Council on Dairy Cattle Breeding
Genotyping-By-Sequencing

2024 CDCB Annual Nominator & Laboratory Workshop

Heather Enzenauer, Applied Geneticist

July 30, 2024

COUNCIL ON DAIRY CATTLE BREEDING

- Review of Sequencing (and more)
- Introduction into GBS (Genotyping-By-Sequencing)
- Future of GBS at CDCB

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Genome Sequence

http://media.economist.com/sites/default/files/cf_images/19980516/3st1.jpg

- Sometimes referred to as “decoding”
- **Genome Sequence**: very long string of letters
  - Reading is not just in the sequence of letters, but also the words constructing the sentence

Order is important! Example: cat vs act
Switching the **c** and **a** change the word and definition, just like changing nucleotides in a codon!
HH1: C → T causes an embryonic lethal! CAA becomes TAA (Stop codon)

- Genes account for < 25% of the DNA in the genome
  - Remaining: regulatory regions that control how genes are turned on and off, long stretches of “non-functional" DNA — so called because it has no known relevant biological function

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Genotyping

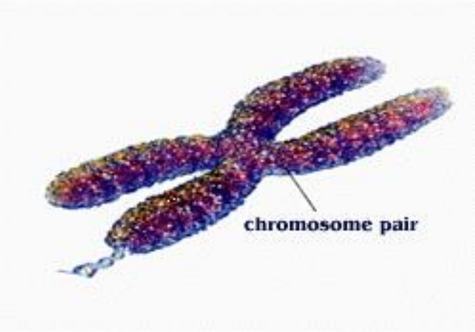
- Process to determine which genetic variants an individual possesses
- Reveals alleles inherited from parents

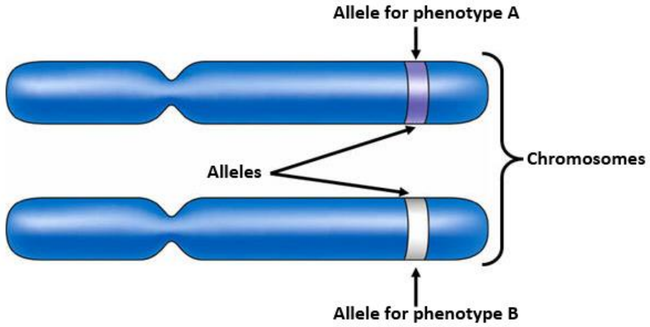
- Genotypes code for phenotypes
  - **Genotype**: organism's genetic composition of alleles
  - **Phenotype**: organism's observable characteristics (results from interaction of genotype and environment)

- **Locus**: specific location/position of a gene on a chromosome
- **Allele**: Alternative forms of the same gene or genetic locus

$$P = G + E$$

phenotypic value | genotypic value | environmental deviation





Heritability: how much variation in phenotype (Vp) is due to genotype (Vg)
$H^2 = Vg/Vp$

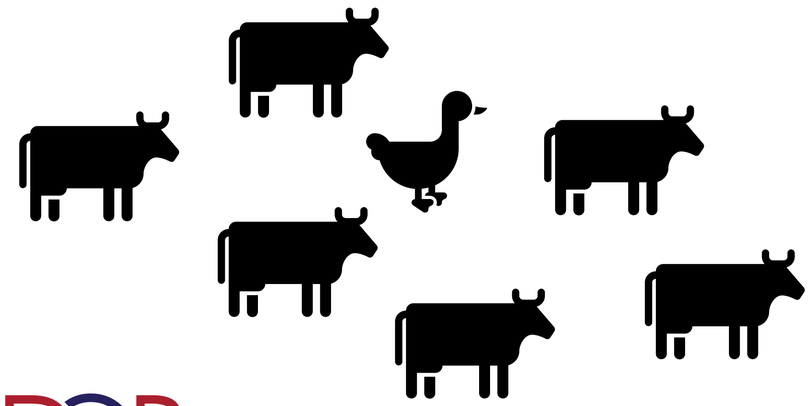
- **Homozygote**: Two copies of the same allele
- **Heterozygote**: Two different alleles

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

What are we looking for?

- Genomic/genetic variation

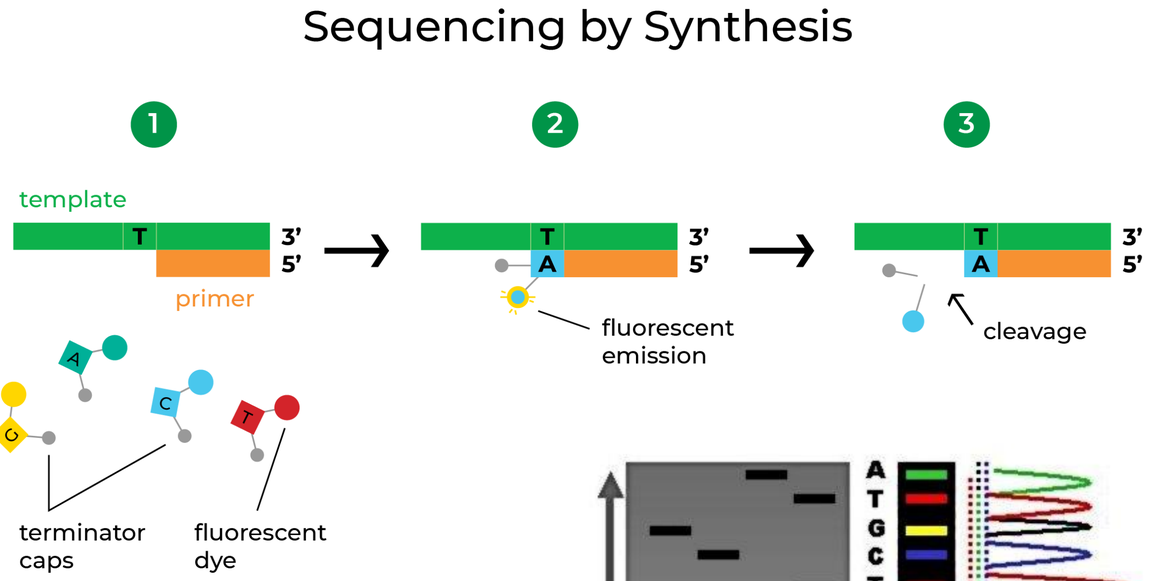
"One of these things is not like the others,
One of these things just doesn't belong,
Can you tell which thing is not like the others,
By the time I finish my song?"

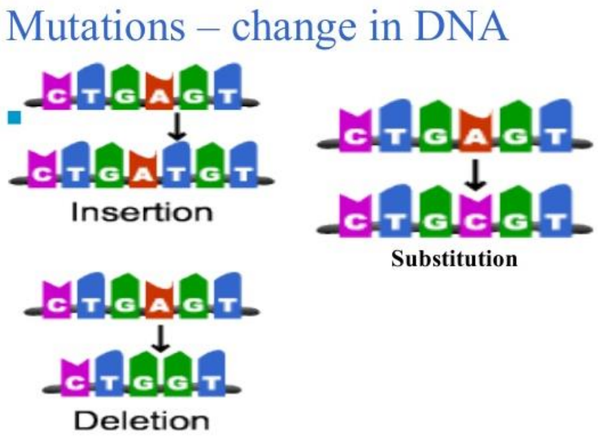


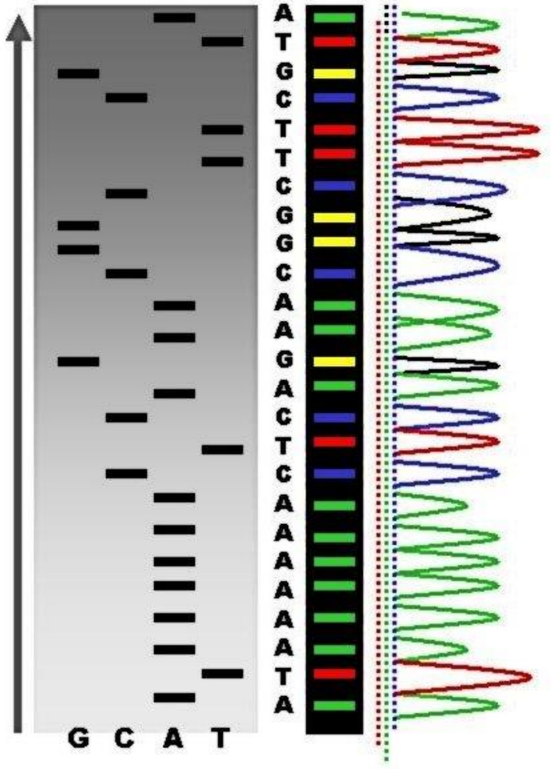
CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Sequencing

- Determining the exact order of the bases in a strand of DNA
- Translate the string of bases into an understanding of how the genome works
- Search for genetic variations and/or mutations that may play a role in the development or progression of a disease/abnormality (substitutions, deletions, or additions of a single base pair or as large as thousands of bases)





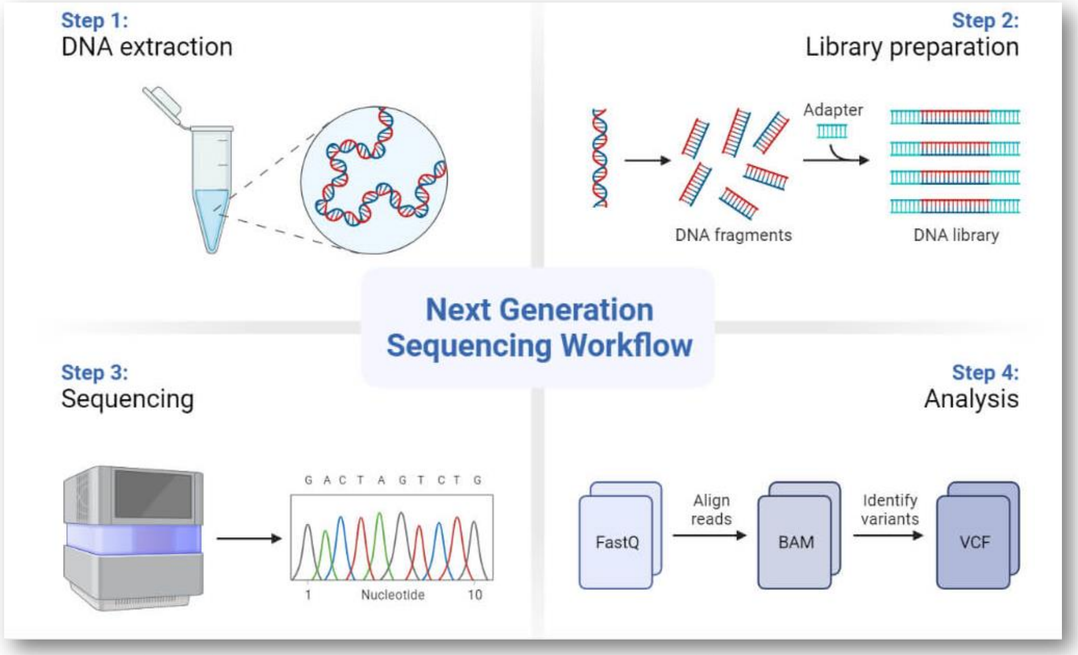


https://en.wikipedia.org/wiki/DNA_sequencing

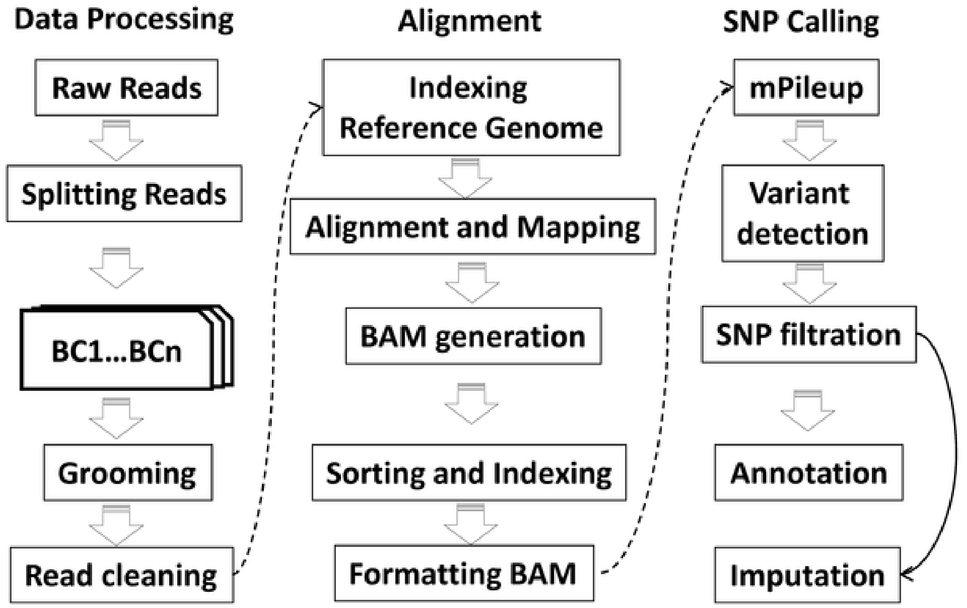
CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Sequencing

- Next Generation Sequencing (NGS)
  - Enables sequencing of thousands to millions of DNA molecules simultaneously

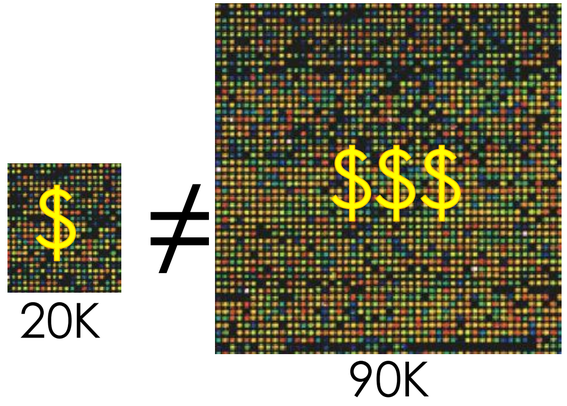


https://www.biorender.com/template/next-generation-sequencing-workflow



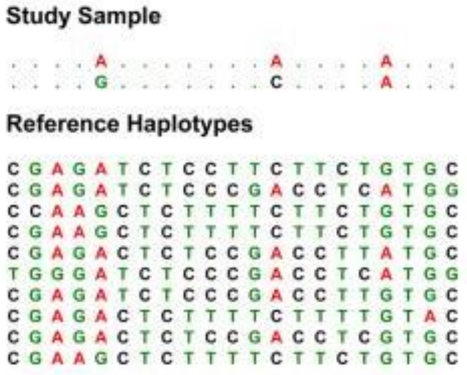
CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Imputation

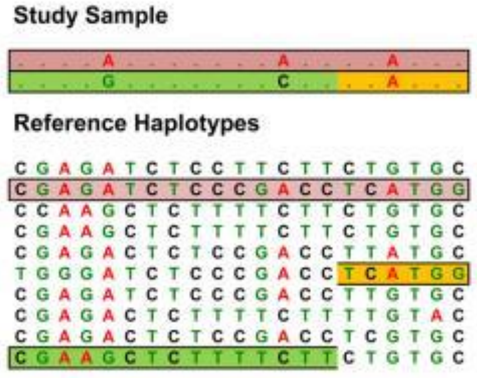


- Statistical inference of unobserved genotypes
  - Fill in missing SNP values
- Combine animals genotyped on different chips (50K, 20K, 6K genotypes, etc.)
- Larger density = more complete info on animal's genome
- Linkage disequilibrium: correlation of SNP values caused by their tendency to travel together during recombination (SNPs close together)

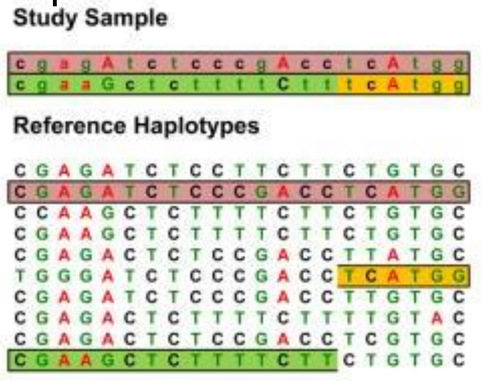
Haplotype: group of variants or genes inherited together from a single parent



Identify shared regions with known alleles between



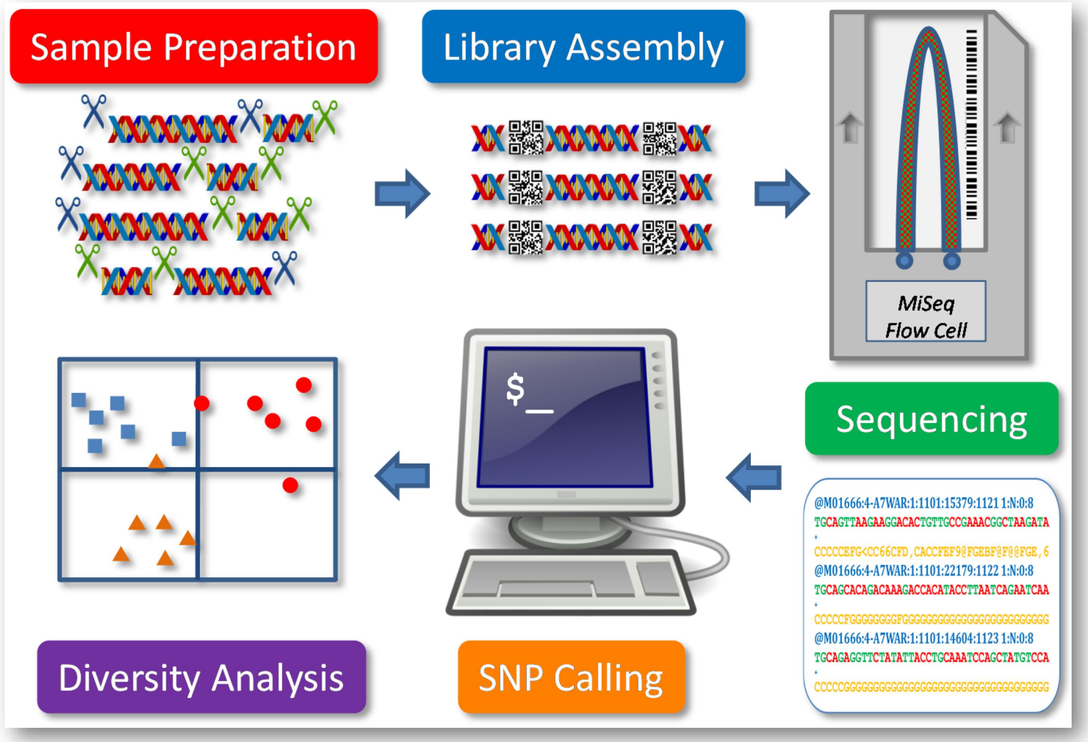
Based on similar 'patterns' between the two, fill in the missing positions in the study sample



CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Genotyping-By-Sequencing

- Next Generation Genotyping (Genotyping-By-Sequencing: GBS)
  - Genetic screening method for discovering novel SNPs and performing genotyping studies
  - Sequence predetermined areas of genetic variation over many samples
  - Ensure sufficient overlap in sequence coverage

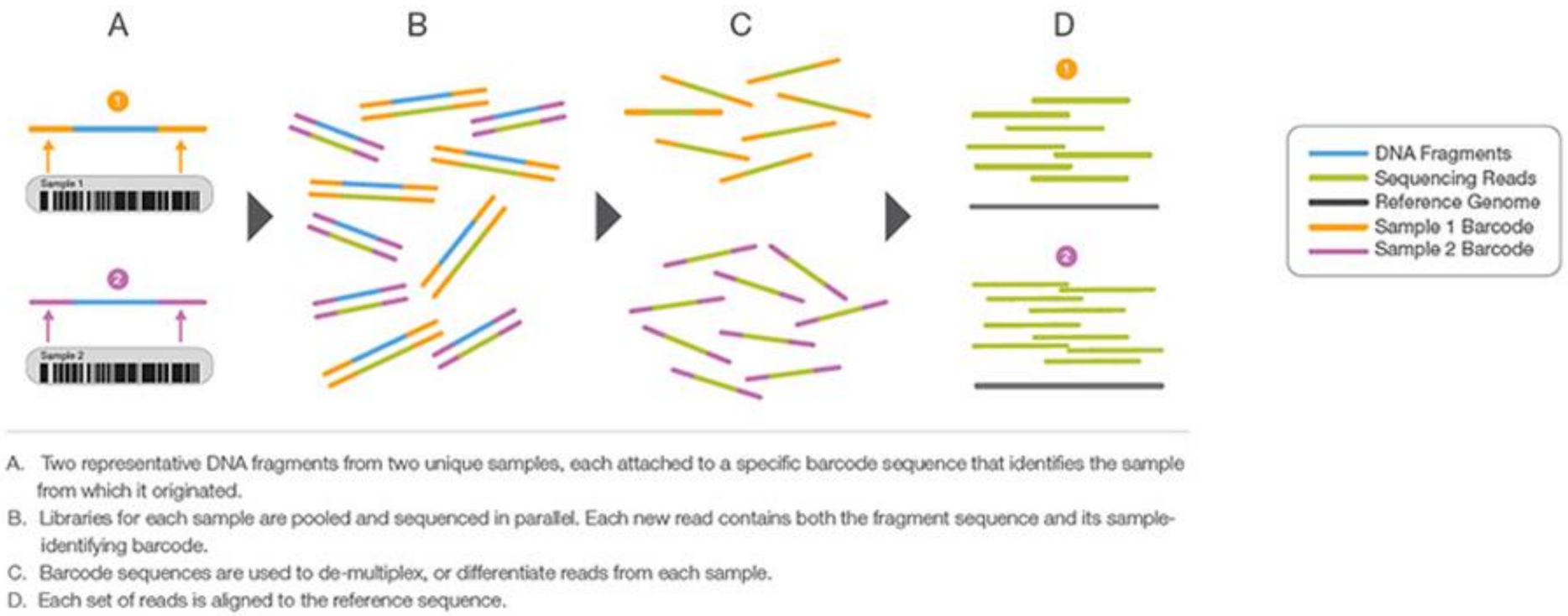


Peterson GW, Dong Y, Horbach C, Fu Y-B. Genotyping-By-Sequencing for Plant Genetic Diversity Analysis: A Lab Guide for SNP Genotyping. *Diversity*. 2014; 6(4):665-680. https://doi.org/10.3390/d6040665

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Multiplex sequencing

- Allows large numbers of libraries to be pooled and sequenced simultaneously
  - Works for smaller genomes OR when targeting specific genomic regions
- Adds a barcode sequence to each DNA fragment for sample identification



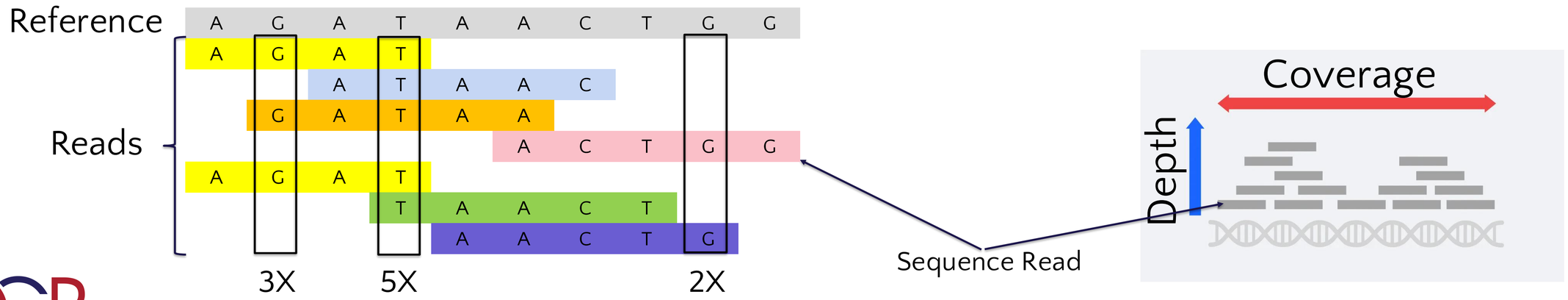
A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
D. Each set of reads is aligned to the reference sequence.

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing.html

Additional Definitions

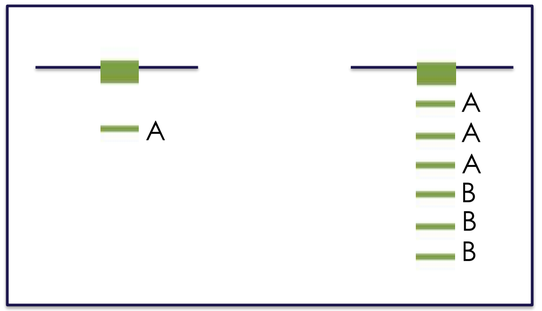
- **Coverage**: percent of genome sequenced at certain depth
- **Depth**: how many reads detected a specific nucleotide at a region or position
- **Read**: sequence from a small section of DNA



CDCB
COUNCIL ON DAIRY CATTLE BREEDING

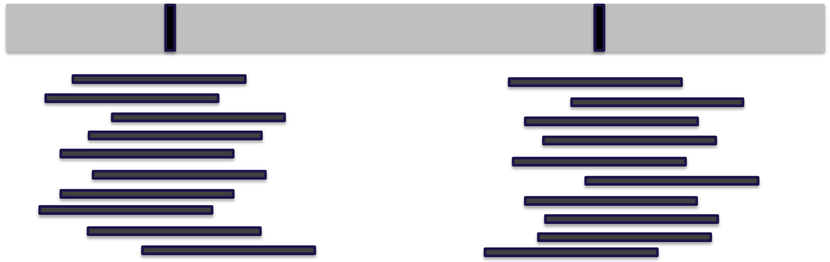
GBS + (?)

- Low-pass sequencing
  - Sequencing a genome to an average depth <1X depth of coverage and applying imputation (Li et al., 2021)
- Target capture/enrichment
  - High-depth sequencing of target loci/regions
  - Benefit: reduce output size; improve accuracy of genotyping target SNPs



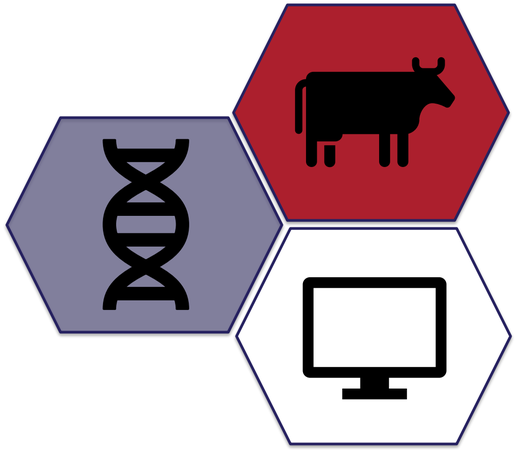
Whole Genome

GBS with Target Enrichment



CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Genotypes



- Array-based genotypes (SNP arrays/chips)
  - Include fixed lists of important, well-studied variants
  - Currently 54 CDCB-approved chips
- Sequencing-based genotypes
  - Flexible lists of variants based on sequence data
  - Currently validating

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

SNP Arrays vs. GBS

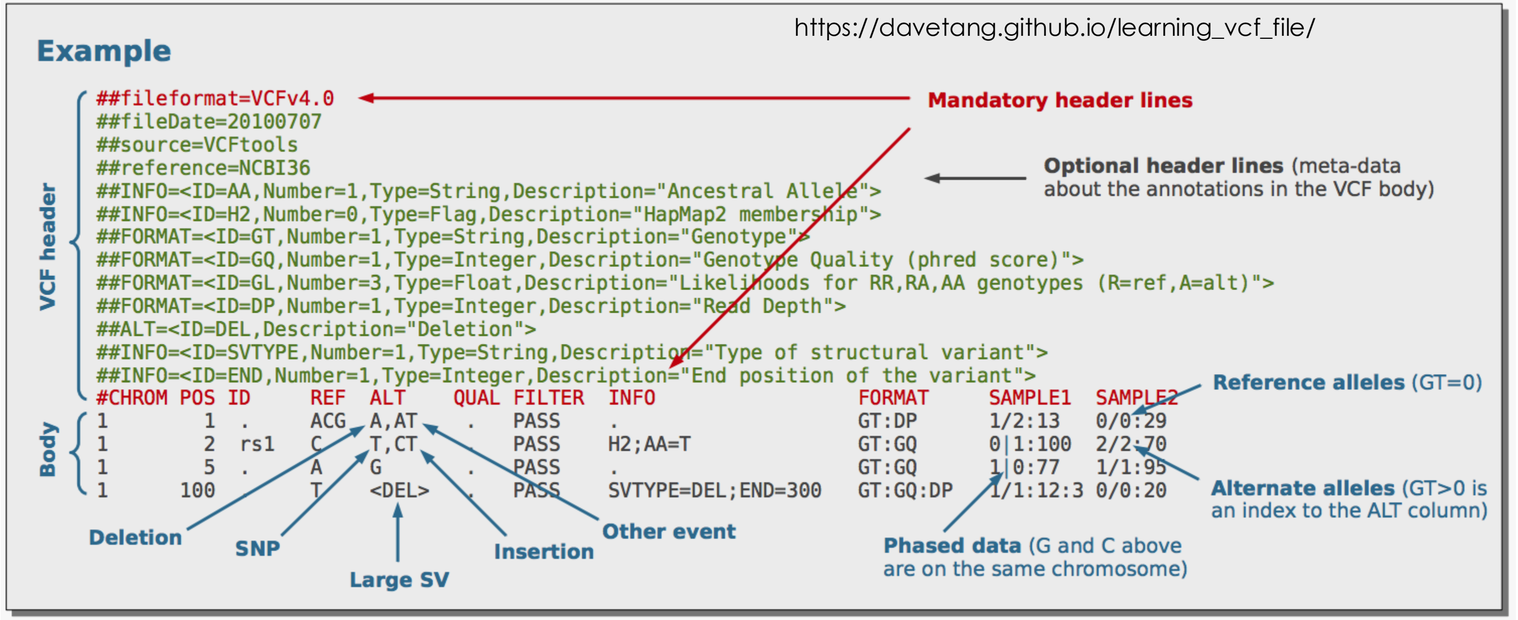
SNP Arrays		GBS	
Pros	Cons	Pros	Cons
• Readily available technology in cattle • Stable markers chosen • Simplified data analysis	• Chip development cost is high for minor breeds with small samples sizes • Need to genotype many individuals for chip to be cost effective • Cannot identify variants outside of pre-determined set	• More flexibility (all SNPs potentially available) • Larger batch sizes for sequencing • Improved accuracy • Identify variants other than SNPs • Lower cost	• Biased reference set for imputation • Sequencing errors at low coverage/depth • Complex data pipeline

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

**More data from GBS = opportunities for research, improved genomic evaluations, dynamic adjustments to Evaluation SNP list

Variant Call Format (VCF) File

- Text file output that includes sequencing information



	Name	Brief description (see the specification for details).
1	CHROM	The name of the sequence (typically a chromosome) on which the variation is being called. This sequence is usually known as 'the reference sequence', i.e. the sequence against which the given sample varies.
2	POS	The 1-based position of the variation on the given sequence.
3	ID	The identifier of the variation, e.g. a dbSNP rs identifier, or if unknown a ".". Multiple identifiers should be separated by semi-colons without white-space.
4	REF	The reference base (or bases in the case of an indel) at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.
6	QUAL	A quality score associated with the inference of the given alleles.
7	FILTER	A flag indicating which of a given set of filters the variation has failed or PASS if all the filters were passed successfully.
8	INFO	An extensible list of key-value pairs (fields) describing the variation. See below for some common fields. Multiple fields are separated by semicolons with optional values in the format: `<key>=<data>[,data]`.
9	FORMAT	An (optional) extensible list of fields for describing the samples. See below for some common fields.
+	SAMPLEs	For each (optional) sample described in the file, values are given for the fields listed in FORMAT

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

SNP Arrays vs. GBS

SNP Arrays		GBS	
Pros	Cons	Pros	Cons
• Readily available technology in cattle • Stable markers chosen • Simplified data analysis	• Chip development cost is high for minor breeds with small samples sizes • Need to genotype many individuals for chip to be cost effective • Cannot identify variants outside of pre-determined set	• More flexibility (all SNPs potentially available) • Larger batch sizes for sequencing • Improved accuracy • Identify variants other than SNPs • Lower cost	• Biased reference set for imputation • Sequencing errors at low coverage/depth • Complex data pipeline

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

**More data from GBS = opportunities for research, improved genomic evaluations, dynamic adjustments to Evaluation SNP list

CDCB Validation Process

SNP Array Data

- Submit validation application and fee
  - $5,000 – new SNP array using previously validated technology
  - $15,000 – new SNP array using new unvalidated genotyping technology
  - At least 50 samples with their own or parental genotypes existing in the CDCB database
- SNP coordinates based on ARS-UCD1.2 assembly
- Inclusion of key SNPs
- Assess data quality and SNP performance
  - Discrepancies in allele frequencies
  - SNPs with no and low call
  - Animals/SNPs with large numbers of conflicts between stored and submitted genotypes, parent-progeny conflicts (PPC)

Fast Discovery, QDisc, ICAR, Y SNPs, …

GBS Data

- Submit validation application and fee
  - $15,000 - new SNP array using new unvalidated genotyping technology
- At least 200 samples with existing genotypes in the CDCB database
- SNP coordinates based on ARS-UCD1.2 assembly
- Inclusion of key SNPs
- Assess data quality and SNP performance
  - Same checks as in SNP array data
- **Imputation quality**
  - **Concordance between CDCB-stored and GBS-submitted genotypes**
- **Data quality (Final Report)**
  - **GenCall (GC) score: Illumina metric that measures the confidence of a genotype's assignment**
- **Sequencing quality (VCF)**
  - **GQ (Genotype Quality): confidence that genotype being assigned is correct**

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

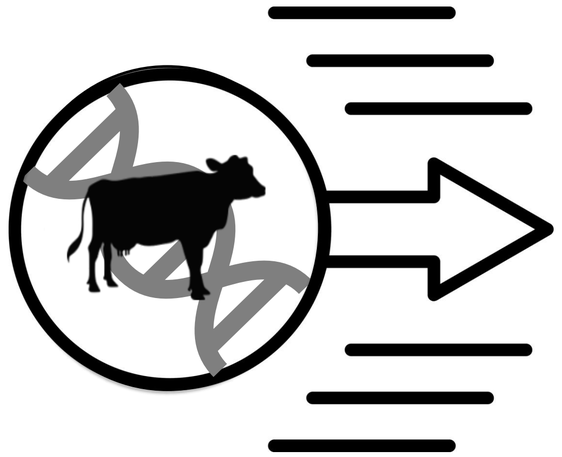
What we've learned…

- Low Pass + Target Enrichment + Imputation
  - High Quality Genotypes
  - Affected by sequencing protocols and imputation
- Filtering is important
  - High standards = High concordance (~99.7%)
- Inconsistency of call rates across samples/submissions
  - Sample and SNP call rates are good indicators of quality for SNP arrays, but for GBS?
- **Each validation approached as a new technology**
  - Variations in protocols used by labs

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Moving forward…

- Establishing requirements for GBS-based genotypes
  - Inclusion of quality metrics
  - VCFs available during the validation process
  - Supplemental information as needed
- Creating new metrics for assessment
  - Consistency and accuracy across submissions
- Develop a pipeline to process more complex files (e.g. VCF files)
- Currently no set date for incorporation of GBS data, but we are making progress!



CDCB
COUNCIL ON DAIRY CATTLE BREEDING

THANK YOU FOR YOUR ATTENTION

www.uscdcb.com

heather.enzenauer@uscdcb.com

CDCB
COUNCIL ON DAIRY CATTLE BREEDING