Developing Accurate Parentage Markers for Cattle: From BSE Traceback to 50k SNP Chips

Gary Bennett

USDA, ARS, US Meat Animal Research Center

Clay Center, Nebraska

What are SNPs?

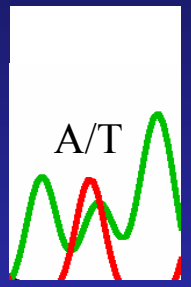
SNPs are sites in the genome where two different nucleotides are observed

DNA trace file

individual #1:

maternal chromosome ...aatggtatcAattaatgctt...

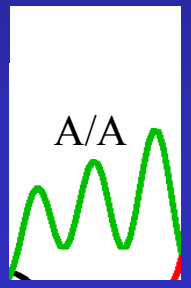
paternal chromosome ...aatggtatcTattaatgctt...



individual #2:

maternal chromosome ...aatggtatcAattaatgctt...

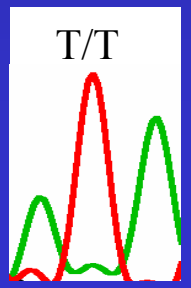
paternal chromosome ...aatggtatcAattaatgctt...



individual #3:

maternal chromosome ...aatggtatcTattaatgctt...

paternal chromosome ...aatggtatcTattaatgctt...



Why SNP?

- Abundant (approximately 30 million in cattle)
- Stable (low back-mutation rate)
- Amenable to high-throughput automatic scoring
- Low cost per SNP genotype
- Many genotyping platforms available
- Alleles easily and universally comparable

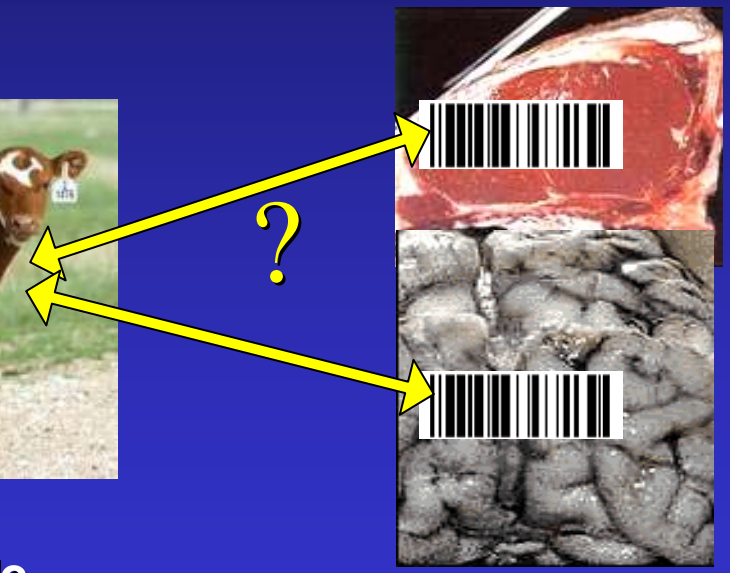
Why not SNP?

- Each microsatellite marker is more powerful (several alleles)
- Each SNP can exclude few parents (2 alleles)
- Several SNP needed to equal one microsatellite
- 30 million SNP not independent

Ways to use DNA for traceback

- **DNA fingerprinting (sample matching)**
  - comparing genotypes between samples
  - resolves disputes if samples were collected at the point of origin *before* a disease outbreak occurred.

- **Advantages:**
  - high degree of power
    - all genotypes used



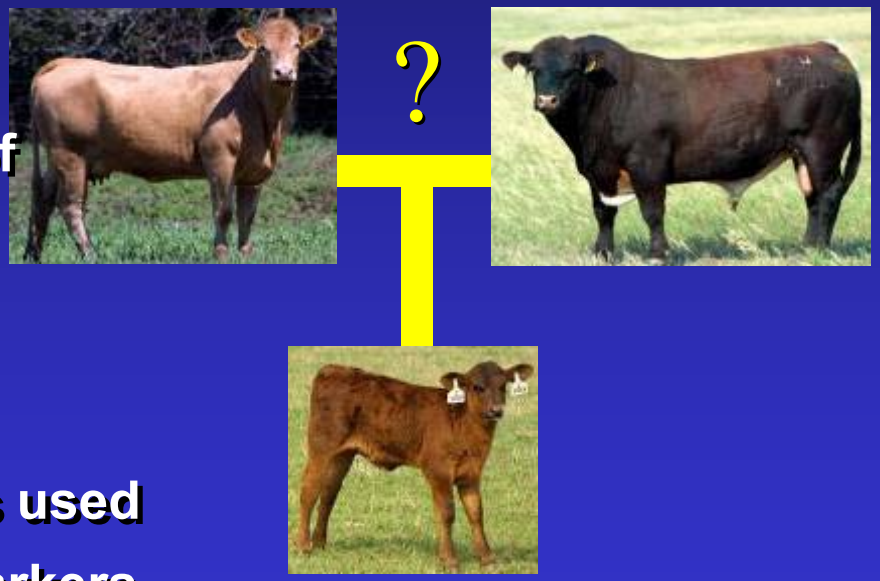
- **Disadvantages:**
  - requires a preexisting sample

Ways to use DNA for traceback

- **Parentage analysis**
  - determining whether alleles are shared between parents and offspring
  - may confirm the origin of a diseased animal if tissues from a parent are available.

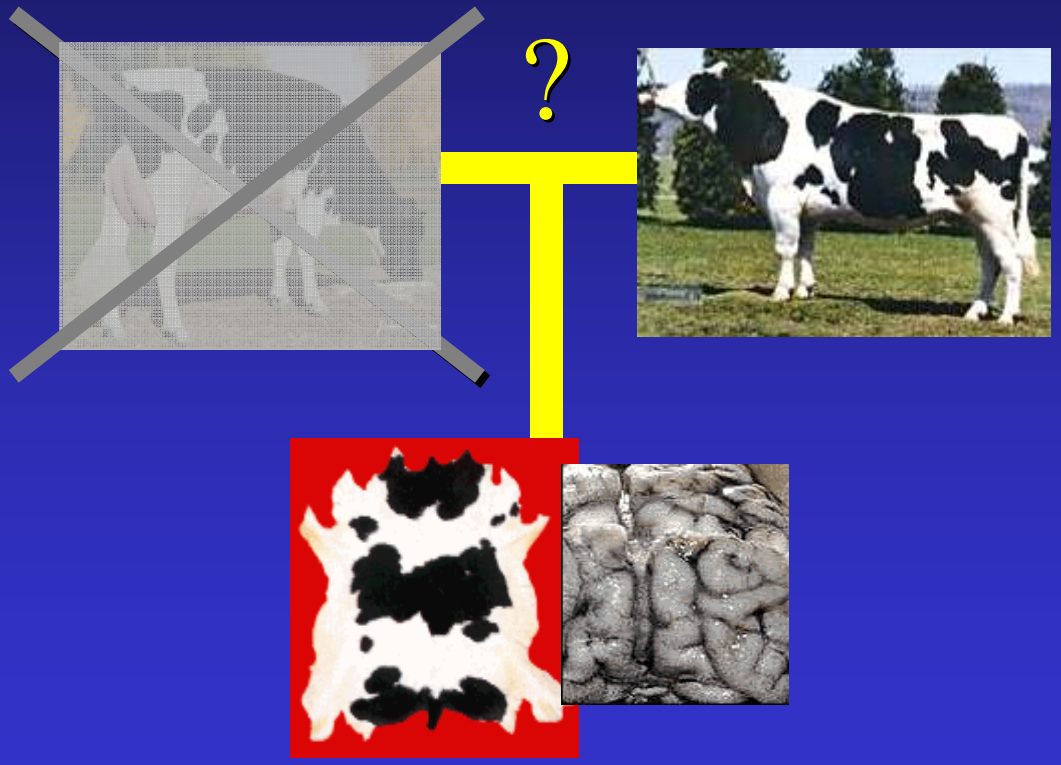
- **Advantages:**
  - preexisting sample of "case" not needed

- **Disadvantages:**
  - not all genotypes used
  - requires more markers
  - requires more samples



Sometimes parentage testing is the last resort for DNA-based traceback

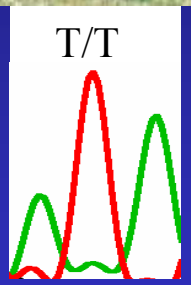
- Worst case scenario: only one parent available
  - Washington State BSE case



SNP markers for parentage

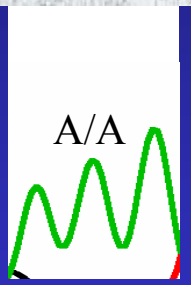
…aatggtatcaTattaatgctt…

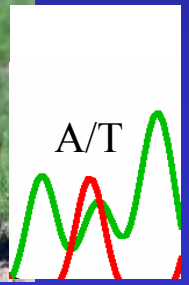
…aatggtatcaTattaatgctt…



…aatggtatcaAattaatgctt…

…aatggtatcaAattaatgctt…





The offspring must share an allele with each parent

…aatggtatcAattaatgctt…

…aatggtatcTattaatgctt…

SNP Exclusion – Sire only

Sire	Progeny			Frequency A		
	AA	AT	TT	0.5	0.3/0.7	0.1/0.9
AA			Exclude	.06	.04	.01
AT						
TT	Exclude			.06	.04	.01
Total				.12	.09	.02

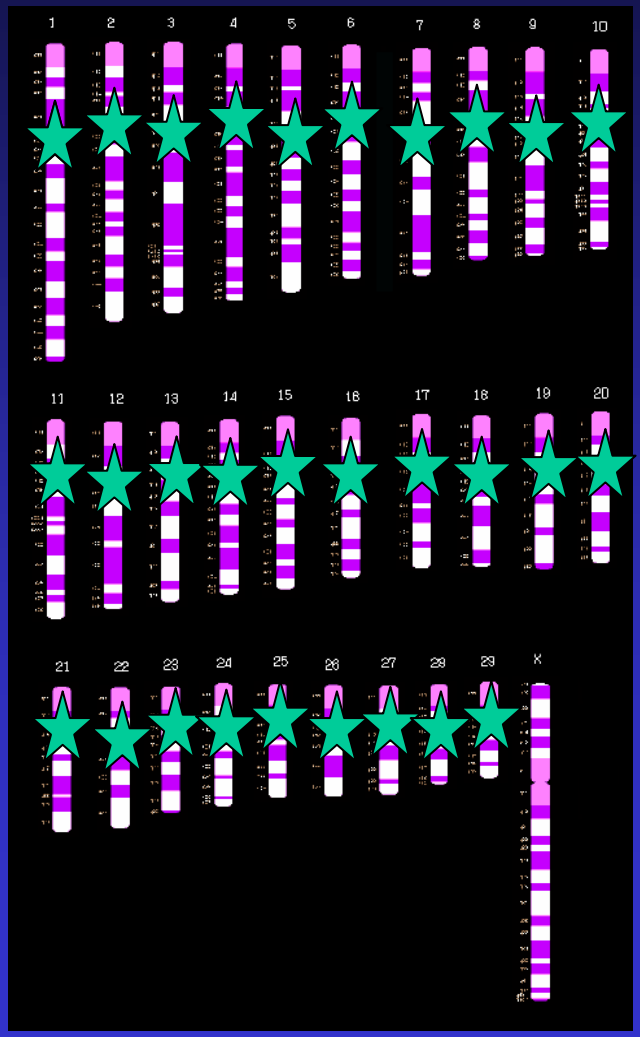
SNP Exclusion – Sire & Dam

Sire	Dam	Progeny			Frequency A		
		AA	AT	TT	0.5	0.3 /0.7	0.1/0.9
AA	AA		X	X	.05	.01	.00
	AT			X	.03	.02	.00
	TT	X		X	.03	.03	.01
AT	AA			X	.03	.02	.00
	AT				0	0	0
	TT	X			.03	.02	.00
TT	AA	X		X	.03	.03	.01
	AT	X			.03	.02	.00
	TT	X	X		.05	.12	.12
Total					.28	.26	.14

Microsatellite Exclusion

Sire	Progeny										Total
	100/100	100/102	100/106	100/108	102/102	102/106	102/108	106/106	106/108	108/108	
100/100					X	X	X	X	X	X	.035
100/102								X	X	X	.031
100/106					X		X			X	.031
100/108					X	X		X			.031
102/102	X		X	X				X	X	X	.035
102/106	X			X						X	.031
102/108	X		X					X			.031
106/106	X	X		X	X		X			X	.035
106/108	X	X			X						.031
108/108	X	X	X		X	X		X			.035
Total											.33

The ideal markers are independently inherited



Problem: there are only 29 autosomes

The ideal marker is frequent in all breeds

A collaborative effort was undertaken to assemble many beef and dairy breeds for testing (screening) allele frequency

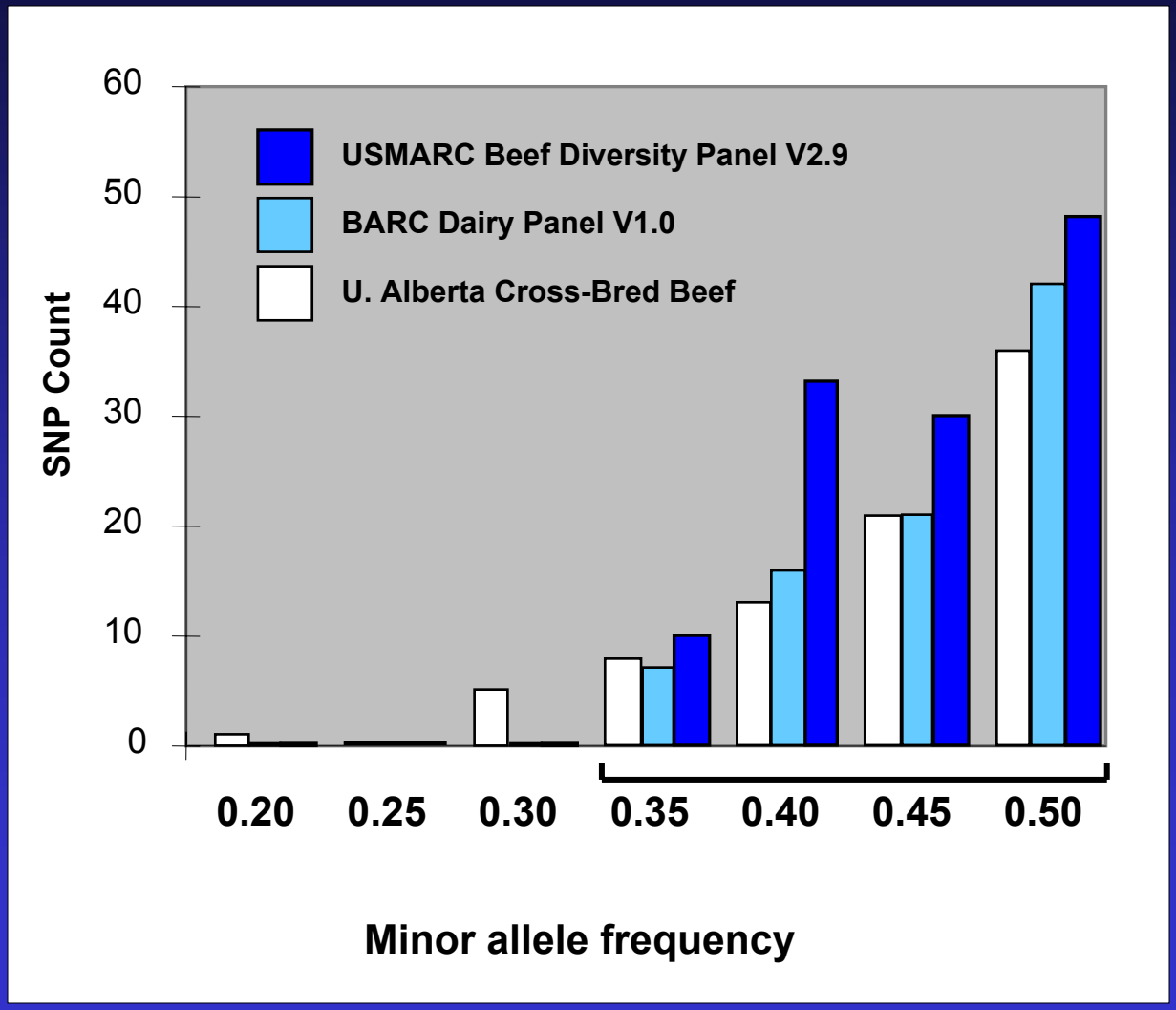
96 diverse sires from 19 beef breeds (Drs. Heaton and Laegreid; ARS, USMARC)

464 cross-bred Canadian beef cattle containing germplasm primarily from Angus, Charolais, Hereford, Simmental, Galloway, and other breeds (Dr. Moore, University of Alberta)

120 prominent sires from 4 dairy breeds (Drs. Van Tassell and Sonstegard; ARS, BARC)

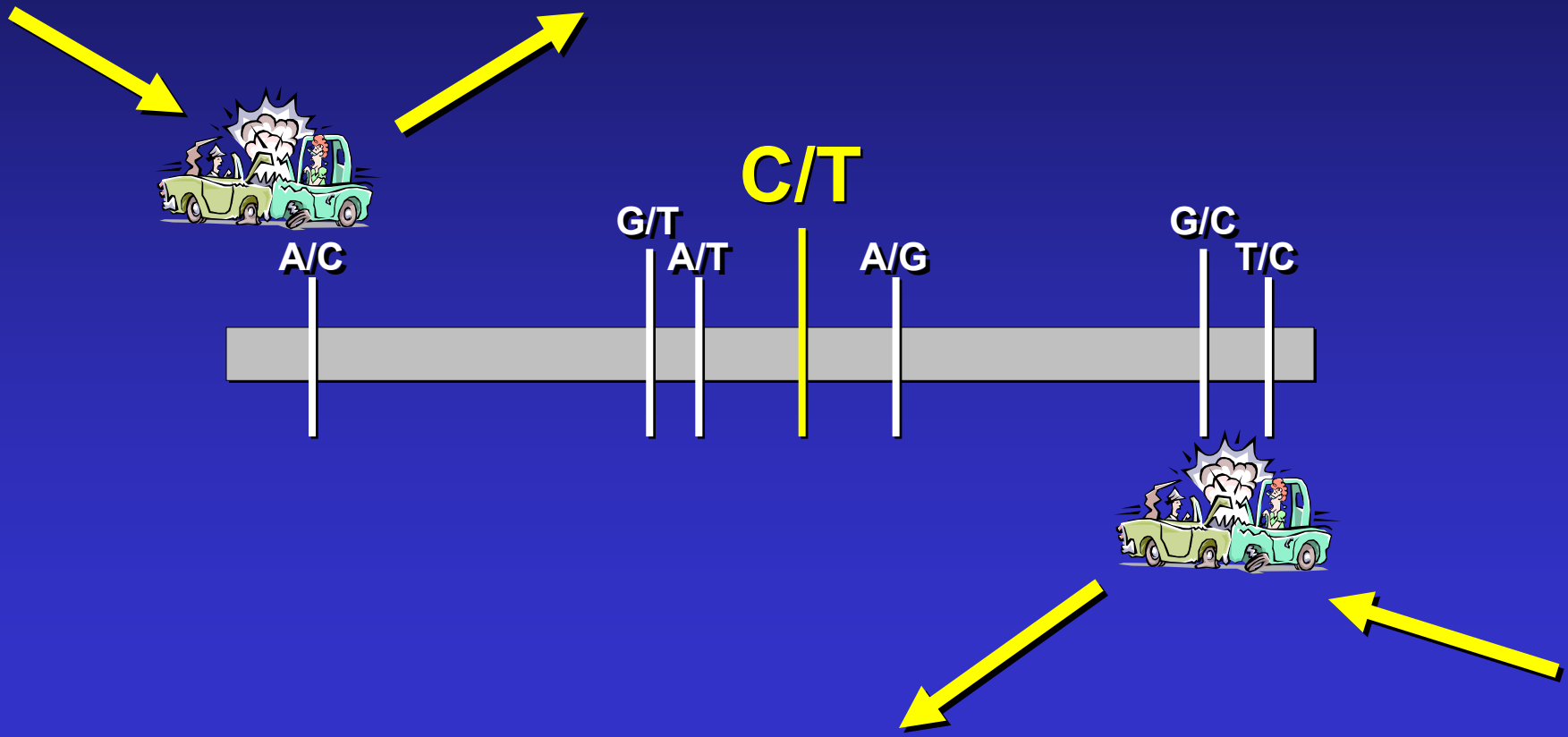
More than 4000 candidate SNPs, mostly from the Bovine Genome Project, were genotyped to select those with best minor allele frequencies (Drs. Heaton, McKay, Moore, and Murdock; MARC and U. Alberta)

Distribution of minor allele frequencies for 122 parentage SNPs in US and Canadian cattle

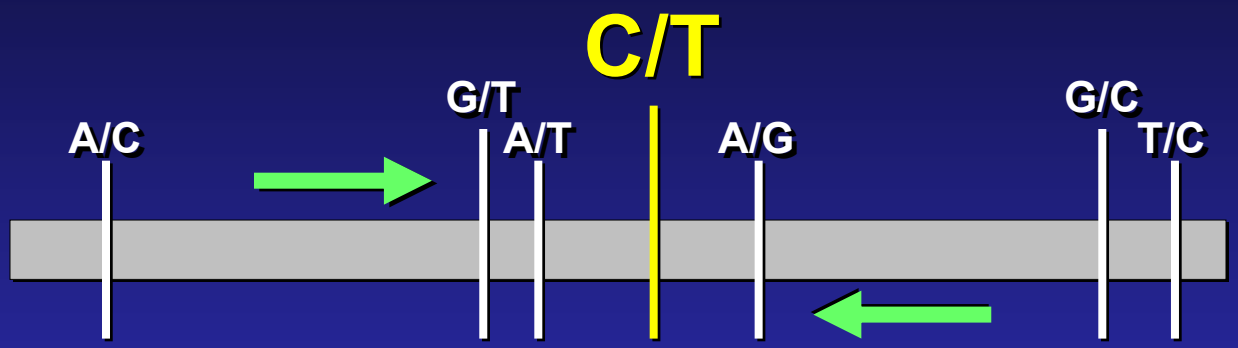


The consequence of 1 SNP every 80 bp

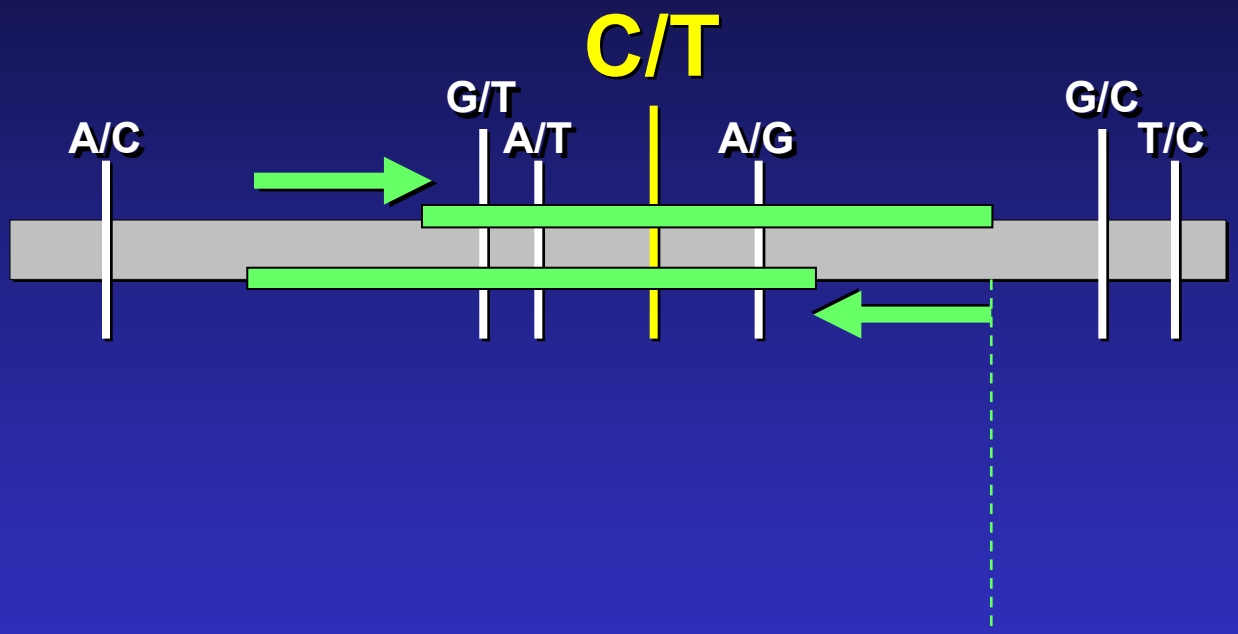
- Wrong genotype assigned to some animals



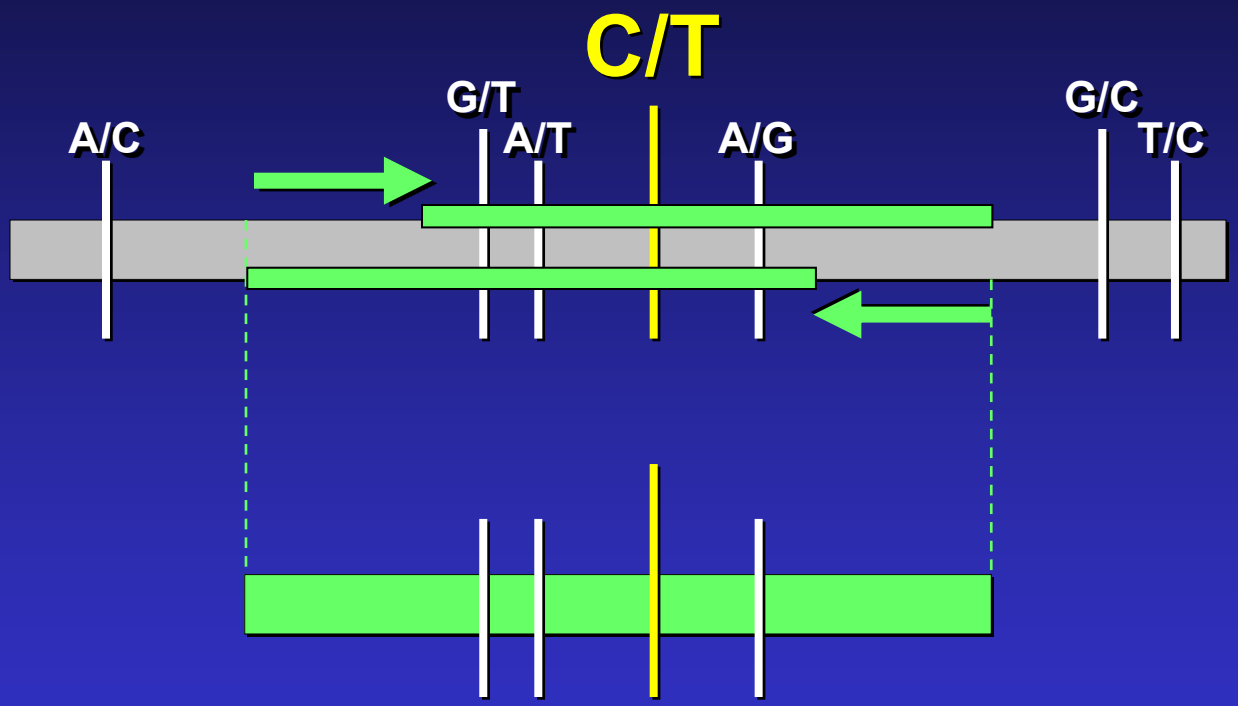
The consequence of 1 SNP every 80 bp



The consequence of 1 SNP every 80 bp

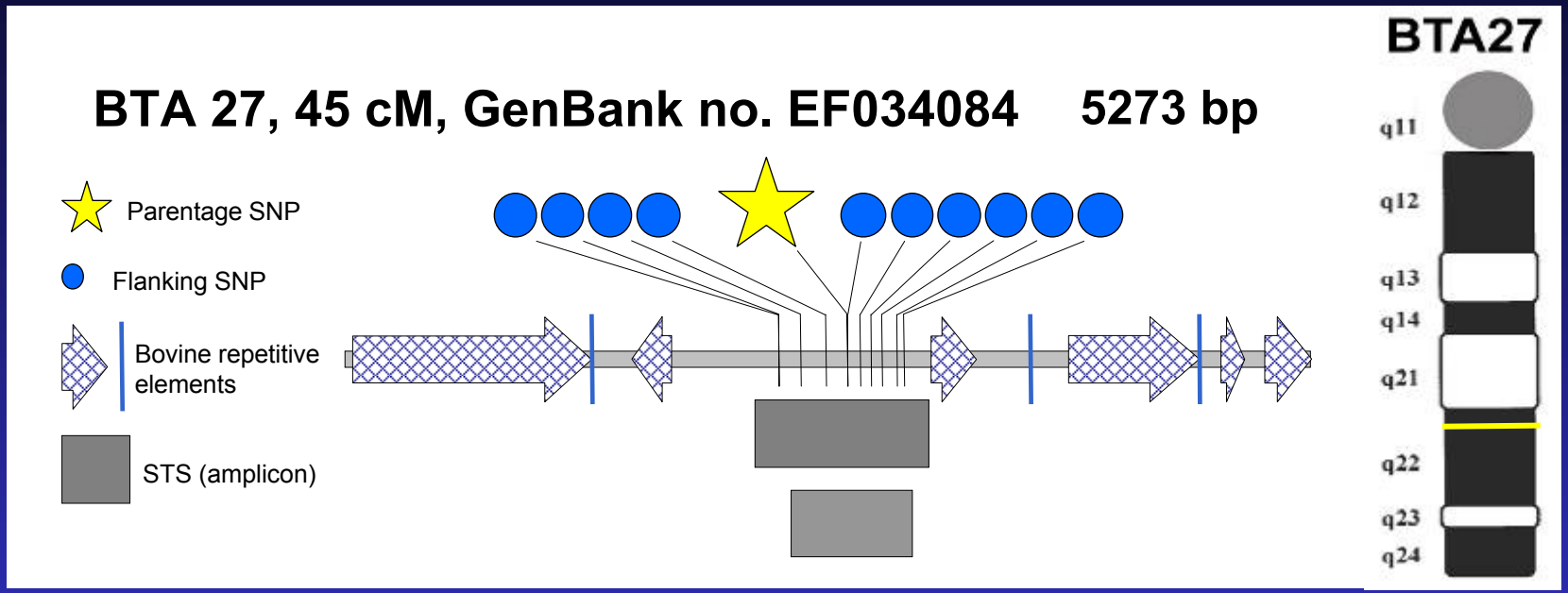


The consequence of 1 SNP every 80 bp



accurate amplification of both maternal and paternal alleles

Physical map



Mean annotation 9.8 kb per file (1.2 Mb total)

1646 adjacent SNPs
259 amplicons
2004 bovine repetitive elements
258 exons
183 CDSs

- **Dr. Anjanette Johnston**
  GenBank Submissions Staff

- **Dr. Ilene Karsch Mizrachi**
  GenBank Coordinator

Public access to SNP information

Ted Kalbfleisch
U. Louisville

http://cgemm.louisville.edu/usmarc/MARC_web_page/traceback.html

CGeMM

University of Louisville

A SNP Marker Set for DNA-based Traceback in North American Beef and Dairy Cattle.

SNP Summary Table: SNPs that have been thoroughly screened to accomplish both DNA fingerprinting and parentage testing in U.S. beef and dairy populations, an ability that only a small fraction of known DNA markers have. These are markers are specially selected for optimum power, genome-wide distribution, accuracy in genotyping, and high-throughput "multiplexibility". The target SNPs and surrounding DNA was sequenced in a group of 216 diverse sires from 19 beef breeds and 4 dairy breeds representing the vast majority of U.S. cattle. The average minor allele frequencies in beef and dairy are greater than 0.410. The DNA diversity in the adjacent regions has been documented for use in designing DNA tests that will be accurate in more than 99.9% of the North American beef and dairy cattle to be tested. Available information includes, flanking context sequence, allele frequency summaries, cross references to other databases, and genomic mapping.

Available Genotype Data: A table of consensus genotypes for the target SNPs scored in the USMARC Beef Cattle Diversity Panels V2.1 and V2.9, and the BARC Dairy Cattle Diversity Panel V1.0

Cattle Population Summaries: A table of animal identifiers, as well as their corresponding breed information.

Research results and information provided by Dr. Michael Heaton of the U.S. Meat Animal Research Center(USMARC), Clay Center, NE.

USDA, ARS Project Number: 5438-32000-023-02 Specific Cooperative Agreement Accession Number: 409004. For more information about the project please see "Project Info".

This project is in collaboration with scientists at the USDA ARS, Bovine Functional Genomics Laboratory in Beltsville, MD, the University of Alberta's Beef Genomics Laboratory in Edmonton Canada, the Center for Genetics and Molecular Medicine at the University of Louisville, KY, and Cogenics in Morrisville, NC.

Center for Genetics and Molecular Medicine (CGeMM)
Director: Kenneth S. Ramos Ph.D.
Director of Bioinformatics Operations: Ted Kalbfleisch Ph.D.
Please direct questions or comments to Ted Kalbfleisch

USDA ARS

Public access to SNP information

Ted Kalbfleisch
U. Louisville

http://cgemm.louisville.edu/usmarc/MARC_web_page/traceback.html

CGeMM

A SNP Marker Set for DNA-based Traceback in North American Beef and Dairy Cattle.

University of Louisville

SNP Summary Table: SNPs both DNA fingerprinting and an ability that only a small fr are specially selected for opt genotyping, and high-throu surrounding DNA was seque breeds and 4 dairy breeds re minor allele frequencies in b in the adjacent regions has b will be accurate in more than be tested. Available informat frequency summaries, cross

Available Genotype Data: scored in the USMARC Beef Dairy Cattle Diversity Panel V1.2

Cattle Population Summaries: A table of animal identifiers, as well as their corresponding breed information.

Research results and information provided by Dr. Michael Heaton of the U.S. Meat Animal Research Center(USMARC), Clay Center, NE.

USDA, ARS Project Number: 5438-32000-023-02 Specific Cooperative Agreement Accession Number: 409004. For more information about the project please see "Project Info".

This project is in collaboration with scientists at the USDA ARS, Bovine Functional Genomics Laboratory in Beltsville, MD, the University of Alberta's Beef Genomics Laboratory in Edmonton Canada, the Center for Genetics and Molecular Medicine at the University of Louisville, KY, and Cogenics in Morrisville, NC.

Center for Genetics and Molecular Medicine (CGeMM)
Director: Kenneth S. Ramos Ph.D.
Director of Bioinformatics Operations: Ted Kalbfleisch Ph.D.
Please direct questions or comments to Ted Kalbfleisch

USDA ARS

Download File

GenBank accession no.	Location of target SNP	dbSNP Accession	Orientation (g)	Linkage Group	Linkage map position (cM)	Chr. No.	Chr. Position(s)	Genomic region identifier	Approximate (SNP skew)	Minor Allele No. Traces at NCBI	SNP at STSs	STS coverage (bp)	USDA/ARC Beef	USDA/ARC Dairy	USMARC Dairy	USMARC Beef	Universal Target SNP and Flanking Segment
DQ383123	1393	rs29012842	rev	BTA_1	7		chr_1	(g)-	SCAFFOLD135069_21121-		856	21	1583	0.49	0.448	nd (g)	0.408	
DQ451555	2281	rs29010795	fwd	BTA_1	34		chr_1	-	SCAFFOLD155476_15644-		1191	14	1035	0.482	0.416	nd	0.422	
AY773474	1589	(g)-		n/a	BTA_1	51	chr_1	-	DVEPC18	-	1192	11	925	0.408	nd	0.355	nd	
DQ404150	2655	rs29012580	rev	BTA_1	70		chr_1	-	SCAFFOLD130061_59457-		1071	4	1053	0.448	0.453	nd	0.4865	
DQ404149	2834	-		n/a	BTA_1	115	chr_1	-	BES6_Contig246_498	-	996	9	1068	0.4375	0.417	nd	0.396	
AY781135	1014	rs29003725	rev	BTA_1	138		chr_1	-	59937	-	1262	10	798	0.364	nd	0.408	nd	
DQ404151	991	rs29019282	fwd	BTA_1	154		chr_1	100281322	SCAFFOLD191372_882	-	1026	13	1030	0.49	0.469	nd	0.3793	

Public access to SNP information

Ted Kalbfleisch
U. Louisville

http://cgemm.louisville.edu/usmarc/MARC_web_page/traceback.html

CGeMM

University of Louisville

A SNP Marker Set for DNA-based Traceback in North American Beef and Dairy Cattle.

SNP Summary Table: SNPs that have been thoroughly screened to acce
both DNA fingerprinting and parentage testing in U.S. beef and dairy popu
an ability that only a small fraction of known DNA markers have. These are
are specially selected for optimum power, genome-wide distribution, accu
genotyping, and high-throughput "multiplexibility". The target SNPs and
surrounding DNA was sequenced in a group of 216 diverse sires from 19
breeds and 4 dairy breeds representing the vast majority of U.S. cattle. The
minor allele frequencies in beef and dairy are greater than 0.410. The DNA
in the adjacent regions has been documented for use in designing DNA te
will be accurate in more than 99.9% of the North American beef and dairy
be tested. Available information includes, flanking context sequence, allele
frequency summaries, cross references to other databases, and genomic n

Available Genotype Data: A table of consensus genotypes for the target
scored in the USMARC Beef Cattle Diversity Panels V2.1 and V2.9, and the
Dairy Cattle Diversity Panel V1.0

Cattle Population Summaries: A table of animal identifiers, as well as t
corresponding breed information.

Research results and information provided by Dr. Michael Heaton of the U.
Animal Research Center(USMARC), Clay Center, NE.

USDA, ARS Project Number: 5438-32000-023-02 Specific Cooperative Agreement Accession Number: 409004. For more information about the project please see "Project Info".

This project is in collaboration with scientists at the USDA ARS, Bovine Functional Genomics Laboratory in Beltsville, MD, the University of Alberta's Beef Genomics Laboratory in Edmonton Canada, the Center for Genetics and Molecular Medicine at the University of Louisville, KY, and Cogenics in Morrisville, NC.

Center for Genetics and Molecular Medicine (CGeMM)
Director: Kenneth S. Ramos Ph.D.
Director of Bioinformatics Operations: Ted Kalbfleisch Ph.D.
Please direct questions or comments to Ted Kalbfleisch

USDA ARS

Download File

Genbank Accession		AY773474	AY776154	AY841151	AY842472	AY842473	AY842474	AY842475
SNP Position		1589	1982	6724	1027	1511	1201	1726
dbSNP ID		(a)-	-	rs29003466	rs29001941	rs29001956	rs29003226	rs29002127
Animal Number	Breed							
176155	Holstein	A/T	G	G	C/G	A/G	C/G	G
182867	-	T	G	T	G	G	G	A
20300442	Holstein	A	A	T	G	G	G	A/G
20300638	Holstein	A	A/G	G/T	C/G	A/G	C/G	A
20300902	Holstein	T	G	G/T	C/G	A/G	C/G	A/G
20300914	Red & White Holstein	T	A/G	T	C/G	A/G	C/G	A/G
20300918	Holstein	T	A/G	G	C/G	G	C/G	A
20300933	Jersey	T	A/G	T	C	A	C	G
20301363	Holstein	A/T	A/G	G/T	C	A/G	G	G
2031217	Red & White Holstein	T	G	G	G	A	C	A/G
2031243	Brown Swiss	T	A	G	C/G	A/G	C/G	A
2031449	Red & White Holstein	A/T	A/G	G/T	C/G	A/G	G	A/G

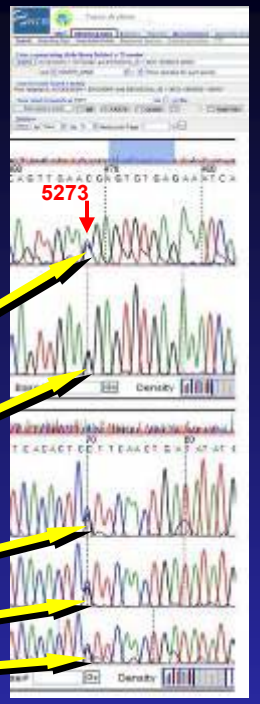
Linking genotypes to tracefiles

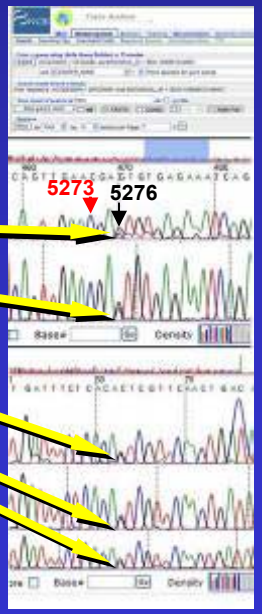
- >350,000 genotypes
- >136,000 publicly viewable tracefiles
- >2,000,000 tracefile genotypes

Table 3. Selected animal genotypes for 11 SNPs in EF034084

		Nucleotide position on GenBank accession number EF078676										
Breed	Animal no.	4910	4914	5025	4697	5273[a]	5276	5341	5378	5456	5558	5574
Angus	19879801	T/T	C/C	G/G	C/C	C/G	G/G					G/G
Angus	19939802	T/T	C/C	G/G	C/C	C/C						
Angus	19999821	T/T	C/C	G/G	C/C	C/G		T/T	C/C	C/C	G/G	G/G
Angus	19999822	T/T	C/C	G/G	C/C	G/G	G/G	T/T	C/C	C/C	G/G	G/G
Angus	19999823	T/T	C/C	G/G	C/C	G/G	G/G	T/T	C/C	C/C	G/G	G/G
Angus	19999825	T/T	C/C	G/G	C/C	G/G	G/G	T/T	C/C	C/C	G/G	G/G
Brahman	19919842	T/T	C/C	G/G	C/C	C/G	G/G	T/T	C/C	C/C	G/G	G/G
Brahman	19999811	T/T	C/T	A/G	C/C	C/G	G/G	T/T	C/C	C/C	G/G	G/G
Brahman	19999812	T/T	T/T	A/A	C/C	C/C	G/G	T/T	C/C	C/C	G/G	G/G
Brahman	19999813	T/T	C/T	A/G	C/C	C/C	C/G				G/G	G/G
Brahman	19999819	T/T	C/C	G/G	C/C	G/G	G/G	T/			G/G	G/G

[a]SNP selected for use in traceback by parentage analysis.





Genetically sound, but …

- Sample integrity
  - Right sample, right label, no contamination
- SNP genotyping technology
  - Accuracy of genotypes, high call rates, error free sample handling, no contamination
- Selection of SNP
  - Evenly spaced, intermediate frequencies, present in many different breeds and populations
- Delivery of information
  - Convenient, when needed, correct
- Easy, Quick, Accurate, Efficient, Economical

The DNA-Based Traceback of the Washington State BSE Case

Announced by USDA on December 23, 2003

First recorded BSE case in the U.S. history

USMARC was asked to help. We designed DNA experiments, decoded the results, and wrote the report.

The dispute

"There's some confusion about the paperwork…."

Which of the 9 downer cattle slaughtered that day had the BSE-infected brain?

"DNA testing by the best experts available could compare samples from the mad cow and its offspring or parents."

Dr. Brian Evans
Chief Veterinary Officer
Canadian Food Inspection Agency

http://www.cbc.ca/stories/2003/12/27/Evans271203

http://www.guardian.co.uk/bse/article/0,2763,1113783,00.html

USDA briefing – December 31

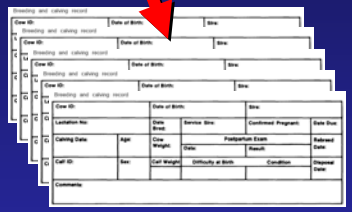
USDA's Chief Veterinarian
Dr. Ron DeHaven of APHIS

"…we are sending multiple samples to two laboratories -- one in Canada and one in the United States."

"… the U.S. laboratory is in Nebraska, …. [and] It's a USDA laboratory that has that expertise."

The situation

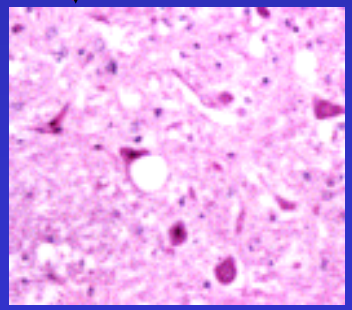
APHIS had physical evidence indicating that the index cow was born in Canada



We asked APHIS to gather pedigree records and tissue samples from all available relatives

Holstein downer cow processed at Moses Lake, WA on December 9th

Its brain tested positive for BSE on December 22nd

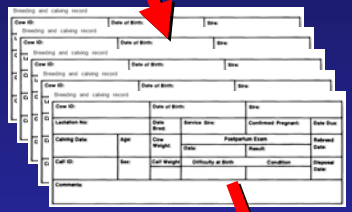


The key question:

Was that ear tag really attached to the animal with the BSE brain?

The situation

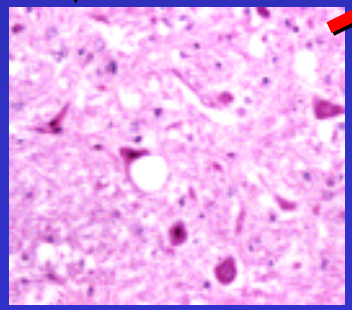
APHIS had physical evidence indicating that the index cow was born in Canada



We asked APHIS to gather pedigree records and tissue samples from all available relatives

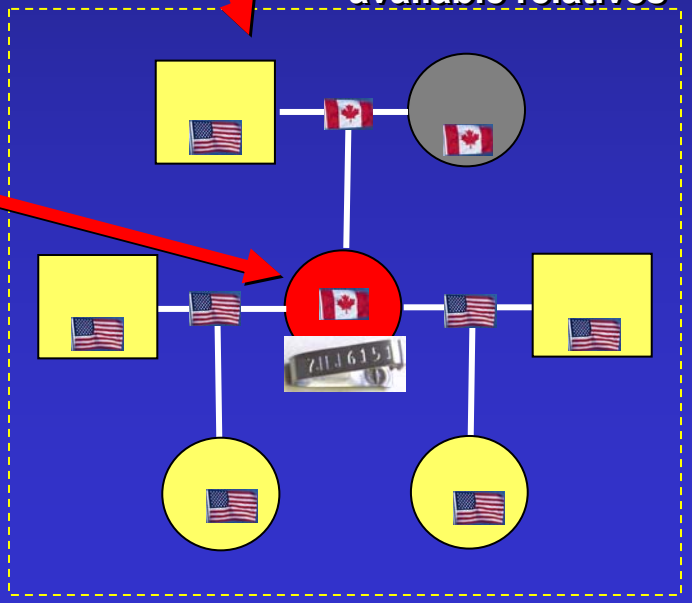
Holstein downer cow processed at Moses Lake, WA on December 9th

Its brain tested positive for BSE on December 22nd



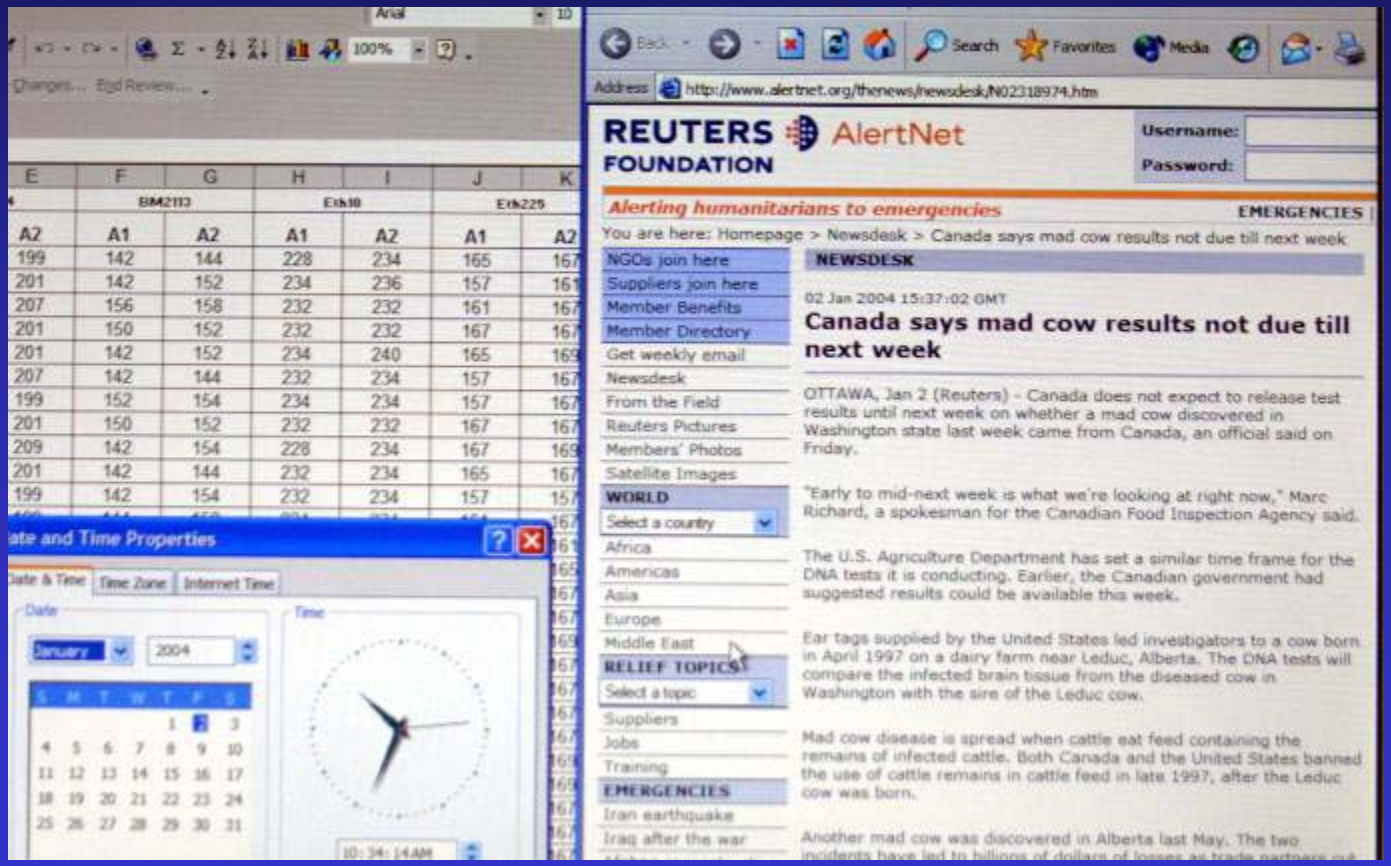
DNA from BSE brain

sent from Ames, IA to test validity of *this* pedigree



January 2, 2004

Test results obtained and decoded within 40 hours
(more that 13,000 genotypes from 66 samples)



January 6

"We now have DNA evidence that allows us to verify with a high degree of certainty, the [Canadian] birthplace of the BSE-infected cow."

Dr. Ron DeHaven

Canadian officials concurred

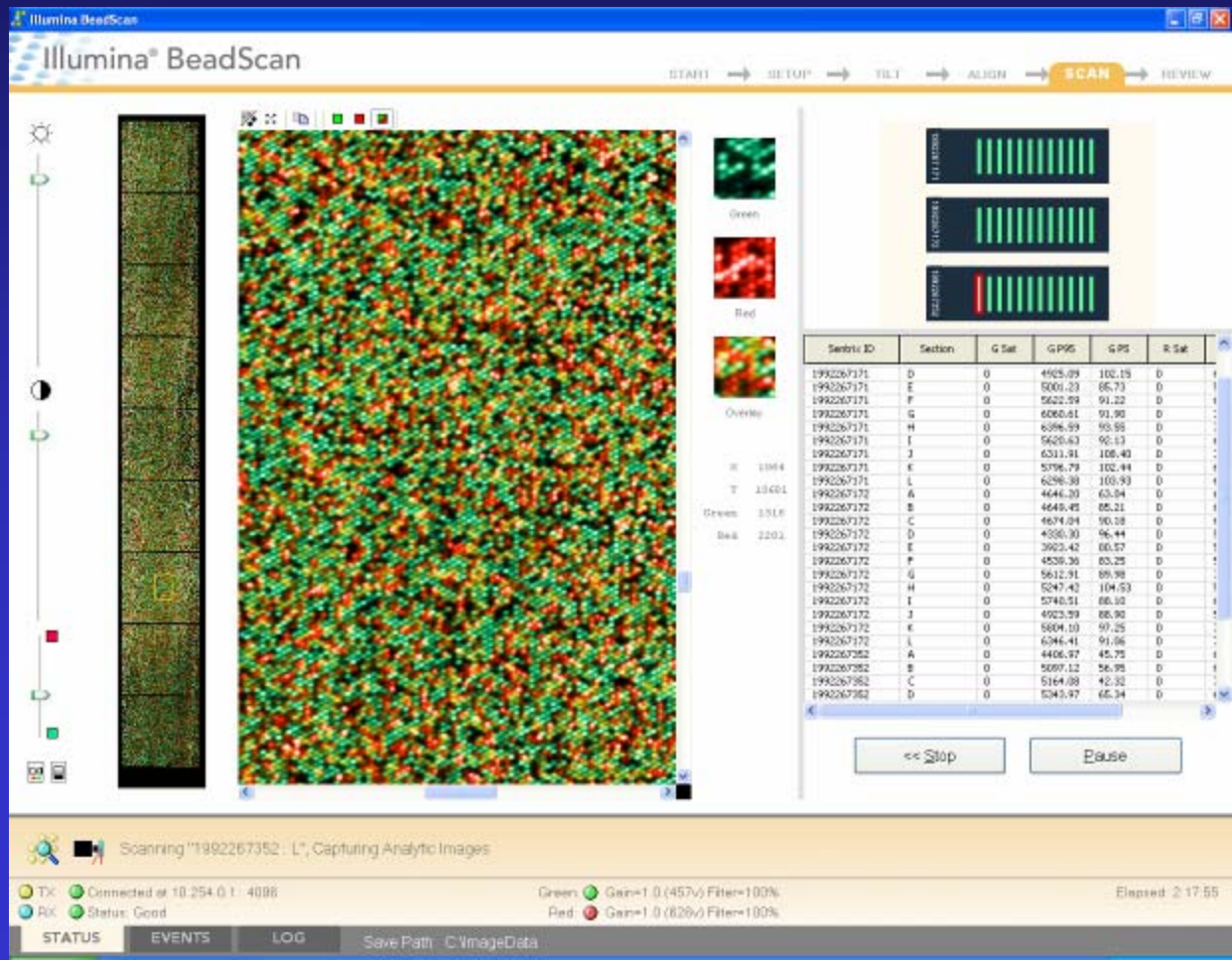
Dr. Brian Evans

http://www.usda.gov/Newsroom/0003.04.html

Cattle SNP-chip Collaboration

- USDA-ARS U.S. Meat Animal Research Center
- USDA-ARS Beltsville Agricultural Research Center (BARC)
- University of Missouri
- University of Alberta
- Illumina / Solexa

Illumina iSelect™ assay (60,800 bead types) -- hope for about 53,000 useful SNP markers



Requirements for Genome Wide Association or Selection

- Markers within each block of Linkage Disequilibrium (LD) that track the main functional haplotypes
- LD blocks of about 100 kilobases in cattle genome achieve r^2 around 0.3
- 3 Gb / 0.1 Mb = 30,000 LD blocks

Goal for the bovine SNP array

- Achieve > 30,000 SNP
- Spread evenly across genome
- Highly informative across cattle breeds and populations

March 2008 | volume 5 | number 3

nature methods

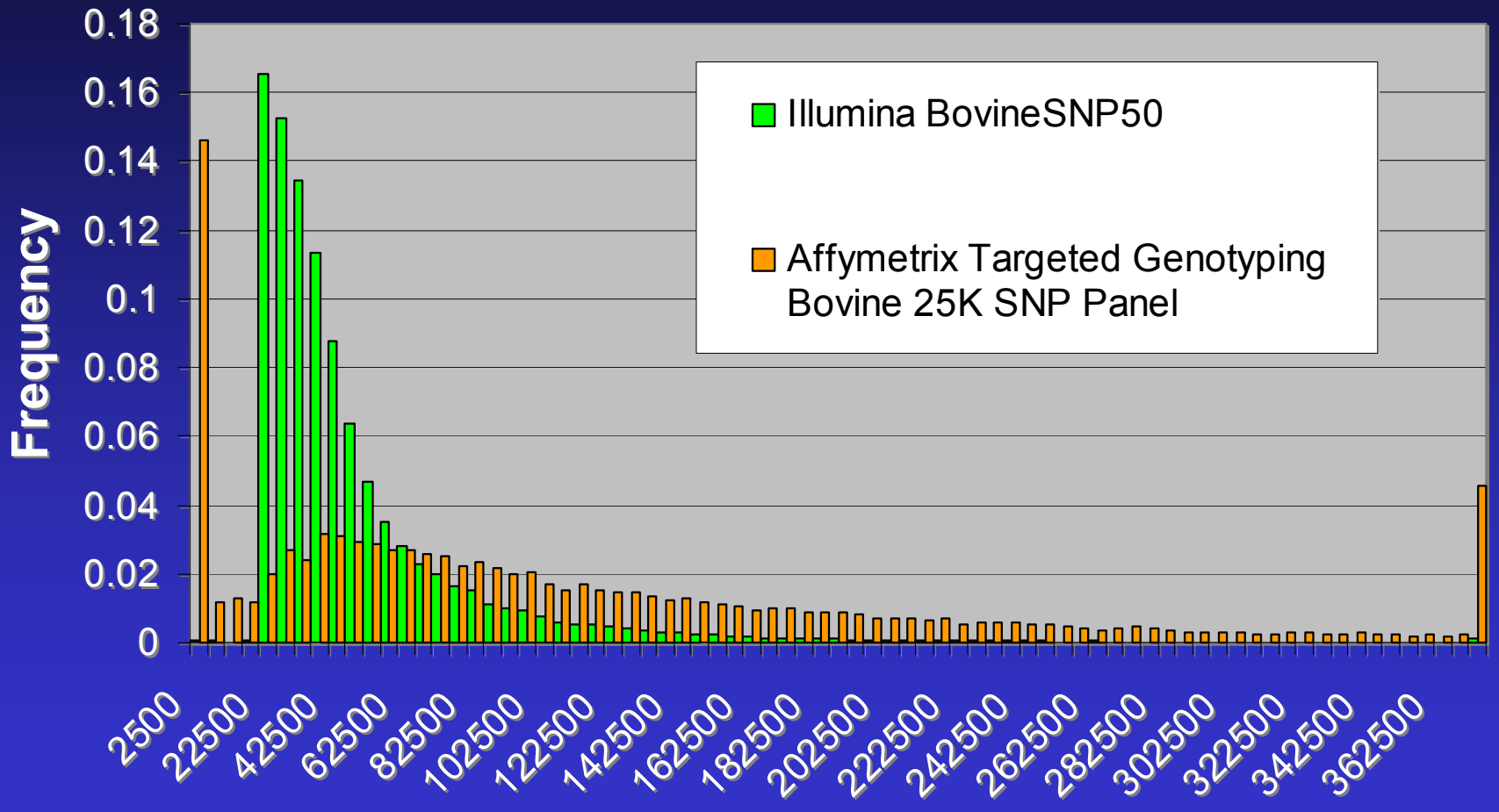
www.nature.com/naturemethods

Techniques for life scientists and chemists

SNP Content on the Chip

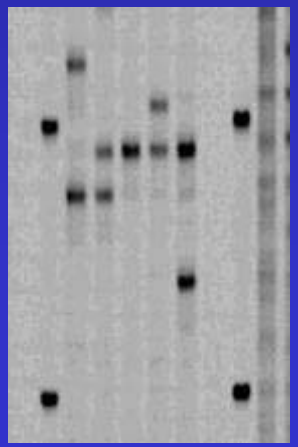
SNPs With MAF (68%)	
Reduced Representation Library	25,125
Bovine Hapmap Consortium	12,641
UA-IFASA	934
Others(US-MARC,DPI)	55
Insilico SNPs (32%)	
Assembly SNPs	10,075
Inter Breed	6,200
BACend Derived	1,484
INRA, DIAS	310
Mandatory Inclusions (0.2%)	
Parentage Markers	118
Selected Genes (BPI,CAPN)	5
Total	**56,947**

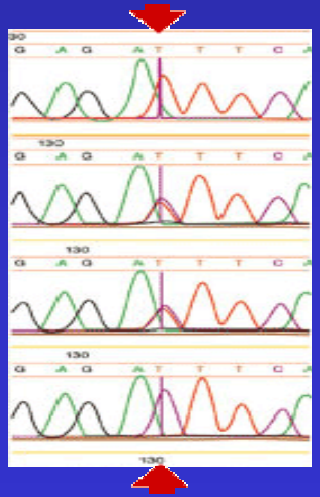
Gap Distribution of SNPs along chromosomes



Definitions

- Genome **is the whole set of DNA**
- Genetic Markers **track regions of the genome**
  - **May be linked/explain phenotypes we observe**
- **A** SNP **is a type of genetic marker**
  - **S**ingle **N**ucleotide **P**olymorphism
- **A** Genotype **is the form of DNA present at a specific location in the genome.**





Whole Genome Selection (WGS)

- Use 1,000's of SNP to predict EPDs
- Like current marker sets but denser
- Genetic differences in DNA that cause phenotypic differences likely close to many markers
- Accounts for small and ambiguous SNP effects on traits
- Should allow WGS to account for more genetic variation
- As of now unproven but promising

Potential looks good

Simulated accuracy of genetic prediction:

Simulated	EPD	Whole Genome
Sires	72%	84%
Progeny with no records	41%	69%
Unrelated with no records	0%	55%

Dr. Warren Snelling

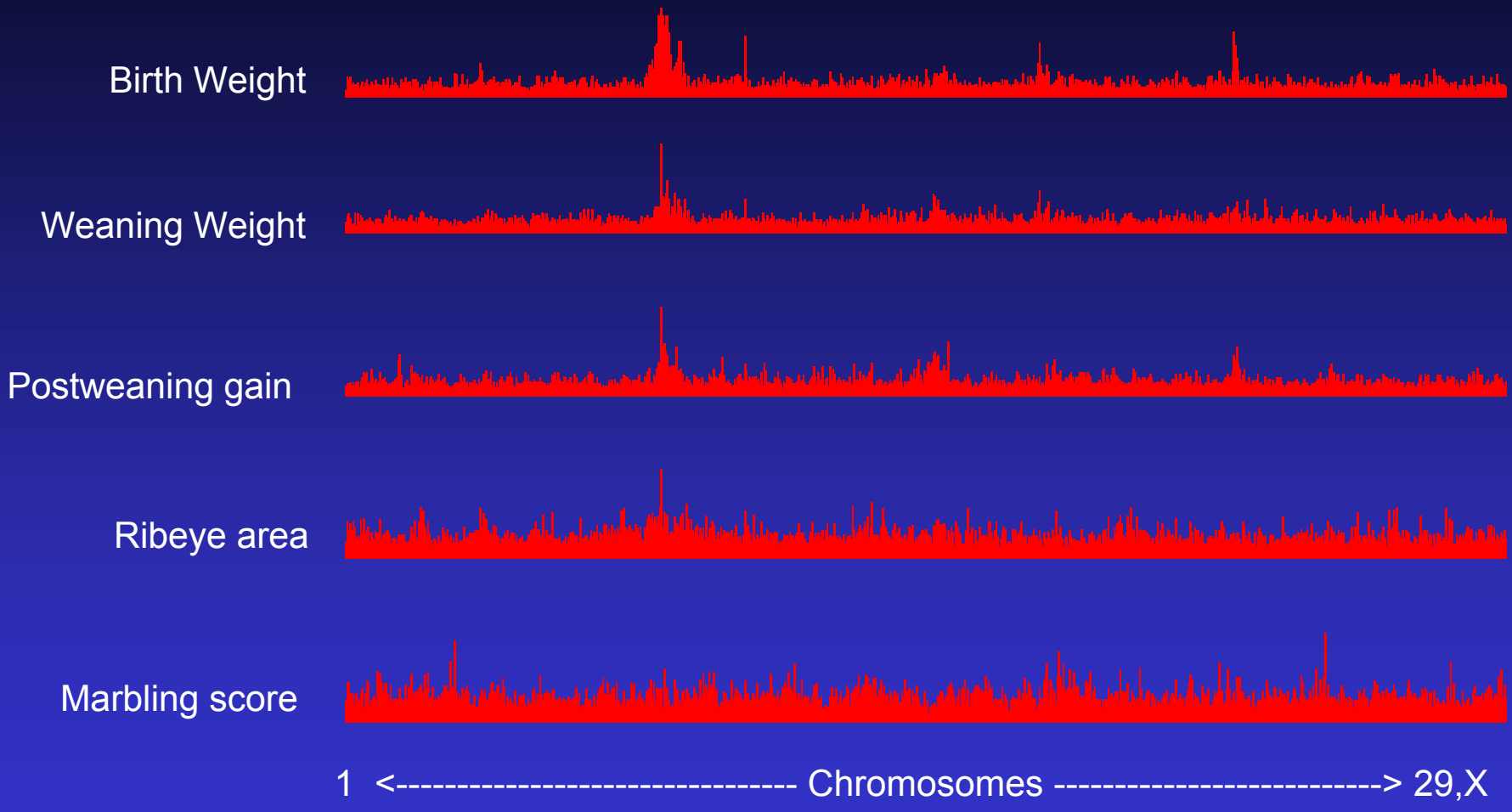
Steps to an initial WGS:

- Low cost for 1,000's of genotypes (done)
- Measure traits on 1,000's of animals or get EPDs on 1,000's of animals (done)
- Genotype these animals (done)
- Analyze the above training data (in progress)
- Genotype breeding animals (in progress)
- Estimate molecular breeding values
- Interpret and disseminate

Training Data

- 3,000+ head of pedigreed animals with extensive phenotypes at USMARC genotyped using the 50K chip:
  - 2,000+ with individual feed intake in finishing or heifer development phase
  - 2,000+ with carcass data, slice shear force, and rib dissection
  - 1300+ with age at puberty, pregnancy rate, and maternal performance
  - 1,100+ that will eventually have individual feed intake as mature cows to estimate maintenance requirements
  - 3,000+ with calving and growth traits

Preliminary SNP Associations



- **Initial results look encouraging**

2,000 Bull Project

2000 Bull Project

- Collaborative effort between USMARC and 16 U.S. beef breed associations that register the most cattle and have a genetic evaluation system.
- Breed associations provide semen for DNA on influential sires
- USMARC runs 50K SNP chip on those 2,000 sires
- USMARC estimates molecular breeding values

Objectives

- Extend genetic predictions from USMARC phenotypes to industry bulls
  - EPDs for traits not typically reported (e.g. feed efficiency) delivered to breed associations
- Validate the effectiveness of WGS using EPDs from the 2000 bulls relative to USMARC data on common traits (e.g., weaning weight)
- Improve accuracy of EPDs for common traits
- Determine to what extent training data must be of the same breed as in which WGS will be applied

Number of Sires to Sample

- Angus 400
- Hereford 282
- Simmental 234
- Charolais 156
- Red Angus 154
- Limousin 145
- Gelbvieh 135
- Shorthorn 91
- Brangus 84
- Beefmaster 83
- Maine-Anjou 59
- Brahman 42
- Chiangus 39
- Santa Gertrudis 39
- Salers 37
- Braunvieh 20

Potential pitfalls

- We don't know that the process of WGS will work
- May need more than 50,000 markers
- Patents might restrict use? Our results will be accessible to all.
- Need to develop complex computational methods
  - 1992-2006: 1.7 million genotypes at USMARC
  - 2007-2008: ~300 million genotypes expected from chip results