An R interface to the Ensembl REST API

Tim Yates

October 12, 2012

Contents

1 Introduction

This package uses the Ensembl REST API[1] (currently in beta) to extract data from Ensembl into R.

As the REST API is in Beta, this package should also be considered to be in flux and functions/-parameters/etc are subject to change as things get finalized.

It could also do with your help. If you find a problem, something you think could be better, or a better way of doing things, please consider visiting the GitHub project at `https://github.com/acbb/EnsemblRest` and posting an issue or a Pull Request. Thanks!

2 Available Methods

To begin with (assuming you have installed this package), you need to load it into your R session:

```
> library( EnsemblRest )
```

The following subsections then list the methods available to you.

[1] `http://beta.rest.ensembl.org/`

2.1 Information

2.1.1 isAlive

Firstly, we can check to see that the REST API is accepting calls:

```
> isAlive()
[1] TRUE
```

2.1.2 infoSpecies

To get a list of available species on the server, you can use the infoSpecies call[2].

```
> infoSpecies()[1:3] # Just the first 3
[[1]]
name    : saccharomyces_cerevisiae
aliases : 4932, saccer, saccharomyces cerevisiae (baker's yeast), baker's yeast, scer, sacchar
groups  : core, otherfeatures, variation
release : 68

[[2]]
name    : ciona_savignyi
aliases : ciosav, 51511, ciona savignyi, csavignyi, c.savignyi, csav, sea squirt ciona savigny
groups  : core, otherfeatures
release : 68

[[3]]
name    : myotis_lucifugus
aliases : little brown bat, mlucifugus, myoluc, mluc, 59463, myotis lucifugus, myotis_lucifugu
groups  : core, otherfeatures
release : 68
```

2.1.3 infoAssembly

The infoAssembly call[3] returns information about the currently available assemblies in the given species.

```
> infoAssembly( 'human' )
assembly_name                 : GRCh37.p8
assembly_date                 : 2009-02
coord_system_versions         : , GRCh37, NCBI36, NCBI34, NCBI35
schema_build                  : 68_37
genebuild_start_date          : 2010-07-Ensembl
genebuild_initial_release_date : 2011-04
genebuild_last_geneset_update  : 2012-07
genebuild_method              : full_genebuild
top_level_seq_region_names    : 1, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 2, 20, 21, 22, 3,
```

[2] http://beta.rest.ensembl.org/documentation/info/species
[3] http://beta.rest.ensembl.org/documentation/info/assembly_info

2.1.4 assemblyDetails

The assemblyDetails call[4] returns information about one of these assemblies.

```
> assemblyDetails( 'X', 'human' )
is_chromosome           : TRUE
length                  : 155270560
assembly_exception_type : REF
coordinate_system       : chromosome
```

2.1.5 infoComparas

The infoComparas call[5] lists the available comparative genomics databases.

```
> infoComparas()
multi
 "68"
```

2.1.6 infoData

The infoData call[6] shows the data releases available to the REST service

```
> infoData()
[1] "68"
```

2.1.7 infoRest

infoRest shows the current version[7] of the REST service

```
> infoRest()
[1] "1.0.0"
```

2.1.8 infoSoftware

And finally in the info section, infoSoftware shows[8] the current version of the Ensembl API.

```
> infoSoftware()
[1] 68
```

[4] http://beta.rest.ensembl.org/documentation/info/assembly_stats
[5] http://beta.rest.ensembl.org/documentation/info/comparas
[6] http://beta.rest.ensembl.org/documentation/info/data
[7] http://beta.rest.ensembl.org/documentation/info/rest
[8] http://beta.rest.ensembl.org/documentation/info/software

2.2 Comparative Genomics

2.2.1 geneTree

This method[9] fetches the gene tree in New Hampshire format for a given Ensembl gene tree identifier.

```
> geneTree( 'ENSGT00390000003602' )
[1] "(((((((ENSGALP00000027524:0.0253,ENSMGAP00000015990:0.0275):0.0922,ENSTGUP00000012130:0.
```

it is also possible to specify the NH format you require (ie: for full format):

```
> geneTree( 'ENSGT00390000003602', nh_format='full' )
[1] "(((((((ENSGALP00000027524:0.0253,ENSMGAP00000015990:0.0275):0.0922,ENSTGUP00000012130:0.
```

2.2.2 homologyById

When given an Ensembl Gene ID, returns the homology[10] information for it.

```
> hResponse = homologyById( 'ENSG00000170037' )
> hResponse                  # The response object
id : ENSG00000170037
containing 42 homologies
> hResponse$homologies[1:2] # Just the top 2 homologies to save room
[[1]]
dn_ds   : 1.01739
type    : ortholog_one2one
subtype : Homininae
source  :
id         : ENSG00000170037
species    : homo_sapiens
protein_id : ENSP00000458251
perc_pos   : 94
perc_id    : 93
cigar_line : 682D155M21D28M17D2M11D2M6D
align_seq  : ----------------------------------------------------------------------------------

target  :
id         : ENSPTRG00000008719
species    : pan_troglodytes
protein_id : ENSPTRP00000014861
perc_pos   : 19
perc_id    : 19
cigar_line : 924M
align_seq  : MATSADSPSSPLGAEDLLSDSSEPPGLNQVSSEVTSQLYASLRLSRQAEATARAQLYLPSTSPPHEGLDGLAQELSRSLSV

[[2]]
dn_ds   :
type    : ortholog_one2one
subtype : Homininae
source  :
id         : ENSG00000170037
```

[9] http://beta.rest.ensembl.org/documentation/info/rest
[10] http://beta.rest.ensembl.org/documentation/info/homology_ensemblgene

```
species    : homo_sapiens
protein_id : ENSP00000458251
perc_pos   : 93
perc_id    : 91
cigar_line : 682D155M22D28M17D2M11D2M6D
align_seq  : ----------------------------------------------------------------------------------------

target  :
id         : ENSGGOG00000006451
species    : gorilla_gorilla
protein_id : ENSGGOP00000006314
perc_pos   : 19
perc_id    : 18
cigar_line : 925M
align_seq  : MATSADSPSSPLGAEDLLSDSSEPPGLNQVSSEVTSQLYASLRLSRQAEATARAQLYLPSTSPPHEGLDGLAQELSRSLSV
```

2.2.3 homologyBySymbol

You can also retrieve homology information[11] given a symbol and a species;

```
> hResponse = homologyBySymbol( 'BRCA2', 'human' )
> hResponse                 # The response object
id : ENSG00000139618
containing 52 homologies
> hResponse$homologies[1:2] # Again, just the top 2 homologies to save room
[[1]]
dn_ds   : 0.29371
type    : ortholog_one2one
subtype : Homininae
source  :
id         : ENSG00000139618
species    : homo_sapiens
protein_id : ENSP00000369497
perc_pos   : 99
perc_id    : 99
cigar_line : 3418M
align_seq  : MPIGSKERPTFFEIFKTRCNKADLGPISLNWFEELSSEAPPYNSEPAEESEHKNNNYEPNLFKTPQRKPSYNQLASTPIIF

target  :
id         : ENSPTRG00000005766
species    : pan_troglodytes
protein_id : ENSPTRP00000009812
perc_pos   : 99
perc_id    : 99
cigar_line : 3418M
align_seq  : MPIGSKERPTFFEIFKTRCNKADLGPISLNWFEELSSEAPPYNSEPAEESEHKNNNYEPNLFKTPQRKPSYNQLASTPIIF

[[2]]
dn_ds   :
type    : ortholog_one2one
subtype : Homininae
```

[11] http://beta.rest.ensembl.org/documentation/info/homology_symbol

```
source  :
id         : ENSG00000139618
species    : homo_sapiens
protein_id : ENSP00000369497
perc_pos   : 95
perc_id    : 94
cigar_line : 22MD3396M
align_seq  : MPIGSKERPTFFEIFKTRCNKA-DLGPISLNWFEELSSEAPPYNSEPAEESEHKNNNYEPNLFKTPQRKPSYNQLASTPII

target  :
id         : ENSGGOG00000015808
species    : gorilla_gorilla
protein_id : ENSGGOP00000015446
perc_pos   : 98
perc_id    : 97
cigar_line : 99M7D563M4D615MD561M7D604M19D891M48D
align_seq  : MPIGSKERPTFFEIFKTRCNKAVDLGPISLNWFEELSSEAPPYNSEPAEESEHKNNNYEPNLFKTPQRKPSYNQLASTPII
```

2.3 Cross References

Cross references are links to other data about the object of interest. It should be noted that these other data hold different fields and datatypes, so the results are returned in a `data.frame` containing a superset of column names, with non-applicable columns for a given result filled with <NA>.

As a single object may have multiple synonyms, this will cause the object to exist in multiple rows, one for each synonym.

2.3.1 xrefsById

Firstly we can get all external references[12] for a given Ensembl ID:

```
> xrefsById( 'ENSG00000170037' )
                                    display_id         primary_id version
1                            OTTHUMG00000172932 OTTHUMG00000172932       2
2                                     Hs.732863          Hs.732863       0
3                                     Hs.348012          Hs.348012       0
4  CENTROSOMAL BRCA2-INTERACTING PROT [*611425]             611425       0
5                                        CNTROB              29616       0
6                                        CNTROB              29616       0
7                                        CNTROB             116840       0
8                                        CNTROB             CNTROB       0
9                                        CNTROB             116840       0
            dbname      info_type           info_text     db_display_name
1             OTTG           NONE                            Havana gene
2          UniGene SEQUENCE_MATCH                                UniGene
3          UniGene SEQUENCE_MATCH                                UniGene
4         MIM_GENE      DEPENDENT                               MIM gene
5             HGNC         DIRECT Generated via ccds         HGNC Symbol
6             HGNC         DIRECT Generated via ccds         HGNC Symbol
7       EntrezGene      DEPENDENT                             EntrezGene
8 Uniprot_genename      DEPENDENT                   UniProtKB Gene Name
9         WikiGene      DEPENDENT                               WikiGene
  ensembl_identity ensembl_start xref_start xref_end ensembl_end score
1             <NA>          <NA>       <NA>     <NA>        <NA>  <NA>
2               99             1          6      752         752  3677
3               99             1         32     3794        3769 18783
4             <NA>          <NA>       <NA>     <NA>        <NA>  <NA>
5             <NA>          <NA>       <NA>     <NA>        <NA>  <NA>
6             <NA>          <NA>       <NA>     <NA>        <NA>  <NA>
7             <NA>          <NA>       <NA>     <NA>        <NA>  <NA>
8             <NA>          <NA>       <NA>     <NA>        <NA>  <NA>
9             <NA>          <NA>       <NA>     <NA>        <NA>  <NA>
       cigar_line
1            <NA>
2 289M6D406M1I51M
3     376M6D3387M
4            <NA>
5            <NA>
6            <NA>
7            <NA>
8            <NA>
9            <NA>
```

[12] http://beta.rest.ensembl.org/documentation/info/xref_id

```
                                                                                       description
1                                                                                             <NA>
2 Transcribed locus, moderately similar to NP_444279.2 centrobin isoform alpha [Homo sapiens]
3                                             Centrobin, centrosomal BRCA2 interacting protein
4                                                CENTROSOMAL BRCA2-INTERACTING PROTEIN; CNTROB
5                                             centrobin, centrosomal BRCA2 interacting protein
6                                             centrobin, centrosomal BRCA2 interacting protein
7                                             centrobin, centrosomal BRCA2 interacting protein
8
9                                             centrobin, centrosomal BRCA2 interacting protein
  xref_identity synonyms
1          <NA>     <NA>
2            99     <NA>
3            90     <NA>
4          <NA>     <NA>
5          <NA>     LIP8
6          <NA>   PP1221
7          <NA>     LIP8
8          <NA>     LIP8
9          <NA>     <NA>
```

2.3.2 xrefsByName

Or, we can look for an external reference primary accession[13] (given a species):

```
> xrefsByName( 'NM_004333', 'human' )

   display_id primary_id version
1 NM_004333.4  NM_004333       4
                                                                  description
1 Homo sapiens v-raf murine sarcoma viral oncogene homolog B1 (BRAF), mRNA.
       dbname info_type          info_text db_display_name
1 RefSeq_mRNA    DIRECT Generated via ccds     RefSeq mRNA
```

2.3.3 xrefsBySymbol

And we can finally look up all Ensembl objects referenced by an external symbol for a given species[14]:

```
> xrefsBySymbol( 'BRAF', 'human' )

        type              id
1       gene ENSG00000157764
2 transcript ENST00000288602
```

[13] http://beta.rest.ensembl.org/documentation/info/xref_name
[14] http://beta.rest.ensembl.org/documentation/info/xref_external

2.4 Lookup

2.4.1 lookupId

To find the database and species containing a known Ensembl id, you can use the lookup function[15] like so:

```
> lookupId( 'ENSG00000170037' )
                id      species object_type db_type
1 ENSG00000170037 homo_sapiens        Gene    core
```

[15] http://beta.rest.ensembl.org/documentation/info/lookup

2.5 Mapping

The mapping functions are used to convert co-ordinates between systems or databases.

There is currently an issue with `mappingCdna`, `mappingCds` and `mappingTranslation` in that the `seq_region_name` is not returned from the REST interface. This has been reported to Ensembl and should be fixed in the next release.

2.5.1 mapping

The `mapping` function[16] converts the co-ordinates in one assembly into another, ie:

```
> mapping( 'NCBI36', '1..10000:1', 'GRCh37', 'human' )
[[1]]
GRanges with 2 ranges and 3 elementMetadata cols:
      seqnames         ranges strand |    assembly coordinate_system
         <Rle>      <IRanges>  <Rle> | <character>       <character>
  [1]        1 [  617, 10000]      + |      NCBI36        chromosome
  [2]        1 [10754, 20137]      + |      GRCh37        chromosome
             type
      <character>
  [1]    original
  [2]      mapped
  ---
  seqlengths:
    1
   NA
```

As you can see, it returns one `GRanges` object per result, with an `original` row and a `mapped` row.

2.5.2 mappingCdna

This function[17] converts CDNA co-ordinates for a given Ensembl Transcript to genomic co-ordinates.

```
> mappingCdna( 'ENST00000288602', '100..300' )
GRanges with 2 ranges and 2 elementMetadata cols:
      seqnames                ranges strand |       gap      rank
         <Rle>             <IRanges>  <Rle> | <numeric> <numeric>
  [1]       NA [140624366, 140624465]      - |         0         0
  [2]       NA [140549912, 140550012]      - |         0         0
  ---
  seqlengths:
   NA
   NA
```

2.5.3 mappingCds

Or you can convert CDS co-ordinates[18] instead of CDNA ones:

```
> mappingCds( 'ENST00000288602', '100..300' )
```

[16] http://beta.rest.ensembl.org/documentation/info/assembly_map
[17] http://beta.rest.ensembl.org/documentation/info/assembly_cdna
[18] http://beta.rest.ensembl.org/documentation/info/assembly_cds

```
GRanges with 3 ranges and 2 elementMetadata cols:
      seqnames                ranges strand |       gap      rank
         <Rle>             <IRanges>  <Rle> | <numeric> <numeric>
  [1]       NA [140624366, 140624404]      - |         0         0
  [2]       NA [140549911, 140550012]      - |         0         0
  [3]       NA [140534613, 140534672]      - |         0         0
  ---
  seqlengths:
   NA
   NA
```

2.5.4 mappingTranslation

And finally, it is possible to convert from protein co-ordinates to genomic ones using the `mappingTranslation` method[19]:

```
> mappingTranslation( 'ENSP00000288602', '100..300' )
GRanges with 5 ranges and 2 elementMetadata cols:
      seqnames                ranges strand |       gap      rank
         <Rle>             <IRanges>  <Rle> | <numeric> <numeric>
  [1]       NA [140534409, 140534615]      - |         0         0
  [2]       NA [140508692, 140508795]      - |         0         0
  [3]       NA [140507760, 140507862]      - |         0         0
  [4]       NA [140501212, 140501360]      - |         0         0
  [5]       NA [140500242, 140500281]      - |         0         0
  ---
  seqlengths:
   NA
   NA
```

[19] http://beta.rest.ensembl.org/documentation/info/assembly_translation

2.6 Sequences

2.6.1 sequenceById

Fetch a sequence based on the stable id of an Ensembl feature[20] (I'm using `str` here to avoid overflowing the pdf too much):

```
> str( sequenceById( 'ENSG00000157764' ), give.head=F, strict.width='cut' )
List of 4
 $ desc    :"chromosome:GRCh37:7:140424943:140624564:-1"
 $ id      :"ENSG00000157764"
 $ seq     :"CGCCTCCCTTCCCCCTCCCCGCCCGACAGCGGCCGCTCGGGCCCCGGCTCTCGGTTATAAGATGG..
 $ molecule:"dna"
```

You can also get different types of sequence, here is an example for the spliced CDNA sequence of a transcript:

```
> str( sequenceById( 'ENST00000408384', type='cdna' ), give.head=F, strict.width='cut' )
List of 4
 $ desc    : NULL
 $ id      :"ENST00000408384"
 $ seq     :"GGATGCCCAGCTAGTTTGAATTTTAGATAAACAACGAATAATTTCGTAGCATAAATATGTCCCAA..
 $ molecule:"dna"
```

And again, for the protein coding

```
> str( sequenceById( 'ENSP00000334393', type='protein' ), give.head=F, strict.width='cut' )
List of 4
 $ desc    : NULL
 $ id      :"ENSP00000334393"
 $ seq     :"MVTEFIFLGLSDSQELQTFLFMLFFVFYGGIVFGNLLIVITVVSDSHLHSPMYFLLANLSLIDLS..
 $ molecule:"protein"
```

2.6.2 sequenceByRegion

You can also just query for a region of a given species[21]:

```
> str( sequenceByRegion( 'X:1_000_000..1_000_100:1', 'human' ), give.head=F, strict.width='cut
List of 4
 $ desc    : NULL
 $ id      :"chromosome:GRCh37:X:1000000:1000100:1"
 $ seq     :"GAAACAGCTACTTGGAAGGCTGAAGCAGGAGGATTGTTTGAGTCTAGGAGTTTGAGGCTGCAGTG..
 $ molecule:"dna"
```

If you pass `format='fasta'` to the above method, it will just return you a character vector containing a FastA formatted sequence.

[20] http://beta.rest.ensembl.org/documentation/info/sequence_id
[21] http://beta.rest.ensembl.org/documentation/info/sequence_region

2.7 Variation

2.7.1 variationAllele

```
> var = variationAllele( 'C', '9:22125503-22125502:1', 'human' )
> var                       # The response object
[[1]]
hgvs        :
  C = 9:g.22125502_22125503insC
transcripts : 10 in total
> var[[1]]$transcripts[1:2] # Just the top 2 transcripts of the first response to save room
[[1]]
data             :
        name         gene_id   transcript_id   biotype cdna_allele_string
1 CDKN2B-AS1 ENSG00000240498 ENST00000585267 antisense                -/C
  is_canonical
1        FALSE

alleles          :
        consequence_terms
1 downstream_gene_variant

[[2]]
data             :
        name         gene_id   transcript_id   biotype cdna_allele_string
1 CDKN2B-AS1 ENSG00000240498 ENST00000580576 antisense                -/C
  is_canonical
1        FALSE

alleles          :
        consequence_terms
1 downstream_gene_variant
```

2.7.2 variationId

```
> var = variationId( 'COSM476', 'human' )
> var                       # The response object
[[1]]
name        : COSM476
is_somatic  : TRUE
hgvs        :
  T = 7:g.140453136A>T
transcripts : 4 in total
> var[[1]]$transcripts[1:2] # Again, just the top 2 transcripts to save room
[[1]]
data             :
  name         gene_id   transcript_id        biotype       ccds
1 BRAF ENSG00000157764 ENST00000288602 protein_coding CCDS5863.1
  cdna_allele_string codon_position translation_stable_id translation_start
1                T/A              2       ENSP00000288602               600
  translation_end exon_number cdna_start cdna_end cds_start cds_end
```

```
1             600       15/18       1860     1860      1799    1799
  is_canonical
1         TRUE

alleles          :
  display_codon_allele_string pep_allele_string codon_allele_string
1                     gTg/gAg               V/E             GTG/GAG
             hgvs_transcript                  hgvs_protein polyphen_score
1 ENST00000288602.6:c.1799T>A ENSP00000288602.6:p.Val600Glu          0.999
  polyphen_prediction sift_score sift_prediction consequence_terms
1   probably damaging       0.02     deleterious  missense_variant

protein_features :
      name                  db
1  PF07714         Pfam domain
2  PF00069         Pfam domain
3 SSF56112 Superfamily domains
4  SM00220       SMART domains
5  SM00219       SMART domains
6  PS50011    PROSITE profiles

[[2]]
data             :
  name         gene_id   transcript_id                biotype
1 BRAF ENSG00000157764 ENST00000479537 nonsense_mediated_decay
  cdna_allele_string codon_position translation_stable_id translation_start
1                T/A              2       ENSP00000418033                28
  translation_end exon_number cdna_start cdna_end cds_start cds_end
1              28         2/6         83       83        83      83
  is_canonical
1        FALSE

alleles          :
  display_codon_allele_string pep_allele_string codon_allele_string
1                     gTg/gAg               V/E             GTG/GAG
            hgvs_transcript                 hgvs_protein polyphen_score
1 ENST00000479537.1:c.83T>A ENSP00000418033.1:p.Val28Glu          0.946
  polyphen_prediction sift_score sift_prediction             V10
1   probably damaging       0.12       tolerated missense_variant
       consequence_terms
1 NMD_transcript_variant

protein_features :
      name                  db
1  PF00069         Pfam domain
2  PF07714         Pfam domain
3 SSF56112 Superfamily domains
4  PS50011    PROSITE profiles
```