1000 bull genomes project

1000 bull genomes project consortium

Victoria
The Place To Be

DEPARTMENT OF
PRIMARY INDUSTRIES

biosciences
research

Outline

- Why do we need sequence data?
- The 1000 bull genomes project
- Results of test run 1 including quality control
- Using the output example : genome wide association studies

Why sequence data?

- The causative mutations are in the data set!

Why sequence data?

- The causative mutations are in the data set!

- Genome wide association studies
  - Straight to causative mutations?
  - Detect rare mutations (SNP chips biased to common SNP)

Why sequence data?

- The causative mutations are in the data set!

- Genome wide association studies
  - Straight to causative mutations?
  - Detect rare mutations (SNP chips biased to common SNP)
- Genomic prediction
  - No longer have to rely on LD with SNP
    - Higher accuracy of prediction (rare variants)?
    - Better persistence of accuracy across generations
  - Better prediction across breeds?
    - No longer need SNP-QTL associations holding across breeds

Why sequence data?

- The causative mutations are in the data set!

- Genome wide association studies
  - Straight to causative mutations?
  - Detect rare mutations (SNP chips biased to common SNP)
- Genomic prediction
  - No longer have to rely on LD with SNP
    - Higher accuracy of prediction (rare variants)?
    - Better persistence of accuracy across generations
  - Better prediction across breeds?
    - No longer need SNP-QTL associations holding across breeds
- Understanding biology

Outline

- Why do we need sequence data?

- **The 1000 bull genomes project**

- Results of test run 1 including quality control

- Using the output example : genome wide association studies

1000 Bull genomes project

- Sequencing still more expensive than SNP chip genotyping
- 100,000s of animals genotyped with SNP chips

- Alternative strategy
  - *Sequence key ancestors and impute genotypes from sequenced animals into all animals genotyped with SNP chips for GWAS, genomic prediction*
- Common need for reference genotype file from sequence

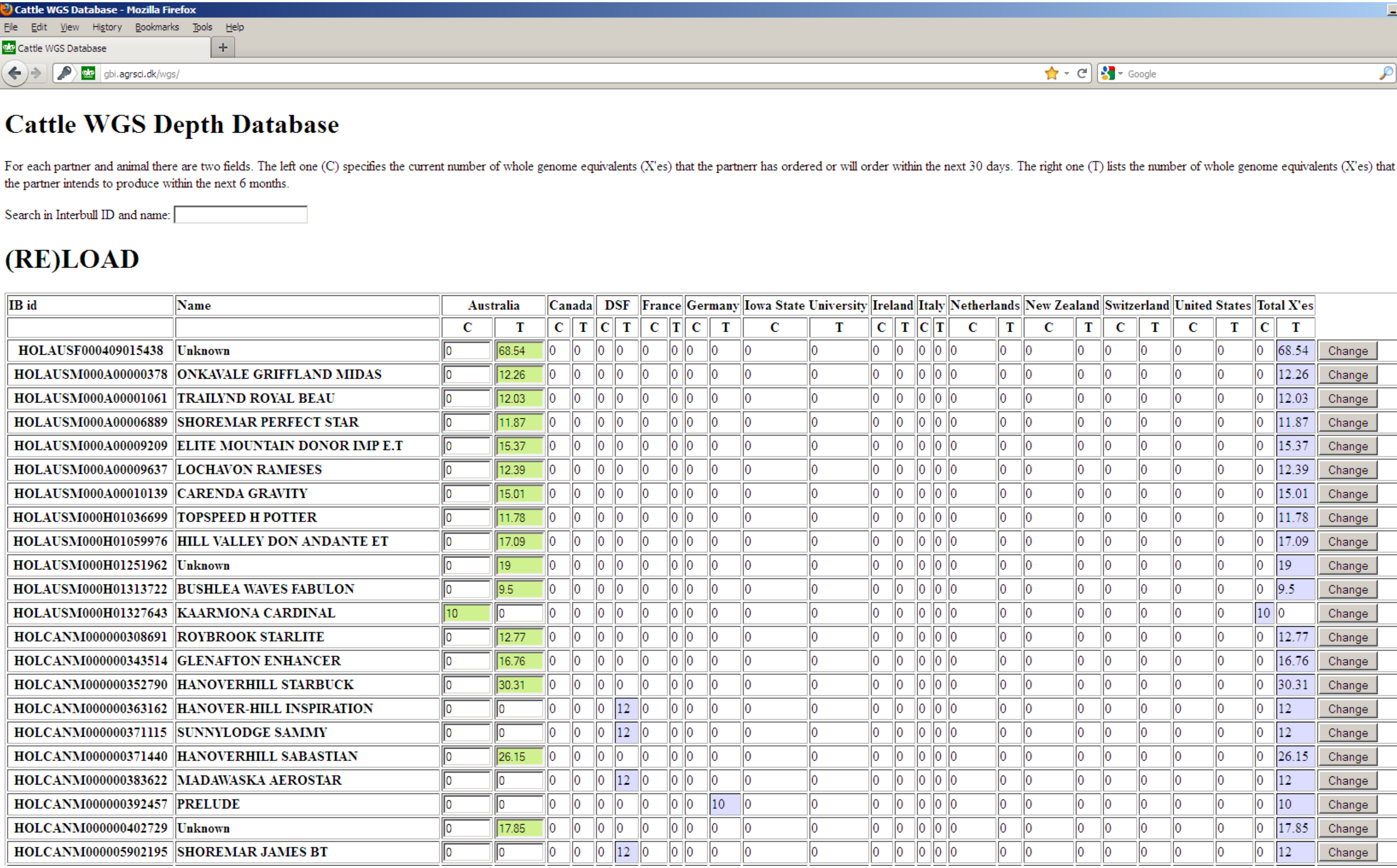
- **1000 bull genomes project**
  - ✓ Provide a database of genotypes from sequenced bulls
  - ✓ Global effort! – groups sequencing can get involved
  - ✓ Receive genotypes for all individuals sequenced

1000 Bull genomes project

- 151 bulls + 1 cow in database
    - ➢ Holstein, Fleckvieh, Jersey, Reds, Angus

- International ID to avoid duplication

- http://gbi.agrsci.dk/wgs/



Cattle WGS Depth Database

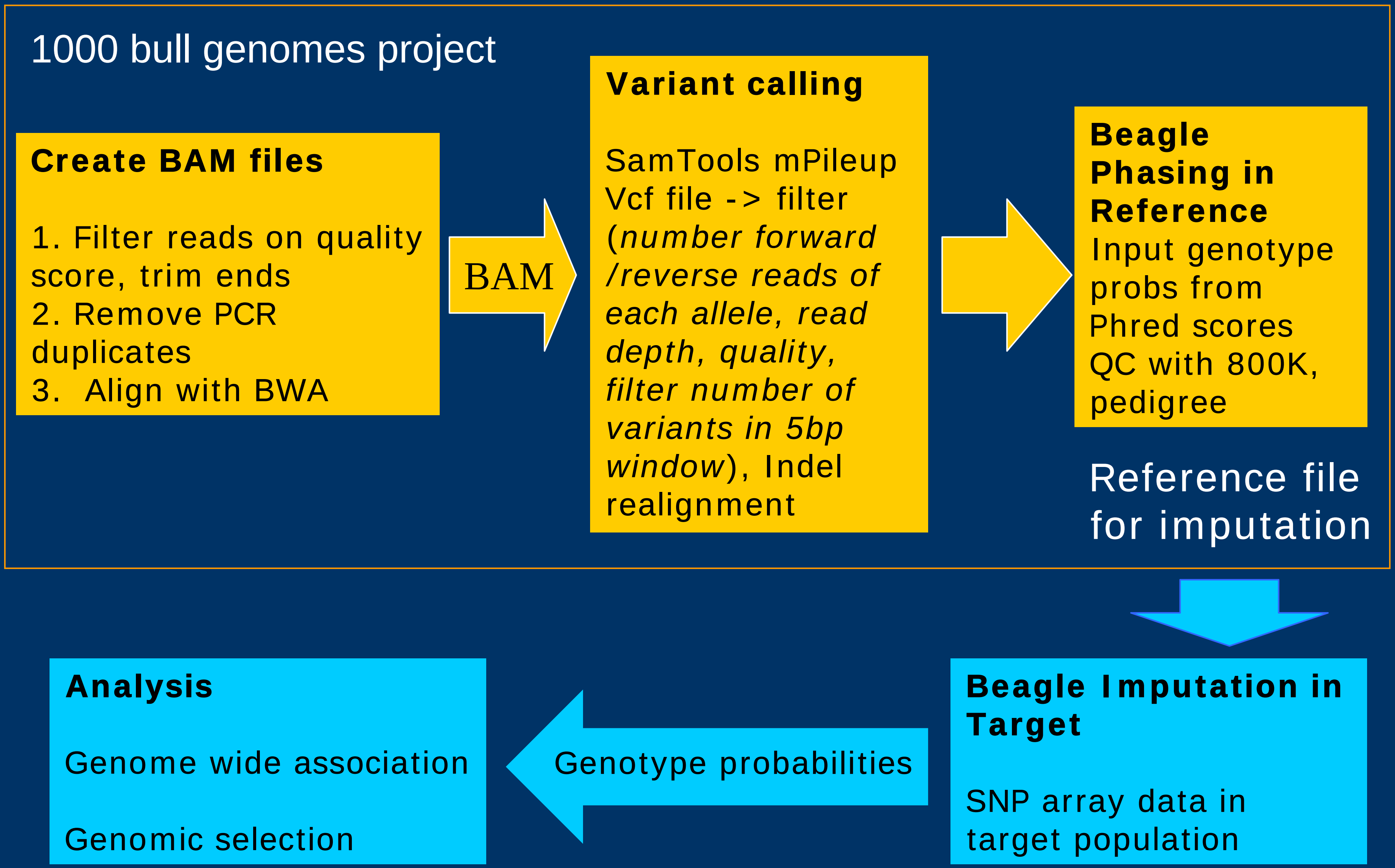
For each partner and animal there are two fields. The left one (C) specifies the current number of whole genome equivalents (X'es) that the partnerr has ordered or will order within the next 30 days. The right one (T) lists the number of whole genome equivalents (X'es) that the partner intends to produce within the next 6 months.

Search in Interbull ID and name:

(RE)LOAD

IB id	Name	Australia		Canada		DSF		France		Germany		Iowa State University		Ireland		Italy		Netherlands		New Zealand		Switzerland		United States		Total X'es		
		C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	
HOLAUSF000409015438	Unknown	0	68.54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	68.54	Change
HOLAUSM000A00000378	ONKAVALE GRIFFLAND MIDAS	0	12.26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12.26	Change
HOLAUSM000A00001061	TRAILYND ROYAL BEAU	0	12.03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12.03	Change
HOLAUSM000A00006889	SHOREMAR PERFECT STAR	0	11.87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11.87	Change
HOLAUSM000A00009209	ELITE MOUNTAIN DONOR IMP E.T	0	15.37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15.37	Change
HOLAUSM000A00009637	LOCHAVON RAMESES	0	12.39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12.39	Change
HOLAUSM000A00010139	CARENDA GRAVITY	0	15.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15.01	Change
HOLAUSM000H01036699	TOPSPEED H POTTER	0	11.78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11.78	Change
HOLAUSM000H01059976	HILL VALLEY DON ANDANTE ET	0	17.09	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17.09	Change
HOLAUSM000H01251962	Unknown	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	Change
HOLAUSM000H01313722	BUSHLEA WAVES FABULON	0	9.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9.5	Change
HOLAUSM000H01327643	KAARMONA CARDINAL	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	Change
HOLCANM000000308691	ROYBROOK STARLITE	0	12.77	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12.77	Change
HOLCANM000000343514	GLENAFTON ENHANCER	0	16.76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16.76	Change
HOLCANM000000352790	HANOVERHILL STARBUCK	0	30.31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30.31	Change
HOLCANM000000363162	HANOVER-HILL INSPIRATION	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	Change
HOLCANM000000371115	SUNNYLODGE SAMMY	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	Change
HOLCANM000000371440	HANOVERHILL SABASTIAN	0	26.15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26.15	Change
HOLCANM000000383622	MADAWASKA AEROSTAR	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	Change
HOLCANM000000392457	PRELUDE	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	Change
HOLCANM000000402729	Unknown	0	17.85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17.85	Change
HOLCANM000005902195	SHOREMAR JAMES BT	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	Change

Imputation of full sequence data



1000 bull genomes project

Create BAM files

1. Filter reads on quality score, trim ends
2. Remove PCR duplicates
3. Align with BWA

BAM

Variant calling

SamTools mPileup
Vcf file -> filter (*number forward /reverse reads of each allele, read depth, quality, filter number of variants in 5bp window*), Indel realignment

Beagle Phasing in Reference
Input genotype probs from Phred scores
QC with 800K, pedigree

Reference file for imputation

Beagle Imputation in Target

SNP array data in target population

Genotype probabilities

Analysis

Genome wide association

Genomic selection

Outline

- Why do we need sequence data?
- The 1000 bull genomes project
- **Results of test run 1 including quality control**
- Using the output example : genome wide association studies

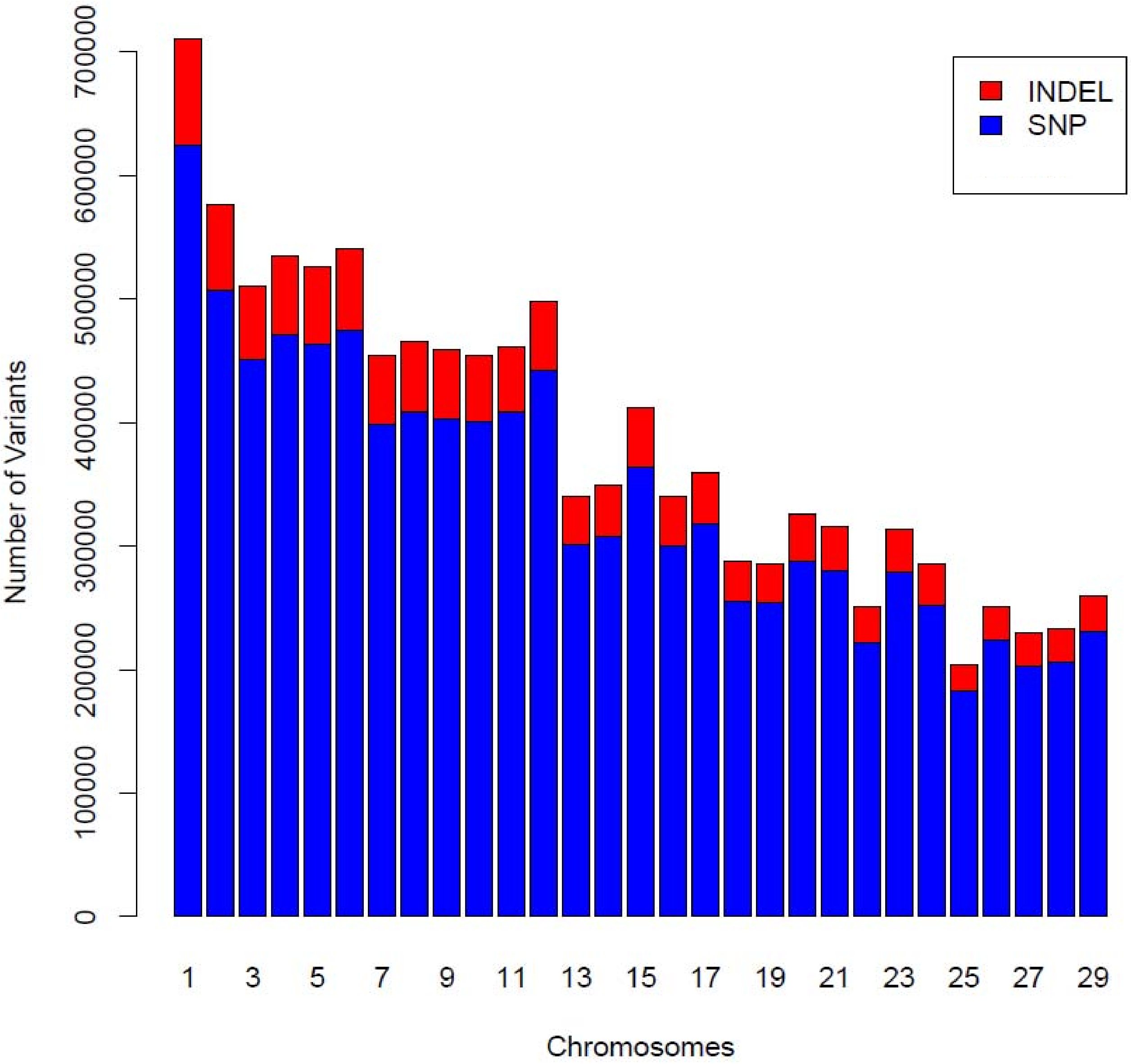
Results of test run 1

- Bull set

International ID	Name	Fold coverage
HOLCANM000000308691	Starlite	12.8
HOLAUSM000A00006889	Shotime	11.9
HOLAUSM000H01036699	Goldsmith	11.8
HOLAUSM000A00010139	Gravita	15
HOLAUSM000H01313722	Orana	9.5
HOLAUSM000A00001061	Beau	12
HOLAUSM000A00000378	OVGM	12.3
HOLCANM000010705608	Goldwyn	22.7
HOLCANM000000352790	Starbuck	30.3
HOLAUSM000A00009637	Rameses	12.4
HOLAUSM000A00009209	Donor	15.4
HOLAUSM000H01059976	Donante	17.1
HOLUSAM000002070579	Mountain	18.9
HOLCANM000000343514	Enhancer	16.8
HOLAUSM000H01251962	Yukon	19
HOLFRAM002991000305	Gibbon	17
HOLFRAM005694028588	Jocko	15.1
HOLUSAM000122358313	Oman	14.7
HOLCANM000000402729	Manhattan	17.9
HOLFRAM002290038601	Fatal	16.9
HOLNLDM000775328514	Cash	16.8
HOLNLDM000829877874	Boudewijn	18.5
HOLCANM000000371440	Sabastian	26.2
HOLUSAM000002005253	Vickai	15.2

Results of test run 1

- 11.23 million filtered variants
- 9.92 million SNP, 1.31 million INDEL detected



Results of test run 1

- Agreement with 800K

Bull	Pre-Beagle	After-Beagle	Difference
HOLAUSM000A00000378	0.988	0.993	0.004
HOLAUSM000A00009209	0.994	0.995	0.002
HOLAUSM000A00010139	0.992	0.995	0.004
HOLAUSM000H01059976	0.973	0.985	0.012
HOLAUSM000H01251962	0.994	0.995	0.002
HOLAUSM000H01313722	0.989	0.995	0.006
HOLCANM000000308691	0.991	0.995	0.004
HOLCANM000000343514	0.994	0.995	0.002
HOLCANM000000352790	0.996	0.997	0.001
HOLCANM000010705608	0.993	0.995	0.002
HOLNLDM000829877874	0.987	0.993	0.006
HOLUSAM000002070579	0.996	0.997	0.001
HOLUSAM000122358313	0.992	0.996	0.003
Mean	0.991	0.994	0.004

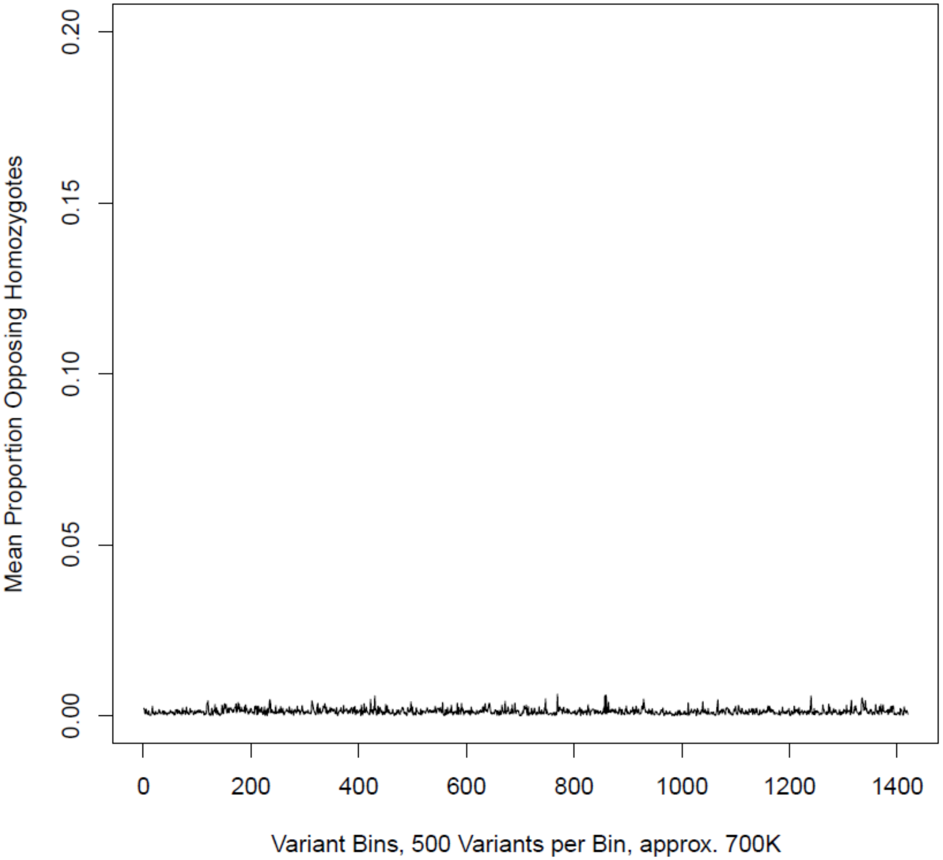
Results of test run 1

- Quality control – opposing homozygotes
  - If sire AA, son must be AA or AT, else if TT genotype calling error! (or denovo mutation….)
  - In data set, 6 sire son pairs
  - How many opposing homozygotes (eg sire -= AA and son = TT?) in windows across genome?

Results of test run 1

- Quality control – opposing homozygotes

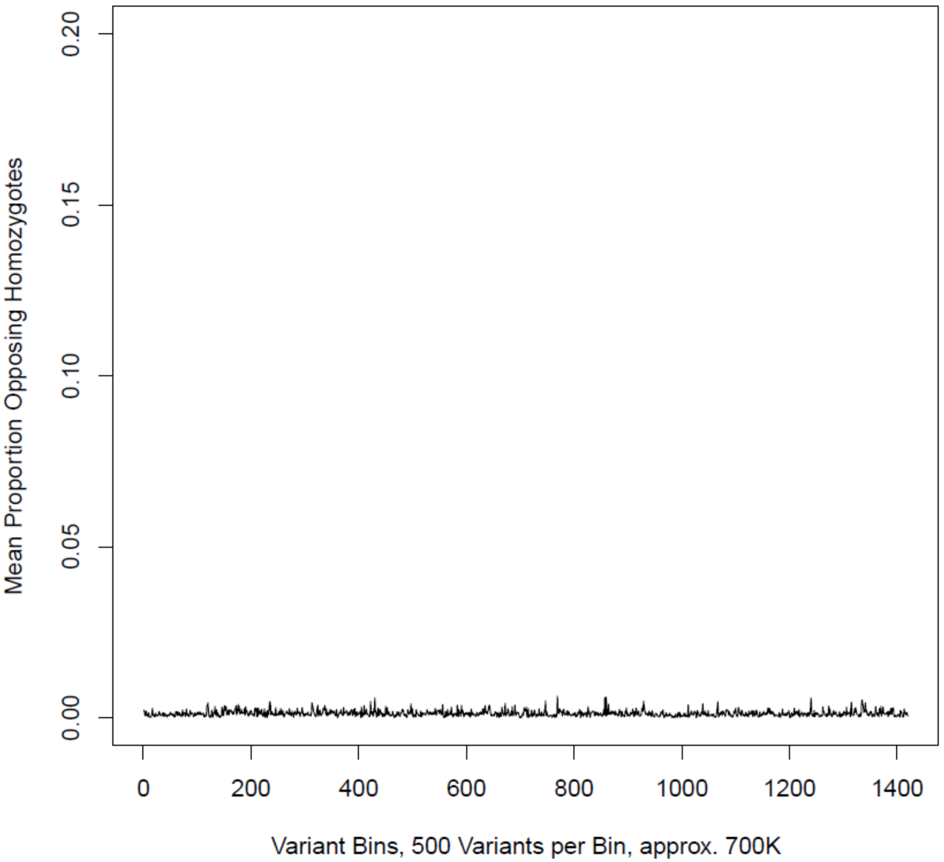
Chromosome 1



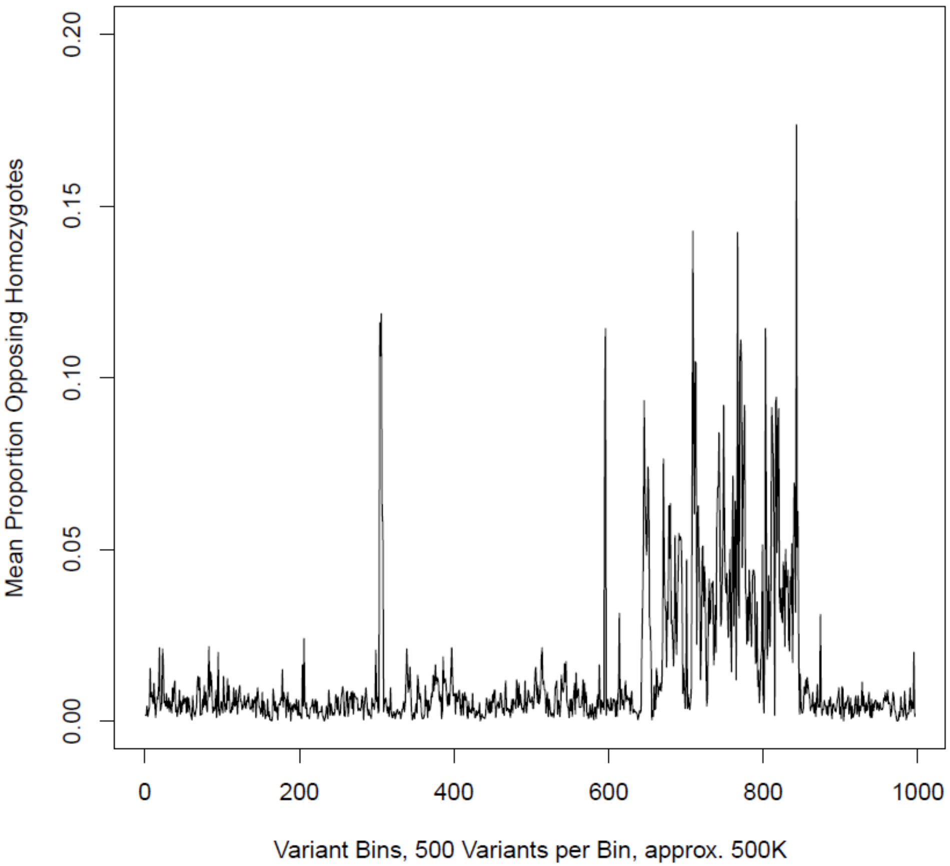
Results of test run 1

- Quality control – opposing homozygotes

Chromosome 1



Chromosome 12

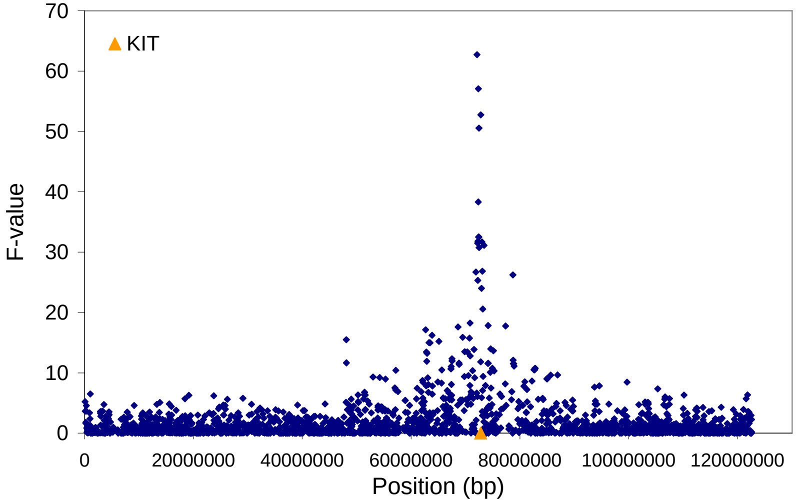


Outline

- Why do we need sequence data?
- The 1000 bull genomes project
- Results of test run 1 including quality control
- **Using the output example : genome wide association studies**

Using imputed full sequence

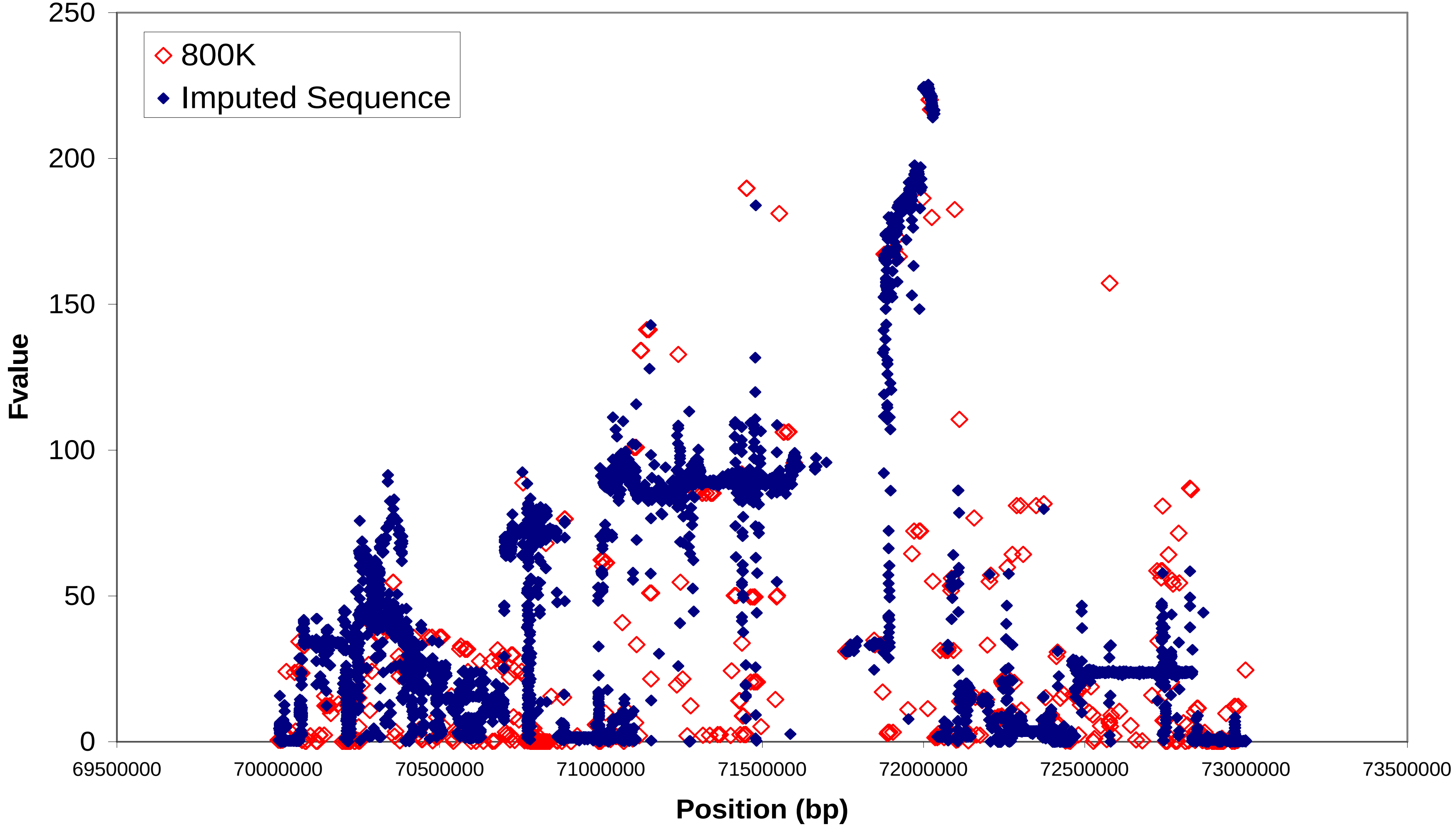
- KIT example
  - — Earlier genome wide association study for proportion of black in Holsteins found association with SNP in KIT locus



  - — Can we impute sequence in this region and re-run association study?

Using imputed full sequence

- KIT example

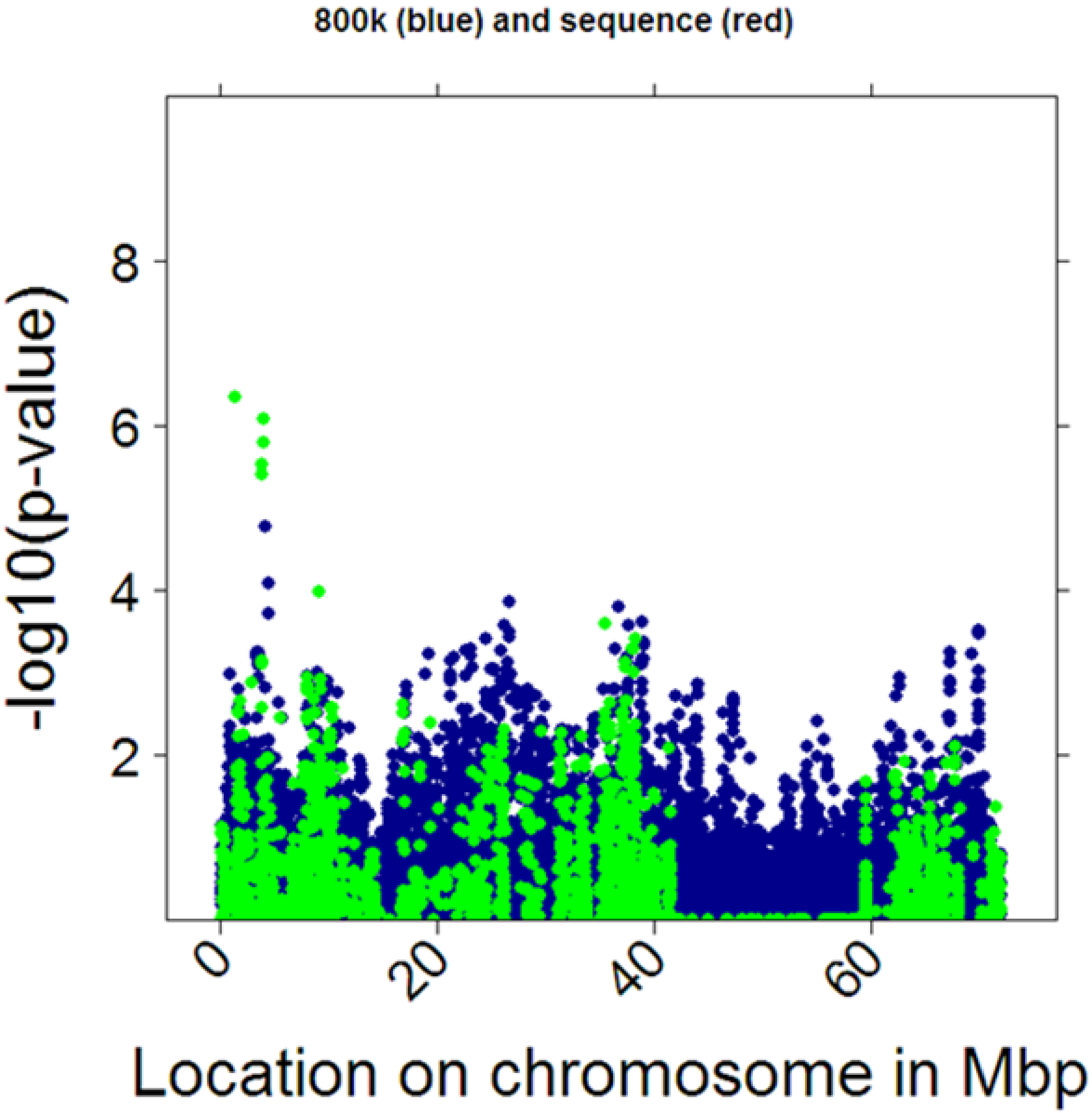


Using imputed full sequence

- Feed conversion efficiency example
  - 848 Holstein heifers with 800K genotypes and feed conversion efficiency phenotypes
  - Genome wide association study with 800K vs Imputed sequence

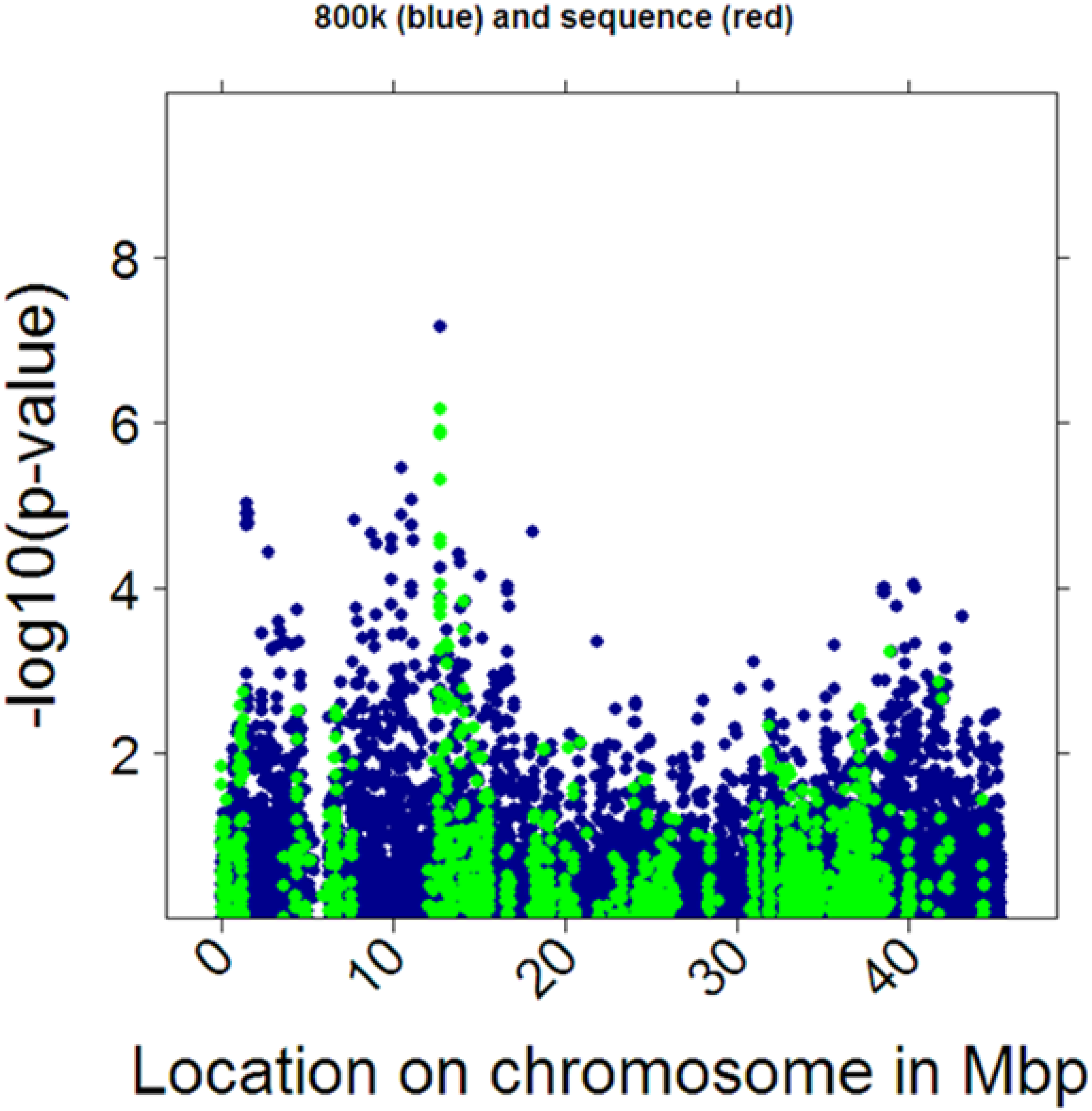
Using imputed full sequence

- Feed conversion efficiency example
- Chr 20



Using imputed full sequence

- Feed conversion efficiency example
- Chr 27



Conclusions

- 1000 bull genomes project underway
  - 151 bulls + 1 cow in data base

- Trial run of pipeline
  - Large numbers of SNP/Indel called
  - Excellent agreement with 800K genotypes
  - Low rate of opposing homozygotes for sire son pairs

- When sequence genotypes used as reference set for imputation
  - SNP detected with higher F-values than original 800K, in some cases
  - Need more bulls!

- Next run in February
- Working groups on variant detection/sequence annotation

- **http://1000bullgenomes.com**

1000 bull genomes project

The 1000 bull genomes project aims to provide, for the bovine research community, a large database for imputation of genetic variants for genomic prediction and genome wide association studies in cattle. The project aims to develop a resource to allow project partners to impute full genome sequence in bulls and cows that have been genotyped with SNP arrays. This could be for example for the purposes of genomic prediction, genome wide association, and discovery of causal mutations.

A database of bulls and cows that have been sequenced can be found here:
http://gbi.agrsci.dk/wgs/

The standard reference genome for the project can be downloaded here:
http://stothard.afns.ualberta.ca/1000_bull_genomes/reference_for_mapping/umd_3_1_reference_1000_bull_genomes.fa.gz

or if you are in Europe http://gbi.agrsci.dk/wgs/umd_3_1_reference_1000_bull_genomes.fa.xz

Sequence alignment guidelines to create BAM files are here: Sequence Alignment Guidelines for producing bam files for the 1000 bull genomes project

The project agreement for new partners, including the list of existing partners is here: 1000 Bull Genomes Project Agreement

And example output files are found here: bovine_variants.txt bovine_dose.txt

With thanks

- Workers
  - Charlotte Anderson, Hans Daetwyler, David Coote, Jennie Pryce
- Steering committee
  - Ruedi Fries (Technische Universität München, Germany)
  - Mogens Lund/Bernt Guldbrandtsent (Aarhus University, Denmark)
  - Didier Boichard (INRA, France)
  - Paul Stothard (University of Alberta, Canada)
  - Roel Veerkamp (Wageningen UR, Netherlands)
  - Ben Hayes/Mike Goddard (DFL)
  - Curt Van Tassell (United States Department of Agriculture)
- Partners
  - Ina Hulsegge , Wageningen UR Livestock Research, Dominique Rocha , INRA, Dirk Hinirichs , Christian-Albrechts-University, D-24098 Kiel, Germany, Alessandro Bagnato , Università degli Studi di Milano, Milano, Italy, Michel Georges/Tom Druet , University of Liege, Richard Spelman , Livestock Improvement Corporation, James Reecy , Iowa State University, Ames, IA, Alan L. Archibald , Roslin Institute, Birgit Gredler , Qualitas AG, Switzerland, Donagh Berry TEAGASC, Sigbjorn Lien, UMB