Computational Methods for SV Characterization

Mark Gerstein

Computational Methods for SV Characterization

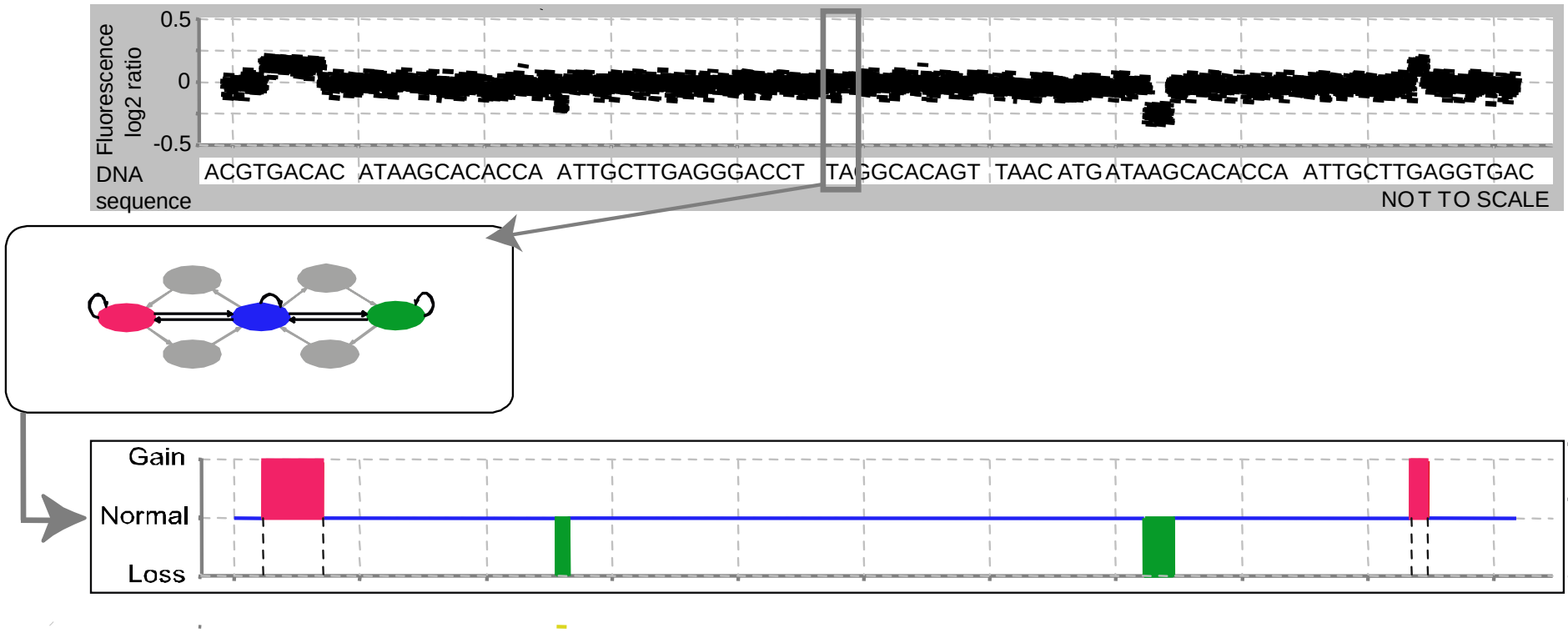
Segmenting Array CGH data

Building a PEM pipeline

Correlating SVs and SDs with Repeats

BreakPtr HMM

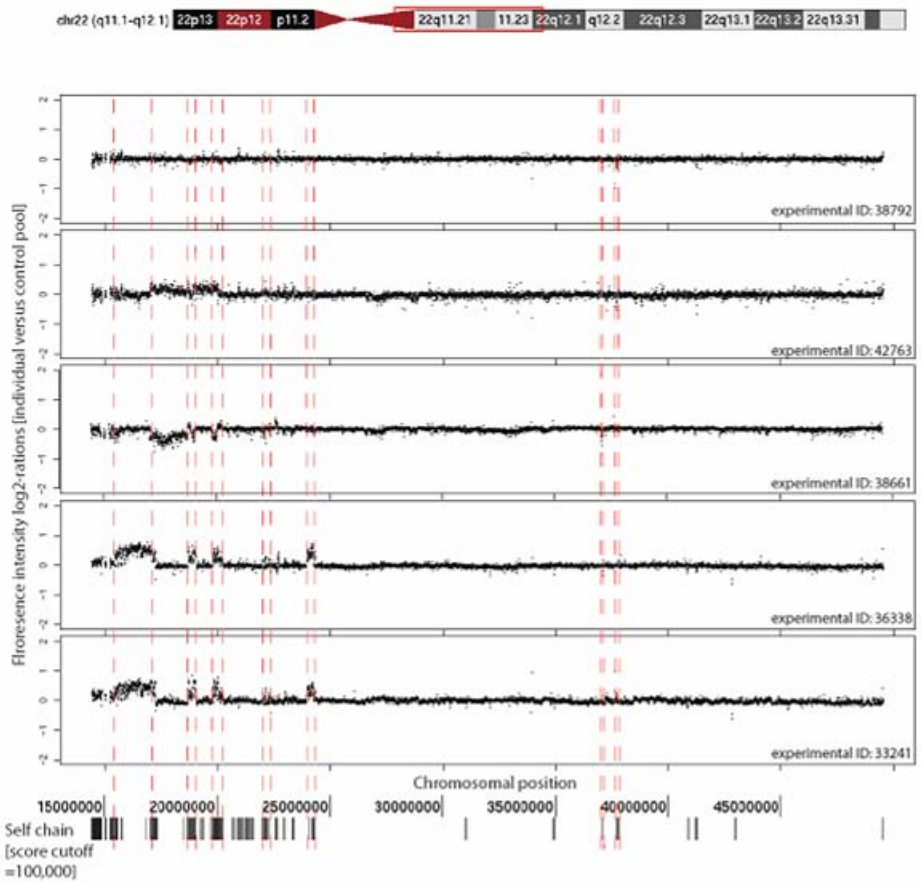
- To get highest resolution on breakpoints need to smooth & segment the signal
- BreakPtr: prediction of breakpoints, dosage and cross-hybridization using a system based on Hidden Markov Models



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

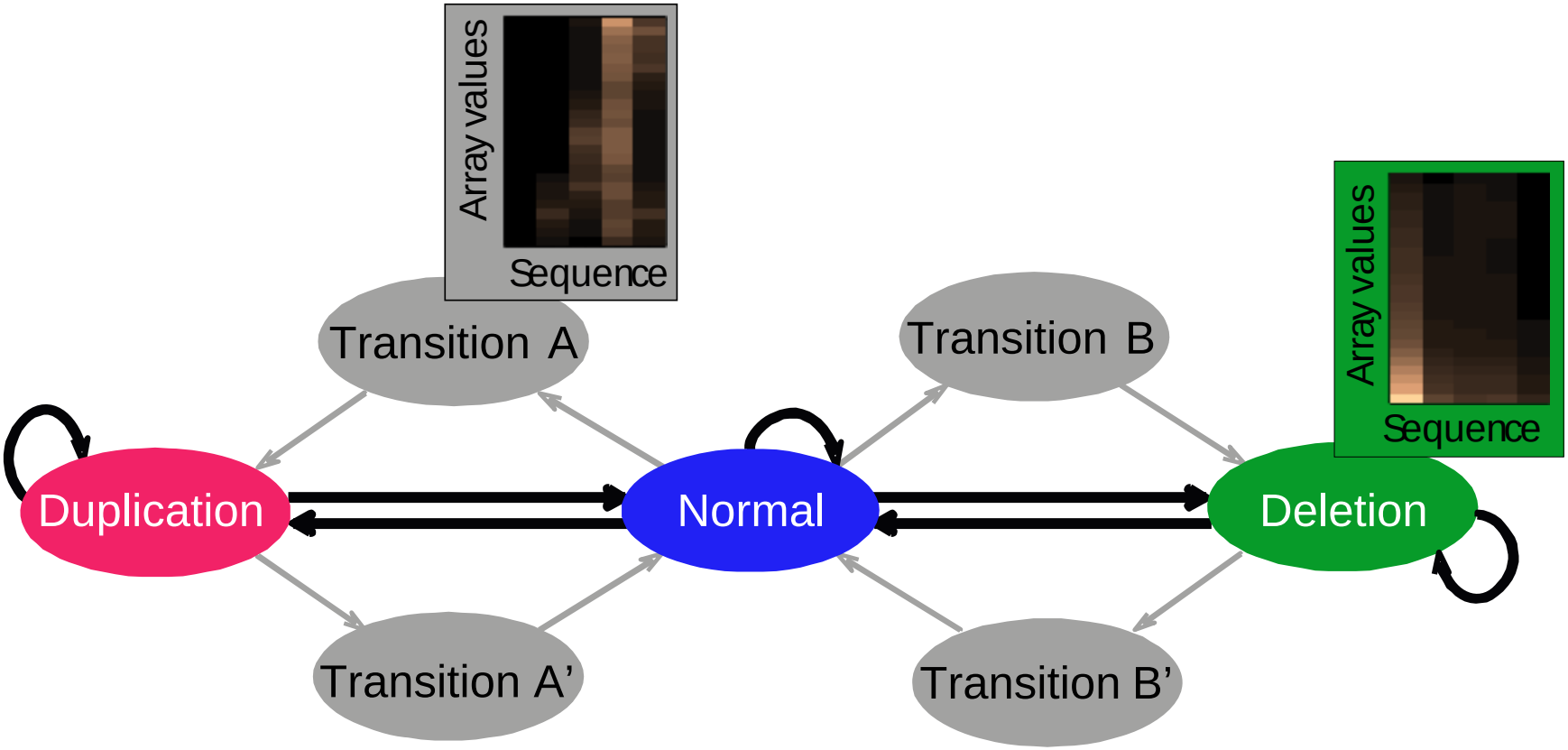
http://breakptr.gersteinlab.org

High resolution of tiling arrays allows statistical integration of nucleotide sequence patterns



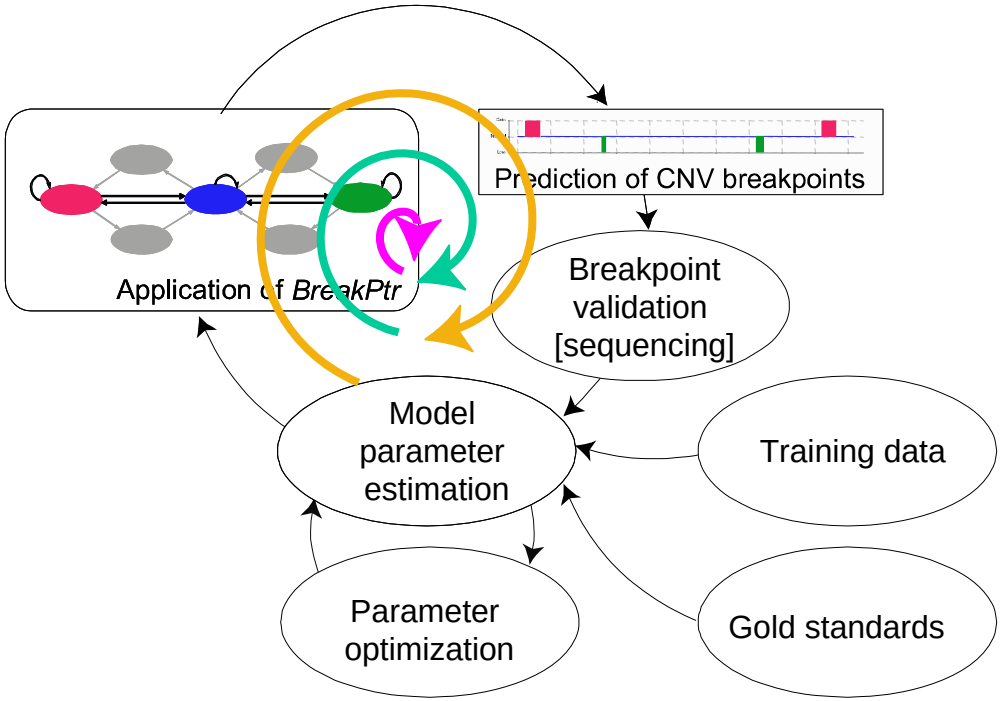
>4-fold enrichment of the breakpoints of copy number variants near segmental duplications (SDs)
[e.g. Sharp *et al.*, *Am. J. Hum. Genet.* 2005; 77:78-88].

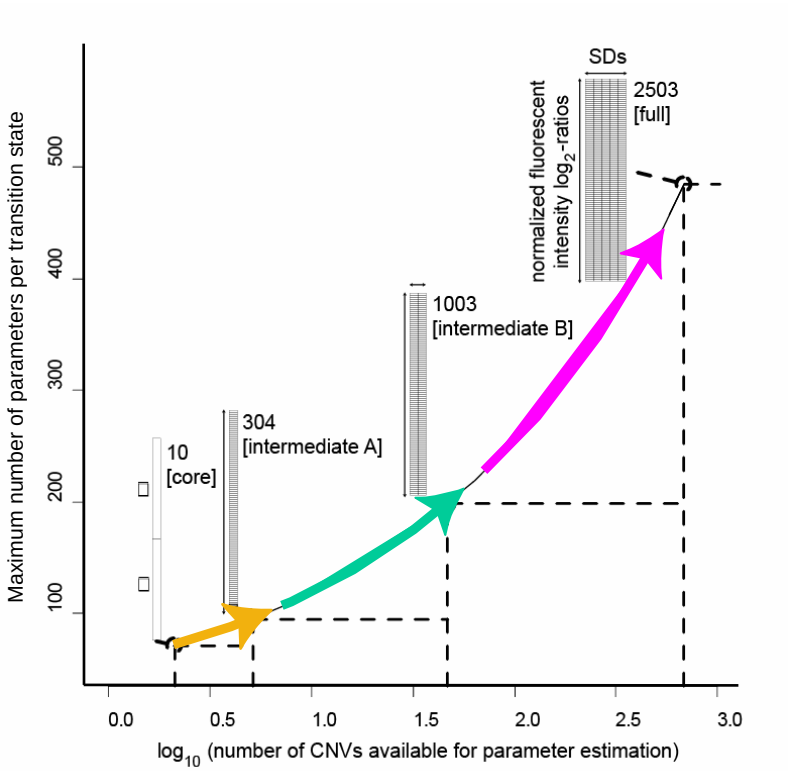
BreakPtr statistically integrates array signal and DNA sequence signatures (using a discrete-valued bivariate HMM)



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

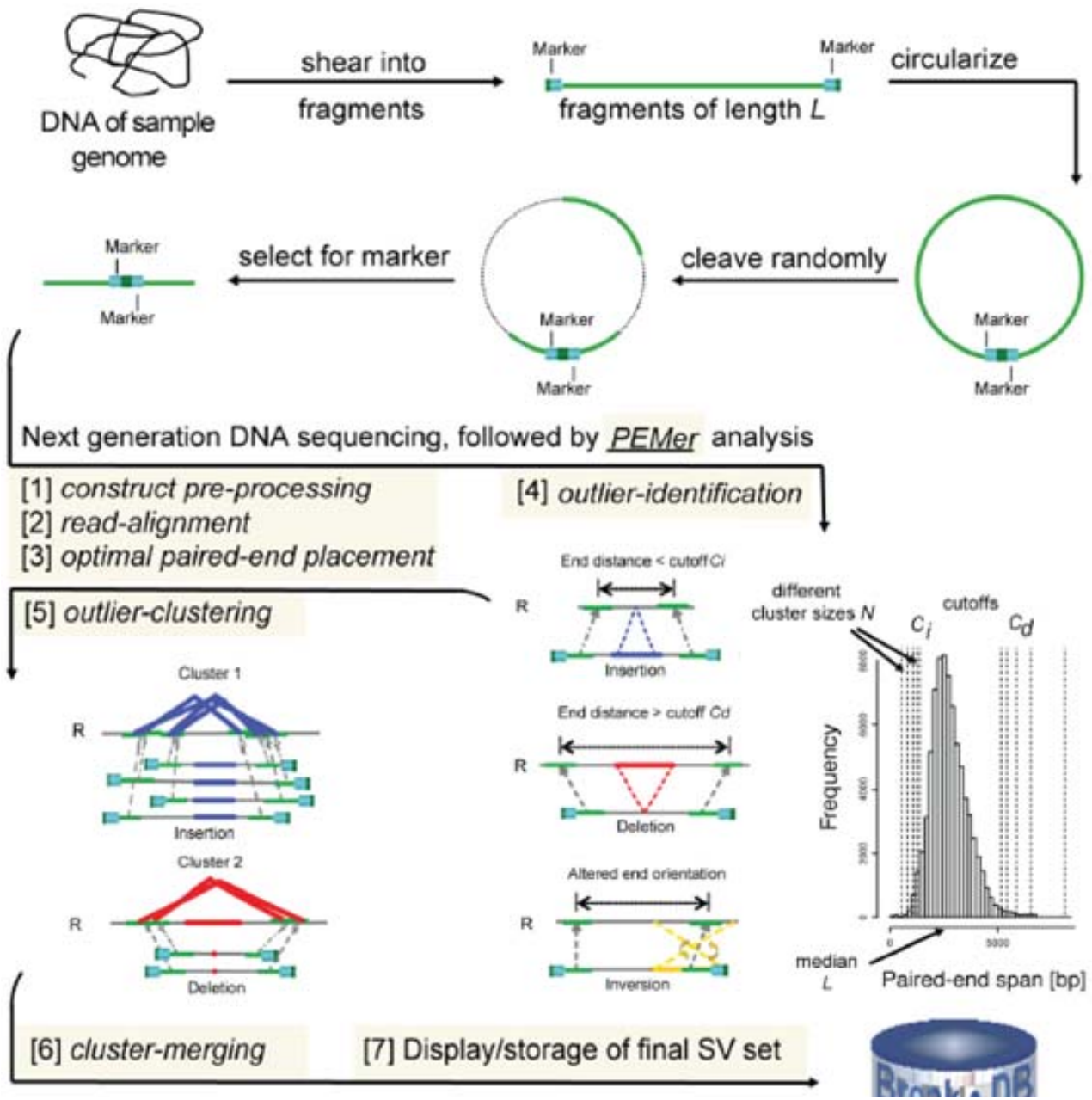
No improvement in accuracy with higher resolution (9nt tiling)

HMM optimized iteratively (using Expectation Maximization, EM)

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

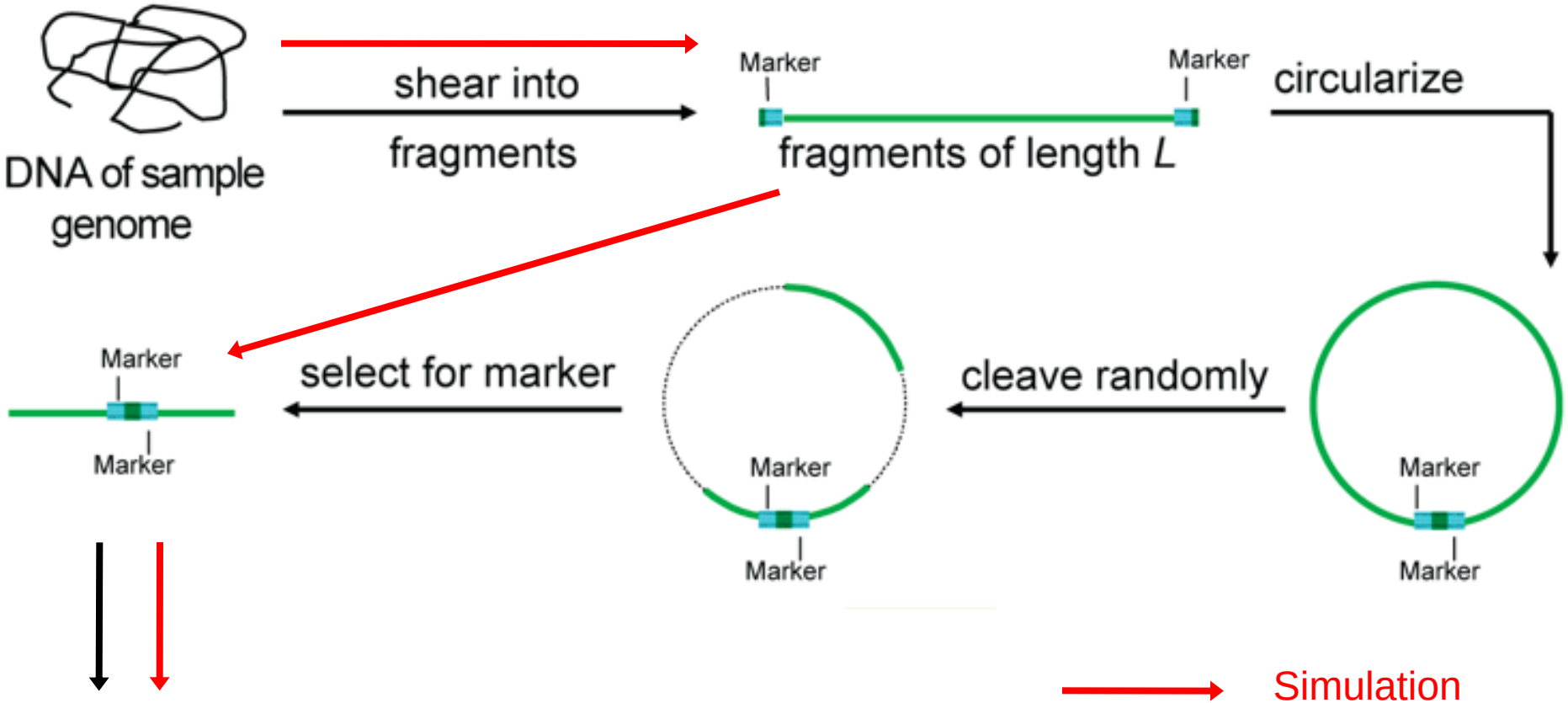
Moving Beyond Arrays, Computational Methods for Next-Generation Sequencing: Paired End Mapping to Find SVs

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



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

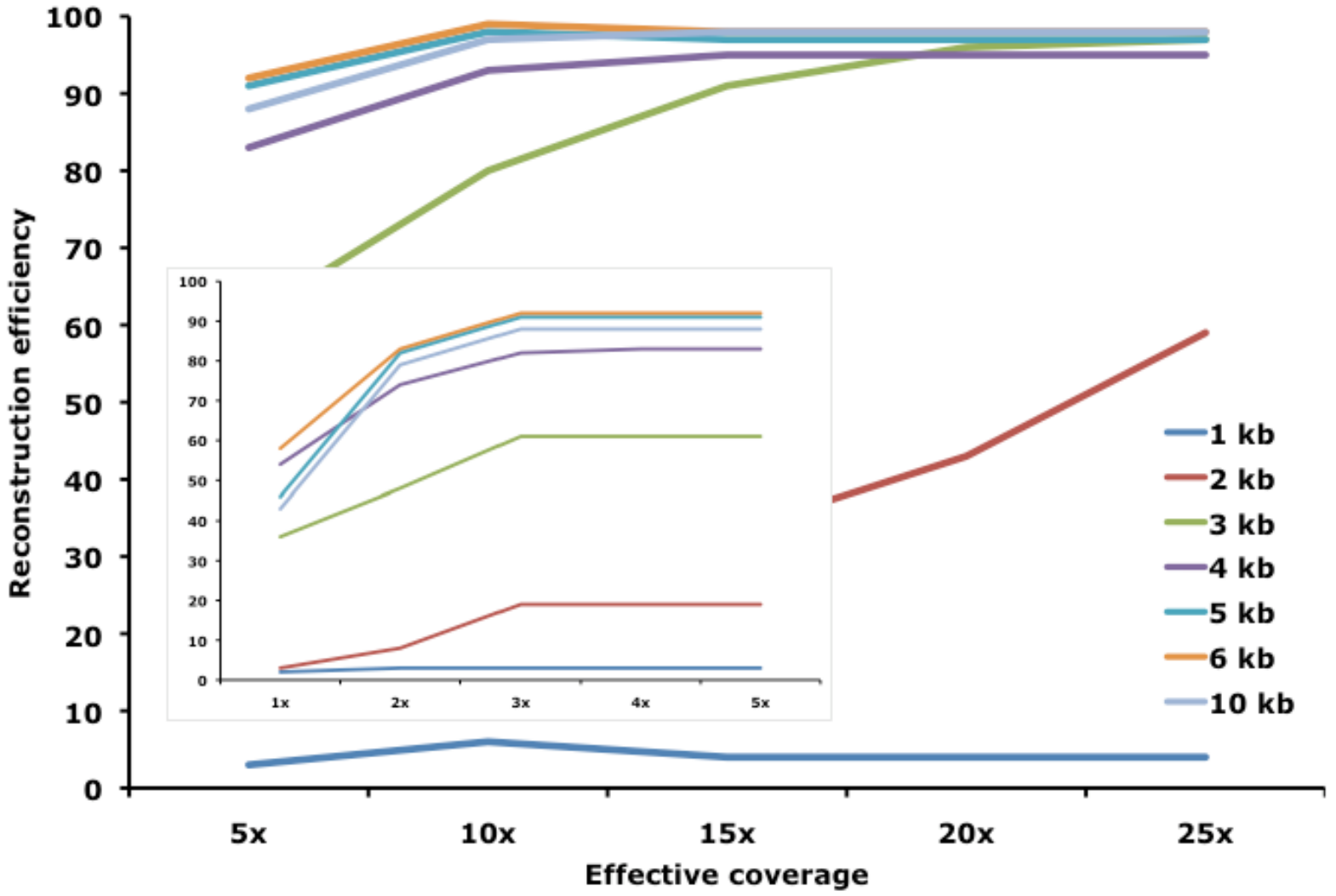
Simulation strategy



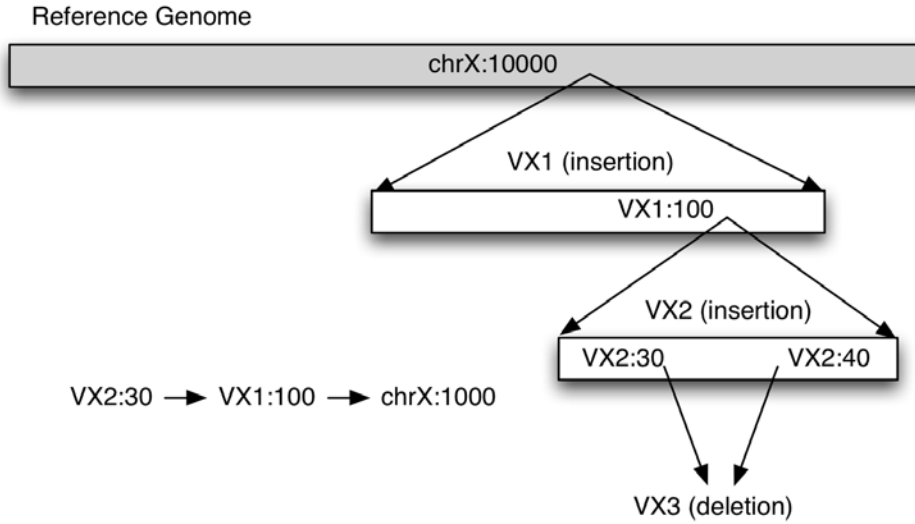
[Korbel et al., GenomeBiol. (submitted)]

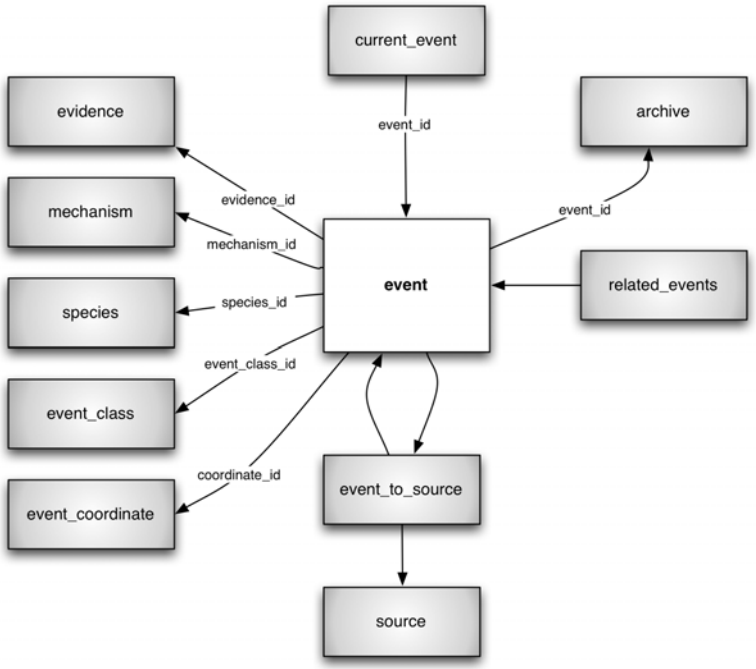
[Korbel et al., GenomeBiol. (submitted)]

Reconstruction efficiency at different coverage



Building a Database of Variants: Complexities



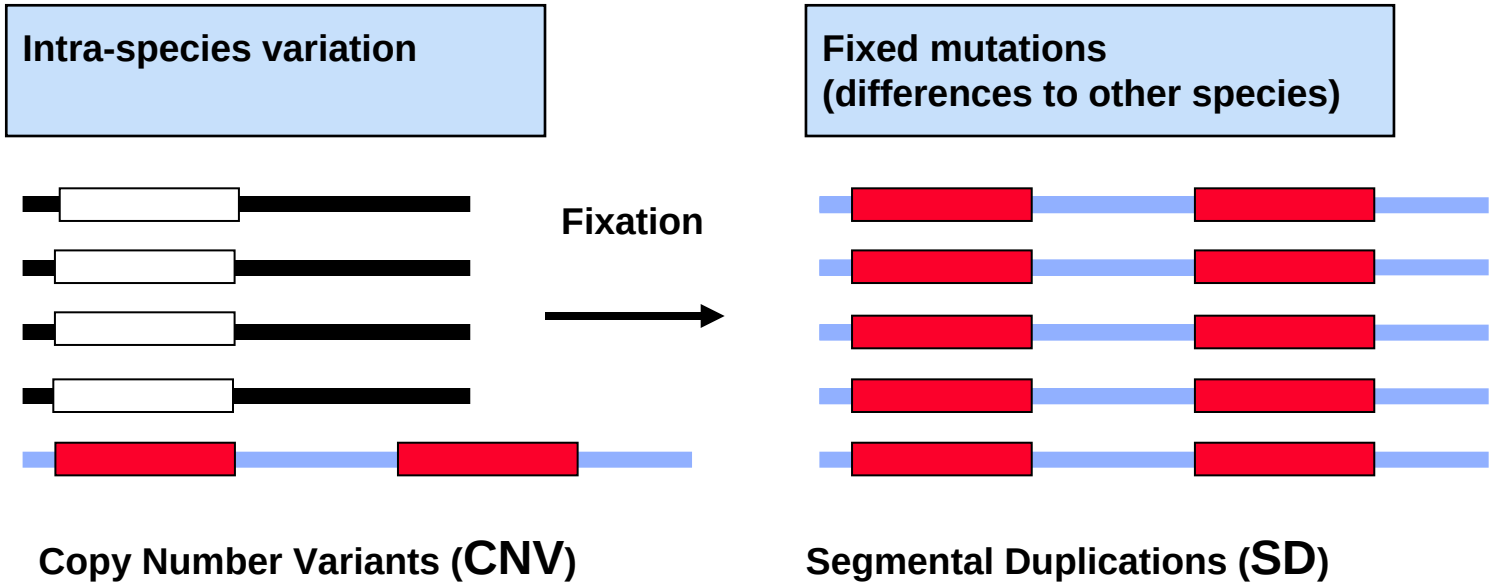


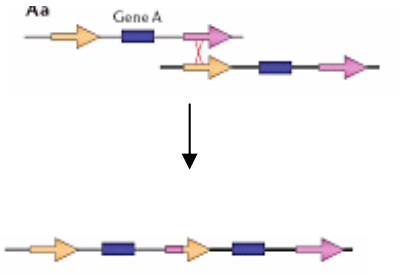
[Korbel et al., GenomeBiol. (submitted)]

Analyzing Duplications in the Genome (SDs & CNVs)

pers. **photo**, see streams.gerstein.info

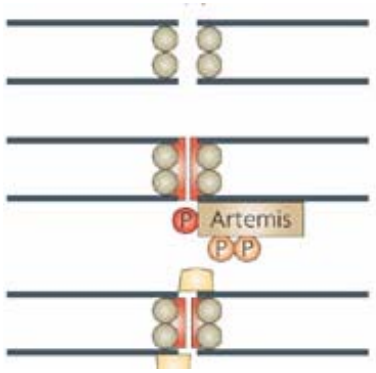
SEGMENTAL DUPLCATIONS AND COPY NUMBER VARIANTS ARE RELATED PHENOMENA AND HAVE BEEN CREATED BY SEVERAL DIFFERENT MECHANISMS





NAHR
(Non-allelic homologous recombination)

Flanking repeat
(e.g. Alu, LINE…)

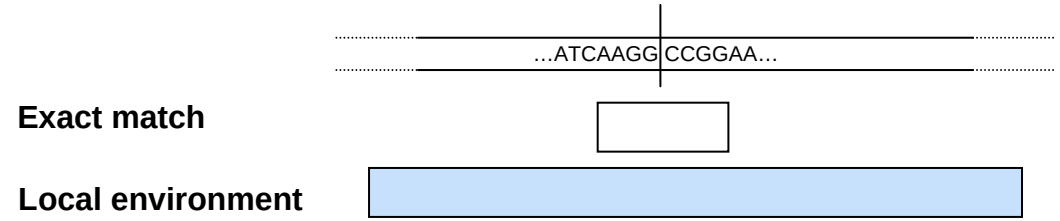


NHEJ
(Non-homologous-end-joining)

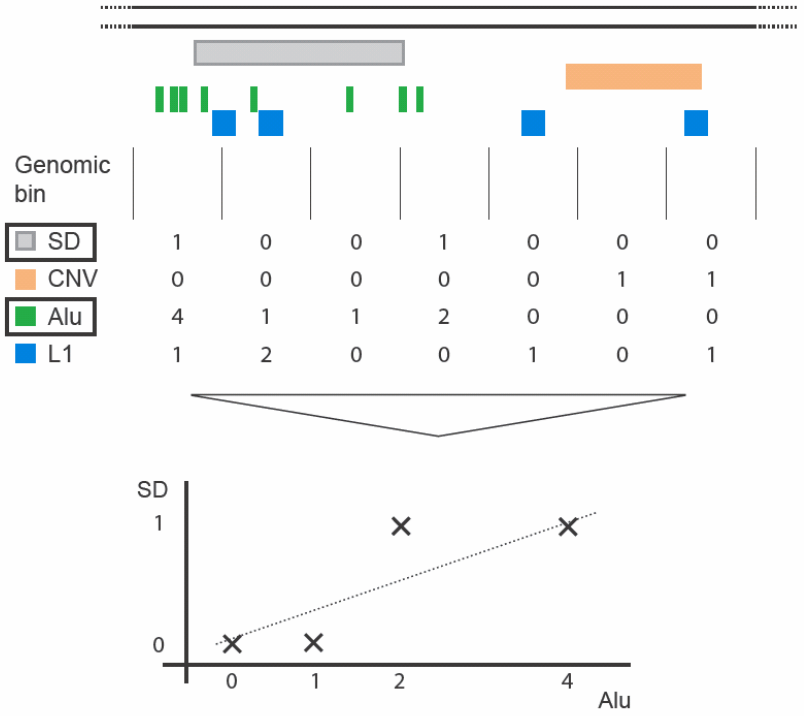
No (flanking) repeats.
In some cases <4bp microhomologies

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

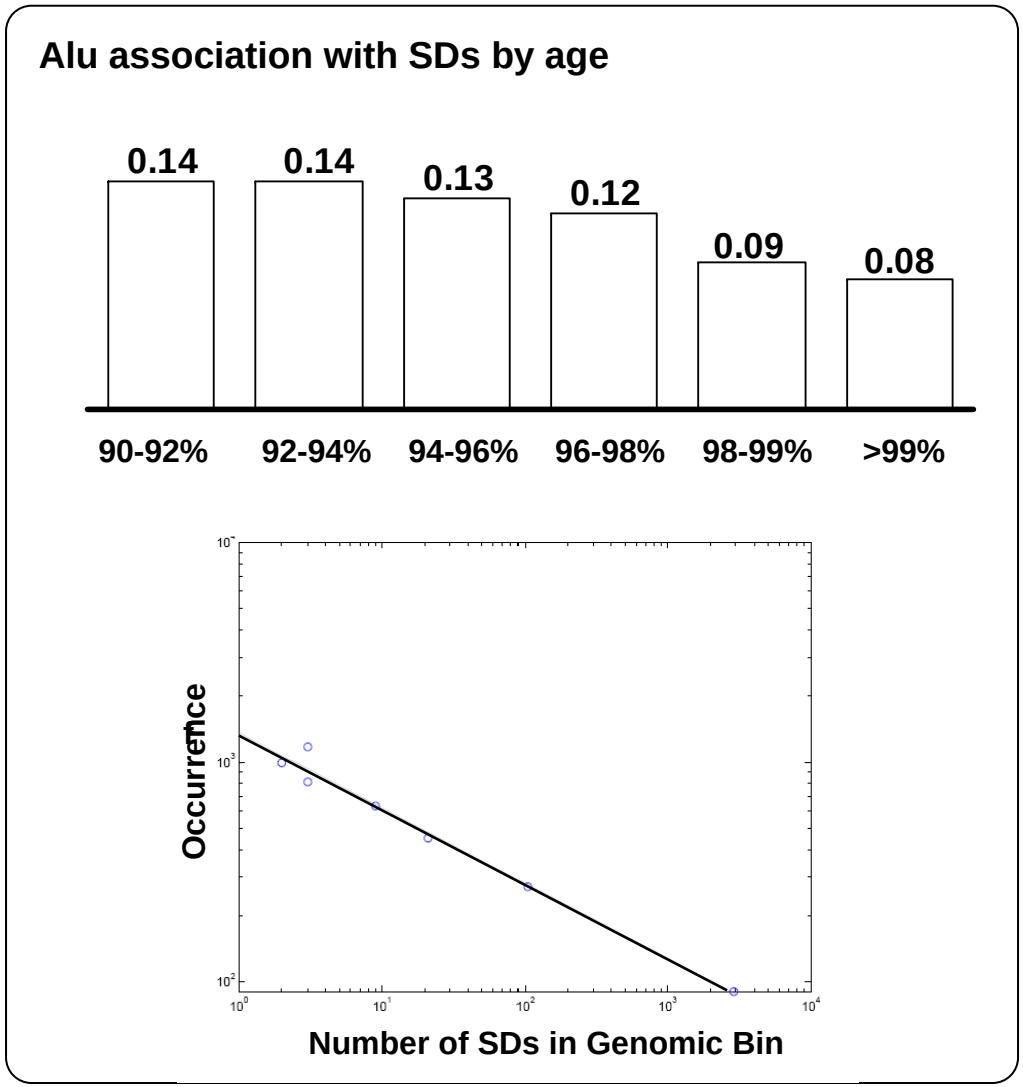


1. Survey a range of genomic features
2. Count the number of features in each genomic bin (100kb)
3. Calculate correlations / enrichments using robust stats



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

SDs ARE CORRELATED WITH ALUS AND OTHER SDs

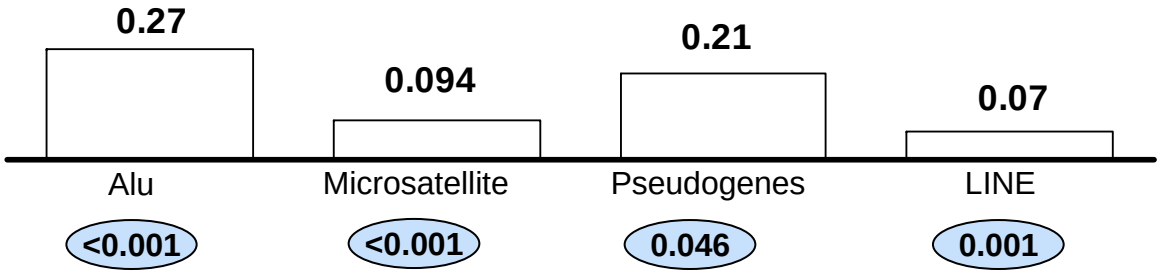


- The co-localization of Alu elements with SDs is highly significant.
- Older SDs have a much higher association with Alus than younger SDs.
- SDs can mediate NAHR and lead to the formation of CNVs
- Such mechanisms ("preferential attachment") are well studied in physics and should leads a very skewed ("power-law") distribution of SDs.
- Hotspots

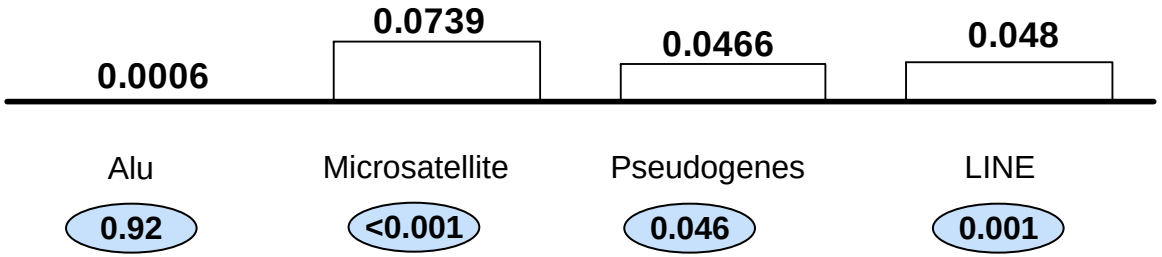
[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1]

ASSOCIATIONS ARE DIFFERENT FOR SDs AND CNVs

SD association with repeats

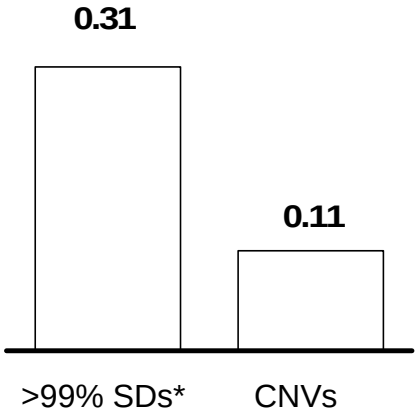


CNV association with repeats



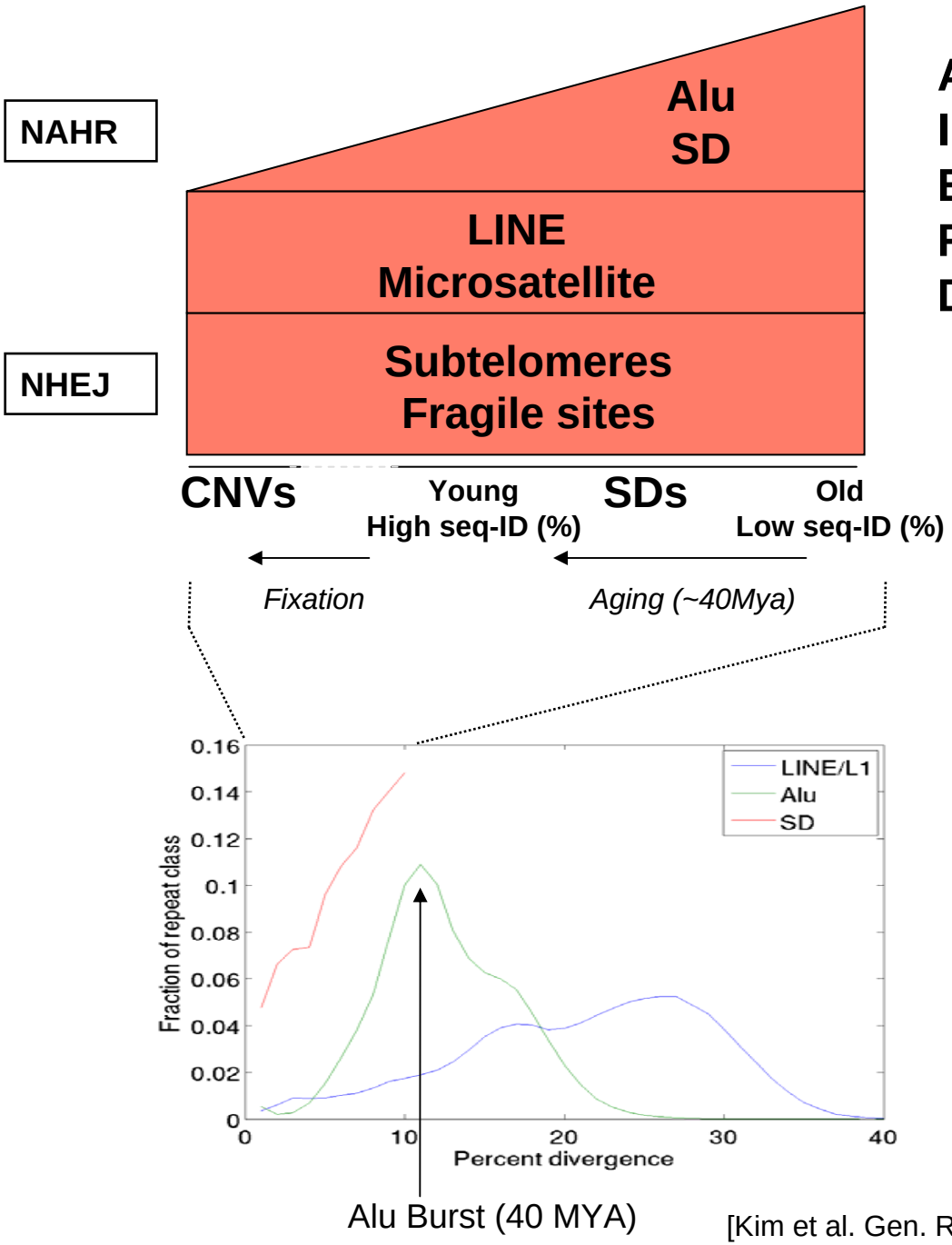
CNVs ARE LESS ASSOCIATED WITH SDs THAN THE GENERAL SD TREND

CNV Association with SDs



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

Future Directions

- Simulations of SV Assembly
- Analysis of Split Reads
- Detailed Analysis of SV and CNVs with Genomic Features

CEGS Informatics Credits

- Array Corrections
  - ◊ J Rozowsky
  - ◊ T Royce
  - ◊ M Seringhaus
- PEMer, SD-CNV, BreakPtr
  - ◊ P Kim
  - ◊ J Korbel
  - ◊ J Du
  - ◊ X Mu
  - ◊ A Abyzov
  - ◊ N Carriero
- Experimental
  - ◊ M Snyder
  - ◊ S Weissman
  - ◊ A Urban