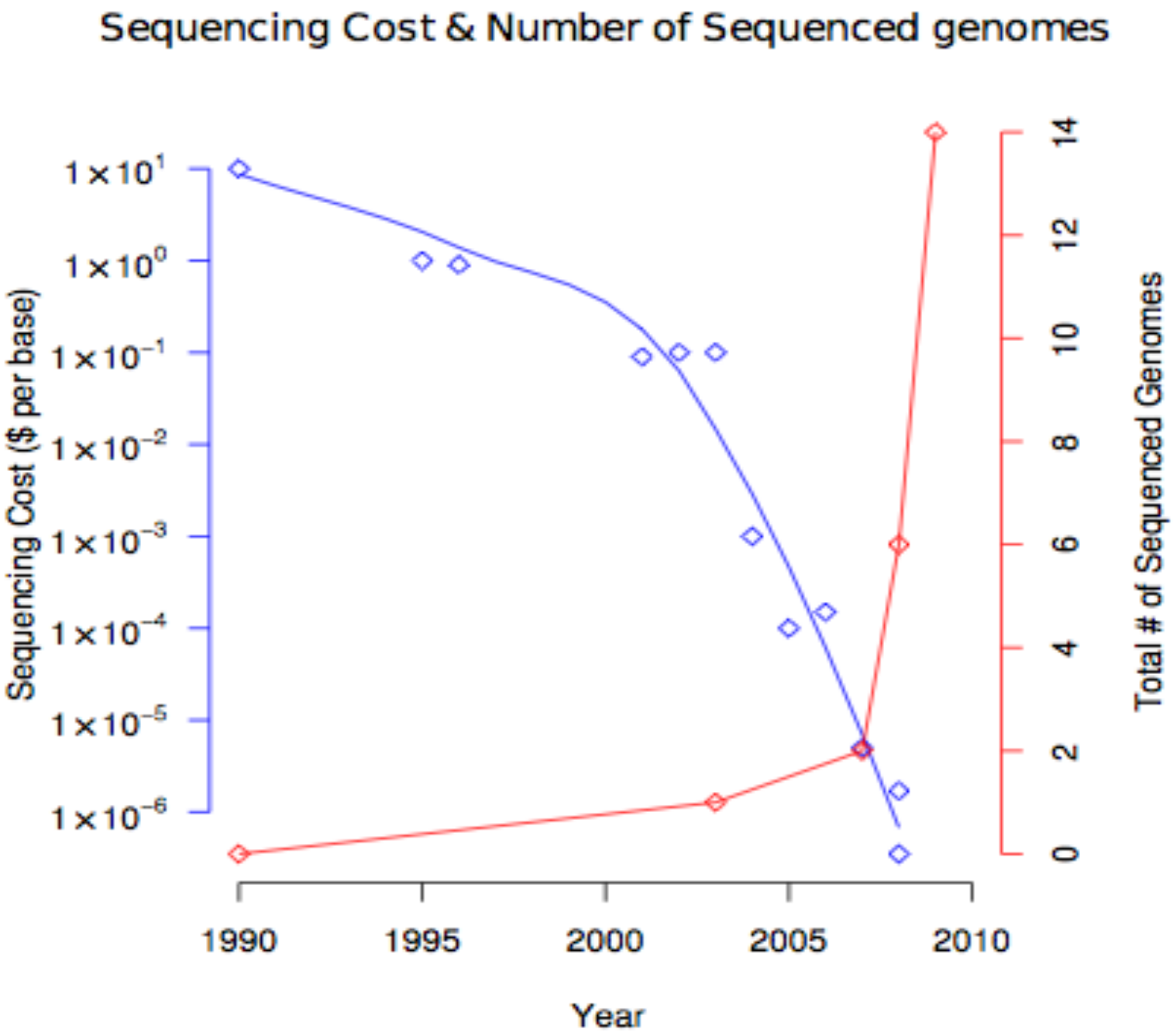
Structural Variation in the Human Genome

Michael Snyder

March 2, 2010


Sequencing Cost & Number of Sequenced genomes
Sequencing Cost ($ per base)
1×10^1
1×10^0
1×10^-1
1×10^-2
1×10^-3
1×10^-4
1×10^-5
1×10^-6
Total # of Sequenced Genomes
0
2
4
6
8
10
12
14
1990
1995
2000
2005
2010
Year

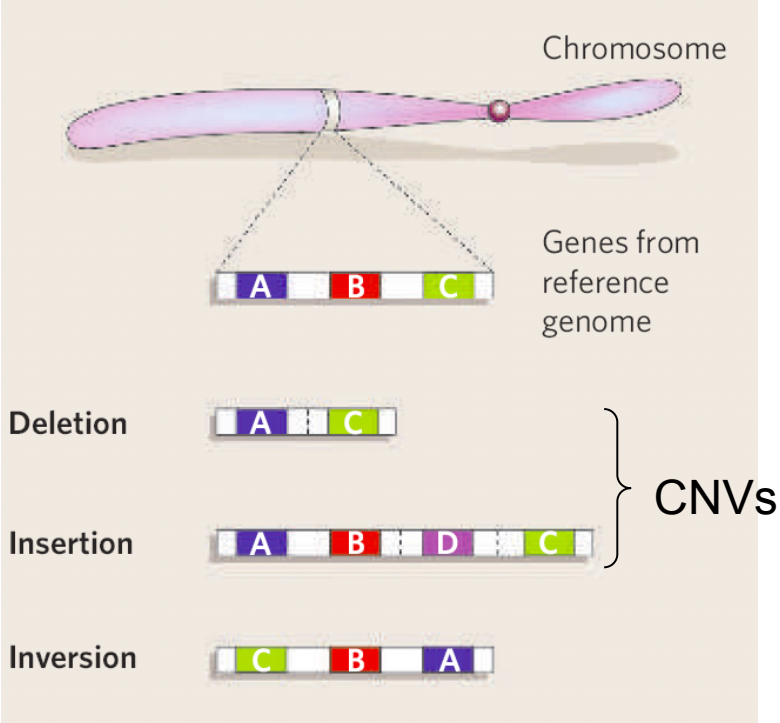
Genetic Variation Among People

Single nucleotide polymorphisms (SNPs)

GATTTAGATCGCGATAGAG
GATTTAGATCTCGATAGAG

0.1% difference among people

Mapping Structural Variation in Humans
>1 kb segments



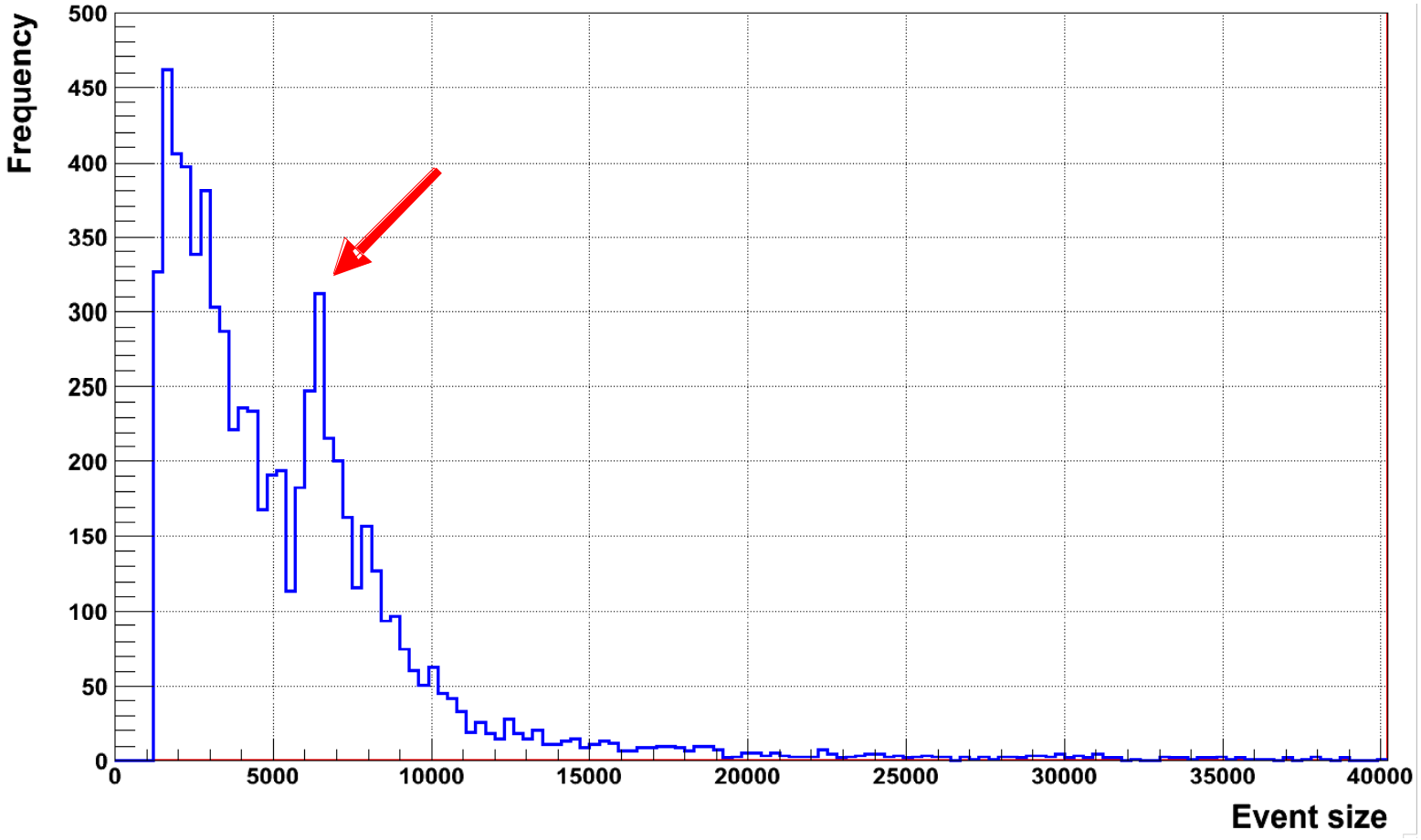
- Thought to be Common
  12% of the genome
  (Redon et al. 2006)

- Likely involved in phenotype variation and disease

- Until recently most methods for detection were low resolution (>50 kb)




Size Distribution of CNV in a Human Genome



Why Study Structural Variation?

- Common in “normal” human genomes--major cause of phenotypic variation
- Common in certain diseases, particularly cancer
- Now showing up in rare disease; autism, schizophrenia

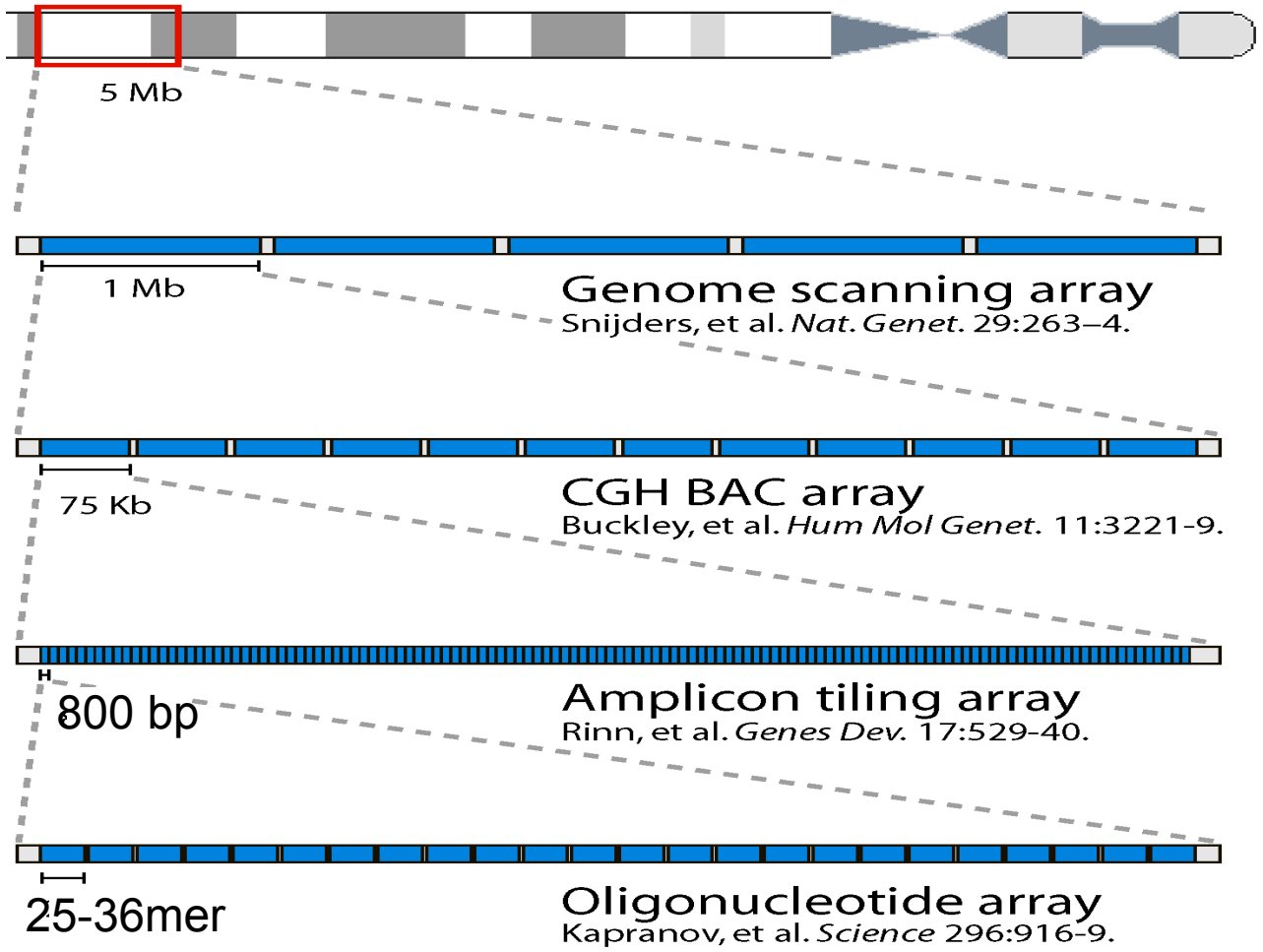
Most Genome Sequencing Projects Ignore SVs

Project	Technology	Paired End	SNPs; Short Indel	SVs	New Seq.	Genotype	Reference
European-Venter	Sanger	Yes	3M; 0.3M	0.2M (> 1000bp)	1M	Limited	Levy et al., 2007
European-Watson	454	No	3M; 0.2M	Limited	No	No	Wheeler et al., 2008
European-Quake	Helicos	No	3M	Limited	No	No	Pushkarev et al., 2009
Asian	Illumina	Partially	3M; 0.1M	2.7K (>100bp)	No	No	Wang et al., 2008
HapMap Sample; Yoruban 18507	Illumina	Yes	4M; 10K	0.1K	No	No	Bentley et al., 2008
HapMap Sample; Yoruban 18507	SOLiD	Partially	4M; 0.2M	5.5K (unknown definition)	No	No	McKernan et al., 2009
Korean	Illumina	Yes	3M	Limited	No	No	Ahn et al., 2009
Korean- AK1	Illumina	Yes	3.45M; 0.17M	~300 CNVs	No	No	Kim et al., 2009
Three human genomes	Complete Genomics	Yes	3.2-4.5M; 0.3-0.5M	Limited (50-90K block substitutions)	No	Limited	Drmanac et al., 2009
AML genome & normal counterpart	Illumina	No	3.8M; 0.7K	Limited	No	No	Ley et al., 2008
AML genome	Illumina	Yes	64	Limited	No	No	Mardis et al., 2009
Melanoma genome	Illumina	Yes	32K;1K	51	No	No	Pleasance et al., 2009a
Lung cancer genome	SOLiD	Yes	23K; 65	392	No	No	Pleasance et al. 2009b

Why Not Studied More?

- Often involves repeated regions
- Rearrangements are complex
- Can involve highly repetitive elements

Genome Tiling Arrays



High-Resolution CGH with Oligonucleotide Tiling Microarrays

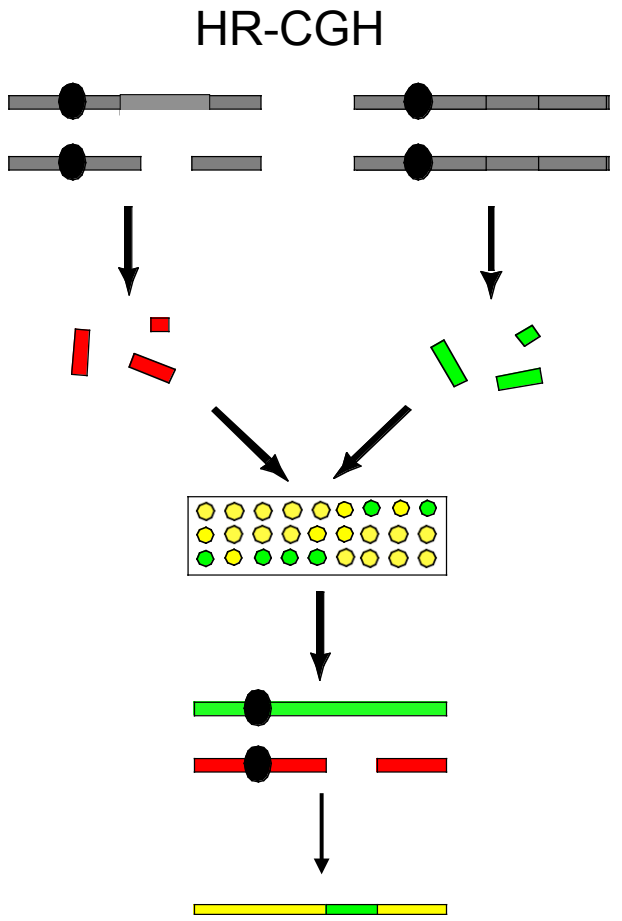
Maskless Array Synthesis

385,000 oligomers/chip

Isothermal oligomers, 45-85 bp

Tiling at ~1/100bp non-repetitive genomic sequence

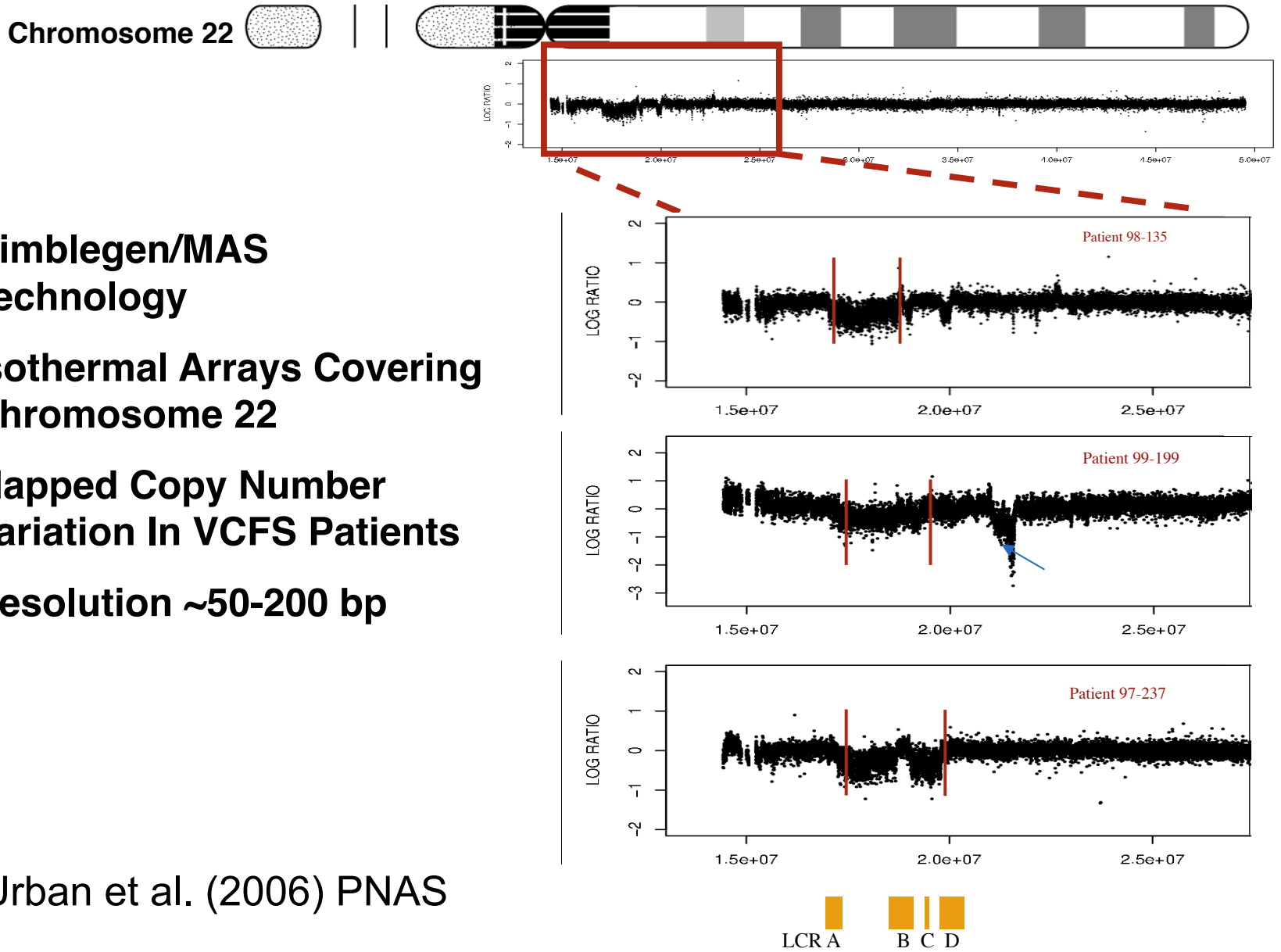
Detects CNVs at <1 kb resolution



Urban et al., 2006

With R. Selzer and R. Green

High Resolution Comparative Genomic Hybridization



Nimblegen/MAS Technology

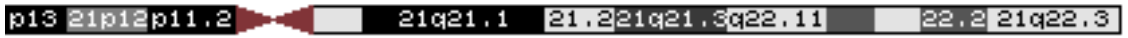
Isothermal Arrays Covering Chromosome 22

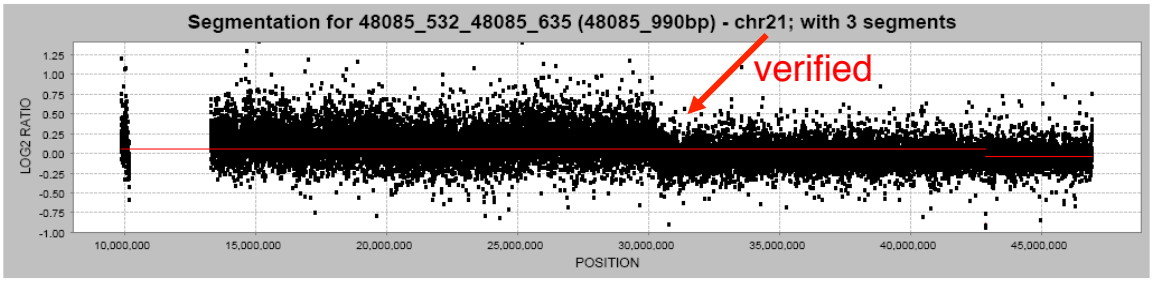
Mapped Copy Number Variation In VCFS Patients

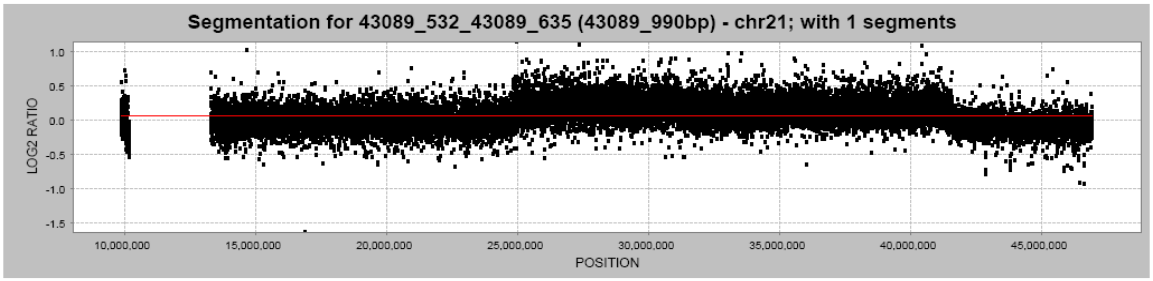
Resolution ~50-200 bp

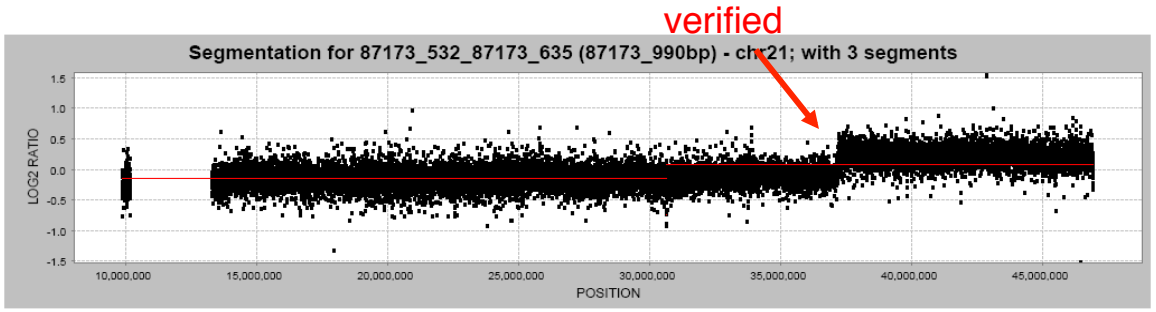
Urban et al. (2006) PNAS

Mapping Breakpoints of Partial Trisomies of Chromosome 21



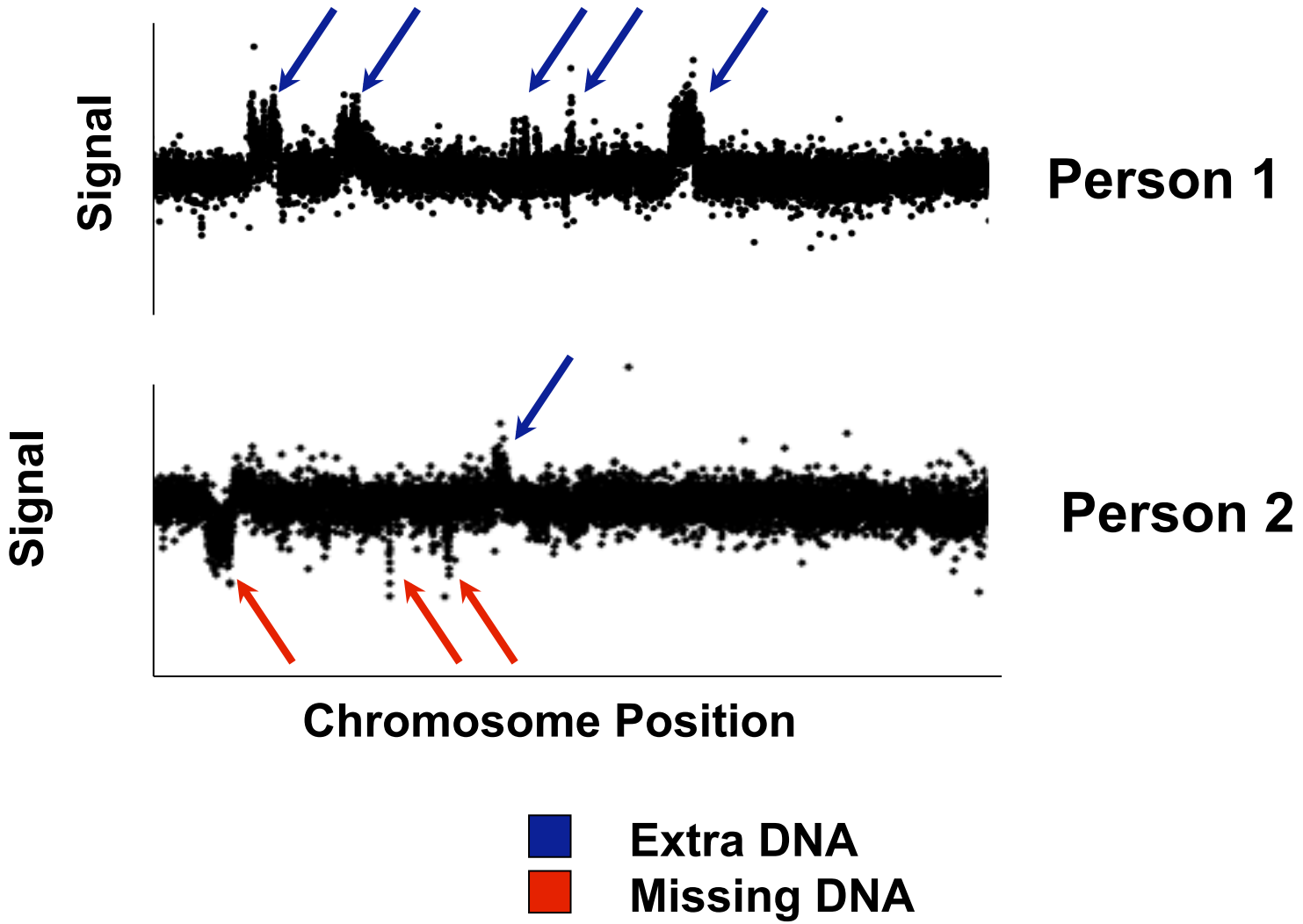




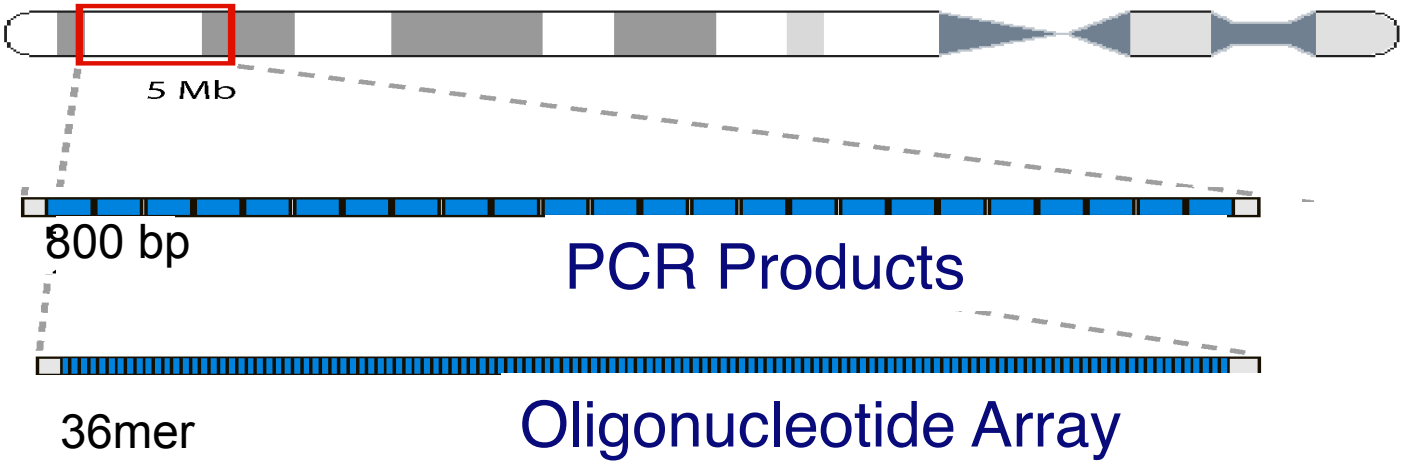


With Korenberg Lab, UCLA

Copy Number Variations in the Human Genome



Genome Tiling Arrays



Massively Parallel Sequencing

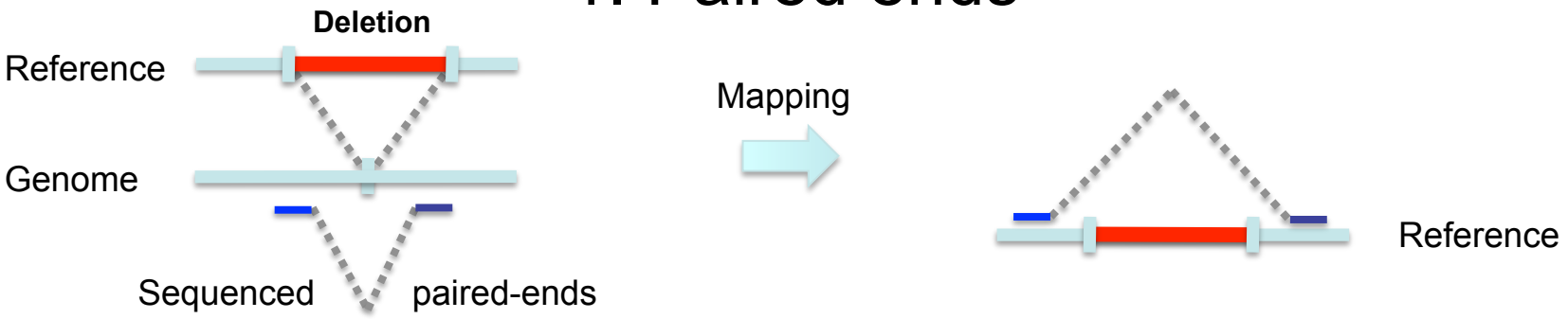
AGTTCACCTAAGA…

CTTGAATGCCGAT…

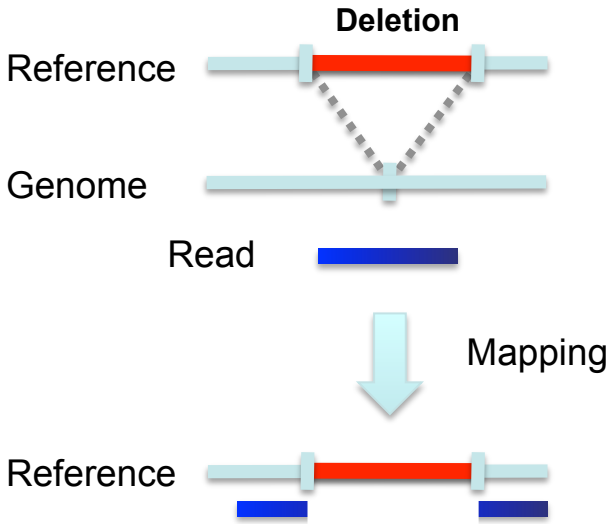
GTCATTCCGCAAT…

High Throughput DNA Sequencing based Methods to detect CNVs/SVs

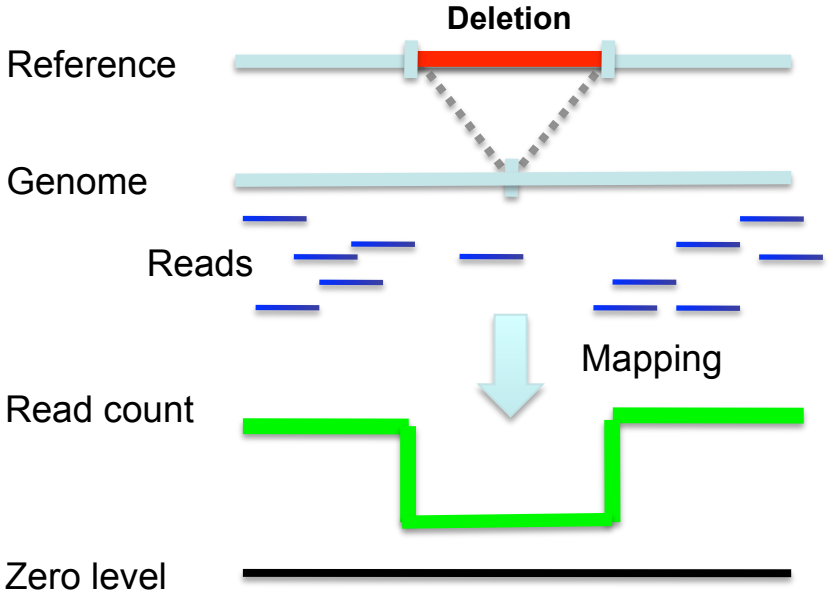
1. Paired ends



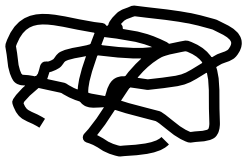
3. Split read



2. Read depth



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



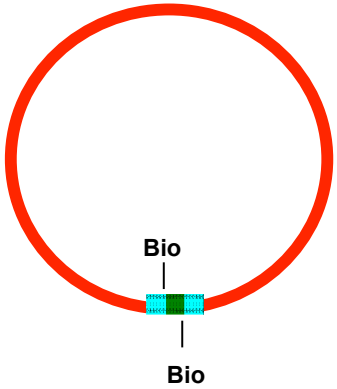
Genomic DNA

Shear to 3 kb

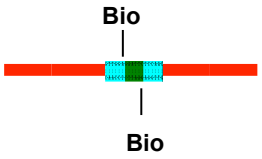
Adaptor ligation

Fragments

Circularize



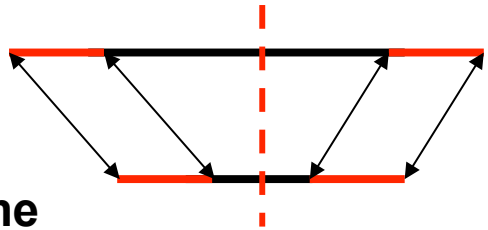
Random Cleavage



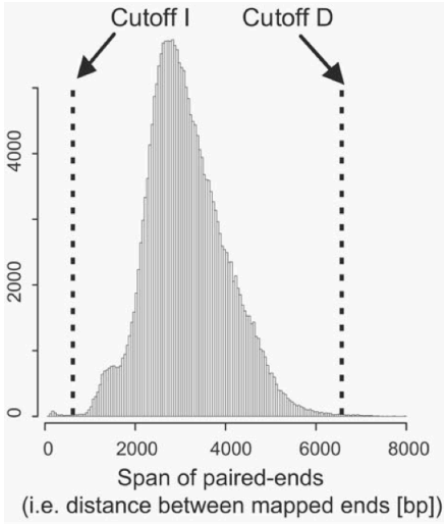
200-300bp

***454* Sequencing**

(250bp reads, 400K reads/run)



Map paired ends to reference genome



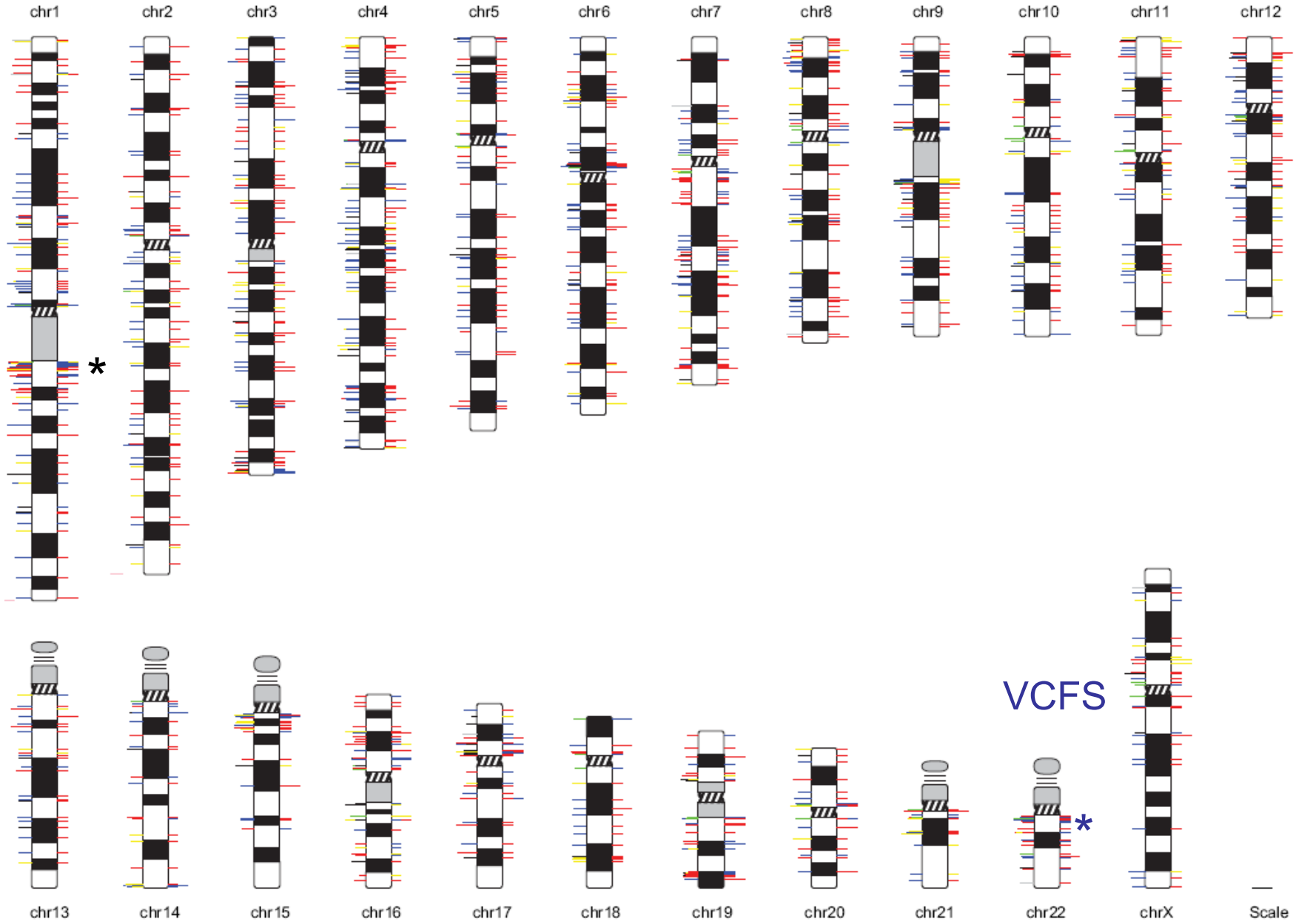
Korbel et al., 2007 Science

Summary of PEM Results

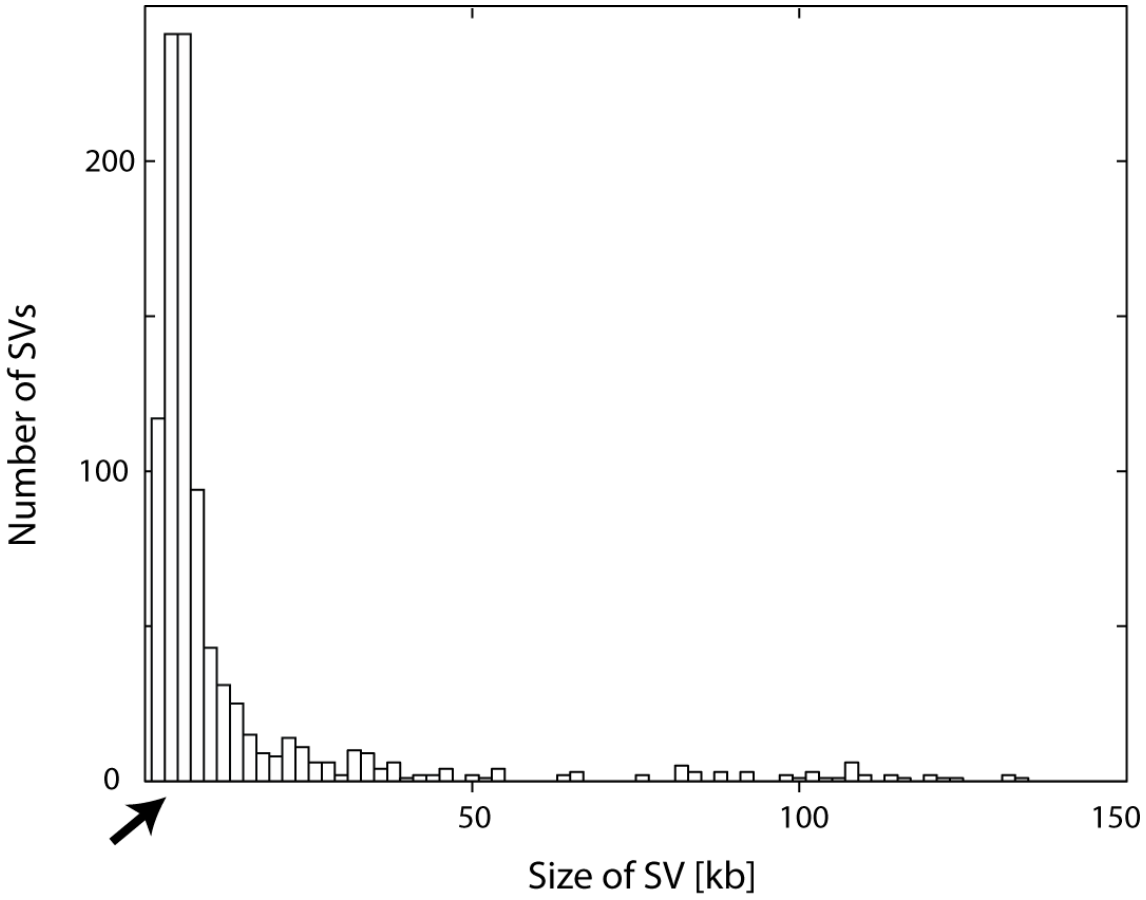
	NA15510 (European?, female)	**NA18505 (Yoruba, female)**
# of sequence reads	> 10 M.	> 21 M.
Paired ends uniquely mapped	> 4.2 M.	> 8.6 M.
Fold coverage	~ 2.1x	~ 4.3x
Predicted Structural Variants*	**473**	**825**
Indels	*422*	*753*
Inversion breakpoints	*51*	*72*
Estimated total variants* genome-wide	**759**	**902**

*at this resolution

~1000 SVs >2.5kb per Person

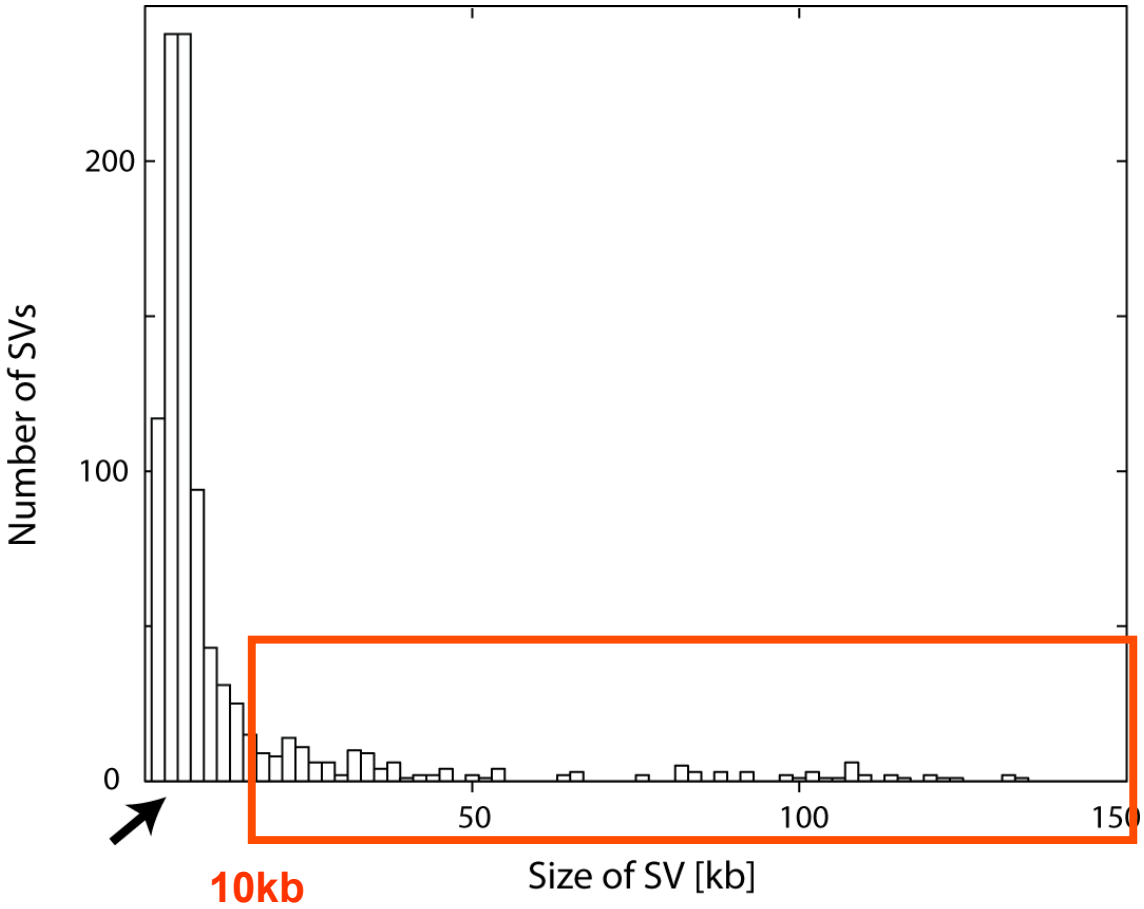


Size distribution of Structural Variants



[Arrow indicates lower size cutoff for deletions]

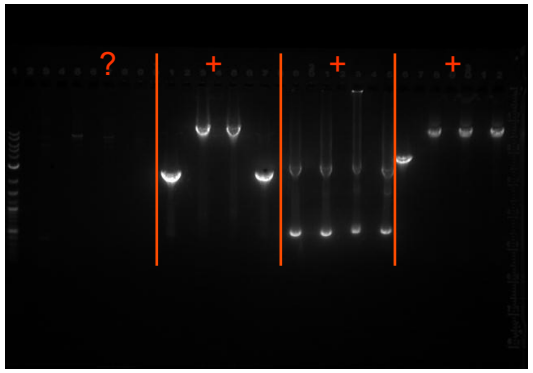
Size distribution of Structural Variants



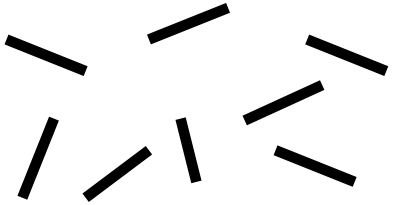
[Arrow indicates lower size cutoff for deletions]

High Throughput Sequencing of Breakpoints

PCR SVs

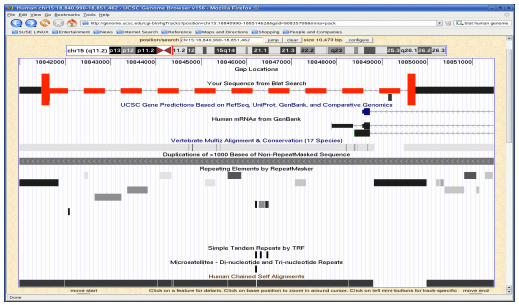


Cut Gel Bands and Pool



Shotgun-sequence PCR Mixture Using 454

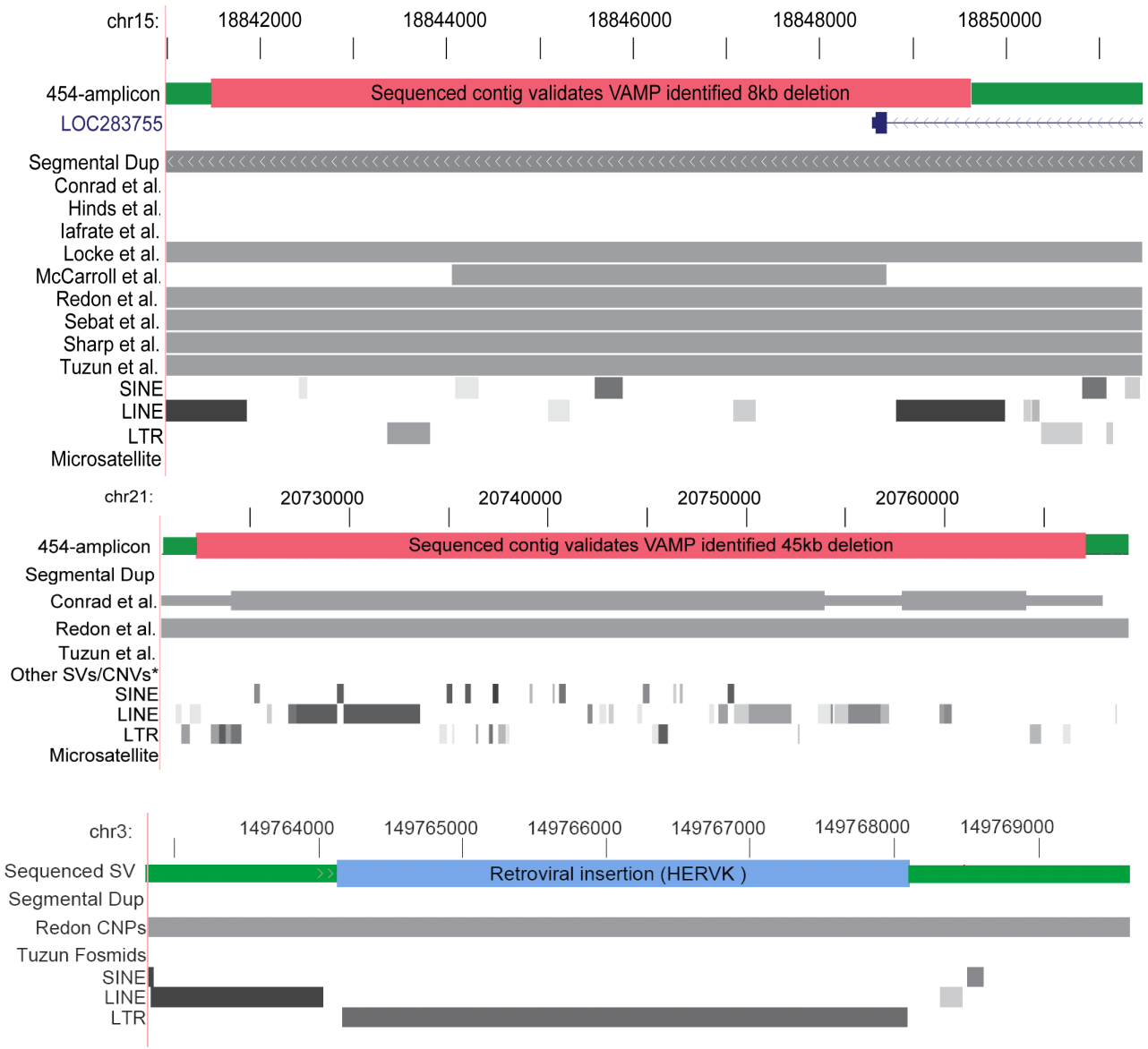
Assemble contigs and determine breakpoints



Genome Sequencer FLX

>200 SVs Sequenced Across Breakpoints

Analysis of Breakpoints

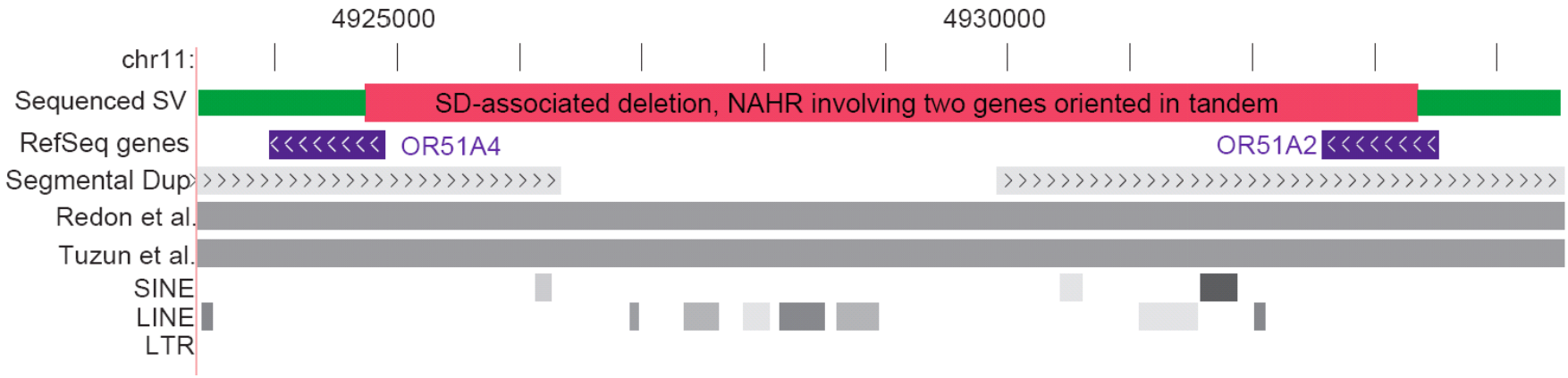


Homologous Recombination 14%

Nonhomologous Recombination 56%

Retrotransposons 30%

17% of SVs Affect Genes



Non-allelic homologous recombination (NAHR; breakpoints in OR51A2 and OR51A4)

← M P E H L S P E T K ← ← M P G H L S P E T K ←

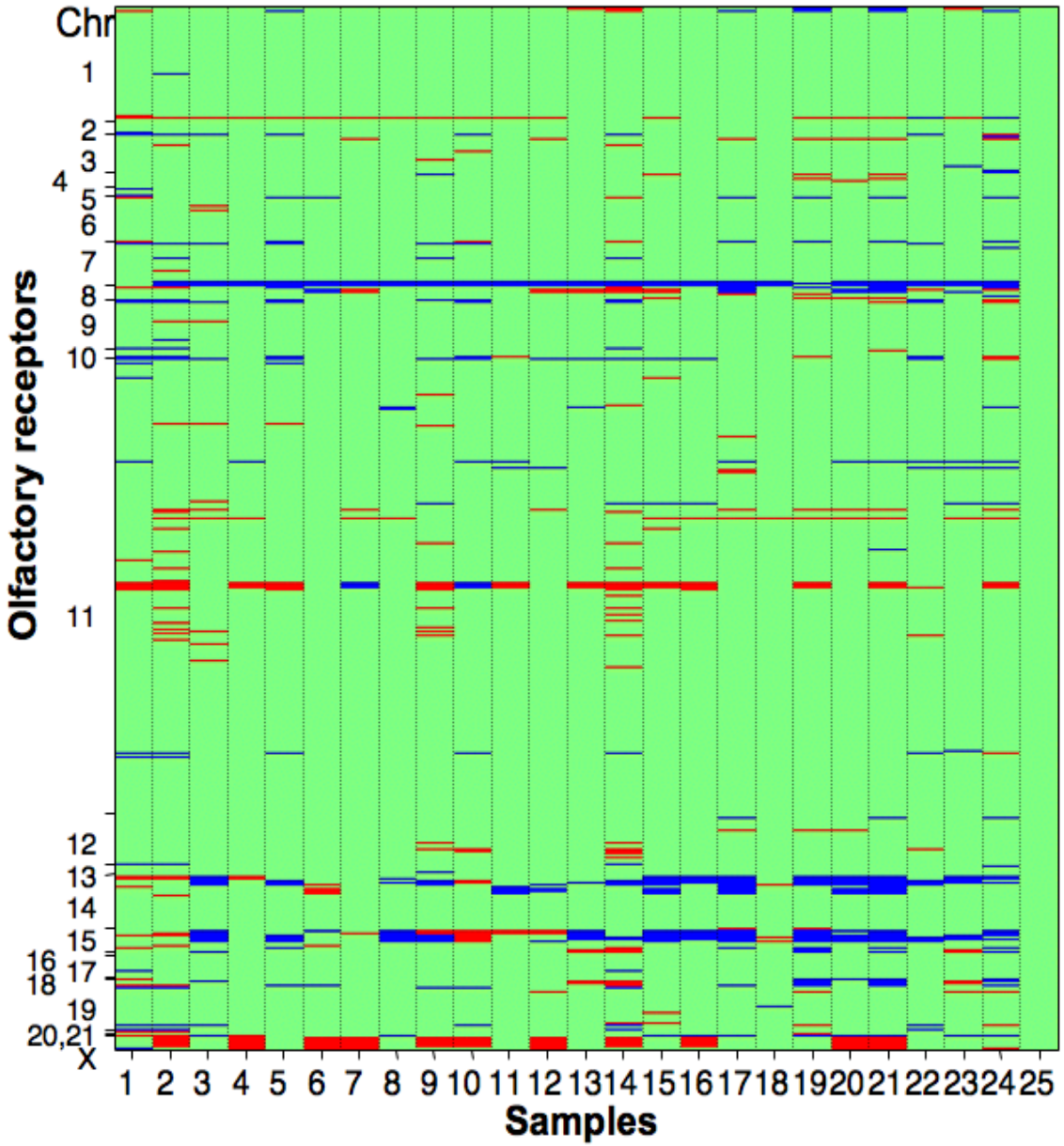
CATGGGCTCATGCAAGgagggctctgtctt - - - catgggcccatgcaagGAGGGCTCTGTCTT

chr11 4924724 4933366

Olfactory Receptor Gene Fusion

Heterogeneity in Olfactory Receptor Genes

(Examined 851 OR Loci)



Gain

Loss

No change

CNVs affect:

93 Genes

151 genes

Paired-end

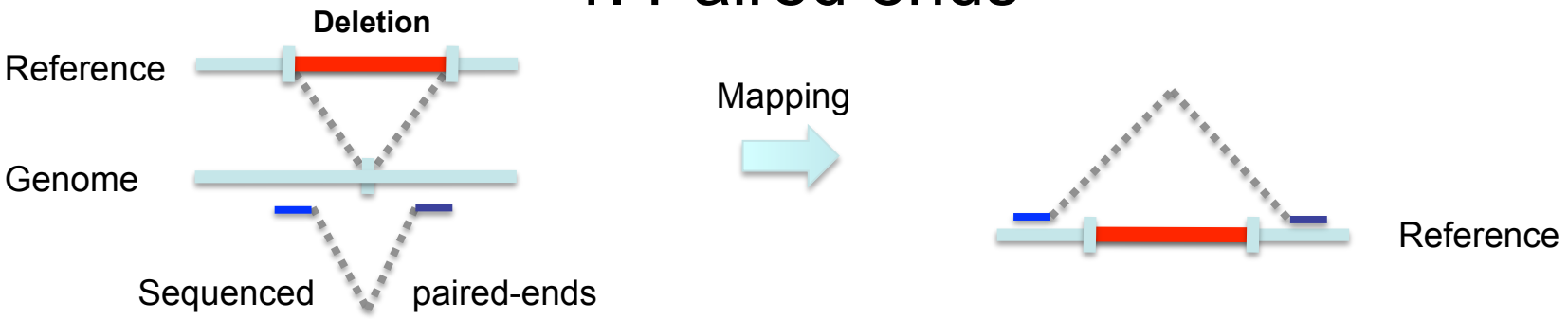
- Variations of the method are available for many platforms: Roche, Illumina, LifeTechnologies
- Long reads are preferable for optimal detection
- Can get different sizes
  - Roche 20 kb, 8kb, 3 kb
  - Ilumina, SOLiD 1.5 kb

Paired-end: Advantages/ Disadvantages

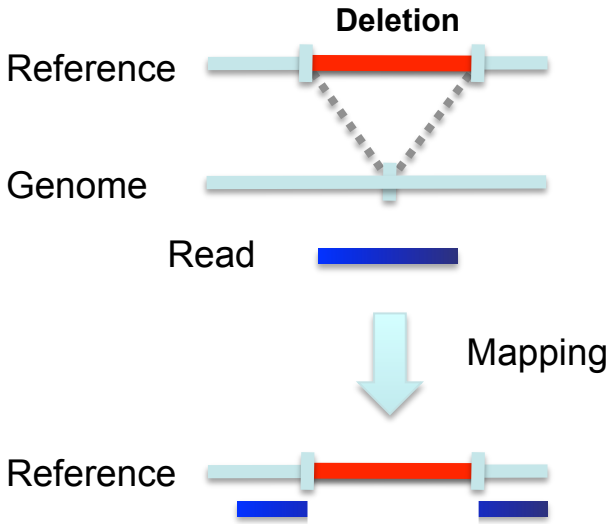
- Can detect highly repetitive CNVs (LINE, SINE, etc.)
- Detect inversions as well as insertions and deletions
- Defines location of CNV
- **Relies on confident independent mapping of each end, problems in regions flanked by repeats**
- Small span between ends limits resolution of complex regions
- Large span between ends limits resolution of break points

High Throughput DNA Sequencing based Methods to detect CNVs/SVs

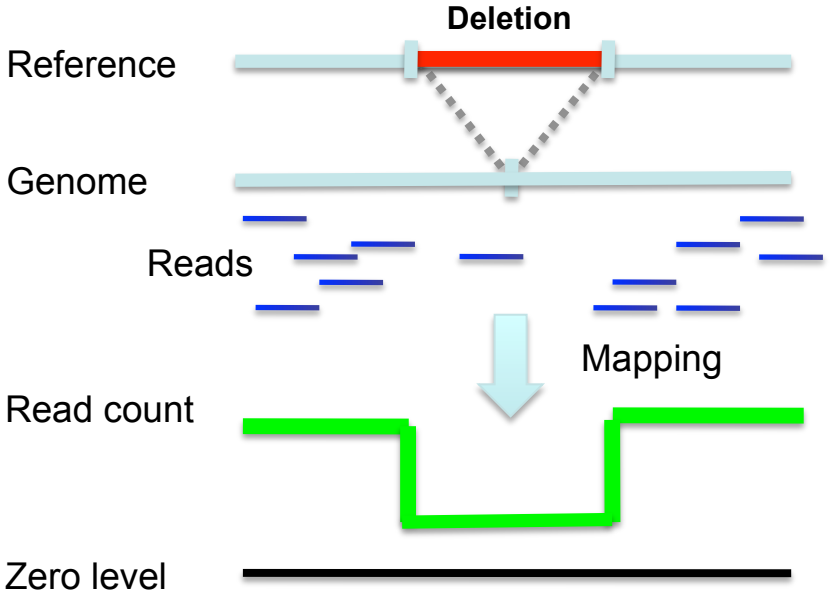
1. Paired ends



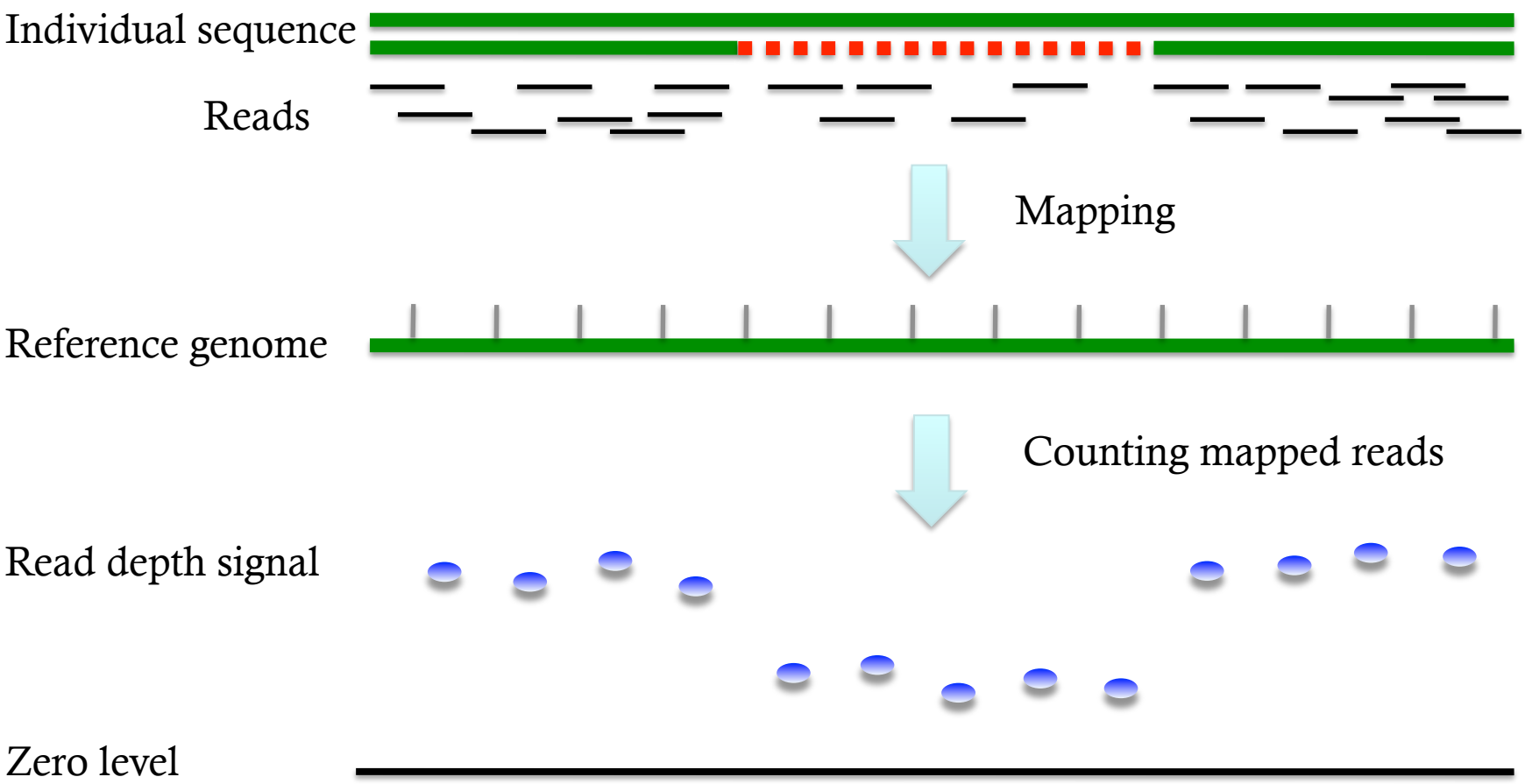
3. Split read



2. Read depth

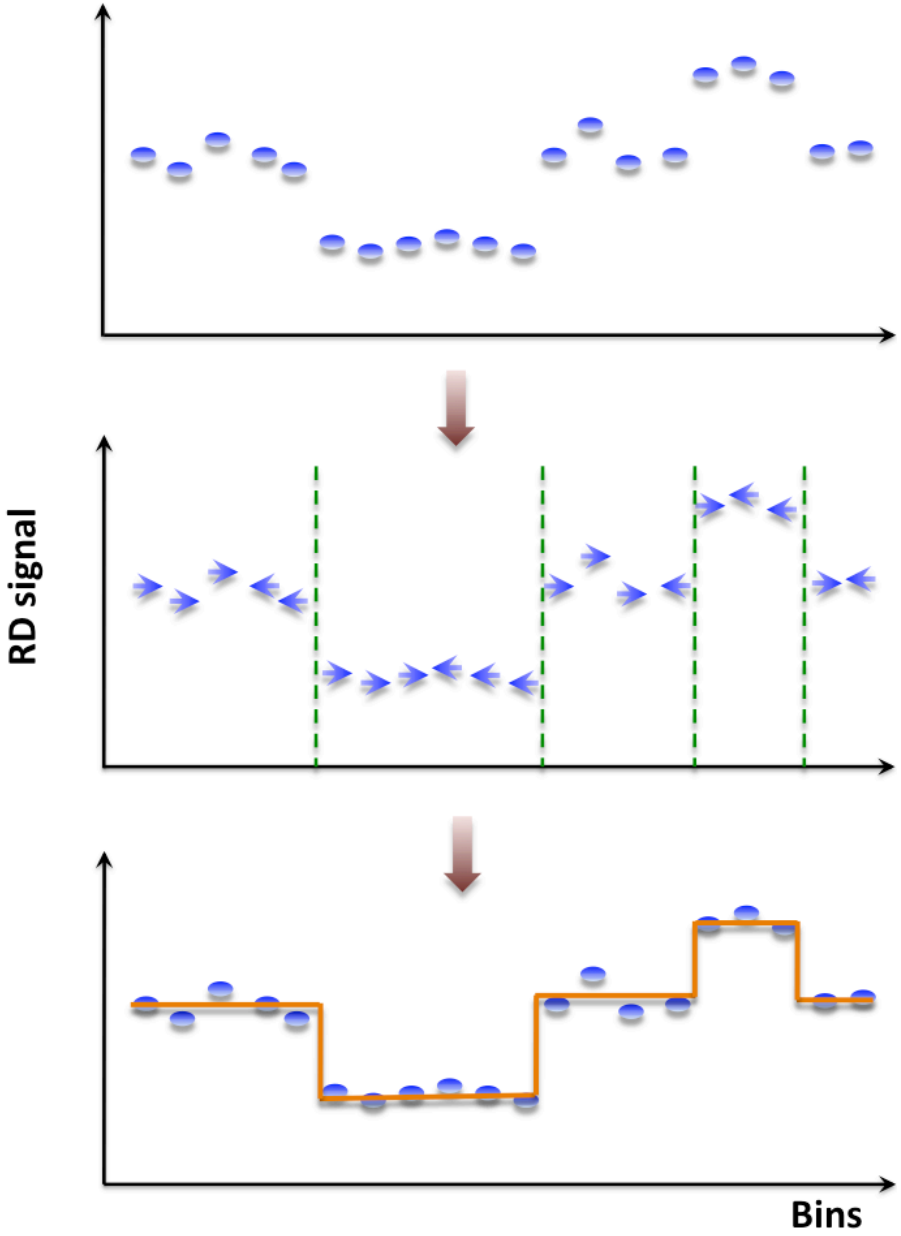


Sequence Read Depth Analysis



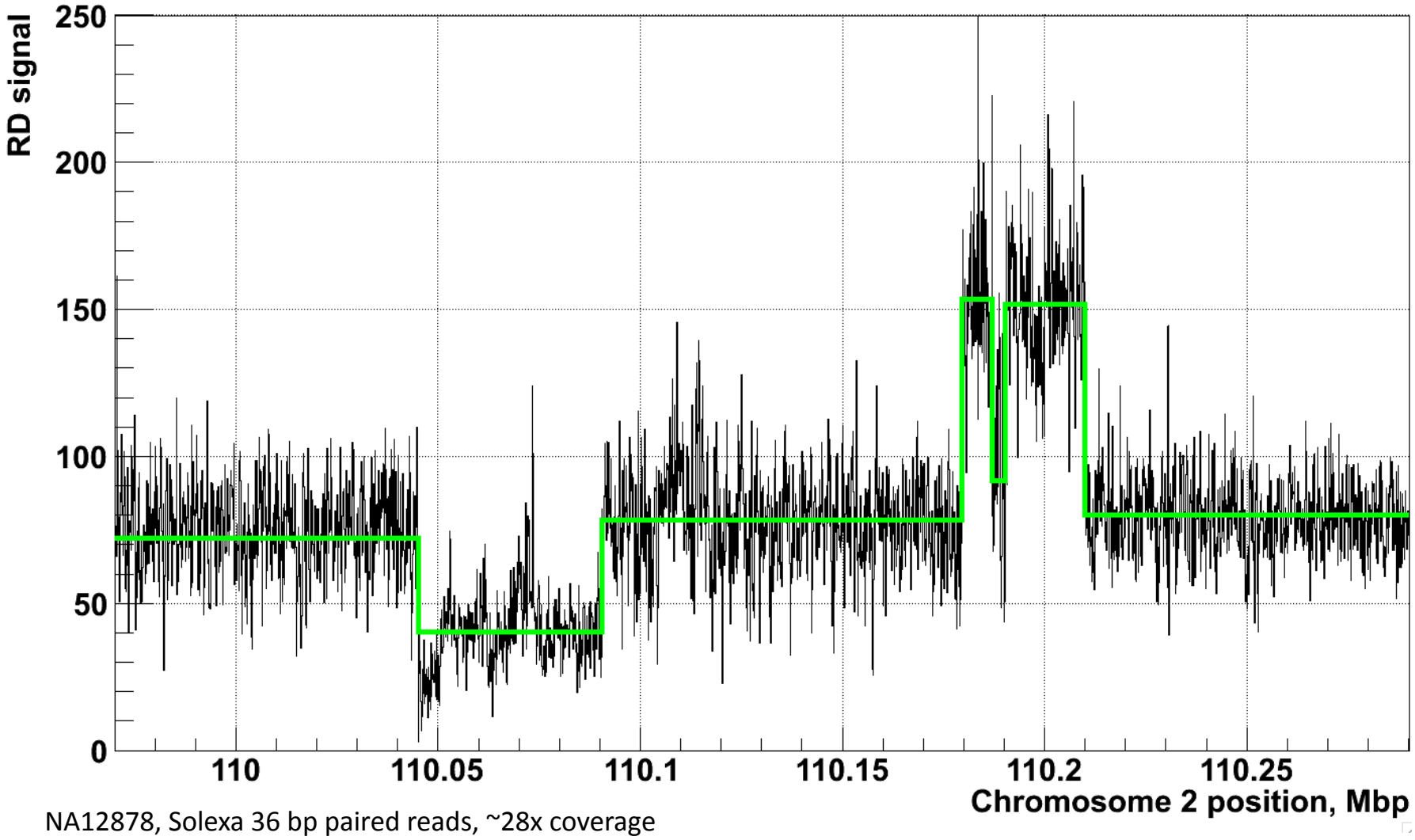
Novel method, CNVnator, mean-shift approach

- 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
- Achieves discontinuity-preserving smoothing
- Derived from image-processing applications



Alexej Abyzov

CNVnator on RD data



NA12878, Solexa 36 bp paired reads, ~28x coverage

Trio predictions

	CEPH trio			Yoruba trio		
	M	F	C	M	F	C
Coverage by mapped reads	~24X	~28X	~28X	~20X	~26X	~32X
Bin size	100	100	100	100	100	100
Power for CNV discovery	4.8	4.7	5.2	4.0	4.4	3.9
Power for CNV discovery (after GC correction)	5.4	5.3	5.8	4.6	5.0	4.9
All deletion calls	3678	3615	5656	3298	4988	2981
Deletion calls larger than 1 kb and excluding chromosomes X and Y	1420	1495	1784	1826	2195	1596
concordant with M	-	803	1008	-	912	878
concordant with F	803	-	1011	912	-	1046
concordant with C	1088	1011	-	878	1046	-
concordant with M or F	-	-	1316	-	-	1251
FDR from validation	19%	16%	19%	22%	26%	19%
FDR corrected for reference individual bias in CGH	5%	4%	10%	14%	16%	10%
Proportion of calls with incorrect boundaries	6%	5%	5% (6%)	6%	6%	5% (5%)
Estimated sensitivity	96% (84-93%)			87% (81-89%)		

RD vs paired-end

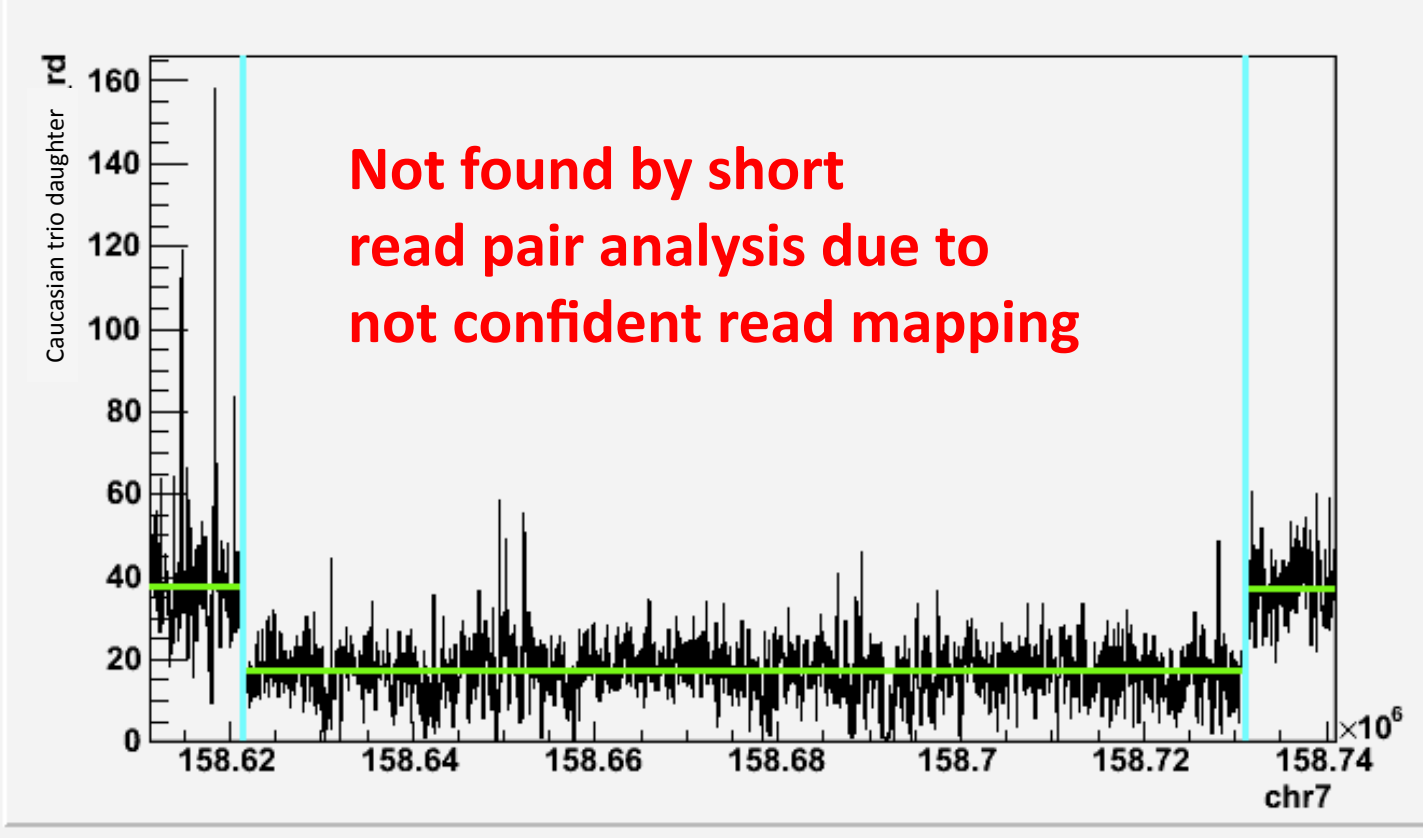
Read Depth

- Difficulty in finding highly repetitive CNVs (LINE, SINE, etc.)
- Uncertain in CNV location
- **Uses mutual information of both ends, better mapping and ascertainment in homologous region**
- **Ascertains complex regions**
- **Can find large insertions**
- Can be used with paired-end, single-end and mixed data

Paired-end

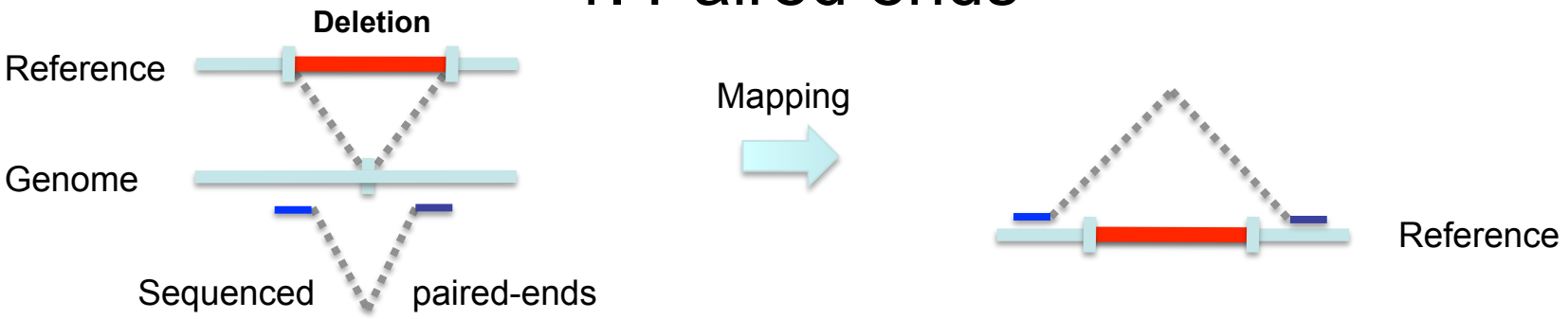
- Can detect highly repetitive CNVs (LINE, SINE, etc.)
- Defines precise location of CNV
- **Relies on confident independent mapping of each end, problems in regions flanked by repeats**
- Small span between ends limits resolution of complex regions
- Large span between ends limits resolution of break points

RD vs read pair (example)

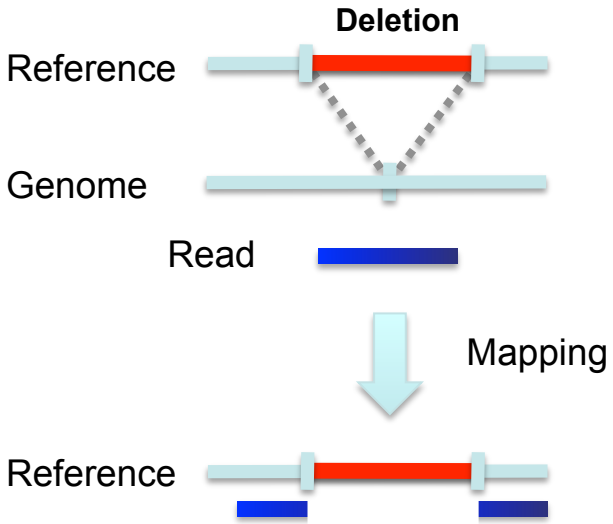


High Throughput DNA Sequencing based Methods to detect CNVs/SVs

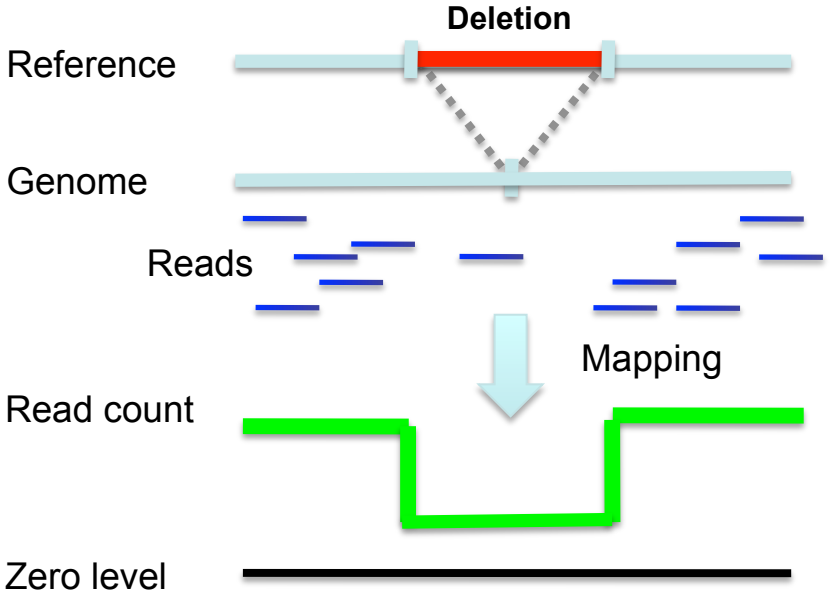
1. Paired ends



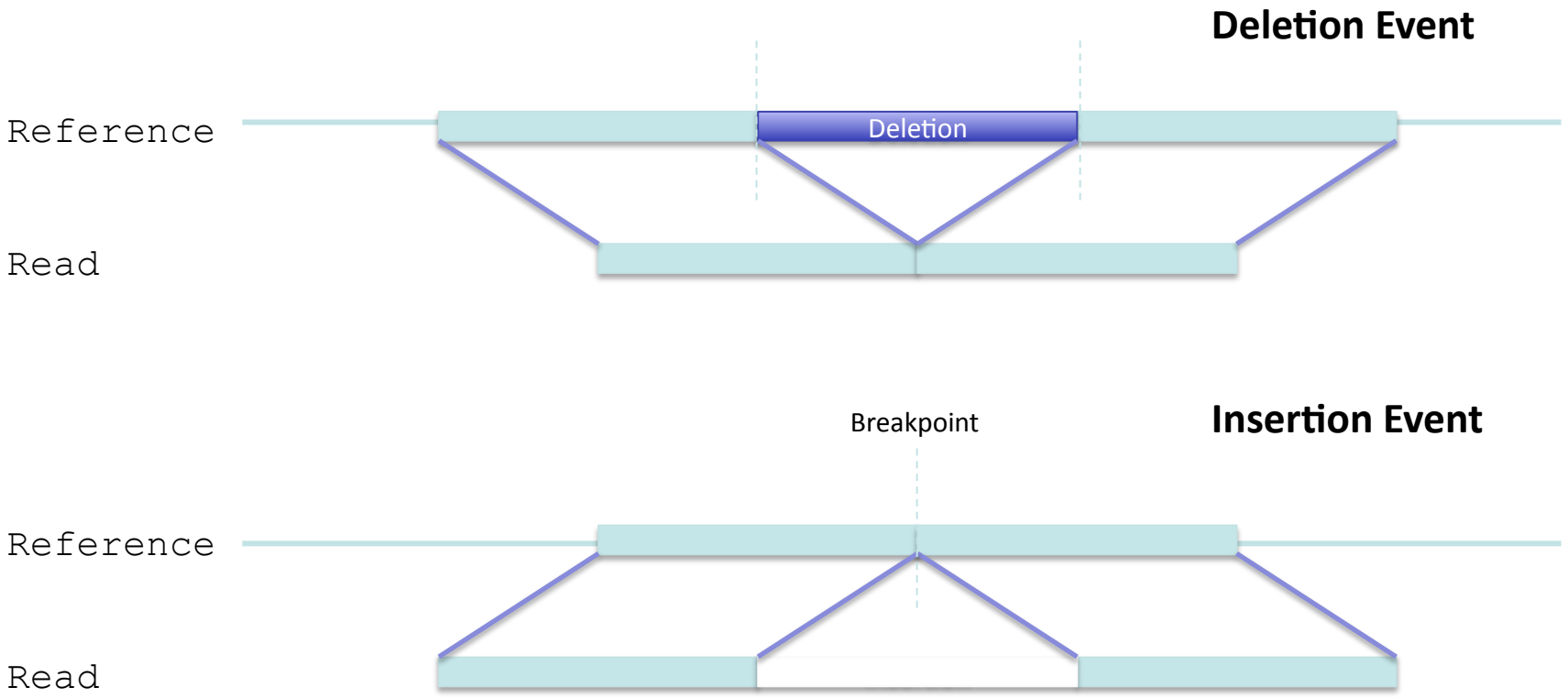
3. Split read



2. Read depth

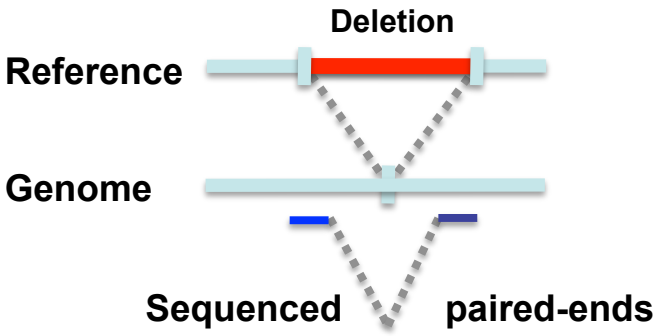


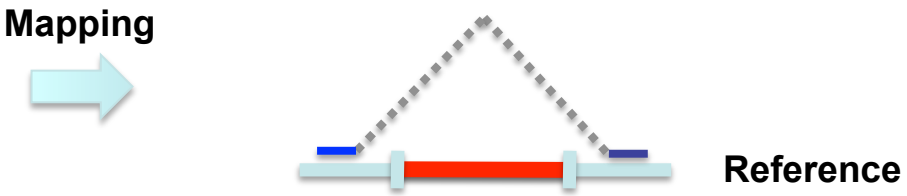
Split-read Analysis



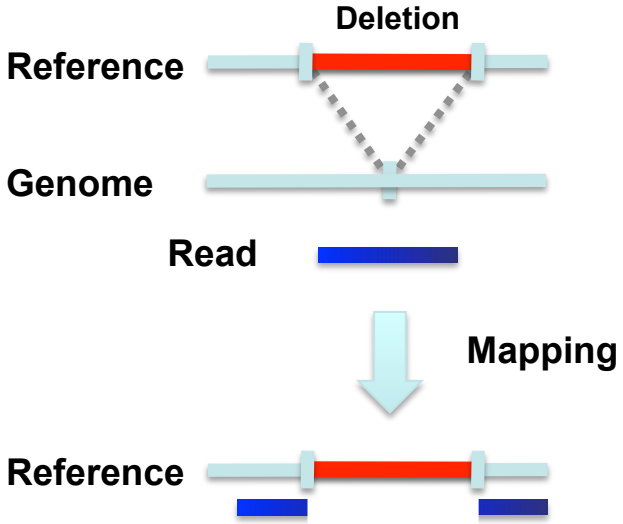
Methods to Find SVs

1. Paired ends

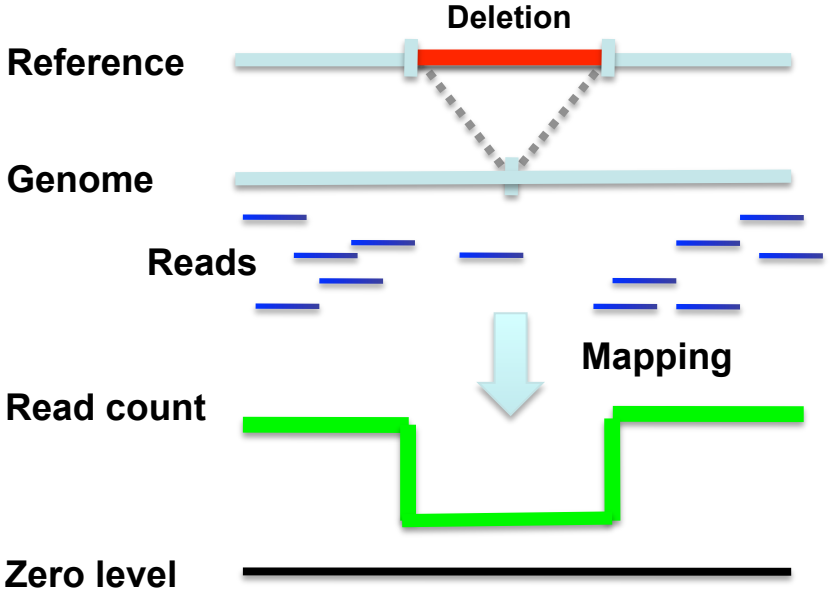




2. Split read



3. Read depth (or aCGH)

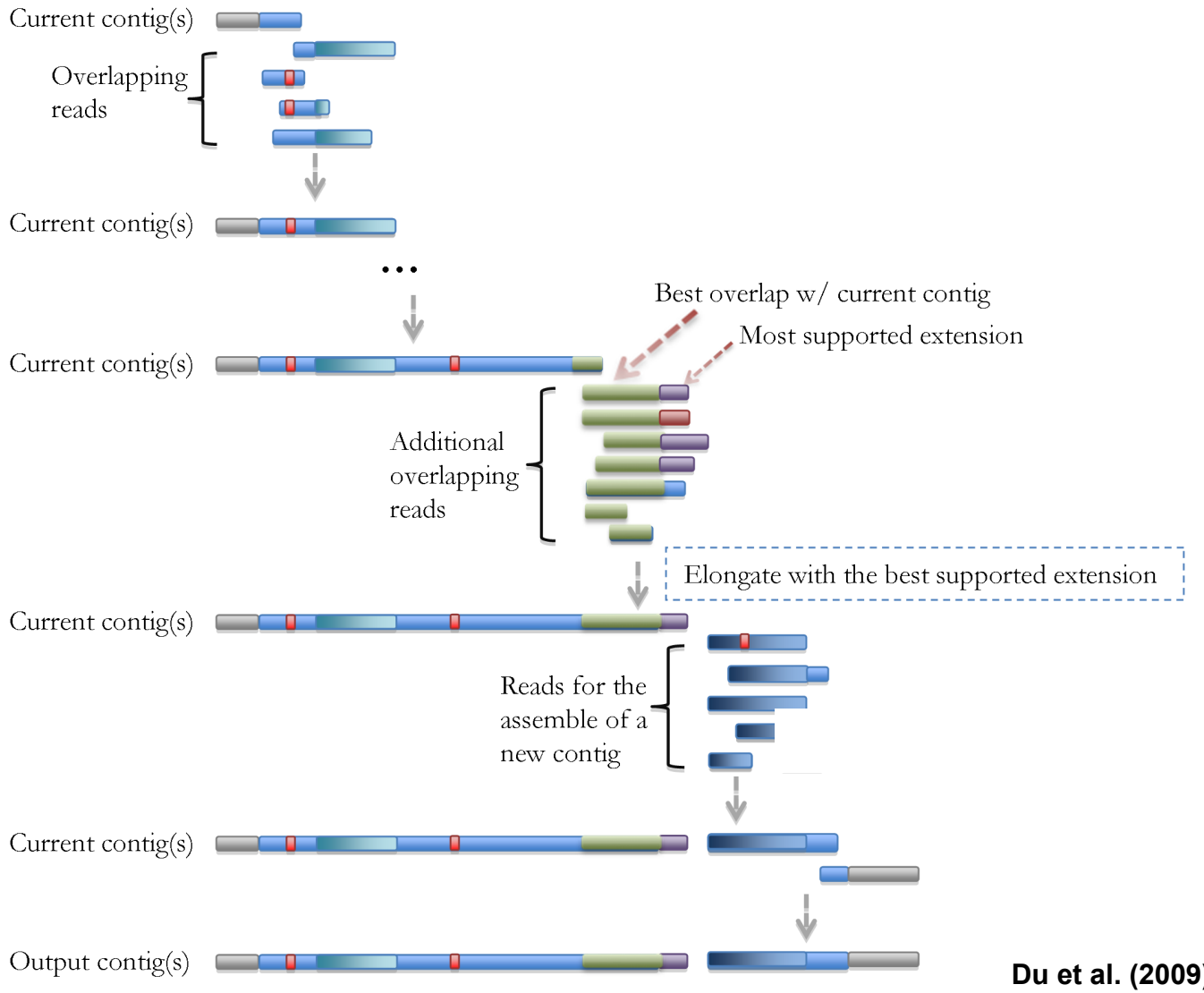


4. Local Reassembly

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

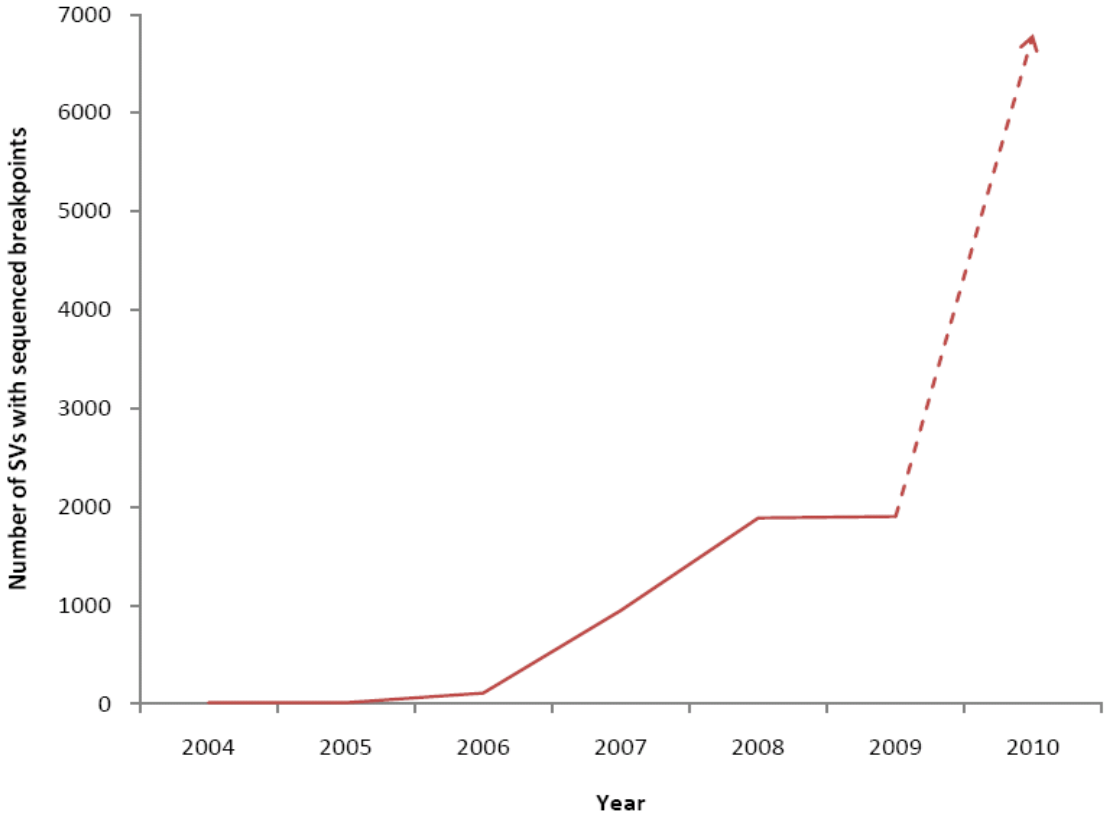
Simple Local Assembly: iterative contig extension

G Iterative contig elongation with the best supported extension -- a mostly greedy approach



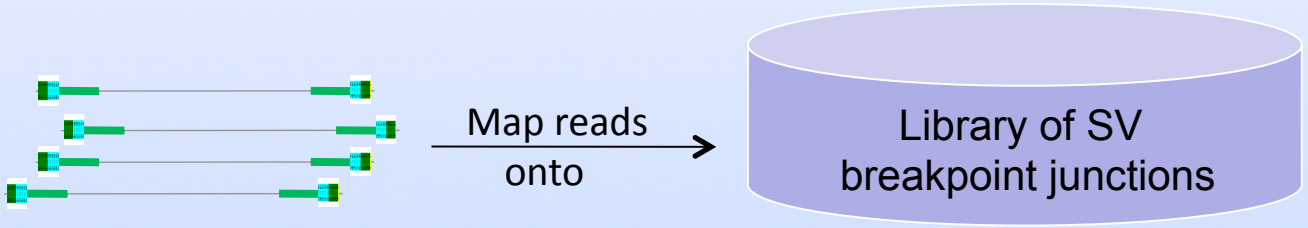
Du et al. (2009), PLoS Comp Biol.

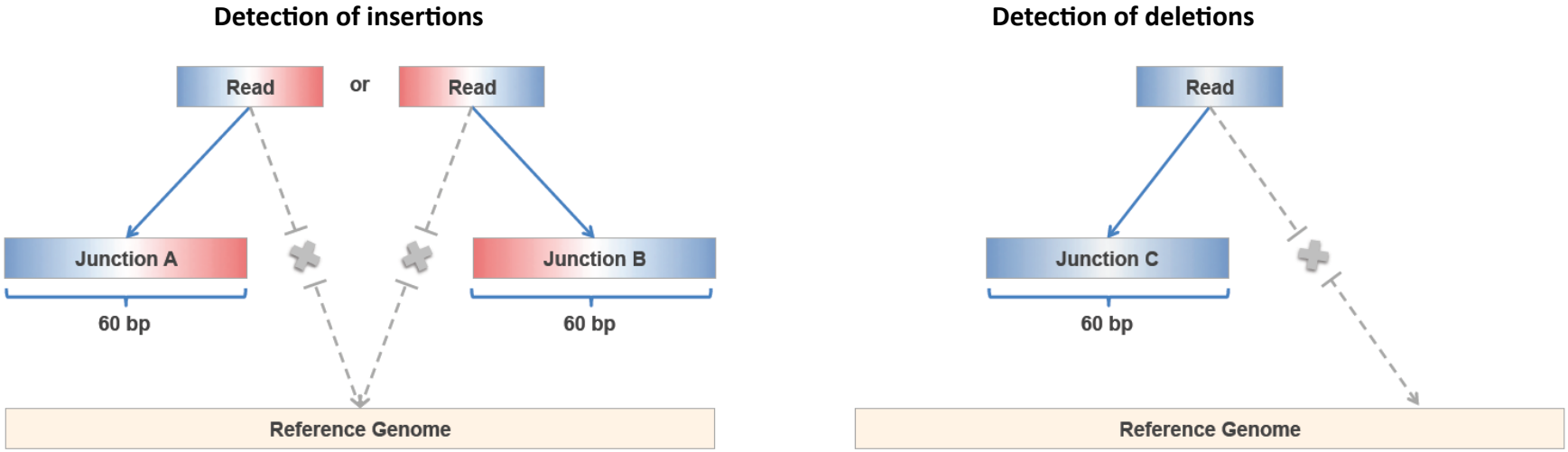
SVs with sequenced breakpoints



BreakSeq enables detecting SVs in Next-Gen Sequencing data based on breakpoint junctions

Leveraging read data to identify previously known SVs ("Break-Seq")





* Read overlaps <10 bp to one side of the breakpoint is discarded and read matches also to the reference genome is classified as non-unique match

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

Applying BreakSeq to short-read based personal genomes

Personal genome (ID)	Ancestry	High support hits (>4 supporting hits)	Total hits (incl. low support)
NA18507*	Yoruba	105	179
YH*	East Asian	81	158
NA12891 [1000 Genomes Project, CEU trio]	European	113	219

***According to the operational definition we used in our analysis (>1kb events) less than 5 SVs were previously reported in these genomes …**

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

Conclusions

1) **SVs are abundant in the human genome**

2) **Different methods are used to detect them: Read pairs, Read Depth, Split reads, New assembly**

3) **Many SV breakpoints are being sequenced; nonhomologous end joining is common. The breakppoint library can be used to identify SVs.**

Acknowledgments

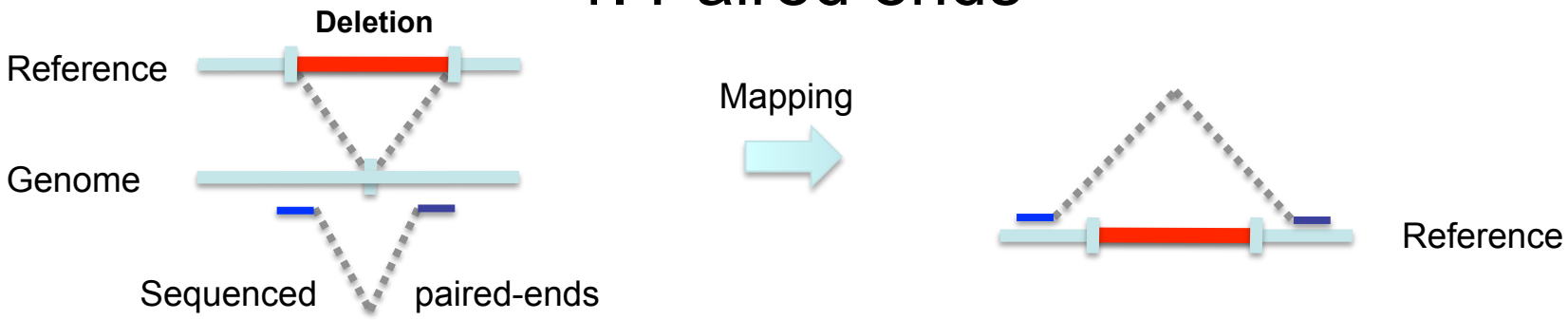
- **Jan Korbel**
- **Alexej Abyzov**
- **Alex Urban**
- **Zhengdong Zhang**
- **Hugo Lam**
- **Mark Gerstein**

454 for Paired End

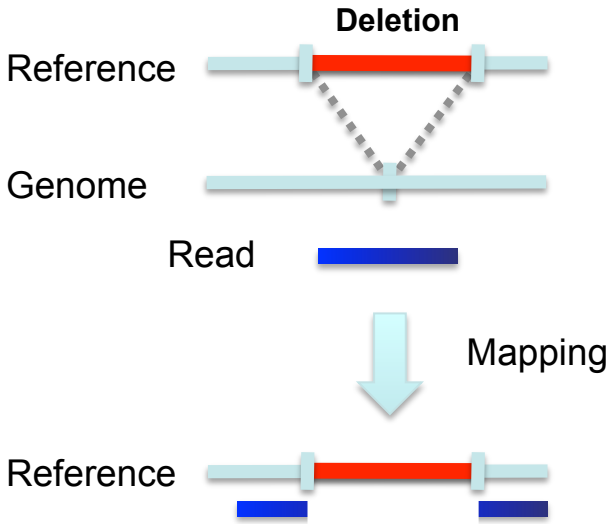
Tim Harkins, Michael Egholm

2nd-Gen Sequencing based Methods to detect CNVs/SVs

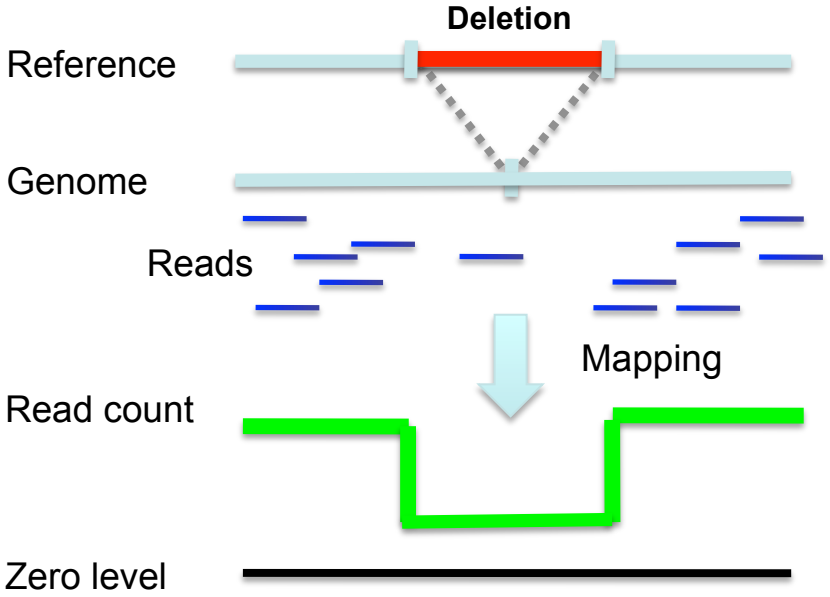
1. Paired ends



2. Split read



3. Read depth



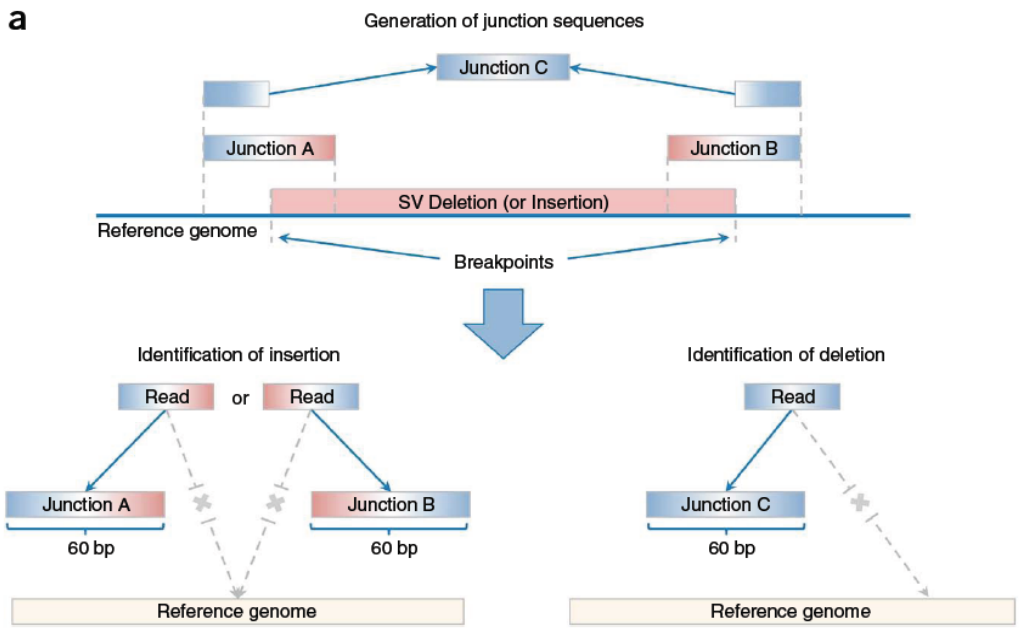
SV-CapSeq v1.0 results for deletions

Data set	Total SVs	Confirmed	Confirmation rate	Confirmation rate (coverage corrected)*
1KG selected events	1839	307	17%	20%
Pre-confirmed	184	134	73%	88%
PCR confirmed	294	101	34%	41%
Pre- & PCR confirmed	56	41	73%	88%
PCR non-validated	940	105	11%	13%
454 PEMer deletions	575	283	49%	59%

Combining 3 captures/elutions (1 per member of CEU trio) and 1+(2x0.5) 454 Titanium runs

*For 2x allelic coverage and breakpoints at least 20 bp away from read ends

SV Junction and Identification



Read overlaps <10 bp to one side of the breakpoint is discarded and read matches also to the reference genome is classified as non-unique match

Figure 2 Mapping breakpoints using the library. (**a**) Overview of the BreakSeq approach. Breakpoints are used to generate junction sequences spanning breakpoints (upper)—the 30 bp of sequence flanking each side of the breakpoint (60 bp total). Then, DNA reads are aligned to the junction sequences (lower). Alignment results are interpreted as follows. In the case of insertions relative to the reference genome (left), sequences A and B represent the left and right breakpoint junction sequences of the nonreference SV allele, respectively. In the case of deletions (right), sequence C represents the junction sequence of the nonreference SV allele. Solid lines with arrows, successful alignments. Dashed lines with crosses, no proper alignment.

For the HCH, CEPH (NA12891) and YRI (NA18507) genomes, we identified 158, 219 and 179 SVs, respectively. 57 SVs were shared between the YRI and HCH genomes, 62 between the YRI and NA12891 genomes, 52 between the HCH and NA12891 genomes, and 42 were common to all three genomes.

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

Contents of the SV-CapSeq array v1.0

2.1 million oligomers tiling the target regions of the genome:

1839 deletion CNVs from (mostly) short read Solexa data (1000 Genome Project)

From long read 454 paired-end data:

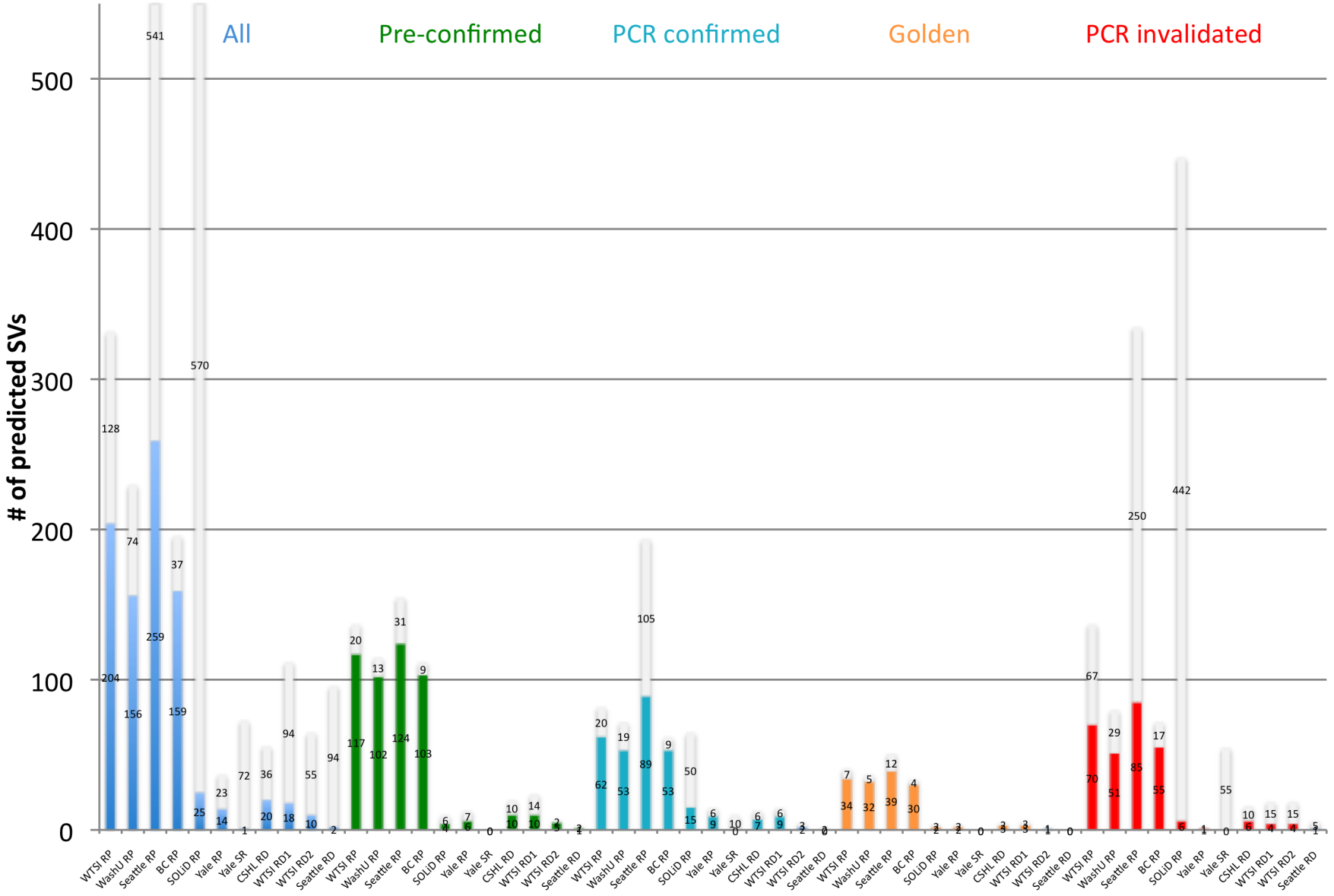
575 deletion CNVs

296 insertions CNVs

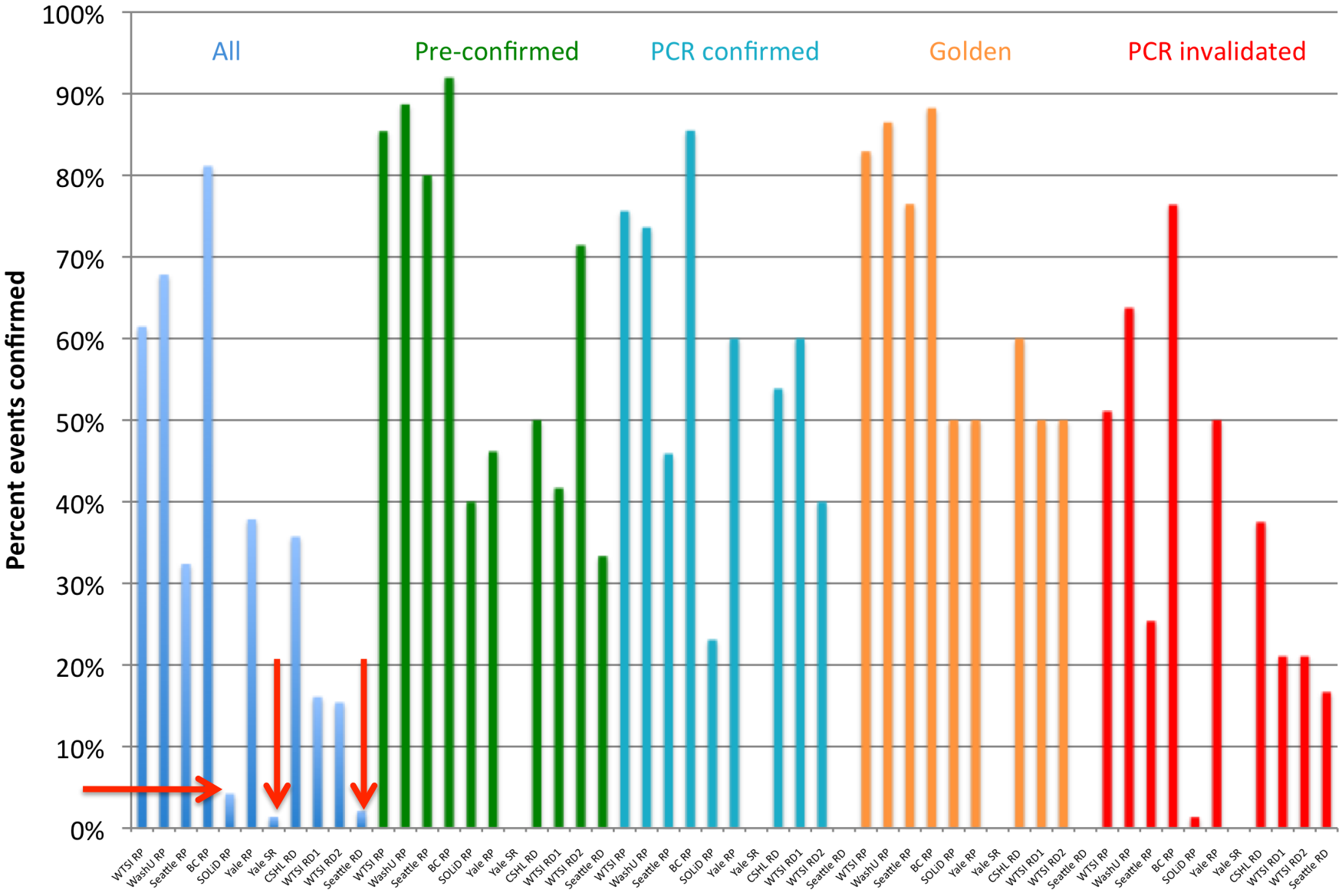
191 inversions SVs

(plus Split-Read indel predictions, Zhengdong Zhang)

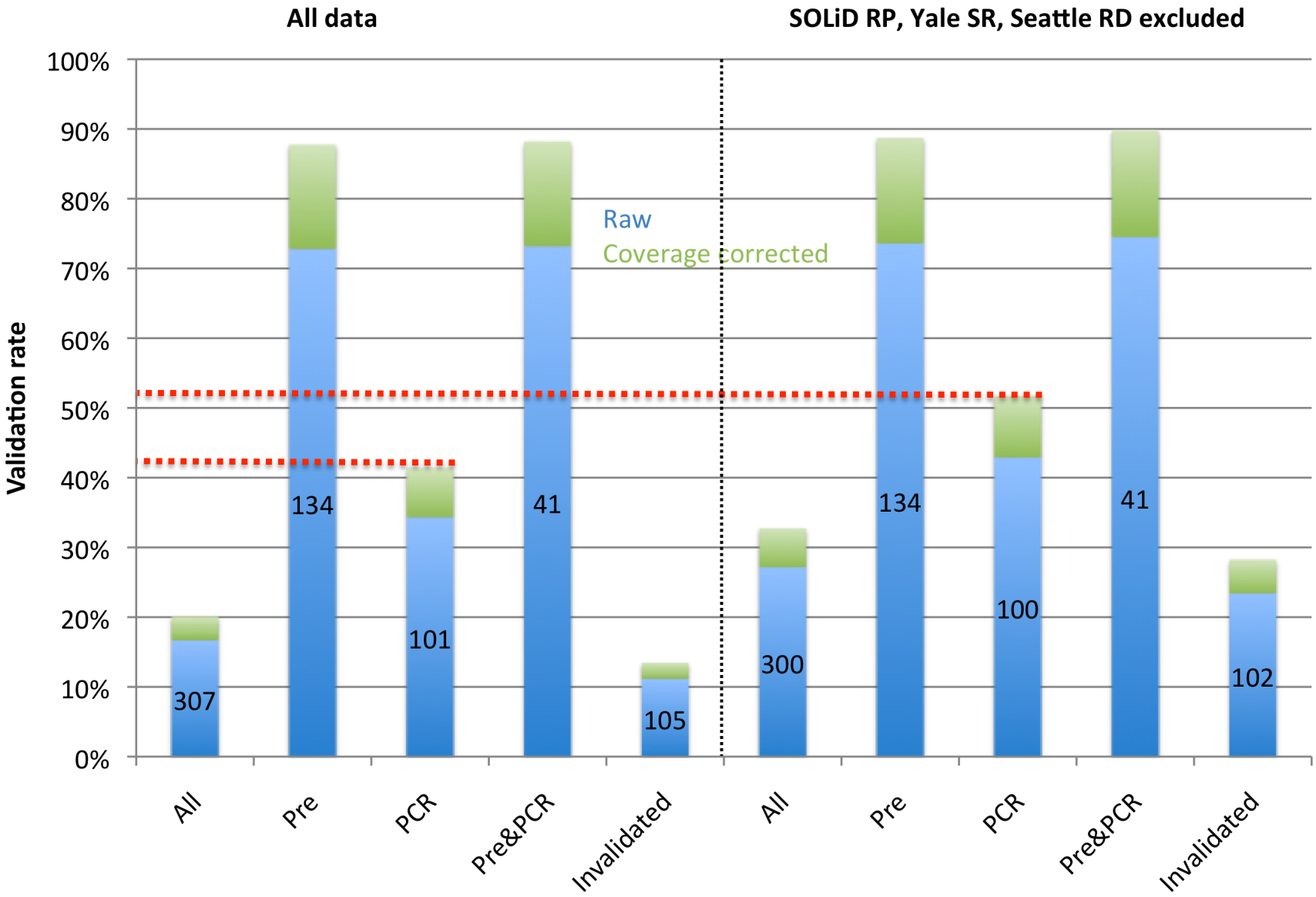
Validations by prediction set

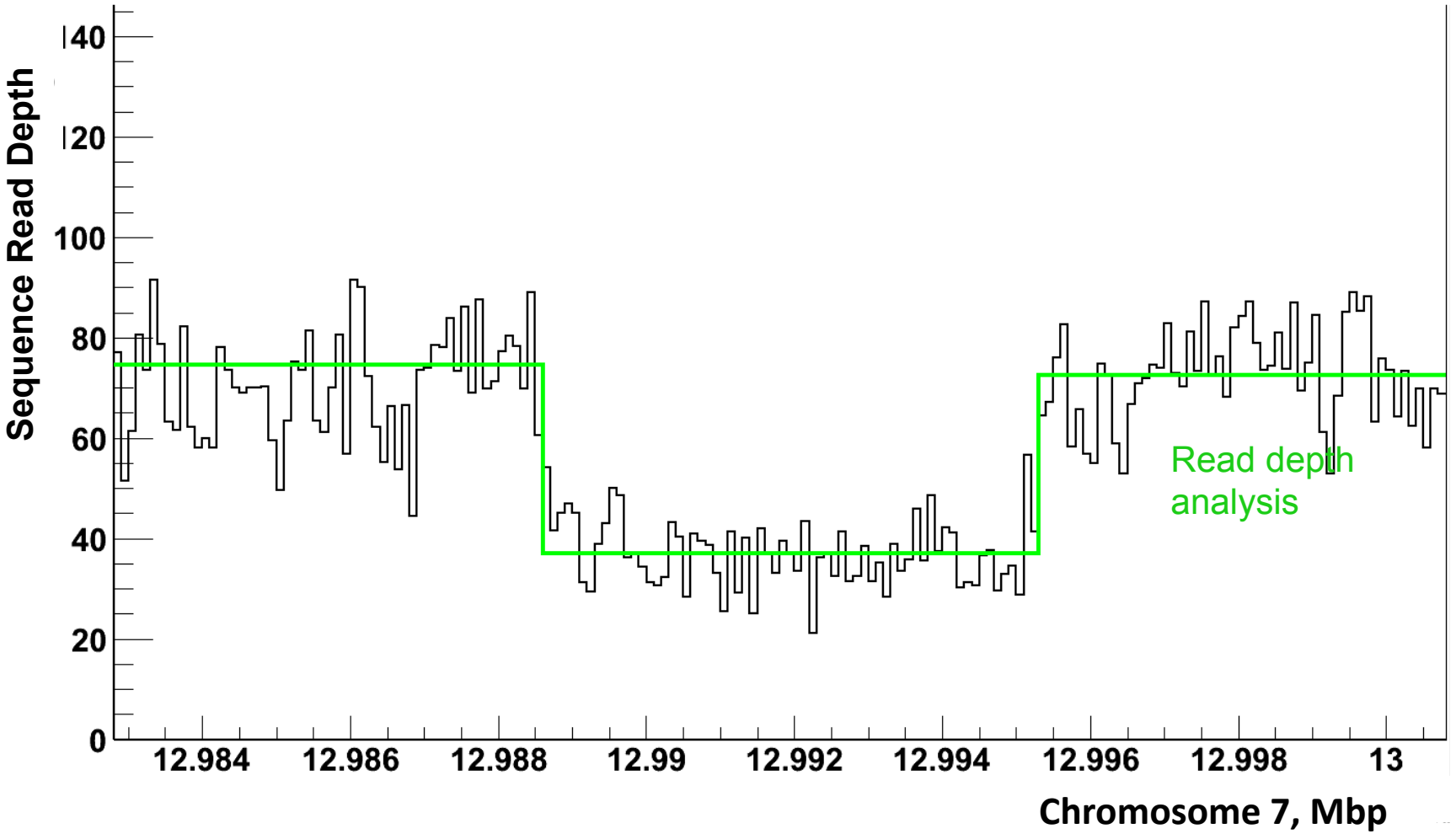


Validation rate by prediction set

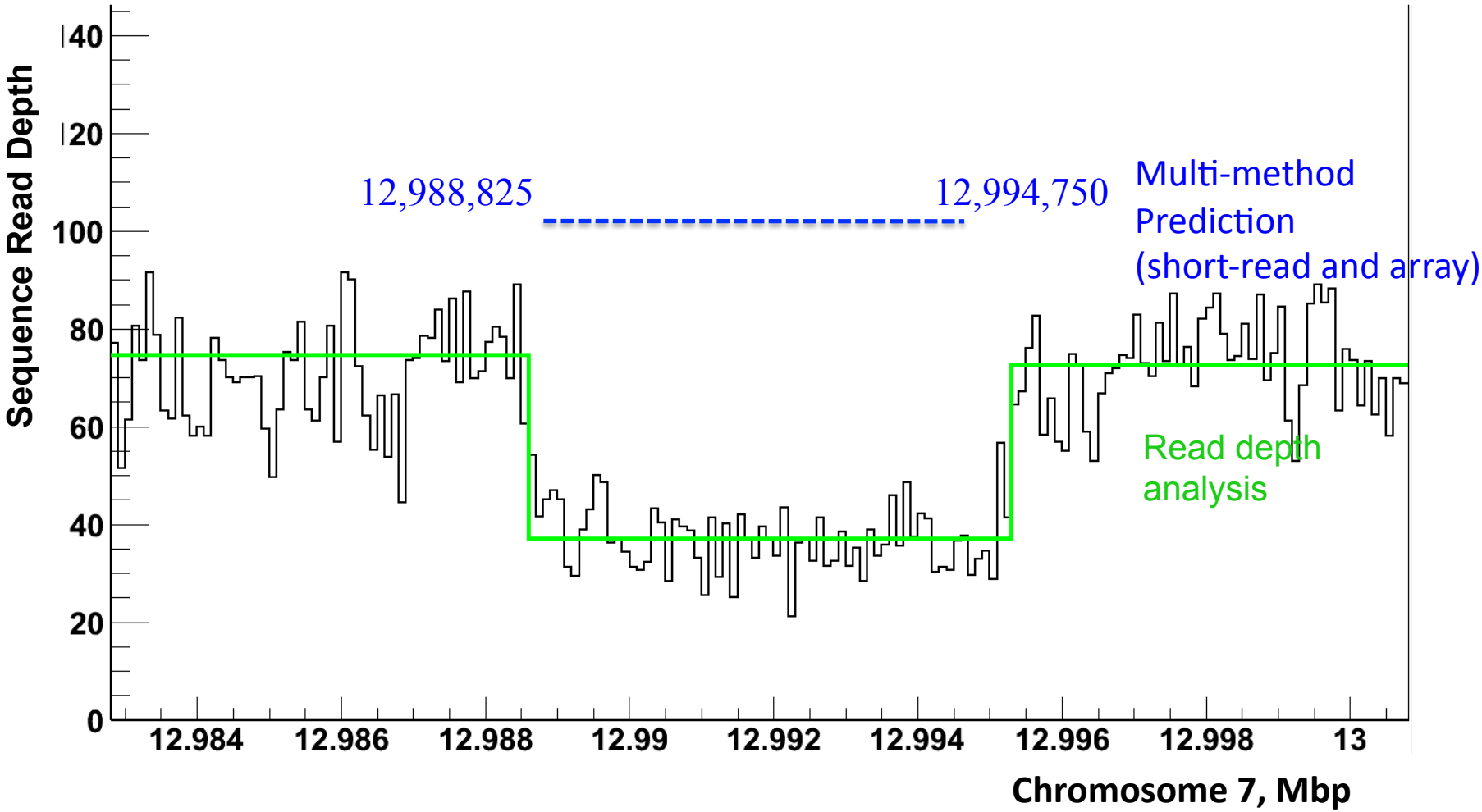


Confirmation rate

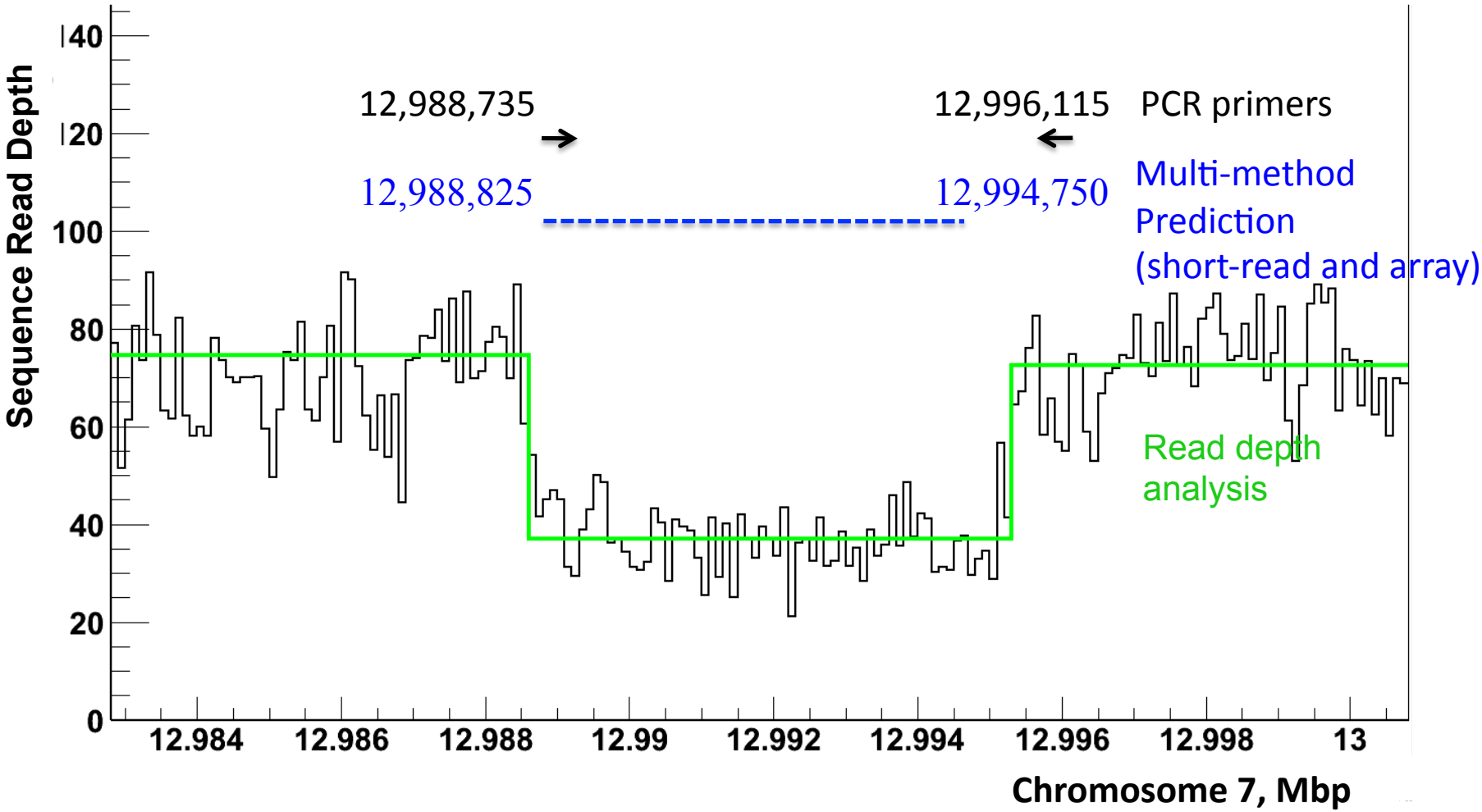




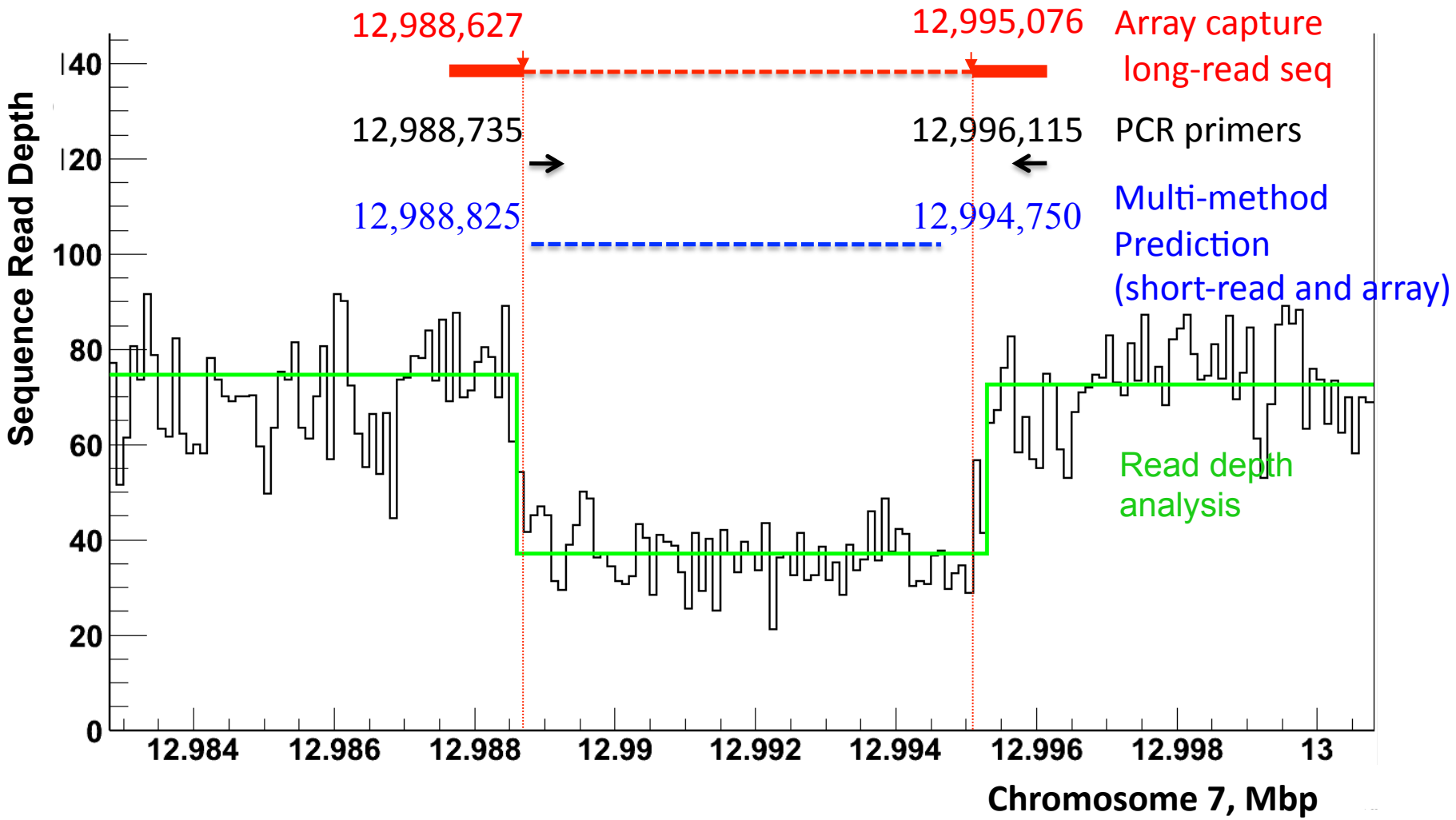
~6500 bp deletion CNV



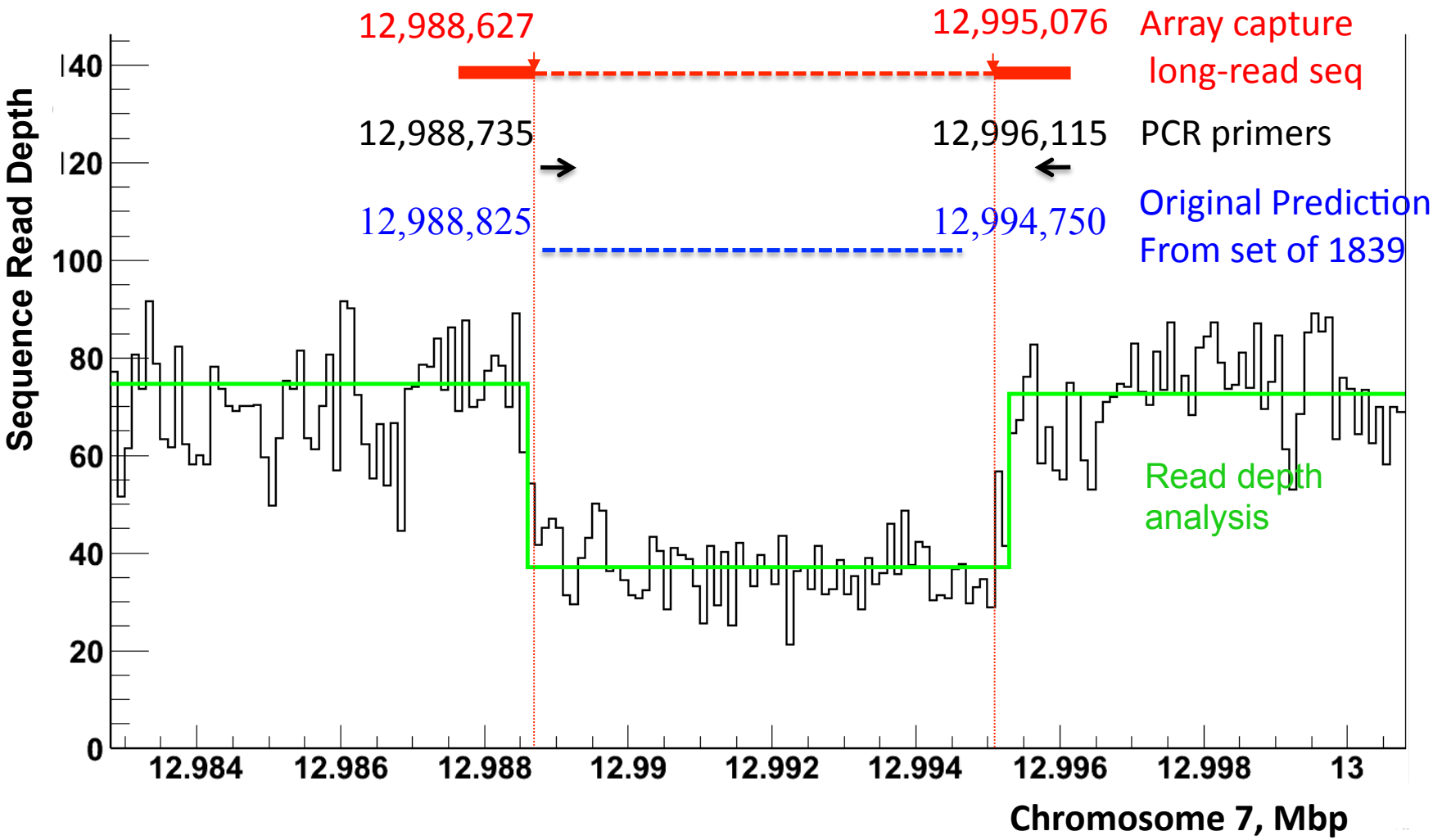
~6500 bp deletion CNV



~6500 bp deletion CNV



~6500 bp deletion CNV



~6500 bp deletion CNV

SV-CapSeq v1.0 results for deletions

Data set	Total SVs	Confirmed	Confirmation rate	Confirmation rate (coverage corrected)*
1KG selected events	1839	307	17%	20%
Pre-confirmed	184	134	73%	88%
PCR confirmed	294	101	34%	41%
Pre- & PCR confirmed	56	41	73%	88%
PCR non-validated	940	105	11%	13%
454 PEMer deletions	575	283	49%	59%

Combining 3 captures/elutions (1 per member of CEU trio) and 1+(2x0.5) 454 Titanium runs

*For 2x allelic coverage and breakpoints at least 20 bp away from read ends

SV-CapSeq Analysis of Structural Variation in the human genome

Ongoing work:

-Develop analysis pipelines for *insertion* and *inversion* SV-CapSeq data

-Analyze nature of off-target CapSeq reads: cross-hybridization and cross-mapping

-Design improved SV-CapSeq array

Goal

Sequence across n x 10,000 SV breakpoints with a single capture and less than one 454 run or ideally using Solexa-Illumina

Important for precision CNV/SV screens and high-quality human genome sequencing

Analysis of Genomic Structural Variation

-exact sizes and breakpoint sequences of CNV/SV are difficult to define but important for functional understanding

-in the absence of long-read deep whole-genome sequencing combining arrays and sequencing allows high-throughput validation and breakpoint analysis

SV-CapSeq Design v2.0:

For Pilot2/DeepCov:

Total SVs -- 3946 (set of CNV used by Jan Korbel for PCR primer design/round 2; only CEU trio)

Deletions -- 2550

Insertions -- 1396 (includes mobile elements)

Total bases to be covered -- 4,784,597

Expected coverage -- 7x (for diploid genome with 500,000 of 400 bp reads by 454)

SV-CapSeq Design v2.0:

For Pilot1/LowCov

NA12003 -- CEPH male

NA18870 -- Yoruba female

NA18953 -- Japanese male

SV selection:

1) All events selected by Jan for PCR validation

2) 250 RD calls from each of the following groups: Yale, CSHL, Einstein

Tiling strategy:

200 bp into outer direction for insertion break point(s)

500 bp into both directions from deletion break points

Total SVs -- 1546

Deletions -- 1438

Mobile elements -- 108

No other insertions

Total bases to be covered -- 2,501,719

Expected coverage -- 8.8x (for diploid genome with 1,000,000 of 400 bp reads by 454)

Computations

- Megablast mapping
  - Mismatch score = -1
  - Hits with > 90% identity
  - At least 40 matching bases
- Best hit placement
  - At least one hit has score > 150
  - No overlapping hits with score difference < 10
- Selecting candidate reads by intersecting placements with predicted regions extended by 1kb
- Needleman-Wunsch alignment of candidate reads with predicted regions (0 gap extend penalty)

Criteria for validation

- Can find two good alignment blocks (see next slide)
- 50% mutual overlap between predicted region and gap between the blocks
- Sum of break-point uncertainty < 5 kb

Acknowledgements

Yale University

Alexej Abyzov

Jan O Korbel

Fabian Grubert

Dejan Palejev

Maya Kasowski

Chandra Erdman

Philip Kim

Nicholas Carriero

Eugenia Saunders

Andrea Tanzer

Mark Gerstein

Sherman Weissman

Michael Snyder
(now Stanford University)

454 / Roche

Jason Affourtit

Brian Godwin

Jan Simons

Lei Du

Bruce Taillon

Zhoutao Chen

Tim Harkins

Michael Egholm

Sanger Centre

Jianxiang Chi

Fengtang Yang

Yujun Zhang

Matthew Hurles

Nigel Carter

UCLA/Cedars-Sinai

Tal Tirosh-Wagner

Julie Korenberg
(now University of Utah)

UPenn

April Hacker

Beverly Emanuel

Cornell

Francesca Demichelis

Sunita Setlur

Mark Rubin

NimbleGen-Roche

Rebecca Selzer

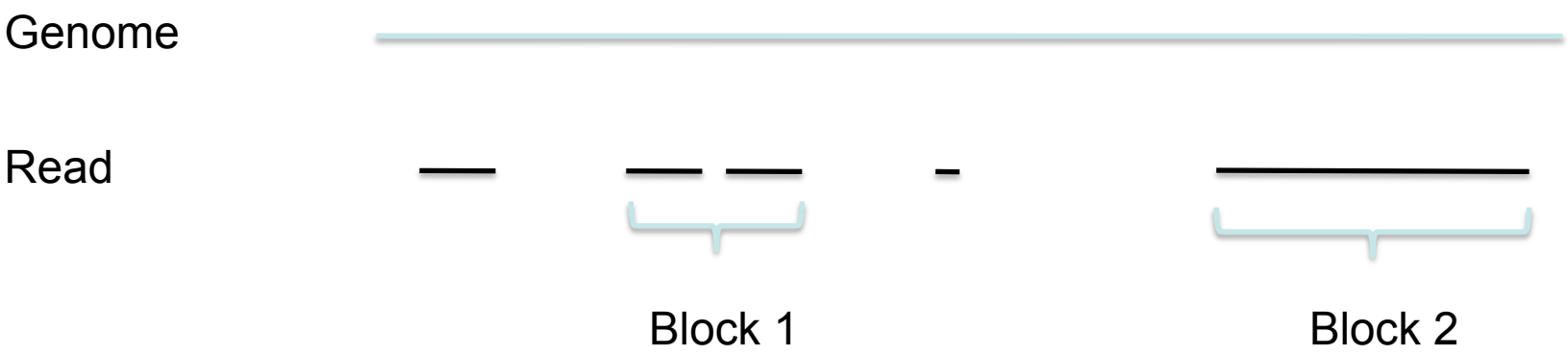
Todd Richmond

Matthew Rodesch

Roland Green

Thomas Alberts

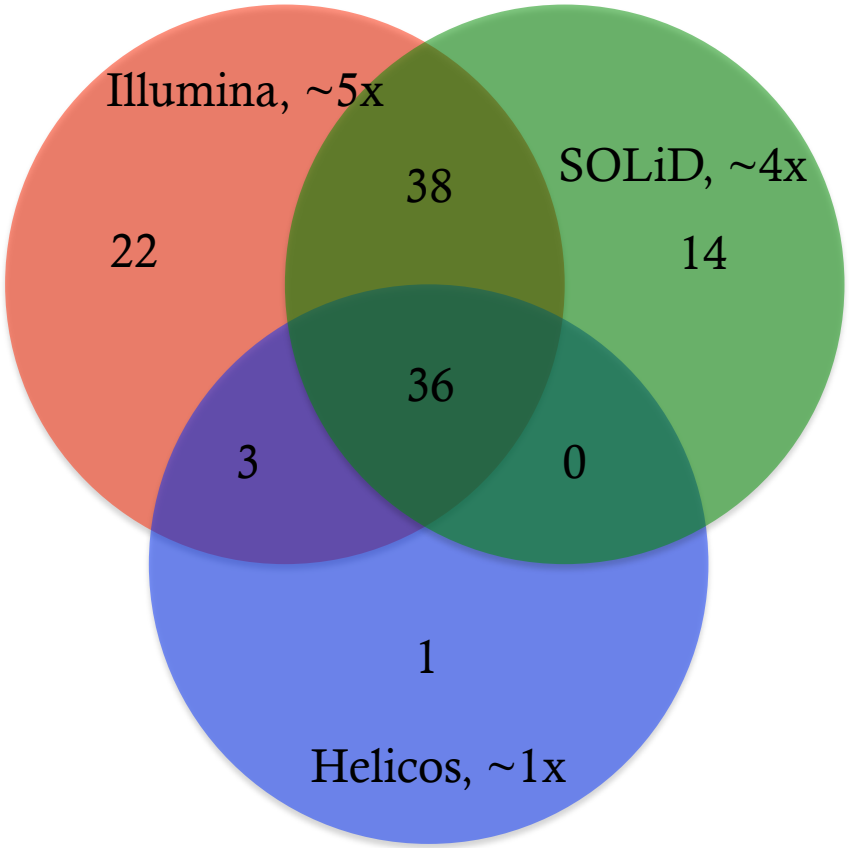
Alignment blocks



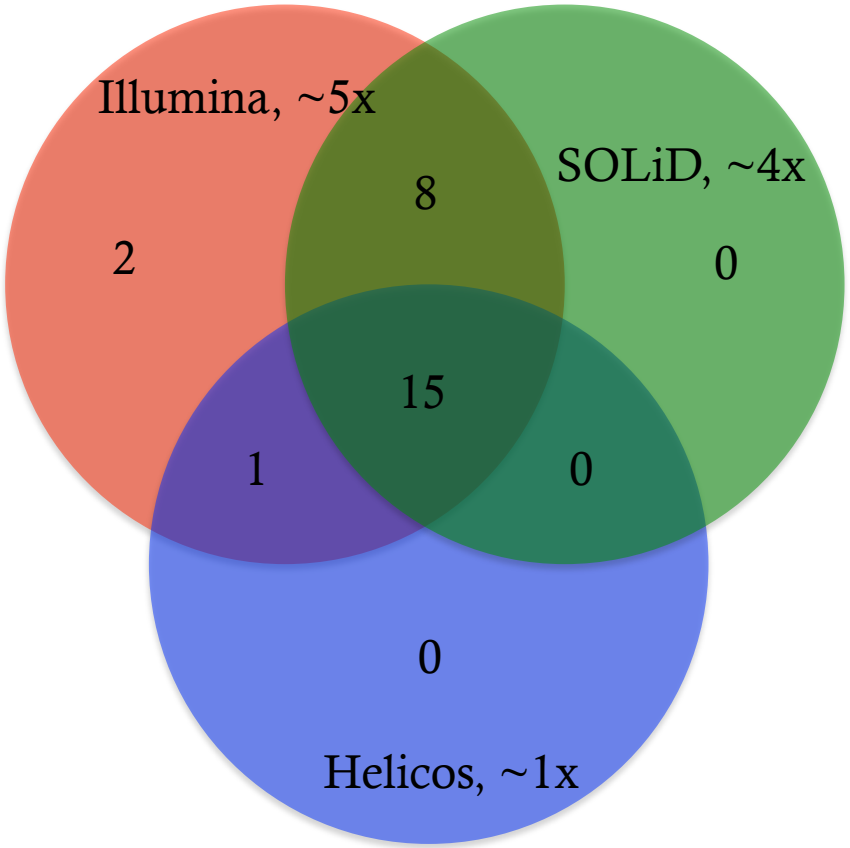
Criteria: gaps < 5 bp, number of aligned nucs > 10

Read-Depth Analysis: Platform comparison (on aCGH calls)

Deletions

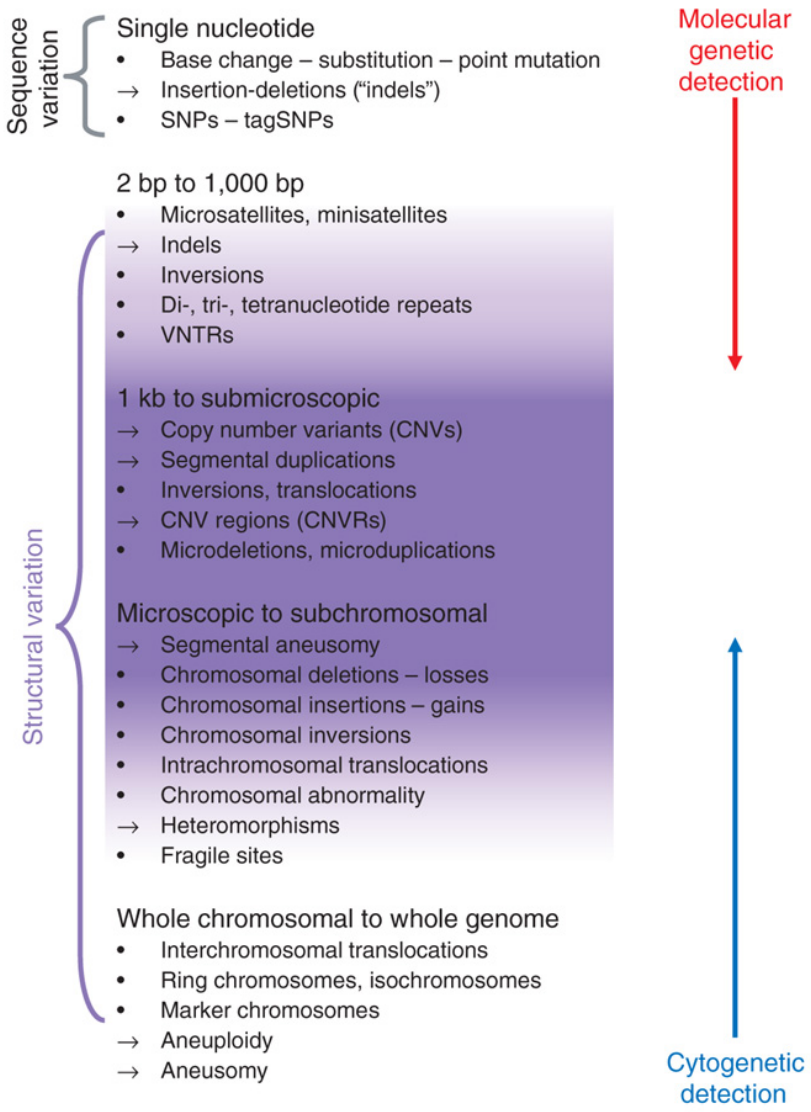


Duplications



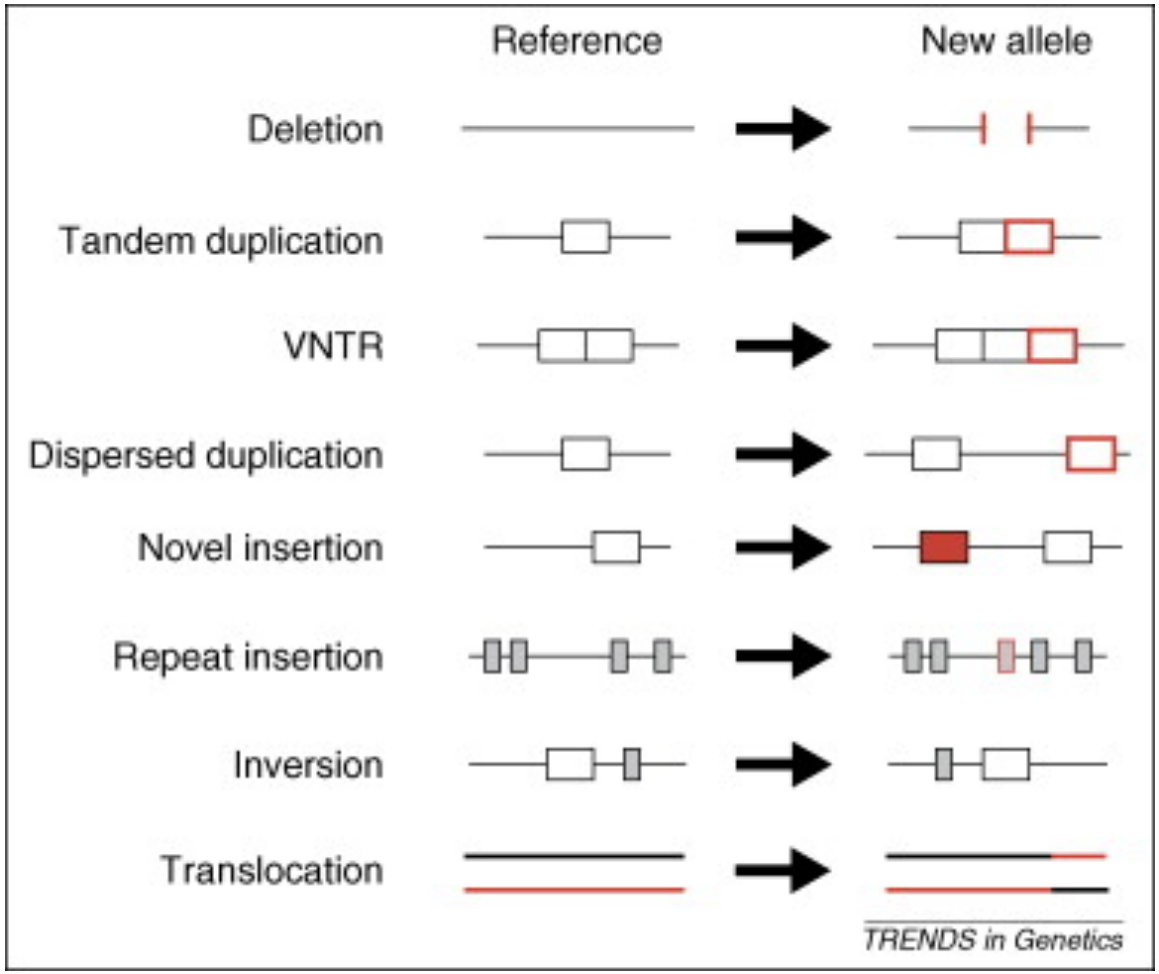
by >50% of reciprocal overlap

Size Spectrum of Human Genomic Variation

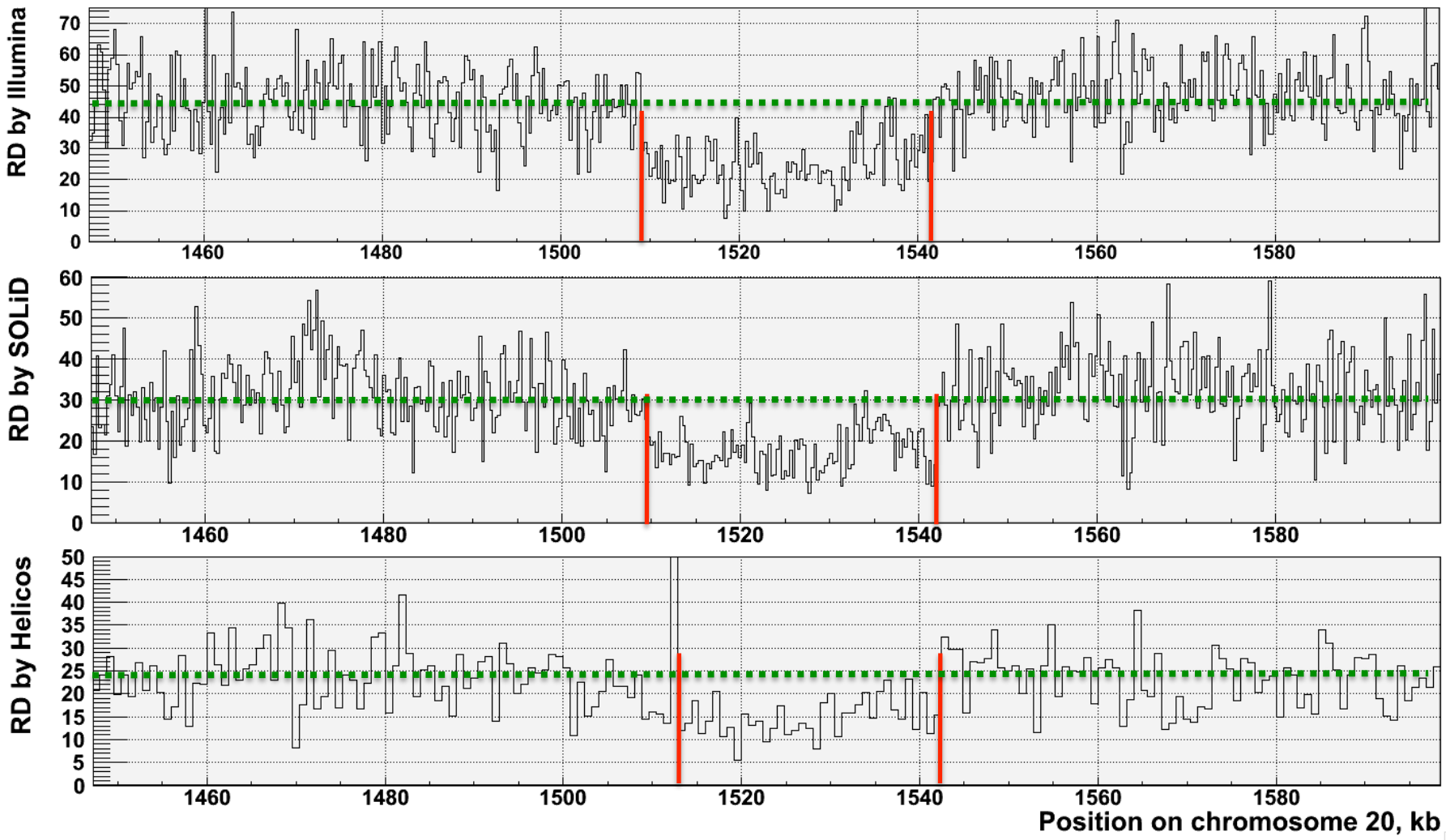


Scherer *et al.* 2007

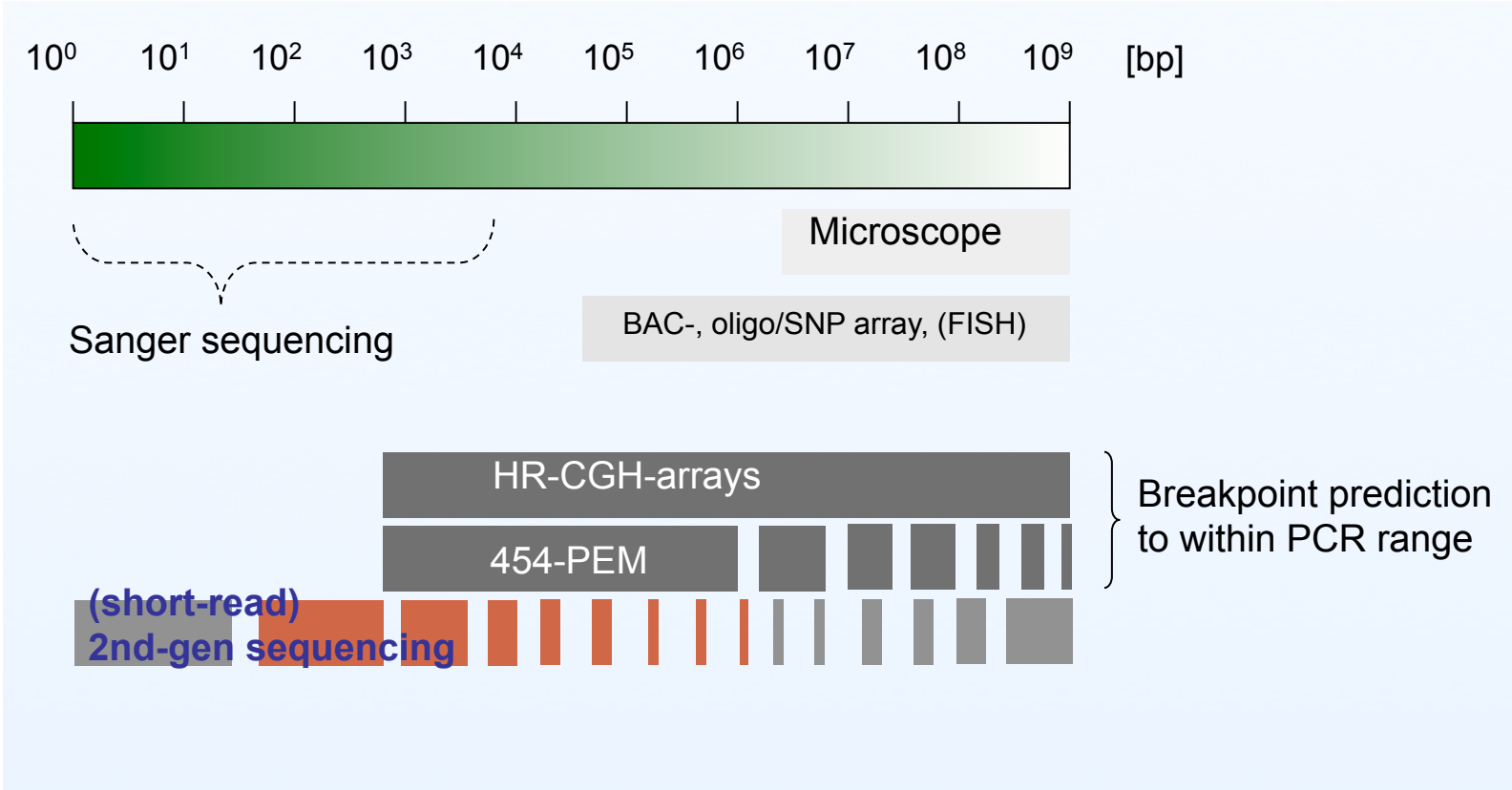
Types of Structural Variation



Hurles *et al.* 2008

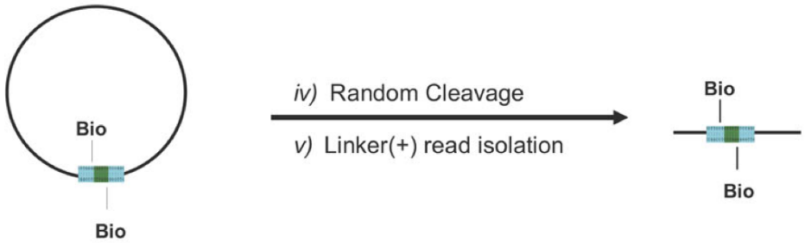

RD by Illumina
RD by SOLiD
RD by Helicos
1460
1480
1500
1520
1540
1560
1580
Position on chromosome 20, kb

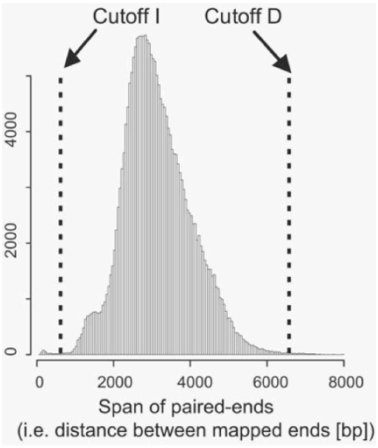
The resolution gap in SV analysis

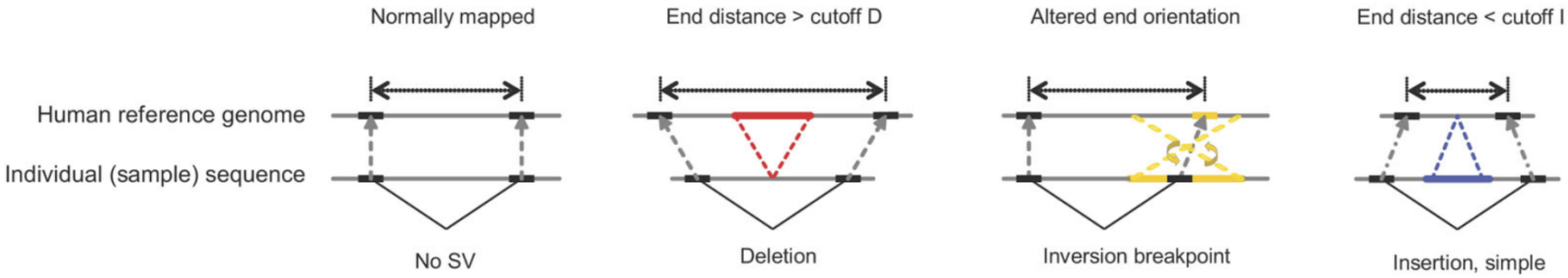


[adapted from Lupski *et al. Nat Genet 2007*]

Paired End Mapping



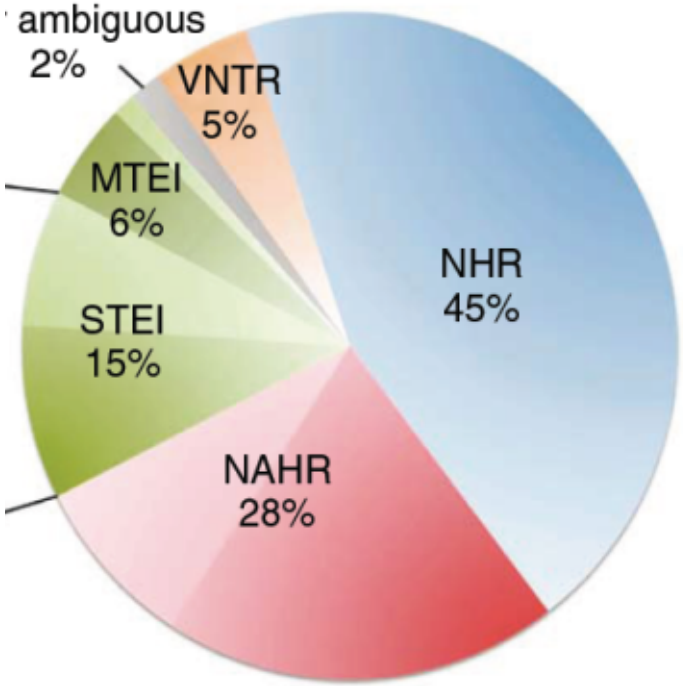




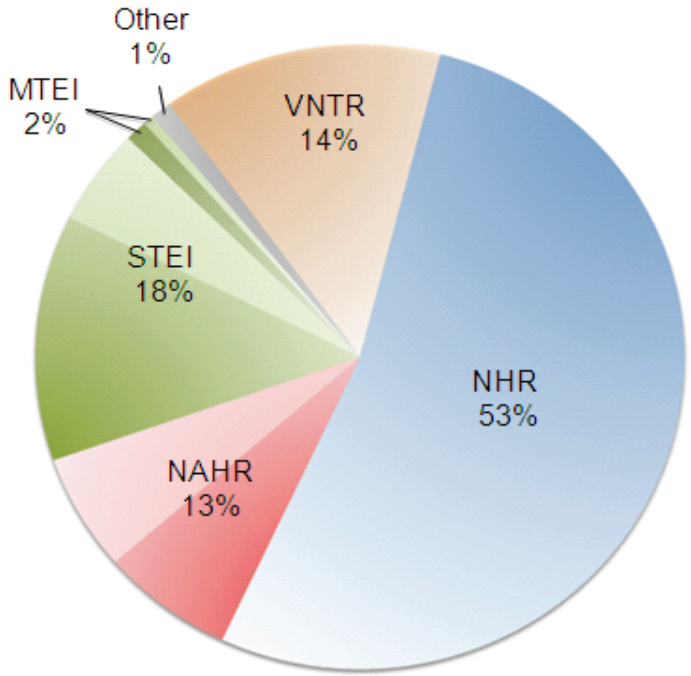
Korbel et al. Science 19 October 2007: Vol. 318. no. 5849, pp. 420 - 426

Mechanism Distribution

Published SVs



1KG SVs



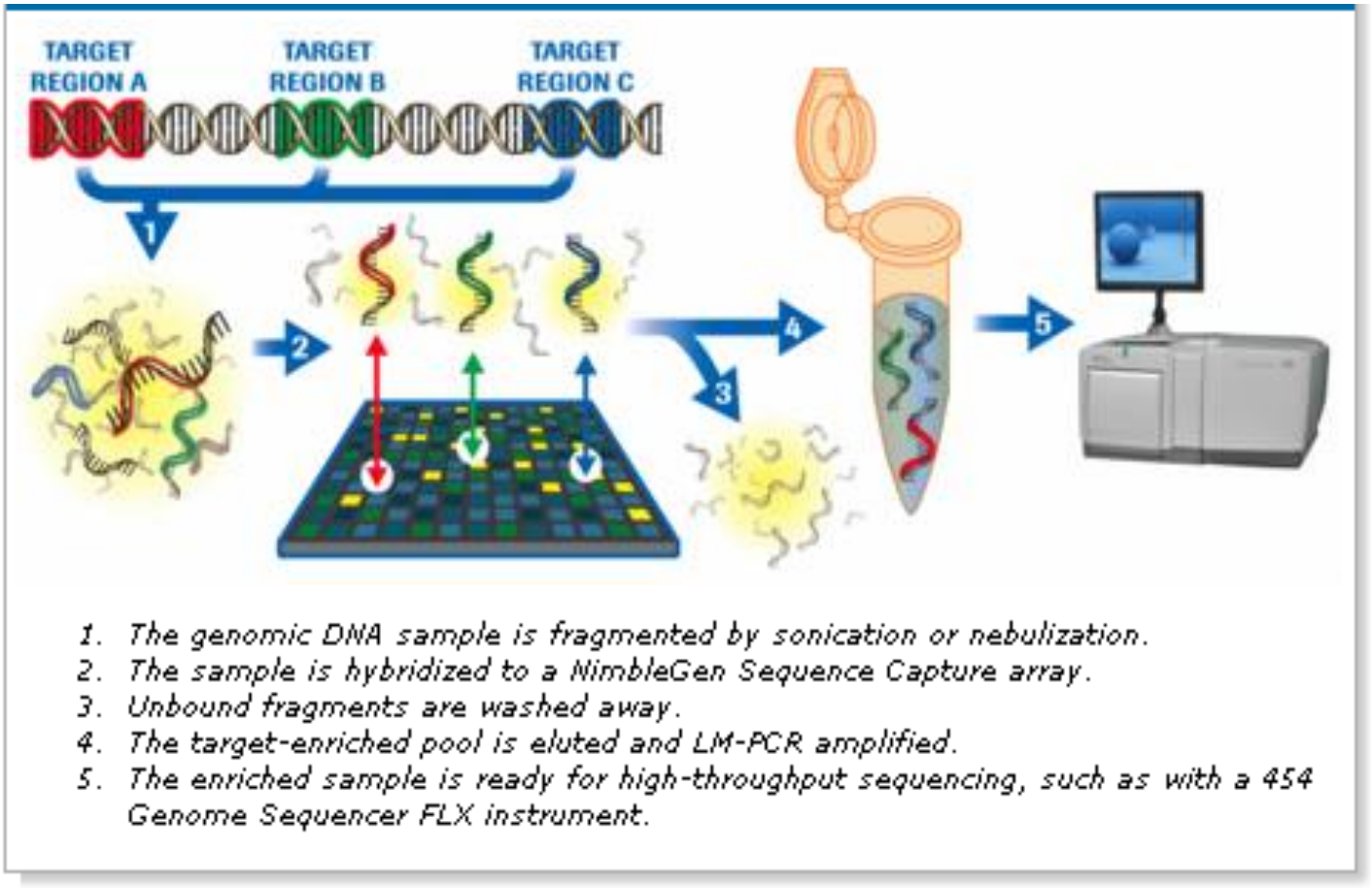
1. Targeted Sequencing

- hybridize genomic DNA to capture array
- wash away unbound fraction
- Elute off-target DNA
- Sequence with 454 Titanium (~400 bp reads)

2. SV-CapSeq analytical pipeline

- Map reads using Megablast; Best hit placement
- Intersect placements with target regions
- Precisely align reads with Needleman-Wunsch to identify split reads: SV validated, breakpoint sequence found

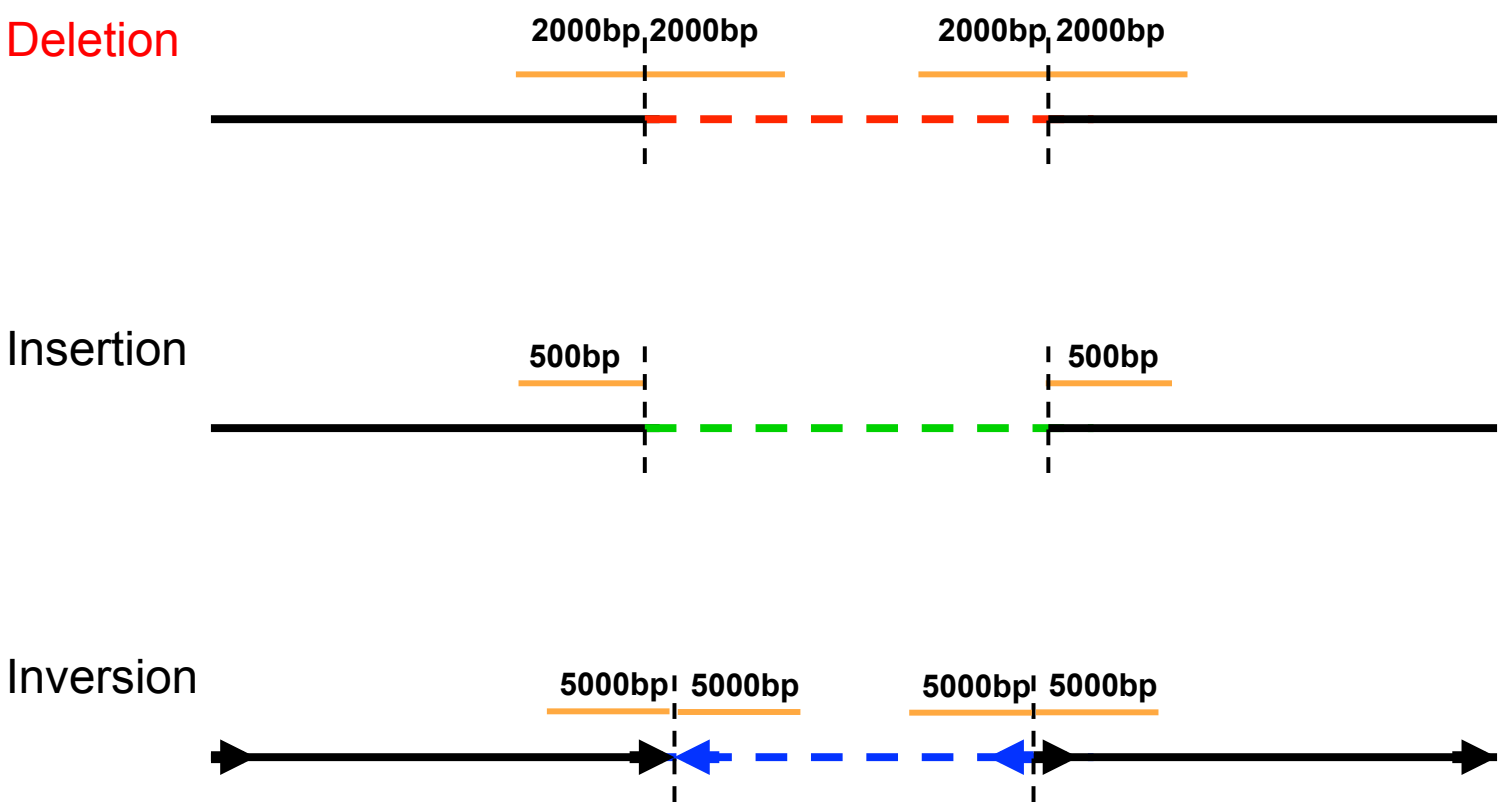
Array Capture Sequencing



1. *The genomic DNA sample is fragmented by sonication or nebulization.*
2. *The sample is hybridized to a NimbleGen Sequence Capture array.*
3. *Unbound fragments are washed away.*
4. *The target-enriched pool is eluted and LM-PCR amplified.*
5. *The enriched sample is ready for high-throughput sequencing, such as with a 454 Genome Sequencer FLX instrument.*

Roche-NimbleGen

SV-CapSeq: Array Design



Contents of the SV-CapSeq array v1.0

2.1 million oligomers tiling the target regions of the genome:

1839 deletion CNVs from (mostly) short read Solexa data (1000 Genome Project)

From long read 454 paired-end data:

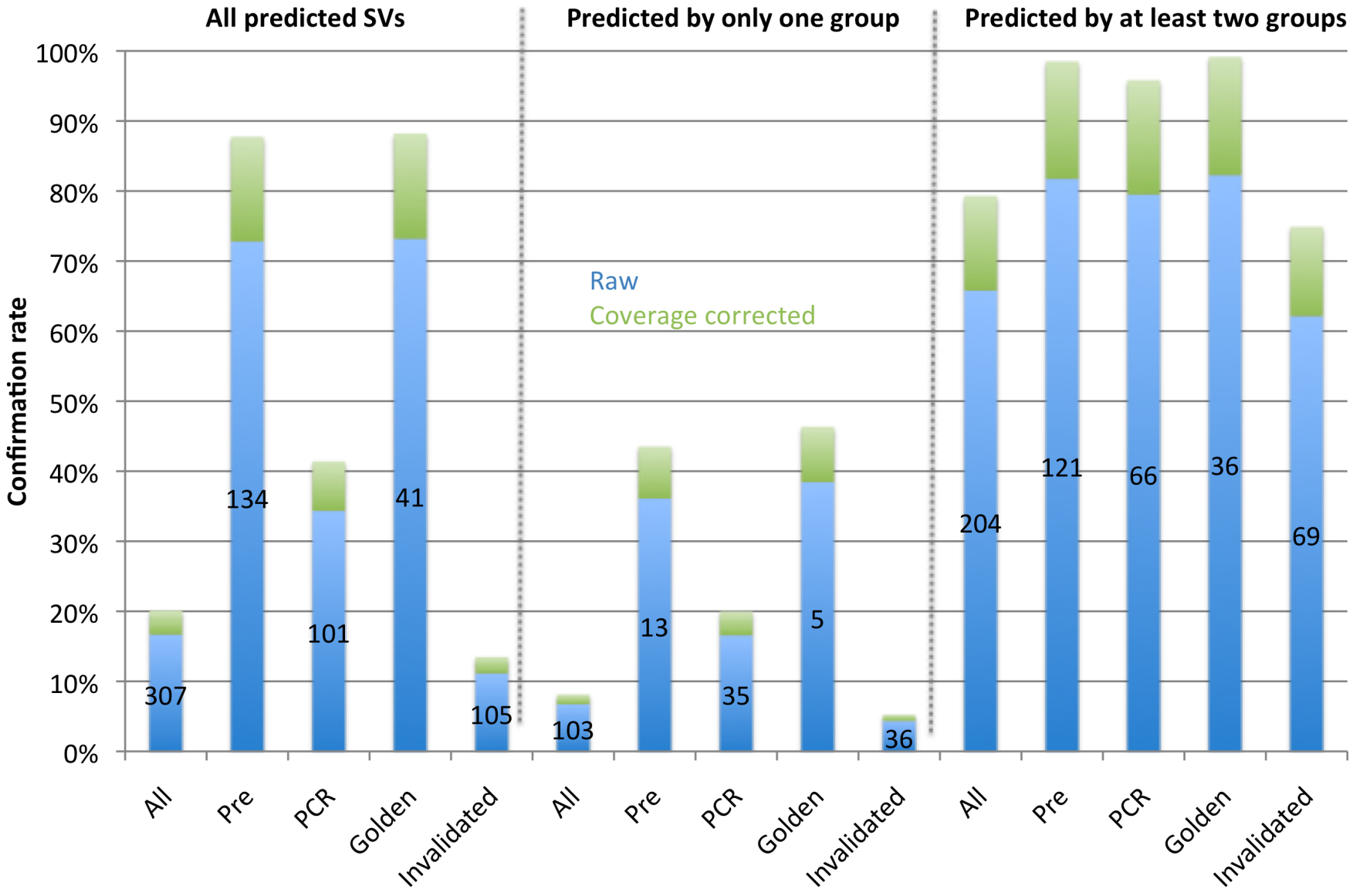
575 deletion CNVs

296 insertions CNVs

191 inversions SVs

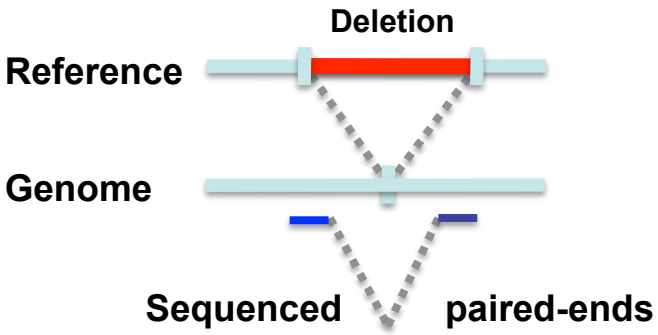
(plus Split-Read indel predictions, Zhengdong Zhang)

Confirmation rate by overlap

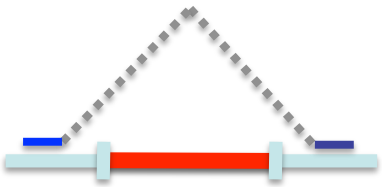


Methods to Find SVs

1. Paired ends

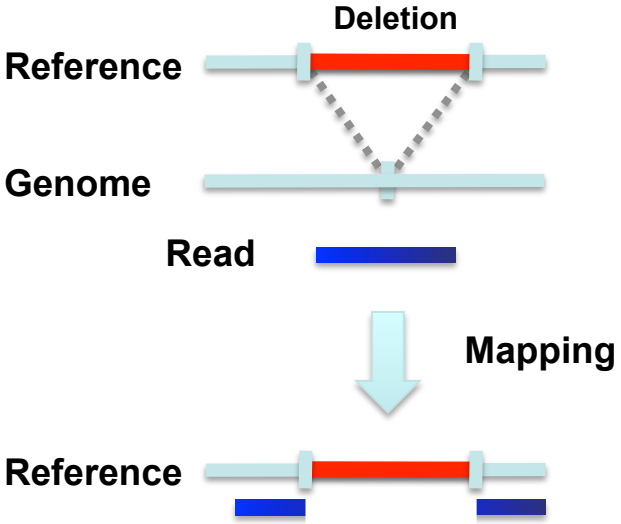


Mapping

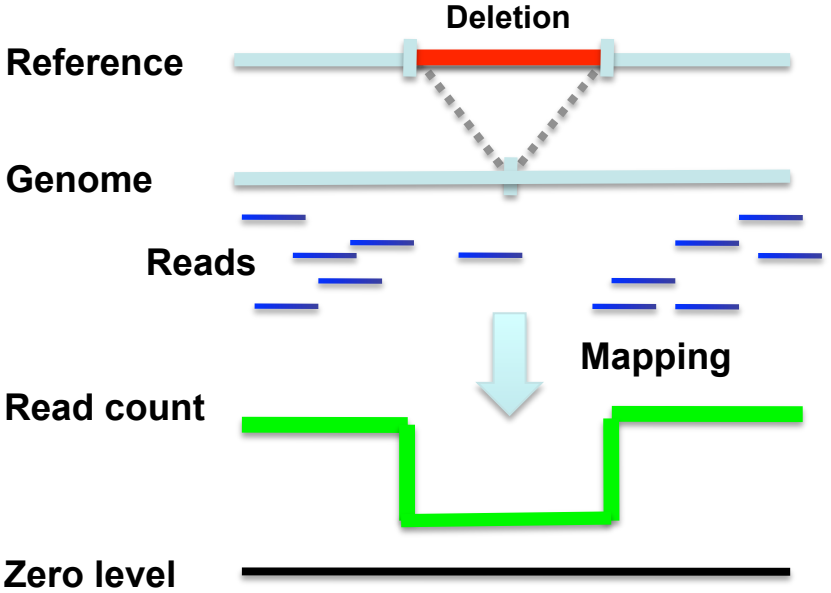


Reference

2. Split read



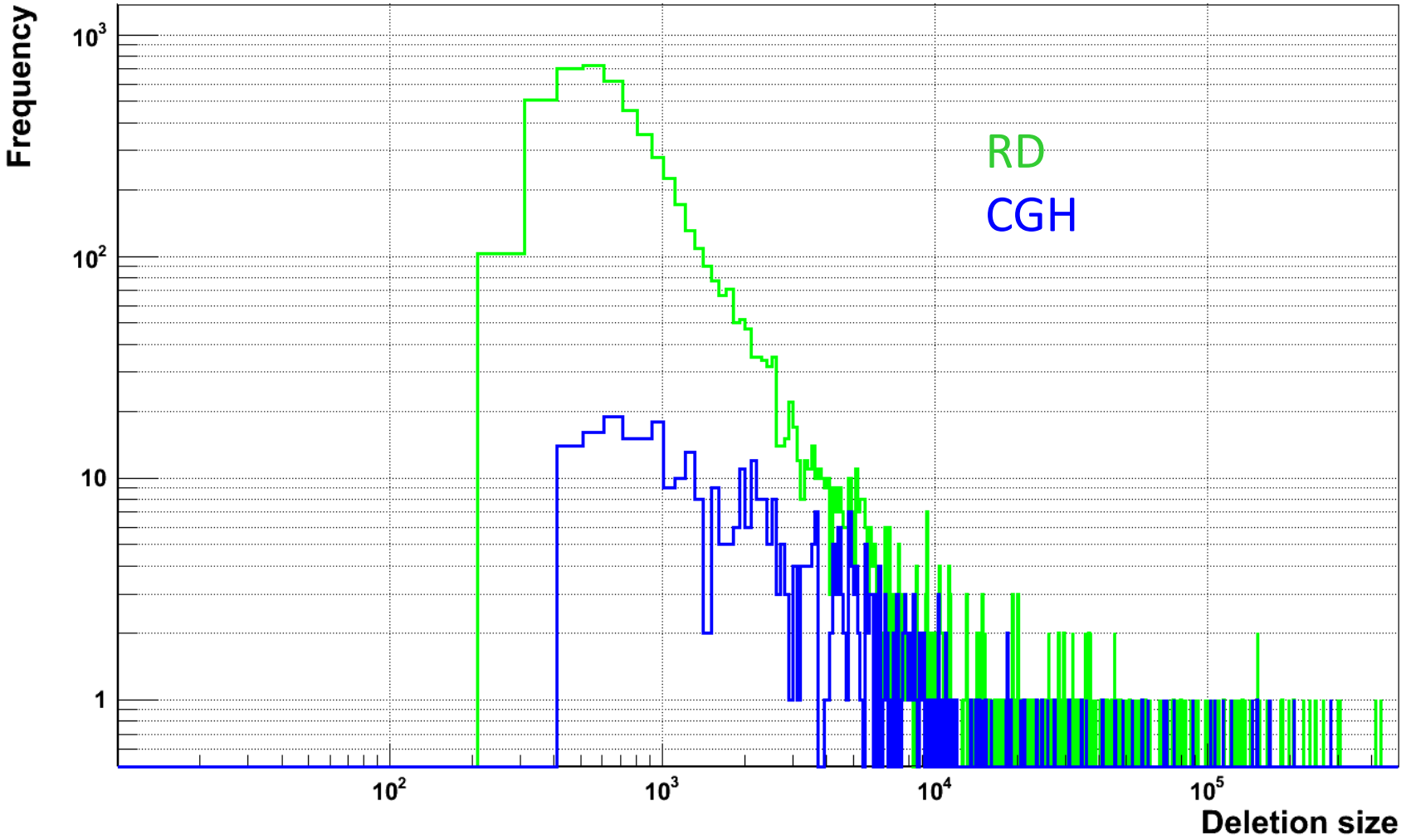
3. Read depth (or aCGH)



4. Local Reassembly

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

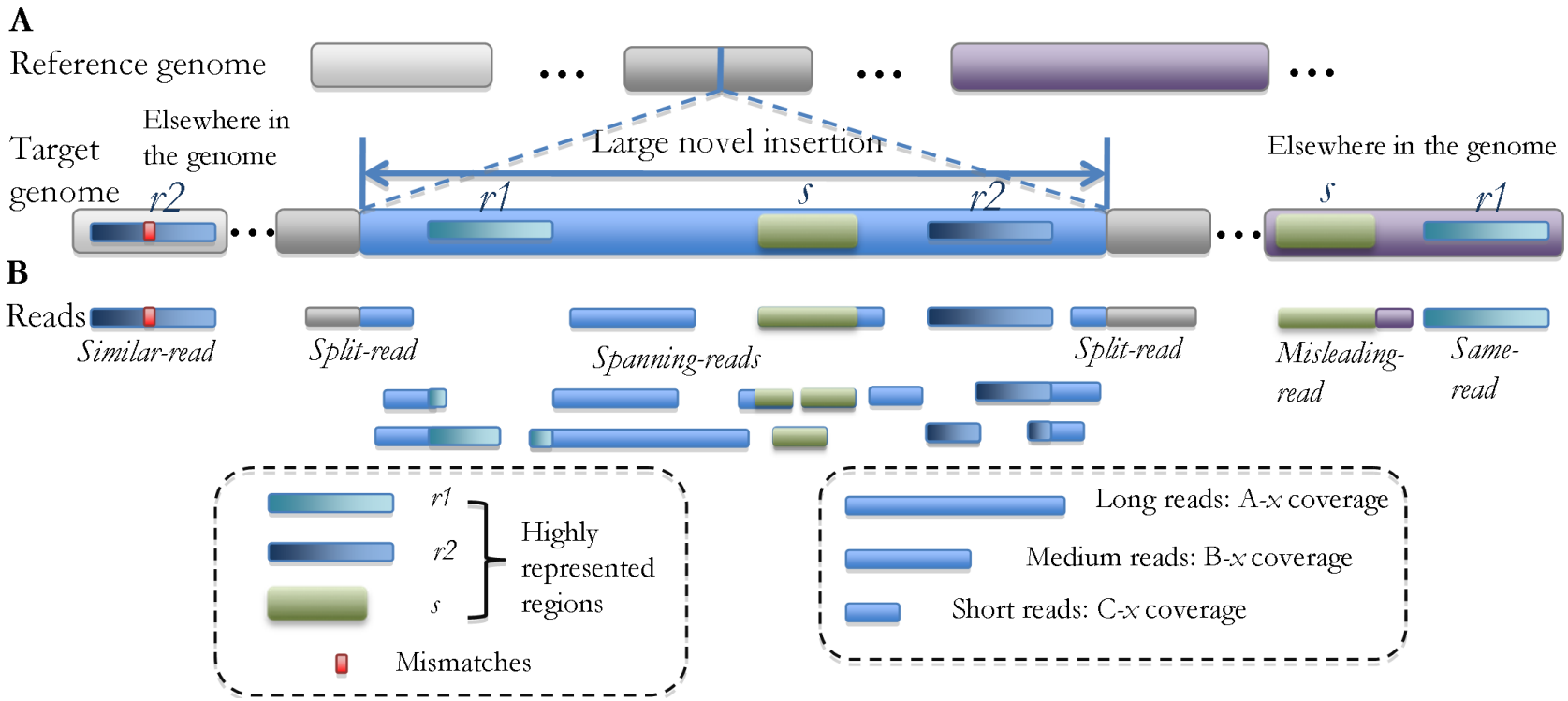
CNV discovery: RD vs CGH



[Daughter in Caucasian trio, NA12878] [CGH prediction are from Conrad et al., Nature, 2009]

Optimal integration of sequencing technologies:
Local Reassembly of large novel insertions

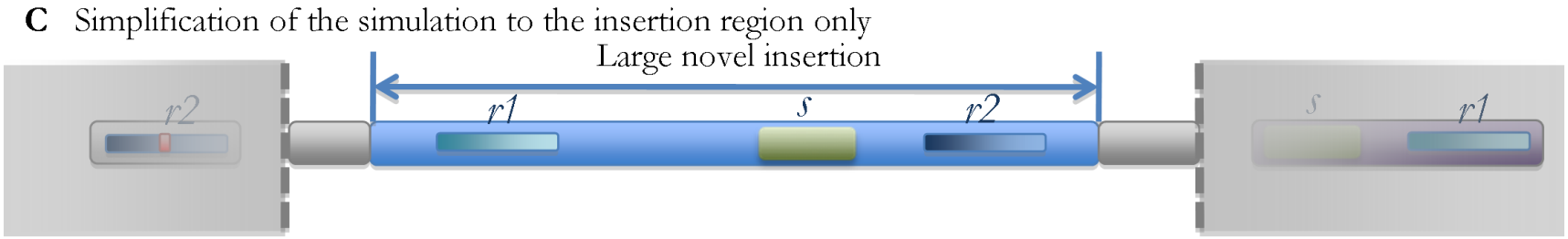
Given a fixed budget, what are the sequencing coverage A, B and C that can achieve the maximum reconstruction rate (on average/worst-case)? Maybe a few long reads can bootstrap reconstruction process.



Du et al. (2009), PLoS Comp Biol, in press

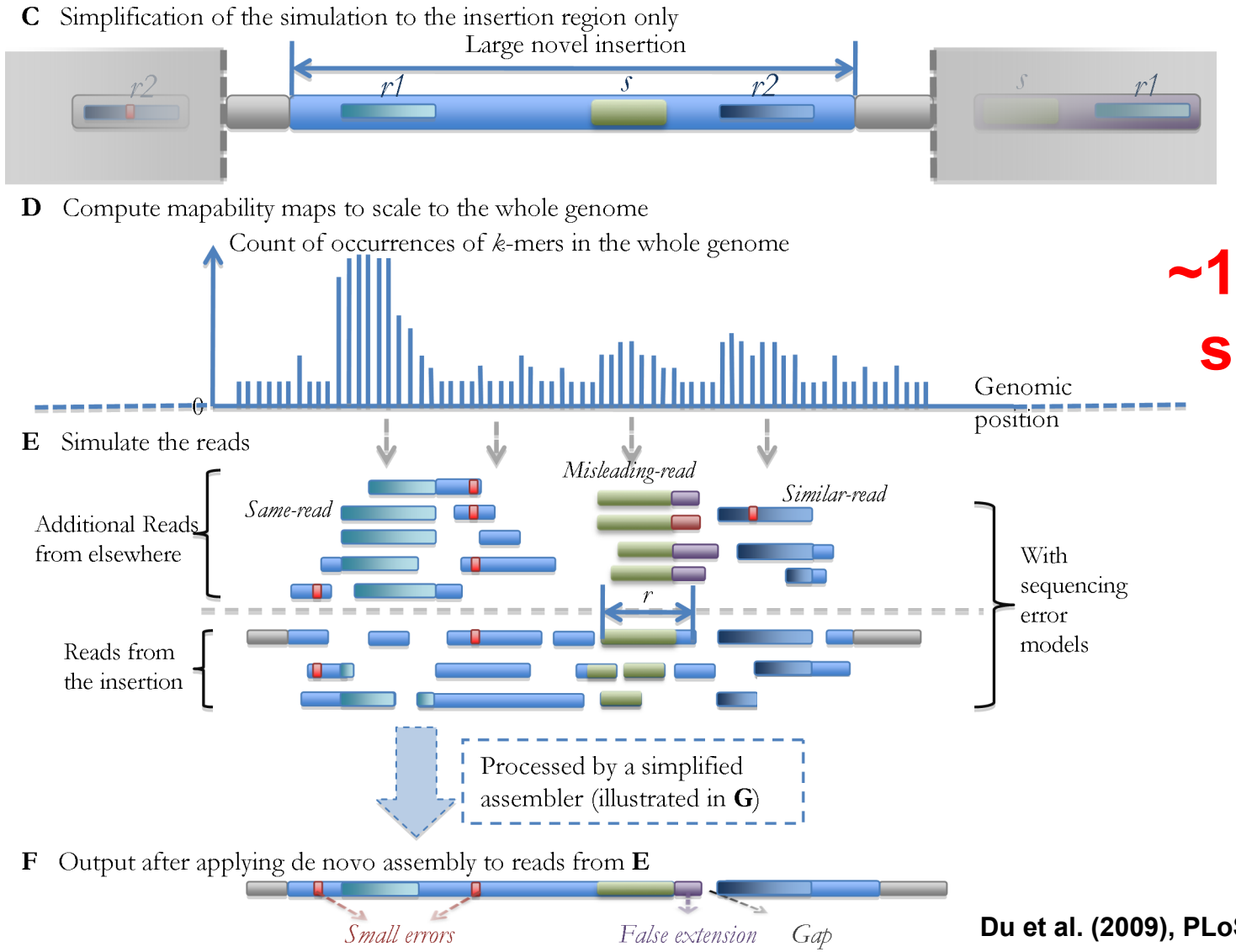
Optimal integration of sequencing technologies:
Need Efficient Simulation

Different combinations of technologies (i.e. read lenghs) very expensive to actually test.
Also computationally expensive to simulate.
(Each round of whole-genome assembly takes >100 CPU hrs; thus, simulation exploring 1K possibilities takes 100K CPU hr)



Du et al. (2009), PLoS Comp Biol, in press

Optimal integration of sequencing technologies:
Efficient Simulation Toolbox using Mappability Maps

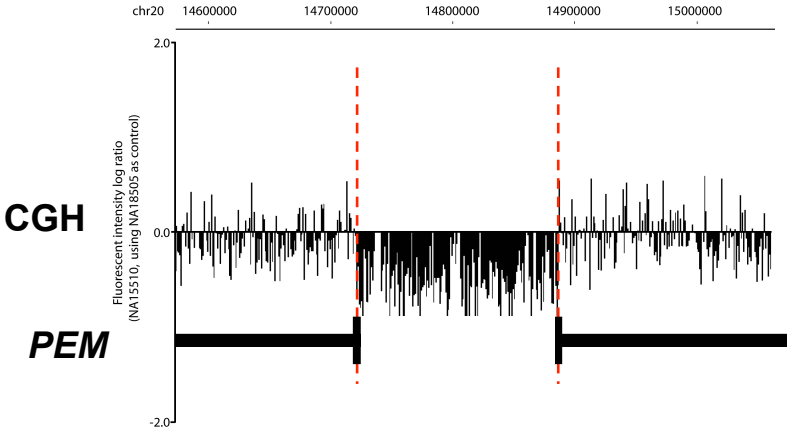


~100,000 X speedup

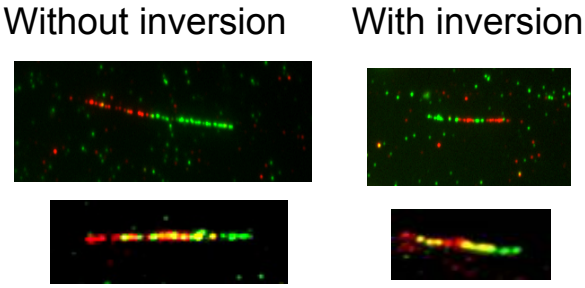
Du et al. (2009), PLoS Comp Biol, in press

Experimental Validation

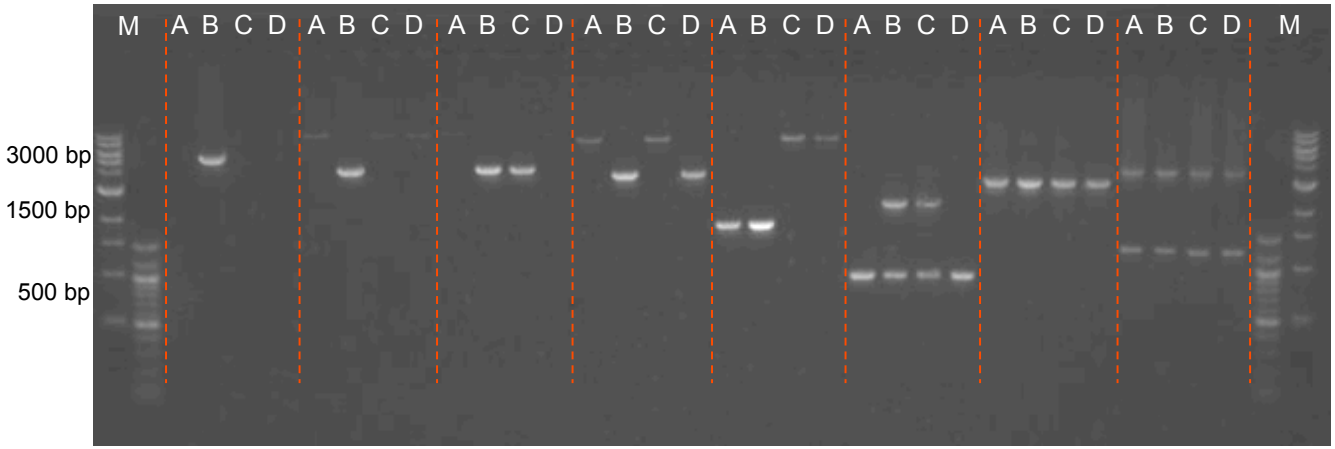
A) CGH



B) Fiber-FISH (For inversions)



C) PCR (Often 4 People)



>500 SVs validated

~50% SV are in more than one ethnic group