Improving SNP Chip Validation and Genomic Data Quality at CDCB

Clarissa Boschiero, Ph.D.
Bioinformatics Scientist

2026 CDCB Genomic Nominator and Laboratory Workshop

Presentation Outline

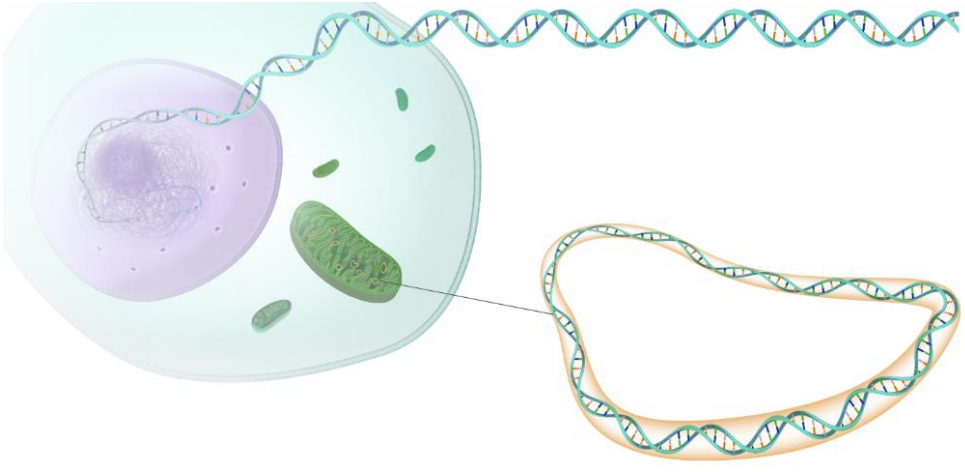
1. Background & CDCB role
2. Cattle genome assembly
3. SNP chip validation process modernization
   I. New technologies
   II. Visualization plots
   III. Verification of SNPs coordinates using flanking sequences
4. SNP usability
5. Key takeaways

Background & CDCB Role

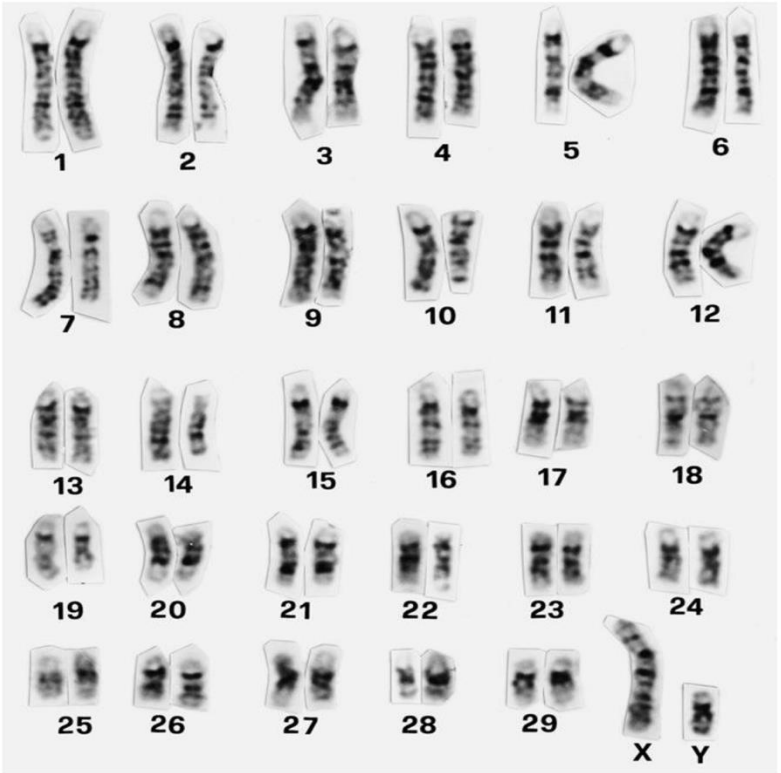
- I have been working at CDCB since December 2024
- I work on 2 teams – the Data Team & R&D
- Sequencing data, identification of genetic variants, bioinformatics and genomics analyses
- SNP chip validation (Heather Enzenauer, George Wiggans)
- SNP usability revision (Heather Enzenauer, Dan Null)
- Identification of genetic variants from sequencing data (John Cole, R&D)

Genome

- A genome is the complete set of genetic material in an organism, containing all the information needed to develop and function.
- In cattle, the genome consists of 30 pairs of chromosomes (29 autosomes and X/Y) located in the cell's nucleus.



https://www.genome.gov/genetics-glossary/Genome

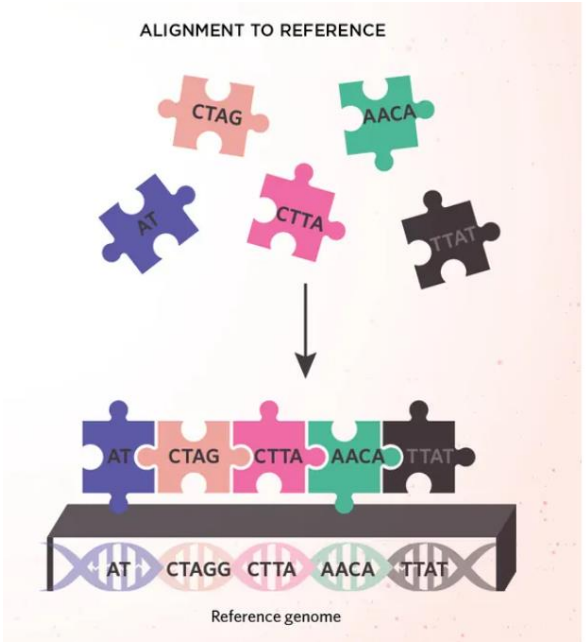


Cattle karyotype (Cribiu et al., 2001).

Reference Genome Importance

A reference genome assembly is a high-quality digital map of a species' DNA that scientists use as a baseline to study and compare individuals of the same species.

Reference genome

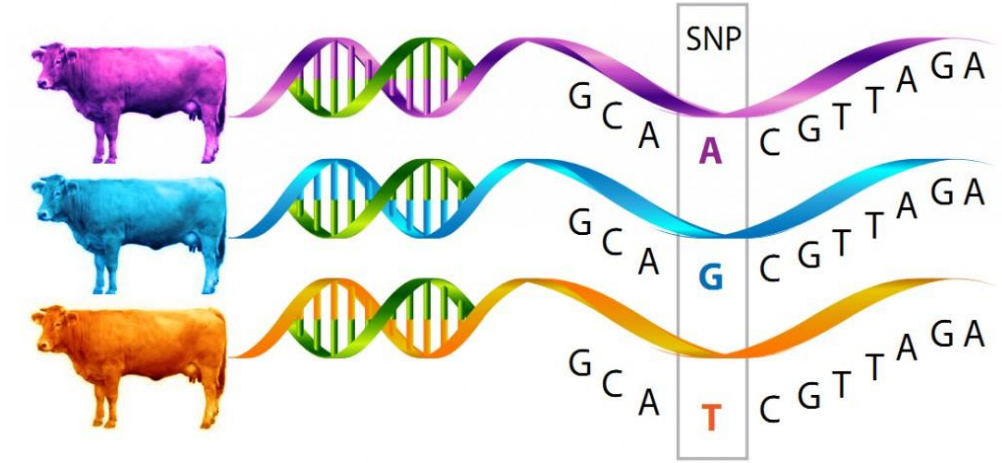


Cattle genome reference file

```
>1 dna:primary_assembly primary_assembly:ARS-UCD1.2:1:1:158534
GTACACTGATCACGTGGCTGATCATGCACAAATCCCATTGCATATCATGCTCTGATCAGG
TTGCTATCATGTACTGATCACTTGGCTGATCATACACTGATCACATGACTGATCATGCAC
TGATCACGTGCCTGATCATGCACTGATCCCGTGGCAGATCATGCACTGATCACGTGCAGA
TCATGCACTCATCATGTGGCTGATCATACAATGATCACGTGGCTGATCATGCACTGATCA
CATGTATGATCGTACACTGATCACGTGGCTGATCATGCACAAATCCCATTGCATATTCAT
GCTCTGATCAGGTTGCTATCATGTACTGATCACGTGGCTGATCATACACTGATCACAATG
ACTGATCATGCACTGATCACGTGCCTGATCATGCACTGATCCCGTGGGGCTGATCTTGCA
CTGATCACGTGGCTGATAATGGCACTGATCACTTGACTGATCATGCACTGATCACGTCTC
TTATCATGCACTACTCACGTGACTGATCATGGACTGATCACGTGACTGATCATGCACTGA
TCATGAGCCTGATCATGCACTGATCACGGGCCTGTTCATGCACTGATCACGTGGCTGATC
CTGCACTGATCACGAGGCTGATCATACACTGATCACGTGACTGTGCATGCAATGATCACT
TGGCTGATCATGTACTGATCACGTGGCTGATCATGCACTGATCACATGACTGAGTTGCAC
TGATCACATAGCTATCATGCACTGATCACGTGACTTCGCATGCCATGATCACGTGACTGA
TCATGCACTGGTCCCGTGGCTATCATGCACTGATAATGTGGCTTATCATACATTGATCAC
CTAAATAATCATGCCACTGATCACGTGGCAATCTTGCTCTGATCACGTGGTTGATCAGGC
ACTGATCACTTAGCTATCATGCACTGATCACATGAACTGCGCAAGAACTGATCACTTAGA
TATCATACACAGGACACGTGACAGATCATGCACTGAATCACGTGACTGATCATGCACTGA
TCTGGTGGCTATCATGCACTGATCCCGTGGCTGATCTTGCACTGATCACGTGGCTTGATA
ATGCACTGATCACTTGACTGATCATGCACTGATCACGTCTCTTATCATGCCACTACTCAC
GTGACTGATCATGGACTGATCACGTGACTGATCATGCACTGATCATGAGCCCTGATCATG
CACTGATCACGGGCCTGTTCATGCACTGATCACGTGGCTGATCCTGCACTTGATCACGAG
GCTGATCATACACTGATCACGTGACTGTGCATGCAATGATCACTTGGCTGATCATGTACT
GATCACGTAGCTGATCATGCTCTGATCACGTGACTGATTATGCACTGATCACATAGCTAT
CATGCACTGATCACATGACTTCGCATGCCATGATCACGTGACTGATCATGCACTGATCAC
GTGGCTATCATGCACTGATAATGTGGCTTATCATACATTGATCACCTAAATATCATGCAC
TGATCACGTGGCAATCTTGCTCTGATCACGTGGTTGATCAGGCACTGATCACTTAGCTAT
CATGCACTGATCACATGACTGCGCAAGAACTGATCACTTAGATATCATACACAGGACACG
TGACAGATCATGCACTGATCACGTGACTGATCATGCACTGATCTGGTGCTATCATGCATT
GATCACGTGGTTGATCATACACTGATCACGTGACTGATCATGCACTGATCACCTGGCTAT
CATGCACTAATCACGTAGCTGATCTGGAACAGATCACGTGACTAATCATGCAGTGATCAC
GTCGCTATCATGCAATGATCTCGAGATTGCGCATGCACTGATCATATGGATGATCACACA
CTGATCACGTGACAGATCATGCACTGAACACGTGACTTATAATGCACTGATCACGTGACT
GCACATGCACTGATCATGTGACTGAGCATGCACTGATCACGTAGCTGATCATGCACTGAT
TCACGAAGTTGATCATACACTGATCATGTGACTGAGCATGCACTGACCACGTAGCTGATC
ATGCACTGATCACGAAGTTGATCATACACTGATCACGTGAGTGTGCATGCACTGATCACG
TGGCGGATAATGCAATGATCACGTGGCTGATCATGCACTGATCACGTAGCTCTCATGCAC
TGATCACATGTCAACCATGCCCGGATCCCGTGGCTGATCATGCACTGATCGCCTGACTGA
TCAGGCACTAGTCACCTGGCTATCATGCACTAATCACGTGGCTGTTCATACACTGACCAC
GTGACTGATCATGCACTGATCTGGTGGCTATCATGCATTGATCACGTGTTTGATCTTACA
CTGATTCACGTGACTGATCATGCACTGATCACCTGGCTTTCATGCACTAATCACGTAGCT
GATCTGGAACAGATCACGTGACTAATCATGCAGTGATCACGTCGCTATCATGCAATGATC
TCGAGATTGCGCATGCACTGATCATATGGATGATCACACACTGATCACGTGACAGATCAT
GCACTGATCACGTGACTTATCATGCACTGATCACGAGACTGCACATGCACTGATCATGTG
ACTGAAGCATGCACTGATCATGTAGCTGATCATGCACTGATCACGAAGTTGATCATCACT
GATCATGTGACTGAGCATGCACTGATCACGTAGCTGATCATGCACTGATCACGAAGTTGA
TCATACACTGATCACGTGAGTGTGCATGCACTGATCACGTGGCGGATAATGCAATGTTCA
CGTAGCTCTCATGCACTGATCACATGTCAATCACGCCCGGATCCCGTGGCTGATCATGCA
CTGATCTCGAAGCTGATCATACACTGATCACGTGAGTGTGCATGCAATGATAACATGGCT
GATTATGCAATTATCACGTGGCTTATCATACACTGATCACGTGGCTGATCATGCACTAAT
CACGTTTTGGATGATGCACTGGTCACGTGGCTGATCATACACTAATCACATGACTGATCA
TGCACTAATCACGTTGTGGATCATGCACTGATCACGTGGCTGTTCTTGCACTTATCACGT
GGCTGATCATACAGTGATTACATGACTGATCATGCACTAATCACGATGTGGATCATGCAC
TGATCACATGAATGTGCATGCATTGATCATGTGGCTGATCATGCACTGATCACGTGGCTG
ATCATGCACTGATCACGAAGCTGATATTACACTGATCACAGGACTGTGCATGCACTGATC
ATGTGGCTGATCATGCACTGATCATGTGGCTGATGATACACTGATCACGTGGCTGATCGT
ACTCTGAACACGAAGCTGAAAATACACTGATCACGAGACTGCATGCACTGATCACTTGAC
```

SNPs Mapped to a Reference Genome

- A single-nucleotide polymorphism (SNP) is a difference in a unique nucleotide and is the most common type of genetic variation
- Having SNPs mapped to a reference genome means we know exactly where each marker is located, and this is essential for making genomic data comparable



https://www.icbf.com/

Cattle genome reference file

```
>1 dna:primary_assembly primary_assembly:ARS-UCD1.2:1:1:158534
GTACACTGATCACGTGGCTGATCATGCACAAATCCCATTGCATATCATGCTCTGATCACG
TTGCTATCATGTACTGATCACTTGGCTGATCATACACTGATCACATGACTGATCATGCAC
TGATCACGTGCCTGATCATGCACTGATCCCGTGGCAGATCATGCACTGATCACGTGCAGA
TCATGCACTCATCATGTGGCTGATCATACAATGATCACGTGGCTGATCATGCACTGATCA
CATGTATGATCGTACACTGATCACGTGGCTGATCATGCACAAATCCCATTGCATATTCAT
GCTCTGATCAGGTTGCTATCATGTACTGATCACGTGGCTGATCATACACTGATCACAATG
ACTGATCATGCACTGATCACGTGCCTGATCATGCACTGATCCCGTGGGGCTGATCTTGCA
CTGATCACGTGGCTGATAATGGCACTGATCACTTGACTGATCATGCACTGATCACGTCTC
TTATCATGCACTACTCACGTGACTGATCATGGACTGATCACGTGACTGATCATGCACTGA
TCATGAGCCTGATCATGCACTGATCACGGGCCTGTTCATGCACTGATCACGTGGCTGATC
```

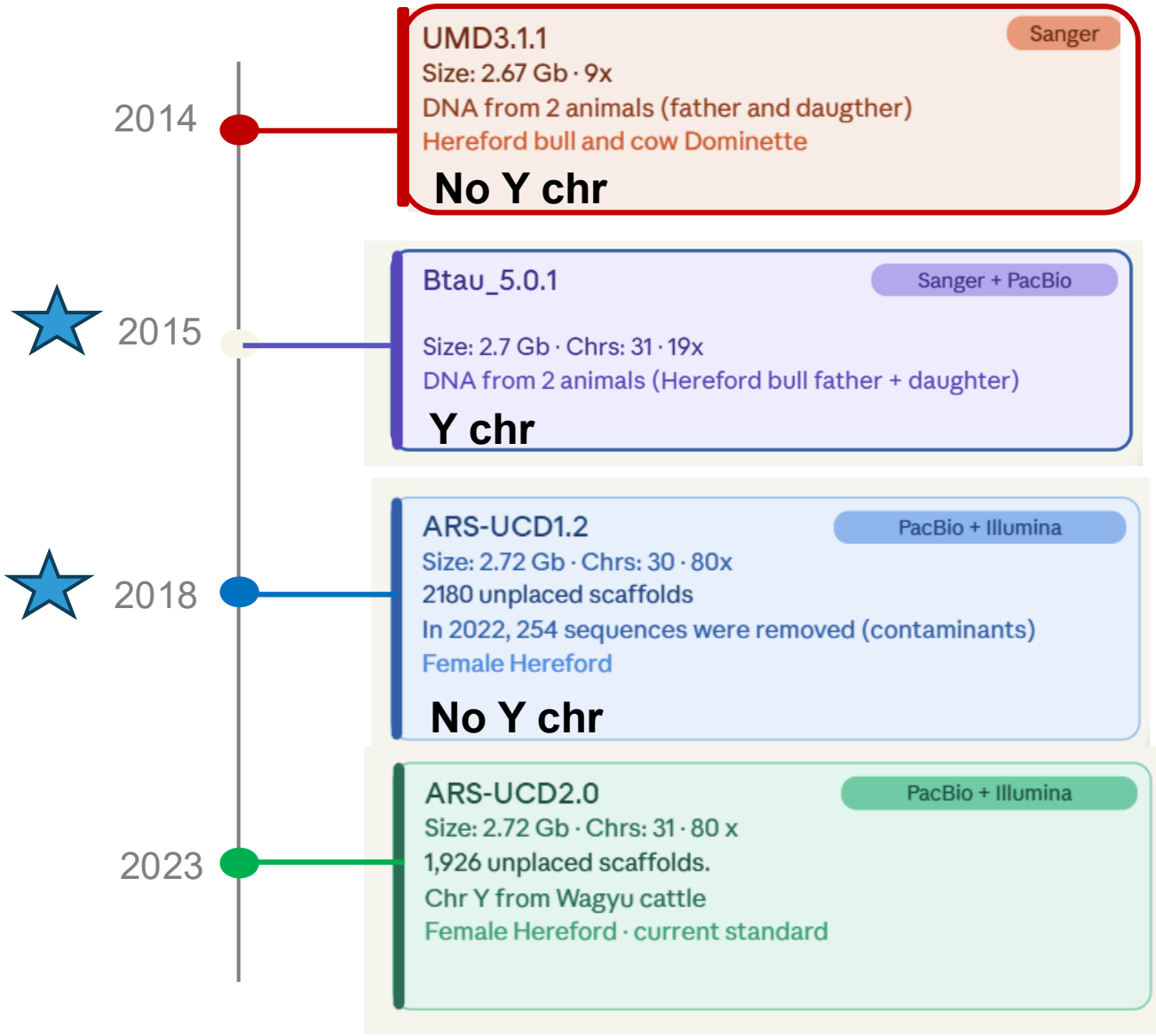
Chr1:60

Cattle Reference Genome

- The first cattle genome assembly was released in 2007 (Btau4.0)
- They sequenced a single inbred Hereford cow named L1 Dominette
- This cow was selected because of her high level of inbreeding
- The low level of genetic variation made it easier to map her genome

https://pmc.ncbi.nlm.nih.gov/articles/PMC2688908/

Timeline of Cattle Genome Assemblies



CDCB uses this assembly for Chr Y SNPs
Reminder: CDCB uses Y SNPs just for gender verification (chips are required to have at minimum 10 Y SNPs).

CDCB uses this assembly for all SNPs (except chr Y SNPs)

Same coordinates!

Cattle Genome References differences

Assembly ARS-UCD 1.2

- 2018
- Issues with 254 sequences
- **No chr Y**

Assembly ARS-UCD 2.0

- 2023
- 254 sequences were removed (located on chromosome unplaced)
- Improved chr Y from a Japanese breed (Wagyu)

These 2 assemblies have the same coordinates for chromosomes 1-29 and X!

- CDCB uses the ARS-UCD1.2 for SNPs on chr 1-29, X, and Btau_5.1 assembly for SNPs on chr Y
- CDCB plans to update the database to the latest cattle genome assembly in the future, but this is not a priority

SNP Chip Validation Process

- Our chip validation process is well established and reliable
- However, in recent years, we have seen an increase in the number of chips and new technologies such as GBS (genotype by sequencing)
- CDCB plans to modernize the chip validation process by providing more informative outputs, such as visualization reports, incorporating SNP coordinate verification, and supporting new technologies

New Technologies - GBS Panels

- Genotyping-by-sequencing (GBS) is a high-throughput sequencing method used to discover and genotype thousands of SNPs across the genome
- Diversity of GBS technologies complicates the definition and analysis of each one – Low pass sequencing (LPS) and target sequencing
- CDCB accepts GBS data

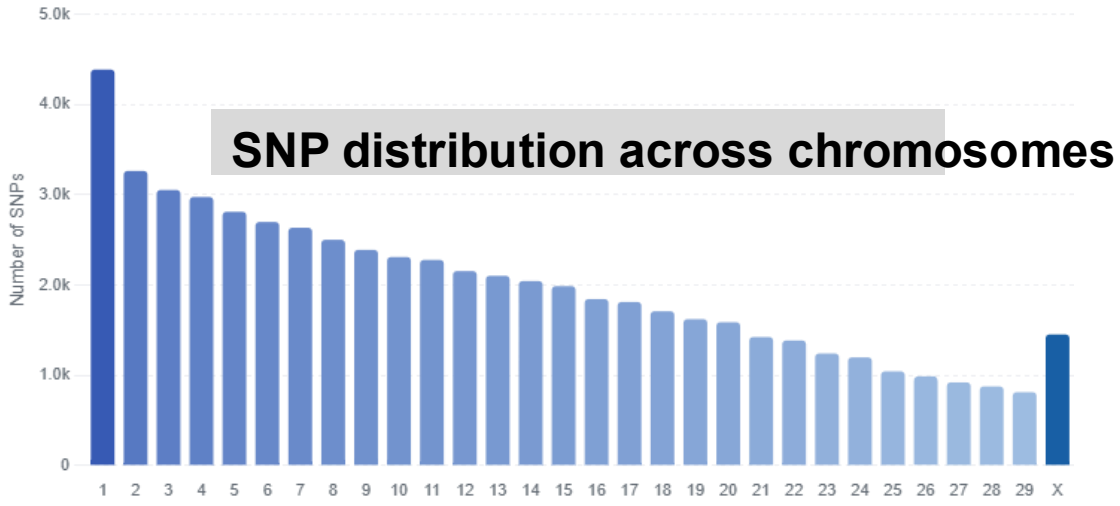
https://gencove.com/blog/low-pass-whole-genome-sequencing

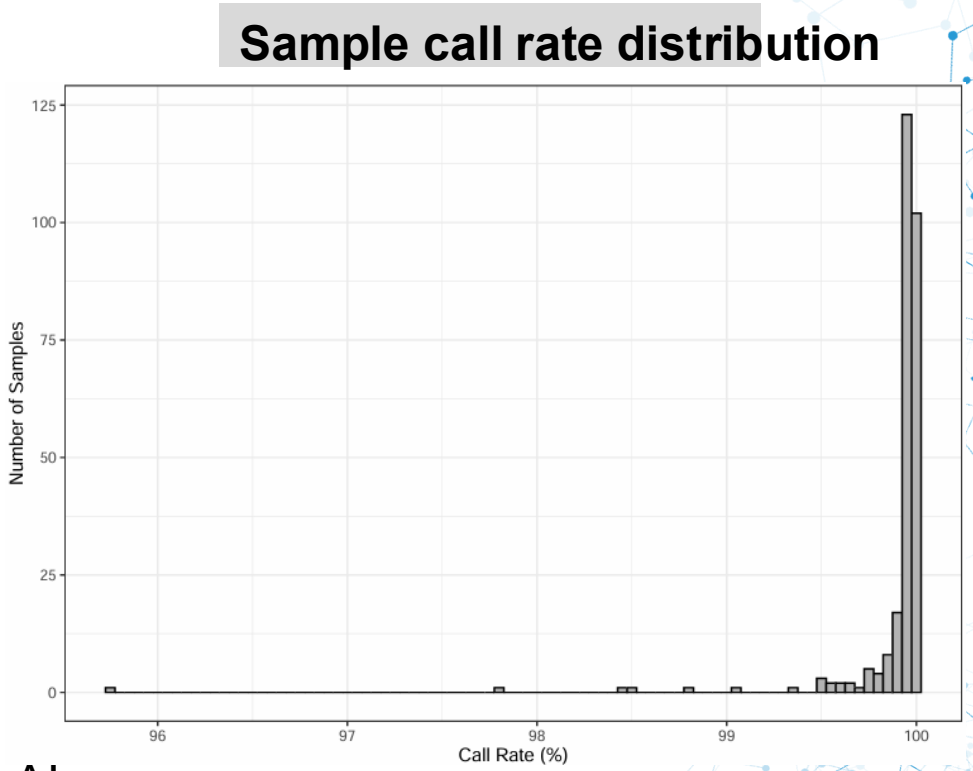
	LPS	Target Sequencing
Genome Coverage	Covers entire genome with low depth	Focused on pre-selected regions (exons, SNPs)
Read Depth	Low coverage (0.1-5x)	High coverage (50-500x)
Variant Detection	Captures genome-wide variation but requires imputation	Detects known and novel variants in targeted regions

CDCB Chip Validation Report with Visualization Plots

Visual Plot examples

- Distribution of SNPs - number of SNPs/chr
- Sample call rate % (>90%)
- SNPs with low calls (>10% missing)





Plot example from a Bovine 50K chip created with Claude.AI

Feedback: please let us know what you would be interested in seeing added!

SNP Coordinates Verification

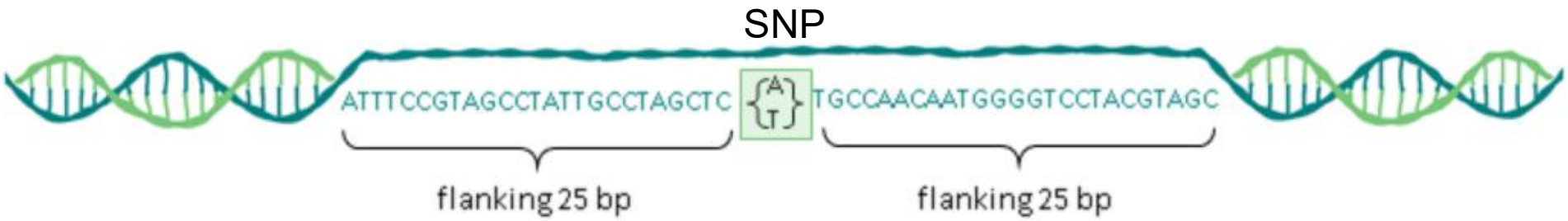
- In the previous CDCB system, genomic locations for new SNPs not present in our database were accepted as provided
- We are currently updating this policy, and we require submission of flanking sequences to enable SNP mapping

CDCB plan

- Verify all SNP locations before submitting them to the database during the chip validation process
- Improve data quality, avoid duplicated SNPs in the system, and detect early issues with SNP locations

Flanking Sequences

- Flanking sequences are nucleotide sequences located adjacent to a specific target sequence of interest — such as a SNP — on either side (upstream and downstream)
- They are landmarks for different applications in molecular biology
- Think of them as the flanking "sides" that border a central genetic marker



https://azifi.tz.agrar.uni-goettingen.de/agreg-snpdb/

Probes vs. Flanking Sequences

SNP array probes used for detecting alleles (genotyping)

- Short DNA sequences designed to bind to a specific SNP allele
- Typically, one probe per allele (to distinguish variants)
- Affymetrix: ~25 bp probes centered on the SNP
- Illumina: ~50 bp probes adjacent to the SNP + single-base extension

Flanking sequences

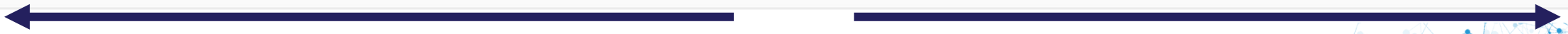
- Longer DNA sequences surrounding the SNP (±50–100 bp or more)
- Include the SNP in the middle: [A/G]
- Used for mapping and coordinate verification, not genotyping

During Chip Validating Process…

- The SNP map file must include flanking sequences for all SNPs to support verification and mapping of SNP positions
- Preference is for at least 50 bp of sequence upstream and downstream of the SNP

Example of a 100 bp flanking sequence:

```
SNP1 (50 bp + 50 bp)
TGAACACATTTACAAGTACTCTTTCTTATATATTACACTTTTCAACAAAA[A/C]AGTTACCATAGTAATAAAGTGAATTTGGTCCAGAAAACTCATGATGACTG
```



Do you have flanking sequences for existing chips that you have validated with CDCB?

During Chip Validating Process… Before Update to Database

```
Name,Index,Chr,Position,FlankingSequence
SNP_Name_1,1,19,123456789,GCAGTGGCACCTGCTCCCTTCTTCCTAGGTGCGCTTCTGTACGCTTACTA[A/T]ATCTCGGCTACATCGGCTACAATTGCGTGTTATGCTCGAGGCTTACACCT
SNP_Name_2,2,30,11223344,CGAGTGGAAATTGCTCACTTATGGCTAGGTGAGATTCTCTAGCCTTAGTA[C/G]CGCCTGGCTAGACTGCATAACCGGTGCGTGTTACGGTCCATTCATAGACA
SNP_Name_3,3,5,98765432,CTTGAGCATGTCGCGAACCTCAGGCAATGTGTGACTCTTTAGTCTGTGTA[C/A]AATCTTACTAGAGGGCATAGTCGATGCGAGTCACTGTACATGCAGAGATT
```

Mapping of the sequences using BWA software

Non-Y SNPs
mapped against ARS-UCD1.2 (chr 1-29, X)

Probe ID	Chr	SNP Position	QUAL
SNP1	1	1,997,582	54
SNP2	21	33,069,961	60
SNP3	19	13,630,556	57

Y SNPs
Mapped against Btau5.0.1

Probe ID	Chr	SNP Position	QUAL
NPF45	Y	3,729,035	60

Decision Summary

We compare our mapping results with SNP coordinates provided and/or coordinates stored at CDCB.

Scenario	Decision Framework & Action
High-Confidence Update (Mapped = Submitted ≠ Stored)	If the mapping result is high-confidence (MAPQ > 20) and aligns with the submitted coordinate but disagrees with the stored coordinate, **update the CDCB stored coordinate**. This addresses legacy database issues.
New SNP High-Confidence Update (Mapped = Submitted, No Stored)	If mapping confirms the submitted coordinate with high-confidence, **add the new SNP coordinate** to the database.
Conflict or Low-Confidence (Inconclusive)	If mapping quality is low (e.g., MAPQ < 20) or the result is non-unique (multiple mappings), **do not update** coordinates.
Full Disagreement (Mapped ≠ Submitted ≠ Stored)	**Do not update** coordinates. Flag the SNP for further review, and **request clarification from the submitting laboratory** to achieve universal agreement on the location.

SNP flagged for further review

- Contact the lab to achieve universal agreement on the SNP coordinate
- If no agreement, a secondary verification is recommended
- If coordinate confirmed, add new SNP coordinate on CDCB database

SNP Usability Evaluation

SNP arrays are not static in quality, even if they have been used for years.

Over time, problems accumulate because:

- Genome assemblies improve/change (SNP coordinates can shift)
- Populations change (new breeds, selection)
- Changes in allele frequencies due to selection
- Accumulation of poorly performing probes
- Technology improves

SNP Usability Evaluation

Currently CDCB performs two SNP usability checks:

1. **Chip-based SNP quality check**
   Monitors all SNPs for all genotypes loaded within the last 90 days
   Focuses on short-term monitoring and early detection of issues

2. **Comprehensive SNP quality check on all stored genotypes**
   Performed periodically on all stored genotypes within the evaluation for deeper analysis and substantial adjustments

CDCB is currently revising these procedures to ensure consistent monitoring

Key Takeaways

- CDCB continues to improve chip validation and SNP usability by re-evaluating processes and criteria and by incorporating additional information such as SNP coordinate verification and visualization plots.
- CDCB accepts GBS data and applies similar quality criteria used for SNP genotyping arrays to this type of data (we are open to discussing new technologies).
- CDCB plans to update its system to the latest cattle genome assembly in the future.

Thank you!

Questions?

Clarissa Boschiero
Bioinformatics Scientist
clarissa.boschiero@uscdcb.com