



Functional information embedded in the unmapped short reads of whole-genome sequencing



Guilherme Neumann¹, Paula Korkuć¹, Monika Reißmann¹, Manuel Wolf², Katharina May², Sven König², Gudrun Brockmann¹

¹Humboldt-Universität zu Berlin, ²Justus-Liebig-Universität Gießen

guilherme.neumann@hu-berlin.de



German Black Pied cattle – Deutsches Schwarzbuntes Niederungsrind (DSN)



- Taurine **dual-purpose** breed
- **Ancestral** population of **Holstein**
- Population around **2,500 animals***
- **Endangered**



Schulze, W. (n.d.). Retrieved February 25, 2021, from <https://www.rind-schwein.de>

*Data from the Central Documentation of Animal Genetic Resources in Germany (TGRDEU), 2022



- **304 sequenced DSN**

Neumann, G. B. *et al.* Design and performance of a bovine 200 k SNP chip developed for endangered German Black Pied cattle (DSN). *BMC Genomics*. 2021.

Neumann, G. B. *et al.* Genomic diversity and relationship analyses of endangered German Black Pied cattle (DSN) to 68 other taurine breeds based on whole-genome sequencing. *Frontiers in Genetics*. 2023.

Know more at 4:30 PM – session 83:
Which diversity measures are important for small
breeds like German Black Pied Cattle (DSN)?
Brockmann, G.A.

- On average 204Mio short reads (150 PE Novaseq 15x coverage) sequenced per animal
- 2Mio reads generally left unmapped (after filtering) – from 144,848 to 40,734,422

Unmapped reads can cover parasites and structural variation



- **Unmapped reads are discarded**, but significant **biological information** and insights **could be uncovered**
- Unmapped reads from DNA and RNA checked for a bird species:
 - found **blood parasites** such as *Plasmodium* and *Trypanosoma*
 - contigs blasted to other close bird species (difference to reference genome)



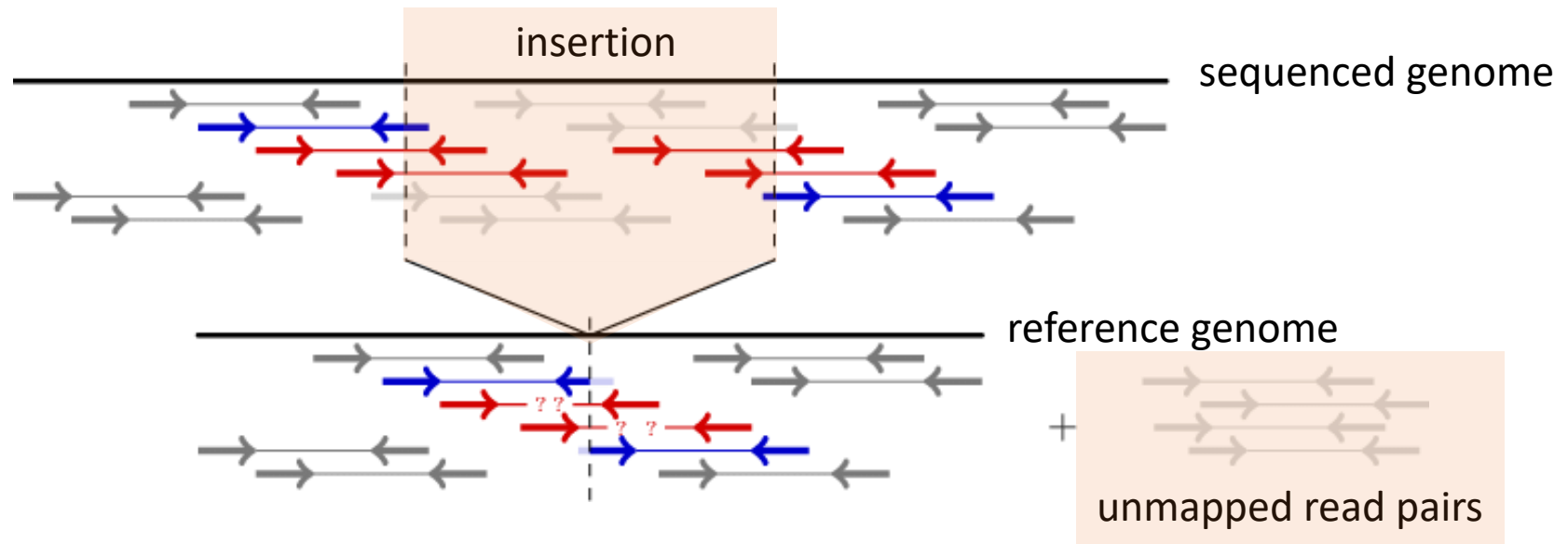
Laine VN, Gossmann TI, Van Oers K, Visser ME, Groenen MAM. Exploring the unmapped DNA and RNA reads in a songbird genome. BMC Genomics. 2019.

- **Missed indels** detected on unmaped reads

Hasan MS, Wu X, Zhang L. Uncovering missed indels by leveraging unmapped reads. Sci Reports. 2019.



- **Unmapped reads are usually discarded**, but might contain significant **biological information**



Retrieved from Kehr B., Melsted P., and Halldorsson B. (2015) *Bioinformatics*.



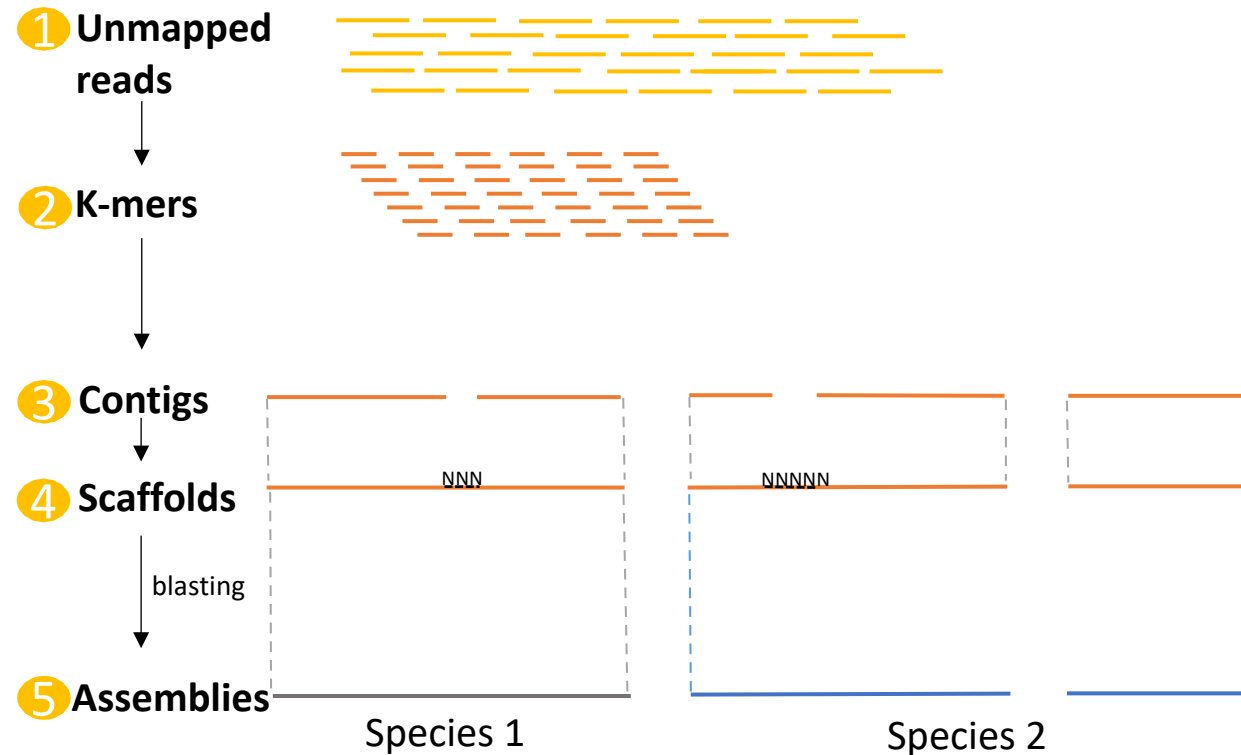
- **Uncover hidden sequences, either from external DNA or from DSN, which do not match the Bovine reference genome**
 - Pathogens' DNA
 - Structural variants



Goal 1 – Detection of pathogens



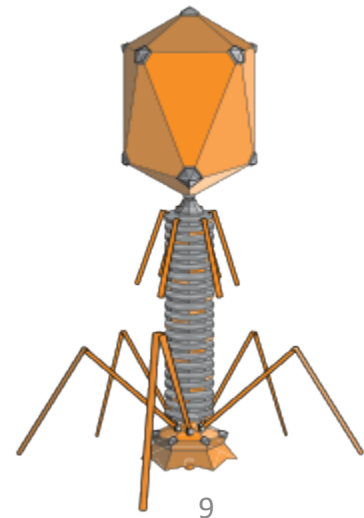
- **Whole-genome sequencing data:**
 - All **unmapped reads**
- **De-novo genome assembly**
 - Abyss 1.5.2 using k-mer size = 90
- **Blastn** to the **representative reference genomes database** from **NCBI**
 - Scaffolds from the same species and animal combined as assembly
 - Only assemblies covering >10% of their reference genomes



- 29 out of 304 DSN animals detected with sequences of 6 viruses :
 - **Bacteriophages** : *Stenotrophomonas phages SMA7* (3 animals)
 - **Plant viruses** : *Helicoverpa armigera densovirus* (1 animal), *Brassica yellows virus* (4 animals)
 - **Insect viruses** : *Deformed wing virus* in bees (16 animals)
 - **Mammals viruses** : ***Bovine parvovirus 3*** (4 animals), ***PreXMRV-1*** (1 animal)

***Bovine parvovirus 3* causes diarrhea in neonatal calves and respiratory and reproductive disease in adult cattle**

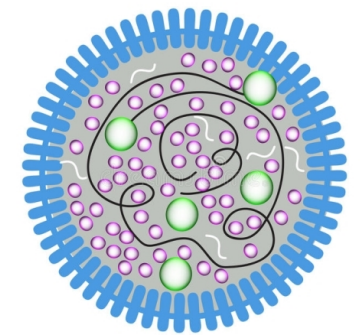
Genome coverage from 97% to 100% with depth from 6 to 41-fold



- 89 out of 304 DSN animals detected with sequences of **11 bacterial species**:
 - **Soil**: *Achromobacter insuavis* (37 animals), *Variovorax gossypii* (18 animals), *Bosea lupini* (9 animals), *Delftia acidovorans* (1 animal)
 - **Water**: *Roseateles aquatilis* (12 animals)
 - **Beef and milk containers**: *Pseudomonas paracarnis* (3 animals), *Pseudomonas carnis* (1 animal), *Pseudomonas lactis* (1 animal), and *Pseudomonas salmasensis* (1 animal)
 - **Pathogens**: ***Mycoplasma wenyonii* str. Massachusetts** (4 animals), ***Candidatus Mycoplasma haemobos*** (2 animals)

***Mycoplasma* can cause mastitis and anemia in cattle**

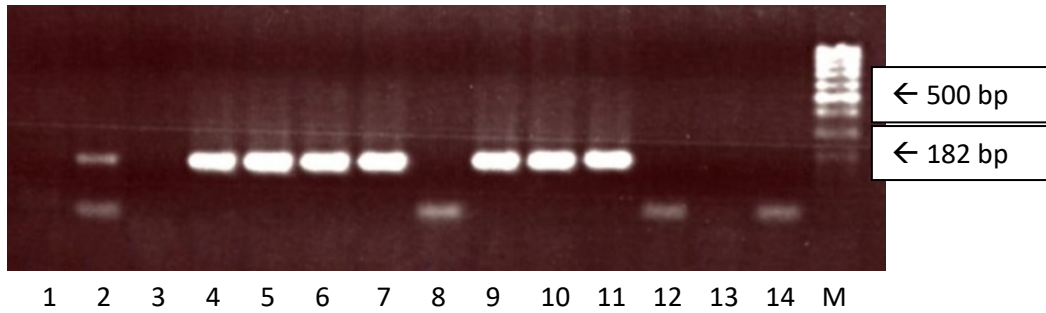
Genome coverage from 11% to 71% with depth from 2 to 1,768-fold



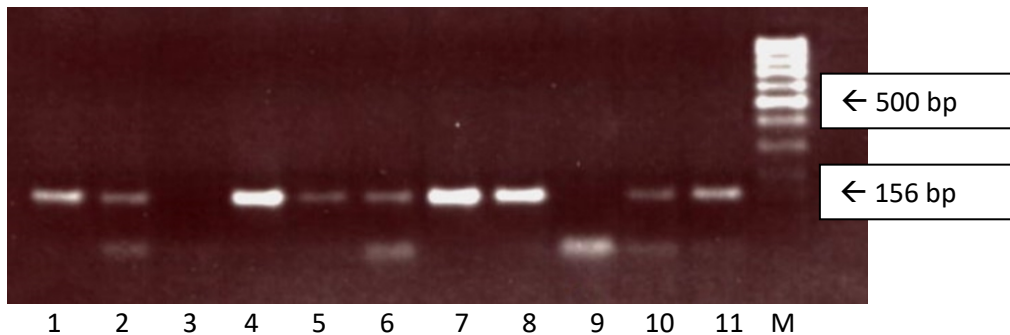
PCR - Validation of results



- Primers for *Mycoplasma wenyonii*, expected fragment length of 182 bp



- Primers for *Candidatus Mycoplasma haemobos*, expected fragment length of 156 bp

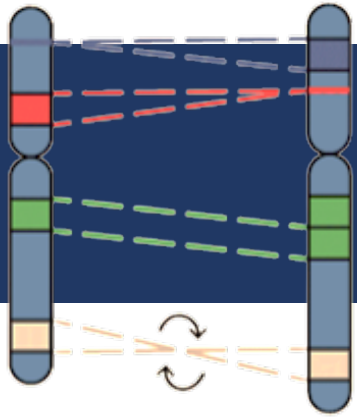


POS	Code	WGS detection	DNA replicate	Tissue
1	BU131	Genome coverage: 34%, depth: 4.1	A	ear
2			A	blood
3	empty	-	-	-
4	Bu131	Genome coverage: 34%, depth: 4.1	B	blood
5	Bu118	Genome coverage: 28%, depth: 27.2	A	blood
6			B	blood
7			C	blood
8	Bu139	Genome coverage: 27%, depth: 1768.8	A	sperm, blood not available
9	Bu120	Genome coverage: 28%, depth: 161.8	A	blood
10			B	blood
11			C	blood
12	Bu121	-	A	blood, control
13	Bu122	-	A	blood, control
14	Bu123	-	A	blood, control

POS	Code	WGS detection	DNA replicate	Tissue
1	BU131	Genome coverage: 11%, depth: 2.3	A	ear
2			A	blood
3	empty	-	-	-
4	Bu131	Genome coverage: 11%, depth: 2.3	B	blood
5	Bu126	Genome coverage: 72%, depth: 4.9	A	ear
6			A	blood
7			B	blood
8			C	blood
9	Bu121	-	A	blood, control
10	Bu122	-	A	blood, control
11	Bu123	-	A	blood, control



- No assemblies >10% of their reference genomes
- But assemblies from different *Bos* species, which covers ~0.01% of the cattle genome
- Structural variants?



Goal 2 – Detection of structural variants



- **All 304 DSN with short-read data:**

- SvABA 1.1.3:

- **unmapped, clipped, and discordant reads**

- *De-novo* genome assembly

- Delly 1.1.5:

- **mapped, clipped, and discordant reads**

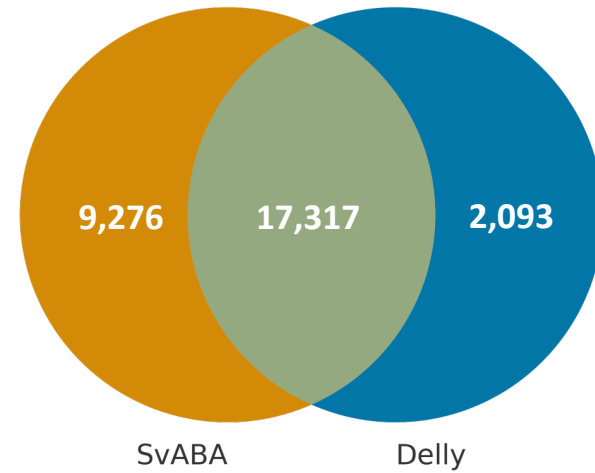
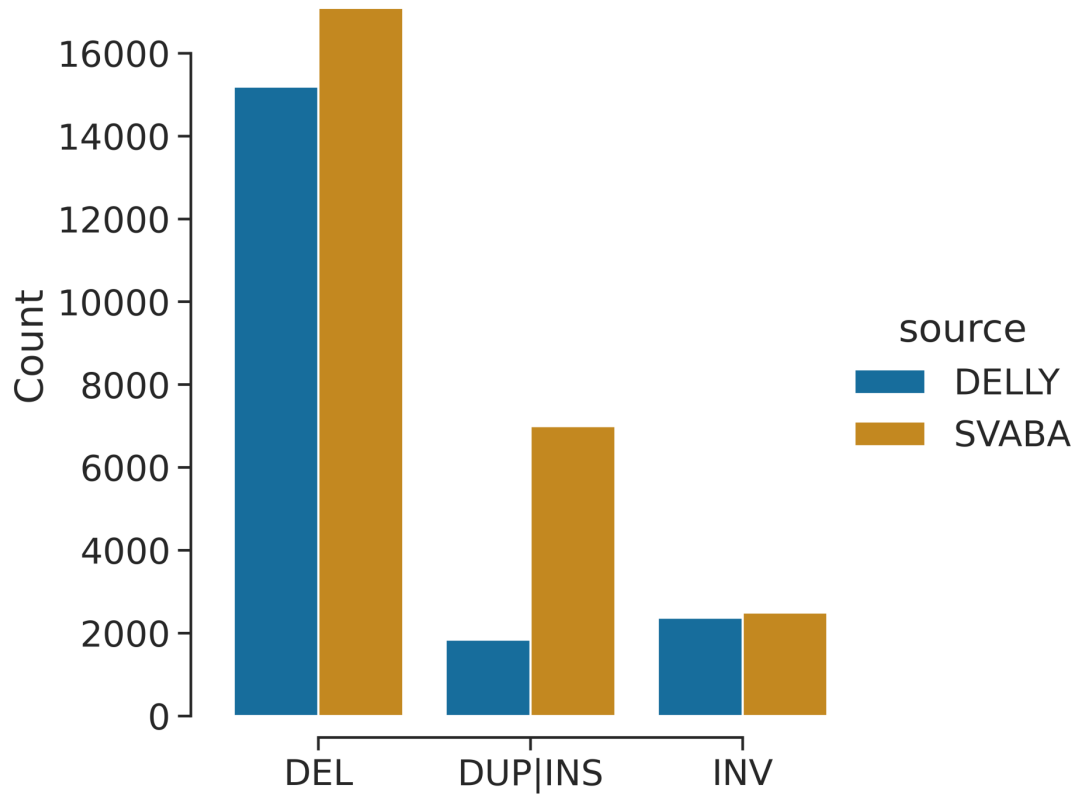
- paired-end mapping, read-depth, and split-read analysis

- **4 DSN with long-read (Hifi) data:**

- Sniffles 2.2, cuteSV 2.0.3, dysgu 1.5 → consensus (two out of three) using SURVIVOR 1.0.7

Structural variants based on short reads

- 19,410 SVs were detected using Delly (**mapped, clipped, and discordant reads**)
- 26,593 SVs using SvABA (**unmapped, clipped, and discordant reads**)
- 32% of *Bos* scaffolds mapped to contigs from SvABA

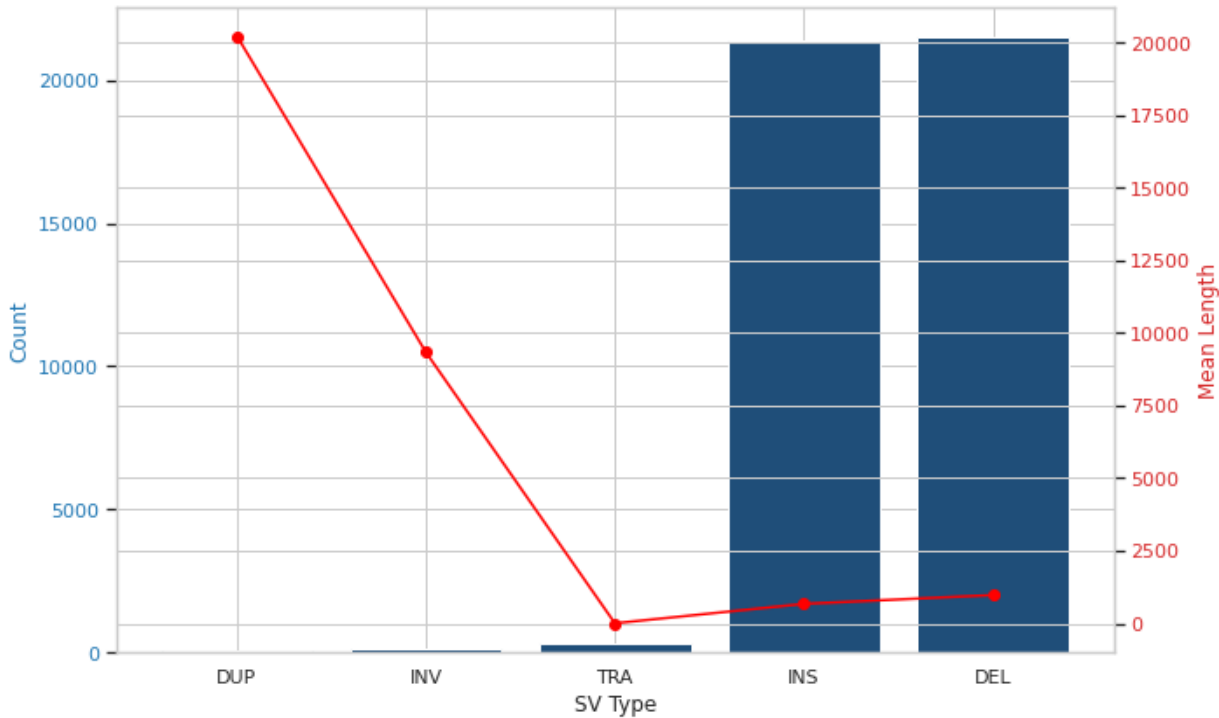


790 SVs were exact matches
(start and end positions and SV type)

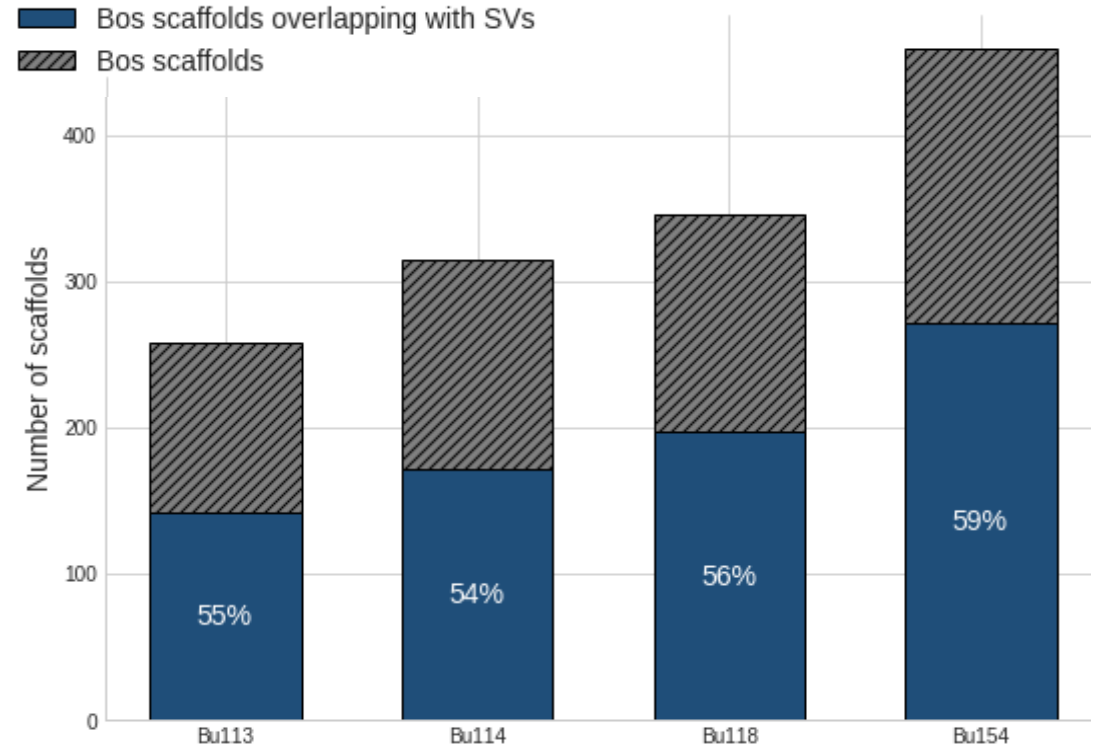
Structural variants based on long reads



- 43,421 SVs were detected in four DSN in at least two out of three software



~35% of SvABA insertions match insertions from long reads



~55% of scaffolds mapped to *Bos* species match SVs detected with long reads

Conclusion



- Very likely occurrence of *Mycoplasma wenyonii* and *Candidatus Mycoplasma haemobos* in DSN detected on the unmapped reads
- Metagenomics is recommended for detecting pathogens, but analysis of **unmapped short reads** could provide **evidence for infections** if other data is lacking
- **Bos scaffolds retrieved from unmapped reads contain structural variants** detected by long-read sequencing

Thank you for your attention!



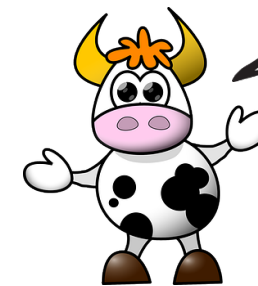
Humboldt-Universität zu Berlin

Gudrun Brockmann
Paula Korkuć
Uwe Müller
Monika Reißmann



Justus-Liebig-Universität Gießen

Sven König
Katharina May
Manuel Wolf



Questions?

RBB Rinderproduktion Berlin-Brandenburg GmbH

Cornelia Buchholz
Maria Thiele

DSN Breeders

Gefördert durch



Bundesministerium
für Ernährung
und Landwirtschaft

Projektträger



Bundesanstalt für
Landwirtschaft und Ernährung

aufgrund eines Beschlusses
des Deutschen Bundestages