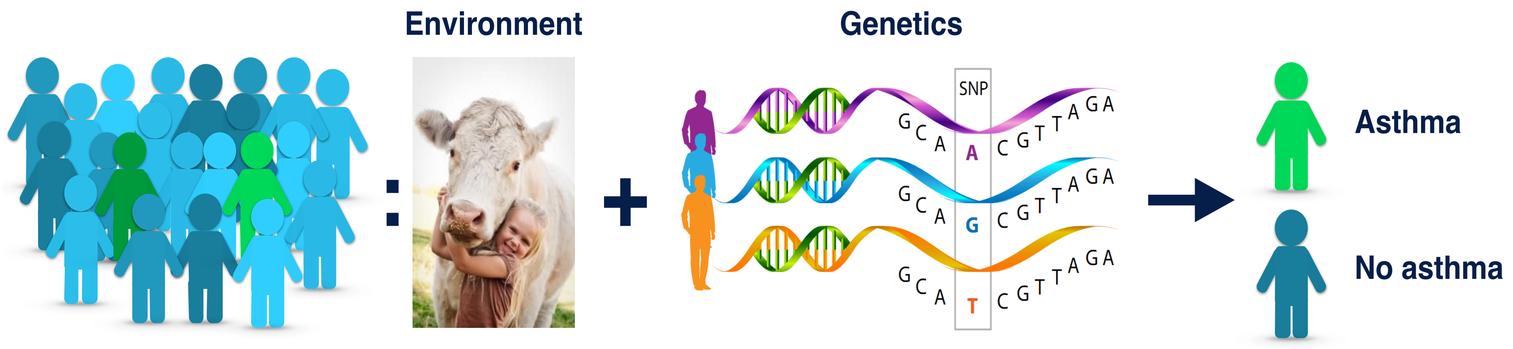
HelmholtzZentrum münchen
German Research Center for Environmental Health

Predicting from high-dimensional molecular data and environmental variables in stratified samples

Norbert Krautenbacher, Christiane Fuchs, Fabian J. Theis
Institute of Computational Biology, Helmholtz Zentrum München, Department of Mathematics, Technische Universität München

Motivation: Prediction of childhood asthma



→ Can childhood asthma be predicted by genetics and environment?

→ *Objective:* **Build risk score** for childhood asthma **based on genetic and environmental predictors**

→ *Challenges:*

- **n << p**
- **biased sample**
- **combining high-dim.** with **low-dim. data**

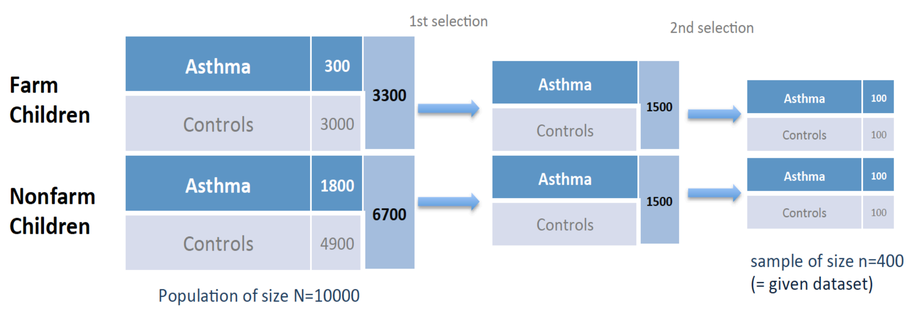
Correcting for the sample bias

Survey design (simplified)

- Interest in **influence of farm exposure** on **asthma**
- For cost reasons: detailed investigations only on **subsample of finite population**

→ How to take sample?

→ **Enrichment of informative observations** by **Stratified two-phase random selection process** (simplified, let N=10000):



→ **Given dataset** is **result of selection process**

→ **Correction for sample bias** necessary

→ otherwise biased estimates

Data

- Data from GABRIELA study
- Project in collaboration with Dr von Mutius, Haunersches Kinderspital

	~40 environmental predictors				2.5 million genetic predictors				Outcome
	Farm	Straw	Stable		SNP1	SNP2	SNP3		Asthma
Child 1	0	0	0		0.45	1	0.21		1
Child 2	1	1	0		1.76	2	1.20		0
Child 3	1	0	1	. . .	1.01	0	0.94	. . .	1
.	.	.	.		.	.	.		.
.	.	.	.		.	.	.		.
.	.	.	.		.	.	.		.
Child 1707	0	1	1		0.22	1	1.99		0

Prediction

→ **n << p** problem

→ **dimension reduction** necessary

Approaches

Correcting for Sample Bias:

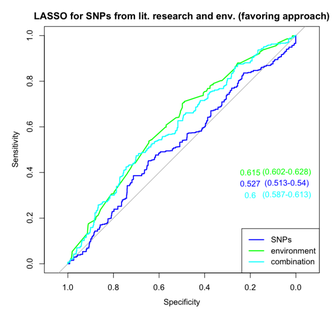
- In **training procedure**: algorithm-specific **weighting approaches**
- In **testing procedur**e: **Validation** by unbiased empirical estimate of the loss calculated by the **weighted mean over losses** per stratum

$$l_u = \frac{1}{m}\sum_{s=1}^{m} w_s \cdot l(h(x_s), y_s)$$

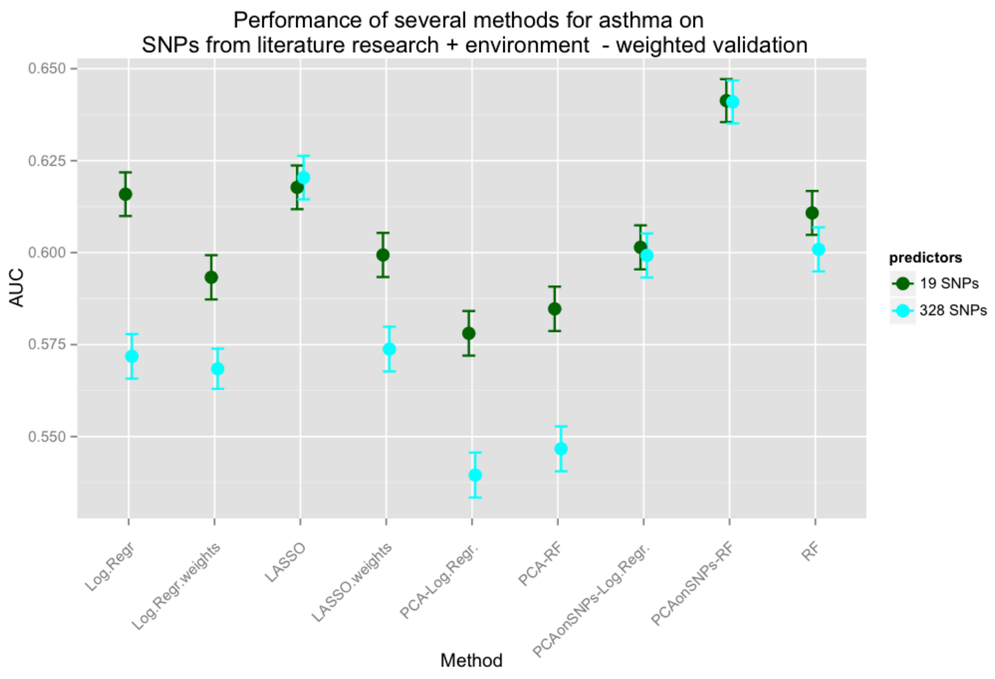
n<<p approaches:

- Reducing dimension of SNP data by
  - Manual selection of **SNPs by literature research**
  - **Univariate feature selection** by survey logistic regression
- Application of appropriate **Learning procedures:**
  - LASSO
  - Random Forest
  - PCA-based approaches

Results



- **SNPs only**: weak prediction power
- **Environment only**: better prediction than for genetics
- **Combination:** no improvement



Conclusions

- **Best prediction** by PCA followed by RF
- **Better prediction** on **environment** compared to genetics
- **Combination** of predictors **doesn't improve** prediction
  → environmental effect covers genetic effect

Outlook

Improve prediction by

- **Combining genetic** with **environmental data** by special approaches
- Incorporation of sample bias
- Taking **all SNPs** into account in further **multivariate methods**

- Ege et al. Gene-environment interaction for childhood asthma and exposure to farming in Central Europe, *JACI*, 2011
- R. De Bin, W. Sauerbrei, A.-L. Boulesteix Investigating the prediction ability of survival models based on both clinical and omics data: two case studies, *Statistics in Medicine*, 2014