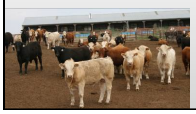
What do we hope to learn from sequencing?

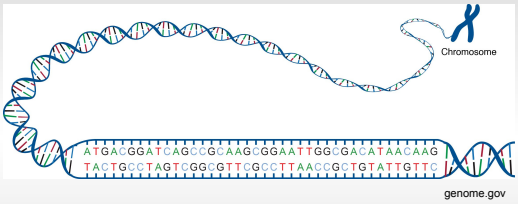
Larry Kuehn and Warren Snelling
Research Geneticists

USDA, ARS, U.S. Meat Animal Research Center

The USDA is an equal opportunity employer.



Genome sequence

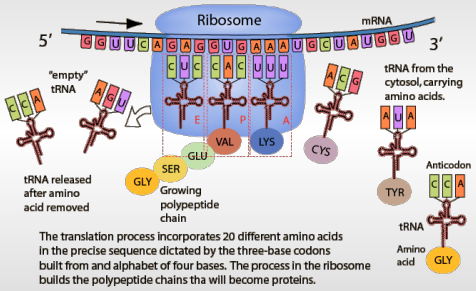


- Trying to read the base pairs (the code) along the whole genome

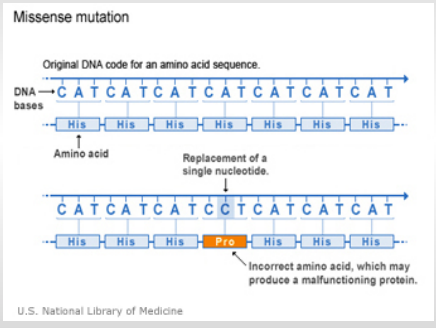
Whole genome sequencing

- Loosely defined as reading and recording all 3 billion bases in the cattle genome
  - Introns and intergenic regions
    - Do not code for any protein
    - Over 98% of genome
    - Likely some regulatory function
  - Exons
    - Remainder
    - Coding regions -- PROTEINS

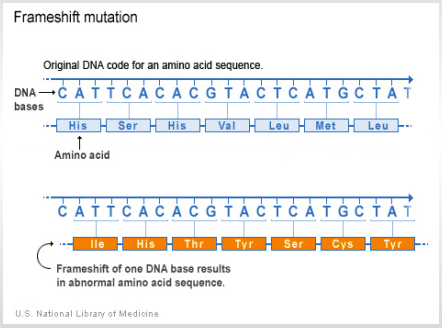
Exon - protein coding



Mutations – genomic variation



Mutations – genomic variation



First cattle genome sequence

- Dominette
  - USDA-ARS – Miles City Montana
  - Inbred Hereford female
- Impact
  - Reference Bovine genome
  - Cost over $53 million
  - Led to DNA chips (50K chip, etc.)
  - Base for further sequencing projects

USMARC sequence technology progression:

1992 Pharmacia manual slab gel – about 20,000 bases wk^{-1} $machine^{-1}$, about $7 per 1,000 base read; used for microsatellite sequencing ($7,000 per Mb)

1997 ABI 377 gel-based system – about 380,000 bases wk^{-1} $machine^{-1}$, about $3 per 800 base read; used for small scale gene sequencing and marker development ($3,750 per Mb)

1999 ABI 3700 capillary system – about 3.5 Mb wk^{-1}, about $2 per 600 base read; used for "large scale" EST sequencing, marker development and start of SNP discovery program ($3,300 per Mb)

2003 ABI 3730 capillary system – about 7.5 Mb wk^{-1}, about $0.80 per 600 base read; used for EST sequencing, BAC-targeted SNP discovery, lots of amplicon resequencing in breed panels ($1,300 per Mb)

Next generation sequencing

- Multiple platforms available

Next generation sequencing

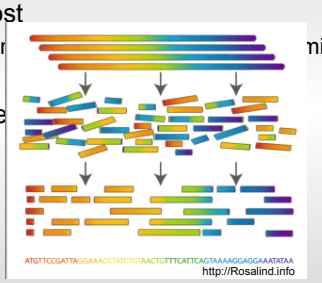
- Accuracy ranges from 87% to 99%
  - General trend is decreased accuracy with longer read-length
  - Generally read 50 to 1000+ bases per read
  - Easier to assemble longer reads
- Net effect is cheaper genome sequencing
  - Costs from $0.05 to $10 per million bases
  - Enough sequence for $150 whole genome

Some complications

- Initial cost
  - Sequencer costs from $80,000 to $1 million
- Genome assembly

Some complications

- Initial cost
  - Sequer illion
- Genome



Some complications

- Initial cost
  - Sequencer costs from $80,000 to $1 million
- Genome assembly
  - Overlap helps
    - Long reads better
    - Helps to compare to base sequence (other cattle)

Complications

- Tremendous amount of data
  - Hard to store, let alone manipulate
  - Makes assembly even tougher
  - Have to develop tools based on results
    - For example, genotyping based on new markers discovered in sequence (50K chip)
    - Hard to decide what 'differences' among animals are important
      - As with chips, phenotypes are still very important

Complications

- Multiple reads required
  - Reading 3 billion bases does not mean whole genome read
    - Each animal has two full genomes (diploid)
    - Genome is fragmented randomly into small pieces
    - Same section of genome likely read numerous times
    - Some advocate sequencing 20-50x when targeting whole genome
      - Focusing on one animal reduces impact of examining multiple animals (diversity, discovery, etc.)

Why bother?

- From a geneticist perspective:
  - Interested in sequencing to improve our chance at finding causal variation
    - Examine differences in sequence
    - Leads to finding markers and mutations
    - Mutations may change protein structure or protein regulation
  - Ultimately differences we see due to genetics lie with differences in the genome or its regulation

Two different strategies

- Compare genomes of diverse animals to see where they are different
  - Marker development, mutational candidates
  - Follow up with genotyping platform
- Associate differences in sequence among several animals with trait variation
  - Requires large numbers of sequenced animals
- First strategy cheaper, more restrictive

Functional variants

Impact	Effect	
High	SPLICE_SITE_ACCEPTOR	FRAME_SHIFT
	SPLICE_SITE_DONOR	STOP_GAINED
	START_LOST	STOP_LOST
	EXON_DELETED	RARE_AMINO_ACID
Moderate	NON_SYNONYMOUS_CODING	CODON_CHANGE
	CODON_INSERTION	CODON_DELETION
	UTR_5_DELETED	UTR_3_DELETED
	CODON_CHANGE_PLUS_CODON_INSERTION	
	CODON_CHANGE_PLUS_CODON_DELETION	
Low	SYNONYMOUS_START	NON_SYNONYMOUS_START
	SYNONYMOUS_CODING	NON_SYNONYMOUS_STOP
	SYNONYMOUS_STOP	START_GAINED
Modifier	All other effects	

Congolani et al., 2012

Summary

- Sequencing offers many possibilities
  - Move toward causal variation
  - Increase selection opportunities
  - Detect microbial interactions with economically relevant traits
  - Lower genotyping costs

Current efforts

- So we have potential for a lot of data
- Now what?

- Mention of a trade name, proprietary product, or specific equipment does not constitute a guarantee or warranty by the USDA and does not imply approval to the exclusion of other products that may be suitable.