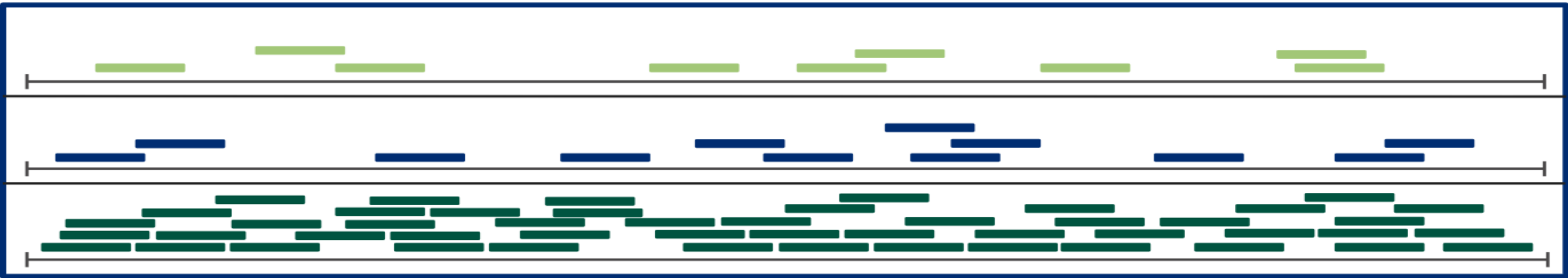
A Vision for the Future of Low-pass Sequencing

U.S. Meat Animal Research Center

R. Mark Thallman and Bailey Engle



Overall Outline

- Constraints to widespread adoption
- Representation of genomic sequence
- Imputation of lowpass from parents with imputed, phased sequence
- Hierarchical statistical model to utilize genomic sequence
  - Motivated by biological processes and molecular biology
- More ambitious opportunities

Constraints to Widespread Adoption

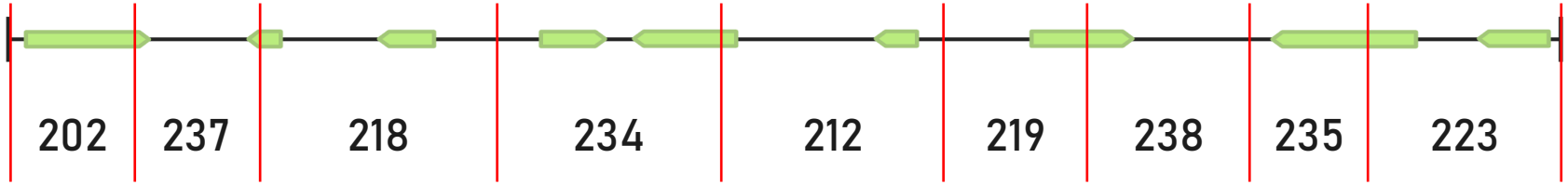
- Cost
  - Tissue sample collection
  - DNA extraction
  - Barcoded library construction
  - Sequencing (function of depth)
  - Storage of Resulting Data
- Value of information provided
  - Breeding
    - Selection
    - Mating
  - Marker Assisted Management
  - Concern that more variants will not translate into greater accuracy of predictions

Thought Experiment

- There are roughly 100 million cattle in the U.S.
- If we had all of their genomic sequences, how much storage would be required to store sufficient information to reconstruct the sequence of any animal?

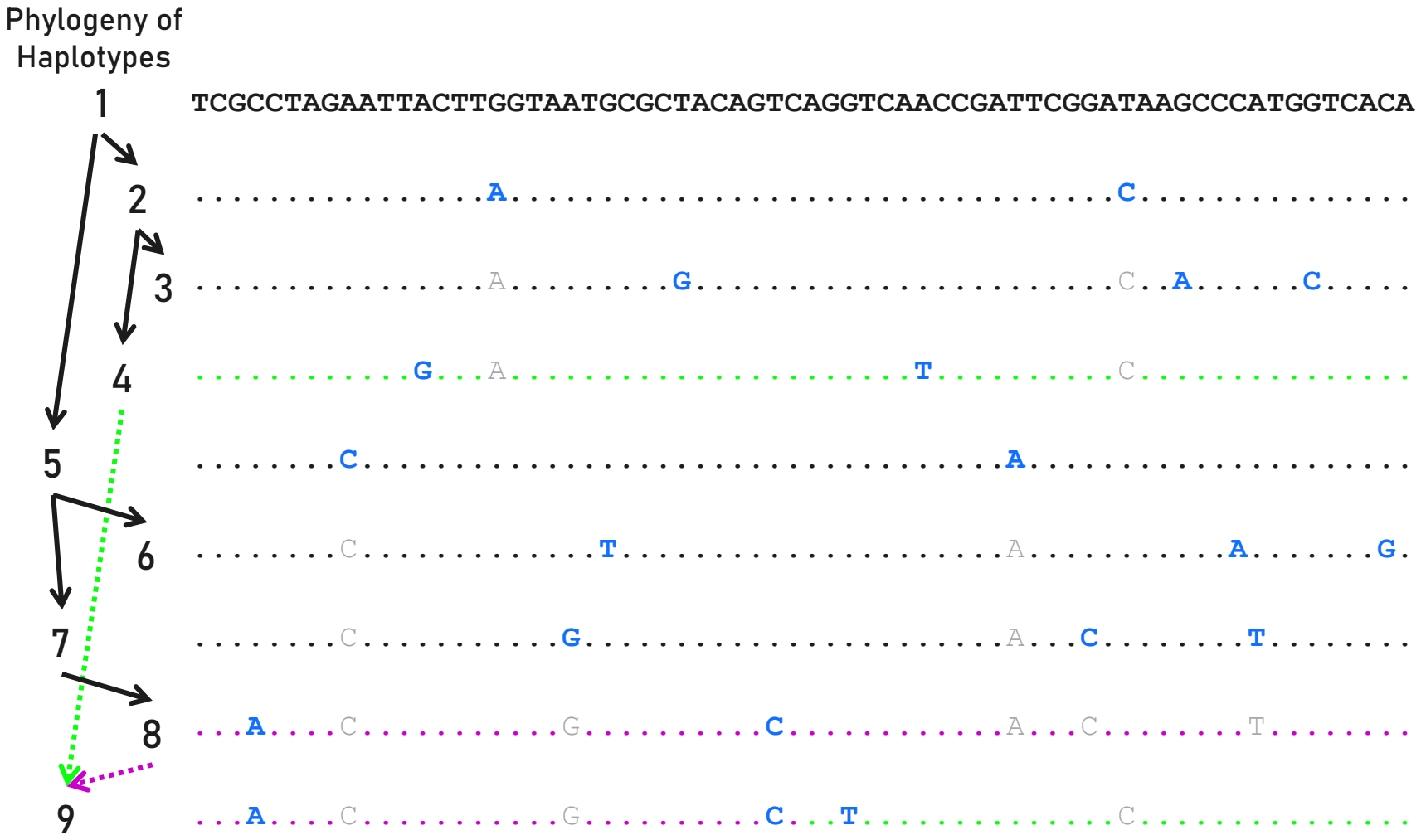
Break the Genome into Segments for Haplotyping

- Break at structural variant boundaries and recombination hotspots
- Maximum of 255 haplotypes / segment in population so haplotype IDs can be stored in one byte
- Maximum of 32,767 bp /segment so offsets can be stored in 16-bit integers



Numbers of haplotypes that occur within population

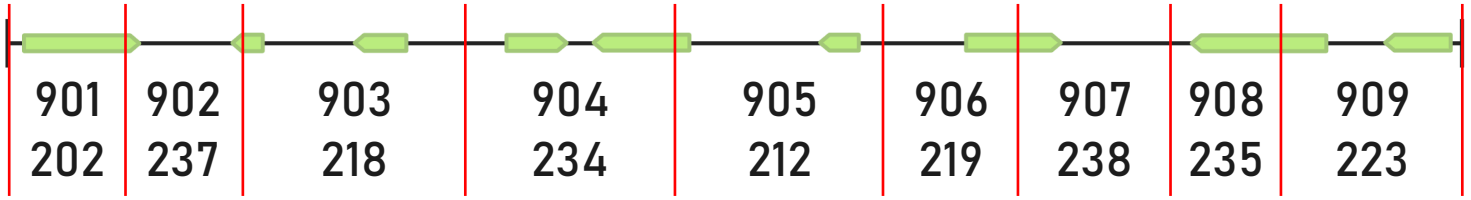
Representation of Haplotype Sequences of One Genome Segment



Reference sequence requires 3 GB of storage across all segments

Assume 250,000 segments, average of 200 haplotypes per segment, 2 new variants per segment per haplotype, and 5 bytes per new variant.
Implies:
250,000 × 200 × 2 × 5
= 500 MB to store variants

Conceptual Representation of Individual Haplotypes



901	902	903	904	905	906	907	908	909
202	237	218	234	212	219	238	235	223

9	9	1	1	1	1	1	1	1
6	6	6	6	2	2	2	2	2

3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	2	2	2

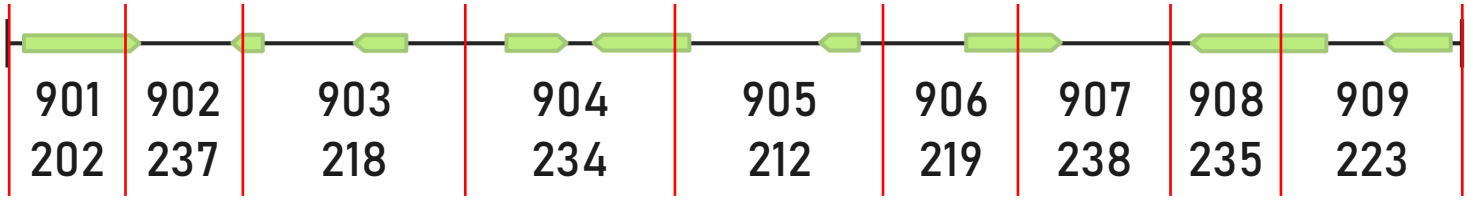
9	9	1	1	1	1	1	1	1
3	3	3	×	4	4	2	2	2

6	6	×	2	2	2	2	2	2
4	4	4	4	4	4	2	2	2

3	3	3	4	4	4	×	1	1
6	×	4	4	4	4	2	2	2

Precursor to format for improved statistical analysis (described later)

Practical Representation of Individual Haplotypes



9	9	1	1	1	1	1	1	1
6	6	6	6	2	2	2	2	2

3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	2	2	2

Paternal
Paternal 907 : 89 : 37,745

Paternal 903 : 28,419 : 80,426
Maternal

$32.99

Maternal 907 : 89 : 10,467
Paternal 902 : 12,185 : 10 (segment : offset : length of ROA)

Assume 250,000 haplotype segments

500 KB/founder × 0.5 M individuals = 250 GB

60 Chromosomes × 1 B + 60 Crossovers × 10 B = 600 bytes / individual

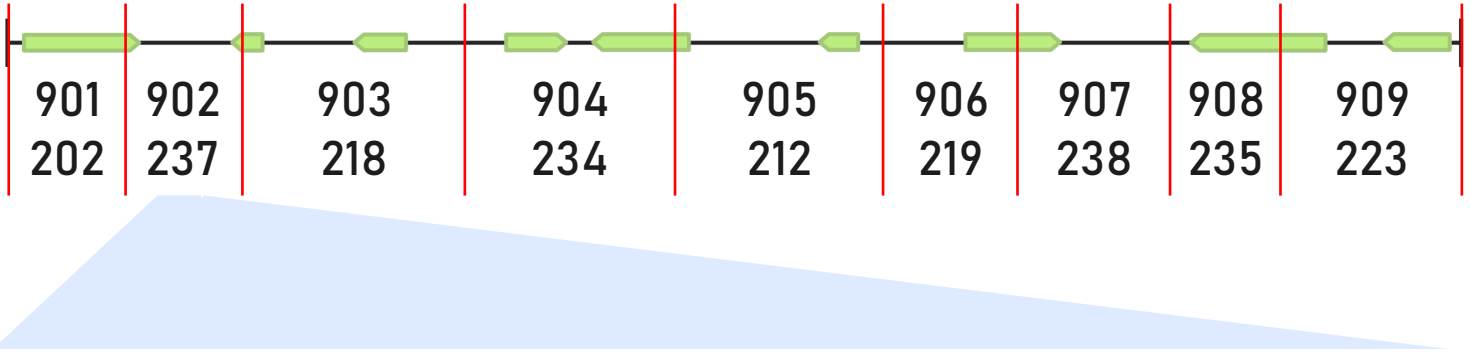
Assume 1 KB / individual × 100 M cattle in U.S. = 100 GB

4 GB + 250 GB + 100 GB = 354 GB

Imputation of Lowpass from Parents with Phased Sequence

- Current low-pass imputation uses algorithms that impute each animal individually without considering pedigree
- Imputation can be far more efficient if both parents are known and have imputed phased genomic sequence
  - Lower depth of sequence coverage
  - Less computational expense

Imputation of Lowpass from Parents with Phased Sequence



9 ...A....C...A...G...G.G...T....C...T...A....T.....C..G..C...G...C.

3 ...C....A...A...A...A.G...G....T...G...A....T.....C..A..C...C...C.

? ? ? 1 0 ? ? 0 ?

~C~ ~A~ ~A~ ~G~ ~.~ ~T~ ~G~ ~T~ ~G~ ~A~ ~.~ ~A~ ~G~ ~G~

? ? ? 1 ? ? 1 ? 1

6 ...C....C...A...G...A.T...T....T...G...A....A.....T..G..A...G...G.

4 ...C....A...G...A...A.G...T....T...G...T....T.....C..G..C...G...C.

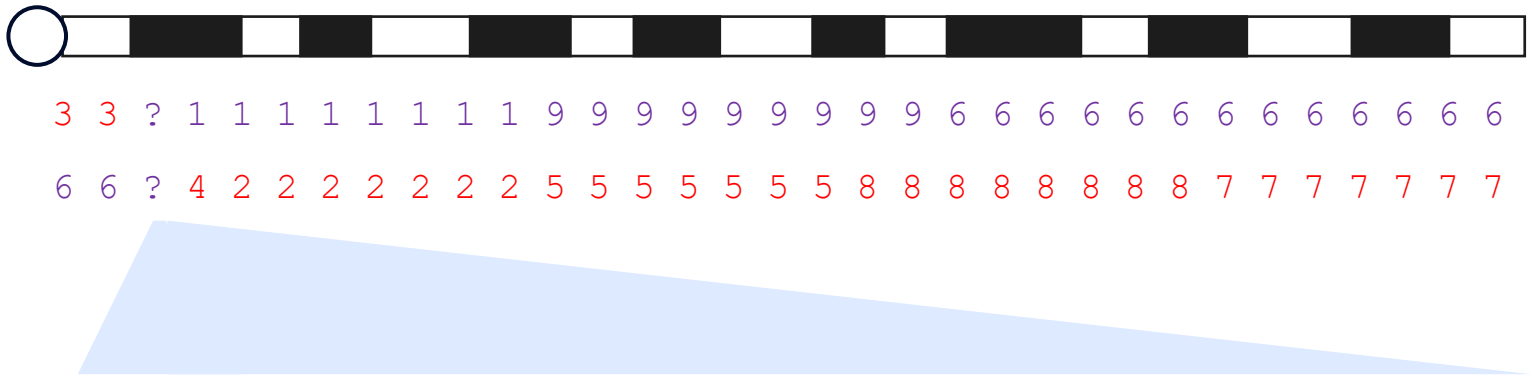
Green paternal 1 due to sequencing error

But we can infer grand-parental origin of some reads based on sequence

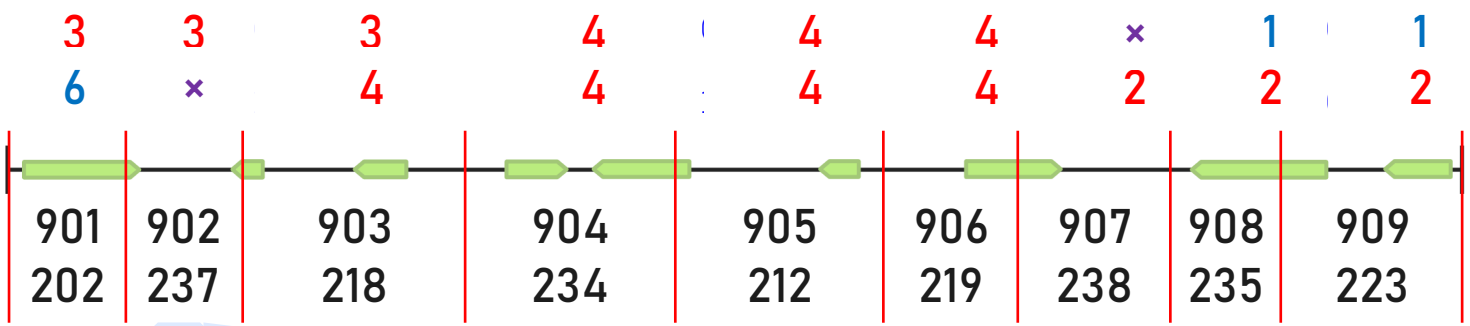
These grand-parental origins are coded as a series of 0s and 1s for each parental chromosome

They look very sparse at this scale

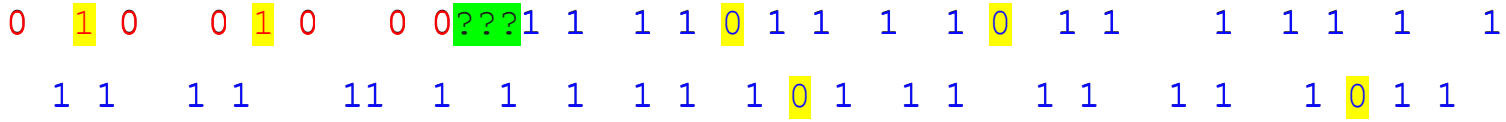
Imputation of Lowpass from Parents with Phased Sequence



That transformation occurs throughout the genome



The grand-parental origins are transformed to haplotypes



The grand-parental origins become more informative as the scale is expanded

Prioritizing Retention of Sequence Reads

- For reads that are stored, store only differences relative to the haplotype that is best matched
- Don't store reads that match an enumerated haplotype
- Most differences from enumerated haplotypes will be sequencing errors, so it does not make sense to store all reads that have errors
- Develop system to accumulate discrepancy counts
  - Store discrepant reads in regions of enumerated haplotypes in which the most discrepancies occur
  - When sufficient evidence of multiple variants of an enumerated haplotype, split it and discard reads that are now concordant with one of the new haplotypes

System of Continuous Improvement of Reference Haplotypes

- In the long run, improvement of reference haplotypes will come from sequence reads generated from low-pass
  - Low pass will find haplotypes that would never be discovered by deep sequencing highly influential bulls
- Improvements to reference haplotypes automatically improve derived sequence of individuals with those haplotypes
- Over time, the regions of ambiguity (ROA) surrounding crossover events will become smaller

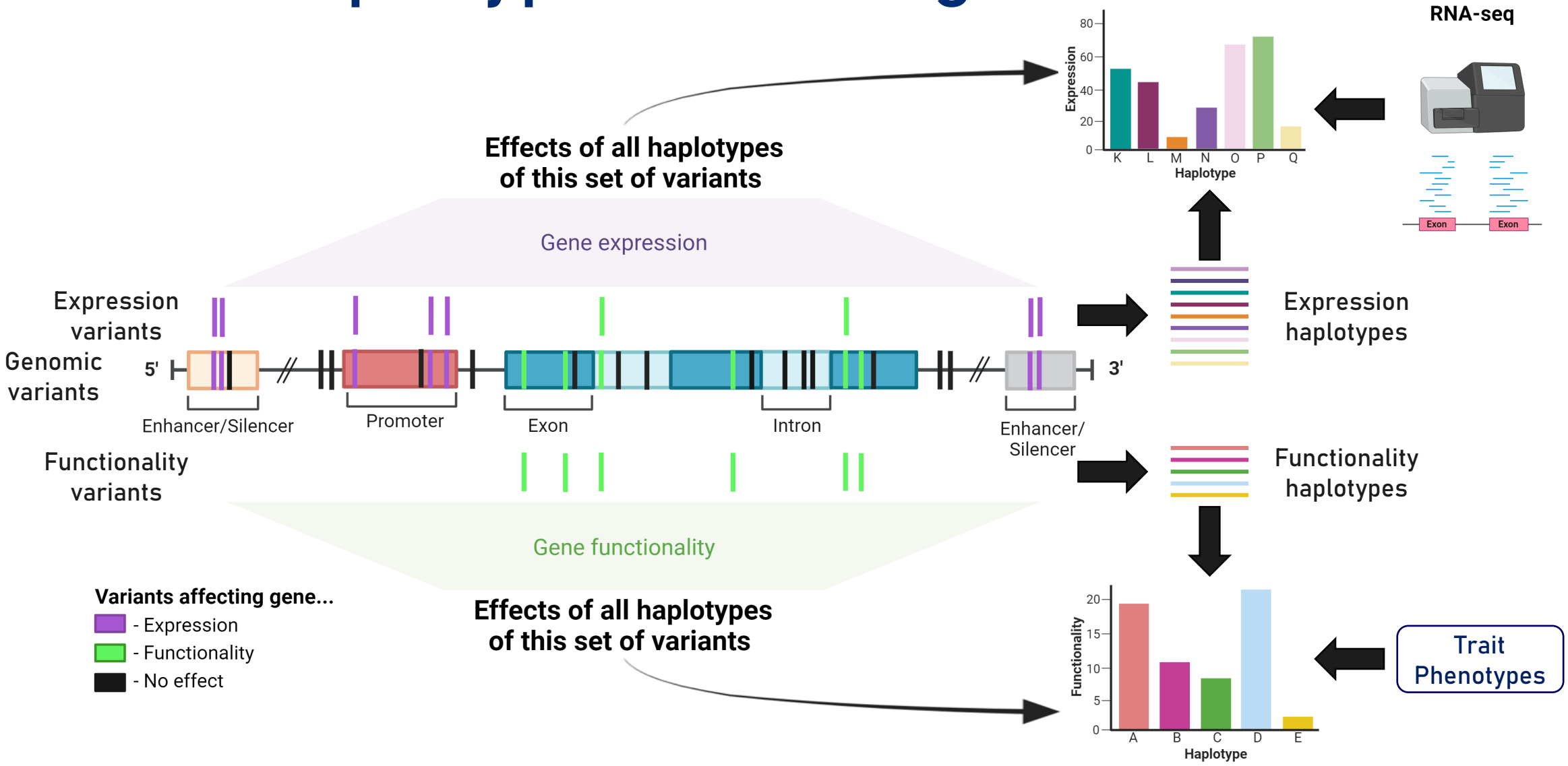
Need 2 different Low-pass Products

- Current low-pass product for founder animals
- A new lower-coverage product for progeny of genome-imputed parents
  - Pedigree-based imputation
  - Much lower depth of sequencing coverage
  - We need a different protocol for barcoded library construction that is optimized for this purpose
    - Should cost minimally more than a PCR reaction
    - Reduced representation that emphasizes sequence flanking genes
  - The advantages of pedigree imputation are far greater if the entire herd or population is sequenced than if just a select few
  - This product needs massive volume to fill low-cost, high-capacity sequencing platforms

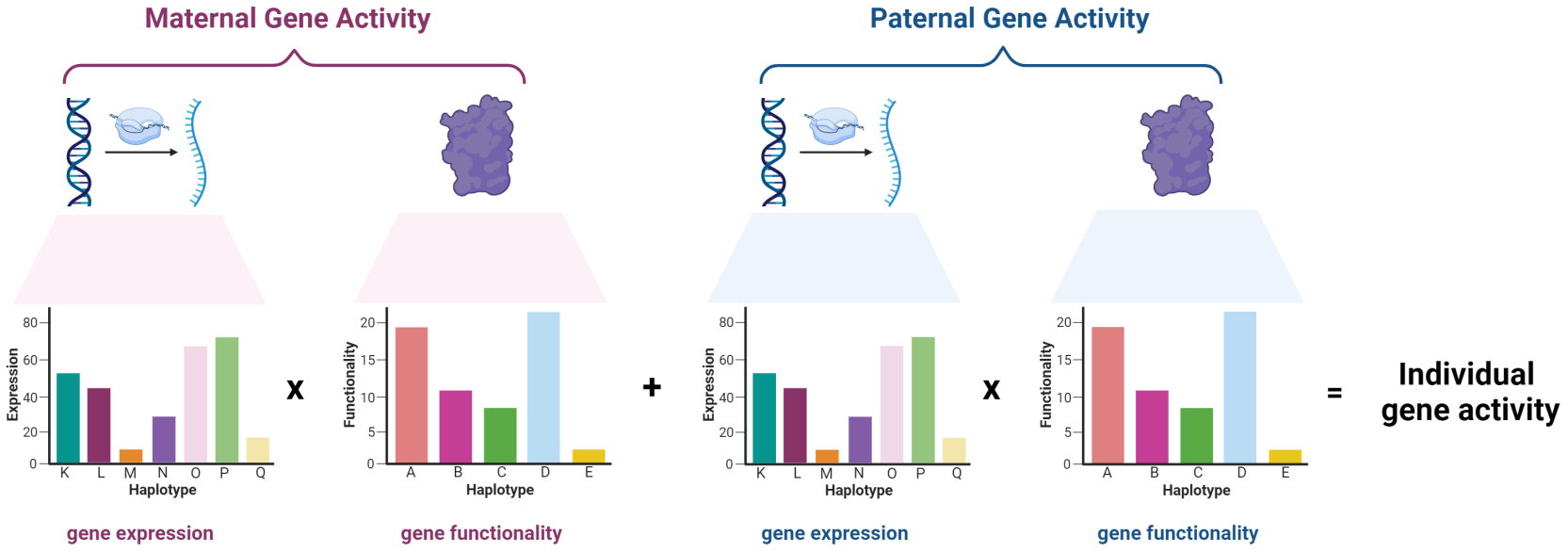
Hierarchical Statistical Model to Utilize Genomic Sequence

- Motivated by biological processes and molecular biology
  - Quite different from linear, additive model
  - Model is multiplicative instead of additive at several levels of hierarchy
  - Gene-based instead of SNP-based
- Reduce the overparameterization problem
  - Use hierarchical model to reduce numerous individual SNP effects to one effect per gene at multiple levels of the hierarchy.
  - Use haplotype model at other levels of heirarchy
  - Use information external to the genomic evaluation to estimate some of the parameters
- Use external information for feature selection and to improve the model

Haplotypes in Biological Model



Model of Gene Activity for Gene *j*



$$\mathbf{X}^{\mathrm{x}}_{\mathrm{m}j}\boldsymbol{x}_j \odot \mathbf{X}^{\mathrm{f}}_{\mathrm{m}j}\boldsymbol{f}_j + \mathbf{X}^{\mathrm{x}}_{\mathrm{p}j}\boldsymbol{x}_j \odot \mathbf{X}^{\mathrm{f}}_{\mathrm{p}j}\boldsymbol{f}_j = \boldsymbol{a}_j$$

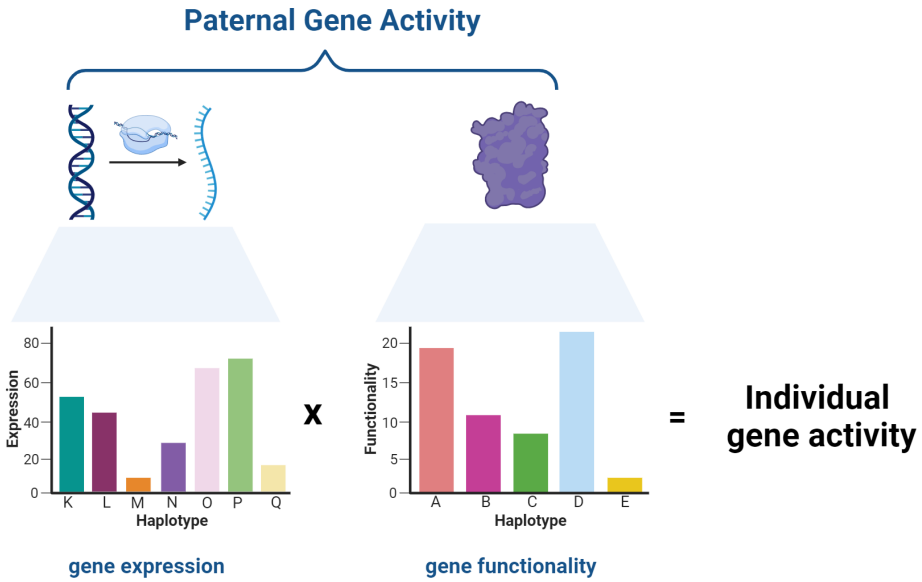
Design matrices relating individuals to maternal haplotypes

Vectors of haplotype effects represented in bar graphs above them

Hadamard product (elementwise multiplication) operator

Vector of individual activity of gene j

Model of Gene Activity with Parental Imprinting



$$\mathbf{X}^{\mathrm{x}}_{\mathrm{p}j}\boldsymbol{x}_j \quad \odot \quad \mathbf{X}^{\mathrm{f}}_{\mathrm{p}j}\boldsymbol{f}_j \quad = \quad \boldsymbol{a}_j$$

Default Model of Dominance for 2 Alleles

Gene Effect:
Assume that for any gene, there is a quantifiable measure of how effective a gene is in any individual.

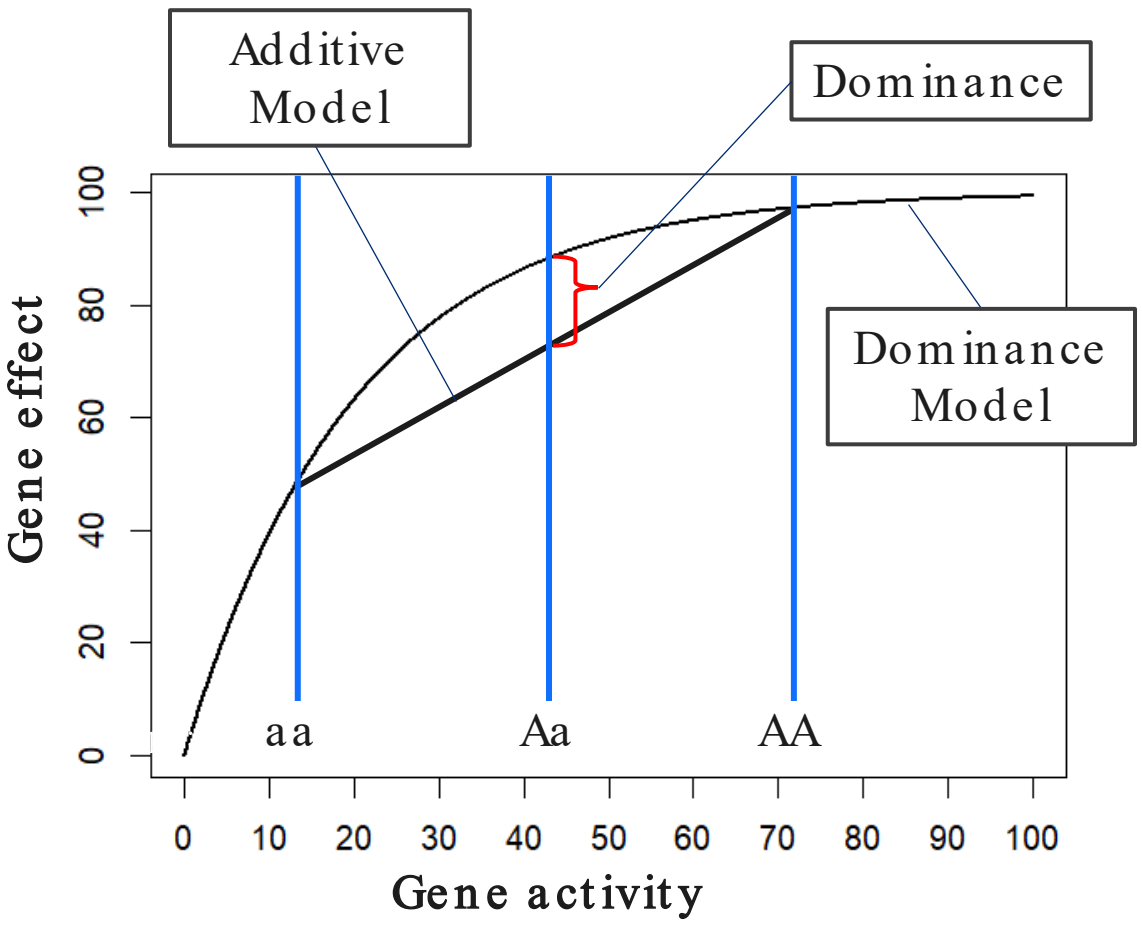
For a gene that produces an enzyme, the measure would be the enzymatic activity.

For a gene that produces a structural protein, the measure might be the amount of the protein.

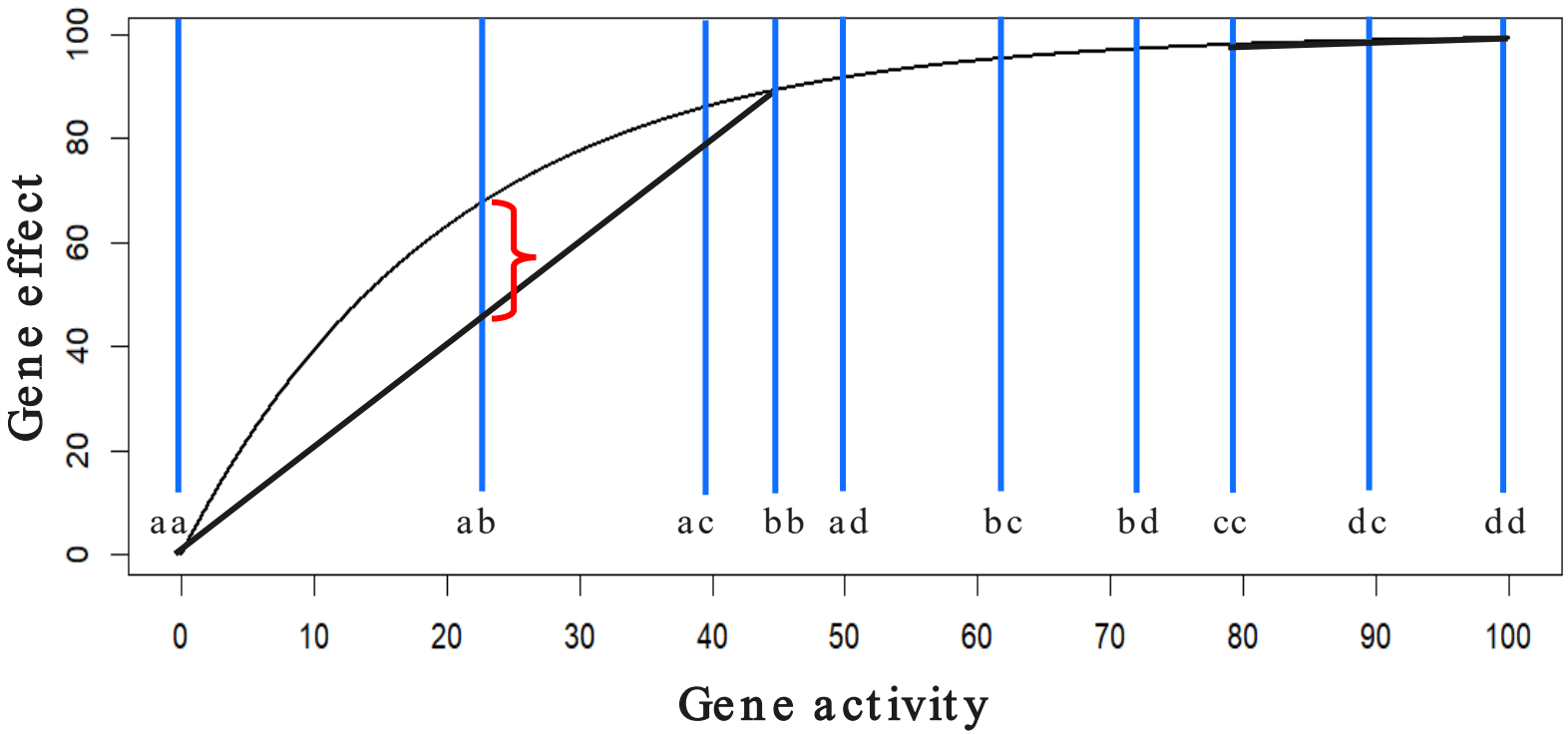
Gene effect is that measure as a percentage of its asymptotic or maximum value.

Gene effect takes into account negative feedback on expression exerted by the gene product.

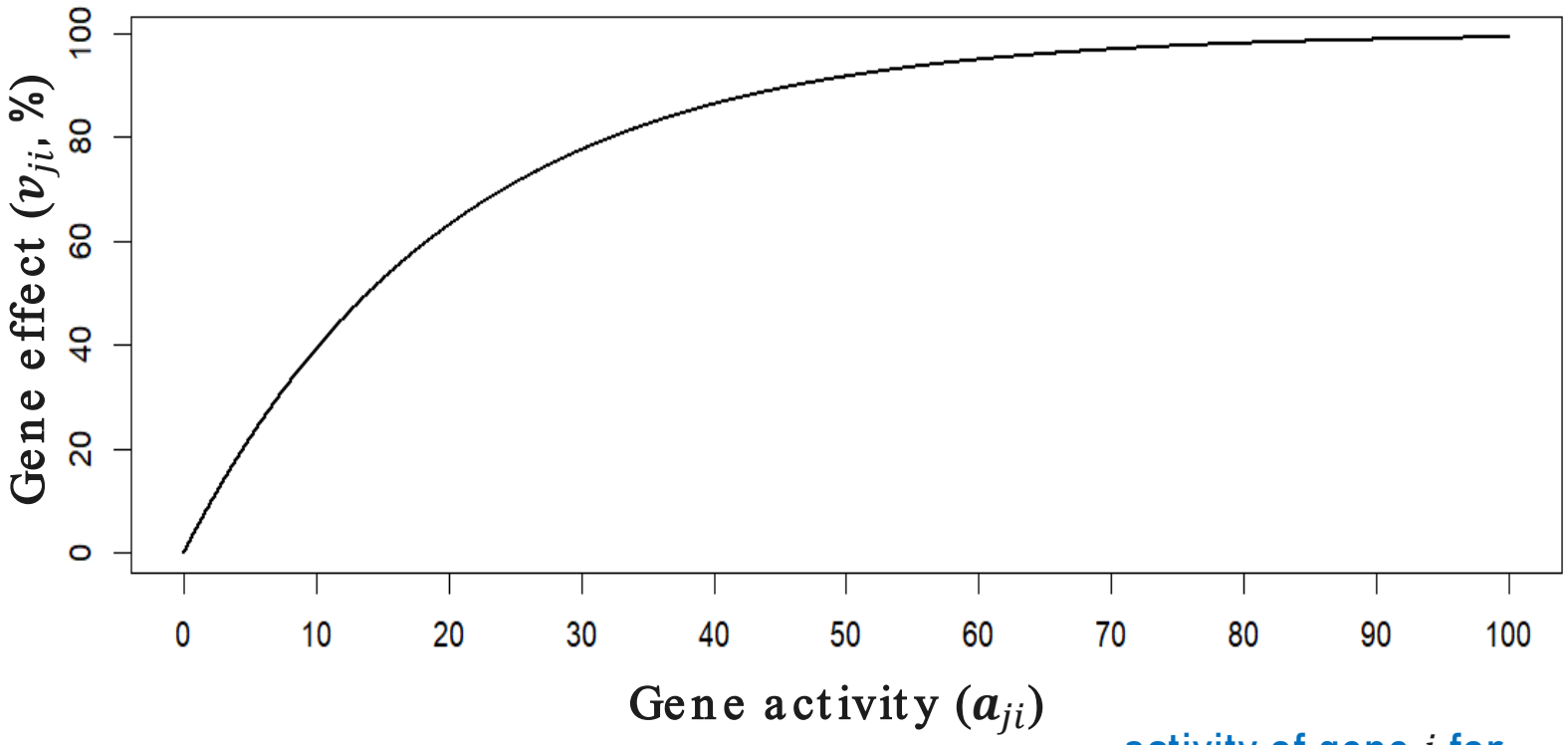
Ideally, phenotypic variation in any trait influenced by a gene would be proportional to its gene effect.



Default Model of Dominance for Multiple Haplotypes



Default Model of Dominance



effect of gene j for individual i

$$v_{ji} = 1 - e^{-k_j a_{ji}}$$

activity of gene j for individual i

decay rate for gene j (parameter to be estimated)

More General Model of Gene Effect

Parameter for linear regression on individual activity of gene j effect

Parameter for regression on the default dominance model for gene j

Parameter for regression on overdominance model for gene j

$$\boldsymbol{v}_{ji} = \boldsymbol{a}_{ji}\alpha_j + \left(1 - \mathrm{e}^{-k_j \boldsymbol{a}_{ji}}\right)\delta_j - \left[\left(\mathbf{X}^{\mathrm{f}}_{\mathrm{m}j} \odot \mathbf{X}^{\mathrm{f}}_{\mathrm{p}j}\right)\mathbf{1}\right]_i o_j \quad \forall i$$

gene effects

Linear effect of individual activity

Default dominance model

Vector equal to 1 for homozygous functional haplotypes and 0 for heterozygous

Term designed to accommodate an overdominance model

Model of Genetic Merit

Vector of total genetic merit of individuals for trait t

Parameter to scale gene j effect to trait t

Parameter to scale product of genes j and k effects to trait t

2-factor epistasis

Parameter to scale product of genes j, k and l effects to trait t

3-factor epistasis

$$\boldsymbol{u}_t = \sum_j \tau_{jt} \boldsymbol{v}_j + \sum_{j,k \in \Psi} \psi_{jkt} \times \boldsymbol{v}_j \odot \boldsymbol{v}_k + \sum_{j,k,l \in \Xi} \xi_{jklt} \times \boldsymbol{v}_j \odot \boldsymbol{v}_k \odot \boldsymbol{v}_k + \cdots$$

gene effects summed over genes

- Pairs of genes in same biochemical pathway
- Pairs in which one gene's product binds to the other gene (e.g. transcription factors) or its product
- etc.

Not any pair of genes that happen to be in the same genome

Triplets of genes in which at least 2 pairs show significant 2-factor epistasis or all 3 genes are in same biochemical pathway

Biologically Motivated Heirarchical Model

$$\boldsymbol{a}_j = \mathbf{X}^{\mathrm{x}}_{\mathrm{m}j}\boldsymbol{x}_j \odot \mathbf{X}^{\mathrm{f}}_{\mathrm{m}j}\boldsymbol{f}_j + \mathbf{X}^{\mathrm{x}}_{\mathrm{p}j}\boldsymbol{x}_j \odot \mathbf{X}^{\mathrm{f}}_{\mathrm{p}j}\boldsymbol{f}_j$$

$$\boldsymbol{v}_{ji} = \boldsymbol{a}_{ji}\alpha_j + \left(1 - \mathrm{e}^{-k_j \boldsymbol{a}_{ji}}\right)\delta_j - \left[\left(\mathbf{X}^{\mathrm{f}}_{\mathrm{m}j} \odot \mathbf{X}^{\mathrm{f}}_{\mathrm{p}j}\right)\mathbf{1}\right]_i o_j \quad \forall i$$

$$\boldsymbol{u}_t = \sum_j \tau_{jt}\boldsymbol{v}_{jt} + \sum_{j,k\in\Psi} \psi_{jkt} \times \boldsymbol{v}_j \odot \boldsymbol{v}_k + \sum_{j,k,l\in\Xi} \xi_{jklt} \times \boldsymbol{v}_j \odot \boldsymbol{v}_k \odot \boldsymbol{v}_k + \cdots$$

$$\boldsymbol{y} = \mathbf{X}^{\beta}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{e}$$

- Genes are meant to include all in the pangenome, not just those in the core genome, as is the case for current evaluations
  - But additional terms are needed in the model to account for copy number variation

Parameters to be estimated:

$\boldsymbol{x}$ = expression haplotype effects (total number over all genes)

$\boldsymbol{f}$ = functionality haplotype effects (total number over all genes)

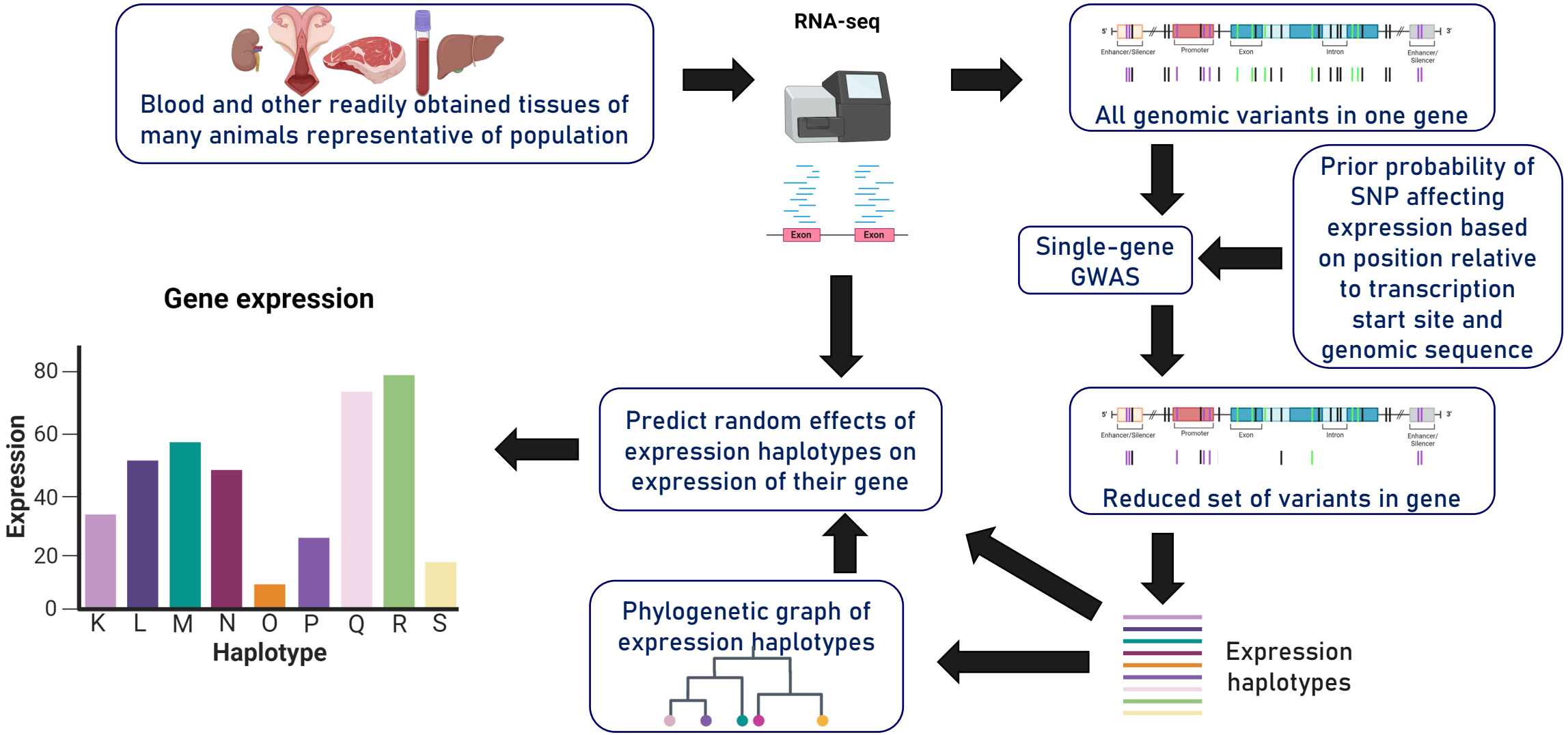
α_j, δ_j, o_j = weighting linear vs dominance (3 × number of genes)

τ_{jt} = scale gene effects to traits (number of traits × number of genes)

ψ_{jkt}, ξ_{jklt} = interactions (number undetermined)

$\boldsymbol{\beta}$ = fixed effects in standard model

Use Population-level RNA-seq to Estimate Effects of Haplotype on Expression



Use Population-level RNA-seq to Estimate Effects of Haplotype on Expression

- Effects of expression haplotypes on gene expression (x) may account for roughly half of the parameters to be estimated in the genomic prediction model.
  - Using information external to the genomic prediction dataset to estimate these parameters should substantially improve the power of genomic prediction.
  - Furthermore, haplotype effects on gene functionality and gene expression are likely to be relatively confounded with one another when estimated from the same data.
- Within-gene estimation of genomic effects on gene expression should require far fewer observations than genome-wide analysis.
- Substituting externally derived estimates of x into the first level of the hierarchical model makes that level of the heirarchy a standard linear model for the estimation of f instead of a multiplicative model.

Computational Considerations

- Probably easiest to fit with MCMC
  - Conditional sampling distributions may be more complicated than for strictly linear models
- Longer term, I want to explore using non-stochastic conditionally linear mixed model computations to run these models

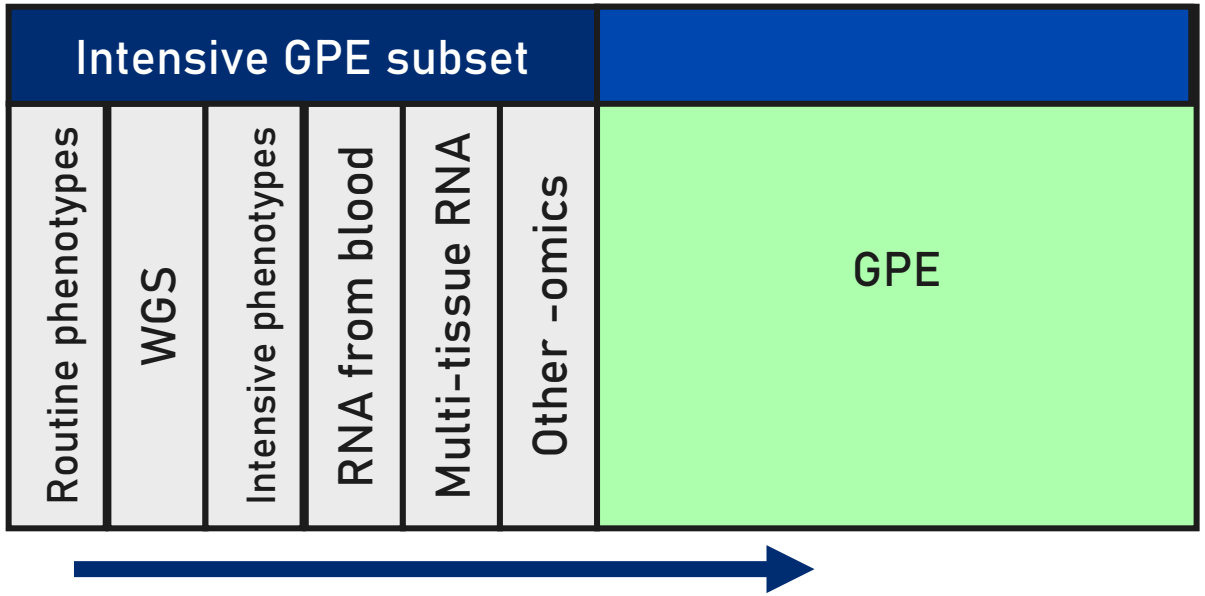
Assumptions of Heirarchical Statistical Model of Genomics

- Default assumption is that effects of haplotypes on gene expression and functionality are proportionally similar across traits for the same gene as is the model for dominance and epistatic effects
- Default assumption is that haplotype effects on gene expression are proportionally similar across tissues, physiological states, and treatments
- Fit random effects that account for departures from these default assumptions when sufficient evidence exists

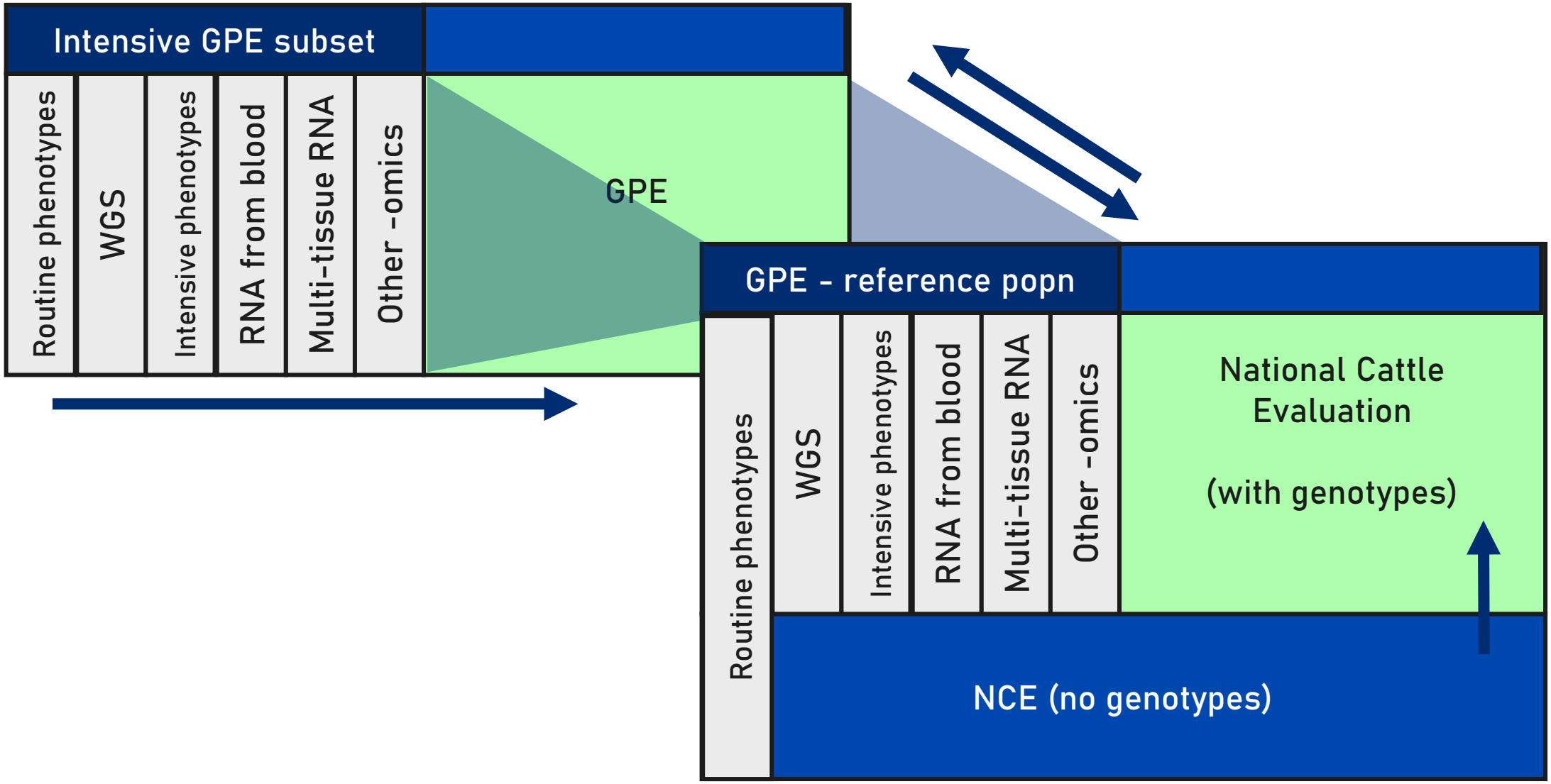
Imputation of Gene Expression

- Imputation of gene expression of many tissues from a reference population with RNA-seq on many tissues and genomics to larger populations with RNA-seq on blood only and genomics
  - Basu, Mahashweta, et al. 2021. Predicting tissue-specific gene expression from whole blood transcriptome. Sci. Adv. 2021; 7 : eabd6991
- Imputation of gene expression from reference population with RNA-seq and genomic sequence to larger populations with genomic sequence but no expression data (within the same tissue(s)
  - Gamazon, E. R., et al. (2015). "A gene-based association method for mapping traits using reference transcriptome data." Nat Genet 47(9): 1091-1098.
- Imputation of multi-omics from reference population with multi-omics and genomic sequence to larger populations with genomic sequence but no multi-omics.
  - Xu, Y., et al. (2023). "An atlas of genetic scores to predict multi-omic traits." Nature 616(7955): 123-131.

Imputation


Intensive GPE subset
Routine phenotypes
WGS
Intensive phenotypes
RNA from blood
Multi-tissue RNA
Other -omics
GPE

Imputation


Intensive GPE subset
Routine phenotypes
WGS
Intensive phenotypes
RNA from blood
Multi-tissue RNA
Other -omics
GPE
GPE - reference popn
Routine phenotypes
WGS
Intensive phenotypes
RNA from blood
Multi-tissue RNA
Other -omics
National Cattle Evaluation
(with genotypes)
NCE (no genotypes)

Livestock as model species for understanding systems biology at a new level

Opportunities for Genotyping Large Numbers of Commerical Cattle for Marker Assisted Selection (A Subset of Precision Livestock Management)

Opportunities for Genotyping Large Numbers of Commerical Cattle for Marker Assisted Selection (A Subset of Precision Livestock Management)

Postdoctoral Position in Variance Component Estimation

Acknowledgments

- Bailey Engle
- Larry Kuehn
- Warren Snelling
- Brittney Keel