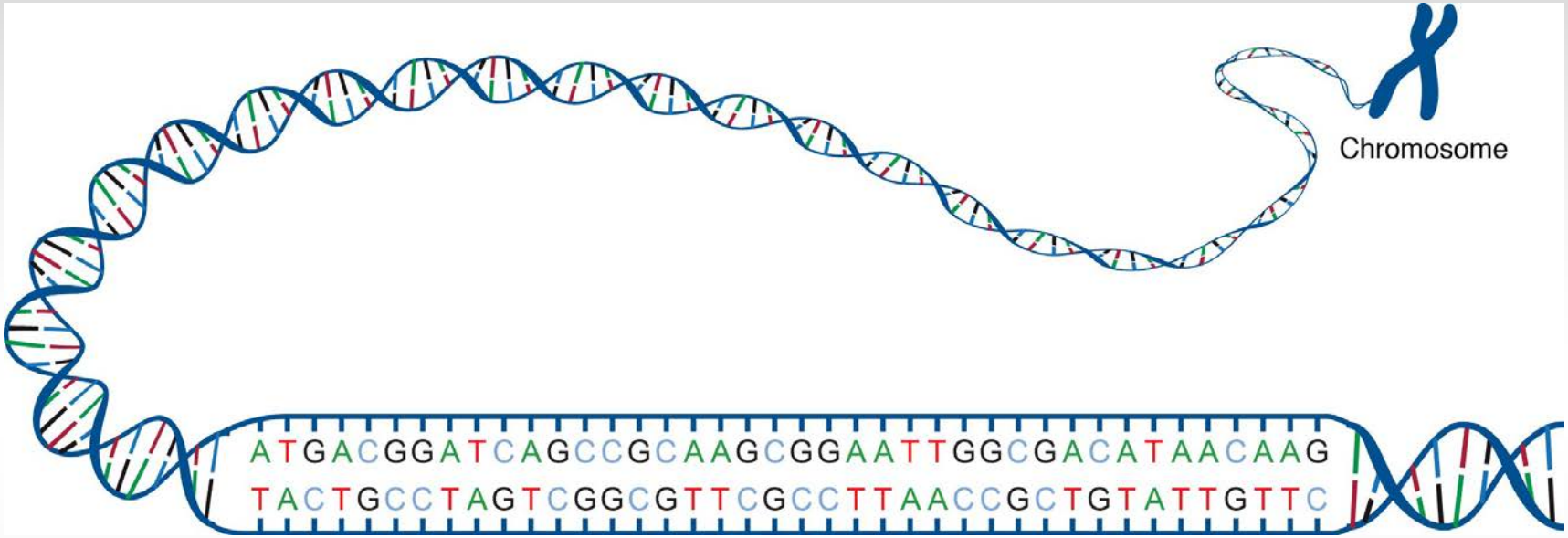
Whole Genome Sequencing: Background and View of Current Efforts

Larry Kuehn, Ph.D.

Research Geneticist

USDA, ARS, U.S. Meat Animal Research Center

Genome sequence



genome.gov

- Trying to read the base pairs (the code) along the whole genome

Whole genome sequencing

- Loosely defined as reading and recording all 3 billion bases in the cattle genome
  - Introns
    - Do not code for any protein
    - Over 98% of genome
    - Likely some regulatory function
  - Exons
    - Remainder
    - Coding regions

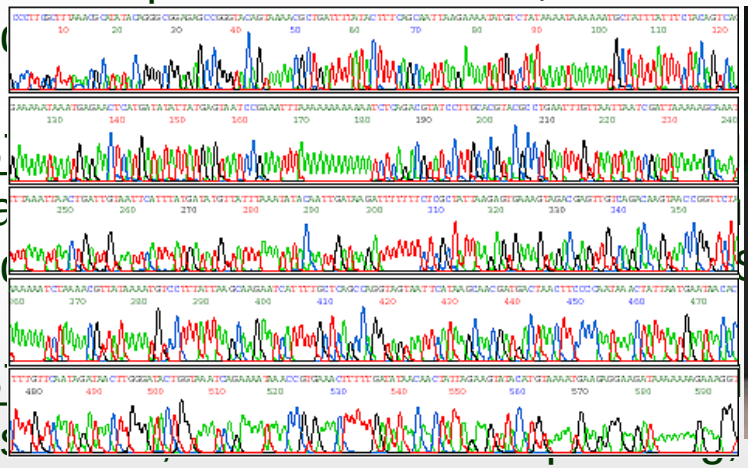
First cattle genome sequence

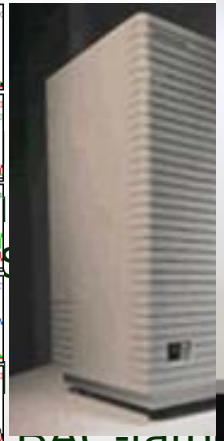
- Dominette
  - USDA-ARS – Miles City Montana
  - Inbred Hereford female
- Impact
  - Reference Bovine genome
  - Cost over $53 million
  - Led to DNA chips (50K chip, etc.)
  - Base for further sequencing projects

USMARC sequence technology progression:

1992 Pharmacia manual slab gel – about 20,000 bases w
$7 per 1,000 base read; used for microsatellite se
($7,000 per Mb)

1997 ABI 377 gel-based system – about 380,000 bases v
about $3 per 800 base read; used for small sca
an





1999 ABI 3
rea
an

2003 ABI 3
ba
BAC-targ
lots of amplicon resequencing in breed panels (

Next generation sequencing

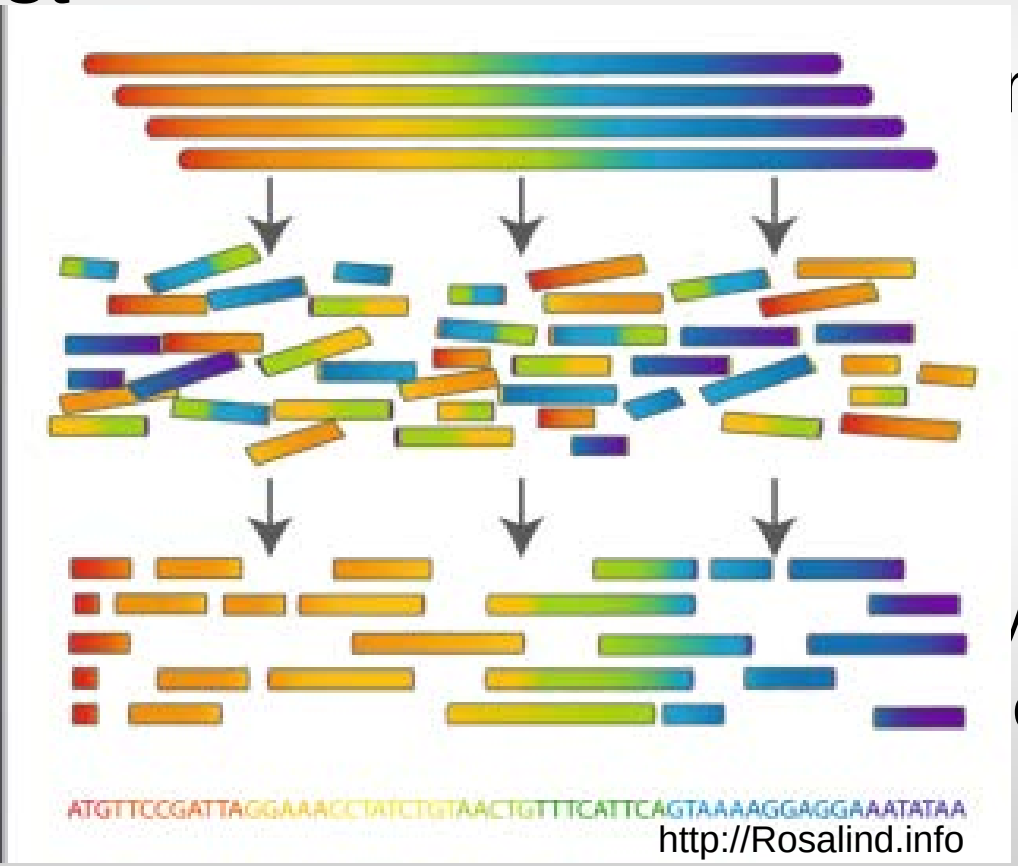
- Multiple platforms available

Next generation sequencing

- Accuracy ranges from 87% to 99%
  - General trend is decreased accuracy with longer read-length
  - Generally read 50 to 1000+ bases per read
  - Easier to assemble longer reads
- Net effect is cheaper genome sequencing
  - Costs from $0.05 to $10 per million bases
  - Enough sequence for $150 whole genome

Some complications

- Initial cost
  - Sequen million
- Genome
  - Overla
    - Long
    - Less / chance
    - Helps er cattle)



http://Rosalind.info

Complications

- Tremendous amount of data
  - Hard to store, let alone manipulate
  - Makes assembly even tougher
  - Have to develop tools based on results
    - For example, genotyping based on new markers discovered in sequence (50K chip)
    - Hard to decide what ‘differences’ among animals are important
      - As with chips, phenotypes are still very important

Complications

- Multiple reads required
  - Reading 3 billion bases does not mean whole genome read
    - Each animal has two full genomes (diploid)
    - Genome is fragmented randomly into small pieces
    - Same section of genome likely read numerous times
    - Some advocate sequencing 20-50x when targeting whole genome
      - Focusing on one animal reduces impact of examining multiple animals (diversity, discovery, etc.)

Current efforts

- So we have potential for a lot of data
- Now what?
- Current efforts at USMARC and some selected programs elsewhere

Why bother?

- From a geneticist perspective:
  - Interested in sequencing to improve our chance at finding causal variation
    - Examine differences in sequence
    - Leads to new markers and mutations
    - Mutations may change protein structure or protein regulation
  - Ultimately differences we see due to genetics lie with differences in the genome or its regulation

Two different strategies

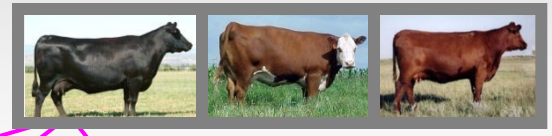
- Compare genomes of diverse animals to see where they are different
  - Marker development, mutational candidates
  - Follow up with genotyping platform
- Associate differences in sequence among several animals with trait variation
  - Requires large numbers of sequenced animals
- First strategy cheaper, more restrictive

GPE Cycle VII Population

AI Sires: AN, HH, AR, SM, CH, LM, GV

×

Base Cows: AN, HH, MARC III



AN & HH only

F_1 Steers

F_1 Bulls

×

F_1 Cows

F_1^2 Steers

F_1^2 Cows

GPE low coverage sequencing

- Foundational bulls from cycle VII
  - 151 AI bulls
  - 45 F_1 bulls with most progeny
- Targeting 2x coverage on each bull
  - Use information from relationships to improve individual bull coverage
  - Assuming higher value from broadly covering all base animals vs. a few bulls

GPE low coverage sequencing

- Plan is to impute large portions of bull sequence to phenotyped descendents
  - Population is currently imputed to 770K
- Functional variants are first target
  - Alison detailed some examples counts in first portion of presentation

Functional variants

Impact	Effect	
High	SPLICE_SITE_ACCEPTOR	FRAME_SHIFT
	SPLICE_SITE_DONOR	STOP_GAINED
	START_LOST	STOP_LOST
	EXON_DELETED	RARE_AMINO_ACID
Moderate	NON_SYNONYMOUS_CODING	CODON_CHANGE
	CODON_INSERTION	CODON_DELETION
	UTR_5_DELETED	UTR_3_DELETED
	CODON_CHANGE_PLUS_CODON_INSERTION	
	CODON_CHANGE_PLUS_CODON_DELETION	
Low	SYNONYMOUS_START	NON_SYNONYMOUS_START
	SYNONYMOUS_CODING	NON_SYNONYMOUS_STOP
	SYNONYMOUS_STOP	START_GAINED
Modifier	All other effects	

Congolani et al., 2012

Next steps

- functional impact of variants identified from low-coverage sequences of 96 bull

Impact	entire genome	AWM genes	top gene sets[a]
High	2,432	54	67
Moderate	27,640	717	1131
Low	62,481	2147	3279
Modifier	10,693,709	332,765	270,748

[a] genes near top SNP from proteolysis,pathway and cellular component gene sets

Next steps

- Same population as well as extreme animals for feed intake and growth are being targeted for whole exon capture
  - Recall exon refers to coding regions
  - Less costly to generate higher coverage
  - Further contribution to loss of function mutations

Other efforts

- Other projects are conducting large sequencing projects to detect further genomic variation
- University of Missouri leading a grant aimed at sequencing for variation in fertility
  - Around 100-200 animals with whole genome sequencing
  - Loss of function mutations in exons are likely candidates for prenatal death

Other efforts

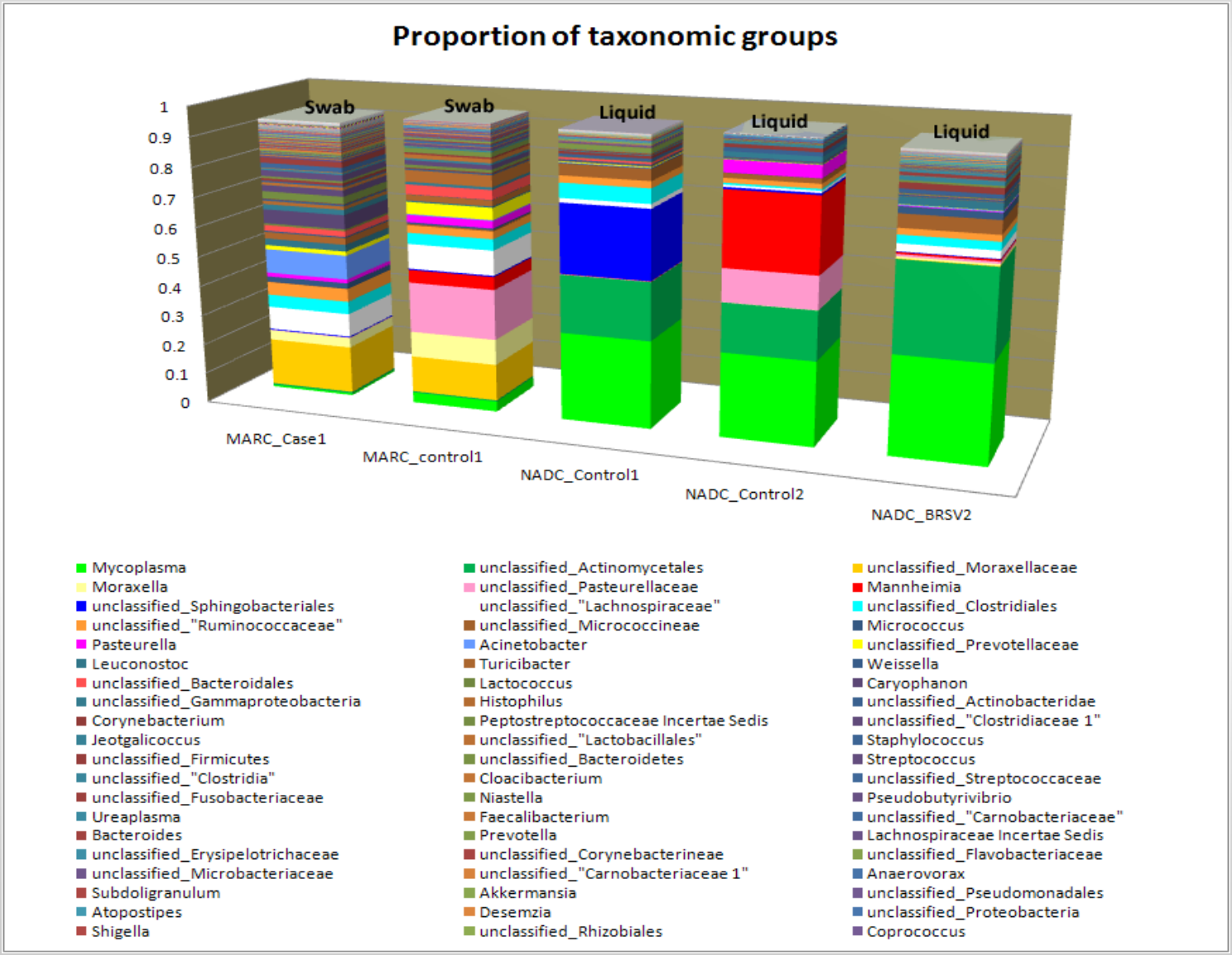
- Genome Canada
  - Approximately 30 animals sequenced at 5X coverage for several breeds

- 1,000 bovine genomes project
  - Ben Hayes (Australia) lead
  - http://www.1000bullgenomes.com/
  - Ultimately ‘public’ exchange of sequence
    - Imputation and mutation discovery objectives

Other sequencing applications

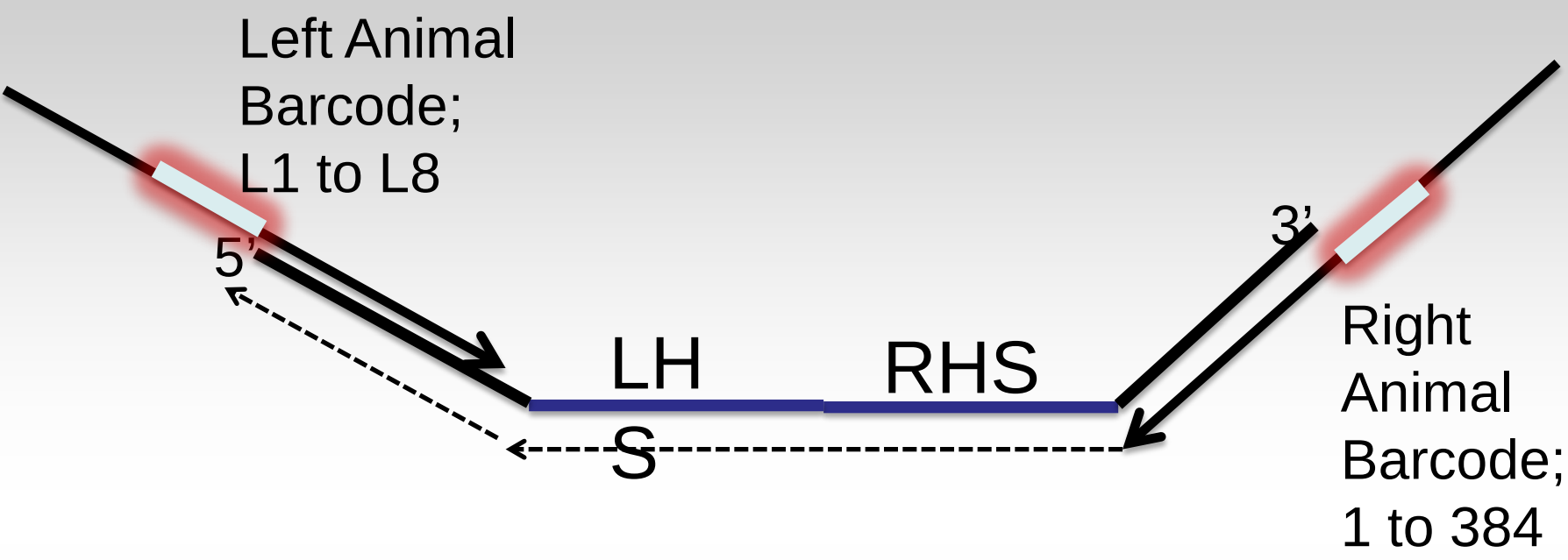
- Metagenomics
  - Loosely defined as whole genome sequencing of all organisms in a sample/population
  - Microorganisms
  - Gut, soil, skin, nasal mucosa, etc.
  - Whole genome cost-prohibitive
    - Often target variable genomic regions that can be used to classify microbes

Nasal microbiome



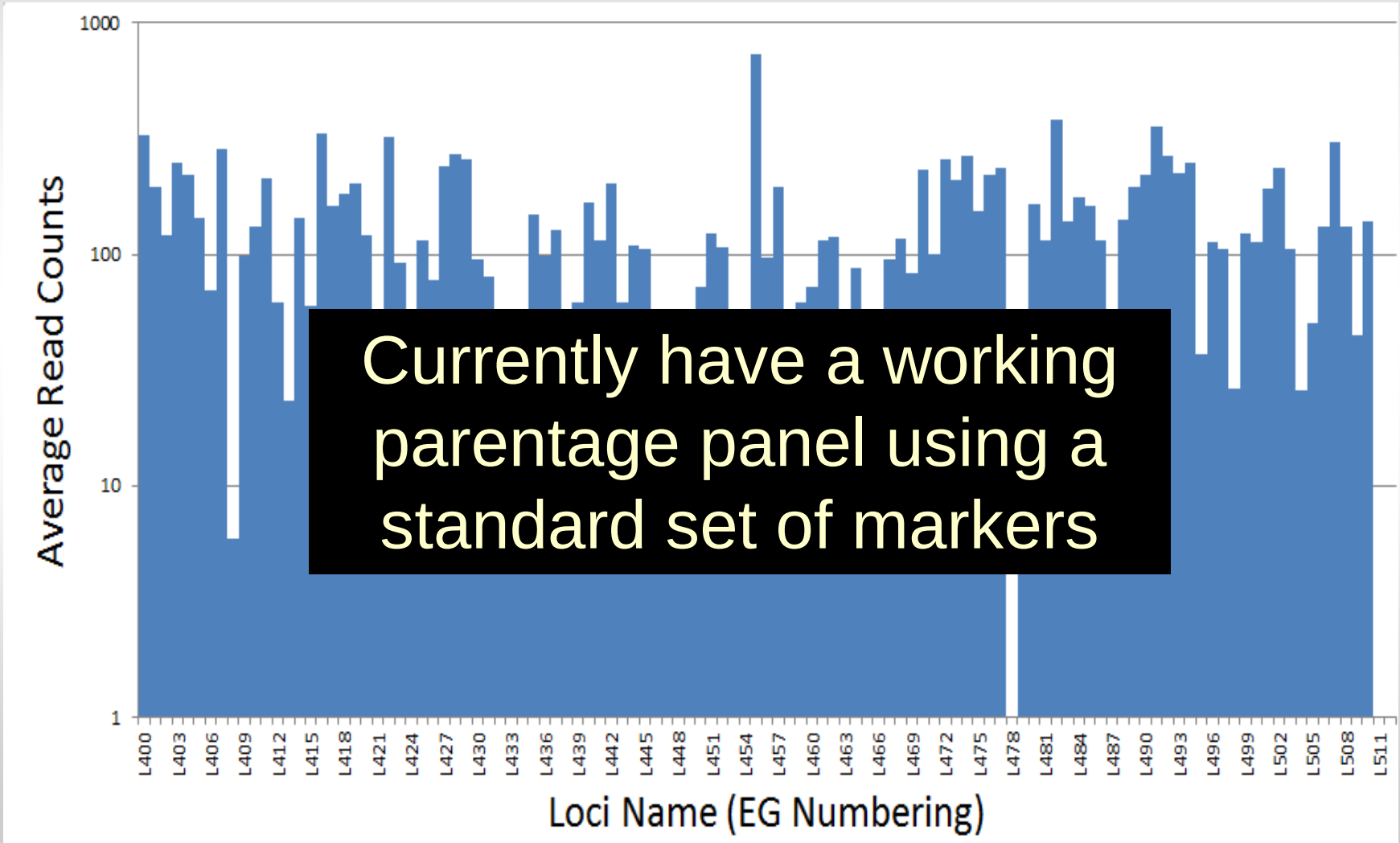
Other sequencing applications

- Using sequencing to genotype
  - Can result in significant cost savings with large numbers of animals and specific targets
  - Currently a collaboration between USMARC (Mark Thallman, Amanda Lindholm-Perry) and Eureka Genomics
  - Must be able to target and maintain identity of animal in the sample



Unlimited animal ID barcodes added with two PCR primers

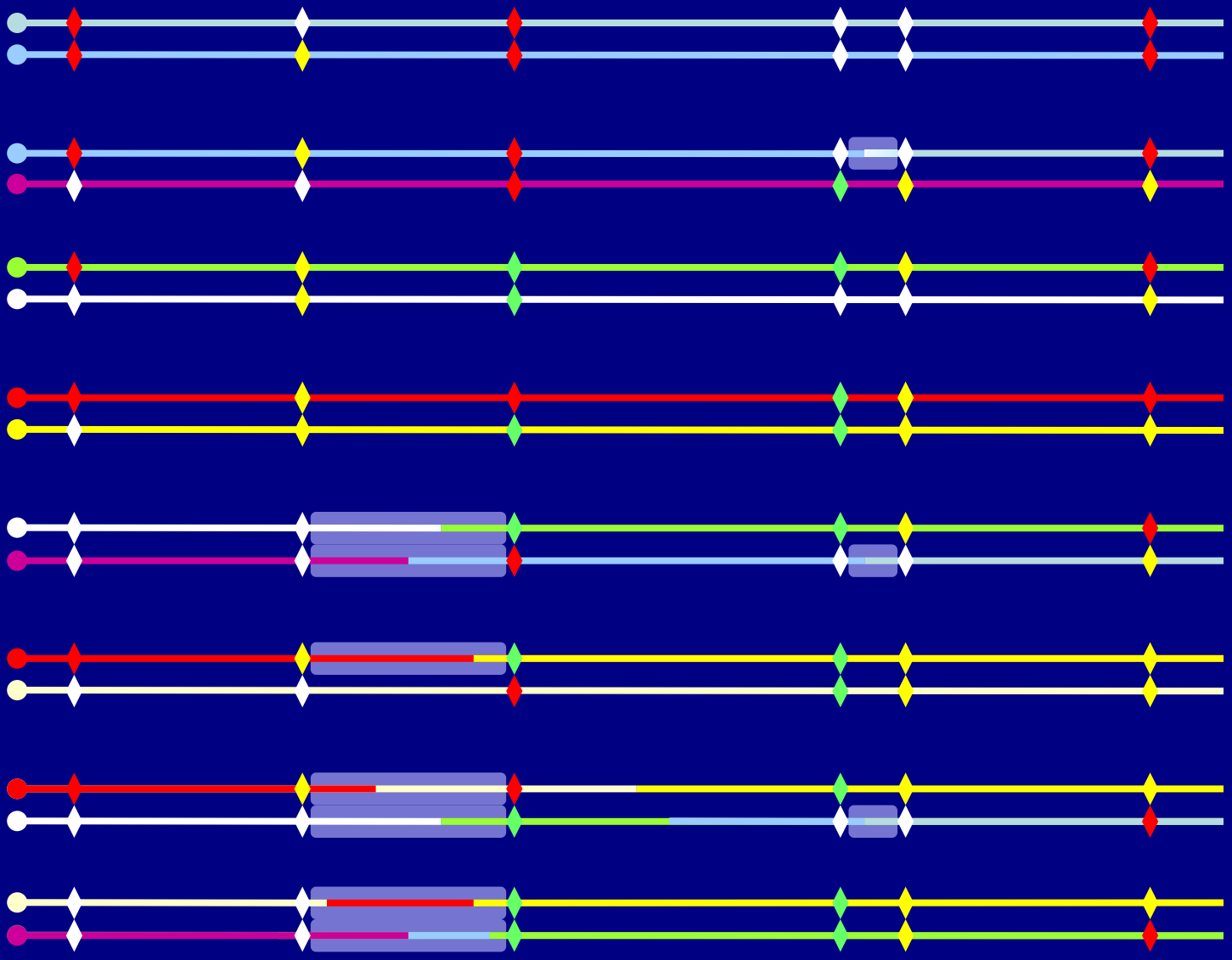
Read Counts per Locus for a Single Animal

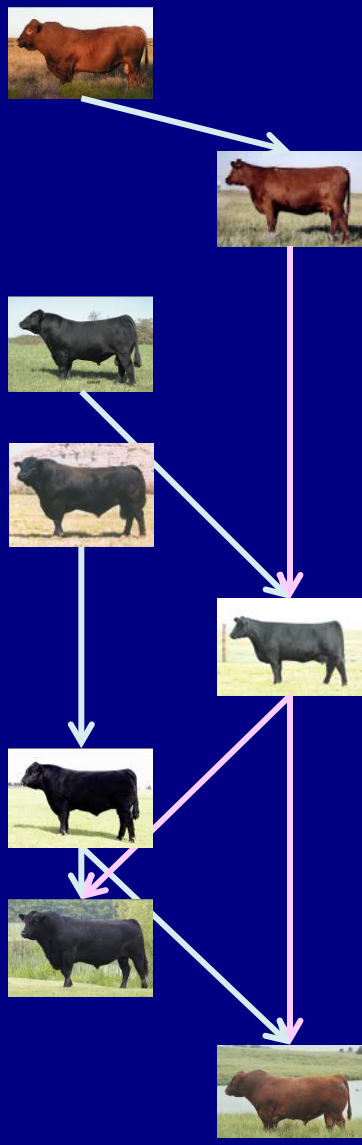


Next step

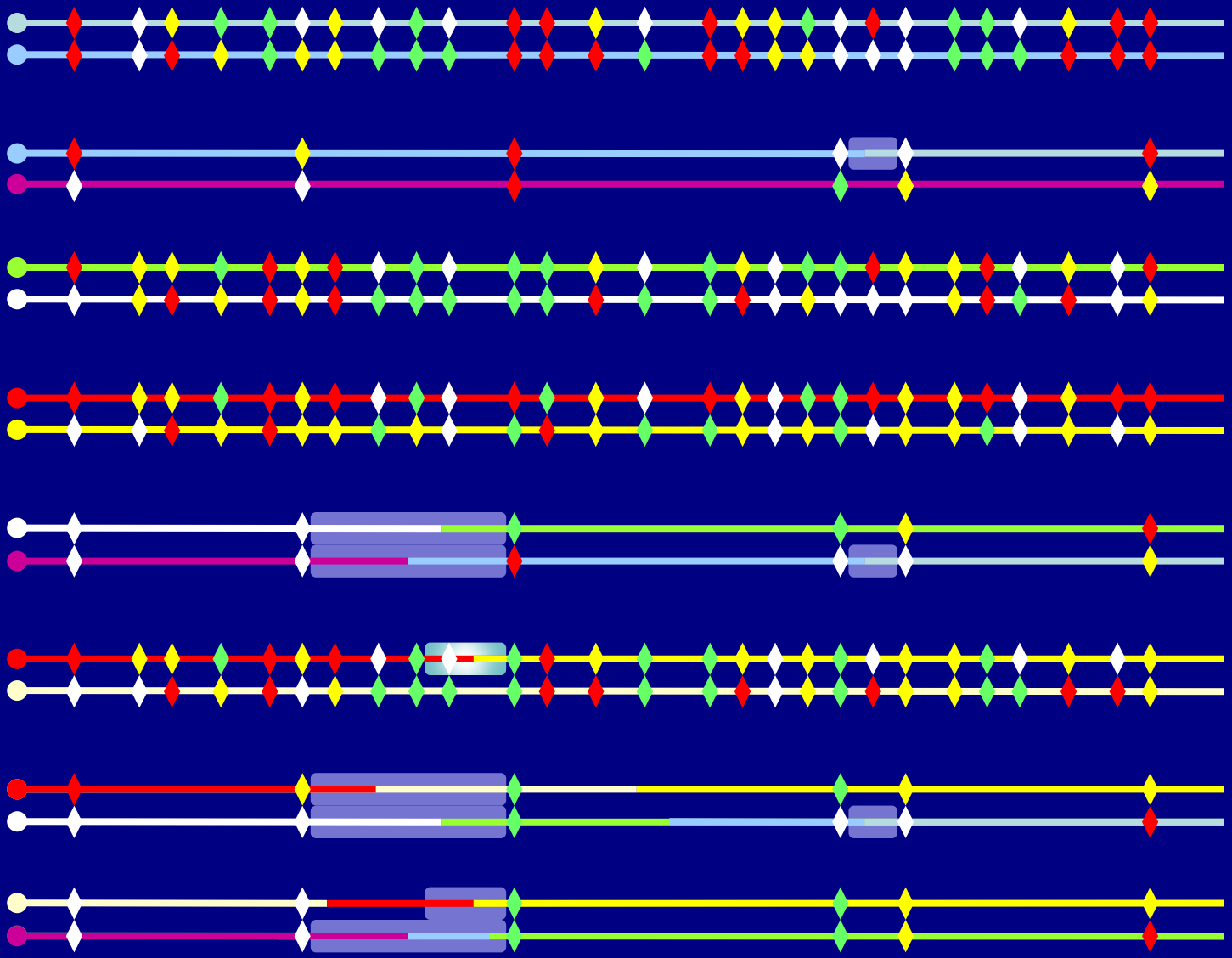
- Sparse genome scan
  - Run sparse genome scan on most or all of the animals in the breed
  - Run the 50K chip on most or all AI sires in the breed
  - Individually sequence as many of the highly influential sires in the breed as is feasible

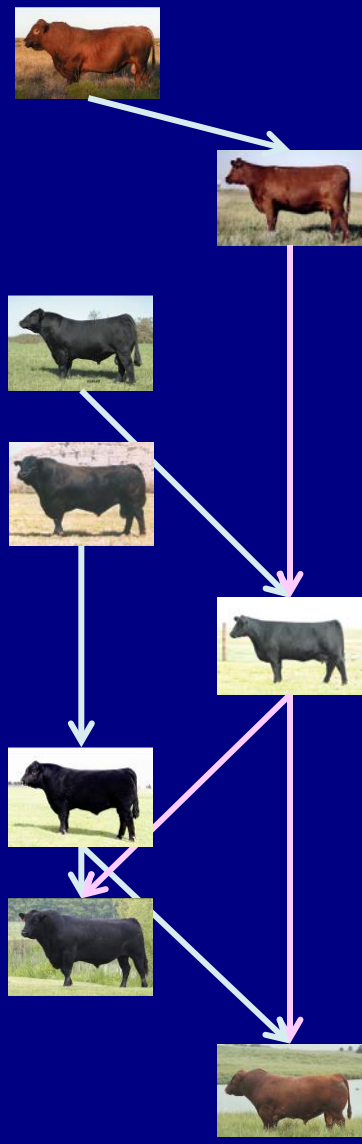
Sparse Genome Scan





Sparse Genome Scan





Summary

- Sequencing offers many possibilities
  - Move toward causal variation
  - Increase selection opportunities
  - Detect microbial interactions with economically relevant traits
  - Lower genotyping costs
- Questions