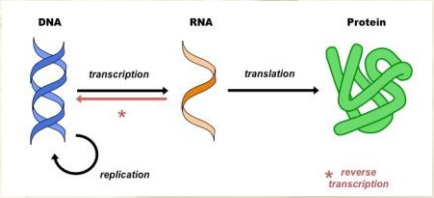
Genomic Selection

Possibilities and Present status

Concept

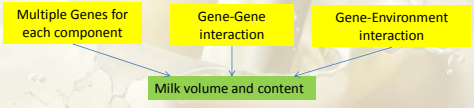
- Why using molecular markers for selection
  - Central dogma of biology



  - Genetic make-up of an individual does not change during lifespan of an individual (except point mutations)
  - Hence markers for genes can tell us how this individual is genetically

Economic traits

- Classical theory of one gene-one protein can mislead in interpreting economic trait inheritance



Around 70-75% variation is due to environment and their interaction

Early efforts

- Identifying Gene or Gene Regions (QTL) which is responsible for genetic variation in population
- However due to limited number of markers available during early years, the analysis lacked power and precision
- Numerous linkage studies were undertaken (Weller 2009), and hundreds of QTL, some of very large effect, were found in livestock species (http://www.animalgenome.org/).
- The number confirmed by repeat studies is much less, and in these the same causal locus may be mapped to somewhat different locations.

Type of markers

- Three types of observable genetic loci for use in QTL detection and MAS can be distinguished, as described by Dekkers (2004):
  - Direct markers: loci that genotype the functional polymorphism for a QTL.
  - LD-markers: loci that are in population-wide linkage disequilibrium with a QTL.
  - LE-markers: loci that are in population-wide linkage equilibrium with the functional mutation but in linkage disequilibrium on a within-family basis.

Markers examples

- Major markers in dairy population
  - DGAT (Grisart et al. 2002)
  - GRH (Blott et al. 2003)
  - κ-Casein (Medrano and Aquilar-Cordova 1990)
  - PRL (Cowan et al. 1990)
  - QTL for milk and protein yields (Spelman et al. 1996)

Markers

- Three strategies can be used to find markers that are in population-wide LD with QTL (Anderson, 2001):
- The candidate gene approach, which involves evaluating markers that are in or close to genes that are thought to be associated with the trait of interest (Rothschild and Plastow 1999)
- QTL fine-mapping approaches, starting from a previously identified QTL region, e.g. based on a cross, by saturating the region with markers.
- A genome scan using population-wide LD based on a high-density marker map, with a marker every 0.5 to 2 cM. (approx 5,00,000 to 20,00,000 Base pairs)

Genomic Selection

- As dense SNP markers were becoming available and affordable, the landmark article by Meuwissen et al. (2001) showed how whole-genome marker data could be incorporated effectively in a breeding programme for a polygenic trait.
- The idea of Meuwissen et al. (2001) was to predict breeding values using trait effects b_k estimated for (i.e., associated with) all the markers as a linear function $\Sigma x_{ik} b_k$ for individual i, where x_{ik} denotes genotype, e.g., 0, 1, 2 at locus k according to its genotype aa, Aa, or AA, utilizing their LD with nearby trait genes.
- They assumed a model in which the trait genes were dispersed throughout the genome. SNP genotypes for all loci are then included in a BLUP or similar analysis, with their associated effects as random variables.

GBLUP

- To predict breeding value using SNP markers, a simple methodology was proposed by VanRaden (VanRaden 2008).
- The pedigree relationship matrix can be replaced by an (additive) genomic relationship matrix (GRM), with elements computed in terms of identity in state (IBS) as a predictor of IBD

Model used for Genomic Breeding Value Estimation – GBLUP

$$y = Xb + Za + e$$

Where:

y = n x 1 vector of de-regressed proof of bulls ; n= number of records
b= p x 1 vector of fixed effects; p=number of levels for fixed effects
a= q x 1 vector of random animal effects; q= number of levels of random effects
e= n x1 vector of random residual effects
X= design matrix of order n x p, which relates records to fixed effects
Z= design matrix of order n x q, which relates records to random animal effects

$$\begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1}\alpha \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

G is the Genomic relational matrix based on marker data and $\alpha = \sigma^2_e / \sigma^2_a$

Requires all animals to be genotyped

Single step GBLUP

$$y = Xb + Za + e$$

Where:

y = n x 1 vector of observations; n= number of records
b= p x 1 vector of fixed effects; p=number of levels for fixed effects
a= q x 1 vector of random animal effects; q= number of levels of random effects
e= n x1 vector of random residual effects
X= design matrix of order n x p, which relates records to fixed effects
Z= design matrix of order n x q, which relates records to random animal effects

$$\begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + H^{-1}\alpha \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

H is the combined Genomic relational and pedigree based relation matrix and $\alpha = \sigma^2_e / \sigma^2_a$

We can use information from genotyped and non genotyped animals together – requires pedigree information

Bayesian approaches

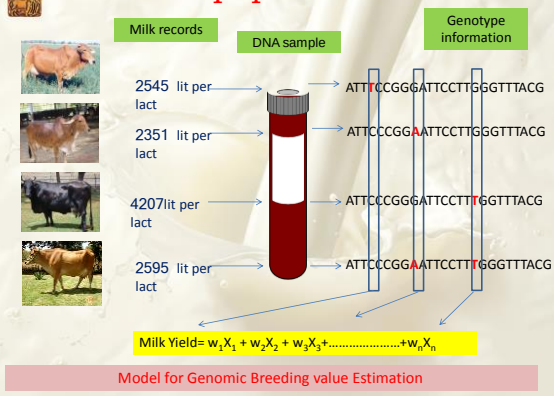
- Bayes A
- Bayes B
- Bayes Lasso

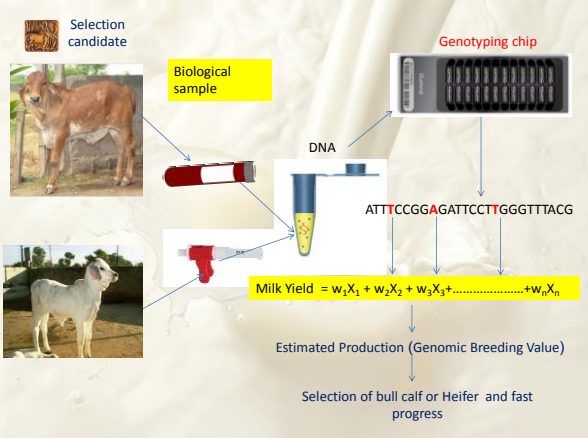
- Gianola (2013) gives a decent-comprehensive review of Bayesian methods for GBV estimation

Genomic Selection – Actual steps for implementation in a population

- Identify a panel of Polymorphic markers
- Genotype large number of animals which has phenotypes (Reference population)
- Training the model / Estimating SNP effects
- Validate the model on validation data set (subset of data which was not used for estimation but has both genotypes and phenotypes
- Use best model/estimates to predict GEBV based on only marker information

Reference population creation





How GS is expected to increase gain??

$$\text{Genetic Gain/Yr} = \frac{\text{Intensity} \times \textbf{Accuracy} \times \text{Genetic St. dev.}}{\text{Generation Interval}}$$

- Sire to Sire Path– Not much change
  - Slight increase in reliabilities with less number of daughters
- Dam to Sire Path- May help in early identification of dams
  - Heifers with high GEBVs can be used as bull dams
- Sire to Dam Path – Much benefit on this path
  - Select young bull with comparatively very high accuracy
- Dam to Dam path – farmers may use GEBV for heifer selection

NDDB's Goal for GS implementation

Genetic Improvement of animals available with farmers
for traits related to milk production and profitability

Phenotype collection

- Under PT and PS projects
  - Traits – Milk, Fat, Protein, SNF, DPR, Type traits – important and feasible traits
  - Breeds- important from milk production perspective
    - Gir, Sahiwal, Kankrej, Rathi, Hariana, Tharparkar
    - HFCB, JCB, HF
    - Jersey - to be included
    - Murrah, Mehsana, Nili-Ravi, Pandharpuri, Jaffarabadi
  - Recording at farmer's doorstep using INAPH
- Large numbers per annum for breeds included in PT projects
- Following SOPs across all projects as notified by DADF, GoI
- Ensuring data quality through robust supervision and monitoring mechanism
- Transaction recording in INAPH right at recorder level

National Dairy Plan I (Since Nov 2011)

- 14 Progeny testing projects
- Annually around 350 bulls are tested
- 35 Lakh test inseminations
- 3.5Lakh daughters registered
- 30,000 daughters milk recorded
- 1,50,000 other animals milk recorded
- 887 bulls have BV with > 70% reliability

A robust information recording system INAPH in place for all projects

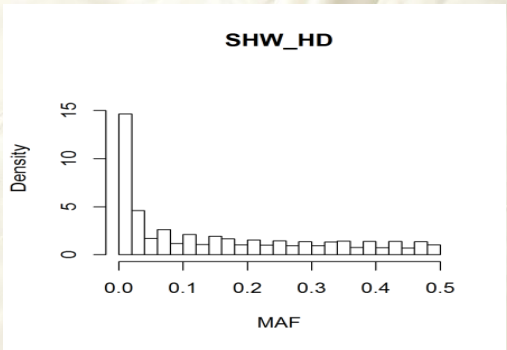
DNA for genotyping

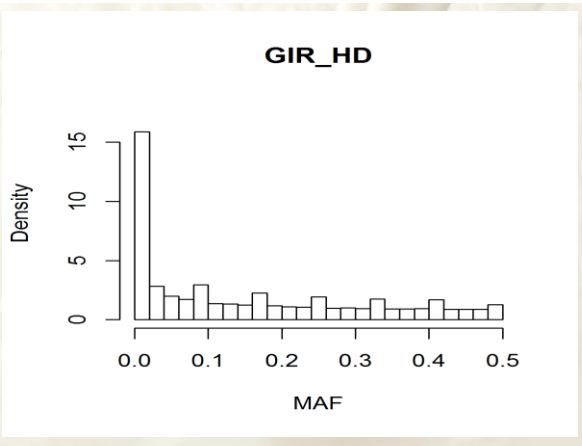
- Started collection of samples since 2014
- Initially major focus was on CB, Murrah, Mehsana – due to sufficient pedigreed observations
- Gir breed focused in 2018
- DNA isolation and creation of repository
- Use of blood, semen and ear tissue samples

Genotypes

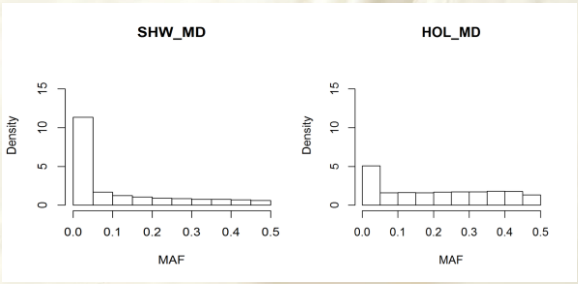
- Some of the existing medium density chips (Bovine 50K and Geneseek 75Ki) not found suitable for Indian breeds or their crosses (Nayee et. al. 2018)

MAF distribution





MAF Illumina 50K



INDUSCHIP

- INDUSCHIP developed for Major cattle breeds and their crosses.
- Process followed – Version 1
  - Genotyping representative samples of Gir, Sahiwal, Kankrej, Red Sindhi and their crosses with Illumina BovineHD
  - Narrowed down SNPs with avg distance of 70Kbps and Polymorphic for all 4/3/2 breeds
  - Identified gap for Individual breeds and filled up with SNP polymorphic for the breed
  - Added these identified SNPs on Bovine LD base
  - Added ISAG parentage SNPs
- INDUSCHIP V1 had 51 K SNPs in design

INDUSCHIP V2

- Genotyping of representative samples of 10 more breeds (Rathi, Tharparkar, Hariana, Ongole, Kangayam, Khillar, Amritmahal, Siri, Hallikar, Deoni) with BovineHD
  - Extracted INDUSCHIP V1 data, evaluated MAF and other QC parameters breed wise
  - Located big gaps of MAF (if any for a breed)
  - Added SNPs suitable to fill gaps (for major breeds)
  - Added ancestry informative SNPs
  - Added SNPs for genetic diseases or known haplotypes (Open source)
- INDUSCHIP V2 have 54 K SNPs

Calculation of GEBV for HFCB and its validation

- 2194 HFCB cows (having 1st lactation test day milk records with known pedigree) and 103 bulls were genotyped.
- GEBV estimated using SS-GBLUP using 10797daughters sired by 258 sires (inclusive of genotyped animals) for milk yield
- Validation against corrected phenotypes (CP)

GS appeared promising in bull selection for HFCB for 1st lactation Milk Yield

Bull category	Correlation Avg. Daughter CP and EBV	Correlation Avg. Daughter CP and GEBV	% increase in correlation
All sires	0.126	0.202	60.3
Genotyped sires	0.127	0.199	56.7
Sires not genotyped	0.029	0.094	224.1

Nayee et. al. in WCGALP 2018

Chromo-painting approach

- Estimated breed-of-origin proportions on chromosomes and use it for GEBV estimation

Category	Correlation of BV and corrected phenotype		Gain/loss
	Conventional ssGBLUP	Breed-of-origin ssGBLUP	
All daughters	0.1247	0.1318	5%
Non-genotyped daughters	0.1097	0.1162	6%
Genotyped daughters	0.2662	0.2758	3%
Genotyped sires	0.1422	0.1670	15%

Gajjar et. al. in WCGALP 2018

With more traits

Trait	Correl EBV and YC	Correl GEBV and YC	% increase in prediction accuracy
Milk Yield	0.075	0.160	113
Fat yield	0.053	0.134	151
Protein Yield	0.079	0.153	94

4200 animals genotyped out of 10267 HFCB animals recorded in first lactation

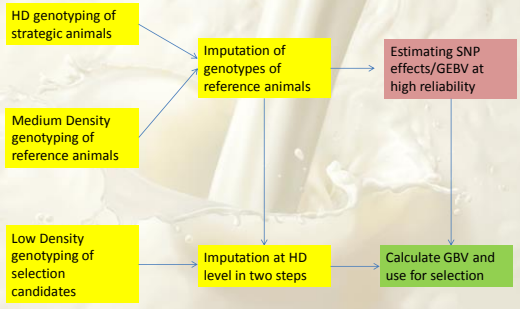
Buffalo Genotyping chip

- Tried Affymetrix Buffalo Axiome genotyping chip
  - Issues with data quality, Breed resolutions
- Work in progress on WGS of major buffalo breeds, identifying SNPs and preparing buffalo chip
- GEBVs will be estimated using genotypes from this chip

Work in progress.....

- Genotyping of more CB (HFCB and JCB) animals and calculation of GEBV for Milk, fat, protein, SNF, DPR.
- Genotyping of 3000 Gir cows (having performance records) samples and estimation of GEBV
- Efforts to include GEBV also (as criteria for bull selection) in MSP for Frozen Semen Production
- Modification of bull production programmes utilizing GEBV and OPU-IVF.

Cost effective Genomic Selection



Probable research areas

- Sequencing various breeds and accurate assemblies
- Breed specific SNPs
- Multi breed reference population and ways to estimate GBV with high accuracy
- Strategies to reduce cost of genotyping
  - Devising LD/MD panel for accurate imputation at HD level
  - Selection of reference population type and size for accurate imputation
  - Imputation methodologies for higher accuracy
- Studying effect of various breed proportions on BV for various traits
- Validation strategies for various breeds (in absence of pedigree data??)

Probable research areas

- **National Phenomic herd!!**
- Identification of major SNPs through GWAS studies on research herd and then verifying in field population –using SNP weightage for GBV estimation
- Sample collection – mechanization in processing hair samples
- Extension on communicating cost-benefit analysis of genotyping to farmers/policy makers
- **Creating national genotype repository**

Thank You