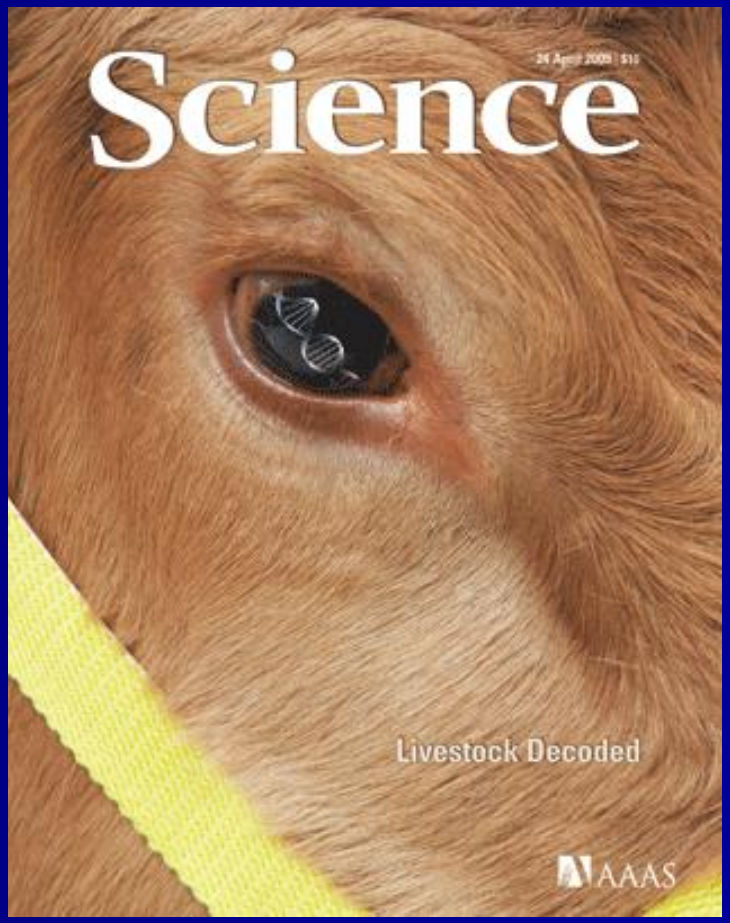
Implications and Application of Whole Genome (re) Sequencing

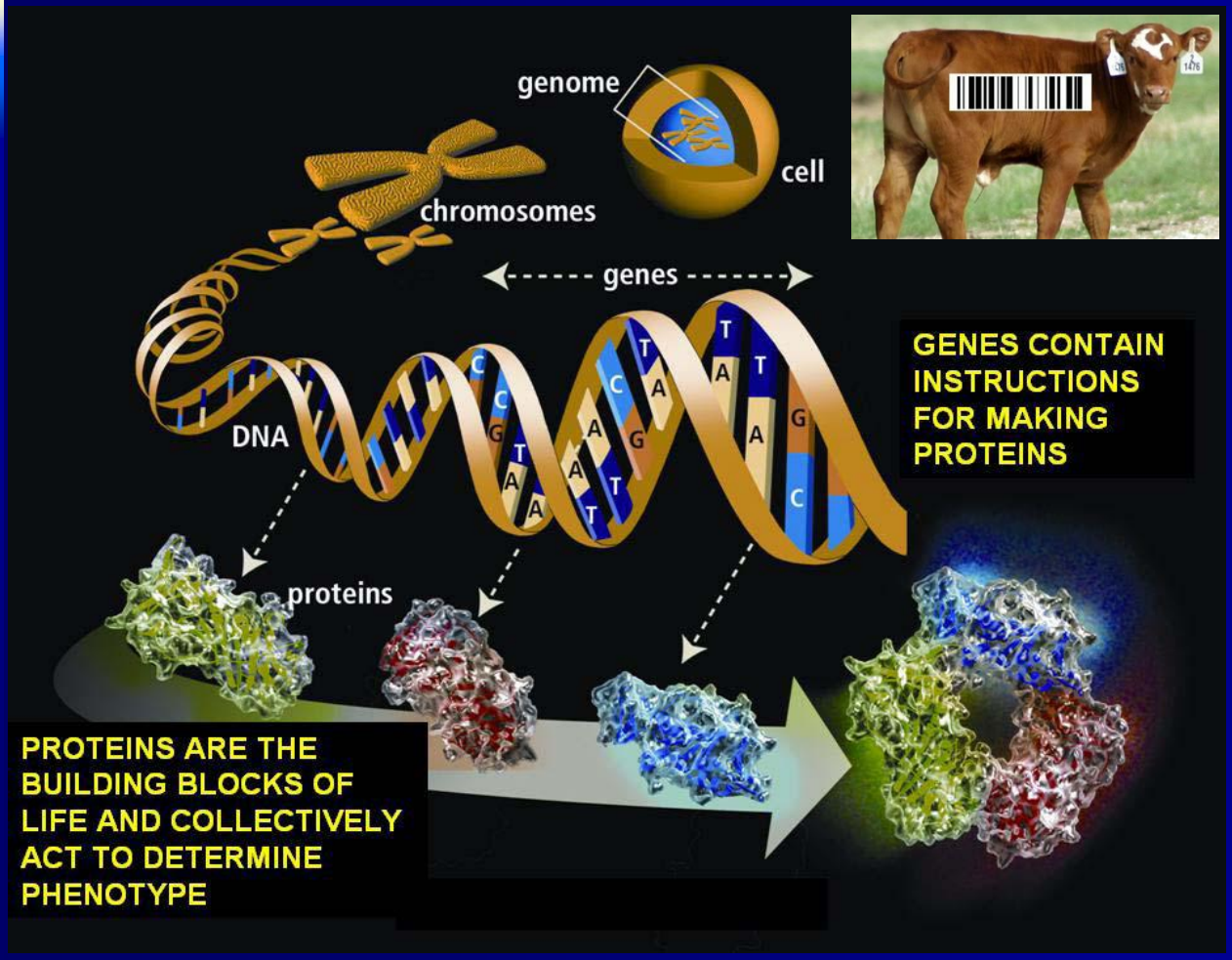
Alison Van Eenennaam

Animal Genomics and Biotechnology
Cooperative Extension Specialist
Department of Animal Science
University of California, Davis
alvaneenennaam@ucdavis.edu
(530) 752-7942



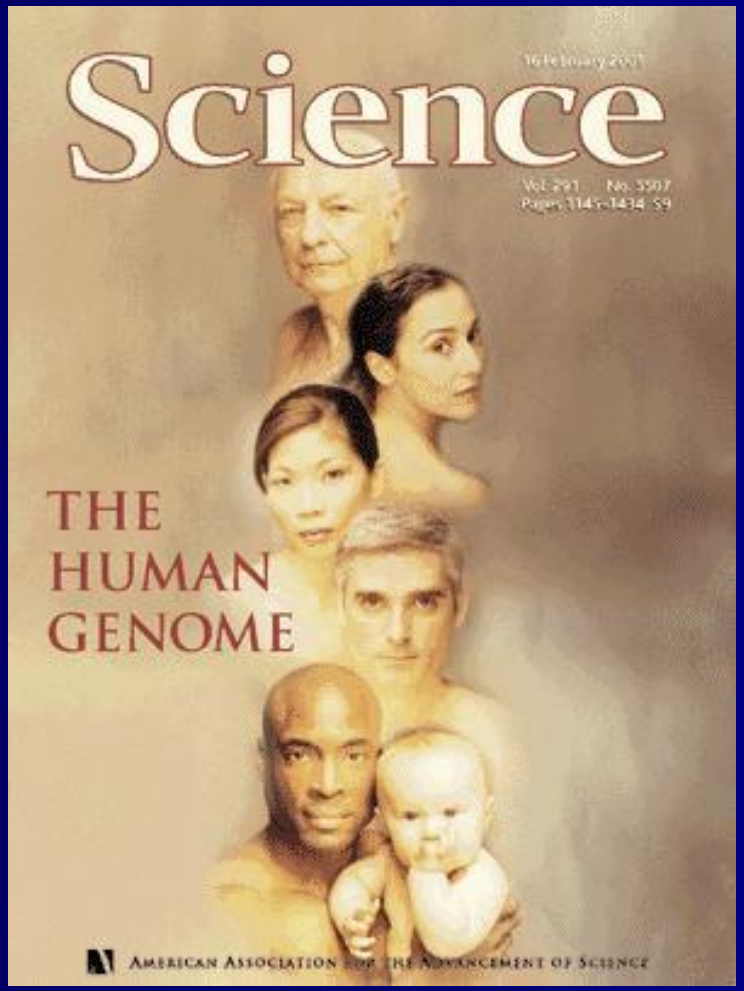
animalscience.ucdavis.edu/animalbiotech

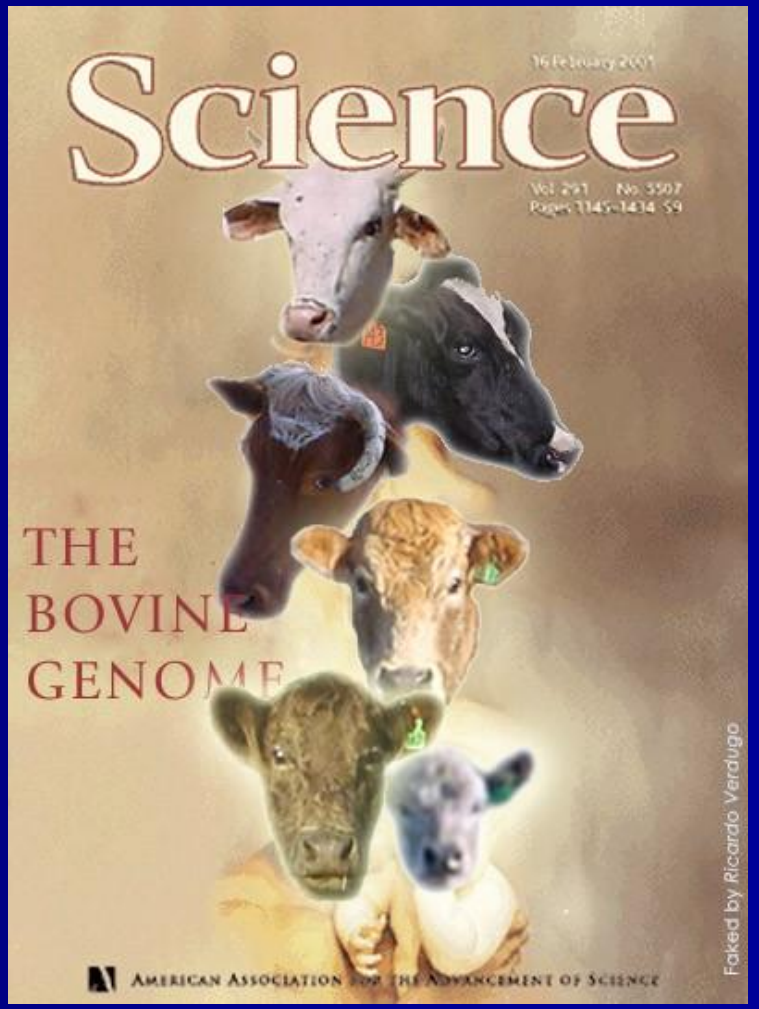
The bovine genome is similar in size to the genomes of humans, with an estimated size of 3 billion base pairs.



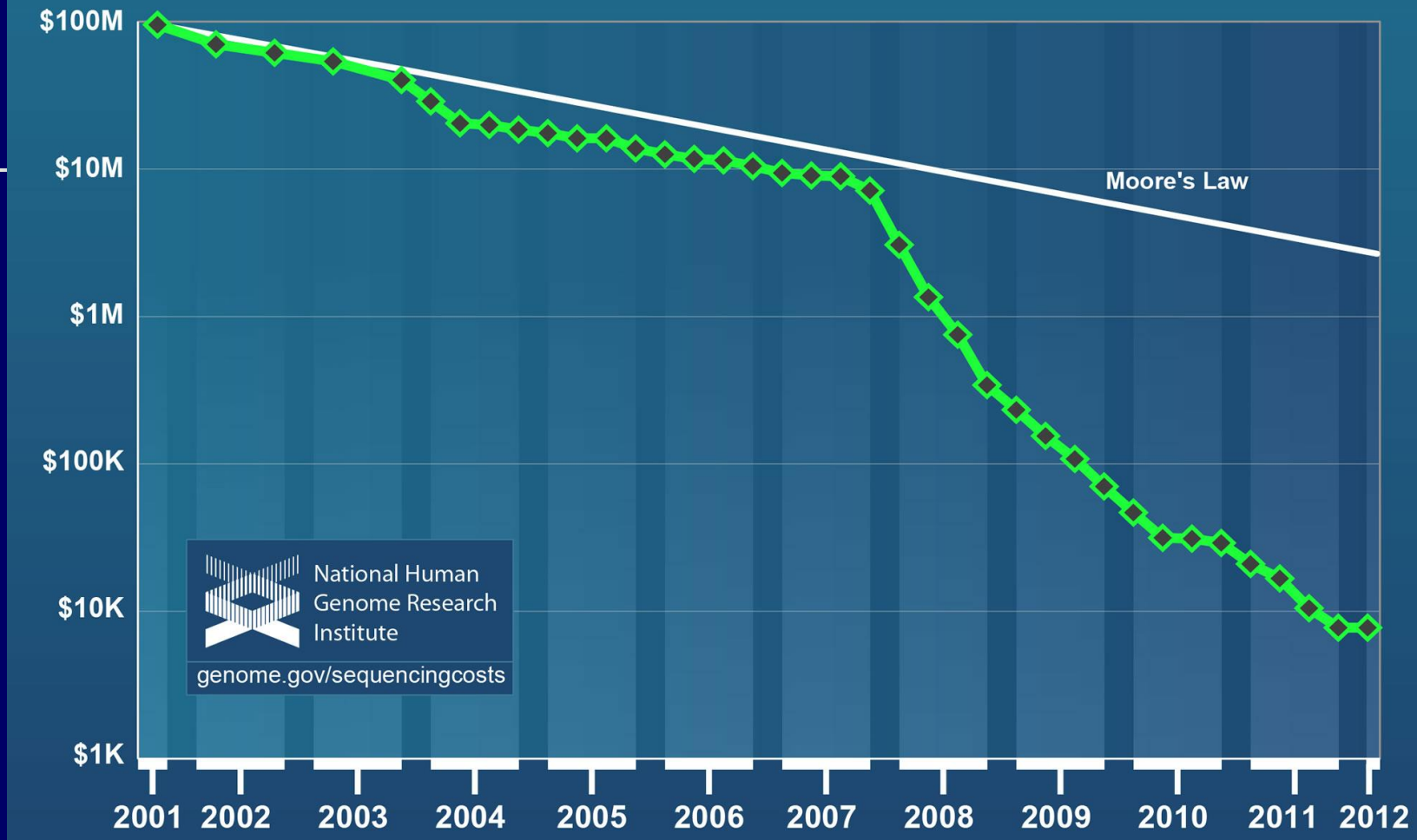
Human & cattle genomes are 83% identical

Human Gemone: 2001
Bovine Genome: 2009





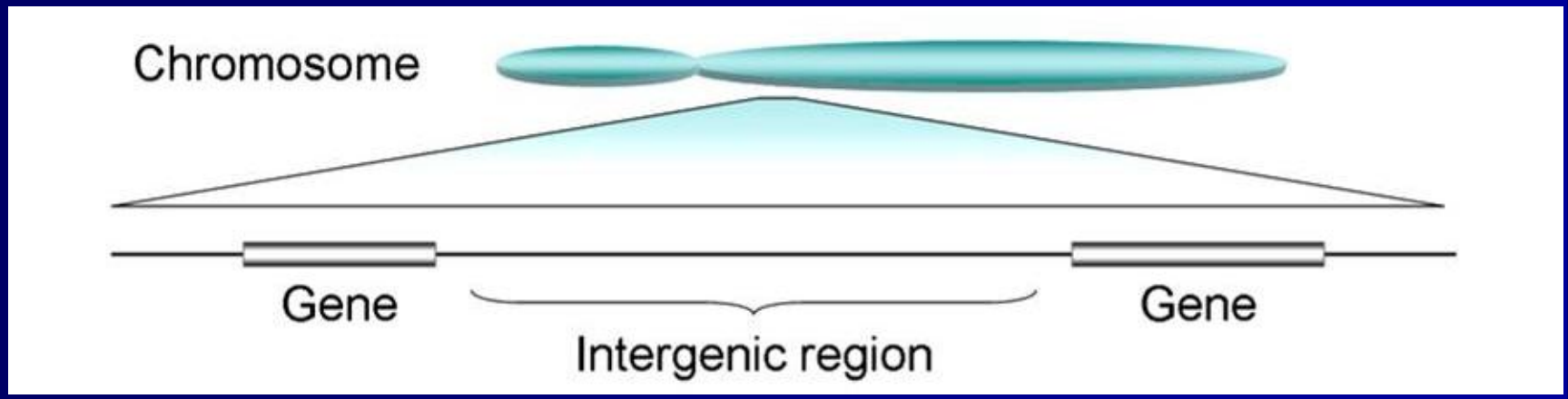
Cost per Genome

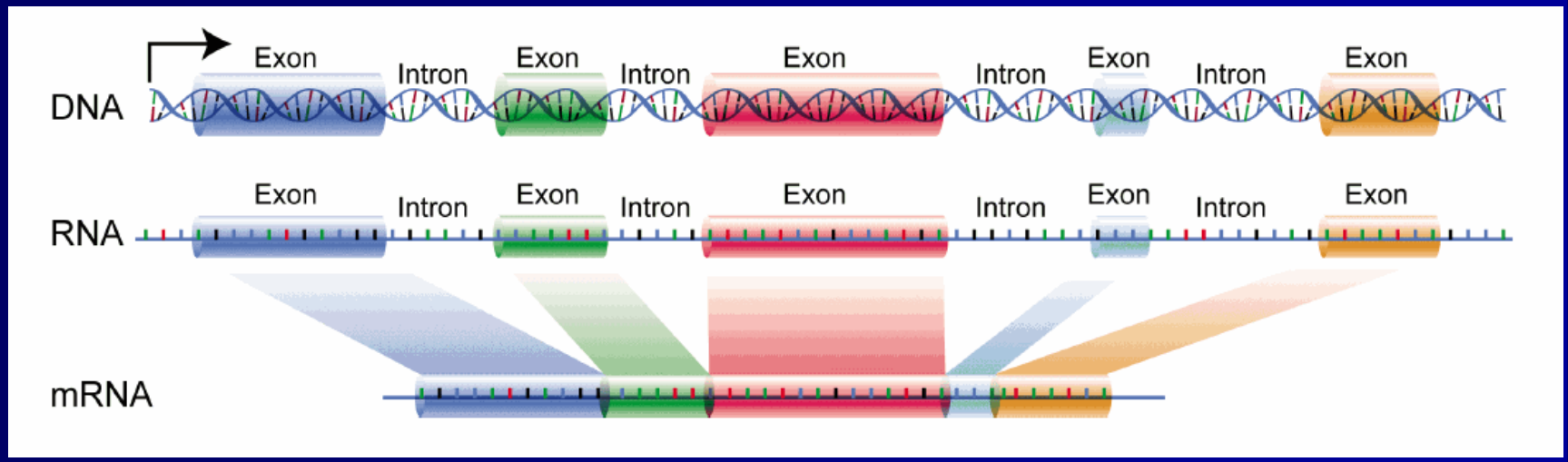


Moore's law is the observation that over the history of computing hardware, the number of transistors on integrated circuits doubles approximately every two years.

1990 – 2003	Human Genome Project	$3,000,000,000
2000 – 2007	Mouse Genome	$250,000,000
2002-2004	Rat Genome	$100,000,000
2003-2008	Bovine Genome	$53,000,000
2004-2006	Macaque Genome	$23,000,000
2007-2008	Baboon	$4,000,000
2006	Watson	$2,000,000
2007	Venter	$35,000,0000
2008	Deer Mouse	$700,000?
2008	1,000 Genomes Project	$350,000
2008	Glioma/Breast Cancer Project	$350,000
2008	Genetic Screen Genomes	<$100,000
2009	Disease case genomes	$5-20,000
2009-10	Your Genome	<$5,000

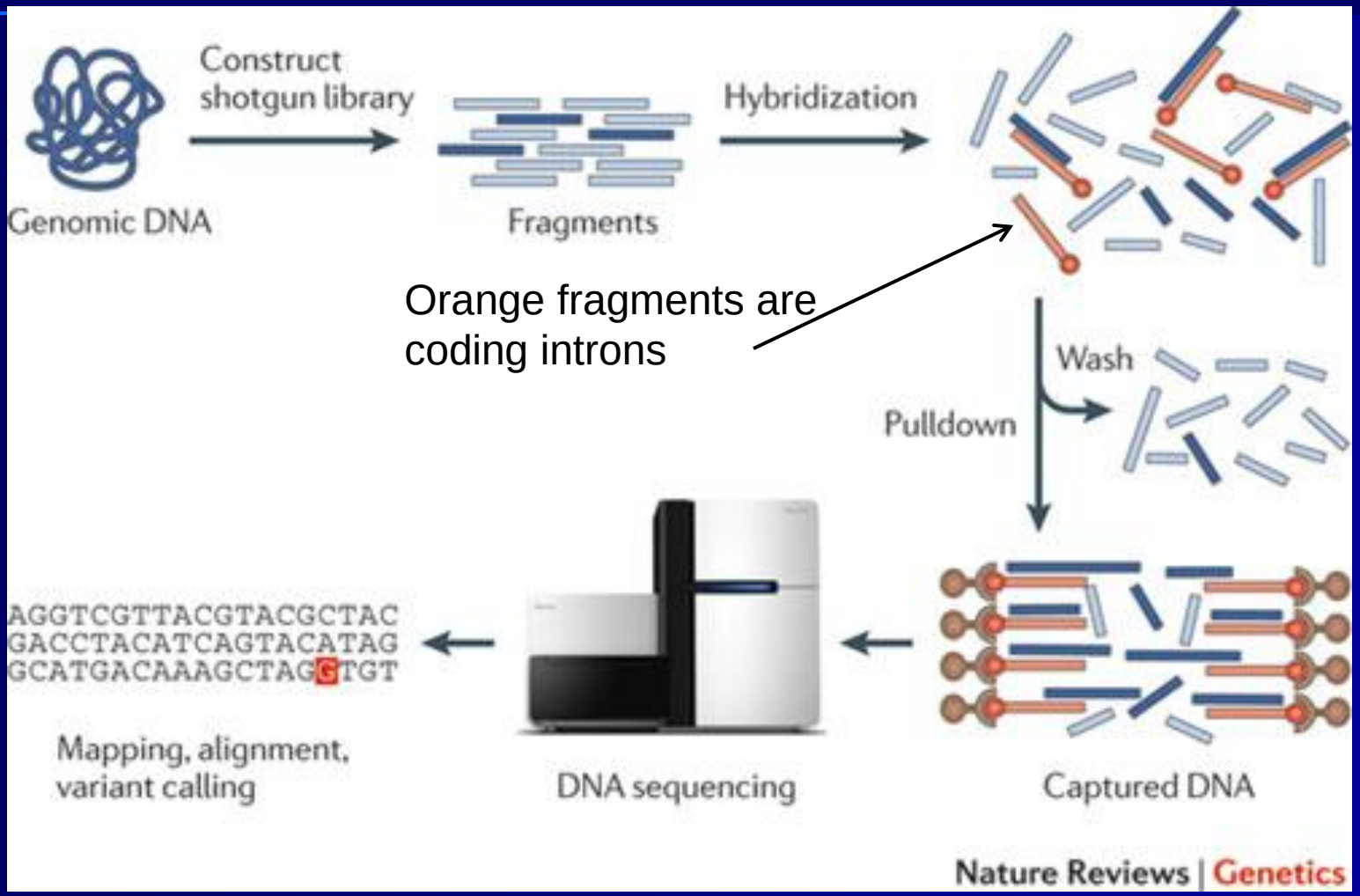
More than 98% of the human genome does not encode protein sequences, including most intergenic DNA and sequences within introns



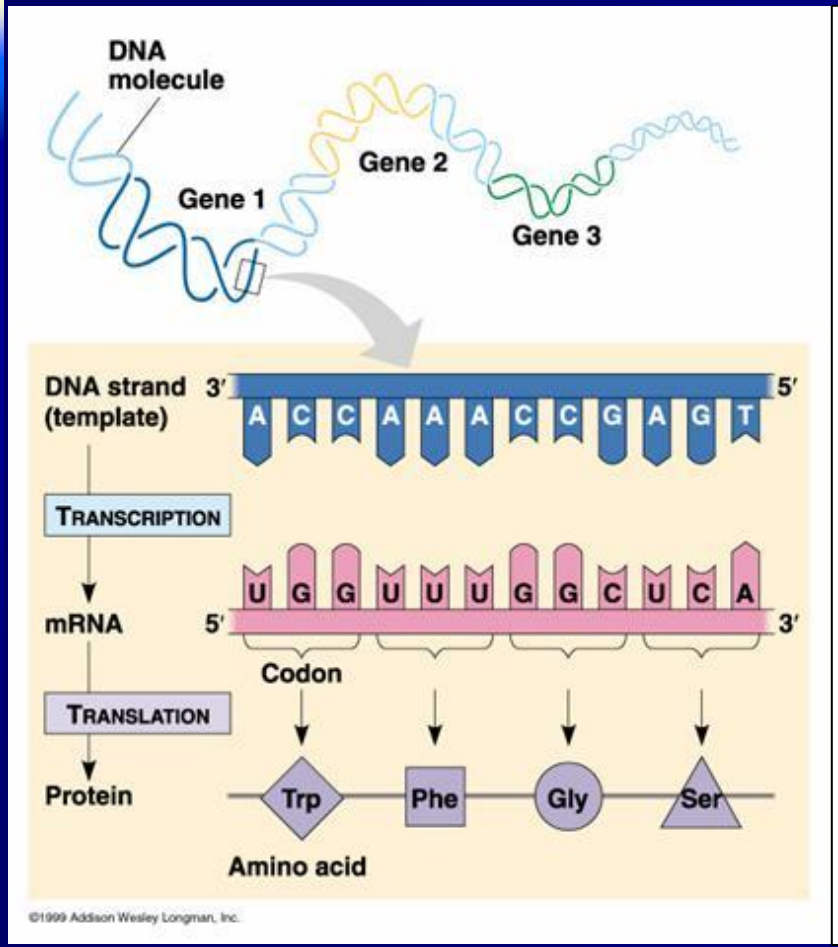


Exome sequencing offers an approach to enrich DNA for exon coding sequences (2%)

Michael J. Bamshad et al. 2011. Exome sequencing as a tool for Mendelian disease gene discovery Nature Reviews Genetics 12, 745-755.



The DNA sequence of a gene can be altered in a number of ways. Gene mutations have varying effects, depending on where they occur and whether they alter the function of essential proteins

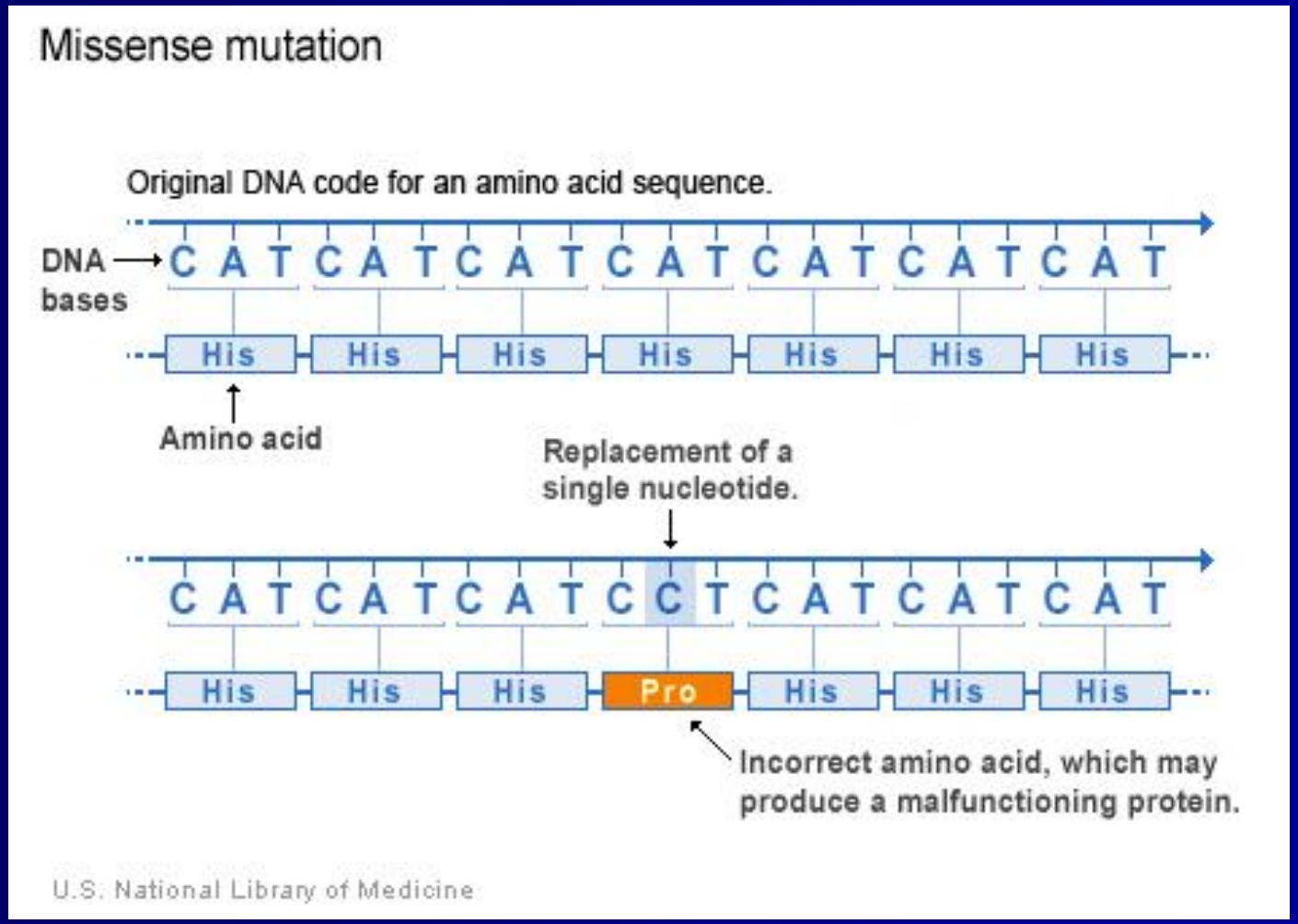


Second letter

First letter	U	C	A	G	Third letter
U	UUU Phe, UUC Phe, UUA Leu, UUG Leu	UCU Ser, UCC Ser, UCA Ser, UCG Ser	UAU Tyr, UAC Tyr, UAA Stop, UAG Stop	UGU Cys, UGC Cys, UGA Stop, UGG Trp	U C A G
C	CUU Leu, CUC Leu, CUA Leu, CUG Leu	CCU Pro, CCC Pro, CCA Pro, CCG Pro	CAU His, CAC His, CAA Gln, CAG Gln	CGU Arg, CGC Arg, CGA Arg, CGG Arg	U C A G
A	AUU Ile, AUC Ile, AUA Ile, AUG Met	ACU Thr, ACC Thr, ACA Thr, ACG Thr	AAU Asn, AAC Asn, AAA Lys, AAG Lys	AGU Ser, AGC Ser, AGA Arg, AGG Arg	U C A G
G	GUU Val, GUC Val, GUA Val, GUG Val	GCU Ala, GCC Ala, GCA Ala, GCG Ala	GAU Asp, GAC Asp, GAA Glu, GAG Glu	GGU Gly, GGC Gly, GGA Gly, GGG Gly	U C A G

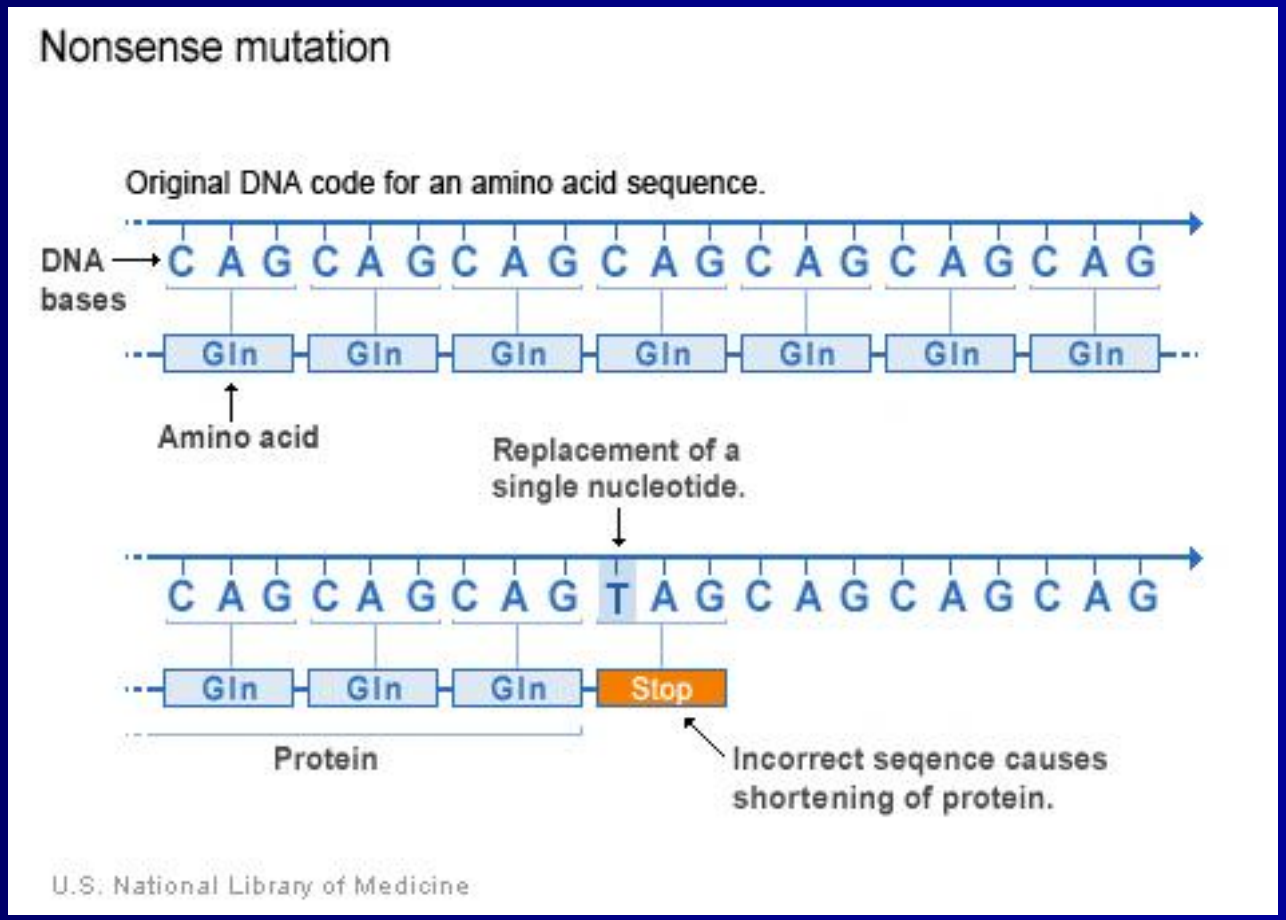
Missense mutation (as compared to synonymous)

This type of mutation is a change in one DNA base pair that results in the substitution of one amino acid for another in the protein made by a gene.



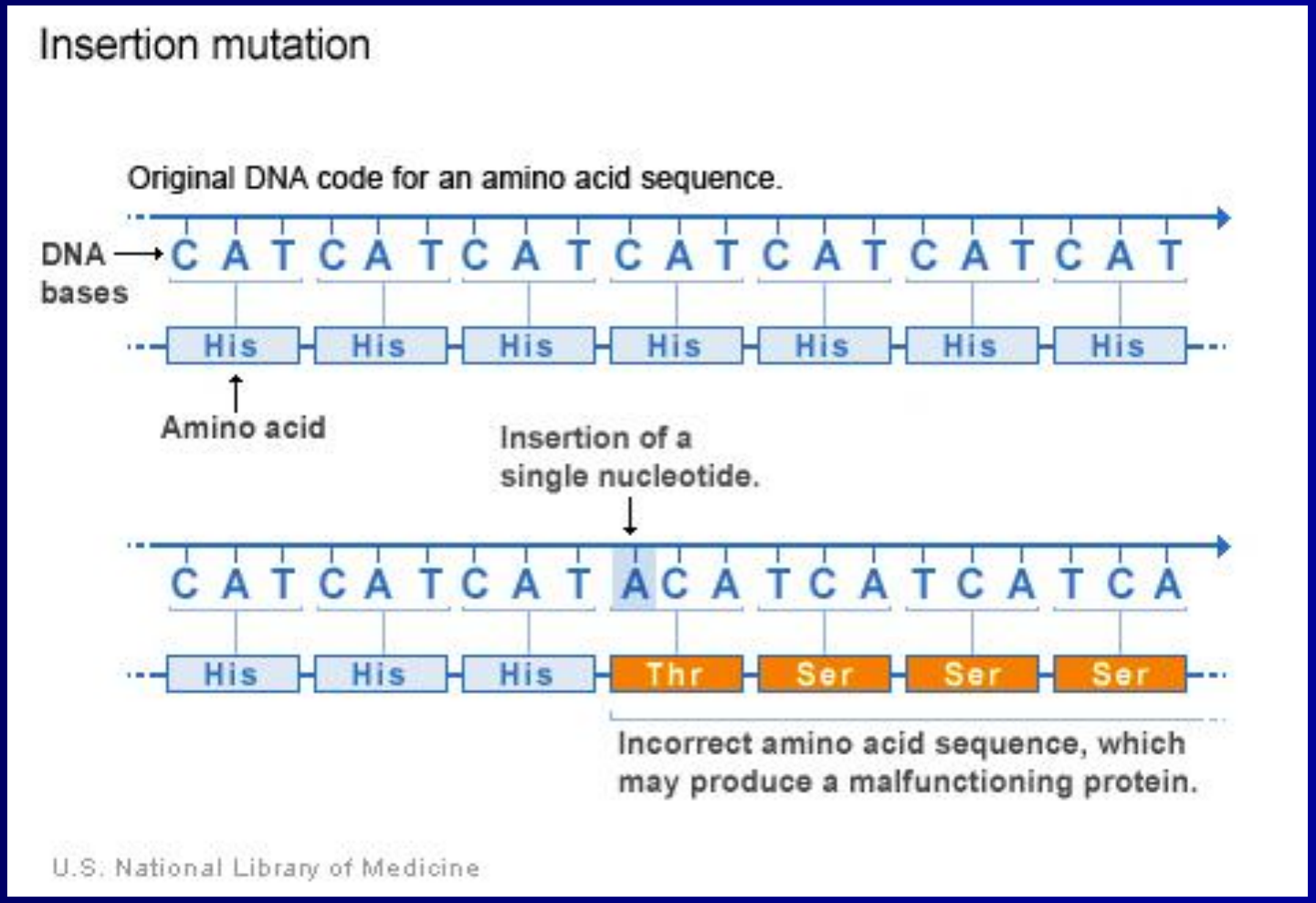
Nonsense mutation

A nonsense mutation is also a change in one DNA base pair. Instead of substituting one amino acid for another, however, the altered DNA sequence prematurely signals the cell to stop building a protein. This type of mutation results in a shortened protein that may function improperly or not at all.



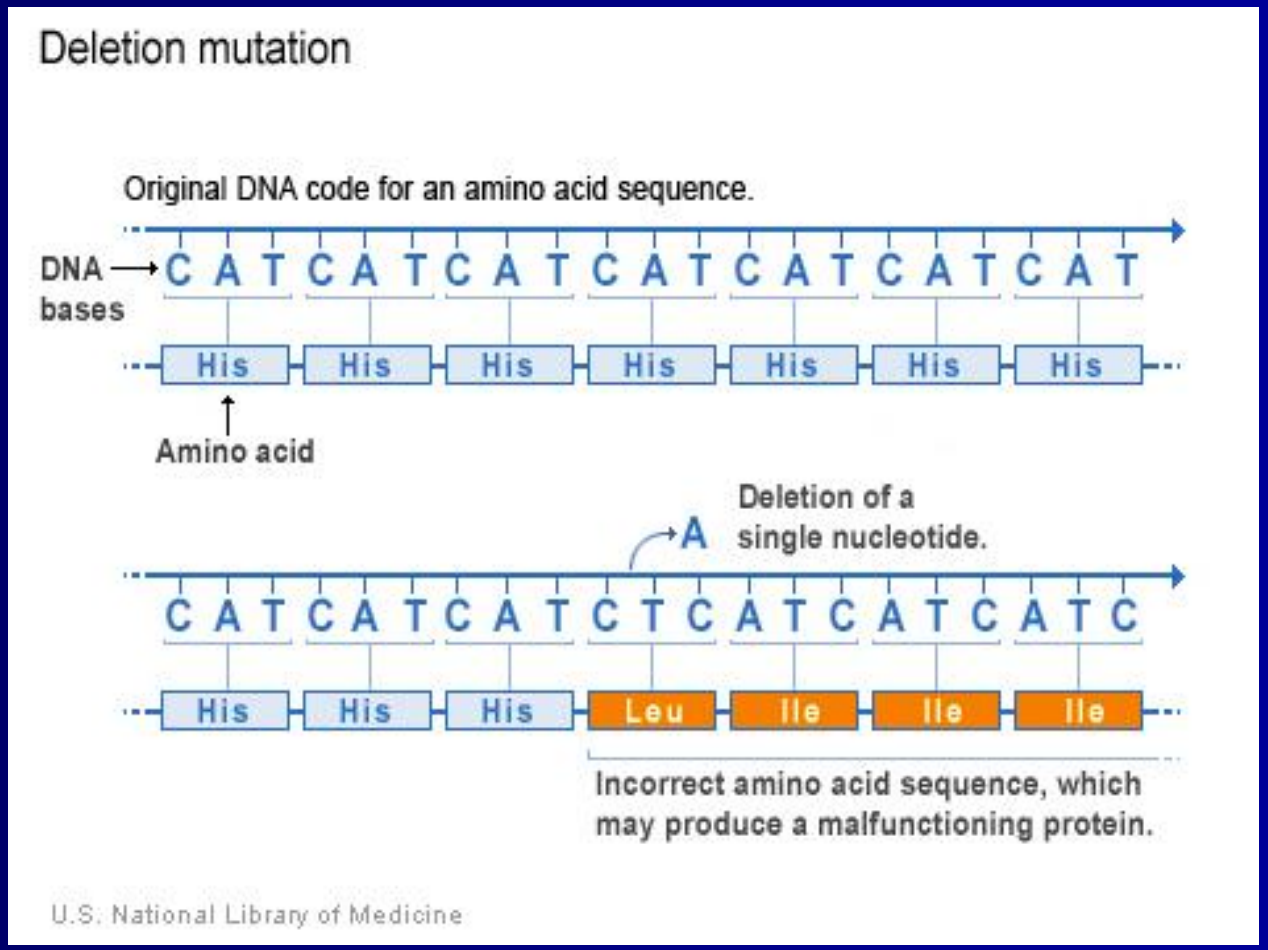
Insertion

An insertion changes the number of DNA bases in a gene by adding a piece of DNA. As a result, the protein made by the gene may not function properly.



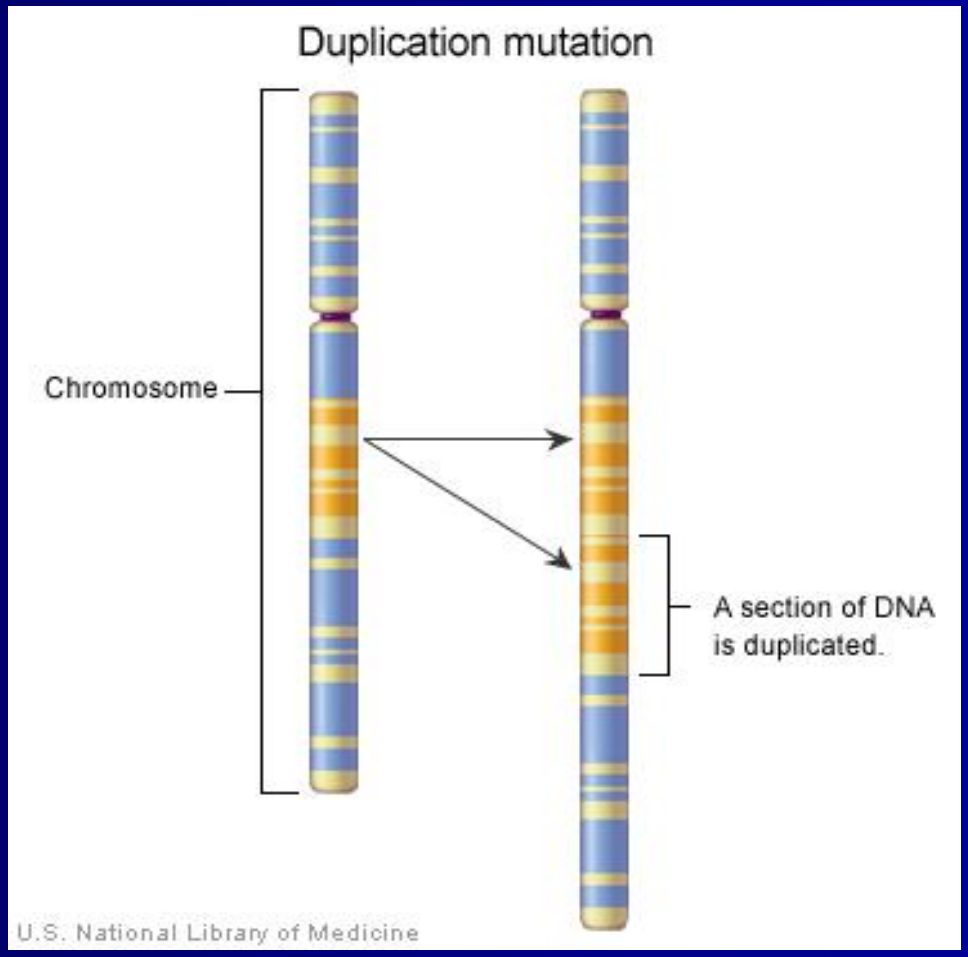
Deletion

A deletion changes the number of DNA bases by removing a piece of DNA. Small deletions may remove one or a few base pairs within a gene, while larger deletions can remove an entire gene or several neighboring genes. The deleted DNA may alter the function of the resulting protein(s).



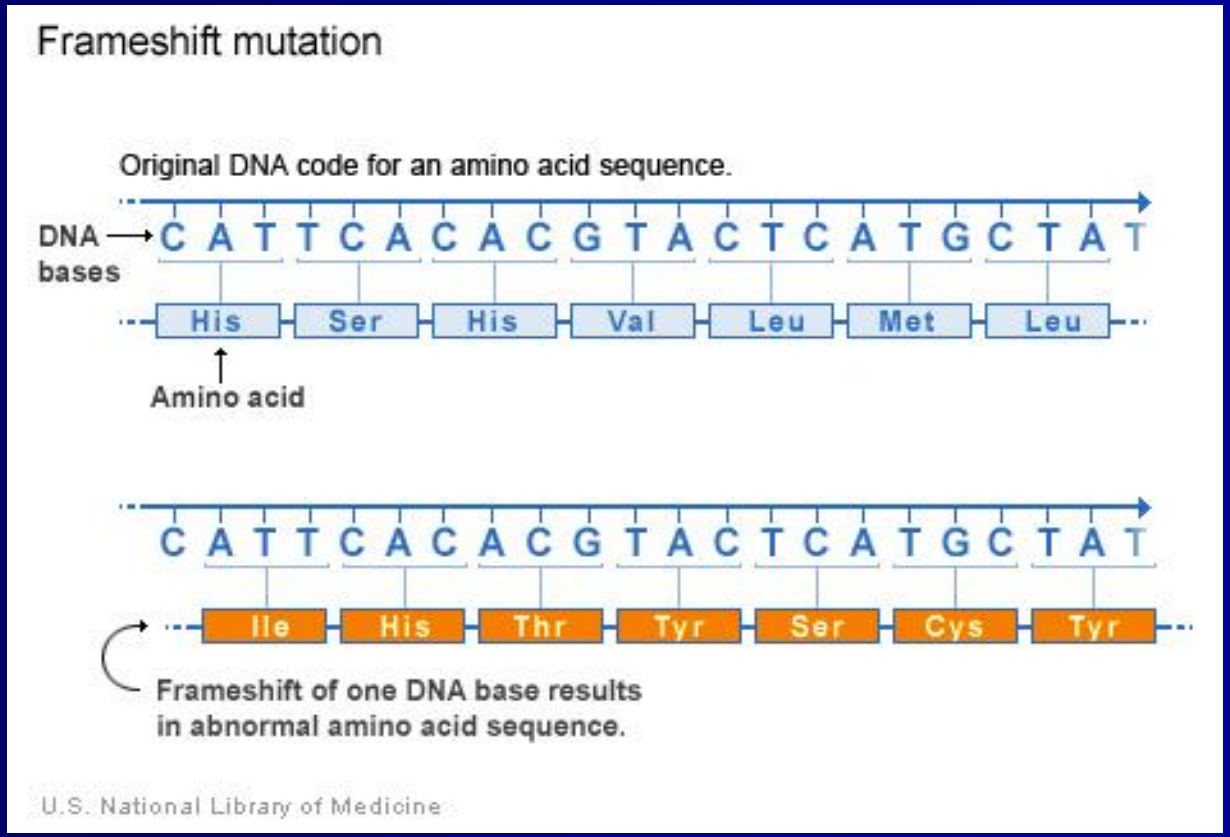
Duplication

A duplication consists of a piece of DNA that is abnormally copied one or more times. This type of mutation may alter the function of the resulting protein.



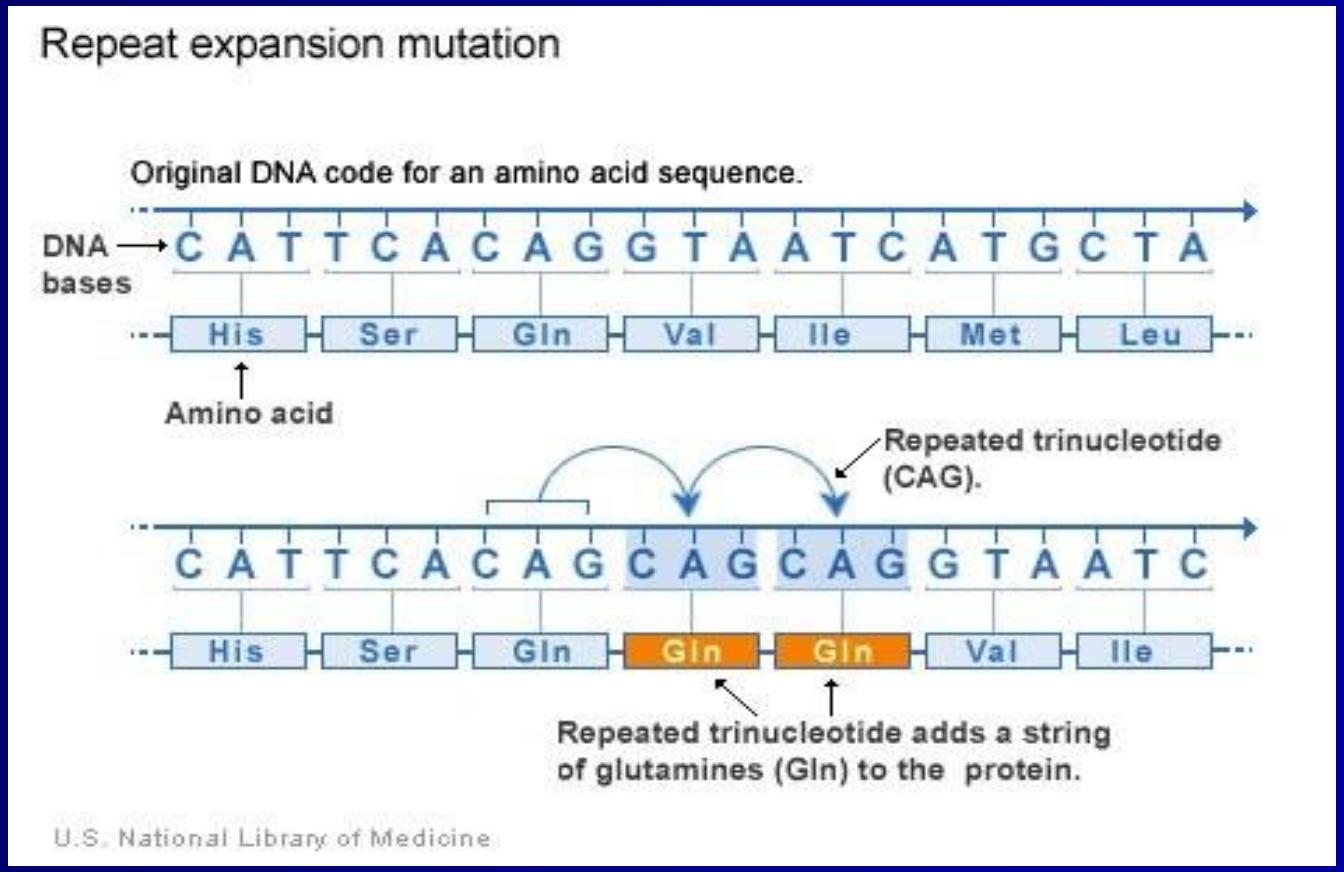
Frameshift mutation

This type of mutation occurs when the addition or loss of DNA bases changes a gene's reading frame. A reading frame consists of groups of 3 bases that each code for one amino acid (i.e. codon). A frameshift mutation shifts the grouping of these bases and changes the code for amino acids. The resulting protein is usually nonfunctional. Insertions, deletions, and duplications can all be frameshift mutations.



Microsatellites (SSR)

Nucleotide repeats are short DNA sequences that are repeated a number of times in a row. For example, a trinucleotide repeat is made up of 3-base-pair sequences, and a tetranucleotide repeat is made up of 4-base-pair sequences. This type of mutation can cause the resulting protein to function improperly.



Background information needed to understand why resequencing might be important/valuable

- Sequencing costs are dropping rapidly
- In the future we might be able to cheaply obtain individual animal sequence(s)
- Not all mutations are going to have an effect
- Even those that change a protein or eliminate a protein may not have an effect
- Need to prioritize variants based on the likelihood that they have an effect

What could be done with genome sequence?

- Discovery of causative SNPs associated with disease
- Discovery of missing homozygotes
- Improve the accuracy of genomic selection?
- Enabling better methods to identify epistasis

SNPS associated with disease….

"It will be essential to develop methods that prioritize SNP variants based on the likelihood that they contribute to disease".

- The frequencies of different classes of variations in ten case and ten control human genomes were compared (K. V. Shianna *et al.*, unpublished data).
- There were 383,913 variants (single nucleotide variants and indels) present in at least two cases and no controls.
- However, if testing is restricted to only variants that affect the coding sequence (i.e. missense mutations), this number drops to 2,354
- If testing is restricted to only protein-truncating variants (i.e. nonsense mutations), the number drops further to 152

Cirulli and Goldstein, 2010. Uncovering the role of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics 11:415.

Average numbers of variants found within the genomes of 11 re-sequenced registered Angus bulls and a comparison to pertinent human 1000 Genome Project findings. (Taylor, Schnabel *et al*., unpublished)

Description	Avg per bull	1000 Genomes
Splice site +/- 2bp	1,055	
UTR	27,883	
Indels non-genic	359,356	
Indels genic	110,633	
Indels (inframe) that affect 1-2-3 amino acids (AA)	100	190-210
Indels that cause frameshifts	656	300-350
Stop codon usage	585	
High quality SNP synonymous AA	23,764	10-12,000
High quality SNP nonsynonymous AA	25,750	10-11,000
High quality SNP genic region	2,028,627	
High quality SNP	1,367,128	
High quality homozygous SNP (differing to Dominette)	2,853,793	
High quality heterozygous SNP	1,055	

Missing homozygotes….

Display Settings: Abstract Send to:

J Dairy Sci. 2011 Dec;94(12):6153-61.

Harmful recessive effects on fertility detected by absence of homozygous haplotypes.

VanRaden PM, Olson KM, Null DJ, Hutchison JL.

Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350, USA. paul.vanraden@ars.usda.gov

Abstract

Five new recessive defects were discovered in Holsteins, Jerseys, and Brown Swiss by examining haplotypes that had a high population frequency but were never homozygous. The method required genotypes only from apparently normal individuals and not from affected embryos. Genotypes from the BovineSNP50 BeadChip (Illumina, San Diego, CA) were examined for 58,453 Holsteins, 5,288 Jerseys, and 1,991 Brown Swiss with genotypes in the North American database. Haplotypes with a length of ≤ 75 markers were obtained. Eleven candidate haplotypes were identified, with the earliest carrier born before 1980; 7 to 90 homozygous haplotypes were expected, but none were observed in the genomic data. Expected numbers were calculated using either the actual mating pattern or assuming random mating. Probability of observing no homozygotes ranged from 0.0002 for 7 to 10^{-4} [5] for 90 expected homozygotes. Phenotypic effects were confirmed for 5 of the 11 candidate haplotypes using 14,911,387 Holstein, 830,391 Jersey, and 68,443 Brown Swiss records for conception rate. Estimated effect for

If allele frequency of SNP is 50% A: 50%T then expect 25% AA; 50% AT, 25% TT

If see 33% AA and 66% AT then have a case of missing homozygotes – likely lethal

Haplotypes Affecting Fertility and their Impact on Dairy Cattle Breeding Programs

Dr. Kent A. Weigel, University of Wisconsin

http://documents.crinet.com/Genex-Cooperative-Inc/Dairy/KWeigel-Haplotypes-Affecting-Fertility.pdf

- The exact genes and their underlying biological roles in fertilization and embryo development are unknown, but it is assumed that the outcome of inheriting the same haplotype from both parents is failed conception or early embryonic loss.
- The reactive approach of attempting to eradicate every animal with an undesirable haplotype is not recommended in light of their economic impact, and is not practical given the likelihood **that many more undesirable haplotypes will be found.**
- Producers should neither avoid using bulls with these haplotypes nor cull cows, heifers, and calves that are carriers, because this will lead to significant economic losses in other important traits.
- Computerized mating programs offer a simple, inexpensive solution for avoiding affected matings, so producers should use these programs and follow through on the mating recommendations.

Improving the accuracy of genomic selection?

- According to a simulation presented by Meuwissen and Goddard a 40% gain in accuracy in predicting genetic values could be achieved by using sequencing data instead of data from 30K SNP arrays alone.
- Furthermore, by using whole-genome sequencing data, the prediction of genetic value was able to remain accurate even when the training and evaluation data were 10 generations apart: observed accuracies were similar to those in which the test and training data came from the same generation.
- According to the authors, "these results suggest that with a combination of genome sequence data, large sample sizes, and a statistical method that detects the polymorphisms that are informative..., high accuracy in genomic prediction is attainable"

Meuwissen and Goddard, 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185(2):623-31.

HOWEVER – did not help *Drosophila* too much

OPEN ACCESS Freely available online

PLoS GENETICS

Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in *Drosophila melanogaster*

PLoS Genet 2012 8(5): e1002685. doi:10.1371/journal.pgen.1002685

Ulrike Ober[1]*, Julien F. Ayroles[2,3], Eric A. Stone[2], Stephen Richards[4], Dianhui Zhu[4], Richard A. Gibbs[4], Christian Stricker[5], Daniel Gianola[6], Martin Schlather[7], Trudy F. C. Mackay[2], Henner Simianer[1]

1 Animal Breeding and Genetics Group, Georg-August-University Göttingen, Göttingen, Germany, **2** Department of Genetics, North Carolina State University, Raleigh, North Carolina, United States of America, **3** Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America, **4** Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, United States of America, **5** agn Genetics GmbH, Davos, Switzerland, **6** Department of Animal Sciences, University of Wisconsin–Madison, Wisconsin, United States of America, **7** Institute for Mathematics, University of Mannheim, Mannheim, Germany

Abstract

Predicting organismal phenotypes from genotype data is important for plant and animal breeding, medicine, and evolutionary biology. Genomic-based phenotype prediction has been applied for single-nucleotide polymorphism (SNP) genotyping platforms, but not using complete genome sequences. Here, we report genomic prediction for starvation stress resistance and startle response in *Drosophila melanogaster*, using ~2.5 million SNPs determined by sequencing the Drosophila Genetic Reference Panel population of inbred lines. We constructed a genomic relationship matrix from the SNP data and used it in a genomic best linear unbiased prediction (GBLUP) model. We assessed predictive ability as the correlation between predicted genetic values and observed phenotypes by cross-validation, and found a predictive ability of 0.239 ± 0.008 (0.230 ± 0.012) for starvation resistance (startle response). The predictive ability of BayesB, a Bayesian method with internal SNP selection, was not greater than GBLUP. Selection of the 5% SNPs with either the highest absolute effect or variance explained did not improve predictive ability. Predictive ability decreased only when fewer than 150,000 SNPs were used to construct the genomic relationship matrix. We hypothesize that predictive power in this population stems from the SNP–based modeling of the subtle relationship structure caused by long-range linkage disequilibrium and not from population structure or SNPs in linkage disequilibrium with causal variants. We discuss the implications of these results for genomic prediction in other organisms.

And these same authors suggest "*the importance of epistasis as a principal factor that determines variation for quantitative traits and provides a means to uncover genetic networks affecting these traits*".

PNAS September 25, 2012 vol. 109 no. 39 15553-15559

Epistasis dominates the genetic architecture of *Drosophila* quantitative traits

Wen Huang[a], Stephen Richards[b], Mary Anna Carbone[a], Dianhui Zhu[b], Robert R. H. Anholt[c], Julien F. Ayroles[a,1], Laura Duncan[a], Katherine W. Jordan[a], Faye Lawrence[a], Michael M. Magwire[a], Crystal B. Warner[b,2], Kerstin Blankenburg[b], Yi Han[b], Mehwish Javaid[b], Joy Jayaseelan[b], Shalini N. Jhangiani[b], Donna Muzny[b], Fiona Ongeri[b], Lora Perales[b], Yuan-Qing Wu[b,3], Yiqing Zhang[b], Xiaoyan Zou[b], Eric A. Stone[a], Richard A. Gibbs[b], and Trudy F. C. Mackay[a,4]

Departments of [a]Genetics and [c]Biology, North Carolina State University, Raleigh, NC 27695; and [b]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030

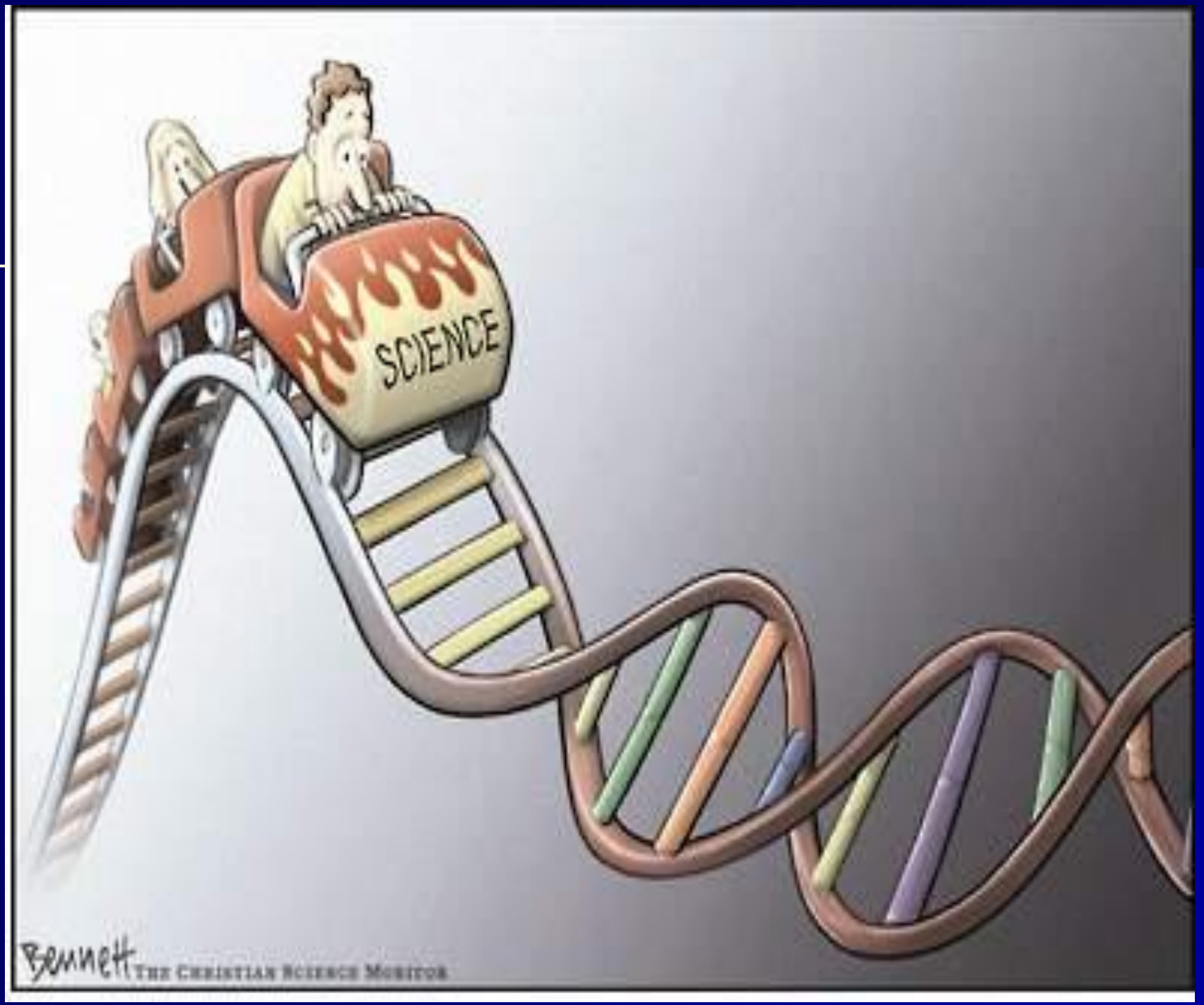
This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2010.

Contributed by Trudy F. C. Mackay, August 6, 2012 (sent for review May 29, 2012)

"We speculate that epistatic gene action is also an important feature of the genetic architecture of quantitative traits in other organisms, including humans. Our analysis paradigm (first identifying loci associated with a quantitative trait in two populations with different allele frequencies and then using these loci as foci for a genome-wide screen for pairwise epistatic interactions) can be applied to any organism for which such populations exist**. For example, human GWASs have been plagued by a lack of replicated associations across populations in even large studies**. We argue that this finding is expected under epistatic gene action and variable allele frequencies."

Conclusions

- In the future we might be able to cheaply obtain individual animal sequence(s)
- This will undoubtedly generate a lot of data
- Will likely need significant improvement in data management and bioinformatics platforms, statistical methods, and development of computer mating software
- Making intelligent/wise use of these data is the challenge!! (i.e. translational genomics)


SCIENCE
Bennett
The Christian Science Monitor