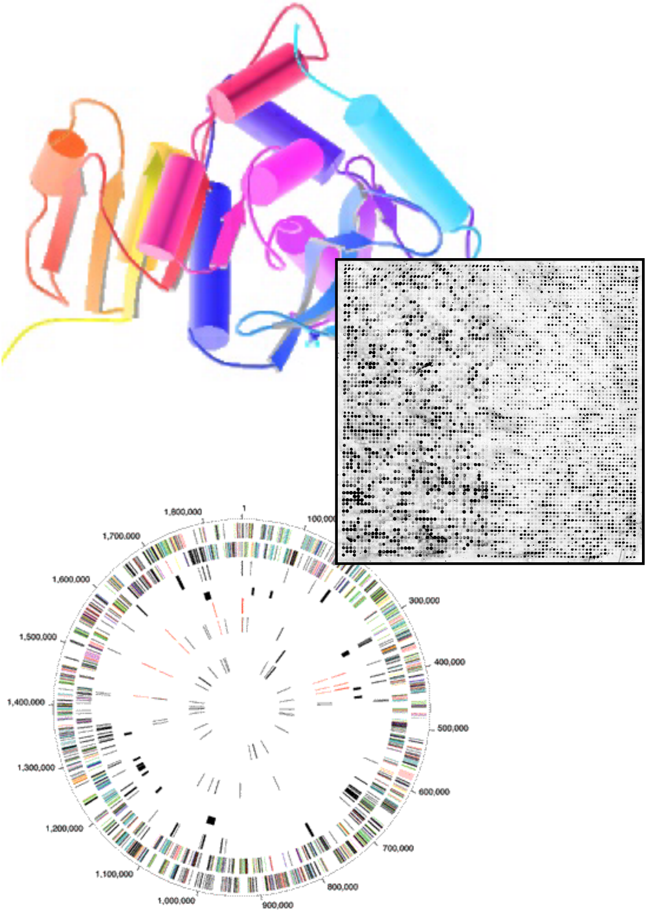
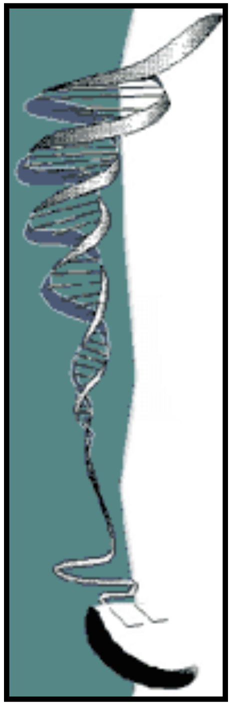
Biomed. Data Sci:

Variant Identification, Focusing on SVs





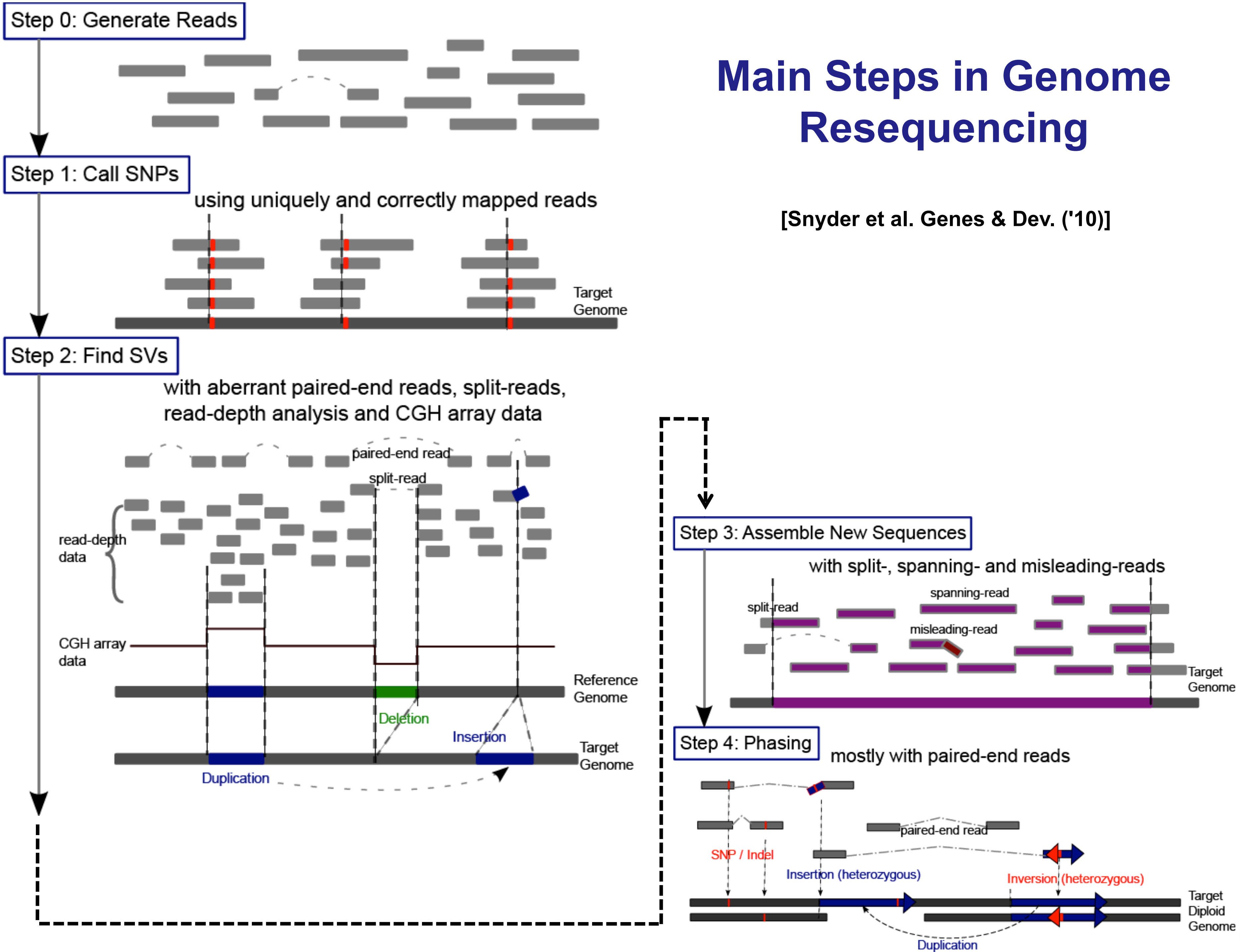
Mark Gerstein, Yale University

gersteinlab.org/courses/452

(last edit in spring ’18)

Main Steps in Genome Resequencing

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



Step 0: Generate Reads

Step 1: Call SNPs using uniquely and correctly mapped reads

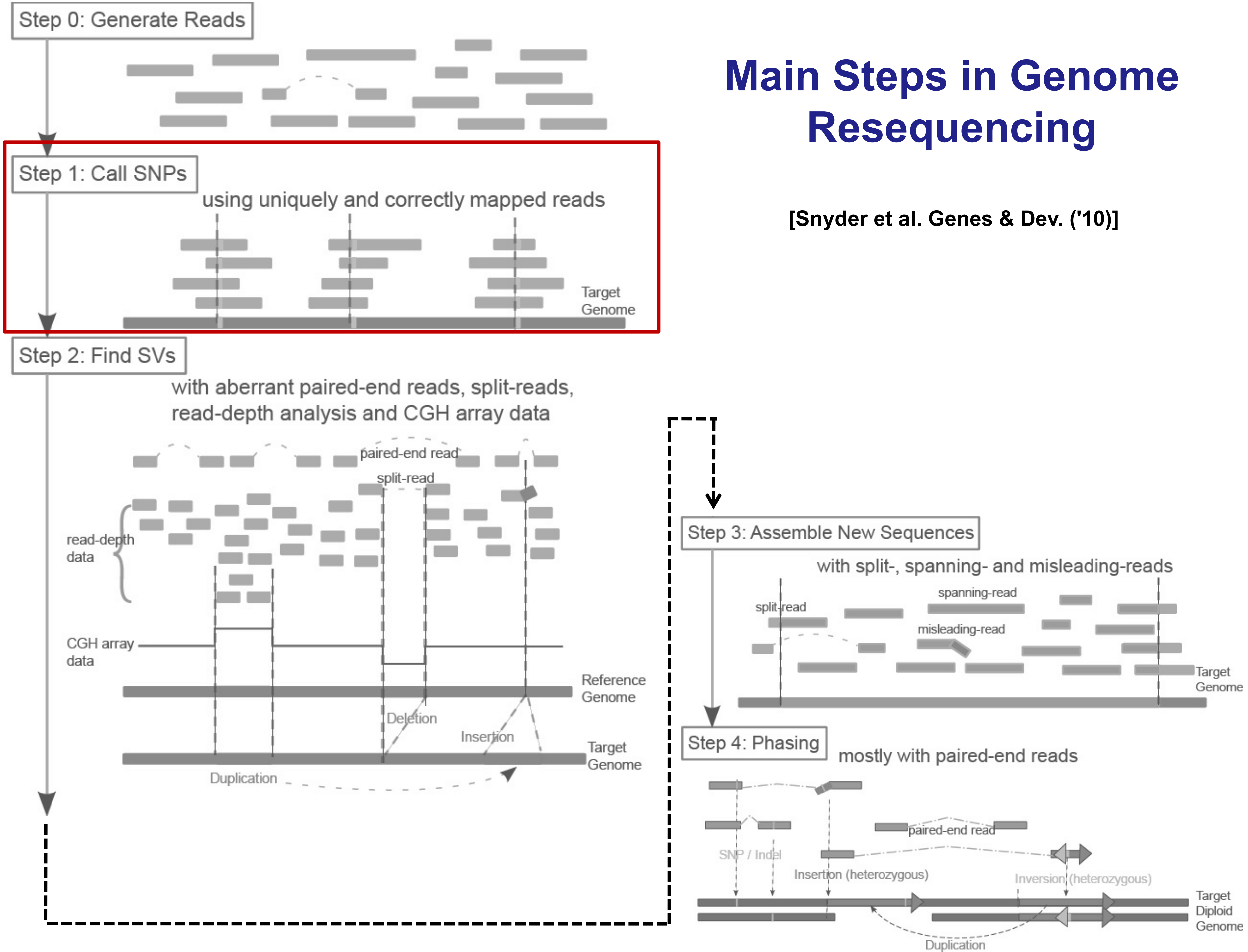
Step 2: Find SVs with aberrant paired-end reads, split-reads, read-depth analysis and CGH array data

Step 3: Assemble New Sequences with split-, spanning- and misleading-reads

Step 4: Phasing mostly with paired-end reads

Main Steps in Genome Resequencing

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



Step 0: Generate Reads

Step 1: Call SNPs

using uniquely and correctly mapped reads

Step 2: Find SVs

with aberrant paired-end reads, split-reads, read-depth analysis and CGH array data

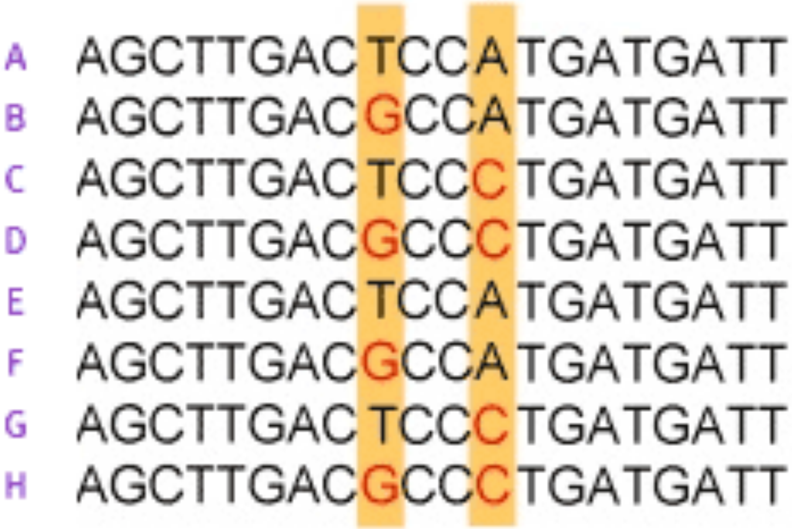
Step 3: Assemble New Sequences

with split-, spanning- and misleading-reads

Step 4: Phasing

mostly with paired-end reads

Bayes' Theorem to detect genomic variant



```
A  AGCTTGACTCCATGATGATT
B  AGCTTGACGCCATGATGATT
C  AGCTTGACTCCCTGATGATT
D  AGCTTGACGCCCTGATGATT
E  AGCTTGACTCCATGATGATT
F  AGCTTGACGCCATGATGATT
G  AGCTTGACTCCCTGATGATT
H  AGCTTGACGCCCTGATGATT
```

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

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

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.

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

With errors, the calculation is more complicated.

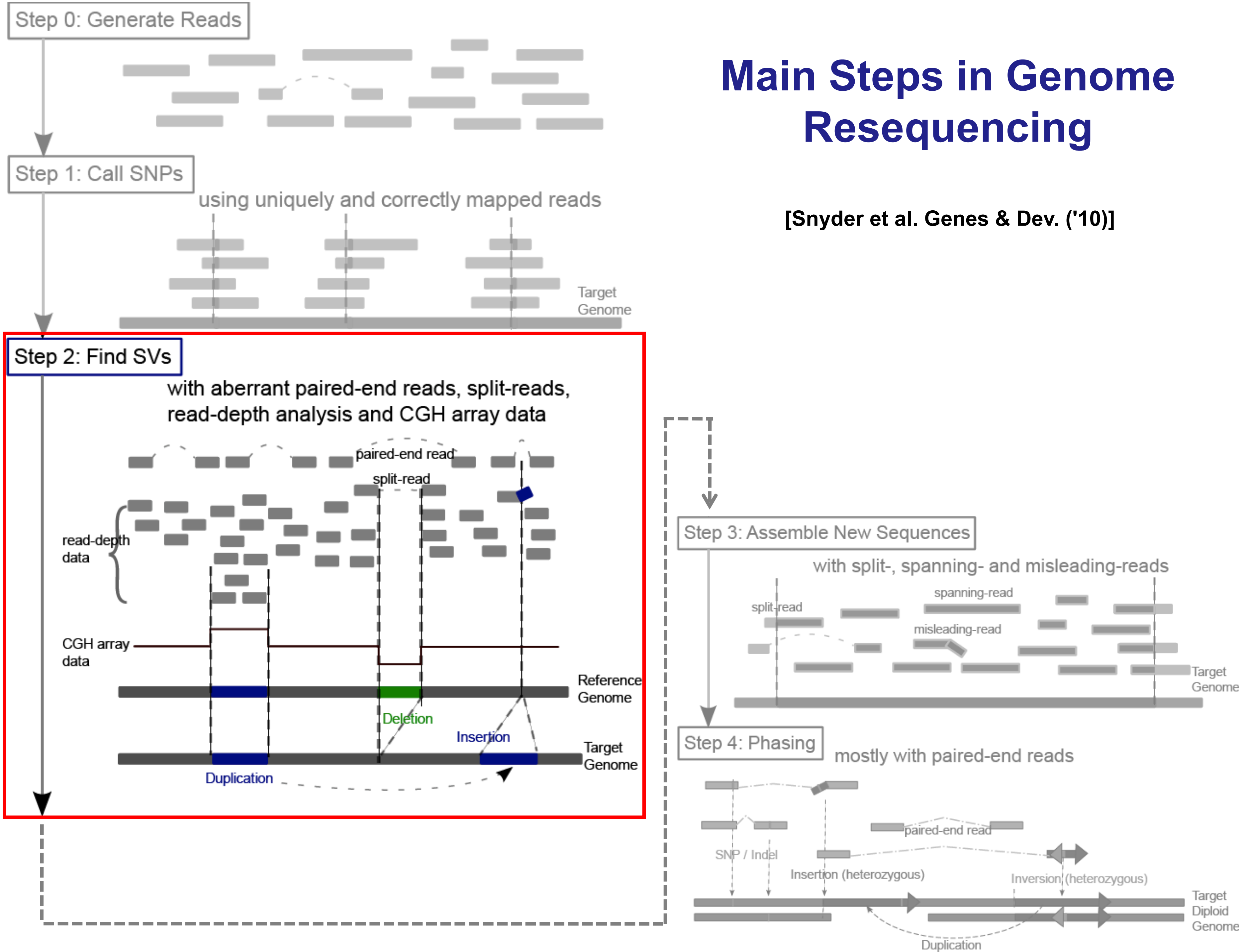
In general:

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

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

Main Steps in Genome Resequencing

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



Step 0: Generate Reads

Step 1: Call SNPs

using uniquely and correctly mapped reads

Step 2: Find SVs

with aberrant paired-end reads, split-reads, read-depth analysis and CGH array data

Step 3: Assemble New Sequences

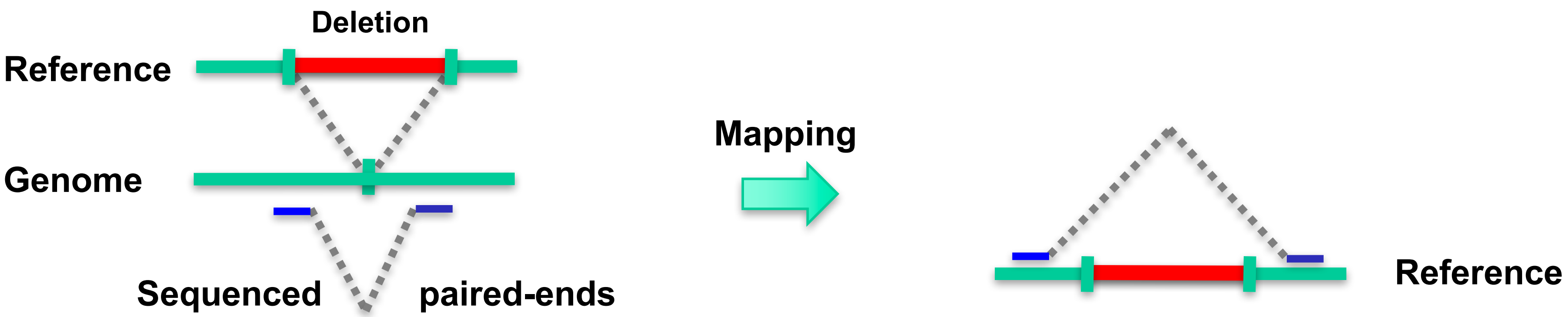
with split-, spanning- and misleading-reads

Step 4: Phasing

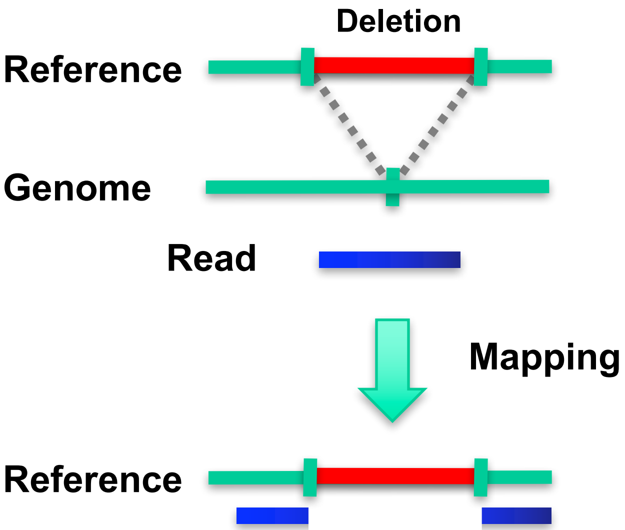
mostly with paired-end reads

Methods to **Find SVs**

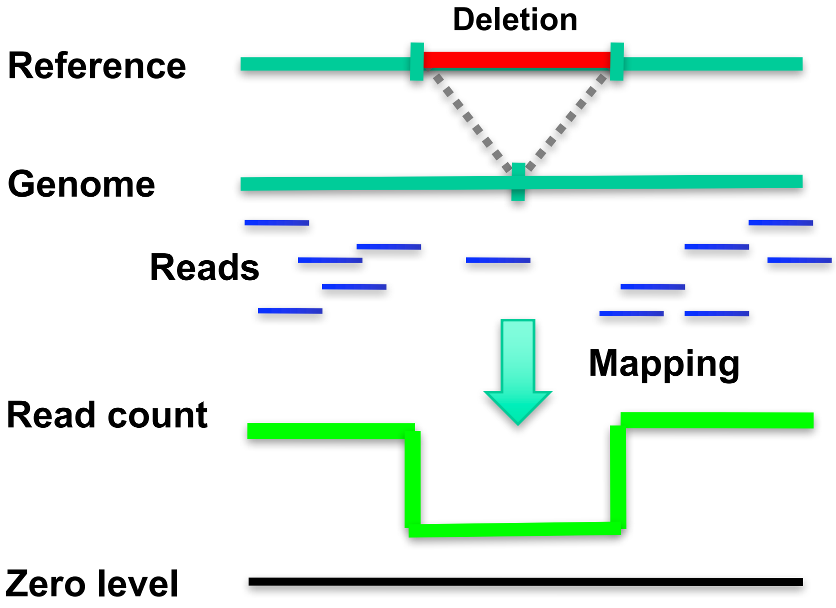
1. Paired ends



2. Split read



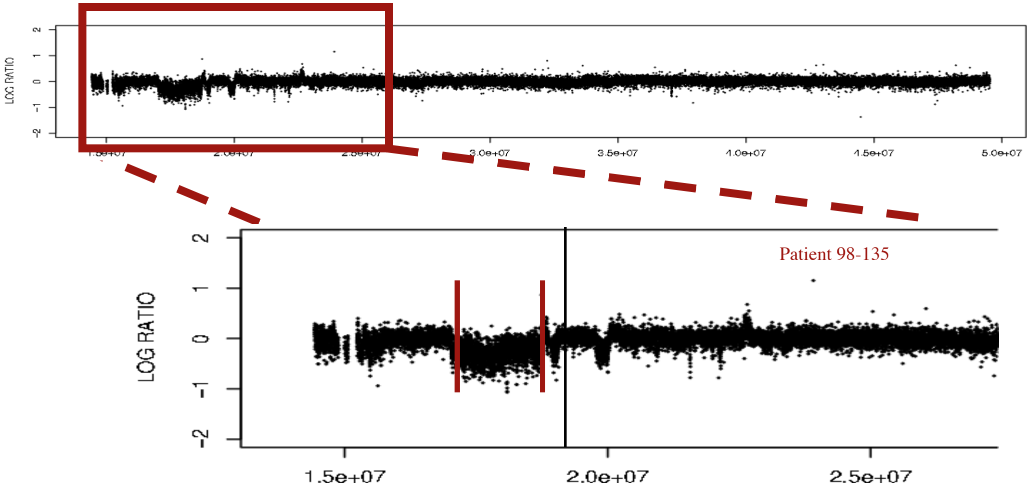
3. Read depth (or aCGH)



4. Local Reassembly

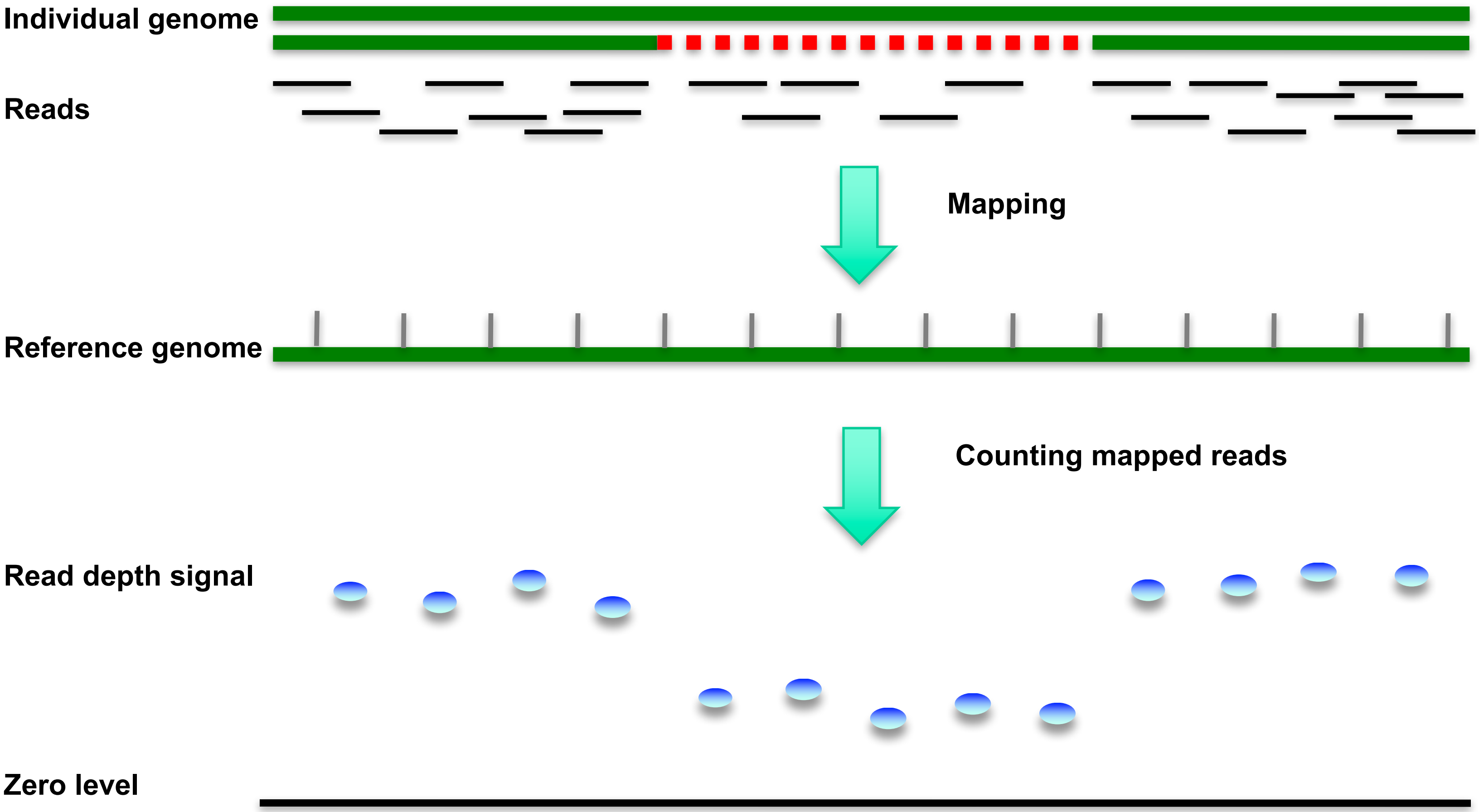
[Snyder et al. Genes & Dev. ('10)]

Read Depth



Array Signal

Read depth



Individual genome

Reads

Mapping

Reference genome

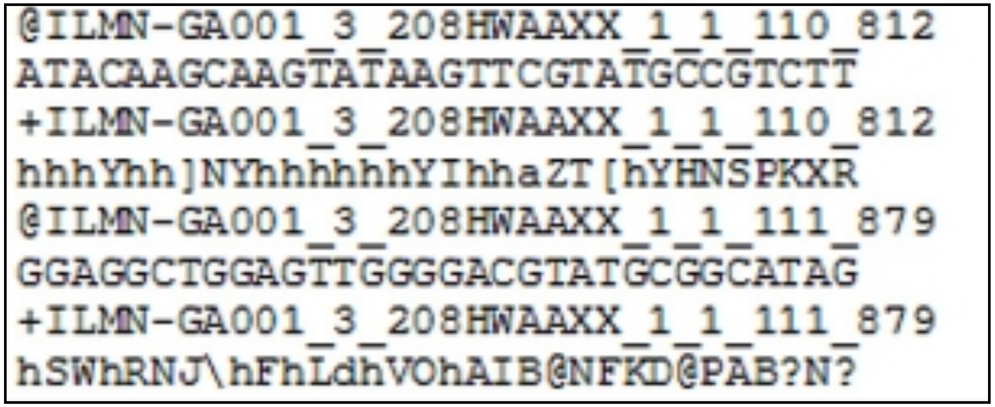
Counting mapped reads

Read depth signal

Zero level

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

Reads to Signal Track

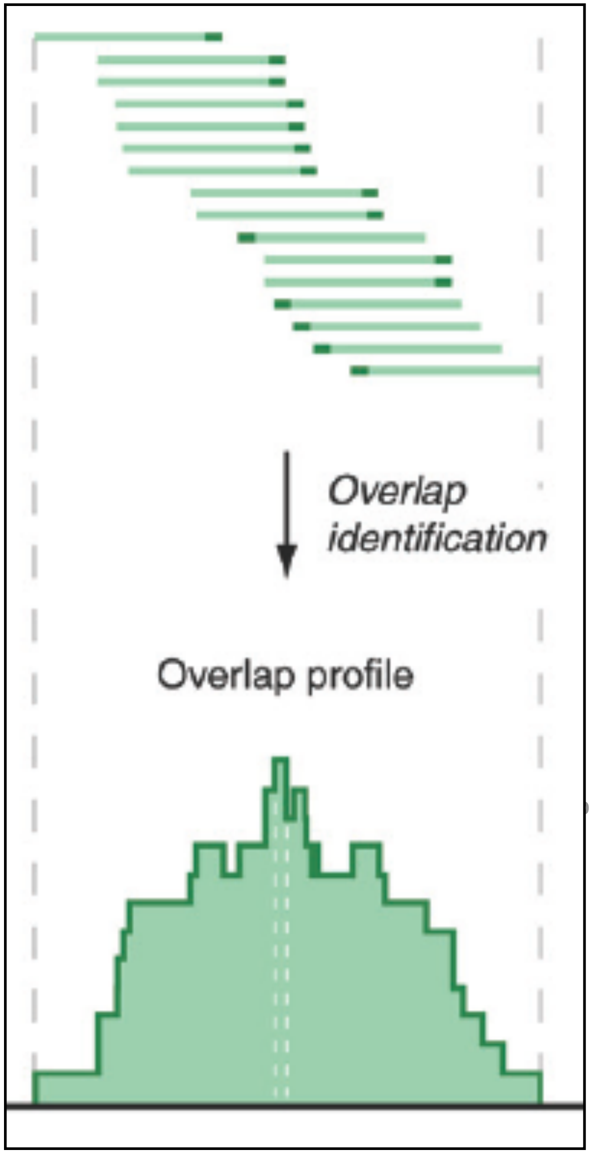


```
@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYhhhhhhYIhhaZT[hYHNSPKXR
@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
+ILMN-GA001_3_208HWAAXX_1_1_111_879
hSWhRNJ\hFhLdhVOhAIB@NFKD@PAB?N?
```

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

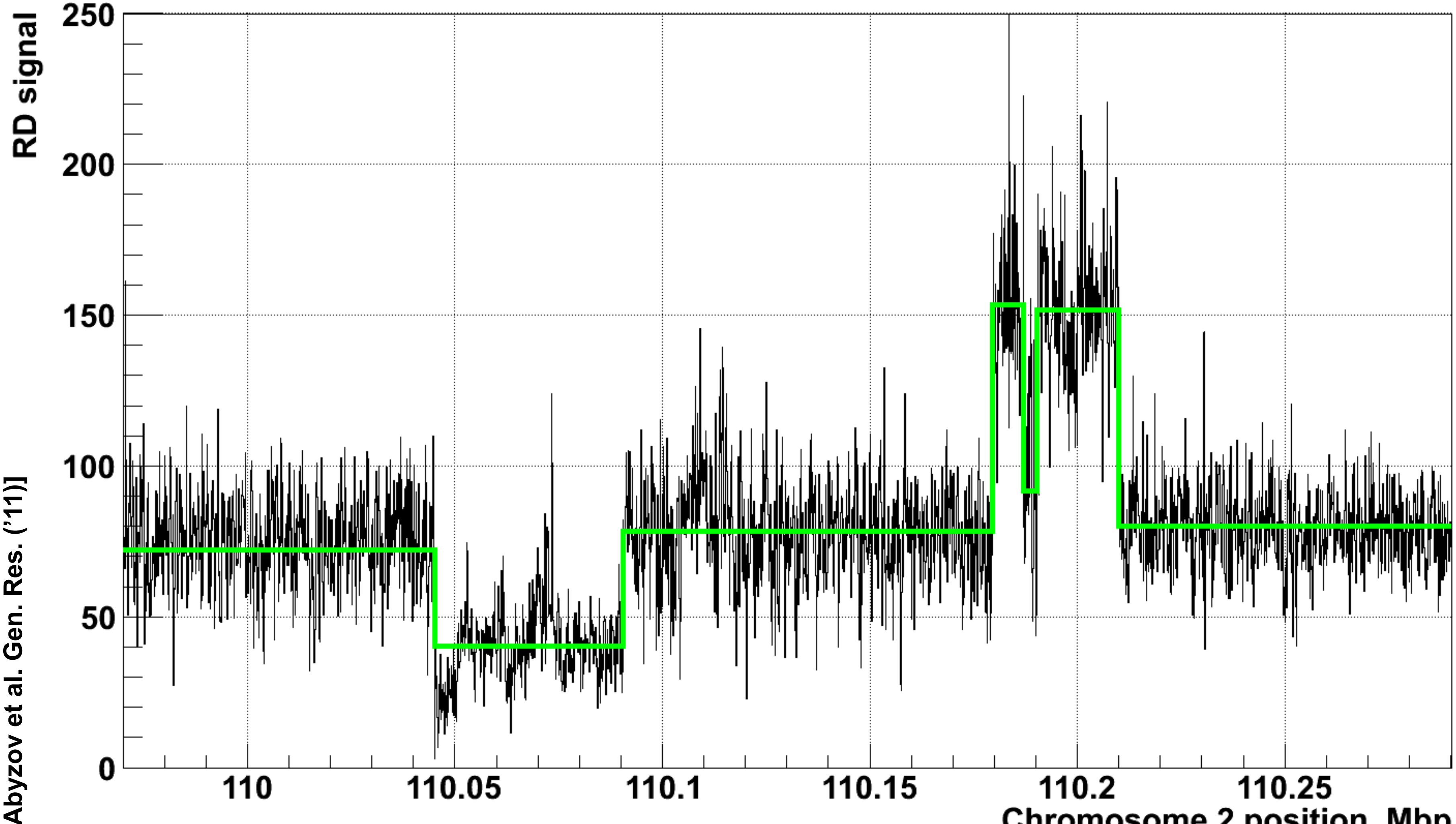
Reads => Signal (Intermediate file)

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



[PLOS CB 4:e1000158]

Example of Application to RD data

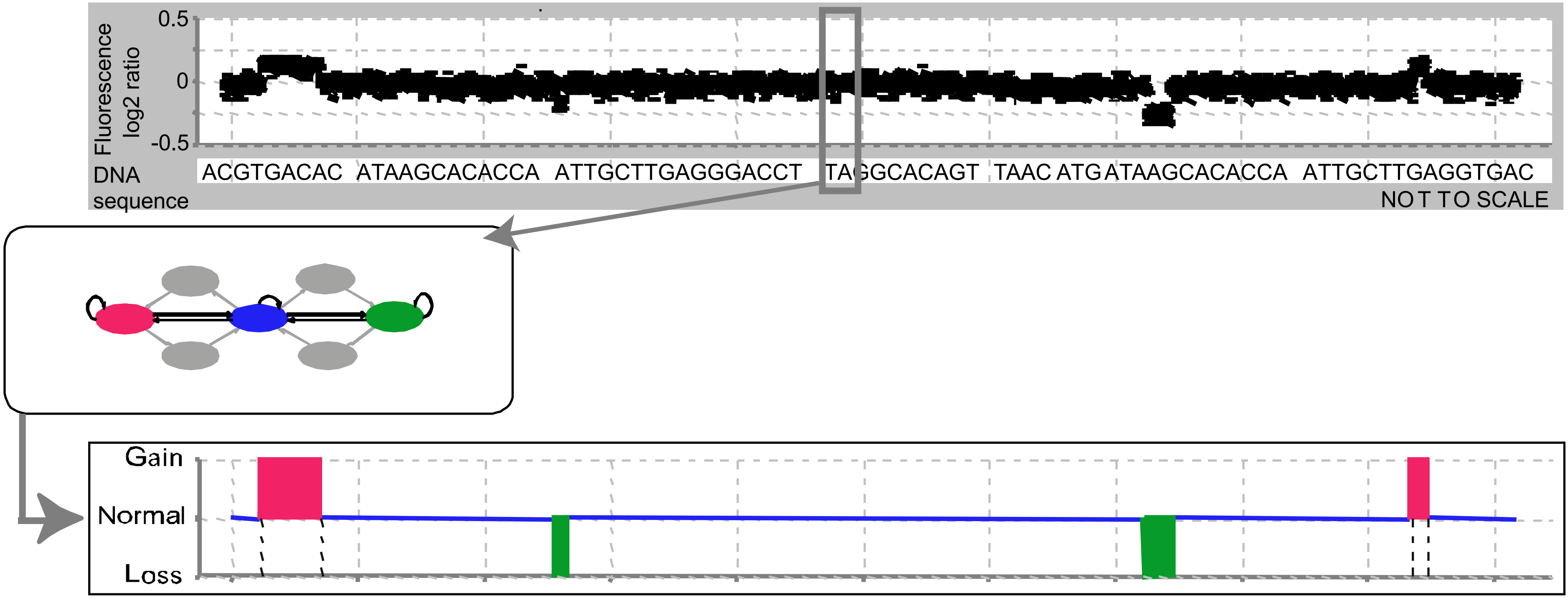


[Abyzov et al. Gen. Res. ('11)]

NA12878, Solexa 36 bp paired reads, ~30x coverage

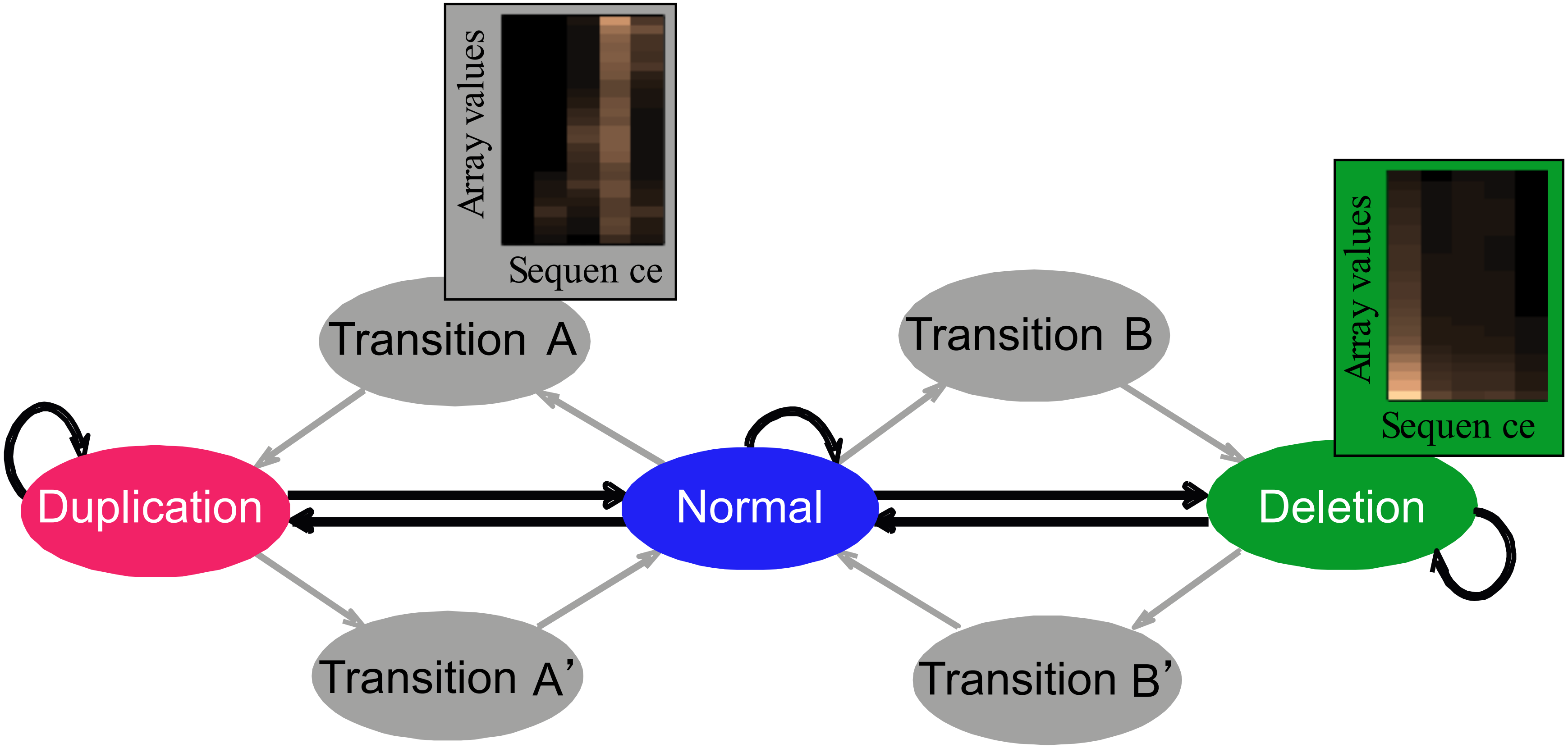
HMM

- 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)

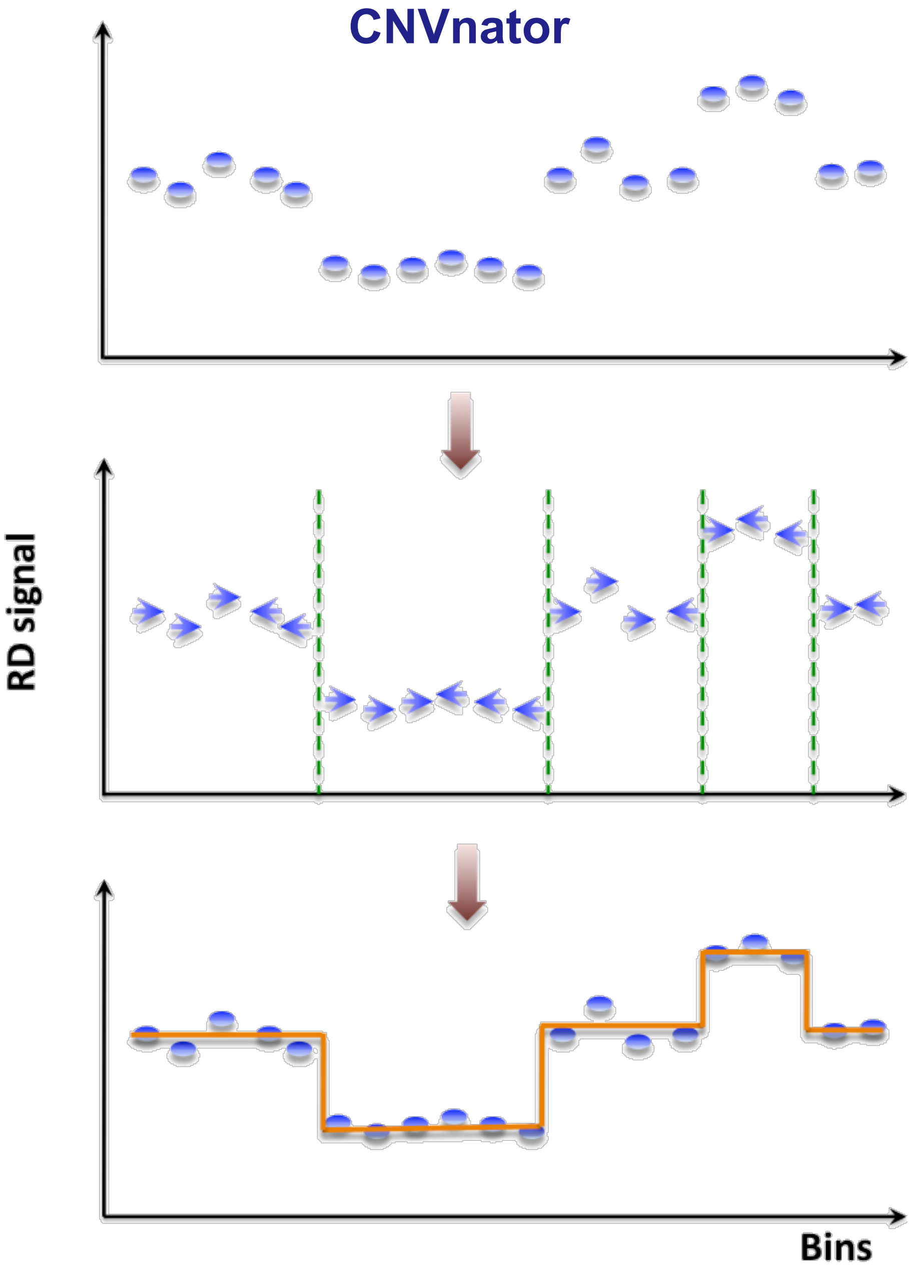
Statistically integrates array signal and DNA sequence signatures (using a discrete-valued bivariate HMM)



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



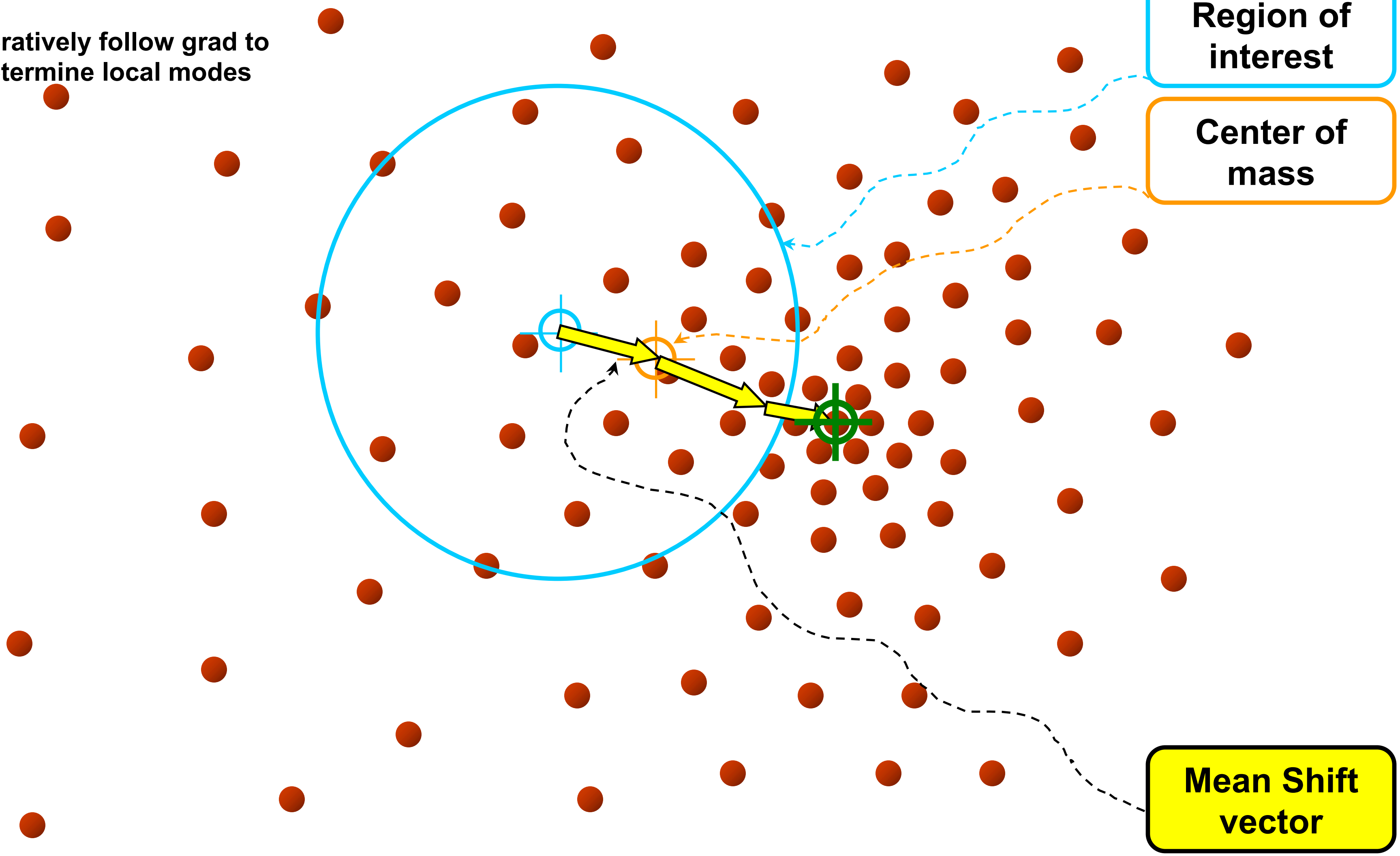
[Abyzov et al. Gen. Res. ('11)]

● Observed depth of coverage counts as samples from PDF

⇨ Kernel-based approach to estimate local gradient of PDF

⌖ Iteratively follow grad to determine local modes

Intuitive Description of MSB



Region of interest

Center of mass

Mean Shift vector

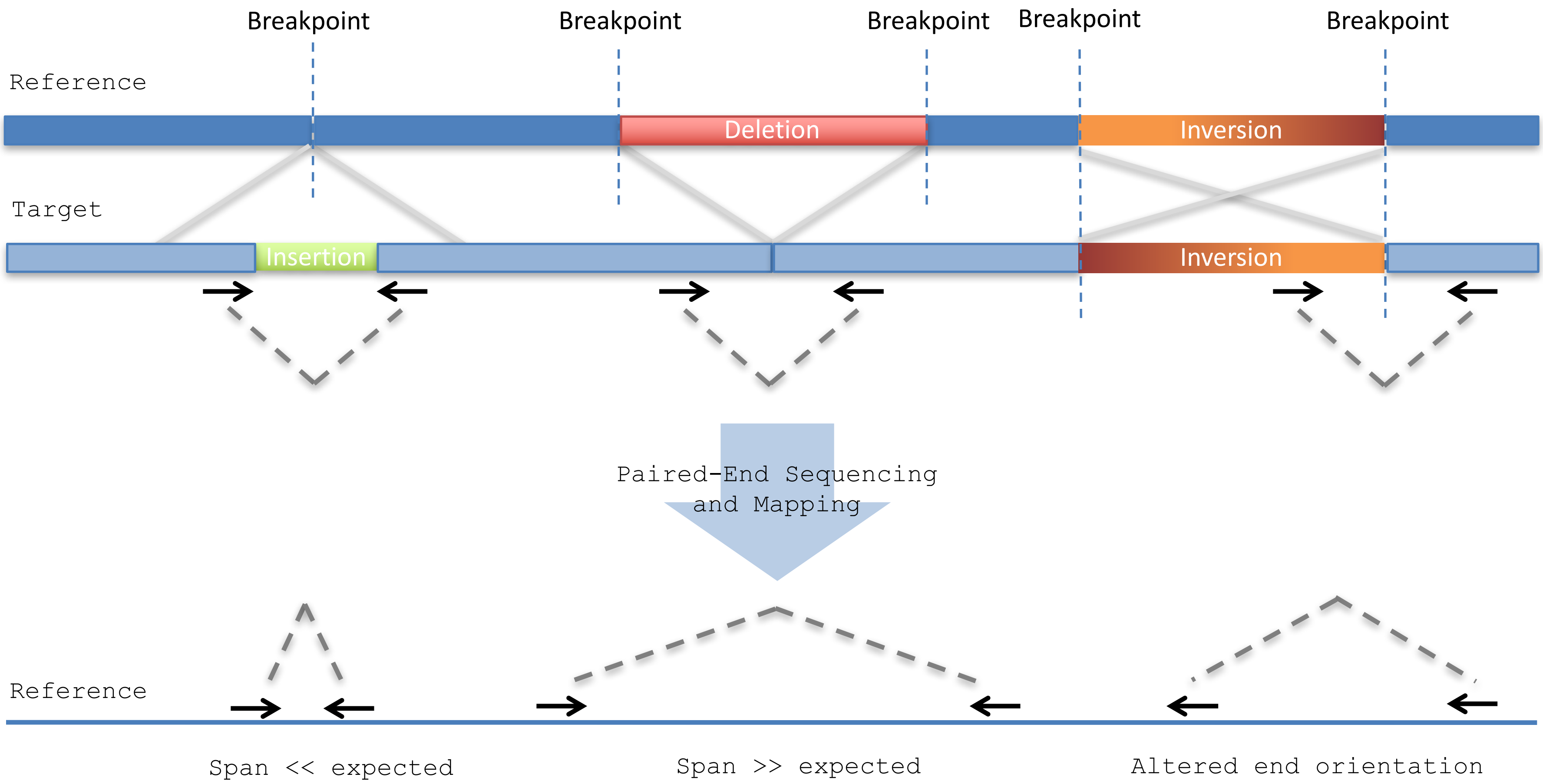
Objective : Find the densest region

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]

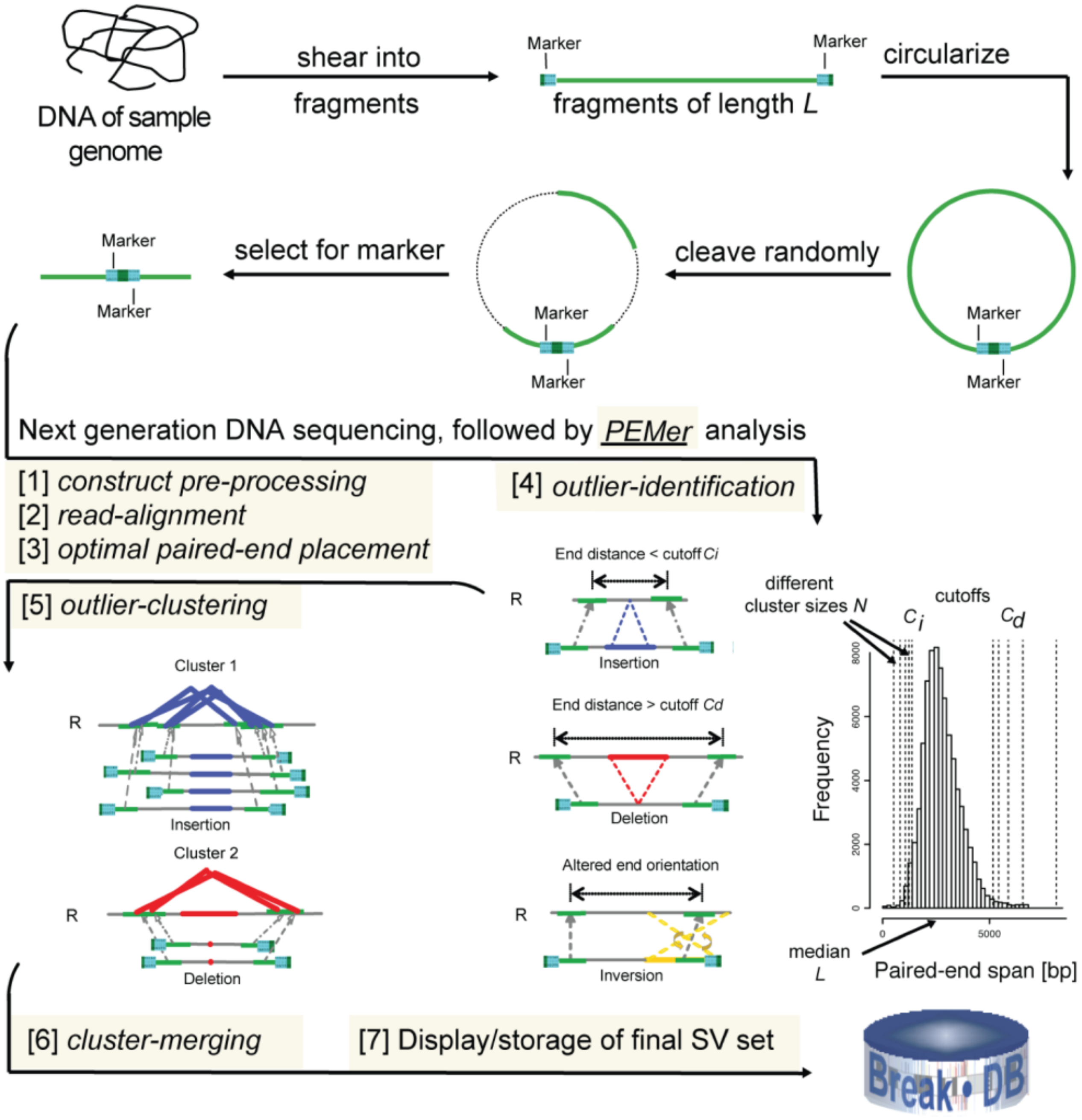
Paired-End

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



Next generation DNA sequencing, followed by *PEMer* analysis

[1] *construct pre-processing*
[2] *read-alignment*
[3] *optimal paired-end placement*

[4] *outlier-identification*

[5] *outlier-clustering*

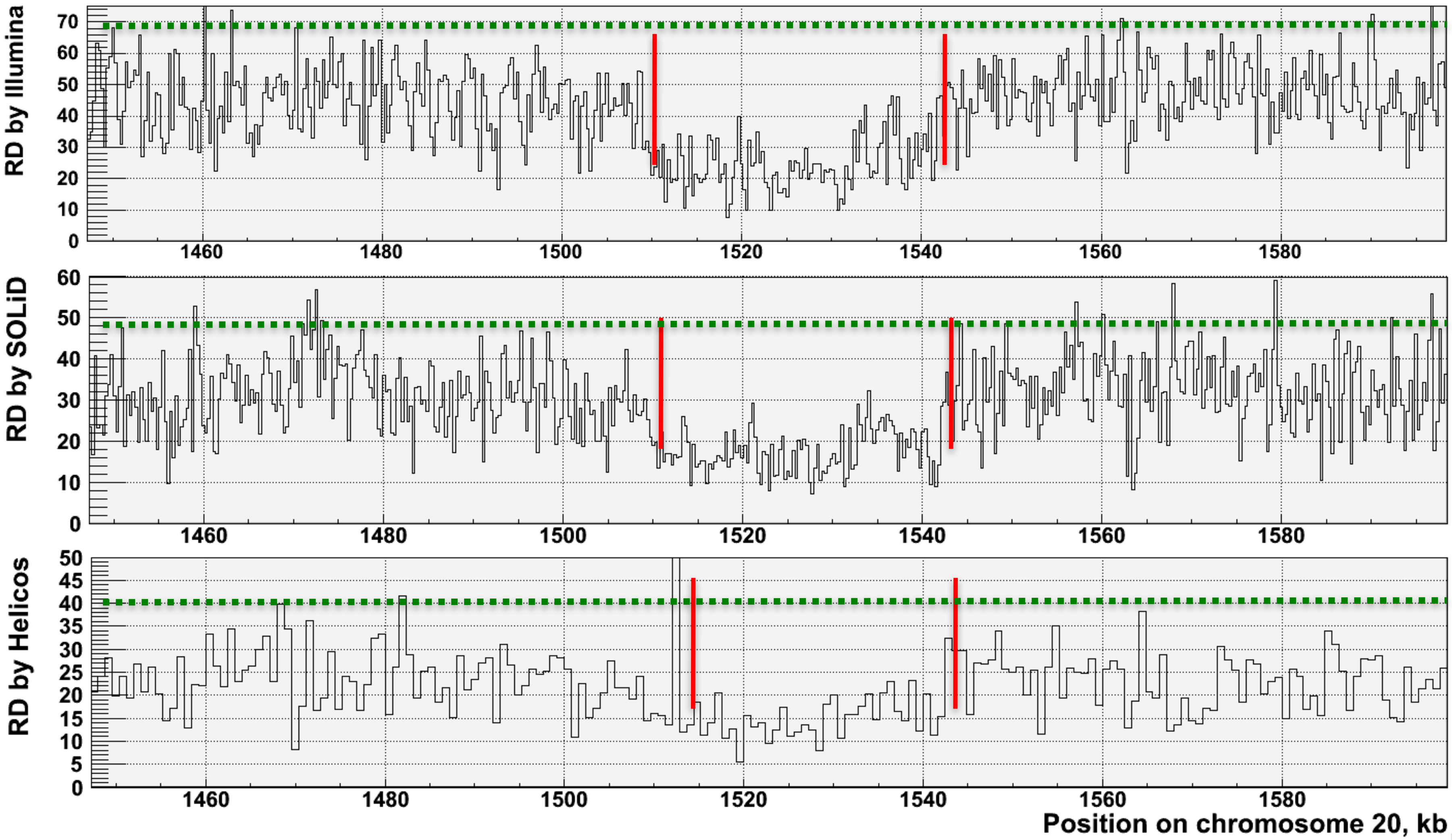
[6] *cluster-merging*

[7] Display/storage of final SV set

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

Split Read

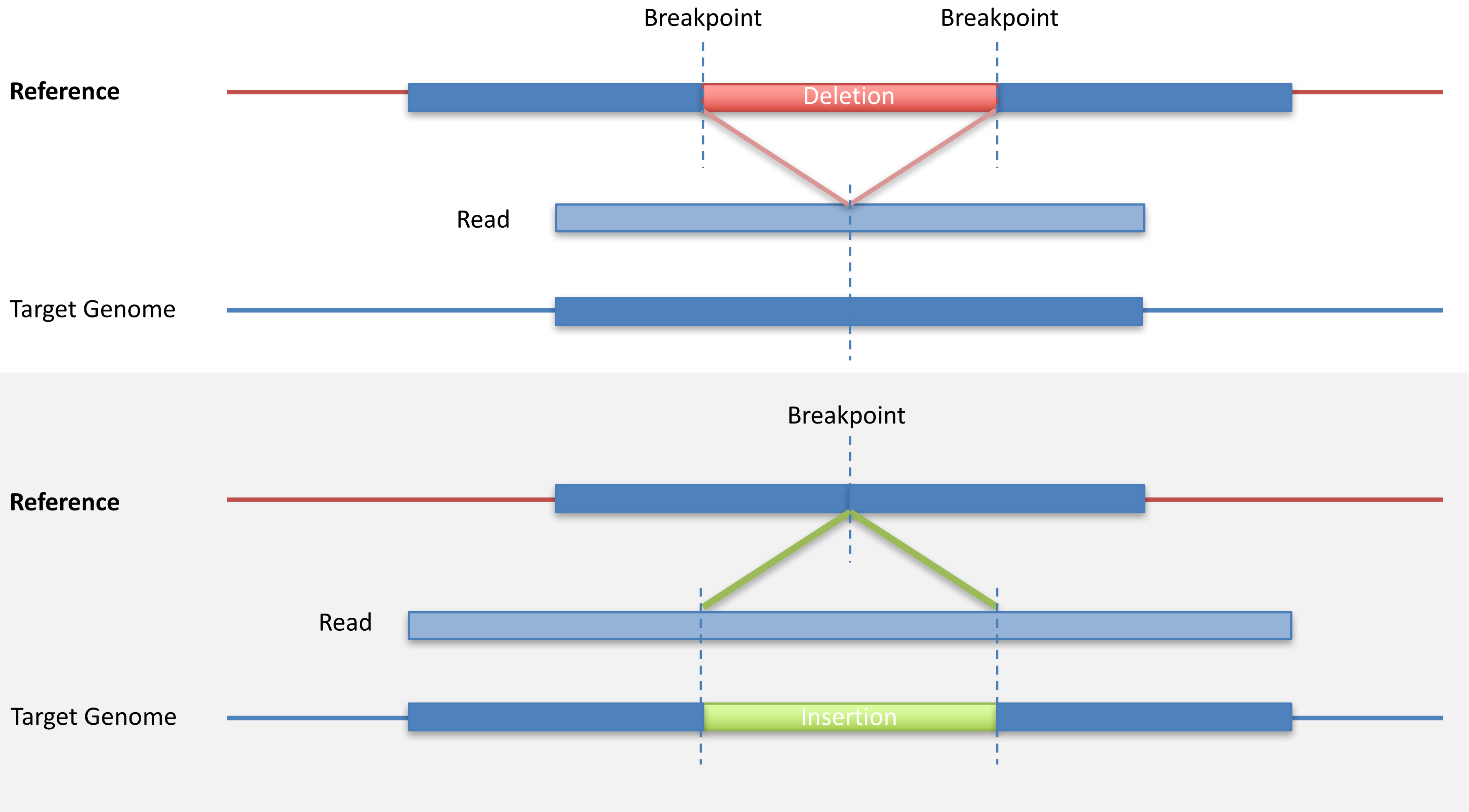
Read-depth works well on a variety of sequencing platforms but provides imprecise breakpoints



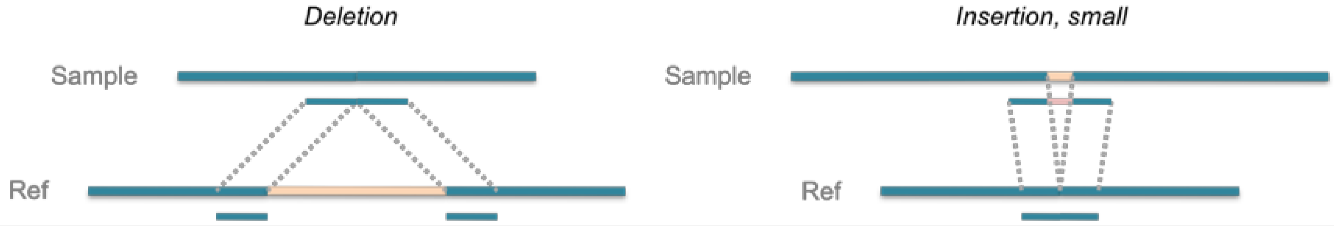
[Abyzov et al. Gen. Res. ('11)]

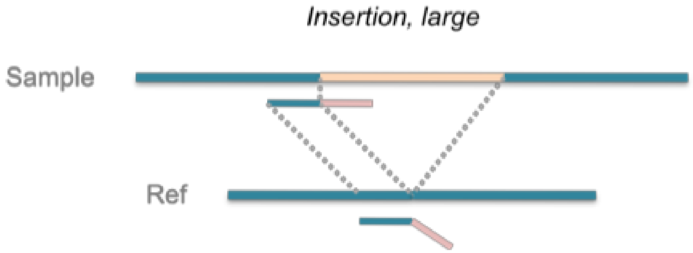
[NA18505]

Split-read Analysis



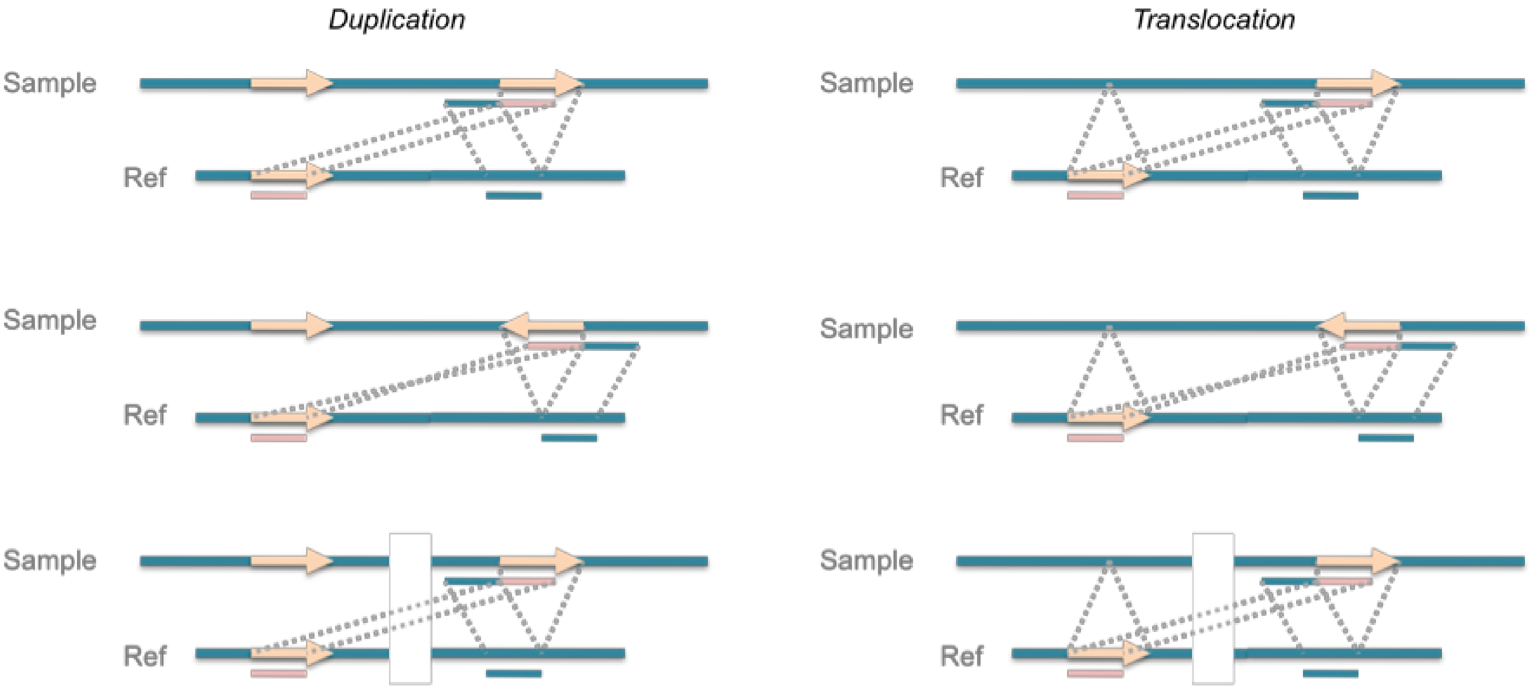
Simple SVs





Deletions are the Easiest to Identify

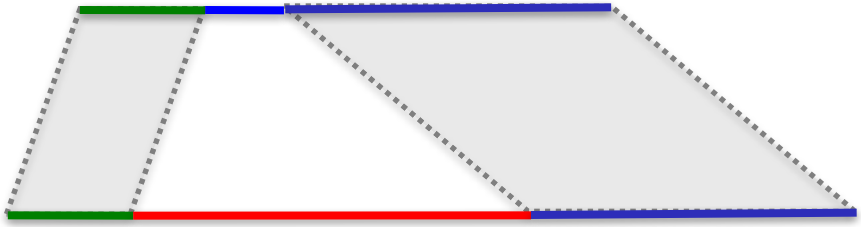
Complex SVs



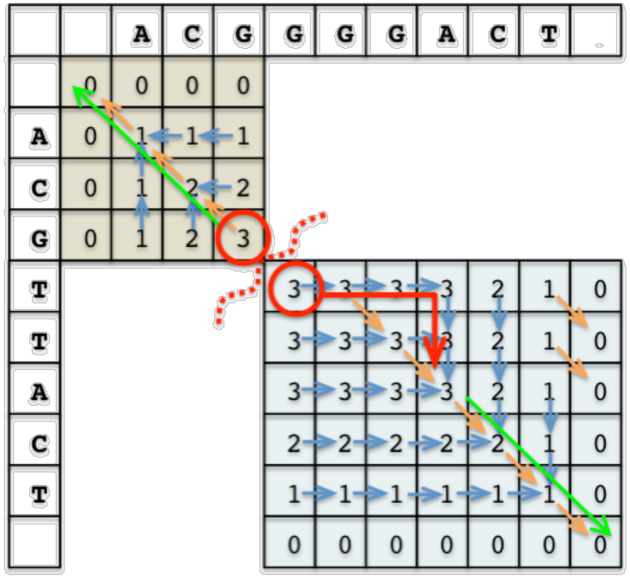
[Zhang et al. ('11) *BMC Genomics*]

Creative application of dynamic programming to a new problem

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



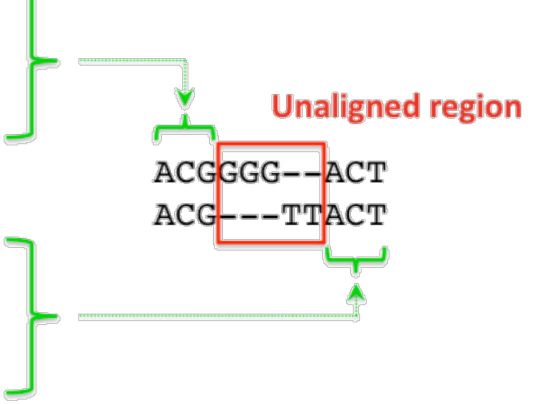
◊ Solution: SW alignment from both ends; combine max scoring alignments



		A	C	G	G	G	G	A	C	T	
	0	0	0	0							
A	0	1	1	1							
C	0	1	2	2							
G	0	1	2	3							
T					3	3	3	3	2	1	0
T					3	3	3	3	2	1	0
A					3	3	3	3	2	1	0
C					2	2	2	2	2	1	0
T					1	1	1	1	1	1	0
					0	0	0	0	0	0	0

AGE Alignment with Gap Excision

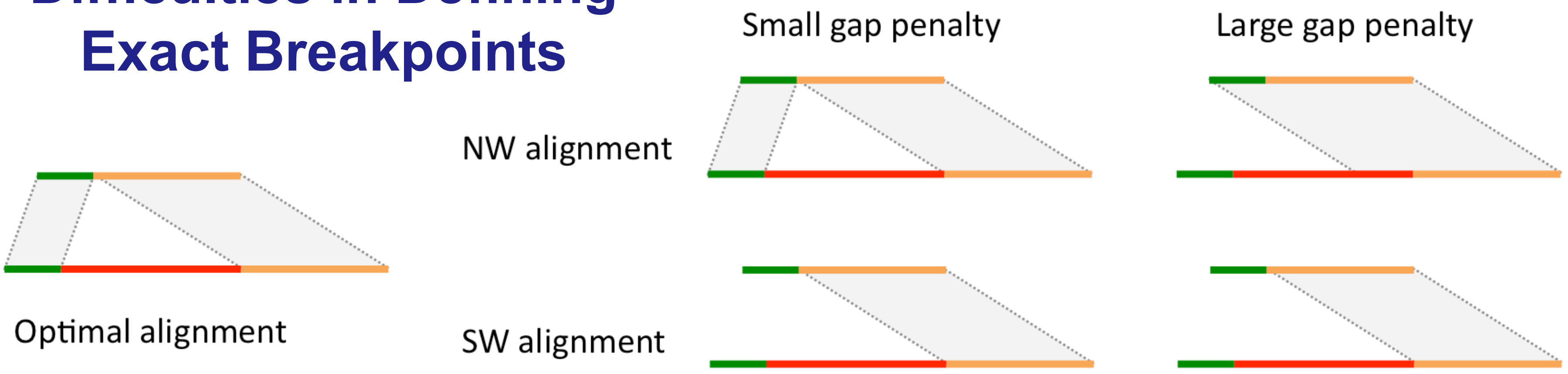
[Abyzov et al. Bioinfo. (’11)]



```
ACGGGG--ACT
ACG---TTACT
```

◊ much more detail in SV section later

Difficulties in Defining Exact Breakpoints



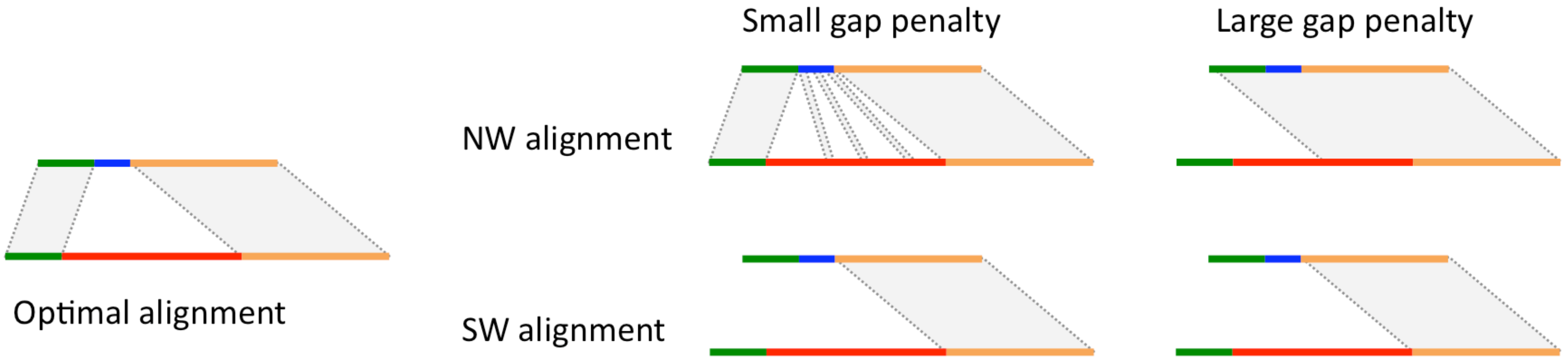
Small gap penalty

Large gap penalty

NW alignment

SW alignment

Optimal alignment



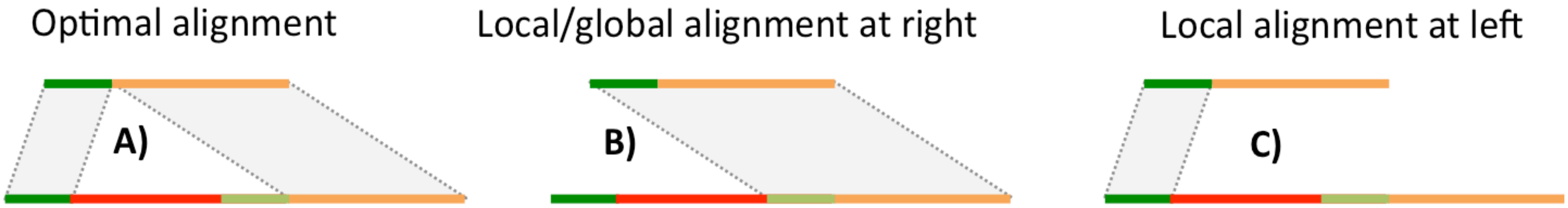
Small gap penalty

Large gap penalty

NW alignment

SW alignment

Optimal alignment



Optimal alignment

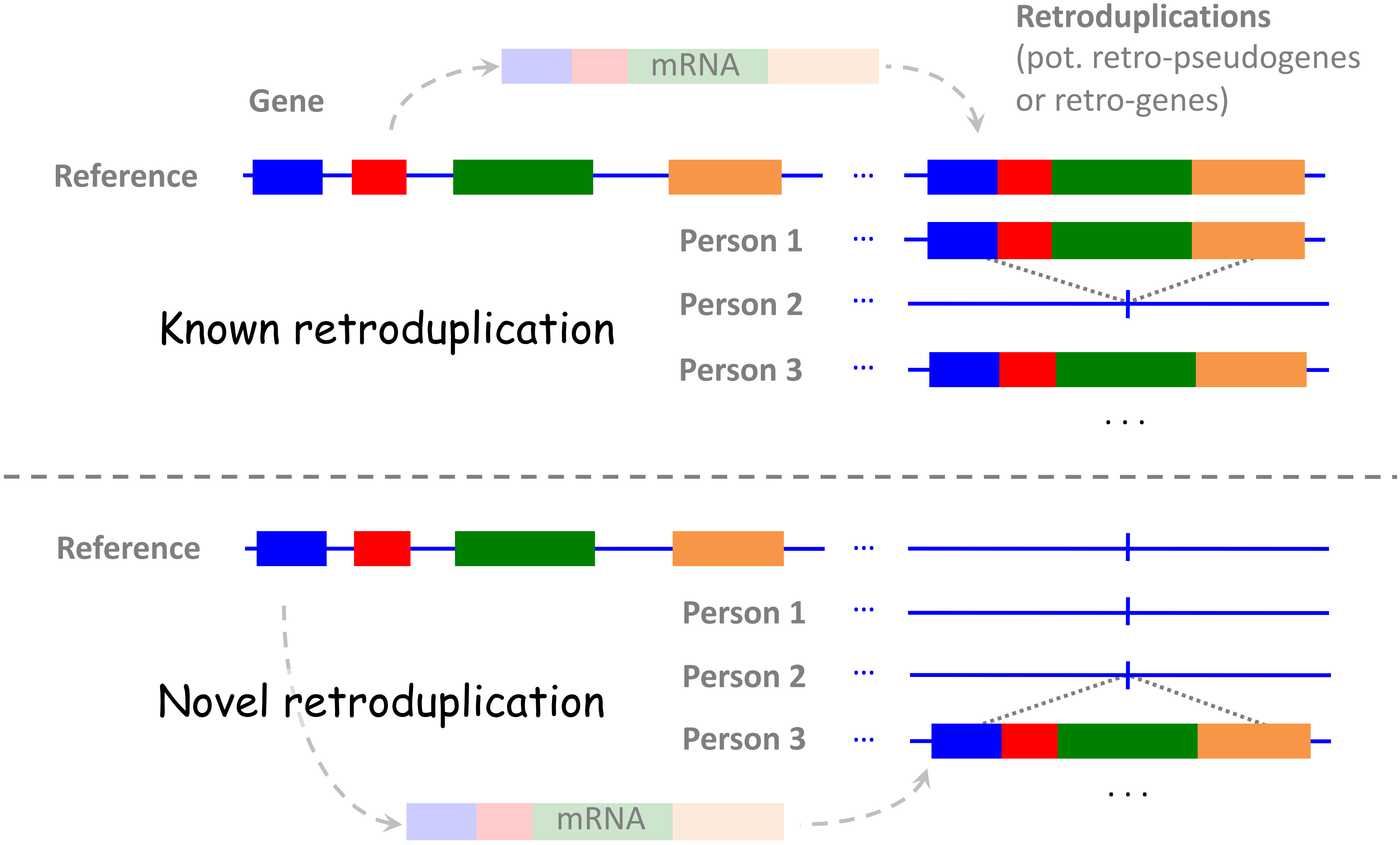
Local/global alignment at right

Local alignment at left

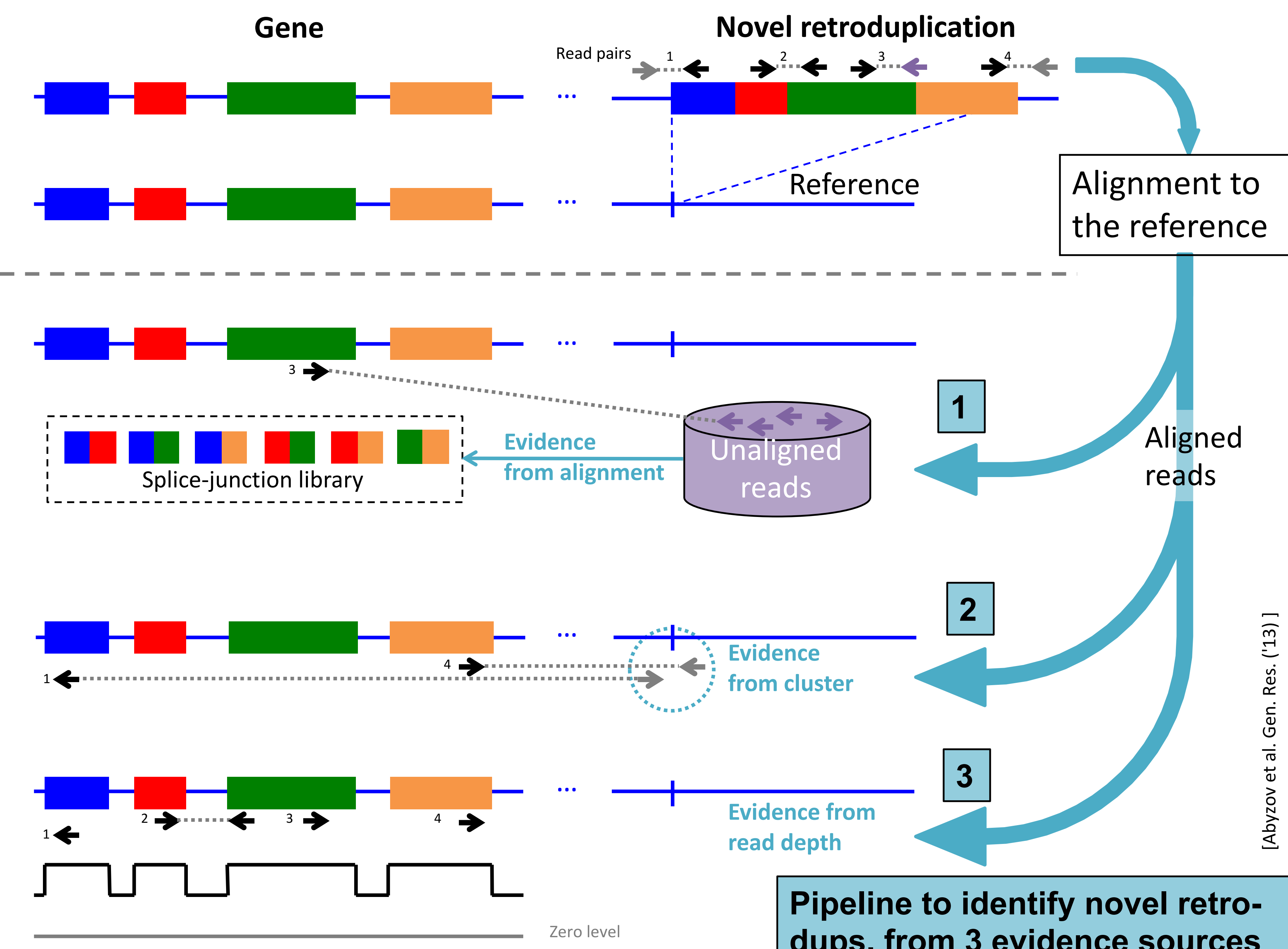
[Abyzov & Gerstein ('11) Bioinfo.]

RDV & Mobile Elements

Retroduplication variation (RDV)

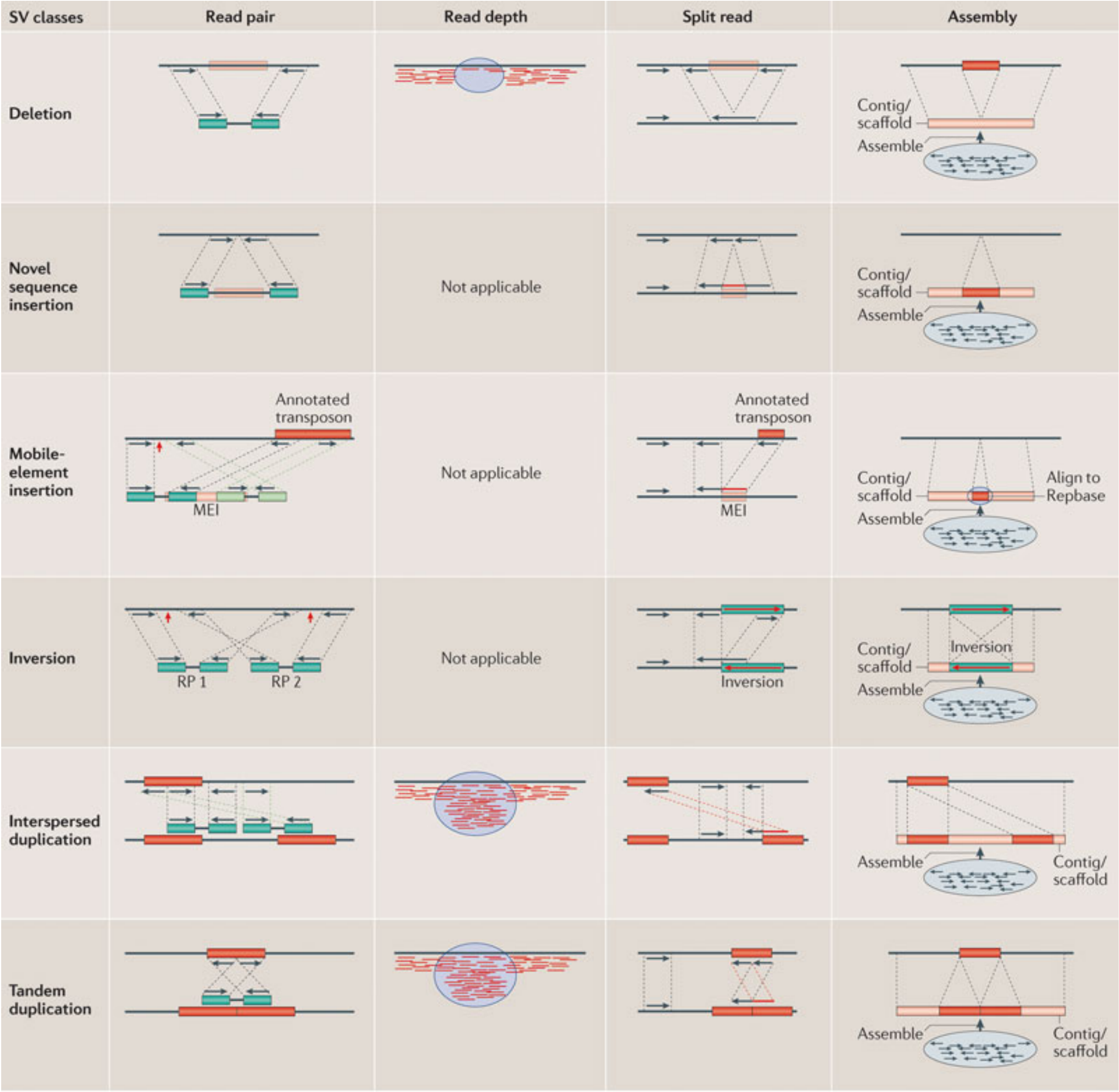


[Abyzov et al. Gen. Res. ('13)]



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

Pipeline to identify novel retro-dups. from 3 evidence sources



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

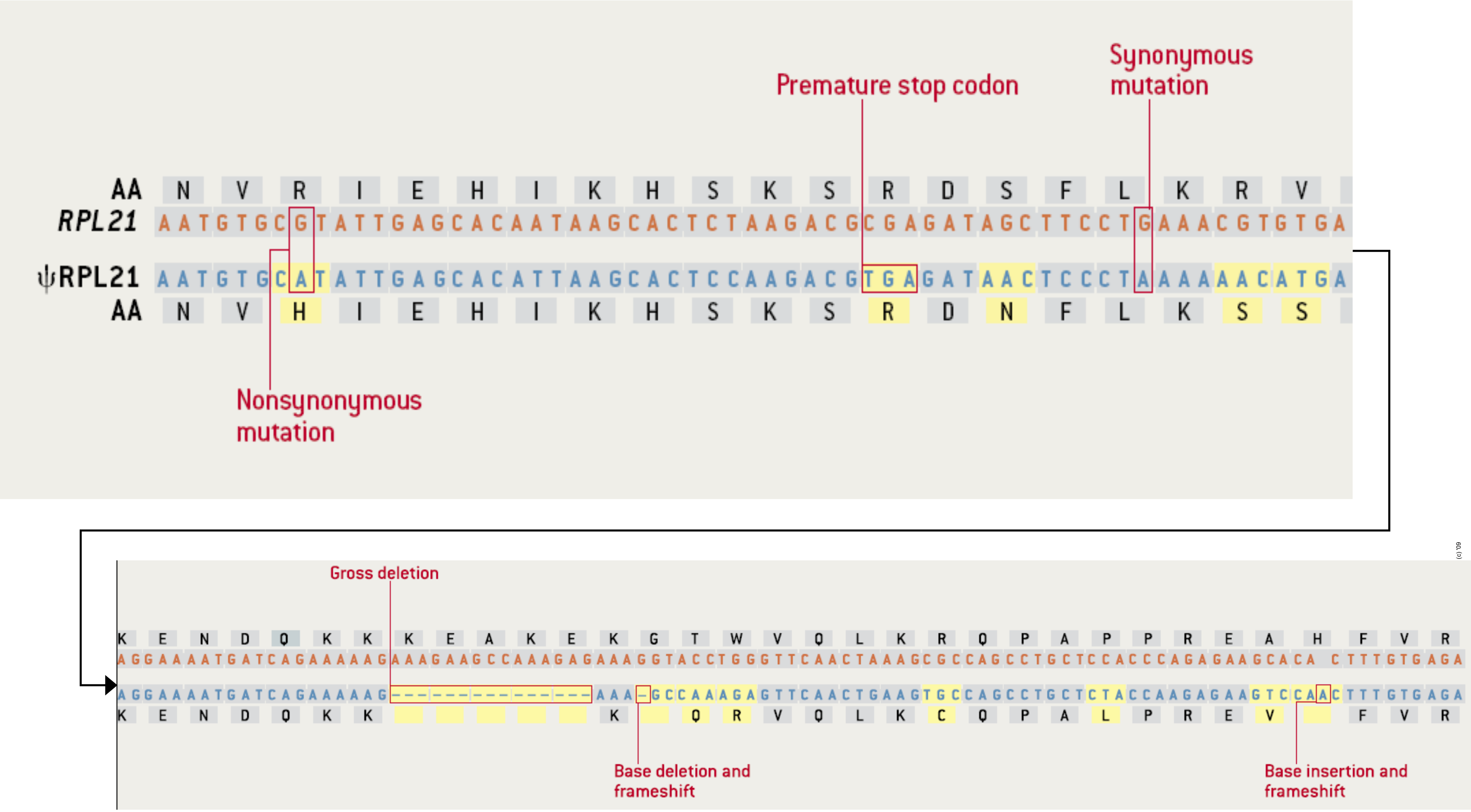
Pseudogenes & Genomic Duplications

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?…

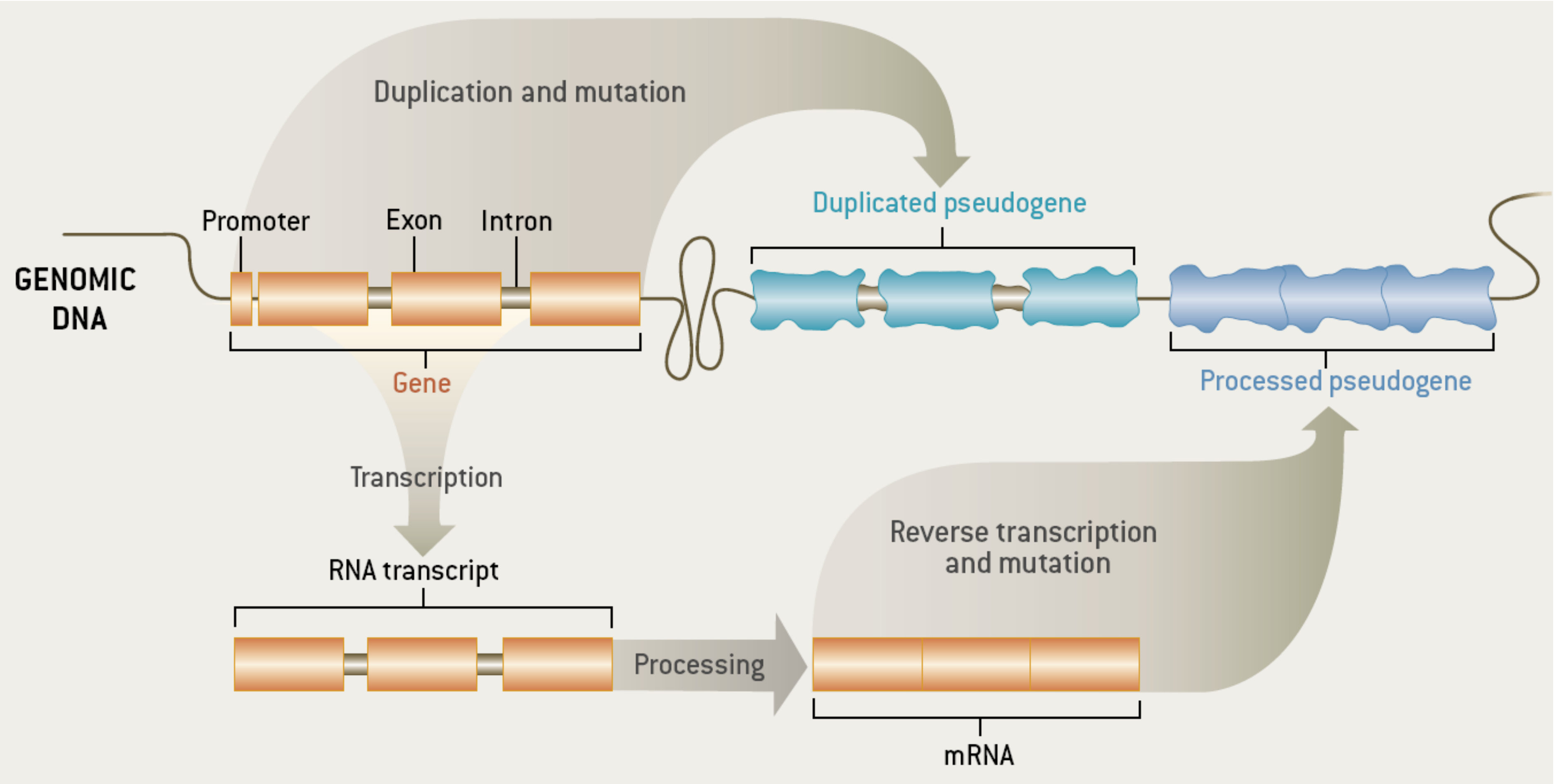
[Mighell et al. *FEBS Letts*, 2000]

Identifiable Features of a Pseudogene (ψRPL21)


Premature stop codon
Synonymous mutation
AA N V R I E H I K H S K S R D S F L K R V
RPL21 A A T G T G C G T A T T G A G C A C A A T A A G C A C T C T A A G A C G C G A G A T A G C T T C C T G A A A C G T G T G A
ψRPL21 A A T G T G C A T A T T G A G C A C A T T A A G C A C T C C A A G A C G T G A G A T A A C T C C C T A A A A A A C A T G A
AA N V H I E H I K H S K S R D N F L K S S
Nonsynonymous mutation
Gross deletion
K E N D Q K K K E A K E K G T W V Q L K R Q P A P P R E A H F V R
A G G A A A A T G A T C A G A A A A A G A A A G A A G C C A A A G A G A A A G G T A C C T G G G T T C A A C T A A A G C G C C A G C C T G C T C C A C C C A G A G A A G C A C A C T T T G T G A G A
A G G A A A A T G A T C A G A A A A A G - - - - - - - - - - - - - - - A A A - G C C A A A G A G T T C A A C T G A A G T G C C A G C C T G C T C T A C C A A G A G A A G T C C A A C T T T G T G A G A
K E N D Q K K K Q R V Q L K C Q P A L P R E V F V R
Base deletion and frameshift
Base insertion and frameshift
(c) '09

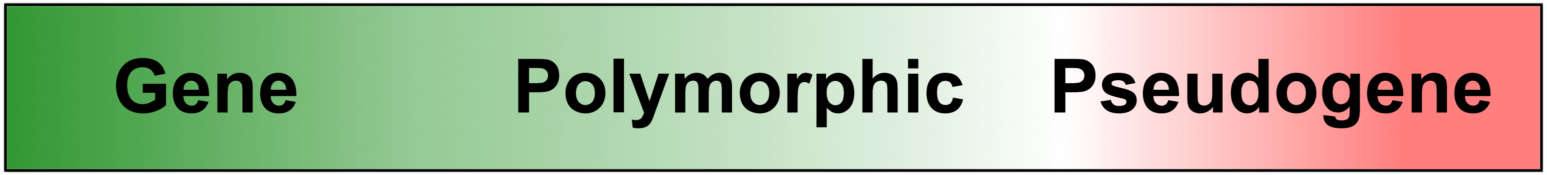
[Gerstein & Zheng. Sci Am 295: 48 (2006).]

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



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

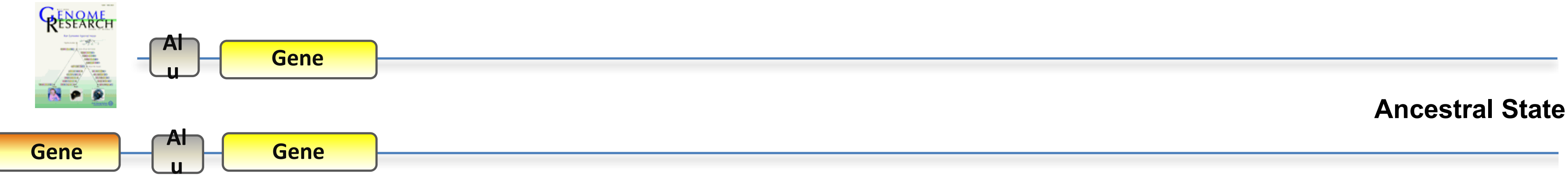
Impact of Genetic Variability: Loss-of-function


Gene Polymorphic Pseudogene

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

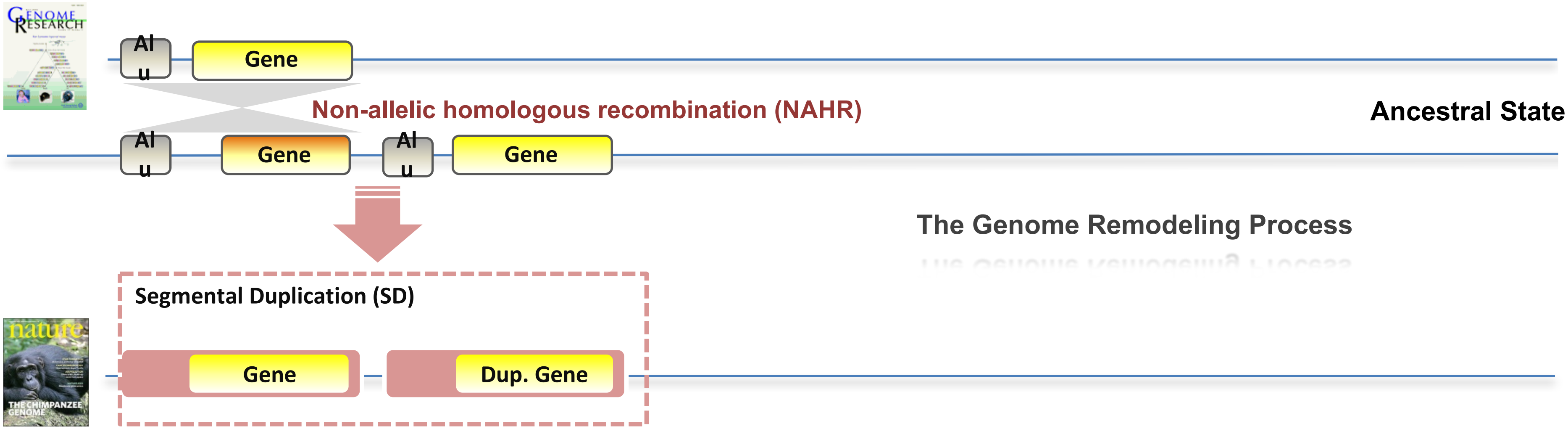
- Previous LoFs are considered as having high probability of being deleterious
- Surprisingly, ~ 100 LoF variants per genome, 20 genes are completely inactivated
- Among ~100 LoFs, we estimate 2 recessive, close to 0 dominant disease nonsense variants per healthy genome.

Genomic Variation



The Genome Remodeling Process

Genomic Variation



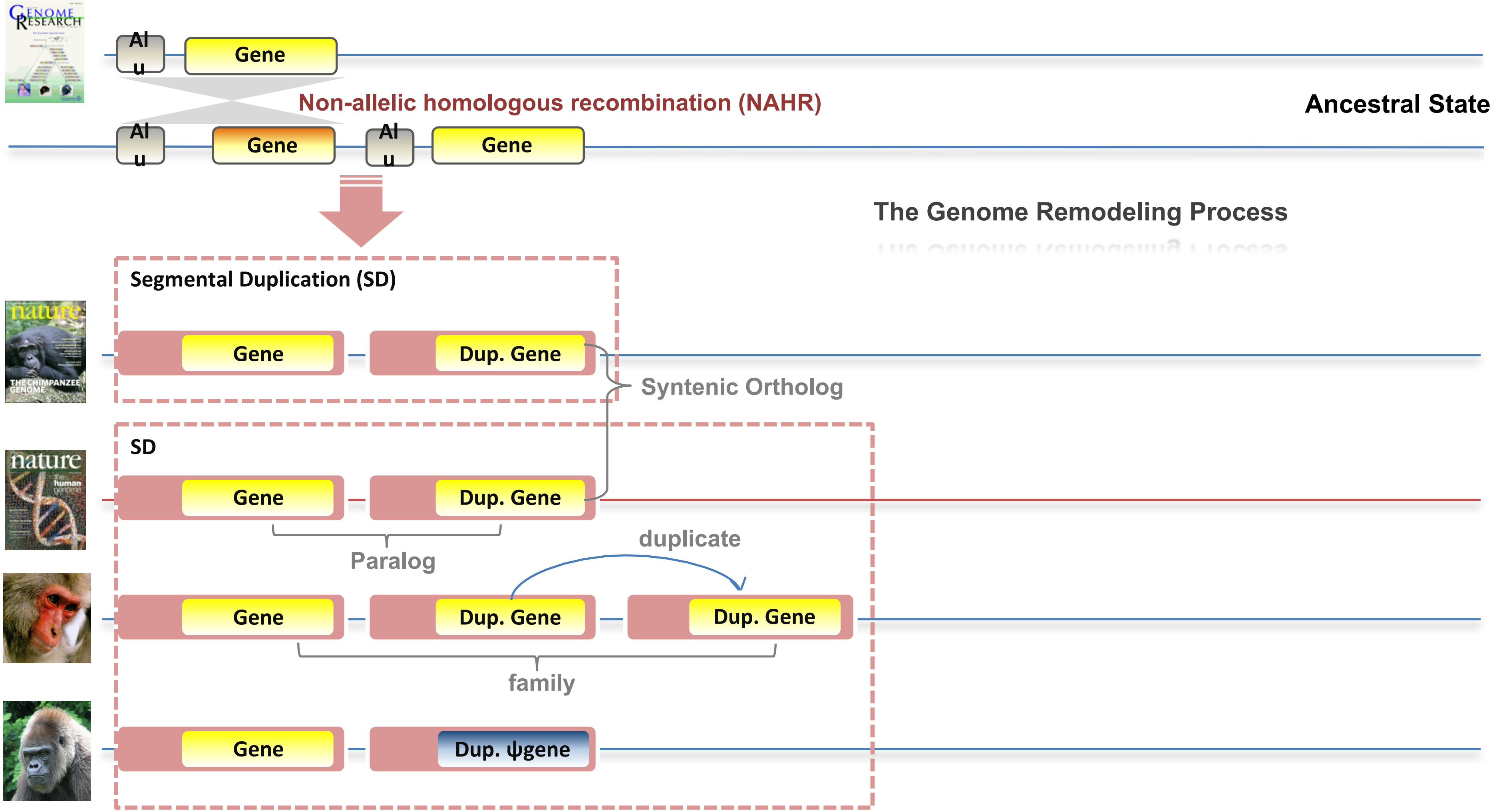
Non-allelic homologous recombination (NAHR)

Ancestral State

The Genome Remodeling Process

Segmental Duplication (SD)

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

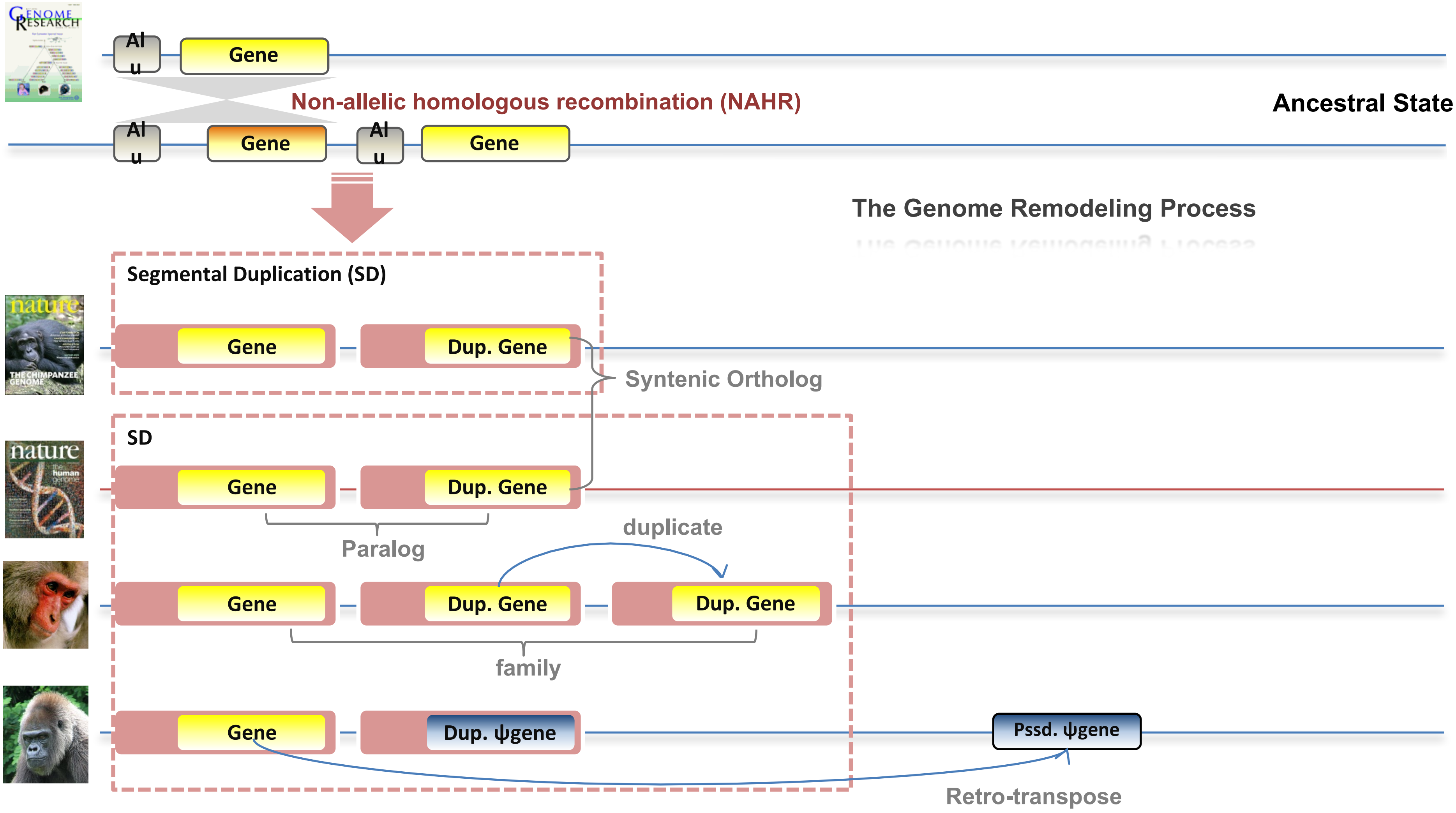
Dup. Gene

family

Gene

Dup. ψgene

Genomic Variation



Al u

Gene

Non-allelic homologous recombination (NAHR)

Ancestral State

Al u

Gene

Al u

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

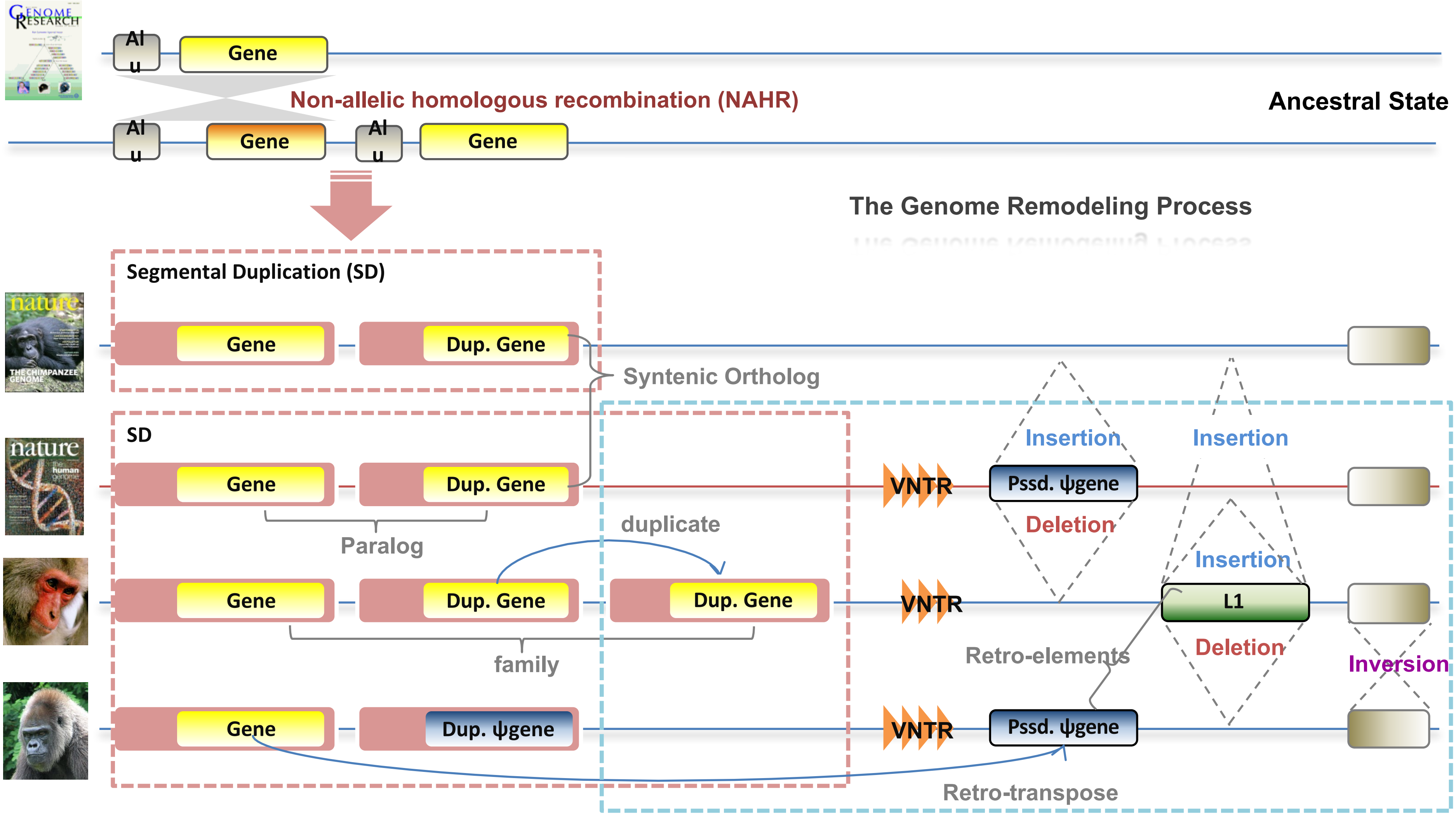
Gene

Dup. ψgene

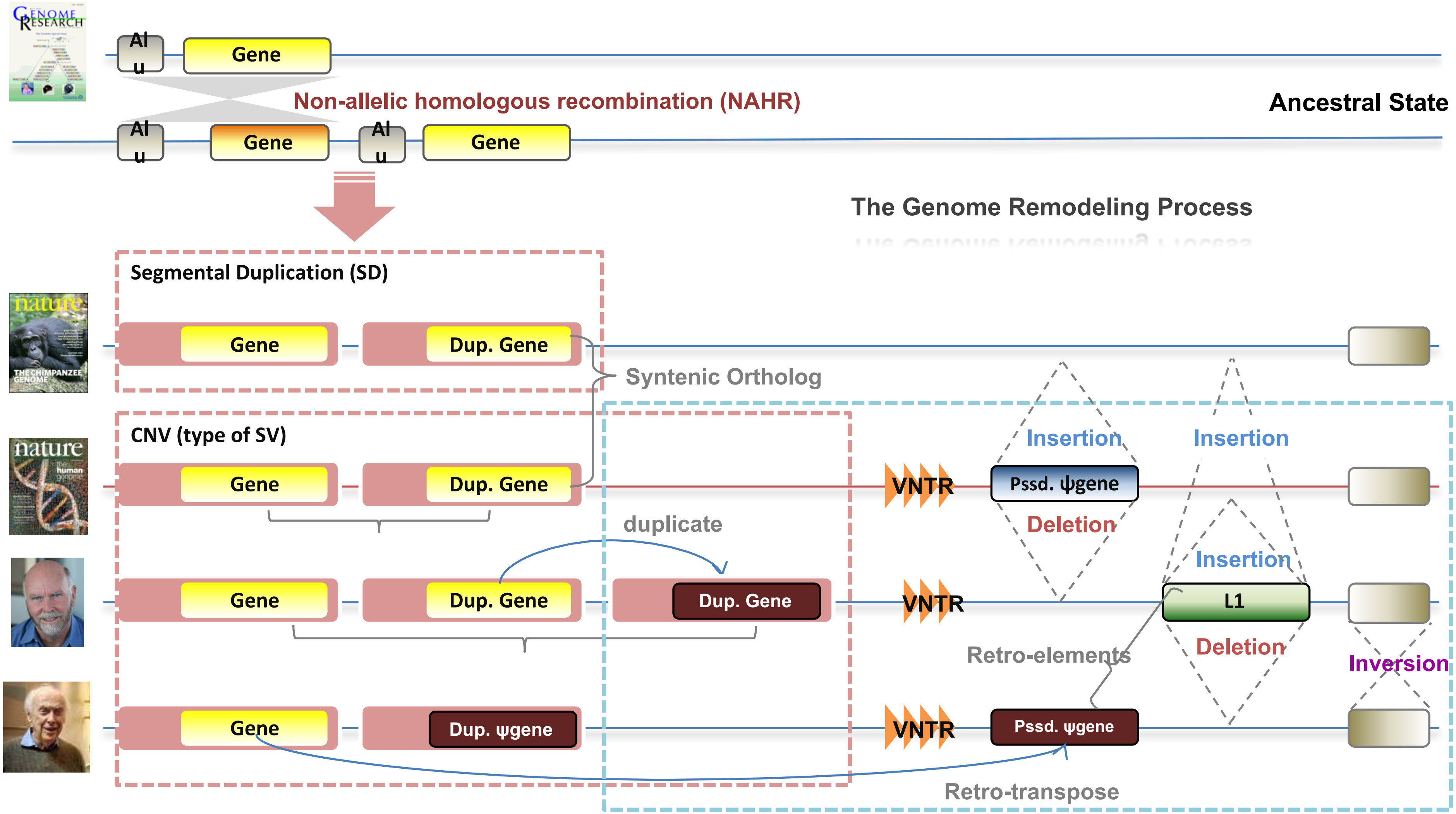
Pssd. ψgene

Retro-transpose

Genomic Variation



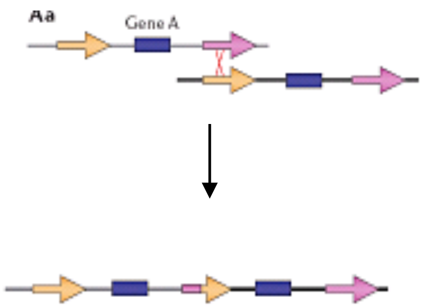
Genomic Variation



"Polymorphic" Genes & Pseudogenes

Exact Breakpoints & Mechanism Classification

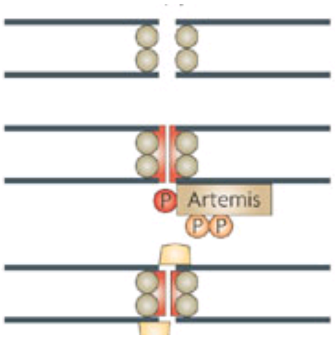
4 mechanisms for SV formation



NAHR

(Non-allelic homologous recombination)

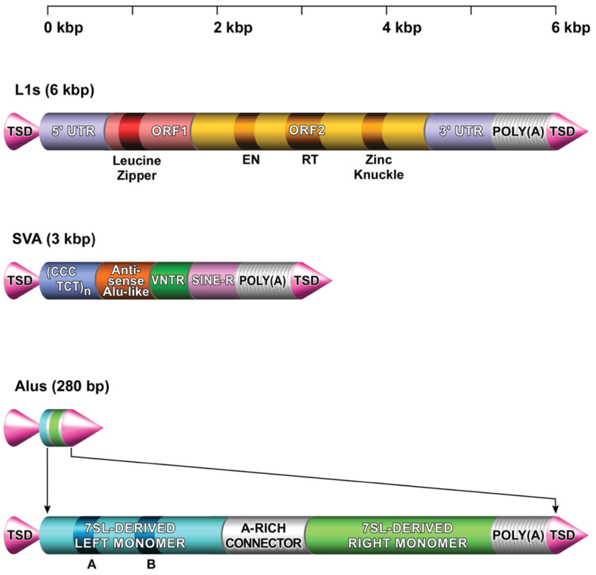
Flanking repeat
(e.g. Alu, LINE…)



NHEJ (NHR)

(Non-homologous-end-joining)

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



TEI

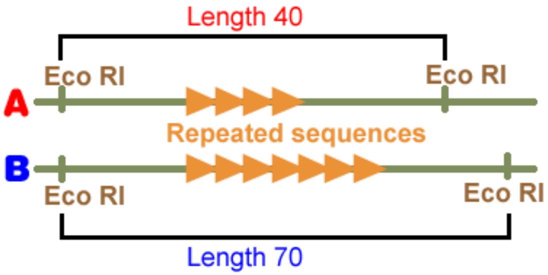
(Transposable element insertion)

L1, SVA, Alus

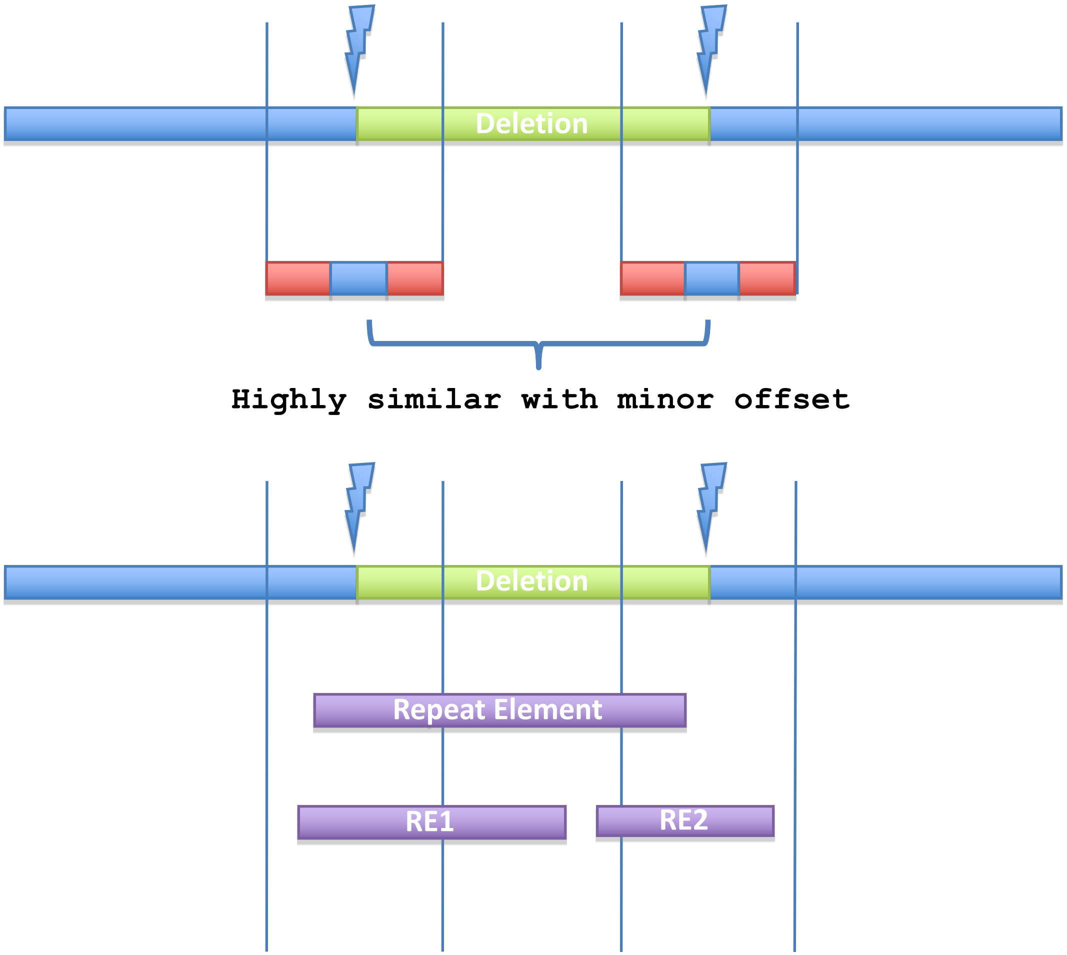
VNTR

(Variable Number Tandem Repeats)

Number of repeats varies between different people



SV Mechanism Classification



NAHR

Deletion

Highly similar with minor offset

Deletion

Single RETRO

Repeat Element

Multiple RETRO

RE1

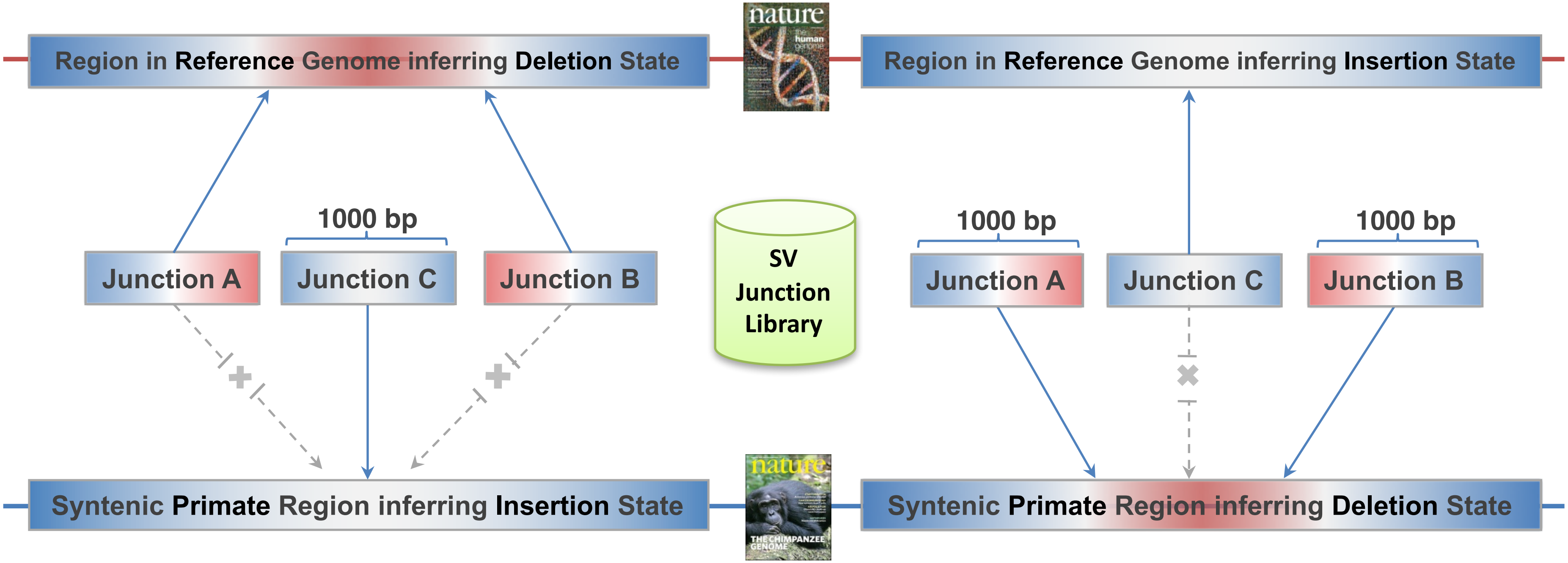
RE2

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

SV Ancestral State Analysis

Inferring Insertion according to Ancestral State

Inferring Deletion according to Ancestral State



Region in Reference Genome inferring Deletion State

1000 bp

Junction A

Junction C

Junction B

Syntenic Primate Region inferring Insertion State

SV Junction Library

Region in Reference Genome inferring Insertion State

1000 bp

Junction A

Junction C

1000 bp

Junction B

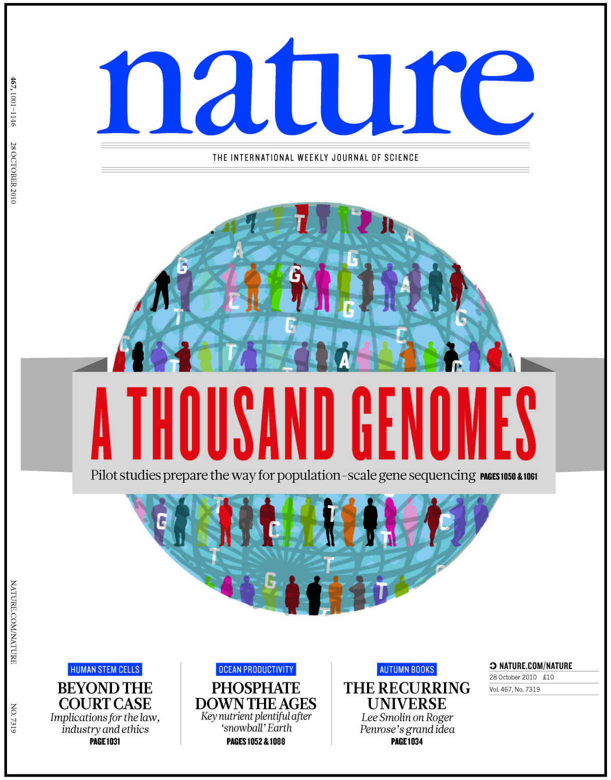
Syntenic Primate Region inferring Deletion State

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

1000G summary

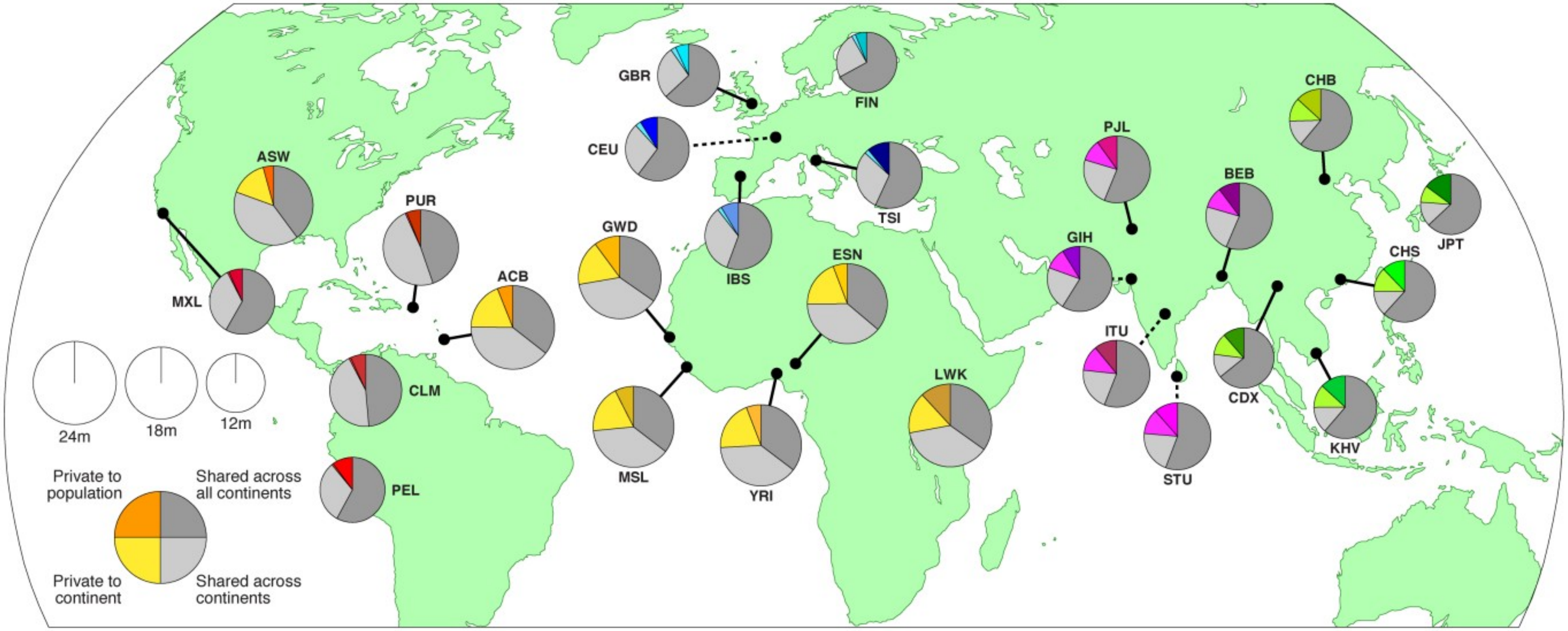
1000G SV
(Pilot, **Phase I** & III)

- **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


nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061
HUMAN STEM CELLS
BEYOND THE COURT CASE
Implications for the law, industry and ethics
PAGE 1031
OCEAN PRODUCTIVITY
PHOSPHATE DOWN THE AGES
Key nutrient plentiful after 'snowball' Earth
PAGES 1052 & 1088
AUTUMN BOOKS
THE RECURRING UNIVERSE
Lee Smolin on Roger Penrose's grand idea
PAGE 1034
NATURE.COM/NATURE
28 October 2010 £10
Vol. 467, No. 7319

[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

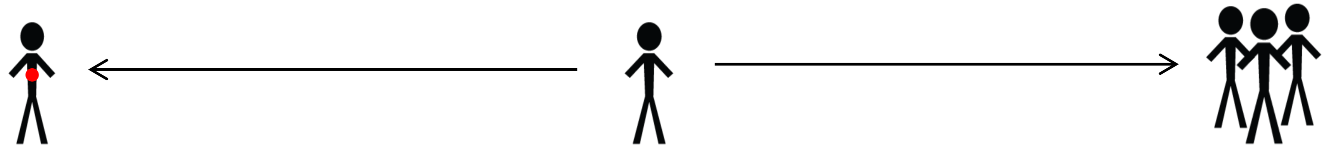
[2] 1000GP Phase3 SV paper. Submid ed to Nature, 2015.
[3] 1000GP ConsorSu m. Submid ed to Nature, 2015.

Human Genetic Variation

A Cancer Genome

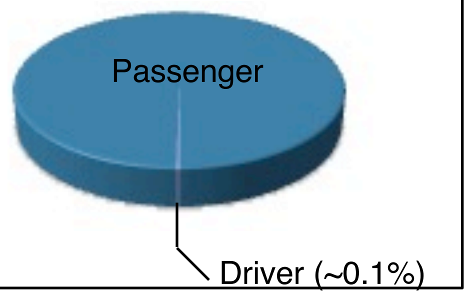
A Typical Genome

Population of
2,504 peoples



Origin of Variants

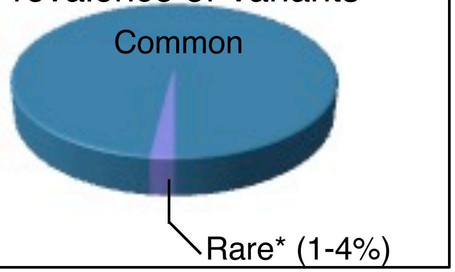
	Coding	Non-coding
Germ-line	22K	4.1 – 5M
Somatic	~50	5K



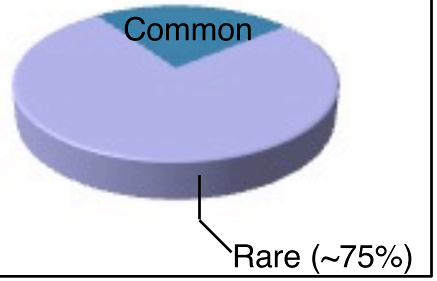
Class of Variants

SNP	3.5 – 4.3M
Indel	550 – 625K
SV	2.1 – 2.5K (20Mb)
Total	4.1 – 5M

Prevalence of Variants



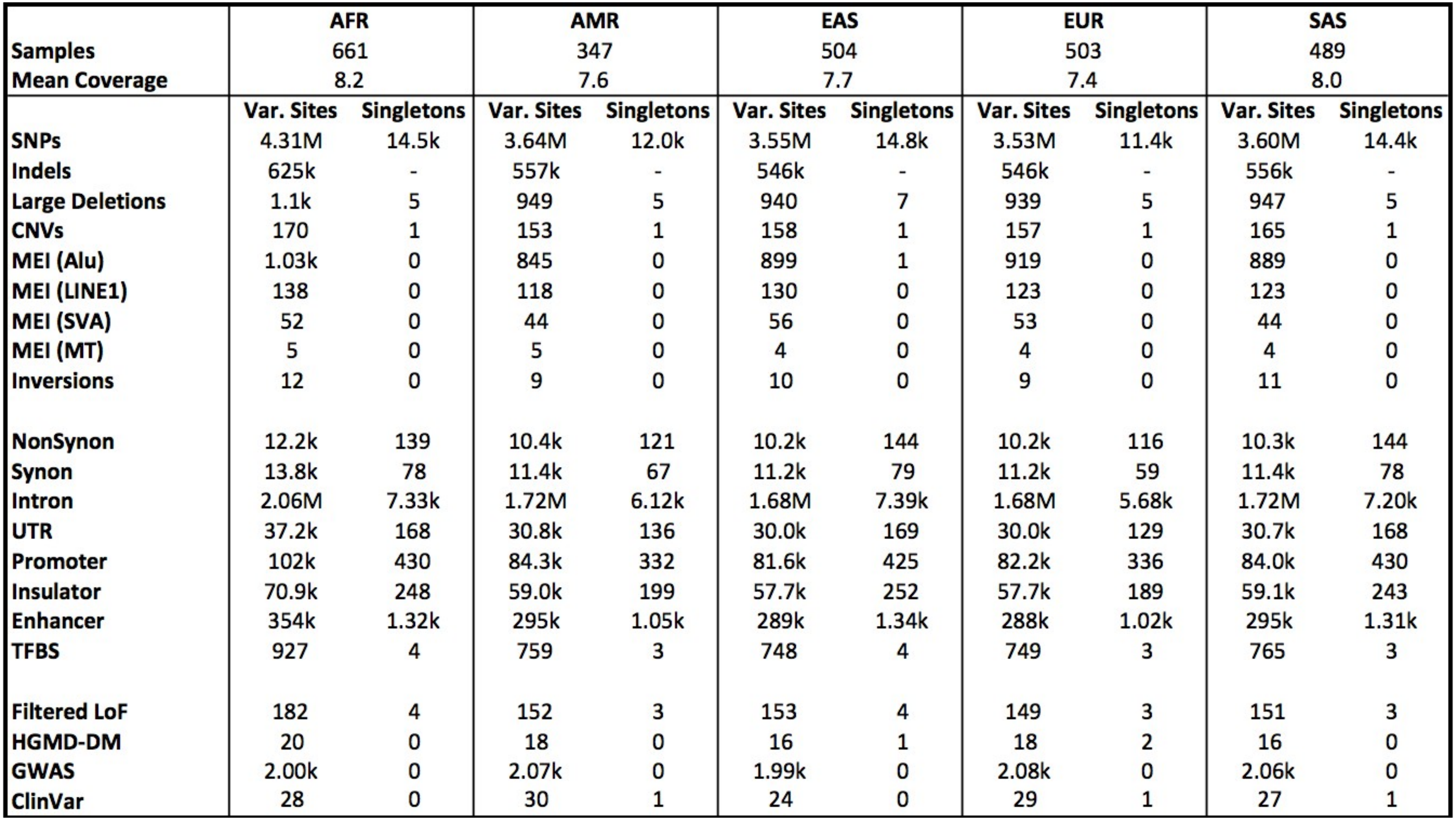
SNP	84.7M
Indel	3.6M
SV	60K
Total	88.3M



* 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

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

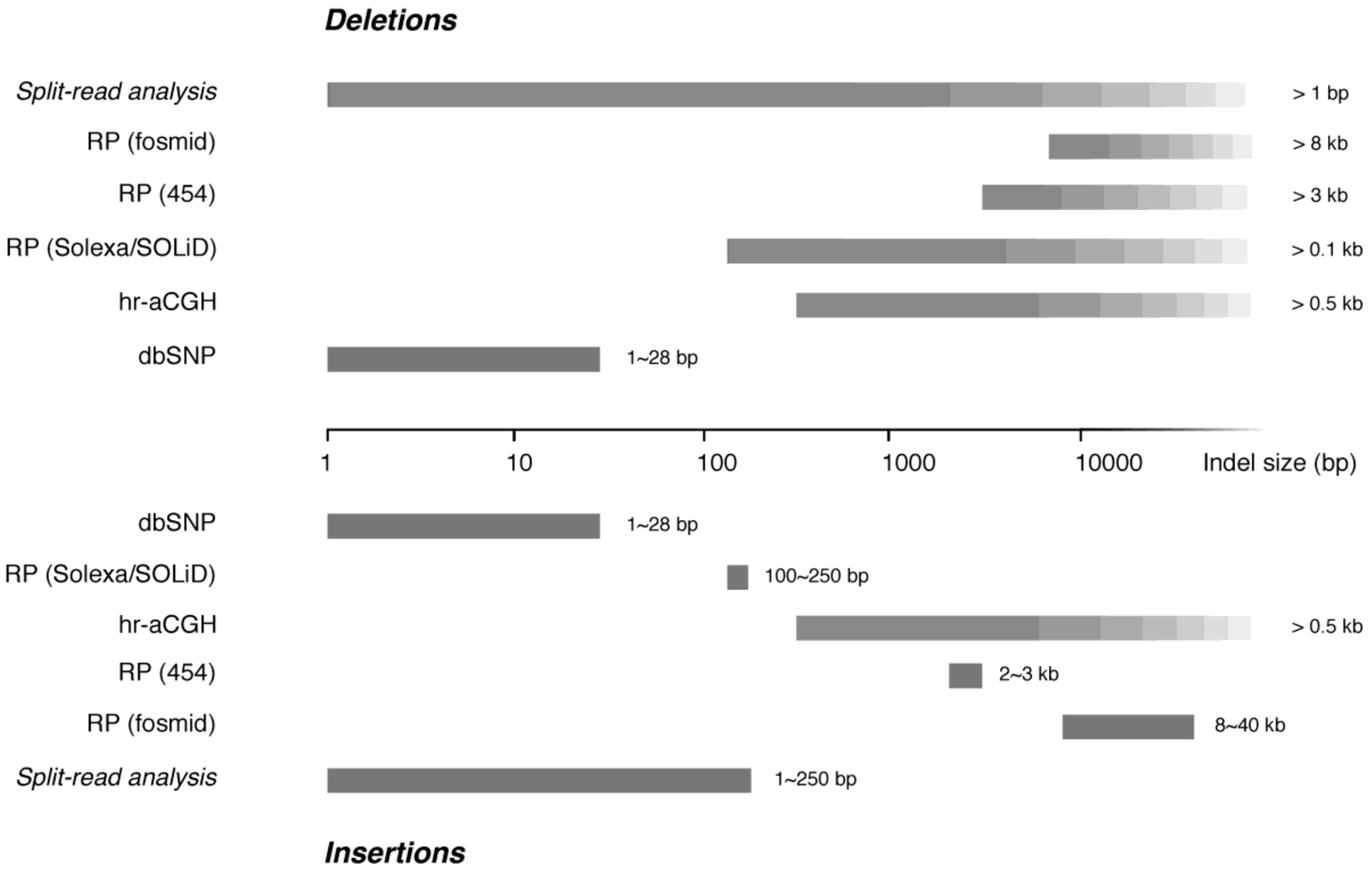
[3] 1000GP Consortium. Submitted to Nature,

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 ConsorSum. Submided to Nature, 2015.

Different Approaches Work Differently on Different Events



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