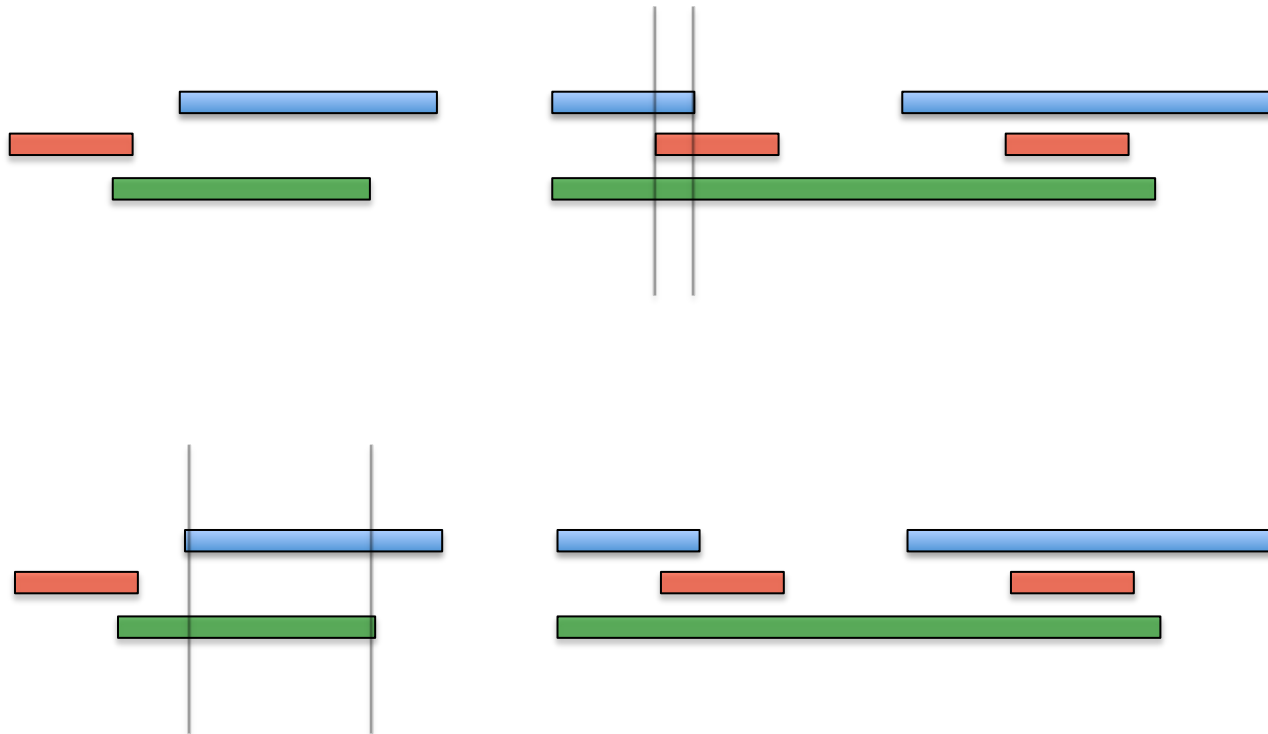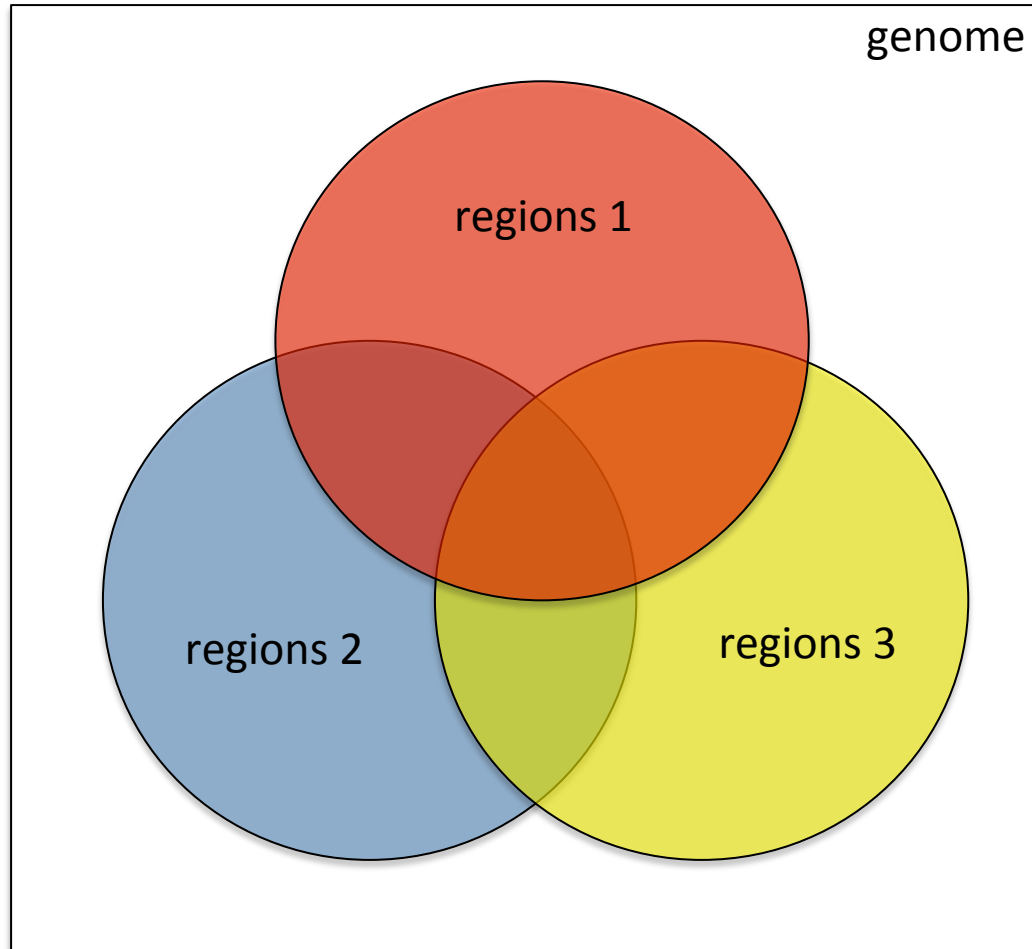# bedtools

intersect, jaccard & merge,
complement & subtract
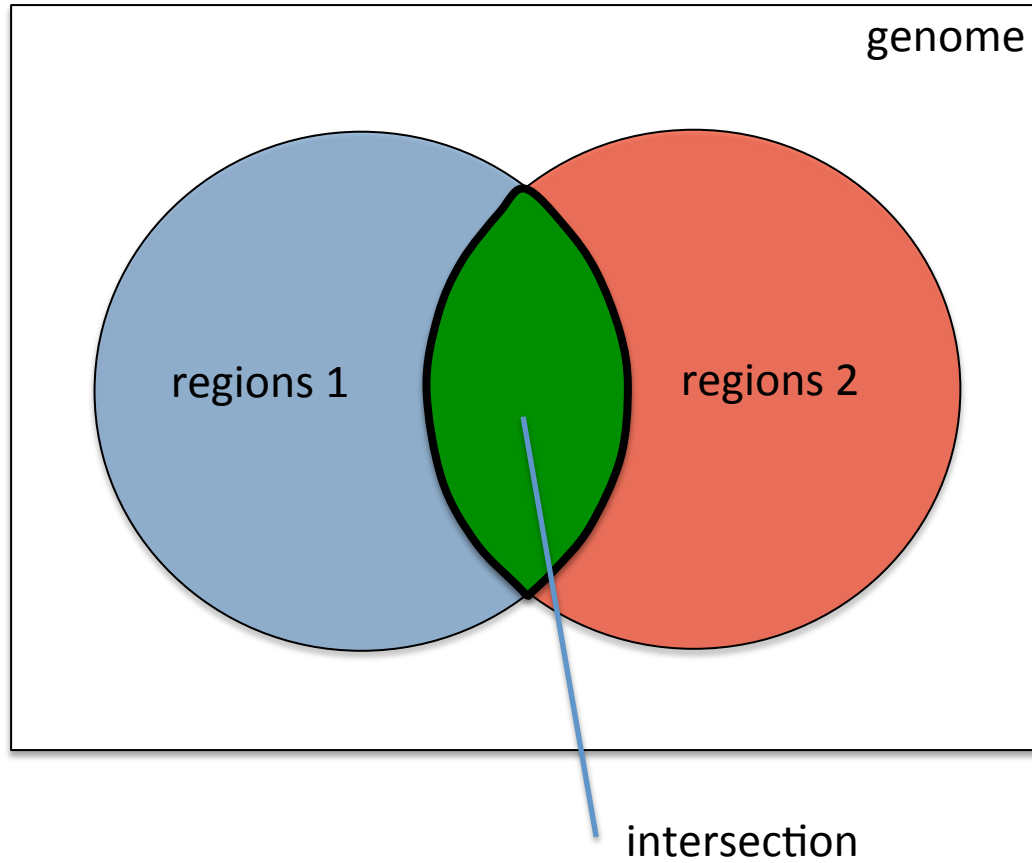
Sets and regions
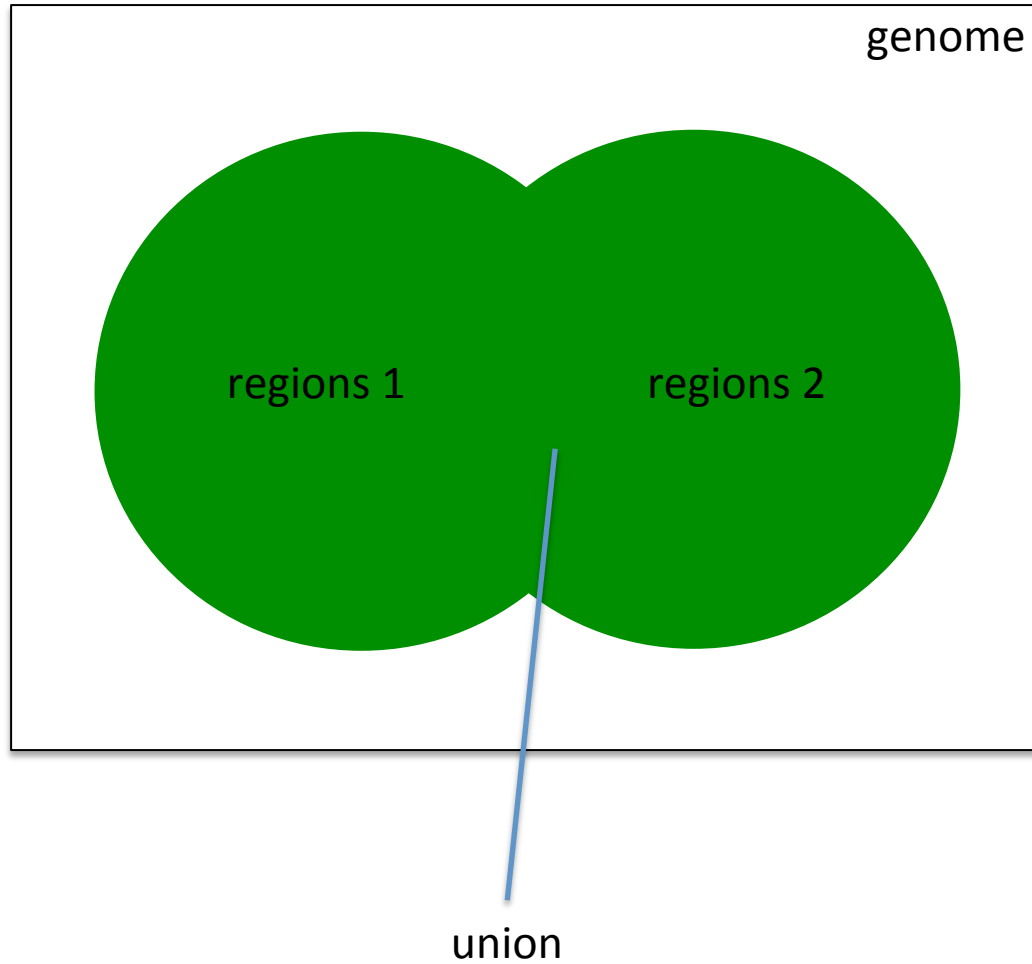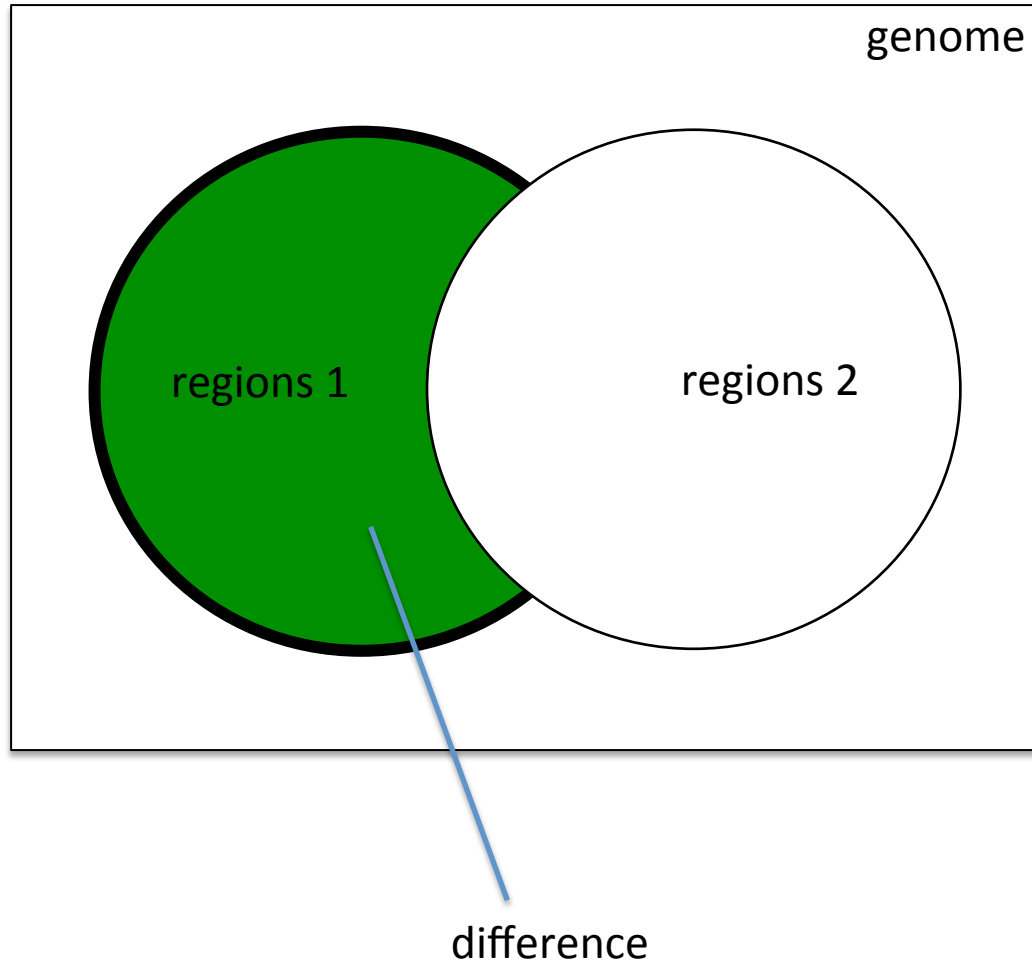
Sets and regions

bedtools intersect

bedtools merge

bedtools subtract

bedtools complement
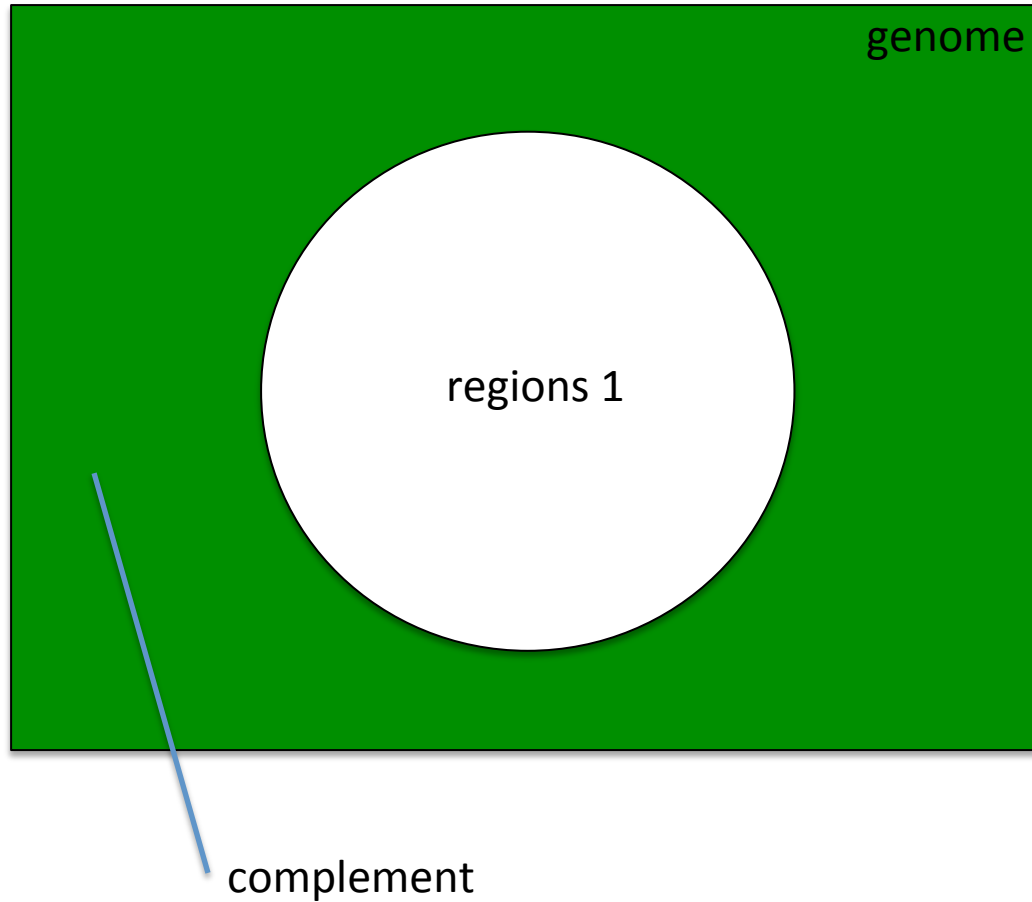
bedtools intersect
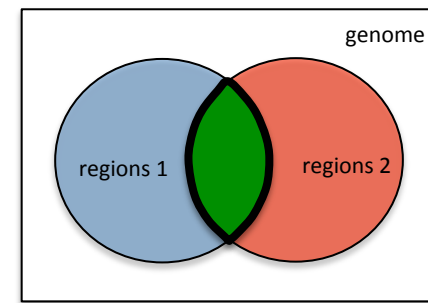
- Finds all possible intersections between regions in A and in B
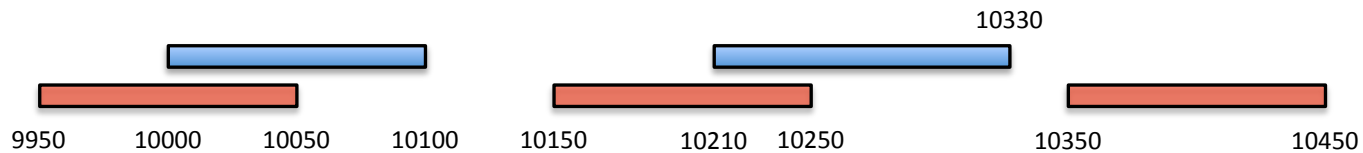
```
bedtools intersect -a <fileA> -b <fileB>
bedtools intersect -wa -a <fileA> -b <fileB>
bedtools intersect -u -a <fileA> -b <fileB>
bedtools intersect -wa -wb -a <fileA> -b <fileB>
bedtools intersect -u -a <fileA> -b <fileB>
bedtools intersect -wao -a <fileA> -b <fileB> | uniq
bedtools intersect -f 0.3 ………
bedtools intersect -f 0.3 -r ………
bedtools intersect -v ………
```

Exercise

CTCF binding motif
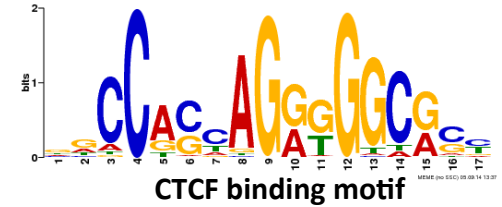
Given motif instances and ChIP-seq peaks:

CTCF_motif_instances_chr15.bed

K562_CTCF_CTCF_ENCFF002CEL_chr15.bed

- What percentage of ChIP-seq peaks coincide with a motif instance?

- What is the maximum number of motif instances found in a single ChIP-seq peak?

bedtools merge


```
bedtools merge -i <file>
bedtools merge -i <file> -d 100
bedtools merge -i <file> -s ........
```
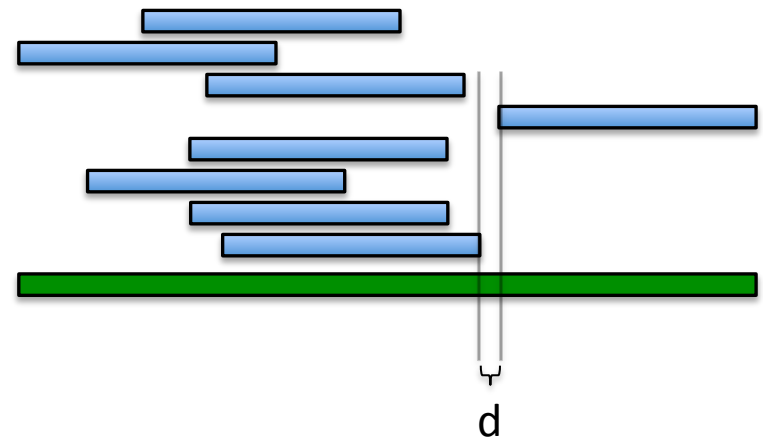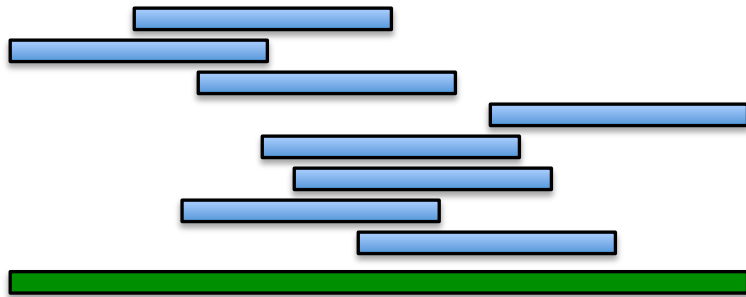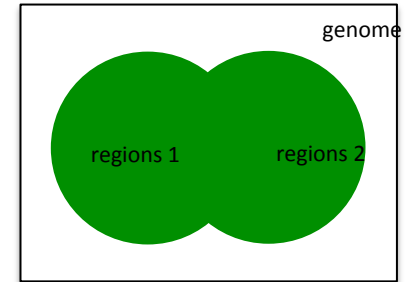

Example: merging multiple files into non-overlapping regions:

```
cat fileA fileB fileB | bedtools sort | bedtools merge
```

Exercise

Given motif instances and two sets of ChIP-seq peaks:

CTCF_motif_instances_chr15.bed
K562_CTCF_CTCF_ENCFF002CEL_chr15.bed
K562_CTCF_CTCF_ENCFF002DBD_chr15.bed

- How many motif instances to not coincide with peaks in either set of ChIP-seq peaks?
- How many motif instances overlap with overlapping ChIP-seq peaks (> 75% reciprocal overlap)
- Using:

  K562_MYC_MYC_ENCFF002CWF_chr15.bed

  How many of these overlapping ChIP-seq peaks that contain a CTCF motif instance also overlap with a ChIP-seq peak for MYC?

bedtools subtract


```
Tool:    bedtools subtract (aka subtractBed)
Version: v2.18.2
Summary: Removes the portion(s) of an interval that is overlapped
         by another feature(s).


Usage:   bedtools subtract [OPTIONS] -a <bed/gff/vcf> -b <bed/gff/vcf>


Options:
        -f      Minimum overlap required as a fraction of A.
                - Default is 1E-9 (i.e., 1bp).
                - (FLOAT) (e.g. 0.50)


        -s      Require same strandedness.  That is, only subtract hits in B
                that overlap A on the _same_ strand.
                - By default, overlaps are subtracted without respect to strand.


        -S      Force strandedness.  That is, only subtract hits in B that
                overlap A on the _opposite_ strand.
                - By default, overlaps are subtracted without respect to strand.


        -A      Remove entire feature if any overlap.  That is, by default,
                only subtract the portion of A that overlaps B. Here, if
                any overlap is found (or -f amount), the entire feature is removed.


        -N      Same as -A except when used with -f, the amount is the sum
                of all features (not any single feature).
```
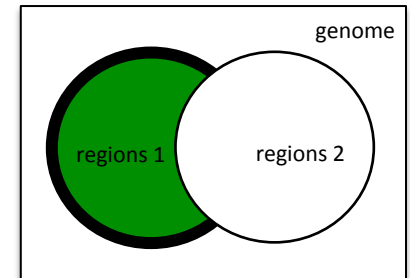
bedtools complement


```
Tool:    bedtools complement (aka complementBed)
Version: v2.18.2
Summary: Returns the base pair complement of a feature file.


Usage:   bedtools complement [OPTIONS] -i <bed/gff/vcf> -g <genome>


Notes:
       (1)  The genome file should tab delimited and structured as follows:
            <chromName><TAB><chromSize>


       For example, Human (hg19):
       chr1    249250621
       chr2    243199373
       ...
       chr18_gl000207_random   4262
```
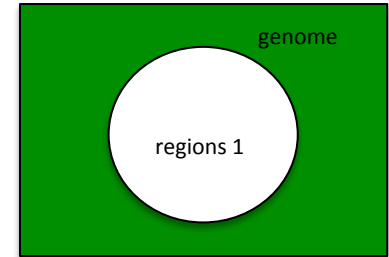
bedtools jaccard

```
Tool:    bedtools jaccard (aka jaccard)
Version: v2.18.2
Summary: Calculate Jaccard statistic b/w two feature files.
         Jaccard is the length of the intersection over the union.
         Values range from 0 (no intersection) to 1 (self intersection).


Usage:   bedtools jaccard [OPTIONS] -a <bed/gff/vcf> -b <bed/gff/vcf>

Options:
        -f      Minimum overlap required as a fraction of A.
                - Default is 1E-9 (i.e., 1bp).
                - FLOAT (e.g. 0.50)

        -r      Require that the fraction overlap be reciprocal for A and B.
                - In other words, if -f is 0.90 and -r is used, this requires
                  that B overlap 90% of A and A _also_ overlaps 90% of B.
```

Similarity of two sets can be calculated as the number of bases in the intersection of the number of bases in the union.