Breakdown of console commands

I. Contents

II. Introduction

For the majority of use-cases, SoFIA's command line consists of three main parts, 1) the target file containing the entities to be annotated, 2) the requested entities and 3) the files containing the provided resources. For detailed information about the command line refer to the documentation on the homepage (https://www.github.com/childsish/sofia). The main article provides three use-cases of how SoFIA could be used. Each following section details the command lines used to produce the examples.

III. Variant annotation (natural variants)

The full command line:

```
> sofia.py aggregate\
      BRCA.vcf:chromosome_id=ucsc\
   -e chromosome\
      position\
      gene_id\
      amino_acid_variant\
      variant_effect\
      variant_impact:provean_score\
      pathway_id\
      variant.info[AF]:1000genomes:resource=1000genomes\
      variant.id:dbSNP:resource=dbSNP\
   -r refseq.gff:gene_id=hugo:transcript_id=ncbi\
      GRCh37.fasta:chromosome_id=ucsc\
      kegg.txt:entity=gene_pathway_map:gene_id=entrez\
      provean.txt:entity=provean_map:transcript_id=ensembl_protein\
      1000genomes.vcf:1000genomes:chromosome_id=ensembl\
      dbSNP.vcf:dbSNP:chromosome_id=ensembl\
   -m chromosome_id=chromosome_id.txt\
      gene_id=gene_id.txt\
      transcript_id=transcript_id.txt
```

The file name with the target entities is passed to SoFIA as the first positional argument. The requested entities are defined after the `-e` flag. The provided resources are defined after the `-r` flag. Each requested entity becomes a column in the output. If identifiers need to be converted, the identifier mappings are defined after the `-m` flag. Entities are pre-defined by the template and a full list of entities defined by the template can be obtained with the command '`sofia.py info entity`'. To aid repetition, requested entities and provided resources can be specified in a text file and passed to the framework using the flags `-E` (file of requested entities) and `-R` (file of provided resources).

Target entities

```
BRCA.vcf:chromosome_id=ucsc
```

This file contains a list of variants derived from the BRCA variant dataset from TCGA. We also specify an extra attribute using '`:`' to separate the fields. This attributes lets SoFIA know that the chromosome identifiers in the resource follow the UCSC style (ie. starting with the characters 'chr').

Requested entities

```
chromosome
```

The name of the chromosome.

```
position
```

The position of the variant on the chromosome (1-indexed).

```
gene_id
```

The name of the gene the variant is in.

```
amino_acid_variant
```

The amino acid change wrought by the variant.

```
variant_effect
```

The effect of the variant on the protein product. Terms conform to the standard defined by the sequence ontology.

```
variant_impact:provean_score
```

The impact of the variant on the protein function. In this case a score produced by the PROVEAN tools is reported. The field after the '`:`' specifies the header.

```
pathway_id
```

The pathway(s) in which the variant can be found.

```
variant.info[AF]:1000genomes:resource=1000genomes
```

The allele frequency of the variant in the 1000 genomes project. This request must specify that the 1000 genomes resource should be used to provide the allele frequency. This requested entity line can be further broken down into three parts separated by the delimiter '`:`'.

1. Requested entity and property access.

   `variant` the base entity being requested.

   `.info` the 'info' property of the entity.

   `[AF]` the 'AF' field of the 'info' property.

2. Header. This will appear in the header row of the output for the appropriate column.

   `1000genomes` the column name to appear in the header row.

3. Requested attributes. The entity must obtain these attributes during resolution.

   `resource=1000genomes` the variant must use the resource named 1000genomes.

```
variant.id:dbSNP:resource=dbSNP
```

The identifiers of the variants from dbSNP which match the target variant. This request specifies that the dbSNP resource should be used to provide the variant identifier. This request can also be broken down into three parts.

1. Requested entity and property access

   `variant` the base entity being requested.

   `.id` the 'id' property of the entity.

2. Header. This will appear in the header row of the output for the appropriate column.

   `dbSNP` the column name to appear in the header row

3. Requested attributes. The entity must obtain these attributes during resolution.

   `resource=dbSNP` the variant must use the resource named dbSNP.

Provided resources

```
gencode.gtf:gene_id=hugo
```

Gene models obtained from the Gencode project. We further specify that the gene identifiers provided are HUGO identifiers. The `gene_id`, `amino_acid_variant`, `variant_effect` entities derive directly from this resource.

```
GRCh37.fasta
```

The human chromosome sequence. The `amino_acid_variant` entity derives directly from this resource.

```
kegg.txt:entity=gene_pathway_map:gene_id=entrez
```

A mapping between KEGG pathways and genes. We need to specify that the file format is a gene/pathway mapping as this can not be inferred from the extension. We further specify that the gene identifiers are Entrez identifiers. The `pathway_id` entity derives directly from this resource. This provided entity line can be further broken down into two parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `kegg.txt` gene to KEGG pathway mapping.

2. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `entity=gene_pathway_map` provided entity is a mapping from gene names to pathways.

   `gene_id=entrez` gene identifiers are Entrez identifiers.

```
provean.txt:entity=provean_map:transcript_id=ensembl_protein
```

A mapping of all possible amino altering variants to their PROVEAN score. We need to specify that the file format is a PROVEAN mapping as this can not be inferred from the extension. We further specify that the transcript identifiers are Ensembl protein identifiers. The `variant_impact` entity derives directly from this resource. This provided entity line can be further broken down into two parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `provean.txt` amino acid change to PROVEAN pathway mapping.

2. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `entity=provean_map` mapping of amino acid changes to PROVEAN scores.

   `gene_id=ensembl_protein` gene identifiers are Ensembl protein identifiers.

```
1000genomes.vcf:1000genomes:chromosome_id=ensembl
```

Data from the 1000 genomes project. A `variant` entity requires this resource and specifically requests it in the attributes. This provided entity line can be further broken down into three parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `1000genomes.vcf` a collection of variants from the 1000 genomes project.

2. A name for the resource that can be referenced by the entities.

   `1000genomes` entities can refer to the resource with this name.

3. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `chromosome_id=ensembl` chromosome identifiers follow the Ensembl style

```
dbSNP.vcf:dbSNP:chromosome_id=ensembl
```

Data from dbSNP. A `variant` entity requires this resource and specifically requests it in the attributes. This provided resource line can be further broken down into three parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `dbSNP.vcf` a collection of variants from dbSNP.

2. A name for the resource that can be referenced by the entities.

   `dbSNP` entities can refer to the resource with this name.

3. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `chromosome_id=ensembl` chromosome identifiers follow the Ensembl style

IV. Variant annotation (cancer mutations)

The full command line:

```
> sofia.py aggregate\
      BRCA.vcf:chromosome_id=ucsc\
   -e chromosome\
      position\
      gene_id\
      amino_acid_variant\
      variant_effect\
      variant_impact:provean_score\
      pathway_id\
      variant.id:cosmic:resource=cosmic\
      number_of_variants:tcga:resource=tcga_coad\
   -r refseq.gff:gene_id=hugo:transcript_id=ncbi\
      GRCh37.fasta:chromosome_id=ucsc\
      kegg.txt:entity=gene_pathway_map:gene_id=entrez\
      provean.txt:entity=provean_map:transcript_id=ensembl_protein\
      cosmic.vcf:cosmic:chromosome_id=ensembl\
      tcga_coad.maf:tcga_coad:chromosome_id=ensembl\
   -m chromosome_id=chromosome_id.txt\
      gene_id=gene_id.txt\
      transcript_id=transcript_id.txt
```

The file name with the target entities is passed to SoFIA as the first positional argument. The requested entities are defined after the `-e` flag. The provided resources are defined after the `-r` flag. Each requested entity becomes a column in the output. If identifiers need to be converted, the identifier mappings are defined after the `-m` flag. Entities are pre-defined by the template and a full list of entities defined by the template can be obtained with the command '`sofia.py info entity`'. To aid repetition, requested entities and provided resources can be specified in a text file and passed to the framework using the flags `-E` (file of requested entities) and `-R` (file of provided resources).

Target entities

```
BRCA.vcf:chromosome_id=ucsc
```

This file contains a list of variants derived from the BRCA variant dataset from TCGA. We also specify an extra attribute using '`:`' separated fields. This attributes lets SoFIA know that the chromosome identifiers in the resource follow the UCSC style (ie. starting with the characters 'chr').

Requested entities

```
chromosome
```

The name of the chromosome.

```
position
```

The position of the variant on the chromosome (1-indexed).

```
gene_id
```

The name of the gene the variant is in.

```
amino_acid_variant
```

The amino acid change wrought by the variant.

```
variant_effect
```

The effect of the variant on the protein product. Terms conform to the standard defined by the sequence ontology.

```
variant_impact:provean_score
```

The impact of the variant on the protein function. In this case a score produced by the PROVEAN tools is reported. The field after the ':' specifies the header.

```
pathway_id
```

The pathway(s) in which the variant can be found.

```
variant.id:cosmic:resource=cosmic
```

The variant id from a COSMIC variant that matches the target variant. This request can be broken down into three parts separated by the delimiter ':'.

1. Requested entity and property access

   `variant` the base entity being requested.

   `.id` the 'id' property of the entity.

2. Header. This will appear in the header row of the output for the appropriate column.

   `cosmic` the column name to appear in the header row

3. Requested attributes. The entity must obtain these attributes during resolution.

   `resource=cosmic` the variant must use the resource named 'cosmic'.

```
number_of_variants:tcga:resource=tcga_coad
```

The number of TCGA adenocarcinoma variants that match the target variant. This request can be broken down into three parts separated by the delimiter ':'.

1. Requested entity and property access

   `number_of_variants` request a number of variants entity.

2. Header. This will appear in the header row of the output for the appropriate column.

   `tcga` the column name to appear in the header row

3. Requested attributes. The entity must obtain these attributes during resolution.

   `resource=tcga_coad` the variant must use the resource named 'tcga_coad'.

Provided resources

```
gencode.gtf:gene_id=hugo
```

Gene models obtained from the Gencode project. We further specify that the gene identifiers provided are HUGO identifiers. The `gene_id`, `amino_acid_variant`, `variant_effect` entities derive directly from this resource.

```
GRCh37.fasta
```

The human chromosome sequence. The `amino_acid_variant` entity derives directly from this resource.

```
kegg.txt:entity=gene_pathway_map:gene_id=entrez
```

A mapping between KEGG pathways and genes. We need to specify that the file format is a gene/pathway mapping as this can not be inferred from the extension. We further specify that the gene identifiers are Entrez identifiers. The `pathway_id` entity derives directly from this resource. This provided entity line can be further broken down into two parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `kegg.txt` gene to KEGG pathway mapping.

2. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `entity=gene_pathway_map` provided entity is a mapping from gene names to pathways.

   `gene_id=entrez` gene identifiers are Entrez identifiers.

```
provean.txt:entity=provean_map:transcript_id=ensembl_protein
```

A mapping of all possible amino altering variants to their PROVEAN score. We need to specify that the file format is a PROVEAN mapping as this can not be inferred from the extension. We further specify that the transcript identifiers are Ensembl protein identifiers. The `variant_impact` entity derives directly from this resource. This provided entity line can be further broken down into two parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `provean.txt` amino acid change to PROVEAN pathway mapping.

2. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `entity=provean_map` mapping of amino acid changes to PROVEAN scores.

   `gene_id=ensembl_protein` gene identifiers are Ensembl protein identifiers.

```
cosmic.vcf:cosmic:chromosome_id=ensembl
```

Data from the COSMIC database. A `variant` entity requires this provided entity and directly requests it in the attributes. This provided entity can be broken down into three parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `cosmic.vcf` a collection of variants from the COSMIC database.

2. A name for the resource that can be referenced by the entities.

   `cosmic` entities can refer to the resource with this name.

3. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `chromosome_id=ensembl` chromosome identifiers follow the Ensembl style

```
tcga_coad.maf:tcga_coad:chromosome_id=ensembl
```

Data from the TCGA colon adenocarcinoma database. A `variant` entity requires this provided entity and directly requests it in the attributes. This provided entity can be broken down into three parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `tcga_coad.vcf` a collection of variants from the TCGA COAD database.

2. A name for the resource that can be referenced by the entities.

   `tcga_coad` entities can refer to the resource with this name.

3. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `chromosome_id=ensembl` chromosome identifiers follow the Ensembl style

V. Variant annotation (drug databases)

The full command line:

```
> sofia.py aggregate\
      BRCA.vcf:chromosome_id=ucsc\
   -e chromosome\
      position\
      gene_id\
      amino_acid_variant\
      variant_effect\
      variant_impact:provean_score\
      pathway_id\
      drug_id:drug_bank:resource=drug_bank\
      drug_id:gdsc:resource=gdsc\
   -r refseq.gff:gene_id=hugo:transcript_id=ncbi\
      GRCh37.fasta:chromosome_id=ucsc\
      kegg.txt:entity=gene_pathway_map:gene_id=entrez\
      provean.txt:entity=provean_map:transcript_id=ensembl_protein\
      db.txt:drug_bank:entity=drug_bank_map:transcript_id=uniprot\
      gdsc.txt:gdsc:entity=gdsc_map:gene_id=hugo\
   -m chromosome_id=chromosome_id.txt\
      gene_id=gene_id.txt\
      transcript_id=transcript_id.txt
```

The file name with the target entities is passed to SoFIA as the first positional argument. The requested entities are defined after the `-e` flag. The provided resources are defined after the `-r` flag. Each requested entity becomes a column in the output. If identifiers need to be converted, the identifier mappings are defined after the `-m` flag. Entities are pre-defined by the template and a full list of entities defined by the template can be obtained with the command '`sofia.py info entity`'. To aid repetition, requested entities and provided resources can be specified in a text file and passed to the framework using the flags `-E` (file of requested entities) and `-R` (file of provided resources).

Target entities

```
BRCA.vcf:chromosome_id=ucsc
```

This file contains a list of variants derived from the BRCA variant dataset from TCGA. We also specify an extra attribute using '`:`' separated fields. This attributes lets SoFIA know that the chromosome identifiers in the resource follow the UCSC style (ie. starting with the characters 'chr').

Requested entities

```
chromosome
```

The name of the chromosome.

```
position
```

The position of the variant on the chromosome (1-indexed).

```
gene_id
```

The name of the gene the variant is in.

```
amino_acid_variant
```

The amino acid change wrought by the variant.

`variant_effect`

The effect of the variant on the protein product. Terms conform to the standard defined by the sequence ontology.

`variant_impact:provean_score`

The impact of the variant on the protein function. In this case a score produced by the PROVEAN tools is reported. The field after the '`:`' specifies the header.

`pathway_id`

The pathway(s) in which the variant can be found.

`drug_id:drug_bank:resource=drug_bank`

A drug identifier from the DrugBank database. This request can be broken down into three parts separated by the delimiter '`:`'.

1. Requested entity and property access

   `drug_id` the base entity being requested.

2. Header. This will appear in the header row of the output for the appropriate column.

   `drug_bank` the column name to appear in the header row

3. Requested attributes. The entity must obtain these attributes during resolution.

   `resource=drug_bank` the variant must use the resource named 'drug_bank'.

`drug_id:gdsc:resource=gdsc`

A drug identifier from the Genomics of Drug Sensitivity in Cancer database (GDSC). This request can be broken down into three parts separated by the delimiter '`:`'.

1. Requested entity and property access

   `drug_id` the base entity being requested.

2. Header. This will appear in the header row of the output for the appropriate column.

   `gdsc` the column name to appear in the header row

3. Requested attributes. The entity must obtain these attributes during resolution.

   `resource=gdsc` the variant must use the resource named 'gdsc'.

Provided resources

`gencode.gtf:gene_id=hugo`

Gene models obtained from the Gencode project. We further specify that the gene identifiers provided are HUGO identifiers. The `gene_id`, `amino_acid_variant`, `variant_effect` entities derive directly from this resource.

`GRCh37.fasta`

The human chromosome sequence. The `amino_acid_variant` entity derives directly from this resource.

```
kegg.txt:entity=gene_pathway_map:gene_id=entrez
```

A mapping between KEGG pathways and genes. We need to specify that the file format is a gene/pathway mapping as this can not be inferred from the extension. We further specify that the gene identifiers are Entrez identifiers. The `pathway_id` entity derives directly from this resource. This provided entity line can be further broken down into two parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `kegg.txt` gene to KEGG pathway mapping.

2. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `entity=gene_pathway_map` provided entity is a mapping from gene names to pathways.

   `gene_id=entrez` gene identifiers are Entrez identifiers.

```
provean.txt:entity=provean_map:transcript_id=ensembl_protein
```

A mapping of all possible amino altering variants to their PROVEAN score. We need to specify that the file format is a PROVEAN mapping as this can not be inferred from the extension. We further specify that the transcript identifiers are Ensembl protein identifiers. The `variant_impact` entity derives directly from this resource. This provided entity line can be further broken down into two parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `provean.txt` amino acid change to PROVEAN pathway mapping.

2. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `entity=provean_map` mapping of amino acid changes to PROVEAN scores.

   `gene_id=ensembl_protein` gene identifiers are Ensembl protein identifiers.

```
db.txt:drug_bank:entity=drug_bank_map:transcript_id=uniprot
```

A mapping of transcript identifiers to DrugBank identifiers. A '`drug_id`' entity requires this provided entity and directly requests it in the attributes. We need to specify that the file is a DrugBank map as this can not be inferred from the extension. We further specify that the transcript identifiers are Uniprot identifiers. This provided entity can be broken down into three parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `db.txt` transcript identifier to drug identifier mapping from DrugBank.

2. A name for the resource that can be referenced by the entities.

   `drug_bank` entities can refer to the resource with this name.

3. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `transcript_id=uniprot` transcript identifiers are Uniprot identifiers

```
gdsc.txt:gdsc:entity=gdsc_map:gene_id=hugo
```

A mapping of transcript identifiers to Genomic of Drug Sensitivity in Cancer (GDSC) identifiers. A '`drug_id`' entity requires this provided entity and directly requests it in the attributes. We need to specify that the file is a GDSC map as this can not be inferred from the extension. We further specify that the transcript identifiers are HUGO identifiers. This provided entity can be broken down into three parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `gdsc.txt` gene identifier to drug identifier mapping from GDSC.

2. A name for the resource that can be referenced by the entities.

   `gdsc` entities can refer to the resource with this name.

3. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `gene_id=hugo` gene identifiers are HUGO identifiers.

VI. Gene model comparison

The full command line:

```
> sofia.py aggregate\
       BRCA.vcf:chromosome_id=ucsc\
    -e chromosome\
       position\
       gene_id:gene_id=hugo:resource=refseq\
       amino_acid_variant:resource=refseq\
       variant_effect:resource=refseq\
       gene_id:gene_id=hugo:resource=gencode\
       amino_acid_variant:resource=gencode\
       variant_effect:resource=gencode\
    -r GRCh37.fasta:chromosome_id=ucsc\
       gencode.gtf:gencode:chromosome_id=ucsc\
       refseq.gff:refseq:chromosome_id=ncbi\
    -m chromosome_id=chromosome_id.txt
```

Target entities

```
BRCA.vcf:chromosome_id=ucsc
```

This file is described in the previous section.

Requested entities

```
chromosome
position
```

These entities do not differ from the previous section.

```
gene_id:gene_id=hugo:resource=refseq
```

The name of the gene the variant is in. The attributes specify that the gene identifier must also be a HUGO identifier and that the entity is derived from the '`refseq`' resource.

```
amino_acid_variant:resource=refseq
```

The amino acid change wrought by the variant. The attributes specify that the entity is derived from the '`refseq`' resource.

```
variant_effect:resource=refseq
```

The effect of the variant on the protein product. Terms conform to the standard defined by the sequence ontology. The attributes specify that the entity is derived from the '`refseq`' resource.

```
gene_id:gene_id=hugo:resource=gencode
```

The name of the gene the variant is in. The attributes specify that the gene identifier must also be a HUGO identifier and that the entity is derived from the '`gencode`' resource.

```
amino_acid_variant:resource=gencode
```

The amino acid change wrought by the variant. The attributes specify that the entity is derived from the '`gencode`' resource.

```
variant_effect:resource=gencode
```

The effect of the variant on the protein product. Terms conform to the standard defined by the sequence ontology. The attributes specify that the entity is derived from the '`gencode`' resource.

Provided resources

```
GRCh37.fasta:chromosome_id=ucsc
```

The human chromosome sequence. The `amino_acid_variant` entity derives directly from this resource. The attributes specify that the chromosome identifiers follow the UCSC style.

```
gencode.gtf:gencode:chromosome_id=ucsc
```

Gene models obtained from the Gencode project. The attributes specify that the chromosome identifiers follow the UCSC style. `gene_id`, `amino_acid_variant`, `variant_effect` entities derive directly from this resource.

```
refseq.gff:refseq:chromosome_id=ncbi
```

Gene models obtained from the Refseq. The attributes specify that the chromosome identifiers are NCBI identifiers. `gene_id`, `amino_acid_variant`, `variant_effect` entities derive directly from this resource.

VII. Sequence feature calculation

The full command line:

```
> sofia.py aggregate\
       ecoli.gbk\
    -e gene_id\
       codon_adaptation_index\
       effective_number_of_codons\
       translation_start_mfe\
       number_of_pest_sequences\
       number_of_upstream_orfs\
    -r ecoli.gbk
```

Target entities

```
ecoli.gbk
```

A GenBank file downloaded directly from the NCBI database. As a target, the genes defined in the file are annotated. The accession number at time of download was U00096.3.

Requested entities

```
gene_id
```

The name of a gene in the GenBank file.

```
codon_adaptation_index
```

The codon adaptation index.

```
effective_number_of_codons
```

The number of effective codons.

```
translation_start_mfe
```

The translation start minimum free energy.

```
number_of_pest_sequences
```

The number of PEST sequences.

```
number_of_upstream_orfs
```

The number of ORFs upstream of the start codon.

Provided resources

```
ecoli.gbk
```

The GenBank file again. As a resource, the genomic sequence found in the GenBank file is used to generate the coding sequences needed to calculate the requested entities.