Validation of New Arrays & Genotyping Technologies

Heather Enzenauer, Ph.D.

2026 CDCB Genomic Nominator and Laboratory Workshop

Validation

- The process of confirming that data are accurate, reliable, and meet required standards.
  - Detect errors before they impact downstream analyses

- Ensures suitability for integration into the National Cooperator Database.
  - Maintains consistency across platforms and technologies
  - Builds confidence in results and decisions
  - Prevents propagation of poor-quality data

Why Validation Matters

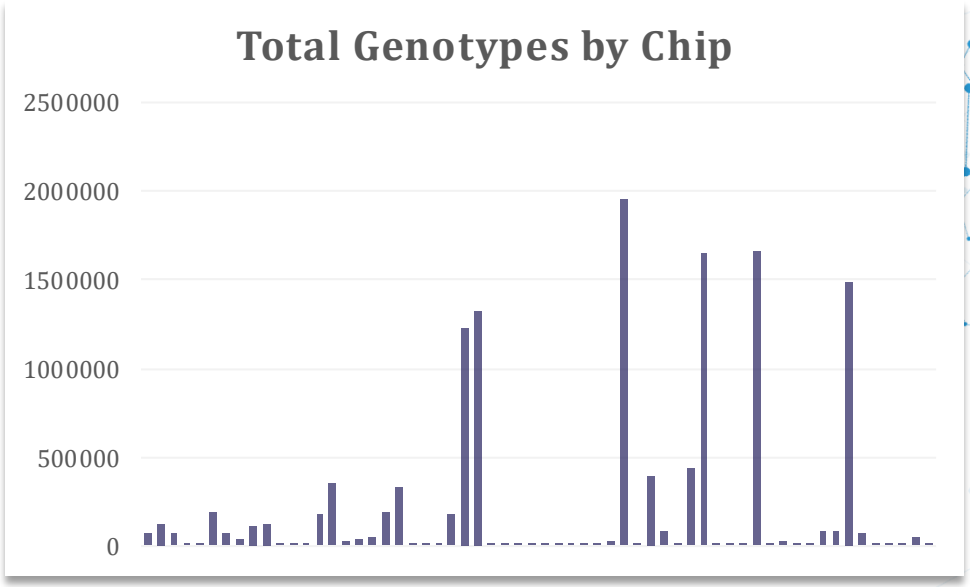
- All CDCB services depend on high-quality input data
- Multiple technologies → variability in data quality
- Validation = gatekeeper to entry into the National Cooperator Database
  - Ensure genomic data entering the National Cooperator Database meets or exceeds established standardized expectations

What Is Validated

- Platform types
  - SNP Arrays (Illumina, Affymetrix)
  - GBS / sequencing panels
- Data providers (laboratories)
- Data itself

Technology Type	Chip Count
Affymetrix SNP Chips	12
Illumina SNP Chips	47
GBS LP + TE Panels	1
GBS TE Panels	1

GBS: Genotyping-by-Sequencing
LP: Low-Pass Sequencing
TE: Target Enrichment / Target Sequencing



Accepted chips: https://redmine.uscdcb.com/projects/cdcb-customer-service/wiki/REFERENCES#Ref162

Validation Goals

- Compatibility with CDCB systems
  - Processes through CDCB pipelines without errors
- SNP-level quality
  - Marker behavior (call rate, allele frequency, inheritance patterns)
- Sample-level quality
  - Genotype behavior (consistency with existing records)
- Long-term performance expectations
  - Perform reliably as data accumulates over time

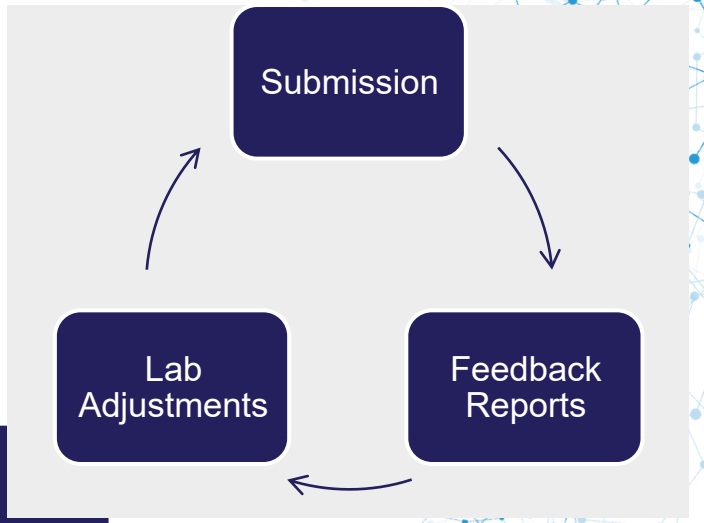
High-Level Workflow

Application Process
- Initial correspondence
- Receive and review application; collect fee; check requirements are fulfilled

Chip Validation / Analysis Process
- Determine the reliability of the SNP on the proposed chip and assess sample quality

Official Test
- Evaluate the chip in a test environment (from submission to report card creation)

Live! + Monitoring long-term performance
- The new chip becomes available to collaborators
- Assessing the new chips performance over time



Validation documentation: https://redmine.uscdcb.com/projects/cdcb-customer-service/wiki/CDCB_SNP_Array_Validation_Process

Data Requirements

- Standard formats (FinalReport, SampleSheet)
- Males + females required
- AB genotypes
  - SNP arrays: 51+ genotypes
  - GBS / sequencing panels: 200+ genotypes; 2+ subsequent batches of 51+
- Map file with SNP coordinates based on ARS-UCD1.2 assembly
  - Now including Flanking Sequences!
- Required SNP lists
  - 4K Fast Discovery (Critical 96) → ≥ 3480/3552 (including all 96)
  - Quick Discovery (using ICAR 550) → ≥ 350/550
  - 195 ICAR parentage list → 195/195
  - Y SNPs for gender verification → 10

0→BB, 1→AB, 2→AA

4K Fast Discovery used for:
Parentage and identical genotype discovery, grandsire unlikely determination, fast sire discovery

550 Quick Discovery (QDisc) used for:
Close relative discovery procedure
- Separate from FULL service and Quick Turnaround (QTurn)

195 ICAR SNP list used for:
Official SNP-based parentage certificate

NEW: Flanking Sequence Requirement

- Used to:
  - Verify SNP identity by mapping to the reference assembly (ARS-UCD1.2)
  - Resolve discrepancies between submitted and stored SNP coordinates

```
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
```

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

SNP-Level Validation

- Call rate (>90%)
- Parent-progeny conflicts (PPC)
- Opposite homozygosity
- Allele frequency comparison
- Mapping consistency

45 female samples where the submitted genotype was 1 (AB), and the stored genotypes was 0 (BB)

checksnp_conflicts.csv

```
snp_key,chrom,anim_gt,self_gt,sire_gt,dam_gt,countF,countM,snp_name
430942,7,1,0,5,5,45,0,ARS-BFGL-NGS-113696
430942,7,1,2,5,5,1,0,ARS-BFGL-NGS-113696
430942,7,2,1,5,5,13,0,ARS-BFGL-NGS-113696
430942,7,2,5,0,5,1,0,ARS-BFGL-NGS-113696
```

Example of Reported Conflicts:

checksnp_bysnp.txt

snp_key	chr	use	compares	conflicts	pct	sire	dam	trio	self	snp_name_FR
1103918	20	1	151	103	68.2	30	6	0	67	ARS-BFGL-NGS-109135
377505	6	1	159	101	63.5	37	3	0	61	Hapmap40629-BTA-102563
430942	7	1	97	60	61.9	1	0	0	59	ARS-BFGL-NGS-113696
1470786	31	1	71	29	40.8	29	0	0	0	BOVINEHD3000025218
503701	8	1	103	28	27.2	0	0	0	28	ARS-BFGL-NGS-15251
720121	11	1	113	25	22.1	1	0	0	24	Hapmap51377-BTA-106302

0→BB, 1→AB, 2→AA, 5→--

Sample-Level Validation

- Call rate (>90%)
- Concordance with existing genotypes (>99%)
- Sex verification
  - Y calls in males
  - Heterozygosity patterns

	X Heterozygosity	Y-Count
Males	hetero <1.4 is valid, 1.4-7.2% is abnormal, >7.2% is a sex conflict	Must have >70% available Y-SNPs
Females	hetero <7.2% is sex conflict†, 7.2-17.2% is abnormal, >17.2% is valid	Must have <20% available Y-SNPs

https://redmine.uscdcb.com/projects/cdcb-customer-service/wiki/CDCB_Genomic_Dictionary#Sex-abnormality

Post-Validation Monitoring

- Submission and Monthly report cards
- 3-month rolling QC checks
- SNP usability updates
- Lab communication

Active lab involvement is critical!

Common Issues We See

- Missing/incorrect SNP coordinates
- Naming inconsistencies
- High PPC SNPs
- Allele reversals
- Low call SNPs

Flanking sequences!

- GBS batch-to-batch variability
  - Call rate differences
  - Conflicts from imputation errors

Chromosome	Definition
0	unassigned
1-29	Autosomal
30	Pseudo-autosomal on X
31	X-specific
32	Y-specific
33	Mitochondrial

Emerging Technologies & Challenges

- Sequencing / GBS
  - Filtered vs. unfiltered data; imputed calls; variable call rate
- Imputation pipelines
  - Breed proportions and reference population size?
- Smaller panels vs. high-density data
- Cross-platform consistency

Discussion	Approach
Compatibility with existing data	Robust genotype processing to flag discrepancies
Validation requirements	Larger submission sample size
External factors affecting quality	Request additional batches of data

Emerging Technologies & Challenges

- Continued considerations for all new technologies
  - Ensure consistency across labs and technologies
  - Integrate new data seamlessly
  - Additional validation steps
- Balancing innovation while safeguarding quality and comparability
- Beyond GBS…whole-genome sequencing
- Beyond genomics …epigenomics, microbiome data

Key Takeaways

Validation ensures trust in genomic evaluations

SNP-level and sample-level QC is critical

Flanking sequences will improve accuracy

Monitoring continues after approval

Questions?

heather.enzenauer@uscdcb.com