Construction and characterization of a comprehensive ovine pangenome from 10 breeds provides new insights into their genetic diversity

V. Sorin, C. Birbes, C. Eché, C. Marcuzzo, J. Sarry, A. Suin, C. Donnadieu, C. Gaspin, C. Iampietro, D. Milan, C. Klopp, M. Boussaha, L. Drouilhet, MP. Sanchez, G. Tosser-Klopp

Projet cofinancé par le Fonds Européen de Développement Régional

> Context and objectives

Context:

⟶ Current ovine genome reference assembly
- Based on a **single Rambouillet sheep**
- **Does not capture all** genetic variations

⟶ **Structural variants**
- Genomic variations ≥ **50 nucleotides**
- Potential **impact** on both **complex** and **Mendelian phenotypic variations**
- **Difficult to detect**

⟶ Advances in **long-read sequencing** technologies
- Construct ***de novo* genome** assemblies

Objectives

Projet cofinancé par le Fonds Européen de Développement Régional

> Context and objectives

Context:

⟶ Current ovine genome reference assembly
- Based on a **single Rambouillet sheep**
- **Does not capture all** genetic variations

⟶ **Structural variants**
- Genomic variations ≥ **50 nucleotides**
- Potential **impact** on both **complex** and **Mendelian phenotypic variations**
- **Difficult to detect**

⟶ Advances in **long-read sequencing** technologies
- Construct ***de novo* genome** assemblies

Objectives

01

Construct an **ovine pangenome graph** using *de novo* assemblies from 10 French sheep breeds

Projet cofinancé par le Fonds Européen de Développement Régional

> Context and objectives

Context:

⟶ Current ovine genome reference assembly
- Based on a **single Rambouillet sheep**
- **Does not capture all** genetic variations

⟶ **Structural variants**
- Genomic variations ≥ **50 nucleotides**
- Potential **impact** on both **complex** and **Mendelian phenotypic variations**
- **Difficult to detect**

⟶ Advances in **long-read sequencing** technologies
- Construct ***de novo* genome** assemblies

Objectives

01 Construct an **ovine pangenome graph** using *de novo* assemblies from 10 French sheep breeds

02 Identify and study **structural variants** (SVs) & **non-reference sequences** (NRSs)

Projet cofinancé par le Fonds Européen de Développement Régional

> Context and objectives

Context:

⟶ Current ovine genome reference assembly
- Based on a **single Rambouillet sheep**
- **Does not capture all** genetic variations

⟶ **Structural variants**
- Genomic variations ≥ **50 nucleotides**
- Potential **impact** on both **complex** and **Mendelian phenotypic variations**
- **Difficult to detect**

⟶ Advances in **long-read sequencing** technologies
- Construct ***de novo* genome** assemblies

Objectives

01 Construct an **ovine pangenome graph** using *de novo* assemblies from 10 French sheep breeds

02 Identify and study **structural variants** (SVs) & **non-reference sequences** (NRSs)

03 Explore the **links between** these **SVs** & **NRSs** and **phenotypes of interest**

Projet cofinancé par le Fonds Européen de Développement Régional

> Strategy to study genetic variations from a sheep pangenome graph

Step 1

Whole genome sequencing

Step 2

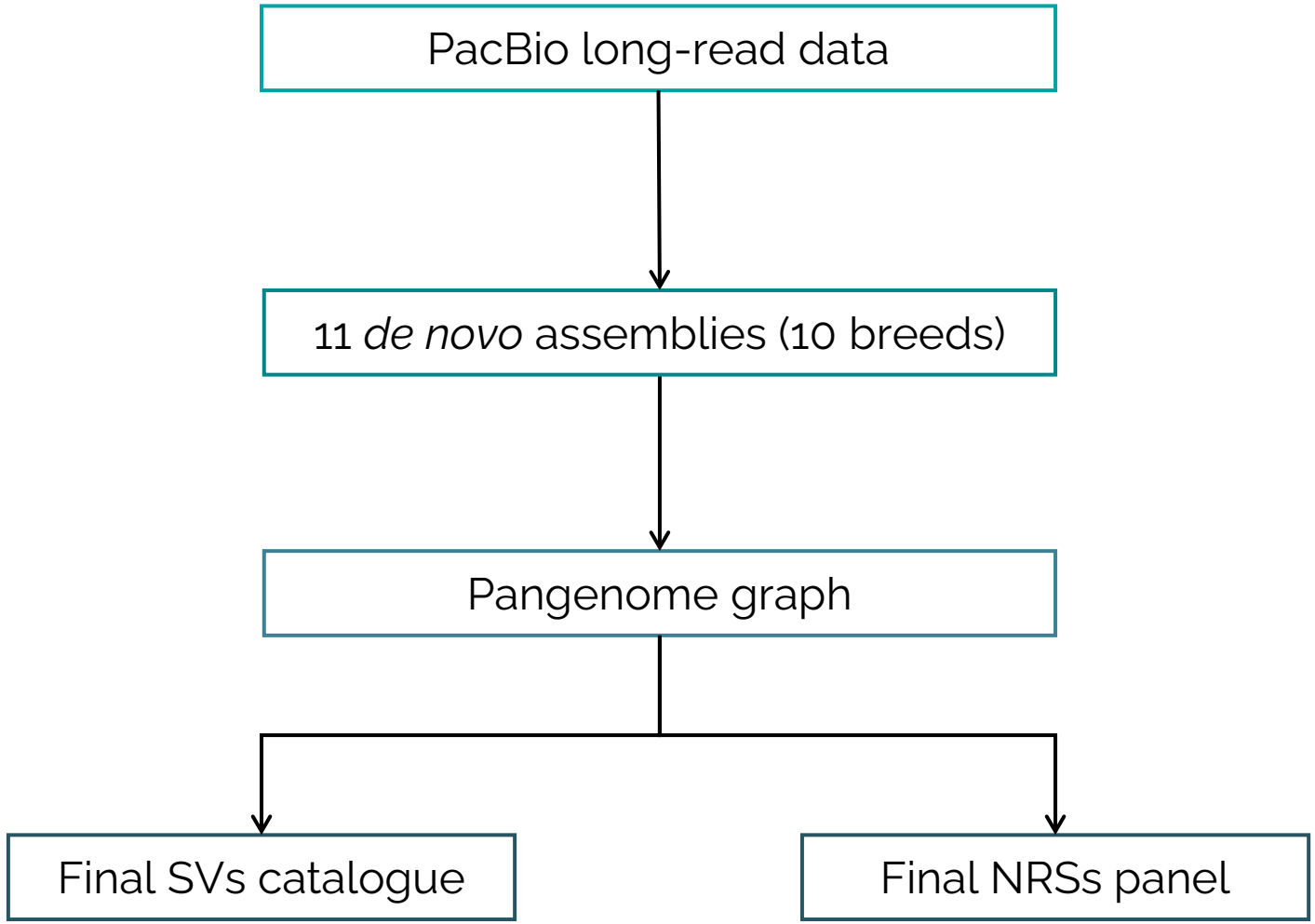
Construction of *de novo* genome assemblies

Step 3

Pangenome graph construction

Step 4

Identification & characterisation of SVs and NRSs



> Step 1: WGS data and phased & *de novo* genome assemblies

Whole genome sequence (WGS) data:

10 animals from 10 breeds
LR (PacBio CLR – 15-20kb & Hi-C and HiFi for phased assemblies)

Assembly construction:

9 *de novo* assemblies (9 breeds)
2 phased assemblies (Lacaune)


Basco Béarnais


Blanche du Massif Central


Corse


Ile de France


Lacaune


Manech Tête Noire


Noir Du Velay


Ouessant

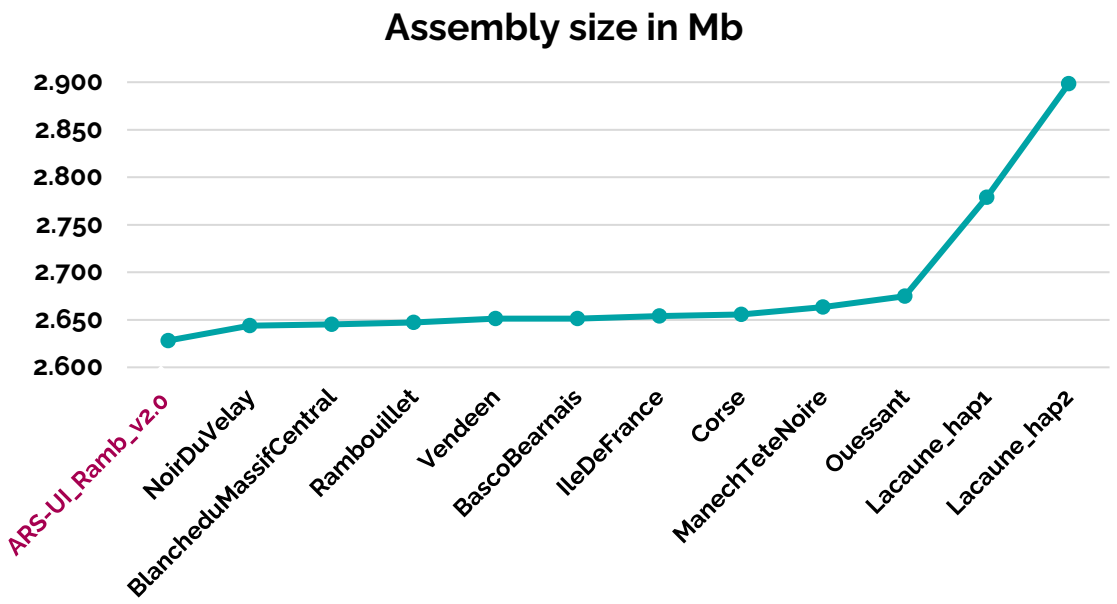

Rambouillet


Vendéen

> Step 2: Assembly metrics

3 main metrics used to assess assembly quality:

Assembly level: scaffolds

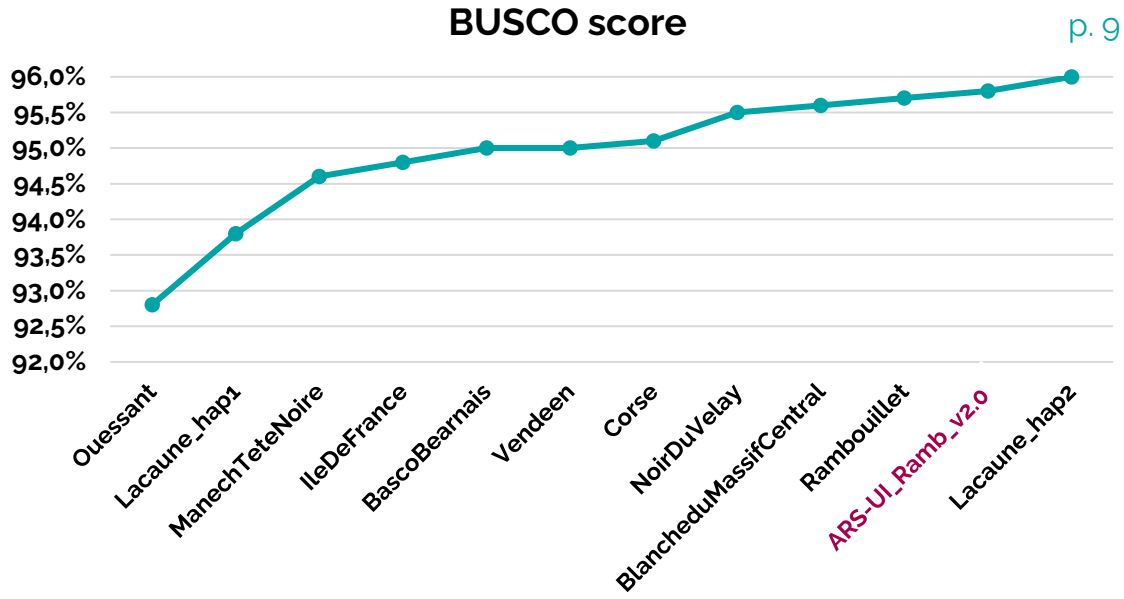


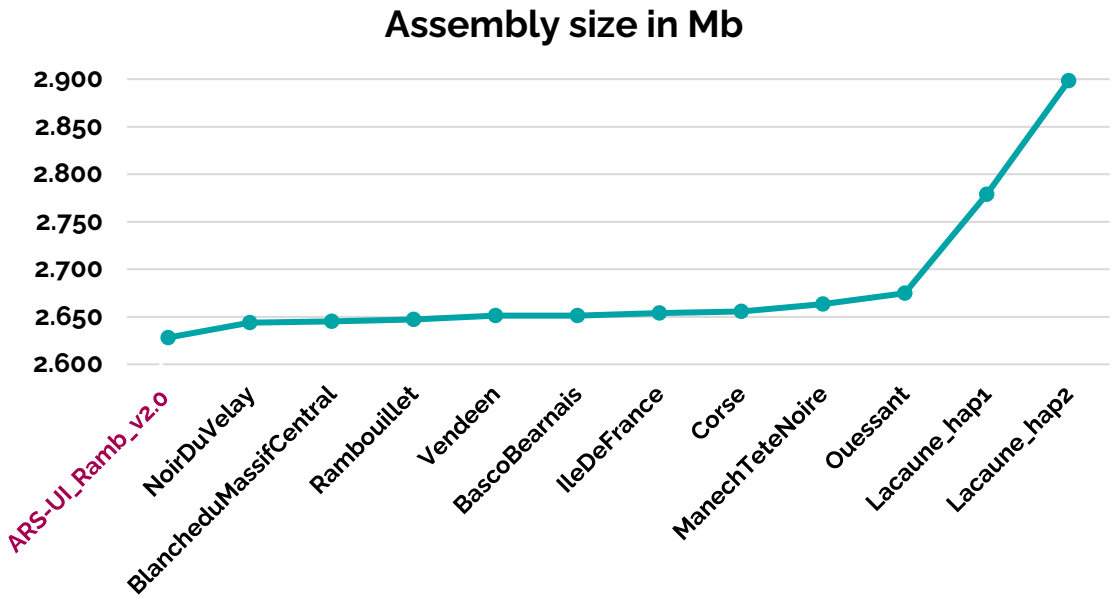
> Step 2: Assembly metrics

3 main metrics used to assess assembly quality:

Assembly level: scaffolds

BUSCO: uses single copies of orthologs to estimate the completeness of genome annotation





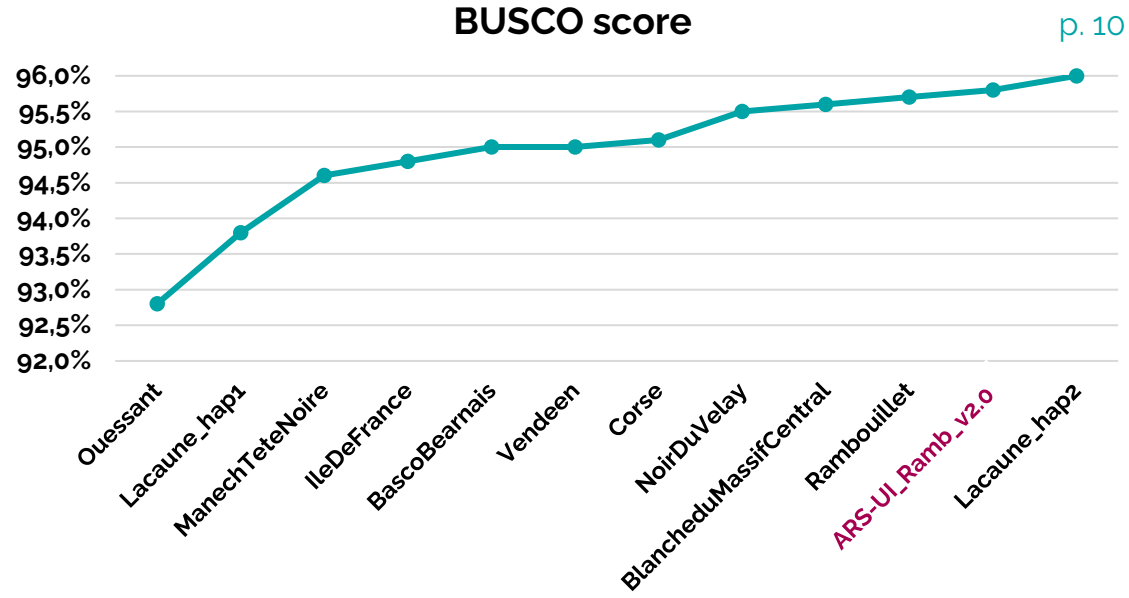
> Step 2: Assembly metrics

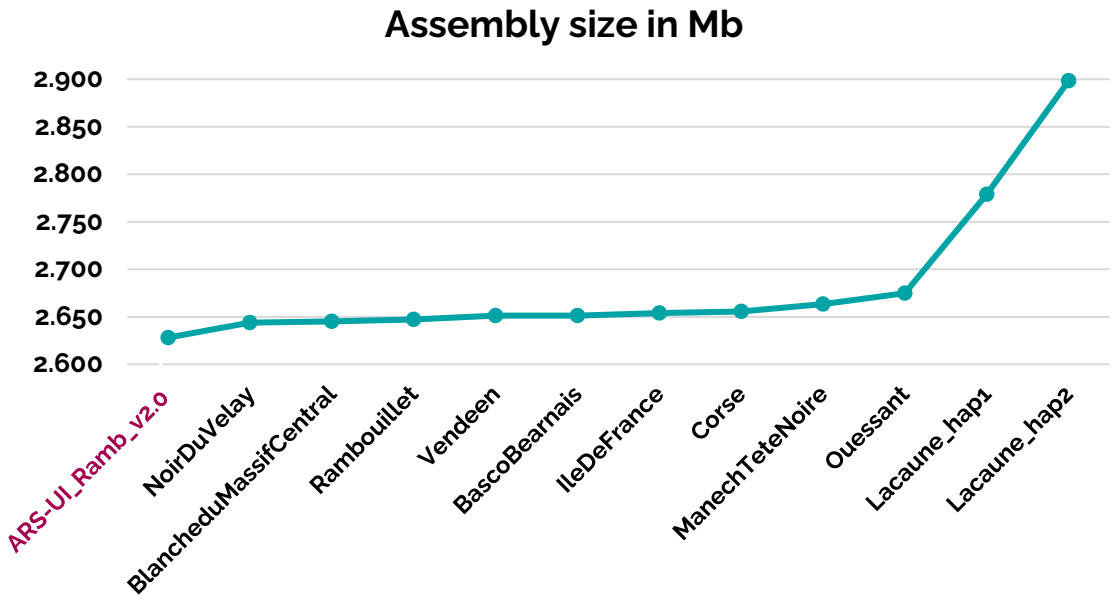
3 main metrics used to assess assembly quality:

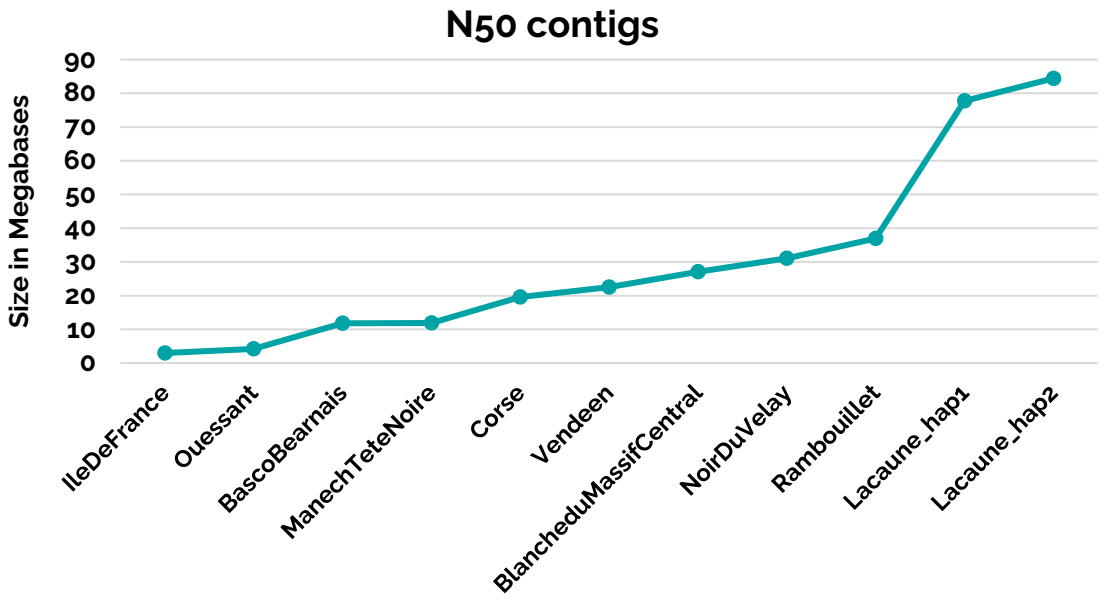
Assembly level: scaffolds

BUSCO: uses single copies of orthologs to estimate the completeness of genome annotation

N50: size of the contig such as 50% of the genome is in contigs of greater size







> Step 3: Construction of pangenome graph

Tools to construct the pangenome graph:

- **Mash version 2.3:** phylogenetic distance
- **Minigraph (MG):** detects only SVs
- **Minigraph-Cactus (MGC):** detects both SNVs* and SVs
  produces 2 graphs
  (SVs / SVs+SNVs)

Genomes used to construct the pangenome graph:

- **Reference:** *GCF_016772045.1_ARS-UI_Ramb_v2.0*
- The autosomes of the 11 *de novo* assemblies

**SNVs = SNPs + InDels*

> Step 3: Construction of pangenome graph

Graph type	Autosomes		
	Minigraph	Minigraph-Cactus	
		Minigraph	Cactus
Graph length (*nt*)	2,577,492,938	2,577,492,938	2,631,736,745
Path length (*nt*)	2,576,522,092	2,577,156,663	-
Core	2,405,918,052	2,404,133,431	-
Flexible	170,604,040	173,023,232	-

Core genome: sequences common to all samples
Flexible genome: sequences present in a subset of the samples

Both tools produce a **structural variation pangenome graph** of the **same size**

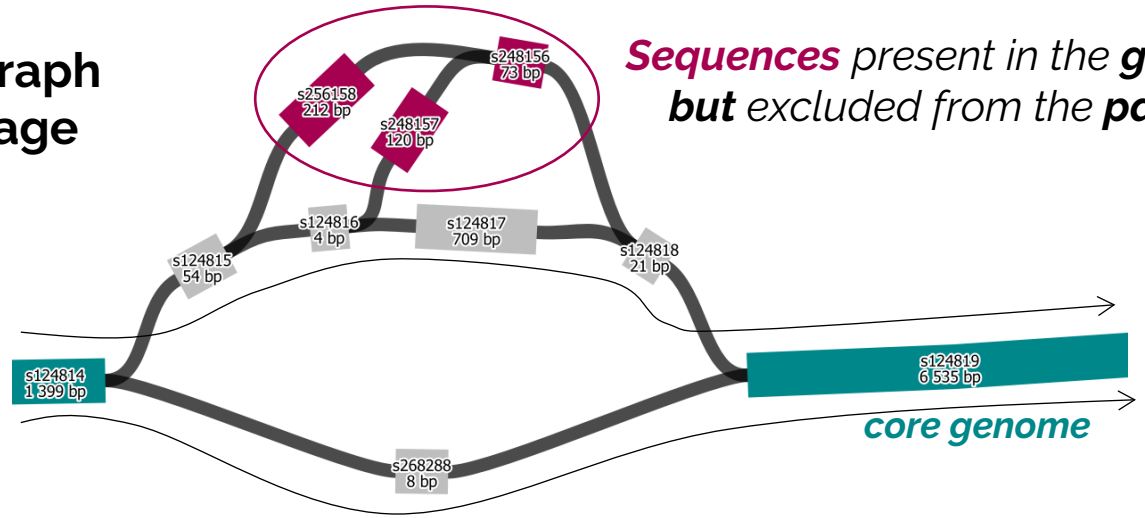
The **cactus graph length** > MG **graph length**

Projet cofinancé par le Fonds Européen de Développement Régional

> Step 3: Construction of pangenome graph

Graph type	Autosomes		
	Minigraph	Minigraph-Cactus	
		Minigraph	Cactus
Graph length (*nt*)	2,577,492,938	2,577,492,938	2,631,736,745
Path length (*nt*)	2,576,522,092	2,577,156,663	-

Example of a Minigraph bubble with Bandage



Sequences** present in the **graph** **but** excluded from the **path

→ **1 407 nodes** were **missing** from the MG **path**

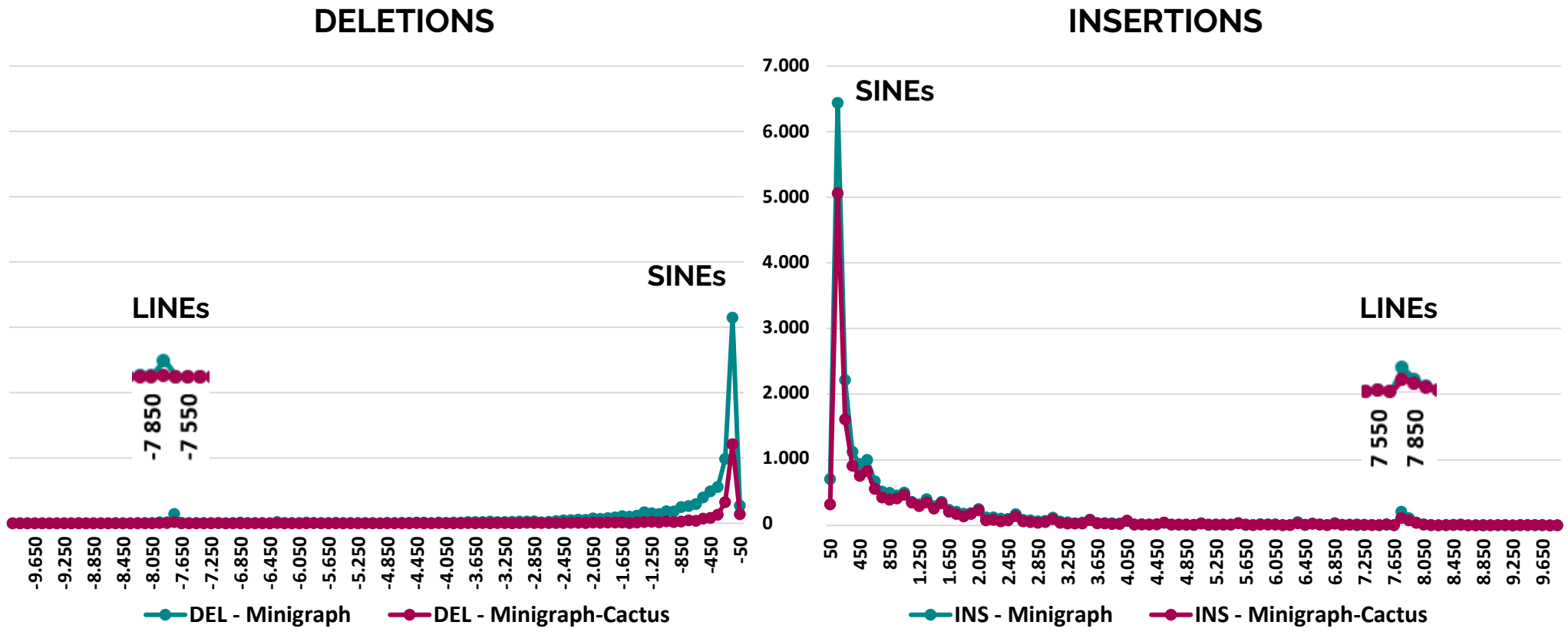
> Steps 4-5: Identification of SVs

	Autosomes	
	Minigraph	*Minigraph-Cactus*
Structural Variants (SVs)	**113,098**	**128,942**
Multi-allelic SVs	11,643	**80,811**
Bi-allelic SVs	**101,455**	**48,131**
Bi-allelic Insertions	**47,045**	**35,974**
Bi-allelic Deletions	**21,245**	**6,707**
Bi-allelic Other variants	33,165	5,450

The vast majority of SVs detected by MGC were **multi-allelic**

Fewer **deletions** were detected by both MG and MGC as compared to **insertions**

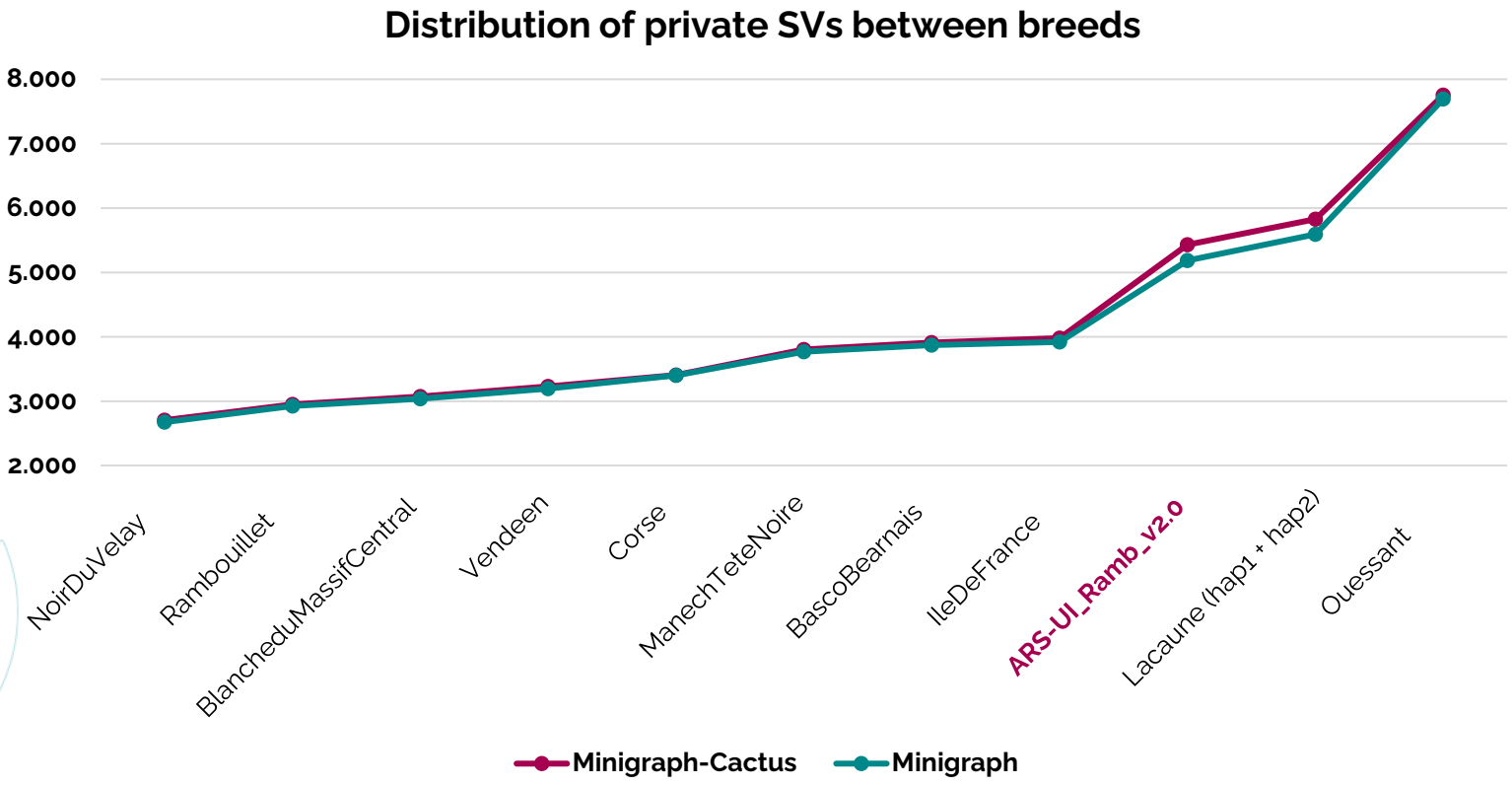
> Steps 4-5: Distribution of the length of deletions and insertions



The **vast majority** of INS and DEL were **small** in **size** (≤*1000 nt*)

2 main peaks corresponding to SVs due to **transposable elements** clearly detected by MG

> Steps 4-5: Identification of SVs

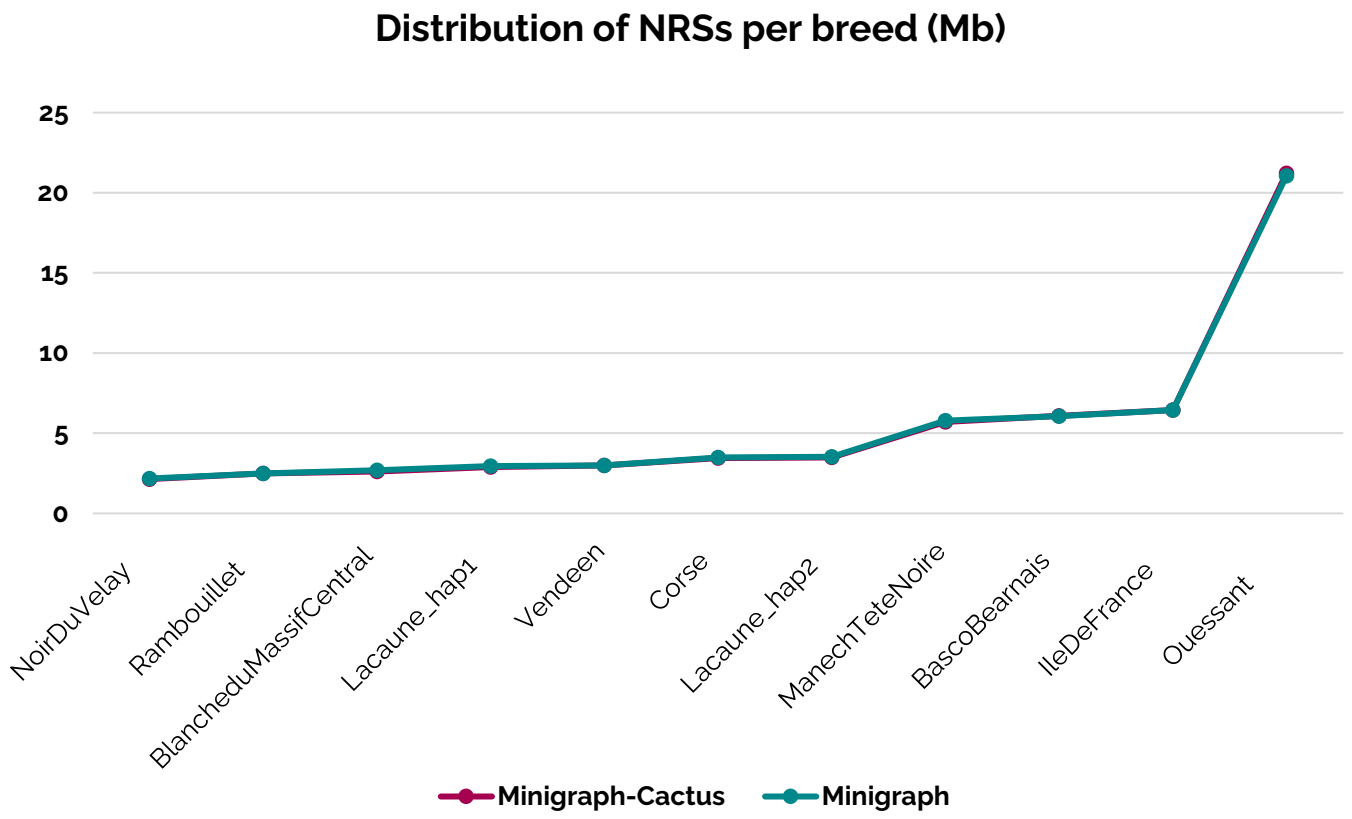


Ouessant breed provides a **large number** of **private SVs**

Lacaune haplotypes have **slightly** the **same number** of **SVs** than the **reference genome**

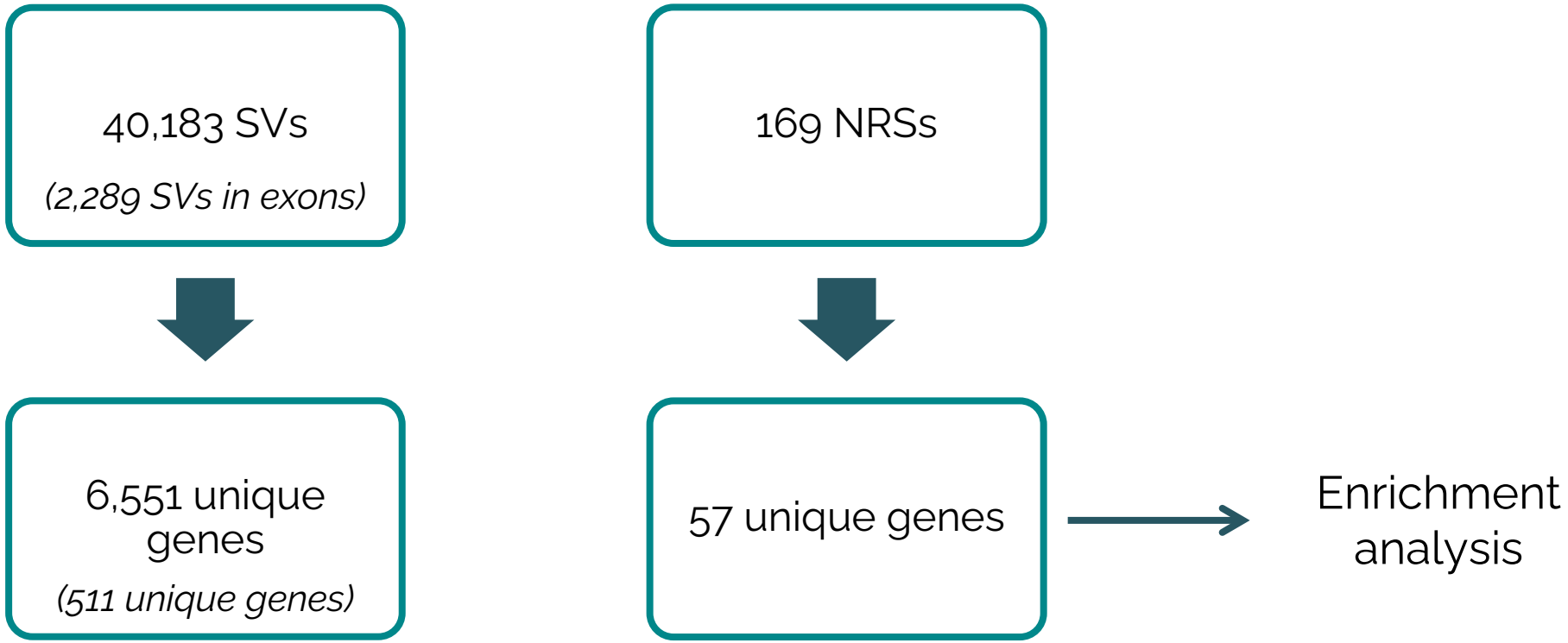
> Steps 4-5: Identification of NRSs from the 2 graphs

	Autosomes	
Pangenome graph	***Minigraph***	***Minigraph-Cactus***
Private NRSs (nt)	**59,700,898**	**59,593,434**



> Annotation of SVs and NRSs

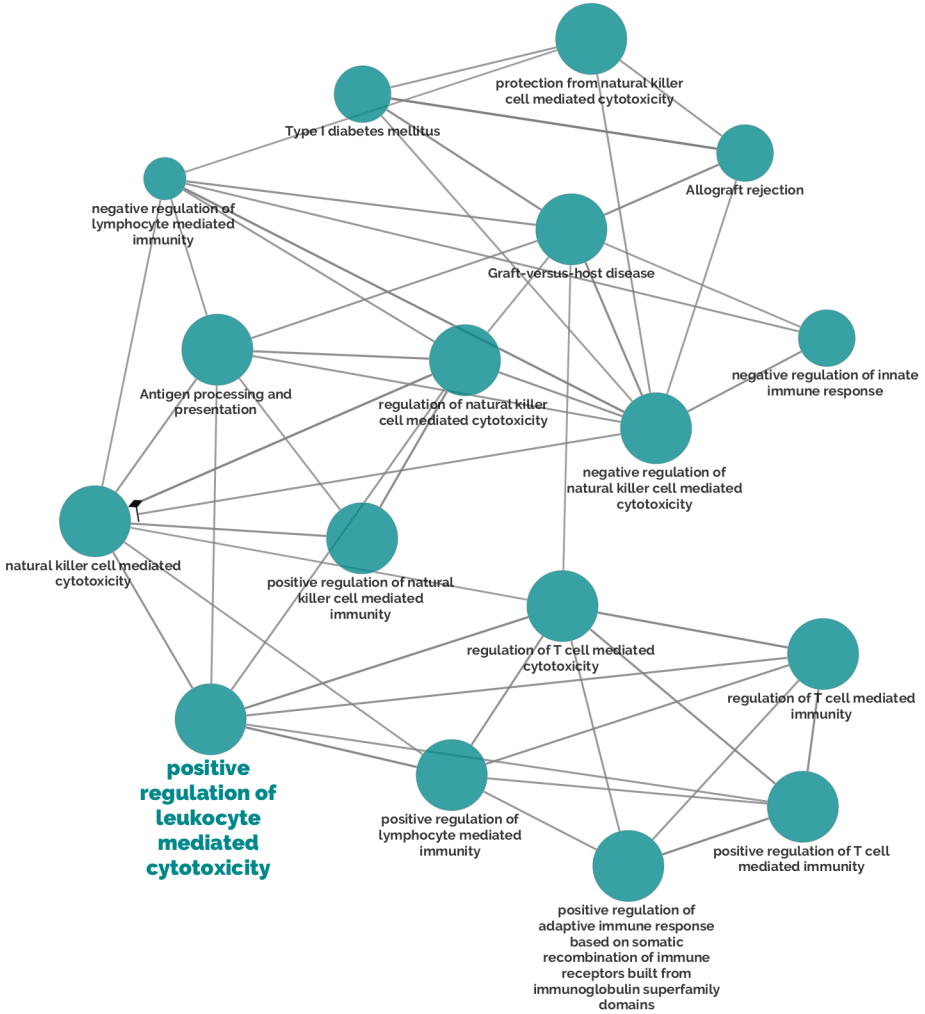
SVs that overlap/contain known sheep genes (exons/introns/regulating regions)

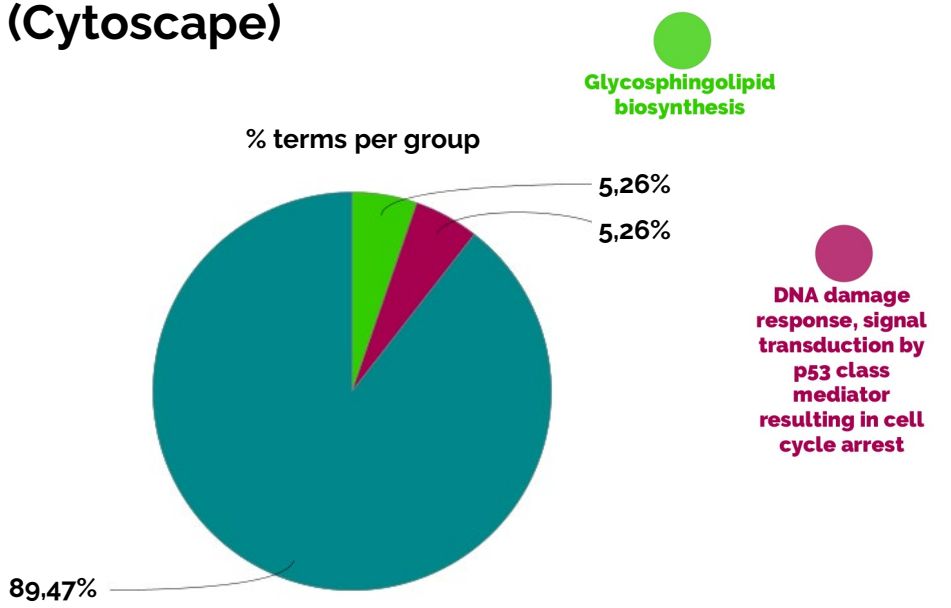


DB: https://ftp.ensembl.org/pub/release-112/gff3/ovis_aries_rambouillet/Ovis_aries_rambouillet.ARS-UI_Ramb_v2.0.112.chr.gff3.gz

> Steps 4-5: Preliminary results for the enrichment analysis

ClueGO (Cytoscape)





- Major Histocompatibility Complex
- Killer Cell Lectin Like Receptor
- Interleukin
- Immunoglobulin

Work is currently in progress to better characterize these results

> Steps 4-5: Repeated sequences in NRSs

		MINIGRAPH	MG - CACTUS
Total length		**104,040,480 bp**	**104,675,562 bp**
GC level		**42.20 %**	**42.21 %**
Bases masked		**47,585,037 bp (45.74 %)**	**47,803,284 bp (45.67 %)**
		% of sequences	**% of sequences**
LINEs		**30.61**	**30.56**
	LINE1	29.87	29.81
	LINE2	0.35	0.35
	L3/CR1	0.04	0.04
	RTE	0.35	0.36

Vast majority of masked sequences → **LINEs subclasses**

~58% of **NRSs** do not contain **repeated sequences**

> Conclusion

- **Assemblies:**
  - CLR LR reads → Construction of 9 *de novo* assemblies of good quality
  - Hi-C and HiFi LR reads of very good quality → Construction of 2 haplotype-resolved *de novo* assemblies

- **Pangenome:**
  - Both MG and MGC tools produce a pangenome of the same size, even if the characteristics differ slightly

- **Structural variants:**
  - MGC identifies more SVs than MG, but with a higher proportion of multi-allelic variants

- **Non-reference sequences:**
  - Approximately 60 Mb of non-reference sequences are identified with both tools and they provide the same distribution per breed

Projet cofinancé par le Fonds Européen de Développement Régional

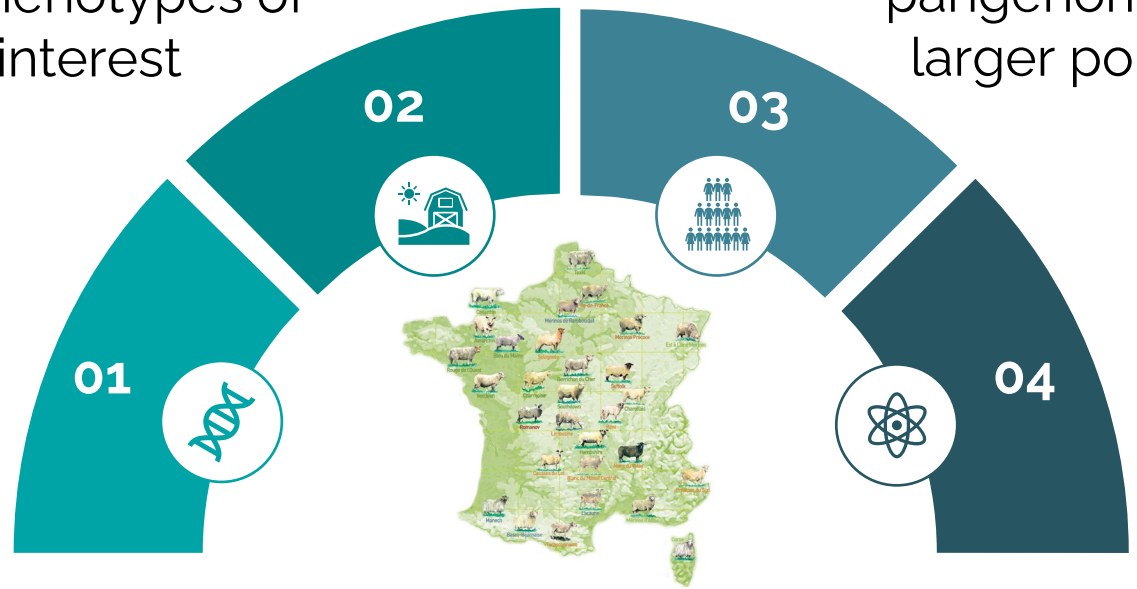
> Future work

01 Characterize SVs & NRSs and their gene content

02 Explore their effect on phenotypes of interest

03 Genotype SVs in the pangenome from a larger population

04 Perform gene prediction and protein homology analyses on NRSs



Same work is currently in progress in goat

Projet cofinancé par le Fonds Européen de Développement Régional

INRAE

Thanks for your attention

This work was financed by APIS-GENE, the Occitanie region and Europe

The SeqOccin group

GABI

GenPhySE
Génétique, Physiologie et Systèmes d'Elevage

Genotoul GeT

Genotoul Bioinfo

All of the private partners who provided the samples