GGP-F250 and Loss of Function Variants

Jerry Taylor, Bob Schnabel, Barry Simpson, Jared Decker, Megan Rolf, Brian Kinghorn, Alison Van Eenennaam, Mike MacNeil, Scott Brown, Mike Smith and Dave Patterson

BIF 2017 June 2, 2017

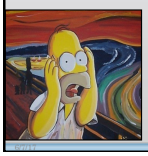
GGPF250

Disclosures

- Stockholder and Scientific Advisory Board Member of Recombinetics

recombinetics

- Scientific Advisory Board Member and Consultant for GeneSeek, a Neogen Company



GENESEEK a Neogen® Company

Evidence For Lethals

USDA AIP RESEARCH REPORT GENOMICS (09-13) Updated May 2, 2013

Haplotype tests for recessive disorders that affect fertility and other traits

J.B. Cole, P.M. VanRaden, D.J. Null, J.L. Hutchison, T.A. Cooper, and S.M. Hubbard
Animal Improvement Program, Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350
301-504-8334 (voice) ~ 301-504-8092 (fax) ~ john.cole@ars.usda.gov ~ http://aipl.arsusda.gov/

Breed	Haplotype	Phenotype	Gene	Frequency (%)	Chromosome	Position (bp)	Reference
Ayrshire	AH1	Ptosis, intellectual disability, retarded growth and mortality (PIRM)	UBE3B	13	17	83923497	Cooper et al. (2014), Venhoranta et al. (2014)
Brown Swiss	BH1	Abortion	—	6.67	7	42,811,272 – 47,002,161	VanRaden et al. (2011)
	BH2	Abortion	TUBD2	7.78	19	1083520	Schwarzenbacher et al. (2016)
Holstein	HH0	Brachyspina and in utero death	FANCI	2.76	21	21,184,869 – 21,188,198	Agerholm et al. (2006), Charlier et al. (2012)
	HH1	Abortion	APAF1	1.92	5	61150400	Adams et al. (2012)
	HH2	Abortion	—	1.66	1	94,860,836 – 96,553,339	VanRaden et al. (2011), McClure et al. (2014)
	HH3	Abortion	SMC2	2.95	8	95410507	Doetschyler et al. (2014), McClure et al. (2014)
	HH4	Abortion	GART	0.37	1	1277227	Fritz et al. (2013)
	HH5	Abortion	TRIM	2.22	9	93,223,651 – 93,370,998	Cooper et al. (2013), Schütz et al. (2016)
Jersey	JH1	Abortion	CWC15	12.1	15	15707169	Sonstegard et al. (2013)
	JH2	Abortion	—	1.3	26	8,812,759 – 9,414,082	VanRaden et al. (2014)

https://aipl.arsusda.gov/reference/recessive_haplotypes_ARR-G3.html#Venhoranta14

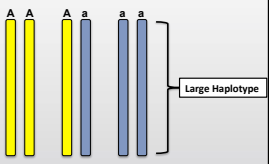
The fact that more have been identified in Holstein may be due to:
1. More sires used in AI to propagate new mutations
2. Larger genotyped population leads to greater power to detect

How Were These Detected?

In large panmictic populations the Hardy-Weinberg Equilibrium principle indicates that:

Gametes	Female			
	Alleles		A	a
		Frequencies	p	q
Male	A	p	p^2	pq
	a	q	pq	q^2

Table : Punnett square for Hardy-Weinberg equilibrium



In a sample of N individuals we would expect to see	Np^2	2Npq	Nq^2	
So…if q=2% and N = 1,000,000 animals genotyped			400	Expected
But what if…			0	Observed?

Power of Detection

$P(0) = (1 - q^2)^N$ = Probability of seeing no homozygotes for rare allele if not lethal

q \ N	1,000	5,000	10,000	20,000	100,000	1,000,000	10,000,000
0.001	0.999000499	0.99501248	0.99004983	0.98019866	0.90483737	0.36787926	4.53997E-05
0.005	0.975309607	0.88249552	0.77879835	0.60652687	0.08208243	1.3884E-11	2.6609E-109
0.010	0.904832894	0.6065155	0.36786105	0.13532175	4.5377E-05	3.7015E-44	0
0.015	0.798496004	0.32461137	0.10537254	0.01110337	1.6876E-10	1.8739E-98	0
0.020	0.670266408	0.13528115	0.01830099	0.00033493	4.2145E-18	1.768E-174	0
0.025	0.535156852	0.04389403	0.00192669	3.7121E-06	7.0487E-28	3.028E-272	0
0.030	0.406404934	0.01108651	0.00012291	1.5107E-08	7.8686E-40	0	0
0.050	0.081828456	3.6688E-06	1.346E-11	1.8117E-22	1.952E-109	0	0
0.100	4.31712E-05	1.4996E-22	2.2488E-44	5.057E-88	0	0	0

<- All Other Breeds -> Angus Holstein

If we perform multiple tests throughout the genome we need to be stringent on the Type I error rate.

Evidence For Lethals

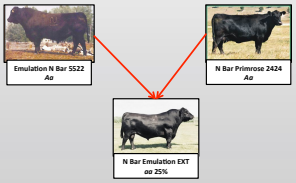
In a sample of N = 4000 registered Angus we find evidence for 7 genomic regions with haplotypes that are never observed as homozygotes (we would expect to see at least 2 based on sample size).

Chr	Location (Mb)	Length (Mb)	Haplotype Frequency[1]
1	27.7-29.0	1.3	0.023
4	82.5-84.0	1.5	0.076
8	62.0-63.0	1.0	0.023
12	60.0-61.2	1.2	0.032
15	82.3-83.1	0.8	0.038
17	46.5-47.5	1.0	0.045
29	43.0-44.2	1.2	0.044

Transmission Disequilibrium Test

If both parents are heterozygotes the probability the son is not *aa* is 75%

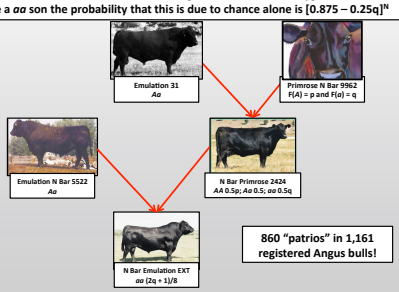
If I had N families with both parents heterozygotes and I never see a *aa* son the probability that this is due to chance alone is 0.75^N



Transmission Disequilibrium Test

If sire and maternal grandsire are heterozygotes the probability the son is not *aa* is [0.875 – 0.25q] (about 87.5% if q is small)

If I had N families with sire and maternal grandsire both heterozygotes and I never see a *aa* son the probability that this is due to chance alone is $[0.875 – 0.25q]^N$



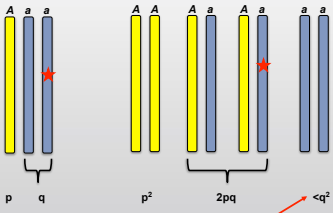
860 "patrios" in 1,161 registered Angus bulls!

TDT in Angus

Chr	Location (Mb)	Length (Mb)	Haplotype Frequency[1]	Number of Patrios[2]	Probability[3]
1	27.7-29.0	1.3	0.023	39	0.0042
4	82.5-84.0	1.5	0.076	127	2.66E-09
8	62.0-63.0	1.0	0.023	35	0.0074
12	60.0-61.2	1.2	0.032	46	0.0014
15	82.3-83.1	0.8	0.038	31	0.011
17	46.5-47.5	1.0	0.045	49	0.00076
29	43.0-44.2	1.2	0.044	118	3.22E-08

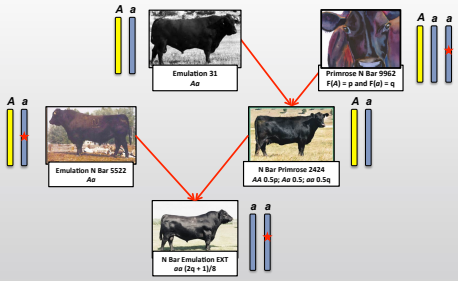
Problem With HWE Analysis

Wild type haplotype and haplotype harboring lethal mutation are both segregating in the population



We will see a deviation from HWE with fewer than expected homozygotes but probably some homozygotes for the haplotype on which the lethal occurs

Same Problem With TDT



A Better Strategy

- **Sequence genomes of bulls that have been widely used in AI**
  - In particular bulls that are carriers for haplotypes deviate strongly from HWE
  - About 3,000 bovines have been sequenced worldwide – we now have raw sequence data for about >700 animals in our database
- **Identify variants likely to be disruptive to gene function**
  - Particularly in genes known to be essential for life
- **Design these onto genotyping assays**
  - Genotype large sample to reduce candidate numbers
  - Migrate these to industry standard assays to generate >100K genotyped animals
- **Use mate selection to avoid double carrier matings**

109 Sequenced Angus Bulls

Chr	Location (Mb)	Length (Mb)	Haplotype Frequency[1]	Number of Patrios[2]	Probability[3]	Sequenced Carriers	Concordant Variants
1	27.7-29.0	1.3	0.023	39	0.0042	1	4
4	82.5-84.0	1.5	0.076	127	2.66E-09	21	9
8	62.0-63.0	1.0	0.023	35	0.0074	5	1
12	60.0-61.2	1.2	0.032	46	0.0014	12	0
15	82.3-83.1	0.8	0.038	31	0.011	10	1
17	46.5-47.5	1.0	0.045	49	0.00076	15	2
29	43.0-44.2	1.2	0.044	118	3.22E-08	16	3

Heterozygous genotypes must be accurately identified in each sequenced individual

Limitations Of Strategy

- Large number of animals must be sequenced to identify rare alleles
- Completely ignores lethal variants that may be in ncRNAs or that are regulatory
  - Flisikowski et al. 2010. A novel mutation in the maternally imprinted PEG3 domain results in a loss of MIMT1 expression and causes abortions and stillbirths in cattle (Bos taurus). PLoS One 5:e15116.
  - FAANG project has just begun
- May have to "test" a large number of candidate variants
  - Real estate issue for current commercial genotyping assays

Motivation For GGP-F250 Design

- USDA NIFA 2013-68004-20364
  - "Identification and management of alleles impairing heifer fertility while optimizing genetic gain in Angus cattle"
  - Loss of function variants in genes essential for life – 4K variants in 10K samples
- USDA NIFA 2011-68004-30214
  - "National program for genetic improvement of feed efficiency in beef cattle"
  - Fine-map Feed Efficiency QTLs – 1K variants in 2K samples
- USDA NIFA 2011-68004-30367
  - "Integrated program for reducing bovine respiratory disease complex in beef and dairy cattle"
  - Fine-map bovine Respiratory Disease susceptibility QTLs – 1K variants in 2K samples
- USDA NIFA 2015-67015-23183
  - "Application of a functional variant assay and sequence imputation to identify large-effect QTL underlying feed efficiency and component traits in beef cattle"
  - Build functional assay and fine-map QTL regions affecting Feed Efficiency – 1K variants in 3.5K samples

GGP-F250 Design: Design Data

Whole Genome Sequence for 262 taurines

Breed	No. Animals	No. Unique Reads	Total Bases	Av. Raw Coverage
Angus	109	82,263,951,806	8,137,666,488,753	25.74
Hereford	18	15,603,339,064	1,501,290,942,627	28.76
Limousin	12	3,704,169,818	357,264,463,240	10.27
Charolais	14	8,560,329,604	858,471,719,367	21.14
Simmental	11	8,902,705,282	885,698,817,042	27.76
Gelbvieh	8	6,366,906,096	633,479,558,830	27.31
Maine Anjou	5	4,061,220,172	403,867,224,031	27.85
Romagnola	4	901,544,762	89,666,842,589	7.73
Shorthorn	2	1,446,405,682	143,863,277,001	24.80
Red Angus	14	4,430,950,144	441,846,880,499	10.88
Holstein	55	13,650,662,246	1,358,163,462,700	8.52
Jersey	9	1,399,450,902	139,150,036,295	5.33
N'Dama	1	739,233,320	73,483,493,461	25.34

GGP-F250 Design: Validation Data

Whole Genome Sequence for 35 indicines/composites

Breed	No. Animals	No. Unique Reads	Total Bases	Av. Raw Coverage
Brahman	11	1,871,667,422	167,772,161,118	5.26
Nelore	8	1,668,006,036	165,728,918,125	7.14
Gir	6	1,583,737,248	157,449,065,756	9.05
Beefmaster	10	8,351,392,646	830,865,082,100	28.65

GGP-F250 Design: Validation Data

RNA-Seq data for 153 animals with 5 tissues/animal

Breed	No. Animals	No. Unique Reads	Total bases	Total Raw Coverage	Av. Raw Coverage
Angus	93	27,406,410,110	1,547,182,582,741	45,072.99	484.66
Hereford	11	3,815,784,459	302,094,954,473	8,800.72	800.07
Holstein	37	1,983,658,906	98,024,059,808	2,855.67	77.18
Crossbred	12	2,307,589,296	113,800,430,204	3,315.27	276.27
Total	153	35,513,442,771	2,061,102,027,226	15,011.16	409.55

GGP-F250 Design: Validation Data

- **Run 4 of the 1000 Bulls Genome Project**
  - 35,431,202 variants called in 1,147 animals
  - 1121 males, 24 females, 2 unknown
  - 24 breeds and composites
- **dbSNP Build 146**
  - 99,453,756 variants
- **UMD3.1 annotation**

GGP-F250 Design Content

Classification	Number of Validation Sources[1] 4	3	2	1	Chip	0	TOTAL
AA Genomic	6485	25455	24545	17732	60	965	75242
AA RNA	0	168	3918	12106	11	19258	35461
AA 1KBulls	0	76	2037	8449	18	102	10682
UTR Genomic	8019	8299	1684	645	0	0	18647
UTR RNA	0	59	792	378	5	0	1234
UTR 1KBulls	0	0	429	2931	8	0	3368
Promoter[2]	28	67	0	0	0	0	95
Splice Genomic	945	3643	4227	2509	0	0	11324
Splice RNA	0	5	205	4102	3	0	4315
Splice 1KBulls	0	0	835	2743	9	8	3595
snRNA Genomic	94	539	592	267	0	0	1492
snRNA RNA	0	3	20	16	0	0	39
Conserved Non-Coding Elements[3]	311	1771	1207	61	0	0	3350
Multi-allelic[4]	105	311	191	237	0	0	844
No Homozygotes Genomic	4	36	48	56	0	0	144
No Homozygotes RNA	0	14	42	8	0	0	64
No Homozygotes 1KBulls	0	0	13	53	1	0	67
BRD/FE QTL regions[5]	3734	12303	5370	2134	0	0	23541
Imputation	0	0	0	0	33729	0	33729
TOTAL	19725	52749	46155	54427	33844	20333	227233

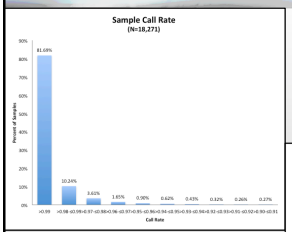
[1]Indicine WGS, RNA-Seq, 1K Bulls and dbSNP
[2]100 bp upstream of gene annotation start (5' UTR or start codon)
[3]Sequence conserved in 9 ruminant genomes. SNPs have PhastCons conservation scores ≥0.5 and CNE length ≥10bp
[4]Coding variants but one sequenced animal had a 3rd allele (possible false positive)
[5]Highest MAF variant within each 5 kb bin across QTL regions

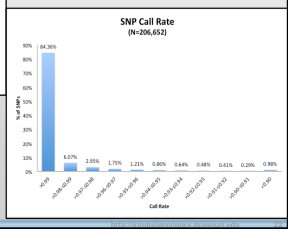
Samples Genotyped

BREED	NO. GENOTYPED	% GENOTYPED
AN	12083	66.13
HFD	945	5.17
LM	219	1.20
CHA	20	0.11
SIM	274	1.50
BRVH	7	0.04
GEL	307	1.68
PIED	9	0.05
RMG	8	0.04
ANR	1255	6.87
CIC	4	0.02
HO	1994	10.91
JER	9	0.05
GNS	7	0.04
NDAM	8	0.04
BR	14	0.08
NEL	8	0.04
GIR	11	0.06
CROS	1073	5.87
BEFM	3	0.02
SGT	11	0.06
SHK	2	0.01
18271/22 Breeds		

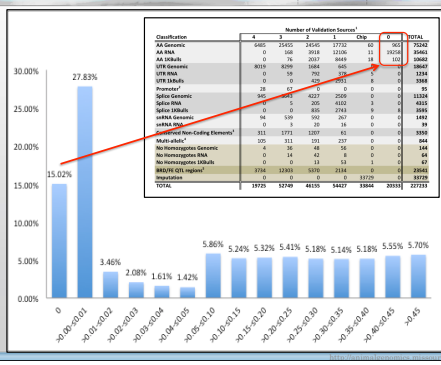
PROJECT	NO. GENOTYPED	% GENOTYPED
Heifer Fertility	11,506	62.97%
Feed Efficiency	4609	25.23%
Bovine Respiratory Disease	1971	10.79%
HapMap/History of Cows	185	1.01%
	18,271	

Sample and SNP Call Rates

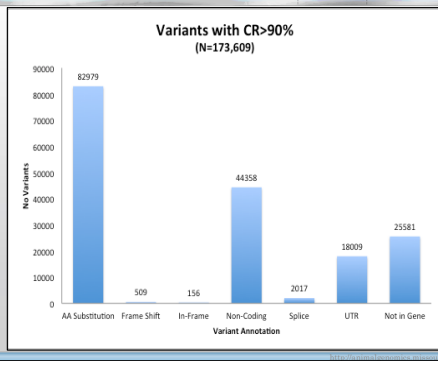


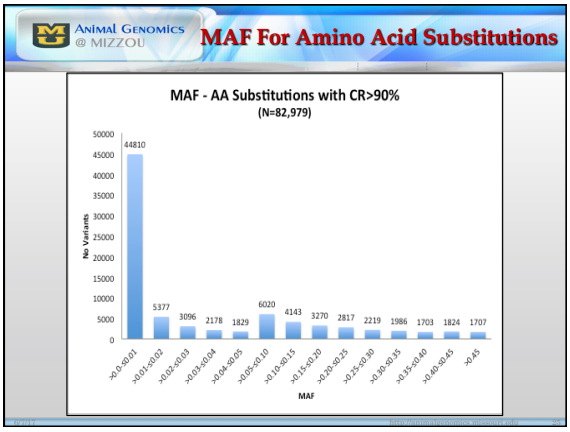


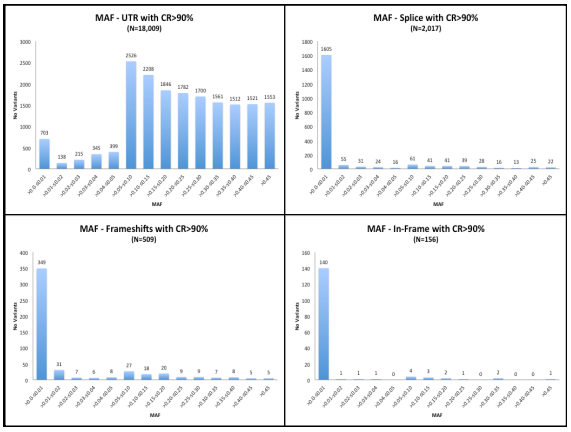
Minor Allele Frequency

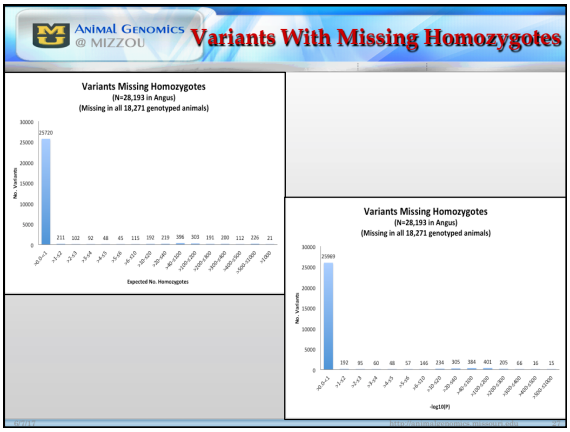


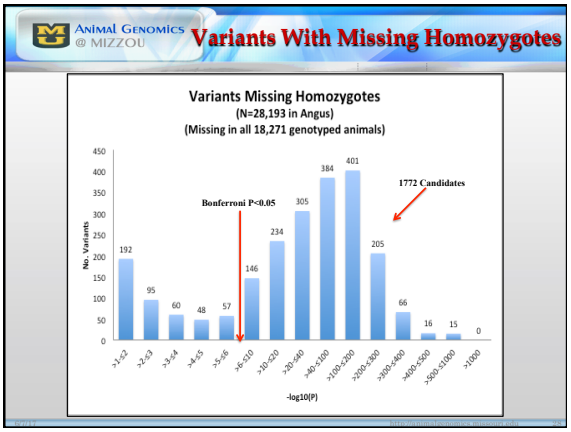
Variant Annotation

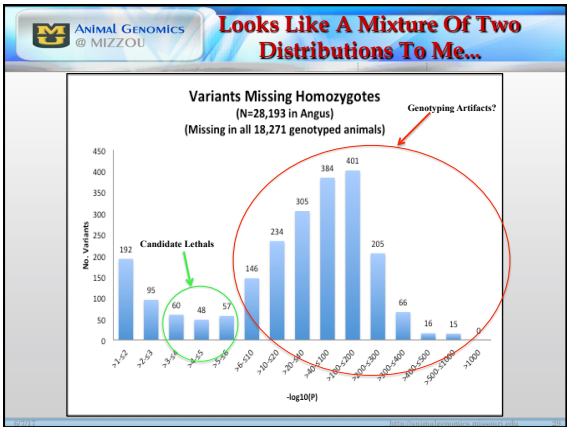


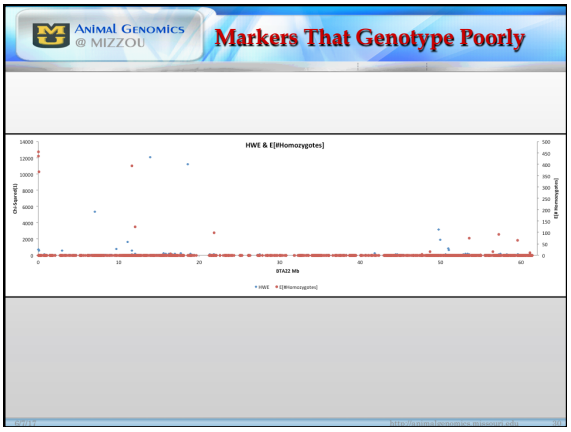

Animal Genomics @ MIZZOU
MAF For Amino Acid Substitutions
MAF - AA Substitutions with CR>90%
(N=82,979)
No. Variants
MAF
44810
5377
3096
2178
1829
6020
4143
3270
2817
2219
1986
1703
1824
1707
>0-0.01
>0.01-0.02
>0.02-0.03
>0.03-0.04
>0.04-0.05
>0.05-0.10
>0.10-0.15
>0.15-0.20
>0.20-0.25
>0.25-0.30
>0.30-0.35
>0.35-0.40
>0.40-0.45
>0.45


MAF - UTR with CR>90%
(N=18,009)
MAF - Splice with CR>90%
(N=2,017)
MAF - Frameshifts with CR>90%
(N=509)
MAF - In-Frame with CR>90%
(N=156)
No. Variants
MAF

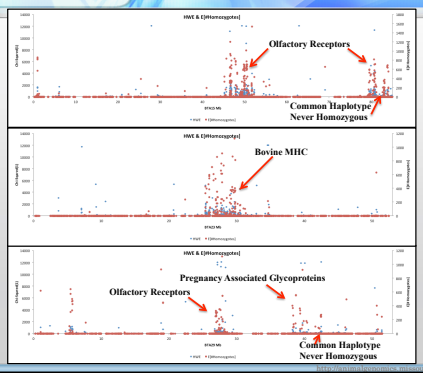

Animal Genomics @ MIZZOU
Variants With Missing Homozygotes
Variants Missing Homozygotes
(N=28,193 in Angus)
(Missing in all 18,271 genotyped animals)
Expected No. Homozygotes
-log10(P)


Animal Genomics @ MIZZOU
Variants With Missing Homozygotes
Variants Missing Homozygotes
(N=28,193 in Angus)
(Missing in all 18,271 genotyped animals)
Bonferroni P<0.05
1772 Candidates
No. Variants
-log10(P)


Animal Genomics @ MIZZOU
Looks Like A Mixture Of Two Distributions To Me...
Variants Missing Homozygotes
(N=28,193 in Angus)
(Missing in all 18,271 genotyped animals)
Genotyping Artifacts?
Candidate Lethals
No. Variants
-log10(P)


Animal Genomics @ MIZZOU
Markers That Genotype Poorly
HWE & E[#Homozygotes]

But There is More Going On...



Next Steps

- **Finish annotation of markers**
  - HGNC ID *versus* Synonyms *versus* Ensembl Gene IDs *versus* NCBI LOC symbols
- **Identify genomic regions with putatively lethal haplotypes**
  - VanRaden et al. (2011) J Dairy Sci 94:6153-61.
  - Use Breed Association chip data
  - Angus (N=200,000), Simmental (N=24,000), etc
- **Identify candidate variants in genes essential for life**
  - Perform patrio analysis for these variants
- **Incorporate candidate lethal variants on commercially utilized chips**
  - GeneSeek GGP products
  - Zoetis i50K, GeneMax® Advantage™
  - Irish Cattle Breeding Federation IBD chip

Conclusions

- **GGP-F250 was designed as a research tool to meet needs of Heifer Fertility, Respiratory Disease and Feed Efficiency Projects**
  - Gene centric but designed to allow genotype imputation into datasets genotyped with BovineSNP50, BovineHD, GGP-HD, GGP-LD, etc
  - Only ~50% of variants detected by sequencing are designable
  - Contains every designable AA substitution discovered in sequence data!
  - Useful research tool e.g. to explore basis of heterosis, inbreeding depression
  - Publicly available now through GeneSeek
- **Lethal Variants**
  - 2,224 candidates
  - Cannot all be lethal
  - Multiple approaches now required to filter data for genotyping artifacts and identification of true lethals (gene essential for life, lack of homozygous haplotypes in large industry datasets)
- **Delivery to Industry**
  - Requires selection indexes
  - Mate selection

Acknowledgements

- **Breed Associations co-sponsoring sequencing:**
  - American Angus Association
  - Australian Angus Association
  - Argentine Angus Association
  - American Hereford Association
  - Beefmaster Breeders United
  - American Gelbvieh Association
  - American International Charolais Association
  - American Simmental Association
  - American Maine-Anjou Association
- **10,000 heifers**
  - Missouri Show-Me-Select Replacement Heifer Program
  - Missouri Angus Association
  - Circle A Angus
- **USDA NIFA grants:**
  - 2011-68004-30214, 2011-68004-30367
  - 2013-68004-20364, 2015-67015-23183
- **GeneSeek for building the GGP-F250!**

USDA NIFA

United States Department of Agriculture
National Institute of Food and Agriculture

The End