annotk user's guide

annotk is a python script used to annotate a text file.
Annotation process consists of adding to a target file, the information contained in a reference file.
Correspondence between target and reference lines is based either on genomic locations (position or region) or equality of values between column(s) of the two files (e.g. a 'Gene symbol' or any other identifiers)

By instance, the process begins with a file containing a set of regions found differentially methylated between samples of two conditions (DMR : differentially methylated regions). These DMRs are specified with 3 values :

1. Chromosome
2. Genomic coordinate for the start of the DMR
3. Genomic coordinate for the end of the DMR

Some other columns maybe specified in the target file (here 'Methylation difference') :

Chromosome	Start	End	Methylation difference
chr1	20108	20256	18.6
chr1	20257	20312	11.8
chr1	20313	20590	18.4
chr1	20618	20819	-52.8
chr1	20618	20819	99.6
chr1	21130	21379	-31.8
chr1	21130	21379	65.4
chr1	47348	47398	-25.3
chr1	58653	58814	-4.6
chr1	70547	70781	-8.8
chr1	75843	75955	-11.9
chr1	115996	116053	93.6

Suppose we want to annotate this file with gene feature information (promoter, 5' UTR, exon, intron, …) defined in a GTF reference file.

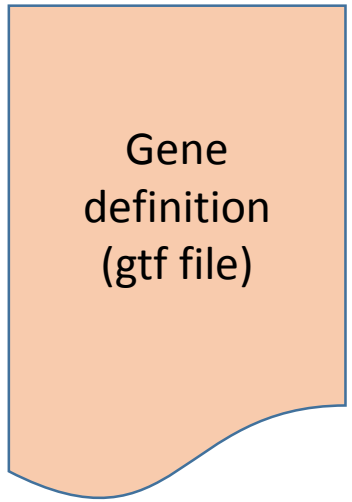
(Nota Bene : Both target and reference files must have a header as first line, with the name of the different columns)

Target file #1:

Chromosome	Start	End	Methylation difference
chr1	20108	20256	18.6
chr1	20257	20312	11.8
chr1	20313	20590	18.4
chr1	20618	20819	-52.8
chr1	20618	20819	99.6
chr1	21130	21379	-31.8
chr1	21130	21379	65.4
chr1	47348	47398	-25.3
chr1	58653	58814	-4.6
chr1	70547	70781	-8.8
chr1	75843	75955	-11.9
chr1	115996	116053	93.6

+

Reference file #1:



Annotated file #1:

Chromosome	Start	End	...	Gene ID	Distance from target	Gene feature
chr1	20108	20256	...	ENSBTAG00000046619	209	
chr1	20257	20312	...	ENSBTAG00000046619	358	
chr1	20313	20590	...	ENSBTAG00000046619	414	
chr1	20618	20819	...	ENSBTAG00000046619	719	
chr1	21130	21379	...	ENSBTAG00000046619	1231	
chr1	47348	47398	...	ENSBTAG00000006858	11790	
chr1	58653	58814	...	ENSBTAG00000039257	-10881	
chr1	70547	70781	...	ENSBTAG00000039257	0	promoter
chr1	70547	70781	...	ENSBTAG00000039257	0	exon
chr1	75843	75955	...	ENSBTAG00000039257	4722	
chr1	115996	116053	...	ENSBTAG00000001753	-8796	

Then, we may want to annotate the resulting file with a second reference file, by instance, a table of gene information (gene name, description, …) extracted from Biomart Ensembl web site. Here the join between the two files will be performed by the match of the 'Gene ID' value in target file with 'Ensembl Gene Id' in reference file :

Target file #2 = Annotated file #1:

Chromosome	Start	End	...	Gene ID	...
chr1	20108	20256	...	ENSBTAG00000046619	...
chr1	20257	20312	...	ENSBTAG00000046619	...
chr1	20313	20590	...	ENSBTAG00000046619	...
chr1	20618	20819	...	ENSBTAG00000046619	...
chr1	21130	21379	...	ENSBTAG00000046619	...
chr1	47348	47398	...	ENSBTAG00000006858	...
chr1	58653	58814	...	ENSBTAG00000039257	...
chr1	70547	70781	...	ENSBTAG00000039257	...
chr1	70547	70781	...	ENSBTAG00000039257	...
chr1	75843	75955	...	ENSBTAG00000039257	...
chr1	115996	116053	...	ENSBTAG00000001753	...

+

Reference file #2:

Ensembl Gene ID	Gene name	Chr	...
ENSBTAG00000046619	5S_rRNA	1	...
ENSBTAG00000006858		1	...
ENSBTAG00000039257		1	...
ENSBTAG00000035349		1	...
ENSBTAG00000001753		1	...
ENSBTAG00000005540	CBX3	1	...

Annotated file #2:

Chromosome	Start	End	...	Gene ID	..	Gene name	...
chr1	20108	20256	...	ENSBTAG00000046619	..	5S_rRNA	...
chr1	20257	20312	...	ENSBTAG00000046619	..	5S_rRNA	...
chr1	20313	20590	...	ENSBTAG00000046619	..	5S_rRNA	...
chr1	20618	20819	...	ENSBTAG00000046619	..	5S_rRNA	...
chr1	21130	21379	...	ENSBTAG00000046619	..	5S_rRNA	...
chr1	47348	47398	...	ENSBTAG00000006858	..		...
chr1	58653	58814	...	ENSBTAG00000039257	..		...
chr1	70547	70781	...	ENSBTAG00000039257	..		...
chr1	70547	70781	...	ENSBTAG00000039257	..		...
chr1	75843	75955	...	ENSBTAG00000039257	..		...
chr1	115996	116053	...	ENSBTAG00000001753	..		...

An annotation process with annotk is specified using a configuration file.
Here is the configuration file corresponding to the two steps annotation process described above:

```
file_to_annotate                    in/DMRs_between_C1_C2.txt
file_format                         tab
output_file                         out/annotated_DMRs.txt
keep_scaffolds                      Yes

theme                               Gene features
          join_type                 gtf
          reference_file            reference/Bos_taurus.UMD3.1.81.gtf
          nb_max_results            All
          max_dist_nearest_gene     10kb

theme                               Gene
          join_type                 value
          target_keys               Gene ID
          reference_file            reference/biomart_Bos_taurus.txt
          reference_keys            Ensembl Gene ID
```

Global parameters

Annotation step #1 parameters

Annotation step #2 parameters

(next slides will detail the content of global and step parameters)

Once configuration file is set up, annotk can be launched using command :

```
RRBS_HOME/Annotation/annotk/annotate.sh <pathname to the configuration file>
```

(N.B : annotk needs python v2.7 and python module bx-python v0.7.3 or a more recent version)

Global parameters :

```
file_to_annotate                    in/DMRs_between_C1_C2.txt
file_format                         tab
output_file                         out/annotated_DMRs.txt
keep_scaffolds                      Yes
```

file_to_annotate pathname of the file you want to annotate (target file of annotations step #1)

file_format either ‘tab’ or ‘fasta’
‘tab’ format corresponds to a text file in which columns are separated by a <TAB> character.
For ‘fasta’, please see below § Annotation of a fasta file.

output_file pathname of the file at the end of the process (annotated file of last annotation step)

keep_scaffolds either ‘Yes’ or ‘No’. Indicates if input lines should be filtered out or not for non conventional chromosome.
A conventional chromosome is specified either by an integer, a letter or ‘Mt’ (prefixed or not by ‘chr’)

Step parameters :

```
...
theme                                   Gene features
          join_type                     gtf
          reference_file                reference/Bos_taurus.UMD3.1.81.gtf
          nb_max_results                All
          max_dist_nearest_gene         10kb
...
```

Each annotation step is defined in a block beginning by the keyword **theme**, followed by a name of the annotation step.

The first parameter to specify for a step annotation is the **join_type**. It indicates how the join between target and reference file should be performed. Three join methods are supported:

1. location — the join is performed by overlaps of genomic locations specified in target and reference file (reference file is a text file with columns separated by a <TAB> character)
2. value — the join is performed by the equality of values contained in one or several columns (reference file is a text file with columns separated by a <TAB> character. value used for the join must be unique in reference file)
3. gtf — the join is performed by overlaps of genomic locations specified in target and reference file In contrast with 'location' join, annotk expects to have gene feature region and gene identifier specified at a fixed position and format, in conformity with gff format (see by instance, Ensembl gtf files ftp://ftp.ensembl.org/pub/release-86/gtf/bos_taurus/Bos_taurus.UMD3.1.86.gtf.gz) (NB: 'gtf' format is different from 'gff3' format which is not supported at the moment)

Next slides describe parameters which can be used in a 'theme' section.

Annotation step parameters	join_type supported	Default value	Comment
theme	location, gtf, value	No (mandatory)	Specifies the beginning of a new annotation step parameter list. Should be followed by the name of the annotation step (e.g. 'Gene features' – see example above)
join_type	location, gtf, value	No (mandatory)	see above
reference_file	location, gtf, value	No (mandatory)	Pathname of the file used as reference
target_keys	location, gtf, value	1,2,3	Name or number of the columns of target file used to establish the join with reference file (separated by a comma)
reference_keys	location, value	1,2,3	Name or number of the columns of reference file used to establish the join with target file (separated by a comma).
min_overlap	location	0%	Minimal value accepted for either overlap_length/target_region_length or overlap_length/reference_region_length for a valid overlap result
interval_shift	location	0	see § Additional comments on annotation step parameters
nb_max_results	location, gtf	all	If several annotation results are possible for a given target location, annotk will report up to **nb_max_results** lines (default value keep all results)
max_dist_nearest_gene	gtf	Infinite	If a target location fails to overlap with a gene feature region, then annotk will report the identifier and distance to nearest gene (located up or downstream), only if this gene is located at less than **max_dist_nearest_gene** base pairs of the target location. If this parameter is set to '0', nothing is reported if overlap with target location fails. (example of possible values : '1000', '1000bp', '1e3', '1kb', '1.5kb')
feature_priorities	gtf	tss,promoter,tts,utr3, utr5,exon,intron	Priority order for reporting overlaps with target location depending on the nature of gene features. If some feature names is missing in this parameter value (e.g. 'intron'), overlap with this feature type will not be reported. ('tss'=transcription start site/'tts'=transcription termination site)
X_upstream, X_downstream (where **X** is 'promoter', 'tss' or 'tts')	gtf	-2000/+100 for 'promoter' +/- 100 for 'tss' +/- 100 for 'tts'	Distance considered for overlaps with 'promoter', 'tss' or 'tts' region. (example of possible values : '1000', '1000bp', '1e3', '1kb', '1.5kb')

Additional comments on annotation step parameters :

- target and reference keys for ‘location’ **join_type** and target file in ‘tab’ format (i.e. not fasta)

If only one column is specified for parameter **target_keys** or **reference_keys**, annotk assume, that this column will contain a genomic region specified as : `chromosome:<start coordinate>-<end coordinate>` (e.g. : ‘10:1234-5678’).

If two columns are specified, annotk assume these columns will contain a genomic position (first column : chromosome ; second column : genomic position – e.g. : ‘10 1234’ represents position ‘1234’ on chromosome ‘10’

If three columns are specified, the first column will contain the chromosome, the second column will contain the start coordinate of the region and the third, the end coordinate.

- Ordering of join results reported (refers to “# overlap” in output file)

If several overlaps between a target location and reference regions can be found, overlaps will be reported according to following rules :

- **join_type** is set to ‘gtf’ : priority is specified in **feature_priorities** parameter value. With the default value (‘tss, promoter, …’), if several overlaps are found for a given target location, annotk will first report overlaps with ‘tss’ feature, then ‘promoter’, and so on (up to **nb_max_results** results will be reported).
- **join_type** is set to ‘location’ : if target location is a genomic region, then annotk will order overlaps by the % of target genomic region included in the overlap (descending order). If target location is a genomic position, then annotk will order overlaps by the absolute distance between the middle of reference region and the target position (ascending order)
- **join_type** is set to ‘value’ : the reference key value must be unique in reference file, therefore, in this case, annotk cannot report several results.

- interval_shift(s) section (only used when **join_type** is set to ‘location’)

In some cases, it may be useful to allow some tolerance(*) in finding overlaps, or try to find overlaps with reference region extended if no overlap are found at first try.

By instance, annotation with CpG island (CGI) regions are often classified as :

- ‘island’ if the target location is located within a CGI region
- ‘shore’ if the target location is located within a CGI region +/- 2kb
- ‘shelves’ if the target location is located within a CGI region +/- 4kb

This can be specified to annotk using a list of interval shifts. Here is an example of a ‘theme’ section with a list of interval shifts :

```
theme                   CpG islands
        join_type       location
        target_keys     Chromosome,Position
        reference_file  reference/list_of_CGI.txt
        reference_keys  1,2,3
        interval_shift  0       island
        interval_shift  2000    shore
        interval_shift  4000    shelves
```

interval_shift value can be followed by a label. This label will be reported in output in column ‘Interval shift’.
If label is missing, the value of the interval shift will be reported.

(*) : by instance if both target and reference files refer to a genomic position, exact matching may be too stringent

- Annotation of fasta file

If you want to annotate fasta file, set **file_format** parameter to ‘fasta’ and verify header of fasta sequences satisfy following rules :

Fasta header begins with a specification of a genomic region. Header should have following format :
`>chr_start_end … anything else …`
where ‘chr’ is a chromosome name, ‘start’ and end are positive integer values.

Fasta header begins with a specification of a genomic position. Header should have following format :
`>chr_position … anything else …`
where ‘chr’ is a chromosome name, ‘position’ is a positive integer values.

A complete example of annotation config file:

```
file_to_annotate        in/DMRs_between_C1_C2.txt
file_format             tab
output_file             out/DMRs_between_C1_C2_annotated.txt
keep_scaffolds          No

theme                           Gene features
        join_type               gtf
        target_keys             Chromosome,Position
        reference_file          reference/Bos_taurus.UMD3.1.81.gtf
        nb_max_results          3
        max_dist_nearest_gene   10kb

theme                   Gene
        join_type       value
        target_keys     Gene ID
        reference_file  reference/bovine_biomart.txt
        reference_keys  Ensembl Gene ID

theme                   Repeats
        join_type       location
        target_keys     Chromosome,Position
        reference_file  reference/bovine_repeats.txt
        reference_keys  1,2,3
        min_overlap     75%


theme                   CpG islands
        join_type       location
        target_keys     Chromosome,Position
        reference_file  reference/bovine_CGI.txt
        reference_keys  1,2,3
        min_overlap     0%
        nb_max_results  all
        interval_shift  0       island
        interval_shift  2000    shore
        interval_shift  4000    shelves
```

Global parameters

GTF annotation

Gene information annotation

Annotation of overlaps with repeat regions

Annotation of overlaps with CpG islands

This example is available in '`annotk_config_example.txt`' file and can be run using '`run_annotk_config_example.sh`' script.

Exemple of output file corresponding to annotation process configured in previous slide :

Chromosome	Position	#overlap	Gene ID	Distance from target	Gene feature	Gene Name	Description	Chromosome Name	Gene Start (bp)	Gene End (bp)	Strand	#overlap	Repeats region	Type	#overlap	CpG islands region	Interval shift	Type
1	180179	1	ENSBTAG00000001753	466			Chloride intr	1	124849	179713	-1				1	1:178297-180180	island	CpG:
1	733760	1	ENSBTAG00000012594	0	tss	MRPS6	Bos taurus m	1	669920	733729	-1				1	1:733575-734527	island	CpG:
1	733760	2	ENSBTAG00000012594	0	promoter	MRPS6	Bos taurus m	1	669920	733729	-1				1	1:733575-734527	island	CpG:
1	1291176	1	ENSBTAG00000009188	0	intron	GART	phosphoribo	1	1265743	1292028	1							
1	1712072																	
1	2029698	1	ENSBTAG00000043993	0	promoter	C21orf62	chromosome	1	2029406	2048813	1							
1	2029698	2	ENSBTAG00000043993	0	utr5	C21orf62	chromosome	1	2029406	2048813	1							
1	2108466	1	ENSBTAG00000003063	-752		SYNJ1	synaptojanir	1	2109219	2196114	1				1	1:2107600-2108663	island	CpG:
1	2205003	1	ENSBTAG00000003059	0	promoter		Uncharacter	1	2204895	2216515	1				1	1:2204145-2205475	island	CpG:
1	2205003	2	ENSBTAG00000003059	0	utr5		Uncharacter	1	2204895	2216515	1				1	1:2204145-2205475	island	CpG:
1	2206311	1	ENSBTAG00000003059	0	promoter		Uncharacter	1	2204895	2216515	1	1	1:2206270-2206362	SINE	1	1:2204145-2205475	shore	CpG:
1	2206311	2	ENSBTAG00000003059	0	intron		Uncharacter	1	2204895	2216515	1	1	1:2206270-2206362	SINE	1	1:2204145-2205475	shore	CpG:
1	3049172											1	1:3048933-3049185	LINE	1	1:3049220-3050865	shore	CpG:
1	5035995	1	ENSBTAG00000039820	0	promoter	CLDN8	claudin 8 [So	1	5035870	5037857	1							
1	5035995	2	ENSBTAG00000039820	0	exon	CLDN8	claudin 8 [So	1	5035870	5037857	1							
1	5239521											1	1:5238611-5241421	Satellite				
1	5244742											1	1:5244169-5246987	Satellite				
1	5304925											1	1:5304653-5307131	Satellite				

genomic position present in input file

Annotation recovered from gtf file

Annotation recovered from gene annotation file

Annotation recovered from repeat annotation file

Annotation recovered from CGI annotation file

‘# overlap’ columns indicate the priority order of annotation result if several overlaps can be found for a target location

- Performance

annotk is really quick ! Annotation of 2.150.000 CpG positions requires :

  - 5’17’’ for a join_type = ‘gtf’ (with 24.600 genes described in the gtf reference file)
  - 0’7’’ for a join_type = ‘value’ (with 24.600 genes described in the gtf reference file)
  - 1’36’’ for a join_type = ‘location’ (with 5.550.000 repeat regions described in the reference file)
  - 0’46’’ for a join_type = ‘location’ (with 37.600 GpG island regions described in the reference file)

The whole process of annotation (4 steps) requires 7’48’’.

- Testing annotk

A set of several use cases exploring all annotk functionalities can be tested using ‘test_annotk.sh’ script.