HapchIQ

Predicting carrier status for genetic conditions with machine learning

Nathan C. Blair, Daniel J. Null, Ezequiel L. Nicolazzi, and John B. Cole*

Council on Dairy Cattle Breeding
Bowie, Maryland, USA

Topics

- Background
- Brown Swiss Haplotype 6
- Brown Swiss, Holstein, and Jersey Polled
- Other haplotypes
- Discussion and conclusions

Background

Background & Terminology

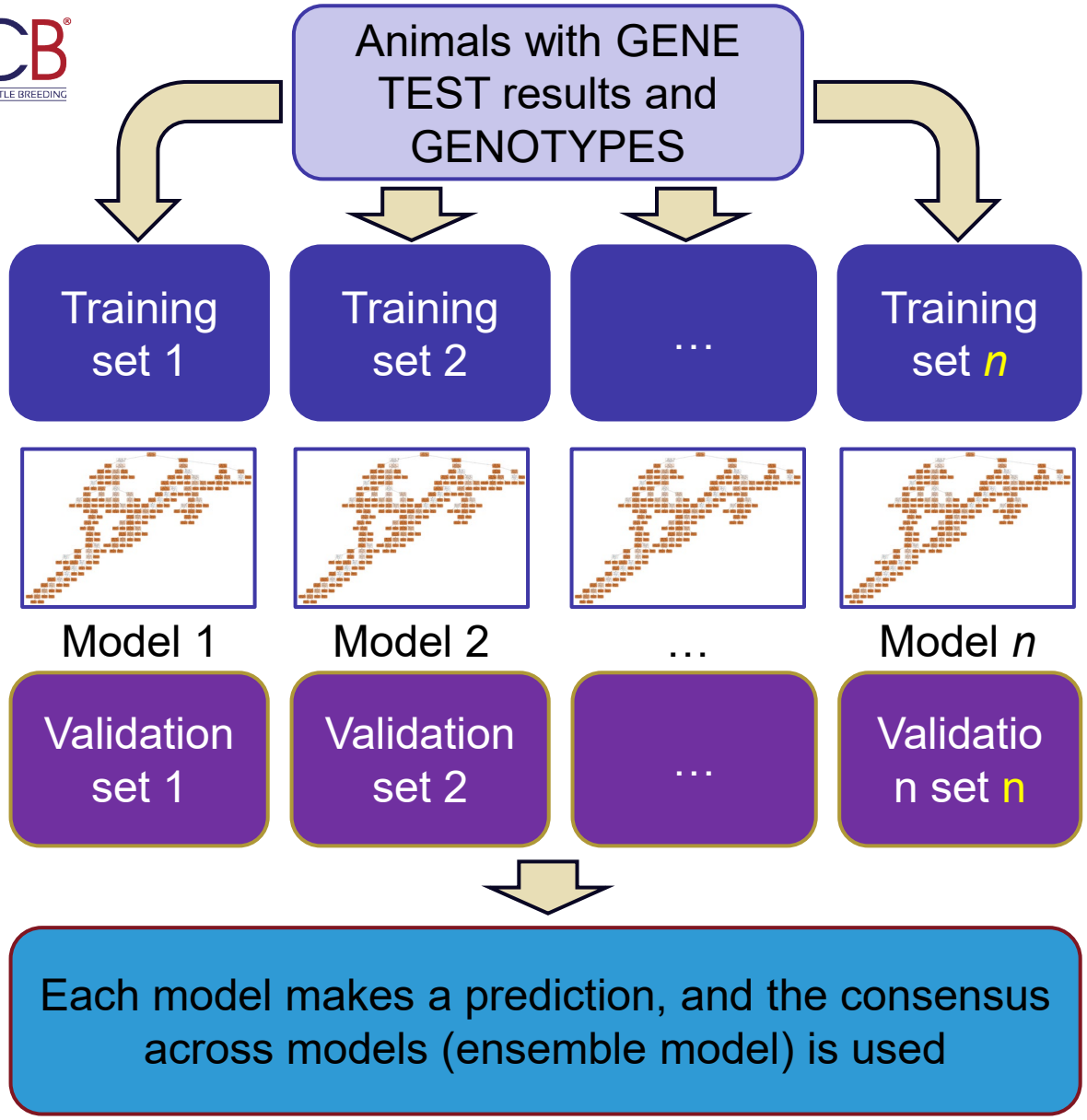
- CDCB currently uses a software package called **Hapcheck** to assign haplotype carrier status using an animal's genotype.
  - Uses the methods described in VanRaden et al. (2011).
- We have recently built a new tool named **HapchIQ** (pronounced "hap-sheek") to improve on **Hapcheck**.
- **HapchIQ** uses a machine-learning approach to assign haplotype carrier status to genotyped animals.

Hapcheck

- Uses genotypes and pedigree information to determine haplotype carrier status
- Sensitive to segmentation in the input data
- Predictions for new haplotypes typically require SNP list updates and re-imputation of all genotypes
- Serial in nature – haplotypes are assigned sequentially
- Many special cases have accumulated in the code and are difficult to validate as the number of genotypes increases

Decision Trees

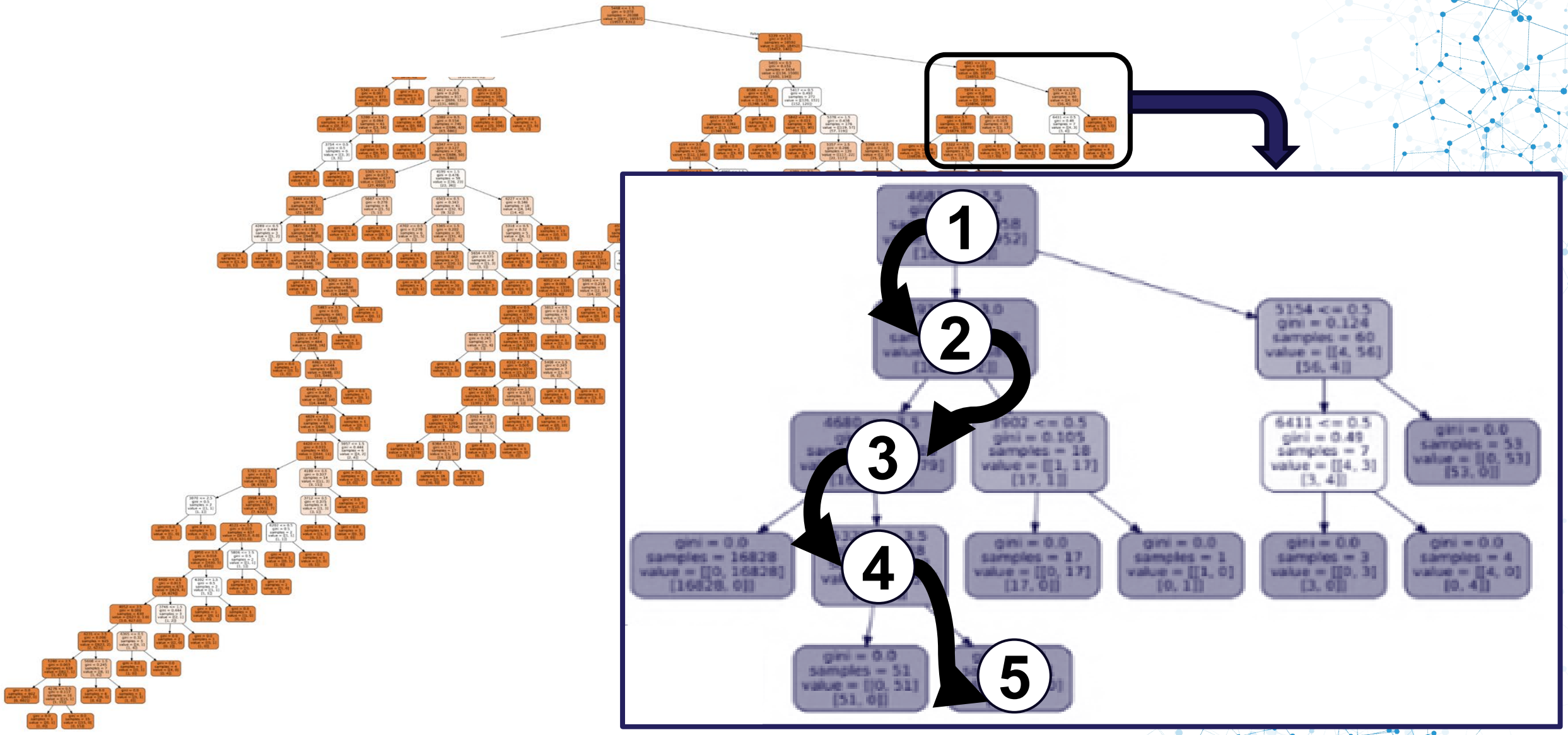
- We developed a gradient boosted tree-based machine learning model using XGBoost.
  - **HapchIQ** is coded in Python (XGBoost 3.1.2, SciKit Learn, and Polars).
  - The **HapchIQ** code base is ~1/2 the size of **Hapcheck**, making maintenance and development easier.
- Decision trees were chosen because they work very well with structured, tabular data, of which we have a lot.
- We also had some success with 2-dimensional convolutional neural networks, but they were more resource intensive.



File of predicted haplotype carrier status for individual animals with genotypes

HOUSA000000000001 CARRIER
HOUSA000000000002 NON-CARRIER
HOUSA000000000003 NON-CARRIER
HOUSA000000000004 NON-CARRIER
HOUSA000000000005 NON-CARRIER
HOUSA000000000006 CARRIER
…
HOUSA000000*nnnnnn* NON-CARRIER

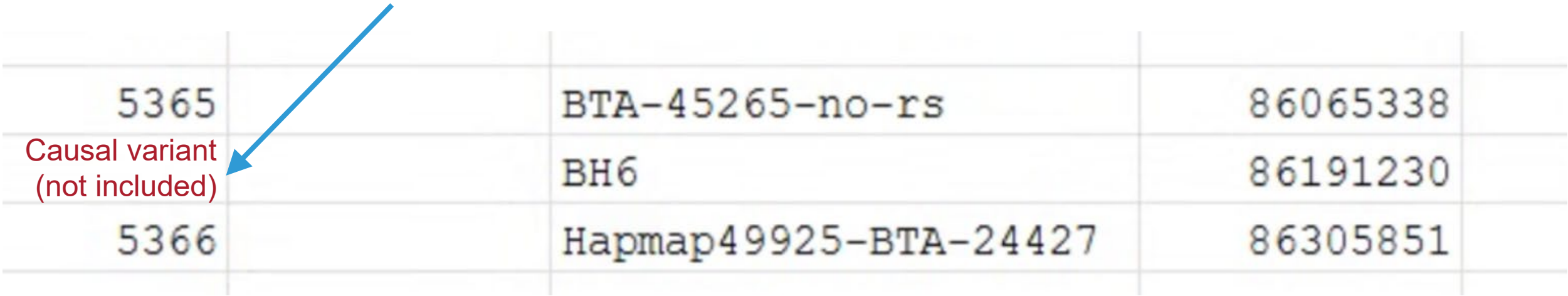
Animals with GENOTYPES but no GENE TEST results


1
2
3
4
5
5154 <= 0.5
gini = 0.124
samples = 60
value = [[4, 56]
[56, 4]]
6411 <= 0.5
gini = 0.49
samples = 7
value = [[4, 3]
[3, 4]]
gini = 0.0
samples = 53
value = [[0, 53]
[53, 0]]
gini = 0.0
samples = 16828
value = [[0, 16828]
[16828, 0]]
gini = 0.0
samples = 17
value = [[0, 17]
[17, 0]]
gini = 0.0
samples = 1
value = [[1, 0]
[0, 1]]
gini = 0.0
samples = 3
value = [[0, 3]
[3, 0]]
gini = 0.0
samples = 4
value = [[4, 0]
[0, 4]]
gini = 0.0
samples = 51
value = [[0, 51]
[51, 0]]

Brown Swiss Haplotype 6

Brown Swiss Haplotype 6 (BH6)

- Affected embryos die, causing fertility reductions.
- The causal allele is located on chromosome 2.
- It's not in the SNP list, but we can still detect it via linkage disequilibrium because it falls in between 2 SNPs we do track.



Concordance of BH6 calls from Hapcheck and HapchIQ with gene test results

Program	SNP list	Concordance
Hapcheck	Official	46.5%
Hapcheck	Modified[1]	92.2%
HapchIQ	Official	90.7%

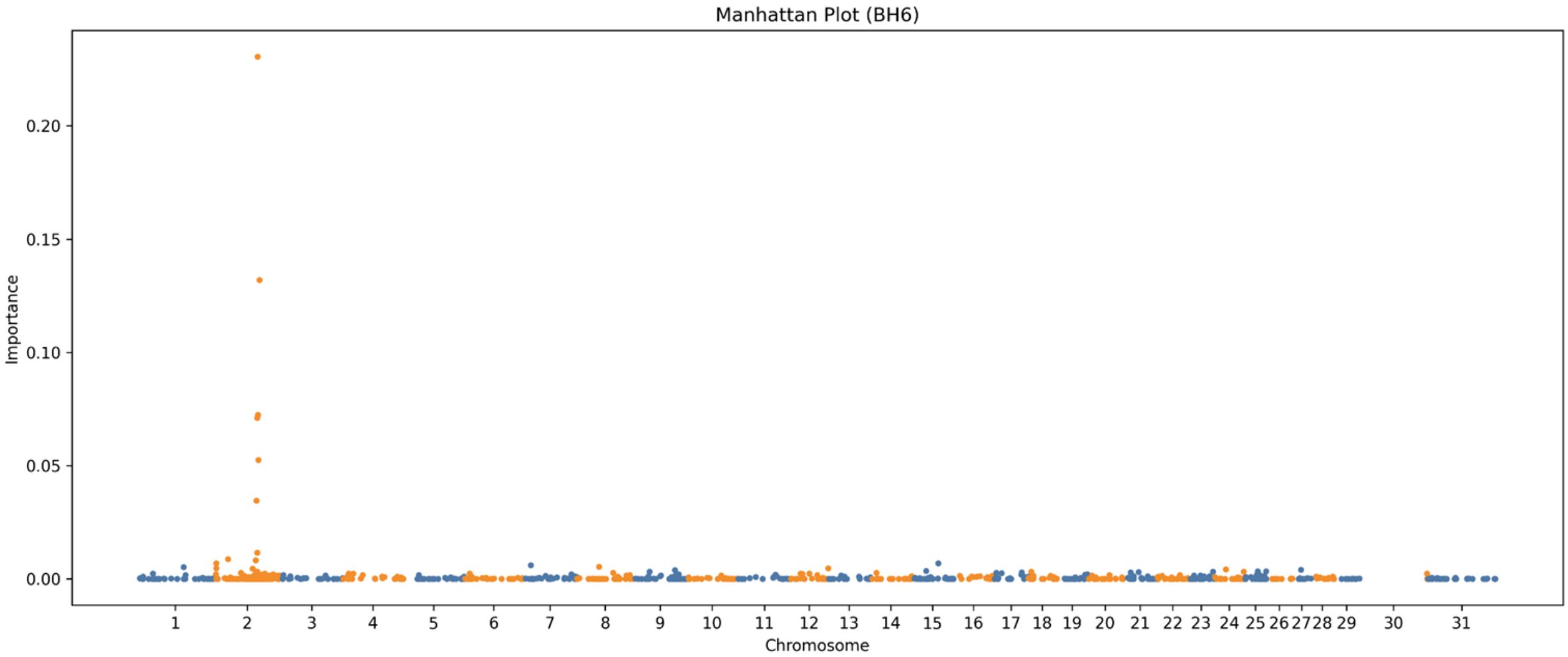
[1]Addition of new DNA markers to the official SNP list requires that all genotypes be reimputed. This was done only for Brown Swiss animals for internal testing.

Run-to-run variation versus training set-to-training set variation

					Changes in status	
Training Set 1	**Evaluation Run 1**	**Training Set 2**	**Evaluation Run 2**	**Number of animals**	**Non-carrier to carrier**	**Carrier to non-carrier**
December 2024	December 2024	—	April 2025	82,553	3	0
December 2024	April 2025	—	August 2025	83,829	1	1
December 2024	August 2025	December 2025	December 2025	90,062	78	46

Comparison of Hapcheck and HapchIQ results for BH6

		HapchIQ (research)	
		Non-carrier	**Carrier**
Hapcheck (official)	**Non-carrier**	87,932	32
	Carrier	51	4,656
	Homozygous	0	2


Manhattan Plot (BH6)
Importance
0.20
0.15
0.10
0.05
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Chromosome

Brown Swiss, Holstein, and Jersey Polled

Challenges with polled calls

- Multiple mutations produce the polled phenotype.
  - More than one variant is segregating in some breeds.
  - Many laboratory gene tests include only the Friesian mutation.
- **Hapcheck** works well for Holstein but not for Brown Swiss and Jersey.
- **HapchIQ** does much better with the current SNP list.
- When additional polled SNP are added, the two methods are comparable.

Breed	SNP list	Program	Matches	Mismatches	Check	Concordance
BS	Official	Hapcheck	17,679	3,047	20,726	85.3%
BS	Official	SNP	16,531	2,949	19,480	84.9%
BS	**Official**	**Hapchiq**	**20,697**	**29**	**20,726**	**99.9%**
BS	SNP name fix	Hapcheck	20,661	65	20,726	99.7%
BS	SNP name fix	SNP	20,721	5	20,726	100.0%
BS	SNP name fix	Hapchiq	20,723	3	20,726	100.0%
JE	Official	Hapcheck	17,559	9,928	27,487	63.9%
JE	Official	SNP	17,643	9,671	27,314	64.6%
JE	**Official**	**Hapchiq**	**27,231**	**256**	**27,487**	**99.1%**
JE	SNP name fix	Hapcheck	26,784	671	27,455	97.6%
JE	SNP name fix	SNP	27,452	33	27,485	99.9%
JE	SNP name fix	Hapchiq	27,470	15	27,485	99.9%
HO	Official	Hapcheck	55,943	2,714	58,657	95.4%
HO	Official	SNP	55,929	2,667	58,596	95.4%
HO	**Official**	**hapchiq**	**58,248**	**409**	**58,657**	**99.3%**
HO	SNP name fix	Hapcheck	58,079	577	58,656	99.0%
HO	SNP name fix	SNP	58,632	24	58,656	100.0%
HO	SNP name fix	Hapchiq	58,590	66	58,656	99.9%

Other haplotypes

Comparison of Hapcheck and HapchIQ results for Holstein from August 2025

Haplotype	Mismatches	Genotypes	Concordance
Holstein Haplotype 1	1,233	9,088,709	99.99%
Holstein Haplotype 2	5,963	9,089,160	99.93%
Holstein Haplotype 5	2,170	9,088,709	99.98%
Holstein Cholesterol Deficiency	44,410	8,914,901	99.50%
Holstein Muscle Weakness	67,024	8,173,668	99.18%

Comparison of Hapcheck and HapchIQ results for other breeds from August 2025

Haplotype	Mismatches	Genotypes	Concordance
Ayrshire Haplotype 2	109	20,448	99.47%
Ayrshire Arthrogryposis Multiplex Congenita	48	20,448	99.77%
Brown Swiss Weaver Syndrome	101	90,081	99.89%
Brown Swiss Spinal Dysmyelination	136	90,081	99.85%
Brown Swiss Haplotype 14	104	90,081	99.88%
Jersey Neuropathy with Splayed Forelimbs	479	930,718	99.95%

Discussion and conclusions

Opportunities for improvement

- Training is very sensitive to class imbalances
  - 2s for many haplotypes are infrequent relative to 0s and 1s, so the algorithm discards them as likely genotyping/imputation errors.
- Training can be computationally intensive
  - We limit input data to 1 or 2 chromosomes, but it's not required.
  - GPUs make a significant difference here but can be more expensive.
- Research is underway to see if pedigree information can be incorporated into **HapchIQ** predictions
- Nodes (rules) in the decision tree can be difficult to interpret

Opportunities for improvement (cont'd)

- Data engineering is 80% of the work
  - Large flat-files of genotypes are hard to parse efficiently and within reasonable amounts of RAM.
- Holstein Muscle Weakness is not performing as well
  - **HapchIQ** may not work as well for haplotypes with incomplete penetrance.
  - We're exploring multi-chromosome inputs and hyperparameter tuning to explore this.

Conclusions

- **HapchIQ** predictions of carrier status perform as well as **Hapcheck** in some cases, and better in others
- A separate **HapchIQ** model is produced for each haplotype, and predictions can run in parallel
- **HapchIQ** can provide a measure of certainty with each prediction, eliminating the need for “3” and “4” calls
- CDCB will distribute test files beginning in August, with a goal of implementation for BH6 and POLLED in December 2026.

Questions?

Thank you for your attention!

john.cole@uscdcb.com

"Yes … I believe there's a question there in the back."