

Computational & Comparative Genomics
Cold Spring Harbor, New York, United States
9-15 November 2011

Ensembl API

Bert Overduin, Ph.D.
Vertebrate Genomics Team
EMBL – European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD, UK



Access to data

- Ensembl web site <http://www.ensembl.org>
- *Pre!* web site <http://pre.ensembl.org>
- *Archive!* web site <http://archive.ensembl.org>
- BioMart <http://www.ensembl.org/biomart/martview>
<http://www.biomart.org/biomart/martview>
- FTP site <ftp://ftp.ensembl.org>
- Amazon Web Services <http://aws.amazon.com/publicdatasets>
- MySQL <http://www.ensembl.org/info/data/mysql.html>
- Perl API <http://www.ensembl.org/info/data/api.html>



Ensembl databases

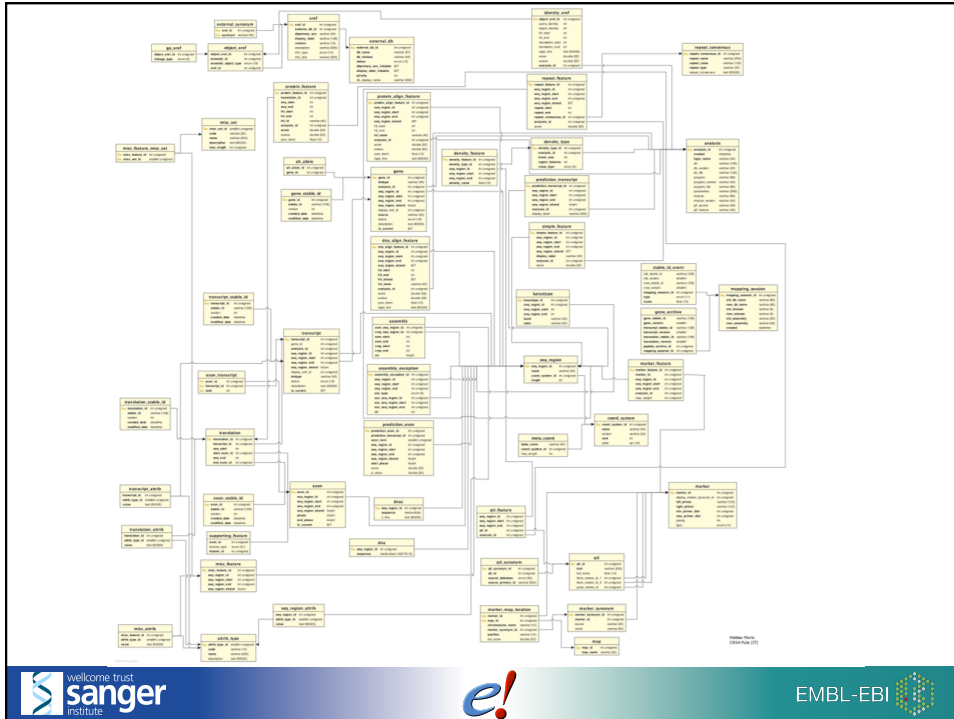
- MySQL
- Species-specific databases:
 - core: genomic sequences and most annotation
 - variation: genetic variation
 - funcgen: regulatory elements
- Cross-species database:
 - compara: all comparative data

MySQL

- SQL = Structured Query Language

Needed:

- MySQL client: <http://www.mysql.com>
- Ability to write MySQL
- Knowledge of database schema
- More info: <http://www.ensembl.org/info/data/mysql.html>



Retrieve the Ensembl Transcript and Protein IDs for the gene ENSG00000164305:

```
mysql -u anonymous -h ensembl.mysql.inf.ed.ac.uk -P 5306
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 395422503 to server version: 5.1.34-log

Type 'help;' or '\h' for help. Type '\c' to clear the buffer.

mysql> use homo_sapiens_core_64_37;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> SELECT gene_stable_id.stable_id AS gene,
transcript_stable_id.stable_id AS transcript,
translation_stable_id.stable_id AS peptide FROM gene, transcript,
translation, gene_stable_id, transcript_stable_id,
translation_stable_id WHERE gene.gene_id = transcript.gene_id AND
transcript.transcript_id = translation.transcript_id AND
gene_stable_id.gene_id = gene.gene_id AND
transcript_stable_id.transcript_id = transcript.transcript_id AND
translation_stable_id.translation_id = translation.translation_id
AND gene_stable_id.stable_id = 'ENSG00000164305';
```

Result:

```
+-----+-----+-----+
| gene          | transcript    | peptide       |
+-----+-----+-----+
| ENSG00000164305 | ENST00000308394 | ENSP00000311032 |
| ENSG00000164305 | ENST00000393585 | ENSP00000377210 |
| ENSG00000164305 | ENST00000523916 | ENSP00000428929 |
| ENSG00000164305 | ENST00000517513 | ENSP00000428372 |
| ENSG00000164305 | ENST00000393588 | ENSP00000377213 |
| ENSG00000164305 | ENST00000447121 | ENSP00000407142 |
| ENSG00000164305 | ENST00000438467 | ENSP00000390792 |
+-----+-----+-----+
7 rows in set (0.26 sec)
```

MySQL

Disadvantages

- Very good knowledge of database schemas needed
- Queries can become very complex
- Not recommended (and not supported) to retrieve sequences

Perl API

- API = Application Programming Interface
- Used by the Ensembl analysis and annotation pipeline and the Ensembl web code
- Robust, reliable and well-supported

Needed:

- Perl
- BioPerl modules (version 1.2.3)
- Ensembl modules
- (Basic) ability to code in Perl

Perl API basics

- Written in Object-Oriented Perl
- Data objects model biological entities, e.g. genes, markers, repeats, variations, ...
- Data objects are retrieved from and stored in the database using Objects adaptors
- Each Object adaptor is responsible for creating Data objects of only one particular type
- Object adaptors are created using the Registry
- The Registry also handles all database connections

Perl API basics

Slice object

- Represents a single continuous region of a genome
- Used to obtain sequence, features or other information from a particular region of interest

Feature object

- Has a defined location on the genome
- E.g. genes, markers, repeats, variations, ...

Retrieve the repeats on human chromosome 4:123370000-123380000:

```
#!/usr/bin/perl

use strict;
use warnings;

use Bio::EnsEMBL::Registry;
my $reg = "Bio::EnsEMBL::Registry";

$reg->load_registry_from_db(
    -host => 'ensemldb.ensembl.org',
    -user => 'anonymous');

my $slice_adaptor = $reg->get_adaptor('human', 'core', 'Slice');

my $slice = $slice_adaptor->fetch_by_region('chromosome', '4', 123370000, 123380000);

my @repeats = @{$slice->get_all_RepeatFeatures};
foreach my $repeat (@repeats){
    print
        $repeat->display_id, "\t",
        $repeat->seq_region_name, ":",
        $repeat->seq_region_start, "-",
        $repeat->seq_region_end, "\n";
}
```

Result:

```
LIMEg      4:123370040-123370344
ERVL-E-int 4:123370958-123371501
dust       4:123371441-123371449
dust       4:123371590-123371598
LTR33      4:123371902-123372337
dust       4:123374412-123374422
dust       4:123374820-123374831
MIR3       4:123375402-123375468
MIRc       4:123375664-123375857
MIR        4:123376647-123376889
dust       4:123378695-123378709
dust       4:123378728-123378745
MIRb       4:123378836-123379025
dust       4:123379336-123379344
AmnSINE2   4:123379539-123379636
```

Documentation & Help

- Installation instructions, schemas, schema descriptions and tutorial:
<http://www.ensembl.org/info/docs/api/core/index.html>
- Inline Perl POD (Plain Old Documentation)
- Web-browsable version of the POD (Doxygen)
<http://www.ensembl.org/info/docs/Doxygen/index.html>
- ensembl-dev mailing list:
<http://www.ensembl.org/info/about/contact/mailing.html>
- Ensembl helpdesk:
helpdesk@ensembl.org

Ensembl API workshops

- Typically 3 days, covering all Ensembl APIs
- Free of charge

- Mar 2012 ? Cambridge, United Kingdom
- 9-11 May 2012 Hinxton, United Kingdom
- 9-11 Oct 2012 Hinxton, United Kingdom
- Nov/Dec 2012 ? Cambridge, United Kingdom

- <http://www.ensembl.info/workshops/calendar/>

- Or host a workshop yourself!


Q U E S T I O N S
A N S W E R S