ISGC call October 24th 2019

Attendees: Shannon Clarke, Rudiger Brauning, Kathryn McRae, John McEwan, Brenda Murdoch, Michelle Mousel, Tim Smith, Ben Rosen, Gwenola Tosser, Emily Clark

Agenda,

1. ISGC meeting at PAG
2. Sheep Genome DB run 3 update (Rudiger Brauning)
3. Update from the Ovine FAANG project (Brenda Murdoch)
4. Other business

Meeting notes:

1. ISGC meeting at PAG

   Confirmed 4 speakers with aim to have ~8 15min presentations. Shannon to forward any goat or sheep abstracts to Ben Rosin for consideration in the sheep/cattle/goat session

2. Sheep Genome DB update (Rudi Brauning):
   - **Run 2**
     - 935 animals
     - utilised GATK and Samtools with OAR v3.1 genome assembly
     - although this is completed with results at EVA (https://www.ebi.ac.uk/eva/?eva-study=PRJEB31241), Run 2 does not currently appear in variant browser.
     - There has been an issue by EVA re rsIDs and file uploads. This is not yet fixed.
     - Summary for the SGD IDs:
       - For run1 and run2 all animals are expected to have a SGD ID, which is a concatenation of
         - 2-letter country code
         - 3-letter breed code, http://sheepgenomesdb.org/files/SheepGenomesDB_SheepBreedCodes.xlsx
         - 1-letter sex code
       - You can find these IDs reported by EVA, e.g.

High quality variant calls from the Sheep genomes project - Run1 (PRJEB14685)

Sample	Genotype
AUDOPU0000000000454	0/0
AUMEHM0000000000001	0/0
AUMEHM0000000000002	0/0
AUMEHM0000000000003	0/0
AUMEHM0000000000004	0/0
AUMEPM0000000000005	0/0
AUMEPM0000000000006	0/1
AUMEPM0000000000007	0/0
AUMEPM0000000000008	0/1
AUMEPM0000000000009	0/0

       - Summary of breed and country codes for run2.

	Australia (AU)	Bangladesh (BD)	Brazil (BR)	China (CN)	Ethiopia (ET)	Finland (FI)	France (FR)	India (IN)	Indonesia (ID)	Iran (IR)	Ireland (IE)	Israel (IL)	Liberia (LR)	Morocco (MA)	New Zealand (NZ)	South Africa (ZA)	Spain (ES)	Switzerland (CH)	Turkey (TR)	United Kingdom (UK)	United States (US)	Grand Total
AFRICAN WHITE DORPER (AWD)																2						2
AFSHARI (AFS)										2												2
AWASSI (AWS)												1							2			3
BANGLADESHI (BAN)		2																				2
BANGLADESHI GAROLE (BAG)		1																				1
BELCLARE (BEC)											2											2
BENI GUIL (BEN)														6								6
BORDER LEICESTER (BOR)	21														1							22
BOUJAD (BOU)														1								1
BRAZILIAN CREOLE (BRA)			2																			2
CASTELLANA (CAS)																	2					2
CHANGTHANGI (CHN)								2														2
CHAROLAIS (CHA)											2											2
CHEVIOT (CHE)															1					1		2
CHURRA (CHU)																	2					2
CINE CAPARI (CCP)																			1			1
COMPOSITE (CMP)	177														98						17	292
COOPWORTH (CPW)	4														38							42
CORRIEDALE (COR)															1							1
D'MAN (DMA)														26								26
DOHNE MERINO (DOM)	1																					1
DOLLGELLAU WELSH MOUNTAIN (DWM)																				1		1
DORPER (DOR)																					6	6
DORPER WHITE (DWI)																					4	4
DORSET (DOW)															1						11	12
DORSET HORNED (DOH)	1																					1
DORSET POLLED (DOP)	30																					30
ETHIOPIAN MENZ (MEN)					1																	1
FINNSHEEP (FIN)						2									2						10	14
GARUT (GAR)									2													2
GULF COAST NATIVE (GUL)																					2	2
INDIAN GAROLE (GAI)								1														1
KARAKAS (KAR)																			2			2
KARYA (KRY)																			1			1
KATAHDIN (KAT)																					8	8
MEAT LACAUNE (LAM)							1															1
MERINO (MER)	3																					3
MERINO HORNED (MEH)	68																					68
MERINO POLLED (MEP)	56																					56
MILK LACAUNE (LAC)													1									1
MORADA NOVA (MOR)			2																			2
NAMAQUA AFRIKANER (MAN)	1																					1
NAVAJO CHURRO (NAC)																					1	1
NORDUZ (NOR)																			2			2
NORWEGIAN WHITE SHEEP (NWS)															2							2
OJALADA (OJA)																	2					2
OULED DJELLAL (OUL)														8								8
RAMBOUILLET (RBM)																					10	10
ROMANOV (RMV)																					10	10
ROMNEY (ROM)															48							48
RONDERIB AFRIKANER (RON)	2																					2
SAKIZ (SAK)																			2			2
SALZ (SAL)																	2					2
SANTA INES (SAN)			2																			2
SARDINIAN ANCESTRAL BLACK (SAR)														24								24
SCOTTISH BLACKFACE (SCB)																				1		1
SUFFOLK BLACKFACE (SUF)	1										2				1						9	13
SUFFOLK WHITE (SUW)	13																					13
SUMATRAN (SUM)									2													2
SWISS MIRROR (MIR)																		1				1
SWISS WHITE ALPINE (SWA)																		4				4
TEXEL (TEX)	2										2				9					1	10	24
TIBETAN (TIB)				2																		2
TIMAHDITE (TIM)														15								15
TREGAON WELSH MOUNTAIN (TWM)																				1		1
UNKNOWN (UKN)	2						10			19				65	10							106
VALAIS BLACK NOSE (VBN)																		1				1
VENDEEN (VEN)											2											2
WELSH HARDY SPECKLED FACE (WHS)																				1		1
WILTSHIRE (WIL)															2							2
Grand Total	382	3	6	2	1	2	11	3	4	21	10	1	1	145	214	2	8	6	10	6	98	936

- Genotype calls
  - Rudi has checked what EVA offers for download.
  - WARNING, this appears to be incomplete for both run1 (chr12 only) and run2 (only chroms 2,8,11,15,17,24,25).

        - Our colleague Sean McWilliam (CSIRO) exchanged files with EVA and is following this up.
        - He has made available the unfiltered merged set of sheep genome variants for run2 at https://doi.org/10.25919/5d3a234da46eb

    - Genotypes on browser
        - Looks okay for run1
        - Not available for run2, only their ENA file browser, https://www.ebi.ac.uk/ena/browser/view/PRJEB31241

    - Links between IDs

        We want a robust unambiguous link between
        • sample ID reported on EVA's browser, e.g. AUCMPF000000000229
        • sample ID reported in vcf files, e.g. AUCMPF000000000229
        • sample ID on BioSample, e.g. SAMN07344217
        • sample ID on SRA, e.g. SRS2478874

        - For run2 there is a problem at EVA if you compare their vcf file and their browser to our records, please contact rudiger.brauning@agresearch.co.nz

          Only 365 records are OKAY
          - sample_alias == SGD ID
          - sample_accession == SGD BioSample ID
          - secondary_sample_accession == SGD SRA ID

          10 records are NOT OKAY but can be fixed
          - NOT sample_alias == SGD ID
          - sample_accession == SGD BioSample ID
          - secondary_sample_accession == SGD SRA ID

          232 records are NOT OKAY but can be fixed
          - NOT sample_alias == SGD ID
          - sample_accession == SGD BioSample ID
          - NOT secondary_sample_accession == SGD SRA ID

          328 records are NOT OKAY cannot be fixed easily
          - NOT sample_alias == SGD ID
          - NOT sample_accession == SGD BioSample ID
          - NOT secondary_sample_accession == SGD SRA ID

        - Run3
        - In **run3 we will no longer assign samples SGD IDs but use BioSampleIDs** instead. This allows sample metadata like sex and breed to be updated without necessitating a change in sample ID. Backwards compatibility is provided via run2's SGDBID/BiosampleID table
        -

- **Run 3**

- WGS data required to be deposited at SRA and project number sent to Rudiger Brauning at AgR for inclusion in Run 3 by August 2019. Need to follow up on samples not yet at SRA.
- New samples:
  - Straight from SRA.
  - A list of SGD2 BioSampleIDs is available, I'm missing BioSampleIDs for some Teagasc animals. Sean confirmed that those BioSampleIDs have not been created. EVA has no list of animals and their BioSampleIDs.
  - Alan Archibald (Roslin): 8 Scottish Blackface and 10 Cheviots. Rudi to follow up.
  - Emily Clark (Roslin): 197 genomes from African sheep https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA523711
  - Cord Drögemüller (University of Bern): 50 individual sheep genomes at ~20x coverage belonging to 14 different local breeds. Rudi to follow up.
  - Gwenola Tosser 58 genomes
    - https://www.ncbi.nlm.nih.gov/bioproject/PRJEB9911 3
    - https://www.ncbi.nlm.nih.gov/bioproject/PRJEB14098 7
    - https://www.ncbi.nlm.nih.gov/bioproject/PRJEB14418 16
    - https://www.ncbi.nlm.nih.gov/bioproject/PRJEB31930 6
    - https://www.ncbi.nlm.nih.gov/bioproject/PRJEB32110 26
  - Possibly from Chinese breeds (Whole-Genome Sequencing of Native Sheep Provides Insights into Rapid Adaptations to Extreme Environments)? 77 animals, sequence data available upon request.
  - Trundle through SRA, use only Illumina and BGI/MGI platform.

- Utilising one pipeline (BWA; GATK haplotypeCaller for SNPs and indels).
  - Avoids +1 problem by generating gvcfs
  - Run3 should create a single set of files with tranche/set information embedded for custom filtering.
  - Unmapped reads are to be captured for separate (unrelated) analysis.
  - Not worrying about lossy CRAM format and sticking with BAM.
  - It has been noted that although the effect of quality trimming is small it does improve the overall results. We will trim and filter fastq, it's more conservative and "better" for scientific reviewers.
- Update at PAG XXVIII

3. Ovine FAANG Project update (Brenda Murdoch)
   - Brenda to update at PAG in the ISGC session and possibly the friday FAANG session prior to PAG
     - 56 tissues in total that will have functional assays performed
     - CAGE (Roslin) is almost complete on 56 tissues.
     - mRNA-seq (Baylor) completed on 30 tissues with the remaining 26 to be sequenced. A subset (8 tissues) also with ISO-Seq
     - ChIP-seq (UIDAHO): has completed 42 tissues for 4 antibodies. Samples from 38 tissues have been sent for sequencing. An additional 5 tissues will be completed bringing total to 47 tissues.
     - ATAC-seq (WSU): assay is still under development.

- WGBS (AgResearch): WGBS underway for the subset of tissues that have mRNA-seq, CAGE, Iso-seq, and ChIP-seq (8 tissues). RRBS carried out on the remaining 48 tissues.
- Rambouillet v1 genome used for all data
- Current grant (4 year) - ending 2020.

- Assay update:
  - CAGE (Emily Clark):
    - completed and hoping to include the mRNA sequence data from USMARC and possibly DNA methylation in the analysis
    - putting together a manuscript to be ready by the end of the year based on the currently available Rambouillet reference
  - WGBS/RRBS (Shannon Clarke):
    - Libraries have been sent for WGBS for 7 tissues
    - DNA required from ovary and liver (due to degradation of DNA) to complete the WGBS set and also in addition to complete the RRBS need Adrenal medulla (AD-M), Spleen (SPL), Lymph node (LN), Uterus caruncle (UTEC), Oviduct (OVI)
    - Tracy Hadfield will double check and send that inventory
  - RNA-seq (Kim Worley)
    - finished RNA-Seq except for metadata and analysis
  - RNA-seq (USMARC/UI)
    - total 32 tissues almost complete
  - ChIP-seq (Brenda Murdoch)
    - 47 tissues, 4 histone marks
    - Received sequence back from 65 libraries and 14 tissues
  - ATAC-seq (Stephen White)
    - Slow freeze protocol not working so have recently received new protocols for flash frozen (David Gorkin) to trial

4. Other business
   - Reference genome updates
     - Rambouilett improvement via long read data - PacBio and Oxford Nanopore data (Kim Davenport and Ben Rosen)
     - Will keep the same accession and update it with the new and improved assembly on NCBI with the revised Rambouillet reference genome release after PAG
   - Grant to be submitted to the next USDA call in March
     - Everyone to read thje RFA and plan a meeting at PAG for people to discuss projects and the grant submission

   - SNP Chips:
     - Rudi Brauning has mapped all ovine SNP chips onto all available reference genomes. Mapping for public chips available at https://doi.org/10.6084/m9.figshare.8424935.v2
     - Illumina have released a new Ovine 50K chip. This was not done with any consultation with the ISGC. Shannon Clarke provided the following feedback to Illumina on behalf of the ISGC:

Dear Illumina

Ovine 50K v2*:*

Firstly the I think that it great that Illumina have updated this chip and onto the XT 96 sample format, however, with my ISGC hat on, I am very disappointed (as are others in the consortium) that the ISGC was not consulted considering

- *the 50K v1 chip was developed by the ISGC in 2007 that was released as an Illumina product*
- *the 15K chip, again developed by the ISGC but was released by the ISGC as an Illumina product.*
    - *There were issues with the scrapie SNPs on this chip which inturn resulted in the generation of a competing chip.*
- *Th HD chip, although a controlled chip by the ISGC, is still a product for Illumina. This chip was utilised for the design of the 15K chip.*
    - *The ISGC are currently considering this chip to be released as a product for Illumina*
- *The continued support the ISGC has given to Illumina and Illumina to the ISGC over the years.*

We realise that there were likely timing restrictions to get a product out quickly, however, Illumina know that AgResearch (on behalf of the ISGC) have carried out many chip designs, cluster file generations, and subsequent validations of many single gene tests/production SNPs that are required for the community utilising

- *The international hapmap samples (therefore many breeds)*
- *The international mapping flock (therefore trios for mendelian inheritance)*
- *Animals carrying homozygous reference and alternate alleles as well as heterozygous animals for the single gene/production traits, including scrapie haplotypes for the 4 codons (ie inclusive of the French breeds)*
- *Samples that we continue to use for each new design and cluster file generation therefore enabling direct comparisons and concordance measures*

I'm hoping that what has been done is that at the very least the SNPs that are on the 15K chip but not on the 50K v1 have been utilised for this updated chip. Having a conversation with myself from an ISGC perspective would have been beneficial to Illumina and defiantly the international sheep researchers.

As discussed at ISAG, would it be possible to have the design file of this updated chip so that we can compare to the 15K, 50K and HD chip and also map the probes to understand if there are new designs for previous SNPs on chips. In addition, on behalf of the ISGC AgResearch would like to obtain at least 500 samples worth (~1000 always better for cluster file generate though) of these chips to generate a cluster file and carry out validation of at least scrapie, but also hopefully all the other SNPs if indeed the 15K design has been utilised. We would then endorse this chip for the community and provide the maf's for the breeds we've genotyped.

Your sincerely,

Shannon Clarke