Structural variation detection and interpretation

Thomas Keane,
EVA + EGA,
EMBL-EBI
@drtkeane
E: tk2@ebi.ac.uk

Genomic structural variation

- Any form of rearrangement of chromosome structure
  - Contribute to genetic diversity and evolution, new gene formation, gene function, phenotypic diversity, rare variants of large effect

- Frequent causes of disease
  - Referred to as genomic disorders
  - Mendelian diseases or complex traits such as behaviors
  - E.g. increase in gene dosage due to increase in copy number

- Several type or categories of structural variation
  - Insertions, deletions, copy number changes, inversions, translocations
  - Complex events contain combinations of these in close proximity

- Breakpoint: as a pair of bases that are adjacent in an experimentally sequenced 'sample' genome but not in the reference genome

- Many experimental techniques to detect SVs

SVs and human disease

Table 1 Copy-number variations and neurogenetic disorders (expanded from References 49 and 50)

Syndrome	OMIM	Locus	Rearrangement	Gene(s)	Reference
Neurodevelopmental					
WBS del(7)q11.23	194050	7q11.23	del	CGS incl. *ELN*	51
dup(7)q11.23	609757		dup		52
AS	105830	15q11–q12	mat del, pat UPD15	*UBE3A*	53
PWS	176270		pat del, mat UPD15	CGS	54
dup(15)	608636	15q11–q13	dup	CGS	55
idic(15)		idic(15)(q13)	trip		56
MDLS	247200	17p13.3	del	CGS incl. *LIS1*	57
SMS	182290	17p11.2	del	CGS incl. *RAI1*	58
PTLS	610883		dup	*RAI1*	59
NF1	162200	17q11.2	del	CGS incl. *NF1*	60
del(17)q21.31	610443	17q21.31	del		61–63
DGS/VCFS	188400	22q11.2	del	CGS incl. *TBX1, COMT*	64
	192430				
dup(22)q11.2	608363		dup	CGS	65
del(22)q13	606232	22q13.3	del	*SHANK3/PROSAP2*	66
RTT	312750	Xq28	del	*MECP2*	67
Rett-like syndrome	300260		dup, trip	*MECP2*	68
PMD	312080	Xq22.2	dup, del	*PLP1*	69
Neurodegenerative					
PD	168601	4q21	dup, trip	*SNCA*	70
SMA	253300	5q13	del, gene conv	*SMN1, SMN2*	71
ADLD	169500	5q23.2	dup	*LMNB1*	72
CMT1A	118220	17p12	dup	*PMP22*	73, 74
HNPP	162500		del		
AD	104300	21q21	dup	*APP*	75

Abbreviations: AD, Alzheimer disease; ADLD, autosomal dominant leukodystrophy; AS, Angelman syndrome; CGS, contiguous gene deletion/duplication syndrome; CMT1A, Charcot-Marie-Tooth type 1 disease; del, deletion; dup(7)q11.23, reciprocal duplication of the WBS region; dup, duplication; gene conv, gene conversion; HNPP, hereditary neuropathy with liability to pressure palsies; MDLS, Miller-Dieker syndrome; NF1, neurofibromatosis type 1; PD, Parkinson disease; PMD, Pelizaeus-Merzbacher syndrome; PWS, Prader-Willi syndrome; RTT, Rett syndrome; SMA, spinal muscular atrophy; trip, triplication; UPD, uniparental disomy; WBS, Williams-Beuren syndrome.

Stankiewicz and Lupski (2010) Ann. Rev. Med.

Methods for detecting SVs

Experimental Approaches

- **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.

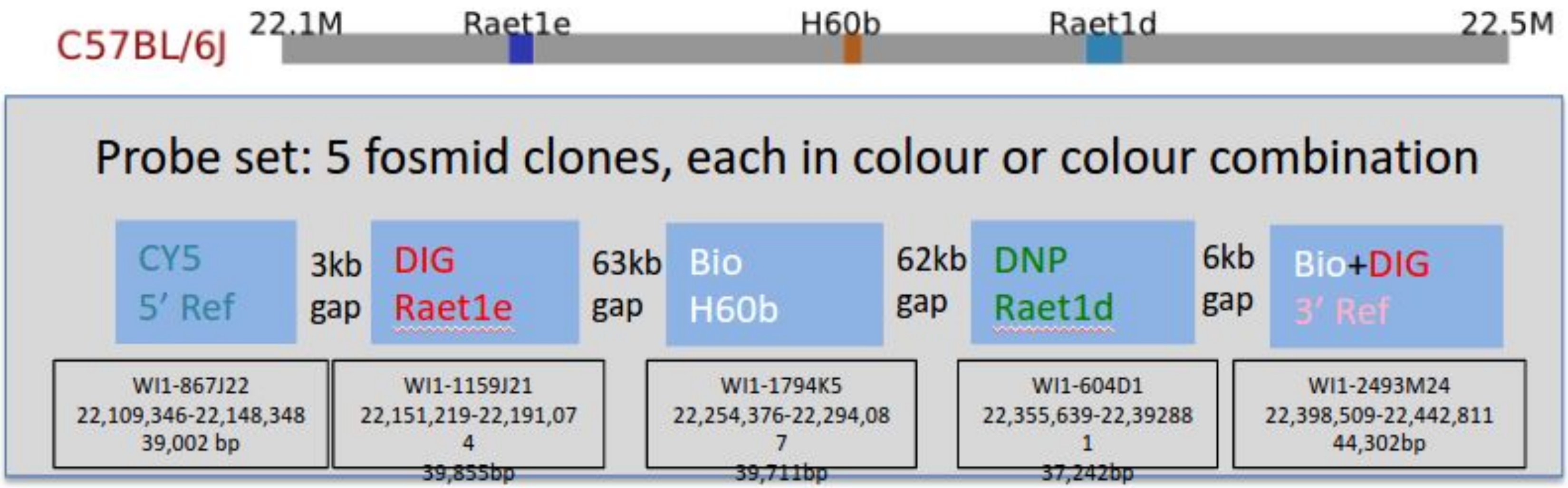
Sequencing Approaches

- **Whole-genome sequencing (WGS)**: Breakpoints of CNV and copy-neutral SV are detectable by paired-end reads that have discordant mappings to the reference genome.
- **Third generation sequencing:** sequencing long molecules of DNA (several kbp) and subsequent alignment to a reference genome to detect SVs

Weckselblatt and Rudd (2015) *Trends in Genetics*

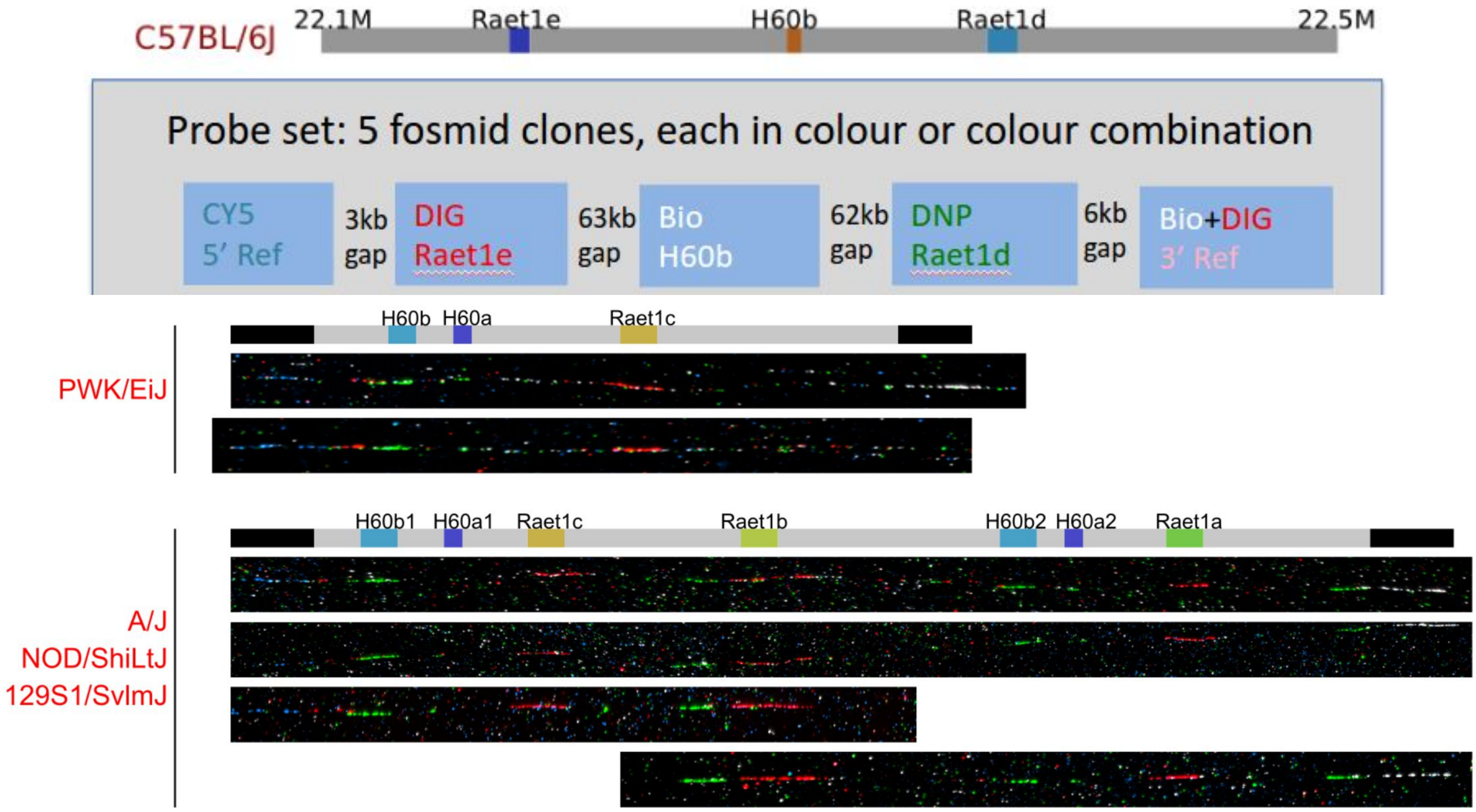
Fiber FISH

Chr10:21,100,000-21,700,000

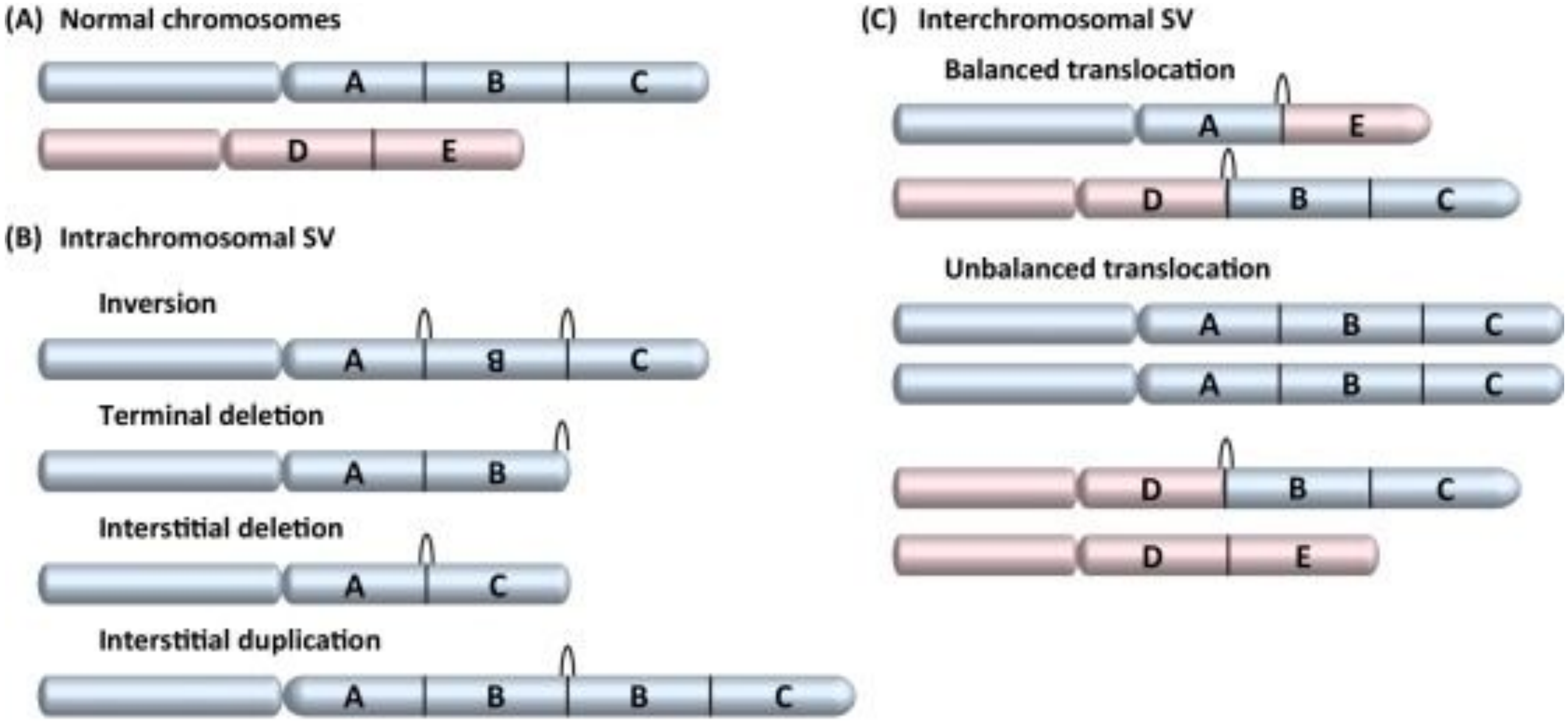


Fiber FISH

Chr10:21,100,000-21,700,000



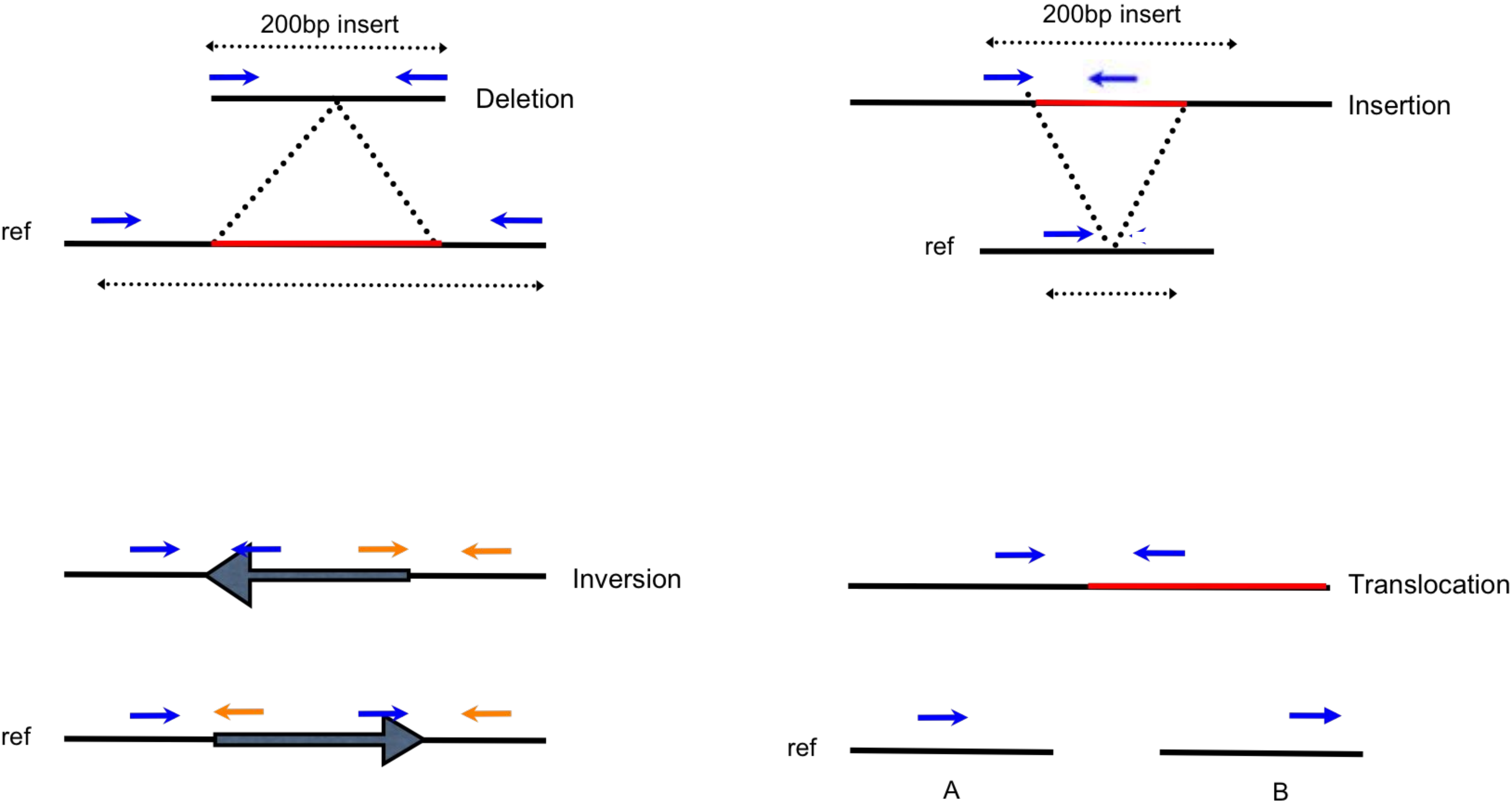
SV types



(A) Two nonhomologous chromosomes shown in blue and pink. Segments are labeled with letters A-E. Black arches indicate structural variation (SV) breakpoint junctions. (B) Intrachromosomal rearrangements include inversions, interstitial and terminal deletions, and interstitial duplications. (C) Simple translocations between two different chromosome ends. Balanced translocations do not result in copy-number variation (CNV), but unbalanced translocations have partial monosomy (segment E) and partial trisomy (segments B,C).

Weckselblatt and Rudd (2015) *Trends in Genetics*

SV types and NGS paired-end sequencing



Retrotransposition

Transposons are segments of DNA that can move within the genome

- ➢ A minimal ‘genome’ - ability to replicate and change location
- ➢ Relics of ancient viral infections

Dominate landscape of mammalian genomes

- ➢ **38-45%** of rodent and primate genomes
- ➢ Genome size proportional to number of TEs

Class 1 (RNA intermediate) and 2 (DNA intermediate)

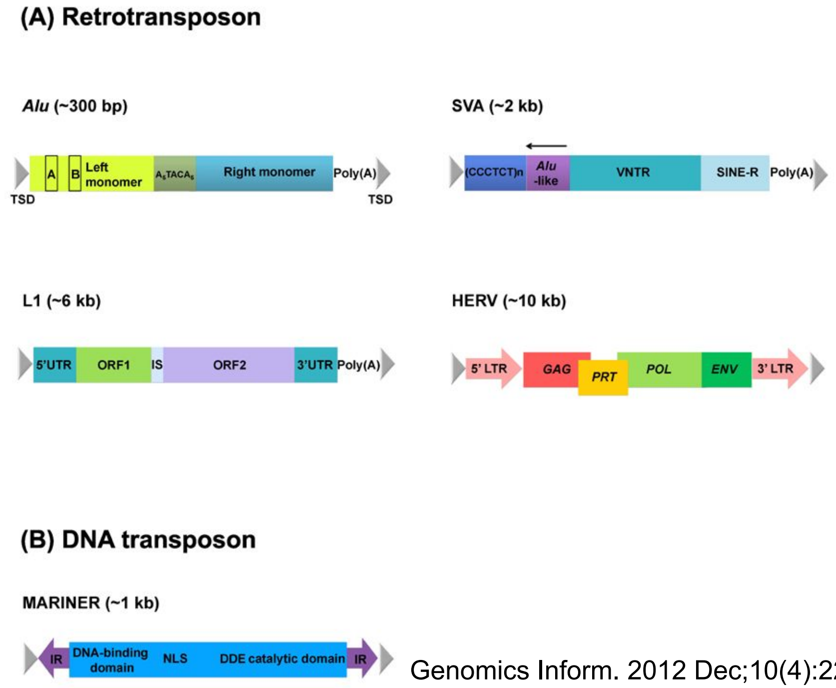
Potent genetic mutagens

- ➢ Disrupt expression of genes
- ➢ Genome reorganisation and evolution
- ➢ Transduction of flanking sequence

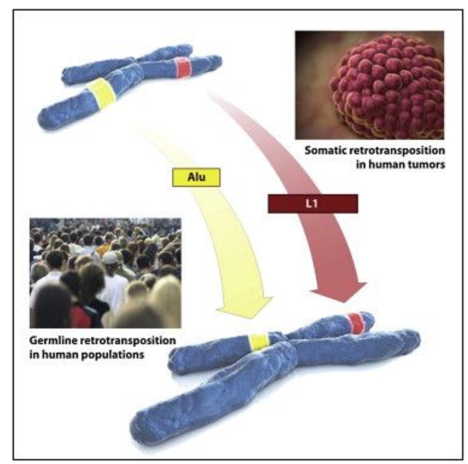
Species specific families

- ➢ Human: Alu, L1, SVA
- ➢ Mouse: SINE, LINE, ERV

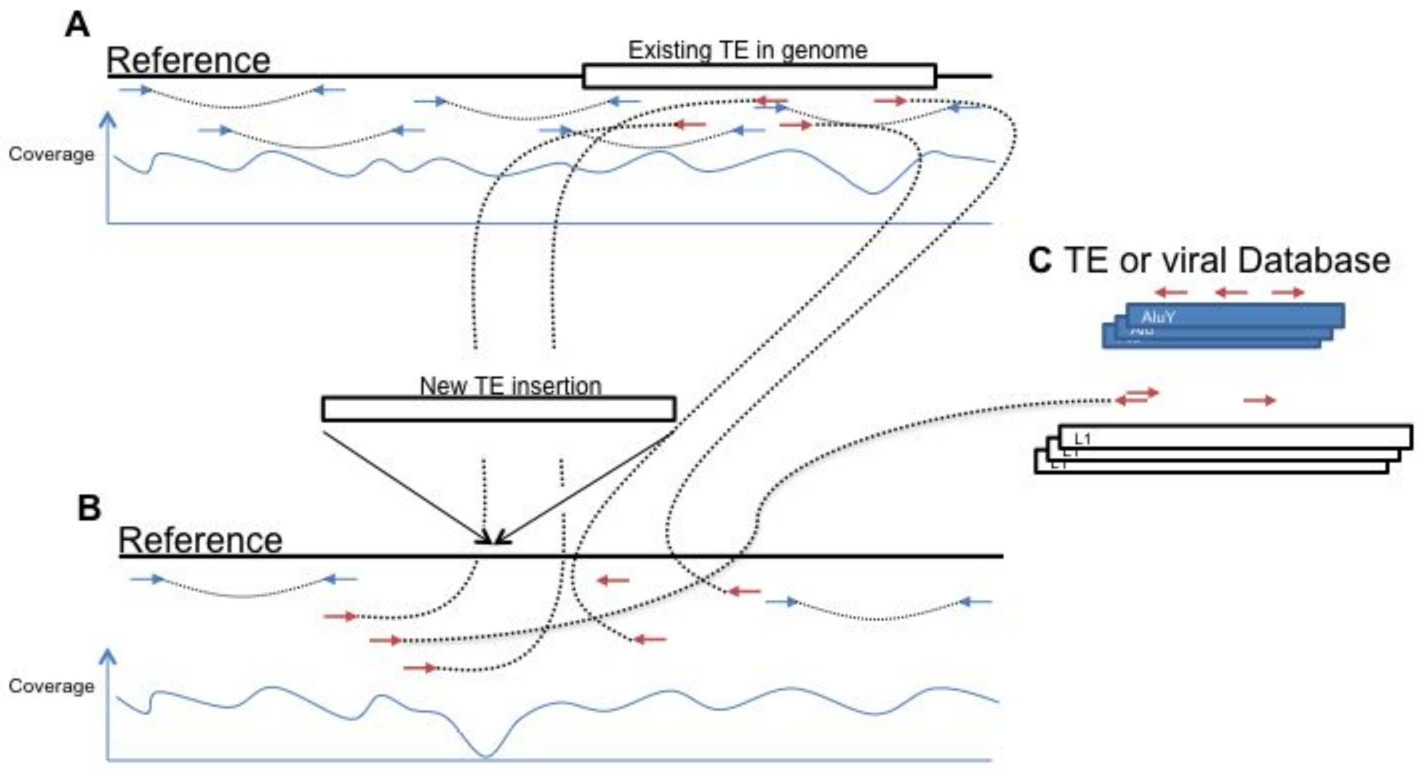
Many other families in other species



Genomics Inform. 2012 Dec;10(4):226-233



NGS and non-reference retrotransposition events



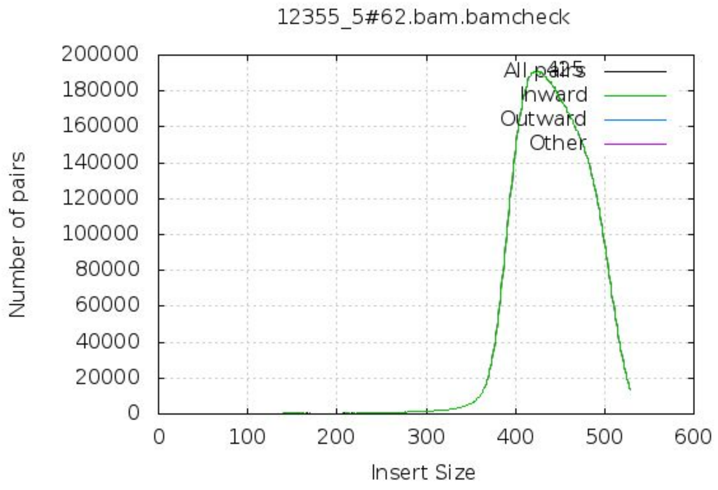
Sources of evidence 1: Read pairs

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



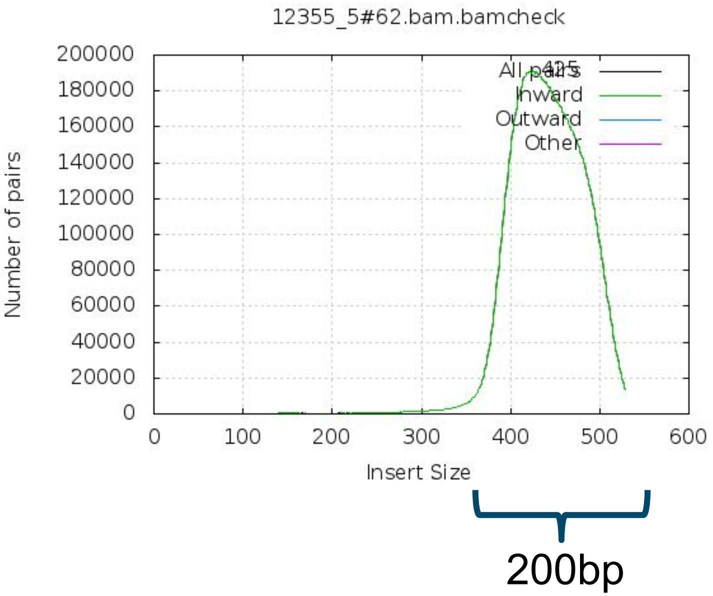
Sources of evidence 1: Read pairs

Several types of structural variations (SVs)

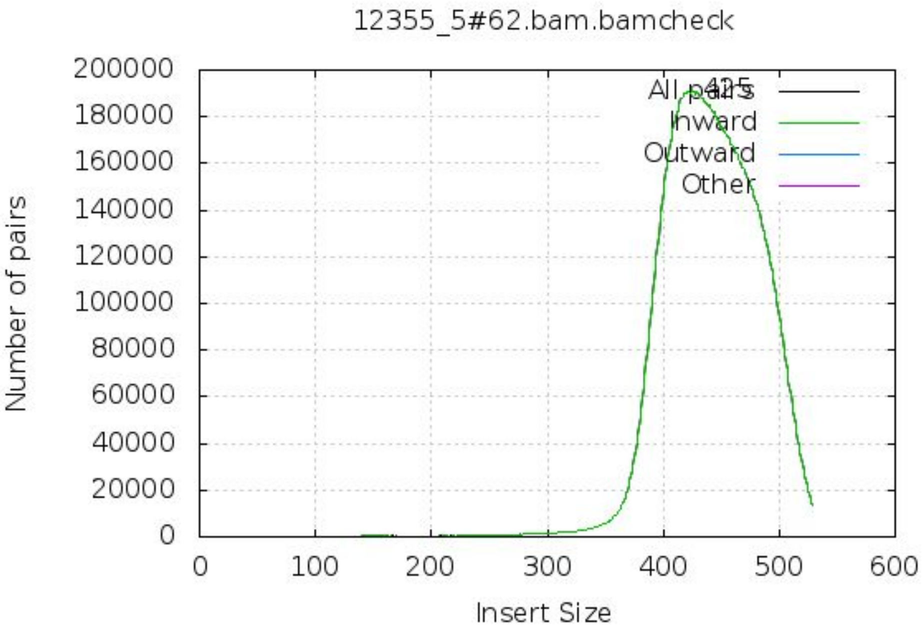
- ➢ Large Insertions/deletions
- ➢ Inversions
- ➢ Translocations

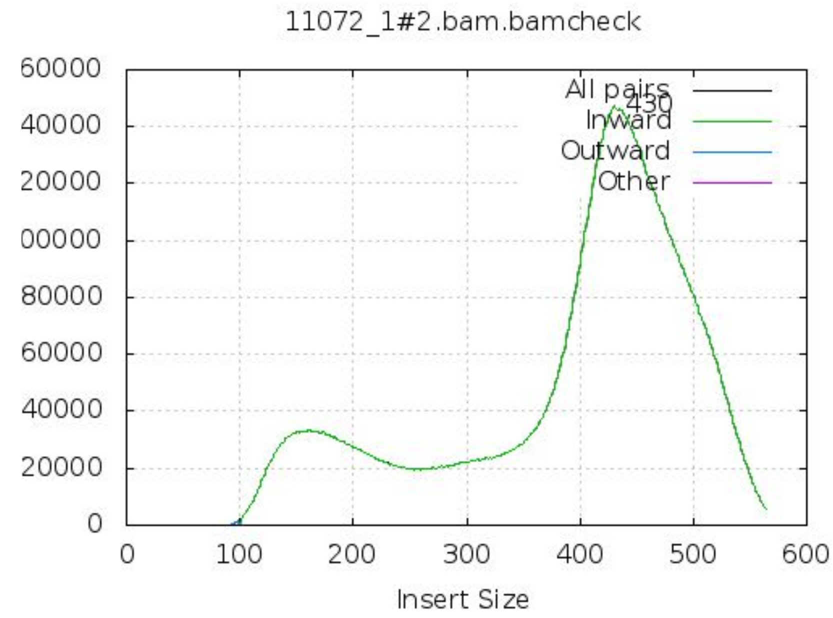
Read pair information used to detect these events

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



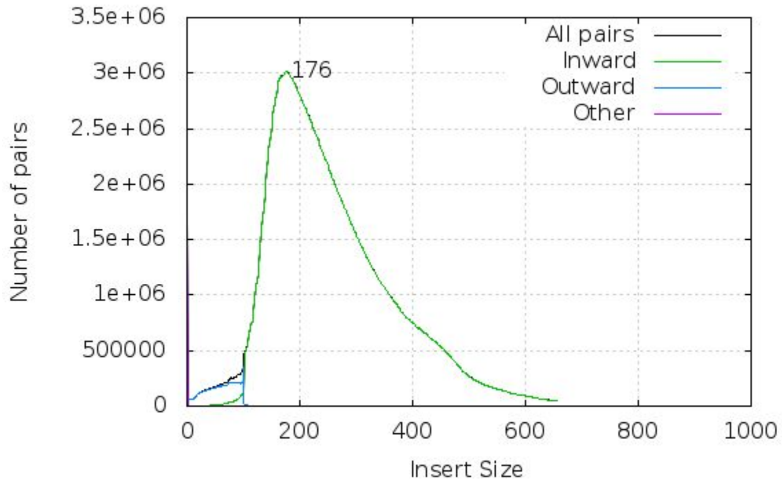
Fragment Size QC



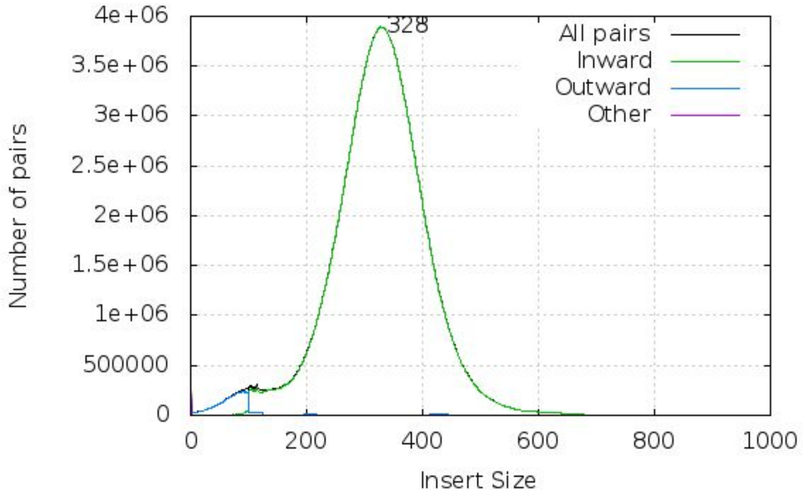


Fragment size again

DBA/1J

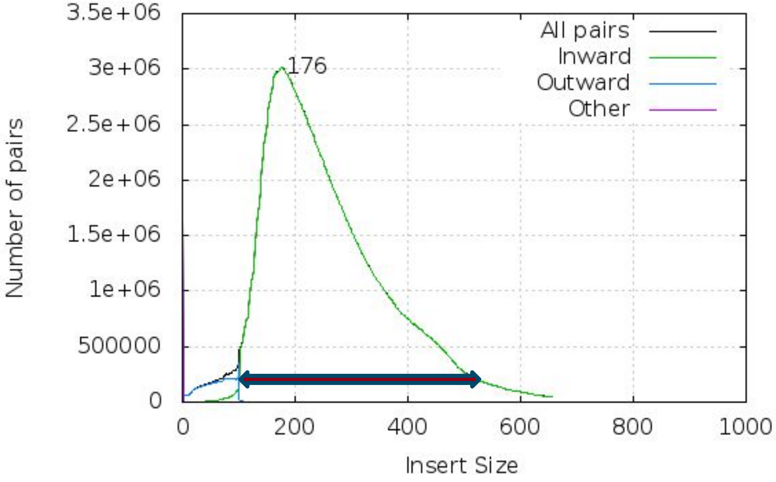


DBA/2J

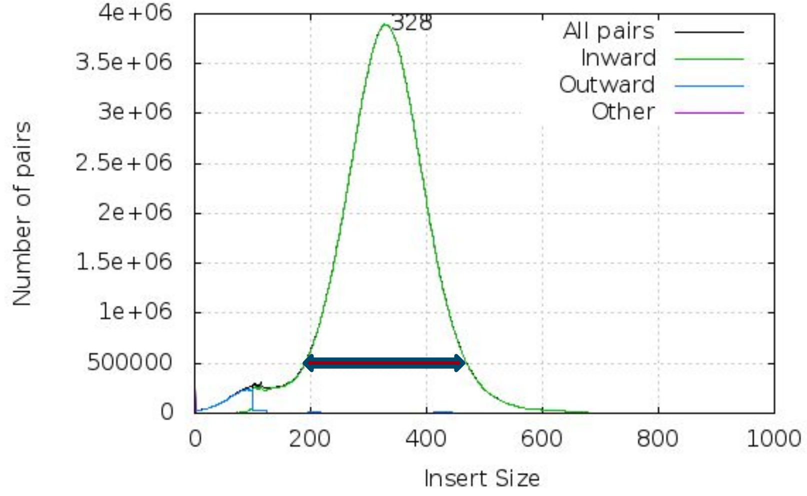


Fragment size again

DBA/1J



DBA/2J



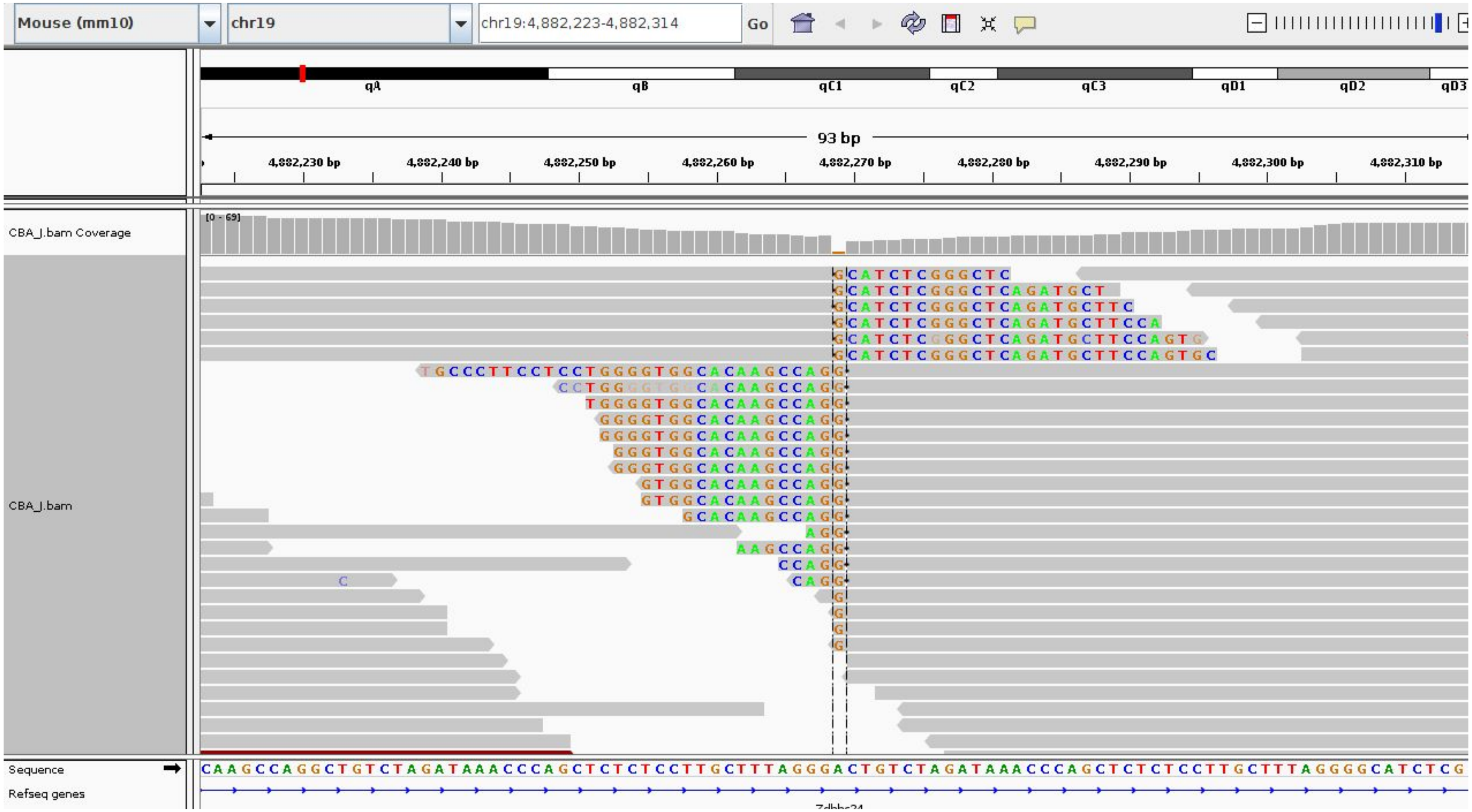
DBA/1J fragment size distribution has larger range (~450bp) vs. DBA/2J (~250bp)

SV caller only considers 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

Sources of evidence 2: 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

Sources of evidence 3: Read depth

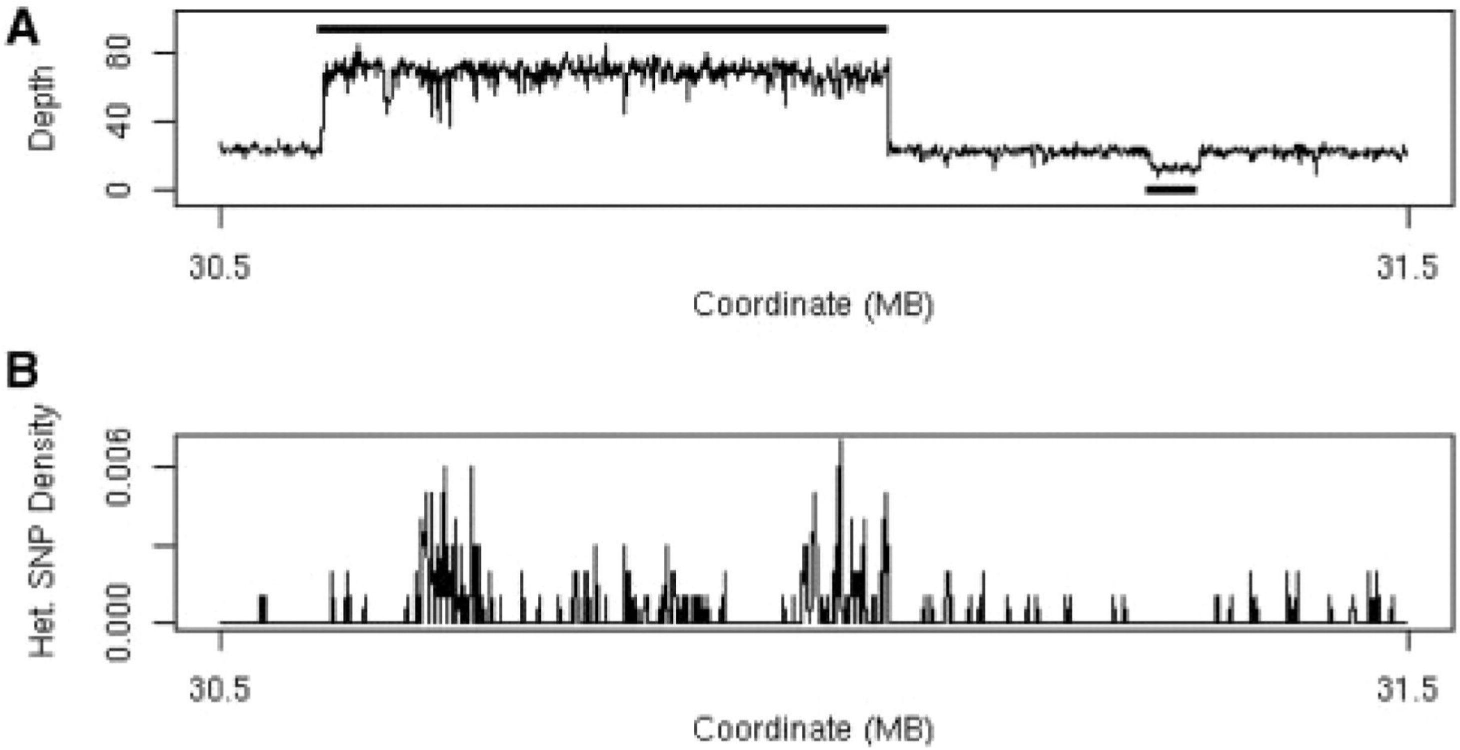


Fig. 1. (A) Plot of sequencing depth across a one megabase region of A/J chromosome 17 clearly shows both a region of 3-fold increased copy number (30.6–31.1 Mb) and a region of decreased copy number (at 31.3 Mb). The solid black line above the depth plot indicates the called copy number gain and the 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 et al. (2009) *Bioinformatics*

Read pairs: Breakdancer

- 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

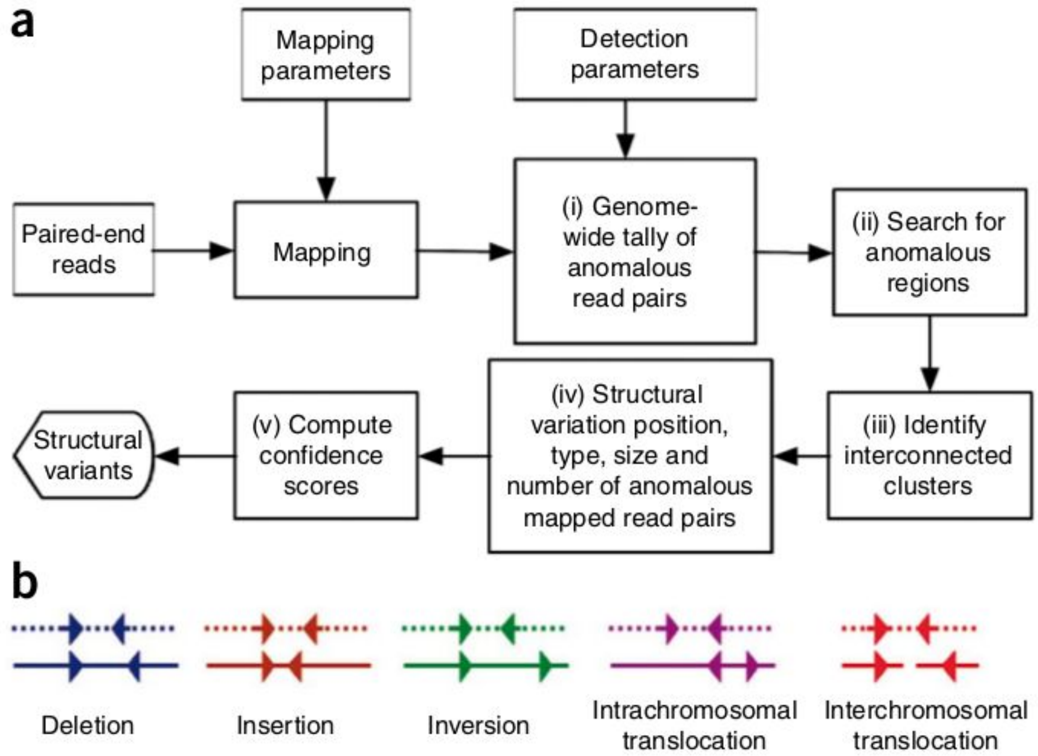


Figure 1 | Overview of BreakDancer algorithm. (**a**) The workflow. (**b**) Anomalous read pairs recognized by BreakDancerMax. A pair of arrows represents the location and the orientation of a read pair. A dotted line represents a chromosome in the analyzed genome. A solid line represents a chromosome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

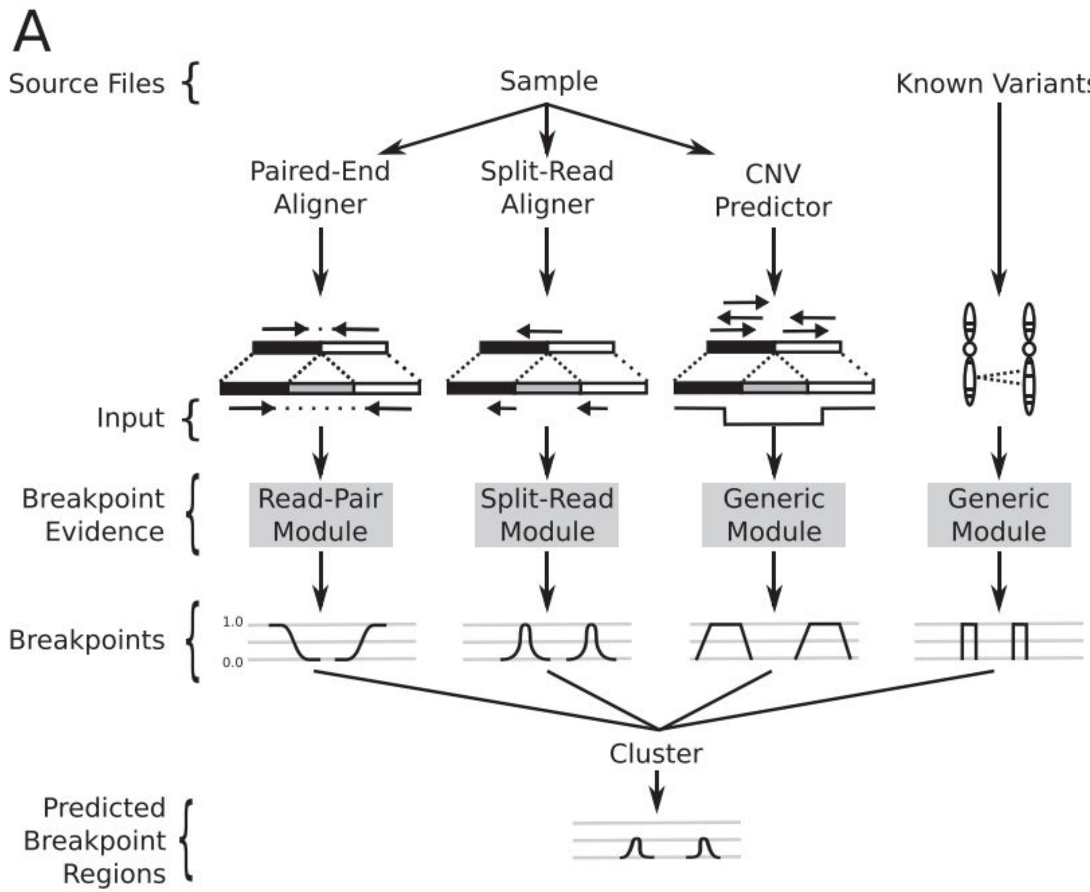
Chen *et al.* (2009) Nature methods

Read pairs + split reads + depth: Lumpy

- ➢ Any number of alignment signals can be integrated into a single discovery process
  - ○ Read pairs, split-reads, depth, user supplied evidence

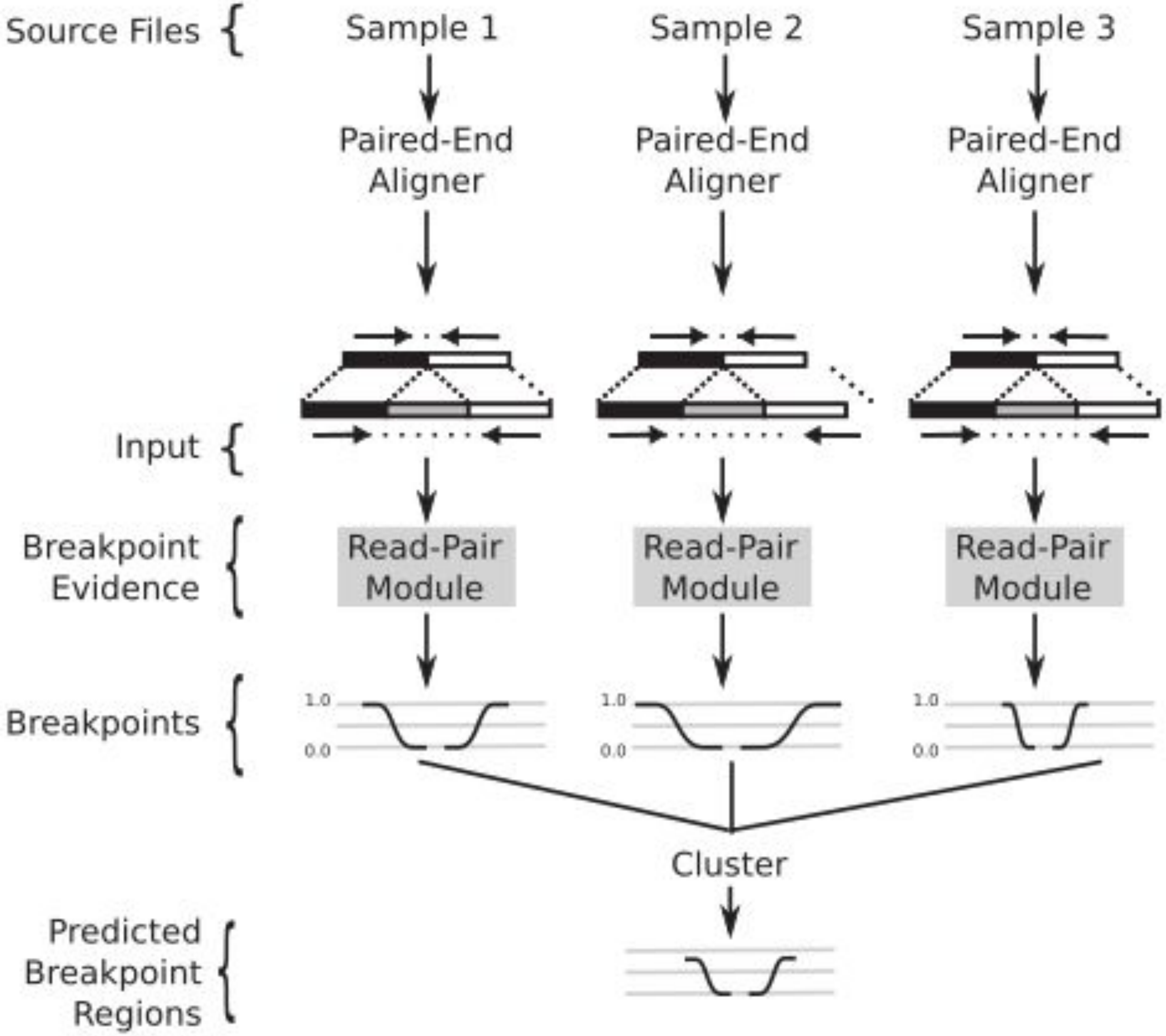
- ➢ Distinct modules that map signals from each alignment evidence type to the common probability interval pair.

- ➢ Evidence from the different alignment signals is mapped to breakpoint intervals, overlapping intervals are clustered and the probabilities are integrated



Layer *et al.* (2014) *Genome Biology*

Lumpy: multi-signal and multi-sample workflows



Layer *et al.* (2014) *Genome Biology*

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
```

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
```

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
```

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
```

- ➢ What does the CIEND info tag describe?
- ➢ How many different types of insertions can be described from the ALT tags?
- ➢ The first and second entries are both deletions, but what is the difference between them?

VCF for SVs

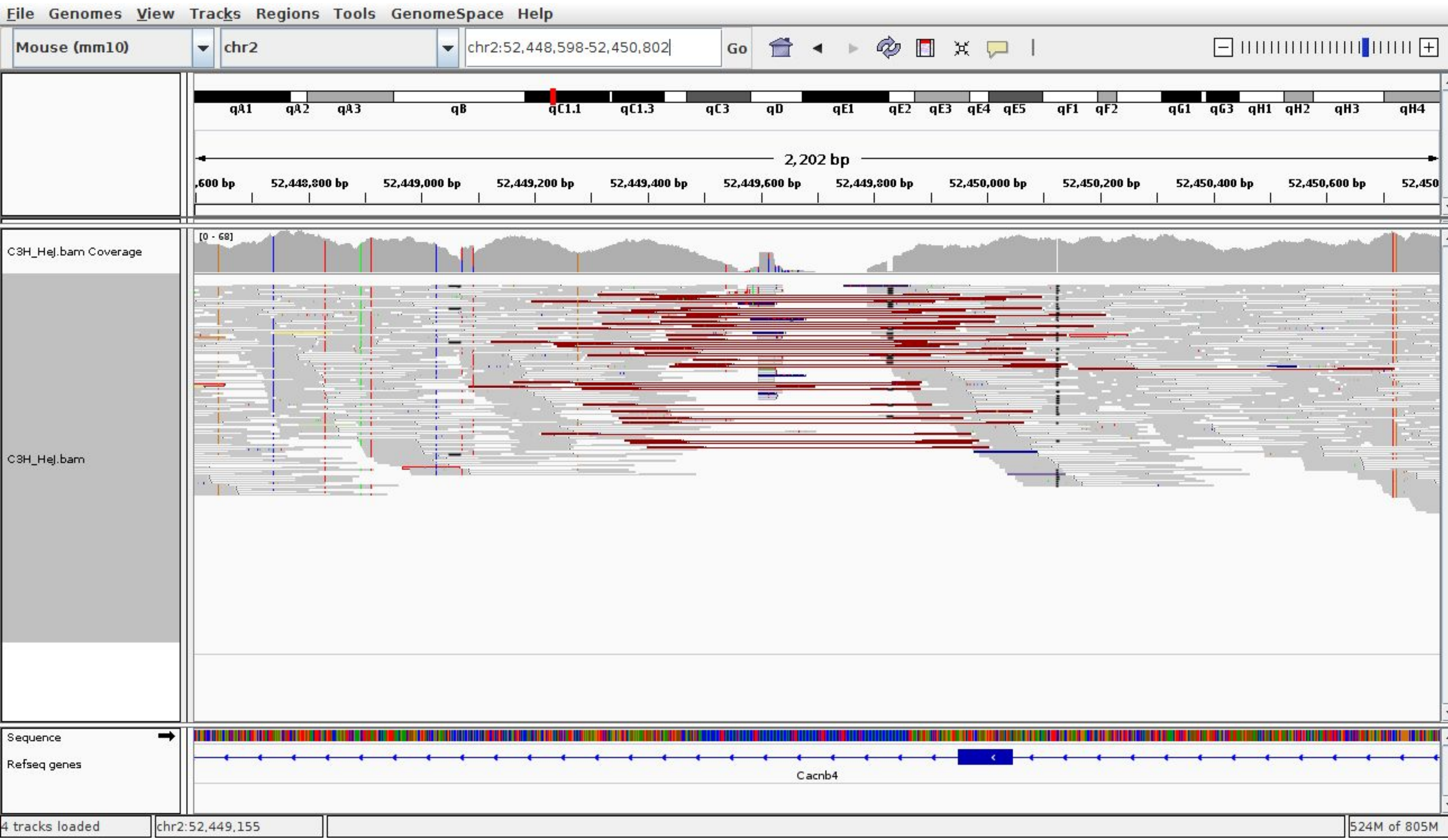
```
##fileformat=VCFv4.1
##fileDate=20100501
##reference=1000GenomesPilot-NCBI36
##assembly=ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/sv/breakpoint_assemblies.fasta
##INFO=<ID=BKPTID,Number=.,Type=String,Description="ID of the assembled alternate allele in the assembly file">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence 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
```

- Can you write out the VCF entry for a Alu insertion at chromosome 7, position 125467, of length 258bp, and a breakpoint confidence interval of +/20bp with one sample that is heterozygous for the insertion and has genotype quality of 40?

SV Visualisation

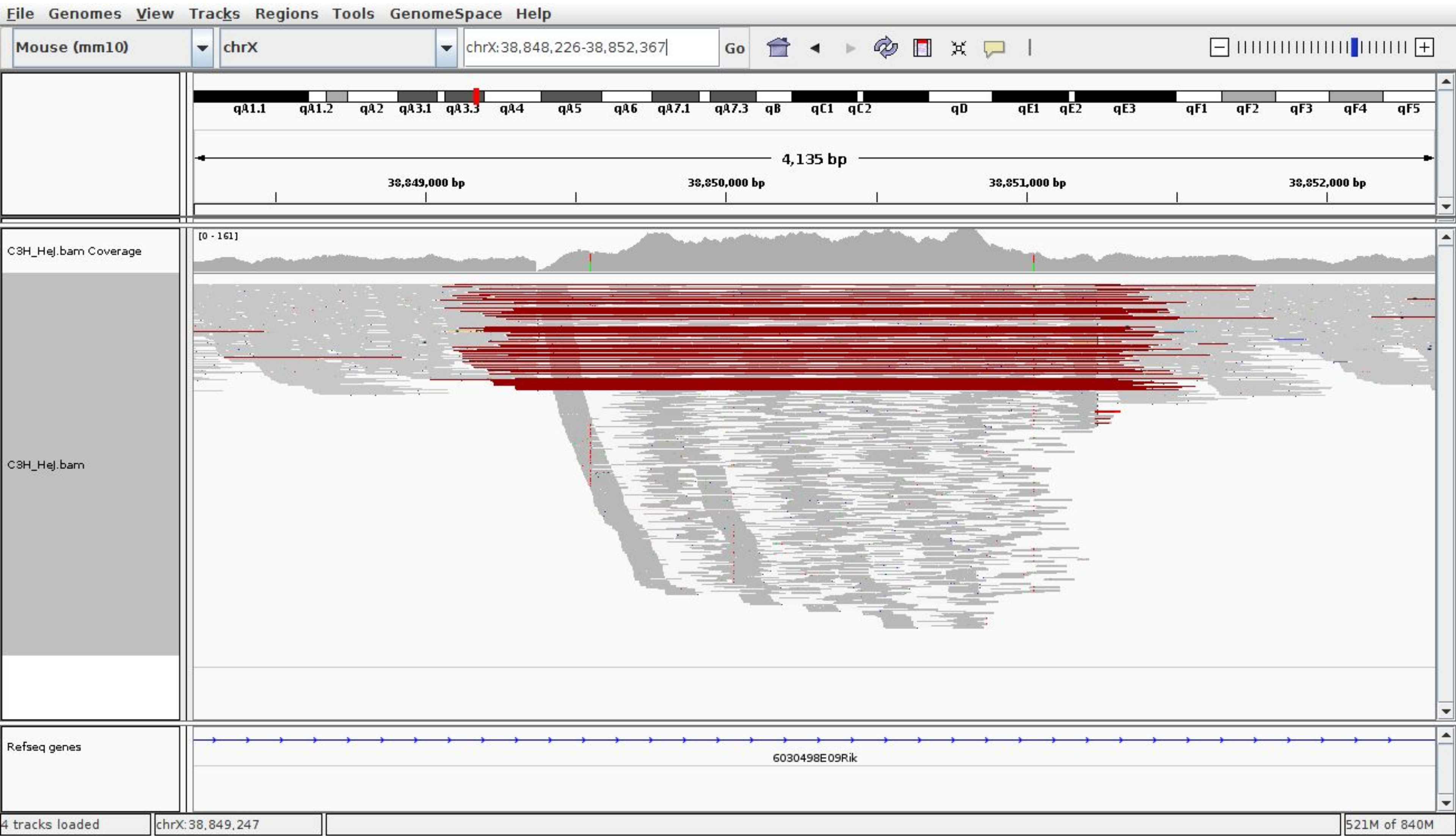
- Structural variation visualisation can be more challenging than SNPs and indels
- Inspect several hundred base pairs or multiple kbp
- Analyse complicated read pair patterns to determine type of SV and sources of error
- Look for soft clipped bases for breakpoint accuracy
- Many NGS visualisation software packages exist
- 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)

IGV - Deletion



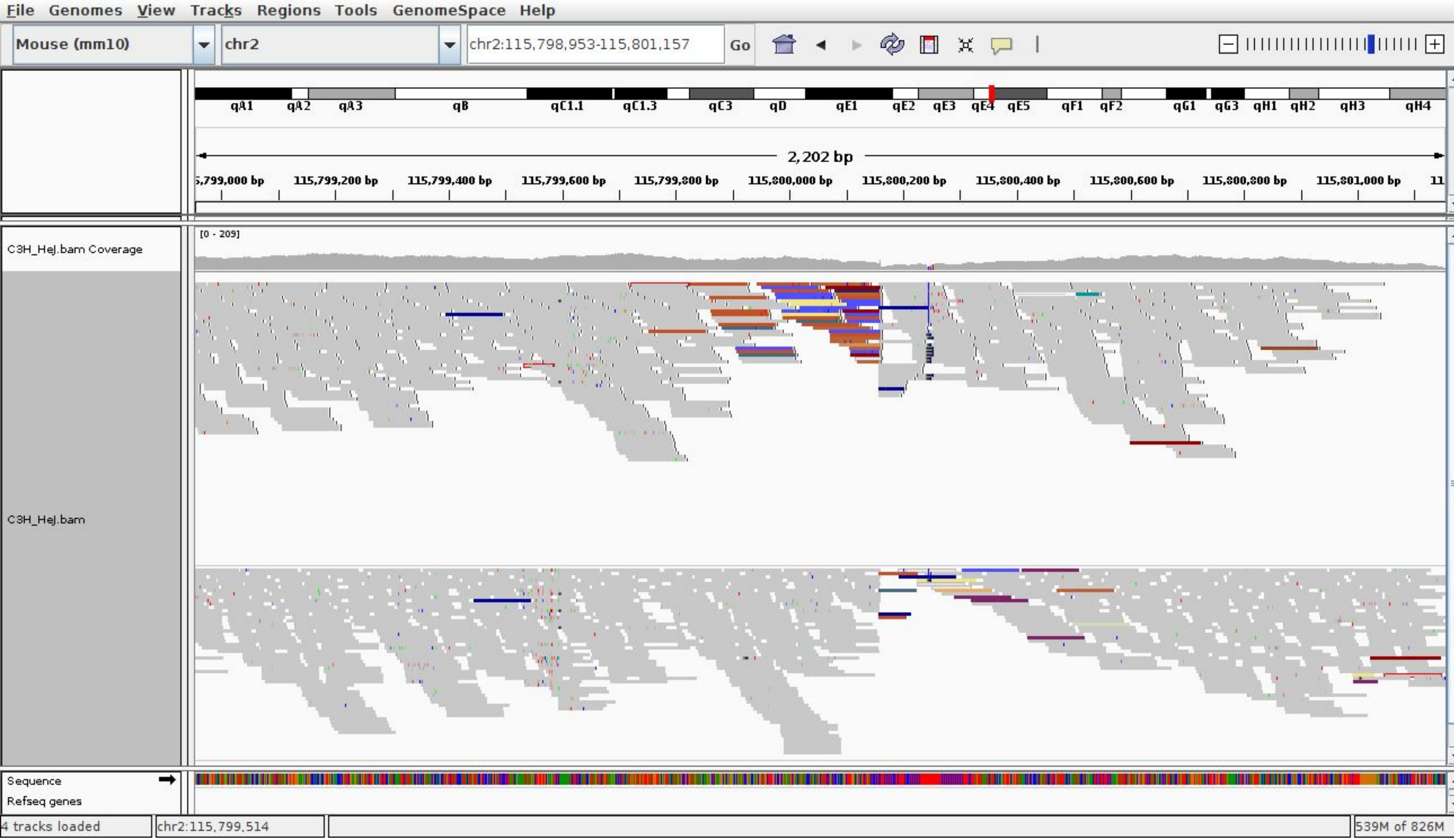
Right click - Squished

IGV - Repeat element deletion



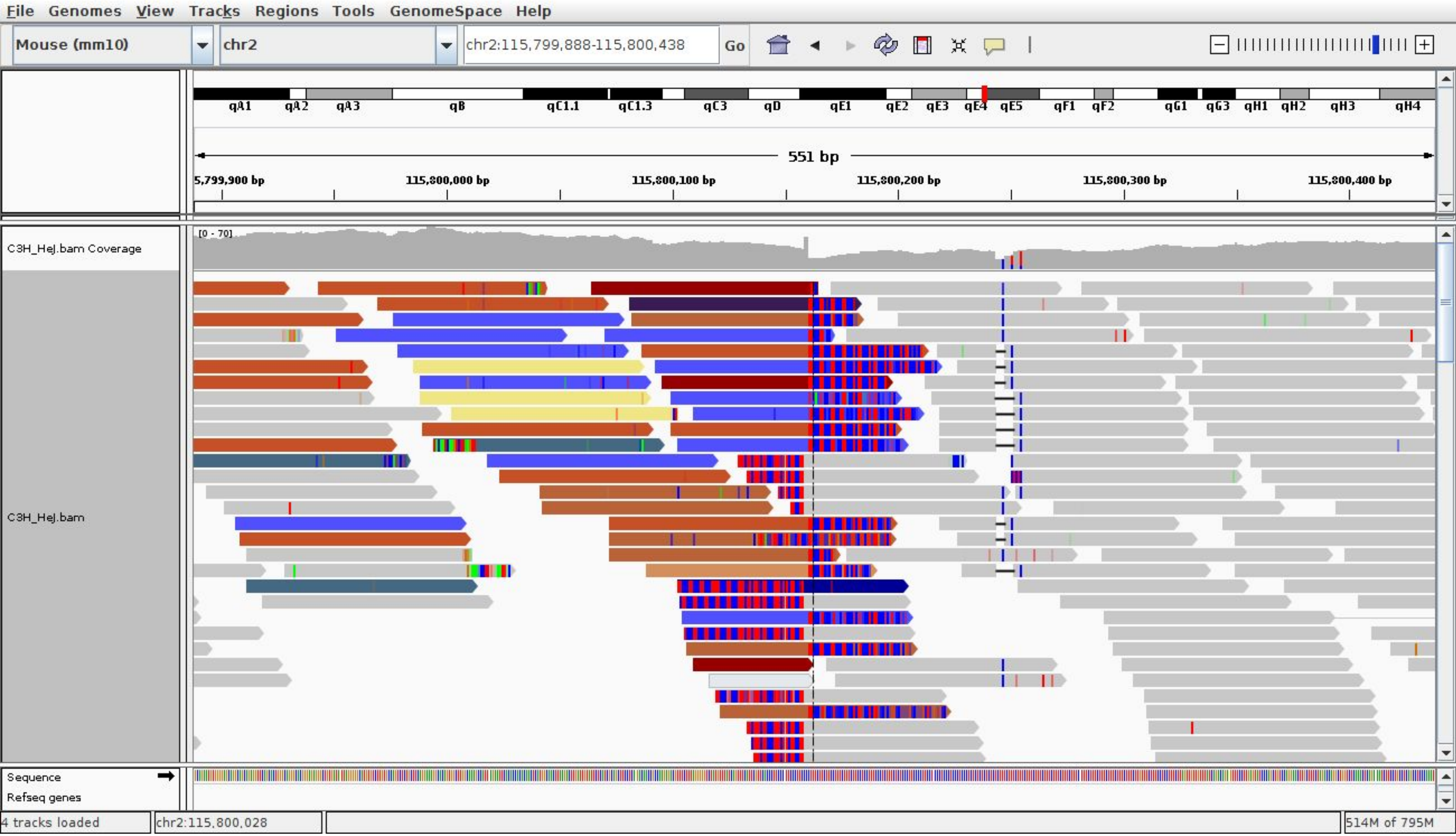
Right click - Squished

IGV - Insertion



Right click - group alignments by 'read strand'

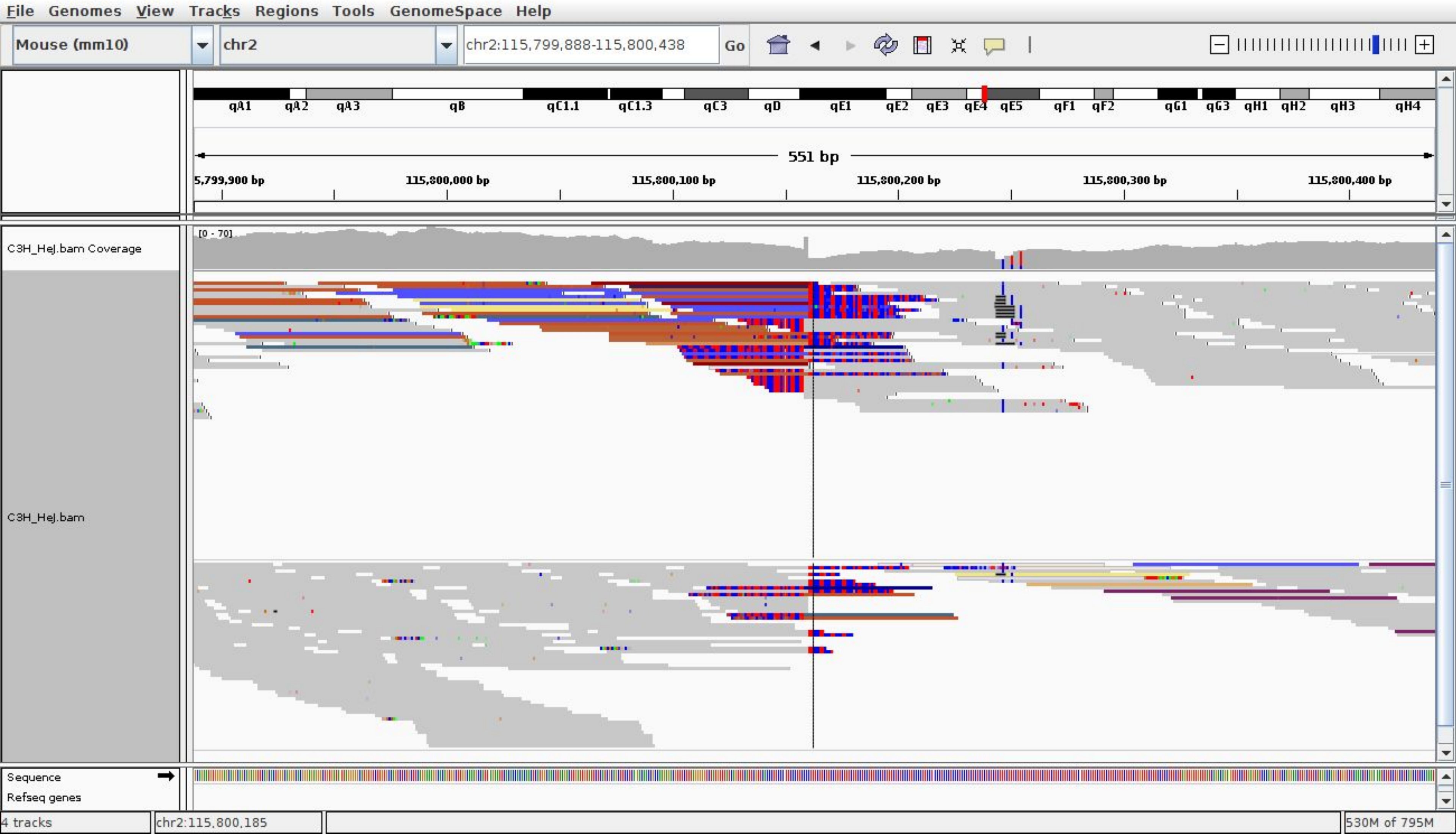
IGV - Insertion (zoomed)



Right click - view mismatch bases

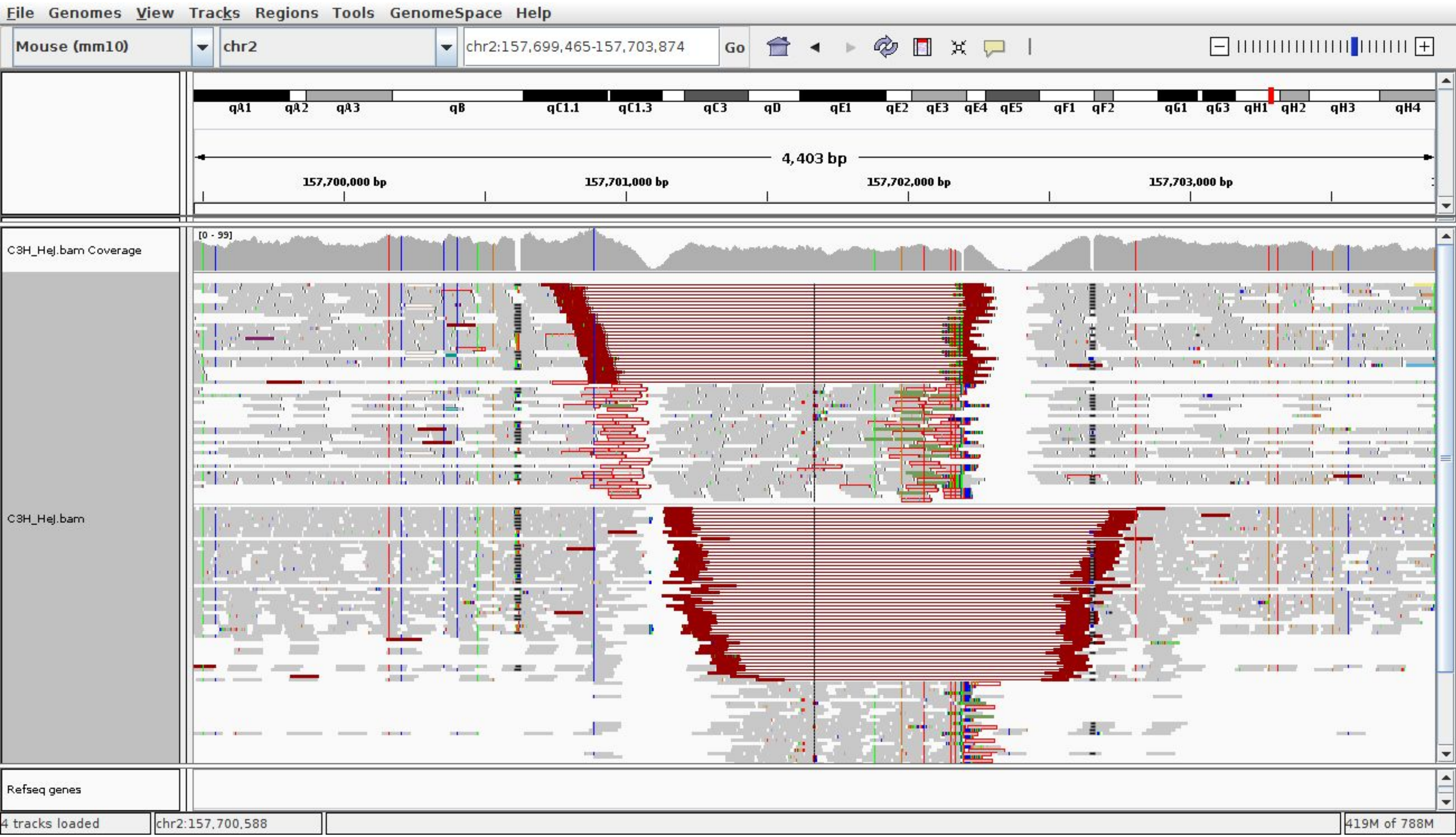
View - preferences - alignments - show soft clipped bases

IGV - Insertion (zoomed)



Right click - Squished

IGV - what is this?



- ➢ Right click - group alignments by - read strand
- ➢ Red lines are reads that are aligned further apart than expected

IGV - mouse over or click on a red read

C3H_HeJ.bam

Left alignment
Read name = HS4_07512:4:2201:4712:44023#7
Sample = C3H_HeJ
Read group = 7512_4#7

Location = chr2:157,700,909
Alignment start = 157,700,856 (+)
Cigar = 100M
Mapped = yes
Mapping quality = 46
Secondary = no
Supplementary = no
Duplicate = no
Failed QC = no

Base = T
Base phred quality = 38

Mate is mapped = yes
Mate start = chr2:157702197 (+)
Insert size = 1343
First in pair
Pair orientation = F1F2

MD = 32T67
RG = 7512_4#7
NM = 1
MQ = 56
AS = 95
XS = 63
CT = 1F100M1242T2F23S77M

Right alignment
Read name = HS4_07512:4:2201:4712:44023#7
Sample = C3H_HeJ
Read group = 7512_4#7

Location = chr2:157,700,909
Alignment start = 157,702,198 (+)
Cigar = 23S77M
Mapped = yes
Mapping quality = 56
Secondary = no
Supplementary = no
Duplicate = no
Failed QC = no

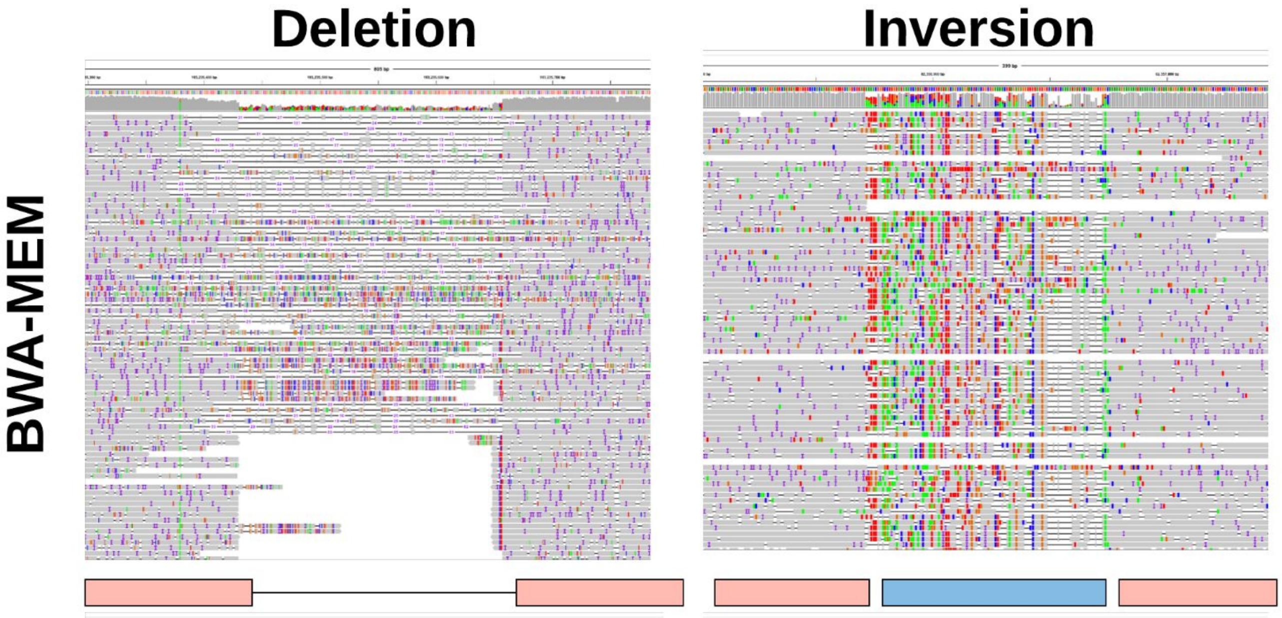
Mate is mapped = yes
Mate start = chr2:157700855 (+)
Insert size = -1343
Second in pair
Pair orientation = F1F2

MD = 52C24
RG = 7512_4#7
NM = 1
MQ = 46
AS = 72
XS = 41

SVs and long read sequencing

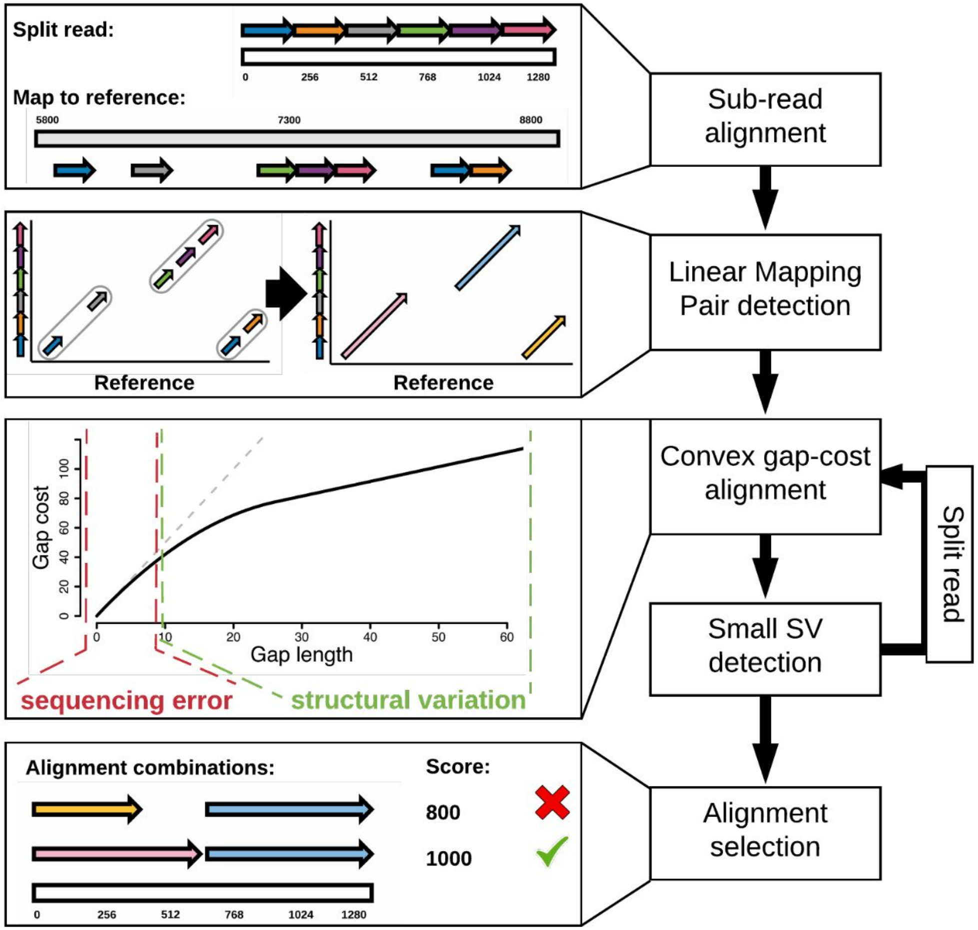
- Single molecule sequencing of large DNA fragments
- Platforms: Oxford nanopore and Pacific Biosciences
- Read lengths 10-20Kbp routinely
- Longer than most common transposable element repeats
- What does it mean for SV detection? Span both breakpoints with single read
- Some new challenges
  - Reads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

Alignment challenges

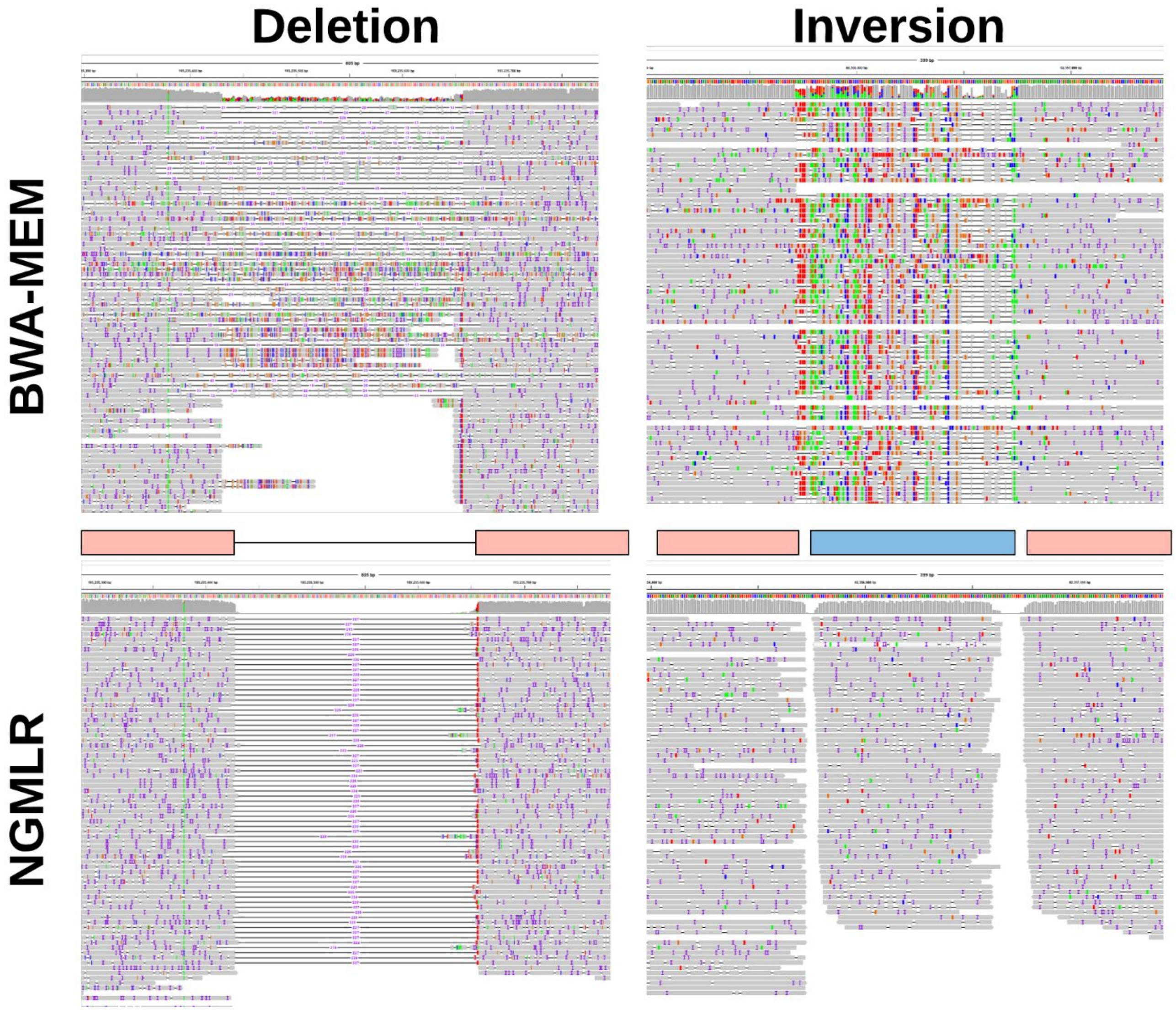


CoNvex Gap-cost alignMents for Long Reads (NGMLR)

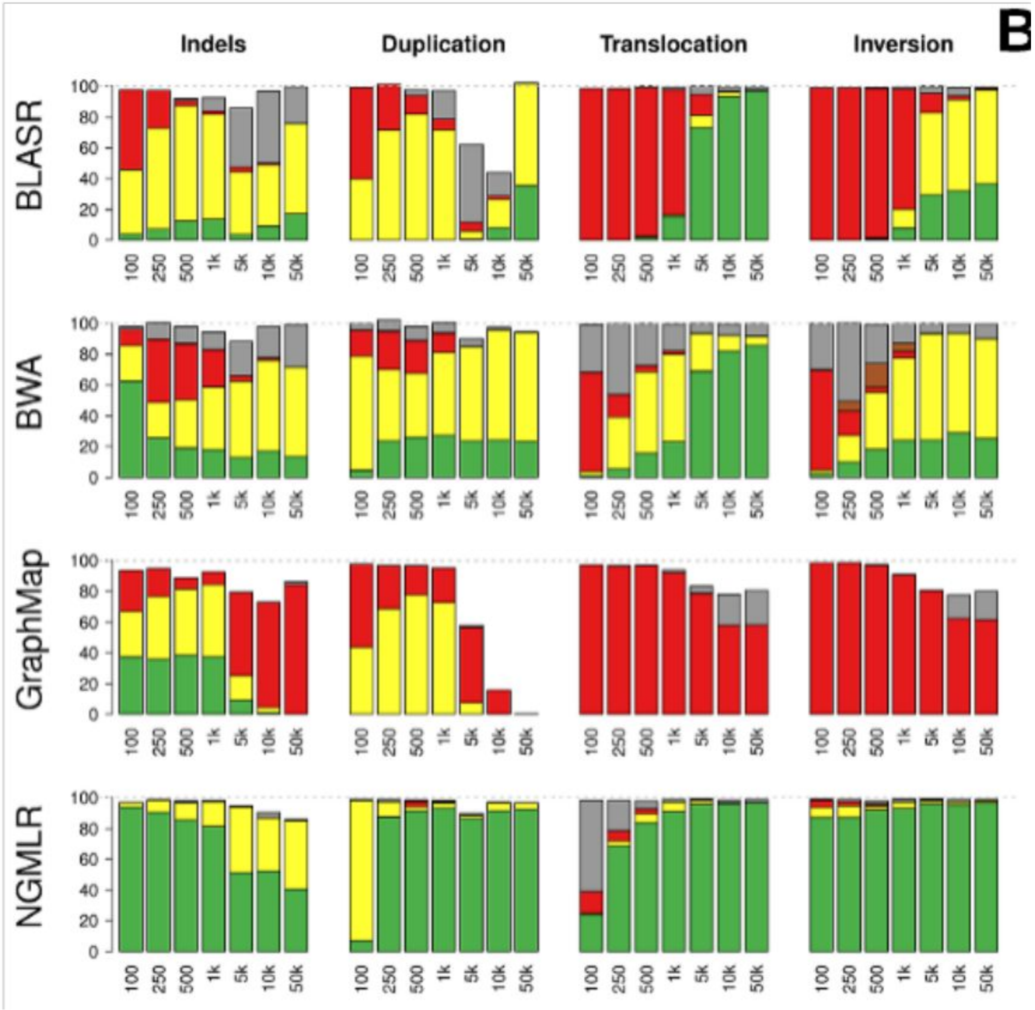
- NGMLR - aligner specifically designed for long reads
- Convex scoring model
  - Extending an indel is penalized proportionally less the longer the indel is



Alignment challenges



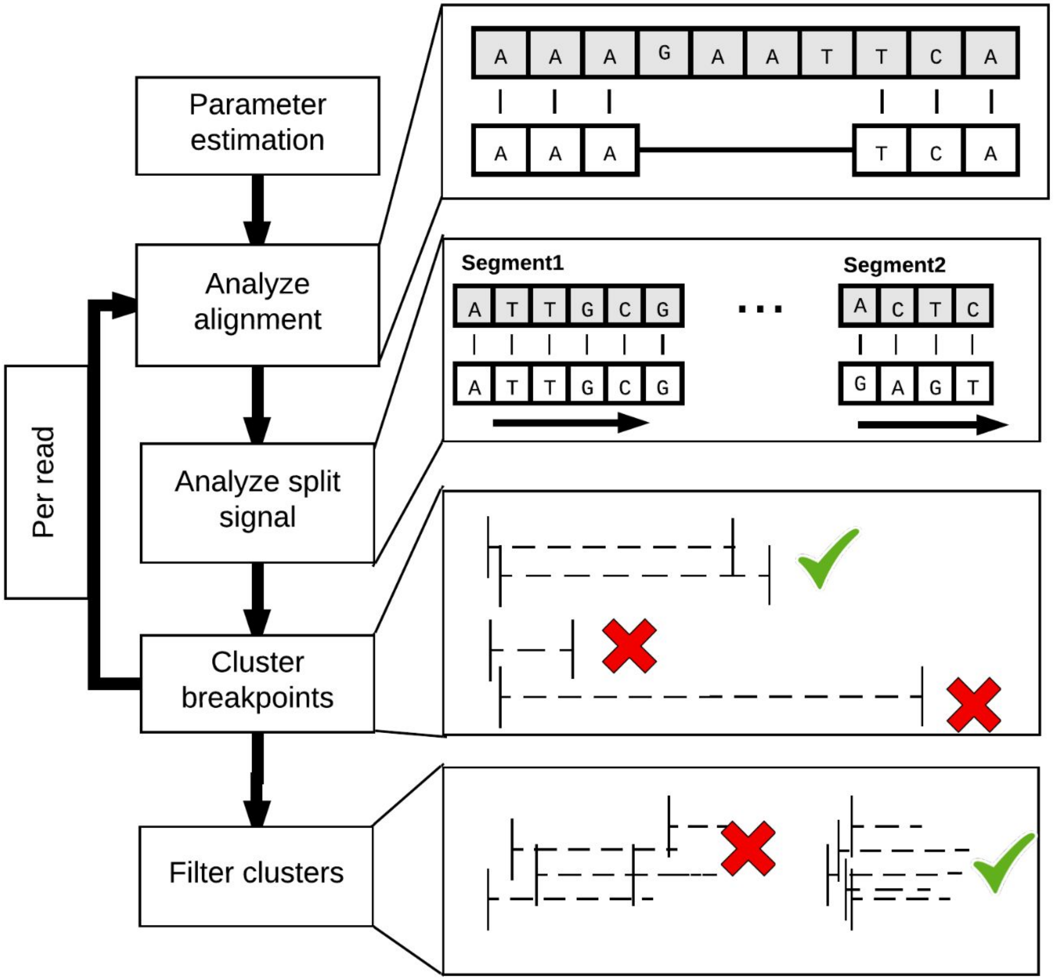
Comparison of aligners (simulated data)



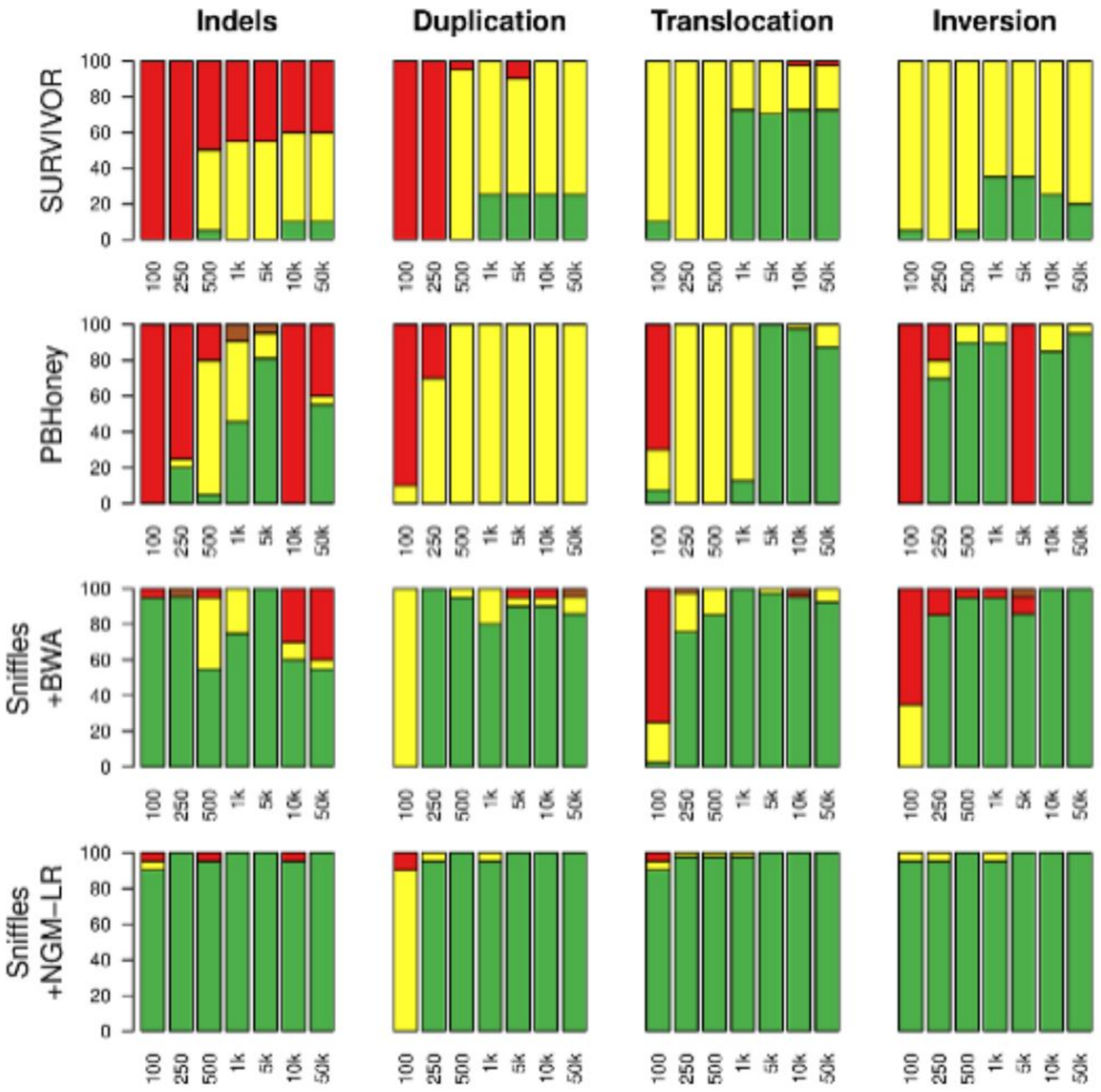
Alignment status: Precise (green), indicated (yellow), forced (red), unaligned reads (white), or trimmed but not aligned through the SV (grey).

Sniffles

SV detection from long read alignments



Sniffles performance



Complex SVs

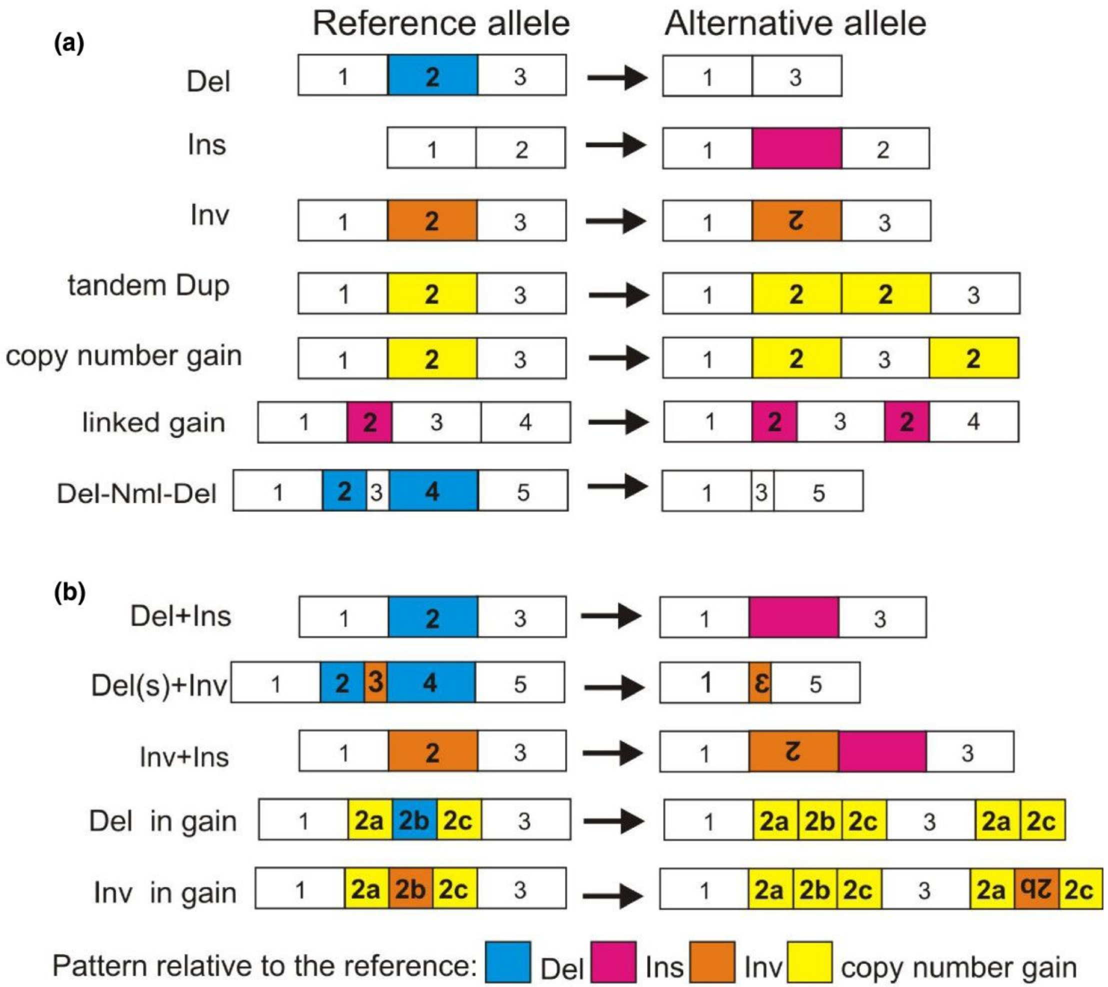


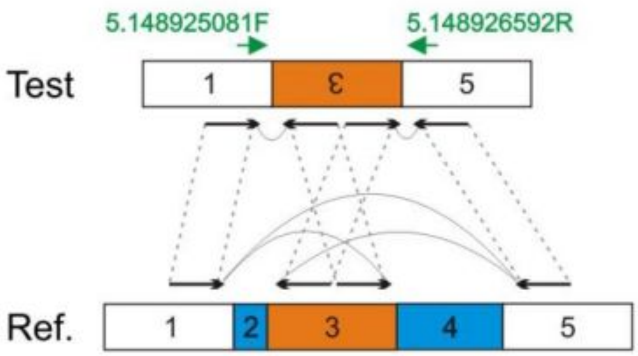
Figure 3. Architecture of structural variants. (a) Simple SVs: deletion (Del), insertion (Ins), inversion (Inv), tandem duplication (tandem Dup) and other types of copy number gains. Linked gain is a small copy number gain at close proximity to its copy. Inverted linked gain (not drawn) is similar to a linked gain but the copy is inverted. Del+Nml+Del is two deletions separated by a normal copy of small size. (b) Complex SVs: deletion co-occurring with insertion (Del+Ins), inversion with flanking deletions (Del(s)+Inv), inversion with insertion (Inv+Ins), deletion within a copy number gain (Del in gain) and inversion within a copy number gain (Inv in gain).

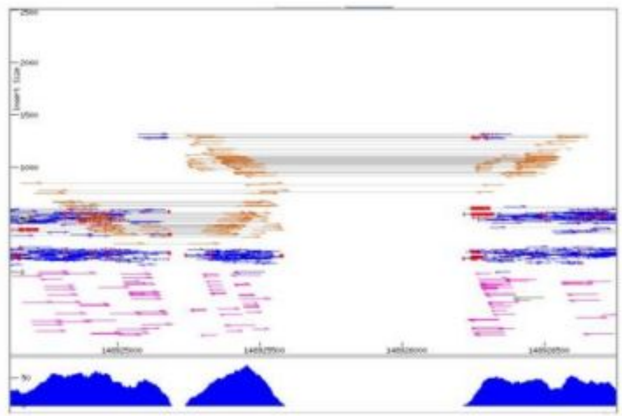
Yalcin *et al.* (2012) Genome Biology

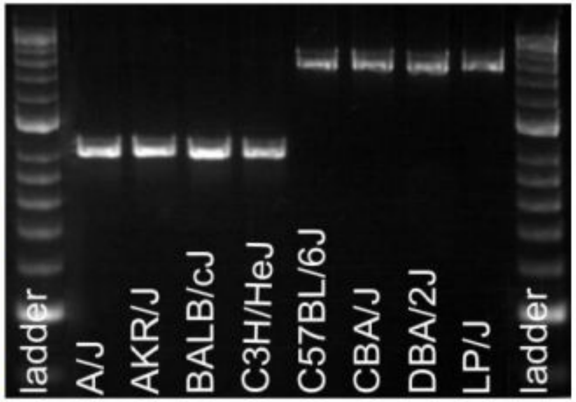
Complex SV Examples

H5

del-71bp_inv-325_del-645 [11110000]
1st del - chr5:148,925,178-148,925,248 bp
inv - ch5:148,925,249-148,925,573 bp
2nd del - chr5:148,925,574-148,926,218 bp

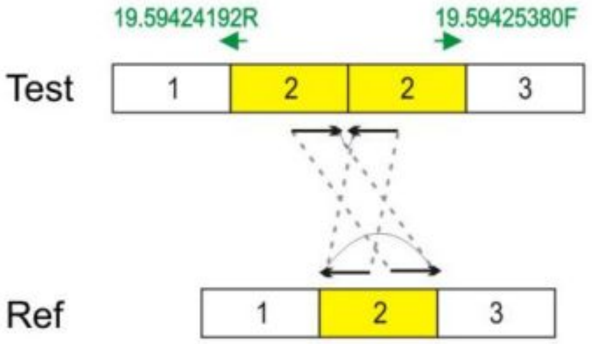


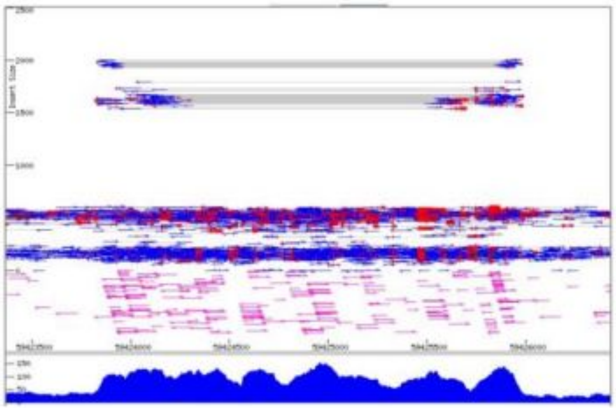


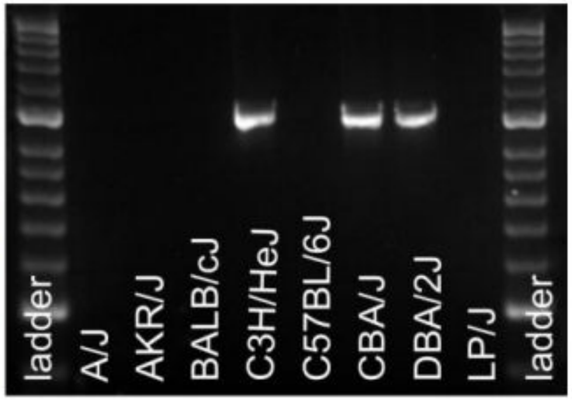


H8

Tandem duplication of 2181 bp [00010110]
chr19:59,423,833-59,425,976 bp

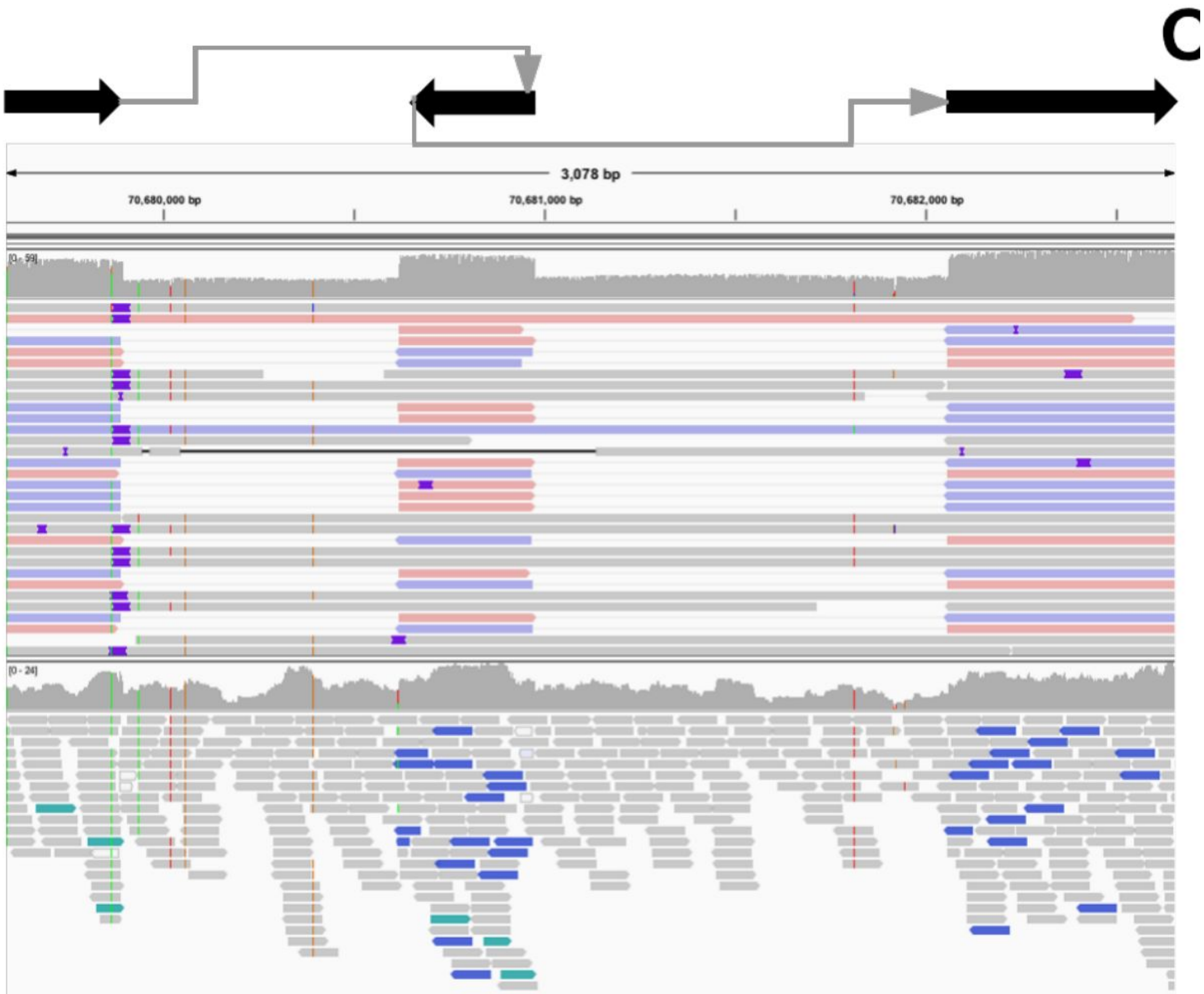






Yalcin *et al.* (2012) Genome Biology

Complex SVs - Long reads



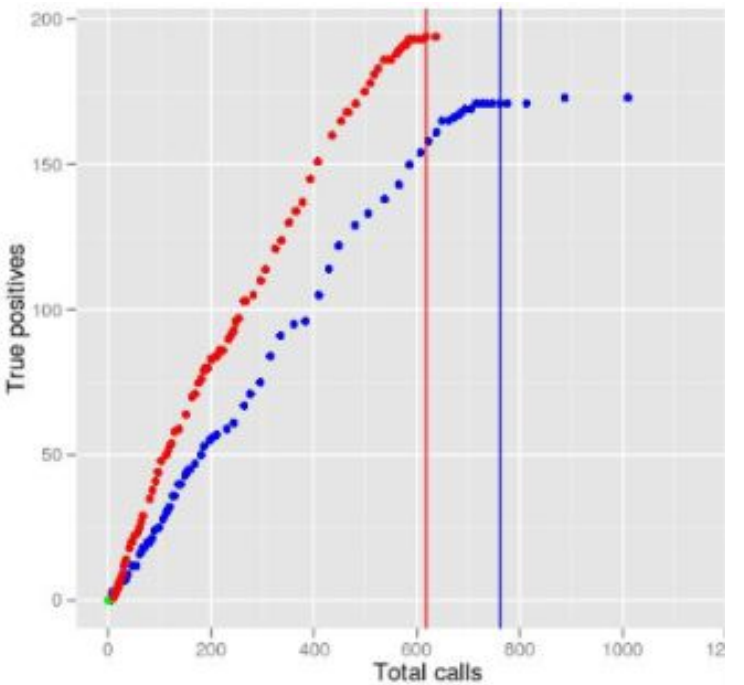
NGLMR

Illumina (BWA-mem)

3kb region: two deletions flanking an inverted sequence

Evaluating SV calls

- ➢ Specificity vs sensitivity
- ➢ False positives vs. false negatives
- ➢ Desirable to have high sensitivity and specificity
- ➢ How to determine sensitivity?
  - ○ External sources of true/known SVs
- ➢ Specificity
  - ○ Validate a random selection of SVs by another technology
  - ○ e.g. PCR products
- ➢ Receiver operator curves to investigate effects of varying parameters



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.