Nucleotide sequence alignments in Compara

Stephen Fitzgerald
stephenf@ebi.ac.uk

What is Ensembl Compara?

A single database which contains precalculated comparative genomics data

Access via perl API and mysql

A production system for generating that database

(not in this presentation)

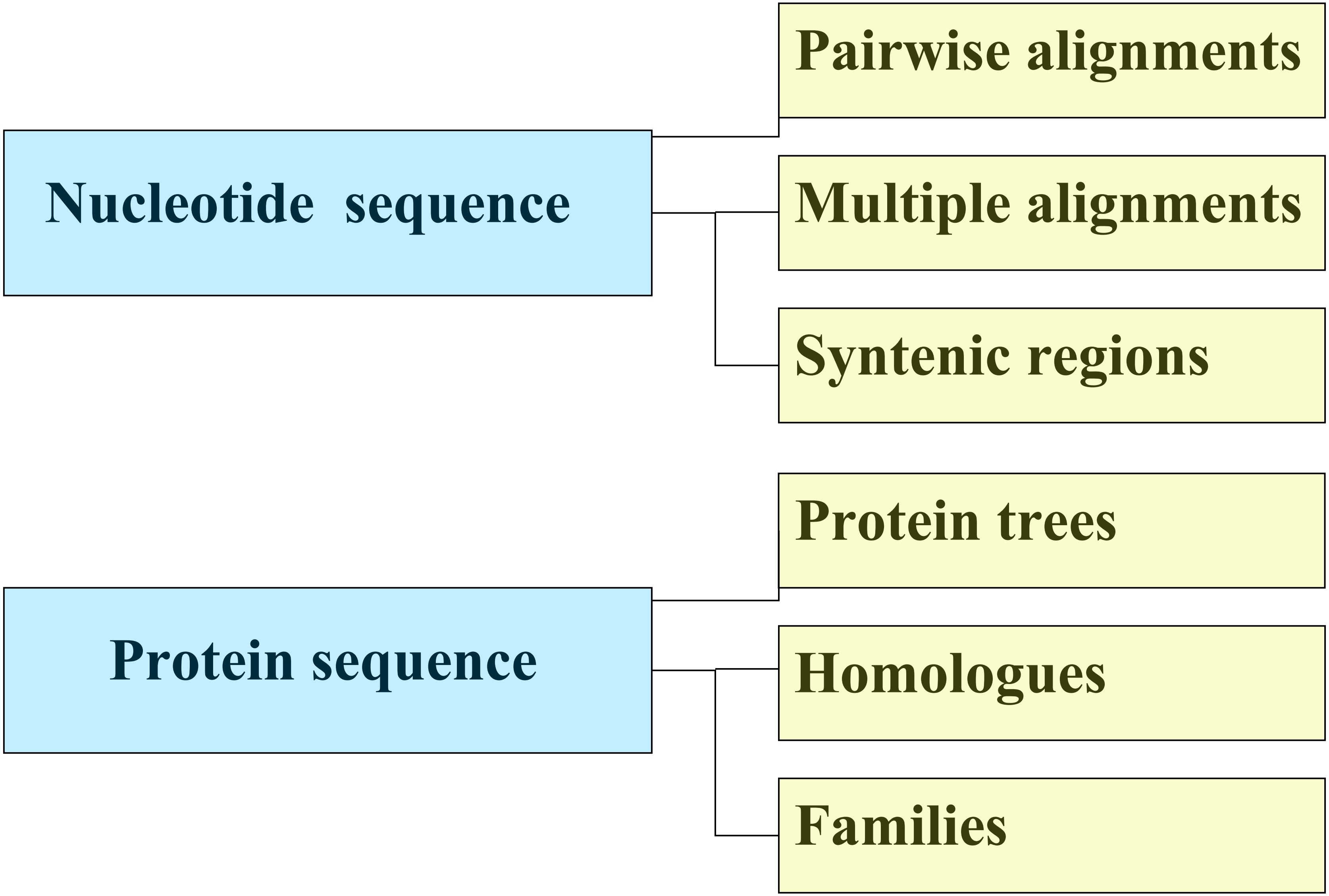
Compara database & the Ensembl core databases

Since there is minimal primary data inside Compara, to gain full access to the data external links with core DBs must be re-established

Example: compara_63 must be linked with the
Ensembl core_63 databases

Proper REGISTRY configuration is critical.
load_registry_from_db is probably the best choice here

Sequence types and outputs


Nucleotide sequence

Pairwise alignments

Multiple alignments

Syntenic regions

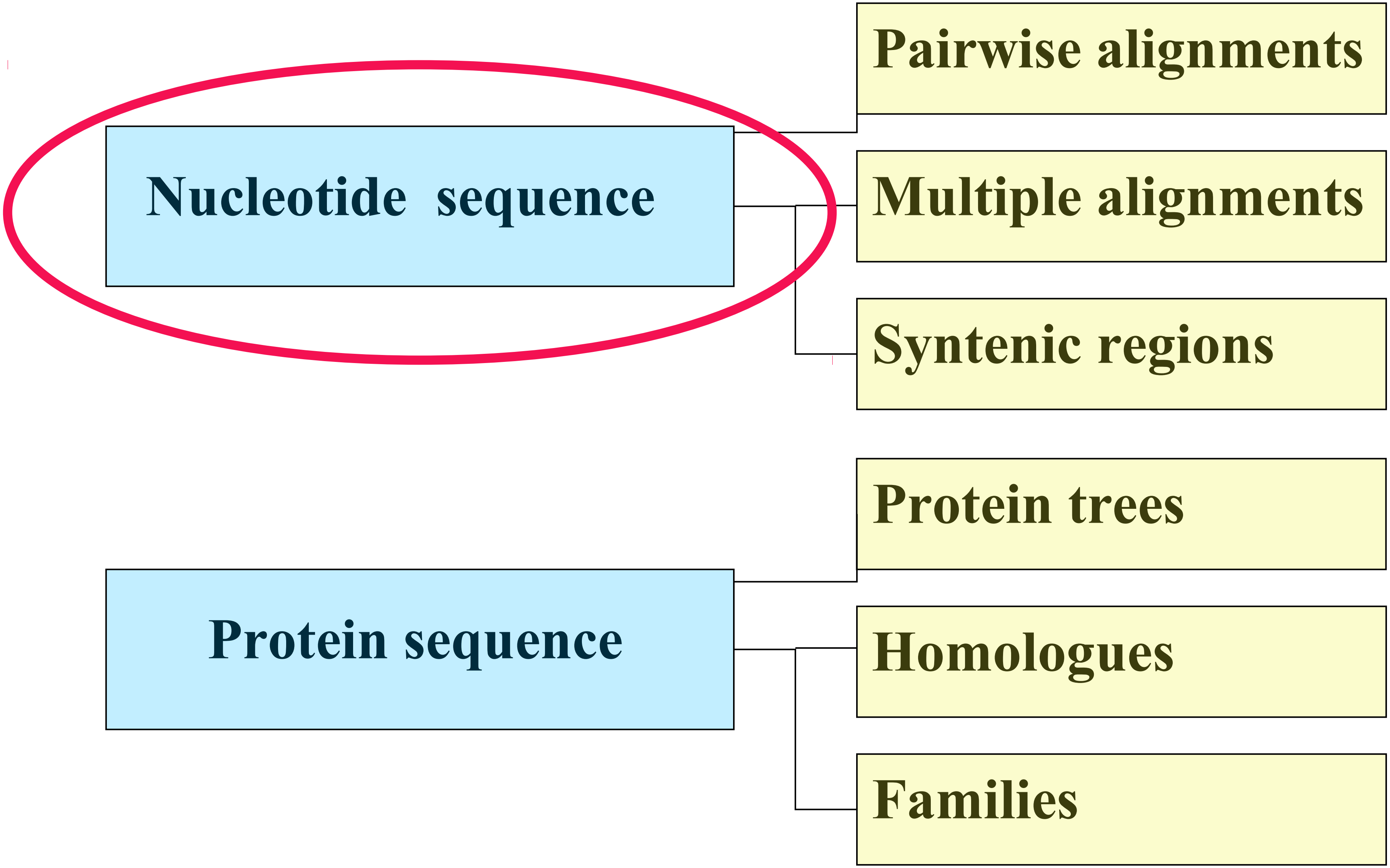
Protein sequence

Protein trees

Homologues

Families

Sequence types and outputs



Nucleotide sequence

Pairwise alignments

Multiple alignments

Syntenic regions

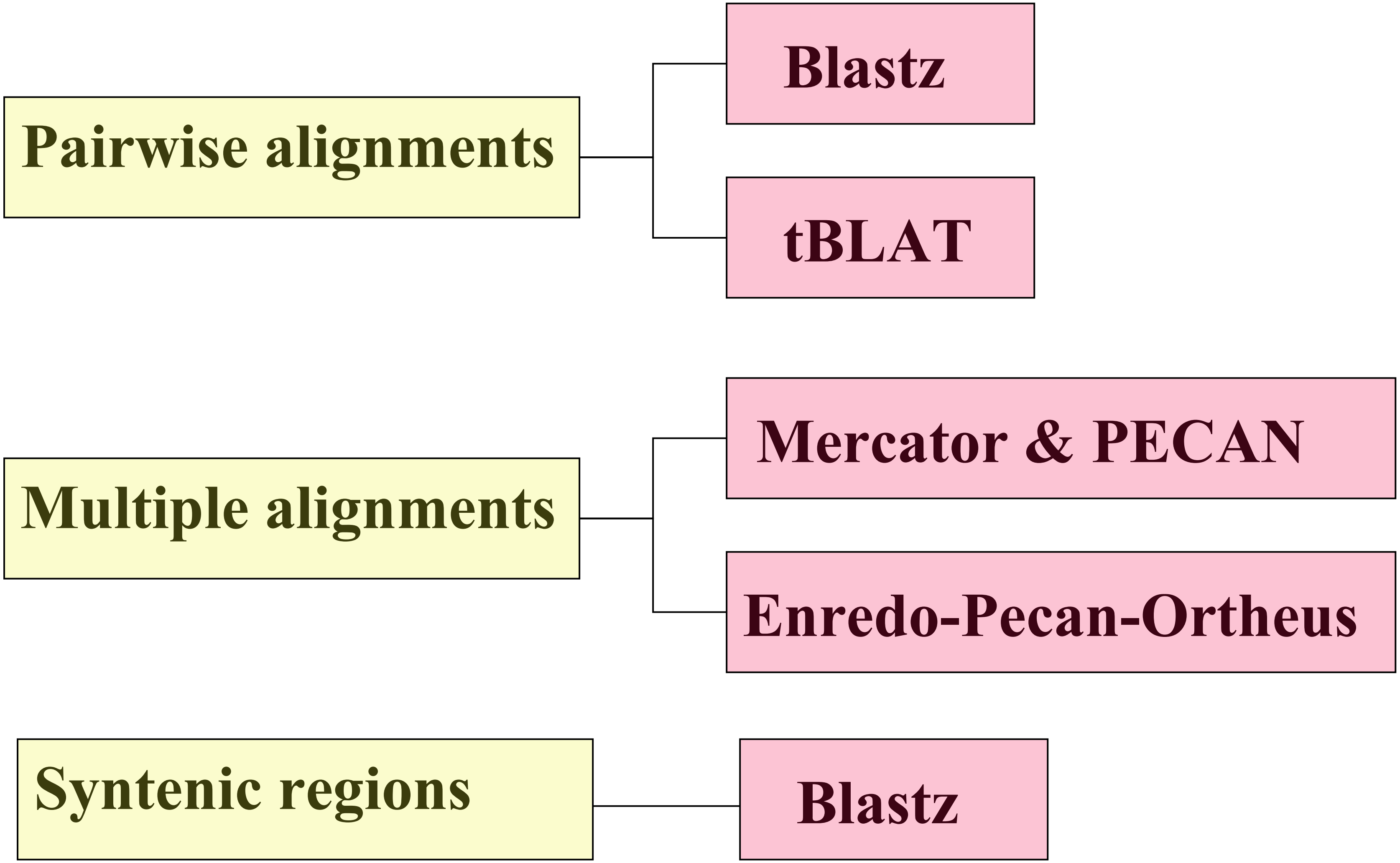
Protein sequence

Protein trees

Homologues

Families

Pipelines and outputs for nucleotide sequence



Pairwise alignments

Blastz

tBLAT

Multiple alignments

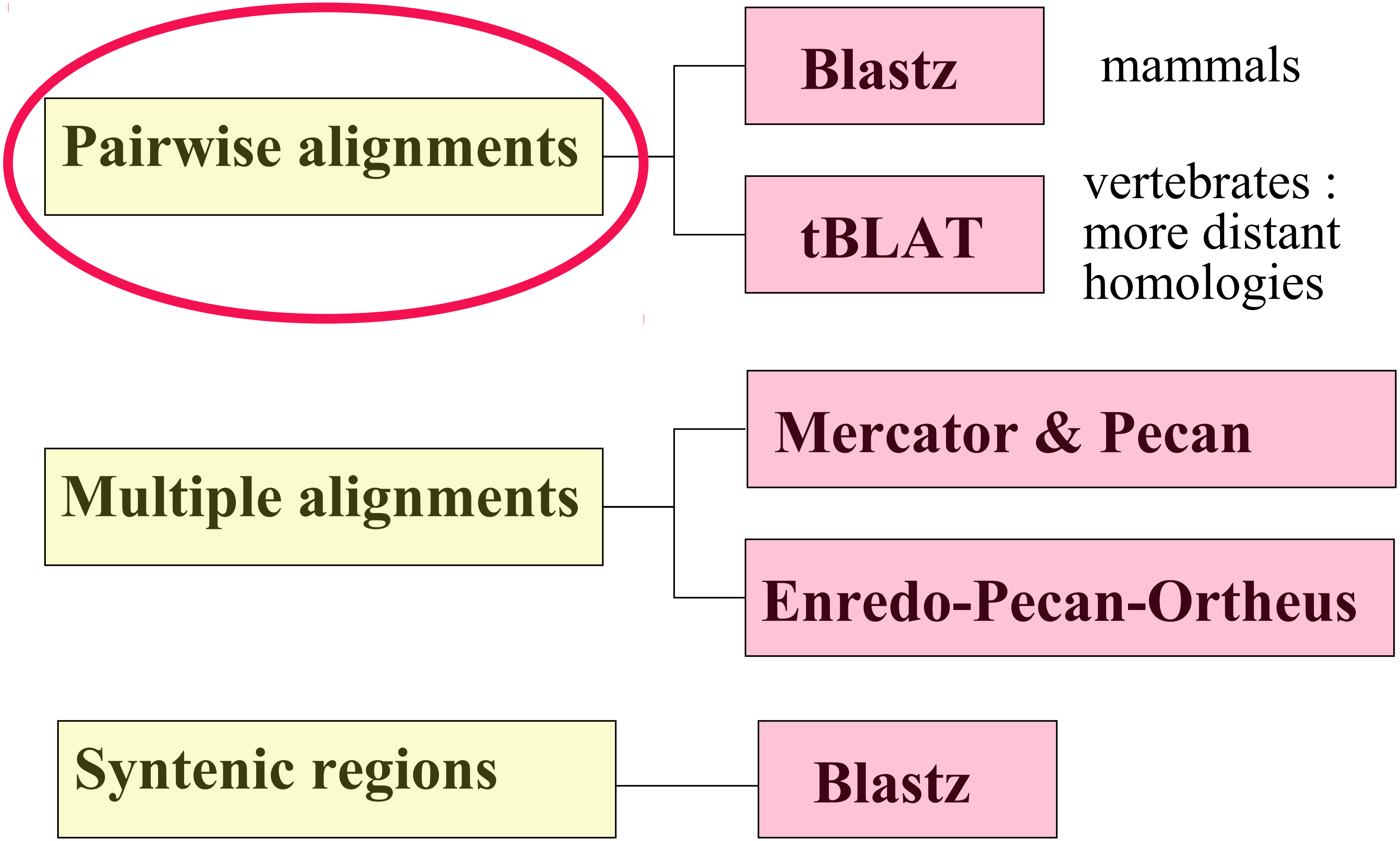
Mercator & PECAN

Enredo-Pecan-Ortheus

Syntenic regions

Blastz

Pipelines and outputs for nucleotide sequence



Pairwise alignments

Blastz

mammals

tBLAT

vertebrates : more distant homologies

Multiple alignments

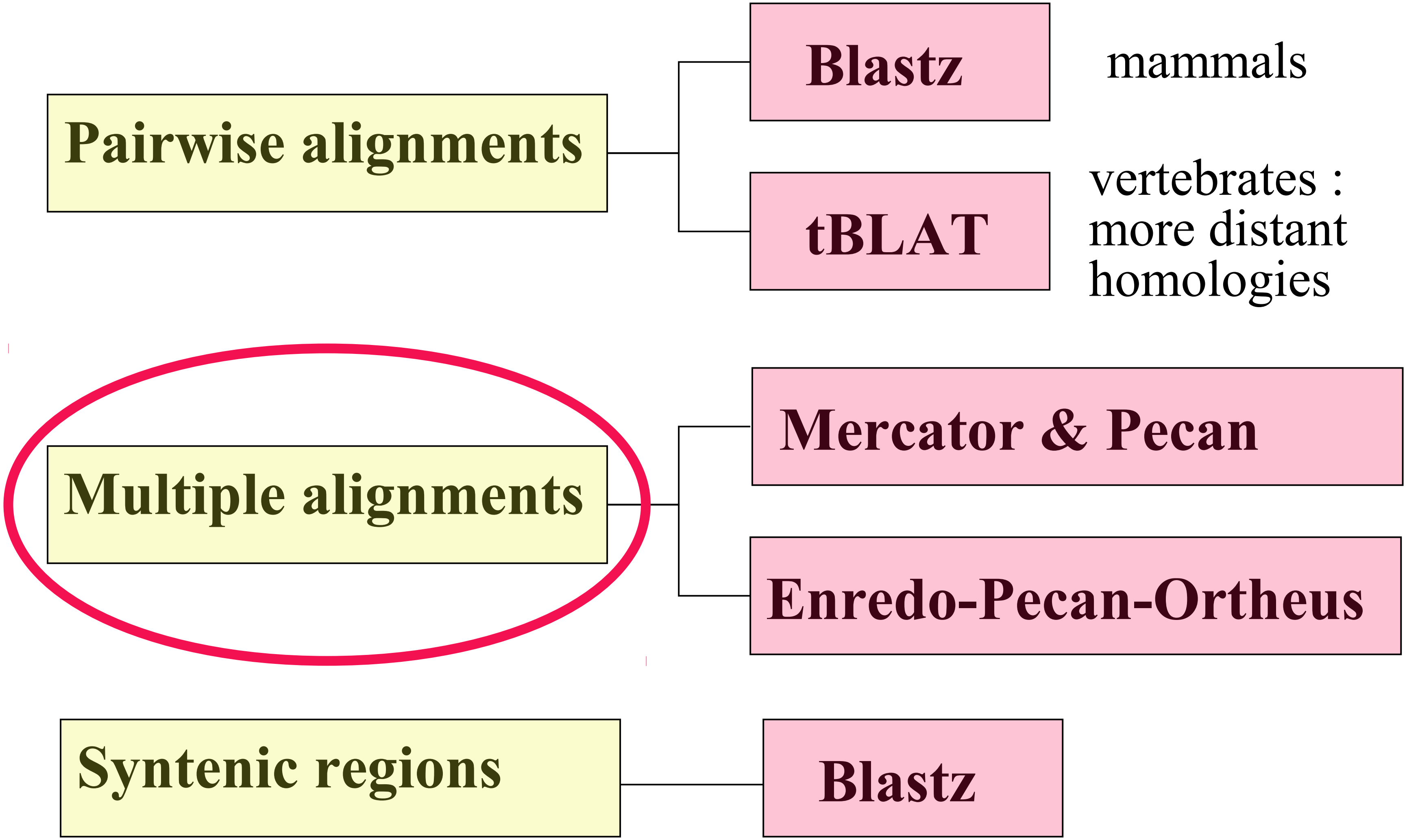
Mercator & Pecan

Enredo-Pecan-Ortheus

Syntenic regions

Blastz

Pipelines and outputs for nucleotide sequence



Pairwise alignments

Blastz

mammals

tBLAT

vertebrates : more distant homologies

Multiple alignments

Mercator & Pecan

Enredo-Pecan-Ortheus

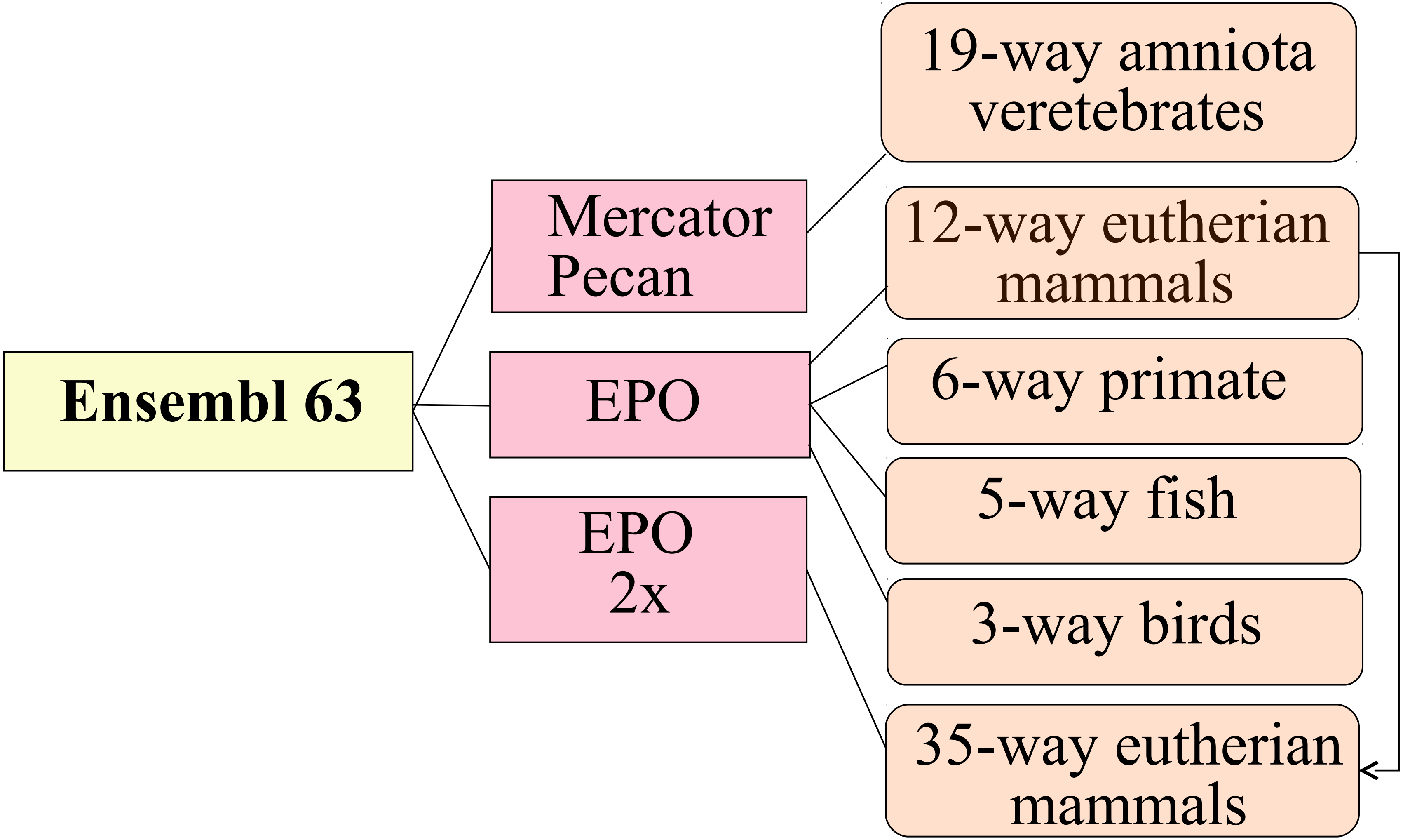
Syntenic regions

Blastz

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 63



Ensembl 63

Mercator Pecan

EPO

EPO 2x

19-way amniota veretebrates

12-way eutherian mammals

6-way primate

5-way fish

3-way birds

35-way eutherian mammals

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-ttaacttaaccccc
pongo_pygmaeus/MT/940-958    actac-ctaactaaaac-ccc
homo_sapiens/MT/1516-1534    gacat-ttaactaaaac-ccc
                                *   ***** **   ***
```

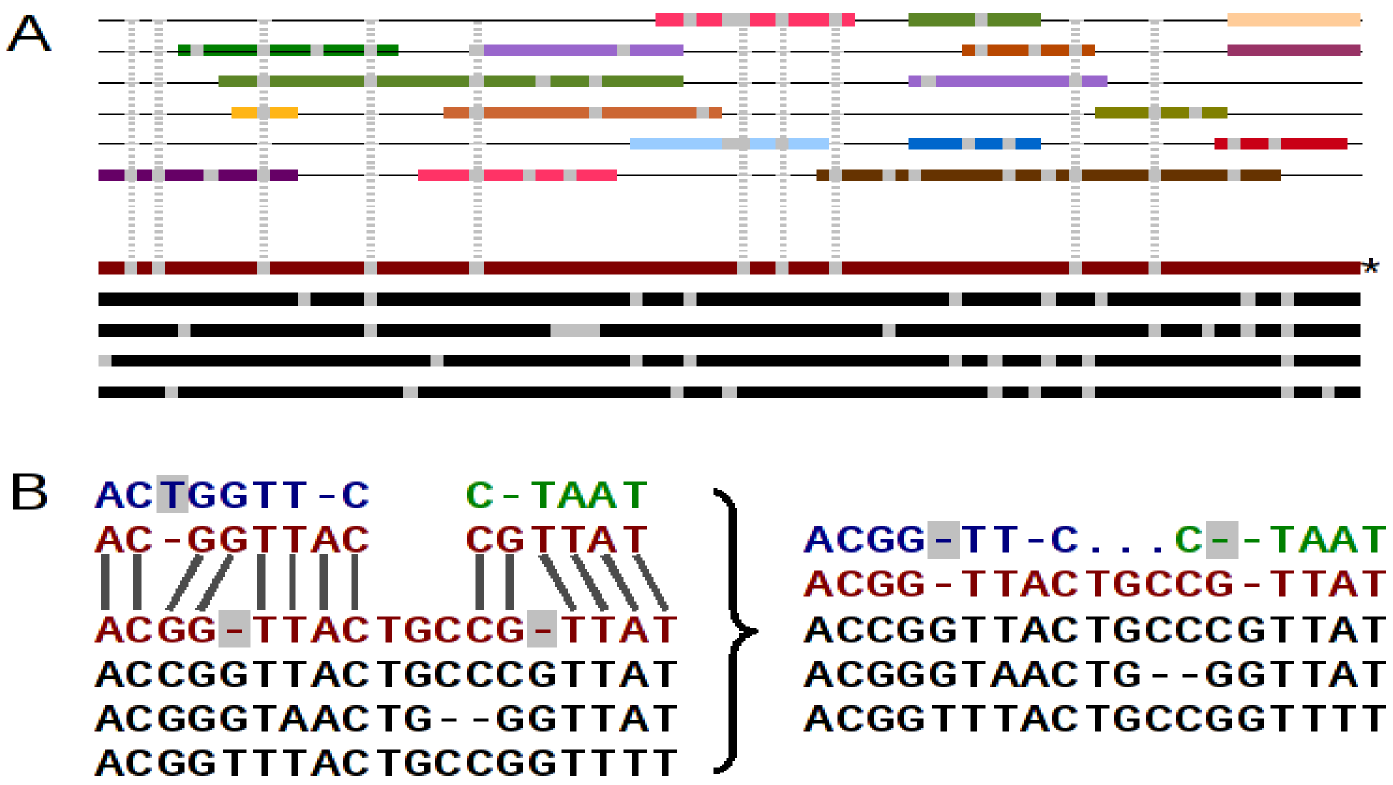
```
GACATTTAACTAAAACCCC     5MD11MD3M
AACATCTTAACTAAACGCCC    17MD3M
GATACTTAACTTAAACCCCC    5MD15M
ACTACCTAACTAAAACCCC     5MD11MD3M
GACATTTAACTAAAACCCC     5MD11MD3M
```

5 genomic_align entries
1 genomic_align_block

Sequences from core

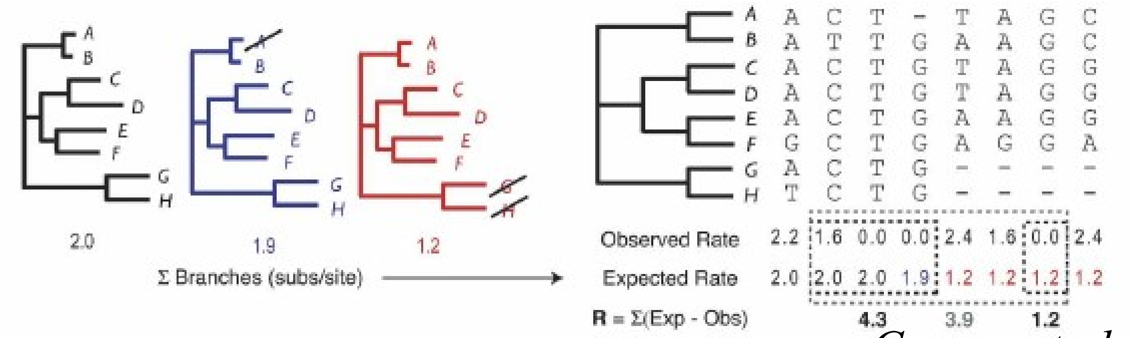
Adding low-coverage (2X) genomes

- 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

ensembl-dev mailing list and HelpDesk

- ensembl-dev mailing list is great for questions around the API and the DB (ensembl-dev@ebi.ac.uk)
- HelpDesk is very helpful
- Give detailed info on what you are trying to do
- Check that you have the modules installed ($PERL5LIB pointing to them)

- Ensembl Compara Team:
  - Javier
  - Kathryn
  - Matthieu
  - Leo
  - Stephen
  - Miguel