Release coordination documentation

Preparation

Preparing this document itself

This document is usually inherited from the previous release cycle and tends to have all the check-boxes ticked. The fastest way to untick them all to start afresh is to go to "Edit...", then switch over to the XML view (a button on the top right with <> on it), and then perform a mass-replace of <ac:task-status>complete</ac:task-status> by <ac:task-status>incomplete</ac:task-status> .

Declaration of Intentions

☑ Once the release coordinator has sent out the email for the declaration of intentions, set up a web page with intentions in the Confluence wiki system to allow easy tracking of the progress.Release plans

☑ Ask compara team members of their intentions
Compara has one extra day to declare their intentions because of the need to know what the genebuilders and associated teams (eg wormbase, ensembl genomes) will declare

☑ Submit the declaration of intentions on the http://admin.ensembl.org/index.html website
Click here to expand...

- Give a short and meaningful title
- Leave "*Species affected*" to "*All species*"
- Describe the change in "*Content*"

- "*Status*" is probably "*Declared*" at this stage, but it may happen that you can already mark it as "*Handed over*"
- Change the "*Site Type*" to "*GRCh37*" if needed
- The default values of the remaining fields are fine

☑ Discuss with the team which declarations are worth being put on the front page (3 at most). For those, leave "*Headline position on the homepage*" to "Not a headline" but fill in a correct *category* ("API/Schema changes", "New alignments", etc)

Environment variables

☑ Define $ENSEMBL_CVS_ROOT_DIR
This is necessary to run the Hive and is used by many scripts/files in this document. Make sure this is defined in your terminal

☑ Define $ENSADMIN_PSW
The password for the mysql 'ensadmin' user also needed for many scripts

☑ Define $COMPARA_REG variable to simplify connecting to databases via registr y

```
export COMPARA_REG="-reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
-reg_type compara -reg_alias"
```

☑ Define $CURR_ENSEMBL_RELEASE variable to simplify naming of databases

```
export CURR_ENSEMBL_RELEASE=`perl -mBio::EnsEMBL::ApiVersion -e 'print
Bio::EnsEMBL::ApiVersion::software_version()."\n"'`

 # make sure that you got the right value (your ensembl checkout has to be
up-to-date)
echo $CURR_ENSEMBL_RELEASE

 # it is also very handy to have the previous release number in a variable:
export PREV_ENSEMBL_RELEASE=`expr $CURR_ENSEMBL_RELEASE - 1`
```

Update your checkouts

Ensure you have up-to-date git checkouts of at least the following repositories, pointing at master branch:

☑ ensembl-compara

☑ ensembl

☑ ensembl-hive

☑ ensembl-analysis

☑ ensj-healthcheck

Compara servers

Check out the current space on the compara servers and delete the last but one release. Leave the previous release for healthchecks. Check with the other compara team members before deleting.

☑ Check space on the admin website

☑ Ask compara team members to tidy up any unwanted databases (run the command below for all compara servers) and inform them of the intention to delete the last but one release

how much space do databases take?

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl/misc-scripts/db/db-space.pl -host compara4
-port 3306 -user ensadmin -password $ENSADMIN_PSW
```

Guidelines for the deletion of Databases

******Before dropping any of the databases, they should be backed up and zipped. At least the eHive tables should be backed on the production databases while research databases should be completely backed up.**
- release database: keep only the last one (for HCs)
- production databases: keep the last production run of each pipeline on each species-set (e.g. keep the primate EPO ***and*** the mammal EPO ***and*** etc

Configuration file

☑ Update production_reg_conf.pl and check back into git:

Click for details

Update the registry configuration file $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl that will be used throughout the release process.

Make sure to have edited the release numbers, added external core databases and fixed name prefixes.

The convention right now (rel.82) is to keep the merged release database on compara5.

DB connection details of production databases (families, nctrees, etc.) can be removed from the file until merge.

Schema preparation

All the schema changes must be ready by the handover of the core databases. See API / Schema changes for the procedure about changing the schema

Update table.sql and create patch files

Here we need to prepare table.sql for the new schema and create the relevant patch files. The general procedure is defined in API / Schema changes.

The other Compara members may have already created patches, so check with them what's there. At the minimum, there should be the schema version increase.

Click here for details

Update the $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/table.sql file and create any patch files.

☑ Create a patch file for the schema_version

☑ Update the schema_version in table.sql and delete the other patch INSERT statements

☑ Update the schema_version in the master database

```
UPDATE meta SET meta_value = XX WHERE meta_key = 'schema_version';
```

☑ Check if any other patch files need creating by looking at the Declaration of Intentions and checking with other compara team members

☑ Add an INSERT statement for the new schema_version in table.sql and for any other new patches

Check the patch files

The schema defined in the current table.sql must be obtainable by patching the previous database. There is a shell script to do the comparison

Click here for details

Run the script $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/production/schema_diff.sh (you may have to press "y" to validate the patches to apply to the old schema) and check the output. The script relies on several things:

- The API version declared by the Core API (in Bio::EnsEMBL::ApiVersion) has been updated
- The meta keys in the live database are correct (they should !)

The only allowed difference is that peptide_align_feature_XX tables are only found in the previous database, not the new one. The script writes 3 files to the current directory: old_schema.sql, patched_old_schema.sql, and new_schema.sql, and automatically runs sdiff. If you need to look at the diff later on, run this:

```
sdiff -w 200 -bs patched_old_schema.sql new_schema.sql
```

git commit table.sql and any patch files

How to patch a database to the latest schema

Click here to expand...

- Use the following script to detect the what schema the database is on and to apply all the required patches to bring it to the latest schema

```
$ENSEMBL_CVS_ROOT_DIR/ensembl/misc-scripts/schema_patcher.pl --host=<>
--user=ensadmin --pass=${ENSADMIN_PSW}
--database=ensembl_compara_database_${CURR_ENSEMBL_RELEASE}
```

☑ Patch the previous production and ancestral databases using the above script

Master database

The Master database can be updated at any time but must be ready before the main pipelines start. For more information about its role, consult Master database.

NCBI taxonomy data

The production team updates the ncbi_taxonomy database on livemirror just before the handover to us (please check that this has been done). We then need to update the tables on our master DB. The current (rel.84) master database is mm14_ensembl_compara_master on compara1

☑ Update the ncbi_taxa_node and ncbi_taxa_name in the master database

Click here for details

The ncbi_taxonomy database is located in mysql://ens-livemirror:3306/ncbi_taxonomy

mysqldump

```
time db_cmd.pl $COMPARA_REG ncbi_taxonomy -executable mysqldump --prepend
--extended-insert --prepend --compress ncbi_taxa_node ncbi_taxa_name |
db_cmd.pl $COMPARA_REG compara_master
```

It usually takes between 30 and 60 seconds.

☑ Load ensembl_aliases.sql onto the master database

⌄ Click here for details

The script will report any discrepancies that need to be resolved ie any nodes which have been deleted from the ncbi_taxonomy database but still have entries in the ensembl_aliases.sql file. Check if these have an entry in the species_set_tag table. If not, it is probably safe to delete them. Check with other compara team members.

load ensembl_aliases

```
db_cmd.pl $COMPARA_REG compara_master <
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/taxonomy/ensembl_aliases.sql
```

☑ Run the CheckTaxon healthcheck

⌄ Click here to expand...

Run the CheckTaxon healthcheck early to find any discrepancies between the ncbi_taxon_name table and the core databases (information about how to set up the healthchecks can be found here)

Run healthcheck

```
#cd to your local healthcheck git repo :
cd ensj-healthcheck/

# make sure you are using the right version of JAVA:
export JAVA_HOME=/software/jdk1.6.0_14

# if you need to recompile (submit to the farm, because you need more memory
than is available on the head) :
bsub -I ant clean jar

# run the healthchecks (submit to the farm, because you need more memory than
is available on the head) :
time bsub -I ./run-configurable-testrunner.sh -h compara1 -d
mm14_ensembl_compara_master --host2 ens-staging2 -t
org.ensembl.healthcheck.testcase.compara.CheckTaxon
```

Add new entries to compara master database

The current master database is called mm14_ensembl_compara_master on compara1. You have to create new genome_dbs and dnafrags when there is a new assembly or a new species. Any new genome_dbs, dnafrags and method_link_species_set_ids need to be added before production starts.

☑ Add a new species / assembly (i.e. a genome_db)

⌄ Click here for details

You will have to run this script once per new / updated species

Add genome_db

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/update_genome.pl
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--compara compara_master --species "gadus_morhua" --release --collection
ensembl
```

If you know that the species won't be ready for this release but the next one, you need to remove "--release --collection ensembl".

Next release, you'll have to rerun update_genome.pl with the --force option

Add the new genome_db_id to the confluence page Release plans. This script may take a while if the species you are adding is new and has a lot of scaffolds. You can check the progress by counting *dnafrag* entries in the master database:

```
SELECT COUNT(*) FROM dnafrag;
```

Add the alias names and other tags to $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/taxonomy/ensembl_aliases.sql

- Each new species must have a 'ensembl alias name'. It should match the "web_name" used by the production team in the "species" table of their "ensembl_production" database on staging1.

  check species/taxa from ensembl_production

  ```
  db_cmd.pl $COMPARA_REG ensembl_production -sql 'select production_name, web_name, taxon from species'
  ```

- Each extant species is anchored to the species tree at a certain taxon. This taxon must be described with two fields in the *ncbi_taxa_name* table. $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/taxonomy/place_species.pl helps in placing the new species in the current compara species tree and getting those tags

  check species/taxa from ensembl_production

  ```
  $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/taxonomy/place_species.pl -master_url mysql://ensro@compara1/mm14_ensembl_compara_master -taxon_ids 9940,9361,7994,7918 -collection ensembl
  ```

  There are two tags to add for each taxon: 'ensembl alias name': a common name or a "simple English" description of the taxon, and 'ensembl timetree mya': the age of the taxon. If the ensembl alias name is missing, try wikipedia and google. If ensembl timetree mya is missing, go to the TimeTree URL given by the script and check whether the data really is missing, or if they changed the format of their URL / HTML pages.

For any new species, update the file $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/taxonomy/ensembl_aliases.sql to add the new tags and reload the file into the master database

☑ Add in extra non-reference patches.

Click here for details

This is currently done when a new patch for either human or mouse is released. This may have already been done, please ask.>

Details about the patches can be found here ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/ e.g. for patch 11: ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37.p11/README

It is first necessary to find if any patches have been deleted or updated since alignments on these need to be deleted from the Compara database. This is done by running the find_assembly_patches.pl script on the new and previous release of the core database

Find assembly patches

```
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/find_assembly_patches.pl -compara "mysql://ensro@compara1/mm14_ensembl_compara_master" \
-new_core
"mysql://ensro@ens-staging1:3306/homo_sapiens_core_81_38?group=core&species=homo_sapiens" -prev_core
"mysql://ensro@ens-livemirror:3306/homo_sapiens_core_80_38?group=core&species=homo_sapiens"
```

Sample ouptut

```
NEW patches
  CHR_MG3231_PATCH 100440 2015-05-12 16:29:47
  CHR_MG4265_PATCH 100420 2015-05-12 16:29:47
  CHR_MG3609_PATCH 100418 2015-05-12 16:29:47
  CHR_MG117_PATCH 100424 2015-05-12 16:29:47
  CHR_MG3562_PATCH 100416 2015-05-12 16:29:47
  CHR_MG4259_PATCH 100430 2015-05-12 16:29:47
  CHR_MG4255_PATCH 100438 2015-05-12 16:29:47
  CHR_MG4266_PATCH 100422 2015-05-12 16:29:47
  CHR_MG4248_PATCH 100412 2015-05-12 16:29:47
  CHR_MG3561_PATCH 100414 2015-05-12 16:29:47
  CHR_MG4249_PATCH 100434 2015-05-12 16:29:47
  CHR_MG4261_PATCH 100432 2015-05-12 16:29:47
  CHR_MG4254_PATCH 100436 2015-05-12 16:29:47
  CHR_MG4264_PATCH 100428 2015-05-12 16:29:47
CHANGED patches
  CHR_MG132_PATCH new=100426 2015-05-12 16:29:47        prev=100407 2014-09-22
11:50:45 dnafrag_id=14025314
DELETED patches
  CHR_MG4237_PATCH 100405 2014-09-22 11:50:45   dnafrag_id=14025313
DnaFrags to delete:
  names: ("CHR_MG132_PATCH","CHR_MG4237_PATCH")
  dnafrag_ids: (14025314,14025313)
Input for create_patch_pairaligner_conf.pl:
--patches
chromosome:CHR_MG3231_PATCH,chromosome:CHR_MG4265_PATCH,chromosome:CHR_MG4259_
PATCH,chromosome:CHR_MG4266_PATCH,chromosome:CHR_MG4248_PATCH,(...)
```

Copy the output to the Intentions for Release page as it will be needed to clean-up the alignments

In this case, there are 14 NEW patches, 1 CHANGED patch and 1 DELETED patch. They can be imported to / deleted from the master database by running update_genome.pl with the --force option

Add patches

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/update_genome.pl
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--compara compara_master --species human --force
```

The steps for running the pairwise alignment pipeline for new patches can be found here:
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/pipelines/READMEs/pair_aligner_patches.txt

☑ Add in the new LRGs

Click here to expand...

LRGs are needed by the Family pipeline, and have to be updated every release. This is done by running the update_genome.pl script on human with the --force option

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/update_genome.pl
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--compara compara_master --species human --force
```

Note that the new LRGs may have already been loaded by the previous step (*add in the human patches*) as the same update_genome.pl command is run

To check if everything loaded OK, compare the output of the following queries:

```
db_cmd.pl --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--reg_type core --reg_alias human -sql 'select count(*) from seq_region join
coord_system cs using(coord_system_id) where cs.name="lrg"'

db_cmd.pl --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--reg_alias compara_master -sql 'select count(*) from dnafrag where
coord_system_name="lrg"'
```

they should be the same.

Add method_link_species_set entries to compara master database

These are usually added by the people that need them, please check.

The release coordinator (or any team member) should create a new method_link_species_set in the master database before starting a new pipeline in order to get a unique method_link_species_set_id. Ideally they can be created before starting to build the new database although new method_link_species_sets can be added later on.

☑ Add dna method_link_species_set entries

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl
--method_link_type LASTZ_NET --genome_db_id 90,142 --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--compara compara_master
```

If you are sure that the mlss will be ready for the next release, you can add --release to the command-line. Otherwise you will have to mark the entry manually as OK before merging

☑ Add synteny method_link_species_set entries

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl
--method_link_type SYNTENY --genome_db_id 90,142 --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
-compara compara_master
```

If you are sure that the mlss will be ready for the next release, you can add --release to the command-line. Otherwise you will have to mark the entry manually as OK before merging

☑ Add species tree method_link_species_set entry

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl
--method_link_type SPECIES_TREE --force --collection "ensembl" --name "species
tree" --source "ensembl" --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
-compara compara_master --release
```

☑ Add protein-trees method_link_species_set entry

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl
--method_link_type PROTEIN_TREES --force --collection "ensembl" --name "protein
trees" --source "ensembl" --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
-compara compara_master --release
```

☑ Add ncRNA-trees method_link_species_set entry

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl
--method_link_type NC_TREES --force --collection "ensembl" --name "nc trees"
--source "ensembl" --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
-compara compara_master --release
```

☑ Add family method_link_species_set entry

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl
--method_link_type FAMILY --force --collection "ensembl" --name "families"
--source "ensembl" --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
-compara compara_master --release
```

☑ Add pairwise ortholog method_link_species_set entry

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl
--method_link_type ENSEMBL_ORTHOLOGUES --force --pw --collection "ensembl"
--source "ensembl" --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
-compara compara_master --release 2>
create_mlss.${CURR_ENSEMBL_RELEASE}.ENSEMBL_ORTHOLOGUES.err >
create_mlss.${CURR_ENSEMBL_RELEASE}.ENSEMBL_ORTHOLOGUES.out
```

Inspect the out and err files for errors

☑ Add singleton paralog method_link_species_set entry

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl
--method_link_type ENSEMBL_PARALOGUES --force --sg --collection "ensembl"
--source "ensembl" --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
-compara compara_master --release 2>
create_mlss.${CURR_ENSEMBL_RELEASE}.ENSEMBL_PARALOGUES.err >
create_mlss.${CURR_ENSEMBL_RELEASE}.ENSEMBL_PARALOGUES.out
```

Inspect the out and err files for errors

☑ Reset the URL of reused mlss_ids
In case the same mlss_id can be reused, the pipeline will probably complain that there is already a URL attached to it. You need to reset these URLs

```
UPDATE method_link_species_set SET url = "" WHERE method_link_species_set_id IN
(<LIST_OF_mlss_ids>);
```

Usually, this happens when there are no new assemblies, in which case you need to give the mlss_id of the Family, ncRNA-tree and protein-tree pipelines

Add new species to phylogenetic tree

☑ Add new species to phylogenetic tree

Click here for details

The easiest way to use this is to use the phylowidget.
Paste in the content of $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/species_tree.eukaryotes.topology.nw, select Arrow and select where you want the new species to go (use ncbi taxonomy or wikipedia etc) eg Gadus morhua.
Then select in the menu "Tree Edit > Add > Sister". Click on the empty node and edit name (add new name). Leave the branch length as it is.
The tree should appear in the Toolbox: save it into
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/species_tree_blength.nh
git commit

Final edits to compara master database

This runs once all the species have been added / updated.

☑ Compare the staging servers to the master database

Click here to expand...

This script will list the genomes of the staging servers, and compare them to the master database.

```
# Run once in dry-run mode to see the changes
perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/update_master_db.pl
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--compara compara_master --dry_run
```

** note: The above might throw a warning about not finding any species name in the ancestral db, this is expected and should not break the test!!

Some things (like the genebuild date is different) can be directly changed by the script (use the --nodry-run option instead). If there is a new assembly / species on the staging servers that is not yet in the master database, run update_genome.pl (see above) to add it, and re-run update_master_db.pl to make sure things are solved.

☑ Healthcheck the master database

⌄ Click here to expand...

See JAVA Healthchecks and use the ComparaMaster group

```
#cd to your local repo of healthcheck
cd ensj-healthcheck/
$ENSEMBL_CVS_ROOT_DIR/ensj-healthcheck/run-configurable-testrunner.sh -h
compara1 -d mm14_ensembl_compara_master --host2 ens-staging2 -g ComparaMaster
```

Production

Now we can run all our production pipelines

End of production window

Create Release Database

Create the new database for the new release and add it to your registry configuration file. Use the $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/sql/table.sql file to create the tables and populate the database with the relevant primary data and genomic alignments that can be reused from the previous release. This can be done with the $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/populate_new_database.pl script. It requires the master database, the previous released database and the fresh new database with the tables already created. The script will copy relevant data from the master and the old database into the new one.

☑ >Create new database

⌄ Click here for details

Create database

```
db_cmd.pl $COMPARA_REG compara_curr -sql "CREATE DATABASE"
db_cmd.pl $COMPARA_REG compara_curr <
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/sql/table.sql
```

☑ Populate the new database

⌄ Click here for details

Before you start copying, make a dry run of the populate_new_database.pl with -intentions flag to review the list of mlss_ids to be copied:

populate_new_database intentions

```
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/populate_new_datab
ase.pl --reg-conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_con
f.pl \
--master compara_master --old compara_prev --new compara_curr
--intentions > populate_new_database.intentions
```

This normally takes less than a minute and produces a long list.

If you believe some of the MLSS or SS entries should NOT be copied, connect to the master database and change the last_release of the unwanted entries to the previous release number. Conversely, if you want to prepare some new MLSS or SS entries to be copied (e.g. the newly run pipelines), change the first_release of the wanted entries to the current release number.

NB: OLD INSTRUCTIONS FOR PRE-first_release/last_release API: *There are cases where the mlss does not change but the underlying data does, e.g. the "patch-to-ref" alignment (H.sap-H.sap lastz-patch and M.mus-M.mus lastz-patch). These have a mlss_id of 556 (H.sap) and 624 (M.mus) and are currently set in the skip_mlss. If there are no new patches, this needs to be removed to allow the existing data to be copied. If there are new patches, please ensure the 'skip_mlss' is set in the meta table. However, the entry in the method_link_species_set table will not be copied and will need to be added manually.*

Start the copying:

populate_new_database

```
bsub -q yesterday -R "select[mem>20000] rusage[mem=20000]" -M20000 -Is bash
time $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/populate_new_database.pl \
    --reg-conf $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl --master compara_master --old compara_prev --new compara_curr > populate_new_database.out
```

Click here for run times

took 3 hours for rel.pre57 (copied from rel.56)
took 3 hours for rel.57 (copied from rel.pre57)
took 2:15 hours for rel.58 (copied from rel.57)
took 2:09 hours for rel.59 (copied from rel.58)
took 3 hours for rel. 60 (copied from rel.59)
rel.64: 2.6h
rel.65: 2.5h
rel.66: 4.8h
rel.67: 2.1h (launched from compara3)
rel.68: 1h40m (run on compara3)
rel.69: 2.5h
rel.70: ~3.5h (compara1 was slow)
rel.71: 4.1h (compara3)
rel.72: 5.1h (compara3)
rel.73: 5.5h (compara2)
rel.74: 2h:3' (compara3)
rel.75: 5.5h (compara5)
rel.77: 9.7h (compara5)
rel.78: 6.0h (compara4)
rel.79:
rel.80:
rel.81: 6h (compara5)

rel.82: 5.5h (compara5)

rel.83: 4.8h (compara5)

rel.85: 7.5h (compara5)

If new method_link_species_sets are added in the master after this, you use this script again to copy the new relevant data. In such case, you will have to:

- skip the old_database in order to avoid trying to copy the dna-dna alignments and syntenies again

- empty ncbi_taxa_name before running

populate_new_database from master only

```
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/populate_new_database.pl \
    --reg-conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl --master compara_master --new compara_curr
```

☑ Delete any pairwise alignments on non-reference patches that have been DELETED or UPDATED.

Click here for details

Find the output of find_assembly_patches.pl script that you ran previously (usually for Human and Mouse) and combine their "Dnafrags to delete" into one common list:

delete patches

```
DNAFRAGS_2_DELETE="(14025314,14025313)"

db_cmd.pl $COMPARA_REG compara_curr -sql "SELECT count(*) FROM genomic_align
ga1, genomic_align ga2, genomic_align_block gab WHERE
ga1.genomic_align_block_id = ga2.genomic_align_block_id AND
ga1.genomic_align_id != ga2.genomic_align_id AND  ga1.genomic_align_block_id =
gab.genomic_align_block_id AND ga1.dnafrag_id in $DNAFRAGS_2_DELETE"

db_cmd.pl $COMPARA_REG compara_curr -sql "DELETE ga1, ga2, gab FROM
genomic_align ga1, genomic_align ga2, genomic_align_block gab WHERE
ga1.genomic_align_block_id = ga2.genomic_align_block_id AND
ga1.genomic_align_id != ga2.genomic_align_id AND  ga1.genomic_align_block_id =
gab.genomic_align_block_id AND ga1.dnafrag_id in $DNAFRAGS_2_DELETE"
```

☑ Run healthchecks on the release database

Click here for details

Run the healthchecks to make sure the the release database is consistent after the initial population of data.

Click here for how to setup and run the healthchecks

Run the compara_external_foreign_keys healthcheck

healthcheck

```
cd $ENSEMBL_CVS_ROOT_DIR/ensj-healthcheck

# make sure you are using the right version of JAVA:
export JAVA_HOME=/software/jdk1.6.0_14

# if you need to recompile (submit to the farm, because you need more memory
than is available on the head) :
bsub -I ant clean jar

# some tests need more memory than the farm3's default:
time bsub -q yesterday -M8000  -R"select[mem>8000]  rusage[mem=8000]" -I
./run-configurable-testrunner.sh -h compara5 -d lg4_ensembl_compara_81 --host2
ens-staging2 -g ComparaShared
```

Click here for run times

rel.83: 13 minutes, 2 expected complaints (CheckSpeciesSetSizeByMethod may complain about Human-on-Human lastz-new and ForeignKeyMasterTables will complain about empty MethodLink entries (this will be deleted later in the merging process))

rel.85: 20mins

and correct any newly detected problems

Merge DNA data

NOTE: All the runs of copy_data.pl (except the last one) should have the flag "-re_enable 0" to avoid recomputing the indices in the end of each run.

Running "-re_enable 1" will add *at least 2 hours* (rel.82) to the merging time (but it is necessary in the final product) so make sure you only do it once.

☑ Pairwise alignments: LASTZ_NET (and, formerly, BLASTZ_NET or TRANSLATED_BLAT_NET)

NOTE : For merging pw alignments involving haplotypes, go to the next point

Click here for details

These data are usually in separate production databases. You can copy them using the $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl script. This script requires write access to the production database if the dnafrag_ids need fixing. Use the flag -re_enable 0 on all calls apart from the last one to avoid recomputing the indices.

Also, check first_release of these databases. In case it hasn't been set, you need to do it **now** on **both** the production database and the master database,

Example:

copy_data

```
# for each source URL: first plug in the --from_url and add --dry_run to check
that the script has found the right MLSS:
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--to_reg_name compara_curr --method_link_type LASTZ_NET --re_enable 0
--from_url mysql://ensadmin:${ENSADMIN_PSW}@compara4/sf5_ggal_falb_lastz_73
--dry_run

    # if happy, remove the --dry_run flag and run it again, preferably on the
farm:
bsub -q yesterday -R "select[mem>5000] rusage[mem=5000]" -M5000 -I time
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--to_reg_name compara_curr --method_link_type LASTZ_NET --re_enable 0
--from_url mysql://ensadmin:${ENSADMIN_PSW}@compara4/sf5_ggal_falb_lastz_73
```

☑ The curious case of LASTZ_PATCH alignments. There is **always** something to copy, even if there are no new patches

Click here for details

You will also have to copy Human_ref_vs_Human_patches and Mouse_ref_vs_Mouse_patches LASTZ_PATCH alignments, but mind the source:

If there were new patches, you'll import them in a way similar to other LASTZ:

copy_data

```
#First run the following pipeline to import the alignments between patches /
haplotypes and primary regions.

init_pipeline.pl
Bio::EnsEMBL::Compara::PipeConfig::ImportPatchAlignmentsToRef_conf -host
comparaX

#then
# note the method_link_type is LASTZ_PATCH !
bsub -q yesterday -R "select[mem>5000] rusage[mem=5000]" -M5000  -I time
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--to_reg_name compara_curr --method_link_type LASTZ_PATCH --re_enable 0
--from_url <the_url_of_the_pipeline_db_from_above>
```

If there were no new patches, you will still have to copy them from compara_prev, since LASTZ_PATCH alignments are automatically skipped by populate_new_database.pl script. You simply have to refer to the previous database as the source:

copy_data

```
# note the method_link_type is LASTZ_PATCH !
bsub -q yesterday -R "select[mem>5000] rusage[mem=5000]" -M5000  -I time
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--to_reg_name compara_curr --method_link_type LASTZ_PATCH --re_enable 1
--from_reg_name compara_prev
```

☑ Pairwise alignments: non-reference patches for the high coverage LASTZ_NET alignments. This is ito be used when merging pw alignments involving haplotypes.

Click here for details

This step is now very similar to the previous.

Do not forget --merge and --patch_merge options.

Also, if it's the last one you might want to switch keys back on

copy_data --merge --patch_merge

```
# first plug in the --from_url and add --dry_run to check that the script has
found the right MLSS:
bsub -q yesterday -R "select[mem>5000] rusage[mem=5000]" -M5000  -I time
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--to_reg_name compara_curr --method_link_type LASTZ_NET --method_link_type
BLASTZ_NET --method_link_type TRANSLATED_BLAT_NET --re_enable 1 --merge
--patch_merge --from_url  mysql://ensro@compara1/mm14_lastz_human --dry_run

    # if happy, remove the --dry_run flag and run it again, preferably on the
farm:
bsub -q yesterday -R "select[mem>5000] rusage[mem=5000]" -M5000  -I time
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--to_reg_name compara_curr --method_link_type LASTZ_NET --method_link_type
BLASTZ_NET --method_link_type TRANSLATED_BLAT_NET --re_enable 1 --merge
--patch_merge --from_url  mysql://ensro@compara1/mm14_lastz_human
```

☑ Multiple alignments: PECAN, EPO, EPO_LOW_COVERAGE, GERP_CONSTRAINED_ELEMENT, GERP_CONSERVATION_SCORE

Click here for details

These data are usually in separate production databases. You can copy them using the $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl script. This script requires write access to the production database if the dnafrag_ids need fixing or the data must be copied in binary mode (this is required for conservation scores).

Some alignments produce conservation scores and constrained elements (check the Release plans) and these need to be copied separately.

eg

copy_data multiple alignment

```
bsub -q yesterday -R "select[mem>5000] rusage[mem=5000]" -M5000 \
    -I time
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl \
    --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--to_reg_name compara_curr \
    --method_link_type EPO --method_link_type EPO_LOW_COVERAGE
--method_link_type PECAN \
    --method_link_type GERP_CONSTRAINED_ELEMENT --method_link_type
GERP_CONSERVATION_SCORE \
    --from_url
mysql://ensadmin:${ENSADMIN_PSW}@compara2/sf5_epo_low_8way_fish_71 -re_enable
0
```

EPO alignments produce ancestral sequences and a separate core database which must also be copied. See below.

Click here for run times

rel 71.
2m: kb3_hsap_ggal_lastz_71 mlss_id=632
1m kb3_mmus_ggal_lastz_71 mlss_id=633
1m kb3_ggal_mgap_lastz_71 mlss_id=634
1m kb3_ggal_xtro_tblat_71 mlss_id=638
1m kb3_hsap_ggal_tblat_71 mlss_id=637
1m sf5_olat_gmor_lastz_71 mlss_id=625
3m kb3_pecan_20way_71   mlss_id=630
4m kb3_pecan_20way_71   mlss_id=631
35m kb3_pecan_20way_71   mlss_id=50045
3m sf5_compara_epo_6way_71 mlss_id=548
1m sf5_olat_onil_lastz_71 mlss_id=626
1m sf5_olat_xmac_lastz_71 mlss_id=627
1m sf5_epo_low_8way_fish_71 mlss_id=628
2m sf5_epo_low_8way_fish_71 mlss_id=629
9m sf5_epo_low_8way_fish_71 mlss_id=50044
1m kb3_ggal_drer_tblat_71   mlss_id=639
1m kb3_ggal_csav_tblat_71   mlss_id=640
1m sf5_ggal_acar_lastz_71   mlss_id=636
91m sf5_ggal_tgut_lastz_7   mlss_id=635 (re-enable 1)
93m sf5_compara_epo_3way_birds_71 mlss_id=641 (re-enable 1)
14m sf5_compara_epo_3way_birds_71 mlss_id=642 (re-enable 1)
16m sf5_compara_epo_3way_birds_71 mlss_id=50046 (re-enable 1)

☑ Check the keys have been re-enabled

Click here for details

Use mysqlshow to highlight if the table still has disabled keys. The text "disabled" will be shown in the Comment column if the key is disabled. An empty Comment column indicates the keys are enabled.

mysqlshow interprets any underscores in the last argument as a wildcard so to get round this, we need to use % as the last argument.

mysqlshow

```
db_cmd.pl $COMPARA_REG compara_curr --executable mysqlshow  -- --keys
genomic_align_block %
db_cmd.pl $COMPARA_REG compara_curr --executable mysqlshow -- --keys
genomic_align %
db_cmd.pl $COMPARA_REG compara_curr --executable mysqlshow  -- --keys
genomic_align_tree %
db_cmd.pl $COMPARA_REG compara_curr --executable mysqlshow  -- --keys
conservation_score %
db_cmd.pl $COMPARA_REG compara_curr --executable mysqlshow  -- --keys
constrained_element %
```

If there are still tables with keys disabled run the following on them:

```
db_cmd.pl $COMPARA_REG compara_curr -sql "ALTER TABLE <table_name> ENABLE
KEYS";
```

☑ Syntenies

Click here for details

First make sure the entries in $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl file point at the latest (staging) versions of the core databases.

Before running the code... Ensure that you check the synteny coverage and if it is less than 1%, It must be deleted from mlss, mlss_tag, dnafrag_region and synteny_region tables. **This should be automated in the synteny pipeline by release 84.

Example

load synteny data

```
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--to_reg_name compara_curr --method_link_type SYNTENY --from_url
mysql://ensro@compara1/cc21_synteny_83 --dry_run
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl --reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--to_reg_name compara_curr --method_link_type SYNTENY --from_url
mysql://ensro@compara1/cc21_synteny_83
```

☑ Build a new ancestral sequence core database

Click here for details

Putting together the database of ancestral sequence is now done using a dedicated Hive-Core mini-pipeline.

Check you have the most recent core checkout ie the correct schema and patch files are added to the meta table.

Go to $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/modules/Bio/EnsEMBL/Compara/PipeConfig and open the PipeConfig file AncestralMerge_conf.pm .

Make sure you have edited/checked the following:

1) current release number

2) names and locations of current and previous ancestral core databases

3) the table of ancestral sequence sources in the second analysis (some entries might point to the previous release ancestral database, some will be new)

For (3), you can run the following query on your release database and on the previous database: (NB: method_link_id=13 is equivalent to method_link_type = "EPO")

EPO query

```
SELECT * FROM method_link_species_set WHERE method_link_id = 13;
```

The new mlss_id should be attached to their production database:
  '641' => 'mysql://ensadmin:$ENSADMIN_PSW@compara3/sf5_3birds_ancestral_sequences_core_71'
The mlss_id that are reused should be linked to the previous database
  '505' => $self->o('prev_ancestral_db'),

The current (as or rel.75) list of ancestral alignments are:

5 teleost fish
6 primates
4 sauropsids ("birds")
15 eutherian mammals

Save the changes, exit the editor and run init_pipeline.pl with this file:

init_pipeline

```
init_pipeline.pl AncestralMerge_conf.pm -host compara5
```

Then run both -sync and -loop variations of the beekeeper.pl command suggested by init_pipeline.pl . This pipeline will merge the separate ancestral core sources into ensembl_ancestral_{rel_number}.

You may want to check the msg table for errors and have a look at the result of the merger:

Which Ancestral sequences do we have?

```
SELECT left(name,12) na, count(*), min(seq_region_id), max(seq_region_id),
max(seq_region_id)-min(seq_region_id)+1 FROM seq_region GROUP BY na;
```

If everything is ok, measure the time:

how much time did running of the pipeline take?

```
call time_analysis('%')
```

Click here for run times

rel.67: 20min
rel.71: 20min
rel.75: 21min

Then drop hive-specific tables:

drop hive tables

```
CALL drop_hive_tables;
```

Make sure all tables are myISAM.

```
 SHOW TABLE STATUS where engine != 'MyISAM';
```

or, if no new multiple alignments were run, copy it over from the previous release

Click here for details

Create a new database for ancestral sequences:

```
 db_cmd.pl $COMPARA_REG ancestral_curr -reg_type core -sql 'CREATE DATABASE'
```

Copy over the data from the previous release:

```
 time db_cmd.pl $COMPARA_REG ancestral_prev -reg_type core -executable
mysqldump | db_cmd.pl $COMPARA_REG ancestral_curr -reg_type core
# took 45 minutes in rel.81
# took 42 minutes in rel.82
# took 38 minutes in rel.83
# took 38 minutes in rel.84
# rel.85: 40mins
```

Patch the database to the current release by applying the relevant patches from $ENSEMBL_CVS_ROOT_DIR/ensembl/sql or use a schema patcher script.

```
$ENSEMBL_CVS_ROOT_DIR/ensembl/misc-scripts/schema_patcher.pl --host=compara5
--user=ensadmin --pass=${ENSADMIN_PSW}
--database=lg4_ensembl_ancestral_${CURR_ENSEMBL_RELEASE}
```

If patches were applied, make sure you have both analyzed and optimized the tables:

```
db_cmd.pl $COMPARA_REG ancestral_curr -reg_type core -executable mysqlanalyze
--verbose
db_cmd.pl $COMPARA_REG ancestral_curr -reg_type core -executable mysqloptimize
--verbose
```

Merge GeneTrees+Families+NCTrees+PatchProjectionsAsHomologies

- ☑ Check that Protein-trees have been run and handed-over ("*Compara hands over Homologies*" date)
- ☑ Check that ncRNA-trees have been run
- ☑ Check that Families have been run
- ☑ Check that LRGs were included in the Families
- ☑ Once Production has updated all the xrefs (incl. gene names and descriptions), the pipeline that merges the gene side (*ImportAltAlleGroupsAsHomologies_conf*) can be run.
  NB: make sure that all the source databases' URLs are correctly registered in the master_db's MLSS table (if the pipelines have moved in between, make sure to edit the relevant MLSS.url columns).

  Click here for details and times

checking the URLs

```
db_cmd.pl -reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
-reg_alias compara_master -sql "SELECT mlss.* FROM method_link_species_set
mlss JOIN method_link ml USING (method_link_id) WHERE ml.type IN ('FAMILY',
'PROTEIN_TREES', 'NC_TREES') AND first_release IS NOT NULL AND last_release IS
NULL"
```

Now initialize and run the *ImportAltAlleGroupsAsHomologies_conf* pipeline.

Running times:

rel.81 : 1h03m

☑ Run the Hive pipeline (*EnsemblMergeDBsIntoRelease_conf*) to merge tables from all the four products into the release database

Click here for details

Go to $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/modules/Bio/EnsEMBL/Compara/PipeConfig/Example and open the PipeConfig file EnsemblMergeDBsIntoRelease_conf.pm

It has a 'urls' hash where you will have to change the names of the databases and possibly their locations:

master_db - is the main compara master
prev_rel_db - should point to the previous release database
curr_rel_db - should point to the current release database being merged into (not the Hive pipeline database, but purely Compara schema product)

protein_db - should point to the current GeneTrees pipeline database
family_db - should point to the current Families pipeline database
ncrna_db - should point to the current ncRNAtrees pipeline database
projection_db - should point to the current PatchProjectionsAsHomologies pipeline database.

Also choose the server to run the merging pipeline on (you don't need a lot of resources or memory, as it is purely Hive book-keeping) and set the 'host' default_option.

Save the changes, exit the editor, init and run the merging pipeline with this file:

running the merging pipeline

```
init_pipeline.pl EnsemblMergeDBsIntoRelease_conf.pm
beekeeper.pl ... -sync
beekeeper.pl ... -loop
```

This pipeline will merge all the protein-side products into the release database.

Click here for times

rel.73 was the first experimental run, code had to be fixed, servers had to be reconfigured, so merging took one whole working day.

In the merging database run: call time_analysis('%');

rel.75 : 5 hours
rel.76 : 5.6 hours

...

rel.82 : 6.1 hours

☑ Load the species-trees (needed for the Species-tree view)

```
$ init_pipeline.pl Bio::EnsEMBL::Compara::PipeConfig::LoadSpeciesTrees_conf
-compara_alias_name compara_curr -host compara5
# Then run beekeeper as suggested by init_pipeline.pl
```

# Note: the last analysis of this pipeline failed in rel.82 (all 4 jobs of this analysis) trying to insert duplicated entries into MLSS_tag table, but all the data was there, so I just carried on. same in rel.85

☑ Drop the three databases used for merging.

⌄ Click here for details

```
# After you are happy about the result of protein side merging you can drop
the "YourName_homology_projections_ThisRelease" database.
$ db_cmd.pl -url
mysql://ensadmin:${ENSADMIN_PSW}@compara5/lg4_homology_projections_${CURR_ENSE
MBL_RELEASE} -sql 'drop database'

 # same for the LoadSpeciesTrees database:
$ db_cmd.pl -url
mysql://ensadmin:${ENSADMIN_PSW}@compara5/lg4_load_species_trees_${CURR_ENSEMB
L_RELEASE} -sql 'drop database'

 # same for the MergeDBsIntoRelease database:
$ db_cmd.pl -url
mysql://ensadmin:${ENSADMIN_PSW}@compara5/lg4_pipeline_dbmerge_${CURR_ENSEMBL_
RELEASE} -sql 'drop database'
```

☐ git commit the changes to the PipeConfig files that you have made.

Final database checks

☑ Remove redundant method_link entries

⌄ Click here for details

In most cases they can be removed, but check with other members of Compara. Remove redundant method_link entries

method_link entries

```
-- prepend with:     db_cmd.pl -reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
-reg_alias compara_curr -sql

SELECT ml.* FROM method_link ml LEFT JOIN method_link_species_set mlss
USING(method_link_id) WHERE mlss.method_link_id IS NULL;
DELETE ml FROM method_link ml LEFT JOIN method_link_species_set mlss
USING(method_link_id) WHERE mlss.method_link_id IS NULL;

note**** we deleted 18 mlss_ids in rel.83
```

☑

☑ Check that all the schema patches have been declared and applied.

- Click here for details

  If unsure, recheck the current schema against the previous schema. See Check the patch files for details

Run the healthchecks

☑ Update the code

- Click here for details

  The healthchecks are written in java and need to be recompiled after a git pull.

compile healthchecks

```
cd $ENSEMBL_CVS_ROOT_DIR/ensj-healthcheck
export JAVA_HOME=/software/jdk1.6.0_14
git pull
bsub -I ant clean jar
```

We don't need to configure a database.properties any more. Everything is done from the command line

☑ Run the healthchecks for ancestral database

- Click here for details

```
time bsub -M8000  -R"select[mem>8000]  rusage[mem=8000]" -I
./run-configurable-testrunner.sh -h compara5 -d sf5_ensembl_ancestral_77 -g
ComparaAncestral
```

It should take less than a minute (if the tables are analyzed / optimized) and usually complains about 1 thing that you can ignore:

```
org.ensembl.healthcheck.testcase.generic.AssemblySeqregion [Team
responsible: GENEBUILD]
  mm14_ensembl_ancestral_80: 0 rows found in assembly table
```

⚠ If healthcheck indicates that tables need to be analysed, follow instructions here: Analyze / Optimize the databases

☑ Update the max_alignment_length IF NECESSARY.

- Click here for details

  Check that the max_alignment_lengths have been computed.

update max_alignment_length

```
time bsub -I ./run-configurable-testrunner.sh -h compara5 -d
sf5_ensembl_compara_77 -t
org.ensembl.healthcheck.testcase.compara.MLSSTagMaxAlign
```

If not (the healthcheck is failing), you can repair it by adding the --repair flag:

update max_alignment_length

```
time bsub -I ./run-configurable-testrunner.sh -h compara5 -d
sf5_ensembl_compara_77 -t
org.ensembl.healthcheck.testcase.compara.MLSSTagMaxAlign --repair 1 --user
ensadmin --password $ENSADMIN_PSW
```

☑ Update the alignment mlss_id of the conservation score IF NECESSARY

⌄ Click here for details

update conservation score mlss_id

```
time bsub -I ./run-configurable-testrunner.sh -h compara5 -d
sf5_ensembl_compara_77 -t
org.ensembl.healthcheck.testcase.compara.MLSSTagGERPMSA
```

If the healthcheck is failing, you can repair it by adding the --repair flag:

update conservation score mlss_id

```
time bsub -I ./run-configurable-testrunner.sh -h compara5 -d
sf5_ensembl_compara_77 -t
org.ensembl.healthcheck.testcase.compara.MLSSTagGERPMSA --repair 1 --user
ensadmin --password $ENSADMIN_PSW
```

☑ Run the ComparaAll group of healthchecks on the release database

⌄ Click here for details

The 'stdbuf -o0' is a trick to prevent the pipe from buffering the output, since in addition to storing it we also want to examine the output visually.

compara_external_foreign_keys

```
time bsub -q yesterday -M8000  -R"select[mem>8000]  rusage[mem=8000]" -I
stdbuf -o0 ./run-configurable-testrunner.sh -h compara5 -d
lg4_ensembl_compara_81 --host2 ens-staging2 -g ComparaAll | tee
healthchecks_after_merge.txt
```

☑ Run the ControlledComparaTables group of healthchecks on the release database

⌄ Click here for details

The 'stdbuf -o0' is a trick to prevent the pipe from buffering the output, since in addition to storing it we also want to examine the output visually.

compara_external_foreign_keys

```
time bsub -q yesterday -M8000  -R"select[mem>8000]  rusage[mem=8000]" -I stdbuf -o0 ./run-configurable-testrunner.sh -h compara5 -d lg4_ensembl_compara_81 --host2 compara1 --compara_master.database mm14_ensembl_compara_master -g ControlledComparaTables | tee healthchecks_controlled_tables_after_merge.txt
```

☑

Test web server

☑ Ask ensembl-production to point the test web server to the compara release database
Upon confirmation from the release coordinator ask other members of Compara to check their data on:
http://staging.ensembl.org/

Final handover of databases [edit]

Analyze / Optimize one last time

☑ This is required for the CopyDbOverServer script to work properly.
Click here to expand...

1. Run ANALYZE_TABLE on compara and ancestral databases
   Click here for details

   analyze table

   ```
   time db_cmd.pl $COMPARA_REG compara_curr   -executable mysqlanalyze --verbose
   time db_cmd.pl $COMPARA_REG ancestral_curr -reg_type core -executable mysqlanalyze --verbose
   ```

2. Run OPTIMIZE_TABLE on compara and ancestral databases
   Click here for details

   optimize table

   ```
   time db_cmd.pl $COMPARA_REG compara_curr   -executable mysqloptimize --verbose
   time db_cmd.pl $COMPARA_REG ancestral_curr -reg_type core -executable mysqloptimize --verbose
   ```

Both operations will take between a few seconds and a few hours, depending on the state of the tables / indexes

Copy databases to staging servers

☑ Logon to ens-staging1 and import the databases

- Click here for details

  The import script MUST be run on the DESTINATION machine with the *mysqlens* user. NB: ask for the password for *mysqlens* well in advance - there may be no-one around you at the right moment!
  You should also check whether there is enough space (~270Gb) on the disk before starting the copy.

  The script assumes that your environment is set ($ENSEMBL_CVS_ROOT_DIR and $ENSADMIN_PSW)

  ```
  ssh ens-staging1
  df -h /mysql
  su -m -c
  $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/production/import_compara_release_databases.pl mysqlens
  ```

  It should take 1h30-2h to run

- ☑ Logon to ens-staging2 and import the databases (same procedure)

File handover

- ☑ If we have generated a new "Age of Base" file, it has to be copied to /nfs/ensnfs-dev/staging/homo_sapiens/GRCh38/compara/ (usually it's named Hsap_ages_${mlss_id}_${release_number}.bb).
  Llet the web-team know if you have copied a new file or if they should consider the file from the previous release .

Final bits

- ☑ Send an email to ensembl-production to announce the handover the databases and cc ensembl-compara@
- ☑ Update the Declaration of Intentions on the admin website to indicate what has been handed over and what didn't make it and has been postponed
- ☑ Dump the master database and place the copy in a safe place
  - Click here to expand...

    It should take a couple of minutes at most to run:

    dump master database

    ```
    db_cmd.pl $COMPARA_REG compara_master --executable mysqldump | gzip - >
    /warehouse/ensembl01/compara/master_dumps/ensembl_compara_master_${RELEASE_VERSION}.mysql.gz
    ```

Post-handover

Update documentation and diagrams [edit]

> ⓘ It is now time to update the static files. This should be done before we branch the code

- ☑ Update the pipeline diagrams for all the pipelines that have been run this release
  - Click here for details

    Go to the docs directory

pipeline diagrams

```
cd $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/pipelines/diagrams
generate_graph.pl $COMPARA_REG compara_ptrees -output ProteinTrees.png
generate_graph.pl $COMPARA_REG compara_nctrees -output ncRNAtrees.png
generate_graph.pl $COMPARA_REG compara_families -output Families.png
generate_graph.pl -url mysql://ensro@compara3/kb3_pecan_20way_71 -output MercatorPecan.png
generate_graph.pl -url mysql://ensro@compara4/sf5_epo_35way_68 -output EpoLowCoverage.png
generate_graph.pl -url mysql://ensro@compara4/sf5_compara_epo_13way_69 -output epo_pt3.png
```

Commit any changed diagrams to git and push.

☑ Update the schema documentation and diagrams
⌄ Click here for details

generate new schema documentation

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-production/scripts/sql2html.pl -i $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/sql/table.sql -o $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/schema/compara_schema.html -d Compara -host compara5 -user ensro -dbname lg4_ensembl_compara_75 -sort_headers 0 -sort_tables 0 -intro $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/schema/intro.html
```

Open the output file $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/schema/compara_schema.html in your browser and check that no example errors are reported.

If everything looks fine, copy this file to public-plugins and commit&push both (the compara one and the webcode one) :

update schema documentation for web

```
cp $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/schema/compara_schema.html $ENSEMBL_CVS_ROOT_DIR/public-plugins/docs/htdocs/info/docs/api/compara/
```

If necessary, update schema diagrams by loading the $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/sql/table.sql schema file into MySQL Workbench, rearrange/colour the nodes and export into PNG.

The schema diagrams will have to be copied both to $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/schema/diagrams and to public-plugins and committed&pushed in both repositories:

update schema diagrams for web

```
cp -r $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/schema/diagrams/*.png $ENSEMBL_CVS_ROOT_DIR/public-plugins/docs/htdocs/info/docs/api/compara/diagrams/
```

☑ Update the API tutorial documentation
⌄ Click here for details

Update the tutorial documentation compara_tutorial.html in this directory:

$ENSEMBL_CVS_ROOT_DIR/ensembl-webcode/htdocs/info/docs/api/compara/

Be careful that the embedded Perl snippets must use HTML-escaped characters (e.g. &lt; and &gt;) and be wrapped in a <pre class="code sh_perl">

Open the URL /info/docs/api/compara/compara_tutorial.html from a sandbox / test website and export it as a PDF in

$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/ComparaTutorial.pdf

To make the pdf look nicer, you can issue a few JavaScript commands to remove the Ensembl headers. See Creating PDF version of VEP docs for more details

- ☑ Update the tutorial about Compara resources

  Click here to expand...

  Do the same with $ENSEMBL_CVS_ROOT_DIR/ensembl-webcode/htdocs/info/website/tutorials/compara.html . This pages can be viewed online at http://staging.ensembl.org/info/website/tutorials/compara.html

- ☑ Update main ensembl species tree if there are any new species

  See $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/species_tree/README

  The end result should go here (check how it should be moved) : http://www.ensembl.org/info/about/species_tree.pdf

- ☑ Check examples work in ensembl-compara/scripts/examples/
- ☑ git commit and push any modified files or added tutorial examples
- ☑ Update the declared intention with removed / deprecated methods

  Click for details

  We need to generate the list of methods exported by the objects / adaptors on the master and release/{$n-1} branches, and compare (diff) them.

Check deprecated / removed methods

```
# on master/ensembl-compara/modules/Bio/EnsEMBL/Compara/
grep "^sub " *pm | sort > ~/MASTER
# on the previous release/ensembl-compara/modules/Bio/EnsEMBL/Compara/
grep "^sub " *pm | sort > ~/RELEASE75
sdiff -w 200 -bs ~/RELEASE75 ~/MASTER | less

# Let's do the same for the adaptors
# on master/ensembl-compara/modules/Bio/EnsEMBL/Compara/
grep "^sub " DBSQL/*pm | sort > ~/MASTER
# on the previous release/ensembl-compara/modules/Bio/EnsEMBL/Compara/
grep "^sub " DBSQL/*pm | sort > ~/RELEASE75
sdiff -w 200 -bs ~/RELEASE75 ~/MASTER | less
```

In both cases, make sure the methods are really removed, and not moved to a base / sub class, etc

Branch the code

- ☑ Check with the rest of Compara that it is ok to branch the code as it is, then create the 'release/THIS_RELEASE_NUMBER' branch in git and push it to the server.

```
cd $ENSEMBL_CVS_ROOT_DIR/ensembl-compara
git pull master
git branch release/<release number>
git checkout release/<release number>
git push origin release/<release number>
```

Data dumps

** All dumps to be done by the release coordinator!!!

These should only be done once ensembl-production has given the go-ahead for this. This is to avoid overloading the databases whilst biomart is being run.

Most dumps (except homology) are currently generated in /lustre/scratchXXX areas and then have to be manually assembled into the /nfs/ensembl/ensembl/ftp_ensembl/release-XX/ tree.
Look at the previous release tree to get the idea. The first level of directories normally defines the file type, and the second level is the team name (except fasta/ where species are mixed).

Ensembl-compara is responsible for the following dumps:

> bed/ensembl-compara (MSA)
> emf/ensembl-compara (MSA and homologies)
> maf/ensembl-compara (multiple_alignments and pairwise_alignments)
> xml/ensembl-compara (homologies)
> fasta/ancestral_alleles (the only one without ensembl-compara in the path)

Make sure they are all either generated by running new dumps, or if some MSAs did not run in this release - sym-linked from the previous release's /warehouse/ens_ftp_arch_03/release-Xx/ tree.

DO NOT LINK OR COPY FROM PREVIOUS RELEASE's /lustre/scratch areas, as they may no longer point to real data!!!
If you do not have enough permissions for reading from /warehouse/ens_ftp_arch_03/release-Xx/ talk to the web-team and ask for it to be fixed.

☑ DNA data dumps
The instructions are in $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/pipelines/READMEs/multi_align.dumps.txt

☑ Gene tree dumps

Click here for details

Go to $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/modules/Bio/EnsEMBL/Compara/PipeConfig and open the PipeConfig file DumpTrees_conf.pm

Check that you are happy about all parameters. In usual cases, they can all be set from the command line and the config file does not need editing.

Make sure you have the XML::Writer module in your PERL5LIB (there is a copy in ~mm14/src/perl/orthoxml/)

Run init_pipeline.pl with this file:

init_pipelline

```
init_pipeline.pl DumpTrees_conf.pm -host compara5 -member_type protein
```

> ! rel.64: testing sqlite mode failed: too many occurrences of "database locked". We should stick to mysql or significantly reduce the analysis_capacities.

Then run the beekeeper.pl -loop command suggested by init_pipeline.pl . This pipeline took 2h8m to run in rel82.

It will produce protein_tree dumps in the directory pointed at by 'target_dir' parameter. ('/lustre/scratch110/ensembl/'.$self->o('ENV', 'USER').'/'.$self->o('pipeline_name'))

It also automatically copies the file to /nfs/ensembl/ensembl/ftp_ensembl/release-XX/, so just check that the files are there with correct sizes, and you're done.

Create the ncRNA pipeline from the same config file:

init_pipeline

```
init_pipeline.pl DumpTrees_conf.pm -host compara5 -member_type ncrna
```

Then run the beekeeper.pl -loop command suggested by init_pipeline.pl . This pipeline took 36m to run in rel82.

This pipeline will produce ncrna_tree dumps in the directory pointed at by 'target_dir' parameter and copy them to /nfs/ensembl/ensembl/ftp_ensembl/release-XX/.

Commit the DumpTrees_conf.pm file into git if you'd like to keep the changes.

☑ Copy the tree content dump for Uniprot

Click here for details

The file 'target_dir'/ensembl.GeneTree_content.{release}.txt.gz needs to be copied to the EBI ftp server, and then MD5 checksum computed and stored next to it:

init_pipeline

```
scp
/lustre/scratch110/ensembl/lg4/protein_81_dumps/ensembl.GeneTree_content.e81.txt.gz #change the files & user for release you are working on.
login.ebi.ac.uk:/nfs/ftp/pub/databases/ensembl/ensembl_compara/gene_trees_for_uniprot
ssh login.ebi.ac.uk #login to ebi
cd /nfs/ftp/pub/databases/ensembl/ensembl_compara/gene_trees_for_uniprot
md5sum ensembl.GeneTree_content.e<CURR_RELEASE_NUMBER>.txt.gz >
ensembl.GeneTree_content.e<CURR_RELEASE_NUMBER>.txt.gz.MD5SUM
```

☑ Ancestral alleles (for the Variation team)

Click here to expand...

> ⓘ Ancestral alleles are computed from EPO data. These can be symlinked from /warehouse/ens_ftp_arch_03/release-XX if no new alignments have been run

Repeat the process for each species of interest and perform a checksum when all data is tarred

```
perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/ancestral_sequences/get_ancestra
l_sequence.pl --conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl
--compara_url mysql://ensro@compara5/sf5_epo_8primates_77 --species
homo_sapiens
dirname=homo_sapiens*
cd $dirname
perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/ancestral_sequences/get_stats.pl
> summary.txt
cd ..
tar -cvzf ${dirname}.tar.gz $dirname
md5sum *.tar.gz > MD5SUM # when archives for all required species are complete
```

☑ Let the production team know that the dumps are ready in their common /nfs/ensembl/ensembl/ftp_ensembl/release-XX/ location.

Patch GRCh37 Databases

For the next 10 years or so, we'll maintain a special archive site for the GRCh37 assembly: http://grch37.ensembl.org

This page explains the procedure to update the Compara databases on it. This has to be done every release and is coordinated with the other teams.

Connection details

The web-team has a Confluence page with all the gory details. To summarize:

- Public MySQL server: ensembldb.ensembl.org:3337
- Internal MySQL server (where we need to update the databases): ens-staging-grch37 in e!84, but it may be different in future releases

Case 1: no data updates

- Patch the ensembl_compara_${CURR_ENSEMBL_RELEASE} database to the newest Compara schema

```
$ENSEMBL_CVS_ROOT_DIR/ensembl/misc-scripts/schema_patcher.pl
--host=ens-staging-grch37 --user=ensadmin --pass=${ENSADMIN_PSW}
--database=ensembl_compara_${CURR_ENSEMBL_RELEASE}
```

- Patch the ensembl_ancestral_${CURR_ENSEMBL_RELEASE} database to the newest Core schema

```
$ENSEMBL_CVS_ROOT_DIR/ensembl/misc-scripts/schema_patcher.pl
--host=ens-staging-grch37 --user=ensadmin --pass=${ENSADMIN_PSW}
--database=ensembl_ancestral_${CURR_ENSEMBL_RELEASE}
```

- That's it, you're ready to hand-over

Case 2: data updates

Data-updates have to be coordinated with the other teams as they can impact them (BioMart, dumps, etc). The next data-update is scheduled to be for e79, but Compara did not participate in it.

Compara updates:

- Upgrade of the protein trees / homologies, to merge with TreeFam. This will impact BioMart, etc.
- Fix the CAFE data ? Actually done at a data update, but I can't remember which one

Final things

☑ Create a word document and a pdf dump of this document

⌄ Click here for details

In the top-right menu of this Confluence page, choose "Tools -> Export to PDF" and "Tools -> Export to Word".

Put these files into $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/production

git commit and push