Detection of SNPs associated with chick mortality in broilers: a machine learning approach

Nanye Long
Daniel Gianola
Kent A. Weigel
Guilherme J. M. Rosa
Santiago Avendaño

THE UNIVERSITY of WISCONSIN MADISON

The challenge of *phenomic* (phenotypic + genomic) data

- Massive phenotypic data exist
- Massive genomic data increasingly available
- SNPs
  - Human: 1.007 million SNPs (The International HapMap Consortium, 2005)
  - Chicken: 2.8 million SNPs (International Chicken Polymorphism Map Consortium, 2004)
  - Salmon: 2,500 SNPs (Hayes et al., 2004)

Background

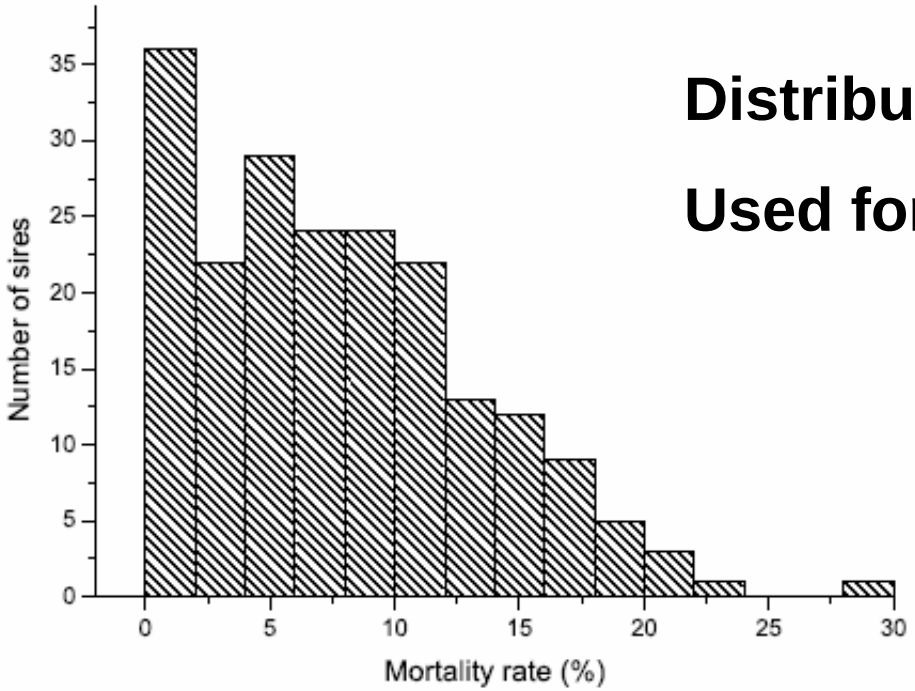
- The “large *p*, small *n* problem” in genome-wide association study.
  - Sift through thousands or even tens of thousands of SNPs, to select those related to the focal trait.
  - Usually there is a small number of phenotypic observations (*n*) and a large number of SNPs (*p*) typed.
- Examination of SNPs one by one neglects information from joint effects.
- Include all markers, model all possible interactions? Unrealistic…

Objective

- Explore model-free techniques that have been used successfully in many domains.
- Machine learning: prediction, mappings from inputs to outputs.
- Use machine learning methods for identifying subset of SNPs associated with chick mortality in broilers.

SNP-mortality data

- Genomics Initiative Project at Aviagen, Ltd.
  - Sire family mortality rates (raw and adjusted) from 0-14d progeny groups of a commercial broiler line.
  - 5,166 SNPs spreading over the chicken genome were typed on 201 sires.
  - 95.5% in HWE at 0.001 significance level

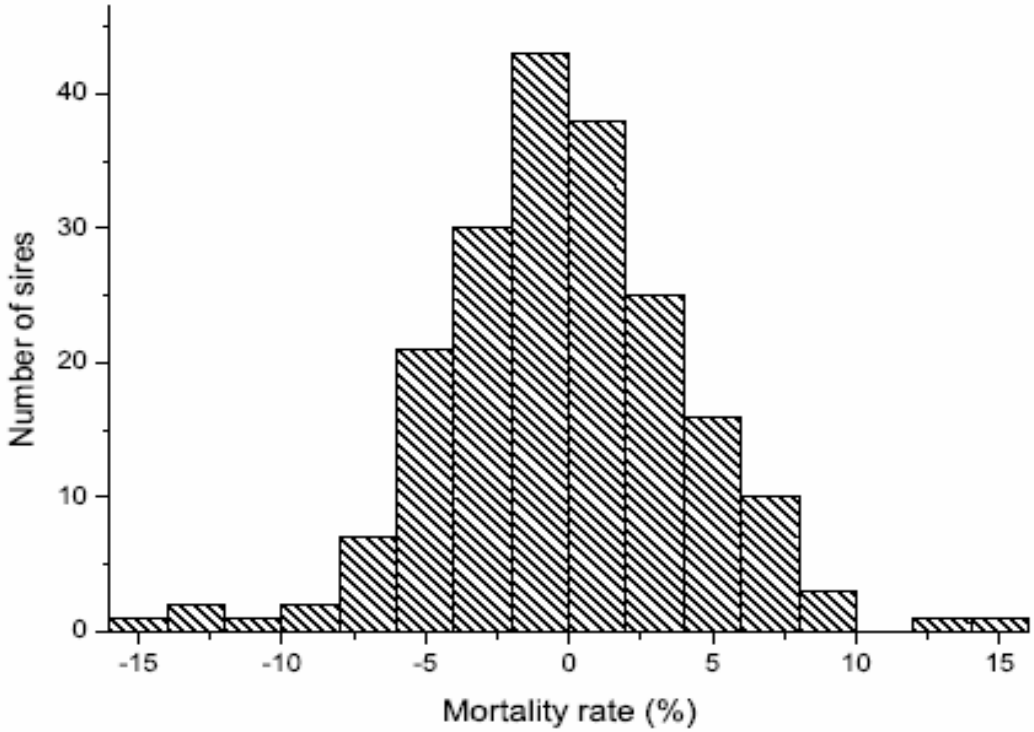


Distribution of raw mortality rate

Used for validation

Distribution of adjusted mortality rate

Used for SNP selection



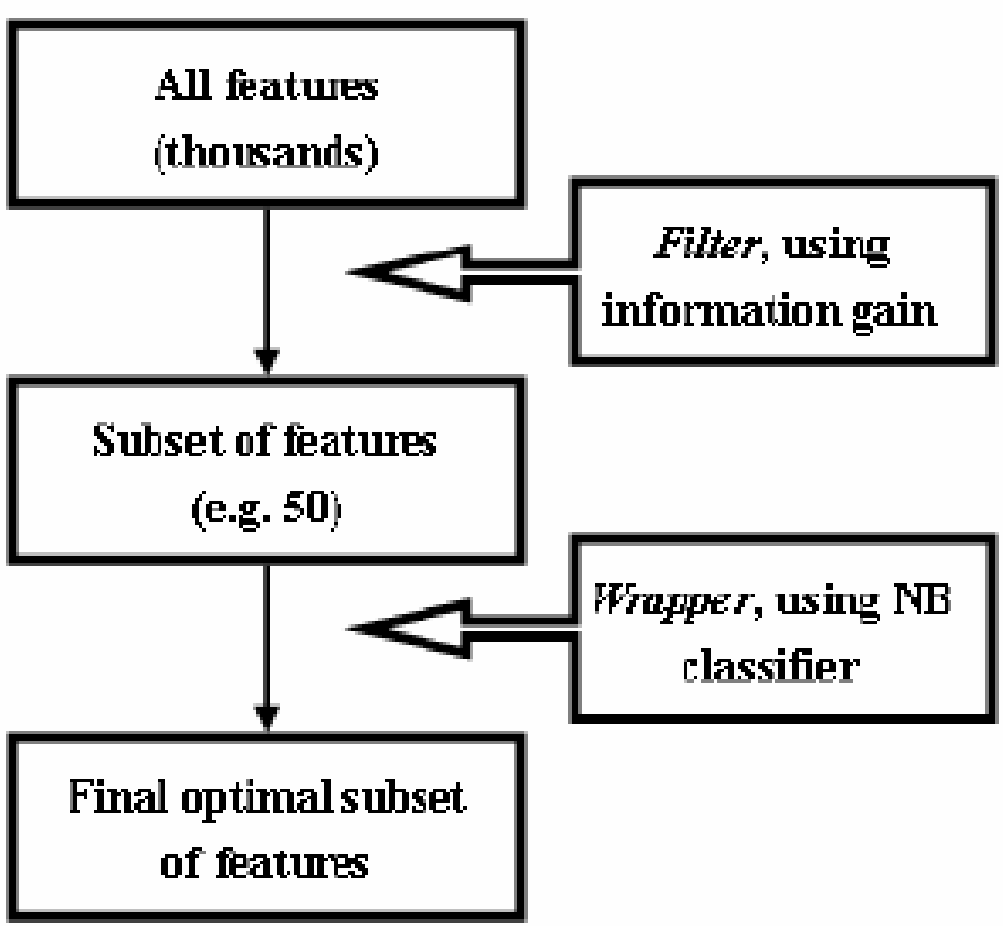
Methods

Discretizing the continuous mortality rates into two classes by two thresholds, c1 and c2, to frame it as a case-control classification problem.

Group	$(\alpha, 1-\alpha)$	$(c_1, c_2)^{\mathrm{E}}$	$(c_1, c_2)^{\mathrm{B}}$	Number of sires
1	$(0.025, 0.975)$	$(-8.92, 7.89)$	$(-8.90, 8.05)$	11
2	$(0.05, 0.95)$	$(-6.31, 6.70)$	$(-6.54, 6.69)$	21
3	$(0.10, 0.90)$	$(-5.09, 5.17)$	$(-5.16, 5.20)$	40
4	$(0.15, 0.85)$	$(-4.34, 4.09)$	$(-4.26, 4.08)$	63
5	$(0.20, 0.80)$	$(-3.50, 3.22)$	$(-3.47, 3.31)$	81
6	$(0.25, 0.75)$	$(-2.77, 2.65)$	$(-2.77, 2.52)$	102
7	$(0.30, 0.70)$	$(-2.19, 1.71)$	$(-2.21, 1.73)$	120
8	$(0.35, 0.65)$	$(-1.70, 1.20)$	$(-1.66, 1.18)$	143
9	$(0.40, 0.60)$	$(-1.20, 0.63)$	$(-1.19, 0.65)$	161
10	$(0.45, 0.55)$	$(-0.76, 0.09)$	$(-0.72, 0.16)$	180
11	(0.5)	(-0.27)	(-0.28)	201

Methods

SNP subset discovery—filter + wrapper



Filter: information gain

$$\mathrm{IG}(F) = I(\frac{N^+}{N}, \frac{N^-}{N}) - \sum_{i=1}^{\nu} \frac{N_i^+ + N_i^-}{N} I(\frac{N_i^+}{N_i}, \frac{N_i^-}{N_i})$$

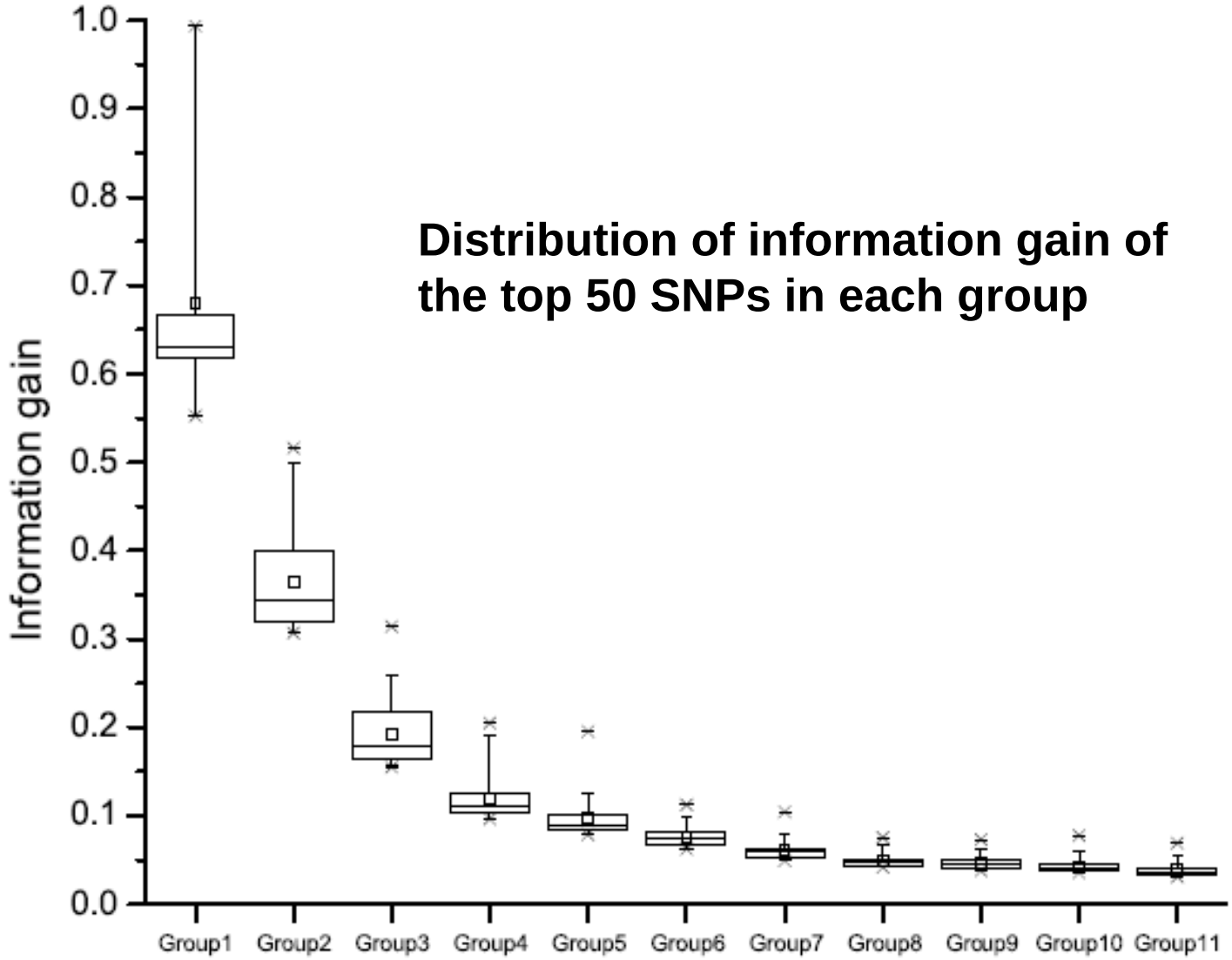
Wrapper: naïve Bayesian classifier

$$\Pr(C = c | A_1 = a_1, \ldots, A_k = a_k) = \frac{\Pr(A_1 = a_1, \ldots, A_k = a_k | C = c) \Pr(C = c)}{\Pr(A_1 = a_1, \ldots, A_k = a_k)}$$

$$\Pr(A_1 = a_1, \ldots, A_k = a_k | C = c) = \Pr(A_1 = a_1 | C = c) \cdots \Pr(A_k = a_k | C = c)$$

$$\widehat{\Pr}(A_j = a_j | C = c) = \frac{\mathrm{count}(A_j = a_j, C = c)}{\mathrm{count}(C = c)}$$

Top scoring SNPs selected from filter



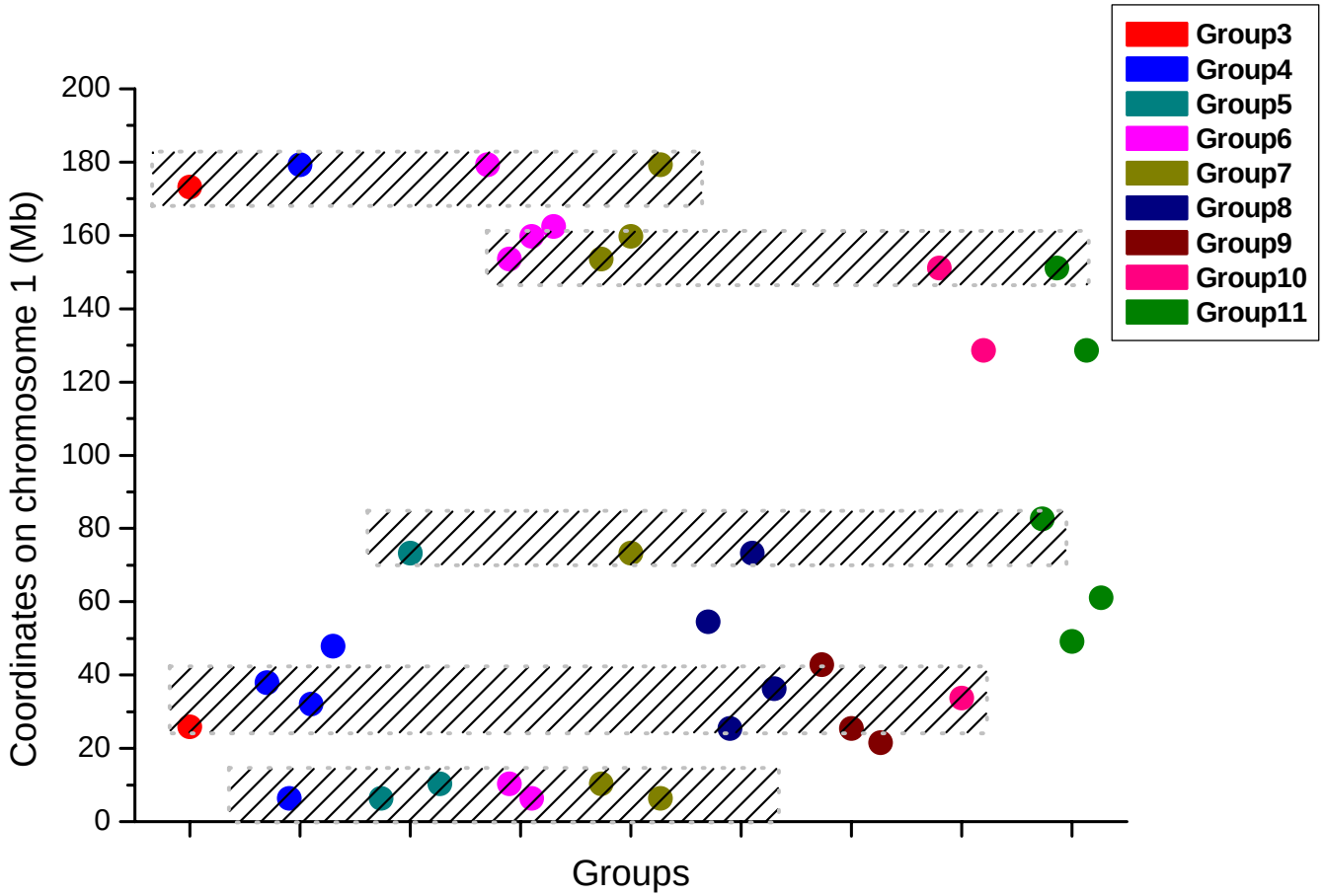
Distribution of information gain of the top 50 SNPs in each group

Wrapper results

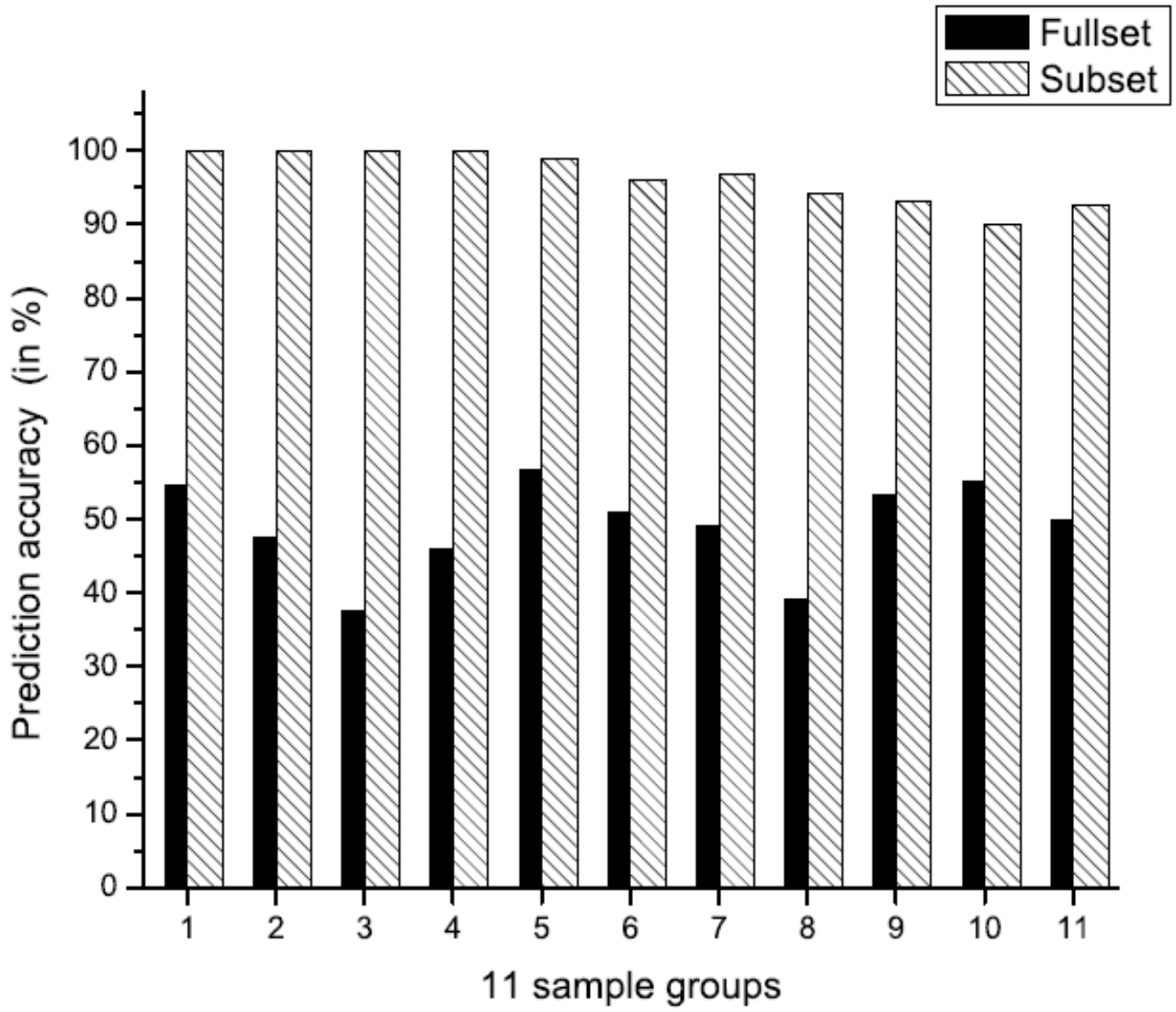
Naïve Bayesian cross-validation prediction error rates of subsets of SNPs selected by four search methods

	Prediction error rate			
Group	FS*	BE†	FSS¶	BSE‡
1	0.091 [1]	**0** [3]	0.091 [1]	0.455 [47]
2	0.095 [3]	**0** [4]	0.095 [3]	0.429 [46]
3	0.250 [2]	**0** [12]	0.250 [2]	0.325 [37]
4	0.270 [6]	**0** [17]	0.270 [4]	0.444 [43]
5	0.284 [4]	**0.012** [19]	0.284 [4]	0.309 [38]
6	0.343 [2]	**0.039** [35]	0.343 [2]	0.373 [37]
7	0.315 [10]	**0.033** [39]	0.317 [8]	0.417 [38]
8	0.402 [4]	**0.059** [42]	0.402 [4]	0.434 [24]
9	0.360 [3]	**0.068** [40]	0.360 [3]	0.342 [38]
10	0.325 [8]	**0.100** [38]	0.325 [8]	0.342 [22]
11	0.376 [4]	**0.075** [39]	0.376 [4]	0.381 [27]

Chromosome 1: the 9 sets of selected SNPs



Prediction accuracy of the top 50 SNPs (full set) and of those in the subset



Ability of predicting mortality rates

$$M_i = \mu + \sum_{j=1}^{n_g} \mathrm{SNP}_{ij} + e_i$$

Predicted residual sum of squares (PRESS)

$$\mathrm{PRESS} = \sum_{i=1}^{N} (M_i - \hat{M}_{(i)})^2$$

Group	Number of SNPs in model	p-value of model	PRESS
1	3	0.7610	0.0026
2	4	0.2805	0.0025
3	12	0.2933	0.0030
4	17	0.0004	0.0027
5	19	0.0027	0.0031
6	35	0.0011	0.0030
7	39	< 0.0001	0.0035
8	42	< 0.0001	0.0040
9	40	0.0016	0.0041
10	38	0.0020	0.0036
11	39	0.0019	0.0043

Statistical models to assess between-sire variance

Without SNP information (random effects logistic model)

$$\text{logit}(\pi_{ij}) = \alpha + u_i$$

Significant between-sire variance

After adding a set of SNPs (mixed effects logistic model)

$$\text{logit}(\pi_{ijk}) = \alpha + \sum_{j=1}^{G} \text{SNP}_{ij} + u_i$$

Between-sire variance decreased as #SNPs increased

Sires nested in SNP configurations (nested random effects model)

$$\text{logit}(\pi_{ijk}) = \alpha + g_i + u_{ij} \qquad Y_{ijk} = \alpha + g_i + u_{ij} + e_{ijk}$$

Between-sire variance was redistributed towards the between-SNP Configuration variance as #SNP configurations increased

Conclusions

- 97-98% of variation in mortality was within families.
- Machine learning procedure was applied to early mortality, as classification problem.
- Predictive ability enhanced by reducing #SNPs
- More advanced curve fitting methods to be explored
  - Generalized additive models
  - Reproducing kernel Hilbert space regression
  - Neural networks