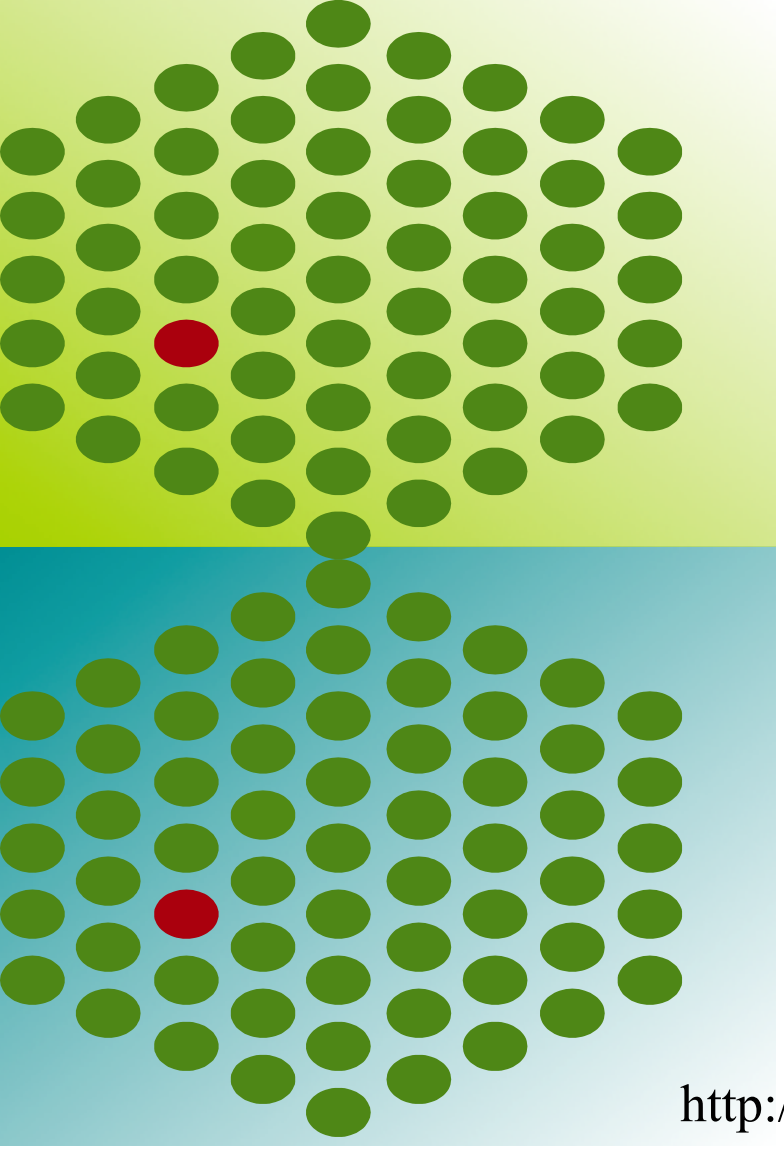


Ensembl Compara Perl API

Stephen Fitzgerald

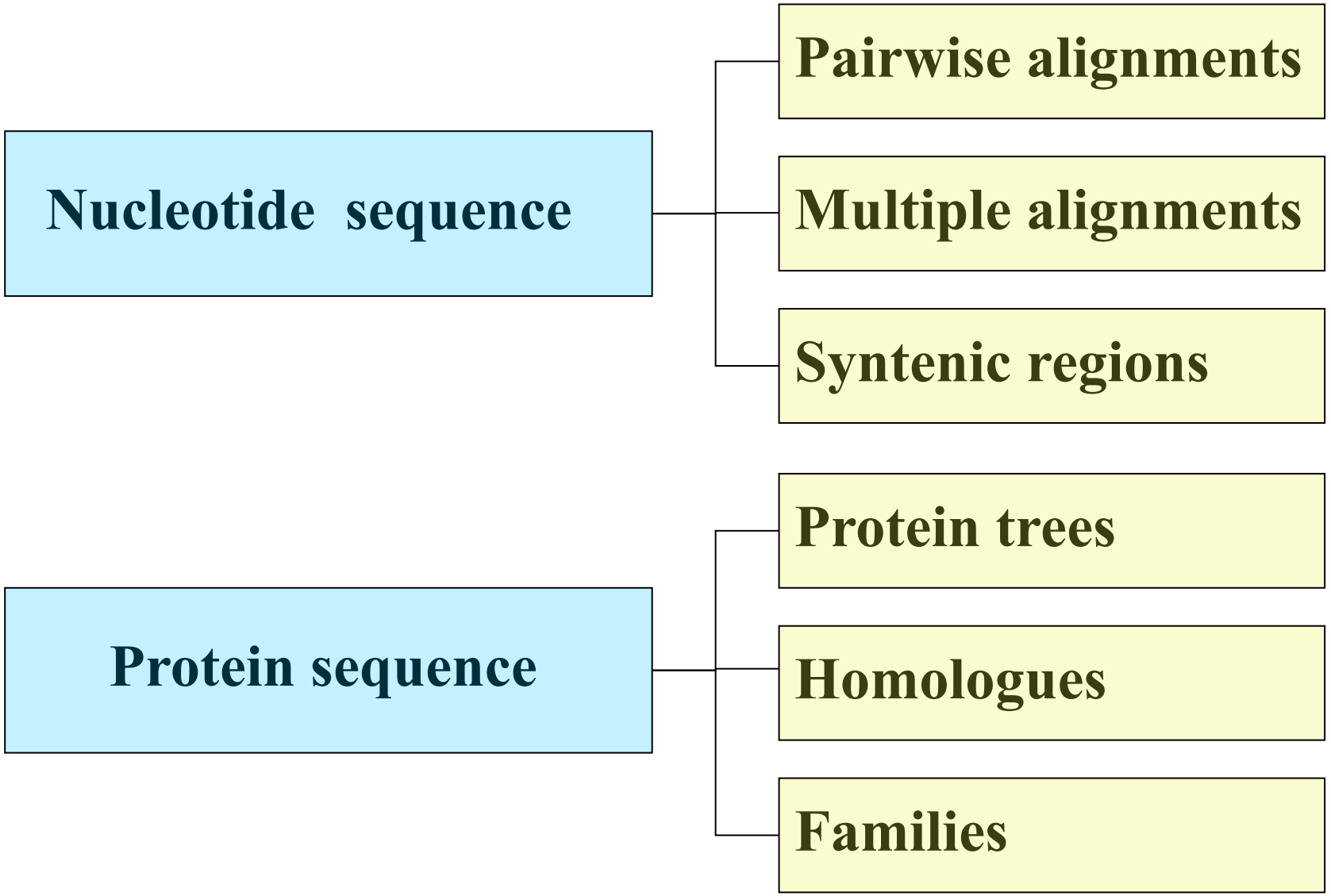
EBI - Wellcome Trust Genome Campus, UK

http://www.ebi.ac.uk/~stephenf/Workshops/Cam_nov_2012/

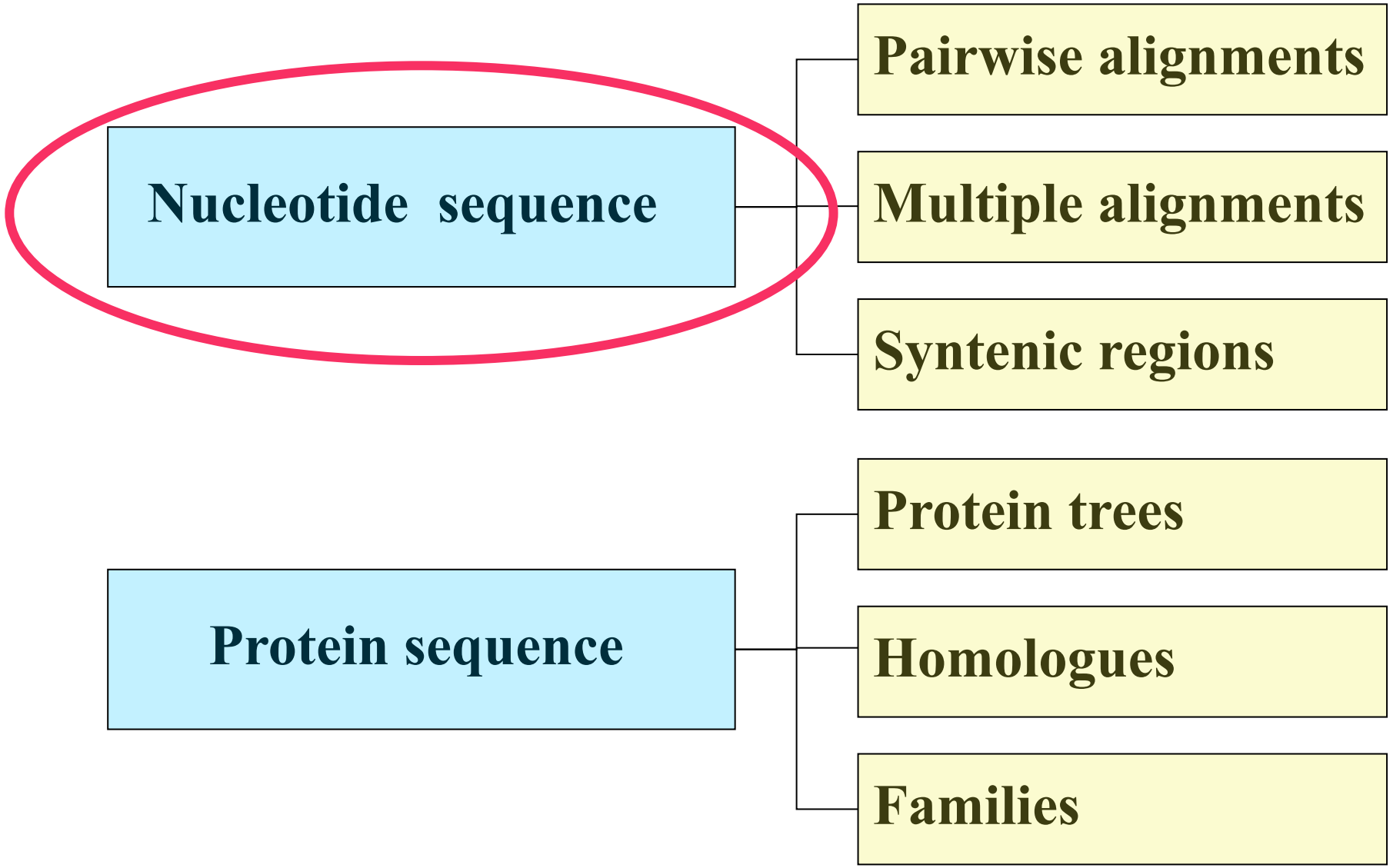
EMBL-EBI

wellcome trust sanger institute

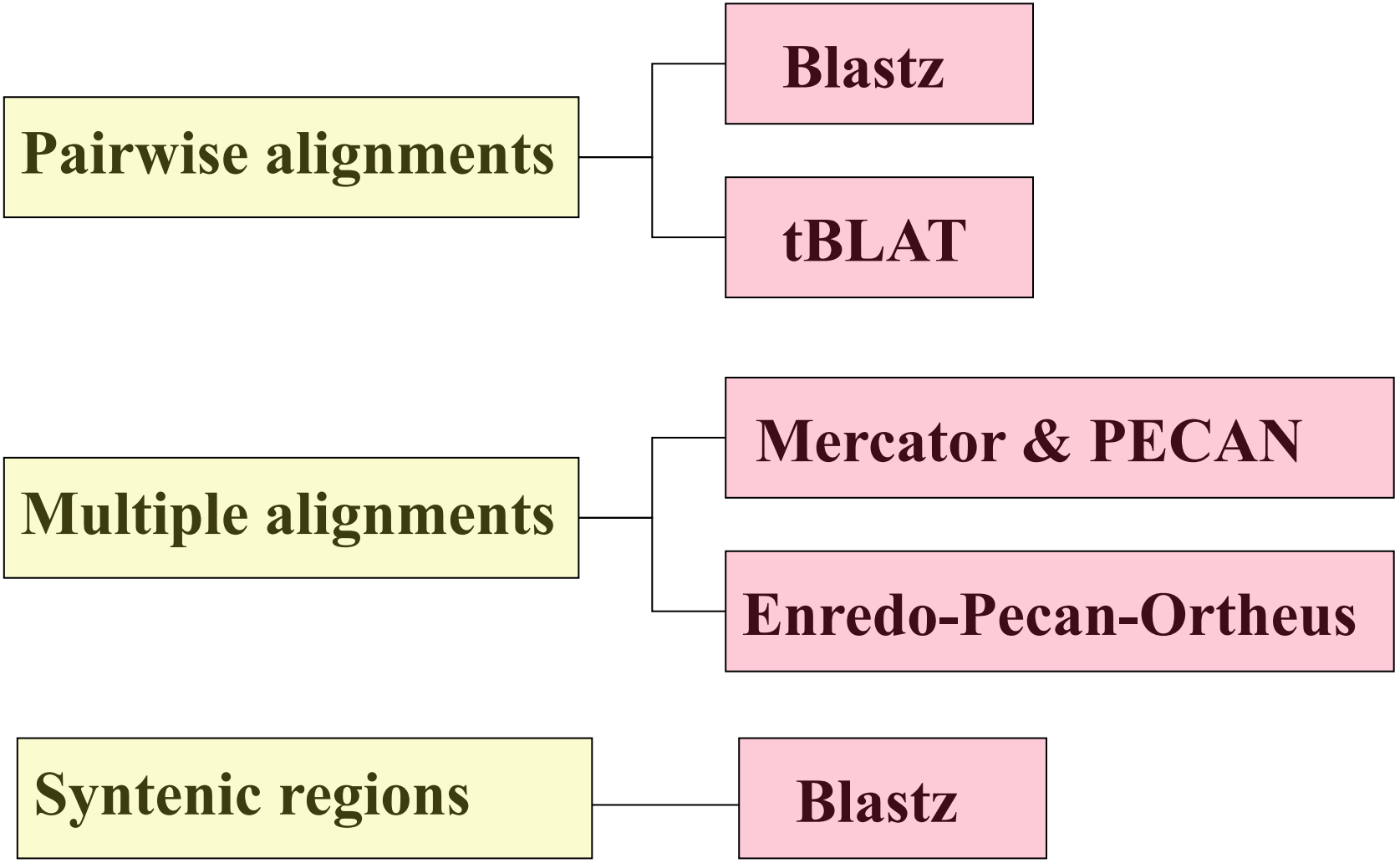
Sequence types and outputs



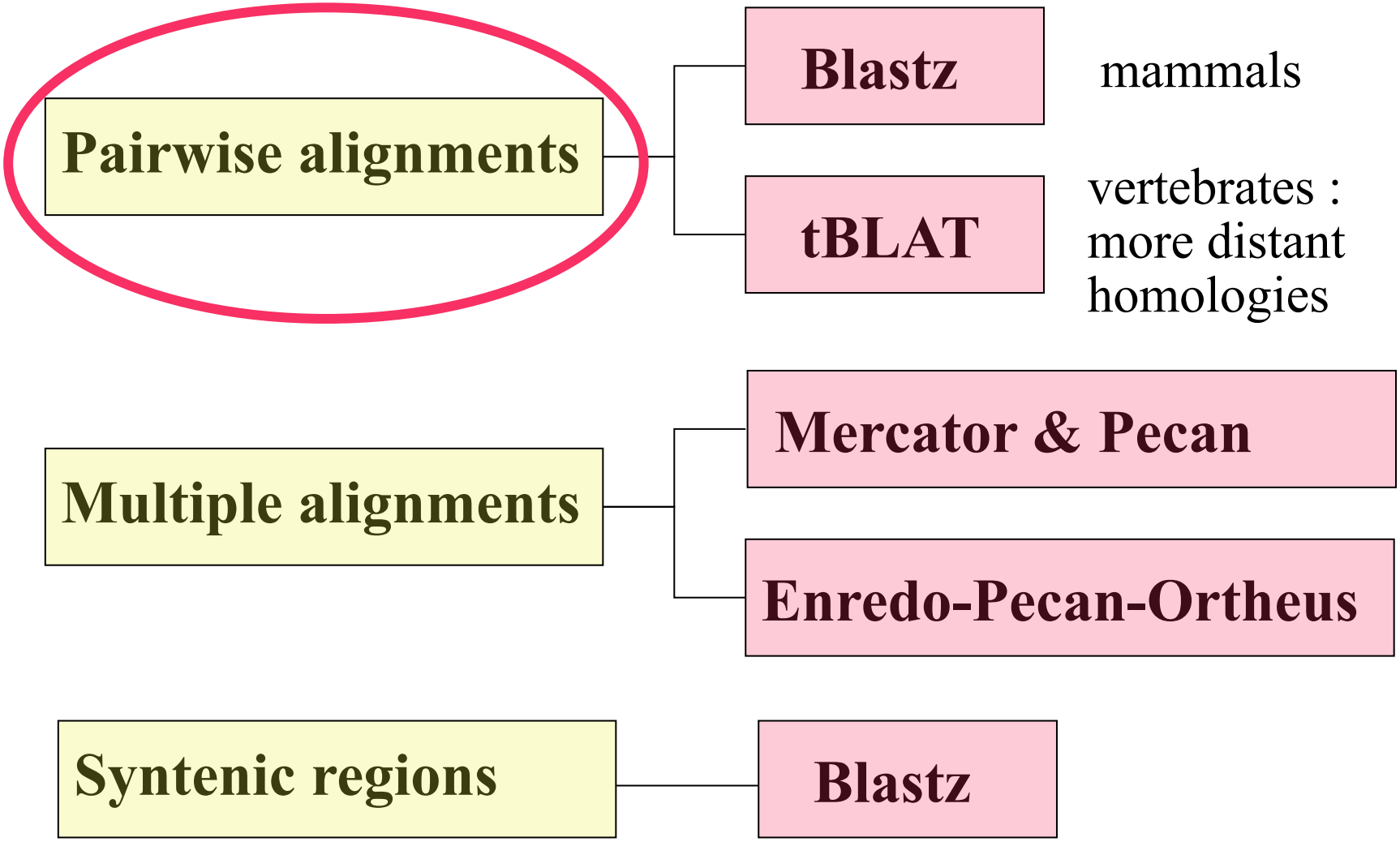
Sequence types and outputs



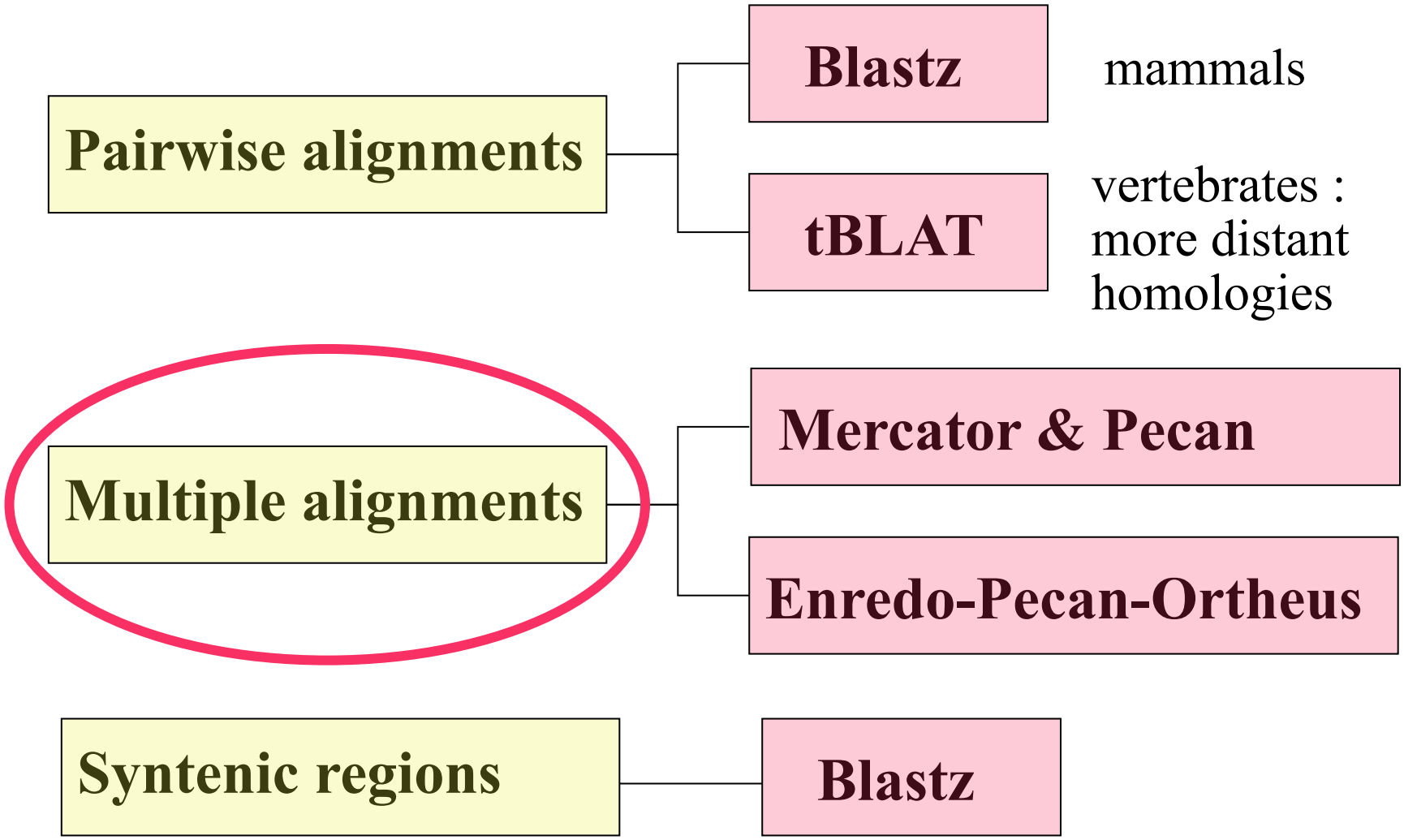
Pipelines and outputs for nucleotide sequence



Pipelines and outputs for nucleotide sequence



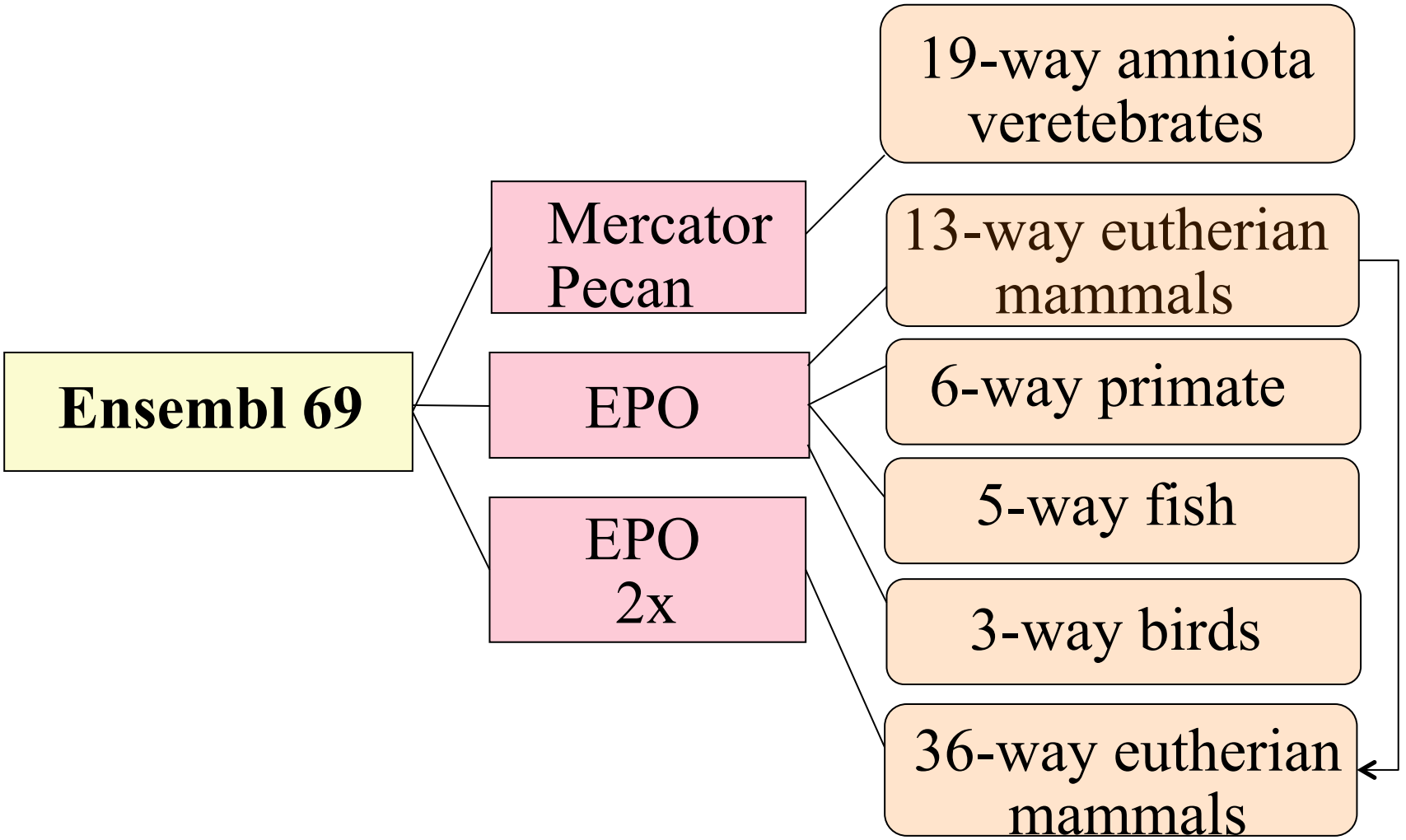
Pipelines and outputs for nucleotide sequence

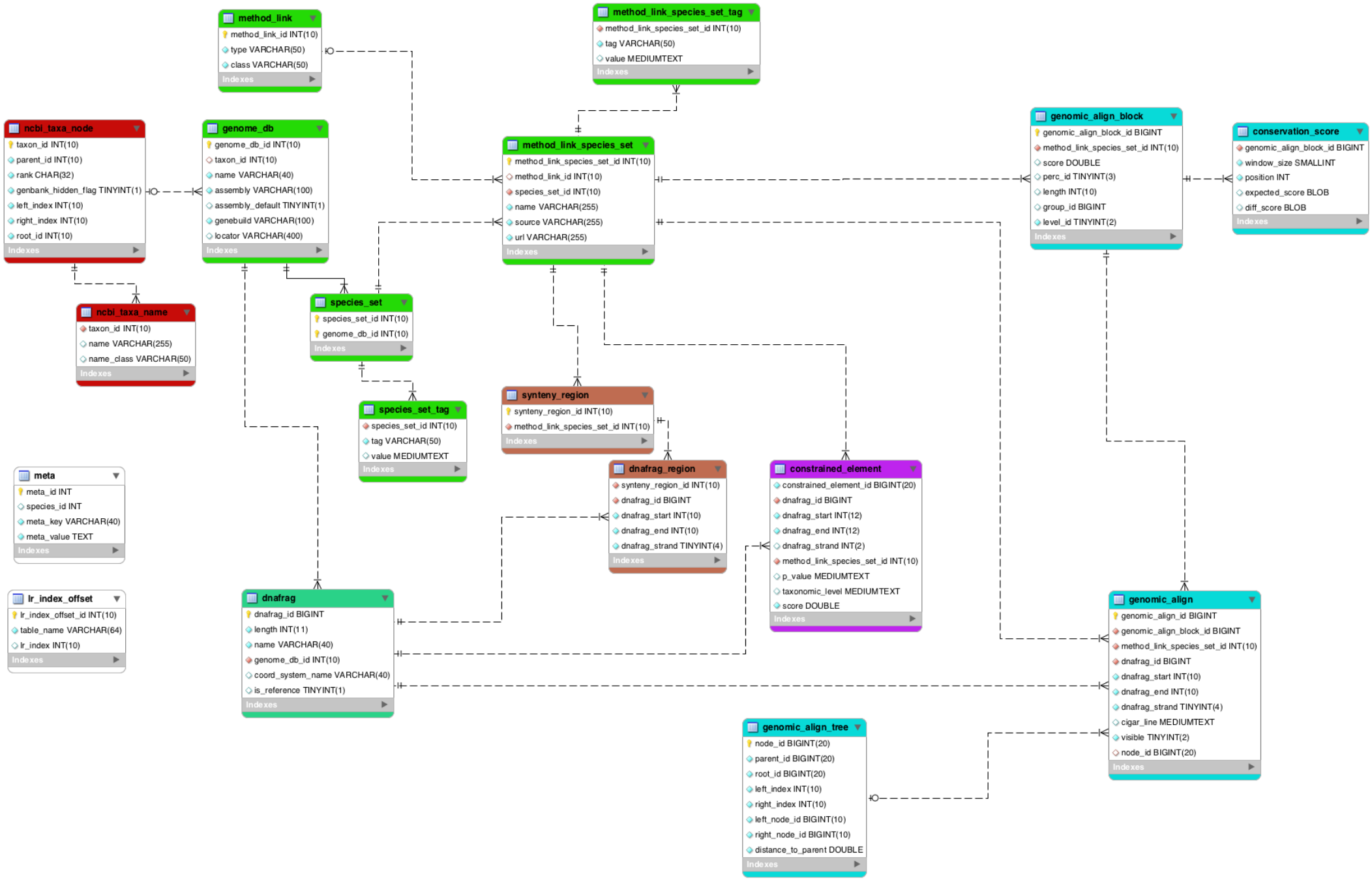


Generating multiple alignments

- **We build homology maps for multiple alignments using**
  - **Mercator** : A graph based program, which uses exon sequences as anchors. It does not allow for the alignment of duplicated regions in a genome.
  - **Enredo** : Also graph based. Use conserved regions from pairwise blastz alignments of whole genomes as anchors. It does allow for the alignment of duplicated regions.
- **Alignment is done using Pecan.**
- **Ancestral sequences are generated using Ortheus.**

MSA in Compara 69





Genomic alignments schema

Exercises – GenomeDB and DnaFrag

- A GenomeDB is used to link the Compara database to each of the Core species databases.

➢ Print the name, assembly version and genebuild version for all the GenomeDBs in the compara db

- A DnaFrag represents a top-level SeqRegion in the Compara database.

➢ Print all the DnaFrags for chimp

Exercises – MethodLinkSpeciesSet

- The MethodLinkSpeciesSet is a central component in the Compara database, it stores information connecting the various analyses (method_link_type) with a set of species (species_set).

➢ Print the total number of MethodLinkSpeciesSet entries stored in the database.

➢ Print a unique list of method_link_types and a count of their number in the database.

➢ Print a list of the species and their internal ids (dbIDs) for the 12 eutherian mammal EPO alignments

Alignments are stored in the genomic_align and genomic_align_block tables

A small example :

```
gorilla_gorilla/MT/935-953    gacat-ttaactaaaac-ccc
macaca_mulatta/MT/1469-1488   aacatcttaactaaacg-ccc
pan_troglodytes/MT/934-953    gatac-ttaacttaaaccccc
pongo_pygmaeus/MT/940-958     actac-ctaactaaaac-ccc
homo_sapiens/MT/1516-1534     gacat-ttaactaaaac-ccc
                                 *   ***** **   ***
```

```
GACATTTAACTAAAACCCC      5MD11MD3M
AACATCTTAACTAAACGCCC     17MD3M
GATACTTAACTTAAACCCCC     5MD15M
ACTACCTAACTAAAACCCC      5MD11MD3M
GACATTTAACTAAAACCCC      5MD11MD3M
```

5 genomic_align entries
1 genomic_align_block

Sequences from core

Exercises - GenomicAlignBlock

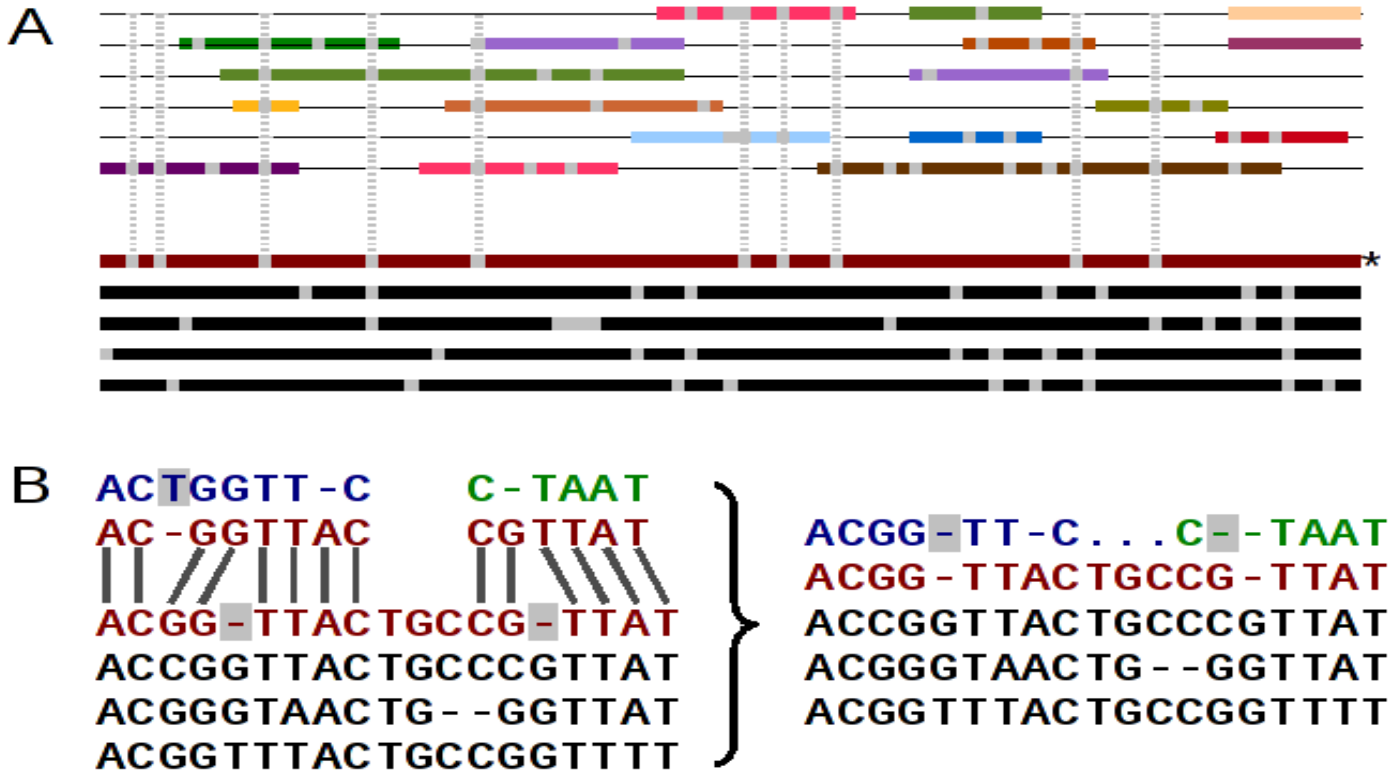
- A GenomicAlignBlock represents an alignment between two or more regions of genomic DNA. Within these blocks every region of genomic DNA is represented by a GenomicAlign object.

➢ Print the LASTZ-NET alignments for pig chromosome 15 with cow (using pig coordinates 105734307 and 105739335).

➢ Change the above example so that it prints the 13-way eutherian mammal (EPO) multiple alignments.

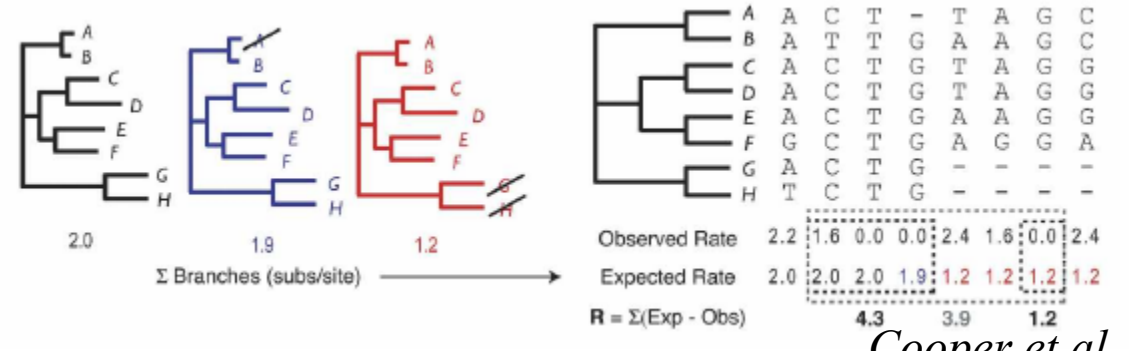
Adding low-coverage (2X) genomes

- Low coverage genomes cannot be fully assembled
- Resulting assembly is too scattered to be used with Enredo
- Run EPO on high-coverage genomes only
- Map 2X genomes using pairwise alignments



Gerp Constrained Elements

- Stretches of the alignment with a high conservation



Cooper et al. Genome Research, 2005

- Constrained elements and coding exons
  - 74% of coding exons are associated with constr. elem.
  - 22% of constr. elem. are associated with coding exons

Exercises – GenomicAlignBlock (Constrained elements)

- A Constrained Elements represent regions in the multiple alignment which appear to be under functional constraint.

➢ Print the constrained element alignments from the above pig locus (use the constrained elements generated from the EPO_LOW_COVERAGE mammals alignments)

Exercises - Synteny

- Print the pig-cow synteny map using pig chromosome 15 as a reference