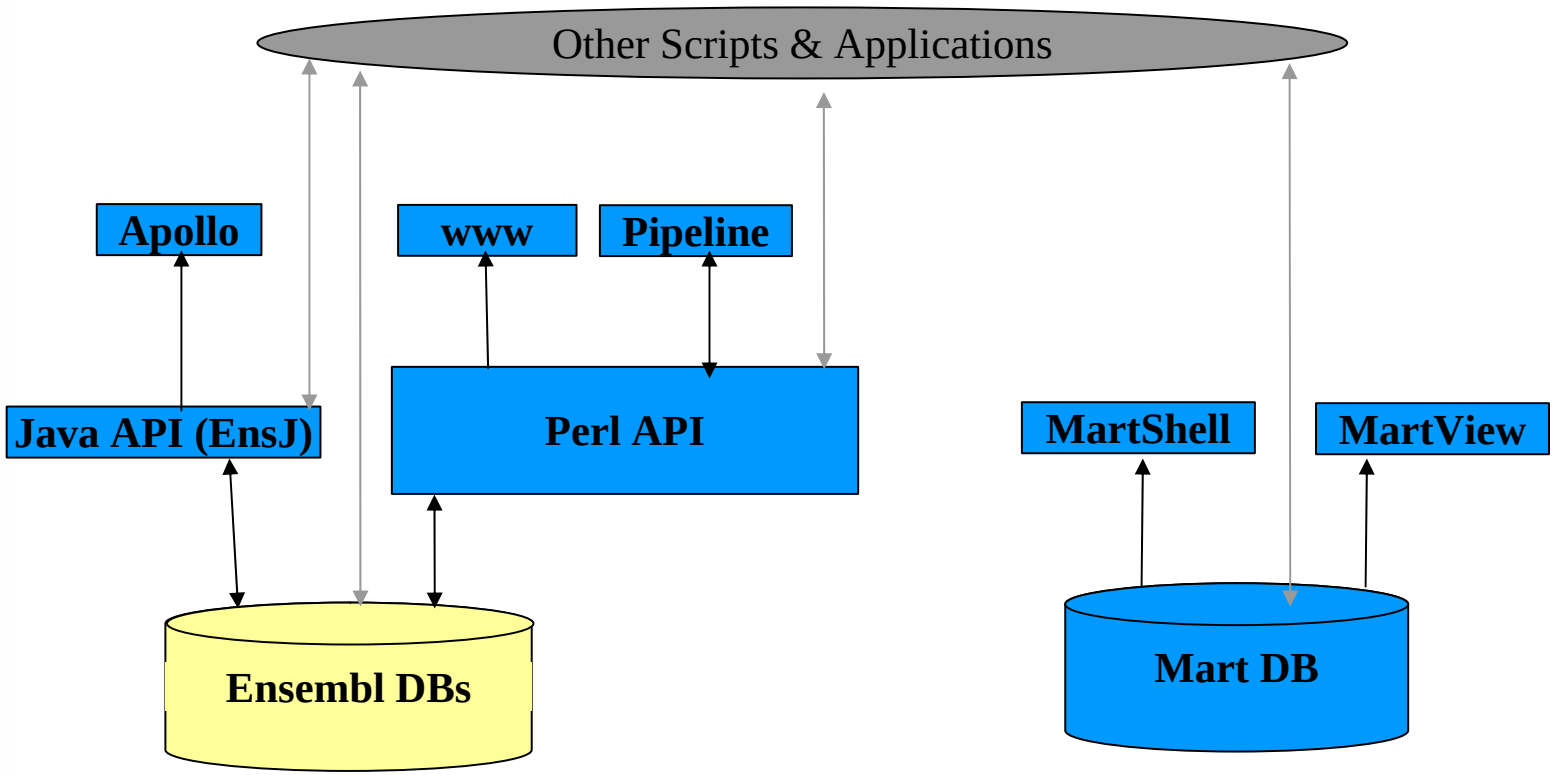
The Ensembl Database Schema

European Bioinformatics Institute

Requirements for the schema

- Store data for human genome
- … and all the other genomes we have
- … and all the genomes we might get
- Flexible to add more data
- Easy to adapt to new genome
- Responds fast enough for web site display and pipelined genebuild

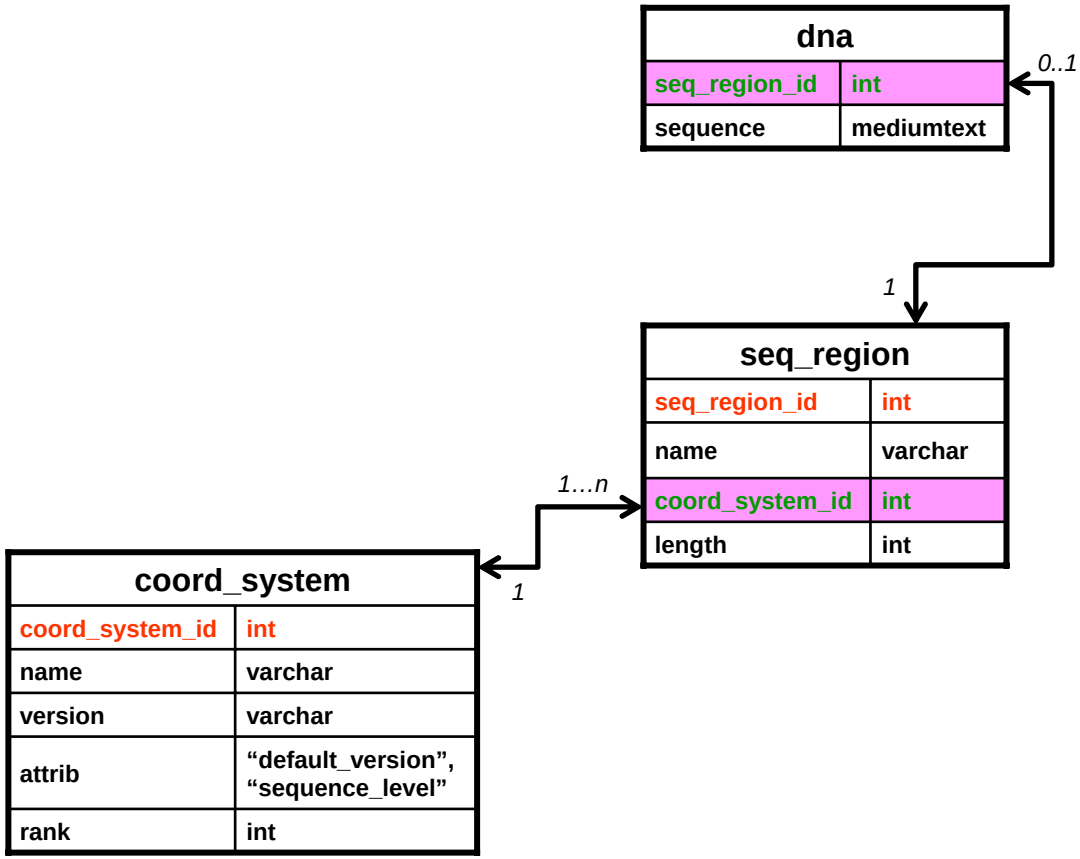
System Context



Sequence regions

- Everything which represents a length of nucleotide sequence is a sequence region.
  - chromosome, BAC-clone, supercontig, scaffold, contig …
- Sequence regions of the same type belong to the same coordinate system.
  - “1”, “2”, and “3” are sequence regions with coordinate system “chromosome”
- Sequence regions have names and lengths.

Sequence regions



Example sequence region

- Chromosome, 1, 200MB
- Clone, AL123123.4, 132KB
- NT_contig, NT_1245675, 17MB
- Contig, AC332232.1.1.123223, 123223

Coordinate system

- The coord_system describes the type of the sequence region
  - Name (“chromosome”, “contig”,…)
  - Version (eg. NCBI35, ZFISH3)
  - Internal id (coord_system_id)
  - Attrib – (default, sequence_level)
  - rank (1..n)
- If you have 2 coordinate systems with the same name, choose a “default” one. They need to have different versions (NCBI34, NCBI35).
- The lower the rank, the bigger the sequence region. Choose 1 for your biggest regions (chromosomes).
- Only one coordinate system is allowed to contain sequence regions with actual sequence attached. Flag it with Attrib = sequence_level.

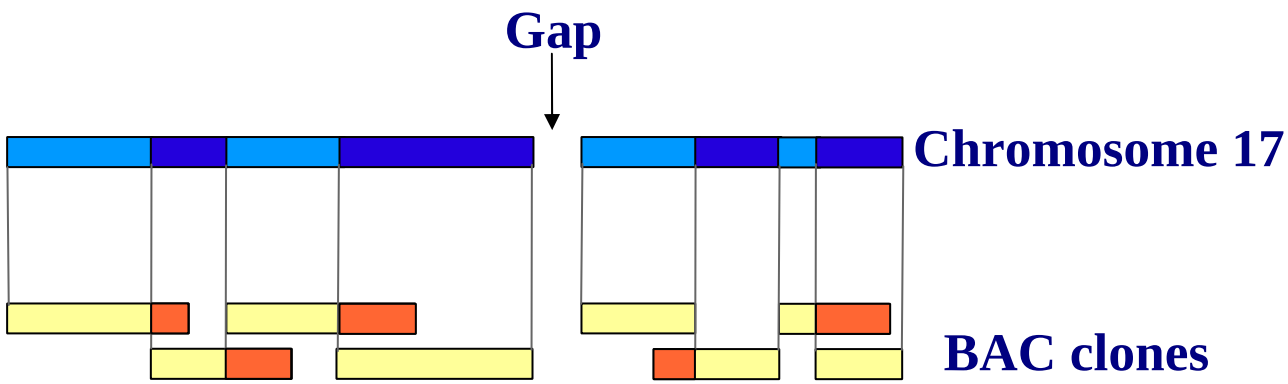
Coordinate system

- “contig”
  - Contiguous sequence.
  - “N”s should be rare and of short length.
  - Can serve as your basic sequence holder
- “clone”
  - Should have a real BAC or PAC or maybe YAC behind it.
  - Might not be contiguous
- “supercontig”
  - Assembled from smaller contiguous sequences.
  - May have small gaps (eg between read pairs)
- “chromosome”
  - Use it only for real chromosomes.
  - or for alternative sequences of reference chromosomes.
- “chunk”
  - Artificial coordinate system to hold sequence regions for technical reasons.
  - Create, when none of the other coordinate systems can hold your sequence (eg. You only have full length chromosomes as coordinate system but they are to long to store)
  - or when you have 2 real sequence containing coordinate systems.

Assemblies

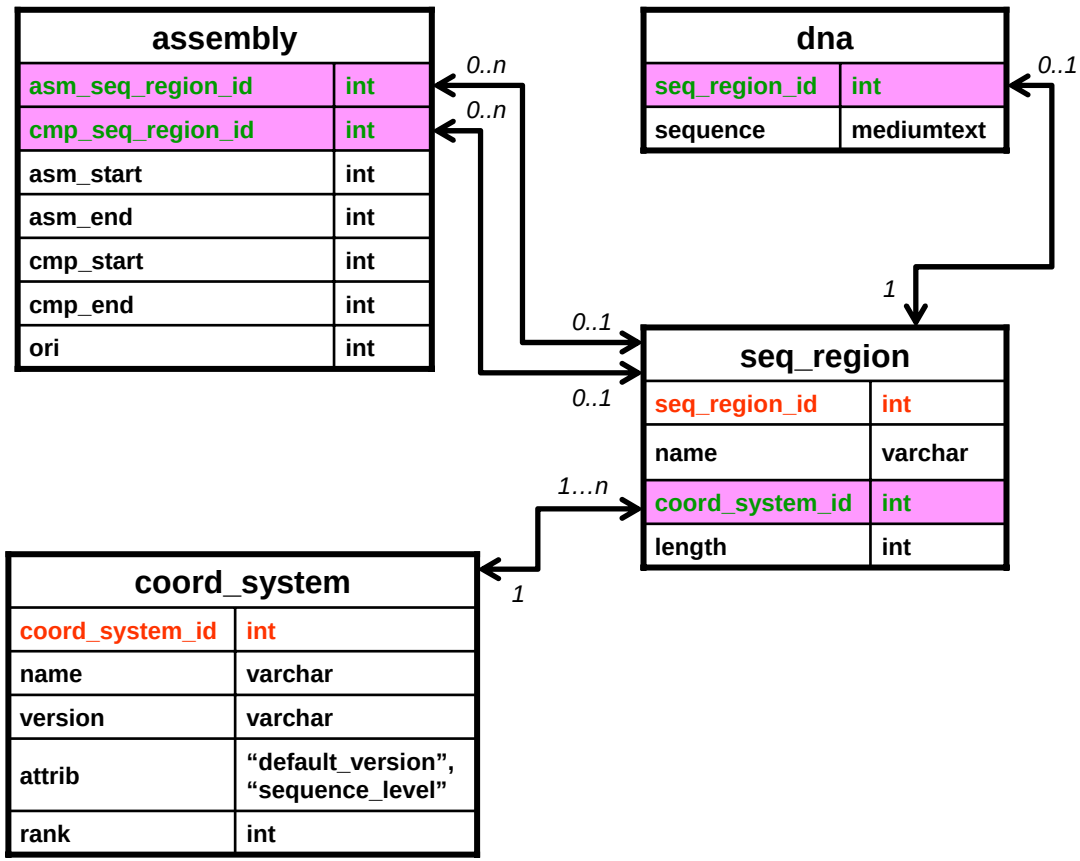
- An assembly defines how sequence regions in one coordinate system are made up of sequence regions from another coordinate system.
- For example human chromosomes are assembled from a “tiling path” of BAC clones.
- Assembly information stored in Ensembl makes it possible to obtain features or sequence from arbitrary sequence regions.

Assemblies



- A row in the assembly table references an assembled and component sequence region.
- How a piece of the assembled sequence region is made from a piece of a component region is defined by a pair of coordinates and an orientation.
- Gaps are represented by the absence of assembly information.

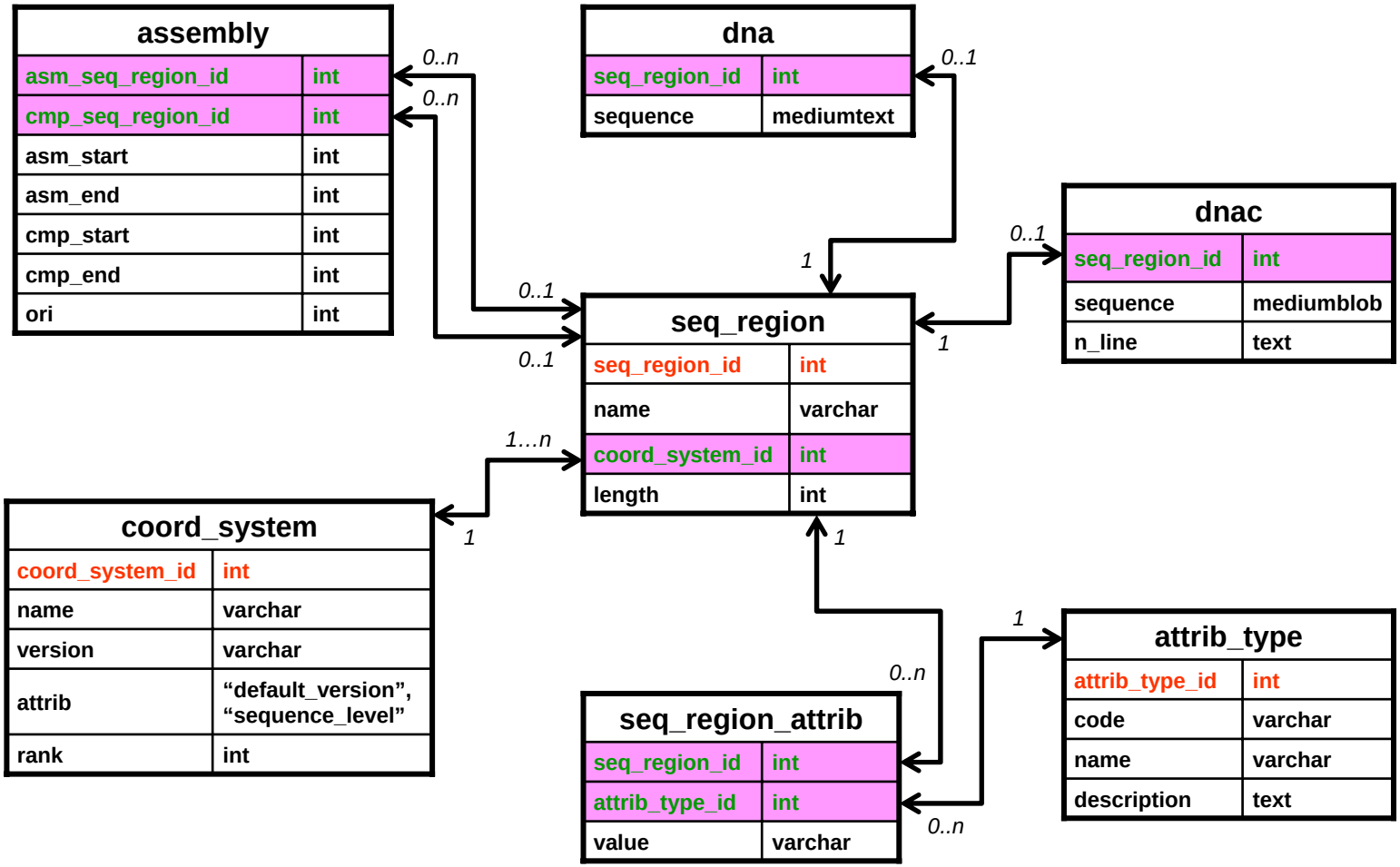
The assembly table



Sequence region attributes

- Arbitrary attributes may be associated with a sequence region via the seq_region_attrib table.
  - sanger ids for certain clones.
  - htg phases for clones.

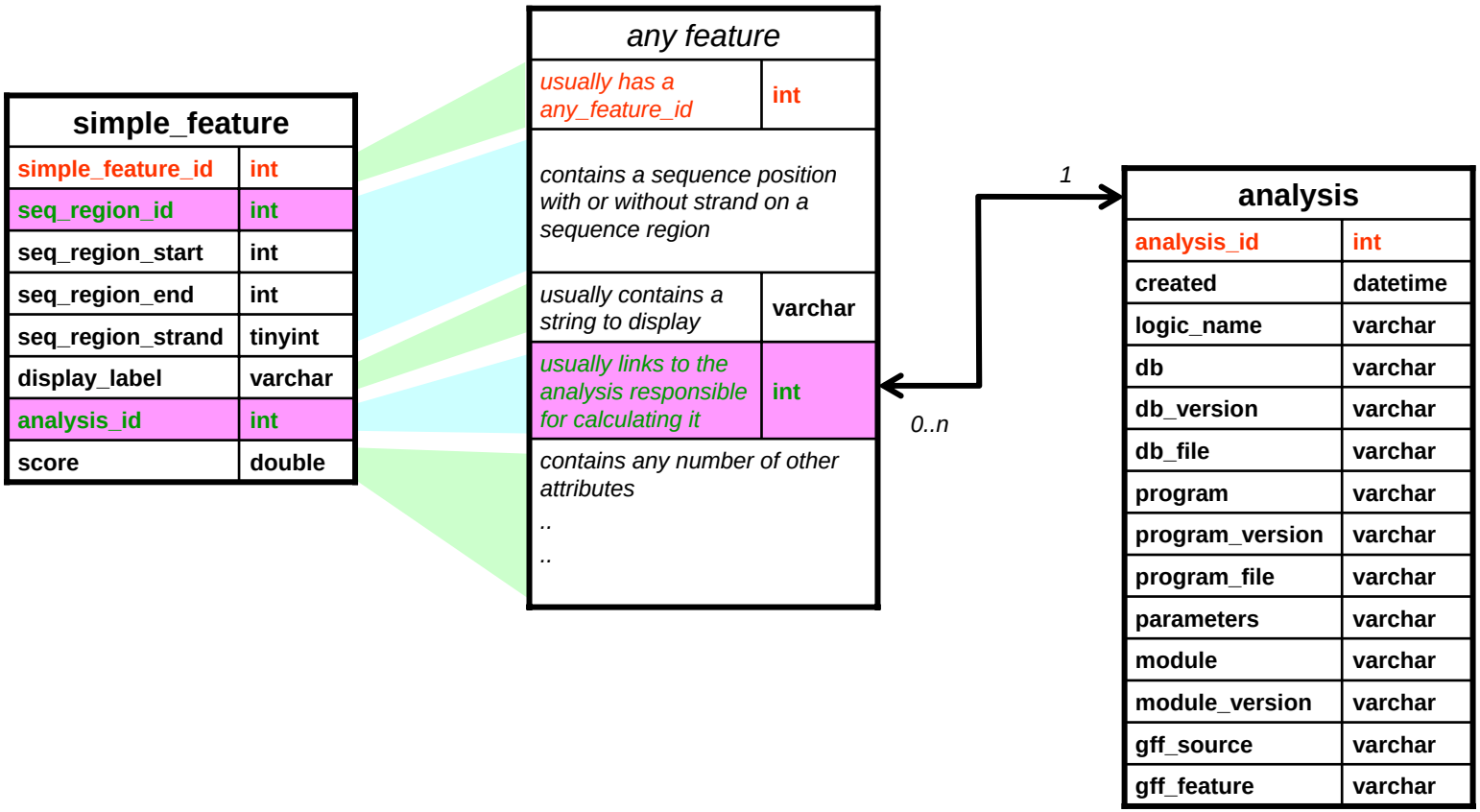
The seq_region_attrib table



Features

- Features are annotation information placed on the genome.
- A feature is stored as a position on a sequence region.

A standard feature



other Features

dna_align_feature	
dna_align_feature_id	int
Sequence position	
hit_start	int
hit_end	int
hit_strand	tinyint
hit_name	varchar
analysis_id	int
score	double
evalue	double
perc_ident	float
cigar_line	text

protein_align_feature	
protein_align_feature_id	int
Sequence position	
hit_start	int
hit_end	int
hit_name	varchar
analysis_id	int
score	double
evalue	double
perc_ident	float
cigar_line	text

repeat_feature	
repeat_feature_id	int
Sequence position	
repeat_start	int
repeat_end	int
repeat_consensus_id	int
analysis_id	int
score	double

1..n

1

repeat_consensus	
repeat_consensus_id	int
repeat_name	varchar
repeat_class	varchar
consensus	text

prediction_exon	
prediction_exon_id	int
prediction_transcript_id	int
exon_rank	int
Sequence position	
start_phase	tinyint
score	double
p_value	double

1

1..n

prediction_transcript	
prediction_transcript_id	int
Sequence position	
analysis_id	int

Genes are features

sequence == seq_region

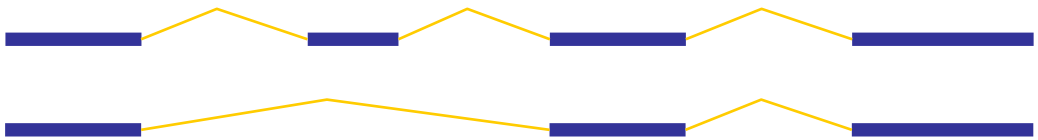
exons

have stable ids

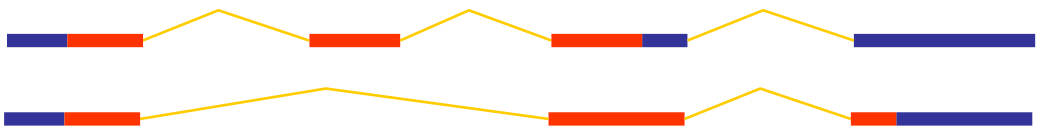
ENSE0001 ENSE0002 ENSE0003 ENSE0004

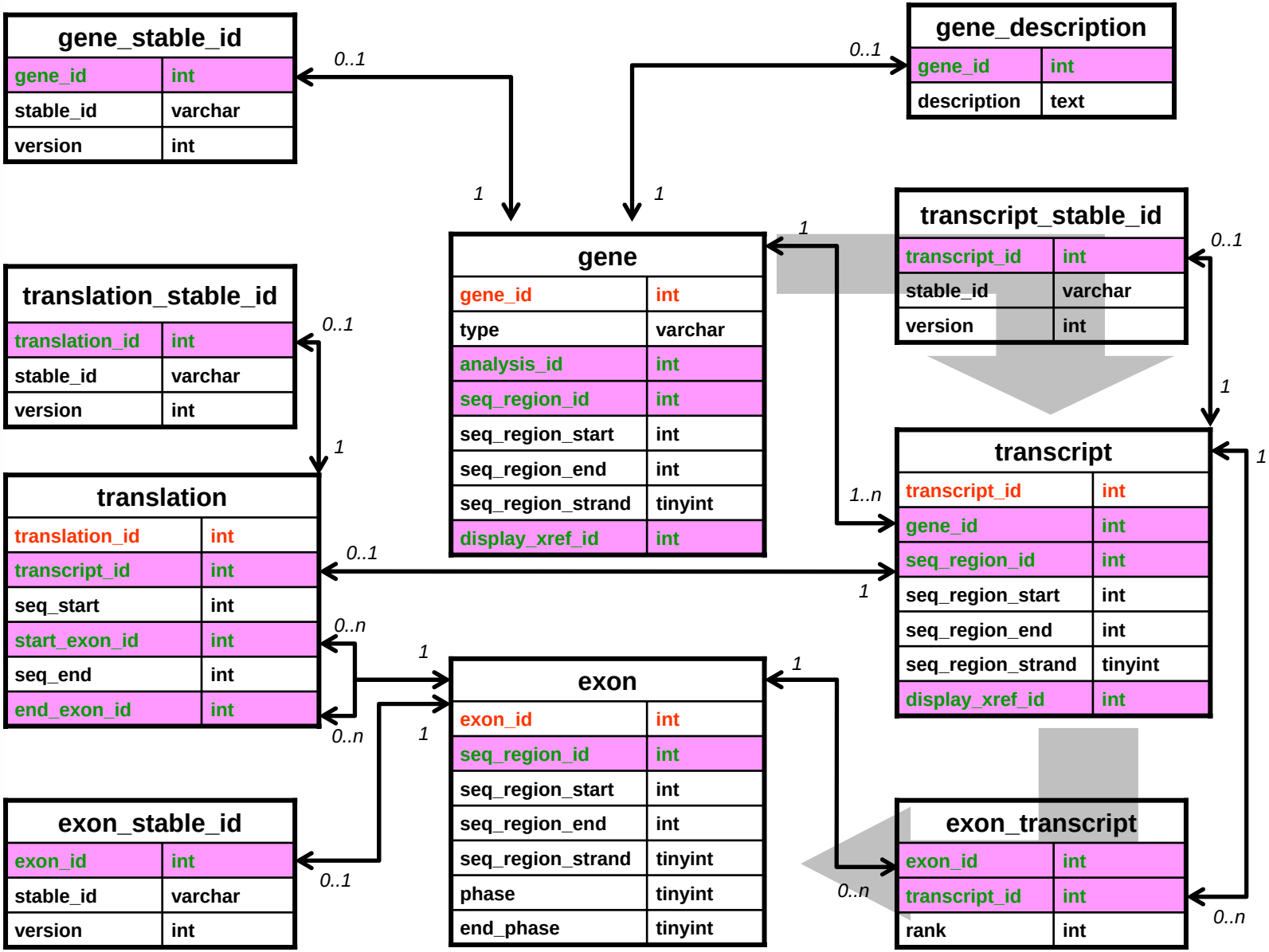
alternative spliced transcripts

stable ids



.. with different translations

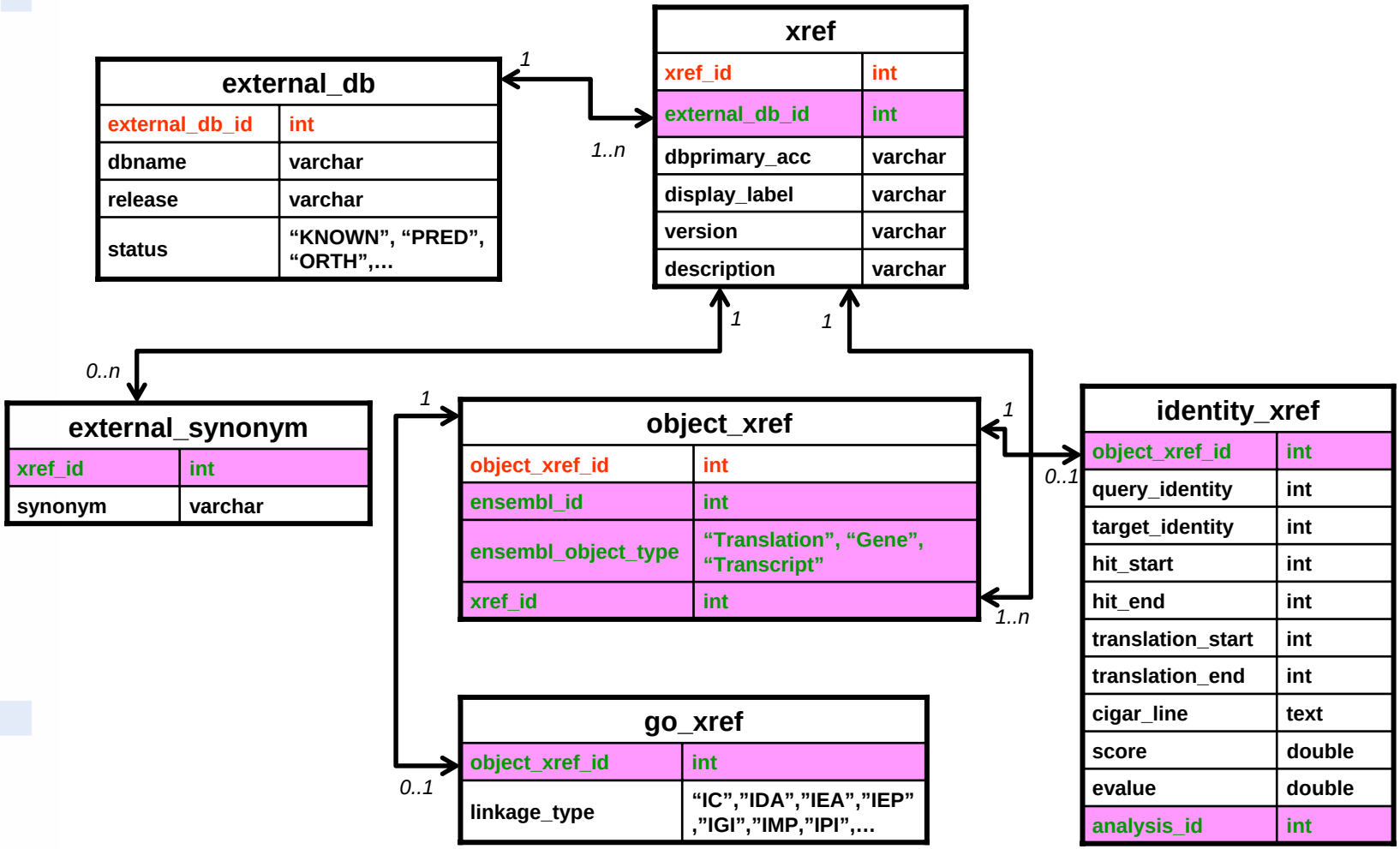



gene_stable_id
gene_id int
stable_id varchar
version int
gene_description
gene_id int
description text
transcript_stable_id
transcript_id int
stable_id varchar
version int
translation_stable_id
translation_id int
stable_id varchar
version int
gene
gene_id int
type varchar
analysis_id int
seq_region_id int
seq_region_start int
seq_region_end int
seq_region_strand tinyint
display_xref_id int
transcript
transcript_id int
gene_id int
seq_region_id int
seq_region_start int
seq_region_end int
seq_region_strand tinyint
display_xref_id int
translation
translation_id int
transcript_id int
seq_start int
start_exon_id int
seq_end int
end_exon_id int
exon
exon_id int
seq_region_id int
seq_region_start int
seq_region_end int
seq_region_strand tinyint
phase tinyint
end_phase tinyint
exon_stable_id
exon_id int
stable_id varchar
version int
exon_transcript
exon_id int
transcript_id int
rank int
0..1
1
0..n
1..n

External references

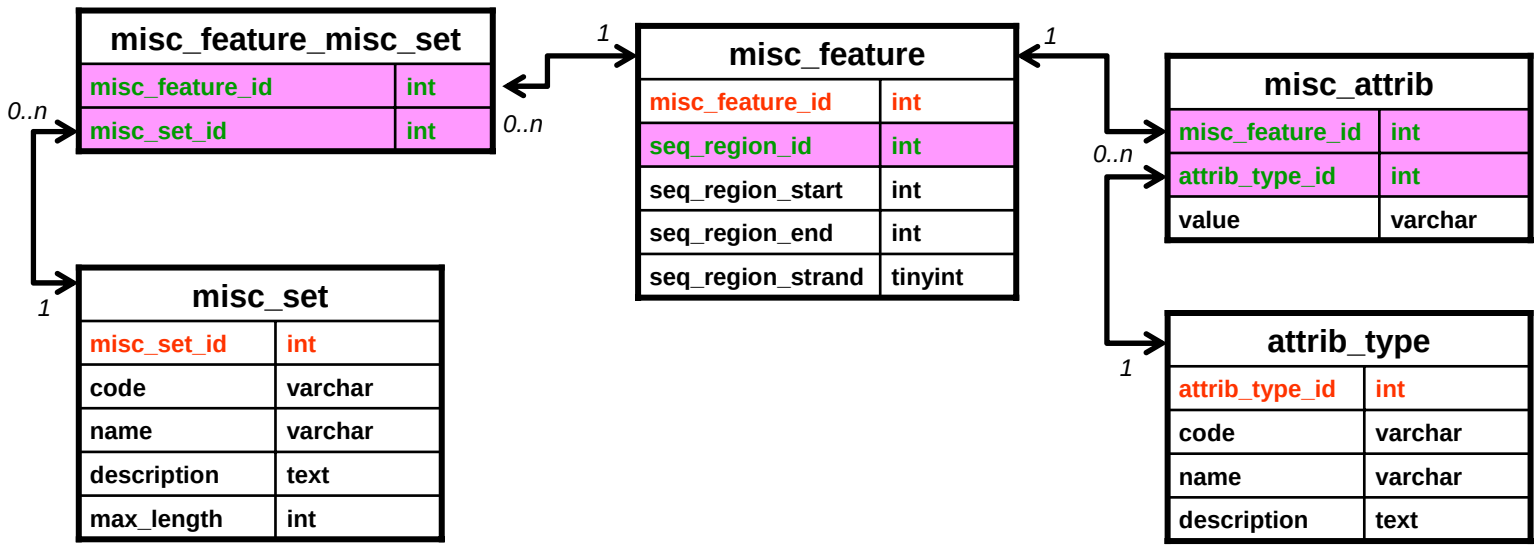
- Ensembl objects can reference objects in other databases.
  - eg a SWISSPROT identifier, GO identifier, Refseq , HUGO, …
- External references are used for display ids in Genes and Transcripts. These links are provided directly in Gene and Transcript.

External references



Misc features

- are features with user definable attributes
- it can belong to a set to provide a trackname for the feature.
- a misc feature can be in more than one set.



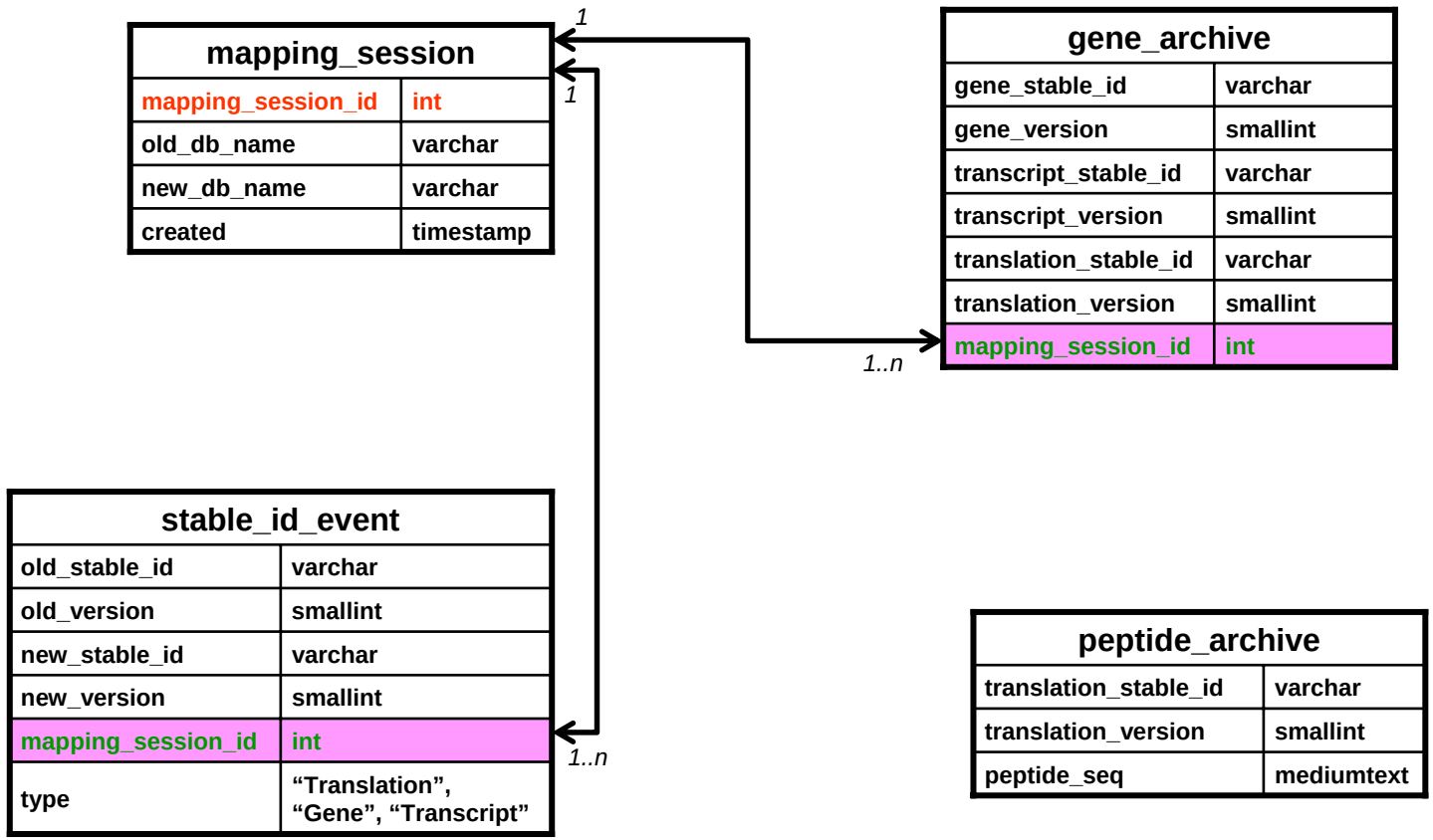
Archive tables

- For some species there is a record of old predictions available.
  - human, mouse
- You can get
  - old peptide sequences
  - how an older gene prediction was made up from transcripts/translations.
  - how genes and transcripts were merged and split from older prediction to newer prediction.

Archive tables

- A mapping session describes the event when a set of ids is mapped from an older database to a newer database
  - Version number are a relatively new addition, so you need the mapping session to uniquely specify a Gene.
- A stable id event for a gene states that some part of the old gene is to be found in the new gene.
  - Same for Transcript.
- The gene archive records the gene structure of the older gene, when the gene has changed during a mapping session.
- The peptide archive records the peptide sequence of he old version of the peptide, when the peptide changes.

Archive tables



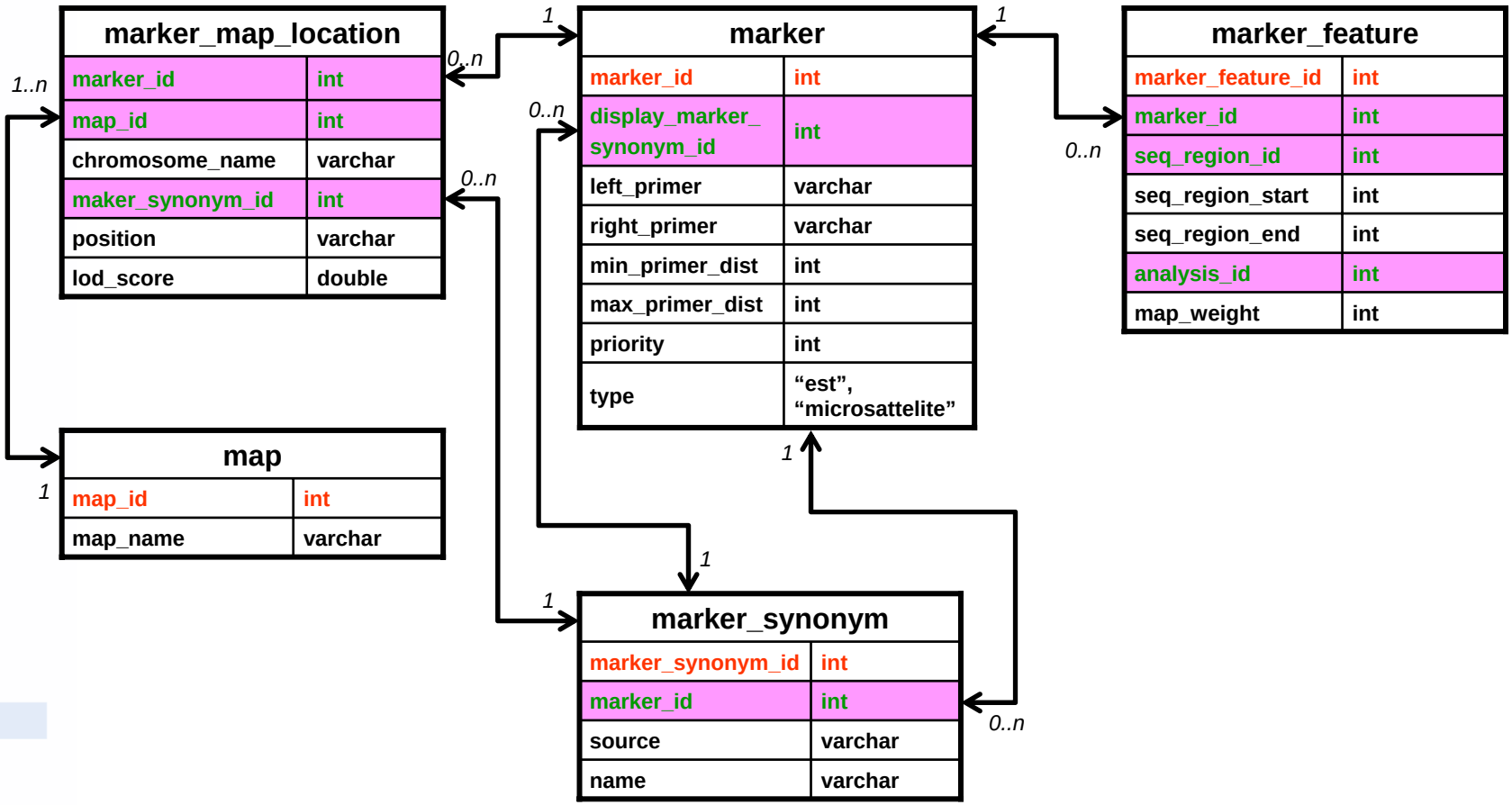
mapping_session	
mapping_session_id	int
old_db_name	varchar
new_db_name	varchar
created	timestamp

gene_archive	
gene_stable_id	varchar
gene_version	smallint
transcript_stable_id	varchar
transcript_version	smallint
translation_stable_id	varchar
translation_version	smallint
mapping_session_id	int

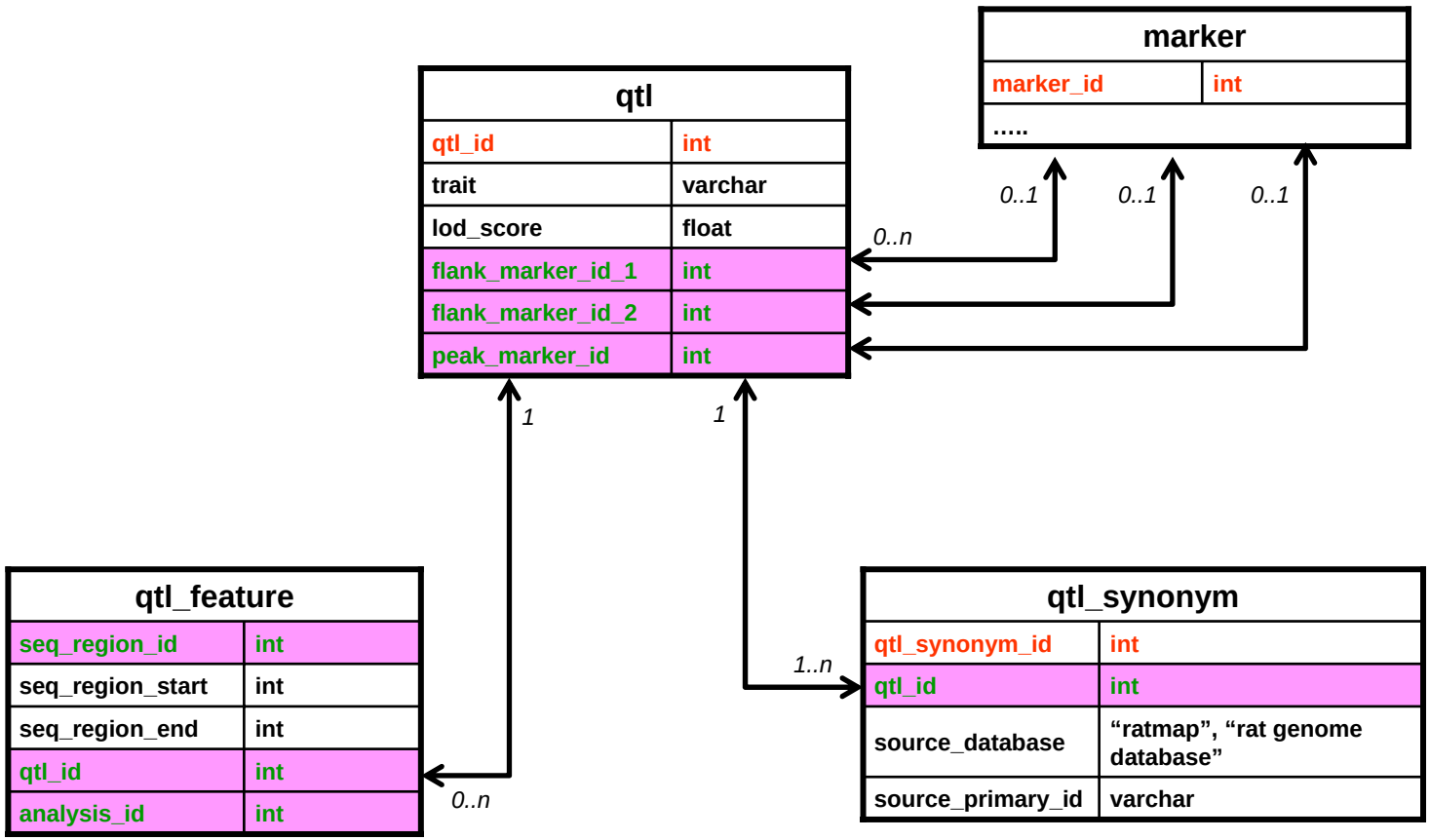
stable_id_event	
old_stable_id	varchar
old_version	smallint
new_stable_id	varchar
new_version	smallint
mapping_session_id	int
type	“Translation”, “Gene”, “Transcript”

peptide_archive	
translation_stable_id	varchar
translation_version	smallint
peptide_seq	mediumtext

Markers and marker features



QTLs


marker
marker_id
int
.....
qtl
qtl_id
int
trait
varchar
lod_score
float
flank_marker_id_1
int
flank_marker_id_2
int
peak_marker_id
int
0..1
0..1
0..1
0..n
1
1
qtl_feature
seq_region_id
int
seq_region_start
int
seq_region_end
int
qtl_id
int
analysis_id
int
0..n
1..n
qtl_synonym
qtl_synonym_id
int
qtl_id
int
source_database
"ratmap", "rat genome database"
source_primary_id
varchar

Meta information

- Meta table contains general key-value pairs
  - eg. species name
  - taxonomy id
- which coordinates can be mapped and how
- future additions likely
- Meta_coord says which feature is stored in which coordinate system
  - more than 1 entry possible
  - no feature retrieval without it.

meta	
meta_id	int
meta_key	varchar
meta_value	varchar

meta_coord	
table_name	varchar
coord_system_id	int

Protein features

- Features can be add to the peptide sequence
- hit_id is usually Pfam, Prosite, prints identifier.

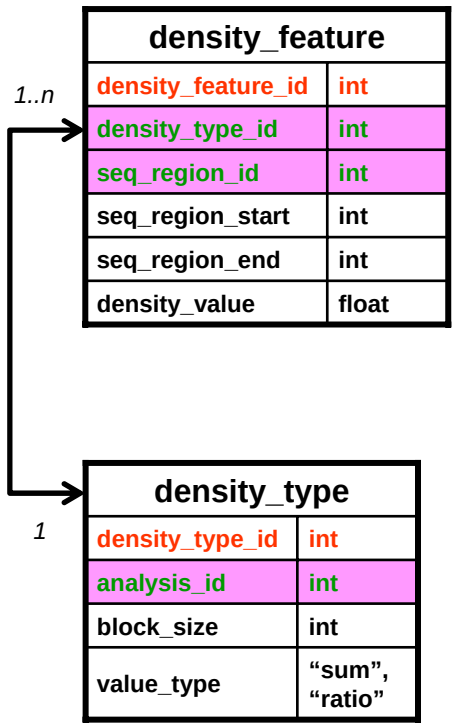
- interpro table links these to interpro ids.
- xrefs have further information for them.

protein_feature	
protein_feature_id	int
translation_id	int
seq_start	int
seq_end	int
hit_start	int
hit_end	int
hit_id	varchar
analysis_id	int
score	double
evalue	double
perc_ident	float

interpro	
interpro_ac	varchar
id	varchar

Density feature

- Density features assign numeric values to regions.
  - GC content
  - gene count
  - repeat coverage
- The blocksize and value_type enable interpolation by API



Karyotype bands

- Karyotype table defines the banding pattern of the chromosomes and how to draw the ideogram.
- A single band is just like any other feature in the database.
- band naming convention depends on species and resolution.
- stain could be (“acen”, “gvar”, “gpos25”, “gpos50”, “gpos75”, “gpos100”, “gneg” ...)

karyotype	
karyotype_id	int
seq_region_id	int
seq_region_start	int
seq_region_end	int
band	varchar
stain	varchar

Supporting Evidence

- Exons can be linked to features.
- These are alignment features that were used as evidence when the exon was created.
  - supporting evidence

supporting_feature	
exon_id	int
feature_type	“protein_align_feature”, “dna_align_feature”
feature_id	int

New tables

transcript_attrib	
transcript_id	int
attrib_type_id	int
value	varchar

translation_attrib	
translation_id	int
attrib_type_id	int
value	varchar

assembly_exception	
assembly_exception_id	int
seq_region_id	int
seq_region_start	int
seq_region_end	int
exc_type	“HAP”, “PAR”
exc_seq_region_id	int
exc_seq_region_start	int
exc_seq_region_end	int
ori	int

alt_allele	
alt_allele_id	int
gene_id	int

Acknowledgements

Ensembl Core Software Team:

- *Arne Stabenau*
- *Glenn Proctor*
- *Craig Melsopp*
- *Ian Longden*

The Rest of the Ensembl Team.