Comparative Genomics and Visualisation – Part 2

Leighton Pritchard

The James Hutton Institute

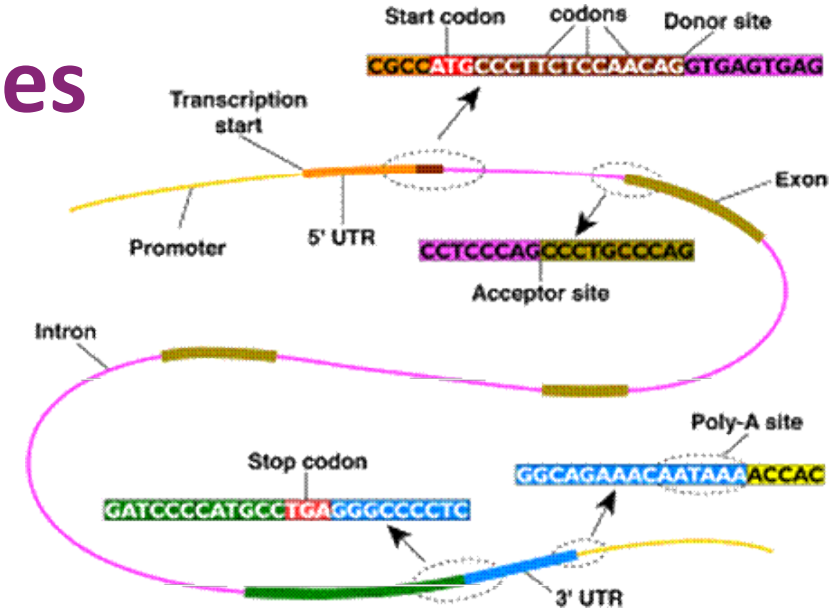
Part 2

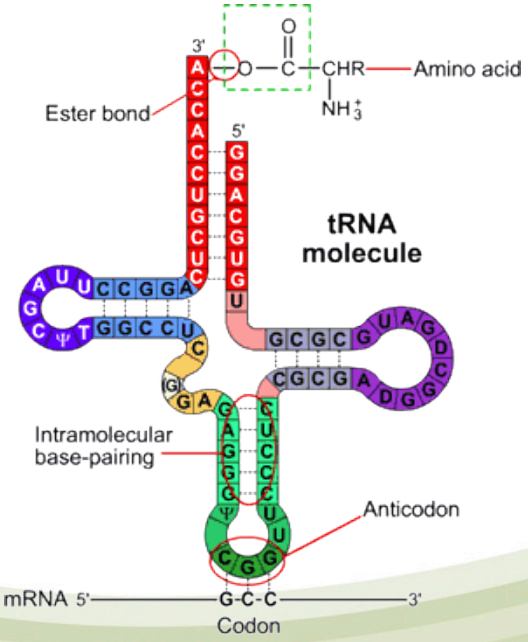
- **Part 1**
  - Experimental Comparative Genomics
  - Bulk and Whole Genome Comparisons
- **Genome Features**
- **Who let the –logues out?**
- **Finishing The Hat**

The James Hutton Institute

Genome Features

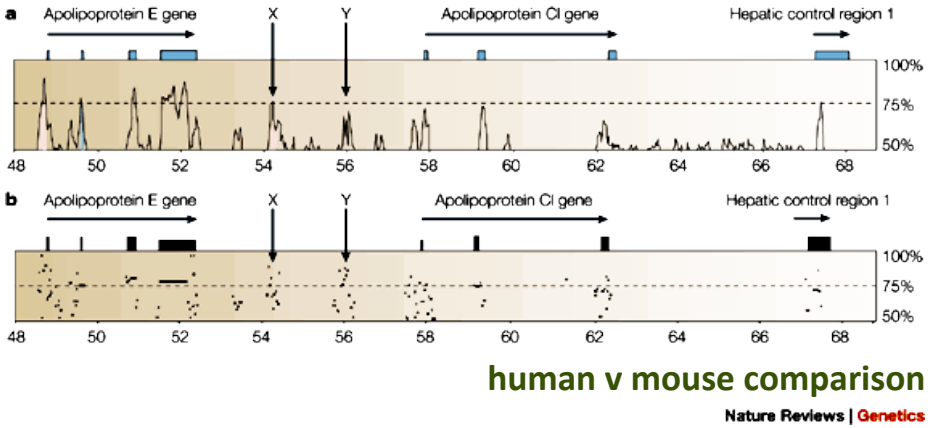
- **Genes:**
  - translation start
  - introns
  - exons
  - translation stop
  - translation terminator
- **ncRNA:**
  - tRNA – transfer RNA
  - rRNA – ribosomal RNA
  - CRISPRs – bacterial and archaeal defence (genome editing)
  - many other classes (including enhancers)





Genome Features

- **Regulatory sites**
  - Transcription start site (TSS)
  - RNA polymerase binding sites
  - Transcription Factor Binding Sites (TFBS)
  - Core, proximal and distal promoter regions



human v mouse comparison

Nature Reviews | Genetics

- **Repetitive Regions and Mobile Elements**
  - Tandem repeats
  - (retro-)transposable elements
    - *Alu* has ≈50,000 active copies in human genome
  - Phage inclusion (bacteria/archaea)

Pennacchio & Rubin (2001) *Nat. Rev. Genet.* doi:10.1038/35052548

Genome Feature Identification

The James Hutton Institute

- **Gene Finding**:
  1. Empirical (evidence-based) methods:
     - Inference from known protein/cDNA/mRNA/EST sequence
     - Inference from mapped RNA reads
  2. *Ab initio* methods:
     - Identification of sequences associated with gene features:
       - TSS, CpG islands, Shine-Dalgarno sequence, stop codons, etc.
  3. Inference from genome comparisons/conservation

Liang *et al.* (2009) *Genome Res.* doi:10.1101/gr.088997.108
Brent (2007) *Nat. Biotech.* doi:10.1038/nbt0807-883
Korf (2004) *BMC Bioinf.* doi:10.1186/1471-2105-5-59

Genome Feature Identification

The James Hutton Institute

- **Finding Regulatory Elements (short, degenerate)**:
  1. Empirical (evidence-based) methods:
     - Inference from protein-DNA binding experiments
     - Inference from coexpression
  2. *Ab initio* methods:
     - Identification of regulatory motifs (profile/other methods):
       - TATA, sigma-factor binding sites, etc.
     - statistical overrepresentation
     - Identification from sequence properties
  3. Inference from sequence conservation/genome comparisons

Zhang *et al.* (2011) *BMC Bioinf.* doi:10.1186/1471-2105-12-238
Kilic *et al.* (2013) *Nucl. Acids Res.* doi:10.1093/nar/gkt1123
Vavouri & Elgar (2005) *Curr. Op. Genet. Devel.* doi:10.1016/j.gde.2005.05.002

Genome Feature Identification

The James Hutton Institute

- All prediction methods result in errors
- All experiments have error
- Genome comparisons can help correct errors
- **[OPTIONAL ACTIVITY] – useful for exercise**
  - `predict_CDS.md` Markdown
- Other options for prokaryotic genecalling:
  - Glimmer (http://ccb.jhu.edu/software/glimmer/index.shtml)
  - GeneMarkS (http://opal.biology.gatech.edu/)
  - RAST (http://rast.nmpdr.org/)
  - BASys (https://www.basys.ca/), etc.
- Options for eukaryotic genecalling:
  - GlimmerHMM (http://ccb.jhu.edu/software/glimmerhmm/)
  - GeneMarkES (http://opal.biology.gatech.edu/gmseuk.html)
  - Augustus (http://augustus.gobics.de/), etc.

Who Let The -logues Out?

The James Hutton Institute

Evolutionary relationships of genome features can be complex.

We require precise terms to describe relationships between genome features.

Comparing Gene Features

The James Hutton Institute

- 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

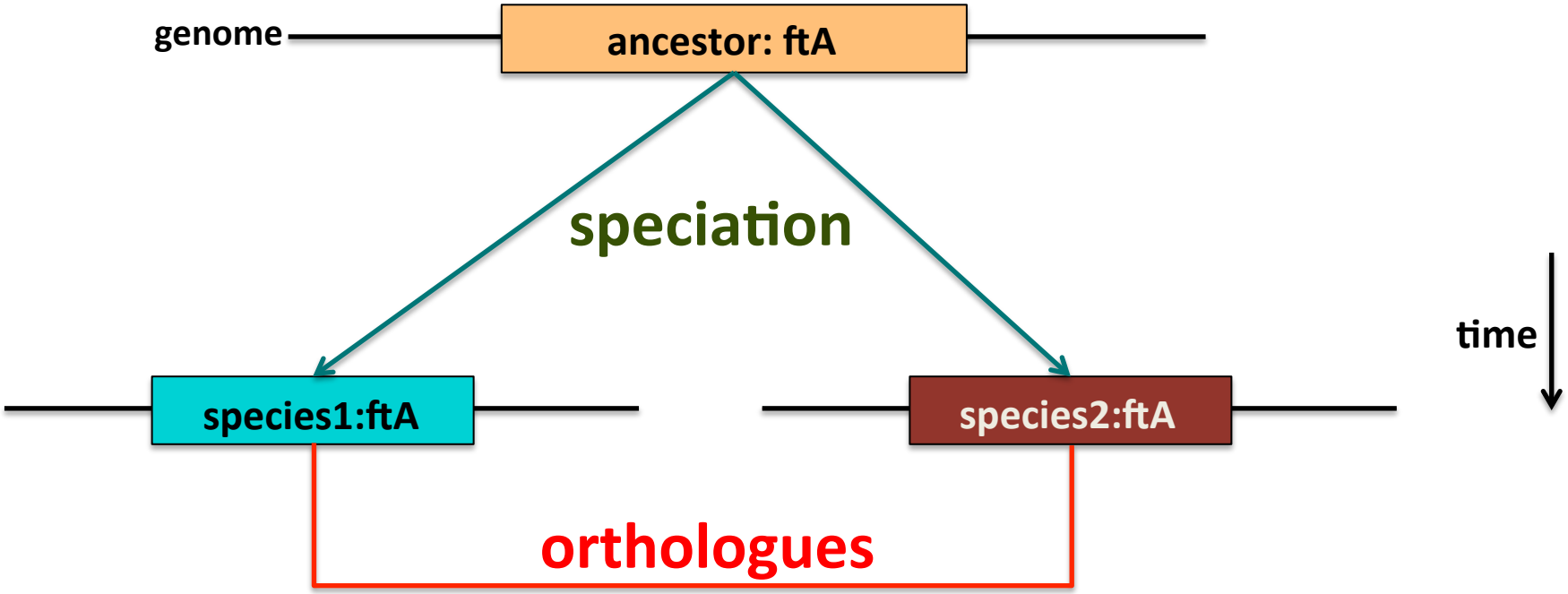
Attack of the –logues

The James Hutton Institute

- 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.)

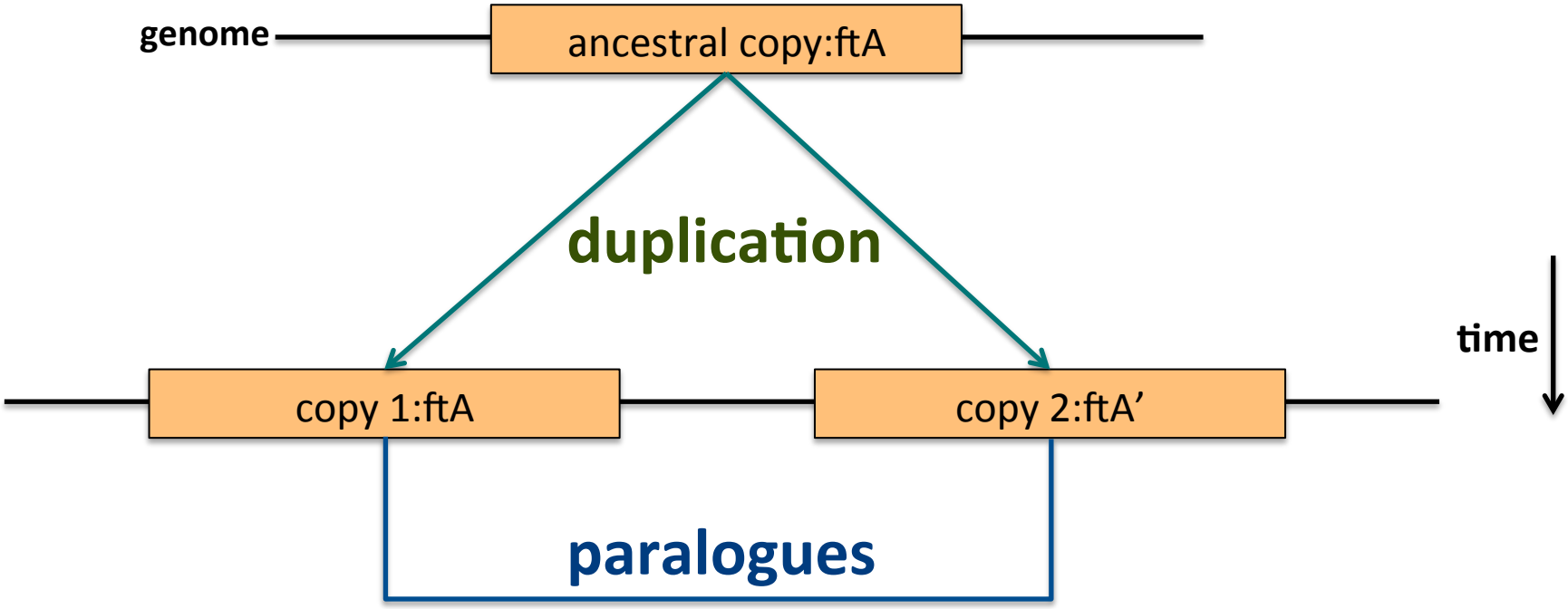
Attack of the –logues

Attack of the –logues



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

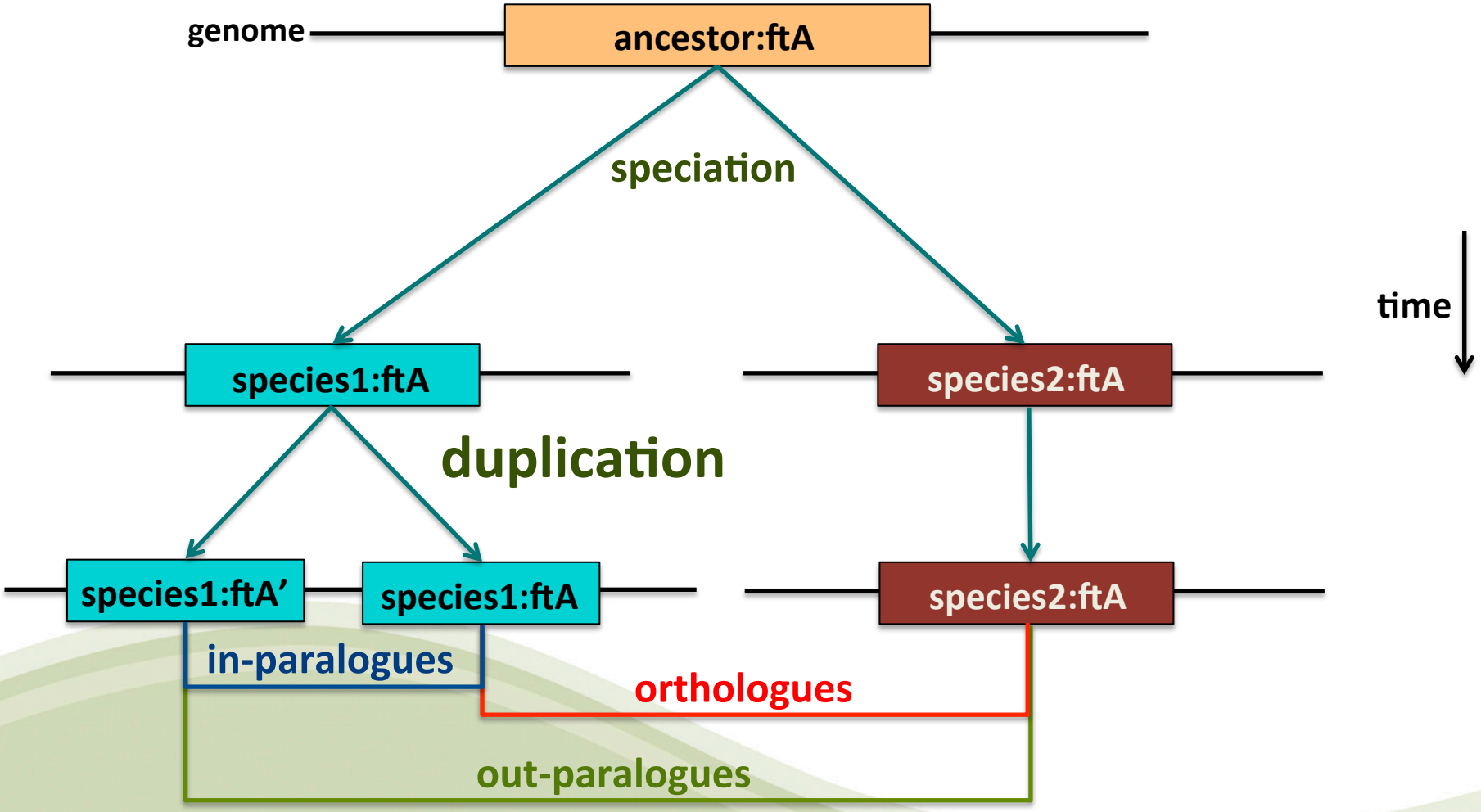
Attack of the –logues



Paralogues: homologues that diverged through duplication within the same genome

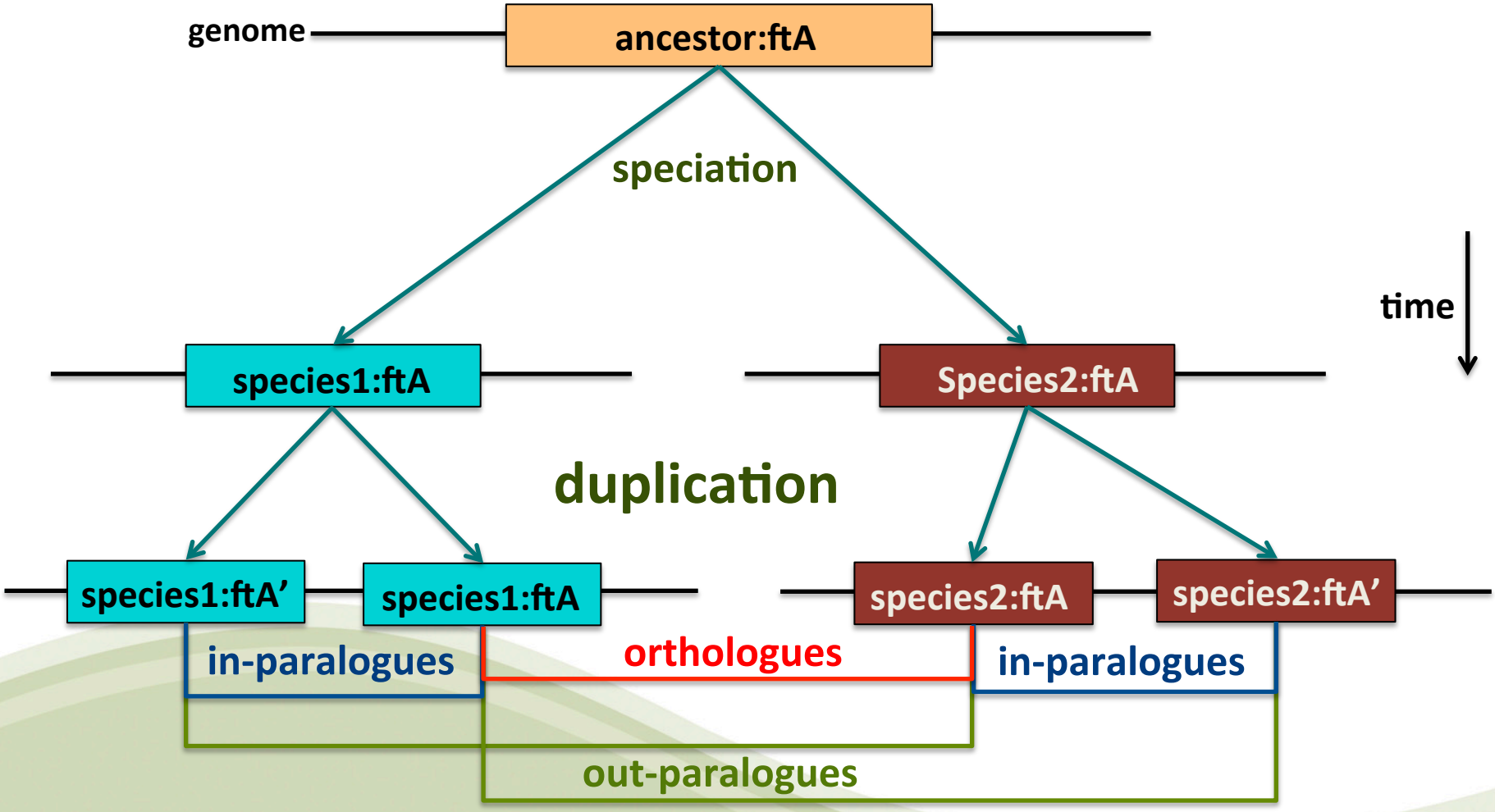
Attack of the –logues

The James Hutton Institute



Attack of the –logues

The James Hutton Institute

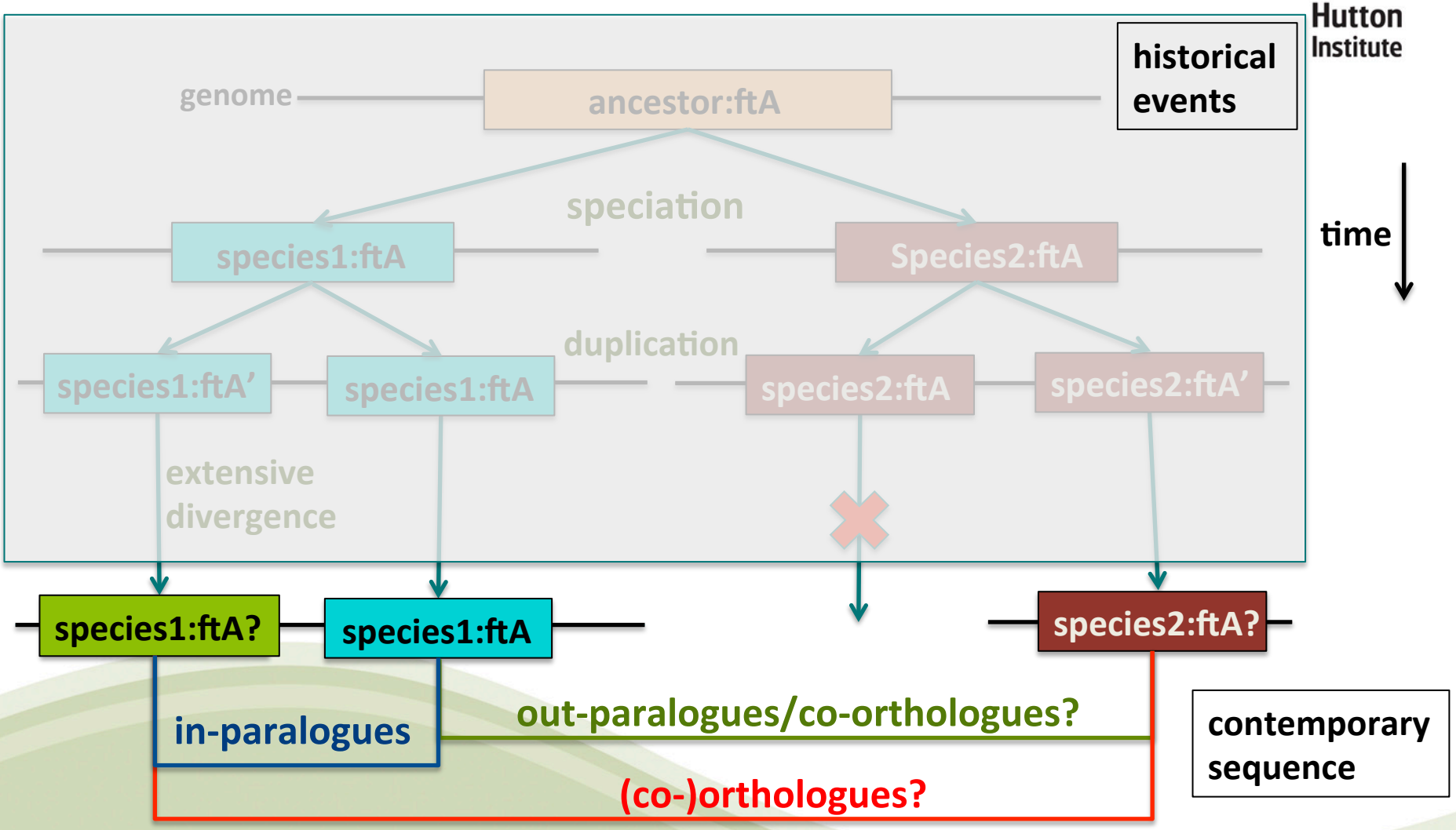


Attack of the –logues

- **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

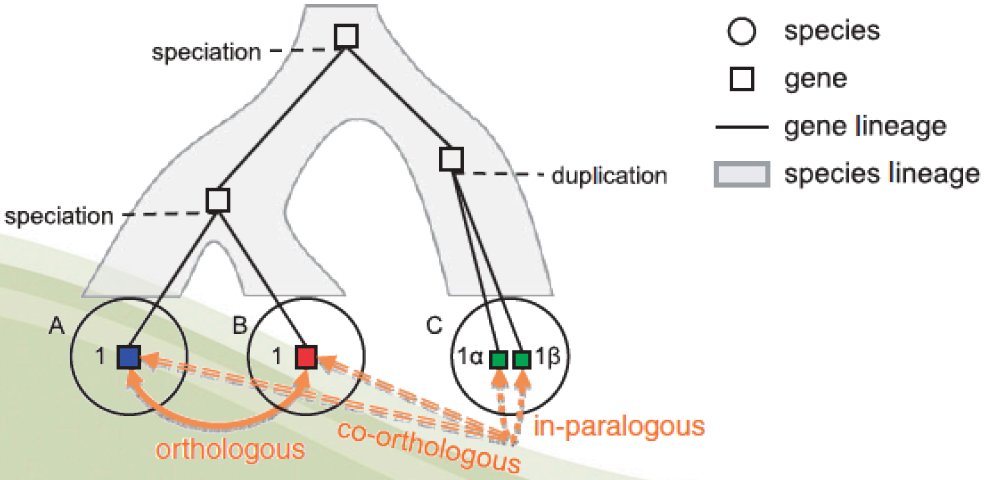
Attack of the –logues



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"

The James Hutton Institute

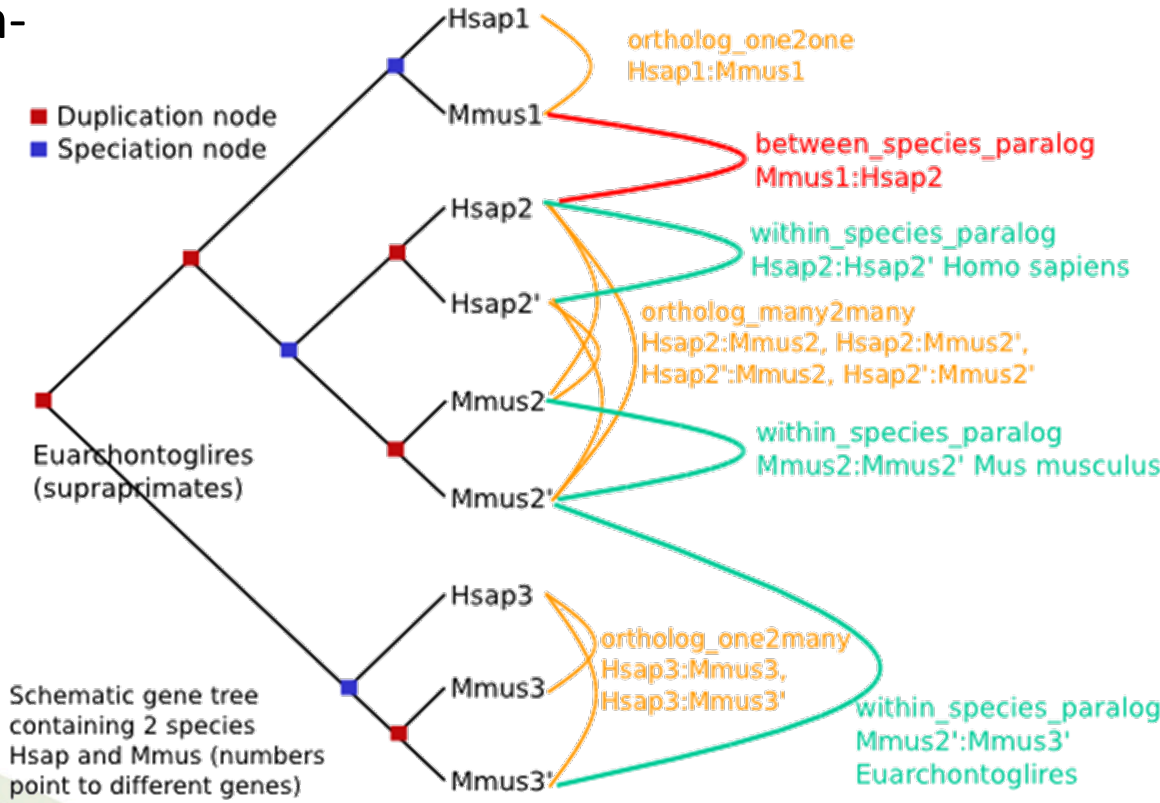
- Frequently abused/misused as a term
- "Orthology" is an evolutionary relationship, often bent into service as a functional descriptor
- Strictly defined only for two species or clades!
  - (cf. OrthoMCL, etc.)
- Orthology is not transitive (A is orthologue of C and B is orthologue of C does not imply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

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

Ensembl Compara definitions

The James Hutton Institute

- `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”

The James Hutton Institute

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)

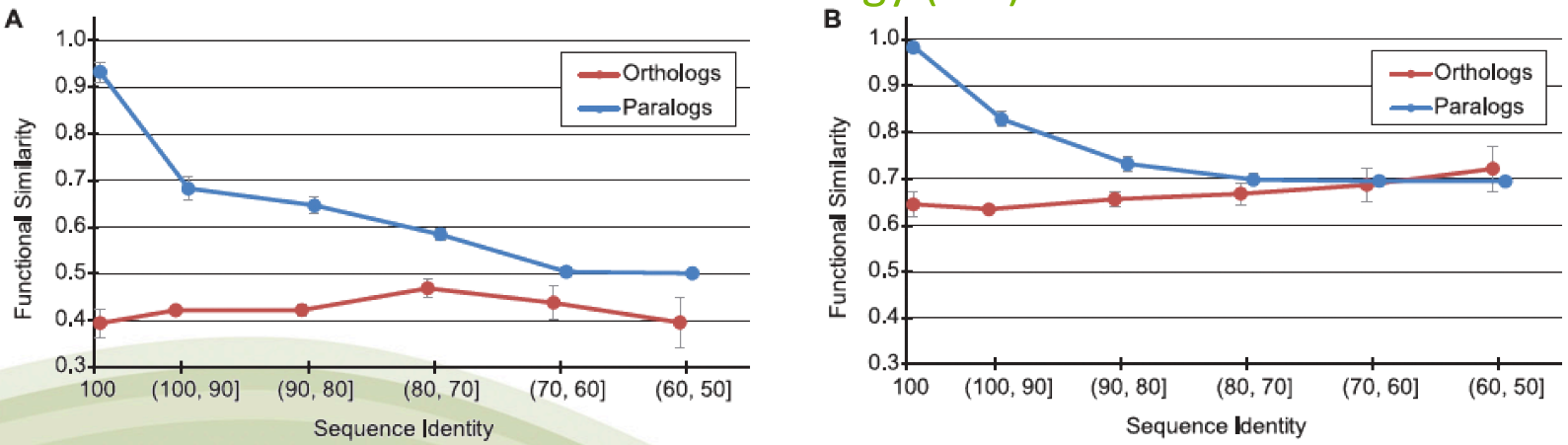


Figure 1. The relationship between functional similarity and sequence identity for human-mouse orthologs (red) and all paralogs (blue). Standard error bars are shown. (**A**) Biological Process ontology, (**B**) Molecular Function ontology.
doi:10.1371/journal.pcbi.1002073.g001

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

The James Hutton Institute

- 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”

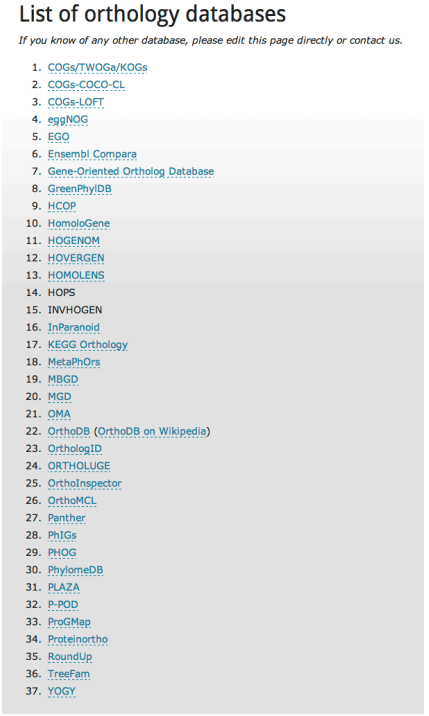
The James Hutton Institute

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

Why are "orthologues" so important?

The James Hutton Institute

- 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?

The James Hutton Institute

- 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

Pairwise approaches

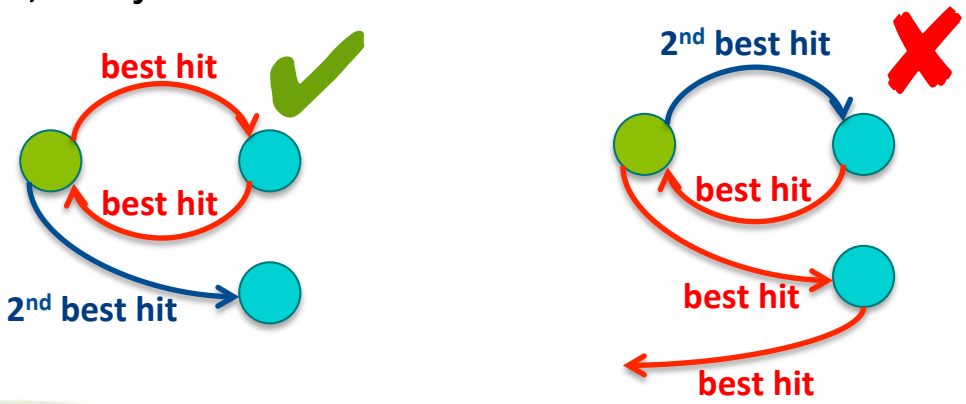
The James Hutton Institute

- S1, S2 are the gene sequence sets from two organisms
- Compare S1 to S2, and identify the most similar pairs of sequences: these are "orthologues" (or "putative orthologues").
- Many similarity measures possible (which threshold: E-value, bit score, coverage…?):
  - Reciprocal best BLAST hit (RBBH) – used by e.g. InParanoid
  - Reciprocal smallest difference (RSD) – used by e.g. RoundUp
  - and so on…
- Can be extended to multi-organism clusters by graph-based approaches

Östlund *et al.* (2009) *Nuc. Acids Res.* doi:10.1093/nar/gkp931
DeLuca *et al.* (2012) *Bioinf.* doi:10.1093/bioinformatics/bts006

Reciprocal Best BLAST Hits

- S1, S2 are the gene sequence sets from two organisms
- BLASTP:
  - Query=S1, Subject=S2
  - Query=S2, Subject=S1



- Optionally filter BLAST hits (e.g. on %identity and %coverage)
- Find all pairs of sequences {G_{S1n}, G_{S2n}} in S1, S2 where G_{S1n} is the best BLAST match to G_{S2n} and G_{S2n} is the best BLAST match to G_{S1n}.

Reciprocal Best BLAST Hits

The James Hutton Institute

- **Advantages:**
  - quick
  - easy
  - performs surprisingly well (see later…)
- **Disadvantages:**
  - misses paralogues
  - not good at identifying gene families or *-to-many relationships without more detailed analysis.
  - no strong theoretical/phylogenetic basis.

COG

- **COG (Clusters of Orthologous Groups; now POG, KOG, eggNOG etc.)**
- Graph extension of RBBH to clusters of mutual RBBH
  - "Any group of at least three proteins from different genomes, more similar to each other than any other proteins from those genomes, are an orthologous family."
  - Conduct RBBH
  - Collapse paralogues
  - Detect "triangles"
  - Merge triangles having common side
  - Manual curation
- Databases have many outparalogues

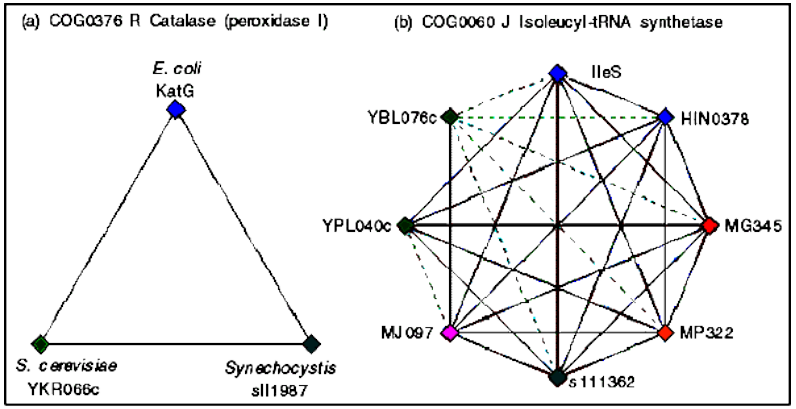
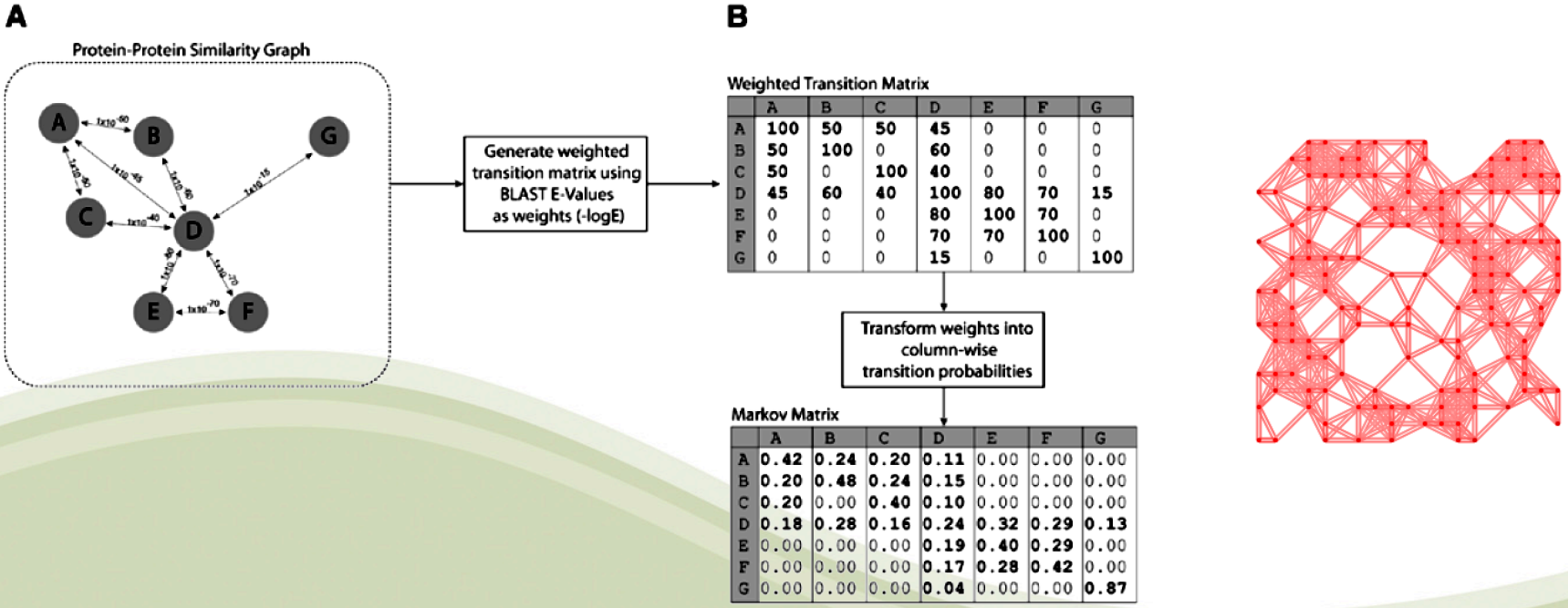


Figure 1. A minimal COG consists of three genes from three different lineages and can be illustrated by a triangle (Figure 1a). COGs were expanded by combining triangles that share sides (Figure 1b). Each of the seven genomes on the COG Web site is represented by a specific color, allowing ready visualization of the orthologous or paralogous relationships between the various genome sequences. Reciprocal best matches between genomes are represented by a solid line. Paralogous relationships, where one genome sequence has a best match to a sequence in another genome but the reverse is not true, are indicated by dashed lines (Figure 1b).

Tatusov *et al.* (2000) *Nucl. Acids Res.* doi:10.1093/nar/28.1.33

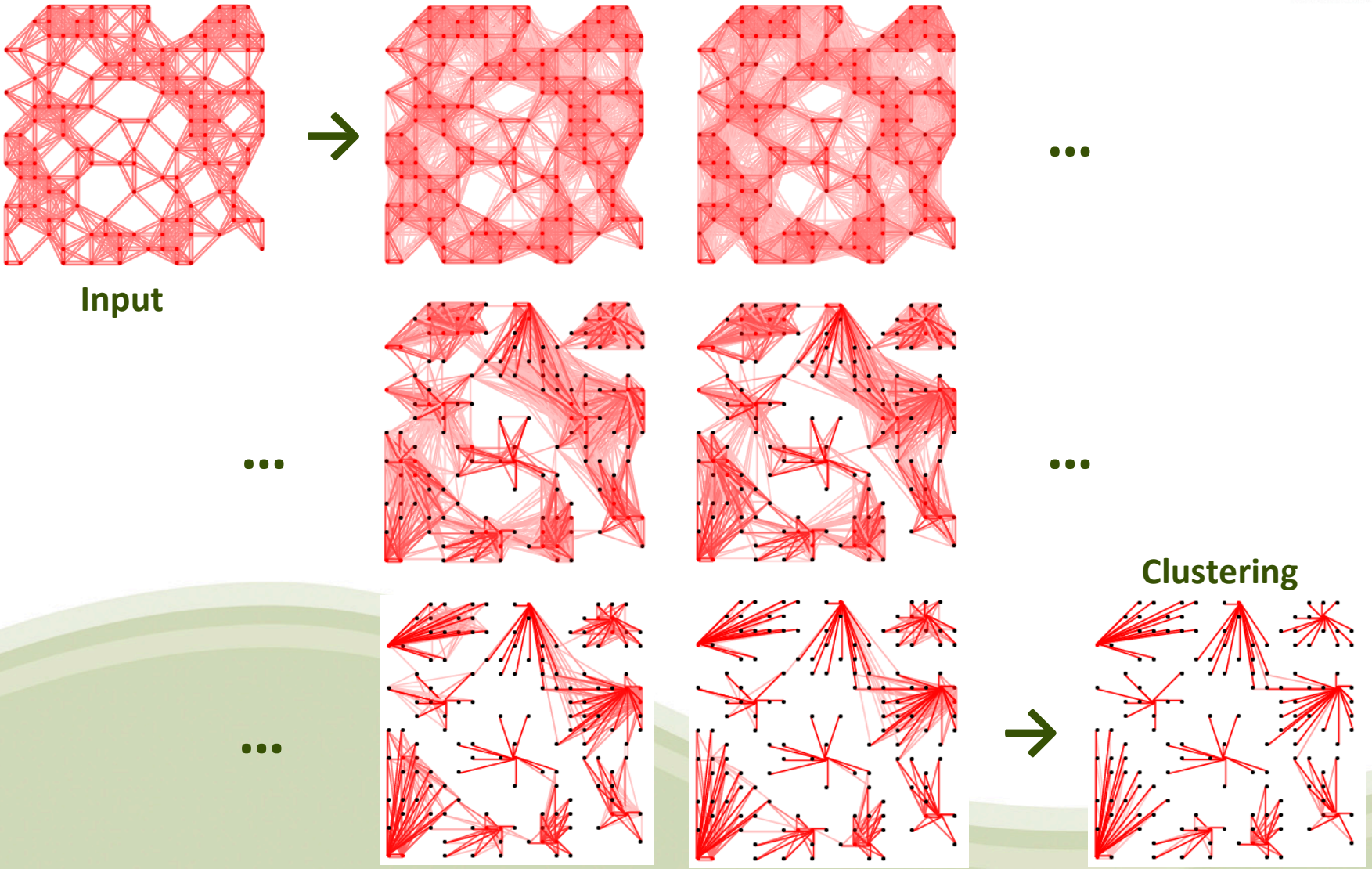
MCL

- MCL constructs a network from all-vs-all BLAST results
- Then applies matrix operations: expansion and inflation
- Iterative *expansion* and *inflation* until network convergence



Enright *et al.* (2002) *Nucl. Acids Res.* doi:10.1093/nar/30.7.1575

MCL

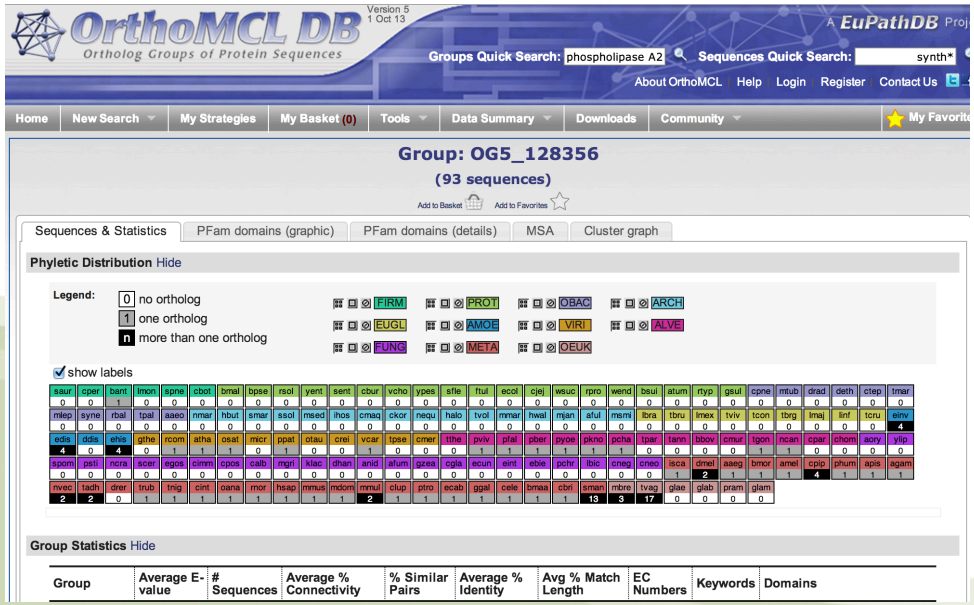


OrthoMCL

- http://orthomcl.org/orthomcl/

  1. Defines potential inparalogue, orthologue and co-orthologue pairs (using RBBH! – see algorithm description in papers directory)
  2. Applies MCL to cluster inparalogue, orthologue, co-orthologue pairs/

- Output clusters include both orthologues and paralogues



Li *et al.* (2003) *Genome Res.* doi:10.1101/gr.1224503

Notes of Caution

The James Hutton Institute

- BLAST-based orthology methods (e.g. RBBH, InParanoid, COG) are fast!
- But they have some drawbacks:
  - No guarantee that sequence matches are transitive (A may match B at a domain differently than B matches C)
  - No evolutionary distance model
  - Multiple domain matches are not accounted for
- These methods find similar sequences, then make assumptions based on similarity and number of matches. **They do not detect orthologues directly!**
- Tree-based methods incorporate:
  - Evolutionary distance
  - Direct orthologue detection

Finding “Orthologues”

The James Hutton Institute

- Pairwise analysis: RBBH
- [ACTIVITY]
  - `find_rbbh.ipynb` iPython notebook
- Multi-organism analysis: MCL
- [ACTIVITY]
  - `mcl_orthologues/README.md` Markdown
  - `mcl_orthologues.ipynb` iPython notebook

Other Methods

The James Hutton Institute

- Synteny-based:
  - **Homologene (NCBI):**
    - http://www.ncbi.nlm.nih.gov/homologene
- Manual curation:
  - **Mouse Genome Database (MGD):**
    - http://www.informatics.jax.org/homology.shtml
- Tree-based:
  - **EnsemblCompara (EMBL-EBI):**
    - http://www.ensembl.org/info/genome/compara/index.html
  - **TreeFam (EMBL-EBI):**
    - http://www.treefam.org/
  - **OrthologID:**
    - http://nypg.bio.nyu.edu/orthologid/

Evaluating Orthologue Predictions

The James Hutton Institute

Which method works best?

(and what do we mean by “best” anyway?)

Evaluating Predictions

The James Hutton Institute

- Works the same way for all prediction tools

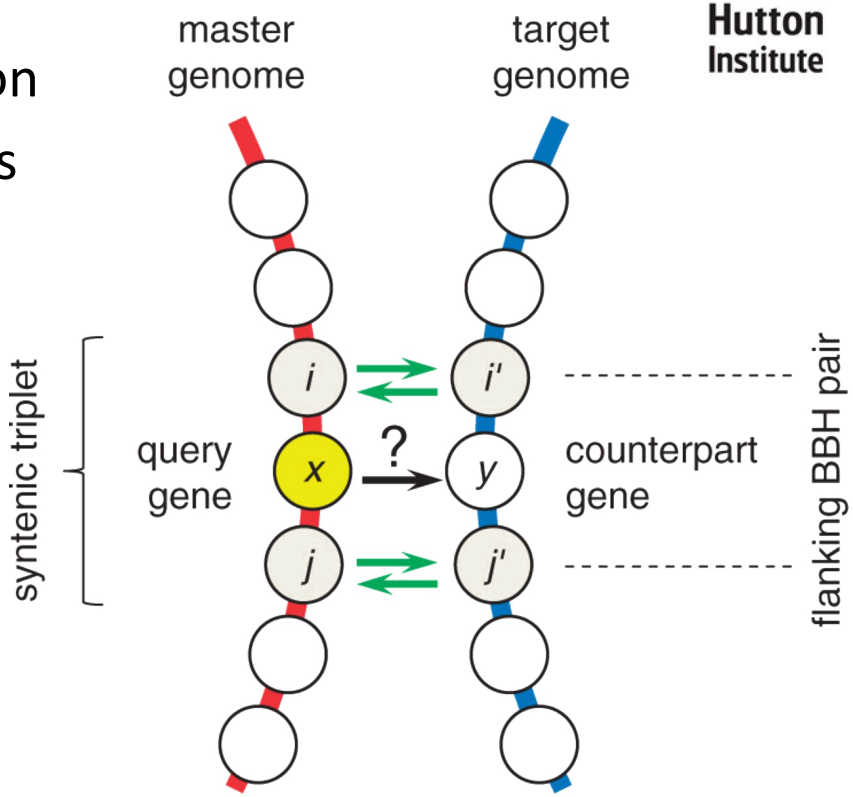
1. Define a "validation set" (gold standard), unseen by the prediction tool
2. Make predictions with the tool
3. Evaluate confusion matrix and performance statistics
   - Sensitivity
   - Specificity
   - Accuracy

Standard:	+ve	-ve
Predict +ve	TP	FP
Predict -ve	FN	TN

False positive rate	FP/(FP+TN)
False negative rate	FN/(TP+FN)
Sensitivity	TP/(TP+FN)
Specificity	TN/(FP+TN)
False discovery rate (FDR)	FP/(FP+TP)
Accuracy	(TP+TN)/(TP+TN+FP+FN)

Evaluating Orthologue Predictions

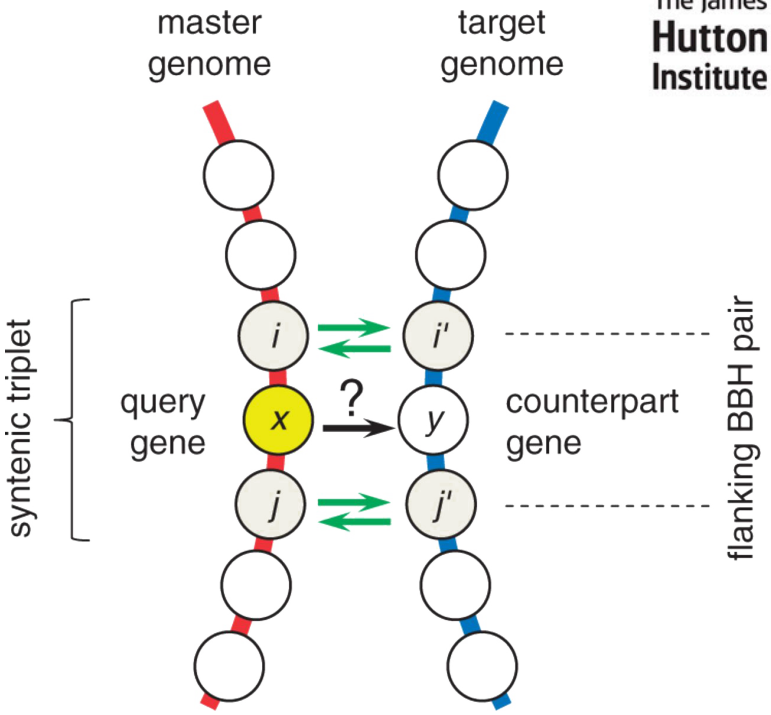
- Take advantage of prokaryotic operon structure: conserved syntenic triplets likely to be orthologous
- Idea: If the outer pair in a syntenic triplet are orthologous, the middle gene is likely to be, too.
  - Middle genes are orthologue "gold standard"
- Do RBBH reliably identify middle genes from syntenic triplets?

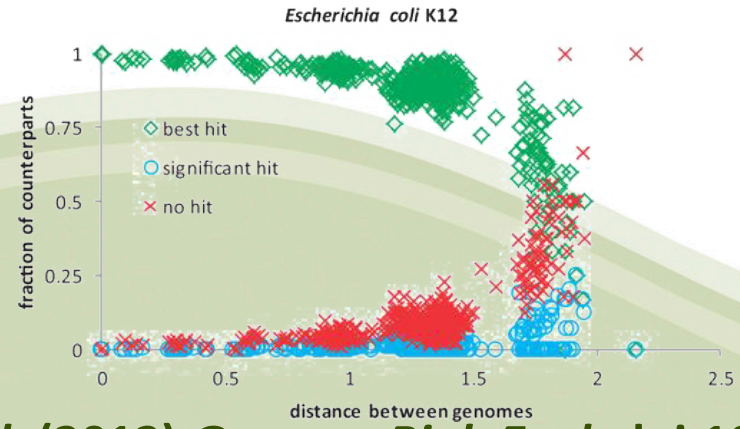


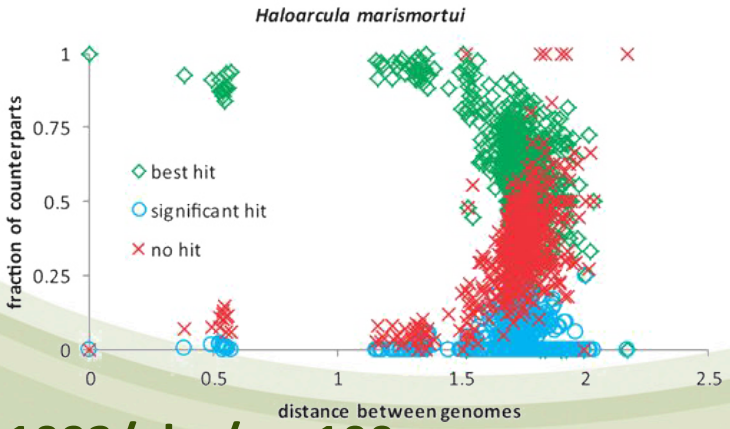
Wolf *et al.* (2012) *Genome Biol. Evol.* doi:10.1093/gbe/evs100

Evaluating Orthologue Predictions

- Two well-characterised genomes compared against 573 prokaryotes
- Identified RBBH (with permissive BLAST settings)
- "Overwhelming majority" of middle genes (counterparts) are BBH
- 88-99% of BBH are in syntenic triplets
- Therefore, RBBH reliably finds orthologues







Wolf *et al.* (2012) *Genome Biol. Evol.* doi:10.1093/gbe/evs100

Evaluating Orthologue Predictions

The James Hutton Institute

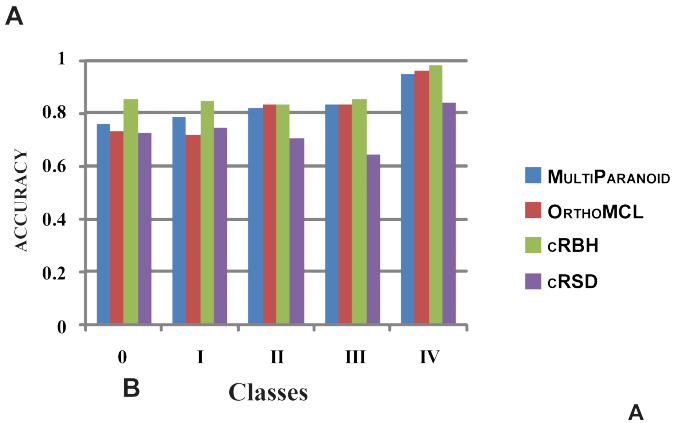
- Four orthologue prediction algorithms:
  - RBBH (and cRBH)
  - RSD (and cRSD)
  - MultiParanoid
  - OrthoMCL
- Tested against 2,723 curated orthologues from six *Saccharomycetes*
- Rated by:
  - Sensitivity: TP/(TP+FN) – what proportion of orthologues are found
  - Specificity: TN/(TN+FP) – how well are non-orthologues excluded
  - Accuracy: (TP+TN)/(TP+TN+FP+FN) – general measure of performance
  - FDR: FP/(FP+TP) – what proportion of predictions are incorrect

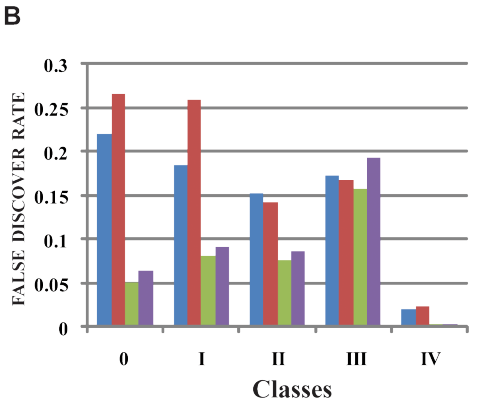
Salichos *et al.* (2011) *PLoS One.* doi:10.1371/journal.pone.0018755

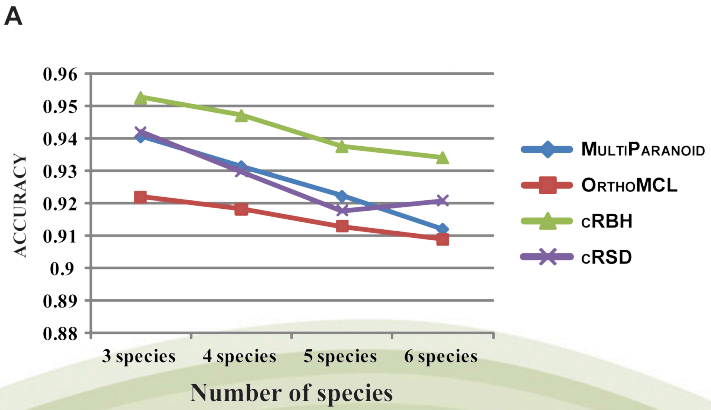
Evaluating Orthologue Predictions

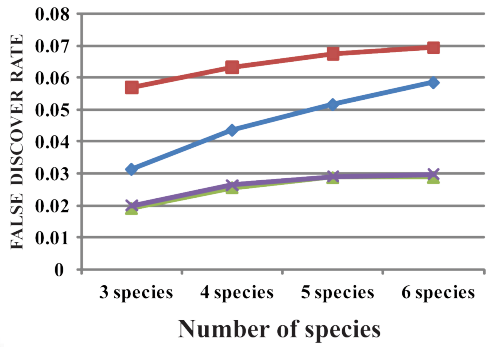
The James Hutton Institute

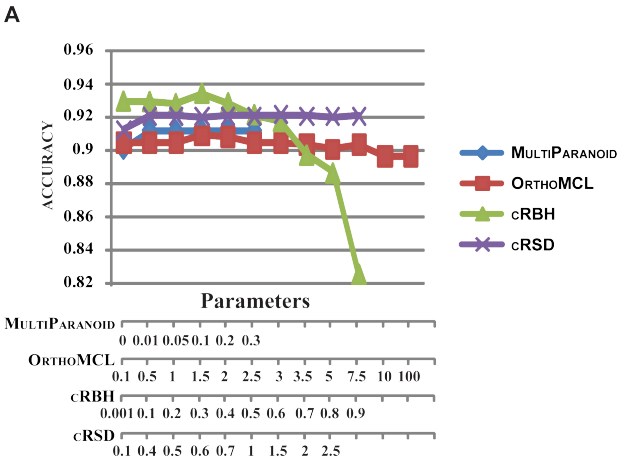
- Four orthologue prediction algorithms:
  - RBBH (cRBH)
  - RSD (cRSD)
  - MultiParanoid
  - OrthoMCL











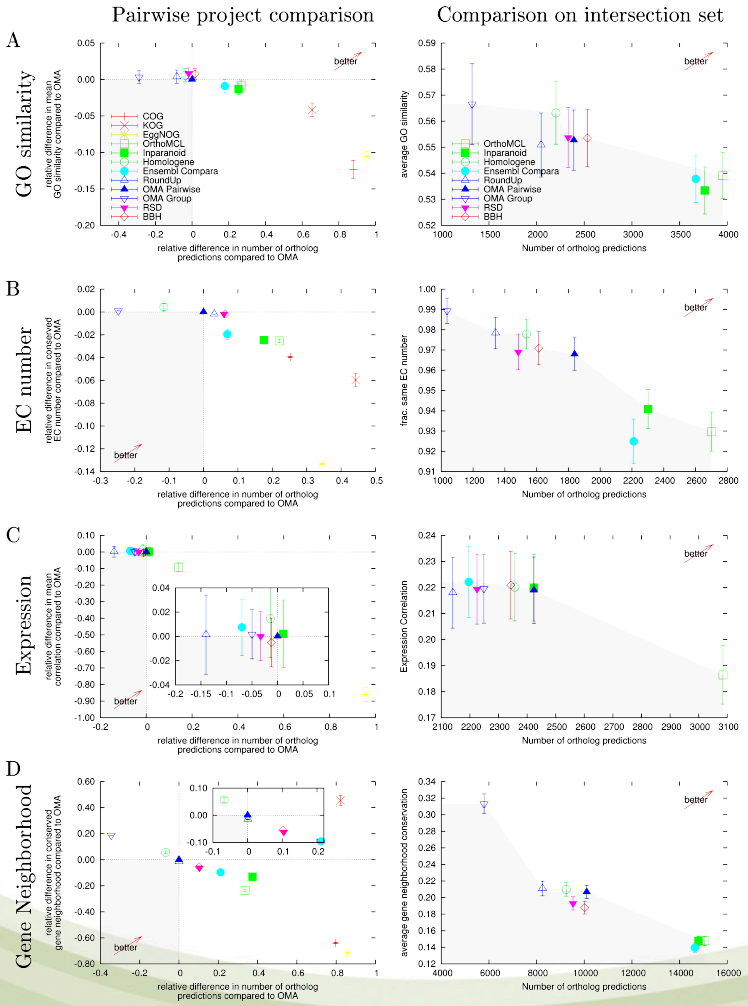
- cRBH most accurate, and specific, with lowest FDR

Salichos *et al.* (2011) *PLoS One.* doi:10.1371/journal.pone.0018755

Evaluating Orthologue Predictions

The James Hutton Institute

- Tests of several methods on a number of literature-based benchmarks for:
  - Correct branching of phylogeny
  - Grouping by function
    - GO similarity
    - EC number
    - Expression level
    - Gene Neighbourhood



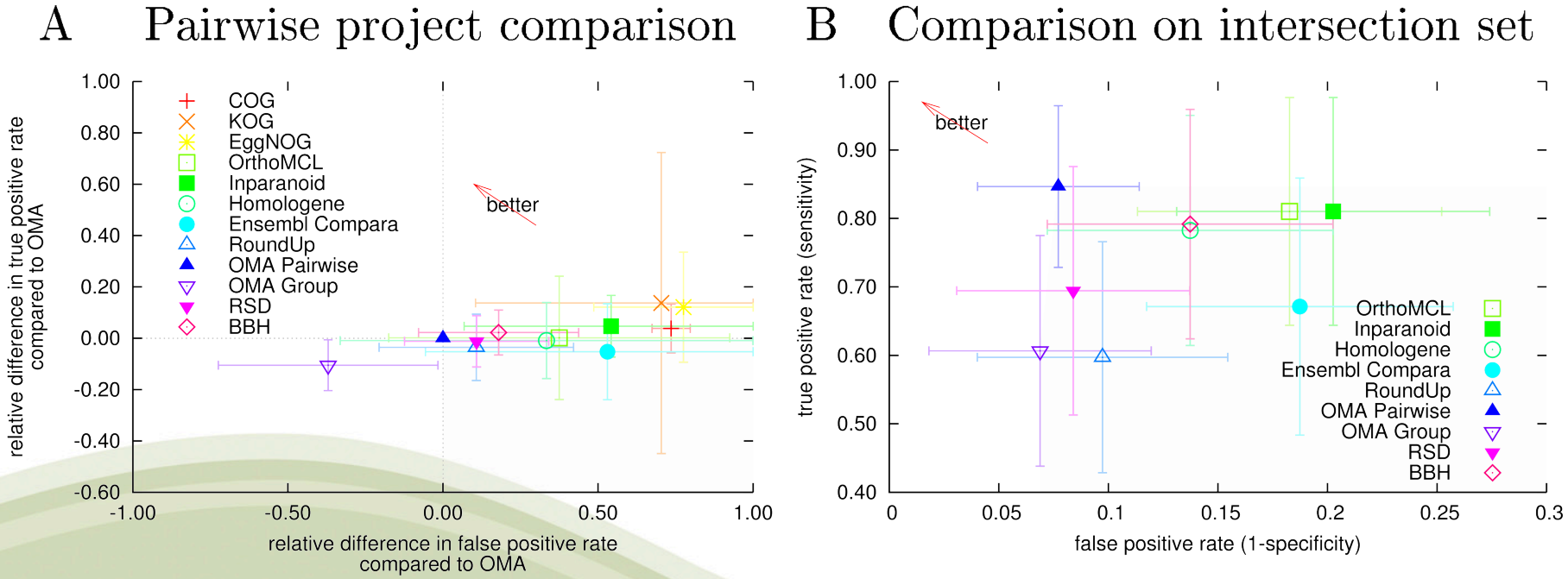
Altenhoff & Dessimoz (2009) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1000262

Evaluating Orthologue Predictions

The James Hutton Institute

phylogenetic tests. Furthermore, we show that standard bidirectional best-hit often outperforms projects with more complex algorithms. First, the present study provides guidance for the broad community of orthology data users as to which database best suits their needs. Second, it introduces new methodology to verify orthology. And third, it sets performance standards for current and future approaches.

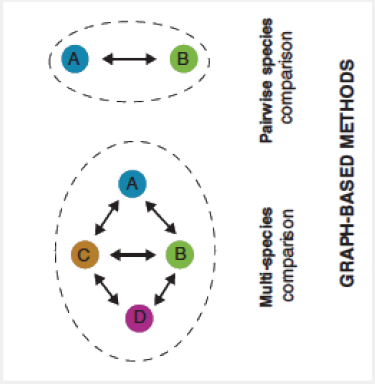
Literature



Altenhoff & Dessimoz (2009) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1000262

Evaluating Orthologue Predictions

- 70 gene family test, multiple evolutionary scenarios
- Tested databases with associated algorithms:



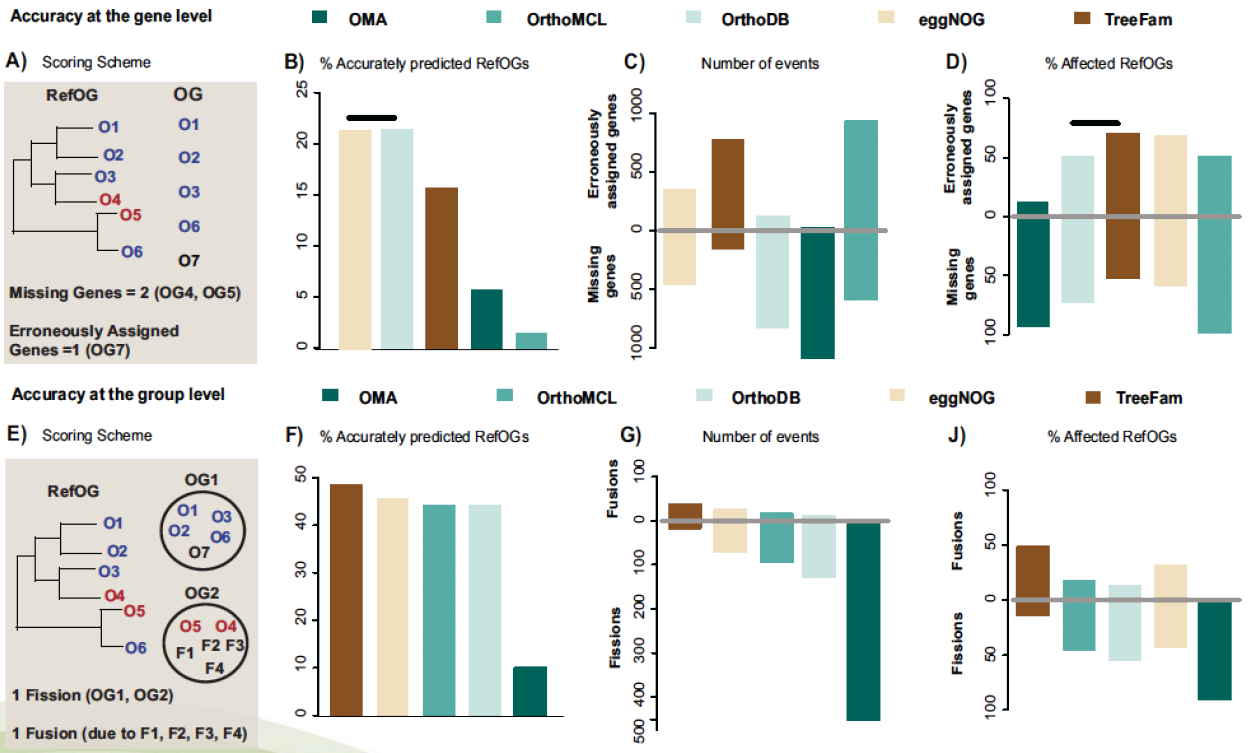
	Phylogenetic distribution	Paralogs	Homology search	Clustering strategy	Hierarchical groups
BRH	ALL*	NO	BLAST	None	-
InParanoid	ALL*	YES	BLAST	None	-
RoundUp	ALL*	NO	Evol. Distance	None	-
COG	ALL	YES	BLAST	Triangles	NO
eggNOG	ALL	YES	BLAST	Triangles	YES
OrthoDB	Eukaryotes	YES	PARALIGN	Triangles	YES
OMA	ALL	YES**	SIMD & Evol. Distance	Maximum weight cliques	YES
OrthoMCL	ALL	YES	BLAST	Markov Clustering	NO
TreeFam	Metazoa	YES	BLAST & HMM	Hierarchical clustering	-
Ensembl Compara	Metazoa	YES	BLAST	Hierarchical clustering	-
PhylomeDB	ALL	YES	BLAST§	None	-

Trachana *et al.* (2011) *Bioessays.* doi:10.1002/bies.201100062

Evaluating Orthologue Predictions

The James Hutton Institute

- 70 gene family test set, multiple evolutionary scenarios



- All methods/dbs have strong scope for improvement.
- OrthoMCL poor performer, TreeFam & eggNOG do best

Trachana *et al.* (2011) *Bioessays.* doi:10.1002/bies.201100062

Orthologue Prediction Performance

The James Hutton Institute

- Performance varies by choice of method and interpretation of “orthology”
- Biggest influence is genome annotation quality
- Relative performance varies with benchmark choice
- (clustering) RBBH outperforms more complex algorithms under many circumstances

Selection Pressures

The James Hutton Institute

Signs of selection pressure identifiable by comparative genomics

Selection Pressures

The James Hutton Institute

- 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

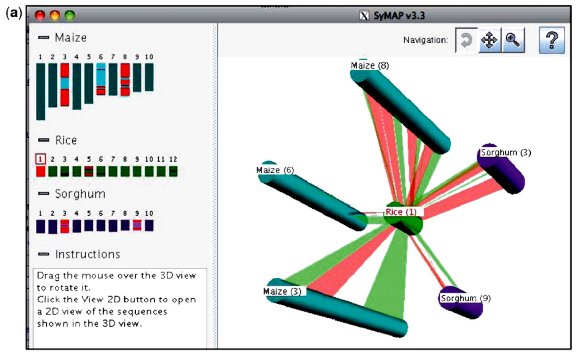
The James Hutton Institute

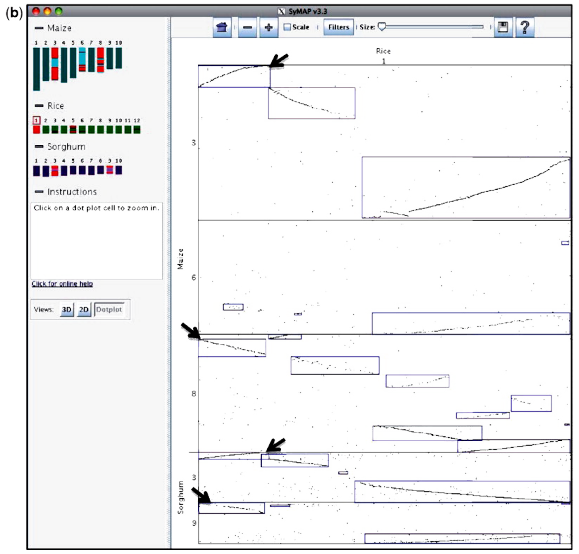
- Selective pressures depend on gene (product) function
- Genes involving physically or functionally-interacting proteins tend to evolve under similar selective constraints
- Particularly in bacteria, this leads to co-expression as *regulons* and collocation in *operons*
- Collocation (and coregulation)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

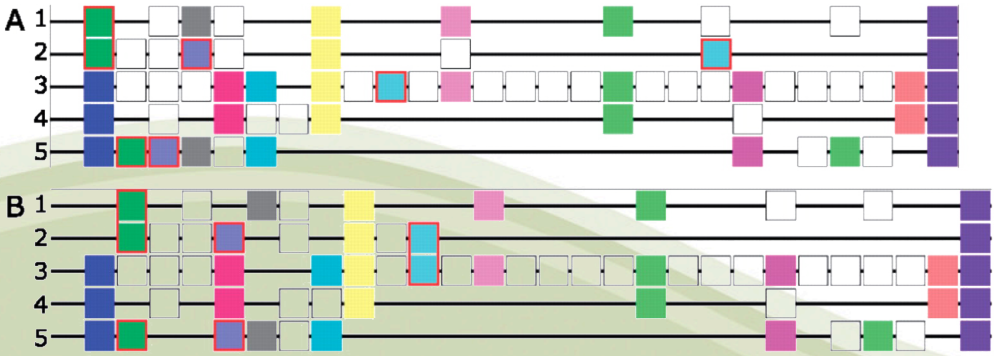
Alvarez-Ponce *et al.* (2011) *Genome Biol. Evol.* doi:10.1093/gbe/evq084

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





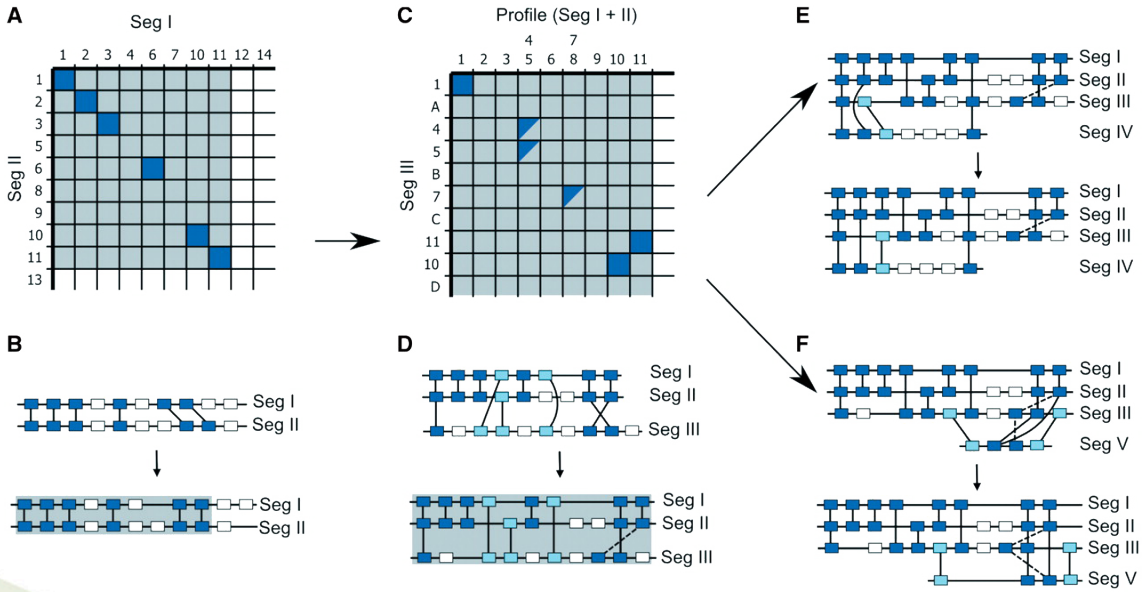


Soderlund *et al.* (2011) *Nucl. Acids. Res.* doi:10.1093/nar/gkr123

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

i-ADHoRe

- Algorithm:
  1. Combine tandem repeats of genes/gene sets
  2. Make *gene homology matrix* (GHM): identify collinear regions (diagonals) for first genome pair
  3. Convert these to *profiles*
  4. Use GG2 algorithm to align profiles
  5. Search next genome with profiles, splitting them where necessary
  6. iterate until complete

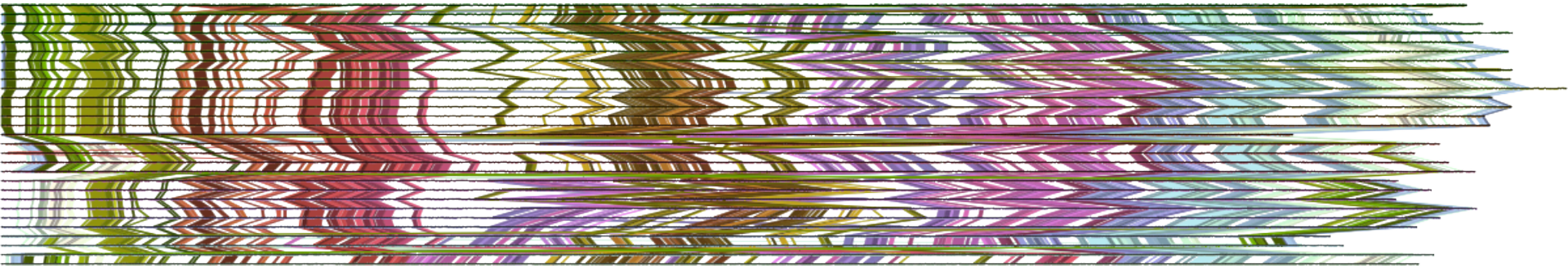


- Gives genome-scale multiple alignments of blocks of genes

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

i-ADHoRe

- [ACTIVITY]
  - `i-ADHoRe/README.md` Markdown
  - `i-ADHoRe.ipynb` iPython notebook



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

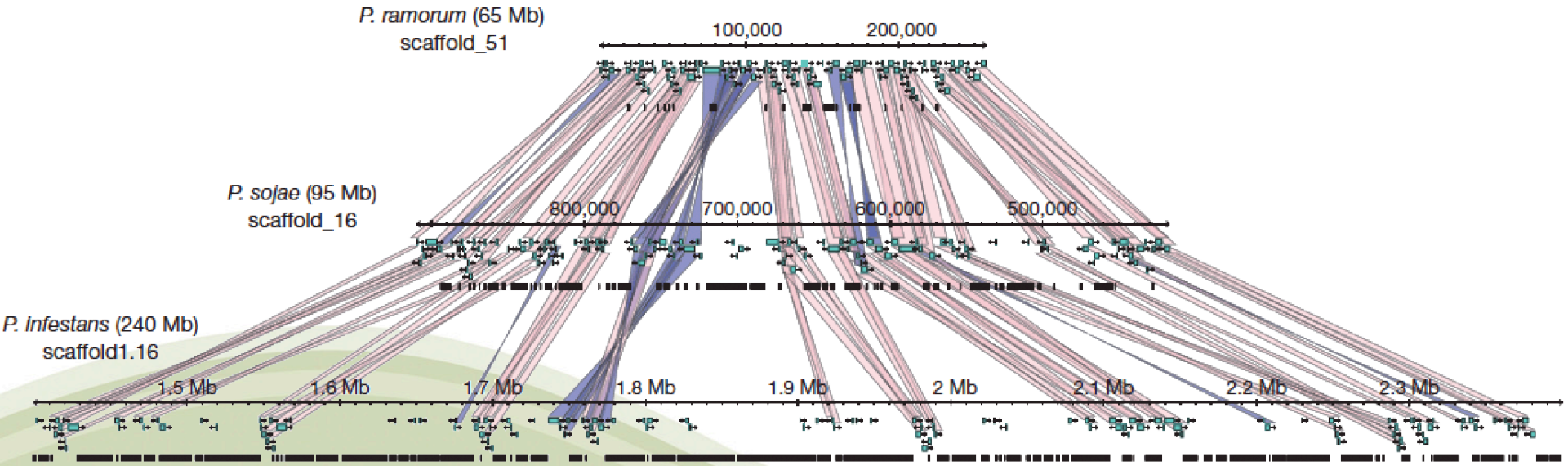
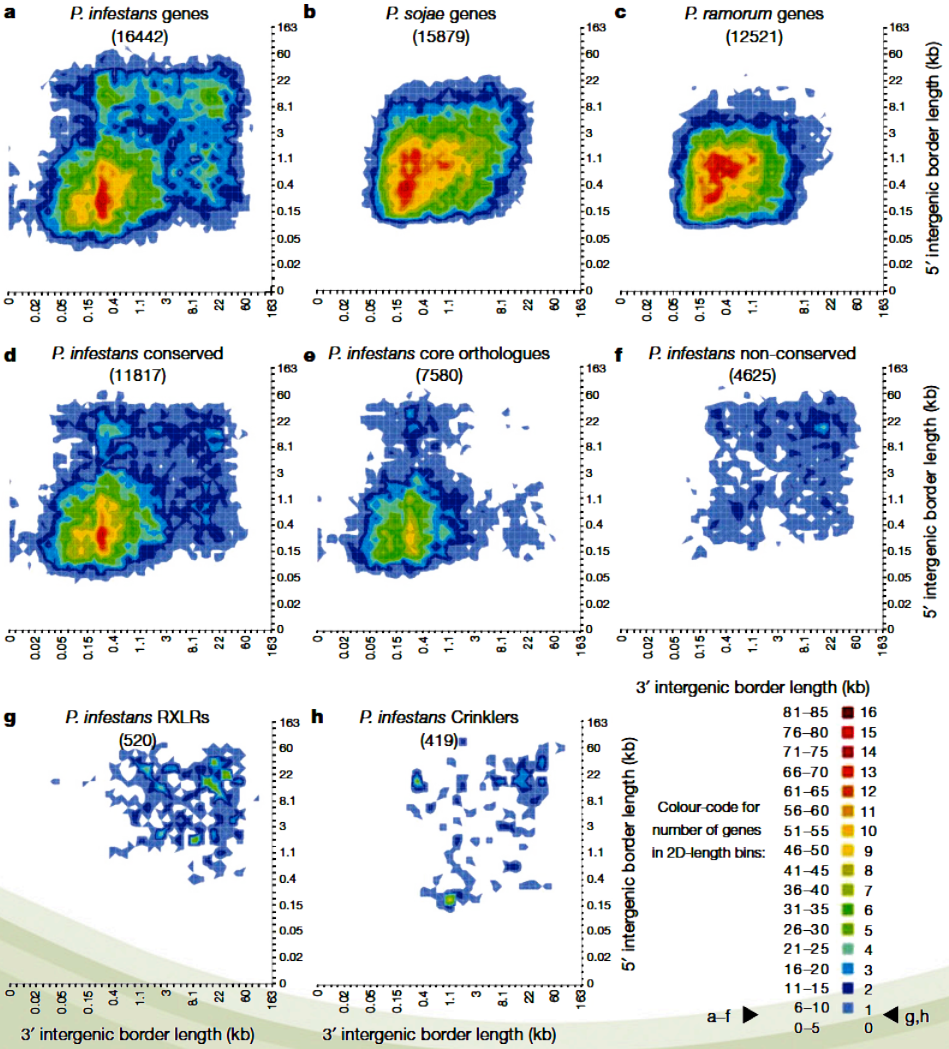


Figure 1 | Repeat-driven genome expansion in *Phytophthora infestans*. Conserved gene order across three homologous *Phytophthora* scaffolds. Genome expansion is evident in regions of conserved gene order, a consequence of repeat expansion in intergenic regions. Genes are shown as turquoise boxes, repeats 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/)

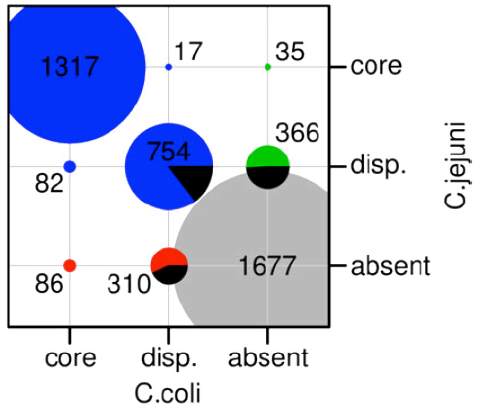


FIG. 4.—Overlap between the core and dispensable (disp.) genomic components of *Campylobacter coli* and *C. jejuni*; core genes were allowed to be missing in one strain per species. The absent/absent section represents genes that were found in other *Campylobacter* species but absent in *C. coli* and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

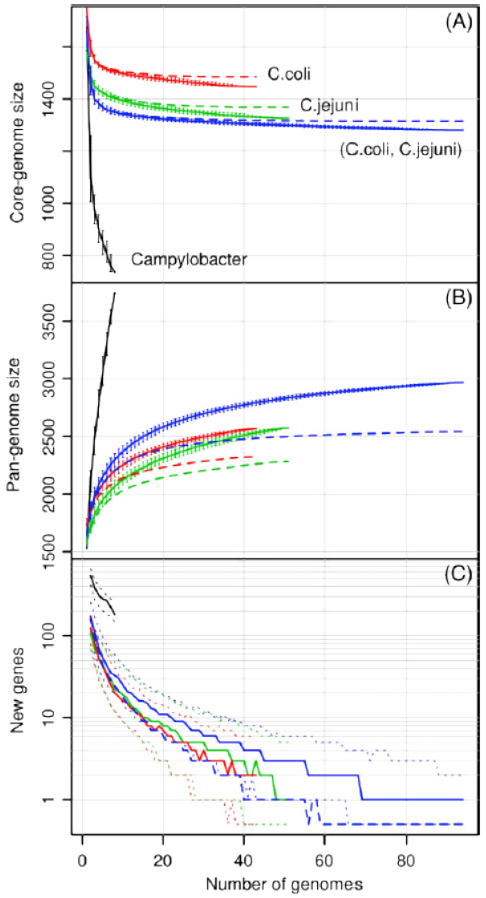


FIG. 2.—Core genome (*A*) and pan-genome (*B*) size estimates, as well as number of newly discovered genes (*C*), as a function of the number of sequenced genomes. The genome input order was randomly permuted 1,000 times. The lines describe the average number of genes (using median statistics), whereas 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

The James Hutton Institute

- 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)

Genome-Wide Positive Selection

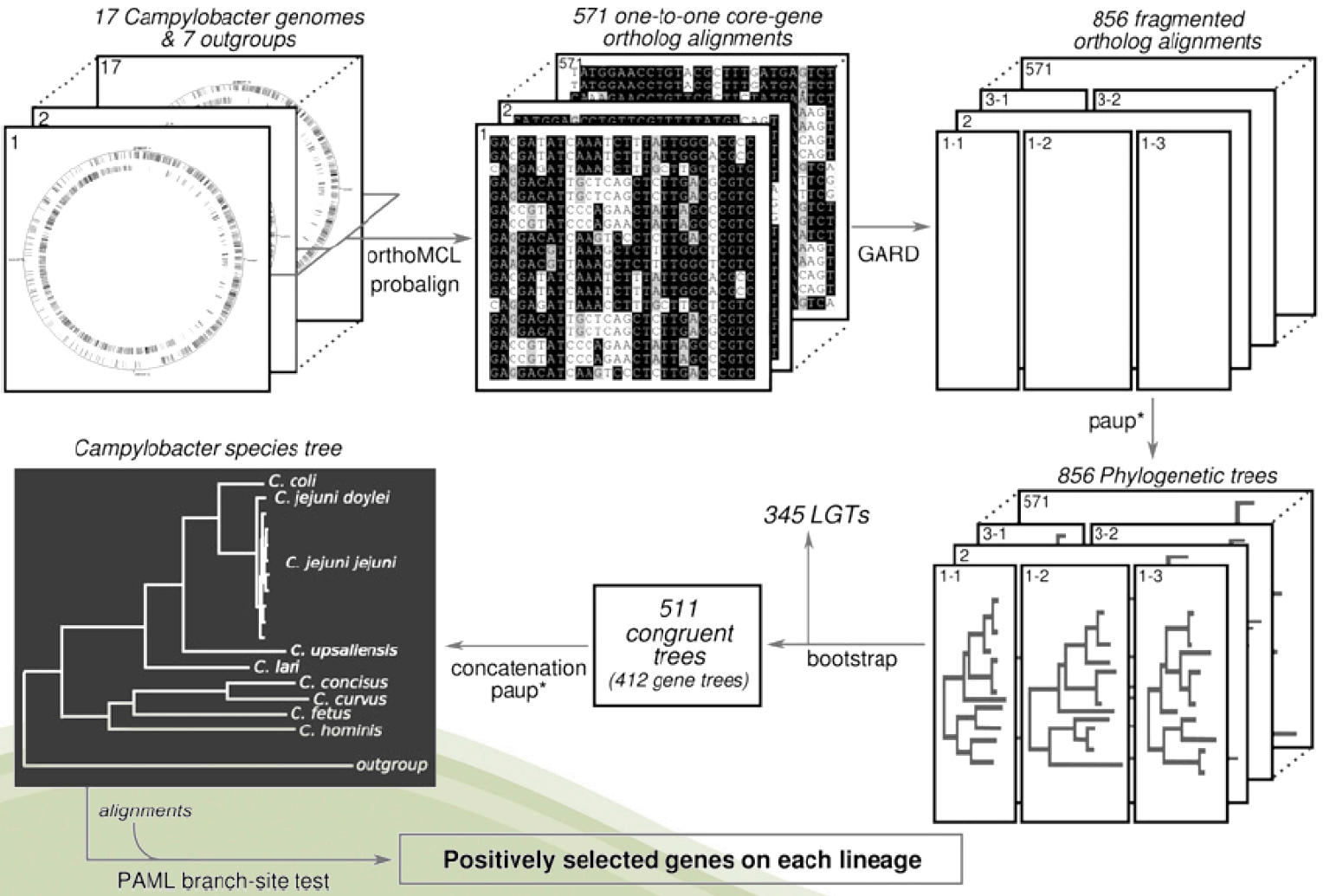
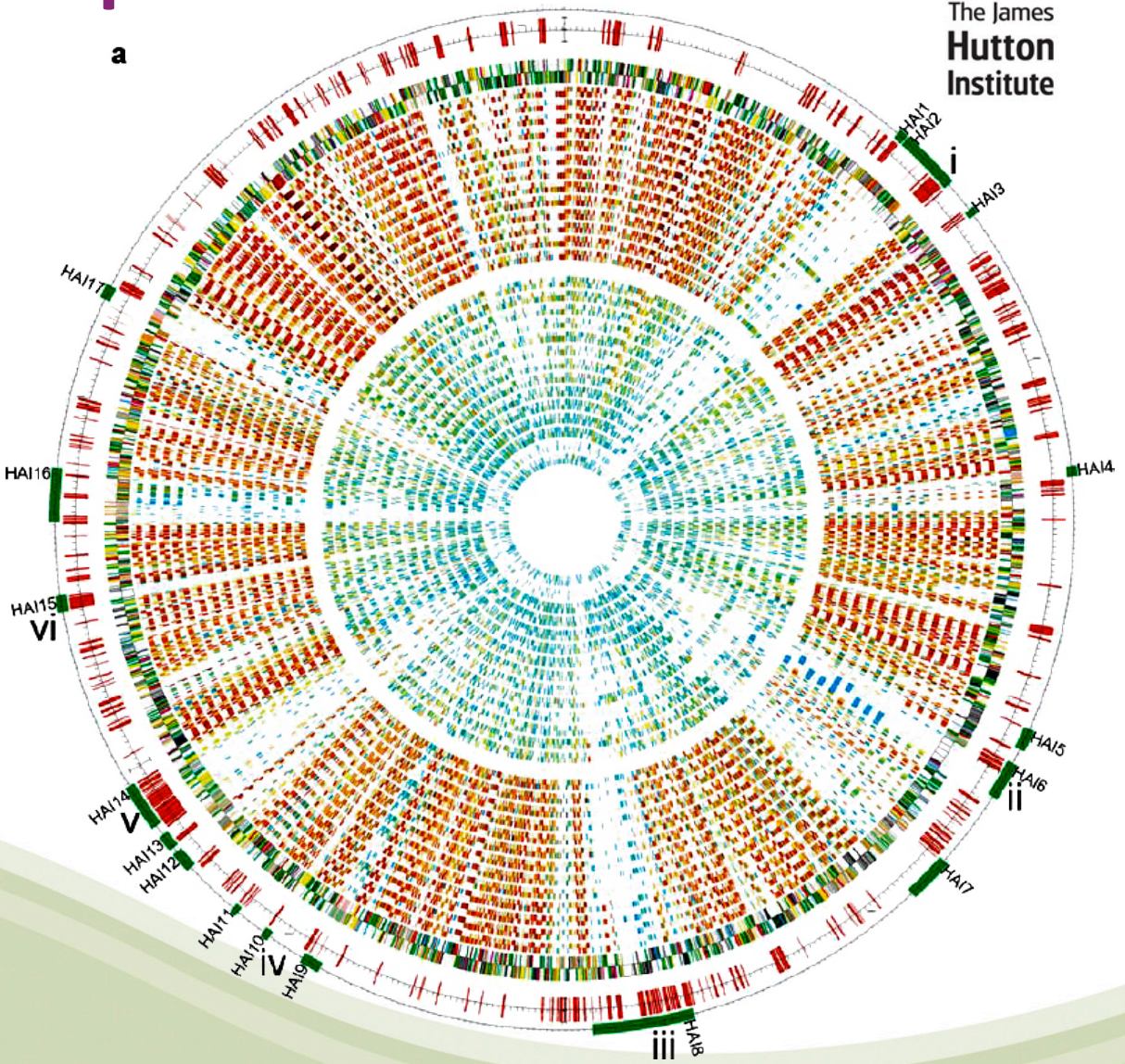


Figure 6. Genome-wide positive selection pipeline.

Lefébure & Stanhope (2009) *Genome Res.* doi:10.1101/gr.089250.108

An Analysis Output

- **Class comparison**: animal-pathogenic (APE) vs plant-associated bacteria (PAB)
- Presence of *horizontally-acquired islands* (HAI)
- Genes with greater similarity to PAB than APE



Toth *et al.* (2006) *Annu. Rev. Phytopath.* doi:10.1146/annurev.phyto.44.070505.143444

Things I Didn't Get To

The James Hutton Institute

- **Genome-Wide Association Studies (GWAS):**
  - Try http://genenetwork.org/ to play with some data
- **Prediction of regulatory elements, e.g.**
  - Kellis *et al.* (2003) *Nature* doi:10.1038/nature01644
  - King *et al.* (2007) *Genome Res.* doi:10.1101/gr.5592107
  - Chaivorapol *et al.* (2008) *BMC Bioinf.* doi:10.1186/1471-2105-9-455
  - CompMOBY: http://genome.ucsf.edu/compmoby/
- **Detection of Horizontal/Lateral Gene Transfer (HGT/LGT), e.g.**
  - Tsirigos & Rigoutsos (2005) *Nucl. Acids. Res.* doi:10.1093/nar/gki187
- **Phylogenomics, e.g.**
  - Delsuc *et al.* (2005) *Nat. Rev. Genet.* doi:10.1038/nrg1603

Finishing The Hat

The James Hutton Institute

Some of the things I hope you have taken away from the lectures/activities

Take-Home Messages

The James Hutton Institute

- **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

Take-Home Messages

The James Hutton Institute

- **Comparative genomics is essentially based around comparisons**
  - What is similar between two genomes? What is different?
- **Comparative genomics is evolutionary genomics**
- **Large datasets benefit from visualisation for effective interpretation**
  - Much scope for improvement in visualisation
- **Tools with the same purpose give different output**
  - BLAST vs MUMmer
  - RBBH vs MCL
  - **Choice of application matters for correctness and interpretation! – understand what the application does, and its limits.**

Take-Home Messages

- Comparative genomics is
  - Fun
  - Indoor work, in the warm and dry
  - Not a job that involves heavy lifting

Credits

The James Hutton Institute

- This slideshow is shared under a Creative Commons Attribution 4.0 License http://creativecommons.org/licenses/by/4.0/)
- Copyright is held by The James Hutton Institute http://www.hutton.ac.uk
- You may freely use this material in research, papers, and talks so long as acknowledgement is made.