

Reconstructing ancestral *Anopheles* mosquitoes gene orders in a model-free approach

Cedric Chauve

Department of Mathematics
Simon Fraser University

Overview



- ❑ So far we have mostly discussed
 - ❑ Several issues in preprocessing data for large-scale comparative genomics
 - ❑ Applications of preprocessing methods (gene families clustering, gene trees reconciliation) to the *Anopheles* genomes dataset
 - ❑ Genome rearrangement distance-based methods for reconstructing ancestral gene orders, with a focus on methods accounting for duplicated genes.

- ❑ In the coming set of lectures, we will see how genome rearrangement methods were applied to the *Anopheles* genomes dataset:
 - ❑ In this lecture we will see a model-free method (i.e. not based on a notion of distance) that was used in a first analysis of the data, while considering only single-copy genes.
 - ❑ In the next lecture we will use the DeCoSTAR method to jointly reconstruct ancestral gene orders and extant scaffolds.

Raw input: the *Anopheles* genomes assemblies



Data.

- ❑ HiSeq2000 libraries from a single female for each species:
 - ❑ 101bp, 180bp insert, 30-170 Depth of Coverage
 - ❑ 101bp, 1.5kb jump, 50-145 DoC
- ❑ Fosill libraries (Williams *et al*, Genome Res. (2012)) from hundred females, 101bp, 35kb jump, 3-11 DoC, 11 species.

Software/strategy.

- ❑ ALLPATHS-LG (Gnerre *et al*, PNAS (2011)) + Pilon (Walker *et al*, PLoS One (2014))
- ❑ *An. gambiae* used as a reference for the Gambia complex

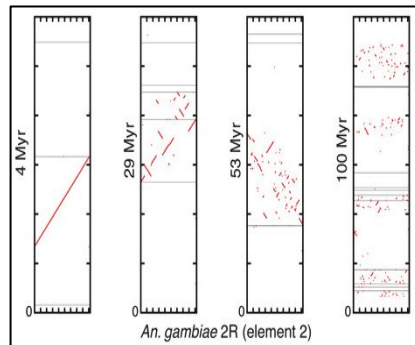
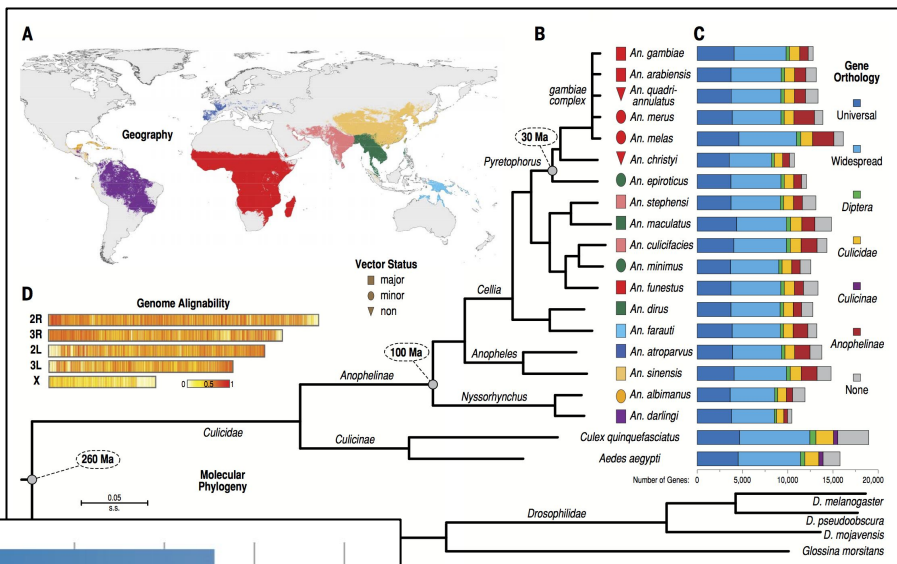
Species	<i>An. alb.</i>	<i>An. atr.</i>	<i>An. chr.</i>	<i>An. col.</i>	<i>An. dar.</i>	<i>An. dir.</i>	<i>An. far.</i>	<i>An. fun.</i>	<i>An. mac.</i>	<i>An. mel.</i>	<i>An. mer.</i>	<i>An. min.</i>	<i>An. qua.</i>
Size (Mb)	170	224	172	224	134	216	181	225	141	227	251	201	283
Gaps	6.9Kb	35Mb	2.6Mb	15Mb	27Kb	18Mb	5Mb	35Mb	10Mb	20Mb	33Mb	15Mb	75Mb
Scf. N50	18Mb	9.2Mb	9Kb	4.4Mb	115Kb	7Mb	1,2Mb	672Kb	4Kb	18Kb	342Kb	1.5Mb	1.6Mb
Scaffolds	204	1,371	30K	10K	2,160	1,266	550	1,392	47K	20K	2,753	678	2,823

The *Anopheles* Science paper (2015)



Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes

Daniel E. Neafsey,^{1,†} Robert M. Waterhouse,^{2,3,4,5,6} Mohammad R. Abai,⁶ Sergey S. Aganezov,⁷ Max A. Alekseyev,⁷ James E. Allen,⁸ James Amon,⁹ Bruno Arcà,¹⁰ Peter Arensburger,¹¹ Gleb Artemov,¹² Lauren A. Assour,¹³ Hamidreza Basseri,⁴ Aaron Berlin,⁴ Bruce W. Birren,⁴ Stephanie A. Blandin,^{14,15} Andrew I. Brockman,¹⁶ Thomas R. Burkot,¹⁷ Austin Burt,¹⁸ Clara S. Chan,^{19,20} Cedric Chanve,²⁰ Joanna C. Chiu,²⁰ Mikkel Christensen,⁴ Carlo Costantini,²¹ Victoria L. M. Davidson,²² Elena Deligdisli,²³ Tania Dottorini,¹⁶ Vicky Dritsou,²⁴ Stacey B. Gabriel,²⁵ Wamdaogo M. Guelbeogo,²⁶ Andrew B. Hall,²⁷ Mira Y. Han,²⁸ Thuang Hlaling,²⁹ Daniel S. T. Hughes,^{8,30} Adam M. Jenkins,³¹ Xiaofang Jiang,^{32,37} Irwin Jungreis,^{2,33} Evdokia G. Kakani,^{33,34} Maryam Kamali,³⁵ Petri Kempainen,³⁶ Ryan C. Kennedy,³⁷ Ioannis K. Kirmizoglu,²² Lizette L. Koekemoer,³⁸ Njoroge Laban,⁴⁰ Nicholas Langridge,⁴⁵ Mara K. N. Lawnczak,¹⁶ Manolis Lirakis,⁴¹ Neil F. Lobo,⁴² Ernesto Lowy,⁴³ Robert M. MacCallum,¹⁶ Chunhong Mao,⁴⁴ Gareth Maslen,⁴⁴ Charles Mbogo,⁴⁴ Jenny McCarthy,¹¹ Kristin Michel,²² Sara N. Mitchell,³⁵ Wendy Moore,⁴⁵ Katherine A. Murphy,²⁹ Anastasia N. Naumenko,³⁵ Tony Nolan,¹⁶ Eva M. Novoa,^{2,3} Samantha O'Loughlin,¹⁵ Chioma Oringanje,⁴⁵ Mohammad A. Oshaghi,⁴ Nazy Pakpour,⁴⁶ Philippos A. Papatheos,^{16,24} Ashley N. Peery,³⁵ Michael Povelones,⁴⁷ Anil Prakash,⁴⁸ David P. Price,^{52,53} Ashok Rajaraman,¹⁹ Lisa J. Reimer,⁵¹ David C. Rinker,⁵² Antonis Rokas,^{52,53} Tanya L. Russell,¹⁷ N'Fale Sagnon,²⁶ Maria V. Sharakhova,³⁵ Terrance Shea,³⁵ Felipe A. Simão,⁴⁹ Frederic Simard,⁵¹ Michel A. Slotman,⁵⁴ Pradya Sombon,³⁵ Vladimir Stegely,¹² Claudio J. Struchiner,^{55,57} Gregg W. C. Thomas,²⁸ Marta Tojo,⁵⁹ Pantelis Topalis,²³ José M. C. Tobo,⁶⁰ Maria F. Unger,⁴² John Vontas,⁴³ Catherine Walton,²⁶ Craig S. Wilding,⁶¹ Judith H. Willis,⁶² Yi-Chieh Wu,^{2,3,63} Guiyun Yan,⁶⁴ Evgeny M. Zdobnov,^{5,5} Xiaofan Zhou,⁵³ Flaminia Catteruccia,^{33,34} George K. Christophides,¹⁶ Frank H. Collins,⁴² Robert S. Corran,⁶² Andrea Crisanti,^{16,24} Martin J. Donnelly,^{31,60} Scott J. Enrlich,¹⁵ Michael C. Fontaine,^{32,66} William Gilbart,⁶⁷ Matthew W. Hahn,^{36,50} Immo A. Hansen,^{39,50} Paul I. Howell,⁶⁸ Fotis C. Kafatos,¹⁶ Manolis Kellis,²³ Daniel Lawson,⁴ Christos Louis,^{41,23,69} Shirley Luckhart,⁴⁴ Marc A. T. Muskavich,^{21,70} José M. Ribeiro,⁷¹ Michael A. Riehle,²⁶ Igor V. Sharakhov,^{35,37} Zhijian Tu,^{27,23} Laurence J. Zwiebel,⁷² Nora J. Besansky^{42,†}



- Synteny conservation at arm level
- Gene order shuffling
- Striking difference of rearrangement rates between sex chromosomes and autosomes, compared to other dipteran genomes.

Figure S8. X versus autosome rearrangement rates.

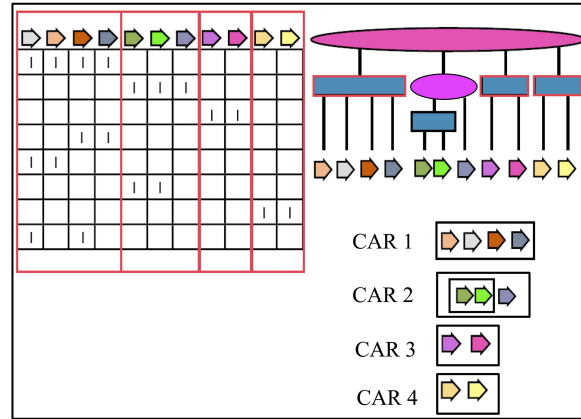
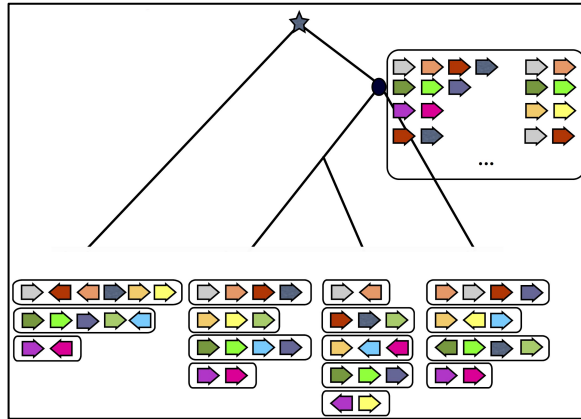
The ratio of X chromosome evolutionary rate to the total rate of rearrangement in anophelines and drosophilids.

Filtered input:



- ❑ When the project started, following the direction of our collaborators, we focused on a smaller subgroup of 11 species:
 - ❑ Gambiae complex: *An. arabiensis*, *An. quadriannulatus*, *An. gambiae*, *An. merus*
 - ❑ *An. funestus*, another African species
 - ❑ *An. minimus*, *An. stephensi*, *An. dirus*, *An. farauti*, a group of Asian species
 - ❑ *An. atroparvus*, *An. albimanus*, two species from Central and South America.
- ❑ Pressed by lack of time to develop novel methods, we applied a well tested method for reconstructing ancestral gene orders that considers only single-copy orthologous gene families, i.e. gene families where there is a single gene copy in each species.
- ❑ This resulted in 5,343 genes per genome, a gene complement of roughly half of the genes expected to be present in each species (read: we threw out half of the gene data).

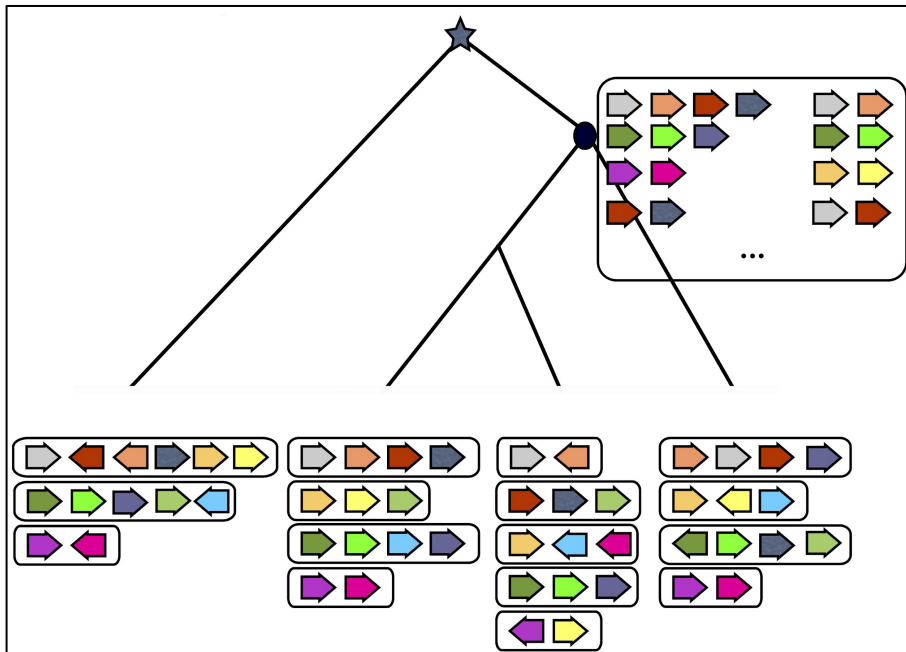
Reconstructing ancestral genomes from unique genes: methods (AnGeS)



Chauve & Tannier, PLoS Comp. Biol. (2008)
 Ouangraoua *et al*, J. Comp. Biol. (2010)
 Gavranovic *et al*, Bioinformatics (2011)
 Wittler *et al*, CPM (2011)
 Jones *et al*, Bioinformatics (2012)

- ❑ Detect group of co-localized **unique** genes conserved in several genomes (Dollo criterion)
- ❑ Weight such groups in based on their conservation pattern
- ❑ Extract a **maximum (weight) subset** that is compatible with a linear order
- ❑ **CARs**: Contiguous Ancestral Regions

Step 1: Dollo detection of syntenic conservation



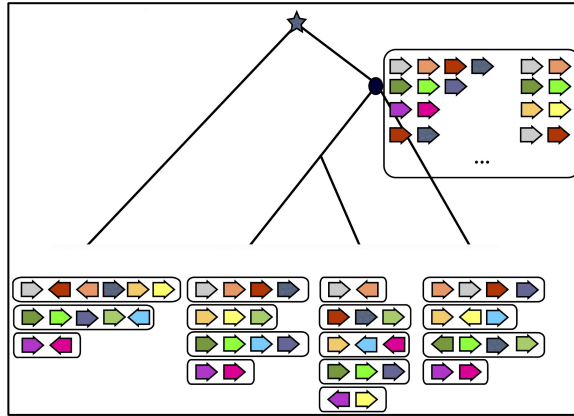
We consider a single ancestral genome of interest (i.e. we process each internal node of the tree independently ... strong departure from the SPP framework). Denote it **A**.

We compare each pair of extant species (**X**,**Y**) such that the path between **X** and **Y** in the tree contains the node **A**.

A **gene adjacency** **gh** is conserved in (**X**,**Y**) if both genes appear consecutively in both extant species.

A **common interval** to **X** and **Y** is a group of at least 3 genes that appear consecutively in **X** and **Y** with no order conservation constraint.

Step 1: encoding conserved features in a binary matrix



	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1						
2					1	1	1			
3								1	1	
4			1	1						
5	1	1								
6					1	1				
7									1	1
8										
9										
10	1		1							

The set of adjacencies and common intervals detected as conserved represent **putatively ancestral groups of co-localized genes**.

This set can be encoded into a **binary matrix**:

- ❑ Columns of the matrix represent genes
- ❑ An adjacency is encoded in a row with two 1s
- ❑ A common interval of k genes is encoded in a row with k 1s.

A “bioinformatics theorem”



1	1	1	1							
				1	1	1				
							1	1		
		1	1							
1	1									
				1	1					
									1	1
1		1								

If all the adjacencies and common intervals were actually present in the ancestral gene order of interest, then the binary matrix has the **Consecutive Ones Property (C1P)**:

The columns of the matrix can be permuted in such a way that along each row, the entries 1 appear consecutively.












Weaker statement: if the adjacencies and common intervals are **conflict-free**, the matrix is C1P.

What can cause conflict?

- ❑ Biology: Convergent evolution for example, that creates independently conserved features along different lineages of the species tree.
- ❑ Data acquisition: Mis-clustering of gene into families, assembly errors.

A discrete mathematics theorem



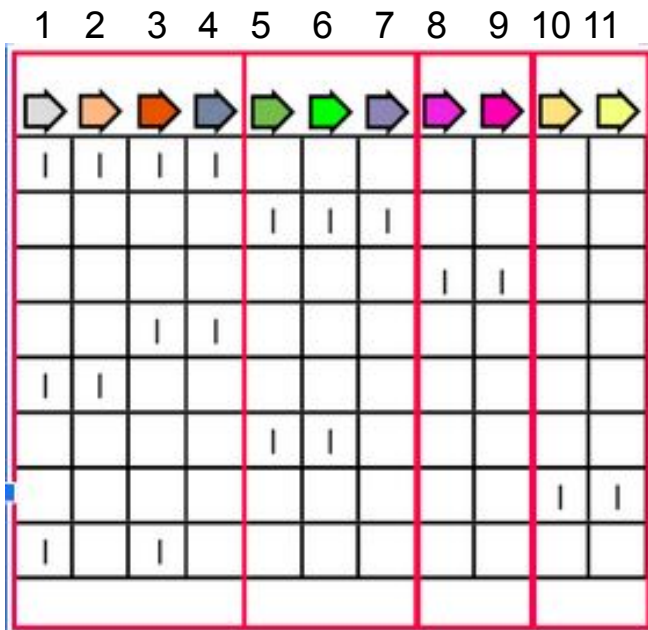
Let \mathbf{M} be a binary matrix with m rows, n columns and e entries 1.

One can decide in linear time and space $O(n+m+e)$ if \mathbf{M} is C1P.

Moreover, the whole set of permutations of the columns of \mathbf{M} that makes the rows of permuted matrix have all entries 1 consecutive can be encoded in a linear space data structure called a **PQ-tree**, that can be computed in linear time and space.

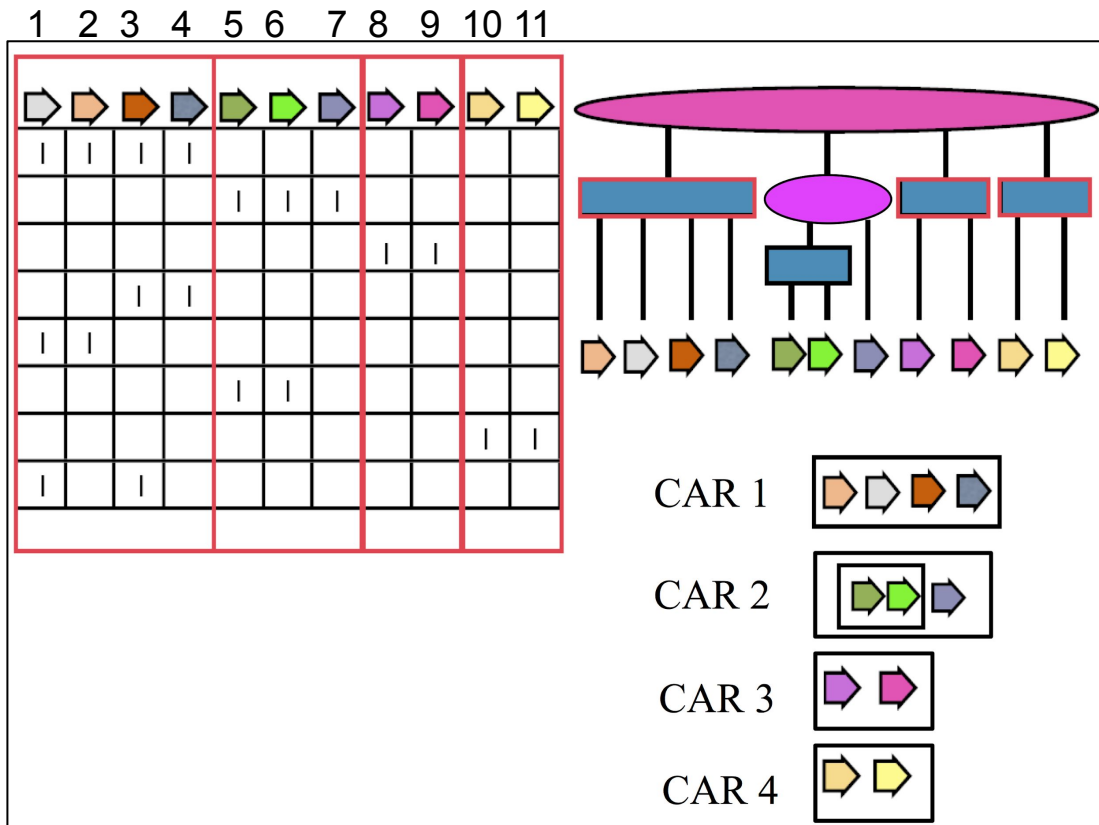
Bioinformatics application: detecting if the conserved adjacencies and common intervals have some conflict can be done in polynomial time, and if there is none, all possible ancestral gene orders can be computed and represented in linear time and space.

The algorithm



1. We compute the overlap graph, defined as the graph whose vertices are the rows of \mathbf{M} and two vertices form an edge if the corresponding sets do overlap (non-empty intersection and non-empty symmetric difference).
2. We then process each connected component independently: each form a maximal subtree of the PQ-tree. These subtrees are related into a *P-node*.
3. For each component, we apply *partition refinement*. If it fails, \mathbf{M} is not C1P. Otherwise, it fixes the relative order of groups of columns, described by a *Q-node*.
4. We repeat on each such group, to expand the tree or detect that \mathbf{M} is not C1P.

Illustration

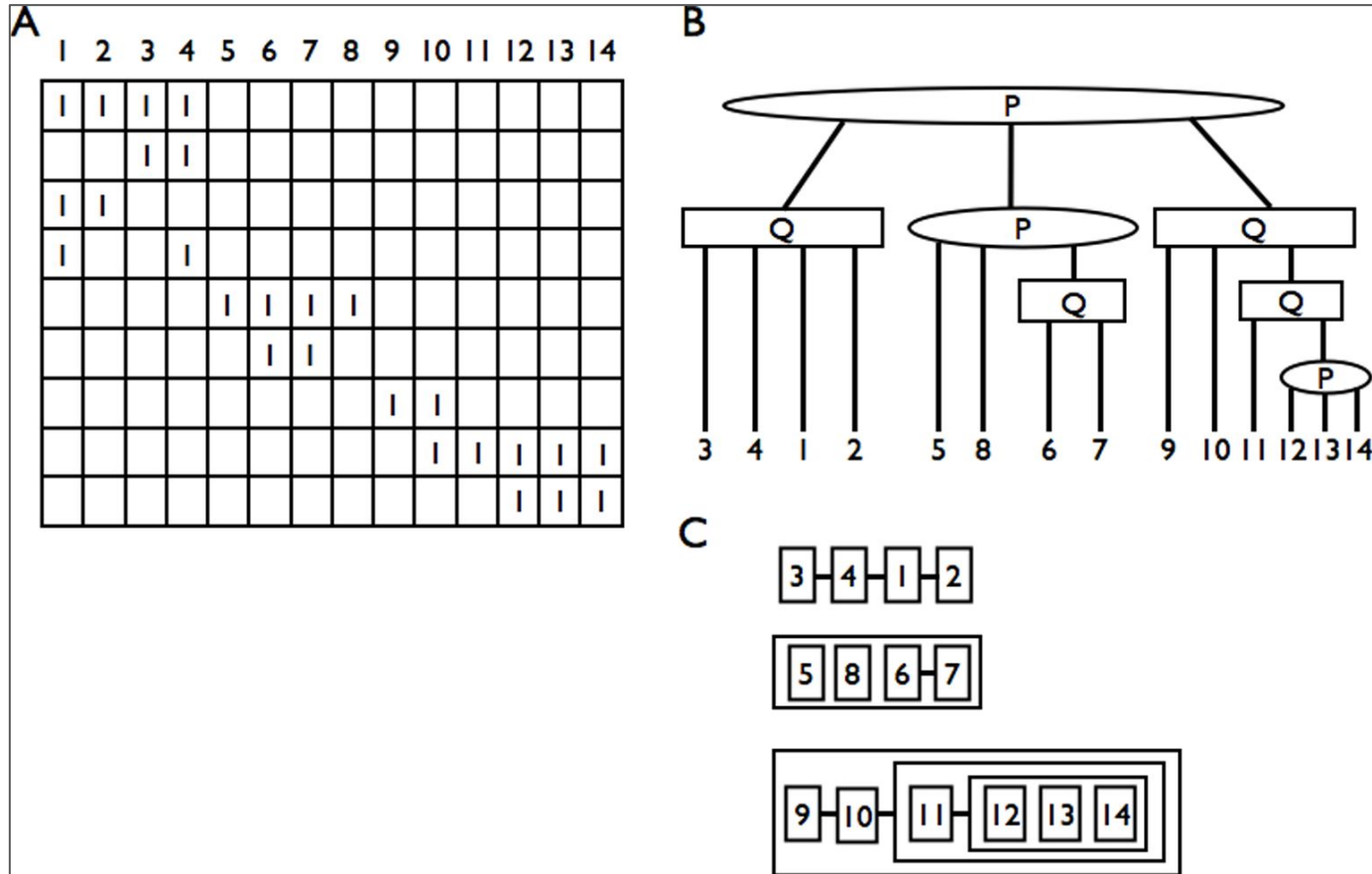


The maximal subtrees for *Contiguous Ancestral Regions* (CARs).

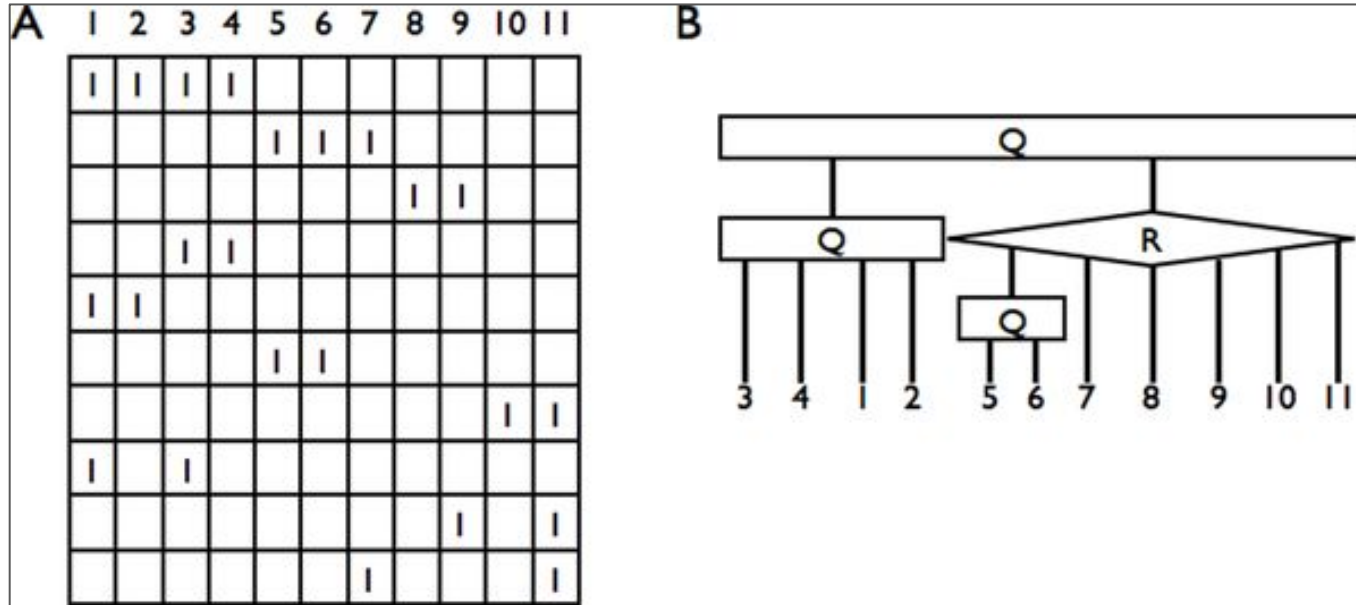
Within CARs, the tree structure encodes alternative gene orders that are compatible with the input matrix.

In our example, CAR 2 encodes 4 gene orders up to a full reversal of the CAR. All other CARs encode a single gene order.

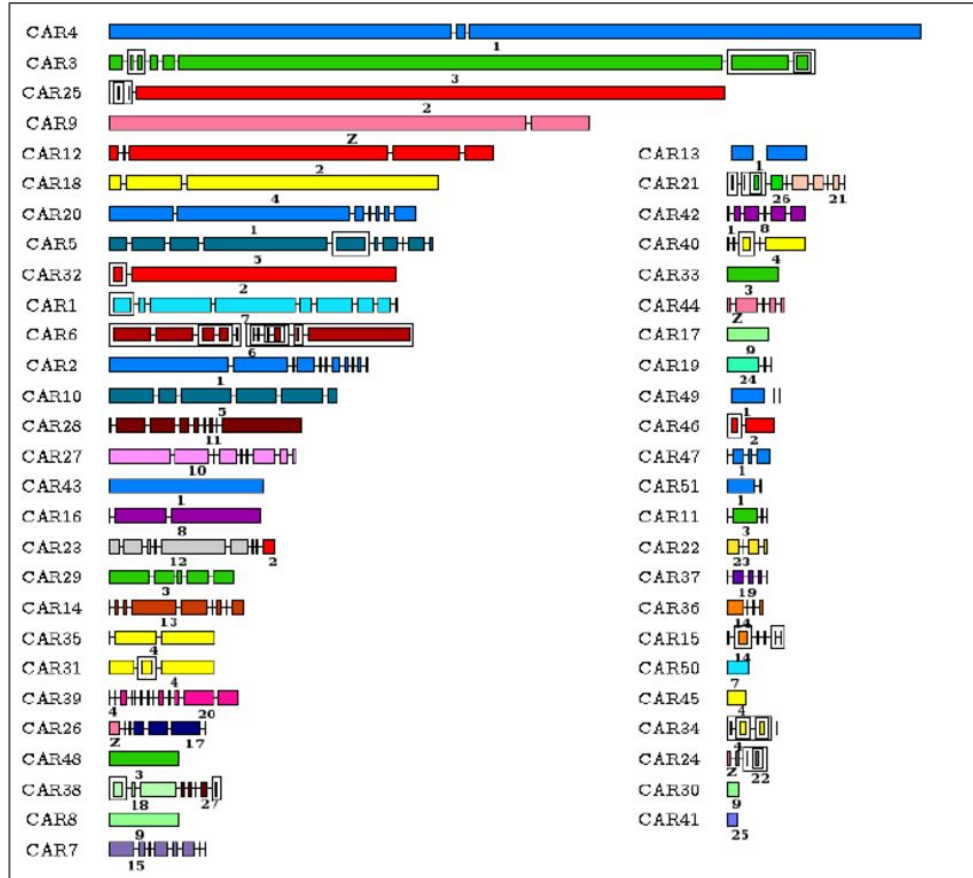
A more complex example



A non C1P matrix



The PQ-tree of the mammalian ancestor



Chauve & Tannier, PLoS Comp. Biol. (2008)

What to do if the matrix is not C1P



With real data, we can expect that the matrix resulting from detecting conserved gene adjacencies and common intervals is not C1P.

The parsimonious approach asks to modify this matrix in a minimal (i.e. parsimonious) way in order to make it C1P. It becomes an optimization problem.

All known versions of this problem are known to be NP-hard:

- ❑ Remove the minimum number of rows to make the matrix C1P.
- ❑ Modify the minimum number of entries to make the matrix C1P.
- ❑ Split rows of the matrix into several rows.
- ❑ ...

In practice, we implemented in AnGeS (1) a greedy heuristic, that takes *weighted rows* (the weight being defined by the number of species where the feature encoded by the row is found) and process rows in decreasing weight order, discarding rows whose addition creates a conflict and (2) a branch-and-bound algorithm that ensures to find an optimal solution.

A few comments



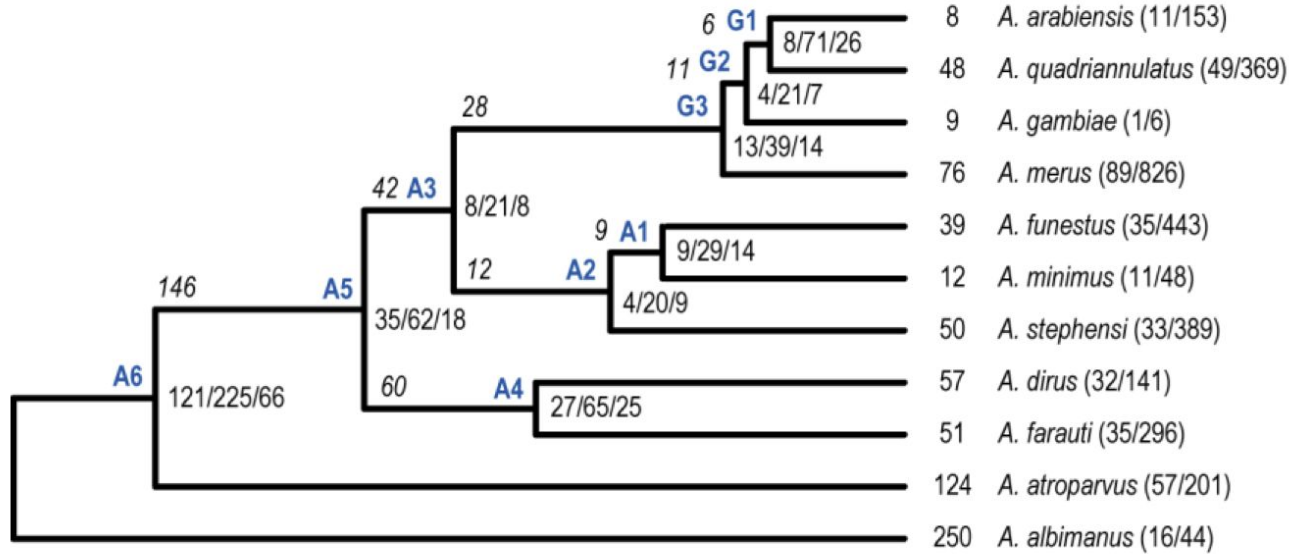
Positive aspects of the method.

- ❑ It is not a single-step method, which allows to optimize every step (detection/weighting of adjacencies and intervals, cleaning of conflicts).
- ❑ It produce a structure (PQR-tree) even before conflict cleaning, which allows to get a good view of the level of conflict in the data and of features that would be common to all possible gene order even prior to cleaning conflicts.
- ❑ Once conflict are cleaned, it produces a compact representation of all gene orders that are consistent with the matrix.

Negative aspects of the method.

- ❑ It does not provide a single gene order, which makes the interpretation more difficult.
- ❑ Computing the distance between sets of gene orders represented by PQ-trees is NP-hard (Jiang, C. Zhu (2010)).
- ❑ It process every ancestral node independently.
- ❑ It does not account for duplicated genes or allow for some flexibility to handle small errors in the adjacencies/intervals.

Results: CARs

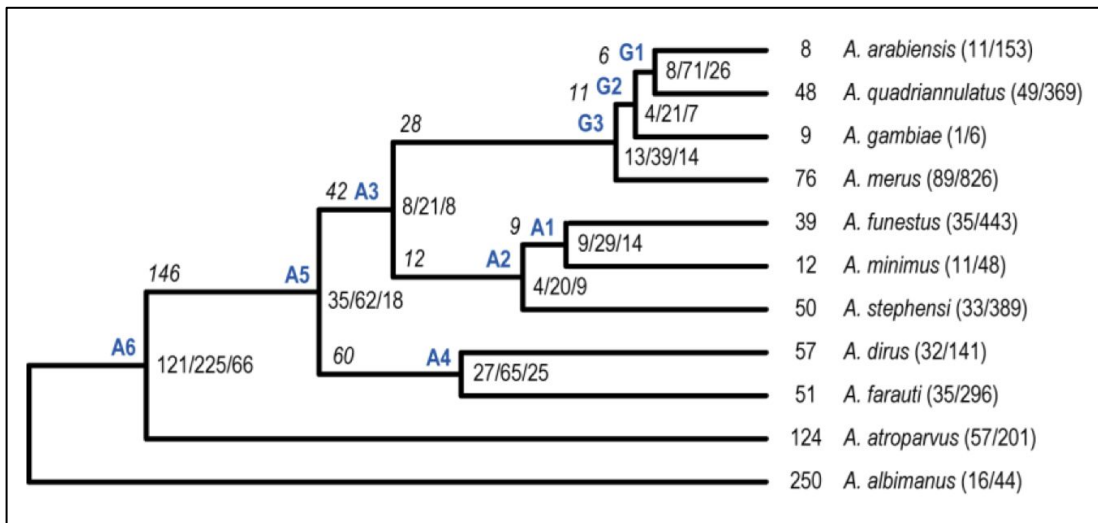


X / W / N90:

1. X = number of X chromosome CARs
2. W = number of WG CARs
3. N90 = number of CARs covering 90% of the 5,343 genes

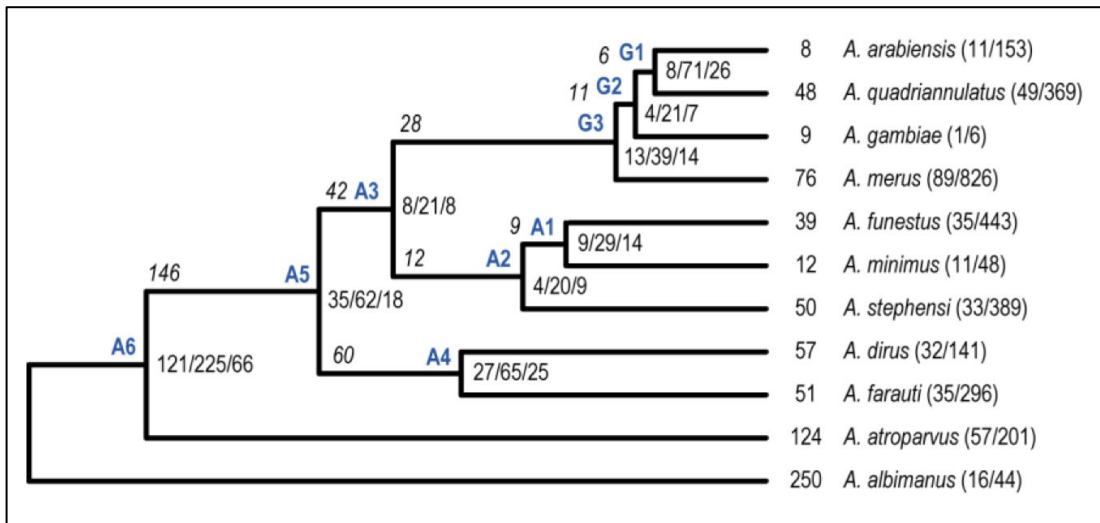
Branches length: rearrangement (DCJ) distance

Results: resolution and conflict



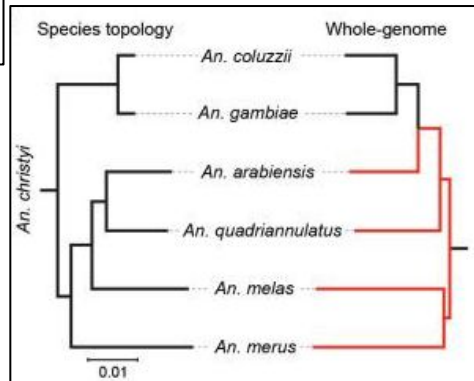
- Very low level of syntenic conflict (for every node, at most 10 discarded adjacencies/intervals are needed to clear all conflicts).
- A few number of CARs cover most genes, showing a pretty good resolution of the ancestral gene orders, even for ancestors of clades of poorly assembled genomes.

Comments: Gambiae complex phylogeny

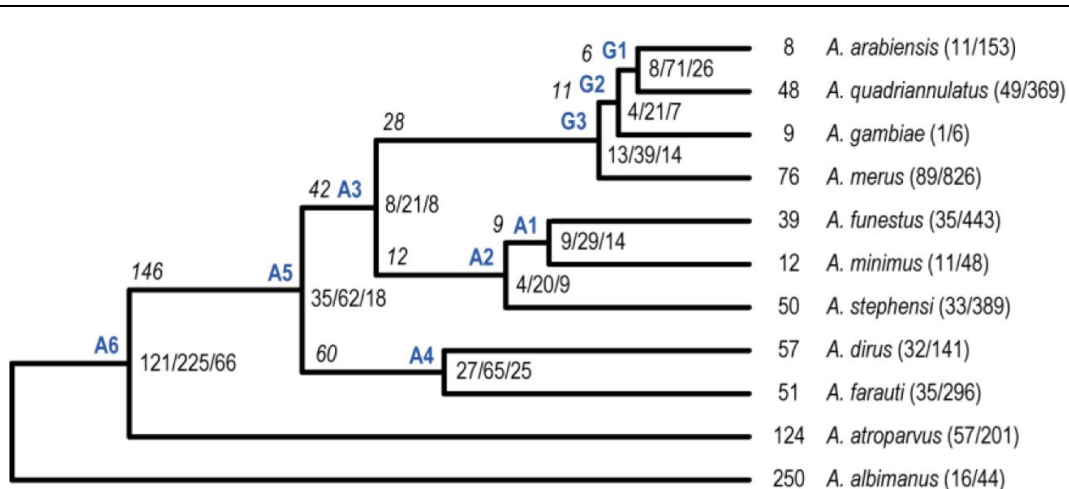


The ancestral reconstruction seems to support the proposed phylogeny of the Gambiae complex:

Assuming the alternative phylogeny that groups *An. gambiae* and *An. Arabiensis* in a clade results in a much higher level of syntenic conflict within the Gambiae complex.



Results: differentiated rearrangement rates



Hypothesis: the rate of genome rearrangements is much higher in the X chromosome than in the autosomes.

To test this we defined CARs as X-CARs or autosome-CARs based on the location of orthologs in the well assembled *An. gambiae* genome.

We corrected the distance to account for gene order fragmentation.

On 11 branches we observe that the rate of rearrangement is at least twice higher on the X chromosome than on the autosome, thus confirming the starting hypothesis.

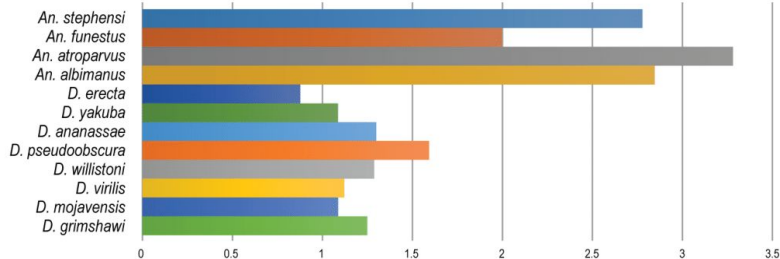


Figure S8. X versus autosome rearrangement rates.

The ratio of X chromosome evolutionary rate to the total rate of rearrangement in anophelines and drosophilids.

Conclusion

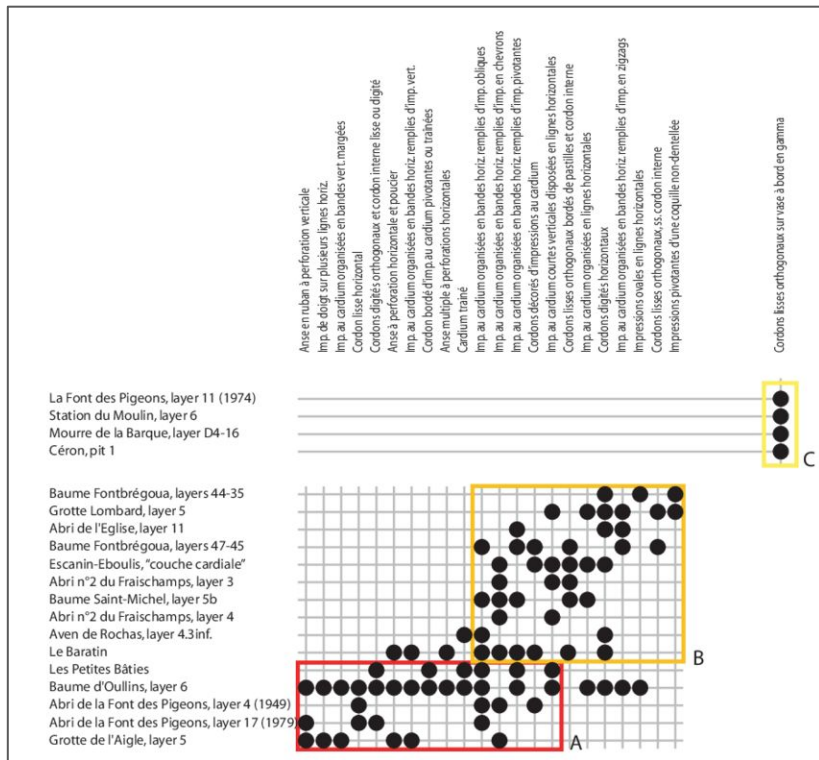


- ❑ We applied an off-the-shelf well tested method (AnGeS) to reconstruct ancestral gene orders that does not account for duplicated genes.
- ❑ This forced us to discard roughly half of the gene data.
- ❑ Nevertheless, we worked completely independently from our biologist collaborators and we obtained several results of interest:
 - ❑ Well-resolved ancestral gene orders, despite the fragmentation of the extant genomes assemblies, obtained with a low level of conflict.
 - ❑ Strong support for the alternative Gambiae complex phylogeny supported by the evolution of genes on the X chromosome (that avoids introgression that add noise to the phylogenetic signal).
 - ❑ Strong support for the hypothesis if a higher rate of genome rearrangements on the X chromosome compared to the autosomes.

Additional comments on the C1P



- ❑ The C1P was actually “invented” a long time ago by archeologists aiming to serialize artifacts found in different ground layers.



Additional comments on the C1P



- ❑ The C1P was actually “invented” a long time ago by archeologists aiming to serialize artifacts found in different ground layers.
- ❑ It was “rediscovered” in the 50s to provide evidence of the linear structure of genes.

A 2	0011111	{	A 2 a	0	1364
			A 2 b	0	386
			A 2 c	0	168
			A 2 d	0	193
			A 2 e	0	168
			A 2 f	0	PT15
			A 2 g	0	H 81
			A 2 h 1	0	
			A 2 h 2	0	
A 3	0001111	{	A 3 a-d	0	PT153
			A 3 e	0	184
			A 3 f	0	250
			A 3 g	0	033
			A 3 h	0	
			A 3 i	0	

Additional comments on the C1P



- ❑ The C1P was actually “invented” a long time ago by archeologists aiming to serialize artifacts found in different ground layers.
- ❑ It was “rediscovered” in the 50s to provide evidence of the linear structure of genes.
- ❑ It was then taken over by mathematicians, who studied the formal problem of recognizing C1P matrices and computing PQ-trees.
- ❑ In the 80s, before sequencing was well developed, it was used to reconstruct physical maps of genomes from hybridization data.
- ❑ In a sense, ancestral gene orders and CARs can be seen as the equivalent, for ancestral species, of genome maps.
- ❑ It has recently been used with success to assemble genomes with long noisy reads, through the technique of spectral seriation.

Variations on the C1P



How could-we extend the C1P approach to handle expected issues with real data:

- ❑ Errors in the acquisition of adjacencies and intervals (e.g. due to incomplete or fragmented genome assembly)
- ❑ Gene duplications
- ❑ ...

Could-we allow some controlled level of gaps (of zeros framed by 1s) in the reordered matrix? The decision problem is NP-hard in most cases, even with very constrained gap structures.

Could-we consider that genes (i.e. columns of the matrix) can-we include in the model a notion of *multiplicity* of occurrences of the gene: the C1P with multiplicity (ask Roland).