La Phylogénie Bayésienne comme Outil d'Etude Archéologique des Peuples de l'Asie du Sud-Est de -10 000 av. J-C à notre ère

Hélie Bazin
sous la direction de Robin Ryder

Mai-Septembre 2022

Dauphine | PSL
EXECUTIVE EDUCATION

Introduction

En 2021, est publié un article intitulé *Triangulation supports agricultural spread of the transeurasian languages* [9] dans la revue Nature. Ce papier propose de croiser trois analyses bayésiennes phylogéniques sur des données génétiques, linguistiques et archéologiques, afin de décrire l'évolution des peuples Est-asiatiques anciens. L'idée de combiner ces trois analyses est de construire une histoire complète reliant l'expansion des technologies agricoles à la diversification génétique et l'évolution des langages.

FIGURE 1 – Asie du Nord-Est

L'approche phylogénique bayésienne s'est imposée depuis quelques années comme une méthode efficace et fiable dans l'étude de l'évolution des langages, des technologies ou des cultures d'une région. L'idée est de reconstruire, à partir d'observations actuelles sur les langues ou la culture, une histoire commune, sous la forme d'un arbre phylogénique. Ainsi chaque taxon est défini par la présence ou non de caractéristiques appelées cognats, permettant de l'écrire comme un vecteur binaire et ainsi de créer une phylogénie. L'approche bayésienne consiste à décrire comment évolue une chaîne de Markov se déplaçant sur l'arbre, partant de la racine, soit l'ancêtre commun le plus récent aux taxons, en calibrant les paramètres d'un modèle par Monte-Carlo Markov Chain.
On prend l'exemple de 20 langages du Pacifique, avec un lexique de 205 mots, chacun décrit par un certain nombre de cognats, allant de 2 à 15. On précise ensuite les caractéristiques du modèle utilisé, et on procède à un MCMC sur les paramètres, qui permet donc d'obtenir un échantillon d'arbres phylogéniques. On peut représenter leur répartition grâce au logiciel DensiTree :

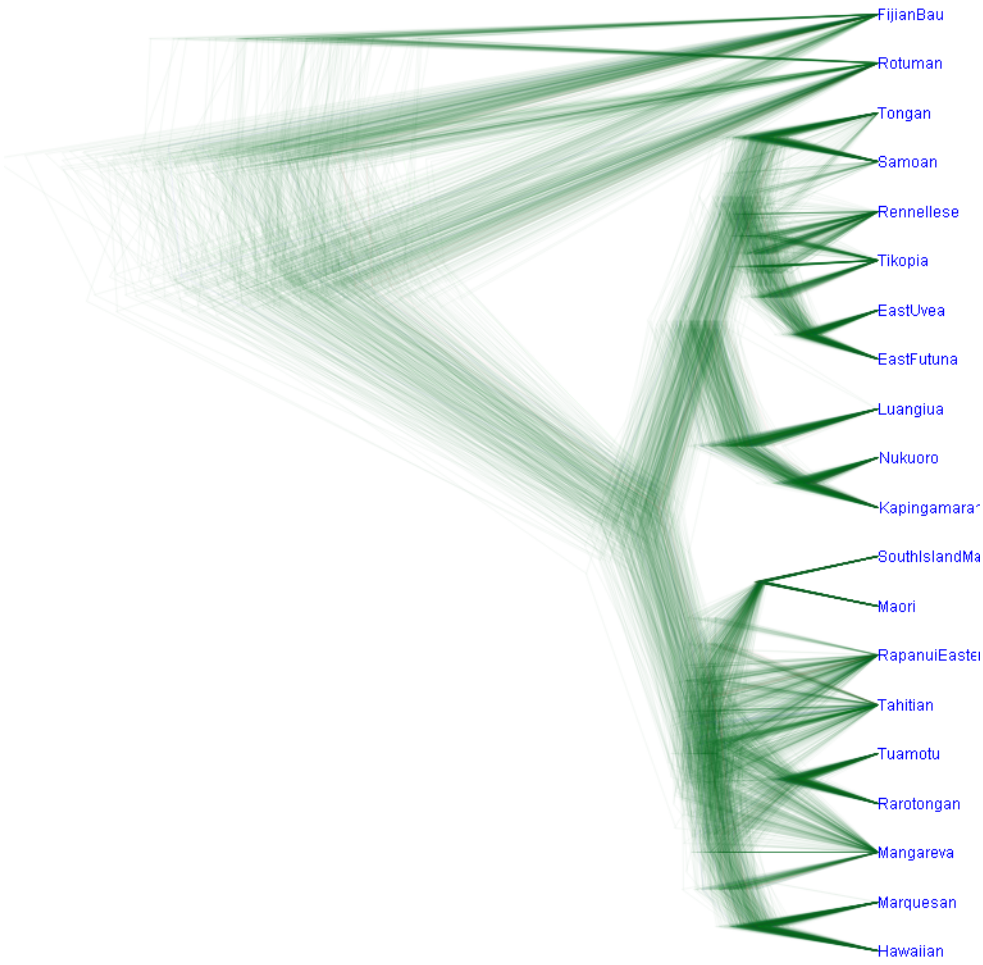


FIGURE 2 – DensiTree Languages Pacifique

Ainsi, en maximisant la posterior, on peut reconstruire une phylogénie, estimer l'âge des bifurcations sur l'arbre et l'âge de l'ancêtre commun le plus récent.

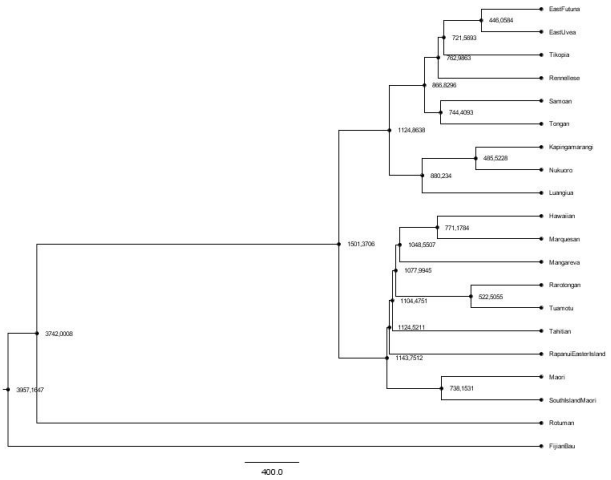


FIGURE 3 – MCC Languages Pacifique

On peut ensuite comparer les résultats en fonction du modèle choisi en calculant leur vraisemblance marginale, ou en testant les relations monophylétiques entre les taxons.

Ces méthodes, très efficaces pour construire une phylogénie présentent plusieurs limites majeures, à commencer par

leur coût en calcul. Les arbres sont décrits par de très nombreux paramètres, ce qui implique un coût élevé de calcul de la posterior et donc un temps de calcul des MCMC très long, d'autant plus que de très nombreuses itérations sont nécessaires pour obtenir un échantillon représentatif, avec un burn-in important.

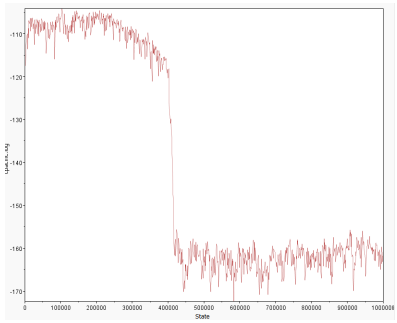


FIGURE 4 – Prior Trace Pacifique

Une autre limite touche le modèle d'arbre binaire lui-même : l'évolution de traits culturels ou linguistiques ne se fait pas uniquement de manière verticale, il est tout à fait raisonnable d'imaginer, par exemple pour des données linguistiques, un transfert horizontal entre deux langages distants sur l'arbre suite à une rencontre entre leurs peuples respectifs.

Dans l'article *Triangulation supports agricultural spread of the transeurasian languages* [9], les auteurs procèdent à trois analyses bayésiennes de ce type sur trois jeux de données : archéologiques, linguistiques et génétiques. En recoupant les arbres construits par MCMC, ils proposent une histoire complète des peuples asiatiques anciens. Cependant, certains doutes subsistent quant à la validité de l'analyse et des conclusions tirées des MCMC. De même, le modèle d'arbre binaire est à interroger. Il est par exemple facile d'imaginer de nombreux transferts horizontaux pour des technologies, particulièrement entre des peuples proches géographiquement.

On se propose donc de reproduire les expériences sur la base de données archéologiques faites dans l'article et, en utilisant plusieurs outils bayésiens mais aussi topologiques, de remettre en cause les conclusions des auteurs. On procède ensuite à plusieurs analyses de clustering pour tenter de trouver des relations entre les données.

Toute la bibliographie, les scripts R, les figures ainsi que les fichiers XML sont retrouvables sur le **Github du mémoire**.

Table des matières

1 Modèle utilisé et méthode de simulation

La base de données archéologiques [9] est composée de 255 sites de fouilles, allant des rives du fleuve Amour jusqu'au Japon, pour lesquels sont indiqués la présence ou non d'une technologie. On trouve ainsi 171 technologies regroupées en six catégories différentes : céramiques, tombeaux, bâtiments, restes de nourriture, objets faits de coquillages et d'os, outils en pierre. Pour chaque site est indiqué son âge approximatif grâce à une datation au carbone 14, ainsi que le grand groupe culturel auquel il appartient.

Ainsi, par exemple, au site Xilongwa situé dans le West Liao, appartenant au groupe culturel Xinglongwa et datant de 7800 avant notre ère, on retrouve entre autres des céramiques sur lesquels sont peintes des motifs en Z à la fois horizontaux et verticaux, mais aussi des outils en obsidienne, des restes de millet ou encore des cerceuils en pierre.

Ces données permettent donc la reconstitution d'une matrice binaire avec en lignes les lieux et en colonne les technologies.

Ces données binaires permettent une approche bayésienne pylogénique. L'idée est de partir de la bien connue relation de Bayes

$$P(T|D,\theta) \propto P(D|T,\theta)P(T,\theta)$$

où T est un arbre binaire, D les données et θ les paramètres du modèle d'arbre [4].
Ce modèle d'abre est caractérisé par un processus de Markov se déplaçant sur l'arbre et décrit par une matrice stochastique infinétisimale $Q = (q_{ij})_{ij}$, telle que si l'état i est à 0 et j à 1, q_{ij} est la probabilité d'acquérir le cognat (ici la technologie) et q_{ji} de le perdre. On rend le processus ensuite dépendant au temps avec la matrice $P(t) = exp(Qt)$.

Le premier modèle considéré est le modèle covarion. Ce modèle ajoute aux variables binaires visibles décrivant l'état du site des variables cachées décrivant si le site est dans en transition rapide ou lente. Il y a donc 4 entrées : 0-lent, 0-rapide, 1-lent, 1-rapide, et on note (f_0, f_1) la fréquence de chaque état, indépendamment de la vitesse de transition. On obtient ainsi la matrice :

$$Q = \begin{pmatrix} . & f_1 & sf_0 & 0 \\ 1 & . & 0 & sf1 \\ sf0 & 0 & . & \alpha f_1 \\ 0 & sf_1 & \alpha f_0 & . \end{pmatrix} \tag{1}$$

où s décrit la probabilité de vitesse de transition, α décrit le taux de mutation (fixé à 1 en état rapide).

Le second modèle est le modèle Pseudo-Dollo. L'idée est que chaque caractéristique peut être gagnée à un taux λ ou être perdue définitivement à un taux μ. Il y a donc trois états : initial, présent, perdu. La matrice obtenue est ainsi :

$$Q = \begin{pmatrix} . & \lambda & 0 \\ 0 & . & \mu \\ 0 & 0 & . \end{pmatrix} \tag{2}$$

Le modèle utilisé dans notre approche est un compromis entre les deux précédents : le modèle Pseudo-Dollo Covarion [1]. L'idée est de garder trois états (initial, présent, perdu), et d'y ajouter la notion de vitesse de transition, avec l'absence de vitesse pour l'état perdu, puisque la caractéristique est supposée perdue à jamais. On a donc cinq états : rapide-présent, rapide-initial, perdu, lent-présent, lent-initial, et on obtient la matrice :

$$Q = \begin{pmatrix} . & \lambda & 0 & s & 0 \\ 0 & . & \mu & 0 & s \\ 0 & 0 & . & 0 & 0 \\ s & 0 & 0 & . & \alpha \\ 0 & s & 0 & 0 & . \end{pmatrix} \tag{3}$$

A ce modèle, on y ajoute la notion d'horloge, c'est à dire d'évolution du taux de mutation dans le temps et dans l'arbre. En effet, il est raisonnable de supposer que les mutations n'ont pas exactement le même taux sur chaque branche de l'arbre. Ainsi on multiplie le taux de mutation d'une branche i par une valeur c_i qui suit une loi log-normale d'espérance 1 et d'écart-type σ, qu'on choisit ici à 5.0E$^-$4.

Enfin pour le prior, le grand nombre de dates ainsi que le risque de sites manquants justifie l'utilisation du modèle dit de Coalescent Bayesian Skyline [3] qui autorise une certaine flexibilité.

Toutes ces caractéristiques peuvent être implémentées dans le logiciel Beauti, qui permet de charger la matrice binaire ainsi que l'âge des taxons. On peut ensuite spécifier le type de modèle d'arbre, le prior, l'horloge et les carcatéristiques du MCMC souhaités. Dans notre cas, on opte pour une méthode de parallel tempering à 4 chaînes avec un delta de température de 0.1.

Le logiciel Beauti renvoie un fichiel .XML qui contient toutes les informations pour lancer les simulations.

C'est ensuite le logiciel BEAST qui s'occupe de cette longue tâche. L'algorithme utilisé est une méthode de Metropolis-Hastings. L'idée est d'obtenir un tirage aléatoires des nombreux paramètres décrits ci-dessus avec la méthode suivante :

1. On choisit un état initial θ_0.
2. On propose à l'itération t un nouvel état x' obtenu d'une distribution $g(x'|x_t)$, telle que $g(x'|x_t) = g(x_t|x')$.
3. On calcule la vraisemblance $P(D|x')$ et le prior $P(x')$ et on calcule le taux d'acceptance :

$$R = \min\Big[\frac{P(D|x')P(x')g(x_t|x')}{P(D|x_t)P(x_t)g(x'|x_t)}, 1\Big].$$

4. On tire aléatoirement $u \in [0, 1]$.
5. Si $u \leq R$, on accepte le candidat en posant $x_{t+1} = x'$.
6. Si $u > R$, on rejete le candidat en posant $x_{t+1} = x_t$.

La limite de cette méthode est que le taux d'acceptance peut être assez bas, et une chaîne peut stagner autour d'une valeur et pas suffisamment explorer toute la distribution postérieure. C'est pourquoi on utilise une technique de parallel tempering [6], en simulant plusieurs chaînes de Markov, et en introduisant pour chaque chaîne un paramètre de température $\beta_i = \frac{1}{1+(i-1)\Delta t}$, où Δt est un paramètre fixé.
On adapte ensuite pour chaque chaîne le taux d'adaptation en utilisant la valeur

$$R_{heated} = \min\Big[\big(\frac{P(D|x')P(x')g(x_t|x')}{P(D|x_t)P(x_t)g(x'|x_t)}\big)^{\beta_i}\frac{g(x_t|x')}{g(x'|x_t)}, 1\Big].$$

Ainsi, plus le paramètre de température β_i est faible, plus les distributions des paramètres sont applaties, ce qui diminue la concentration des valeurs simulées en un point. On procède enfin à des échanges entre les différentes chaînes avec une méthode d'accepation-rejet avec un taux d'acceptance

$$R_{ij} = \min\Big[\frac{P(x_i|D)^{\beta_j}P(x_j|D)^{\beta_i}}{P(x_i|D)^{\beta_i}P(x_j|D)^{\beta_j}}, 1\Big].$$

Dans notre cas, on opte pour une méthode de parallel tempering à 4 chaînes avec un delta de température de 0.1.

Dans notre étude, il est apparu nécessaire de faire des MCMC d'au moins 5 000 000 itérations. En sortie de logiciel, on possède un fichier .log qui contient la chaîne pour chaque paramètre, ainsi qu'un fichier .trees qui contient les arbres simulés.

Dans notre cas, la base de données fournies par les auteurs sur leur site n'étant pas la bonne, il a d'abord fallu retraiter ces données, en ajoutant les données manquantes, et créer un fichier XML avec les âges de chaque site en prenant la datation carbone moyenne pour chaque site.
Le parallel tempering couplé au modèle pseudo-Dollo covarion nécessita de créer directement les fichiers XML pour Beauti, en précisant donc les dates, le modèle, le prior, l'horloge, le type de MCMC et les fichiers de sortie. Cette tâche s'avéra difficile tant le nombre de paramètres est élevé et les fichiers XML suivent des règles bien précises pour pouvoir être lus par le logiciel BEAST. Tous les fichiers XML sont déposés sur le GitHub du papier.
On peut finalement lancer l'analyse dans le logiciel BEAST. Le temps de simulation est très long, chaque MCMC prenant plusieurs heures à lui seul.

2 Analyse

La nature des données permet de s'interroger sur la pertinence du modèle d'arbre. Les auteurs du papier se basent sur un arbre dit de Maximum Clade Credibility obtenu à partir du MCMC sur l'ensemble des données. Si ce modèle est consistant, on s'attend à pouvoir reconstruire une histoire similaire avec les seules données céramiques, nourriture, etc... C'est pourquoi notre analyse se base sur quatre MCMC : un avec toutes les données, un avec les données céramiques, un avec les données outils et un avec les données restantes mises ensembles (bâtiments, tombeaux, coquillages, nourriture). Pour chacune de ces dernières, la petite taille de l'échantillon conduit à des MCMC avec une trop grande variance, ce qui justfie la mise en commun des quatre groupes pour obtenir un échantillon de taille similaire aux autres.

Pour le groupe avec toutes les données, 50 000 000 itérations ont été réalisées, 5 000 000 dans les deux autres.

Pour pouvoir analyser les fichiers MCMC, il faut d'abord traîter les fichiers .log dans le logiciel LogCombiner, avant de les ouvrir dans le logiciel Tracer qui permet de visualiser les MCMC pour chaque paramètre. On peut observer que chaque MCMC a convergé, et qu'un burn-in d'environ 20% est nécessaire.

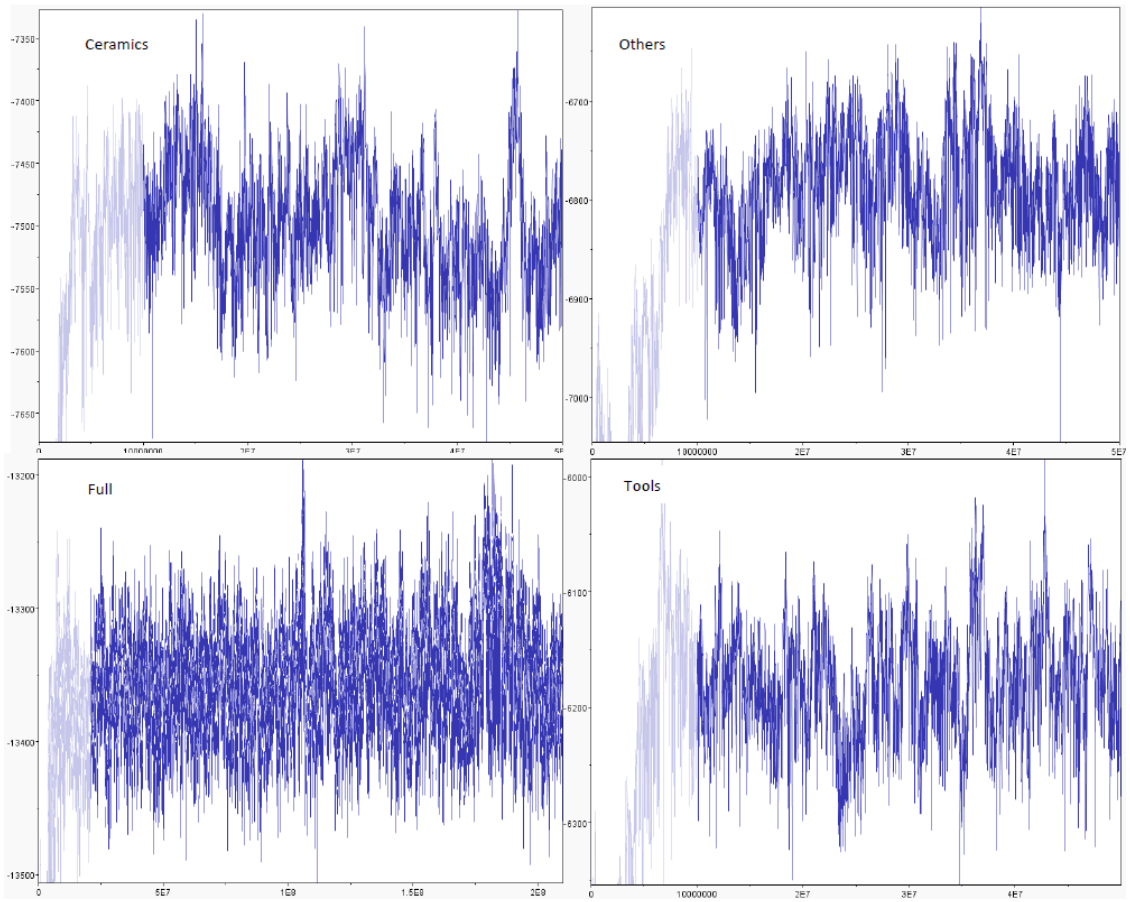


FIGURE 5 – Posterior Trace

2.1 Topologie des arbres

L'idée est de calculer par Monte Carlo la distance entre les distributions d'arbre. Ainsi, si

$$T_1 \sim \pi_i,\ T_2 \sim \pi_j,$$

alors on peut calculer par Monte-Carlo :

$$\mathbb{E}_{ij}[d(T_1, T_2)]$$

mais aussi

$$\mathbb{E}_{ii}[d(T_1, T_2)]$$

ce qui pourrait nous donner une sorte d'idée de variance. La question maintenant est la métrique d utilisée.
On utilise sur le package R ape la fonction treedist, qui nous donne trois métriques différentes :

Définition 2.1.1 (Symmetric Difference, Unweighted Robinson Foulds)**.** *La différence symmétrique [5] compte le nombre de branches dans T1 qui définissent une bifurcation absente de T2, plus le nombre de branches dans T2 qui définissent une bifurcation absente de T1.*

Définition 2.1.2 (Branch Score Difference)**.** *On construit pour chaque arbre la liste de toutes les partitions possibles. On donne ensuite à chacque branche de la liste un score : 0 si la branche est dans l'arbre, la longueur de la branche sinon. Enfin, on fait la somme sur les branches de la différence quadratique entre les scores.*
Ainsi, si les deux arbres ont la branche $\{A, D \mid B, C, E\}$, la somme contient la différence quadratique entre les longueurs de cette branche sur chaque arbre. Si un seul arbre possède cette branche, la somme contient le carré de la longueur de la branche sur cet arbre. Si les deux arbres n'ont pas cette branche, rien n'est ajouté à la somme car la différence vaut alors 0 (voir [5]).

Définition 2.1.3 (Path Difference)**.** *Soit Σ un ensemble de n taxons et soit $T1$ et $T2$ deux arbres phylogéniques sur Σ.*
La difference-distance [11] d'une paire de feuilles i et j, où $i \neq j$ et $i, j \in \Sigma$, sur $T1$ et $T2$ est $|d(T1, i, j) - d(T2, i, j)|$.
Quelque soit l'entier $p \leq 1$, la distance l_p-path-difference entre T1 et T2, notée $\delta_p(T1, T2)$ est :

$$\delta_p(T1, T2) = \Big(\sum_{i \neq j \ et \ i,j \in \Sigma} |d(T1, i, j) - d(T2, i, j)|^p \Big)^{\frac{1}{p}}$$

Les résultats du calcul par Monte-Carlo de la distance moyenne entre les arbres sont donnés dans les tableaux suivants :

	Full	Ceramics	Tools	Others
Full	375.2			
Ceramics	472.3	440.9		
Tools	488.0	497.9	475.9	
Others	479.6	498.3	501.9	453.6

Table 1 – Symetric Difference

	Full	Ceramics	Tools	Others
Full	13104.6			
Ceramics	15373.5	13816.2		
Tools	18112.9	17827.4	18158.3	
Others	18965.5	18970.1	20686.6	20031.5

Table 2 – Branch Score Difference

	Full	Ceramics	Tools	Others
Full	745.8			
Ceramics	1425.3	1208.7		
Tools	1564.7	1746.1	1321.8	
Others	1445.1	1672.3	1641.9	1117.6

TABLE 3 – Path Difference

Ces résultats ne sont pas très concluants pour conclure à une différence notable entre les arbres avec des écarts relatifs qui ne sont pas particulièrement marquants. On observe néanmoins une corrélation entre la taille de l'échantillon et la distance au sein d'un même groupe. Cette analyse, bien que nécessaire, ne permet donc pas de conclure sur la pertinence du modèle proposé.

2.2 Vraisemblance Marginale

La vraisemblance marginale semble être un outil intéréssant pour tester la pertinence du découpage des données en sous-groupes. En effet, en calculant la vraisemblance marginale de chaque jeu de données, on pourra comparer le modèle complet contre le modèle aux trois sous-groupes indépendants en calculant le facteur de Bayes.
Néanmoins, à cause du grand nombre de données sur lesquels on travaille ainsi que leur nature, calculer la vraisemblance marginale peut s'avérer difficile. On utilise ici une méthode implémentée dans Beast : le Stepping-Stone Sampling [12] que l'on décrit ici.

Stepping-Stone Sampling

On pose d'abord $p_\beta = f(y|\theta, M)^\beta f(\theta|M)$ et

$$c_\beta = \int_\Theta p_\beta \mathrm{d}\theta$$
$$q_\beta = \frac{p_\beta}{c_\beta}.$$

où $f(y|\theta, M)^\beta$ est la vraisemblance et $f(\theta|M)$ le prior.
Ainsi, la vraiemblance marginale $f(y|M) = \frac{c_{1.0}}{c_{0.0}}$ puisque $c_{0.0}$=1.
On note ensuite

$$r_{SS} = \frac{c_{1.0}}{c_{0.0}} = \prod_{k=1}^{K} \frac{c_{\beta_k}}{c_{\beta_{k-1}}} = \prod_{k=1}^{K} r_{SS,k}$$

où $0 = \beta_0 < \beta_1 < ... < \beta_K = 1$.
L'idée est donc d'estimer les $r_{SS,k}$ par importance sampling en utilisant $p_{\beta_{k-1}}$ comme densité de sampling, puisqu'elle est proche de p_{beta_k} mais possède des queues de distributions plus denses. Ainsi

$$\begin{aligned} r_{SS,k} &= \frac{c_{\beta_k}}{c_{\beta_{k-1}}} \\ &= \int_\Theta \frac{f(y|\theta, M)^{\beta_k}}{f(y|\theta, M)^{\beta_{k-1}}} p_{\beta_{k-1}}(\theta|y, M)\mathrm{d}\theta \end{aligned}$$

On estime cette quantité par Monte-Carlo :

$$\hat{r}_{SS,k} = \frac{1}{n}\sum_{i=0}^{n} f(y|M, \theta^i_{\beta_{k-1}})^{\beta_k - \beta_{k-1}}$$

où $(\theta^i_{\beta_{k-1}})_{i=1,..,n}$ est un échantillon de loi $q_{\beta_{k-1}}$ simulé par MCMC.
On obtient finalement un estimateur la vraisemblance maximale en faisant le produit des $r_{SS,k}$.

Pour implémenter cette méthode, il a fallu créer de nouveaux fichiers XML. On procède encore à une méthode de parallel tempering pour les MCMC intermédiaires pour obtenir les échantillons $(\theta^i_{\beta_{k-1}})_{i=1,..,n}$, avec $n = 10,000,000$ et un burn-in sur les MCMC de 50%. On utilise un stepping-stone à 8 pas dont les écarts sont obtenus par simulation d'une loi Beta(α,1), avec $\alpha = 0.3$ On implémente ensuite cette méthode dans Beast dont voici un exemple de sortie :

```
marginalLs[7 ] = -151838.48630185952

Step        theta         likelihood   contribution ESS
0            1            -9042.9878   0            27.6458
1            0.5982       -9451.3829   -3759.4728   49.0825
2            0.3258       -10146.4262  -2729.4043   50.3895
3            0.1548       -11324.3343  -1891.9636   32.3914
4            0.0593       -13528.9029  -1242.2819   36.7658
5            0.0154       -15652.8321  -677.7543    106.4912
6            0.0015       -17050.8844  -217.873     77.3172
7            0            -151838.4863 -43.6702     3.8985
sum(ESS) = 383.9819

marginal L estimate = -10562.420019646032


Total wall time: 15562 seconds
Done
```

On obtient les résultats suivants :

	ML	ESS
Full	-10562.42	384
Ceramics	-4317.22	377
Tools	-2750.16	403
Others	-3449.7	393
Ceramics x Tools x Others	-10517.08	

TABLE 4 – Vraisemblances Marginales

On obtient donc une valeur de Bayes $BF = -45.34$, en faveur du modèle aux données compartementées, et remet en cause la pertinence d'une analyse où les données archéologiques sont mises ensembles.

2.3 Arbres MCC

Les deux résultats précédents ne sont pas suffisants pour arriver à une conclusion, et ne nous donnent pas d'information sur la nature même des arbres phylogéniques simulés et la pertinence du modèle d'arbre. Il nous faut donc aller plus en détail, et le premier type d'arbre à observer est le maximum clade credibility (MCC).
Le principe du MCC est simple : sur chaque arbre simulé, on donne à chaque clade unscore basé sur le nombre d'occurences de celle-ci dans m'échantillon des arbres. On reconstruit ainsi un arbre qui maximise la posterior pour chaque clade. L'arbre obtenu est une manière rapide de résumer l'échantillon d'arbres obtenu après MCMC.
On utilise le logiciel TreeAnnotator, qui permet de concaténer les milliers d'arbre contenus dans le fichier .trees, en un seul arbre MCC avec un burn-in de 20 %, contenu dans un fichier .tree, qu'on upload ensuite dans R.
Nous avons donc tracé ensuite pour chaque simulation l'arbre MCC correspondant, en coloriant chaque feuille en fonction du groupe culturel auquel le taxon appartient. Les groupes culturels sont regroupés ainsi :

1. Yayoi (Japon, Rouge)
2. Mumun (Corée, Bleu)
3. Chulmun (Corée, Vert Olive)
4. Xiaozhushan (Liaodong, Orange)
5. Shuangtozi (Liaodong, Jaune)
6. Xiajiadian (Ouest Liao et Amur, Violet Sombre)
7. Hongshan (Ouest Liao et Amur)
8. Zuojiashan (Amur, Vert Clair)
9. Bajinbao (Amur, Cyan)
10. Xiaohexi (Ouest Liao, Doré)
11. Xinglongwa (Ouest Liao, Amur, Rivière Jaune, Corail)
12. Xinkailiu (Amur, Violet Clair)
13. Zaisanovka (Primorye, Turquoise)
14. Zhaobaogou (Ouest Liao, Rivière Jaune, Vert Foncé)
15. Reste des cultures (divers, Noir)

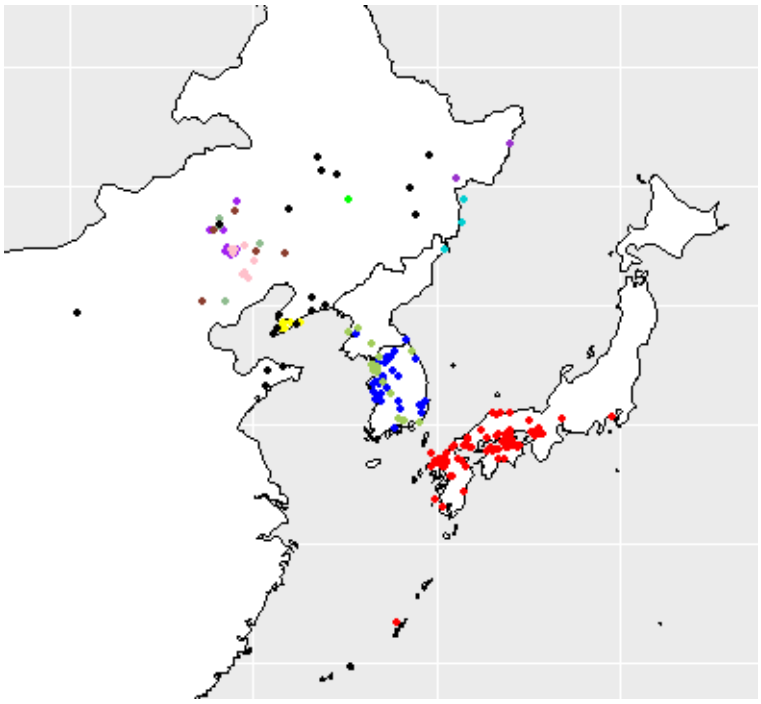


FIGURE 6 – Carte des Sites Archéologiques

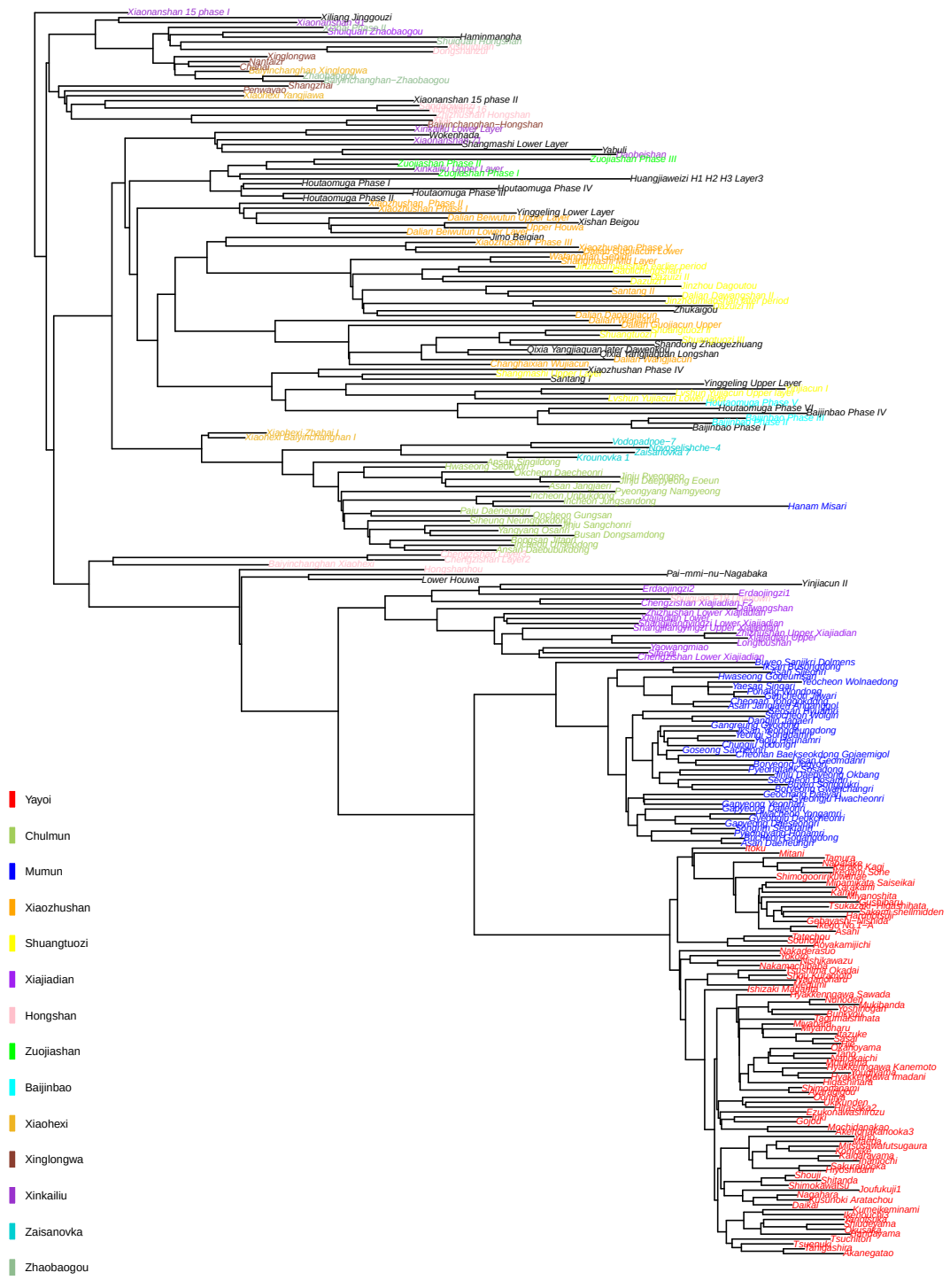


FIGURE 7 – MCC Données complètes

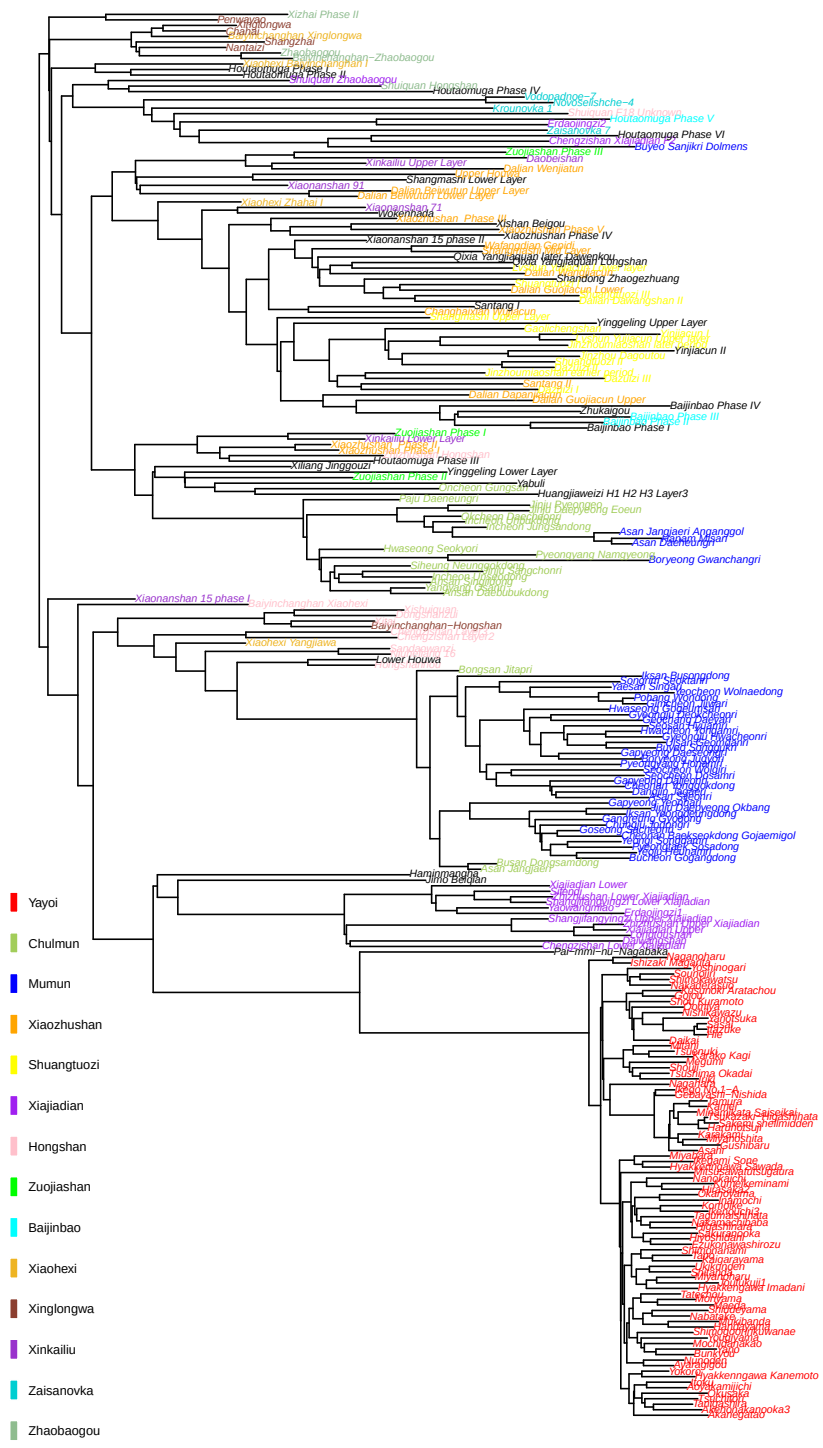


FIGURE 8 – MCC Céramiques

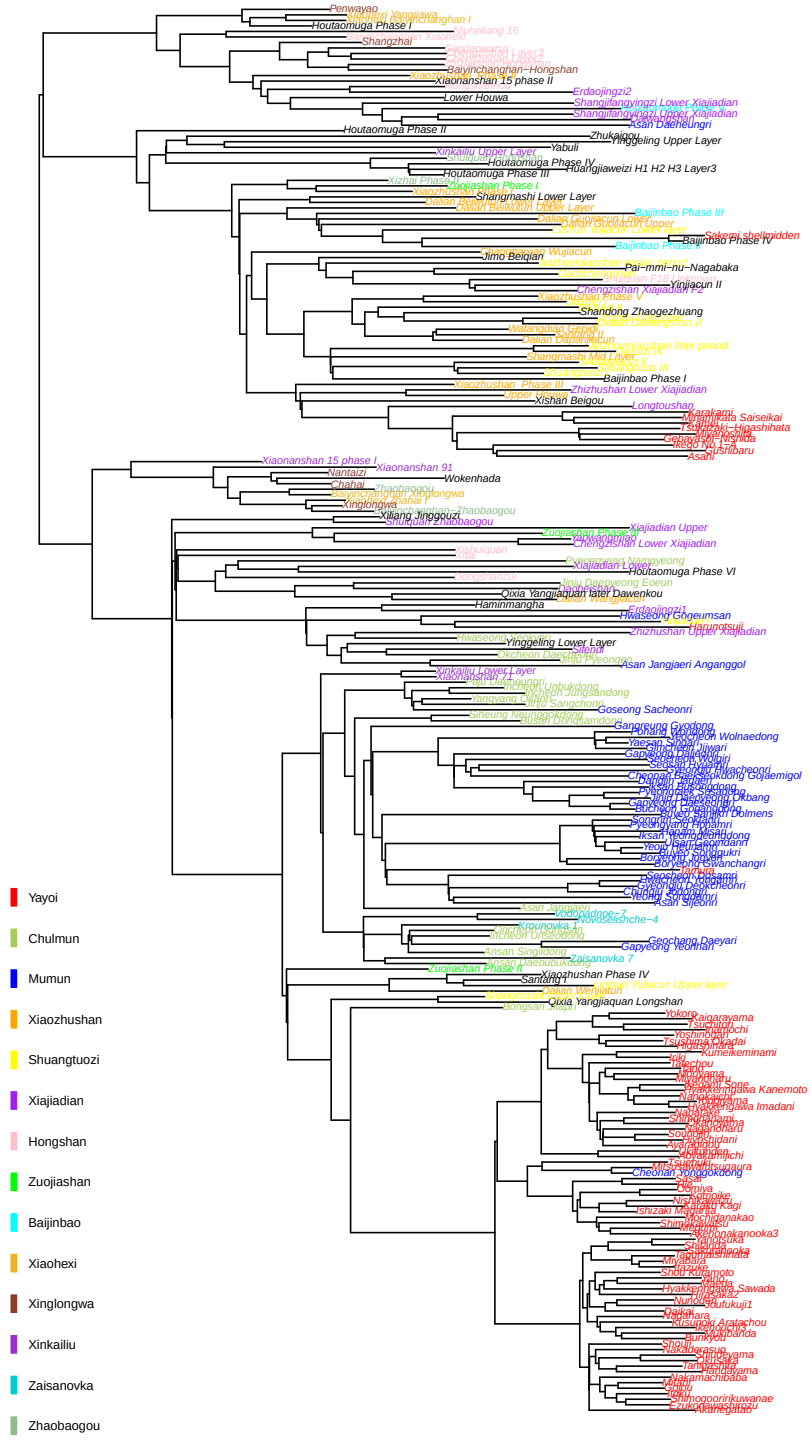


FIGURE 9 – MCC Outils

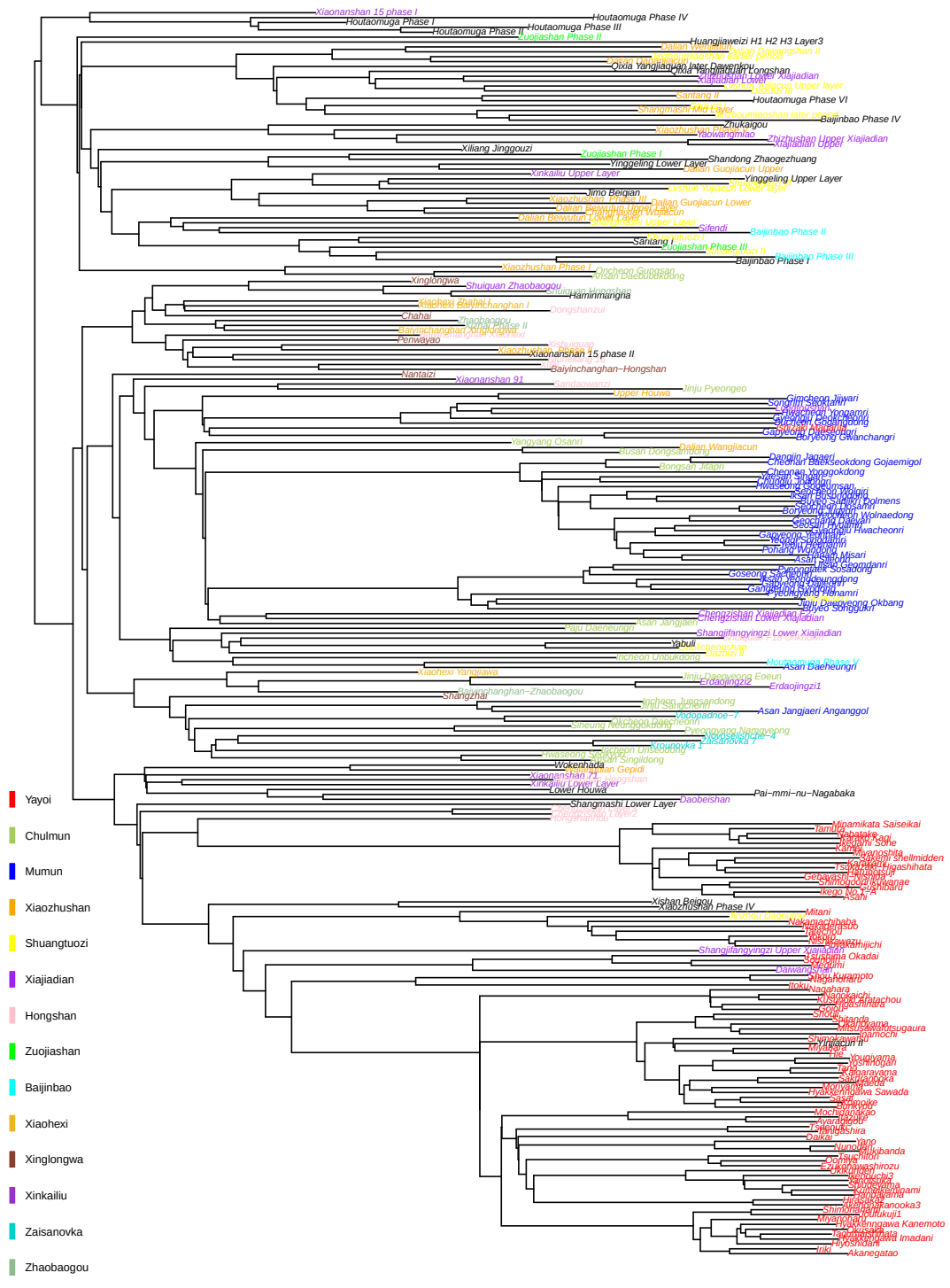


FIGURE 10 – MCC Autres

Le premier arbre tracé est le MCC des données complètes. Les auteurs se basent sur ce tracé pour établir leur analyse archéologique. Si l'on suppose comme eux que ce modèle d'arbre est le bon et qu'un MCC est un bon objet pour tirer des conclusions, on peut effectivement observer les conclusions du papier, à savoir des populations originaires des rives du fleuve Amour qui se sont réparties dans toute la péninsule du Liaodong, avec deux épisodes migratoires vers la Corée, un des Mumuns, qui s'est ensuite étendu au Japon pour créer la culture Yayoi, et un des Chulmuns.
C'est cela dit la pertinence de ce modèle que nous interrogeons, et la première étape est d'observer si les MCC sur les sous-groupes racontent la même histoire.
On observe sur l'arbre MCC Cérmamique une histoire assez similaire, ce qui est consistant avec les calculs topologiques faits plus haut, où la distance entre arbre complet et arbre céramique était toujours la plus faible.
Il est beaucoup plus difficile sur l'arbre MCC Outils d'identifier clairement la même histoire. Si les cultures Yayoi et Mumun sont à peu près au même endroit, on trouve des taxons de ces deux cultures complètement isolés, et les autres cultures ne dessinent aucune structure claire.

Enfin, L'histoire racontée sur l'arbre MCC Autres est ici assez chaotique, avec aucune culture clairement regroupée, à l'exception de la culture Yayoi, qui présentent tout de même des disparités notables. La culture Mumun est quant à elle regroupée, mais pas du tout située au même niveau sur l'arbre que pour les autres MCC. Enfin, la structure globale d'arbre est assez éloignée de l'arbre complet et plus chaotique.

La première approche serait de se dire que les écarts entre les arbres sont avant tout dûs au nombre de taxons, et que plus celui ci est faible, plus la variance est élevée. On a 171 taxons dans le modèle complet, 69 dans le modèle céramique, 37 dans le modèle outils, et 64 dans le modèle autre. Il n'est donc pas bien clair que le nombre de données suffise à expliquer ces arbres MCC différents puisque il y a à peu près le même nombre de taxons entre céramiques et autres pour des arbres très différents, et bien moins de données pour l'arbre outils et pourtant une structure assez similaire à l'arbre céramique et l'arbre complet.
Si chaque arbre technologique ne raconte pas exactement la même histoire, c'est peut-être d'avantage le modèle d'arbre qui est à interroger. On touche ici à la limite de l'arbre MCC, à savoir qu'on force les données à nous sortir un arbre phylogénique en maximisant la posterior, sans donner aucune information sur la qualité de ce modèle phylogénique ni sa vraisemblance.

2.4 Groupes Monophylétiques

Sur le MCC du modèle complet, on observe des groupes culturels identifiables et isolés. De plus, dans le papier, les auteurs regroupent ensemble les cultures Mumun et Yayoi, sur l'hypothèse tirée d'un arbre MCC qu'ils proviennent du même ancètre. Ainsi, plusieurs hypothèses monohylétiques sont faites pour tenter de reconstruire une histoire archéologique à partir d'une observation très visuelle sur un type d'arbre que nous avons remis en cause plus haut.

On propose donc de tester ces hypothèses monophylétiques sur 1000 arbres de chaque sous-groupe grâce à la fonction R is.monophyletic du package *ape*. Voici les résultats obtenus :

	Full	Ceramics	Tools	Others
Yayoi	0.997	0.984	0	0
Mumun	0	0	0	0
Chulmun	0	0	0	0
Xiaozhushan	0	0	0	0
Shuangtozi	0	0	0	0
Xiajiadian	0	0	0	0
Hongshan	0	0	0	0
Zuojiashan	0.21	0.012	0	0.051
Bajinbao	0	0	0	0
Xiaohexi	0	0	0	0.001
Xinglongwa	0	0	0	0
Xinkailiu	0	0	0	0
Zaisanovka	0.011	0	0.001	0.222
Zhaobaogou	0.002	0.006	0	0

TABLE 5 – Tests Monophylétiques

Comme supposé par les arbres MCC, peu de relations de parenté sont clairement identifiables, excepté pour la culture Yayoi dans le groupe complet et le groupe céramique.
En observant l'arbre MCC du modèle complet, on s'aperçoit de la présence du taxon Hanam Misari du groupe Mumun qui est isolé dans le groupe Chulmun, et en le retirant du groupe, on obtient les résultats suivants :

	Full	Ceramics	Tools	Others
Yayoi	0.997	0.984	0	0
Mumun	0.959	0	0	0
Chulmun	0.46	0	0	0
Yayoi x Mumun	0.709	0	0	0

TABLE 6 – Tests Monophylétiques sans taxon Hanam-Misari

Ces tests nous permettent de sérieusement remettre en cause la pertinence de ce modèle d'arbre phylogénique pour ces données archéologiques, puisque presque aucune relation de parenté apparaît clairement entre les cutures, à l'exception des données Coréennes et Japonaises dans le modèle complet.

2.5 Arbres de Consensus

On propose désormais de construire les arbres de consensus de chaque jeu de données. Un arbre de consensus est un arbre qui contient uniquement les clades qui ont une posterior supérieure à un certain seuil. L'avantage de cet arbre est qu'il n'est pas binaire et permet vraiment de visualiser à quel point les données suivent une structure phylogénique. Pour notre analyse, on a choisi un seuil de 50%. On crée les arbres consensus dans R grâce à la fonction consensus() du package *ape*.
On observe comme attendu sur l'arbre complet de bons résultats pour les cultures Yayoi, Mumun et Chulmun, autrement dit les groupes culturels de la péninsule Coréenne et du Japon. Les résultats sont assez similaires dans l'arbre céramique. En revanche, les structures phylogéniques sont inexistantes dans les deux autres arbres, et il est impossible d'identifier des groupes culturels distincts.

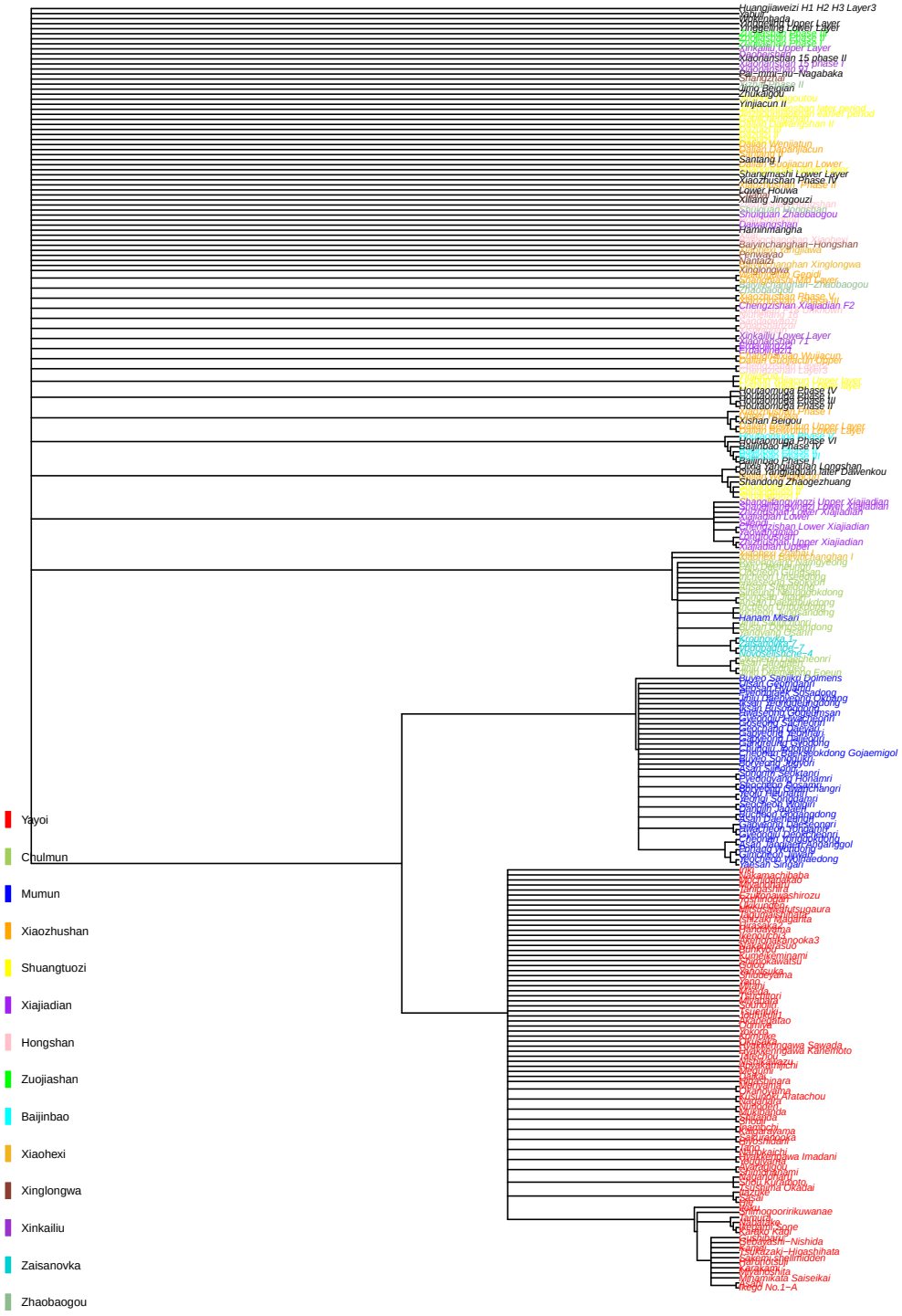


FIGURE 11 – Consensus Données Complètes

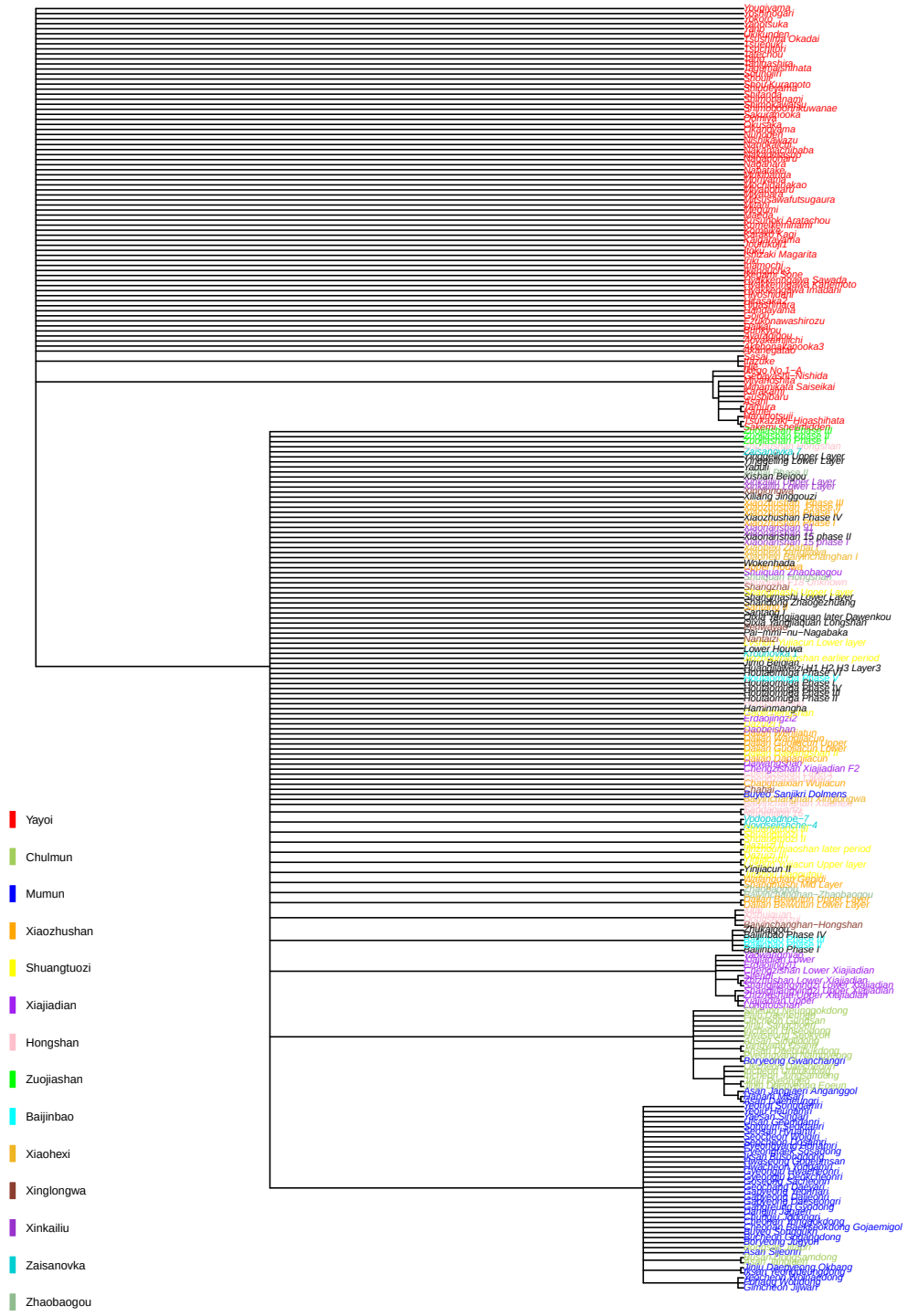


FIGURE 12 – Consensus Céramiques

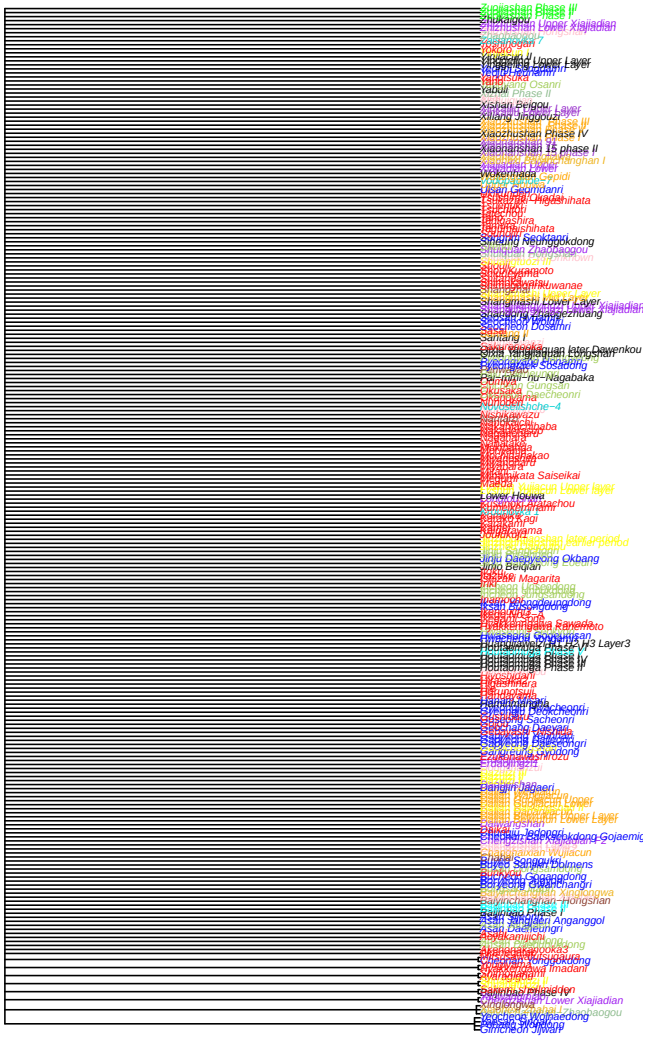


FIGURE 13 – Consensus Outils

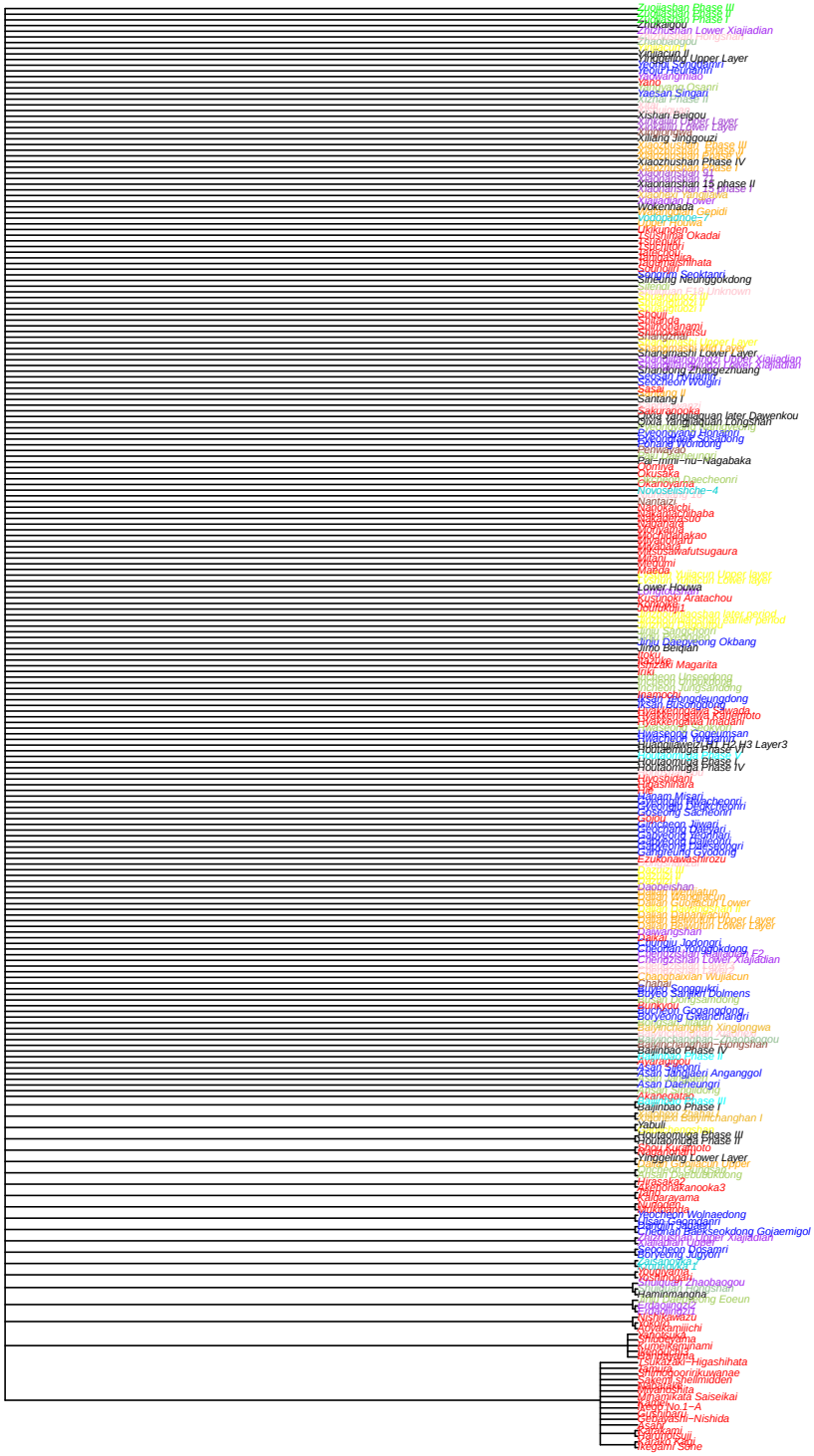


FIGURE 14 – Consensus Autres

L'ensemble des analyses faites à ce stade permet plusieurs constats :

1. Les résultats du calcul de la vraisemblance marginale et des tests monophylétiques semblent pointer vers une histoire qui diffère selon les technologies qu'on regarde. Ce constat est appuyé par la grande volatilité du groupe Autres, composé de familles technologiques regroupées, et qui malgré un nombre de données similaires au groupe céramique, ne présente aucune structure monophylétique ou phylogénique.
2. A l'exception des groupes culturels japonais et coréens, il est impossible avec ce modèle de conclure à quoi que ce soit sur les dynamiques de population de la péninsule du Liaodong. Les arbres reconstruit ne sont pas monophylétiques, et un modèle phylogéniques sur les technologies est insuffisant pour déterminer quoi que ce soit.
3. Les arbres de consensus semblent en faveur de l'hypothèse de deux épisodes migratoires vers la Corée que dessinent les arbres MCC : un premier épisode entre -6000 et -4000 ans donnant naissance à la culture Chulmun, suivi d'un second entre -3000 et -2500 ans donnant naissance à la culture Mumun et qui s'est prolongé vers le Japon entre -2700 et -1800, qui a vu naître la culture Yayoi. Notons d'ailleurs que cette hypothèse est principalement portée par les données céramiques, qui représentent en effet l'essentiel des trouvailles archéologiques faîtes dans la région.

3 Japon-Corée

On propose donc d'utiliser les mêmes outils d'analyse en se concentrant sur les données japonaises et coréennes, soit les cultures Yayoi, Mumun, Chulmun. On a donc effectué plusieurs MCMC sur Beast : Yayoi, Mumun, Chulmun, Yayoi-Mumun, Corée et Japon-Corée.

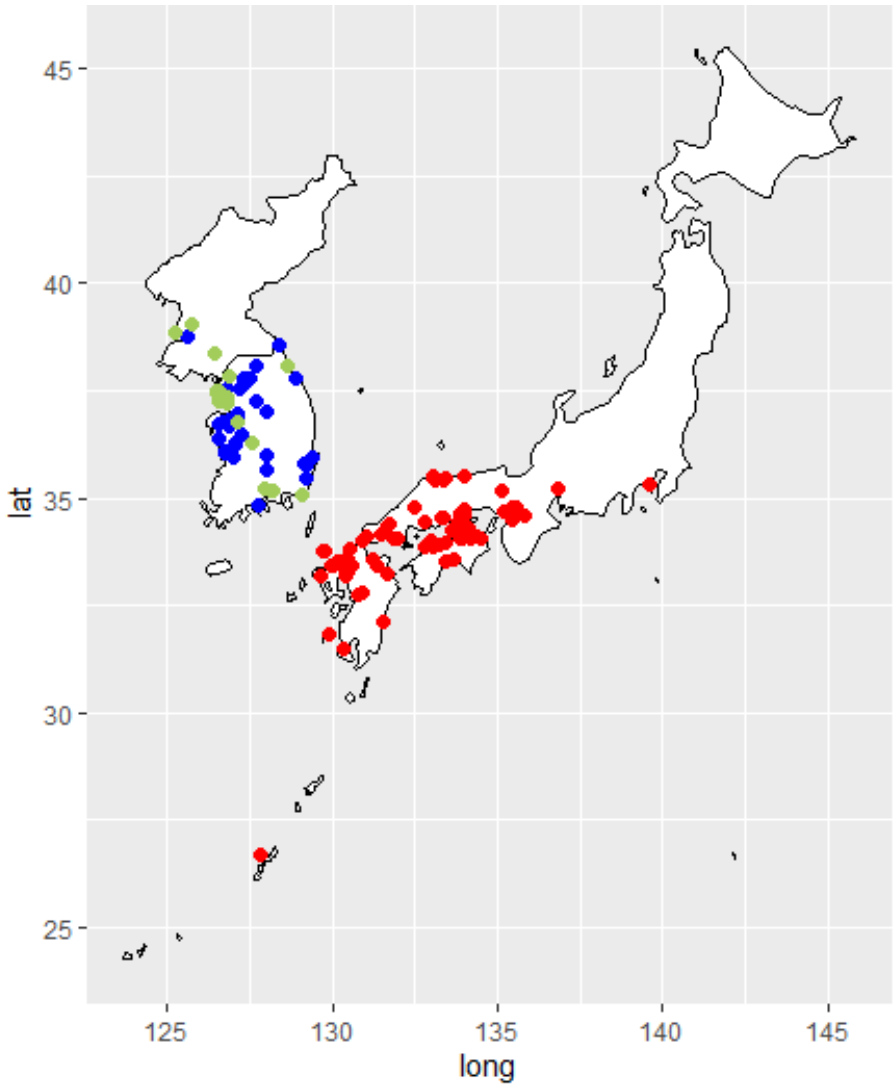


FIGURE 15 – Carte Sites Japon-Corée

3.1 Arbres MCMC

Voici les MCC interculturels obtenus :

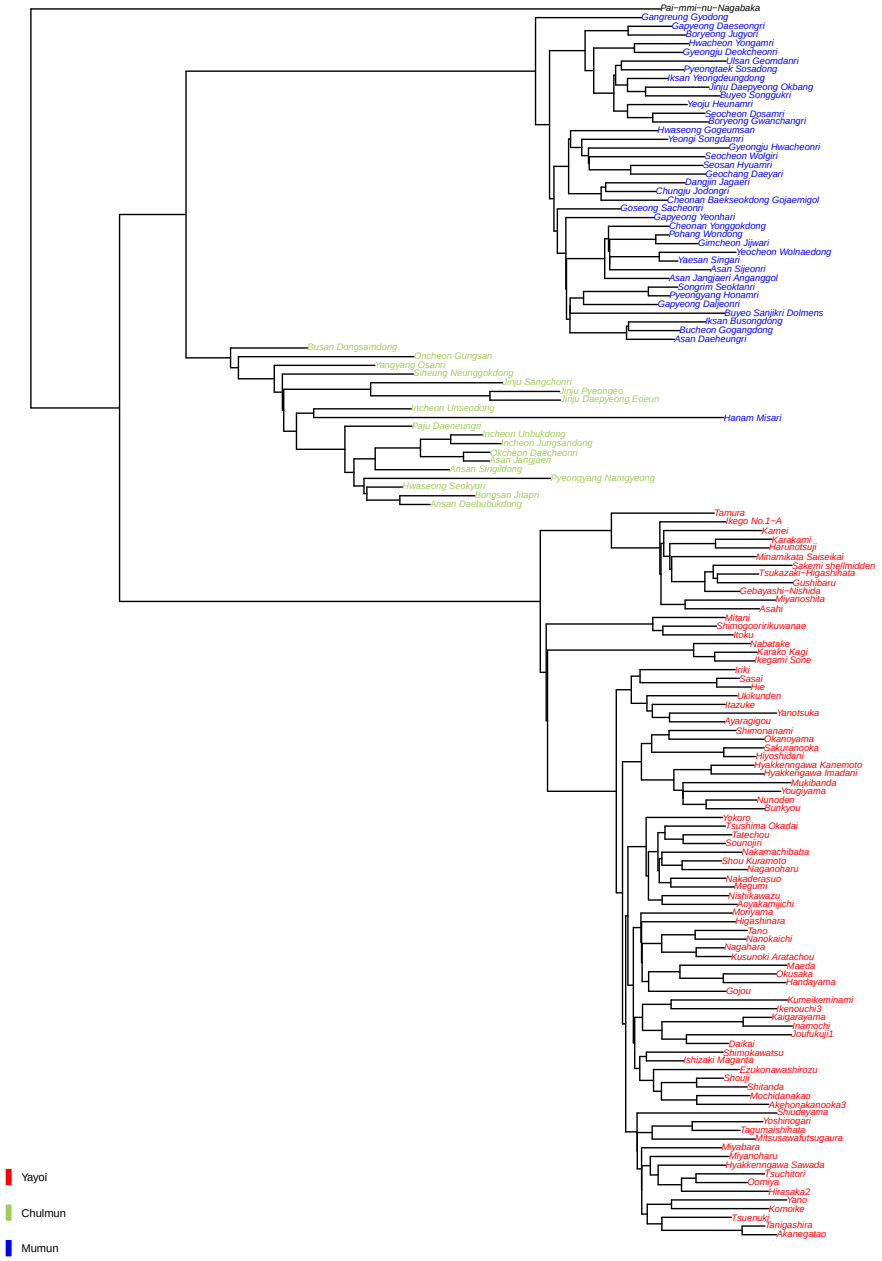


FIGURE 16 – MCC Japon Corée

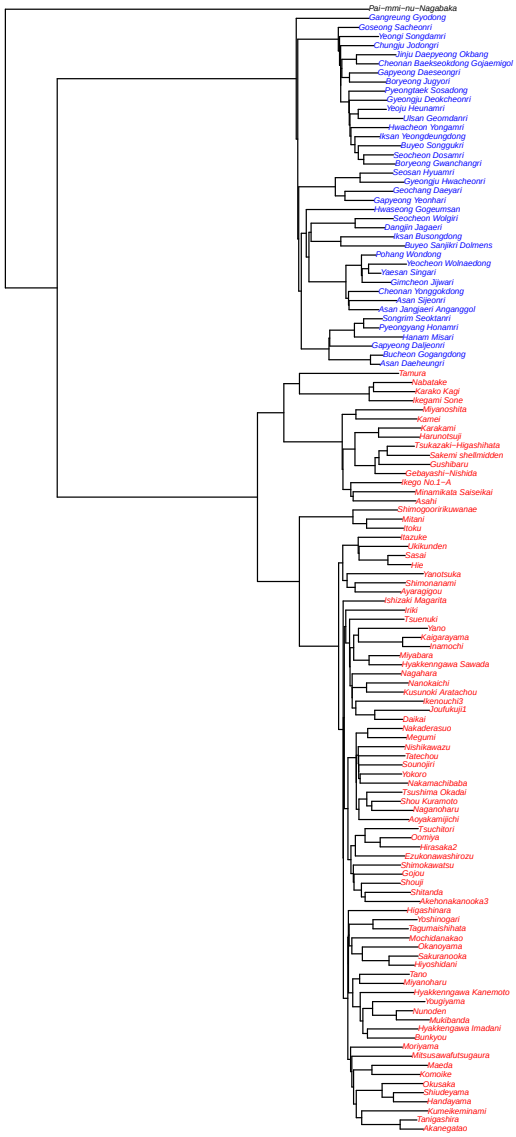


FIGURE 17 – MCC Yayoi-Mumun

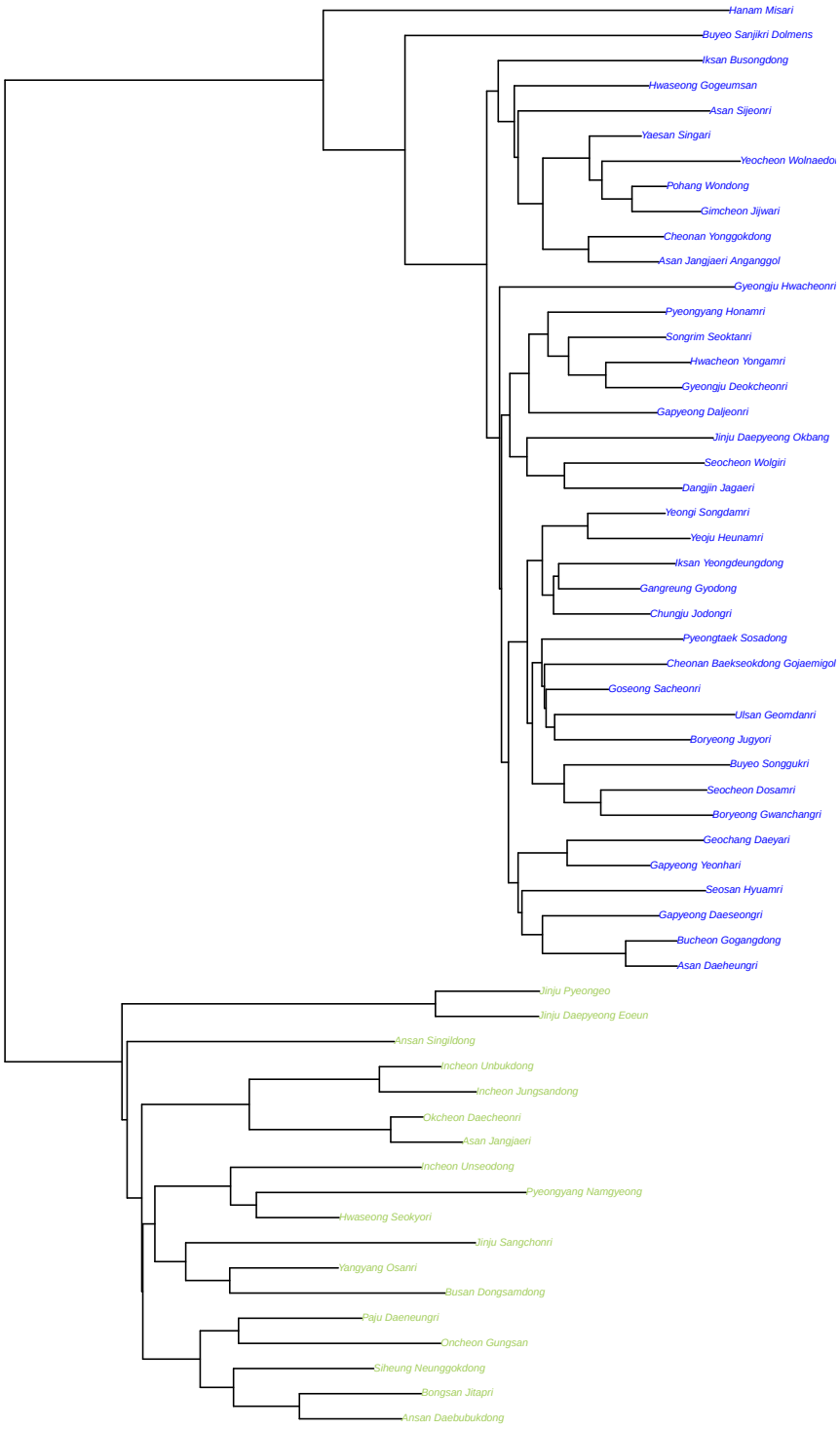


FIGURE 18 – MCC Corée

On retrouve bien les structures attendues, avec une vraie césure entre les cultures. On note néanmoins que les données coréennes forment un groupe distinct du groupe japonais, tandis que l'analyse plus haut montrait un groupe Yayoi-Mumun et un groupe Chulmun. De plus, on retrouve encore le taxon Hanam-Misari qui est isolé au milieu du groupe Chulmun.
Néanmoins, l'histoire racontée par cet arbre est différente de celle racontée par le modèle complet. En effet, il semble ici que l'on ait un groupe Corée et un groupe Japon distincts, alors qu'on avait précédemment un groupe Chulmun et un groupe Mumun-Yayoi. Cet arbre MCMC soutiendrait d'avantage l'hypothèse de deux migrations, une vers la Corée qui aurait donnée naissance à deux cultures, Chulmun et Mumun, et une seconde vers le Japon qui aurait donné naissance à la culture Yayoi.

3.2 Monophylétie et Vraisemblance Marginale

On teste l'impression visuelle donnée par les MCC en testant la monophylétie des groupes culturels au sein de l'arbre Japon-Corée, en mettant le taxon Hanam-Misari dans le groupe Chulmun pour ne pas fausser les résultats. De même, on associe le taxon PaimminuNagabaka situé dans les îles japonaises à la culture Yayoi.

Yayoi	Mumun	Chulmun	Yayoi-Mumun	Mumun-Chulmun	Japon-Corée
0.657	0.997	0.887	0.042	0.849	1

Table 7 – Tests Monophylétiques Japon-Corée

Ces résultats confirment l'impression visuelle donnée par les MCC, d'une bien meilleure relation monophylétique entre ces taxons, particulièrement dans les sous-groupes culturels et géographiques. On note aussi une meilleure relation monophylétique pour le groupe Chulmun-Mumun que pour le groupe Yayoi-Mumun.
On calcule ensuite la vraisemblance marginale pour chaque MCMC avec la même méthode stepping-stone sampling sur Beast.

	ML	ESS
Yayoi	-1918.64	526
Mumun	-1167.39	1169
Chulmun	-574.95	1398
Yayoi-Mumun	-3326.48	598
Mumun-Chulmun	-1907.19	1015
Yayoi-Mumun-Chulmun	-4088.15	489
Yayoi x Mumun x Chulmun	-3660.98	

Table 8 – Vraisemblances Marginales

Ces résultats, comme précédemment, tendent à favoriser le modèle qui sépare chaque culture avec un facteur de Bayes de -427.17.

3.3 Consensus

Enfin, on trace les arbres de consensus :

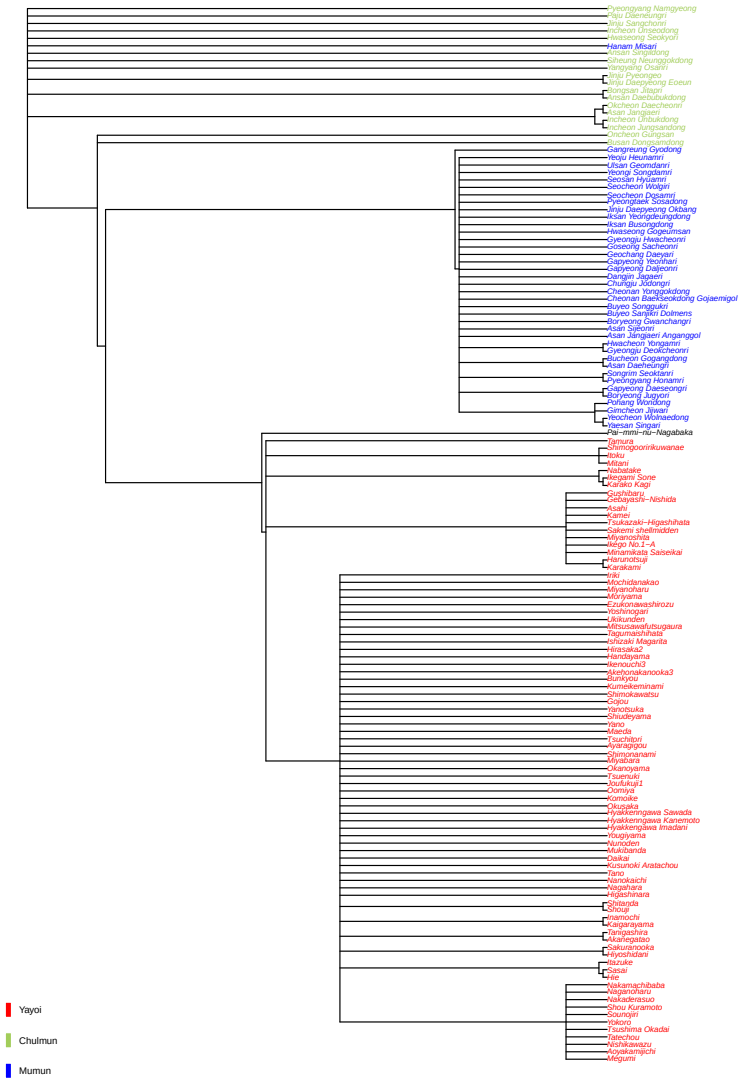


FIGURE 19 – Consensus Japon-Corée

Ce premier arbre de consensus est surprenant car il entre en contradiction avec le MCMC et les tests monophylétiques, puisqu'il supporte l'hypothèse initiale d'une première migration Chulmun vers la Corée, suivie d'une migration Mumun qui se serait étendue vers le Japon.

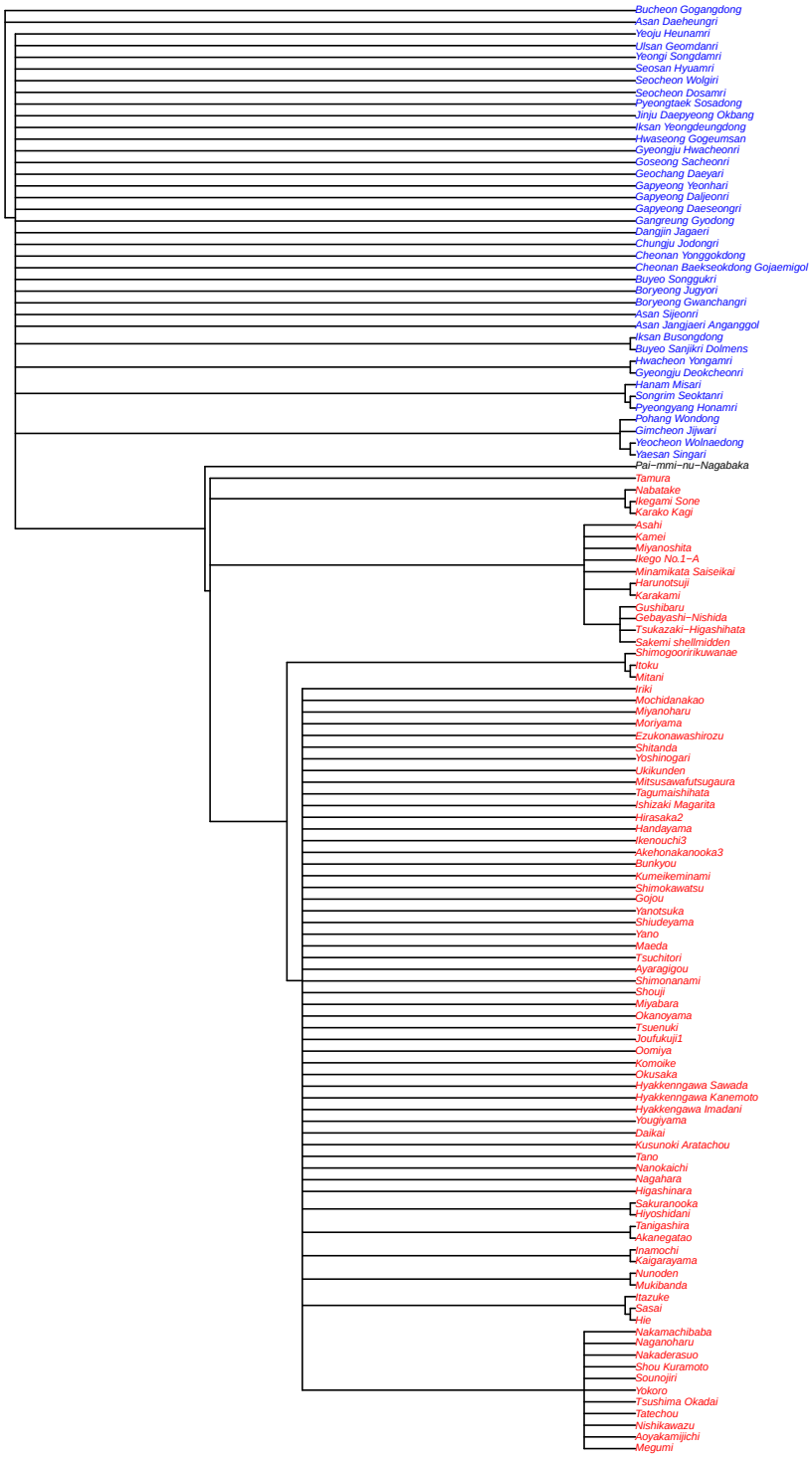


FIGURE 20 – Consensus Yayoi-Mumun

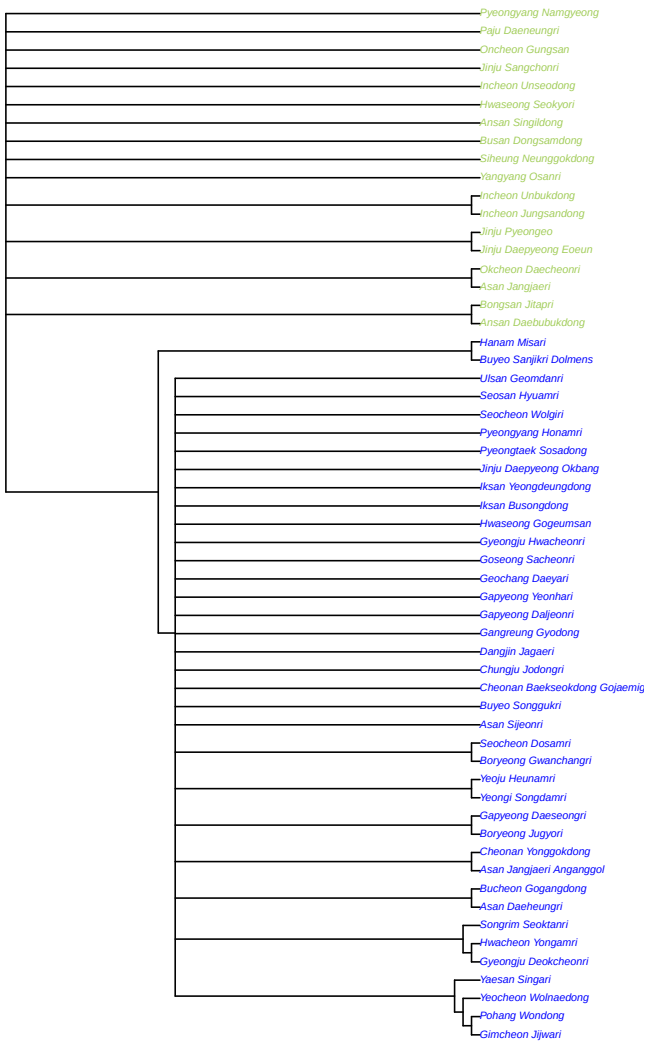


FIGURE 21 – Consensus Corée

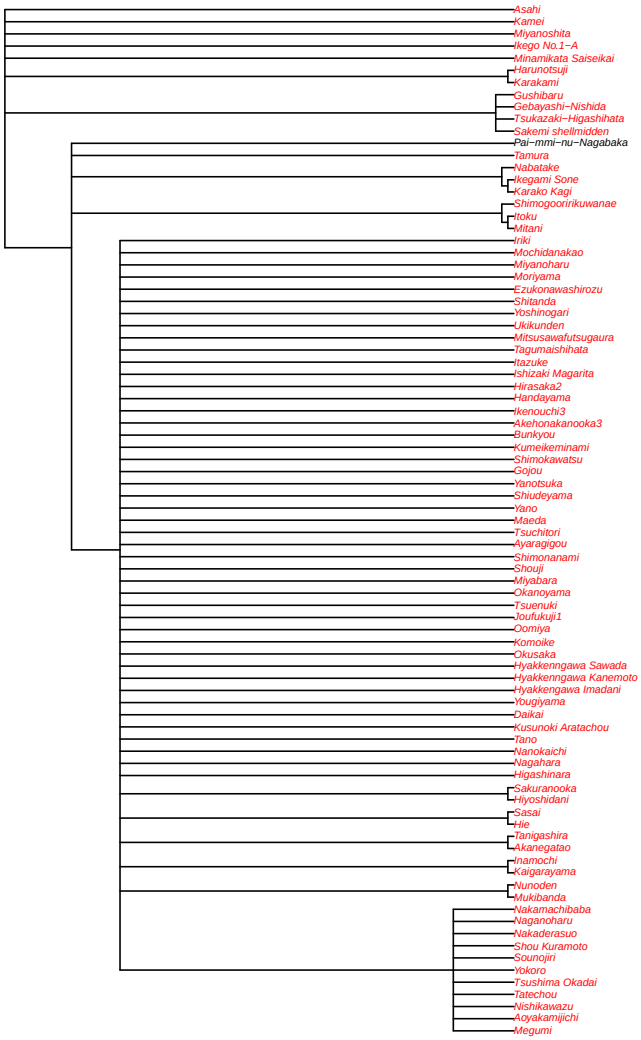


FIGURE 22 – Consensus Yayoi

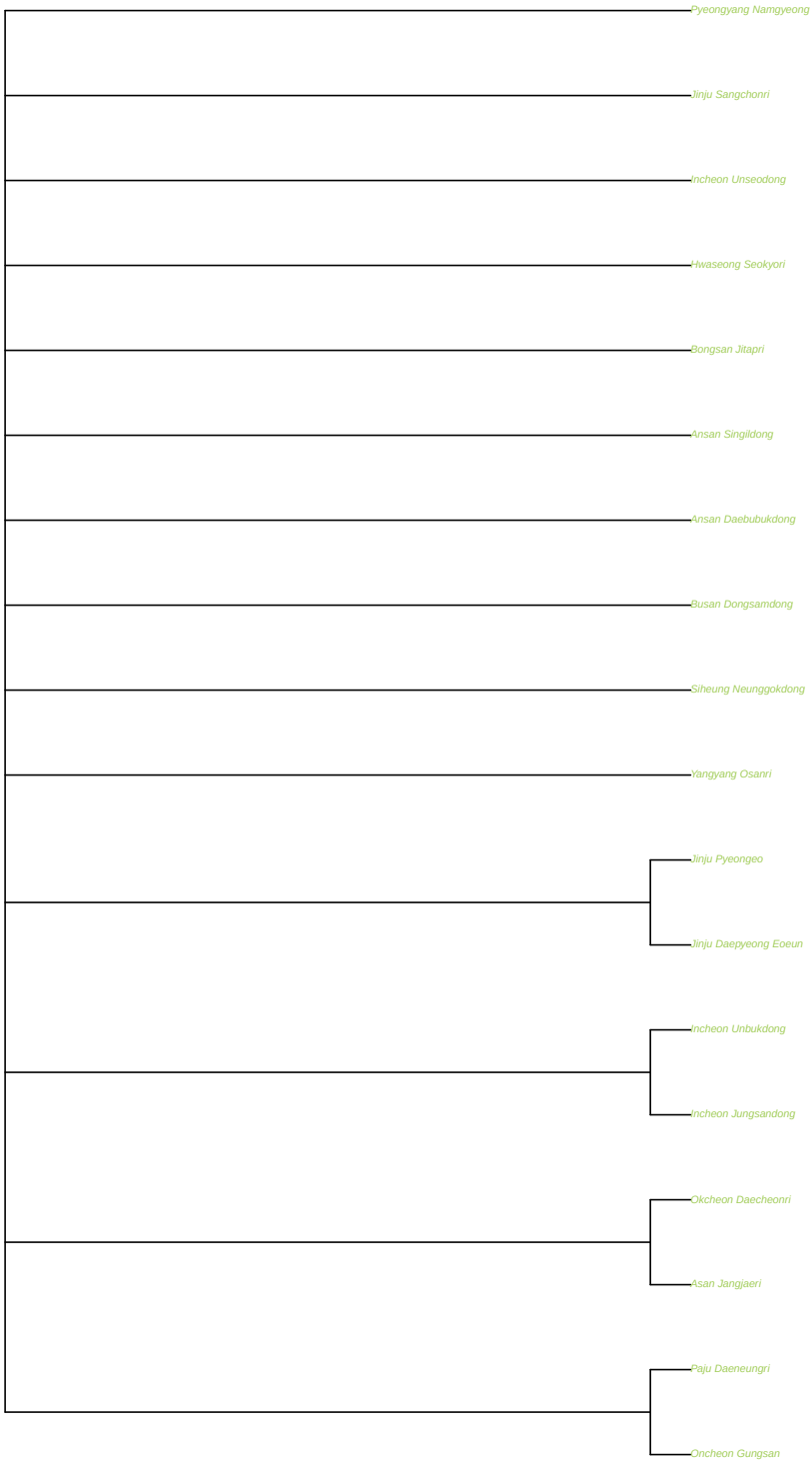


FIGURE 23 – Consensus Chulmun

FIGURE 24 – Consensus Mumun

Ces arbres de consensus sont assez déterminants puisqu'ils prouvent qu'il n'existe pas ou peu de relations phylogéniques au sein même des cultures. Cela remet donc en cause de manière assez définitive la pertinence de ce modèle d'arbres phylogéniques pour décrire l'évolution des technologies agricoles préhistoriques dans la péninsule du Liao.

4 Diffusion technologique

A ce stade, il apparaît clairement que l'approche phylogénie bayésienne telle qu'elle est proposée par les auteurs de [9] échoue à décrire l'évolution des peuples de l'Asie de l'Est. On a donc essayé plusieurs méthodes pour tenter d'expliquer la diffusion des différentes technologies dans la région.
Notre première approche a été de rester dans un cadre phylogénique Bayésien en s'intéréssant d'abord aux travaux de Nico Neureiter [7] sur la phylogéographie. L'idée de ce type de modèle est d'estimer la position des ancètres à partir de la position connue des taxons. On espère ainsi pouvoir représenter géographiquement l'évolution du langage, de la génétique ou, dans notre cas, des technologies. L'idée est de faire évoluer sur un arbre phylogénique à $2n$ feuilles un mouvement brownien X_t^i à deux dimensions qui à chaque noeud se sépare en deux autres browniens $X_t^{i_1}$ et $X_t^{i_2}$, avec comme contrainte que si $x_i \in \mathbb{R}^2$ désigne les coordonnées de la feuille $i \in \{1, ...2n\}$, et T_i sa date estimée, $X_{T_i}^i = x_i$.
Cette méthode nécessite néanmoins une phylogénie sur laquelle faire évoluer les mouvements browniens et on a vu précédemment qu'on ne possédait pas de résultats fiables sur la relation phylogénique de ces données archéologiques.

4.1 Analyse en Composantes Principales

On s'inspire ici des travaux de John Novembre et co. [8] qui proposent une Analyse en Composantes Principales (ACP) sur un ensemble de génomes prélevés sur des populations européennes. En projetant ces données sur le plan et en procédant à une rotation, les auteurs parviennent à une reconstruction assez claire du continent européen. On procède ici à deux types d'ACP : l'ACP exponentielle et l'ACP logistique.

On sppose qu'il y a n observations de vecteurs binaires de dimensions d, qu'on représente par une matrice $n \times d$ X, telle que chaque élément x_{ij} est une Bernouilli de probabilité p_{ij}.
On pose de plus

$$\theta_{ij} = \log \frac{p_{ij}}{1 - p_{ij}}$$

ACP Exponentielle
Cette méthode à été développée par Collins et co. [2]. L'idée est de trouver une base $v_1, ..., v_l \in \mathbb{R}^d$ et des vecteurs $(\hat{\theta}_i)_{i=1,...,n}$ avec $\hat{\theta}_i = \sum_k a_{ik} v_k$, tels que $\hat{\theta}_i$ soit "proche" de x_i.
On note $V \in \mathbb{R}^{l \times d}$ la matrice dont la k-ième ligne est v_k et $A \in \mathbb{R}^{n \times l}$ la matrice composée des a_{ik}.
Le but est donc de minimiser la fonction

$$L(V|A) = -\log \mathbb{P}(X|A, V) = -\sum_i \sum_j \log \mathbb{P}(x_{ij}|\hat{\theta}_{ij}) = C + \sum_i \sum_j (-x_{ij}\hat{\theta}_{ij} + G(\hat{\theta}_{ij}))$$

où $G(\theta) = \log(1 + e^\theta)$. Ainsi, le but de l'ACP exponentielle est de trouver

$$a \in \text{argmin}_{a \in \mathbb{R}^l} \log(1 + e^{-x^* \dot{a} V})$$

avec $x^* = 2x - 1$.
ACP Logistique
On sait que le paramètre naturel d'une distribution normale réduite $\mathcal{N}(\mu, 1)$ est μ. Une ACP standard peut s'écrire comme la minimization de

$$\sum_i \left\| x_i - VV^T x_i \right\|^2 = \left\| X - VV^T \right\|_F^2$$

avec $VV^T = I_l$, ce qui revient donc à trouver la meilleure projection sur un espace de dimension l de la matrice des paramètres naturels du modèle saturé, soit X.

En adaptant ce résultat issu des modèles linéaires généralisés à des données binaires, en observe que le paramètre naturel pour une distribution Bernouilli est $logit(p_{ij})$, et donc dans le cas du modèle saturé, $\infty \times (2x_{ij} - 1)$. Pour pouvoir rendre cette méthode implémentable, on choisit un paramètre m élevé plutot que ∞, et on définit $\tilde{\theta}_{ij} = m \times (2x_{ij} - 1)$.
L'ACP logistique est donc la recherche de la quantité qui minimise la déviation de Bernouilli, soit

$$V \in \text{argmin}_{V \in \mathbb{R}^{d \times l}} D(X|\tilde{\Theta}VV^T) = -2 \sum_{ij} \left(x_{ij}\hat{\theta}_{ij} - \log(1 + exp(\hat{\theta}_{ij})) \right)$$

avec $\hat{\Theta} = \tilde{\Theta}VV^T$ et $VV^T = I_l$

Voici les résultats obtenus :

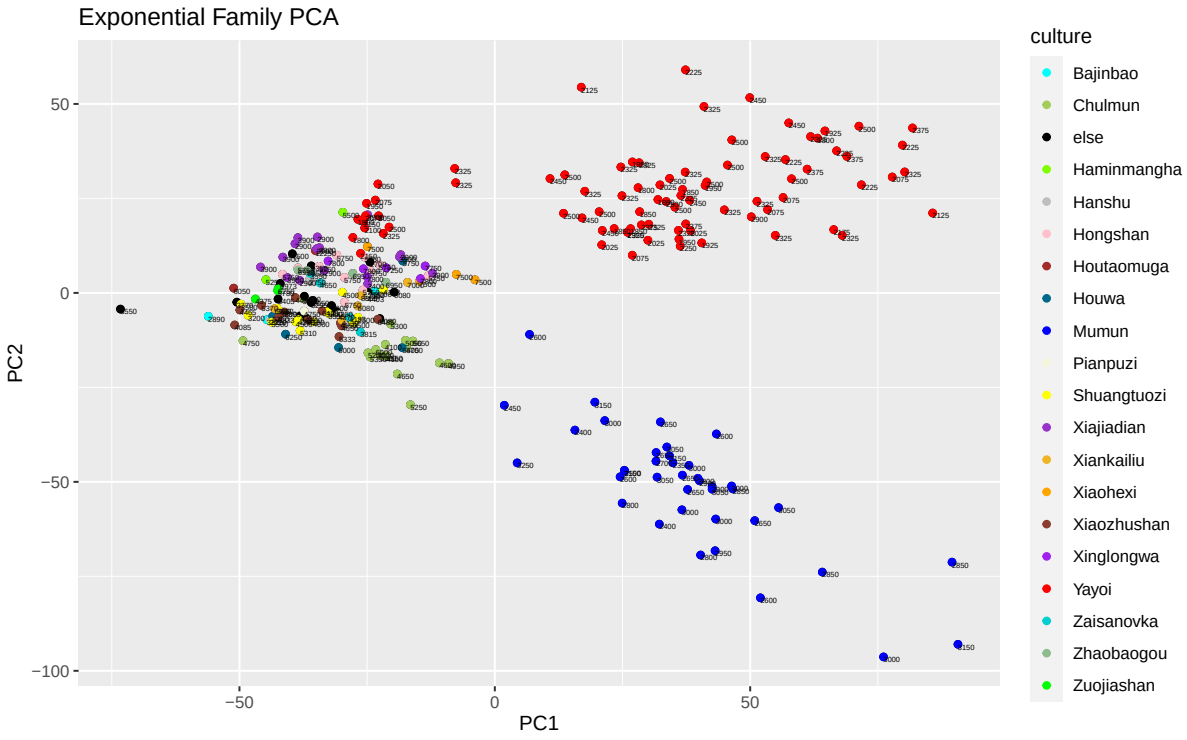


FIGURE 25 – Exonential PCA

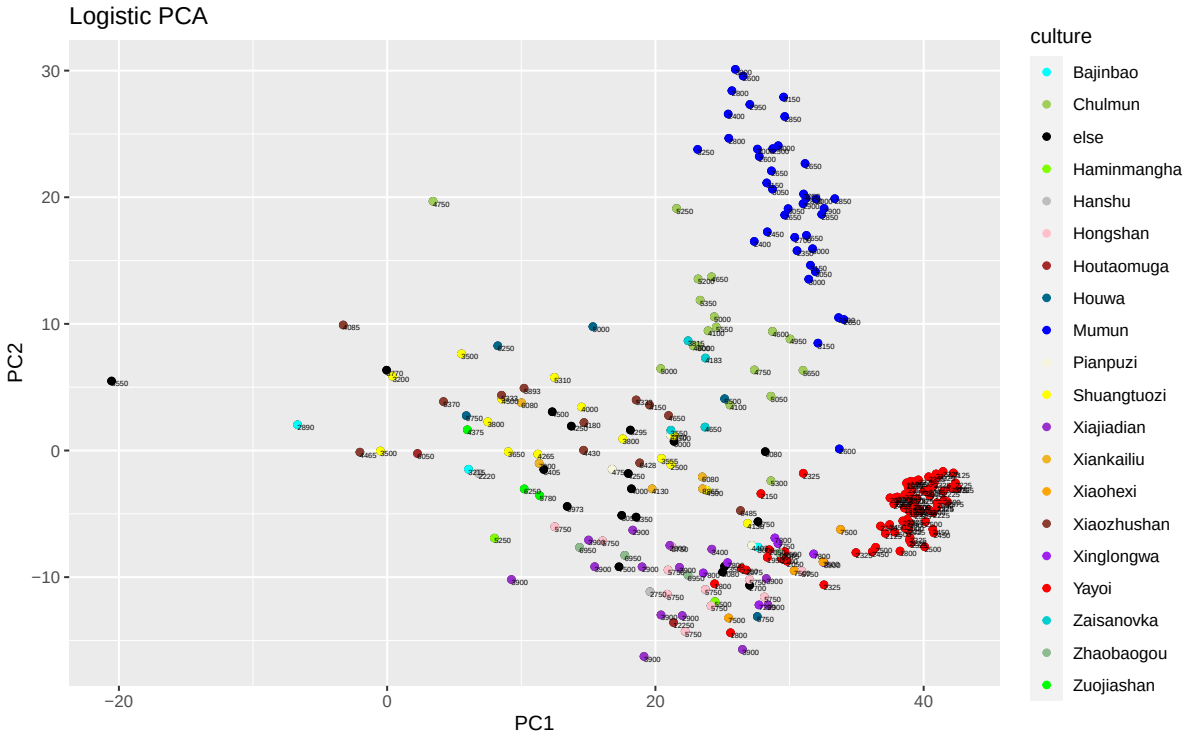
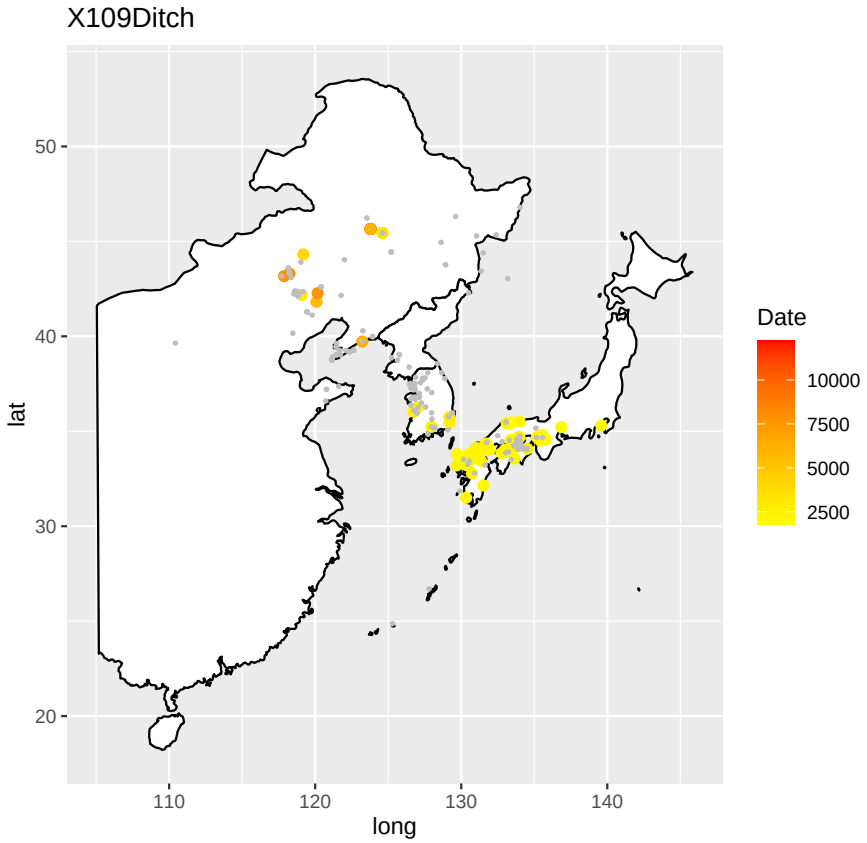


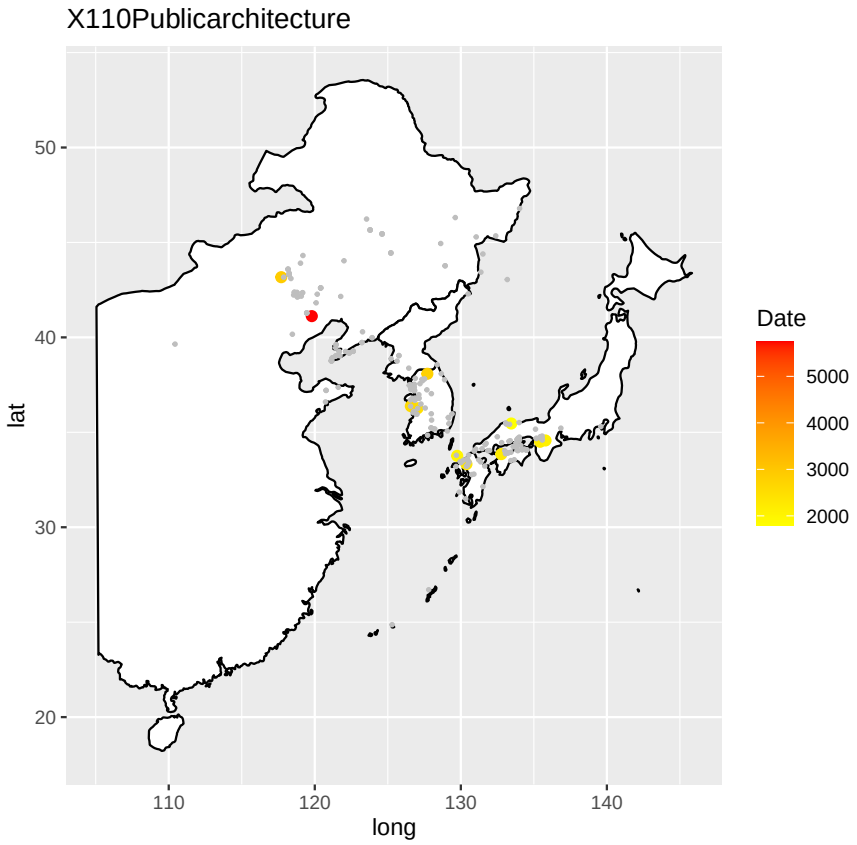
FIGURE 26 – Logistic PCA

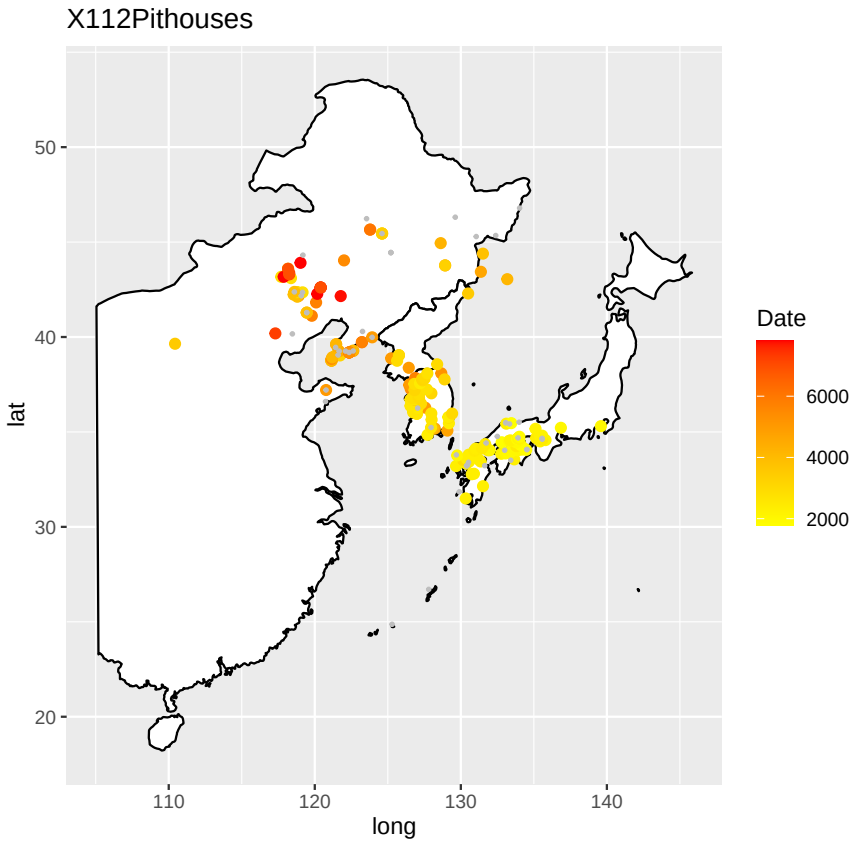
Ces résultats sont en accord avec ce qui précédait, à savoir une bonne inférence pour les groupes Yayoi, Mumun et Chulmun, mais peu d'informations sur les autres cultures. On observe de plus pas de corrélation entre l'âge des sites et leur placement sur l'ACP, puisque les dates semblent réparties aléatoirement au sein des groupes japonais et coréens. La trop grande similarité technologique entre les cultures autres que japonaises et coréennes pointent vers la nécessité d'une autre approche, en se concentrant d'avantage l'évolution de chaque technologie indépendamment, plutôt qu'une analyse par groupe culturel.

4.2 Diffusion Technologique

On se propose donc de représenter pour chaque technologie du sous-groupe bâtiments sa position et son âge, pour tenter d'observer une manière particulière de la diffusion technologique. Voici ce qu'on obtient :



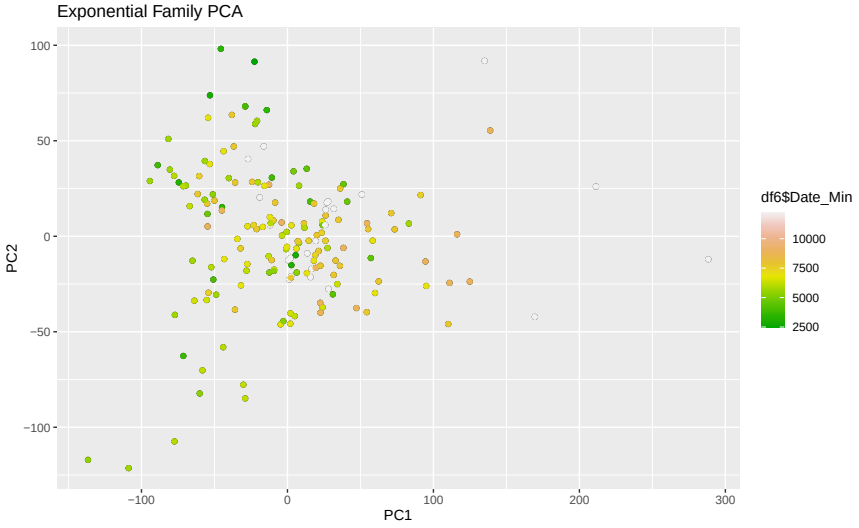

X110Publicarchitecture
50
40
30
20
lat
110
120
130
140
long
Date
5000
4000
3000
2000

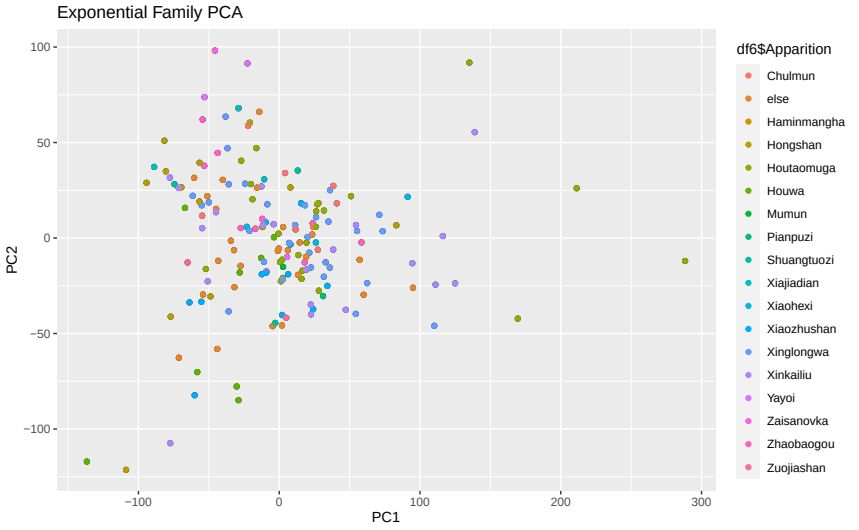

X112Pithouses
50
40
30
20
lat
110
120
130
140
long
Date
6000
4000
2000

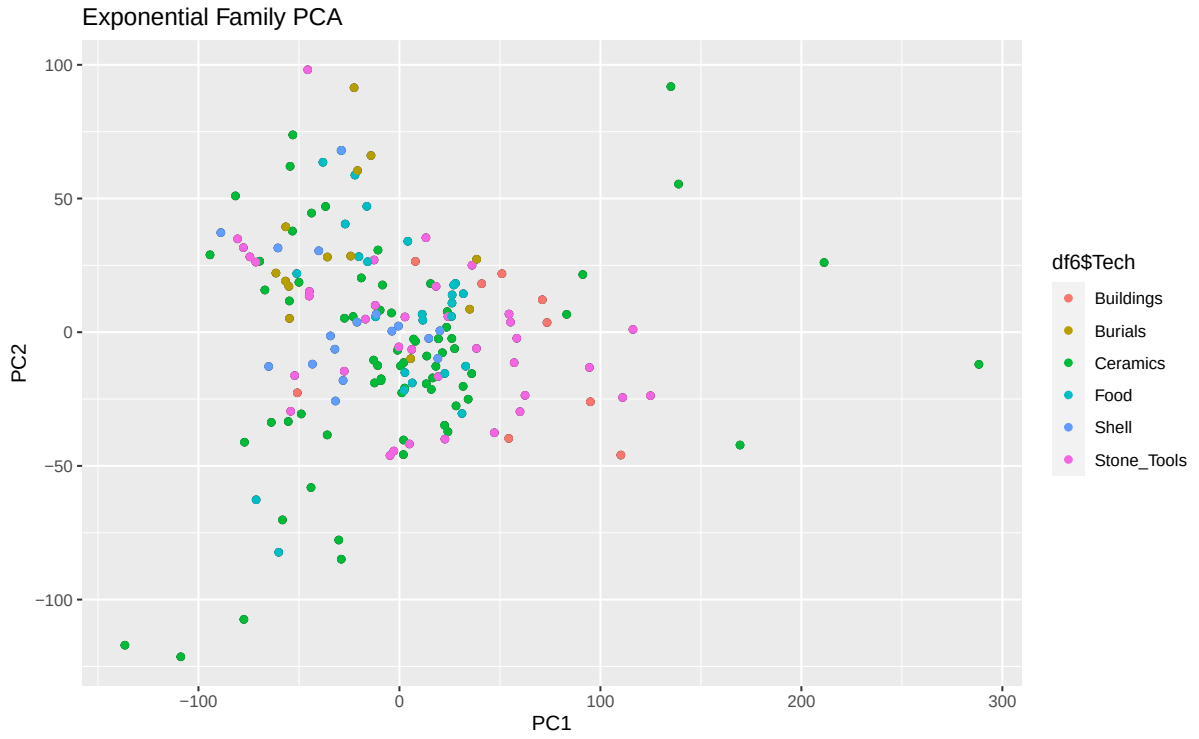
Cette représentation semble montrer des aspérités dans la fluidité de la diffuion de technologies. En effet, si toutes apparaissent dans le Nord-Est de la Chine et se diffusent ensuite dans la péninsule Coréenne et le Japon, il semble clair que leur mode de diffusion n'est pas le même. Par exemple, si la technologie Door Entrance est présente sur un très grand nombre de site, ce n'est pas le cas de la technologie Public Architecture, tandis que la technologie Ditch se diffuse peu dans l'Asie Continentale mais relativement bien au Japon. Bien entendu, l'absence d'une trouvaille technologique en un lieu à une époque ne siginifie pas que la technologie ne s'y trouvait pas, mais la simple représentation sur la carte de ces diffusions technologiques semble pointer vers une variation dans la vitesse de diffusion selon la technologie, ainsi qu'une dépendance au groupe culturel ou à la région.

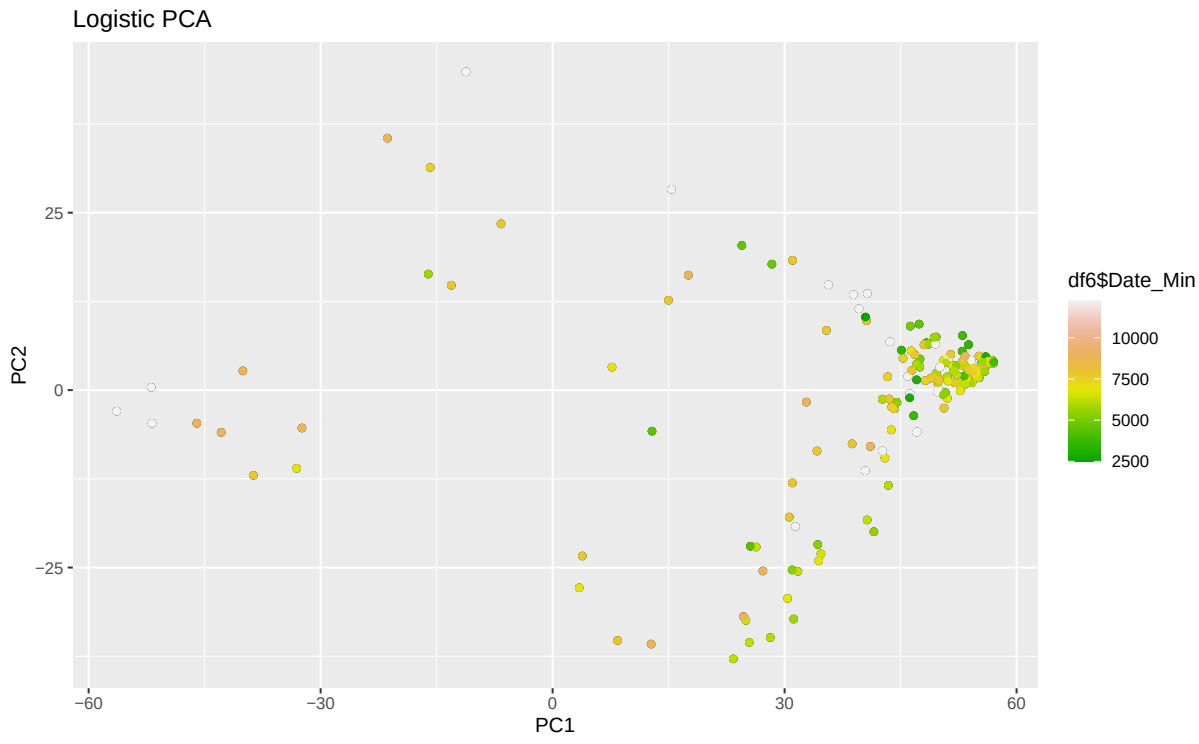
4.3 Clustering Technologique

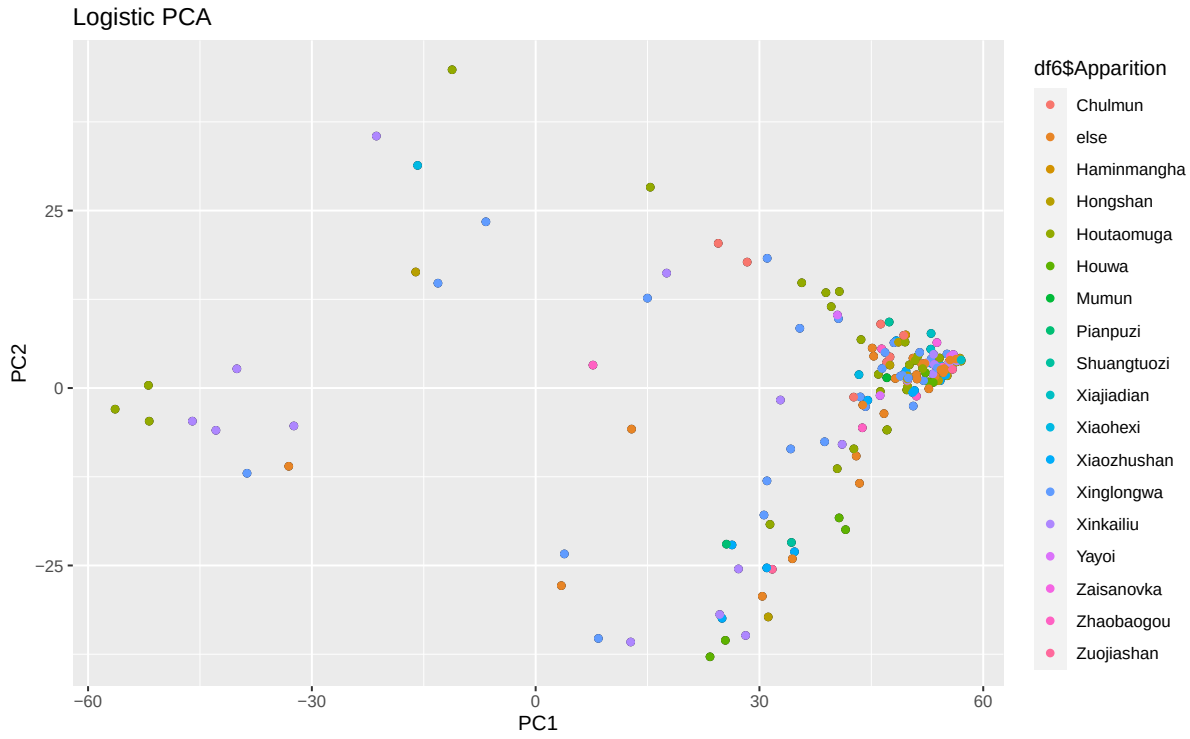
On propose donc de procéder exactement de la même manière que dans la section 4.1 en transposant la matrice de données. De cette manière, les observations sont les technologies et les variables les sites archéologiques. On y ajoute aussi la date d'apparition, la culture d'apparition et le type de chaque technologie. Voici les résultats obtenus.

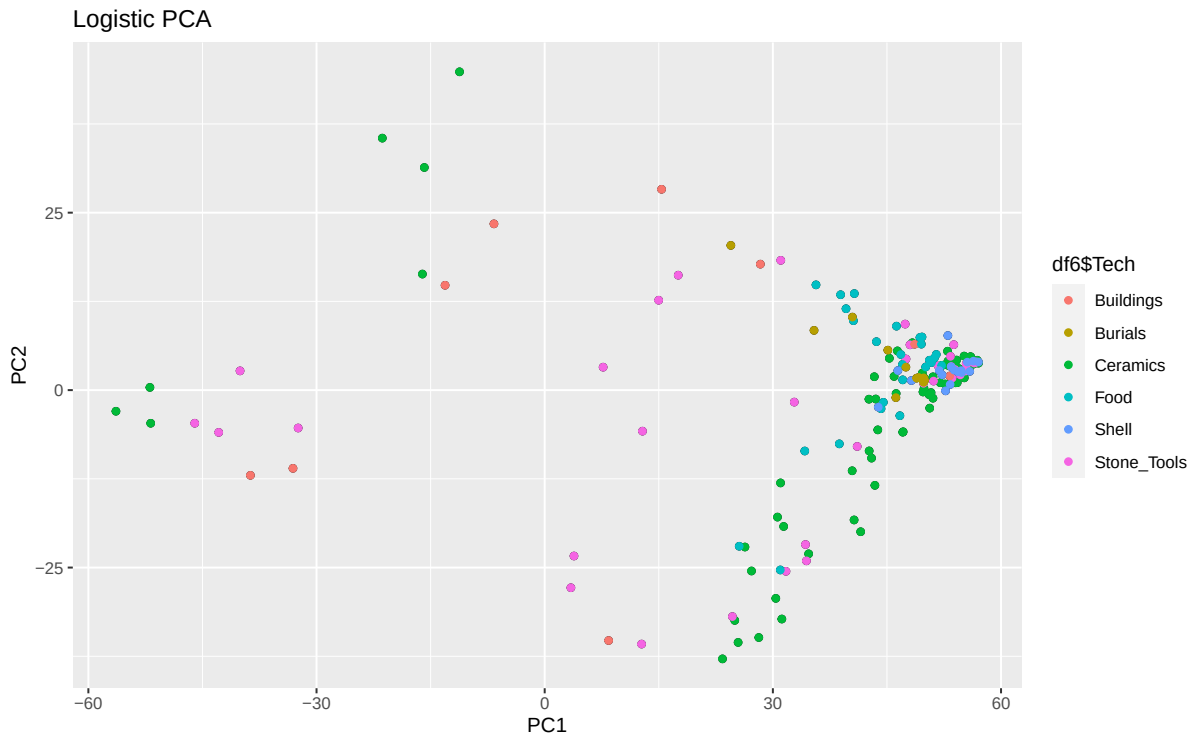





Exponential Family PCA
PC2
PC1
100
50
0
−50
−100
−100
0
100
200
300
df6$Tech
Buildings
Burials
Ceramics
Food
Shell
Stone_Tools


Logistic PCA
PC2
PC1
25
0
−25
−60
−30
0
30
60
df6$Date_Min
10000
7500
5000
2500


Logistic PCA
PC2
PC1
25
0
−25
−60
−30
0
30
60
df6$Apparition
Chulmun
else
Haminmangha
Hongshan
Houtaomuga
Houwa
Mumun
Pianpuzi
Shuangtuozi
Xiajiadian
Xiaohexi
Xiaozhushan
Xinglongwa
Xinkailiu
Yayoi
Zaisanovka
Zhaobaogou
Zuojiashan


Logistic PCA
PC2
PC1
25
0
−25
−60
−30
0
30
60
df6$Tech
Buildings
Burials
Ceramics
Food
Shell
Stone_Tools

Ces graphiques nous indiquent l'absence de relation évidente dans les données et nous indiquent que les covariables sont insuffisantes pour expliquer l'évolution des technologies. En effet, deux technologies sont d'autant plus proches qu'elles sont présentes dans les mêmes lieux. Or ne voir aucune logique dans l'ACP, montre qu'il ne suffit pas qu'une technologie apparaîsse à un instant t ou à un lieu x pour qu'on la retrouve à un instant $t' > t$ ou à un lieu x' proche de x. On en déduit donc que la logique des données est plus complexe que la simple diffusion temporelle ou géographique.

On propose ensuite une méthode de clustering hiérarchique. On calcule une matrice de distance entre les données avec la dissimilarité de Pearson : on calcul le coefficient de corrélation r entre chaque variables, puis on utilise comme "métrique" (voir [10]) la valeur $d = \frac{1-r}{2}$. On représente ensuite les clusters de cette matrice à l'aide d'un dendogramme qui les représente dans un arbre.

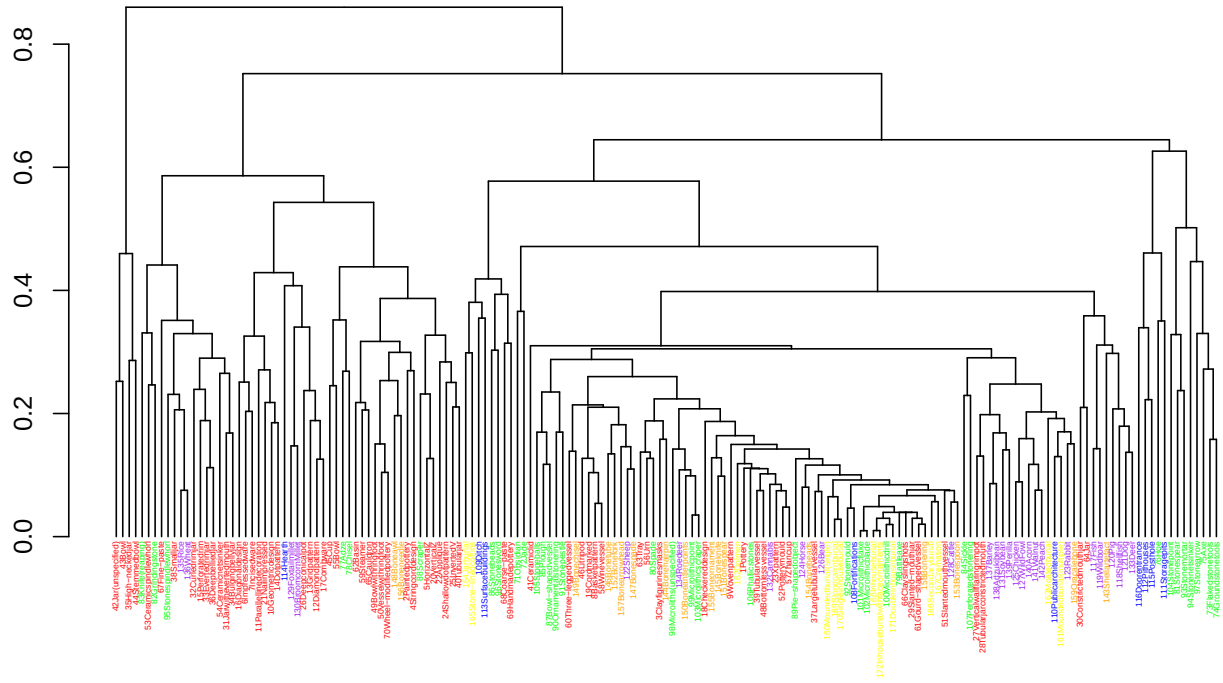


FIGURE 27 – Dendogramme Technologique

Cette visualisation nous apporte plusieurs informations intéressantes sur les données. On observe en premier lieu plusieurs clusters : tout à droite un cluster outils en pierre à droite (en vert), un cluster nourriture (en violet), un cluster céramique à gauche (en rouge). Mais on observe d'autres clusters qui ne sont pas liés à la famille technologique mais possèdent tout de même une certaine logique. Par exemple, on observe dans le deuxième cluster en partant de la droite un cluster contenant des données nourritures dont des poissons, ainsi que des coquillages (famille Shell and bones) et des jarres (famille céramique), ce qui semble montrer une corrélation logique entre les données qui dépasse les familles technologiques. De même, on observe un cluster contenant deux technologies mortuaires, la technologie Architecture publique et Oracle Bone. On observe encore un cluster contenant deux technologies mortuaires, la technologie Ditch et Surface Buildings et des céramiques.
Cette méthode de clustering indique dans un premier temps la non-indépendance des technologies, ce qui explique partiellement les difficultés rencontrées lors des analyses phylogéniques. Cette visualisation montre également que la diffusion de technologies se fait d'avantage par la pratique ou non d'un certain rite, ou d'une certaine consommation.

En mettant ensembles toutes les analyses et visualisations faites dans cette partie, il apparaît clair que les technologies de notre jeu de données suivent une logique plus complexe que la simple dynamique temporelle ou géographique. Ainsi, il ne suffit pas qu'une technologie apparaîsse à un endroit ou un instant pour qu'elle se diffuse ensuite autour. Cette diffusion semble d'avantage liée à des logiques sociales et culturelles plus complexes, et varie au sein même des propres groupes culturels.

Conclusion et remerciements

En partant des données utilisées dans l'article *Triangulation supports agricultural spread of the transeurasian languages* [9], nous avons reproduit les expériences faîtes par les auteurs, en simulant nos propres MCMC, qui ont bien convergé. On est vite arrivé à la conclusion qu'il existait peu de signal phylogénique sur ces données, en particulier entre les peuples continentaux, et qu'il était par conséquent difficile de tirer les conclusions que font les auteurs dans leur article. On a ensuite proposé de recentrer l'analyse sur les peuples qui présentaient un meilleur signal, japonais et coréens, mais nous sommes heurtés au même problème, à savoir pas assez de signal pour tirer des conclusions sur l'évolution des peuples asiatiques anciens.
On a donc ensuite tenté d'ouvir notre approche en quittant la phylogénie bayésienne et en cherchant à comprendre quels sont les moteurs de la diffusion technologique dans nos données, et il est apparu clair que des dynamiques plus profondes régissent l'évolution de la technologie dans l'Asie de l'Est. Il nous aurait fallu plus de temps pour mieux investiguer ces relations et mettre le doigt dessus.
Plusieurs pistes de réflexion appparaissent désormais. Dans un premier temps, la disparité entre les sites continentaux que les différents outils d'analyse peinent à différentier, et les sites coréens et japonais justifie d'augmenter le nombre de sites continentaux pour avoir plus de signal. Les grands écarts géographiques et temporels entre les continentaux (7000 BC à 4000 BC) et les japonais-coréens (3000 BC à 2000 BC) rendent la tentative de reconstruire l'histoire du continent en une seule analyse trop ambitieuse.
Néanmoins, nous avons essayé de particulariser dans la section 3 l'approche aux sites japonais-coréens sans succès, ce qui indique que le modèle phylogénique est à améliorer. Dans la partie 4.2, les différences dans la diffusion des différentes technologies montrent non seulement qu'une technologie peut apparaître à différentes endroits et moments de manière indépendante, mais aussi une dépendance au lieu d'aparition dans la facilté de diffusion de la technologie. Une approche serait donc de compléter la matrice en rajoutant une variable géographique et/ou climatique latente quantifiée qui contrôlerait les taux de la matrice de transition de la chaîne de Markov telle que sa vitesse de déplacement dépende de sa situation sur l'arbre. L'introduction de cette variable nécessiterait une conversation avec archéologues et géographes pour bien la calibrer. Enfin, la non-indépendance des technologies nécéssite une complexification des modèles. Par exemple, on sait qu'avec la sédentarisation de l'Homme, les technologies agricoles et de stockage ont évoluées, menant à l'introduction de religions qui font évoluer les technologies mortuaires. Ainsi, des relations complexes existent entre les technologies qui dépassent leur famille. Un modèle qui prend en compte ces relations serait à imaginer.
Ce mémoire aura été l'occasion pour moi de mettre en pratique les compétences acquises en statistiques et plus particulièrement en statistiques bayésiennes appliquées dans mon année de Master. J'ai donc pu me familiariser avec les outils de la phylogénie bayésienne, plus particulièrement le logiciel Beast et les différents packages R qui permettent une analyse approfondie des arbres linguistiques, ou dans notre cas, archéologiques. Bien qu'on n'ait pu arriver à une conclusion sur les peuples asiatiques anciens, les contributions du mémoire ne sont pas négligeables puisqu'elles réfutent une hypothèse défendue dans une revue majeure et proposent une piste d'investigation pour comprendre mieux ce sujet majeur dans la recherche archéologique contemporaine.
Je tiens à remercier Robin Ryder pour son tutorat tout au long du mémoire et la proposition de travailler sur ce sujet passionnant. Je remercie également Fabrice Rossi et Julien Poisat pour leurs conseils.

Références

[1] Remco R Bouckaert and Martine Robbeets. Pseudo dollo models for the evolution of binary characters along a tree. *BioRxiv*, page 207571, 2017.

[2] Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal components analysis to the exponential family. *Advances in neural information processing systems*, 14, 2001.

[3] Alexei J Drummond, Andrew Rambaut, BETH Shapiro, and Oliver G Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, 22(5) :1185–1192, 2005.

[4] Konstantin Hoffmann, Remco Bouckaert, Simon J Greenhill, and Denise Kühnert. Bayesian phylogenetic analysis of linguistic data using beast. *Journal of Language Evolution*, 6(2) :119–135, 2021.

[5] Mary K Kuhner and Joseph Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular biology and evolution*, 11(3) :459–468, 1994.

[6] Nicola F Müller and Remco R Bouckaert. Adaptive metropolis-coupled mcmc for beast 2. *PeerJ*, 8 :e9473, 2020.

[7] Nico Neureiter, Peter Ranacher, Rik van Gijn, Balthasar Bickel, and Robert Weibel. Can bayesian phylogeography reconstruct migrations and expansions in linguistic evolution ? *Royal Society open science*, 8(1) :201079, 2021.

[8] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218) :98–101, 2008.

[9] Martine Robbeets, Remco Bouckaert, Matthew Conte, Alexander Savelyev, Tao Li, Deog-Im An, Ken-ichi Shinoda, Yinqiu Cui, Takamune Kawashima, Geonyoung Kim, et al. Triangulation supports agricultural spread of the transeurasian languages. *Nature*, 599(7886) :616–621, 2021.

[10] Victor Solo. Pearson distance is not a distance. *arXiv preprint arXiv :1908.06029*, 2019.

[11] Biing-Feng Wang and Chih-Yu Li. Fast algorithms for computing path-difference distances. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(2) :569–582, 2018.

[12] Wangang Xie, Paul O Lewis, Yu Fan, Lynn Kuo, and Ming-Hui Chen. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic biology*, 60(2) :150–160, 2011.