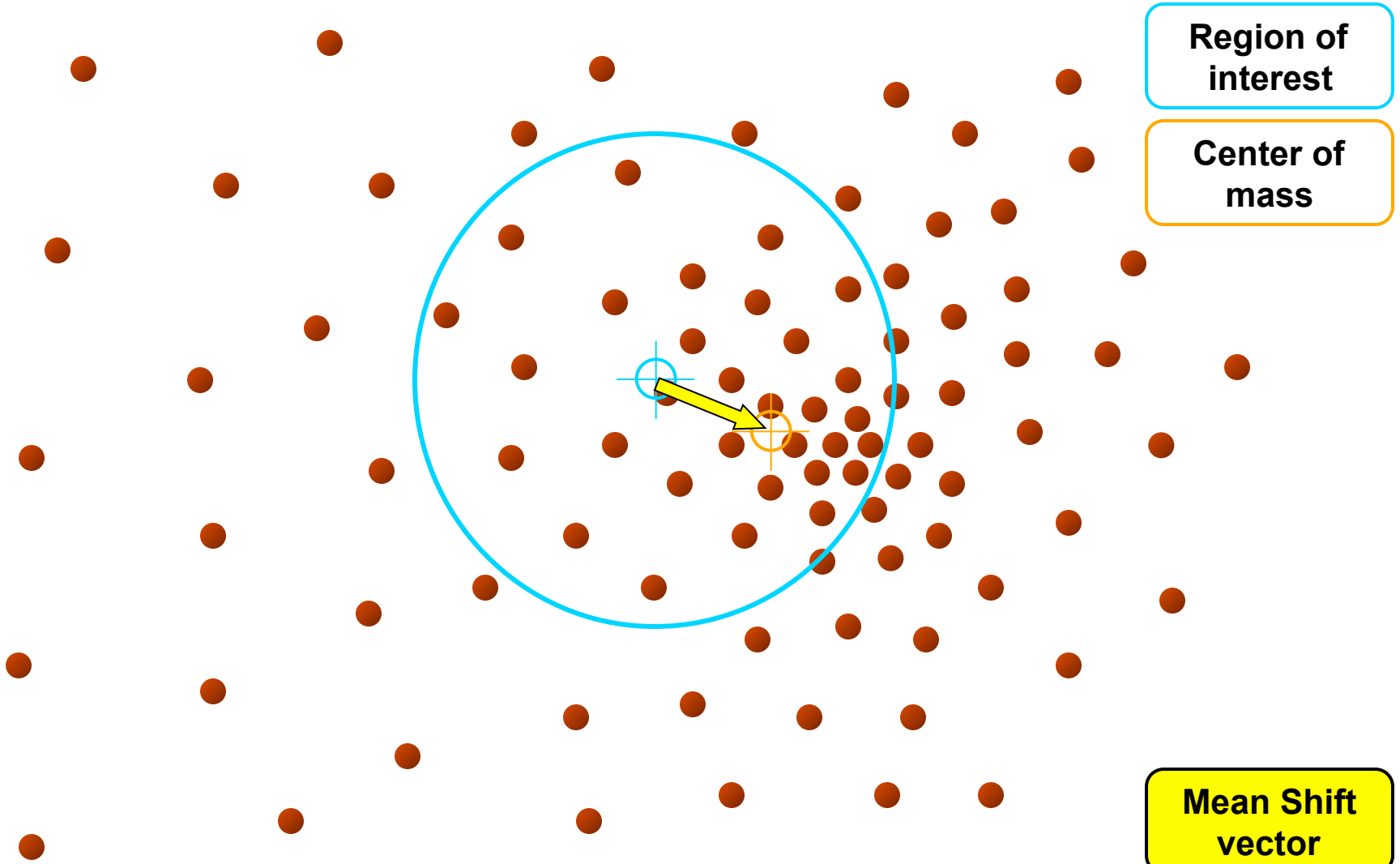


Objective : Find the densest region

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

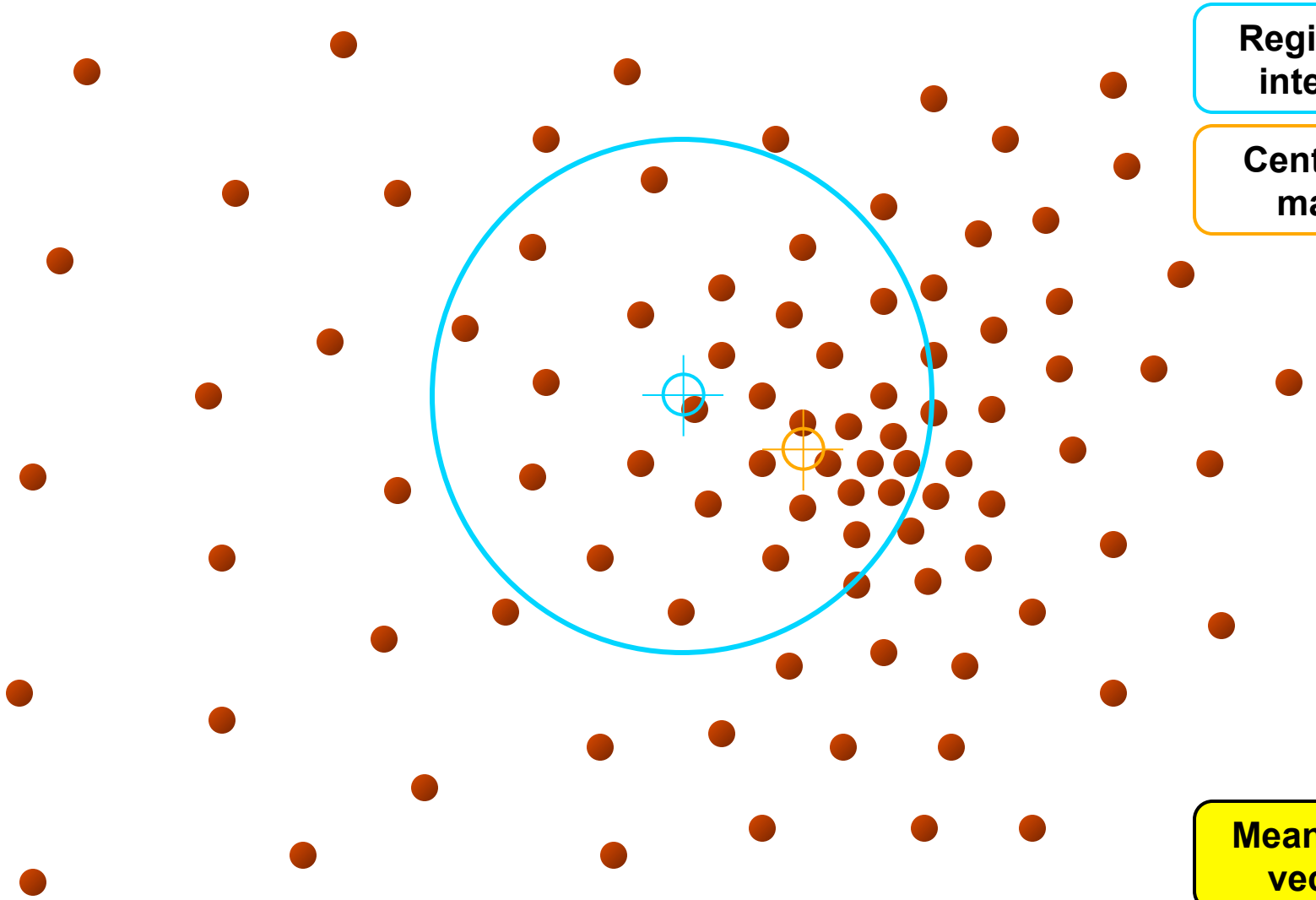


Objective : Find the densest region

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

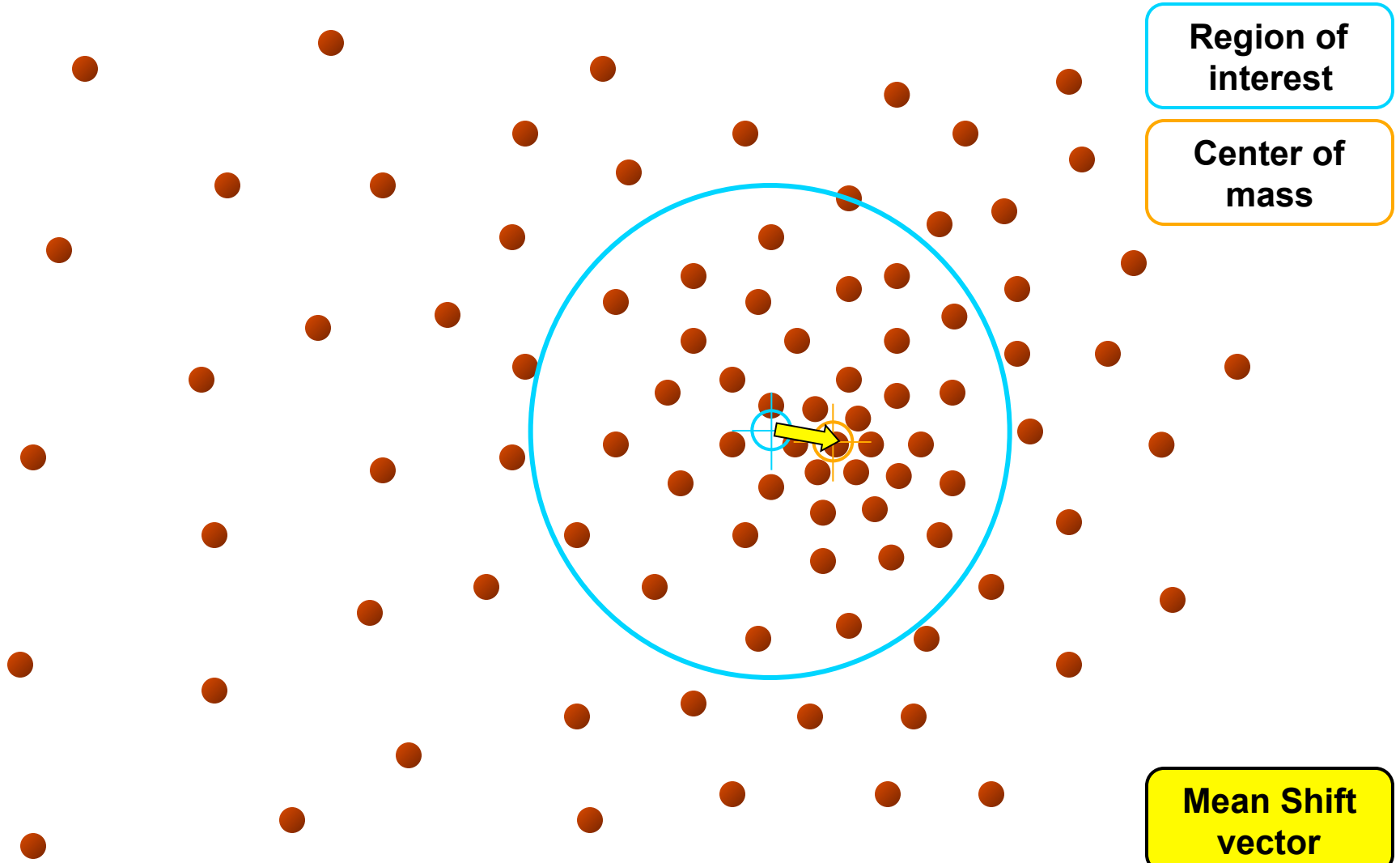


Objective : Find the densest region

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

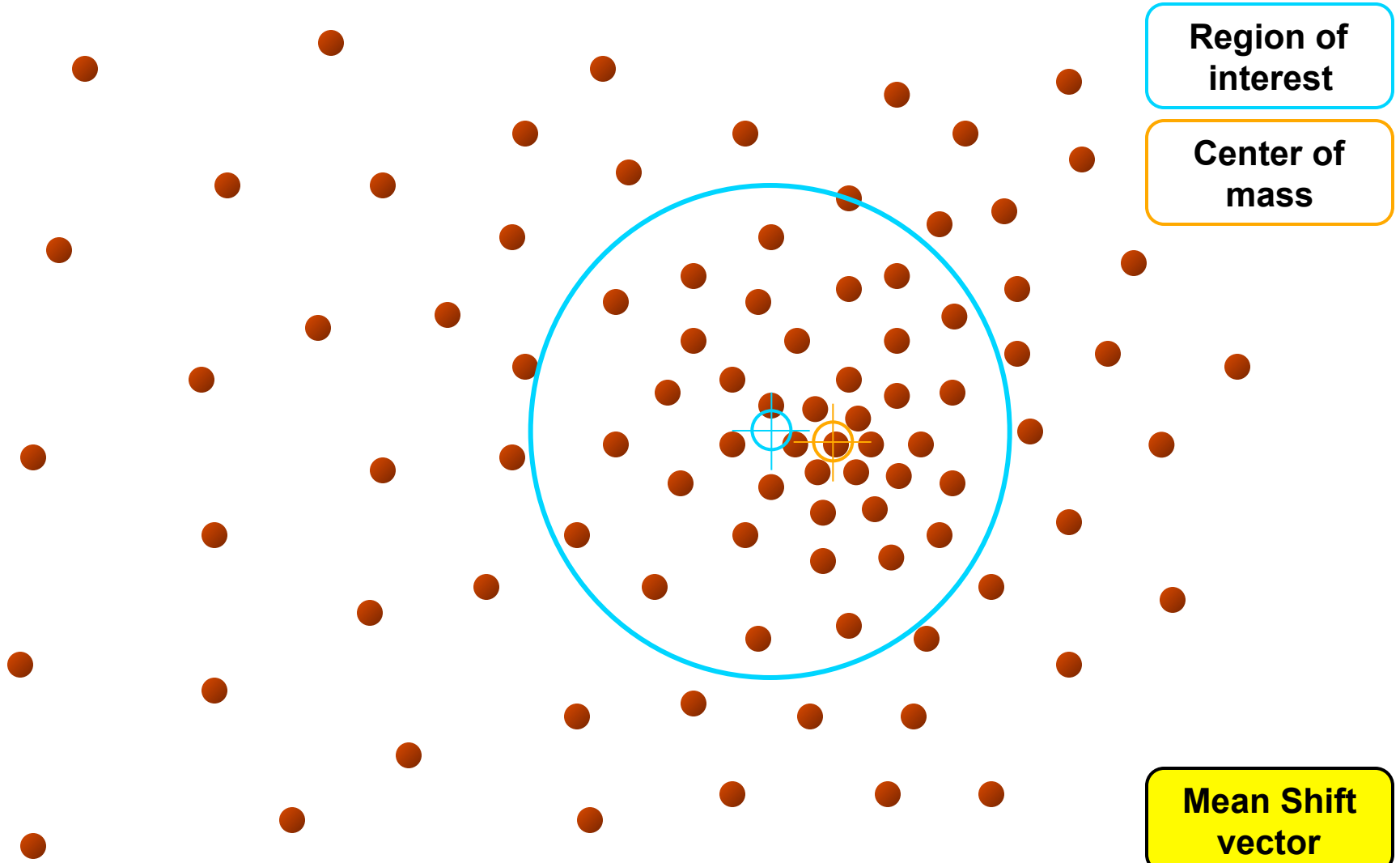


Objective : Find the densest region

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

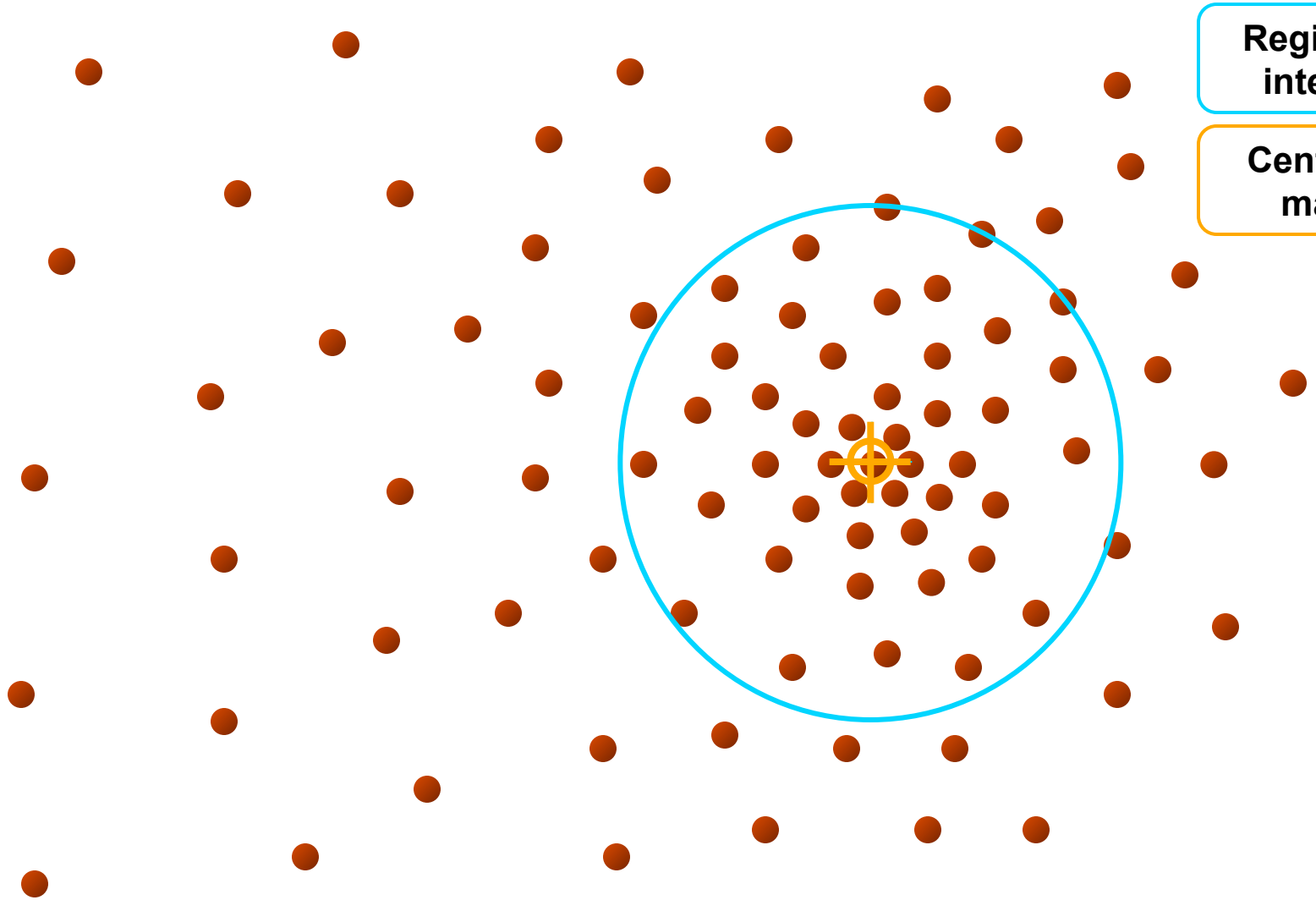


Objective : Find the densest region

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

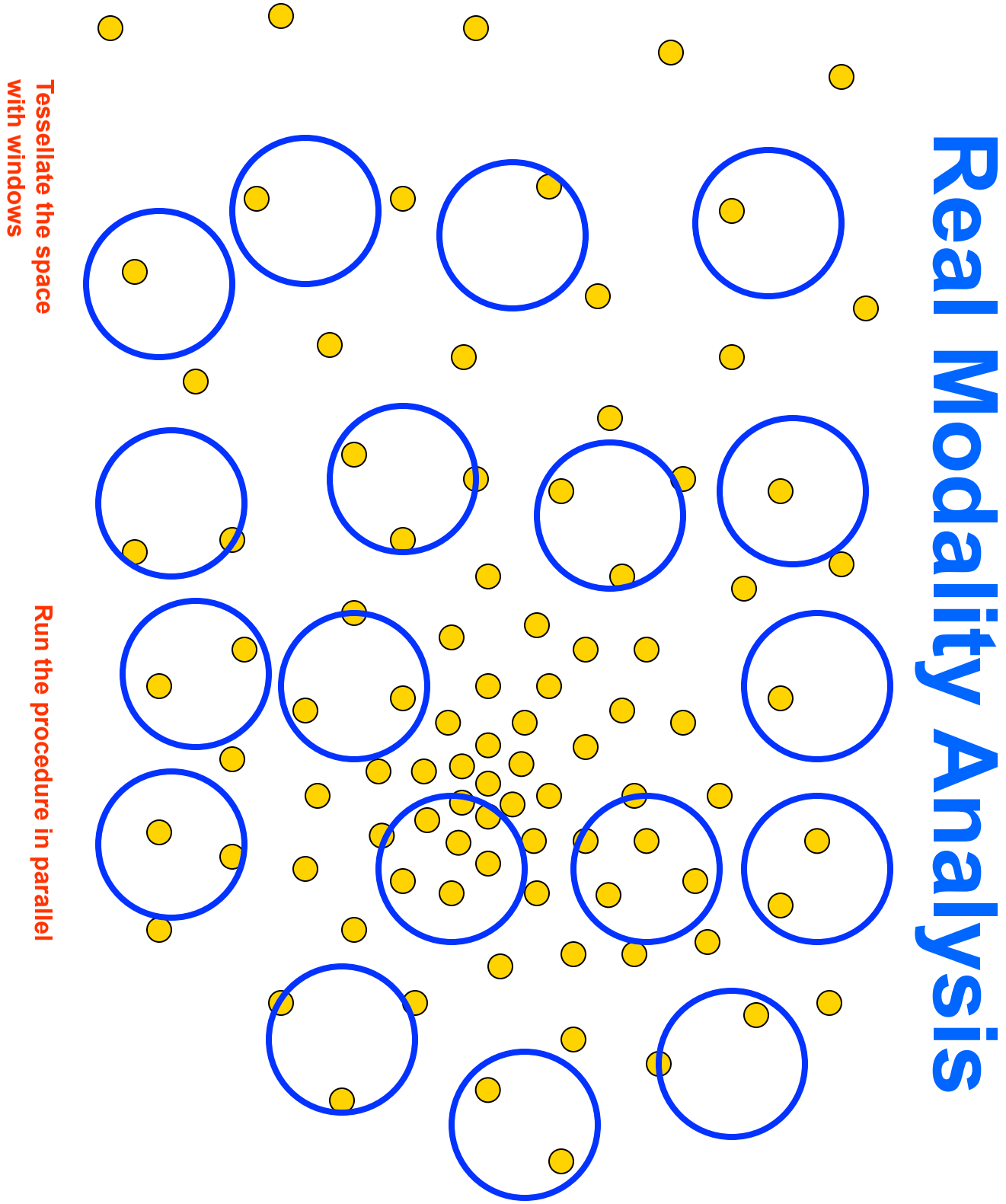


Objective : Find the densest region
Distribution of identical billiard balls

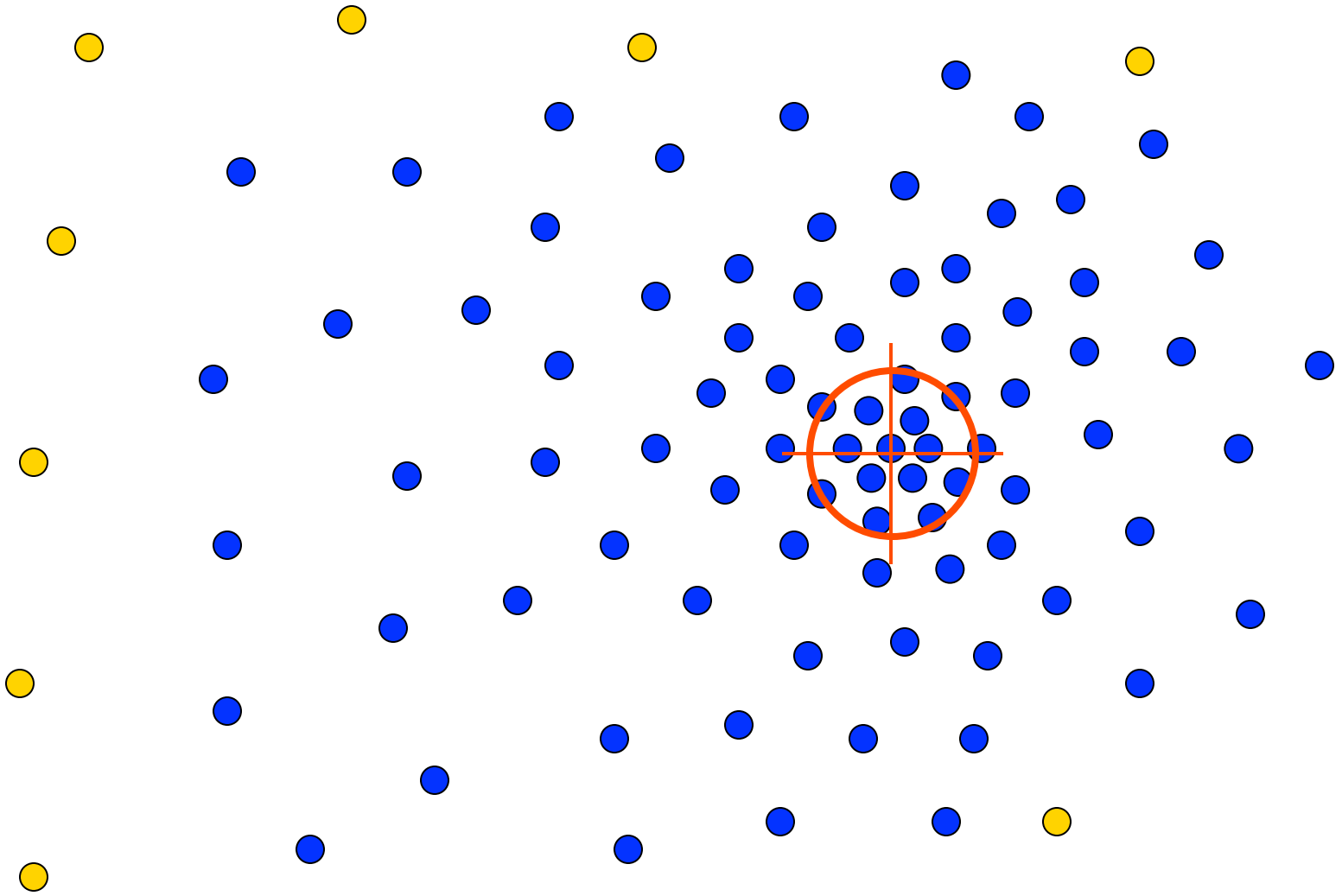
[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

[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



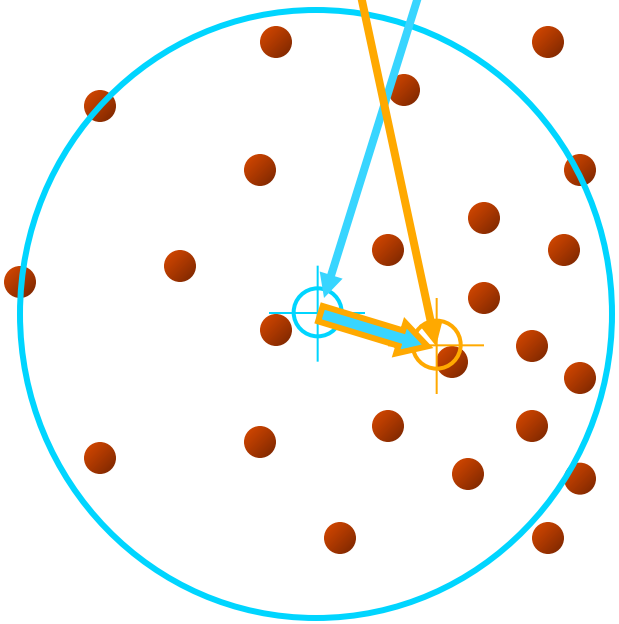
The blue data points were traversed by the windows towards the mode

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

Computing The Mean Shift

$$\nabla P(\mathbf{x}) = \frac{c}{n}\sum_{i=1}^{n}\nabla k_i = \frac{c}{n}\left[\sum_{i=1}^{n} g_i\right]\left[\frac{\sum_{i=1}^{n}\mathbf{x}_i g_i}{\sum_{i=1}^{n} g_i} - \mathbf{x}\right]$$

Yet another Kernel density estimation !



Simple Mean Shift procedure:

- Compute mean shift vector

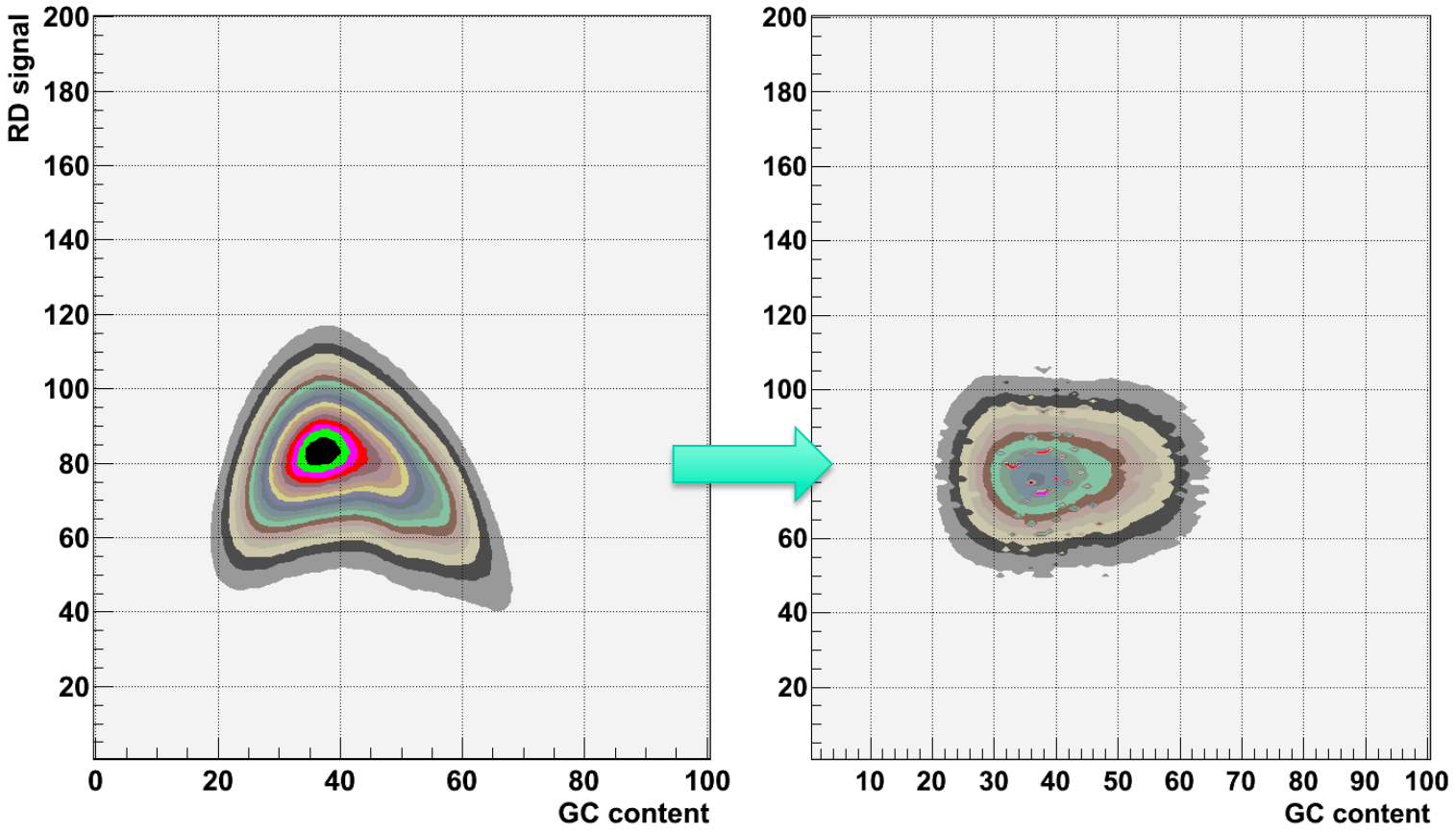
$$\mathbf{m}(\mathbf{x}) = \left[\frac{\sum_{i=1}^{n}\mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x}\right]$$

- Translate th

$$g(\mathbf{x}) = -k'(\mathbf{x})$$

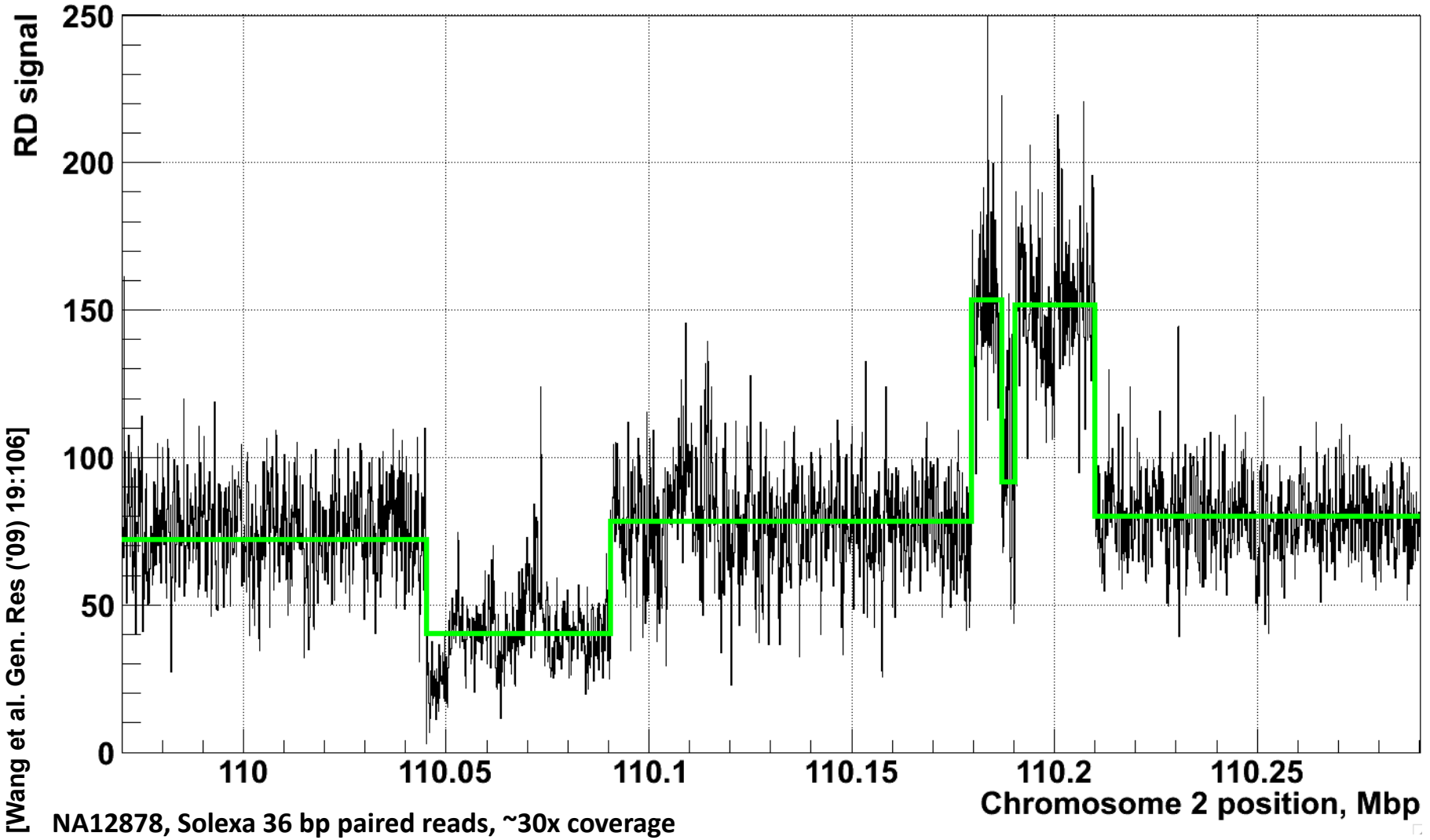
[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

GC bias correction

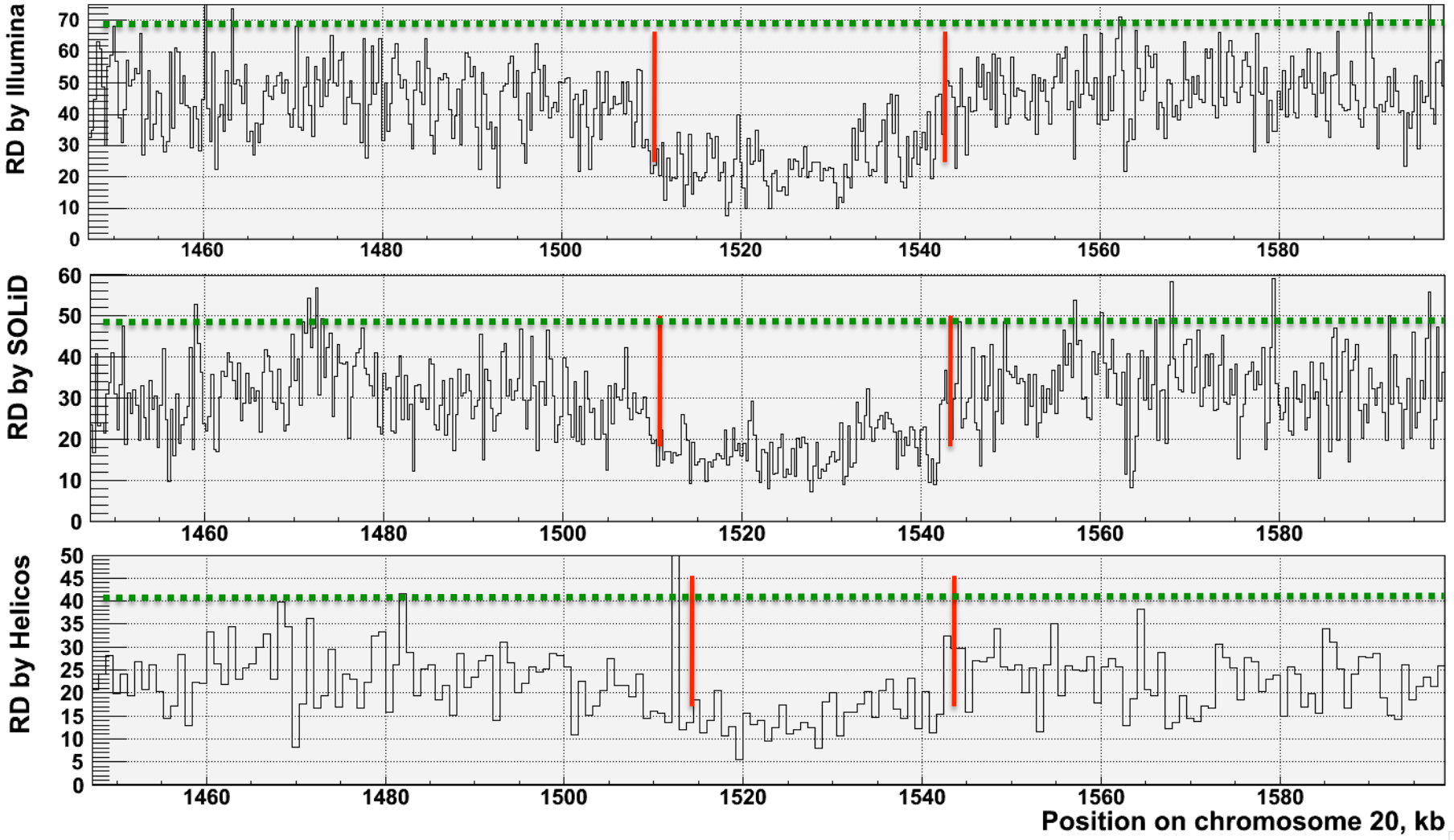


$$RD_{corrected} = \overline{RD}_{global}\ RD/\overline{RD}_{GC}$$

Example of Application of MSB to RD data

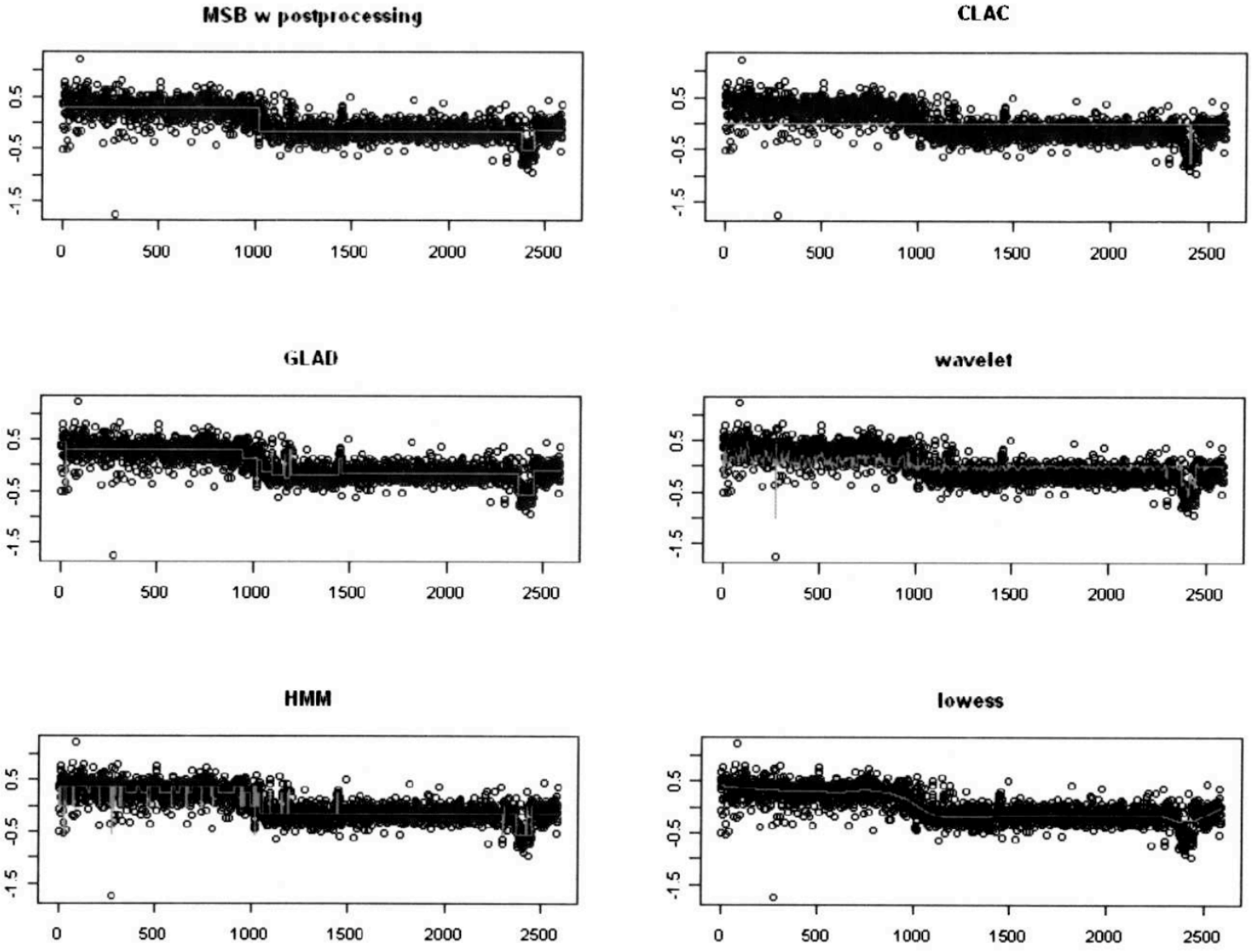


RD works well on a variety of sequencing platforms



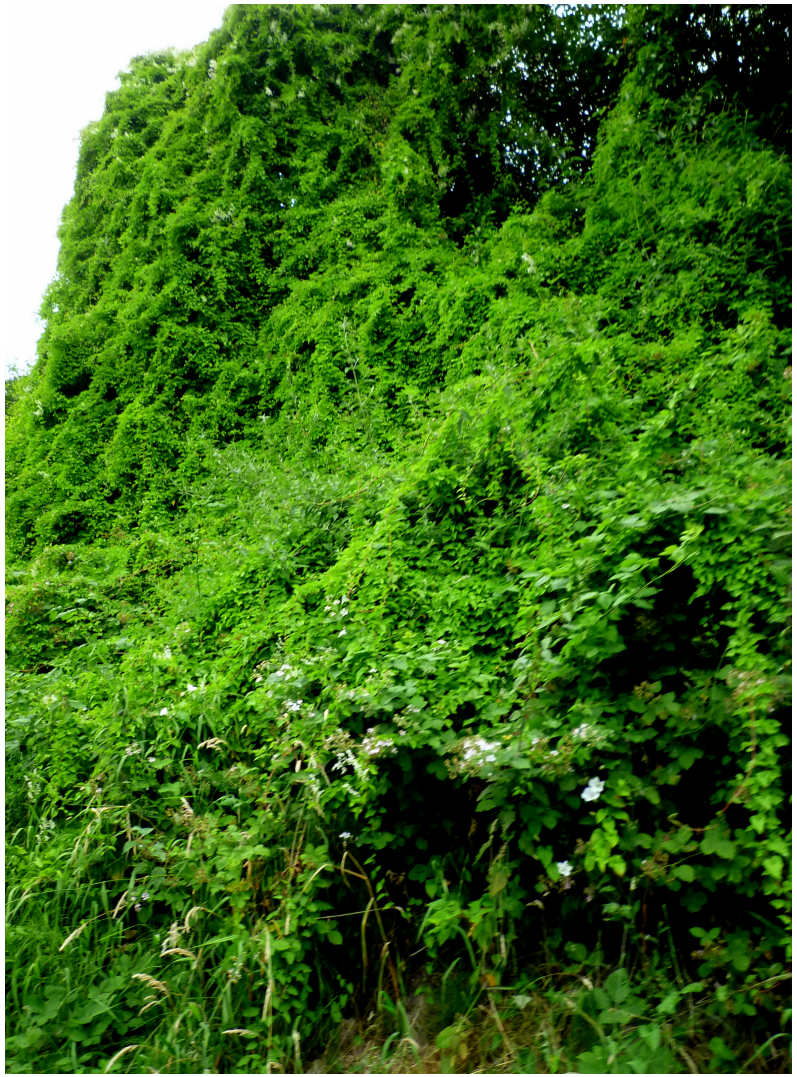
[NA18505]

MSB works well on array data too

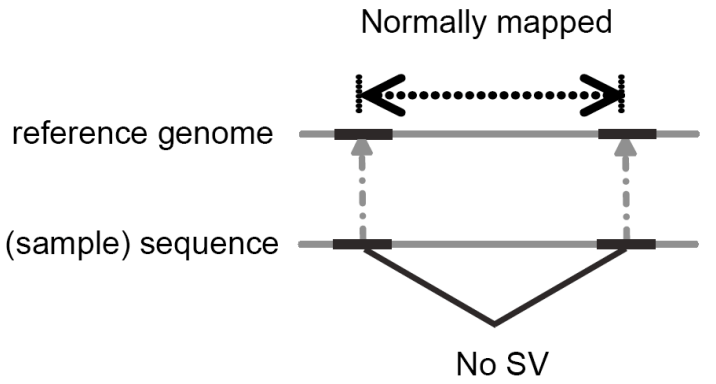


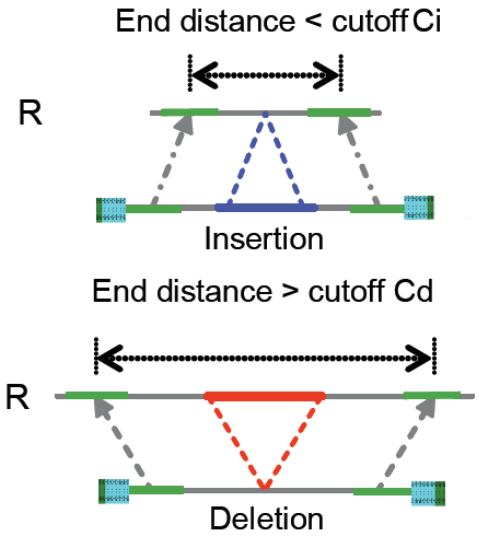
[Wang et al. Gen. Res ('09) 19:106]

Looking for Aberrantly Placed Paired Ends



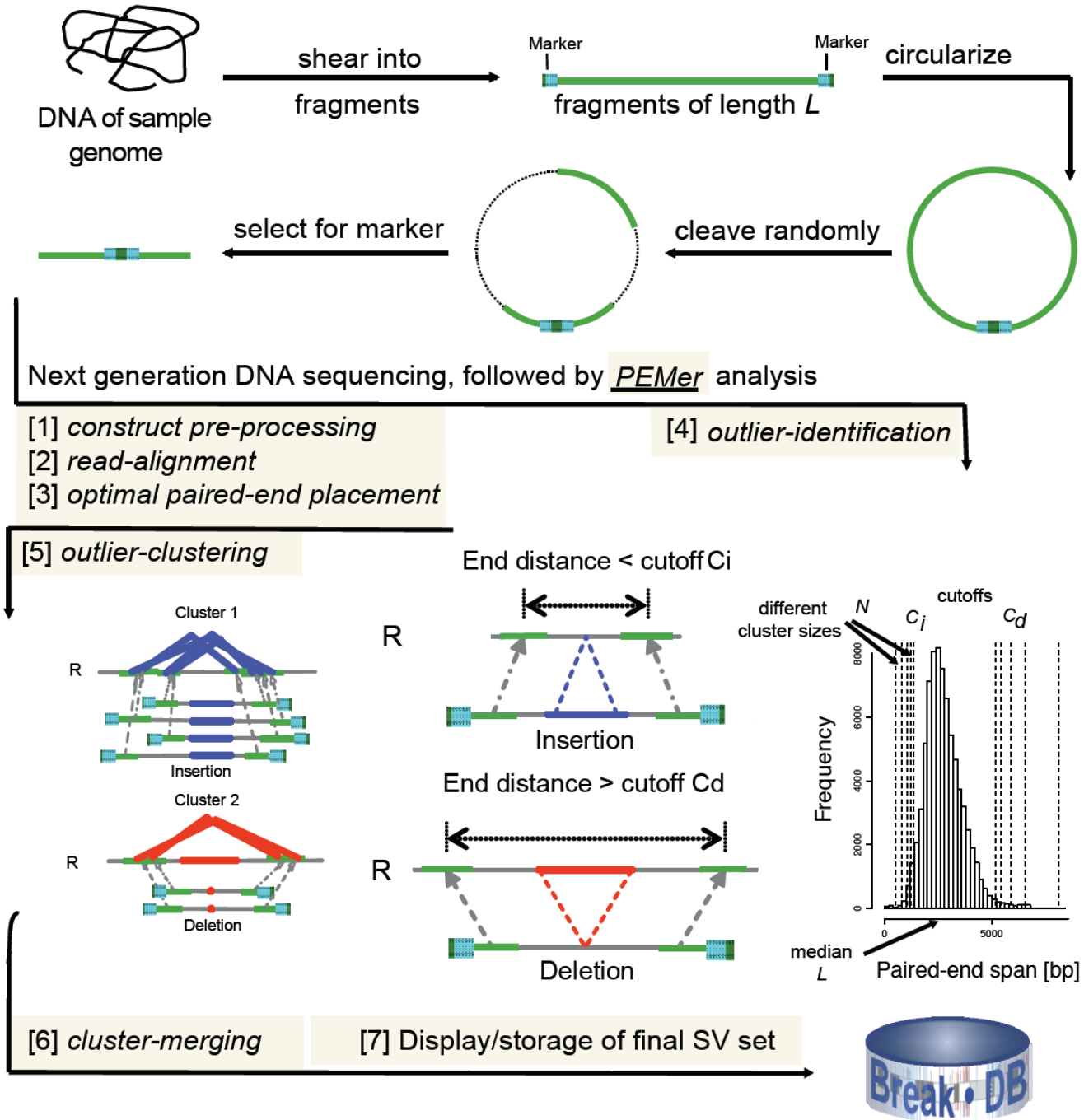
PEMer: Detecting Structural Variants from Discordant Paired Ends in Massive Sequencing





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

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

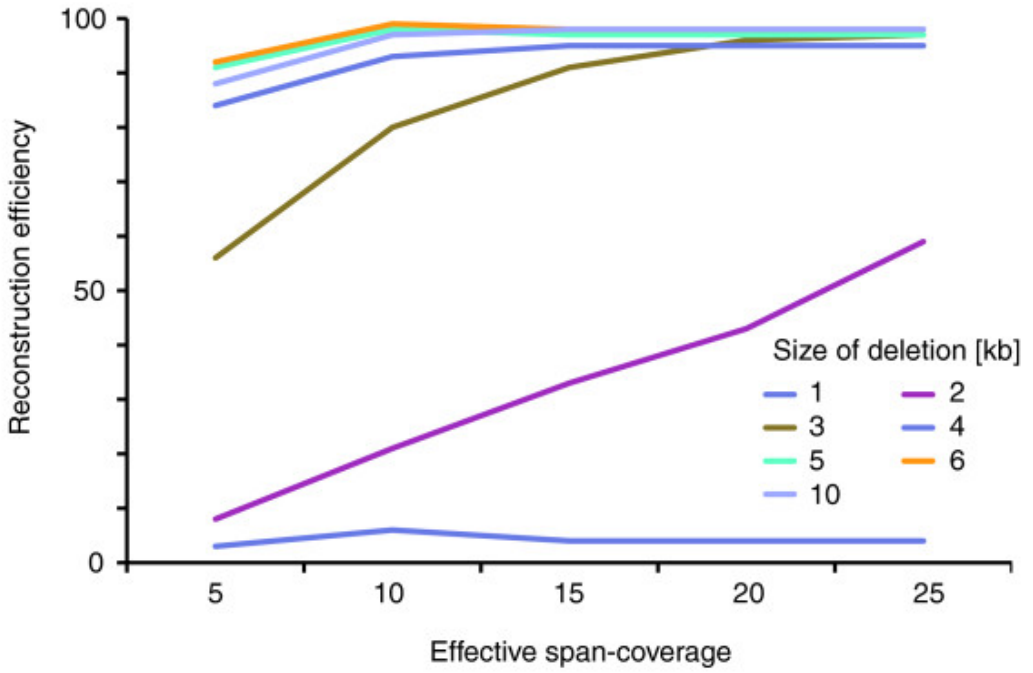


[Korbel et al., Science ('07); Korbel et al., GenomeBiol. ('09)]

Parameterize Error Models through Simulation

Reconstruction efficiency at different coverage

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



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

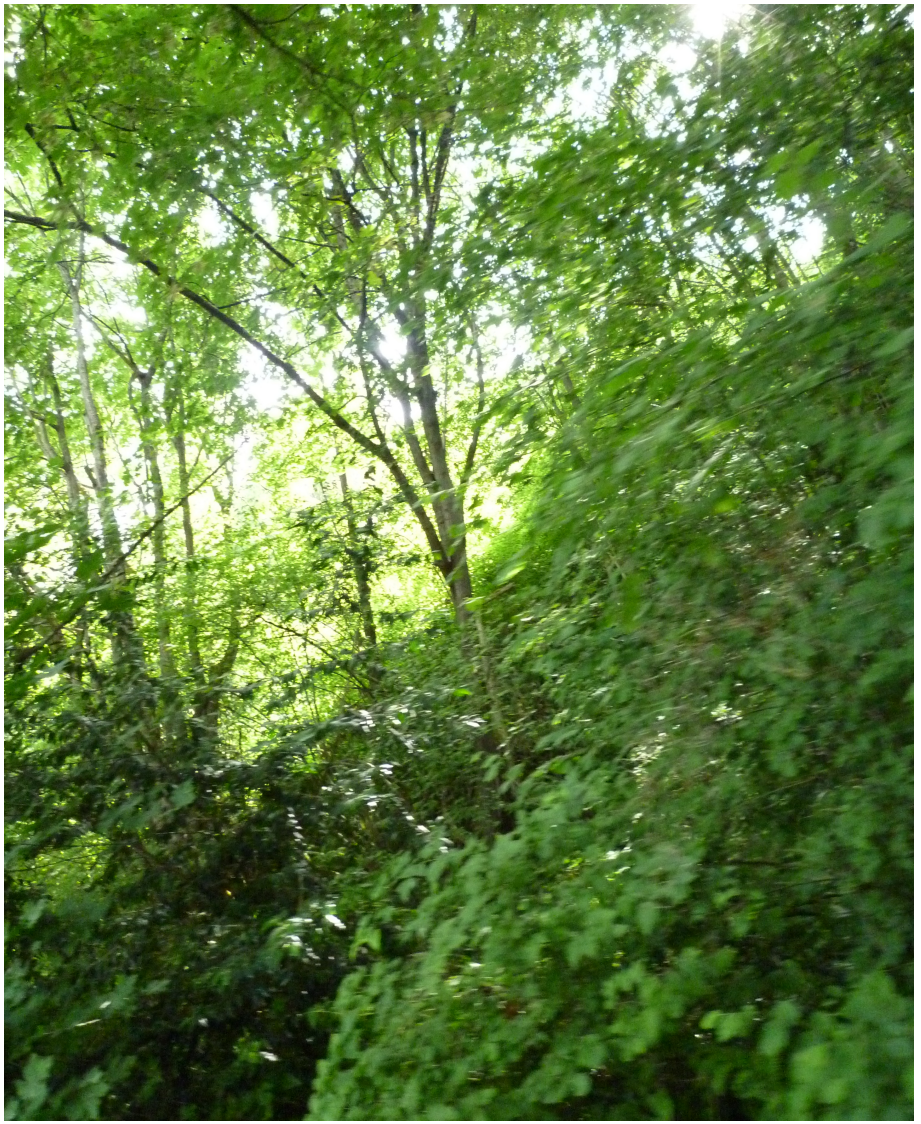
Reconstruction of heterozygous insertions

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

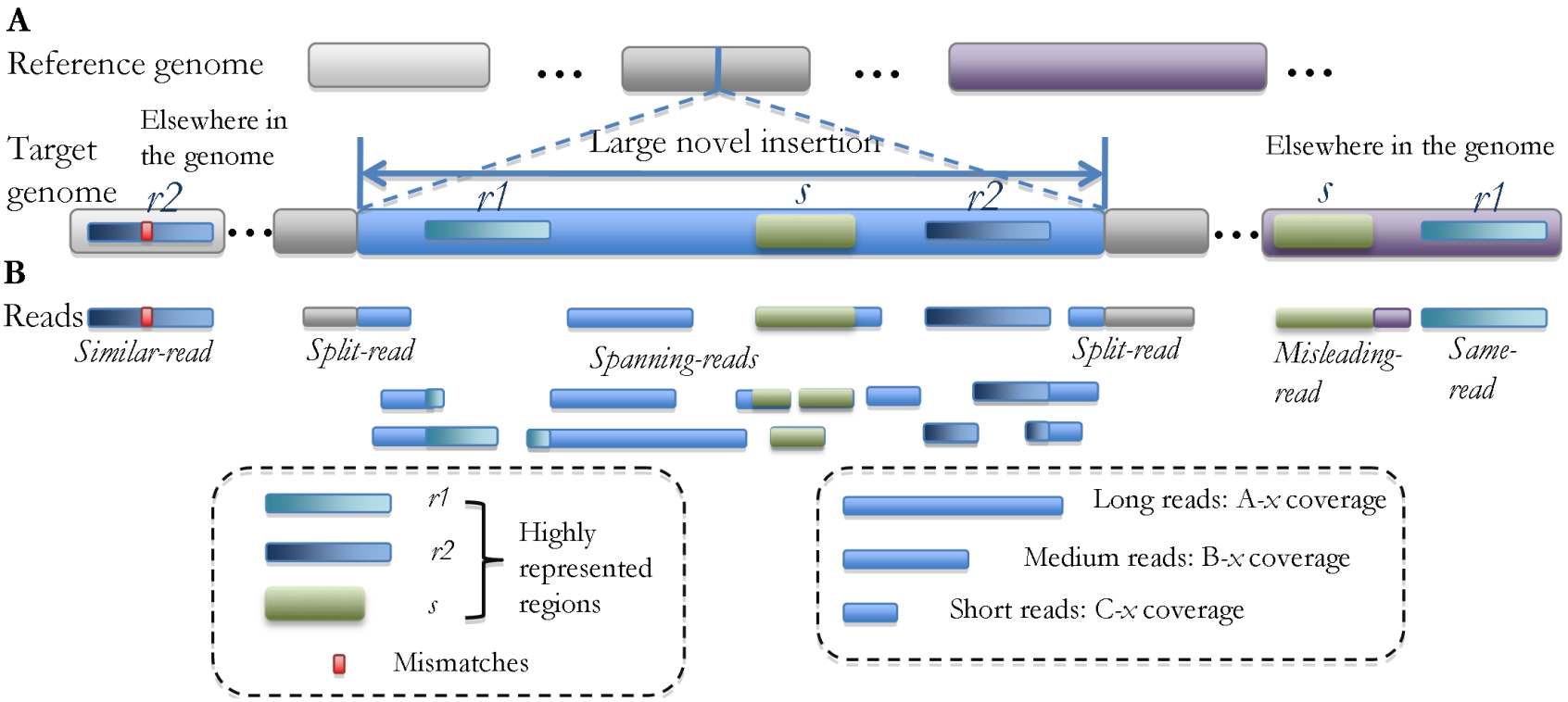
[Korbel et al., GenomeBiol. ('09)]

Local Reassembly



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

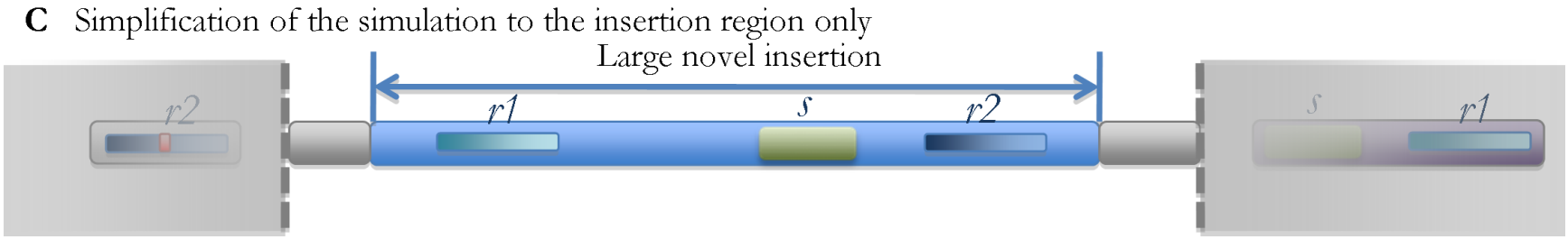
Given a fixed budget, what are the sequencing coverage A, B and C that can achieve the maximum reconstruction rate (on average/worst-case)? Maybe a few long reads can bootstrap reconstruction process.



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

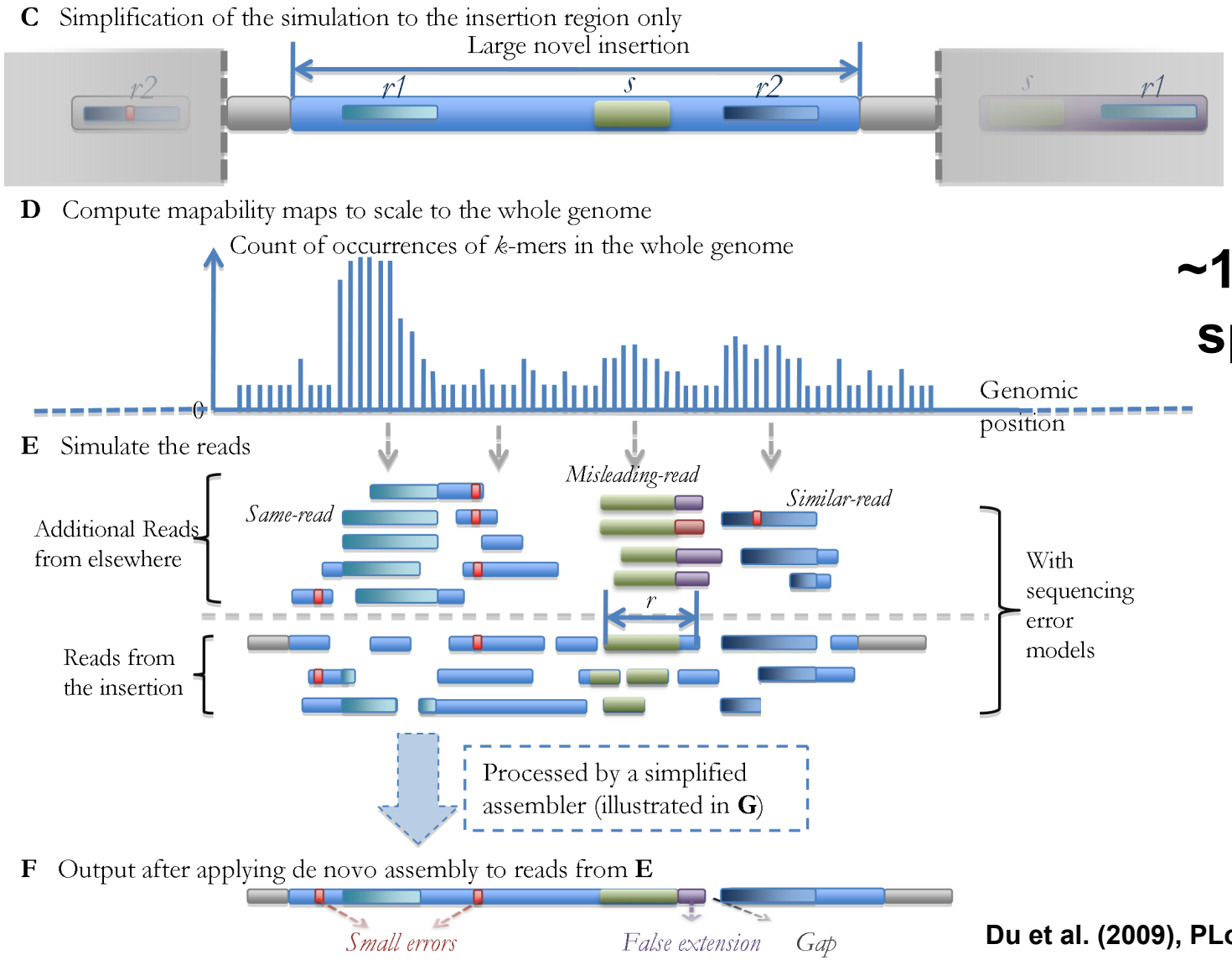
Optimal integration of sequencing technologies:
Need Efficient Simulation

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

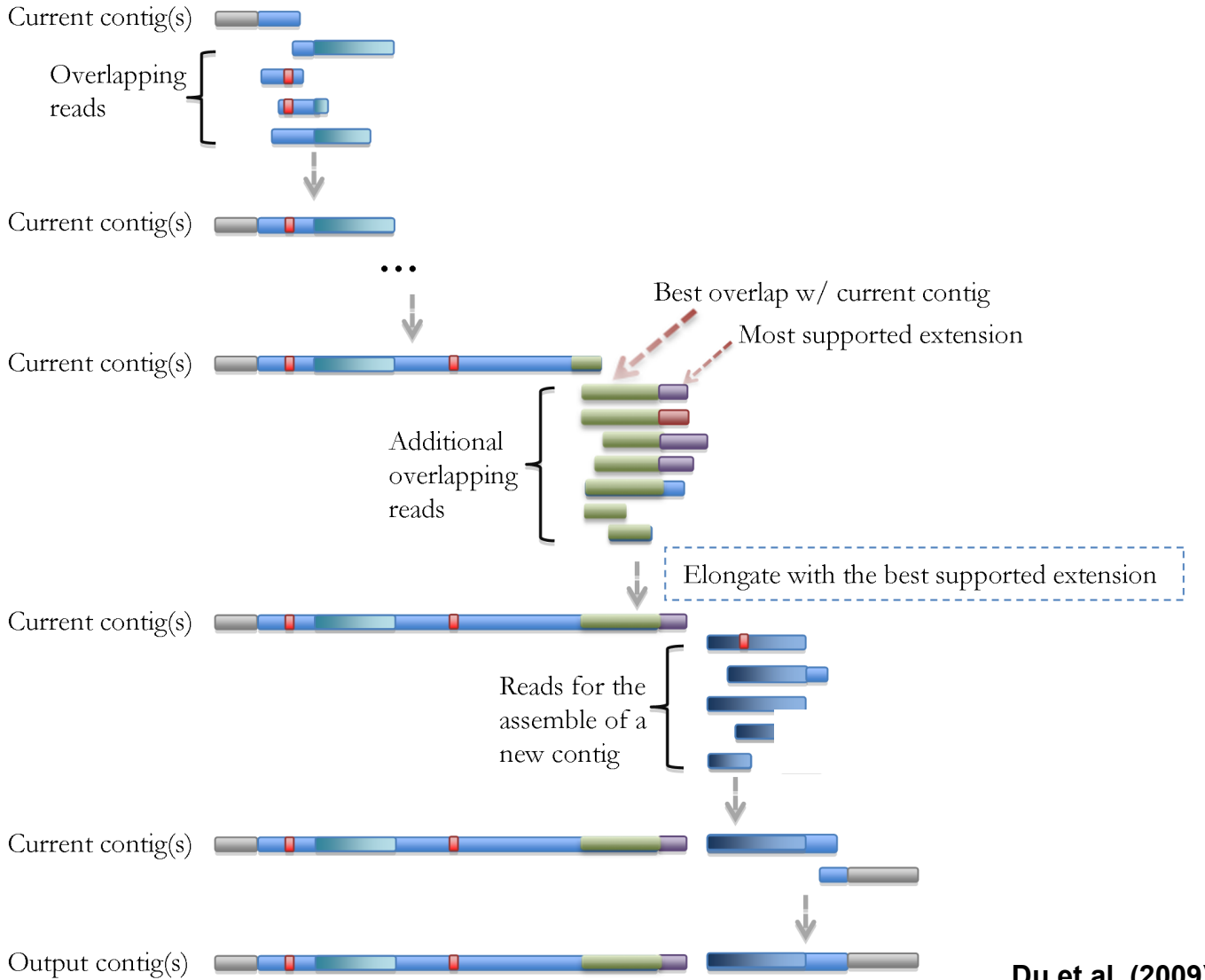


~100,000X speedup

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

Optimal integration of sequencing technologies:
Efficient Simulation using A Simplified Assembler

G Iterative contig elongation with the best supported extension

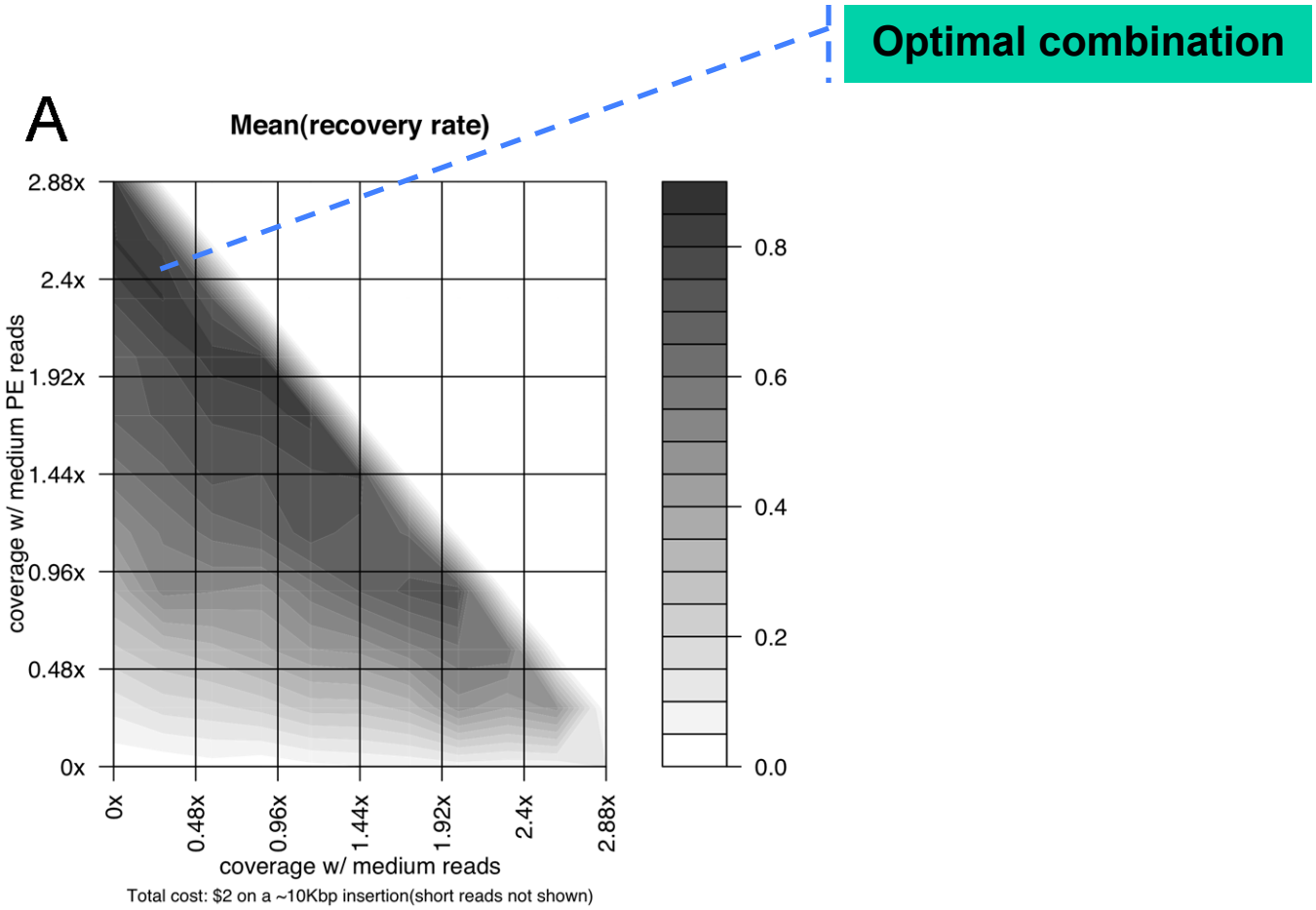


Du et al. (2009), PLoS Comp Biol, in press

Optimal integration of sequencing technologies:

Simulation shows power of PEs

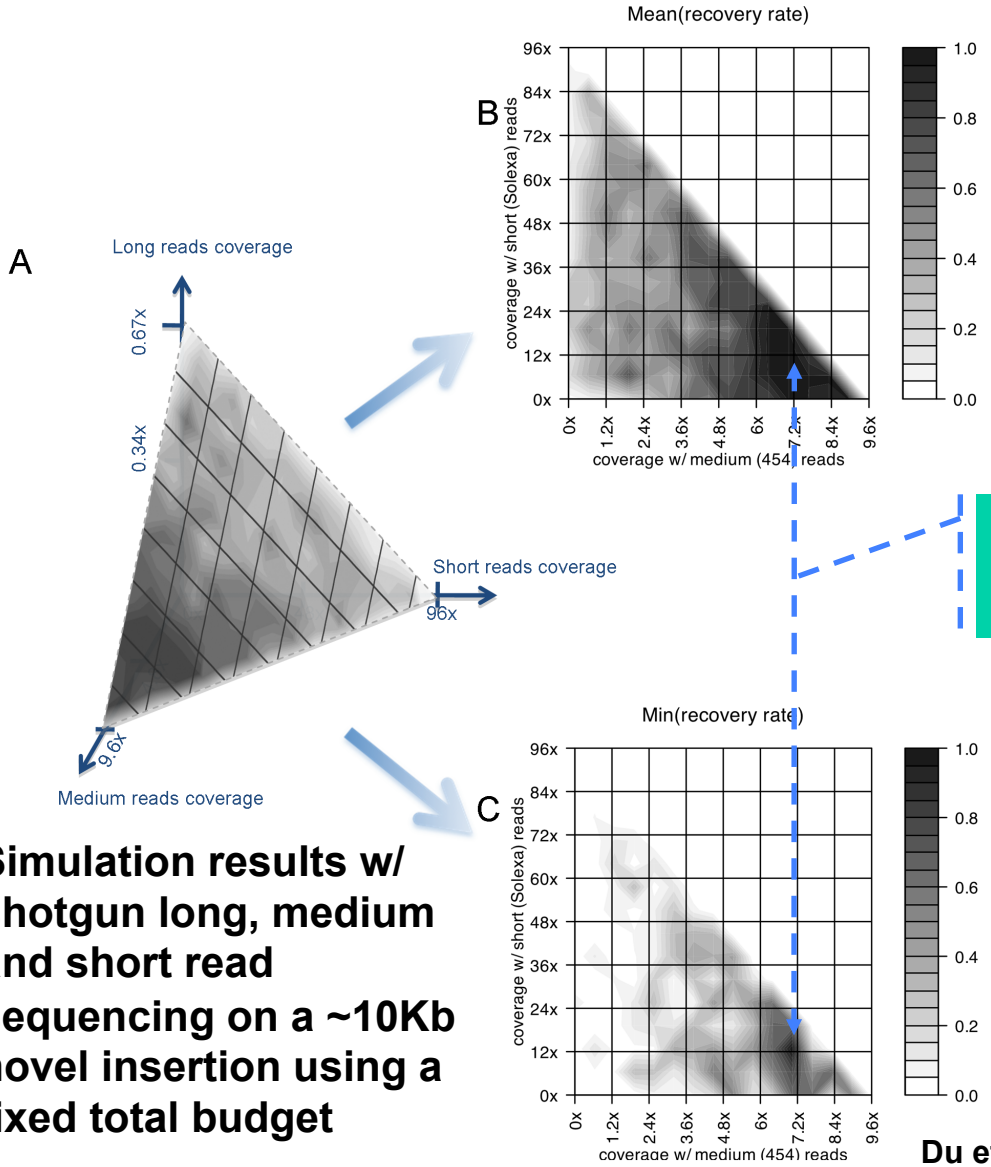
Simulation results w/ shotgun & paired-end reads on the same ~10Kb insertion



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

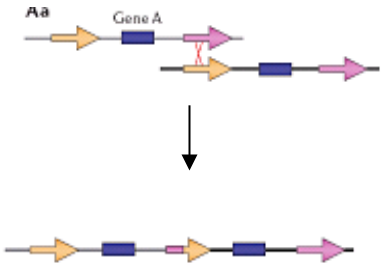
Optimal combination of different technologies

Result dependent on specific parameter setting of different sequencing technologies

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

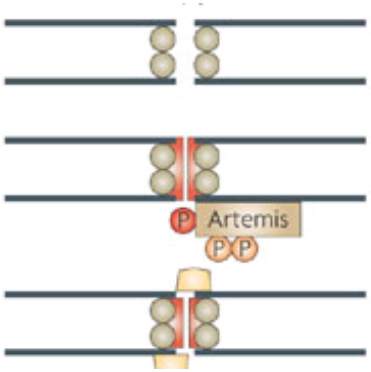
Analyzing Repeated Blocks in the Genome (SDs & CNVs)

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…)

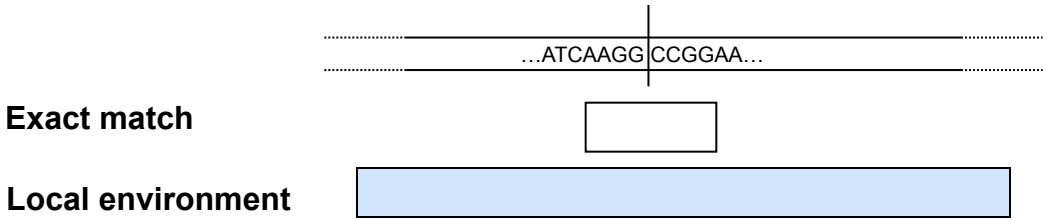


NHEJ
(Non-homologous-end-joining)

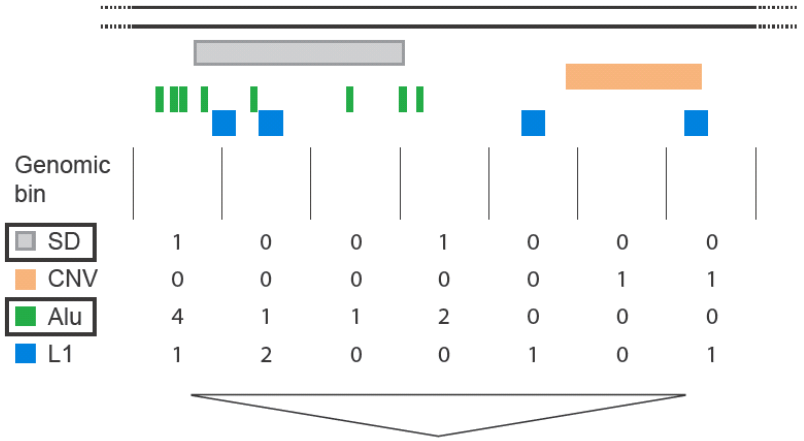
No (flanking) repeats.
In some cases <4bp microhomologies

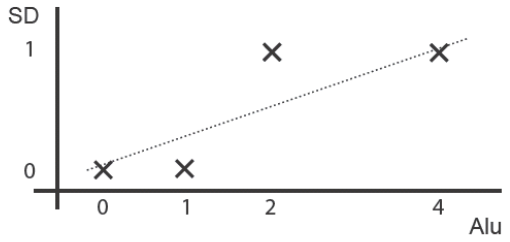
PERFORM LARGE SCALE CORRELATION ANALYSIS TO DETECT REPEAT SIGNATURES OF SDs AND CNVs

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



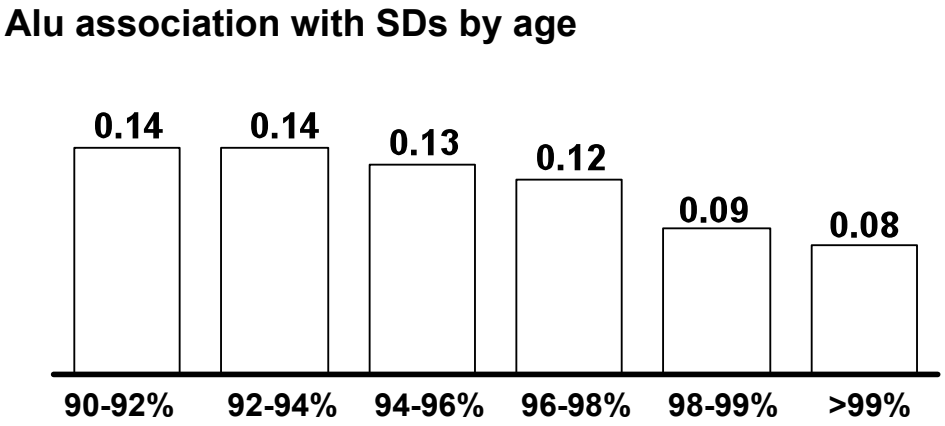
1. Survey a range of genomic features
2. Count the number of features in each genomic bin (100kb)
3. Calculate correlations / enrichments using robust stats





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

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

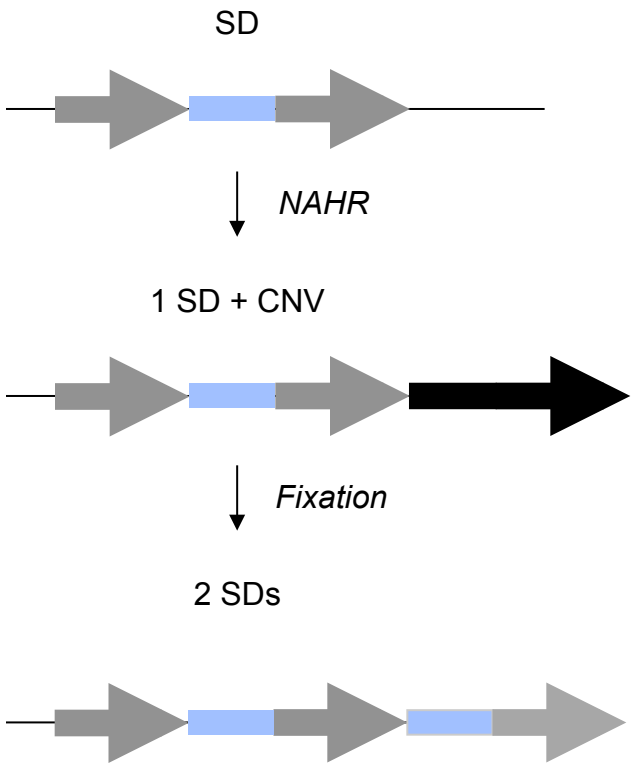


- The co-localization of Alu elements with SDs is highly significant.
- Older SDs have a much higher association with Alus than younger SDs.
- Hence it is likely, that Alu elements were more active in mediating NAHR in the past (consistent with the Alu burst)

[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

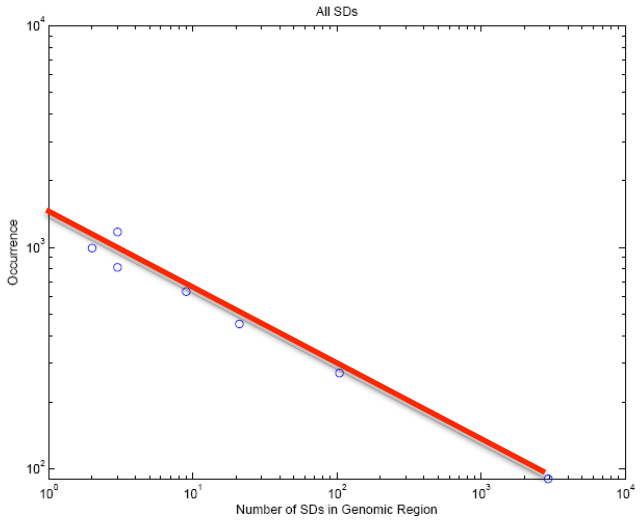
Hypothesis



"SD selfpropagation"

Corollary

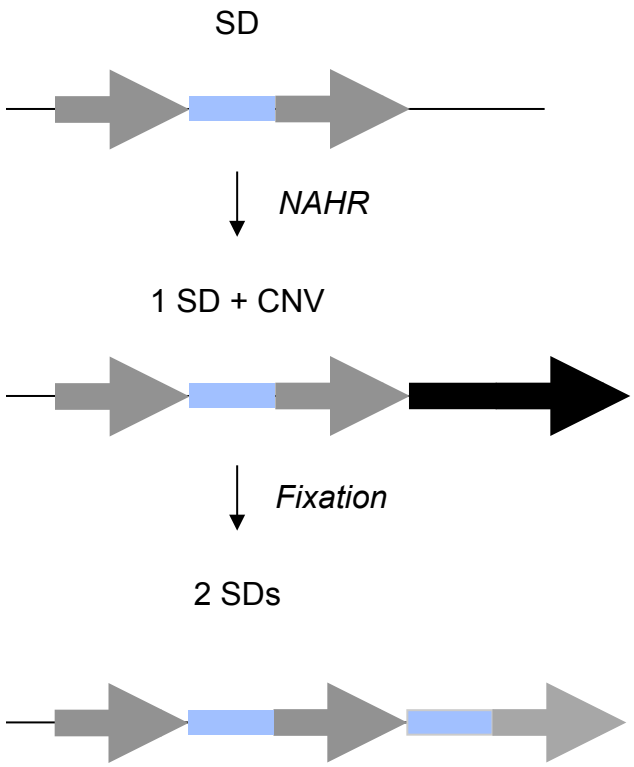
- SDs can mediate NAHR and lead to the formation of CNVs
- CNVs can become fixed and then be SDs
- Such mechanisms ("preferential attachment") are well studied in physics and should leads a very skewed ("power-law") distribution of SDs.



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

FOCUSSING ON SDS: SDs COLOCALIZE WITH EACH OTHER

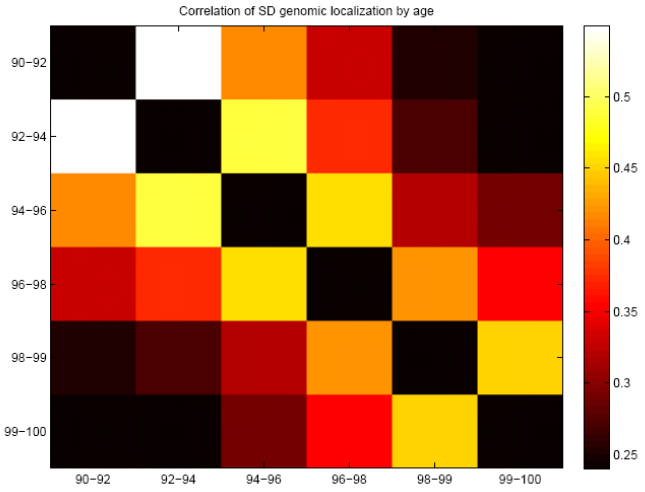
Hypothesis



"SD self-propagation"

Corollary

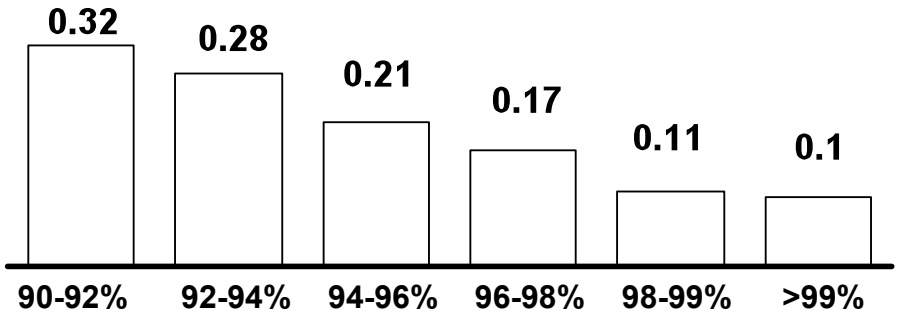
- SDs can mediate NAHR and lead to the formation of CNVs
- CNVs can become fixed and then be SDs
- SDs of similar age should co-localize better with each other:



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

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

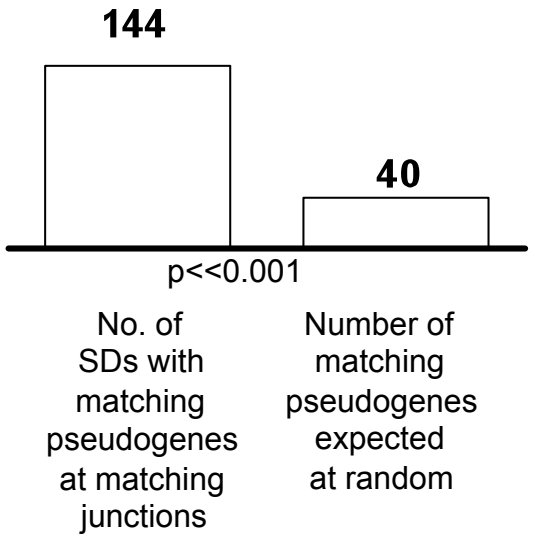
Pseudogene association with SDs by age



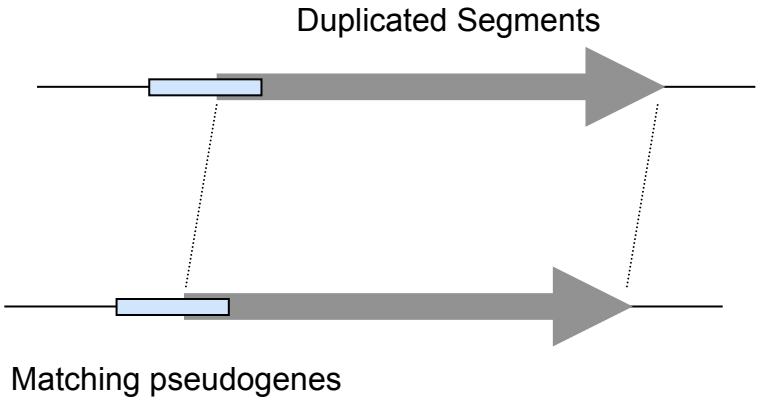
Duplicated pseudogenes associated with SDs, particularly older ones

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

Processed pseudogenes at SD junctions



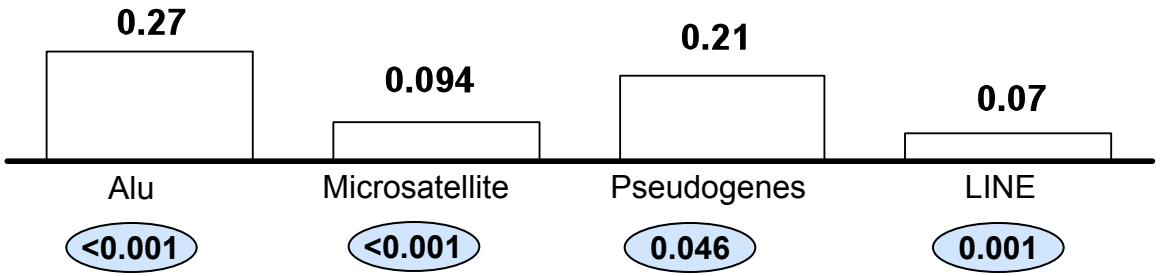
Processed Pseudogenes: serving as repeats for mediating NAHR



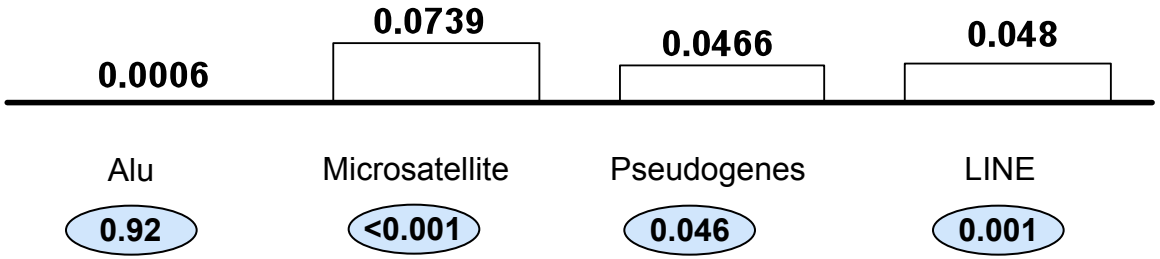
[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1]

ASSOCIATIONS ARE DIFFERENT FOR SDs AND CNVs

SD association with repeats

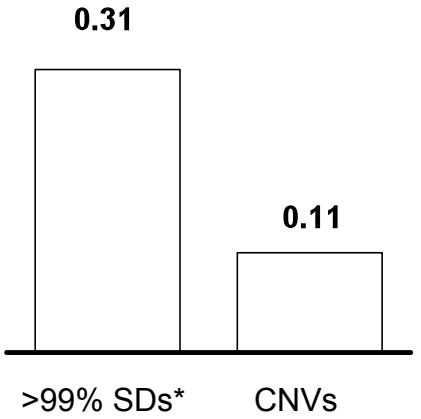


CNV association with repeats

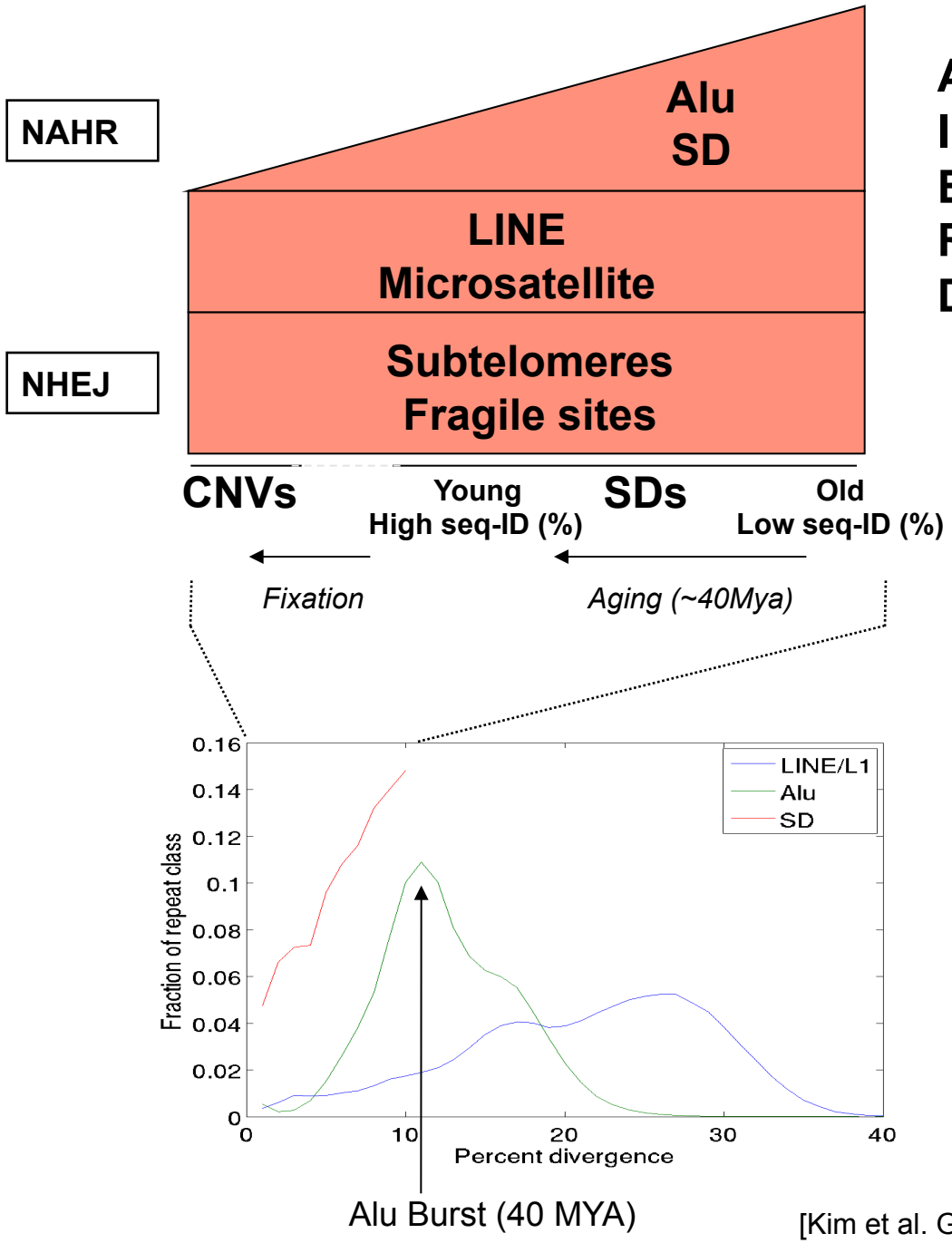


CNVs ARE LESS ASSOCIATED WITH SDs THAN THE GENERAL SD TREND

CNV Association with SDs



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



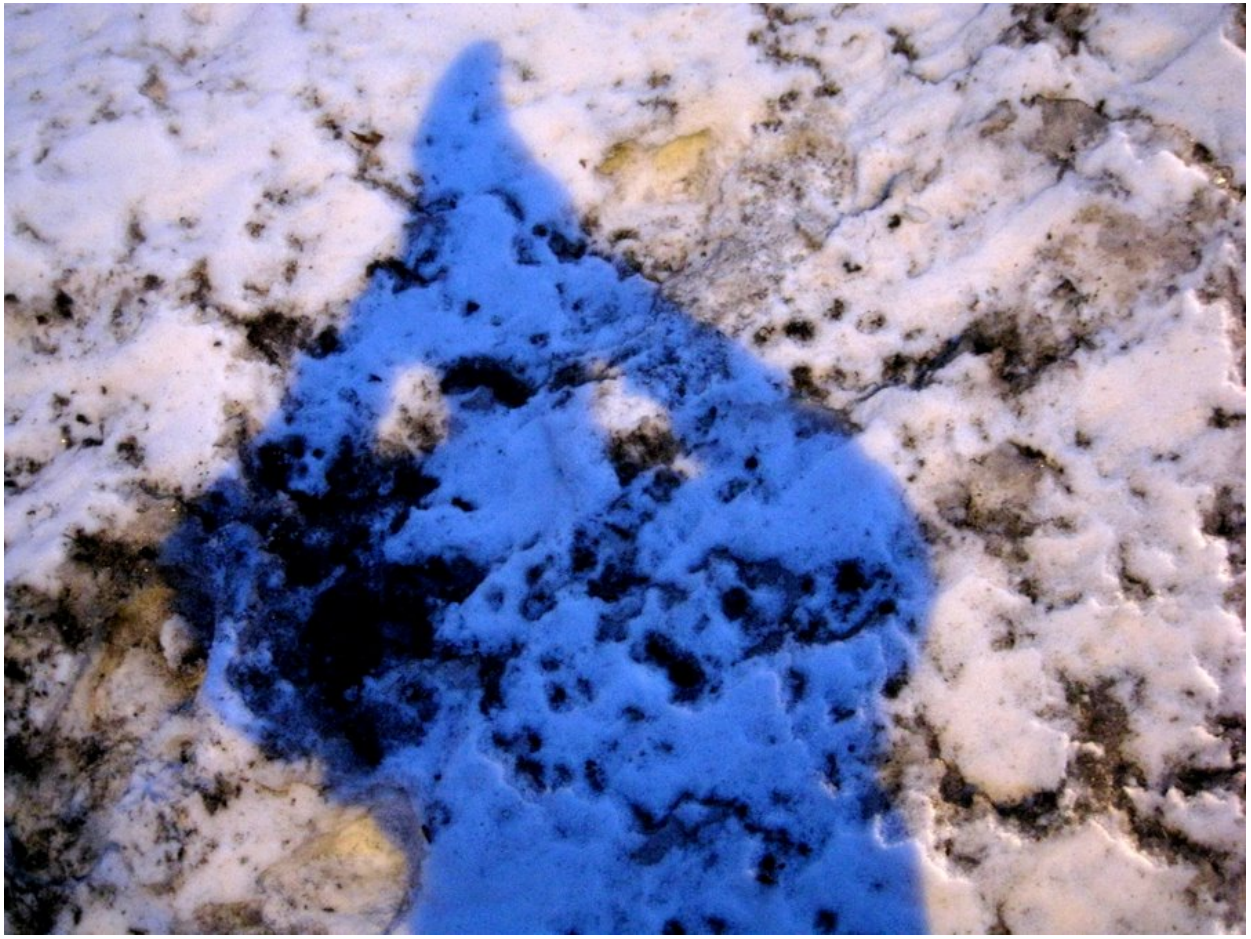
AFTER THE ALU BURST, THE IMPORTANCE OF ALU ELEMENTS FOR GENOME REARRANGEMENT DECLINED RAPIDLY

- 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

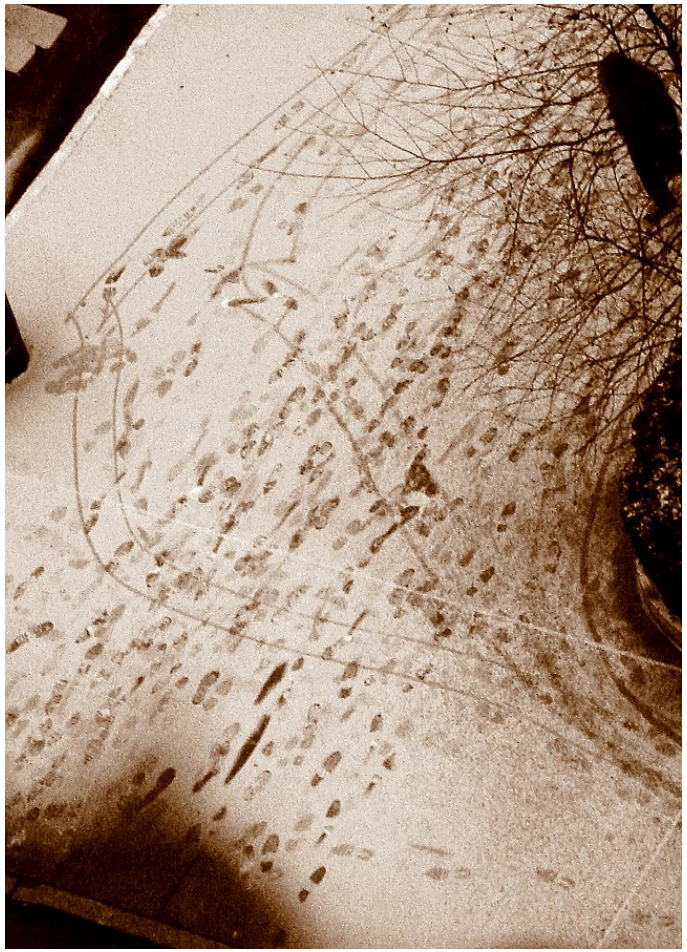
[Kim et al. Gen. Res. ('08), arxiv.org/abs/0709.4200v1]

Summary:

Looking Back Over the Talk



Outline



- Calling Variable Blocks in Genome (CNVs,SDs)
  - Calling them with various signal processing approaches
- Analyzing Association of Variable Blocks with repeats, in relation to formation mechanisms

Signal Processing #2: 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

GenomeTECH.gersteinlab.org

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

More Information on this Talk

TITLE: Human Genome Annotation

SUBJECT: GenomeAssembly

DESCRIPTION:
IEEE International Conference on Bioinformatics & Biomedicine (BIBM-2009), 2009.11.02, 15:45-16:15; [I:**BIBM**] (Short adaption of GenomeTechAnnote talk, building on [I:**UCSC**] focusing just on SV reconstruction and analysis, includes updates **msb*** . Takes 29' with 2 questions, or ~24' of talk time with **sdcnvcorr*** sect.)

(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** .