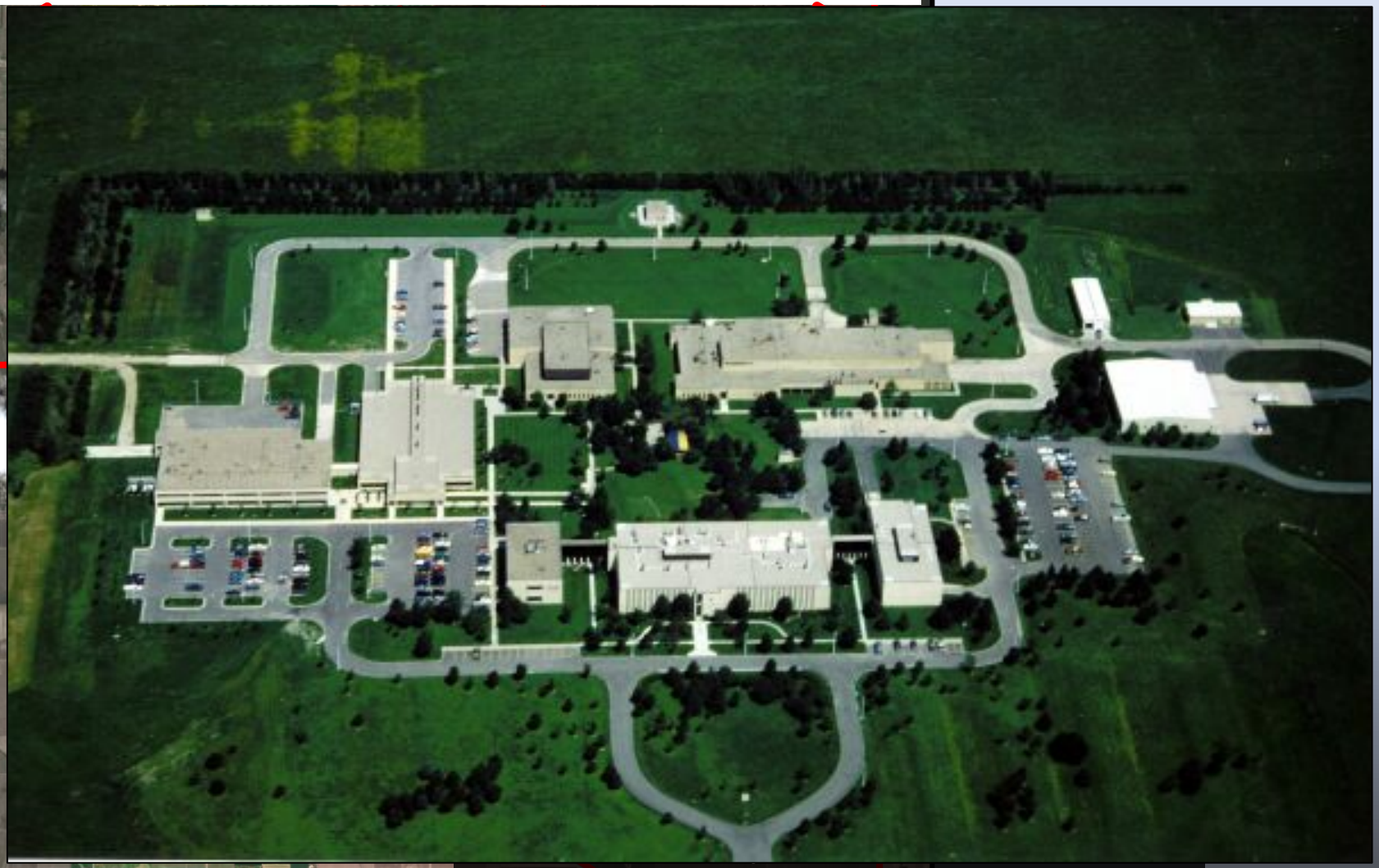
The evolution of reference assembly:
Improving animal genomes using long reads and high heterozygosity

January 17, 2018
PacBio Developer's Workshop
San Diego, CA

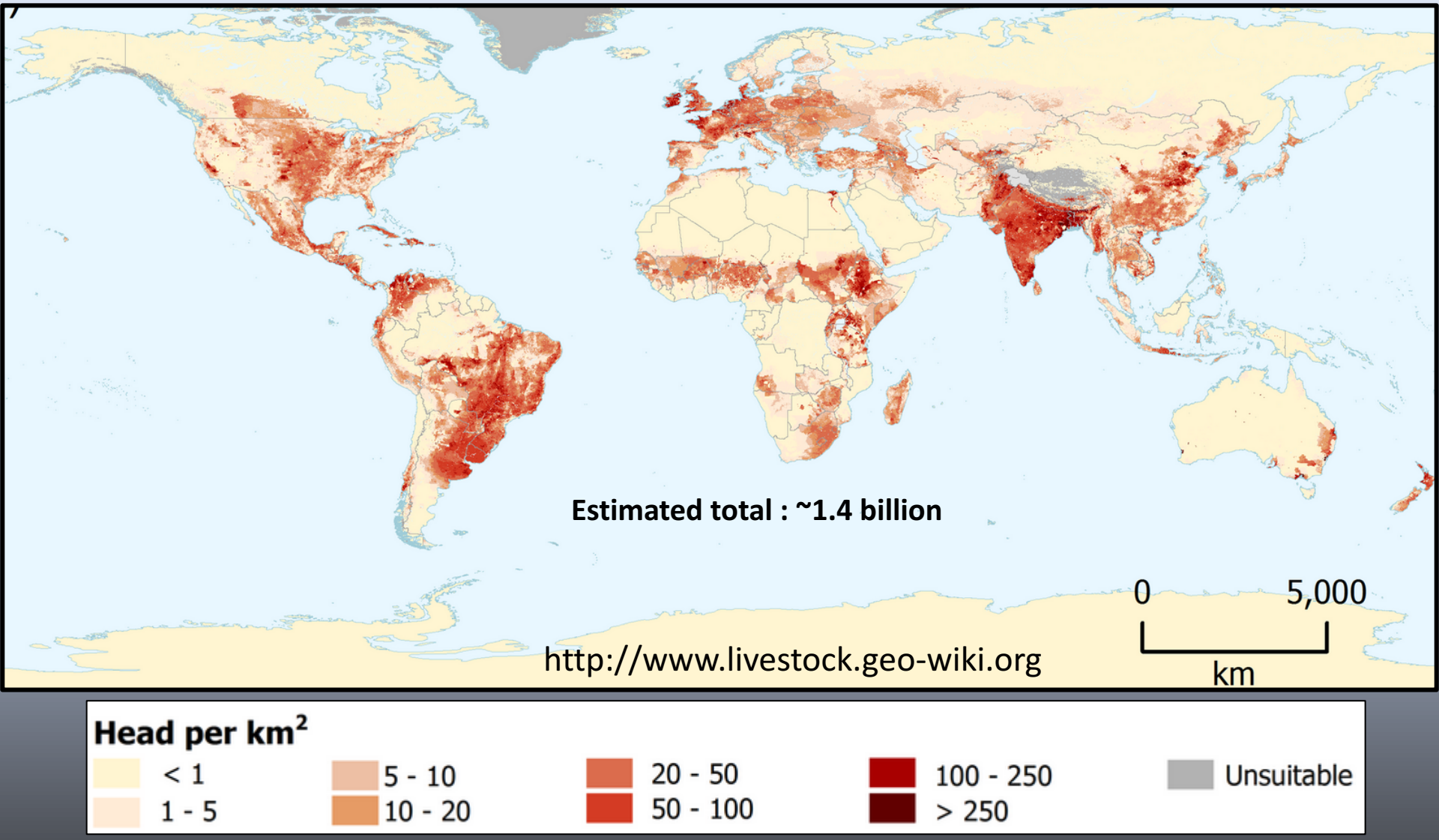
USDA
AGRICULTURAL RESEARCH SERVICE

USDA Agricultural Research Service
U.S. Meat Animal Research Center Clay Center, Nebraska



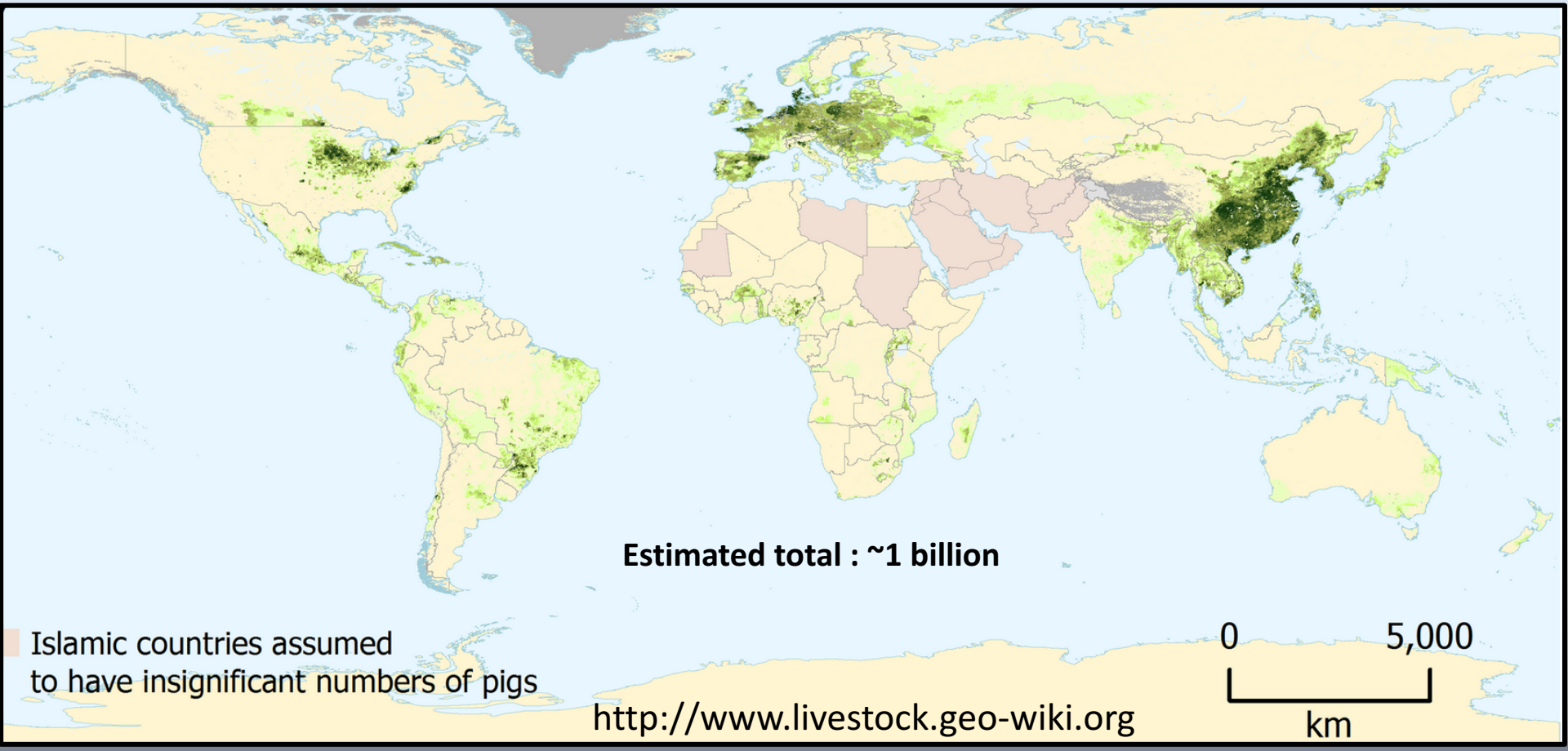
2000 breeding ewes

Mapping global cattle density (2014)



Mapping the global distribution of livestock.
Robinson et al., PLoS ONE 9:e96084. 2014.

Mapping global swine density (2014)



Head per km^2

< 1	5 - 10	20 - 50	100 - 250	Unsuitable
1 - 5	10 - 20	50 - 100	> 250	

Mapping the global distribution of livestock.
Robinson et al., PLoS ONE 9:e96084. 2014.

The evolution of reference assembly

The “human” genome

- The original reference human genome, and the current GRCh38p12, do not represent any existing real-world genome
  - Estimated individual haploid genome = 2.8-2.9 Gb
  - GRCh38p12 = 3.26 Gb

- The use of multiple individuals to provide the sequence data massively complicates the assembly process

- Adding sequence found in additional donors to move to a “pan-genome” reference assembly

Individual human genomes

- Genbank has (Jan. 10) eleven assemblies of individual humans (not cell lines)
  - 9 short-read assemblies, with 40-300 kb contig N50
  - 2 long-read assemblies, with 8.3 and 29 Mb contig N50

- Hundreds of thousands of unassembled “resequenced” genomes

- The points being :

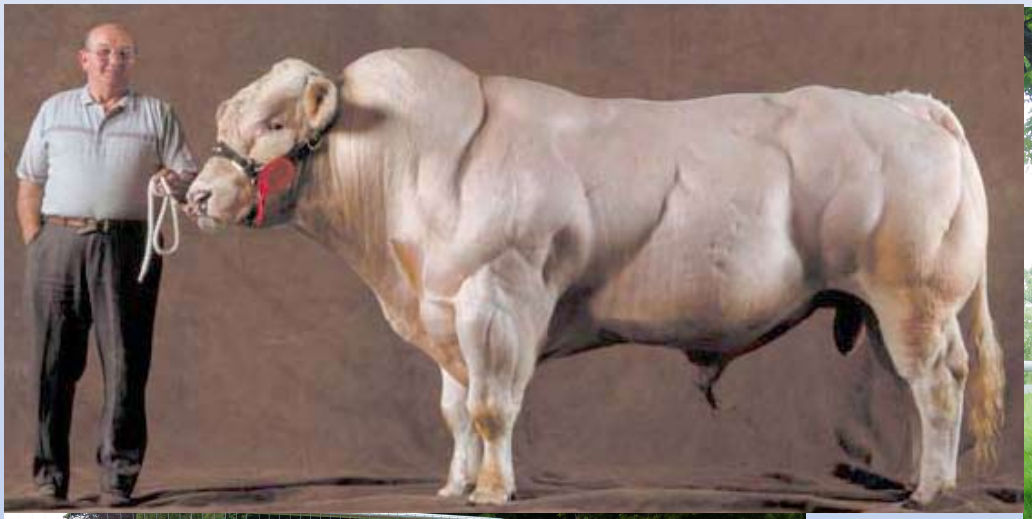
  1. no effort to reduce heterozygosity at any step

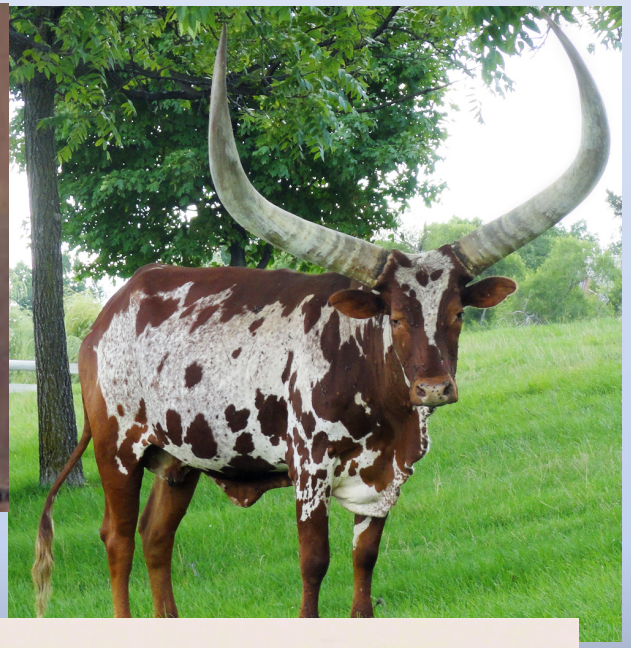
  2. no species have yet had multiple high quality reference genomes for comparisons

Range of human phenotypic variation

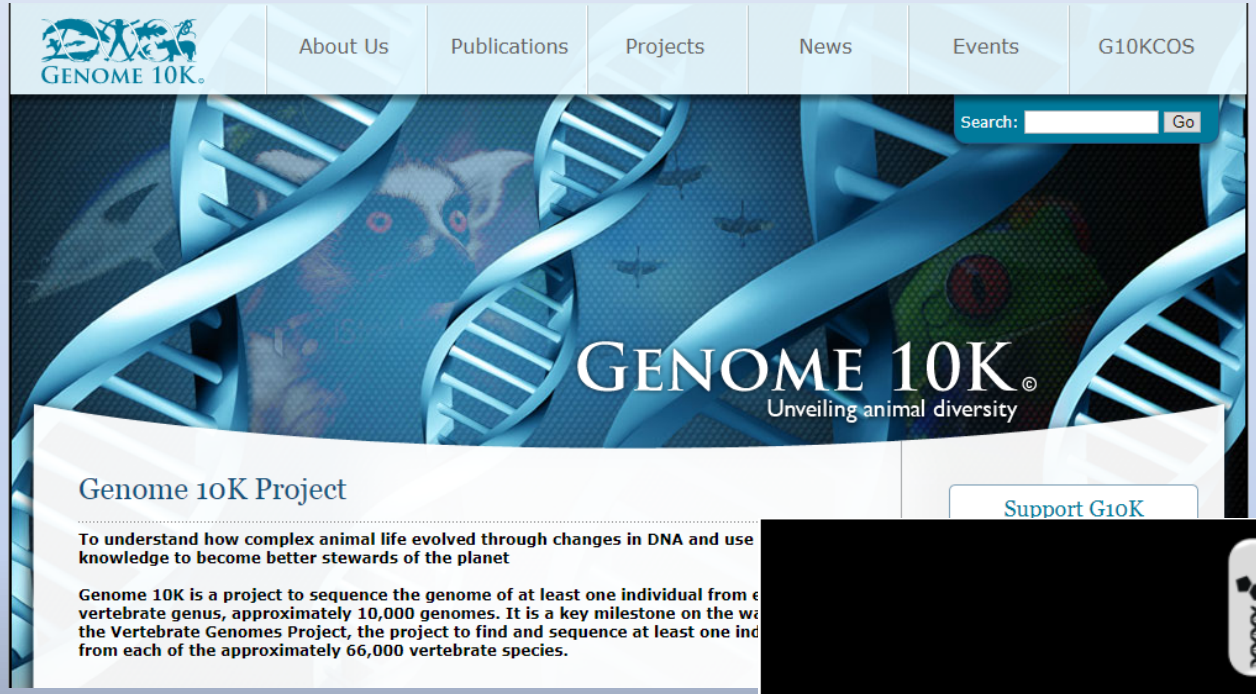
- No doubt there is phenotypic variation among human populations
  - How many high-quality genomes are optimal to inform the study of variation?

Phenotypic variation in animals





Non-human genomes



Non-human genomes

- For non-human species, inbred individuals favored to simplify assembly

- For some species, multiple individuals required to get sufficient DNA
  - e.g. lesser grain borer, mealworm, roundworm

Goat genome

- Selected animal from “stable inbred” line called San Clemente goats

The first livestock long-read assembly

Bickhart et al., *Nature Genetics* 49:643-50. April 2017

Approximately 500x improved continuity over the short read-based assembly

The goat reference assembly is GOAT

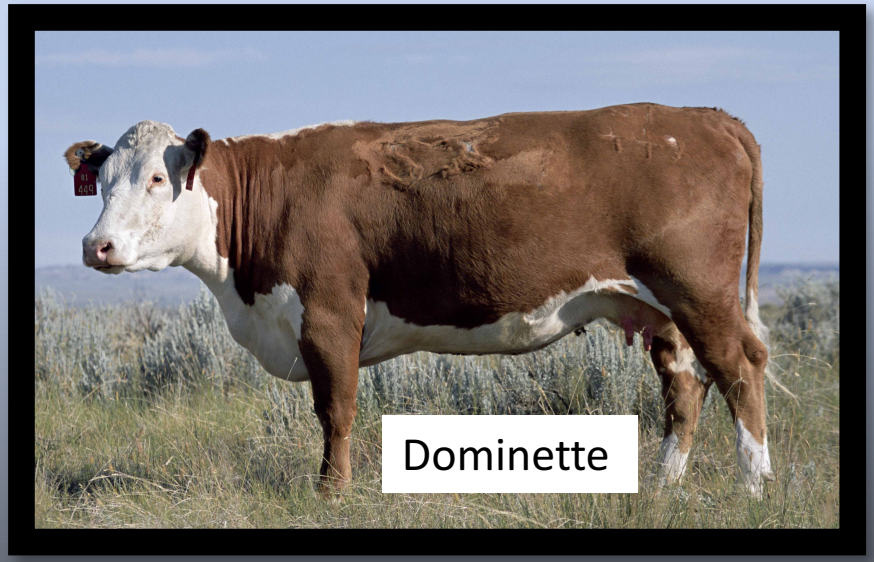
Assembly performed using predecessor to Canu

	Human	Mouse	Goat
Total sequence length (bp)	3,253,848,404	2,818,974,548	2,922,813,246
Total assembly gap length (bp)	161,368,351	79,435,572	38,187
Number of contigs	1,519	885	30,399
Contig N50 (bp)	56,413,054	32,273,079	26,244,591
Contig L50	19	26	32
Number of scaffolds	858	336	29,907
Scaffold N50 (bp)	59,364,414	52,589,046	87,277,232
Scaffold L50	17	18	13

Cattle reference genome

- >$50 million project by Baylor HGSC (ca. 2005)

- Animal selected to be the most documented homozygous available (genetic relationship of sire and daughter 93%)



Skip details of the short read assembly – we have now a long-read version

Cattle reference genome

- Long read assembly of Dominette going well – final polishing after gap filling

Description	Dominette
Total sequence length (bp)	2,715,862,177
Number of contigs	2628
Contig N50	25.9 Mb
Contig L50	32
Number of scaffolds	30
Scaffold N50	105 Mb
Scaffold L50	17

Additional cattle genome assemblies

- Cattle subspecies – *Bos taurus taurus* and *Bos taurus indicus*

Heat tolerant
Parasite resistant
Decreased meat quality
Lower “retail product yield”

Heat stress susceptible
Parasite susceptible
High meat quality
Higher “retail product yield”

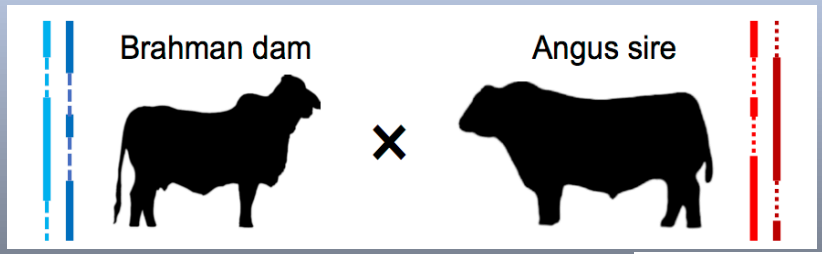
“2 for 1” cattle genomes

- Proposal : sequence an F1 offspring Angus x Brahman
  - Preliminary sequence data indicates one breed-specific base per 80-100 bp

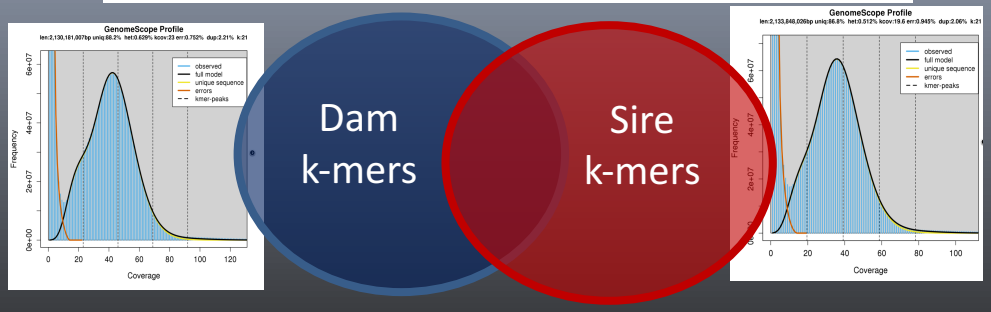
Strategy

- Male F1 fetus from Angus x Brahman (so Angus Y chromosome, Brahman X)
  - Generated 134x PacBio data (almost all Sequel) > 1kb subread (65x each haplotype)
  - Obtained 12x Hi-C coverage from Phase Genomics
  - Also collected 60x 2x150 PE short read sequence from each parent

Just add talent from NHGRI !

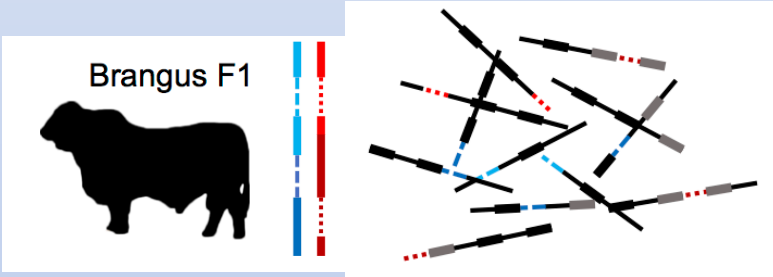


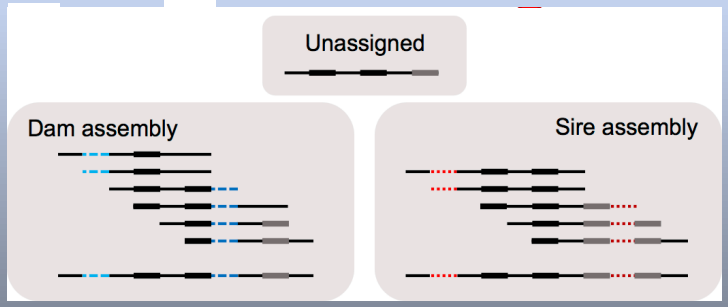
Courtesy: Arang Rhie



Strategy

- After separating the reads based on parent-specific k-mers, perform separate assembly for each haplotype (leaving out unassigned reads during contig formation)



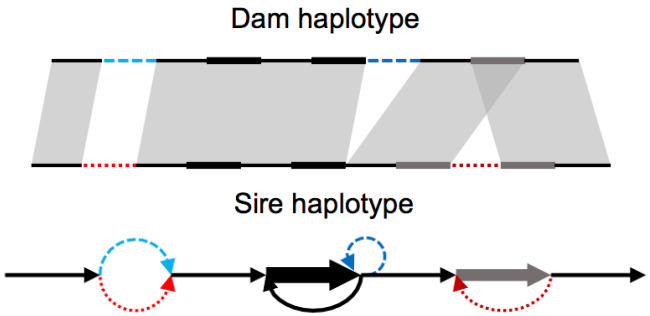


Courtesy: Arang Rhie

Dam (Brahman) haplotigs

Sire (Angus) haplotigs

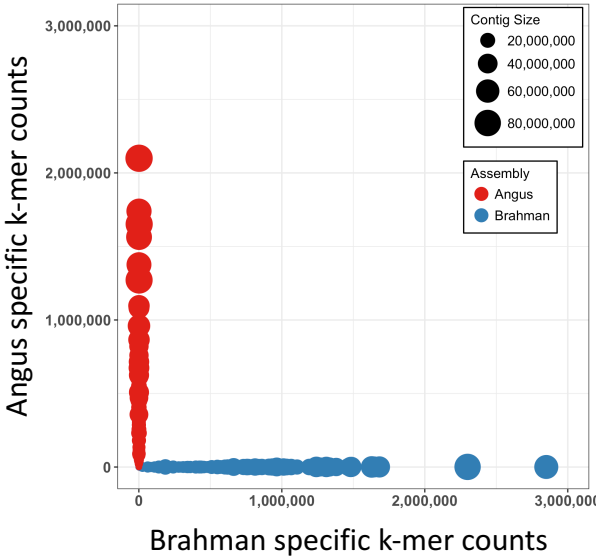
Comparison to FALCON-unzip



Courtesy : Arang Rhie

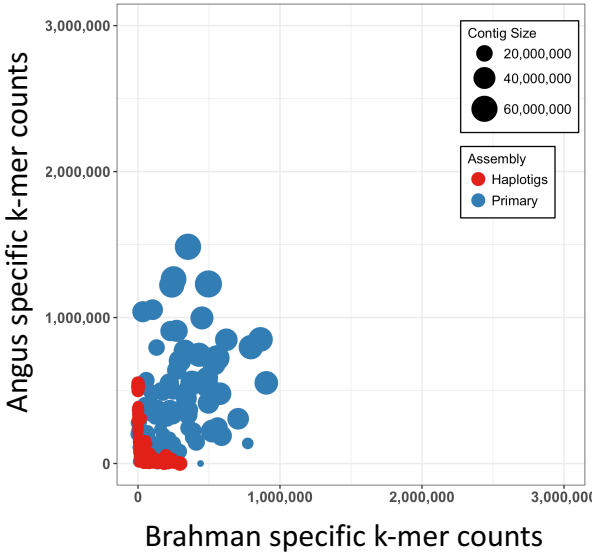
trio binning

Haplotigs = Contigs in each assembly
agree with parental haplotypes (Phased)



FALCON-unzip

Primary = Longest path in the graph (pseudo-hap)
Alternate haplotigs = Alternate path in the bubble



Result

- Still preliminary because the use of Hi-C data for the F1 for scaffolding still being worked out
  - Generally, each assembly represents a fully resolved haplotype of the fetus
  - Each assembly has contig N50 >20 Mb before any gap-filling steps
  - One Angus, one Brahman assembly

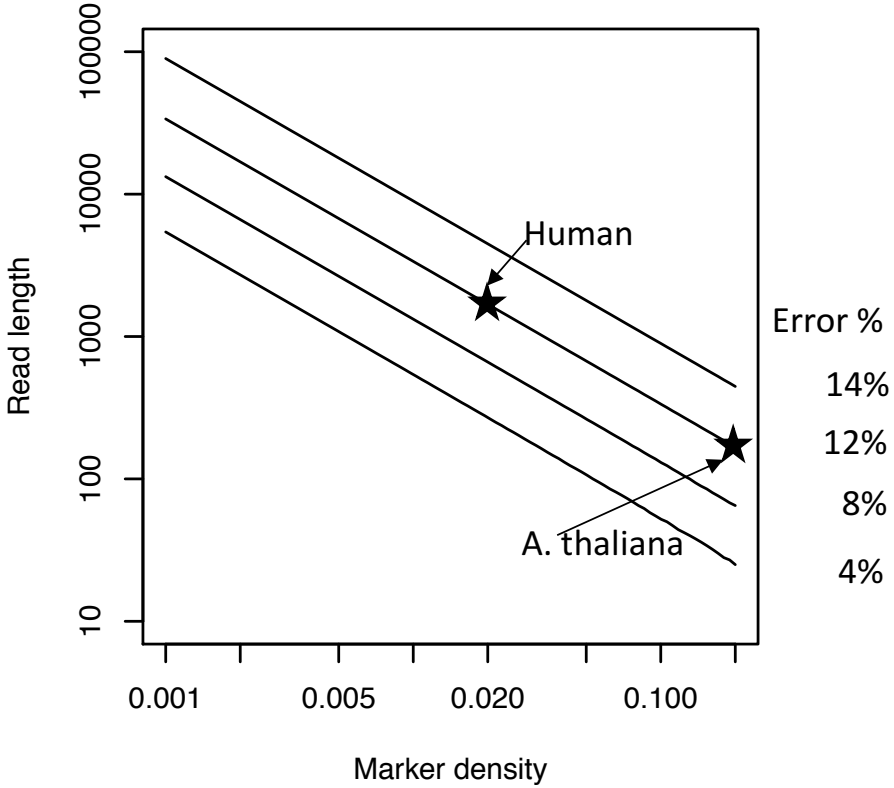
First haplotig N50 > 20Mb ever!!

Your mileage may vary

- Success depends on :
  - Degree of sequence variation between parental genomes
  - Read length
  - Sequence depth
  - Ploidy

Classification with sequencing error

- K-mers sensitive to SVs and SNPs
  - Each SNP == k k-mers
- Expect
  - 90% confidence reads ≥ 5 kbp have at least one k-mer
- Observe
  - 87.4% of all bases
    - avg read length 12 kbp
  - 90% of all bases >= 5kbp



k-mer size should be selected to balance need for unique k-mers in the genome (depends on genome size) and read error rate

Courtesy : Sergey Koren

Conclusion

- Out with homozygosity !! Grab all the heterozygosity you can find !!
  - Caveat : composites won't work quite as well even if highly heterozygous, because the parental haplotypes may not have unique k-mers everywhere

Meishan

White Composite

Conclusion

Maybe interspecies crosses ?

Liger

Mule

Yakalo

NHGRI
Arang Rhie
Sergey Koren
Brian Walenz
Alexander Dilthey
Brian Ondov
Adam Phillippy

ARS
Ben Rosen
Derek Bickhart
Warren Snelling

University of Adelaide
John Williams
Stefan Hiendleder
Cynthia Liu
Lloyd Low

University of Maryland
Aleksey Zimin
Jay Guhrye

University of Missouri
Bob Schnabel

Pacific Biosciences
Sarah Kingan
Marty Badgett

Phase Genomics
Ivan Liachko
Shawn Sullivan
Zev Kronenberg

Dovetail Genomics
Nicholas Putnam
Richard (Ed) Green

Computomix
Sebastian Schultheiss
Christian Dreisher