1000 bull genomes project

1000 bull genomes project consortium
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 partner 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.

#### (RE)LOAD

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Australia</th>
<th>Canada</th>
<th>DSF</th>
<th>France</th>
<th>Germany</th>
<th>Iowa State University</th>
<th>Ireland</th>
<th>Italy</th>
<th>Netherlands</th>
<th>New Zealand</th>
<th>Switzerland</th>
<th>United States</th>
<th>Total X’es</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOLAUSF000409015438</td>
<td>Unknown</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>65.54</td>
</tr>
<tr>
<td>HOLAUSM000A00000378</td>
<td>ONKAVELE GRIFFLAND MIDAS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12.26</td>
</tr>
<tr>
<td>HOLAUSM000A000001661</td>
<td>TRAILYN ROYAL BEAU</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12.03</td>
</tr>
<tr>
<td>HOLAUSM000A000006889</td>
<td>SHOREMAR PERFECT STAR</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11.87</td>
</tr>
<tr>
<td>HOLAUSM000A000009209</td>
<td>ELITE MOUNTAIN DONOR IMP ET</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15.37</td>
</tr>
<tr>
<td>HOLAUSM000A000009637</td>
<td>LOCHAVON RAMESSES</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12.39</td>
</tr>
<tr>
<td>HOLAUSM000A00010139</td>
<td>CARENDA GRAVITY</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15.01</td>
</tr>
<tr>
<td>HOLAUSM000H0156699</td>
<td>TOPSPEED H POTTER</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11.78</td>
</tr>
<tr>
<td>HOLAUSM000H01059976</td>
<td>HILL VALLEY DON ANDANTE ET</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17.09</td>
</tr>
<tr>
<td>HOLAUSM000H01251962</td>
<td>Unknown</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>HOLAUSM000H01331722</td>
<td>BUSHLEA WAVES FABULON</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9.5</td>
</tr>
<tr>
<td>HOLAUSM000H01337643</td>
<td>KAARMONA CARDINAL</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>HOLCAAN000000308691</td>
<td>ROYBROOK STARLITE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12.77</td>
</tr>
<tr>
<td>HOLCAAN000000344554</td>
<td>GLENAFTON ENHANCER</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16.76</td>
</tr>
<tr>
<td>HOLCAAN000000355790</td>
<td>HANOVERHILL STARBUCK</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30.31</td>
</tr>
<tr>
<td>HOLCAAN000000363162</td>
<td>HANOVER-HILL INSPIRATION</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>HOLCAAN000000371115</td>
<td>SUNNYLodge SAMMY</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>HOLCAAN000000371440</td>
<td>HANOVERHILL SABASTIAN</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>26.15</td>
</tr>
<tr>
<td>HOLCAAN000000383632</td>
<td>MADAWASKA AEROSTAR</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>HOLCAAN000000392457</td>
<td>PRELUDE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>HOLCAAN000000402729</td>
<td>Unknown</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17.85</td>
</tr>
<tr>
<td>HOLCAAN0000005902195</td>
<td>SHOREMAR JAMES BT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>32</td>
</tr>
</tbody>
</table>
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

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

Analysis
Genome wide association
Genomic selection

Genotype probabilities

Beagle Imputation in Target
SNP array data in target population
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

<table>
<thead>
<tr>
<th>International ID</th>
<th>Name</th>
<th>Fold coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOLCANM0000000308691</td>
<td>Starlite</td>
<td>12.8</td>
</tr>
<tr>
<td>HOLAUSM000A00006889</td>
<td>Shotime</td>
<td>11.9</td>
</tr>
<tr>
<td>HOLAUSM000H01036699</td>
<td>Goldsmith</td>
<td>11.8</td>
</tr>
<tr>
<td>HOLAUSM000A00010139</td>
<td>Gravita</td>
<td>15</td>
</tr>
<tr>
<td>HOLAUSM000H01313722</td>
<td>Orana</td>
<td>9.5</td>
</tr>
<tr>
<td>HOLAUSM000A00001061</td>
<td>Beau</td>
<td>12</td>
</tr>
<tr>
<td>HOLAUSM000A0000378</td>
<td>OVGM</td>
<td>12.3</td>
</tr>
<tr>
<td>HOLAUSM00010705608</td>
<td>Goldwyn</td>
<td>22.7</td>
</tr>
<tr>
<td>HOLAUSM00000352790</td>
<td>Starbuck</td>
<td>30.3</td>
</tr>
<tr>
<td>HOLAUSM000A0009637</td>
<td>Rameses</td>
<td>12.4</td>
</tr>
<tr>
<td>HOLAUSM000A0009209</td>
<td>Donor</td>
<td>15.4</td>
</tr>
<tr>
<td>HOLAUSM000H01059976</td>
<td>Donante</td>
<td>17.1</td>
</tr>
<tr>
<td>HOLUSAM00002070579</td>
<td>Mountain</td>
<td>18.9</td>
</tr>
<tr>
<td>HOLAUSM0000343514</td>
<td>Enhancer</td>
<td>16.8</td>
</tr>
<tr>
<td>HOLAUSM000H01251962</td>
<td>Yukon</td>
<td>19</td>
</tr>
<tr>
<td>HOLFRAM002991000305</td>
<td>Gibbon</td>
<td>17</td>
</tr>
<tr>
<td>HOLFRAM005694028588</td>
<td>Jocko</td>
<td>15.1</td>
</tr>
<tr>
<td>HOLUSAM000122358313</td>
<td>Oman</td>
<td>14.7</td>
</tr>
<tr>
<td>HOLAUSM00000402729</td>
<td>Manhattan</td>
<td>17.9</td>
</tr>
<tr>
<td>HOLFRA00229008601</td>
<td>Fatal</td>
<td>16.9</td>
</tr>
<tr>
<td>HOLNLDM000775328514</td>
<td>Cash</td>
<td>16.8</td>
</tr>
<tr>
<td>HOLNLDM000829877874</td>
<td>Boudewijn</td>
<td>18.5</td>
</tr>
<tr>
<td>HOLCAN000000371440</td>
<td>Sebastian</td>
<td>26.2</td>
</tr>
<tr>
<td>HOLUSAM00002005253</td>
<td>Vickai</td>
<td>15.2</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Bull</th>
<th>Pre-Beagle</th>
<th>After-Beagle</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOLAUSM000A00000378</td>
<td>0.988</td>
<td>0.993</td>
<td>0.004</td>
</tr>
<tr>
<td>HOLAUSM000A00009209</td>
<td>0.994</td>
<td>0.995</td>
<td>0.002</td>
</tr>
<tr>
<td>HOLAUSM000A00010139</td>
<td>0.992</td>
<td>0.995</td>
<td>0.004</td>
</tr>
<tr>
<td>HOLAUSM000H01059976</td>
<td>0.973</td>
<td>0.985</td>
<td>0.012</td>
</tr>
<tr>
<td>HOLAUSM000H01251962</td>
<td>0.994</td>
<td>0.995</td>
<td>0.002</td>
</tr>
<tr>
<td>HOLAUSM000H01313722</td>
<td>0.989</td>
<td>0.995</td>
<td>0.006</td>
</tr>
<tr>
<td>HOLCANM000000308691</td>
<td>0.991</td>
<td>0.995</td>
<td>0.004</td>
</tr>
<tr>
<td>HOLCANM000000343514</td>
<td>0.994</td>
<td>0.995</td>
<td>0.002</td>
</tr>
<tr>
<td>HOLCANM000000352790</td>
<td>0.996</td>
<td>0.997</td>
<td>0.001</td>
</tr>
<tr>
<td>HOLCANM000010705608</td>
<td>0.993</td>
<td>0.995</td>
<td>0.002</td>
</tr>
<tr>
<td>HOLNLDM000829877874</td>
<td>0.987</td>
<td>0.993</td>
<td>0.006</td>
</tr>
<tr>
<td>HOLUSAM000002070579</td>
<td>0.996</td>
<td>0.997</td>
<td>0.001</td>
</tr>
<tr>
<td>HOLUSAM000122358313</td>
<td>0.992</td>
<td>0.996</td>
<td>0.003</td>
</tr>
<tr>
<td>Mean</td>
<td>0.991</td>
<td>0.994</td>
<td>0.004</td>
</tr>
</tbody>
</table>
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
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
Conclusions

- 1000 bull genomes project underway
  - 151 bulls + 1 cow in database
- Trial run of pipeline
  - Large numbers of SNP/Indel called
  - Excellent agreement with 800K genotypes
  - Low rate of opposing homozygotes for sire son pairs
  - Need more bulls!
- When sequence genotypes used as reference set for imputation
  - SNP detected with higher F-values than original 800K, in some cases
- Next run in February
- Working groups on variant detection/sequence annotation
- [1000bullgenomes.com](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/](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](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](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