Apport du séquençage nanopore pour l'identification de variants de structure et l'étude de la méthylation

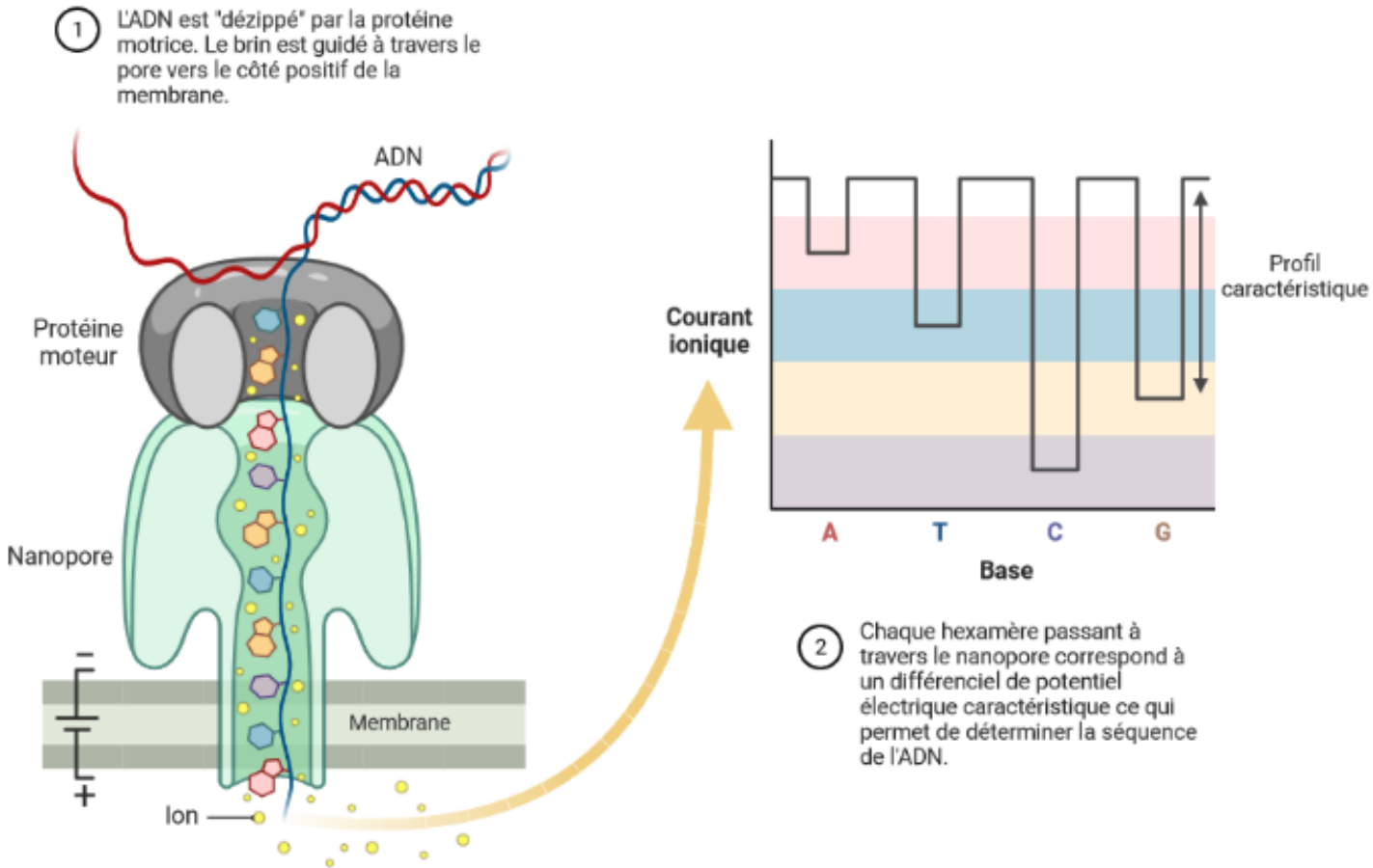
C. Bardel, N. Chatron, Q. Testard
G. Ben-Hassen, W. Desaintjean, Y. Kamili

2ème séminaire Bioinfo-Diag

16 mai 2023

Le séquençage nanopore

Principe du séquençage nanopore :



Le séquençage nanopore (2)

Intérêt du séquençage nanopore

- Détection de points de cassures
- Phasing
- Analyse des maladies à expansion
- Étude directe de la méthylation
- Adaptive sampling
- Assemblage de génomes
- Analyse de l'ARN
- Séquençage "sur le terrain"
- ...

Objectif de l'étude

Traitement de données de génomes

Séquençage sur un PromethION

- Prise en main de ces données, faisabilité de leur gestion dans un CHU
- Étude des variants de structure (SV) à l'échelle du génome
- Étude de la méthylation
- Analyse combinée pour étudier l'impact des SV sur la méthylation

Matériel

Les patients

- 14 patients atteints de syndromes malformatifs/déficience intellectuelle
- Points de cassure caractérisés par une analyse “short read” dans le cadre du projet ANI
  - 7 translocations X - autosome
  - 3 translocations entre autosomes
  - 2 remaniements complexes
  - 2 expensions

Données

- Séquençage du génome des 14 patients, 15 à 20x
- Séquençage réalisé sur un PromethION, flowcells 9.4.1, par la société KeyGene
- Fichiers fournis : données brutes (fast5) et fichiers fastq

Les données (suite)

Taille des données

Taille des données en Go

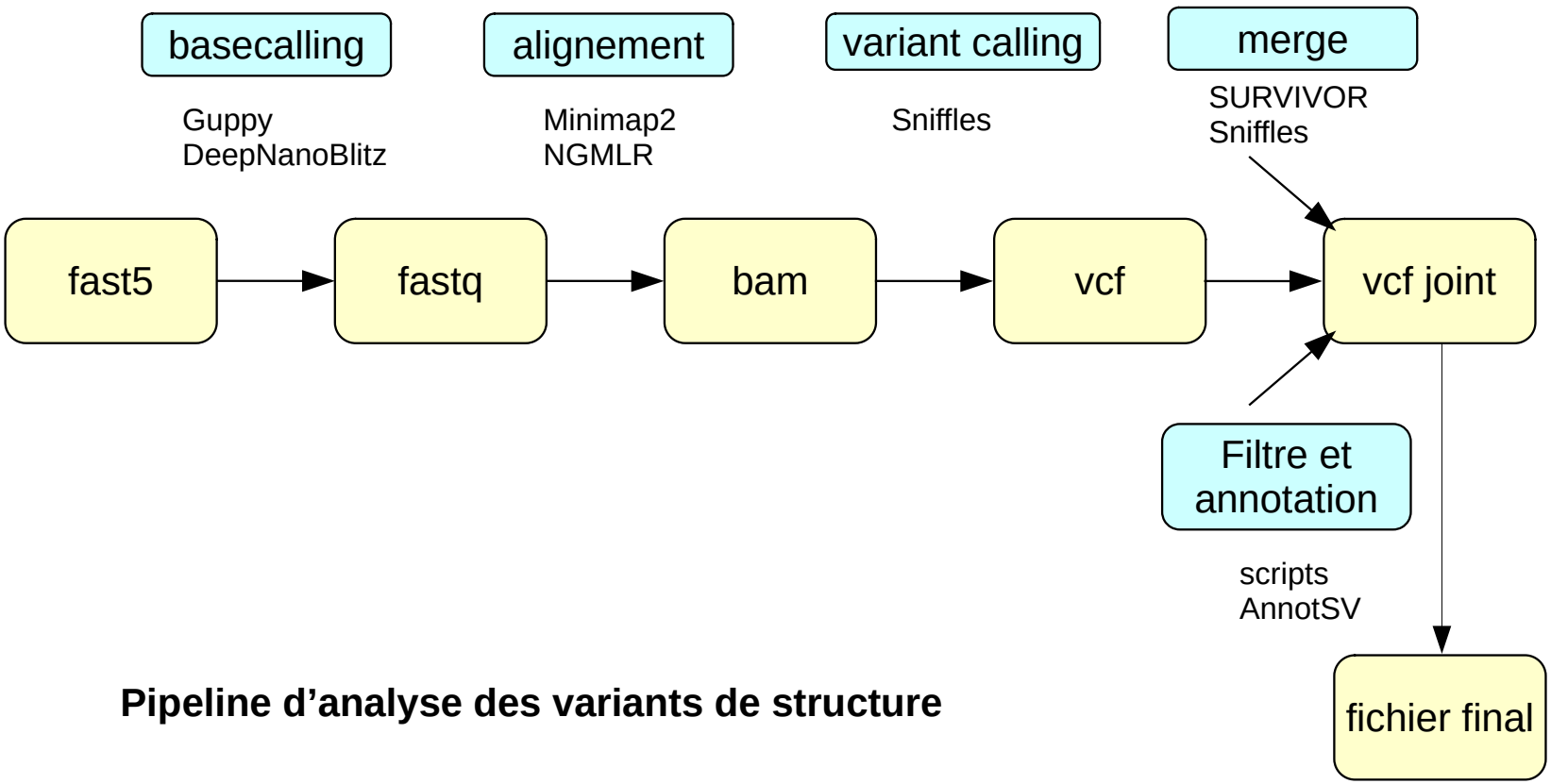
type	min	mediane	max
fast5	714	923	1900
fastq.gz	35	69	88

$\Rightarrow$ Au total : $\sim$ 16 To de données pour le projet

- Fichiers fastq.gz : même ordre de grandeur que du WGS short read
- Fichiers fast5 : beaucoup plus gros que les données brutes de WGS short read
- Amélioration constante des outils de basecalling
  $\rightarrow$ conservation des données

Prévoir de l'espace (stockage + travail) et un réseau rapide pour les transferts de données

Détection des variants de structure



Pipeline d'analyse des variants de structure

Basecalling

Choix de l'outil de basecalling

- Test de 2 outils : Guppy et DeepNanoBlitz (DNB)
- DNB : Pls modèles testés, du plus (7) au moins (1) précis
- Test réalisé sur 4000 reads, extrapolé pour 4 000 000 read
  - Puis test sur genome entier des plus rapides
- 1 patient : entre 4 et 9 millions de reads

Analyses réalisées sur 25 coeurs 2.9GHz

Outils	tps 4 000	tps 4M (estimé)
Guppy CPU	2h	83,3 jours
DNB CPU 7	25h	...
DNB CPU 6	1h	41,6 jours
DNB CPU 1	12 min	8,3 jours
Guppy GPU	2 min	1,4 jours

Basecalling (2)

Comparaison des 2 outils de basecalling

- Comparaison de Guppy GPU et DNB CPU 1
- Qualité des données :

	nb reads	Q7	longuest read	N50	frac non-sense
Guppy	9 083 052	86.2%	433 263	17 077	0.043
DNB	9 082 883	65.7%	200 475	15 128	0.241

- Nombre de SV : 2 à 3 fois plus avec DNB qu'avec Guppy

Nouvel outil proposé par ONT : Dorado

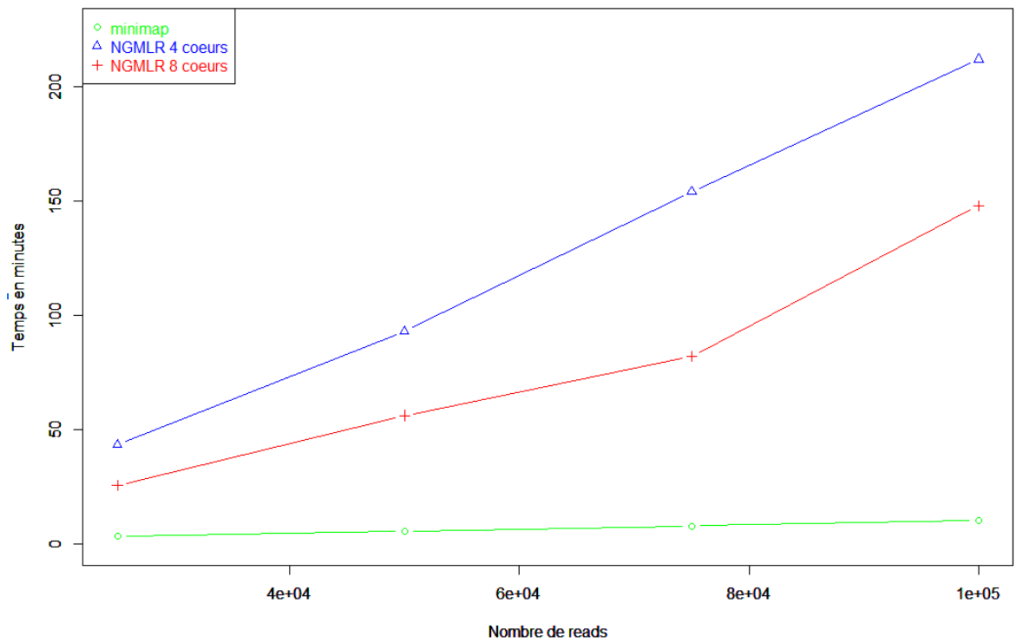
Plusieurs niveaux de précision (infos ONT)

- FAST : peu précis, mais rapide (analyse en temps réel)
- HAC : High Accuracy, précision intermédiaire. Temps de calcul : $\times 5$ à $\times 8$ /FAST. Utilisation en GPU.
- SUP : Super Accuracy, précision haute. Temps de calcul : $\times 3$ /HAP. Utilisation en GPU

Choix de l'outil d'alignement

Étude du temps d'alignement

- Données : fichiers fastq fournis par KeyGene
- Alignement sur GRCH38
- Test sur des sous échantillons de 25k, 50k, 75k et 100k reads
- Test de NGMLR (4 et 8 coeurs) et minimap2 (4 coeurs)



Sur un génome (9,1 M) :

- minimap2 : 14h
- NGMLR : $>$ 6,5 jours

Suite du pipeline

Logiciel d'appel de variant

- Test de Sniffles, CuteSV et SVIM
- Performance et temps de calcul correctes de Sniffles

Filtre et annotation

Combinaison des 14 VCF en un VCF de cohorte : SURVIVOR

- SV identiques si dmax $<$ 1000 bp et SV du même type
- Conservation de tous les SV, pas de tri sur la taille

Filtre du VCF unique :

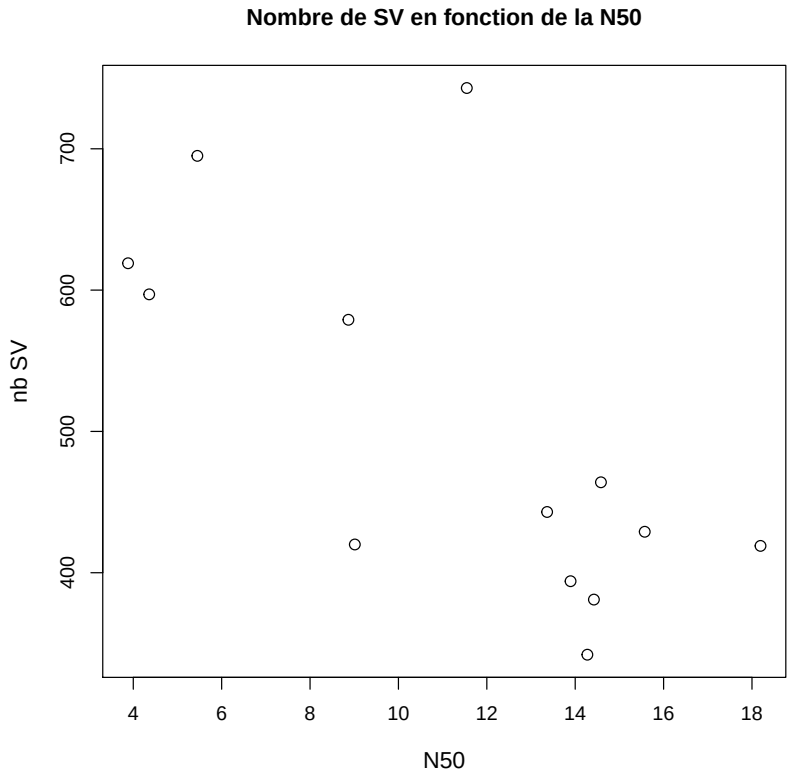
- Conservation des SV uniques $\rightarrow$ Passage de plus de 30 000 variants à 1 000 à 2 000 environ
- Élimination des variants génotypés 0/0 ou ./.
- Conservation des SV touchant uniquement les chr 1-22XYM
- Élimination des SV $<$ 100 bp $\rightarrow$ supprime près de 2/3 des DEL et INS

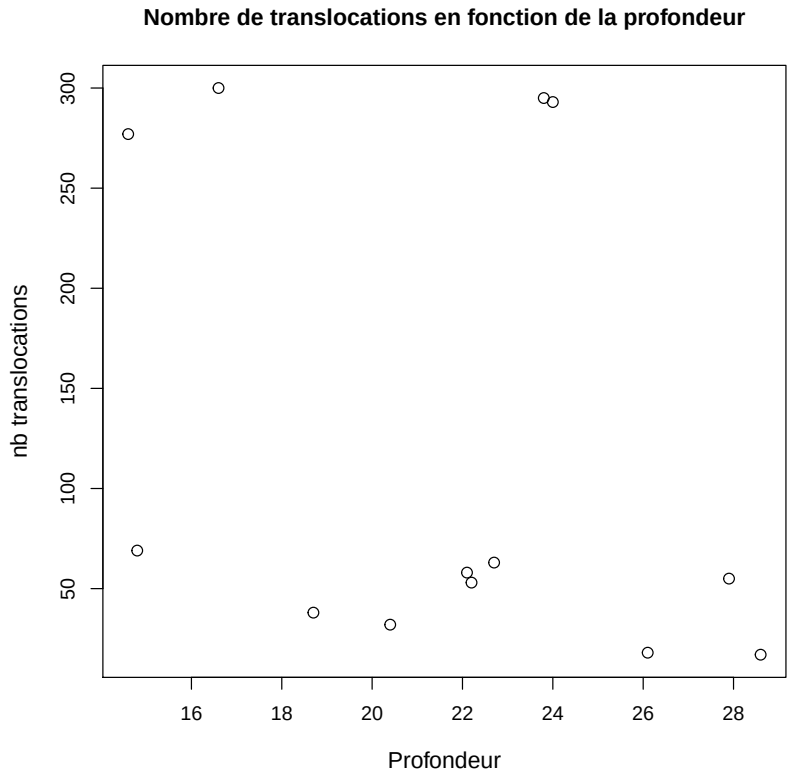
Variants de structure : description

Patient	Prof	N50	TRA	DEL	DUP	INS	INV	Total
15A1981	14.8x	15.6	69	140	1	180	39	429
15A2679	22.2x	14.6	53	206	10	191	4	464
15A2738	20.4x	14.3	32	135	3	169	3	342
15A2920	22.7x	9.8	63	153	10	190	4	420
15A3579	28.6x	14.4	17	163	4	193	4	381
15A4381	27.9x	13.4	55	195	2	186	5	443
15A79	16.6x	5.5	300	180	3	202	10	695
16A2542	22.1x	11.5	58	314	7	359	5	743
16A5984	18.7x	13.9	38	160	4	192	0	394
17A6136	23.8x	4.4	295	136	6	154	6	597
18A3551	10.5x	4.2	3319	151	13	169	21	3673
19A6796	14.6x	8.9	277	149	5	144	4	579
DUC-F	26.1x	18.2	18	171	6	223	1	419
PARS	24.0x	3.9	293	146	6	170	4	619

Variants de structure : description

Échantillon outlier éliminé



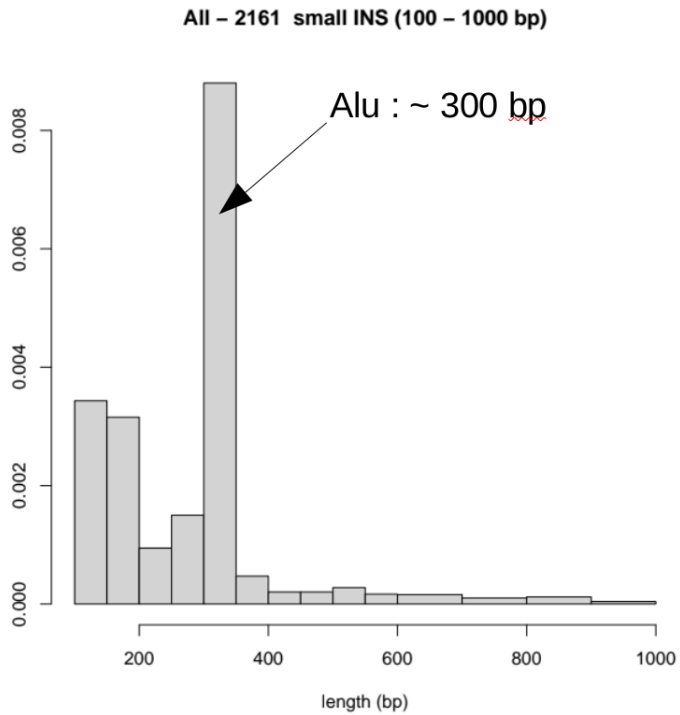


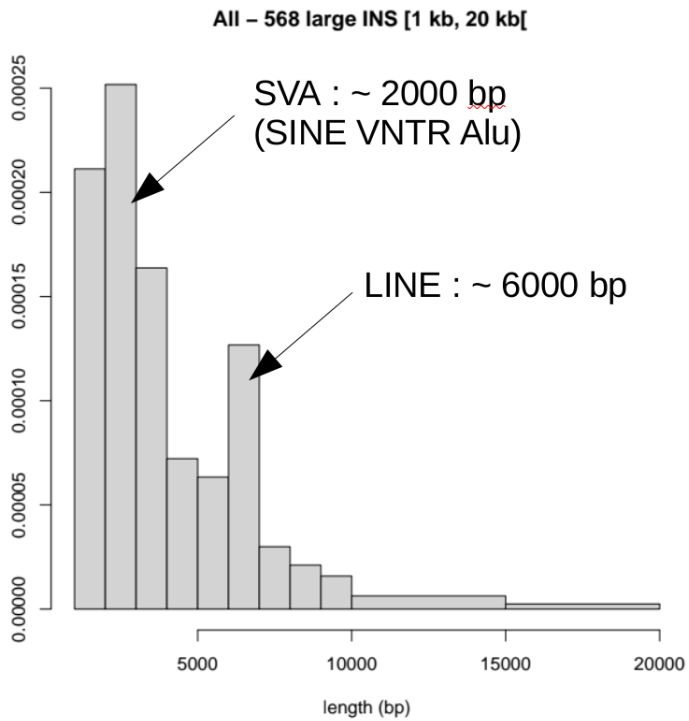
r=-0.68 (p=0.01032)

r=-0.34 (p> 0.05)

Reads + longs → moins de SV détectés, probablement moins d'artefacts

Variants de structure : description





- Profil de fréquence de taille conforme aux attentes
- Retrouvé de façon individuelle sur les patients
- Pour les DEL : pic des séquences Alu retrouvés, SVA et LINE peu clair

Comparaison au séquençage short read

Parmi les 30 points de cassure étudiés

- 29 avaient été trouvés avec le séquençage short read
- 23 ont été retrouvés de façon exacte ($<$ 1000 bp, mais quasi toujours à quelques bases max) avec les long reads
- 1 a été trouvé en plus
- 3 ont été trouvés avec des positions éloignées ($>$ 1000 bp)
- 2 ont été retrouvés exactement, mais appelés 0/0

Quelques perspectives

- Filtrer plus de variants
  - Travailler sur l'annotation
  - Enrichir la base pour filtrer les variations uniques
- Alignement T2T

Analyse de méthylation : test d'outils

Outils testés

- Guppy fast5mod
- Megalodon
- Nanopolish
- Tombo
- Deepsignal2 (version CPU et GPU)

Données

Test sur données publiques de NA12878

- Objectif : comparaison à une référence issue d'encode (Yuen et al., 2021), coefficient de corrélation de Spearman

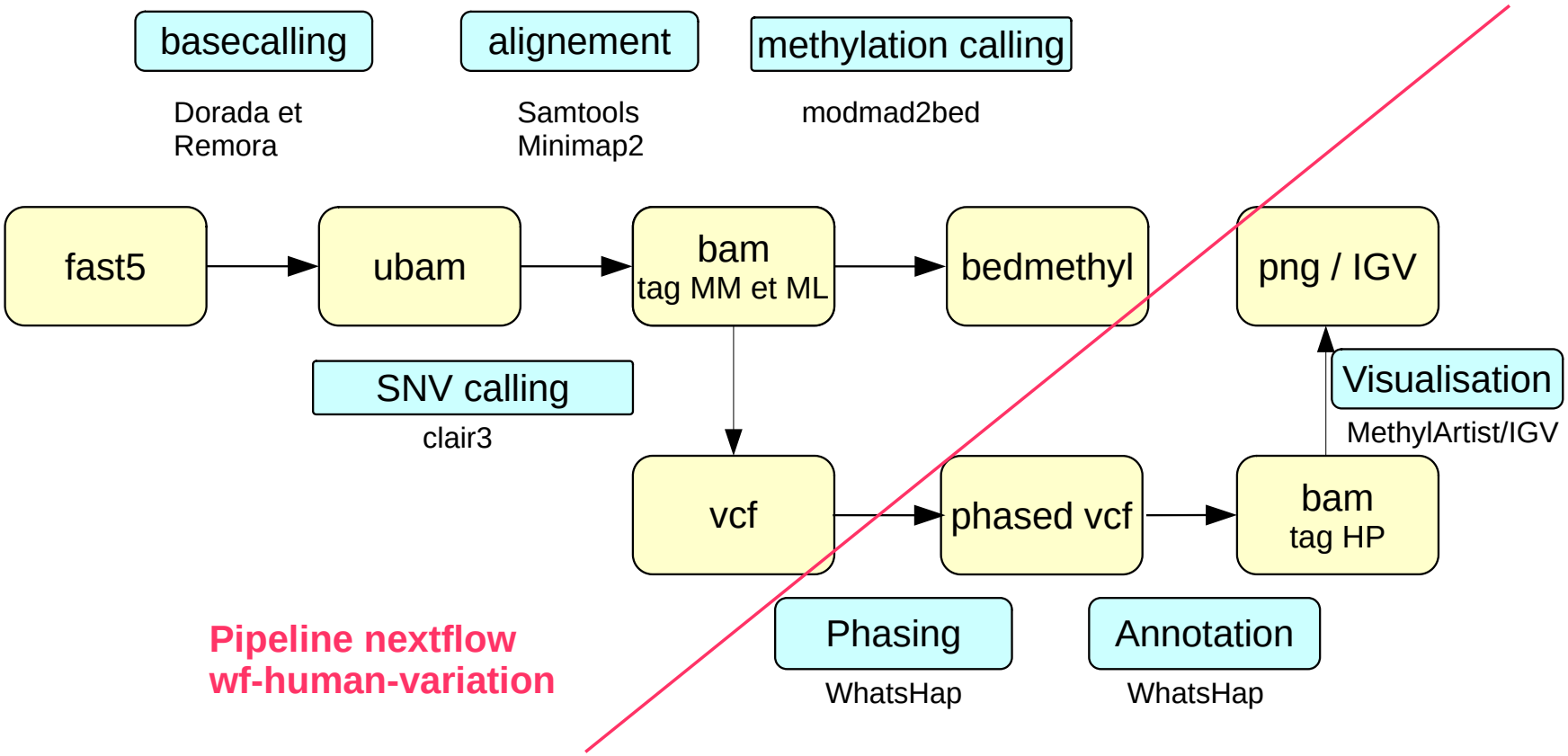
Étude de la méthylation (2)

Résultats

- Élimination de Tombo pour l'extraction des méthylations
  - identifie 3 fois moins de sites potentiellement méthylés que les autres
- Élimination de Nanopolish, Megalodon et Deepsignal2-CPU
  - Temps de calcul trop long ($>$ 7 jours)
- Comparaison Guppy fast5mod et Deepsignal2-GPU

Outils	Temps calcul	Corrélation	Nb sites non étudiés
Guppy fast5mod (CPU)	5j 5h 18min	0,78	42 millions
DeepSignal2-GPU	1j 1h 41 min	0,61	37 millions

Pipeline pour l'analyse de méthylation



Pipeline d'analyse de la méthylation

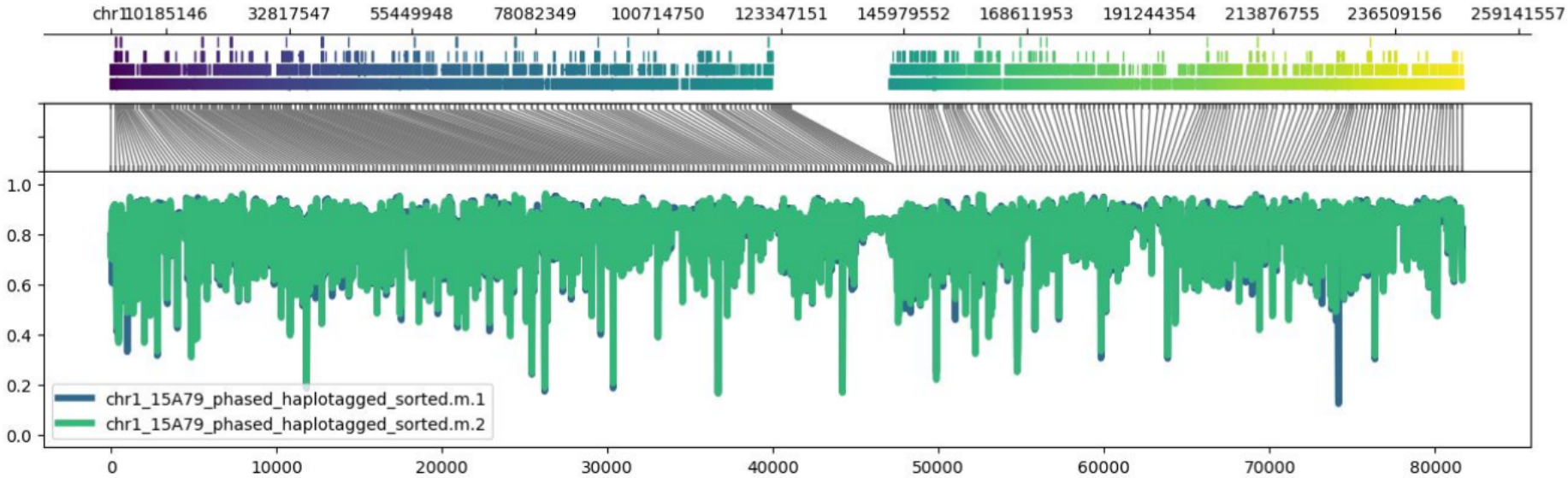
Temps de calcul :

- Basecalling : $\sim$ 16h (GPU)
- SNP calling : $\sim$ 16h (CPU)
- SV calling : $\sim$ 10h (CPU)

Caractéristiques du serveur :

- GPU : A6000
- CPU : 37 coeurs/74 threads
- 432 Go RAM

Methylation : 1er résultat

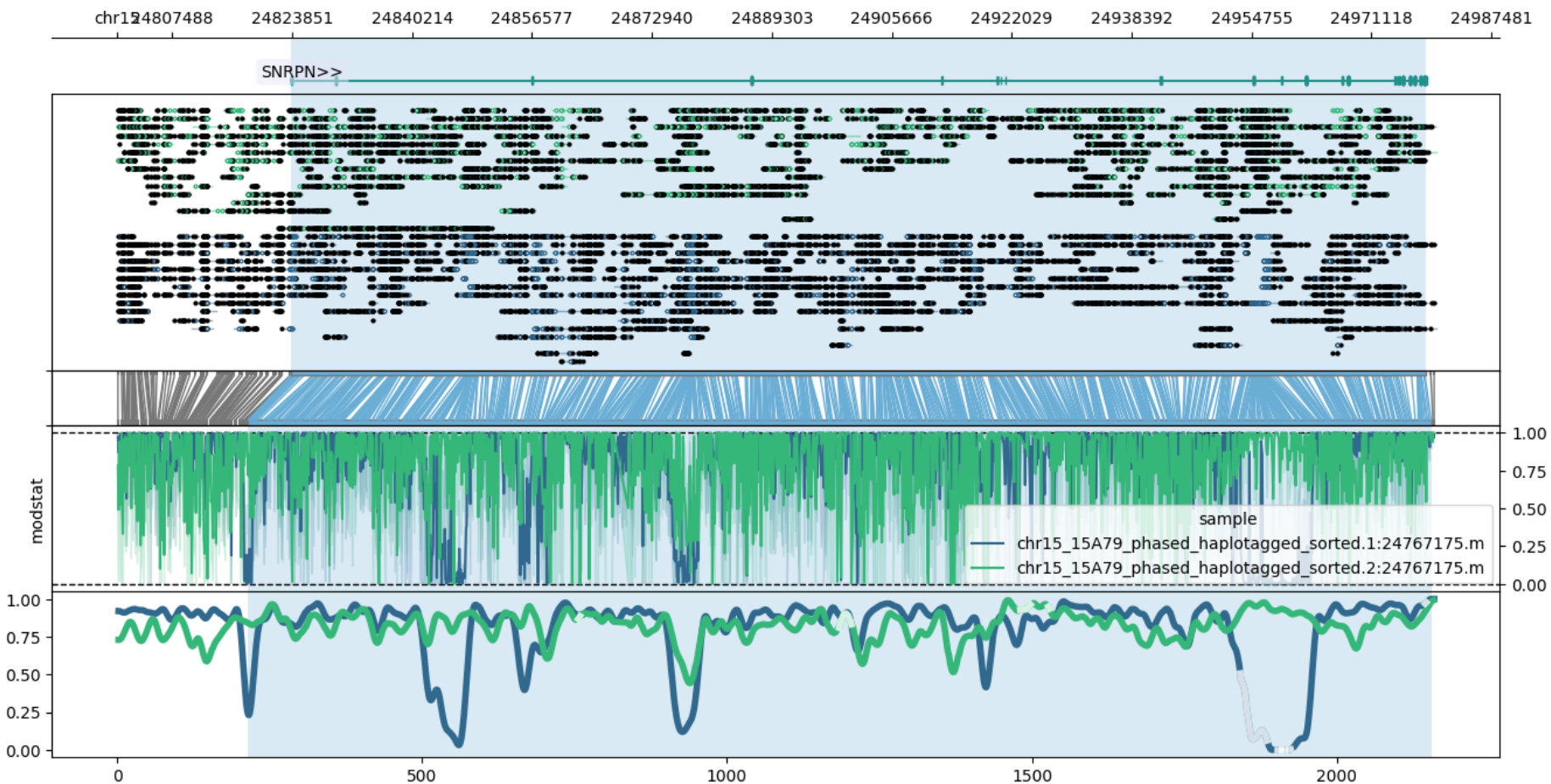


- Méthylation moyenne chr 1 : 72.8%
- Méthylation moyenne ensemble patients : 57.8%

Gershman Science. 2022 April ; 376(6588)

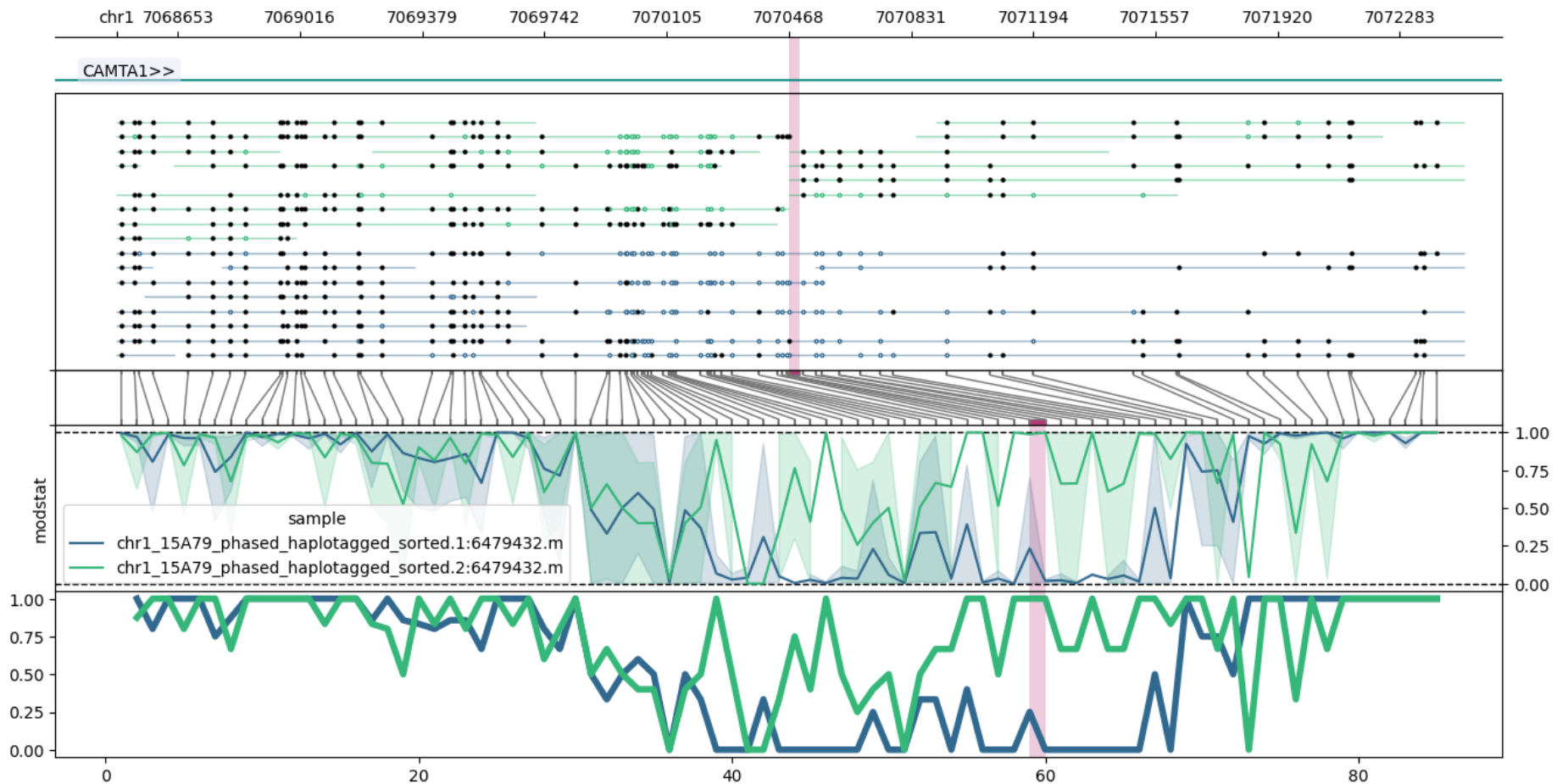
- Méthylation médiane cellules différenciées : 75%
- Méthylation médiane cellules peu différenciées : 36.8%

Région contrôle : SNRPN



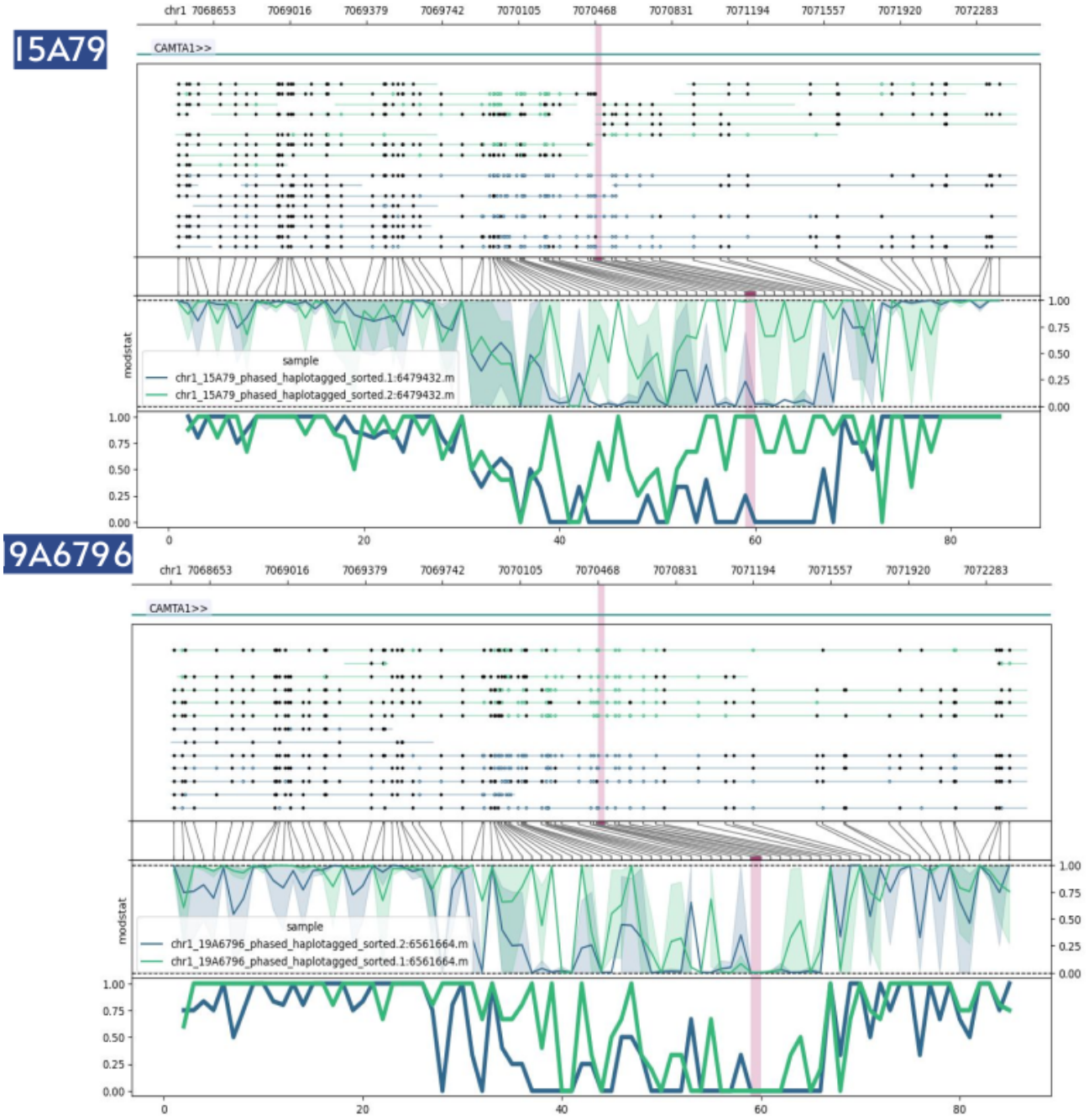
- Empreinte maternelle → différence de méthylation entre les 2 haplotypes
- Même profile chez les 2 patients étudiés

Analyse au niveau des points de cassure



Profil différent des 2 autres patients étudiés

Analyse au niveau des points de cassure



Conclusion et perspectives

Conclusion

Le traitement de ces données constitue un **challenge** pour un CHU

- Nécessité d'un espace de stockage conséquent
- Nécessité d'accéder à des GPU pour les étapes de basecalling

Points forts de ces données

- Possibilité d'analyser la méthylation à partir des mêmes données que pour l'étude des variants (nucléotidiques ou de structure)
- Identification de points de cassure dans des régions difficiles en SR

Perspectives

- Intégration des données de méthylation et de SV, comparaison des profils des 14 patients
- Étude de l'influence de différents facteurs sur l'appel de méthylation
- Tests d'outils de détection automatique des régions avec méthylation différentielle des haplotypes

Remerciements

Laboratoire de Cytogénétique des HCL

- Dr Nicolas Chatron
- Pr Damien Sanlaville
- Pr Caroline Schluth-Bolard
- Mme Flavie Diguet
- Mme Audrey Labalme

Cellule Bioinformatique des HCL

- Mr Quentin Testard
- Mme Genna Ben-Hassen (stage M2)
- Mr William Desaintjean (stage M1)
- Mme Yasmine Kamili (stage M1)

Moyens de calcul

- Cluster de calcul de l'IN2P3 : utilisation de la ferme GPU en 2021-2022
- DSN des HCL : financement du serveur de calcul avec carte GPU (2023)