Apport des approches phylogénétiques pour expliquer l'origine des génomes mosaïques, exemple chez le Riz

Charles-Elie Rabier

Vincent Berry, Fabio Pardi et Céline Scornavacca

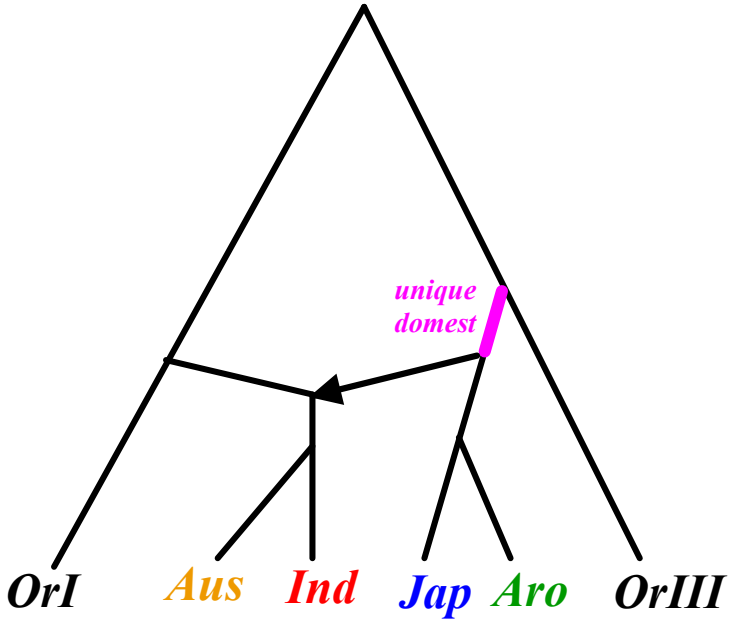
ISEM, Institut des Sciences de l'Evolution de Montpellier
LIRMM, Laboratoire d'informatique, de Robotique et de Microélectronique
Genome Harvest

Jean-Christophe Glaszmann, Joao Santos

AGAP, Amélioration Génétique et Adaptation des Plantes, CIRAD

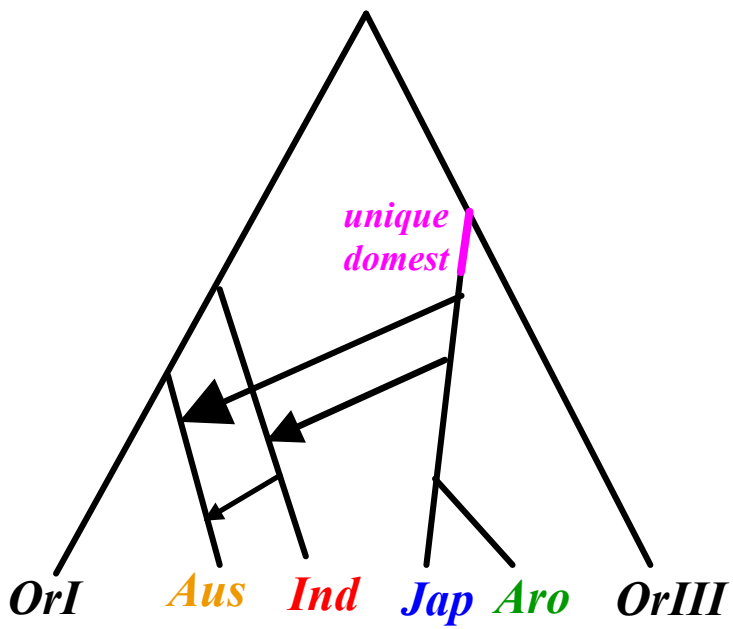
Quelques thèses sur la domestication

- Huang et al. (Nature, 2012) : japonica domestiqué à partir d'un riz sauvage dans le sud de la Chine, puis croisé à un sauvage dans le sud est de l'Asie, générant indica



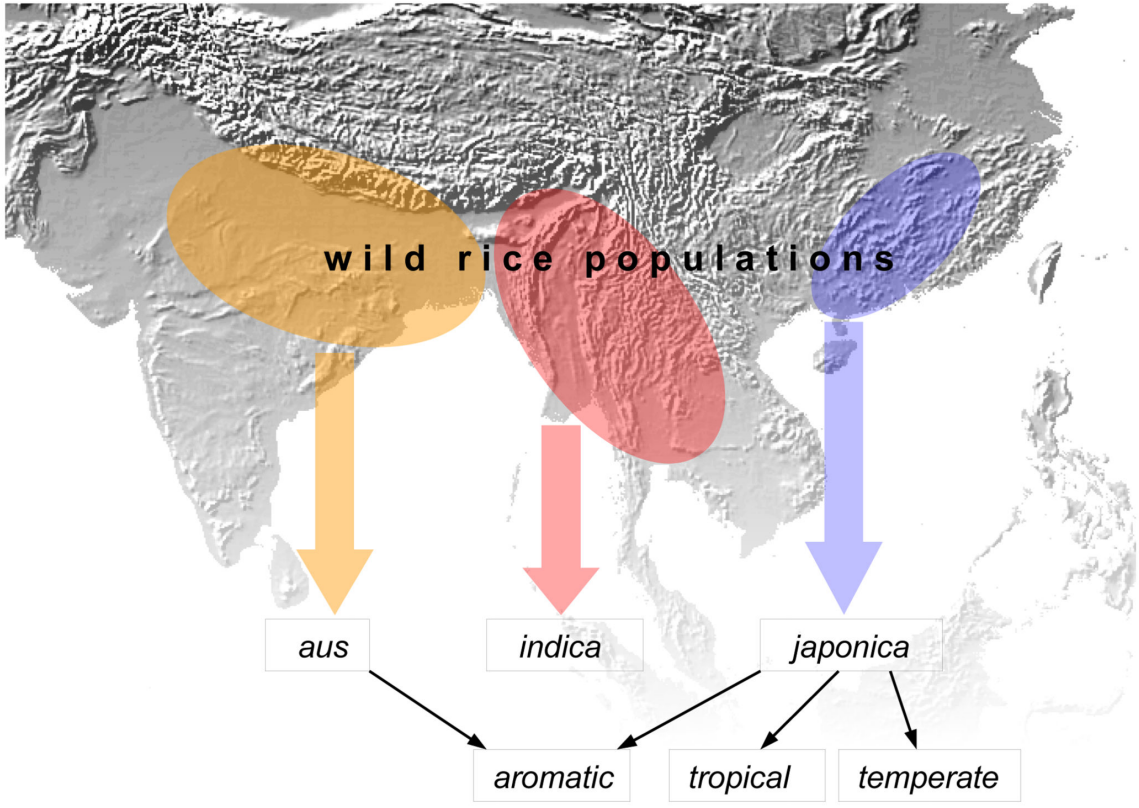
Quelques thèses sur la domestication

- Choi et al. (MBE, 2017) soutiennent aussi un seul évènement de domestication (japonica). Introgression par hybridation de japonica et proto-indica et proto-aus, générant indica et aus



Quelques thèses sur la domestication

- Civan et al. (Nature Plants, 2015) : indica, japonica et aus domestiqués séparément dans différentes parties d'Asie



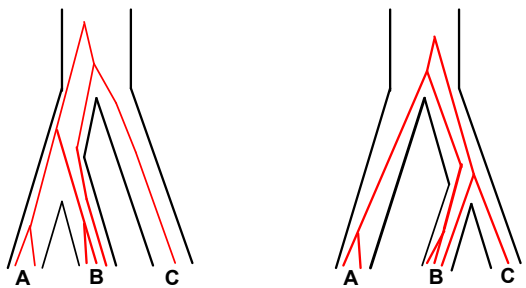
Notre approche méthodologique

On s'intéresse à un modèle qui, outre le tri de lignées, considère explicitement les mutations et hybridation. Modélisation Bayésienne plus fine.

Nos pistes :

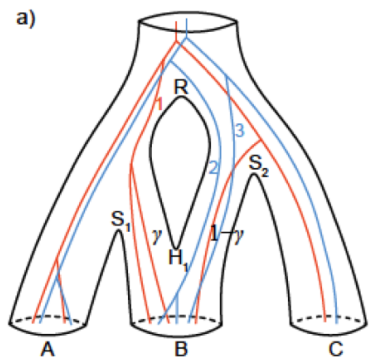
1. Inférence d'arbres d'espèces + arbres résumés en réseaux phylogénétiques

SNAPP (Bryant et al. 2012, MBE) + SplitsTree



2. Inférence directe de réseaux

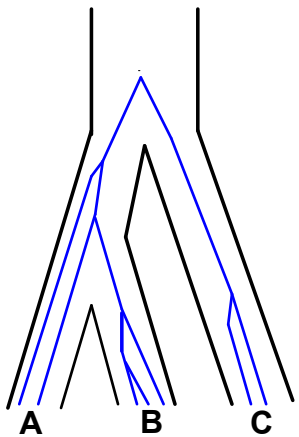
Extension de SNAPP aux réseaux



Logiciel SNAPP pour l'inférence Bayésienne d'arbres (Bryant et al. 2012, MBE)

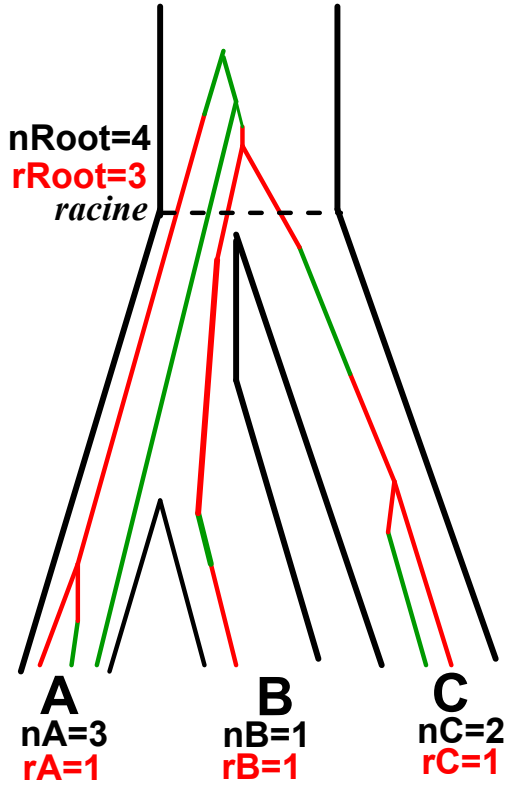
- Marqueurs biallélliques (SNPs) indépendants sachant l'arbre d'espèces

- Modélisation de l'arbre de locus (backward)
  - Processus de coalescence évoluant à l'intérieur d'un arbre d'espèces (MultiSpecies Coalescent)
  - Processus autorisant la discordance entre arbres de locus et arbres d'espèces (tri de lignées incomplet)



Les mutations interviennent au cours du temps

- Modélisation des séquences (forward)
  - mutation (rouge $\leftrightarrow$ vert) : modèle markovien évoluant le long des branches de l'arbre de locus



- V.a. : rRoot, nRoot, rA, rB, rC
- pas d'aléa dans nA, nB, nC
- Data=(rA, rB, rC)
- Vraisemblance : $\mathbb{P}(Data \mid S)$

Calcul de vraisemblance dans un arbre

- $\mathbb{P}(n_{root} = i \mid Count)$ calculé récursivement en remontant dans le temps (postorder)

  Tavaré (Theor Pop Biol, 1984), Watterson (Theor Pop Biol, 1984), Takahata and Nei (Genetics, 1985) ...

- $\mathbb{P}(Data \mid Count, n_{root} = i, r_{root} = j)$ calculé récursivement en remontant dans le temps (postorder)

  Slatkin (Genetics, 1996) vs. Griffiths and Tavaré (Springer, 1997)

- $\mathbb{P}(r_{root} = j \mid n_{root} = i)$ calculé par
  - la loi Binomiale : $\mathbb{P}(r_{root} = j \mid n_{root} = i) = C_i^j p^j (1-p)^{i-j}$
  - la loi $\beta(\theta, \theta)$ sur le paramètre p de la Binomiale :
    $\mathbb{P}(r_{root} = j \mid n_{root} = i) = C_i^j B(j+\theta, i-j+\theta)/B(\theta, \theta)$

- Astuces afin de raccourcir les calculs : Vraisemblances partielles...

La statistique Bayésienne dans SNAPP

- S : arbre d'espèces (topologie, longueurs de branches, tailles de populations)
- X_i : alignements pour le locus i
- G_i : arbre de locus pour le locus i
- m loci

$$\mathbb{P}(S|X_1,\ldots,X_m) \propto \left(\prod_{i=1}^{m}\int_{\psi}\mathbb{P}(X_i|G_i)\mathbb{P}(G_i|S)dG_i\right)P(S)$$
$$\propto \mathbb{P}\left(Data \mid S\right)P(S)$$

SNAPP intègre sur tous les arbres de locus

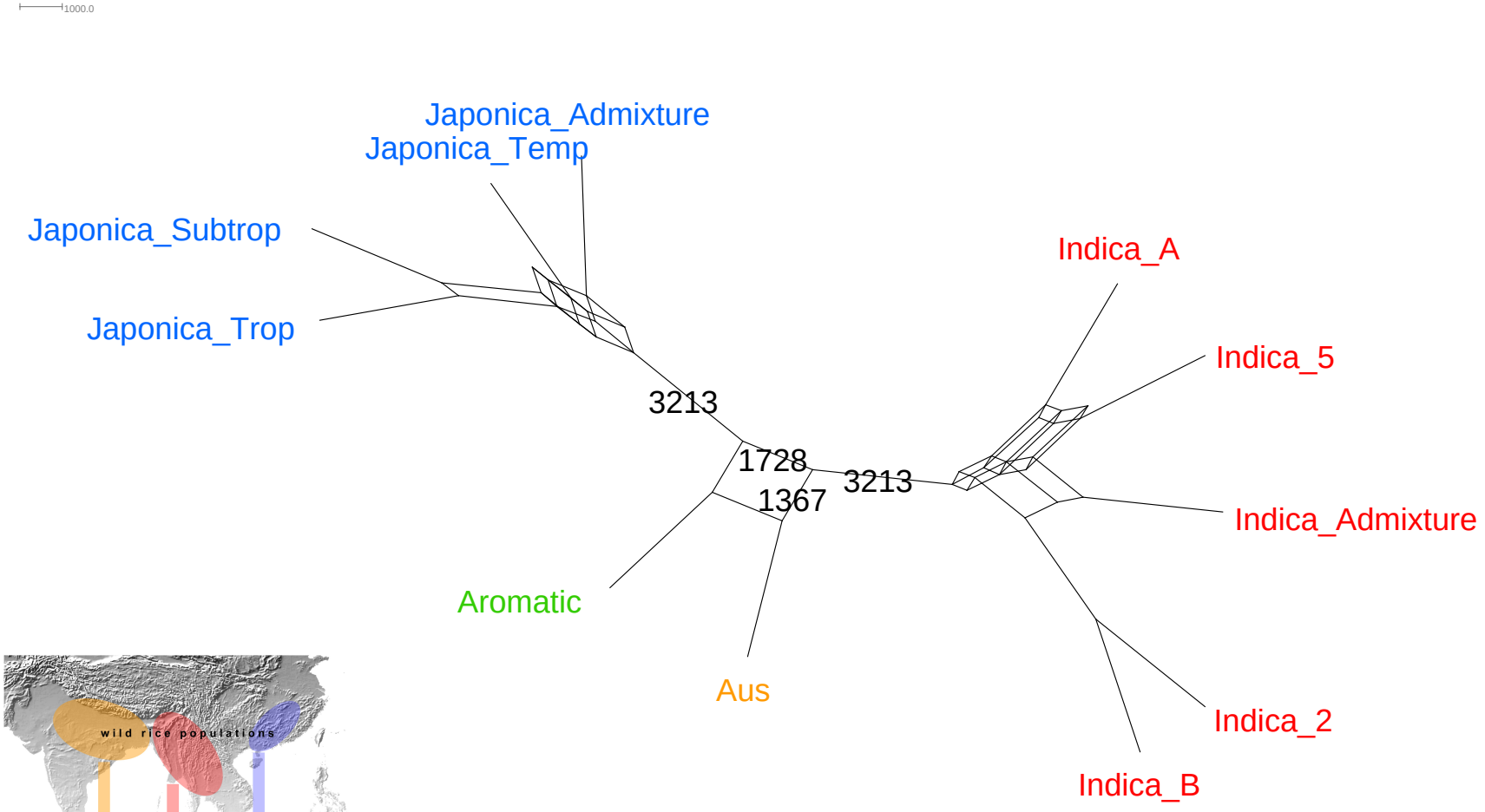
Calcul de la *prior* P(S) par le processus de naissances

$\Rightarrow$ Markov Chain Monte Carlo (MCMC) afin d'estimer la distribution à posteriori de $\mathbb{P}(S|X_1,\ldots,X_m)$

Implémenté dans BEAST

Chromosome 6 (données J. Santos, J-C. Glaszmann)

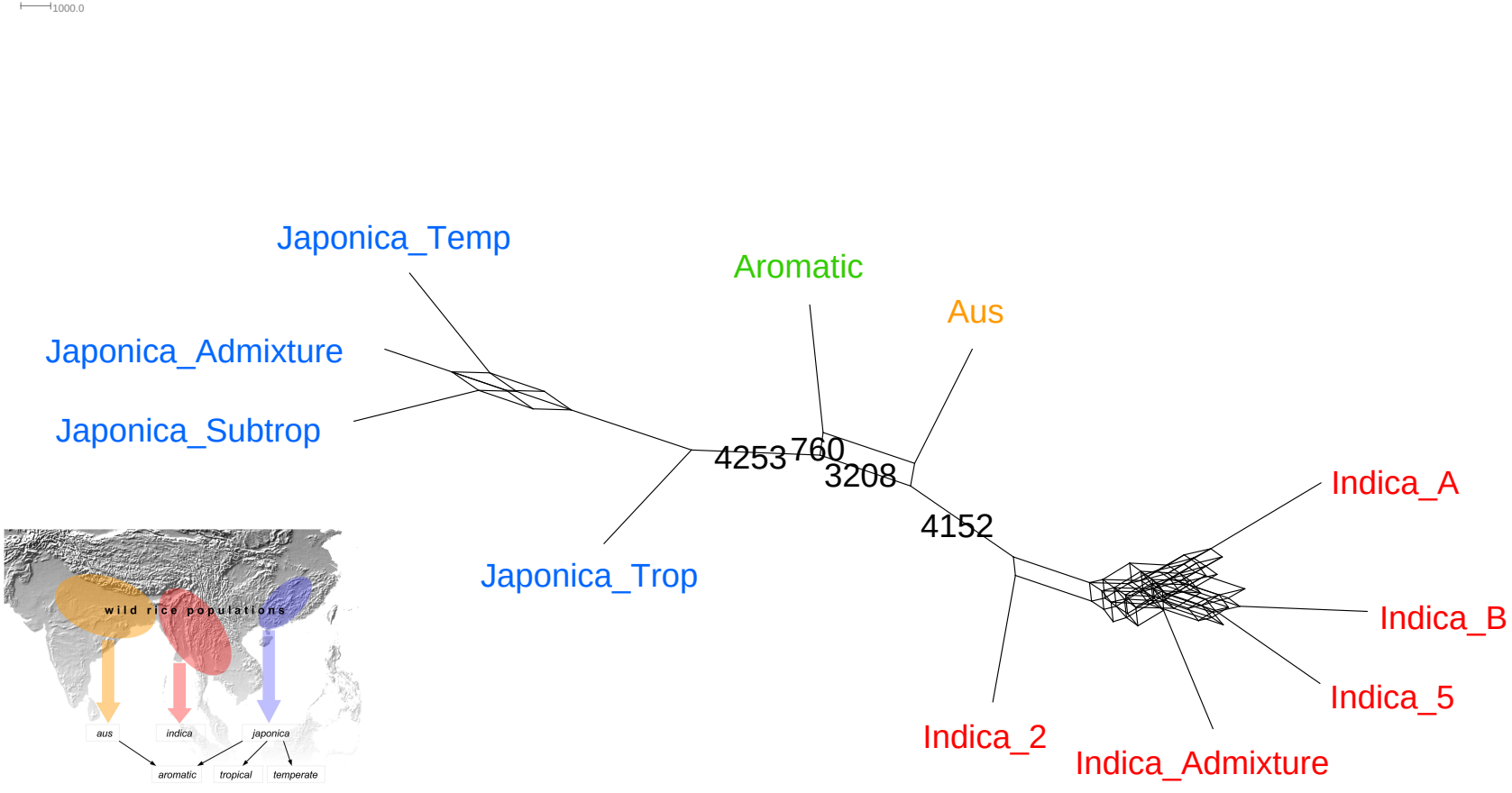
Conservation de 1550 SNPs (un SNP tous les 500)



Chromosome 10 (données J. Santos, J-C. Glaszmann)

Conservation de 1089 SNPs (un SNP tous les 500)

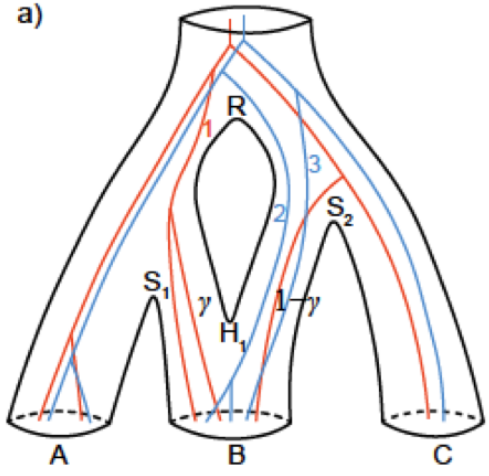
- JDD2 (1er SNP= 50ème SNP du chromosome 10)



Simulateur basé sur un réseau (Genome Harvest)

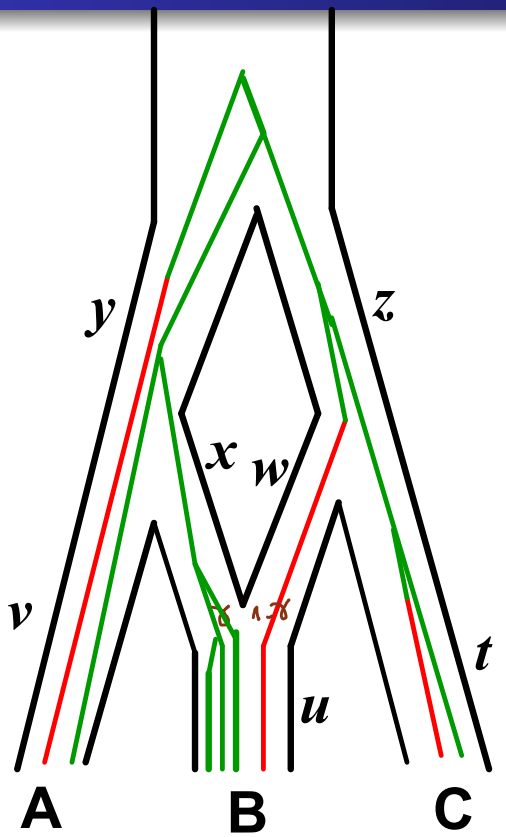
SNAPPSimNet construit sur la base du simulateur SNAPPSim de Bryant et al. (2012)

- Génération d'arbres de locus évoluant à l'intérieur d'un réseau selon un processus de coalescence



- Snapp est fortement attiré par un scénario sous-jacent au réseau

Piste 2 : une méthode Bayésienne directe d'inférence de réseaux



$Data_z$: proportion de rouge/vert dans les espèces sous la branche z

$Data_y$: proportion de rouge/vert dans les espèces sous la branche y

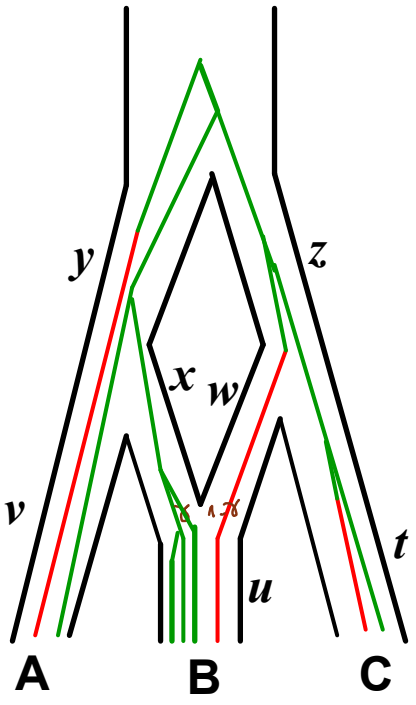
$Data_{z^T}$ et $Data_{y^T}$ ne sont plus indépendantes...

$Data_{z^T}$ et $Data_{y^T}$ comprennent les allèles rouges et verts de l'espèce hybride B !!!

On ne peut plus effectuer le produit des probabilités

$\mathbb{P}(Data_{z^T}) \times \mathbb{P}(Data_{y^T})$!!!

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

(7) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

- Optimisation combinatoire des calculs en cours (dimension des matrices)

Conclusion

- Implémentation de la méthode Bayésienne pour les réseaux
- L'inférence de réseau est un sujet compétitif : Tanja Stadler (ETH Zurich), Luay Nakhleh (Rice University, USA). Notre approche devrait être plus performante sur des réseaux aux nombreuses hybridations
- Afin de comprendre l'histoire des riz cultivés, nécessité de disposer de riz sauvages, à l'instar de Choi et al. (MBE, 2017), Wang et al (Genome Research, 2017) ...

Les données de riz

1. Données disponibles à l'état brut sur le site de l'IRRI
2. Prétraitement de Joao (données manquantes ...)
3. 3023 variétés avec 895 977 marqueurs disponibles sur le chromosome 6 (Merci Joao !)
4. 2 jeux de données proposés par JC Glazmann (core collections)
   - 20 variétés
   - 50 variétés (7 aromatic, 7 aus, 13 indica, 17 japonica, 4 indéterminés)

Constitutions de 4 jeux de données de manière empirique

- Conservation de 1550 SNPs (un SNP tous les 500)
  - JDD1 (1er SNP= 1er SNP du chromosome 6)
  - JDD2 (1er SNP= 50e SNP du chromosome 6)
- Conservation de 7749 SNPs (un SNP tous les 100)
  - JDD3 (1er SNP= 1er SNP du chromosome 6)
  - JDD4 (1er SNP= 50e SNP du chromosome 6)

Analyse des chromosomes 2 et 10

Chromosome 2 :

- 1 129 426 marqueurs
- Conservation de 2026 SNPs (un SNP tous les 500)
  - JDD1 (1er SNP= 1er SNP du chromosome 2)
  - JDD2 (1er SNP= 50e SNP du chromosome 2)

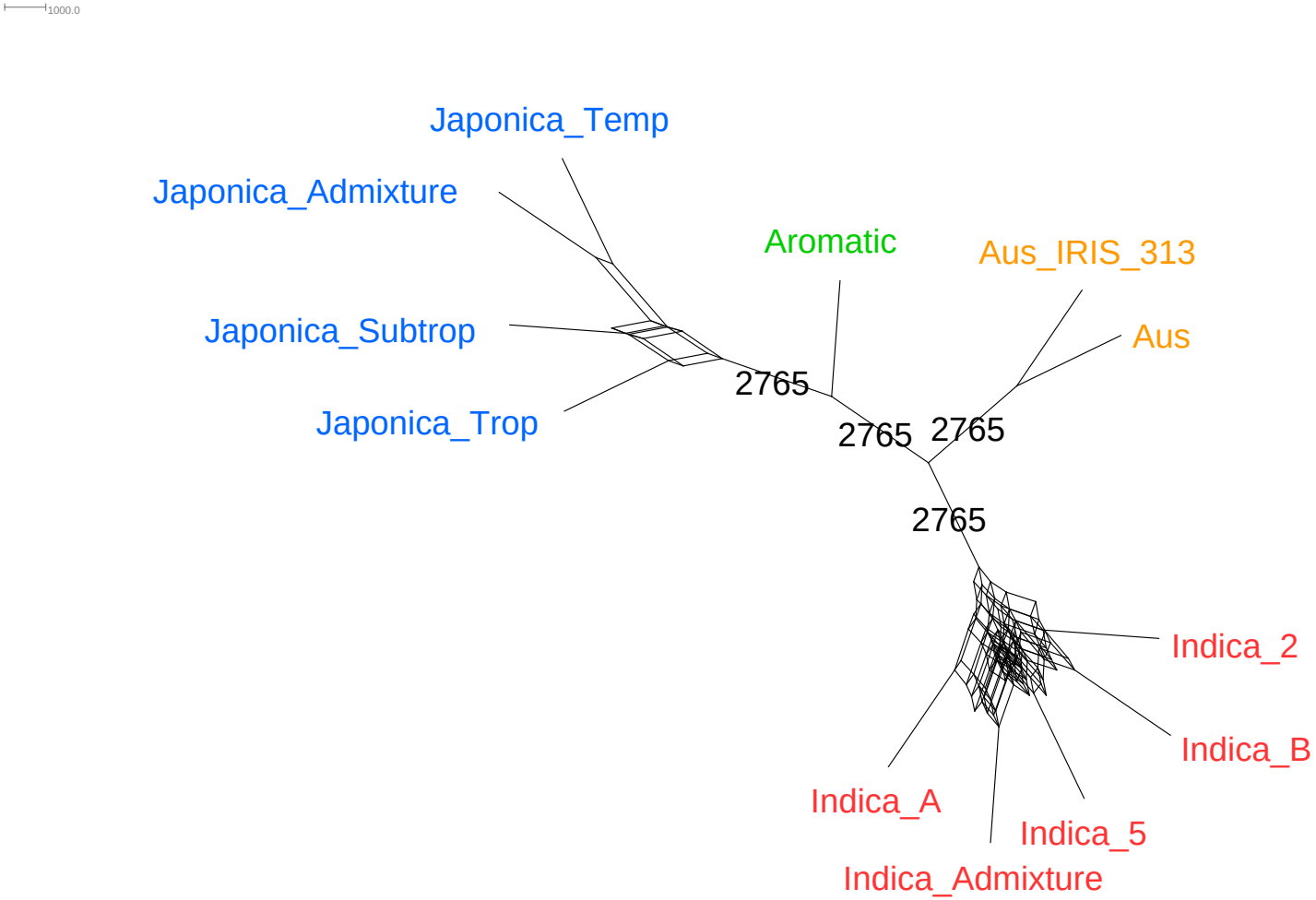
Chromosome 10 :

- 635 037 marqueurs
- Conservation de 1089 SNPs (un SNP tous les 500)
  - JDD1 (1er SNP= 1er SNP du chromosome 10)
  - JDD2 (1er SNP= 50e SNP du chromosome 10)

Chromosome 2 (données J. Santos, J-C. Glaszmann)

Conservation de 2026 SNPs (un SNP tous les 500)

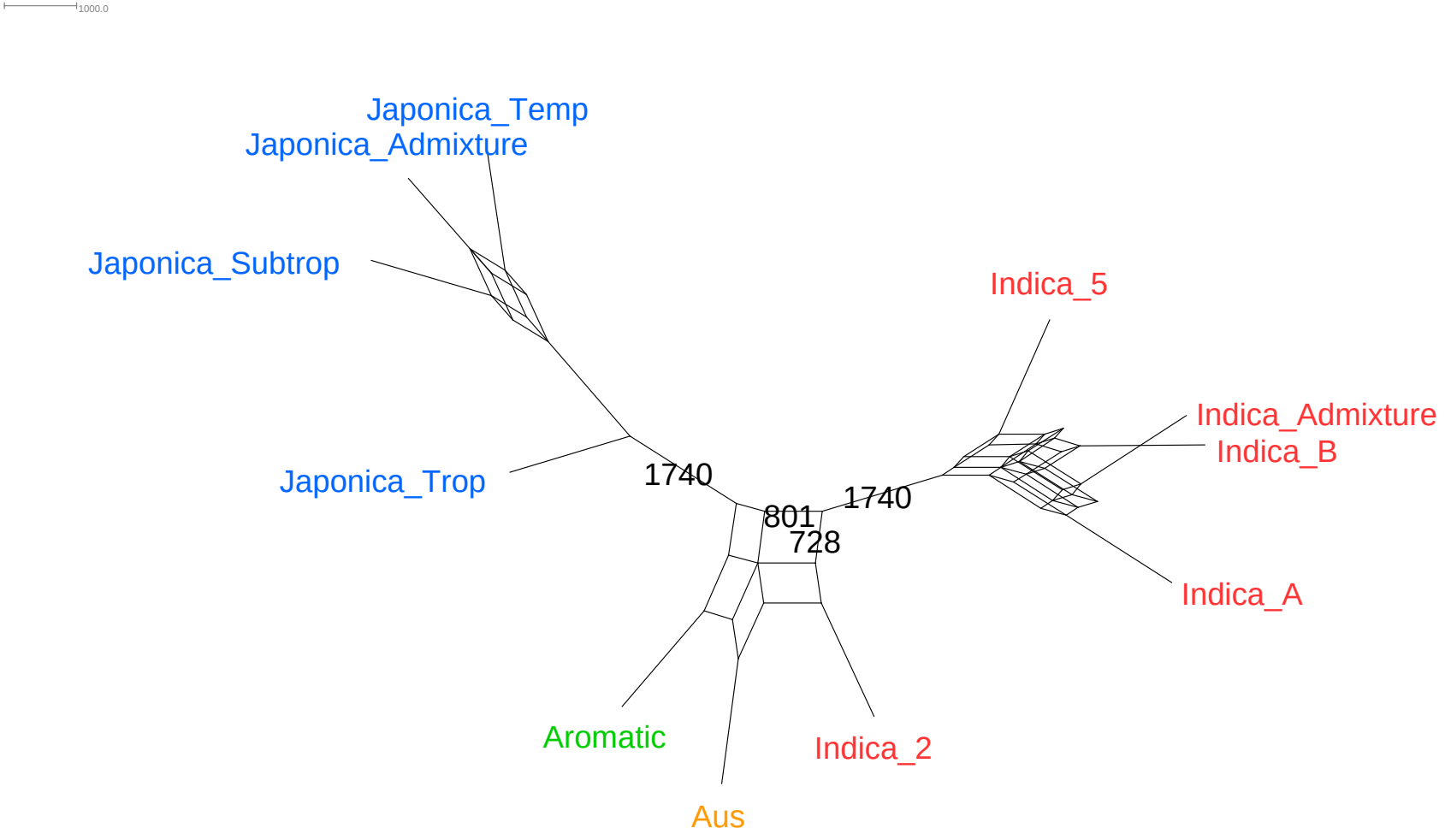
- JDD1 (1er SNP= 1er SNP du chromosome 2)



Chromosome 10 (données J. Santos, J-C. Glaszmann)

Conservation de 1089 SNPs (un SNP tous les 500)

- JDD1 (1er SNP= 1er SNP du chromosome 10)



Calcul de vraisemblance dans un arbre

Où se situe l'aléatoire dans le modèle ?

- Modèle de coalescence $\rightarrow$ pas d'aléatoire dans le nombre de lignées (Count) dans chaque espèce !
- L'aléatoire réside dans la répartition d'allèles rouges et verts dans chaque espèce

$$\mathbb{P}(Data)$$

$$= \sum_i \sum_j \mathbb{P}(Data \mid Count, n_{root} = i, r_{root} = j)\, \mathbb{P}(n_{root} = i, r_{root} = j \mid Count)$$

$$= \sum_i \sum_j \mathbb{P}(Data \mid Count, n_{root} = i, r_{root} = j)\, \mathbb{P}(r_{root} = j \mid n_{root} = i)$$

$$\times\, \mathbb{P}(n_{root} = i \mid Count)$$