getfasta

`bedtools getfasta` extracts sequences from a FASTA file for each of the intervals defined in a BED/GFF/VCF file.

Tip

1. The headers in the input FASTA file must *exactly* match the chromosome column in the BED file.

2. You can use the UNIX `fold` command to set the line width of the FASTA output. For example, `fold -w 60` will make each line of the FASTA file have at most 60 nucleotides for easy viewing.

3. BED files containing a single region require a newline character at the end of the line, otherwise a blank output file is produced.

See also

[maskfasta](#)

Usage and option summary

Usage

```
$ bedtools getfasta [OPTIONS] -fi <input FASTA> -bed <BED/GFF/VCF>
```

(or):

```
$ getFastaFromBed [OPTIONS] -fi <input FASTA> -bed <BED/GFF/VCF>
```

Option	Description
-fo	Specify an output file name. By default, output goes to stdout.
-name	Use the "name" column in the BED file for the FASTA headers in the output FASTA file.

v: latest ▾

Option	Description
-tab	Report extract sequences in a tab-delimited format instead of in FASTA format.
-bedOut	Report extract sequences in a tab-delimited BED format instead of in FASTA format.
-s	Force strandedness. If the feature occupies the antisense strand, the sequence will be reverse complemented. *Default: strand information is ignored*.
-split	Given BED12 input, extract and concatenate the sequences from the BED "blocks" (e.g., exons)

Default behavior

`bedtools getfasta` will extract the sequence defined by the coordinates in a BED interval and create a new FASTA entry in the output file for each extracted sequence. By default, the FASTA header for each extracted sequence will be formatted as follows: "<chrom>:<start>-<end>".

```
$ cat test.fa
>chr1
AAAAAAAACCCCCCCCCCCCCCGCTACTGGGGGGGGGGGGGGGGGGGG

$ cat test.bed
chr1 5 10

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

# optionally write to an output file
$ bedtools getfasta -fi test.fa -bed test.bed -fo test.fa.out

$ cat test.fa.out
>chr1:5-10
AAACC
```

-name Using the BED "name" column as a FASTA header.

Using the `-name` option, one can set the FASTA header for each extracted sequence to be the "name" columns from the BED feature.

```
$ cat test.fa
>chr1
AAAAAAAACCCCCCCCCCCCCCGCTACTGGGGGGGGGGGGGGGGGGGG

$ cat test.bed
chr1 5 10 myseq

$ bedtools getfasta -fi test.fa -bed test.bed -name
>myseq
AAACC
```

v: latest ▾

-tab Creating a tab-delimited output file in lieu of FASTA output.

Using the -tab option, the -fo output file will be tab-delimited instead of in FASTA format.

```
$ cat test.fa
>chr1
AAAAAAAACCCCCCCCCCCCCCGCTACTGGGGGGGGGGGGGGGGGGGG

$ cat test.bed
chr1 5 10 myseq

$ bedtools getfasta -fi test.fa -bed test.bed -name -tab
myseq AAACC
```

-bedOut Creating a tab-delimited BED file in lieu of FASTA output.

Using the -tab option, the -fo output file will be tab-delimited instead of in FASTA format.

```
$ cat test.fa
>chr1
AAAAAAAACCCCCCCCCCCCCCGCTACTGGGGGGGGGGGGGGGGGGGG

$ cat test.bed
chr1 5 10 myseq

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

-s Forcing the extracted sequence to reflect the requested strand

bedtools getfasta will extract the sequence in the orientation defined in the strand column when the "-s" option is used.

```
$ cat test.fa
>chr1
AAAAAAAACCCCCCCCCCCCCCGCTACTGGGGGGGGGGGGGGGGGGGG

$ cat test.bed
chr1 20 25 forward 1 +
chr1 20 25 reverse 1 -

$ bedtools getfasta -fi test.fa -bed test.bed -s -name
>forward
CGCTA
>reverse
TAGCG
```

v: latest ▾

-split Extracting BED "blocks".

One can optionally request that FASTA records be extracting and concatenating each block in a BED12 record. For example, consider a BED12 record describing a transcript. By default, getfasta will extract the sequence representing the entire transcript (intons, exons, UTRs). Using the -split option, getfasta will instead produce separate a FASTA record representing a transcript that splices together each BED12 block (e.g., exons and UTRs in the case of genes described with BED12).

```
$ cat genes.bed12
chr1    164404  173864  ENST00000466557.1       0       -       173864  173864
chr1    235855  267253  ENST00000424587.1       0       -       267253  267253
chr1    317810  328455  ENST00000426316.1       0       +       328455  328455

$ bedtools getfasta -fi chr1.fa -bed genes.bed12 -split -name
>ENST00000466557.1
gaggcgggaagatcacttgatatcaggagtcgaggcgggaagatcacttgacgtcaggagttcgagactggcccggcc
>ENST00000424587.1
ccaggaagtgaaaatgacactttactgttttaatttgcatttctctgcttacaagtggattacacacattttcgtgtg
>ENST00000426316.1
AATGATCAAATTATGTTTCCCATGCATCAGGTGCAATGGGAAGCTCTTctggagagtgagagaagcttccagttaagg

# use the UNIX fold command to wrap the FASTA sequence such that each line
# has at most 60 characters
$ bedtools getfasta -fi chr1.fa -bed genes.bed12 -split -name | \
      fold -w 60
>ENST00000466557.1
gaggcgggaagatcacttgatatcaggagtcgaggcgggaagatcacttgacgtcaggag
ttcgagactggcccggccaacatggtgaaaccgcatctccactaaaaatacaaaaattag
cctggtatggtggtgggcacctgtaatcccagtgacttgggaggctaaggcaggagaatt
tcttgaacccaggaggcagaggttgcagtgaccagcaaggttgcgccattgcacccccagc
ctgggcgataagagtgaaaactccatctcaaaaaaaaaaaaaaaaaaaaaaaTTCCTTTGG
GAAGGCCTTCTACATAAAAATCTTCAACATGAGACTGGAAAAAAGGGTATGGGATCATCA
CCGGACCTTTGGCTTTTACAGCTCGAGCTGACAAAGTTGATTTATCAAGTTGTAAATCTT
CACCTGTTGAATTCATAAGTTCATGTCATATTTTCTTTCAGACAATTCTTCAGTTTGTTT
ACGTAGATCAGCGATACGATGATTCCATTTCTtcggatccttgtaagagcagagcaggtg
atggagagggtgggaggtgtagtgacagaagcaggaaactccagtcattcgagacgggca
gcacaagctgcggagtgcaggccacctctacggccaggaaacggattctcccgcagagcc
tcggaagctaccgaccctgctcccaccttgactcagtaggacttactgtagaattctggc
cttcagacCTGAGCCTGGCAGCTCTCTCCAACTTTGGAAGCCCAGGGGCATGGCCCCTGT
CCACAGATGCACCTGGCATGAGGCGTGCCCAGAGGGACAGAGGCAGATGAGTttcgtctc
ctccactggattgtgagggcCAGAGTTGAACTCCCTCATTTTCCGTTCCCCAGCATTGGC
AGGTTCTGGGACTGGTGGCTGTGGTGGCTCGTTGGTCTTTGTCTCTTAGAAGGTGGGGAA
TAATCATCATCT
>ENST00000424587.1
ccaggaagtgaaaatgacactttactgttttaatttgcatttctctgcttacaagtggat
tacacacattttcgtgtgctgttggctacttatTCATTCAGAAAACATACTAAGTGCTGG
CTCTTTTTCATGTCCTTTATCAAGTTTGGATCATGTCATTTGCTATTTTCTTTCTGATGT
AAACTCTCAAAGTCTGAAGTGTATTGTCTTTTCCTGACACATATGTTGTAAATAATTTTC
TGGCTTACATTTTGACTTTTAATTTCATTCACGATGTTTTTAATGAATAATTTTAATTTT
TATGAATGCAAGTTAAAATAATTCTTTCATTGTGGTCTCTGACATGTCATGCCAATAAGG
GTCTTCTCCTCCAAGAGCACAGAAATATTTGCCAATACTGTCCTTAAAATCGGTCACAGT
TTCATTTTTTATATATGCATTTTACTTCAATTGGGGCTTCATTTTACTGAATGCCCTATT
TGAAGCAAGTTTCTCAGTTAATTCTTTTTCTCAAAGGGCTAAGTATGGTAGATTGCAAACA
TAAGTGGCCACATAATGCTCTCACCTCcttttgcctcctctcccaggaggagatagcgtcc
```

v: latest

```
atctttccactccttaatctgggcttggccgtgtgacttgcactggccaatgggatatta
acaagtctgatgtgcacagaggctgtagaatgtgcacgggggcttggtctctcttgctgc
cctggagaccagctgccCCACGAAGGAACCAGAGCCAACCTGCTGCTTCCTGGAGGAAGA
CAGTCCCTCTGTCCCTCTGTCTCTGCCAACCAGTTAACCTGCTGCTTCCTGGAGGGAGAC
AGTCCCTCAGTCCCTCTGTCTCTGCCAACCAGTTAACCTGCTGCTTCCTGGAGGAAGACA
GTCACTCTGTCTCTGccacccagttgaccgcagacatgcaggtctgctcaggtaagacc
agcacagtccctgccctgtgagccaaaccaaatggtccagccacagaatcgtgagcaaat
aagtgatgcttaagtcactaagatttgggCAAAAGCTGAGCATTTATCCCAATCCCAATA
CTGTTTGTCCTTCTGTTTATCTGTCTGTCCTTCCCTGCTCATTTAAAATGCCCCCACTGC
ATCTAGTACATTTTTATAGGATCAGGGATCTGCTCTTGGATTAATGTTGTGTTCCCACCT
CGAGGCAGCTTTGTAAGCTTCTGAGCACTTCCCAATTCCGGGTGACTTCAGGCACTGGGA
GGCCTGTGCATCAGCTGCTGCTGTCTGTAGCTGACTTCCTTCACCCCTCTGCTGTCCTCA
GCTCCTTCACCCCTGGGCCTCAGGAAATCAATGTCATGCTGACATCACTCTAGATCTAAA
AGTTGGGTTCTTGgaccaggcgtggtggctcacacctgtaatcccagcactttgggaggc
cgaggcgggtggatcacaaggtcaggagatcaagacgattctggctaacacggtgaaacc
ccgtctctactaaaaatacaaaaaaattagccgggtgtggtggcaggtgcctgtagcccc
agctacttgggaggctgaggcaggagaatggcttgaacctgggaggtggagcttgcagtg
agccaagatcacgccactgcactccagaatgggagagagagcgagactttctcaaaaaaa
aaaaaaaaaCTTAGGTTCTTGGATGTTCGGGAAAGGGGGTTATTATCTAGGATCCTTGAA
GCACCCCAAGGGCATCTTCTCAAAGTTGGATGTGTGCATTTTCCTGAGAGGAAAGCTTT
CCCACATTATACAGCTTCTGAAAGGGTTGCTTGACCCACAGATGTGAAGCTGAGGCTGAA
GGAGACTGATGTGGTTTCTCCTCAGTTTCTCTGTGCAGCACCAGGTGGCAGCAGAGGTCA
GCAAGGCAAACCCGAGCCCGGGGATGCGGAGTGGGGGCAGCTACGTCCTCTCTTGAGCTA
CAGCAGATTCACTCTGTTCTGTTTCATTGTTGTTTAGTTTGCGTTGTGTTTCTCCAACTT
TGTGCCTCATCAGGAAAGCTTTGGATCACAATTCCCAGtgctgaagaaaaggccaaact
cttggttgtgttctttgattAGTgcctgtgacgcagcttcaggaggtcctgagaacgtgt
gcacagtttagtcggcagaaacttagggaaatgtaagaccaccatcagcacataggagtt
ctgcattggtttggtctgcattggtttggtCTTTTCCTGGATACAGGTCTTGATAGGTCT
CTTGATGTCATTTCACTTCAGATTCTTCTTTAGAAAACTTGGACAATAGCATTTGCTGTC
TTGTCCAAATTGTTACTTCAAGTTTGCTCTTAGCAAGTAATTGTTTCAGTATCTATATCA
AAAATGGCTTAAGCCTGCAACATGTTTCTGAATGATTAACAAGGTGATAGTCAGTTCTTC
ATTGAATCCTGGATGCTTTATTTTTCTTAATAAGAGGAATTCATATGGATCAG
>ENST00000426316.1
AATGATCAAATTATGTTTCCCATGCATCAGGTGCAATGGGAAGCTCTTctggagagtgag
agaagcttccagttaaggtgacattgaagccaagtcctgaaagatgaggaagagttgtat
gagagtggggagggaaggggggaggtggagggaTGGGGAATGGGCCGGGATGGGATAGCGC
AAACTGCCCGGGAAGGGAAACCAGCACTGTACAGACCTGAACAACGAAGATGGCATATTT
TGTTCAGGGAATGGTGAATTAAGTGTGGCAGGAATGCTTTGTAGACACAGTAATTTGCTT
GTATGGAATTTTGCCTGAGAGACCTCATTCCTCACGTCGGCCATTCCAGGCCCCGTTTTT
CCCTTCCGGCAGCCTCTTGGCCTCTAATTTGTTTATCTTTTGTGTATAAATCCCAAAATA
TTGAATTTTGGAATATTTCCACCATTATGTAAATATTTTGATAGGTAA
```

📗 v: latest ▾