

Comparative Genomics and Visualisation

BS32010

4. Genome Feature Comparisons



**The James
Hutton
Institute**

Leighton Pritchard^{1,2,3}

¹Information and Computational Sciences,

²Centre for Human and Animal Pathogens in the Environment,

³Dundee Effector Consortium,

The James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA



Acceptable Use Policy

Recording of this talk, taking photos, discussing the content using email, Twitter, blogs, etc. is permitted (and encouraged), providing distraction to others is minimised.

These slides will be made available on SlideShare.

These slides, and supporting material including exercises, are available at https://github.com/widdowquinn/Teaching-2015-03-17-UoD_compgenvis



Table of Contents

Comparisons of genome features

Genome Features

Who Let the -logues Out?

What's so important about orthologues?

Evaluating orthologue prediction

Using orthologue predictions

Selection Pressure

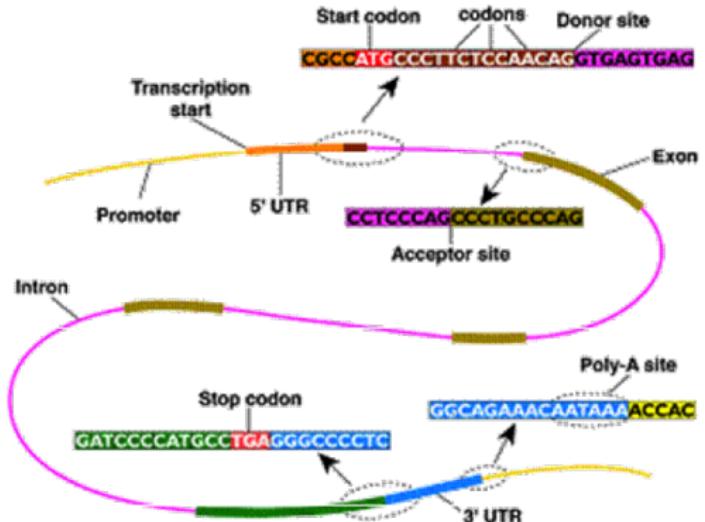
Finishing the Hat



Gene features

Significant substructure, especially in eukaryotes

- translation start
- introns
- exons
- translation stop
- translation terminator

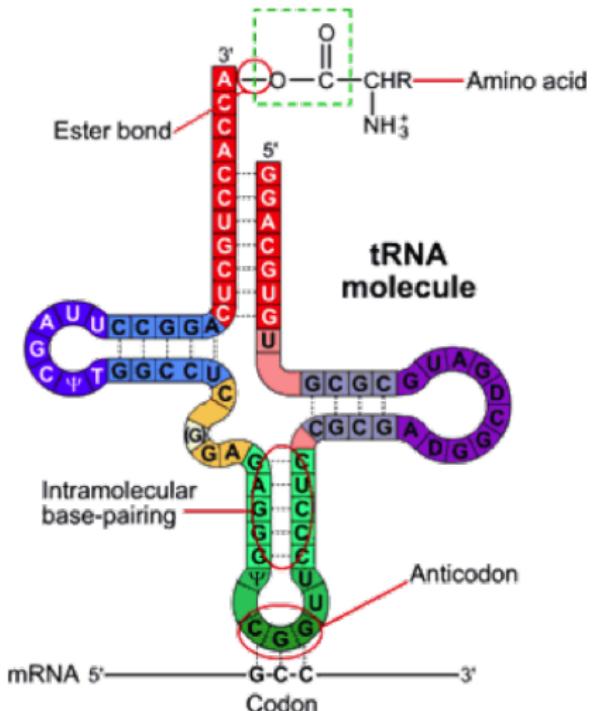




RNA features

RNA/ncRNA: characterised by complex secondary structure

- tRNA - transfer RNA
- rRNA - ribosomal RNA
- CRISPRs - prokaryotic defence, and genome editing
- many other functional classes, including enhancers

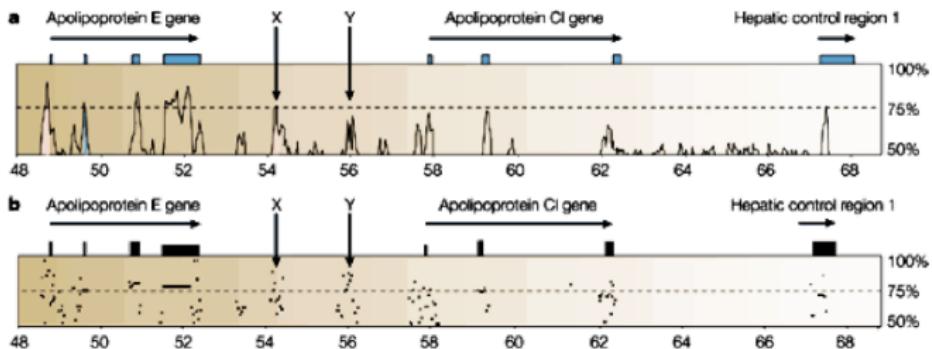




Regulatory features ^a

^a Pennacchio & Rubin (2001) *Nature Rev. Genet.* doi:10.1038/35052548

- transcription start sites (TSS)
- RNA polymerase (RNAP) binding sites
- transcription factor binding sites (TFBS)
- core, proximal and distal promoter regions

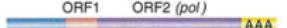
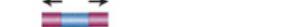




Repeat/mobile elements

- tandem repeats (VNTR, etc.)
- (retro-) transposable elements (over 50k *Alu* copies in human genome!)
- phage inclusion (prokaryotes)

Types of transposable elements in the human genome

Element	Transposition	Structure	Length	Copy number	Fraction of genome
LINEs	Autonomous	ORF1 ORF2 (<i>pol</i>) 	1–5 kb	20,000–40,000	21%
SINEs	Nonautonomous		100–300 bp	1,500,000	13%
DNA transposons	Autonomous		2–3 kb	300,000	3%
	Nonautonomous		80–3000 bp		



Gene finding ^{a b c}

^a Liang et al. (2009) *Genome Res.* doi:10.1101/gr.088997.108

^b Brent (2007) *Nat. Biotech.* doi:10.1038/nbt0807-883

^c Korf (2004) *BMC Bioinf.* doi:10.1186/1471-2105-5-59

At genome scales, we need to automate functional prediction

Empirical (evidence-based) methods:

- Inference from known protein/cDNA/mRNA/EST sequence
- Interference from mapped RNA reads (e.g. RNAseq)

Ab initio methods:

- Prediction on the basis of gene features (TSS, CpG islands, Shine-Dalgarno sequence, stop codons, nucleotide composition, etc.)

Inference from genome comparisons/sequence conservation



Regulatory element finding ^{a b c}



The James
Hutton
Institute

^aZhang *et al.* (2011) *BMC Bioinf.* doi:10.1186/1471-2105-12-238

^bKilic *et al.* (2013) *Nucl. Acids Re.* doi:10.1093/nar/gkt1123

^cVavouris & Elgar (2005) *Curr. Op. Genet. Deve.* doi:10.1016/j.gde.2005.05.002

Empirical (evidence-based) methods:

- Inference from protein-DNA binding experiments
- Interference from co-expression

Ab initio methods:

- Identification of regulatory motifs (profile/other methods; TATA, σ -factor binding sites, etc.)
- Statistical overrepresentation of motifs
- Identification from sequence properties

Inference from genome comparisons/sequence conservation



Multiple genome alignment

EXERCISE 7:

`predict_CDS/bacterial_CDS_prediction.md`



Genecalling software

Many options for this, including... Prokaryotes

- **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/>
- **Prokka:** <http://www.vicbioinformatics.com/software.prokka.shtml>

Eukaryotes

- **GlimmerHMM:** <http://ccb.jhu.edu/software/glimmerhmm/>
- **GeneMarkES:** <http://opal.biology.gatech.edu/gmseuk.html>
- **Augustus:** <http://augustus.gobics.de/>
- **SNAP:** <http://korflab.ucdavis.edu/software.html>



Feature identification

All prediction methods give you errors

- **False positive:** predicts features where there are none
- **False negative:** fails to predict a feature that is present
- **Magnitude:** does not identify correct bounds on/value for feature
- **Category:** predicts a feature to belong to the wrong class

All experiments have errors

Genome comparisons can help correct for these errors



What makes genome features equivalent?



When we compare two features (e.g. genes) between two or more genomes, there must be some basis for making the comparison. That is, they have to be *equivalent* in some way, such as:

- common evolutionary origin
- functional similarity
- a family-based relationship

It's common to define equivalence of genome features in terms of evolutionary relationship.



Why look at equivalent features?



The real power of genomics is comparative genomics!

- Makes catalogues of genome components comparable between organisms
- Differences, e.g. presence/absence of equivalents may support hypotheses for functional or phenotypic difference
- Can identify characteristic signals for diagnosis/epidemiology
- Can build parts lists and wiring diagrams for systems and synthetic biology



Table of Contents

Comparisons of genome features

Genome Features

Who Let the -logues Out?

What's so important about orthologues?

Evaluating orthologue prediction

Using orthologue predictions

Selection Pressure

Finishing the Hat



Who let the -logues out?

Genome features can have complex evolutionary relationships

We need precise terms to describe these relationships





The -logues drop ^a

^aFitch et al. (1970) *Syst. Zool.* doi:10.2307/2412448



How do we understand the relationships between features in more than one genome?

- Functional similarity: **analogy**
- Evolutionary common origin: **homology, orthology, etc.**
- Evolutionary/functional/family relationship: **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



Definitions

Technical terms describing evolutionary relationships

- **Homologues**: share a common ancestor **NOTE: there are NOT degrees of homology**
- **Analogues**: are functionally similar. Analogues may or may not share common ancestry
- **Orthologues**: are homologues that *diverged through speciation*
- **Paralogues**: are homologues that *diverged through duplication* within the same genome

(also *co-orthologues*, *xenologues*, etc.)



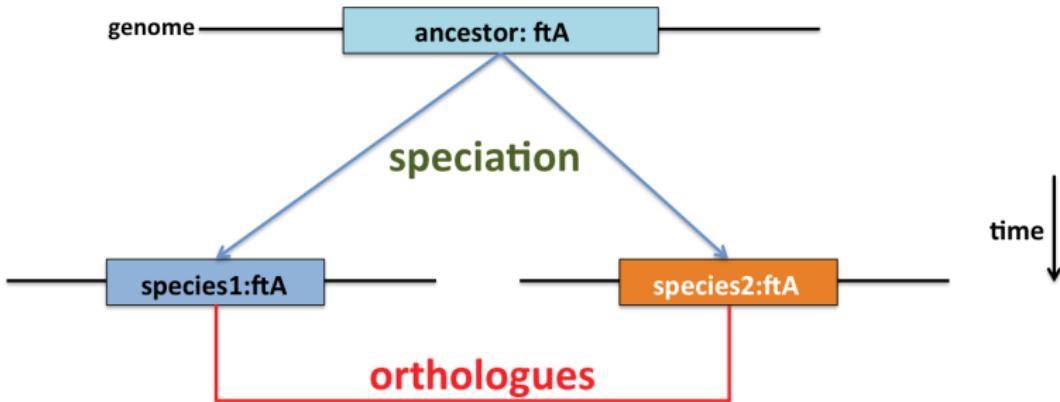
Who let the -logues out?



time
↓



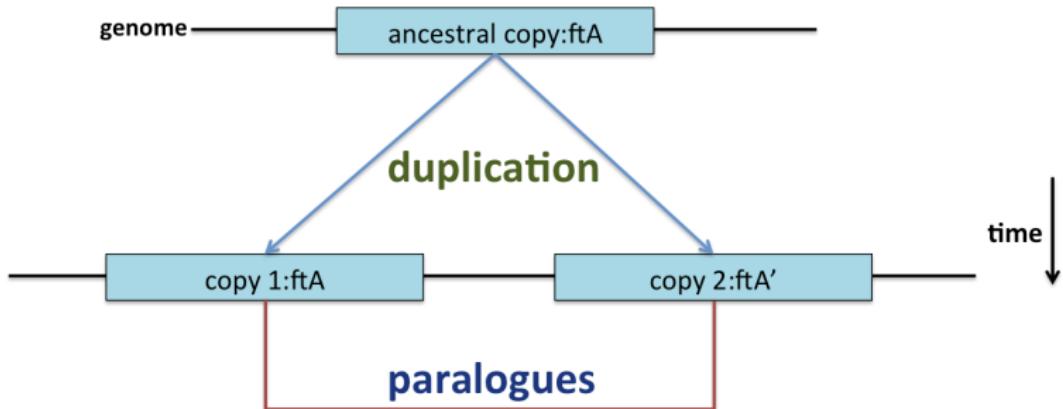
Who let the -logues out?



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



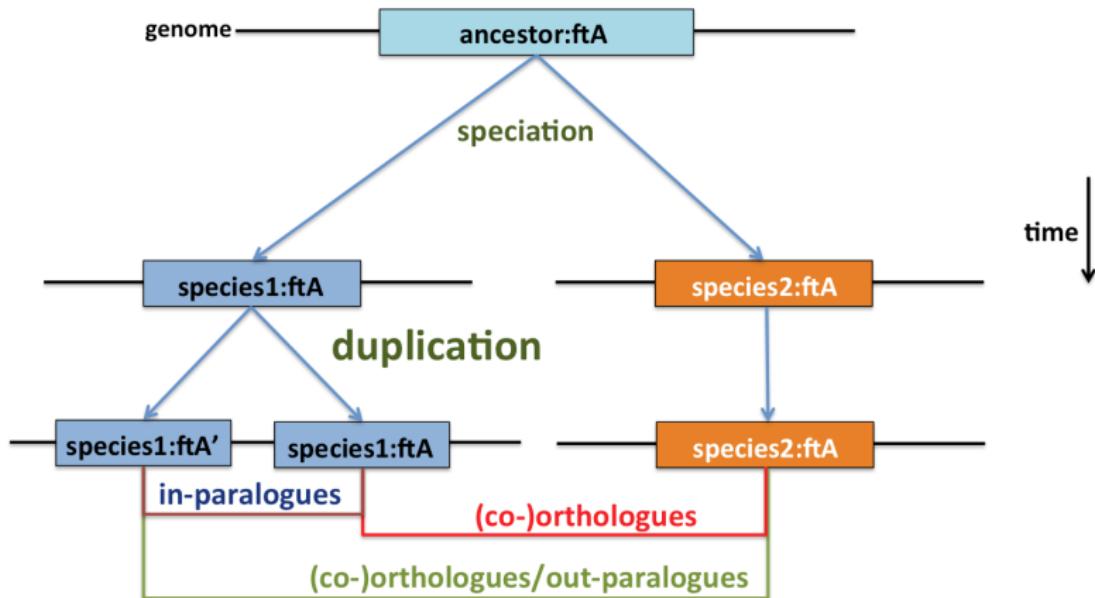
Who let the -logues out?



Paralogues: homologues that diverged through duplication within the same genome

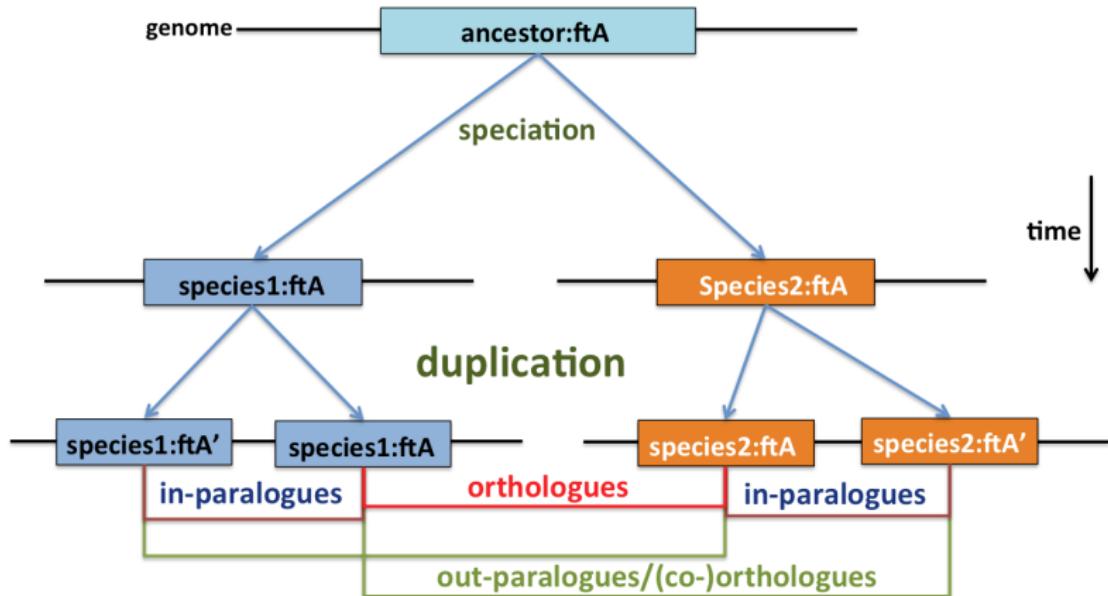


Who let the -logues out?





Who let the -logues out?



But it's a little more complicated than that.

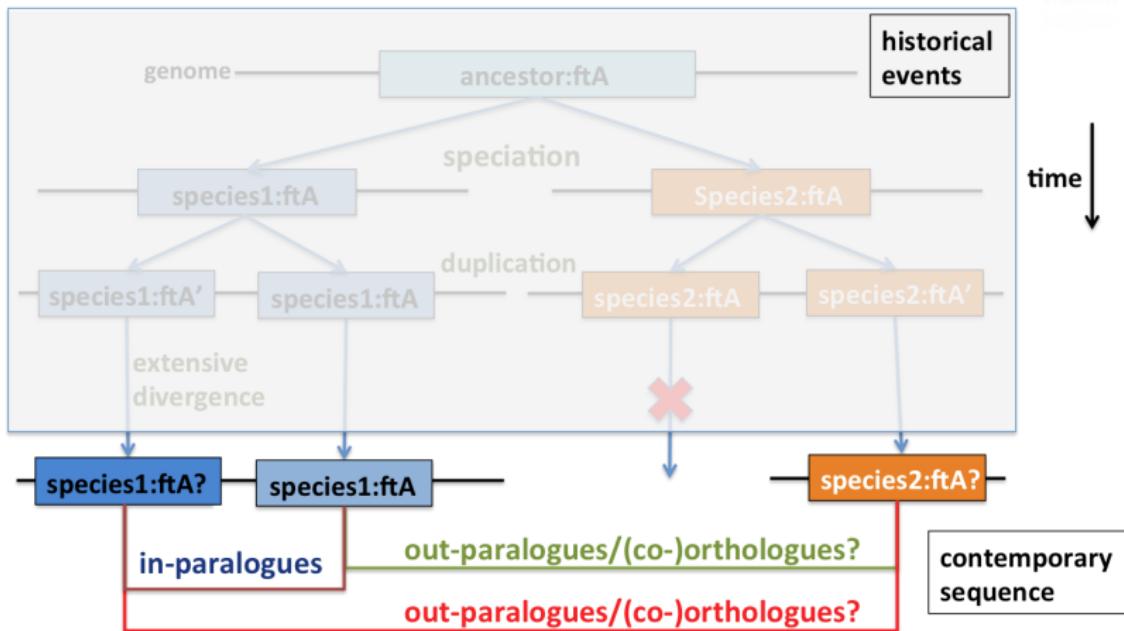
Biology is not well-behaved.

- Gene loss
- Homologues may diverge so widely that they can be hard to recognise
- Reconstructed evolutionary trees may not be robust inferences of speciation (or relevant to it, in prokaryotes)
- There is no record of history - we can only make inferences

All classifications of orthology/paralogy are inferences!



^a Kristensen et al. (2011) *Brief. Bioinf.* 12:379-391 doi:10.1093/bib/bbr030



All classifications of orthology/paralogy are inferences!

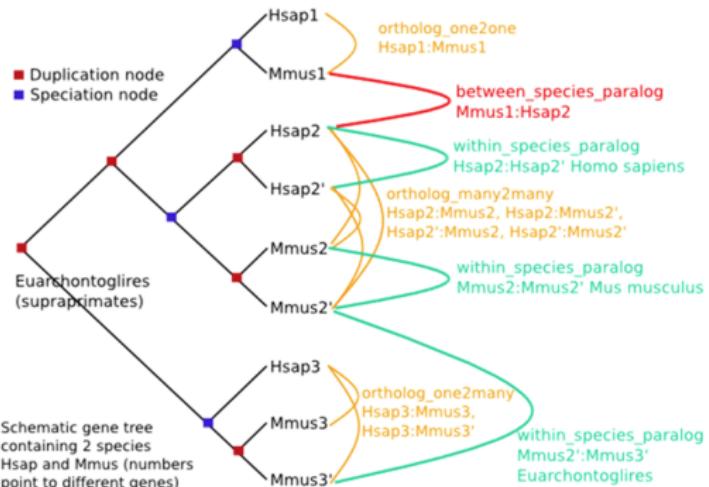


Ensembl Compara ^a

^aVilella et al. (2009) *Genome Res.* **19**:327-335 doi:10.1101/gr.073585.107

Some tools/databases, e.g. Ensembl Compara, use slightly different definitions (almost everything's an “orthologue”)

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





Orthology ^a

^aStorm & Sonnhammer (2002) *Bioinformatics* doi:10.1093/bioinformatics/18.1.92

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

All classifications of orthology/paralogy are inferences!



Table of Contents

Comparisons of genome features

Genome Features

Who Let the -logues Out?

What's so important about orthologues?

Evaluating orthologue prediction

Using orthologue predictions

Selection Pressure

Finishing the Hat



The Ortholog Conjecture ^a ^b

^a Nehrt et al. (2011) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002073

^b Chen et al. (2012) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002784



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



The Ortholog Conjecture ^a

^aNehrt et al. (2011) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002073

- Paralogues are better predictors of function than orthologues
 \therefore the conjecture is false
- Cellular context is better for protein function inference

(function defined in Gene Ontology (GO) terms)

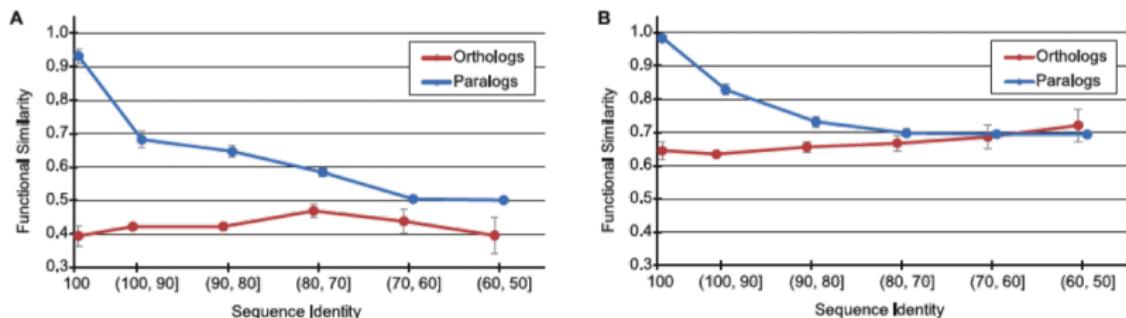


Figure 1. The relationship between functional similarity and sequence identity for human-mouse orthologs (red) and all paralogs (blue). Standard error bars are shown. (A) Biological Process ontology, (B) Molecular Function ontology.
doi:10.1371/journal.pcbi.1002073.g001



The Ortholog Conjecture ^a

^aChen et al. (2012) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002784



We do not understand function well enough to test the conjecture

- “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?)



Why focus on orthologues? ^{a b c}

^aChen and Zhang (2012) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002784

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

^cAltenhoff and Dessimoz (2009) *PLoS Comp. Biol.* 5:e1000262 doi:10.1371/journal.pcbi.1000262

Formalisation of the idea of *corresponding genes* in different organisms.

Orthologues serve two purposes:

- **Evolutionary equivalence**
- **Functional equivalence** ("The Ortholog Conjecture")

Applications in comparative genomics, functional genomics and phylogenetics.

Over 30 databases attempt to describe orthologous relationships (http://questfororthologs.org/orthology_databases)



Finding “Orthologues”

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



Finding orthologues ^{a b c}

^a Kristensen *et al.* (2011) *Brief. Bioinf.* **12**:379-391 doi:10.1093/bib/bbr030

^b Trachana *et al.* (2011) *Bioessays* **33**:769-780 doi:10.1002/bies.201100062

^c Salichos and Rokas (2011) *PLoS One* **6**:e18755 doi:10.1371/journal.pone.0018755.g006

Multiple methods and databases

- **Pairwise genome**

- RBBH (aka BBH, RBH),
RSD, InParanoid, RoundUp

- **Multi-genome**

- *Graph-based*: COG, eggNOG,
OrthoDB, OrthoMCL, OMA,
MultiParanoid
- *Tree-based*: TreeFam,
Ensembl Compara,
PhylomeDB, LOFT

List of orthology databases

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

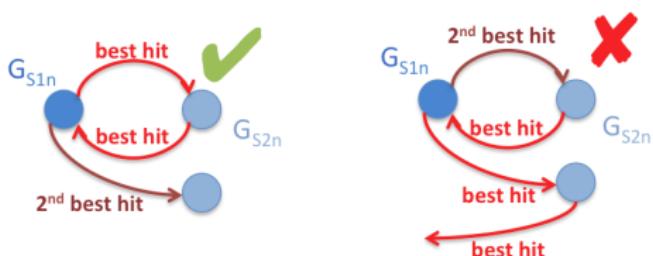
1. COG/WEGeKOGs
2. COGe-COCO-CL
3. COGe-LOFT
4. eggNOG
5. EGO
6. Ensembl Compara
7. Gene-Oriented Ortholog Database
8. GreenPhyDB
9. HCDP
10. HemisVogene
11. HOGENOM
12. HOVERGEN
13. HOMOLENS
14. HOPS
15. INVHOGEN
16. JnParanoid
17. KEGG Orthology
18. MetaphOrn
19. MiGD
20. MiGD
21. OMA
22. OrthoDB (OrthoDB on Wikipedia)
23. OrthoLogID
24. ORTHOLUGE
25. OrthoMepster
26. OrthoMCL
27. Panther
28. PhOG
29. PHOG
30. PhylomeDB
31. PLAZA
32. P-POD
33. ProMMap
34. Proteinortho
35. RoundUp
36. TreeFam
37. YOBY



Reciprocal Best BLAST Hits ^a

^aOn Reciprocal Best BLAST Hits 19/7/2012

- S_1, S_2 are the gene sequence sets from two organisms
- Use sequence search tool (BLAST/FASTA):
 - Query= S_1 , Subject= S_2
 - Query= S_2 , Subject= S_1



- Optionally filter hits (e.g. on %identity and %coverage)
- Find all pairs of sequences $\{G_{S1n}, G_{S2n}\}$ in S_1, S_2 where G_{S1n} is the best BLAST match to G_{S2n} and G_{S2n} is the best BLAST match to G_{S1n} .



Finding “Orthologues”: RBBH

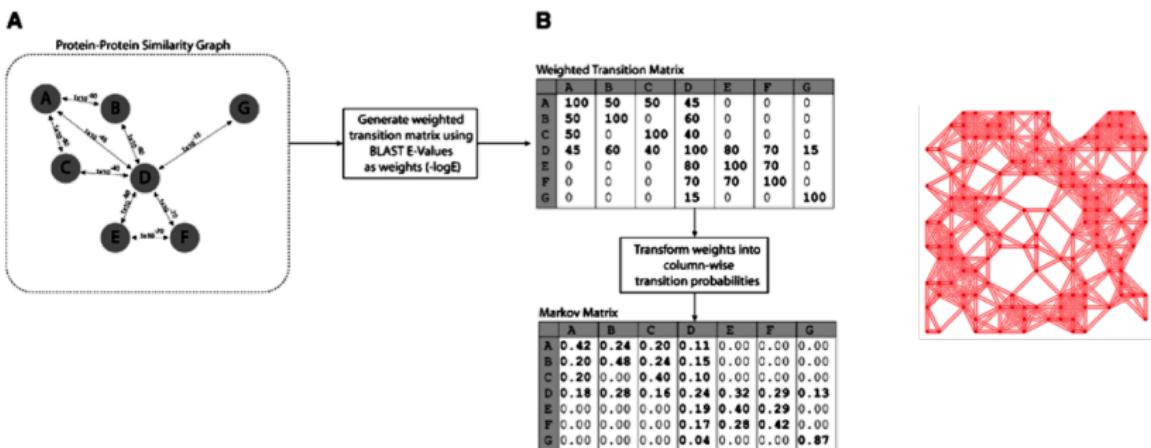


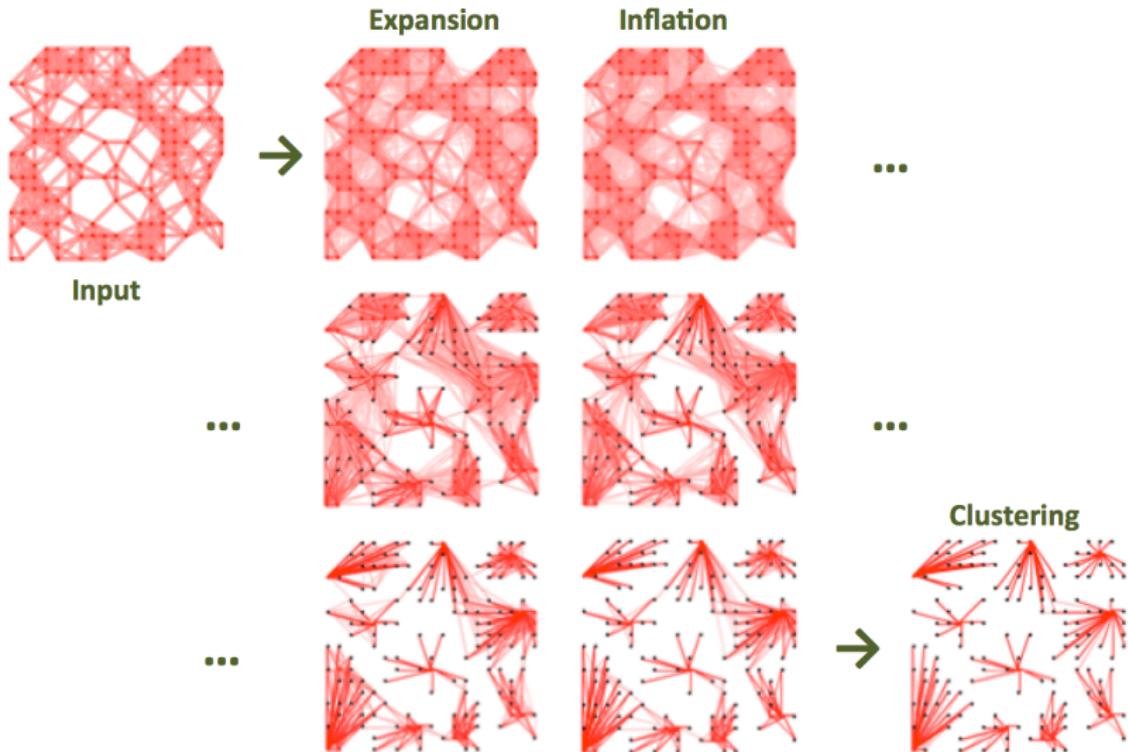
EXERCISE 8:

`ex08_find_rbbh.ipynb`



- MCL constructs a network (*graph*) from all-against-all BLAST results
- Matrix operations (*expansion, inflation*) are applied
- Expansion, inflation iterated until the network converges





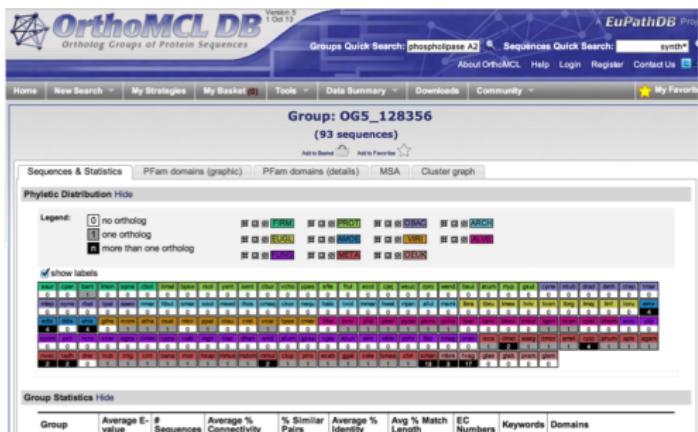


OrthoMCL ^{a b}

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

^b<http://orthomcl.org/orthomcl/>

- Defines potential inparologue, orthologue and co-orthologue pairs, **using RBBH**
- Applies MCL to cluster these pairs of sequences
- Output clusters include both orthologues and paralogues





Finding “Orthologues”: MCL

EXERCISE 9:

`mcl_orthologues/ex_09a_mcl_orthologues.md`

`mcl_orthologues/ex09b_mcl_orthologues.ipynb`



Notes of caution

BLAST-based orthology methods are fast!

. . . but there are some drawbacks:

- No guarantee that sequence matches are transitive
(A may match B at a different domain than B matches C)
- No evolutionary distance model
- Do not account for multiple domain matches

Orthologues are proposed on the basis of sequence similarity, **not predicted directly**.



Orthologue prediction methods

Homologene (NCBI): *synteny*

- <http://www.ncbi.nlm.nih.gov/homologene>

Mouse Genome Database (MGD): *manual curation*

- <http://www.informatics.jax.org/homology.shtml>

EnsembleCompara (EMBL-EBI): *tree-based*

- <http://www.ensembl.org/info/genome/compara/index.html>

TreeFam (EMBL-EBI): *tree-based*

- <http://www.treefam.org>

OrthologID: *tree-based*

- <http://nypg.bio.nyu.edu/orthologid>



Table of Contents

Comparisons of genome features

Genome Features

Who Let the -logues Out?

What's so important about orthologues?

Evaluating orthologue prediction

Using orthologue predictions

Selection Pressure

Finishing the Hat

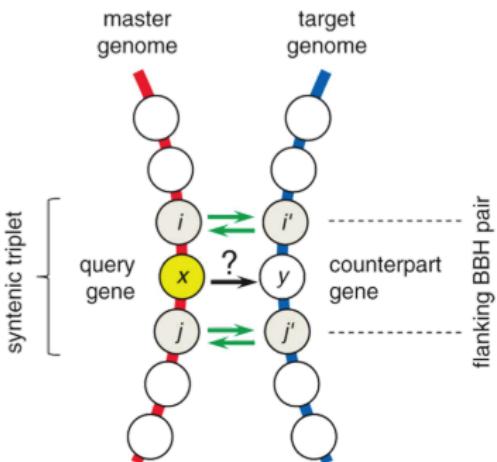


Which prediction methods work best? ^a

^aWolf and Koonin (2012) *Genome Biol. Evol.* 4:1286-1294 doi:10.1093/gbe/evs100

Taking advantage of prokaryotic operon structure: **if the outer pair of a syntenic triplet of genes are orthologous, the middle gene is also likely to be orthologous.**

Specifically testing reciprocal best hits (RBH).



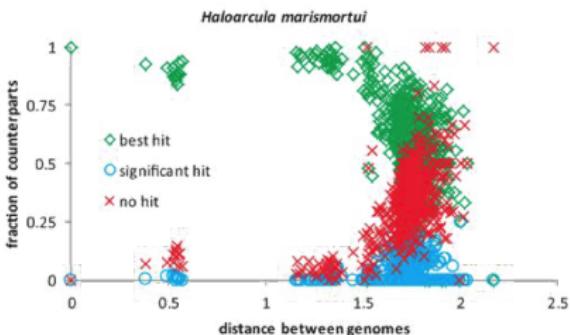
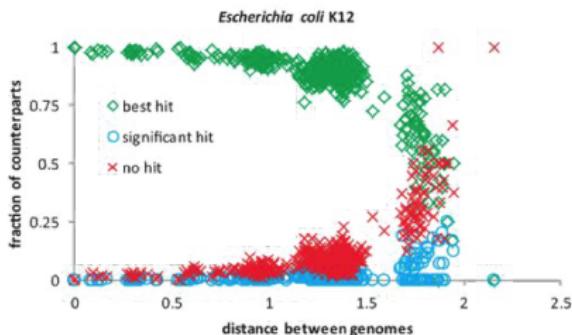


Which prediction methods work best? ^a

^aWolf and Koonin (2012) *Genome Biol. Evol.* 4:1286-1294 doi:10.1093/gbe/evs100

- Tested on 573 prokaryotic genomes
- 88-99% of RBH found in syntenic triplets
- Overwhelming majority of middle genes are RBH

RBH reliably finds “orthologues”.





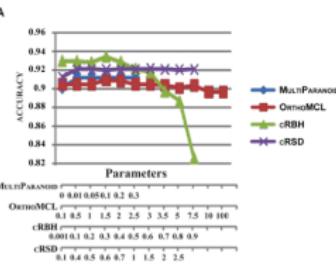
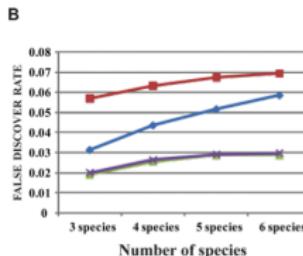
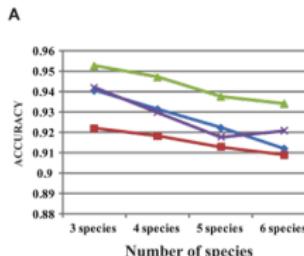
Which prediction methods work best? ^a

^aSalichos and Rokas (2011) *PLoS One* 6:e18755 doi:10.1371/journal.pone.0018755.g006

Four methods tested against 2,723 curated orthologues from six *Saccharomycetes*

- RBBH (and cRBH); RSD (and cRSD); MultiParanoid; OrthoMCL
- Rated by statistical performance metrics: sensitivity, specificity, accuracy, FDR

cRBH most accurate and specific, with lowest FDR.





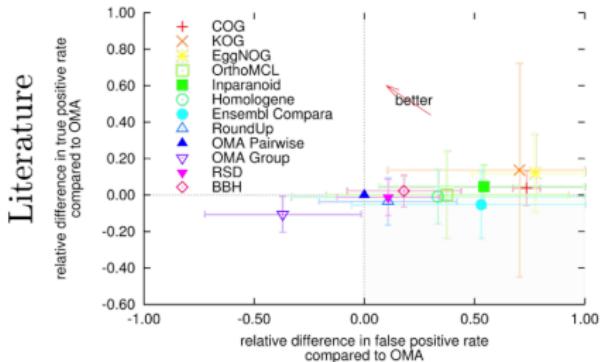
Which prediction methods work best? ^a

^a Altenhoff and Dessimoz (2009) *PLoS Comp. Biol.* 5:e1000262 doi:10.1371/journal.pcbi.1000262

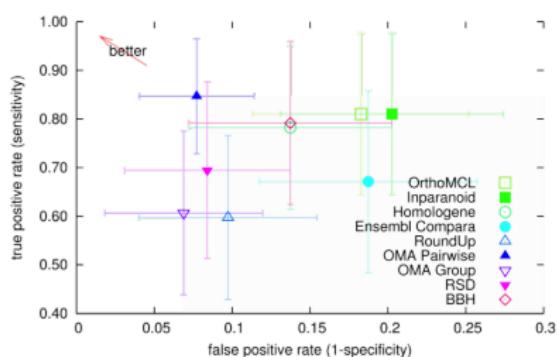
Testing on literature-based benchmarks for grouping by function and correct branching of phylogeny.

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.

A Pairwise project comparison



B Comparison on intersection set





Which prediction methods work best?

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



Table of Contents

Comparisons of genome features

Genome Features

Who Let the -logues Out?

What's so important about orthologues?

Evaluating orthologue prediction

Using orthologue predictions

Selection Pressure

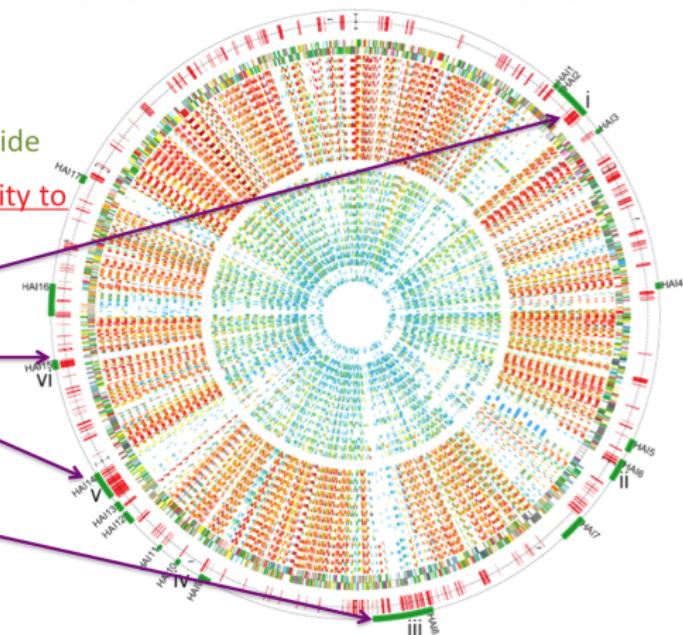
Finishing the Hat



Functional adaptation in Pba^a

^aToth et al. (2006) Ann. Rev. Phytopath. 44:305-336 doi:10.1146/annurev.phyto.44.070505.143444

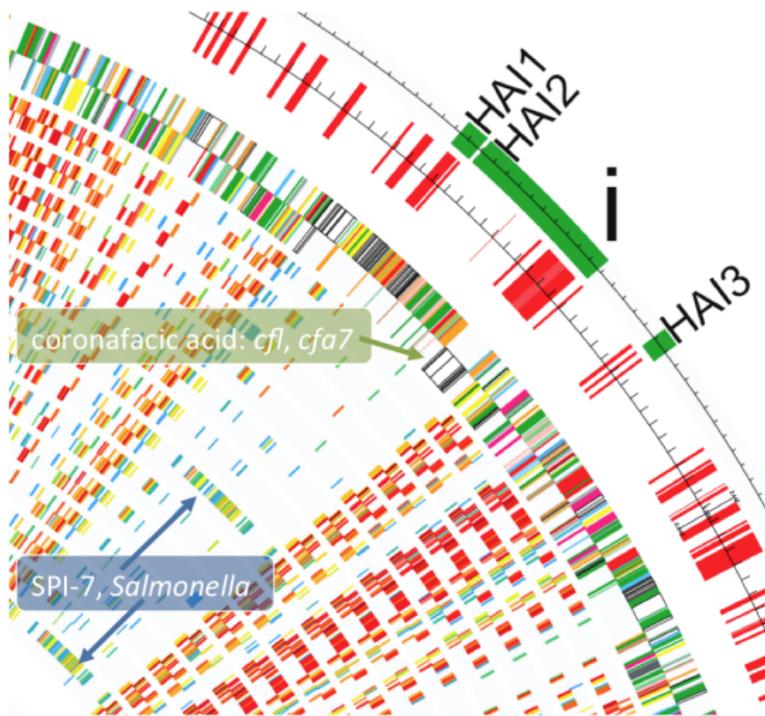
- Comparison against plant- (13) and animal-associated (14) bacteria
- Plant-associated in centre
- Animal-associated on outside
- Red marks: greater similarity to plant-associated bacteria
- HAI2: Phytotoxin
- HAI15: Adherence
- HAI14: Nitrogen fixation
- HAI8: T3SS





Functional adaptation in Pba^a

^a Toth et al. (2006) Ann. Rev. Phytopath. 44:305-336 doi:10.1146/annurev.phyto.44.070505.143444



Coronatine (*P. syringae*) interferes with jasmonate responses in host, as a jasmonate mimic

Coronafacic acid –
Pseudomonas syringae phytotoxin precursor
(coronatine)
- payload

SPI-7 -
Salmonella Typhi
Pathogenicity island
- delivery system



Table of Contents

Comparisons of genome features

Genome Features

Who Let the -logues Out?

What's so important about orthologues?

Evaluating orthologue prediction

Using orthologue predictions

Selection Pressure

Finishing the Hat



Selection Pressure

Comparative genomics helps identify selection pressures at a structural and sequence level



How orthologues help

Defining core groups of genes as “orthologue” allows analysis of groups of genes together by:

- synteny/collocation
- gene neighbourhood changes (e.g. *genome expansion*)
- pan genome (core/accessory genomes)

and of individual genes within those groups, by:

- multiple alignment
- domain detection
- identification of functional sites
- inference of directional selection (stabilising/positive selection)



Synteny ^a

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



Selection pressures depend on gene (product) function

- Genes involving physically or functionally-interacting proteins tend to involve under selective constraints
In bacteria, this leads to coexpression in *regions* and collocation in *operons*
- Collocation (and correlation) may be identified by genome comparisons
- Also true for regulatory and metabolic networks



Synteny ^a ^b

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

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

Several tools for synteny detection, e.g.

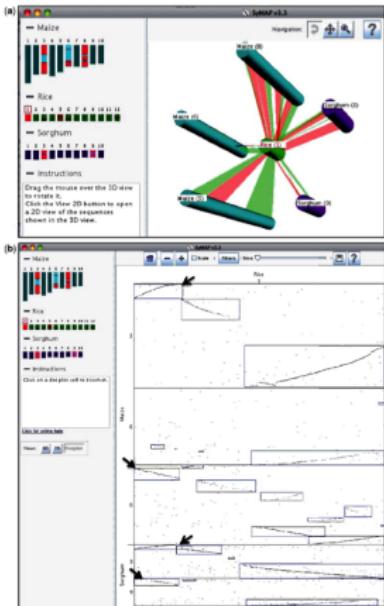
- SyMAP

<http://www.agcol.arizona.edu/software/symp/>

- i-ADHoRe

[http://bioinformatics.psb.ugent.be/software/details/i-
ADHoRe](http://bioinformatics.psb.ugent.be/software/details/i-ADHoRe)

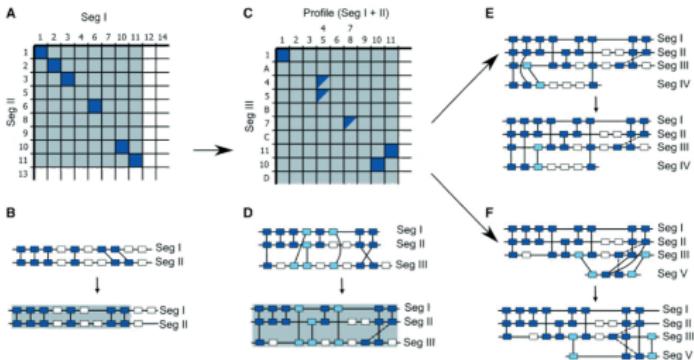
- MCScan, Cyntenator, etc.





Algorithm: starts from defined equivalent genes to produce genome-scale multiple alignments of blocks of genes

1. Combine tandem repeats of gene sets
2. Make *gene homology matrix* (GHM)
3. Convert collinear GHMs to *profiles*
4. Align *profiles* (GG2 algorithm)
5. Search next genome with profiles, and iterate until complete





Synteny: i-ADHoRe

EXERCISE 10:

[i-ADHoRe/ex10a_i-ADHoRe.md](#)

[i-ADHoRe/ex10b_i-ADHoRe.ipynb](#)



Genome expansion ^a

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

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

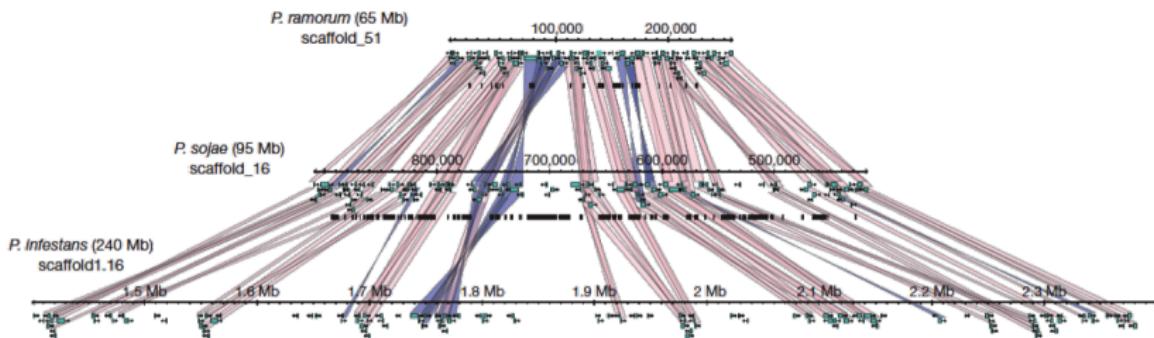


Figure 1 | Repeat-driven genome expansion in *Phytophthora infestans*. Conserved gene order across three homologous *Phytophthora* scaffolds. Genome expansion is evident in regions of conserved gene order, a

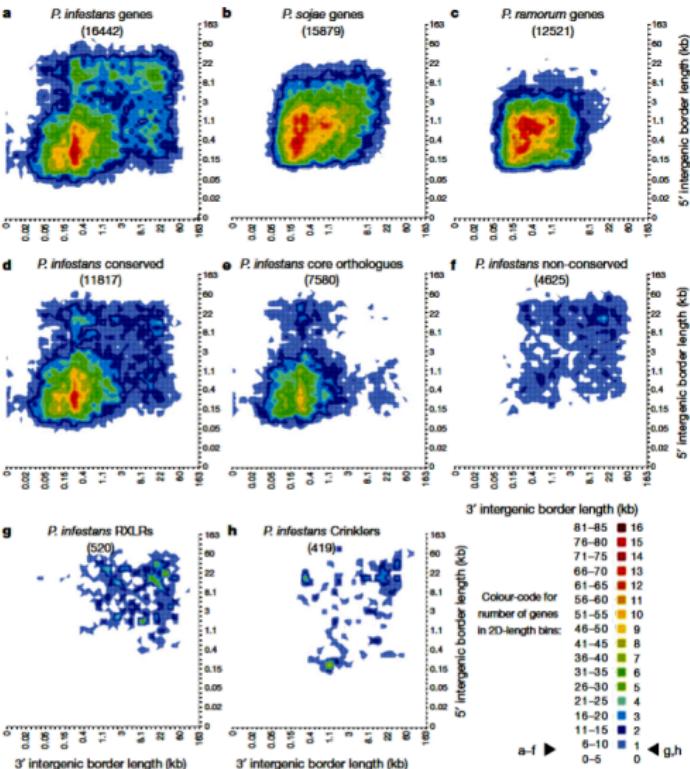
consequence of repeat expansion in intergenic regions. Genes are shown as turquoise boxes, repeats as black boxes. Collinear orthologous gene pairs are connected by pink (direct) or blue (inverted) bands.



Genome expansion ^a

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

- Mobile elements (MEs) are large, and duplicate/carry genes with them
- Larger intergenic regions in MEs
- Effector proteins found preferentially in regions with large gaps
- Two-speed genome associated with adaptability





Two-speed genome visualisation

EXERCISE 11:

`ex11_pi_two_speed.ipynb`



The Pangenome

The Core Genome Hypothesis:
“The core genome is the primary cohesive unit defining a bacterial species”



Core genome ^{a b}

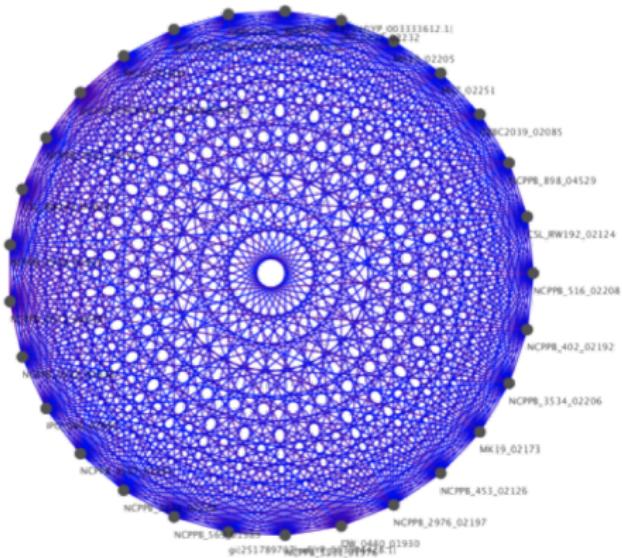
^aLaing (2010) *BMC Bioinf.* doi:10.1186/1471-2105-11-461

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



The James
Hutton
Institute

Once equivalent genes have been identified, those present in all related isolates can be identified: **the core genome**.



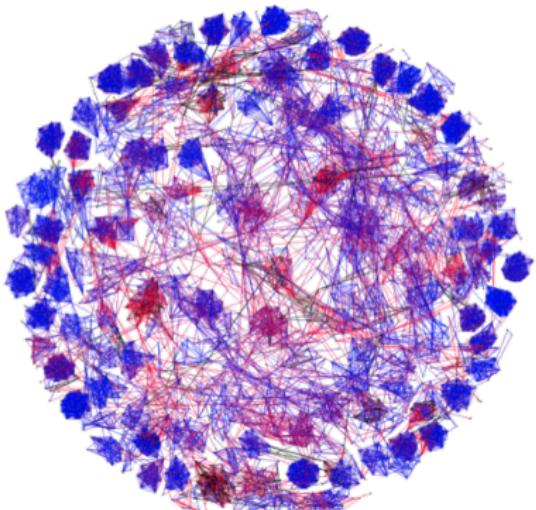


Accessory genome ^a ^b

^aLaing (2010) *BMC Bioinf.* doi:10.1186/1471-2105-11-461

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

The remaining genes are **the accessory genome**, and are expected to mediate function that distinguishes between isolates.

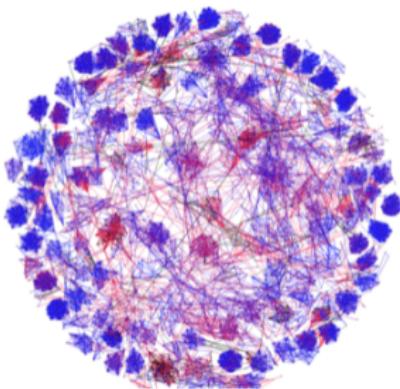




Accessory clusters

Accessory RBH clusters can be pruned, to identify the accessory genome specific to subgroups of isolates:

Species	Weak Pruning	Full Pruning
Core Genome	2201	2201
<i>D. chrysanthemi</i>	32	36
<i>D. dadantii</i>	11	14
<i>D. dianthicola</i>	102	127
<i>D. paradisiaca</i>	404	441
<i>D. solani</i>	120	157
<i>D. zeae</i>	33	40



- Accessory: RBBH with all other members of same species, but no other *Dickeya*
- Weak pruning: remove all RBBH <80% identity, <40% coverage
- Full pruning: trim graph (by Mahalanobis distance) until minimal cliques found

These genes may be responsible for subgroup-specific phenotypes

Accessory genome ^{a b}

^a Croll and McDonald (2012) *PLoS Path.* 8:e1002608 doi:10.1371/journal.ppat.1002608

^b Baltrus et al. (2011) *PLoS Path.* 7:e1002132 doi:10.1371/journal.ppat.1002132.t002

Accessory genomes are a cradle for adaptive evolution
This is particularly so for bacterial pathogens, such as
Pseudomonas spp.

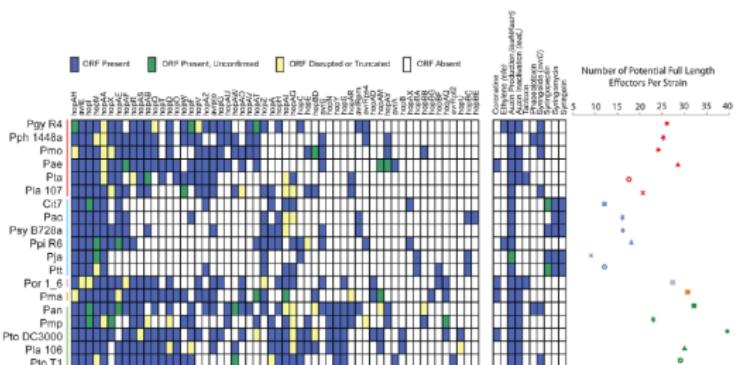


Figure 3. *P. syringae* isolates harbor extensive diversity in virulence gene repertoires. TTE, toxin, and plant hormone biosynthesis genes are listed across the top. *P. syringae* genomes, color-coded by phylogenetic group as in Figure 1. At the left, a blue box indicates presence of full-length ORFs or complete pathways within each genome. Green boxes indicate that genes or pathways are present by similarity searches, but the presence of full-length genes could not be verified by PCR, or the pathways are potentially incomplete. Yellow boxes indicate that genes are either significantly truncated or are disrupted by insertion sequence elements. White boxes indicate absence of genes or pathways from the strains based on homology searches. At the far right, the total number of potentially functional TTE proteins is shown for each genome and displayed according to the color-coded strain and group symbols shown in Figure 1.
doi:10.1371/journal.ppat.1002132.g003



Core genome synteny ^a

^aProost et al. (2012) *Nuc. Acids Res.* **40**:e11 doi:10.1093/nar/gkr955

Using tools like i-ADHoRe that identify synteny and collinearity, the structural organisation of the core genome can be determined:



For *Dickeya*, the core genome appears to be structurally well-conserved across all isolates.



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

^b<https://lfz.corefacility.ca/panseq/>

Panseq is a tool for identification of core and accessory genomes

Pan Seq → pan~genomic sequence analysis

Home Analyses Contact FAQ

Welcome

Panseq is an easy-to-use, web-based group of tools
for pan-genomic analyses.

Novel Region Finder

Discover genomic regions unique to a sequence, or group of sequences.

Pan-genome Analyses

Identify the pan-genome among your sequences. Find SNPs in the core genome
and determine the distribution of accessory genomic regions.

Loci Selector

Identify loci to offer the best discrimination among your dataset.

Panseq is open source. Get the standalone version from:

<https://github.com/chadlaing/Panseq>



^aTreangen et al. (2014) *Genome Biol.* **15**:524 doi:10.1186/s13059-014-0524-x

Visualising and organising comparison/pangenome data across thousands of bacteria is difficult.

Harvest suite of tools, for alignment and visualisation of thousands of genomes:

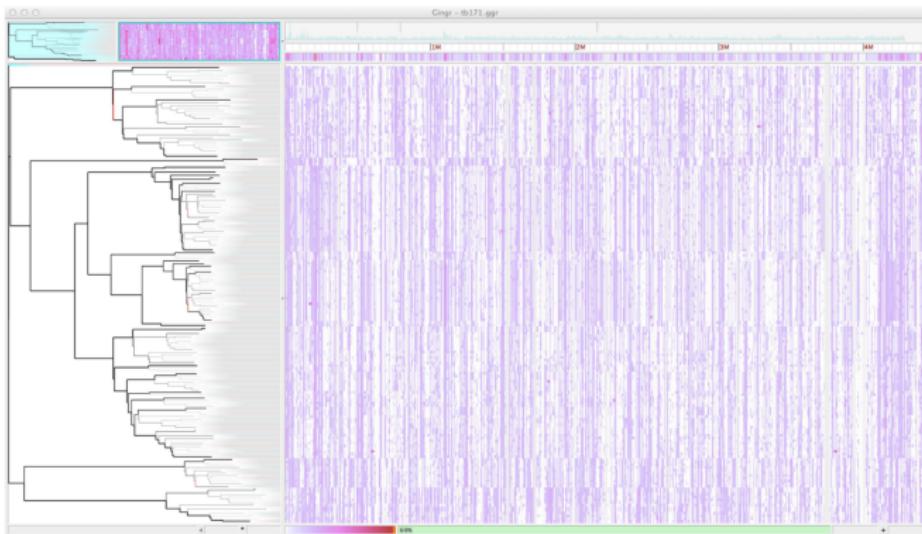




Table of Contents

Comparisons of genome features

Genome Features

Who Let the -logues Out?

What's so important about orthologues?

Evaluating orthologue prediction

Using orthologue predictions

Selection Pressure

Finishing the Hat



What didn't I get to?

- 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
 - AMPHORA <https://phylogenomics.wordpress.com/software/amphora/>



Messages to take away

- 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



Messages to take away

- 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



Messages to take away

- Comparative genomics is comparisons
 - What is *similar* between two genomes?
 - What is *different* between two genomes?
- Comparative genomics *is* evolutionary genomics
 - Lots of scope for improvement in tools
- Tools that 'do the same thing' can give different output
 - BLAST vs MUMmer
 - RBBH vs MCL
 - The choice of application matters for correctness and interpretation



Messages to take away

Comparative genomics is

- Fun
- Indoor work, in the warm and dry
- Not a job that involves a lot of heavy lifting



Licence: CC-BY-SA

By: Leighton Pritchard

This presentation is licensed under the Creative Commons Attribution ShareAlike license

<https://creativecommons.org/licenses/by-sa/4.0/>