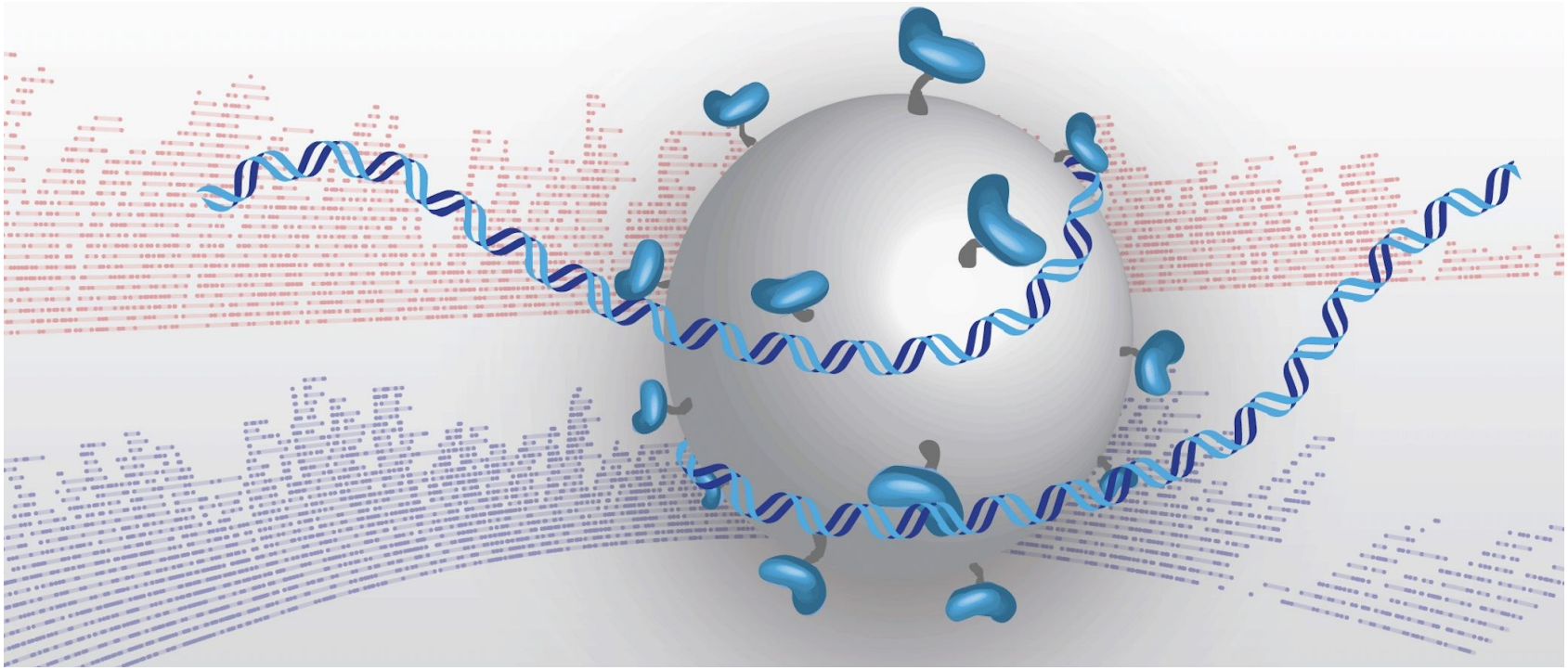
Haplotagging: état de l'art

Phasage de génomes à bas coût pour les études populationnelles



UNIVERSITÉ DE MONTPELLIER

cnrs

BiodivOc
Biodiversité Occitanie

Projet ***DevOCGen***

Montpellier
Ecologie
Evolution
Biodiversité

BiodivOc
Biodiversité Occitanie

Projet *DevOCGen* : Développement et applications de nouveaux outils pour la gestion et la conservation des populations naturelles à partir de données génomiques

Porteurs
Simon BOITARD
Raphaël LEBLOIS

Plateforme **GenSeq**
UAR MEEB

CBGP
Centre de Biologie pour la Gestion des Populations

Montpellier Ecologie Evolution Biodiversité

Un projet multi-partenaires Occitanie

isem
Institut des Sciences de l'Evolution-Montpellier

edb LABORATOIRE Évolution et Diversité Biologique
Comprendre la biodiversité

Centre d'Ecologie Fonctionnelle & Evolutive

GenPhySE

Theoretical and Experimental Ecology Station

IMAG
Institut Montpellierain Alexander Grothendieck

Laboratoire Genome et Developpement des Plantes

ISYEB
Institut de Systématique, Evolution, Biodiversité

Institut de Mathématiques de Toulouse

IEES Paris
Institut d'Ecologie et des Sciences de l'Environnement de Paris

NATURE EN OCCITANIE

Projet *DevOCGen* - Plateforme GenSeq

Axe méthodologique : Développement de l'haplotagging

Librairies à « lectures liées » permettant de reconstituer l'information de la phase haplotypique

Érick Desmarais

Frédérique Cerqueira

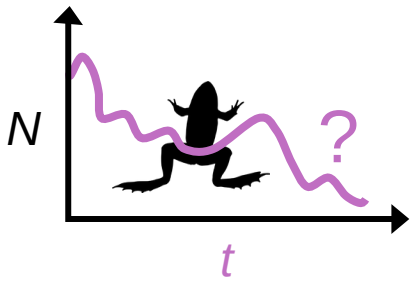
Anais Bordes

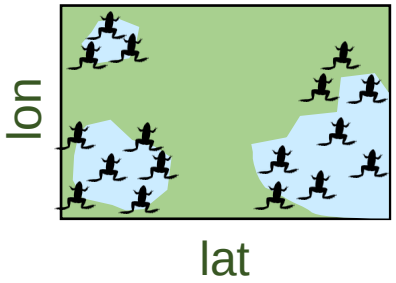
Cathy Liautard-Haag

Valentina Neglia

Plusieurs dimensions à renseigner :

le temps, l'espace, le génome



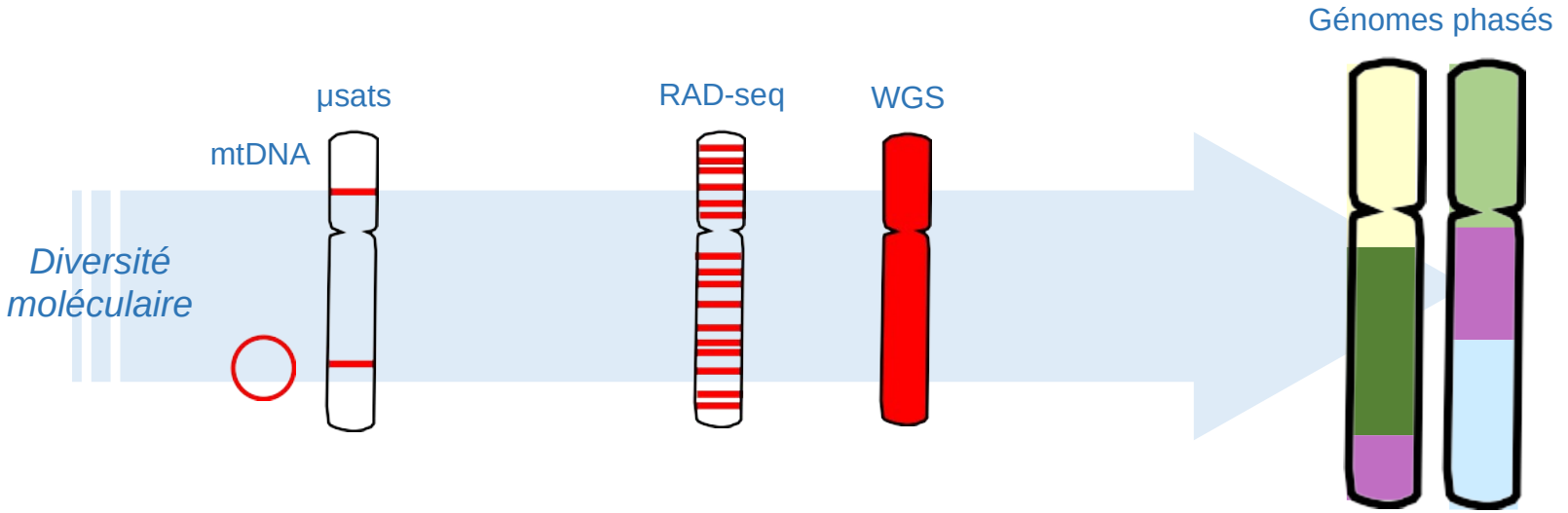


TEMPS ***ESPACE*** ***GENOME***

Défis :

- suivis temporels rares, partiels, ou inexistants
- respecter la démarche « ***éviter-réduire***-compenser »
- maximiser l'information extraite de chaque échantillon

Quelle information peut-on espérer extraire des données génomiques ?



Structure & Histoire des populations

Hybridation

Apparentement

Structure & Histoire des populations

Hybridation

Apparentement

Sélection

Ascendance & démographie récente

Recombinaison

Utilités de l’information haplotypique pour l’étude de la biodiversité

POUR FAIRE QUOI?

- Étude de la **recombinaison**
- Détection de la **sélection**
- Reconstruction de l’**histoire évolutive**
- Inférence de l’**ascendance** et de la **démographie récente**
- Reconstruction du **Graphe de Recombinaison Ancestrale** (ARG)

- Étude des **variants structuraux**
- **Assemblage** de génomes et de pangénomes

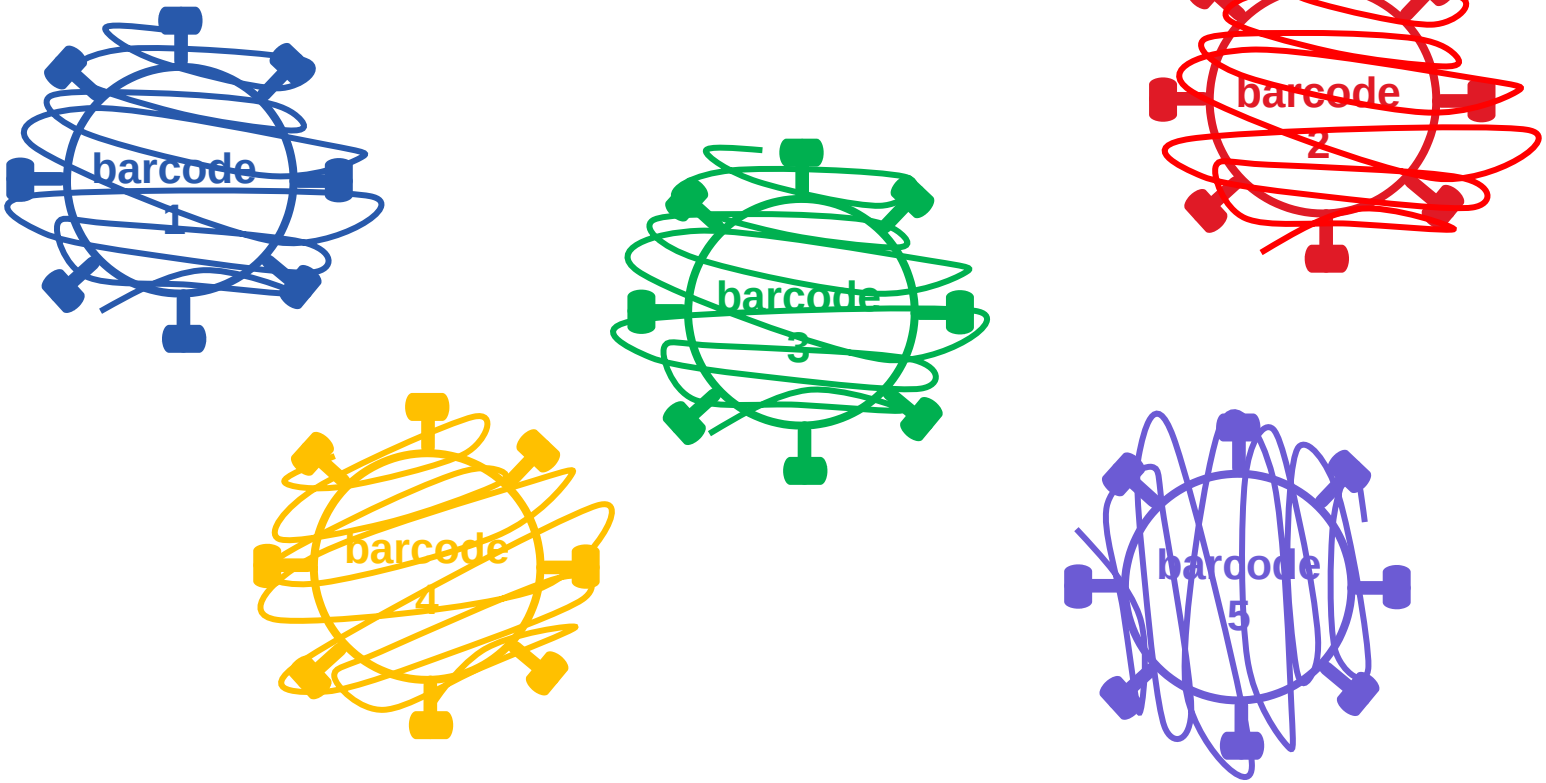
Utilités de l'information haplotypique pour l'étude de la biodiversité

COMMENT PHASER LES GENOMES ?

- **Séquençage short-read** (*Illumina*):
  - **+** économique, taux d'erreur faible
  - **-** perte de l'information haplotypique au-delà de ~1kb
- **Séquençage long-read** (*PacBio*, *ONT*):
  - **+** phase physique préservée (limitée par qualité de l'ADN)
  - **-** coûts mal adaptés aux études populationnelles
- **Séquençage linked-read** (Haplotagging):
  - **+** combine les deux avantages
  - **-** coûteux et débit limités

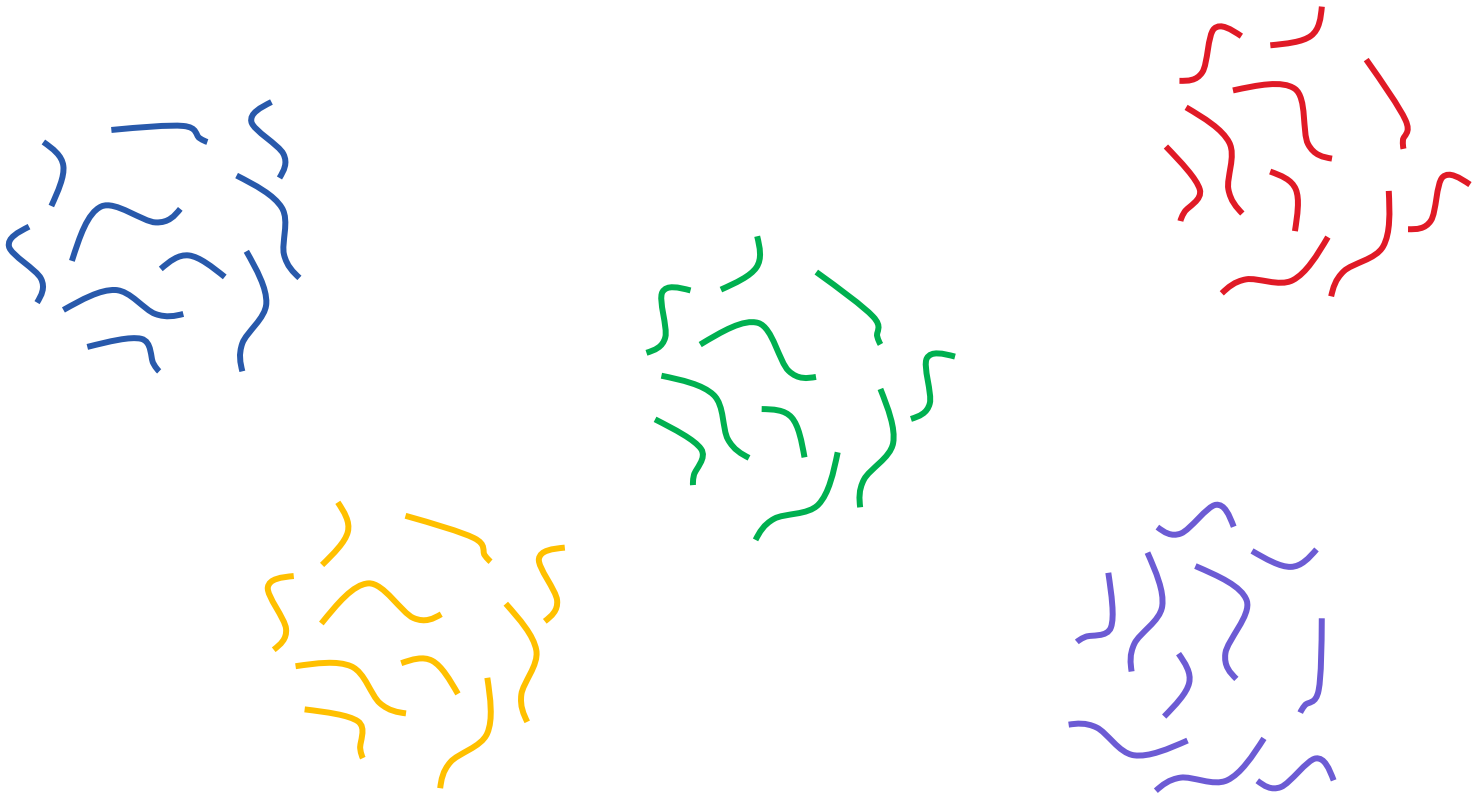
L’haplotagging

LINKED-READ QU’EST-CE QUE C’EST?

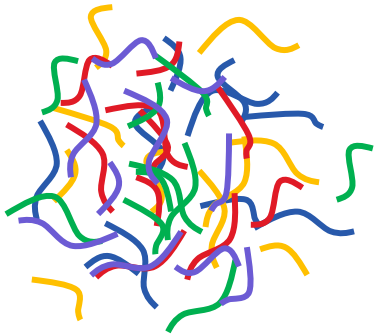


Meier et al., *PNAS*, 2021

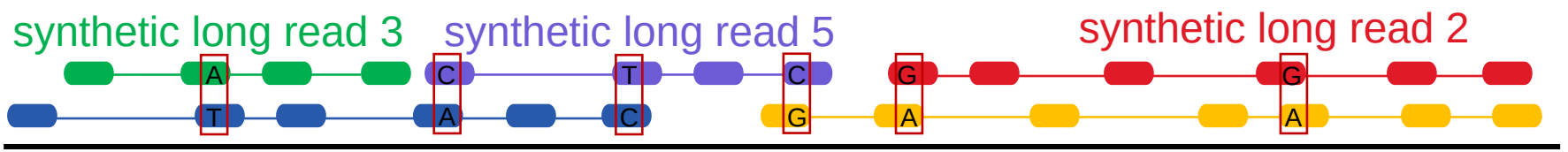
Un barcode/index par molécule



Reconstruction bioinformatique des reads longs synthétiques



séquençage short-read et mapping



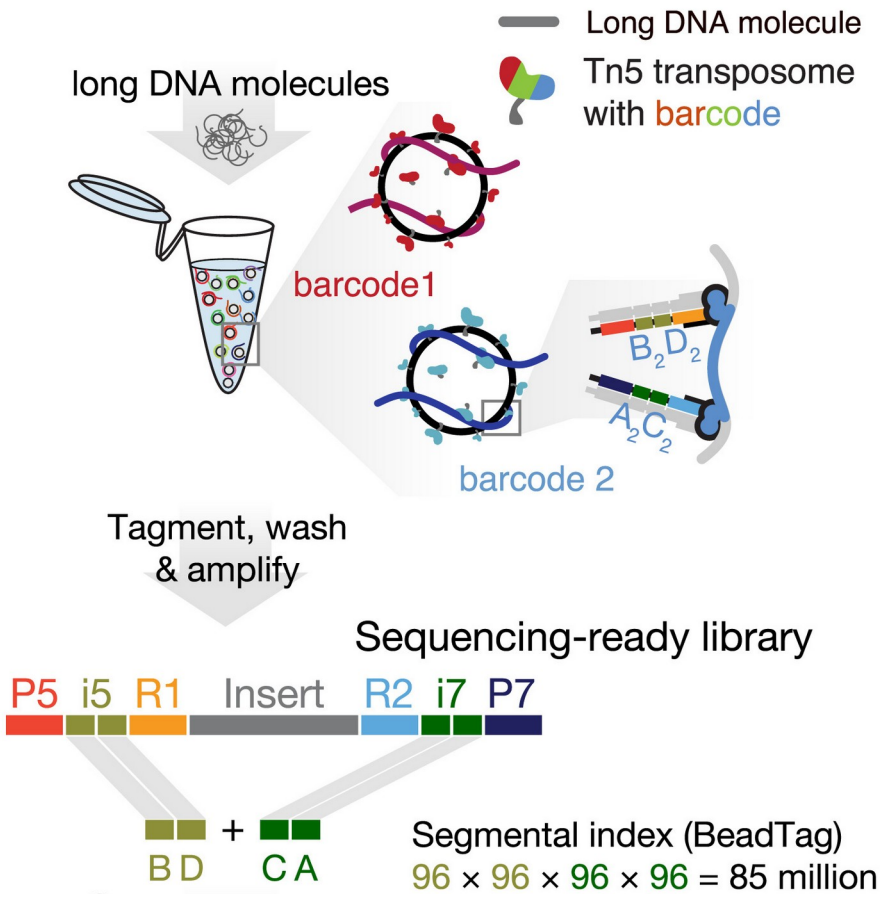
génome de référence

Phasage physique

ACTCGG *haplotype maternel*

haplotype paternel **TACGAA**

Récap

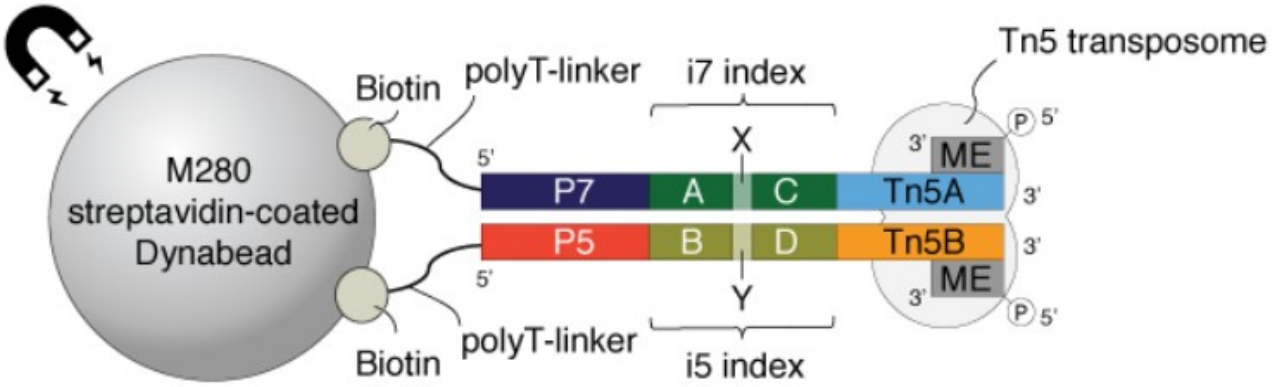


Meier et al., *PNAS*, 2021

L’haplotagging

COMMENT CA MARCHE?

- **Combine fragmentation et marquage** de l’ADN (transfert d’adaptateurs indexés) en une seule réaction rapide de tagmentation
- Utilise la **transposase Tn5 liée à des billes**
- **Chaque bille contient** des complexes Tn5-adaptateur avec **une combinaison unique d’index** (modules A-C sur i7, B-D sur i5)



Meier et al., *PNAS*, 2021

Un jeu de 96^4 billes différentes !

L’haplotagging

AVANTAGES

- Reconstitution de **reads longs synthétiques**
- **Phasage de blocks mégabasiques** à l’échelle individuelle
- **Réaction enzymatique rapide** avec PCR
- **Bas coût**, moins cher que les librairies Illumina commerciales
- Applicable chez une **large gamme d’organismes**

Meier et al., *PNAS*, 2021

Les étapes

1. Extraction/quali/quanti d’ADN de haut poids moléculaire
2. Construction des libraires d’haplotagging et séquençage
3. **Analyses bio-informatiques** pour le phasage des génomes

Matériel = HMW DNA

QUALITÉ DE L'ADN EST CRUCIALE POUR LA RÉUSSITE DU PHASAGE

Échantillonnage :

- Optimiser les conditions échantillonnage et de conservation
- Flash freezing Azote liquide, éthanol 96% ou RNAlater à -20/-80°C

Extraction :

- HMW MagBead (Zymo research)
- Monarch genomic DNA (NEB)
- Nanobind Tissue (PacBio)
- ¨Phénol-chloroforme

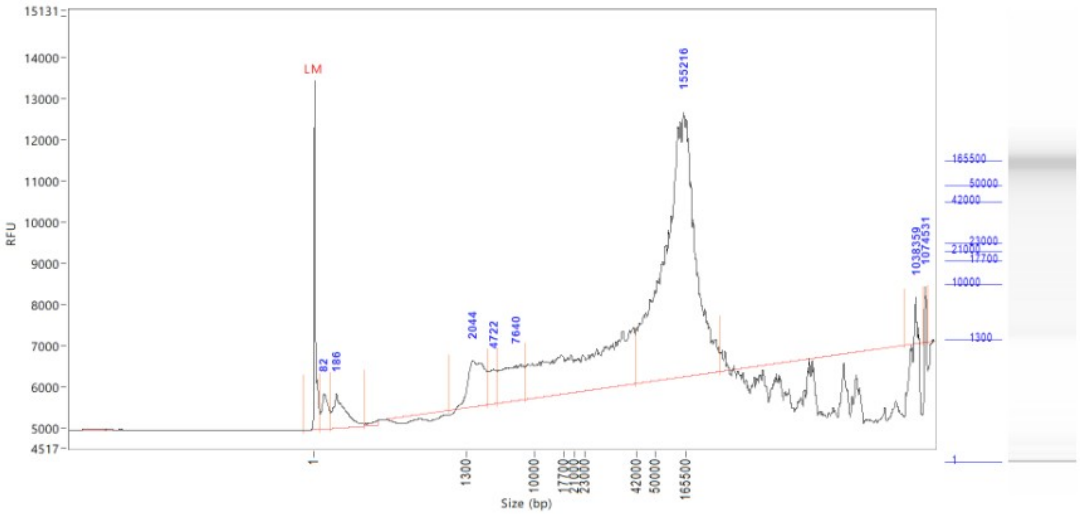
Élimination des fragments courts - sélection de taille :

- SRE kit (PacBio) pour éliminer le molécules <10 kb (moyen)

QUALITE - QUANTITE ADN

Vérification sur Femto Pulse (Agilent):

- Maximiser les fragments >10kpb



Agilent Femto Pulse
DNA profile

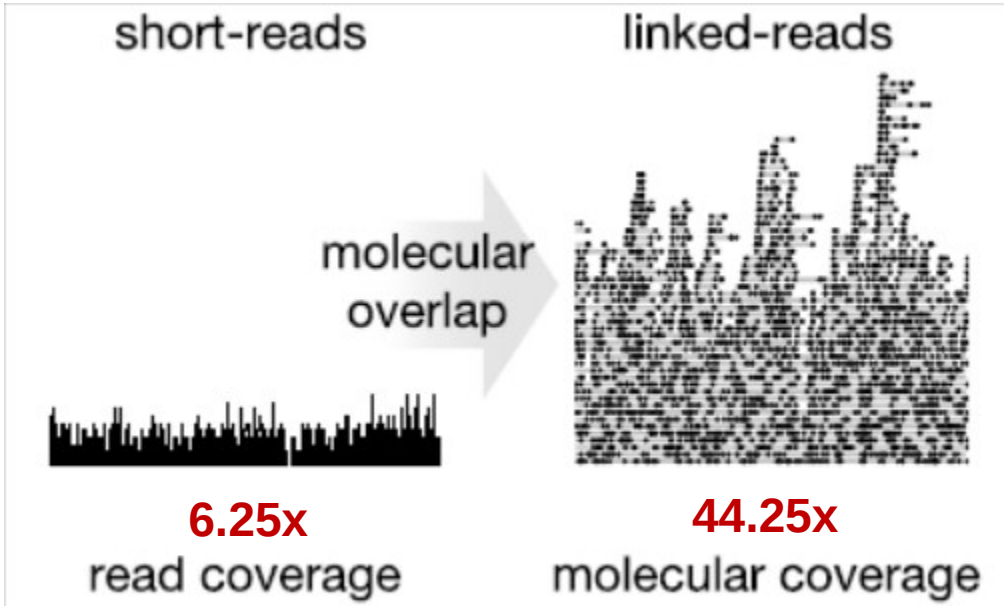
≥ 50% of DNA ≥ 30 kb

Très très peu d'ADN nécessaire: <1ng ADN/indiv

Préparation des ADN un grosse partie du temps et du budget

Notion de couverture moléculaire

Pour un génome de 1Gb - ligne S4
En utilisant des molécules d'ADN de 50kb en entrée :
Les molécules haplotaggées devraient couvrir 50^3 x 885000 = 44.2Gb
>> Couverture moléculaire théorique de 44.2x par individu



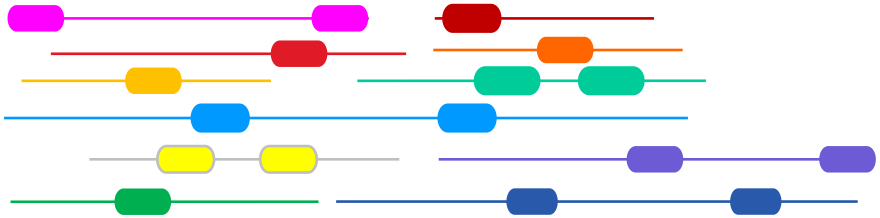
Objectif de couverture moléculaire de 50X
Avec couverture de reads de 5x et des ADN de taille moyenne de 50kb:
En moyenne 10 reads par molécule

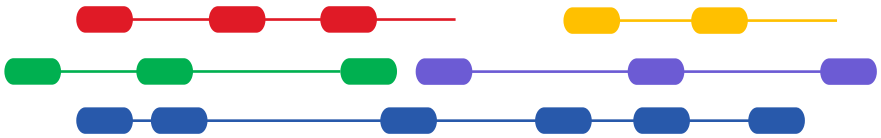
Design de construction des librairies

Choix FlowCell : xx million reads

Taille génome: Couverture moyenne par position génome (5x)

Nombre et taille moyenne des molécules ADN génomique





IMPORTANCE DE LA TAILLE DES MOLÉCULES
IMPORTANCE DE LA QUANTITÉ D'ADN UTILISÉ /
DIMINUER LE NOMBRE DE BILLES SÉQUENCÉES

Design de construction des librairies

Input Parameters

Number of individuals multiplexed:

0 — 96

0 10 20 30 40 50 60 70 80 90 96

Choose a flowcell type:

S4

Choose read length:

150

Haploid genome size (Mb):

0 — 451 — 5,000

0 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000

Median size of HMW DNA molecules (bp):

0 — 30,000 — 200,000

0 20,000 40,000 60,000 80,000 100,000 120,000 140,000 160,000 180,000 200,000

Number of beads used:

1,000,000 — 100,000,000

0 10,000,000 20,000,000 30,000,000 40,000,000 50,000,000 60,000,000 70,000,000 80,000,000 90,000,000 100,000,000

Number of HMW DNA molecules used:

4,000,000 — 100,000,000

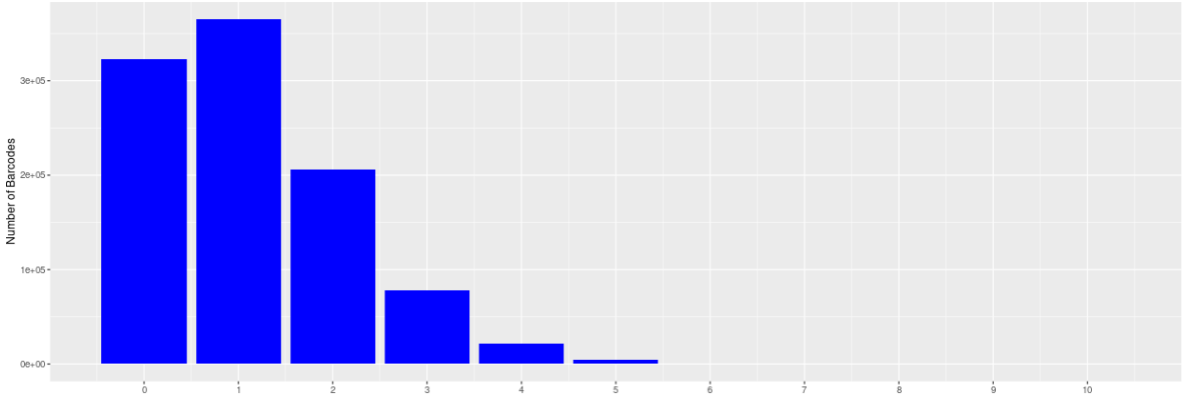
0 10,000,000 20,000,000 30,000,000 40,000,000 50,000,000 60,000,000 70,000,000 80,000,000 90,000,000 100,000,000

Fraction of the tagmentation volume used:

0 — 0.08 — 1

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Nombre de copies de chaque barcode :



Espèces déjà analysées en happlotagging à *GenSeq*

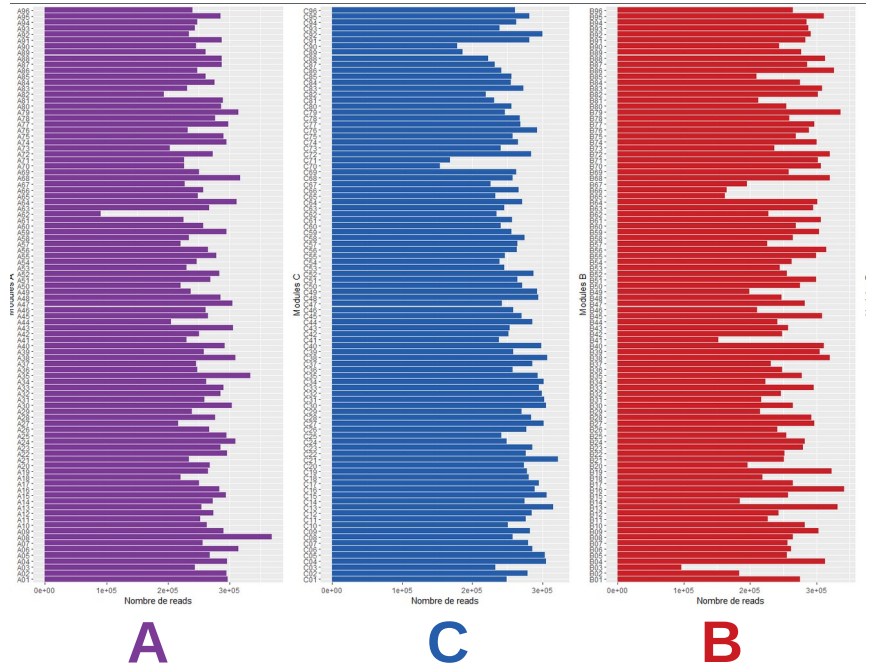








Plusieurs optimisations du protocole



A C B

Les coûts

1. **Extraction d'ADN de haut poids moléculaire (+ *Femto*)**

   Extraction : ~10€/échant

   Femto: ~10€/échant

   SRE: ~5€/échant

2. **Construction des libraires d'haplotagging**

   ~700€ par plaque de 96 échantillon

3. **Séquençage : ~ 3200€ pour un génome de 1Gb**

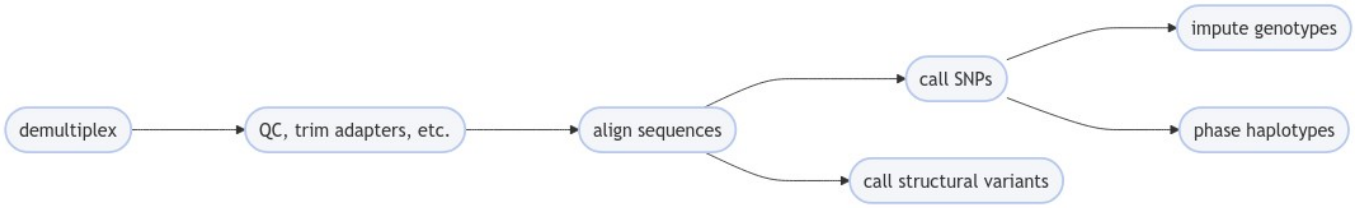
4. **Bioinformatique (stockage + analyse)**

Pipeline bio-informatique

Harpy: a pipeline for processing haplotagging linked-read data

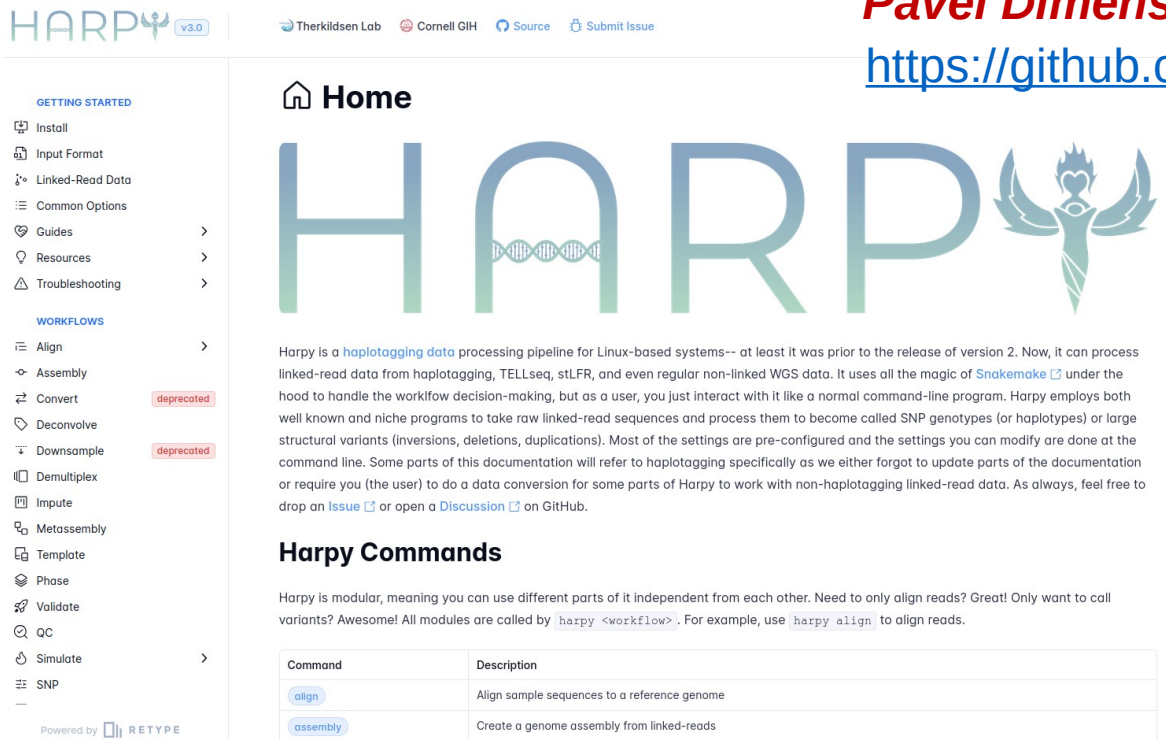
Pavel V. Dimens[1,*], Ryan P. Franckowiak[1], Azwad Iqbal[1], Jennifer K. Grenier[2], Paul R. Munn[2], Nina Overgaard Therkildsen[1]

Bioinformatics Advances, 2025, **00**, vbaf133
https://doi.org/10.1093/bioadv/vbaf133
Advance Access Publication Date: 5 June 2025
Application Note

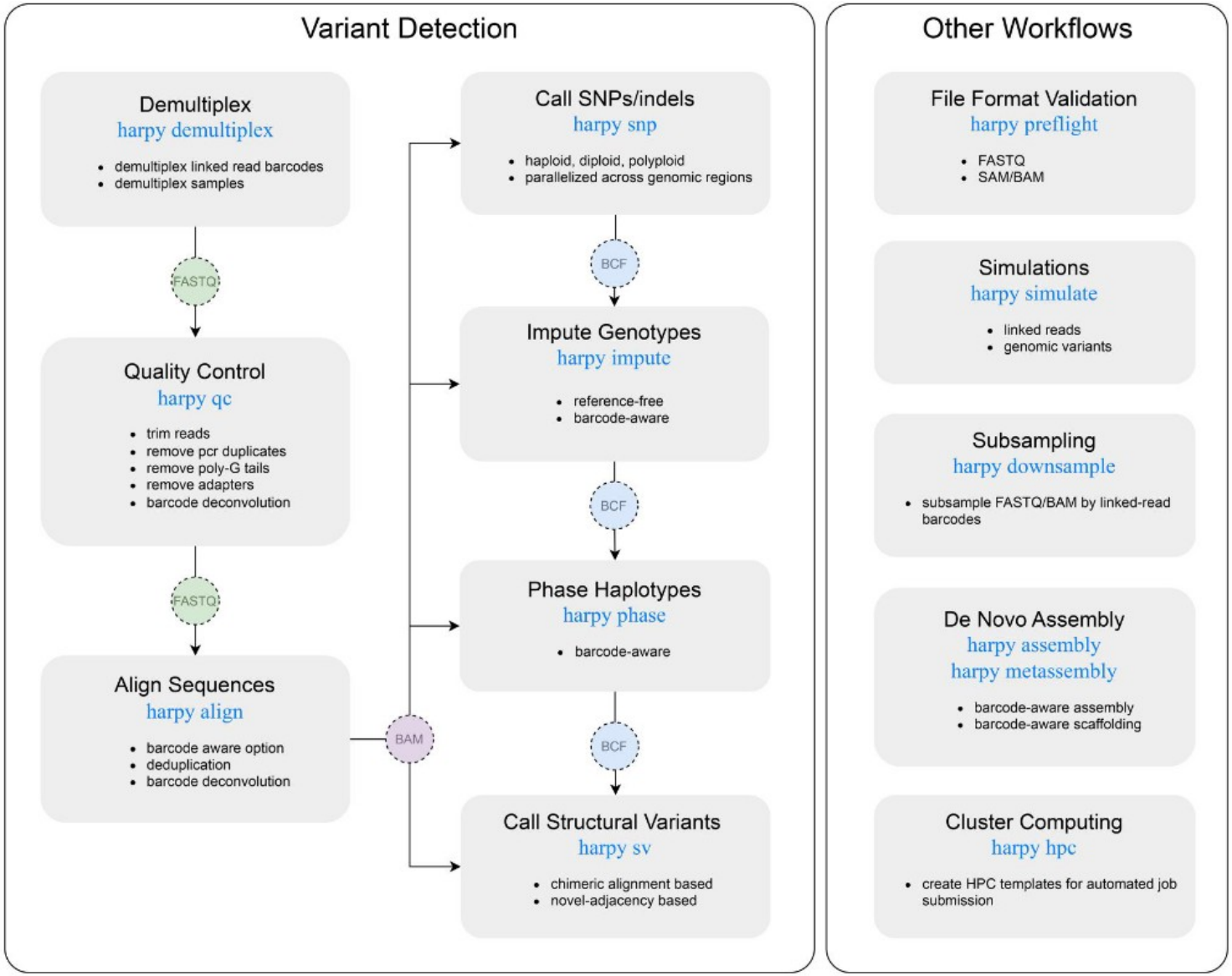


Pavel Dimens

https://github.com/pdimens/HARPY

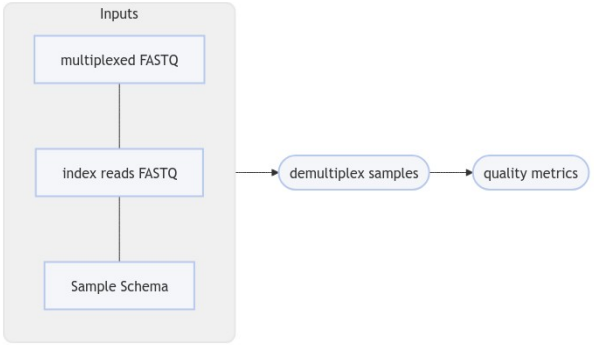


Pipeline Harpy



Dimens et al. 2025

Rapports Harpy : démultiplexage



Samples	Min % Valid	Max % Valid	Min BX Absent	Max BX Absent
96	75.773	98.521	0	0

General Per-Sample Haplotag Barcode Statistics

Below is a table listing all the samples Harpy processed and their associated haplotag barcode statistics as determined by the reads in the **forward read only**. If for some reason `TotalBarcodes` equals `0`, then there may be an issue with the format of your FASTQ headers.

CSV

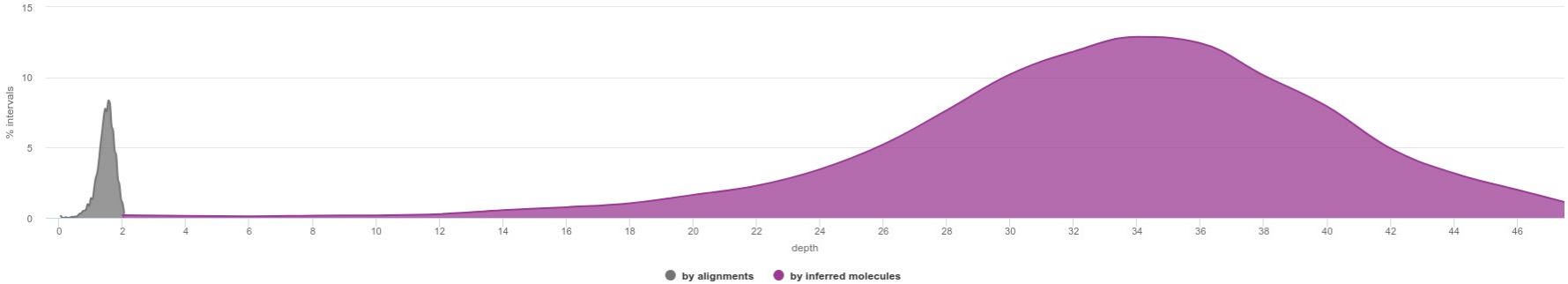
Sample	TotalReads	TotalBarcodes	ValidBarcodes	ValidPercent	InvalidBarcodes	InvalidPercent
All	All	All	All	All	All	All
Lsaxa_TRA_VPA_010	1856694	1856694	1807087	97.328	49607	2.672
Lsaxa_TRA_VPA_011	8699420	8699420	8518240	97.917	181180	2.083
Lsaxa_TRA_VPA_021	13392423	13392423	13130035	98.041	262388	1.959
Lsaxa_TRA_VPA_034	6180942	6180942	6044892	97.799	136050	2.201

Overall barcode quality per sample

Rapports Harpy : alignement

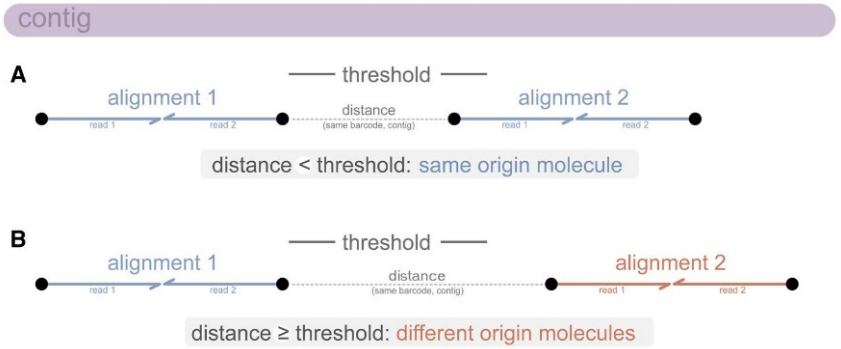
Contigs	Intervals	Average Depth	Stdev Depth	Mol. Average Depth	Mol. Stdev Depth
63	50 kb	1.46	0.3	32.29	7.08

couverture reads *vs* molécules synthétiques :



inferred molecules calculate "effective" or "molecule" coverage

Etape de déconvolution :

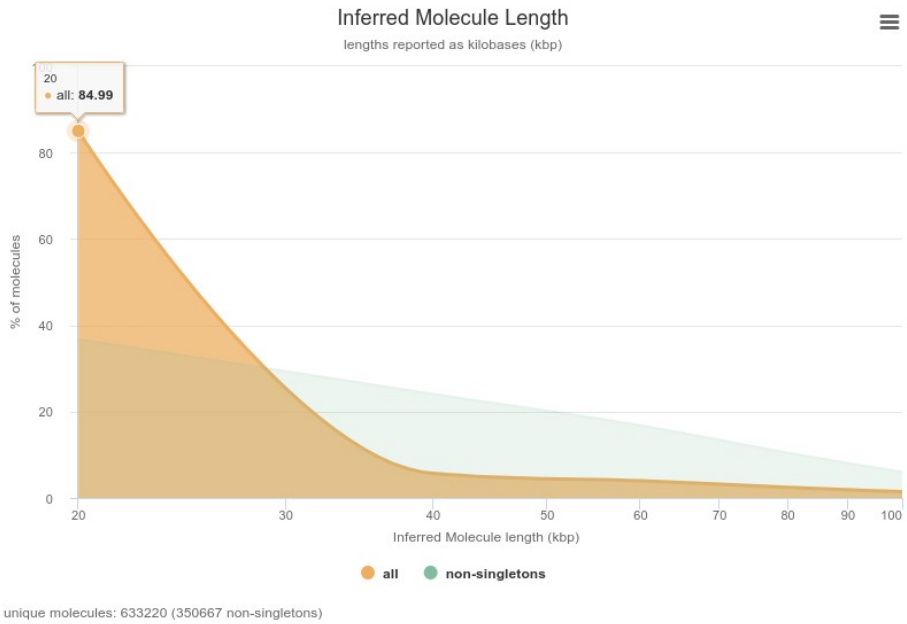


Rapports Harpy : alignement

Contigs	Unique Barcodes	Molecule Threshold	Unique Molecules	Valid BX Records	Invalid BX Records	Singletons (%)	Avg molecules per contig
2,751	633,220	200,000	1,477,736	2,229,336	130,732	76.27	538

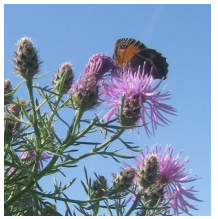
CSV

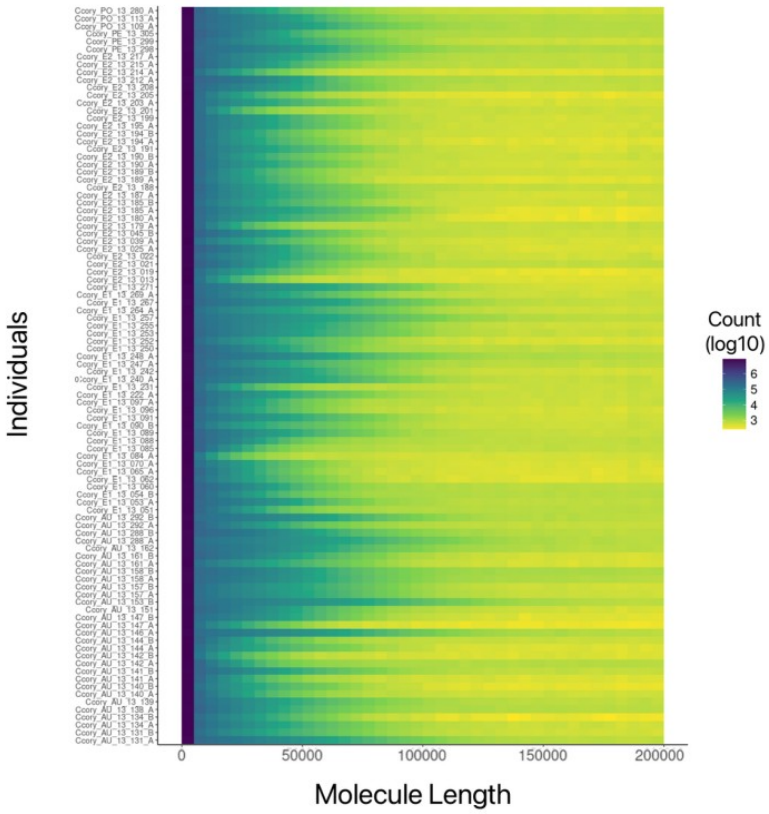
contig	valid BX records	invalid BX records	molecules	N50	N75	N90
7	140434	8161	91487	61403	37267	19609
0	131344	7845	84806	61214	37206	19892
3	129261	7754	83637	61014	37051	19967
6	120673	6981	78249	61018	37419	20024
2	116393	6882	75984	60772	36652	19529
4	106623	6190	69440	60383	36279	19315



unique molecules: 633220 (350667 non-singletons)

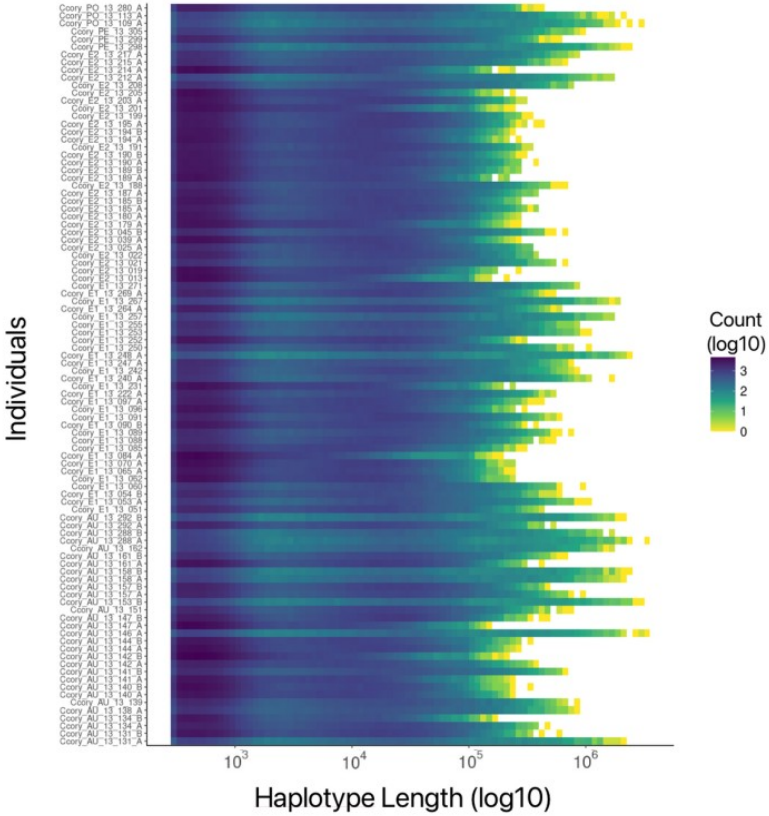
Imputation et phasage physique





STITCH imputation

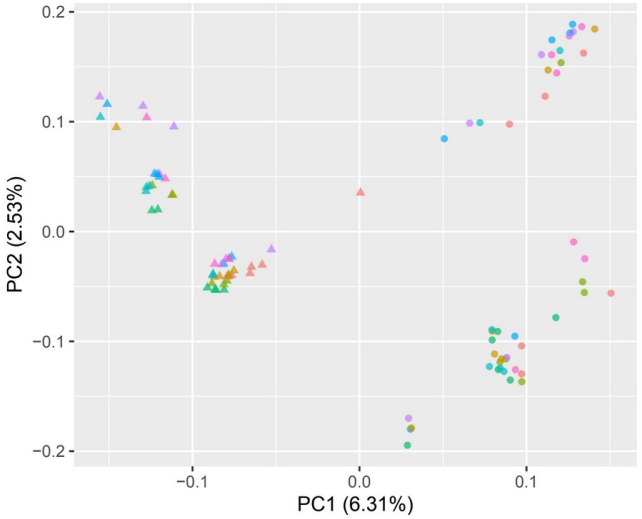
HAPCUT2 phasing

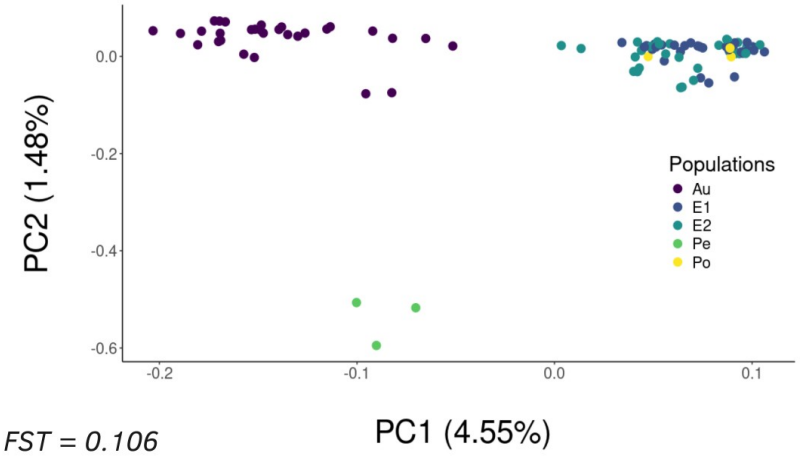


Synthetic molecule length

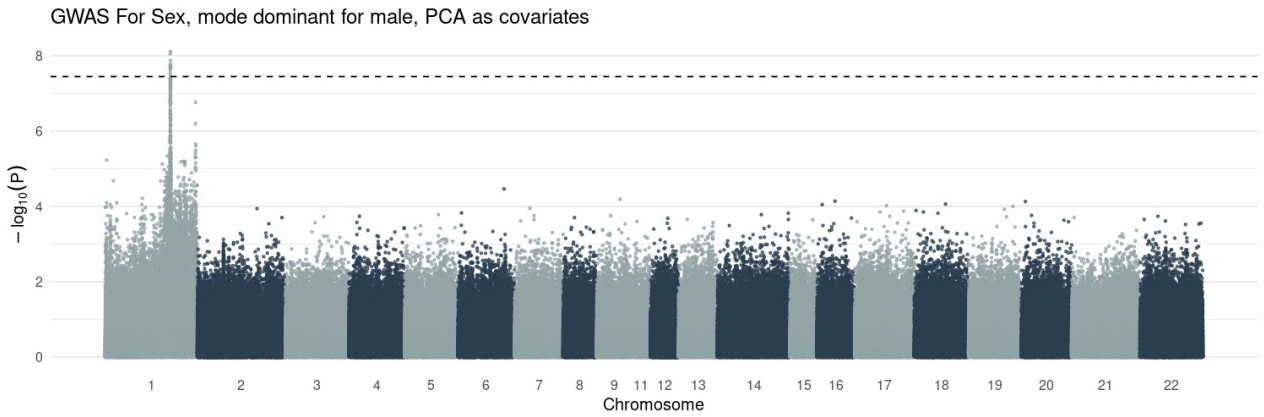
Phased haplotype block length

Analyses de structure génétique



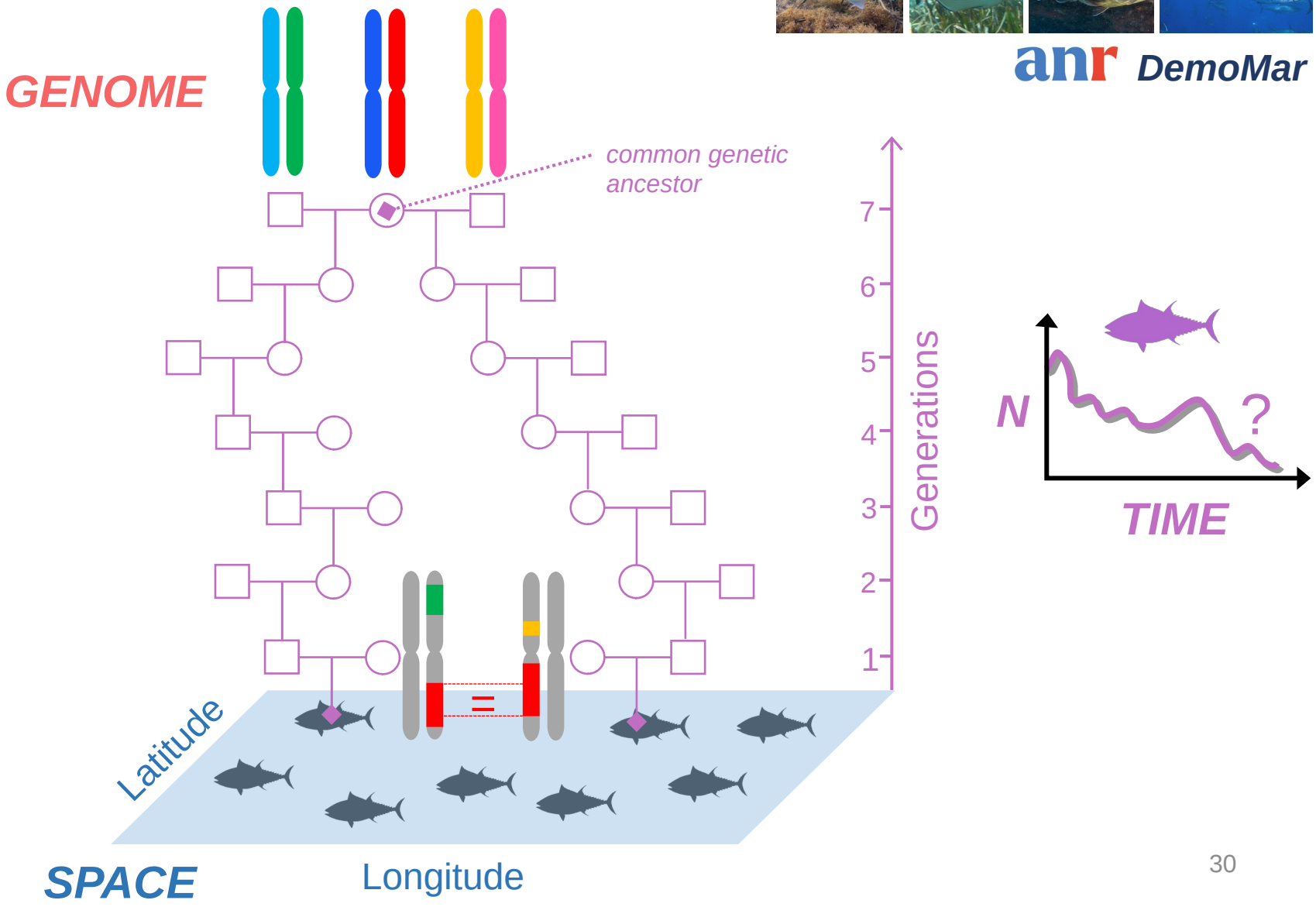


FST = 0.106



Analyse des co-ascendances récentes

anr DemoMar



CONCLUSIONS : exemples d’applications de l’haplotagging

- ✓ Détection des segments IBD pour étudier la démographie récente (centaurée)

- ✓ Détection de signaux d’introgression adaptative (cione)

- ✓ Détection d’inversions chromosomiques et de leurs points de cassure (hippocampes)

Remerciements

L'équipe *GenSeq* (UAR MEEB)

Partenaires du projet *DevOCGen*

Groupe de discussion haplotagging

MPI : Cecile Molinier, Lauric Reynes, Frank Chan, Marek Kucka

Roscoff : Claire Daguin, Thomas Broquet, Alexis Simon

Rennes : Claire Mérot, Claire Lemaitre, Fabrice Legeai

Paris : Pierre de Villemereuil, Mélanie Januario, Pascaline Chifflet-Belle, Elise Gay

Développement et soutien bioinfo

Harpy **:** Pavel Dimens

Plateforme MBB : Khalid Belkhir, Benjamin Penaud, Iago Bonnici