Ресеквенирование
Поиск полиморфизмов у человека
Продолжение

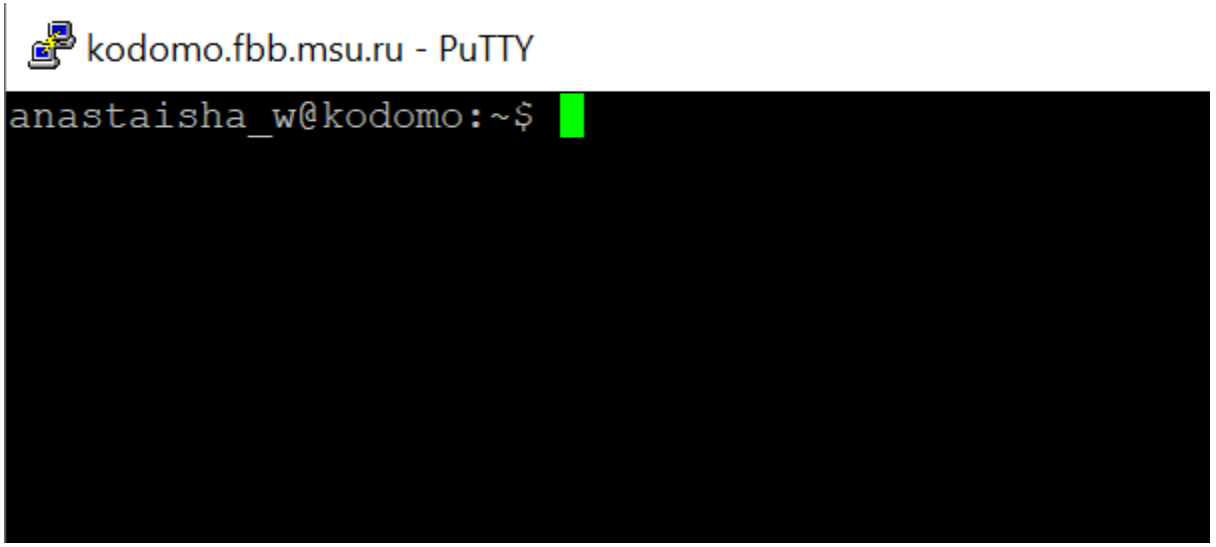
Анастасия Жарикова

24 ноября 2020

azharikova89@gmail.com

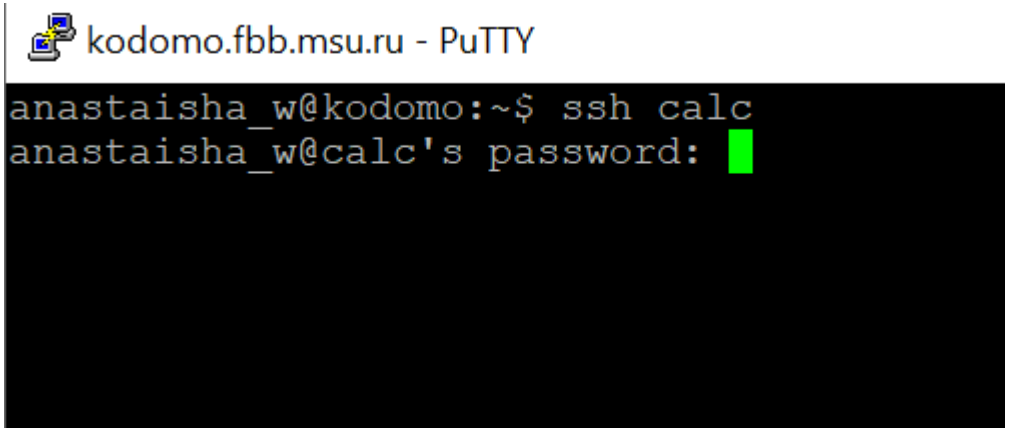
Работа на кластере

Заходим на kodomo



Работа на кластере

Заходим на узел calc



Используем свой пароль от kodomo

Работа на кластере

Идем в рабочую директорию

Здесь есть папка DATA с нашими данными
ТУТ нужно создать свою директорию и работать только в ней!

Поиск вариантов

.bam

.gvcf

```
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  NA12878
20  10001567    .   A   <NON_REF>   .   .   END=10001616    GT:DP:GQ:MIN_DP:PL  0/0:38:99:34:0,101,11
20  10001617    .   C   A,<NON_REF> 493.77  .   BaseQRankSum=1.632;ClippingRankSum=0.000;DP=38;Excess
20  10001618    .   T   <NON_REF>   .   .   END=10001627    GT:DP:GQ:MIN_DP:PL  0/0:39:99:37:0,105,15
20  10001628    .   G   A,<NON_REF> 1223.77 .   DP=37;ExcessHet=3.0103;MLEAC=2,0;MLEAF=1.00,0.00;RAW_
20  10001629    .   G   <NON_REF>   .   .   END=10001660    GT:DP:GQ:MIN_DP:PL  0/0:43:99:38:0,102,12
```

.vcf

```
chr1  28563 . A G 139.90  PASS  AC=2;AF=1.00;AN=2;DP=5;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ
chr1  49298 . T C 515.77  PASS  AC=2;AF=1.00;AN=2;DP=17;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;M
chr1  52238 . T G 716.77  PASS  AC=2;AF=1.00;AN=2;DP=22;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;M
chr1  55926 . T C 120.90  PASS  AC=2;AF=1.00;AN=2;DP=5;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ
chr1  61442 . A G 314.77  PASS  AC=2;AF=1.00;AN=2;DP=10;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;M
chr1  61947 . C T 397.77  PASS  AC=1;AF=0.500;AN=2;BaseQRankSum=3.01;ClippingRankSum=0.00;DP=33;Excess
chr1  61987 . A G 703.77  PASS  AC=1;AF=0.500;AN=2;BaseQRankSum=0.426;ClippingRankSum=0.00;DP=42;Exces
chr1  61989 . G C 703.77  PASS  AC=1;AF=0.500;AN=2;BaseQRankSum=0.125;ClippingRankSum=0.00;DP=41;Exces
chr1  69511 . A G 358.77  PASS  AC=2;AF=1.00;AN=2;DP=13;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;M
chr1  83084 . T A 204.80  PASS  AC=2;AF=1.00;AN=2;DP=7;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ
```

.filt.vcf

```
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA000
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:4
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:4
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:3
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:6
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:4
```

Target

Construction protocol: Genomic DNA was extracted using the DNeasy Blood and Tissue kit (Qiagen) according to the manufacturer's protocol. All the exon genes were captured using a SureSelectXT Human All Exon kit v3 (Agilent) according to the SureSelectXT Target Enrichment for Illumina Paired-End Multiplexed Sequencing Protocol 1.1.1. Enriched libraries were sequenced using Illumina Genome Analyzer IIx. whole-exome sequencing

bam -> gvcf

gatk3 -T HaplotypeCaller

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCF block">
##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, describing how the alternate allele
 another">
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample
cts records within a phasing group">
```

bam -> gvcf

gatk3 -T HaplotypeCaller

```
chr1    565205  .       G       <NON_REF>       .       .       END=565222      GT:DP:GQ:MIN_DP:PL      0/0:3:9:3:0,9,68
chr1    565223  .       A       <NON_REF>       .       .       END=565223      GT:DP:GQ:MIN_DP:PL      0/0:1:3:1:0,3,41
chr1    565224  .       A       <NON_REF>       .       .       END=565269      GT:DP:GQ:MIN_DP:PL      0/0:3:9:3:0,9,68
chr1    565270  .       G       <NON_REF>       .       .       END=565275      GT:DP:GQ:MIN_DP:PL      0/0:4:12:4:0,12,93
chr1    565276  .       T       <NON_REF>       .       .       END=565285      GT:DP:GQ:MIN_DP:PL      0/0:4:9:4:0,9,135
chr1    565286  .       C       T,<NON_REF>     71.03   .       DP=4;ExcessHet=3.0103;MLEAC=2,0;MLEAF=1.00,0.00;RAW_MQ=14400.00
 GT:AD:DP:GQ:PL:SB       1/1:0,4,0:4:12:99,12,0,99,12,99:0,0,2,2
chr1    565287  .       A       <NON_REF>       .       .       END=565317      GT:DP:GQ:MIN_DP:PL      0/0:3:9:3:0,9,62
chr1    565318  .       T       <NON_REF>       .       .       END=565343      GT:DP:GQ:MIN_DP:PL      0/0:2:6:2:0,6,73
chr1    565344  .       G       <NON_REF>       .       .       END=565344      GT:DP:GQ:MIN_DP:PL      0/0:4:12:4:0,12,131
chr1    565345  .       C       <NON_REF>       .       .       END=565405      GT:DP:GQ:MIN_DP:PL      0/0:3:9:3:0,9,68
chr1    565406  .       C       T,<NON_REF>     63.18   .       DP=3;ExcessHet=3.0103;MLEAC=2,0;MLEAF=1.00,0.00;RAW_MQ=10800.00
 GT:AD:DP:GQ:PL:SB       1/1:0,3,0:3:9:90,9,0,90,9,90:0,0,1,2
chr1    565407  .       C       <NON_REF>       .       .       END=565409      GT:DP:GQ:MIN_DP:PL      0/0:3:9:3:0,9,86
chr1    565410  .       C       <NON_REF>       .       .       END=565453      GT:DP:GQ:MIN_DP:PL      0/0:2:6:2:0,6,43
chr1    565454  .       T       C,<NON_REF>     63.56   .       DP=2;ExcessHet=3.0103;MLEAC=2,0;MLEAF=1.00,0.00;RAW_MQ=7200.00
  GT:AD:DP:GQ:PGT:PID:PL:SB       1/1:0,2,0:2:6:0|1:565454_T_C:90,6,0,90,6,90:0,0,1,1
chr1    565455  .       T       <NON_REF>       .       .       END=565463      GT:DP:GQ:MIN_DP:PL      0/0:2:6:2:0,6,43
chr1    565464  .       T       C,<NON_REF>     63.56   .       DP=2;ExcessHet=3.0103;MLEAC=2,0;MLEAF=1.00,0.00;RAW_MQ=7200.00
```

gvcf -> vcf

gatk3 -T GenotypeGVCFs

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=FSs,Description="FS > 60.0">
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=MQRSs,Description="MQRankSum < -12.5">
##FILTER=<ID=MQs,Description="MQ < 40.00">
##FILTER=<ID=QDs,Description="QD < 2.00">
##FILTER=<ID=RPRSs,Description="ReadPosRankSum < -8.0">
##FILTER=<ID=SORs,Description="SOR > 3.0">
##FILTER=<ID=SnpCluster,Description="SNPs found in clusters">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCF block">
##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, describing how the alternate alleles
 are phased in relation to one another">
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample
(but not across samples) connects records within a phasing group">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF spe
cification">
```

gvcf -> vcf

gatk3 -T GenotypeGVCFs

```
chr1    832160  chr1:832160_T/G T       G       62.74   SnpCluster      AC=2;AF=1;AN=2;DP=2;ExcessHet=3.
0103;FS=0;MLEAC=2;MLEAF=1;MQ=57;QD=31.37;SOR=0.693      GT:AD:DP:GQ:PGT:PID:PL  1/1:0,2:2:6:1|1:832152_C
_A:90,6,0
chr1    832318  chr1:832318_C/A C       A       42.74   LowQual AC=2;AF=1;AN=2;DP=2;ExcessHet=3.0103;FS=
0;MLEAC=2;MLEAF=1;MQ=60;QD=21.37;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:70,6,0
chr1    832398  chr1:832398_T/C T       C       21.77   LowQual AC=2;AF=1;AN=2;DP=2;ExcessHet=3.0103;FS=
0;MLEAC=2;MLEAF=1;MQ=60;QD=10.88;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:49,6,0
chr1    849998  chr1:849998_A/G A       G       21.77   LowQual AC=2;AF=1;AN=2;DP=2;ExcessHet=3.0103;FS=
0;MLEAC=2;MLEAF=1;MQ=60;QD=10.88;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:49,6,0
chr1    858691  chr1:858691_TG/T        TG      T       42.7    LowQual AC=2;AF=1;AN=2;DP=2;ExcessHet=3.
0103;FS=0;MLEAC=2;MLEAF=1;MQ=60;QD=21.35;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:79,6,0
chr1    1148419 chr1:1148419_C/T        C       T       21.77   LowQual AC=2;AF=1;AN=2;DP=2;ExcessHet=3.
0103;FS=0;MLEAC=2;MLEAF=1;MQ=60;QD=10.88;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:49,6,0
chr1    1254436 chr1:1254436_A/G        A       G       776.77  PASS    AC=1;AF=0.5;AN=2;BaseQRankSum=-1
.167;ClippingRankSum=0;DP=68;ExcessHet=3.0103;FS=0.984;MLEAC=1;MLEAF=0.5;MQ=60;MQRankSum=0;QD=11.77;Read
PosRankSum=-0.035;SOR=0.909     GT:AD:DP:GQ:PGT:PID:PL  0/1:43,23:66:99:0|1:1254436_A_G:805,0,2318
chr1    1254443 chr1:1254443_G/A        G       A       762.77  PASS    AC=1;AF=0.5;AN=2;BaseQRankSum=0.
067;ClippingRankSum=0;DP=67;ExcessHet=3.0103;FS=0;MLEAC=1;MLEAF=0.5;MQ=60;MQRankSum=0;QD=11.73;ReadPosRa
nkSum=-0.195;SOR=0.776  GT:AD:DP:GQ:PGT:PID:PL  0/1:42,23:65:99:0|1:1254436_A_G:791,0,2319
chr1    1369029 chr1:1369029_C/G        C       G       21.77   LowQual AC=2;AF=1;AN=2;DP=2;ExcessHet=3.
0103;FS=0;MLEAC=2;MLEAF=1;MQ=60;QD=10.88;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:49,6,0
chr1    1686040 chr1:1686040_G/T        G       T       296.77  PASS    AC=1;AF=0.5;AN=2;BaseQRankSum=-1
```

Vcf filtration

gatk3 -T SelectVariants

https://gatk.broadinstitute.org/hc/en-us/articles/360035531112--How-to-Filter-variants-either-with-VQSR-or-by-hard-filtering

```
gatk3 -T VariantFiltration \
      -R $g \
      -o $namespace.snp_filt.vcf \
      --variant $namespace.raw_snps.vcf \
      --clusterSize 3 \
      --clusterWindowSize 10 \
      --filterExpression "MQ < 40.00" \
      --filterName "MQs" \
      --filterExpression "QD < 2.00" \
      --filterName "QDs" \
      --filterExpression "FS > 60.0" \
      --filterName "FSs" \
      --filterExpression "SOR > 3.0" \
      --filterName "SORs" \
      --filterExpression "MQRankSum < -12.5" \
      --filterName "MQRSs" \
      --filterExpression "QUAL < $QuAL_tHreSH" \
      --filterName "LowQual" \
      --filterExpression "ReadPosRankSum < -8.0" \
      --filterName "RPRSs" 2>> logs/$namespace.gatk3_filt.log
```

```
gatk3 -T VariantFiltration \
      -R $g \
      -o $namespace.indel_filt.vcf \
      --variant $namespace.raw_indels.vcf \
      --filterExpression "QD < 2.00" \
      --filterName "QDi" \
      --filterExpression "FS > 200.0" \
      --filterName "FSi" \
      --filterExpression "QUAL < $QuAL_tHreSH" \
      --filterName "LowQual" \
      --filterExpression "ReadPosRankSum < -20.0" \
      --filterName "RPRSi" 2>> logs/$namespace.gatk3_filt.log
```

Объединение snp + indel

bcftools

http://samtools.github.io/bcftools/bcftools.html

?Аннотация?

ANNOVAR

https://doc-openbio.readthedocs.io/projects/annovar/en/latest/

refgene

dbsnp

1000 genomes

GWAS

Clinvar

Далее – клиническая интерпретация

Bedtools

http://bedtools.readthedocs.io/en/latest/index.html

https://media.readthedocs.org/pdf/bedtools/latest/bedtools.pdf

Очень хороший инструмент для работы с геномными интервалами

Более 35 опций + параметры

Bamtobed

```
$ bedtools bamtobed -i reads.bam | head -3
chr7    118970079    118970129    TUPAC_0001:3:1:0:1452#0/1    37    -
chr7    118965072    118965122    TUPAC_0001:3:1:0:1452#0/2    37    +
chr11   46769934     46769984     TUPAC_0001:3:1:0:1472#0/1    37    -
```

Bamtofastq

```
$ bedtools bamtofastq -i NA18152.bam -fq NA18152.fq

$ head -8 NA18152.fq
@NA18152-SRR007381.35051
GGAGACATATCATATAAGTAATGCTAGGGTGAGTGGTAGGAAGTTTTTTCATAGGAGGTGTATGAGTTGGTCGTAGCGGAATCGGGGGTATGCTGTTCG
+
<<<;;<;<;;<;;;;;;;;;;;;<<<:;;;;;;;;;;;;;;;;;;;;;:::::;;;;<<;;;;;;;;;;;;;;;;;;;;;;;;;;;;<<<<<;;;;;;;;;;
@NA18152-SRR007381.637219
AATGCTAGGGTGAGTGGTAGGAAGTTTTTTCATAGGAGGTGTATGAGTTGGTCGTAGCGGAATCGGGGGTATGCTGTTCGAATTCATAAGAACAGGGAC
+
<<<<<<<<<<;;<;<;;;;<<;<888888899<;;;;;;<;;;;;;;;;;;;;;;;;;;;;;;;;;<<<<<;;;;;;;;;<;<<<<<;;;;;;;;;;;;
```

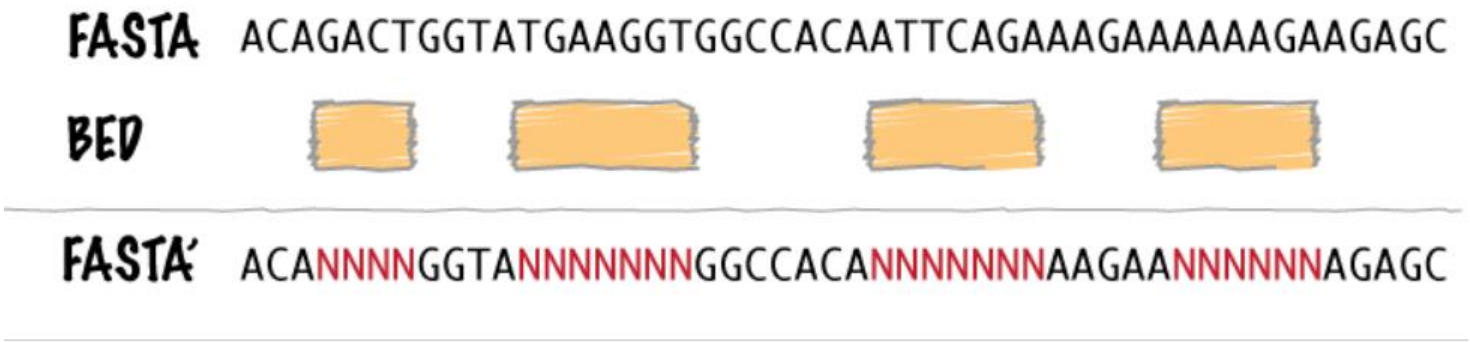
Bedtobam

```
head -5 rmsk.hg18.chr21.bed
chr21 9719768  9721892  ALR/Alpha  1004  +
chr21 9721905  9725582  ALR/Alpha  1010  +
chr21 9725582  9725977  L1PA3 3288 +
chr21 9726021  9729309  ALR/Alpha  1051  +
chr21 9729320  9729809  L1PA3 3897 -

bedToBam -i rmsk.hg18.chr21.bed -g human.hg18.genome > rmsk.hg18.chr21.bam

samtools view rmsk.hg18.chr21.bam | head -5
ALR/Alpha  0   chr21 9719769  255  2124M *  0  0  *  *
ALR/Alpha  0   chr21 9721906  255  3677M *  0  0  *  *
L1PA3      0   chr21 9725583  255  395M  *  0  0  *  *
ALR/Alpha  0   chr21 9726022  255  3288M *  0  0  *  *
L1PA3      16  chr21 9729321  255  489M  *  0  0  *  *
```

Getfasta


```
$ cat test.fa
>chr1
AAAAAAAACCCCCCCCCCCCCGCTACTGGGGGGGGGGGGGGGGG

$ cat test.bed
chr1 5 10

$ bedtools getfasta -fi test.fa -bed test.bed
>chr1:5-10
AAACC
```

Maskfasta



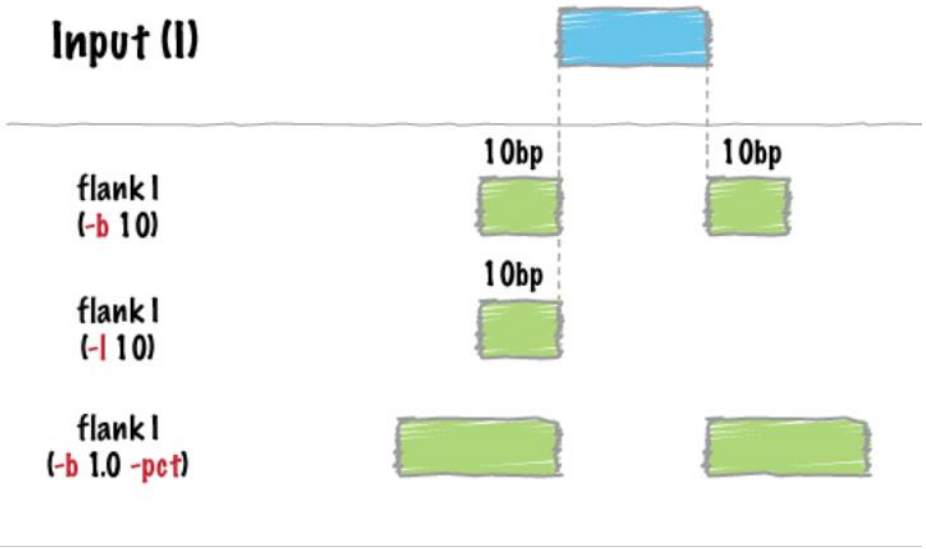

```
$ cat test.fa
>chr1
AAAAAAAACCCCCCCCCCCCCGCTACTGGGGGGGGGGGGGGGGG

$ cat test.bed
chr1 5 10

$ bedtools maskfasta -fi test.fa -bed test.bed -fo test.fa.out

$ cat test.fa.out
>chr1
AAAAANNNNNCCCCCCCCCCGCTACTGGGGGGGGGGGGGGGGG
```

Maskfasta

```
$ bedtools maskfasta -fi test.fa -bed test.bed -fo test.fa.out -soft

$ cat test.fa.out
>chr1
AAAAAaaaccCCCCCCCCCCGCTACTGGGGGGGGGGGGGGGGG
```

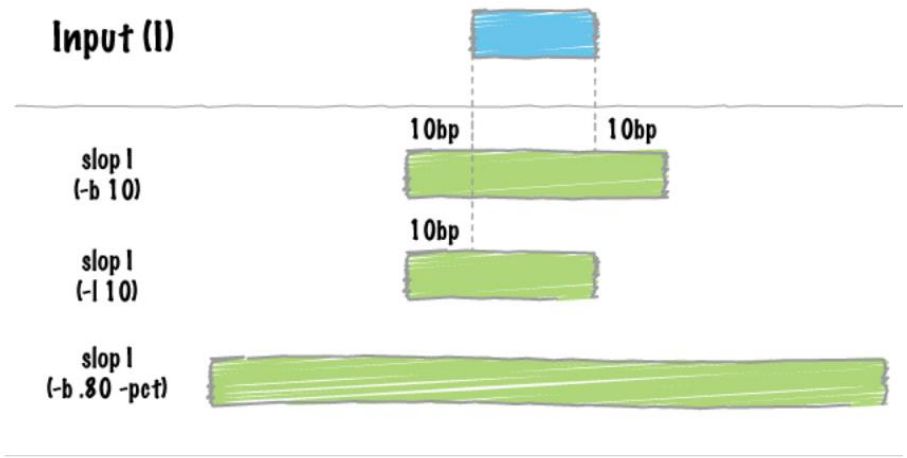
```
$ bedtools maskfasta -fi test.fa -bed test.bed -fo test.fa.out -mc X

$ cat test.fa.out
>chr1
AAAAAXXXXXCCCCCCCCCCGCTACTGGGGGGGGGGGGGGGGG
```

Flank




```
$ cat A.bed
chr1 100 200
chr1 500 600

$ cat my.genome
chr1 1000

$ bedtools flank -i A.bed -g my.genome -b 5
chr1  95      100
chr1  200     205
chr1  495     500
chr1  600     605

$ bedtools flank -i A.bed -g my.genome -l 2 -r 3
chr1  98      100
chr1  200     203
chr1  498     500
chr1  600     603
```

Slop

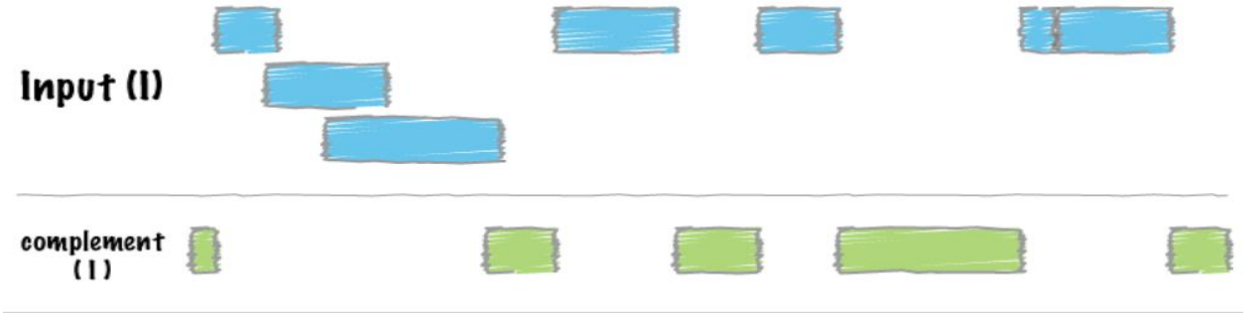



```
$ cat A.bed
chr1 5 100
chr1 800 980

$ cat my.genome
chr1 1000

$ bedtools slop -i A.bed -g my.genome -b 5
chr1 0 105
chr1 795 985

$ bedtools slop -i A.bed -g my.genome -l 2 -r 3
chr1 3 103
chr1 798 983
```

Shift

```
$ cat A.bed
chr1 5 100 +
chr1 800 980 -

$ cat my.genome
chr1 1000

$ bedtools shift -i A.bed -g my.genome -s 5
chr1 10 105 +
chr1 805 985 -

$ bedtools shift -i A.bed -g my.genome -p 2 -m -3
chr1 7 102 +
chr1 797 977 -
```

Complement

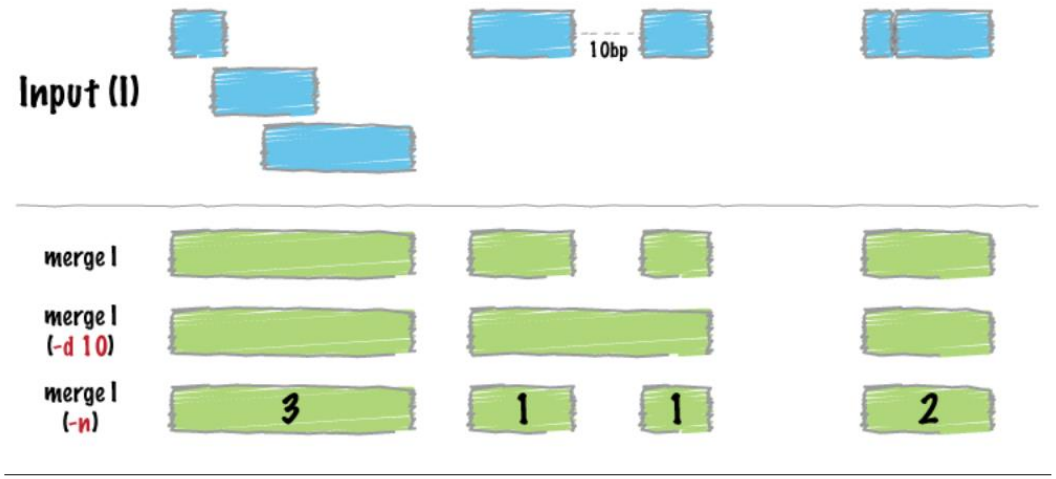



```
$ cat A.bed
chr1  100  200
chr1  400  500
chr1  500  800

$ cat my.genome
chr1  1000
chr2  800

$ bedtools complement -i A.bed -g my.genome
chr1  0   100
chr1  200 400
chr1  800 1000
chr2  0   800
```

Subtract

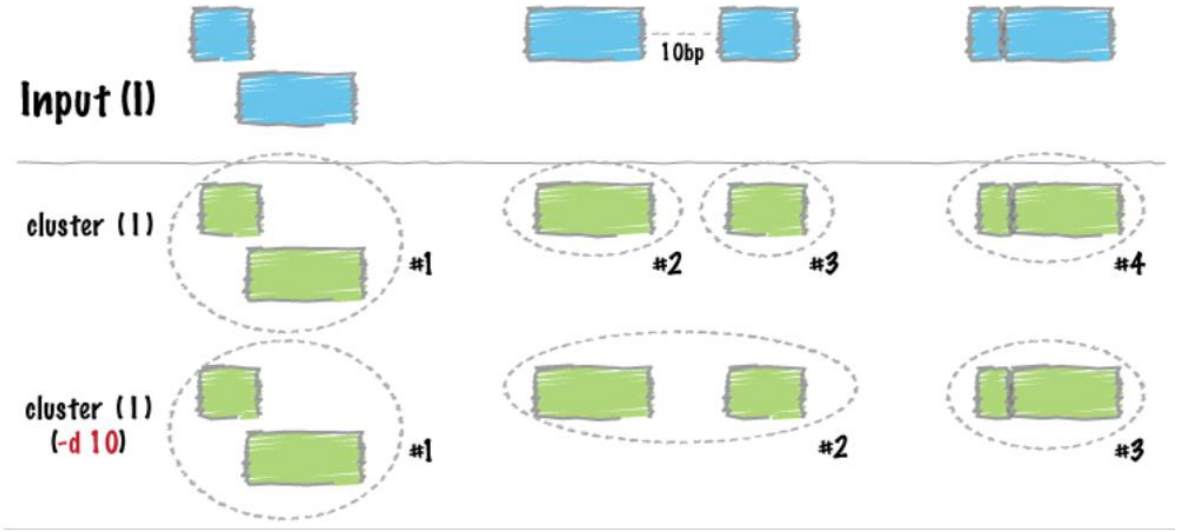

```
$ cat A.bed
chr1  10   20
chr1  100  200

$ cat B.bed
chr1  0    30
chr1  180  300

$ bedtools subtract -a A.bed -b B.bed
chr1  100  180
```

Merge



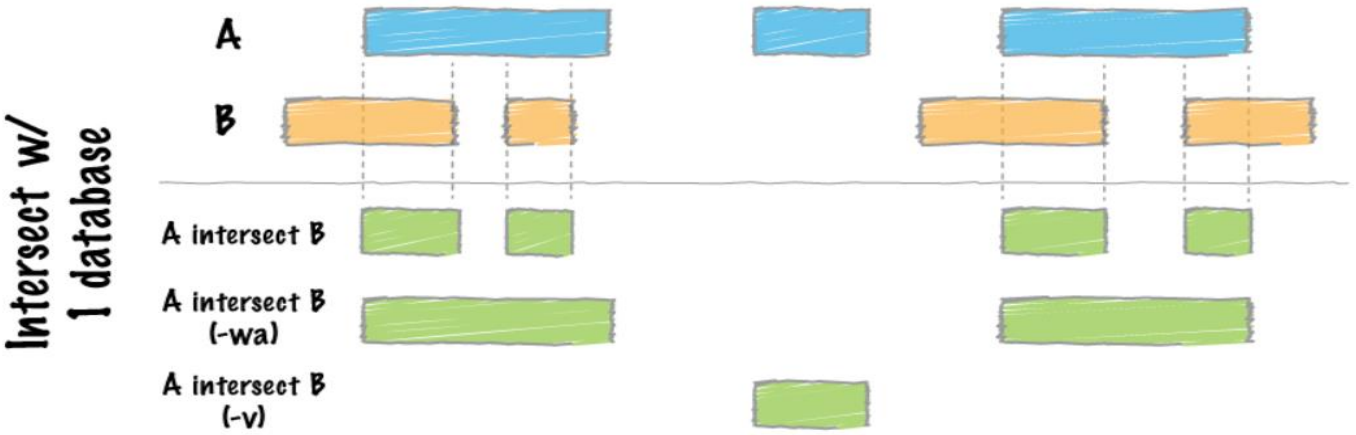

```
$ cat A.bed
chr1  100  200
chr1  180  250
chr1  250  500
chr1  501  1000

$ bedtools merge -i A.bed
chr1  100  500
chr1  501  1000
```

Cluster




```
$ cat A.bed
chr1  100  200
chr1  180  250
chr1  250  500
chr1  501  1000

$ bedtools cluster -i A.bed
chr1  100     200     1
chr1  180     250     1
chr1  250     500     1
chr1  501     1000    2
```

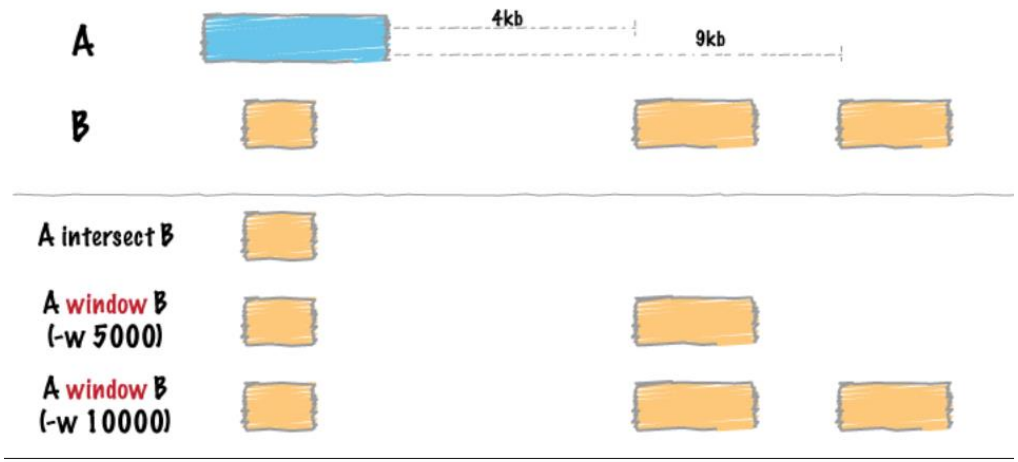
Intersect




```
$ cat A.bed
chr1  10  20
chr1  30  40

$ cat B.bed
chr1  15  20

$ bedtools intersect -a A.bed -b B.bed
chr1  15  20
```

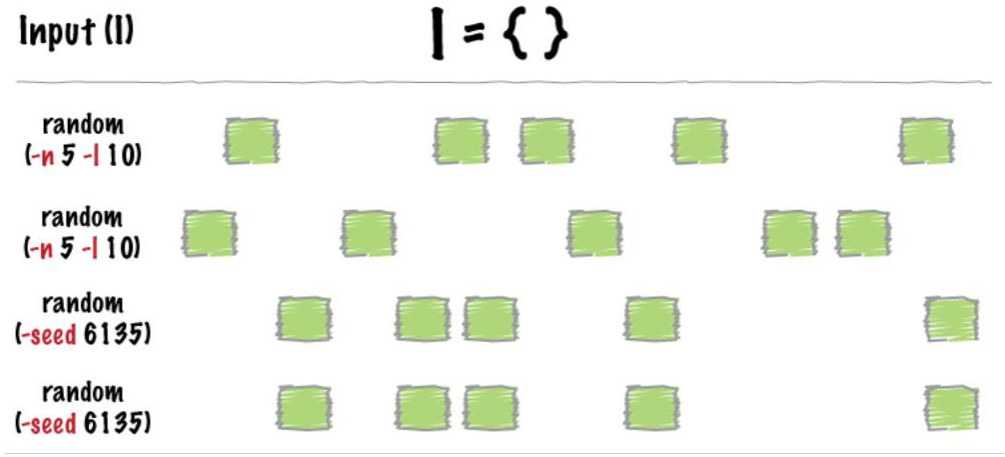
```
$ cat A.bed
chr1    10    20
chr1    30    40

$ cat B.bed
chr1    15  20
chr1    18  25

$ bedtools intersect -a A.bed -b B.bed -c
chr1    10    20    2
chr1    30    40    0
```

Window




```
$ cat A.bed
chr1  100  200

$ cat B.bed
chr1  500  1000
chr1  1300 2000

$ bedtools window -a A.bed -b B.bed -w 5000
chr1  100  200  chr1  500   1000
chr1  100  200  chr1  1300  2000
```

Random




```
$ bedtools random -g hg19.genome -n 3
chr20 47975280        47975380        1       100     -
chr16 23381222        23381322        2       100     +
chr3  104913816       104913916       3       100     -
```

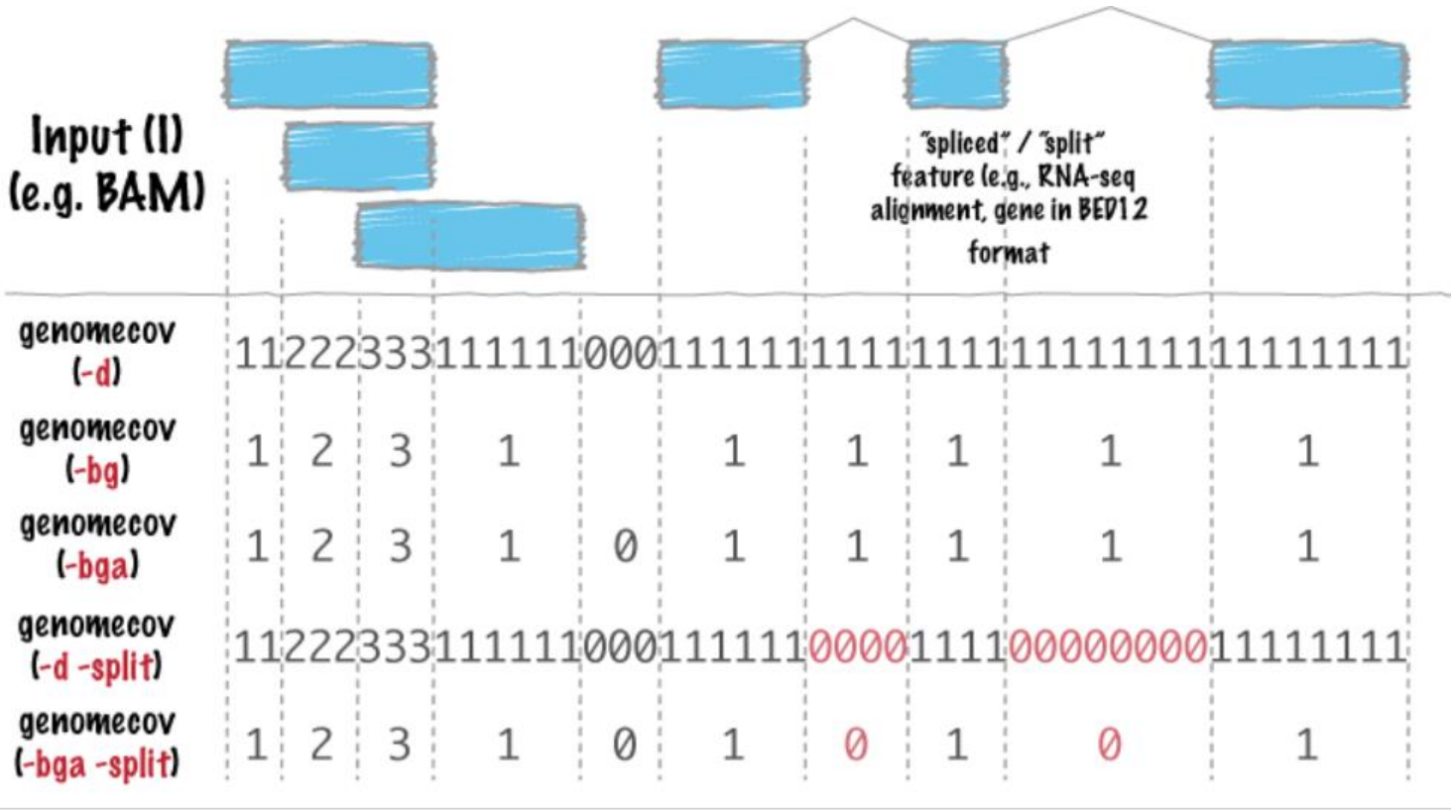
Shuffle


```
$ cat A.bed
chr1  0  100  a1  1  +
chr1  0  1000 a2  2  -

$ cat my.genome
chr1  10000
chr2  8000
chr3  5000
chr4  2000

$ bedtools shuffle -i A.bed -g my.genome -chrom
chr1  9560  9660  a1  1  +
chr1  7258  8258  a2  2  -
```

Coverage

```
$ cat A.bed
chr1  0   100
chr1  100 200
chr2  0   100

$ cat B.bed
chr1  10  20
chr1  20  30
chr1  30  40
chr1  100 200

$ bedtools coverage -a A.bed -b B.bed
chr1  0   100  3  30  100 0.3000000
chr1  100 200  1  100 100 1.0000000
chr2  0   100  0  0   100 0.0000000
```

Genomecov


Input (I)
(e.g. BAM)
"spliced" / "split"
feature (e.g., RNA-seq
alignment, gene in BED12
format
genomecov
(-d)
11222333111111000111111111111111111111111111111
genomecov
(-bg)
1 2 3 1 1 1 1 1 1
genomecov
(-bga)
1 2 3 1 0 1 1 1 1 1
genomecov
(-d -split)
11222333111111000111111000011110000000011111111
genomecov
(-bga -split)
1 2 3 1 0 1 0 1 0 1

Genomecov

1. chromosome (or entire genome)
2. depth of coverage from features in input file
3. number of bases on chromosome (or genome) with depth equal to column 2.
4. size of chromosome (or entire genome) in base pairs
5. fraction of bases on chromosome (or entire genome) with depth equal to column 2.

```
$ cat A.bed
chr1  10  20
chr1  20  30
chr2  0   500

$ cat my.genome
chr1  1000
chr2  500

$ bedtools genomecov -i A.bed -g my.genome
chr1   0  980  1000  0.98
chr1   1  20   1000  0.02
chr2   1  500  500   1
genome 0  980  1500  0.653333
genome 1  520  1500  0.346667
```

Genomecov

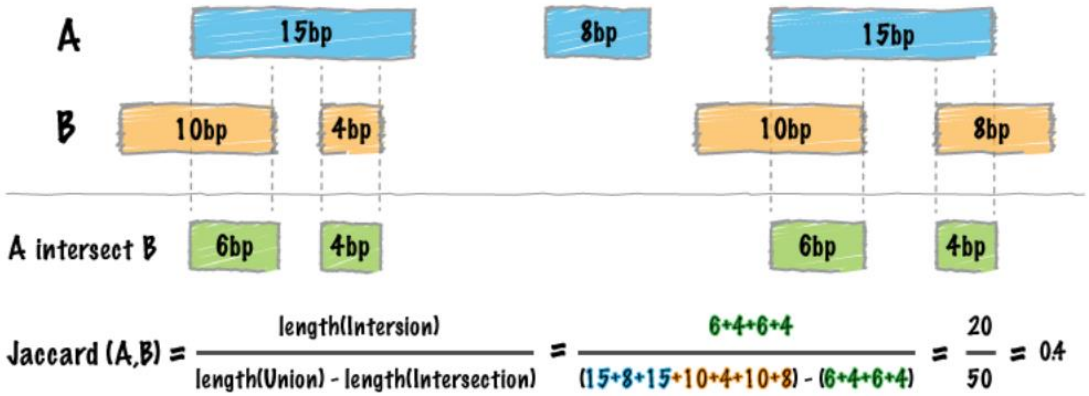
```
$ cat A.bed
chr1  10  20
chr1  20  30
chr2  0   500

$ cat my.genome
chr1  1000
chr2  500

$ bedtools genomecov -i A.bed -g my.genome -d | \
      head -15 | \
      tail -n 10
chr1  6  0
chr1  7  0
chr1  8  0
chr1  9  0
chr1  10 0
chr1  11 1
chr1  12 1
chr1  13 1
chr1  14 1
chr1  15 1
```

Genomecov

```
$ bedtools genomecov -ibam NA18152.bam -bg | head
chr1  554304  554309  5
chr1  554309  554313  6
chr1  554313  554314  1
chr1  554315  554316  6
chr1  554316  554317  5
chr1  554317  554318  1
chr1  554318  554319  2
chr1  554319  554321  6
chr1  554321  554323  1
chr1  554323  554334  7
```

```
$ bedtools genomecov -ibam NA18152.bam -bga | head
chr1  0       554304  0
chr1  554304  554309  5
chr1  554309  554313  6
chr1  554313  554314  1
chr1  554314  554315  0
chr1  554315  554316  6
chr1  554316  554317  5
chr1  554317  554318  1
chr1  554318  554319  2
chr1  554319  554321  6
```

Jaccard




```
$ cat a.bed
chr1  10  20
chr1  30  40

$ cat b.bed
chr1  15   20

$ bedtools jaccard -a a.bed -b b.bed
intersection  union   jaccard n_intersections
5     20      0.25    1
```

Makewindows

hg19.txt:
chr1 249250621
chr2 243199373
...
chr18_gl000207_random 4262

bedtools makewindows -g hg19.txt -w 1000000
chr1 0 1000000
chr1 1000000 2000000
chr1 2000000 3000000
chr1 3000000 4000000
chr1 4000000 5000000

Sort

```
cat A.bed
chr1 800 1000
chr1 80  180
chr1 1   10
chr1 750 10000

sortBed -i A.bed
chr1 1   10
chr1 80  180
chr1 750 10000
chr1 800 1000
```

```
cat A.bed
chr1 800 1000
chr1 80  180
chr1 1   10
chr1 750 10000

sortBed -i A.bed -sizeD
chr1 750 10000
chr1 800 1000
chr1 80  180
chr1 1   10
```

```
sort -k 1,1 -k2,2n a.bed
chr1 1   10
chr1 80  180
chr1 750 10000
chr1 800 1000
```

И это всё еще далеко не всё …