EasyMirror and EasyImport: Simplifying the setup of a custom Ensembl database and webserver for any species

Richard J Challis[1], Sujai Kumar[1], Lewis Stevens[1], Mark Blaxter[1]

1 - Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, EH9 3FL, UK

Author for correspondence: richard.challis@ed.ac.uk

Abstract

As the generation and use of genomic datasets is becoming increasingly common in all areas of biology, the need for resources to collate, analyse and present data from one or more genome projects is becoming more pressing. The Ensembl platform is a powerful tool to make genome data and cross-species analyses easily accessible through a web interface and a comprehensive API. Here we introduce the EasyMirror and EasyImport pipelines to facilitate the setup and hosting of custom Ensembl genome browsers.

EasyMirror (https://github.com/lepbase/easy-mirror) makes it possible to set up a mirror of any Ensembl or Ensembl Genomes (including Bacteria, Metazoa, Fungi, Plants and Protists) species in four simple steps that can be run in less than an hour on a fresh Ubuntu installation. This tool exploits the modular nature of the Ensembl codebase to allow a site to be set up with none, some or all of the data hosted locally.

EasyImport (https://github.com/lepbase/easy-mirror) extends this approach to simplify the import of genomic data for any species from standard flat files into the Ensembl database schema, ready to be deployed using EasyMirror. All that is needed to get started is a genome fasta file and the gene models in GFF format.

Documentation for both pipelines is available at http://easy-import.readme.io

Introduction

An increasing number of approaches in all areas of biology and ecology now rely on access to genomic sequence data for the taxa under investigation. This has fuelled rapid increases both in the number of genome sequencing projects, and in the number of research groups generating, assembling and annotating draft genome sequences. In order to maximise the value of these data, it is important to ensure that they are made accessible to the widest possible community of users.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2401v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

Academic publications require submission of draft genomes to the public databases (Cochrane et al. 2016) but often don't require submission of additional analyses such as gene prediction, functional annotation, variant-calling, and orthology discovery. Although the results of these analyses are published, it can be hard to get access to the underlying data in a consistent format. Some labs make these data available on their individual lab websites as simple downloadable files, but others make them more accessible by providing a genome browser, blast interface, and a way to search for specific genes by name, orthology, function, or protein domain annotation.

There are a number of tools that have been developed to host and share genomic data such as GBrowse (Donlin 2009), JBrowse (Skinner et al. 2009), UCSC Genome Browser (Speir et al. 2016), InterMine (Kalderimis et al. 2014), PlantGenIE (Sundell et al. 2015), and Ensembl (Yates et al. 2016). Ensembl is the most complete of these tools as it supports both single species and multi-species comparative views alongside variation and functional data. By using an Ensembl genome browser, projects can offer a familiar and standardised interface with access to a powerful application programming interface (API) to facilitate large-scale comparative analysis and data-mining. Genomic data also continue to be valuable well beyond the initial funding cycle. From an archival perspective data imported to an Ensembl database are stored in a format that is likely to have have long-term support with a mature database structure and codebase. The term "Ensembl" is colloquially used for the Ensembl websites hosted by the European Bioinformatics Institute (EBI). However, in this paper, we mainly use the term to refer to the Ensembl webserver code that is used to host these websites, and to the Ensembl database schema that stores genomic data which is displayed by the Ensembl webserver.

A small number of high-profile taxonomically oriented sites, such as VectorBase (Giraldo-Calderón et al., 2015), WormBase ParaSite (Howe et al., 2016) and AvianBase (Eöry et al., 2015), have been set up using Ensembl, however despite the advantages of using the Ensembl webserver, its adoption as a browser for genome project or lab-hosted databases has been limited. Ensembl has an extensive list of dependencies that must be installed before it can run and the complexity of the code and need to edit a number of interconnected configuration files can make it difficult to trace the cause of problems during installation.

Lepbase (Challis et al. 2016) is a taxon-oriented genomic resource for the Lepidoptera and includes an Ensembl providing access to most publicly available Lepidopteran genome assemblies and gene models. Here we present the EasyMirror and EasyImport pipelines that we have developed as part of the Lepbase project to facilitate the setup and hosting of a custom Ensembl site. EasyMirror ensures that all dependencies are met and enables a mirror site of any existing Ensembl species (i.e., a species already available at ensembl.org, including the Ensembl Genomes sites) to be set up with locally or remotely hosted data. EasyImport extends this approach and allows import of new sequence and gene model data into an Ensembl database from FASTA and GFF files, ready to be hosted using EasyMirror. EasyImport also includes scripts for import of additional annotations, verification and file export. While initially developed for a taxon-oriented resource, we also present examples of project-based and lab-specific Ensembls that we have created using these pipelines.

Pipeline description

The process of setting up a custom Ensembl webserver involves two stages, handled by the EasyMirror and EasyImport pipelines, respectively: (i) installing Ensembl code and dependencies, including local database- and webservers to host an Ensembl mirror site and (ii) importing additional datasets into the Ensembl schema for hosting alongside or in addition to mirrored data. Because both pipelines share common dependencies and require similar final configurations, they are packaged together (with EasyMirror as a submodule of the EasyImport git repository) and are described as a combined workflow below. The combined EasyMirror and EasyImport pipelines are split into 13 steps (Figure 1), of which only four are required to set up an Ensembl mirror site and five of the remaining steps are optional when importing a new genome assembly and gene models.

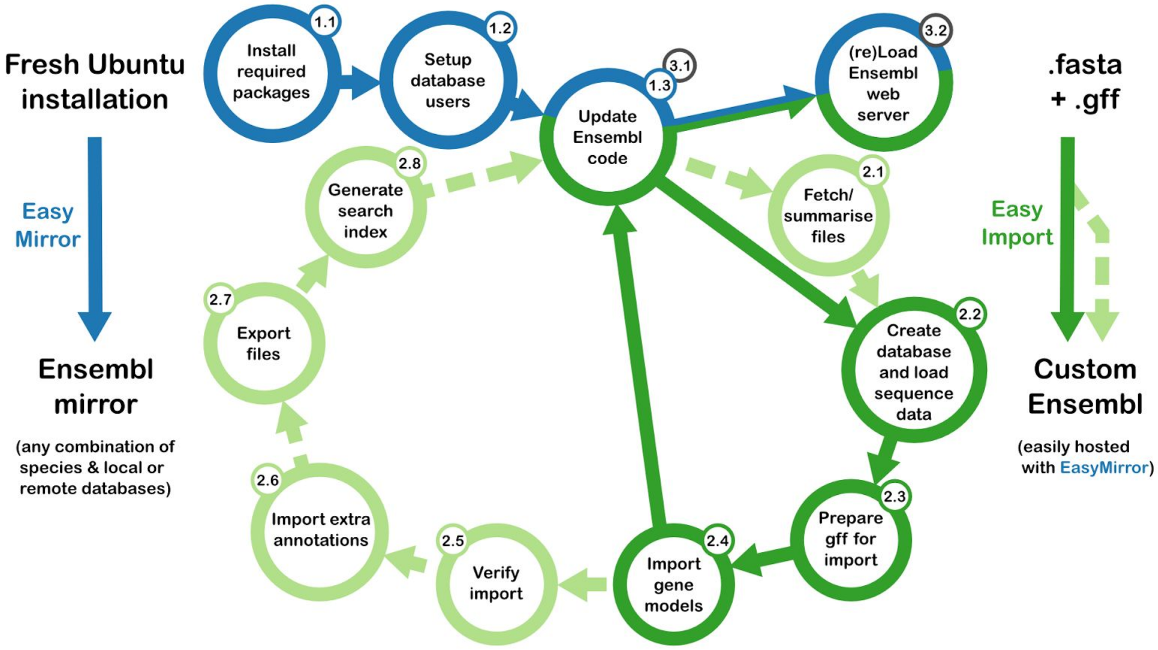
Fresh Ubuntu installation
Easy Mirror
Ensembl mirror
(any combination of species & local or remote databases)
1.1 Install required packages
1.2 Setup database users
1.3 3.1 Update Ensembl code
3.2 (re)Load Ensembl web server
2.1 Fetch/ summarise files
2.2 Create database and load sequence data
2.3 Prepare gff for import
2.4 Import gene models
2.5 Verify import
2.6 Import extra annotations
2.7 Export files
2.8 Generate search index
.fasta + .gff
Easy Import
Custom Ensembl
(easily hosted with EasyMirror)

Figure 1. Overview of the EasyMirror/EasyImport pipelines.

This paper provides an overview of the EasyMirror/EasyImport pipelines and is intended to provide an introduction to our more comprehensive and regularly updated documentation at http://easy-import.readme.io. The online documentation includes full descriptions of the commands used to run the steps below and detailed information on available configuration options for each step.

Configuration

All parameters for both EasyMirror and EasyImport are set and passed to scripts through INI format configuration files. This mirrors the use of INI files in ensembl-webcode allowing users setting up Ensembl instances using EasyImport to gain familiarity with the syntax conventions needed to further customise an instance. Use of configuration files rather than command line parameters also improves reproducibility since all parameters must be saved for the pipeline to be run.

We have taken the decision to use separate configuration files for conceptually separate parts of the pipeline, which prevents individual INI files from having to contain too much information and facilitates reuse of common parameters when importing more than one species. The drawback is that some details (database connection data in particular) are repeated across multiple INI files so care must be taken to use the correct template when making changes to the default settings.

Mirroring Ensembl with EasyMirror

1.1 Server setup

The Ensembl codebase has extensive software and perl module dependencies. EasyMirror offers a single script to install these dependencies on a fresh installation of Ubuntu 14.04. This is the only step within either EasyMirror or Easy import that is specific to a single Linux distribution, and the only one to require root access. It should be straightforward to adapt EasyMirror (and therefore EasyImport) to run on an alternate Linux distribution, using the package lists within the script install_dependencies.sh and exchanging the package manager (e.g. using yum instead of apt), adapting any version numbers as necessary.

1.2 Database creation

An Ensembl mirror can be run using only remotely hosted data, however there is still a requirement for a local database to hold session information. Mirrored data may also be hosted locally and this is often desirable as it offers improved performance, especially in locations with intermittent internet access. EasyMirror simplifies database and user setup, allowing data to be imported from remote database dumps and ensuring that database users are created with appropriate read/write or read-only permissions.

1.3 Update Ensembl code

With all the dependencies installed and databases set up, setting up an Ensembl site requires several Ensembl repositories hosted at https://github.com/ensembl and (optionally) at https://github.com/ensemblgenomes. EasyMirror automatically clones essential repositories, allowing control of which branch to checkout, and uses the plugin architecture of ensembl-webcode to allow additional plugins to cloned from alternate repositories. This allows custom plugins to be developed containing site-specific modifications and loaded as part of the EasyMirror pipeline. Database connection parameters and basic species configuration files are written automatically, with files stored in the Ensembl public-plugins/mirror repository. Once the Ensembl code has been updated to reflect the required configuration, the webserver is ready to be started as a mirror site by proceeding directly to step 3.2. Alternatively, new data can be imported ready to be hosted in an Ensembl site, as described below.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2401v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

Importing data with EasyImport

2.1 (optional) Fetch and summarise sequence files

EasyImport can retrieve files directly from local or remote directories, based on entries in the INI configuration file, ready to be imported. This ensures that the location of the source file used in the import is preserved to improve reproducibility. These can be files hosted on any ftp or http server such as the NCBI ftp servers, individual lab webservers, or sequencing provider servers. Subsequent steps will automatically retrieve the required files so this step is optional. However it can be useful to fetch FASTA, GFF and additional annotation files for inspection prior to importing. If run, this step will report a summary of GFF attribute counts by feature type, which is particularly useful in determining which fields may contain IDs, descriptions, etc. and whether any discrepancies in the counts of each attribute suggest that the GFF needs to be repaired by the GFF parser.

2.2 Database creation and sequence data import

Consistency with the latest Ensembl database schema is ensured by importing the database schema for each new core database from the Ensembl table.sql file and using an existing database with the same schema version as a template with the ensembl-production populate_production_db_tables.pl script. Scaffold/contig data will almost always be available as FASTA format (or possibly FASTA plus AGP). This is also the starting point for the Ensembl import pipeline.

2.3 Parsing and repairing GFF

A significant challenge for a file-based approach to data import is dealing with the fragmentation of the GFF format and the diverse ways that names, IDs, descriptions and annotations can be captured both in GFF attributes and across a range of other file types. The design of EasyImport is centered around a flexible GFF parser (https://github.com/rjchallis/gff-parser) that is designed to embrace the diversity of real world GFF files by allowing full customisation of expected relationships and properties with functions to repair, warn or ignore errors during validation.

Our GFF parser is a perl module that provides a mechanism to assign expectations and validation rules to specific GFF feature types while having very little hard-coded dependence on official gff specifications, allowing the flexibility to handle many more edge cases than other parsers, which allows gene models and annotations to be extracted from diverse, often invalid files. For EasyImport (in which the GFF parser is included as a git submodule) a subset of the full functionality can be controlled through a meta-syntax in the core import INI files and while the GFF parser can accommodate departures from GFF specification, EasyImport requires a GFF3 compatible format in column nine. The flexibility of this approach to GFF parsing initially adds some complexity to parameter specification, but many patterns can be reused across most GFF3 files and the benefit is that a record of all modifications to the original GFF3 file can be preserved in the INI file, maintaining a complete and unambiguous record of the import procedure to ensure full reproducibility.

2.4 Gene model import

While the GFF file format is standardised by the GFF parser, additional functionality in EasyImport allows for retrieval of gene IDs, synonyms and descriptions from GFF attributes as well as from FASTA headers and simple text files. EasyImport uses a simple match and replace syntax, specified in the INI file, to allow names to be matched across files where they may have different formats depending on whether they are gene, transcript or protein. For a community resource such as Lepbase, this provides the flexibility to incorporate information from diverse sources as supplied by individual labs rather than demanding a standardised format, which could act as a deterrent to full data sharing. For others implementing this pipeline, it offers the flexibility to integrate with existing protocols without the need to reformat data prior to import.

2.5 Import verification

The most common problems with data import from GFF files can be detected through comparison of expected protein sequences with translations exported from an Ensembl database. EasyImport checks that the same IDs are present in each of these sets, and that the sequences are identical. The most common causes of differences are alternate interpretations of phase and manual edits in the expected sequences file that terminate translations at the first stop codon.

2.6 Additional annotations

Some xrefs can be imported via Dbxref attributes in a GFF file, however, we have deliberately limited the extent to which additional annotations can be imported from GFF due to the complexity of validating additional feature types and of mapping from potentially variable attribute names to specific fields in the Ensembl core database tables. Several xref types can be richly represented in the Ensembl database if all required attributes are provided and this is easiest to ensure by working directly with program outputs. This also fits most closely with the lepbase.org model of ensuring consistency across genomes from diverse sources through annotating features with consistent databases/parameters. EasyImport currently supports direct import of blastp, InterproScan and RepeatMasker output files.

2.7 File export

Sequence file export (scaffold, protein and cds) provides access to bulk data files for analysis or to provide file downloads. Detailed summary statistics can also be exported in JSON format, which are used at ensembl.lepbase.org to populate the assembly statistics tables and in assembly stats (Challis 2016a) and codon usage (Challis 2016b) visualisations.

2.8 Search indexing

Ensembl supports a very basic (direct MySQL) search out of the box, this is best replaced with a search plugin so we have made index_database.pl available as part of EasyImport to generate an autocomplete/search index compatible with the lepbase-search (https://github.com/lepbase/lepbase-search) plugin.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2401v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

3.1 Update Ensembl code

Following species import, the ensembl-webcode configuration needs to be updated as in step 1.3 so this step reuses the same code which, after modification of the INI file to include the new species, will setup the code ready to host the new species data.

3.2 (re)Load Ensembl site

The final step is to load the ensembl site, which will become available on the port specified in the configuration file.

Example

Using the default example and running the commands in Box 1 will give a working Ensembl with a newly imported species assembly and gene models of the winter moth *Operophtera brumata* (Derks et al. 2015).

```
sudo apt-get update
sudo apt-get upgrade
sudo apt-get install git
git clone --recursive https://github.com/lepbase/easy-import ei
cd ei/em
sudo ./install-dependencies.sh ../conf/setup.ini
./setup-databases.sh ../conf/setup-db.ini
./update-ensembl-code.sh ../conf/setup.ini
mkdir ~/import
cd ~/import
perl ../ei/core/import_sequences.pl ../ei/conf/core-import.ini
perl ../ei/core/prepare_gff.pl ../ei/conf/core-import.ini
perl ../ei/core/import_gene_models.pl ../ei/conf/core-import.ini
cd ~/ei/em
./update-ensembl-code.sh ../conf/setup.ini
./reload-ensembl-site.sh ../conf/setup.ini
```

Box 1. Complete list of required commands to set up an Ensembl mirror site on a fresh Ubuntu 14.04 installation using default configuration files. The resulting Ensembl site will contain five Lepidoptera species, four mirrored from EnsemblMetazoa plus a direct import of sequences and gene models for the winter moth *Operophtera brumata* and will be available at http://localhost:8080.

Case studies

EasyImport facilitates the import of diverse data into Ensembl databases, making it relatively straightforward to set up and host Ensembl sites for any set of assemblies for which sequence data and gene models are available. We have used the pipeline described here to set up three distinct Ensembl instances reflecting some of the alternate use cases for

EasyMirror/EasyImport: (i) a taxon-oriented resource at ensembl.lepbase.org; (ii) project data sharing at ensembl.caenorhabditis.org; and (iii) a site to host assemblies and annotations generated by a genomics lab at ensembl.ngenomes.org.

ensembl.lepbase.org

Lepbase (Challis et al. 2016) is a taxon-oriented genomic resource for the Lepidoptera. One of the core services Lepbase provides is the Ensembl genome browser at ensembl.lepbase.org. EasyMirror was developed as part of the Lepbase project to allow us to quickly replicate server setup while deploying new instances for development and testing and moving our site between virtual machines. The generalisation of EasyMirror to accommodate any combination of species and database locations has allowed us to implement various features of the Lepbase Ensembl as plugins, which aids debugging as we can revert to Ensembl code and/or data by changing lines in an INI file.

As a taxon-oriented resource, Lepbase hosts genomes from a variety of sources with many different variations in file formats. Our initial species import scripts required modification to accommodate the idiosyncrasies of each new assembly dataset, which was an approach that scaled poorly as the number of available assemblies increased. EasyImport is the result of extending the EasyMirror approach to the import of new data into an Ensembl, allowing us to import new data sets as they become available while maintaining a reproducible record of the import process in the import INI file. Incorporation of new species data into our production site then only requires adding the new database name to the EasyMirror INI file.

ensembl.caenorhabditis.org

The Caenorhabditis Genomes Project (CGP) aims to produce genome assemblies and annotations for all *Caenorhabditis* species (~50) currently in culture. These data are shared between all project partners and made publicly available through the CGP Ensembl site at ensembl.caenorhabditis.org. This instance combines mirrored pre-existing Ensembl databases for several species (eg. *C. elegans*) and custom databases based on imported data. This project-level genome browser is essential to success of the CGP, allowing for the release of data which is often preliminary and hence not suitable for release at WormBase (Howe et al., 2016) and other public databases.

ensembl.ngenomes.org

The Blaxter Lab is a genomics and bioinformatics group at the University of Edinburgh leading or collaborating on genomics projects across a large number of animals, plants, fungi and bacteria, and hosted across a number of lab and project websites. The EasyMirror/EasyImport pipelines provide an opportunity to rationalise many of these resources into a common interface. We are beginning to add our published projects to a lab Ensembl setup sat ensembl.ngenomes.org and will also begin adding unpublished datasets as they become available to accelerate the dissemination of our research.

Future development

Ensembl provides the architecture to combine individual genomes, comparative analyses, variation and functional information in a single resource. EasyMirror already allows for hosting these additional features and we plan to facilitate the required data import in the next stages of development for EasyImport. We have been actively working on data import to the compara database schema. Unlike our approach to core species databases, the process of comparative analysis import is contingent on the analysis used and files generated so this is not part of universal pipeline described above but will be described separately alongside the analytical methods (Kumar et al., in prep). For many additional features, Ensembl scripts are already available to load data from a standard file format (e.g. variation data), so we plan to incorporate these into the EasyImport pattern using the existing INI config format to set up database and set parameters to allow the scripts to be run in the same straightforward and reproducible way as the rest of the EasyImport pipeline.

Conclusions

Making genomic datasets available through a human and machine accessible online portal increases the value of those datasets to the wider community. Several tools are available for groups producing or collating assemblies and annotations that fall outside the criteria for inclusion in major public databases to host their own online resources. For these groups we would recommend Ensembl (Yates et al. 2016), often considered to be one of the more difficult tools to implement but arguably the most powerful and supporting the widest suite of integrated data views. EasyMirror and EasyImport make the setup of a custom Ensembl and addition of new data relatively easy, and are suitable for creating taxon-oriented, project-based and lab-specific ensembl sites.

Acknowledgements

Lepbase is funded by a Biotechnology and Biological Sciences Research Council Tools and Resources fund award (BB/K020161/1, BB/K019945/1, BB/K020129/1). We wish to thank Paul Kersey, Daniel Zerbino, James Allen, Gareth Maslen, Kevin Howe and Bruce Bolt for helpful discussions during our visit to the EBI.

References

Challis, R. 2016a. assembly-stats 1.5. . Available at: https://zenodo.org/record/56996#.V8A4oJMrKuM.

Challis, R. 2016b. codon-usage v1. . Available at: https://zenodo.org/record/56681?ln=en#.V8A42ZMrKuM.

Challis, R.J., Kumar, S., Dasmahapatra, K.K.K., Jiggins, C.D. and Blaxter, M. 2016.

Lepbase: the Lepidopteran genome database. *BioRxiv*.

Cochrane, G., Karsch-Mizrachi, I., Takagi, T. and International Nucleotide Sequence Database Collaboration 2016. The international nucleotide sequence database collaboration. *Nucleic Acids Res* 44(D1), pp. D48–50.

Derks, M.F.L., Smit, S., Salis, L., Schijlen, E., Bossers, A., Mateman, C., Pijl, A.S., de Ridder, D., Groenen, M.A.M., Visser, M.E. and Megens, H.-J. 2015. The Genome of Winter Moth (Operophtera brumata) Provides a Genomic Perspective on Sexual Dimorphism and Phenology. *Genome Biol Evol* 7(8), pp. 2321–2332.

Donlin, M.J. 2009. Using the Generic Genome Browser (GBrowse). *Curr Protoc Bioinformatics* Chapter 9, p. Unit 9.9.

Kalderimis, A., Lyne, R., Butano, D., Contrino, S., Lyne, M., Heimbach, J., Hu, F., Smith, R., Stěpán, R., Sullivan, J. and Micklem, G. 2014. InterMine: extensive web services for modern biology. *Nucleic Acids Res* 42(Web Server issue), pp. W468–72.

Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. 2009. JBrowse: a next-generation genome browser. *Genome Res* 19(9), pp. 1630–1638.

Speir, M.L., Zweig, A.S., Rosenbloom, K.R., Raney, B.J., Paten, B., Nejad, P., Lee, B.T., Learned, K., Karolchik, D., Hinrichs, A.S., Heitner, S., Harte, R.A., Haeussler, M., Guruvadoo, L., Fujita, P.A., Eisenhart, C., Diekhans, M., Clawson, H., Casper, J., Barber, G.P. and Kent, W.J. 2016. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res* 44(D1), pp. D717–25.

Sundell, D., Mannapperuma, C., Netotea, S., Delhomme, N., Lin, Y.-C., Sjödin, A., Van de Peer, Y., Jansson, S., Hvidsten, T.R. and Street, N.R. 2015. The plant genome integrative explorer resource: plantGenIE.org. *New Phytol* 208(4), pp. 1149–1156.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F.J. and Flicek, P. 2016. Ensembl 2016. *Nucleic Acids Res* 44(D1), pp. D710–6.