Release coordination documentation

Declaration of Intentions

- [x] Set up a web page

  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

- [x] 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
- [x] Submit the declaration of intentions on the http://admin.ensembl.org/index.html website.

Environment variables

- [x] 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
- [x] Define $ENSADMIN_PSW
  The password for the mysql 'ensadmin' user also needed for many scripts

Update your cvs checkout

Ensure you have new checkouts of at least the following repositories

- [x] ensembl-compara
- [x] ensembl
- [x] ensembl-hive
- [x] ensembl-analysis
- [x] ensj-healthcheck

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 (e72) master database is sf5_ensembl_compara_master on compara1

- [x] 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 mysqldump -u ensro -h ens-livemirror -P3306 --extended-insert
    --compress --delayed-insert ncbi_taxonomy ncbi_taxa_node ncbi_taxa_name |
    mysql -u ensadmin -p${ENSADMIN_PSW} \
     -h compara1 sf5_ensembl_compara_master
    ```

    - Times

      rel.64: 45 sec

      rel.65: 47 sec

      rel.66: 47 sec

      rel.67: 30 sec

      rel.68: 30 sec

      rel.69: 35 sec

      rel.70 32 sec

      rel.71 34 sec

      rel.72 36 sec
- [x] Check new extant taxon names

  Each new species must have a 'ensembl alias name' tag in the $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/taxonomy/ensembl_aliases.sql file. It should match the "web_name" used by the production team in the "species" table of their "ensembl_production" database on staging1. This may have already been added by Ensembl Production.
- [x]

- [ ] Update the ancestral taxon names

  Each new 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 alias name': a "simple English" description of the taxon
  - 'ensembl timetree mya': the age of the taxon. It can be obtained from the TimeTree database (http://www.timetree.org)

  For any new species, update the file $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/taxonomy/ensembl_aliases.sql to add the two new tags

- [x] 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

```
mysql -u ensadmin -p${ENSADMIN_PSW} -h compara1 sf5_ensembl_compara_master
<
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/taxonomy/ensembl_aliases.sql
```

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.

- [x] Check space on http://www.ebi.ac.uk/~mp/compara_servers_disk_load.html
- [x] Ask compara team members to tidy up any unwanted databases and inform them of the intention to delete the last but one release
- [x] Delete ???_ensembl_compara_xx
- [x] Delete ???_ensembl_compara_ancestral_xx

Configuration file

- [x] Update production_reg_conf.pl and check back in to cvs

  ▾ 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 (since 66) is to have the release database in compara3.

Update schema version

- [x]

- [x] Update the schema_version in the master database

```
update meta set meta_value = XX where meta_key = 'schema_version';
```

Update table.sql and create patch files

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

- [x] Create a patch file for the schema_version
- [x] Check if any other patch files need creating by looking at the Declaration of Intentions and checking with other compara team members
- [x] Update the schema_version in table.sql
- [x] Delete the previous patch INSERT statements from table.sql
- [x] Add an INSERT statement for the new schema_version in table.sql and for any other new patches

Check the patch files

- [x] Create schema from previous database
  - Click here for details

mysqldump

```
mysqldump --defaults-group-suffix=_compara3 --no-data --skip-add-drop-table
kb3_ensembl_compara_71 | sed 's/AUTO_INCREMENT=[0-9]*\b//' >old_schema.sql
```

- [x] Create a new database from the current schema
  - Click here for details

create new database

```
mysql --defaults-group-suffix=_compara2  -e "create database
kb3_current_schema_test"
mysql --defaults-group-suffix=_compara2 kb3_current_schema_test <
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/sql/table.sql
mysqldump --defaults-group-suffix=_compara2 --no-data --skip-add-drop-table
kb3_current_schema_test | sed 's/AUTO_INCREMENT=[0-9]*\b//' >new_schema.sql
```

- [x] Check the differences and create any missing patch files
  - Click here for details

Check for differences

```
sdiff -b old_schema.sql new_schema.sql | less
```

- [x] Generate an empty database from the old schema, populate the meta table, apply the patches, dump it, and check that you get the new schema.

Click for details

We need to populate the meta table from the previous release to allow the schema_patcher.pl script to work. The final sdiff will display the peptide_align_feature_xxx tables that are in the previous release. These can be ignored.

Check patches

```
mysql --defaults-group-suffix=_compara2 -e 'create database
kb3_schema_patch_test'
mysql --defaults-group-suffix=_compara2 kb3_schema_patch_test <
old_schema.sql
mysqldump --defaults-group-suffix=_compara3 kb3_ensembl_compara_71 meta |
mysql --defaults-group-suffix=_compara2 kb3_schema_patch_test
$ENSEMBL_CVS_ROOT_DIR/ensembl/misc-scripts/schema_patcher.pl --host
compara2 --port 3306 --user ensadmin --pass $ENSADMIN_PSW --database
kb3_schema_patch_test --type compara --from 71 --release 72 --verbose
mysqldump --defaults-group-suffix=_compara2 --no-data --skip-add-drop-table
kb3_schema_patch_test | sed 's/AUTO_INCREMENT=[0-9]*\b//'
>patched_old_schema.sql
sdiff -w 200 -bs patched_old_schema.sql new_schema.sql | less
```

- [x] cvs commit table.sql and any patch files

After Handover

Run CheckTaxon healthcheck

- [x] Run the CheckTaxon healthcheck early to find any discrepancies between the ncbi_taxon_name table and the core databases

Run healthcheck

```
./run-healthcheck.sh -d sf5_ensembl_compara_master -type compara -d2 .+_core_72_.+
CheckTaxon
```

We can use the compara master database as the source, before the creation of the release database.

(see xxxx on how to set run the healthchecks)

Add new genome_db to compara master database

The current master database (e72) is called sf5_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.

- [x] Add genome_db

Click here for details

This may have already been done if the dna guys started early, please check.

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" --collection ensembl
```

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. You can check the progress by counting dnaFrag entries in the master database:

```
select count(*) from dnafrag;
```

If it is a new genebuild and the assembly hasn't changed, you can just edit the entry in master (genome_db table) introducing the new genebuild from meta.genebuild.start_date in the core database.

```
update genome_db set genebuild ="2011-07-FlyBase" where genome_db_id = 105;
```

- [x] 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/

eg 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 these need to be deleted from the master before the new and replacement patches are added. 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_patche
s.pl -new_core
"mysql://ensro@ens-staging1:3306/homo_sapiens_core_72_37?group=core&species
=homo_sapiens" -prev_core
"mysql://ensro@ens-livemirror:3306/homo_sapiens_core_71_37?group=core&speci
es=homo_sapiens"
```

The output looks like this (format: name, seq_region_id, date)

Sample ouptut

```
NEW patches
  HG29_PATCH 1001061122 2013-02-18 14:16:55
  HG1592_PATCH 1001061114 2013-02-18 14:16:55
  HG385_PATCH 1001061116 2013-02-18 14:16:55
  HSCHR6_2_CTG5 1001061112 2013-02-18 14:16:55
  HG1079_PATCH 1001061118 2013-02-18 14:16:55
CHANGED patches
  HG1436_HG1432_PATCH new=1001061124 2013-02-18 14:16:55
prev=1000859885 2012-10-08 16:48:36
  HG1292_PATCH new=1001061108 2013-02-18 14:16:55       prev=1000759258
2012-04-27 12:12:07
  HSCHR22_1_CTG1 new=1001061120 2013-02-18 14:16:55     prev=1000057052
2010-09-07 14:22:38
  HG1287_PATCH new=1001061110 2013-02-18 14:16:55       prev=1000859831
2012-10-08 16:48:36
DELETED patches
Patches to delete:
("HG1436_HG1432_PATCH","HG1292_PATCH","HSCHR22_1_CTG1","HG1287_PATCH")
```

In this case, there are 5 NEW patches, 4 CHANGED patches and no DELETED patches. Any CHANGED or DELETED patches must be deleted from the master before importing the new patch set.

To add extra non-reference patches to an existing assembly, you need the -force option to just add those dnafrags which aren't already in the database.

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
```

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

Pairwise method_link_species_set

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

- [ ] Add synteny method_link_species_set entries

Synteny method_link_species_set

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

- [ ] Add homology method_link_species_set_entries

▾ Click here for details

Choose a temp. directory where the output will be generated:

Choose temp directory

```
export MLSS_DIR="/tmp/mlss_creation"
mkdir $MLSS_DIR
```

Run the loading script several times:

--pw stands for all pairwised genome_db_ids in the list provided

--sg stands for keep genome_db_id in the list alone (singleton)

Protein method_link_species_set

```
# orthologues
echo -e "201\n" | perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl --f \
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.
pl \
--pw --collection ensembl
1>$MLSS_DIR/create_mlss.ENSEMBL_ORTHOLOGUES.201.out
2>$MLSS_DIR/create_mlss.ENSEMBL_ORTHOLOGUES.201.err

# paralogues btw
echo -e "202\n" | perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl --f \
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.
pl \
--pw --collection ensembl
1>$MLSS_DIR/create_mlss.ENSEMBL_PARALOGUES.btw.202.out
2>$MLSS_DIR/create_mlss.ENSEMBL_PARALOGUES.btw.202.err

# paralogues wth
echo -e "202\n" | perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl --f \
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.
pl \
--sg --collection ensembl
1>$MLSS_DIR/create_mlss.ENSEMBL_PARALOGUES.wth.202.out
2>$MLSS_DIR/create_mlss.ENSEMBL_PARALOGUES.wth.202.err

# proteintrees
echo -e "401\n" | perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl --f \
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.
pl \
--name "protein trees" --collection ensembl
1>$MLSS_DIR/create_mlss.PROTEIN_TREES.401.out
2>$MLSS_DIR/create_mlss.PROTEIN_TREES.401.err

# nctrees
echo -e "402\n" | perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl --f \
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.
pl \
--name "nc trees" --collection ensembl
1>$MLSS_DIR/create_mlss.NC_TREES.402.out
2>$MLSS_DIR/create_mlss.NC_TREES.402.err

# families
echo -e "301\n" | perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/create_mlss.pl --f \
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.
pl \
--name "families" --collection ensembl
1>$MLSS_DIR/create_mlss.FAMILY.301.out
2>$MLSS_DIR/create_mlss.FAMILY.301.err
```

If output/error files are ok, remove them all:

Remove files

```
rm -rf $MLSS_DIR
```

Otherwise make yourself a nice cup of tea and then *PANIC*

Update species_set_tags in compara master database

- [x] Update the species_set_tags

Click here for details

Run the script: update_species_sets.pl

Update species_set_tags

```
perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/update_species_sets.
pl --conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.
pl --dbname compara_master > update_species_sets.out 2>
update_species_sets.err
```

Add/Update any dna tags if necessary (they may be already present). There should be a tag for each of the multiple alignments. The create_mlss.pl script (see above) adds this using the --species_set_name option. Check to make sure this has happened.

Add new species to phylogenetic tree

- [x] Add new species to phylogenetic tree

Click here for details

The easiest way to use this is to use the phylowidget.

From the EnsEMBL home page:

View full list of all Ensembl species

Species tree (Requires Java)

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) and branch length

The tree should appear in the Toolbox but if not, then save the tree

Copy the new tree into

$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/species_tree_blength.nh

cvs commit

rel.65 -- Added Gadus morhua

rel.66 -- Added Latimeria chalumnae (Coelacanth)

Also included 'ensembl timetree mya' in ncbi_taxa_name for

taxon_id: 8287 -- value: 414.9 in the master and final database.

rel.69 added Xiphophorus_maculatus and Mustela_putorius_furo

Final checks to compara master database

- [x] Check if any new species have been postponed
  If a new species is postponed for this release, check that no entries (genome_db, dnafrags, etc) were added to the master database. If they were, you can simply switch the assembly_default value in the genome_db table.
  [1] for species making it / used in the pipeline
  [0] for species not making it / or old assemblies
- [x] Drop method_link_species_set entries for alignments which did not make it.
  - Click here for details
    Check with other members of compara. Remove redundant entries.

mlss

```
SELECT ml.* FROM method_link ml LEFT JOIN method_link_species_set mlss ON
ml.method_link_id=mlss.method_link_id WHERE mlss.method_link_id IS NULL;
```

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.

- [x] Create new database
  - Click here for details

Create database

```
mysql --defaults-group-suffix=_compara3 -e "CREATE DATABASE
kb3_ensembl_compara_72"
mysql --defaults-group-suffix=_compara3 kb3_ensembl_compara_72 <
$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_databas
e.pl --reg-conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.
pl \
--master compara_master --old compara_prev --new compara_curr --intentions
> populate_new_database.intentions
```

This takes about a minute and produces a long list.

If you believe some of the entries should NOT be copied, you should manually add 'skip_mlss' and 'skip_ss' entries into master database meta table.

NB There are cases where the mlss does not change but the underlying data does eg 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

```
time
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/populate_new_databas
e.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)

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_databas
e.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

For the UPDATED patches, you need to use the original dnafrag_ids.

delete patches

```
SELECT count(*) FROM genomic_align ga1, genomic_align ga2,
genomic_align_block gab WHERE ga1.dnafrag_id in
(13768162,13728785,13768189,12967785) AND 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;
99906

DELETE ga1, ga2, gab FROM genomic_align ga1, genomic_align ga2,
genomic_align_block gab WHERE ga1.dnafrag_id in
(13768162,13728785,13768189,12967785) AND 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;
Query OK, 299718 rows affected (34.34 sec)
99906*3=299718
```

☑ 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
time ./run-healthcheck.sh -d kb3_ensembl_compara_72 -type compara -d2
.+_core_72_.+ compara_external_foreign_keys
```

and correct any problems, if any

Merge DNA data

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

☑ Pairwise alignments: LASTZ_NET, TRANSLATED_BLAT_NET

▾ 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.

Example:

copy_data

```
Using the mlss_id:
bsub -q yesterday -R "select[mem>5000] rusage[mem=5000]" -M5000000 \
    -I time
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl \
    --from_url mysql://ensadmin:${ENSADMIN_PSW}@host/production_db \
    --to_url mysql://ensadmin:${ENSADMIN_PSW}@host/release_db --mlss 268
-re_enable 0

Using method_link_id:
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl \
    --from_url
mysql://ensadmin:${ENSADMIN_PSW}@compara2/kb3_cint_csav_lastz_72\
    --to_url
mysql://ensadmin:${ENSADMIN_PSW}@compara3/kb3_ensembl_compara_72
--method_link LASTZ_NET --re_enable 0
```

- [x] Pairwise alignments: non-reference patches for the high coverage LASTZ_NET alignments

▼ Click here for details

Make sure that all the pairwise alignment are in the database before loading these. First find all the relevant method_link_species_set ids for each patch set (currently we have patches for human, mouse and mouse patches against human):

Example:

find list of mlss_ids

```
export AFFECTED_MLSSS=`mysql --defaults-group-suffix=_compara2
kb3_hsap_lastz_hap_72 -N -e "select
group_concat(distinct(method_link_species_set_id) SEPARATOR ' ') from
genomic_align join dnafrag using (dnafrag_id) where genome_db_id=90 and
method_link_species_set_id<1000"`

export AFFECTED_MLSSS=`mysql --defaults-group-suffix=_compara2
kb3_mmus_lastz_hap_70 -N -e "select
group_concat(distinct(method_link_species_set_id) SEPARATOR ' ') from
genomic_align join dnafrag using (dnafrag_id) where genome_db_id=134 and
method_link_species_set_id<1000"`
```

Use the --merge and --patch_merge option of copy_data to add the alignments.The patch_merge option ensures merging will work even if the genomic_align_block_ids or genomic_align_ids are not sequential.

copy_data --merge

```
time for i in $AFFECTED_MLSSS; do echo $i;
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl
--from_url mysql://ensro@compara2/kb3_hsap_lastz_hap_72 --to_url
mysql://ensadmin:${ENSADMIN_PSW}@compara3/kb3_ensembl_compara_72 --mlss $i
--merge --patch_merge; done

time for i in $AFFECTED_MLSSS; do echo $i;
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl
--from_url mysql://ensro@compara2/kb3_mmus_lastz_hap_70 --to_url
mysql://ensadmin:${ENSADMIN_PSW}@compara3/sf5_ensembl_compara_70 --mlss $i
--merge --patch_merge; done
```

- [ ] Multiple alignments: PECAN, EPO, EPO_LOW_COVERAGE

  ▾ 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]" -M5000000 \
    -I time
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl \
    --from_url
mysql://ensadmin:${ENSADMIN_PSW}@compara2/sf5_epo_low_8way_fish_71 \
    --to_url
mysql://ensadmin:${ENSADMIN_PSW}@compara3/kb3_ensembl_compara_71
--method_link EPO_LOW_COVERAGE -re_enable 0

    bsub -q yesterday -R "select[mem>5000] rusage[mem=5000]" -M5000000 \
    -I time
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl \
    --from_url
mysql://ensadmin:${ENSADMIN_PSW}@compara2/sf5_epo_low_8way_fish_71 \
    --to_url
mysql://ensadmin:${ENSADMIN_PSW}@compara3/kb3_ensembl_compara_71
--method_link GERP_CONSTRAINED_ELEMENT -re_enable 0

    bsub -q yesterday -R "select[mem>5000] rusage[mem=5000]" -M5000000 \
    -I time
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/copy_data.pl \
    --from_url
mysql://ensadmin:${ENSADMIN_PSW}@compara2/sf5_epo_low_8way_fish_71 \
    --to_url
mysql://ensadmin:${ENSADMIN_PSW}@compara3/kb3_ensembl_compara_71
--method_link GERP_CONSERVATION_SCORE -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)

- [x] 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

```
mysqlshow -u ensro -h compara3 --keys kb3_ensembl_compara_72
genomic_align_block %
mysqlshow -u ensro -h compara3 --keys kb3_ensembl_compara_72 genomic_align
%
mysqlshow -u ensro -h compara3 --keys kb3_ensembl_compara_72
genomic_align_tree %
mysqlshow -u ensro -h compara3 --keys kb3_ensembl_compara_72
conservation_score %
mysqlshow -u ensro -h compara3 --keys kb3_ensembl_compara_72
constrained_element %
```

- [ ] 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.

Load the synteny data by running the $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/synteny/LoadSytenyData.pl script. This requires a synteny file. The location of this should be on the Release plans.

Example

load synteny data

```
perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/synteny/LoadSyntenyData.pl
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.
pl \
    --dbname compara_curr -ref "Homo sapiens" -nonref "Callithrix jacchus"
-mlss_id 10052 \

/lustre/scratch101/ensembl/kb3/scratch/hive/release_64/kb3_hsap_cjac_synten
y_64/synteny/all.100000.100000.BuildSynteny
```

☑ 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 list of alignments with ancestral alignments are:

3-way birds

5-way fish

6-way primates

13-way eutherian mammals

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

init_pipeline

```
init_pipeline.pl AncestralMerge_conf.pm -password $ENSADMIN_PSW
```

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:

Check msg table

```
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, drop hive-specific tables:

drop hive tables

```
CALL drop_hive_tables;
```

Make sure all tables are myISAM.

▾ Click here to see times

rel.67 20min

rel.71 20min

Merge GeneTrees+Families+NCTrees

☑

Check the StableID mapping and the TreeFam mapping steps on the ProteinTrees have been done

- [x] Merge protein databases together

Click here for details

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

It has 6 sections for connecting to databases where you will have to change the names of the databases and possibly their locations:

pipeline_db - is your intermediate target (all protein side pipelines merged together)

master_db - is the main compara master

prevrel_db - should point to the previous release database

genetrees_db - should point to the current GeneTrees pipeline database

families_db - should point to the current Families pipeline database

nctrees_db - should point to the current ncRNAtrees pipeline database

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

init_pipeline

```
init_pipeline.pl MergeHomologySideTogether_conf.pm
```

Then run both -sync and -loop variations of the beekeeper.pl command suggested by init_pipeline.pl .

This pipeline will create a database with protein side pipeline databases merged together.

Click here for times

2.7h to execute for release 63

9.6h to execute for release 65

~9h to execute for release 66

~3h to execute for release 71

- [x] Project genes as homologies for human and mouse

Click here for details

First update the AUTO_INCREMENT of the following tables to make is easier to spot new entries

alter table

```
ALTER TABLE member   AUTO_INCREMENT=300000001;
ALTER TABLE sequence AUTO_INCREMENT=300000001;
ALTER TABLE homology AUTO_INCREMENT=300000001;
```

Run the script on human and mouse.

NB: The underscore in the species name is necessary to match the genome_db name.

There is a "-no_store 1" flag to avoid writing to the compara database.

convert_patch_to_compara_homologies

```
perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/convert_patch_to_com
para_homologies.pl \
    -reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.
pl \
    -comp_alias compara_homology_merged -species homo_sapiens

perl
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/convert_patch_to_com
para_homologies.pl \
    -reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.
pl \
    -comp_alias compara_homology_merged -species mus_musculus
```

▼ Click here for previous results

For e70 - human

#new compara entries:

#578 ref genes

#1028 projected genes

#1459 new homologies

#For e71 mouse

new compara entries:

10 ref genes

22 projected genes

65 new homologies

☑ Updating member.display_label fields

This step has to happen ASAP, but AFTER the Core name projections have been done. And before you analyze/optimize the tables.

▼ Click here for details

The order of proceedings:

1. Matthieu runs the Homology pipeline

2. The homology database is given to Production, who uses it to run name projections on Core databases (display_labels and gene_descriptions change in Core databases)

3. We use the information in Core databases (derived from Homology, i.e. Compara) to fix the display_labels and gene_descriptions for Compara

!MAKE SURE YOU ARE IN SYNC WITH THE REST OF THE WORLD!

☑ Ensure your registry is correct

▾ Click here for details

The registry file which is used should point to the server where all updated Xref projections are located. This will mean staging servers 1 and 2. To load the data from these two servers you can use the

Bio::EnsEMBL::Registry->load_registry_from_multiple_dbs() call on both servers.

Check the file

$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_conf.pl

☑ Run the script

▾ Click here for details

NB: Please note that your --compara_db will have to be set to the database where you want to perform the member update;

It can be the intermediate homology database if it hasn't been merged into the release database yet, or the release database itself.

Run the script (once for display_label, and once for description):

MemberDisplayLabelUpdater

```
bsub -q yesterday -R "select[mem>5000] rusage[mem=5000]" -M5000000 -I
\
    time standaloneJob.pl
Bio::EnsEMBL::Compara::RunnableDB::MemberDisplayLabelUpdater
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_
conf.pl --compara_db compara_homology_merged --debug 1 --replace 1

bsub -q yesterday -R "select[mem>5000] rusage[mem=5000]" -M5000000 -I
\
    time standaloneJob.pl
Bio::EnsEMBL::Compara::RunnableDB::MemberDisplayLabelUpdater
--reg_conf
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/production_reg_
conf.pl --compara_db compara_homology_merged --debug 1 --replace 1
--mode description
```

You can use "--compara_db compara_curr" if it is run on the final merged db,
or "--compara_db mysql://ensadmin:${ENSADMIN_PSW}@compara3:3306/lg4_compara_homology_merged_64" if you do not have the databases in the Registry

▾ Click here for times

rel59 real 9m44.754s

rel60 real 10m23.002s

rel62 real 11m

rel63 real 11m

rel.64 12m

rel.65 24m

rel.66 5m

rel.67 5m

rel.69 10m

rel.70 13min

rel.71 20min

rel. 72 13m and 6m

- [x] Merge protein databases into release database

  - Click here for details

    Run another mini-pipeline to merge the protein side into the release database.

    Go to (or stay in) $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/modules/Bio/EnsEMBL/Compara/PipeConfig and open the PipeConfig file MergeHomologyIntoRelease_conf.pm

    Usually, only the release version and the server names have to be changed. You can still double-check the 3 sections:

    pipeline_db - is a hive database that is only used for job tracking - it may/should be removed right after the pipeline is done

    merged_homology_db - is the result of the previous step (protein side databases merged together)

    rel_db - is the main release database

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

    init_pipeline

    ```
    init_pipeline.pl MergeHomologyIntoRelease_conf.pm -password $ENSADMIN_PSW
    -hive_driver sqlite
    ```

    Then run both -sync and -loop variations of the beekeeper.pl command suggested by init_pipeline.pl .

Click here for times

2.6h for rel65

~1.0h for rel66

2.5h for rel 72

- [x] Populate the member_production_counts table

  Click here for details

  Populate the member_production_counts table using the script at $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/production/populate_member_production_counts_table.sql

  member_production_counts

  ```
  time mysql -hcompara3 -uensadmin -p$ENSADMIN_PSW kb3_ensembl_compara_72 < $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/production/populate_member_production_counts_table.sql
  ```

  Click here for times

  Reckon about an hour

- [ ] Clean up merged databases

  Click here for details

  After you are happy about the result of both mergers you can drop both "compara_homology_merged" and "compara_full_merge" databases.

- [x] CVS commit the changes to the PipeConfig files that you have made.

Final database checks

- [x] 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

  ```
  SELECT ml.* FROM method_link ml LEFT JOIN method_link_species_set mlss ON ml.method_link_id=mlss.method_link_id WHERE mlss.method_link_id IS NULL;
  DELETE ml FROM method_link ml LEFT JOIN method_link_species_set mlss ON ml.method_link_id=mlss.method_link_id WHERE mlss.method_link_id IS NULL;
  ```

- [x] 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

- [x] Update the code

  Click here for details

  The healthchecks are written in java and need to be recompiled after a CVS update.

  compile healthchecks

  ```
  cd $ENSEMBL_CVS_ROOT_DIR/ensj-healthcheck
  find build -name '*.class' -delete
  export JAVA_HOME=/software/jdk1.6.0_01
  ant
  ```

  Need something here about the database.properties file.....

- [x] Run the healthchecks for ancestral database

  Click here for details

  time $ENSEMBL_CVS_ROOT_DIR/ensj-healthcheck/run-healthcheck.sh -d kb3_ensembl_ancestral_71 compara-ancestral

  Click here for times

  # rel.62: 4sec, all successful

  # rel.63: 14sec, all successful after analyzing 3 tables

  # rel.65: 13sec, all successful after analyzing the tables.

  # rel.67: 8sec, all successful

  # rel.68: 8sec, all successful

  # rel.71: 10sec all successful apart from UTR.java (ignore)

- [x] Update the Meta table

  Click here for details

  You can use the corresponding healthcheck with the -repair option:

  update meta table

  ```
  $ENSEMBL_CVS_ROOT_DIR/ensj-healthcheck/run-healthcheck.sh -d
  kb3_ensembl_compara_71 -type compara -repair Meta
  ```

- [x] Update the max_alignment_length.

  Click here for details

  Update the max_alignment_length. You can use the corresponding healthcheck with the -repair option:

update max_alignment_length

```
$ENSEMBL_CVS_ROOT_DIR/ensj-healthcheck/run-healthcheck.sh -d
kb3_ensembl_compara_71 -type compara -repair MLSSTagMaxAlign
```

- [ ] Update the alignment mlss_id of the conservation score

  - Click here for details

update conservation score mlss_id

```
$ENSEMBL_CVS_ROOT_DIR/ensj-healthcheck/run-healthcheck.sh -d
kb3_ensembl_compara_71 -type compara -repair MLSSTagGERPMSA
```

- [x] Run the compara_external_foreign_keys healthcheck

  - Click here for details

compara_external_foreign_keys

```
time $ENSEMBL_CVS_ROOT_DIR/ensj-healthcheck/run-healthcheck.sh -d
kb3_ensembl_compara_72 -type compara -d2 .+_core_72_.+
compara_external_foreign_keys
```

  - Click here for previous results

rel.56 everything passed apart from CheckTaxon - according to Javier in this particular case it was not a problem

rel.pre57 it took 20 minutes (all passed).

rel.57 it took ?? minutes ('genbank common name' for 4 species had to be copied from their 'ensembl common name' in ncbi_taxa_name table)

rel.58: 22m

rel.61: 32m, 1 failure

rel.63: 23m, 1 failure (taeniopygia_guttata_core_63_1: common_name::zebra finch is not in lg4_ensembl_compara_63)

rel.65: 25m, 1 failure (taeniopygia_guttata_core_65_1: common_name::zebra finch is not in mp12_ensembl_compara_65)

rel.66: 12m, 1 failure (taeniopygia_guttata_core_66_1: common_name::zebra finch is not in mp12_ensembl_compara_66)

rel.71. 1 failure (ncbi_taxa_name: Lepidion inosimae: common_name "Morid cod" and "morid cod", this was agreed to be OK in rel.70).

rel.72 12m

- [x] Run the compara_genomic healthcheck
  - Click here for details

    compara_genomic

    ```
    time $ENSEMBL_CVS_ROOT_DIR/ensj-healthcheck/run-healthcheck.sh -d
    sf5_ensembl_compara_69 -type compara compara_genomic
    ```

    - Click here for previous results

      rel.58: 47m, 7 failures

      rel.61: 51m, 3 failures

      rel.63: 84m, (2.5 errors that are "ok")

      rel.65: 75m, (5 errors that are "ok")

      rel.66: 23m, (4 errors that are "ok")

      rel.72: 57m
- [x] Run the compara_homology healthcheck
  - Click here for details

    compara_homology

    ```
    time $ENSEMBL_CVS_ROOT_DIR/ensj-healthcheck/run-healthcheck.sh -d
    sf5_ensembl_compara_69 -type compara compara_homology
    ```

    - Click here for previous results

      rel.58: 14m, 5 failures

      rel.61: 30m, 3 failures

      rel.62: 3m, 1 failure (CheckSpeciesSetTag, ok?)

      rel.63: 52m, success (after some fixing, of course)

      rel.65: 48m, 3 errors

      rel.66: 18m

      rel.72: 1h6m
- [x] Check for empty tables
  - Click here for details

    Until a health check exists, it is best to check that there are no empty tables in the release db. A comparison between the current and previous release database WRT table sizes and/or row-numbers would be wise, check for missing tables / new tables and large, unexpected row number changes.
- [ ] CVS commit database.properties

Click here for details

Hide any password / sensitive data from the configuration file database.properties (for example, using the string "${ENSADMIN_PSW}") and commit it back to ensembl-compara/scripts/pipeline/

Test web server

- [x] 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/

Run ANALYZE_TABLE and OPTIMIZE TABLE

This is required for the CopyDbOverServer script to work properly.

- [x] Run ANALYZE_TABLE on compara and ancestral databases

  Click here for details

  analyze table

  ```
  time mysqlcheck --analyze --verbose --host=compara3 --port=3306
  --user=ensadmin --password=$ENSADMIN_PSW --databases kb3_ensembl_compara_71
  time mysqlcheck --analyze --verbose --host=compara3 --port=3306
  --user=ensadmin --password=$ENSADMIN_PSW --databases
  kb3_ensembl_ancestral_71
  ```

  Click here for times

  rel.56 12min

  rel.pre57 30+105min

  rel.57 9+4+5min

  rel.58 3min

  rel.62 6min

  rel.63 25m

  rel.64 16m

  rel.65 21m

  rel.66 9m

  rel.71 7m

  rel.72 10m49.630s

- [x] Run OPTIMIZE TABLE on compara and ancestral databases

  Click here for details

optimize table

```
time mysqlcheck --optimize --verbose --host=compara3 --port=3306
--user=ensadmin --password=$ENSADMIN_PSW --databases kb3_ensembl_compara_71
time mysqlcheck --optimize --verbose --host=compara3 --port=3306
--user=ensadmin --password=$ENSADMIN_PSW --databases
kb3_ensembl_ancestral_71
```

Click here for times

rel.56 2.5 hours

rel.pre57 : took several iterations (not all tables were MyISAM initially), last one 132min.

rel.57 2+1.6 hours

rel.62 32min

rel.63 61m

rel.64 1.5h

rel.65 3h

rel.66 2h

rel.71 2h

rel.72 104m53.572s

Copy databases to staging servers

- [x] Logon to ens-staging1 and create a file containing the copying options

Click here for details

First, ssh into the DESTINATION machine and switch to the bash shell

NB: ask for the password for mysqlens well in advance - there may be no-one around you at the right moment!

ssh

```
ssh mysqlens@ens-staging1
bash
```

Create a file that will contain one line for each database with the source/destination parameters, like this:

copy_options

```
cat <<EOF >/tmp/kb3_ensembl_compara_71.copy_options
#from_host      from_port   from_dbname                to_host
to_port     to_dbname
#
compara3        3306        kb3_ensembl_compara_72     ens-staging1
3306        ensembl_compara_72
compara3        3306        kb3_ensembl_ancestral_72   ens-staging1
3306        ensembl_ancestral_72
EOF
```

Set the password into the environment variable:

export ENSADMIN_PSW='...'

- [x] Copy the databases to ens-staging1

Click here for details

You should check whether there is enough space on the disk before starting the copy.

CopyDBoverServer

```
time perl ~kb3/src/ensembl_main/ensembl/misc-scripts/CopyDBoverServer.pl
-pass $ENSADMIN_PSW \
    -noflush /tmp/kb3_ensembl_compara_72.copy_options >
/tmp/kb3_ensembl_compara_72.copy.err 2>&1
```

Click here for times

copying of rel_56 took 2 hours (SUCCESSFUL for both databases - you should check the output file)

copying of ensembl_compara_pre57 took 2 hours (SUCCESSFUL)

copying of ensembl_compara_57 took 2 hours (SUCCESSFUL)

copying of ensembl_ancestral_57 took 20 minutes (only SUCCESSFUL after analyzing/optimizing)

copying of ensembl_compara_58 and ensembl_ancestral_58 together took 1:30h (SUCCESSFUL)

copying of ensembl_compara_62 and ensembl_ancestral_62 took 2h38

copying of ensembl_compara_63 and ensembl_ancestral_63 took 2h

copying of ensembl_compara_64 and ensembl_ancestral_64 took 2h15m

rel65 2h51m

rel66 1h12m to copy over ensembl_compara_66

0h11m to copy over ensembl_ancestral_66

rel68 90m39.996s

r70 ens-staging2 (1h29m4s)

rel71: 1h39m

rel72 1h40m50s

- [x] Logon to ens-staging2 and create a file containing the copying options

  ▼ Click here for details

  First, ssh into the DESTINATION machine and switch to the bash shell

  # NB: ask for the password for mysqlens well in advance - there may be no-one around you at the right moment!

  ssh

  ```
  ssh mysqlens@ens-staging2
  bash
  ```

  Create a file that will contain one line for each database with the source/destination parameters, like this:

  copy_options

  ```
  cat <<EOF >/tmp/kb3_ensembl_compara_71.copy_options
  #from_host      from_port   from_dbname                to_host
  to_port     to_dbname
  #
  compara3        3306        kb3_ensembl_compara_72     ens-staging2
  3306        ensembl_compara_72
  compara3        3306        kb3_ensembl_ancestral_72   ens-staging2
  3306        ensembl_ancestral_72
  EOF
  ```

  Set the password into the environment variable:

  export ENSADMIN_PSW='...'

- [x] Copy the databases to ens-staging2

  ▼ Click here for details

  CopyDBoverServer

  ```
  time perl ~kb3/src/ensembl_main/ensembl/misc-scripts/CopyDBoverServer.pl
  -pass $ENSADMIN_PSW \
      -noflush /tmp/kb3_ensembl_compara_72.copy_options >
  /tmp/kb3_ensembl_compara_72.copy.err 2>&1
  ```

  ▼ Click here for times

  copying of ensembl_compara_58 took 1:15h

copying of ensembl_compara_62 took 1:34h

copying of ensembl_compara_63 took 3:44h

copying of ensembl_compara_64 took 3h51m

rel65 3h28m

rel66 ~2h

rel68 104m11.046s

rel 72 1h40m24s

Final handover of databases

- [x] Send an email to ensembl-production to announce the handover the databases
- [ ] Update the Declaration of Intentions http://admin.ensembl.org/index.htmlto indicate what has been handed over and what didn't make it and has been postponed

Update documentation and diagrams

- [x] Update pipeline diagrams

  ▾ Click here for details

  Update the pipeline diagrams in the docs directory

  pipeline diagrams

```
cd $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/pipelines/diagrams
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-hive/scripts/generate_graph.pl -url
mysql://ensro@compara3/mp12_compara_homology_72 -output ProteinTrees.png
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-hive/scripts/generate_graph.pl -url
mysql://ensro@compara2/mp12_compara_nctrees_72 -output ncRNAtrees.png
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-hive/scripts/generate_graph.pl -url
mysql://ensro@compara4/lg4_compara_families_7204 -output Families.png
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-hive/scripts/generate_graph.pl -url
mysql://ensro@compara3/kb3_pecan_20way_71 -output MercatorPecan.png
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-hive/scripts/generate_graph.pl -url
mysql://ensro@compara4/sf5_epo_35way_68 -output EpoLowCoverage.png
perl $ENSEMBL_CVS_ROOT_DIR/ensembl-hive/scripts/generate_graph.pl -url
mysql://ensro@compara4/sf5_compara_epo_13way_69 -output Epo.png
```

  CVS commit these files

- [x] Update the schema documentation

  ▾ Click here for details

schema documentation

```
perl $ENSEMBL_CVS_ROOT_DIR/ensembl/misc-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 compara3 -user ensro -dbname kb3_ensembl_compara_71
-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 $ENSEMBL_CVS_ROOT_DIR/ensembl-webcode/htdocs/info/docs/api/compara/ and commit both (the compara one and the webcode one).

- [ ] Update the 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/

    Perl source files should be in $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/tutorial_examples/

    They can be transformed into HTML files with the following command:

    $ highlight --syntax=perl --style zellner -f --class-name=comp_tut TUTORIAL_EXAMPLE.pl > TUTORIAL_EXAMPLE.inc

    If you don't have 'highlight' installed, you can ask Matthieu to do it

    The .inc files can then be moved to $ENSEMBL_CVS_ROOT_DIR/ensembl-webcode/htdocs/info/docs/api/compara/
- [ ] Save tutorial documentation as pdf
  - Click here for details

    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
- [ ] Update main ensembl species tree if there are any new species

  Ask Matthieu how this is done. http://www.ensembl.org/info/about/species_tree.pdf
- [x] CVS commit any modified files or added tutorial examples

Update the .inc files

cd $ENSEMBL_CVS_ROOT_DIR/public-plugins/ensembl/htdocs/info/docs/compara

- [x] Protein trees

Click here for details

tree-stats

```
cat $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/sql/tree-stats.sql \
    | mysql -hcompara3 -uensadmin -p${ENSADMIN_PSW}
mp12_compara_homology_72 -H \
    | sed 's/\/TABLE>/\/TABLE>\n/g' | grep -v optimize | \
    bash
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/get_stats_trees.sh
pt > protein_trees.inc
```

- [x] ncRNA trees

Click here for details

tree-stats

```
cat $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/sql/tree-stats.sql | sed
's/ENSEMBLPEP/ENSEMBLTRANS/' | sed 's/pep/trans/g' \
    | mysql -hcompara2 -uensadmin -p${ENSADMIN_PSW} mp12_compara_nctrees_72
-H \
    | sed 's/\/TABLE>/\/TABLE>\n/g' | grep -v optimize | \
    bash
$ENSEMBL_CVS_ROOT_DIR/ensembl-compara/scripts/pipeline/get_stats_trees.sh
nc > nc_trees.inc
```

- [x] CVS commit these files

Data dumps

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.

- [x] DNA data dumps
  Generally the person who ran the pipeline will also do the data dumps. The instructions are in $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/README.multi_align.dumps

- [x] Gene tree dumps
  Generally the person who ran the pipeline will also do the data 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, only these 2 parameters need to be changed:

rel being the current release number

rel_coord username of the release coordinator (defaults to you !)

You can still check that the parameters are correct

rel_db pointing at the release database,

target_dir suitable for creating and storing the dumps in

If not happy, edit the changes, save the config file and exit the editor.

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 Bio::EnsEMBL::Compara::PipeConfig::DumpTrees_conf
-tree_type protein -pipeline_db -host=compara2
```

rel.64: testing sqlite mode failed: too many occurrences of "database locked". We should stick to mysql.

Then run both -sync and -loop variations of the beekeeper.pl command suggested by init_pipeline.pl .

This pipeline will produce protein_tree dumps in the directory pointed at by 'target_dir' parameter.

▾ Click here for times

rel_60: took 5 hours on a "bad lustre" day. On one of such days you're better off pointing at your home directory!

rel_64: 48m

rel65: 2h5m

rel66: more than 1 week (lustre was very slow, and dump_all_homologies takes ages)

rel67 and rel68: much faster (< 2 days)

rel70: < 3 hours

Create the ncRNA pipeline from the same config file:

init_pipeline

```
init_pipeline.pl Bio::EnsEMBL::Compara::PipeConfig::DumpTrees_conf
-tree_type ncrna -pipeline_db -host=compara2
```

Then run both -sync and -loop variations of the beekeeper.pl command suggested by init_pipeline.pl .

This pipeline will produce ncrna_tree dumps in the directory pointed at by 'target_dir' parameter.

Click here for times

rel60: 8m, the other extreme.

rel64: 5m

rel65: 7m

rel66: 1h in total

rel67 and rel68: a few hours

rel70: < 2 hours

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

- [x] 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. You can scp the file to login.ebi.ac.uk:/nfs/ftp/pub/databases/ensembl/ensembl_compara/ and from there, create its MD5 checksum

- [x] Report locations of the dumps to ensembl-production

Click here for details

Dna dumps:

Not all the multiple alignments are run for each release. The data dumps for any multiple alignments not run this release should be copied from the previous release. Currently (e72), the full set of mulitple alignments with the corresponding dump files, are:

6 primates EPO (emf)

13 eutherian mammals EPO (emf)

5 teleost fish EPO (emf)

3 neognath birds EPO (emf)

8 teleost fish EPO_LOW_COVERAGE (emf, bed)

36 eutherian mammals EPO_LOW_COVERAGE (emf, bed)

20 amniota vertebrates Pecan (emf, bed)

Gene tree dumps:

'target_dir'/emf should go to /emf/ensembl-compara/homologies/ on the ftp

'target_dir'/xml should go to /xml/ensembl-compara/homologies/ on the ftp

Final things

- [x] Create a word ducument and a pdf dump of this document

- Click here for details

  Create a word document and pdf file of this document and place it in $ENSEMBL_CVS_ROOT_DIR/ensembl-compara/docs/

  CVS commit this

- [x] Dump the master database and place the copy in a safe place
  - Click here for details

    prev (rel65) /warehouse/ensembl01/sf5/Compara_master_dumpps/sf5_ensembl_compara_master.sql.04_01_12.gz

    rel65 /warehouse/ensembl01/compara/master_dumps/sf5_ensembl_compara_master.sql.04_01_12.gz

    rel66 /warehouse/ensembl01/compara/master_dumps/sf5_ensembl_compara_master.66.gz (3min to dump)

    rel71 /warehouse/ensembl01/compara/master_dumps/sf5_ensembl_compara_master.71.gz (3min to dump)