Low Pass Sequence Data in Genetic Evaluation

A joint UNL/USMARC project

Larry Kuehn, Warren Snelling,
Mark Thallman, Matt Spangler

Current genomically-enhanced EPD

- Generally based on genotyping arrays (20-100K depending on iteration)
- Inserted into EPD prediction using a single-step approach that is generally unweighted (but could be weighted)
  - May or may not be based on a reduced set
- Rarely takes advantage of functional variants or other possible causal variants

Functional variants

- Gene annotation
  - Understanding the coding regions
    - Identifying mutations that alter gene products or stop protein formation completely
    - Advances in next generation sequencing and genome annotations have significantly improved discovery of these mutations
  - Deleterious mutations that stop protein coding could certainly affect fertility
    - These and protein changing mutations could impact several trait complexes
  - First generation functional chip in cattle (F250K)

Could functional variants be more effective?

Genetic correlations between birth weight and GPE-trained birth weight MBV					
			Evaluated population		
Marker set	size	GPE h^2	SFA	Red Angus	Simmental
F250 shared with 50K	33,869	0.45	0.35	0.44	0.25
Significant GPE effects	279	0.34	0.44	0.43	0.25
LD reduced	12	0.30	0.49	0.47	0.28
NCAPG	1	0.06	0.31	0.32	0.22

- Small sets of functional variants can explain meaningful phenotypic variation within and across populations
  - depends on number and size of effects - difficult to identify variants causing small effects, especially for traits influenced by many variants with small effects

Problems with F250K

- Approximately 120,000 usable variants in USMARC populations after screening no calls, monomorphic loci, excess male calls
  - 703/5,751 loss of function remaining (651 genes)
  - 32,057/94,641 non-syn SNP (10,985 genes)
  - Around 15,000 potentially regulatory SNP
- Many genes missing – could do better

New potential

- Genotyping by sequencing with low-coverage sequencing
  - 40 to 60 million variants
  - Cost has scaled down with sequencing
    - No need for 1x coverage/animal
  - Will continue to improve with pedigree and improved reference haplotypes
  - Low-pass or skim-sequencing
  - Accuracy upward of 99% on many breeds
    - Warren Snelling will cover later

UNL/USMARC

- Current Proposal Objectives:
  - Enhancing the portability of genomic predictors
  - Increasing the accuracy of genomic predictors

- Both accomplished through evaluation of the use of low-coverage sequencing in genetic evaluation systems

Current Plan

- Through increased genotyping on UNL populations and USMARC GPE and SFA populations, evaluate accuracy gains from evaluating new marker sets from low-pass sequencing
  - Genotyping will be a combination of array and low-coverage sequencing with the opportunity to impute millions of markers through both populations

Animals

- Approximately 5,000 UNL animals/year
  - Partly an earlier Nebraska Beef Systems project
  - Includes all UNL cow herds and animals entering UNL owned feedlots

- Another 5,000 USMARC animals/year
  - Germplasm Evaluation Program (GPE)
  - Selection for Function Alleles Project (SFA)
  - Commercial populations with important phenotypes

Traits collected on GPE (UNL in red)

Calving
- Dystocia
- Survival

Growth
- Gestation Length
- Birth Weight
- Weaning Weight
- Postweaning growth
- Mature weight, height, and condition

Maternal
- Birth Weight
- Dystocia
- Survival
- Weaning Weight
- Milk Production

Carcass & Meat Quality
- Shear force
- Yield Grade factors
- Marbling
- Color Stability
- Ultrasound carcass

Efficiency
- Feed utilization of finishing steers
- Feed utilization of pre-breeding heifers
- Mature cow maintenance requirements
- Rumen microbial composition

Reproduction
- Heifer age at puberty
- AFC
- Heifer pregnancy rate
- Cow pregnancy rate
- Fetal death loss
- Postpartum interval

Longevity

Disease Resistance (IBK, BRD)

Adaptation

Analysis

- Not straightforward
  - P >>>>> N
  - Will need to design strategies that give prior weighting to different marker types (e.g., functional variants, regulatory variants)
  - Plan includes funding for research support

- Mark Thallman will cover some initial ideas

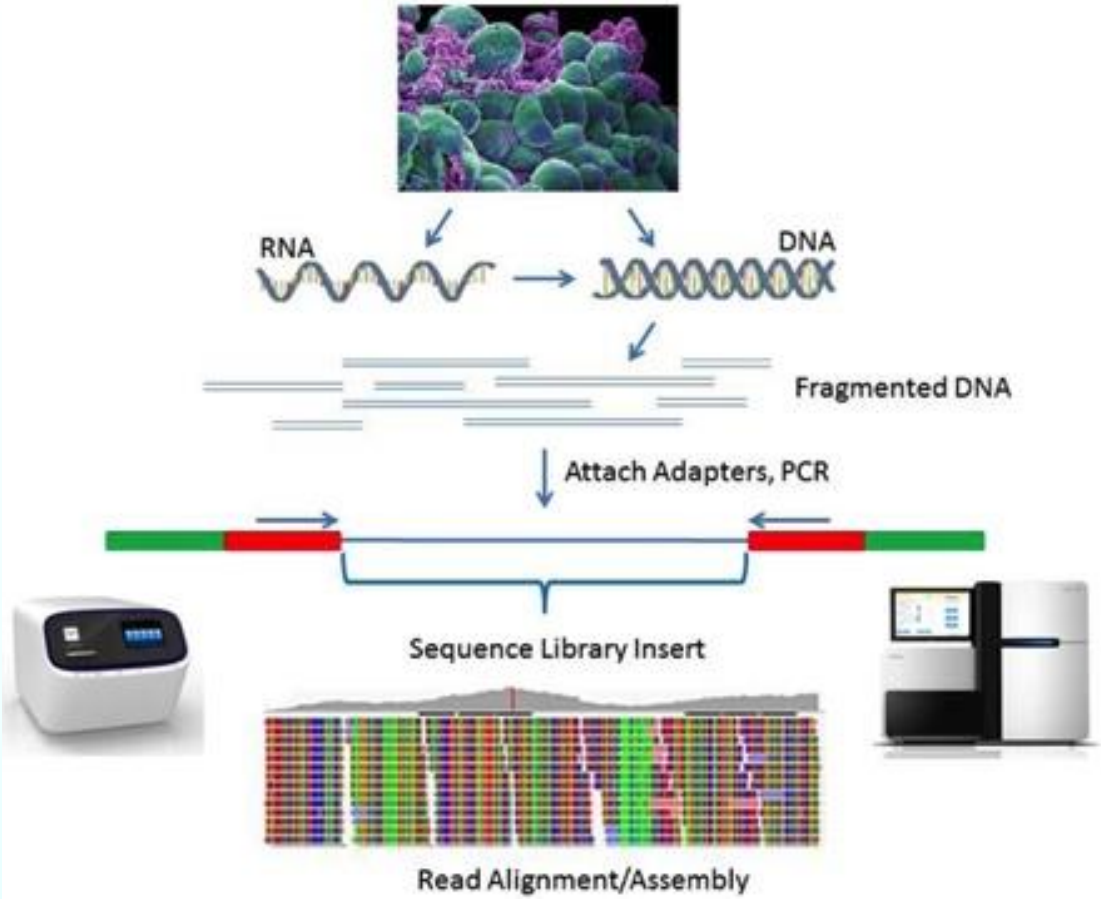
Byproducts

- Potential for GWAS of some novel traits
  - Extension of novel traits to genetic evaluation will depend on success of weight traits
    - Primary goal is increasing utility of genetic evaluation
    - Most important strategy is to help make novel traits less novel
- Understanding of imputation and storage requirements for low-coverage sequence
  - Will help with implementation in genetic evaluation service providers

Low-pass sequence data in genetic evaluation

Genome sequencing

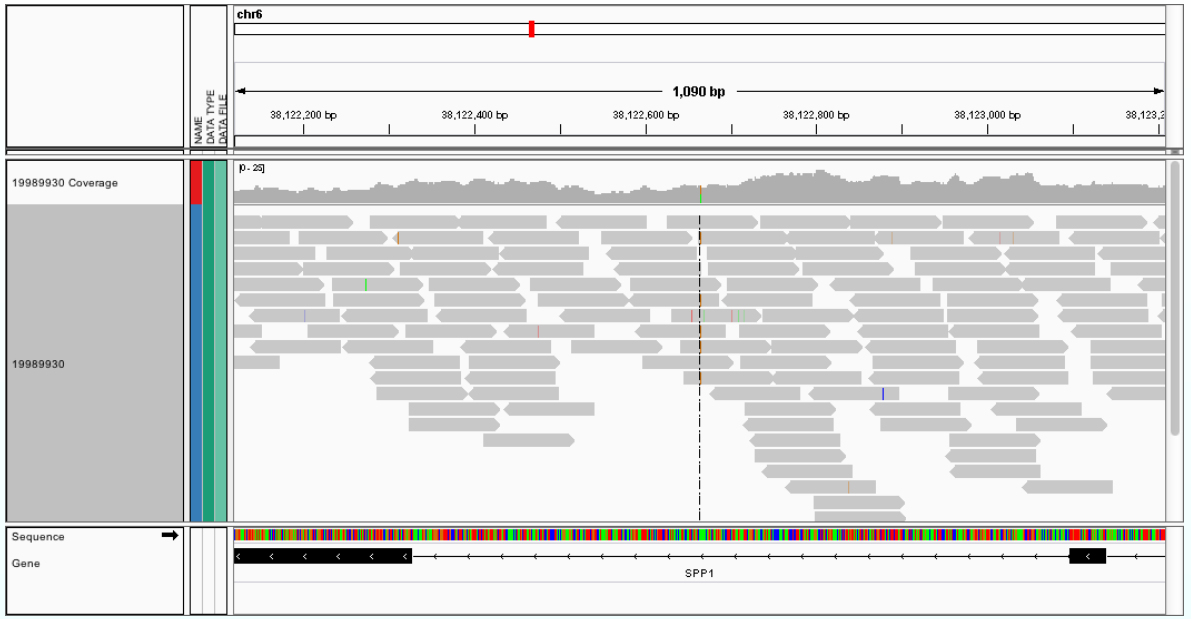
- cannot read chromosome sequence from end to end
- can read fragments
  - 50-300 bp short reads
  - 5-20 Kbp long reads
- random process
  - "library" of randomly fragmented DNA
  - read ends of fragments
  - align reads to reference assembly



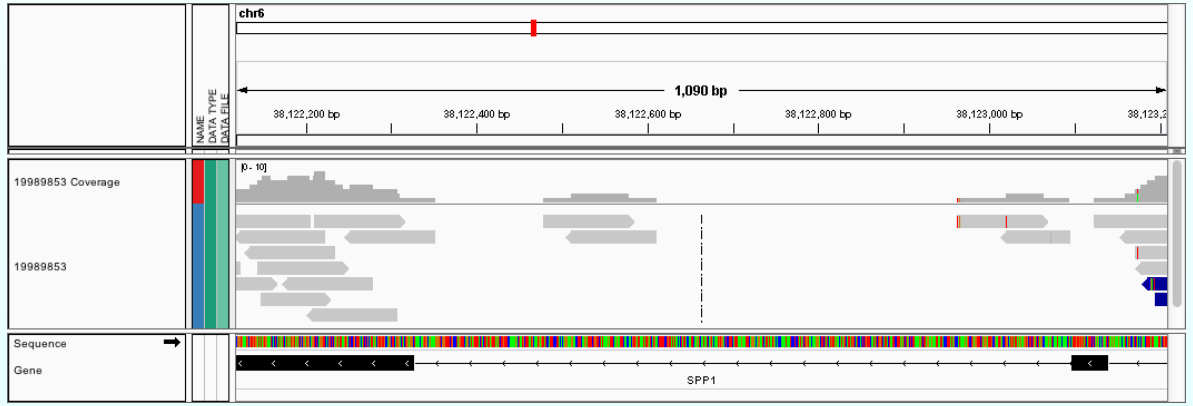
Head et al., 2014 BioTechniques 56:61-77

Genome coverage

- x = bases read / genome length
- substantial variation around average coverage
- portion of genome read increases with coverage



10x



2.5x

using low-pass (<2x) sequence

- variant discovery
  - similar cost and effort to sequence many individuals at low coverage or few individuals at high coverage
    - broader sampling to detect sequence variation in population

A survey of polymorphisms detected from sequences of popular beef breeds[1,2,3]

W. M. Snelling,*[4] G. L. Bennett,* J. W. Keele,* L. A. Kuehn,*
T. G. McDaneld,* T. P. Smith,* R. M. Thallman,* T. S. Kalbfleisch,† and E. J. Pollak*

*USDA, ARS, U. S. Meat Animal Research Center, Clay Center, NE 68933; and †Department of Biochemistry and Molecular Biology, School of Medicine, University of Louisville, Louisville, KY 40202

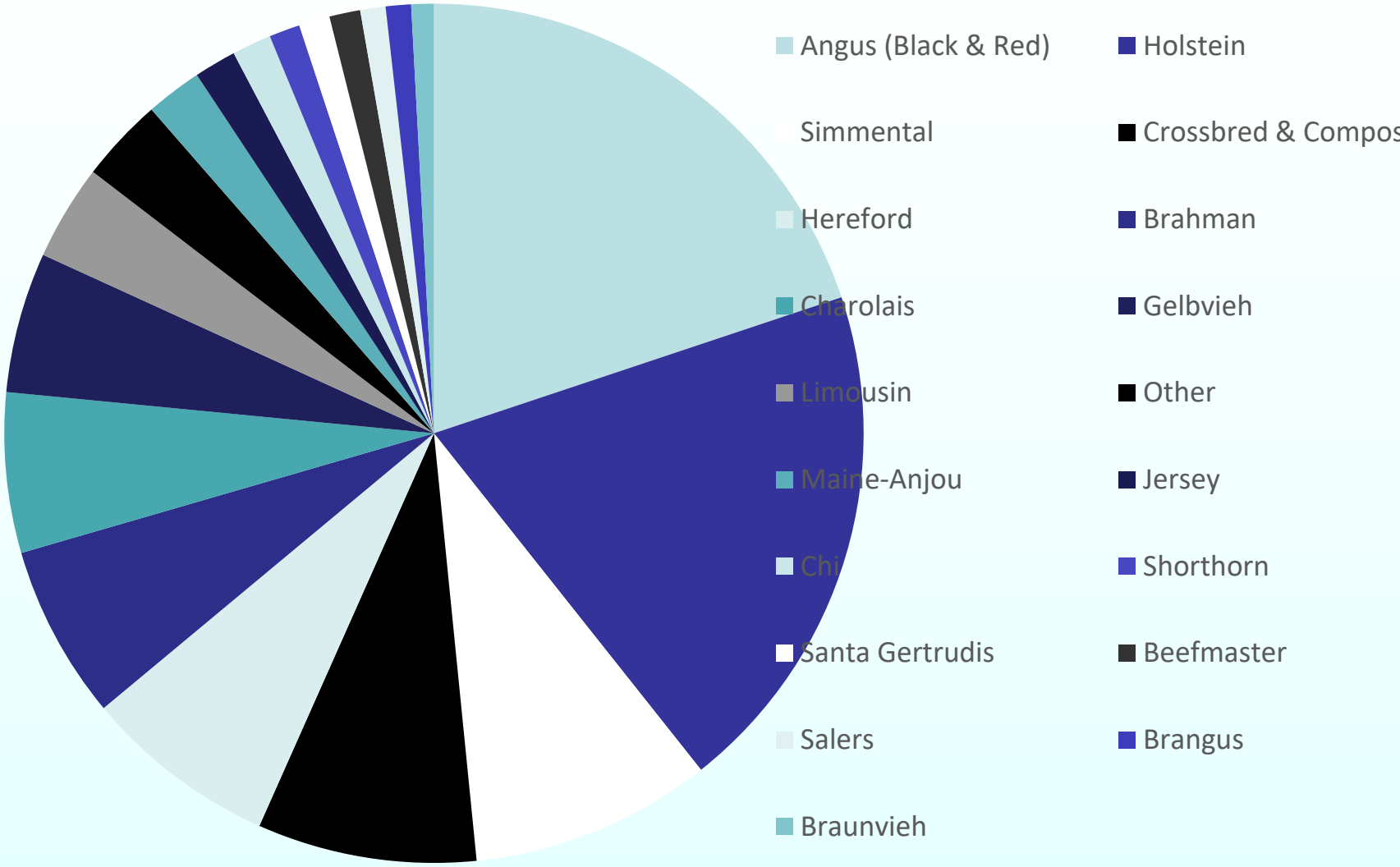
270 bulls, 28.8 million variants, 158,000 interesting variants

using low-pass sequence

- genotyping?
  - low direct call rate
    - few sites covered by enough reads to call genotype from sequence
    - little overlap among sites called from different samples
  - imputation – match low-coverage reads to reference haplotypes
    - genotypes imputed for all variants detected in reference
    - lower per-sample costs than deep sequence or genotyping arrays for human GWAS
      - Li et al., 2011; Pasanuic et al., 2012; Gilly et al., 2018

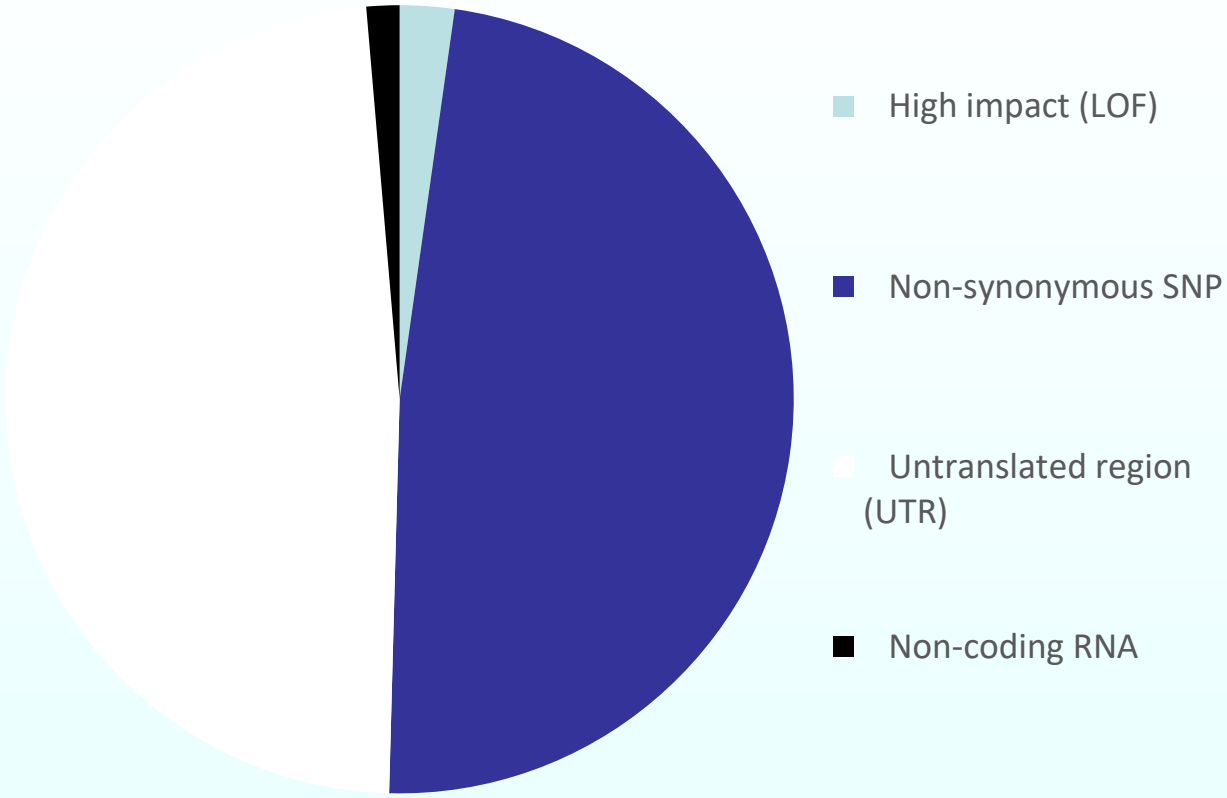
Gencove imputation – reference panel

- 947 cattle with > 4X



Gencove imputation – reference panel

- 59,198,025 variants
- 660,071 interesting
  - change or regulate proteins



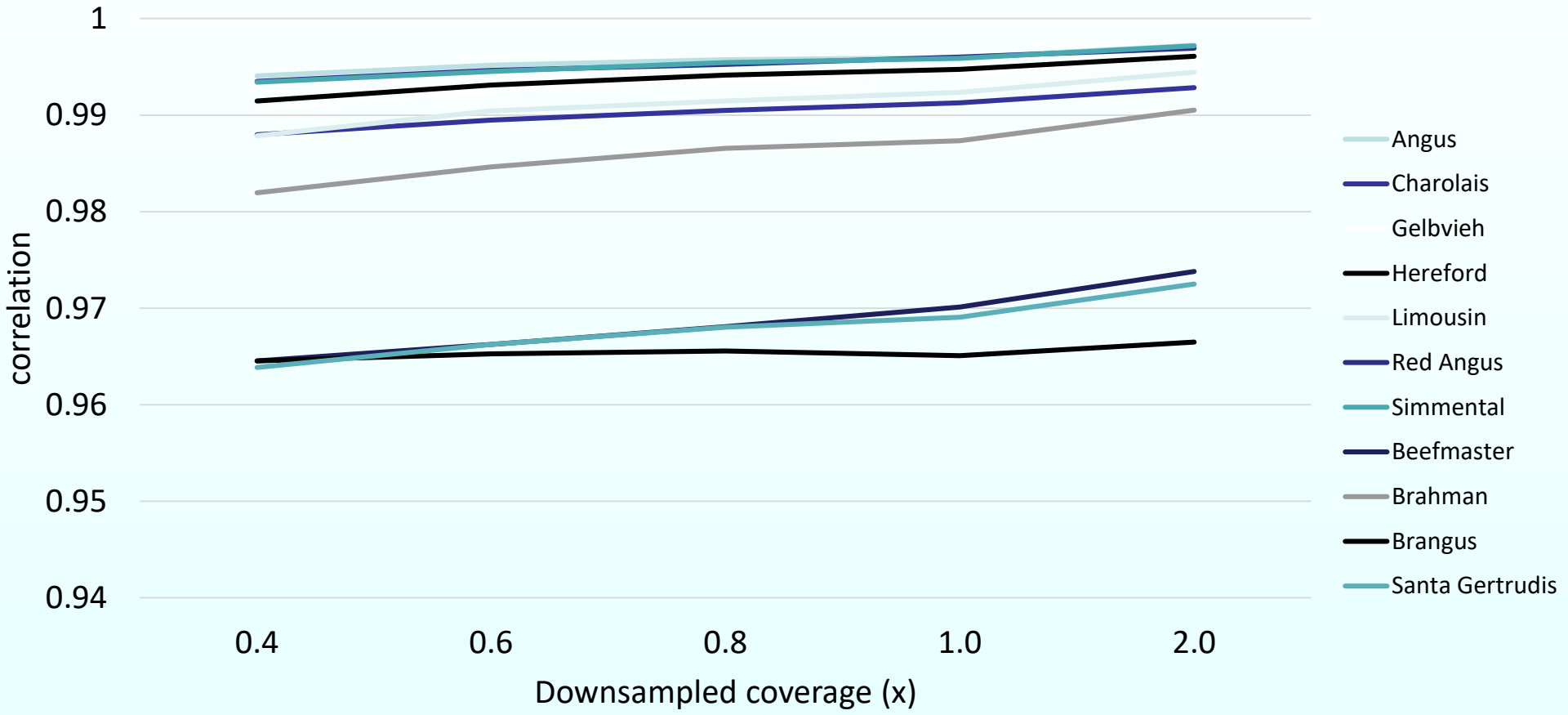
GPE sequence – Gencove imputation

Evaluate low-pass by downsampling

- mimic low-pass sequencing by sampling reads from deeper sequence
- GPE sires
  - one bull from each Cycle VII breed, Brahman, indicus-influenced composites
  - > 4x downsampled to 0.4x, 0.6x, 0.8x, 1x, 2x
- Feed efficiency steers
  - 79 steers with extreme intake or gain
  - ~ 10x downsampled to 1x

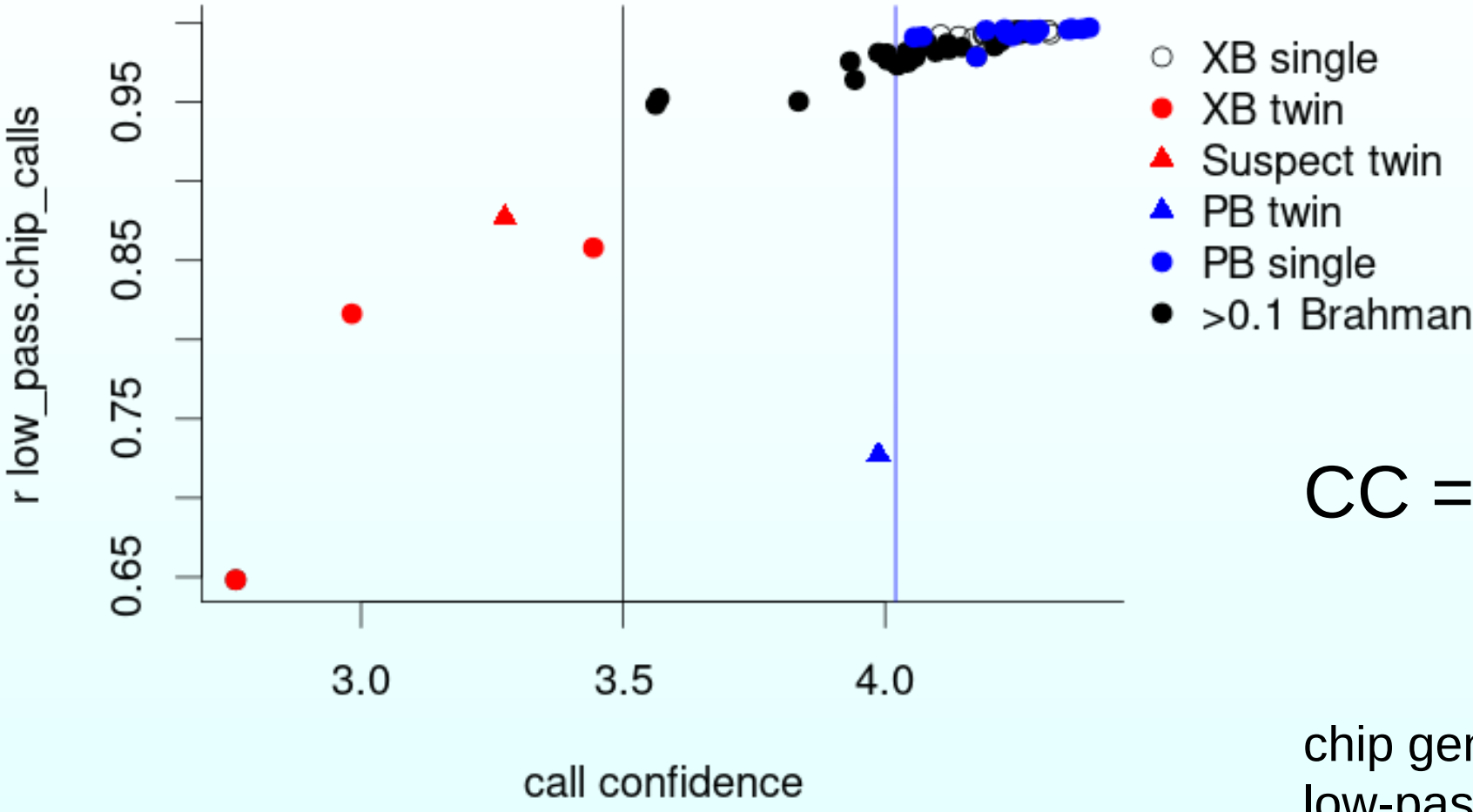
GPE sire sequence – Gencove imputation

Agreement between BovineHD and genotypes imputed from downsampled sequence



GPE steer sequence – Gencove imputation

”Call Confidence”, based on imputed genotype probabilities, indicates agreement between chip and imputed genotypes



$$CC = \text{mean}(-\log_{10} (1-GP_{max}) \text{ for } GP_{max} < 1$$

chip genotypes from twin ear notch
low-pass sequence from twin blood

GPE steer sequence – Gencove imputation

Genomic prediction

- (G)BLUP including all steer records
  - pedigree BLUP without genotypes
  - genomic BLUP with available chip genotypes
    - pedigree used to impute lower density chips to BovineHD + F250
- Marker effects for steer MBV trained by GPE without steer data
  - MBV from marker effects applied to chip genotypes and genotypes imputed from downsampled sequence

GPE steer sequence – Gencove imputation

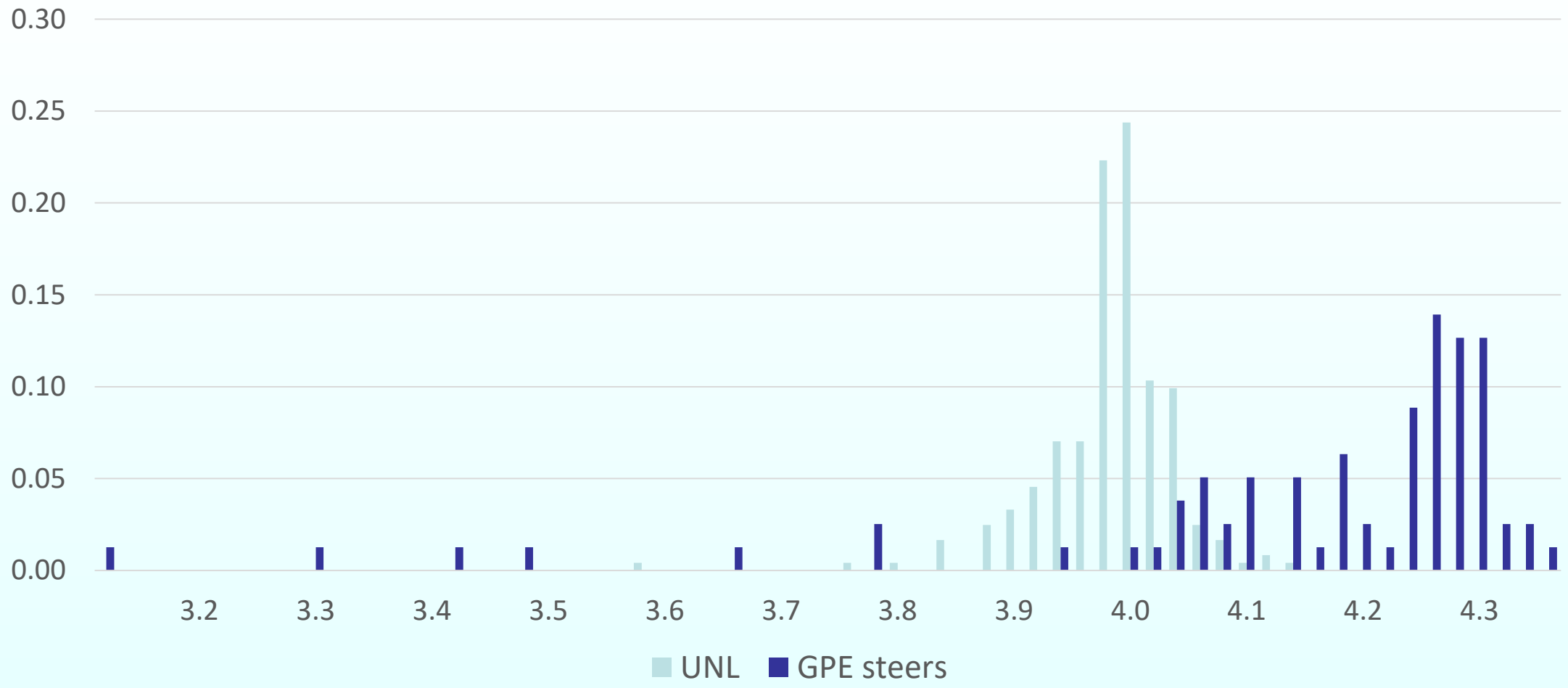
Correlations between steer EBV and MBV

		Birth weight		PWG		Marbling score	
	MBV	BLUP	GBLUP	BLUP	GBLUP	BLUP	GBLUP
Chip	F250[a]	0.73	0.90	0.78	0.88	0.77	0.93
	F250s[b]	0.56	0.68	0.65	0.71	0.66	0.75
	50K[c]	0.71	0.89	0.79	0.89	0.79	0.95
Seq	F250	0.71	0.88	0.77	0.88	0.75	0.91
	F250s	0.54	0.64	0.63	0.71	0.59	0.69
	50K	0.70	0.84	0.80	0.90	0.76	0.93

[a] 116,472 (102,931) functional variants from F250; [b] 551 to 698 (532 to 668) selected functional variants;
[c] 51,496 (48,573) variants shared by F250 and BovineHD

UNL low-pass sequence – Gencove imputation

Call confidence distribution



low-pass sequencing & imputation

- current results suggest sequence variant genotypes can be accurately imputed from low-coverage sequence
  - accuracy is not perfect, but imperfect accuracy recognized by genotype probabilities

- genotype calls for comprehensive set of known sequence variants
  - 50K, HD, functional variant panels can be extracted
  - eventually replace 50K with variants more likely to affect phenotypic variation
    - reduce dependence on LD between 50K & QTL
    - enable more accurate genomic predictions across breeds, crosses, generations

low-pass sequencing & imputation

- cost competitive with existing SNP chips
  - encourage complete genotyping
    - reduce bias in genetic evaluations due to selective genotyping
  - justify genotyping commercial calves
    - incorporate commercial data into genetic evaluation
    - genomic predictions to support calf management and marketing decisions
- Imputation from low-coverage sequenced can avoid chip-related issues
  - probe design and manufacturing costs
  - large sample size needed to train genotype calls
  - limited shelf-life

low-pass sequencing & imputation

Concerns and future work

- rare defect variant genotypes
  - reference panel needs to include known defect carriers
- “gaps” in reference panel
  - industry cattle with weak relationships to reference panel – low accuracy imputation
    - need systematic approach to identify and fill gaps with informative haplotypes
- imputation from chip genotypes to sequence variants
  - leverage existing genotypes

Acknowledgments

iGENOMX

Paul Doran
Keith Brown

NEOGEN

Stewart Bauck
J R Tait
Ben Pejsar

gencove

Joe Pickrell
Jeremy Li
Jesse Hoff
Tomaz Berisa

USDA

Entire crew involved with GPE, tissue sampling & repository, sequencing, …
(too many to name)

Opportunities for Low-pass Sequencing of Pedigreed Populations and How it May Fit into Genomic Evaluation

Mark Thallman

Premises of Current Genomic EPDs

- Markers are spread evenly across the genome at intermediate frequencies or are selected from sets of such markers
- Assume some markers may directly affect traits, but most do not
- Assumes causative variation is closely associated with markers
- All genotyped animals either have, or can be imputed to a common set of markers
- Current genomic predictions are more accurate than predictions without genomics

Challenges in Current Genomic EPDs

- Some, but limited, increase in accuracy available from improving utilization of the markers on current chips
- Limited increase in accuracy available from increasing number of markers on chip of same type as are on current chips
- The high-hanging fruit is causative variation not on current chips that often has low minor allele frequency
  - There are millions of candidates and only limited opportunities for prioritizing them without having genotypes to evaluate effects
  - Nonetheless, Warren has shown benefits of screening putative functional variants from a relatively small subset of the entire pool of such variants

Approach to Improve Genomic Prediction Accuracy

- Sequence influential bulls
  - Discover SNP
  - Impute sequence to descendants using chip genotypes
  - Identify most promising sequence variants to improve accuracy
    - Use functional information and preliminary associations with traits
  - Develop new chips that include the promising new variants
  - Determine which promising variants appear most predictive
  - Include most predictive variants in genomic prediction models and future chips
  - Repeat
- If this looks hard, that's because the high hanging fruit is most of what is left to do and it is hard.
  - But, Matt Spangler calls this iterative redesign of chips "untenable" when considered in the context of low-pass sequencing as an alternative.

Goals of Low-pass Sequencing

- Sequencing a random sample of the genome of an animal in lieu of genotyping a specific set of markers
  - Short term goal is to impute to the standard set of markers used in current analyses at cost competitive with genotyping
  - Intermediate goal is to identify markers that are more predictive of important traits
  - Long-term goal is to replace genotyping by imputing entire population to full genomic sequence

Comparison:

Chip Genotyping

- High accuracy without imputation
- High call rate without imputation
- If genotype called, get both paternal and maternal alleles
- Focused on genotypes
- Mature technology

Low-pass Sequencing

- Accuracy depends on imputation
- Call rate depends on imputation
- May impute paternal allele, but not maternal (or vica versa)
- Focused on haplotypes
- In early stages of development

Concerns Over Low-pass Sequencing

- How will it integrate with existing SNP chips and the subsets of SNP used in current genetic evaluations?
  - Warren showed it is feasible (within limits)
- Will genetic defects and other “must have” variants (e.g., polled, color) be reported reliably?
  - Several approaches available to enhance representation in the library
- Requires imputation to produce a useful result
  - Imputation is already part of genomic evaluation pipeline
- Requires more sophisticated imputation than SNP chips
  - Warren showed it is feasible
- Will it work for parentage determination?
  - SNP chips are great for parentage determination, but low-pass will be far superior, extending into pedigree reconstruction

So, why consider low-pass sequencing?

- It will make the process of SNP discovery, promising variant identification, adding to evaluation, validating in field data, dropping dropouts, returning to SNP discovery, and repeating far more seamless, continuous, and less time consuming than iteratively redesigning SNP chips.
- Current cost is somewhat greater than that of 50K chips.
- Cost may decrease to below SNP chips.
- SNP discovery will be far more thorough than if it is limited to higher coverage of relatively few influential bulls.

Information from Sequence Compared with 50K Chip

Yellow represents locations of markers on 50K Chip. There are about 60,000 bases between them.

Only positions in yellow can be observed through the chip

Blue represents locations of variable bases that affect an important trait. We generally don't know how many or where they are.


...TC-G-C-C-T~A-GA-A-T-T~A-C-T-T-G~GT-A-A-T-G...
...TC-A-A-C-G~G-GT-A-T-A~G-C-C-C-A~GA-A-C-C-A...


...TA-G-C-C-G~G-CT-T-T-T~G-G-C-T-A~GA-G-C-T-A...
...CC-A-C-G-G~G-GT-A-G-T~G-C-C-T-G~CA-A-C-C-G...


...TC-A-A-C-G~G-GT-A-T-A~G-C-C-C-A~GT-A-A-T-G...
...TA-G-C-C-G~G-CT-T-T-T~G-G-C-T-A~GA-G-C-T-A...


...TC-A-A-C-G~G-GT-A-T-A~G-C-C-C-A~GA-A-C-C-A...
...CC-A-C-G-G~G-GT-A-G-T~G-C-C-T-G~CA-A-C-C-G...


...TA-G-C-C-G~G-CT-T-T-T~G-C-C-C-A~GT-A-A-T-G...
...CC-A-C-G-G~G-GT-A-G-T~G-C-C-T-G~CA-A-C-C-G...

Letters represent variable positions in the genome

"-" represent stretches of constant bases that do not vary in cattle. They could be from 1 to >1,000 bases (about 50 on average)

"~" represent stretches of constant and variable bases too long to represent in detail in the diagram (generally > 10,000 bases)

"…" represent the remainder of the chromosome to the right (or left) of this region (average about 50,000,000 bases)

A Few Cautions About the Example

- If you are watching the recording at your own pace for a deeper understanding of the concepts:
  - This is a contrived example intended to illustrate a few key concepts
  - The frequencies of errors, uncalled sequence, informative sequence reads, and crossovers are therefore higher than might occur in practice
    - All of these are concentrated in a few very short stretches of sequence in order to illustrate concepts associated with them
  - The example assumes no sequencing errors and mutations and obscures many of the other complexities of real data, including determining phase and grandparental origin
  - The example uses over-simplified logic including single base exclusions and matches
    - It is **not** representative of any algorithm that would be used in practice

Low-Pass Sequencing Reads

... C-G- ~ A-A-T- ~A-C-T- ~ -A-A-T- ...
... -A-A- ~G-GT- ~ -C-C-A~ -C-C-A...

...TA-G- ~ -T-T-T~ -G-C-T-A~GA-G- ...
... C-A-C-C- ~ -GT-A- ~ -C-T-G~ -A-C-C-G...

...TC-A- ~ T-A-T- ~G-C-C- ~ -A-T-G...
... A-G-C- ~G-CT- ~ -G-C-T- ~ -G-C-T- ...

Reference Haplotype Imputation of Low-Pass

1 ...TC-G-C-C-T~A-GA-A-T-T~A-C-T-T-G~GT-A-A-T-G...
2 ...TC-A-A-C-G~G-GT-A-T-A~G-C-C-C-A~GA-A-C-C-A...
3 ...TA-G-C-C-G~G-CT-T-T-T~G-G-C-T-A~GA-G-C-T-A...
4 ...CC-A-C-C-G~A-GT-A-G-T~A-G-C-C-G~CA-G-A-C-G...

Ø ...CC-A-C-G-G~G-GT-A-G-T~G-C-C-T-G~CA-A-C-C-G...

Cow's maternal haplotype (not included in reference haplotype panel)

These sequences match Haplotype 4, so surrounding sequence is imputed to it

3 3 3 3
TA-G- -T-T-T -G-C-T-A GA-G-
C-A-C- -GT-A- -C-T-G -A-C-C-G
4 4 None None

These sequences do not match any haplotype in reference, so surrounding sequence is missing

TA-G-C-C-G~G-CT-T-T-T~G-G-C-T-A~GA-G-C-T-A
CC-A-C-**C**-G~**A**-GT-A-G-T~.-.-C-T-G~..-A-C-C-G

2 Imputation errors due to cow's maternal haplotype not being included in reference panel

Dots represent bases that cannot be imputed unambiguously

Add Sparse Coverage of Descendants

...TC-G-C-C-T~A-GA-A-T-T~A-C-T-T-G~GT-A-A-T-G...
...TC-A-A-C-G~G-GT-A-T-A~G-C-C-C-A~GA-A-C-C-A...

...TA-G-C-C-G~G-CT-T-T-T~G-G-C-T-A~GA-G-C-T-A...
...CC-A-C-**C**-G~**A**-GT-A-G-T~.-.-C-T-G~..-A-C-C-G...

...TC-A-A-C-G~G-GT-A-T-A~G-C-C-.-.~G.-A-A-T-G...
...TA-G-C-C-G~G-CT-T-T-T~G-G-C-T-A~GA-G-C-T-A...

... -GT-A-T-A ...
... G-GT-A-G- ...

... A-G-C- -C-C-A ...
... -A-C-G-G CA-A- ...

Determine Grandparental Origin of Descendants

...TC-G-C-C-T~A-GA-A-T-T~A-C-T-T-G~GT-A-A-T-G...
...TC-A-A-C-G~G-GT-A-T-A~G-C-C-C-A~GA-A-C-C-A...

...TA-G-C-C-G~G-CT-T-T-T~G-G-C-T-A~GA-G-C-T-A...
...CC-A-C-**C**-G~**A**-GT-A-G-T~.-.-C-T-G~..-A-C-C-G...

...TC-A-A-C-G~G-GT-A-T-A~G-C-C-.-.~G.-A-A-T-G...
...TA-G-C-C-G~G-CT-T-T-T~G-G-C-T-A~GA-G-C-T-A...

... -GT-A-T-A ...
... G-GT-A-G- ...

... A-G-C- -C-C-A ...
... -A-C-G-G CA-A- ...

Fill Non-Recombinants with Parental Haplotypes

...TC-G-C-C-T~A-GA-A-T-T~A-C-T-T-G~GT-A-A-T-G...
...TC-A-A-C-G~G-GT-A-T-A~G-C-C-C-A~GA-A-C-C-A...

...TA-G-C-C-G~G-CT-T-T-T~G-G-C-T-A~GA-G-C-T-A...
...CC-A-C-**C**-G~**A**-GT-A-G-T~.-.-C-T-G~..-A-C-C-G...

...TC-A-A-C-G~G-GT-A-T-A~G-C-C-.-.~G.-A-A-T-G...
...TA-G-C-C-G~G-CT-T-T-T~G-G-C-T-A~GA-G-C-T-A...

...TC-A-A-C-G~G-GT-A-T-A~G-C-C-C-A~GA-A-C-C-A...
...CC-A-C-**C**-G~G-GT-A-G-T~.-.-C-T-G~..-A-C-C-G...

...TA-G-C-C-G~G-.T-.-T-.~G-.-C-C-A~G.-A-A-T-G...
...CC-A-C-G-G~G-GT-A-G-T~.-.-C-T-G~CA-A-C-C-G...

Impute from Progeny to Parents


...TC-G-C-C-T~A-GA-A-T-T~A-C-T-T-G~GT-A-A-T-G...
...TC-A-A-C-G~G-GT-A-T-A~G-C-C-C-A~GA-A-C-C-A...


...TA-G-C-C-G~G-CT-T-T-T~G-G-C-T-A~GA-G-C-T-A...
...CC-A-C-.-G~.-GT-A-G-T~.-.-C-T-G~CA-A-C-C-G...


...TC-A-A-C-G~G-GT-A-T-A~G-C-C-C-A~G.-A-A-T-G...
...TA-G-C-C-G~G-CT-T-T-T~G-G-C-T-A~GA-G-C-T-A...

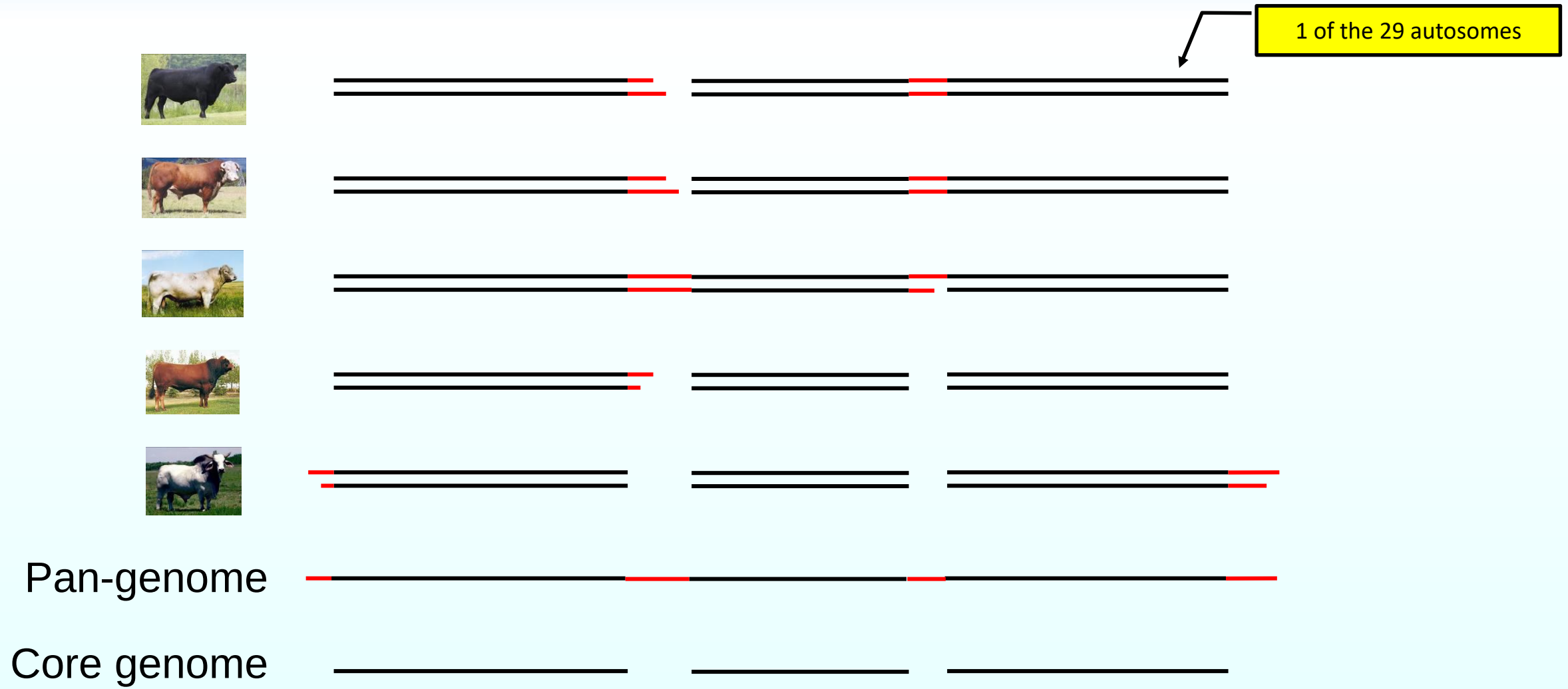

...TC-A-A-C-G~G-GT-A-T-A~G-C-C-C-A~GA-A-C-C-A...
...CC-A-C-.-G~.-GT-A-G-T~.-.-C-T-G~CA-A-C-C-G...


...TA-G-C-C-G~G-.T-.-T-.~G-.-C-C-A~G.-A-A-T-G...
...CC-A-C-.-G~.-GT-A-G-T~.-.-C-T-G~CA-A-C-C-G...

Summary of Imputation Approaches

- Off-the-shelf low-pass works amazingly well
- It could work better combined with pedigree imputation
- It could be less expensive with pedigree imputation
- The advantages of pedigree imputation are far greater if the entire herd or population is sequenced than if just a select few
- Low-pass captures far more genetic variation than current chips can

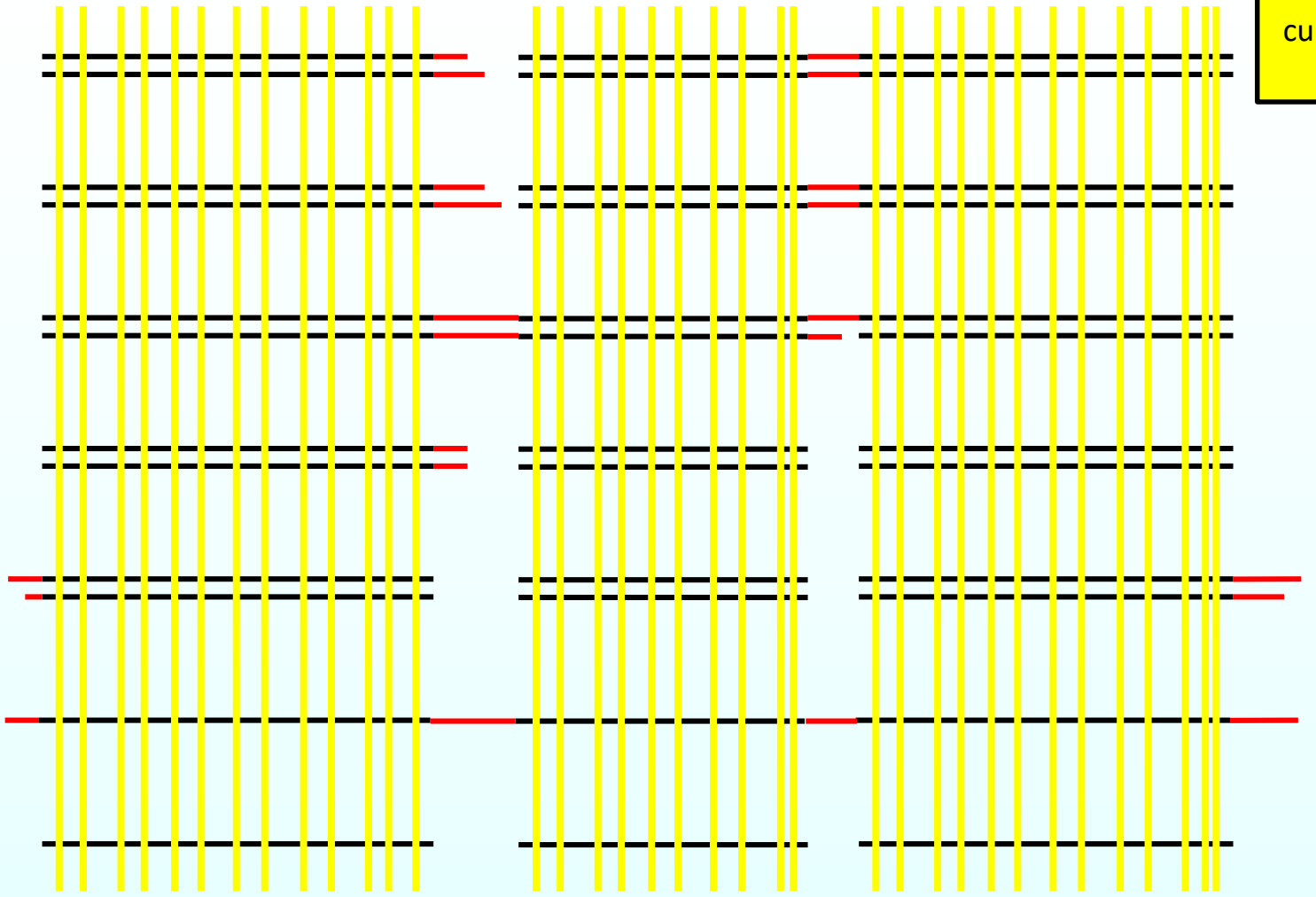
Structural Variation in Genomes



Structural Variation in Genomes

Pan-genome

Core genome



Yellow lines represent chip markers. Because they are selected for high call rate, almost all markers on current chips are probably in the core genome

Structural Variation in Genomes

- We are just getting started in cattle
- There is much more we don’t know than we do know
- We do know some genes that vary in copy number
- It seems likely there are at least some genes that are expressed in some animals and absent in others
  - Such genes seem likely to contribute to functional variation
- It is likely to account for a substantial amount of the “missing heritability”
- It is detected much more effectively through long-read technology than with the short reads used in low-pass
- Once detected and added to reference haplotypes, it should be feasible to impute structural variation with short-read low-pass sequence generated now

Implementation of Low-Pass in the Germplasm Evaluation (GPE) Population

- Have sequenced 397 sires influential in GPE comprising 20 breeds at 2X-4X depth
  - Contribute to reference haplotypes, along with other sources
  - Much of that sequence is on sire-son pairs to enhance haplotyping
- Have genotyped much of the GPE population with chips of various densities
- Have prioritized 3,000 animals for low-pass and thousands of others for additional low density chips
  - Animals designated for low-pass are those expected to fill the most holes in the reference haplotypes
- Evaluate quality of imputation
- Do additional sequencing to fill most important holes
- Develop analyses to utilize the imputed sequence data to identify predictive markers not on the chips and improve genomic predictions

Strategy for Implementation of Low-Pass in Seedstock Breeding

- Begin with a collection of reference haplotypes
- Use low-pass instead of chips as it becomes cost-competitive or can be demonstrated to provide sufficient accuracy to justify cost
- Verify that concerns listed above are addressed
- Evaluate quality of imputation and accuracy of prediction
- Collect additional sequence on individuals that would most effectively fill the most important holes in the reference sequence

What Might Genomic Evaluation Look Like With Low-Pass Sequencing?

- Short-term
  - Keep current marker sets and models until low-pass comprises a substantial proportion of the data
  - Monitor quality of imputed genotypes for those markers

What Might Genomic Evaluation Look Like With Low-Pass Sequencing?

- Intermediate term
  - Identify and sequence influential ancestors which, if low-pass sequenced, would provide imputed (through chip genotypes) sequence to the greatest number of phenotyped individuals
  - Use non-production genetic evaluation runs to continuously screen for variants not in the model that have greatest predictive ability
  - Continuously, but gradually, add loci with greatest predictive ability to the production model and drop those that are least predictive
    - Include loci outside core genome
    - Functional and putative regulatory SNP weighted higher than intergenic SNP
  - Impute the genotypes of loci in the production genomic evaluation model not included on chips back to animals genotyped only with chips

What Might Genomic Evaluation Look Like With Low-Pass Sequencing?

- Long term
  - Perhaps an hierarchical model in which:
    - Part of model relates a haplotype layer to an unobserved gene activity layer informed by prior probabilities of variants influencing gene product function or gene expression level
    - Default assumption that variants not in immediate region of gene affect gene only through their own gene products
    - Second part of model relates gene activity layer to phenotype layer of many different traits with priors based on physiological gene networks and other concepts from systems biology
    - Gene activity layer is not trait-specific and is informed by low-pass RNA sequencing of many tissues under various conditions, proteomics, metabolomics, low-pass metagenomics, and other physiological indicator traits; low-pass RNA sequencing replaces some of coverage requirement for low-pass genomic sequence
    - Dominance and epistasis expressed at gene activity layer
    - Reduces dimensions of parameter space and incorporates many additional sources of information relative to current model in which each variant is potentially and separately related to each trait.
  - Many other possibilities

The p >> n Problem

- We have many times more marker effects (p = # parameters) than animals (n = # observations)
- It is sometimes called model overfitting
- If not accounted for, it causes predictions to appear more accurate than they are
- Many ways to deal with it; won't cover here
- This was a serious problem in the early days of genomic EPDs based on SNP chips, but has become much less of a concern as several breeds now have substantially more animals genotyped than SNP available for inclusion in the model
- As we consider selecting markers from tens of millions of candidates, p >> n reemerges.
- But, our best chance to improve accuracy is to consider all variants, so we will have to return to dealing with p >> n.

Conclusions

- In 2015, I presented a poster arguing that successful widespread utilization of low-pass sequencing was dependent on technological advances in two areas:
  - Methods for cost effective construction of sequencing libraries
  - Algorithms, data structures, and software to efficiently impute low-pass data to genomic sequence throughout populations
  - Although much work remains to be done, Warren demonstrated substantial progress on both fronts and that low-pass is competitive
- There is far more information in an incomplete and imperfect view of the majority of the genome (low-pass) than there is in a near-perfect view of a minute fraction of the genome (chips)