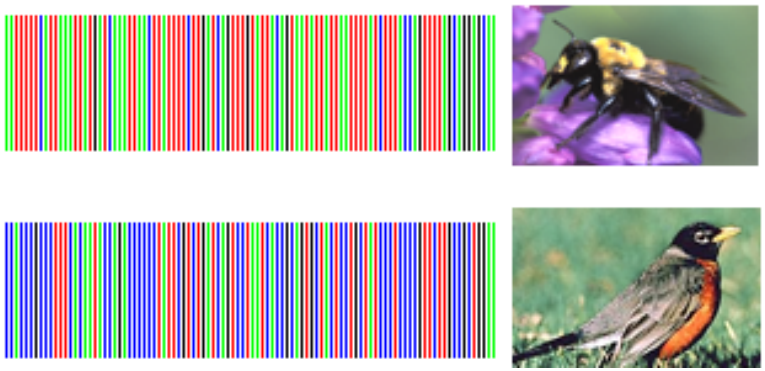
Module « Barcode ADN »

Méthodes d'analyse (2) : Extension des outils de génétique des populations aux espèces

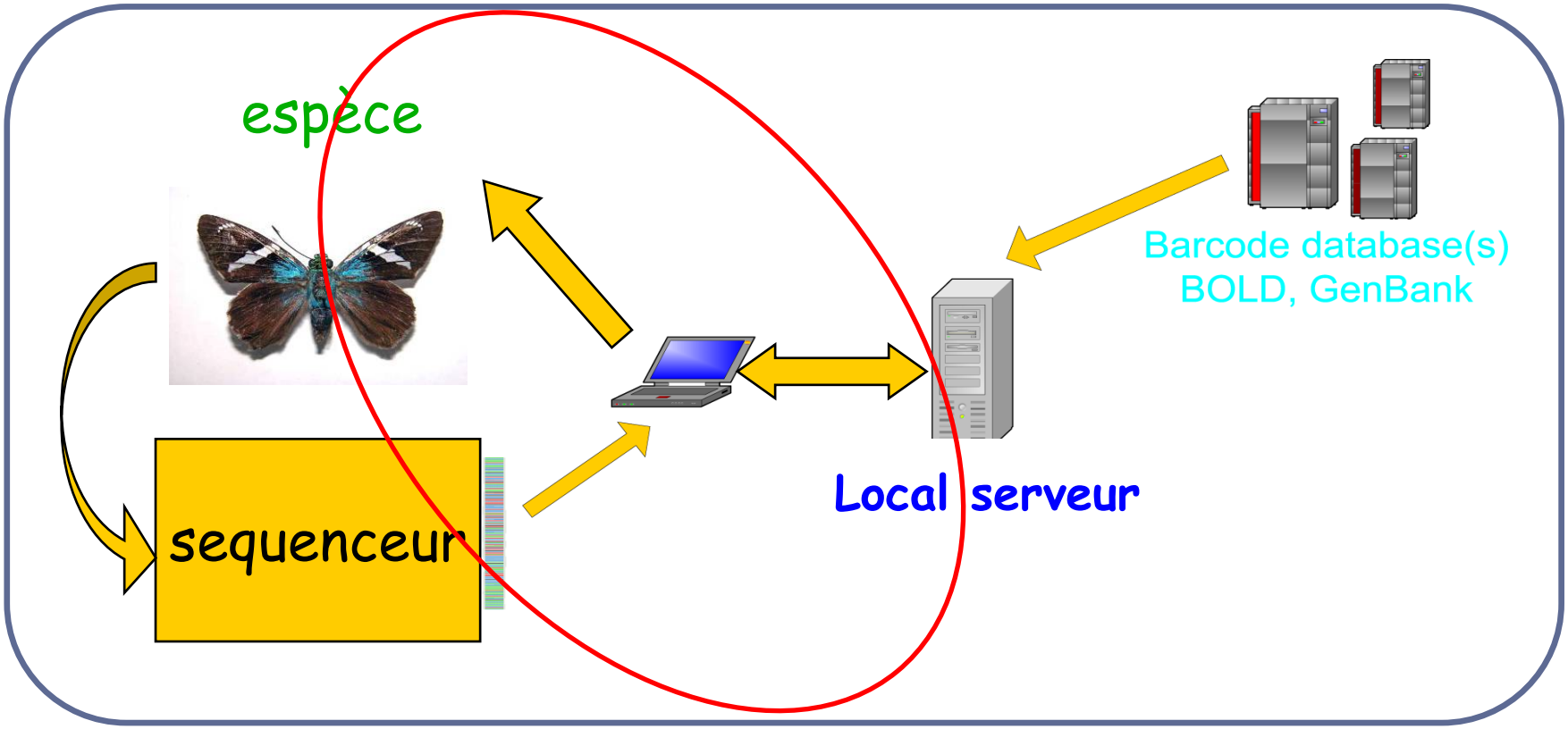


Raphaël Leblois, MC MNHN, dep^{t} Systématique & Evolution

Origine, Structure et Evolution de la biodiversité

(UMR CNRS/MNHN/IRD 5202)

Méthodes d'analyse des données du projet "Barcode ADN"



Outils d'analyse :

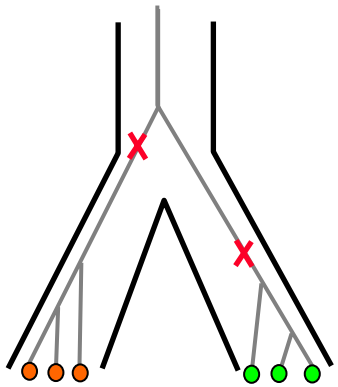
Permet d'interroger la base de données et de proposer une espèce à partir de la séquence présentée

Plan de l'exposé

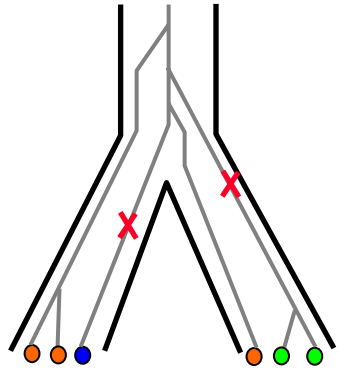
1. **Problèmes & Limites du «Barcode ADN»**
2. **Notion de tri des lignées ancestrales**
3. **Les outils disponibles**
4. **Extension des outils de génétique des pops pour :**
   a. Assignation d'une séquence à une espèce
   b. Délimitation d'espèces

1.Problèmes et limites du "Barcode ADN"

- Il n'existe pas de gène "universel/idéal" qui serait toujours variable entre espèce et invariant au sein de toutes espèces



Cas du gène idéal



Mais la réalité est plus complexe

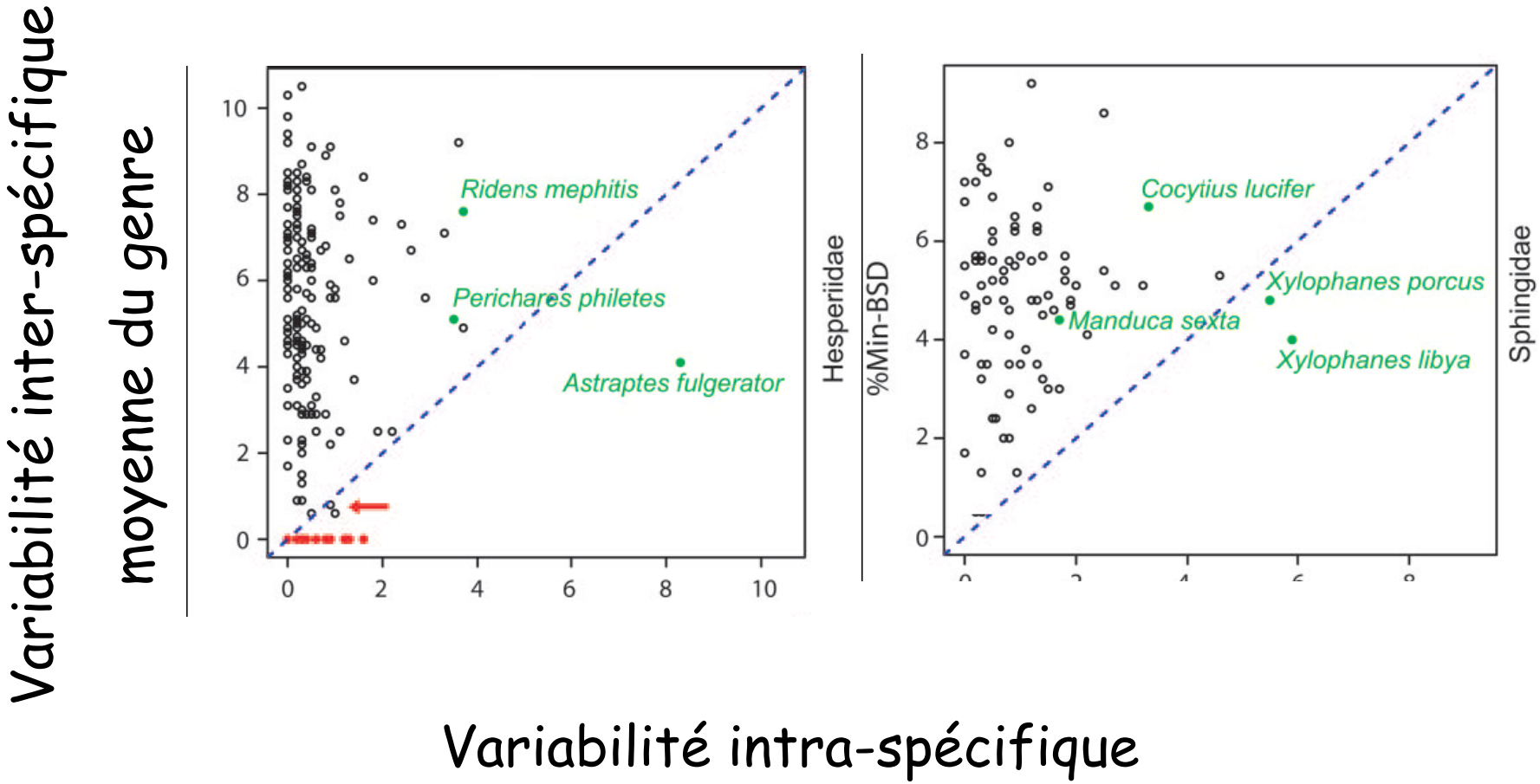
1.Problèmes et limites du "Barcode ADN"

- Il n'existe pas de gène "universel/idéal" qui serait toujours variable entre espèce et invariant au sein de toutes espèces

- Plus globalement, il n'est pas toujours vrai que la variabilité génétique intra-spécifique est très faible par rapport à la variabilité génétique inter-spécifique, même avec COI

1.Problèmes et limites du "Barcode ADN"

- **Variabilité intra- et inter-spécifique**



Un exemple de cas problématique :

Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*

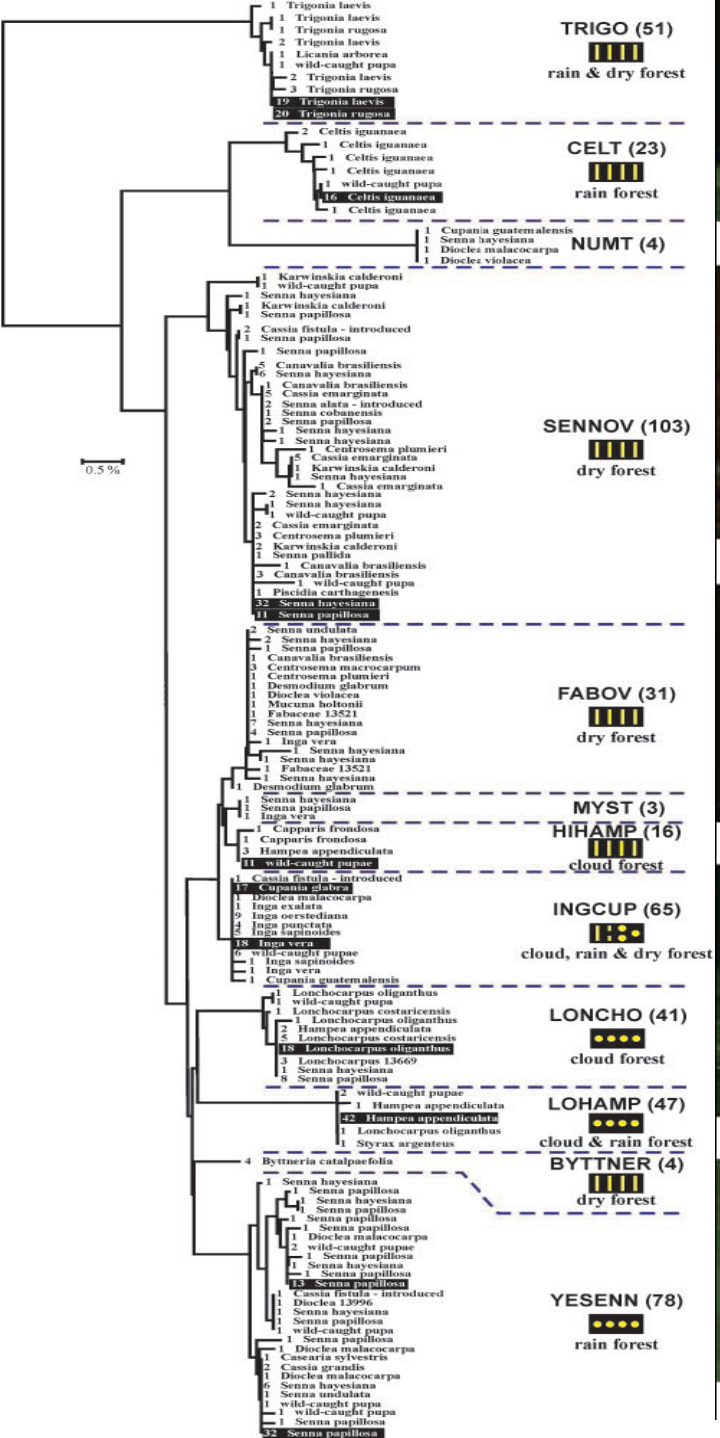
Paul D. N. Hebert*†, Erin H. Penton*, John M. Burns‡, Daniel H. Janzen§, and Winnie Hallwachs§

*Department of Zoology, University of Guelph, Guelph, ON, Canada N1G 2W1; ‡Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560-0127; and §Department of Biology, University of Pennsylvania, Philadelphia, PA 19104

Contributed by Daniel H. Janzen, August 20, 2004

Astraptes fulgerator, first described in 1775, is a common and widely distributed neotropical skipper butterfly (Lepidoptera: Hesperiidae). We combine 25 years of natural history observations in northwestern Costa Rica with morphological study and DNA barcoding of museum specimens to show that *A. fulgerator* is a complex of at least 10 species in this region. Largely sympatric, these taxa have mostly different caterpillar food plants, mostly distinctive caterpillars, and somewhat different ecosystem preferences but only subtly differing adults with no genitalic divergence. Our results add to the evidence that cryptic species are prevalent in tropical regions, a critical issue in efforts to document global species richness. They also illustrate the value of DNA barcoding, especially when coupled with traditional taxonomic tools, in disclosing hidden diversity.

Fig. 1. Newly eclosed female *A. fulgerator* (species LOHAMP, voucher code 02-SRNP-9770) from the ACG.

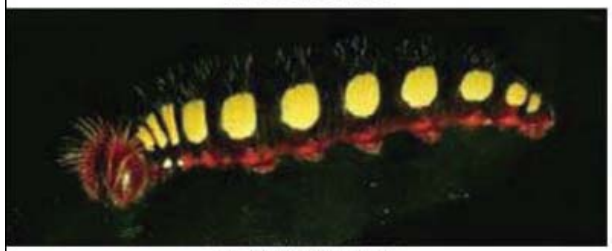


TRIGO

CELT

LONCHO

INGCUP



LOHAMP

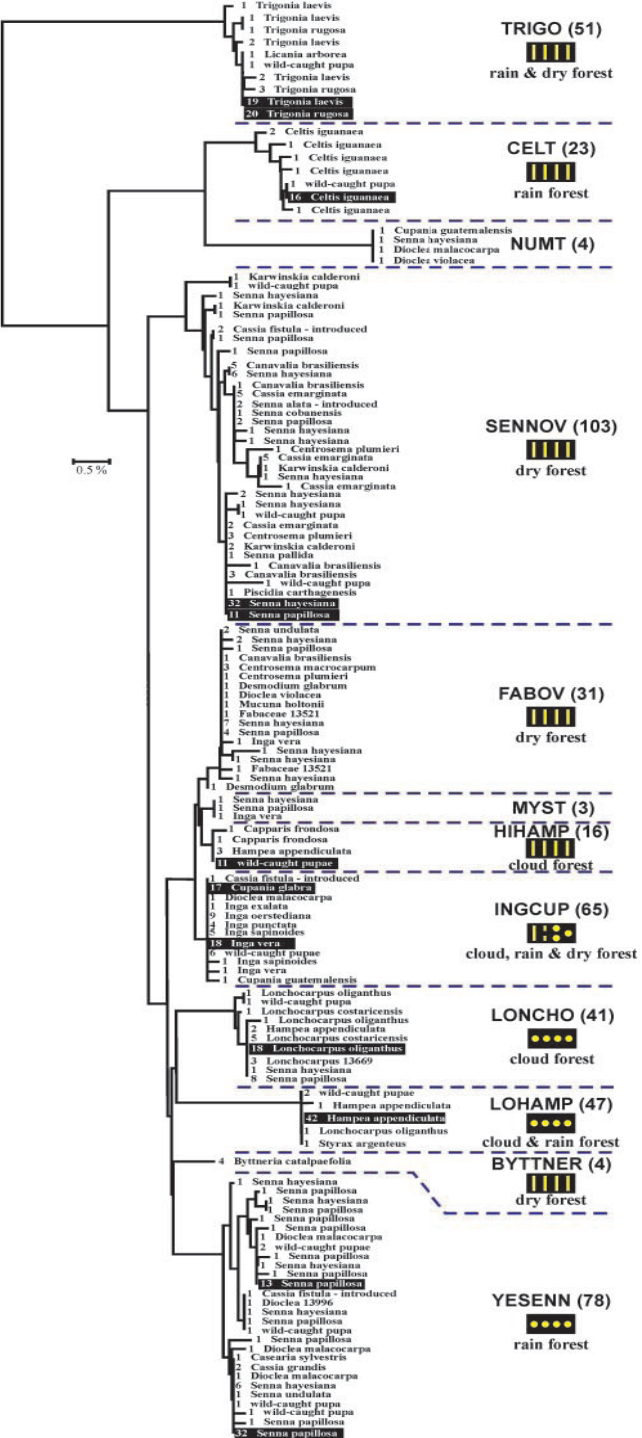
HIHAMP

BYTTNER

FABOV

YESENN

SENNOV

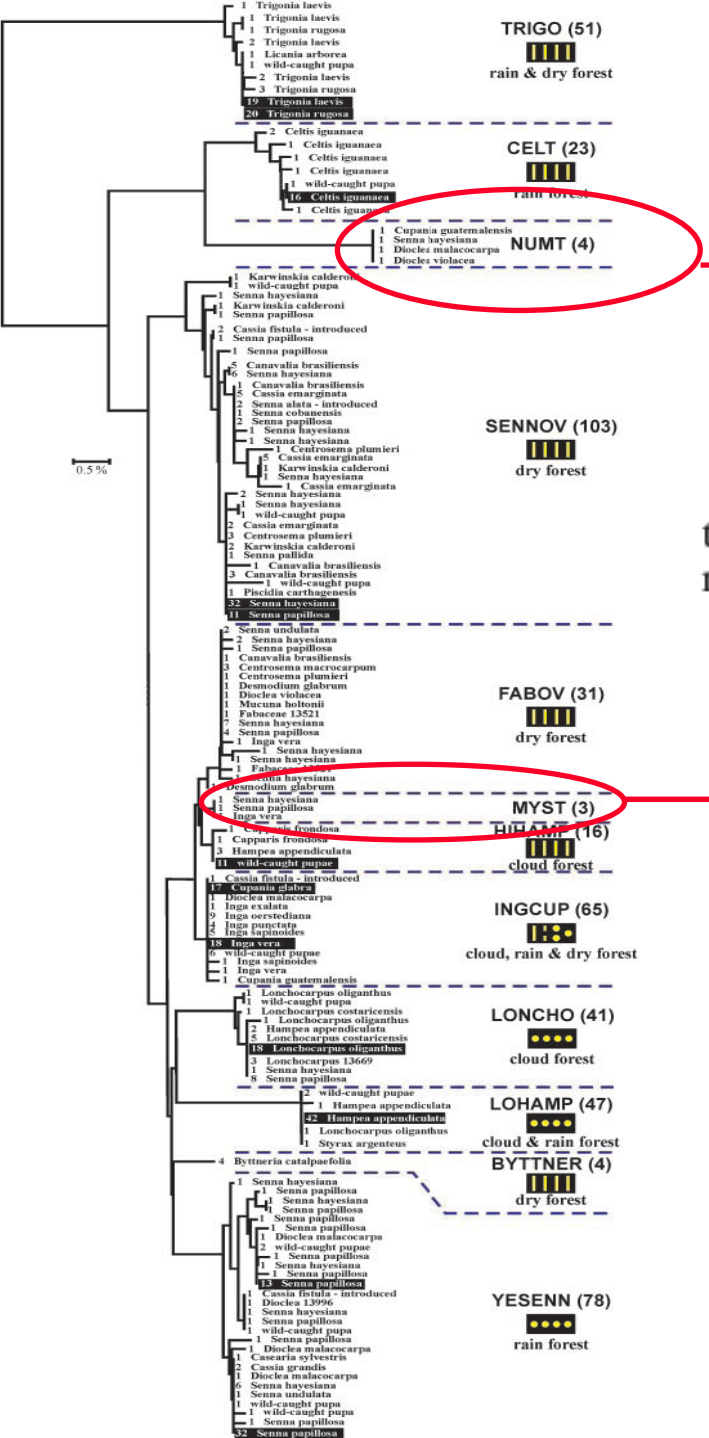


→ dans 1 espèce (déjà décrite et étudiée), ils en trouvent 10. Chacune de ces 10 espèces serait associée à une famille spécifique de plantes.

Pb potentiel : Femelles choisissent leur lieux de ponte (= une plante) et ADN mt transmit exclusivement par femelles

-> si les femelles ne change pas/peu de milieu de ponte mais que les mâles se reproduisent avec toutes les femelles sans préférences, on aura un signal de structuration sur l'ADN mt qui ne se retrouvera pas sur l'ADN nucléaire…

Intéressant mais pas forcément 10 espèces différentes… Nécessité de confirmer/infirmer avec des données nucléaires ou des expériences de croisement…

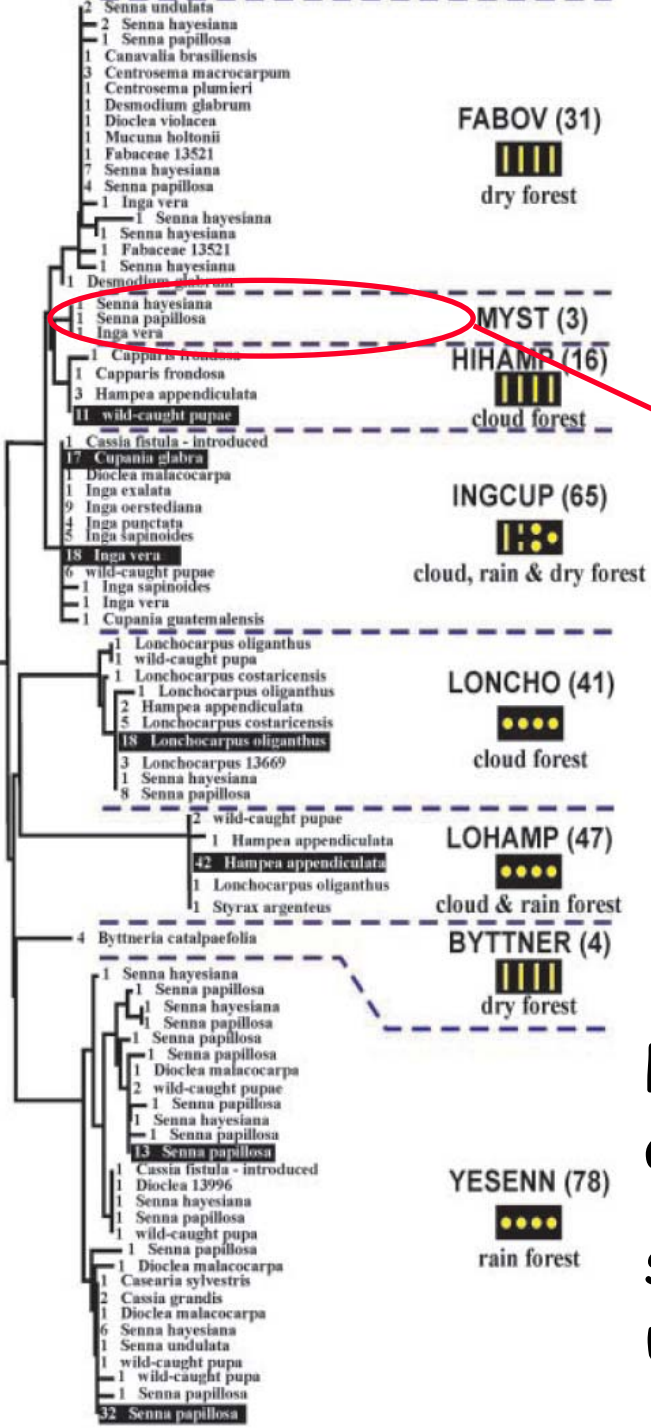


Au passage on découvre aussi des problèmes techniques …

4 individus mal classés

Two small COI groups of three (MYST) and four (NUMT) individuals are treated separately for reasons that are justified later.

3 individus mal classés



Au passage on découvre aussi des problèmes techniques …

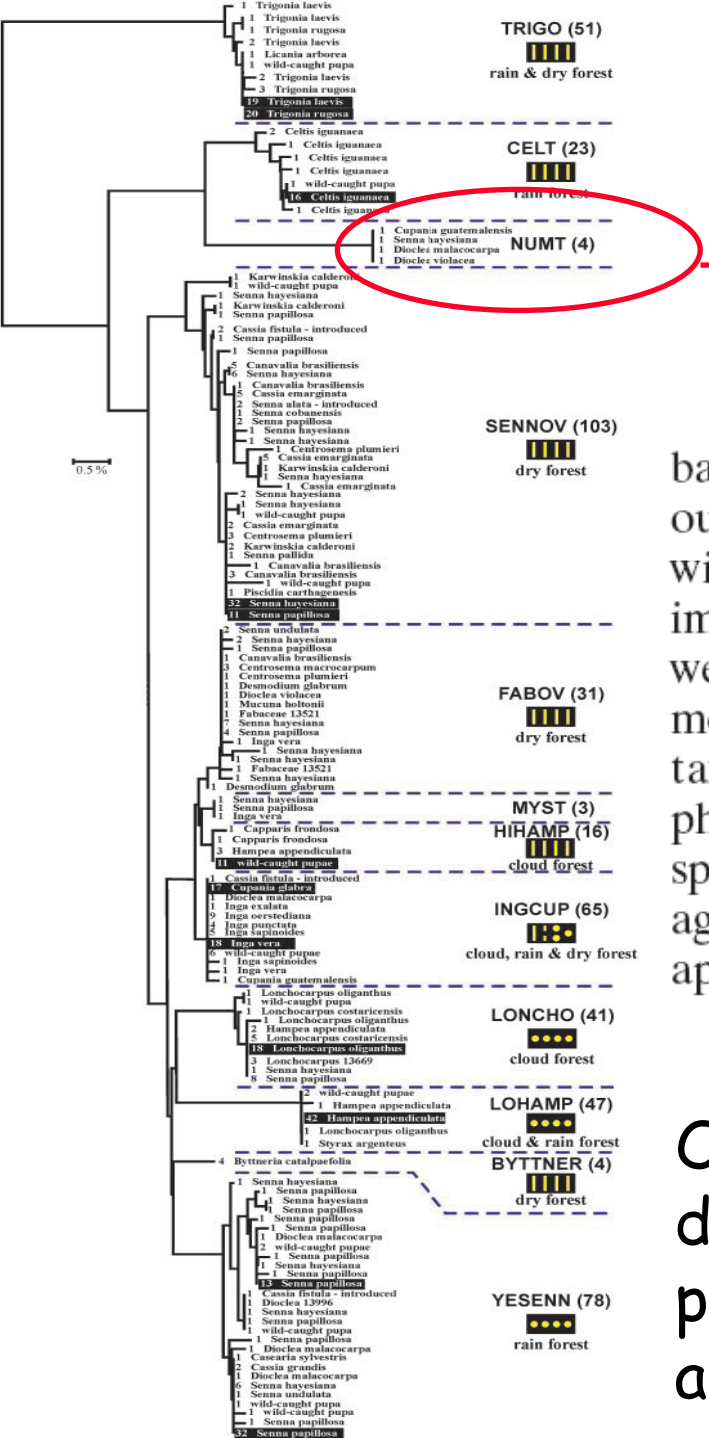
Two small COI groups of three (MYST) and four (NUMT) individuals are treated separately for reasons that are justified later.

3 individus mal classés

The MYST lineage apparently represents an exception to this pattern of mitochondrial divergence as it includes one individual from *Inga* and two from *Senna*, suggesting that it is shared by INGCUP and FABOV. As such, this likely represents a case in which lineage sorting is incomplete, a result that might have been expected given the low genetic divergence (and presumed recent origin) of these groups.

Problème dû au tri des lignées ancestrales?

seule manière de le savoir = tester sous un modèle adéquate (i.e. coalescence)



Au passage on découvre aussi des problèmes techniques …

→ Detection de copie NUMT

Numts pose a potential interpretational hazard for any PCR-based survey of mitochondrial DNA diversity (25), and 2.8% of our COI sequences showed probable coamplification of a Numt with its mitochondrial counterpart. However, the taxonomic impact of these coamplifications was small; all such individuals were identified as belonging to the *A. fulgerator* complex, and most individuals could be assigned to one of its 10-component taxa when the pseudogene sequence was determined. We emphasize, as well, that when sequencing is done with fresh specimens, the use of RT-PCR provides strong protection against Numt amplification (26), suggesting the use of this approach in taxa with COI pseudogenes.

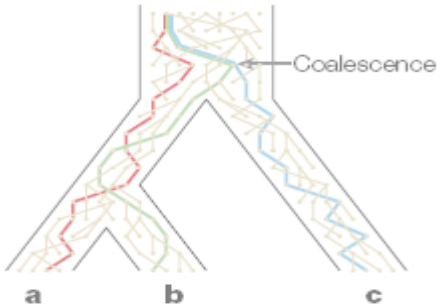
OK car copie récentes (i.e. pas tres divergente des séquences Mt) mais plus problématique si copies NUMTS anciennes…

1.Problèmes et limites du "Barcode ADN"

- Plus généralement, les principales sources d'erreurs (= la séquence focale est assignée à la mauvaise espèce) avec le Barcode ADN sont :

  1. L'espèce recherchée n'est pas représentée dans la base de données
  2. La séquence focale est plus proche génétiquement d'une séquence appartenant à une autre espèce à cause :
     a. D'un tri des lignées ancestrale (lineage sorting) "imparfait"
     b. du processus aléatoire de mutation

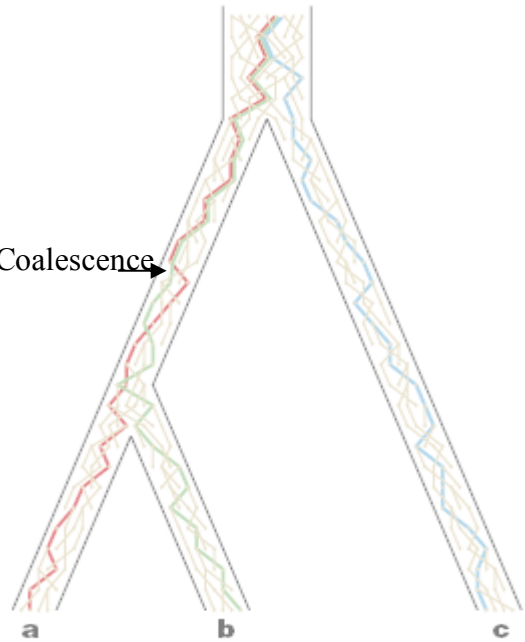
2. Arbres d'espèces et arbres de gènes = Tri des lignées ancestrales



Histoire généalogique d'un gène = processus stochastique de forte variance (coalescent)

-> dans quelle mesure l'arbre d'un gène et l'arbre des espèces sont concordants???

= tri des lignées ancestrales (lineage sorting)



2. Arbres d'espèces et arbres de gènes = Tri des lignées ancestrales

OPEN ACCESS Freely available online

PLoS GENETICS

Discordance of Species Trees with Their Most Likely Gene Trees

James H. Degnan[1], Noah A. Rosenberg[2]

1 Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of America, **2** Department of Human Genetics, Bioinformatics Program, and the Life Sciences Institute, University of Michigan, Ann Arbor, Michigan, United States of America

Because of the stochastic way in which lineages sort during speciation, gene trees may differ in topology from each other and from species trees. Surprisingly, assuming that genetic lineages follow a coalescent model of within-species evolution, we find that for any species tree topology with five or more species, there exist branch lengths for which gene tree discordance is so common that the most likely gene tree topology to evolve along the branches of a species tree differs from the species phylogeny. This counterintuitive result implies that in combining data on multiple loci, the straightforward procedure of using the most frequently observed gene tree topology as an estimate of the species tree topology can be asymptotically guaranteed to produce an incorrect estimate. We conclude with suggestions that can aid in overcoming this new obstacle to accurate genomic inference of species phylogenies.

Citation: Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genet 2(5): e68. DOI: 10.1371/journal.pgen.0020068

2. Arbres d'espèces et arbres de gènes = Tri des lignées ancestrales

Histoire généalogique d'un gène = processus stochastique de forte variance (coalescent)

= tri des lignées ancestrales (lineage sorting)

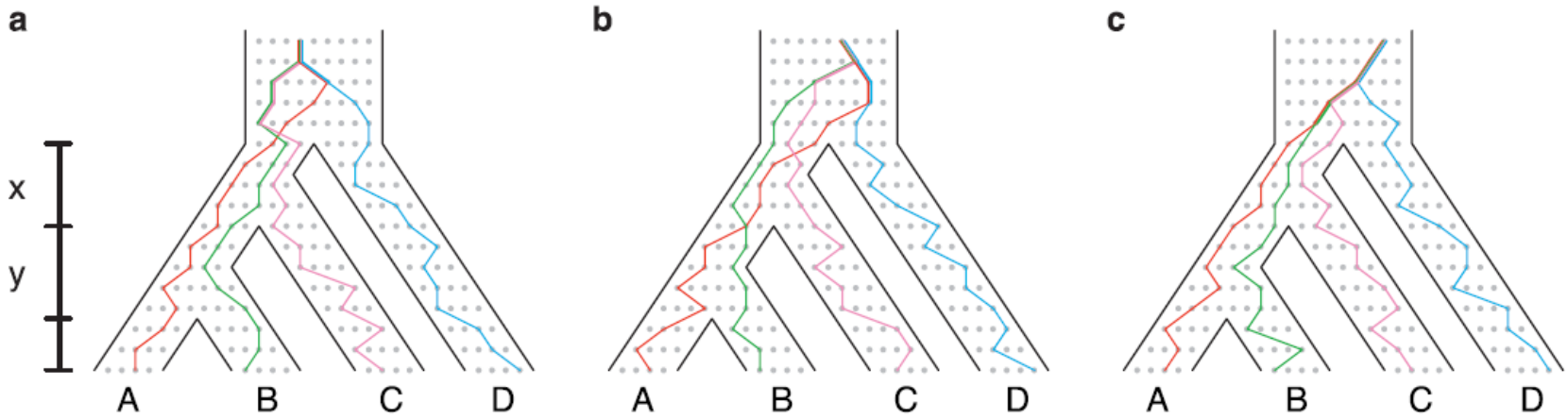


Figure 1. Anomalous Gene Trees for Four Taxa

Colored lines represent gene lineages that trace back to a common ancestor along the branches of a species tree with topology (((AB)C)D). The figure illustrates how a gene tree can have a higher probability of having a symmetric topology, in this case ((AD)(BC)), than of having the topology that matches the species tree. If the internal branches of the species tree—*x* and *y*—are short so that coalescences occur deep in the tree, the two sequences of coalescences that produce a given symmetric gene tree topology together have higher probability than the single sequence that produces the topology that matches the species tree.

(a) and (b) Two coalescence sequences leading to gene tree topology ((AD)(BC)). In (a), the lineages from B and C coalesce more recently than those from A and D, and in (b), the reverse is true.

(c) The single sequence of coalescences leading to gene tree topology (((AB)C)D).

DOI: 10.1371/journal.pgen.0020068.g001

2. Arbres d'espèces et arbres de gènes = Tri des lignées ancestrales

Histoire généalogique d'un gène = processus stochastique de forte variance (coalescent)

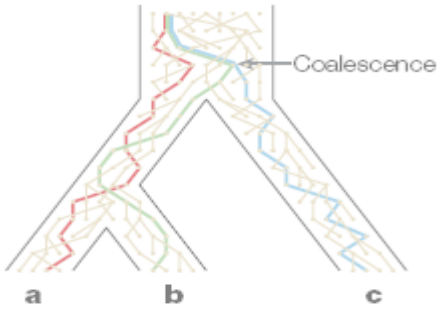
= tri des lignées ancestrales (lineage sorting)

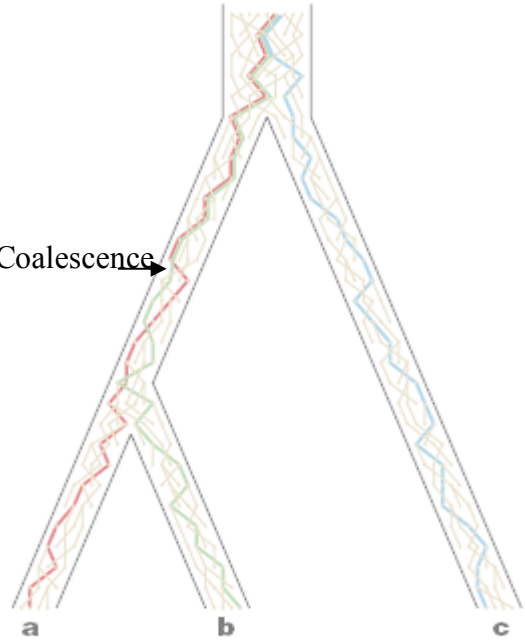
Mieux avec fort échantillonnage intra-spécifique :

> One strategy that may circumvent the occurrence of AGTs is the use of a sample with multiple individuals per species. Because many lineages from each species may persist reasonably far into the past, the chance of coalescences on a short branch is higher if many lineages are present [7,14,30]. Thus, increasing the sample size has a similar effect to lengthening short branches near the tips. As multiple sampled lineages from a species will coalesce on recent branches of the species tree, however, increased sample sizes will not assist the inference if recent branches are long but deep branches in the species tree are short.

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genet 2(5): e68. DOI: 10.1371/journal.pgen.0020068

2. Arbres d'espèces et arbres de gènes = Tri des lignées ancestrales





Histoire généalogique d'un gène = processus stochastique de forte variance (coalescent)

-> dans quelle mesure l'arbre d'un gène et l'arbre des espèces sont concordants???

= tri des lignées ancestrales (lineage sorting)

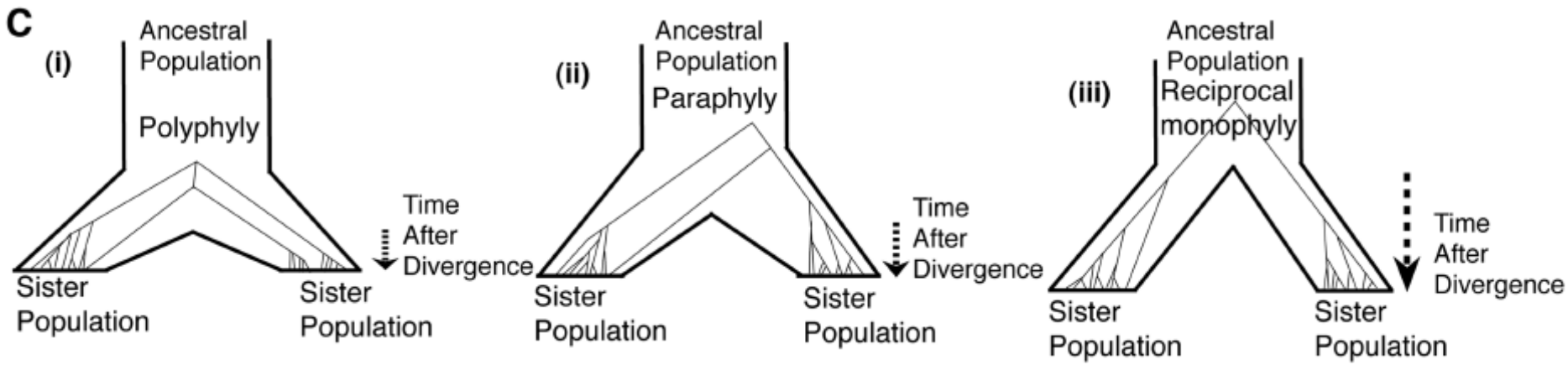
Concordants si :

- Temps spéciation >> temps de coalescence
- Pas/peu de flux de gènes pendant l'événement de spéciation
- Pas de transfert de gènes horizontaux

2.Arbres d'espèces et arbres de gènes ="phylogénie et coalescence"

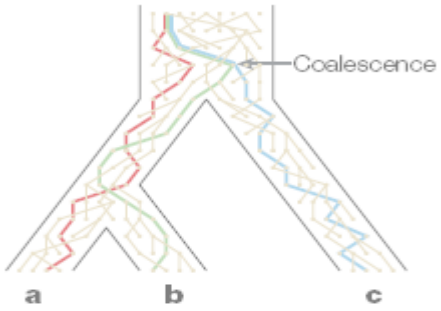
Tri des lignées ancestrales (lineage sorting)

Pour avoir une monophyly réciproque, il faut des temps de divergence long / temps de coalescence

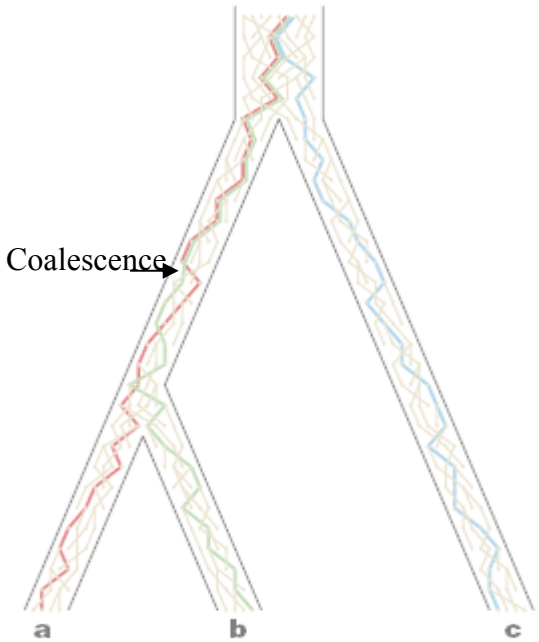


HICKERSON ET AL.—DNA BARCODING WILL FAIL TO DISCOVER ANIMAL SPECIES
Syst. Biol. 55(5):729–739, 2006

2.Arbres d'espèces et arbres de gènes



genetic theory. For simplicity, we will assume that there are two possible database species, a 'true' and a 'wrong' species. Assuming that both species are panmictic and of constant size, the probability that the query sequence does not share a most recent common ancestor with the true species, before either of them share a common ancestor with the wrong species, is simply $(2/3)e^{-T/N}$, where T is the divergence time between the two species in number of generations and N is the effective chromosomal population size of the true species (see, e.g., Hudson, 1990). When taking into account the mutational process,



Concordants si temps spéciation >> temps de coalescence

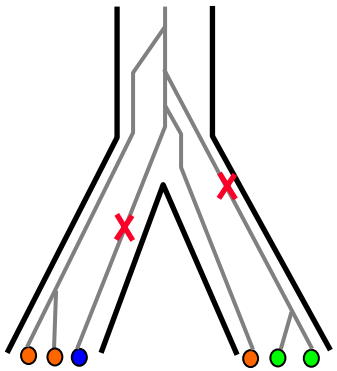
Fortement influencé
par facteurs populationnels,
Temps de coalescence courts si :
Petites tailles de populations
et/ou Sélection

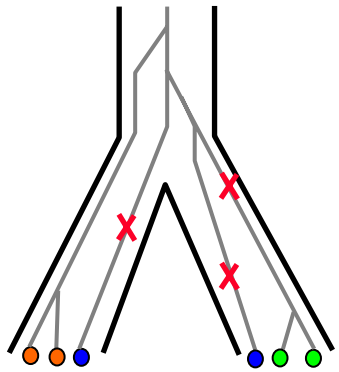
1.Problèmes et limites du "Barcode ADN"

- erreurs dues au processus aléatoire de mutation

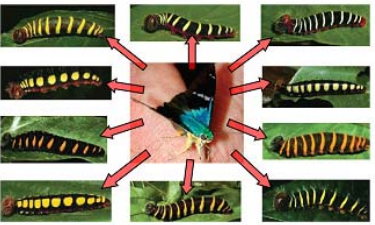
  notamment phénomène d'homoplasie : identité par état mais pas identité par descendance

Ex de cas problématiques





- Les problèmes dus à la mutation sont important quand le nombre de mutation est grand (e.g. fort θ =fort taux de mutation ou grande tailles de populations)



1.Problèmes et limites du "Barcode ADN"

Les cas problématiques seront donc :

- Divergence récente
- Grandes populations
- Diversité très faible ou très forte

Mais aussi :

- Hybridation/Introgression
- Populations fortement structurées dans l'espace

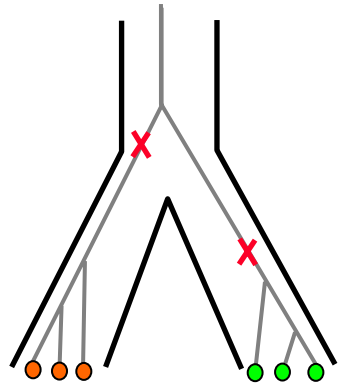
+ cas spécifiques : ex. dispersion femelles ≠ dispersion mâles

(cf exemple Astraptes)

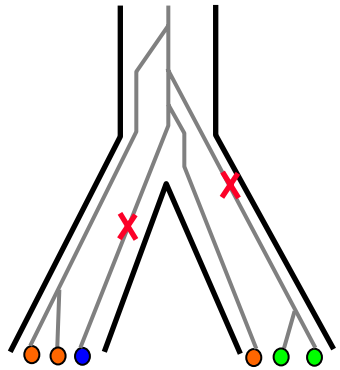
1.Problèmes et limites du "Barcode ADN"

- pas de gène "universel/ideal"
- problème de la diversité génétique intra-spécifique, même avec COI
- problème du tri des lignées ancestrales

-> nécessité de développer un cadre d'analyse rigoureux permettant de séparer variabilité intra et inter-spécifique et de préciser l'incertitude de chaque détermination plutôt que de se fier à l'appariement plus ou moins parfait des séquences Barcode (e.g. Blast)



Cas du gène idéal



Mais la réalité est plus complexe

1.Problèmes et limites du "Barcode ADN"

En plus de la nécessité de développer des outils d'analyse adaptés, il est potentiellement utile de considerer **des marqueurs supplémentaires et un échantillonnage plus poussé dans les cas problématiques :**

- Marqueurs plus variables (divergence récentes, hybridation, faible diversité) : séq. hypervariables, microsatellites
- Marqueurs nucléaires si possible recombinants (hybridation, forte diversité, dispersion sexe-biaisée)

  recombinaison = même espèce
- Echantillonnage d'un grand nombre d'individus (>10) sur plusieurs aires géographiques

3. Le projet "Barcode ADN" quelles méthodes d'analyse?

But = assignation d'une séquence à un groupe taxonomique génétiquement caractérisé

= distinguer variabilité intra- *vs.* inter-spécifique

4 grandes classes de méthodes :

1. Similitudes/Appariement de séquences
2. Classification statistiques
3. Phylogénétiques (Neighbors-Joining, Maximum Vraisemblance)
4. Modèle populationnels : Spéciation + Coalescence (Maximum Vraisemblance, Bayesien)

3. Le projet "Barcode ADN" quelles méthodes d'analyse?

1. Similitudes/Appariement de séquences : Blast, Google gene

L'espèce est donnée par la séquence de la base de données la plus proche de la séquence focale

+ rapidité

- pas de distinction variabilité intra-/inter-spé

- pas de signification/modèle biologique

- pas de résolution des ambigüités

Ca marche quand même dans >70% des cas testés!

Google Gene

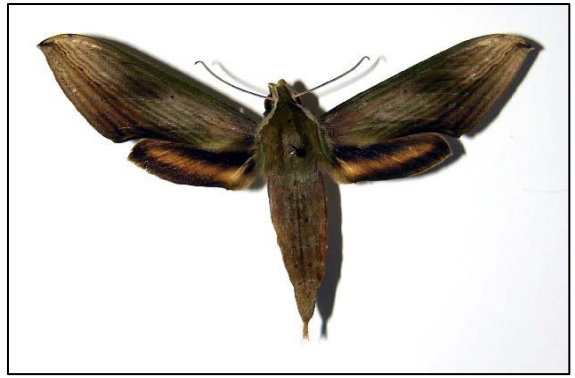
NNAACATTATATTTTATTTTTGGAATTTGAGCAGGAATAGTTGGAACCT
CACTAAGATTACTAATTCGAGCAGAA

GoogleGene Search | Clear

21 bases = 1 character (word)

```
>Query sequence [651 bases (31 characters) out of 660 original bases]
NNAACATTATATTTTATTTTT GGAATTTGAGCAGGAATAGTT GGAACCTCACTAAGATTACTA
ATTCGAGCAGAATTAGGAACC CCCGGATCTTTAATTGGAGAT GACCAAATTTATAACACAATT
GTTACAGCTCATGCATTTATT ATAATTTTTTTTATAGTAATA CCAATTATAATTGGAGGATTT
GGTAATTGATTAGTACCTTTA ATATTAGGAGCACCTGATATA GCATTCCCACGAATAAATAAC
ATAAGATTTTGACTTTTACCC CCTTCATTAACTCTTTTAATT TCTAGAAGTATTGTAGAAAAC
GGAGCAGGAACTGGTTGAACA GTTTACCCCCCTCTCTCTTCT AACATTGCTCATAGTGGAAGT
TCTGTAGATTTAGCTATTTTT TCCCTTCATTTAGCTGGTATT TCTTCAATTATAGGAGCTGTA
AATTTTATTACTACTATTATT AATATGCGAATTAATAATTTA TCATTTGATCAAATACCATTA
TTTGTTTGAGCTGTTGGAATC ACAGCCTTTTTATTATTACTA TCTTTACCAGTATTAGCTGGT
GCAATTACAATATTATTAACT GATCGAAATCTTAATACATCA TTTTTTGACCCTGCTGGAGGG
GGAGACCCTATTCTATATCAA
```

Sequences matching your query:



```
>MHASC043 05 02 SRNP 19434 Xylophanes Janzen01 [31 out of 31 characters match]
NNAACATTATATTTTATTTTT GGAATTTGAGCAGGAATAGTT GGAACCTCACTAAGATTACTA
ATTCGAGCAGAATTAGGAACC CCCGGATCTTTAATTGGAGAT GACCAAATTTATAACACAATT
GTTACAGCTCATGCATTTATT ATAATTTTTTTTATAGTAATA CCAATTATAATTGGAGGATTT
GGTAATTGATTAGTACCTTTA ATATTAGGAGCACCTGATATA GCATTCCCACGAATAAATAAC
ATAAGATTTTGACTTTTACCC CCTTCATTAACTCTTTTAATT TCTAGAAGTATTGTAGAAAAC
GGAGCAGGAACTGGTTGAACA GTTTACCCCCCTCTCTCTTCT AACATTGCTCATAGTGGAAGT
TCTGTAGATTTAGCTATTTTT TCCCTTCATTTAGCTGGTATT TCTTCAATTATAGGAGCTGTA
AATTTTATTACTACTATTATT AATATGCGAATTAATAATTTA TCATTTGATCAAATACCATTA
TTTGTTTGAGCTGTTGGAATC ACAGCCTTTTTATTATTACTA TCTTTACCAGTATTAGCTGGT
GCAATTACAATATTATTAACT GATCGAAATCTTAATACATCA TTTTTTGACCCTGCTGGAGGG
GGAGACCCTATTCTATATCAA CATTTATTT
```

```
>MHASC038 05 02 SRNP 18238 Xylophanes Janzen01 [30 out of 31 characters match]
NNAACATTATATTTTATTTTT GGAATTTGAGCAGGAATAGTT GGAACCTCACTAAGATTACTA
ATTCGAGCAGAATTAGGAACC CCCGGATCTTTAATTGGAGAT GACCAAATTTATAACACAATT
GTTACAGCTCATGCATTTATT ATAATTTTTTTTATAGTAATA CCAATTATAATTGGAGGATTT
GGTAATTGATTAGTACCTTTA ATATTAGGAGCACCTGATATA GCATTCCCACGAATAAATAAC
ATAAGATTTTGACTTTTACCC CCTTCATTAACTCTTTTAATT TCTAGAAGTATTGTAGAAAAC
GGAGCAGGAACTGGTTGAACA GTTTACCCCCCTCTCTCTTCT AACATTGCTCATAGTGGAAGT
TCTGTAGATTTAGCTATTTTT TCCCTTCATTTAGCTGGTATT TCTTCAATTATAGGAGCTGTA
AATTTTATTACTACTATTATT AATATGCGAATTAATAATTTA TCATTTGATCAAATACCATTA
TTTGTTTGAGCTGTTGGAATC ACAGCCTTTTTATTATTACTA TCTTTACCAGTATTAGCTGGT
GCAATTACAATATTATTAACT GATCGAAATCTTAATACATCA TTTTTTGACCCTGCTGGAGGA
GGAGACCCTATTCTATATCAA CATTTATTT
```

3. Le projet "Barcode ADN" quelles méthodes d'analyse?

2. Modèles statistiques de classification (cf. cours FRED AUSTERLITZ)

Etape1 : Recherche de critères de classification (= les nucléotides les plus discriminants ou des distances génétiques) pour former des groupe homogènes (= les espèces) à partir de sur toutes les séquences de la base de données

Etape2 : Classer la séquence focale selon les mêmes critères dans un des groupes

+ meilleure classification

+ incertitude possible

- parfois lent

- pas de signification/modèle biologique

3. Le projet "Barcode ADN" quelles méthodes d'analyse?

3. Phylogénétiques (neighbor joining, maximum vraisemblance)

+ relativement rapide
+ prendre en compte des modèles d'évolution moléculaire
+ cadre d'analyse de données génétique connu
+ marche mieux que les autres quand polymorphisme fort
- Marche moins bien que les autres quand polymorphisme faible
- Pas de critère d'incertitude (juste ambigüités)

Marche mieux que les méthodes précédentes quand forte variabilité intra-spécifique

3. Le projet "Barcode ADN" quelles méthodes d'analyse?

Ce qui est fait dans la majorité des études de Barcode ADN = l'outils par défaut de BOLD

- Recherche par similarité (BLAST) des 100 séquences les plus proches
- Construction d'un arbre Neighbour Joining avec ces 100 séquences
- La séquence focale est assignée à l'espèce la plus proche
- Critère = distance génétique entre les séquences (ex: K2P)

3. Le projet "Barcode ADN" quelles méthodes d'analyse?

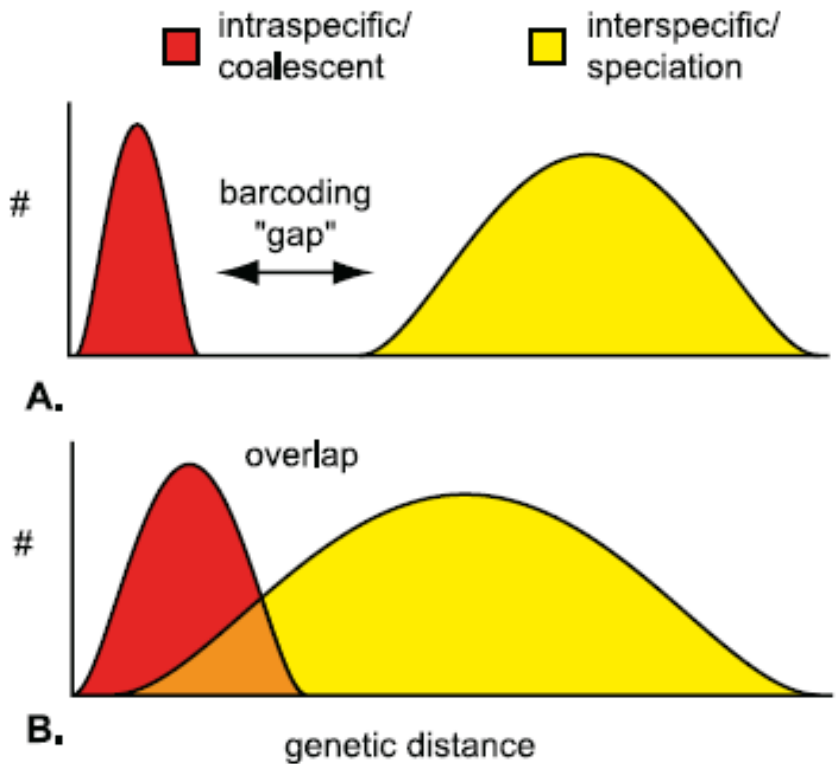
Distance génétiques intra- et inter-spécifiques

- Critère = distance génétique entre les séquences (ex: K2P)

  - ✓ Idée de seuil : globalement 2-3% semble être le maximum pour des distances intra-spécifiques (3% rule), ou la distance entre 2 espèces doit être plus de 10x supérieure à la distance intra-spécifique (10x rule)

  - ✓ Comparaison des distances intra- et interspécifique dans la base de données par rapport à la distance entre la séquence focale et l'espèce la plus proche

Distance génétiques intra- et inter-spécifiques : le Barcoding gap

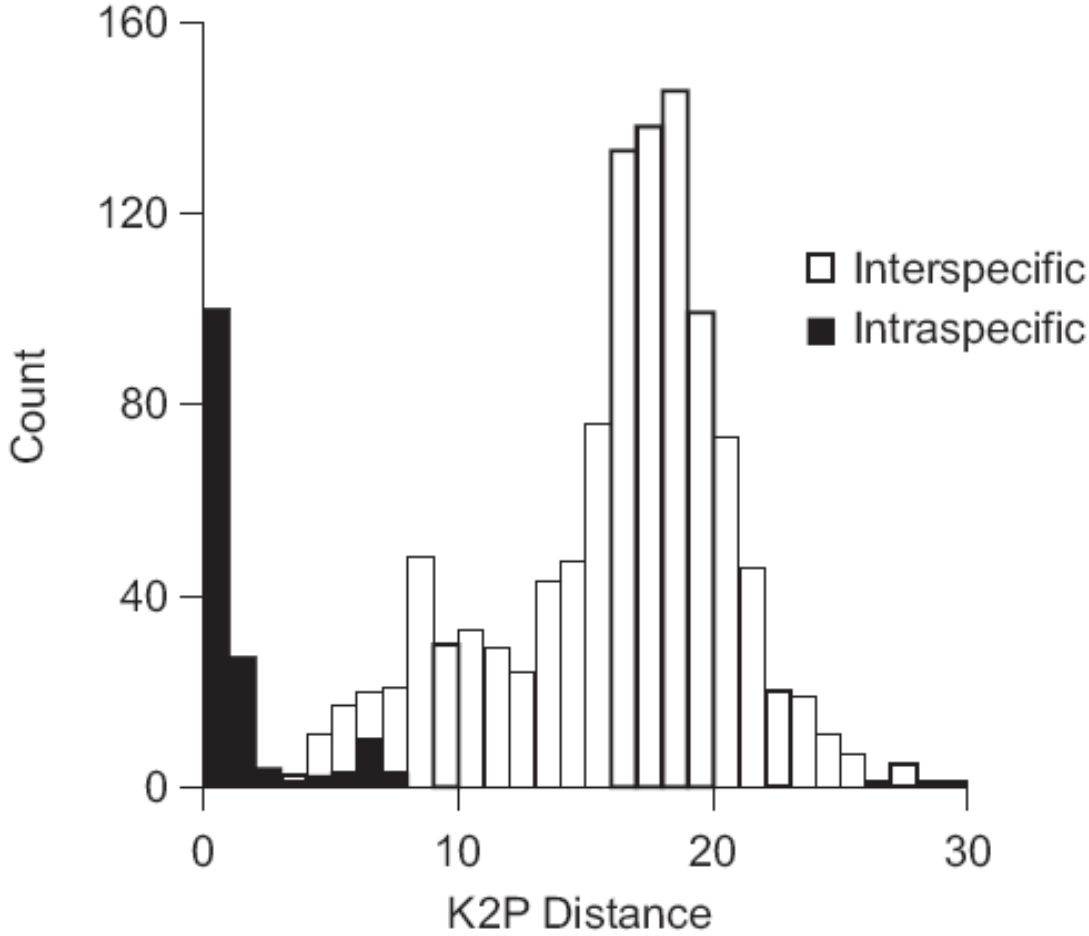


Distance génétiques intra- et inter-spécifiques

exemple chez les arachnides

Barrett & Hebert 2005 Can. Journ. Of Zool.

Fig. 3. Histogram of intraspecific and interspecific (congeneric) genetic divergence across 203 arachnid species. Divergences were calculated using Kimura's two parameter (K2P) model.



Distance génétiques intra- et inter-spécifiques

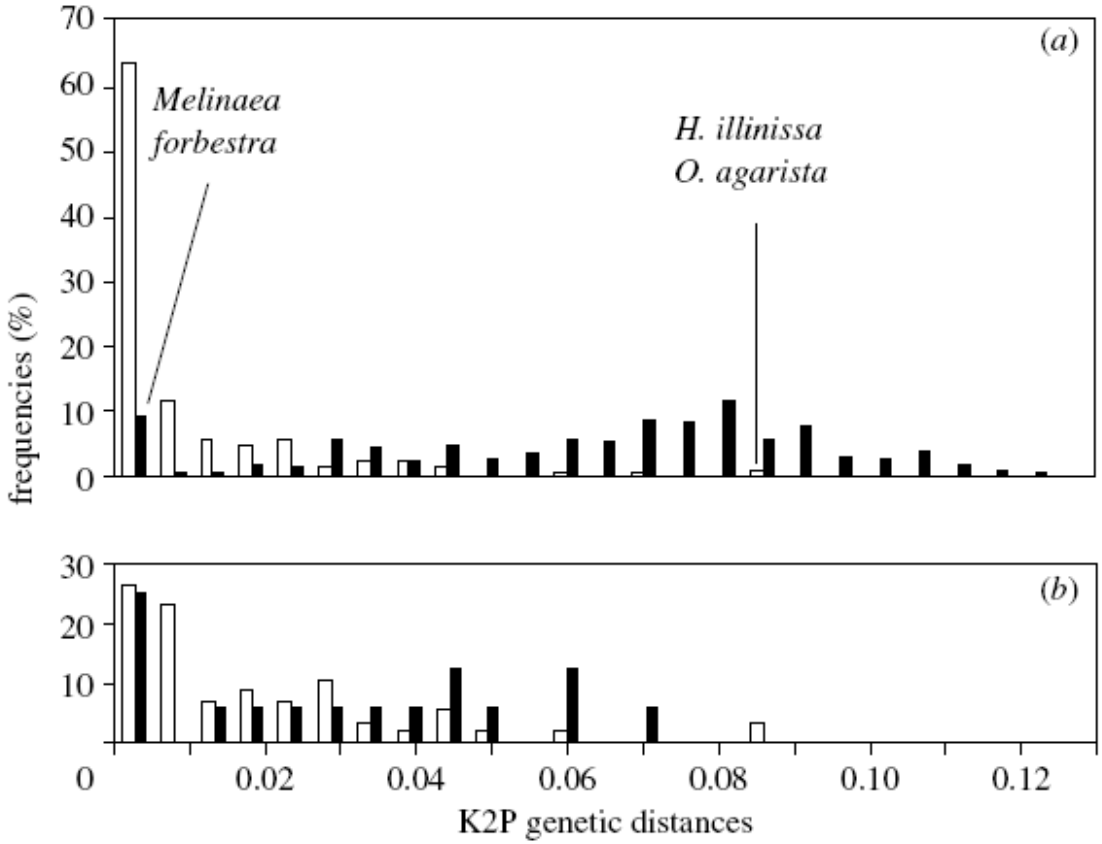
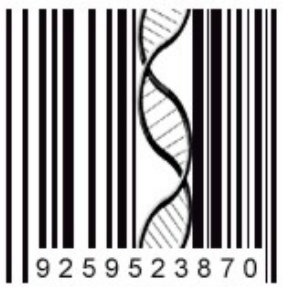


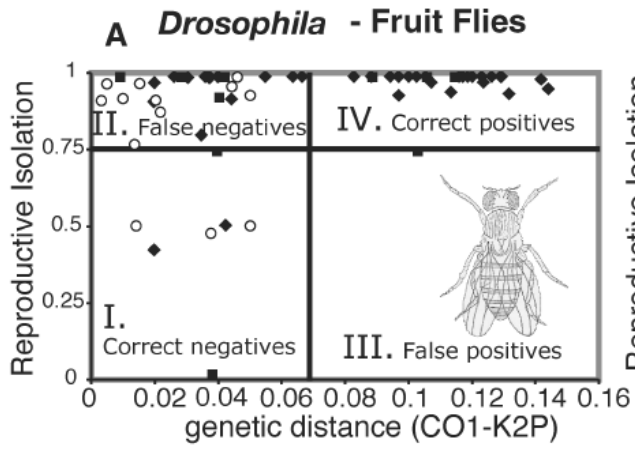
Figure 2. Distribution of within-species (white) and between congeneric species (black) K2P distances of barcode sequences. Only species and genera represented locally are considered. (*a*) All pairwise distances and (*b*) only maximum intraspecific and minimum interspecific distances.

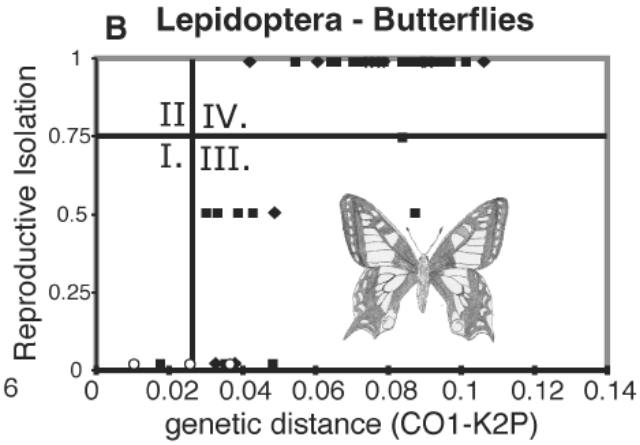


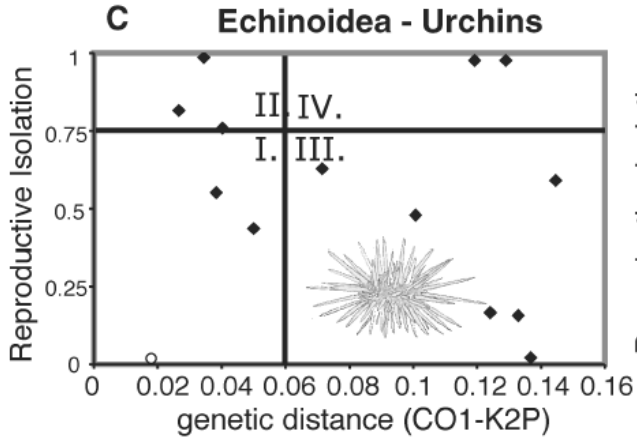
Distances génétique et isolement reproducteur

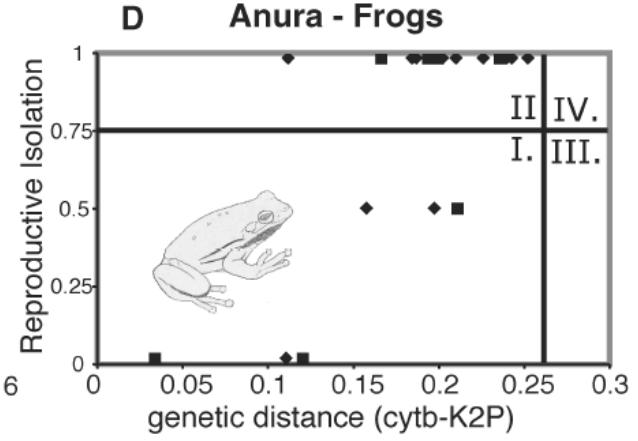
Hickerson et Al. 2006 Syst. Biol.

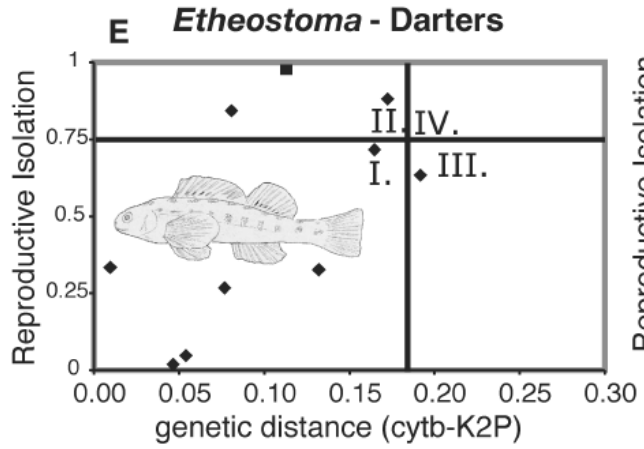
(10x rule)

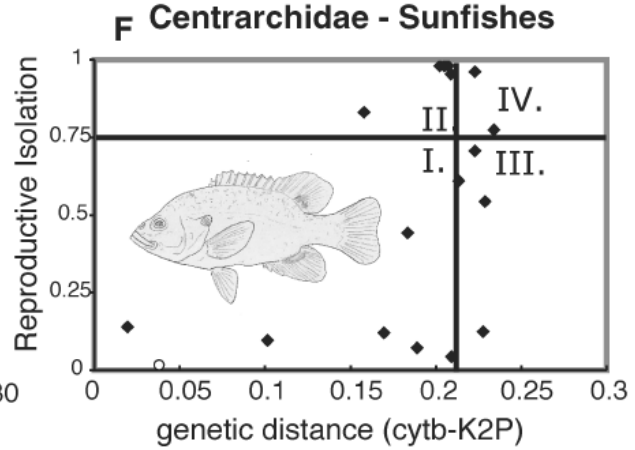












Distance génétiques intra- et inter-spécifiques

Critère = distance génétique entre les séquences (ex: K2P)

- ✓ Idée de seuil : globalement 2-3% semble être le maximum pour des distances intra-spécifiques , ou la distance entre 2 espèces doit être plus de 10x supérieure à la distance intra-spécifique
- ✓ Comparaison de les distances intra- et interspécifique dans la base de données par rapport à la distance entre la séquence focale et l'espèce la plus proche

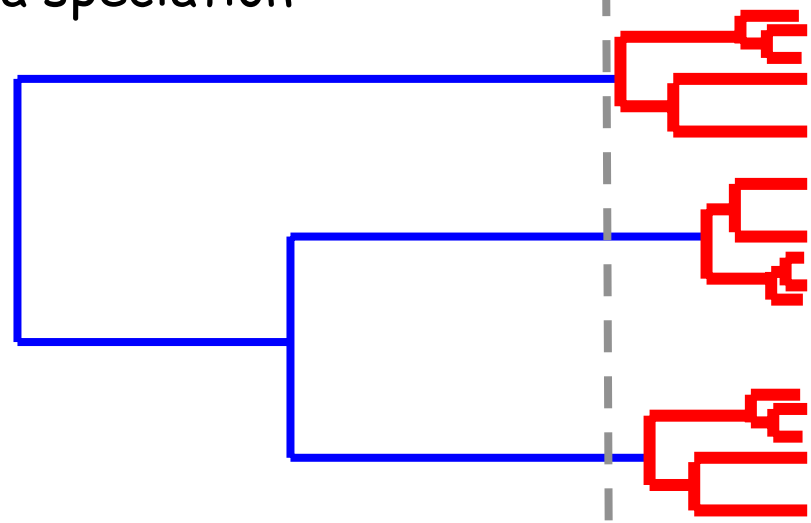
Pas de bons critère car aucune justification scientifique sous jacente : les distance génétiques ne sont pas reliées à l'isolement reproducteur

Ce sont uniquement des observations empiriques qui ne sont pas toujours vrai

4. Extension des outils de génétique des populations

Modèle populationnels : Spéciation + Coalescence

Extension des approches de génétique des populations pour prendre en compte la spéciation



Branchements inter-espèce

Modèles de spéciation :

Taux de spéciation, d'extinction

Branchements intra-espèce

Théorie de la coalescence :

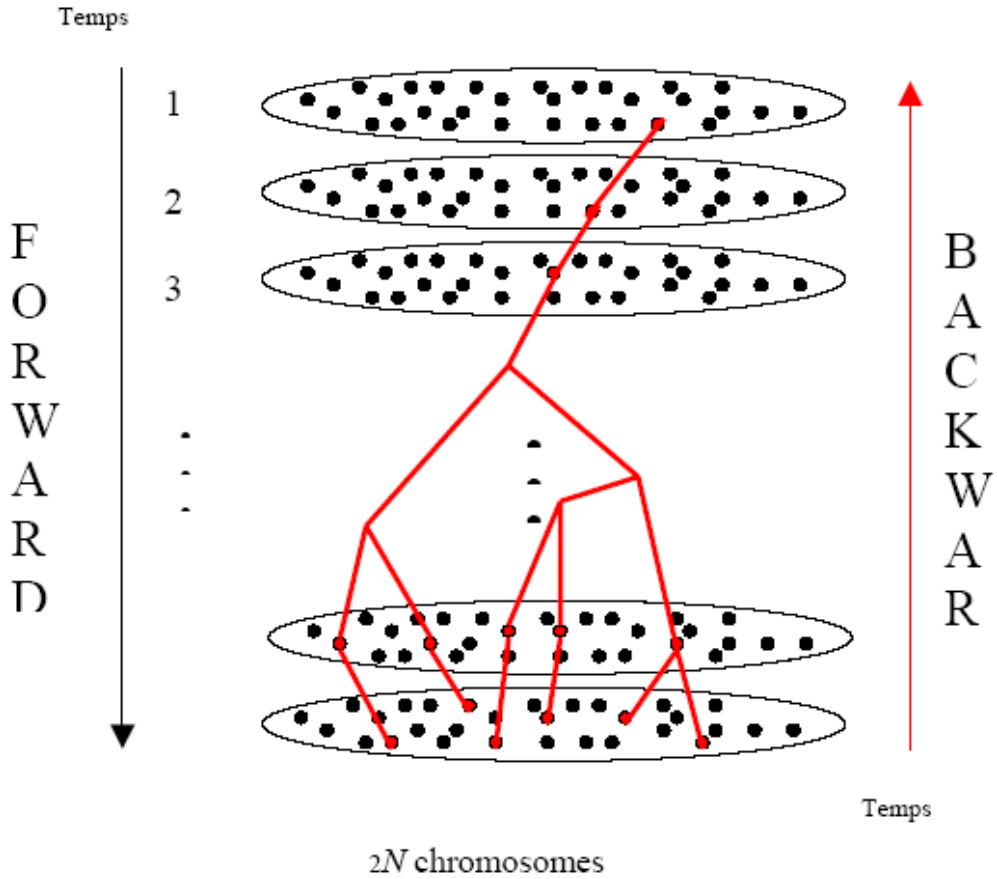
Taille de pops, flux de gènes,

Histoire démographique et selective

4. Extension des outils de génétique des populations

- Rappel des principales sources d'erreurs :
  1. L'espèce recherchée absente de la base de données
  2. Tri des lignées ancestrale "imparfait"
  3. Processus aléatoire de mutation
- La génétique des population, et notamment la théorie de la coalescence, permet de calculer les probabilités des 2 derniers facteurs et pourrait aussi permettre de calculer la probabilité d'appartenance à une espèce

Théorie de la coalescence
et
estimations de paramètres démographiques



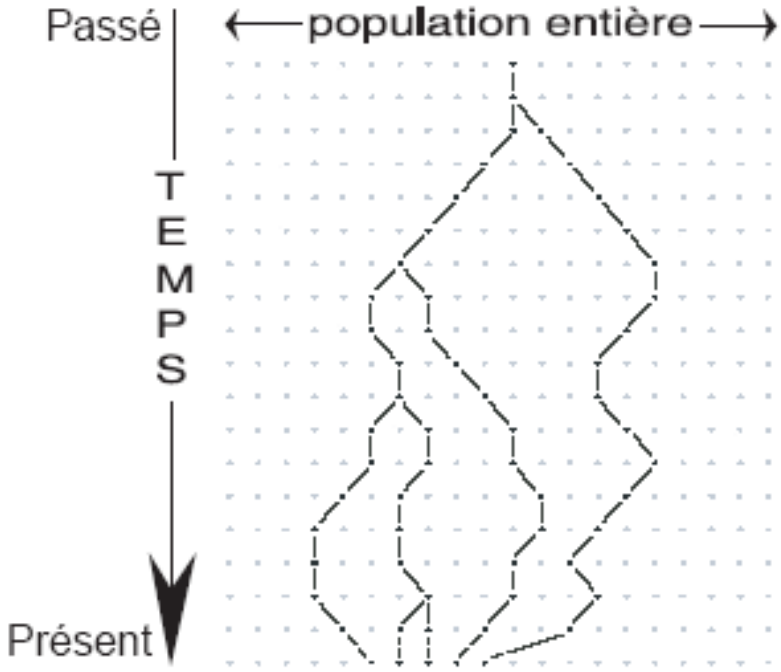
Origine de la théorie de la coalescence :

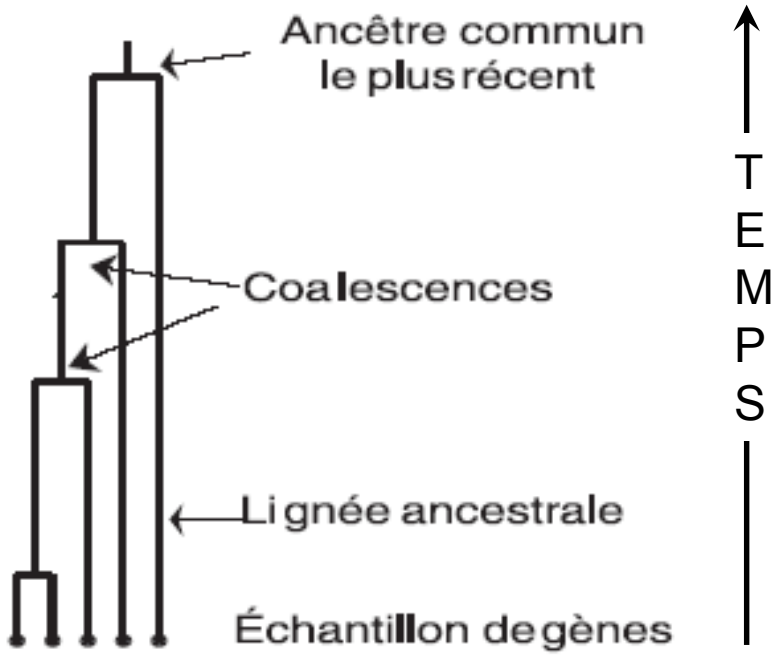
- 1974 –1982 gestation (Kingman, Ewens, Watterson)

- 1982 Kingman & Tajima

- depuis 1990, nombreux développements

  par Griffiths, Tavaré, Hudson, Donnelly, Felsenstein, Nielsen, Hey, Wakeley et beaucoup d'autres…





→ **Nouvelle approche de génétique des populations théorique**

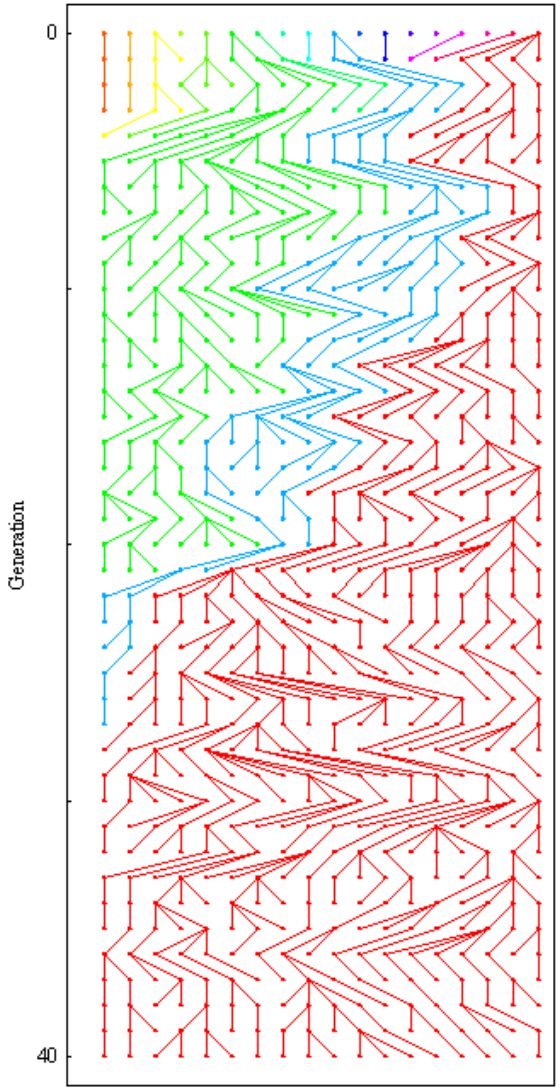
Approche classique

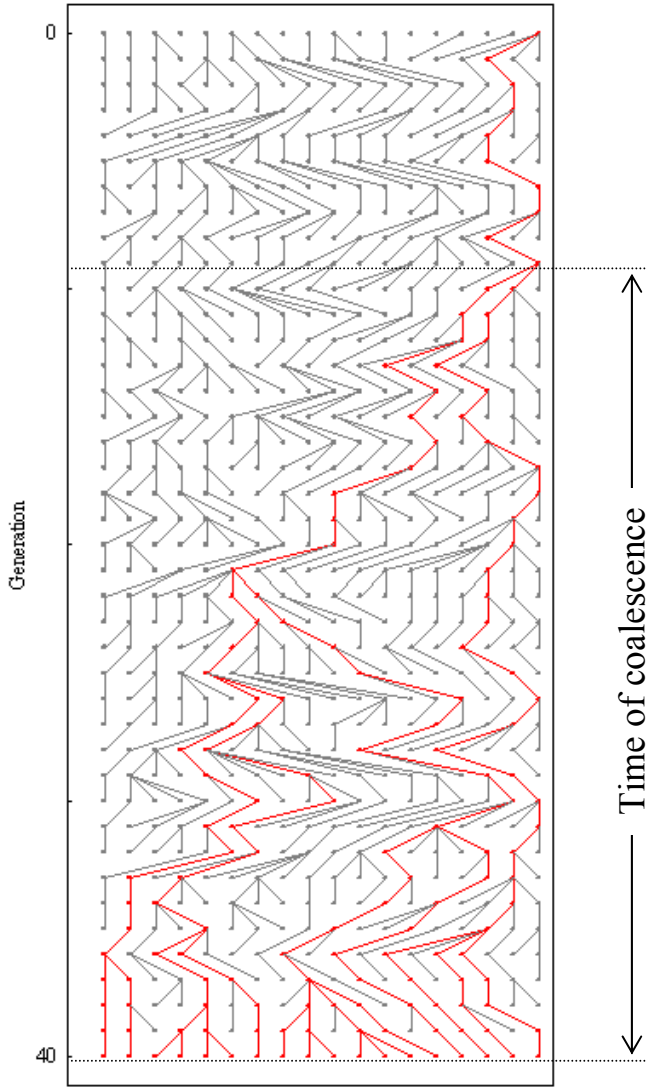
- POPULATION
- Fréquences
- Vision avant (Forward)

Approche « coalescence »

- ECHANTILLON
- Généalogie des gènes
- Vision arriere (Backward)

Another way of looking at Genetic drift: *the coalescence theory*





Les différentes lignées fusionnent (coalescent) au fur et à mesure que l'on remonte vers le passé

La coalescence : la dérive vue en remontant le temps

Principaux avantages de la coalescence

- Simplification de l'analyse quantitative des modèles stochastiques et réinterprétation des résultats théoriques

- La structure des données génétiques reflète pour une large part la généalogie sous-jacente → meilleure compréhension de la variabilité génétique observée

- **Méthodes de simulation extrêmement efficaces**

- **Elle fournit des techniques d'inférences sur les données génétiques qui permettent l'usage complet de l'information**

coalescence et estimation de paramètres évolutifs

- Par maximum de vraisemblance (**ML**) ou Bayesien

Utilise des algorithmes spécifiques, fondés sur la coalescence, pour calculer la vraisemblance ou la probabilité postérieure d'un échantillon

→ **utilise toute l'information des données**

- Approximate Bayesian Computation (ABC)

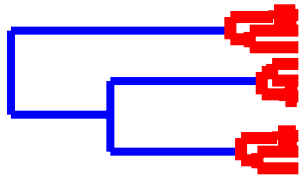
Utilise les algorithmes de simulation de données décrit précédemment et des statistiques résumées (ex: nA, He, Fst, VarAll…) pour approcher la vraisemblance d'un échantillon

→ **résume l'information des données en plusieurs statistiques**

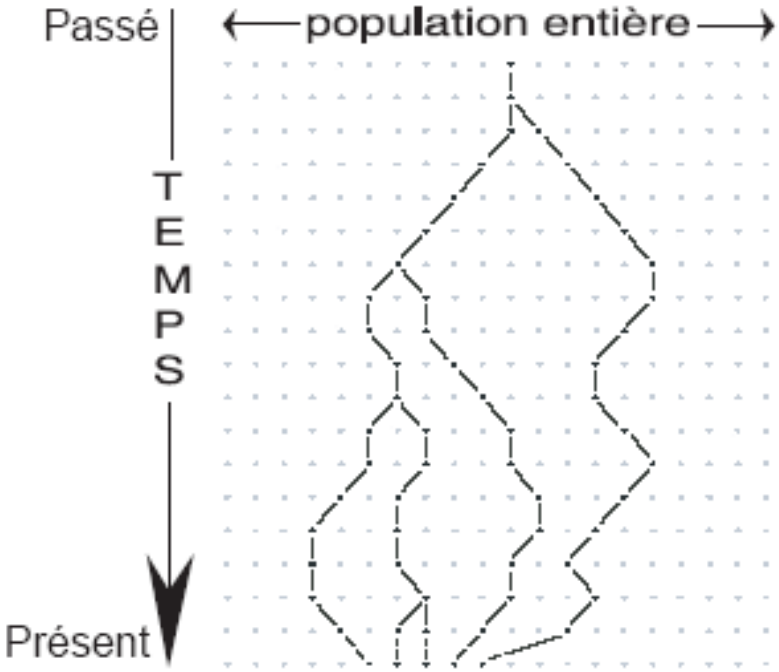
Il existe 2 approches principales pour le calcul de la vraisemblance ou de la probabilité postérieure d'un échantillon sachant les valeurs de paramètres :

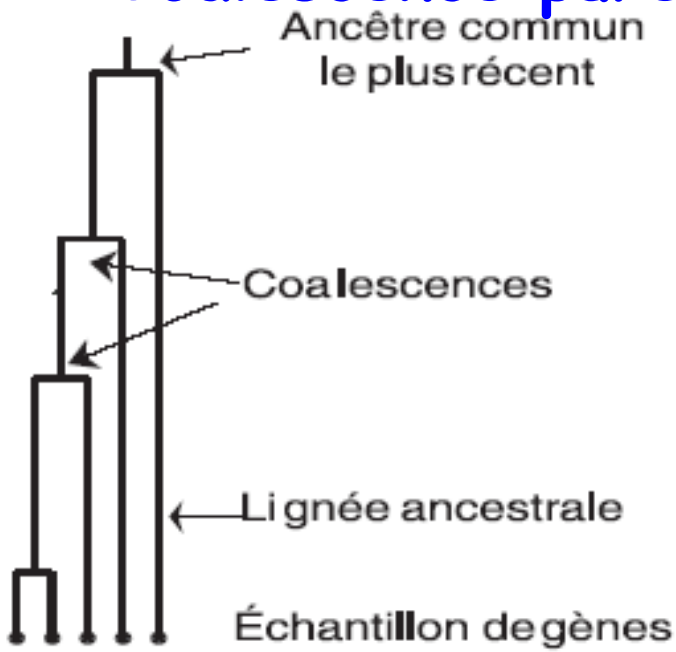
- L'approche de ***Griffiths et coll.*** (**GENETREE**) utilisant des chaînes de Markov absorbantes et de l'*Importance Sampling (IS)*
- L'approche de ***Felsenstein et coll.*** **(FLUCTUATE, MIGRATE)** utilisant un algorithme de *Monte Carlo par Chaînes de Markov (MCMC)*
  Utilisée par R. Nielsen & Jody Hey dans les logiciels **MDiv, IM, IMa**

Différences = échantillonnage des généalogies (IS/MCMC)



4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure





Sous la forme la plus simple il y a 1 paramètres par population :

$\Theta = 4 N_e \mu$

Autres intérêts de la coalescence :

- Certaine formes de structuration géographique et de fluctuations démographiques peuvent être prise en compte

4. Extension des outils de génétique des populations

Généralités sur l'utilisation de la coalescence pour le Barcode

+ Cadre statistique optimal (max vraisemblance)
  -> utilisation de toute l'info des données, incertitude, test hyp....
+ Cadre theorique très développé, bien connu et **puissant**
+ Signification biologique forte, modèles populationnels réalistes
+ Inférence biologiques possibles (limitées içi car 1 seul ou peu de gènes)

- Fondées sur marqueurs neutres (mais certaines selections +- OK)
- Développée pour multilocus, ici un seul ou peu de gènes
  -> moindre puissance
- Demande des moyen de calcul importants si modèles complexes

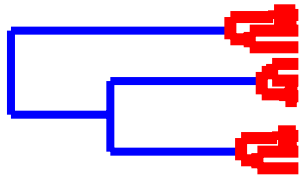
4. Extension des outils de génétique des populations

Généralités sur l'utilisation de la coalescence pour le Barcode

- Devrait marcher quand les autres méthodes ne marche pas

- Nécessite une bonne description de la variabilité intra-spécifique
  (-> échantillonnage intra-spécifique important)

- Assignation lente mais la plus précise : Utilisation pour l'étape finale d'assignation à une espèce quand elle est problématique



4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

Utilisation de la coalescence sans hypothèse de spéciation

On s'interesse uniquement à ce qui se passe au niveau intra-spécifique

Probabilité que la séquence focale *x* appartienne à l'espèce *i* sachant les données *D* (i.e. les séquences de l'espèce *i*):
$Pr(x \in i/D)$

La théorie de la coalescence permet de calculer cette probabilité en faisant 2 hypothèses fortes :

1- l'espèce est considére comme une population avec comme seul paramètre populationnel $\theta=4Ne\mu$

2- on connaît θ (ou on l'estime)

On s'interesse donc à $Pr(x \in i/D, \theta) = P(x+D/\theta)$

4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

1er exemple : The Coalescent Assigner (disponible sur le web)

Syst. Biol. 56(1):44–56, 2007

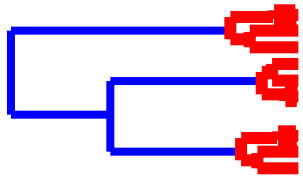
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150601167005

A Step Toward Barcoding Life: A Model-Based, Decision-Theoretic Method to Assign Genes to Preexisting Species Groups

ZAID ABDO[1] AND G. BRIAN GOLDING[2]

[1]*Department of Mathematics and Department of Statistics, University of Idaho, Moscow, Idaho 83844, USA*
[2]*Department of Biology, McMaster University, Life Science Building, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada; E-mail: golding@mcmaster.ca (G.B.G.)*

Abstract.—A major part of the barcoding of life problem is assigning newly sequenced or sampled individuals to existing groups that are preidentified externally (by a taxonomist, for example). This problem involves evaluating the statistical evidence towards associating a sequence from a new individual with one group or another. The main concern of our current research is to perform this task in a fast and accurate manner. To accomplish this we have developed a model-based, decision-theoretic framework based on the coalescent theory. Under this framework, we utilized both distance and the posterior probability of a group, given the sequences from members of this group and the sequence from a newly sampled individual to assign this new individual. We believe that this approach makes efficient use of the available information in the data. Our preliminary results indicated that this approach is more accurate than using a simple measure of distance for assignment. [Assignment; barcoding; coalescent; decision theory.]



4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

Syst. Biol. 56(1):44–56, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150601167005

A Step Toward Barcoding Life: A Model-Based, Decision-Theoretic Method to Assign Genes to Preexisting Species Groups

ZAID ABDO[1] AND G. BRIAN GOLDING[2]

Ils utilisent un programme de coalescence existant (FLUCTUATE, Felsenstein et al.) pour estimer le Θ de toutes les espèces candidates et calculer la vraisemblance que la séquence focale appartienne à chacune de ces espèces $L(x \in i/D, \theta) = P(x+D/\theta)$

4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

Quelques résultats encourageants :

Coalescent assigner

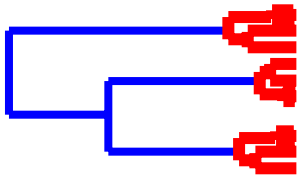
$n = 5$	$\nu \setminus \theta$	0.001	0.01	0.1
Divergence	0.001	93%	89%	82%
	0.01	99%	97%	92%
	0.1	100%	100%	99%
	1	100%	100%	100%

Diversité genetique

distance method

$n = 5$	$\nu \setminus \theta$	0.001	0.01	0.1
	0.001	92%	61%	56%
	0.01	99%	93%	60%
	0.1	100%	100%	92%
	1	100%	100%	99%

Abdo & Golding, 2007, Systematic Biology



Coalescent assigner

globallement meilleur que

Distance method

sauf pour des grandes tailles d'échantillon

non intuitif…

pb de configuration de FLUCTUATE?? Pb de temps de calculs trop courts?? Oui chaine de markov trop courte meilleur résultats avec des chaines plus longues

Abdo & Golding, 2007, Systematic Biology

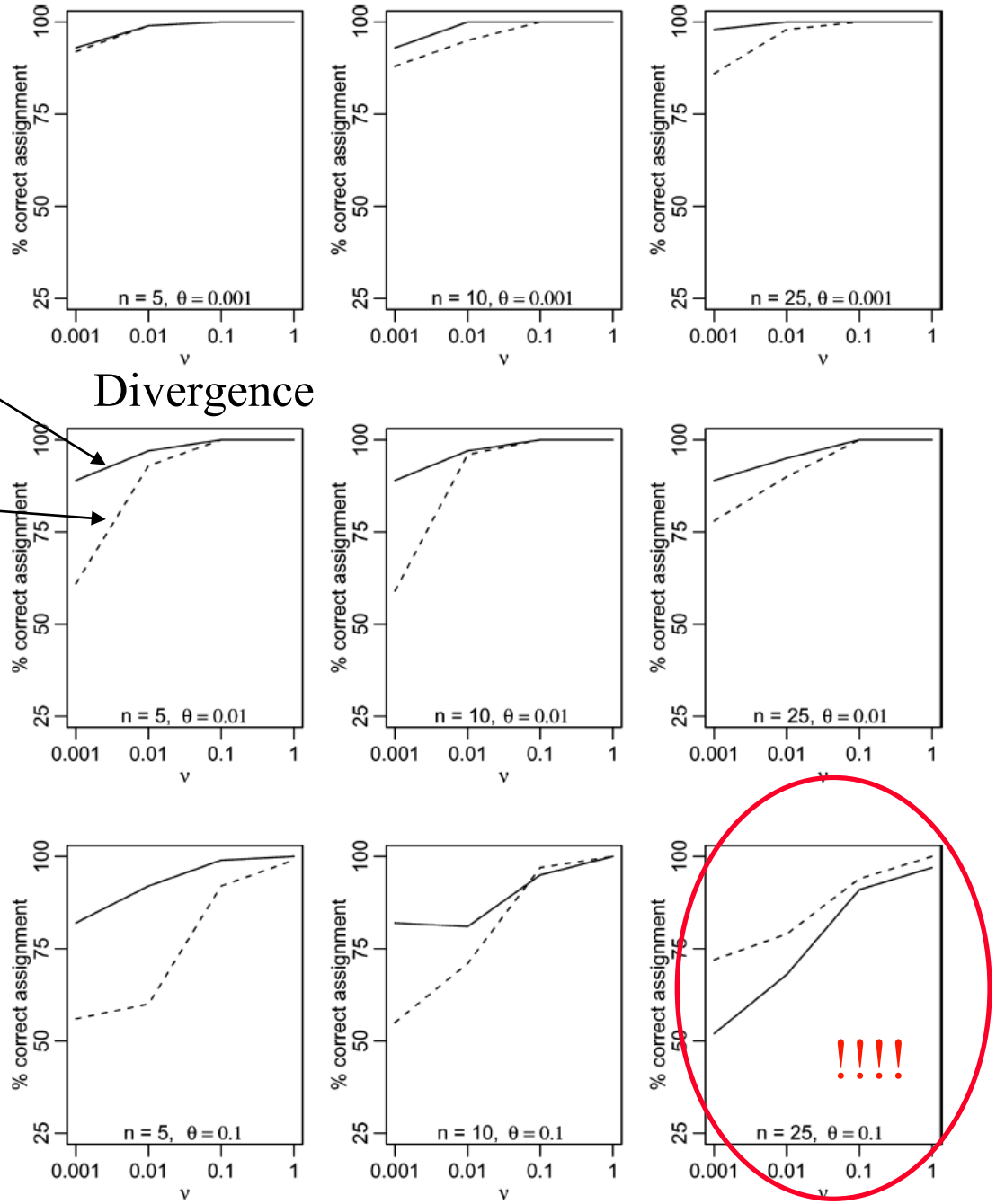


FIGURE 2. Power of the coalescent assigner (solid line) compared to that of the distance method (dotted line) as ν increases at different levels of θ and different sample sizes.

4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

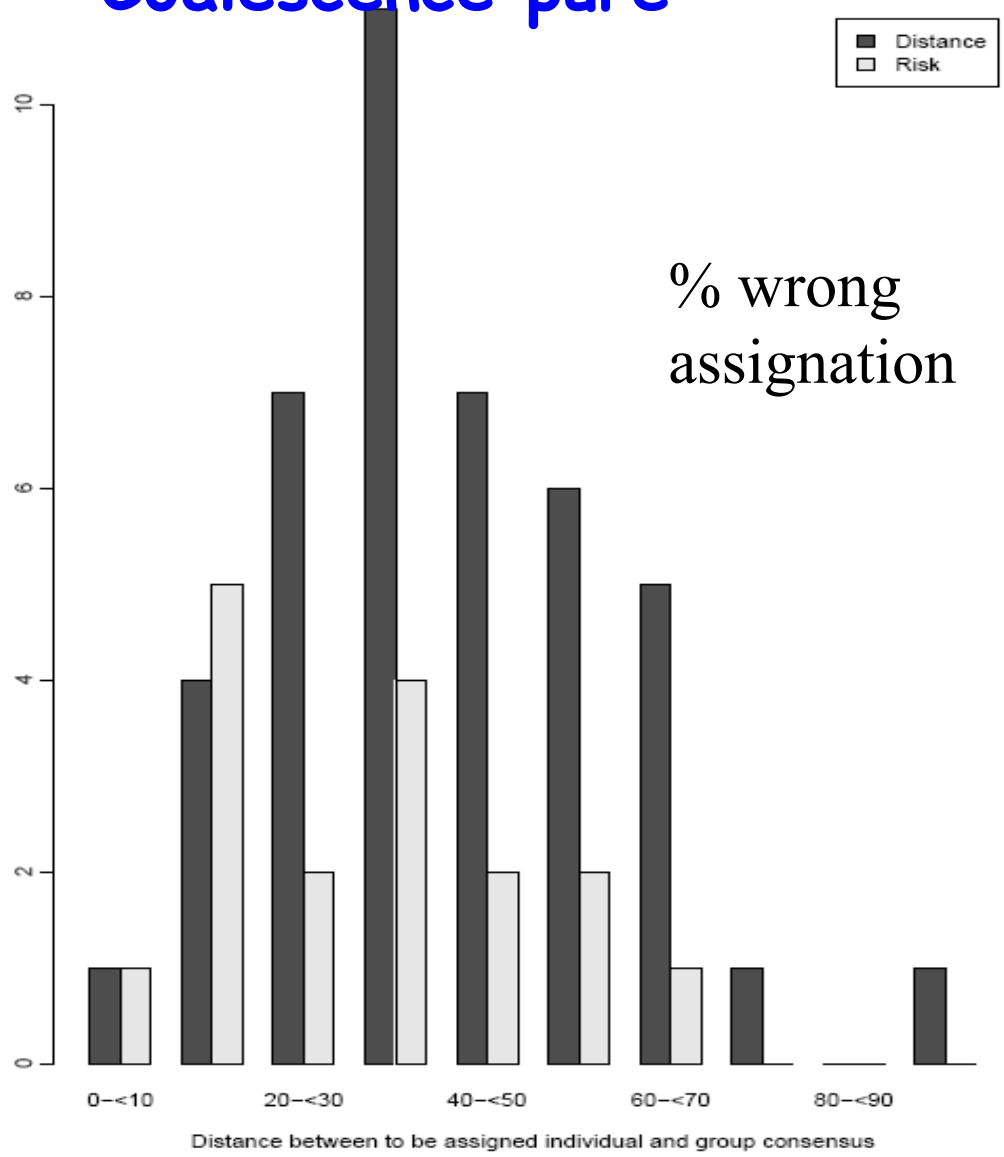
Coalescent assigner (= risk)

se trompe moins souvent que

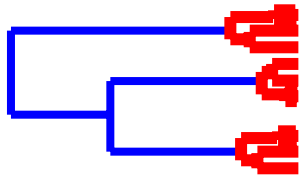
Distance method

quand la distance entre l'individu focal et son espèce est grande

Abdo & Golding, 2007, Systematic Biology



% wrong assignation



4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

Article général sur l'apport de la coalescence dans le contexte du Barcode ADN

Syst. Biol. 55(1):162–169, 2006

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500431239

Statistical Approaches for DNA Barcoding

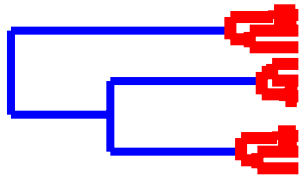
RASMUS NIELSEN[1] AND MIKHAIL MATZ[2]

[1]*Department of Biological Statistics and Computational Biology, Center for Bioinformatics, University of Copenhagen Universitetsparken 15, 2100 Copenhagen, Denmark; E-mail: rn28@cornell.edu*

[2]*Whitney Laboratory and Department of Molecular Genetics and Microbiology, University of Florida, 9505 Ocean Shore Blvd, Saint Augustine, FL 32080, USA*

The weakest spot of DNA barcoding is the obvious fact that no gene can serve as an ideal barcode, i.e., be always invariant within species but different among species. It has been pointed out by several authors that DNA-based identification, if it is to become a rigorous analysis, should be concerned about distinguishing intraspecific from interspecific variation rather than simply recording perfect and imperfect sequence matches (Dunn, 2003; Lipscomb et al., 2003; Stoeckle, 2003). To get around this problem, at present it is assumed that the interspecific sequence variation should exceed a certain threshold, say, 2% or 3% dissimilarity, set on the basis of empirical observations of sequence differences among congeneric species (Hebert et al., 2003a, 2003b). Such an approach seems to be too simplistic to avoid inconclusive or erroneous results. Clearly, there is a need for statistical methods for assessing if a sampled query sequence is sufficiently similar to a particular data base sequence to justify a species assignment of the query.

There are several possible statistical approaches to this problem. In the classical hypothesis-based (frequentist) approach, the null hypothesis could be that the query sequence is a member of a particular species. Such a test would control the rate at which the query is assigned to the true species, i.e., it controls the rate of false negatives. In the Bayesian approach the posterior probability of a species assignment is calculated by assuming a prior distribution of species assignments. We will discuss the problems and merits of these approaches and provide some guidelines towards the use of statistics in DNA barcoding experiments.



4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

2ème exemple de méthode d'assignation Barcode utilisant la coalescence : Utilise un programme de coalescence adhoc par MCMC sous un model de divergence simple

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2005) 360, 1969–1974
doi:10.1098/rstb.2005.1728
Published online 14 September 2005

A likelihood ratio test for species membership based on DNA sequence data

Mikhail V. Matz[1] and Rasmus Nielsen[2,*]

[1]*Whitney Laboratory for Marine Bioscience, Department of Molecular Genetics and Microbiology, University of Florida, 9505 Ocean Shore Blvd, Saint Augustine, FL 32080, USA*
[2]*Center for Bioinformatics, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark*

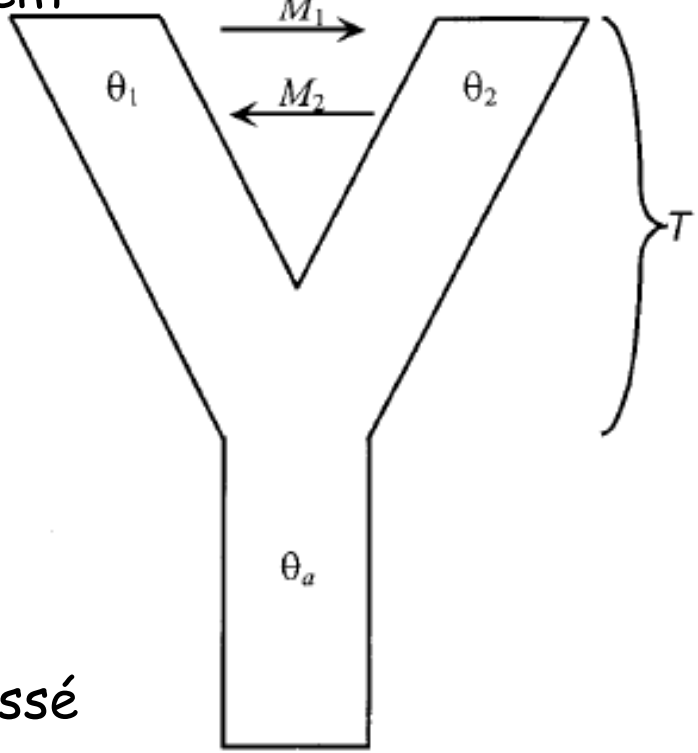
DNA barcoding as an approach for species identification is rapidly increasing in popularity. However, it remains unclear which statistical procedures should accompany the technique to provide a measure of uncertainty. Here we describe a likelihood ratio test which can be used to test if a sampled sequence is a member of an *a priori* specified species. We investigate the performance of the test using coalescence simulations, as well as using the real data from butterflies and frogs representing two kinds of challenge for DNA barcoding: extremely low and extremely high levels of sequence variability.

Keywords: DNA barcoding; likelihood ratio tests; assignment of individuals; coalescent simulations

4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

Un modèle démographique intéressant pour le Barcode ADN : Le modèle de divergence avec ou sans flux de gènes

présent

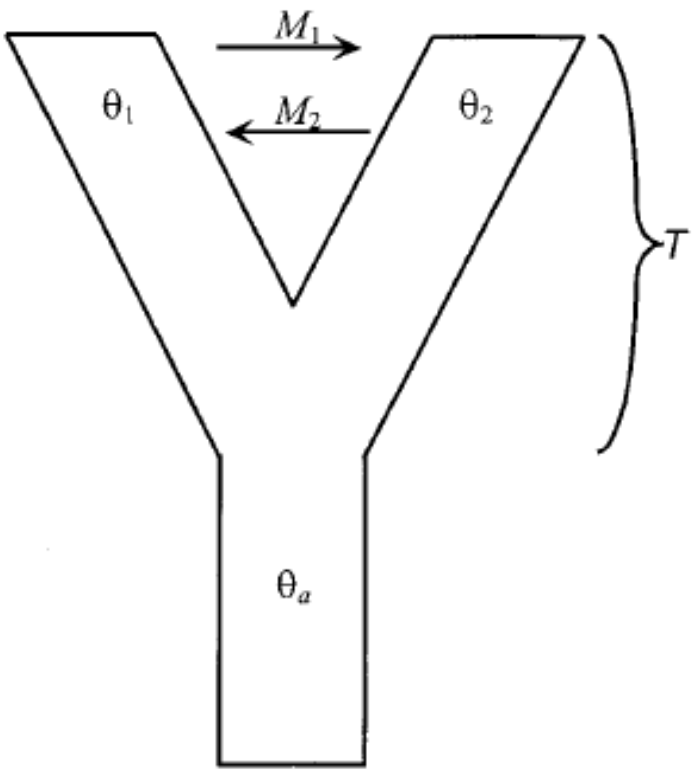


passé

Dans le cadre de l'assignation pure, une des populations est représenté uniquement par la séquence focale, l'autre représente l'espèce testée (échantillon de séquence de l'espèce présentent dans BOLD).

4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

Un modèle démographique intéressant pour le Barcode ADN : Le modèle de divergence avec ou sans flux de gènes



Indices permettant de dire si la séquence focale appartient ou non a l'espèce :

Il y a de la migration (M1 et M2 ne sont pas nuls)

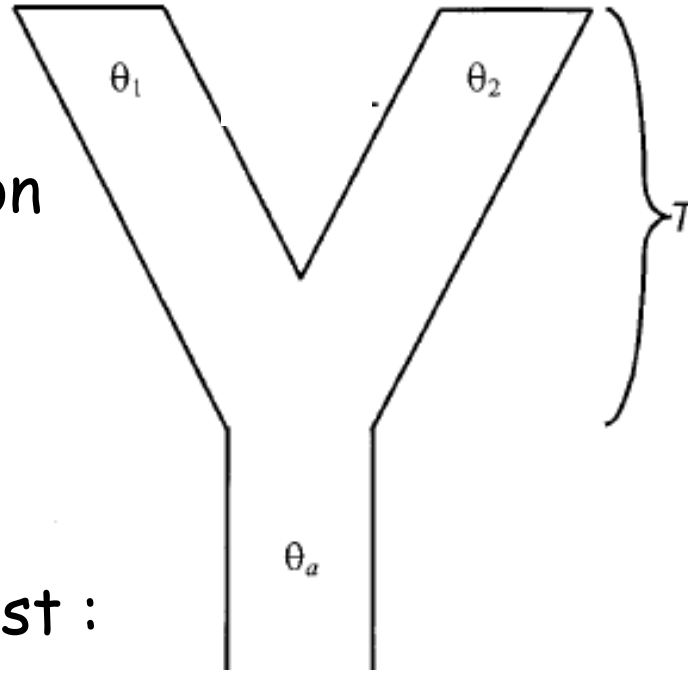
Et/ou pas de divergence importante entre la séquence focale et l'espèce testée (T=0)

= tests d'hypothèses sous maximum de vraisemblance

4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

A likelihood ratio test for species membership based on DNA sequence data

Mikhail V. Matz[1] and Rasmus Nielsen[2,*]



Modèle de divergence pure sans migration

Test d'hypothèse : T = 0

(= la séquence focale appartient à cette espèce)

Indice de confiance = Likelihood ratio test :

$$-2\,\mathrm{Log}\left(\frac{L_{T=0}(T)}{\max_T\{L(T)\}}\right),$$

null hypothesis is rejected at the 5% significance level if the likelihood ratio statistic exceeds ~2.71.

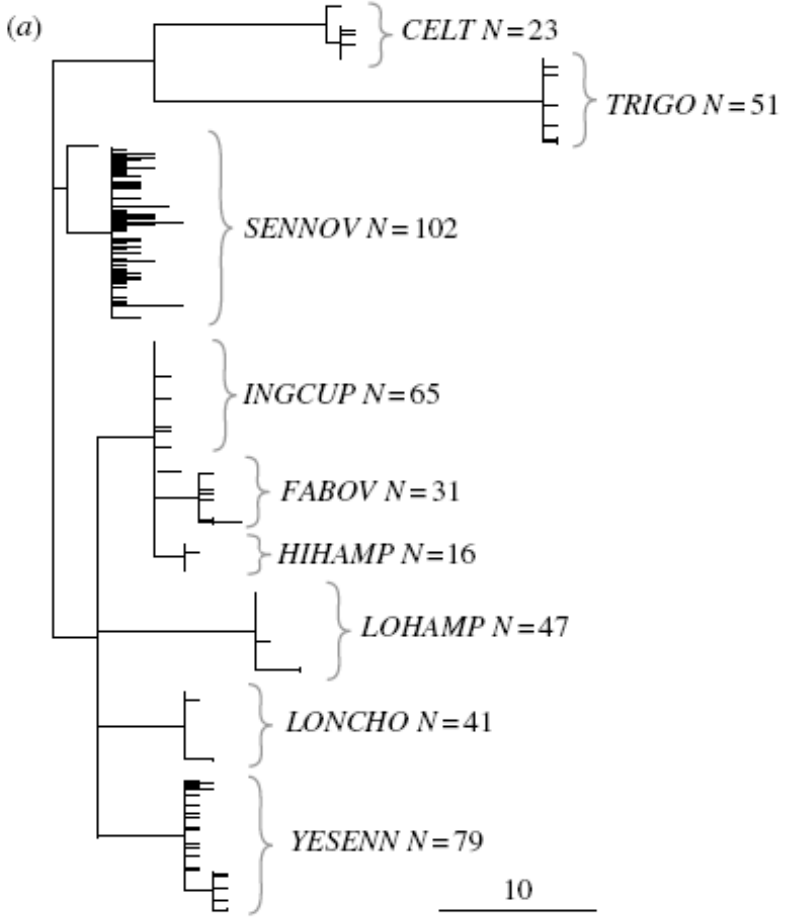
where $L(T)$ is the integrated likelihood function for T.

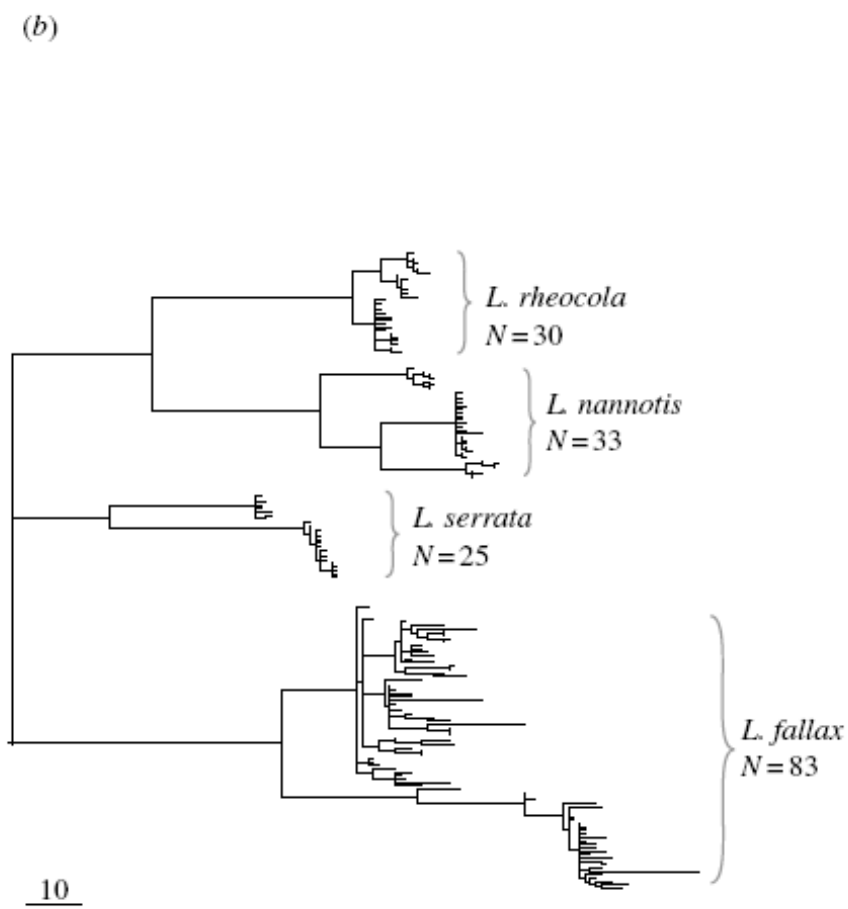
4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

Tests sur 2 jeux de données réels :

a) Astraptes : faible diversité et faible divergence

b) Littoria : forte diversité et forte divergence





4. Extension des approches populationnelles à l'étude des espèces :

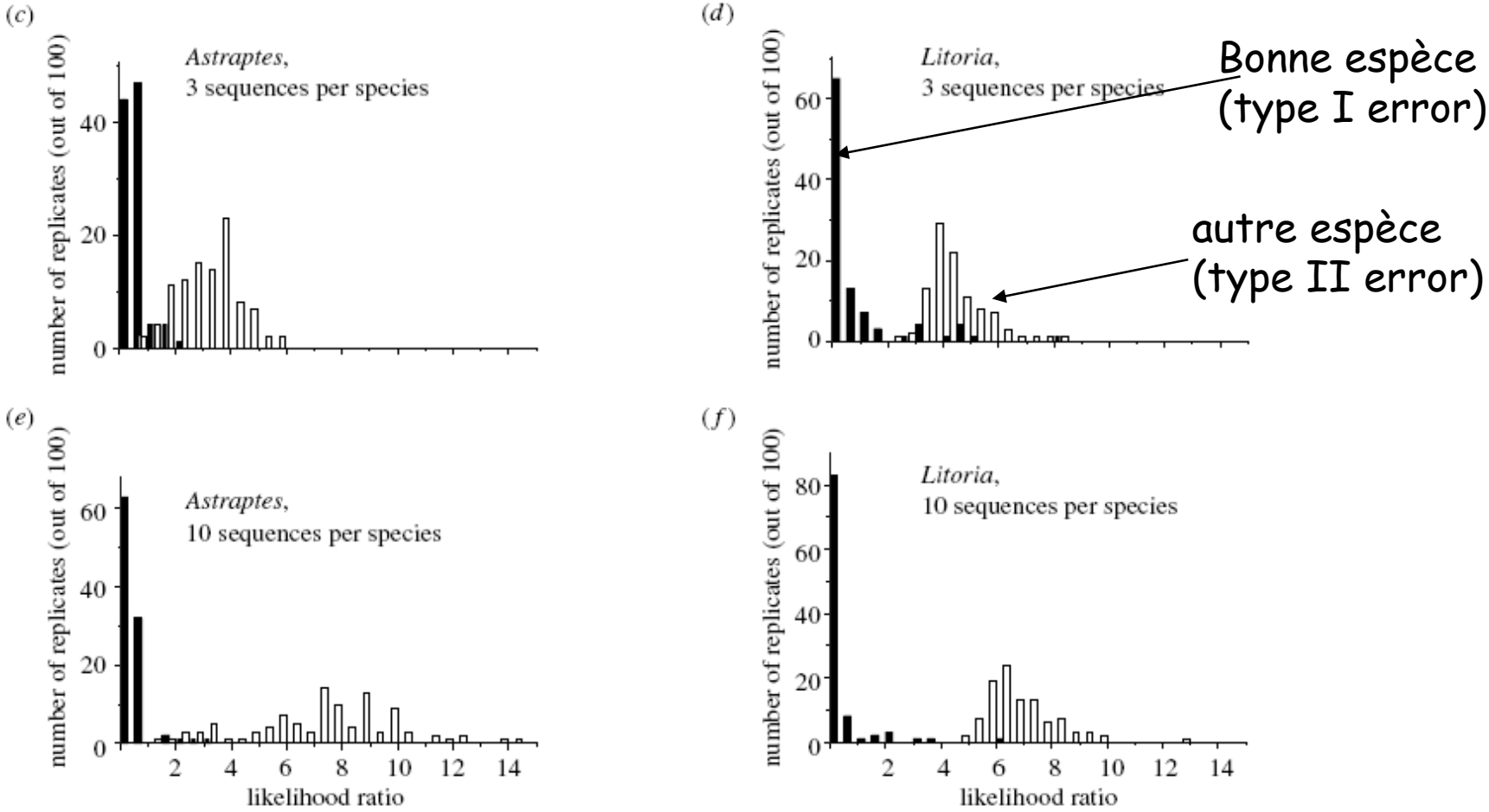


Figure 2. Consensus maximum parsimony trees for *cox1* sequences from the two real data sets: (*a*) skipper butterfly *Astraptes fulgerator* species complex, and (*b*) four species of the tree frogs of the genus *Litoria*. Scale bars: 10 nucleotide changes. The number of individual sequences per species is indicates near the species names. (*c–f*): frequency distributions of the likelihood ratio test statistic in simulations with these datasets. The number of sequences used to represent a true or sister species in the test was either 3 (*c*, *d*) or 10 (*e*, *f*). Filled bars, test with correct species to assess type I error rate; open bars, test with sister species to assess type II error rate.

4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

Résumé du test sur données réelles :

Faible divergence -> erreurs de type II : faux négatif

Forte divergence -> erreurs de type I : faux positif

Marche mieux quand forte divergence et gros échantillon

10 séquences OK

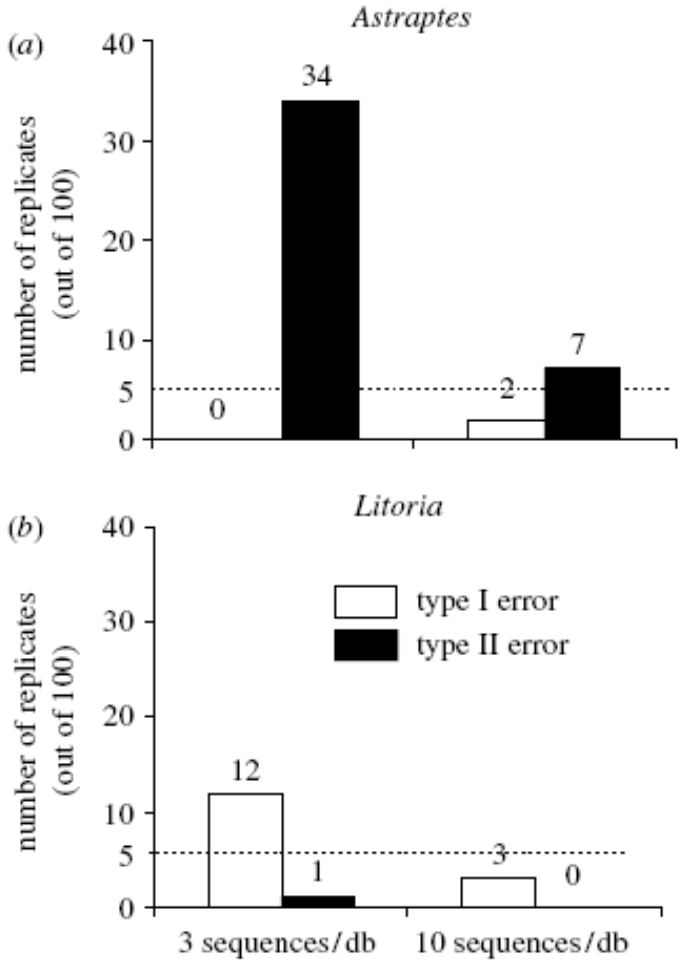
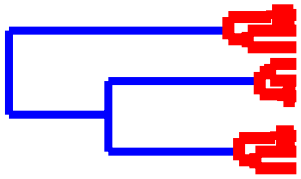


Figure 3. Summary of error rates obtained for *Astraptes* (*a*) and *Litoria* (*b*) datasets with different number of sequences per species in the database, assuming the critical value of 2.7. Open bars, type I error rate; filled bars, type II error rate.



4. Extension des approches populationnelles à l'étude des espèces : Coalescence pure

Conclusion sur l'assignation a l'aide de la coalescence :

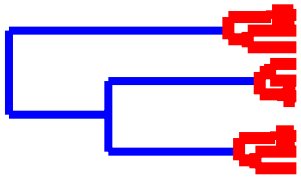
Pas besoin de prendre en compte des modèles de spéciation -> modèle relativement simple (mais peut être trop simpliste)

Semble bien marcher mais nécessite de gros jeux de données pour avoir une bonne précision

Trade-off précision/temps de calculs

Nécessité de faire plus des test (par simulations et sur de nombreux jeux de données réelles)

Assez peu de méthodes développées à ce jour (2!)…



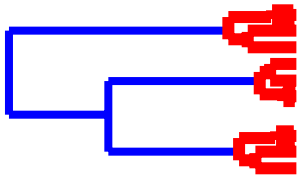
4. Extension des approches populationnelles à l'étude des espèces : Coalescence et spéciation

Analyses de données barcode ADN et délimitation d'espèce :

C'est de la taxonomie moléculaire.

Beaucoup plus compliqué et plus controversé que l'assignation par Barcode ADN car nécessite des critères fiables de différenciation d'espèces

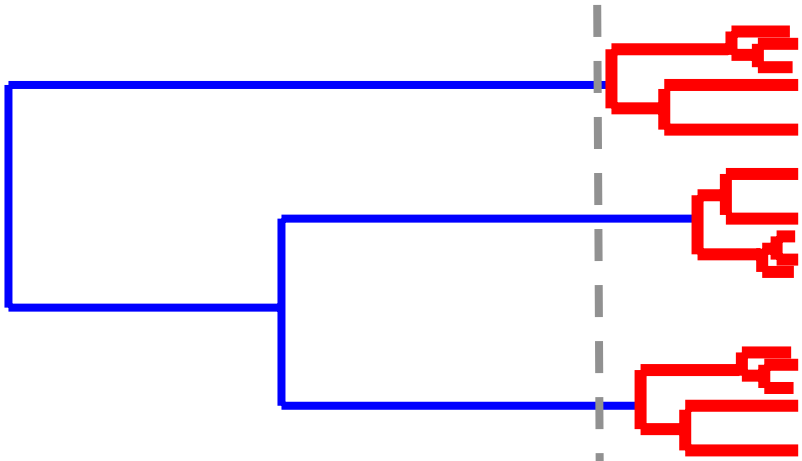
Pour l'instant, les critères de distance génétiques et seuils (3-5% ou 10x rule) sont utilisés mais pas de support biologique



4. Extension des approches populationnelles à l'étude des espèces : Coalescence et spéciation

Extension des outils de génétiques des population pour faire de la délimitation d'espèce :

Beaucoup plus compliqué que l'assignation car nécessite des modèles de spéciations couplés à des modèles populationnels



Branchements inter-espèce

Modèles de spéciation :

Taux de spéciation, d'extinction

Branchements intra-espèce

Théorie de la coalescence :

Taille de pops, flux de gènes,
Histoire démographique et selective

4. Délimitation d'espèces : taxonomie moléculaire

Syst. Biol. 55(4):595–609, 2006

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150600852011

Sequence-Based Species Delimitation for the DNA Taxonomy of Undescribed Insects

JOAN PONS,[1,2,10] TIMOTHY G. BARRACLOUGH,[2,3] JESUS GOMEZ-ZURITA,[1,2,7] ANABELA CARDOSO,[1,4,7] DANIEL P. DURAN,[1,8] STEAPHAN HAZELL,[1,2,9] SOPHIEN KAMOUN,[5] WILLIAM D. SUMLIN,[6] AND ALFRIED P. VOGLER[1,2]

[1]*Department of Entomology, The Natural History Museum, London SW7 5BD, United Kingdom; E-mail: a.vogler@nhm.ac.uk (A.P.V.)*
[2]*Division of Biology and NERC Centre for Population Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire SL5 7PY, United Kingdom*
[3]*Jodrell Laboratory, Royal Botanic Gardens, Kew TW9 3DS, United Kingdom*
[4]*Faculdade de Ciências da Universidade de Lisboa, Departamento de Biologia Animal, Centro de Biologia Ambiental, Rua Ernesto Vasconcelos, 1749-016, Campo Grande, Lisboa, Portugal*
[5]*Department of Plant Pathology, Ohio State University, Ohio Agricultural Research and Development Center, Wooster, Ohio, 44691, USA*
[6]*Department of Entomology, Texas A&M University, College Station, Texas 77843, USA*
[7]*Present Address: Area de Biología Animal, Departamento de Zoología y Antropología Física, Facultad de Biología, Universidad de Murcia–Campus de Espinardo, 30071 Murcia, Spain*
[8]*Present Address: Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee, 37235, USA*
[9]*Present Address: Division of Zoology, School of Animal and Microbial Sciences, University of Reading, Reading RG6 6AJ, United Kingdom*
[10]*Present Address: Unitat de Biologia Evolutiva, Facultat de Ciencias de la Salut i e de la Vida, Universitat Pompeu Fabra, C/Dr. Aiguader 80, 08003 Barcelona, Catalonia, Spain*

Abstract.— Cataloging the very large number of undescribed species of insects could be greatly accelerated by automated DNA based approaches, but procedures for large-scale species discovery from sequence data are currently lacking. Here, we use mitochondrial DNA variation to delimit species in a poorly known beetle radiation in the genus *Rivacindela* from arid Australia. Among 468 individuals sampled from 65 sites and multiple morphologically distinguishable types, sequence variation in three mtDNA genes (cytochrome oxidase subunit 1, cytochrome *b*, 16S ribosomal RNA) was strongly partitioned between 46 or 47 putative species identified with quantitative methods of species recognition based on fixed unique ("diagnostic") characters. The boundaries between groups were also recognizable from a striking increase in branching rate in clock-constrained calibrated trees. Models of stochastic lineage growth (Yule models) were combined with coalescence theory to develop a new likelihood method that determines the point of transition from species-level (speciation and extinction) to population-level (coalescence) evolutionary processes. Fitting the location of the switches from speciation to coalescent nodes on the ultrametric tree of *Rivacindela* produced a transition in branching rate occurring at 0.43 Mya, leading to an estimate of 48 putative species (confidence interval for the threshold ranging from 47 to 51 clusters within 2 log*L* units). Entities delimited in this way exhibited biological properties of traditionally defined species, showing coherence of geographic ranges, broad congruence with morphologically recognized species, and levels of sequence divergence typical for closely related species of insects. The finding of discontinuous evolutionary groupings that are readily apparent in patterns of sequence variation permits largely automated species delineation from DNA surveys of local communities as a scaffold for taxonomy in this poorly known insect group. [Phylogenetic species concept; coalescence; mtDNA; Cicindelidae; Australia; paleoclimate.]

4. Délimitation d'espèces : taxonomie moléculaire

tree, x_i. We combine standard models that separately consider branching within populations (Hudson, 1991; Wakeley, 2006) and branching between species (Nee, 1994, 2001; Nee et al., 1994). Under a neutral coalescent, the likelihoods of the waiting times within a single population with effective population size Ne and n_i lineages present during waiting time i are given by:

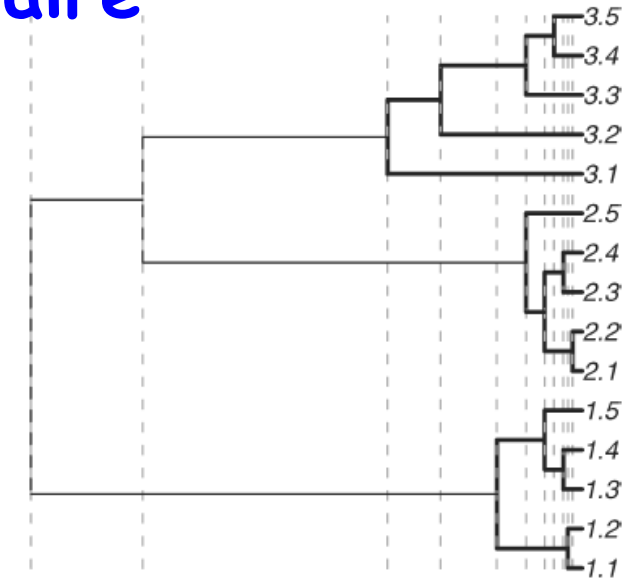
$$L_{(x_i)} = \lambda n_i (n_i - 1) e^{-\lambda n_i (n_i - 1) x_i} \quad (1)$$

where the birth rate

$$\lambda = \frac{1}{2N_e} \quad (2)$$

The simplest standard approach for considering branching between species is as a Yule model (Yule, 1924); i.e., a stochastic birth-only model. The likelihoods of the waiting times in a species phylogeny (one tip per species) of a clade with constant average speciation rate, λ, and no extinction are given by:

$$L_{(x_i)} = \lambda n_i e^{-\lambda n_i x_i} \quad (3)$$



Waiting intervals	x1	x2	x3	x4	x5	etc.
Diversification (n)	2	3	2	2	1	etc.
Species 1 (n)	0	0	2	3	3	etc.
Species 2 (n)	0	0	0	0	0	etc.
Species 3 (n)	0	0	0	0	2	etc.

FIGURE 2. Schematic illustration of the waiting times in a calibrated tree and the numbers of lineages present for each type of diversification process (interspecies diversification and within-species coalescence) during each waiting interval. Branches are categorized as either between species (thin lines) or within species branching (bold lines) according to the procedures described in Material and Methods.

4. Délimitation d'espèces : taxonomie moléculaire

Mixage des modèle de coa et de spéciation :

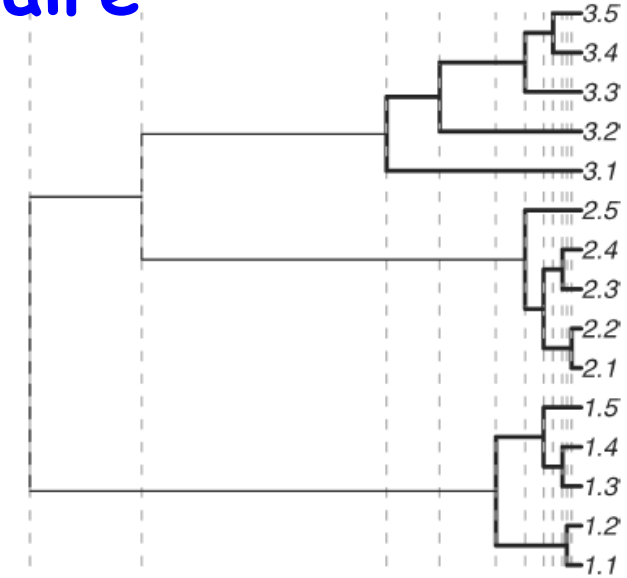
X_i : temps d'attentes entre brachements

$$L(x_i) = be^{-bx_i}$$

$$b^* = \beta(n_i) + \sum_{j=1,k}(\lambda_j(n_{i,j}(n_{i,j}-1)))$$

Taux de spéciation

Taux de coalescence dans la pop j = $1/2N_j$



Waiting intervals	x1	x2	x3	x4	x5	etc.
Diversification (n)	2	3	2	2	1	etc.
Species 1 (n)	0	0	2	3	3	etc.
Species 2 (n)	0	0	0	0	0	etc.
Species 3 (n)	0	0	0	0	2	etc.

FIGURE 2. Schematic illustration of the waiting times in a calibrated tree and the numbers of lineages present for each type of diversification process (interspecies diversification and within-species coalescence) during each waiting interval. Branches are categorized as either between species (thin lines) or within species branching (bold lines) according to the procedures described in Material and Methods.

4. Délimitation d'espèces : taxonomie moléculaire

Les taux de coalescence plus forts que les taux de spéciation -> rupture entre le régime de coalescence et le régime de spéciation

But : détection de cette rupture et localisation dans le temps

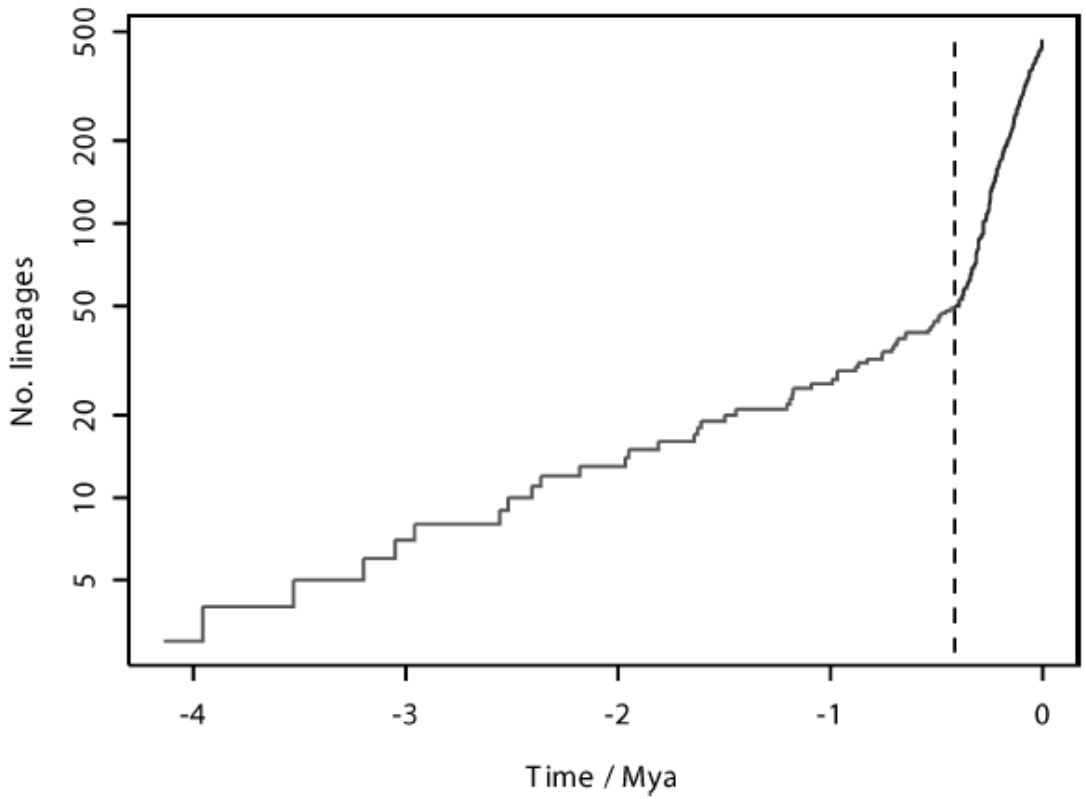
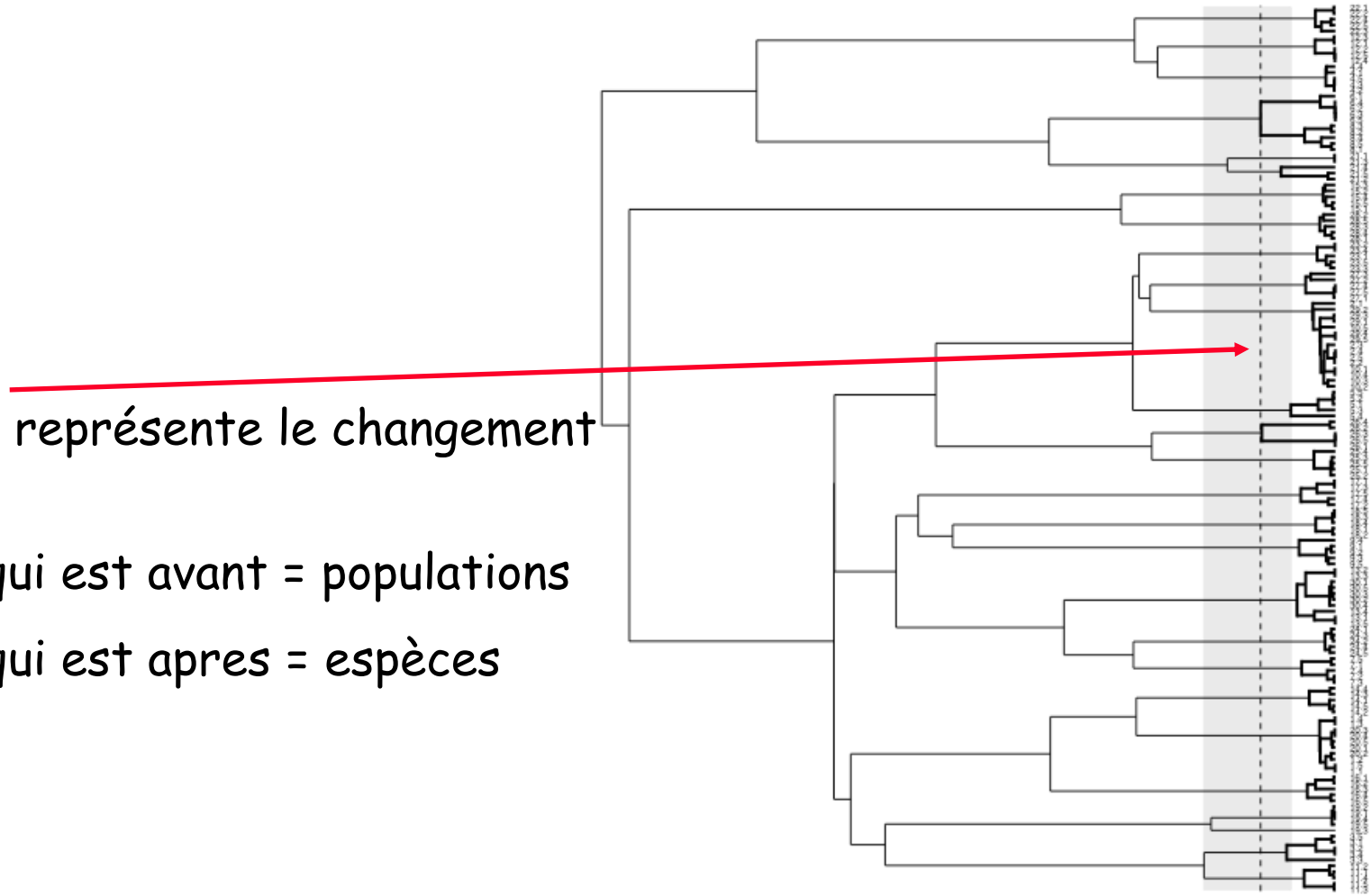


FIGURE 5. Lineages-through-time plot based on the time calibrated tree obtained from all 468 haplotypes. The sharp increase in branching rate, corresponding to the transition from interspecies to intraspecies branching events, is indicated by the dotted line.

4. Délimitation d'espèces : taxonomie moléculaire



Cette ligne représente le changement de régime

-> tout ce qui est avant = populations

-> tout ce qui est apres = espèces

4. Délimitation d'espèces : taxonomie moléculaire

Application à un jeux de données réelles

Mais pas de tests rigoureux de la méthode

On s'attend à ce que ca marche mal pour des évènements de spéciation récents (pb car c'est la que c'est intéressant!)

A tester sur Astraptes par exemple...

4. Délimitation d'espèces : taxonomie moléculaire

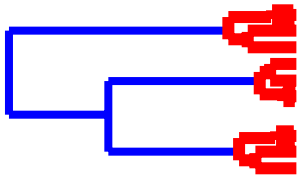
Sequence-Based Species Delimitation for the DNA Taxonomy of Undescribed Insects

JOAN PONS,[1,2,10] TIMOTHY G. BARRACLOUGH,[2,3] JESUS GOMEZ-ZURITA,[1,2,7] ANABELA CARDOSO,[1,4,7] DANIEL P. DURAN,[1,8] STEAPHAN HAZELL,[1,2,9] SOPHIEN KAMOUN,[5] WILLIAM D. SUMLIN,[6] AND ALFRIED P. VOGLER[1,2]

Approche intéressante mais pour l'instant assez limité :

- un seul de taux de coalescence pour toutes les pops et espèces (i.e. taille de pop identique pour tous les groupes)
- un seul taux de spéciation

Possibilité de considérer des taux variables selon les groupes mais augmente grandement le nombre de paramètres à estimer, et donc les tailles d'échantillons necessaires et les temps de calculs



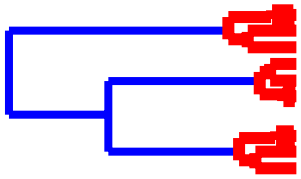
4. Délimitation d'espèces : taxonomie moléculaire

Possibilité d'étendre les approches de coalescence par maximum de vraisemblance (Matz & Nielsen 2005, Abdo & Golding 2006) pour faire de la délimitation d'espèces mais toujours les mêmes problèmes :

Plus de paramètres à estimer - > plus de données nécessaires et temps de calculs plus longs

La délimitation d'espèce devrait être plus efficaces si on considère en plus de COI des gènes nucléaires (dans la phase de construction de la base de donnée, pas pour l'assignation pure) et notamment des gènes recombinants.

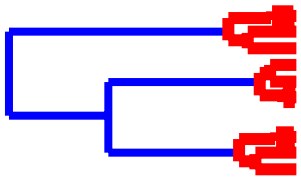
On s'éloigne des buts du barcode pour aller vers de la systématique moléculaire (moins standard et non faisable par des non spécialistes)…



4. Délimitation d'espèces : taxonomie moléculaire

Mais le barcode n'est pas une nouvelle méthode pour faire de la taxonomie et de la systématique mais surtout un outils de détermination (cf. clé de détermination) :

La délimitation d'espèces à partir des données Barcode pourra seulement suggérer de ré-analyser selon les approches classiques de taxonomie certains taxons pour lesquels le "Barcode ADN" suggérerait une séparation d'une espèce en plusieurs ou le regroupement de plusieurs espèces en une seule…



5. Conclusions sur l'utilisation d'outils de génétique des populations pour les analyses de données barcode ADN

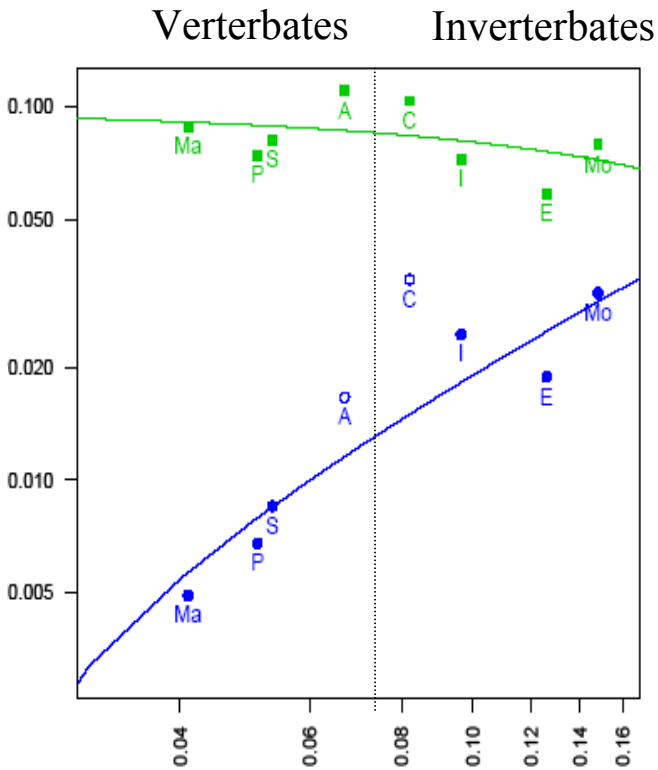
Les problèmes des analyses par coalescence des données Barcode ADN :

- Tx de spéciation variables + Θ variables (tailles de pops, structuration, fluctuations démographiques,…),

  -> beaucoup de paramètres de nuisance mais possibilité de les estimer uniquement avec la base de données de référence, réutilisable ensuite pour l'assignation de séquences focales

- Coalescent = modèle neutre, COI fonctionne t il comme un gène neutre??

Bazin et al. 2006 Science



Allozyme heterozygosity

- population size influences nuclear, but not mitochondrial DNA diversity

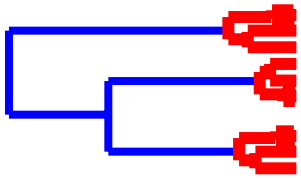
mtDNA

nuclear DNA

-recurrent adaptive evolution (**selective sweeps**) can explain the homogeneous mtDNA pattern

Peut on utiliser la coalescence pour modéliser des gènes mitochondriaux?

Sans doute mais il faut tester si c'est robuste…



3. Besoin de nouvelles méthodes d'analyse

Les problèmes des modèles Coa + spéciation :

•Peu de locus mais ok pour l'assignation car on ne cherche pas estimer beaucoup de paramètres, uniquement tester l'appartenance d'une séquence à une espèce. Plus problématique pour la délimitation d'espèce.

• temps de calcul long mais nouveau algorithmes plus efficaces pour modèles relativement simples

Conclusion

Le projet "Barcode ADN" potentiellement intéressant mais beaucoup de points restent a tester pour que ca marche à 99%

->important de tester sur des jeux de données "ambigus"

Extension des approches populationnelles aux espèces

bien pour comprendre la répartition intra- et inter-spécifique de la diversité génétique

bon cadre pour quantifier, tester et prédire les erreurs et incertitudes de l'approche

perspective : intéressant pour étudier de façon combinée la micro et macro évolution et les mécanismes de spéciation avec des modèles populationnels réalistes…