UNIVERSITÉ Clermont Auvergne

MASTER 1 Bio-informatique

Parcours Analyse et Modélisation des Données

RAPPORT DE STAGE PRÉSENTÉ PAR :

Laura PLAYA PARIENTE

Titre

Construction et comparaison de pangénomes en graphe du virus de l'hépatite delta

Titre court

Pangénome en graphe du VHD

Title

Hepatitis delta virus pangenome graphs construction and comparison

Short title

HDV Pangenome graph

Responsable du Stage : **Xavier GRAND**
Institut d'hépatologie de Lyon
UMR 1350 PaThLiv - équipe HeLip
151 cours Albert Thomas
69424 Lyon Cedex 03

Juillet 2025

Remerciements

Je remercie vivement mon tuteur de stage, Xavier Grand, de m'avoir permis d'effectuer ce stage sur le sujet passionnant des graphes de pangénome viraux. Un immense merci pour sa patience, ses précieux conseils, sa rigueur, sa bienveillance, et tout le temps qu'il consacre à faire grandir ses étudiants.

Merci également à mes jeunes collègues stagiaires en bioinformatique, Safa, Lou-Sahra et Leslie, pour la superbe ambiance d'entraide et de convivialité.

Merci enfin à tout le personnel du laboratoire PaThLiv pour leur accueil chaleureux et leur professionnalisme. Mention spéciale à Marie-Laure, ma camarade de bureau, pour son énergie débordante et contagieuse, qui a rempli de bonne humeur ces huit semaines de stage.

Résumé

Les pangénomes en graphe contiennent toute la diversité génomique d'une espèce dans une structure de données en forme de graphe, avec des nœuds contenant des segments de séquence nucléotidique reliés par des arêtes de façon à pouvoir restituer chaque génome entier comme un chemin sur le graphe. Les pangénomes en graphe permettent un stockage compact et une représentation claire des régions du génome conservées (collapsées dans un seul nœud) et des régions variables (en forme de bulles). Des nombreux outils bioinformatiques ont été développés pour la construction de pangénomes en graphe d'espèces eucaryotes. L'objectif de ce stage est d'adapter trois de ces outils - Pangénome Graph Builder (PGGB), Minigraph-Cactus et Cuttlefish- pour la construction du pangénome du virus de l'hépatite D (VHD), un virus avec un génome circulaire d'ARN simple brin d'environ 1700 bases et une importante diversité dans sa séquence nucléotidique. Le travail au cours du stage a permis d'écarter le constructeur Cuttlefish, qui produit des graphes en forme de pelote ne permettant pas d'analyse de la structure du génome. Il a également permis d'identifier des critères de topologie et de qualité d'alignement pour la comparaison de pangénomes et de mettre en lumière le fort impact du choix de l'origine des séquences du VHD sur la structure des graphes.

Abstract

Pangenome graphs capture the entire genomic diversity of a species within a graph-shaped data structure, where nodes contain segments of nucleotide sequences connected by edges in such a way that each complete genome can be reconstructed as a path through the graph. These graphs allow for compact storage and a clear, intuitive representation of conserved genomic regions (collapsed into a single node) and variable regions (which form bubbles). Numerous bioinformatics tools have recently been developed for the construction of pangenome graphs of eukaryotic species, particularly the human pangenome. The aim of this internship is to adapt three of these tools — Pangenome Graph Builder (PGGB), Minigraph-Cactus, and Cuttlefish — to build the pangenome of the Hepatitis D virus (HDV), a virus with a circular single-stranded RNA genome of approximately 1700 bases and substantial sequence diversity. The work carried out during the internship led to the exclusion of the Cuttlefish graph constructor, which produces tangled graphs unsuitable for genome structure analysis. It also enabled the identification of topological and alignment quality criteria for comparing pangenome graphs and highlighted the strong impact of sequence origin on graph structure.

Sommaire

Abréviations

DBG :	De Bruijn Graph
cDBG :	compacted De Bruijn Graph
HBsAg :	Hepatitis B surface Antigen
L-HDAg :	Large Hepatitis Delta Antigen
PGGB :	Pangenome Graph Builder
S-HDAg :	Small Hepatitis Delta Antigen
VCF :	Variant Call Format
VG :	Variation Graph
VHB :	Virus de l'hépatite B
VHD :	Virus de l'hépatite Delta

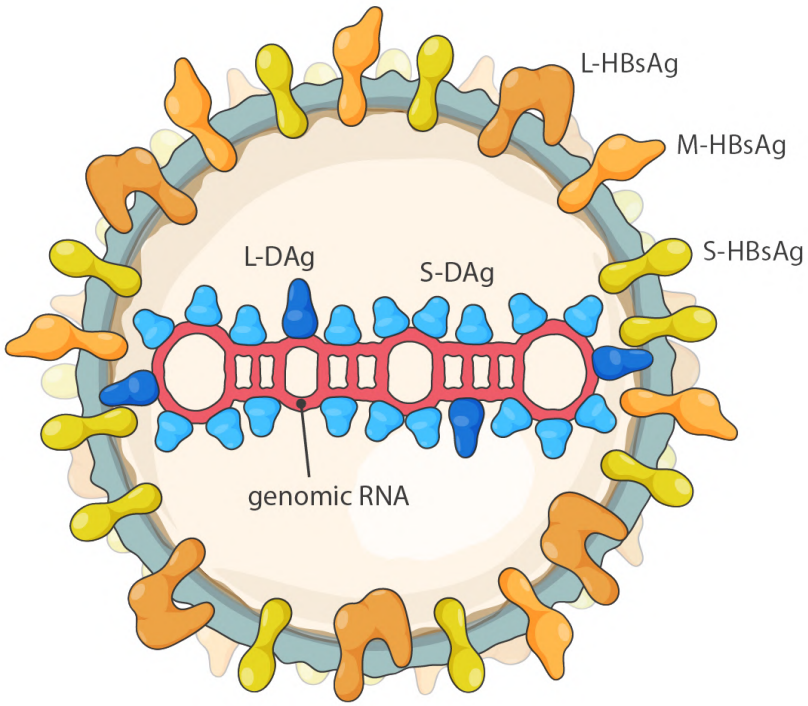


Figure 1. Virus de l'hépatite delta (VHD)
Représentation schématique d'un virion du VHD avec le complexe ribonucléique composé du génome circulaire à ARN simple brin (rouge) associé aux antigènes delta (bleu) et l'enveloppe lipidique avec les antigènes de surface du virus de l'hépatite B (jaune et orange)
L-DAg : Large Delta Antigen
S-DAg : Small Delta Antigen
L-HBsAg : Large Hepatitis B surface Antigen
M-HBsAg : Medium Hepatitis B surface Antigen
S-HBsAg : Small Hepatitis B surface Antigen
Source : https://ictv.global/report/chapter/kolmioviridae/kolmioviridae/deltavirus

INTRODUCTION ET ANALYSE BIBLIOGRAPHIQUE

Le virus de l'hépatite delta (VHD)

L'hépatite D est une maladie du foie provoquée par le virus de l'hépatite delta (VHD). L'infection par le VHD se produit uniquement en présence du virus de l'hépatite B (VHB) car le VHD a besoin des protéines de l'enveloppe du VHB pour compléter son cycle de vie. La surinfection par le VHD aggrave l'hépatite causée par le VHB, augmentant le risque de développer des complications comme l'insuffisance hépatique dans les infections aiguës et la cirrhose et le carcinome hépatocellulaire dans les infections chroniques[1]. Le traitement actuel contre le VHD repose principalement sur l'administration d'interféron, avec un taux de réponse très faible chez les patients et un risque élevé de rechute après l'arrêt du traitement[2]. Au moins 15 à 20 millions de patients infectés par le VHB sont coinfectés par VHD[3].

Le VHD est un petit virus sphérique de la famille des kolmioviridés dont l'enveloppe contient les antigènes de surface du virus de l'hépatite B (HBsAg). Cette enveloppe entoure un complexe ribonucléoprotéique constitué du génome du VHD associé aux deux isoformes de l'antigène delta : la petite (S-HDAg, 195 aa) et la grande isoforme (L-HDAg, 214 aa) (Figure 1). Le génome du VHD est un ARN circulaire simple brin de polarité négative d'une taille de 1.7kb environ. Il est caractérisé par un taux élevé de GC (> 60 %) avec plus de 70 % de complémentarité dans sa séquence, conduisant à une structure en forme de bâtonnet (Figure 1). L'ARN complémentaire de ce génome (ARN anti-génomique) contient une seule phase ouverte de lecture fonctionnelle codant les deux isoformes de l'antigène delta S-HDAg et L-HDAg.

Les séquences génomiques du VHD identifiées jusqu'à présent présentent une importante variabilité dans leur séquence nucléotidique pouvant aller jusqu'à 35 à 40 % sur l'ensemble du génome. Ces séquences ont été classées en huit génotypes (VHD-1 à VHD-8), chaque génotype regroupant des séquences avec 80% de similarité sur la séquence du génome entier. Chaque génotype est divisé en 2 à 5 sous-types avec une similarité inter-sous-type ≥ 90 %[4]. Il a été démontré in vitro que chaque génotype présente des différences remarquables dans sa capacité de réplication[5] et que l'évolution clinique et la réponse aux traitements lors de l'infection chez l'humain varient également selon les génotypes[6].

Du fait de la variabilité entre les séquences, il est difficile de trouver une séquence consensus ou une séquence de référence représentative de l'ensemble des génotypes du VHD, ce qui a mené à la multiplication de séquences de référente entre les zones géographiques et les laboratoires.

(A)

```
CCTGAGCCAA--TCCGAAAACGTA
CCTGAGCCAAGTTCCGATTTCGTA
CCTGGGCCAAGTTCCGAAAACGTA
CCTGGGCCAA--TCCGATTTCGTA
```

(B)

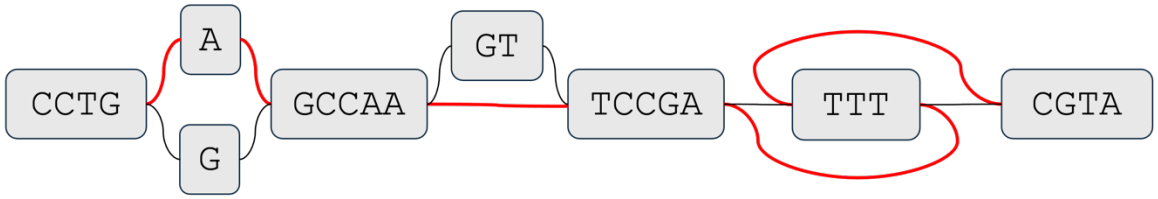


Figure 2. Illustration schématique du principe de pangénome en graphe de séquences

A. Génomes de l'espèce sous forme d'alignement multiple avec les régions variables indiquées en gras.

B. Représentation du pangénome sous forme de graphe. Certains nœuds sont communs à toutes les séquences (parties conservées), d'autres sont visités par certains chemins seulement (polymorphismes).

Le graphe est orienté de façon que chaque nœud peut être lu dans les deux sens (une séquence et son complément inverse).

Chaque génome peut être reconstituée comme un chemin dans le graphe. Les arêtes en rouge indiquent le chemin qui reconstruit le premier génome en (A).

Un effort pour regrouper cette diversité génomique a été fait avec la mise en place d'une base de données[7], mais cette base n'est plus maintenue (dernière mise à jour en 2020). Aucune ressource ne centralise aujourd'hui l'ensemble de l'information génomique du VHD.

Pangénome de séquences sous forme de graphe

En microbiologie le mot pangénome peut designer l'ensemble de gènes d'une espèce ou plus généralement les gènes d'un clade ou d'un échantillon, avec une division en génome cœur (gènes partagés par tous les individus) et génome accessoire (gènes spécifiques à certains sous-groupes). Dans le cadre de ce stage, un pangénome fait référence à un pangénome de séquences, l'ensemble de génomes entiers d'une espèce [8]. Cette approche permet d'analyser non seulement la présence ou l'absence de gènes, mais aussi la variabilité à chaque position du génome, y compris dans les régions non codantes.

D'un point de vue bioinformatique, le stockage et la représentation du pangénome peuvent se faire avec différentes structures de données. Historiquement, dû à la rareté d'assemblages de génomes complets de qualité, la représentation de la diversité génomique se fait avec un génome complet de référence et une description de tous les variants par rapport à ce génome, typiquement dans un fichier VCF[8]. Mais cette approche induit un biais de référence avec une sous-représentation des variants éloignés de la référence et des difficultés de prise en compte de variants structuraux.

Le pangénome peut aussi simplement être stocké comme une collection de génomes entiers, sous la forme d'un alignement multiple ou avec un stockage individuel de chaque génome. Mais lorsque le pangénome inclut un grand nombre de séquences ces représentations sont inefficaces pour le stockage et l'analyse[9].

Une structure de données particulièrement bien adaptée aux pangénomes est le graphe, un ensemble de nœuds reliés par des arêtes. Dans le cas des pangénomes, chaque nœud du graphe est un segment de séquence nucléotidique et chaque génome un chemin sur le graphe[8] (Figure 2).

Les pangénomes en graphe rassemblent dans une seule structure toute la diversité génomique de l'espèce avec un stockage compact et une représentation claire des régions conservées dans le génome, collapsées dans un seul nœud, et des régions variables, représentées en forme de bulles. Les graphes de pangénome permettent aussi d'aligner davantage de lectures que les séquences de référence, tout en évitant le biais de référence et améliorant la détection de variants rares et de variants structuraux[10 11].

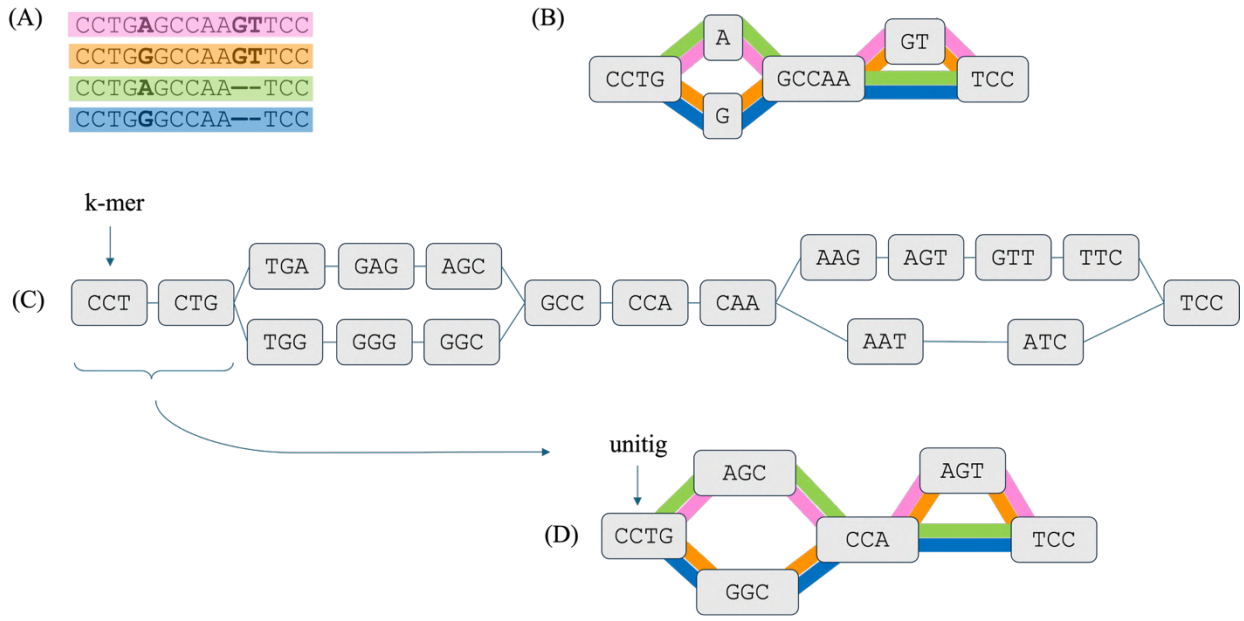


Figure 3. Types de pangénomes en graphe

A. Ensemble de séquences sous forme d'alignement multiple avec les variations indiquées en gras.

B. Graphe à partir de l'alignement (appelé graphe de variation par les auteurs de l'outil Pangénome Graph Builder). Chaque chemin en couleur correspond à une des séquences en (A)

C. Graphe de De Bruijn avec k = 3 des séquences en (A).

D. Graphe de De Bruijn compacté (k initial =3) des séquences en (A). Chaque chemin en couleur correspond à une des séquences en (A).

Avec l'augmentation rapide du nombre de génomes entiers disponibles, les pangénomes en graphe sont en plein essor et ont récemment été appliqués à de nombreuses espèces[12] [13] dont l'humain[9]. Mais l'application à des espèces procaryotes et de virus est encore modeste avec peu d'exemples dans la littérature par rapport aux espèces eucaryotes.

Stratégies pour la construction de pangénomes en graphe.

De nombreux outils bioinformatiques sont en cours de développement pour la construction de pangénomes en graphe avec des approches différentes, mais qui peuvent être regroupées en trois grandes catégories :

Graphes de variants prédéterminés

Un graphe de variants est généré à partir d'une séquence de référence unique et de la liste des variations connues par rapport à cette référence (fichier VCF). L'utilisation des variations prédéterminées a pour avantage de ne nécessiter que peu de ressources, mais elle perd aussi une partie de l'intérêt du pangénome car elle ne s'affranchit pas du biais de référence.

Ces graphes de variants peuvent être générés avec l'outil VG (Variation Graph Toolkit)[14].

Graphes à partir d'alignements

Les algorithmes dans cette catégorie génèrent des graphes à partir de l'alignement des séquences du pangénome. Deux outils se distinguent par leur popularité : Pangenome Graph Builder (PGGB)[15], qui fait référence aux graphes produits comme graphes de variation (Figure 3) et Minigraph-Cactus[16]. Les deux sont des pipelines qui enchaînent des étapes d'alignement, de génération du graphe et d'optimisation de la topologie. L'enchaînement est linéaire dans le cas de PGGB qui commence par une étape d'alignement par paires de toutes les séquences (avec l'outil wfmash), puis la génération du graphe (outil seqwish) ; alors que l'enchaînement des étapes est itératif dans Minigraph-Cactus, qui aligne et intègre progressivement des séquences au graphe à partir d'un premier alignement à une séquence de référence.

Graphes de De Bruijn compactés (cDBG)

Les graphes de De Bruijn (De Bruijn Graph : DBG) sont des graphes orientés dont les nœuds représentent des k-mers et les arêtes le chevauchement entre le suffixe et le préfixe (de taille k-1) des k-mers (Figure 3). Chaque nœud est unique dans le graphe, si la séquence se répète, le nœud sera visité plusieurs fois dans le chemin qui reconstruit la séquence. Les DBG peuvent être compactés en cDBG, en fusionnant les chemins linéaires sans embranchement (suites de nœuds avec une seule arête entre chaque nœud) en nœuds plus longs appelés unitigs.

Tableau 1. Numéros d'accession des séquences de virus l'hépatite delta (VHD) pour la construction et la validation des pangénomes.
(*) - génome de référence du laboratoire PathLiv UMR1350
En gras séquences explicitement identifiées comme complete genome dans GenBank (séquences de référence). En bleu : séquences de validation. En gris : séquences redondantes entre sources (déjà sélectionnées)

Génotype VHD	ICVT 2025 Genus Deltavirus	Miao 2019 [17] Subtype | numéro d'accession		Chowdhury 2025 [18] Table S5	Charre 2023 [19] Table S1
VHD-1 **AJ000558** (*)	**AF104263**	1a	JX888100 **KY463677**	**NC_001653**	**U81989**
		1b	JX888098 **KJ744242** **KJ744255**	**KJ744223**	M84917
		1c			M58629
		1d			**M21012**
		1e			AM779575
VHD -2	**AF104264**	2a	X60193	**MZ671233**	X60193
		2b	AJ309879		AJ309879
VHD -3	**AB037948**	3A	LT604954		L22063
		3b	**AB037947**		LT604954
		3c	**KC590319**	KF786346 RC	
VHD -4	AF018077	4a	AF018077	**AB118822**	AF018077
		4b	**AB118818**		AB088679
VHD -5	**AM183331**	5a	JX888103	LT604957	AJ584848
		5b	AM183331		AM183331
VHD -6	AX741164	6a	AJ584847	AJ584847	AJ584847
		6b	JX888102		**AM183329**
		6C	**AM183332**		
VHD -7	**AM183333**	7a	AJ584844	**MG711711**	AJ584844
		7B	AM183333		AM183333
VHD -8	**AM183330**	8a	AJ584849	AM183330	AJ584849
		8B	LT594488		LT604973

Des outils comme Pantools[20], SplitMEM[21], Bifrost[22], TwoPaCo[23] ou Cuttlefish[24] se basent sur des cDBG pour la construction de pangénomes. L'utilisation de k-mers uniques est à la fois leur force (excellente performance en termes de vitesse de calcul et d'espace de stockage) et leur faiblesse (difficulté à restituer des variations de taille inférieure à k et structure générale à faible linéarité et lisibilité)[8]. L'outil Cuttlefish été retenu dans le cadre du stage pour son format de sortie standard (GFA 1.0), sa facilité d'installation et de prise en main, ainsi que pour le travail continu de maintenance dont il fait objet.

Objectifs du stage

Ce stage s'inscrit dans un projet de création d'un centre de ressources de référence pour les génomes des virus hépatiques, dont le génome du virus de l'hépatite D. Dans ce cadre, un pangénome en graphe du VHD permettrait une analyse fine de la structure du génome et faciliterait la caractérisation de la diversité génétique.

L'objectif du stage est de construire et comparer des graphes de pangénome du VHD produits avec trois constructeurs : PGGB, Minigraph-Cactus et Cuttlefish

MATÉRIELS ET MÉTHODES

Séquences du VHD

Sélection des séquences

Les séquences du VHD utilisées pour ce travail ont été choisies parmi celles proposées comme référence pour les différents génotypes par le Comité International pour la Taxonomie de Virus (ICTV d'après ses sigles en anglais) et par 3 études récentes[17 18 19] , auxquelles s'ajoute la référence utilisée par le laboratoire d'accueil du stage, correspondant au génotype VHD-1 (Tableau 1). Cet ensemble correspond à 44 séquences dont 22 sont explicitement décrites comme génome complet et ont été sélectionnées pour la construction des graphes de pangénome.

Le pourcentage d'identité entre ces séquences se trouve entre 67% et 99% d'après des calculs de distance réalisés avec le package R Biostrings (Figure S1) avec des groupes d'identité plus importante par génotype (Figure S2). La distance mash entre ces séquences a été calculée avec l'outil Mash[25], qui fournit une estimation rapide de distance entre séquences avec l'algorithme MinHash.

Les 22 autres séquences, non décrites explicitement comme génome complet mais qui pourraient l'être par leur longueur, ont été utilisées comme séquences de validation.

Linéarisation des séquences

Le génome du VHD est circulaire et le choix de l'origine pour sa linéarisation n'est pas toujours le même selon les séquences. Une convention de numération a été proposée lors des premiers séquençages avec une origine au site unique de restriction HindIII[26], mais ce site s'est révélé très polymorphe et des séquences avec différentes origines se retrouvent dans les bases de données. Nous avons choisi l'origine utilisée par la séquence référence du laboratoire d'accueil dans son dépôt genbank (accession AJ000558.1, longueur 1678 bases) où le CDS se trouve sur la séquence complémentaire du fragment entre les positions 953 et 1597. Pour le test de l'effet réindexation des séquences, le nouvel origine est pris à la position 920 de l'alignement multiple (comprenant des gaps) généré par Multalin[27] avec les 22 séquences de référence.

Génération des graphes de pangénome

PGGB

La version v0.7.4 de l'outil Pangénome Graph Builder (PGGB) a été utilisée avec des valeurs des paramètres *-p* = 77 et *-s* = 1000 (sauf indication contraire).

Minigraph-Cactus

La version v2.9.9 de l'outil Minigraph-Cactus a été utilisée. La commande minimale et les paramètres disponibles en ligne de commande au lancement de l'outil n'ont pas permis de générer de graphe à partir de nos séquences du VHD. C'est après de nombreuses tentatives que la modification de plusieurs paramètres simultanément via le fichier de configuration a permis aux calculs d'aboutir. Les principales modifications concernent les tailles des minimiseurs et des fenêtres pour l'alignement (*-k* = 3, *-w* = 5), ainsi que les paramètres qui filtrent la taille des alignements à inclure dans le graphe (*-d* 10, *-l* 300) et le preset (profil prédéfini) *-x* lr recommandé pour l'analyse de long-reads nécessaire pour nos calculs. Le graphe de pangénome retenu pour la comparaison finale a été généré avec les options *- - noSplit - - permissiveContigFilter - - configFile*

Cuttlefish

La version v2.2 de l'outil Cuttlefish a été utilisée avec une taille de k-mer de 13 (sauf indication contraire).

Évaluation et visualisation

Les outils gfatools (version 0.5-r292) et PGGE (Pangenome Graph Evaluation, version unique dans le dépôt https://github.com/pangenome/pgge) ont été utilisés pour l'évaluation des graphes.

L'outil Bandage version 0.8.1 a été utilisé pour la visualisation et l'extraction des métriques topologiques telles que le nombre de fragments (connected components) et le nombre d'extrémités libres (dead ends). Les graphiques pour la présentation les résultats ont été générés avec R version 4.3.1 dans R Studio 2025.05.1 Build 513 à partir de scripts rédigés au cours du stage, dont une adaptation de beehave.R, le script utilisé par PGGE pour la génération de violin plots.

Général

Les outils PGGB, Minigraph-Cactus et PGGE ont été exécutées dans des containers Singularity version 3.11 et l'ensemble du projet a été mené en local sur une machine du laboratoire d'accueil avec Linux Ubuntu 22.04.3. Les graphes de pangénome ont été générés au format GFA 1.0 ou 1.1. Tous les scripts produits sont en libre accès dans le dépôt git https://github.com/laplayap/hdv_pangenome.

RESULTATS

L'objectif du stage était la construction et la comparaison de graphes de pangénome du VHD avec 3 outils : Pangenome Graph Builder (PGGB), Minigraph-Cactus et Cuttlefish. Pour cela, les valeurs par défaut des paramètres dans chaque outil ont été conservés autant que possible en utilisant des commandes minimales. Cependant, certains paramètres doivent être choisis ou d'autres adaptés aux des séquences d'entrée. La première partie du stage a été dédiée à l'exploration des paramètres influents et l'impact de leur variation sur les graphes produits.

Paramétrage des outils

Pangénome Graph Builder – PGGB

Deux paramètres sont signalés dans la documentation PGGB comme étant clé pour la définition de la structure du graphe : la longueur de la graine (*-s*, valeur par défaut 10k) et l'identité minimale entre graines (*-p*, valeur par défaut 95), les deux utilisés pour l'étape d'alignement de séquences

Paramètre -p

La valeur de *-p* a été fixée à 77 d'après le calcul de distance mash recommandé dans la documentation. La distance mash maximale trouvée entre les séquences de référence du VHD a été de 0.23, et donc l'identité minimale calculée comme 1 – 0,23 = 0,77.

Paramètre -s

La valeur de -s était dans notre cas bornée par la limite inférieure de l'outil (100) et la longueur des séquences VHD (1680 nt) car la graine ne peut pas être plus longue que la séquence alignée.

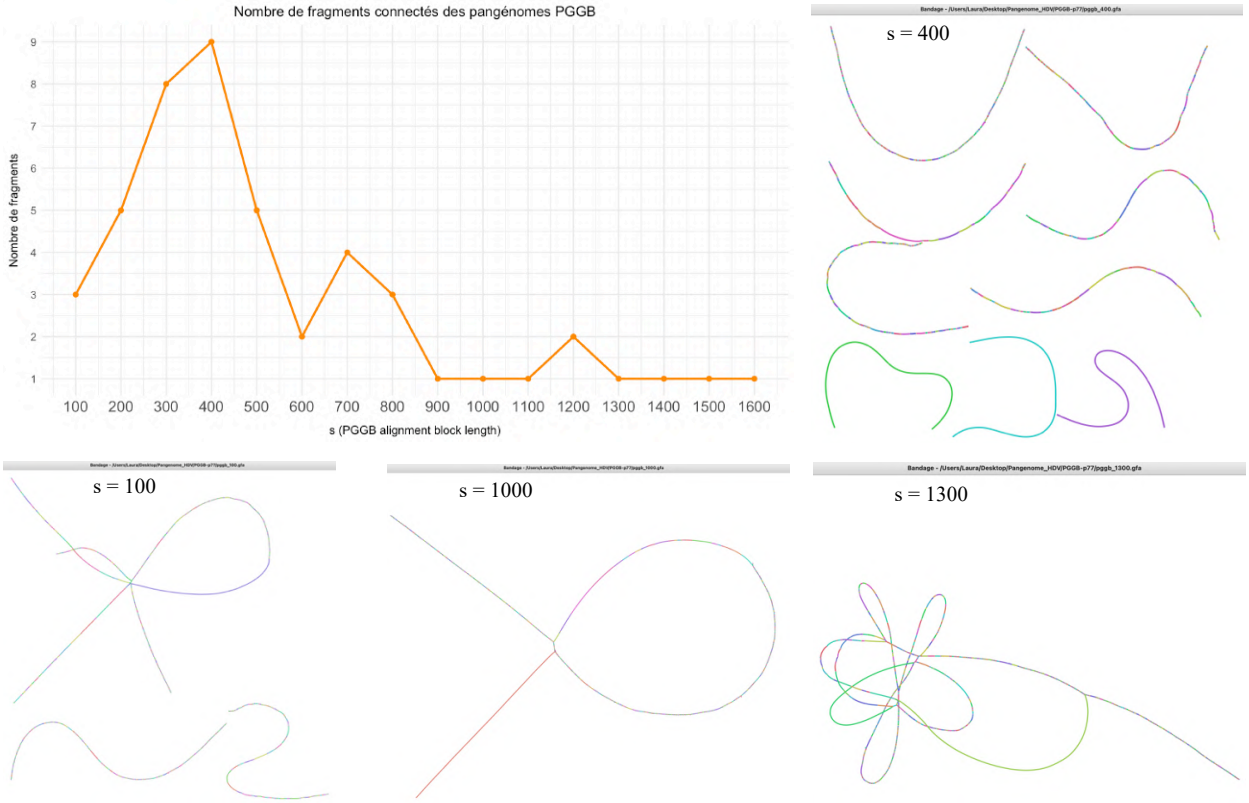


Figure 4. Graphes générés avec PGGB – variation avec paramètres *-s*
Nombre de fragments en fonction du paramètre s (haut à gauche) et visualisation des graphes pour *-s* = 100 (bas à gauche), *-s* = 1000 (bas centre) et *-s* = 1300 (bas à droite). Paramètre p fixé à 77.

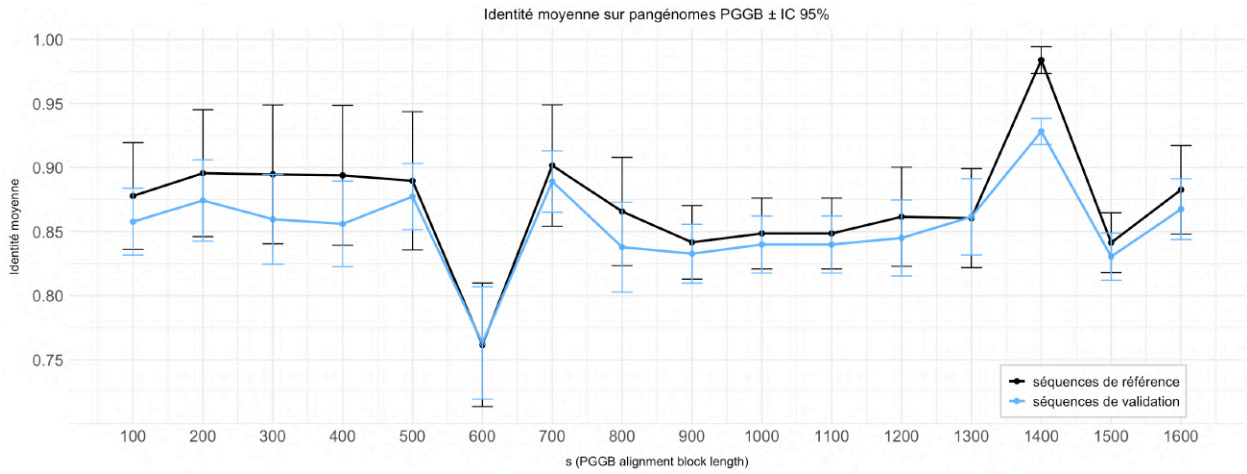


Figure 5. Identité moyenne des alignements sur les graphes PGGB – variation avec paramètre *-s*
Identité lors de l'alignement des séquences de référence ayant servi à la construction des graphes (noir) et de validation (bleu) en fonction du paramètre *-s*. Paramètre *-p* fixé à 77.

Des graphes de pangénome ont été générés pour des valeurs de -s entre 100 et 1600 par pas de 100. L'exploration visuelle de ces graphes a révélé que seulement 7 des 16 graphes générés (-*s* = 900, 1000, 1100, 1300, 1400, 1500 et 1600), ont une structure unique connexe, les autres sont fragmentés en plusieurs composants (Figure 4). Le nombre de fragments est plus important pour les plus faibles valeurs de -*s*, avec une valeur maximale de 9 fragments pour -s = 400.

Les 44 séquences de VHD ont été alignées sur les graphes générés, sauf pour le graphe -s = 1400, où seulement 60% des séquences ont pu être alignées, et la qualité de ces alignements a été évaluée à travers la métrique aln.id calculée avec l'outil PGGE. Cette métrique correspond au pourcentage d'identité nucléotidique des séquences alignées sur le graphe.

Ce pourcentage oscille sans tendance claire avec le paramètre -*s*, avec des valeurs proches entre les séquences de référence et de validation (Figure 5). Un maximum marqué pour -*s* = 1400 est observé, du fait que les séquences les plus difficiles à aligner n'ont pas été prises en compte pour le calcul. Malgré sa performance modérée en alignement des séquences, le pangénome avec -*s* =1000 a été retenue pour la suite par sa linéarité (Figure 4), sa structure en un seul fragment et une valeur de longueur de graine significativement plus petite que la longueur totale du génome (contrairement à des valeurs de -*s* ≥ 1500).

Minigraph-Cactus

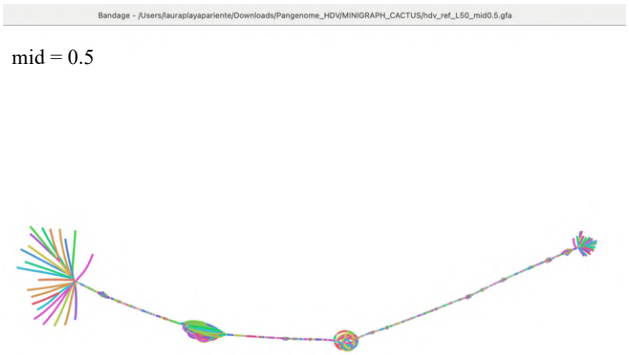
Les paramètres de l'outil Minigraph-Cactus ont déjà été ajustés pour permettre aux calculs d'aboutir avec les séquences de référence du VHD. Devant le grand nombre de paramètres disponibles et l'absence de préconisation dans la documentation, seul le paramètre *mid* (« minimal identity », valeur par défaut 0.5) a été ajusté. Ce paramètre filtre les alignements à intégrer au graphe de façon à retenir seulement les alignements avec une identité > *mid*.

Des graphes de pangénome générés pour des valeurs de *mid* entre 0.3 et 0.8 ont tous une topologie similaire, en un seul fragment, avec beaucoup d'extrémités libres (Figure 6A). Lors de l'alignement de séquences sur ces graphes, tous les pourcentages d'identité nucléotidique se situent entre 90% et 100% (Figure 6B). En raison des faibles différences lors de la variation de *mid*, la valeur par default de 0.5 est conservée pour la suite.

Cuttlefish

L'outil Cuttlefish nécessite la définition du paramètre *k*, la taille du k-mer pour la construction du graphe de De Bruijn. *k* doit être un nombre impair entre 1 et 63, la valeur par défaut dans l'outil est de 27. Des graphes de pangénome du VHD sont générés pour des valeurs de *k* entre 1 et 45.

(A)



(B)

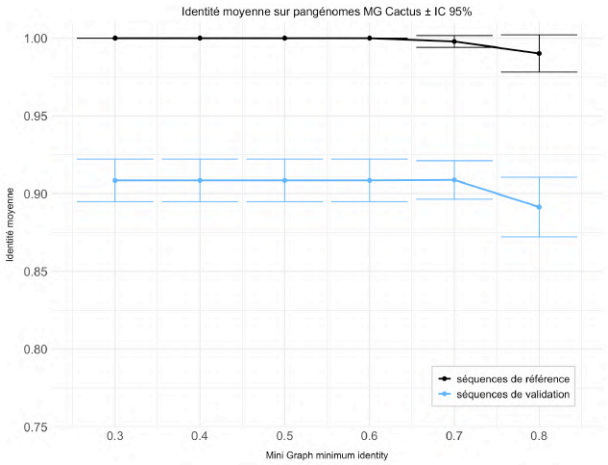
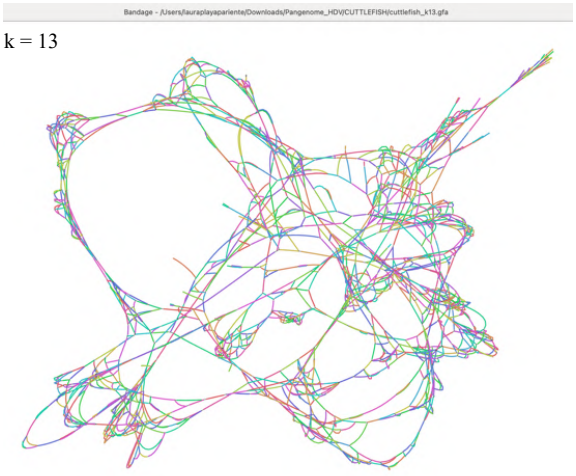


Figure 6. Graphes générés avec Minigraph-Cactus
A. Visualisation du graphe pour minimal identity (*mid*) = 0.5 (valeur par défaut)
B. Identité lors de l'alignement des séquences en fonction du paramètre *mid*.
Valeurs moyennes pour les séquences de référence (noir) ou de validation (bleu).

(A)



(B)

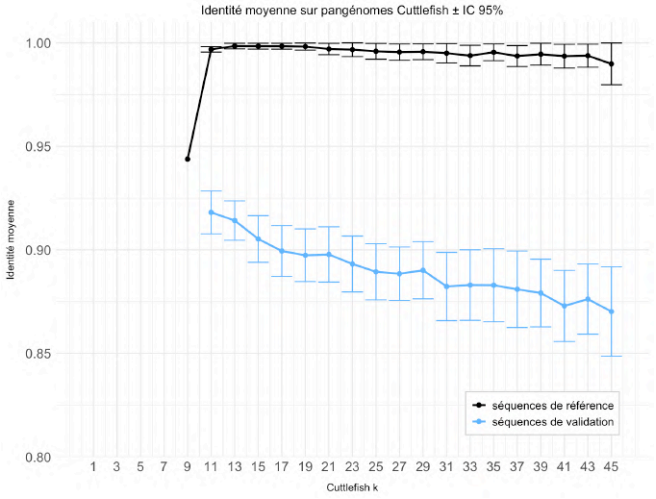


Figure 7. Graphes générés avec Cuttlefish
A. Visualisation du graphe avec $k = 13$.
B. Identité lors de l'alignement des séquences en fonction de *k*, la taille du k-mer du graphe de De Bruijn.
Valeurs moyennes pour les séquences de référence (noir) ou de validation (bleu).

Tous les graphes générés sont formés d'un seul fragment, mais leur structure devient rapidement complexe, avec de nombreuses boucles entre les nœuds et un rendu visuel en forme de pelote qui rend difficile son interprétation (Figure 7A).

L'alignement de l'ensemble des 44 séquences a été possible seulement pour des valeurs de $k \geq 13$. A partir de cette valeur le pourcentage d'identité des alignements est proche de 100% pour les séquences de référence, mais baisse avec k pour les séquences de validation (figure 7B).

Le graphe avec k = 13 a été retenu pour la suite car il permet l'alignement de l'ensemble de séquences tout en limitant la perte en qualité d'alignement pour les séquences de validation.

Comparaison des outils

Deux types de critères ont été utilisés pour la comparaison des graphes générés avec les trois outils : des critères de topologie et des critères de qualité d'alignement sur les graphes des 44 séquences VHD sélectionnées (Tableau S1).

Critères topologiques

Les pangénomes sélectionnés pour la comparaison avaient déjà tous passé le premier filtre topologique pour éviter la fragmentation et étaient tous constitués d'un seul fragment. Une analyse plus fine de la topologie a été menée avec la comparaison de métriques autour du nombre et de la taille des nœuds : distribution des longueurs des nœuds, nombre de nœuds et longueurs totale, ainsi que le nombre d'extrémités libres. Le graphe généré avec PGGB a presque 3000 nœuds, alors que les graphes générés avec Minigraph-Cactus et Cuttlefish en ont autour de 2000 et 1000 respectivement (Tableau S2). Mais la plupart des nœuds du graphe PGGB étant de longueur 1 (Figure 8A), la longueur totale des nœuds du graphe PGGB est nettement inférieure aux deux autres (Figure 8B).

Concernant les extrémités libres, le graphe Minigraph-Cactus en comporte beaucoup plus que les autres, concentrées au début et fin de graphe (Figure 6A).

Critères de qualité d'alignement

La qualité des alignements sur les graphes a été évaluée à l'aide de cinq métriques fournies par le pipeline PGGE : l'identité moyenne des alignements (aln.id), la couverture des séquences d'entrée (qsc), la proportion d'alignements uniques (uniq), la proportion d'alignements multiples (multi) et la proportion de séquences non alignées (nonaln). Des valeurs élevées de aln.id, qsc et uniq, et faibles de multi et nonaln reflètent une bonne qualité d'alignement

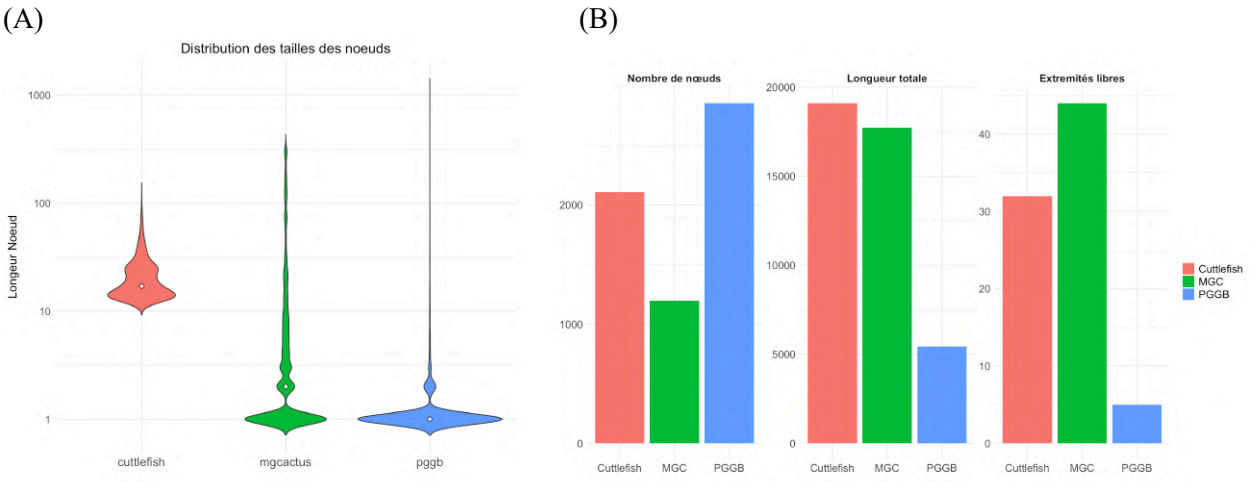


Figure 8. Comparaison graphes de pangénome – critères topologiques

A. Comparaison des distributions des tailles des nœuds. La médiane des distributions est indiquée par un point blanc. B. Critères de topologie entre les graphes de pangénome du VHD générés avec les outils PGGB, Minigraph-Cactus et Cuttlefish.

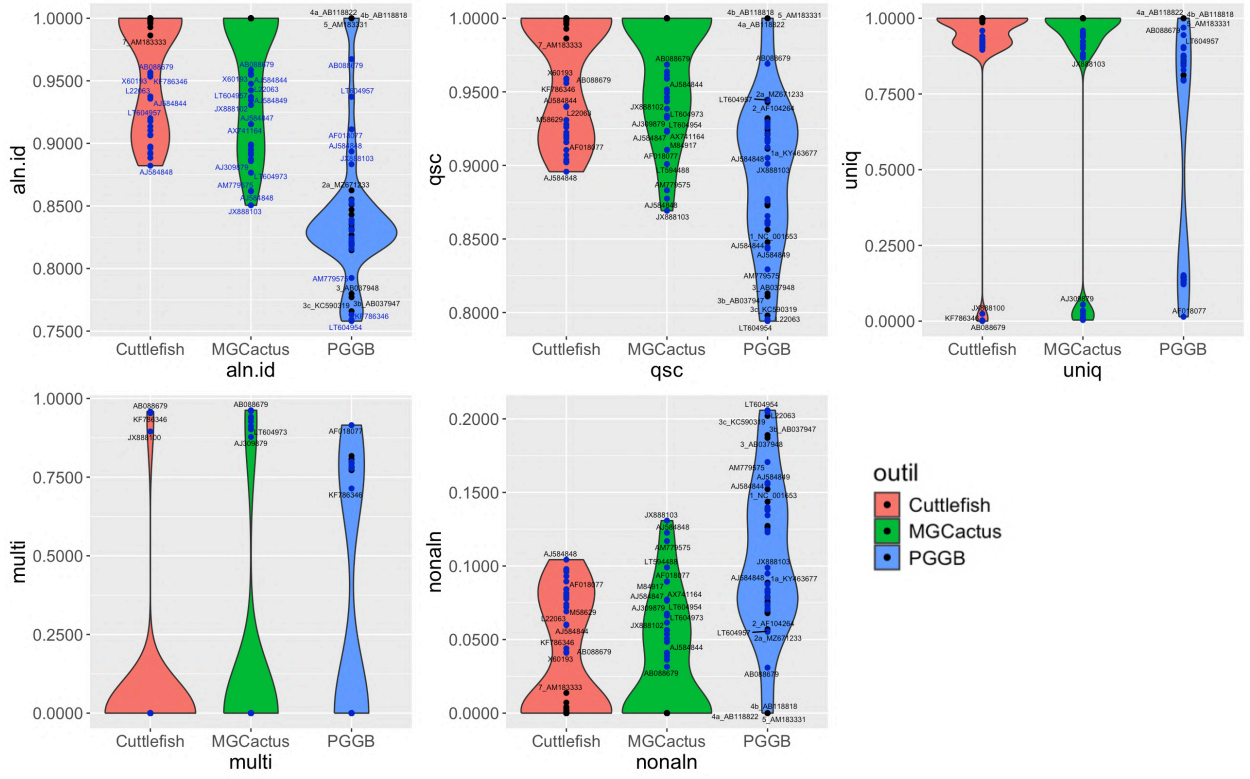


Figure 9. Comparaison des graphes de pangénome – critères d'alignement

Métriques calculées sur les alignements de 44 séquences du VHD sur les graphes, chaque point correspond à une séquence de référence (noir) ou validation (bleu). **aln.id** : pourcentage de nucléotides identiques parmi ceux alignées, **qsc** : pourcentage de nucléotides de la séquence alignés sur le graphe, **uniq** : alignements uniques sur le graphe, **multi** : alignement d'une même portion de la séquence query sur plus d'une région du graphe, **nonaln** : pourcentage de nucléotides non alignées sur le graphe.

La qualité des alignements sur le graphe généré avec Cuttlefish est la meilleure avec des pourcentages d'identité et de couverture entre 88 et 100% pour toutes les séquences. La performance du graphe produit avec Minigraph-Cactus est excellente pour les séquences de référence (identité et couverture à 100%), mais se dégrade pour les séquences de validation, avec des valeurs entre 85 et 97%. Les alignements sur le graphe PGGB sont ceux de moins bonne qualité, avec des valeurs d'identité et couverture entre 76 et 100% sans distinction entre les types de séquences (Figure 9).

Effet du changement d'origine des séquences

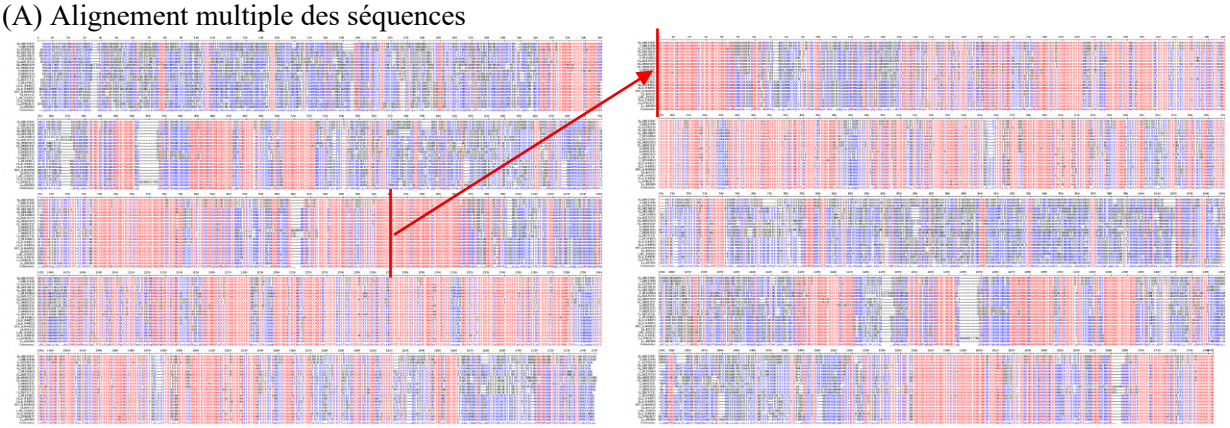
Le grand nombre d'extrémités libres en début et fin du pangénome construit avec Minigraph-Cactus a conduit à des interrogations sur l'effet d'un changement de l'origine dans la linéarisation des séquences données en entrée de l'outil. Des graphes ont été générés avec les 3 outils (PGGB, Minigraph-Cactus et Cuttlefish) avec les mêmes séquences d'entrée, mais où l'origine avait été choisi au sein d'une région conservée (Figure 10A). Ce changement transforme comme attendu les extrémités libres en début et fin de graphe du graphe Minigraph-Cactus en une bulle (Figure 10C), mais il a également un fort impact dans la topologie du graphe PGGB (Figure 10B). L'impact sur le graphe issu de Cuttlefish paraît faible mais est difficile à évaluer visuellement (Figure 10D).

DISCUSSION, CONCLUSIONS ET PERSPECTIVES

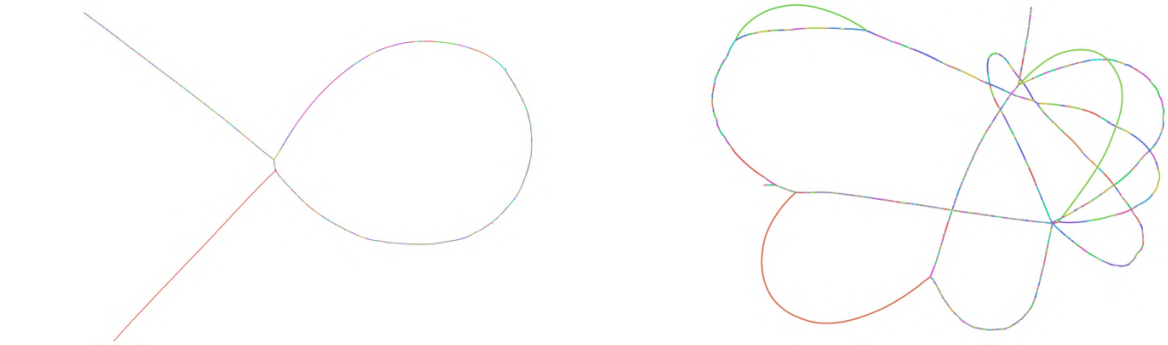
Ce stage a été l'occasion d'explorer la construction de pangénomes de séquences pour le virus de l'hépatite D. Des pangénomes en graphes issus de trois constructeurs, PPGB, Minigraph-Cactus et Cuttlefish, ont été comparés. Cette comparaison a nécessité d'une part un important travail de paramétrage pour adapter les constructeurs à la petite taille du génome du VHD, et d'autre part la définition de critères pour l'évaluation et l'analyse des différences entre les graphes produits.

Deux critères topologiques ont rapidement émergé comme essentiels pour le choix des graphes: **l'intégrité structurelle**, un graphe unique et non fragmenté, et **la linéarité**. Ces critères sont particulièrement importants dans le cadre du stage, qui fait partie d'un projet plus large ayant comme objectif à long terme la création d'un explorateur de pangénome interactif, « pangenome browser » en anglais. Des graphes en un seul fragment avec une structure linéaire sont requis pour l'exploration des régions conservées et des régions variables du VHD avec le futur browser. Les graphes de pangénome produits par Cuttlefish ont été écartés sur ce critère de linéarité: leur topologie dense en forme de pelote les rendait inadaptés à une exploration visuelle ou à une analyse fonctionnelle.

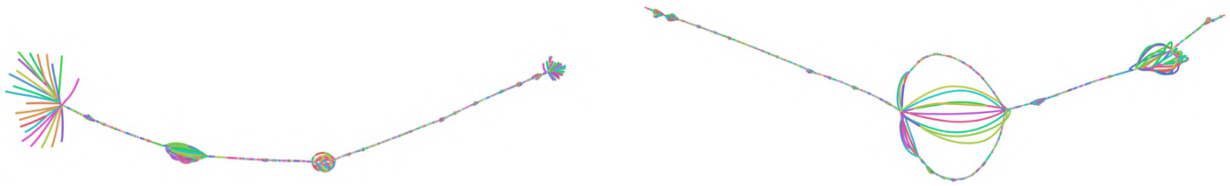
(A) Alignement multiple des séquences



(B) Graphes PGGB



(C) Graphes Minigraph-Cactus



(D) Graphes Cuttlefish

Figure 10. Effet du changement d'origine des séquences du VHD du pangénome.
Comparaison point d'origine = 1 (gauche) et = 920 (droite) des 22 séquences de référence du VHD utilisées pour la constrcution des pangénomes. (A) Alignement Multiple des séquences (B) Graphes générés avec PGGB. (C) Graphes générés avec Minigraph-Cactus. (D) Graphes générés avec Cuttlefish.

À ce jour, l'évaluation de la linéarité repose sur une inspection visuelle des graphes, ce qui constitue une des principales limites de ce travail. Le développement d'un critère quantitatif pour mesurer cette propriété constitue une perspective importante. Ce critère pourrait être un score composé de plusieurs métriques de complexité topologique comme la distribution de longueurs et le degré (nombre de connexions) moyen des nœuds.

Pour PGGB et Minigraph-Cactus, les paramètres de l'étape d'alignement préalable à la construction du graphe ont été ajustés pour permettre la production de pangénomes du VHD. Dans le cas de PGGB, la taille et l'identité des graines utilisées lors de l'alignement ont été modifiées, avec le constat qu'une faible taille de graine entraînait la génération de graphes fragmentés. Une valeur intermédiaire de taille de graine a été retenue, permettant la production d'un graphe unique, globalement linéaire. Pour Minigraph-Cactus, plusieurs paramètres ont dû être modifiés, en particulier la taille des *minimizers* et des fenêtres d'indexation, qui ont été réduites au minimum pour garantir la production du graphe.

Plusieurs critères ont été proposés pour la comparaison des graphes produits avec les différents constructeurs. D'un côté des **métriques topologiques** : le nombre total de nœuds, la longueur totale du graphe, et le nombre d'extrémités libres, dont une valeur élevée reflète un défaut de connectivité. Des graphes trop longs et peu connectés perdent en lisibilité, compacité et facilité d'interprétation biologique. D'un autre côté des **métriques reflétant la qualité des alignements** de génomes complets du VHD sur les graphes, comme le pourcentage de longueur de chaque séquence alignée, le pourcentage d'identité sur les régions alignées et la proportion d'alignements uniques. Une bonne qualité d'alignement est importante car le pangénome doit restituer fidèlement la diversité génomique du virus sous forme de différents chemins sur le graphe.

Au vu de l'ensemble de ces critères, Minigraph-Cactus présente le meilleur compromis entre linéarité et qualité d'alignement, surtout après un choix adapté de l'origine pour la linéarisation des séquences qui a permis de réduire drastiquement le nombre d'extrémités libres.

En conclusion, ce travail a permis d'écarter les constructeurs de pangénomes basées sur des graphes de De Bruijn comme Cuttlefish, et d'identifier des critères pertinents dans le choix du constructeur pour le VHD. Il a également mis en lumière le fort impact de l'origine des séquences sur la topologie des graphes. Cette observation, non documentée dans les outils existants, souligne l'importance de la prise en compte de la circularité des génomes pour la conception du futur explorateur de pangénome.

RÉFÉRENCES BIBLIOGRAPHIQUES

1. Farci, P., and Niro, G. (2012). Clinical Features of Hepatitis D. Semin Liver Dis *32*, 228–236. https://doi.org/10.1055/s-0032-1323628.
2. Wedemeyer, H., Yurdaydin, C., Hardtke, S., Caruntu, F.A., Curescu, M.G., Yalcin, K., Akarca, U.S., Gürel, S., Zeuzem, S., Erhardt, A., et al. (2019). Peginterferon alfa-2a plus tenofovir disoproxil fumarate for hepatitis D (HIDIT-II): a randomised, placebo controlled, phase 2 trial. The Lancet Infectious Diseases *19*, 275–286. https://doi.org/10.1016/S1473-3099(18)30663-7.
3. Chen, H.-Y., Shen, D.-T., Ji, D.-Z., Han, P.-C., Zhang, W.-M., Ma, J.-F., Chen, W.-S., Goyal, H., Pan, S., and Xu, H.-G. (2019). Prevalence and burden of hepatitis D virus infection in the global population: a systematic review and meta-analysis. Gut *68*, 512–521. https://doi.org/10.1136/gutjnl-2018-316601.
4. Le Gal, F., Brichler, S., Drugan, T., Alloui, C., Roulot, D., Pawlotsky, J.-M., Dény, P., and Gordien, E. (2017). Genetic diversity and worldwide distribution of the deltavirus genus: A study of 2,152 clinical strains. Hepatology *66*, 1826–1841. https://doi.org/10.1002/hep.29574.
5. Wang, W., Lempp, F.A., Schlund, F., Walter, L., Decker, C.C., Zhang, Z., Ni, Y., and Urban, S. (2021). Assembly and infection efficacy of hepatitis B virus surface protein exchanges in 8 hepatitis D virus genotype isolates. J Hepatol *75*, 311–323. https://doi.org/10.1016/j.jhep.2021.03.025.
6. Roulot, D., Brichler, S., Layese, R., BenAbdesselam, Z., Zoulim, F., Thibault, V., Scholtes, C., Roche, B., Castelnau, C., Poynard, T., et al. (2020). Origin, HDV genotype and persistent viremia determine outcome and treatment response in patients with chronic hepatitis delta. Journal of Hepatology *73*, 1046–1062. https://doi.org/10.1016/j.jhep.2020.06.038.
7. Usman, Z., Velkov, S., Protzer, U., Roggendorf, M., Frishman, D., and Karimzadeh, H. (2020). HDVdb: A Comprehensive Hepatitis D Virus Database. Viruses *12*, 538. https://doi.org/10.3390/v12050538.
8. Andreace, F., Lechat, P., Dufresne, Y., and Chikhi, R. (2023). Comparing methods for constructing and representing human pangenome graphs. Genome Biol *24*, 274. https://doi.org/10.1186/s13059-023-03098-2.
9. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. Nature *617*, 312–324. https://doi.org/10.1038/s41586-023-05896-x.
10. Sirén, J., Monlong, J., Chang, X., Novak, A.M., Eizenga, J.M., Markello, C., Sibbesen, J.A., Hickey, G., Chang, P.-C., Carroll, A., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science *374*, abg8871. https://doi.org/10.1126/science.abg8871.
11. Duchen, D., Clipman, S.J., Vergara, C., Thio, C.L., Thomas, D.L., Duggal, P., and Wojcik, G.L. (2024). A hepatitis B virus (HBV) sequence variation graph improves alignment and sample-specific consensus sequence construction. PLoS One *19*, e0301069. https://doi.org/10.1371/journal.pone.0301069.
12. Leonard, A.S., Crysnanto, D., Mapel, X.M., Bhati, M., and Pausch, H. (2023). Graph construction method impacts variation representation and analyses in a bovine super-pangenome. Genome Biol *24*, 124. https://doi.org/10.1186/s13059-023-02969-y.
13. Meng, Q., Xie, P., Xu, Z., Tang, J., Hui, L., Gu, J., Gu, X., Jiang, S., Rong, Y., Zhang, J., et al. (2025). Pangenome analysis reveals yield- and fiber-related diversity and interspecific gene flow in Gossypium barbadense L. Nat Commun *16*, 4995. https://doi.org/10.1038/s41467-025-60254-x.
14. Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol *36*, 875–879. https://doi.org/10.1038/nbt.4227.

15. Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., et al. (2024). Building pangenome graphs. Nat Methods *21*, 2008–2012. https://doi.org/10.1038/s41592-024-02430-3.

16. Hickey, G., Monlong, J., Ebler, J., Novak, A.M., Eizenga, J.M., Gao, Y., Human Pangenome Reference Consortium, Marschall, T., Li, H., and Paten, B. (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol *42*, 663–673. https://doi.org/10.1038/s41587-023-01793-w.

17. Miao, Z., Zhang, S., Ma, Z., Hakim, M.S., Wang, W., Peppelenbosch, M.P., and Pan, Q. (2019). Recombinant identification, molecular classification and proposed reference genomes for hepatitis delta virus. Journal of Viral Hepatitis *26*, 183–190. https://doi.org/10.1111/jvh.13010.

18. Chowdhury, S., Jacobsen, C., Depledge, D.P., Wedemeyer, H., Sandmann, L., and Kefalakes, H. (2025). Sequence analysis of the hepatitis D virus across genotypes reveals highly conserved regions amidst evidence of recombination. Virus Evolution *11*, veaf012. https://doi.org/10.1093/ve/veaf012.

19. Charre, C., Regue, H., Dény, P., Josset, L., Chemin, I., Zoulim, F., and Scholtes, C. (2023). Improved hepatitis delta virus genome characterization by single molecule full-length genome sequencing combined with VIRiONT pipeline. Journal of Medical Virology *95*, e28634. https://doi.org/10.1002/jmv.28634.

20. Jonkheer, E.M., van Workum, D.-J.M., Sheikhizadeh Anari, S., Brankovics, B., de Haan, J.R., Berke, L., van der Lee, T.A.J., de Ridder, D., and Smit, S. (2022). PanTools v3: functional annotation, classification and phylogenomics. Bioinformatics *38*, 4403–4405. https://doi.org/10.1093/bioinformatics/btac506.

21. Marcus, S., Lee, H., and Schatz, M.C. (2014). SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. Bioinformatics *30*, 3476–3483. https://doi.org/10.1093/bioinformatics/btu756.

22. Holley, G., and Melsted, P. (2020). Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. Genome Biol *21*, 249. https://doi.org/10.1186/s13059-020-02135-8.

23. Minkin, I., Pham, S., and Medvedev, P. (2017). TwoPaCo: an efficient algorithm to build the compacted de Bruijn graph from many complete genomes. Bioinformatics *33*, 4024–4032. https://doi.org/10.1093/bioinformatics/btw609.

24. Khan, J., and Patro, R. (2021). Cuttlefish: fast, parallel and low-memory compaction of de Bruijn graphs from large-scale genome collections. Bioinformatics *37*, i177–i186. https://doi.org/10.1093/bioinformatics/btab309.

25. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

26. Kuo, M.Y., Goldberg, J., Coates, L., Mason, W., Gerin, J., and Taylor, J. (1988). Molecular cloning of hepatitis delta virus RNA from an infected woodchuck liver: sequence, structure, and applications. J Virol *62*, 1855–1861. https://doi.org/10.1128/JVI.62.6.1855-1861.1988.

27. Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. Nucleic Acids Research *16*, 10881–10890. https://doi.org/10.1093/nar/16.22.10881.

ANNEXES

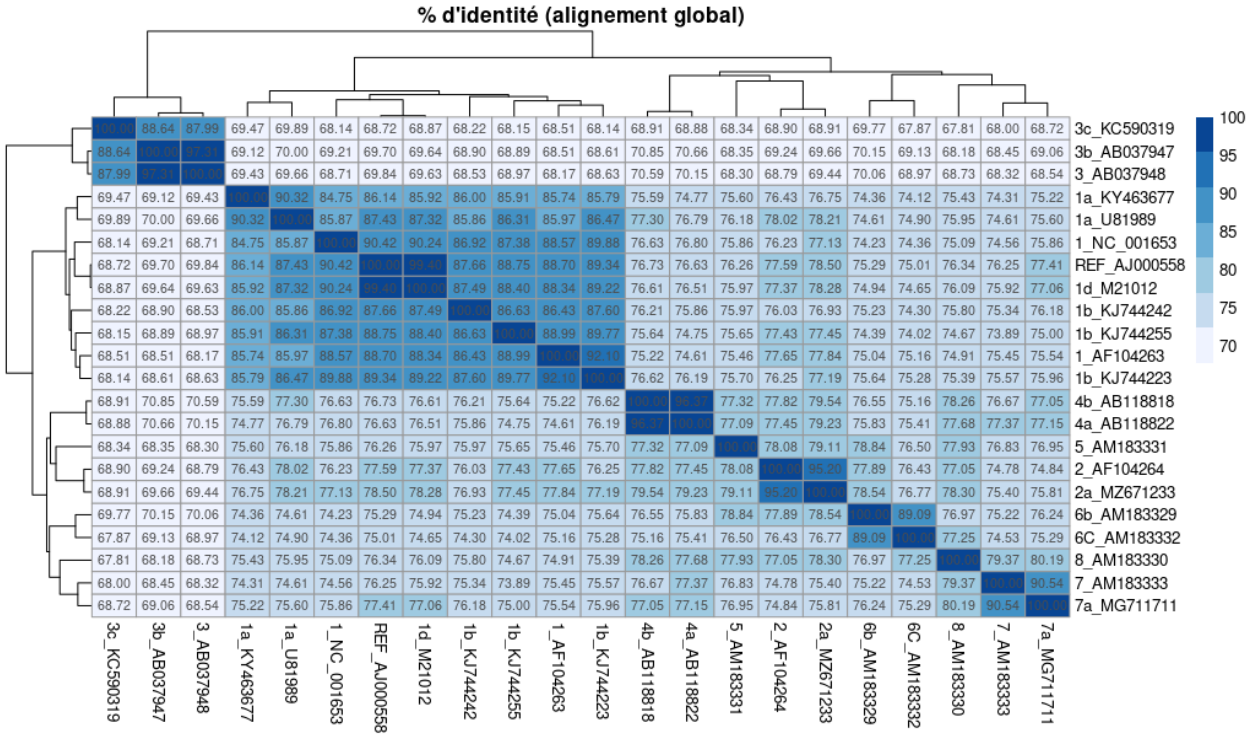


Figure S1. Pourcentage d'identité entre séquences du VHD de référence
Calculs avec la fonction pairwiseAlignment du package R Biostrings.
Representation heatmap générée avec le package R pheatmap.

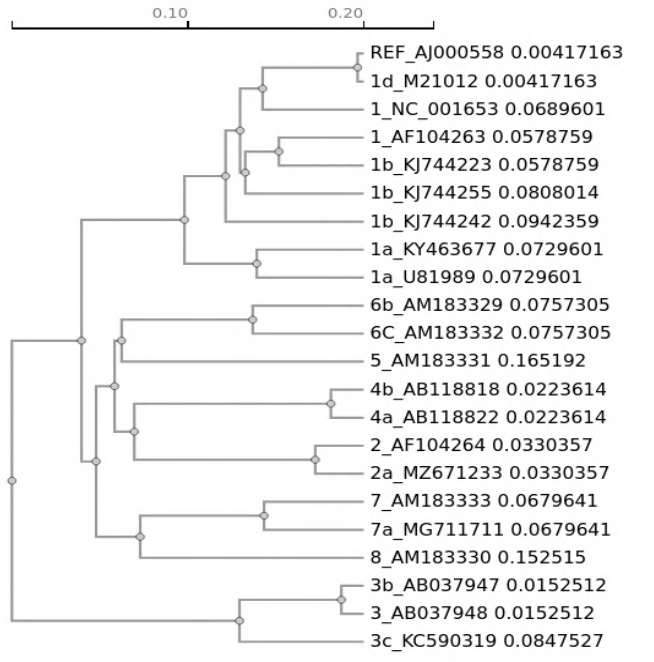


Figure S2. Arbre guide généré après l'alignement multiple avec clustal omega
Online sur EBI web server
(https://www.ebi.ac.uk/jdispatcher/msa/clustalo)

Tableau S1. Métriques proposées pour l'évaluation et la comparaison de graphes de pangénome

	Métrique	Définition	Calcul
Topologie	Nombre de noeuds	Nombre de nœuds du graphe	3 possibilités : • Nombre de lignes « S » du fichier GFA • gfatools stats / Number of segments • Bandage / Graph Information / Nodes
	Longueur totale	Somme des longueurs de tous les nœuds du graphe	3 possibilités : • Somme du 4ème champ des lignes « S » du fichier GFA • gfatools stats / Total segment length • Bandage / Graph Information / Total length
	Nombre d'extrémités libres	Nombre de nœuds connectés seulement dans une direction (arêtes seulement entrantes ou sortantes, mais pas les deux)	Bandage / Graph Information / More info / Dead ends
Alignement	aln.id (alignment identity)	Pourcentage de nucléotides identiques parmi ceux alignées	Pipeline PGGE : outil Peanut à partir du fichier .GAF produit par Graph Aligner
	qsc (query sequence coverage)	Pourcentage de nucléotides de la séquence alignés sur le graphe	
	uniq (unique alignments)	Alignements uniques sur le graphe	
	multi (multiple alignments)	Alignement d'une même portion de la séquence query sur plus d'une région du graphe	
	nonaln (non aligned)	Pourcentage de nucléotides non alignés sur le graphe	

Tableau S2. Métriques topologiques des graphes de pangénome générés avec les outils PGGB

Outil	PGGB	Minigraph Cactus	Cuttlefish
Nombre de nœuds	2854	1195	2107
Longueur totale	5437	17714	19095
Nombre d'extrémités libres	5	44	32