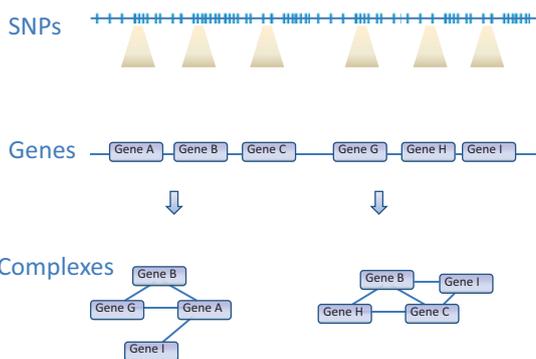# Gene prioritization by data integration

**Li Jiang, Luc Janss, Stefan M. Edwards, Axel Skarman, Bo Thomsen,  Peter Sørensen**

Centre for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University (DK).

A network-based gene prioritization approach for ranking of genes associated with a quantitative trait is presented. The approach integrates information from trait phenotypes, SNP data and protein-protein associations and uses a Bayesian mixture model for quantifying the variance explained by genetic variants linked to each candidate gene and gene complexes. The approach is illustrated on milk production traits in dairy cattle. It provides a general framework for prioritizing genes associated with various quantitative traits in different species.

## Variance component analysis using a Bayesian Mixture Model

SNPs

Genes
Gene A · Gene B · Gene C · Gene G · Gene H · Gene I

Complexes
Gene B · Gene G · Gene A · Gene I
Gene B · Gene I · Gene H · Gene C

### Data
Phenotypes was conventional estimated breeding values for milk, fat and protein yield on 4497 Danish Holstein bulls.
Genotypes was avaliable on all bulls based imputed HD chip data (>600k SNPs, MAF>0.01).
SNPs were mapped to the UMD3.1 Bovine Genome assembly containing 26,352 genes with an Entrez Gene ID.
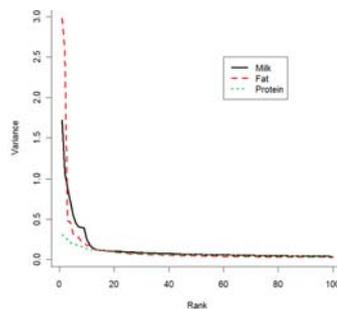Protein association was obtained from the STRING database

### Analysis
Data was analysed using the following model: $y = Xb + e$, where $y$ is the vector of phenotypes, $X$ is the design matrix relating SNP genotypes to individuals, $b$ is the vector of SNP effects and $e$ is the residual. Sampling of $b$ was based on mixture distributions with a small and large variance as prior distribution for the SNP effects.

Genomic variances was computed as $Var(X_i b_i)$ over MCMC cycles for the ith marker subset including whole genome, genes or biological complexes/pathways.

## Large overlap among high ranking genes across traits
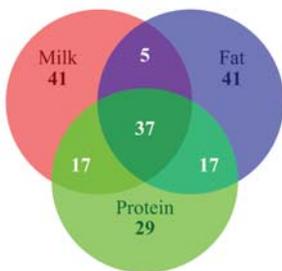
## Many genes with small variance contributions

## High ranking genes are connected in biological complexes and pathways



Venn diagram:
Milk 41
Fat 41
Protein 29
5
37
17
17

| KEGG Pathway | FAT | MILK | PROTEIN |
|---|---|---|---|
| Metabolic pathway | 4.52 | 4.26 | 2.50 |
| Glycerolipid metabolism | 3.08 | 1.85 | 0.23 |
| Retinol metabolism | 3.03 | 1.75 | 0.13 |
| Fat digestion and absorption | 3.02 | 1.77 | 0.13 |
| mRNA surveillance pathway | 2.74 | 1.00 | 0.43 |