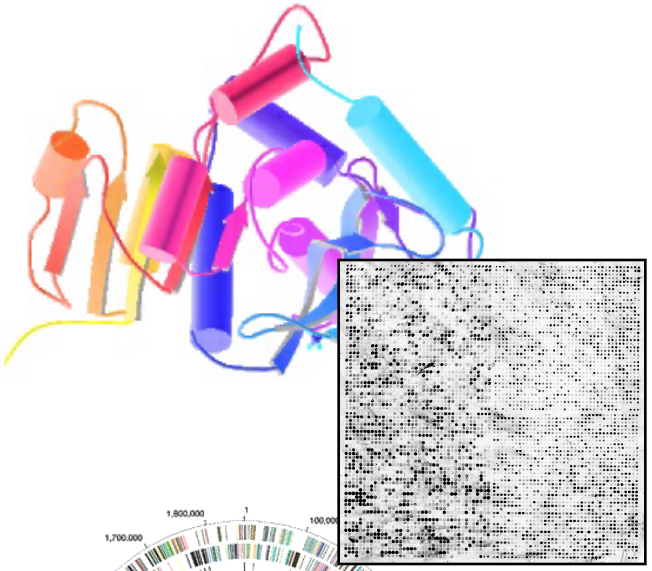
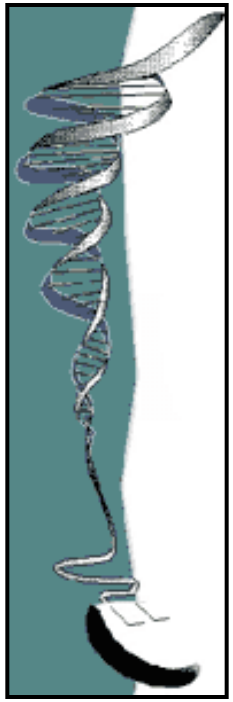
Bioinformatics:

Variant Identification, Focusing on SVs





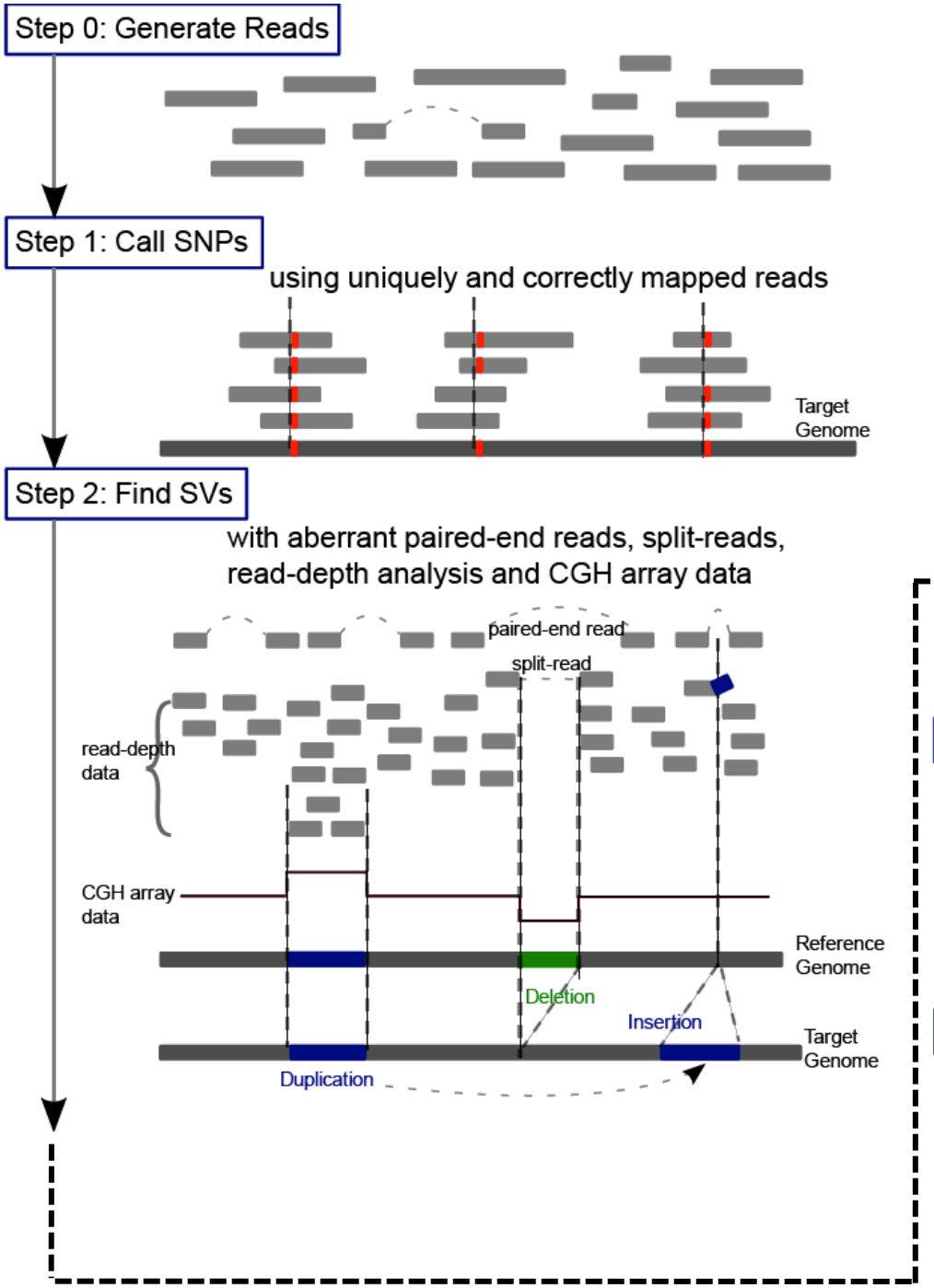
Mark Gerstein, Yale University

gersteinlab.org/courses/452

(last edit in spring '17)

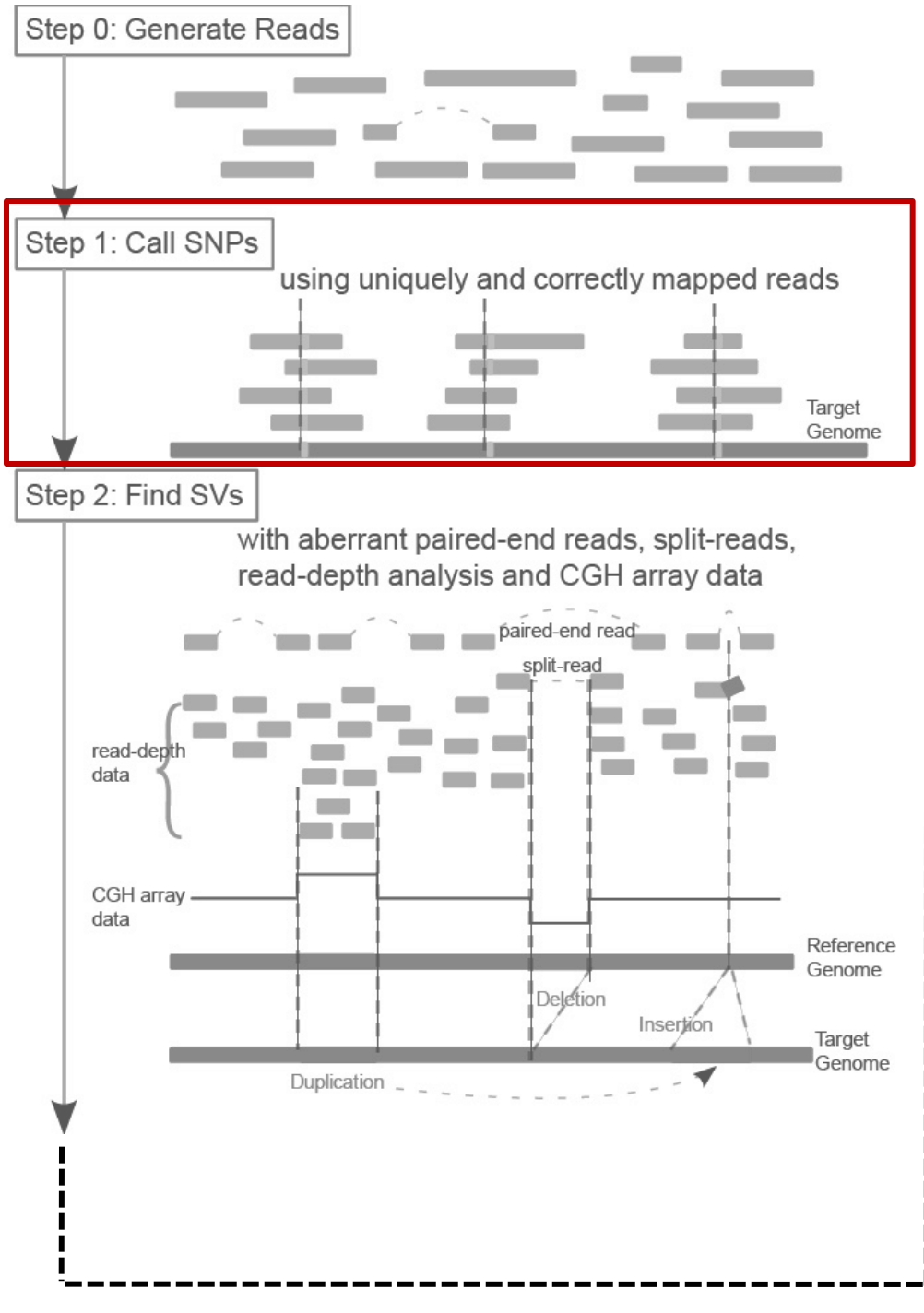
Main Steps in Genome Resequencing

[Snyder et al. Genes & Dev. ('10)]



Main Steps in Genome Resequencing

[Snyder et al. Genes & Dev. ('10)]



Bayes' Theorem to detect genomic variant

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)}$$

$$= \frac{P(D|G)\,P(G)}{\sum_{i=1}^{n} P(D|G_i)\,P(G_i)}$$

In the above equation:

- D refers to the observed data
- G is the genotype whose probability is being calculated
- G_i refers to the ith possible genotype, out of n possibilities

Calculating the conditional distribution $P(D|G)$:

Assuming an error free model, for each heterozygous SNP site of the diploid genome, covered by K reads, the number of reads i representing one of the two alleles follows binomial distribution.

$$\boldsymbol{P_{err\_free}(D|G) = f(i|k, 0.5) = \binom{k}{i} 0.5^k}$$

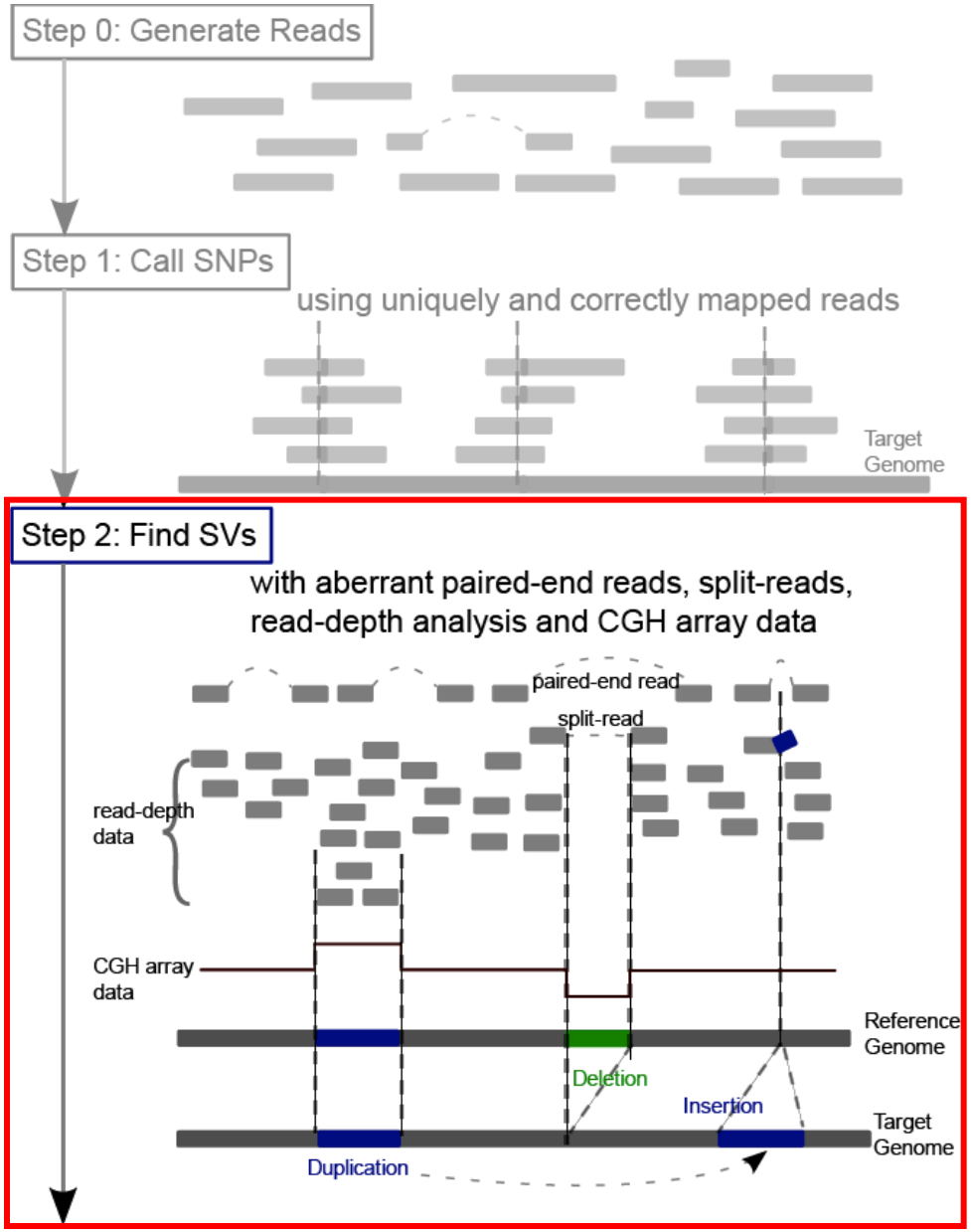
With errors, the calculation is more complicated.

In general:

$$\boldsymbol{P(D|G) = P_{err\_free}(D|G) + P_{err}(D|G)}$$

Main Steps in Genome Resequencing

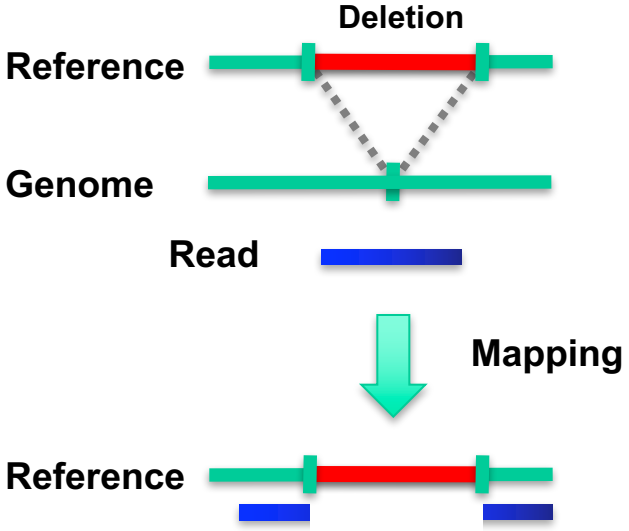
[Snyder et al. Genes & Dev. ('10)]



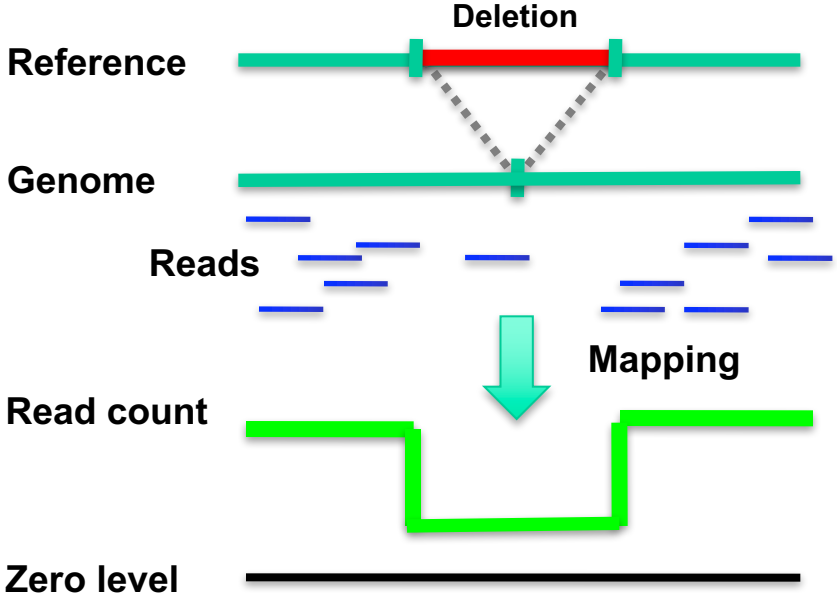
Methods to **Find SVs**

1. Paired ends

2. Split read



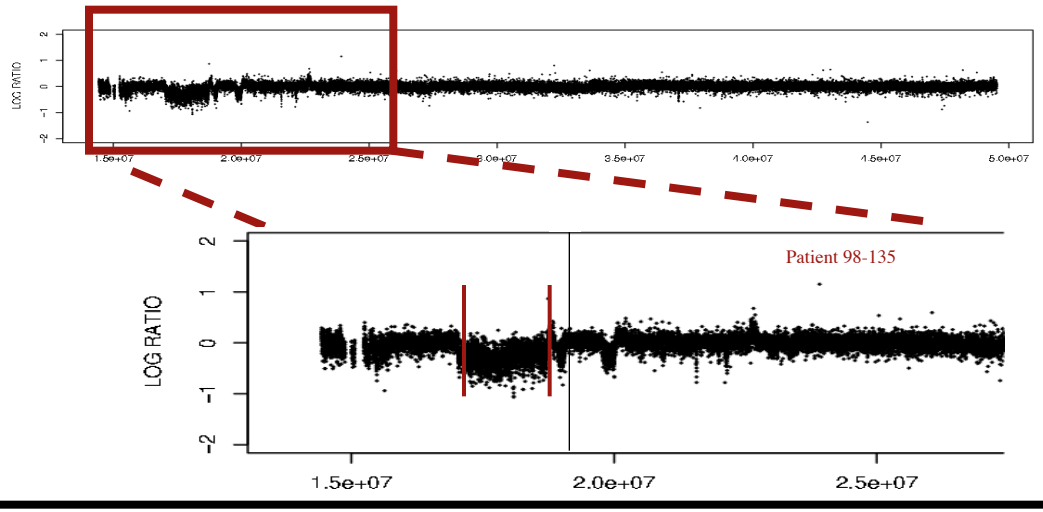
3. Read depth (or aCGH)



4. Local Reassembly

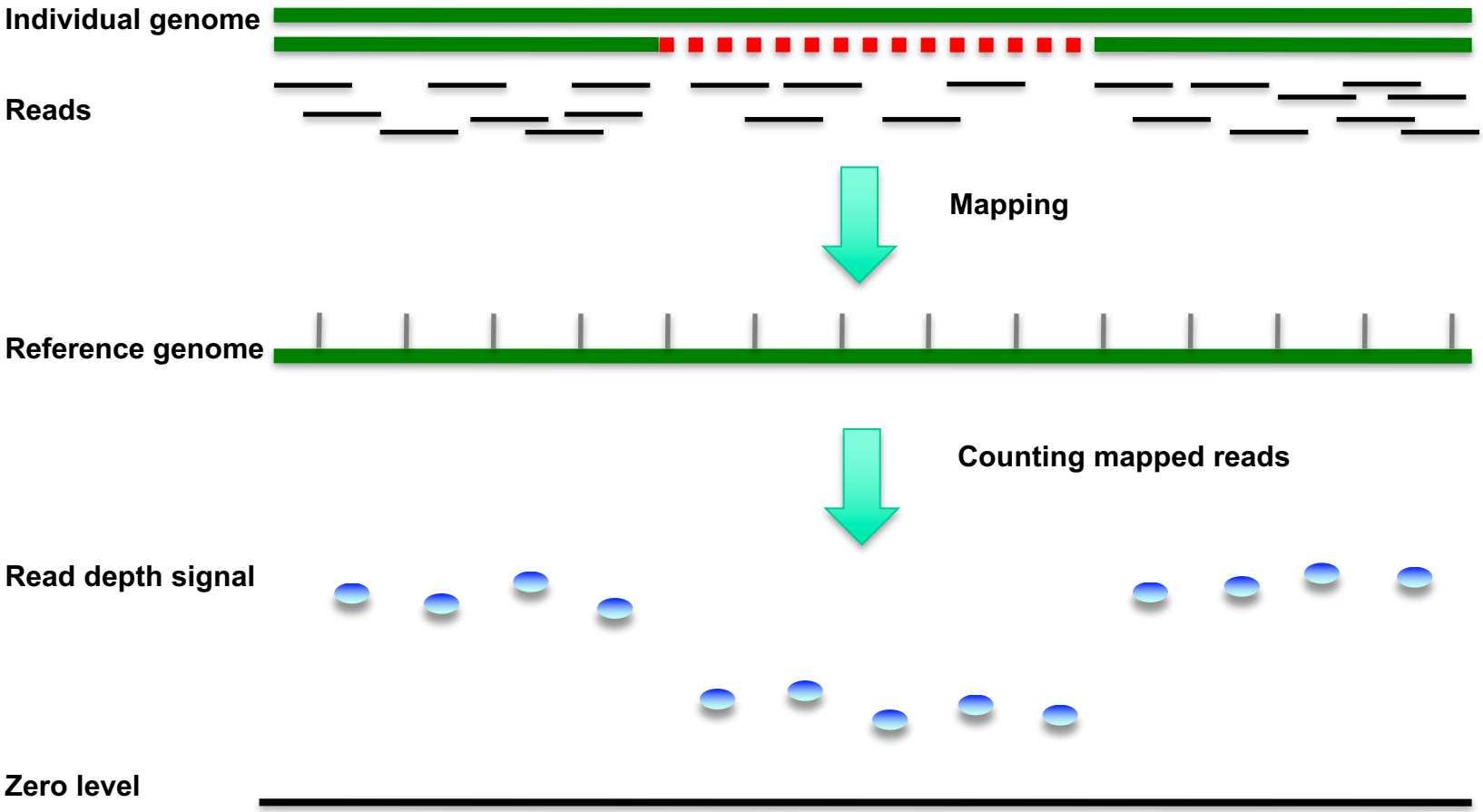
[Snyder et al. Genes & Dev. ('10)]

Read Depth



Array Signal

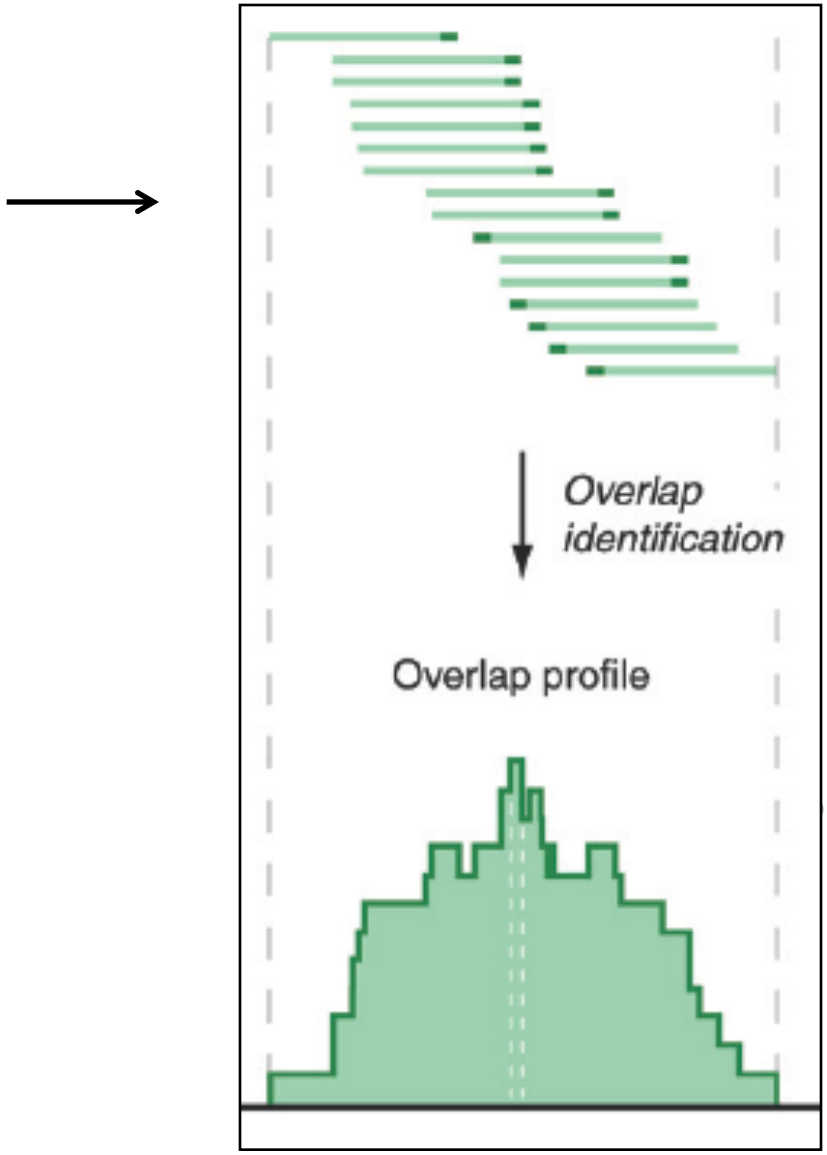
Read depth



[Urban et al. ('06) PNAS; Wang et al. Gen. Res. ('09); Abyzov et al. Gen. Res. ('11)]

Reads to Signal Track

```
@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYhhhhhhYIhhaZT[hYHNSPKXR
@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
+ILMN-GA001_3_208HWAAXX_1_1_111_879
hSWhRNJ\hFhLdhVOhAIB@NFKD@PAB?N?
```



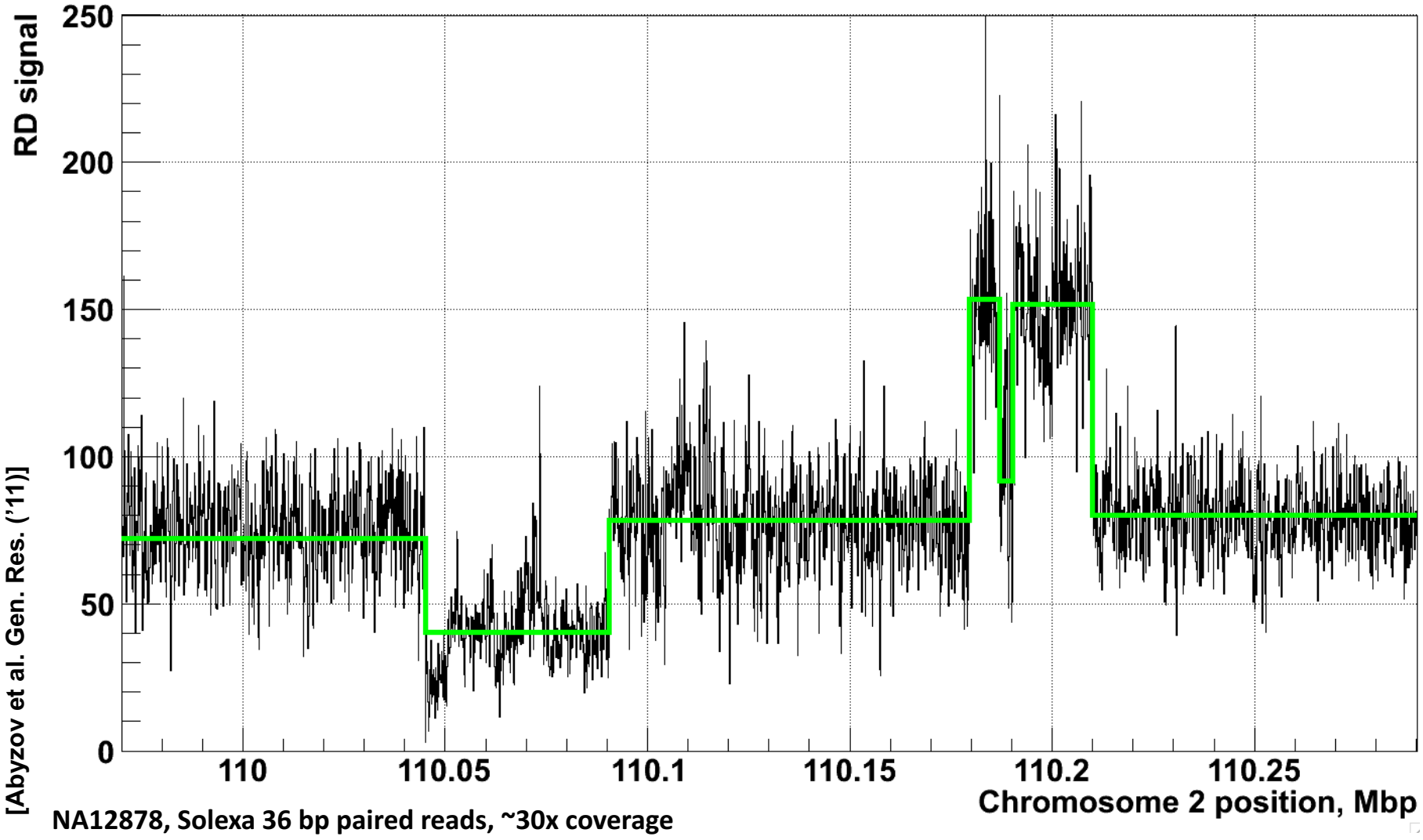
Reads (fasta)
+ quality scores (fastq)
+ mapping (BAM)

Reads => Signal (Intermediate file)

Accumulating @ >1 Pbp/yr (currently),
~20% of tot. HiSeq output

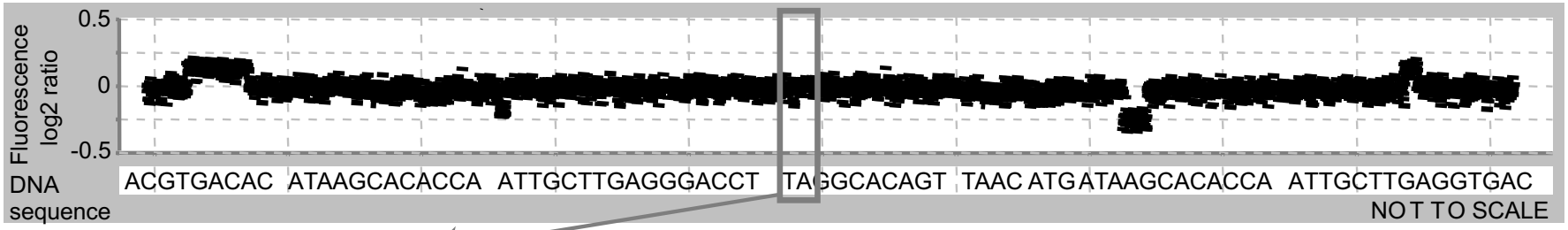
[*PLOS CB* 4:e1000158]

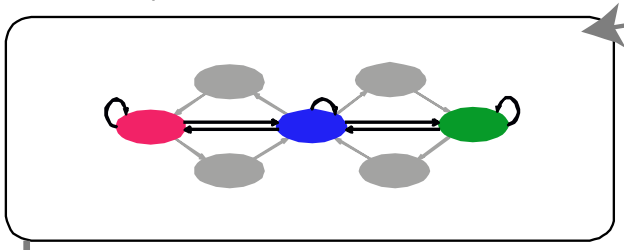
Example of Application to RD data

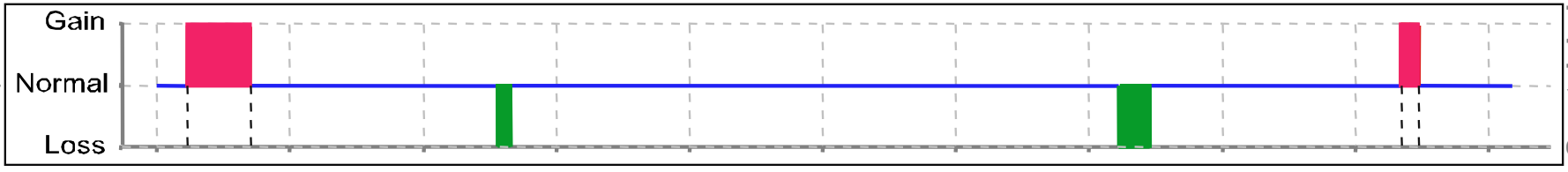


HMM

- 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

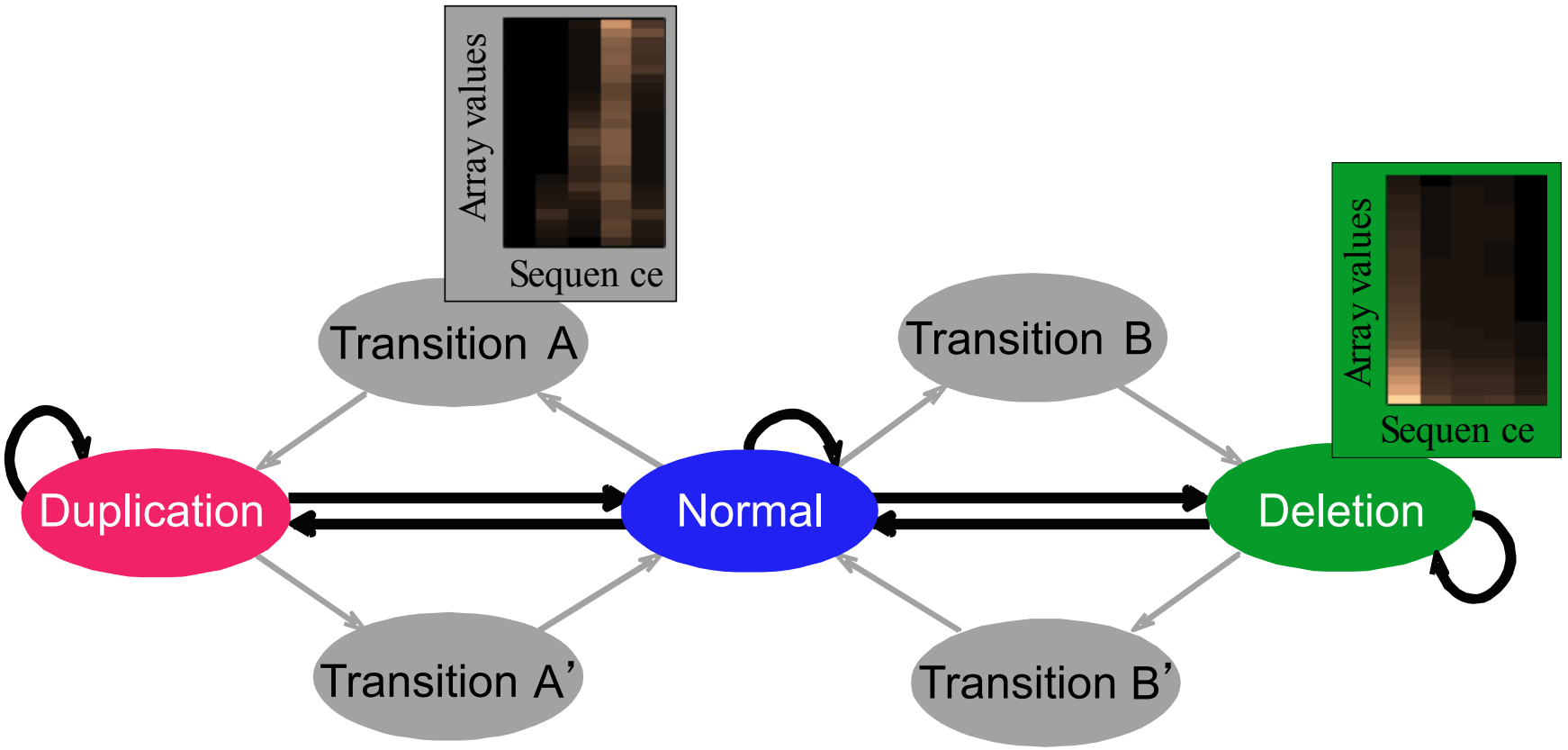






Korbel*, Urban* *et al.*, PNAS (2007)

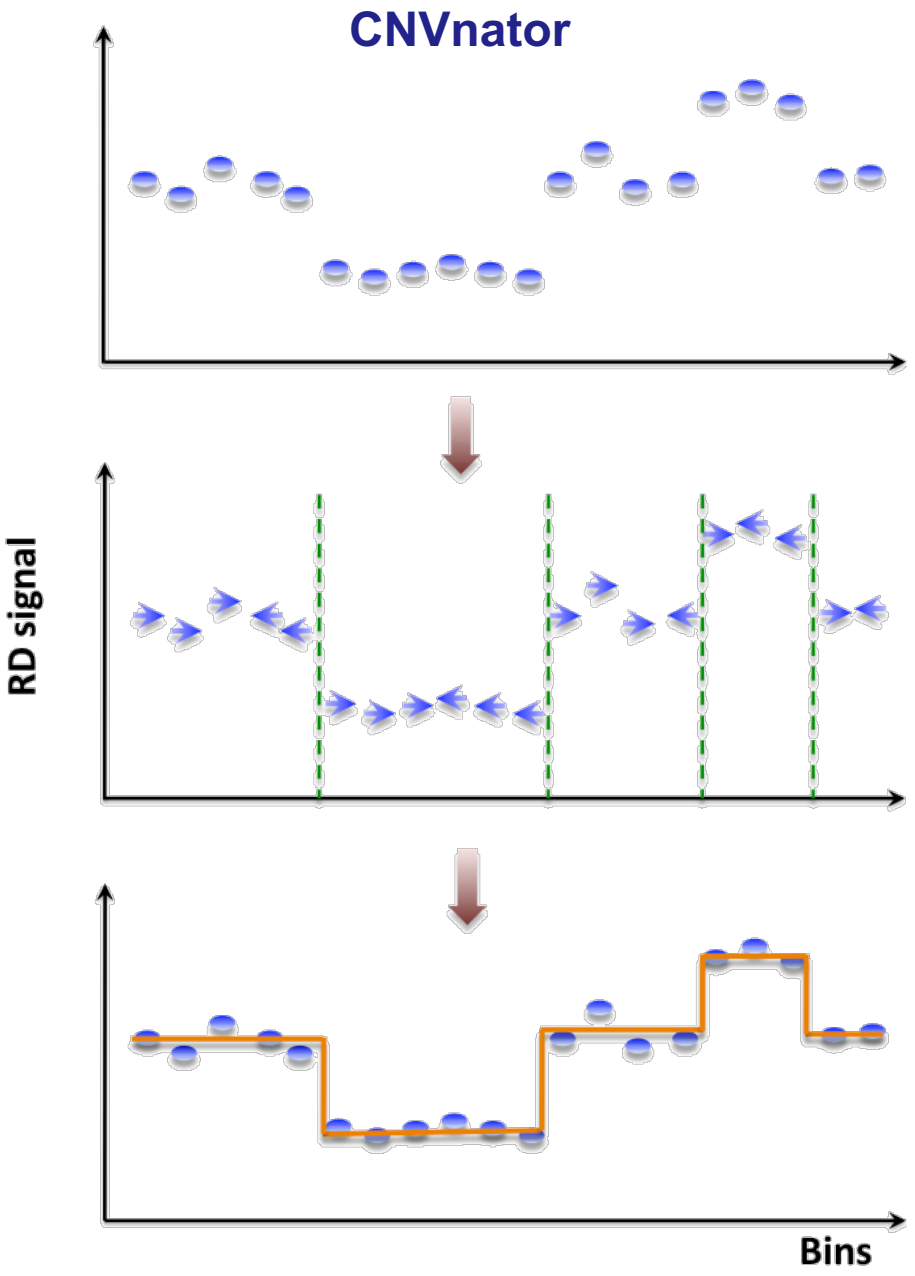
Statistically integrates array signal and DNA sequence signatures (using a discrete-valued bivariate HMM)



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

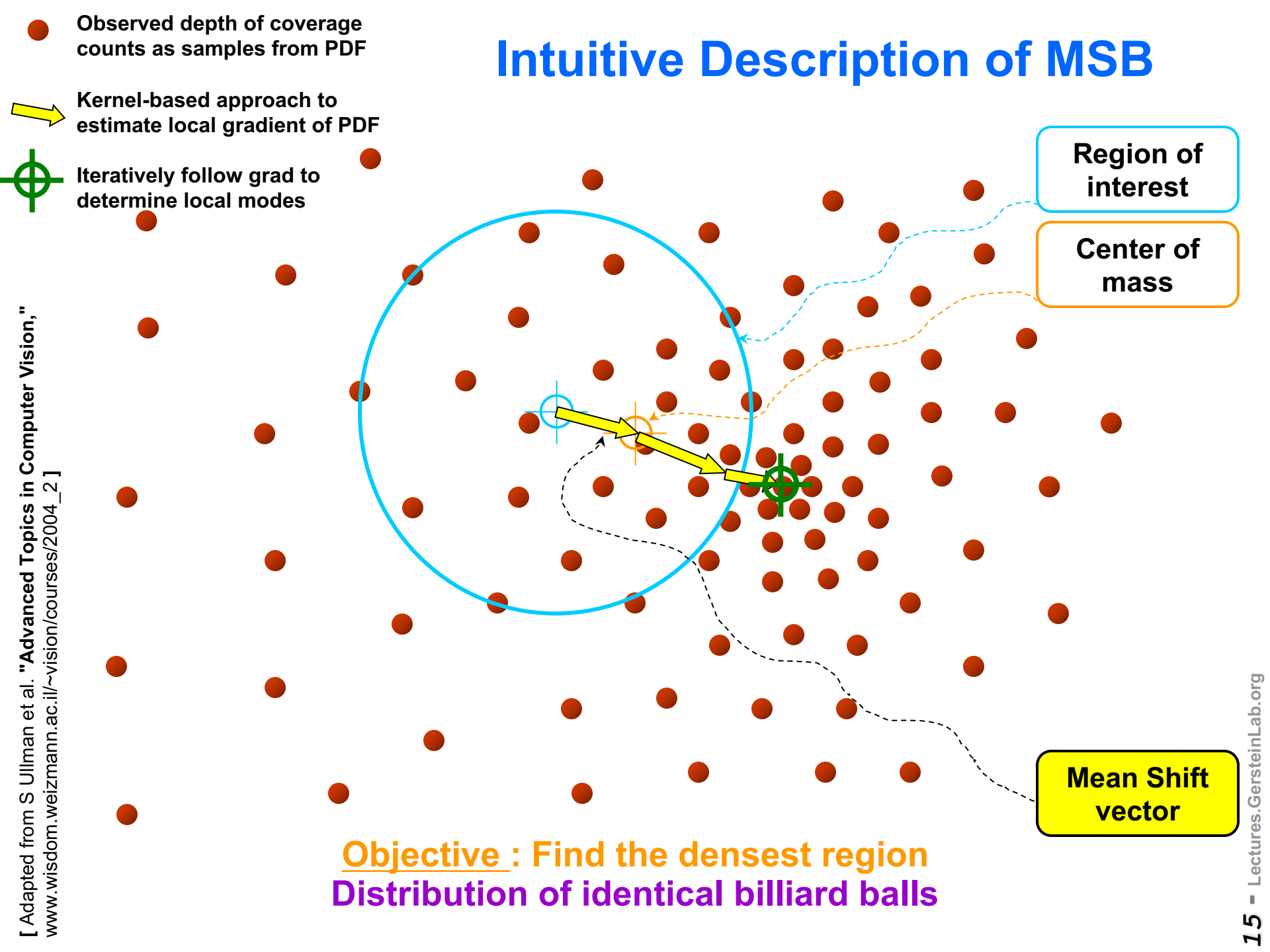
Mean-shift-based (MSB) segmentation: no explicit model

- 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



[Abyzov et al. Gen. Res. ('11)]

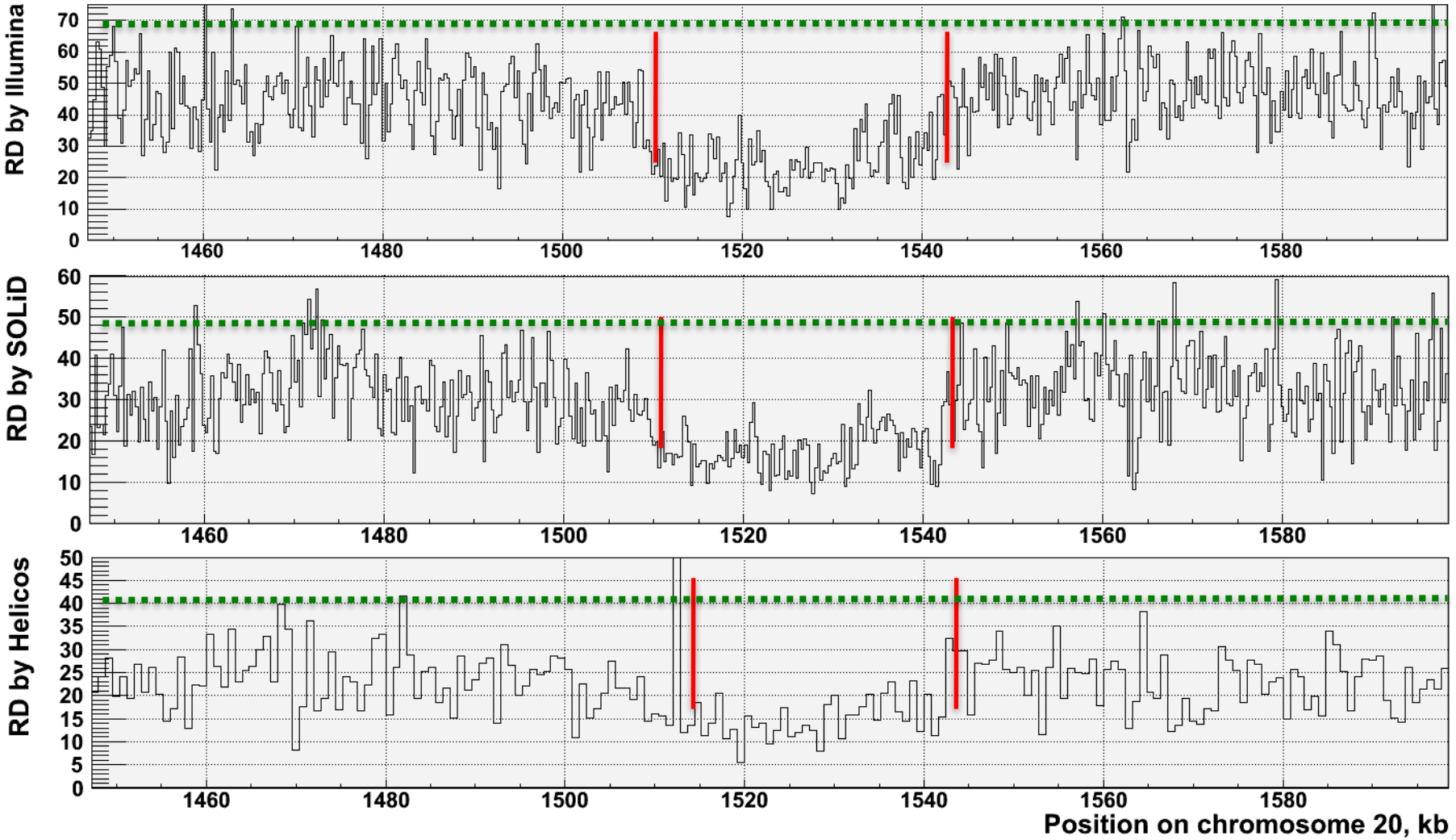
Intuitive Description of MSB



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

Split Read

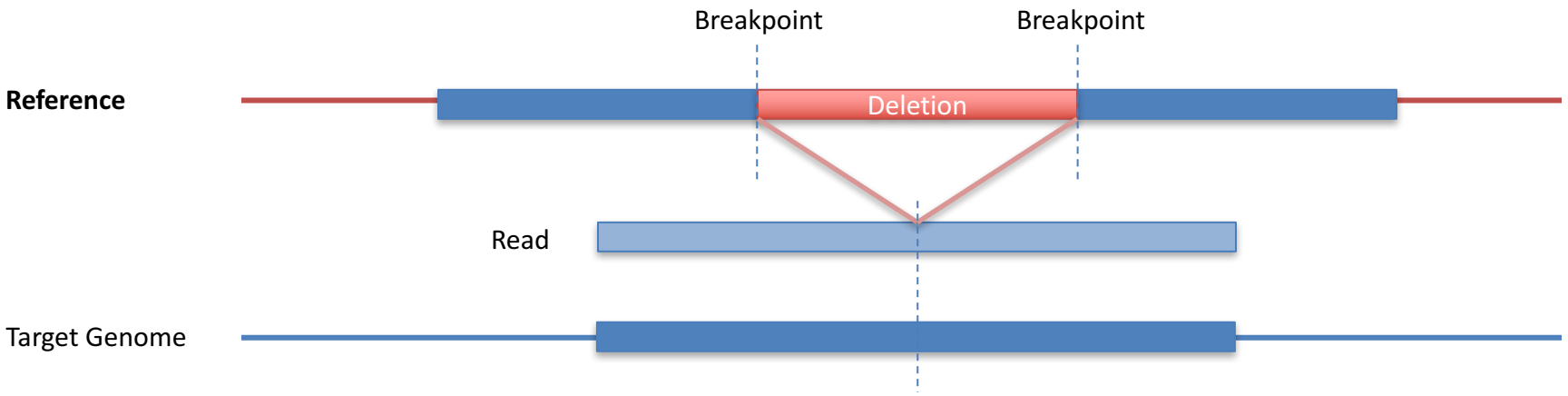
Read-depth works well on a variety of sequencing platforms but provides imprecise breakpoints

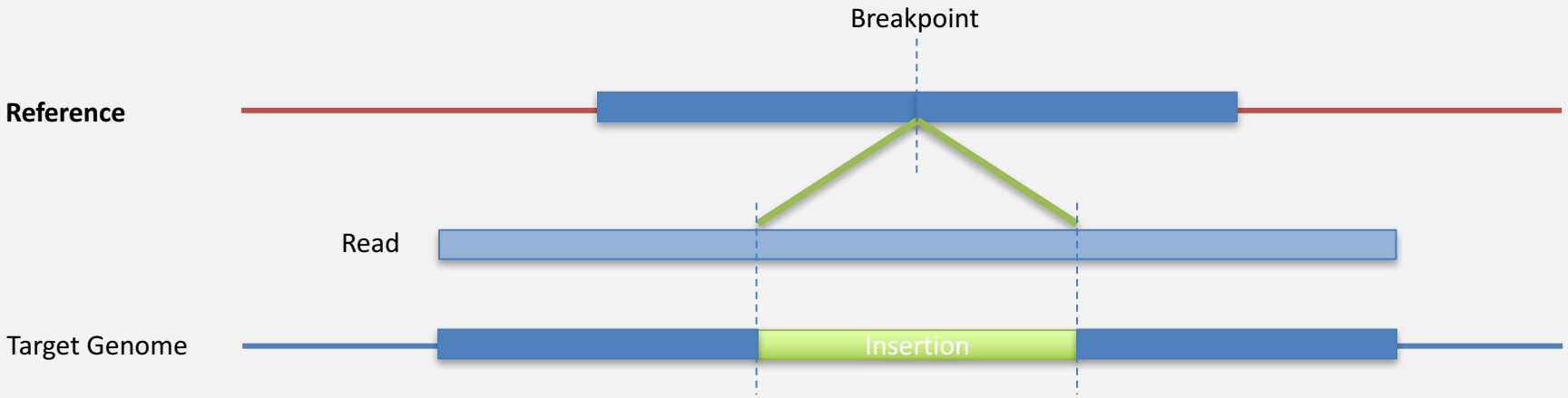


[Abyzov et al. Gen. Res. ('11)]

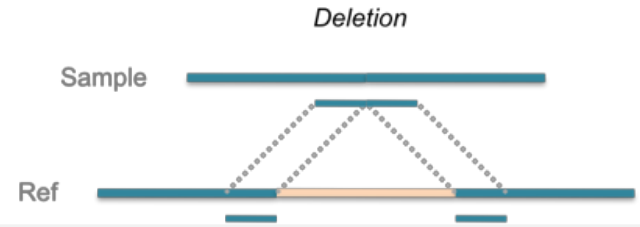
[NA18505]

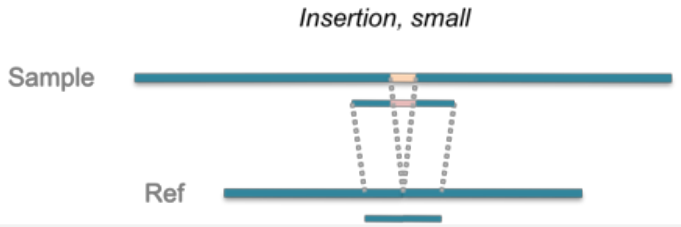
Split-read Analysis

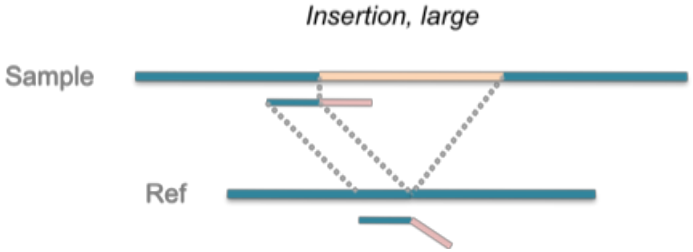




Simple SVs

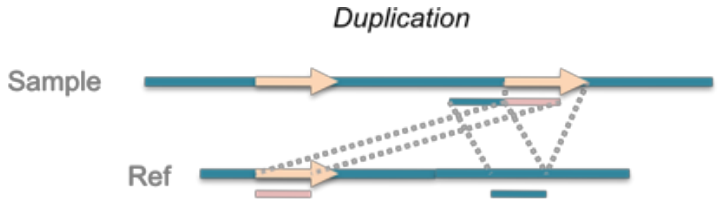


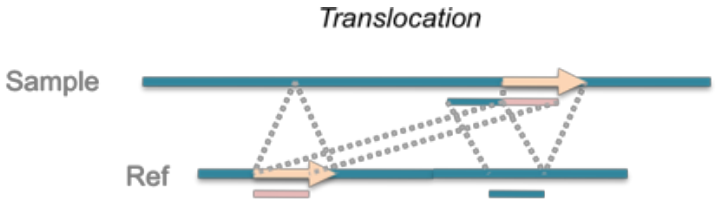


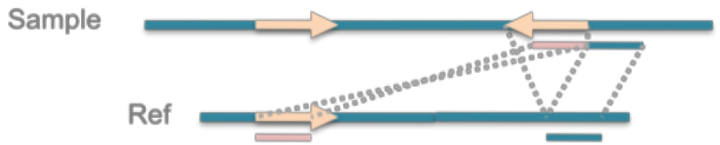


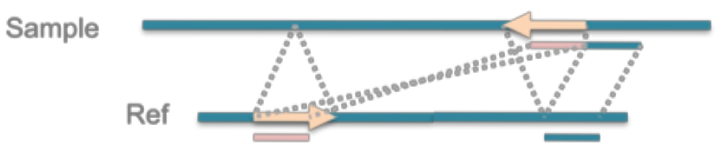
Deletions are the Easiest to Identify

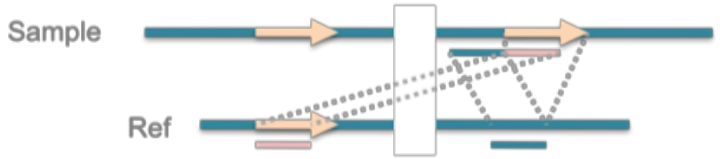
Complex SVs

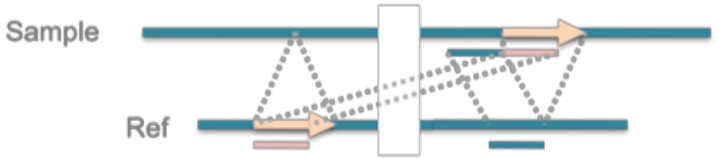








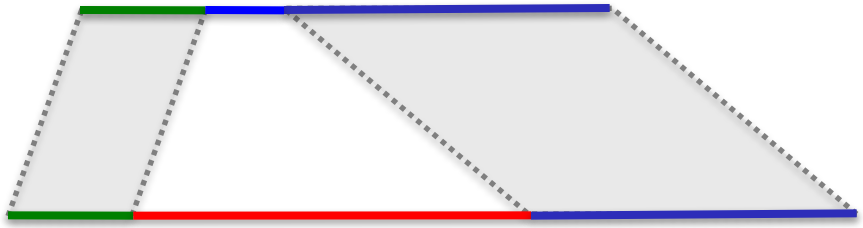




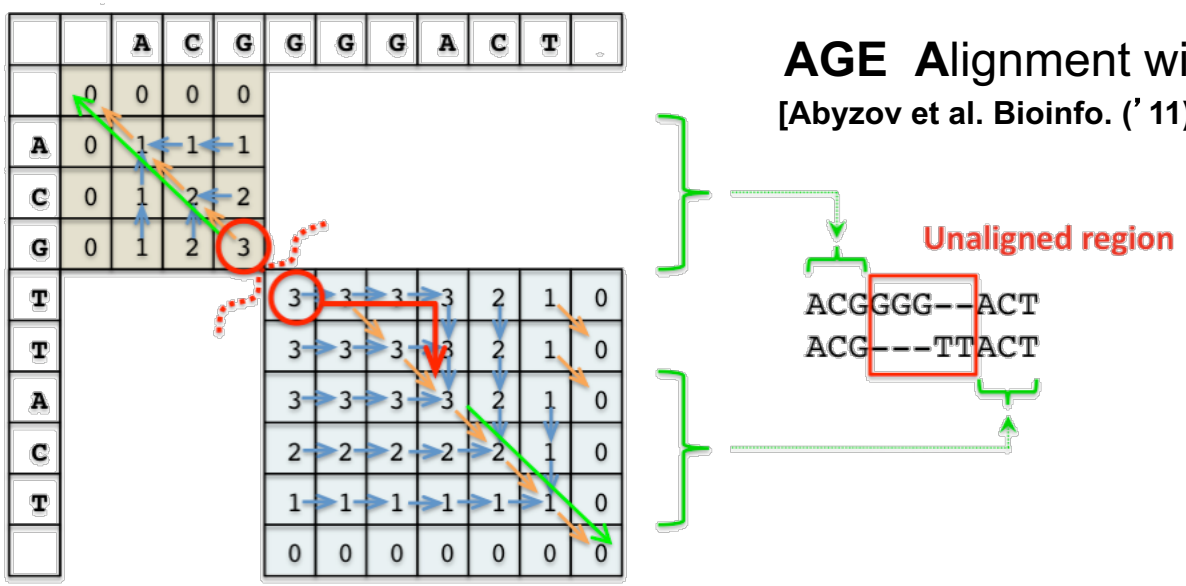
[Zhang et al. ('11) *BMC Genomics*]

Creative application of dynamic programming to a new problem

– Problem: Map insertions and deletions to a reference genome:



◊ Solution: SW alignment from both ends; combine max scoring alignments



AGE **A**lignment with **G**ap **E**xcision
[Abyzov et al. Bioinfo. (' 11)]

◊ much more detail in SV section later

Difficulties in Defining Exact Breakpoints

Optimal alignment

Small gap penalty

Large gap penalty

NW alignment

SW alignment

Optimal alignment

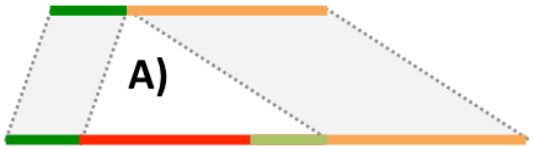
Small gap penalty

Large gap penalty

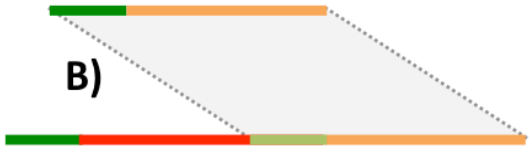
NW alignment

SW alignment

Optimal alignment



Local/global alignment at right

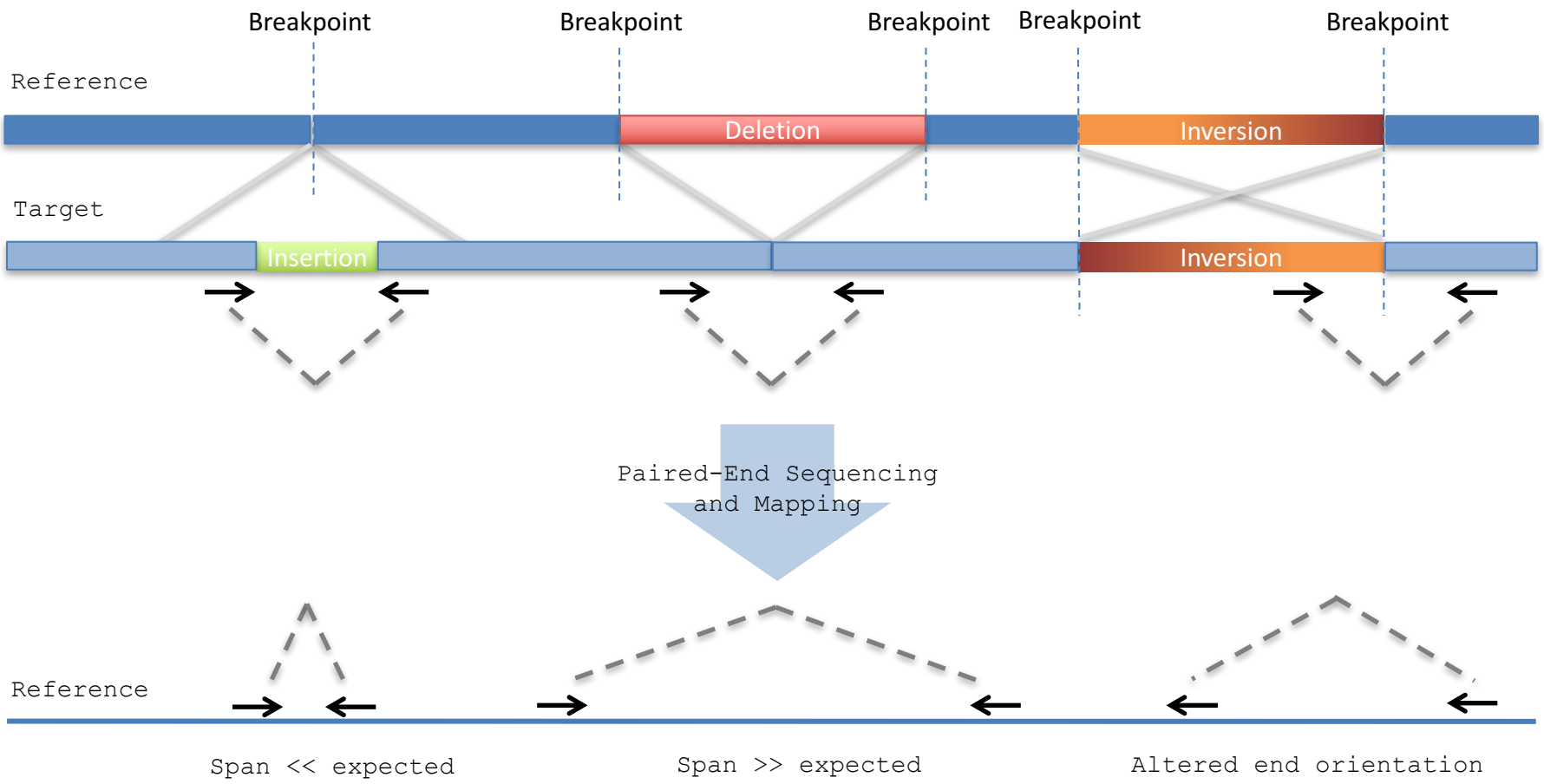


Local alignment at left

[Abyzov & Gerstein ('11) Bioinfo.]

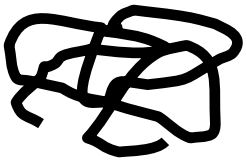
Paired-End

Paired-End Mapping



- Both paired-ends map within repeats.
- Limited the distance between pairs; therefore, neither large nor very small rearrangements can be detected

High-Resolution Paired-End Mapping (HR-PEM)



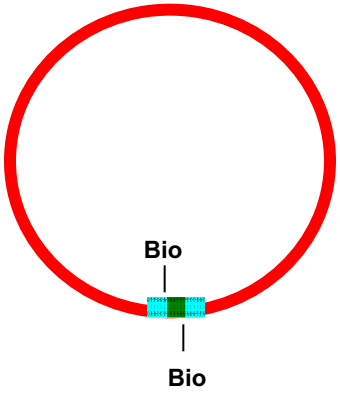
Genomic DNA

Shear to 3 kb

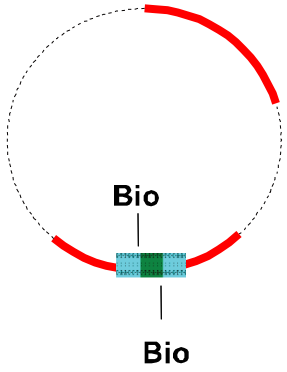
Adaptor ligation

Fragments

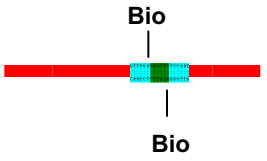
Circularize



Random Cleavage



Select

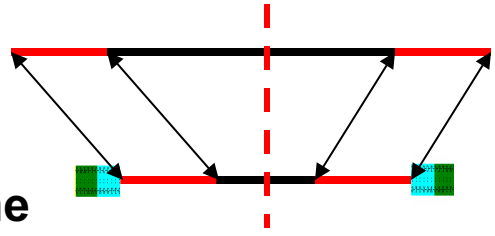


200-300bp

454 Massively Parallel Sequencing

(250bp/reads, 400k reads/run)

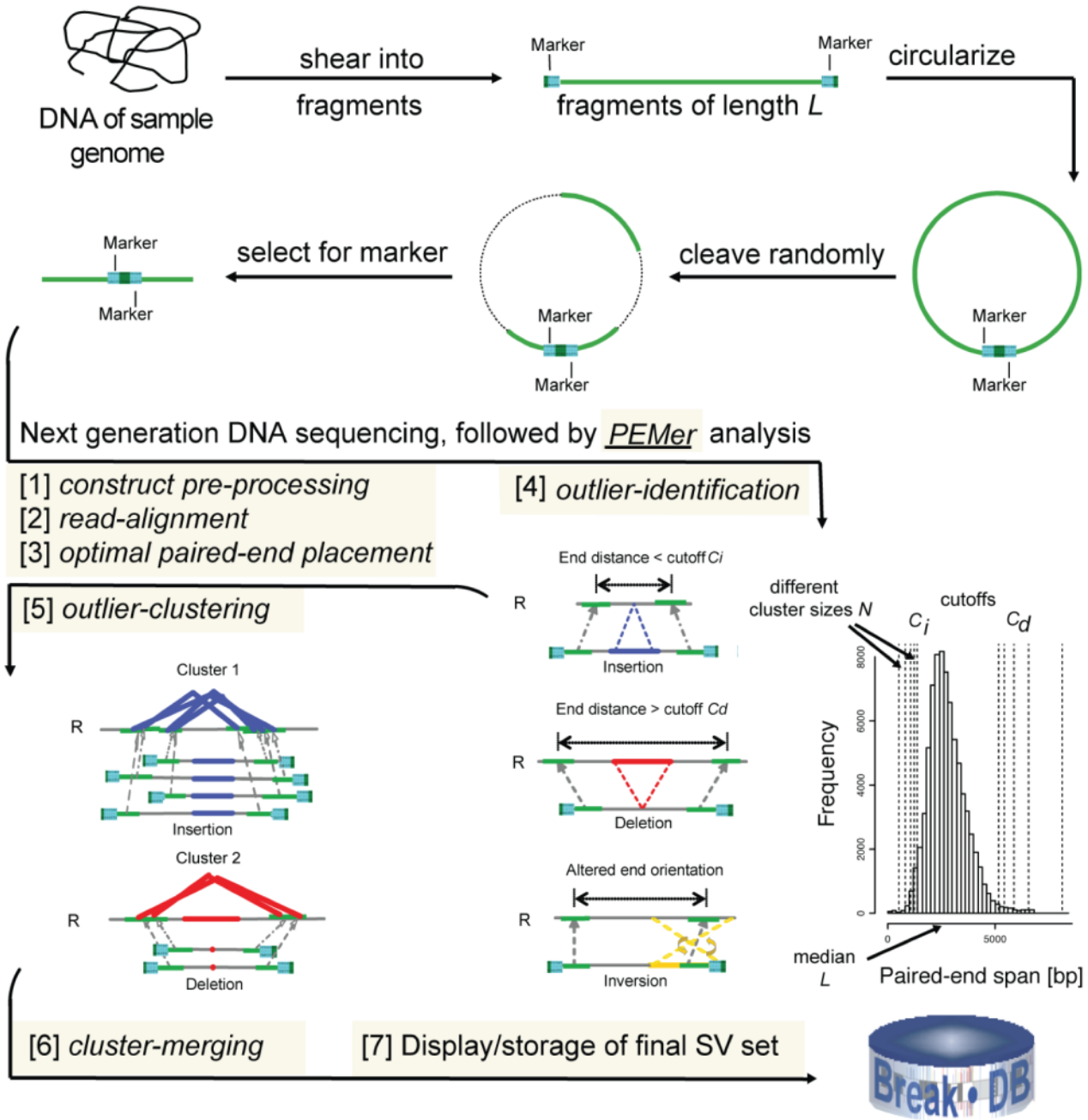
Map paired ends to human reference genome



3kb

Korbel et al., 2007 *Science*

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



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

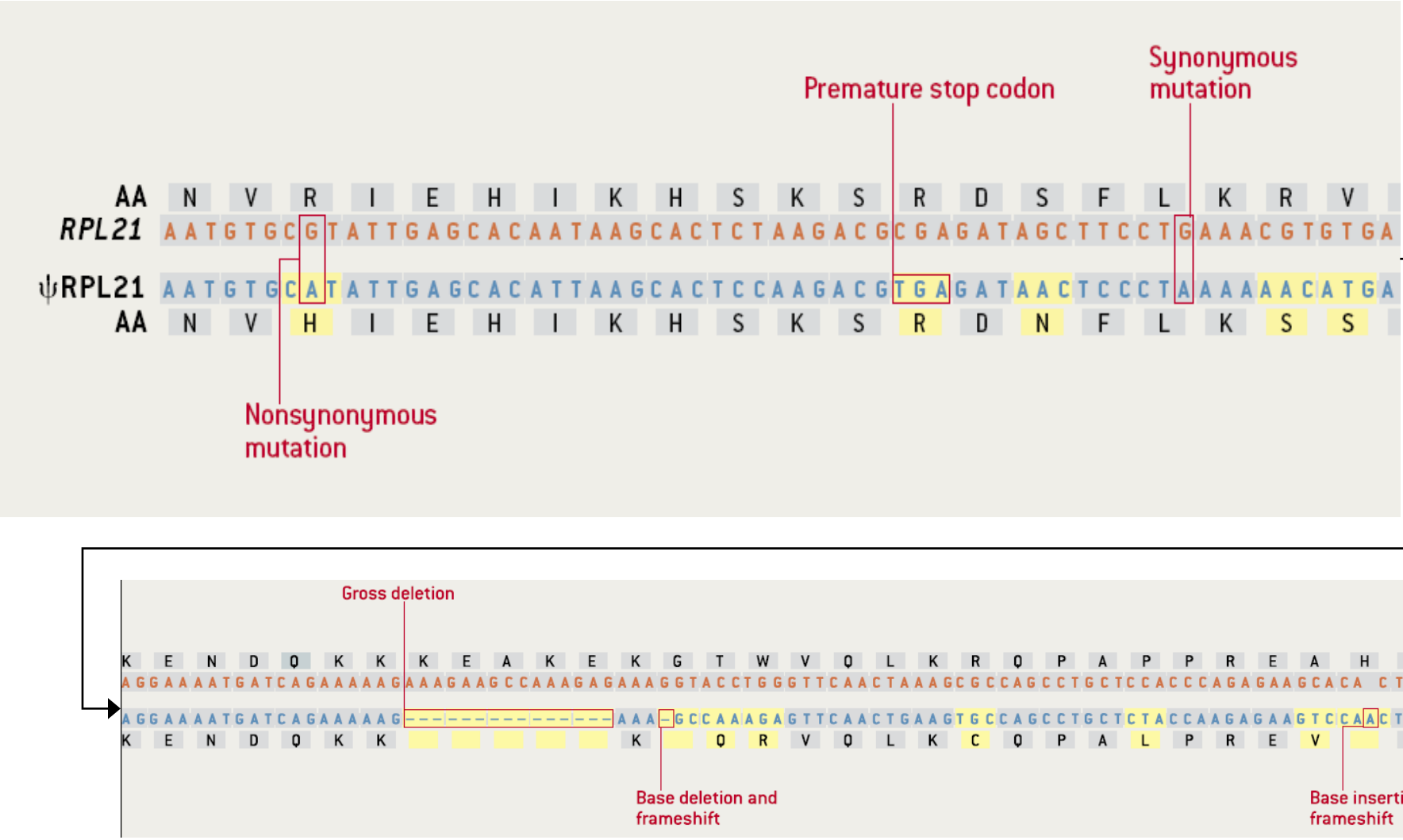
Pseudogenes & Genomic Duplications

Pseudogenes are among the most interesting intergenic elements

- Formal Properties of Pseudogenes (ΨG)
  - Inheritable
  - Homologous to a functioning element – ergo a repeat!
  - Non-functional
    - No selection pressure so free to accumulate mutations
      - Frameshifts & stops
      - Small Indels
      - Inserted repeats (LINE/Alu)
    - **What does this mean?** no transcription, no translation?…

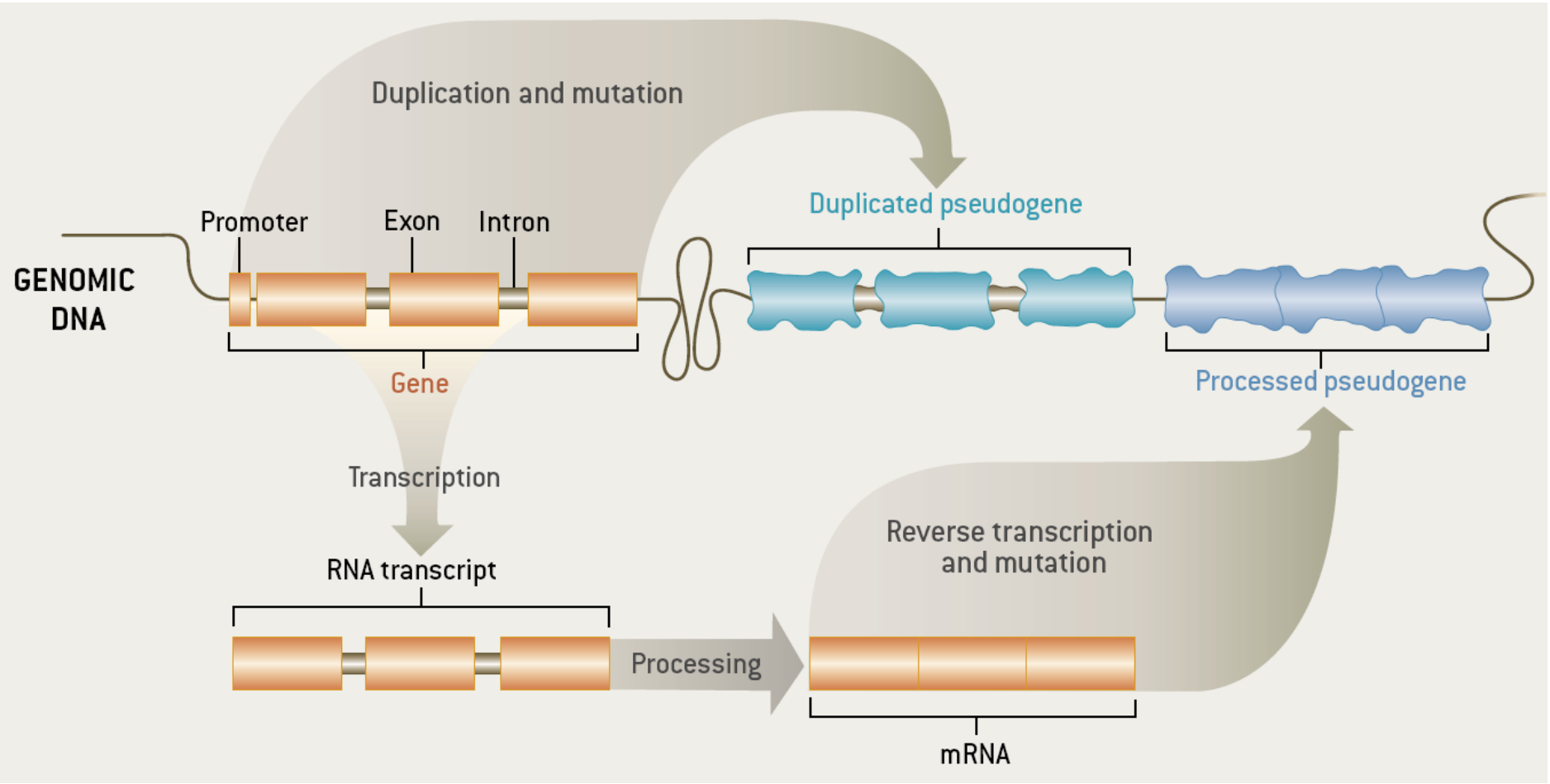
[Mighell et al. *FEBS Letts*, 2000]

Identifiable Features of a Pseudogene (ψRPL21)



[Gerstein & Zheng. Sci Am 295: 48 (2006).]

Two Major Genomic Remodeling Processes Give Rise to Distinct Types of Pseudogenes



[Gerstein & Zheng. Sci Am 295: 48 (2006).]

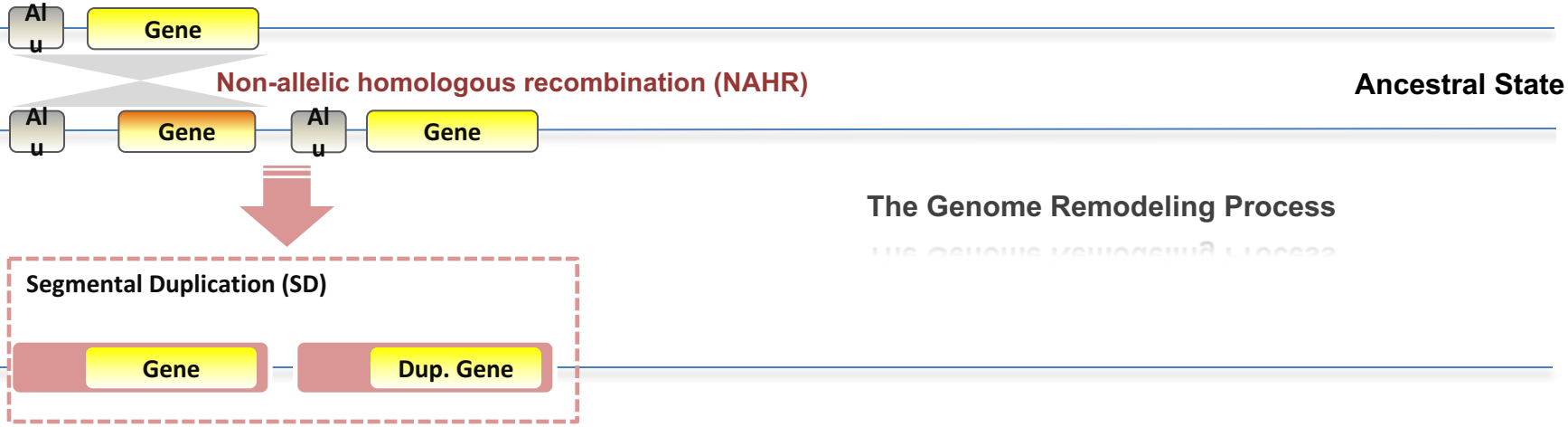
Impact of Genetic Variability: Loss-of-function

- - Truncating nonsense SNPs
- - Splice-disrupting SNPs
- - Frameshift-causing indels
- - Disrupting structural variants

- Previous LoFs are considered as having high probability of being deleterious
- Surprisingly, ~ 100 LoF variants per genome, 20 genes are completely inactivated
- Among ~100 LoFs, we estimate 2 recessive, close to 0 dominant disease nonsense variants per healthy genome.

Genomic Variation

Genomic Variation



Genomic Variation

Alu Gene

Non-allelic homologous recombination (NAHR)

Ancestral State

Alu Gene Alu Gene

The Genome Remodeling Process

Segmental Duplication (SD)

Gene Dup. Gene

Syntenic Ortholog

SD

Gene Dup. Gene

Paralog

duplicate

Gene Dup. Gene Dup. Gene

family

Gene Dup. ψgene

Genomic Variation

Alu Gene

Non-allelic homologous recombination (NAHR)

Ancestral State

Alu Gene Alu Gene

The Genome Remodeling Process

Segmental Duplication (SD)

Gene Dup. Gene

Syntenic Ortholog

SD

Gene Dup. Gene

Paralog

duplicate

Gene Dup. Gene Dup. Gene

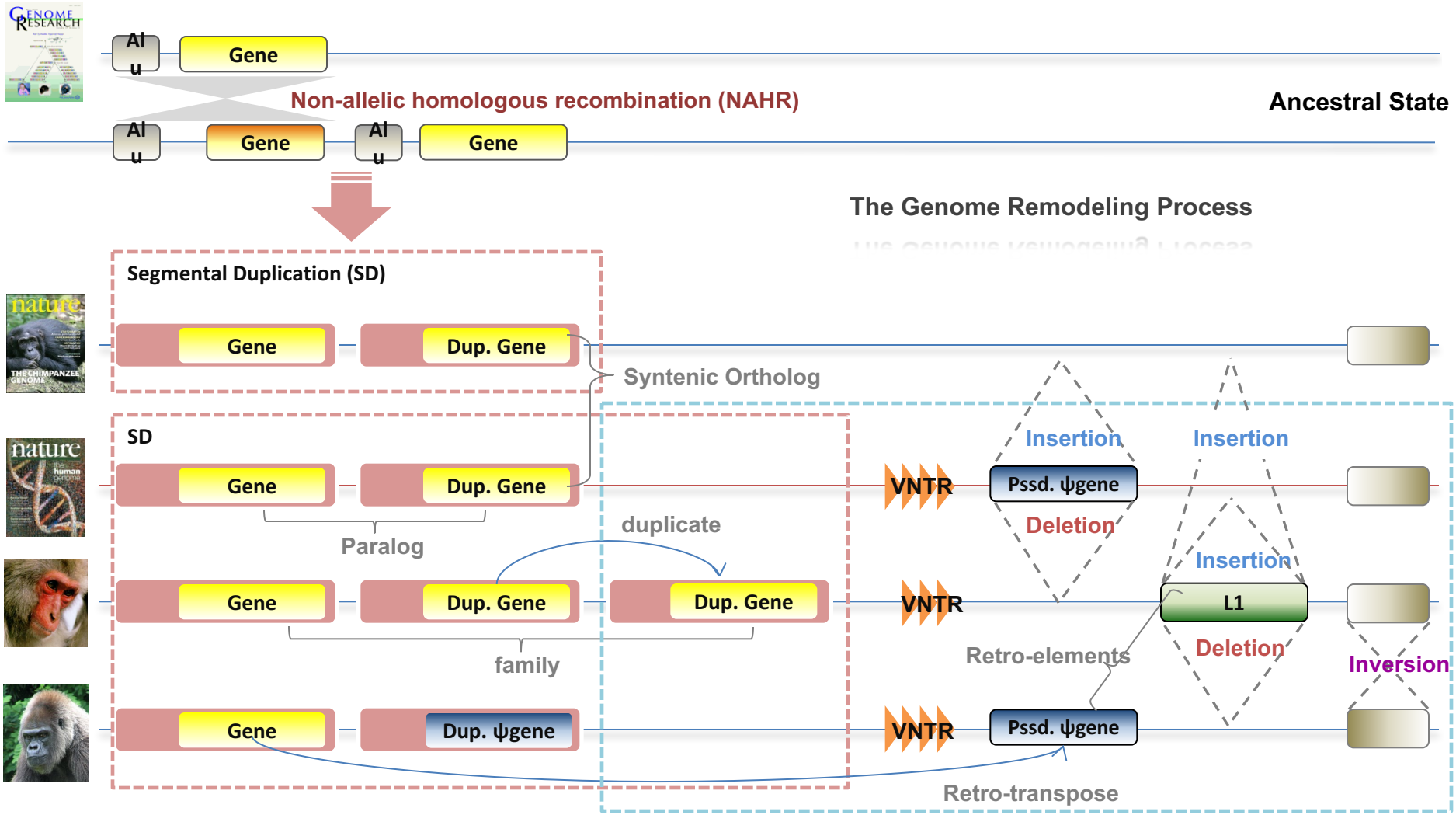
family

Gene Dup. ψgene

Pssd. ψgene

Retro-transpose

Genomic Variation



Genomic Variation

Alu Gene

Non-allelic homologous recombination (NAHR)

Ancestral State

Alu Gene Alu Gene

The Genome Remodeling Process

Segmental Duplication (SD)

Gene Dup. Gene

Syntenic Ortholog

CNV (type of SV)

Gene Dup. Gene

duplicate

Gene Dup. Gene Dup. Gene

Gene Dup. ψgene

VNTR Pssd. ψgene

Insertion

Deletion

Insertion

VNTR L1

Insertion

Deletion

Retro-elements

Inversion

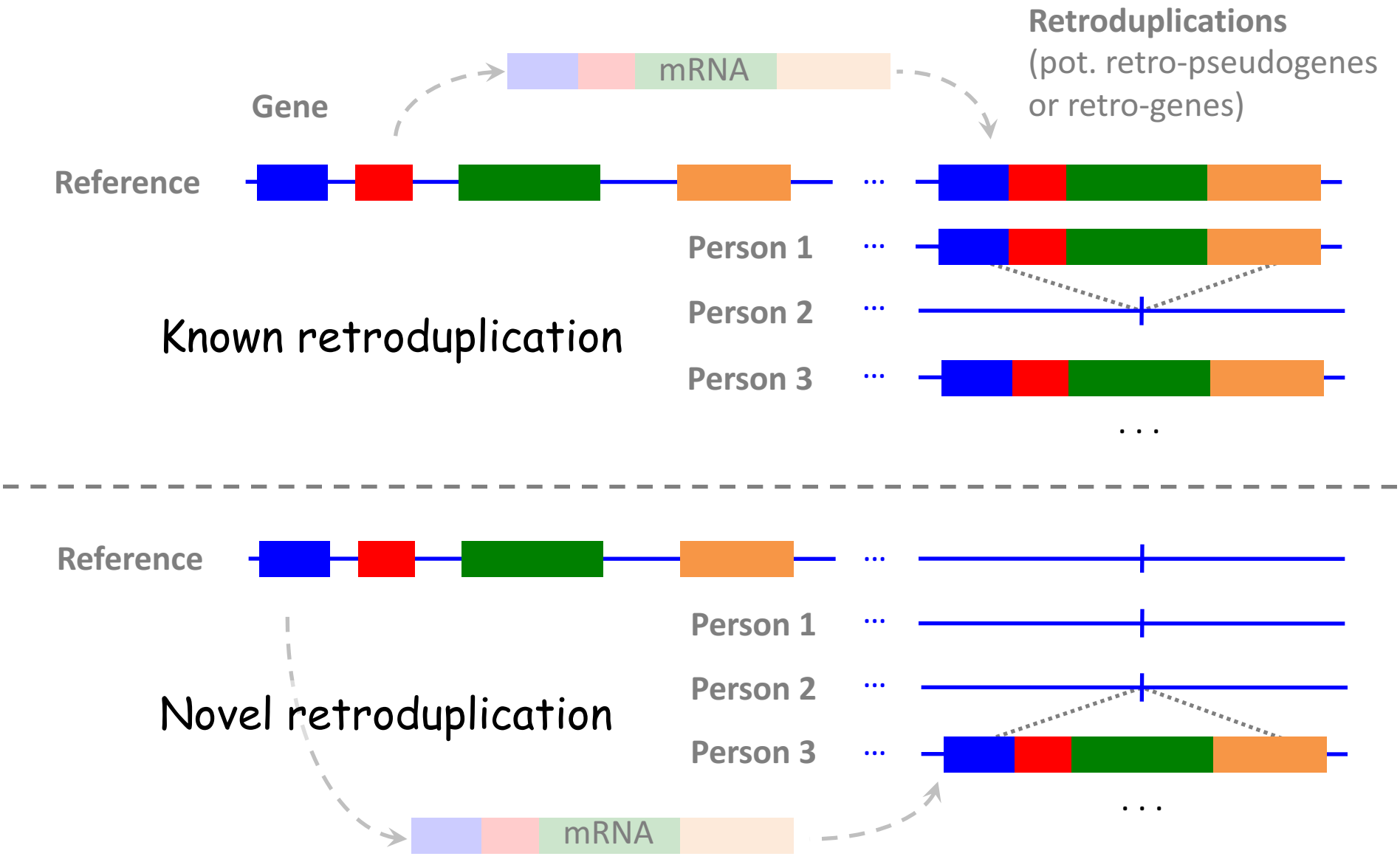
VNTR Pssd. ψgene

Retro-transpose

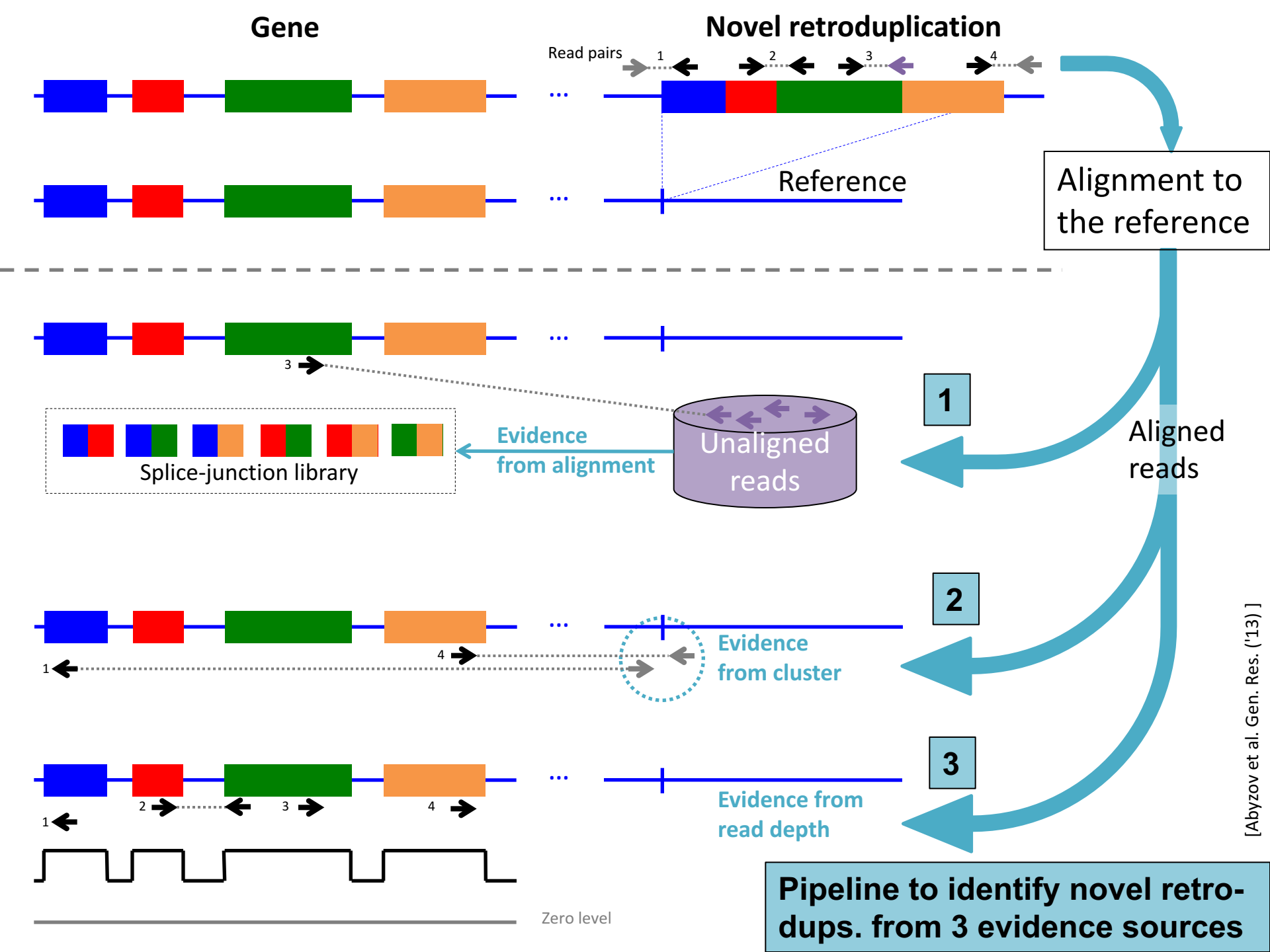
"Polymorphic" Genes & Pseudogenes

RDV & Mobile Elements

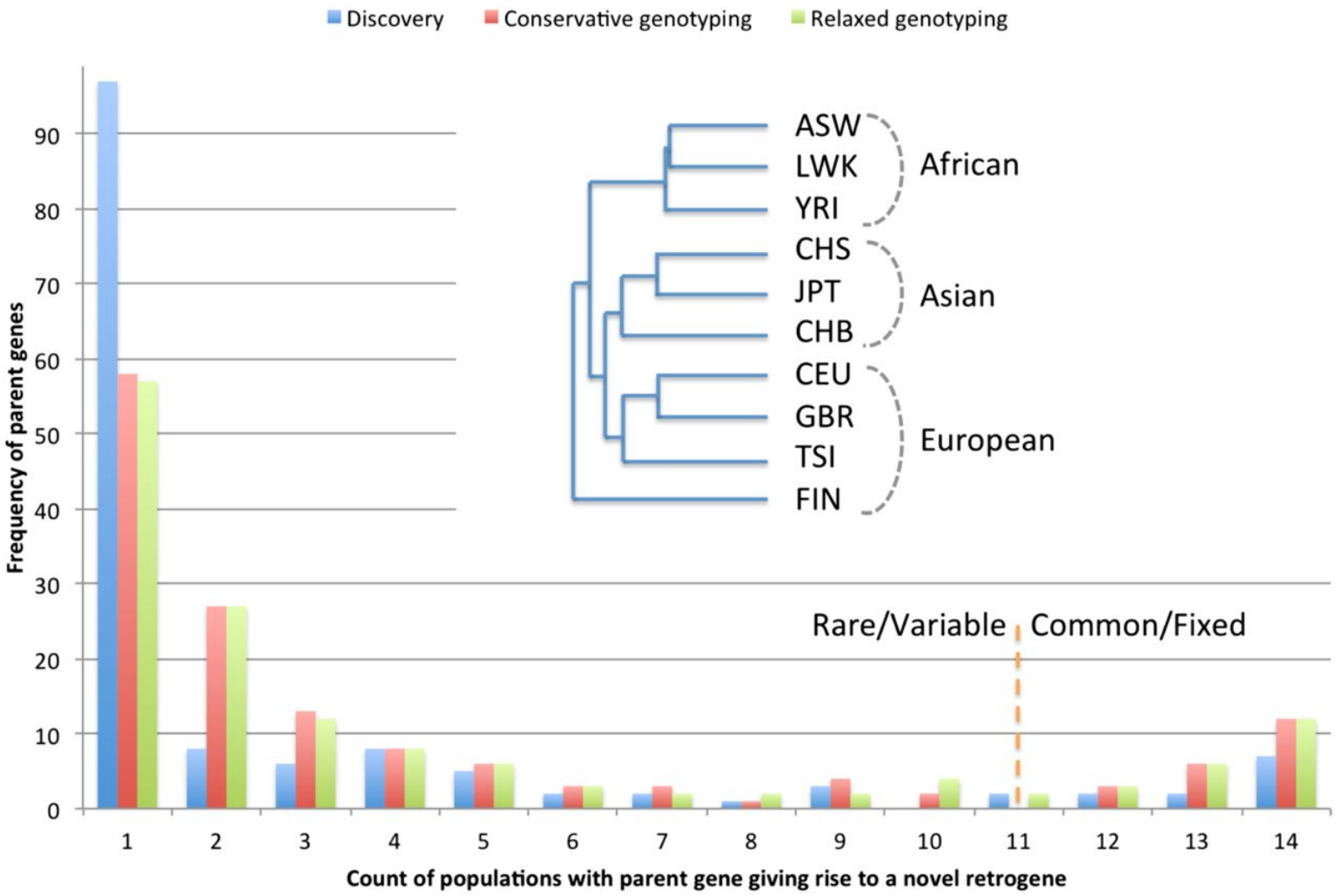
Retroduplication variation (RDV)



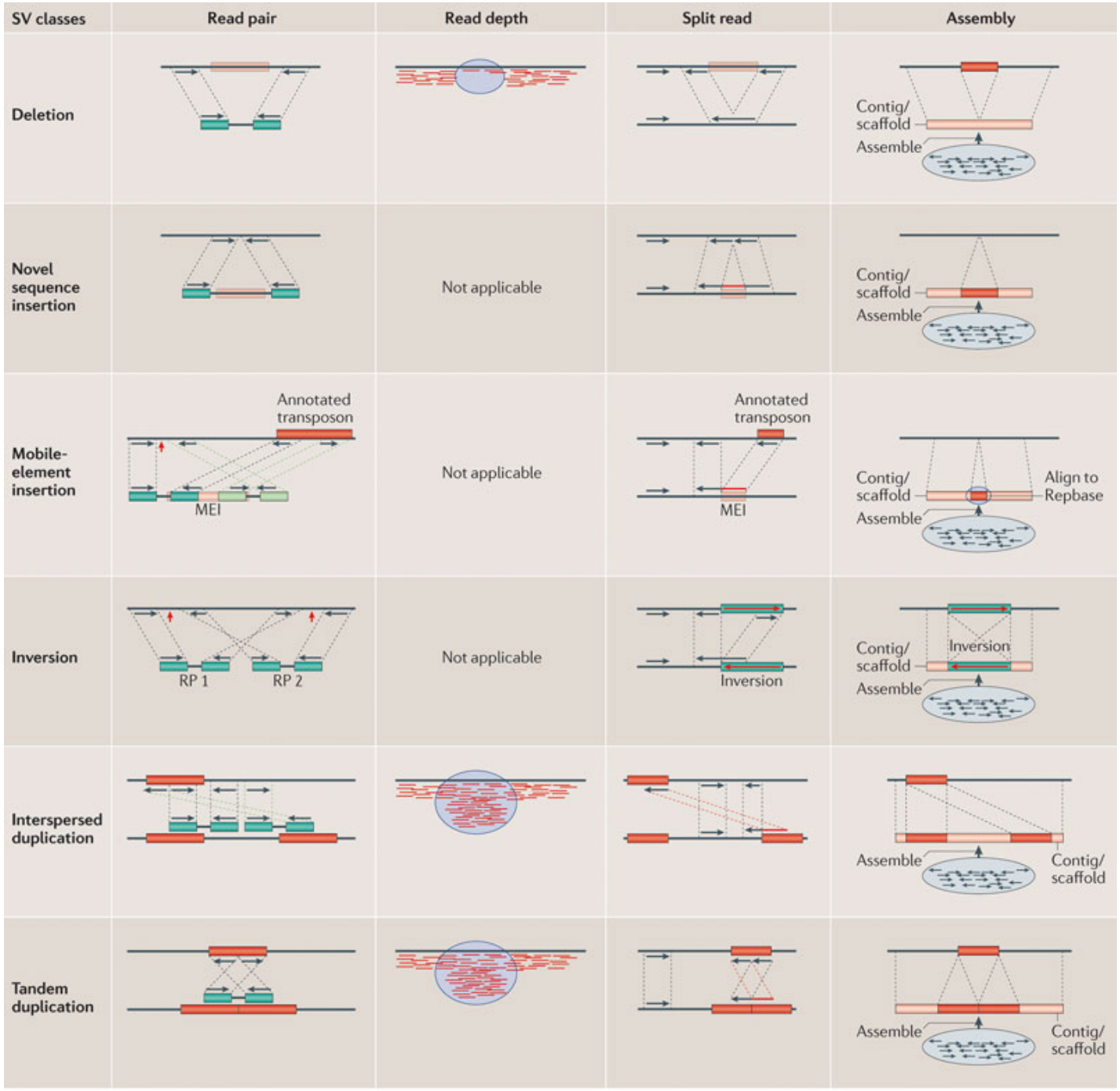
[Abyzov et al. Gen. Res. ('13)]


Gene
Novel retroduplication
Read pairs
1
2
3
4
Reference
Alignment to the reference
Aligned reads
Splice-junction library
Evidence from alignment
Unaligned reads
1
2
Evidence from cluster
3
Evidence from read depth
Zero level
[Abyzov et al. Gen. Res. ('13)]
Pipeline to identify novel retro-dups. from 3 evidence sources

Frequency of novel retroduplications by populations.



Abyzov A et al. Genome Res. 2013;23:2042-2052



Can Alkan, Bradley P. Coe & Evan E. Eichler
Nature Reviews Genetics 12, 363-376 (May 2011)

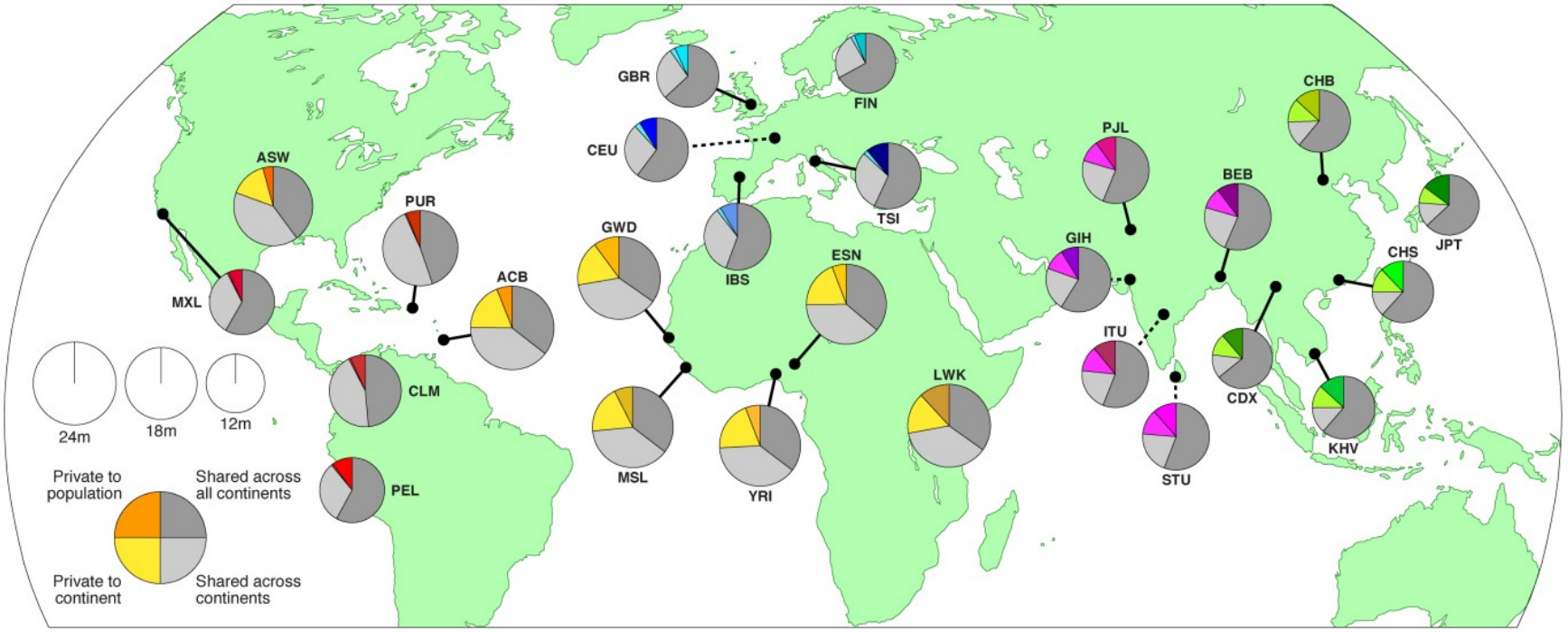
1000G summary

1000G SV (Pilot, **Phase I** & III)

- **Many** different callers compared & used
  - including SRiC & CNVnator but also VariationHunter, Cortex, NovelSeq, PEMer, BreakDancer, Mosaik, Pindel, GenomeSTRiP, mrFast….
- Merging
- Genotyping (GenomeSTRiP)
- Breakpoint assembly (AGE & Tigra_SV)
- Mechanism Classification

[1000 Genomes Consortium, Nature (2010, 2012); Mills et al., Nature (2011)]

Summary Stats of 1000GP SV Phase3



- 68,818 SVs
- 2,504 unrelated individuals
- 26 populaSons
- 37,250 SVs with resolved breakpoints

[2] 1000GP Phase3 SV paper. Submided to Nature, 2015.
[3] 1000GP ConsorSum. Submided to Nature, 2015.

Phase 3: Median Autosomal Variant Sites Per Genome

	AFR		AMR		EAS		EUR		SAS	
Samples	661		347		504		503		489	
Mean Coverage	8.2		7.6		7.7		7.4		8.0	
	Var. Sites	**Singletons**	**Var. Sites**	**Singletons**	**Var. Sites**	**Singletons**	**Var. Sites**	**Singletons**	**Var. Sites**	**Singletons**
SNPs	4.31M	14.5k	3.64M	12.0k	3.55M	14.8k	3.53M	11.4k	3.60M	14.4k
Indels	625k	-	557k	-	546k	-	546k	-	556k	-
Large Deletions	1.1k	5	949	5	940	7	939	5	947	5
CNVs	170	1	153	1	158	1	157	1	165	1
MEI (Alu)	1.03k	0	845	0	899	1	919	0	889	0
MEI (LINE1)	138	0	118	0	130	0	123	0	123	0
MEI (SVA)	52	0	44	0	56	0	53	0	44	0
MEI (MT)	5	0	5	0	4	0	4	0	4	0
Inversions	12	0	9	0	10	0	9	0	11	0
NonSynon	12.2k	139	10.4k	121	10.2k	144	10.2k	116	10.3k	144
Synon	13.8k	78	11.4k	67	11.2k	79	11.2k	59	11.4k	78
Intron	2.06M	7.33k	1.72M	6.12k	1.68M	7.39k	1.68M	5.68k	1.72M	7.20k
UTR	37.2k	168	30.8k	136	30.0k	169	30.0k	129	30.7k	168
Promoter	102k	430	84.3k	332	81.6k	425	82.2k	336	84.0k	430
Insulator	70.9k	248	59.0k	199	57.7k	252	57.7k	189	59.1k	243
Enhancer	354k	1.32k	295k	1.05k	289k	1.34k	288k	1.02k	295k	1.31k
TFBS	927	4	759	3	748	4	749	3	765	3
Filtered LoF	182	4	152	3	153	4	149	3	151	3
HGMD-DM	20	0	18	0	16	1	18	2	16	0
GWAS	2.00k	0	2.07k	0	1.99k	0	2.08k	0	2.06k	0
ClinVar	28	0	30	1	24	0	29	1	27	1

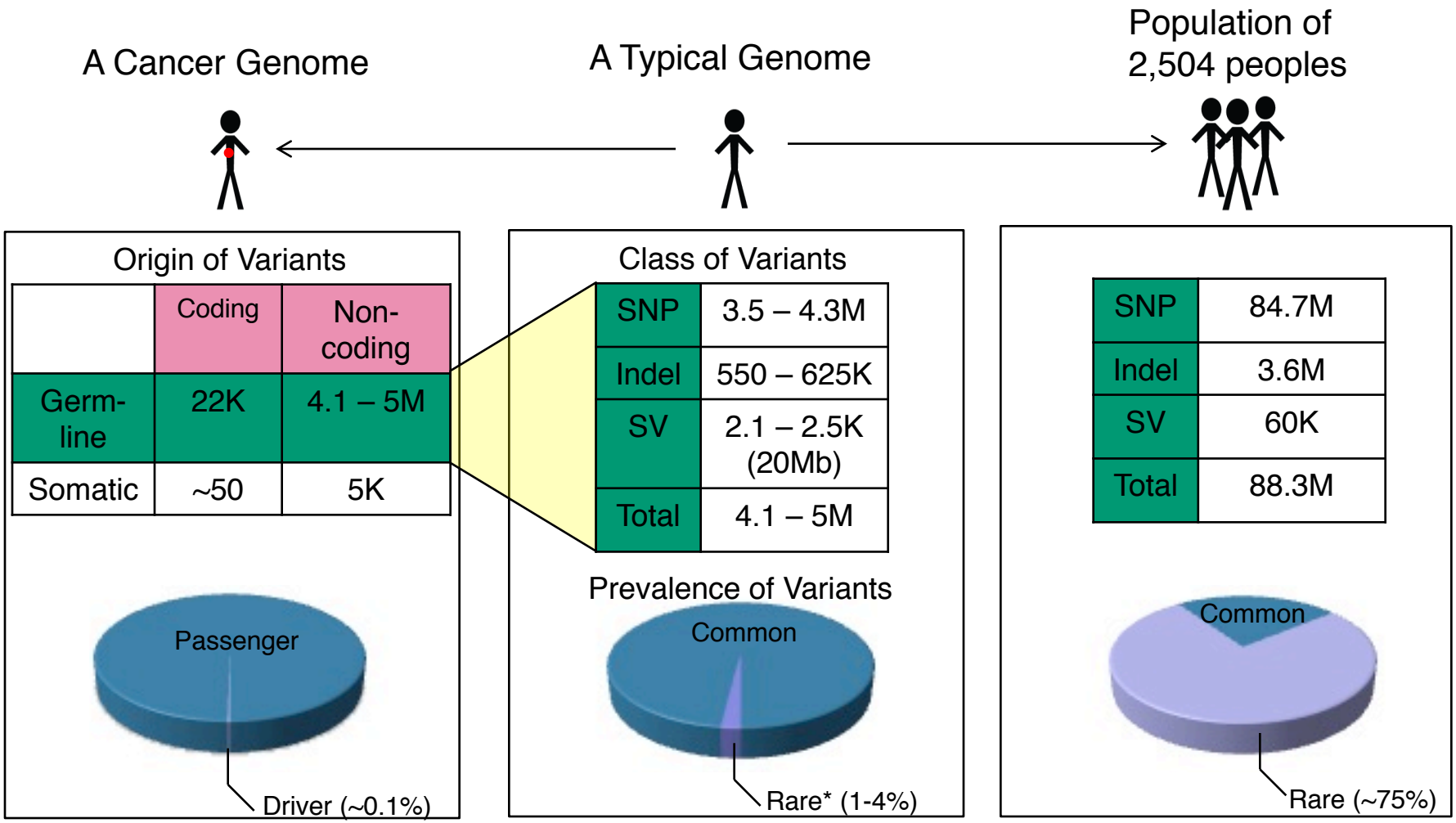
[3] 1000GP Consortium. Submitted to Nature,

A Typical Genome

- A typical genome differs from the reference genome at 4.09 – 5.02 million sites.
- The typical genome contains 2,100 – 2,500 SVs, covering ~20 million bases.
- A typical genome contains 149 – 182 sites with protein truncating variants, 10 – 12 thousand sites with peptide sequence altering variants, and 459 – 565 thousand variant sites overlapping regulatory regions.

[3] 1000GP Consortium. Submitted to Nature, 2015.

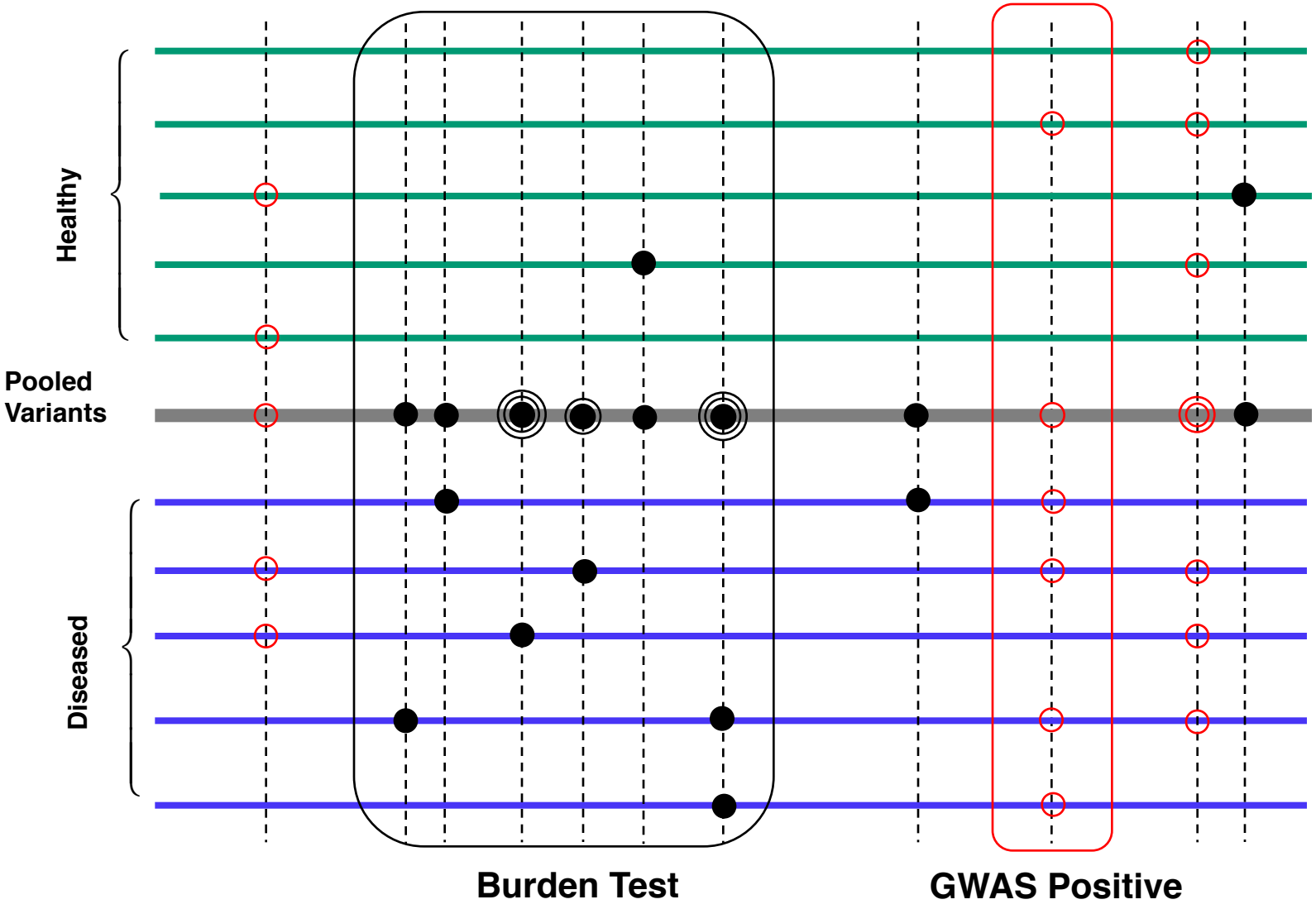
Human Genetic Variation



* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

Association of Variants with Diseases



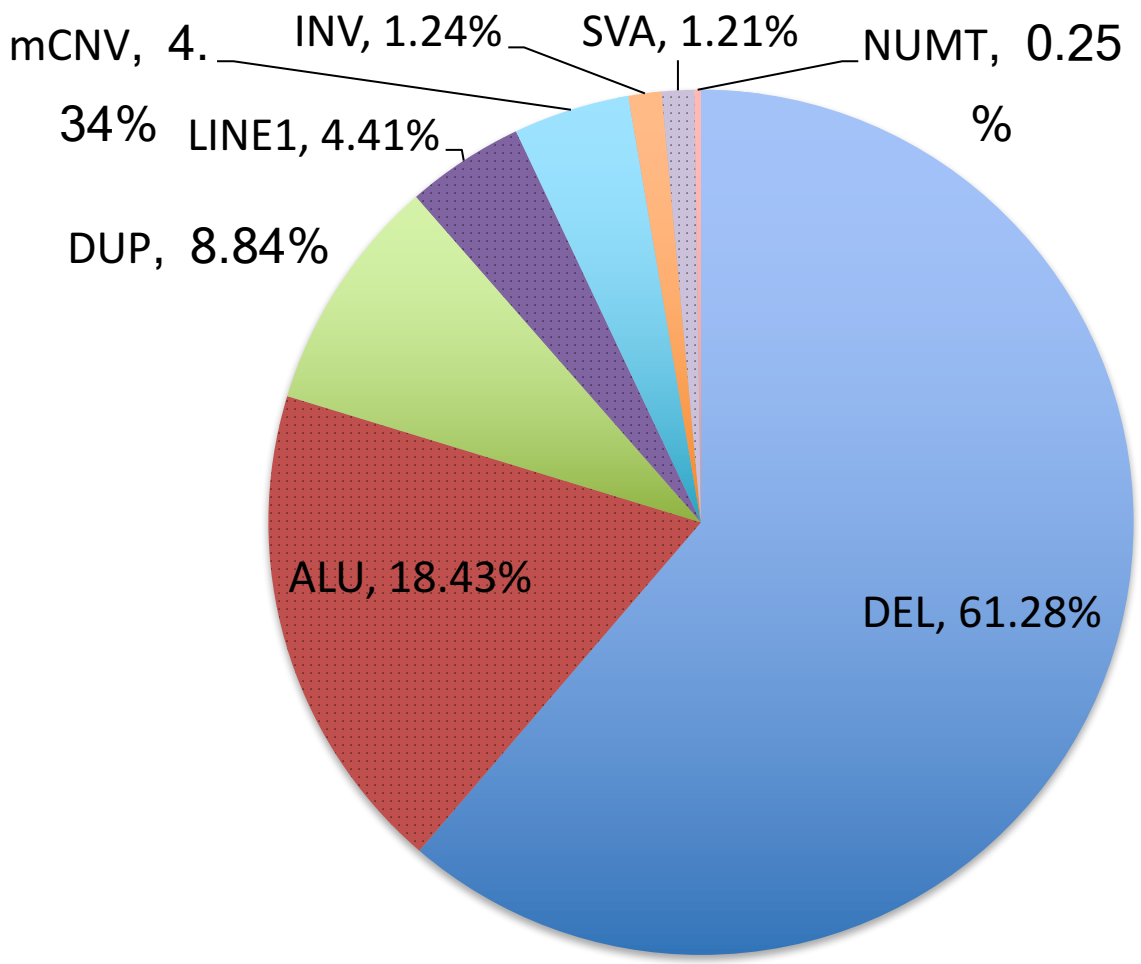
Structural Variations (SVs)

- SVs make up the majority of varying nucleotides among humans.

- More base pairs are altered as a result of SVs, than of single-nucleotide variations.
  - On the haploid reference assembly, a medium of 8.9 Mbp are affected by SVs, while 3.6 Mbp affected by SNPs.

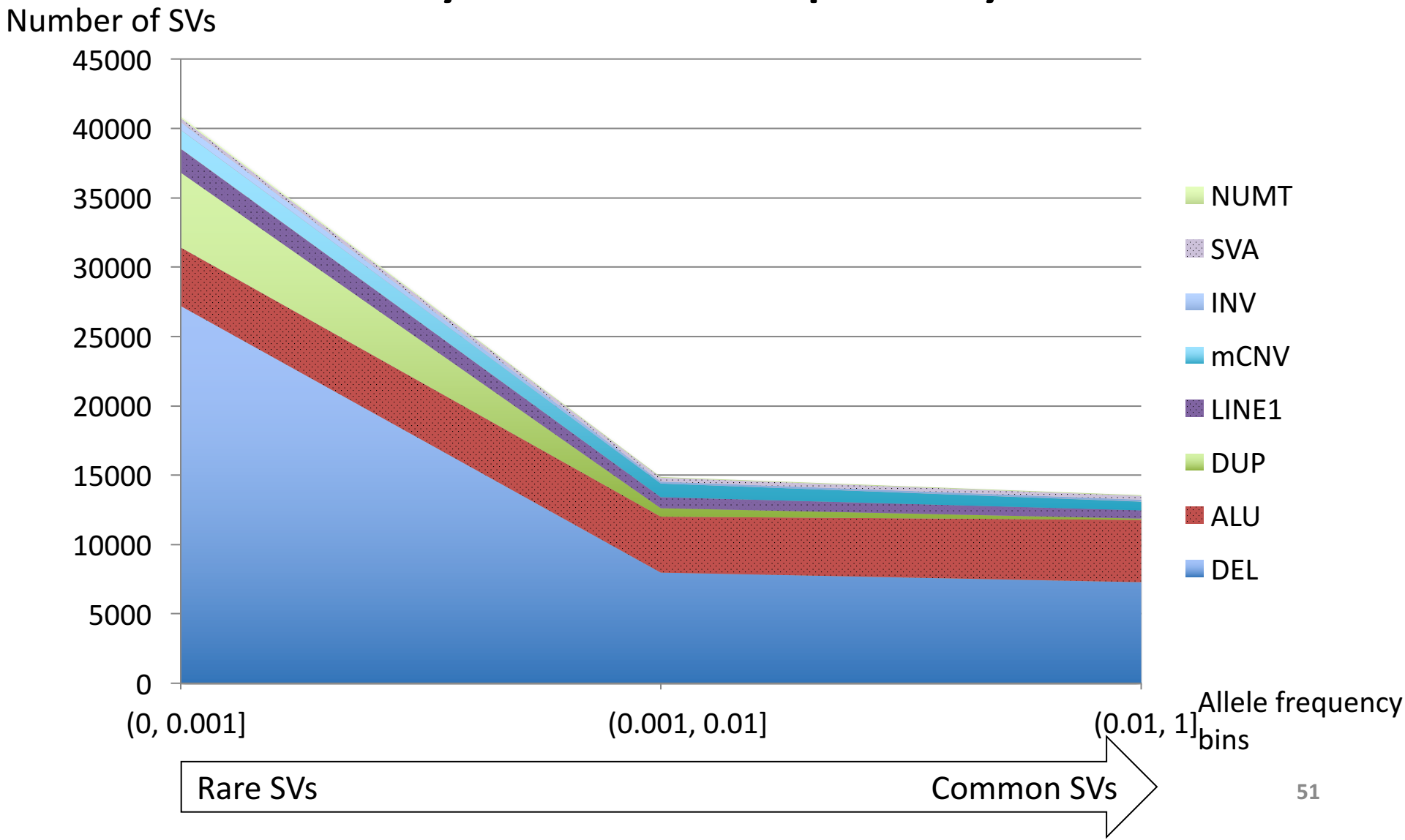
[1] Weischenfeldt J, et al. Nat Rev Genet, 2013.
[2] 1000GP Phase3 SV paper. Submided to Nature, 2015.

Distribution of Different SVs in Normal Human Populations

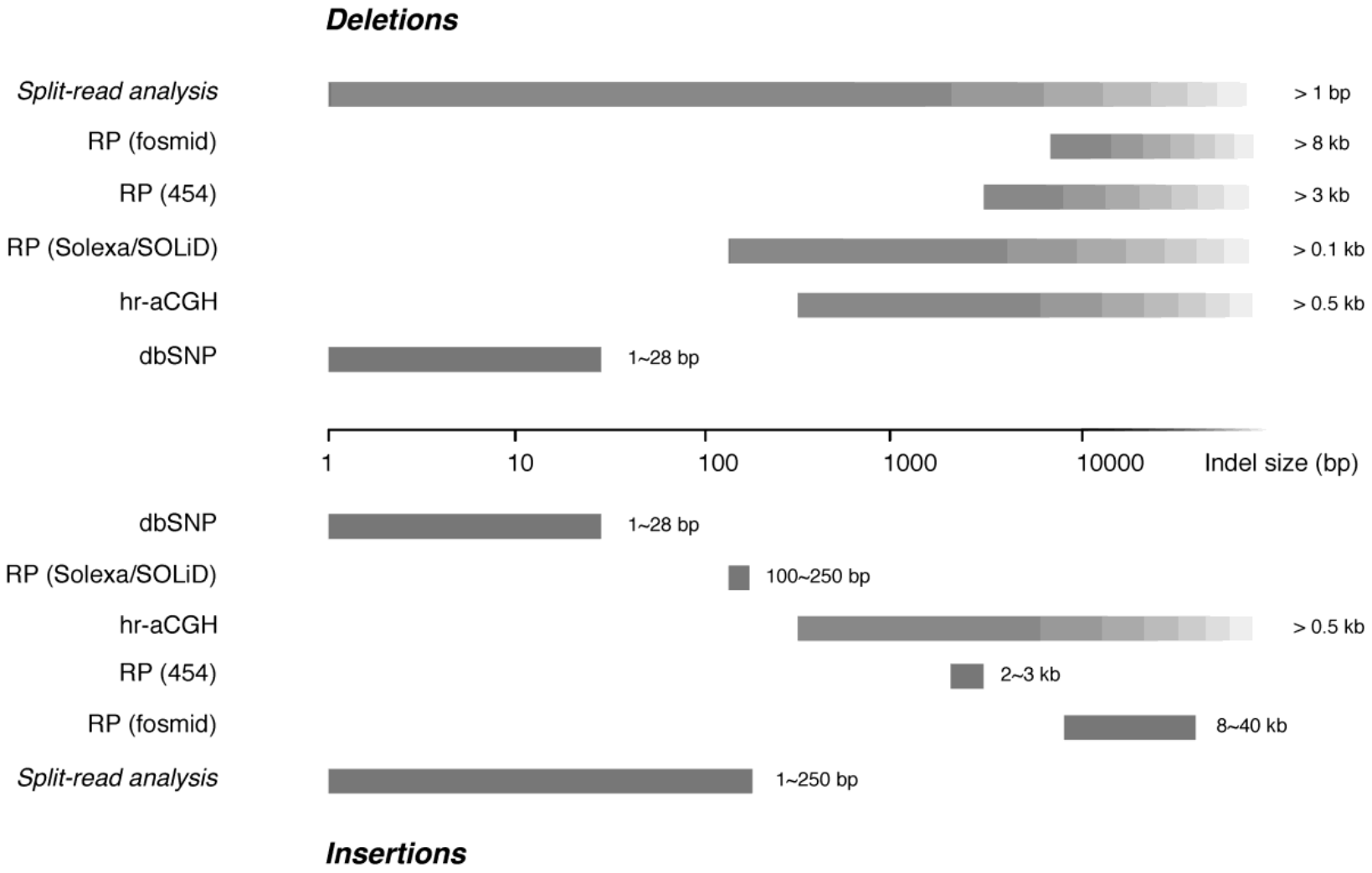


Total ~70K SVs from over 2,500 normal individuals (the 1000 Genomes Project)

Distribution of Different SVs Stratified by Allele Frequency

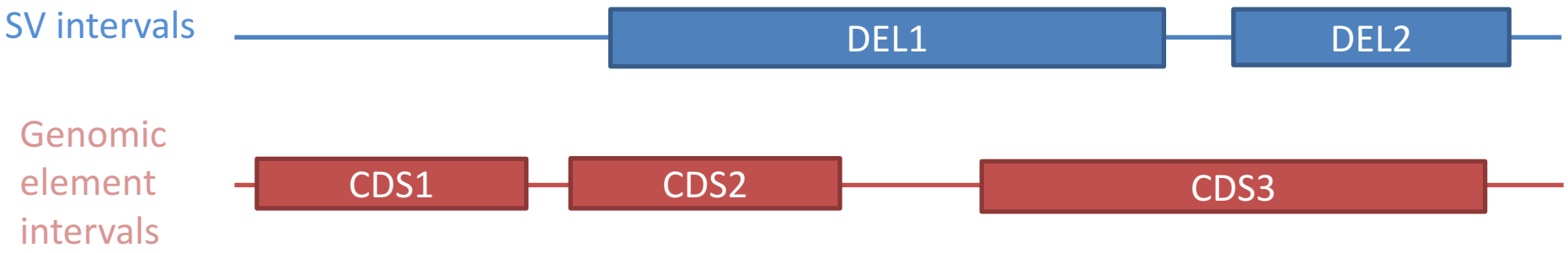


Different Approaches Work Differently on Different Events



[Zhang et al. ('11) *BMC Genomics*]

Measure of Overlap between SVs and Genomic Elements



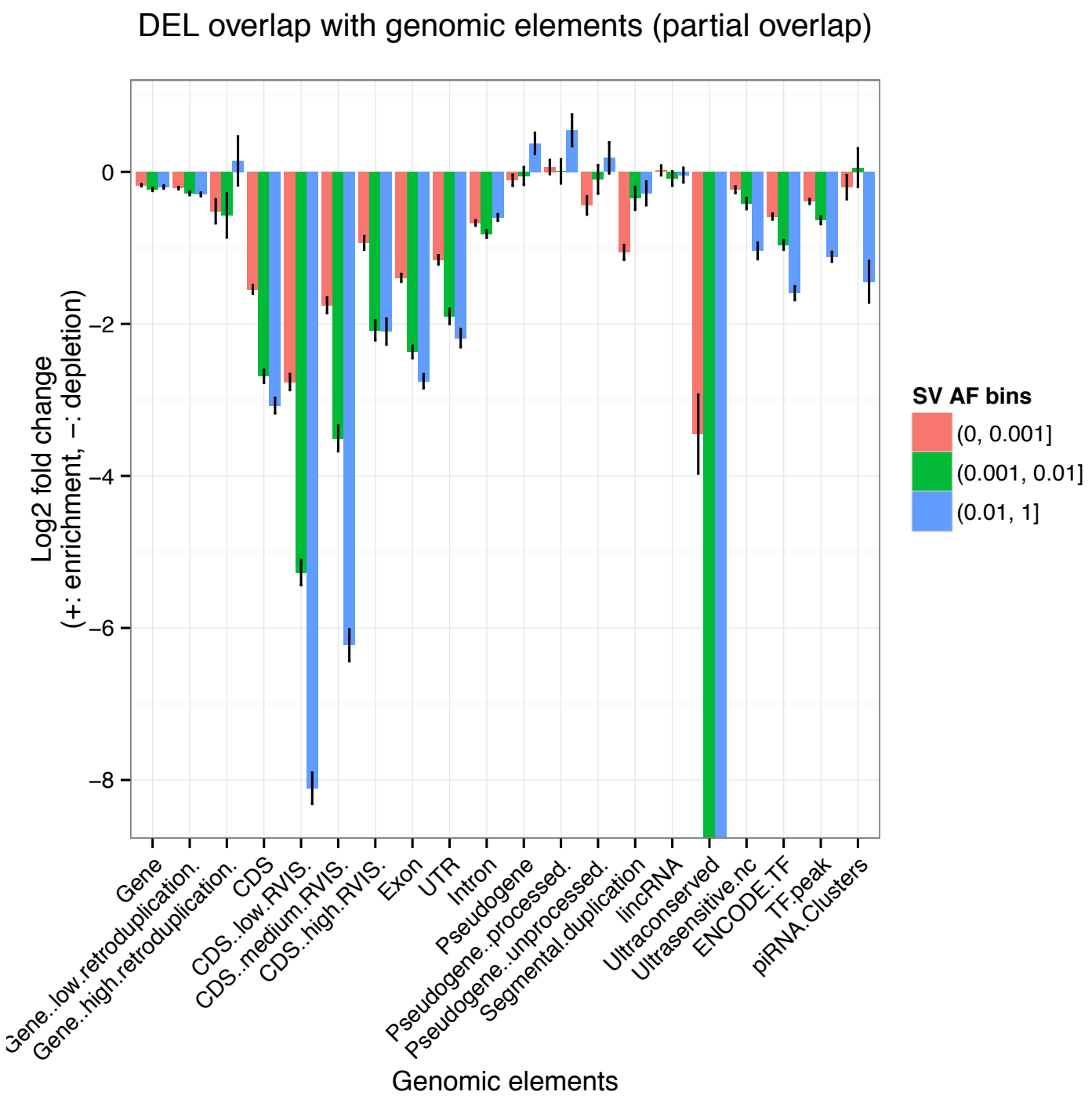
Partial overlap statistic:

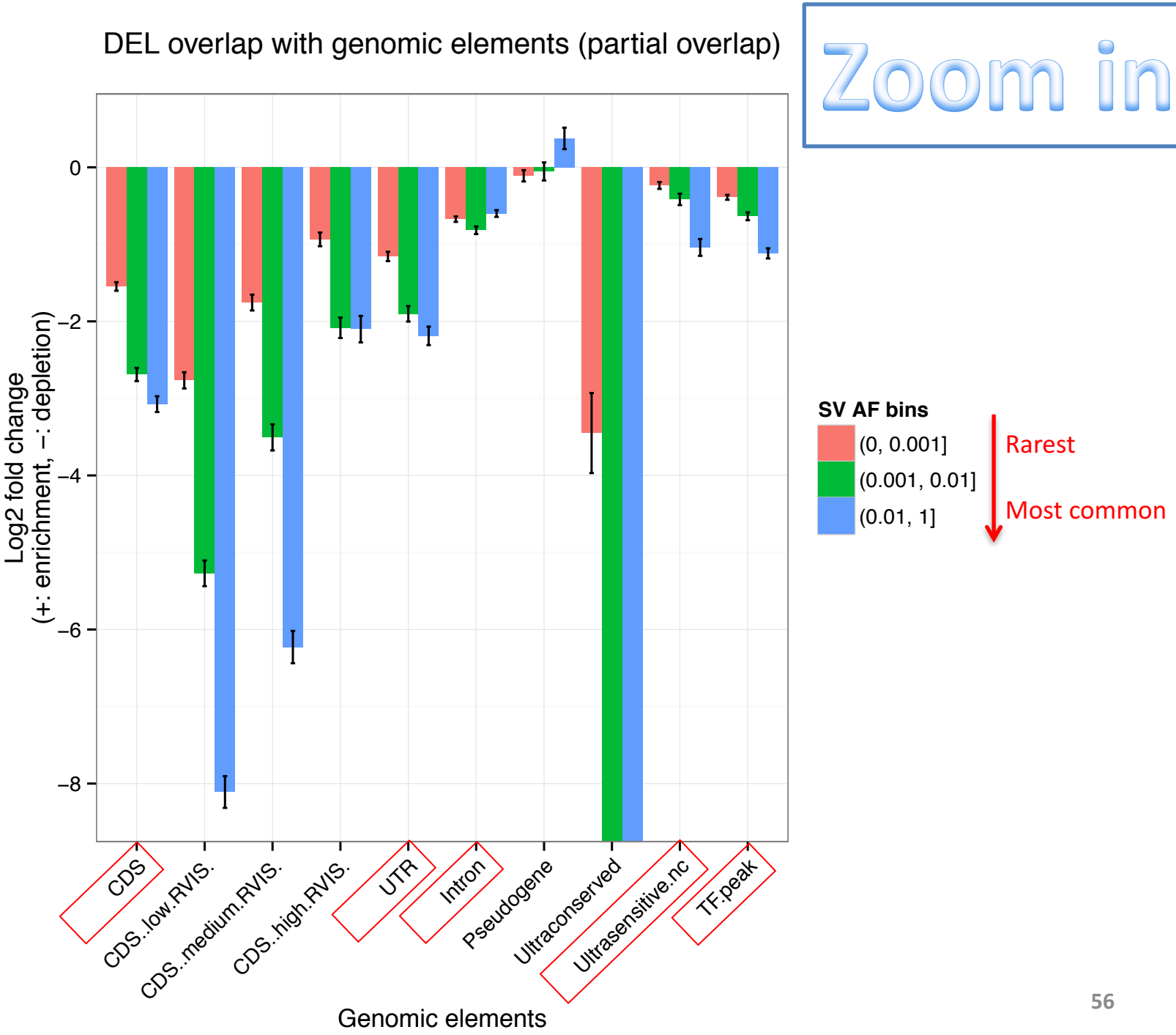
Count the number of genomic elements that have at least 1 bp overlap with SVs.

Permutation Tests

- Permutation scheme
  - Randomly shuffle SV locations while maintaining the local structure
    - Same number of SVs, same length distribution
    - Shuffled SVs still locate on the same chromosome
    - Hg19 gap removed
  - Log2 fold change and empirical p-values

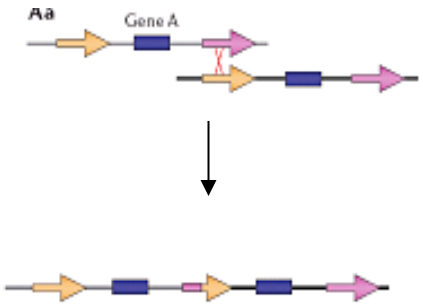
- Datasets
  - 8 types of SVs from the 1000 Genomes Project
  - 20 types of genomic elements from GENCODE, ENCODE, and other literature


DEL overlap with genomic elements (partial overlap)
Log2 fold change
(+: enrichment, −: depletion)
0
−2
−4
−6
−8
Gene
Gene..low.retroduplication.
Gene..high.retroduplication.
CDS
CDS..low.RVIS.
CDS..medium.RVIS.
CDS..high.RVIS.
Exon
UTR
Intron
Pseudogene
Pseudogene..processed.
Pseudogene..unprocessed.
Segmental.duplication
lincRNA
Ultraconserved
Ultrasensitive.nc
ENCODE.TF
TF.peak
piRNA.Clusters
Genomic elements
SV AF bins
(0, 0.001]
(0.001, 0.01]
(0.01, 1]


DEL overlap with genomic elements (partial overlap)
Zoom in
Log2 fold change
(+: enrichment, −: depletion)
0
−2
−4
−6
−8
CDS
CDS..low.RVIS.
CDS..medium.RVIS.
CDS..high.RVIS.
UTR
Intron
Pseudogene
Ultraconserved
Ultrasensitive.nc
TF.peak
Genomic elements
SV AF bins
(0, 0.001]
(0.001, 0.01]
(0.01, 1]
Rarest
Most common

Exact Breakpoints & Mechanism Classification

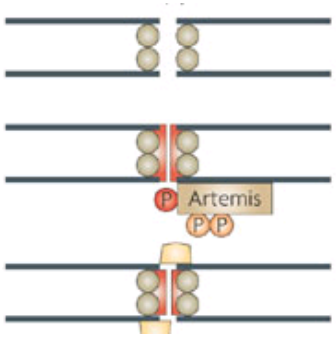
4 mechanisms for SV formation



NAHR

(Non-allelic homologous recombination)

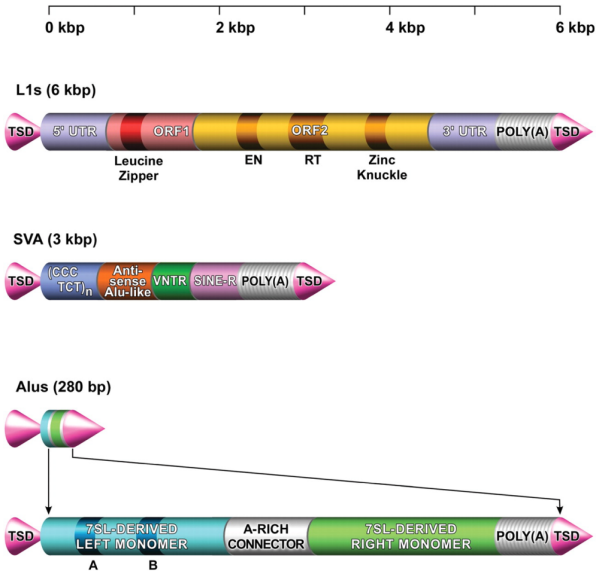
Flanking repeat
(e.g. Alu, LINE…)



NHEJ (NHR)

(Non-homologous-end-joining)

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



TEI

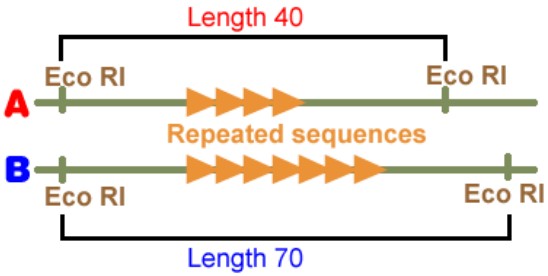
(Transposable element insertion)

L1, SVA, Alus

VNTR

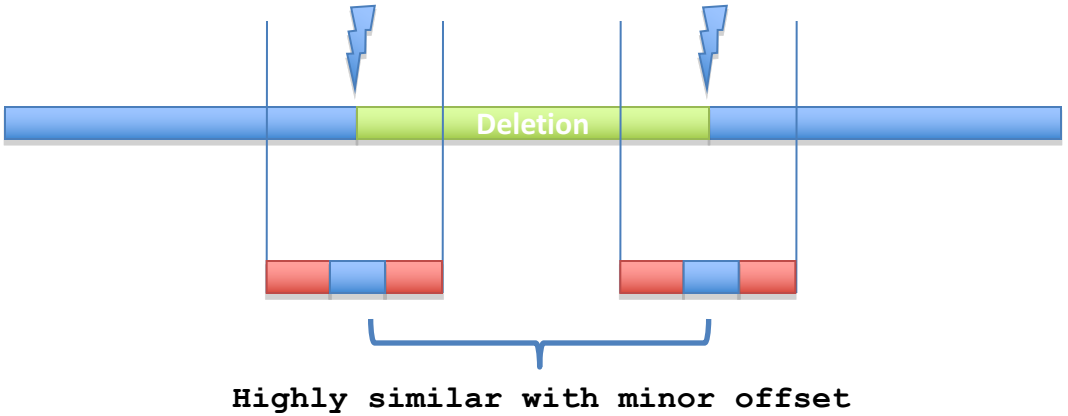
(Variable Number Tandem Repeats)

Number of repeats varies between different people



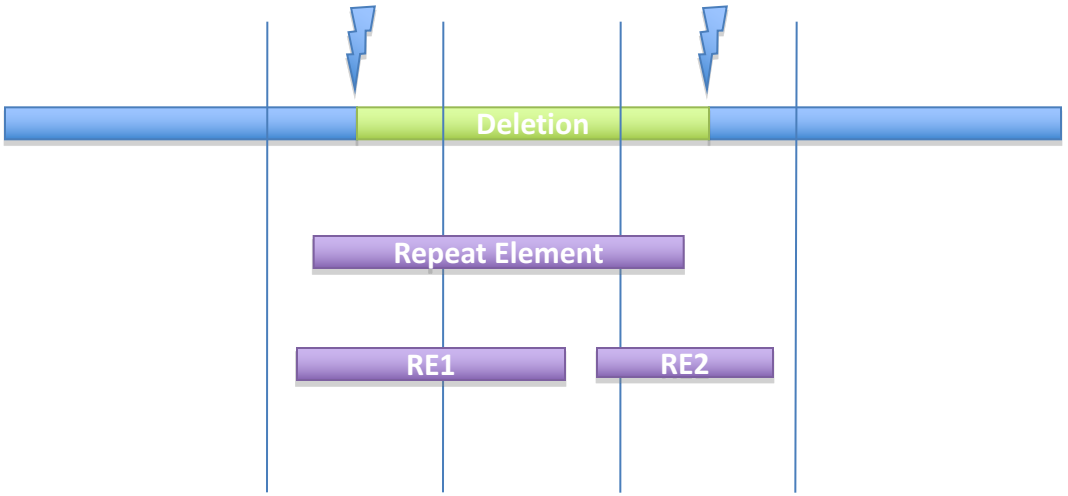
SV Mechanism Classification

NAHR



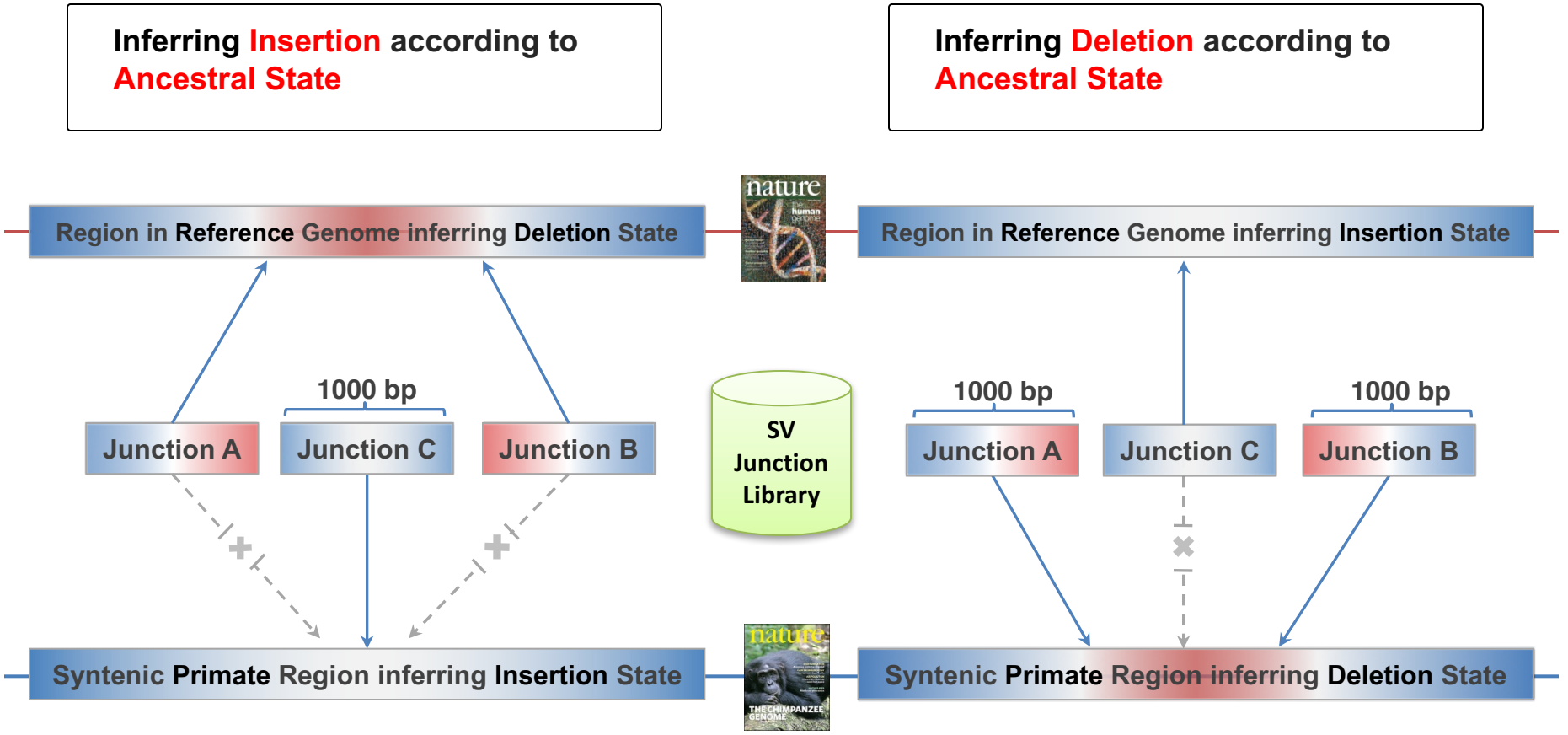
Single RETRO

Multiple RETRO



[Lam et al., ('10) *Nat. Biotech.*]

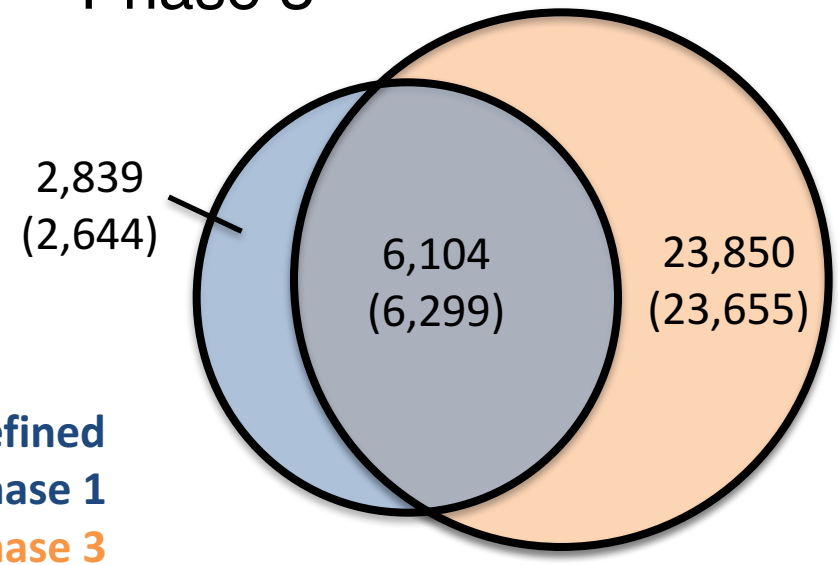
SV Ancestral State Analysis



[Lam et al., ('10) *Nat. Biotech.*]

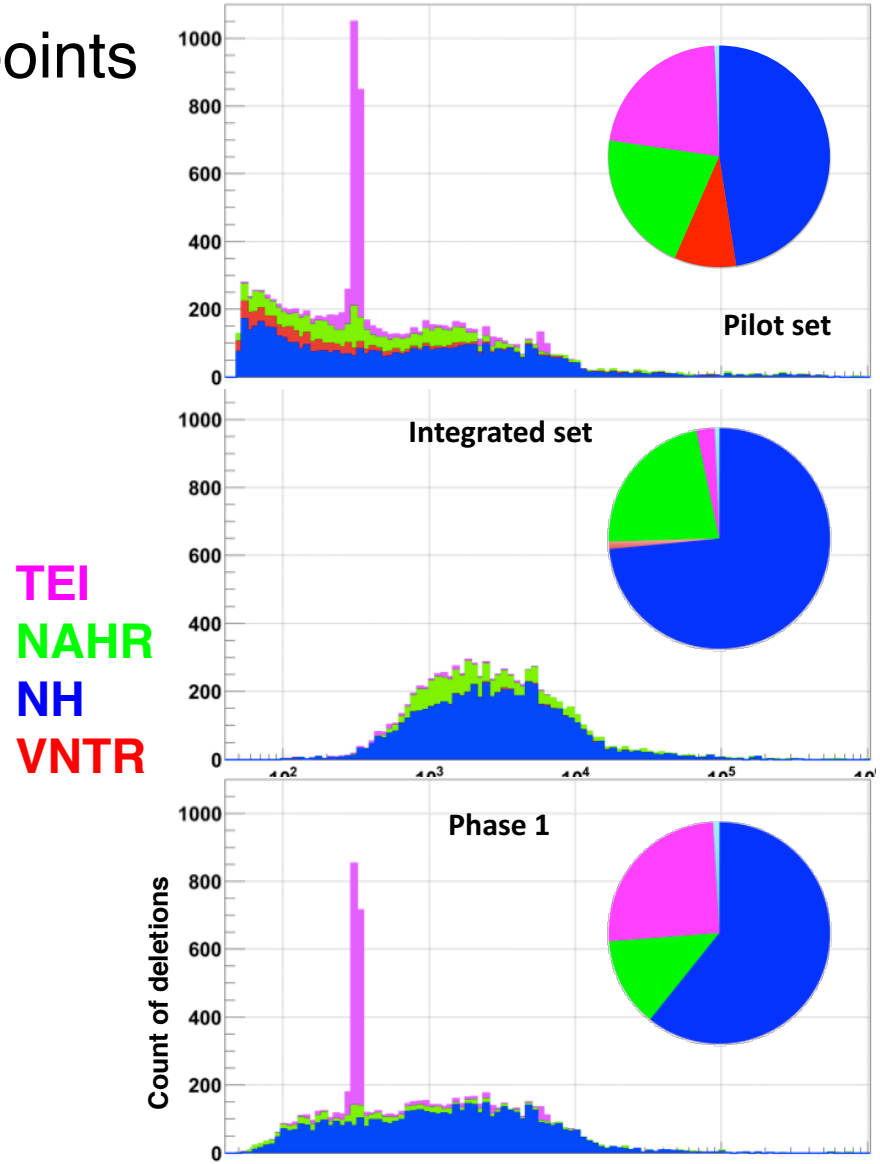
Breakpoint characterization in 1000G

- Breakseq #1 w/ ~2000 breakpoints [Lam et al. Nat. Biotech. (‘10)]
- Pilot
- Phase 1 “Integrated” & Phase 1 refined
- Phase 3



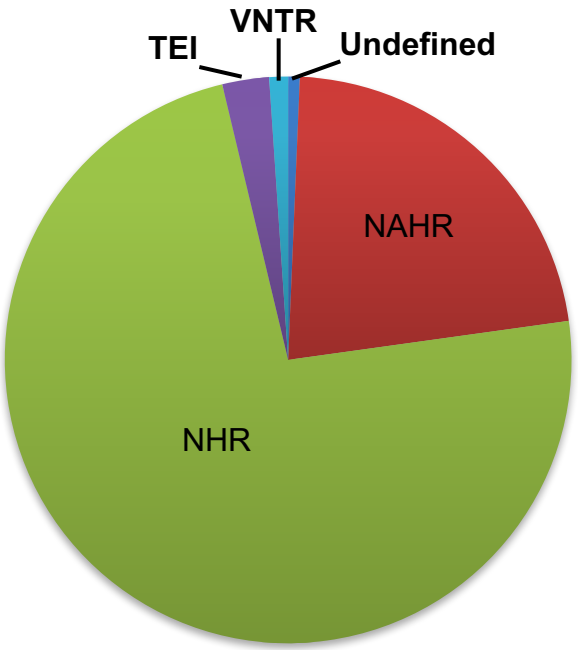
Refined
Phase 1
Phase 3

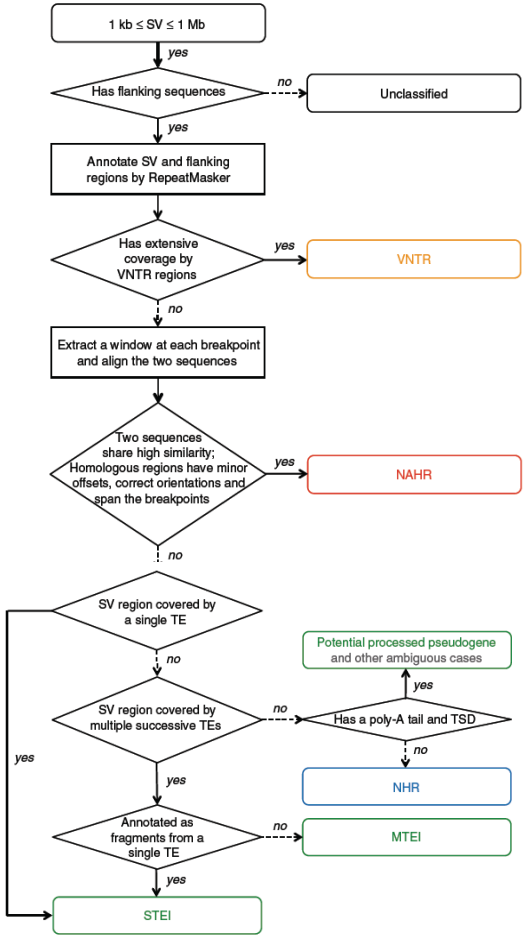
Exact match
Number in parentheses: >50% reciprocal match



[Abyzov et al. (‘15) *Nature Comm.*]

Summary of Mechanism Classification of ~8900 Deletion Breakpoints in 1000G Phase I



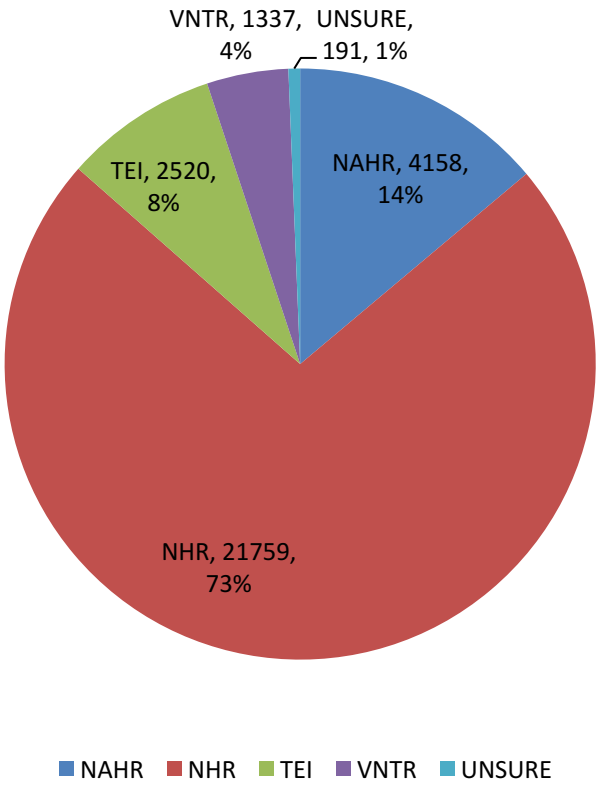


[1000 Genomes Consortium, Nature (2012)]
[Lam et al., ('10) *Nat. Biotech.*]

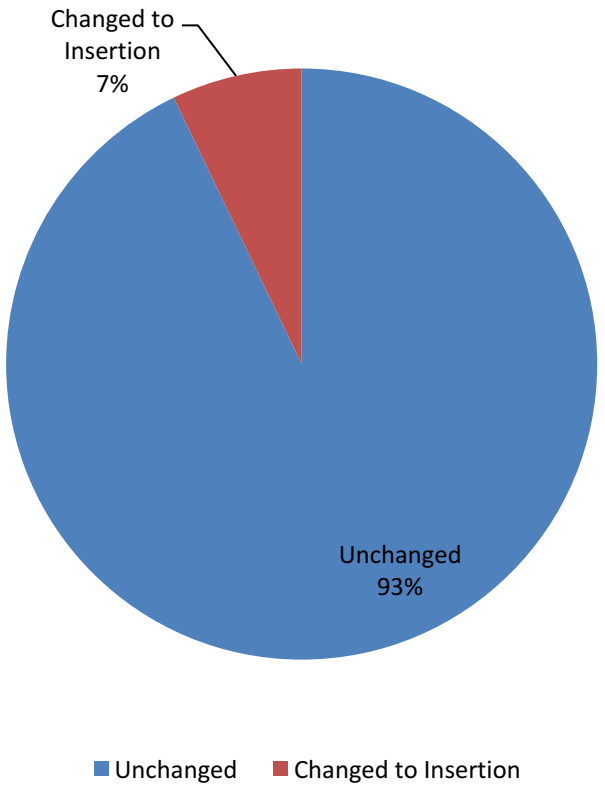
Mechanism	<500 bps	500-1000 bps	1-10 kbps	>10 kbps
NAHR	9 (2.6%)	294 (23.3%)	1420 (22.6%)	255 (24.7%)
NHR	284 (82.8%)	889 (70.4%)	4642 (73.7%)	748 (72.4%)
MEI	47 (13.7%)	67 (5.3%)	124 (2.0%)	0 (0%)
VNTR	2 (0.6%)	7 (0.6%)	64 (1.0%)	23 (2.2%)
Undefined	1 (0.3%)	6 (0.5%)	45 (0.7%)	7 (0.7%)
Total	343 (100%)	1263 (100%)	6295 (100%)	1033 (100%)

BreakSeq Annotation

Formation Mechanisms



Ancestral States



Remarks: There are 79 STEI_NAH events, i.e. 79 events were changed from NAHR to STEI based on our new criteria in the enhanced BreakSeq. Extended annotations from BreakSeq such as NAHR_EXT, STEI_NAH, etc are grouped into their corresponding mechanisms in the above.

Hugo Lam

References

- **Depth-of-coverage**

  CNVnator (Abyzov et al., 2011)

- **Paired-end mapping**

  PEMer (Korbel et al., 2009): For discovery of CNVs and inversions; could also be implemented for translocations

  Breakdancer (Chen et al., 2009): For discovery of CNVs, inversions, and translocations

  GenomeSTRiP (Broad institute): whole-genome, integrating read depth, paired end; population level feature

- **Programs for analysis of longer reads that directly sequence breakpoints**

  CREST (Wang et. al., 2011): Detects small and large structural variants by direct sequencing of breakpoints.

  SRiC (Zhang et al., 2011): Similar to CREST

  Algorithm for strobe reads (Ritz et al., 2010)