

RACEdb

version 14Dec05

This document describes the schema of a database, RACEdb, designed to store and retrieve information related to RACE experiments and their fragments hybridized on a genome tiling array. The overall organization of the data consists of three main levels:

- **projects:** a project is a set of RACE experiments.
- **RACE experiments:** a RACE experiment is a single RACE reaction that starts from a specific index region in the genome using a set of primers and going in either 5' or 3' direction.
- **RACE fragments:** a RACE fragment, RACEfrag or RF hereafter, is a genomic interval between two positions in a chromosome which has hybridized with a RACE product resulting of one or more of the RACE experiments.

Following this organization of the data a relational schema, shown in Figure 1, was designed which consists of the following SQL tables:

- **projects:** information about each project
- **race_exps:** information about each RACE experiment
- **primers:** information about the primers used in the RACE experiments.
- **race_frags:** genomic intervals (RFs) where some RACE product coming from a specific tissue or cell line has hybridized.
- **cell_lines:** information about the cell lines used in the RACE experiments (when available).
- **tissues:** information about the tissues used in the RACE experiments (when available).
- **race_frags_clusters:** pairs of RFs and RACE experiments identifiers that establish what set of RFs is assigned as being the result of a particular RACE experiment. This assignment can be specified with some degree of confidence.

and the following relationships between them:

- **projects—race_exps:** one project has one or more RACE experiments.
- **projects—race_frags:** one project has one or more RFs.
- **race_exps—primers:** one RACE experiment has one or more primers.
- **race_exps—race_frags_clusters:** one RACE experiment has one or more RFs.
- **cell_lines—race_frags:** one cell line produces one or more RFs.
- **tissues—race_frags:** one tissue produces one or more RFs.
- **race_frags—race_frags_clusters:** one RF can belong to one or more RACE experiments (normally with different degrees of confidence).

The rest of this document contains a detailed description of the fields that form each of the previously described SQL tables. All fields called #id and described as `internal identifier` correspond to auto-incremented identifiers generated by the database.

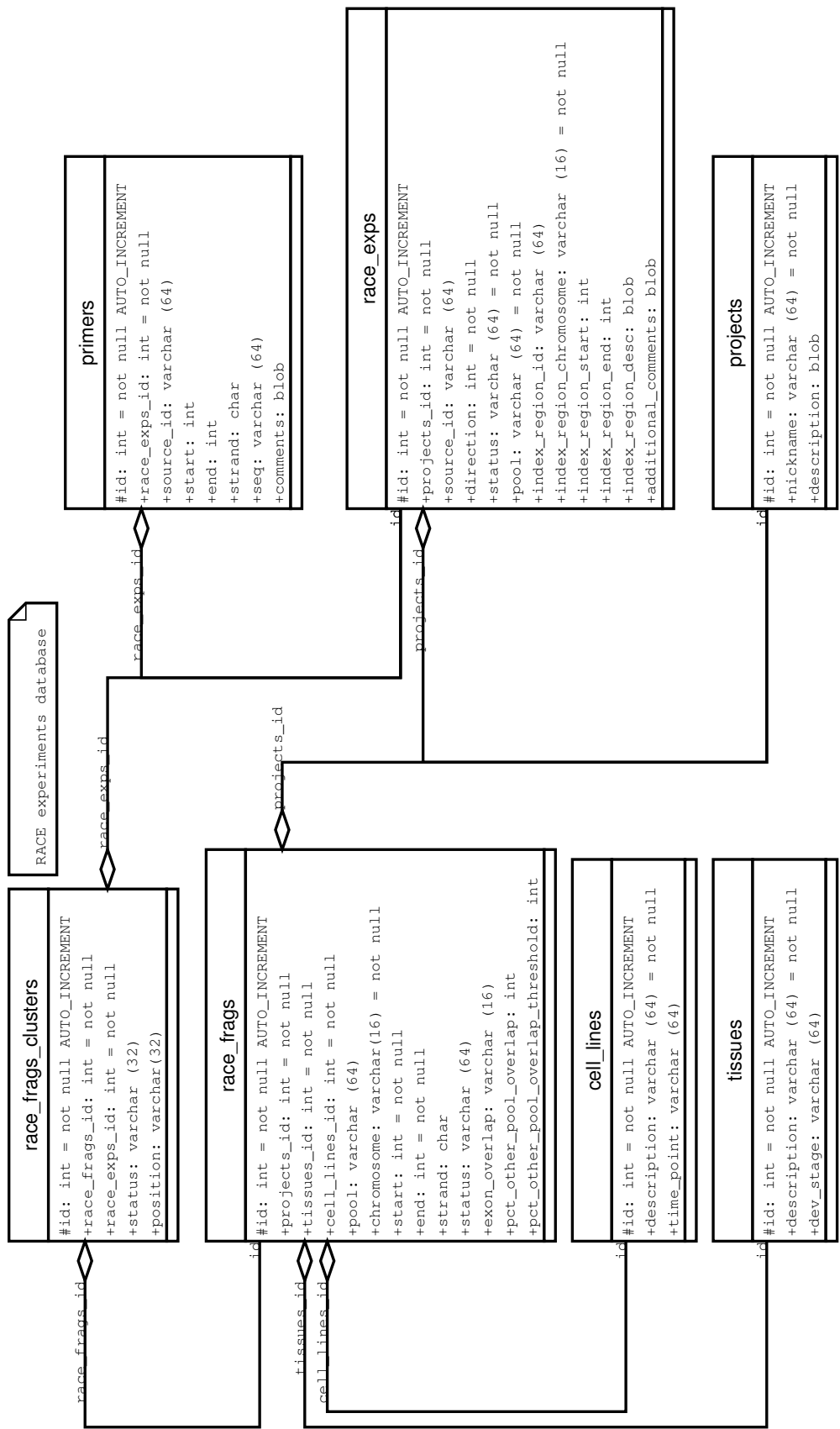


Figure 1: Relational schema of RACEdb.

projects

#id internal identifier
nickname one-word identifier of the project
description one-sentence description of the project

race_exps

#id internal identifier
projects_id internal identifier of the related project
source_id one-word identifier of the RACE experiment given by the data producer
direction direction of the RACE specified as a number, either 5 or 3
status one-word keyword informing on the extent to which the index region is interrogated by the probes in the tiling array. Values, for instance, can be INTERROGATED, NOT_INTERROGATED and PARTLY_INTERROGATED
pool pool or pools to which the index region belongs to
index_region_id one-word identifier of the index region, e.g., a locus or transcript identifier
index_region_chromosome chromosome of the index region
index_region_start start position of the index region
index_region_end end position of the index region
index_region_desc description of the index region
additional_comments information to help understanding the data, e.g., when something specific to this experiment happen

primers

#id internal identifier
race_exps_id internal identifier of the related RACE experiment
source_id one-word identifier of the primer given by the data producer
start start position of the primer
end end position of the primer
strand strand of the primer
seq DNA sequence of the primer
comments additional information as, e.g., whether the primer is spliced

race frags

#id internal identifier
projects_id internal identifier of the related project
tissues_id internal identifier of the related tissue
cell_lines_id internal identifier of the related cell line
pool one-word identifier of the pool to which it belongs
chromosome chromosome of the RF
start start position of the RF
end end position of the RF
strand end position of the RF
status one-word keyword to help us using properly this RF, e.g., KEPT or DISCARDED depending on whether it shows up in more than one pool
exon_overlap whether it overlaps an annotated exon
pct_other_pool_overlap percentage of overlap with other pools

pct_other_pool_overlap_threshold threshold on pct_other_pool_overlap above we decided something about the status of this RF

cell_lines

#id internal identifier
description one-word description of the cell line
time_point time point of development of this cell line (if applicable)

tissues

#id internal identifier
description one-word description of the tissue
dev_stage developmental stage of this tissue (if applicable)

race_fragments_clusters

#id internal identifier
race_fragments_id internal identifier of the related RF
race_experiments_id internal identifier of the related RACE experiment
status status (confidence) of the relationship, e.g., ASSIGNED, ASSIGNED_AMBIGUOUS1_CONFIDENTLY, etc.
position location of the RF with respect to the index region of the related RACE experiment as, e.g., INTERNAL, EXTERNAL, EXONIC, INSIDEIDXREG, OVLPNGIDXREG, OUTSIDEIDXREG, etc.