Structural Variants

Structural variations

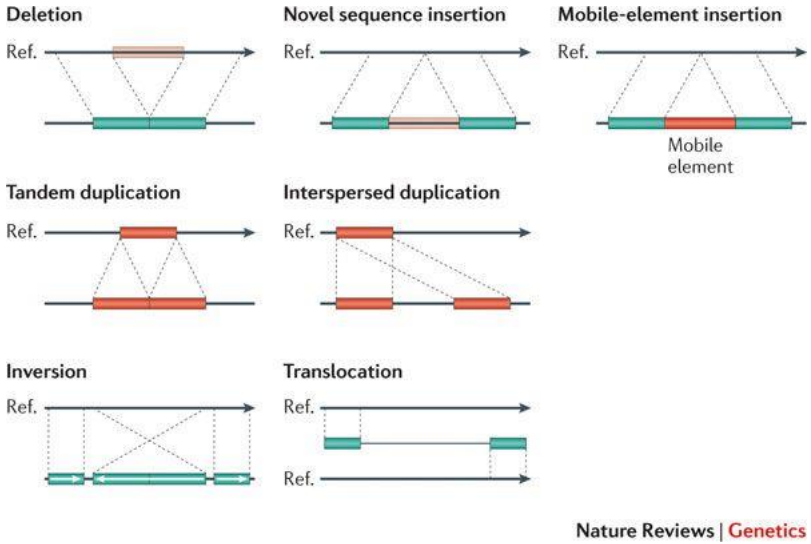
Structural genomic events > 50 bp

CNVs, but also structural rearrangements

Common in human genomes in normal population

Major cause of phenotypic variation

Common in some diseases, cancer +++



Growing effects on rare disease, autism, schizophrenia

How to detect Structural variations ?

Detection of genomic rearrangements

Read depth signals

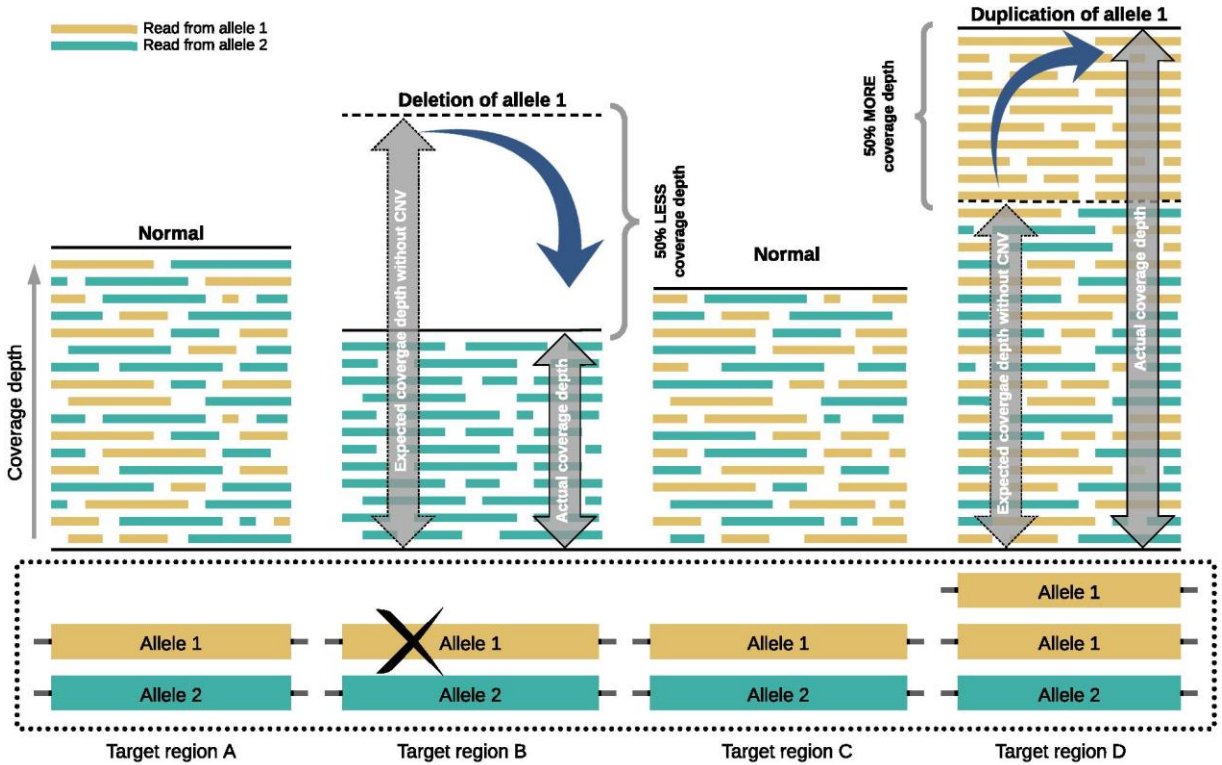
Mate-pair or paired-end mapping abnormalities

Split-Read alignments

Read depth signals

Mate-pair or paired-end mapping abnormalities

Split-Read alignments



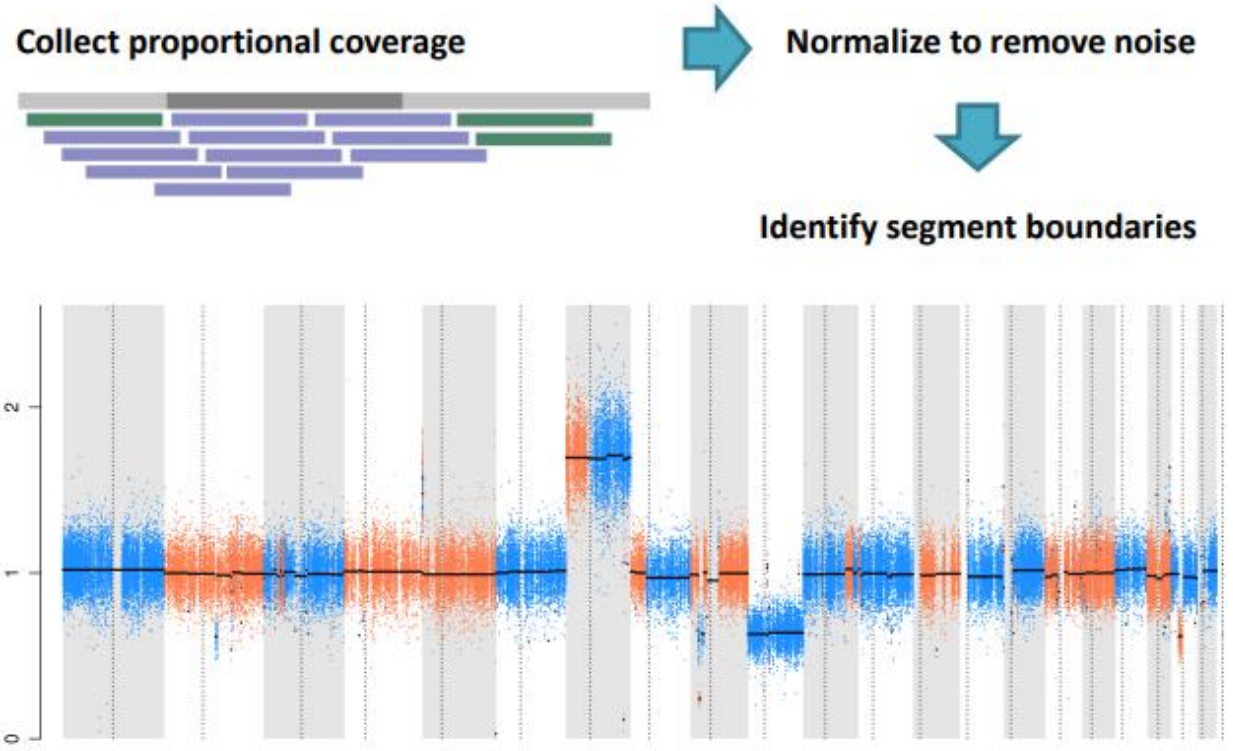
Singh, A.K., Olsen, M.F., Lavik, L.A.S. et al. Detecting copy number variation in next generation sequencing data from diagnostic gene panels. BMC Med Genomics

Read depth signals

Mate-pair or paired-end mapping abnormalities

Split-Read alignments

Copy number: it's all about coverage and normalization



from Introduction to Somatic Variant Discovery, GATK Best Practices for Variant Discovery

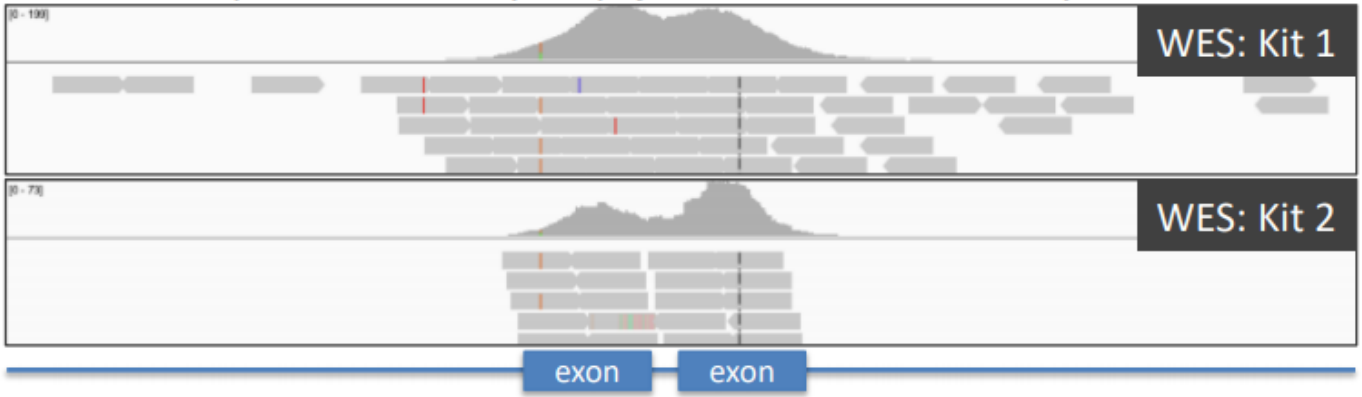
Read depth signals

Mate-pair or paired-end mapping abnormalities

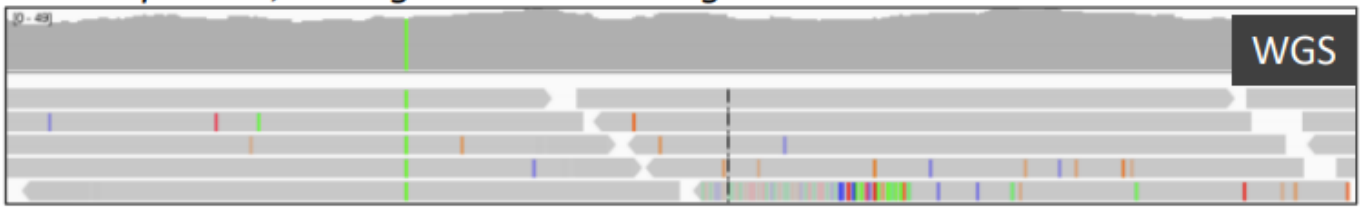
Split-Read alignments

Coverage is variable across WES targets and kits

WES bait-capture and library amplification add to variability.



In comparison, WGS gives even coverage.



from Somatic Copy Number Alterations, GATK Best Practices for Variant Discovery

Read depth signals

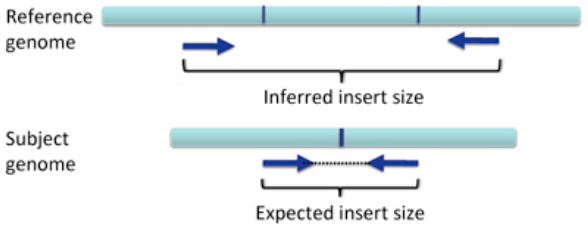
Mate-pair or paired-end mapping abnormalities

Split-Read alignments

Paired reads can yield evidence of structural events.

Alignment coloring options help highlight these events based on:

- Inferred insert size (template length)
- Pair orientation (relative strand of pair)



https://www.igv.org/workshops/March2017/SlidesHandouts/IGV_SlideDeck.pdf

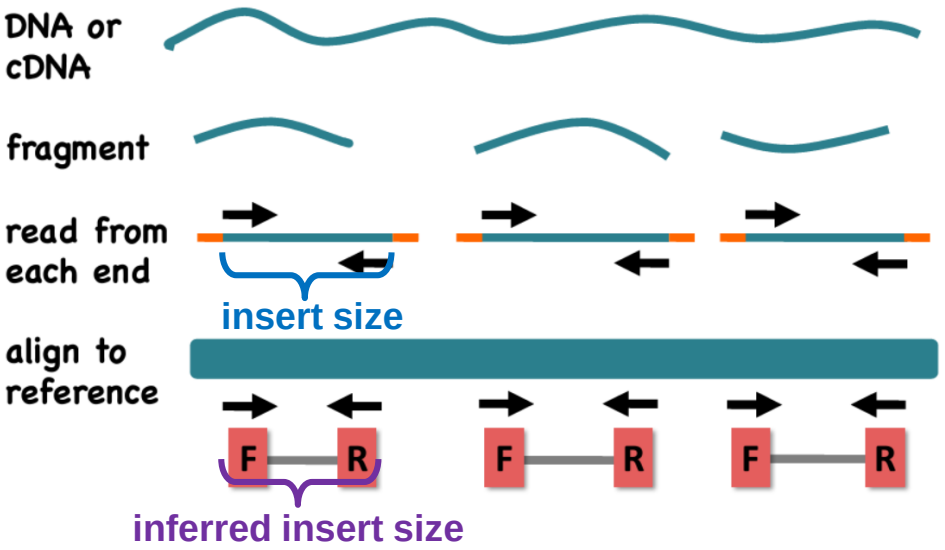
Read depth signals

Mate-pair or paired-end mapping abnormalities

Split-Read alignments

Differences between **insert size** and **inferred insert size** can be used to detect evidence of structural variants, including :

- Deletions
- Insertions
- Inter-chromosomal rearrangements

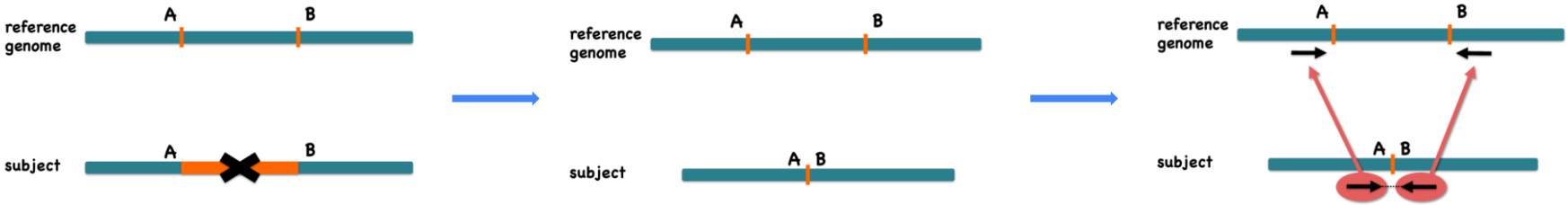


Read depth signals

Mate-pair or paired-end mapping abnormalities

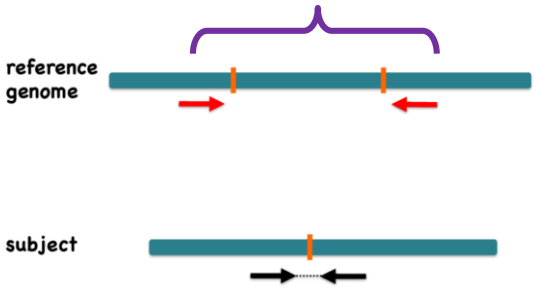
Split-Read alignments

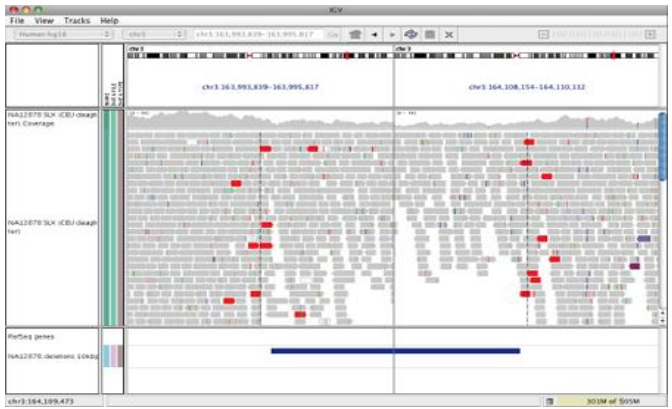
DELETION



inferred insert size

greater than expected value





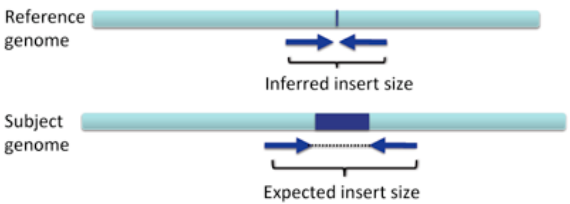
http://software.broadinstitute.org/software/igv/interpreting_insert_size

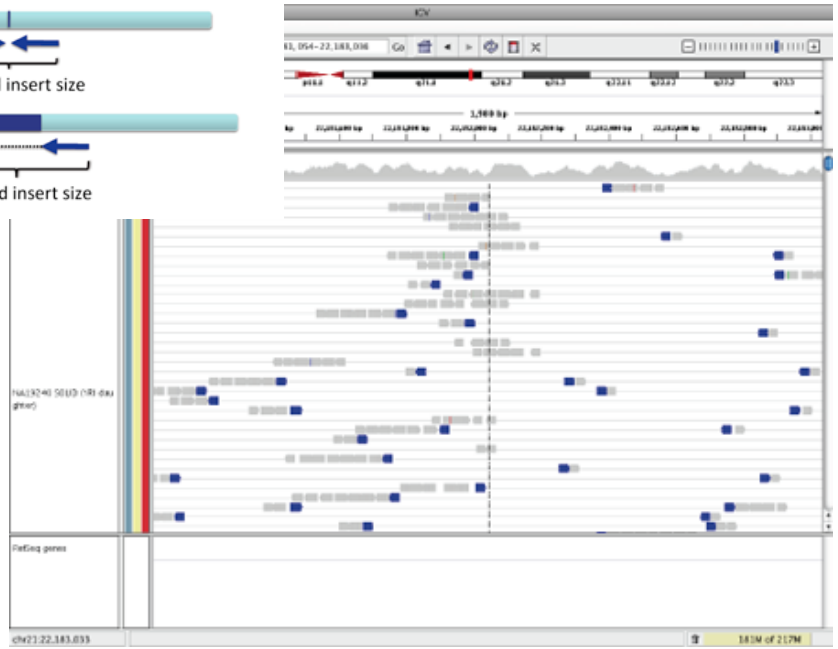
Read depth signals

Mate-pair or paired-end mapping abnormalities

Split-Read alignments

inferred insert size smaller than expected value (possible evidence of an insertion)





INSERTION

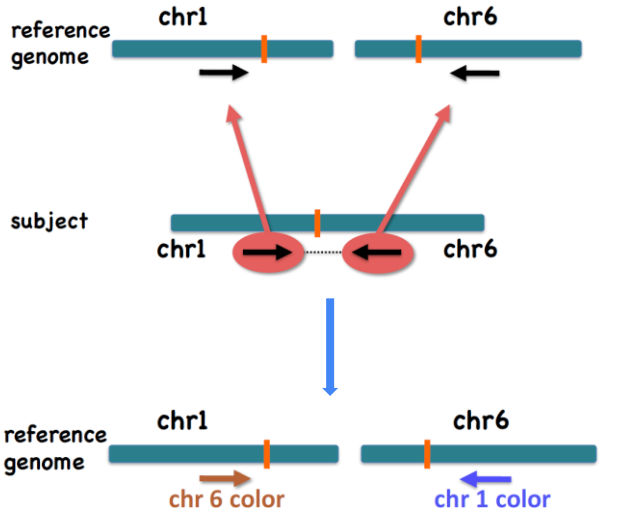
http://software.broadinstitute.org/software/igv/interpreting_insert_size

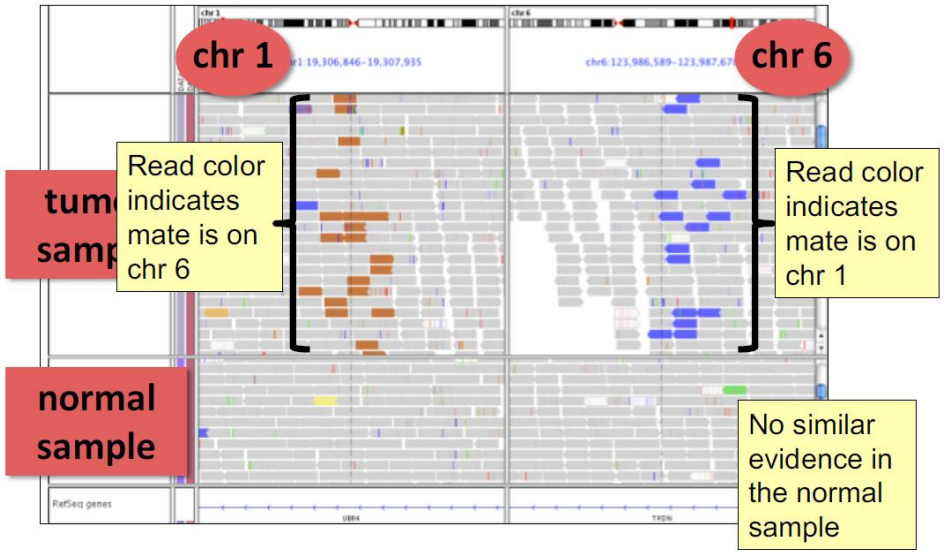
Read depth signals

Mate-pair or paired-end mapping abnormalities

Split-Read alignments

Inter-chromosomal fusion



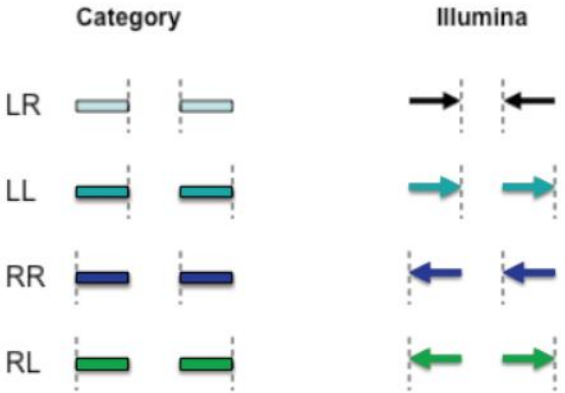


Read depth signals

Mate-pair or paired-end mapping abnormalities

Split-Read alignments

IGV Interpreting Color by Pair Orientation



LR	Normal reads. The reads are left and right (respectively) of the unsequenced part of the sequenced DNA fragment when aligned back to the reference genome.
LL,RR	Implies inversion in sequenced DNA with respect to reference.
RL	Implies duplication or translocation with respect to reference.

Orientation of paired reads can reveal evidence of structural events, including:

- Inversions
- Duplications
- Translocations

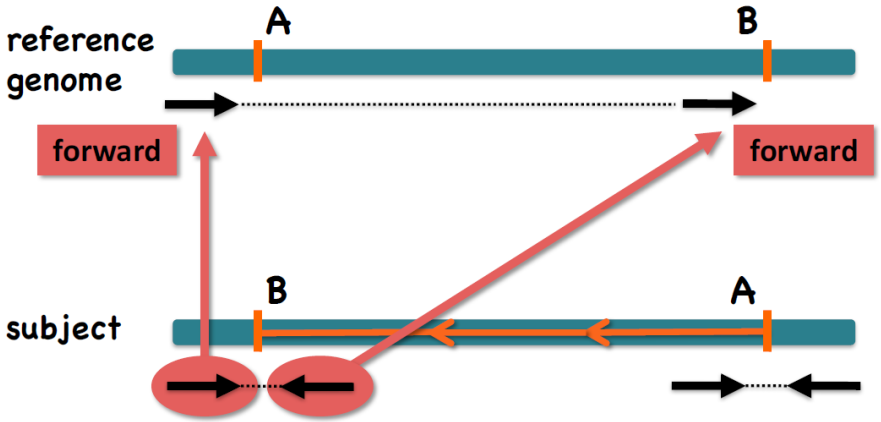
http://software.broadinstitute.org/software/igv/interpreting_insert_size

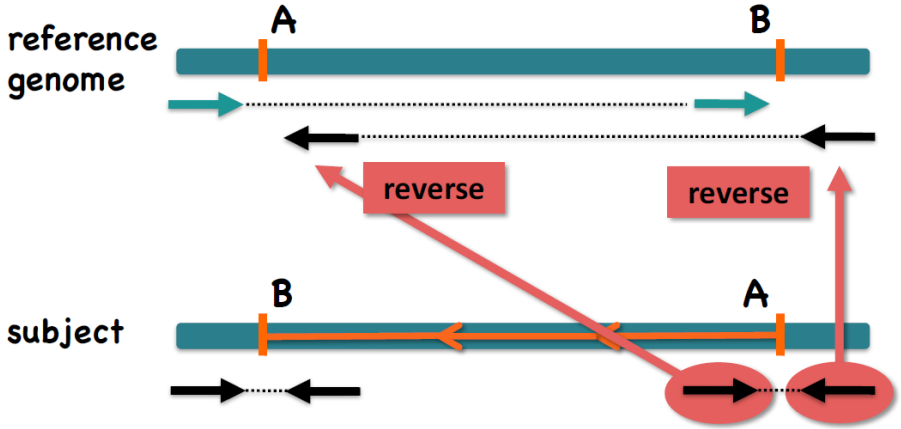
Read depth signals

Mate-pair or paired-end mapping abnormalities

Split-Read alignments

Inversion





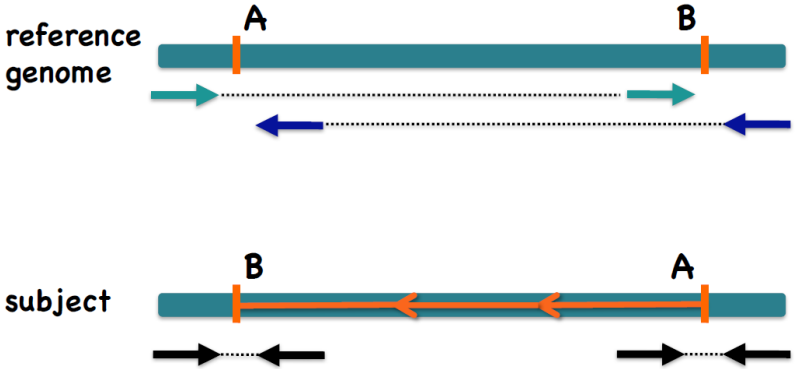
http://software.broadinstitute.org/software/igv/interpreting_insert_size

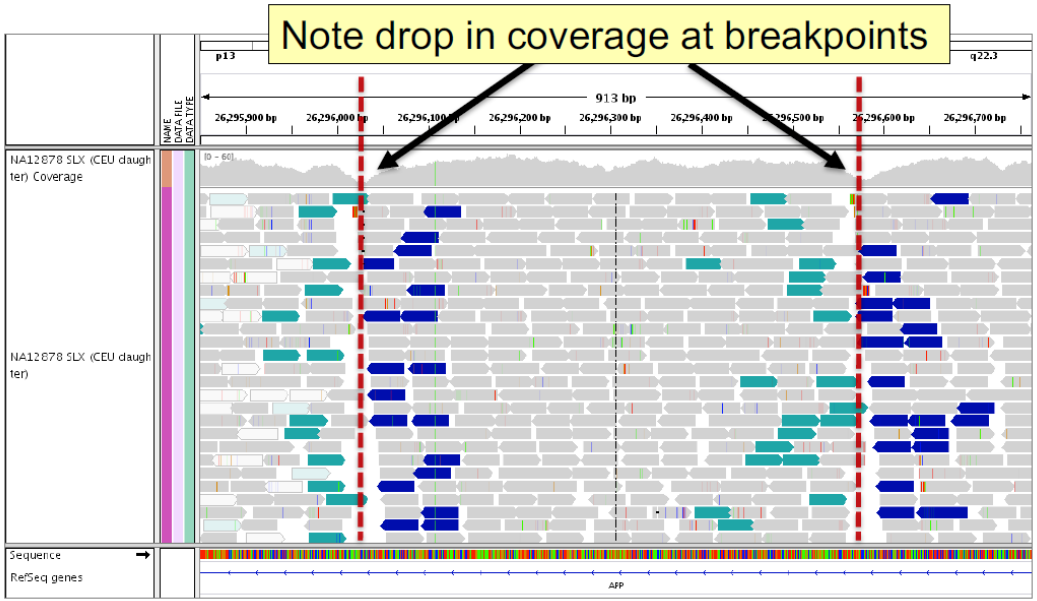
Read depth signals

Mate-pair or paired-end mapping abnormalities

Split-Read alignments

Inversion





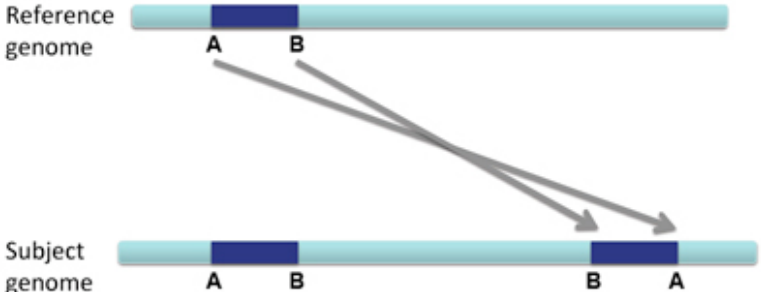
http://software.broadinstitute.org/software/igv/interpreting_insert_size

Read depth signals

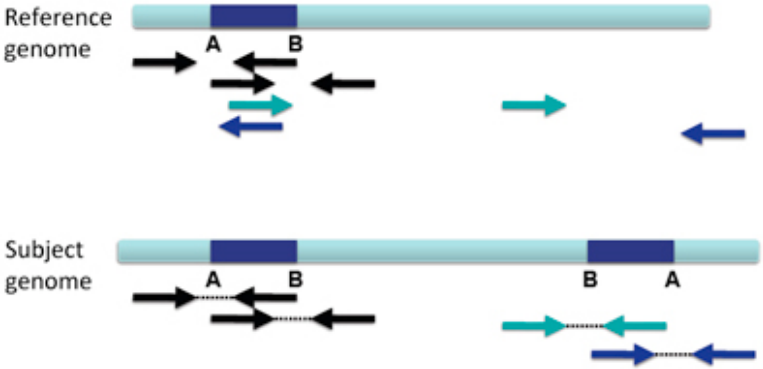
Mate-pair or paired-end mapping abnormalities

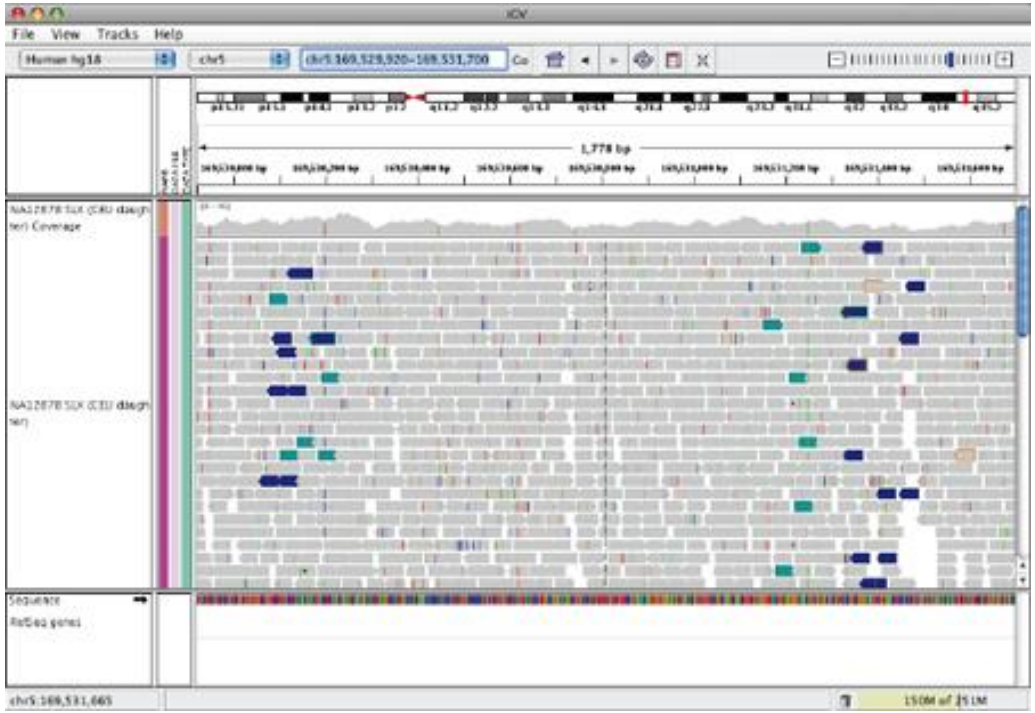
Split-Read alignments

Inverted Duplication



There will be overlapping left and right reads, and there will likely be alte





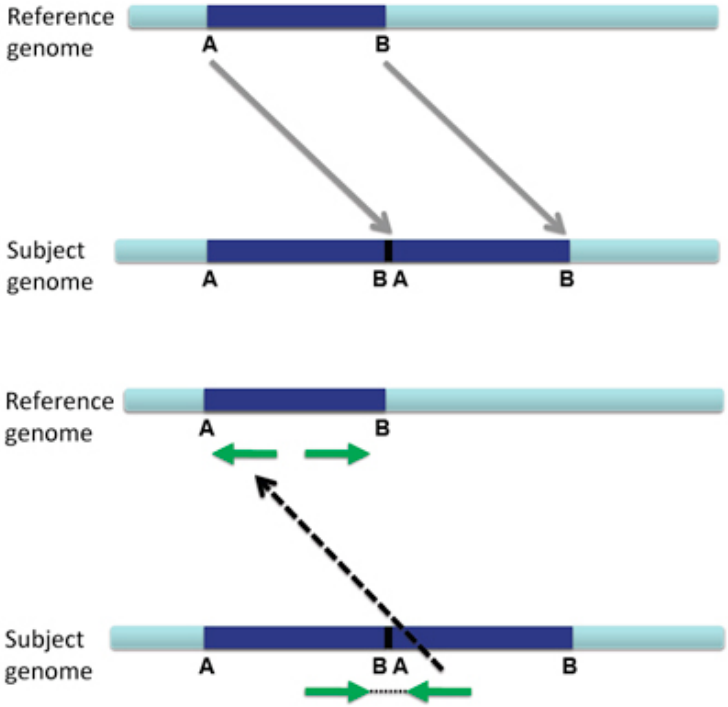
http://software.broadinstitute.org/software/igv/interpreting_insert_size

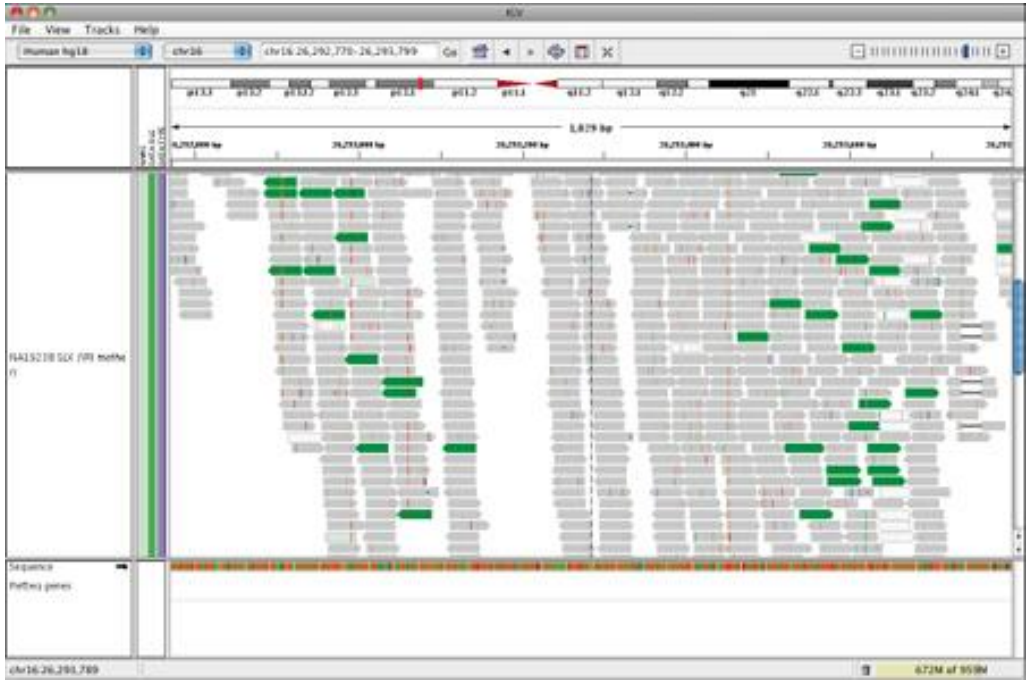
Read depth signals

Mate-pair or paired-end mapping abnormalities

Split-Read alignments

Tandem Duplication



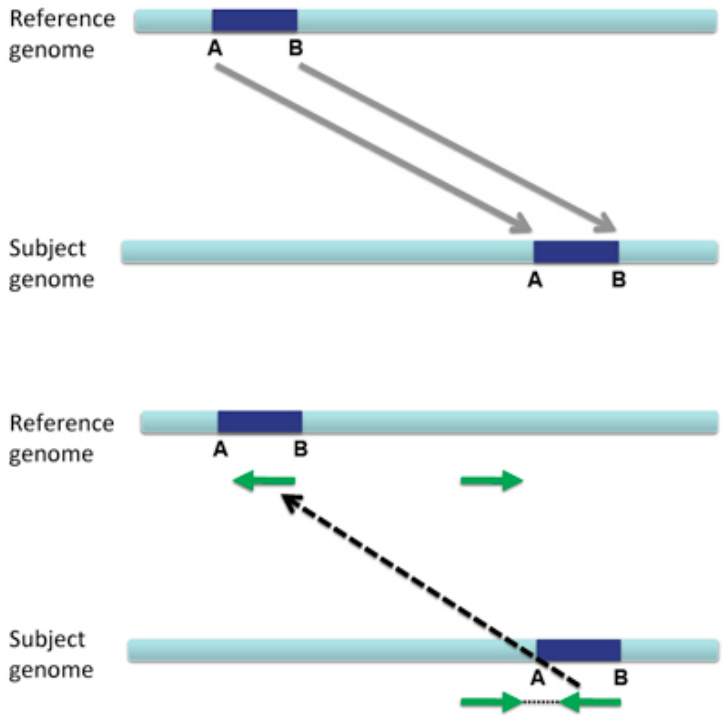


http://software.broadinstitute.org/software/igv/interpreting_insert_size

Read depth signals

Mate-pair or paired-end mapping abnormalities

Split-Read alignments



Translocation on the Same Chromosome

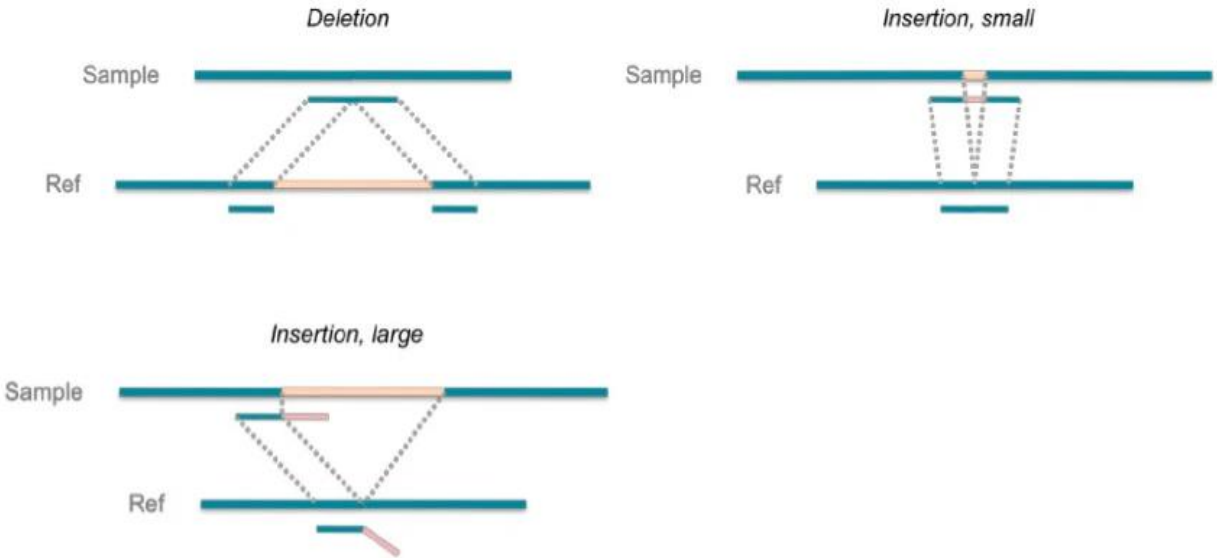
http://software.broadinstitute.org/software/igv/interpreting_insert_size

Mate-pair or paired-end mapping abnormalities

Read depth signals

Split-Read alignments

Reads spanning the exact breakpoint of a structural variation

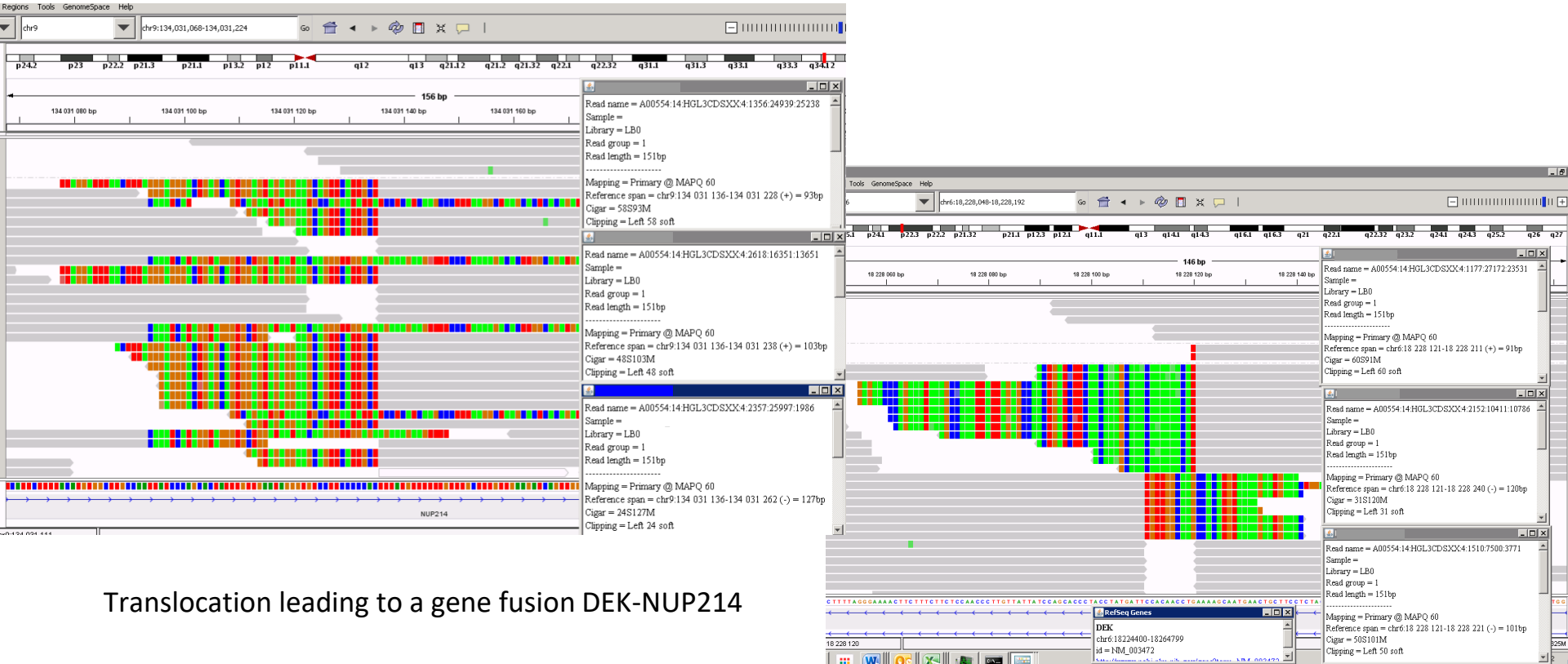


Zhang *et al.* Identification of genomic indels and structural variations using split reads. *BMC Genomics* 12, 375 (2011)

Mate-pair or paired-end mapping abnormalities

Read depth signals

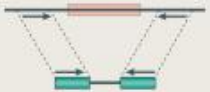
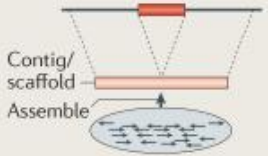
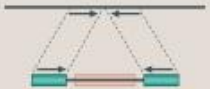
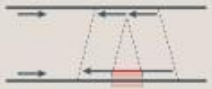
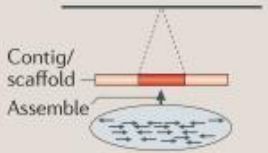
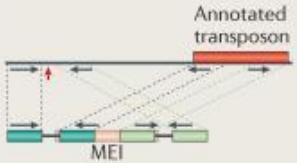
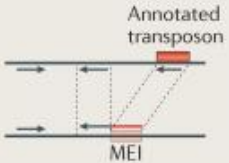
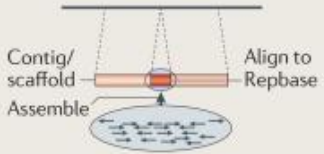
Split-Read alignments



Translocation leading to a gene fusion DEK-NUP214

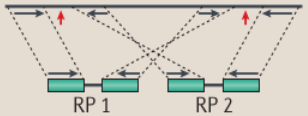
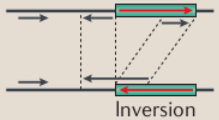
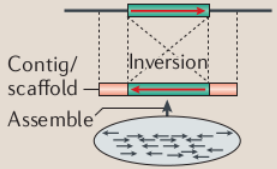
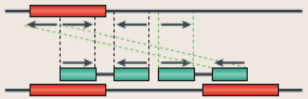
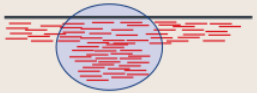
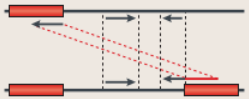
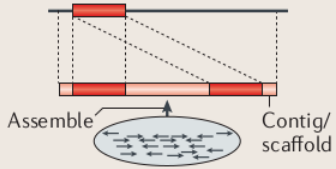
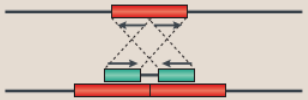
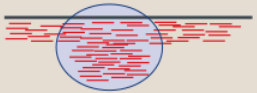
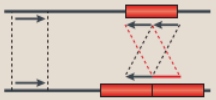
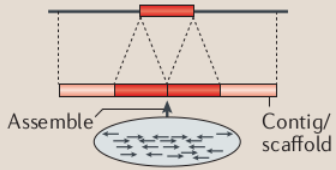
Structural Variation Calling

SNV INDEL		misalignment	< 50bp	

SV classes	Read pair	Read depth	Split read	Assembly
Deletion				
Novel sequence insertion		Not applicable		
Mobile-element insertion		Not applicable		

Genome structural variation discovery and genotyping. Nature Review Genetics 2014

Structural Variation Calling

SNV INDEL		misalignment	< 50bp	
SV classes	**Read pair**	**Read depth**	**Split read**	**Assembly**
Inversion		Not applicable		
Interspersed duplication				
Tandem duplication				

Genome structural variation discovery and genotyping. Nature Review Genetics 2014

How to represent Structural Variations in VCF ?

Encoding Structural Variants in VCF format

```
##fileformat=VCFv4.3
##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=Integer,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        1/1:14
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
```

https://samtools.github.io/hts-specs/VCFv4.3.pdf

Encoding Structural Variants in VCF format

```
##fileformat=VCFv4.3
##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=Integer,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        1/1:14
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
```

precise deletion with known breakpoint

https://samtools.github.io/hts-specs/VCFv4.3.pdf

Encoding Structural Variants in VCF format

```
##fileformat=VCFv4.3
##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=Integer,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        1/1:14
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
```

imprecise deletion of approximately 205 bp

https://samtools.github.io/hts-specs/VCFv4.3.pdf

Encoding Structural Variants in VCF format

```
##fileformat=VCFv4.3
##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=Integer,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        1/1:14
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
```

imprecise deletion of an ALU element

https://samtools.github.io/hts-specs/VCFv4.3.pdf

Encoding Structural Variants in VCF format

```
##fileformat=VCFv4.3
##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=Integer,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        1/1:14
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
```

imprecise insertion of an L1 element

https://samtools.github.io/hts-specs/VCFv4.3.pdf

Encoding Structural Variants in VCF format

```
##fileformat=VCFv4.3
##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=Integer,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        1/1:14
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
```

imprecise duplication of approximately 21Kb

https://samtools.github.io/hts-specs/VCFv4.3.pdf

Encoding Structural Variants in VCF format

```
##fileformat=VCFv4.3
##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=Integer,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        1/1:14
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
```

imprecise tandem duplication of 76bp

https://samtools.github.io/hts-specs/VCFv4.3.pdf

Encoding complex rearrangements with breakends in VCF

Rearrangement breakpoint defined as 2 breakends → novel adjacency

Breakend is encoded by SVTYPE=BND in the INFO field

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
2	321681	bnd_W	G	G]17:198982]	6	PASS	SVTYPE=BND
2	321682	bnd_V	T	]13:123456]T	6	PASS	SVTYPE=BND
13	123456	bnd_U	C	C[2:321682[	6	PASS	SVTYPE=BND
13	123457	bnd_X	A	[17:198983[A	6	PASS	SVTYPE=BND
17	198982	bnd_Y	A	A]2:321681]	6	PASS	SVTYPE=BND
17	198983	bnd_Z	C	[13:123457[C	6	PASS	SVTYPE=BND

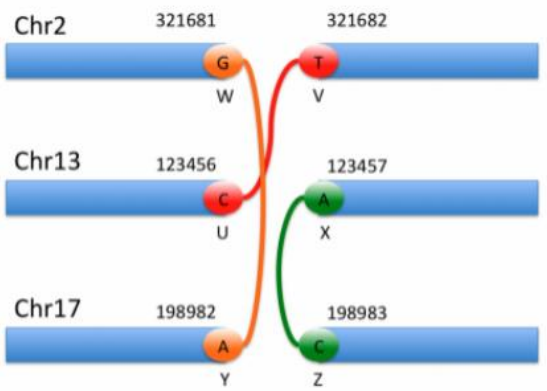


Figure 1: All possible orientations of breakends

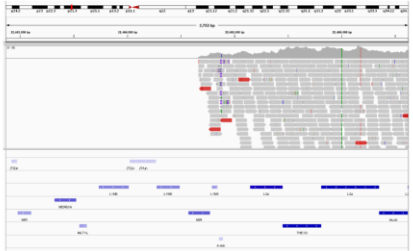
REF	ALT	Meaning
s	t[p[	piece extending to the right of p is joined after t
s	t]p]	reverse comp piece extending left of p is joined after
s	]p]t	piece extending to the left of p is joined before t
s	[p[t	reverse comp piece extending right of p is joined before t

https://samtools.github.io/hts-specs/VCFv4.3.pdf

How to visualize Structural Variations ?

Visualization of structural variants

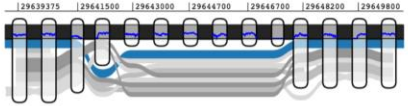
A. Linear Genome Browser



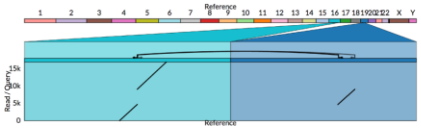
F. Linear Coordinate Plot

D. SV Table

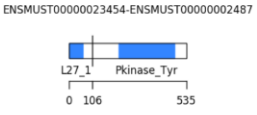
I. Graph View



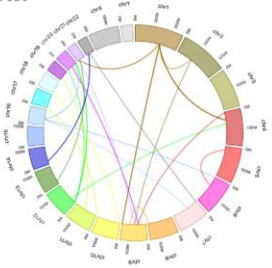
B. Dot Plot



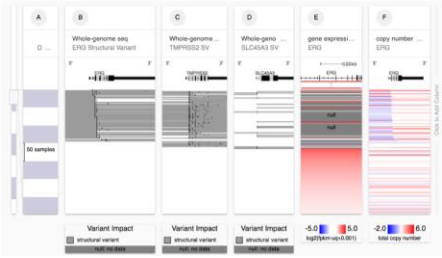
G. Two-way View



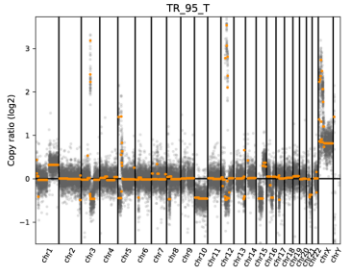
E. Circos Plot



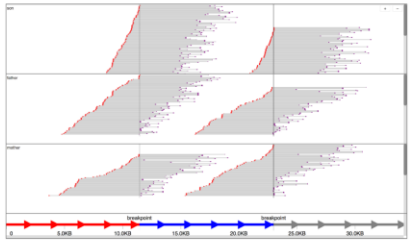
J. Population View



C. Scatter Plot

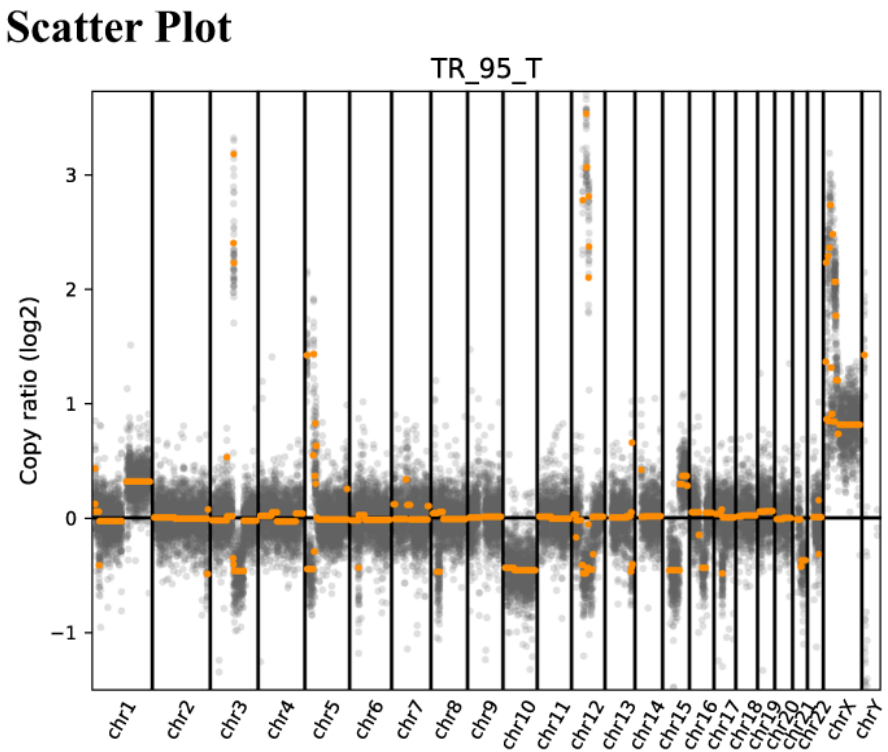


H. Multi-way View



Yokoyama, T.T., Kasahara, M. Visualization tools for human structural variations identified by whole-genome sequencing. *J Hum Genet* **65**, 49–60 (2020).

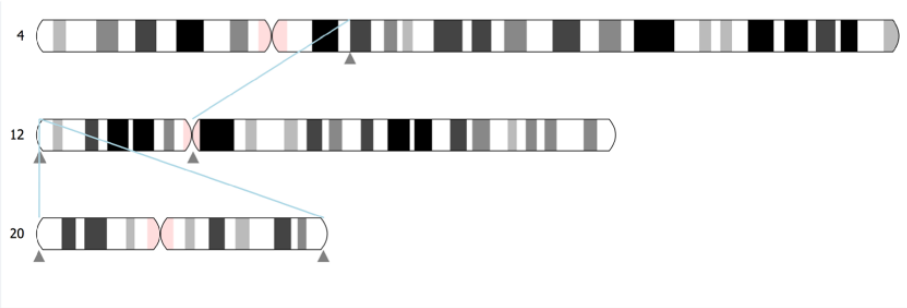
Visualization of structural variants



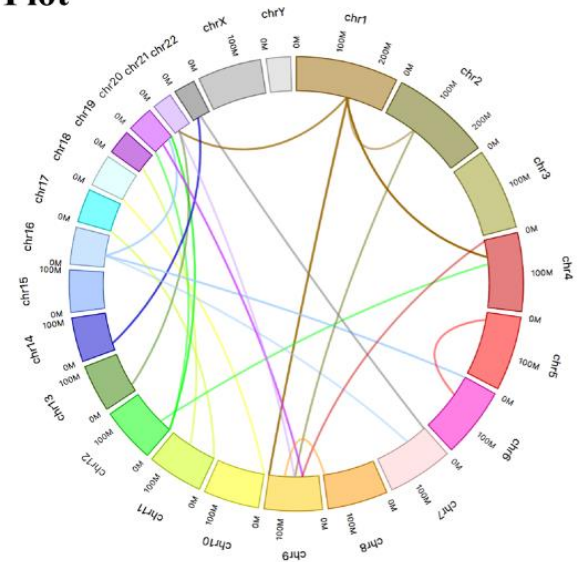
Yokoyama, T.T., Kasahara, M. Visualization tools for human structural variations identified by whole-genome sequencing. *J Hum Genet* **65**, 49–60 (2020).

Visualization of structural variants

Linear Coordinate Plot

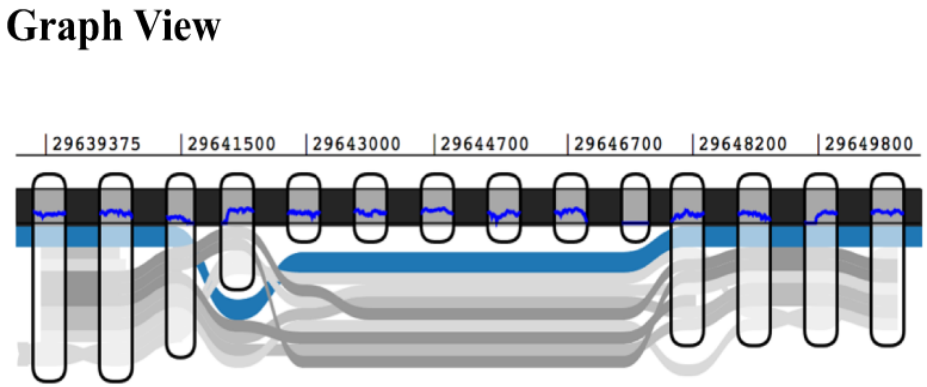


Circos Plot



Yokoyama, T.T., Kasahara, M. Visualization tools for human structural variations identified by whole-genome sequencing. *J Hum Genet* **65**, 49–60 (2020).

Visualization of structural variants



Yokoyama, T.T., Kasahara, M. Visualization tools for human structural variations identified by whole-genome sequencing. *J Hum Genet* **65**, 49–60 (2020).