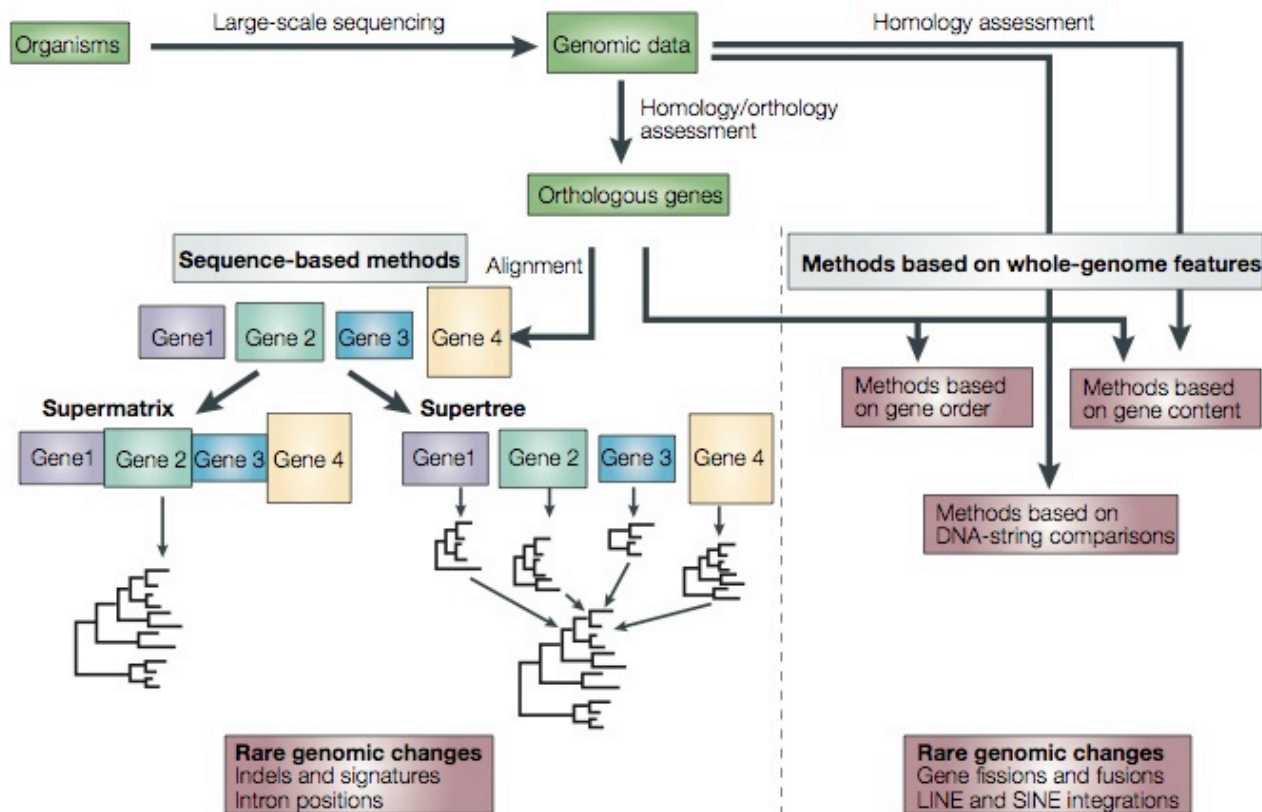
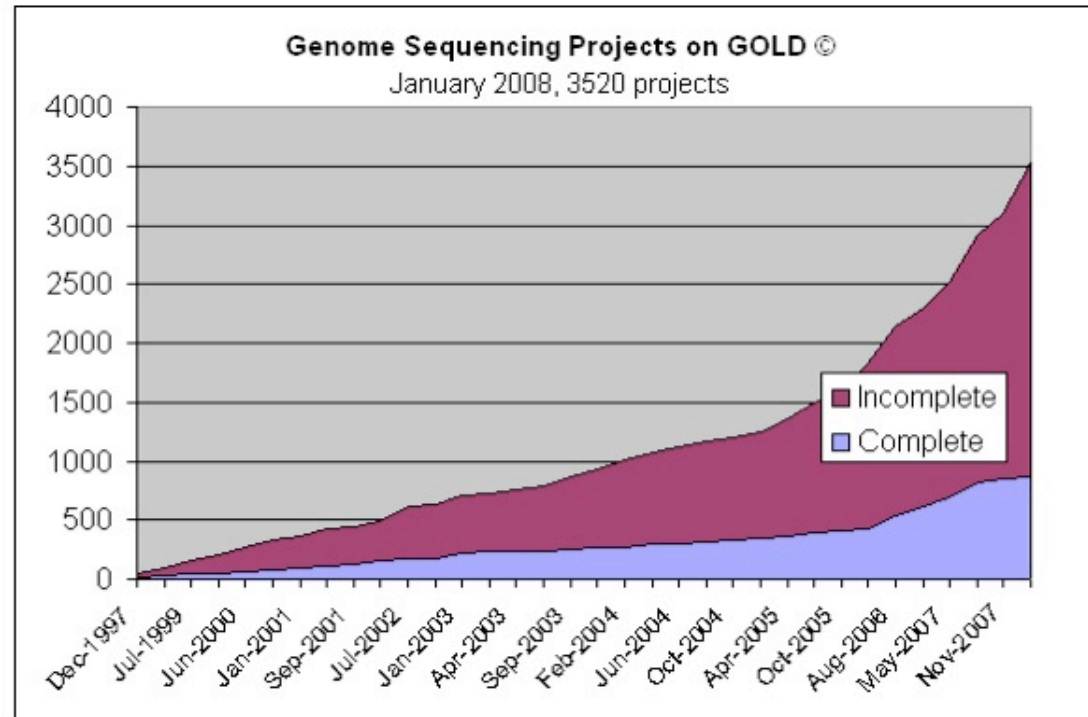


Algorithmes pour la prise en compte d'arbres de familles multigéniques en phylogénomique.

V. Berry (MAB-LIRMM) & V. Ranwez (PPP-ISEM)



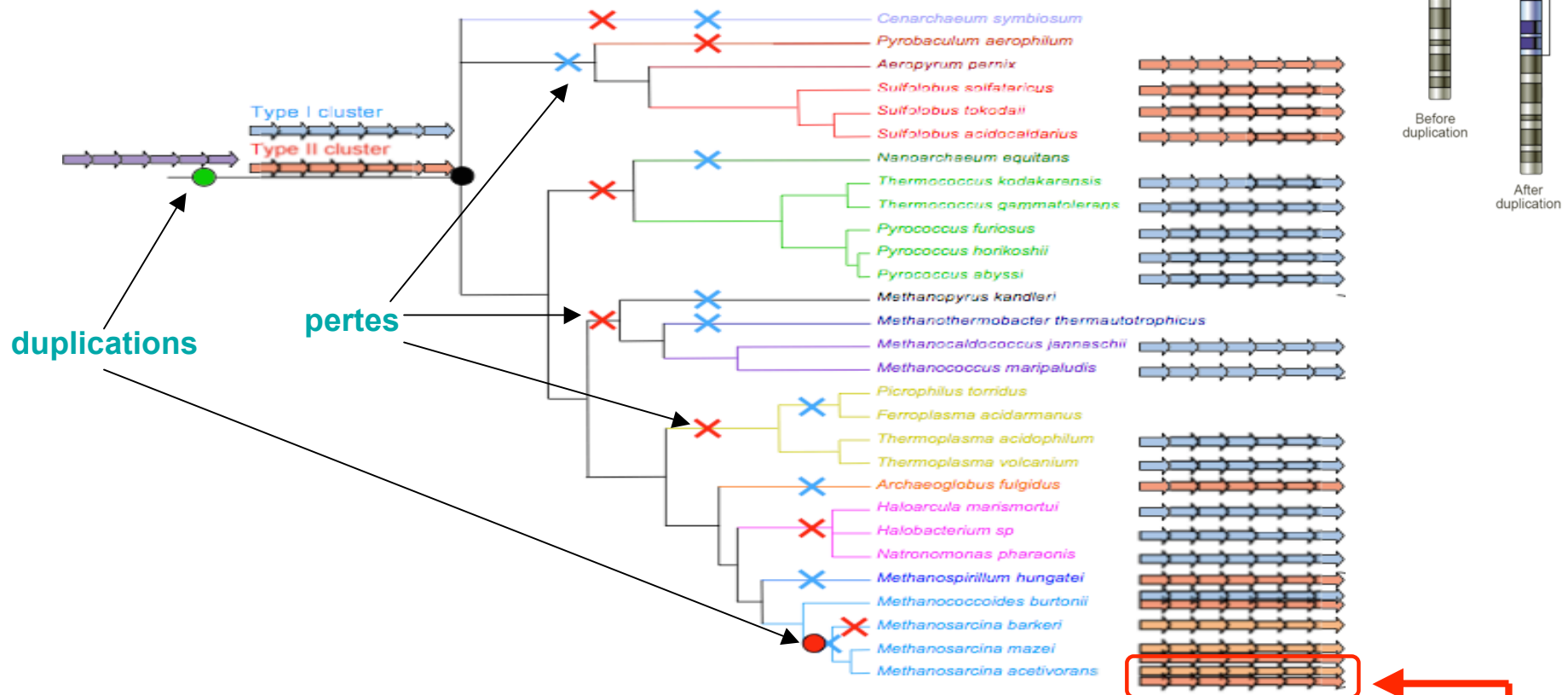
Contexte : la phylogénomique



- De nombreux génomes complets disponibles
- Banques de données de familles de gènes de plus en plus exhaustives
- **Wanted** : méthodes pour prendre en compte ces quantités énormes de données (>50 000 familles de gènes)

Contexte : les familles multigéniques

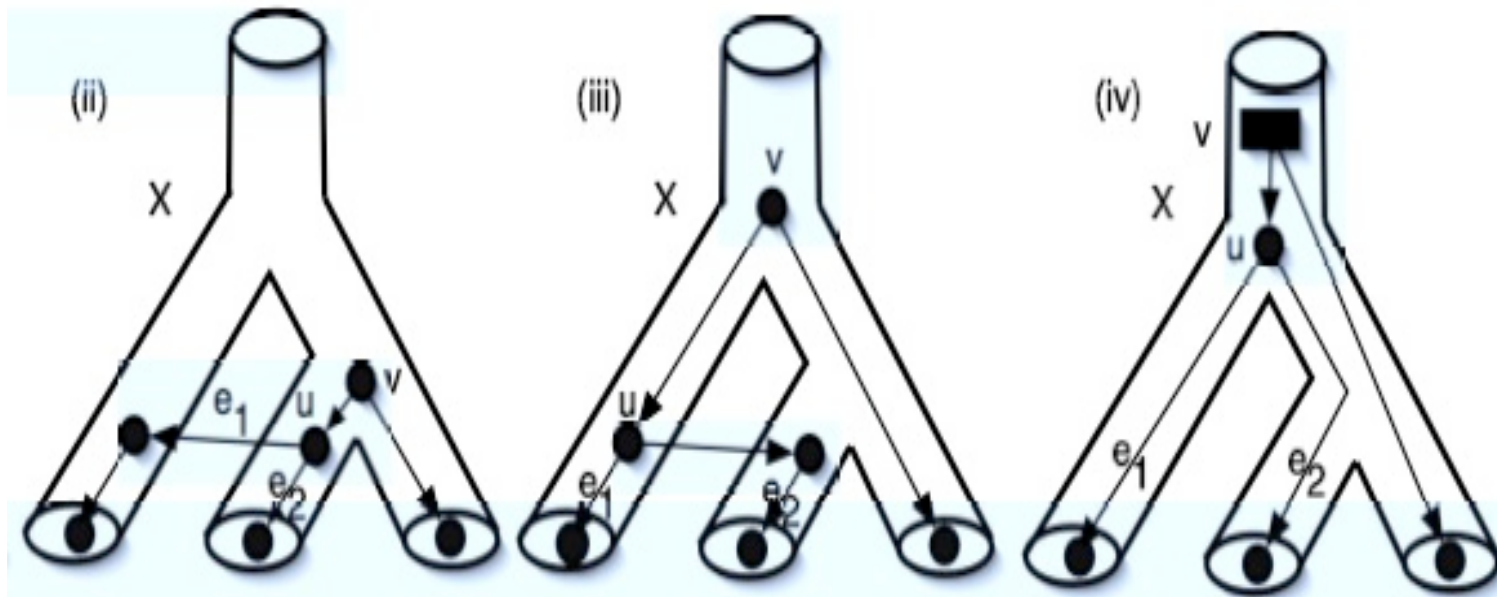
- Les arbres de gènes sont utilisés pour la construction d'arbres de grande taille et l'inférence de phénomènes macro-moléculaires : duplications, pertes et transferts de gènes



Ces événements engendrent des familles multigéniques : gènes présents en plusieurs exemplaires chez certaines espèces

Constat

- Il existe *peu* de méthodes d'inférence exploitant les familles multigéniques (35% des gènes !!!)
- Il n'existe *pas* de méthode prenant en compte toute la complexité des phénomènes sous-jacents





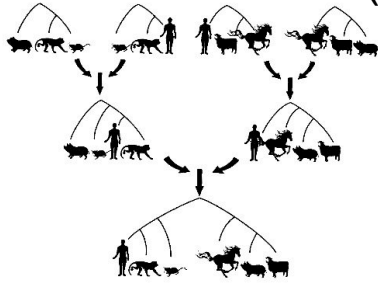
La concurrence

Soit T l'arbre des espèces et $G_1 \dots G_k$ des arbres de gènes. On considère les duplications (D), pertes (P) et transferts horizontaux (H)

Données	Questions	Résultats <i>(FPT)</i>
$T + G_1$	Inférer réconciliation -D/P -H -D/P/H (mais peu de gènes)	(Hallet Lagergren 01, Lyubetsky et al) (Adario-Berry et al 03, Hallet Lagergren 04 & 06) (Vernot et al 07,)
$T +$ séquences de $G_1 \dots G_k$	Inférer $G_1 \dots G_k$ et réconciliation	(Durand et al 05, Arvestad et al 04)
$G_1 \dots G_k$	Inférer T sous hypothèse de D/P (mais pas H)	(Page 94, Mirkin et al 95, Guigo et al 96, Zhang 97, Stege 99, Hallet et Lagergren 00, Ma et al 00,

Axes du sujet de thèse

- **Aspects théoriques** : algorithmes sur les arbres multi-labels (algorithmes polynomiaux ou FPT) :



- Proposer une **méthode de super-arbres** acceptant en entrée des arbres multi-labels et prenant en compte les événements biologiques provoquant l'apparition de copies multiples d'un gène (transferts horizontaux, duplications).
 - Proposer une **mesure de distances** biologiquement raisonnable entre arbres multi-labels. Sur la base de simulations et de données concrètes, évaluer à l'aide de cette mesure de distance les performances des méthodes de reconstruction.
 - Étendre les **algorithmes de consensus** majoritaire, MCT et MAST pour obtenir un arbre représentant un compromis entre les arbres multi-labels.
- **Validations sur données réelles et apports pratiques** :
 - Mammifères (duplications): équipe de V. Ranwez (ISEM)
 - Bactéries (duplications + HGT) : collaboration en cours avec le LBBE (Lyon).

Argumentaire encadrants

1. Niveau de publication :

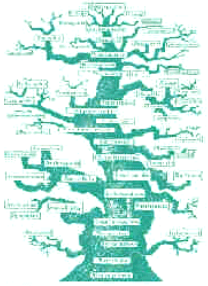
	2007	2008	In press
VB	2 revues int. + 1 conf int. LNCS	2 revues int.	1 revue int.
VR	2 revues int. (dont PNAS)		1 revue int.
VB & VR	1 revue int.	1 revue int.	

2. Niveau d'encadrement :

- VB : P.G. (50%, fin en 2010) + C.S. (50%, fin en 2009)
- VR : C.S. (50%, fin en 2009)

3. Autres :

- Vincent B. : MCF, HDR en cours
- Vincent R. : MCF, HDR à venir bientôt
- Financements communs (UM2, ESPT, ANR)



Quelques publications récentes



- GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Galtier, N., L. Duret, S. Glémin, and **V. Ranwez**. 2009. *Trends Genet.* in press.
- On the Approximation of Computing Evolutionary Trees. **V. Berry**, F. Nicolas, S. Guillemot and C. Paul, *Trans. In Algorithms*, in press.
- [PhySIC_IST : healing source trees to infer healthy supertrees.](#) [Scornavacca C.](#), **V. Berry**, E. J. P. Douzery and **V. Ranwez**. *BMC Bioinformatics* 9:413, 2008.
- From constrained to unconstrained maximum agreement subtree in linear time. **V. Berry**, Z.S. Peng and H.F. Ting, *Algorithmica*, 2008.
- OrthoMaM : A database of orthologous genomic markers for placental mammal phylogenetics **Ranwez V.**, Delsuc F., Ranwez S., Belkhir K. Tilak M. & Douzery E. J. P. *BMC Evolutionary Biology* 7 : 241, 2007.
- [PhySIC : A Veto Supertree Method with Desirable Properties.](#) **Ranwez V.**, **Berry V.**, Criscuolo A., Fabre P.H., Guillemot S, Scornavacca C., Douzery E.J.P. *Systematic Biology*, 56(5), pp 798-817, 2007
- Multiple molecular evidences for a living mammalian fossil. Huchon D, Chevret P, Jordan U, Kilpatrick CW, **Ranwez V**, Jenkins PD, Brosius J, Schmitz J, *Proc Natl Acad Sci U S A*. 2007 May 1 ;104(18):7495-9, 2007.
- Extending the maximum agreement subtree and maximum compatible tree problems to the supertree context. **V. Berry**, F. Nicolas, *Journ. Discrete Algorithms*, 2007.
- Improved parametrized complexity of Maximum Agreement Subtree and Maximum Compatible Tree problems. **V. Berry** and F. Nicolas, *IEEE/ACM Trans. on Comp. Biol. and Bioinf.*, 3(3), 2006.
- Fast Computation of Supertrees for Compatible Phylogenies with Nested Taxa. **V. Berry**, C. Semple, *Systematic Biology*, 55(2), U108-U126, Apr 2006.
- Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. J. Dutheil, S. Gaillard, E. Bazin, S. Glémin, **V. Ranwez**, N. Galtier and K. Belkhir. *BMC Bioinformatics* 7:188, 2006.