Dr. Jack Shere APHIS
Dr. Alecia Naugle APHIS
TVMDL Texas A&M:
Drs Akey, Ferro, Legare
CVM TAMU, AgriLife

Decoding Genome Sequence

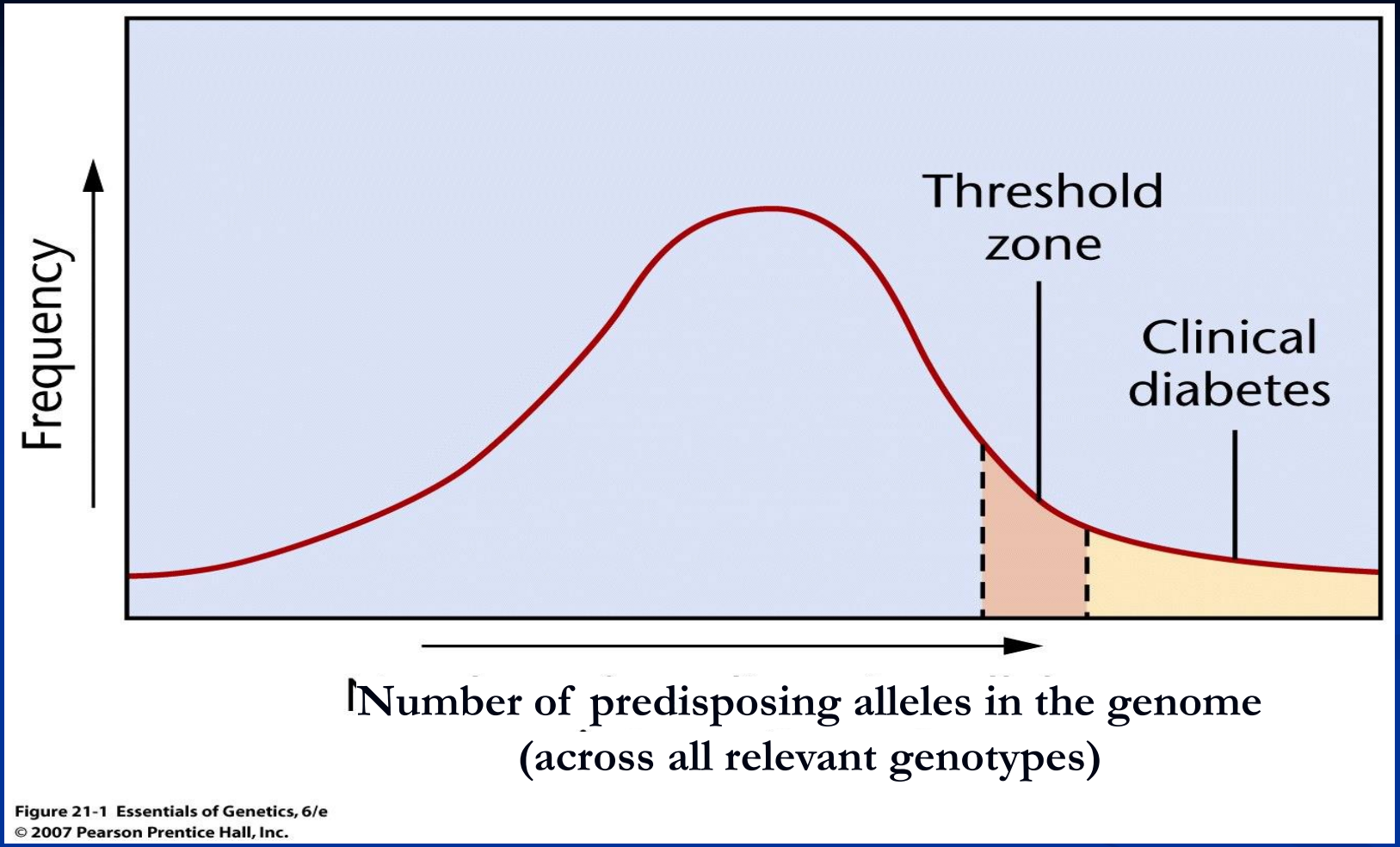


Figure 21-1 Essentials of Genetics, 6/e
© 2007 Pearson Prentice Hall, Inc.

- Genetic Liability: Unobserved Susceptibility to disease; the sum of environmental and genetic contributions. Forms continuous range, but only two phenotypes here!

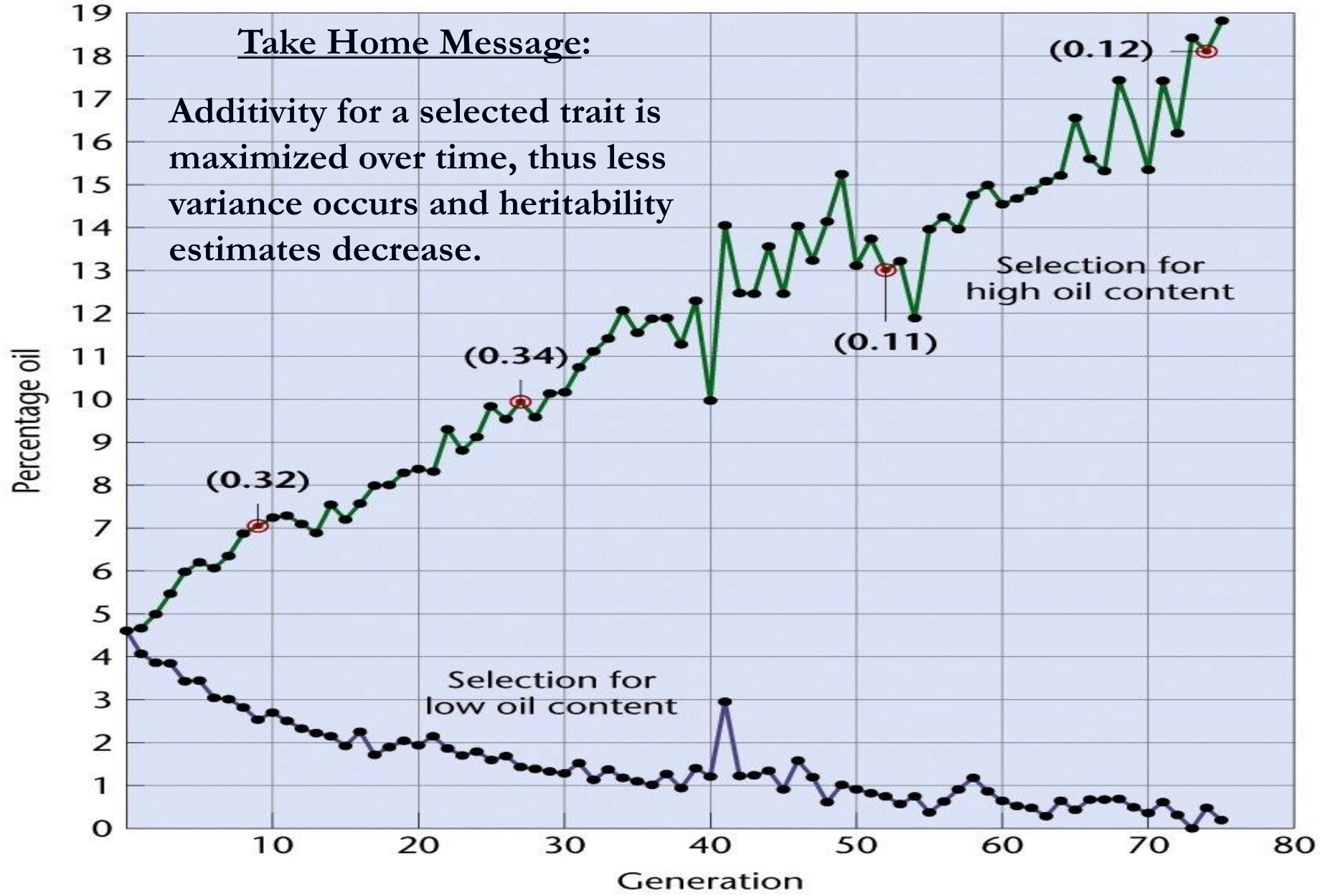
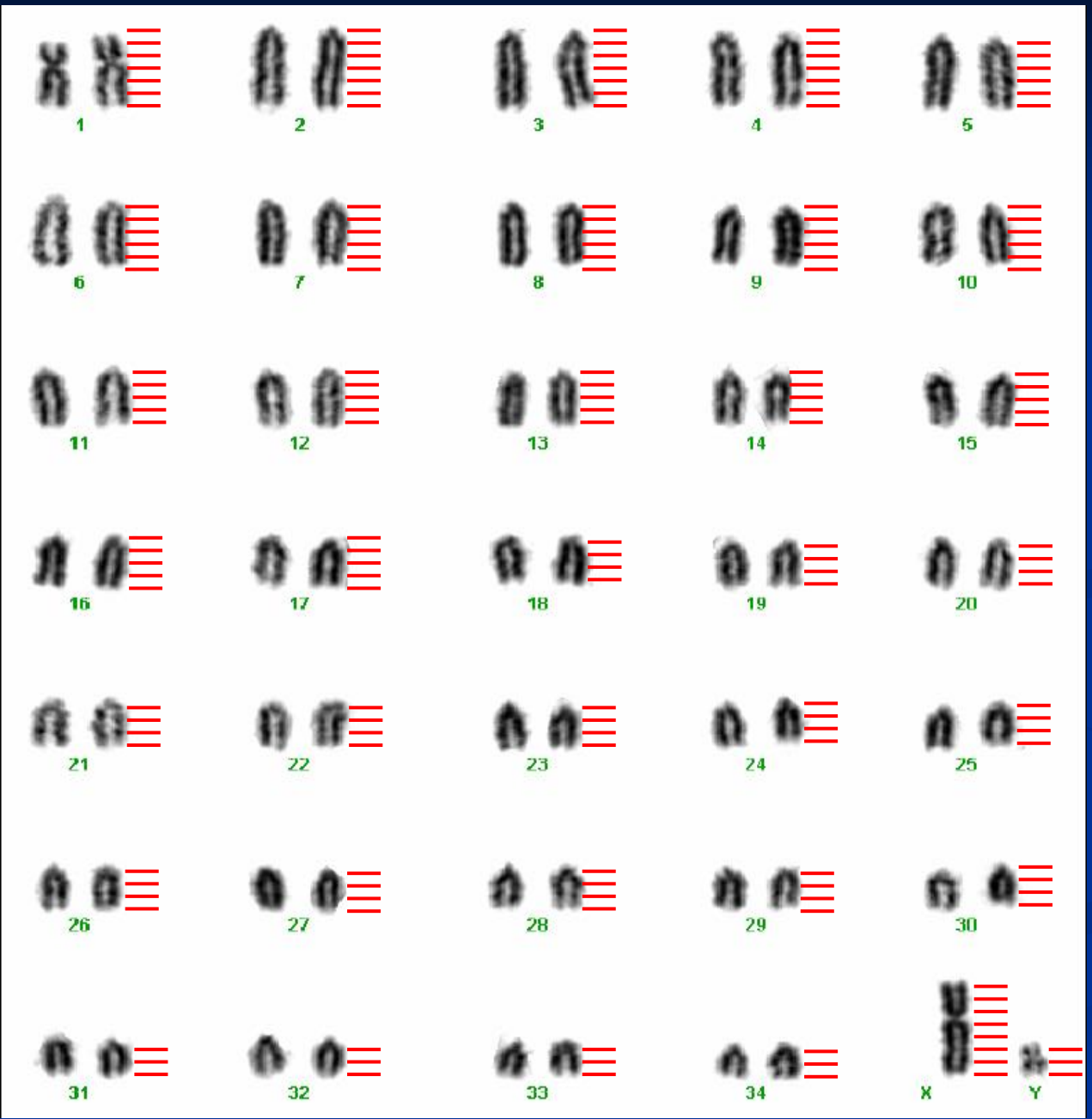


Figure 21-8 Essentials of Genetics, 6/e

© 2007 Pearson Prentice Hall, Inc.

Genome-Wide Association



- Uses SNPs
- SNPs more or less evenly spaced across all chromosomes
- Allows us to test for associations with traits

G3 Genes | Genomes | Genetics

INVESTIGATION

Accurate Genomic Predictions for Chronic Wasting Disease in U.S. White-Tailed Deer

Christopher M. Seabury,*,[1] David L. Oldeschulte,* Eric K. Bhattarai,* Dhruti Legare,[†] Pamela J. Ferro,[†] Richard P. Metz,[‡] Charles D. Johnson,[‡] Mitchell A. Lockwood,[§] and Tracy A. Nichols**

*Department of Veterinary Pathobiology, Texas A&M University, College Station, Texas, [†]Texas A&M Veterinary Medical Diagnostic Laboratory, College Station, Texas, [‡]Genomics Core, Texas A&M AgriLife Research, College Station, Texas, [§]Texas Parks and Wildlife Department, Austin, Texas, and **USDA-APHIS-VS-Cervid Health Program, Fort Collins, CO

ABSTRACT The geographic expansion of chronic wasting disease (CWD) in U.S. white-tailed deer (*Odocoileus virginianus*) has been largely unabated by best management practices, diagnostic surveillance, and depopulation of positive herds. Using a custom Affymetrix Axiom single nucleotide polymorphism (SNP) array, we demonstrate that both differential susceptibility to CWD, and natural variation in disease progression, are moderately to highly heritable ($h^2 = 0.337 \pm 0.079 - 0.637 \pm 0.070$) among farmed U.S. white-tailed deer, and that loci other than *PRNP* are involved. Genome-wide association analyses using 123,987 quality filtered SNPs for a geographically diverse cohort of 807 farmed U.S. white-tailed deer (n = 284 CWD positive; n = 523 CWD non-detect) confirmed the prion gene (*PRNP*; G96S) as a large-effect risk locus (*P*-value $<$ 6.3E-11), as evidenced by the estimated proportion of phenotypic variance explained (PVE $\geq$ 0.05), but also demonstrated that more phenotypic variance was collectively explained by loci other than *PRNP*. Genomic best linear unbiased prediction (GBLUP; n = 123,987 SNPs) with *k*-fold cross validation ($k = 3$; $k = 5$) and random sampling (n = 50 iterations) for the same cohort of 807 farmed U.S. white-tailed deer produced mean genomic prediction accuracies $\geq$ 0.81; thereby providing the necessary foundation for exploring a genomically-estimated CWD eradication program.

KEYWORDS
genome-wide association
chronic wasting disease
white-tailed deer
genomic prediction
heritability

The fatal wasting syndrome known as chronic wasting disease (CWD) was first observed among captive mule deer (*Odocoileus hemionus*) and black-tailed deer (*Odocoileus hemionus columbianus*) within several Colorado wildlife facilities during the late 1960s, and histologically recognized as a prion disease by the late 1970s (Williams and Young 1980; Moreno and Telling 2018). Thereafter, CWD was described in free-ranging U.S. mule deer, elk (*Cervus elaphus nelsoni*), white-tailed deer (*Odocoileus virginianus*; hereafter WTD) and moose (*Alces alces shirasi*), with subsequent diagnostic surveillance suggesting an irreversible geographic expansion of the disease among farmed and free-ranging populations of these species (Moreno and Telling 2018; Gavin *et al.* 2019; Osterholm *et al.* 2019).

Copyright © 2020 Seabury *et al.*
doi: https://doi.org/10.1534/g3.119.401002
Manuscript received December 17, 2019; accepted for publication February 25, 2020; published Early Online March 2, 2020.
This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Supplemental material available at dryad: https://doi.org/10.5061/dryad.xd2547dcw.
[1]Corresponding author: E-mail: cseabury@cvm.tamu.edu

At present, at least 26 U.S. states and multiple Canadian provinces are known to be affected by CWD (Moreno and Telling 2018; Gavin *et al.* 2019; Osterholm *et al.* 2019). Likewise, Norway, Finland, and the Republic of Korea have also reported CWD in free-ranging reindeer (*Rangifer tarandus*; Norway), moose (Norway, Finland), and imported elk (Korea) (Moreno and Telling 2018; Gavin *et al.* 2019; Osterholm *et al.* 2019). The implementation of modern best management practices, including containment and depopulation of positive herds, has not prevented the emergence of CWD in new geographic areas (Moreno and Telling 2018; Gavin *et al.* 2019; Osterholm *et al.* 2019). Therefore, a need currently exists to develop novel strategies to reduce the prevalence of CWD among farmed deer and elk.

MATERIALS AND METHODS

Study overview

Herein, we investigate differential susceptibility to CWD among farmed U.S. WTD by utilizing genomic DNA samples from CWD positive and CWD non-detect WTD to perform next-generation sequencing with variant prediction for the construction and validation of a medium density SNP array. Thereafter, we use the array in conjunction with *PRNP* genotypes to conduct genome-wide association

analyses (GWAA's) and produce marker-based heritability estimates. Finally, we conclude our study by utilizing the genome-wide SNP data to deploy genomic prediction equations with *k*-fold cross validation to assess the potential for developing a genomically-estimated CWD eradication program.

Animal resources, CWD diagnostics, and DNA isolation

Frozen whole blood samples (n = 448) and rectal biopsies (n = 37) from farmed U.S. white-tailed deer (WTD; both sexes) were available within an existing USDA APHIS repository that was created via federal CWD surveillance activities; including depopulations of CWD positive herds (USDA APHIS, Fort Collins, CO). All herds included both CWD positive (n = 256) and CWD non-detect (n = 229) WTD (Thomsen *et al.* 2012), with geographic representation that included WTD farms located in the U.S. Midwest, Northeast, and South. All diagnostic classifications were based upon immunohistochemistry (*i.e.*, IHC of lymph node, obex), as implemented and performed at USDA National Veterinary Services Laboratory (NVSL) in Ames Iowa (Thomsen *et al.* 2012). Genomic DNA was isolated from frozen whole blood using the Applied Biosystems MagMAX DNA Multi-Sample Ultra Kit with the KingFisher 96 Purification System (ThermoFisher), as recommended by the manufacturer, at the Texas Veterinary Medical Diagnostic Laboratory (TVMDL; College Station, TX). Genomic DNA was isolated from WTD rectal biopsies using the LGC sbeadex tissue purification kit (LGC) with automation at GeneSeek Neogen (Lincoln, NE). Hair samples (n $>$ 700) from farmed WTD (both sexes) were also available within an existing Texas Parks and Wildlife Department (TPWD; Austin, TX) repository created via surveillance and depopulation efforts after the initial detection of CWD in Texas. At the time of study, sample repositories for these herds included CWD positive (n $\geq$ 100) and CWD non-detect (n $\geq$ 600) WTD, with a geographic representation that included WTD farms located in the U.S. South (Texas). All diagnostic classifications for TPWD samples were based upon IHC (*i.e.*, lymph node, obex, and one tonsil) initially performed at TVMDL, with further confirmation at NVSL (Thomsen *et al.* 2012). Genomic DNA was isolated from WTD hair follicles using the LGC sbeadex tissue purification kit (LGC) with automation at GeneSeek Neogen (Lincoln, NE). All genomic DNAs were quantified and evaluated for purity (260/280 ratio) via Nanodrop (ThermoFisher).

Reduced representation libraries and sequencing

Pooled DNA samples were previously shown to be effective for variant prediction; thus enabling downstream genotyping in WTD (Seabury *et al.* 2011). Therefore, pooled DNA samples were created for CWD positive (WT1) and CWD non-detect (WT0) WTD acquired from the USDA APHIS repository (frozen blood). Briefly, WT1 (n = 190) and WT0 (n = 184) genomic DNAs with concentrations $\geq$ 15 ng / μl were used to construct sequencing pools representing depopulated WTD from the U.S. Midwest, Northeast, and South by targeting 50 ng of genomic DNA per WTD in each respective pool (WT1, WT0). Genomic DNAs with concentrations $<$ 15 ng / μl were retained for downstream Affymetrix Axiom and *PRNP* genotyping. Aliquots from each genomic DNA pool (WT0, WT1) were digested with EcoRI, HindIII, and PstI (NEB) in 1X CutSmart buffer for 4 hr at 37°. Enzymes were heat inactivated at 80° for 20 min and held at 10° until ligation. Ligase buffer, ligase (NEB) and barcoded enzyme-specific adapters compatible with DNA possessing EcoRI, HindIII or PstI overhangs were added to the digested samples, and incubated at 16° for 8 hr. Following heat inactivation at 80° for 20 min, 1/10th volume of 3M NaAc (pH 5.2) and two volumes of 100% ethanol were added to each sample, and then held at -20° for 1 hr before spinning at high speed for 10 min in a bench-top microfuge. Pellets were washed twice in 1 ml 70% ethanol and resuspended in 130 μl 1X TE. Samples were then sheared to an average size of 350 bp on the Covaris E220 sonicator, and AMPure XP bead purified as per the manufacturers protocol (Beckman Coulter). Sheared DNA fragments were size selected using a Pippin Prep 2% dye-free agarose gel with internal size markers (Sage Bioscience); aiming for 300-800 bp inserts. Recovered samples were cleaned with 1X AMPure XP beads and end-repaired first with Bst DNA Polymerase (NEB), then with a DNA End Repair Kit (NEB), and A-Tailed using Klenow Fragment (3′ → 5′ exo-) (NEB) in the presence of 50 nM ATP. An Illumina P7-adapter (Adapter B) was ligated to the A-tailed ends as described above. Following two rounds of AMPure XP bead purification, 150 ng of each pool was then subjected to "pre-selection PCR" (PreCR) in which a biotinylated forward primer (P5-Select) and unique indexed reverse primers (TDX) were used to amplify and tag desired DNA fragments. Reactions (200μl total) contained 200 nM dNTPs, biotinylated forward and two P7-index primers per pool, 4 units Q5 Hi-Fidelity Taq (NEB), and were split into 2 X 100 μl volumes for thermocycling. Following an initial denaturation at 98° for 30 sec, samples were subjected to 15 cycles of 98° for 10 sec, 72° for 30 sec then a final elongation at 72° for 5 min and held at 4°. PCR products were purified using Qiagen PCR purification columns, then 1X AMPure XP beads, and quantified via DeNovix. Removal of undesirable fragments (P5 to P5 and P7 to P7 ligated products) was achieved with Dynabeads M-270 Streptavidin coupled magnetic beads (ThermoFisher). Briefly, 50 μl of beads per sample were captured and washed twice with 1X Bead Washing Buffer (1X BWB, 10 mM Tris-HCl with pH 7.5, 1 mM EDTA, 2 M NaCl). Beads were resuspended in 100μl 2X BWB and mixed with 2000 ng of PreCR product in 100 μl EB. After 20 min at room temperature, beads were captured and washed three times in 200 μl 1X BWB, twice in 200 μl water, and once in 100 μl 1X SSC. Beads were then resuspended in 50 μl 1X SSC, heated at 98° for 5 min, and placed on a magnet, with the supernatant removed thereafter. This elution was repeated and the final supernatants were purified with Qiagen PCR columns, as recommended by the manufacturer. The eluted ssDNA was DeNovix quantified, and diluted to 1 ng/μl with EB. A final PCR was performed on 10 ng of input DNA using FiLi-F1 and FiLi-R1 primers in a 50 μl reaction, with only 8 cycles. Final PCR products representing WT0-EcoRI, WT0-HindIII, WT0-PstI, WT1-EcoRI, WT1-HindIII, and WT1-PstI were purified with 1X AMPure XP beads, quantified and assessed for quality on a Fragment Analyzer (Advanced Analytics), and sequenced (2 × 125 bp, paired end) on the Illumina HiSeq 2500 at the Texas AgriLife Genomics and Bioinformatics Core Facility at Texas A&M University. Raw reads generated for each library were as follows: WT1 EcoRI (134,299,714); WT1 PstI (175,412,740); WT1 HindIII (152,371,052); WT0 EcoRI (145,989,752); WT0 PstI (120,598,450); WT0 HindIII (137,148,058). Primers used were as follows: For ligation to restriction enzyme cut DNA, adapters were made by mixing equimolar amounts of top (T) and bottom (B) oligos in 1X oligo hybridization buffer (50 mM NaCl, 1 mM EDTA, 10 mM Tris-HCl, pH 8.0), heating them to 98° for 1 min, and allowing them to cool to room temperature at a rate of 0.1° per second. Primer sequences used were as follows (X denotes bases used for barcoding): Eco_T,5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACX XXX XXX XAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T-3′; Eco_B,5′-AAT TAG ATC GGA AGA GCG TCG TGT AGG GAA AGA GTG TXX XXX XXX GTG

TAG ATC TCG GTG GTC GCC GTA TCA TT-3′; Hind_T, 5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACX XXX XXX XAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T-3′; Hind_B, 5′-AGC TAG ATC GGA AGA GCG TCG TGT AGG GAA AGA GTG TXX XXX XXX GTG TAG ATC TCG GTG GTC GCC GTA TCA TT-3′; Pst_T, 5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACX XXX XXX XAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TTG CA-3′; Pst_B, 5′- AGA TCG GAA GAG CGT CGT GTA GGG AAA GAG TGT XXX XXX XXG TGT AGA TCT CGG TGG TCG CCG TAT CAT T-3′; Adaptor-B_T, /5Phos/GAT CGG AAG AGC ACA CGT CTG AAC TCC AGT CAC-3′; Adaptor-B_B, 5′-GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC T-3′; P5_Select, /5BiotinTEG/AAT GAT ACG GCG ACC ACC GAG ATC TAC AC-3′; FiLi-F1: 5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC AC-3′; FILi-R1: 5′-CAA GCA GAA GAC GGC ATA CGA GAT-3′; TDX, 5′-CAA GCA GAA GAC GGC ATA CGA GAT XXX XXX XGT GAC TGG AGT TCA-3′.

Sequence analysis and affymetrix axiom array design

All Illumina sequences were trimmed for quality and adapters using CLC Genomics Workbench 10.1.1 (Qiagen), as previously described (Halley *et al.* 2014; Halley *et al.* 2015; Sollars *et al.* 2017). All trimmed reads were mapped to the WTD genome assembly (GCF_002102435.1 Ovir.te_1.0; https://www.ncbi.nlm.nih.gov/assembly/GCF_002102435.1/) using the CLC Genomics Workbench 10.1.1 reference mapping algorithm (Seabury *et al.* 2011; Halley *et al.* 2014; Halley *et al.* 2015; Sollars *et al.* 2017). Variant prediction was performed using a probabilistic approach implemented within CLC Genomics Workbench 10.1.1 (Halley *et al.* 2014; Halley *et al.* 2015; Oldeschulte *et al.* 2017; Sollars *et al.* 2017). This algorithm estimates error probabilities from quality scores, and uses these probabilities to determine the most likely allele combination per nucleotide position, thus facilitating a user-specified minimum probability threshold ($P \geq 0.95$) for variant prediction, and variant quality scores (Halley *et al.* 2014; Halley *et al.* 2015; Oldeschulte *et al.* 2017; Sollars *et al.* 2017). Additional variant prediction parameters and filters were similar to those recently described (Seabury *et al.* 2011; Oldeschulte *et al.* 2017), and the probabilistic approach produced evidence for 6,268,706 variants, which included 5,561,550 putative SNPs ($P \geq 0.95$; Minor Allele Frequency ≥ 0.01). Variant prediction results were exported from CLC as a single variant call formatted file (VCF), which was used for SNP array design. Briefly, the VCF file was filtered according to the Affymetrix Axiom myDesign guidelines for SNP submission (http://www.affymetrix.com/support/technical/byproduct.affx?product=axiom_custom_agrigenomics) using a custom python script as follows: Retain only biallelic SNPs with minimum depth = 10, maximum depth = 150, minimum minor allele frequency (MAF) ≥ 0.15, minimum SNP quality score = 30, identify probe overlaps for exclusion, and prioritize variants that maximize array density (*i.e.*, A/T and C/G take up two spaces on the array). The python script as well as more detailed documentation are available in Additional File 30 (DRYAD). The targeted number of SNPs for submission to Affymetrix was $> 200,000$; to facilitate internal Affymetrix scoring (*i.e.*, by pconvert, best_pconvert; recommended, neutral, or not recommended) that would enable the design of a final Affymetrix Axiom 200K SNP array. Collectively, 200,000 SNPs were favorably scored (n = 179,508 recommended; n = 20, 492 neutral) and used for array fabrication.

PRNP and affymetrix axiom array genotyping

PRNP genotyping for missense variants at codons 37, 95, 96, 116, and 226 was performed at GeneSeek Neogen (Lincoln, NE) as part of an existing commercial genotyping by sequencing (GBS) service. Briefly, the functional *PRNP* gene was PCR amplified using primers designed to be exclusionary to a processed pseudogene as previously described (O'Rourke *et al.* 2004), and the resulting amplicons were purified via AMPure XP beads as recommended by the manufacturer (Beckman Coulter); thus allowing for the creation of barcoded Illumina Nextera XT DNA libraries and amplicon sequencing on an Illumina MiSeq. *PRNP* genotypes were called from the aligned read pileups at GeneSeek Neogen, and delivered in text format. Affymetrix Axiom 200K genotyping was also performed at GeneSeek Neogen using the established Affymetrix best practices workflow; with genotypes delivered in text format. Affymetrix quality control thresholds were DQC ≥ 0.82, QC call rate $\geq$ 97%, passing samples in the project $\geq 95\%$, and average call rate for passing samples $\geq 97\%$. Collectively, 860 WTD samples with the desired metadata (*i.e.*, sex, age, U.S. general region) passed all Affymetrix QC filters; each with 125,968 SNP array genotypes, and paired *PRNP* genotypes, thus yielding a combined set of 125,973 SNP genotypes for analysis. Fifty-three WTD did not have CWD diagnostic data at the time of study. SNPs which did not convert on the Affymetrix Axiom 200K WTD array were primarily due to call rates below the best practices threshold (n = 37,197), and failures to meet thresholds in two or more best practices criteria (n = 36,045). All SNP conversion types are comprehensively summarized in DRYAD (Additional File 32).

GWAA and genomic prediction with cross validation

Prior to all analyses, a comparative marker map was created by aligning the WTD *PRNP* sequence and all Affymetrix Axiom 200K probe sequences with ARS-UCD1.2 (GCA_002263795.2) via blastn, thus providing comparative evidence for the origin of the array SNPs (*i.e.*, autosomal *vs.* non-autosomal); which was necessary because the draft *de novo* WTD genome assembly (GCF_002102435.1 Ovir.te_1.0) is unanchored (*i.e.*, by maps or *in situ* hybridization). The comparative marker map was joined to the combined set of all genotypes (*PRNP* + Affymetrix Axiom array), and quality control analyses were performed in SVS v8.8.2 or v8.8.3 (Golden Helix). Initially, pairwise IBS distances were computed to identify twins and duplicate samples. Eight samples present in both repositories (USDA APHIS, TPWD) were purposely duplicated for use as process controls, and correctly identified by IBS/IBD estimates (Additional File 31 in DRYAD). Eight additional samples were also identified as either duplicates or potential twins. In all cases, only the sample with the highest call rate was retained for further analyses. Additional quality control analyses and filtering were as follows: sample call rate ($< 97\%$ excluded), and thereafter, SNP filtering by call rate ($> 15\%$ missing excluded), MAF (< 0.01 excluded), polymorphism (monomorphic SNPs excluded), and Hardy-Weinberg Equilibrium (excludes SNPs with HWE *P*-value $<$ 1e-25), which yielded 123,987 SNPs for all analyses. *PRNP* SNPs which failed to endure quality control filtering included codons 95 (MAF < 0.01) and 116 (monomorphic), whereas codons 37, 96, and 226 remained. All GWAA's and genomic predictions with *k*-fold cross validation were performed on 807 WTD with recorded metadata that included sex, age, U.S. general region of origin (*i.e.*, Midwest, Northeast, South), and CWD diagnostics. CWD phenotypes used in all analyses were: CWD Scores (0 = non-detect, 1 = lymph node positive, 2 = lymph node and obex positive); CWD Binary (0 = non-detect, 1 = lymph node positive and/or obex positive). However, at the time of study, one WTD (sample ID: MQ6Q) only possessed diagnostic data for a

CWD positive tonsil biopsy, and thus was assigned a CWD Score and a CWD Binary phenotype of "1". All WTD GWAA's were performed using a mixed linear model with variance component estimates, as described and implemented in EMMAX, and executed in SVS v8.8.2 or v8.8.3 (Golden Helix), where all genotypes are also recoded as 0, 1, or 2, based on the incidence of the minor allele (Kang *et al.* 2010; Segura *et al.* 2012; Seabury *et al.* 2017; Smith *et al.* 2019). Briefly, the general mixed model can be specified as: $y = X\beta + Zu + \epsilon$, where y represents a $n \times 1$ vector of observed CWD phenotypes, X is a $n \times f$ matrix of fixed effects, β is a $f \times 1$ vector representing the coefficients of the fixed effects, u represents the unknown random effect, and Z is a $n \times t$ matrix relating the random effect to the CWD phenotypes of interest (Kang *et al.* 2010; Segura *et al.* 2012; Seabury *et al.* 2017; Smith *et al.* 2019). Herein, it is necessary to assume that $Var(u) = \sigma_g^2 K$ and $Var(\epsilon) = \sigma_e^2 I$, whereby $Var(y) = \sigma_g^2 ZKZ' + \sigma_e^2 I$, but in this study Z represents the identity matrix I, and K represents a relationship matrix of all WTD samples. To solve the mixed model equation using a generalized least squares approach, we must first estimate the relevant variance components (*i.e.*, σ_g^2 and σ_e^2) as previously described (Kang *et al.* 2010; Segura *et al.* 2012; Seabury *et al.* 2017; Smith *et al.* 2019). Variance components were estimated using the REML-based (restricted maximum likelihood) EMMA approach (Kang *et al.* 2010), with stratification accounted for and controlled using a genomic relationship matrix (G) (VanRaden 2008), as computed from the WTD genotypes. Genomic relationship matrix (GRM) heritability estimates ($[h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)]$) were produced as previously described (Kang *et al.* 2010; Segura *et al.* 2012; Seabury *et al.* 2017; Smith *et al.* 2019). Moreover, because previous WTD studies indicate that the probability of CWD infection increases with age (Grear *et al.* 2006), and that CWD may disparately affect male and female WTD in different U.S regions, including differences in clinical disease progression and mortality (Grear *et al.* 2006; Edmunds *et al.* 2016), the possibility for different CWD strains must be considered (Bistaffa *et al.* 2019). Therefore, the following fixed-effect covariates were specified for comparison of GWAAs: sex, age, U.S. region of origin, and the total number as well as types (0 = none-detected; 1 = lymph node only; 2 = lymph node and obex) of CWD positive diagnostic tissues, with one exception (*i.e.*, sample ID: MQ6Q), as noted above.

For all genomic prediction analyses involving k-fold cross validation, we used GBLUP as previously described (Taylor 2014) and implemented in SVS v8.8.2 or v8.8.3 (Golden Helix), where the variance components were again estimated using the REML-based EMMA technique (Kang *et al.* 2010) with a genomic relationship matrix (G) (VanRaden 2008; Taylor 2014). For WTD GBLUP analyses, consider the general mixed model equation: $y = X_f B_f + u + \epsilon$, across n WTD samples where fixed effects specified as B_f include the intercept and any additional covariates (*i.e.*, U.S. region, sex, age); but also assume $Var(\epsilon) = \sigma_e^2 I$, as above, and that the random effects u are additive genetic merits (*i.e.*, genomically estimated breeding values or GEBVs) for these WTD samples, which are produced from m markers as $u = M\alpha$, where M is a $n \times m$ matrix, and α is a vector where α_k is the allele substitution effect (ASE) for marker k. In this study, we used overall normalization for matrix M, as implemented in SVS v8.8.2 or v8.8.3 (Golden Helix), and explored solutions with and without gender corrections (*i.e.*, full dosage compensation, no dosage compensation) (Taylor 2014), to produce GEBVs for all WTD samples as well as estimates of ASE for all SNPs. Moreover, considering that all training set samples precede the validation set, we define $Z = [I|0]$, where the width and height of I is given as n_t, the width of the zero matrix is given as n_v, and the height of the zero matrix is n_t. Thus we can partition u, X_f, and y according to their origin (*i.e.*, training *vs.* validation set) as $u = \begin{bmatrix} u_t \\ u_v \end{bmatrix}$, $X_f = \begin{bmatrix} X_{ft} \\ X_{fv} \end{bmatrix}$, $y = \begin{bmatrix} y_t \\ y_v \end{bmatrix}$, and compute a genomic relationship matrix using all samples for use with the EMMA technique (Kang *et al.* 2010); to implement a mixed model for the training set as follows: $y_t = X_{ft} B_f + Zu + \epsilon_t$, where $Var(u) = \sigma_G^2 G$ and $Var(Zu) = \sigma_G^2 ZGZ'$. Finally, we predict the validation set phenotypes as: $\hat{y}_v = X_{fv}\hat{B}_f + \hat{u}_v$, from the intercept and any validation covariates X_{fv} as well as the predicted values of $\hat{u}_v$. Additional formulae and supporting documentation are available at https://doc.goldenhelix.com/SVS/latest/svsmanual/mixedModelMethods/overview.html#gblupproblemstmt. Notably, prior to this study, GEBVs were not estimated or utilized in WTD, and thus they cannot be expected to be intuitive or easily understood by U.S. WTD farmers or relevant regulatory agencies. However, the predicted CWD binary phenotypes are both intuitive and easily understood as estimates of enhanced or reduced susceptibility to CWD. Because GBLUP predicts CWD binary phenotypes across a range of values (*i.e.*, 0 to 1), SVS v8.8.2 and v8.8.3 (Golden Helix) considers predicted values of 0.5 and higher as "1", and < 0.5 as "0", thus facilitating the calculation of important summary statistics which require binary classifications. Justification for rounding is evident within the histograms representing the frequency distributions of the predicted CWD binary phenotypes (Fig. S1), the relevant GEBVs (Fig. S2), and the relationship between the predicted CWD binary phenotypes and the relevant GEBVs (Fig. S3); with an obvious break that occurs at 0.50 (Fig. S1-S3). Binary summary statistics for GBLUP-based genomic predictions with k-fold ($k = 3$; $k = 5$) cross validations (n = 50 iterations) were computed in SVS v8.8.2 or v8.8.3 (Golden Helix) as follows: Area Under the Curve as $AUC = \frac{U_1}{n_1 n_2}$, where n_1 is the sample size of observations with CWD binary phenotypes of 0, n_2 is the sample size of observations with CWD binary phenotypes of 1, and $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$, where R_1 is the sum of the ranks for the predicted binary CWD phenotypes with actual phenotypes of 0 (from CWD diagnostics); Mathews Correlation Coefficient as $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN)\cdot(FP+TN)\cdot(TP+FP)\cdot(FN+TN)}}$, where TP is the number of true CWD positives (from CWD diagnostics), TN is the number of true CWD non-detects (from CWD diagnostics), FP is the number of false positives, and FN is the number of false non-detects among the predicted CWD binary phenotypes; Genomic Prediction Accuracy as $ACC = \frac{TP+TN}{TP+FN+FP+TN}$; Sensitivity (true positive rate) as $TPR = \frac{TP}{TP+FN}$; Specificity (true negative rate) as $SPC = \frac{TN}{FP+TN}$; Root Mean Square Error as $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$. For the GBLUP predicted CWD Scores (0, 1, 2), we also produced and report relevant summary statistics from the k-fold ($k = 3$; $k = 5$) cross validations (n = 50 iterations) computed in SVS v8.8.2 or v8.8.3 (Golden Helix) as follows: Pearson's Product-Moment Correlation Coefficient as $r_{y,\hat{y}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{(n-1)s_y s_{\hat{y}}}$ where s_y and $s_{\hat{y}}$ are the standard deviations; Residual Sum of Squares as $RSS = \sum_{i=1}^{n}(y_i - \hat{y})^2$; Total Sum of Squares $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$; R-Squared as $R^2 = 1 - \frac{RSS}{TSS}$; Root Mean Square Error as $RMSE = \sqrt{\frac{RSS}{n}}$; Mean Absolute Error as $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$.

Randomizing and blinding

All GBLUP-based k-fold ($k = 3$; $k = 5$) cross validations (*i.e.*, CWD binary; CWD-scores) were performed using automated random sampling to define the validation set (*i.e.*, to predict on) and the training set, for the specified values of k.

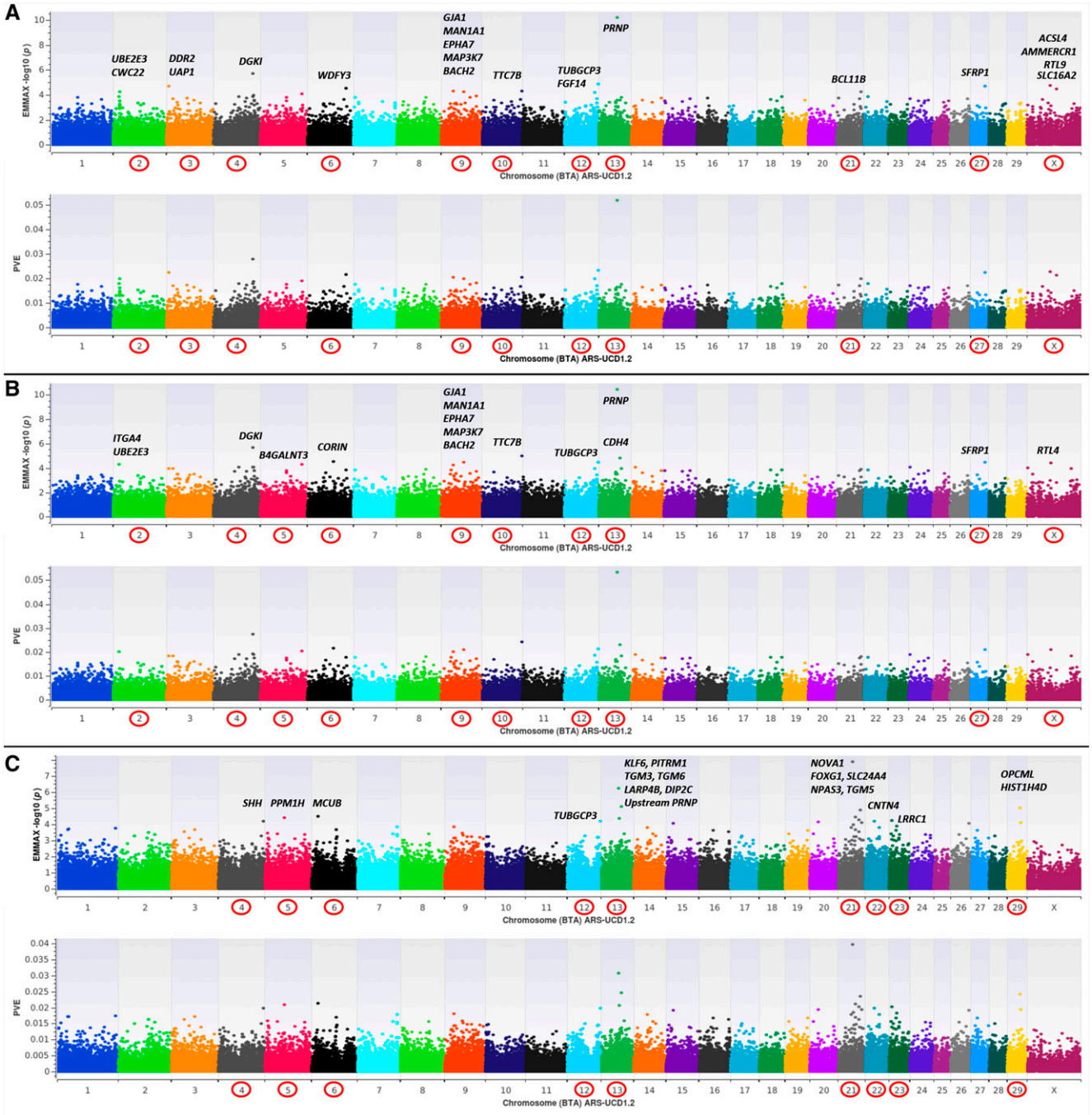


Figure 1 EMMAX binary case-control (0, 1) genome-wide association analyses (GWAA) for Chronic Wasting Disease (CWD) in farmed U.S. white-tailed deer (*Odocoileus virginianus*; hereafter WTD). All dual-panel manhattan plots depict -log10 *P*-values and the proportion of phenotypic variance explained (PVE) by white-tailed deer marker-effects on the y-axis, and the comparative position of all probe sequences on the x-axis, as inferred by blastn alignment with the bovine genome (ARS-UCD1.2). All analyses include diagnostically confirmed CWD positive (n = 284) and CWD non-detect (n = 523) WTD, and marker-based GRM heritability estimates ($[h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)]$) (Kang *et al.* 2010; Segura *et al.* 2012; Seabury *et al.* 2017; Smith *et al.* 2019). **a**, EMMAX GWAA for CWD with no fixed-effect covariates, high GRM heritability estimates ($h^2 = 0.637 \pm 0.070$) and relevant positional candidate genes. **b**, EMMAX GWAA for CWD with fixed-effect covariates (sex, age, U.S. region), high GRM heritability estimates ($h^2 = 0.546 \pm 0.076$) and relevant positional candidate genes. **c**, EMMAX GWAA for CWD with fixed-effect covariates (sex, age, U.S. region, CWD-scores), moderate GRM heritability estimates ($h^2 = 0.337 \pm 0.079$) and relevant positional candidate genes.

Data availability

Accession codes are as follows: Illumina sequence data (SRA: SRR10313416-SRR10313421); genotype and phenotype data (DRYAD https://doi.org/10.5061/dryad.xd2547dcw).

RESULTS AND DISCUSSION

Using three reduced representation libraries (n = 374 farmed U.S. WTD) and Illumina paired-end sequencing for reference mapping and variant prediction, we successfully constructed a custom

Affymetrix Axiom 200K SNP array for the WTD (See Methods). All probe sequences were aligned to the new PacBio long-read bovine genome assembly (ARS-UCD1.2; GCA_002263795.2), thus creating a comparative marker map. Thereafter, we genotyped a cohort of farmed WTD diagnostically classified (See Methods) as CWD positive (n = 284) and CWD non-detect (n = 523) from three U.S. geographic regions (Midwest, Northeast, South). Genome-wide association analyses were conducted using a mixed linear model with genomic relationship matrix and variance component analysis, thus yielding marker-based heritability estimates (GRM heritability), as implemented in EMMAX (Kang *et al.* 2010; Segura *et al.* 2012). All GWAA's were carried out using 123,987 quality filtered SNPs for two dependent variables including a binary case-control variable (0 = non-detect; 1 = CWD positive) (Figure 1), and an interval variable (CWD-scores) which simultaneously reflects both the total number of CWD positive diagnostic tissues (*i.e.*, 0, 1, 2) as well as the positive tissue types (*i.e.*, 1 = lymph node only; 2 = lymph node and obex; Figure 2); with non-zero CWD-scores accurately modeling the natural progression of disease (Thomsen *et al.* 2012). Across all GWAA's (Figure 1, Figure 2), GRM heritability estimates were moderate to high (*i.e.*, $h^2 = 0.337 \pm 0.079 - 0.637 \pm 0.070$); thus confirming that differential susceptibility to CWD in WTD is under genetic control, and that host genomic background also influences variation in disease progression. Herein we also confirm the *PRNP* gene as a major risk locus, and specifically, the codon 96 missense variant (G96S; binary case-control *P*-value $<$ 6.30E-11; CWD-scores *P*-value $<$ 1.49E-15) as well as one upstream promoter SNP (Affx-574071595; CWD-scores *P*-value $\leq$ 5.40E-06) as being significantly associated with differential susceptibility to CWD, and with variation in natural disease progression among WTD (Figure 1, Figure 2, Table S1) (O'Rourke *et al.* 2004; Seabury *et al.* 2007). However, it should also be noted that 11 CWD-positive WTD possessed the codon 96SS genotype, and the proportion of phenotypic variance explained (PVE) by even the largest-effect *PRNP* SNPs detected across all GWAA's (G96S, Promoter Affx-574071595) was $<$ 0.11 (Figure 1, Figure 2), indicating that loci other than *PRNP* influence differential susceptibility and disease progression. These results are compatible with prior analyses performed in mice; where incubation times for transmissible spongiform encephalopathies were largely influenced by a genetic architecture independent of *PRNP*

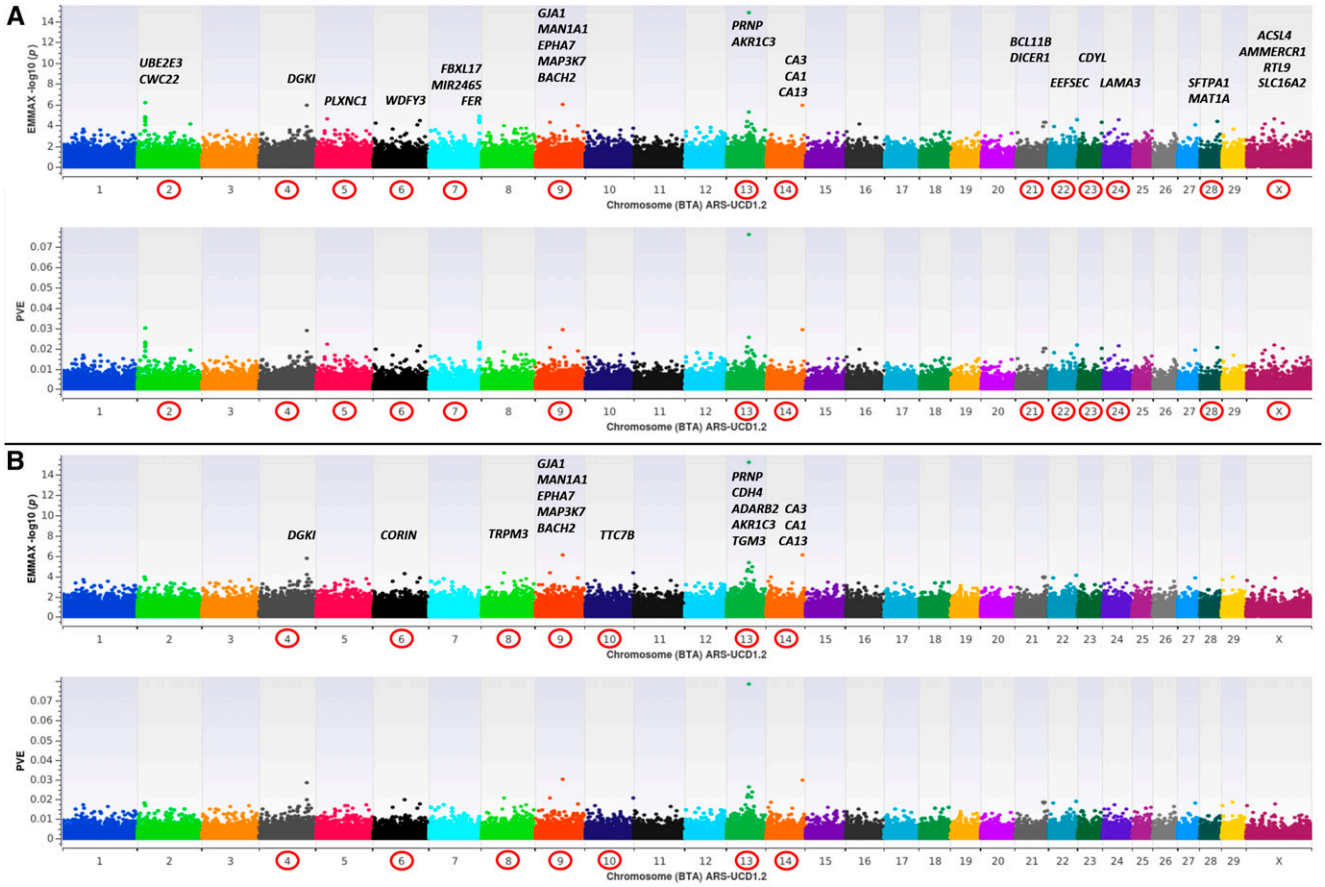


Figure 2 EMMAX genome-wide association analyses (GWAA) for Chronic Wasting Disease (CWD) in farmed U.S. white-tailed deer (*Odocoileus virginianus*; hereafter WTD) using an interval variable (CWD-scores) to simultaneously reflect both the total number of CWD positive diagnostic tissues (*i.e.*, 0, 1, 2) as well as the positive tissue types (*i.e.*, 1 = lymph node only; 2 = lymph node and obex). All dual-panel manhattan plots depict -log10 *P*-values and the proportion of phenotypic variance explained (PVE) by white-tailed deer marker-effects on the y-axis, and the comparative position of all probe sequences on the x-axis, as inferred by blastn alignment with the bovine genome (ARS-UCD1.2). All analyses include diagnostically confirmed CWD positive (n = 284) and CWD non-detect (n = 523) WTD, and marker-based GRM heritability estimates ($[h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)]$) (Kang *et al.* 2010; Segura *et al.* 2012; Seabury *et al.* 2017; Smith *et al.* 2019). **a**, EMMAX GWAA for CWD-scores with no fixed-effect covariates, high GRM heritability estimates ($h^2 = 0.589 \pm 0.069$) and relevant positional candidate genes. **b**, EMMAX GWAA for CWD-scores with fixed-effect covariates (sex, age, U.S. region), high GRM heritability estimates ($h^2 = 0.515 \pm 0.075$) and relevant positional candidate genes.

Table 1 Summary of chronic wasting disease genomic predictions in farmed U.S. white-tailed deer (*Odocoileus virginianus*)

k-Fold Subsample	GBLUP Model Covariates	Mean AUC (*SD*)[a]	Mean Matthews Coefficient (*SD*)	Mean Genomic Prediction Accuracy (*SD*)	Mean Sensitivity (*SD*)	Mean Specificity (*SD*)	Mean RMSE (*SD*)[b]
***k* = 3**	None	0.8471 (0.0068)	0.5871 (0.0152)	0.8167 (0.0066)	0.6447 (0.0117)	0.9101 (0.0091)	0.3768 (0.0043)
***k* = 3**	Sex, Age, U.S. Region	0.8534 (0.0063)	0.5787 (0.0158)	0.8119 (0.0070)	0.6735 (0.0122)	0.8870 (0.0081)	0.3746 (0.0040)
***k* = 5**	None	0.8485 (0.0053)	0.5940 (0.0132)	0.8198 (0.0057)	0.6496 (0.0110)	0.9121 (0.0060)	0.3754 (0.0028)
***k* = 5**	Sex, Age, U.S. Region	0.8542 (0.0047)	0.5870 (0.0123)	0.8159 (0.0055)	0.6716 (0.0110)	0.8942 (0.0066)	0.3734 (0.0025)

[a] Area Under the Curve (AUC; Wilcoxon Mann Whitney Method; See Methods).
[b] Root Mean Square Error (RMSE; See Methods).

(Iyegbe *et al.* 2010). Across all GWAA's (n = 123,987 SNPs), only 61 SNPs met a nominal significance level for polygenic traits (*P*-value ≤ 5E-05) (Wellcome Trust Case Control Consortium 2007; Seabury *et al.* 2017), with 17 detected in more than one GWAA. This is relatively unsurprising since EMMAX is known to produce conservative *P*-values; particularly when large-effect regions exist (Zhou and Stephens 2012). Moreover, the architecture of both investigated CWD traits (Figure 1, Figure 2) is such that few large-effect regions exist (*i.e.*, PVE ≥ 0.03); but together with many smaller-effect regions, a significant proportion of the phenotypic variance can be explained. Interestingly, an investigation of all GWAA's revealed many of the same positional candidate genes (Figure 1, Figure 2; PVE ≥ 0.02); the majority of which have been implicated in the pathophysiology of various prion diseases including scrapie (*i.e.*, *ACSL4*, *CA3*, *KLF6*), bovine spongiform encephalopathy (*i.e.*, *NPAS3*, *BACH2*, *EPHA7*, *ITGA4*), spontaneous and familial Creutzfeldt-Jakob disease (*i.e.*, *HIST1H4D/OPCML*, *LAMA3*, *TTC7B*), and various other neurodegenerative conditions including Alzheimer's (*i.e.*, *DGKI*, *SFRP1*, *SLC24A4*) and Parkinson's disease (*BCL11B*) (Scherzer *et al.* 2007; Tang *et al.* 2009; Tortosa *et al.* 2011; Silver *et al.* 2012; Cohen *et al.* 2013; Lambert *et al.* 2013; Filali *et al.* 2014; Lee *et al.* 2014; Kipkorir *et al.* 2015; Xerxa *et al.* 2016; Esteve *et al.* 2019; Majer *et al.* 2019). Summary data for all GWAA's and positional candidate genes (Figure 1, Figure 2) as well as the corresponding P-P plots are provided in the supplementary information (Table S1; Additional Files 1-5 in DRYAD).

To investigate the potential for developing a genomically-estimated CWD eradication program for farmed WTD, we used genomic best linear unbiased prediction (GBLUP) in conjunction with *k*-fold cross validation and random sampling (Taylor 2014). Briefly, WTD data (*i.e.*, genotypes, CWD diagnostic phenotypes, and other metadata) were randomly partitioned into *k* -subsamples (*k* = 3, *k* = 5), and one of these subsamples was then selected (*i.e.*, within a discrete fold) to serve as the validation set for genomic prediction; thus the GBLUP model was fit using the remaining (*i.e.*, *k* -1) subsamples within that fold (*i.e.*, the training data); until all subsamples were used for both training and prediction. All cross validations with random sampling were run for 50 iterations, with each iteration consisting of either three or five folds (*k* = 3, *k* = 5), and summary statistics were produced (See Table 1; Methods; Additional Files 6-29 in DRYAD). Binary GBLUP models fit with no fixed-effect covariates (*k* = 3, *k* = 5) produced high mean genomic prediction accuracies (≥ 0.8167) and specificities (≥ 0.9101), with small standard deviations, but lower mean sensitivities (≤ 0.6496), indicating that false negatives pose the greatest challenge for reducing the prevalence of CWD via genomic prediction (Table 1; See Methods). Similar results were also obtained when binary GBLUP models were fit using sex, age, and U. S. region of origin as fixed-effect covariates (*k* = 3, *k* = 5; Table 1). However, in addition to false negatives, some false positive genomic predictions were also observed; most likely due to underlying genomic susceptibility coupled with either very early stages of disease (*i.e.*, CWD non-detect diagnostically) and/or differences in exposure (Tsairidou *et al.* 2014). All results were robust to the inclusion or exclusion of non-autosomal loci (*i.e.*, X, MT; 123,987 SNPs *vs.* 120,808 SNPs, respectively), and to full dosage compensation *vs.* no dosage compensation when putative X-linked SNPs were included (Table 1; Additional Files 6-29 in DRYAD). GBLUP models fit with the CWD-scores (0, 1, 2), thus reflecting the natural progression of disease, produced lower mean genomic prediction accuracies (*i.e.*, ≤ 0.6007; See Methods; DRYAD), regardless of the inclusion or exclusion of fixed-effect covariates (sex, age, U.S. region), non-autosomal loci, or the implementation of full dosage compensation *vs.* no dosage compensation (*k* = 3, *k* = 5).

CONCLUSIONS

Herein, we demonstrate that differential susceptibility to CWD and variation in natural disease progression are both heritable, polygenic traits in farmed U.S. WTD, and that genome-wide SNP data can be used to produce accurate genomic predictions for risk (≥ 0.8167); thereby providing the first novel strategy for reducing the prevalence of CWD. Moreover, given the genomic architecture of these traits, we also demonstrate that *PRNP* genotyping alone cannot be expected to facilitate an eradication program, or to rapidly reduce the overall prevalence of CWD in farmed U.S. WTD.

AUTHOR CONTRIBUTIONS

C.M.S designed the research with input from T.A.N. M.A.L. and T.A.N. provided CWD diagnostic data, other relevant animal metadata, and biological samples for DNA isolation from existing agency repositories. C.M.S. and D. L. isolated and quantified DNA with assistance from P.F. E.K.B. prepared hair cards and quantified DNA. R.P.M. and C.D.J. prepared and sequenced reduced representation libraries. C.M.S. performed trimming, reference mapping, and variant prediction. C.M.S. designed the Affymetrix Axiom 200K SNP array with assistance from D.L.O. D.L.O. performed blastn searches to create comparative marker maps for the array. D.L.O. submitted reads to NCBI SRA. C.M.S. performed quality filtering of genotypes, genome wide association analyses with marker-based heritability estimates, alignment with positional candidate genes, and genomic predictions with *k*-fold cross validation and summary statistics. C.M.S. wrote the manuscript; incorporating input from D.L.O., E.K.B., D.L., P.F., R.P.M., C.D.J., M.A.L., and T.A.N. All authors edited and approved the manuscript.

ACKNOWLEDGMENTS

This material was made possible, in part, by a Cooperative Agreement from the United States Department of Agriculture's Animal and Plant Health Inspection Service (APHIS). It may not necessarily express APHIS' views. CMS acknowledges this support from USDA-MIS-Animal and Plant Health Inspection Service (grant AP17VSSPRS00C126)

as well as support from Texas Parks and Wildlife Department (grant 475613) and a charitable gift from the American Chronic Wasting Disease Foundation.

LITERATURE CITED

Bistaffa, E., T. T. Vuong, F. A. Cazzaniga, L. Tran, G. Salzano *et al.*, 2019 Use of different RT-QuIC substrates for detecting CWD prions in the brain of Norwegian cervids. Sci. Rep. 9: 18595. https://doi.org/10.1038/s41598-019-55078-x

Cohen, E., D. Avrahami, K. Frid, T. Canello, E. Levy Lahad *et al.*, 2013 Snord 3A: A molecular marker and modulator of prion disease progression. PLoS One 8: e54433. https://doi.org/10.1371/journal.pone.0054433

Edmunds, D. R., M. J. Kauffman, B. A. Schumaker, F. G. Lindzey, W. E. Cook *et al.*, 2016 Chronic Wasting Disease Drives Population Decline of White-tailed Deer. PLoS One 11: e0161127. https://doi.org/10.1371/journal.pone.0161127

Esteve, P., J. Rueda-Carrasco, M. Inés Mateo, M. J. Martin-Bermejo, J. Draffin *et al.*, 2019 Elevated levels of Secreted-Frizzled-Related-Protein 1 contribute to Alzheimer's disease pathogenesis. Nat. Neurosci. 22: 1258–1268. https://doi.org/10.1038/s41593-019-0432-1

Filali, H. I., Martín-Burriel I., Harders F., Varona L., Hedman C. *et al.*, 2014 Gene expression profiling of mesenteric lymph nodes from sheep with natural scrapie. BMC Genomics 15: 59. https://doi.org/10.1186/1471-2164-15-59

Gavin, C., D. Henderson, S. L. Benestad, M. Simmons, and A. Adkin, 2019 Estimating the amount of Chronic Wasting Disease infectivity passing through abattoirs and field slaughter. Prev. Vet. Med. 166: 28–38. https://doi.org/10.1016/j.prevetmed.2019.02.016

Grear, D. A., M. D. Samuel, J. A. Langenberg, and D. Keane, 2006 Demographic patterns and harvest vulnerability of chronic wasting disease infected white-tailed deer in Wisconsin. J. Wildl. Manage. 70: 546–553. https://doi.org/10.2193/0022-541X(2006)70[546:DPAHVO]2.0.CO;2

Halley, Y. A., D. L. Oldeschulte, E. K. Bhattarai, J. Hill, R. P. Metz *et al.*, 2015 Northern Bobwhite (*Colinus virginianus*) Mitochondrial Population Genomics Reveals Structure, Divergence, and Evidence for Heteroplasmy. PLoS One 10: e0144913. https://doi.org/10.1371/journal.pone.0144913

Halley, Y. A., S. E. Dowd, J. E. Decker, P. M. Seabury, E. Bhattarai *et al.*, 2014 A draft de novo genome assembly for the northern bobwhite (*Colinus virginianus*) reveals evidence for a rapid decline in effective population size beginning in the Late Pleistocene. PLoS One 9: e90240. https://doi.org/10.1371/journal.pone.0090240

Iyegbe, C. O., O. O. Abiola, C. Towlson, J. F. Powell, and S. A. Whatley, 2010 Evidence for Varied Aetiologies Regulating the Transmission of Prion Disease: Implications for Understanding the Heritable Basis of Prion Incubation Times. PLoS One 5: e14186. https://doi.org/10.1371/journal.pone.0014186

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42: 348–354. https://doi.org/10.1038/ng.548

Kipkorir, T., C. M. Colangelo, and L. Manuelidis, 2015 Proteomic analysis of host brain components that bind to infectious particles in Creutzfeldt-Jakob disease. Proteomics 15: 2983–2998. https://doi.org/10.1002/pmic.201500059

Lambert, J. C., C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims *et al.*, 2013 Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. 45: 1452–1458. https://doi.org/10.1038/ng.2802

Lee, S. M., M. Chung, K. J. Hwang, Y. R. Ju, J. W. Hyeon *et al.*, 2014 Biological network inferences for a protection mechanism against familial Creutzfeldt-Jakob disease with E200K pathogenic mutation. BMC Med. Genomics 7: 52. https://doi.org/10.1186/1755-8794-7-52

Majer, A., S. J. Medina, D. Sorensen, M. J. Martin, K. L. Frost *et al.*, 2019 The cell type resolved mouse transcriptome in neuron-enriched brain tissues from the hippocampus and cerebellum during prion disease. Sci. Rep. 9: 1099. https://doi.org/10.1038/s41598-018-37715-z

Moreno, J. A., and G. C. Telling, 2018 Molecular Mechanisms of Chronic Wasting Disease Prion Propagation. Cold Spring Harb. Perspect. Med. 8. https://doi.org/10.1101/cshperspect.a024448

O'Rourke, K. I., T. R. Spraker, L. K. Hamburg, T. E. Besser, K. A. Brayton *et al.*, 2004 Polymorphisms in the prion precursor functional gene but not the pseudogene are associated with susceptibility to chronic wasting disease in white-tailed deer. J. Gen. Virol. 85: 1339–1346. https://doi.org/10.1099/vir.0.79785-0

Oldeschulte, D. L., Y. A. Halley, M. L. Wilson, E. K. Bhattarai, W. Brashear *et al.*, 2017 Annotated Draft Genome Assemblies for the Northern Bobwhite (*Colinus virginianus*) and the Scaled Quail (*Callipepla squamata*) Reveal Disparate Estimates of Modern Genome Diversity and Historic Effective Population Size. G3 (Bethesda) 7: 3047–3058. https://doi.org/10.1534/g3.117.043083

Osterholm, M. T., C. J. Anderson, M. D. Zabel, J. M. Scheftel, K. A. Moore *et al.*, 2019 Chronic Wasting Disease in Cervids: Implications for Prion Transmission to Humans and Other Animal Species. MBio 10. https://doi.org/10.1128/mBio.01091-19

Scherzer, C. R., A. C. Eklund, L. J. Morse, Z. Liao, J. J. Locascio *et al.*, 2007 Molecular markers of early Parkinson's disease based on gene expression in blood. Proc. Natl. Acad. Sci. USA 104: 955–960. https://doi.org/10.1073/pnas.0610204104

Seabury, C. M., C. A. Gill, J. W. Templeton, J. B. Dyar, J. N. Derr *et al.*, 2007 Molecular characterization of the Rocky Mountain elk (*Cervus elaphus nelsoni*) *PRNP* putative promoter. J. Hered. 98: 678–686. https://doi.org/10.1093/jhered/esm091

Seabury, C. M., D. L. Oldeschulte, M. Saatchi, J. E. Beever, J. E. Decker *et al.*, 2017 Genome-wide association study for feed efficiency and growth traits in U.S. beef cattle. BMC Genomics 18: 386. https://doi.org/10.1186/s12864-017-3754-y

Seabury, C. M., E. K. Bhattarai, J. F. Taylor, G. G. Viswanathan, S. M. Cooper *et al.*, 2011 Genome-wide polymorphism and comparative analyses in the white-tailed deer (*Odocoileus virginianus*): a model for conservation genomics. PLoS One 6: e15811. https://doi.org/10.1371/journal.pone.0015811

Segura, V., B. J. Vilhjálmsson, A. Platt, A. Korte, Ü. Seren *et al.*, 2012 An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. 44: 825–830. https://doi.org/10.1038/ng.2314

Silver, M., E. Janousova, X. Hua, P. M. Thompson, G. Montana *et al.*, 2012 Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. Neuroimage 63: 1681–1694. https://doi.org/10.1016/j.neuroimage.2012.08.002

Smith, J. L., M. L. Wilson, S. M. Nilson, T. N. Rowan, D. L. Oldeschulte *et al.*, 2019 Genome-wide association and genotype by environment interactions for growth traits in U.S. Gelbvieh cattle. BMC Genomics 20: 926. https://doi.org/10.1186/s12864-019-6231-y

Sollars, E. S., A. L. Harper, L. J. Kelly, C. M. Sambles, R. H. Ramirez-Gonzalez *et al.*, 2017 Genome sequence and genetic diversity of European ash trees. Nature 541: 212–216. https://doi.org/10.1038/nature20786

Tang, Y., W. Xiang, S. A. Hawkins, H. A. Kretzschmar, and O. Windl, 2009 Transcriptional changes in the brains of cattle orally infected with the bovine spongiform encephalopathy agent precede detection of infectivity. J. Virol. 83: 9464–9473. https://doi.org/10.1128/JVI.00352-09

Taylor, J. F., 2014 Implementation and accuracy of genomic selection. Aquaculture 420–421: S8–S14. https://doi.org/10.1016/j.aquaculture.2013.02.017

Thomsen, B. V., D. A. Schneider, K. I. O'Rourke, T. Gidlewski, J. McLane *et al.*, 2012 Diagnostic accuracy of rectal mucosa biopsy testing for chronic wasting disease within white-tailed deer (*Odocoileus virginianus*) herds in North America: Effects of age, sex, polymorphism at *PRNP* codon 96, and disease progression. J. Vet. Diagn. Invest. 24: 878–887. https://doi.org/10.1177/1040638712453582

Tortosa, R., X. Castells, E. Vidal, C. Costa, M. Ruiz de Villa *et al.*, 2011 2011 Central nervous system gene expression changes in a transgenic mouse model for bovine spongiform encephalopathy. Vet. Res. (Faisalabad) 42: 109. https://doi.org/10.1186/1297-9716-42-109

Tsairidou, S., J. A. Wooliams, A. R. Allen, R. A. Skuce, S. H. McBride *et al.*, 2014 Genomic Prediction for Tuberculosis Resistance in Dairy Cattle. PLoS One 9: e96728. https://doi.org/10.1371/journal.pone.0096728

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423. https://doi.org/10.3168/jds.2007-0980

Wellcome Trust Case Control Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678. https://doi.org/10.1038/nature05911

Williams, E. S., and S. Young, 1980 Chronic wasting disease of captive mule deer: A spongiform encephalopathy. J. Wildl. Dis. 16: 89–98. https://doi.org/10.7589/0090-3558-16.1.89

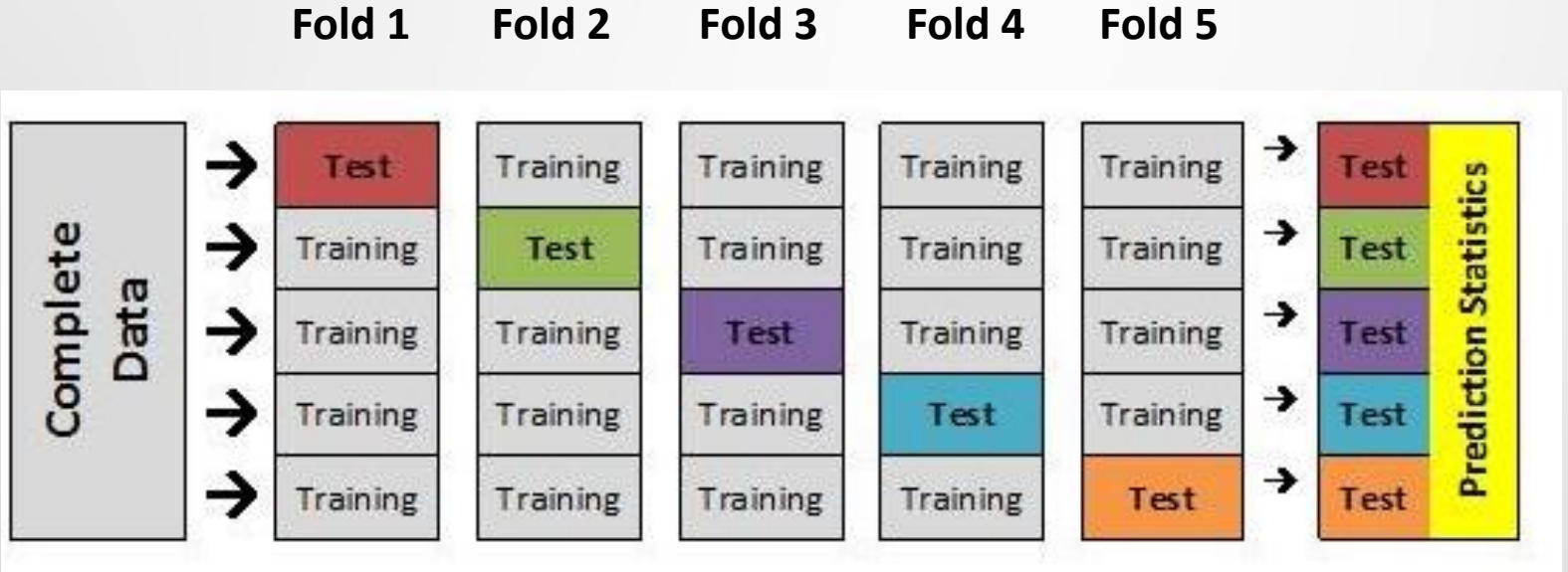
Xerxa, E., M. Barbisin, M. N. Chieppa, H. Krmac, E. Vallino Costassa *et al.*, 2016 Whole Blood Gene Expression Profiling in Preclinical and Clinical Cattle Infected with Atypical Bovine Spongiform Encephalopathy. PLoS One 11: e0153425. https://doi.org/10.1371/journal.pone.0153425

Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44: 821–824. https://doi.org/10.1038/ng.2310

Communicating editor: D. J. de Koning

So What Are the Details?

Machine Learning: Cross Validation



This is considered 1 Iteration. We did 50 Iterations of k= 5 and k = 3.

So What Are the Details?

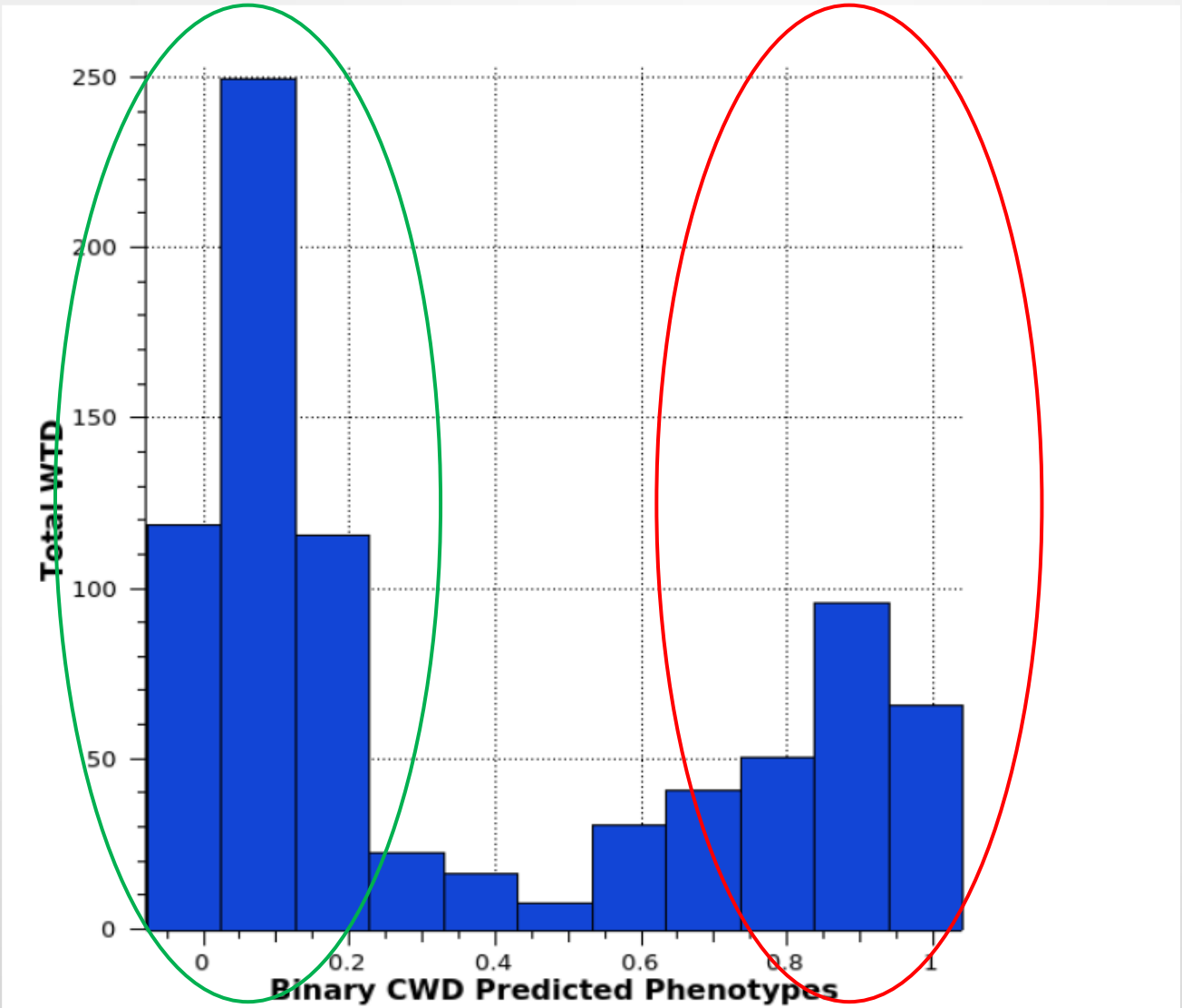


Figure S1 (A). Histogram of binary CWD phenotypes predicted using GBLUP with full dosage compensation and no fixed-effect covariates (n = 123,987 SNPs) for 807 farmed U.S. white-tailed deer (*Odocoileus virginianus*).

So What Are the Details?

Summary of chronic wasting disease genomic predictions in farmed U.S. white-tailed deer (*Odocoileus virginianus*)

K-Fold Subsample	GBLUP Model Covariates	Mean AUC (*SD*)[a]	Mean Matthews Coefficient (*SD*)	Mean Genomic Prediction Accuracy (*SD*)	Mean Sensitivity (*SD*)	Mean Specificity (*SD*)	Mean RMSE (*SD*)[b]
K = 3	None	0.8471 (0.0068)	0.5871 (0.0152)	0.8167 (0.0066)	0.6447 (0.0117)	0.9101 (0.0091)	0.3768 (0.0043)
K = 3	Sex, Age, U.S. Region	0.8534 (0.0063)	0.5787 (0.0158)	0.8119 (0.0070)	0.6735 (0.0122)	0.8870 (0.0081)	0.3746 (0.0040)
K = 5	None	0.8485 (0.0053)	0.5940 (0.0132)	0.8198 (0.0057)	0.6496 (0.0110)	0.9121 (0.0060)	0.3754 (0.0028)
K = 5	Sex, Age, U.S. Region	0.8542 (0.0047)	0.5870 (0.0123)	0.8159 (0.0055)	0.6716 (0.0110)	0.8942 (0.0066)	0.3734 (0.0025)

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \quad TPR = \frac{TP}{TP + FN} \quad SPC = \frac{TN}{FP + TN}$$