Chiara M[1], Pesole G[2,3], Horner DS[1]

1 Univeristà degli Studi di Milano. 2 Università degli Studi di Bari. 3 IBBE CNR Bari

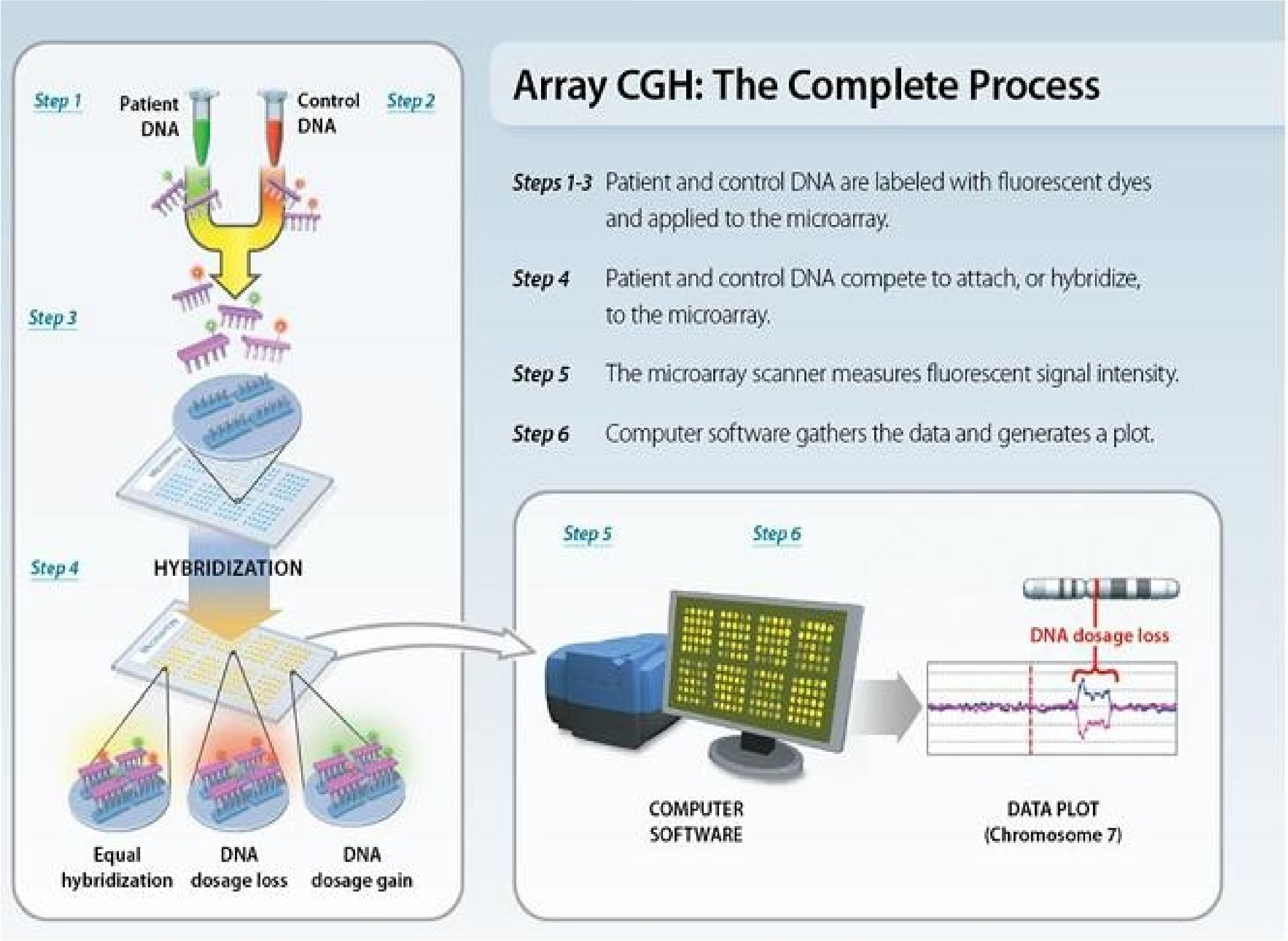
Improved detection of SV using Support Vector Machines (SVM)

- Next generation Sequencing (Illumina PE)
- Structural Variations
- Support Vector Machines
- Bioinformatics

Structural Variations

- Any DNA sequence alteration other than a single nucleotide substitution:
  - Copy number variations (CNV);
  - Insertions/deletions (indels);
  - Translocations;
  - Inversions
- Human genomes differ more as a consequence of structural variation than of single-base-pair differences
  - Contribute to heritable genetic diseases and cancers
  - role in speciation?

High troughput SV detection



Array CGH: The Complete Process

Steps 1-3 Patient and control DNA are labeled with fluorescent dyes and applied to the microarray.

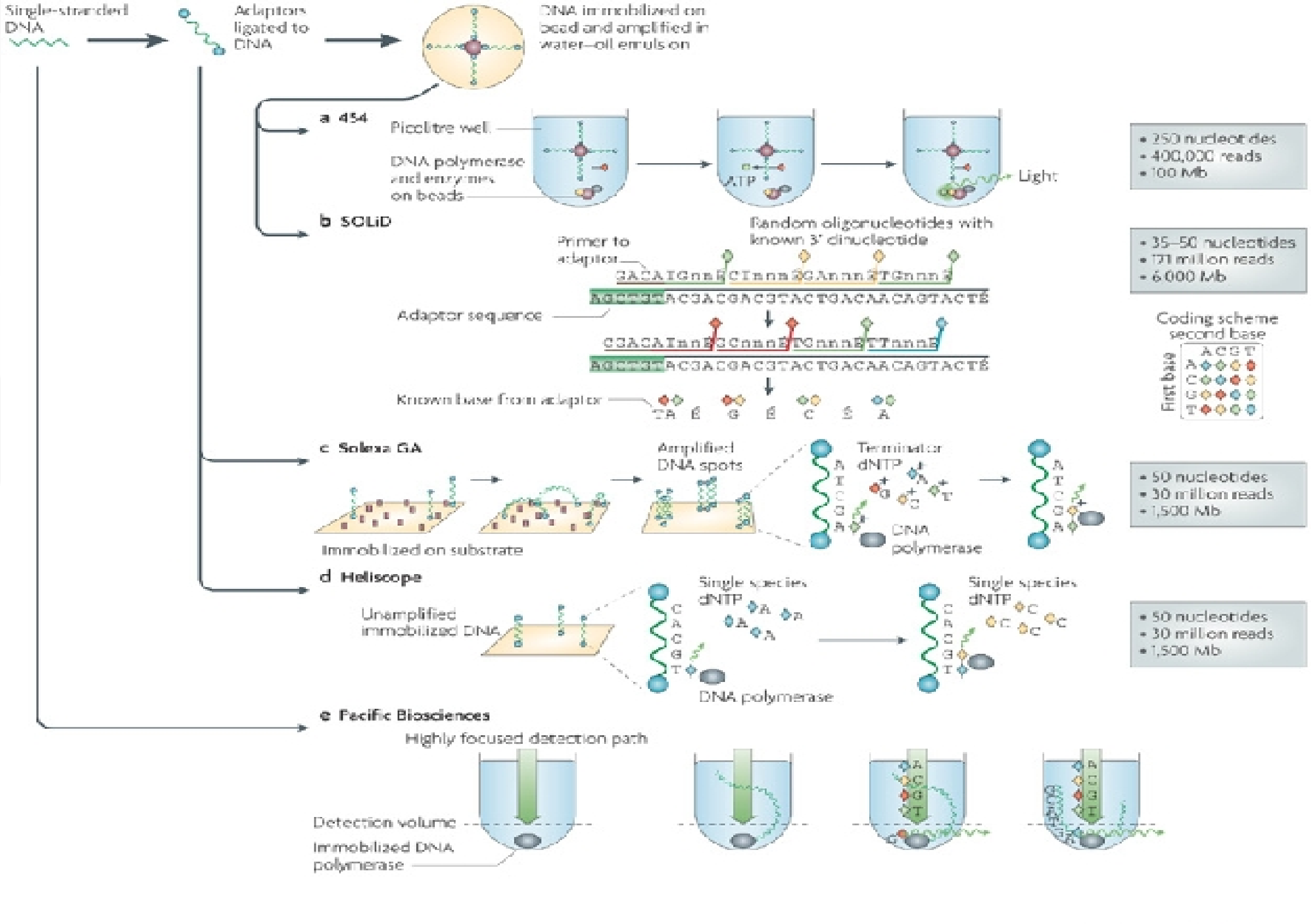
Step 4 Patient and control DNA compete to attach, or hybridize, to the microarray.

Step 5 The microarray scanner measures fluorescent signal intensity.

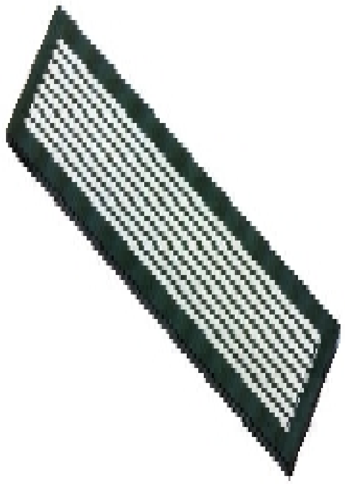
Step 6 Computer software gathers the data and generates a plot.

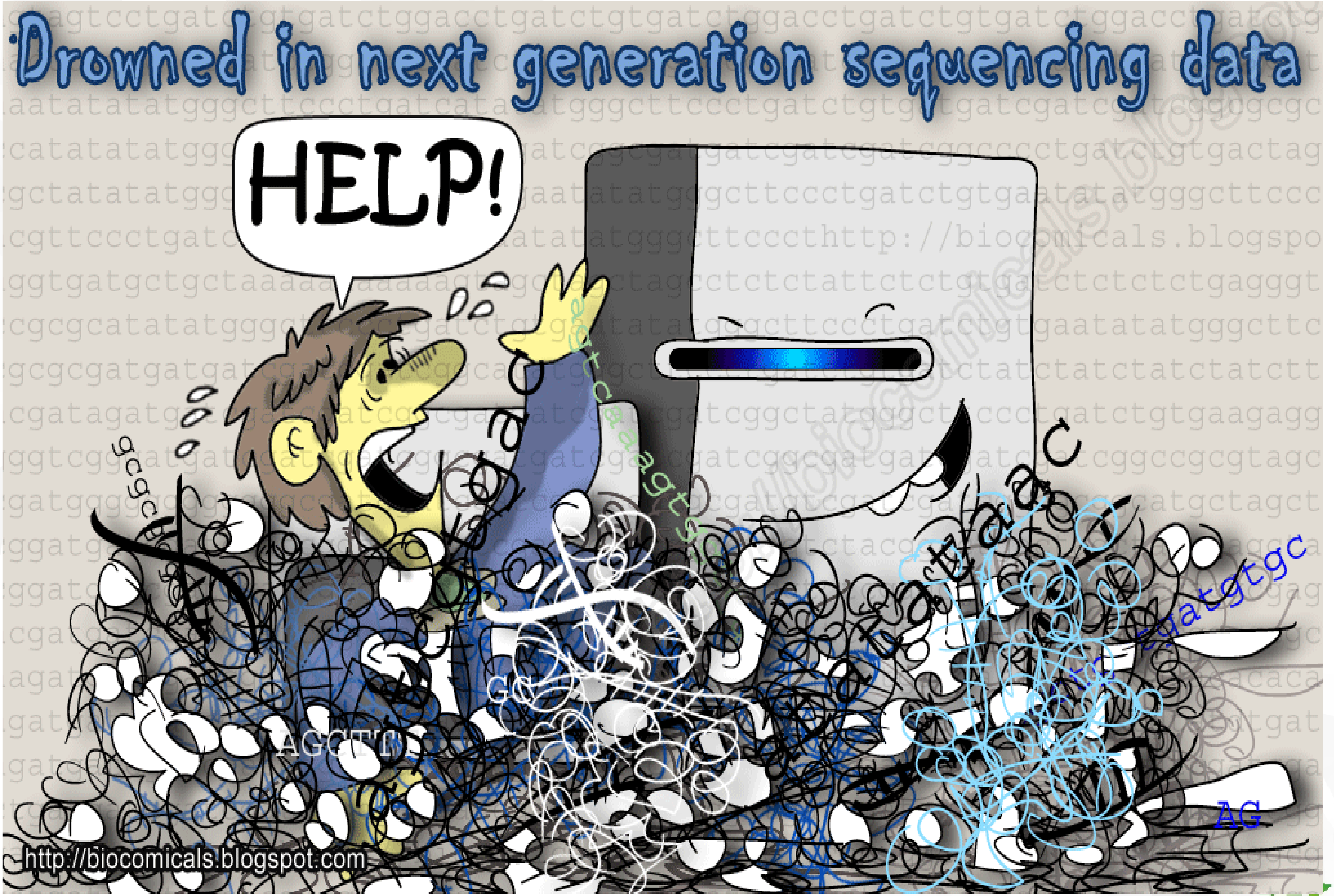
© David Kawei

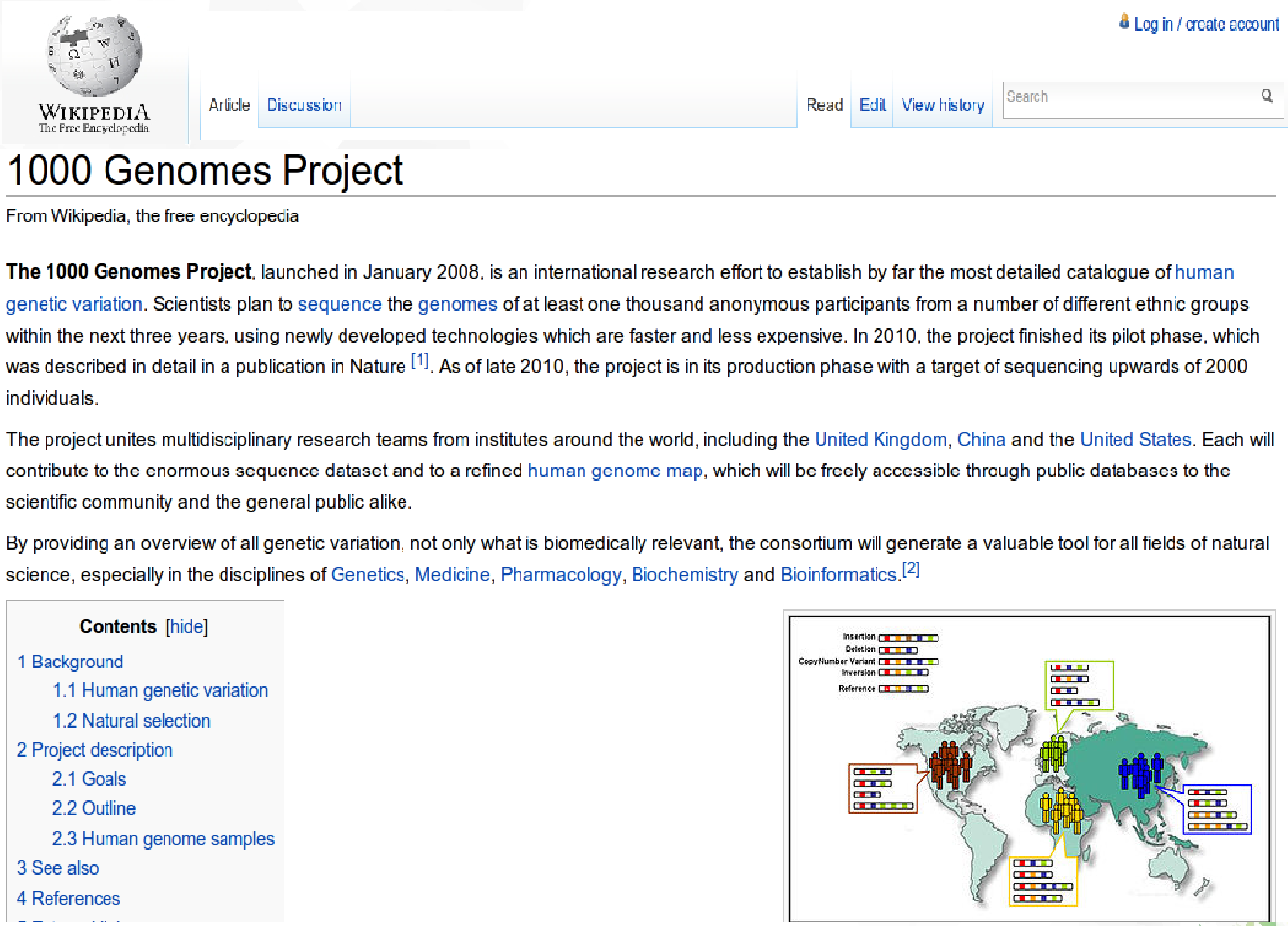
High throughput SV detection



Nature Reviews | Microbiology




Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com



1000 Genomes Project

From Wikipedia, the free encyclopedia

The 1000 Genomes Project, launched in January 2008, is an international research effort to establish by far the most detailed catalogue of human genetic variation. Scientists plan to sequence the genomes of at least one thousand anonymous participants from a number of different ethnic groups within the next three years, using newly developed technologies which are faster and less expensive. In 2010, the project finished its pilot phase, which was described in detail in a publication in Nature [1]. As of late 2010, the project is in its production phase with a target of sequencing upwards of 2000 individuals.

The project unites multidisciplinary research teams from institutes around the world, including the United Kingdom, China and the United States. Each will contribute to the enormous sequence dataset and to a refined human genome map, which will be freely accessible through public databases to the scientific community and the general public alike.

By providing an overview of all genetic variation, not only what is biomedically relevant, the consortium will generate a valuable tool for all fields of natural science, especially in the disciplines of Genetics, Medicine, Pharmacology, Biochemistry and Bioinformatics.[2]

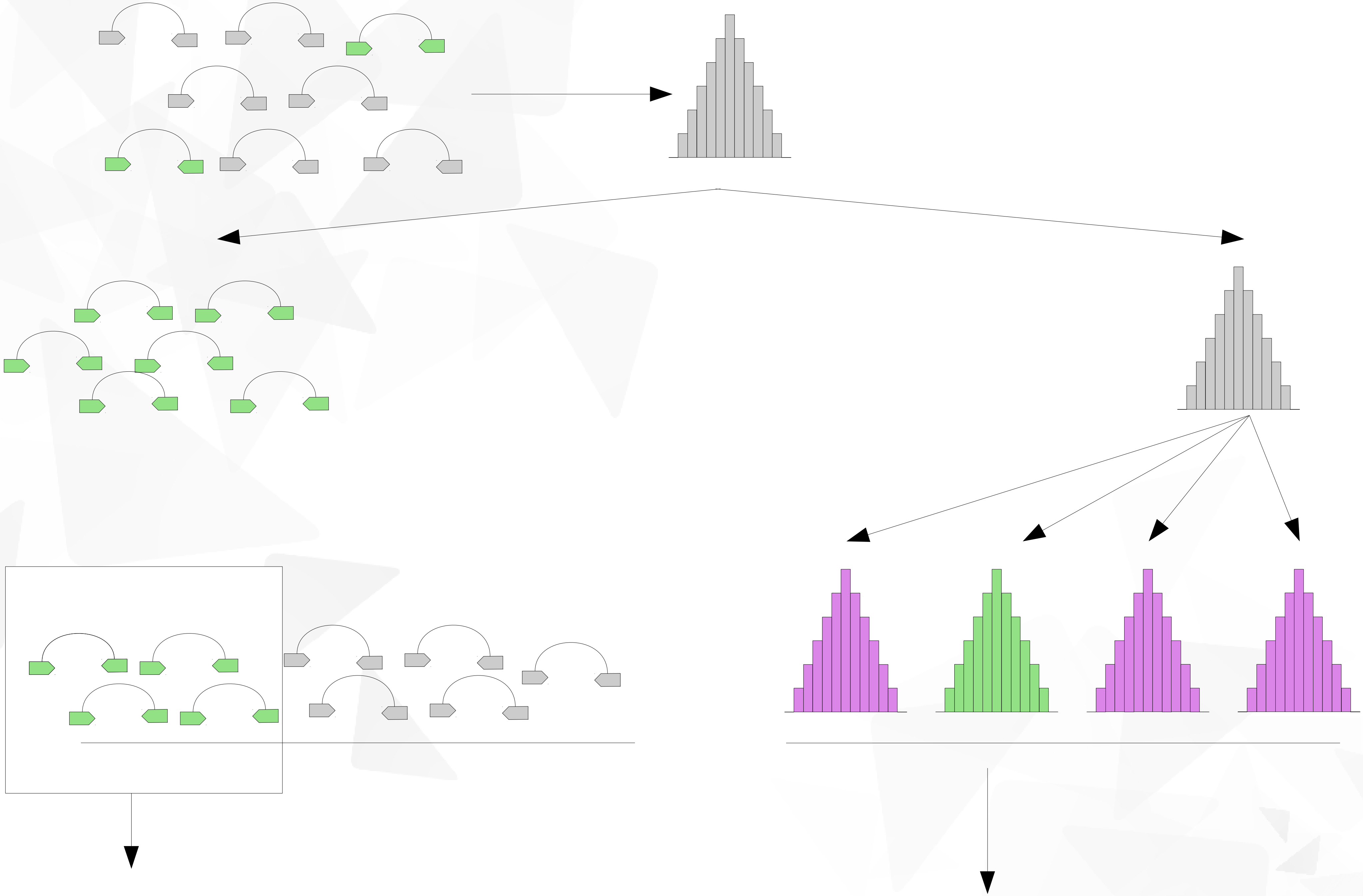
Contents [hide]

SV detection with NGS data (1): Alignment

- Whole Genome assembly (WGA)
  - High computational resources required;
  - Most assembler are “graph based”;
  - Alignment after assembly may not be trivial;
  - High resolution and precision for predictions
- Split read mapping
  - Problems in repetitive/low complexity regions;
  - Can't find large insertions;
  - Read length more an issue than coverage
  - PE/MP are better

SV detection with NGS (2): Statistics

- Read-depth:
  - Limited resolution;
  - Difficult to locate insertions/inversion;
  - Problems with highly repetitive regions;
  - Good for CNV (especially low copy)
- Insert size:
  - Usually not sensitive to small events;
  - Assumptions made on “insert-size” distribution;
  - How do you identify aberrant pairs?

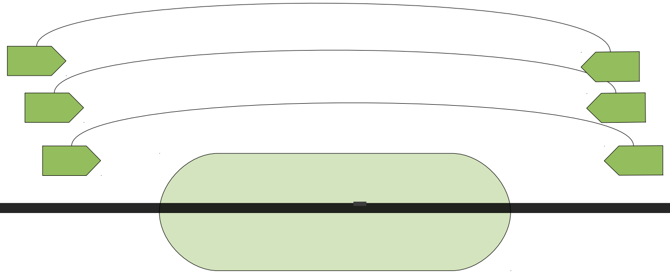


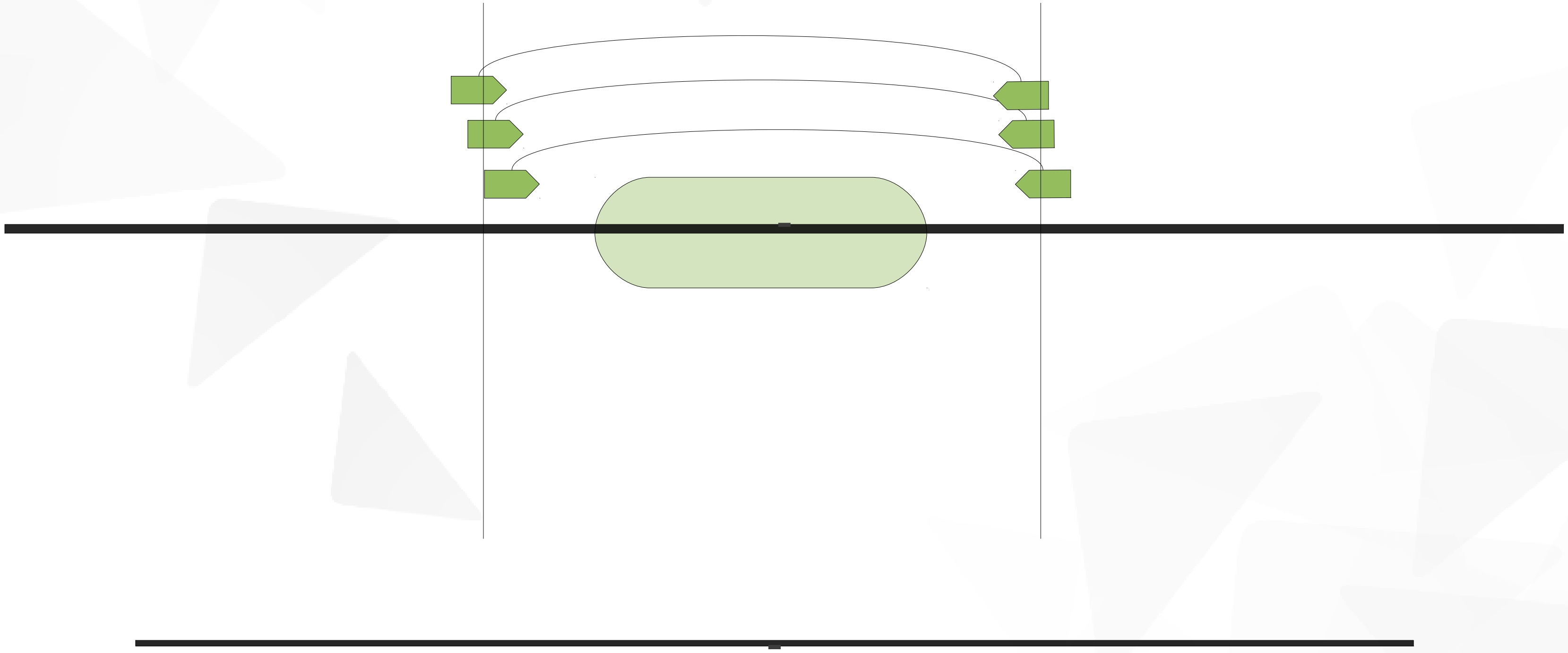
INDEL

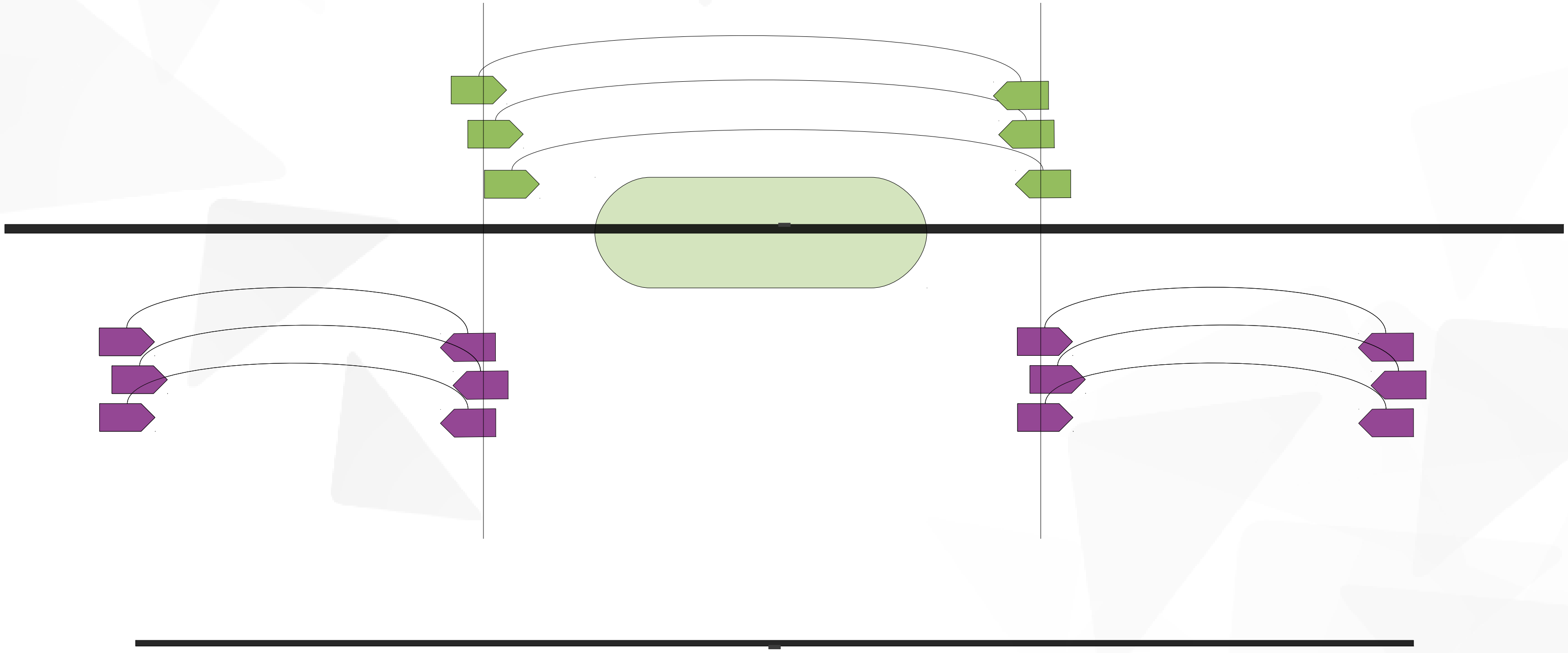
INDEL

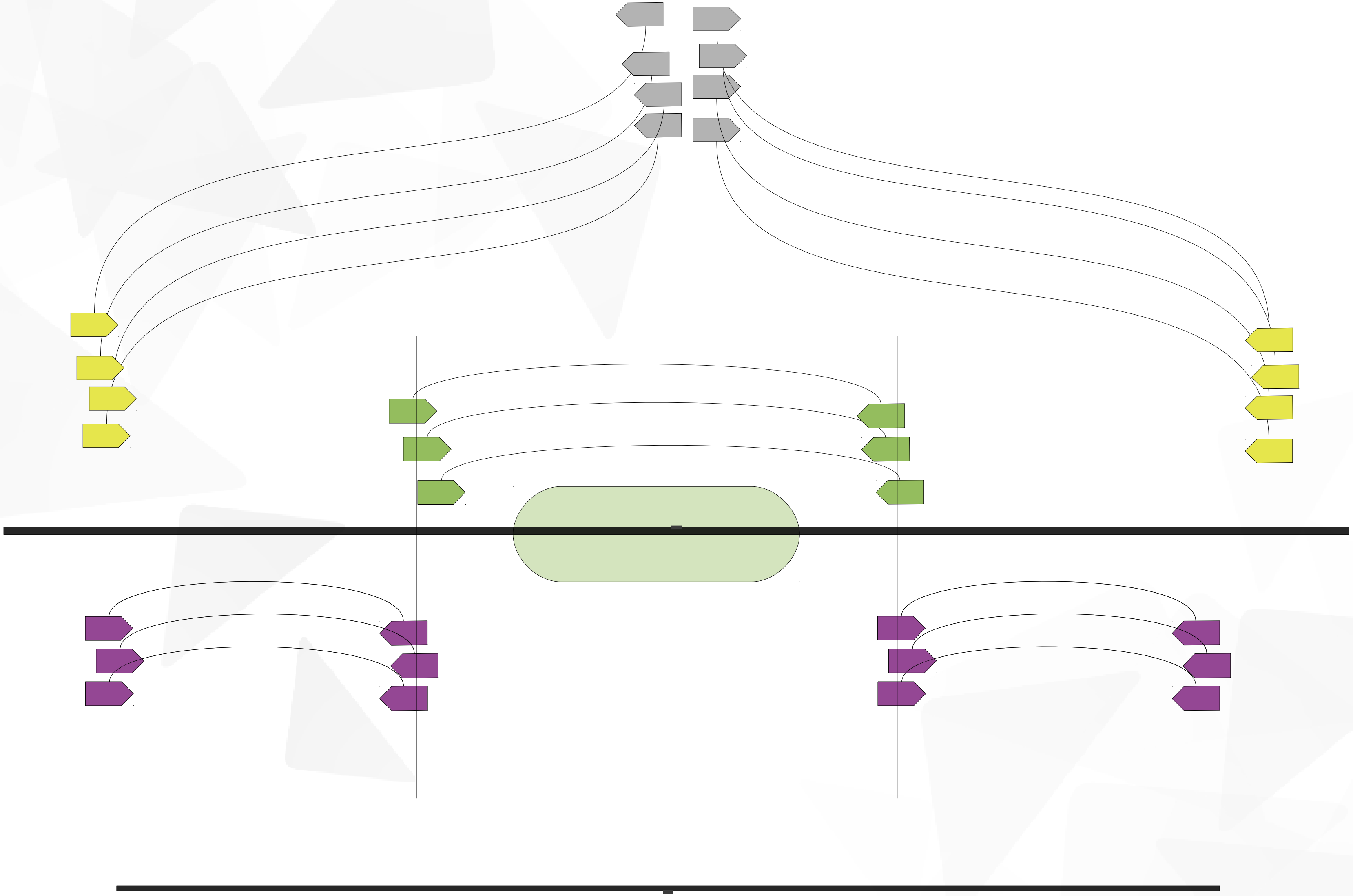
Improving “insert-size” methods

- Insert size distribution alone is unlikely to be sensitive to small events (depending on the mean and variance of global insert size)
- Context specific variations in insert size for particular genomic regions?
- Can we incorporate other types of information into insert-size analysis?
- And which ones?

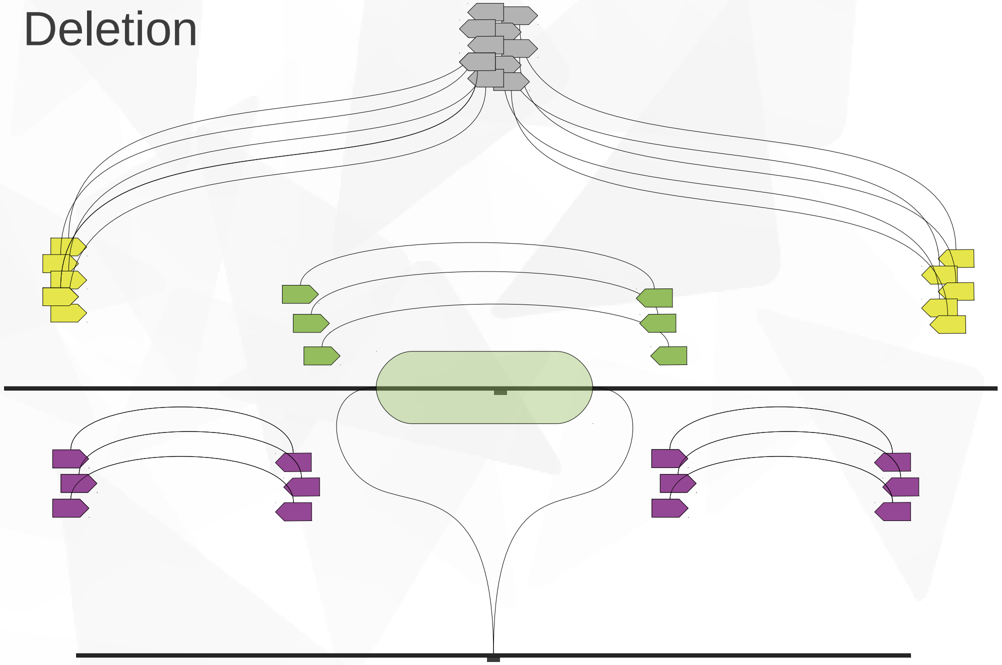




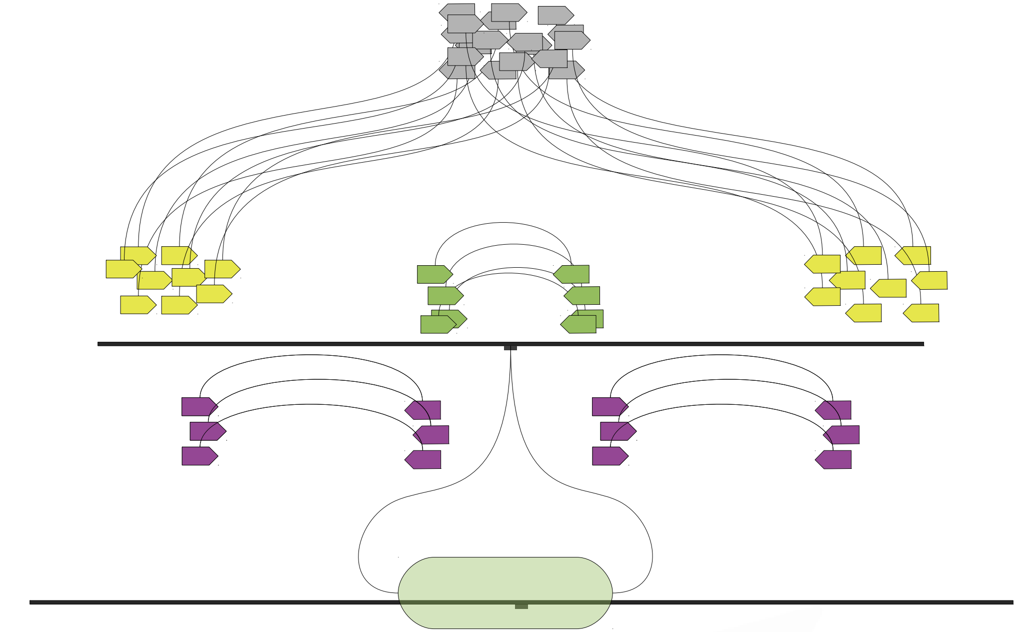




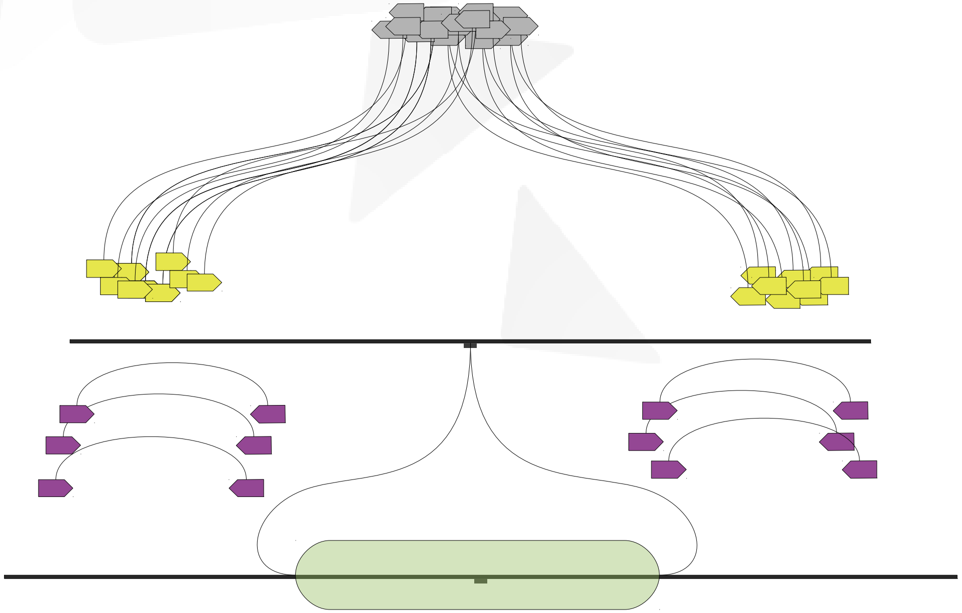
Deletion



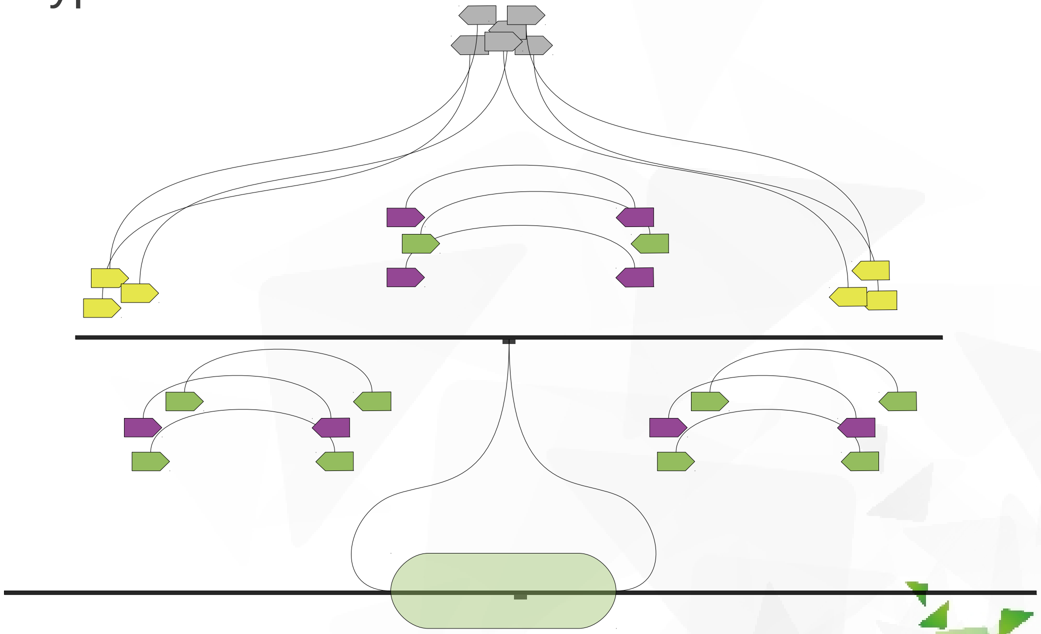
Short insertion



Long insertion



Hyper-variable

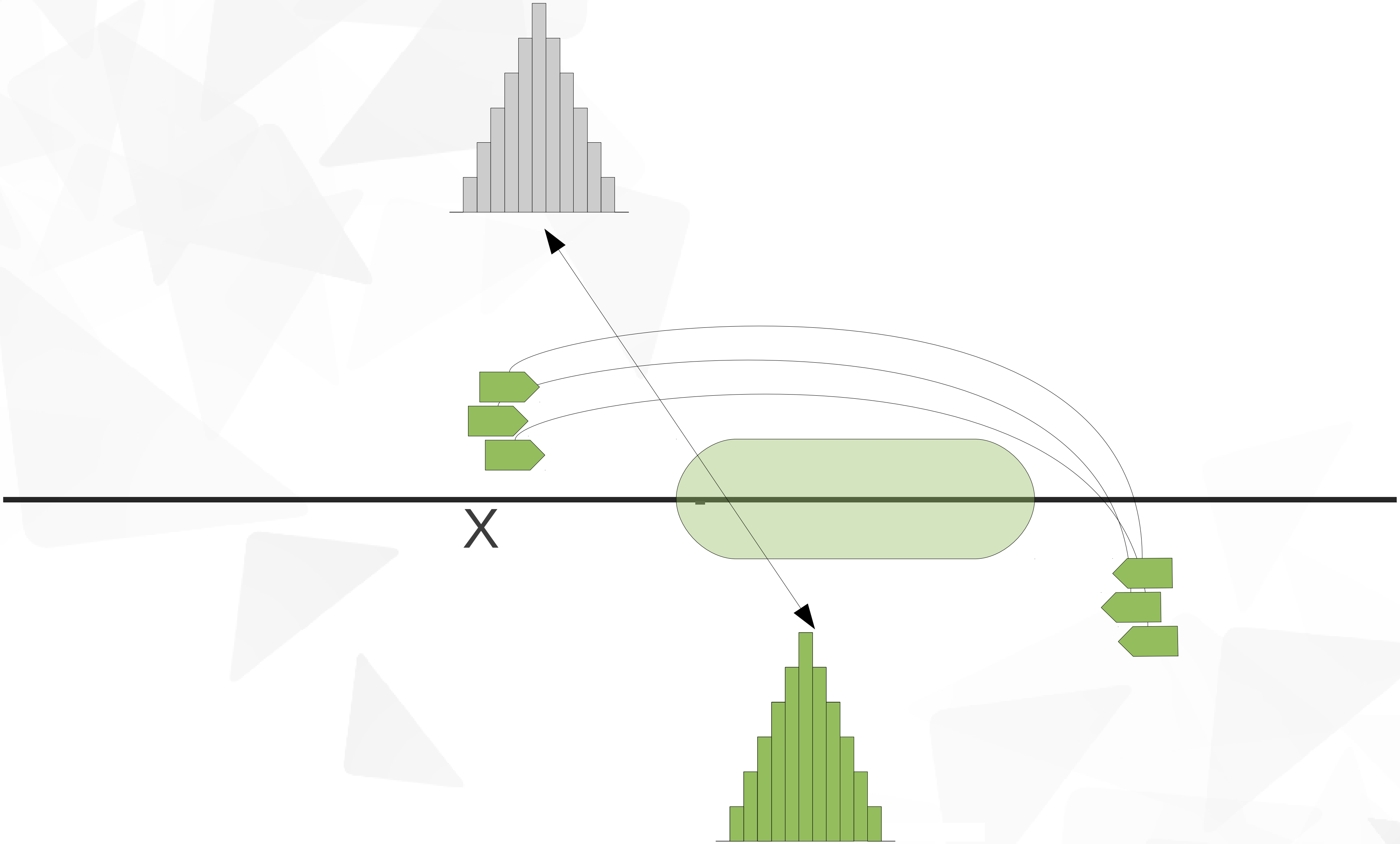


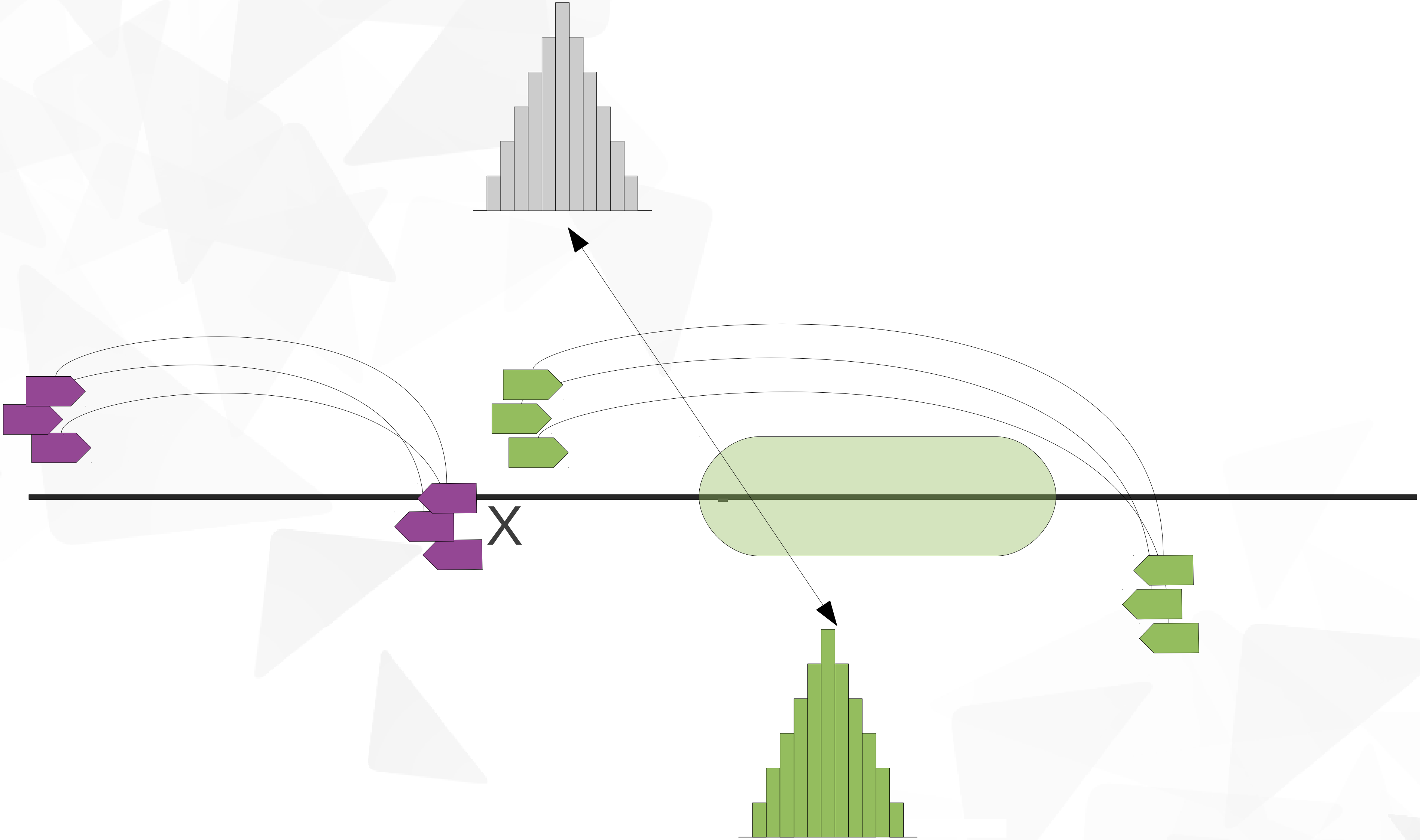
Can we improve indel detection by integrating more data?

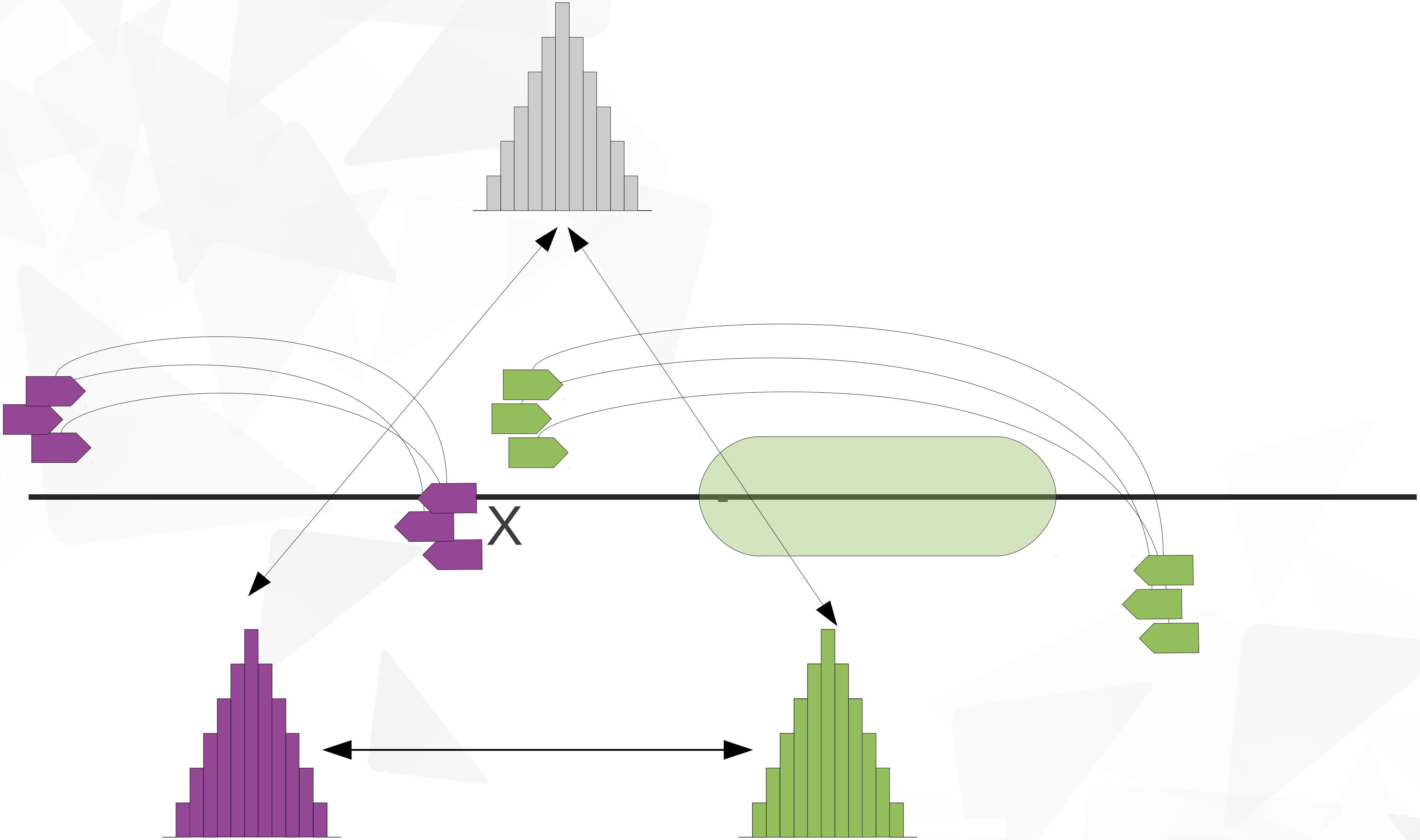
- size distributions alone may not find small events
- presence and position “broken” pairs may be informative
- take advantage of the asymmetric nature of paired reads

A novel strategy for SV detection

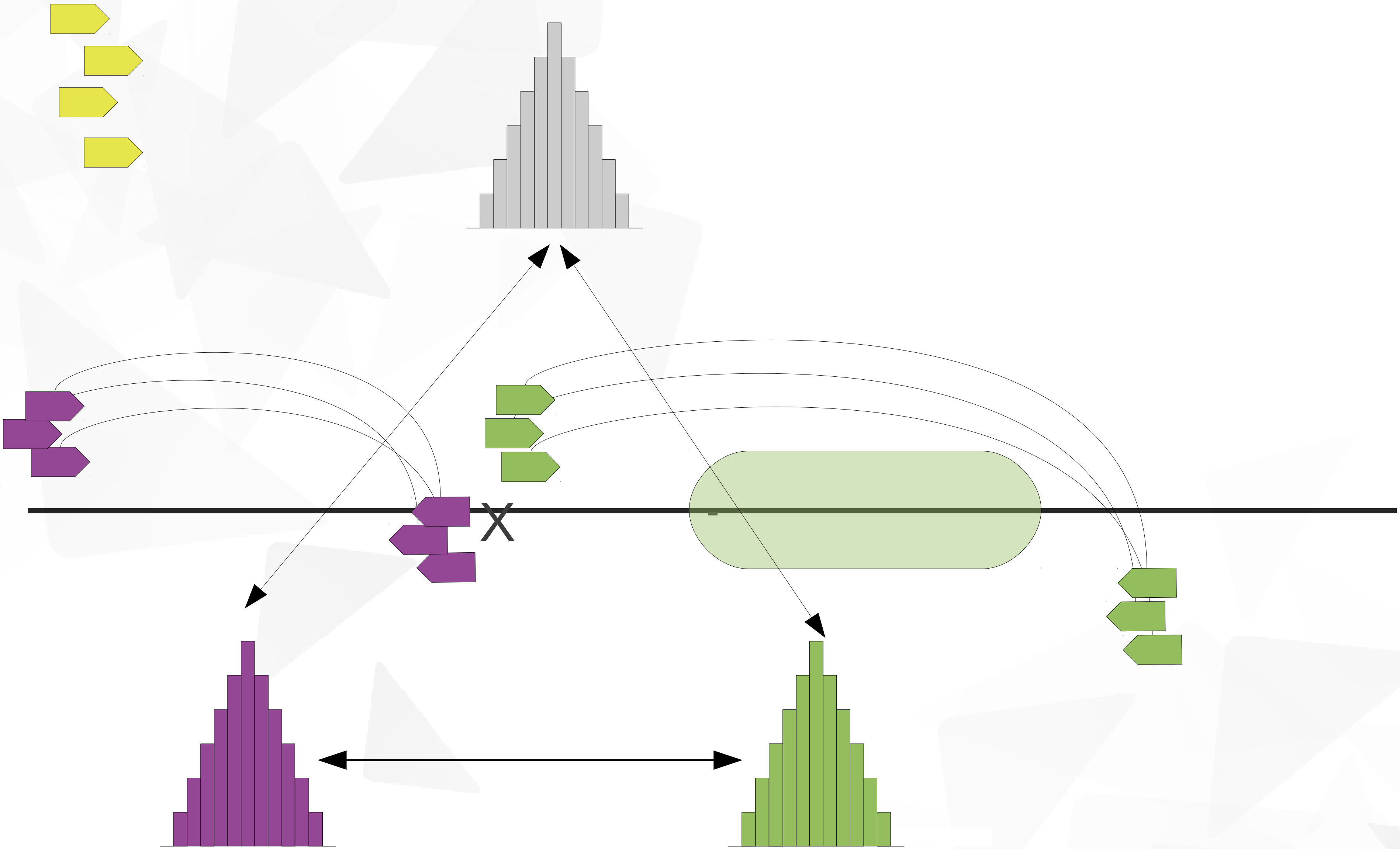
- based on supervised learning (Support Vector Machines)
- Multi-class SVM from libsvm (www.csie.ntu.edu.tw/~cjlin/libsvm/)
- 5 distinct categories:
    - Long insertion (longer than insert size), Short insertion, Deletion, No event, Variable region

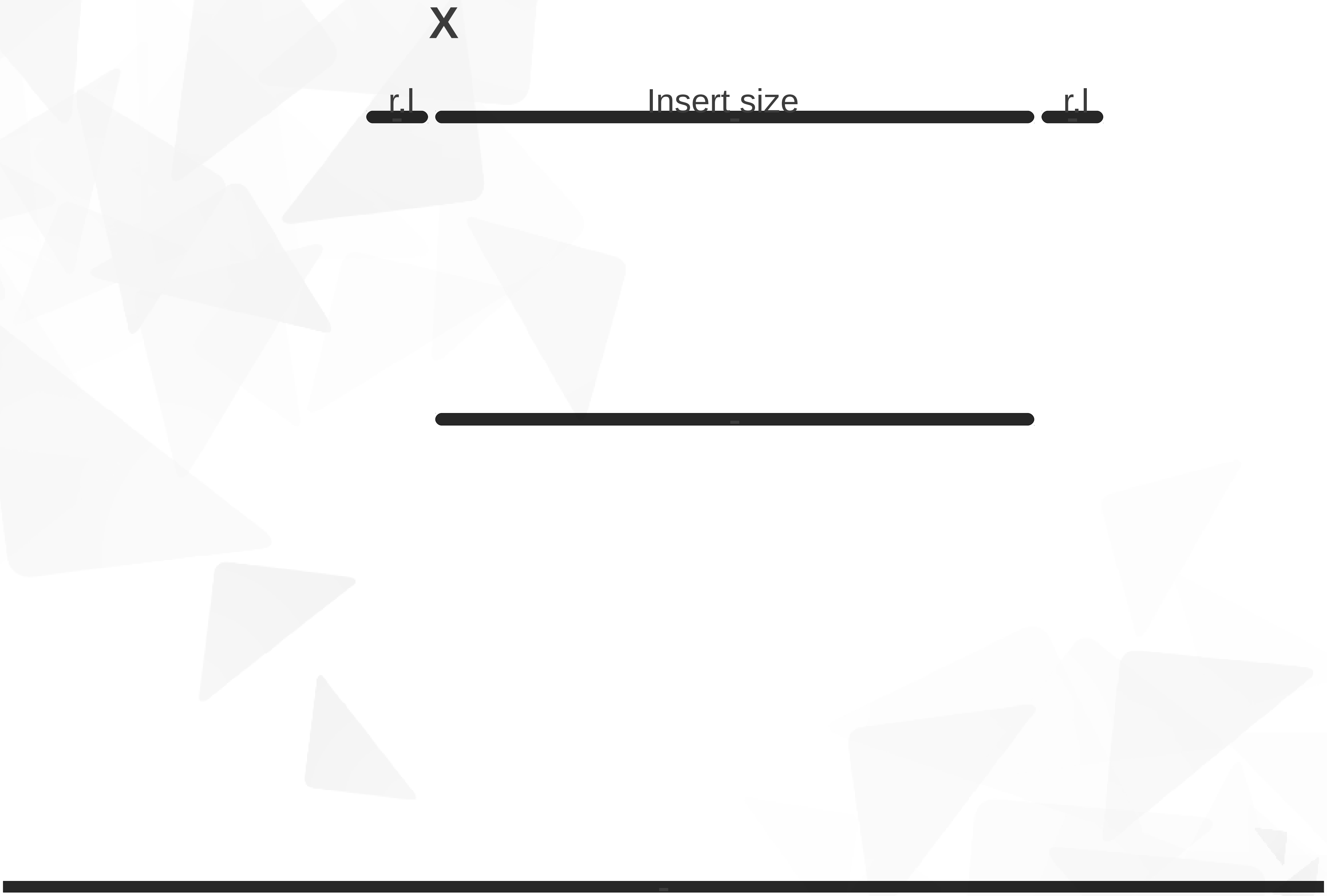

X


X



X


X



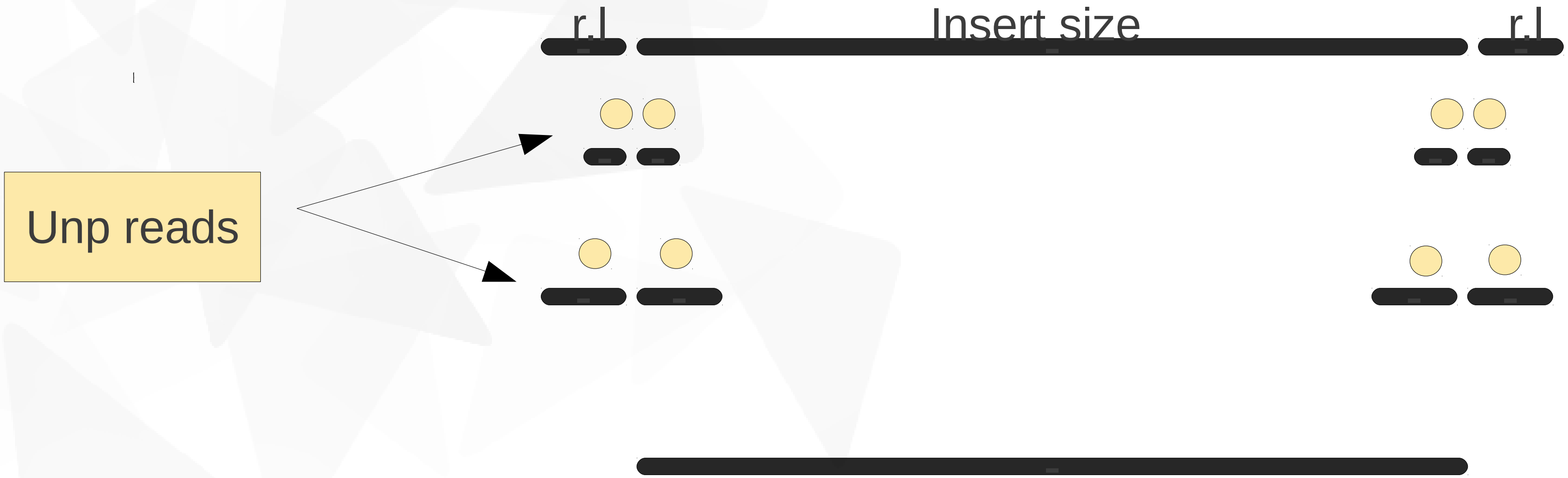
X

r.l

Insert size

r.l

X

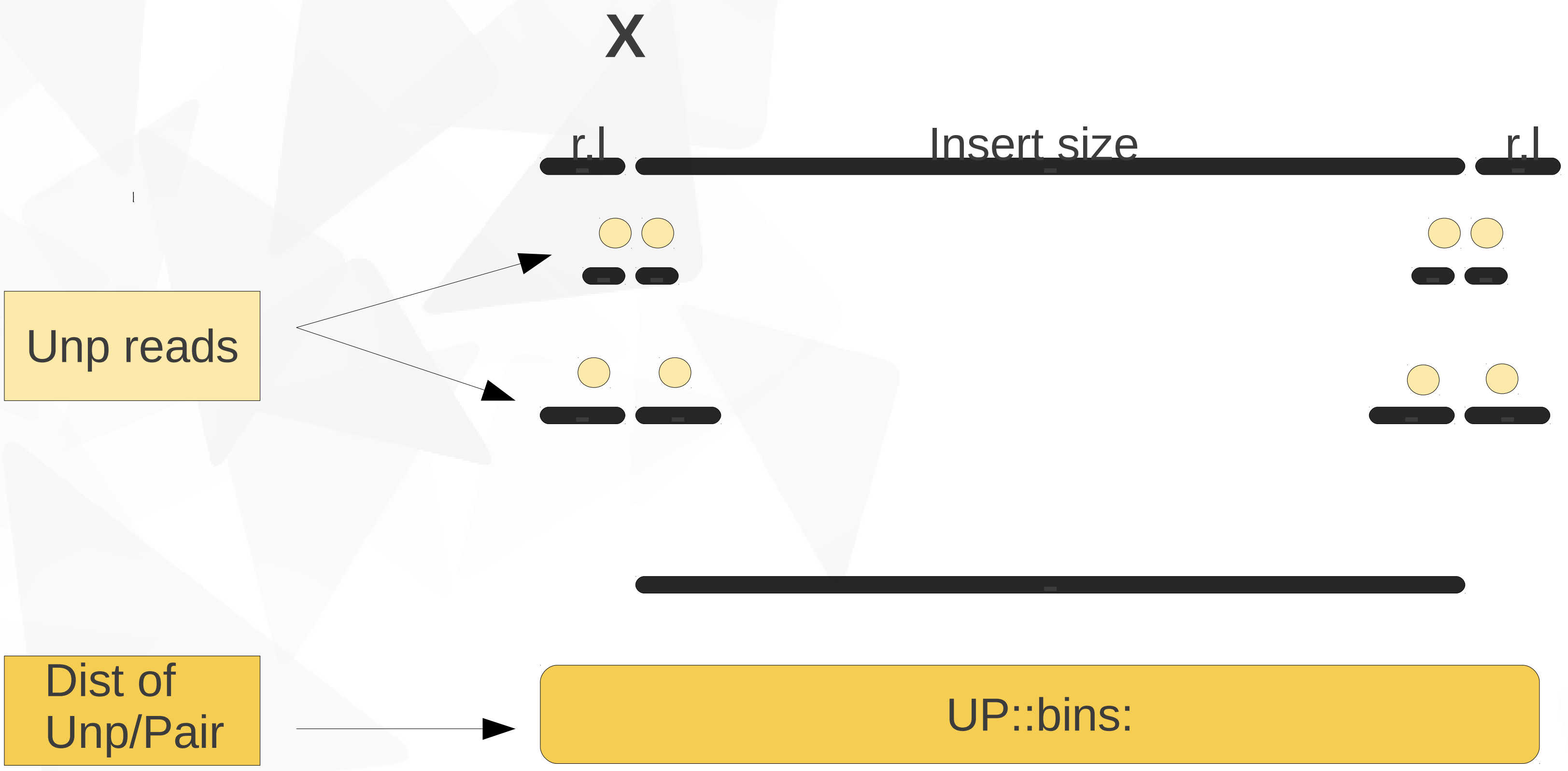


r.l

Insert size

r.l

Unp reads



X

r.l

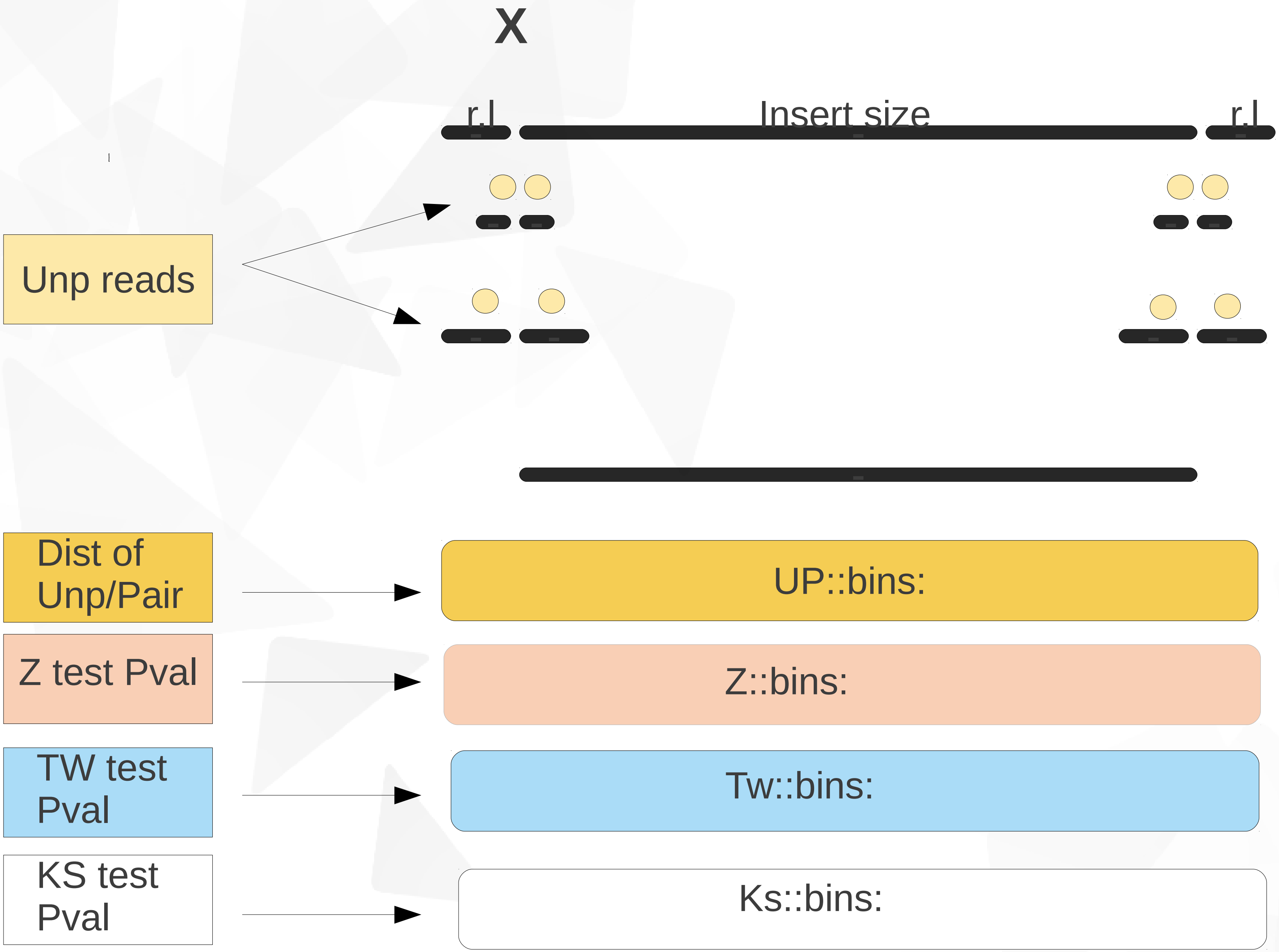
Insert size

r.l

Unp reads

Dist of Unp/Pair

UP::bins:



X

r.l

Insert size

r.l

Unp reads

Dist of Unp/Pair

UP::bins:

Z test Pval

Z::bins:

TW test Pval

Tw::bins:

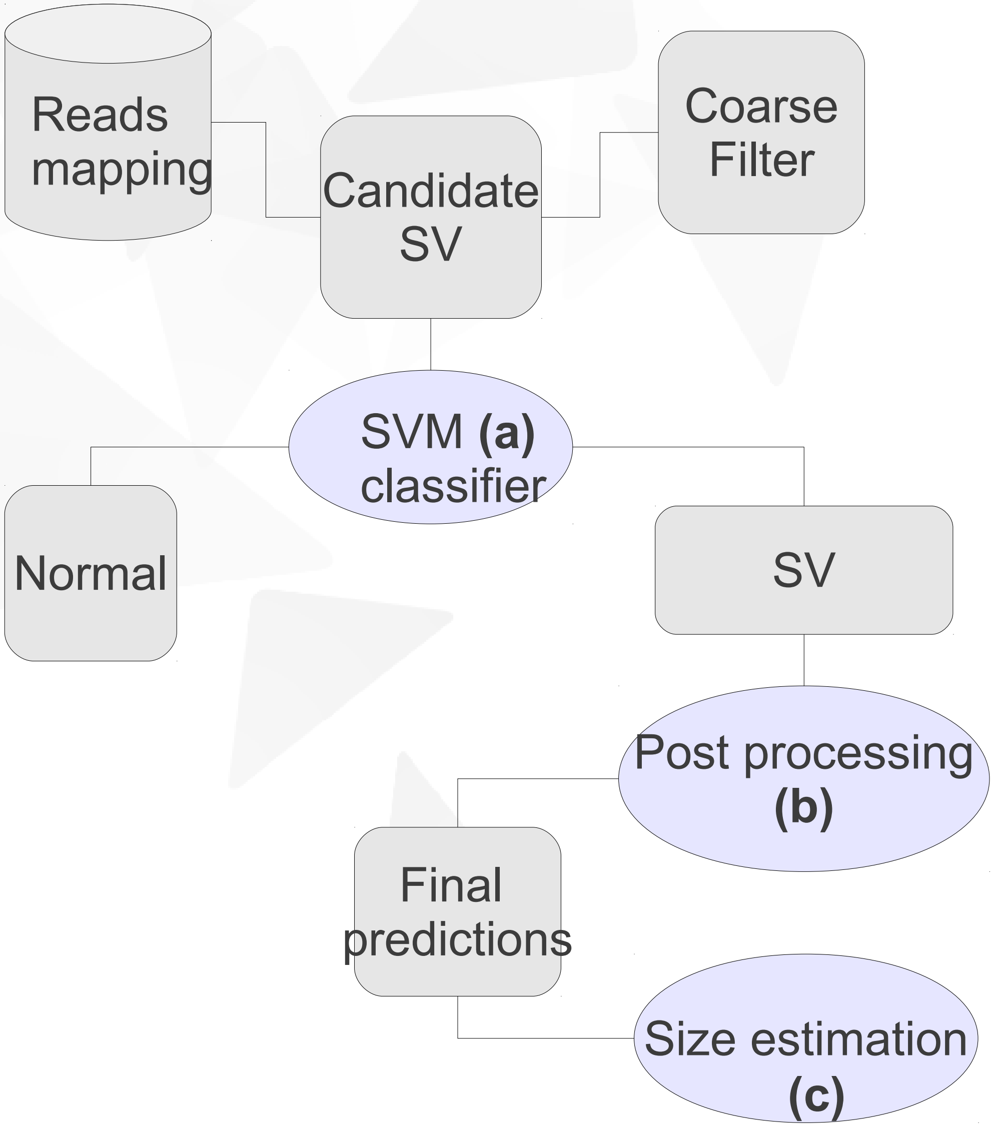
KS test Pval

Ks::bins:

Training

- Choose regions where donor and reference genomes show good coverage and no evidence of anomalous insert size or peaks of broken pairs
- Introduce events *in-silico*
- *Remap reads*
- *Calculate features*
- *Use known positions of indels to generate positive and negative training sets*

The SVM2 algorithm



Reads mapping

Candidate SV

Coarse Filter

SVM **(a)** classifier

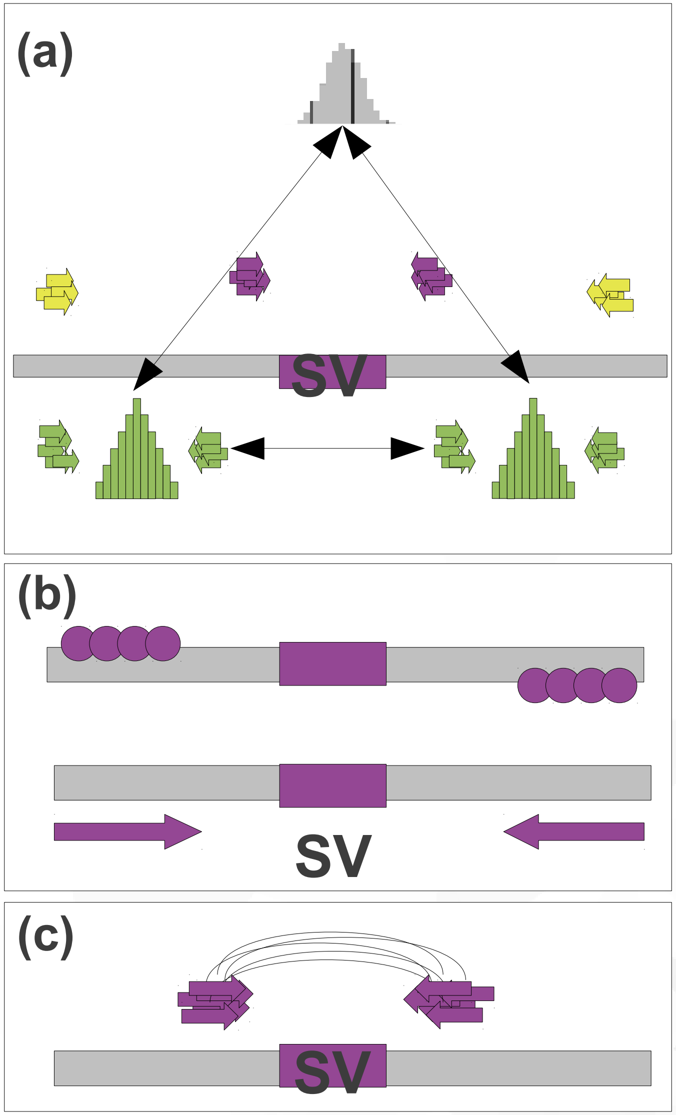
Normal

SV

Post processing **(b)**

Final predictions

Size estimation **(c)**



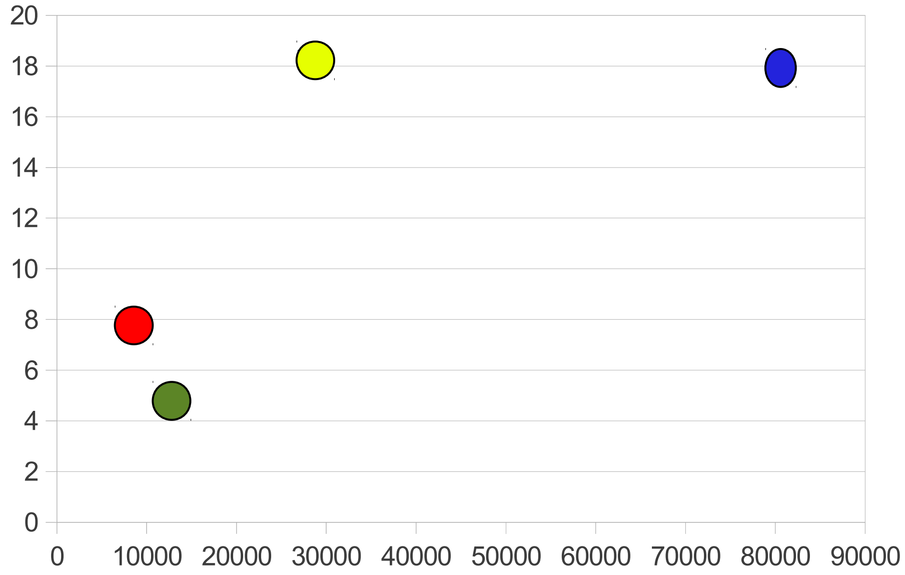
Evaluation with real human genome data

- Human genome, African male sequenced both with:
  - Illumina (=~40 X, PE, I.S 208 bp):
    - Bentley et al. Nature. 2008 456: 53-9
  - Sanger (=~0.3 X, fosmids, I.S. 40 Kb)
    - Kidd et al. 2008 Nature 453:56-64
    - SV of 1 to 100 bp w.r.t the reference human genome.
    - 116170 deletions
    - 107719 insertions
    - released clone mapping

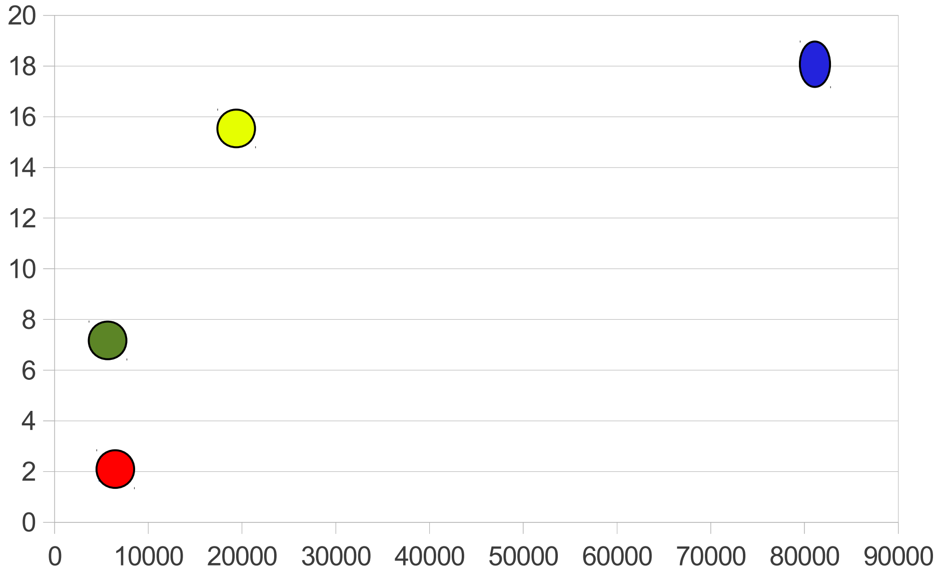
Comparison with similar tools

	Deletions	Valid by Kidd	Insertions	Valid by Kidd
🟢 **Modil**	13147	622 (5%)	3981	282(7%)
🔴 **Variation Hunter**	8537	703(8%)	7142	100(1.5%)
🟡 **Break Dancer**	27092	4970(18%)	19305	2983(15%)
🔵 **SVM**	80520	14387(18%)	81121	14870(18%)

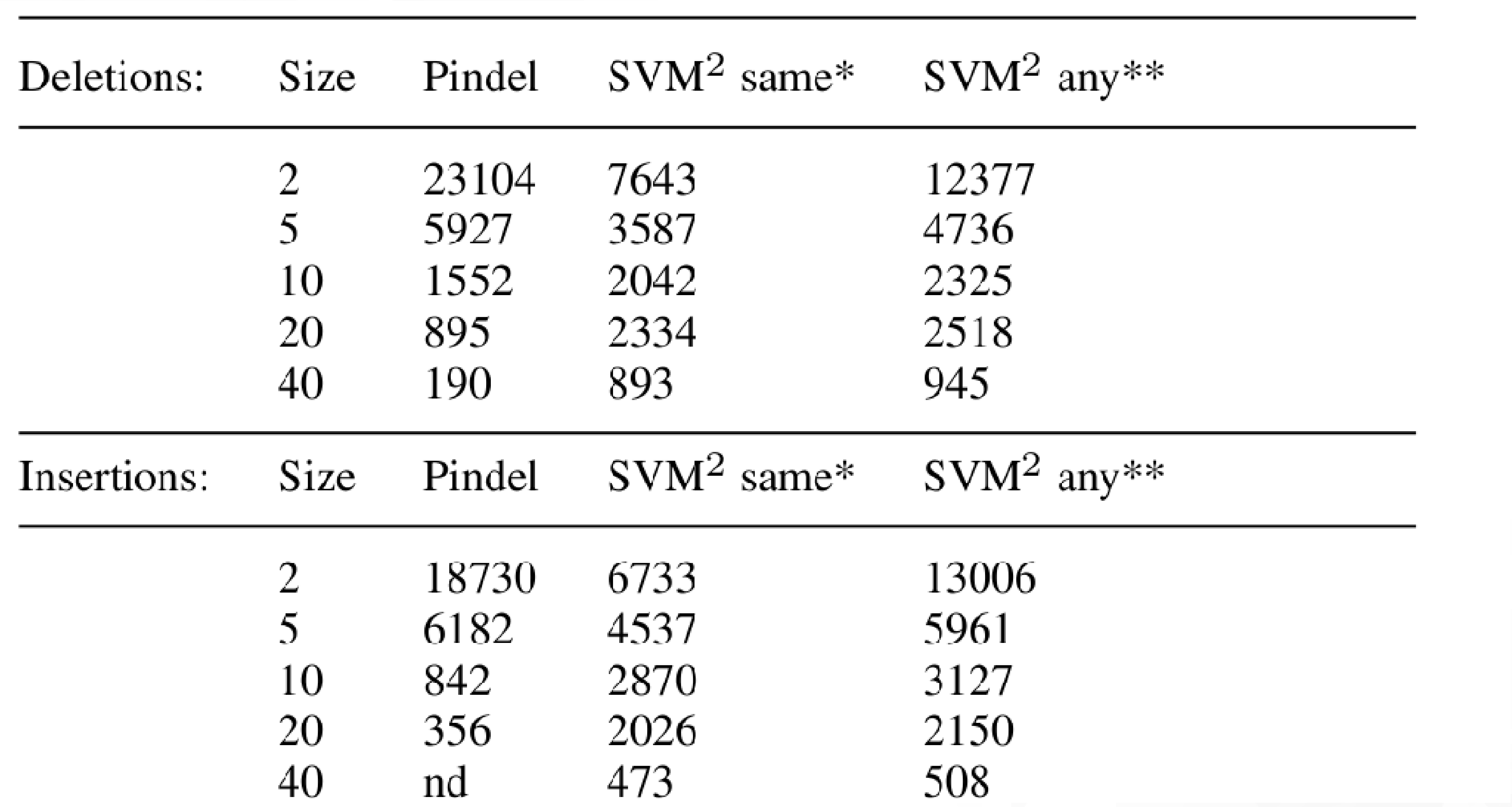
Deletions



Insertions



Comparison with split mapping



Deletions:	Size	Pindel	SVM^2 same*	SVM^2 any**
	2	23104	7643	12377
	5	5927	3587	4736
	10	1552	2042	2325
	20	895	2334	2518
	40	190	893	945

Insertions:	Size	Pindel	SVM^2 same*	SVM^2 any**
	2	18730	6733	13006
	5	6182	4537	5961
	10	842	2870	3127
	20	356	2026	2150
	40	nd	473	508

Conclusions

- Same specificity as Breakdancer but 3X to 4X more sensitivity;
- 23.4% “validation rate” (maximum possible 30%) if consider only positional validation;
- Perform better than split mapping at >= 5 bp;
- Customization (trained on your data);
- Not based on a single metric;
- More robust than Pindel in repetitive/low complexity regions;

Perspectives/problems

- Heterozygosity
  - In principle we can use expectation maximization algorithms to find loci where insert sizes can best be modeled by two distributions
  - In practice we have little genome-wide information on heterozygous SV for evaluation
- Mate Pair libraries
  - Refers to the construction of libraries with large “inserts” by circularization step.
  - For now, such libraries tend to have large insert size ranges!
- Combining split mapping and sophisticated insert size methods into a single tool

Acknowledgments

Graziano Pesole

Giulio Pavesi

Federico Zambelli

Massimiliano Borsani

Gian Marco Prazzoli

David S. Horner

Matteo Chiara

Carmela Gissi

Francesca Griggio

BEaCON

Bioinformatics Evolution and COmparative GeNomics