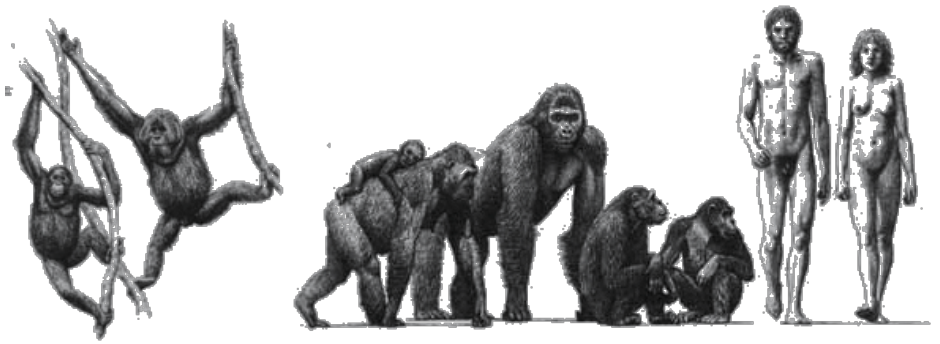
Structural variation

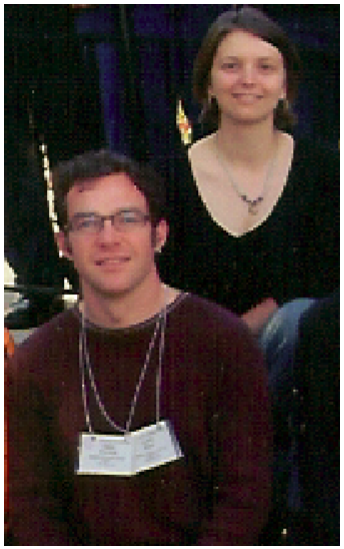
Programming for Biology

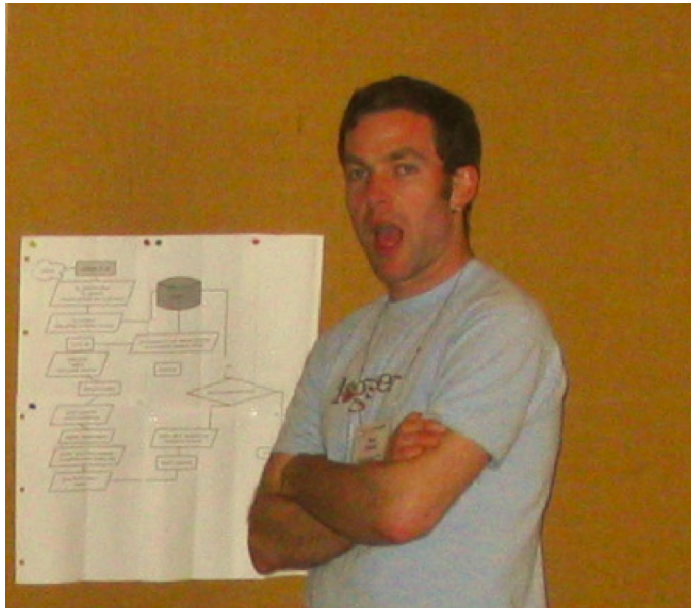
CSH, October 2014

Tomas Marques-Bonet
ICREA Research Professor
Institut de Biologia Evolutiva



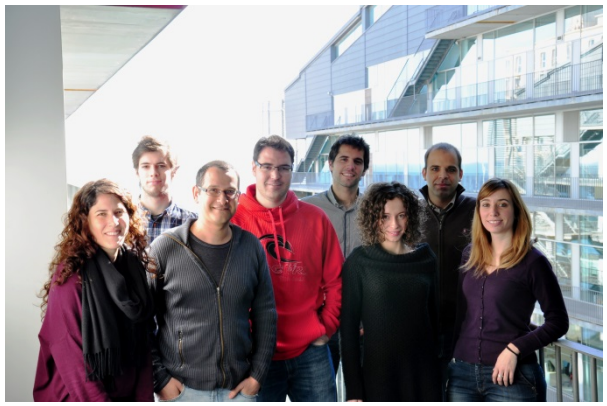
11 years from my 1st CSH!





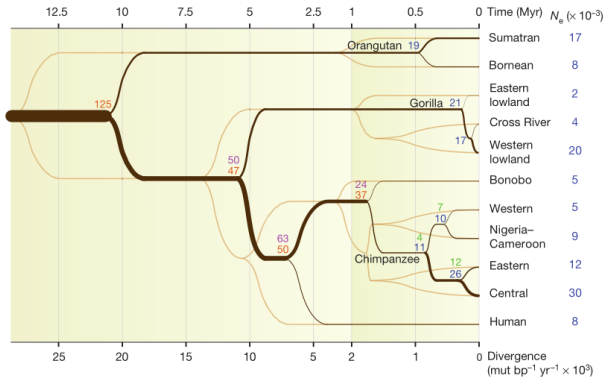
Who we are?

- Evolutionary genomics
  - Barcelona, Biomedical Research Park



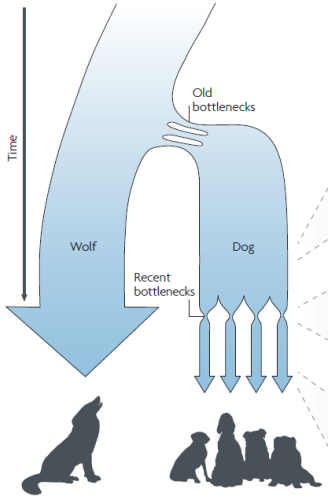
What do we do?

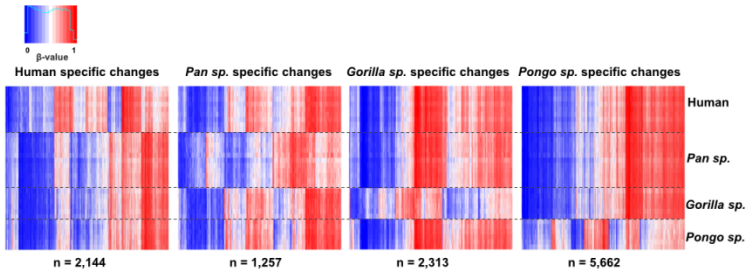
- Natural selection on human evolution



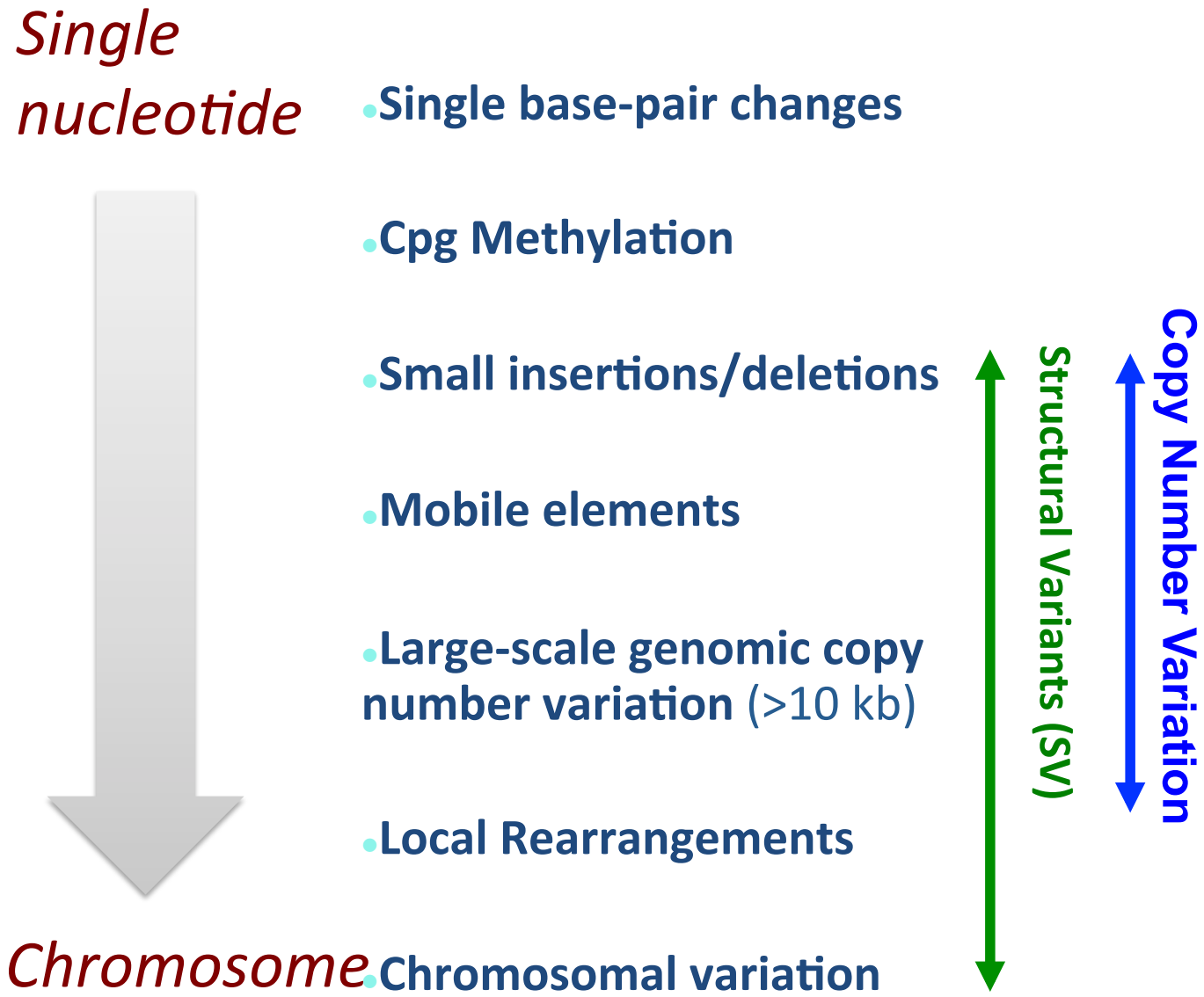
- Transcriptome and Epigenetics in Primates

- Canid evolution

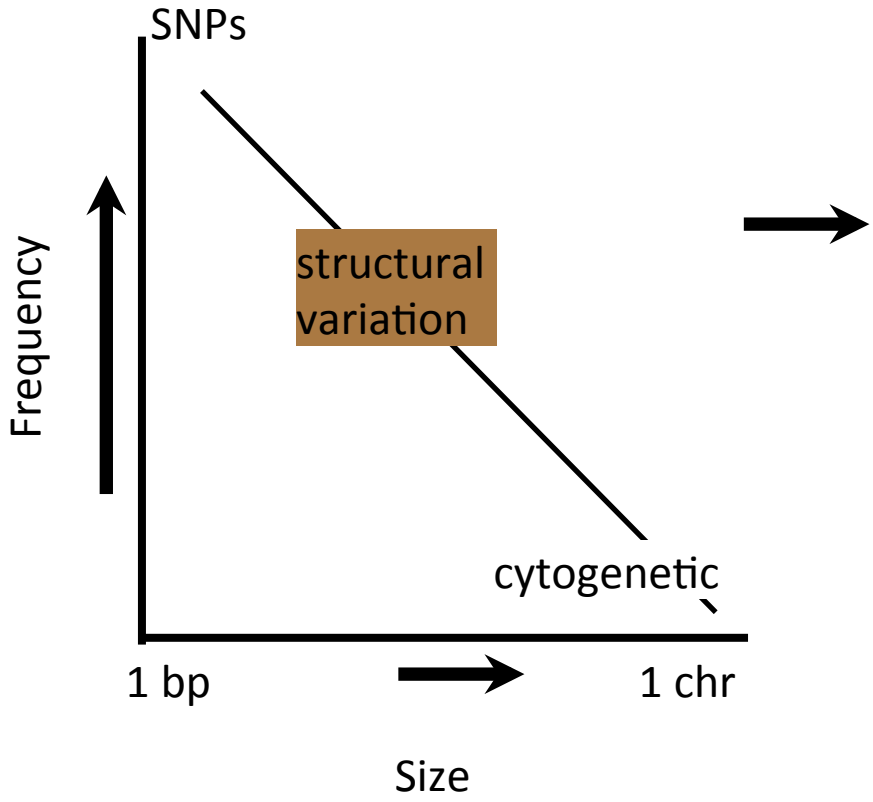




Continuum of Genomic Variation

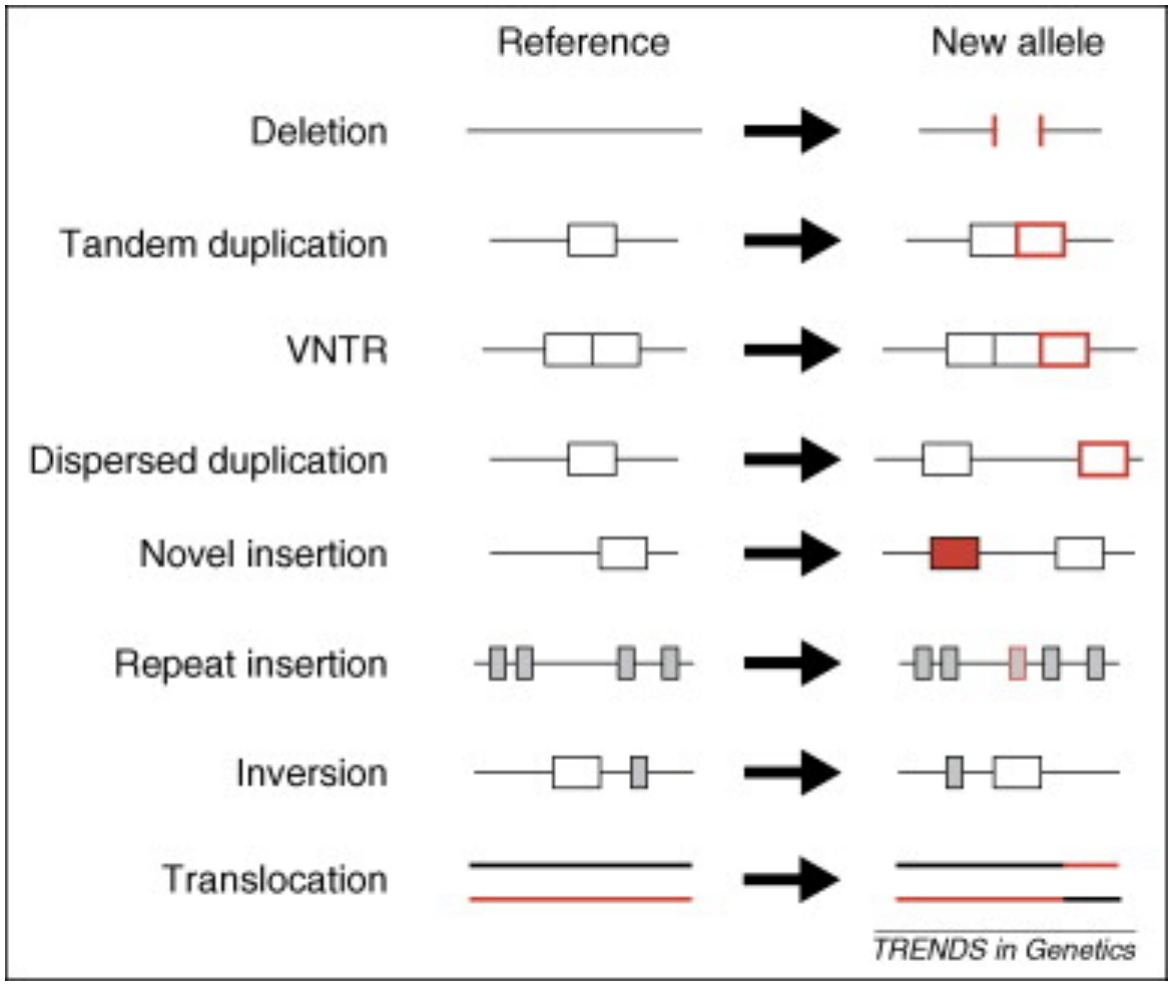


Genomic Structural Variation



- Gene-altering, *e.g.* immune response, drug metabolism
- Abundant: majority of human heterozygosity
- Numerous plausible functional consequences

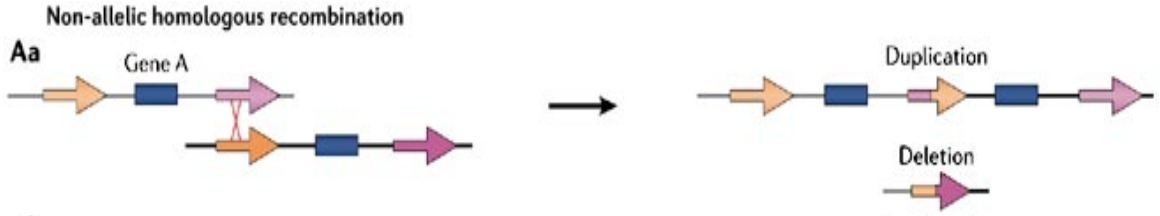
Types of Structural Variation



Hurles *et al.* 2008

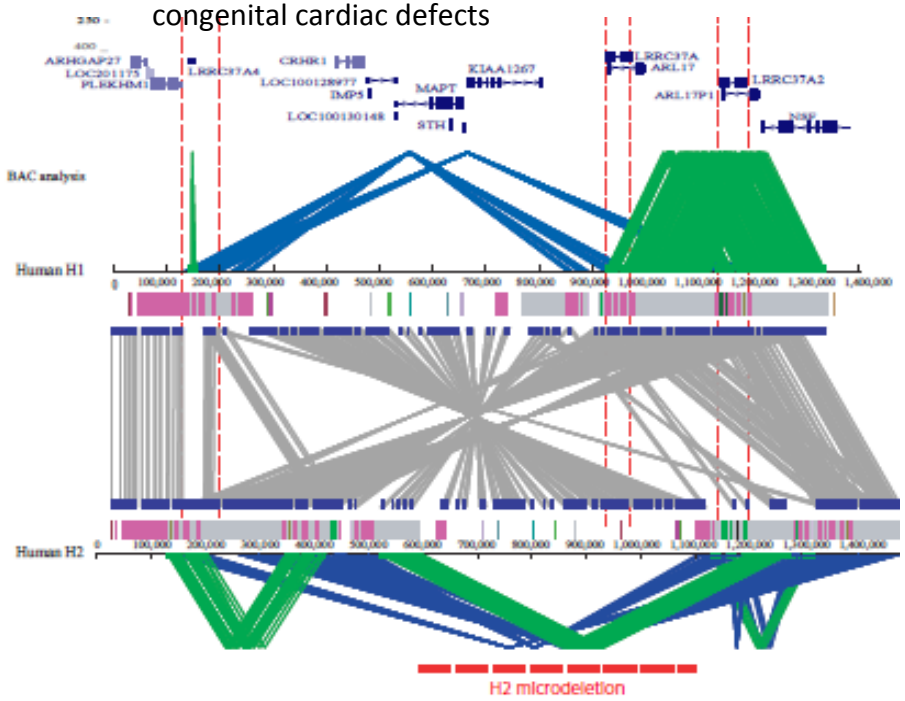
Why Study Structural Variation?

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



17q21.31 deletion syndrome

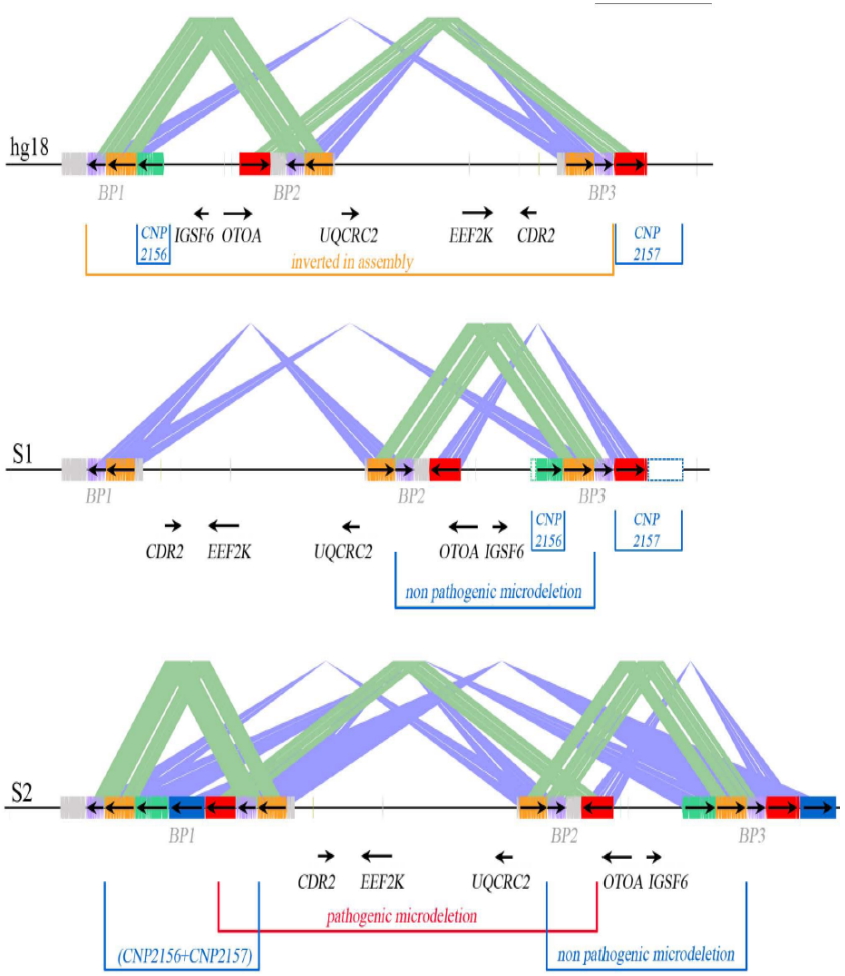
MR, global delay and
congenital cardiac defects



Zody *et al.* *Nature* Genetics (2008)

16p12.1 deletion syndrome

childhood intellectual disability, developmental delay



Antonacci *et al.* *Nature* Genetics (2010)

Challenges of CNV studies

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

Methods to Find SVs

Experimental approach

ArrayCGH (SNP based and genomic)

Sequence based

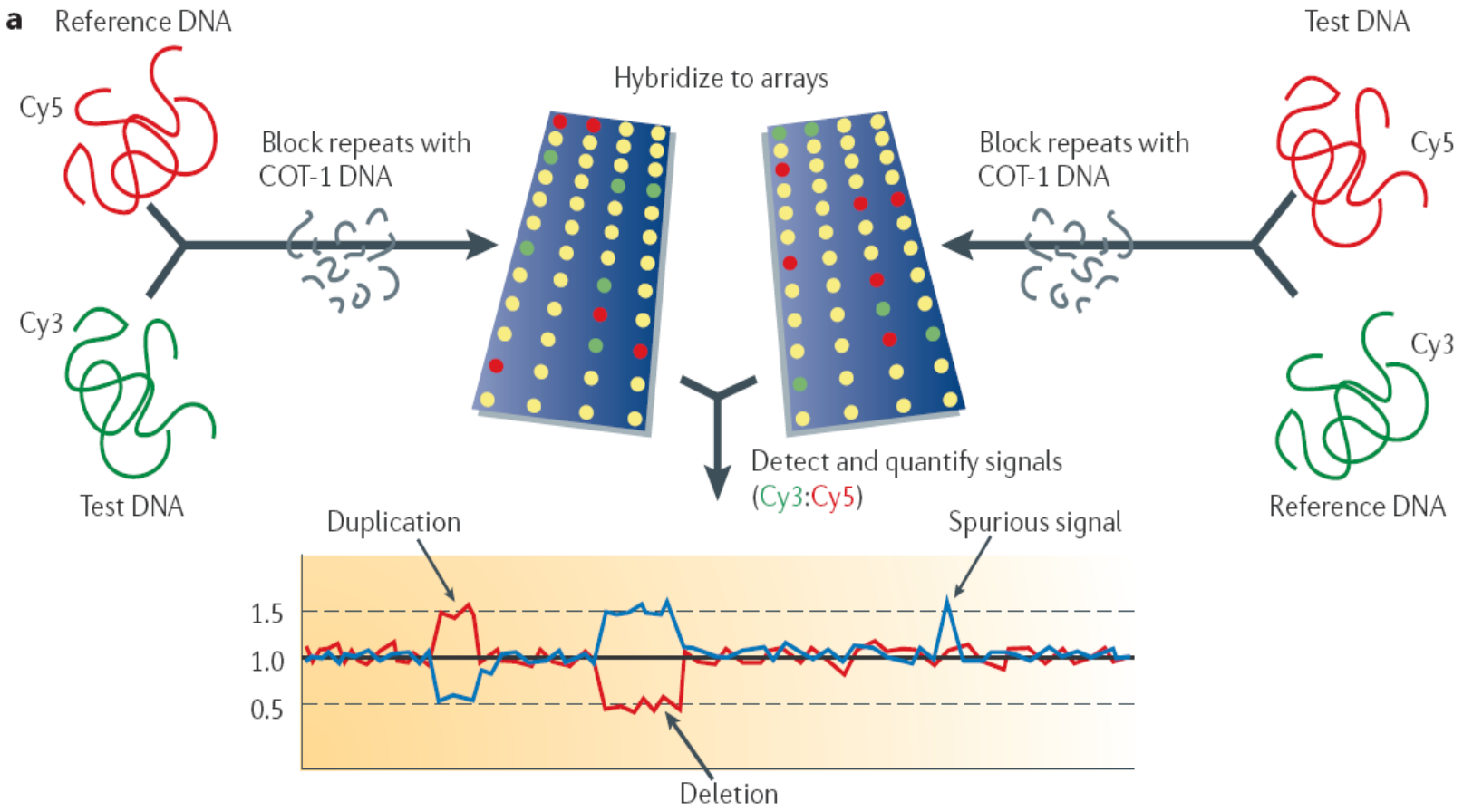
Local and *de novo* assembly

Read pair analysis

Read depth analysis

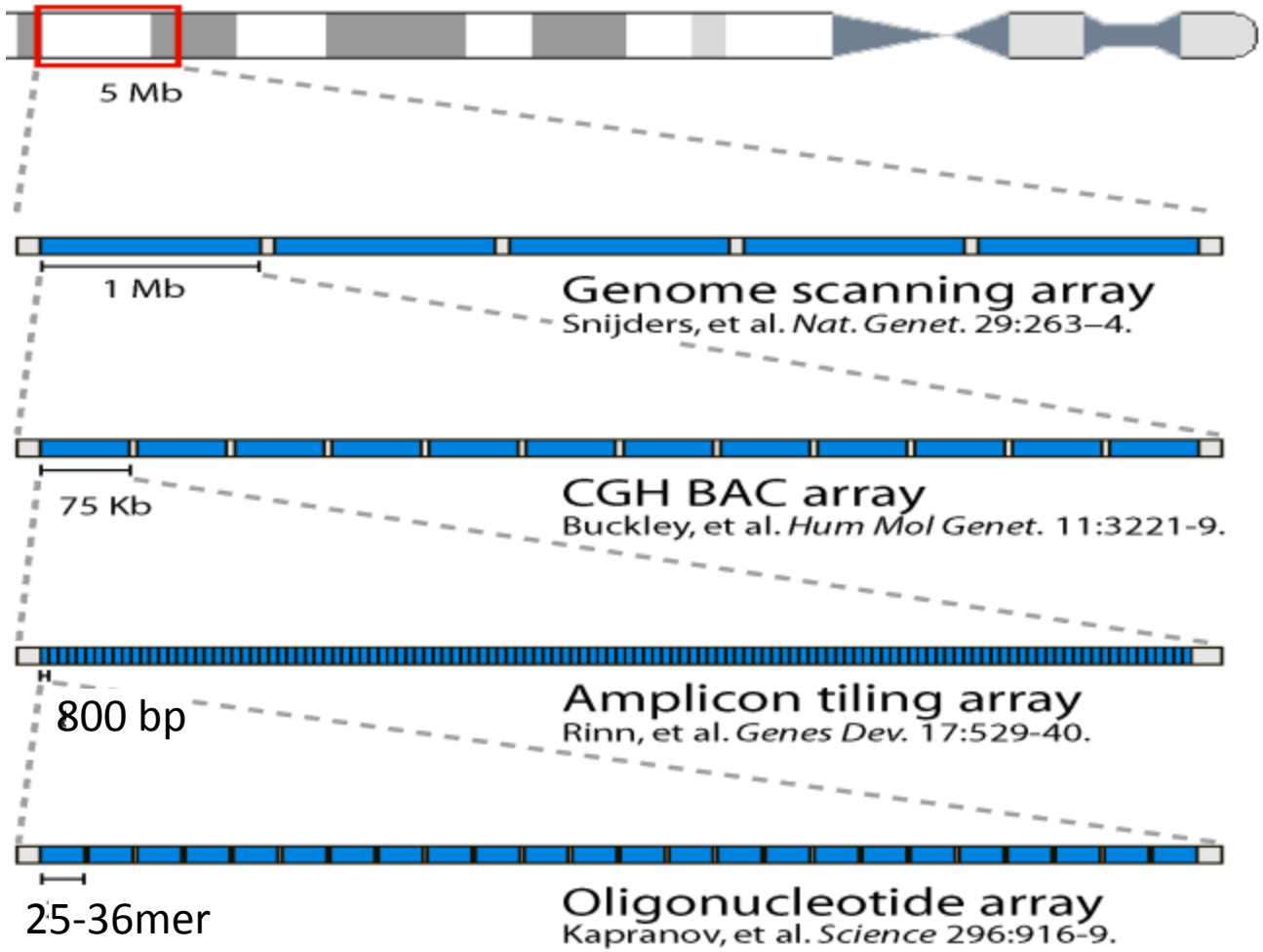
Split read analysis

METHOD 1: Copy Number Variation: Array Comparative Genomic Hybridization



*Modified:*Feuk et al. *Nat Rev Genet* 2006

Genome Tiling Arrays



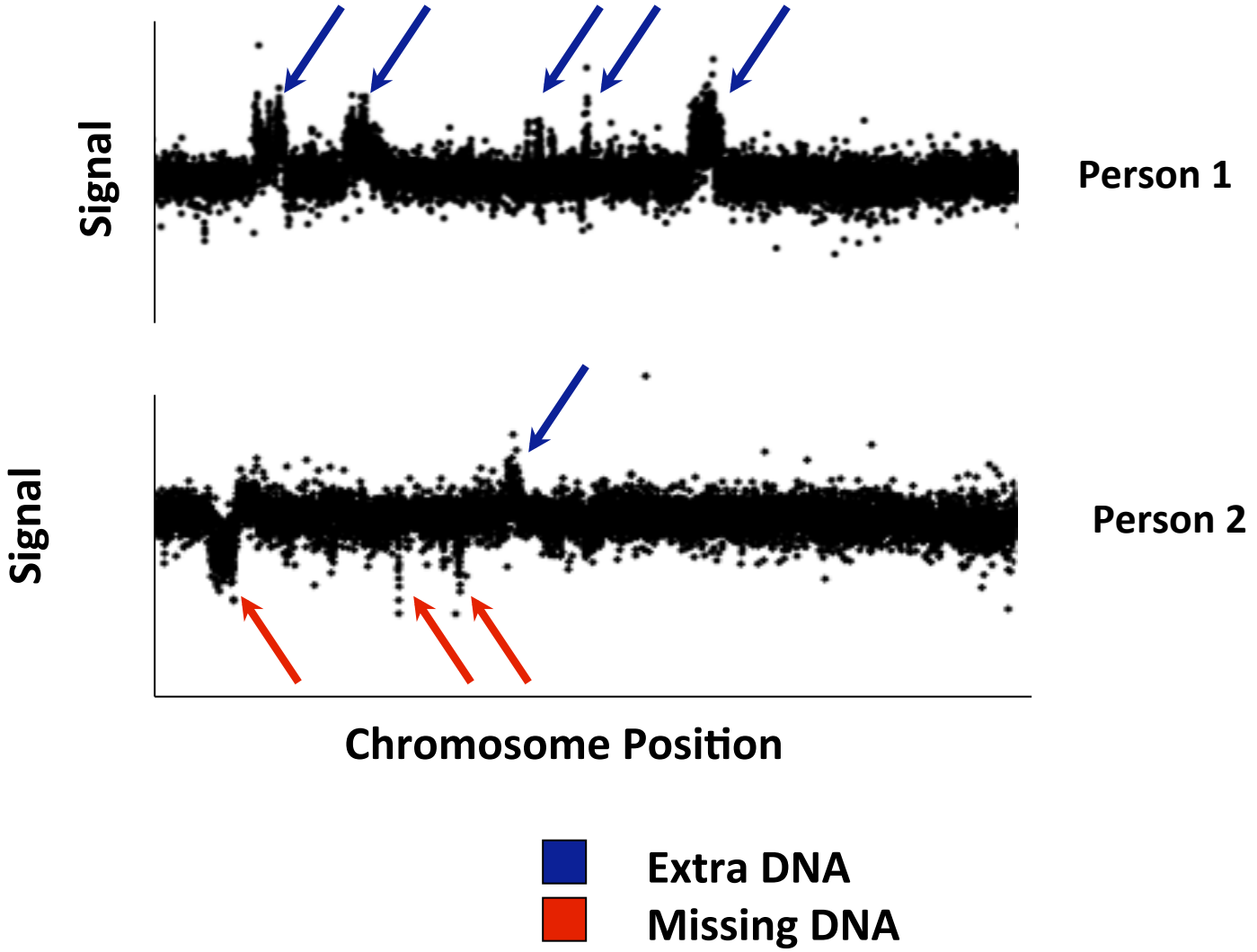
Typical Analysis Procedure

- For each probe, calculate a log2 ratio of test/reference
  - Log2 serves to center values around 0
  - Hemizygous deletion in test: log2(test/reference)=log2(1/2)=-1
  - Duplication in test:

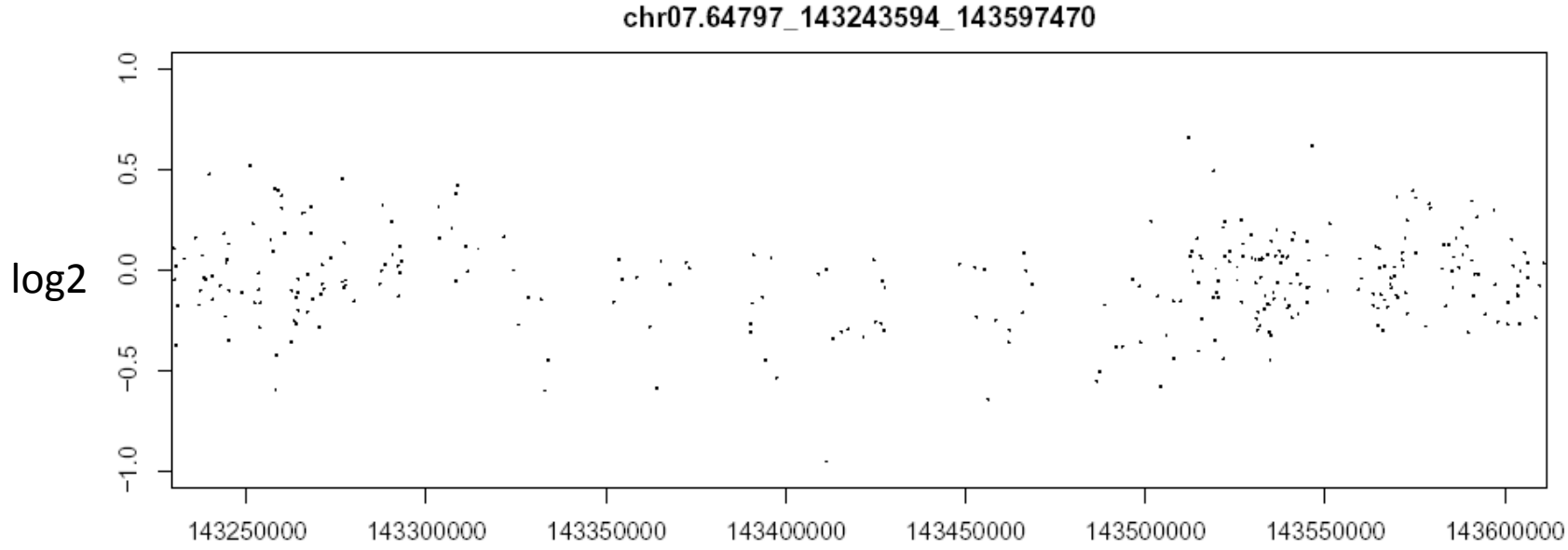
    log2(test/reference)=log2(3/2)=0.59
  - Homozygous duplication:

    log2(test/reference)=log2(4/2)=1

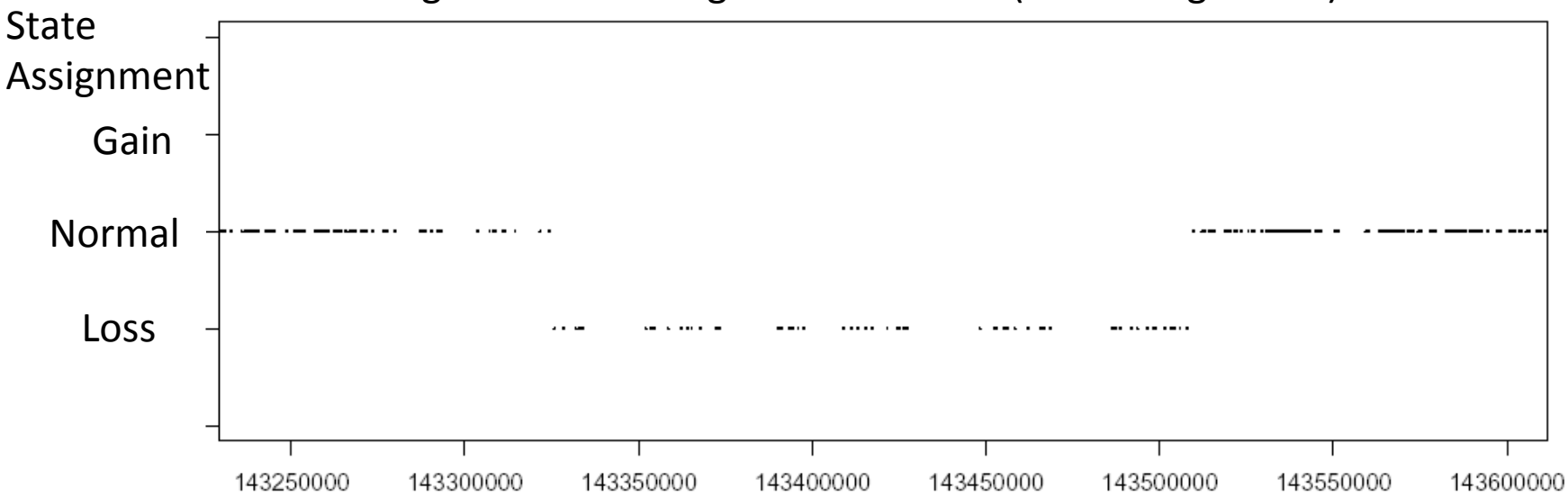
Copy Number Variations in the Human Genome



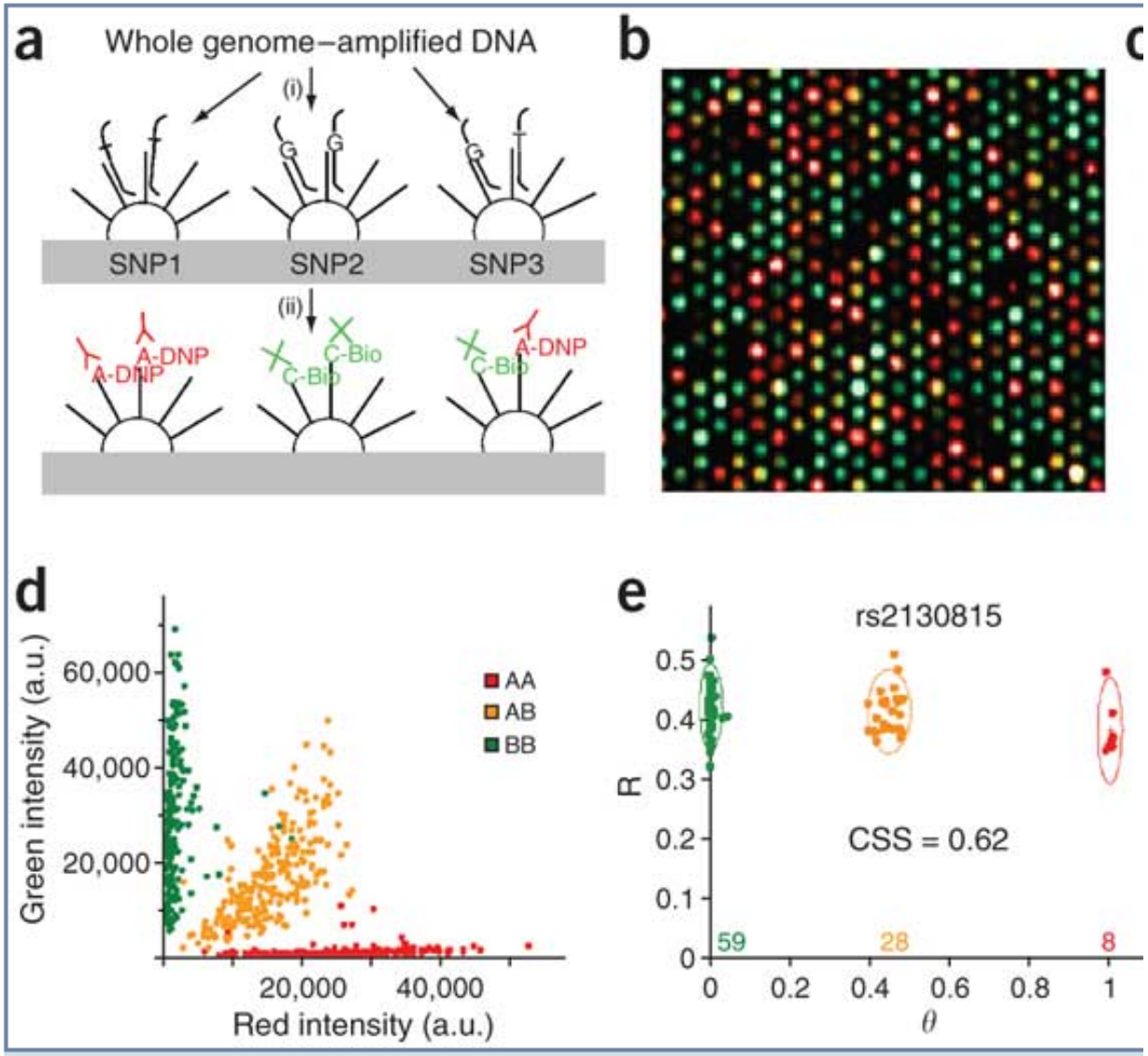
3-State HMM



Segmentation using a 3-state HMM (Viterbi Algorithm)

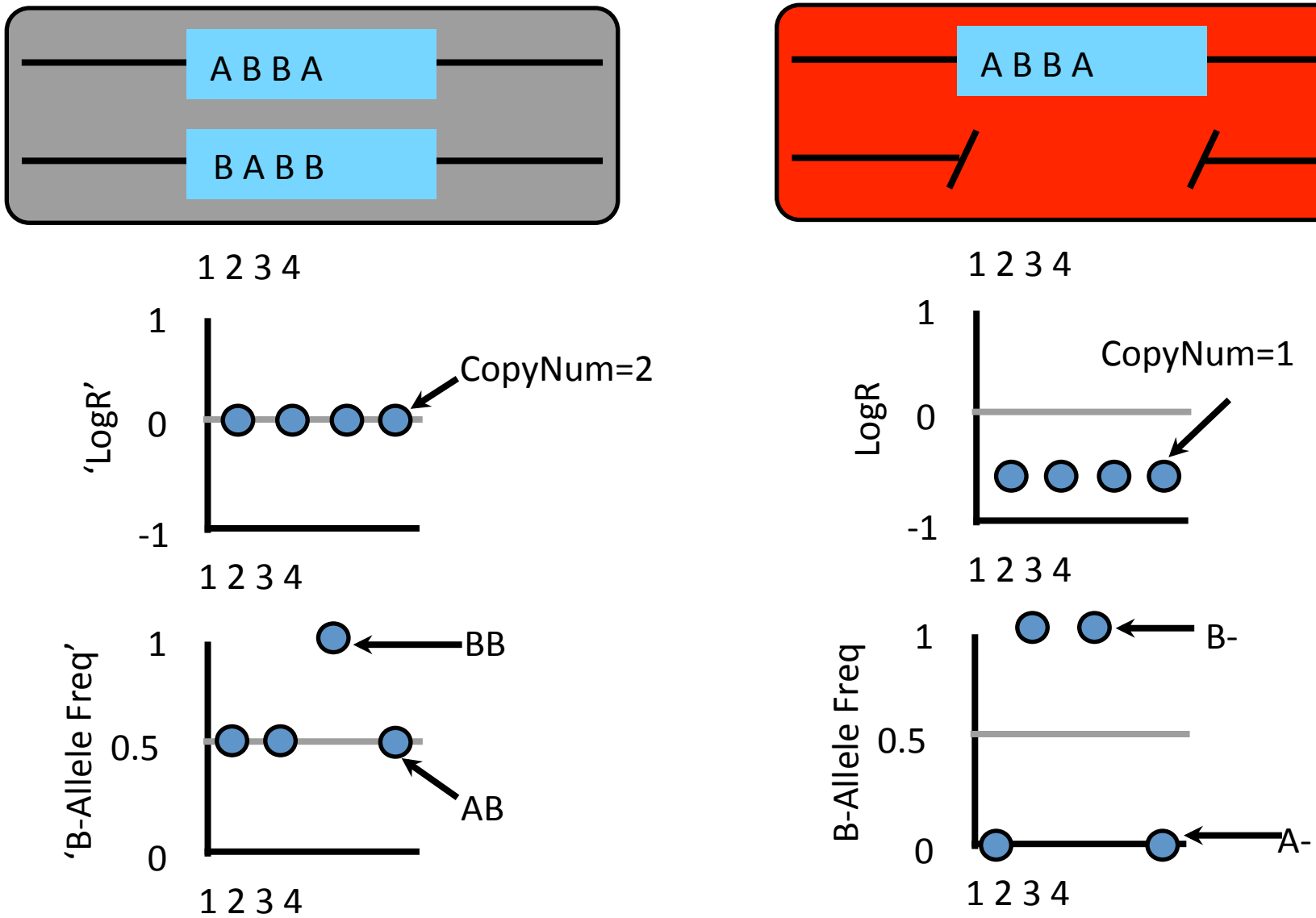


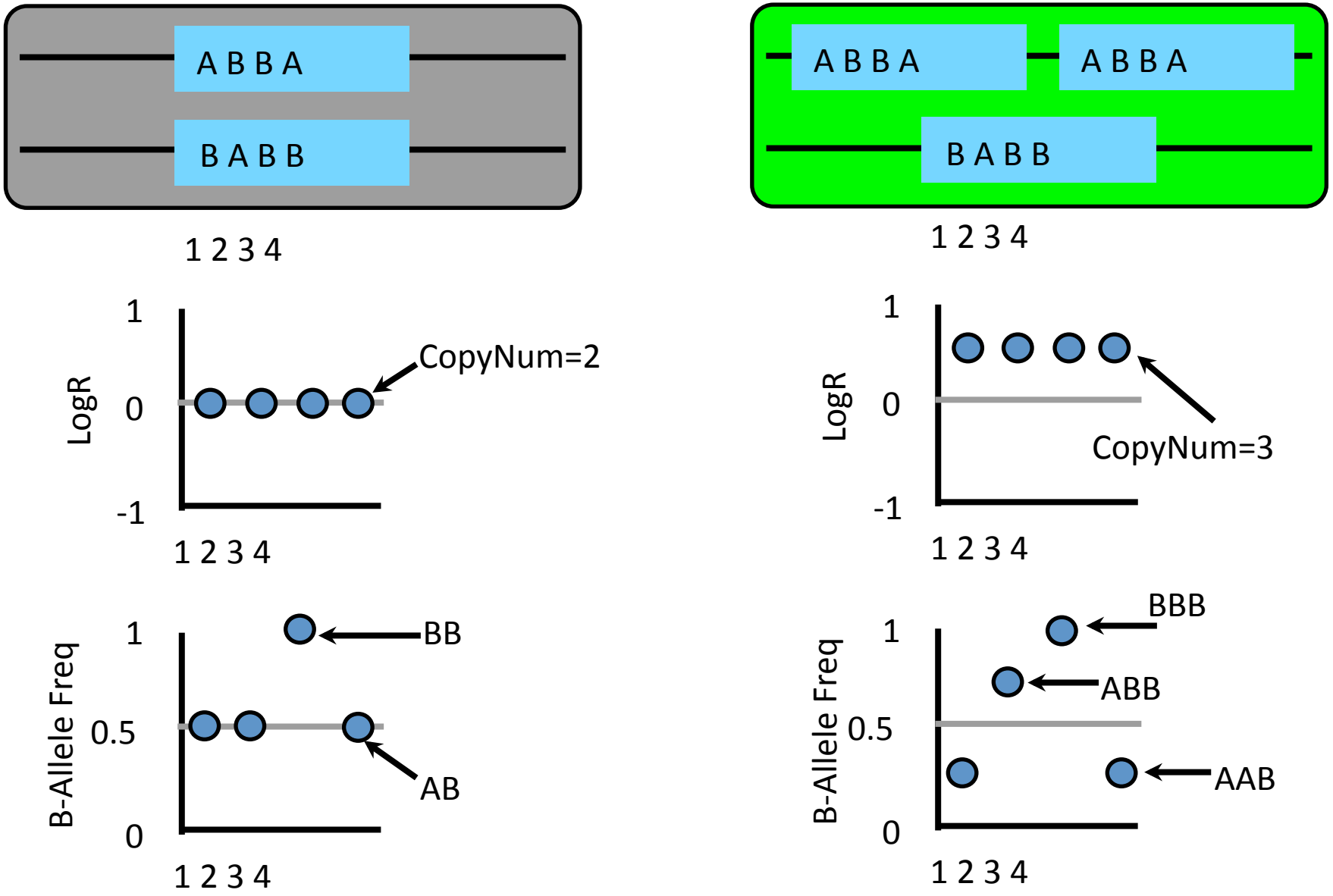
METHOD 2: Copy Number Variation: SNP genotyping Array

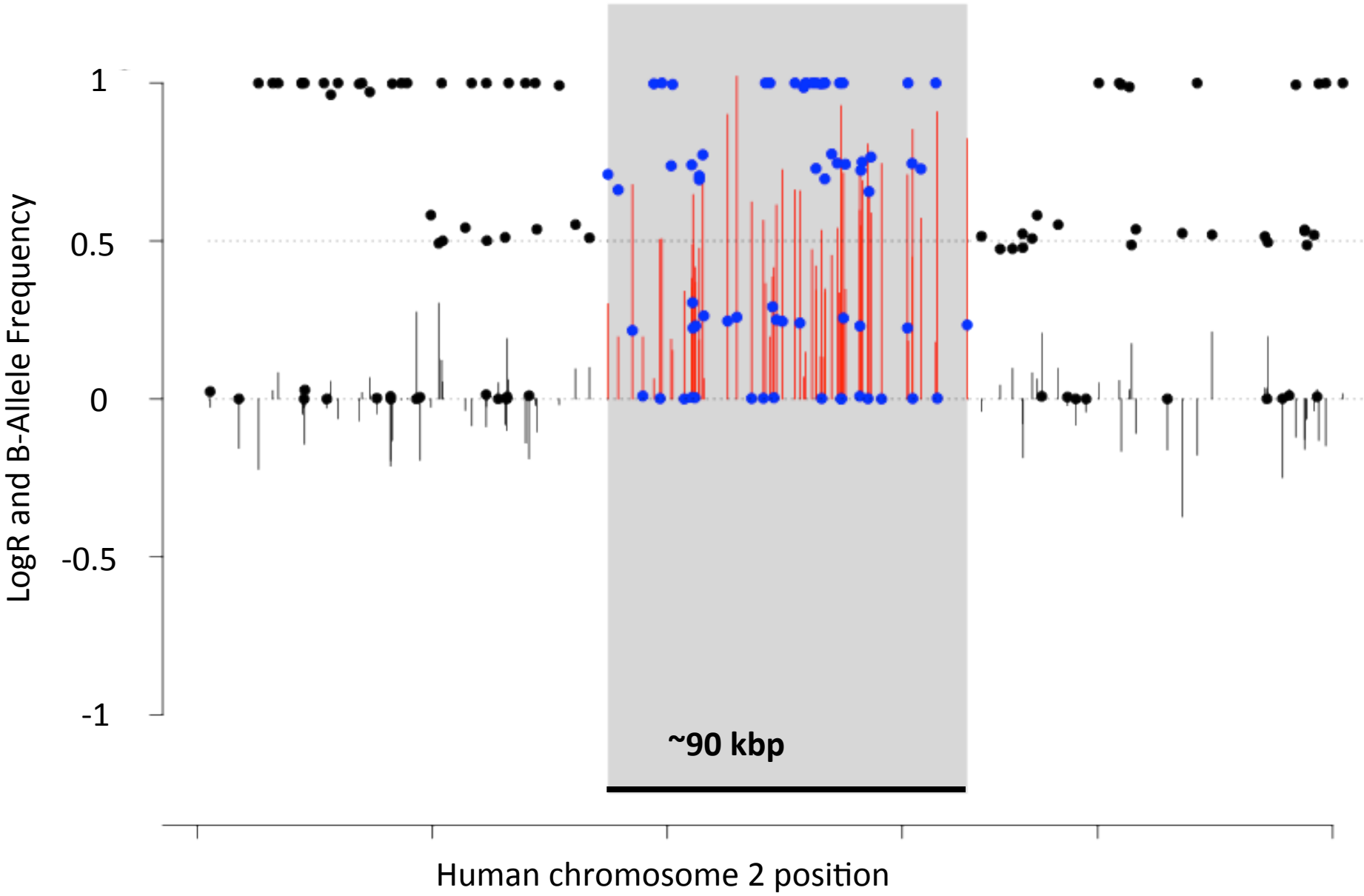


Steemers *et al.*

SNP Fluorescence-Based Deletion Discovery




A B B A
B A B B
1 2 3 4
LogR
1
0
-1
CopyNum=2
1 2 3 4
B-Allele Freq
1
0.5
0
BB
AB
1 2 3 4
A B B A
A B B A
B A B B
1 2 3 4
LogR
1
0
-1
CopyNum=3
1 2 3 4
B-Allele Freq
1
0.5
0
BBB
ABB
AAB
1 2 3 4


1
0.5
0
-0.5
-1
LogR and B-Allele Frequency
~90 kbp
Human chromosome 2 position

Sequencing Methods

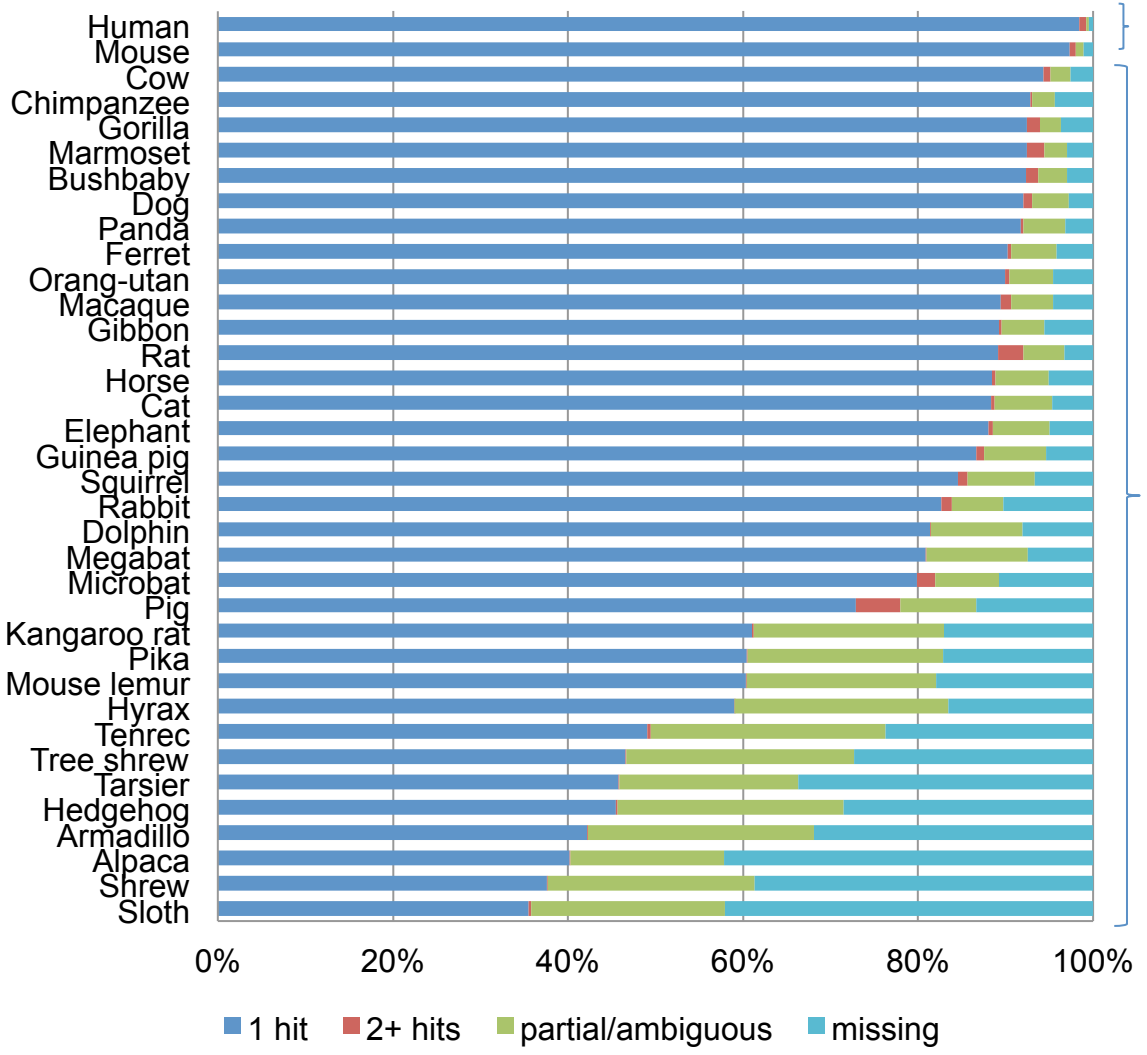
- Going Backwards… Sanger, 454, Illumina…..

- CNV and SV are hotspots of research… but reality is:
  - Limitations of the methods
    - Indirect methods. ALL have problems!!
  - What do we want?
    - Clone sequencing/Phasing (Moleculo?)
    - Finish sequence and better assemblies (PACBIO?)

De novo assemblies

- Theory vs. Reality
- Most assemblies (even with Sanger technology!) are collapsed.

Quality of “old days” assemblies



FINISHED

Phasing (Clone sequencing)

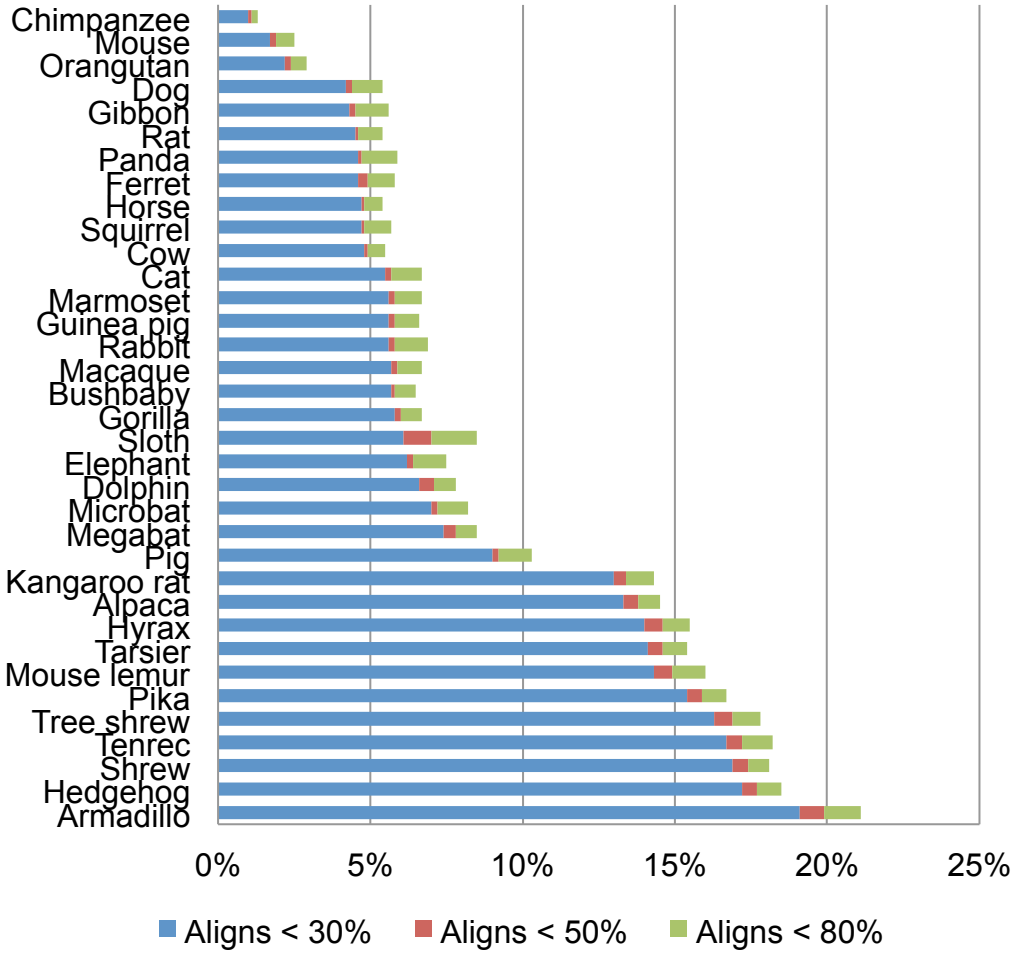
Clone end (BACs)

Long reads (Sanger)

DRAFT

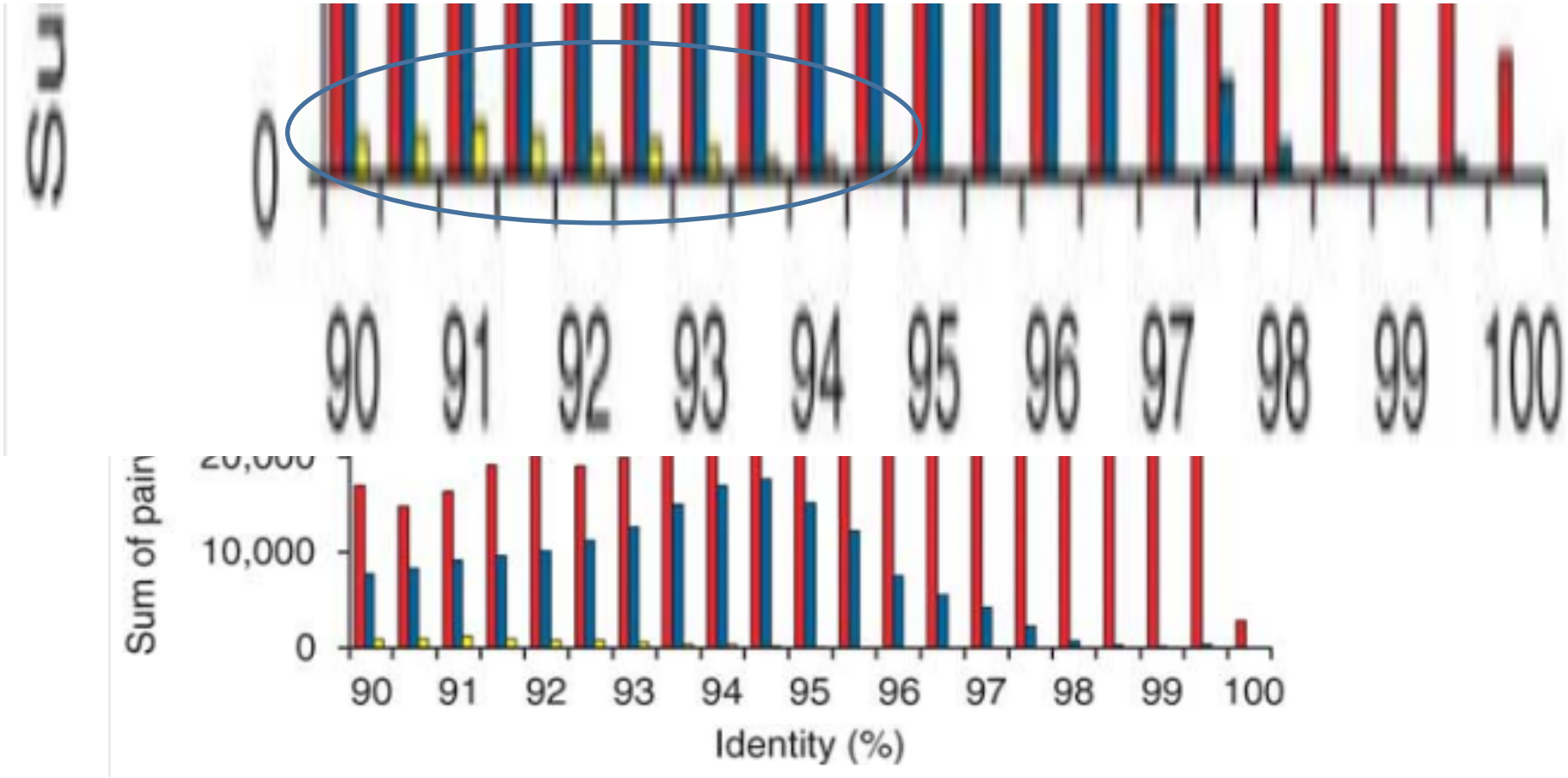
Proportion of 11K mammalian genes

Quality of assemblies (II)



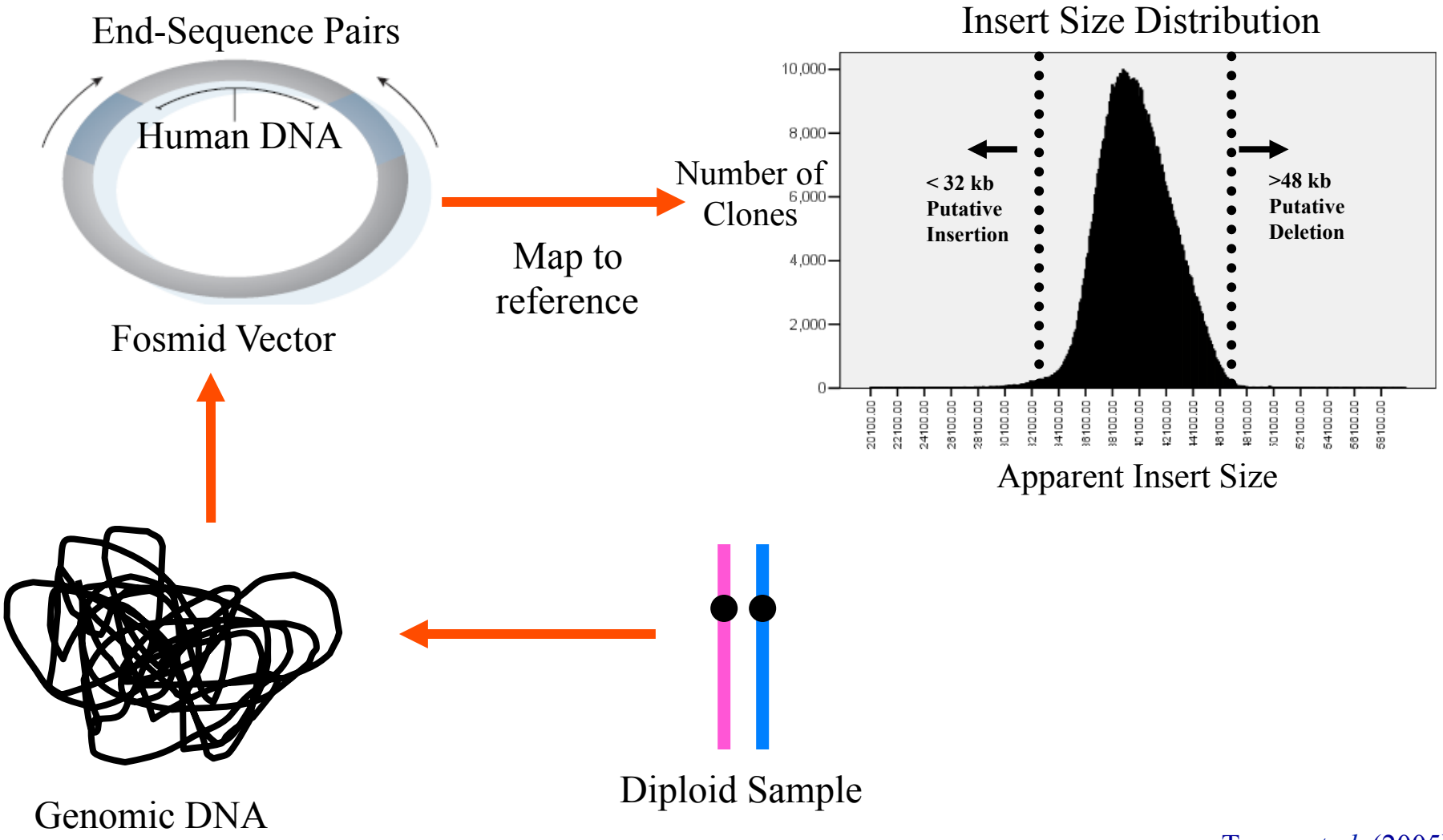
Incomplete respresentation of human genes

Limitations of NGS assemblies



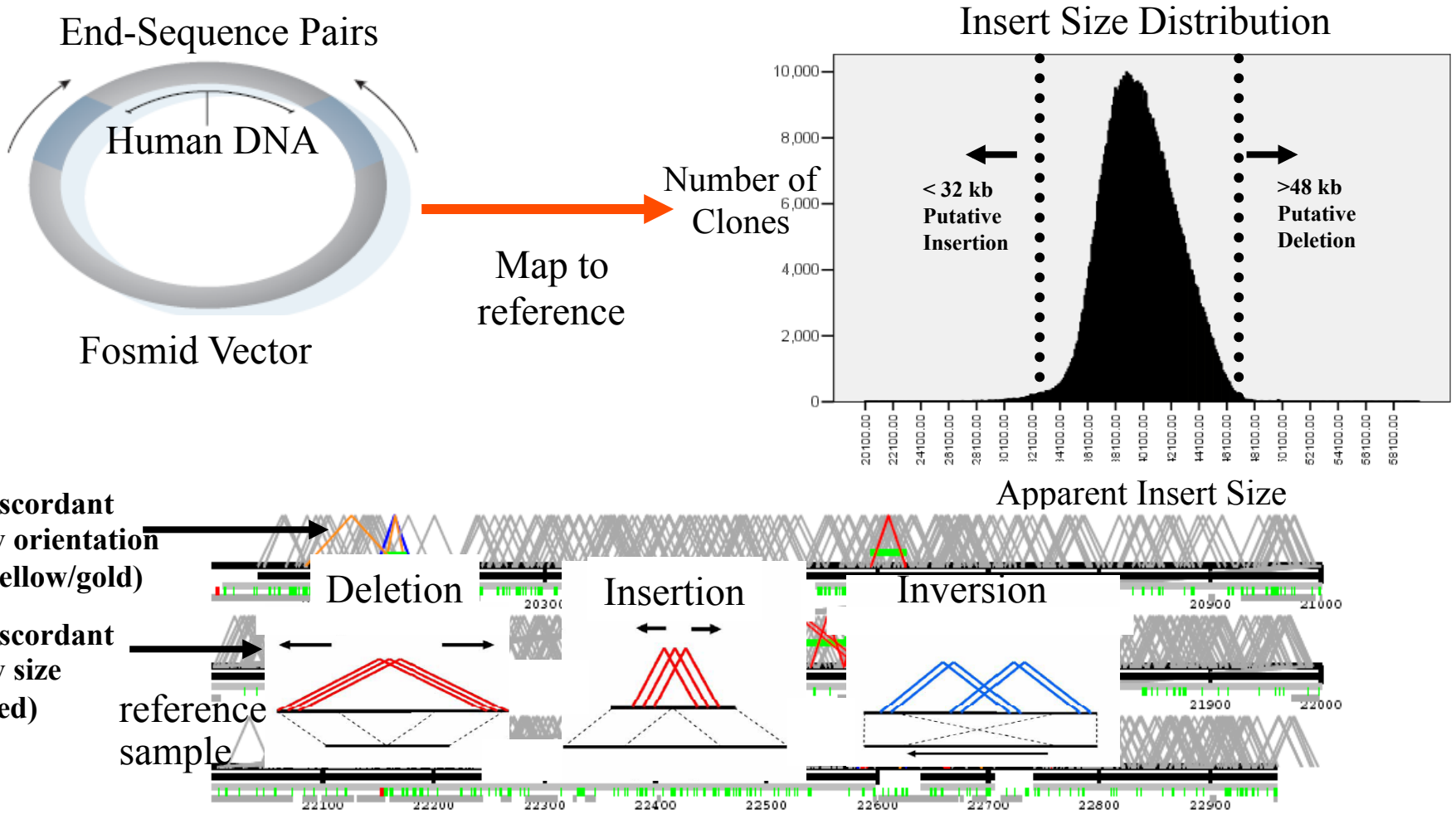
Alkan et al. Nature Methods 2010

Method 2: End-Sequence Pair (ESP) Analysis



Tuzun *et al.* (2005)

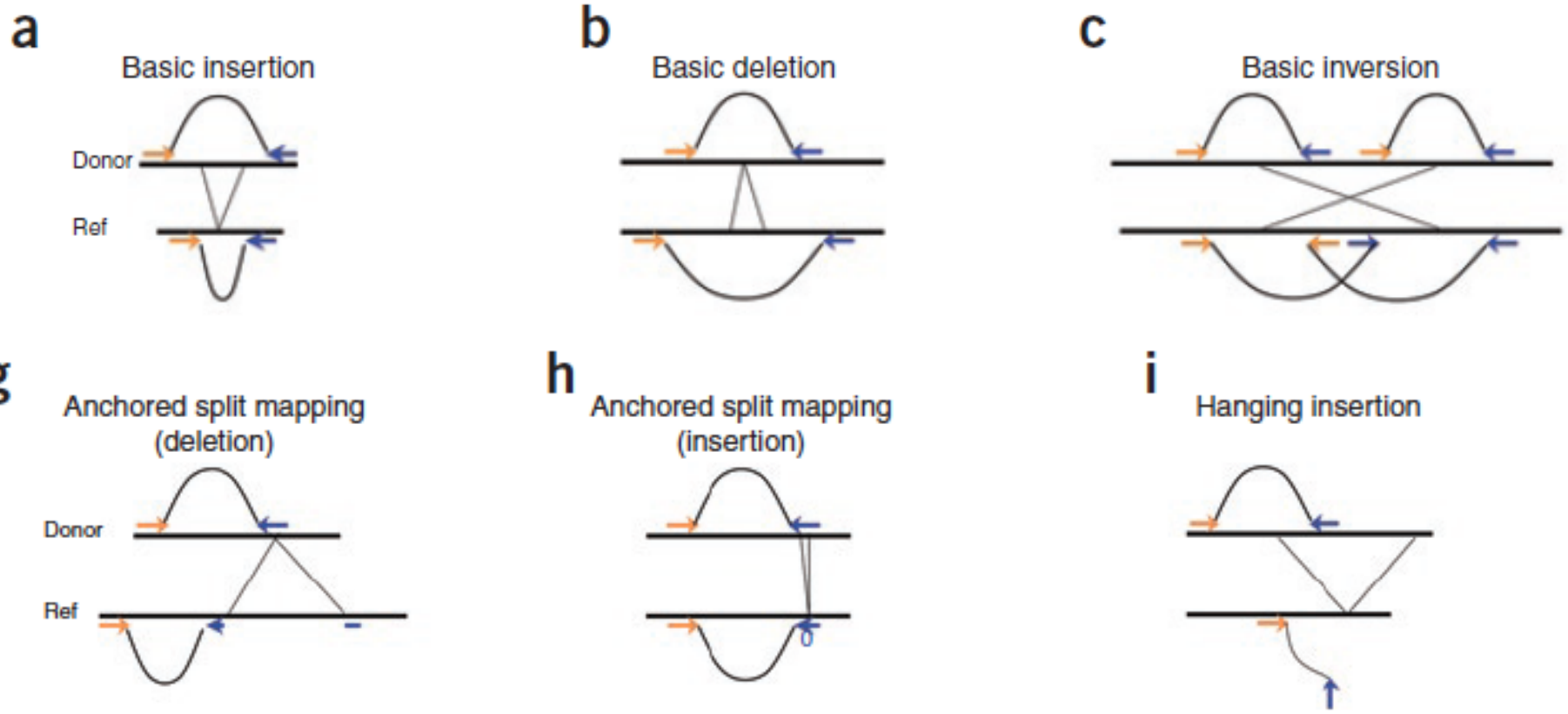
Method 2: End-Sequence Pair (ESP) Analysis



Tuzun *et al.* (2005)

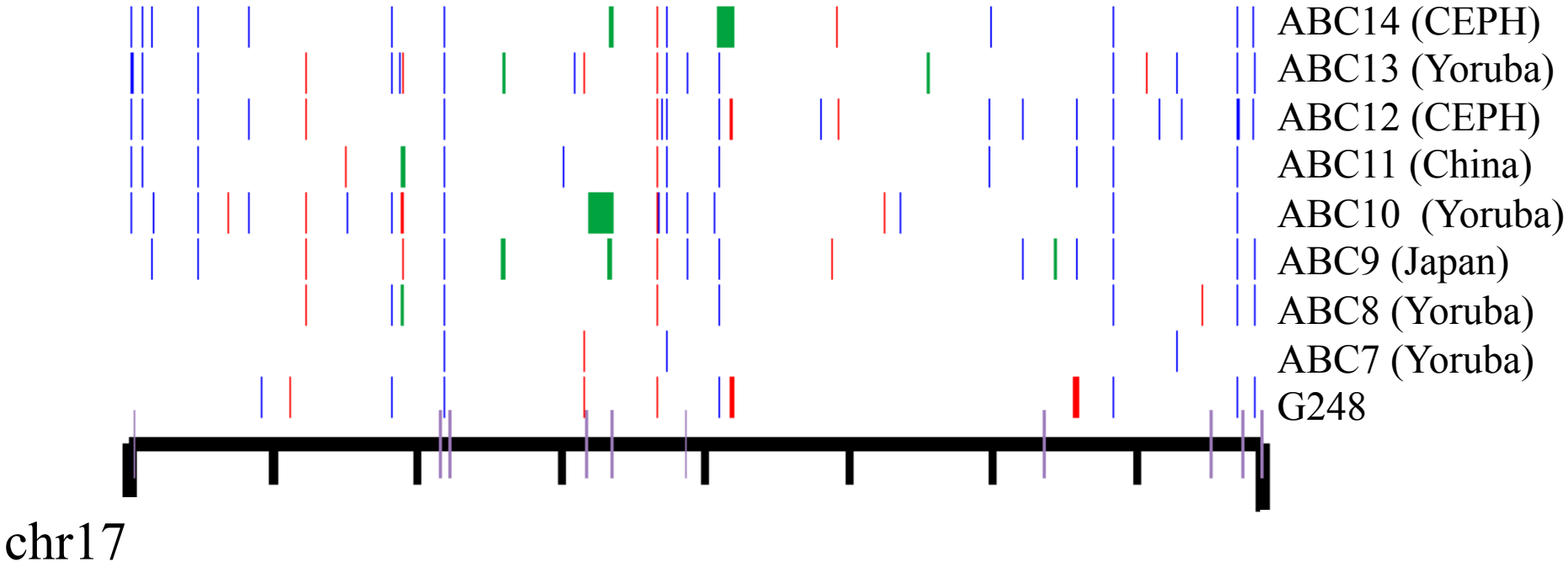
What can we find?

Structural variation detection:

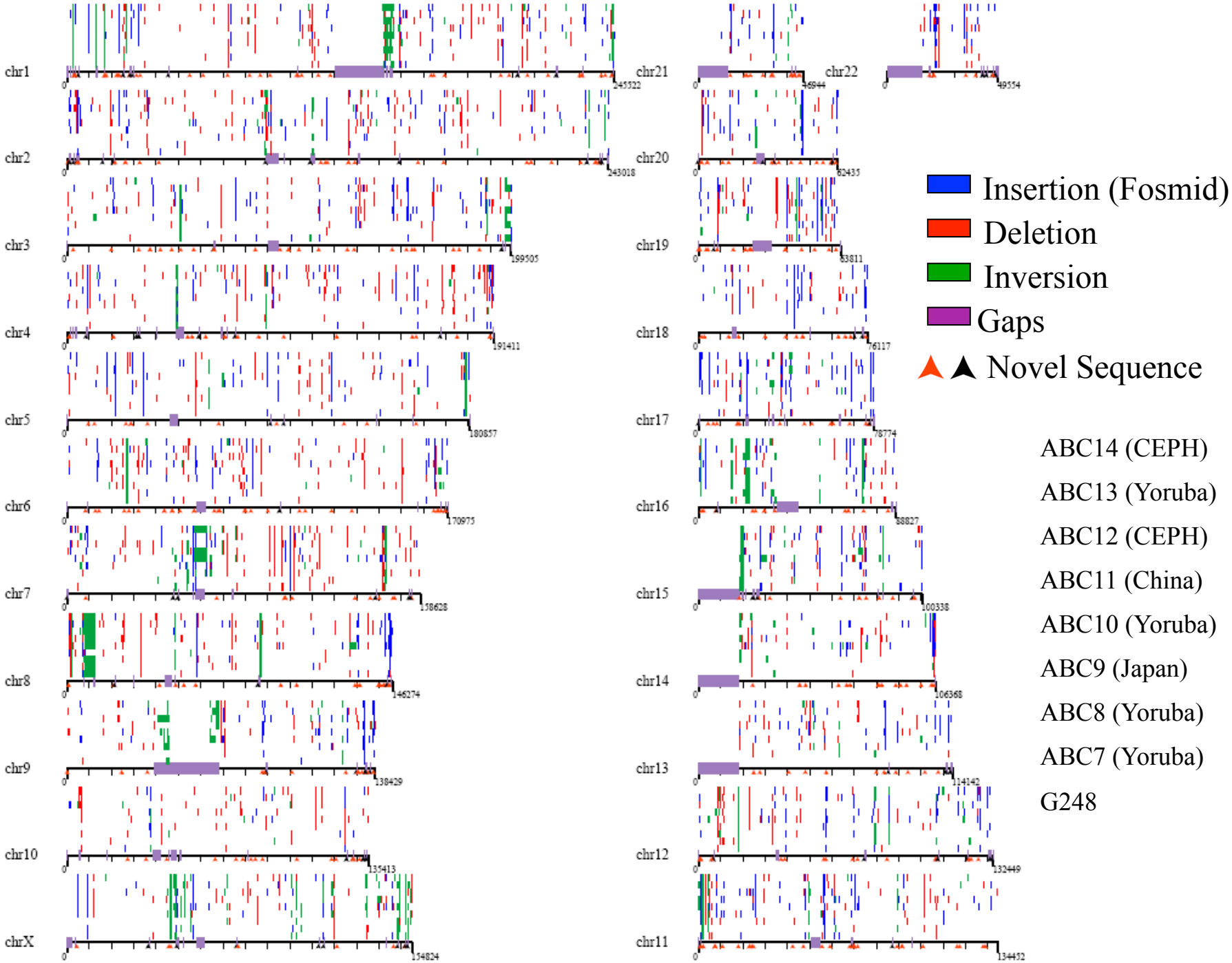


Alkan et al. Nature Review Genetics 2011

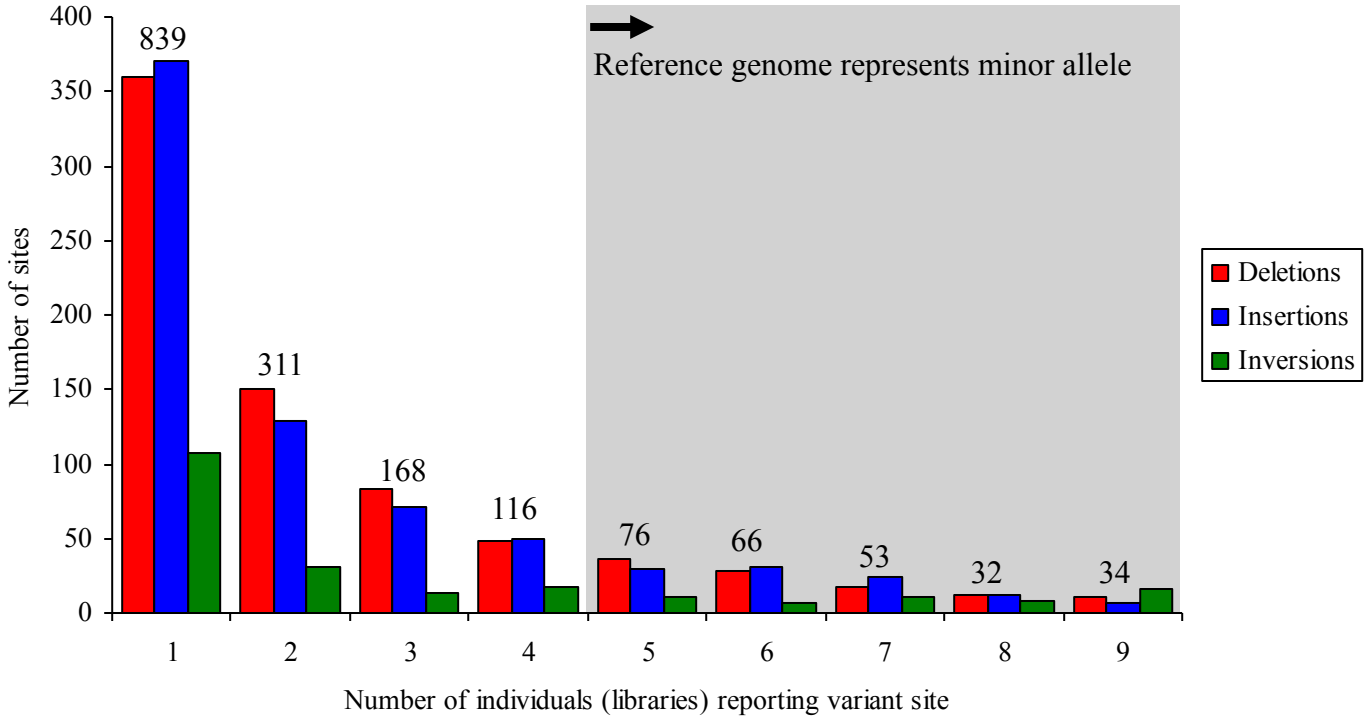
Map of Validated Variants



- **Genome wide map of variants**
- **Ability to resolve structure of individual haplotypes**

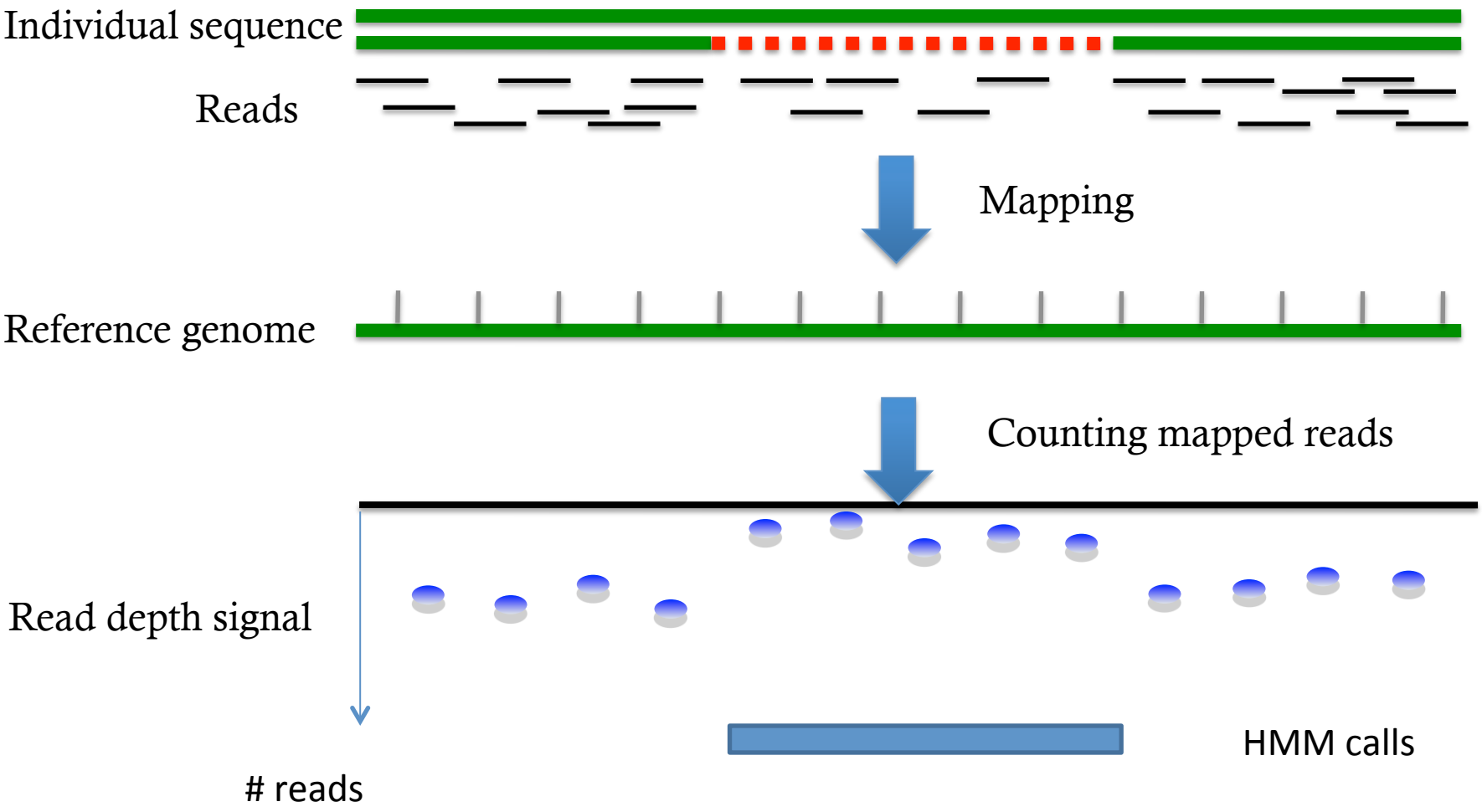

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chrX
chr21
chr22
chr20
chr19
chr18
chr17
chr16
chr15
chr14
chr13
chr12
chr11
Insertion (Fosmid)
Deletion
Inversion
Gaps
Novel Sequence
ABC14 (CEPH)
ABC13 (Yoruba)
ABC12 (CEPH)
ABC11 (China)
ABC10 (Yoruba)
ABC9 (Japan)
ABC8 (Yoruba)
ABC7 (Yoruba)
G248

Frequency of Validated Sites

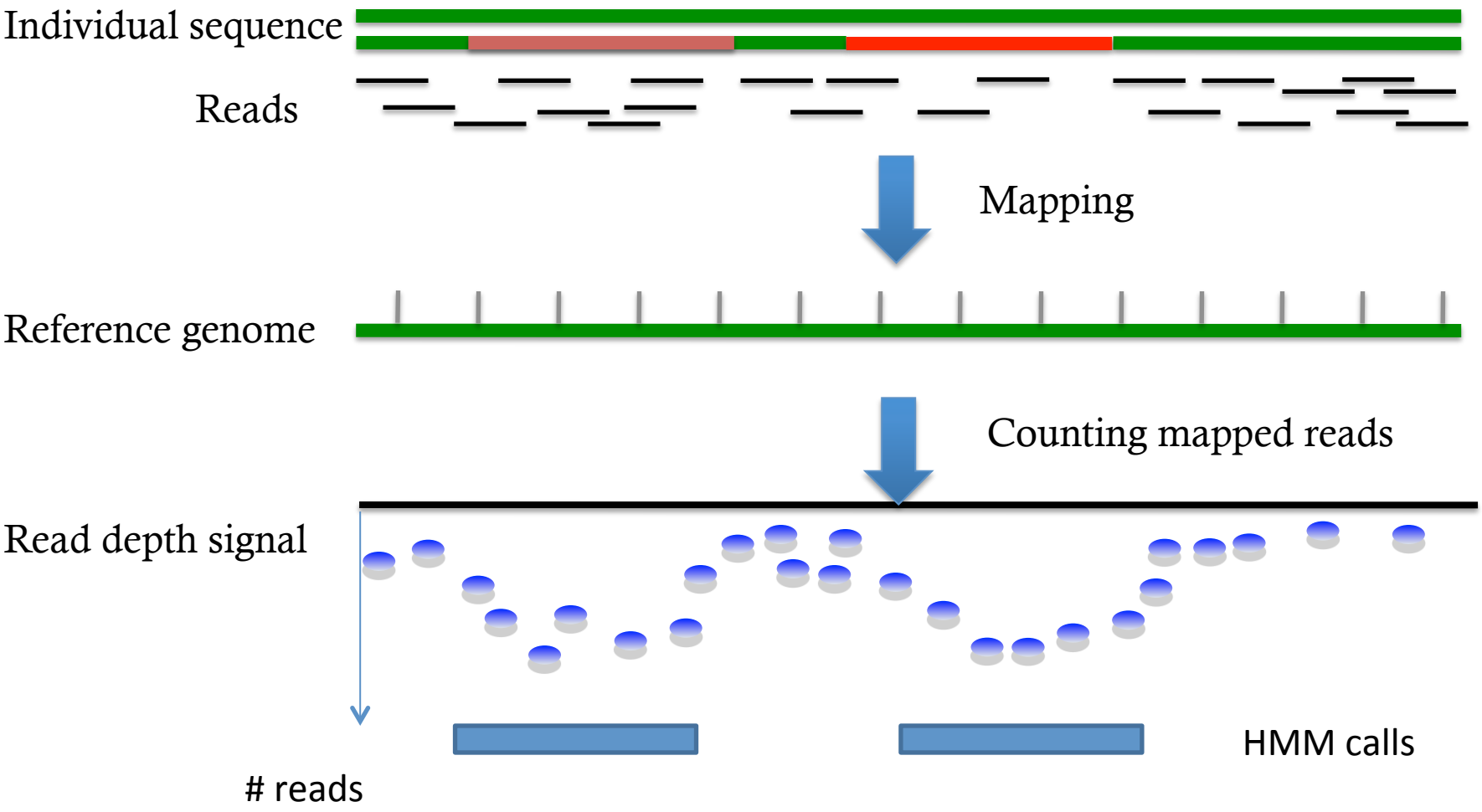


261 (15%) sites where reference genome represents a minor allele

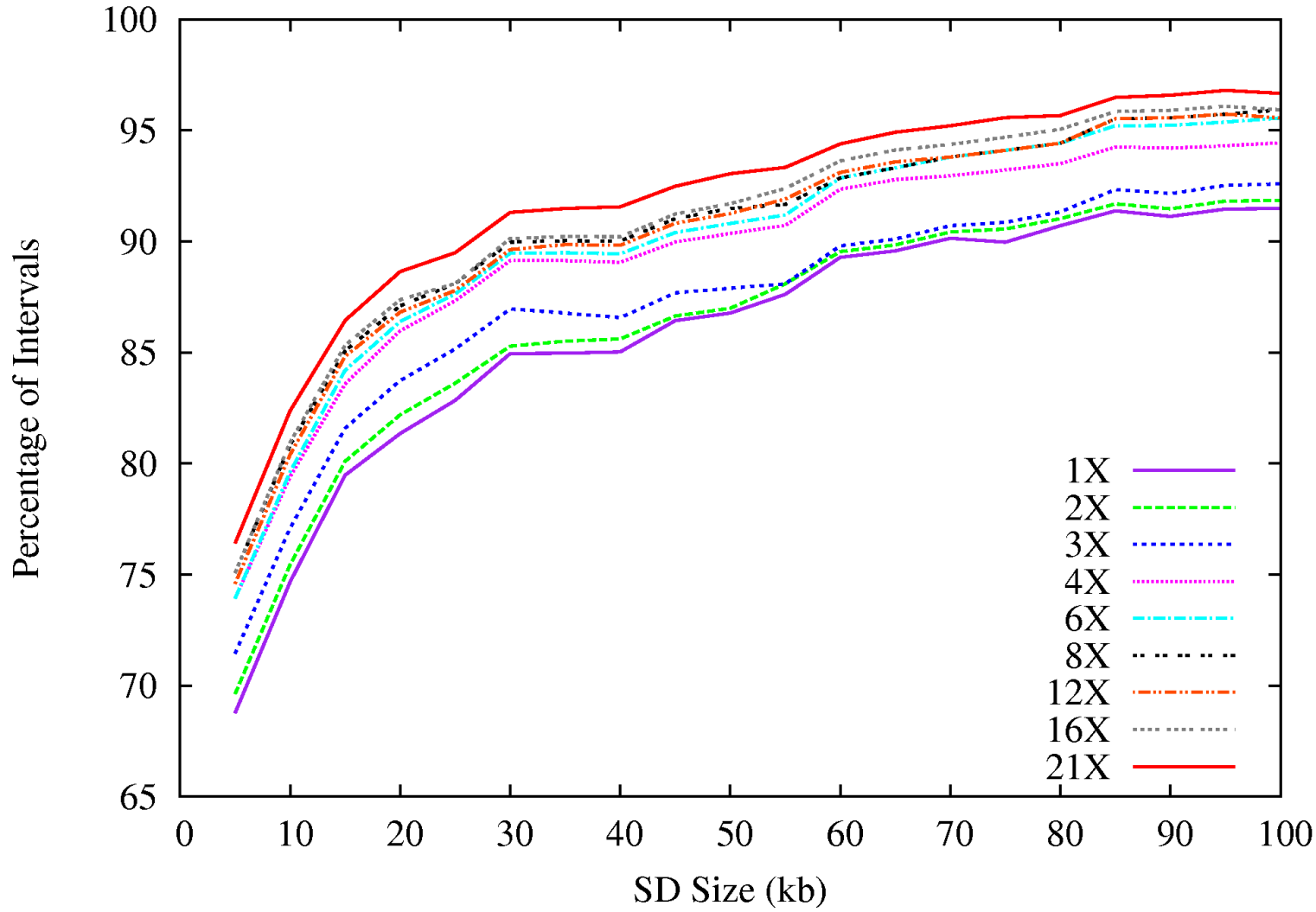
Method 3: Sequence Read Depth Analysis



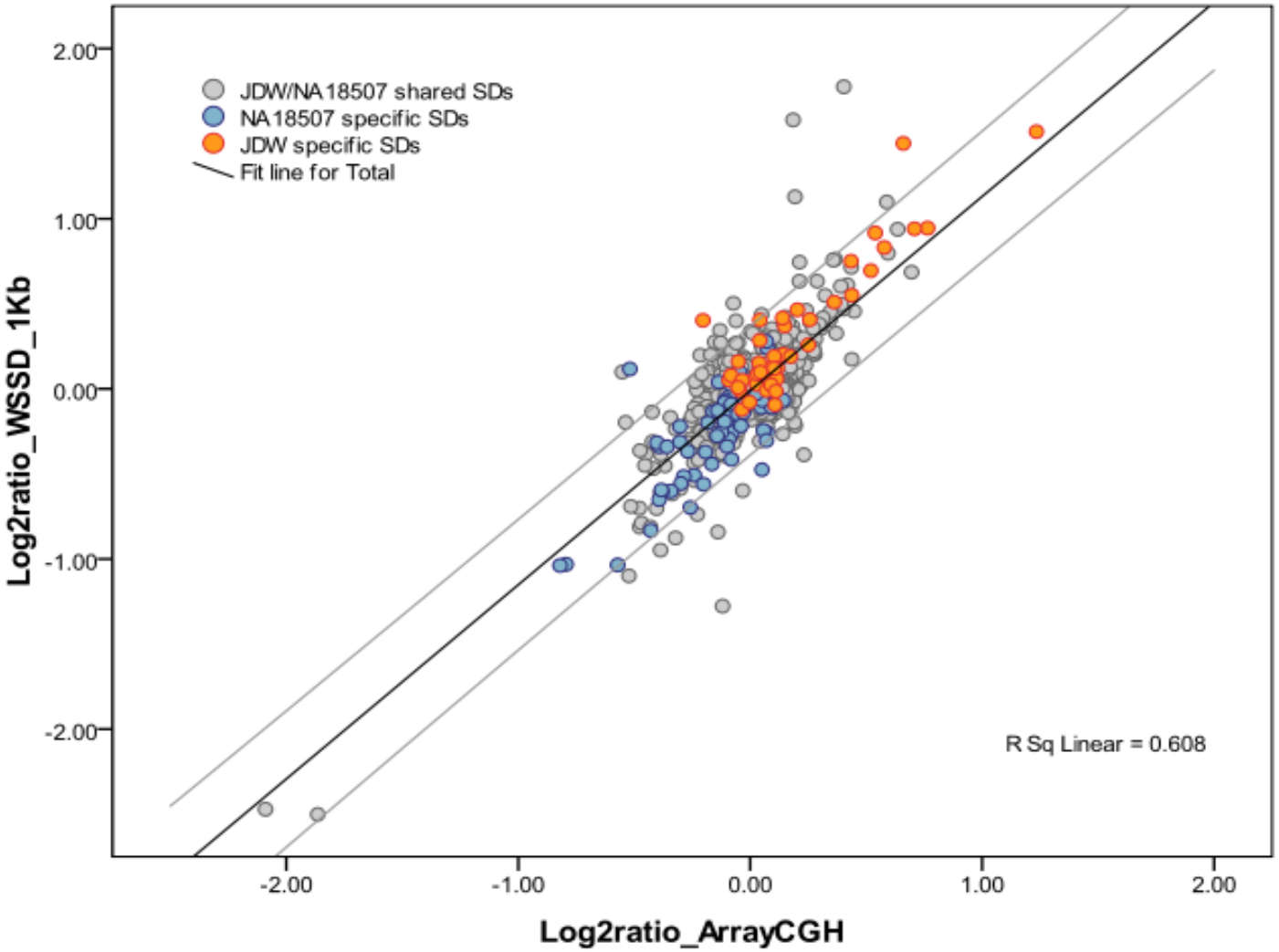
Method 3: Sequence Read Depth Analysis



Sequence coverage and detection power

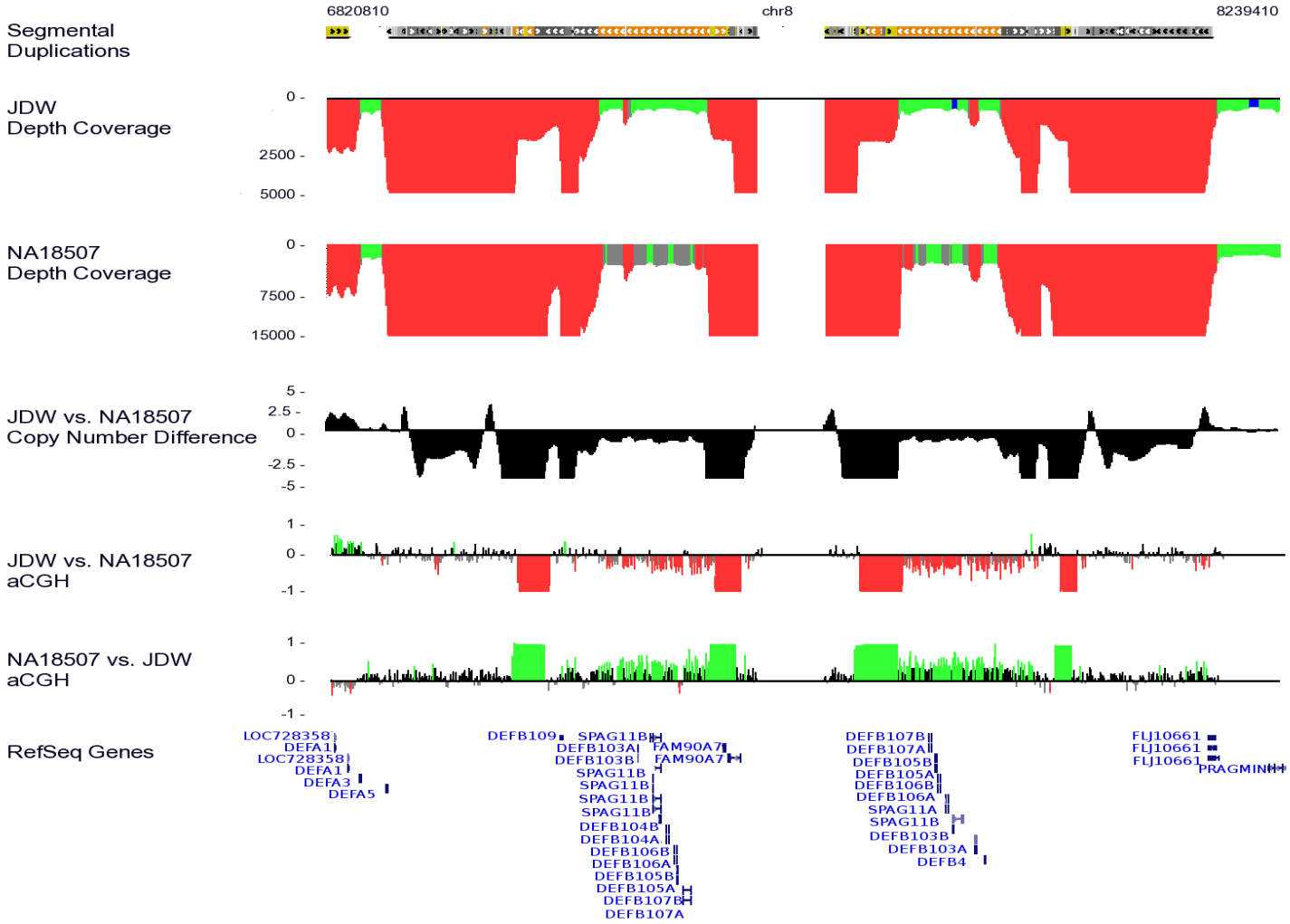


Validation of copy-number estimations



Alkan et al., Nature Genetics, 2009

Defensin gene cluster + *FAM90A7*

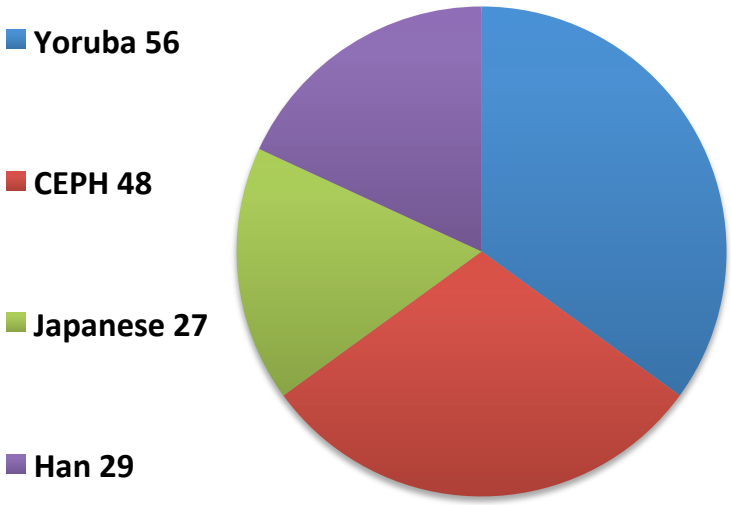


Associated with psoriasis and Crohn's disease

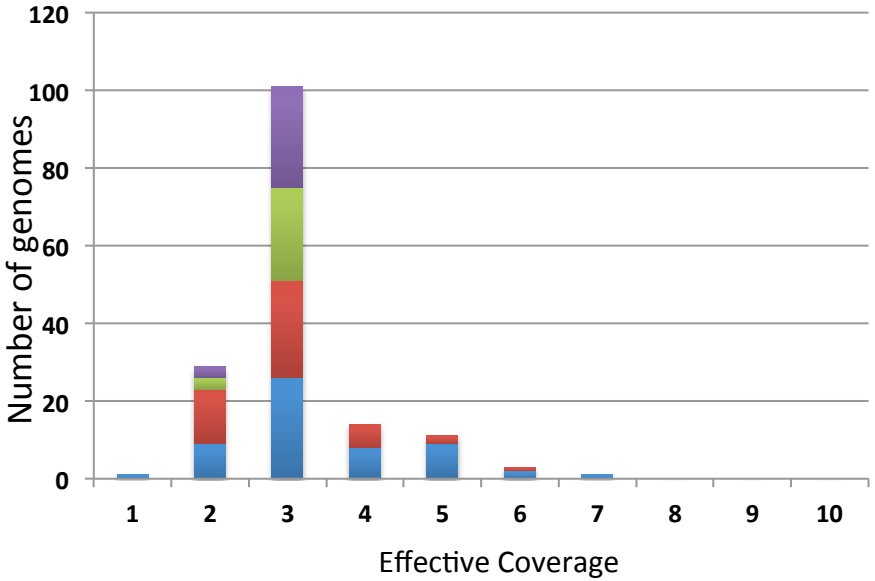
Alkan et al., Nature Genetics, 2009

Scaling up: 1000 Genomes and more

Individuals sequenced in Pilot 1



Histogram of Pilot 1 Illumina effective coverage



Individuals sequenced in Pilot 2

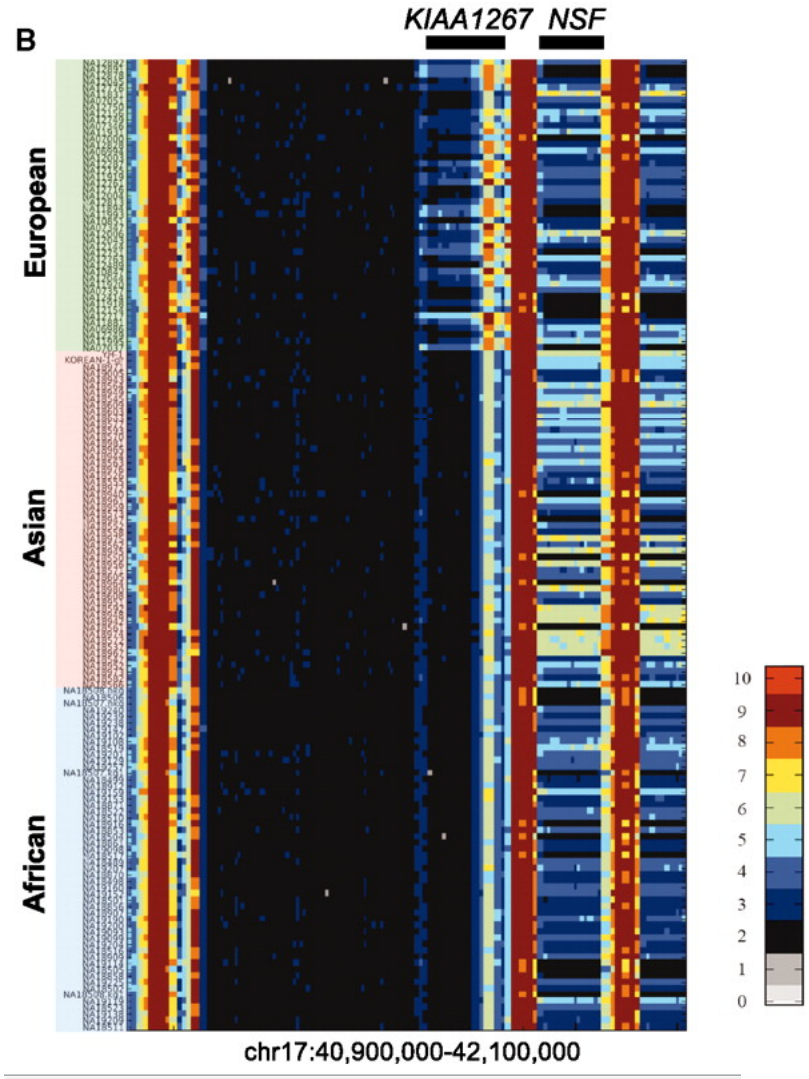
ID	Effective Coverage	Population
NA19240	24	YORUBA
NA19239	19	YORUBA
NA19238	13	YORUBA
NA12891	21	CEPH
NA12892	18	CEPH
NA12878	22	CEPH

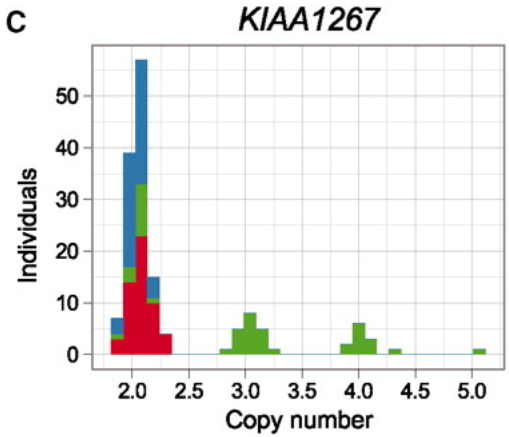
Other Genomes

ID	Effective Coverage	Population
YH-1	22	HAN CHINESE°
NA18507	29	YORUBA ⌘
NA18506	30	YORUBA *
NA18508	25	YORUBA *
KOREAN	12	KOREAN ❖

Sudmant, Kitzman, et al., Science, 2010

Copy number variation in human populations



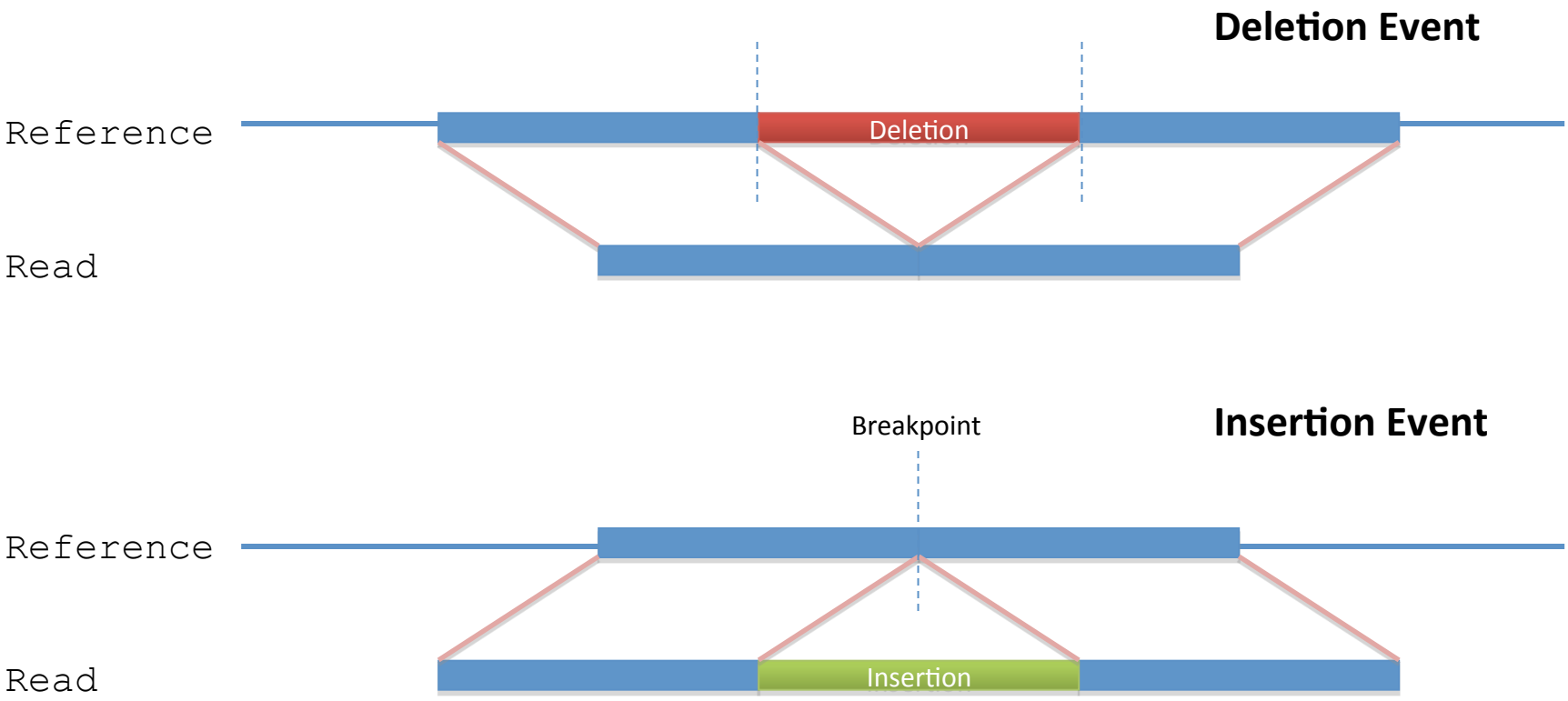


Sudmant et al. Science 2010

Deletions

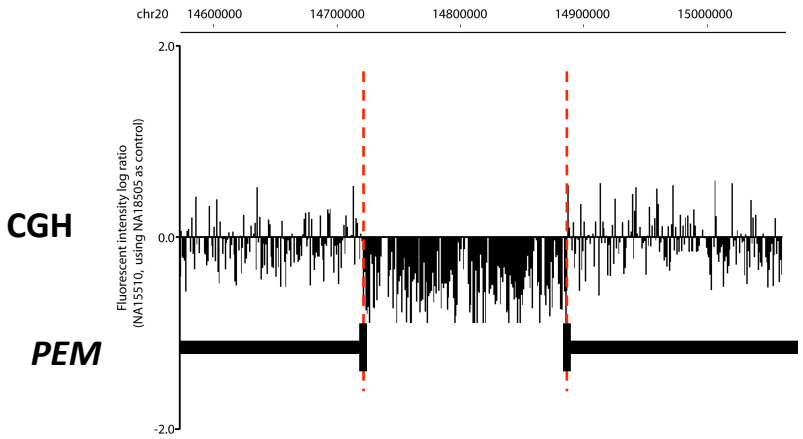

chr13:38,390,001-38,407,895
p12 p11.2 q11 q12.12 q12.3 q13.3 q14.11 q14.2 q21.1 q21.31 q21.33 q22.2 q31.1 q31.2 q32.1 q33.1 q33.3 q34
17 kb
38,392 kb
38,394 kb
38,396 kb
38,398 kb
38,400 kb
38,402 kb
38,404 kb
38,406 kb
NAME
DATA FILE
DATA TYPE
floquet_gem_no_split.bam Coverage
floquet_gem_no_split.bam
RefSeq Genes
chr13_hg19_repeatmasked.bed
WGAC_WSSD.bed
annotated_repeats_SD_Sr
final_paired_end_clusteri
prova_annotated_VH.bed
TRPC4
LINE
Simple_repeat
14
6

Split-read Analysis

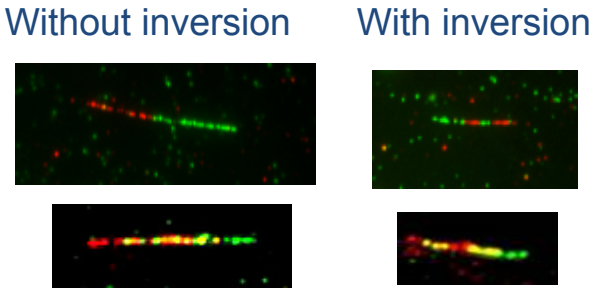


Experimental Validation

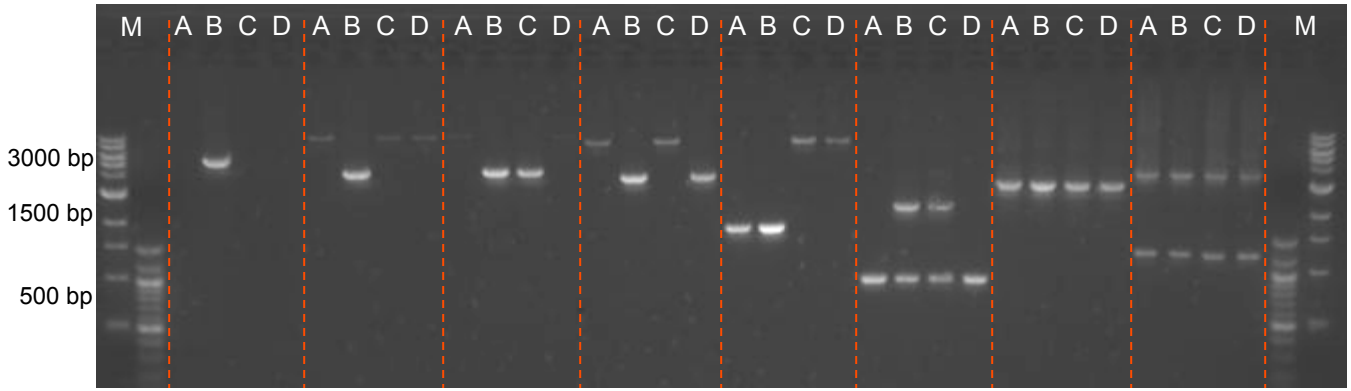
A) CGH



B) Fiber-FISH (For inversions)



C) PCR



Methods to Find SVs

Experimental approach

ArrayCGH (SNP based and genomic)

Based on ratios, Saturate quite fast, poor breakpoint resolution

Sequence based

Read pair analysis

- Deletions, small novel insertions, inversions, transposons
- Size and breakpoint resolution dependent to insert size

Read depth analysis

- Deletions and duplications
- Relatively poor breakpoint resolution

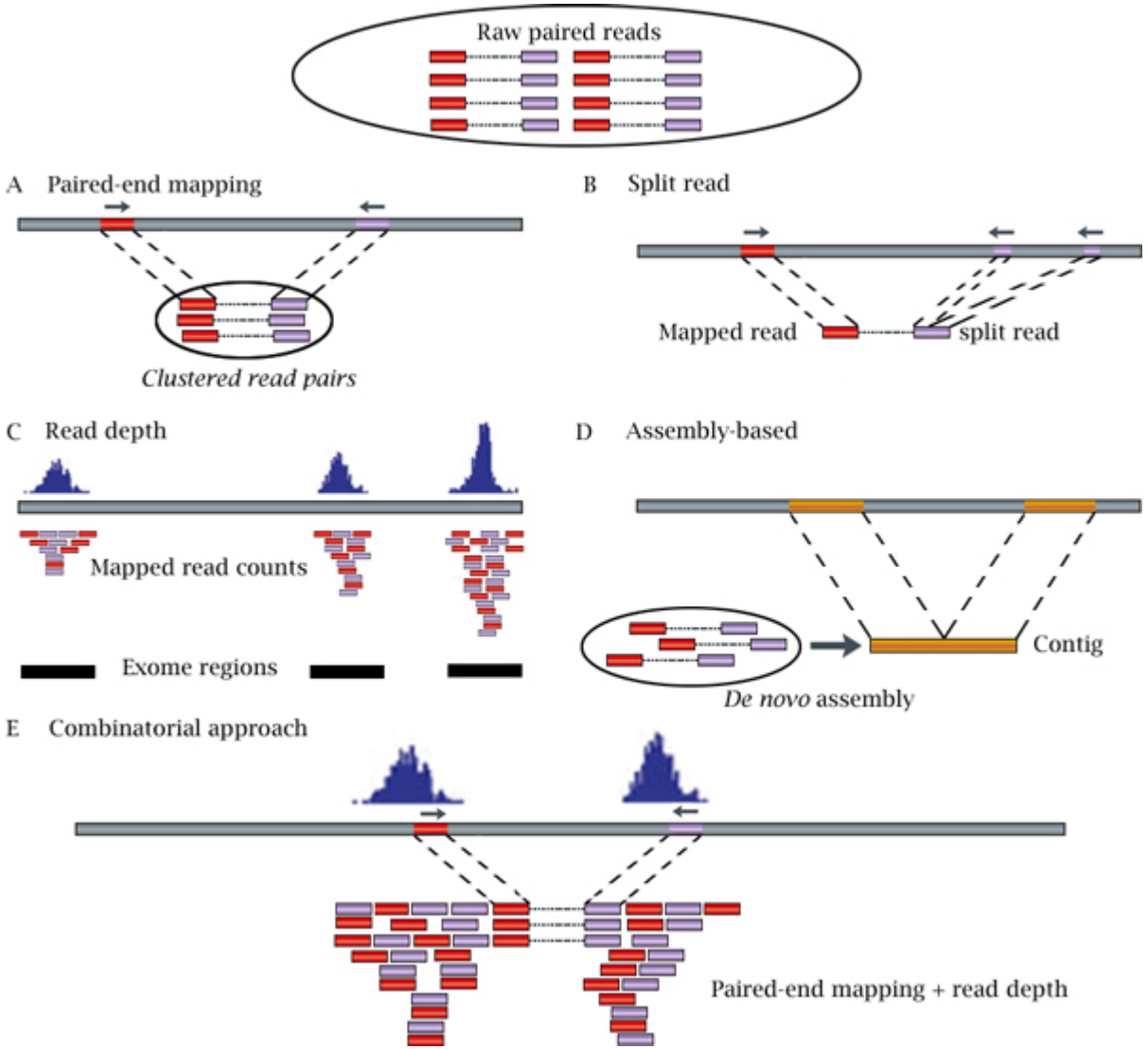
Split read analysis

- Small novel insertions/deletions, and mobile element insertions
- 1bp breakpoint resolution

Local and *de novo* assembly

- SV in unique segments
- 1bp breakpoint resolution

Review software



Software I

Method	Reference	Language	Control required?	Input format	GC correction	single-end/ pair-end	Methodology characteristics
CNV-seq	[15]	R, perl	Yes	hits	No	single-end	statistical testing
FREEC	[21]	C	Optional	SAM,BAM,bed,*etc.*	Optional	both	LASSO regression
readDepth	[22]	R	No	bed	Yes	both	CBS, LOESS regression
CNVnator	[23]	C	No	BAM	Yes	both	mean shift algorithm
SegSeq	[14]	Matlab	Yes	bed	No	single-end	statistical testing,CBS
EWT (RDXplorer)	[11]	R, python	No	BAM	Yes	single-end	statistical testing
cnD	[16]	D	No	SAM,BAM	No	both	HMM, Viterbi algorithm
CNVer	[17]	C	No	BAM	Yes	pair-end	maximum-likelihood, graphic flow
CopySeq	[18]	Java	No	BAM	Yes	pair-end	MAP estimator
rSW-seq	[19]	NA	Yes	NA	Yes	single-end	Smith-Waterman algorithm
CNAseg	[20]	R	Yes	BAM	No	pair-end	wavelet transform and HMM
CNAnorm	[24]	R	Yes	SAM,BAM	Yes	both	linear regression or CBS
cn.MOPS	[26]	R, C++	multiple samples	BAM or data matrix	No	both	mixture of Poissons, MAP, EM, CBS
JointSLM	[27]	R, Fortran	multiple samples	data matrix	Yes	both	HMM, ML estimator, Viterbi algorithm

doi:10.1371/journal.pone.0059128.t001

break point position estimation: readDepth = EWT>CNVnator>FREEC>CNV-seq>SegSeq;
copy number estimation: CNVnator>CNV-seq>readDepth>FREEC>EWT>SegSeq;

Zhao et al. BMC Bioinformatics 2013
Duan et al. Plos One 2013