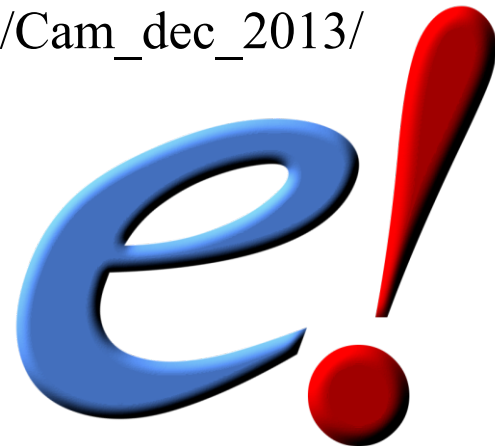


Ensembl Compara Perl API

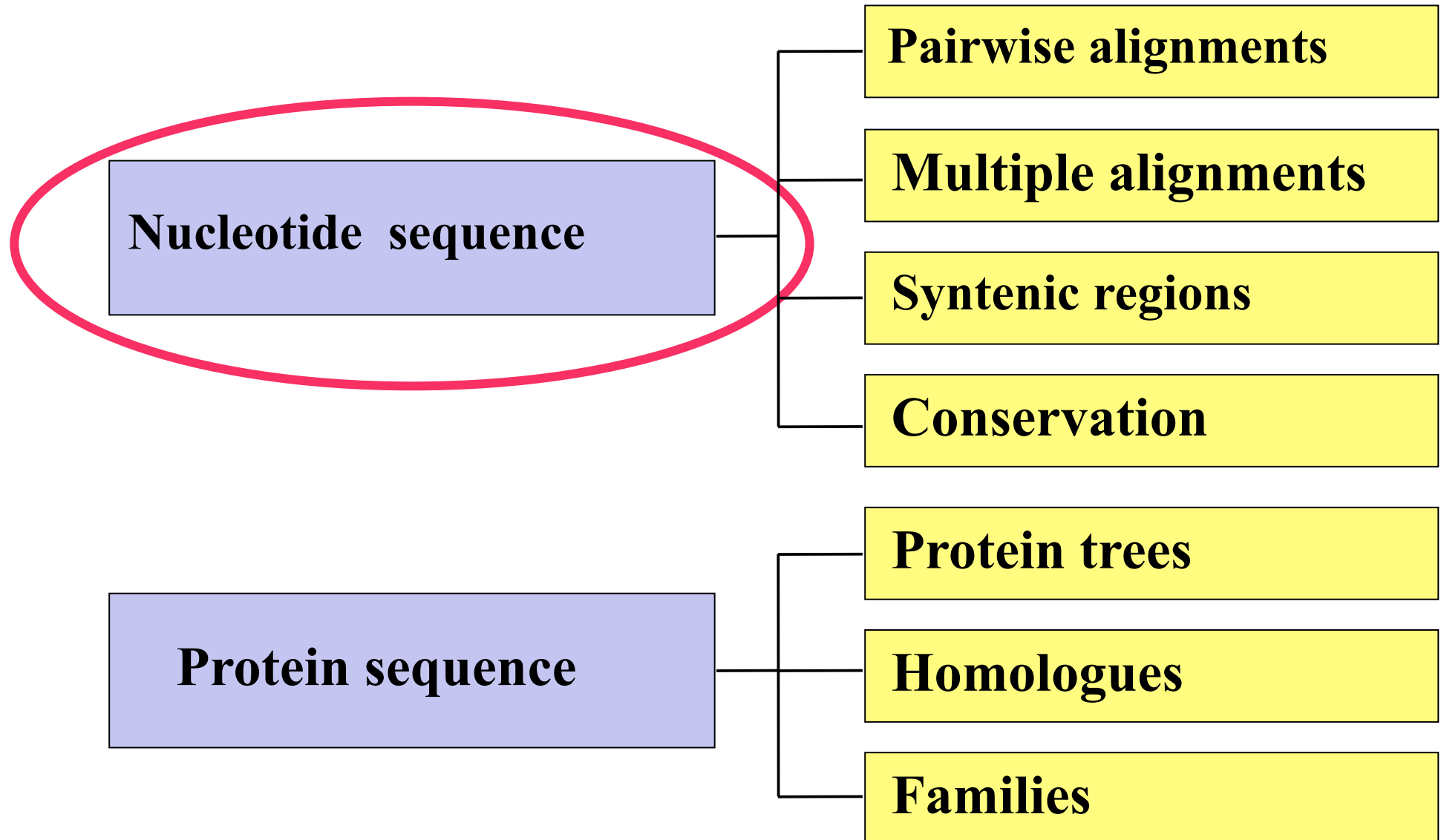
Stephen Fitzgerald

EBI - Wellcome Trust Genome Campus, UK

http://www.ebi.ac.uk/~stephenf/Workshops/Cam_dec_2013/



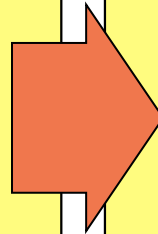
Sequence types and outputs



Nucleotide sequence analyses

Pairwise Alignments

BLASTZ-net
LASTZ-net
t-BLAT-net

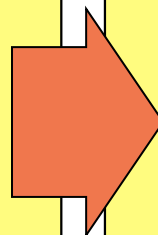


Syntenic regions

Only for species with
chromosomal mappings

Multiple alignments

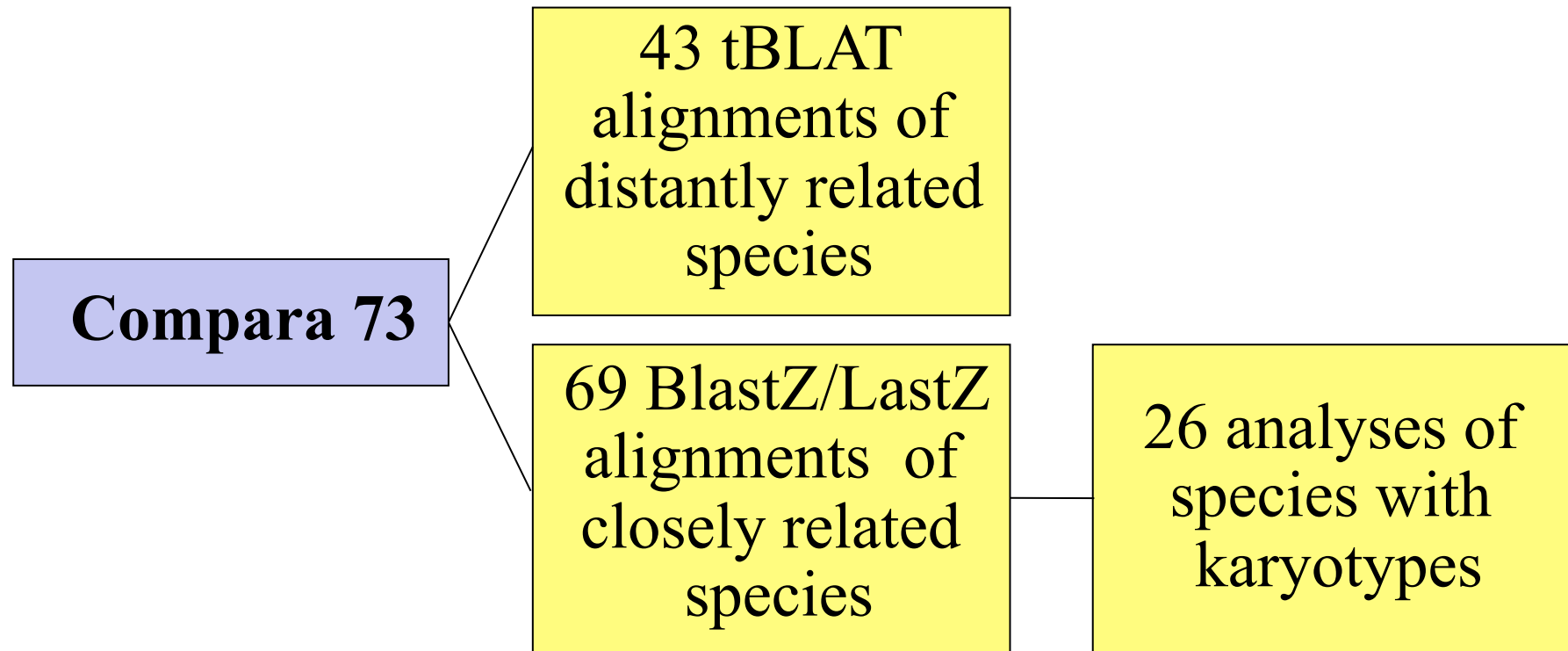
Mercator-Pecan
Enredo-Pecan-Ortheus



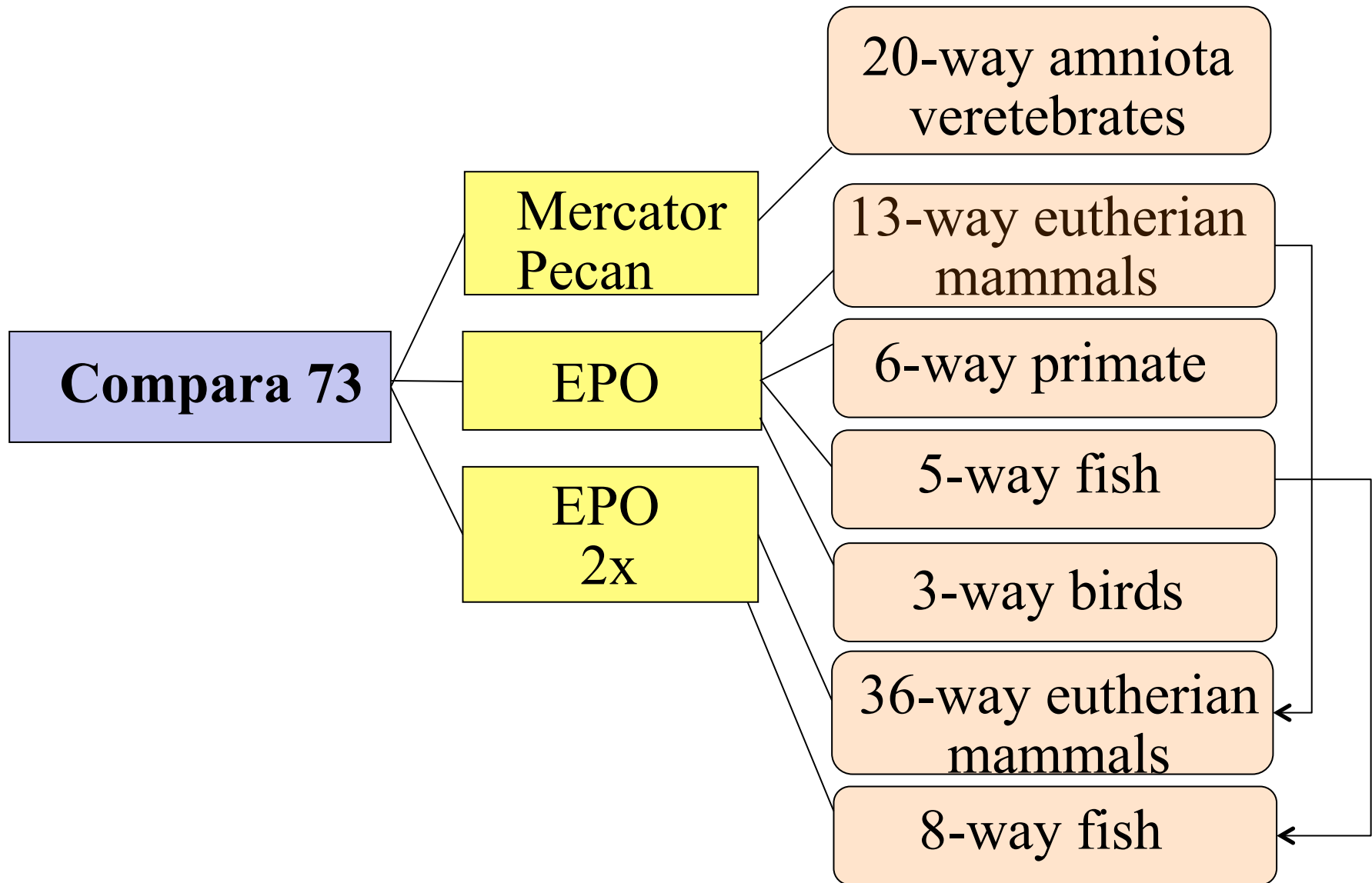
Conservation

GERP Cons. Scores
GERP Constr. Elements

Pairwise Alignments in Compara 73



Multiple Alignments in Compara 73





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 DnaFrag 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

Compara database is coupled to Ensembl core databases

Compara stores relationships between the genomes by loading references or 'handles' to external data.

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

Example: compara_70 must be linked with the Ensembl core_70 databases

Proper REGISTRY configuration is needed

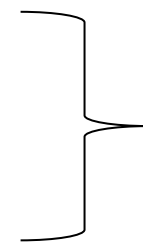
Alignments are stored in the genomic_align and genomic_align_block tables

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
AACATCTTAACTAAACGCC
GATACTTAACTTAAACCCCC
ACTACCTAACTAAAACCCC
GACATTTAACTAAAACCCC

5MD11MD3M
17MD3M
5MD15M
5MD11MD3M
5MD11MD3M



5 genomic_align entries
1 genomic_align_block

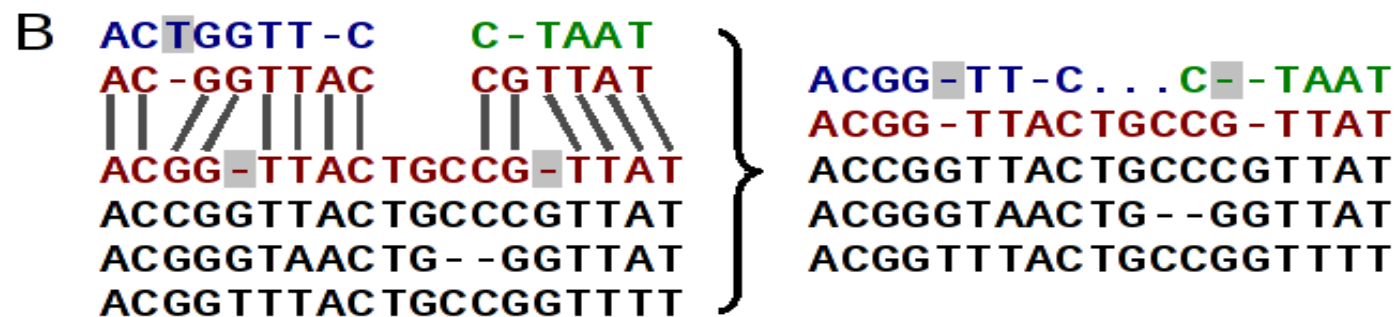
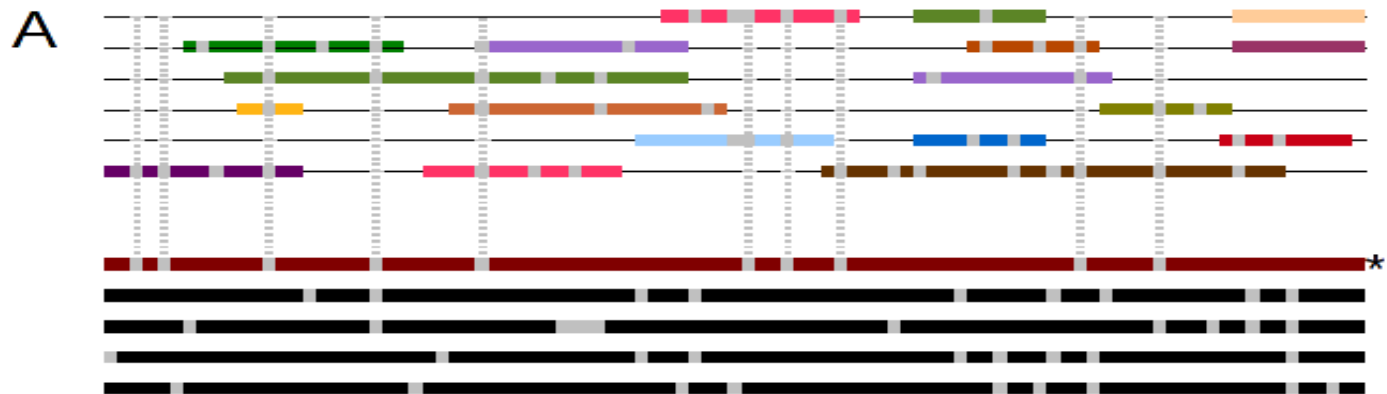
Sequences from core

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



Exercises – GenomicAlignBlock (Constrained elements)

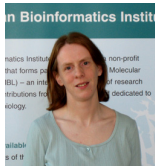
- A Constrained Elements represent regions in the multiple alignment which appear to be under functional constraint.
- Print the constrained element alignments from the above pig locus (use the constrained elements generated from the EPO_LOW_COVERAGE mammals alignments)

Exercises - Synteny

- Synteny blocks are derived from Lastz-net alignments
 - group syntenic alignments closer than 200Kb
 - link syntenic groups closer than 3Mb
 - minimum length of the syntenic block: 100 kb
- Print the pig-cow synteny map using pig chromosome 15 as a reference

Acknowledgements

Compara Team



Kathryn



Leo



Miguel



Stephen



Matthieu



European Commission
Framework Programme 7



Quantomics
From Sequence to Consequence :
Tools for the Exploitation of Livestock Genomes

wellcometrust

