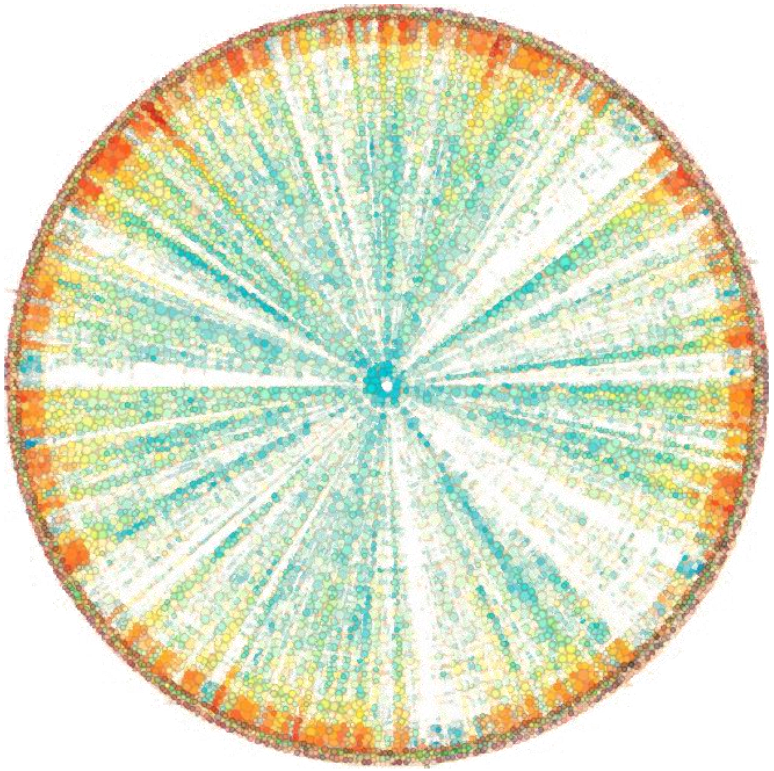
Comparative Genomics and Visualisation

- Evolutionary relationships of genome features can be complex.
- We require precise terms to describe relationships between genome features.



Comparing Gene Features

- Given gene annotations for more than one genome, how can we organise and understand relationships?
  - Functional similarity (analogy)
  - Evolutionary common origin (homology, orthology, etc.)
  - Evolutionary/functional/family relationships (paralogy)

DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS

WALTER M. FITCH

Abstract

Fitch, W. M. (Dept. Physiological Chem., U. Wisconsin, Madison 53706) 1970. Distinguishing homologous from analogous proteins. Syst. Zool., 19:99–113.—This work provides a means by which it is possible to determine whether two groups of related proteins have a common ancestor or are of independent origin. A set of 16 random

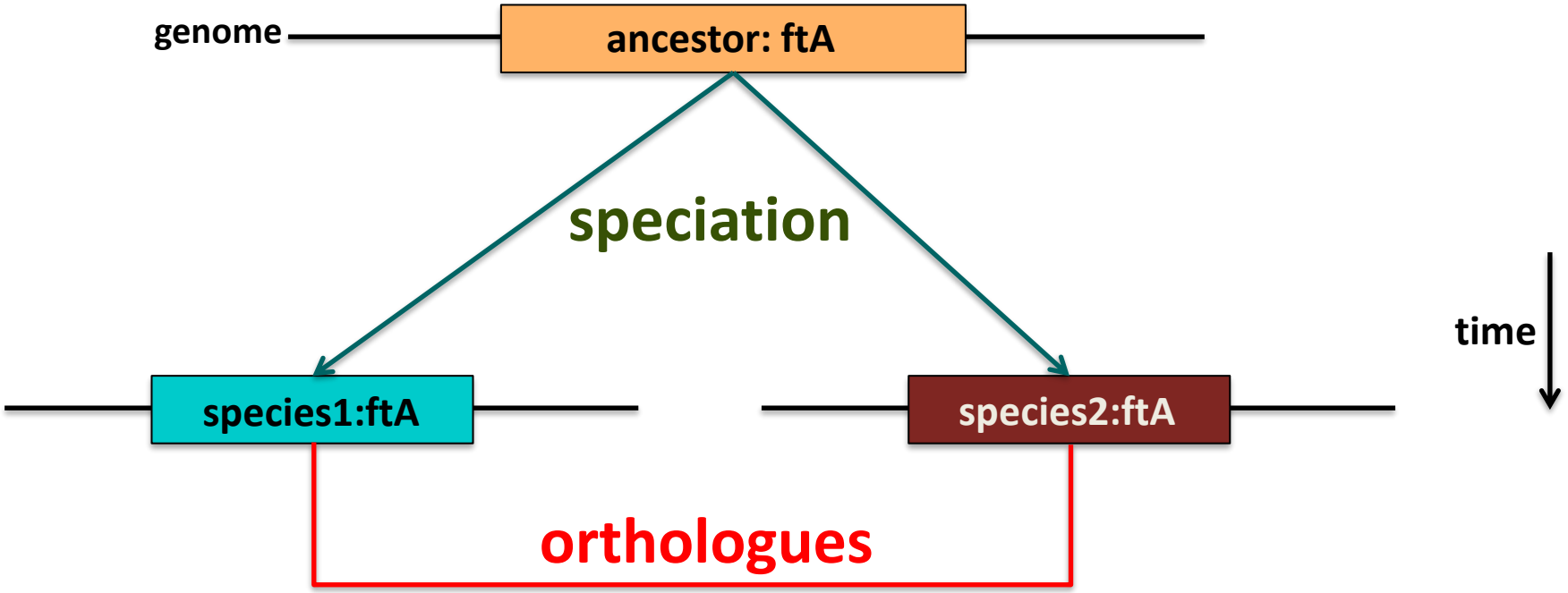
Terms first suggested by Fitch (1970) *Syst. Zool.* doi:10.2307/2412448

Attack of the –logues

- Technical terms describing evolutionary relationships
- **Homologues:** elements that are similar because they share a common ancestor (**NOTE: There are NOT degrees of homology!**)
- **Analogues:** elements that are (functionally?) similar, possibly through convergent evolution and not by sharing common ancestry
- **Orthologues:** homologues that diverged through speciation
- **Paralogues:** homologues that diverged through duplication within the same genome
- (also co-orthologues, xenologues, etc.)

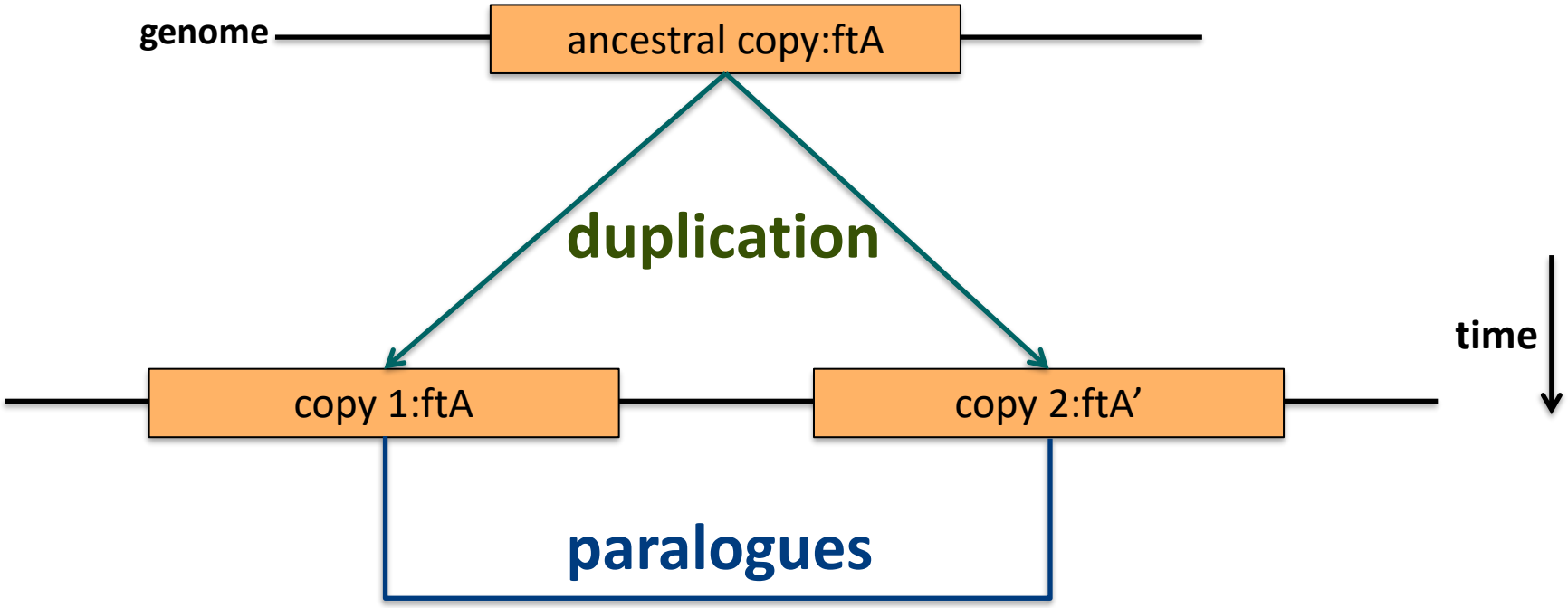
Attack of the –logues

Attack of the –logues



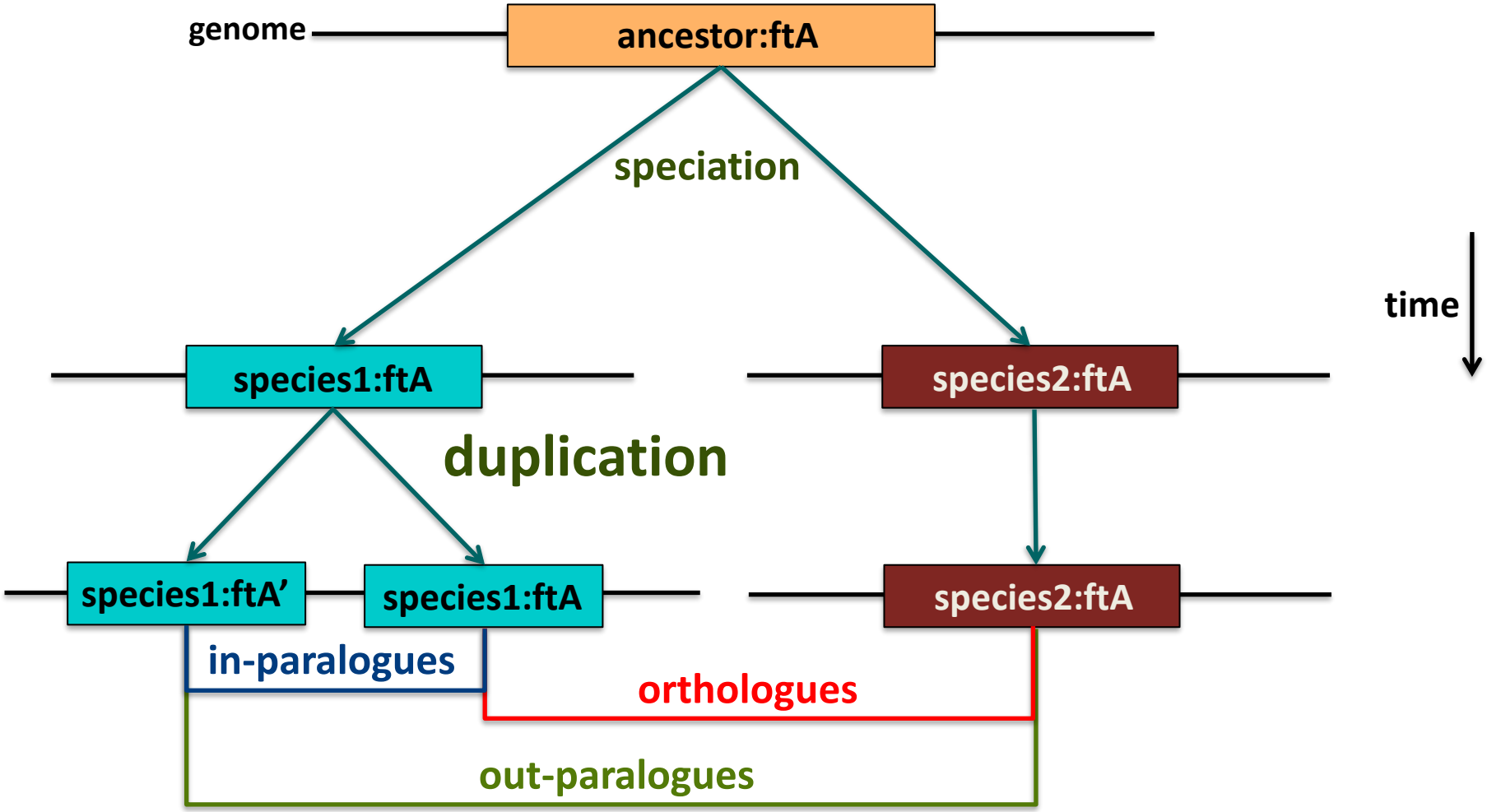
- **Orthologues:** homologues that diverged through speciation

Attack of the –logues

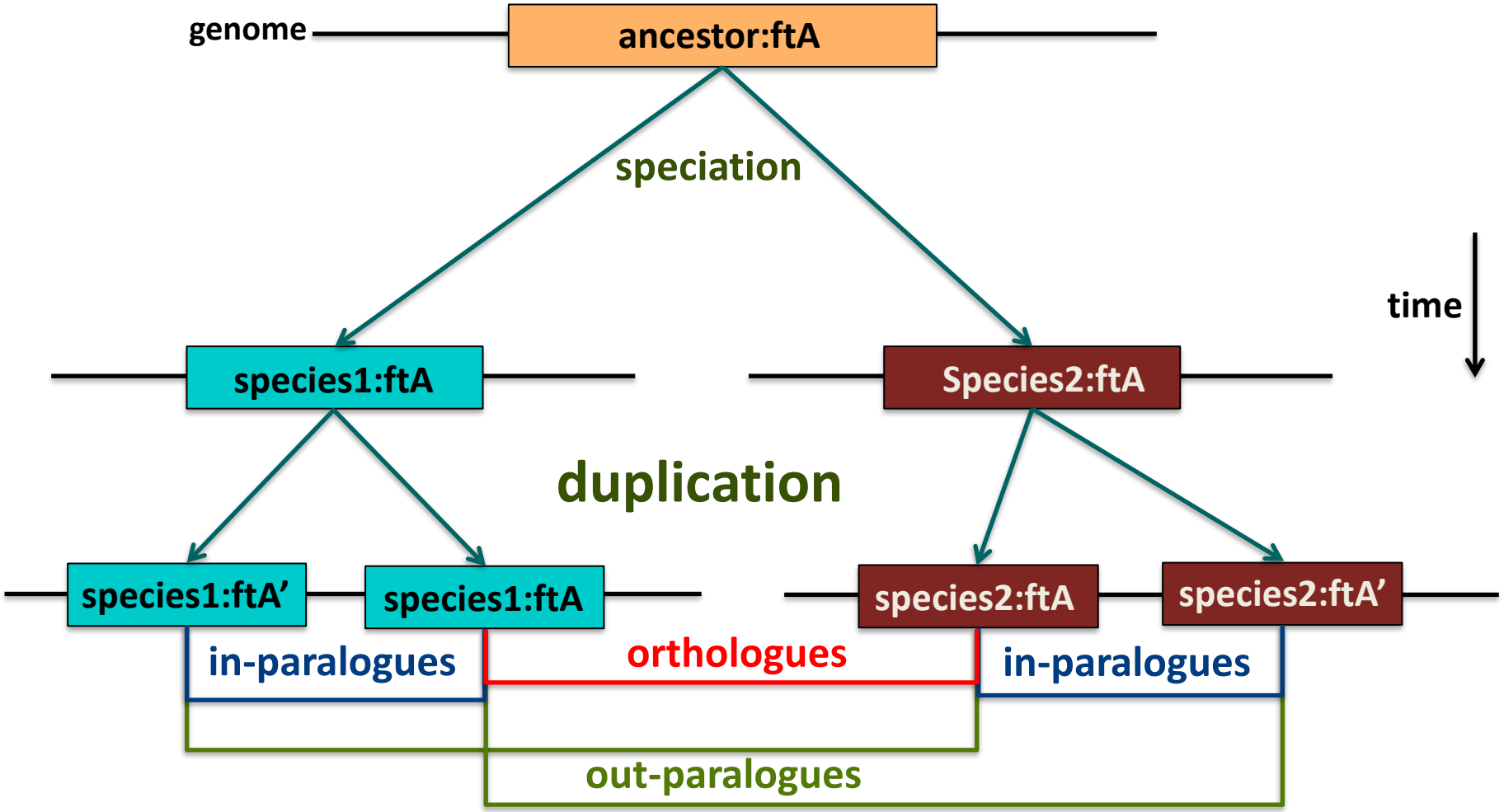


Paralogues: homologues that diverged through duplication within the same genome

Attack of the –logues



Attack of the –logues

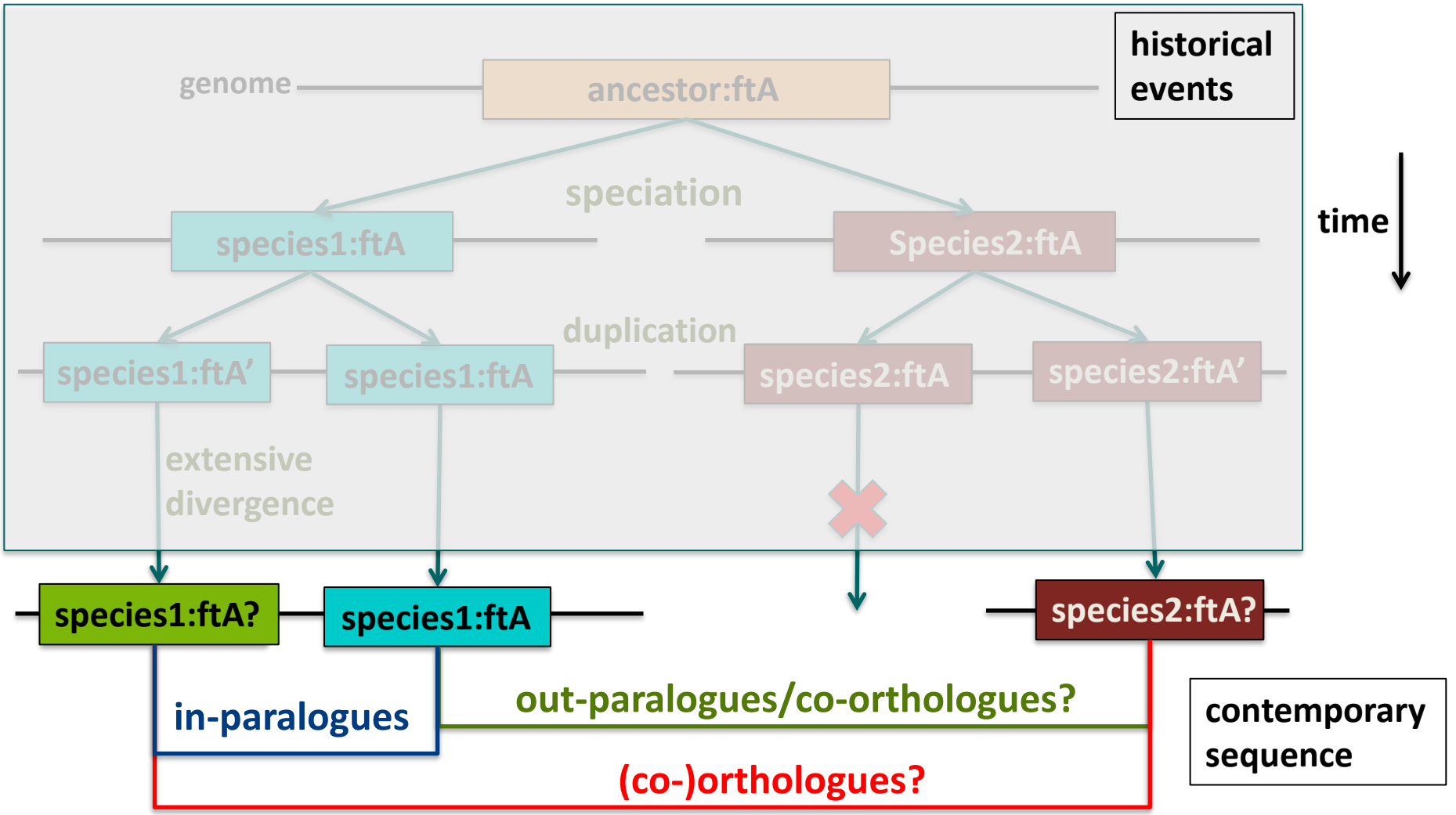


Attack of the –logues

- **BUT:** biology is not well-behaved: relationships can be difficult to infer
  - Gene loss occurs
  - Homologues can diverge – sometimes very widely: hard to recognise
  - Reconstructed evolutionary trees for speciation events may not be robust

Kristensen et al. (2011) *Brief. Bioinf.* doi:10.1093/bib/bbr030

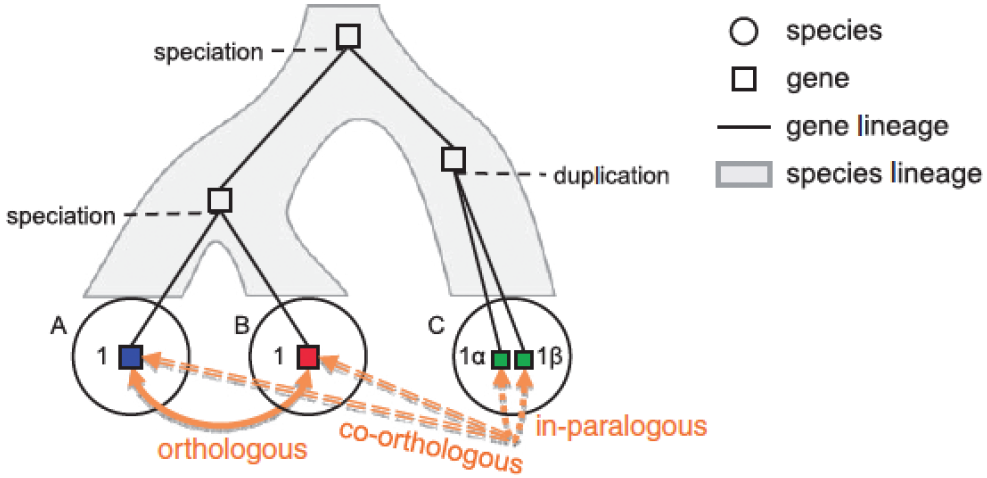
Attack of the –logues



Current classifications of orthology/paralogy are inferences

Attack of the –logues

- **BUT:** biology is not well-behaved: relationships can be difficult to infer
  - Gene loss occurs
  - Homologues can diverge – sometimes very widely: hard to recognise
  - Reconstructed evolutionary trees for speciation events may not be robust
- Some resources and tools 'bend' definitions, e.g. Ensembl Compara and OrthoMCL.



http://www.ensembl.org/info/genome/compara/ homology_method.html

Kristensen et al. (2011) *Brief. Bioinf.* doi:10.1093/bib/bbr030

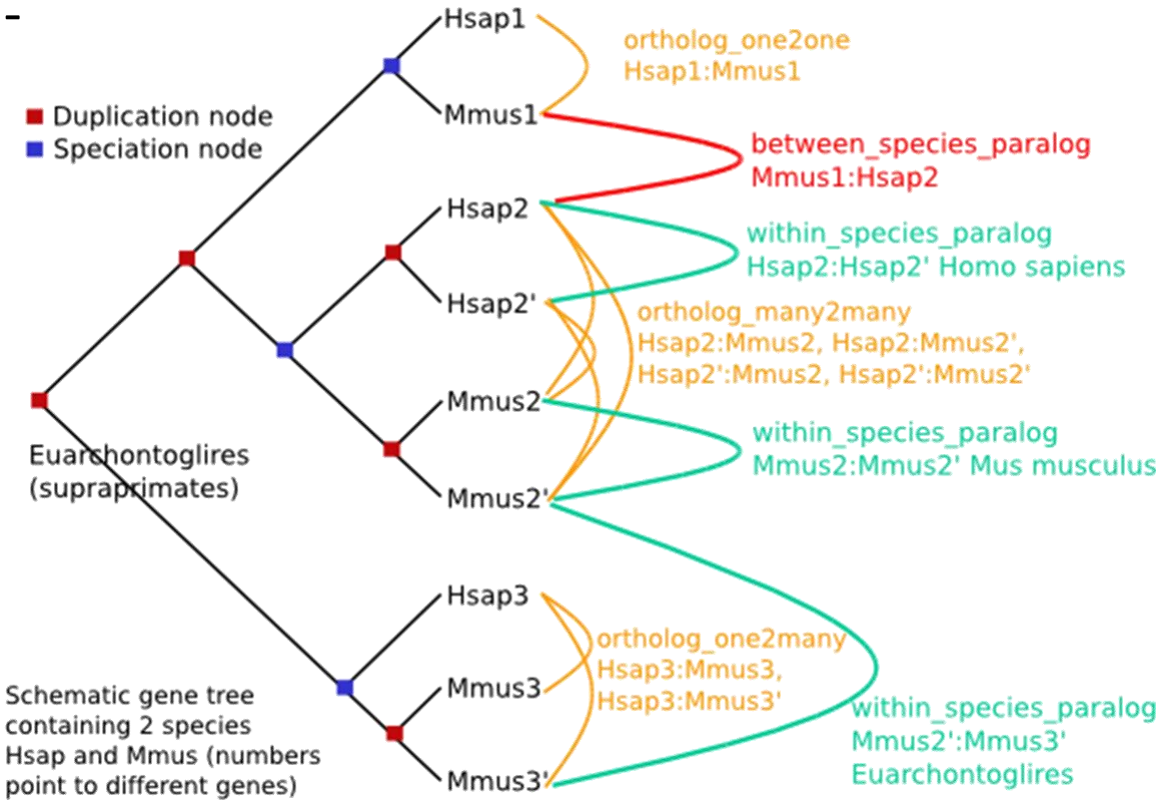
Note on “Orthology”

- Frequently abused/misused as a term
- “Orthology” is an evolutionary relationship, often bent into service as a functional descriptor
- Strictly defined only for two species or clades!
  - (cf. OrthoMCL, etc.)
- Orthology is not transitive (A is orthologue of C and B is orthologue of C does not imply A is an orthologue of B)
  - (cf. EnsemblCompara definitions)

Storm & Sonnhammer (2002) *Bioinformatics.* doi:10.1093/bioinformatics/18.1.92

Ensembl Compara definitions

- `within_species_paralog:`
  same-species paralogue (in-paralogue)
- `ortholog_one2one:`
  orthologue
- `ortholog_one2many:`
  orthologue/paralogue relationship
- `orthology_many2many:`
  orthologue/paralogue relationship



NOTE: the taxonomy may not always be correct…

Vilella *et al.* (2009) *Genome Res.* doi:10.1101/gr.073585.107

“The Ortholog Conjecture”

Without duplication, a gene is unlikely to change its basic function, because this would lead to loss of the original function, and this would be harmful.

Problems with the Ortholog Conjecture

- Nehrt *et al.* (2011) say:
  - Paralogues better predictor of function than orthologues
    - ∴ conjecture is false!
  - Cellular context better for protein function inference
  - Function defined from Gene Ontology (GO)

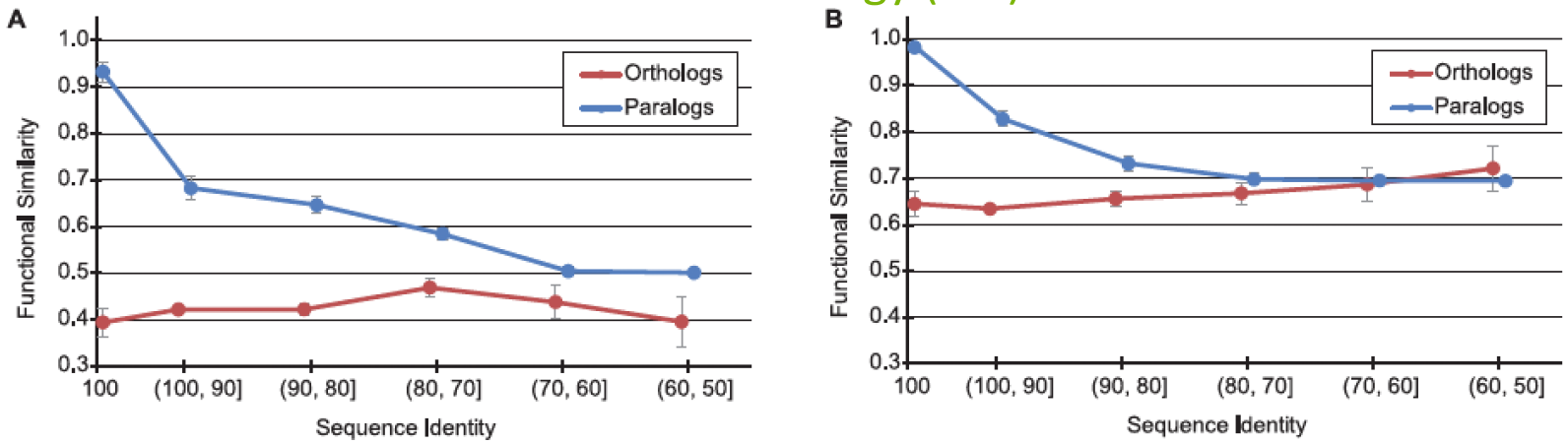


Figure 1. The relationship between functional similarity and sequence identity for human-mouse orthologs (red) and all paralogs (blue). Standard error bars are shown. (**A**) Biological Process ontology, (**B**) Molecular Function ontology.
doi:10.1371/journal.pcbi.1002073.g001

Nehrt *et al.* (2011) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002073
Chen *et al.* (2012) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002784

Problems with the Ortholog Conjecture

- But do we understand function well enough to test the conjecture?
- Chen *et al.* (2012) say: **"No"**
  - "examination of functional studies of homologs with identical protein sequences reveals experimental biases, annotation errors, and homology-based functional inferences that are labeled in GO as experimental. These problems […] make the current GO inappropriate for testing the ortholog conjecture"
  - Expression level similarity is more similar for orthologues than paralogues **(but is *this* "function"…?)**

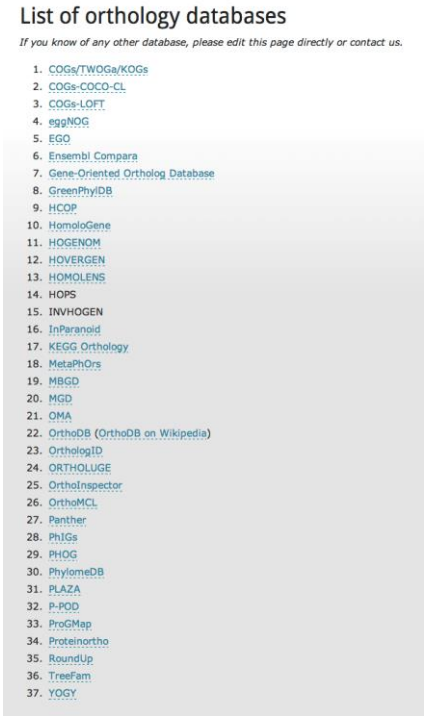
Nehrt *et al.* (2011) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002073
Chen *et al.* (2012) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002784

Finding “Orthologues”

The process of finding evolutionary (and/or functional) equivalents of genes across two or more organisms’ genomes.

Why are “orthologues” so important?

- Orthology formalises the concept of **corresponding genes** across multiple organisms.
  - Evolutionary
  - Functional? (**“The Ortholog Conjecture”**)
- Applications in:
  - Comparative genomics
  - Functional genomics
  - Phylogenetics, …
- Many (>35) databases attempt to describe orthologous relationships
  - http://questfororthologs.org/orthology_databases



List of orthology databases

If you know of any other database, please edit this page directly or contact us.

1. COGs/TWOGs/KOGs
2. COGs-COCO-CL
3. COGs-LOFT
4. eggNOG
5. EGO
6. Ensembl Compara
7. Gene-Oriented Ortholog Database
8. GreenPhylDB
9. HCOP
10. HomoloGene
11. HOGENOM
12. HOVERGEN
13. HOMOLENS
14. HOPS
15. INVHOGEN
16. InParanoid
17. KEGG Orthology
18. MetaPhOrs
19. MBGD
20. MGD
21. OMA
22. OrthoDB (OrthoDB on Wikipedia)
23. OrthologID
24. ORTHOLUGE
25. OrthoInspector
26. OrthoMCL
27. Panther
28. PhIGs
29. PHOG
30. PhylomeDB
31. PLAZA
32. P-POD
33. ProGMap
34. Proteinortho
35. RoundUp
36. TreeFam
37. YOGY

Dessimoz (2011) *Brief. Bioinf.* doi:10.1093/bib/bbr057

How to find orthologues?

- Many published methods and databases:
  - Pairwise between two genomes:
    - RBBH (aka BBH, RBH, etc.), RSD, InParanoid, RoundUp
  - Multi-genome
    - Graph-based: COG, eggNOG, OrthoDB, OrthoMCL, OMA, MultiParanoid
    - Tree-based: TreeFam, Ensembl Compara, PhylomeDB, LOFT
- Methods may apply different - or refined - definitions of orthology, paralogy, etc.

Salichos *et al.* (2011) *PLoS One.* doi:10.1371/journal.pone.0018755
Trachana *et al.* (2011) *Bioessays* doi:10.1002/bies.201100062
Kristensen *et al.* (2011) *Brief. Bioinf.* doi:10.1093/bib/bbr030

Selection Pressures

Signs of selection pressure identifiable by comparative genomics

Selection Pressures

- Defining core groups of genes by “orthology” allows analysis of those groups:
  - **Synteny/collocation**
  - Gene neighbourhood changes (e.g. **genome expansion**)
  - **The pangenome: core and accessory genomes**
- and sequences in those groups:
  - Multiple alignment
  - Domain detection
  - Identification of functional sites
  - **Inference of evolutionary pressures**

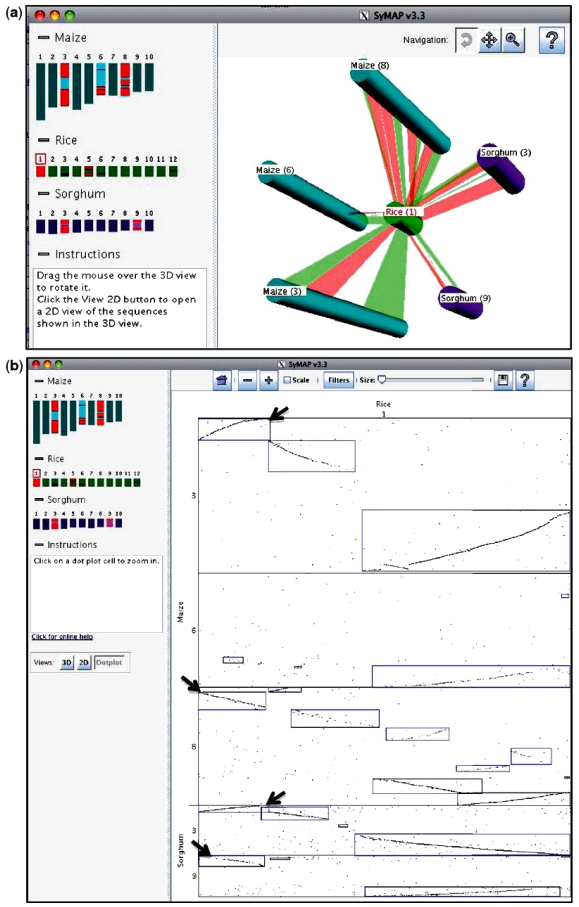
Synteny

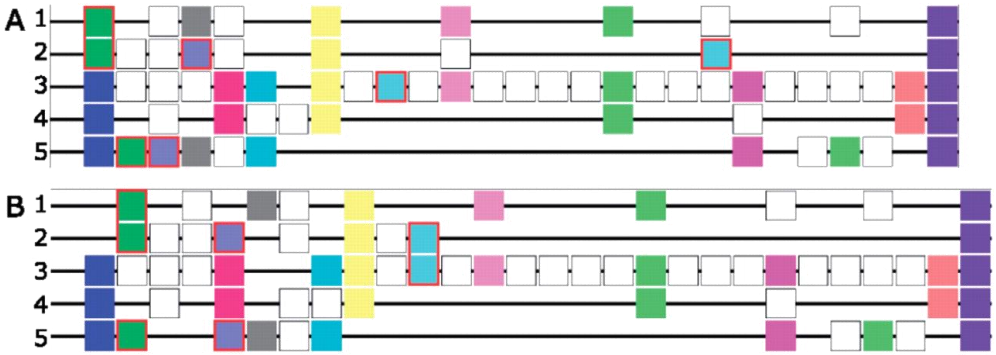
- Selective pressures depend on gene (product) function
- Genes involving physically or functionally-interacting proteins tend to evolve under similar selective constraints
- Particularly in bacteria, this leads to co-expression as *regulons* and collocation in *operons*
- Collocation (and coregulation) may be identified by comparative genomics
- (This is also true when considering regulatory or metabolic networks, similarly to genome organisation)

Alvarez-Ponce *et al.* (2011) *Genome Biol. Evol.* doi:10.1093/gbe/evq084

Synteny

- Many tools/packages/services for synteny detection, e.g.
  - SyMAP
    - http://www.agcol.arizona.edu/software/symap/
  - i-ADHoRe
    - http://bioinformatics.psb.ugent.be/software/details/i--ADHoRe
  - MCScan, Cyntenator, etc





Soderlund *et al.* (2011) *Nucl. Acids. Res.* doi:10.1093/nar/gkr123

Proost *et al.* (2011) *Nucl. Acids Res.* doi:10.1093/nar/gkr955

Genome Expansion

- Mobile/repeat elements reproduce and expand during evolution
- Generates sequence "laboratory" for variation and experiment
- e.g. *Phytophthora infestans* effector protein expansion and arms race

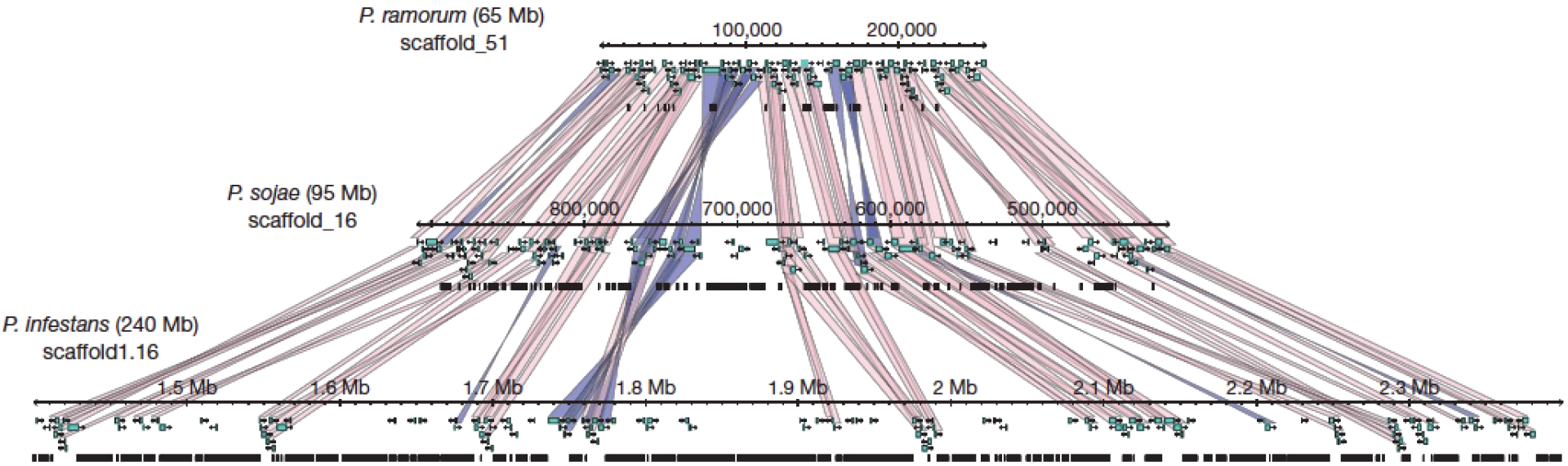
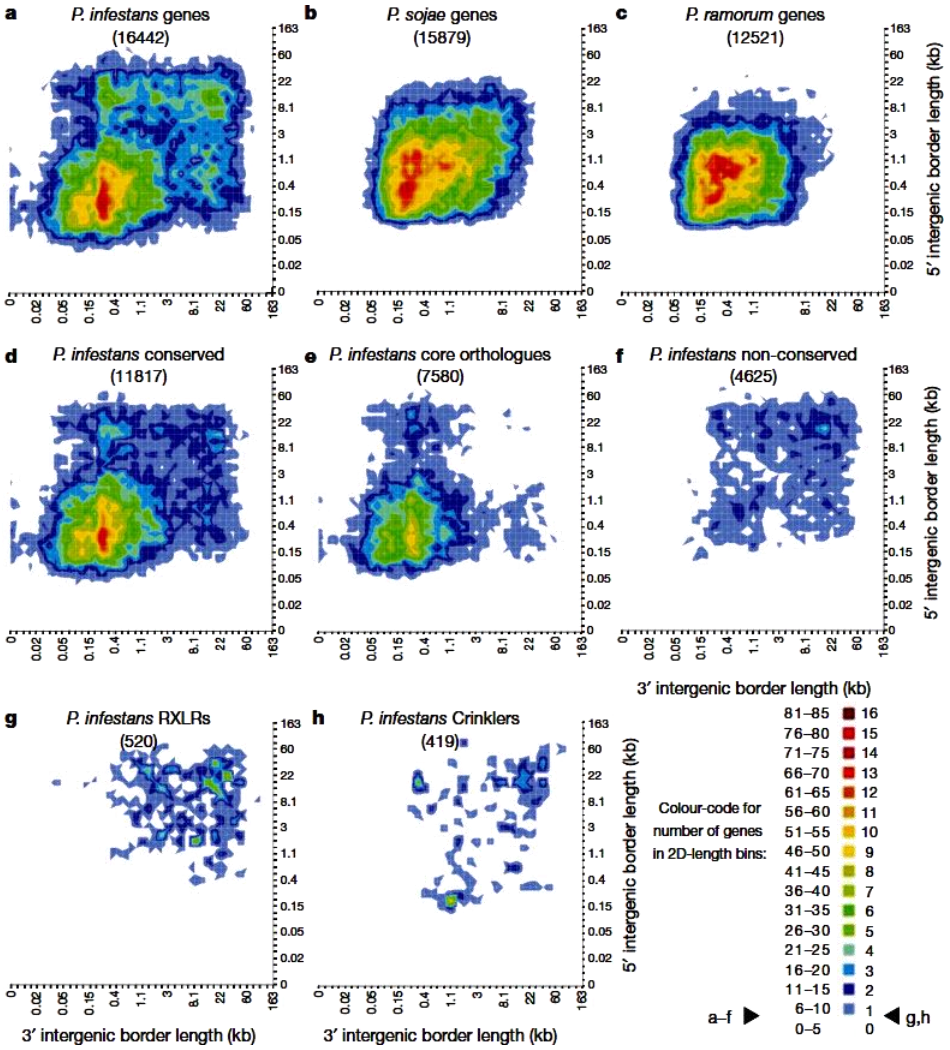


Figure 1 | Repeat-driven genome expansion in *Phytophthora infestans*. Conserved gene order across three homologous *Phytophthora* scaffolds. Genome expansion is evident in regions of conserved gene order, a consequence of repeat expansion in intergenic regions. Genes are shown as turquoise boxes, repeats as black boxes. Collinear orthologous gene pairs are connected by pink (direct) or blue (inverted) bands.

Haas *et al.* (2009) *Nature.* doi:10.1038/nature08358

Genome Expansion

- Mobile elements (MEs) are large, carry genes with them.
- Regions rich in MEs have larger gaps between consecutive genes
- Effector proteins are found preferentially in regions with large gaps, also show increased rates of evolutionary divergence.
- "Two-speed genome" associated with adaptability to new hosts/escape from evolutionary "bottleneck"



Haas *et al.* (2009) *Nature.* doi:10.1038/nature08358

The Pangenome

- The gene complement of a set of organisms (e.g. species group) is ***the pangenome***, defined by the union of two gene sets:
  - **Core genes**: genes present in all examples (define common species characteristics)
  - **Accessory genes**: genes only present in a subset of examples (relevant to adaptation of individuals)
- Definition depends on composition of organism set
- **Core genome hypothesis:**
  - "The *core genome* is the primary cohesive unit defining a bacterial species."
- Online tools available, e.g.
  - Panseq (http://lfz.corefacility.ca/panseq/)

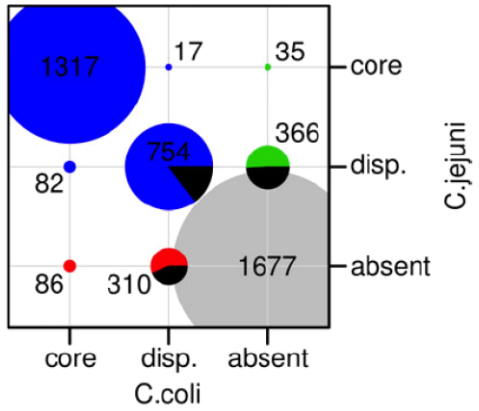


FIG. 4.—Overlap between the core and dispensable (disp.) genomic components of *Campylobacter coli* and *C. jejuni*; core genes were allowed to be missing in one strain per species. The absent/absent section represents genes that were found in other *Campylobacter* species but absent in *C. coli* and *C. jejuni*. Cirle radii are proportional to the number of genes. The black surface represents the proportion of putative pseudogenes.

Laing *et* al. (2010) *BMC Bioinf.* doi:10.1186/1471-2105-11-461

***Lefébure et al.* (2010) *Genome Biol. Evol.* doi:10.1093/gbe/evq048**

Defining a species' core genome

- "Orthologue groups" with a representative in (nearly) every member of the set
- But we only have a sample of the species, not every member…
- …so use rarefaction curves to estimate core genome size.
    1. Randomly order organisms, and count number of 'core' and 'new' genes seen with each new genome addition.
    2. Repeat until you have a reasonable estimate of error/no new genes found

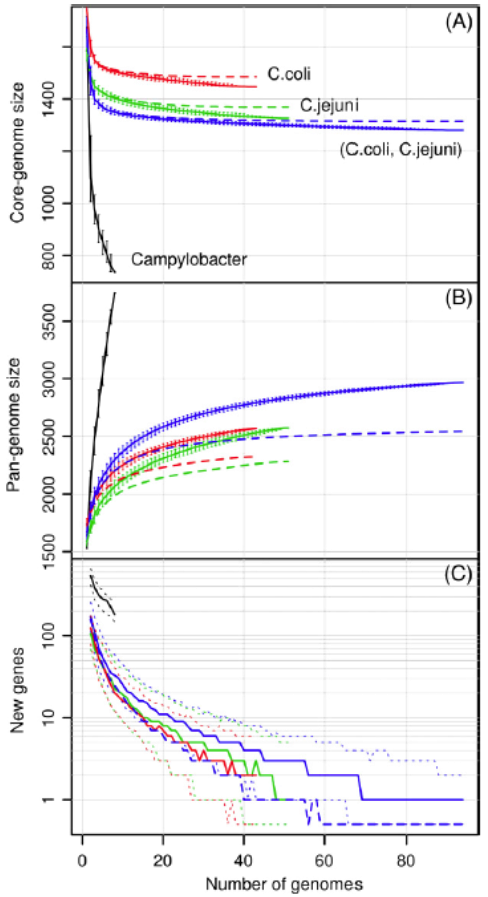


FIG. 2.—Core genome (*A*) and pan-genome (*B*) size estimates, as well as number of newly discovered genes (*C*), as a function of the number of sequenced genomes. The genome input order was randomly permuted 1,000 times. The lines describe the average number of genes (using median statistics), whereas the vertical bars delimit the second and third quartiles, with the exception of panel (*C*), where quartiles are represented by short dashed lines. On panel (*A*), the long dashed lines correspond to the average core genome size when one taxon is allowed a missing core gene, whereas on the (*B* and *C*) panels, they describe the pan-genome size or number of new genes for the combined species data set when the putative pseudogenes are excluded.

Lefébure *et al.* (2010) *Genome Biol. Evol.* doi:10.1093/gbe/evq048

Directional Selection

- Several statistical tests for directional selection, e.g.
  - QTL sign
  - **K_a/K_s (d_N/d_S) ratio test – most commonly applied**
  - Relative rate test
- **K_a/K_s ratio:**
  - **K_a (or d_N)**: number of non-synonymous substitutions per non-synonymous site
  - **K_s (or d_S)**: number of synonymous substitutions per synonymous site
  - Ka/Ks > 1 ⇒ positive selection; Ka/Ks < 1 ⇒ stabilising selection
  - Several methods/tools for calculation
    - PAML (http://abacus.gene.ucl.ac.uk/software/paml.html)
    - SeqinR (http://cran.r-project.org/web/packages/seqinr/index.html)

Take-Home Messages

- **Comparative genomics is a powerful set of techniques for:**
  - Understanding and identifying evolutionary processes and mechanisms
  - Reconstructing detailed evolutionary history of a set of organisms
  - Identifying and understanding common genomic features of organisms
  - Providing hypotheses about gene function for experimental investigation
- **A huge amount of data is available to work with**
  - And it's only going to get much, much larger
- **Results feed into many areas of study:**
  - Medicine and health
  - Agriculture and food security
  - Basic biology in all fields
  - Systems and synthetic biology