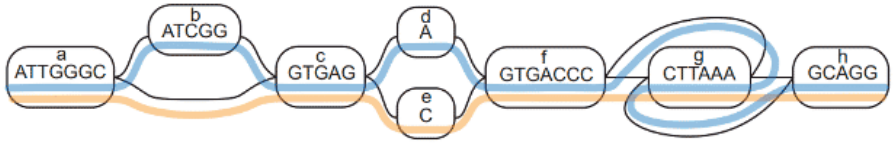
Construction of a cattle pangenome for 14 French dairy and beef breeds provides new insights into their genetic diversity



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

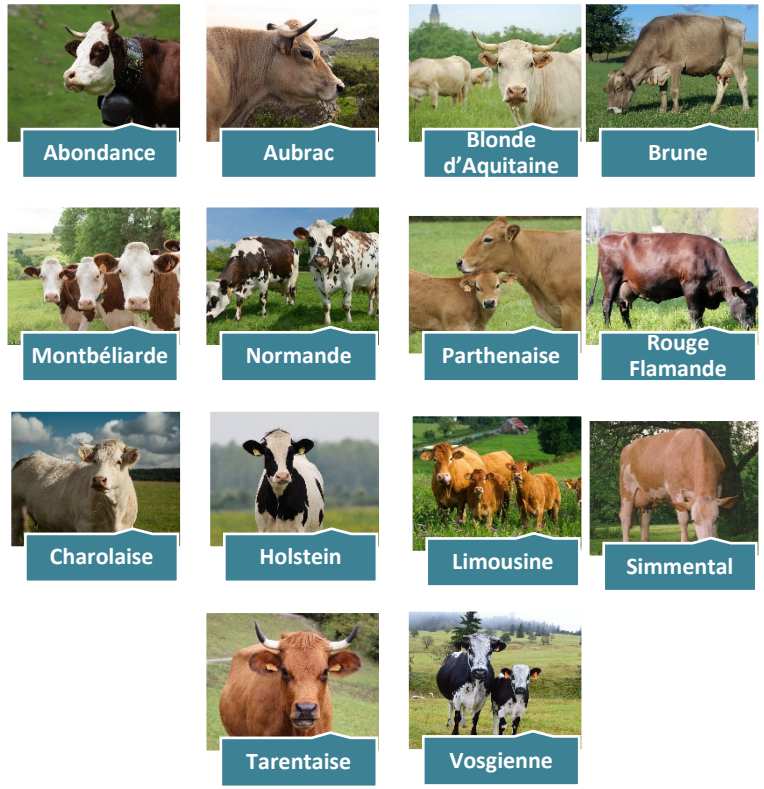
* Université Paris Saclay, INRAE, AgroParisTech, GABI, Domaine de Vilvert, 78350 Jouy-en-Josas, France

Main objectives

- **Objective 1:** Construct a cattle **pangenome graph** using *de novo* genome assemblies from 14 French breeds

- **Objective 2:** Identify and study the functional spectrum of **structural variants** (SVs) and **non-reference sequences** (NRSs)
  - ✓ **SVs** are defined as genomic variations longer than 50 nucleotides
  - ✓ **NRSs** are defined as sequences absent from the current bovine genome assembly

- **Objective 3:** Identify **novel small genetic variants** from the NRS sequences



De novo genome assemblies

De novo genome assemblies

- 169 animals corresponding to 14 breeds were sequenced with the PacBio CLR technology

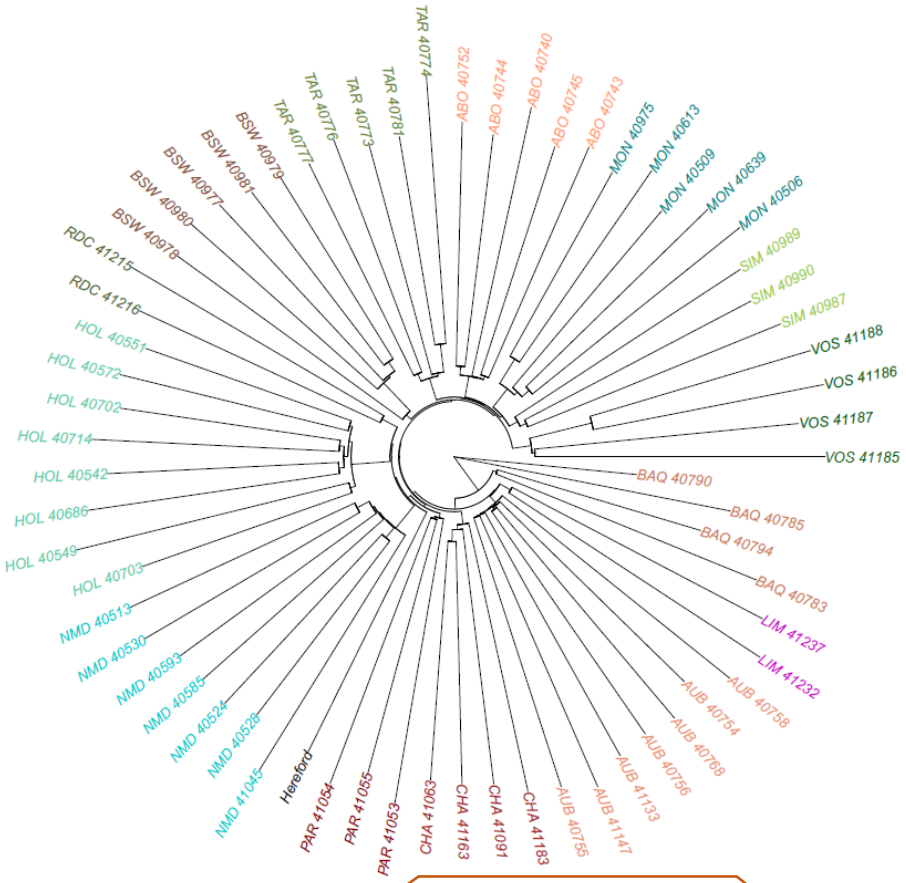
  - 64 chromosome-level genome assemblies polished with both PacBio long-reads and Illumina short-reads
    - ✓ Pangenome construction
    - ✓ Identification of SVs and NRSs

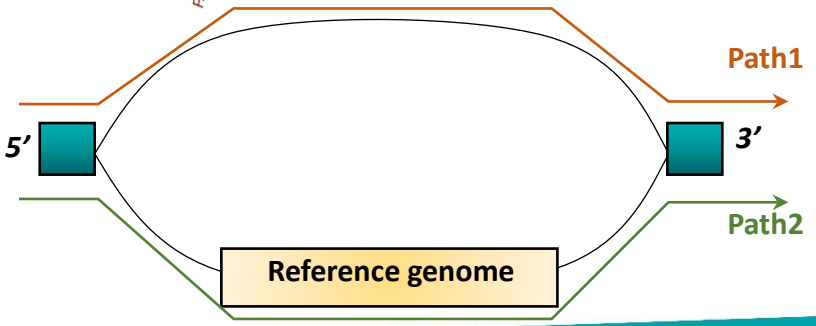
  - 105 contig-level genome assemblies
    - ✓ Genotyping of pangenome-derived SVs

Breed	Number of samples	Pangenome panel	Genotyping panel
Holstein	31	8	23
Monbéliarde	28	5	23
Normande	26	7	19
Charolais	17	4	13
Abondance	11	5	6
Aubrac	11	7	4
Blonde d'Aquitaine	10	4	6
Limousine	10	2	8
Brown Swiss	5	5	0
Simmental	5	3	2
Tarentaise	5	5	0
Vosgienne	4	4	0
Parthenaise	3	3	0
Rouge Flamande	3	2	1
Total	**169**	**64**	**105**

Construction of the pangenome graph

- **Genomes to construct the pangenome graph**
  - ***Reference*** : ARS-UCD1.2 (autosomes + BTAX)
  - ***Assemblies*** : 64 *de novo* genome assemblies
- **Tools to construct the pangenome graph**
  - ***Mash version 2.3*** : phylogenetic distance
  - ***Minigraph*** : produce a graph composed of chains of bubbles with the reference as the backbone
- **Tools to identify and genotype SVs**
  - ***gfatools (bubble parameter) :*** bubbles in the graph represents SVs
  - ***minigraph (-cxasm --call parameters) :*** Find the path/allele of this assembly in each bubble





Construction of a pangenome graph

❖ Characteristics of the pangenome graph

	Minigraph
Pangenome size (nt)	2,933,608,906
core genome (nt)*	2,562,959,040
flexible genome (nt)**	370,649,866

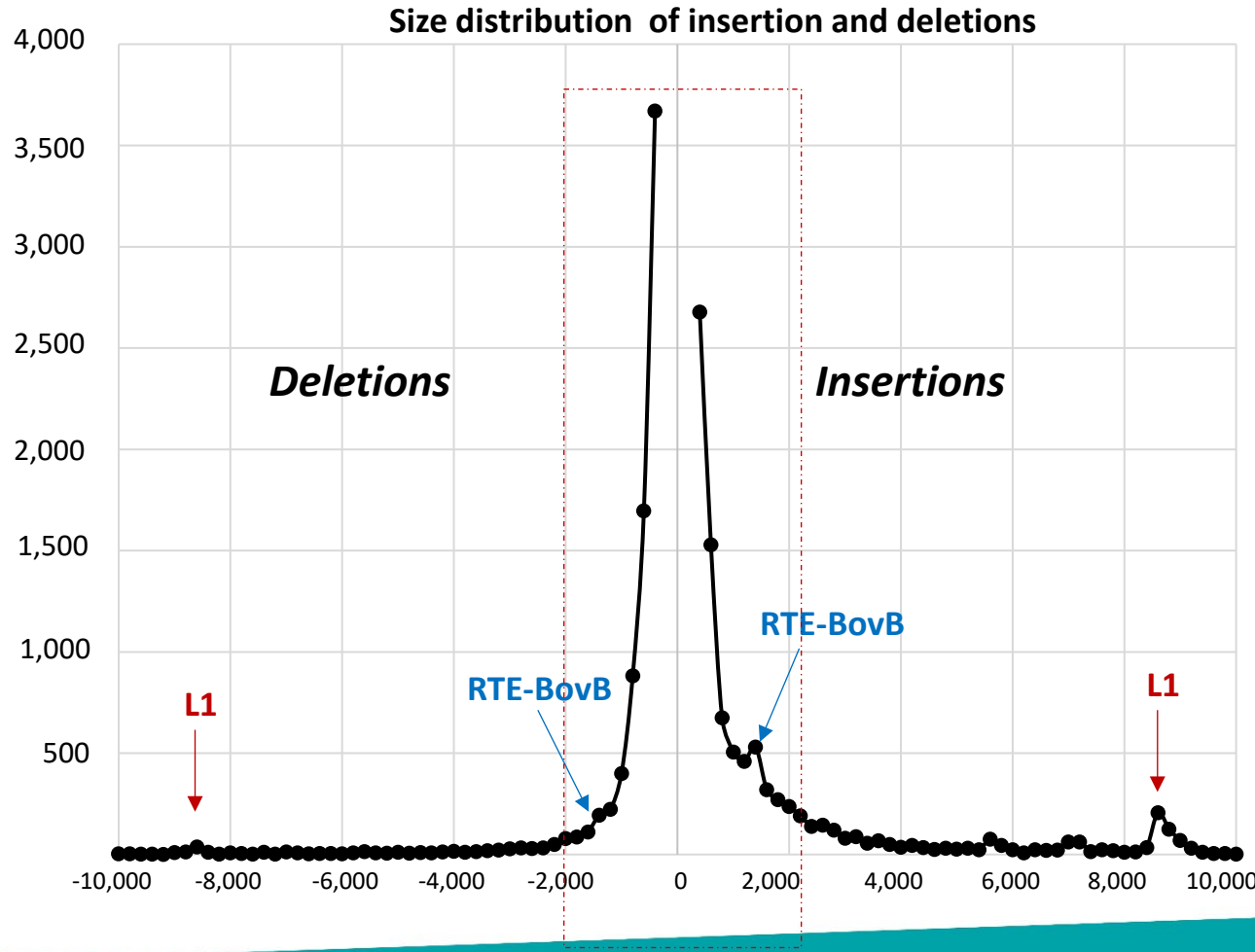
* Core genome : genomic sequences shared by all samples

** Flexible genome : genomic sequences shared by a subset of samples

Identification of structural variations

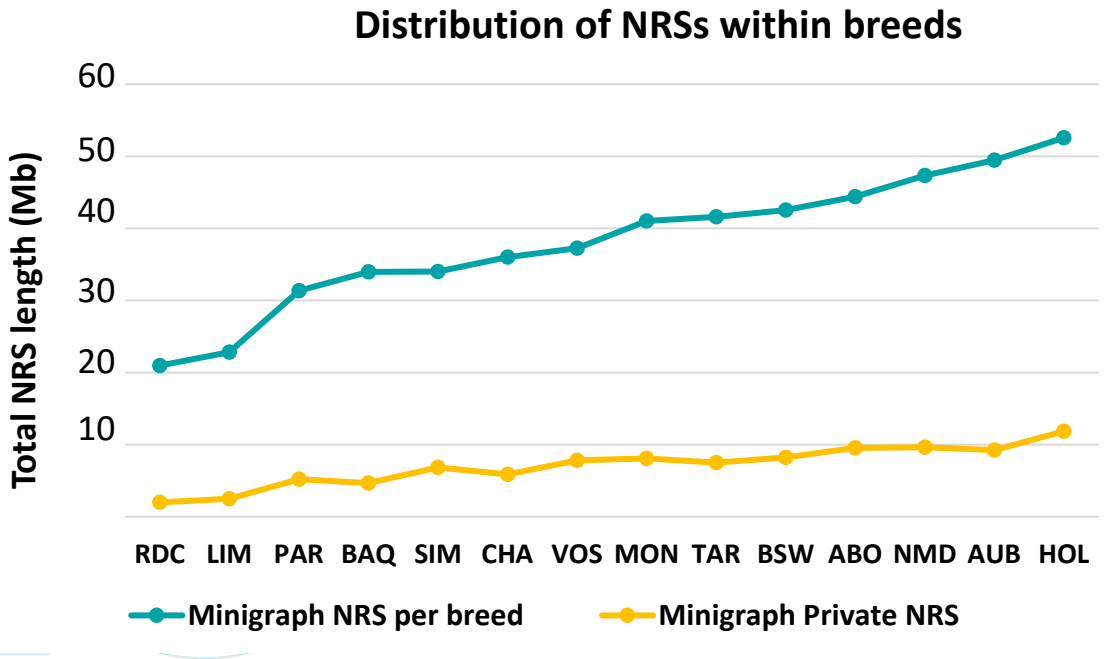
❖ Identification of structural variants

SVs	Total
All	109,267
Bi-allelic SVs	84,614
Bi-allelic insertions	21,831
Bi-allelic deletions	21,342
Other bi-allelic SVs	41,441



Identification of NRSs

- **Length of NRS sequences**: **160 Mb**



GC percent	41.63
Repeated sequences	49 Mb (30.7 %)
	% sequences
LINEs:	18.96
LINE1	17.59
LINE2	0.56
L3/CR1	0.06
RTE	0.75

- The total length of **NRSs** is proportional to the number of animals per breed

- A higher proportion of masked sequences corresponded to the LINEs classes of TEs

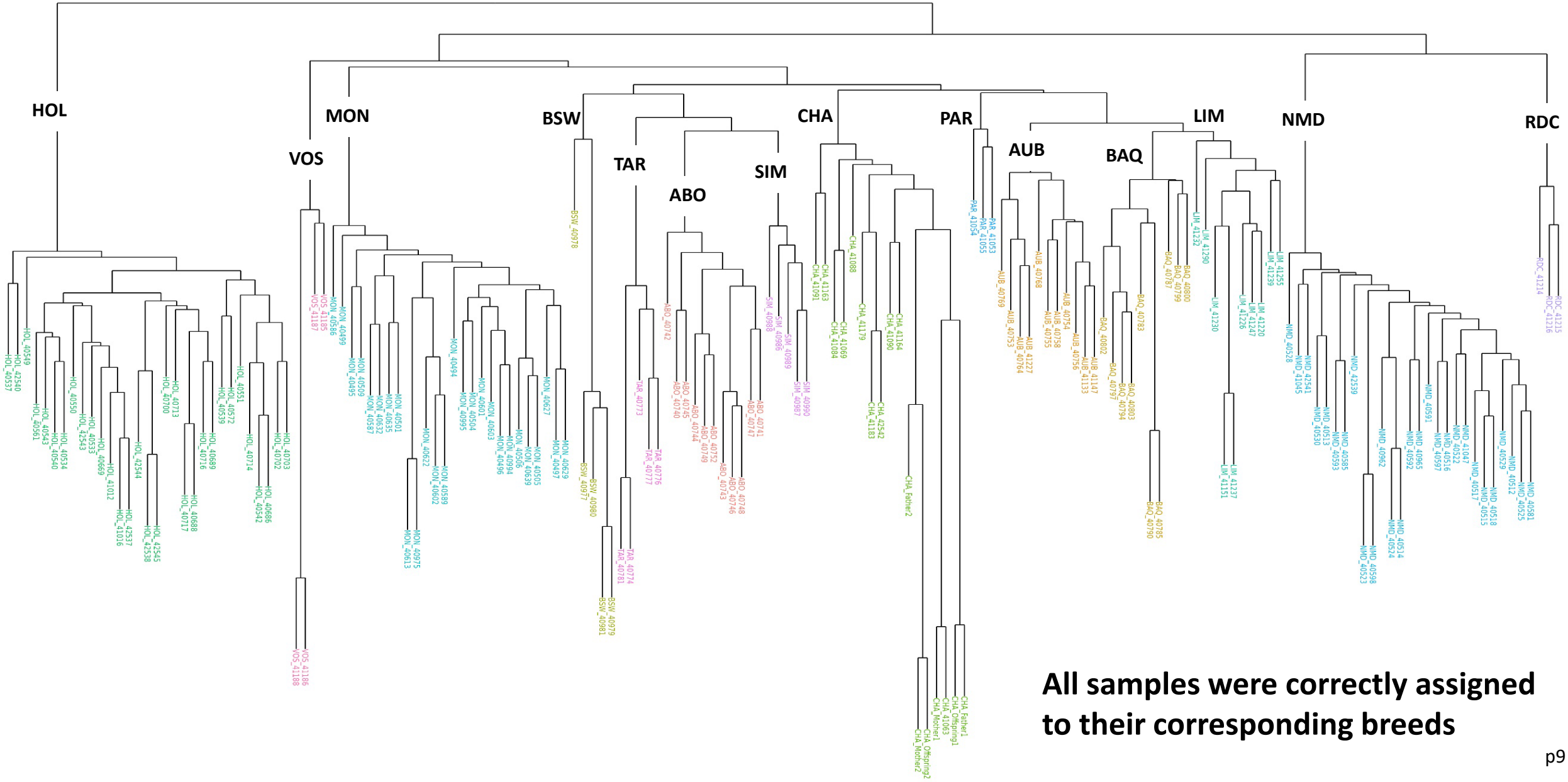
- ~70% of NRSs don't contain repeated sequences and may code for potential functional elements

Population structure

- **Main goal :** Assessment of population structure using presence/absence variation (**PAV**) of SVs or NRSs

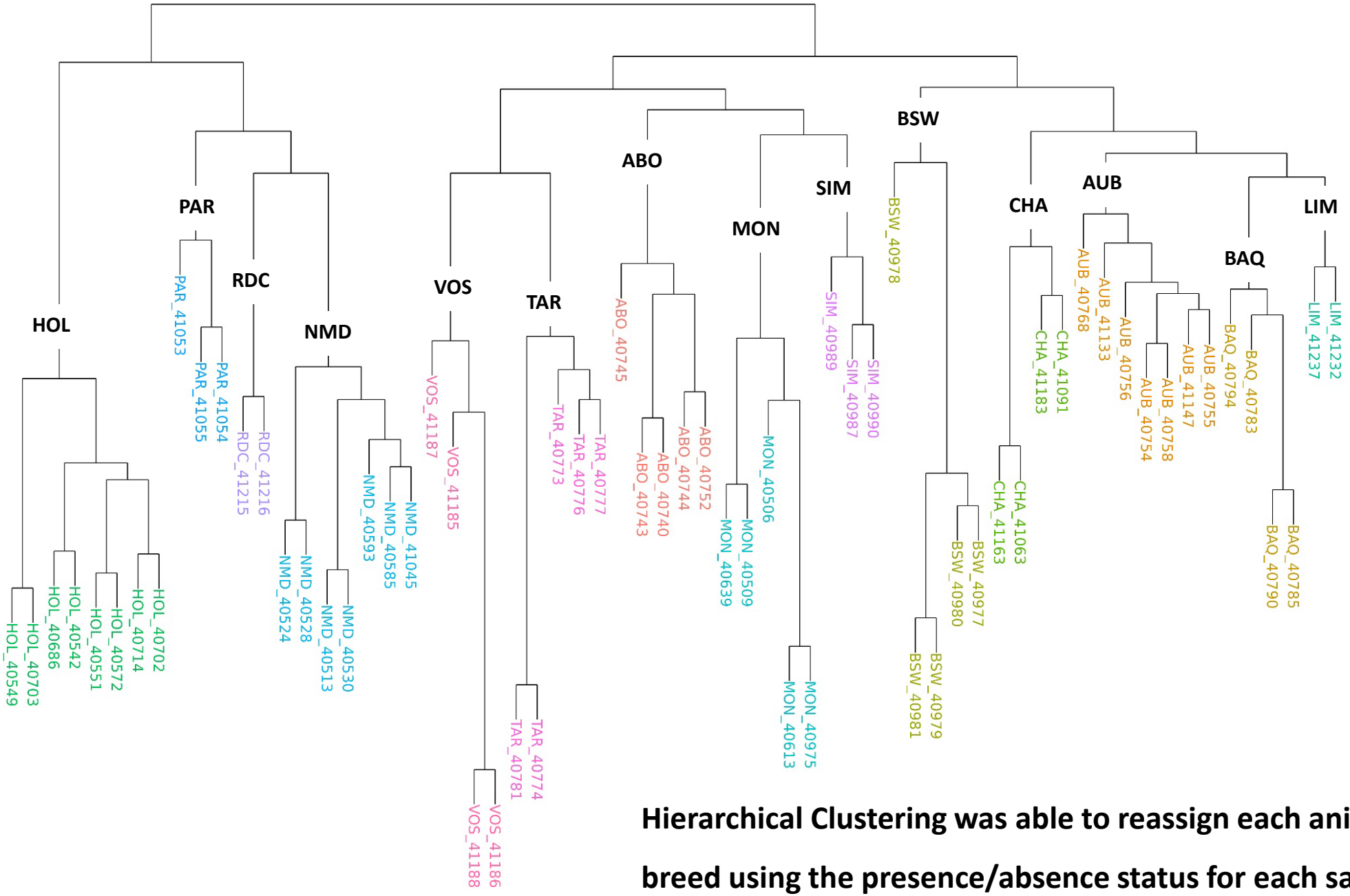
- **Strategy**
  - Production of a PAV-matrix with the presence/absence status of SVs or NRSs in each sample
  - Hierarchical clustering analysis of the PAV-matrix using the R function « **HCLUST** »

Hierarchical clustering of SV presence/absence in the 169 samples



All samples were correctly assigned to their corresponding breeds

Hierarchical clustering of presence/absence NRSs



Hierarchical Clustering was able to reassign each animal to its corresponding breed using the presence/absence status for each sample

Calling novel small genomic variations from NRSs

❖ **Main goal :** Search for new genetics variants from NRS sequences

❖ **Strategy**

➢ Get unmapped short reads from 252 BAM files

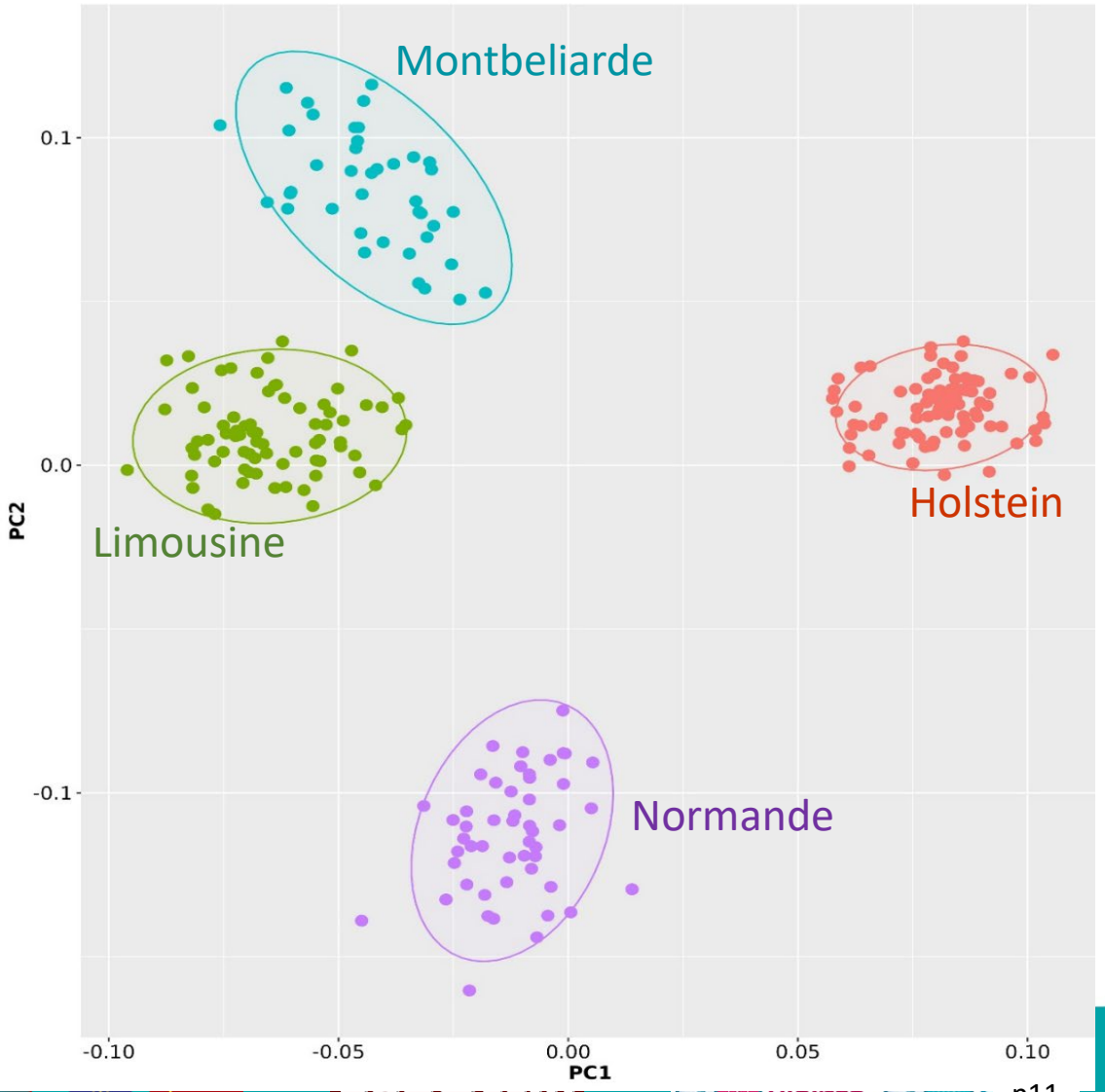
Breed	Total animals
Holstein	88
Montbéliarde	40
Normande	50
Limousine	74

➢ Align the unmapped reads with BWA MEM against the NRSs
➢ Call SNPs and InDels using GATK
➢ Investigate population structure using PLINK

❖ **Results**

Variants	Total
Variants	25,095
Bi-allelic	24,004
Bi-allelic SNPs	21,246
Bi-allelic small insertions	1,653
Bi-allelic small deletions	1,105

❖ **Population structure**



Conclusions

- Construction of a pangenome graph using 64 high quality *de novo* genome assemblies
- Identification and genotyping 109,000 SVs
- Identification of 159 Mb of NRSs
- Hierarchical clustering using SVs and NRS correctly assigned samples to their corresponding breeds
- Identification of 25,000 novel small genomic variants from NRSs
- PCA analysis using these variants grouped all samples in distinct clusters according to their breed of origin

Same work is conducted in sheep and goat (see next presentation of Valentin Sorin)

Future work

- Evaluate the biological impact of both SVs and the novel small genomic variants
- Search for potential new functional elements using the NRSs
- Explore the link between SVs and the novel small genomic variants on certain phenotypes of interest
- Submission of a scientific article in the near future

Acknowledgments

- This work is part of the PhD project of Valentin Sorin.
  - The PhD is co-supervised by Mekki Boussaha, Gwenola Tosser-Klopp, Marie-Pierre Sanchez and Laurence Drouilhet

- This work is part of the SeqOccIn project, financed by APIS-GENE, the Occitanie region and Europe

- The SeqOccIn group

- All the private partners who provided the samples

Thank you for your attention