Validation process of the Goat_IGGC_HTS_v3 Illumina chip

83,951 markers were submitted

Start from the previous design validated through the genotyping of 384 animals = 59727 SNPs (Goat_IGGC_65K_v2)

REMOVE 2703 SNP (GenTrain Score =0 or useless duplicates or GBS SNPs performing badly)

ADD 3,472 (362 duplicates either within the Add-On or in the final design) new GBS SNPs (Rudiger Brauning)

ADD 10,133 (10,010 on autosomes) representative SNPs from VarGoats VCF file (Andrea Talenti)

ADD 1,529 (398 duplicates=SNPs present in the available 10K IMAGE chip) ancestral SNPs from IMAGE chip design (Licia Colli & Gwenola Tosser-Klopp)

ADD 48 new variations in casein cluster (Guilherme Neumann & Siham Rahmatalla)

ADD 20 new variations to genotype CSN1S1 null alleles 01 & 02 (Get-PlaGe sequencing platform, Mathieu Charles & Gwenola Tosser-Klopp)

ADD 803 (10 duplicates) SNPs in candidate genes or candidate regions (Gwenola Tosser-Klopp & Philippe Bardou)
SHBG/TMEM154/PAEP/orthology with paratuberculosis region in cattle (kindly provided by Mekki Boussaha and M.P Sanchez, GABI, INRAE)
casein cluster ~1MB

ADD 1,654 LOF mutations from VarGoats (Marcel Amills & Maria Luigi)

ADD 4,267 gap fillers (Philippe Bardou)

ADD 4,993 probes related to 265 SV (Illumina)
Probe density at 1/1000 or 1/1500.
3 probes upstream/downstream of each region.
Max probes per region = 50
Min probes per region = 3
We then made probe selections weighted more heavily on a better probe (unique sequence) than even spacing.

ADD 8 PRNP variants (Gwenola, 6 duplicates)

Goat_IGGC_75K_v3 = 57024+21926+4993+8= 83951 sequences (submitted)

83,511 markers were synthetized

Origin	Number
v1	51919
v2	4857
v3-submitted	26735
Total	**83511**

74624 markers were clustered

This work was done in collaboration between Hélène Larroque, Isabelle Palhière Gwenola Tosser-Klopp (INRAE), Shannon Clarke, Hayley Baird (AgResearch) and Ashley Lee, Yuting Bai, André Eggen, Shu Zhang and Casey Turk (Illumina).
384 samples were genotyped on the chip. 192 (most of them already genotyped on v2) of them are from LECA, UNIMI, VarGoats project and an INRAE inhouse project, 96 are from AgReseach and 96 from UCDavis.

Purpose:
Gentrain is the process of generating a manifest and/or reference cluster file for an array product

Primary Goals:

1. Ensure accurate genotype calls
2. Maximize sample call rate and SNP frequency
3. Minimize errors (reproducibility and heritability)
4. Remove poorly-performing assays from the manifest

General Protocol (on 284 samples):

1. Check sample-dependent and sample-independent controls. Evaluate sample quality and assay performance.
2. Evaluate samples and remove any outliers.
3. Leverage GenomeStudio metrics to filter out SNPs with robust clustering.
4. Manually adjust clusters missed by GenomeStudio algorithm
    a. Optimize call frequency
    b. Call missed clusters
    c. Adjust miscalled clusters
    d. Evaluate sex chromosomes
        i. Avoid calling females on Y SNPs
        ii. Check for male hets on X SNPs
5. Cut SNPs where the likelihood of miscalls is high. Examples of cuts include:
    a. Low call frequency (percent of samples called)
    b. Poor concordance (percent of samples that received the same call on another experiment)
    c. Rep errors (number of replicate pairs with disagreeing calls)
    d. PC/PPC errors (number of parent-child pairs or parent-parent-child trios with impossible genotypes)
    e. Poor cluster separation (measure of theta distance between clusters)
    f. Low intensity (increased likelihood of inaccurate calls)

g. Monomorphic SNPs that sit far off axis

6. Check concordance with previous versions or experiments

Additional Steps (Goat_IGGC_HTS_v3)

1. Screen through list of "CSN1S1" markers and rescue, as needed
    a. Note: Not all samples will be called for these markers
2. Import cluster positions for 3 critical loci (provided by Gwenola)
3. Rescue 86 markers (Hayley Baird) through the analysis of 384 samples

Description of the markers

Origin	Number
v1	51499
v2	4528
v3-submitted	18597
Total	**74624**

Category	Number	Duplicates
10133_VarGoats-SNP	8720	
Ancestral	846	398
candidate_region	726	10
CSN1S1_null-alleles	17	10
gap_filler	3592	
GBS	3138	347
LoF	1507	
new-variations-in-casein_cluster	43	
PRP_new	8	6
SV	0	
V1 or V2	56027	707
Total général	**74624**	**1478**

Description of the downloadable files

The manifest file (bpm and csv formats) and the cluster file contain the 83,511 synthetized markers.
This csv file gives additional information (origin of the markers, a more accurate SNP location…).
Category-v1 is detailed in the "50K goat SNP chip" section (this webpage).
Category-v2 is detailed in the "50K SNP chip version 2" section (this webpage).

Upcoming work

The 384 genotypes were not sufficient to accurately validate the SV probes.
They were thus zeroed in the cluster file.
We intend to improve this through the accumulation of genotypes.
Collaboration on this project is encouraged. Please contact gwenola.tosser@inrae.fr

Acknowledgements

We thank the SNP providers and the scientists & engineers who analyzed the genotyping data on a set of 384 relevant animals (kindly provided by AgResearch, LECA, UNIMI, VarGoats project, INRAE inhouse project and UCDavis).
We thank Sigenae for bioinformatics analysis, data and information releases.
We thank Illumina for providing chips for genotyping 384 animals
We thank Illumina and AgResearch for the cluster file generation