VarGoats international initiative, a 1000 goat genomes project

Licia Colli, Paola Crepaldi, Paolo Ajmone-Marsan, Alessandra Stella, Gwenola Tosser-Klopp, . Consortium Vargoats

VarGoats international initiative, a 1000 goat genomes project

Licia Colli[1,2], *Paola Crepaldi*[3], *Paolo Ajmone-Marsan*[1,2], *Alessandra Stella*[4], *Gwenola Tosser-Klopp*[5], *The VarGoats Consortium*[6]

[1]*Dipartimento di Scienze Animali, della Nutrizione e degli Alimenti, Università Cattolica del S. Cuore, Piacenza, Italy;* [2]*BioDNA Centro di Ricerca sulla Biodiversità e sul DNA Antico, Università Cattolica del S. Cuore, Piacenza, Italy;* [3]*Dipartimento di Medicina Veterinaria, University of Milan, Milan, Italy;* [4]*Istituto di Biologia e Biotecnologia Agraria, National Research Council, Milan, Italy;* [5]*GenPhySE, Université de Toulouse, INRA, ENVT, Castanet Tolosan, France;* [6]*http://www.goatgenome.org/vargoats.html*

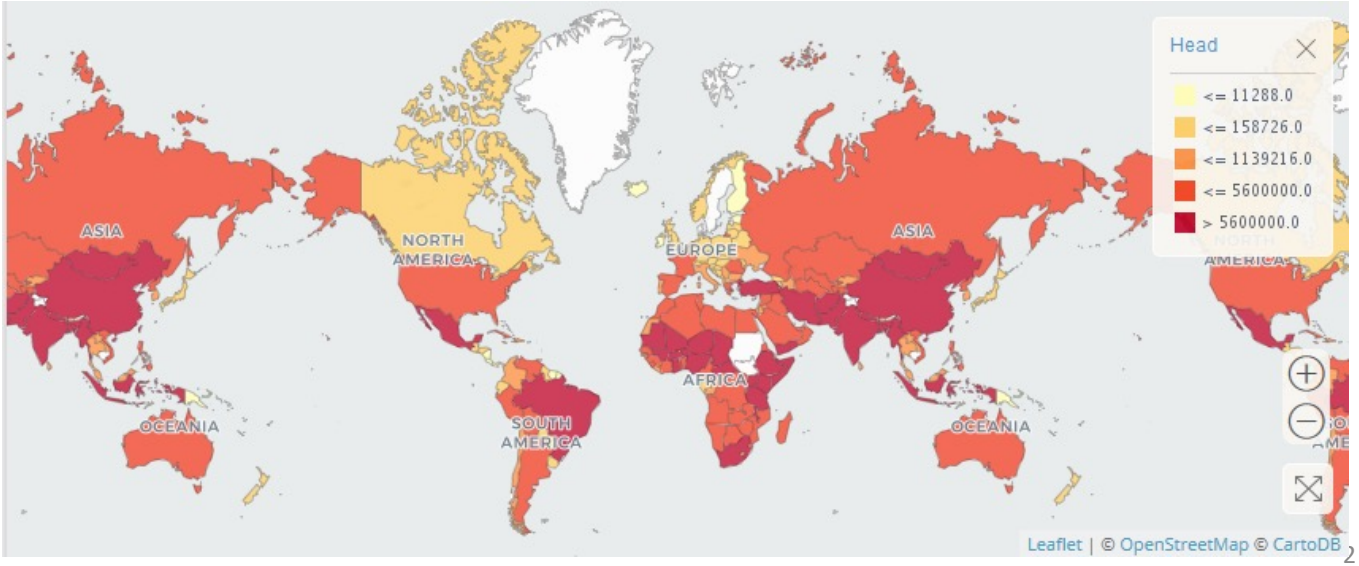
licia.colli@unicatt.it.

23° Congress of the Animal Science and Production Association ASPA – June, 13th 2019

Domestic goats:

- Goat was domesticated from bezoar (*Capra aegagrus*) ca. 10,000-15,000 years ago.
- Goats are adapted to various (sometimes harsh) environments.
- Breed concept is ca. 200 years old.
- Milk, meat and fiber breeds.
- Nowadays a few breeds represent most of the animals, particularly in developed countries. Ex: in France Alpine & Saanen ca. 80% animals.

1 billion goats in the world, 18% endangered.



VarGoats project:

France Genomique Project launched in 2016

Gwenola Tosser-Klopp, Université de Toulouse, INRA
VarGoats coordinator

Goat Genome

VarGoats

Identification of Variations in Goat genomes related to domestication and adaptation

VarGoats is the first step of a 1000 goat genomes project and is lead by Gwenola Tosser-Klopp (INRA, France).

It is supported by FRANCE GENOMIQUE through a call for Large Scale DNA Sequencing projects. This means the scientific Consortium provides DNA and gets back genome sequences, generated at Genoscope (Evry, France). TGCC (Très Grand Centre de calcul du CEA) is the bioinformatic infrastructure where sequences will be stored and available for the Consortium. The data will be made available to VarGoats participants and data analysis will be performed in working groups already created in ADAPTMAP program or if needed in new working groups. Data will be used only for academic purposes, specifically for performing population genetics studies, for the investigation of diversity, domestication and adaptation traits, the discovery of variants (SNPs, CNVs, structural variants, causal mutations), the detection of selective sweeps, with the final goal to develop breeding solutions. Hybridization between species will also be studied, thanks to the availability of sequences from various capra species. At the end of the project, data will be released in a public database for research purpose only, even in case of no publication.

http://www.goatgenome.org/vargoats.html

VARGOATS aims:

- Use of **whole genome sequences:**
  - to investigate **domestication**, human-mediated **selection** and **adaptation** at the genomic level.
  - to study **genetic diversity**.
  - to **detect variants** (SNPs, CNVs, structural variants, causal mutations) and identify **Loss of Function** mutations.
  - To detect **selective sweeps**.

Final goal: to develop **breeding solutions.**

VARGOATS working context:

- **International Goat Consortium** (<2011)
- **Core working group** (INRA/USDA-AGIN/LECA/PTP later joined by Roslin)
- **Bioinformatics skills**
- **Genomic tools**
  - **high quality assembly** (ARS1 = golden goat genome)
    - **MD 50K chip**

- **Data availability from previous studies:**
  - **ADAPTmap project** = **>4,000** 50K SNP chip **genotypes**.
  - **NextGen project** = ca. **190 WG sequences** from Iran and Morocco.
  - **ClimGen project** = 50K SNP chip **genoptypes** and **methylation studies**.
  - Publicly available data.

VarGoats sampling strategy:

Relevant gene pools

Detailed description of sampling

Gene pools | Inbred breeds | Other Capra species

Relevant gene pools

Determined by Working Groups from ADAPTmap project (lead by Licia Colli and Paola Crepaldi)

1. Pakistan
2. Northern Africa
3. Western Africa
4. Eastern Africa
5. Southeastern Africa
6. Madagascar
7. Boer
8. Spain & France
9. Saanen
10. Alpine
11. Mediterranean
12. Northern Europe
13. America
14. Australia
15. Wild goats, Turkey, Iran

Gene pools | Inbred breeds | Other Capra species

Inbred breeds

To explore deleterious mutations

1. Icelandic
2. Palmera

Detailed description of sampling

Gene pools | Inbred breeds | Other Capra species

Other Capra species

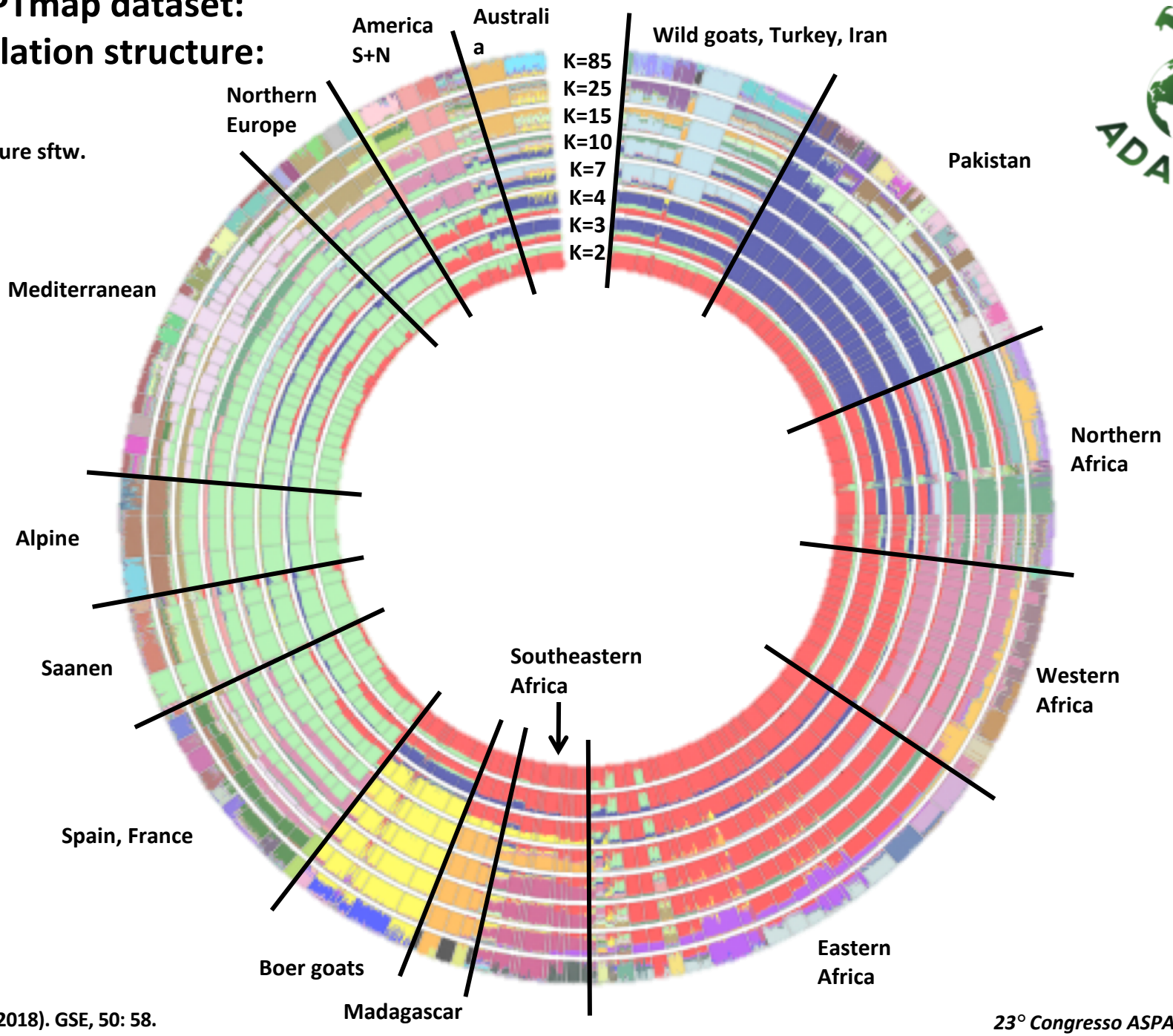
1. Capra falconeri
2. Capra ibex
3. Capra falconeri

Global diversity, post-domestication history, and selection signatures.

Aims:

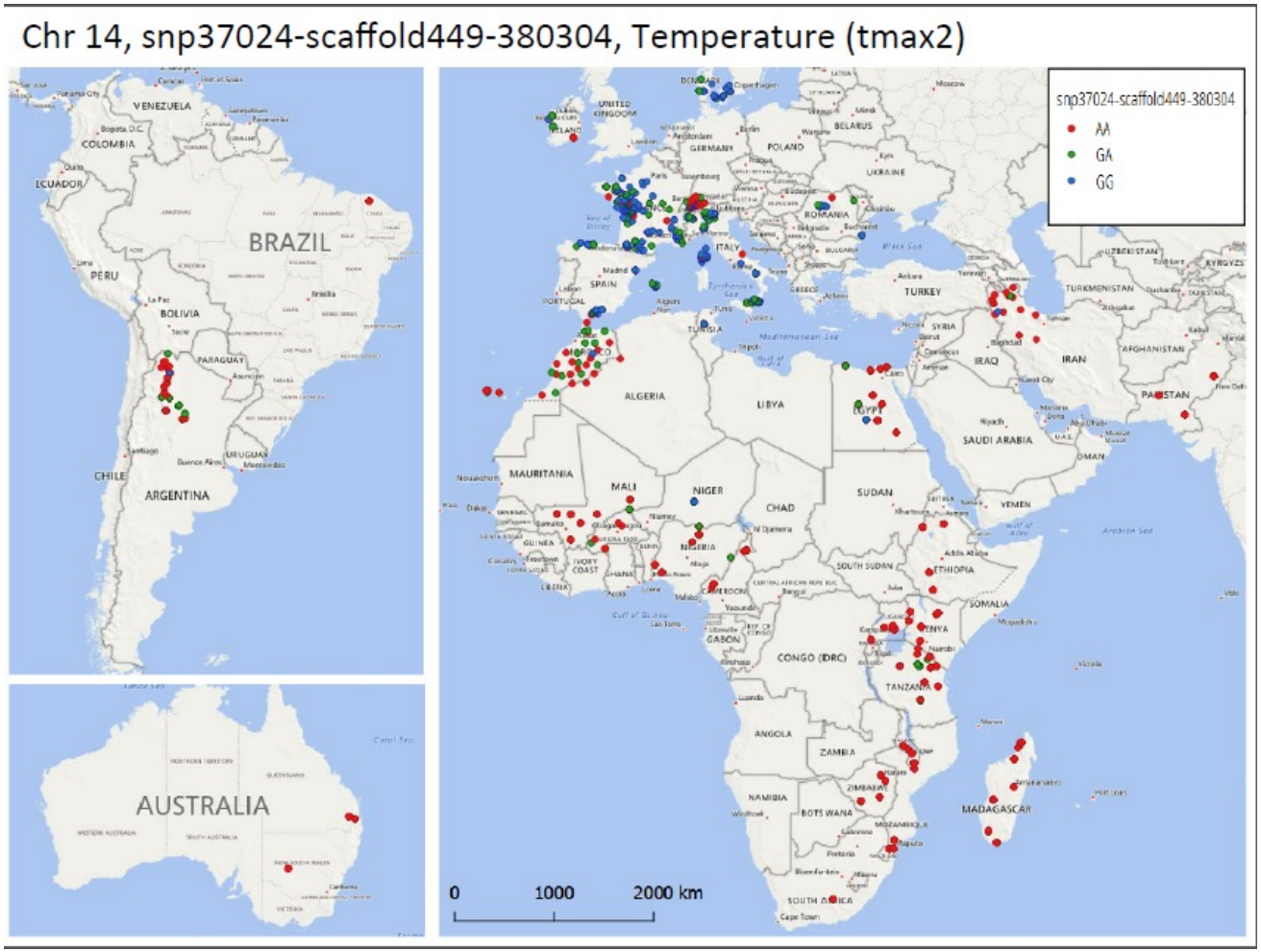
- **to evaluate diversity, gene flow, population structure and migrations.**
- **to identify signatures of natural or human-mediated selection to the environmental or productive conditions.**

ADAPTmap dataset: population structure:



Colli et al. (2018). GSE, 50: 58.

ADAPTmap dataset: Selection signatures



Annotated gene: LYPLA1 (feed intake).

VarGoats sampling strategy:

Detailed description of sampling

Gene pools | Inbred breeds | Other Capra species

Relevant gene pools

Determined by Working Groups from ADAPTmap project (lead by Licia Colli and Paola Crepaldi)

1. Pakistan
2. Northern Africa
3. Western Africa
4. Eastern Africa
5. Southeastern Africa
6. Madagascar
7. Boer
8. Spain & France
9. Saanen
10. Alpine
11. Mediterranean
12. Northern Europe
13. America
14. Australia
15. Wild goats, Turkey, Iran

Gene pools | Inbred breeds | Other Capra species

Inbred breeds

To explore deleterious mutations

1. Icelandic
2. Palmera

Inbred breeds

Aim: to study ROHs and deleterious mutations.

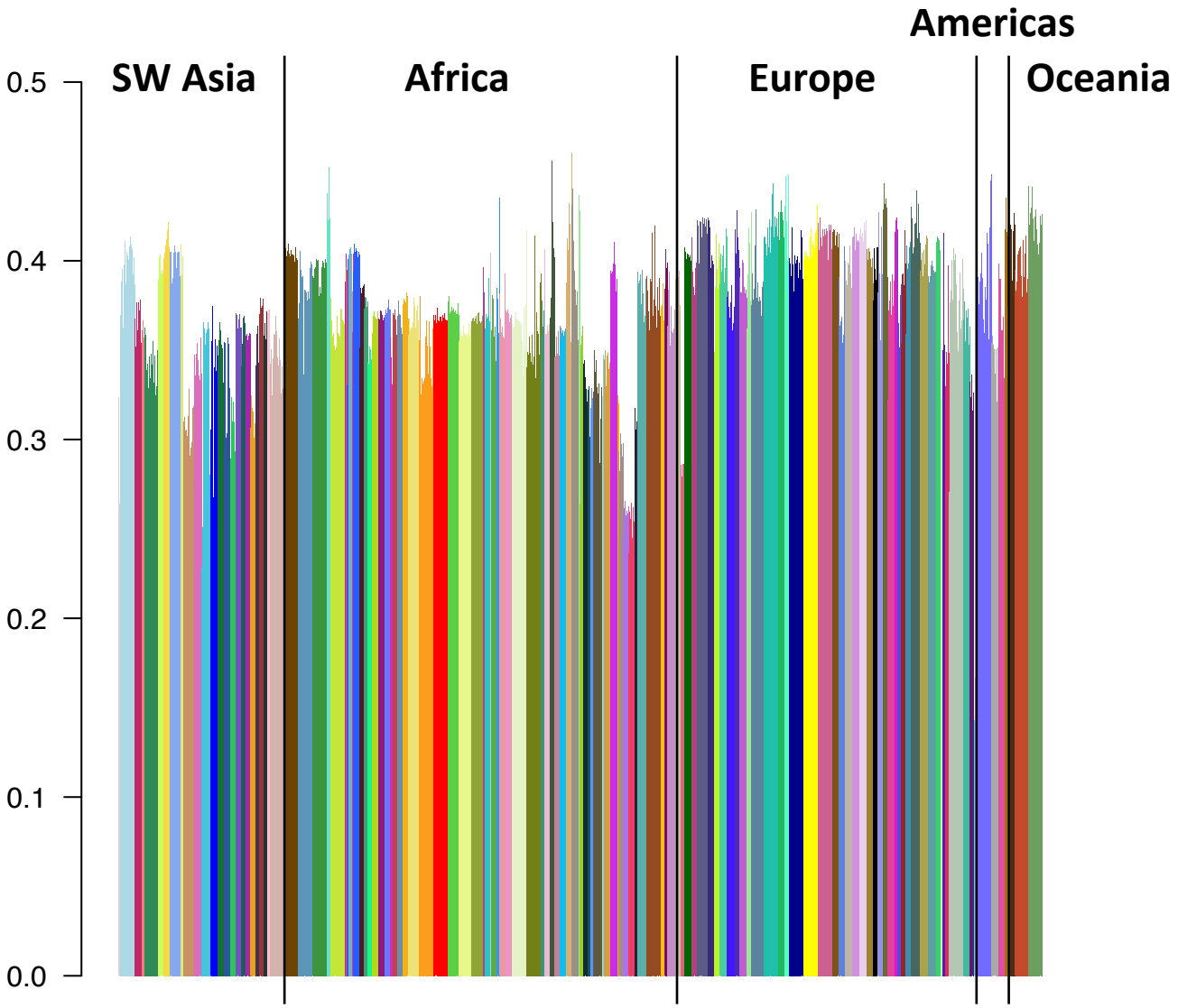
Detailed description of sampling

Gene pools | Inbred breeds | Other Capra species

Other Capra species

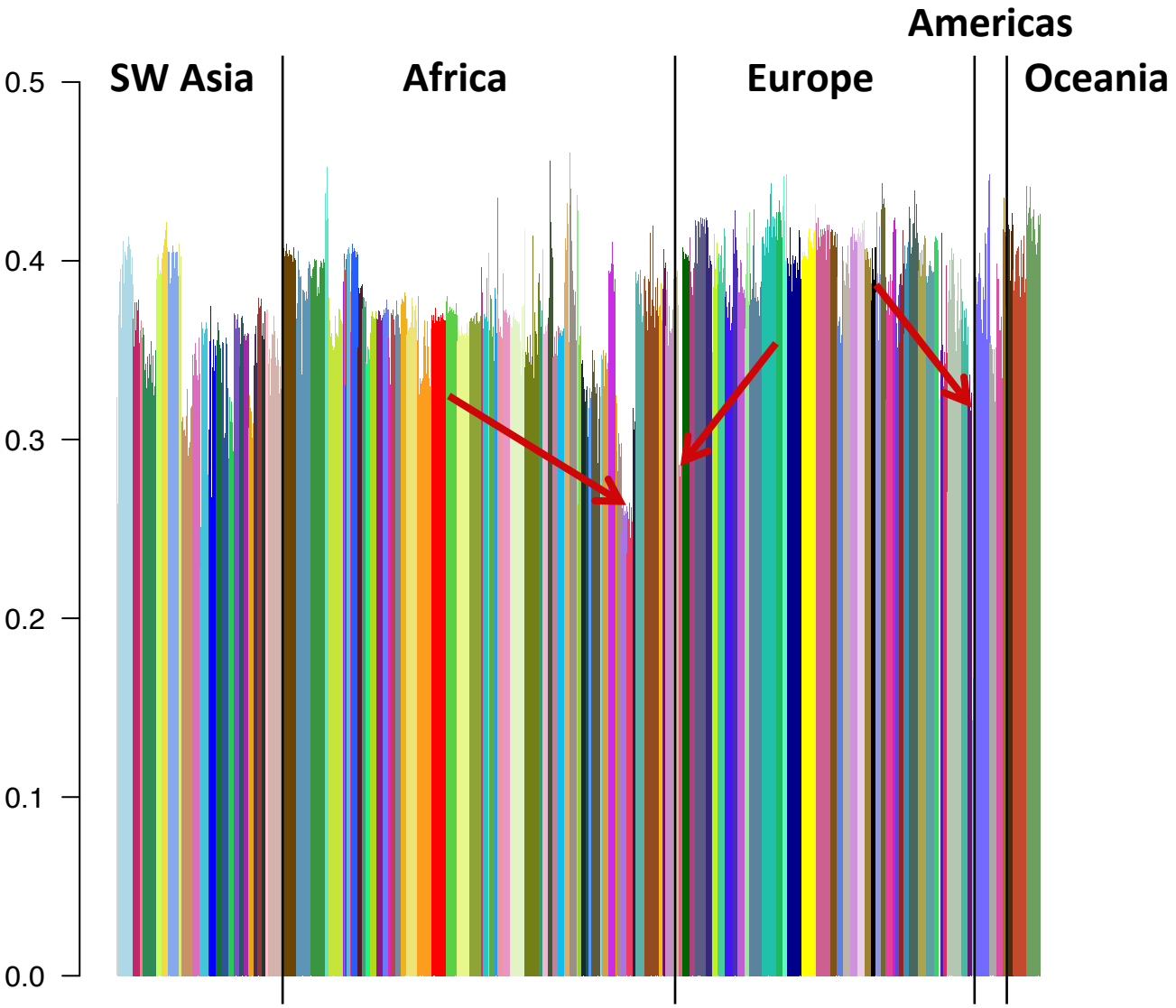
1. Capra falconeri
2. Capra ibex
3. Capra falconeri

ADAPTmap dataset: Observed Heterozygosity



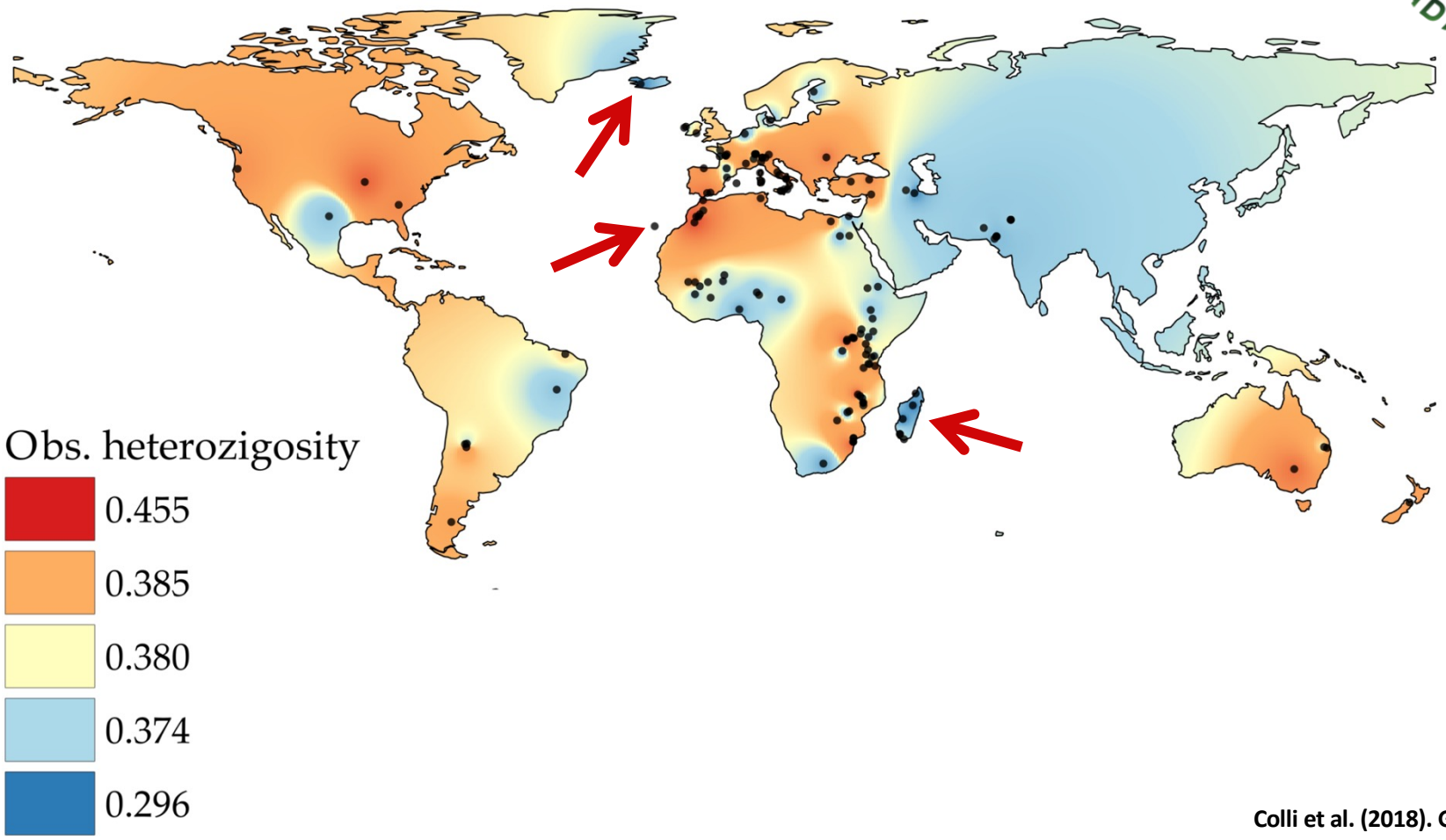
Analysis performed
by M. Milanesi & E. Vajana.

ADAPTmap dataset: Observed Heterozygosity



Analysis performed
by M. Milanesi & E. Vajana.

ADAPTmap dataset: Observed Heterozygosity map



Colli et al. (2018). GSE, 50: 58.

VarGoats sampling strategy:

Detailed description of sampling

Gene pools | Inbred breeds | Other Capra species

Relevant gene pools

Determined by Working Groups from ADAPTmap project (lead by Licia Colli and Paola Crepaldi)

1. Pakistan
2. Northern Africa
3. Western Africa
4. Eastern Africa
5. Southeastern Africa
6. Madagascar
7. Boer
8. Spain & France
9. Saanen
10. Alpine
11. Mediterranean
12. Northern Europe
13. America
14. Australia
15. Wild goats, Turkey, Iran

Gene pools | Inbred breeds | Other Capra species

Inbred breeds

To explore deleterious mutations

1. Icelandic
2. Palmera

Detailed description of sampling

Gene pools | Inbred breeds | Other Capra species

Other Capra species

1. Capra falconeri
2. Capra ibex
3. Capra falconeri

Other *Capra* species

Aim: to study wild x domestic hybridization and adaptive introgression.

UNIVERSITÀ CATTOLICA del Sacro Cuore

VarGoats WGs:

4 WGs follow up from ADAPTmap + 4 new WGs

VarGoats WG11 – "SNP calling and CNV detection": B. Rosen & T. Faraut.
VarGoats WG12 – "Methods (demographic models, imputation)": not started.
VarGoats WG13 – "Extent of loss of function alleles": M.Amills & G. Tosser-Klopp.
VarGoats WG14 – "Hybridization between species": L. Colli & P.Crepaldi.

ADAPTmap WG1 – "Improvement of genome assembly" & ADAPTmap GROUP 2 – "Genome annotation" now called "Genome annotation/Pan Genome Analysis" & **ADAPTmap WG6 –** "Integration, standardization and visualization of genomic data": not started.
ADAPTmap WG3 – "Comparative genomics (with other ruminants)": Clet Wandui Masiga & E.Clark.
ADAPTmap WG7 – "Population genetics analyses and population history domestication reconstruction": L. Colli & P. Crepaldi & F.Pompanon.
ADAPTmap WG8 – "Selection signatures (landscape genomics, iHS, CLL, EHH, XPEHH, Fst, Visible genetic profile)": L. Colli & P. Crepaldi & F.Pompanon.

VarGoats workflow:

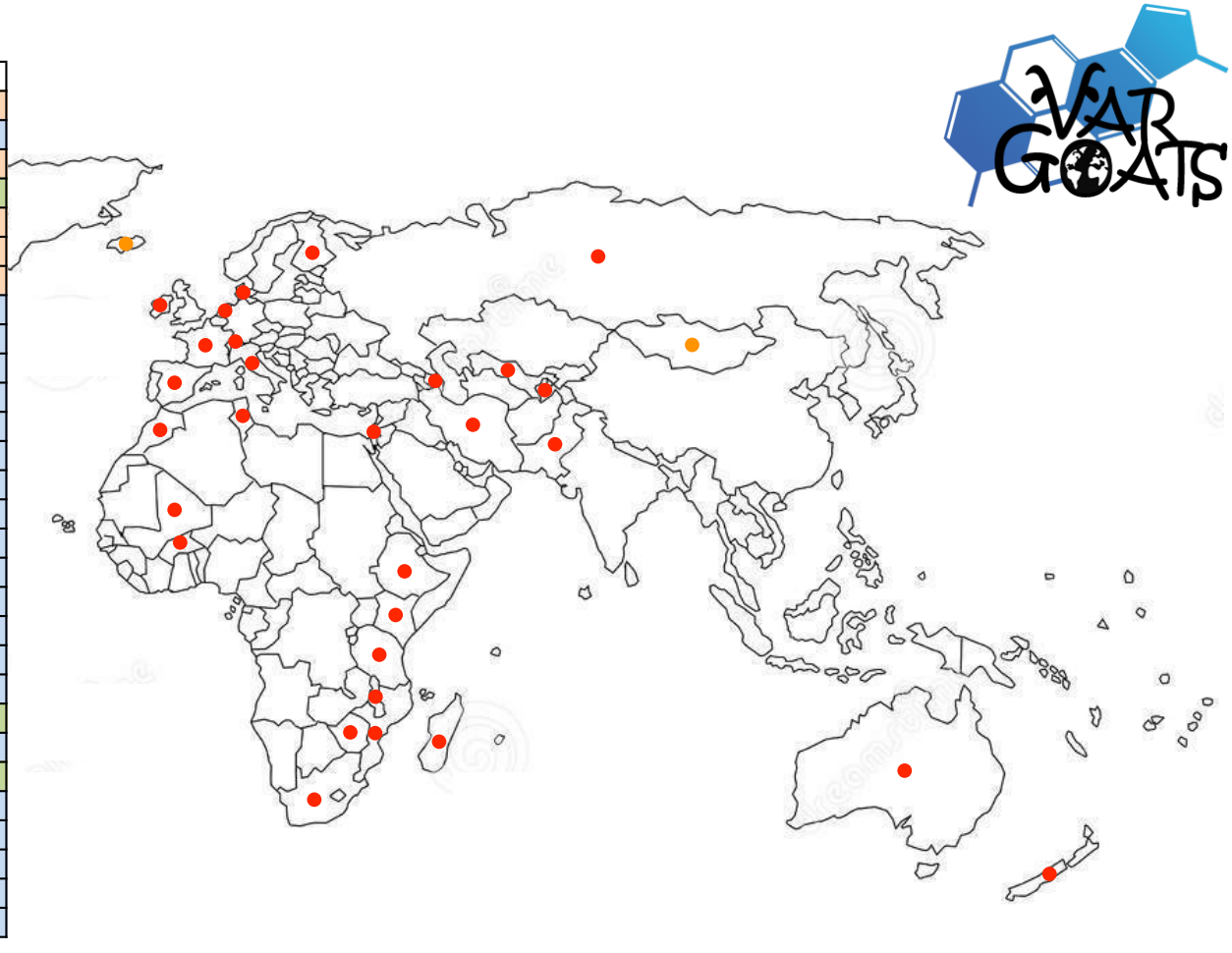
- Sequencing started in Jul. 2016.
- **Several sequence data sources** (Genoscope / CEA / Roslin / Public data).
- Data are being produced in several steps:
  - 248 animals in Dec. 2017 → 830 animals in Nov. 2018 → **ca. 1000 animals in Dec. 2019.**

VarGoats dataset:

To date: 829 individuals representing **8 species**, **84 populations, 30 countries** and **4 continents**.

VarGoats dataset:

continent	country	populations	animals
Asia	Azerbaijan	1	1
	Iran	1	20
	Israel	1	1
	Pakistan	7	31
	Russian Federation	1	1
	Tajikistan	1	1
	Uzbekistan	1	1
Africa	Burkina Faso	1	1
	Ethiopia	4	28
	Kenya	3	16
	Madagascar	6	43
	Malawi	5	25
	Mali	6	36
	Morocco	1	163
	Mozambique	1	163
	South Africa	1	3
	Tanzania	9	66
	Tunisia	1	5
	Zimbabwe	3	28
Europe	Denmark	1	6
	Finland	1	1
	France	18	222
	Ireland	1	5
	Italy	10	39
	Netherlands	1	5
	Spain	4	22
	Switzerland	3	13
Oceania	Australia	3	5
	New Zealand	1	8



815 sequences from **84** local and transboundary **domestic populations**, and **14 sequences** from **7 wild goat species**.

UNIVERSITÀ CATTOLICA del Sacro Cuore

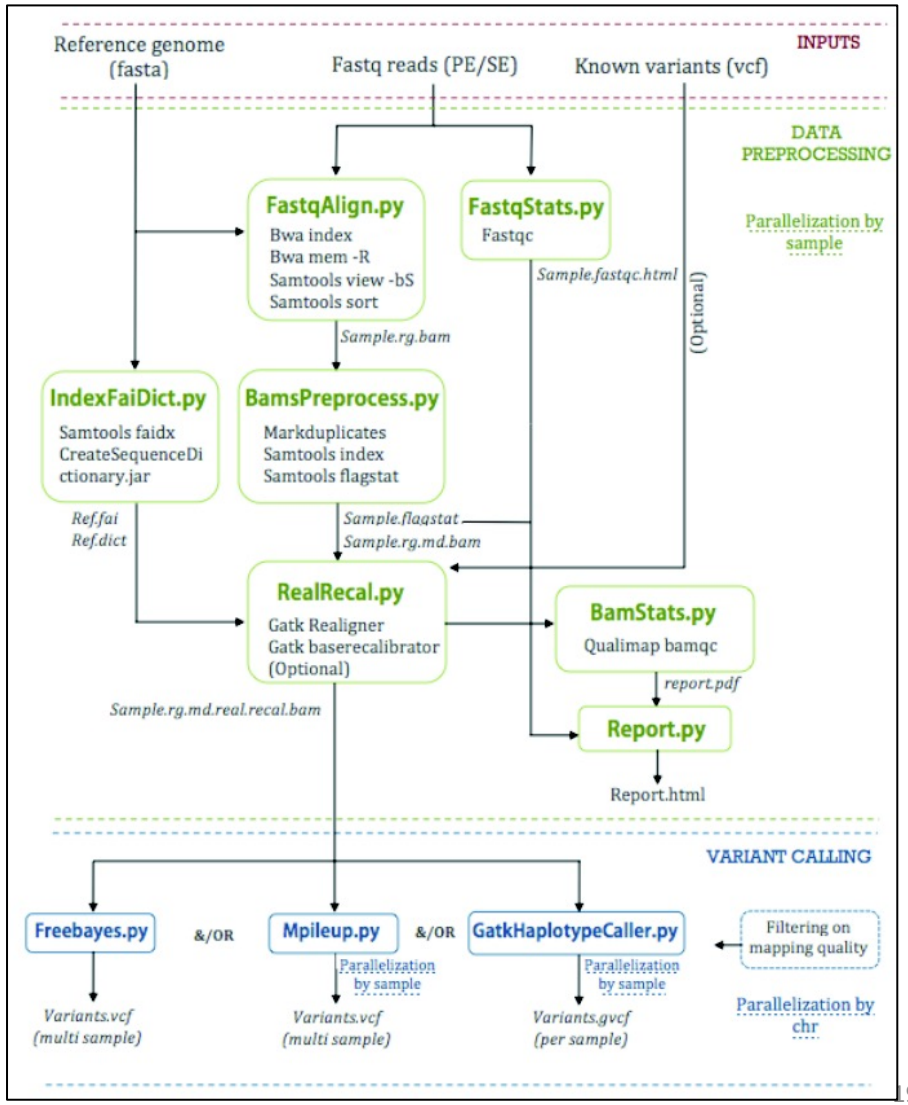
VarGoats workflow:

- Sequencing started in Jul. 2016.
- **Several sequence data sources** (Genoscope / CEA / Roslin / Public data).
- Data are being produced in several steps:
  - 248 animals in Dec. 2017 → 830 animals in Nov. 2018 → **ca. 1000 animals in Dec. 2019.**

- **One single bioformatics pipeline** for SNP calling at TGCC run by Philippe Bardou (INRA-Toulouse, Sigenae team)
- **Filtering criteria and analyses pipelines are being optimized on intermediate datasets.**

Pipeline - alignment and variant calling:

- Based on Jflow workflow manager
- http://jflow.toulouse.inra.fr
- Main steps :
  - **Mapping fastq file on the reference genome ARS1 (BWA)**
  - Mapping post-processing (Picard tools, GATK): **bam file**
  - Variant discovery (GATK HaplotypeCaller, ...): **VCF/gVCF file**
  - Generate a vcf file by chr.
    - Generate a multiple-sample gVCF file (GATK CombineGVCF)
    - Perform "genotyping" and generate **30 vcf files** (GATK GenotypeGVCF)
  - Filter and annotate (snpeff) vcf files and generate **30 vcf annotated files**
  - Data release.



UNIVERSITÀ CATTOLICA del Sacro Cuore

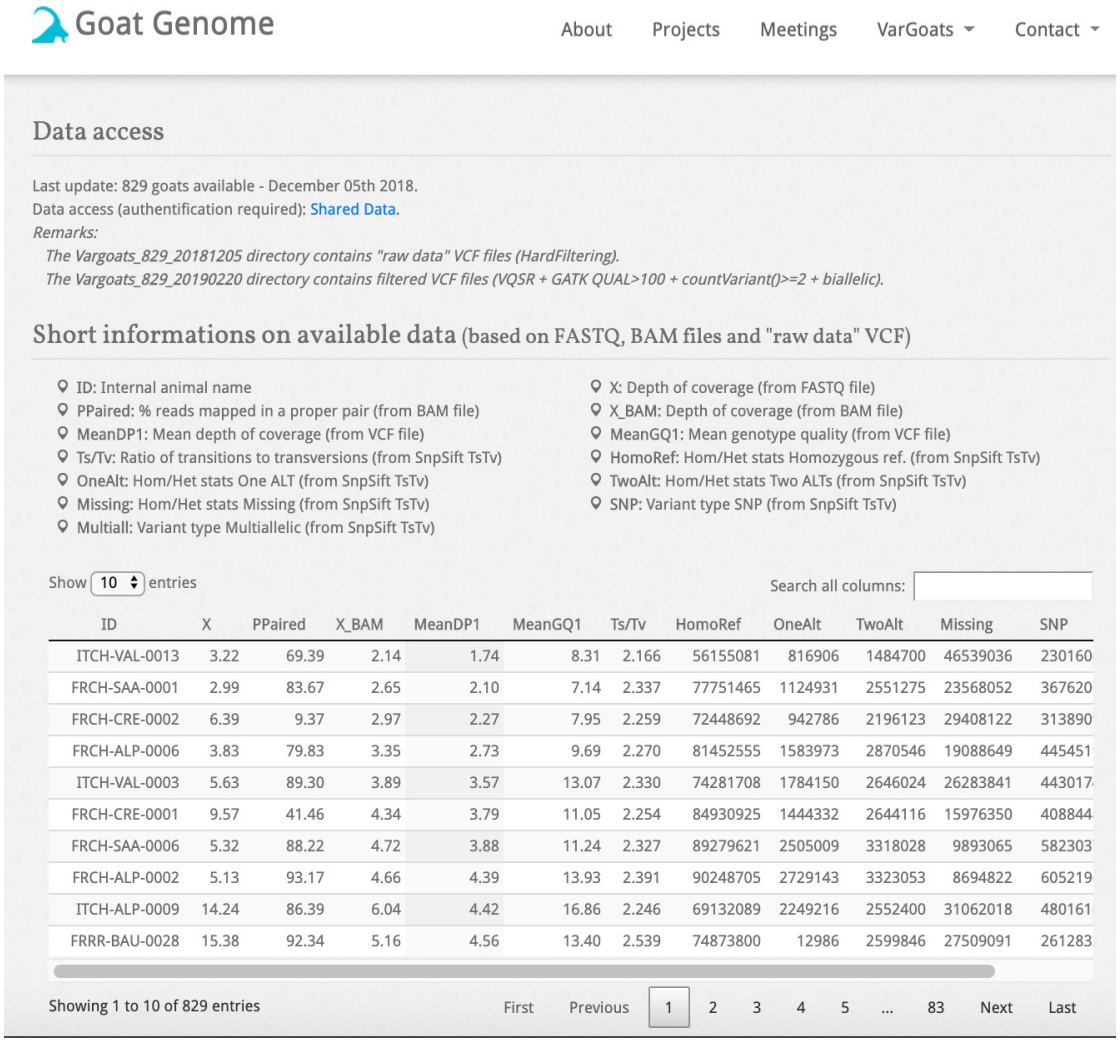
VarGoats workflow:

- Sequencing started in Jul. 2016.
- **Several sequence data sources** (Genoscope / CEA / Roslin / Public data).
- Data are being produced in several steps:
  - 248 animals in Dec. 2017 → 830 animals in Nov. 2018 → **ca. 1000 animals in Dec. 2019.**

- **One single bioformatics pipeline** for SNP calling at TGCC run by Philippe Bardou (INRA-Toulouse, Sigenae team)
- **Filtering criteria and analyses pipelines are being optimized on intermediate datasets.**

- **Data will be released in the public domain** (expected at the end of 2019).

VarGoats searchable database:



Goat Genome

About Projects Meetings VarGoats Contact

Data access

Last update: 829 goats available - December 05th 2018.
Data access (authentification required): Shared Data.
Remarks:
The Vargoats_829_20181205 directory contains "raw data" VCF files (HardFiltering).
The Vargoats_829_20190220 directory contains filtered VCF files (VQSR + GATK QUAL>100 + countVariant()>=2 + biallelic).

Short informations on available data (based on FASTQ, BAM files and "raw data" VCF)

- ID: Internal animal name
- PPaired: % reads mapped in a proper pair (from BAM file)
- MeanDP1: Mean depth of coverage (from VCF file)
- Ts/Tv: Ratio of transitions to transversions (from SnpSift TsTv)
- OneAlt: Hom/Het stats One ALT (from SnpSift TsTv)
- Missing: Hom/Het stats Missing (from SnpSift TsTv)
- Multiall: Variant type Multiallelic (from SnpSift TsTv)
- X: Depth of coverage (from FASTQ file)
- X_BAM: Depth of coverage (from BAM file)
- MeanGQ1: Mean genotype quality (from VCF file)
- HomoRef: Hom/Het stats Homozygous ref. (from SnpSift TsTv)
- TwoAlt: Hom/Het stats Two ALTs (from SnpSift TsTv)
- SNP: Variant type SNP (from SnpSift TsTv)

Show 10 entries Search all columns:

ID	X	PPaired	X_BAM	MeanDP1	MeanGQ1	Ts/Tv	HomoRef	OneAlt	TwoAlt	Missing	SNP
ITCH-VAL-0013	3.22	69.39	2.14	1.74	8.31	2.166	56155081	816906	1484700	46539036	230160
FRCH-SAA-0001	2.99	83.67	2.65	2.10	7.14	2.337	77751465	1124931	2551275	23568052	367620
FRCH-CRE-0002	6.39	9.37	2.97	2.27	7.95	2.259	72448692	942786	2196123	29408122	313890
FRCH-ALP-0006	3.83	79.83	3.35	2.73	9.69	2.270	81452555	1583973	2870546	19088649	445451
ITCH-VAL-0003	5.63	89.30	3.89	3.57	13.07	2.330	74281708	1784150	2646024	26283841	443017
FRCH-CRE-0001	9.57	41.46	4.34	3.79	11.05	2.254	84930925	1444332	2644116	15976350	408844
FRCH-SAA-0006	5.32	88.22	4.72	3.88	11.24	2.327	89279621	2505009	3318028	9893065	582303
FRCH-ALP-0002	5.13	93.17	4.66	4.39	13.93	2.391	90248705	2729143	3323053	8694822	605219
ITCH-ALP-0009	14.24	86.39	6.04	4.42	16.86	2.246	69132089	2249216	2552400	31062018	480161
FRRR-BAU-0028	15.38	92.34	5.16	4.56	13.40	2.539	74873800	12986	2599846	27509091	261283

Showing 1 to 10 of 829 entries First Previous 1 2 3 4 5 ... 83 Next Last

http://www.goatgenome.org/vargoats_data_access.html

UNIVERSITÀ CATTOLICA del Sacro Cuore

VarGoats preliminary results:

- **829 genomes** to date.
- **Sequencing depth** between **1.74x** and **35.38x**.
- 16 WGS < 5x depth → 9 sequences with <4.5x depth will be discarded.
- Avg. Number of SNPs:
  - **SNPs overall**: domestic goats 2.3M-9.6M; wild goats 2.6M-21.7M.

UNIVERSITÀ CATTOLICA del Sacro Cuore

Variant selection:

Biallelic SNPs ✓

Multiallelic SNPs ?

- Number?
- Species-specific differences?
- Distribution along the genome?
- Distribution with respect to geographical origin?

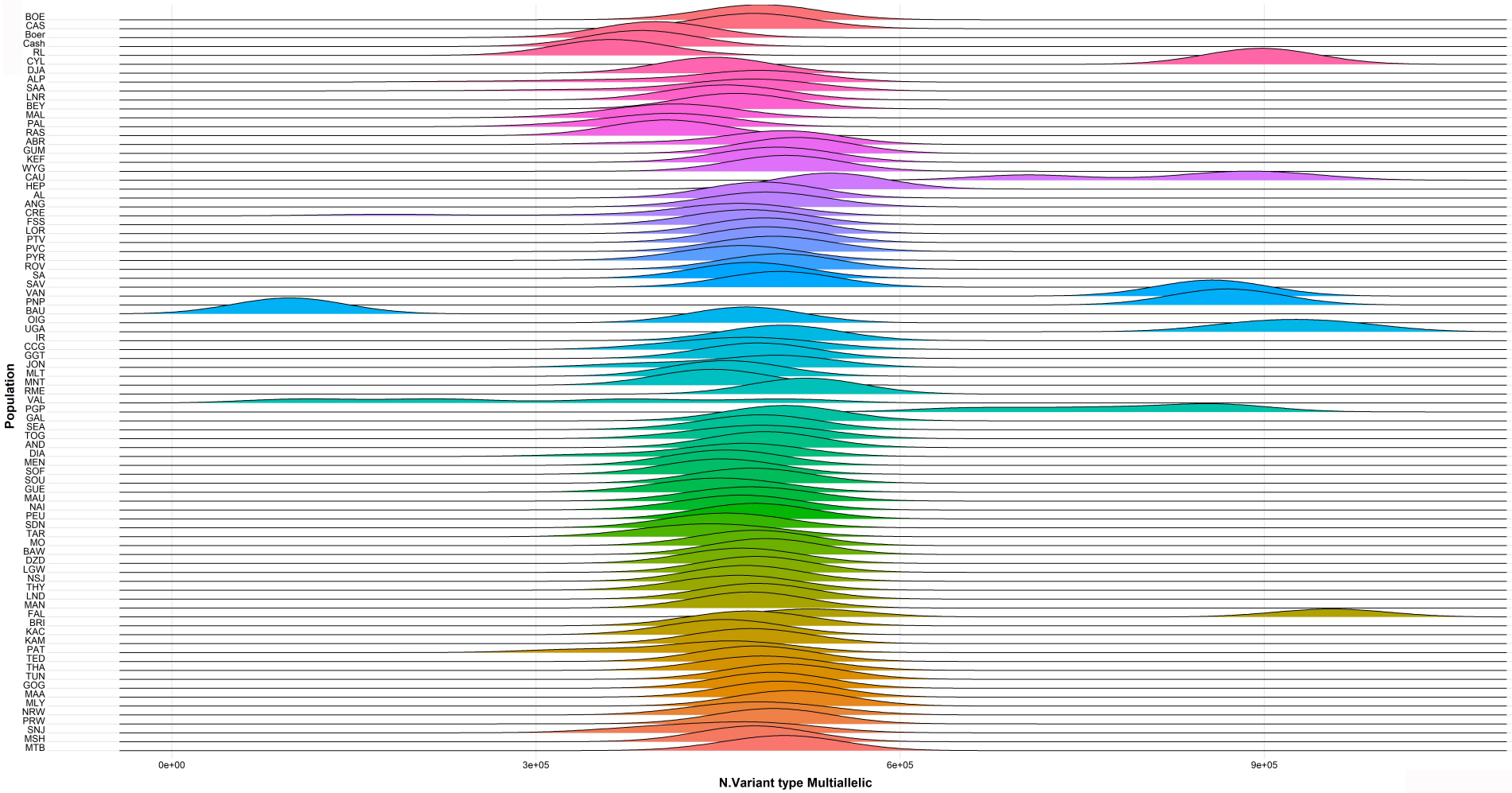
UNIVERSITÀ CATTOLICA del Sacro Cuore

VarGoats preliminary results:

- **829 genomes** to date.
- **Sequencing depth** between **1.74x** and **35.38x**.
- 16 WGS < 5x depth → 9 sequences with <4.5x depth will be discarded.
- Avg. Number of SNPs:
  - **SNPs overall**: domestic goats 2.3M-9.6M; wild goats 2.6M-21.7M.
  - **Multiallelic SNPs**: domestic goats 0.12M-0.57M; wild goats 0.09M-1.04M.

Results – Multiallelic variants:

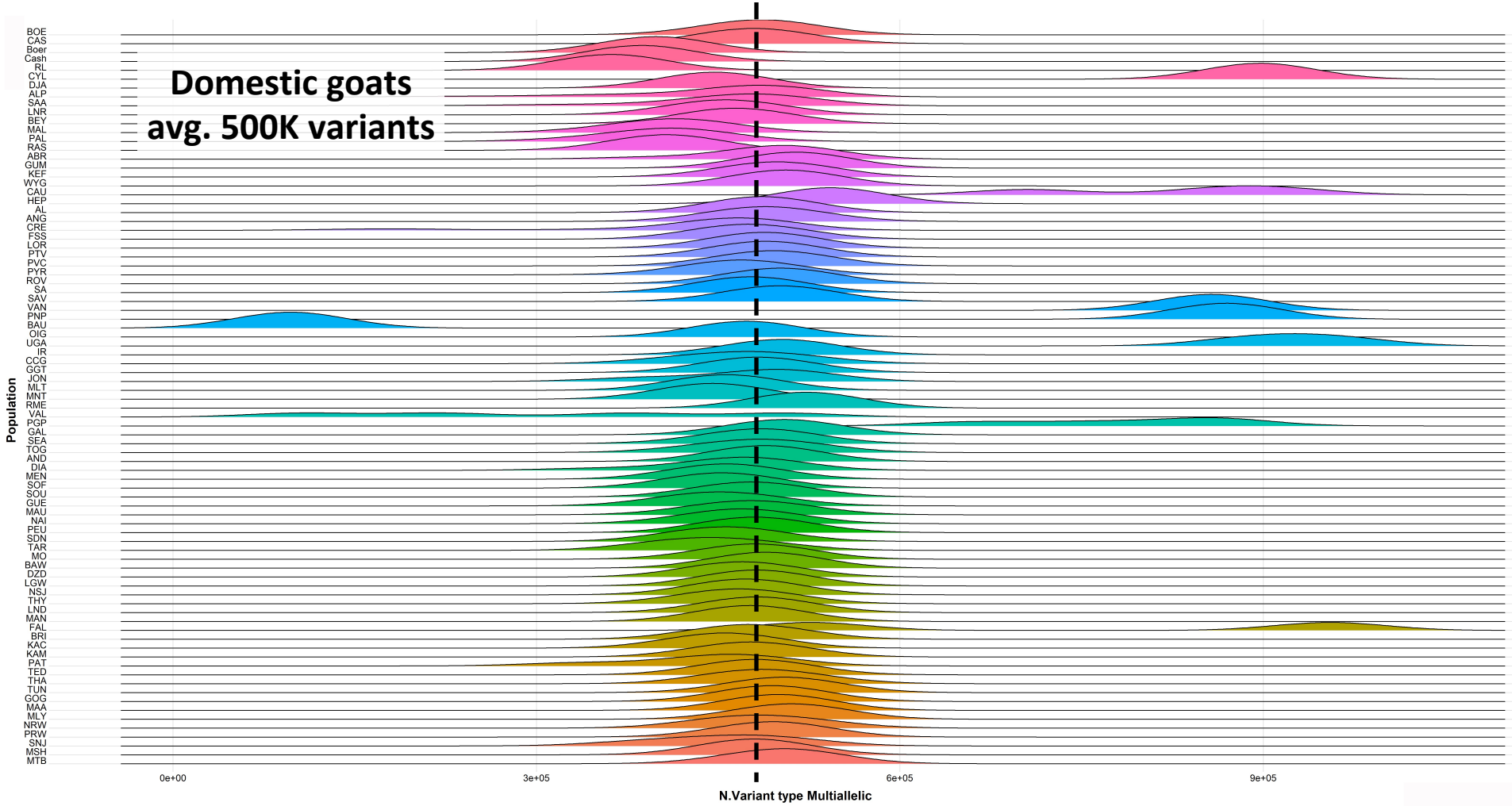
Multiallelic variants per population



UNIVERSITÀ CATTOLICA del Sacro Cuore

Results – Multiallelic variants:

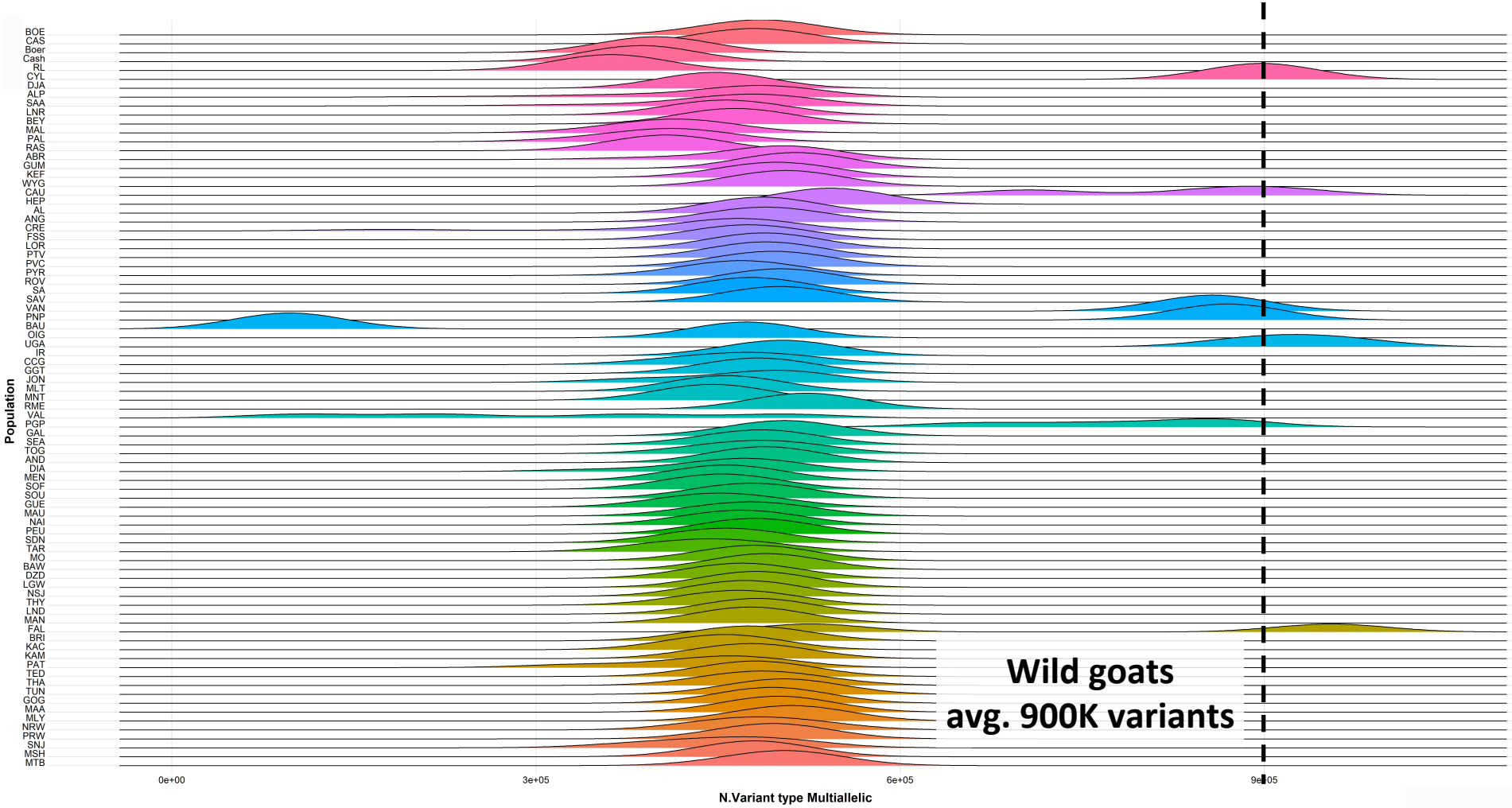
Multiallelic variants per population



UNIVERSITÀ CATTOLICA del Sacro Cuore

Results – Multiallelic variants:

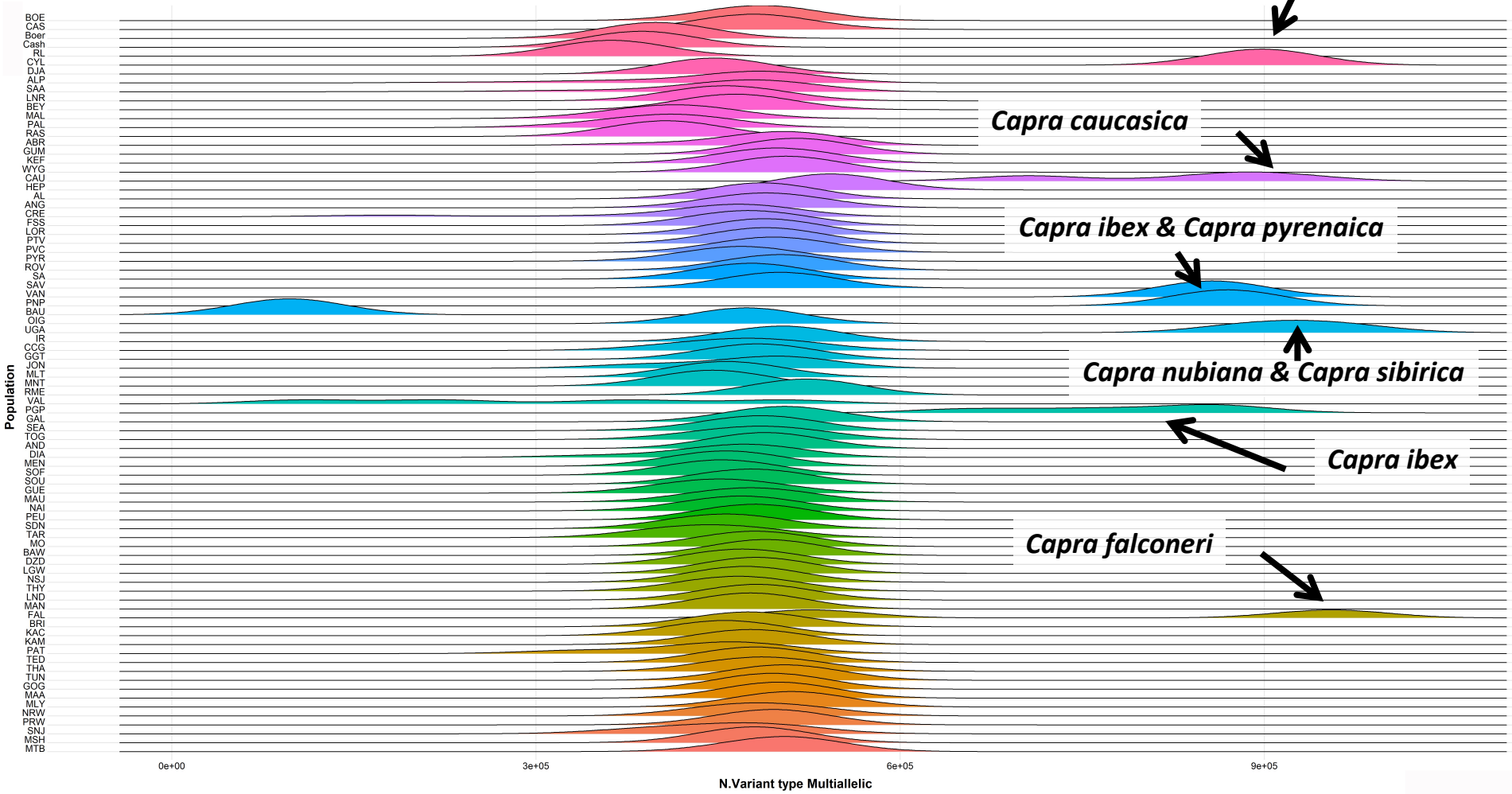
Multiallelic variants per population



UNIVERSITÀ CATTOLICA del Sacro Cuore

Results – Multiallelic variants:

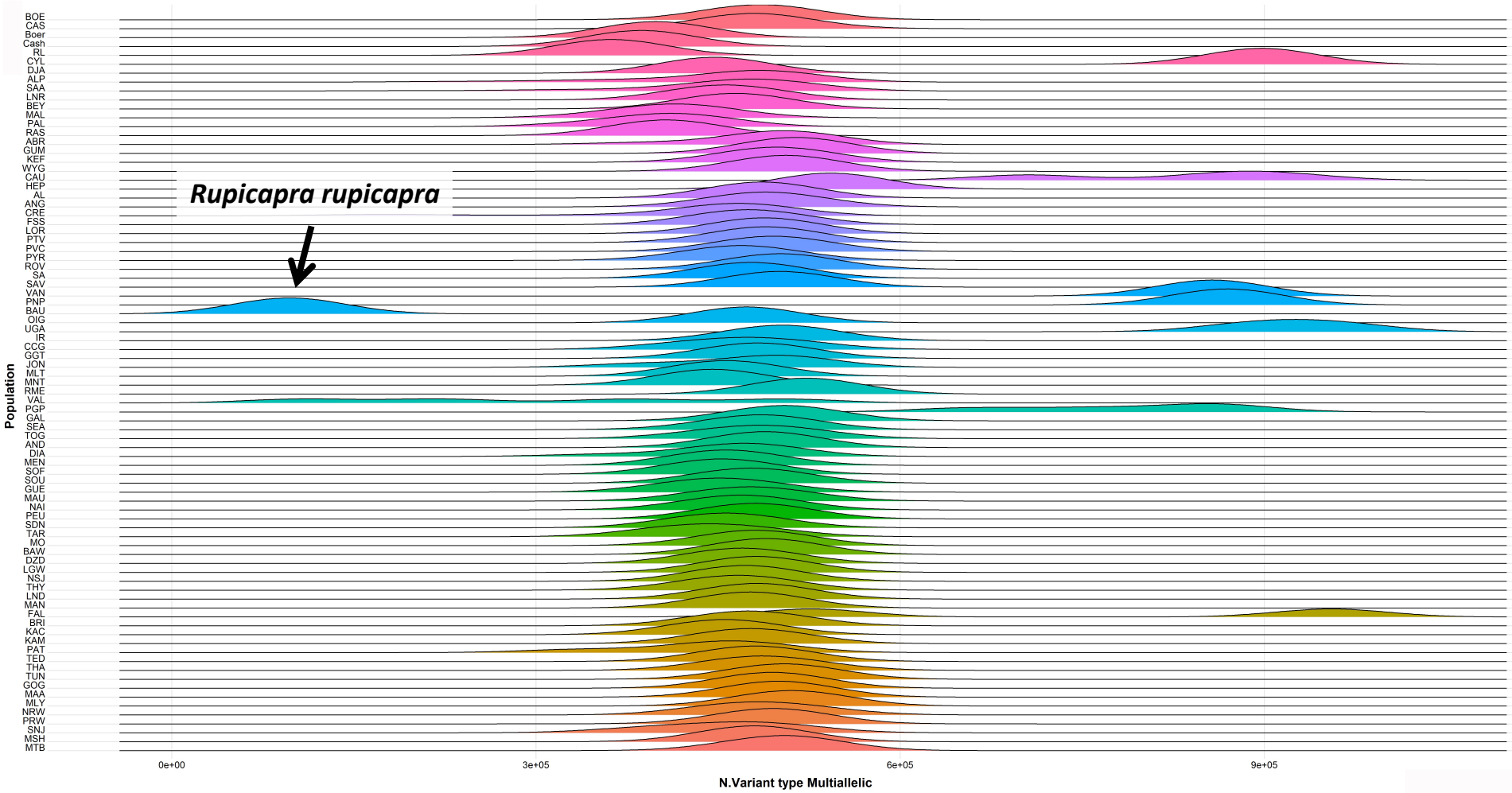
Multiallelic variants per population



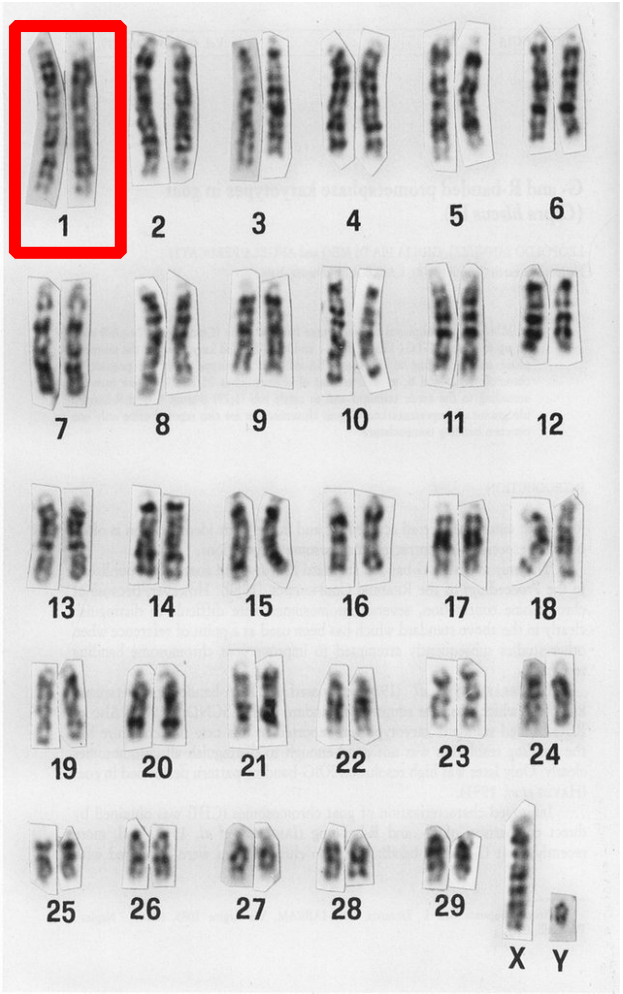
UNIVERSITÀ CATTOLICA del Sacro Cuore

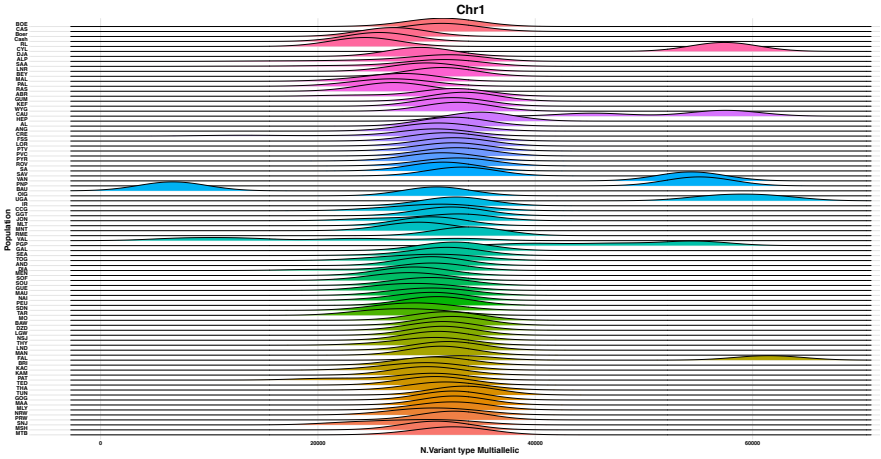
Results – Multiallelic variants:

Multiallelic variants per population



Multiallelic variants per chromosome:

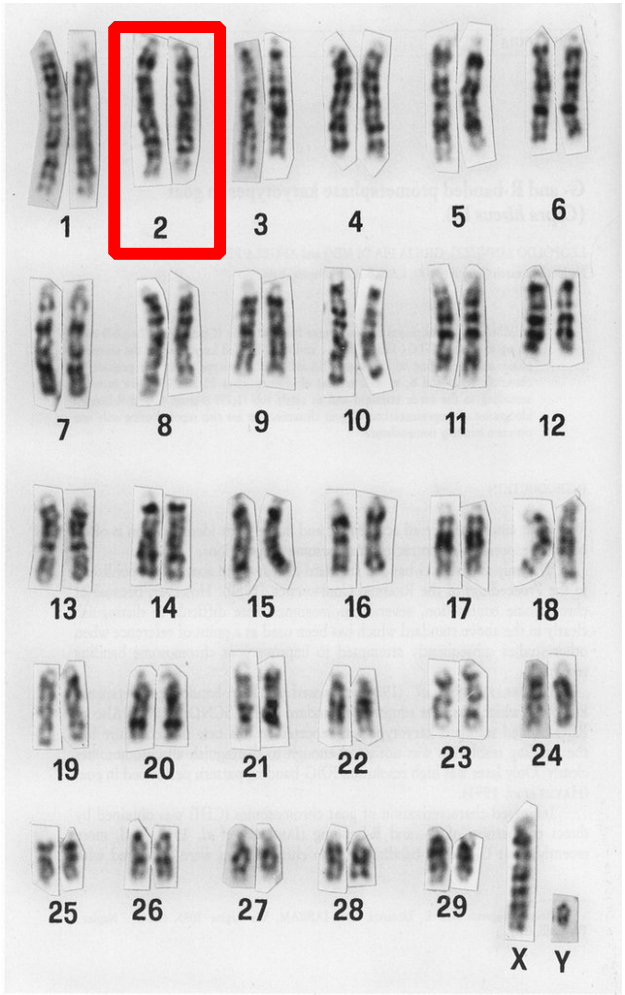


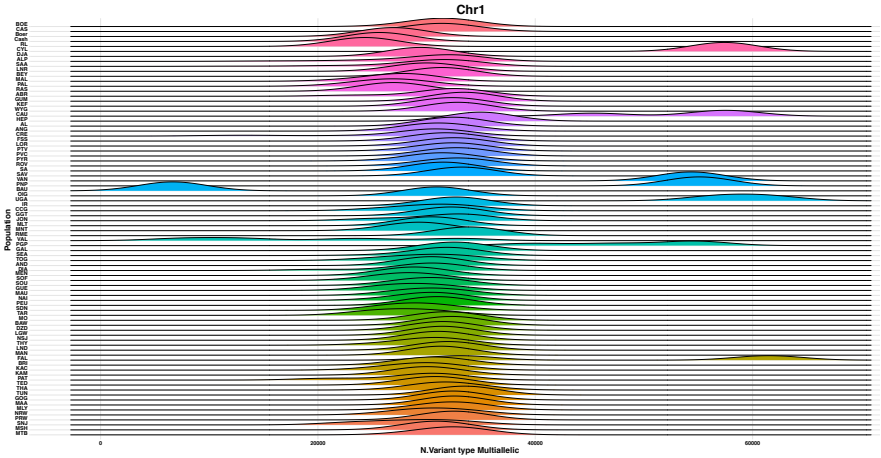


CHR1 = 32K multiallelic SNPs on avg.

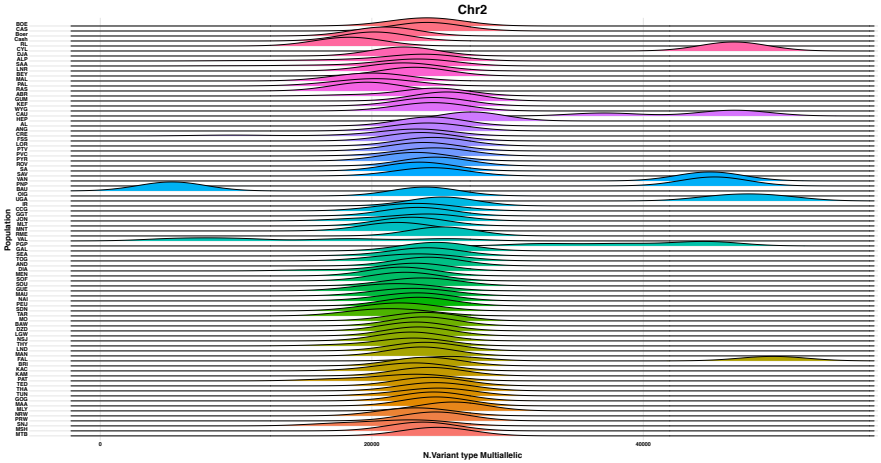
UNIVERSITÀ CATTOLICA del Sacro Cuore

Multiallelic variants per chromosome:



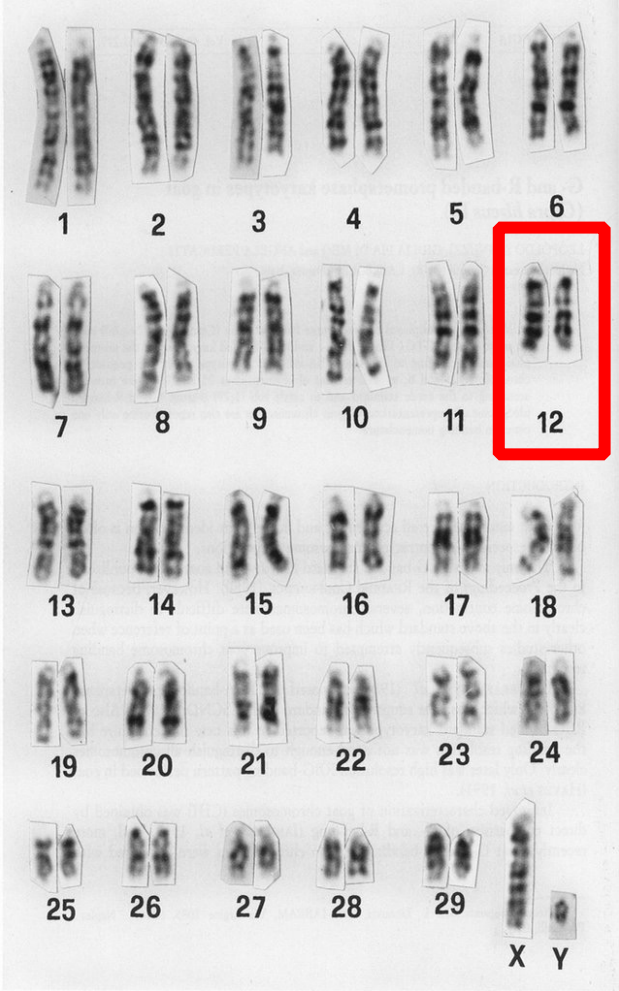


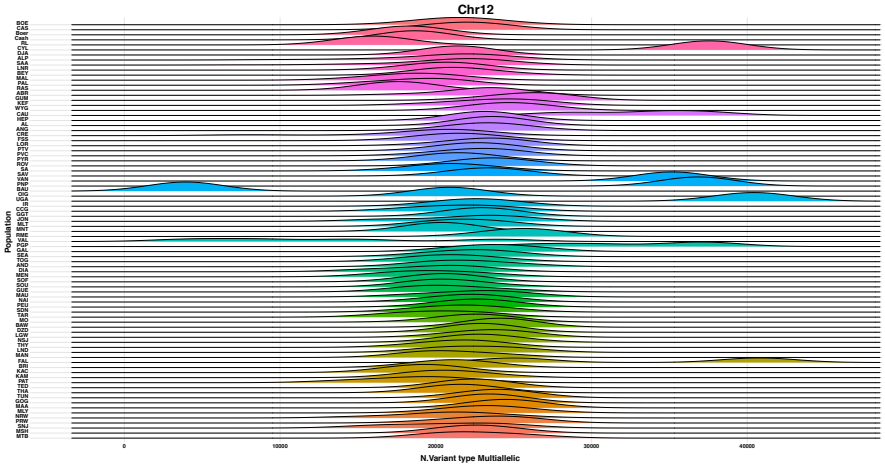
CHR1 = 32K multiallelic SNPs on avg.



CHR2 = 25K multiallelic SNPs on avg.

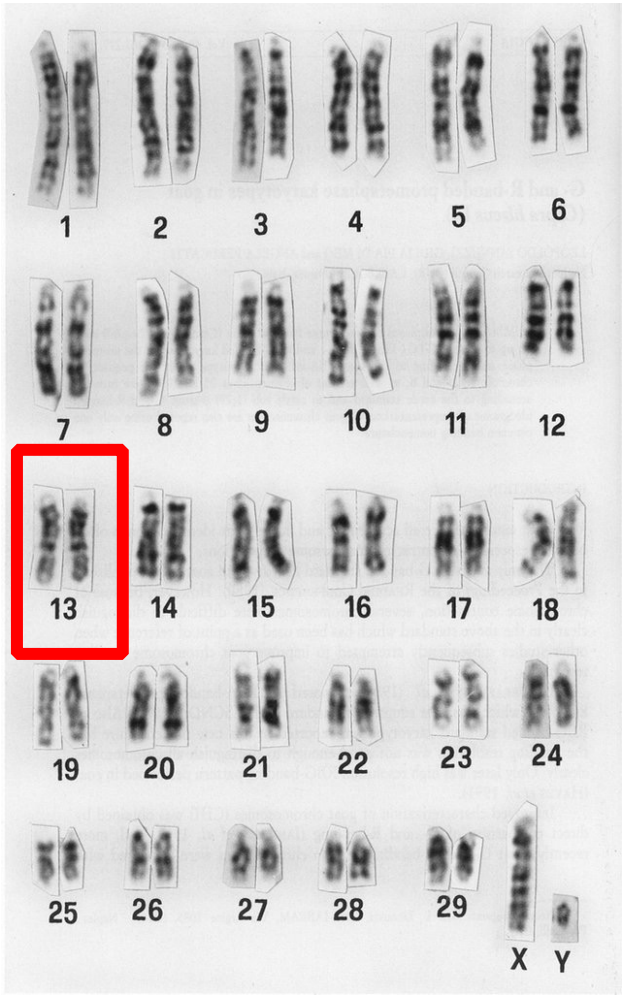
Multiallelic variants per chromosome:

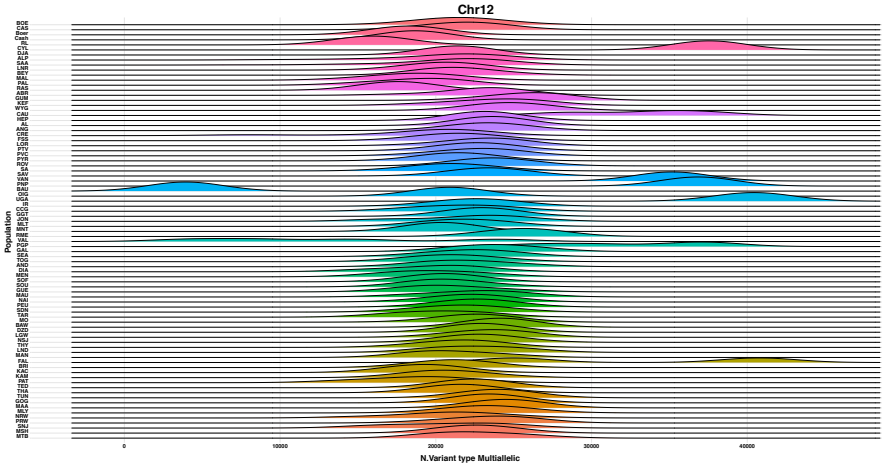




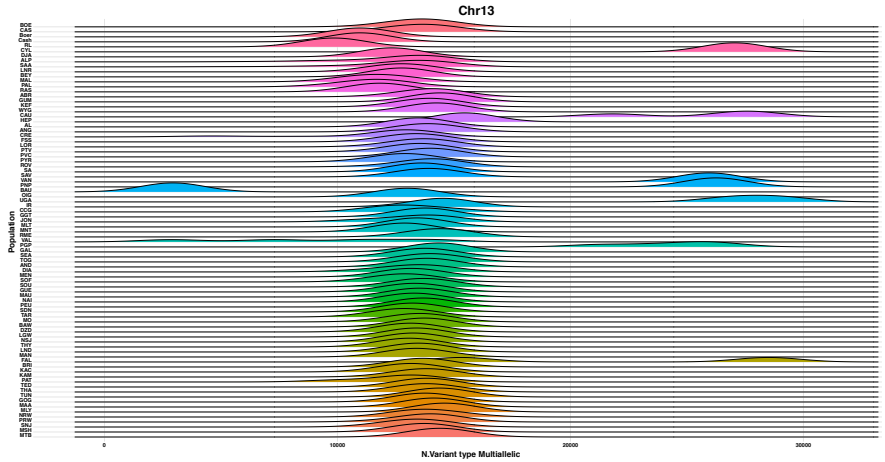
CHR12 = 22K multiallelic SNPs on avg.

Multiallelic variants per chromosome:





CHR12 = 22K multiallelic SNPs on avg.



CHR13 = 14K multiallelic SNPs on avg.

UNIVERSITÀ CATTOLICA del Sacro Cuore

Conclusions:

• The **preliminary evaluation of diversity** suggests that **wild goat species** have a **higher variation** (10-20M SNPs) compared to domestic goats (<10M SNPs).

• **Multiallelic SNPs** show a **similar trend** and represent **ca. 1/20 of the total variants**.

• **Multialleic SNPs number** is generally **proportional to chromosome size** (some exceptions → further investigation).

Next steps → evaluate their genomic and within-chromosome distribution, and their functional role.

We want you(r samples):

60 WG sequencing slots are still available

(sequencing completion deadline Dec. 2019).

We'd like to improve the number of Asian goats

→ Samples?

Deadline for providing DNA: Sept. 2019.

Acknowledgements:

- Institutions: INRA, USDA, PTP, CNRS, UGA
- Platforms: Adriana Alberti, Patrick Wincker (Génoscope),
- Bioinformaticians: Philippe Bardou (INRA), S.Engelen (Génoscope)
- Scientists (non exhaustive):
  - Jim Reecy and Muhammad Moaeen-ud-Din, Pakistan
  - Marcel Amills, Spain
  - Alessandra Stella, Paola Crepaldi , Italy
  - Thomas Faraut, François Pompanon, Gwenola Tosser-Klopp, France
  - Ben Rosen, Curt Van Tassell, USDA, AGIN
  - Emily Clark, Clet Wandui Masiga

Special thanks:
INRA: Laure Denoyelle, Thomas Faraut.
UCSC: Marcello Del Corvo.
Roslin: Andrea Talenti.
DTU: Francesca Bertolini.

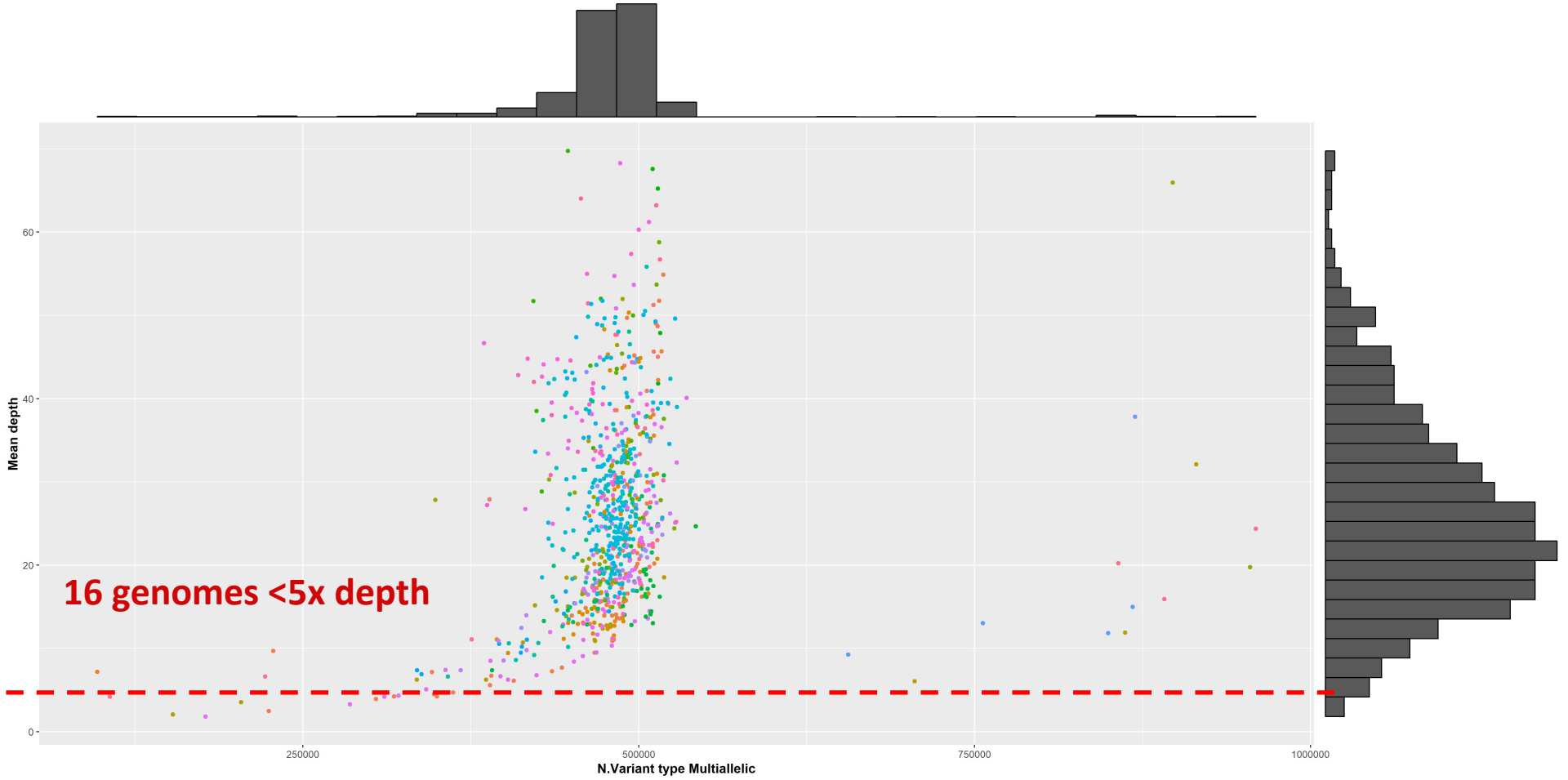
- Technicians: Julien Sarry (INRA), Céline Orvain (Génoscope)
- Animal selection: Isabelle Palhière (French goats), Licia Colli (International goats)
- Sample providers (non exhaustive):
  - Jim Reecy and Muhammad Moaeen-ud-Din, Pakistan
  - Marcel Amills, Amparo Martinez, Vincenzo Landi, Felix Goyache and Isabelle Alvarez & Armand Sanchez, Spain
  - Alessandra Stella, Paola Crepaldi and the Italian Goat Consortium: Alessandra Crisà, (IGC-Crea, Italy), Donata Marletta (IGC-UNICT Italy), Tonello Carta (IGC-Agris-Sardegna, Italy) and Paolo Ajmone Marsan (IGC-UNICATT, Italy)
  - James Kijas, Australia
  - Christine Flury, Cord Droegemueller, Switzerland
  - Capgenes, Daniel Allain, Michel Naves, Isabelle Palhière, Rachel Rupp & Gwenola Tosser-Klopp, France
  - Vivi Hunnicke Nielsen & Bernt Gudbrandtsen, Denmark
  - Hans Lenstra, The Netherlands
  - Jon Hallsteinn Hallsson, Iceland
  - Carina Visser, South Africa
  - Seán Carolan, Old Irish Goat Society, Ireland
  - Dylan, Duby, Museum d'histoire naturelle, France
  - Michele Ottino, Parco National Gran Paradiso, Italy
  - Ben Rosen, AGIN, USDA, Africa

Thank you for your attention.

UNIVERSITÀ CATTOLICA del Sacro Cuore

Results – Multiallelic variants:

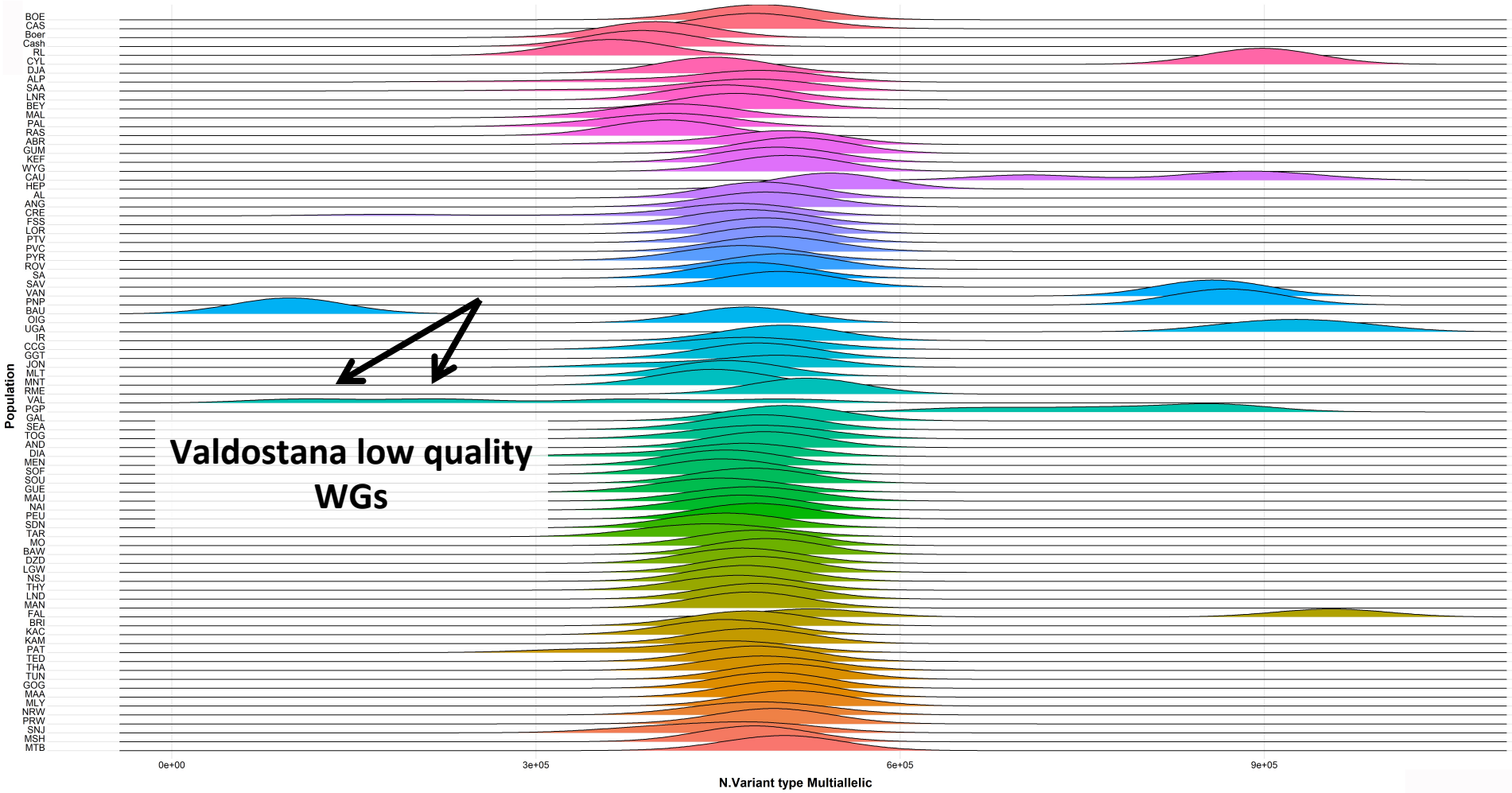
Mean depth vs. number of multiallelic variants



UNIVERSITÀ CATTOLICA del Sacro Cuore

Results – Multiallelic variants:

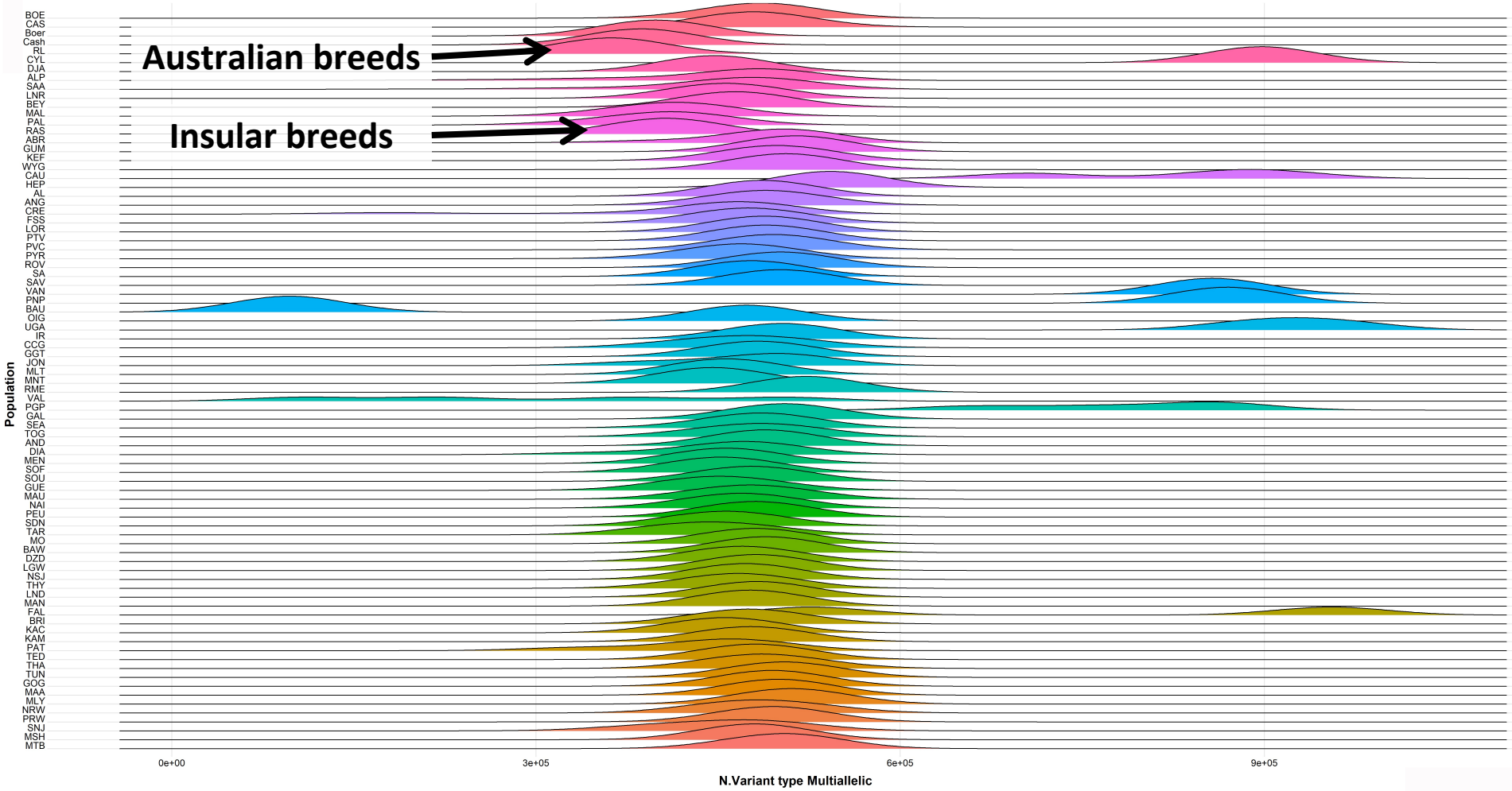
Multiallelic variants per population



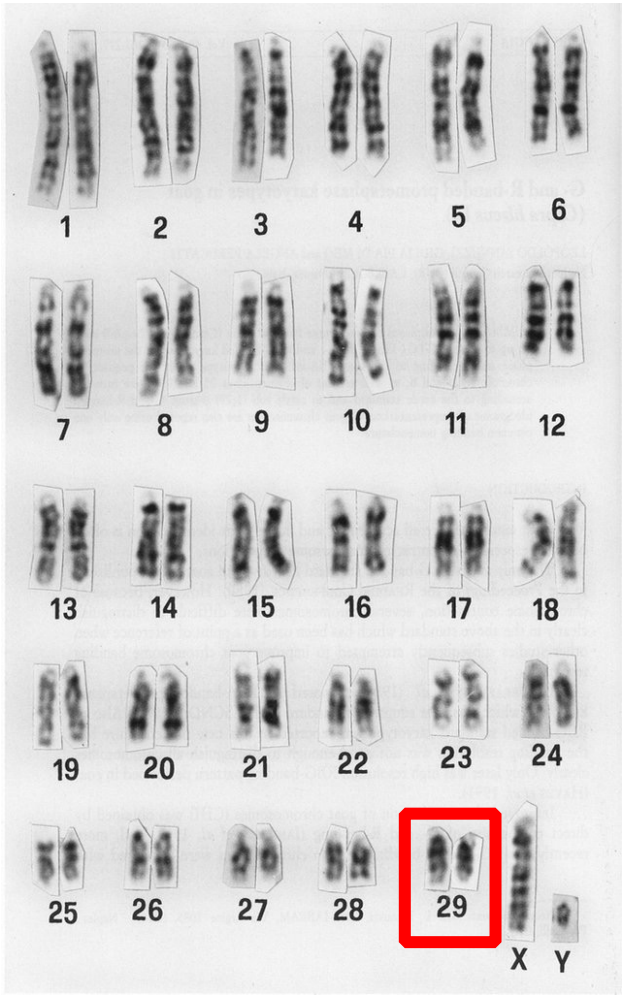
UNIVERSITÀ CATTOLICA del Sacro Cuore

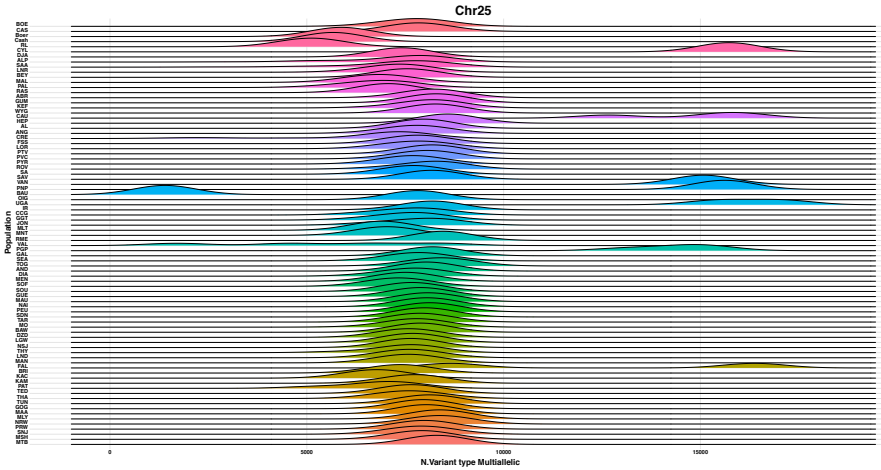
Results – Multiallelic variants:

Multiallelic variants per population

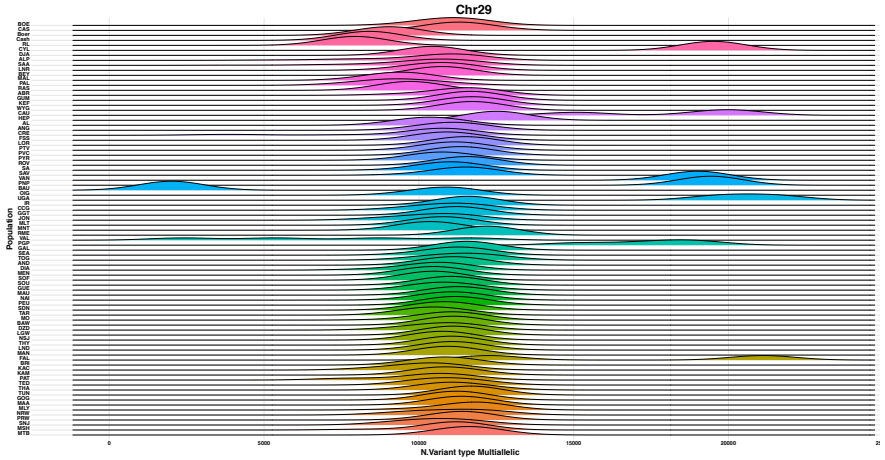


Multiallelic variants per chromosome:





CHR25 = 8K multiallelic SNPs on avg.



CHR29 = 12K multiallelic SNPs on avg.