Annotation of selection signatures in the bovine breed Asturiana de Valles

Cyriel Paris (1), **Simon Boitard** (1), Bertrand Servin (1),
Natalia Sevane (2), Susana Dunner (2)

(1) INRA GenPhySE, Castanet-Tolosan, France. (2) UCM, Madrid, Spain.

EAAP 2019

Introduction

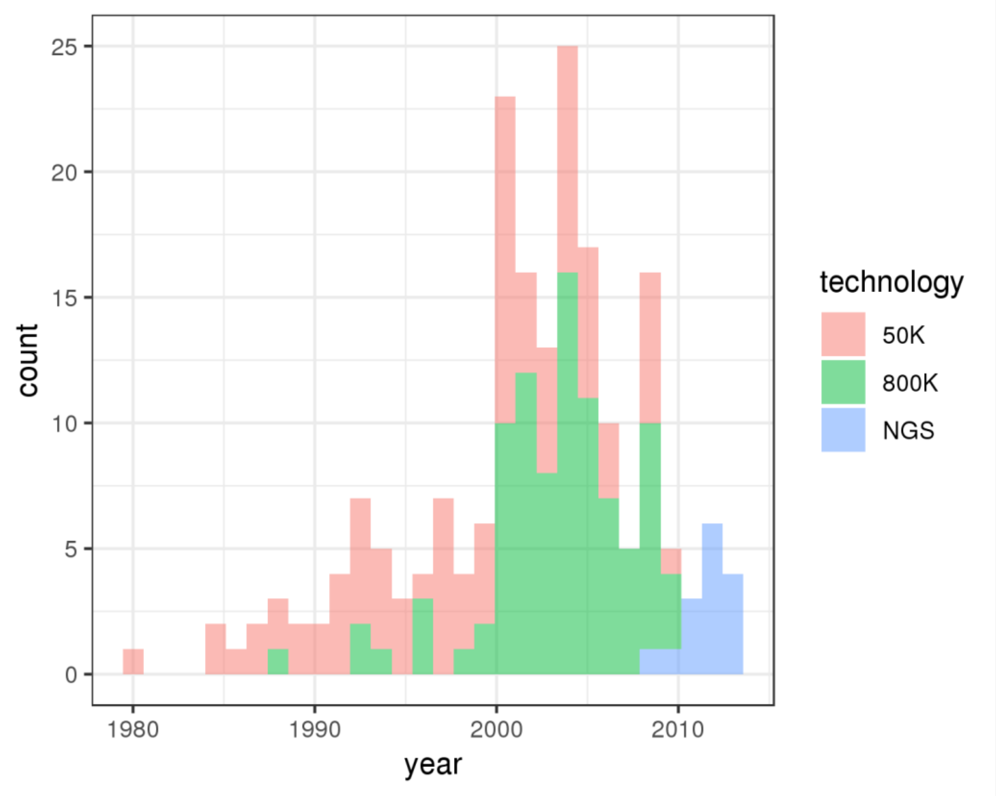
- **Past positive selection** affects genetic diversity and can be detected from present time molecular data.
- **Annotation of selection events**: adaptive trait? Onset and intensity of selection?
- Difficult from present time data, **easier from time series data**.
- **Usefulness of gene banks** to (i) detect recent selection or (ii) annotate signatures of historical selection?

Case study: Asturiana de los Valles

- Spanish autochtonous **beef cattle breed**.
- North of Spain (Asturias).
- Semi-extensive breeding conditions.
- Evolution of genetic diversity along 35 years (1980 – 2015).

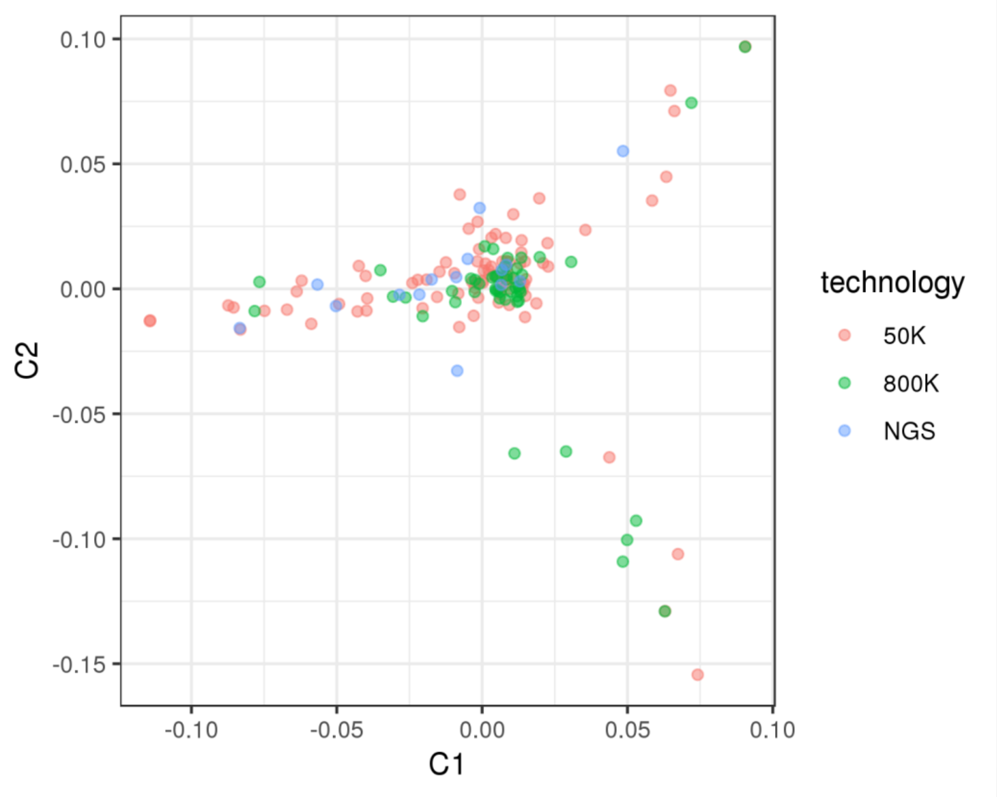
Sampling

- **Genotype data** (50K or 800K) already available for **137 animals**.
- **15 animals sequenced** within IMAGE at ≈ 8X coverage.



Merging data

- After **quality filters:**
  - 14,328,987 SNPs detected from IMAGE NGS data.
  - 43,950 SNPs from chip data.
  - 35,656 SNPs in common.
- **No detectable effect of the technology.**

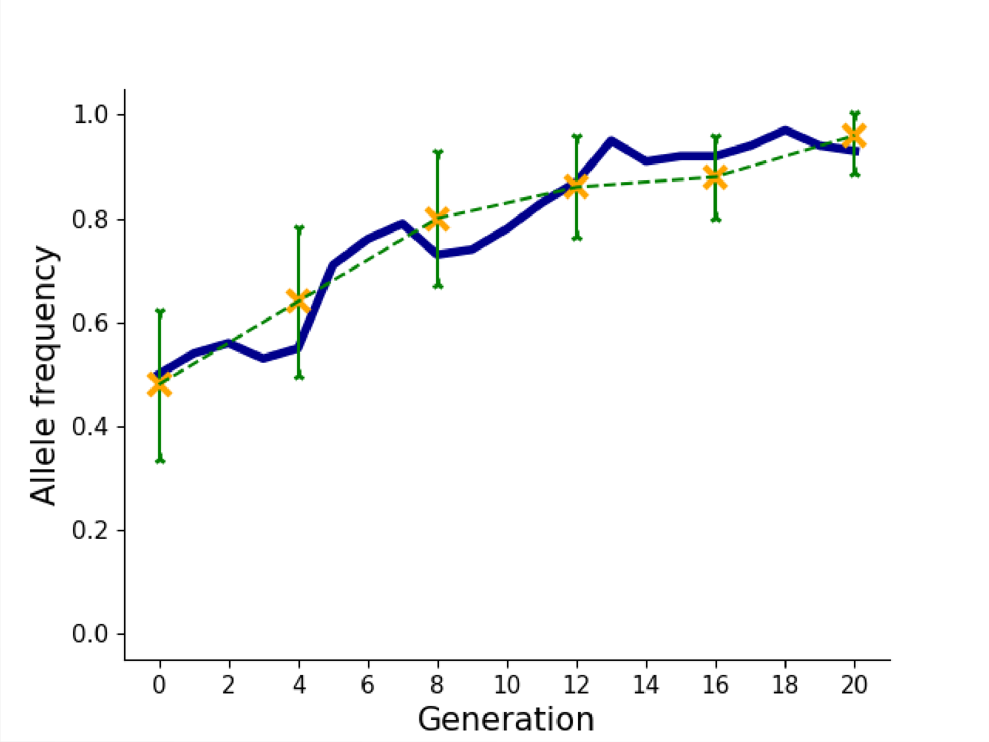


Final time series data set

- Generation time of 4 years.
- Inbred or related animals (within generations) removed.

Generation	1	2	3	4	5	6	7	8	9
First year	1980	1984	1988	1992	1996	2000	2004	2008	2012
Initial sample size	1	5	11	15	21	40	38	12	10
Final sample size	0	4	8	13	17	28	29	9	9

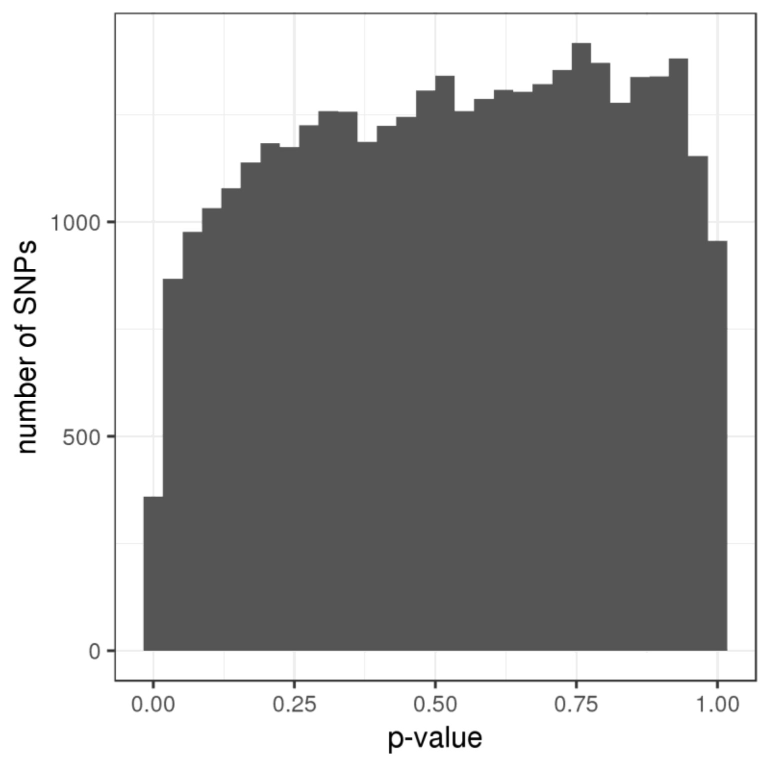
Selection signatures: time series approach



- Population allele frequency trajectory at a locus informative about selection.
- Only a few samples of limited size observed.
- **New method** to **detect selected loci** and **estimate selection intensity** from **temporal samples** (Paris et al, bioRxiv).
- **Fast** (scalable to genome-wide analyzes) and accurate.

Time series approach: results

- Estimated **effective population size** of **400** animals (Ne R package, Hui and Burt 2015).
- No evidence of selection at the single SNP level.
- **5 regions enriched in low p-values** using a local score approach (Fariello *et al*, 2017).



Time series approach: candidate regions

Chr	**Start (bp)**	**End (bp)**	**Length (bp)**	**Nb SNP**	**Genes**
10	45387461	45564676	177215	7	PLEKH02, PIF1, *RBPMS2*, *OAZ2*, *ZNF609*, *RF00413*, TRIP4
13	41414256	41529941	115685	3	-
17	4675045	4750693	75648	4	*FHDC1*, *ARFIP1*
17	31268164	31632465	364301	8	-
22	39414833	39491373	76540	3	*PTPRG*

- Candidate genes related to **carcass and meat traits** (RBPMS2, OAZ2) or **milk traits** (ARFIP1).

Selection signatures: present time data

- **Signatures of historical selection** detected from **image NGS data** only (15 animals from generations 8 and 9).
- nSL statistic (Ferrer-Admetlla et al, 2014): looks for **long haplotypes** at **high frequency**.
- No proper p-value associated to a given nSL score: **outlier approach.**

Present time data: SNPs shared with the chip

Chr	Start (bp)	End (bp)	Length (bp)	Nb SNP	Log10 (pval)	Genes
2	7169804	7270116	100312	2	9.82	COL5A2, COL3A1
2	8476975	8476975	1	1	8.12	-
6	55360713	55360713	1	1	7.85	-
10	98290813	98290813	1	1	10.10	FLRT2
25	13647777	13647777	1	1	7.10	PARN, BFAR, PLA2G10

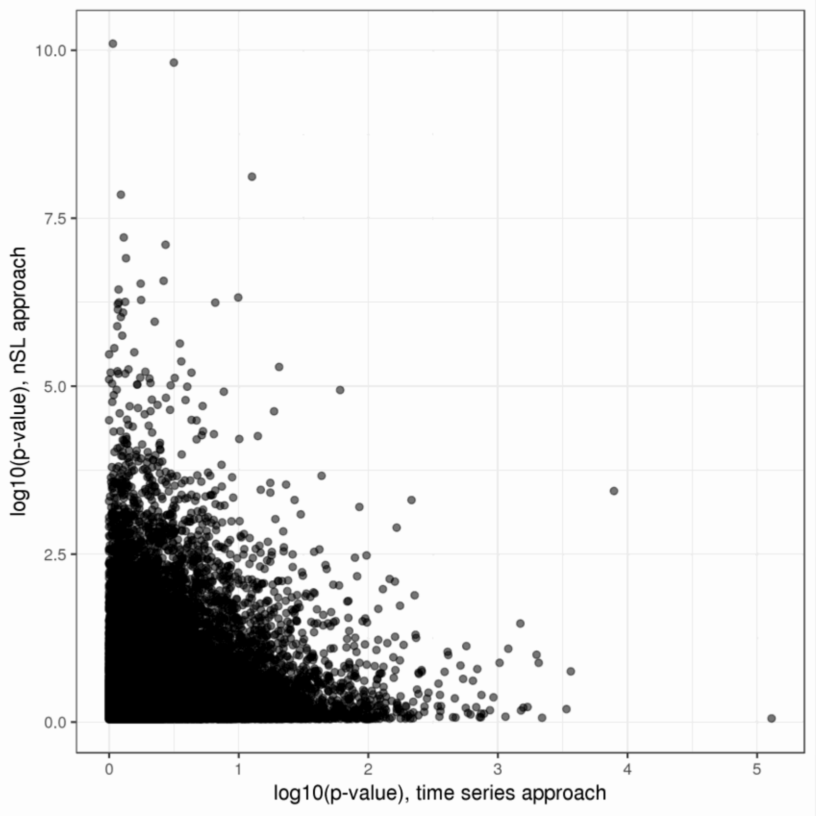
- One region close to the **MSTN gene** (Chr2, ≈ 6.2Mb), one allele nt821(del11) associated to **double muscling** selected in Asturiana (Dunner et al, 2003).
- FLRT2: **embryonic development**, a **calf birth weight** in Holstein.

Present time data: all NGS SNPs

- **Many more signals** due to **higher SNP density** (≈ 10.5 millions).
- The 4 most extreme regions include previous candidate regions close to **MSTN and FLRT2**.
- **Two additional regions** on Chr 4 (≈ 11.3 Mb) with **immunity** related genes and Chr15 (≈ 46.5Mb) with **olfactory receptor** genes.

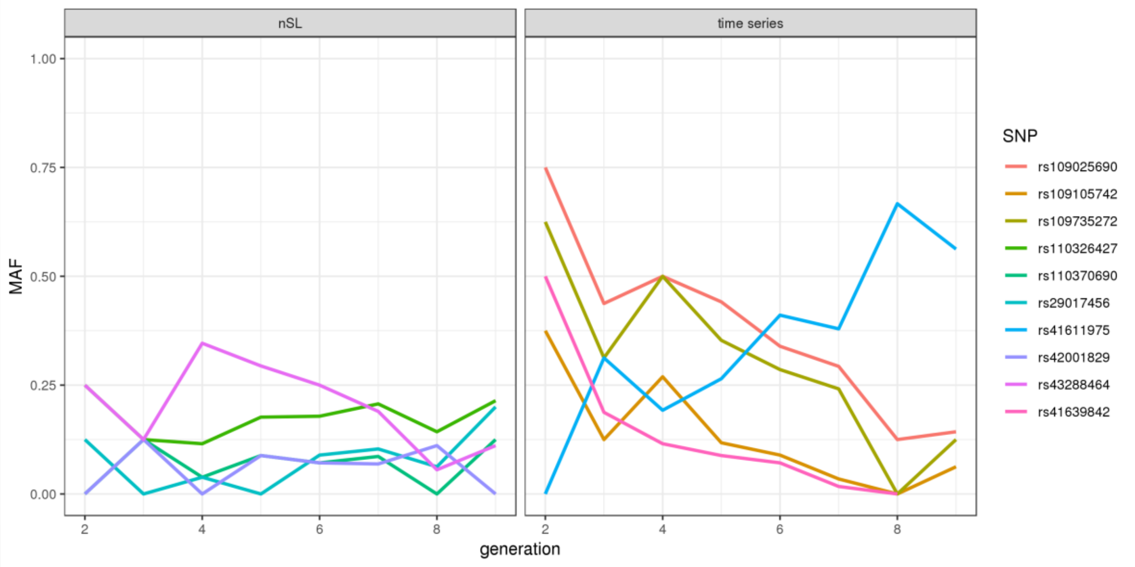
Comparison between approaches

- **Almost no correlation between the p-values** obtained from the time series and the nSL approach.



Comparison between approaches

- **Top 5 SNPs** of the **time series approach** (right): **clear shift in allele frequency since 1980**.
- **Top 5 nSL SNPs** (left): **stable allele frequency since 1980**, suggesting selection at these loci is older.



Conclusions

- **Time series** approach **complimentary** to standard **present time sampling** approach:
  - 5 new loci detected when using LD information.
  - Better understanding of selection history at nSL candidate loci, e.g. MSTN.
- **Relevant candidate genes** identified with the two approaches, related to **meat or milk production, immunity, olfaction** . . .
- Time series material stored in **gene banks** can provide **rich information about breed history and biology**.

Ackowledgements

- **IMAGE project**, European Union's Horizon 2020 Research and Innovation Programme, grant agreement no. 677353
- **Haptitude project**, INRA métagrogramme SelGen
- **Kenza Bazi-Kabbaj & Maria Bernard** (INRA GABI)
- **Genotoul bioinformatics platform** Toulouse Midi-Pyrénées

Detection power

- **No single SNP** under recent selection significantly **detected** in Asturiana:
  - **experimental design:** sample sizes, number of time points?
  - **biological reasons**: low effective population size, weak selection apart from double muscling?
  - Computer simulations: selection can be detected if selection intensity above 0.2.
  - **SNP density.**