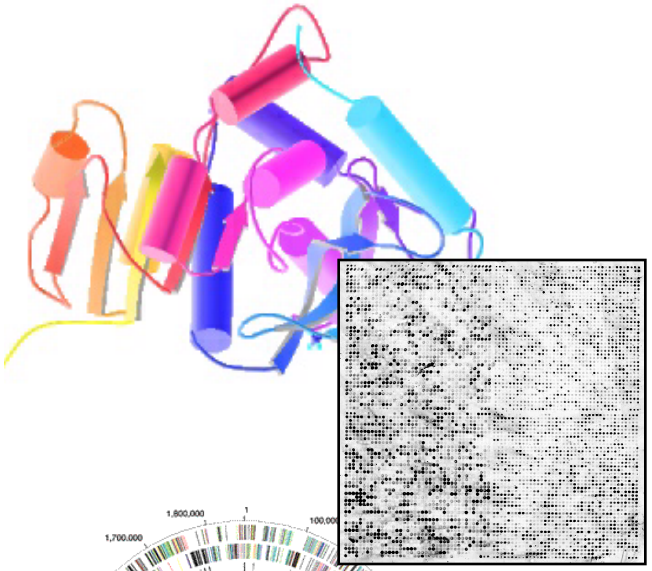
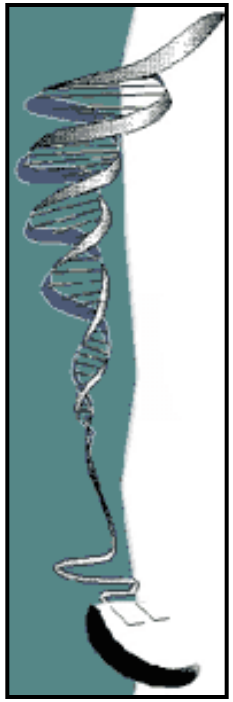
Biomed. Data Sci:

Variant Identification, Focusing on SVs





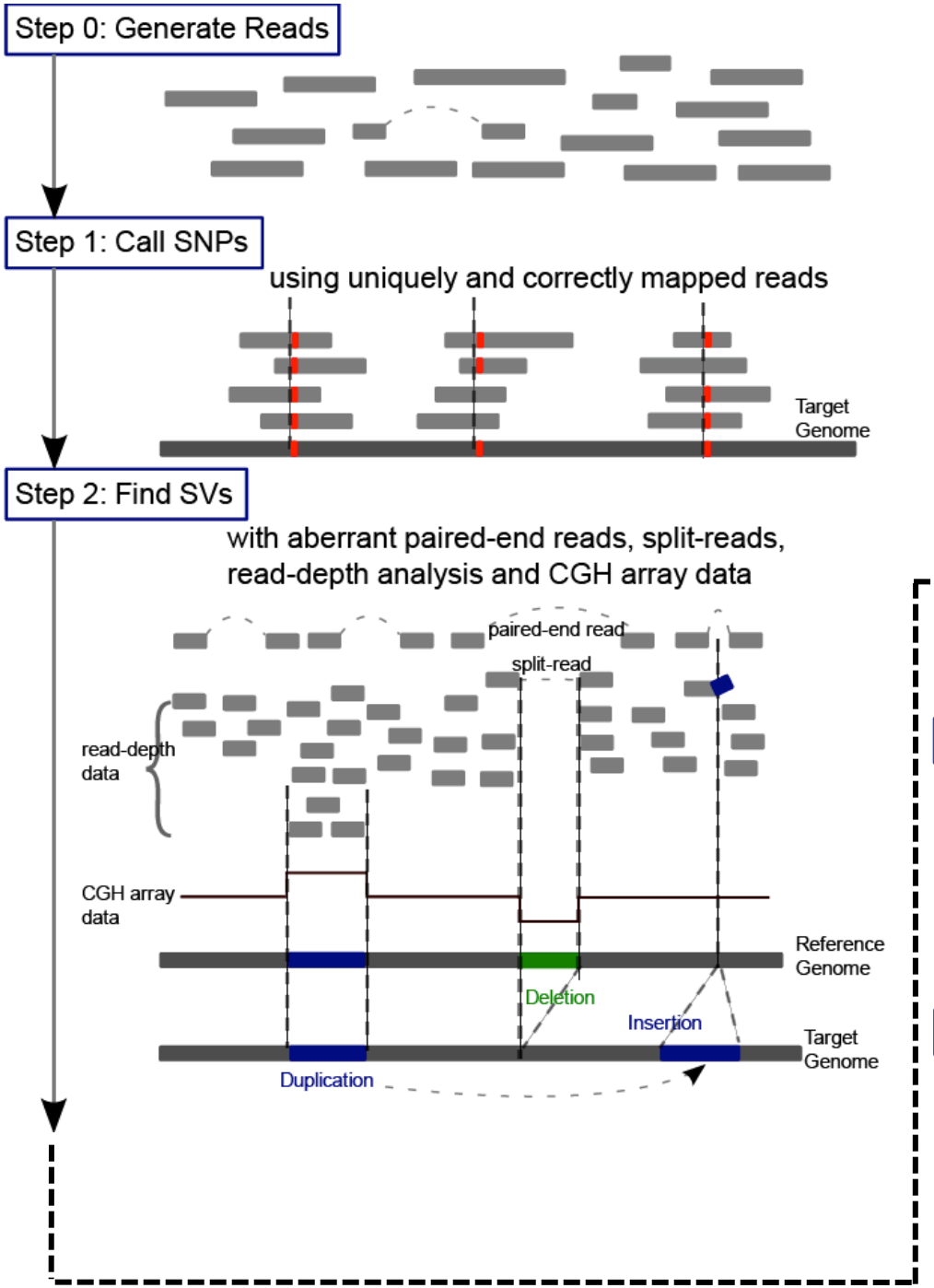
Mark Gerstein, Yale University

gersteinlab.org/courses/452

(last edit in spring ’20, pack #6)

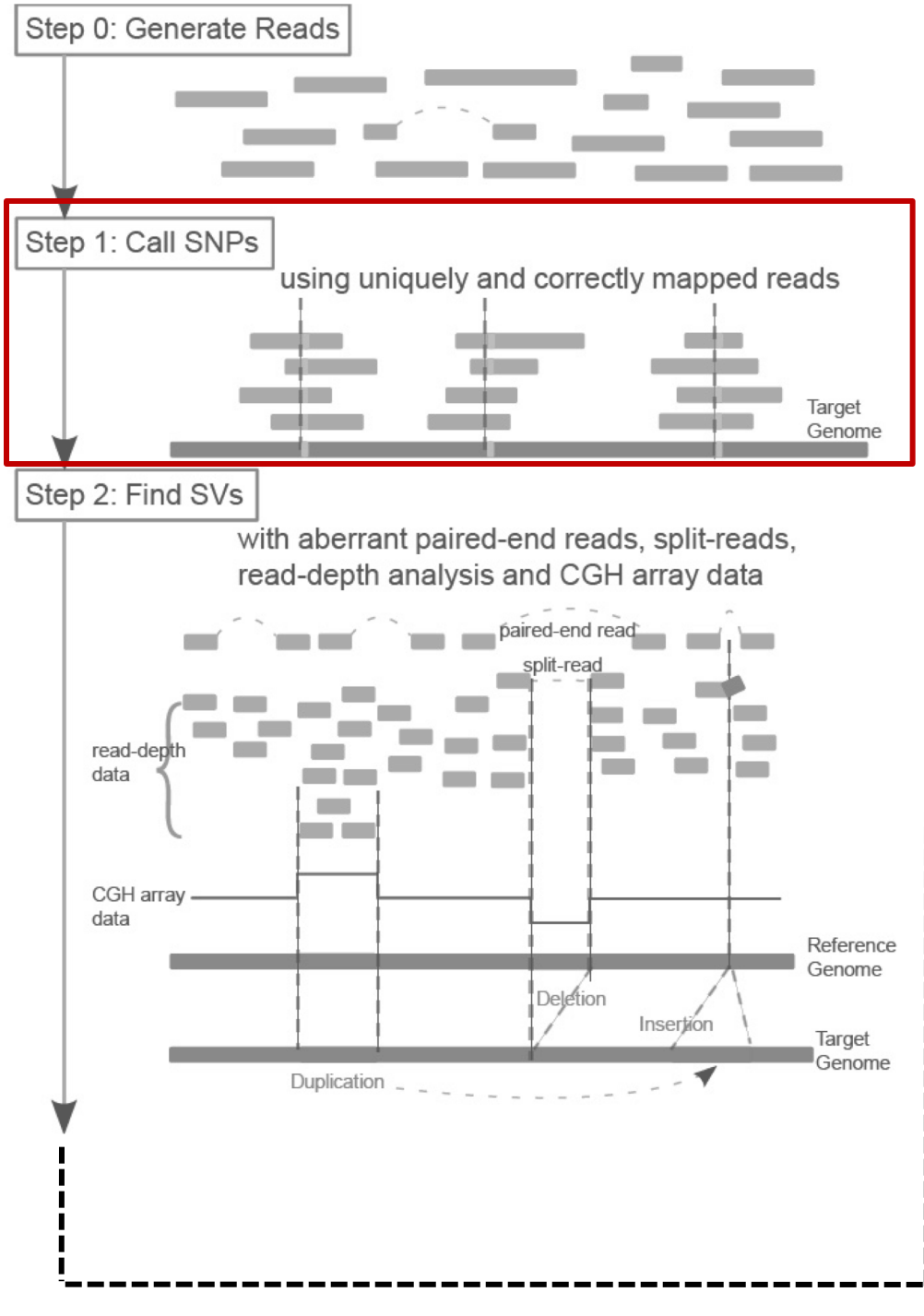
Main Steps in Genome Resequencing

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

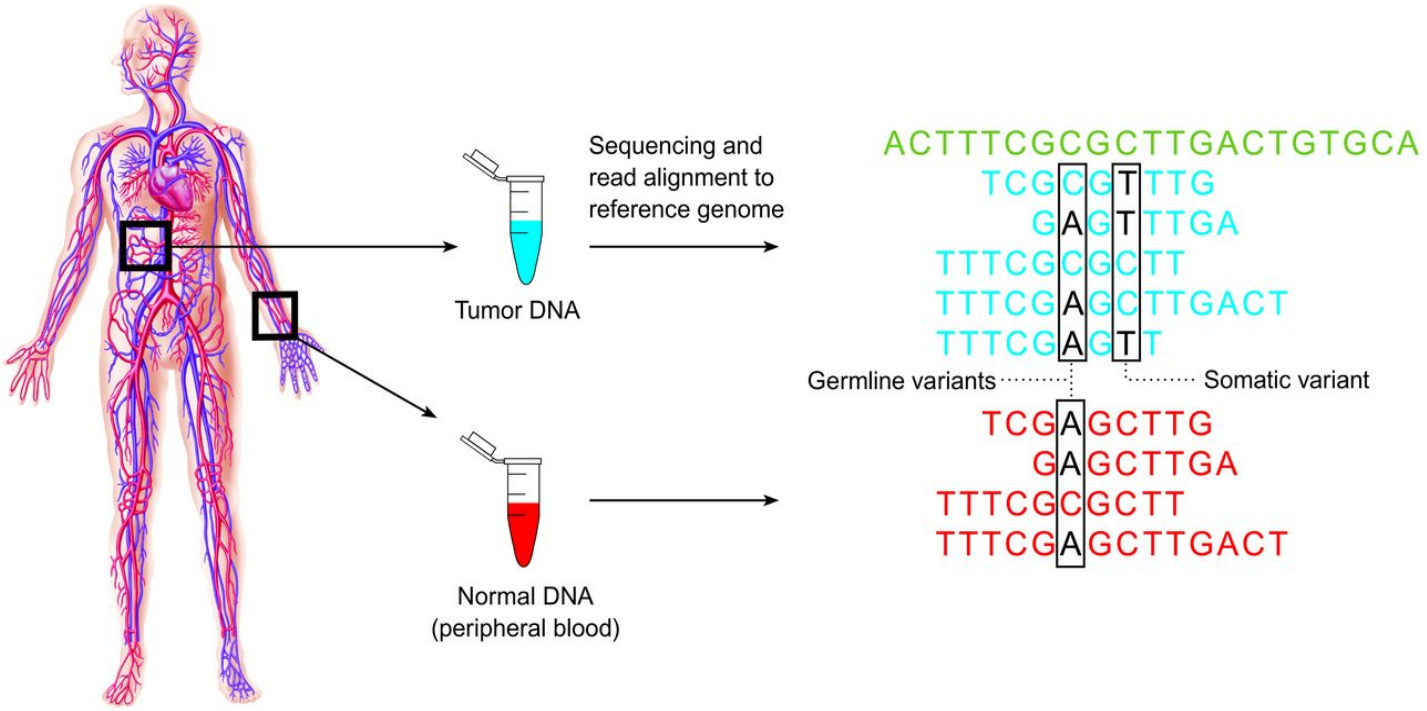


Main Steps in Genome Resequencing

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



Characterization of genomic variations: somatic vs germline



Sequencing tumor and normal samples from cancer patients provide insight into somatic and germline variation profile.

Samuel et al, Clinical chemistry, 2013

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

[You et al. ('12) Bioinformatics 28: 643 doi: 10.1093/bioinformatics/bts001]

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.

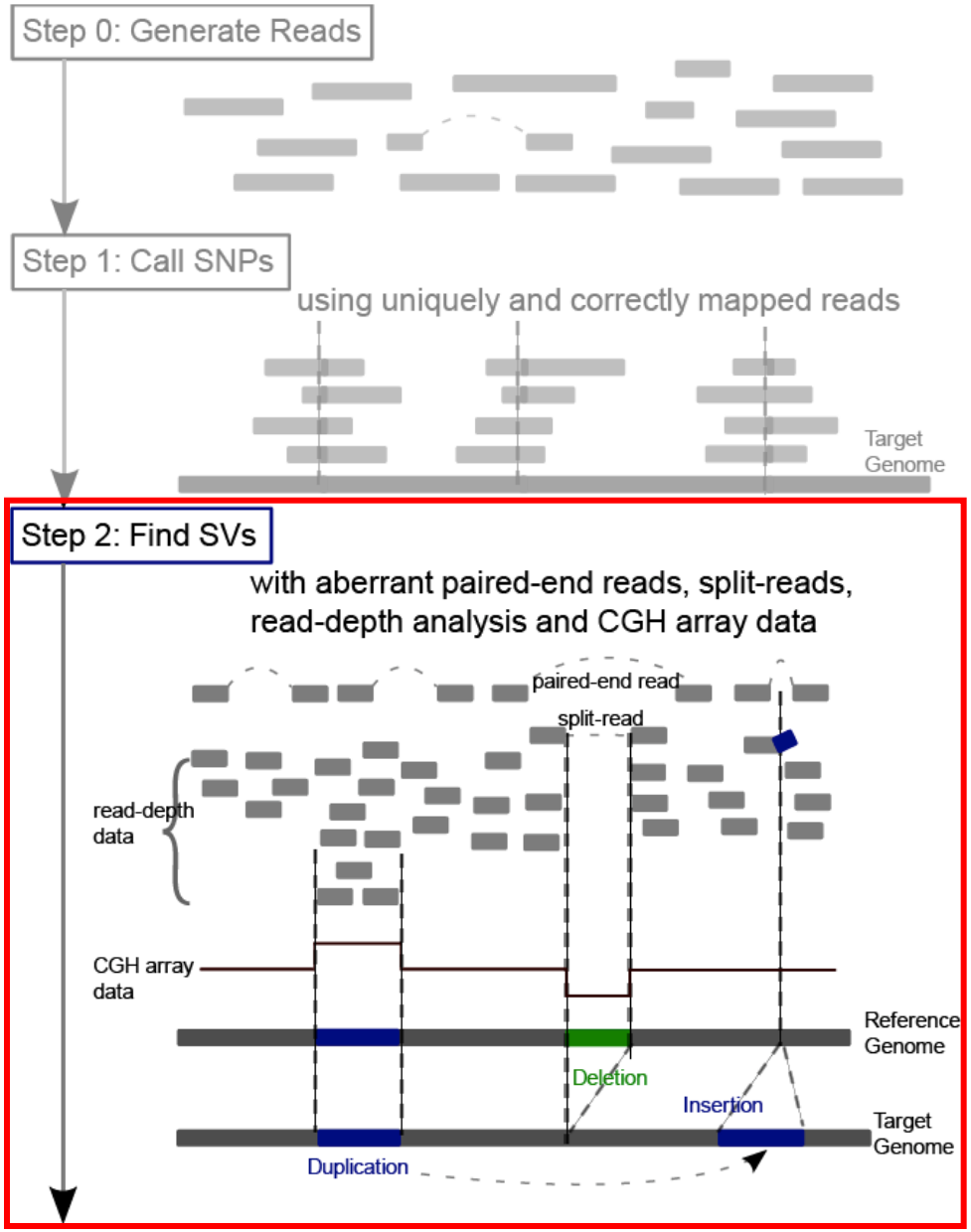
In general:

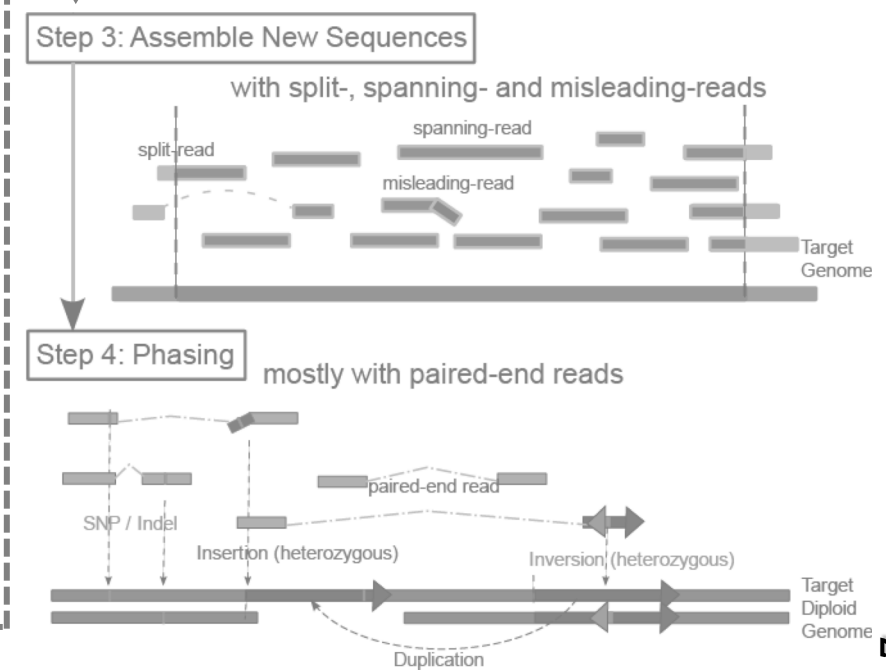
$$\boldsymbol{P(D|G) = P_{err\_free}(D|G) + P_{err}(D|G)}$$

[You et al. ('12) Bioinformatics 28: 643 doi: 10.1093/bioinformatics/bts001]

Main Steps in Genome Resequencing

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

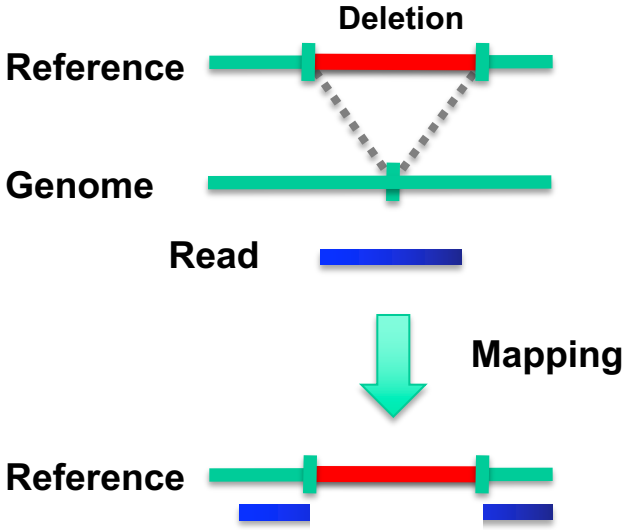




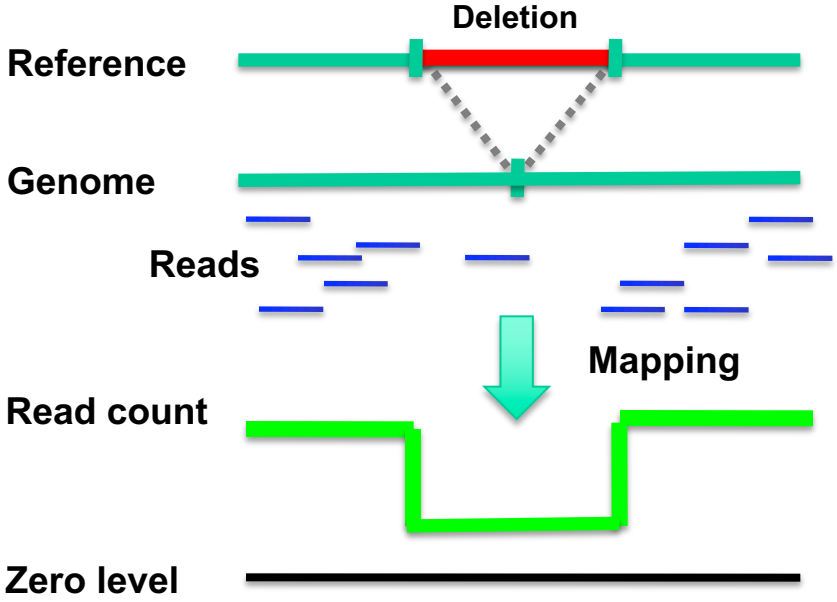
Methods to Find SVs

1. Paired ends

2. Split read



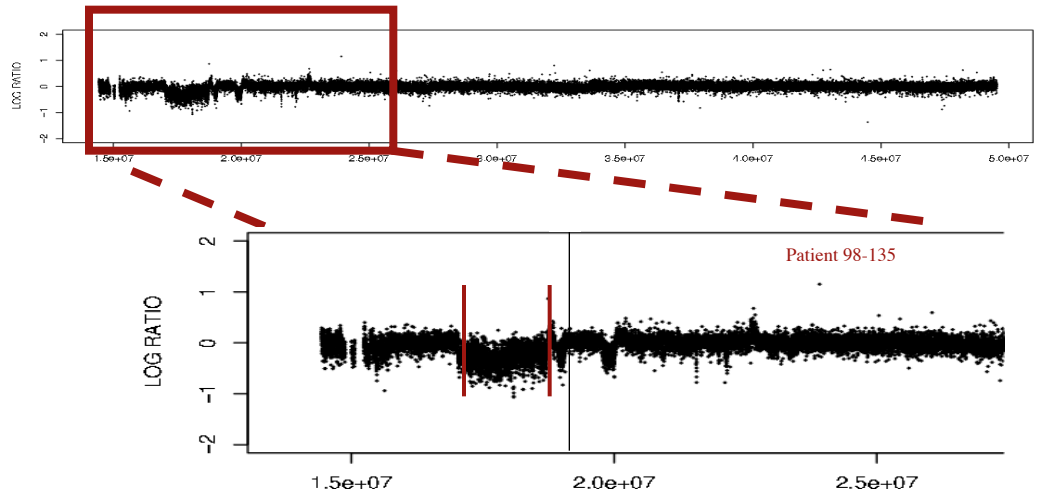
3. Read depth (or aCGH)



4. Local Reassembly

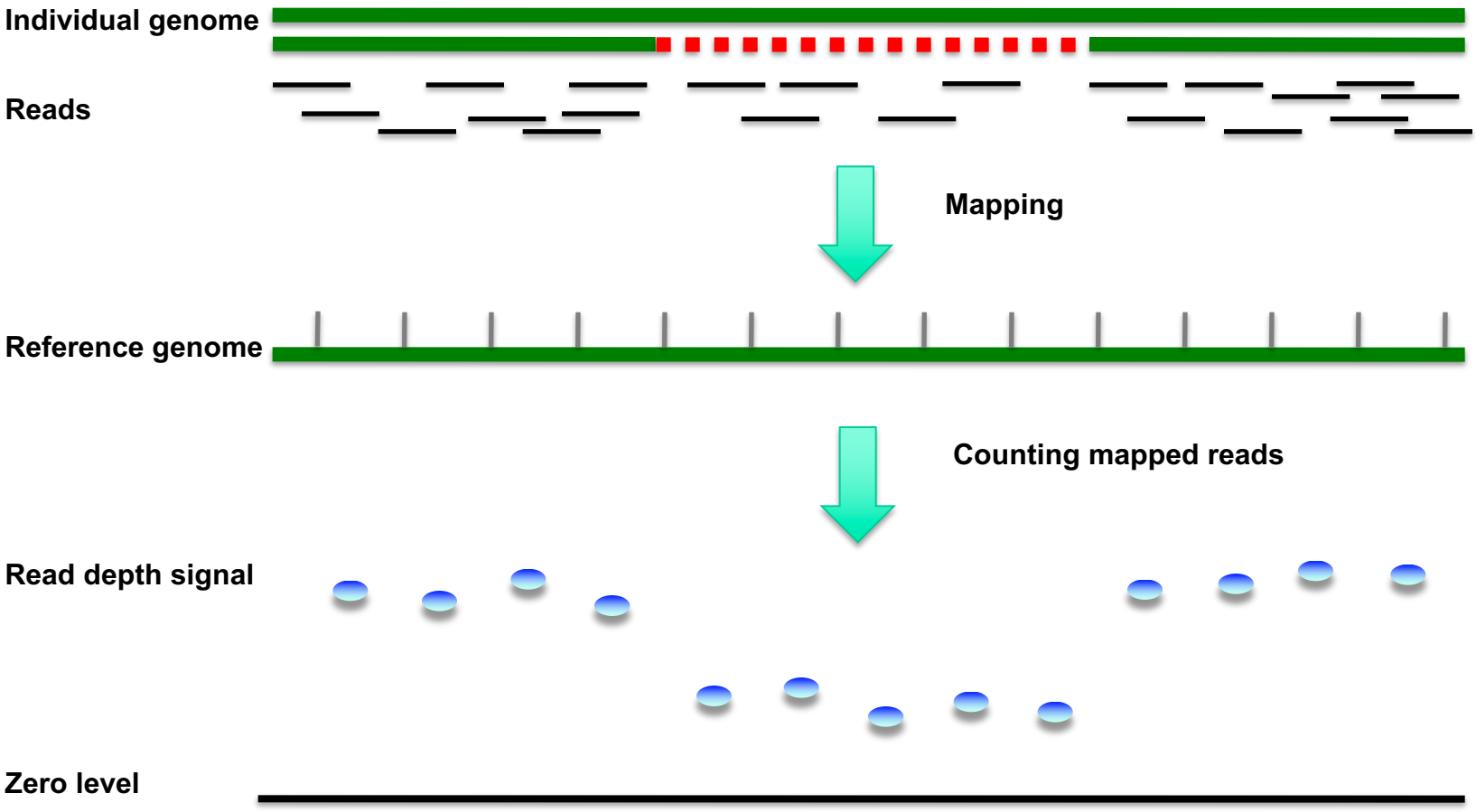
[Snyder et al. Genes & Dev. ('10)]

Read Depth



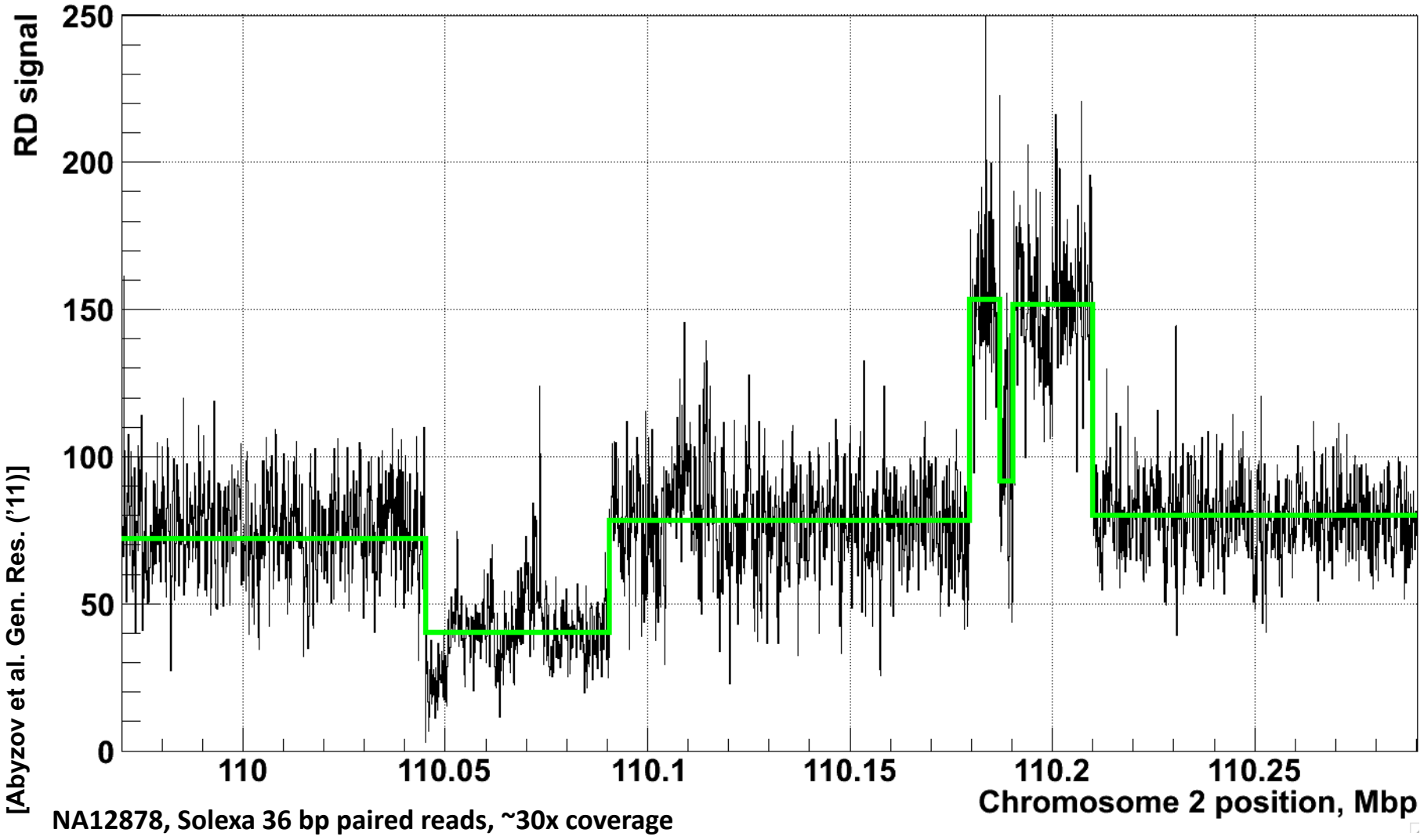
Array Signal

Read depth



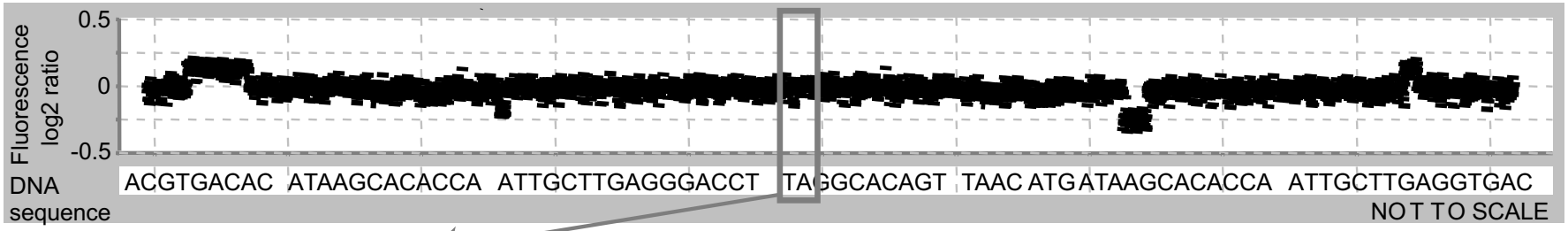
[Urban et al. ('06) PNAS; Wang et al. Gen. Res. ('09); Abyzov et al. Gen. Res. ('11)]

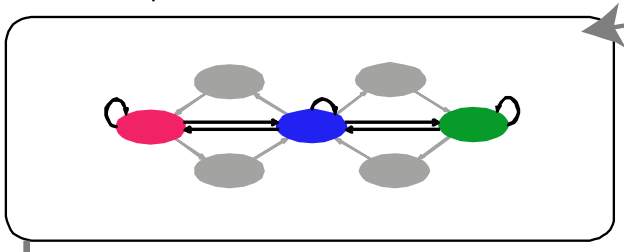
Example of Application to RD data

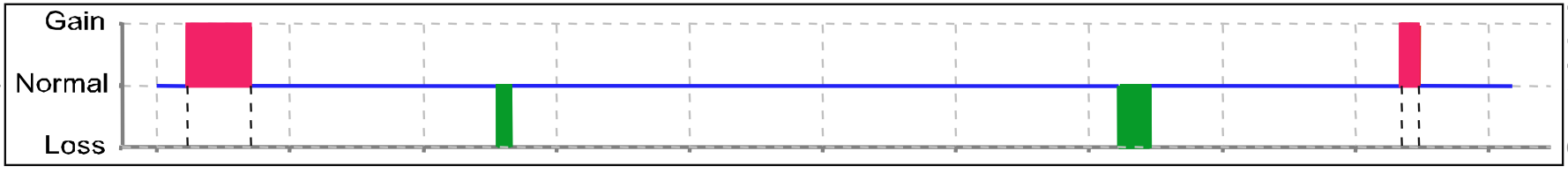


HMM

- 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

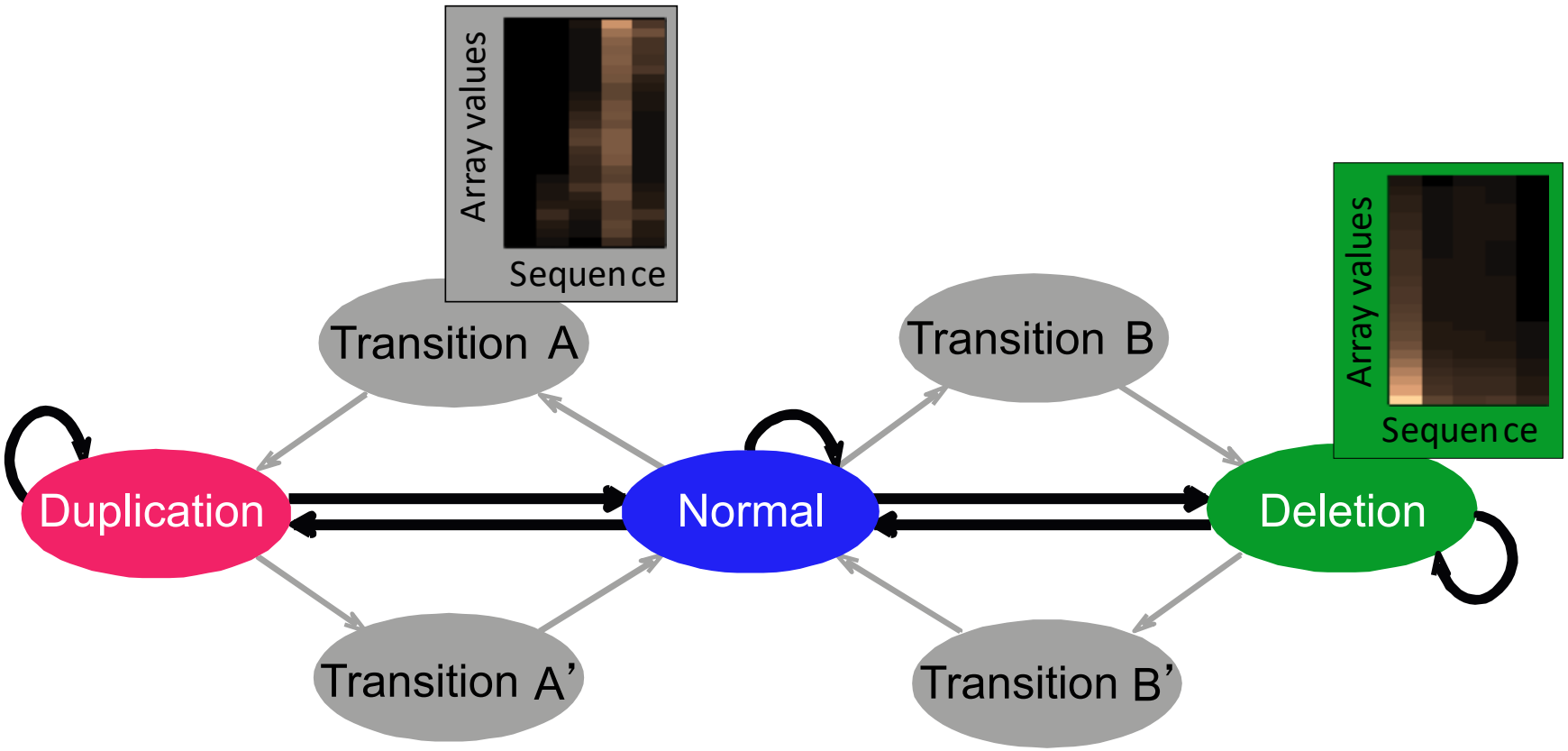






Korbel*, Urban* *et al.*, PNAS (2007)

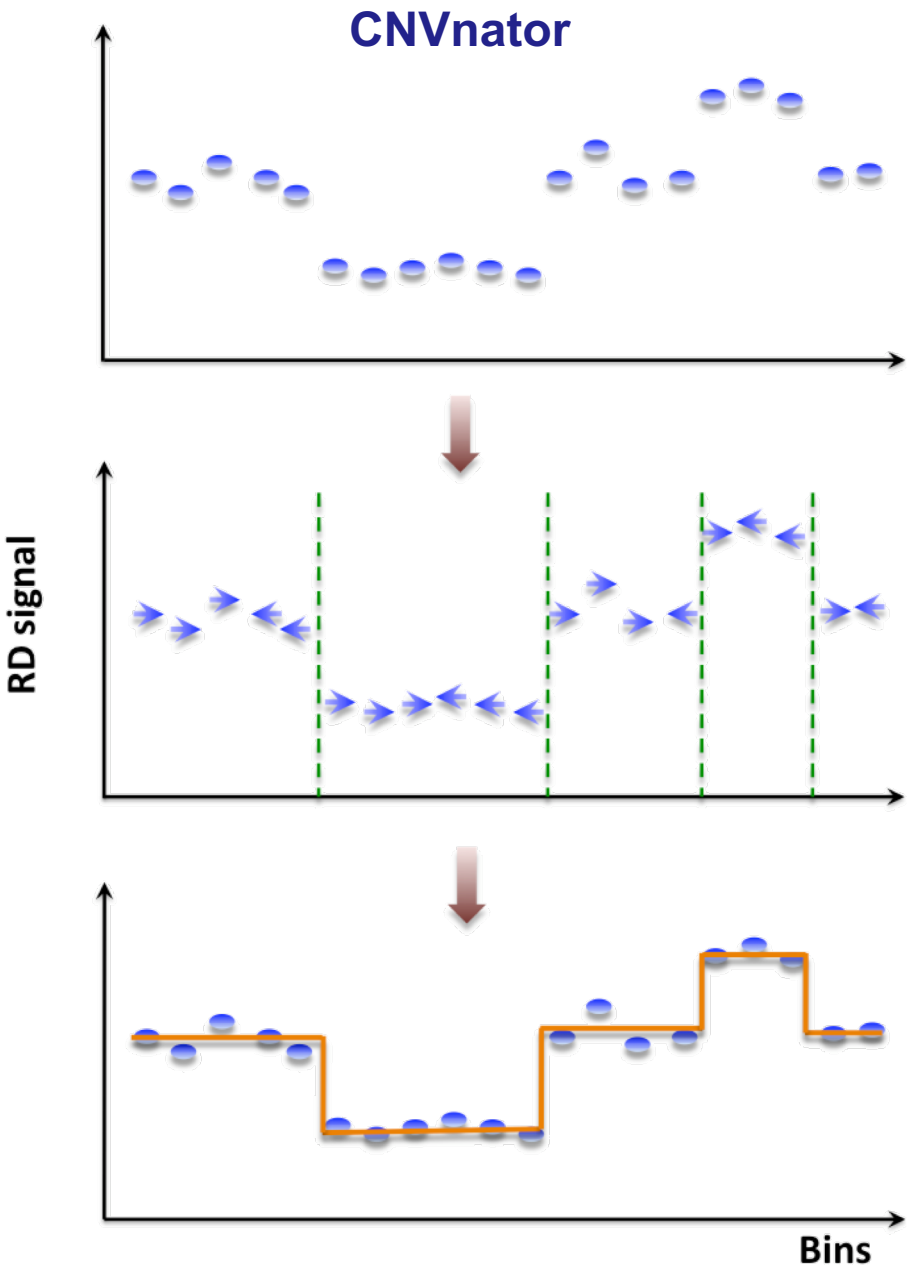
Statistically integrates array signal and DNA sequence signatures (using a discrete-valued bivariate HMM)



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

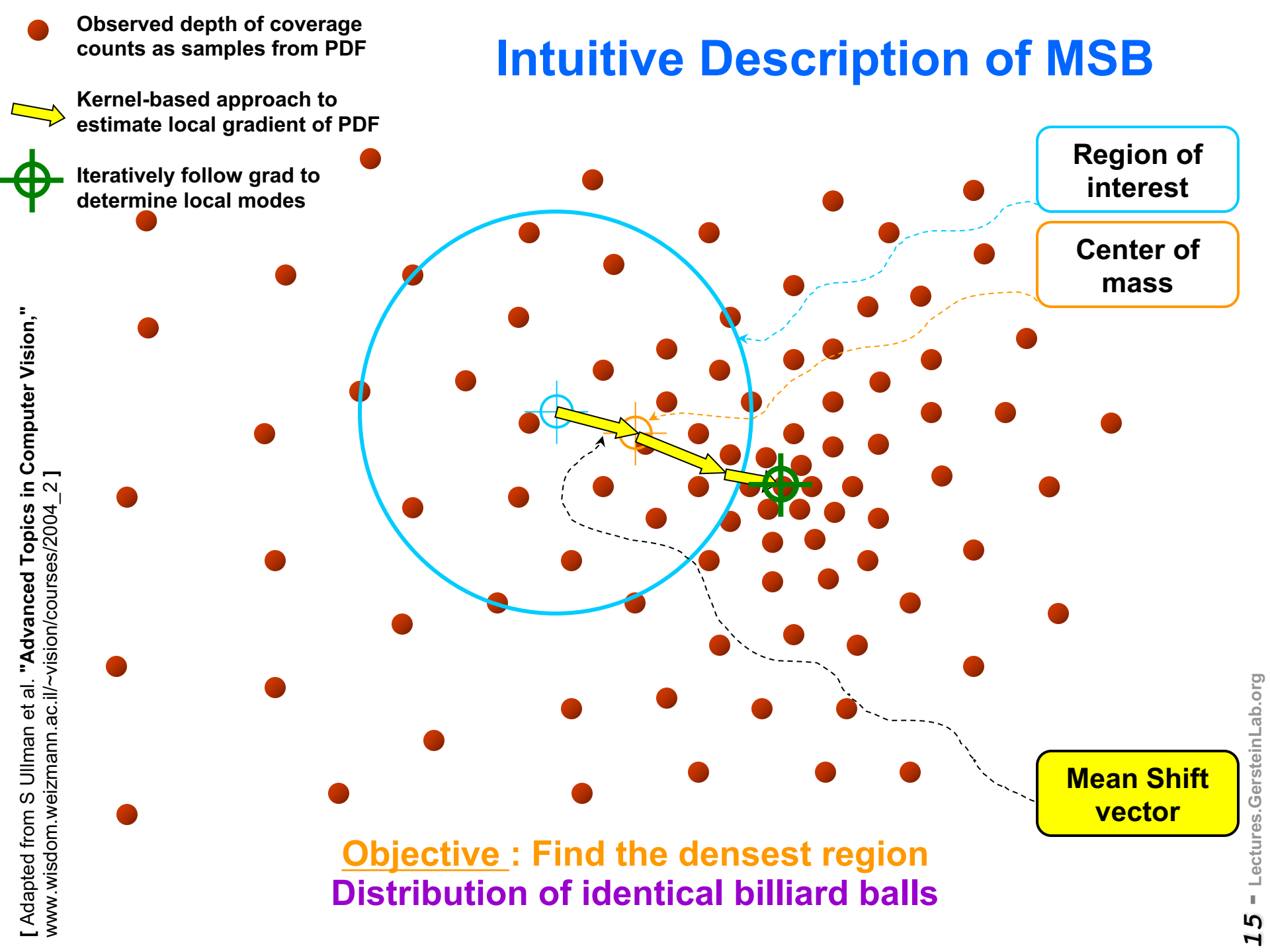
Mean-shift-based (MSB) segmentation: no explicit model

- 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



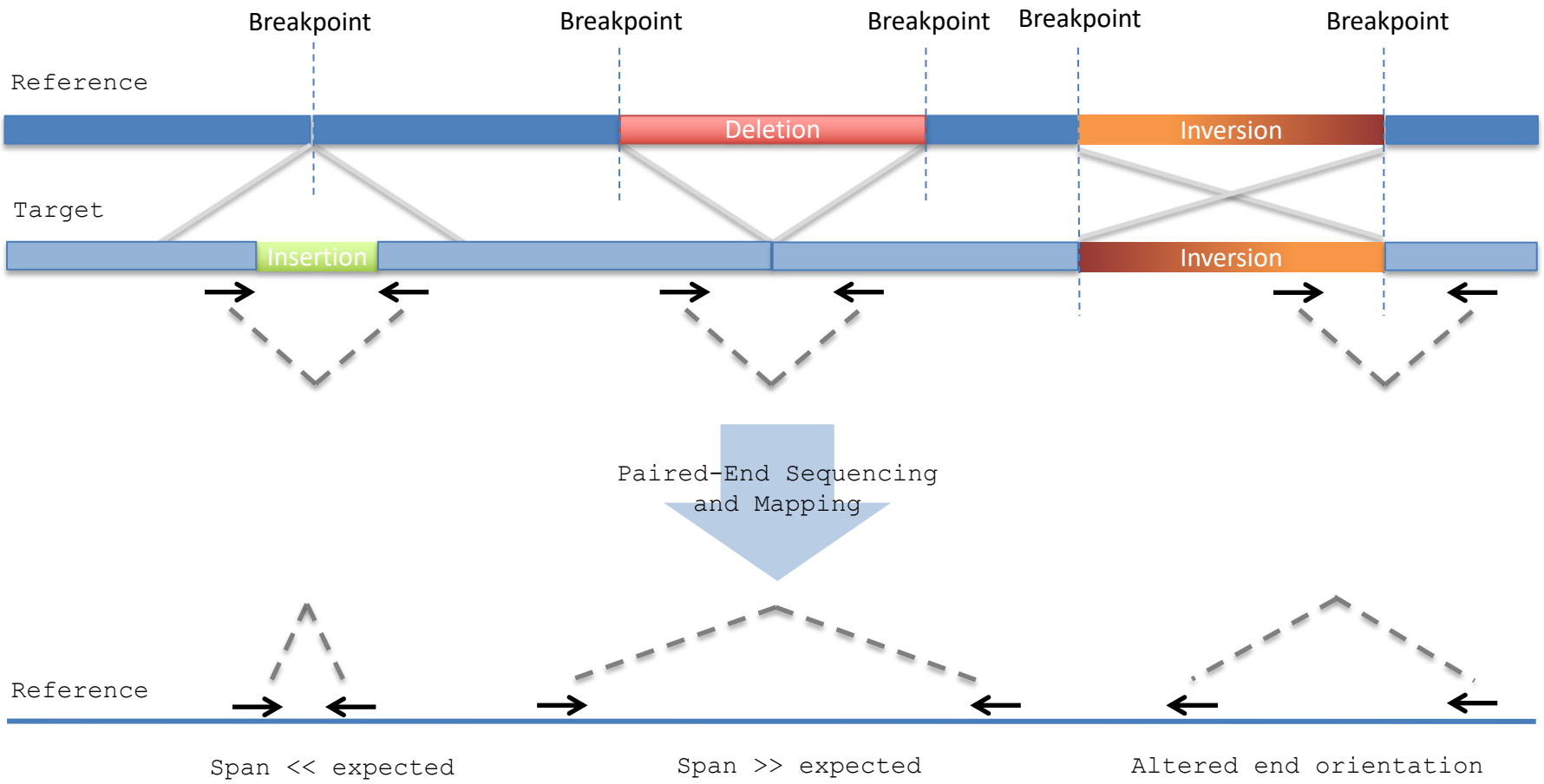
Objective : Find the densest region

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]

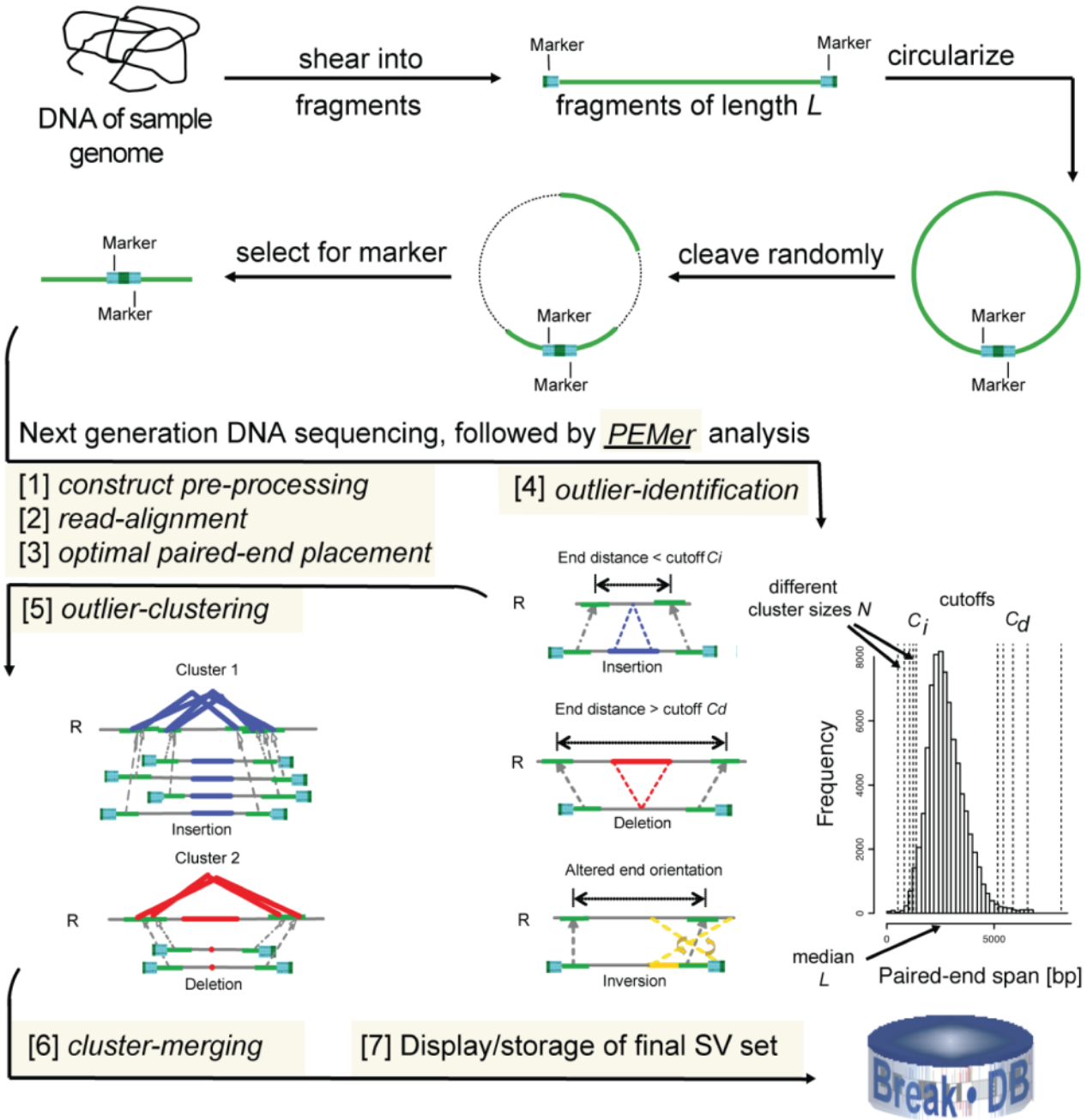
Paired-End

Paired-End Mapping



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

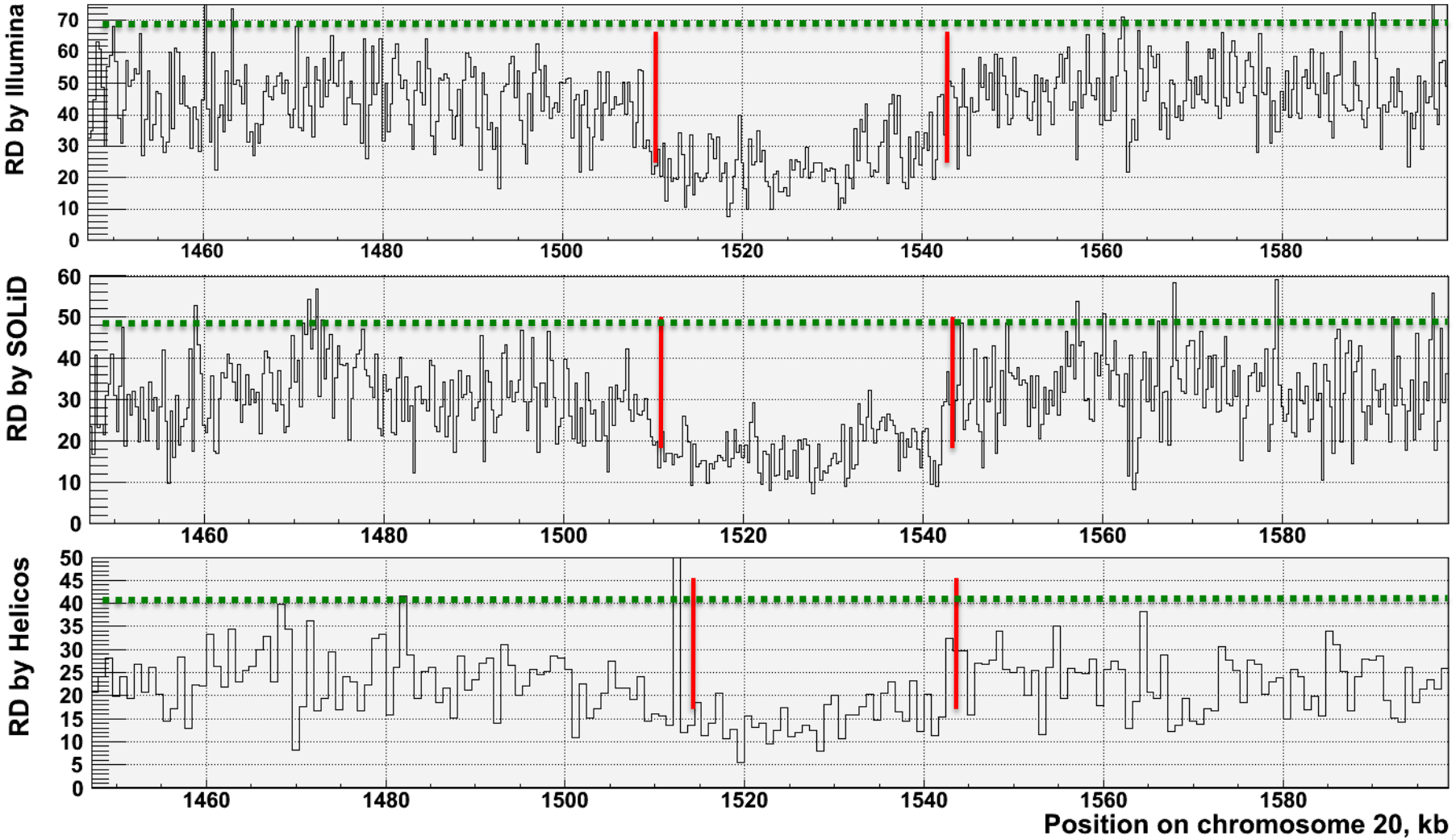
Overall Strategy for Analysis of NextGen Seq. Data to Detect Structural Variants



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

Split Read

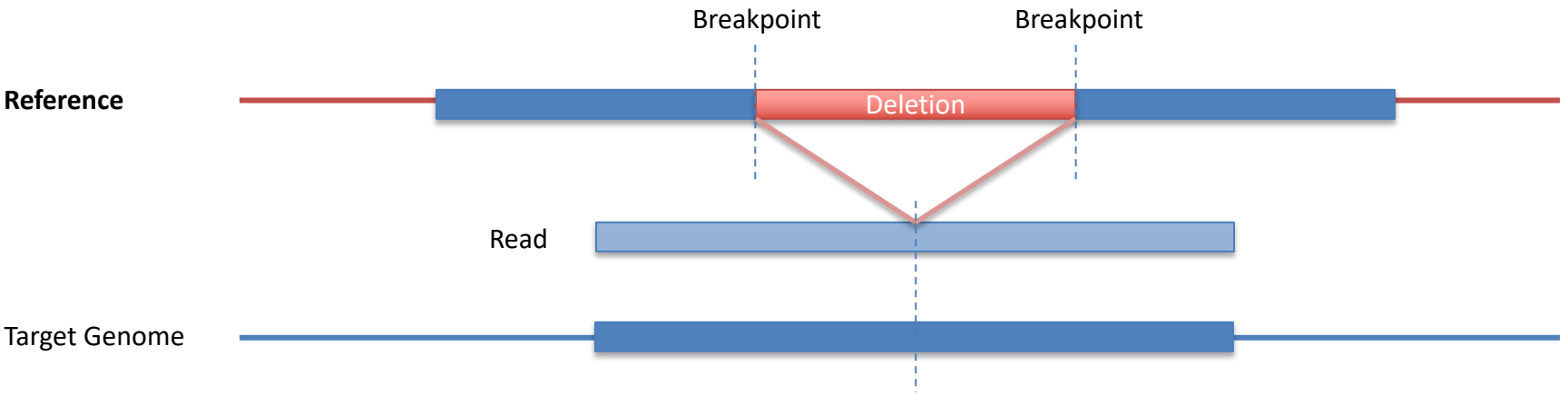
Read-depth works well on a variety of sequencing platforms but provides imprecise breakpoints

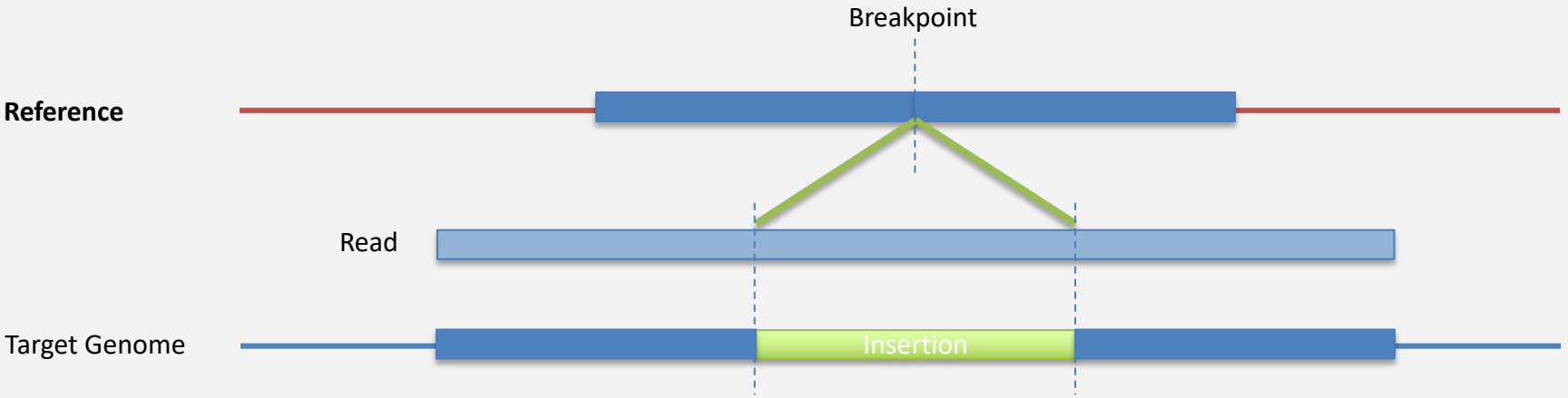


[Abyzov et al. Gen. Res. ('11)]

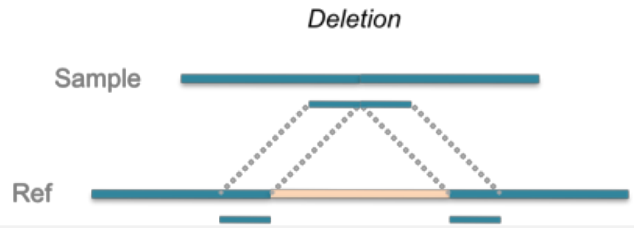
[NA18505]

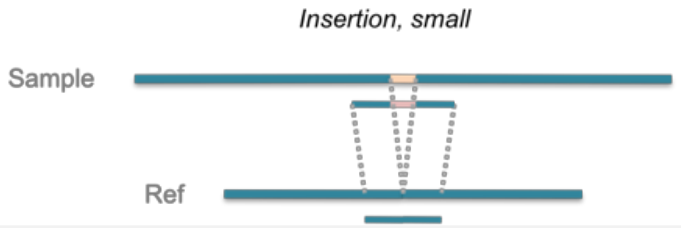
Split-read Analysis

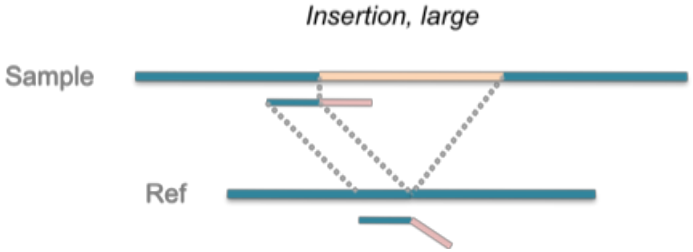




Simple SVs

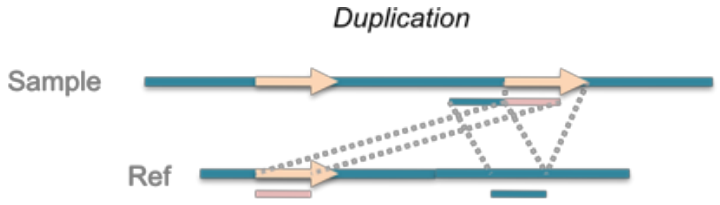


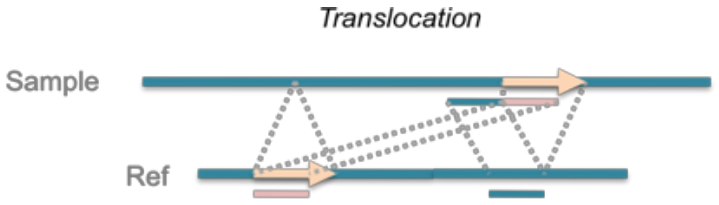


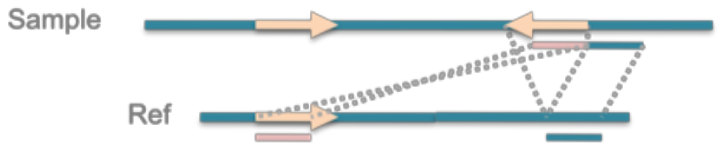


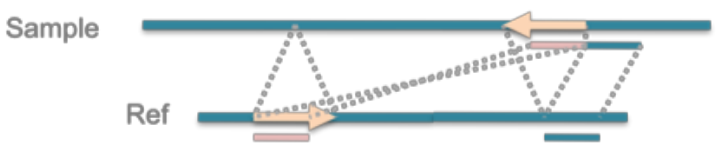
Deletions are the Easiest to Identify

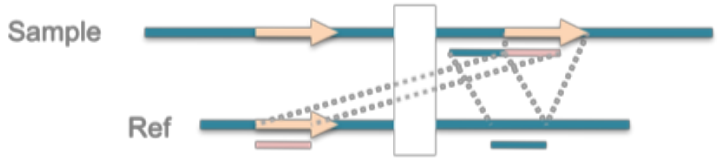
Complex SVs

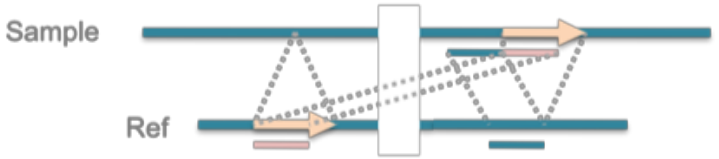








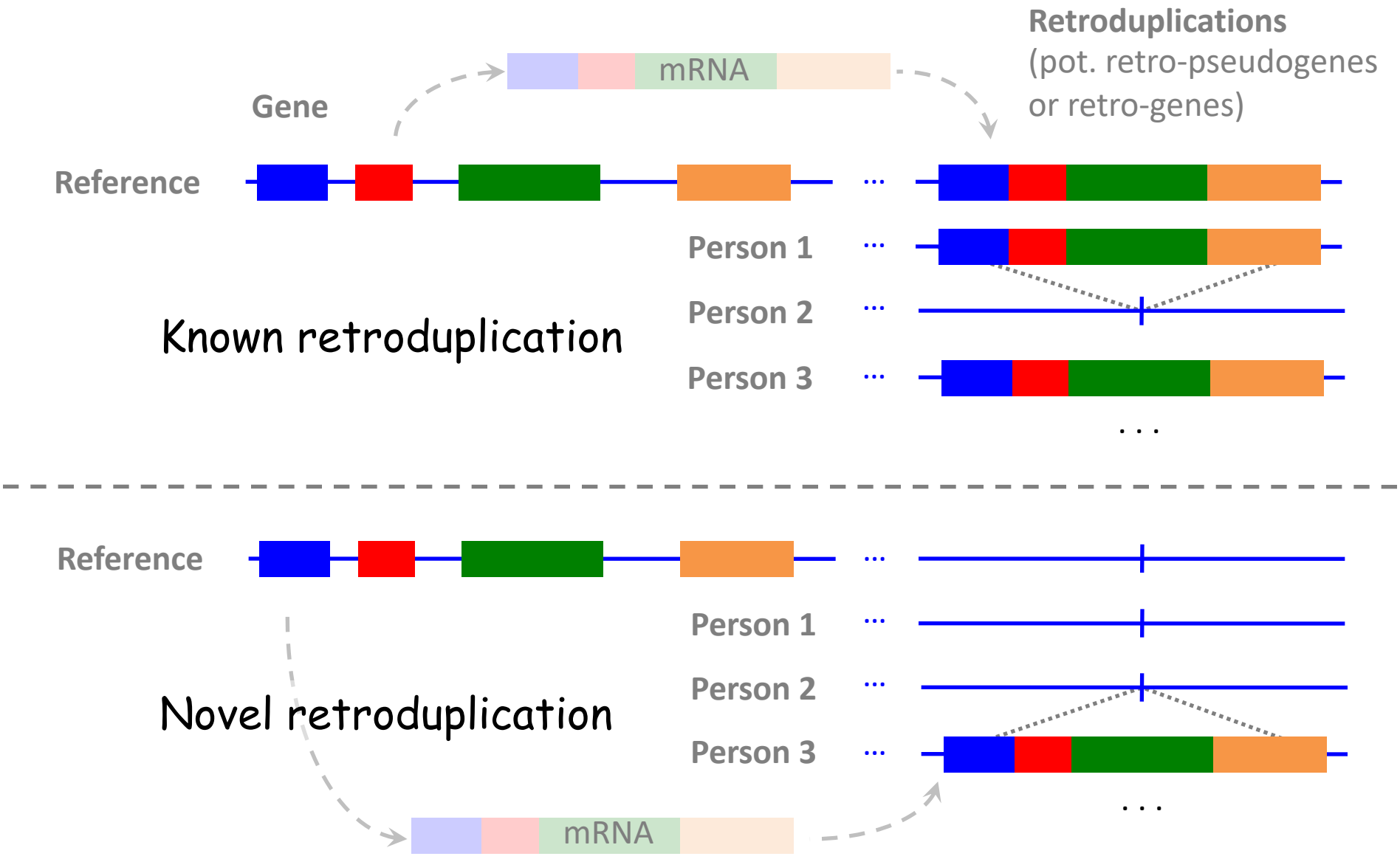




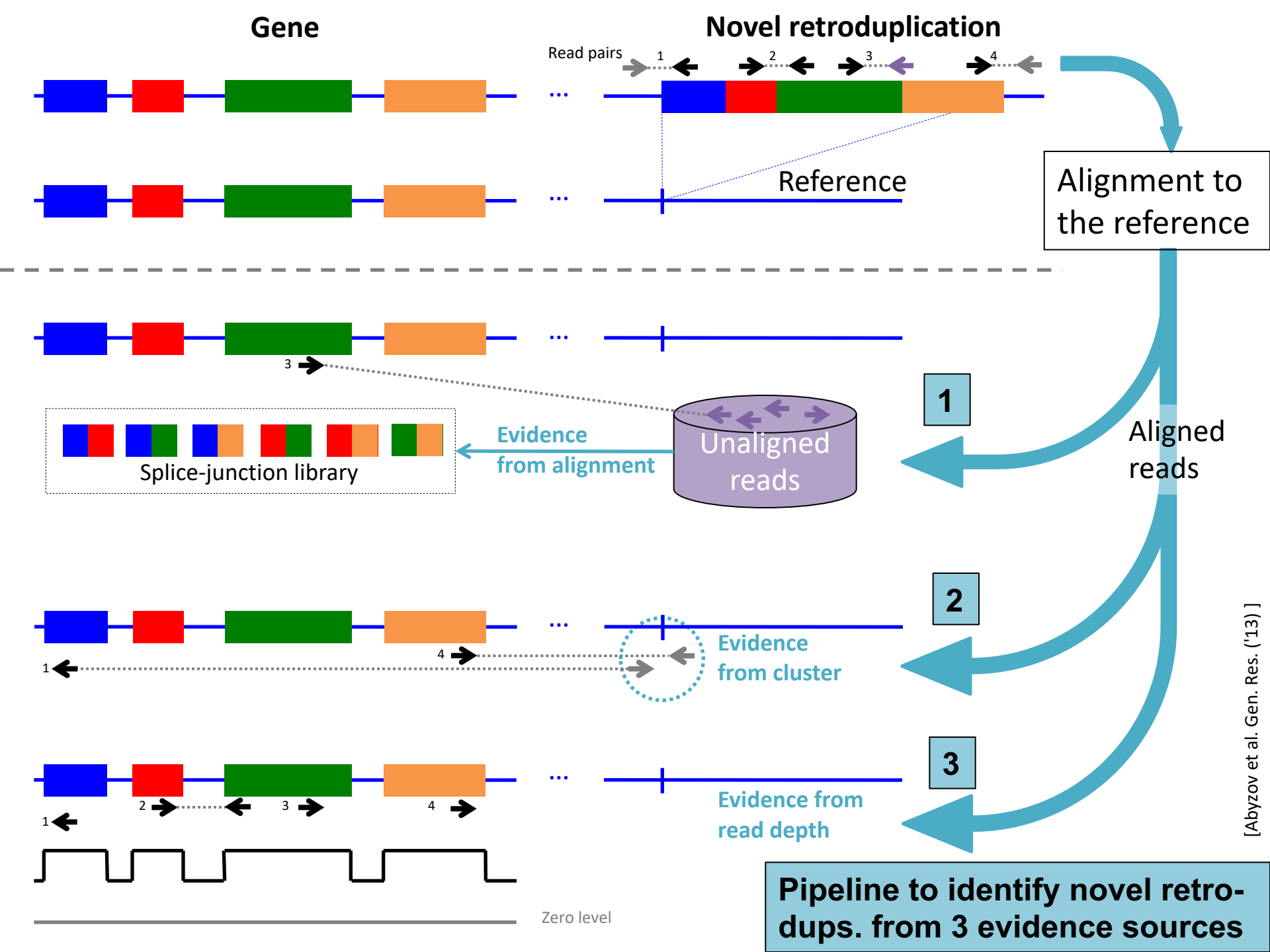
[Zhang et al. ('11) *BMC Genomics*]

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

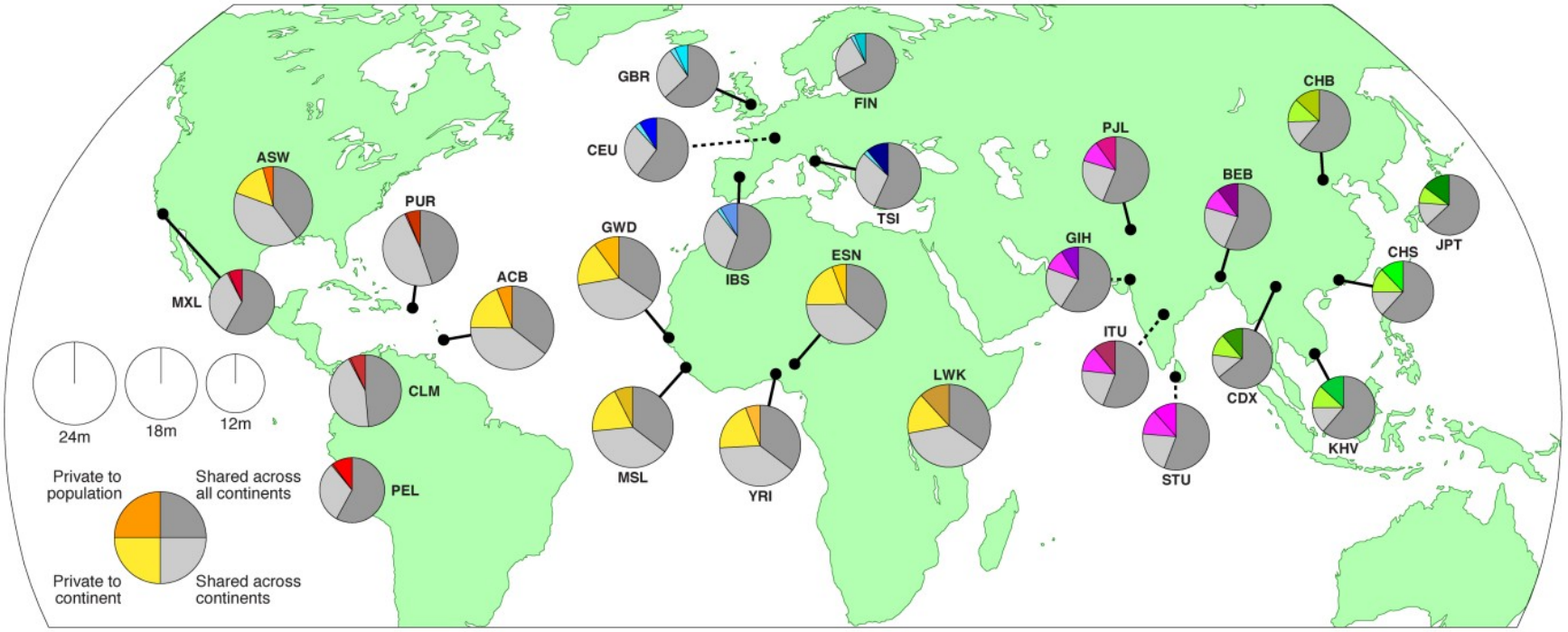
1000G summary

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

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

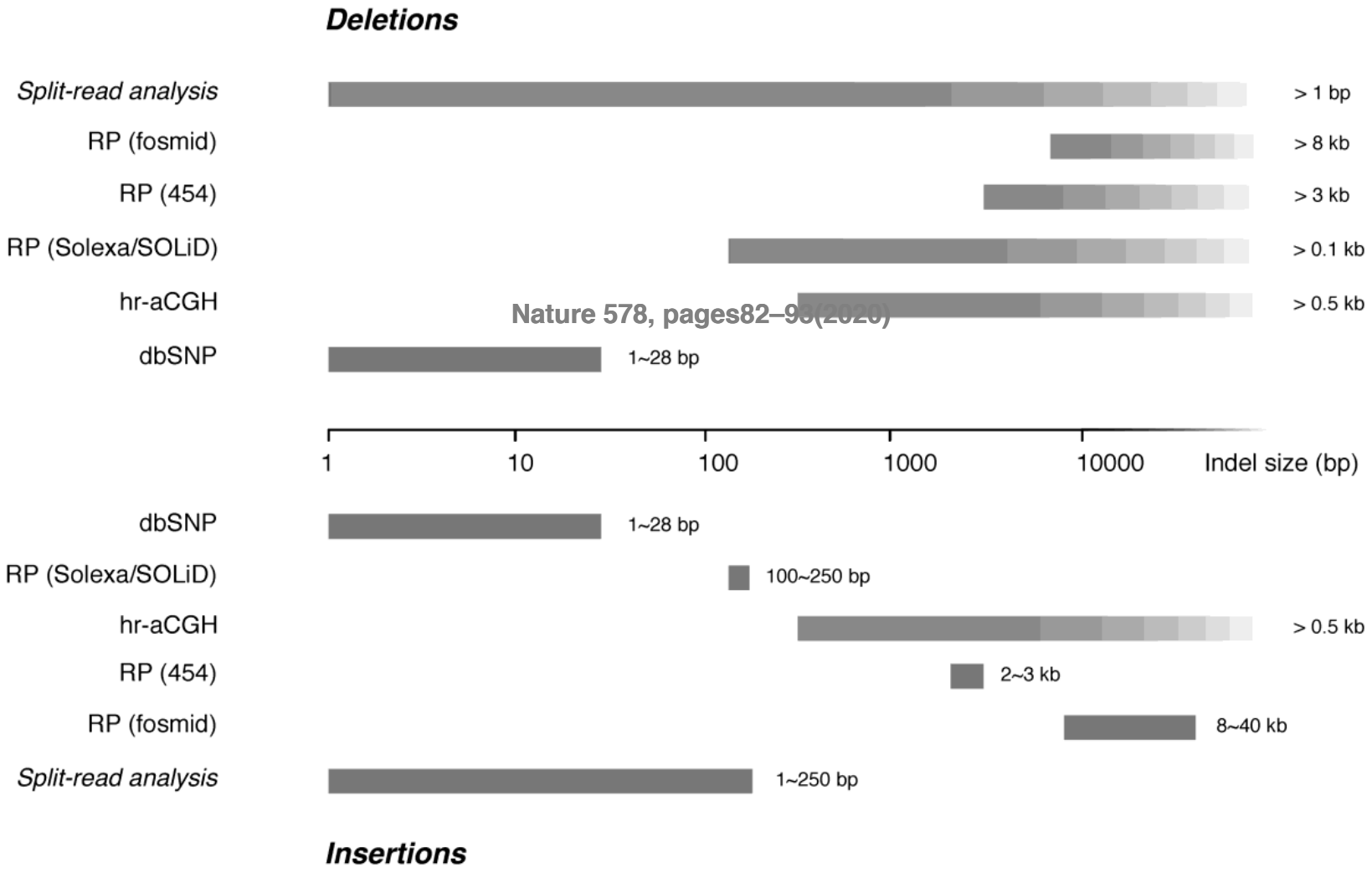
[2] 1000GP Phase3 SV paper. Submided to Nature, 2015.
[3] 1000GP ConsorSum. Submided to Nature, 2015.

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	12	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,

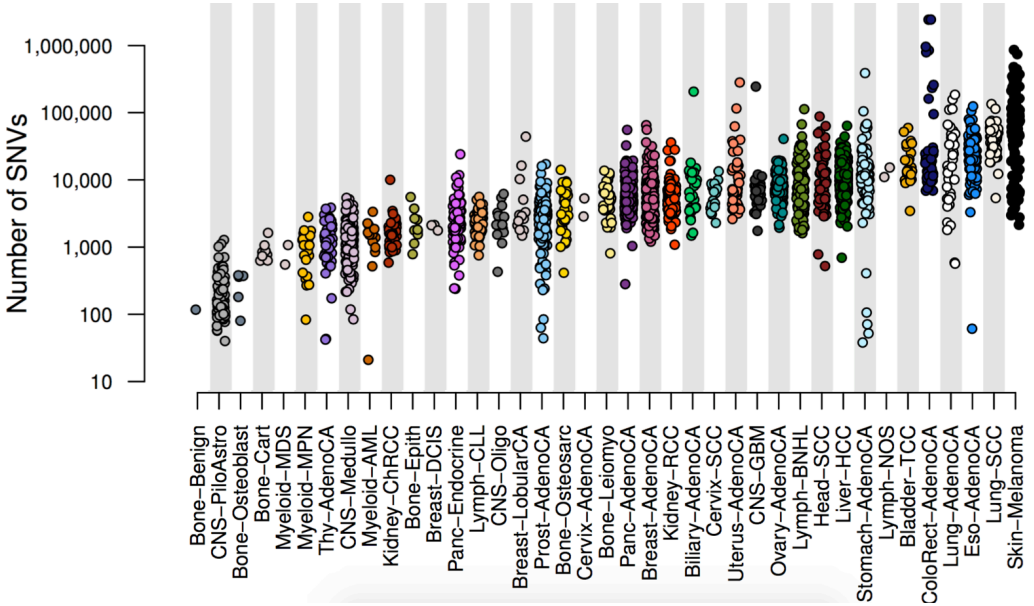
Different Approaches Work Differently on Different Events



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

PCAWG summary

PCAWG : most comprehensive resource for cancer whole genome analysis



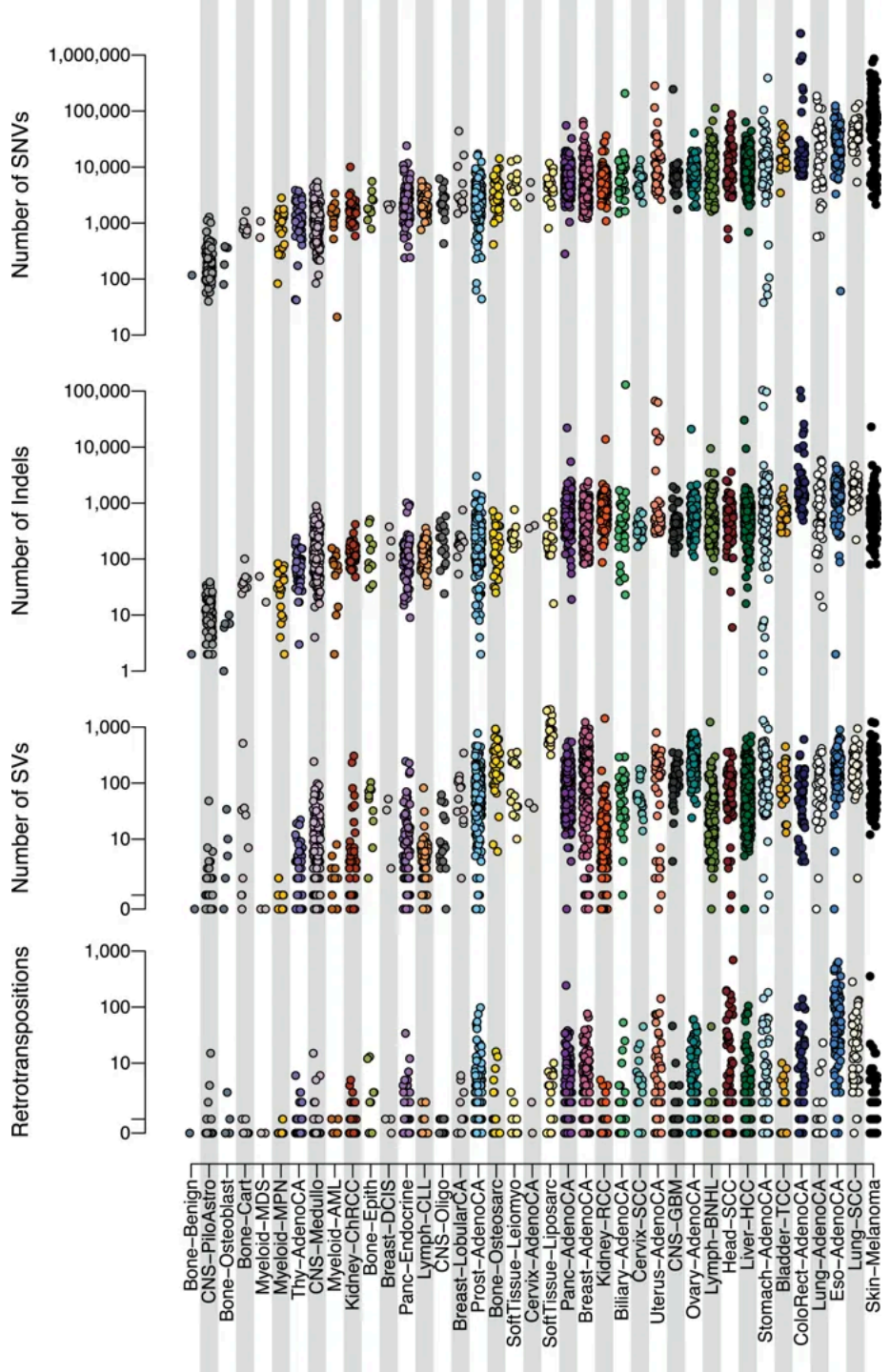
Adapted from Campbell et. al., bioRxiv ('17).
Now published as Nature 578: 82–93 (2020)

Project Goals:

- **To understand role of non-coding regions of cancer genomes in disease progression.**
- **Union of TCGA-ICGC efforts**

- **Jointly analyzing ~2800 whole genome tumor/normal pairs**
  - **> 580 researchers**
  - **16 thematic working groups**
  - **~30M total somatic SNVs**

PCAWG
PanCancer Analysis
OF WHOLE GENOMES



PCAWG Summary Variant Totals by Cancer

Nature 578: 82–93 (2020)