Aalto University
School of Science

CS-E5865 Computational genomics

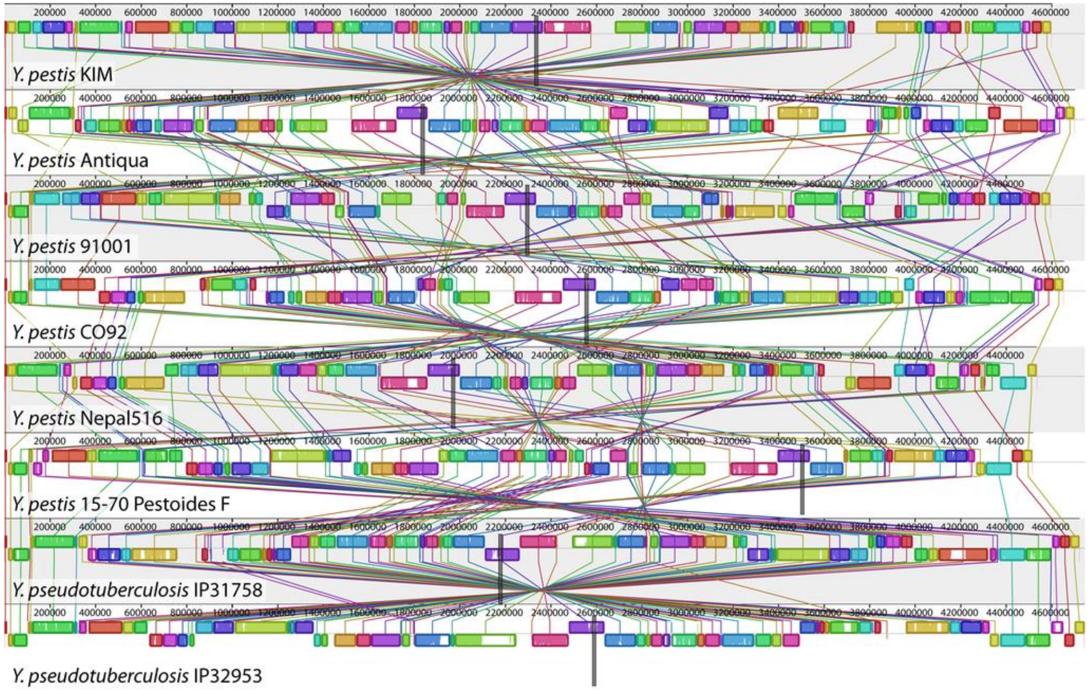
Autumn 2020, Lecture 8: Comparative genomics

Lecturer: Pekka Marttinen

Assistants: Alejandro Ponce de León, Zeinab Yousefi, Onur Poyraz

Comparative genomics

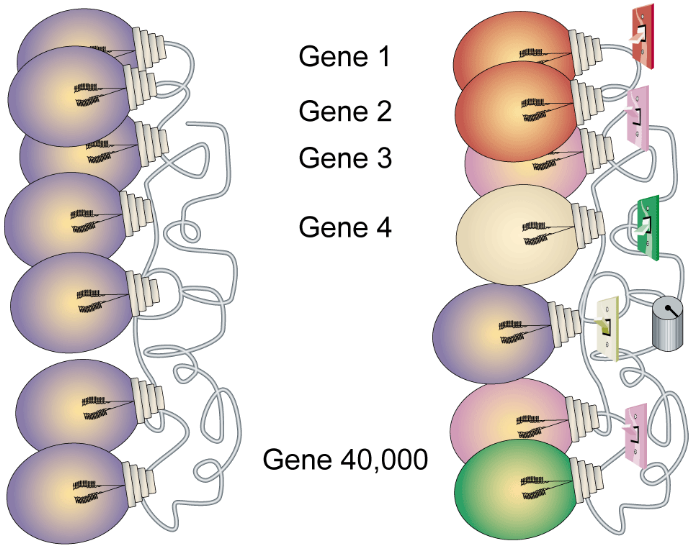
- Comparison of genomes across multiple organisms
- Large scale variation: gene duplications, transpositions, deletions, inversions, horizontal gene transfers

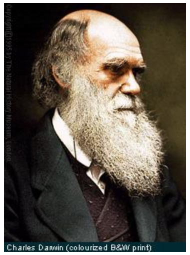


https://en.wikipedia.org/wiki/Comparative_genomics

Goals of comparative genomics

- improved identification of genes and other conserved regions
- knowledge of what each gene does
- improved understanding of evolution

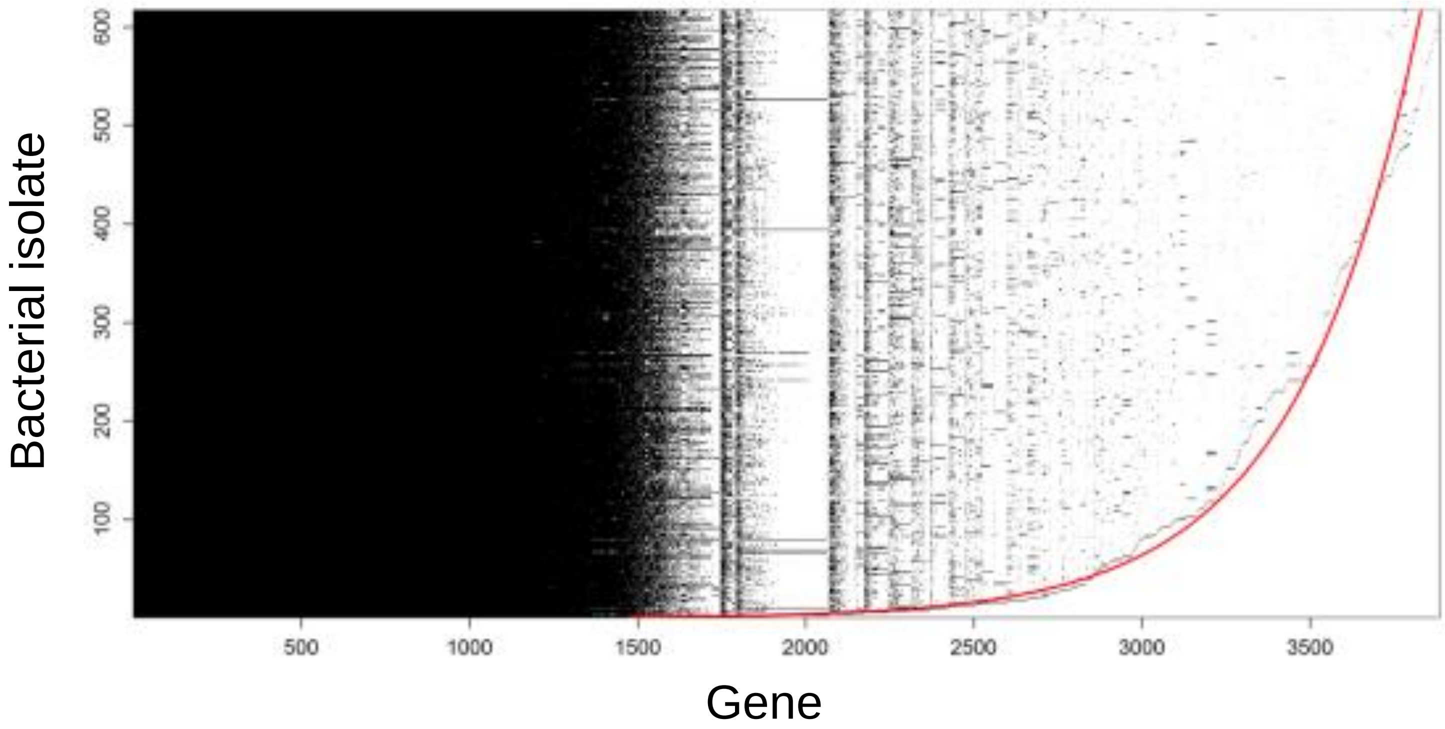




Variation in gene content

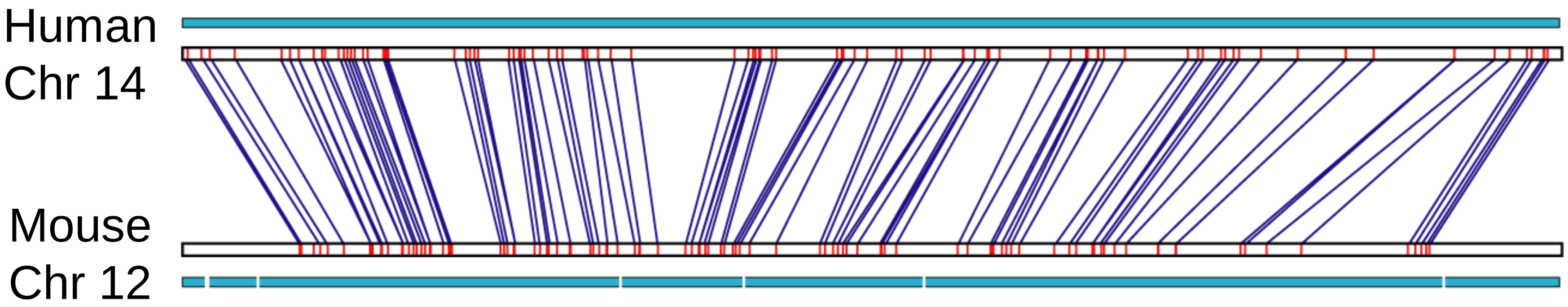
Core genome
(genes present in all)

Accessory genome
(genes present in a subset)



Synteny

- Synteny: relative ordering of genes on chromosomes
- The comparison of mouse and human genomes reveals regions of synteny
  - Almost identical order of genes along the chromosomes
  - the sizes of the intergenic regions vary



Orthologs and Paralogs

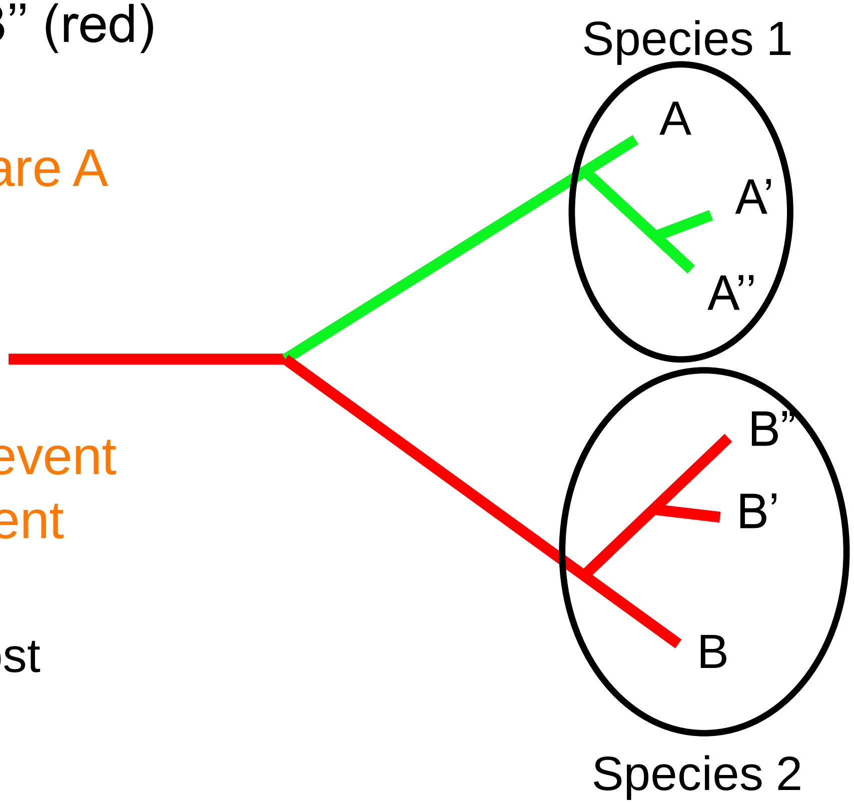
- For the comparison of two genes to be meaningful one must know the evolutionary relationship between them

- Distinguish between:
  - Orthologs genes found in two species that have been inherited from a common ancestor
  - Paralogs copies of a gene created by gene duplication events

- Observation: paralogous genes may have different structure and function

Orthologs and Paralogs

- Genes A and B have a common ancestor
- Gene duplication within species has produced genes A’,A’’ (green) and B’, B’’ (red)
- A and B are most closely related to the ancestral gene → reasonable to compare A and B
- A’, A’’, B’ and B’’ may be more distantly related and have new functions
- Are two genes related by a speciation event (orthologs) or by a gene duplication event (paralogs)?
  - Check how they diverged from their most recent common ancestor.



Homology searches

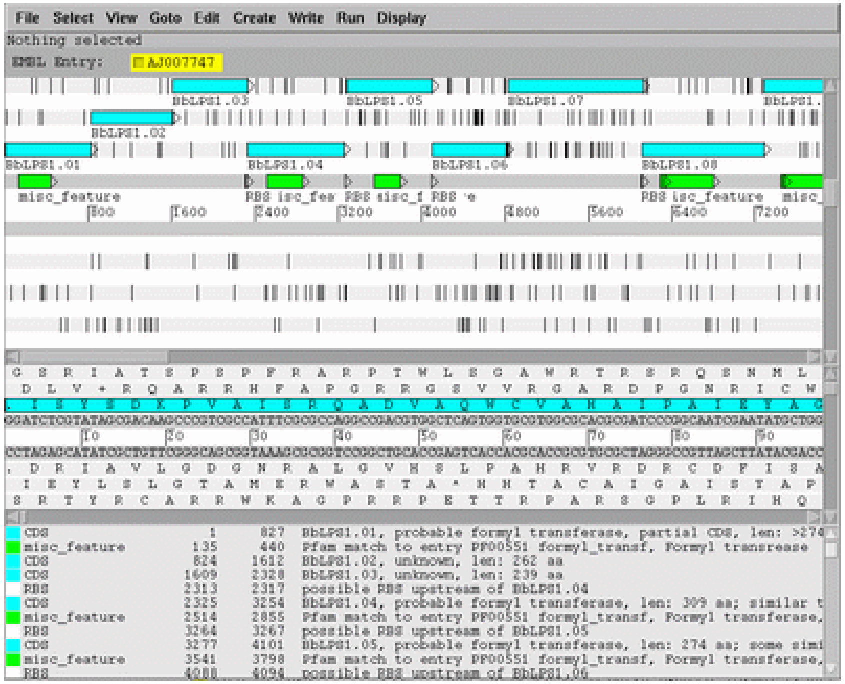
- Search databases for matches with a given DNA sequence

- Use a computational method to align sequences
  - Perfect match not required
    - Allow for insertions, deletions and base changes

- Most common method: BLAST (discussed briefly in an earlier lecture)

Genome annotation

- Comparative genomics can help identify:
  - Gene structure: e.g. exon/intron boundaries
  - Gene function: by similarity to a gene with a known function
  - Regulatory sequences, e.g., promoters and enhancers
  - Interactions between gene products



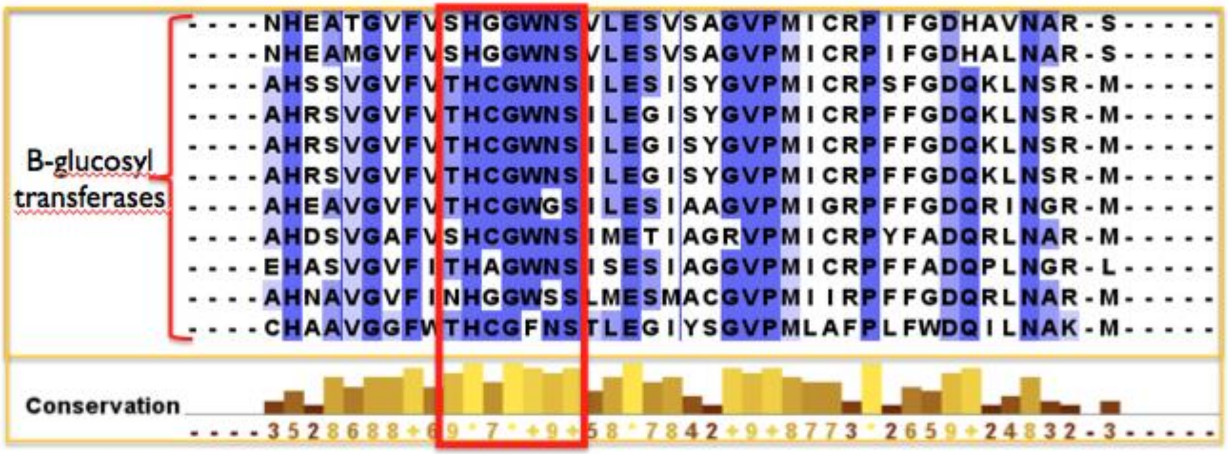
```
CDS                1     827  BbLPS1.01, probable formyl transferase, partial CDS, len: >274
misc_feature     135     440  Pfam match to entry PF00551 formyl_transf, Formyl transrease
CDS              824    1612  BbLPS1.02, unknown, len: 262 aa
CDS             1609    2328  BbLPS1.03, unknown, len: 239 aa
RBS             2313    2317  possible RBS upstream of BbLPS1.04
CDS             2325    3254  BbLPS1.04, probable formyl transferase, len: 309 aa; similar t
misc_feature    2514    2855  Pfam match to entry PF00551 formyl_transf, Formyl transferase,
RBS             3264    3267  possible RBS upstream of BbLPS1.05
CDS             3277    4101  BbLPS1.05, probable formyl transferase, len: 274 aa; some simi
misc_feature    3541    3798  Pfam match to entry PF00551 formyl_transf, Formyl transferase,
RBS             4088    4094  possible RBS upstream of BbLPS1.06
```

http://en.bio-soft.net/dna/Artemis.html

Conserved regions in genomes

- Conservation of DNA is usually for a reason
  - Without constraints random mutations would occur

- Non-coding DNA without any functional importance diverges more rapidly than protein coding DNA



```
----NHEATGVFVSHGGWNSVLESVSAGVPMICRPIFGDHAVNAR-S------
----NHEAMGVFVSHGGWNSVLESVSAGVPMICRPIFGDHALNAR-S------
----AHSSVGVFVTHCGWNSILESISYGVPMICRPSFGDQKLNSR-M------
----AHRSVGVFVTHCGWNSILEGISYGVPMICRPFFGDQKLNSR-M------
----AHRSVGVFVTHCGWNSILEGISYGVPMICRPFFGDQKLNSR-M------
----AHRSVGVFVTHCGWNSILEGISYGVPMICRPFFGDQKLNSR-M------
----AHEAVGVFVTHCGWGSILESIAAGVPMIGRPFFGDQRINGR-M------
----AHDSVGAFVSHCGWNSIMETIAGRVPMICRPYFADQRLNAR-M------
----EHASVGVFITHAGWNSISESIAGGVPMICRPFFADQPLNGR-L------
----AHNAVGVFINHGGWSSLMESMACGVPMIIRPFFGDQRLNAR-M------
----CHAAVGGFWTHCGFNSTLEGIYSGVPMLAFPLFWDQILNAK-M------
----3528688+69*7*+9+58*7842+9+8773*2659+24832-3------
```

http://2013.igem.org/Team:Berkeley/Methods/Protocols

Genome comparisons for gene finding

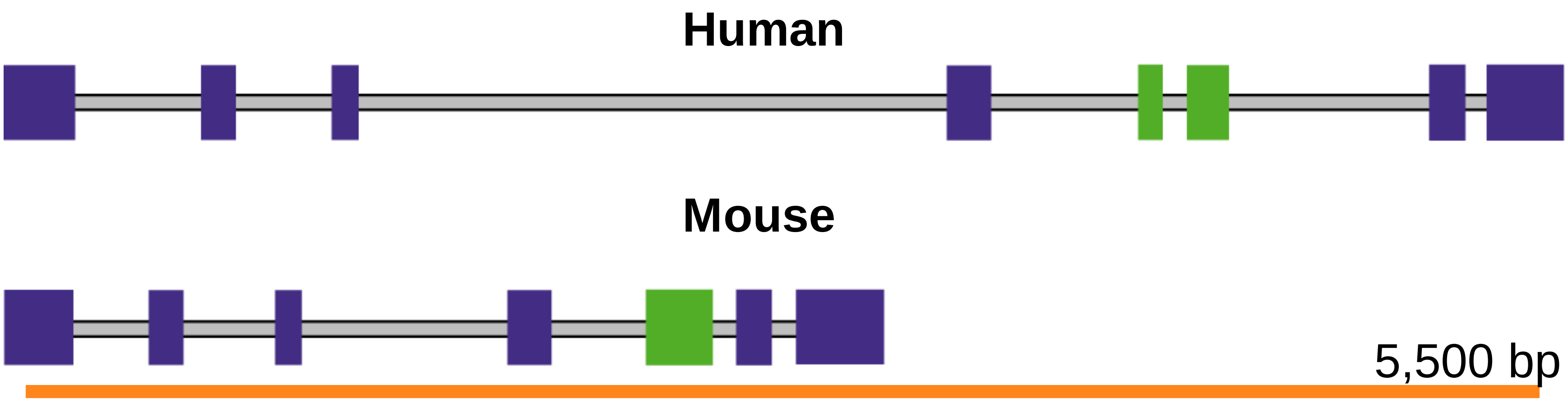
- Coding regions and exon/intron structures tend to be conserved in different species

- Look for ORFs conserved in both genomes

- Cross-genome comparisons improve accuracy
  - Fugu/human comparison found >1000 genes that had been missed by the annotation programs
  - Mouse/human comparison reduced the number of false-positive predictions
    - Twinscan, a dual-genome gene finder, predicted 25,600 human genes versus 45,000 predicted by Genscan

Sequence comparison example: evolution of gene structure

- Comparison of the human and mouse spermidine synthase genes (involved in the synthesis of polyamines)
- Revealed an additional intron in the human gene not found in the mouse homolog



Nature. 2002 Dec 5;420(6915):520-62

Aalto University
School of Science

Genome sizes

- Comparison of the total number of genes:
  - Smaller than originally expected
  - Ex: Human genome initially thought to have 100,000 genes
  - Now around 21000 genes
  - Fruit fly ~13600 genes, dog ~25000 genes, mouse ~23000 genes
- Suggests that many new functions arise from altering transcriptional regulation
  - Use old genes in new ways

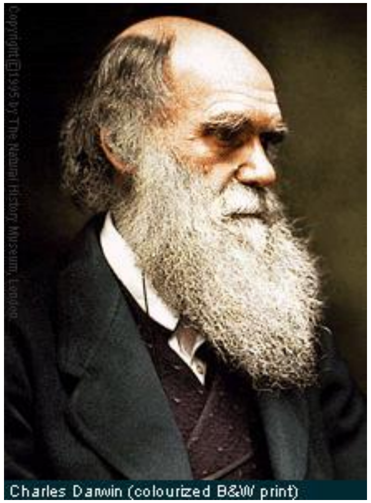
Evolution of new traits

- Genome comparisons can also help to gain a better understanding of evolution

- Evolution of adaptive traits occurs through:
  - Evolution of new genes
  - Changing when and where genes are expressed
  - Horizontal gene transfer

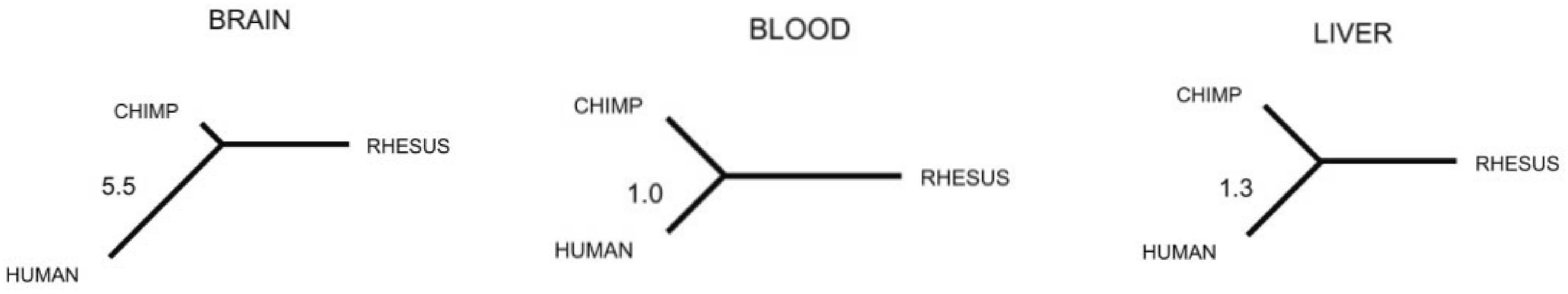
What is difference between man and chimpanzee?

- Man and chimpanzee have a genome wide similarity greater than 95%.
- Studies suggest differences between the species are due to differences in gene expression
  - Striking differences found especially in brain



Human/ape gene expression comparisons

- A comparison of the expression of 12,000 genes in rhesus monkeys, chimpanzees, and humans



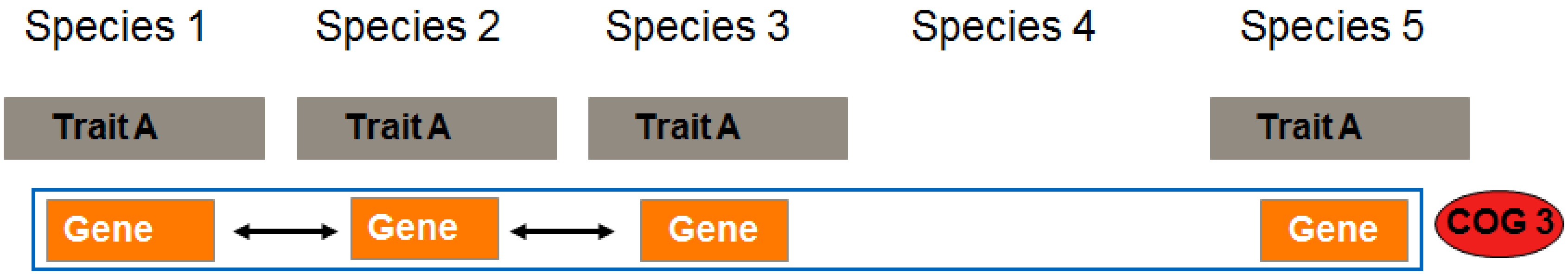
Distance trees based on the relative changes in the expression levels.
Numbers represent the ratio of changes between humans and chimpanzees.

Enard W et al. (2002) “Intra- and Interspecific variation in primate gene expression patterns” Science 296

Relating traits to genes

- The Trait-to-gene approach uses the “COG database”
  - COG= Cluster of Orthologous Group which identifies likely orthologues when making whole genome comparisons among multiple species.

- When all species share a COG →it is likely to play an important role in a basic housekeeping function.



Species 1 Species 2 Species 3 Species 4 Species 5

Trait A Trait A Trait A Trait A

Gene ⟷ Gene ⟷ Gene Gene

COG 3

From genomes to populations

- *population_genetic_modeling.pdf*

Thanks!

Please complete the course feedback questionnaire, it only takes a few minutes. By doing this, you will gain 1 bonus point for the exam.

Exam from 9 am to 1 pm on October 20th, 2020. Check the details from oodi and make sure you have a registration!

The exam is arranged remotely. Allowed equipment: calculator (non-symbolic, memory erased), course material (lectures, exercises, model solutions), other regular equipment (pencil, eraser, ...). Further details about the arrangement will be announced in MyCourses in due time.

Retake on December 11th, 2020.