Structural Variation Detection and Interpretation

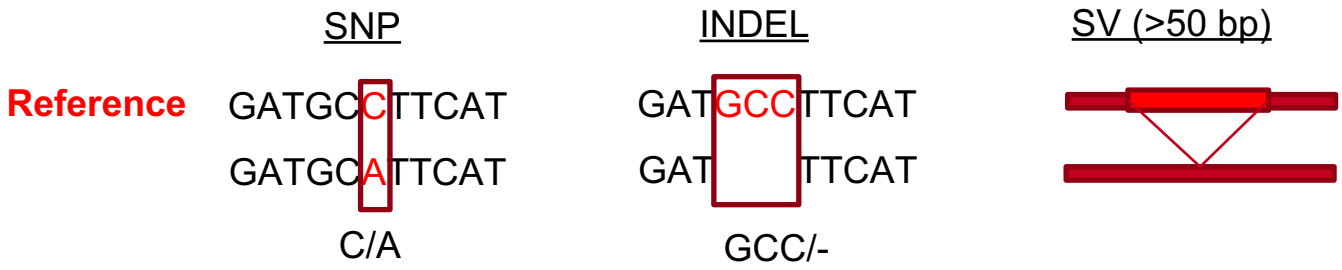
Eugene Gardner
Postdoctoral Fellow
Wellcome Sanger Institute
@DrGeneUK
E: eg15@sanger.ac.uk

Presentation originally developed by:

Thomas Keane
EVA, EGA, ENA archive infrastructure team
EMBL-EBI

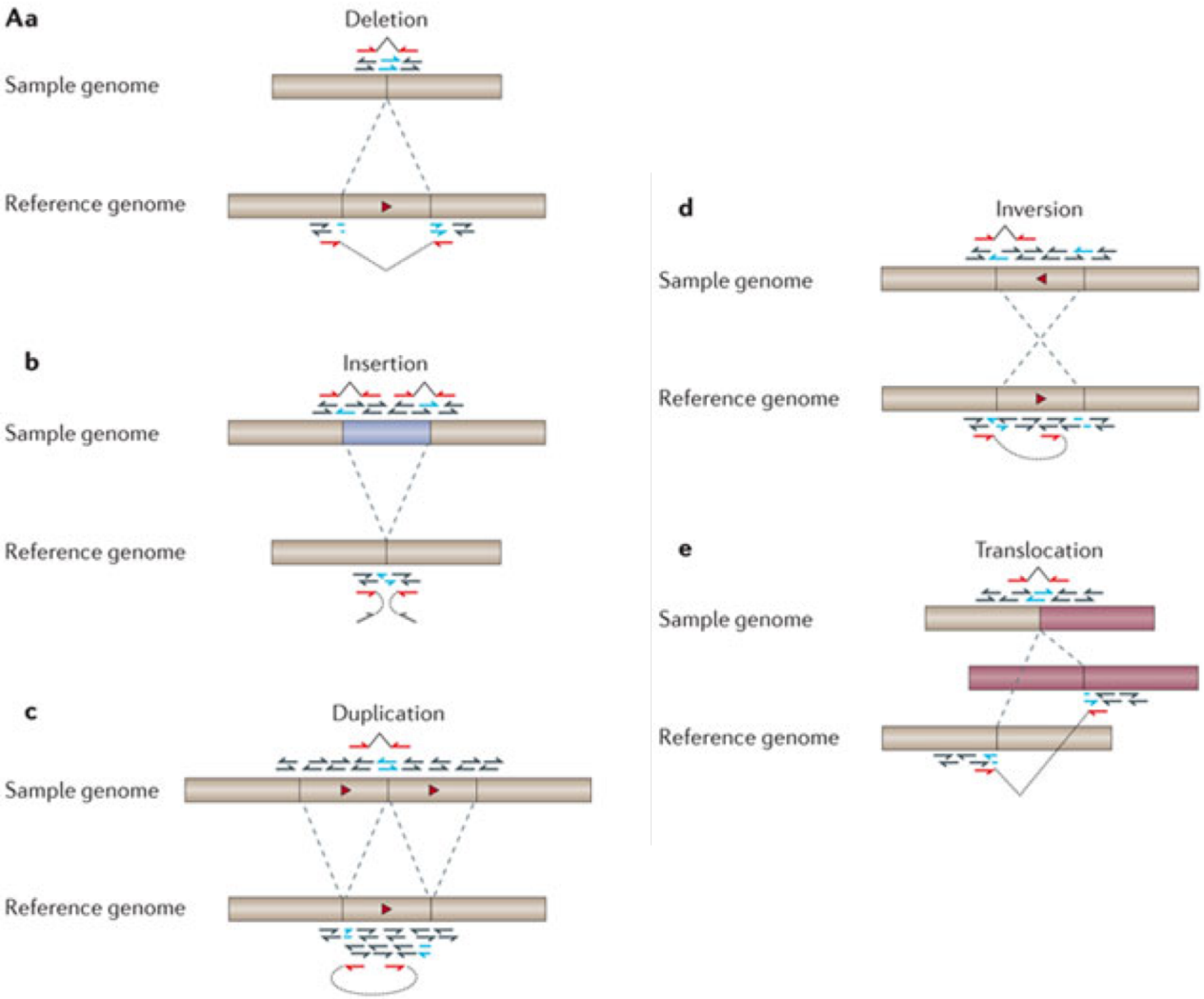
Human Genetic Variation

Variation class	Size	# of polymorphic sites per human
Single nucleotide variant (SNV)	1 bp	3.53M – 4.31M
Small insertions and deletions (InDels)	1 – 50 bp	546k – 625k
Structural Variants (SVs)	>50 bp – 1 Mbp+	2.5k – 3.2k
Copy Number Variants (CNVs)		*4.8k*
Insertions (incl. Mobile Element Insertions)		*2.6k*
Inversions		*14*
Translocations		*<1*



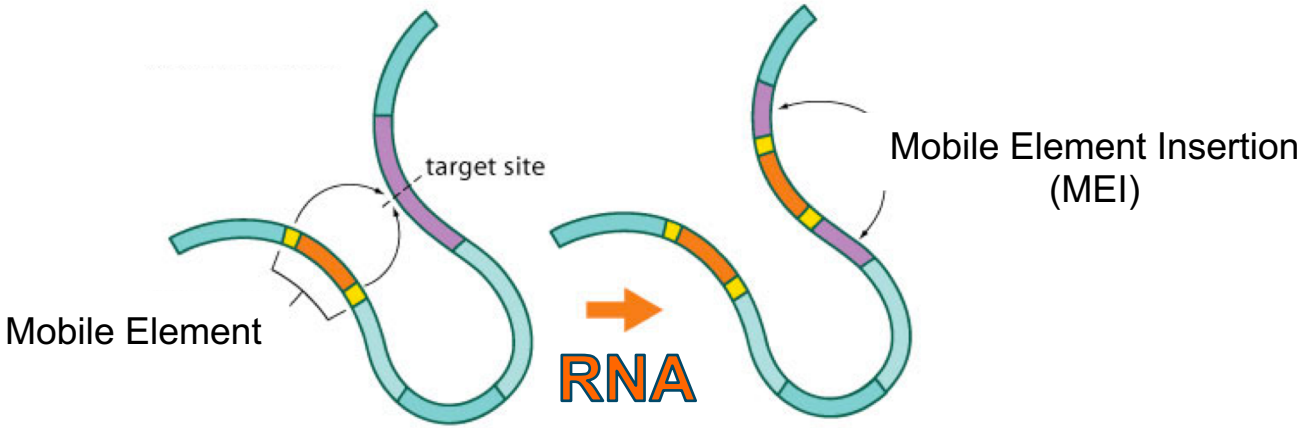
Collins, R. L. *et al.* A structural variant reference for medical and population genetics. (2020). Nature. PMID: 32461652.
1000 Genomes Project Consortium. A global reference for human genetic variation. (2015). *Nature.* PMID: 26432245.

Structural Variant Types



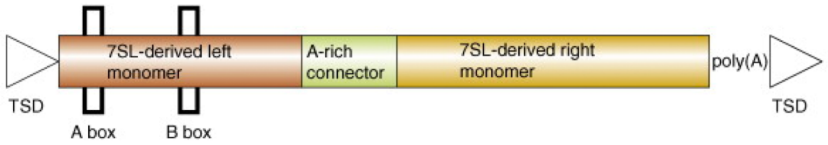
Weischenfeldt J. et al. *Phenotypic impact of genomic structural variation: insights from and for human disease*. (2013). *Nature Reviews Genetics*. PMID: 23329113.

Retrotransposition

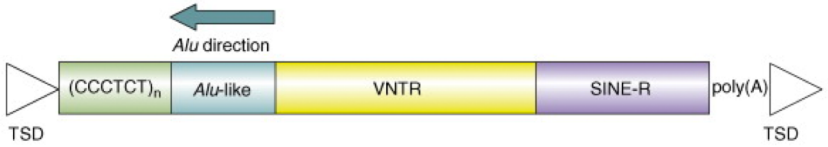


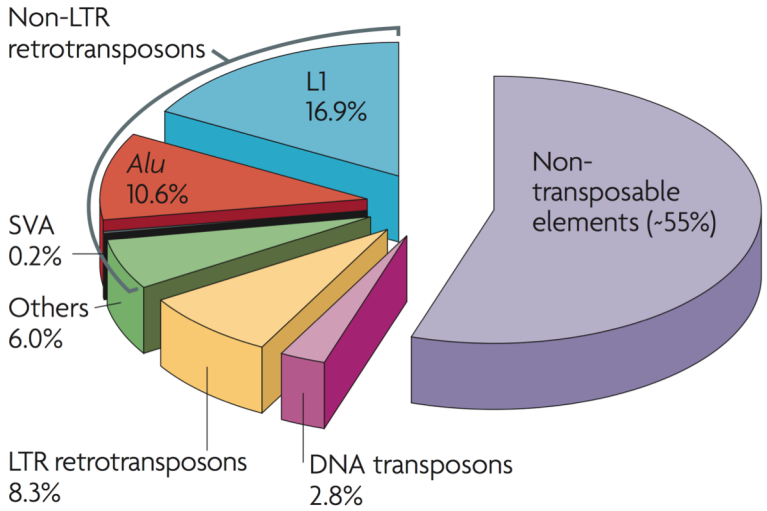
L1-6kbp

Alu-300bp



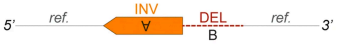
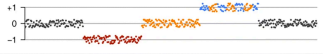
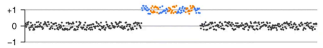
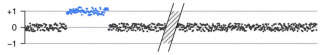
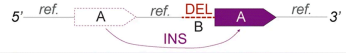
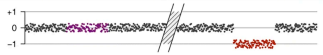
SVA-3kbp





Cordaux, R., & Batzer, M. A. *The impact of retrotransposons on human genome evolution.* (2009). *Nat Rev Genet*. PMID: 19763152.
Mills, R. E., Bennett, E. A., Iskow, R. C., et al. *Which transposable elements are active in the human genome?* (2007). *Trends Genet*. PMID: 17331616.

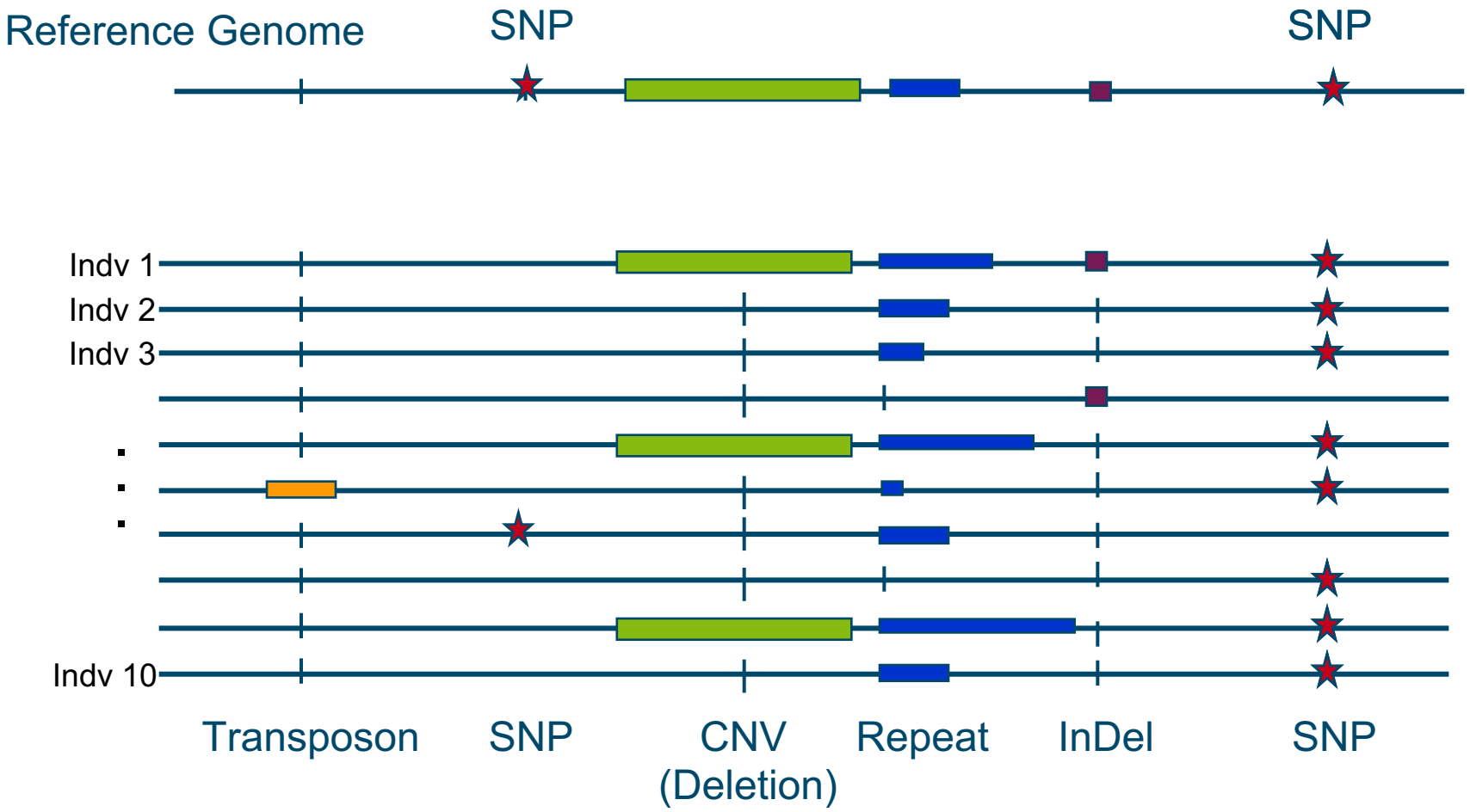
Complex SVs

cxSV Subclass	Abbreviation	SVs	Validation Rate	Total Observations	Subjects with ≥ 1 SV	Median Size (kb)	DEL DUP INS INV	Rearrangement Schematic	Simulated Copy Number Profile
Inversion with 5' Flanking Deletion	delINV	38	100% (19/19)	2,301	99.7%	12.2			
Inversion with 3' Flanking Deletion	INVdel	40	100% (19/19)	2,242	100.0%	9.3			
Paired-Deletion Inversion	delINVdel	58	96% (25/26)	2,288	98.0%	15.2			
Inversion with 5' Flanking Duplication	dupINV	7	75% (3/4)	62	8.7%	54.6			
Inversion with 3' Flanking Duplication	INVdup	3	100% (1/1)	6	0.9%	82.9			
Paired-Duplication Inversion	dupINVdup	45	96% (27/28)	151	19.0%	112.5			
Inversion with 5' Flanking Duplication and 3' Flanking Deletion	dupINVdel	6	100% (2/2)	9	1.3%	27.3			
Inversion with 5' Flanking Deletion and 3' Flanking Duplication	delINVdup	10	100% (5/5)	90	12.2%	67.5			
Inverted Duplication with Flanking Triplication	dupTRIPdup-INV	5	100% (5/5)	5	0.7%	113.9			
Inverted Repeat / Inverted Tandem Duplication	IR	11	88% (7/8)	36	5.1%	73.8			
Compound CNV	cpdCNV	22	100% (17/17)	2,085	99.4%	29.1		Various	
Dispersed Duplication	dDUP	10	100% (2/2)	42	6.1%	17.5			
Dispersed Duplication with Deletion	dDUPdel	9	100% (4/4)	60	8.5%	32.1			
Insertion with Deletion	INSdel	4	100% (1/1)	12	1.7%	5.9			
Compound Insertion	cpdINS	5	100% (2/2)	251	36.0%	3.1		Various	
Compound Insertion with Deletion	cpdINSdel	1	NA (0/0)	5	0.7%	9.6		Various	
Compound/Complex Rearrangement (or Other)	CCR	15	100% (11/11)	21	3.1%	239.8		Various	
All cxSV	-	289	97% (150/154)	9,666 (14/subject)	100.0%	27.3	DEL: 61.8% DUP: 47.8% INV: 84.8% INS: 11.8%		

Collins R. L. et al. *Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome.* (2017). *Genome Biology*. PMID: 28260531

Organisms Are Variable in Their Genetic Variation

Organisms Are Variable in Their Genetic Variation



Methods for detecting SVs

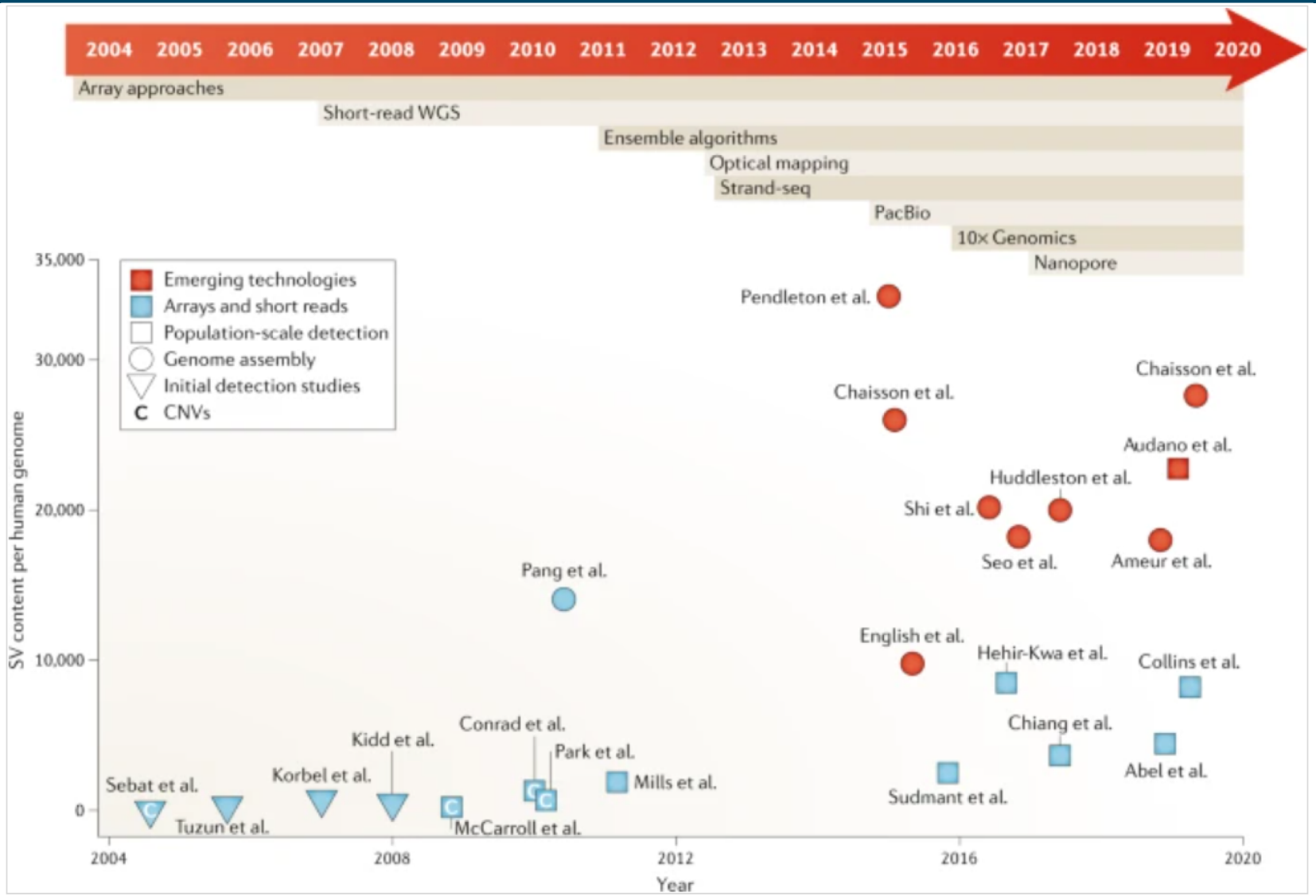
- **Chromosome banding:** chromosomes are prepared from dividing cells, stained, and viewed with a microscope. Large deletions, duplications, and translocations are detected if the banding pattern or chromosome structure is altered.

- **Fluorescence in situ hybridization (FISH)**: fluorescent-labeled DNA probes hybridize to metaphase or interphase cells to visualize a locus on a chromosome and determine copy number. FISH can determine the location of chromosomal segments identified by microarray, NGS, and WGS.

- **Microarray**: array comparative genome hybridisation (array CGH) detects copy-number differences between abnormal and reference genomes. SNP arrays detect changes in copy-number and allelic ratios. CNV location and SV organization are not determined by microarray methods.

- **Whole Genome Sequencing (WGS)**: The use of massively parallel, reversible terminated chemistry sequencing of the entire genome of a single individual. Today, two approaches are generally used for SV discovery
  - **short read WGS**: Breakpoints of SVs are detectable by short (typically ≤250 bp) paired-end reads that have discordant mappings to the reference genome.
  - **long read WGS**: sequencing long molecules of DNA (several kbp) and subsequent alignment to a reference genome to detect SVs

New Technologies Allow Identification of More SVs

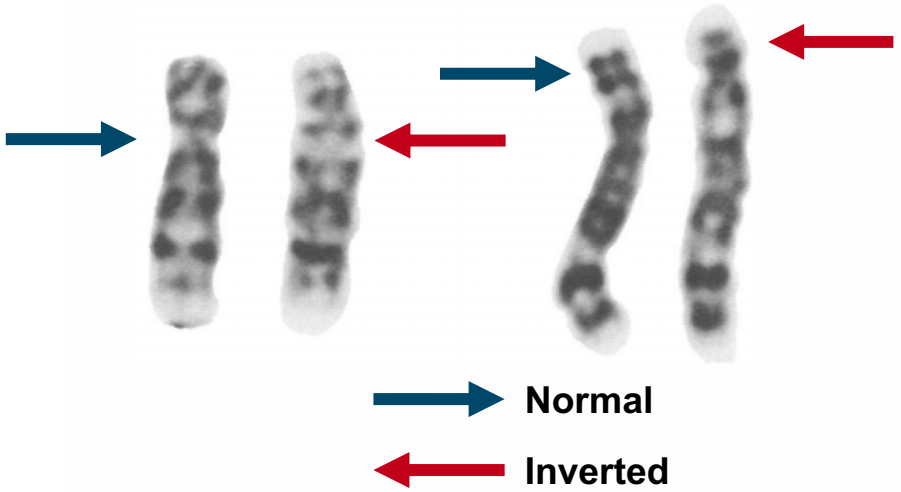


Ho S. S., Urban A. E., and Mills R.E. *Structural variation in the sequencing era.* (2020). *Nature Reviews Genetics*. PMID: 31729472

Chromosome Banding

The first structural variants were visible through a microscope!

It's hard to see, but notice the banding between the left and right chromosomes don't match!



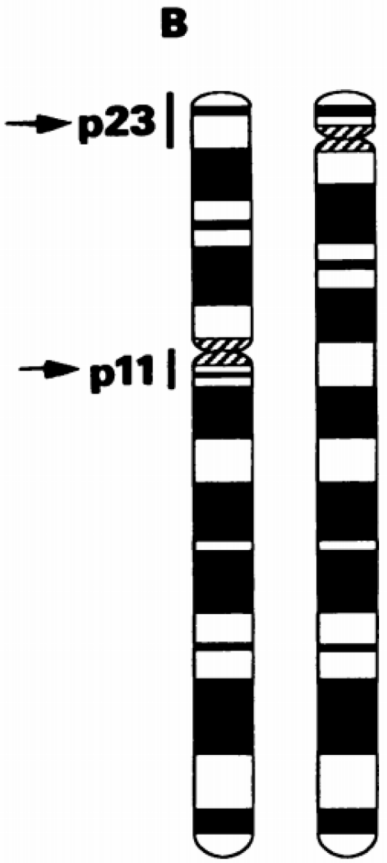
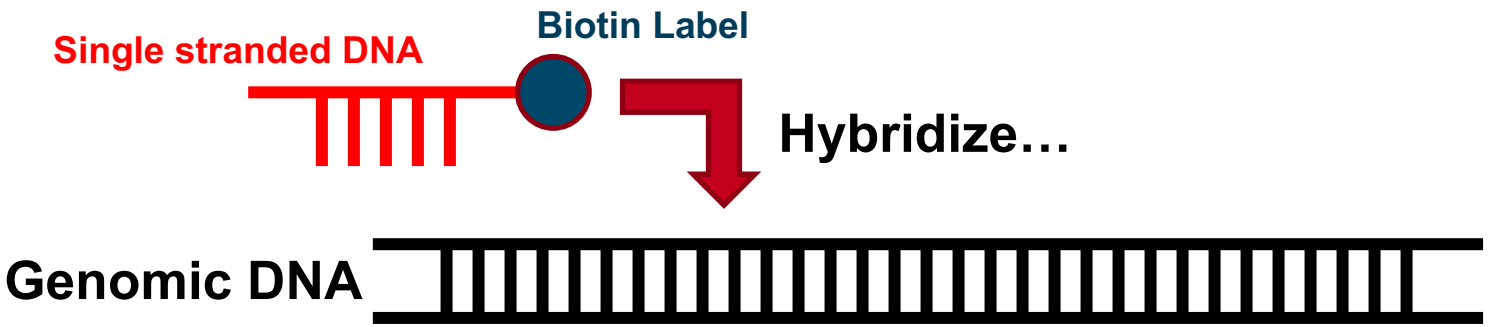


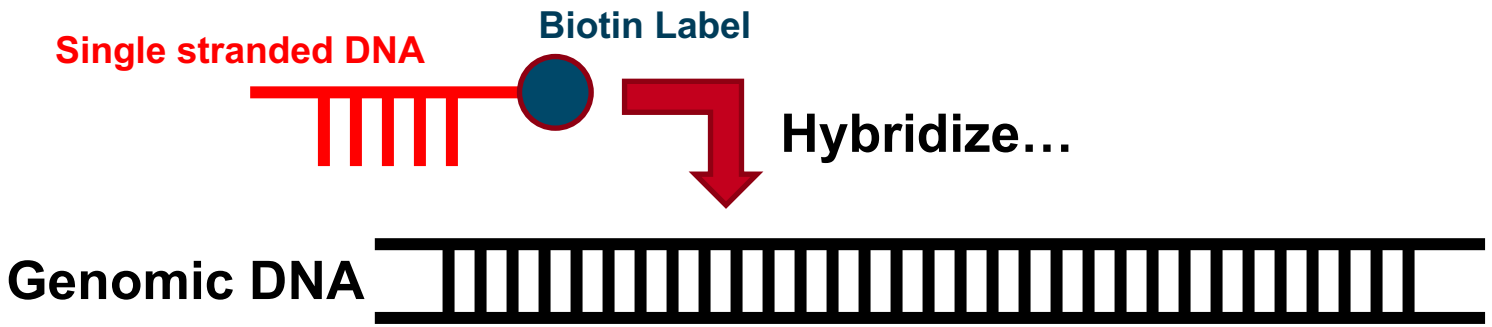
Figure 2 (A) G banded chromosomes 8 from two carriers of the inv(8)(p23q11). The inversion chromosomes are to the right in each pair. (B) Schematic presentation of inv(8)(p23q11). Arrows indicate the breakpoints in 8p23 and 8q11.

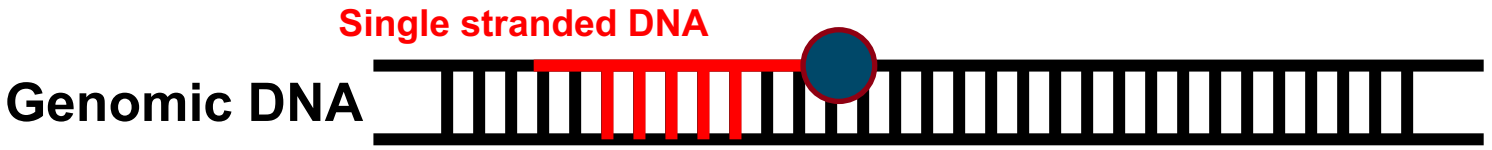
Boyd H. et al. *Familial pericentric inversion inv(8)(p23q11).* (1994). Journal of Medical Genetics. PMID: 8014967

Fiber Flourescent *in* situ Hybridization (FISH)

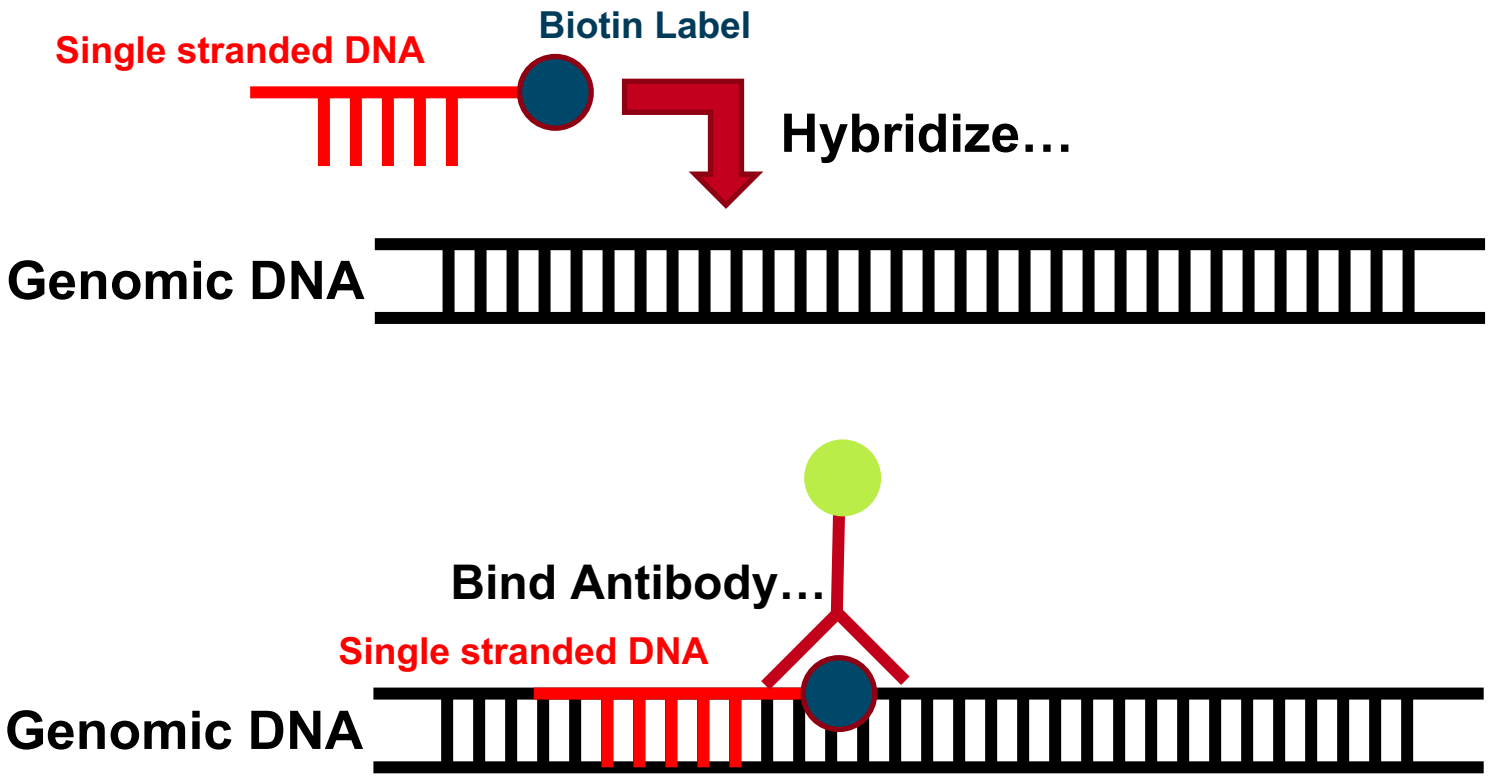


Fiber FISH

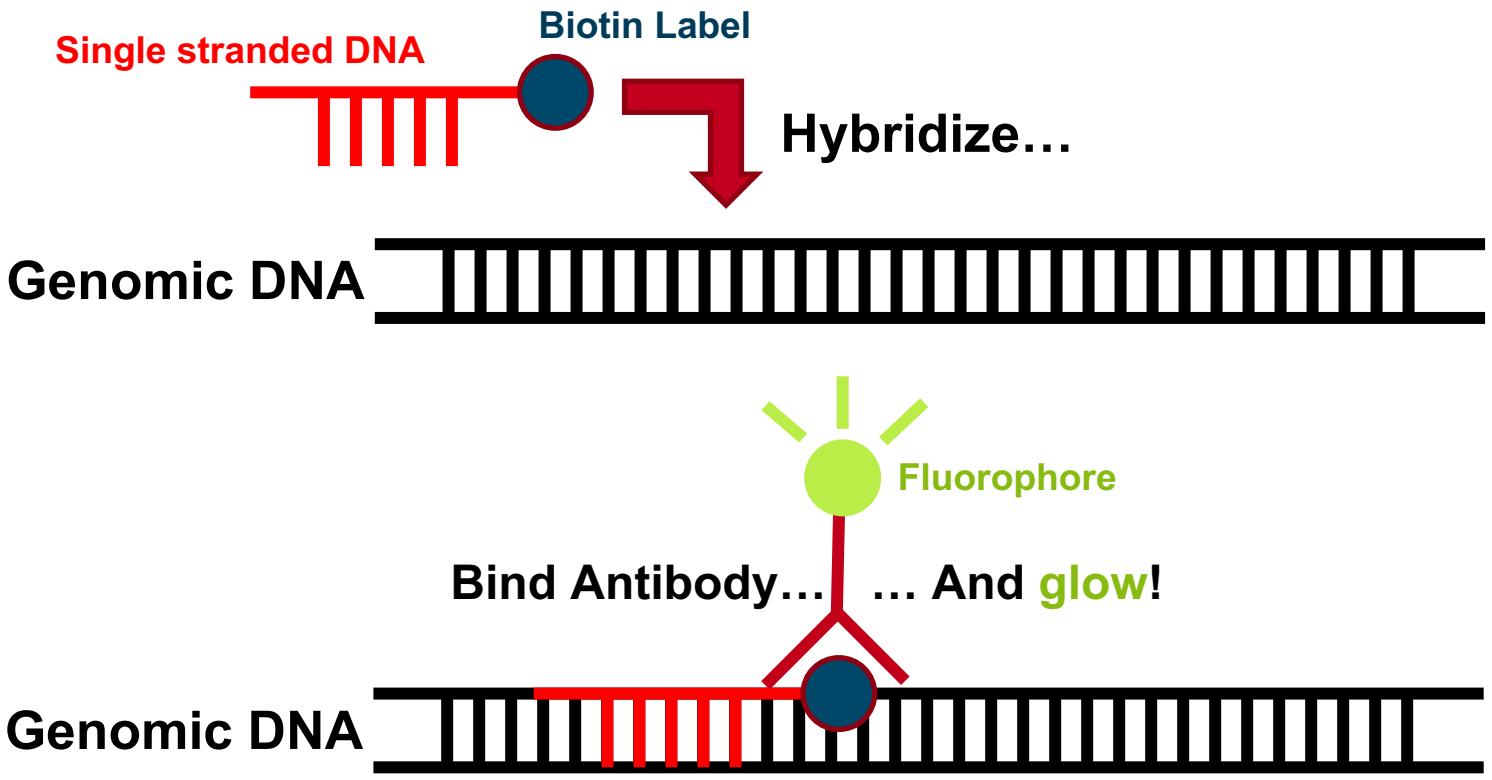

Single stranded DNA
Biotin Label
Hybridize…
Genomic DNA


Single stranded DNA
Genomic DNA

Fiber FISH


Single stranded DNA
Biotin Label
Hybridize…
Genomic DNA
Bind Antibody…
Single stranded DNA
Genomic DNA

Fiber FISH



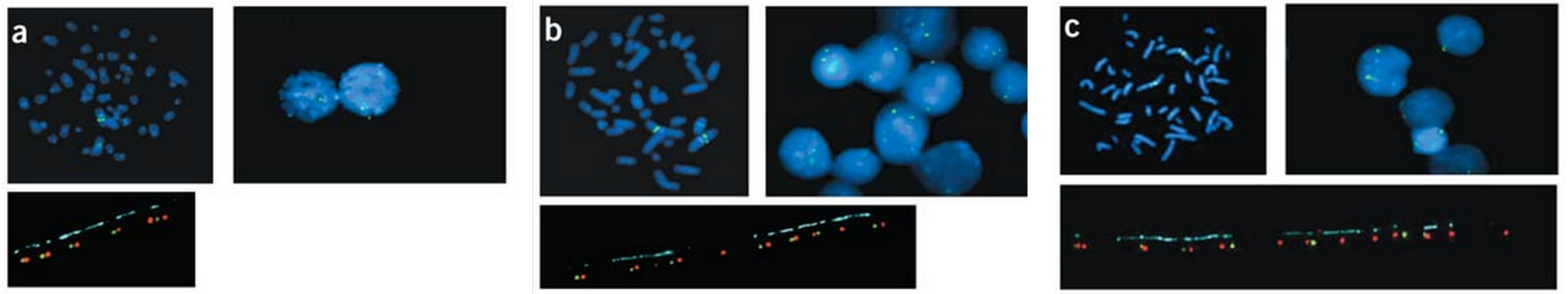
Please note: this is a simple cartoon example. The reality is more complex!

Fiber FISH

Can we identify regions in the human genome that have more/less copies in different individuals using fiber FISH?

Yes! One example is at the Amylase (*AMY1*) gene locus in humans.

Fewer Copies **Average Copies** **More Copies**

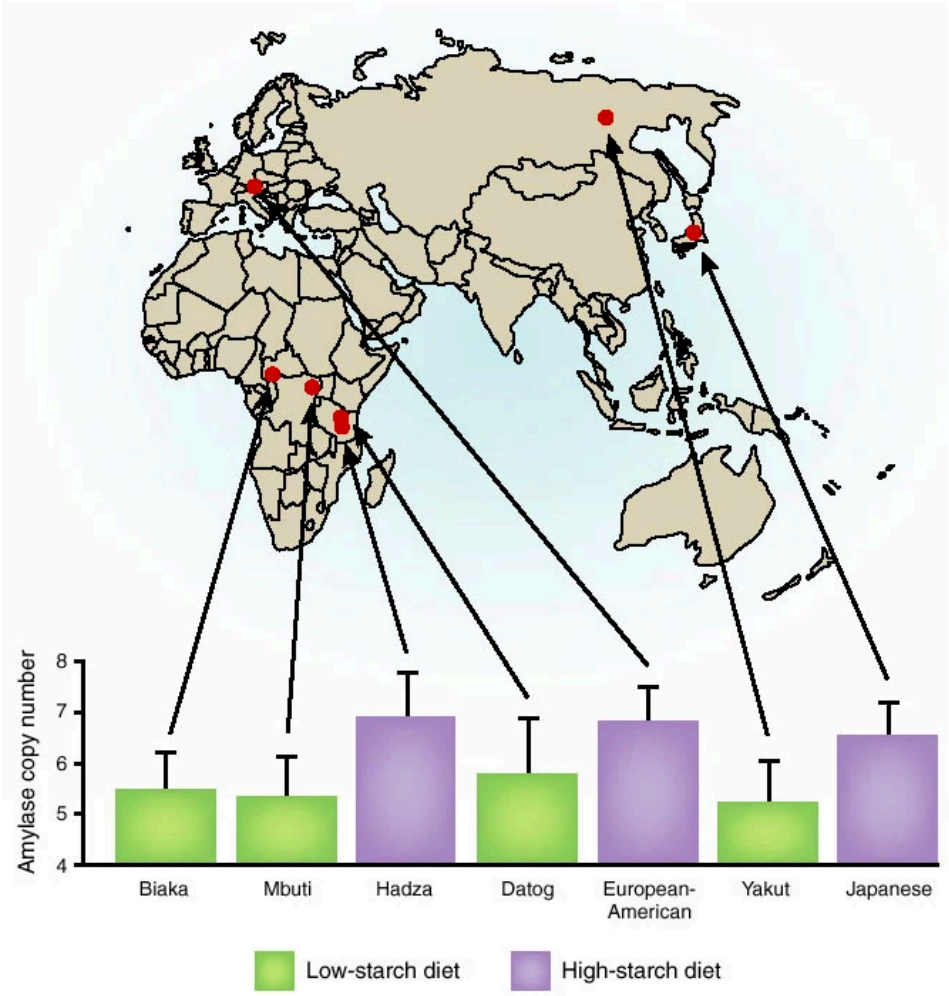


If we used SV types we just learned…

Deletion **Reference** **Duplication**

Iafrate A. J. et al. *Detection of large-scale variation in the human genome*. (2004). *Nature Genetics*. PMID: 15286789

Fiber FISH



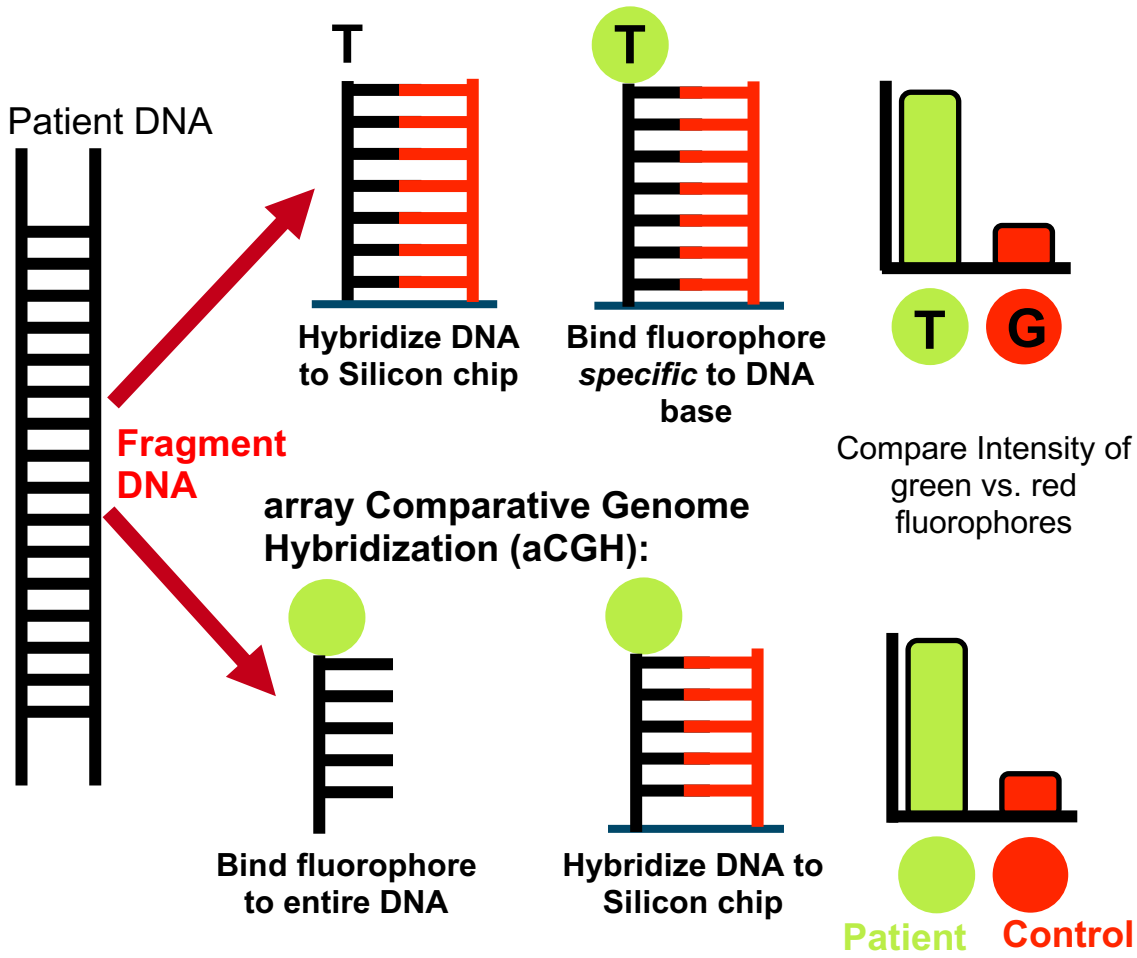
The copy number of amylase varies based on population, and likely depends on the amount of starch (plant material) that a given population includes in their diet.

More Starch = Need More Amylase!

Perry G. H. et al. *Diet and the evolution of human amylase gene copy number variation.* (2007). *Nature Genetics*. PMID: 17828263
Novembre J., Pritchard J. K., Coop G. *Adaptive drool in the gene pool*. (2007). *Nature Genetics*. PMID: 17898775

Microarrays

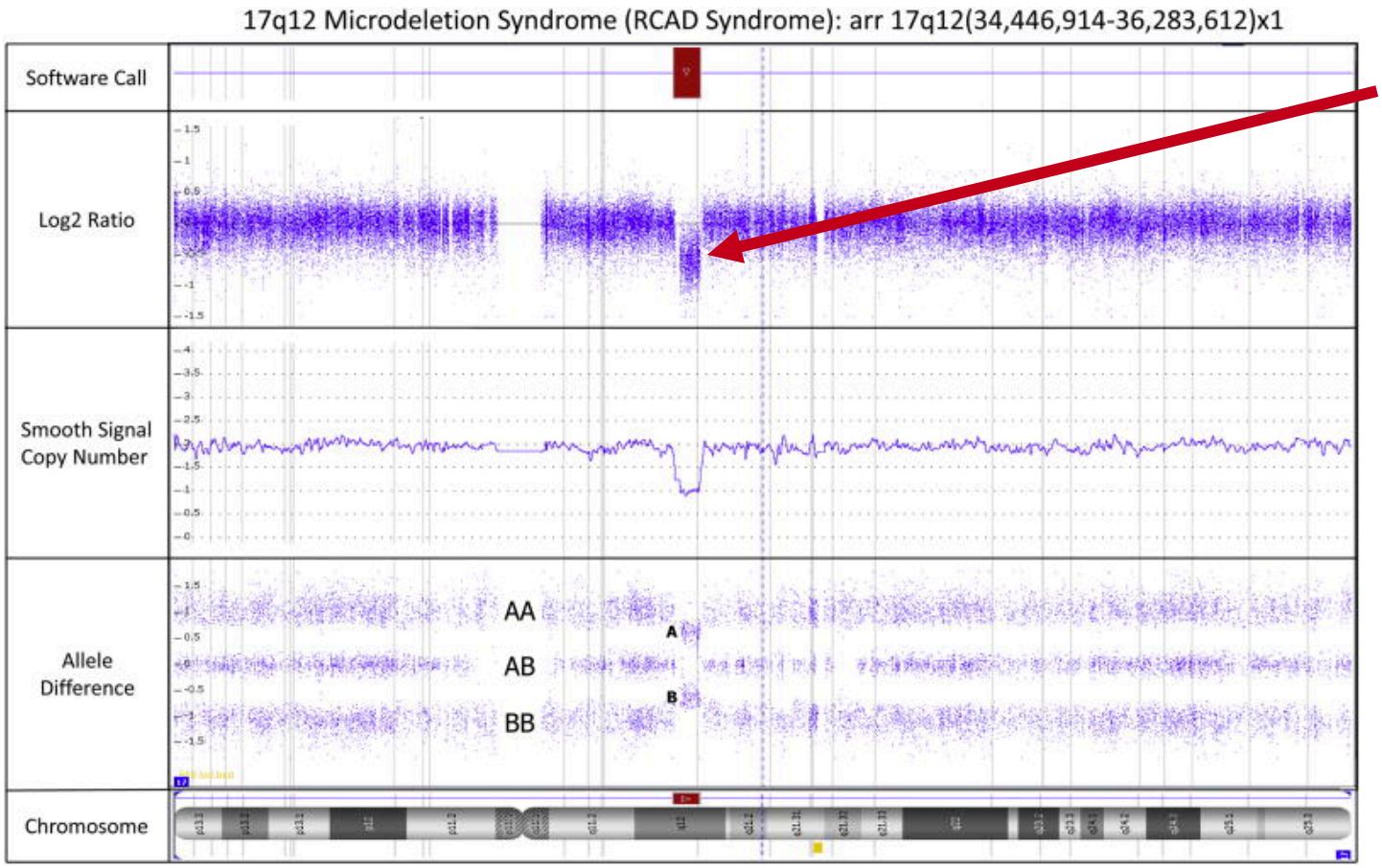
SNP Microarray:



Cons	Pros
• Low density – can only identify very large events (>100kbp) • Can be difficult to quality control • Can only find copy gain / loss (CNVs), no other SV types	• Relatively low cost –can analyze thousands of individuals at a time • Can analyze SNVs and CNVs at the same time
• Cannot identify SNVs (specialized) • Can only find copy gain / loss (CNVs), no other SV types	• Can identify events invisible to Microarrays • Relatively low cost – can analyze thousands of individuals at a time

Both techniques hybridize DNA to a chip (silicon/glass/plastic/etc) and look for differences in fluorescence. They differ in how the fluorophore is bound…

Microarray & aCGH



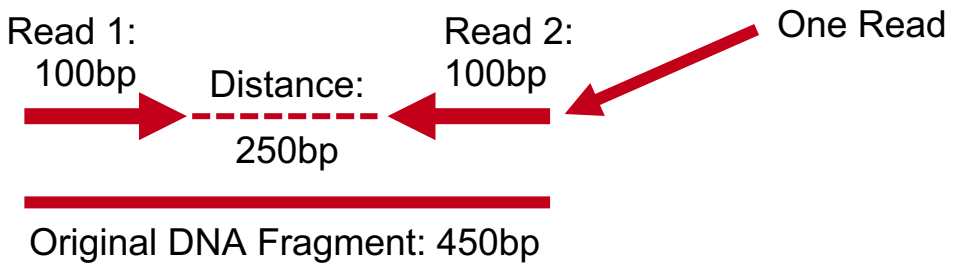
- Still used today to analyze large numbers of individuals for large Structural Variants (>100kbp) for relatively low cost:
  - UK Biobank has assessed 500k individuals with this approach. More than any other study!

Levy V. et al. *Prenatal diagnosis by chromosomal microarray analysis.* (2018). *Fertility and Sterility*. PMID: 29447663

Short Read Sequencing for SV Detection

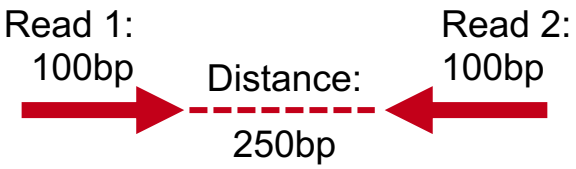
You learnt in the last module to identify SNVs using short read sequencing data:

You may also remember that these reads are sequenced in PAIRS…

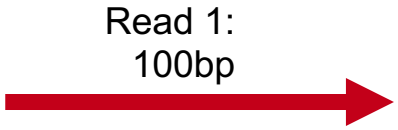


Types of Structural Variant Evidence from Short Read Data

Pair Orientation / Distance:

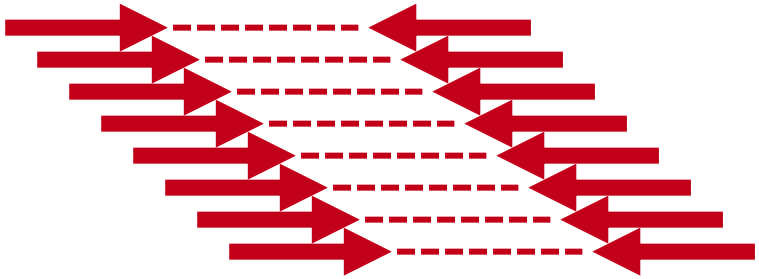


Split Reads:



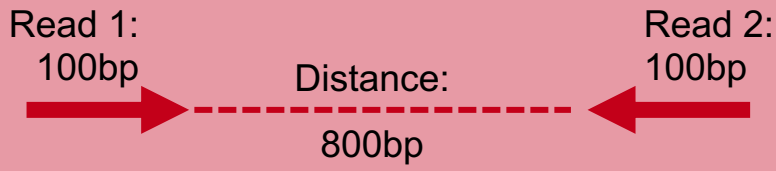
Entire read aligns to reference genome

Read Depth:



8 read pairs align to same location

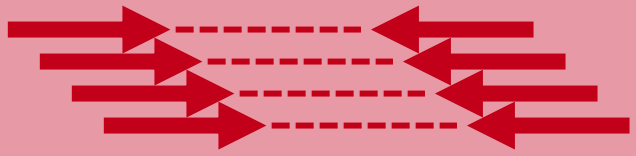
Possible SV?



Reads align further/closer together

Reads align in wrong direction

Only red part aligns to reference genome

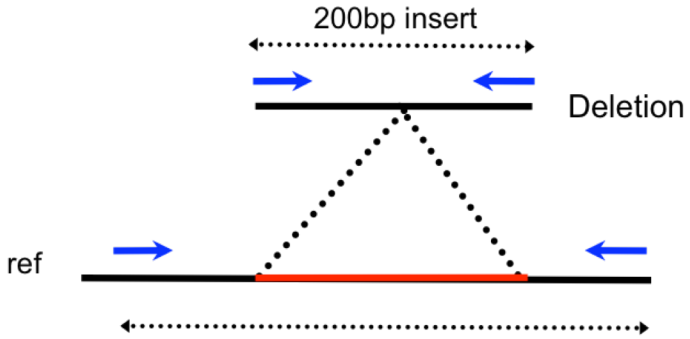


4 read pairs align to same location

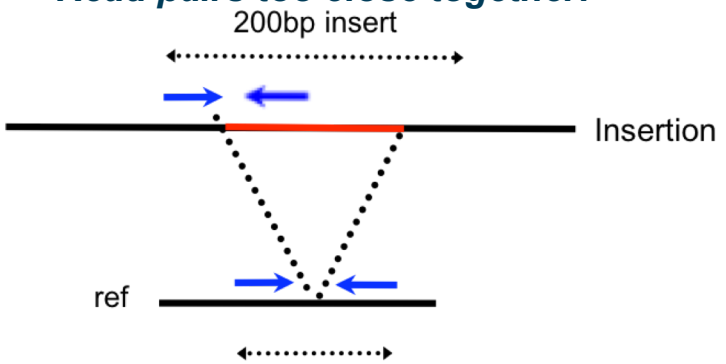
SV Types and Read Pair Information

We use how read pairs are oriented when they align to the reference to identify Structural Variants:

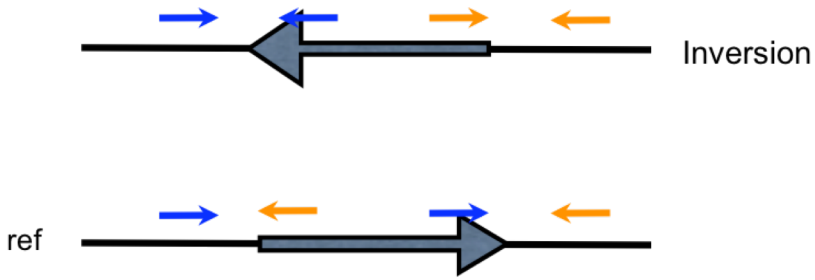
Read pairs too far apart:



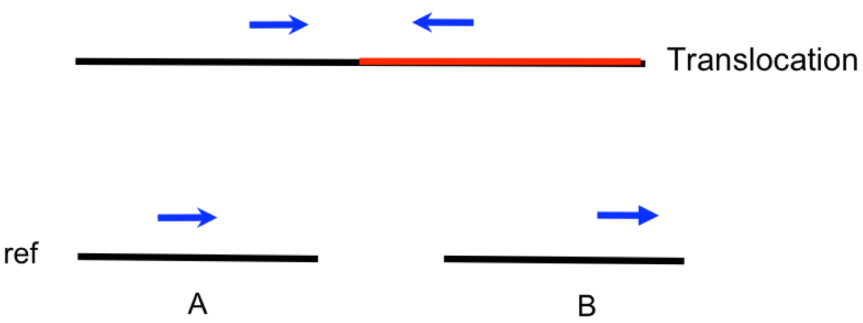
Read pairs too close together:



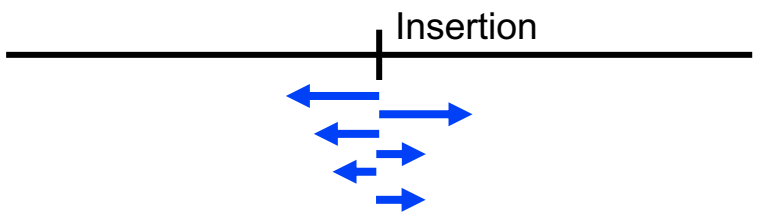
Read pairs align next in the same direction:



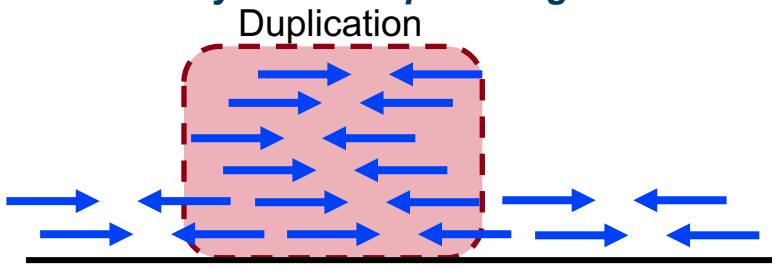
Read pairs align to different chromosomes:



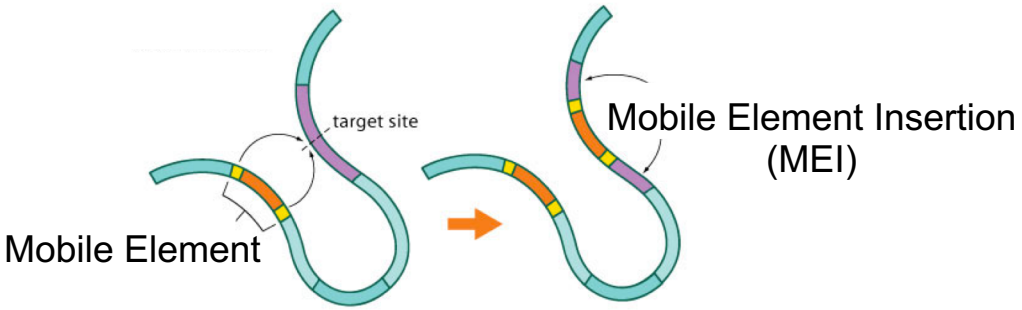
Reads align only partially to the reference:

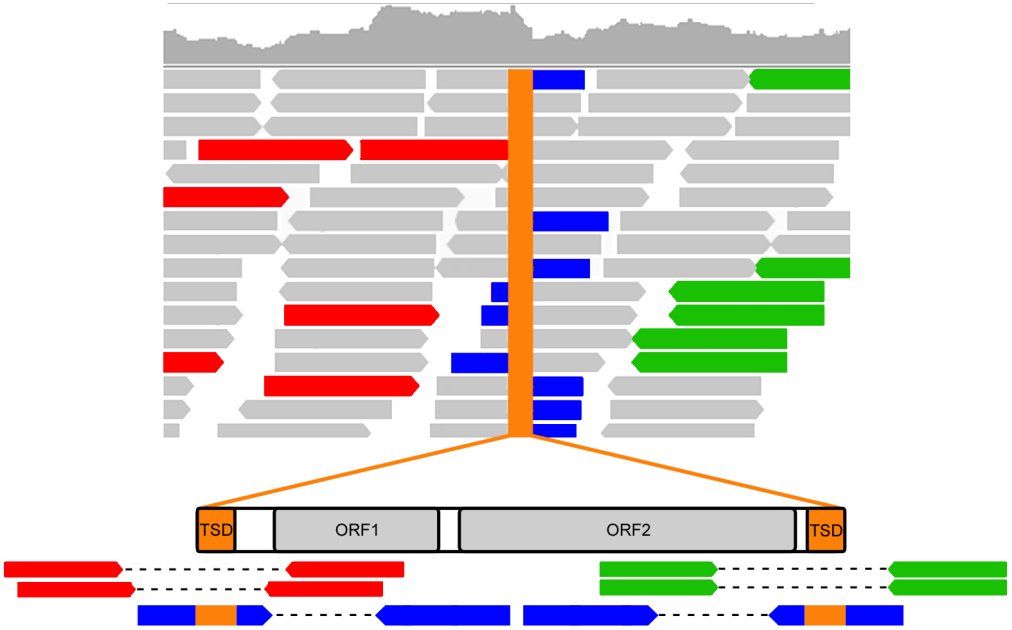


Too many/few read pairs align:



Pair Orientation / Distance – More on non-reference retrotransposition events





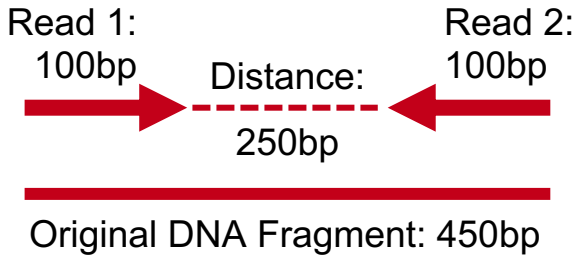
- Retrotransposons leave a unique read pair signature during alignment
- Some read pairs align to the reference and the insertion (green/red)

Gardner, E. J., et al. *The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology.* (2017). *Genome Research*. PMID: 28855259

Pair Orientation / Distance – Fragment Size Distribution

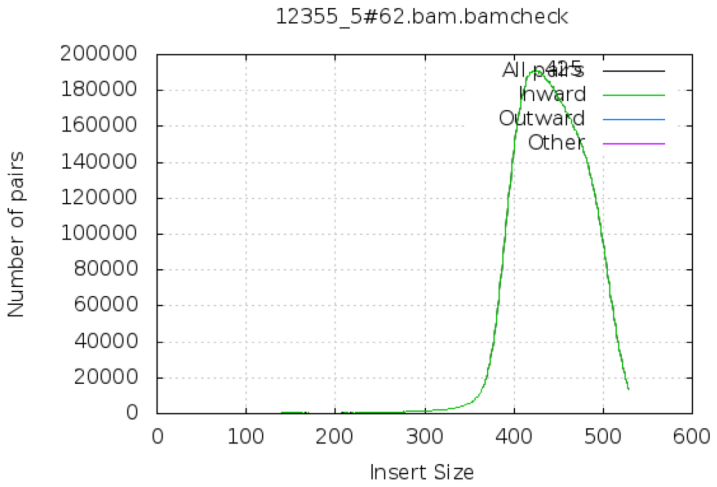
Several types of structural variations (SVs)

- Large Insertions/deletions
- Inversions
- Translocations



Read pair information used to detect these events

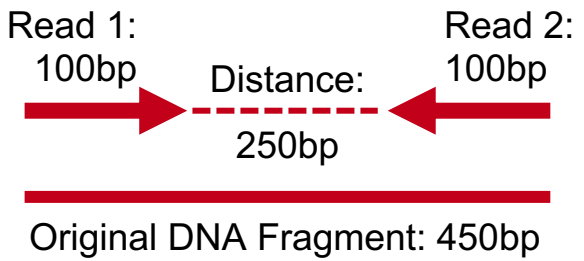
- Paired end sequencing of either end of DNA fragment
- Observe deviations from the expected fragment size
- Presence/absence of read pairs



Pair Orientation / Distance – Fragment Size Distribution

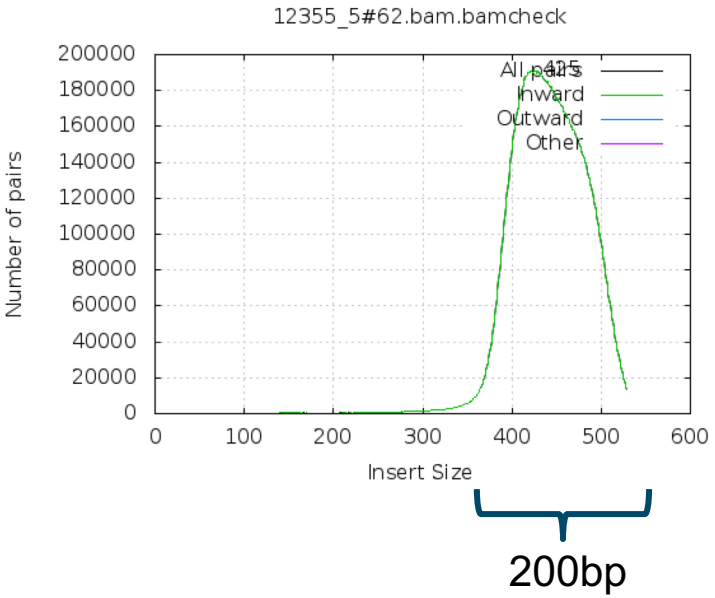
Several types of structural variations (SVs)

- Large Insertions/deletions
- Inversions
- Translocations

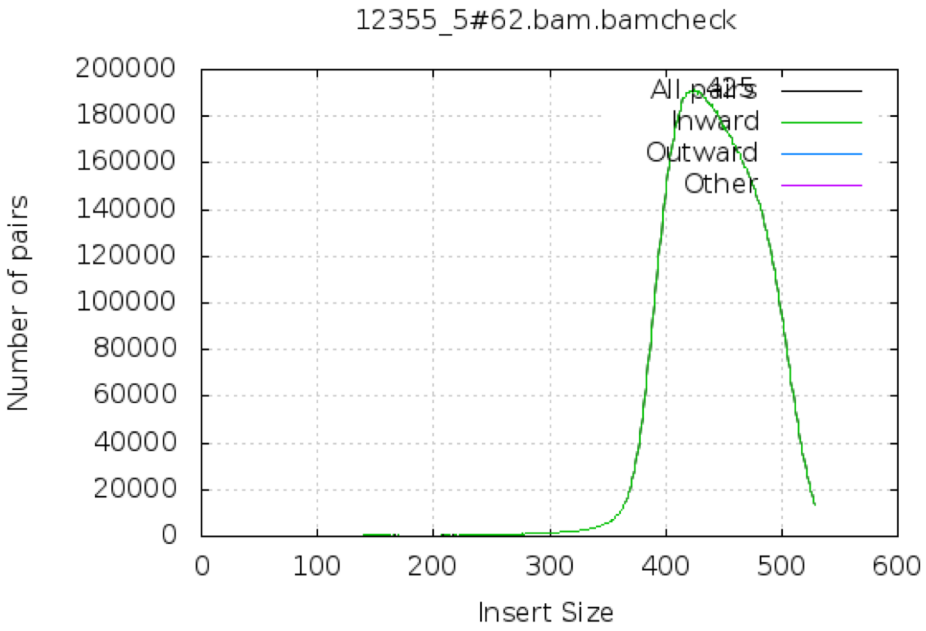


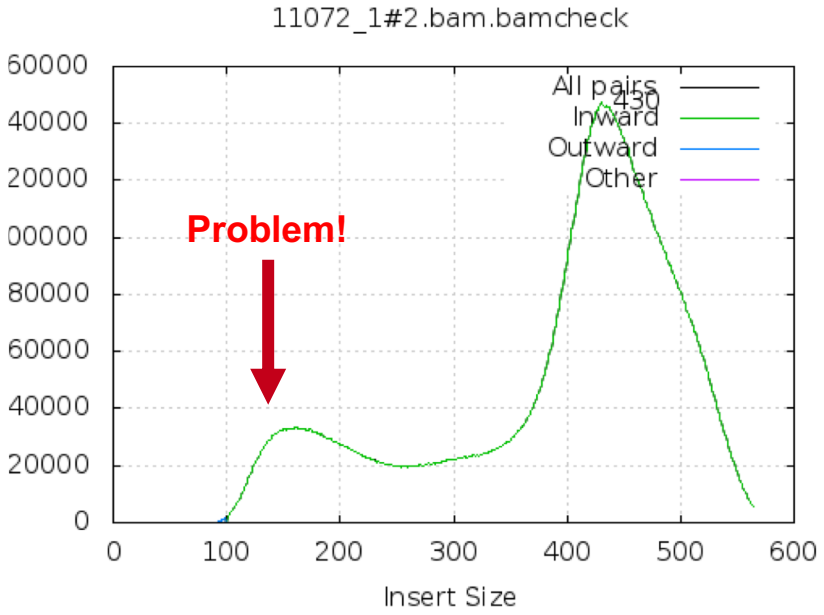
Read pair information used to detect these events

- Paired end sequencing of either end of DNA fragment
- Observe deviations from the expected fragment size
- Presence/absence of read pairs

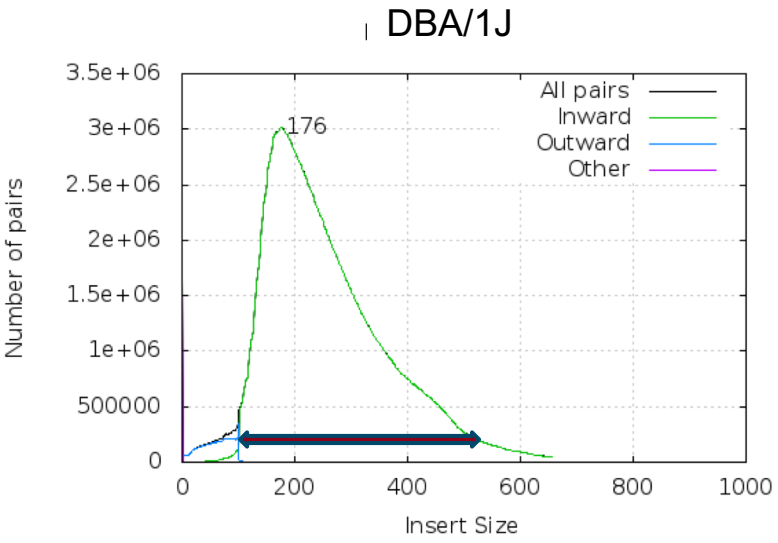


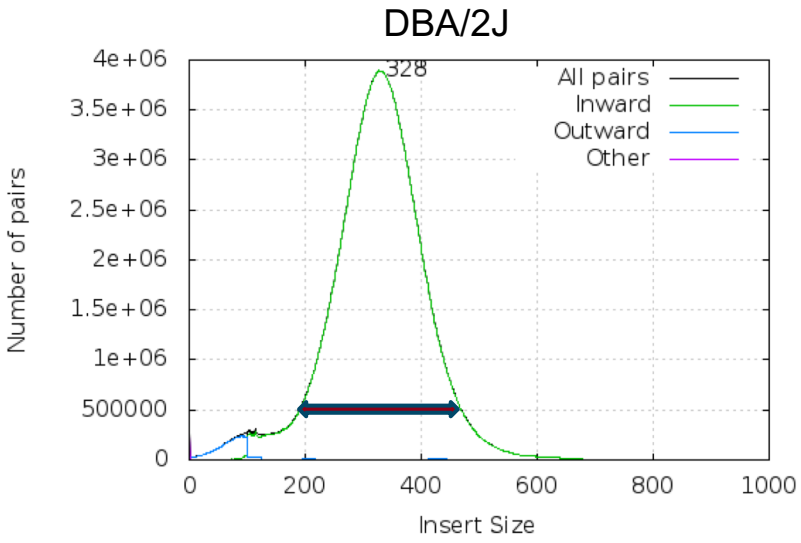
Pair Orientation / Distance – Fragment Size QC





Pair Orientation / Distance – Fragment Size Distribution





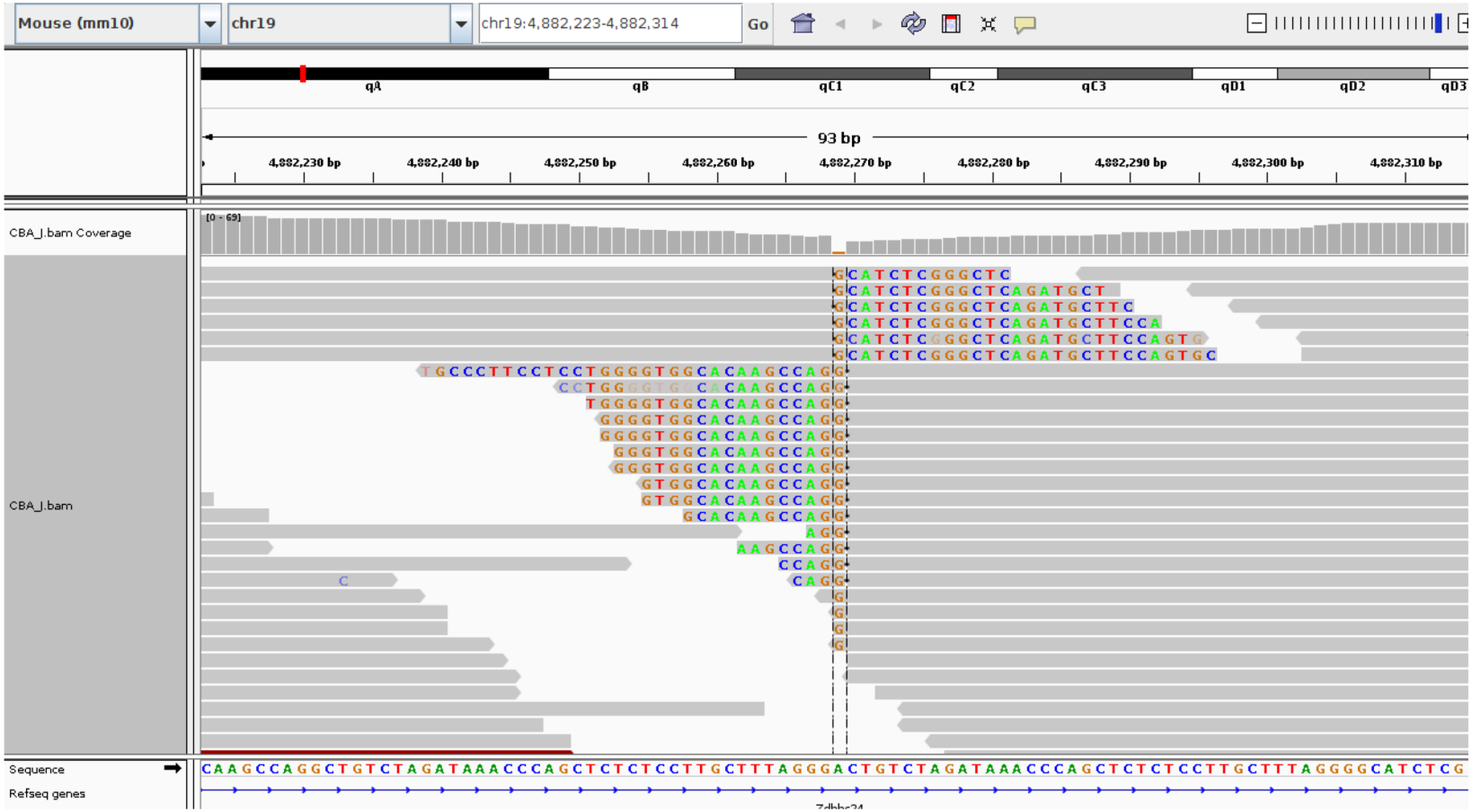
DBA/1J fragment size distribution has larger range (~450bp) vs. DBA/2J (~250bp)

SV callers only consider read pairs discordant if they fall outside of the extremes of the fragment size distribution

- Observed in DBA/1J that we had lower sensitivity to call SVs in the 300-500bp range compared to DBA/2J

Split Reads

- A split-read alignment is a single DNA fragment that spans a breakpoint and therefore does not contiguously align to the reference genome
- Errors in the sequencing and alignment processes creates some ambiguity in the exact location of the breakpoint associated with a split-read alignment



Right click - view mismatch bases

View - preferences - alignments - show soft clipped bases

Read Depth

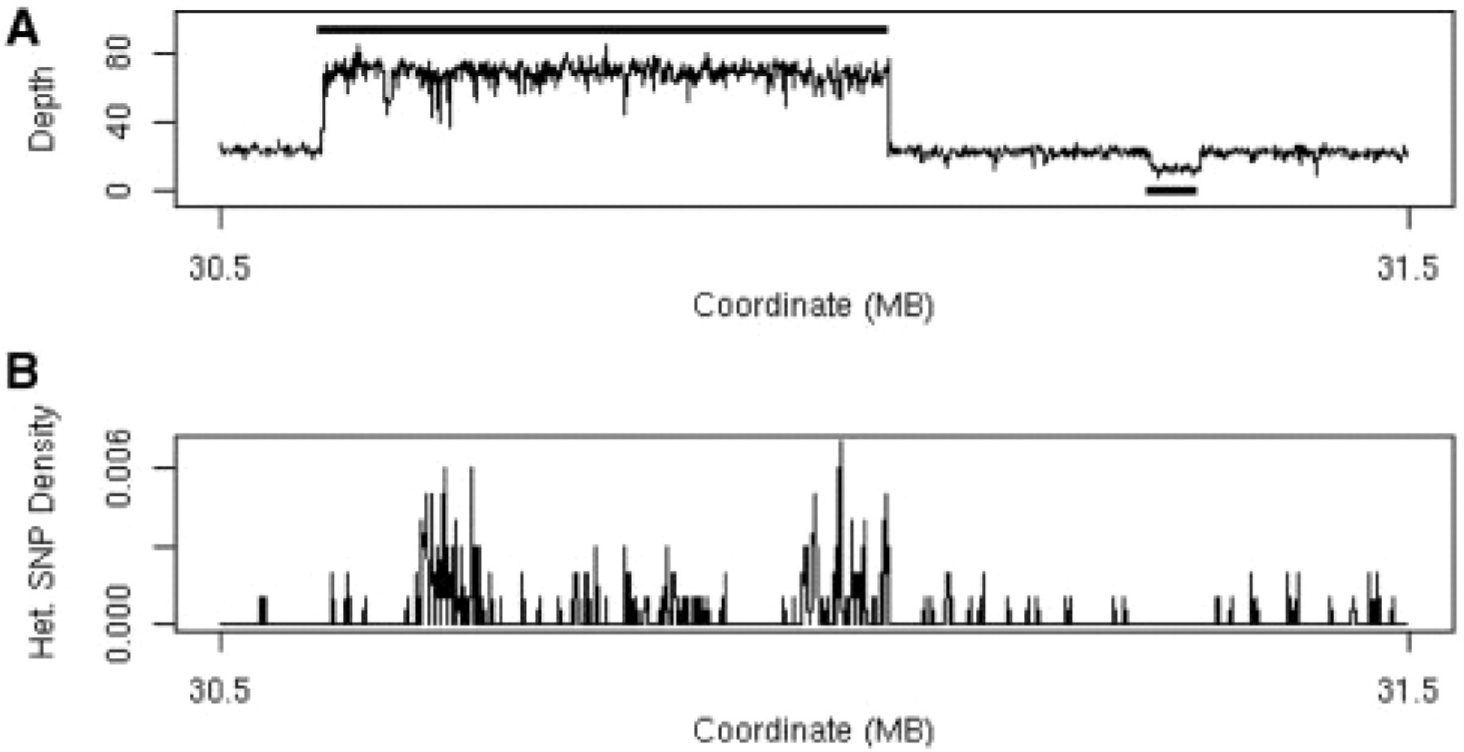


Fig. 1. (A) Plot of sequencing depth across a one megabase region of A/J chromosome 17 clearly shows both a region of 3-fold increased copy number (30.6–31.1 Mb) and a region of decreased copy number (at 31.3 Mb). The solid black line above the depth plot indicates the called copy number gain and the solid black line below the plot indicates the called copy number loss. (B) Plot of the heterozygous SNP rate for the same region showing the high number of apparent heterozygous SNPs associated with the copy number gain.

Simpson J. T. et al. *Copy number variant detection in inbred strains from short read sequence data.* (2009). *Bioinformatics.* PMID: 20022973

Identifying Structural Variants from Short Read Data

- A number of tools have been developed that can find SVs from short read sequencing data

- They use any combination of the three approaches we just discussed:
  - Pair Orientation / Distance
  - Split Reads
  - Depth

- We are now going to talk about two specific callers that you will have a chance to use yourself during the practical exercise:
  - BreakDancer: uses **Pair Orientation / Distance**
  - LUMPY: uses **Pair Orientation / Distance**, **Split Reads**, and **Depth**

- Please note! There are many additional tools that have been developed to identify SVs, we are just looking at two today.

BreakDancer: Read pairs

- Identifies deletions, insertions, inversions and intrachromosomal and interchromosomal translocations
- Input: BAM file
- Algorithm:
  - Analyse a subset of reads from each sequencing library (determine mean and standard deviation of fragment size)
  - Walk along each chromosome to identify all of the anomalous read pairs (a). Identify interconnected clusters.
  - Assign anomalous clusters into categories (b)
- Output
  - Text with one SV event per line
  - Filter by: minimum number of reads, quality score, type of SV

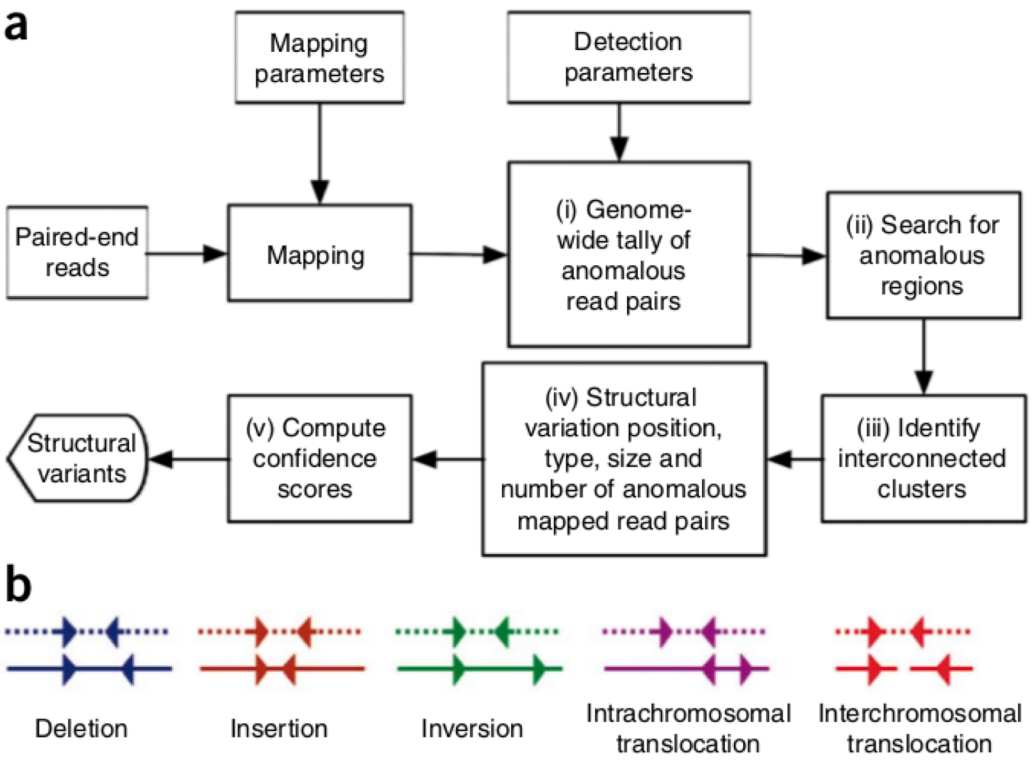
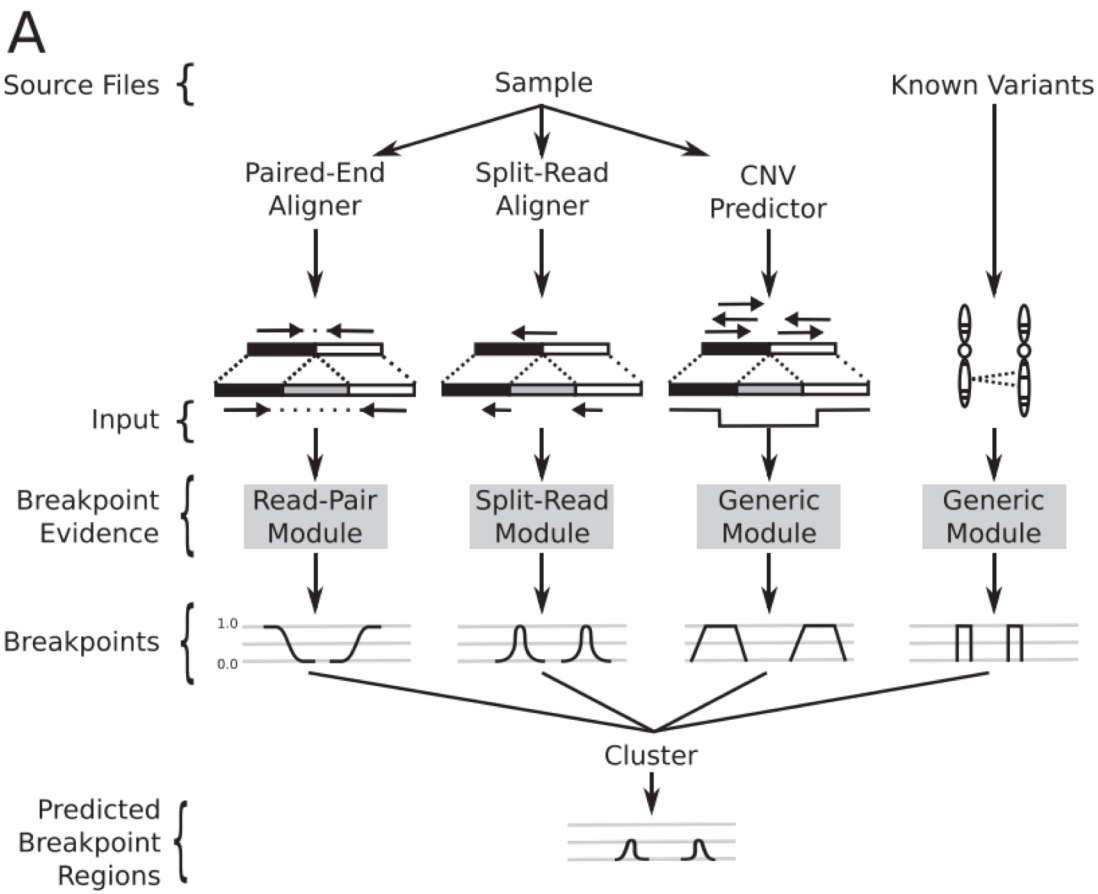


Figure 1 | Overview of BreakDancer algorithm. (**a**) The workflow. (**b**) Anomalous read pairs recognized by BreakDancerMax. A pair of arrows represents the location and the orientation of a read pair. A dotted line represents a chromosome in the analyzed genome. A solid line represents a chromosome in the reference genome.

Chen K *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. (2009) *Nature methods.* PMID: 19668202

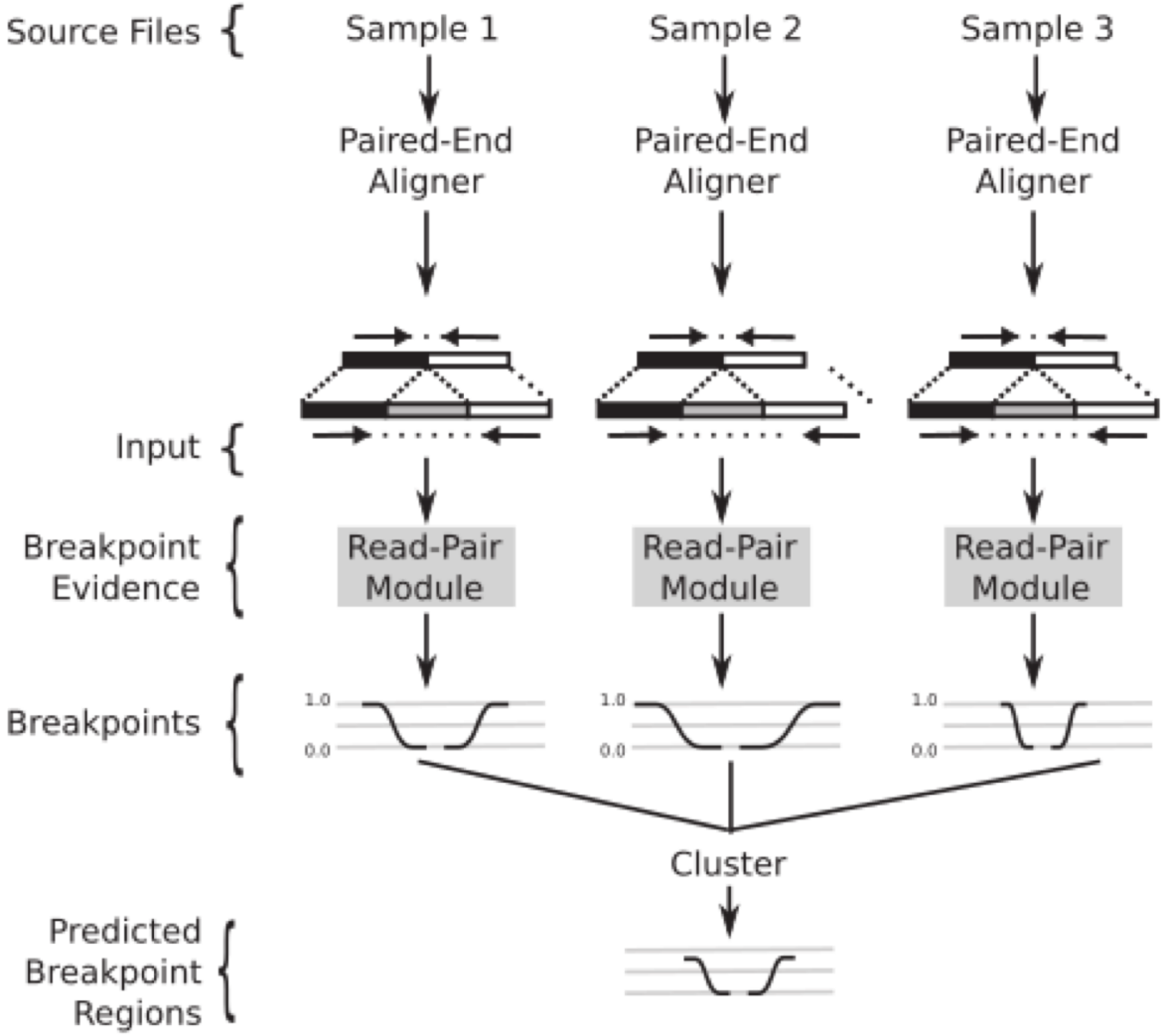
LUMPY: Read pairs + split reads + depth

- Any number of alignment signals can be integrated into a single discovery process
  - Read pairs, split-reads, depth, user supplied evidence
- Distinct modules that map signals from each alignment evidence type to the common probability interval pair.
- Evidence from the different alignment signals is mapped to breakpoint intervals, overlapping intervals are clustered and the probabilities are integrated



Layer R. M. *et al. LUMPY: a probabilistic framework for structural variant discovery.* (2014). *Genome Biology*. PMID: 24970577

LUMPY: multi-signal and multi-sample workflows



Layer R. M. *et al. LUMPY: a probabilistic framework for structural variant discovery.* (2014). *Genome Biology.* PMID: 24970577

VCF for SVs

```
##fileformat=VCFv4.1
##fileDate=20100501
##reference=1000GenomesPilot-NCBI36
##assembly=ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/sv/breakpoint_assemblies.fasta
##INFO=<ID=BKPTID,Number=.,Type=String,Description="ID of the assembled alternate allele in the assembly file">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=HOMLEN,Number=.,Type=Integer,Description="Length of base pair identical micro-homology at event breakpoints">
##INFO=<ID=HOMSEQ,Number=.,Type=String,Description="Sequence of base pair identical micro-homology at event breakpoints">
##INFO=<ID=SVLEN,Number=.,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DEL:ME:ALU,Description="Deletion of ALU element">
##ALT=<ID=DEL:ME:L1,Description="Deletion of L1 element">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=DUP:TANDEM,Description="Tandem Duplication">
##ALT=<ID=INS,Description="Insertion of novel sequence">
##ALT=<ID=INS:ME:ALU,Description="Insertion of ALU element">
##ALT=<ID=INS:ME:L1,Description="Insertion of L1 element">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=CNV,Description="Copy number variable region">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype quality">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise events">
##FORMAT=<ID=CNQ,Number=1,Type=Float,Description="Copy number genotype quality for imprecise events">
#CHROM POS     ID        REF              ALT          QUAL FILTER INFO                                                              FORMAT       NA00001
1      2827694 rs2376870 CGTGGATGCGGGGAC  C            .    PASS   SVTYPE=DEL;END=2827708;HOMLEN=1;HOMSEQ=G;SVLEN=-14                GT:GQ        1/1:13.9
2       321682 .         T                <DEL>        6    PASS   SVTYPE=DEL;END=321887;SVLEN=-205;CIPOS=-56,20;CIEND=-10,62         GT:GQ        0/1:12
2     14477084 .         C                <DEL:ME:ALU> 12   PASS   SVTYPE=DEL;END=14477381;SVLEN=-297;CIPOS=-22,18;CIEND=-12,32       GT:GQ        0/1:12
3      9425916 .         C                <INS:ME:L1>  23   PASS   SVTYPE=INS;END=9425916;SVLEN=6027;CIPOS=-16,22                    GT:GQ        1/1:15
3     12665100 .         A                <DUP>        14   PASS   SVTYPE=DUP;END=12686200;SVLEN=21100;CIPOS=-500,500;CIEND=-500,500 GT:GQ:CN:CNQ ./.:0:3:16.2
4     18665128 .         T                <DUP:TANDEM> 11   PASS   SVTYPE=DUP;END=18665204;SVLEN=76;CIPOS=-10,10;CIEND=-10,10       GT:GQ:CN:CNQ ./.:0:5:8.3
```

Danecek P. et al. *The variant call format and VCFtools.* (2011). *Bioinformatics*. PMID: 21653522

VCF for SVs

```
##fileformat=VCFv4.1
##fileDate=20100501
##reference=1000GenomesPilot-NCBI36
##assembly=ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/sv/breakpoint_assemblies.fasta
##INFO=<ID=BKPTID,Number=.,Type=String,Description="ID of the assembled alternate allele in the assembly file">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=HOMLEN,Number=.,Type=Integer,Description="Length of base pair identical micro-homology at event breakpoints">
##INFO=<ID=HOMSEQ,Number=.,Type=String,Description="Sequence of base pair identical micro-homology at event breakpoints">
##INFO=<ID=SVLEN,Number=.,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DEL:ME:ALU,Description="Deletion of ALU element">
##ALT=<ID=DEL:ME:L1,Description="Deletion of L1 element">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=DUP:TANDEM,Description="Tandem Duplication">
##ALT=<ID=INS,Description="Insertion of novel sequence">
##ALT=<ID=INS:ME:ALU,Description="Insertion of ALU element">
##ALT=<ID=INS:ME:L1,Description="Insertion of L1 element">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=CNV,Description="Copy number variable region">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype quality">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise events">
##FORMAT=<ID=CNQ,Number=1,Type=Float,Description="Copy number genotype quality for imprecise events">
#CHROM POS     ID        REF              ALT          QUAL FILTER INFO                                                              FORMAT       NA00001
1      2827694 rs2376870 CGTGGATGCGGGGAC  C            .    PASS   SVTYPE=DEL;END=2827708;HOMLEN=1;HOMSEQ=G;SVLEN=-14                GT:GQ        1/1:13.9
2       321682 .         T                <DEL>        6    PASS   SVTYPE=DEL;END=321887;SVLEN=-205;CIPOS=-56,20;CIEND=-10,62         GT:GQ        0/1:12
2     14477084 .         C                <DEL:ME:ALU> 12   PASS   SVTYPE=DEL;END=14477381;SVLEN=-297;CIPOS=-22,18;CIEND=-12,32       GT:GQ        0/1:12
3      9425916 .         C                <INS:ME:L1>  23   PASS   SVTYPE=INS;END=9425916;SVLEN=6027;CIPOS=-16,22                    GT:GQ        1/1:15
3     12665100 .         A                <DUP>        14   PASS   SVTYPE=DUP;END=12686200;SVLEN=21100;CIPOS=-500,500;CIEND=-500,500 GT:GQ:CN:CNQ ./.:0:3:16.2
4     18665128 .         T                <DUP:TANDEM> 11   PASS   SVTYPE=DUP;END=18665204;SVLEN=76;CIPOS=-10,10;CIEND=-10,10       GT:GQ:CN:CNQ ./.:0:5:8.3
```

Danecek P. et al. *The variant call format and VCFtools.* (2011). *Bioinformatics*. PMID: 21653522

VCF for SVs

```
##fileformat=VCFv4.1
##fileDate=20100501
##reference=1000GenomesPilot-NCBI36
##assembly=ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/sv/breakpoint_assemblies.fasta
##INFO=<ID=BKPTID,Number=.,Type=String,Description="ID of the assembled alternate allele in the assembly file">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=HOMLEN,Number=.,Type=Integer,Description="Length of base pair identical micro-homology at event breakpoints">
##INFO=<ID=HOMSEQ,Number=.,Type=String,Description="Sequence of base pair identical micro-homology at event breakpoints">
##INFO=<ID=SVLEN,Number=.,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DEL:ME:ALU,Description="Deletion of ALU element">
##ALT=<ID=DEL:ME:L1,Description="Deletion of L1 element">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=DUP:TANDEM,Description="Tandem Duplication">
##ALT=<ID=INS,Description="Insertion of novel sequence">
##ALT=<ID=INS:ME:ALU,Description="Insertion of ALU element">
##ALT=<ID=INS:ME:L1,Description="Insertion of L1 element">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=CNV,Description="Copy number variable region">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype quality">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise events">
##FORMAT=<ID=CNQ,Number=1,Type=Float,Description="Copy number genotype quality for imprecise events">
#CHROM POS     ID        REF              ALT          QUAL FILTER INFO                                                              FORMAT       NA00001
1      2827694 rs2376870 CGTGGATGCGGGGAC  C            .    PASS   SVTYPE=DEL;END=2827708;HOMLEN=1;HOMSEQ=G;SVLEN=-14                GT:GQ        1/1:13.9
2       321682 .         T                <DEL>        6    PASS   SVTYPE=DEL;END=321887;SVLEN=-205;CIPOS=-56,20;CIEND=-10,62         GT:GQ        0/1:12
2     14477084 .         C                <DEL:ME:ALU> 12   PASS   SVTYPE=DEL;END=14477381;SVLEN=-297;CIPOS=-22,18;CIEND=-12,32       GT:GQ        0/1:12
3      9425916 .         C                <INS:ME:L1>  23   PASS   SVTYPE=INS;END=9425916;SVLEN=6027;CIPOS=-16,22                    GT:GQ        1/1:15
3     12665100 .         A                <DUP>        14   PASS   SVTYPE=DUP;END=12686200;SVLEN=21100;CIPOS=-500,500;CIEND=-500,500 GT:GQ:CN:CNQ ./.:0:3:16.2
4     18665128 .         T                <DUP:TANDEM> 11   PASS   SVTYPE=DUP;END=18665204;SVLEN=76;CIPOS=-10,10;CIEND=-10,10       GT:GQ:CN:CNQ ./.:0:5:8.3
```

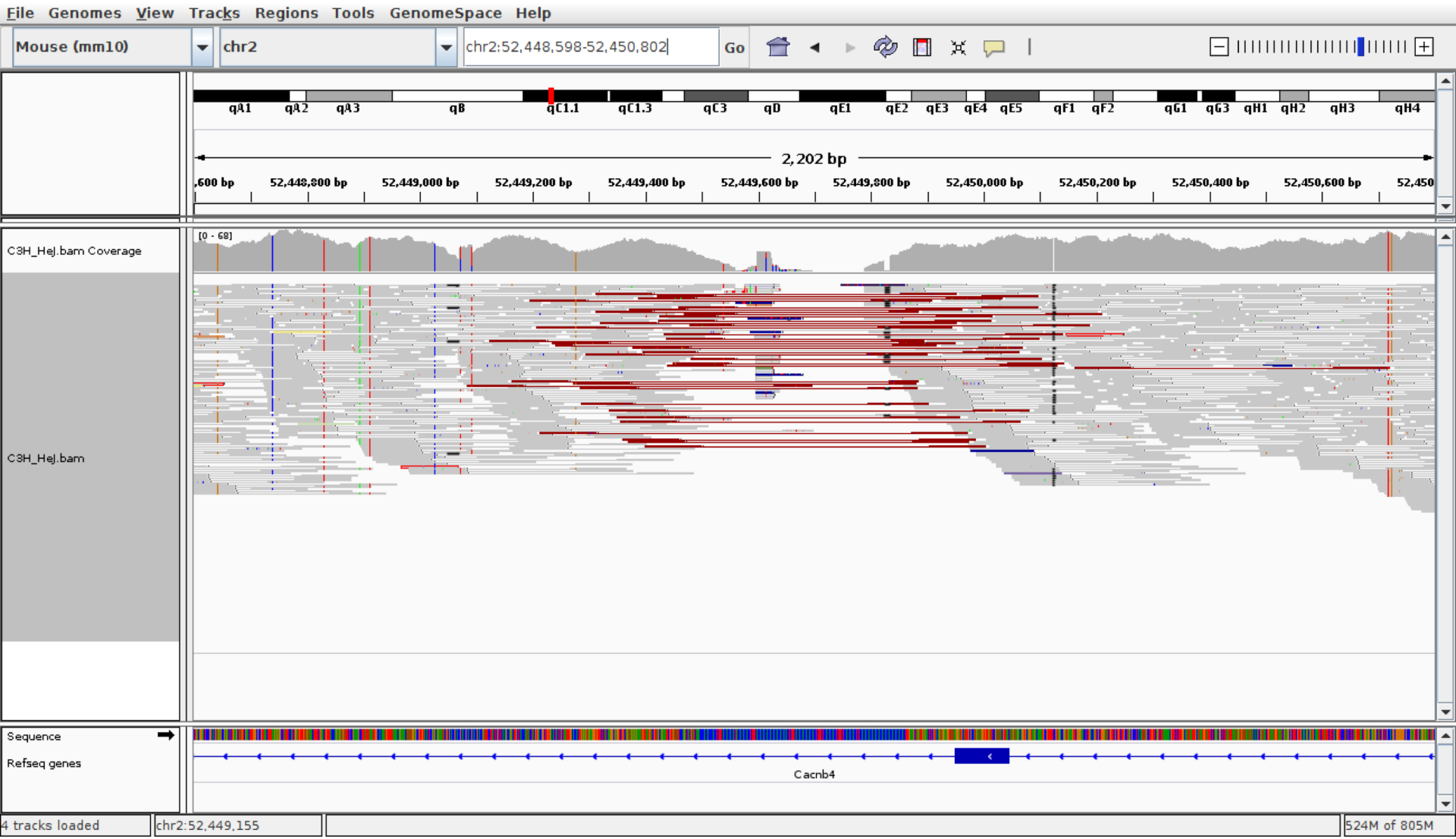
Danecek P. et al. *The variant call format and VCFtools.* (2011). *Bioinformatics*. PMID: 21653522

SV Visualisation

- Structural variation visualisation can be more challenging than SNPs and indels
- Inspect several hundred base pairs or multiple kbp
- Analyse complicated read pair patterns to determine type of SV and sources of error
- Look for soft clipped bases for breakpoint accuracy
- Many NGS visualisation software packages exist
- The Integrative Genomics Viewer (**IGV**) from Broad institute is a popular and easy to use visualisation software
  - Requires BAM file and fasta file of the reference genome
  - Viewing settings need to be tailored for the type of SV being visualised (see notes below each screenshot)

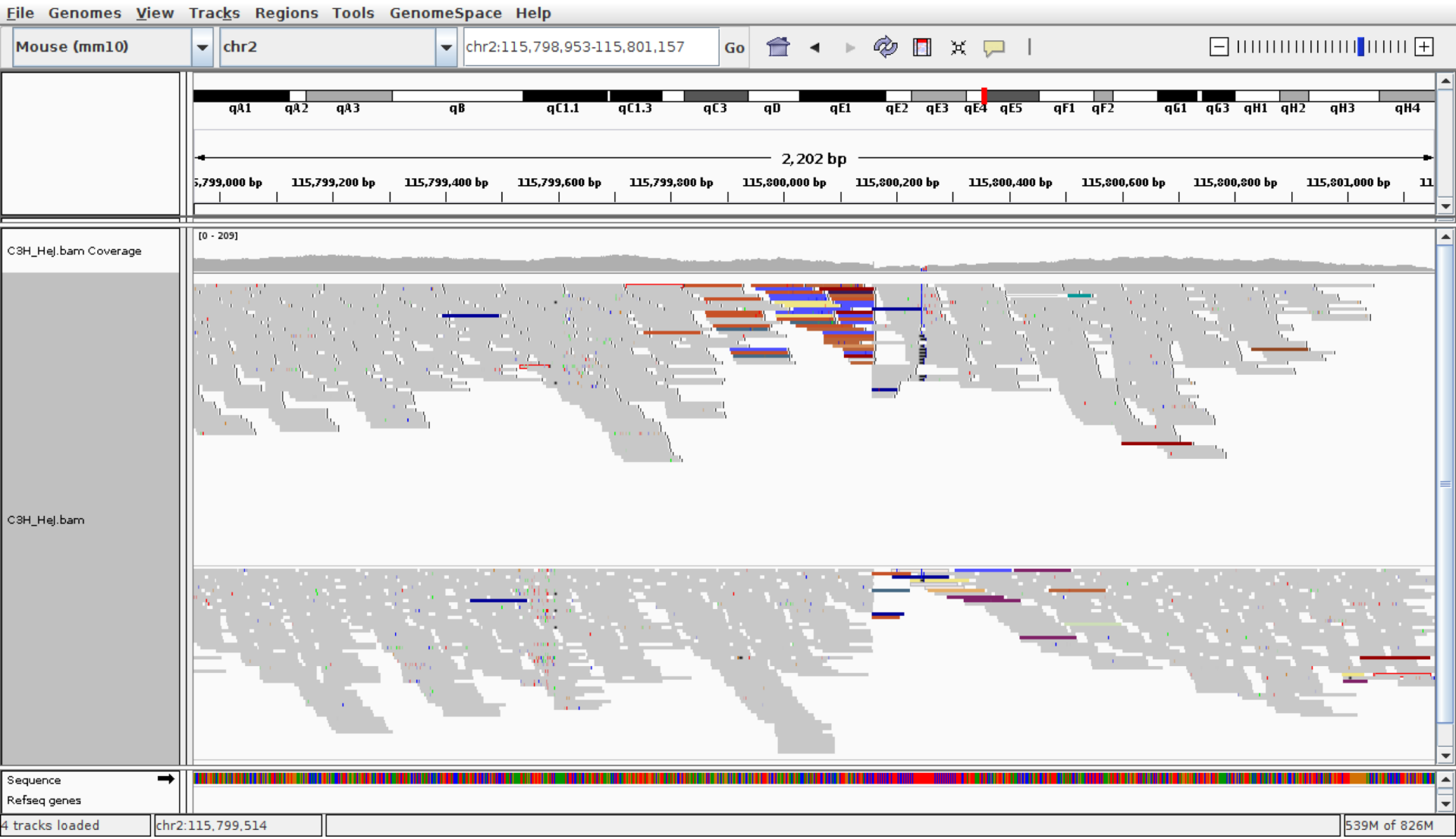
Thorvaldsdóttir H. et al. *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration*. (2013). *Brief. Bioinformatics.* PMID: 22517427

IGV - Deletion



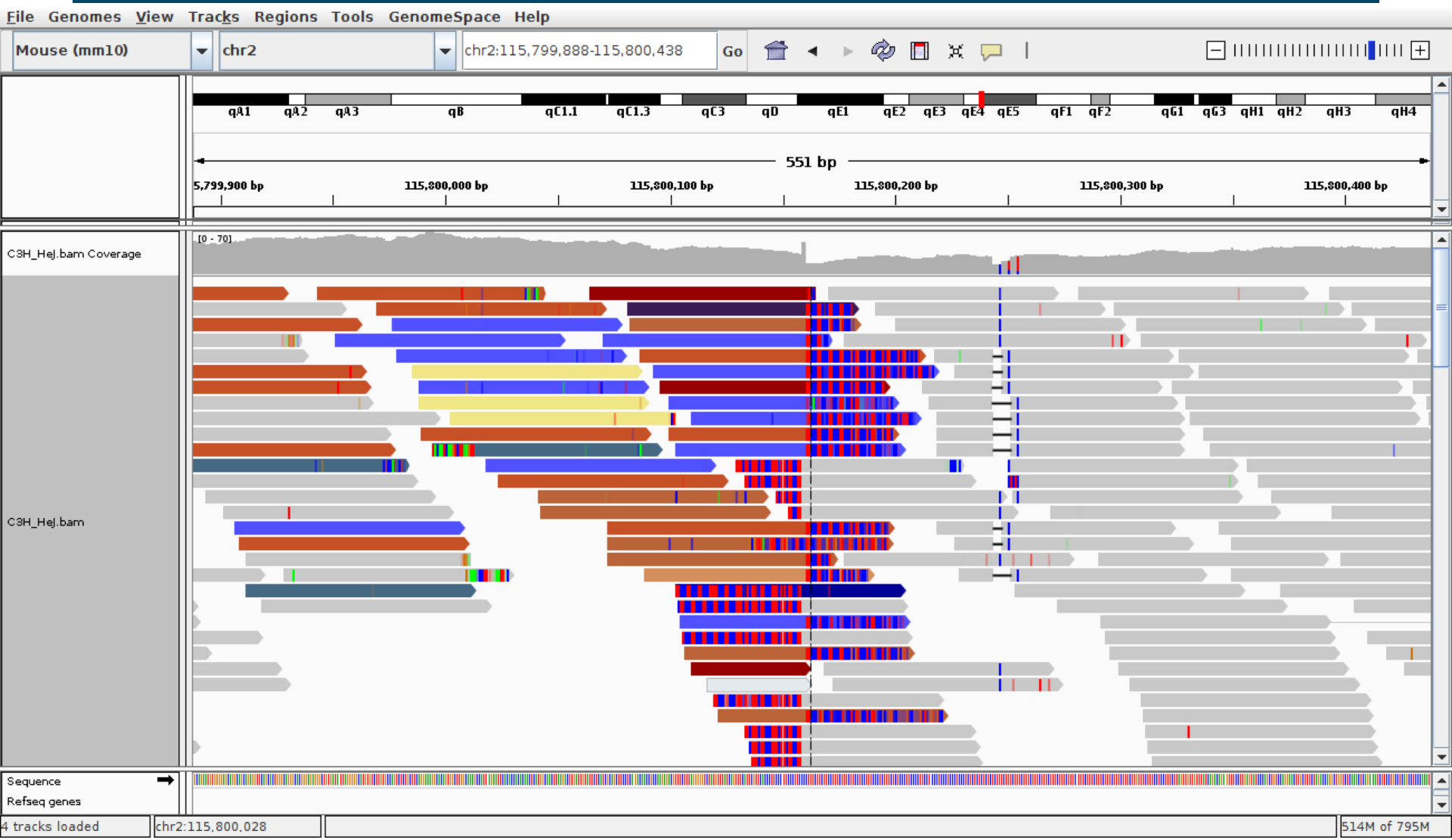
Right click - Squished

IGV - Insertion



Right click - group alignments by 'read strand'

IGV - Insertion (zoomed)

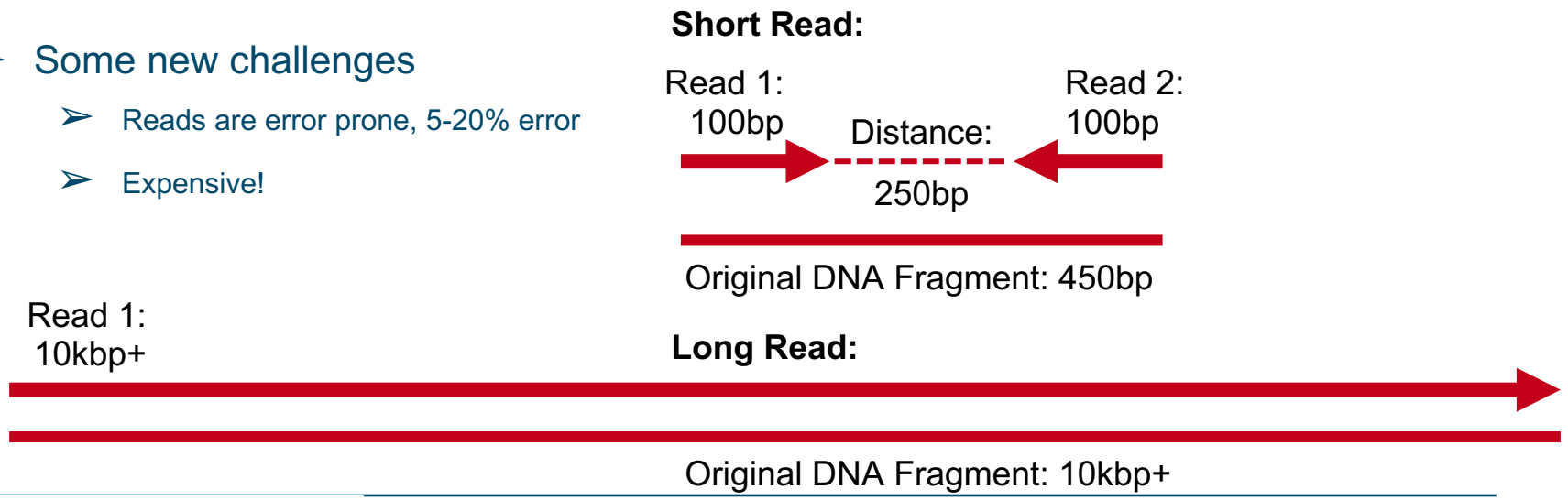


Right click - view mismatch bases

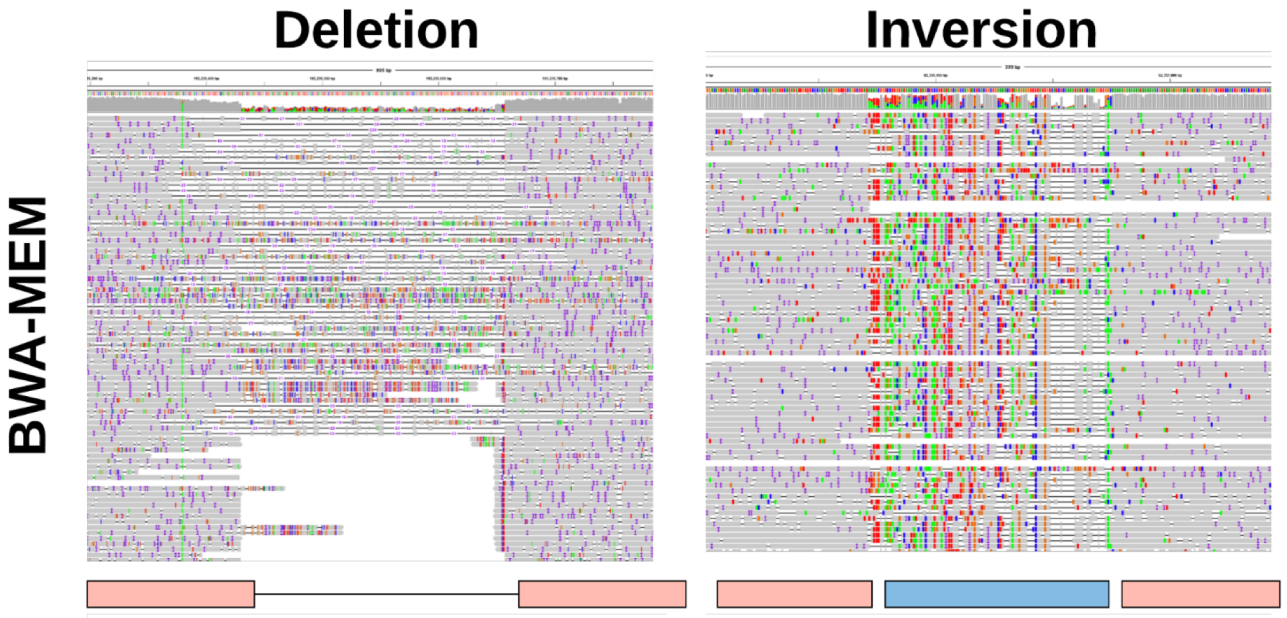
View - preferences - alignments - show soft clipped bases

SVs and long read sequencing

- Single molecule sequencing of large DNA fragments
- Platforms: Oxford nanopore and Pacific Biosciences
- Read lengths routinely 10-20Kbp
- Longer than most common transposable element repeats
- What does it mean for SV detection? Span both breakpoints with single read
- Some new challenges
  - Reads are error prone, 5-20% error
  - Expensive!



Alignment challenges



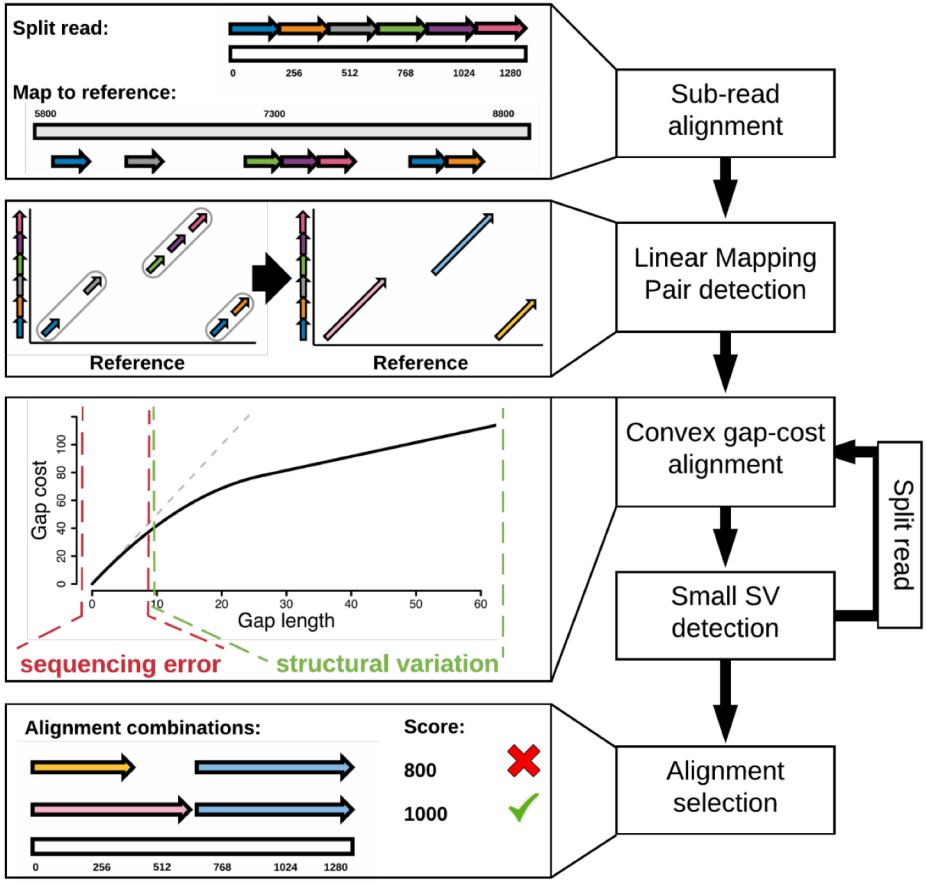
Sedlazeck F. et al. *Accurate detection of complex structural variations using single-molecule sequencing*. (2018). *Nature Methods*. PMID: 29713083

Alignment challenges

Sedlazeck F. et al. *Accurate detection of complex structural variations using single-molecule sequencing*. (2018). *Nature Methods*. PMID: 29713083

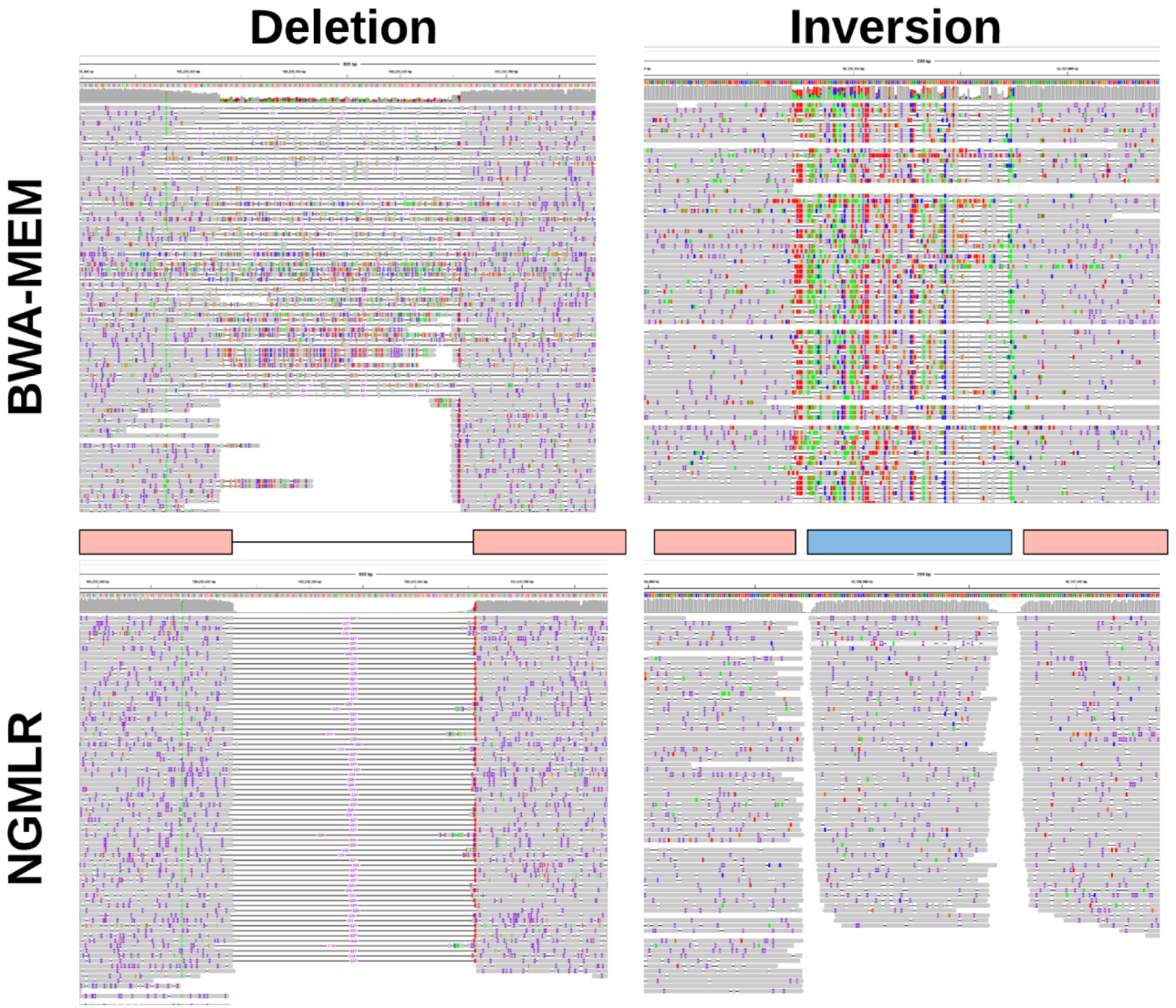
coNvex Gap-cost alignMents for Long Reads (NGMLR)

- NGMLR - aligner specifically designed for long reads
- Convex scoring model
  - Extending an indel is penalized proportionally less the longer the indel is



Sedlazeck F. et al. *Accurate detection of complex structural variations using single-molecule sequencing*. (2018). *Nature Methods*. PMID: 29713083

Alignment challenges

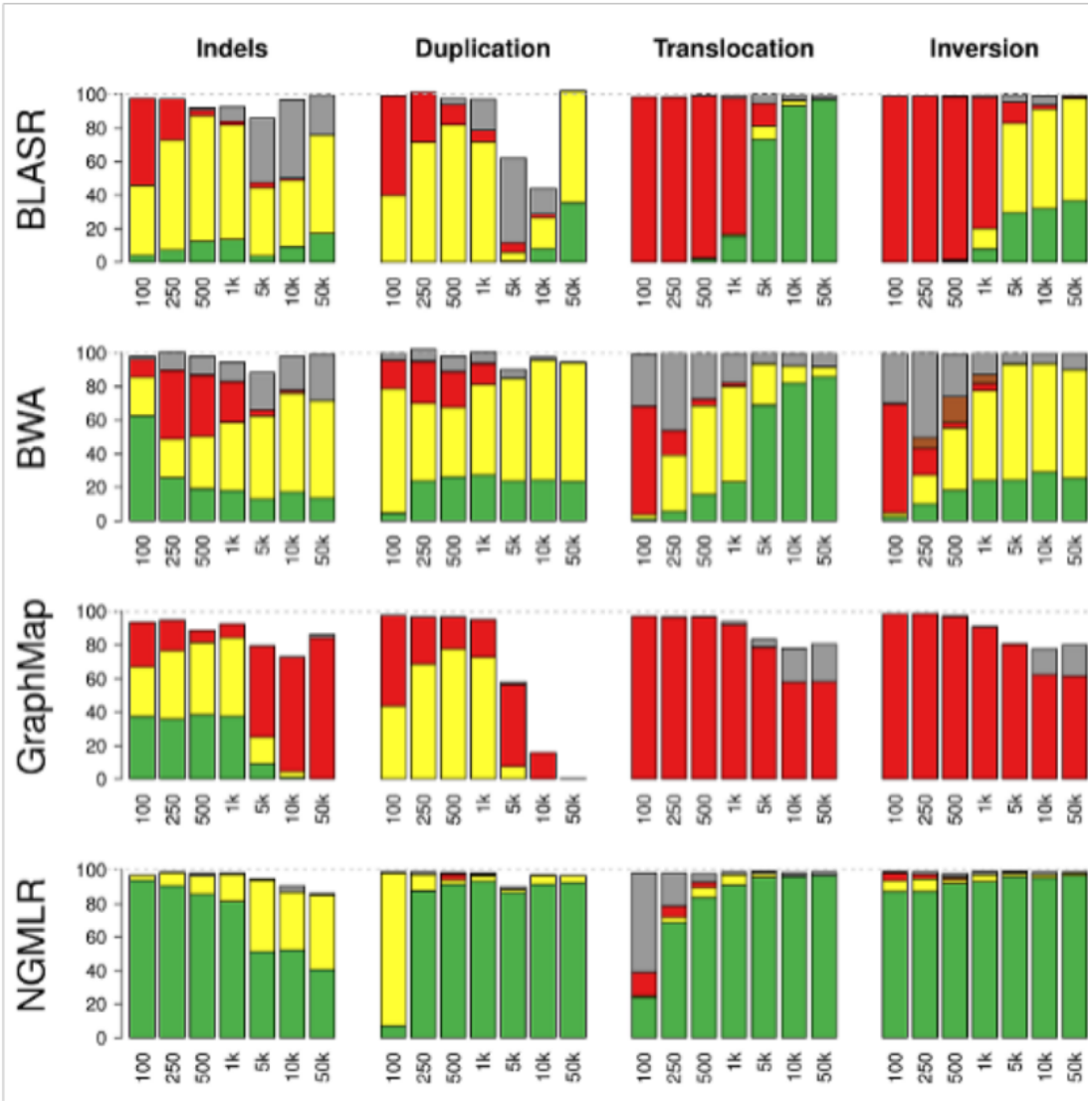


Sedlazeck F. et al. *Accurate detection of complex structural variations using single-molecule sequencing*. (2018). *Nature Methods*. PMID: 29713083

Comparison of aligners (simulated data)

Alignment Status:

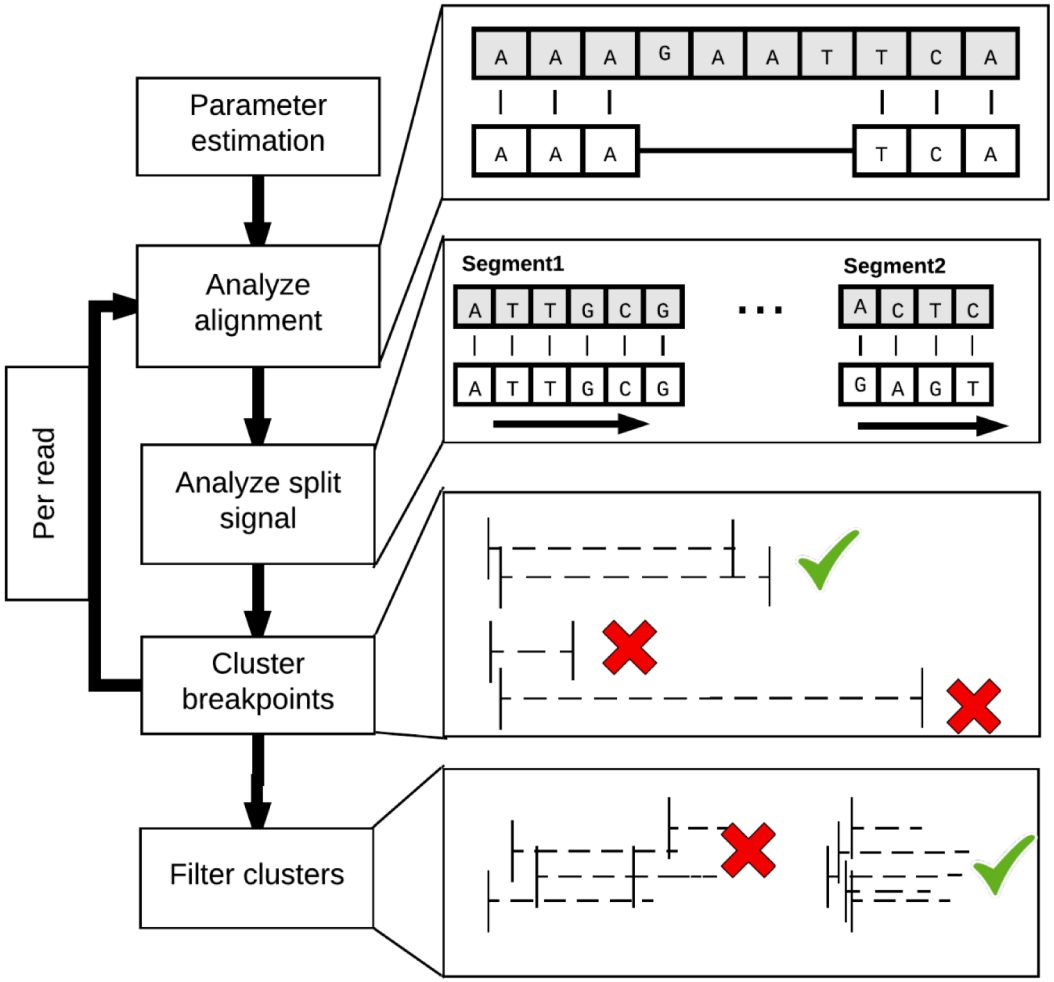
- Precise
- Indicated
- Forced
- Unaligned reads
- Trimmed but not aligned through the SV



Sedlazeck F. et al. *Accurate detection of complex structural variations using single-molecule sequencing*. (2018). *Nature Methods*. PMID: 29713083

Sniffles

SV detection from long read alignments

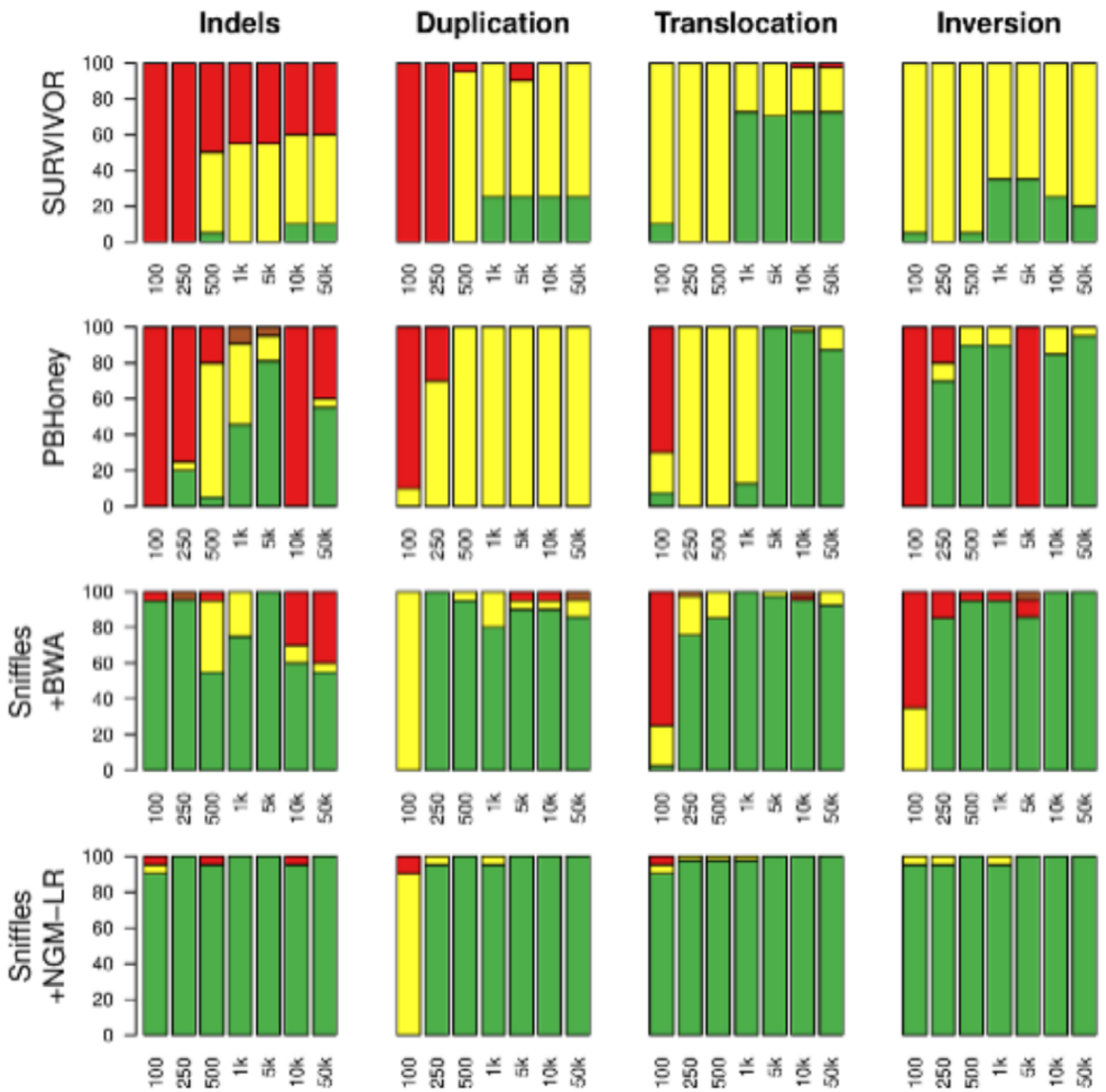


Sedlazeck F. et al. *Accurate detection of complex structural variations using single-molecule sequencing*. (2018). *Nature Methods*. PMID: 29713083

Sniffles performance

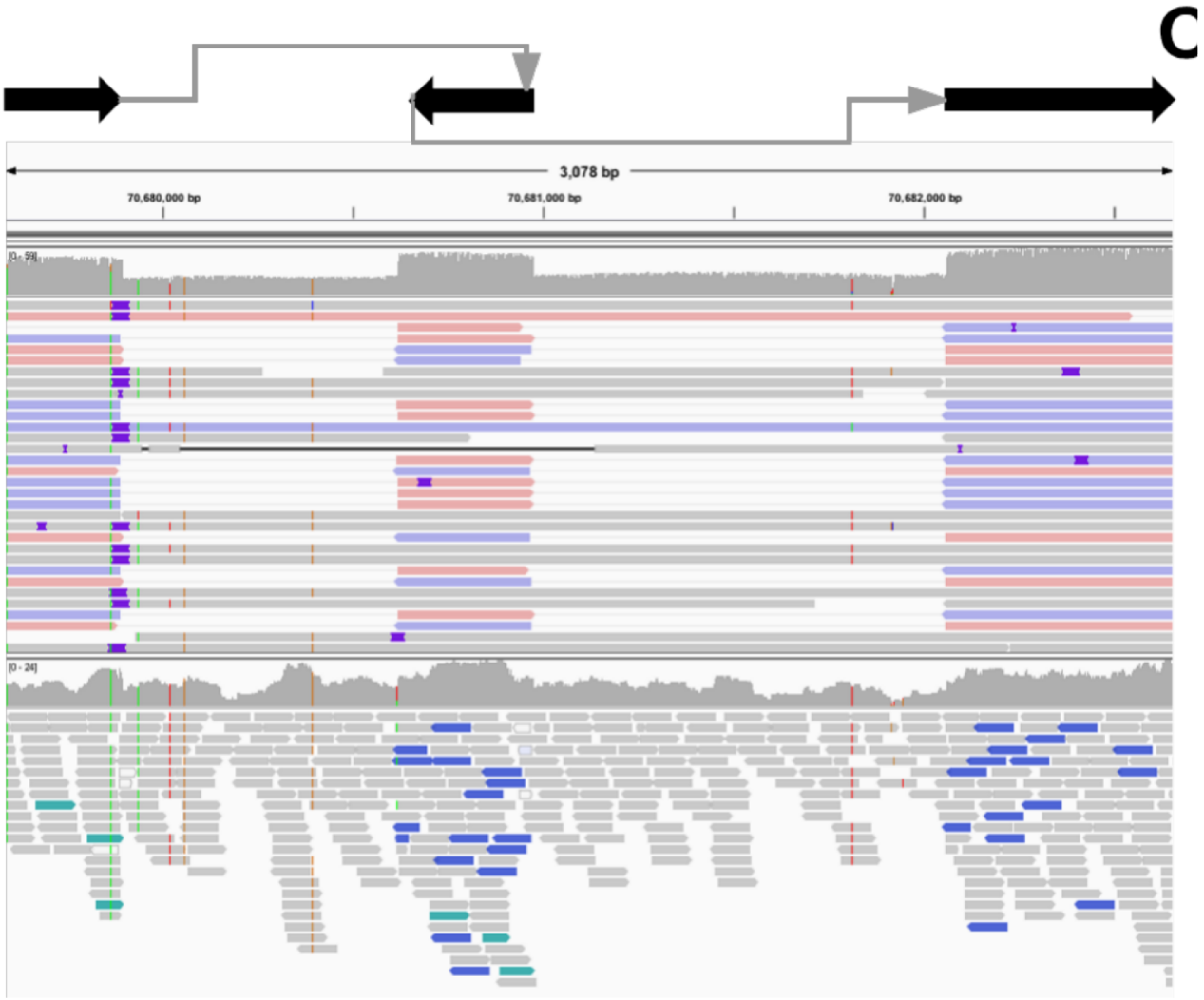
Alignment Status:

- Precise
- Indicated
- Forced
- Unaligned reads
- Trimmed but not aligned through the SV



Sedlazeck F. et al. *Accurate detection of complex structural variations using single-molecule sequencing*. (2018). *Nature Methods*. PMID: 29713083

Complex SVs - Long reads

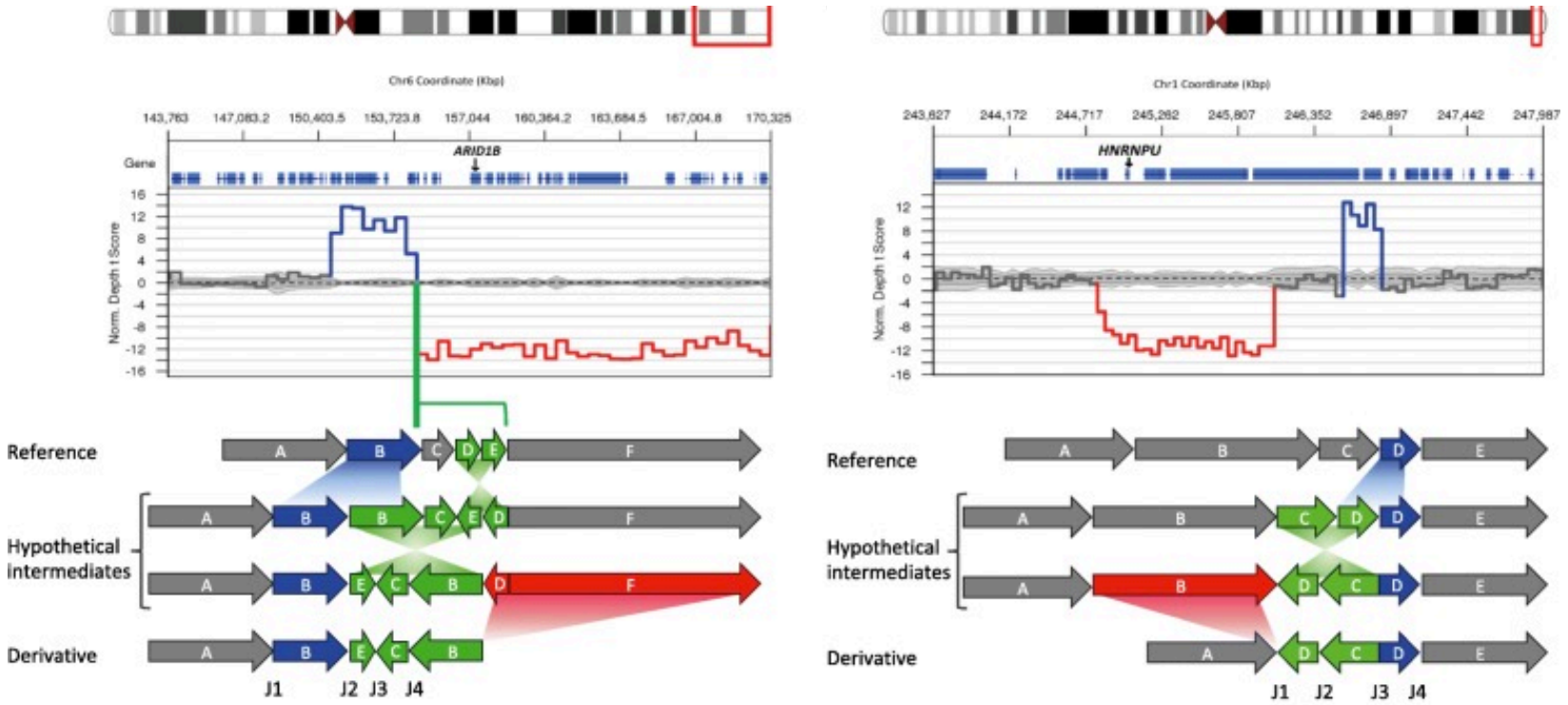


NGLMR

Illumina (BWA-mem)

3kb region: two deletions flanking an inverted sequence

Complex SV Examples in Diagnostic Sequencing



Phenotype:

Coffin-Siris syndrome; Atrial septal defect; Cleft soft palate

Size:

~23 Million Base Pairs (23 Mbp)

Phenotype:

Tonic-clonic seizures; Intellectual disability; Learning difficulties

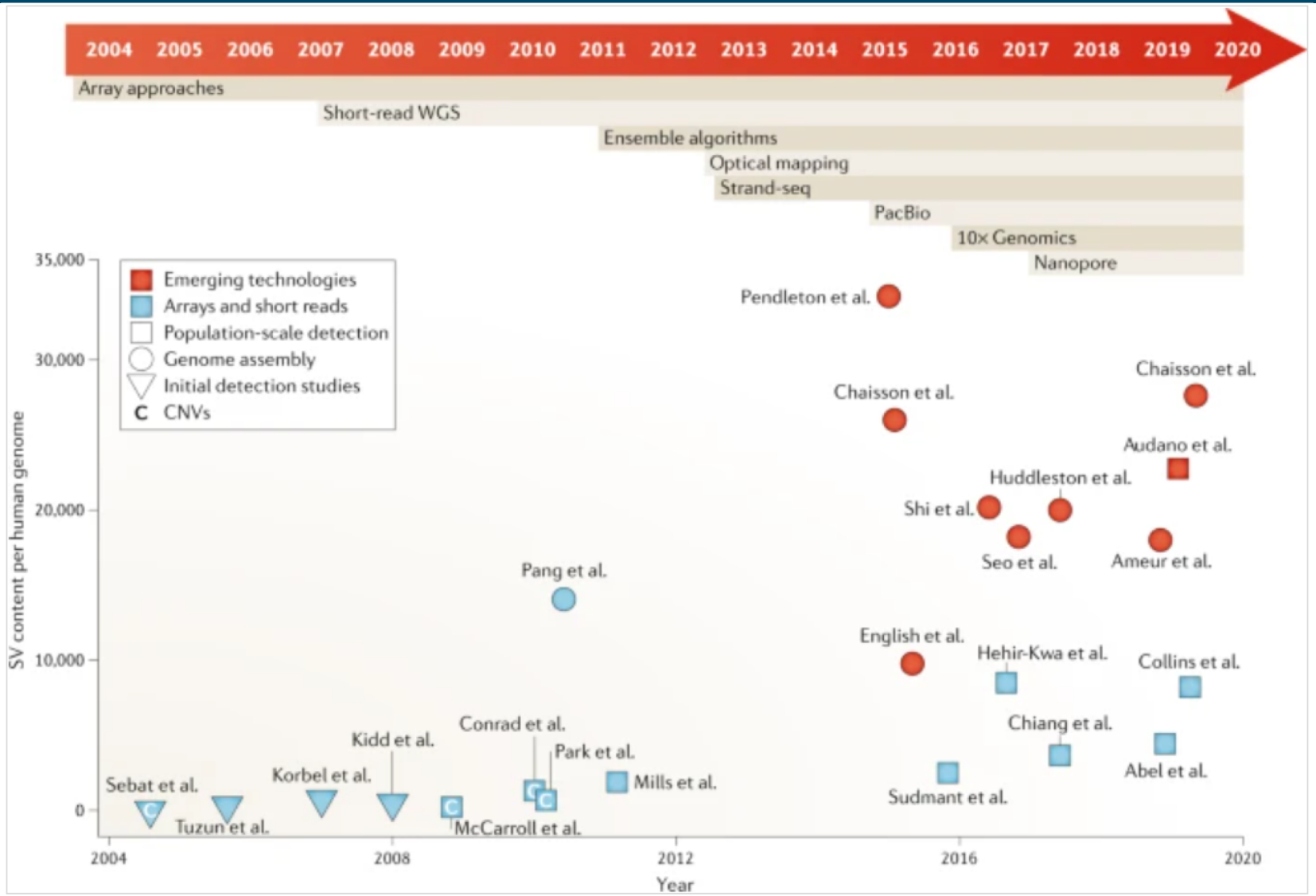
Size:

~2 Million Base Pairs (2Mbp)

Both were associated with patient phenotype!

Sanchis-Juan A. et al. *Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing.* (2018). *Genome Medicine.* PMID: 30526634

New Technologies Allow Identification of More SVs



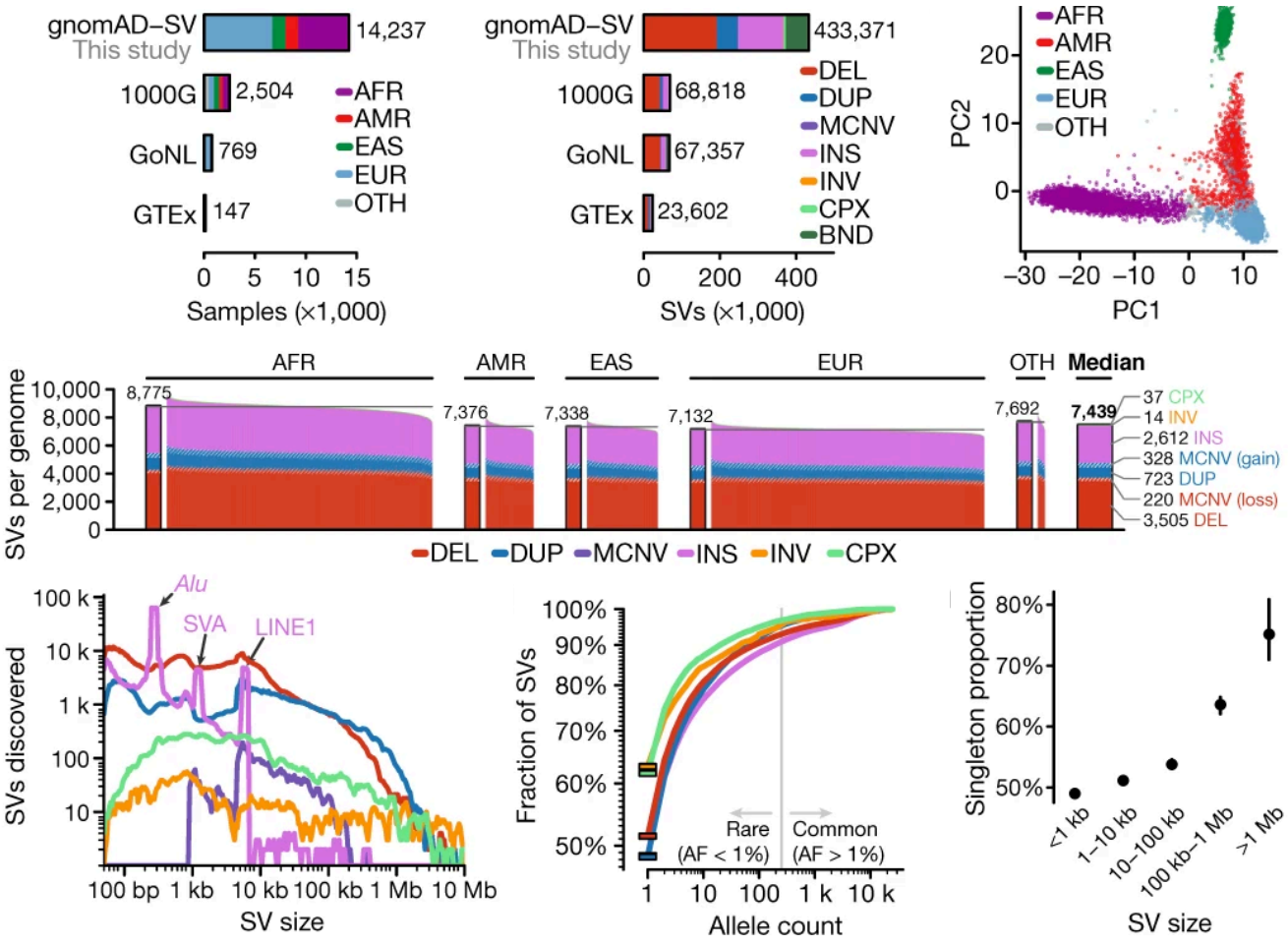
Ho S. S., Urban A. E., and Mills R.E. *Structural variation in the sequencing era.* (2020). *Nature Reviews Genetics*. PMID: 31729472

The Future of Structural Variation – Short Reads

The gnomAD SV project

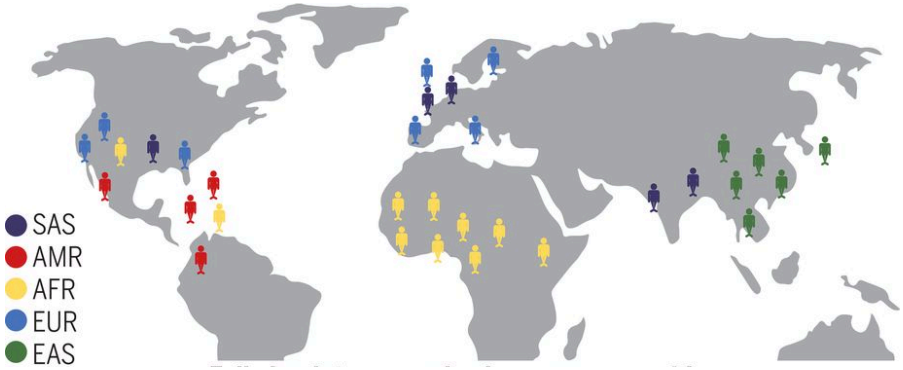
Collins RL et al. assessed >14,000 human genomes from a diverse set of populations for Structural Variation with short read technology as part of the gnomAD project.

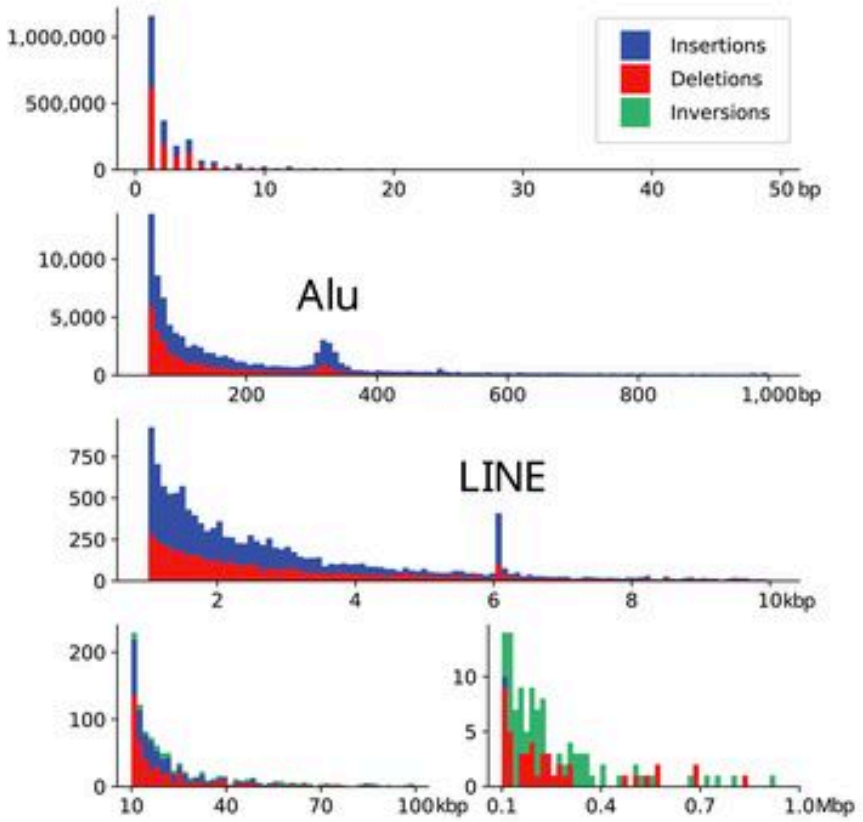
They identified 433,371 SVs.



Collins, R. L. *et al.* A structural variant reference for medical and population genetics. (2020). Nature. PMID: 32461652

The Future of Structural Variation – Long Reads





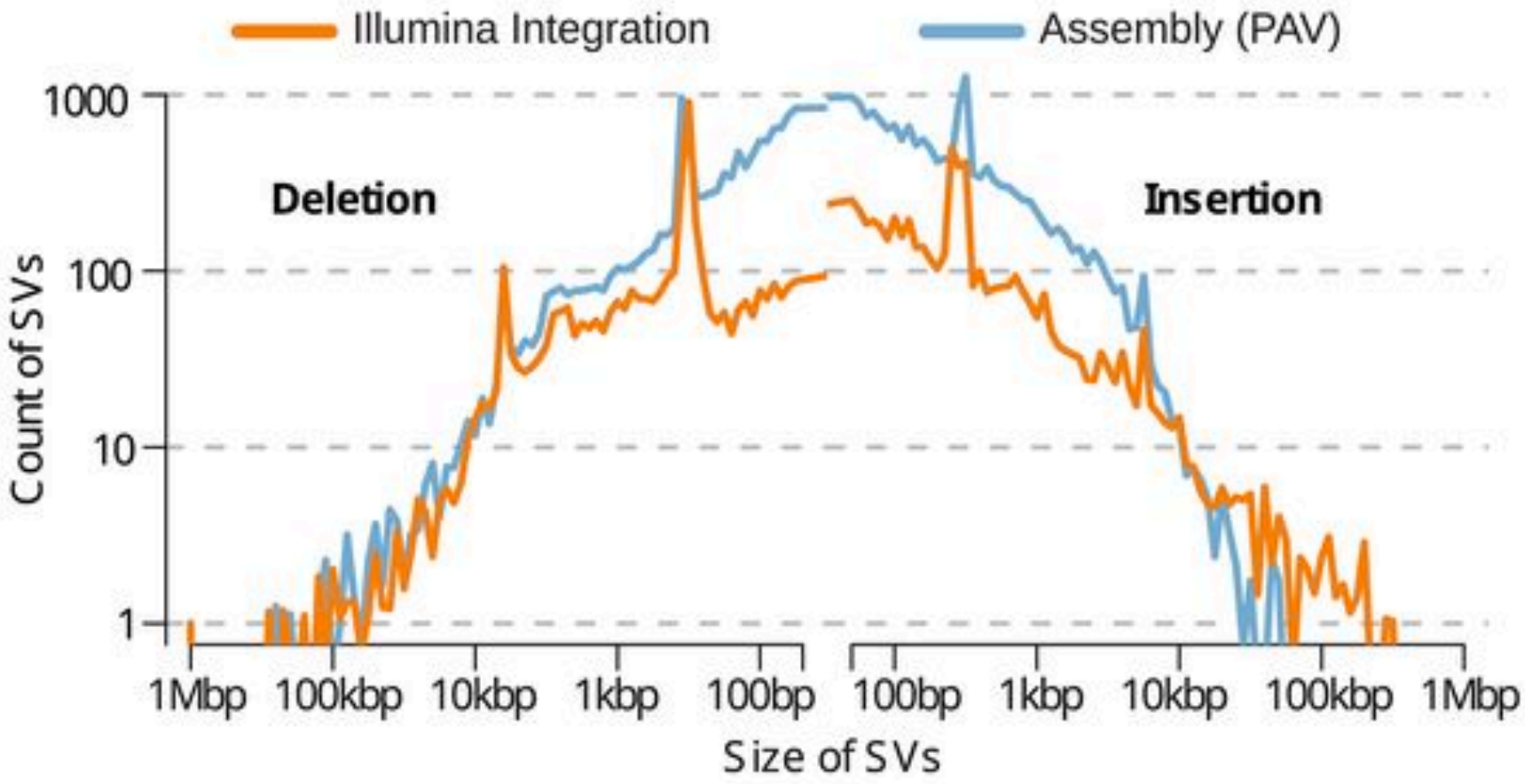
The Human Genome Structural Variation Consortium

Ebert P. et al. used long read sequencing to identify SVs in 64 humans from around the world.

They identified a total of 107,590 SVs

Ebert P. et al. *Haplotype-resolved diverse human genomes and integrated analysis of structural variation*. (2021). *Science*. PMID: 33632895

The Future of Structural Variation – Which Technology?



Ebert P. et al. *Haplotype-resolved diverse human genomes and integrated analysis of structural variation*. (2021). *Science*. PMID: 33632895

The Future of Structural Variation – Which Technology?

- Long Reads or Short Reads?
  - Long Reads are still expensive
  - Short Reads don't work well at some size ranges

So do we…

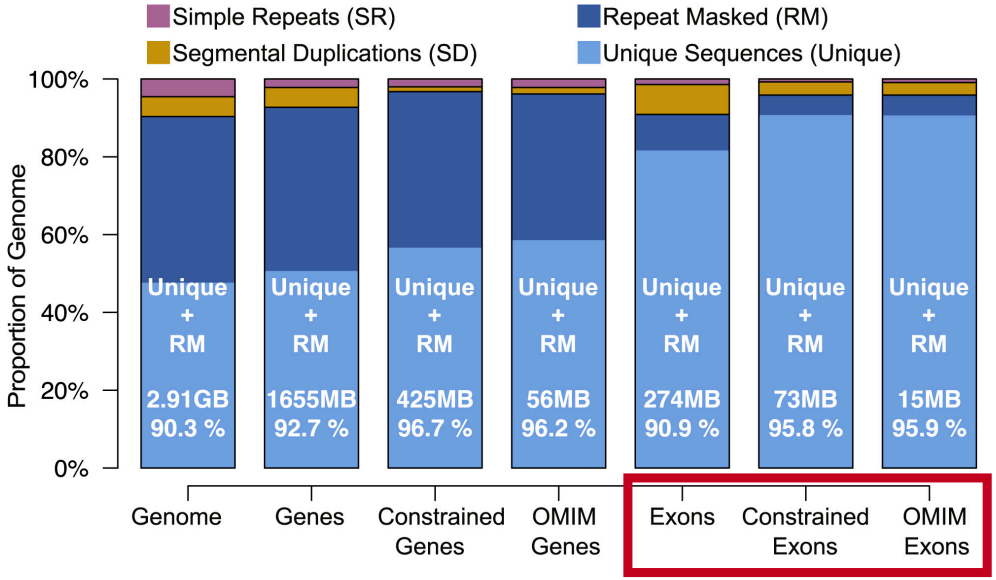
… Sequence small number of samples really well (long reads), or…
… large number of samples adequately (short reads)?

Or: How do we best trade-off number of samples vs. accuracy of SV identification…?

Ebert P. et al. *Haplotype-resolved diverse human genomes and integrated analysis of structural variation*. (2021). *Science*. PMID: 33632895

The Future of Structural Variation – Which Technology?

Do we actually need all of those structural variants that long read sequencing finds…?



These are where variants which cause disease are found (currently!)

"[split read sequencing] captures virtually all high-quality deletions derived from [long read sequencing] assembly in the regions of the genome that encompass over 95% of currently annotated coding sequence in genes with existing evidence for dominant-acting pathogenic mutations from OMIM. We therefore anticipate that a minority of "unsolved" cases will be explained by novel and readily interpretable deletions that can be captured by [long read sequencing] but remain cryptic to [split read sequencing] in known disease-associated genes…

… However, given that the most highly repetitive regions of the genome have been traditionally inaccessible in human disease studies, it is anticipated that new disease-associated genes and sequences will emerge as functional annotation of these repetitive sequences and duplicated genes continues to improve."

Zhao X et al. *Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies.* (2021). *AJHG*. PMID: 33789087

Computer exercises

1. Trivia questions about a VCF output file from the LUMPY SV caller.
   a. http://www.genomebiology.com/2014/15/6/R84

2. Use the BreakDancer software package to call structural variants on a yeast sample that was paired-end sequenced on the illumina Hiseq.

3. Use the LUMPY software package to call structural variants on a yeast sample that was paired-end sequenced on the illumina Hiseq.

4. Call SVs using the Sniffles caller on a yeast sample that was sequenced on the Pacbio platform.

5. Introduction to BEDtools for doing regional comparisons over genomic co-ordinates.