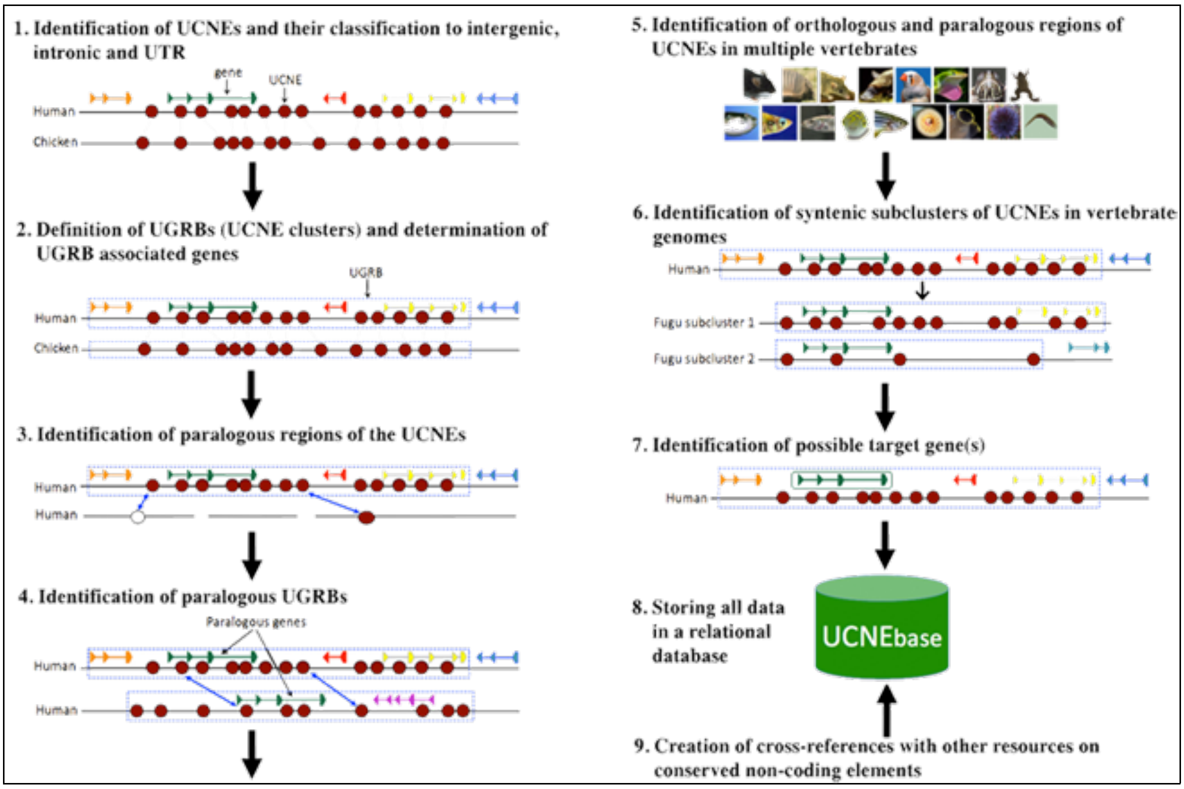
UCNEbase documentation content

UCNEbase methodology

The information provided by UCNEbase is generated by a combination of automatic procedures and manual curation steps. The methodology used for the creation of UCNEbase is schematically shown in the figure below.



1. Identification of UCNEs and their classification to intergenic, intronic and UTR
gene
UCNE
Human
Chicken
2. Definition of UGRBs (UCNE clusters) and determination of UGRB associated genes
UGRB
Human
Chicken
3. Identification of paralogous regions of the UCNEs
Human
Human
4. Identification of paralogous UGRBs
Paralogous genes
Human
Human
5. Identification of orthologous and paralogous regions of UCNEs in multiple vertebrates
6. Identification of syntenic subclusters of UCNEs in vertebrate genomes
Human
Fugu subcluster 1
Fugu subcluster 2
7. Identification of possible target gene(s)
Human
8. Storing all data in a relational database
UCNEbase
9. Creation of cross-references with other resources on conserved non-coding elements

Definition of Ultra-Conserved Non-coding Elements (UCNEs)

We defined UCNEs as non-coding human DNA regions that exhibit ≥ 95% sequence identity between human and chicken and are longer than 200bp. The sequence identity threshold corresponds to a base substitution rate of approximately 1% per 100 million years. We have previously shown that sequences fulfilling such stringent criteria exist only in vertebrates. To compile a list of human UCNEs we scanned whole genome alignments between human and chicken downloaded from UCSC genome browser with a sliding window technique. Human and chicken were selected as reference species for two main reasons: (i) their evolutionary distance provides high specificity in detecting functional elements and (ii) both genome assemblies are of high quality and thus suitable for identifying large syntenic regions. From the initially extracted set of ultraconserved sequence elements, we eliminated coding regions, and a few human repetitive sequences aligning with the same chicken sequence. The remaining 4’351 sequences composed our reference set of UCNEs. Each element of this set was then classified as either “intergenic”, “intronic” or “UTR-associated” according to the human gene annotation from RefSeq. The length of the UCNEs identified in this way ranged from 200 - 1419 bp with a mean = 325 bp and a median = 283 bp. The total length is 1.4 Mbp.

Definition of Ultraconserved Genomic Regulatory Blocks (UGRBs)

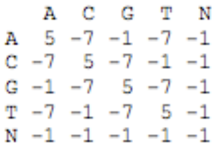
We defined UGRBs (also referred to as UCNE clusters) as arrays of UCNEs that are syntenically conserved between the human and chicken genomes. Syntenic conservation means that the orthologs of the individual UCNEs of a human UGRB occur in the same order within a restricted area of a chicken chromosome. During the initial scan, we required that neighbouring UCNEs must not be separated by more than 0.5 Mb in both human and chicken. However, a few exceptions to this rule were made during subsequent manual curation based on visual inspections of the genomic context.

Currently, the UCNEbase comprises 239 UGRBs encompassing 3868 UCNEs. The number of UCNEs within a UGRB varies considerably from 134 in the ZEB2 cluster to only two in the ONECUT2 cluster (with an average of 16 and a median of 8 UCNEs). The genomic size of the identified UGRBs also varies significantly from 4.9 Mb (IRXB cluster) to ~2 kb (CPEB4 cluster).

For each UGRB, we defined a corresponding set of **UGRB associated genes** comprising all genes that fall within or overlap with genomic region spanned by the block. If the UGRB starts or ends with an intergenic UCNE, the upstream and downstream flanking genes were also included. The set UGRB-associated genes was sometimes expanded during subsequent manual curation steps, for instance by including paralogs of genes that from a paralogous UGRB.

Identification of human UCNE paralogs

Human genomic regions which exhibit significant sequence similarity to a UCNE are considered paralogs of that UCNE. In general, conserved non-coding elements have very fewer paralogs compared to protein-coding genes. However, the relatively rare cases of UCNE paralogs are highly informative with regard to the origin of UGRBs. To identify human paralogous regions of UCNEs we first split the human genome sequence (hg19) into fragments of 10’000 bp overlapping by 2’000 bp. We then used the program SSEARCH v34 from the FASTA package to compute the optimal local alignment score between each human UCNE and each other human genomic fragment. The following base substitution matrix was used for this purpose:



	A	C	G	T	N
A	5	-7	-1	-7	-1
C	-7	5	-7	-1	-1
G	-1	-7	5	-7	-1
T	-7	-1	-7	5	-1
N	-1	-1	-1	-1	-1

The gap opening and gap extension penalties (SSEARCH options –f and –g) were set to −20 and −3. At this stage, we retained all matches with E-values ≤ 0.02 for further tests. We then used PRSS3 version 3.4t26 (also from the FASTA package) to compute a base composition-adjusted E-values for each match by shuffling the DNA sequence from the corresponding region 1000 times in windows of 20 bp. All genomic regions matching a UCNE with an E-value ≤ 10^{-4} were accepted as paralogs.

Our systematic search for paralogs confirmed the expectation that most UCNEs are unique. Only 464 UCNEs have at least one human paralog. Of the 1252 paralogous regions found, only 177 were UCNEs themselves.

Identification of paralogous UGRBs

The identification of paralogous genomic regulatory blocks was mostly done by manual curation. As a minimal condition, we required that two UGRBs share at least one paralogous gene. However, most paralogous blocks also share paralogous UCNEs. In some cases, synteny across paralogous blocks was used to redefine the extension of individual UGRBs. In total, 82 UGRBs were found to have at least one paralogous UGRB forming 39 groups.

Detection of UCNE orthologs in other species

Currently, UCNEbase contains information about UCNE homologs (orthologs and paralogs) in 18 vertebrate genomes:

- four mammals:
  - ○ Mouse (mm10)
  - ○ Armadillo (dasNov1)
  - ○ Opossum (monDom5)
  - ○ Platypus (ornAna1)
- two birds:
  - ○ Chicken (galGal3)
  - ○ Zebra finch (taeGut1)
- two reptiles:
  - ○ Lizard (anoCar2)
  - ○ Painted turtle (chrPic1)
- one amphibian:
  - ○ Xenopus (xenTro3)
- five fishes:
  - ○ Fugu (fr2)
  - ○ Medaka (oryLat2)
  - ○ Stickleback (gasAcu1)
  - ○ Tetraodon (tetNig2)
  - ○ Zebrafish (danRer7)

We also identified a few UCNEs (mainly located in UTRs) that have orthologs in:

- ○ Lamprey (petMar1)
- ○ Ciona intestinalis (ci2)
- ○ Sea urchin (strPur2)
- ○ Lancelet (braFlo1).

To find homologs (orthologs and paralogs) of UCNEs in other species, we used essentially the same protocol as for finding paralogs in the human genome. We split the genomes of a given target species into fragments of 10’000 bp overlapping by 2’000 bp and then used SSEARCH to compute optimal local alignment scores between each human UCNE and each genomic fragment from the target species. For purely technical reasons, we used SSEARCH version v36.3.5 instead of version v34 for this task. We further used SSEARCH instead of PRSS3 to compute base composition-adjusted E-values for each match by shuffling the DNA sequence from the target genome 500 times in windows of 20 bp. All genomic regions matching a UCNE with an E-value $\leq 10^{-4}$ were accepted as homologs. To distinguish orthologs from paralogs we proceeded as follows. If the human UCNE had no paralogs, the matching region was immediately classified as an ortholog. Otherwise, the homologous region from the target genome was compared to all human paralogous regions of the UCNE under consideration. Base composition-adjusted E-values were compared as described above. If the sequence matched a paralogous regions with a lower E-value then the UCNE itself, it was considered a paralog. Otherwise, it was considered an ortholog.

Identification of syntenic subclusters of UCNEs in vertebrate genomes

For each human UGRB, we identified orthologous syntenic subclusters of UCNEs in other vertebrates. An orthologous syntenic subcluster is a set of UCNE orthologs that occurs as a cluster on the same chromosome, scaffold, or contig in another vertebrate genome assembly such that any two neighbouring UCNEs are separated by ≤ 0.5 Mbp. For most species we would expect only one orthologous cluster per UGRB. In reality, we often find one cluster plus a few isolated orthologous UCNEs located on sequence contigs not assigned to chromosomes. The situation could be different in the five fish species that have undergone a lineage-specific whole genome duplication.

Identification of possible target genes

Genomic regulatory blocks are generally assumed to control only one target gene belonging to the so-called *trans-dev* family. With all the information on orthologous and paralogous regions in other genomes at hand, we tried to identify the most likely target gene for each cluster. To this end, we primarily relied on a genomic context analysis approach. We reasoned that target genes will always be conserved together with UCNEs after whole genome duplication events. Based on the analysis of the gene content of paralogous UGRBs in human, and the fate of UGRB-associated UCNEs in duplicated fish genomes, we were often able to identify a single target gene. In the cases where we were left with several candidates, we gave preference to genes encoding transcription factors. In fact, the overwhelming majority of target genes uniquely defined genomic context analysis turned out to be transcription factors, most of them containing either zinc fingers or homeodomains, or both.

UGRB and UCNE nomenclature

In UCNEbase, we try to define names that carry some information about the function and genomic location of UCNEs, as well as its evolutionary relationship to other UCNEs. UCNE names are typically composed of two parts:

- UGRB name and
- element name.

For example, *DACH1_Ava* and *DACH1_Benjamin* are two UCNEs that belong to the *DACH1 cluster*.
UGRBs have the same name as their putative target genes. Elements are identified by common people's names or names from mythology. Within a UGRB, the alphabetical order of the elements reflects the linear arrangement of the elements along chromosome. Importantly, paralogous UCNEs share the same element name (e.g. *DACH1_Hana* is a paralog of *DACH2_Hana*).
For elements that are not part of a UGRB, the corresponding chromosome name replaces the block name, e.g. *chr2_Nemo*. The rule that paralogs should have the same name extends to non-clustered UCNEs (e.g. *chr10_Sherlock* is a paralog of *CPEB2_Sherlock*).
A small number of UCNEs are very close to each other and thus could be part of the same functional entity. To specifically mark such cases, UCNEs that are separated by ≤ 50 bp in both human and chicken are given the same element name, however extended by different serial numbers (e.g. *DACH1_Scheherazade_1* and *DACH1_Scheherazade_2*).

Database content

As UCNEbase is organized along two hierarchical levels, there are two types of entries, UCNEs and UGRBs. About 90% of UCNE entries are related to UGRBs. There is only one entry per UCNE or UGRB containing information for all vertebrate species covered by the resource.

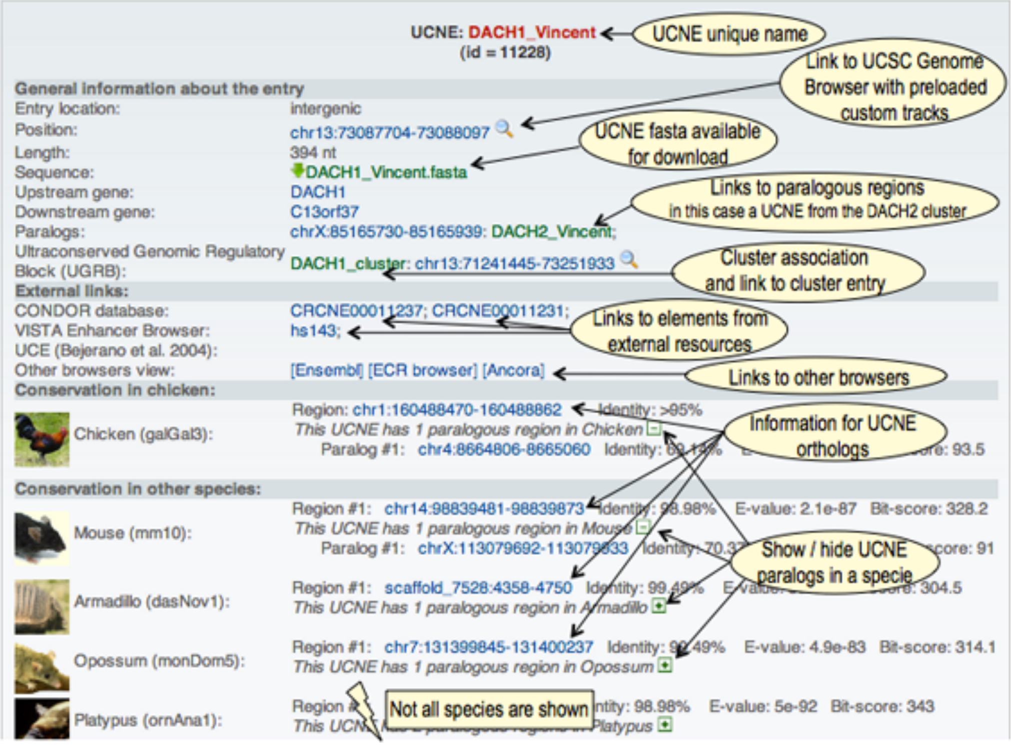
Content of a UCNE entry

Each UCNE entry has two parts, one providing detailed information relating to the human genome, and a second part providing information on homologous elements in other species. The first part of the UCNE entry contains the following data items:

- a unique name
- the location relative to the nearest genes (intergenic, intron, or UTR)
- the genome coordinates in UCSC format
- the length of the UCNE
- the sequence in FASTA
- the names of overlapping genes (for intronic and UTR-associated UCNE), or the nearest upstream

and downstream genes (for intergenic UCNE)
- a list of human paralogous UCNEs (identified by name and genomic coordinates) and other paralogous regions (identified by genomic coordinates only)
- the name of the corresponding UGRB (if any)
- cross-references to overlapping entries from the CONDOR database, VISTA Enhancer Browser and Bejerano's ultraconserved element (UCE) collection

The second part contains information about homologous regions in other vertebrates. The regions are defined by genomic coordinates, and classified as either orthologs or paralogs. In addition, the sequence identity, E-value and bitscore of the local alignments are stored. A web display of a UCNE entry is shown below:



UCNE: DACH1_Vincent
(id = 11228)

General information about the entry
Entry location: intergenic
Position: chr13:73087704-73088097
Length: 394 nt
Sequence: DACH1_Vincent.fasta
Upstream gene: DACH1
Downstream gene: C13orf37
Paralogs: chrX:85165730-85165939: DACH2_Vincent;
Ultraconserved Genomic Regulatory Block (UGRB): DACH1_cluster: chr13:71241445-73251933

External links:
CONDOR database: CRCNE00011237; CRCNE00011231;
VISTA Enhancer Browser: hs143;
UCE (Bejerano et al. 2004):
Other browsers view: [Ensembl] [ECR browser] [Ancora]

Conservation in chicken:
Chicken (galGal3): Region: chr1:160488470-160488862 Identity: >95%
This UCNE has 1 paralogous region in Chicken ⊟
Paralog #1: chr4:8664806-8665060 Identity: 69.14% E ore: 93.5

Conservation in other species:
Mouse (mm10): Region #1: chr14:98839481-98839873 Identity: 98.98% E-value: 2.1e-87 Bit-score: 328.2
This UCNE has 1 paralogous region in Mouse ⊟
Paralog #1: chrX:113079692-113079833 Identity: 70.3 core: 91
Armadillo (dasNov1): Region #1: scaffold_7528:4358-4750 Identity: 99.49% E-value ore: 304.5
This UCNE has 1 paralogous region in Armadillo ⊞
Opossum (monDom5): Region #1: chr7:131399845-131400237 Identity: 99.49% E-value: 4.9e-83 Bit-score: 314.1
This UCNE has 1 paralogous region in Opossum ⊞
Platypus (ornAna1): Region # ntity: 98.98% E-value: 5e-92 Bit-score: 343
This UCN Platypus ⊞

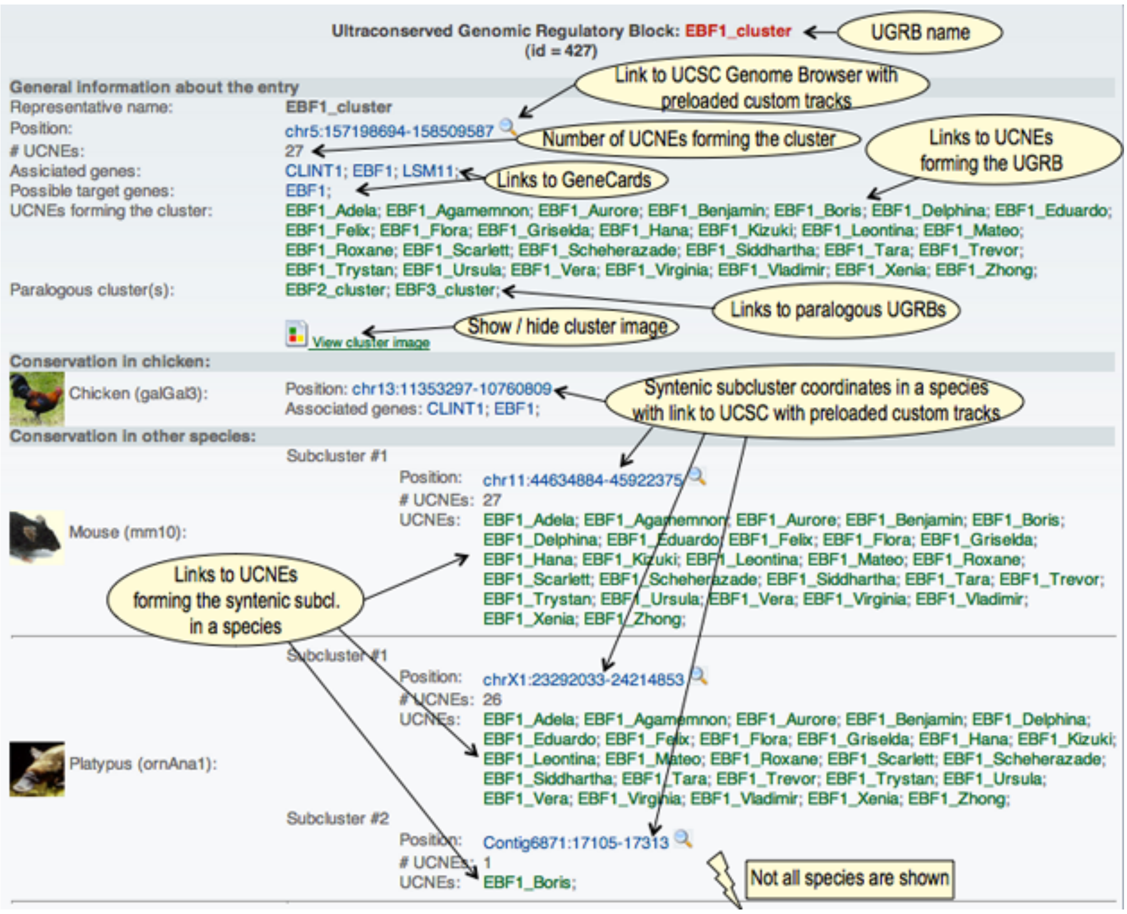
UCNE unique name
Link to UCSC Genome Browser with preloaded custom tracks
UCNE fasta available for download
Links to paralogous regions in this case a UCNE from the DACH2 cluster
Cluster association and link to cluster entry
Links to elements from external resources
Links to other browsers
Information for UCNE orthologs
Show / hide UCNE paralogs in a specie
Not all species are shown

Content of a UGRB entry

Each UGRB entry also has two parts, one providing detailed information relating to the human genome, and a second part providing information on homologous elements and clusters of elements in other species. The main section of a UGRB entry contains the following data items:

- a unique name corresponding to the most likely target gene
- the genome coordinates in UCSC format
- the number of UCNEs forming the block
- a list of all human genes associated with the block
- a list of possible target gene (in most cases only one)
- a list of all UCNEs forming the block
- a list of paralogous UGRBs

The section on sequence conservation contains synteny maps of UGRBs across multiple vertebrate genomes. A complete synteny map for a given species consists of one or several syntenic blocks referred to as "subclusters". The information associated with a subcluster comprises the genomic coordinates, the number of orthologous UCNEs, and the names of these UCNEs. An example of a web display of a UGRB entry is show below:



Ultraconserved Genomic Regulatory Block: EBF1_cluster
(id = 427)

General information about the entry

Representative name: EBF1_cluster
Position: chr5:157198694-158509587
# UCNEs: 27
Assiciated genes: CLINT1; EBF1; LSM11;
Possible target genes: EBF1;
UCNEs forming the cluster: EBF1_Adela; EBF1_Agamemnon; EBF1_Aurore; EBF1_Benjamin; EBF1_Boris; EBF1_Delphina; EBF1_Eduardo; EBF1_Felix; EBF1_Flora; EBF1_Griselda; EBF1_Hana; EBF1_Kizuki; EBF1_Leontina; EBF1_Mateo; EBF1_Roxane; EBF1_Scarlett; EBF1_Scheherazade; EBF1_Siddhartha; EBF1_Tara; EBF1_Trevor; EBF1_Trystan; EBF1_Ursula; EBF1_Vera; EBF1_Virginia; EBF1_Vladimir; EBF1_Xenia; EBF1_Zhong;
Paralogous cluster(s): EBF2_cluster; EBF3_cluster;
View cluster image

Conservation in chicken:

Chicken (galGal3): Position: chr13:11353297-10760809
Associated genes: CLINT1; EBF1;

Conservation in other species:

Mouse (mm10):
Subcluster #1
Position: chr11:44634884-45922375
# UCNEs: 27
UCNEs: EBF1_Adela; EBF1_Agamemnon; EBF1_Aurore; EBF1_Benjamin; EBF1_Boris; EBF1_Delphina; EBF1_Eduardo; EBF1_Felix; EBF1_Flora; EBF1_Griselda; EBF1_Hana; EBF1_Kizuki; EBF1_Leontina; EBF1_Mateo; EBF1_Roxane; EBF1_Scarlett; EBF1_Scheherazade; EBF1_Siddhartha; EBF1_Tara; EBF1_Trevor; EBF1_Trystan; EBF1_Ursula; EBF1_Vera; EBF1_Virginia; EBF1_Vladimir; EBF1_Xenia; EBF1_Zhong;

Platypus (ornAna1):
Subcluster #1
Position: chrX1:23292033-24214853
# UCNEs: 26
UCNEs: EBF1_Adela; EBF1_Agamemnon; EBF1_Aurore; EBF1_Benjamin; EBF1_Delphina; EBF1_Eduardo; EBF1_Felix; EBF1_Flora; EBF1_Griselda; EBF1_Hana; EBF1_Kizuki; EBF1_Leontina; EBF1_Mateo; EBF1_Roxane; EBF1_Scarlett; EBF1_Scheherazade; EBF1_Siddhartha; EBF1_Tara; EBF1_Trevor; EBF1_Trystan; EBF1_Ursula; EBF1_Vera; EBF1_Virginia; EBF1_Vladimir; EBF1_Xenia; EBF1_Zhong;
Subcluster #2
Position: Contig6871:17105-17313
# UCNEs: 1
UCNEs: EBF1_Boris;

UGRB name
Link to UCSC Genome Browser with preloaded custom tracks
Number of UCNEs forming the cluster
Links to UCNEs forming the UGRB
Links to GeneCards
Links to paralogous UGRBs
Show / hide cluster image
Syntenic subcluster coordinates in a species with link to UCSC with preloaded custom tracks
Links to UCNEs forming the syntenic subcl. in a species
Not all species are shown

User interfaces

Data access

UCNEbase provides several query mechanisms to find UCNEs and UGRBs based on different search criteria. All entries can be accessed by their chromosomal location in the human genome or by proximity to particular genes via the web links **"Browse UCNE clusters"** and **“Browse individual UCNEs”**.
The **“Advanced search”** page allows searches by additional criteria, including genomic location in other vertebrate species.
The page **"Search by external IDs"** provides access via external database IDs from the CONDOR database, VISTA Enhancer Browser, and Bejerano’s UCEs collection.
UCNEbase also provides three fully hyperlinked summary tables:

- **Paralogous clusters** - containing a list of paralogous UGRBs
- **Paralogous UCNEs** - containing a list paralogous UCNEs
- **Species cluster summary** - showing the numbers of conserved UCNEs for each UGRB in all species.

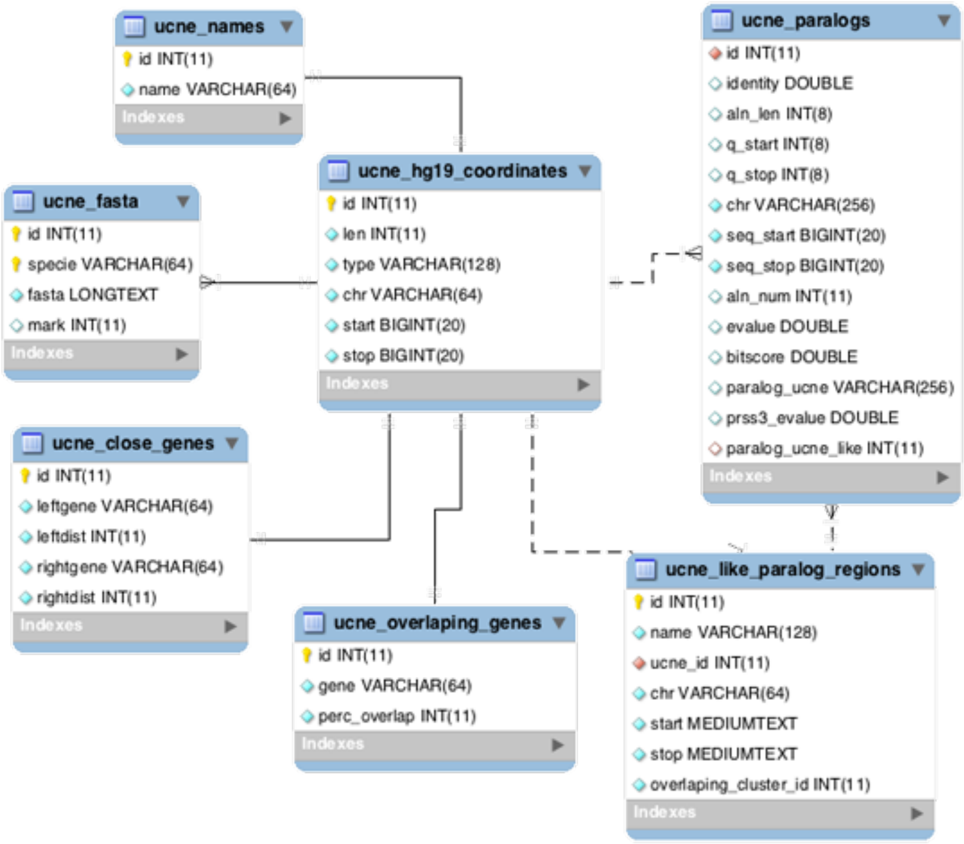
Data visualization

UCNEbase relies on the UCSC Genome Browser for data visualization. A large part of the information content is available as custom track files. This has the principle advantage that information from UCNEbase can be explored together with a great variety of genome annotations from other sources. The UCSC browser also serves as a navigation platform. All data items from UCNEbase that can be displayed in a browser window are back-linked to the corresponding UCNE and UGRB entries.
For the human genome, UCNEbase provides custom tracks for UCNEs, UGRBs, UCNE paralogs, CONDOR CNEs, Vista elements, and UCEs from Bejerano’s collections. In addition, there is a group of tracks showing the subset of UCNEs conserved in different species.

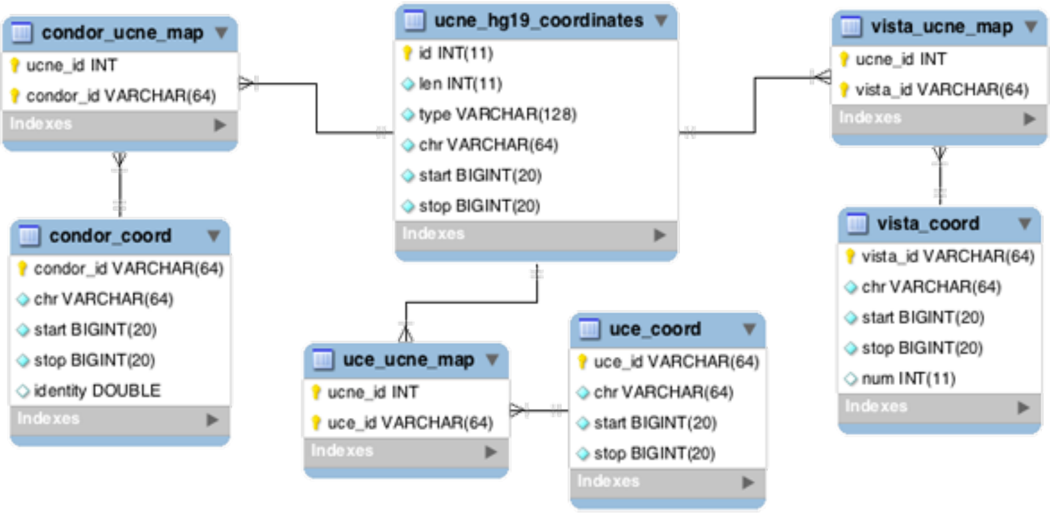
For non-human species, there are tracks for UCNE orthologs, UCNE paralogs, and subclusters of UCNEs corresponding to human UGRBs.

UCNEbase schema diagram (ER)

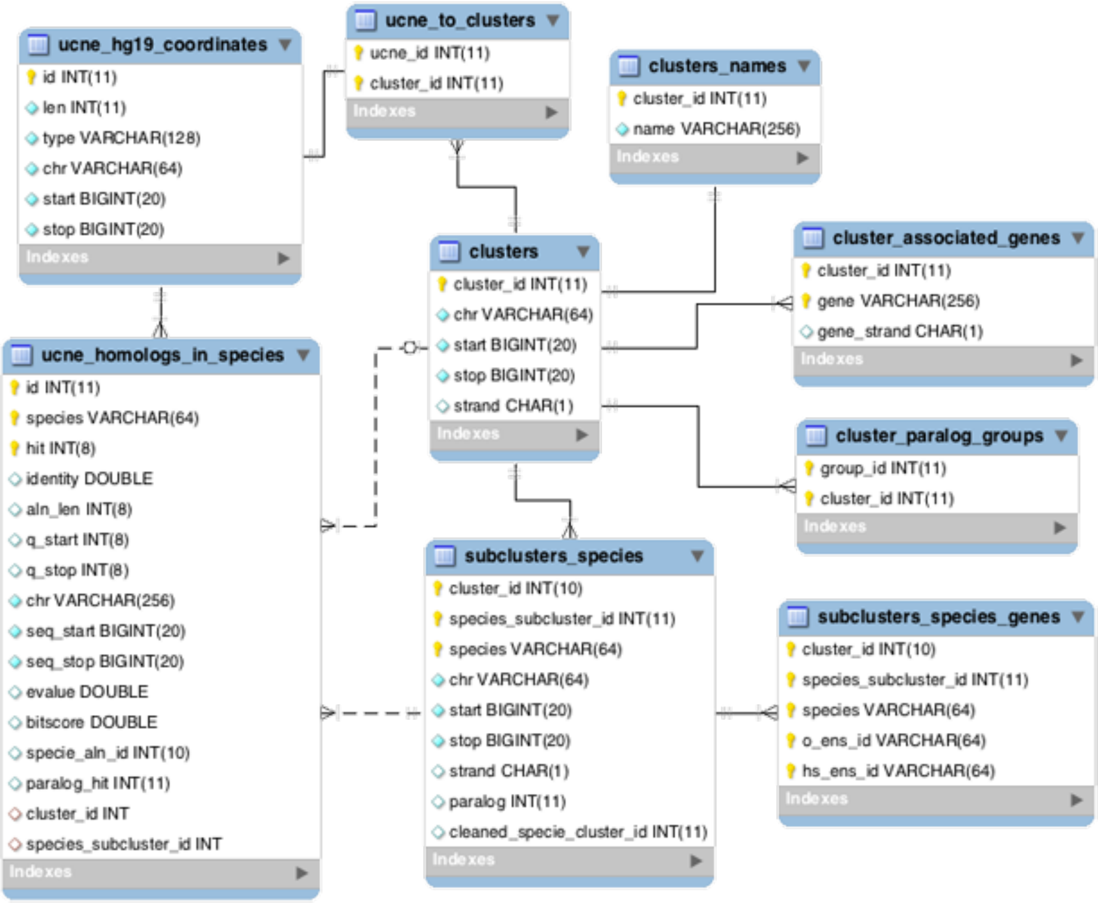
UCNE general information



UCNE cross-references



Information on UCNE clusters (UGRBs), UCNE homologs in species and UCNE subclusters in species



ucne_hg19_coordinates
id INT(11)
len INT(11)
type VARCHAR(128)
chr VARCHAR(64)
start BIGINT(20)
stop BIGINT(20)
Indexes

ucne_to_clusters
ucne_id INT(11)
cluster_id INT(11)
Indexes

clusters_names
cluster_id INT(11)
name VARCHAR(256)
Indexes

clusters
cluster_id INT(11)
chr VARCHAR(64)
start BIGINT(20)
stop BIGINT(20)
strand CHAR(1)
Indexes

cluster_associated_genes
cluster_id INT(11)
gene VARCHAR(256)
gene_strand CHAR(1)
Indexes

ucne_homologs_in_species
id INT(11)
species VARCHAR(64)
hit INT(8)
identity DOUBLE
aln_len INT(8)
q_start INT(8)
q_stop INT(8)
chr VARCHAR(256)
seq_start BIGINT(20)
seq_stop BIGINT(20)
evalue DOUBLE
bitscore DOUBLE
specie_aln_id INT(10)
paralog_hit INT(11)
cluster_id INT
species_subcluster_id INT
Indexes

cluster_paralog_groups
group_id INT(11)
cluster_id INT(11)
Indexes

subclusters_species
cluster_id INT(10)
species_subcluster_id INT(11)
species VARCHAR(64)
chr VARCHAR(64)
start BIGINT(20)
stop BIGINT(20)
strand CHAR(1)
paralog INT(11)
cleaned_specie_cluster_id INT(11)
Indexes

subclusters_species_genes
cluster_id INT(10)
species_subcluster_id INT(11)
species VARCHAR(64)
o_ens_id VARCHAR(64)
hs_ens_id VARCHAR(64)
Indexes