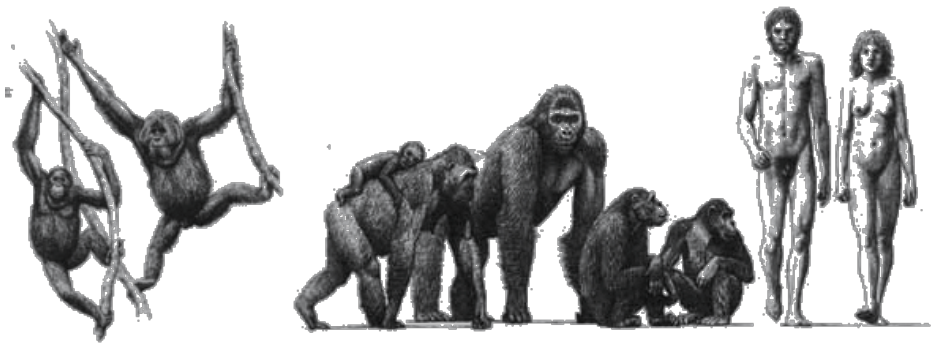
Structural variation

Programming for Biology

CSH, October 2017

Tomas Marques-Bonet
ICREA Research Professor
Institut de Biologia Evolutiva

Lukas Kuderna

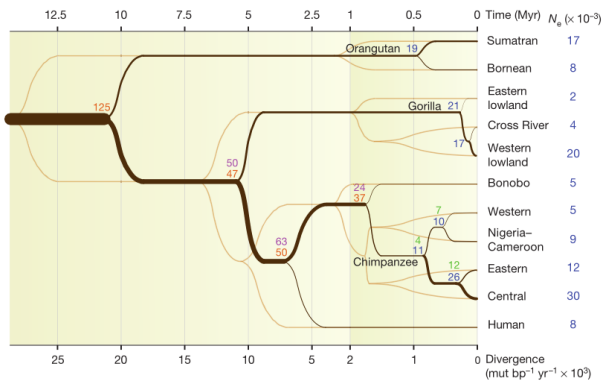


Who are we?

- Evolutionary genomics
  - Barcelona, Biomedical Research Park

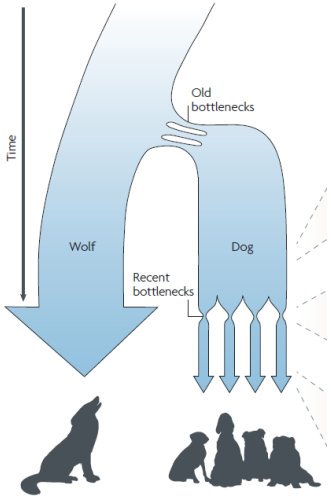
What do we do?

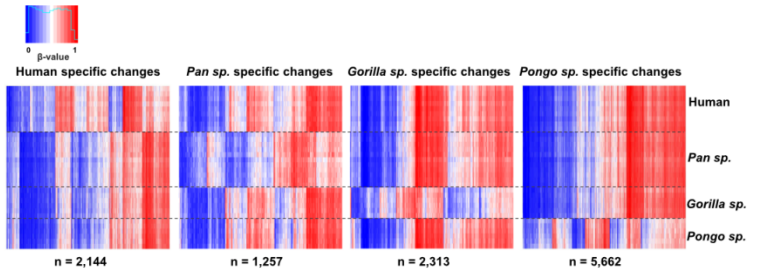
- Natural selection on human evolution



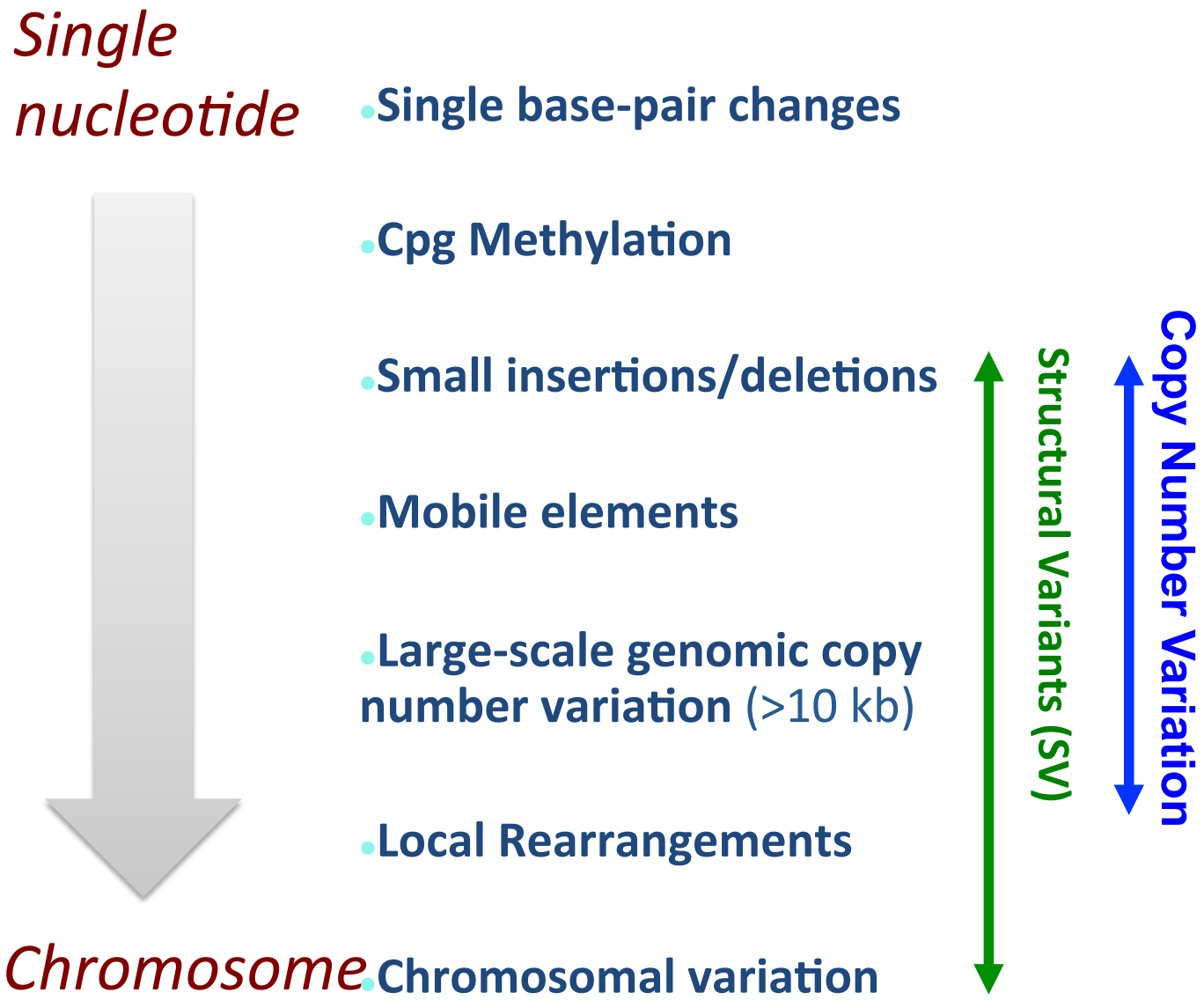
- Transcriptome and Epigenetics in Primates

- Canid evolution

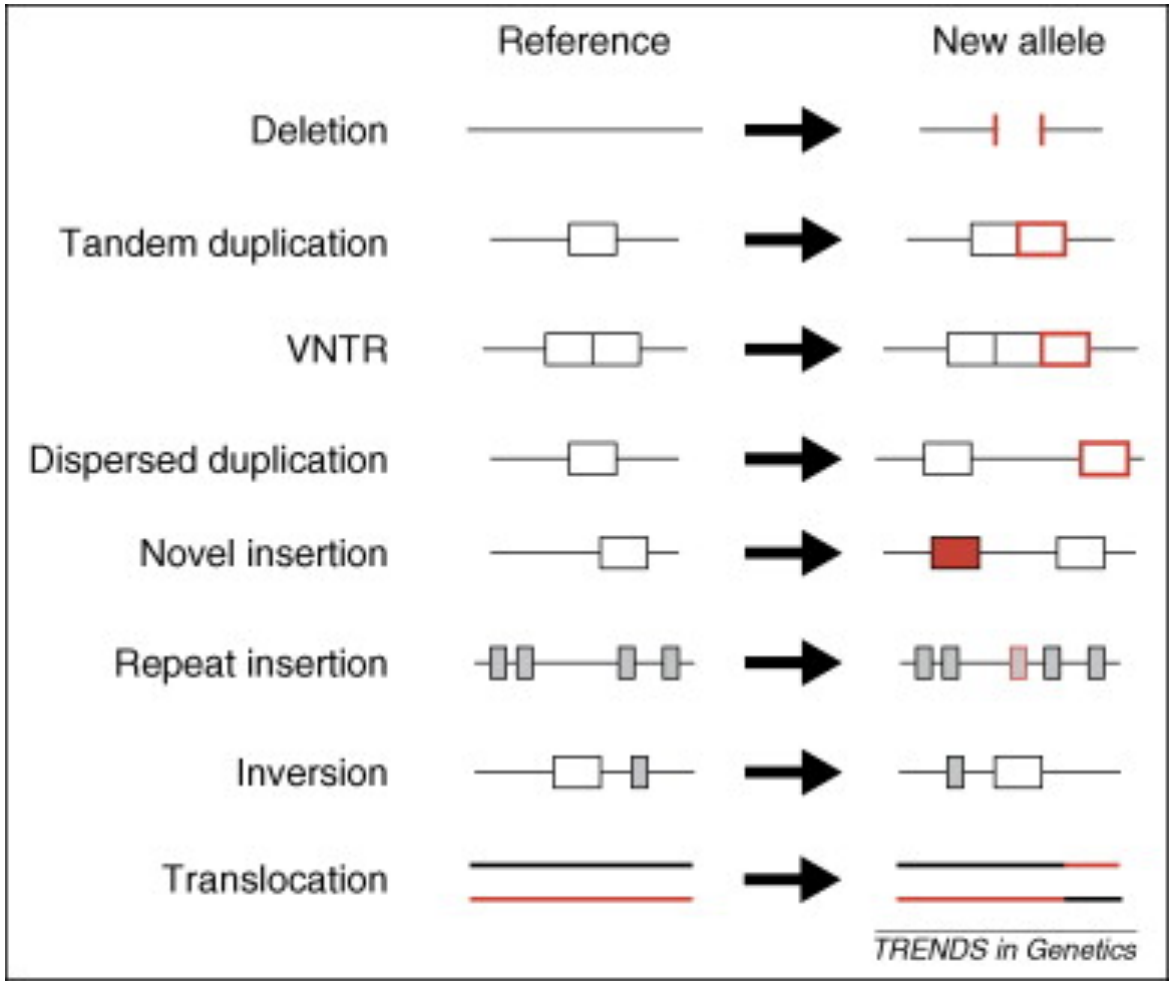




Continuum of Genomic Variation



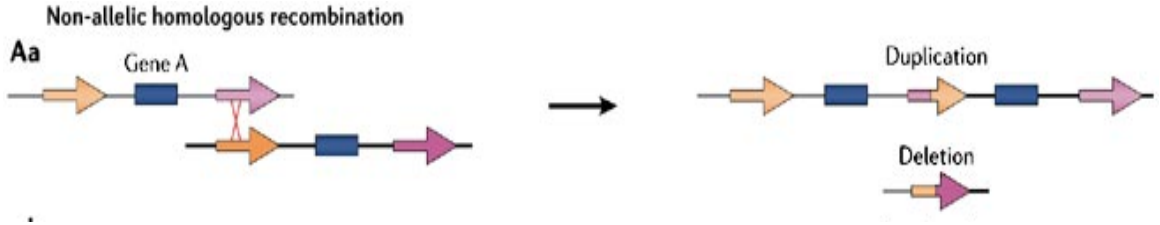
Types of Structural Variation



Hurles *et al.* 2008

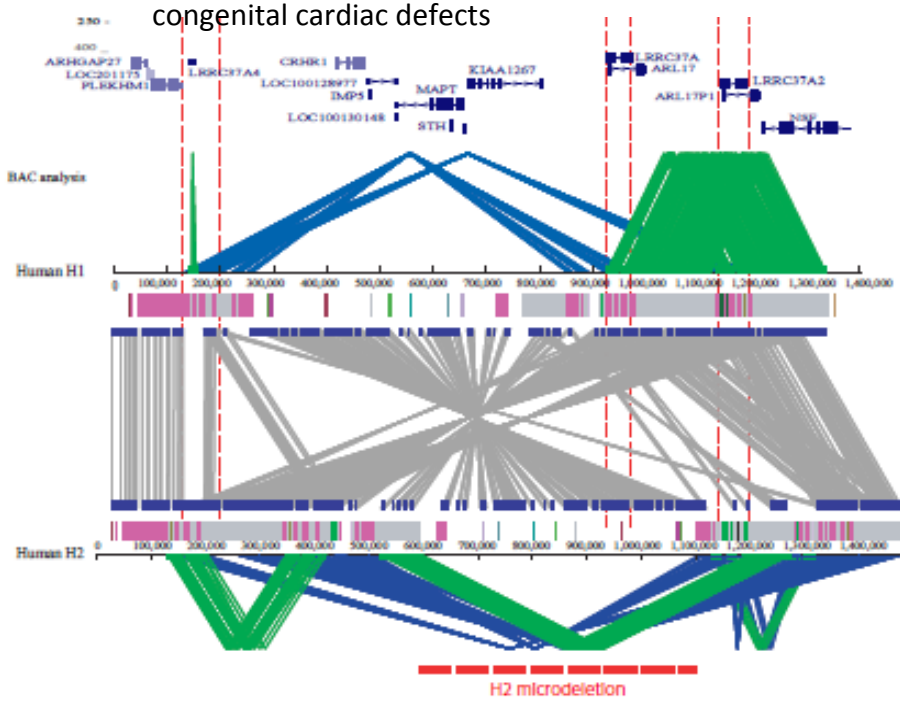
Why Study Structural Variation?

- 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

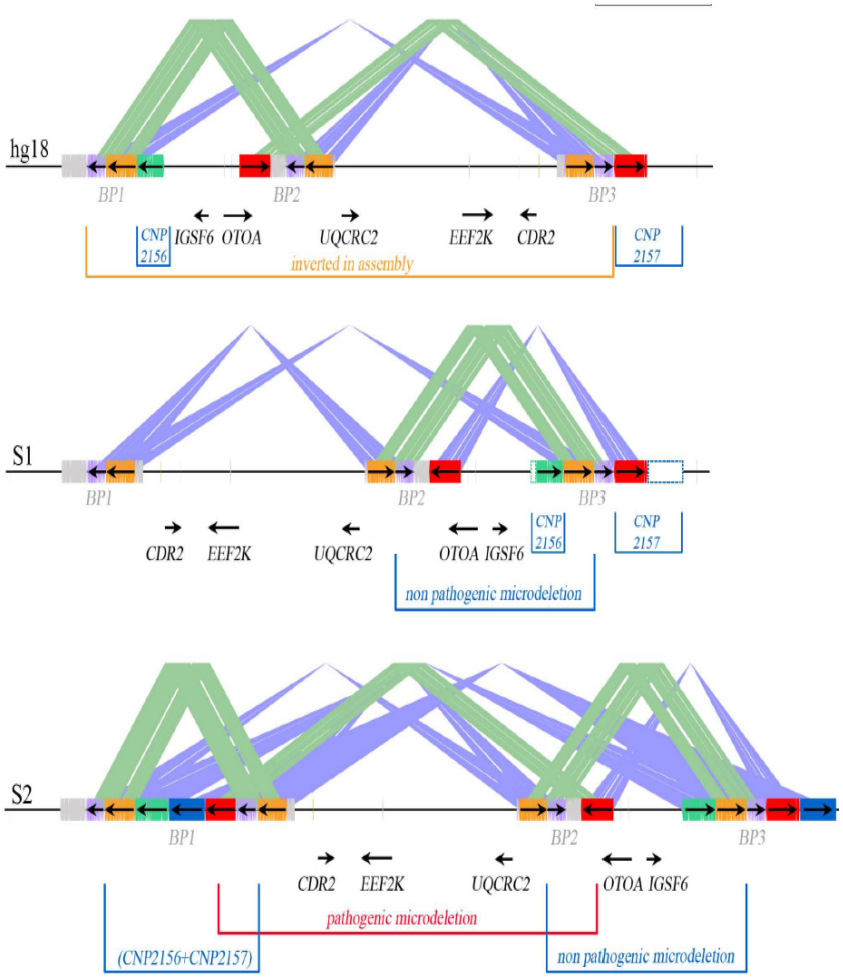
MR, global delay and
congenital cardiac defects



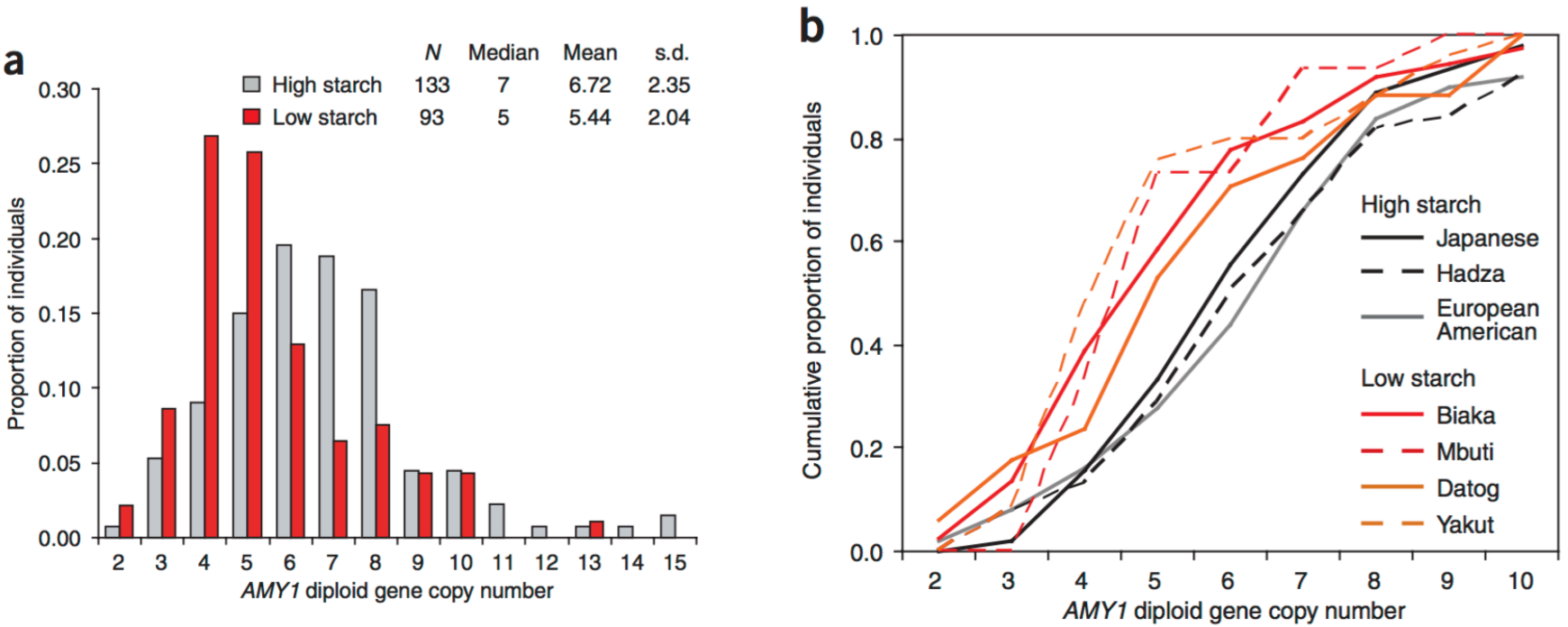
Zody *et al.* *Nature* Genetics (2008)

16p12.1 deletion syndrome

childhood intellectual disability, developmental delay



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


a
N Median Mean s.d.
High starch 133 7 6.72 2.35
Low starch 93 5 5.44 2.04
Proportion of individuals
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2 3 4 5 6 7 8 9 10 11 12 13 14 15
AMY1 diploid gene copy number
b
Cumulative proportion of individuals
1.0
0.8
0.6
0.4
0.2
0.0
2 3 4 5 6 7 8 9 10
AMY1 diploid gene copy number
High starch
Japanese
Hadza
European American
Low starch
Biaka
Mbuti
Datog
Yakut

Methods to Find SVs

Experimental approach

ArrayCGH (SNP based and genomic)

Sequence based

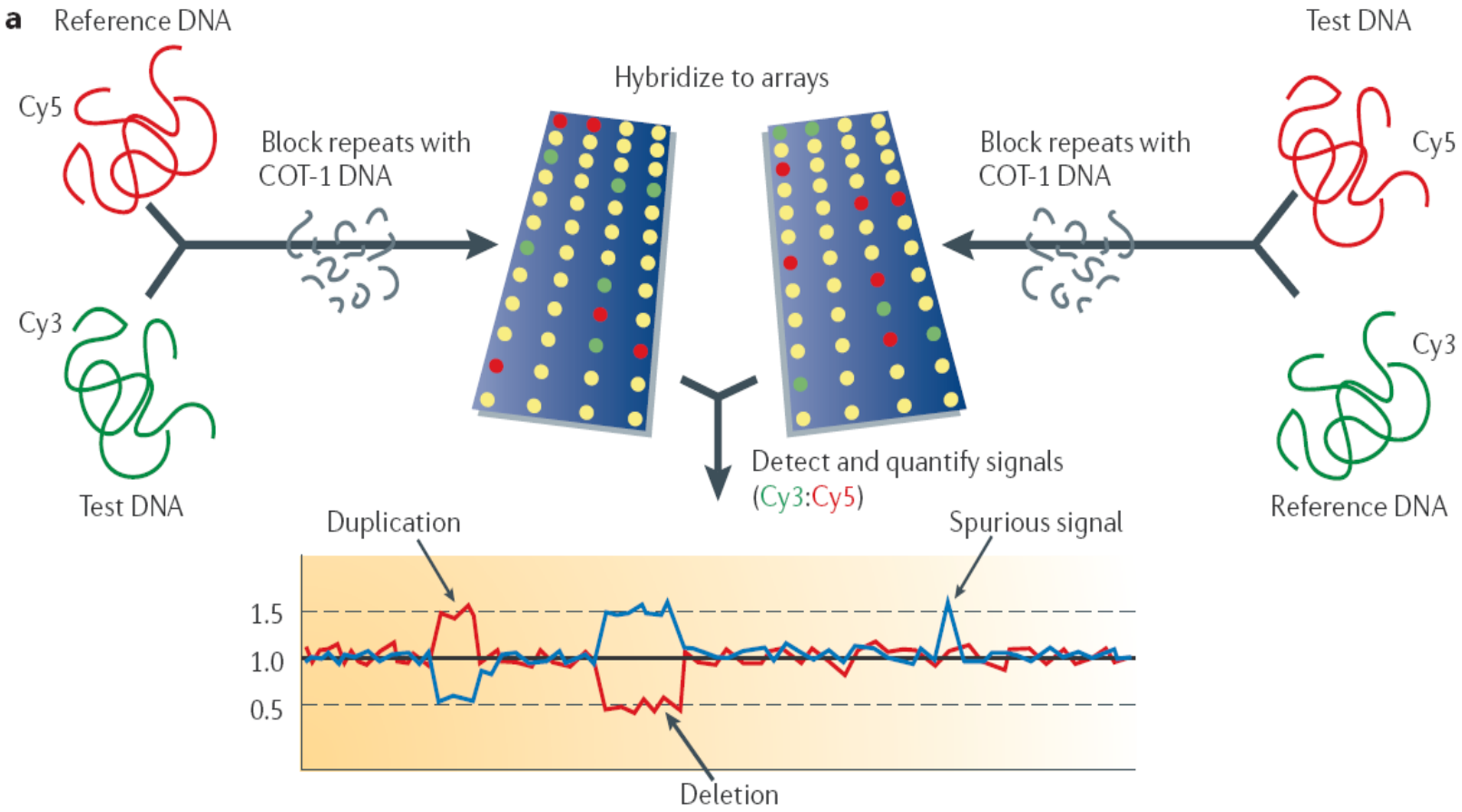
Local and *de novo* assembly

Read pair analysis

Read depth analysis

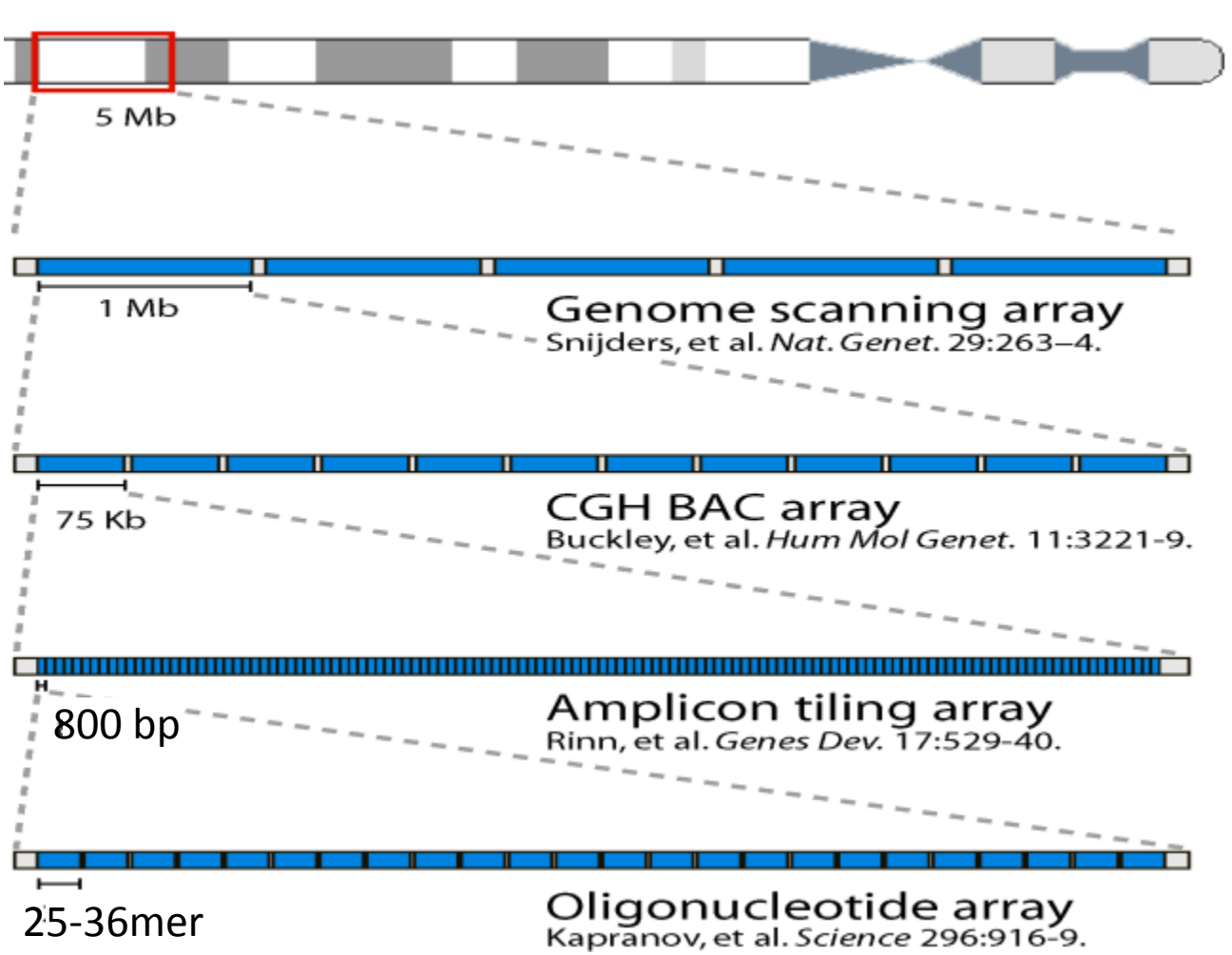
Split read analysis

METHOD 1: Copy Number Variation: Array Comparative Genomic Hybridization



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

Genome Tiling Arrays



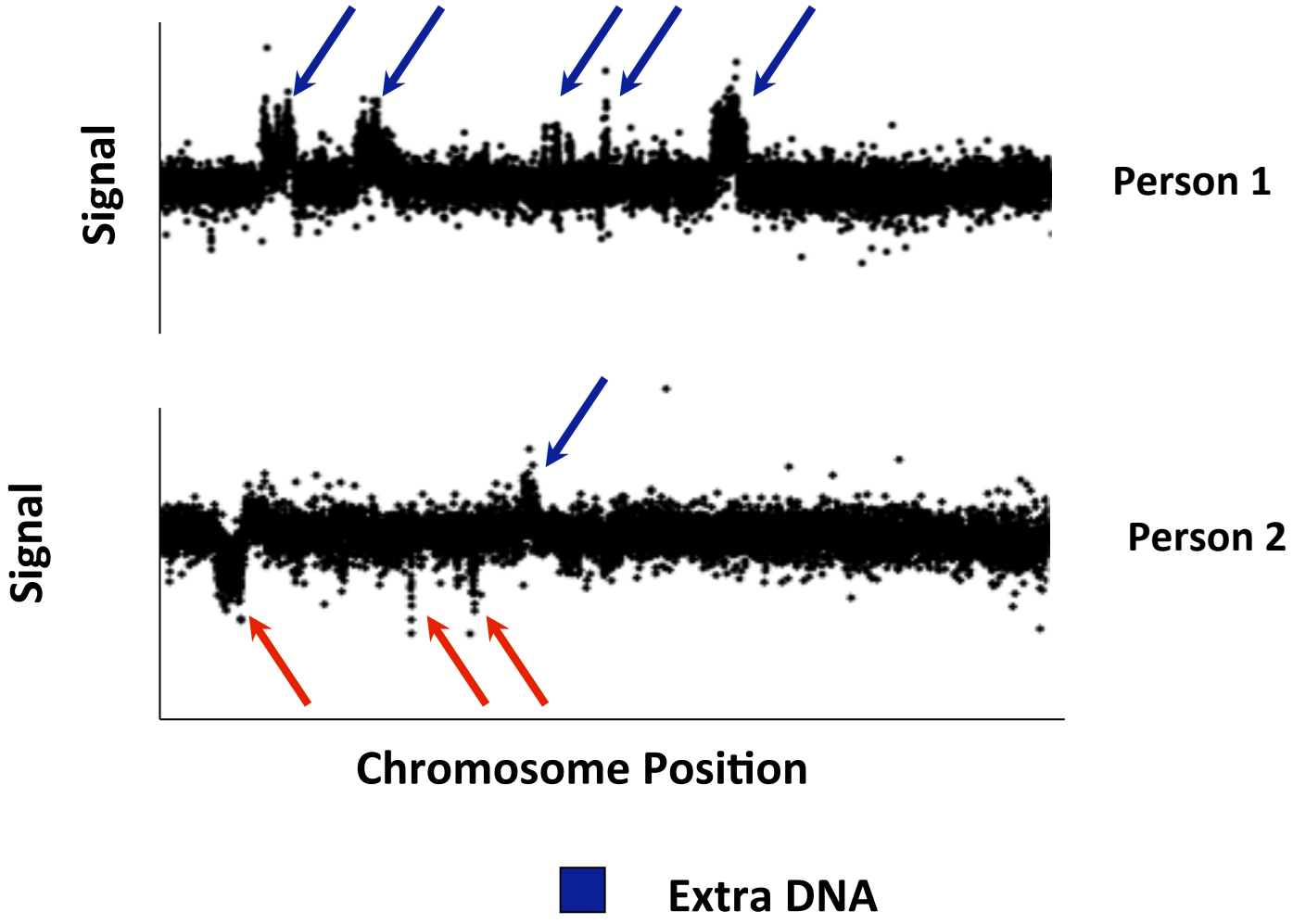
Typical Analysis Procedure

- 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:

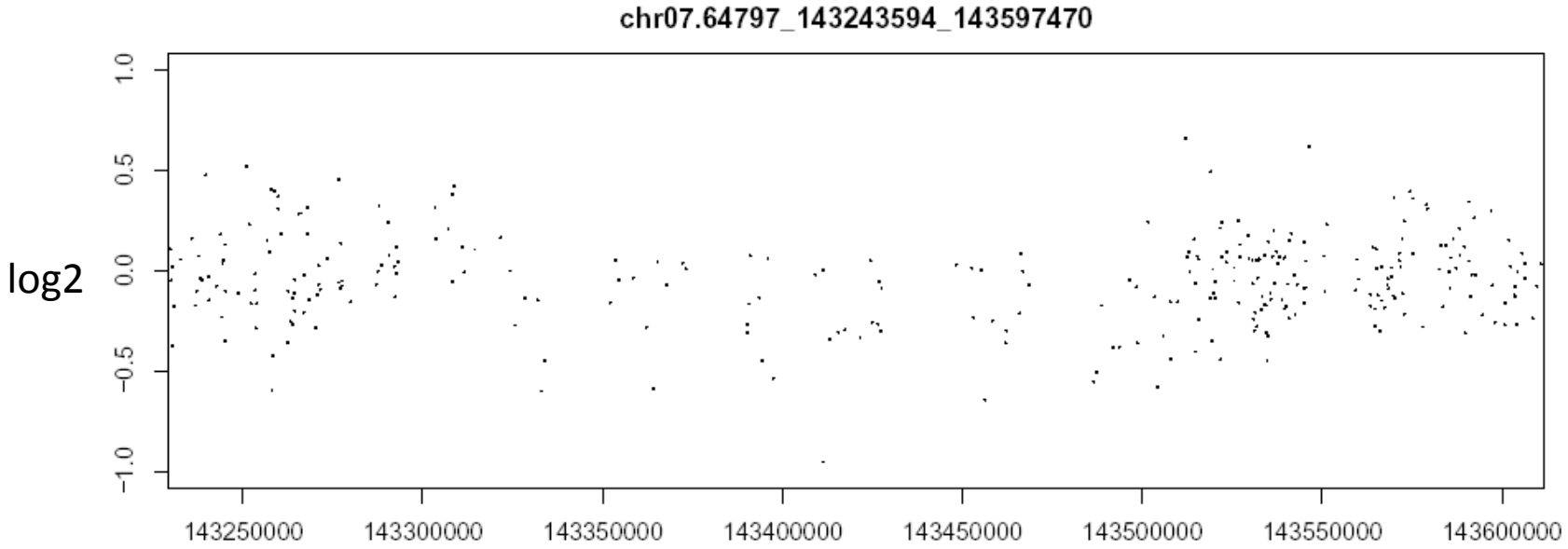
    log2(test/reference)=log2(3/2)=0.59
  - Homozygous duplication:

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

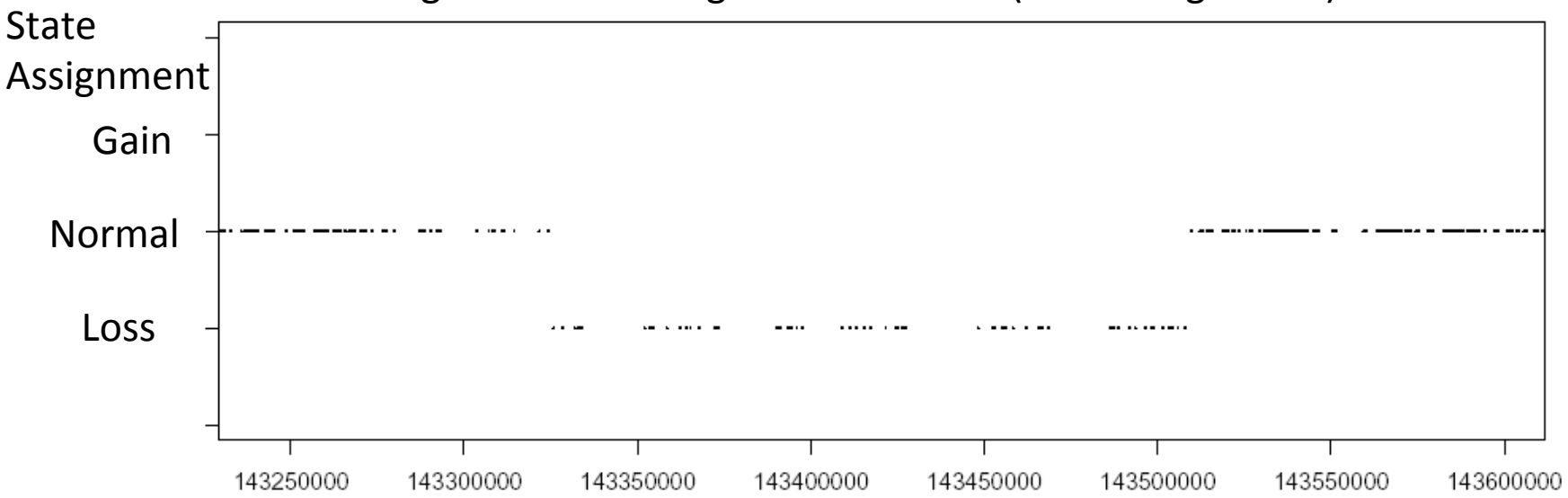
Copy Number Variations in the Human Genome



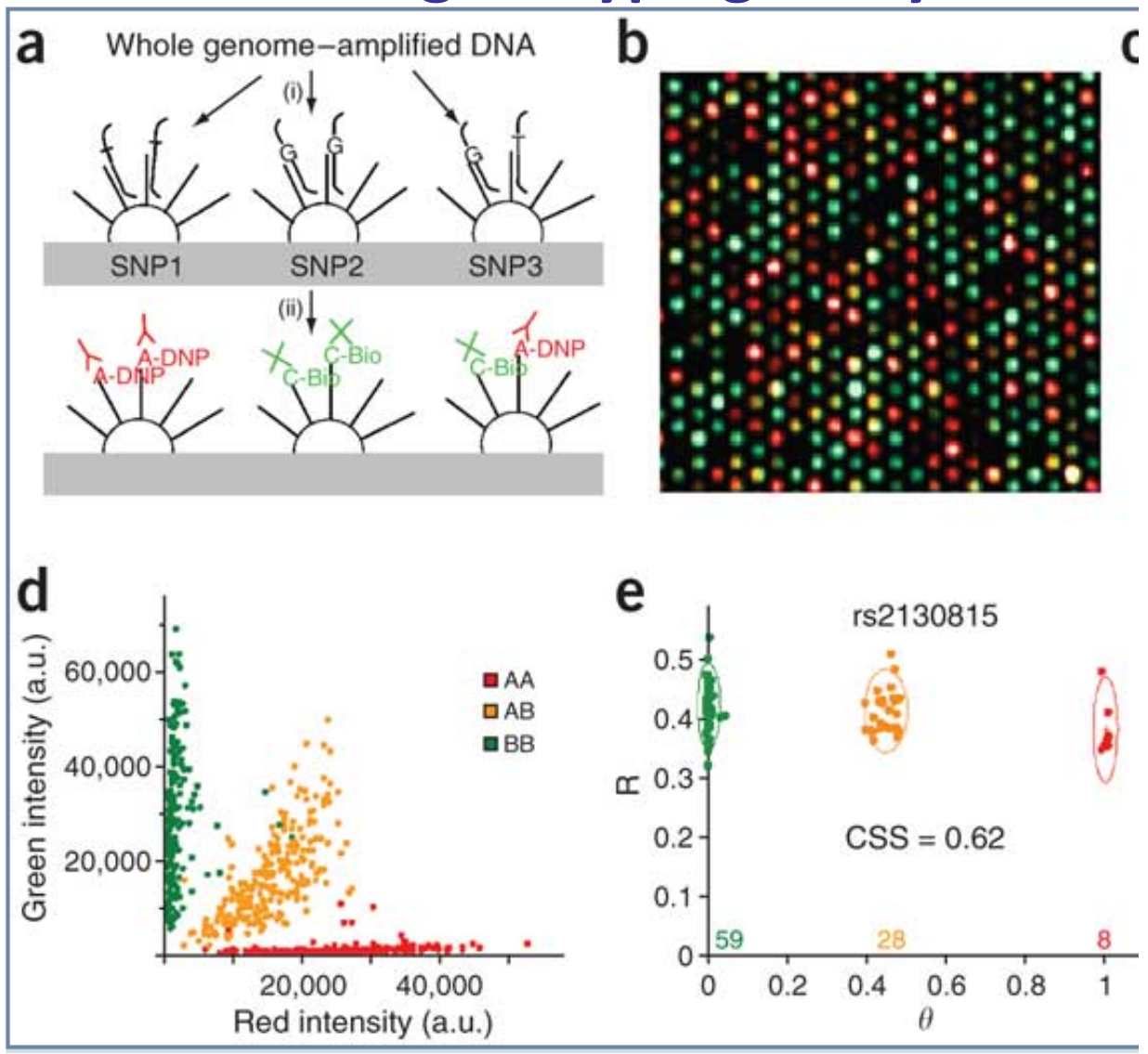
3-State HMM



Segmentation using a 3-state HMM (Viterbi Algorithm)

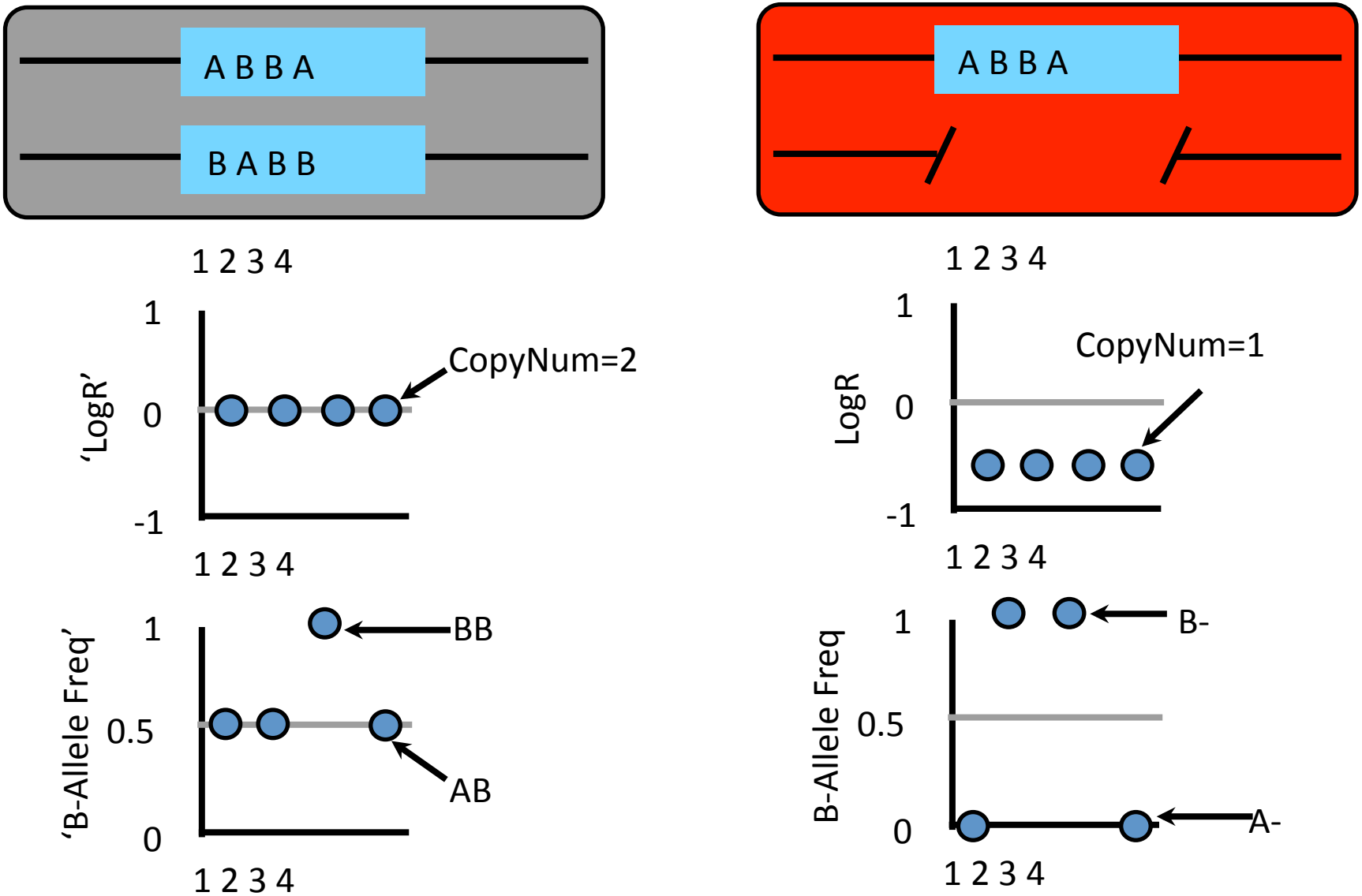


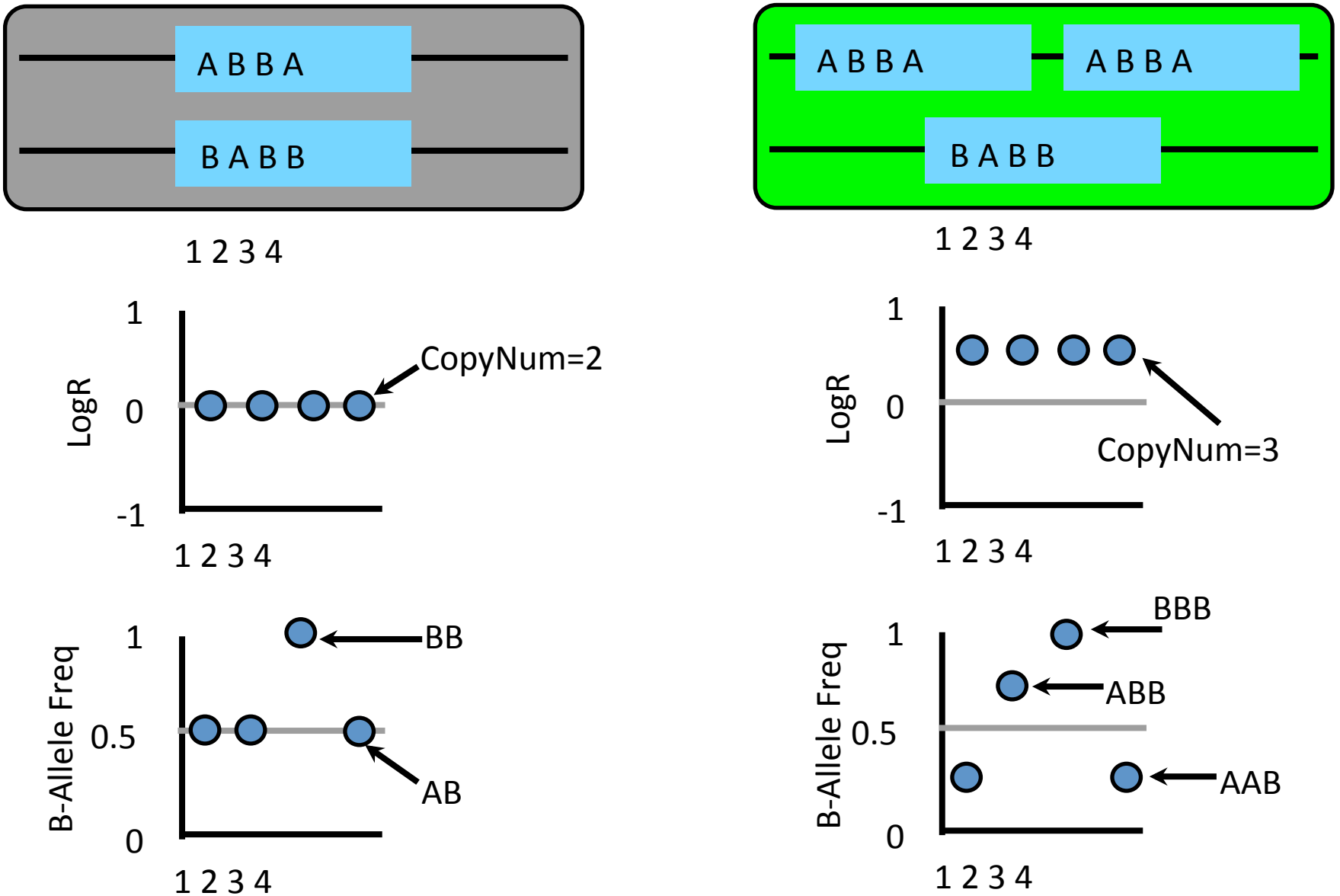
METHOD 2: Copy Number Variation: SNP genotyping Array

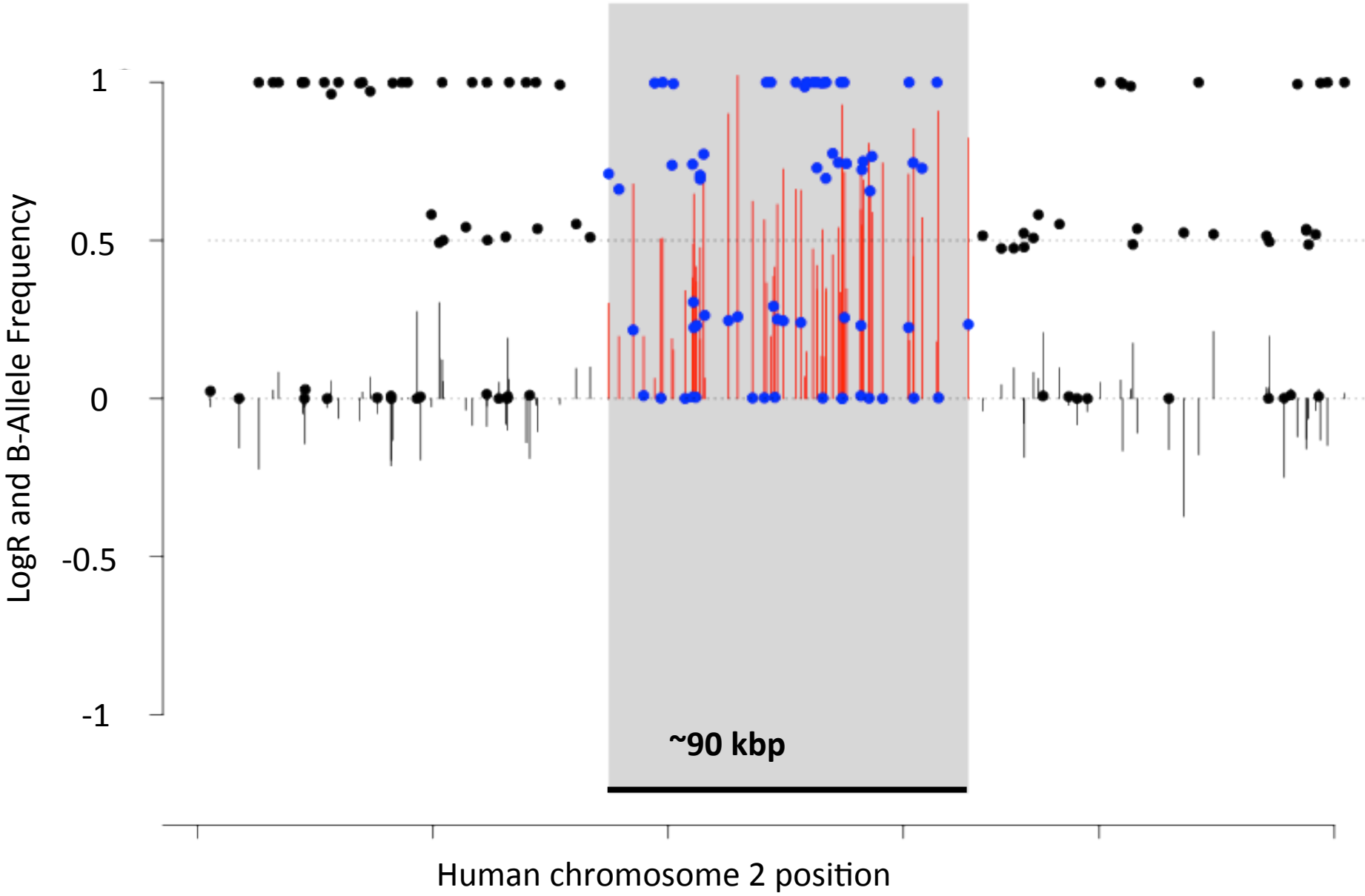


Steemers *et al.*

SNP Fluorescence-Based Deletion Discovery



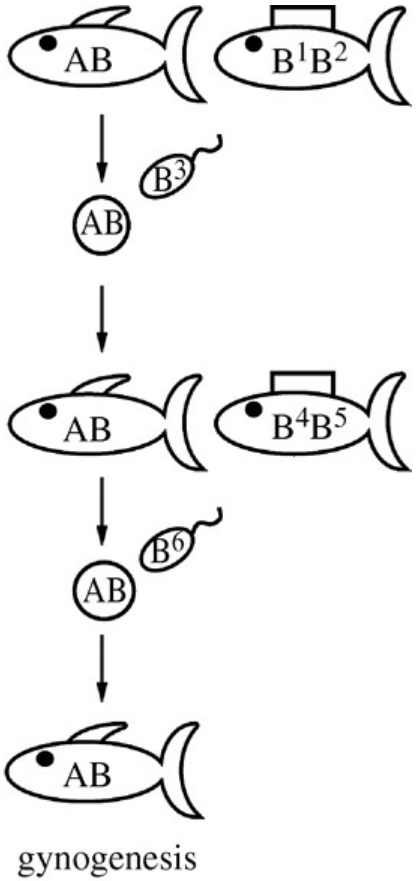

A B B A
B A B B
1 2 3 4
LogR
1
0
-1
CopyNum=2
1 2 3 4
B-Allele Freq
1
0.5
0
BB
AB
1 2 3 4
A B B A
A B B A
B A B B
1 2 3 4
LogR
1
0
-1
CopyNum=3
1 2 3 4
B-Allele Freq
1
0.5
0
BBB
ABB
AAB
1 2 3 4


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

Amazon Molly (*Poecilia formosa*)

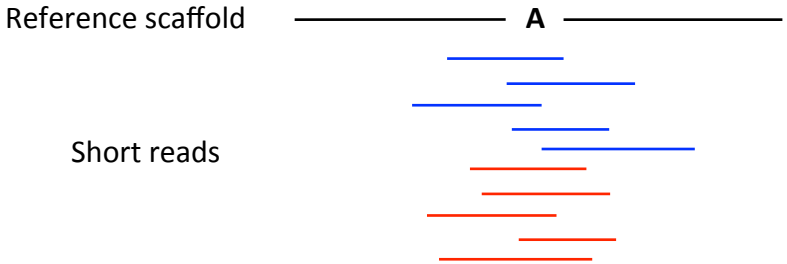
What is it that makes the Amazon molly so unique?

- The Amazon molly is one of the few extant asexual vertebrates
- All individuals in the species are females
- They reproduce by gynogenesis (sperm-dependant parthenogenesis)



Lampert, K.P. 2006

Detecting paternal introgression

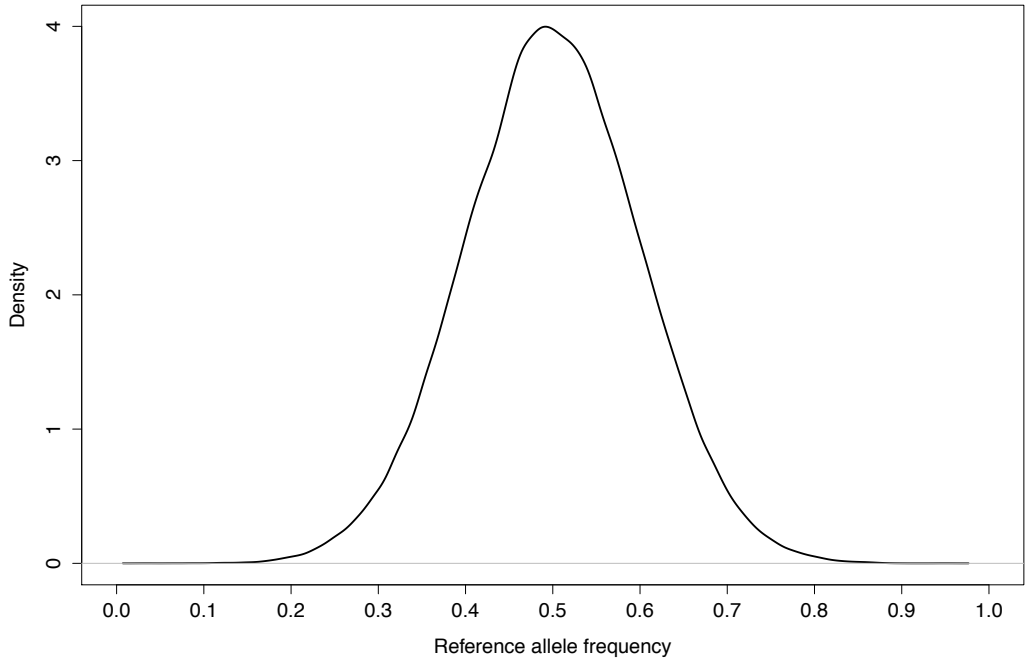


Reads that carry the reference allele **A** : 5

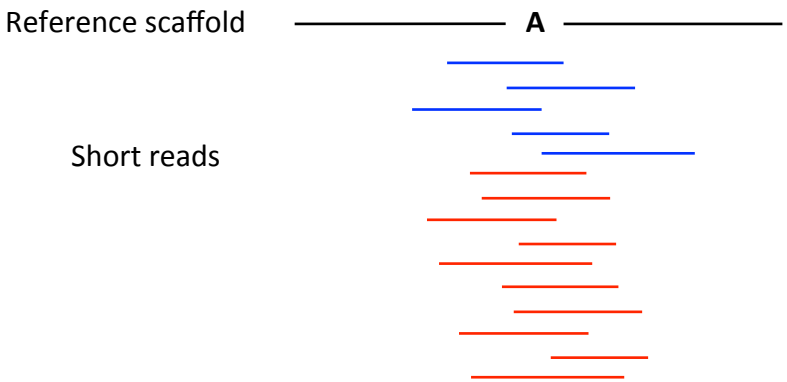
Reads that carry the alternative allele **C** : 5

Frequency of the alternative allele: 5/10 = 0.5

Scaffold dotation



Detecting paternal introgression

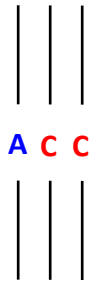


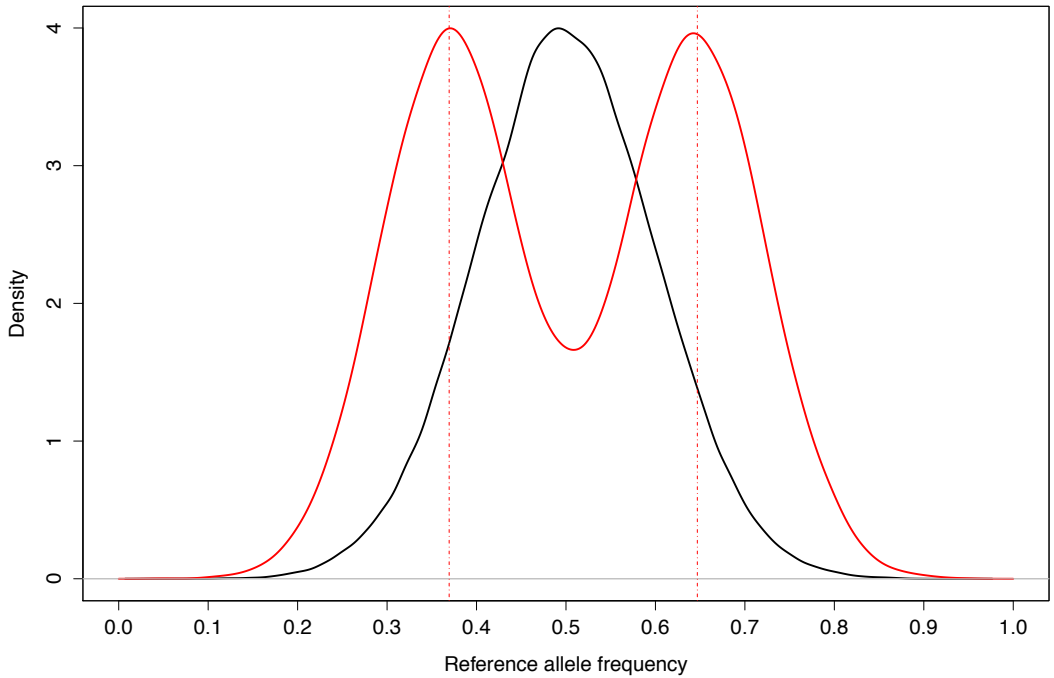
Reads that carry the reference allele **A** : 5

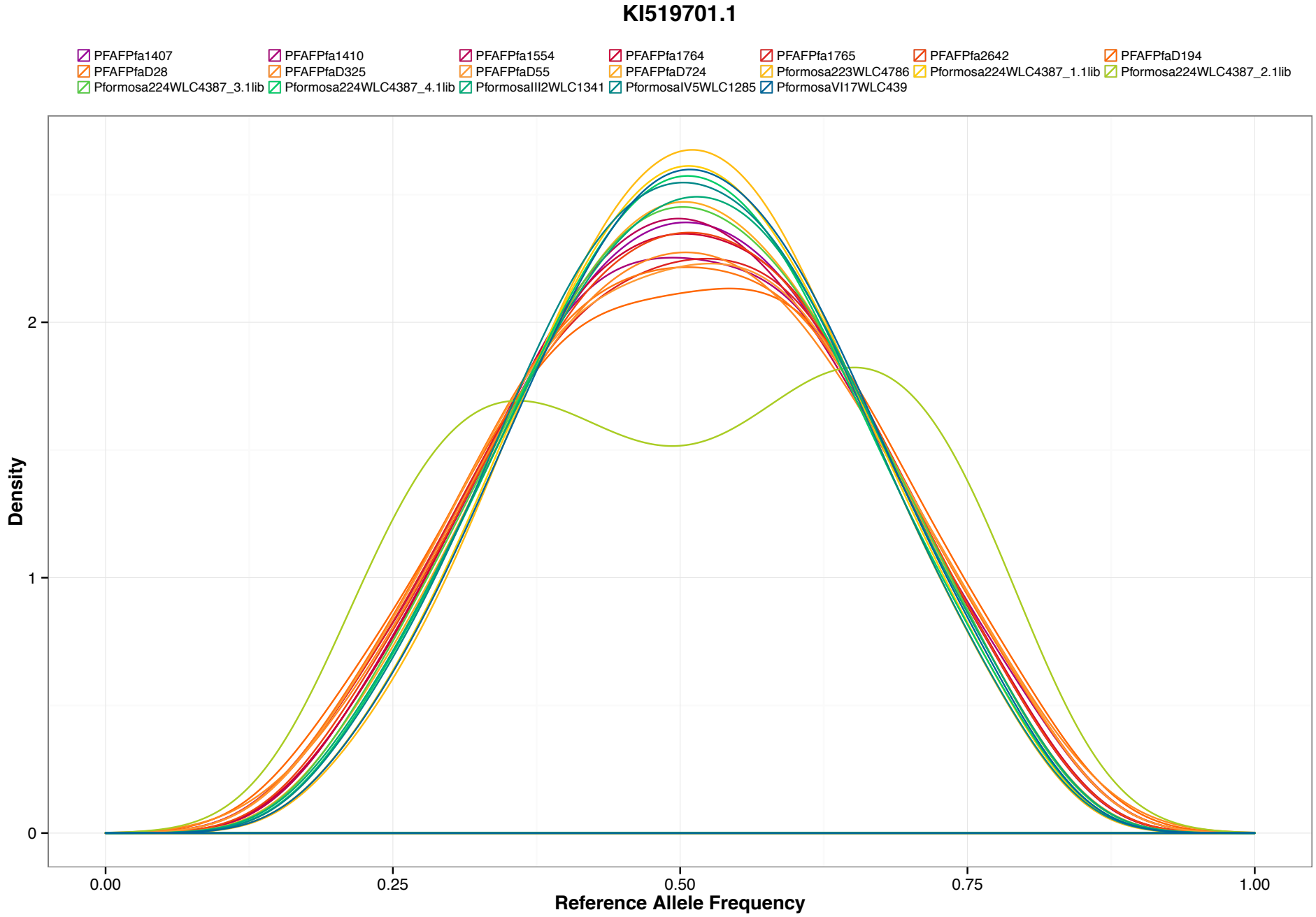
Reads that carry the reference allele **C** : 10

Frequency of the alternative allele: 5/15 = 0.33

Scaffold dotation

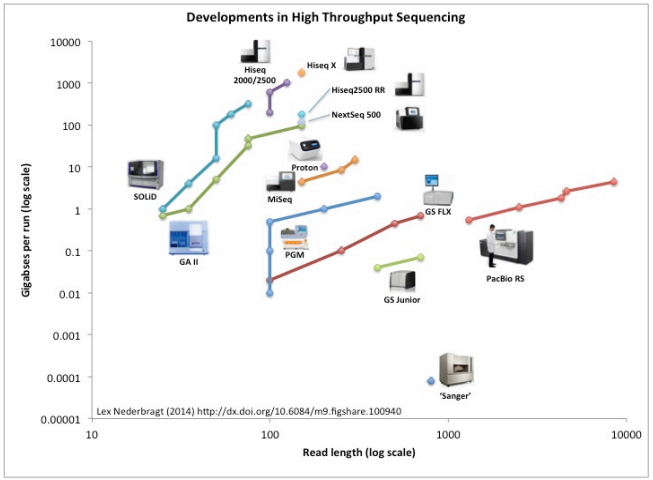





KI519701.1
PFAFPfa1407
PFAFPfa1410
PFAFPfa1554
PFAFPfa1764
PFAFPfa1765
PFAFPfa2642
PFAFPfaD194
PFAFPfaD28
PFAFPfaD325
PFAFPfaD55
PFAFPfaD724
Pformosa223WLC4786
Pformosa224WLC4387_1.1lib
Pformosa224WLC4387_2.1lib
Pformosa224WLC4387_3.1lib
Pformosa224WLC4387_4.1lib
PformosaII2WLC1341
PformosaIV5WLC1285
PformosaVI17WLC439
Density
0
1
2
0.00
0.25
0.50
0.75
1.00
Reference Allele Frequency

Sequencing Methods

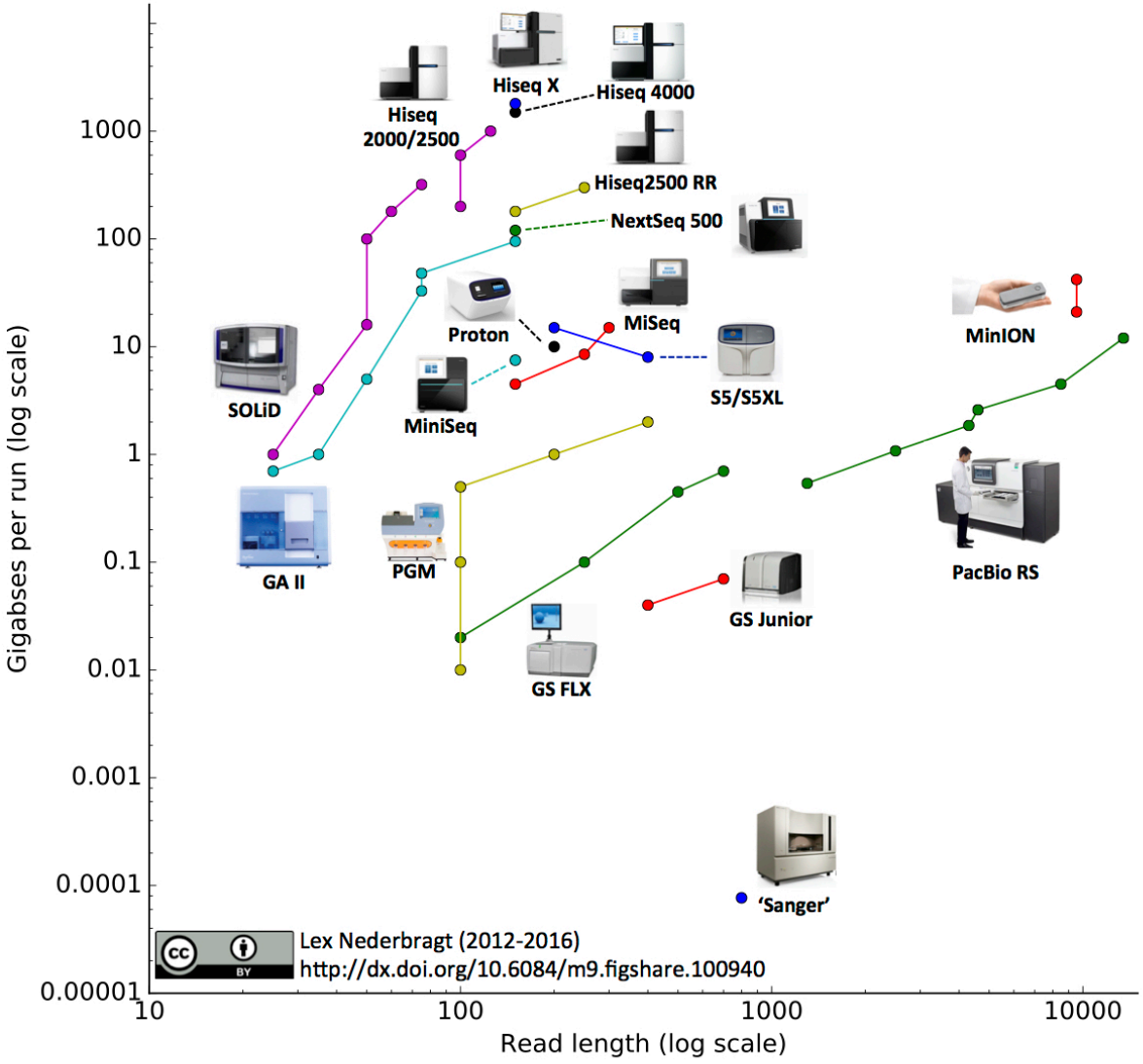
•Going Backwards…
Sanger, 454, Illumina…..



•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
  - •Finish sequence and better assemblies

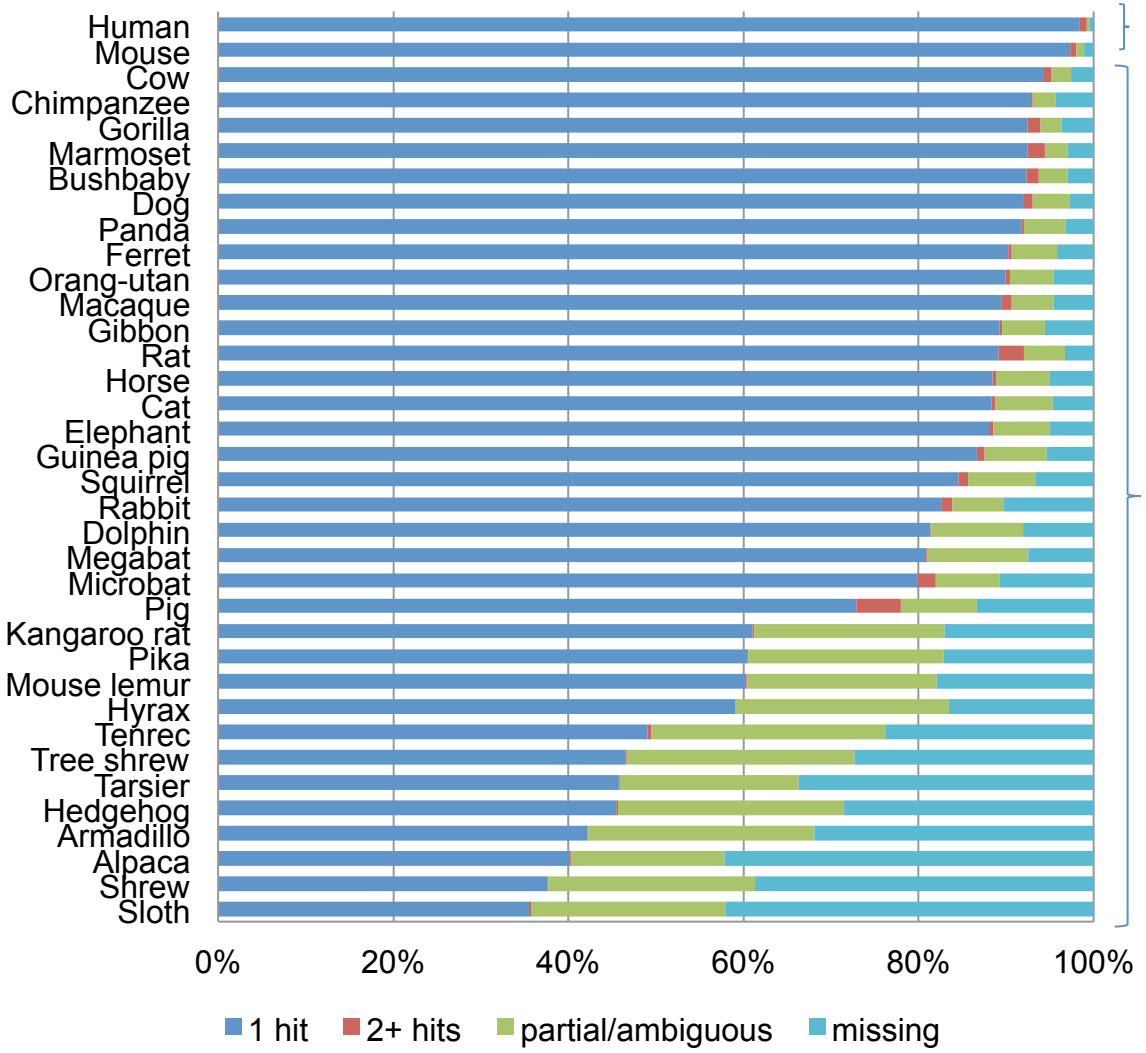
Sequencing Methods



De novo assemblies

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

Quality of “old days” assemblies



FINISHED

Phasing (Clone sequencing)

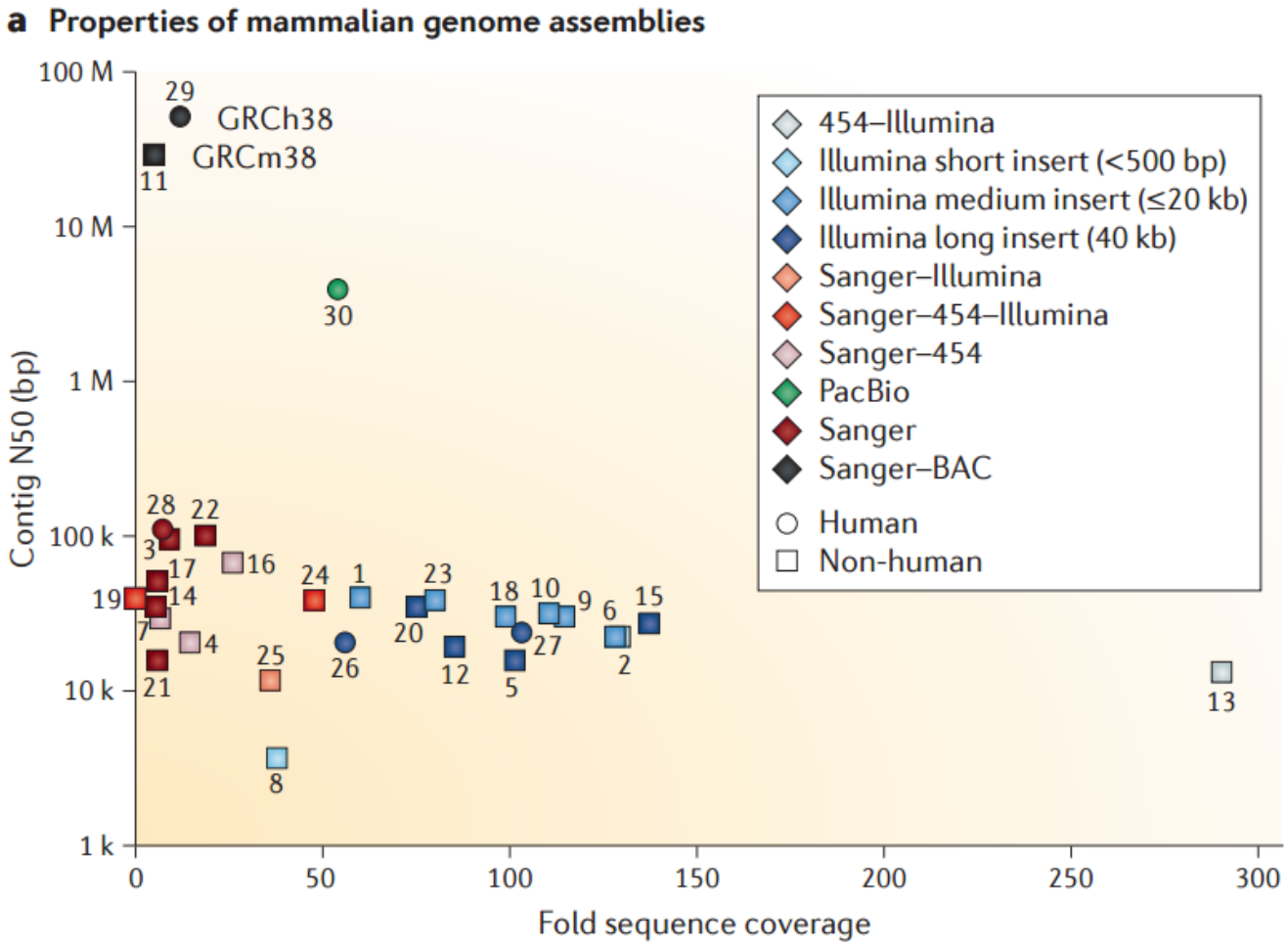
Clone end (BACs)

Long reads (Sanger)

DRAFT

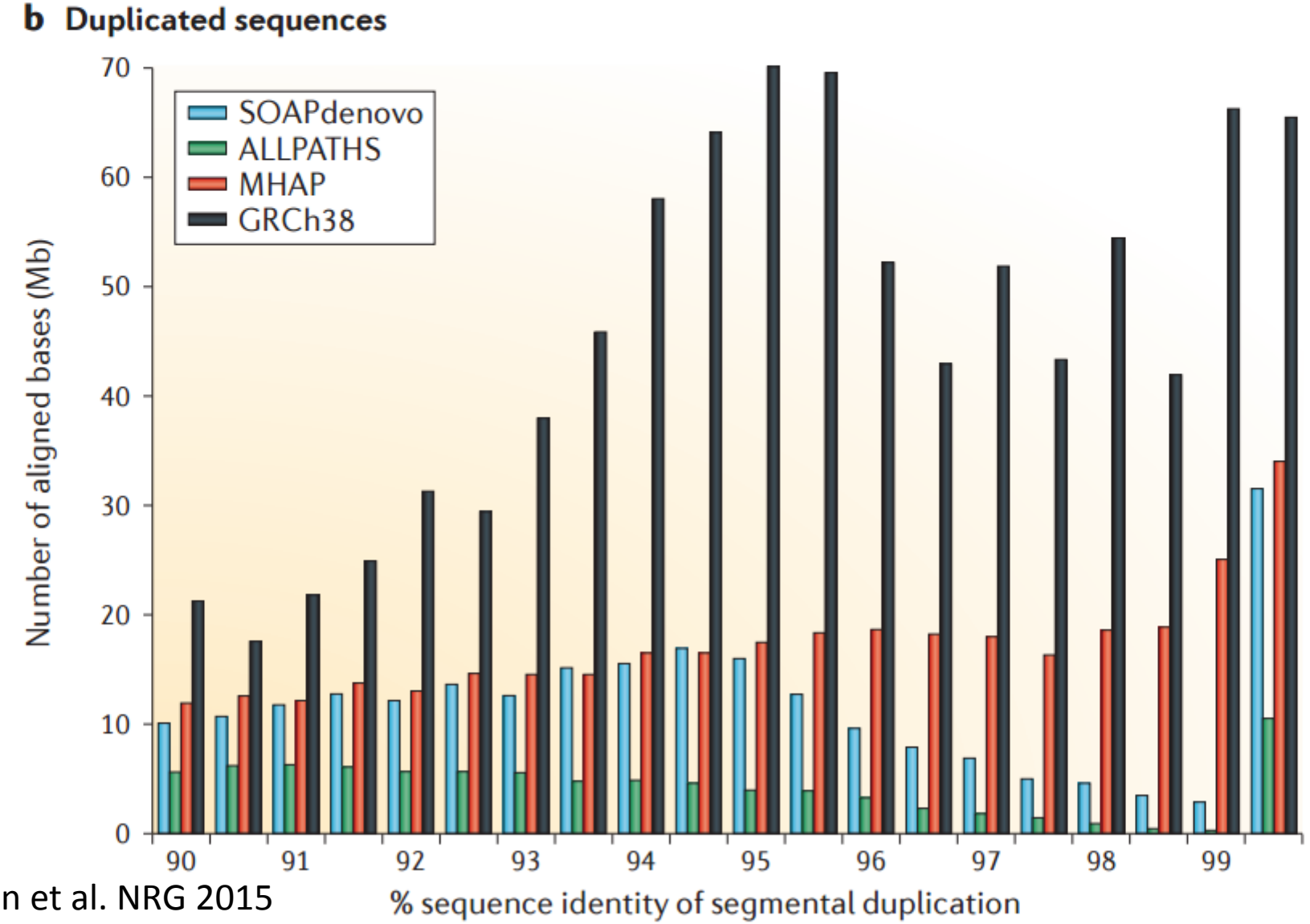
Proportion of 11K mammalian genes

Contigs are small!



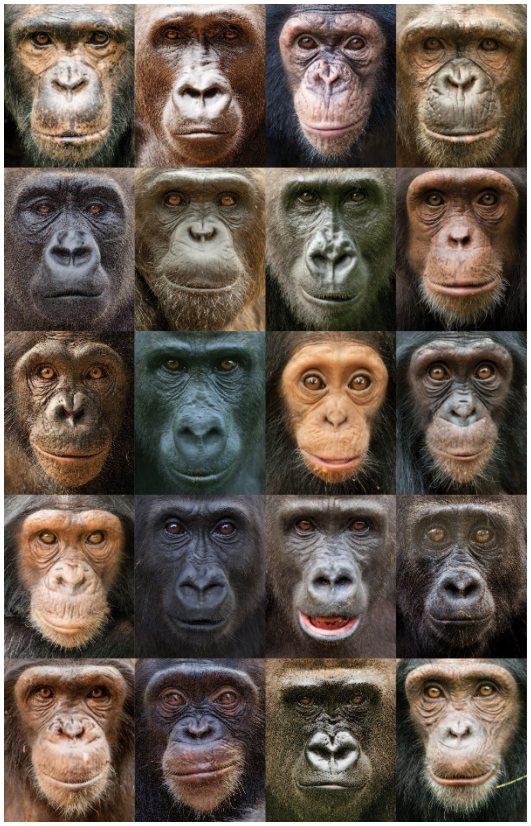
Chaisson et al. NRG 2015

Limitations of NGS assemblies



Chaisson et al. NRG 2015

Chimpanzee data to improve the reference



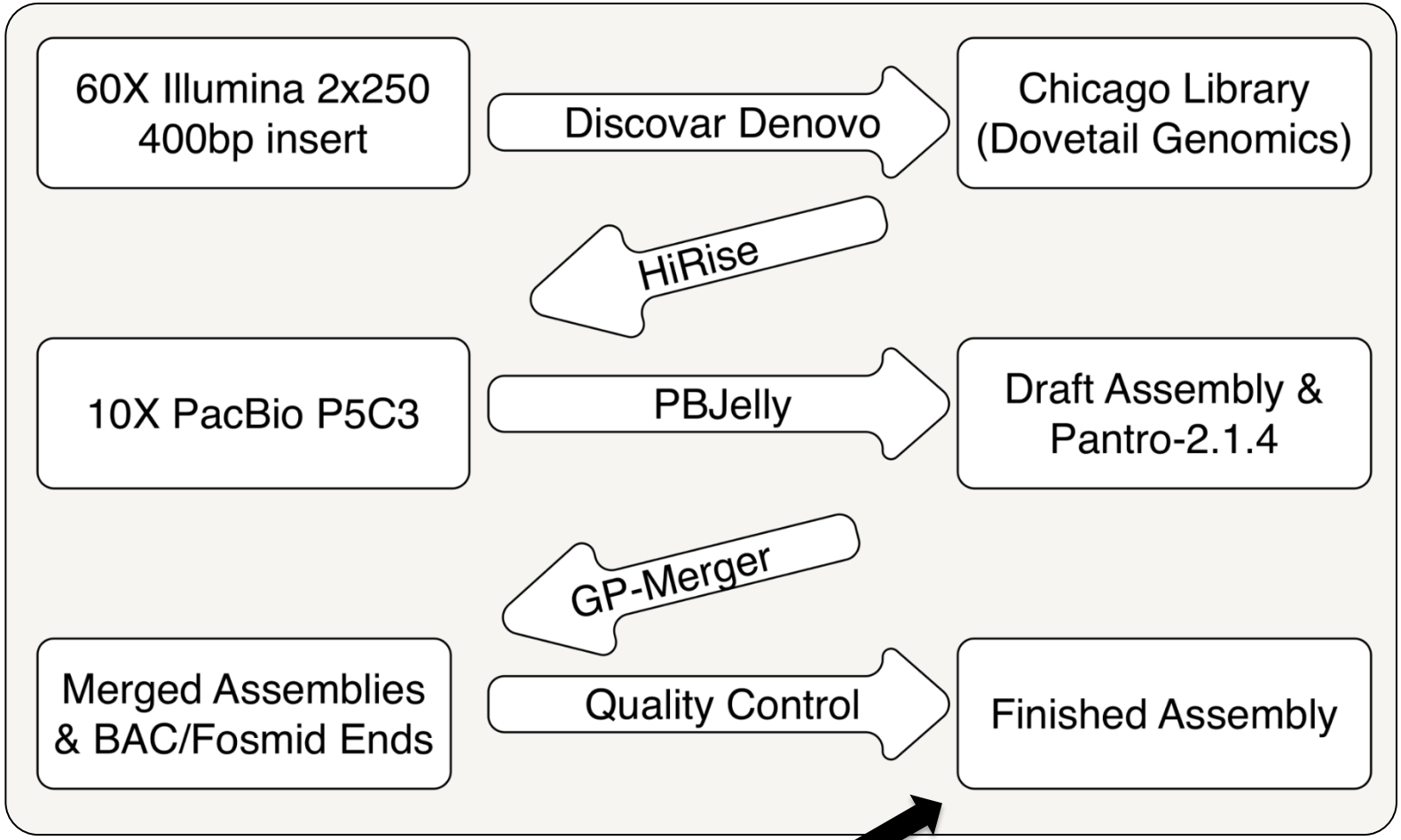
- 6x Sanger (plasmids, fosmids)
- BAC end reads
- 1000 finished BACs

New generated data

- Fosmid pool sequencing ($$$)
- Bionano (optical mapping) ($$)
- 80x Illumina (short insert) ($)
- 20 X Illumina mate pairs ($)
- ~100X Illumina short ins size, PE overlap ($)
- 10x PacBio from a 20k library ($$)*
- 3x Moleculo ($$)*
- 2 lanes HiC (Chicago) Dovetail ($$)

Wes Warren (WashU),
Andrew Sharp (Mount Sinai),
Lars Feuk (Uppsala)

Our assembly pipe



Break large misassemblies (N=17)
Error correction (base substitutions, small indels)

What is each data good for?

- **60 X Illumina 2x250**: Initial contiguous fragments (contigs) of the genome

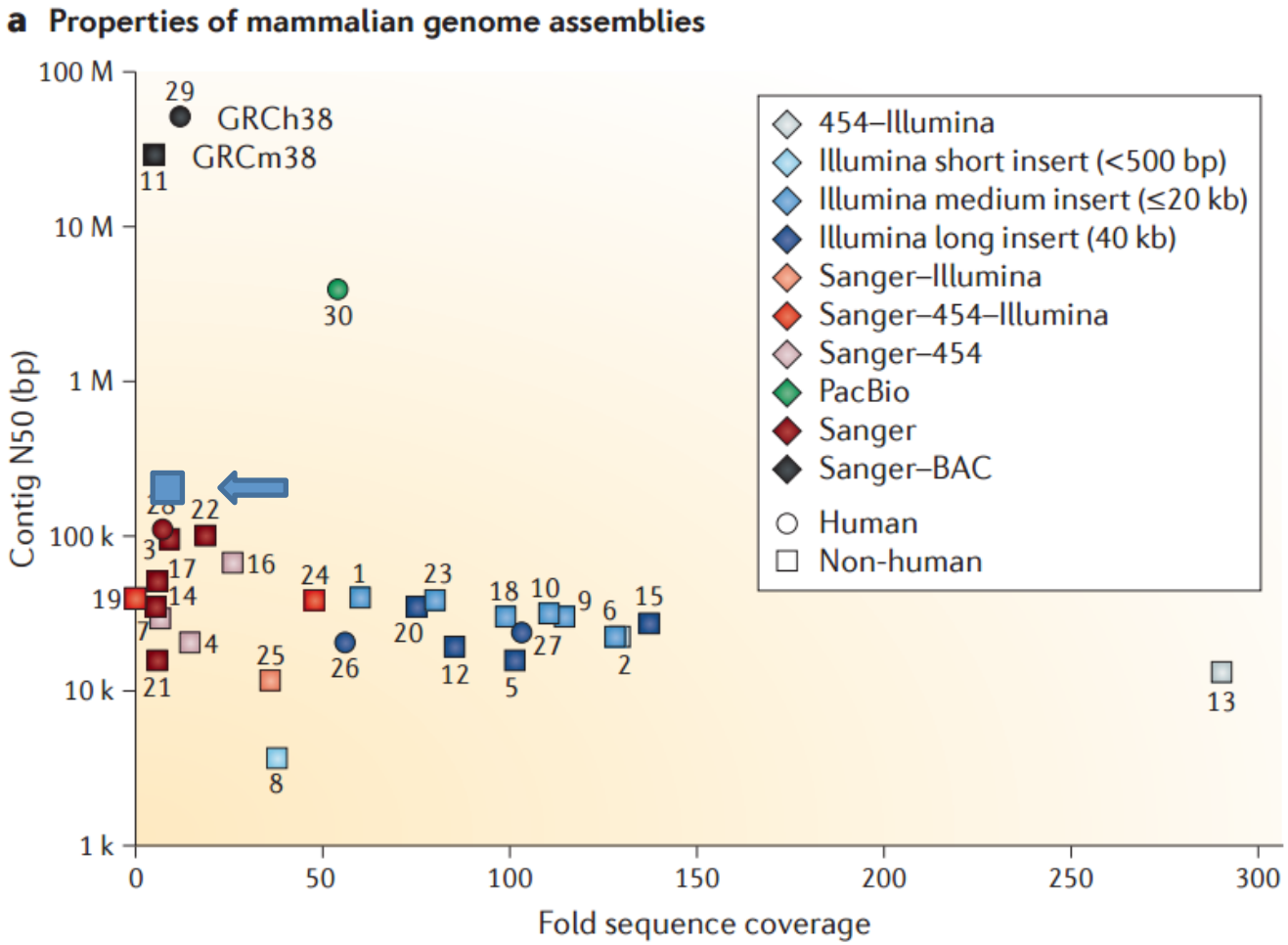
- **Dovetail Genomics:** Long range information of order & orientation of contigs (scaffolds) (closed system)

- **10X PacBio:** Fill in remaining gaps (stretches of unknown sequence within scaffolds, between 2 contigs)

What do we get?

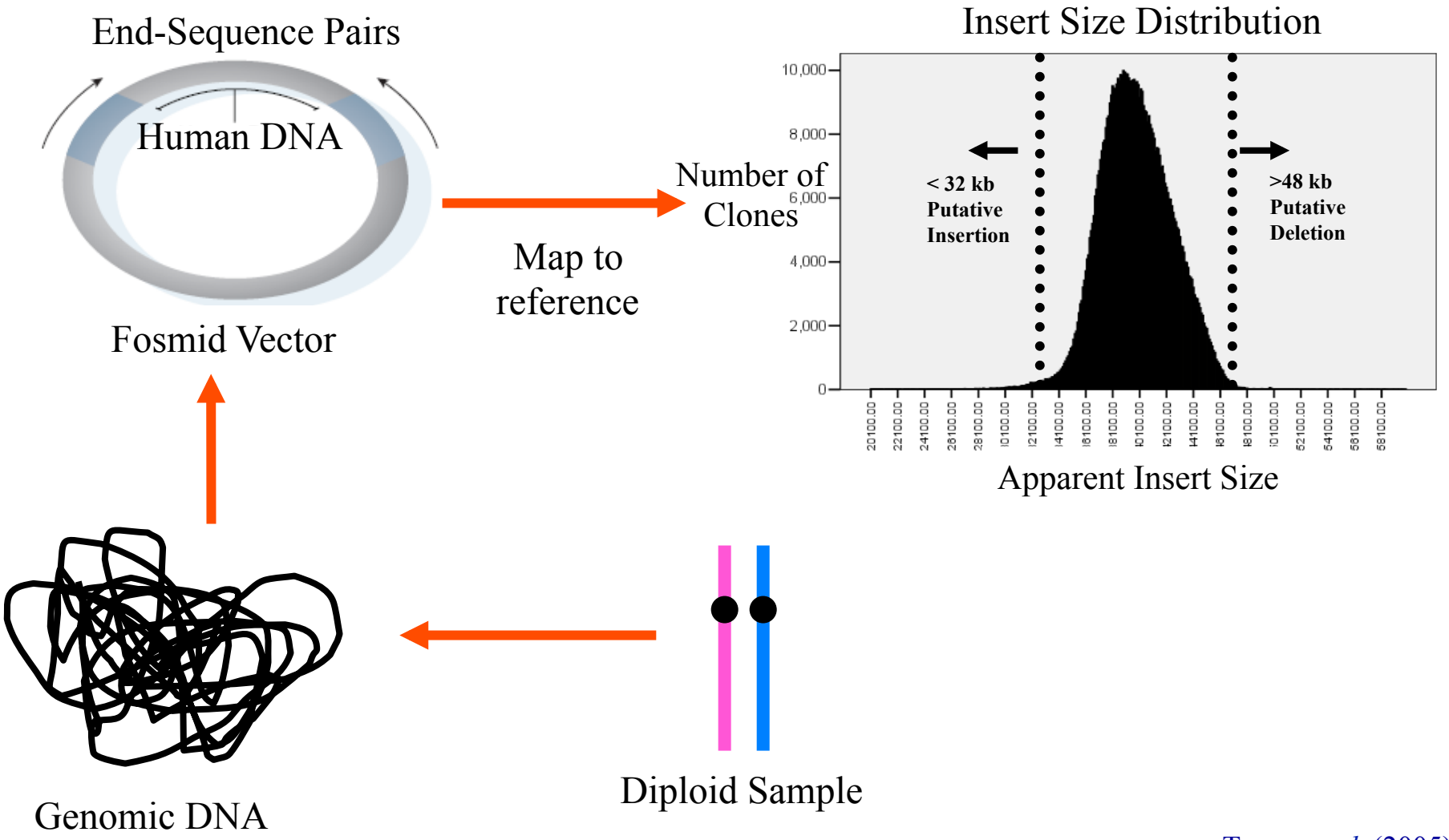
	Pantro-2.1.4	Pantro5V0.2 preliminary
Number of scaffolds	24,129	45,000
Scaffold N50 (bp)	8,925,874	26,673,241
Number of contigs	148,553	72,817
Contig N50 (bp)	64,231	334,510

Where will the chimp go?



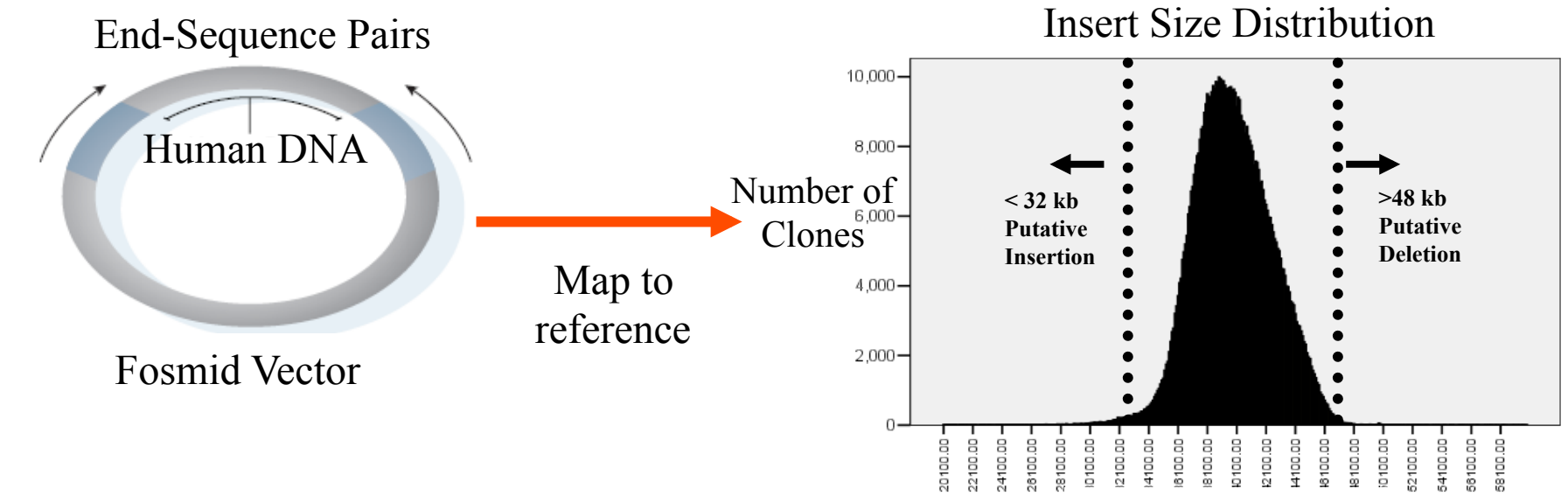
Chaisson et al. NRG 2015

Method 2: End-Sequence Pair (ESP) Analysis



Tuzun *et al.* (2005)

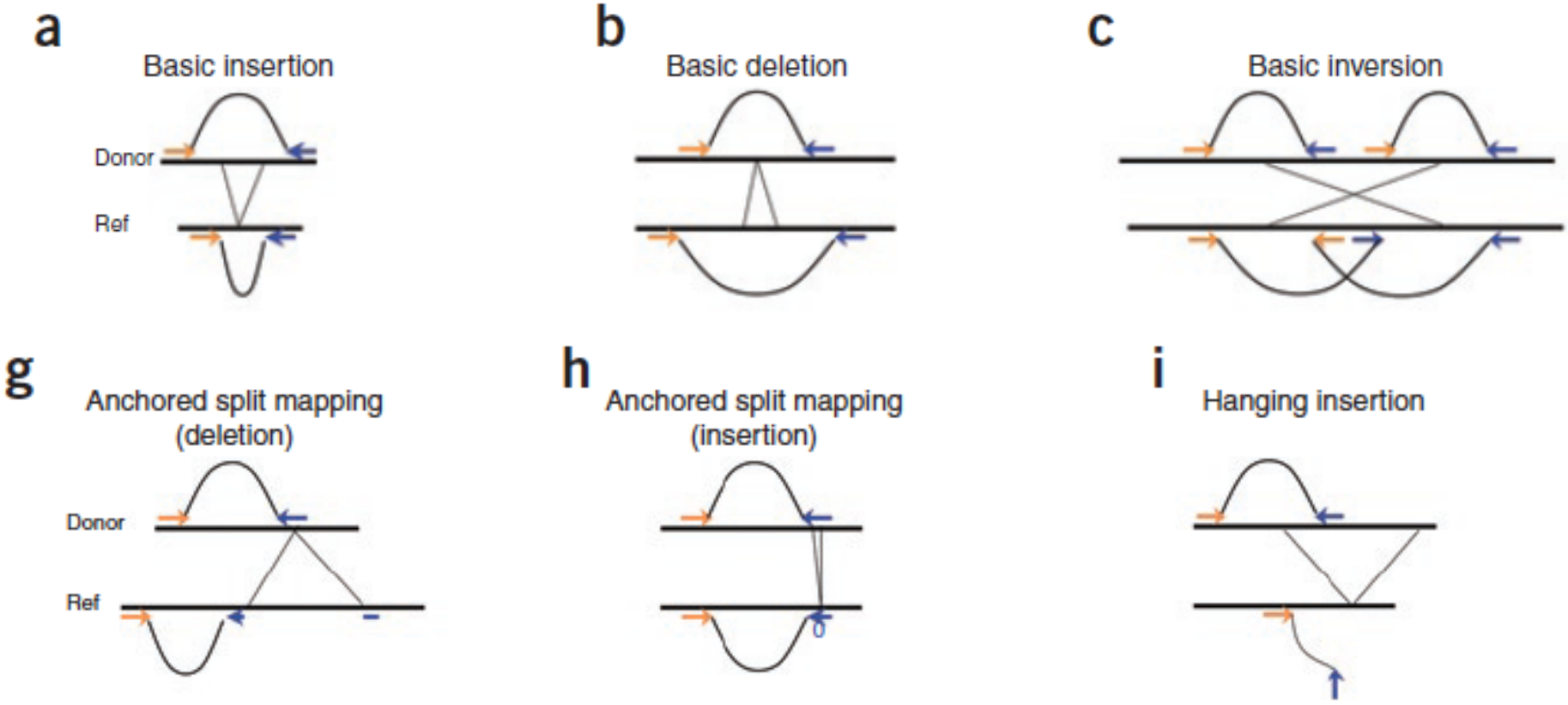
Method 2: End-Sequence Pair (ESP) Analysis



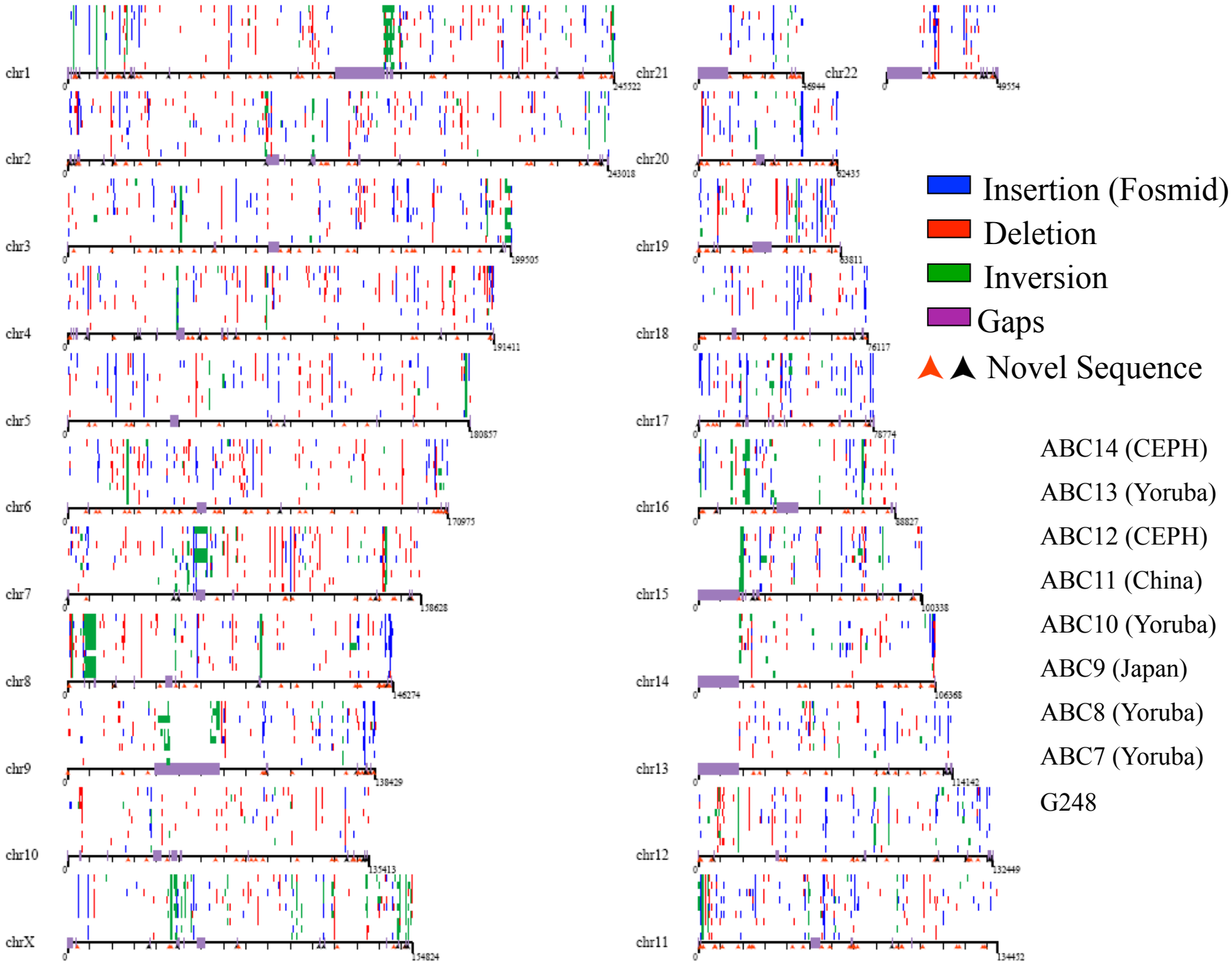
Tuzun *et al.* (2005)

What can we find?

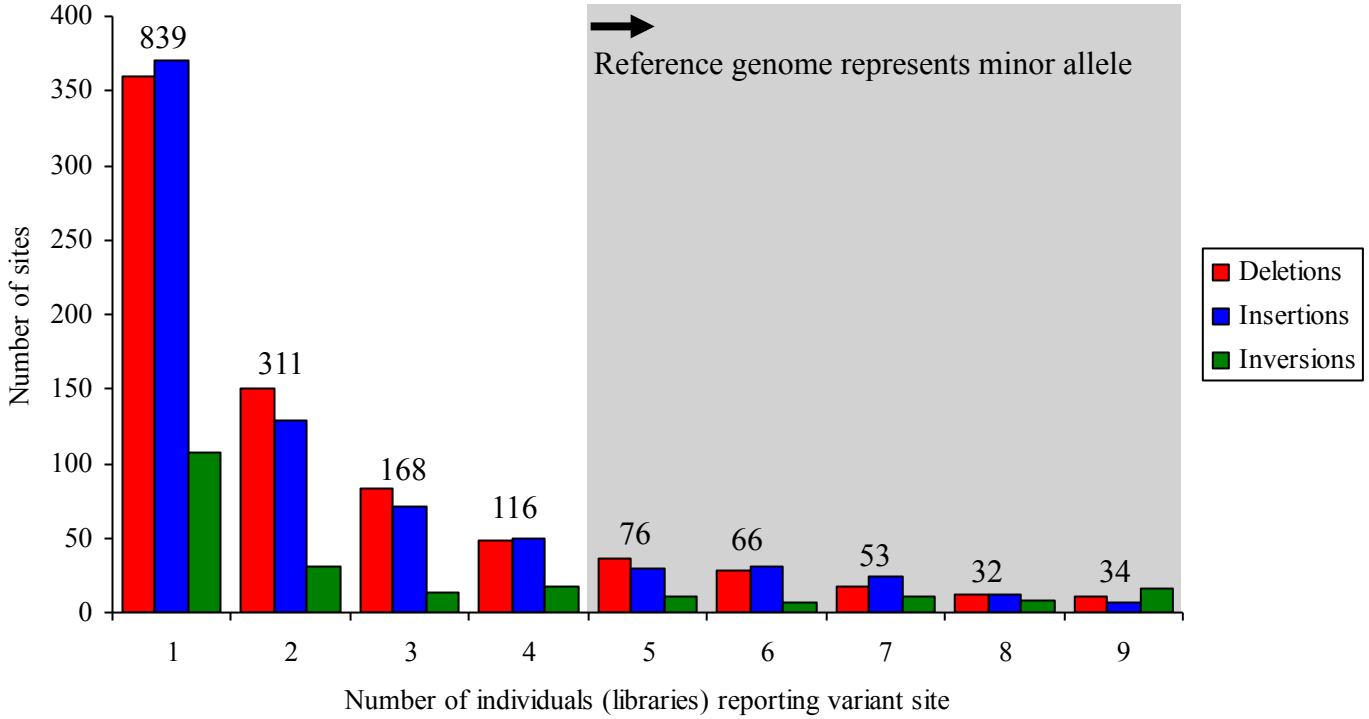
Structural variation detection:



Alkan et al. Nature Review Genetics 2011

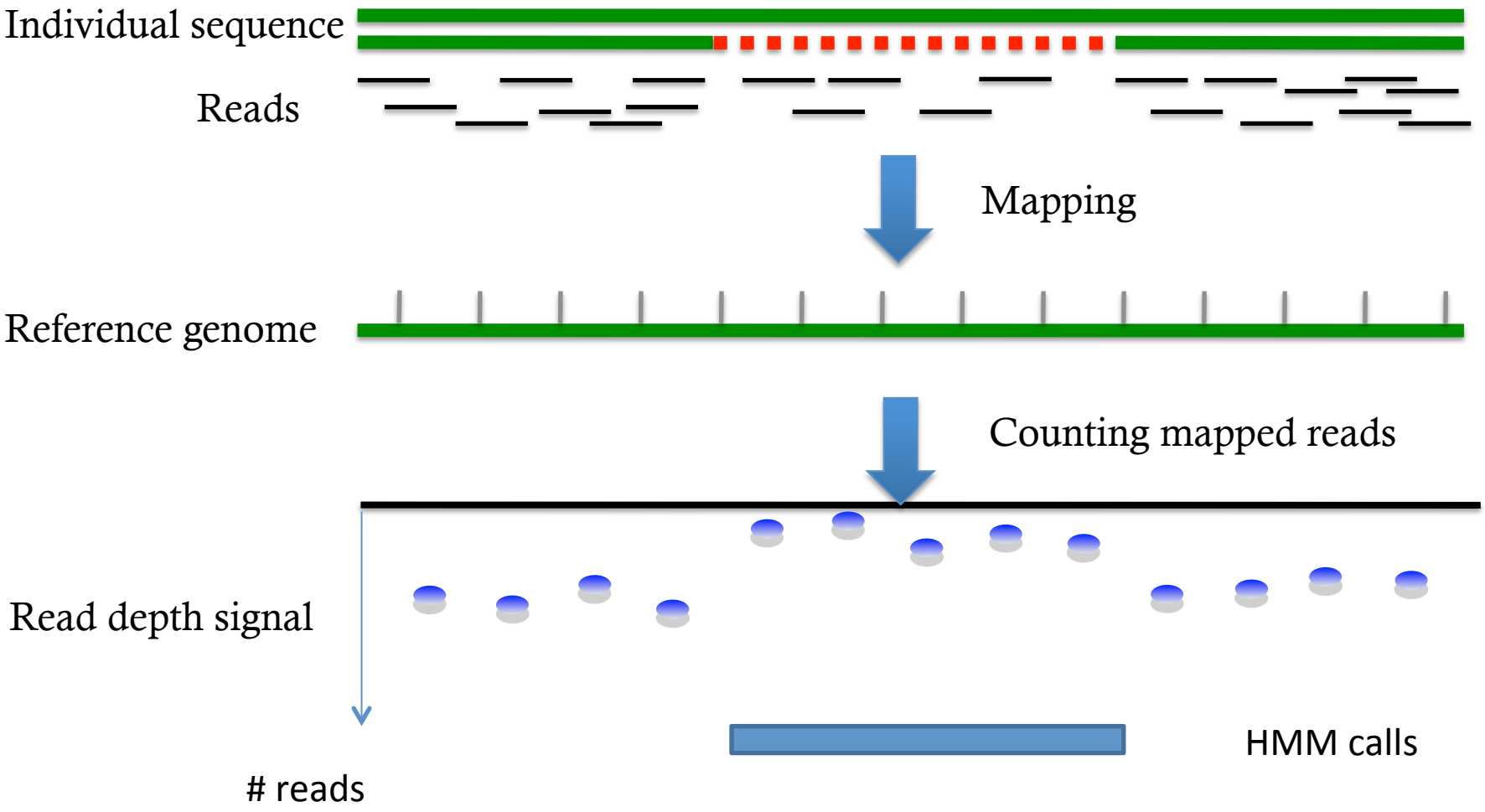

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chrX
chr21
chr22
chr20
chr19
chr18
chr17
chr16
chr15
chr14
chr13
chr12
chr11
Insertion (Fosmid)
Deletion
Inversion
Gaps
Novel Sequence
ABC14 (CEPH)
ABC13 (Yoruba)
ABC12 (CEPH)
ABC11 (China)
ABC10 (Yoruba)
ABC9 (Japan)
ABC8 (Yoruba)
ABC7 (Yoruba)
G248

Frequency of Validated Sites

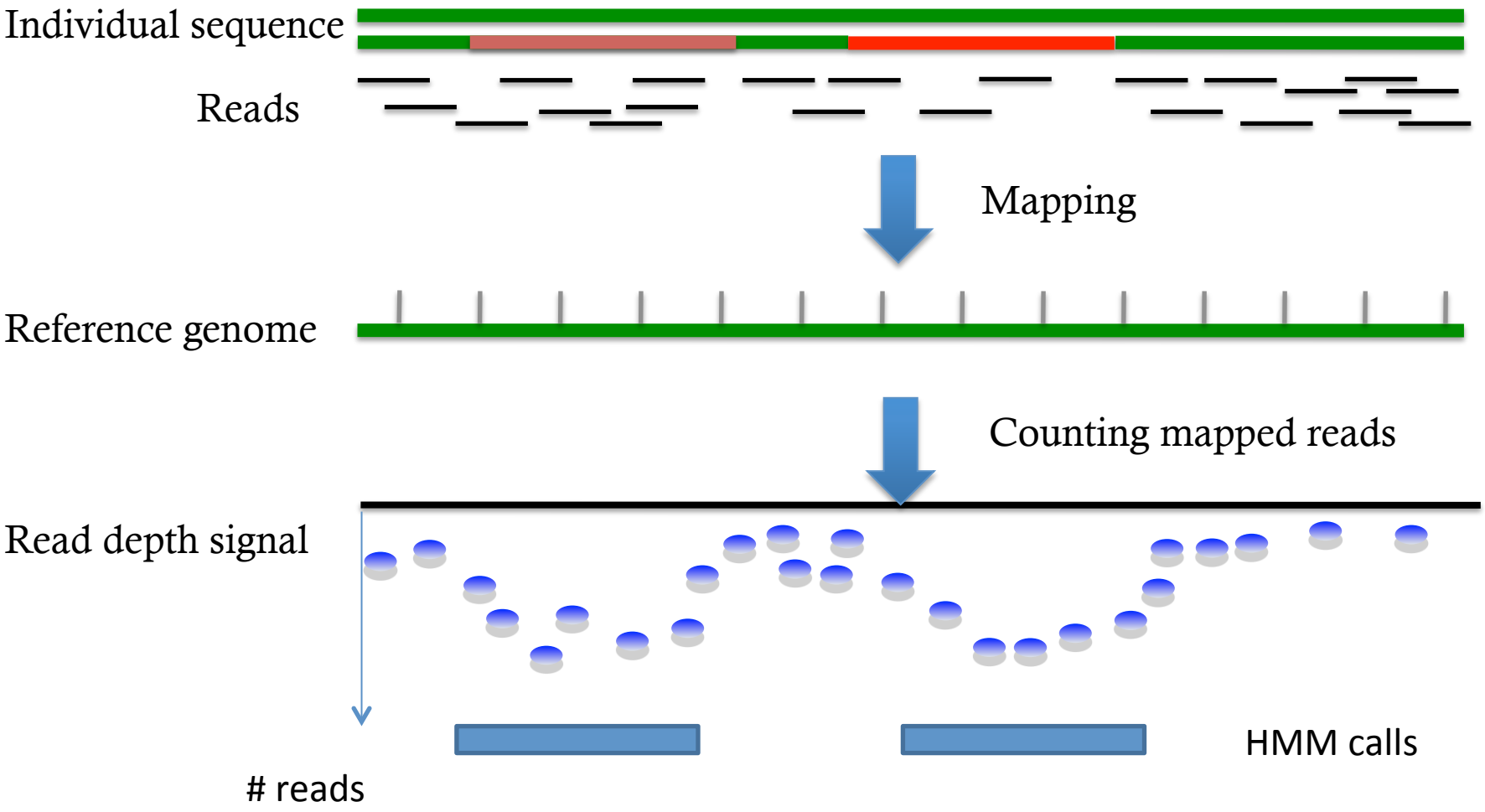


261 (15%) sites where reference genome represents a minor allele

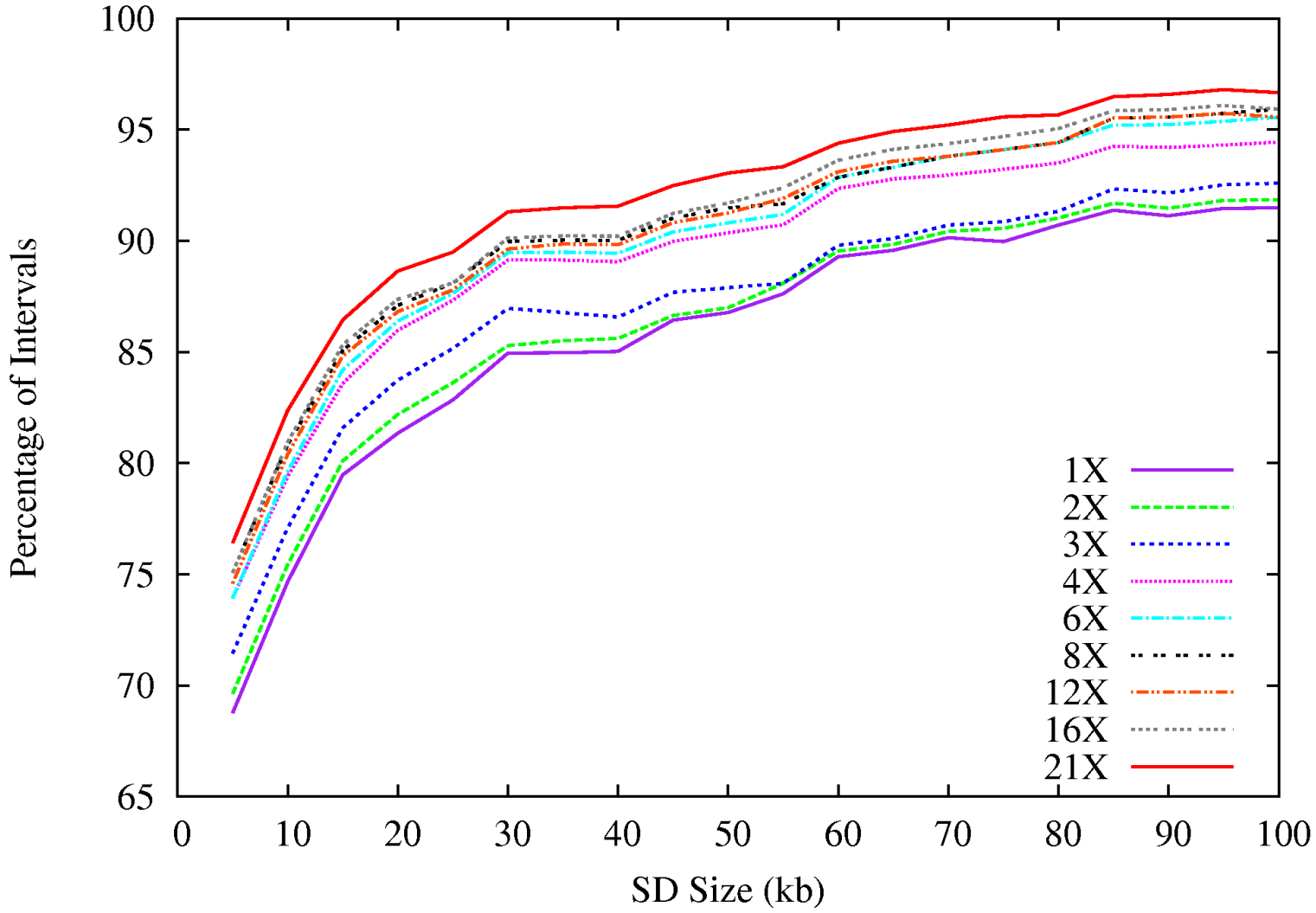
Method 3: Sequence Read Depth Analysis



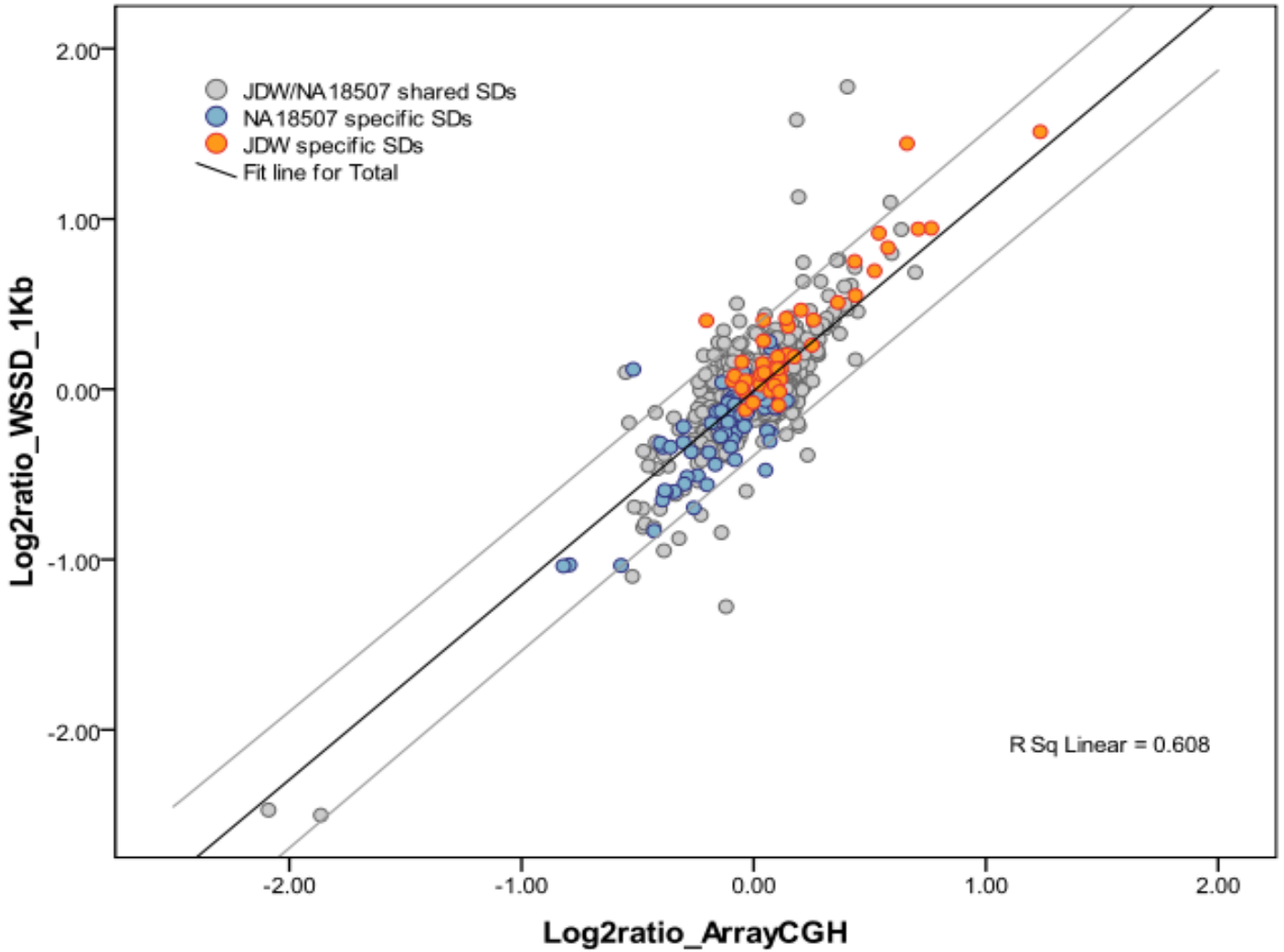
Method 3: Sequence Read Depth Analysis



Sequence coverage and detection power



Validation of copy-number estimations

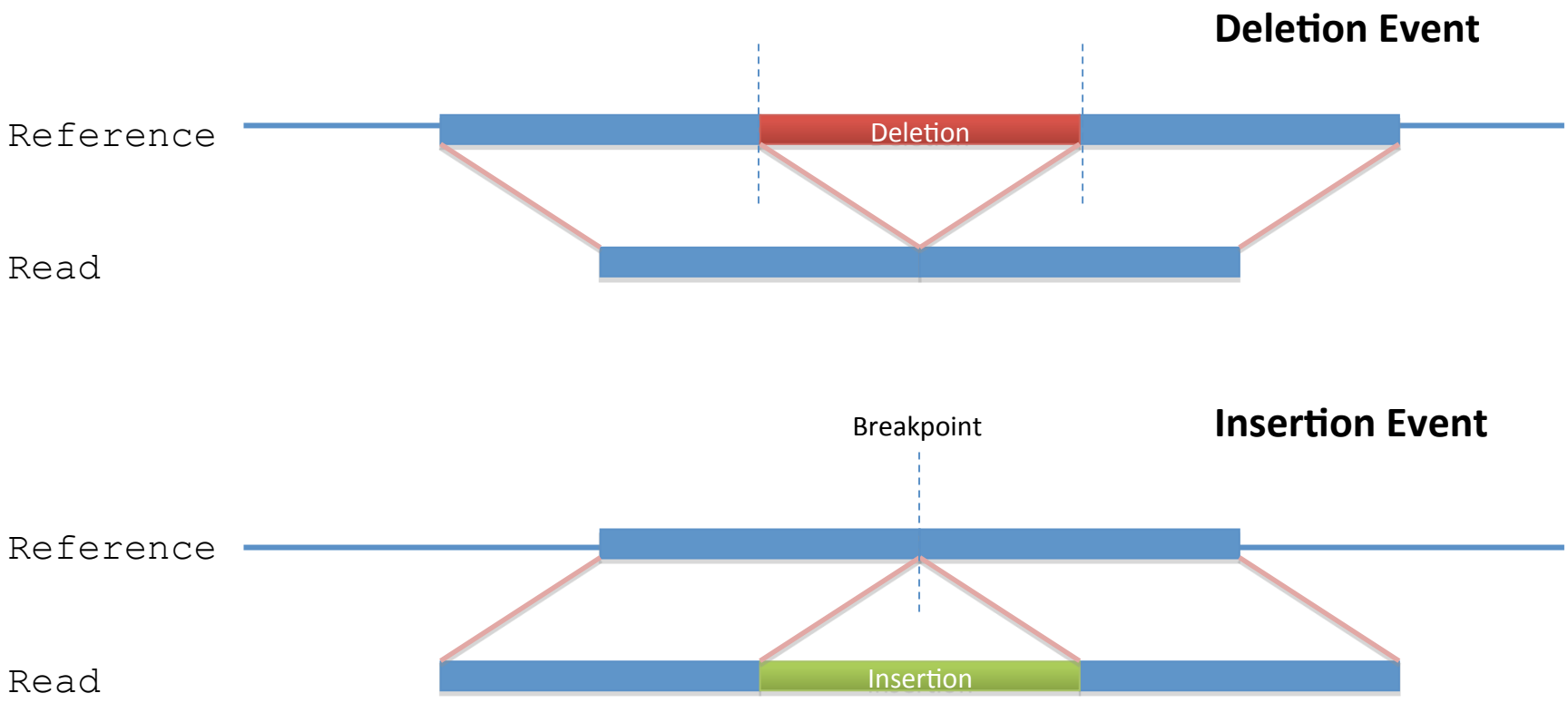


Alkan et al., Nature Genetics, 2009

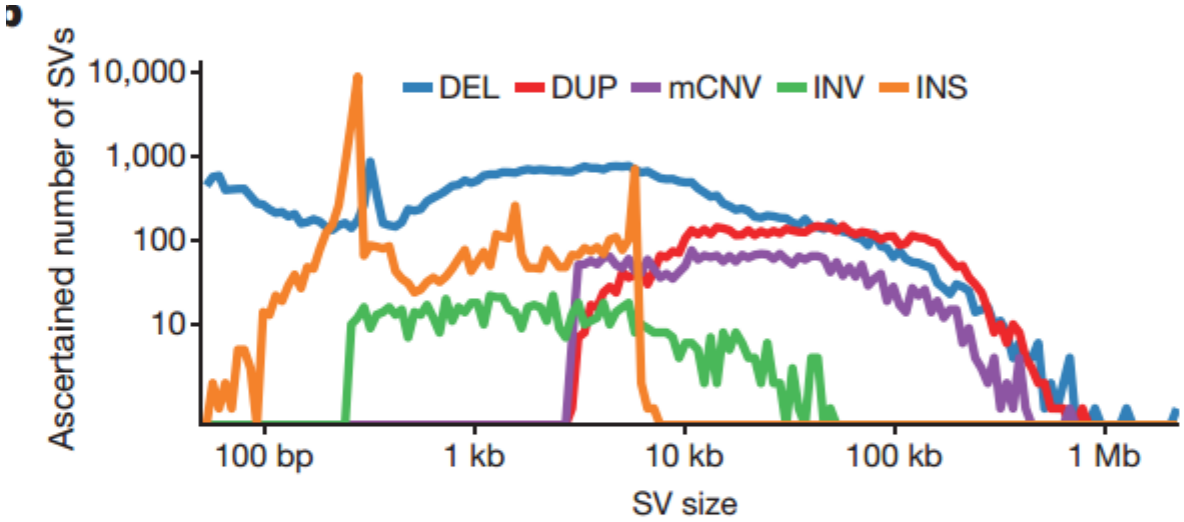
Deletions

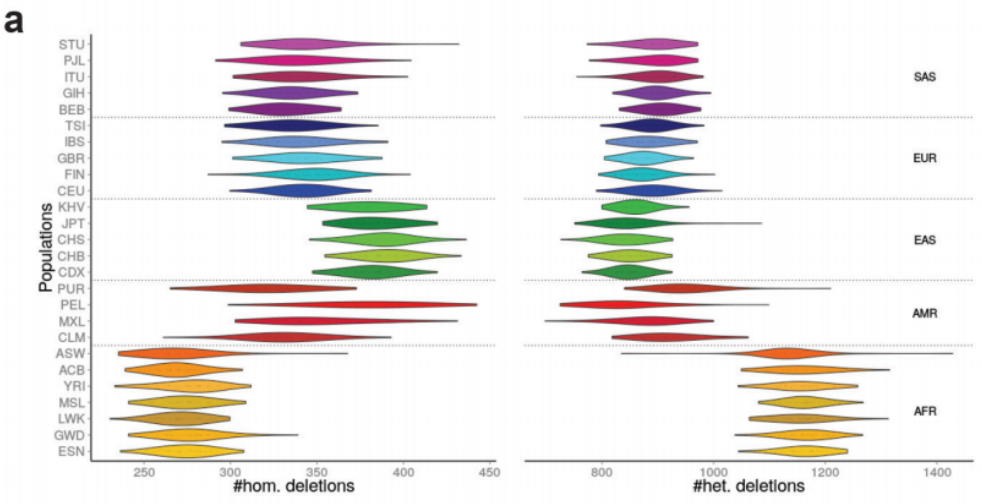

chr13:38,390,001-38,407,895
p12 p11.2 q11 q12.12 q12.3 q13.3 q14.11 q14.2 q21.1 q21.31 q21.33 q22.2 q31.1 q31.2 q32.1 q33.1 q33.3 q34
17 kb
38,392 kb
38,394 kb
38,396 kb
38,398 kb
38,400 kb
38,402 kb
38,404 kb
38,406 kb
NAME
DATA FILE
DATA TYPE
floquet_gem_no_split.bam Coverage
floquet_gem_no_split.bam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

Split-read Analysis



Moving to populations

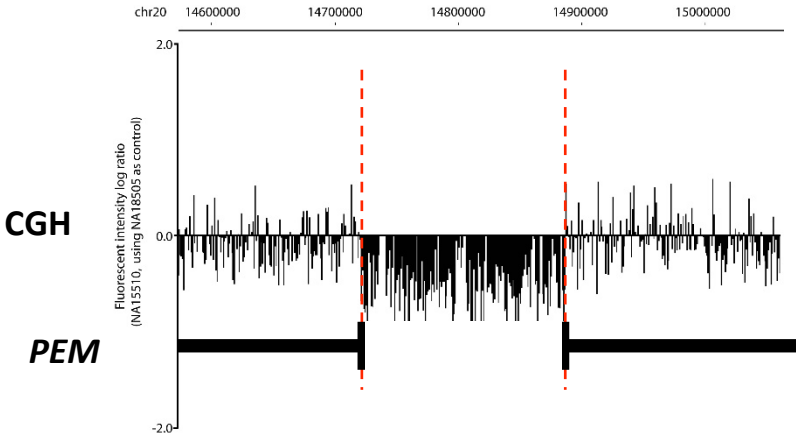




1KG Phase3 SV consortium. Nature 2015

Experimental Validation

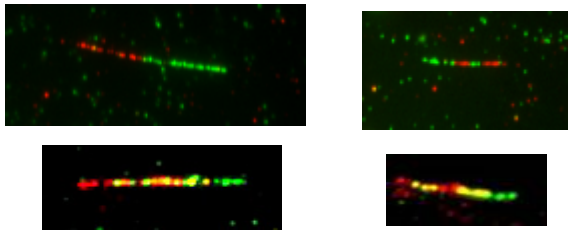
A) CGH



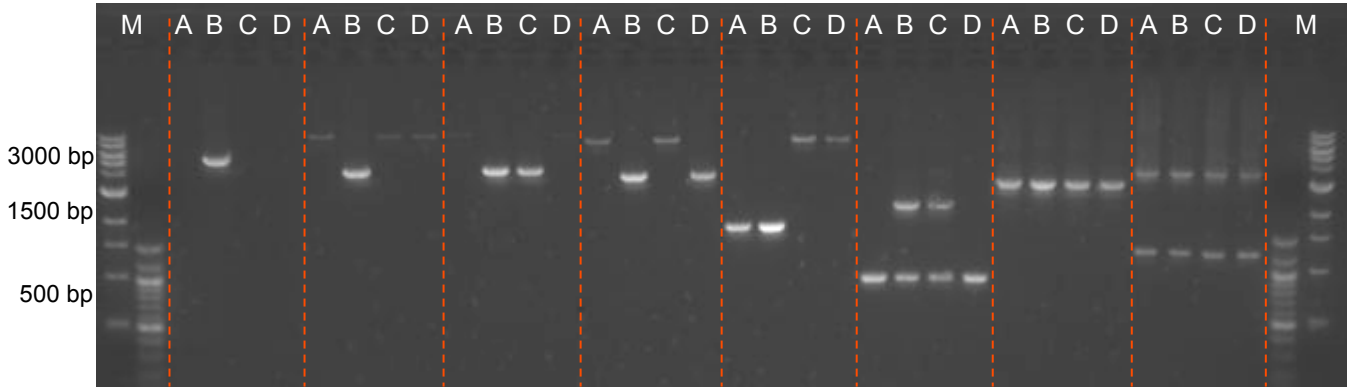
B) Fiber-FISH (For inversions)

Without inversion

With inversion



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

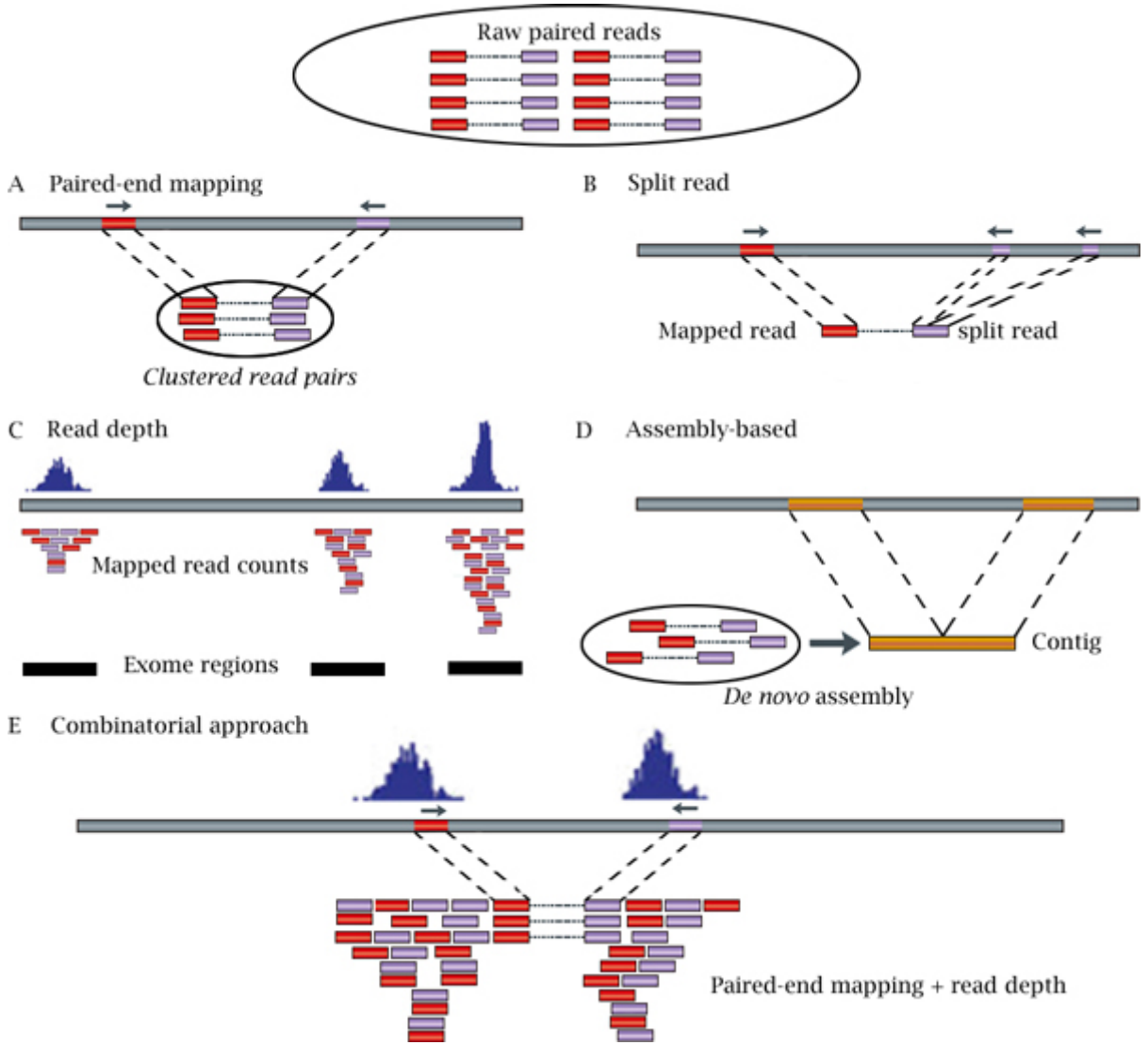
Split read analysis

- 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