# COMBI - Combining high-dimensional classification and multiple hypotheses testing for the analysis of big data in genetics

Thorsten Dickhaus

University of Bremen
Institute for Statistics

AG DANK Herbsttagung 2016
WIAS Berlin, 18.11.2016

# Outline

1. Motivation: Genetic association studies
2. Statistical setup and challenges
3. Combined two-stage procedure

Reference:

Mieth, B., Kloft, M., Rodriguez, J.A., Sonnenburg, S., Vobruba, R., Morcillo-Suarez, C., Farre, X., Marigorta, U.M., Fehr, E., Dickhaus, T., Blanchard, G., Schunk, D., Navarro, A. and Müller, K.-R. (2016): Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies. *Scientific Reports, in press.*

# Outline

1 Motivation: Genetic association studies

2 Statistical setup and challenges

3 Combined two-stage procedure

Reference:

Mieth, B., Kloft, M., Rodriguez, J.A., Sonnenburg, S., Vobruba, R., Morcillo-Suarez, C., Farre, X., Marigorta, U.M., Fehr, E., Dickhaus, T., Blanchard, G., Schunk, D., Navarro, A. and Müller, K.-R. (2016): Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies. *Scientific Reports, in press.*

# Outline

1. Motivation: Genetic association studies
2. Statistical setup and challenges
3. Combined two-stage procedure

Reference:

Mieth, B., Kloft, M., Rodriguez, J.A., Sonnenburg, S., Vobruba, R., Morcillo-Suarez, C., Farre, X., Marigorta, U.M., Fehr, E., Dickhaus, T., Blanchard, G., Schunk, D., Navarro, A. and Müller, K.-R. (2016): Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies. *Scientific Reports, in press.*

# Outline

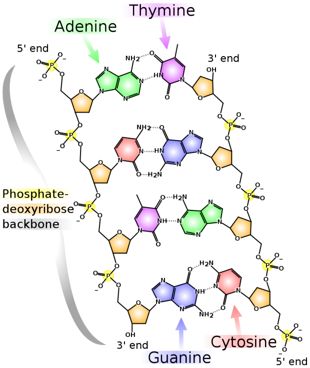
1 Motivation: Genetic association studies

2 Statistical setup and challenges

3 Combined two-stage procedure

# Deoxyribonucleic acid (DNA)

- Genetic information is coded in the base pairs at loci of the DNA
- Different possible realizations at one locus: alleles
- Body cells are diploid, i. e., consisting of two sets of chromosomes
- Individual with same alleles on both chromosomal double-helices at a particular locus $i$: homozygous at $i$, otherwise heterozygous at $i$



# What is a SNP (single nucleotide polymorphism) ?

## Bi-allelic SNPs: Exactly two possible alleles

```
Locus     1   2   3   4    ...    i    ...    M
```

# What is a SNP (single nucleotide polymorphism) ?

## Bi-allelic SNPs: Exactly two possible alleles

| Locus | 1 | 2 | 3 | 4 | ... | i | ... | M |
|---|---|---|---|---|---|---|---|---|
| Tom | A | A | G | T | ... | A | ... | G |

# What is a SNP (single nucleotide polymorphism) ?

## Bi-allelic SNPs: Exactly two possible alleles

| Locus | 1 | 2 | 3 | 4 | ... | i | ... | M |
|---|---|---|---|---|---|---|---|---|
| Tom | A | A | G | T | ... | A | ... | G |
| Andrew | A | A | G | C | ... | A | ... | C |

# What is a SNP (single nucleotide polymorphism) ?

## Bi-allelic SNPs: Exactly two possible alleles

| Locus | 1 | 2 | 3 | 4 | ... | i | ... | M |
|---|---|---|---|---|---|---|---|---|
| Tom | A | A | G | T | ... | A | ... | G |
| Andrew | A | A | G | C | ... | A | ... | C |
| Rachel | A | A | G | C | ... | G | ... | G |

# What is a SNP (single nucleotide polymorphism) ?

## Bi-allelic SNPs: Exactly two possible alleles

| Locus | 1 | 2 | 3 | 4 | ... | i | ... | M |
|---|---|---|---|---|---|---|---|---|
| Tom (m) | A | A | G | T | ... | A | ... | G |
| Tom (p) | A | A | G | T | ... | A | ... | C |
| Andrew | A | A | G | C | ... | A | ... | C |
| Rachel | A | A | G | C | ... | G | ... | G |

# What is a SNP (single nucleotide polymorphism) ?

## Bi-allelic SNPs: Exactly two possible alleles

| Locus | 1 | 2 | 3 | 4 | ... | i | ... | M |
|---|---|---|---|---|---|---|---|---|
| Tom (m) | A | A | G | T | ... | A | ... | G |
| Tom (p) | A | A | G | T | ... | A | ... | C |
| Andrew | A | A | G | C | ... | A | ... | C |
|  | A | A | G | C | ... | G | ... | C |
| Rachel | A | A | G | C | ... | G | ... | G |
|  | A | A | G | T | ... | G | ... | G |

# Outline

1 Motivation: Genetic association studies

2 Statistical setup and challenges

3 Combined two-stage procedure

# Contingency table layout in association studies

Assume a bi-allelic marker (SNP) at a particular locus and a binary phenotype of interest, e. g., a disease status.

| Genotype | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ | $\Sigma$ |
|---|---|---|---|---|
| Phenotype 1 | $x_{1,1}$ | $x_{1,2}$ | $x_{1,3}$ | $n_{1.}$ |
| Phenotype 0 | $x_{2,1}$ | $x_{2,2}$ | $x_{2,3}$ | $n_{2.}$ |
| Absolute count | $n_{.1}$ | $n_{.2}$ | $n_{.3}$ | $N$ |

In case of allelic tests:

| Genotype | $A_1$ | $A_2$ | $\Sigma$ |
|---|---|---|---|
| Phenotype 1 | $x_{1,1}$ | $x_{1,2}$ | $n_{1.}$ |
| Phenotype 0 | $x_{2,1}$ | $x_{2,2}$ | $n_{2.}$ |
| Absolute count | $n_{.1}$ | $n_{.2}$ | $N$ |

# Formalized association test problem

Multiple test problem with system of hypotheses $\mathcal{H} = (H_j : 1 \leq j \leq M)$, where $H_j$ : Genotype$_j$ $\perp$ Phenotype with two-sided alternatives $K_j$.

# Formalized association test problem

Multiple test problem with system of hypotheses $\mathcal{H} = (H_j : 1 \leq j \leq M)$, where $H_j$ : Genotype$_j \perp$ Phenotype with two-sided alternatives $K_j$.

Abbreviated notation (one particular position):

$\mathbf{n} = (n_{1.}, n_{2.}, n_{.1}, n_{.2}, n_{.3}) \in \mathbb{N}^5$ resp. $\mathbf{n} = (n_{1.}, n_{2.}, n_{.1}, n_{.2}) \in \mathbb{N}^4$ ,

$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{pmatrix} \in \mathbb{N}^{2\times3}$ resp. $\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \in \mathbb{N}^{2\times2}$.

# Hypergeometric table probability

In both cases, the probability of observing $\mathbf{x}$ given $\mathbf{n}$ is under the null given by

$$f(\mathbf{x}|\mathbf{n}) = \frac{\prod_{n\in\mathbf{n}} n!}{N! \prod_{x\in\mathbf{x}} x!}.$$

# Tests for association of marker and phenotype

## (i) Chi-squared test

$$Q(\mathbf{x}) = \sum_r \sum_s \frac{(x_{rs} - e_{rs})^2}{e_{rs}}, \text{ where } e_{rs} = n_{r.} n_{.s}/N.$$

Resulting ”exact” (non-asymptotic) $p$-value:

$$p_Q(\mathbf{x}) = \sum_{\tilde{\mathbf{x}}} f(\tilde{\mathbf{x}}|\mathbf{n}), \text{ with}$$

summation over all $\tilde{\mathbf{x}}$ with marginals $\mathbf{n}$ such that $Q(\tilde{\mathbf{x}}) \geq Q(\mathbf{x})$.

(Local) level $\alpha$ test: $\varphi_Q(\mathbf{x}) = \mathbb{1}_{p_Q(\mathbf{x}) \leq \alpha}$

# Tests for association of marker and phenotype

## (ii) Tests of Fisher-type

$$p_{\text{Fisher}}(\mathbf{x}) = \sum_{\tilde{\mathbf{x}}} f(\tilde{\mathbf{x}}|\mathbf{n}), \text{ with}$$

summation over all $\tilde{\mathbf{x}}$ with marginals $\mathbf{n}$ such that $f(\tilde{\mathbf{x}}|\mathbf{n}) \leq f(\mathbf{x}|\mathbf{n})$.

Corresponding level $\alpha$ test: $\varphi_{\text{Fisher}}(\mathbf{x}) = \mathbb{1}_{p_{\text{Fisher}}(\mathbf{x}) \leq \alpha}$

# Tests for association of marker and phenotype

## (ii) Tests of Fisher-type

$$p_{\text{Fisher}}(\mathbf{x}) = \sum_{\tilde{\mathbf{x}}} f(\tilde{\mathbf{x}}|\mathbf{n}), \text{ with}$$

summation over all $\tilde{\mathbf{x}}$ with marginals $\mathbf{n}$ such that $f(\tilde{\mathbf{x}}|\mathbf{n}) \leq f(\mathbf{x}|\mathbf{n})$.

Corresponding level $\alpha$ test: $\varphi_{\text{Fisher}}(\mathbf{x}) = \mathbb{1}_{p_{\text{Fisher}}(\mathbf{x}) \leq \alpha}$

$\varphi_Q$ and $\varphi_{\text{Fisher}}$ keep the (local) significance level $\alpha$ conservatively for any sample size $N$.

# Challenges for the statistical methodology

1. $M >> 1$ simultaneous association tests (high multiplicity, ”big data”)
2. Interactions between genes (genetic networks)

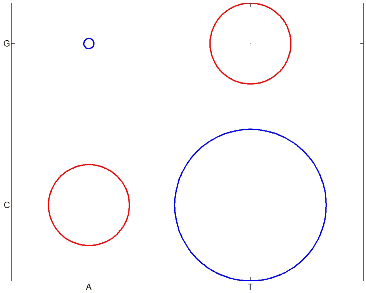
**$\Rightarrow$ A simple locus-by-locus analysis with a Bonferroni adjustment for multiplicity is inappropriate here!**

(We want to control the family-wise error rate (FWER), i. e., the probability of at least one type I error among the $M$ individual tests.)

# Interactions: First example

**Statistically non-significant! ($\alpha = 0.025$)**

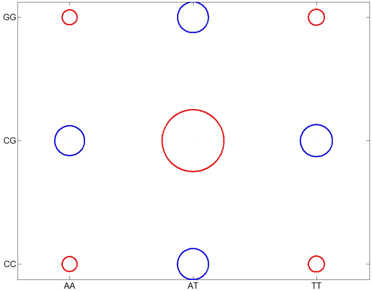
|  | AC | AG | TC | TG | $\sum$ |
|---|---|---|---|---|---|
| Y = 0 | 0 | 1 | 15 | 0 | 16 |
| Y = 1 | 8 | 0 | 0 | 8 | 16 |



# Interactions: Second example

**Statistically non-significant! ($\alpha = 0.05$)**

| | AACC | AACG | AAGG | ATCC | ATCG | ATGG | TTCC | TTCG | TTGG | $\sum$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Y = 0 | 0 | 120 | 0 | 125 | 0 | 124 | 0 | 129 | 0 | 498 |
| Y = 1 | 60 | 0 | 60 | 0 | 249 | 0 | 64 | 0 | 64 | 497 |



# Outline

1 Motivation: Genetic association studies

2 Statistical setup and challenges

3 Combined two-stage procedure

# The ”combi method” in a nutshell

1. Fix a number $1 \leq k \leq M$ of hypothesized informative positions.
2. Run a support vector machine-based classification. Record the $k$ largest of the SVM weights in absolute value.
3. Compute the $k$ corresponding $p$-values for testing association.
4. For a pre-defined FWER level $\alpha$, decide that position $j$ is informative if $j$ is among the ”top $k$” SVM positions **and** its $p$-value is below a threshold $t^* \equiv t^*(k, \alpha)$.

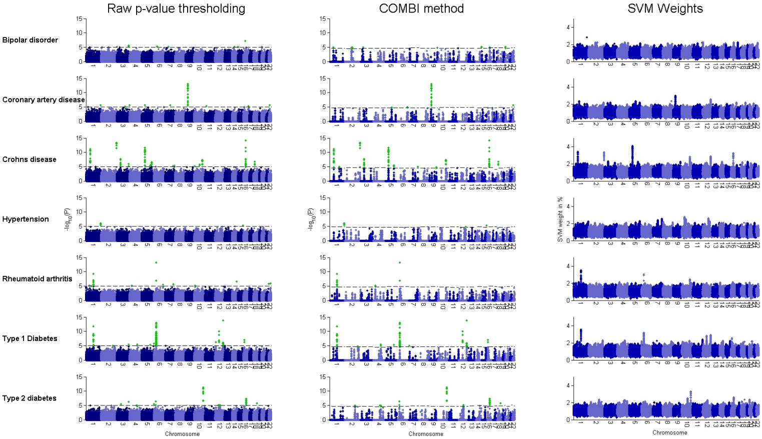
The threshold $t^*$ has to be chosen such that the FWER is controlled at level $\alpha$ for the entire procedure.

# Computation of $t^*$

- We developed a fully resampling-based method for calibrating $t^*$ on the basis of the ascertained data.
- Essentially, it employs an estimation of the correlation resp. affinity of SVM weights and $p$-values for association.
- The resampling method is a generalization of the 'min P' procedure (Westfall and Young, 1993).
- Writing $t^*$ in the form $\alpha/M_{\text{eff.}}$, we conjecture that $k < M_{\text{eff.}} << M$ for most of the relevant applications.

  We call $M_{\text{eff.}}$ the "effective number of tests" in the second step of the combi method.

# Application of the combi method (WTCCC 2007 data)


Raw p-value thresholding
COMBI method
SVM Weights
Bipolar disorder
Coronary artery disease
Crohns disease
Hypertension
Rheumatoid arthritis
Type 1 Diabetes
Type 2 diabetes
$-\log_{10}(P)$
SVM weight in %
Chromosome