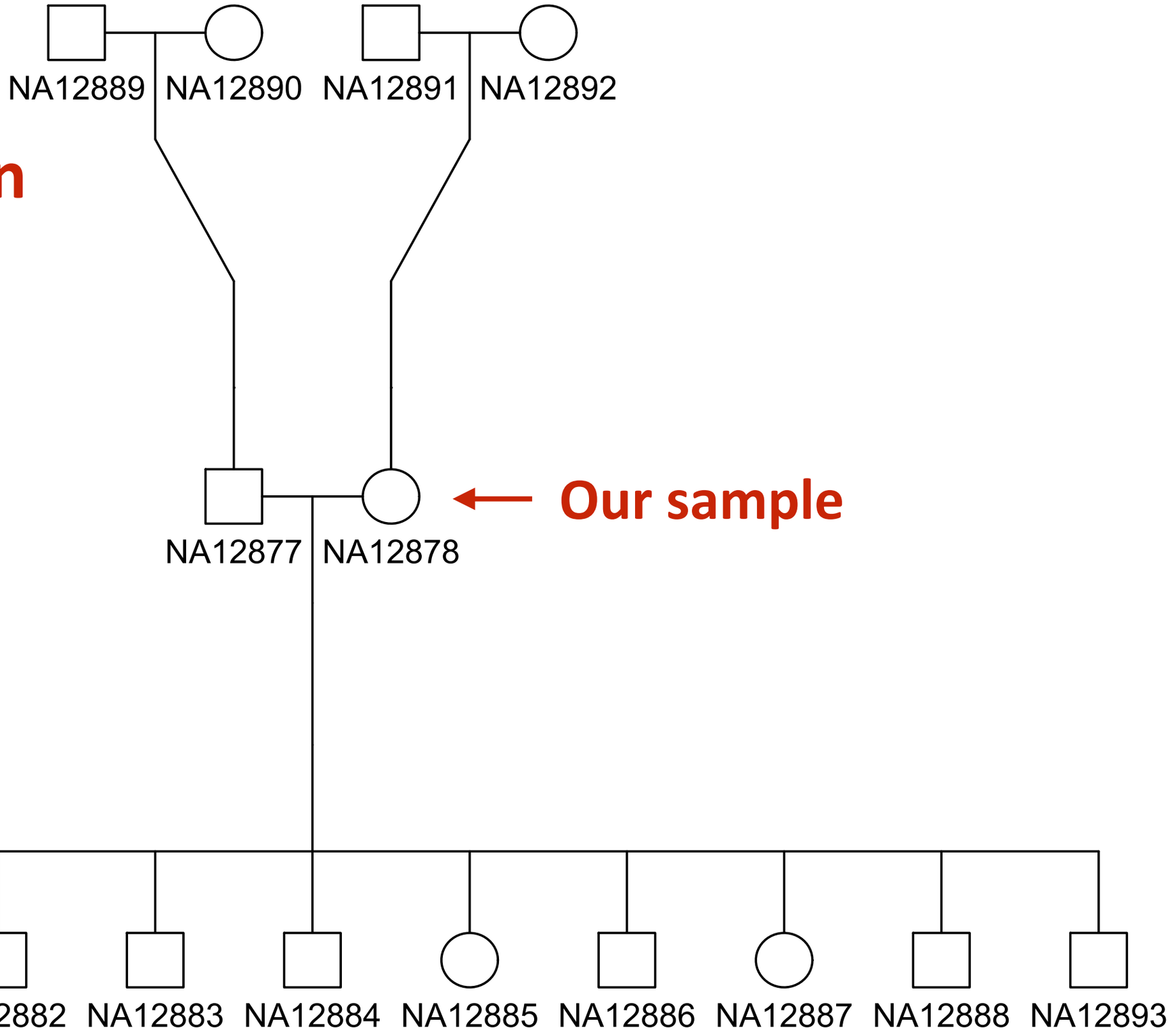
Practical session: mapping structural variation from whole genome sequence (WGS) data

A three-generation human pedigree (CEPH 1463)



Working with BAM files

Fields

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

Flags

Flag	Chr	Description
0x0001	p	the read is paired in sequencing
0x0002	P	the read is mapped in a proper pair
0x0004	u	the query sequence itself is unmapped
0x0008	U	the mate is unmapped
0x0010	r	strand of the query (1 for reverse)
0x0020	R	strand of the mate
0x0040	1	the read is the first read in a pair
0x0080	2	the read is the second read in a pair
0x0100	s	the alignment is not primary
0x0200	f	the read fails platform/vendor quality checks
0x0400	d	the read is either a PCR or an optical duplicate

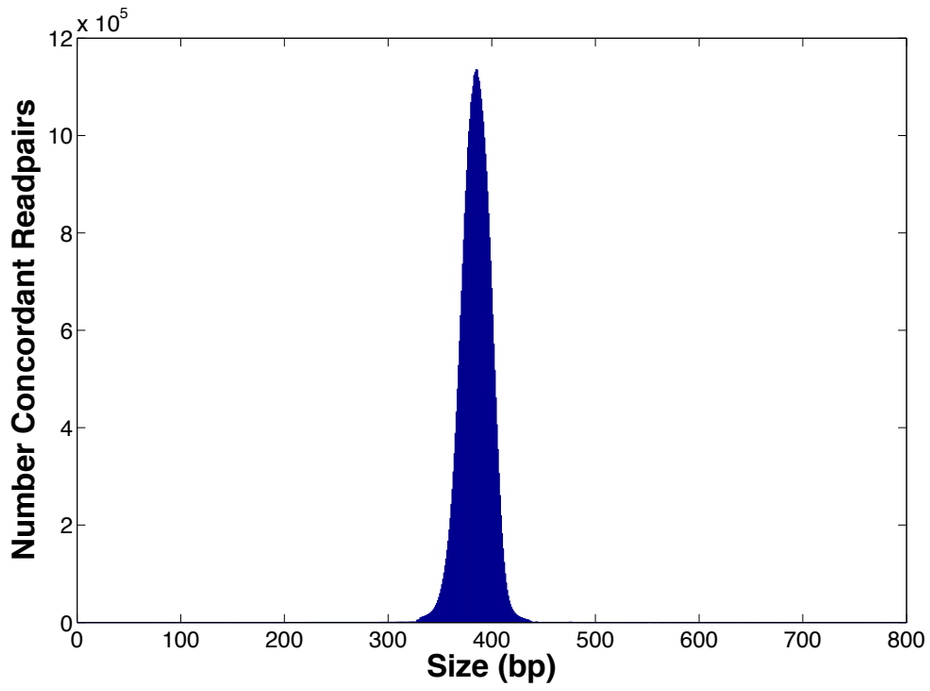
To look at a bam file: samtools view file.bam

To select for entries that have a given flag:
samtools view -f 0x0002 file.bam

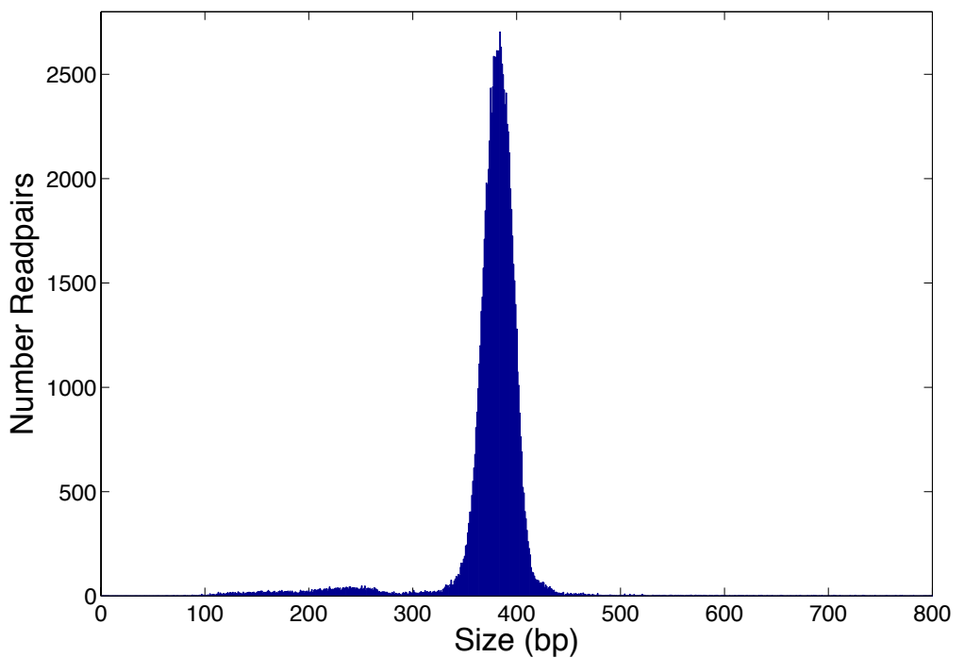
To select for entries that lack a given flag:
samtools view -F 0x0002 file.bam

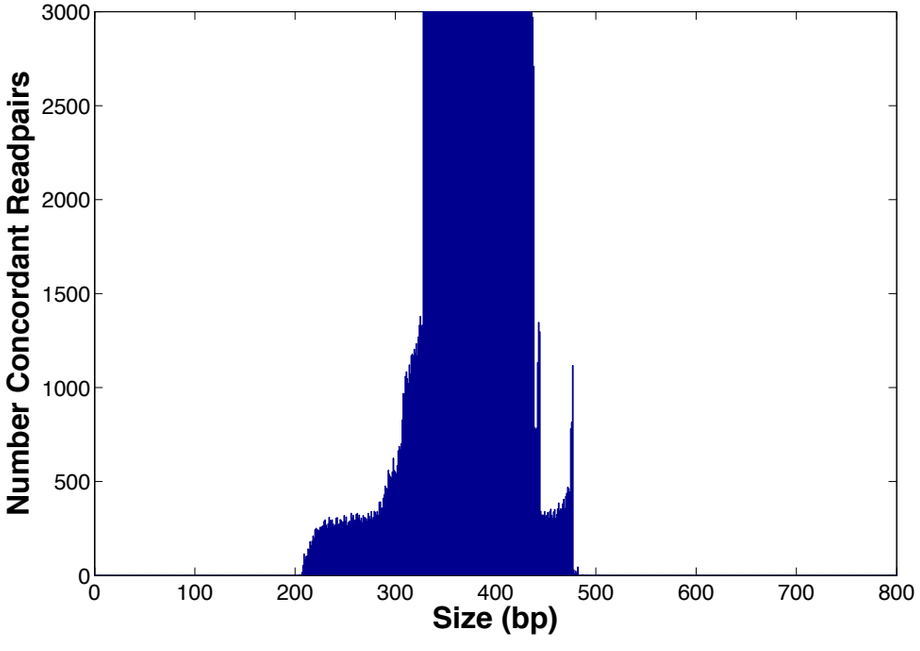
Insert Size

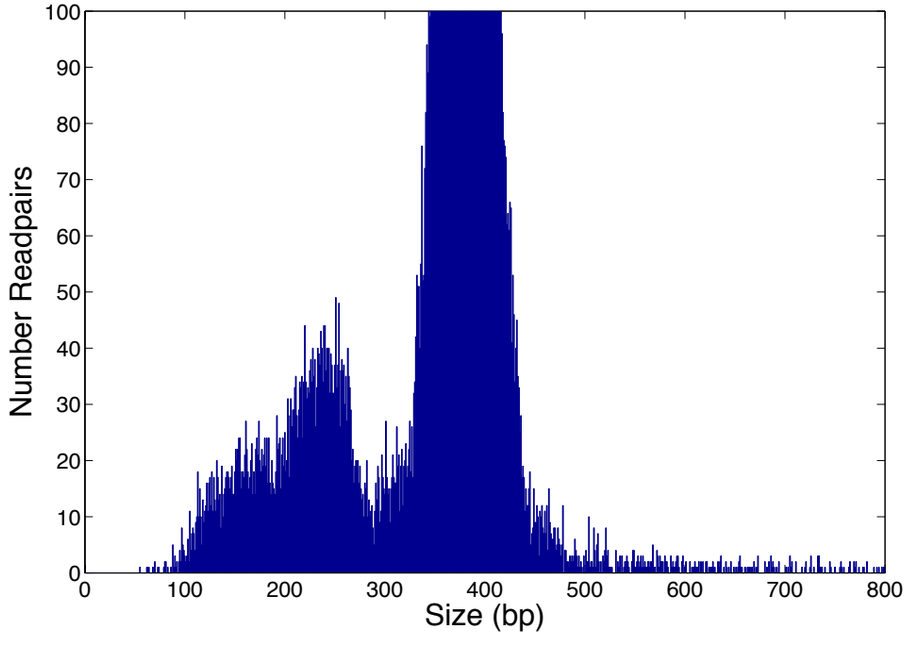
39 million concordant readpairs



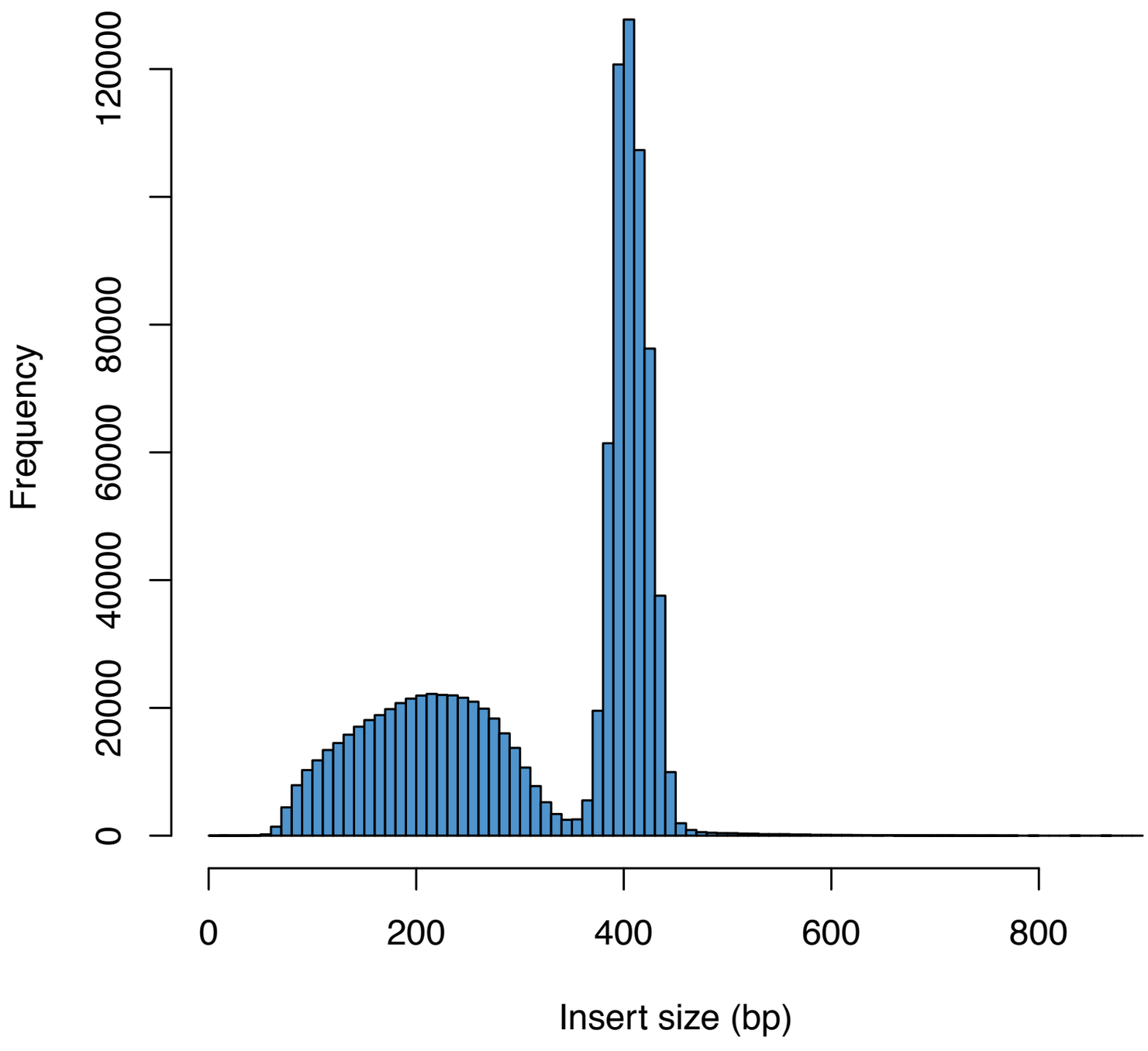
100,000 readpairs +/- orientation, < 1000 bp



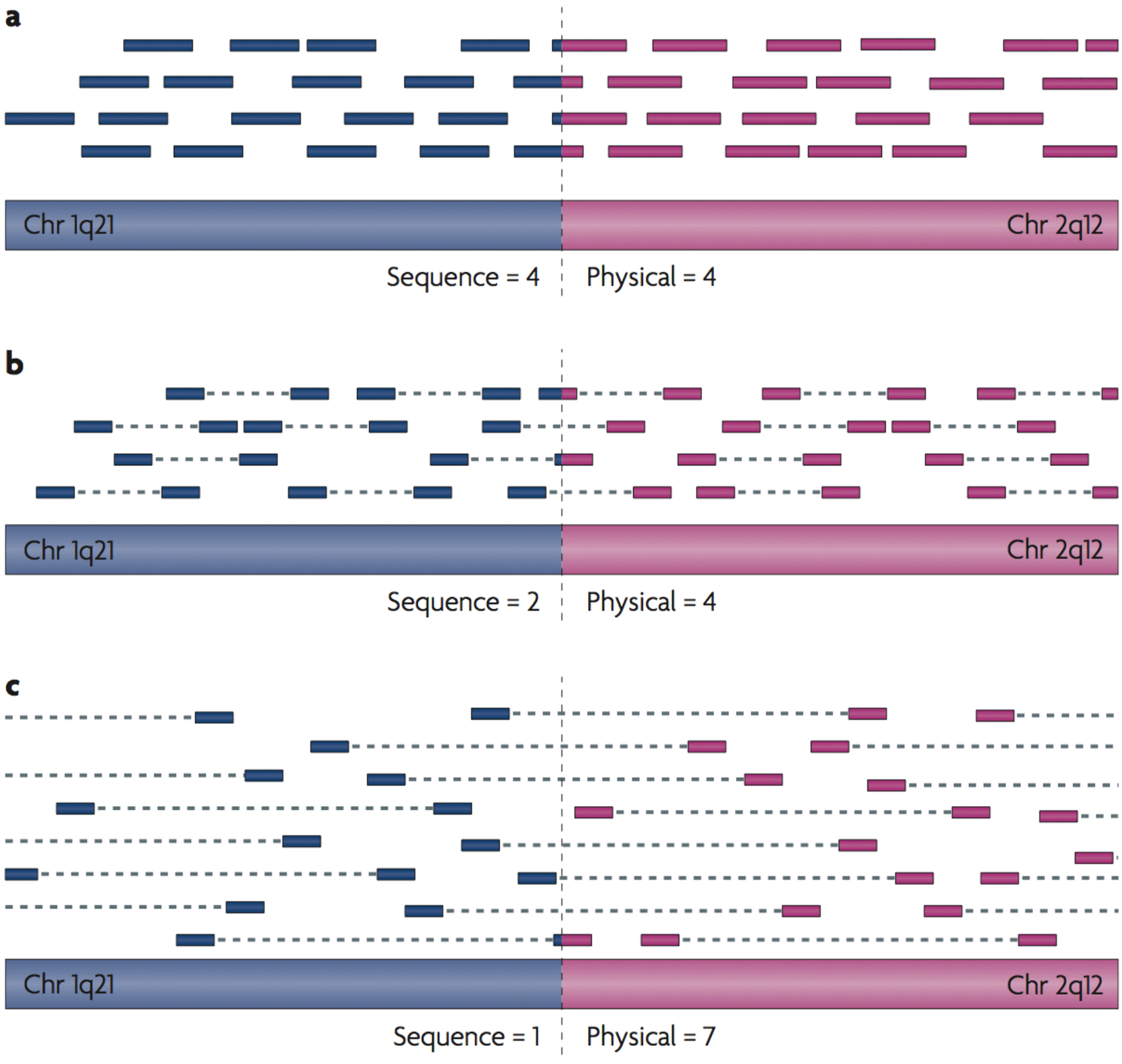




A recent WGS dataset from the GTEX project



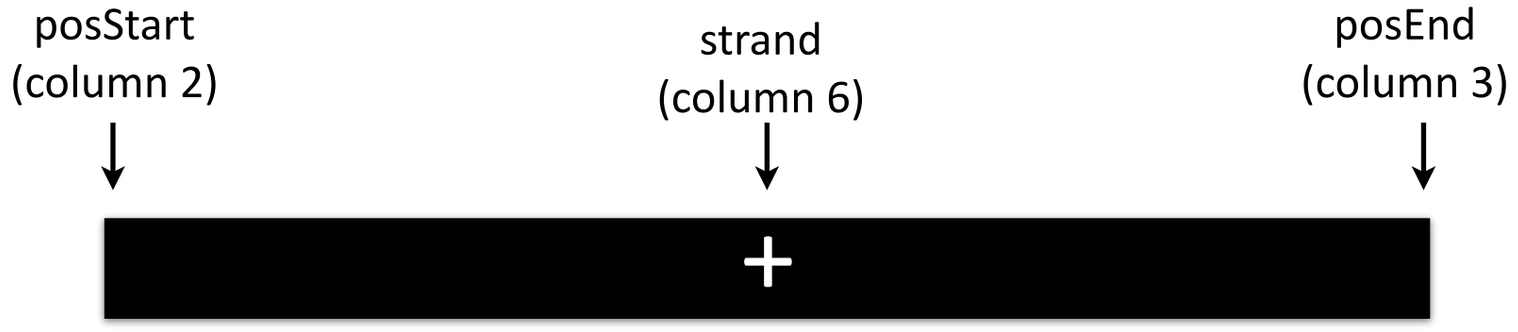
Sequence vs. physical coverage



Meyerson et al., Nature Reviews, 2010

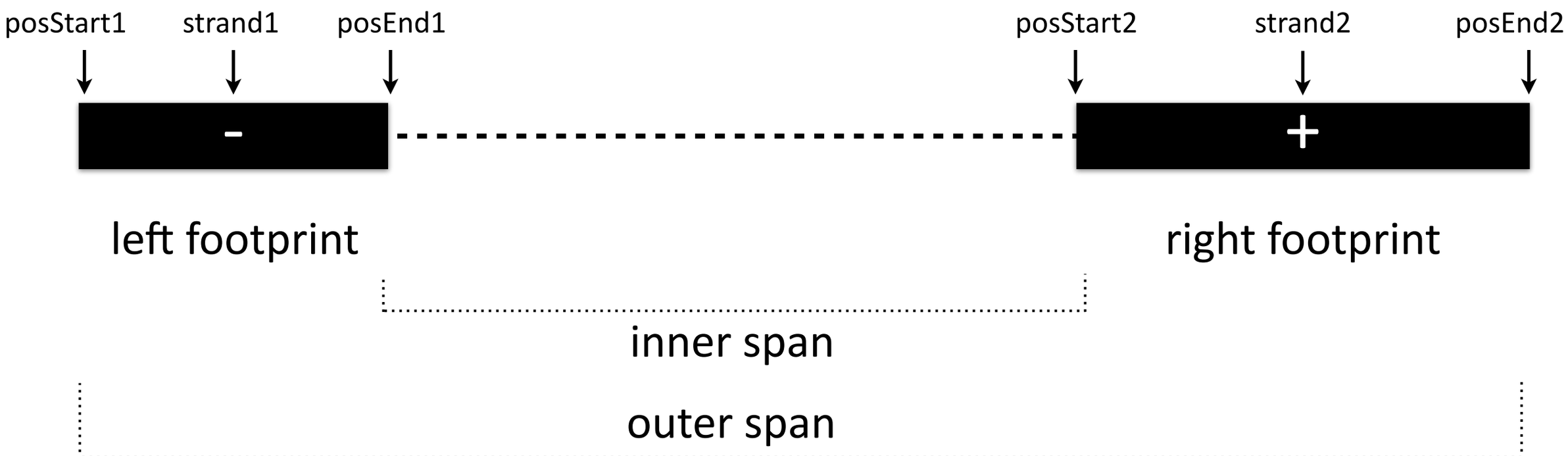
Bed format (.bed):

1=chrom
2=posStart
3=posEnd
4=name
5=score
6=strand (+/-)



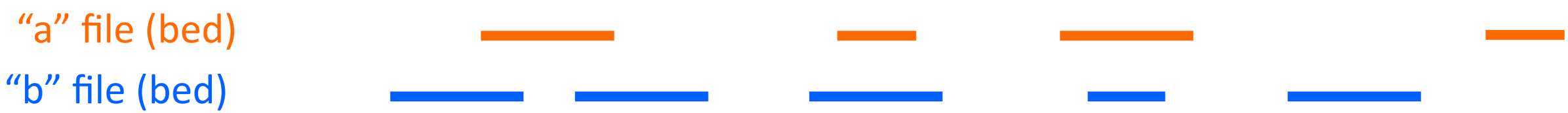
BEDPE format (.bedpe)

1=chrom1 ; **2**=posStart1 ; **3**=posEnd1 ; **4**=chrom2 ; **5**=posStart2 ; **6**=posEnd2 ;
7=featureName ; **8**=score ; **9**=strand1 ; **10**=strand2

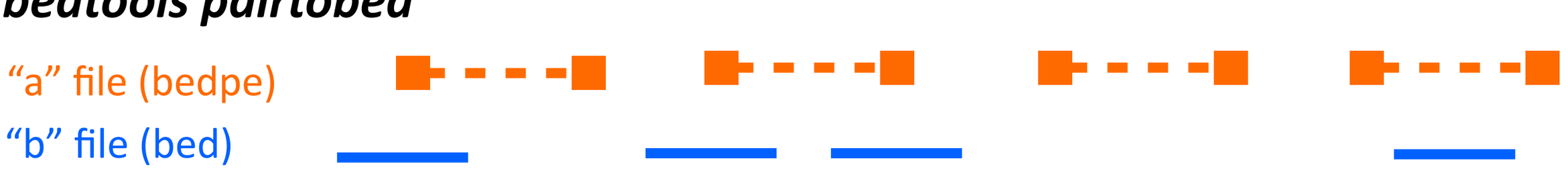


Bedtools: intersect, pairtobed & pairtopair

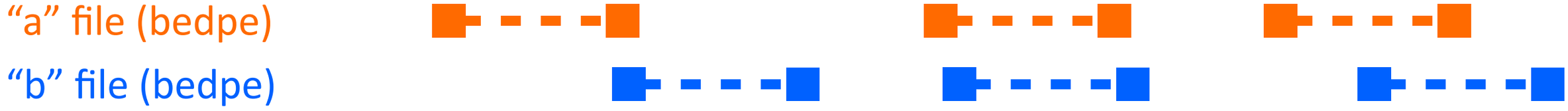
bedtools intersect



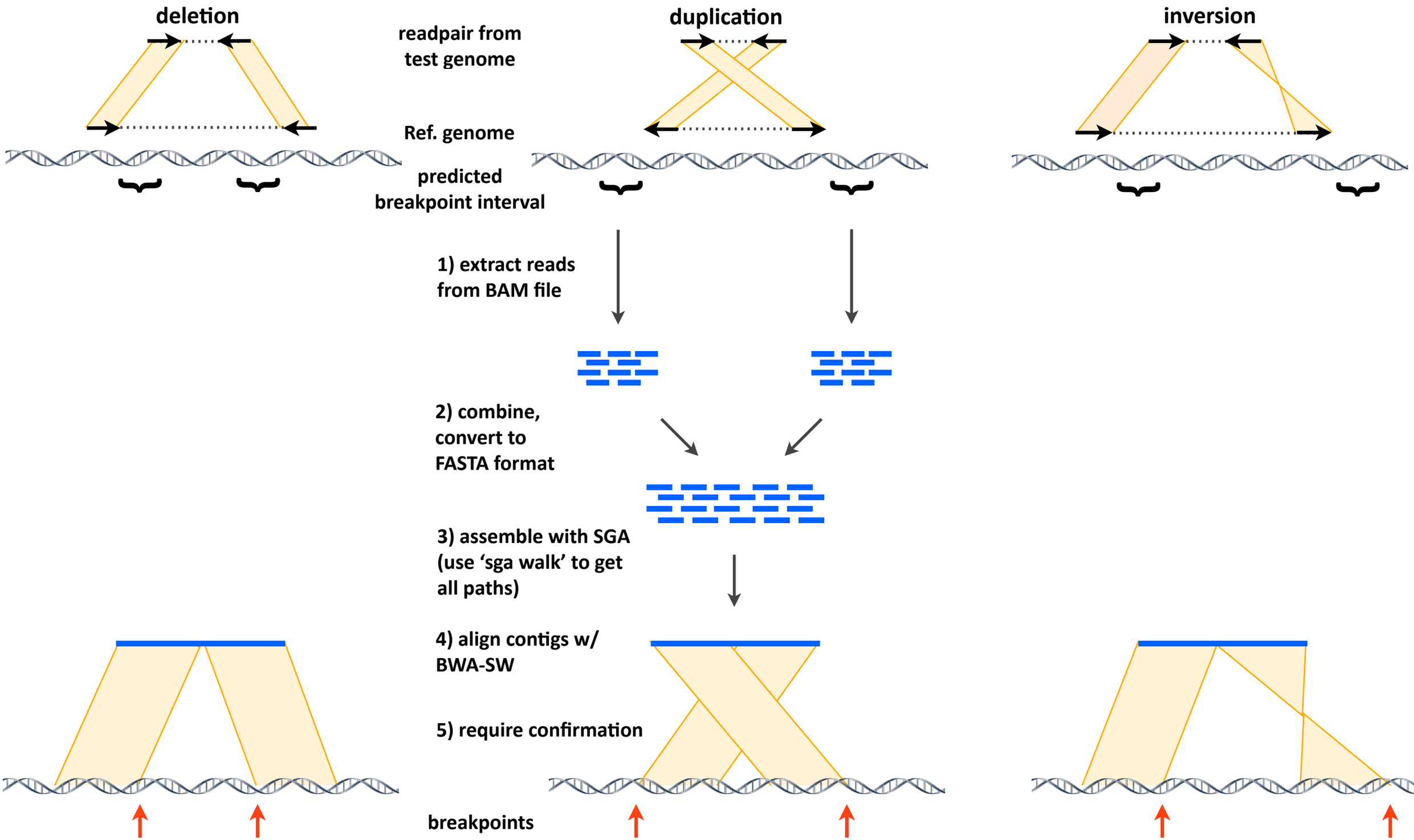
bedtools pairtobed



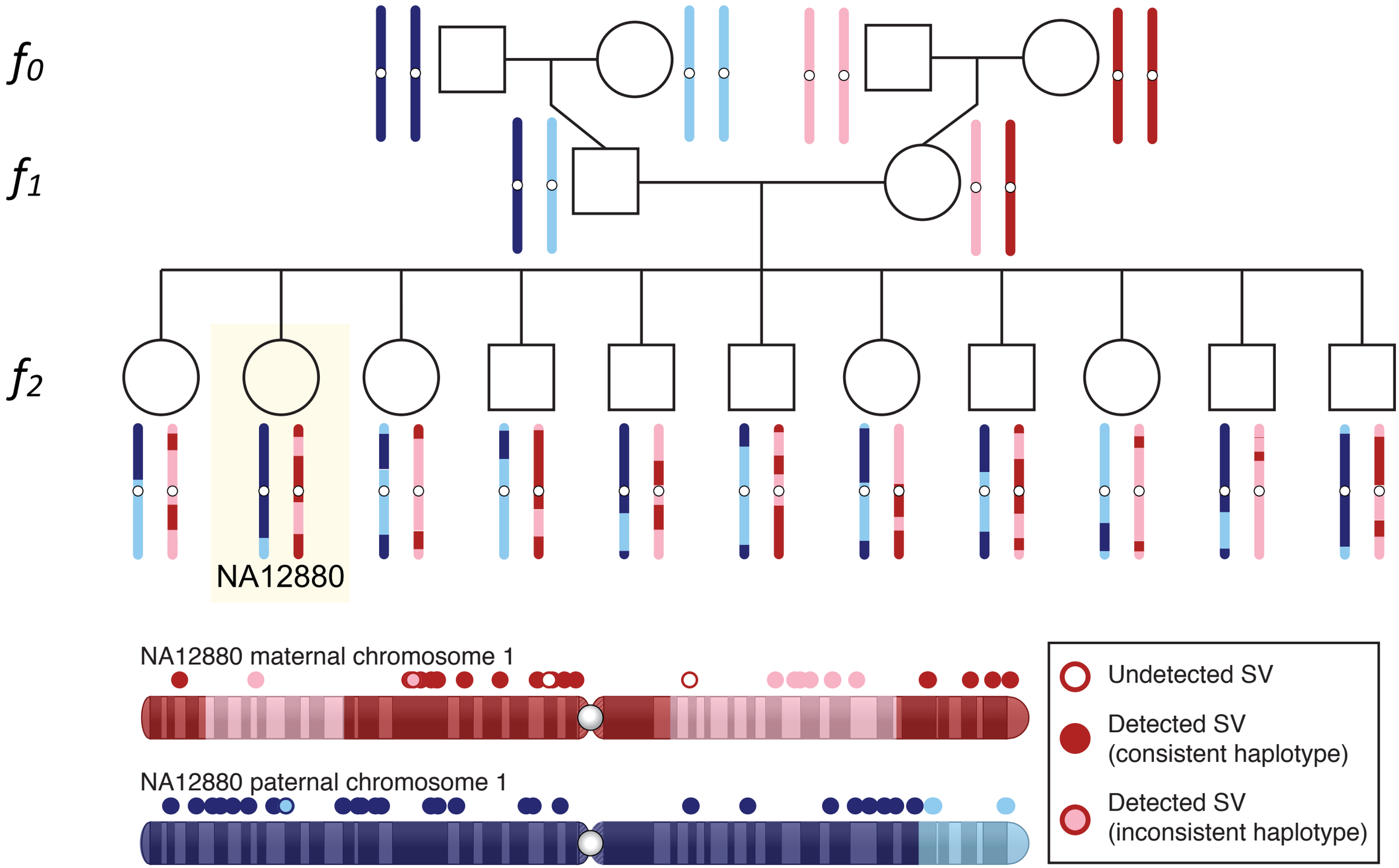
bedtools pairtopair



Validation by assembly



SV “validation” by familial segregation



<u>Measures</u>: 1) mendelian violation; 2) detection rate; 3) haplotype concordance

Tuning variant detection performance using ROC curves