



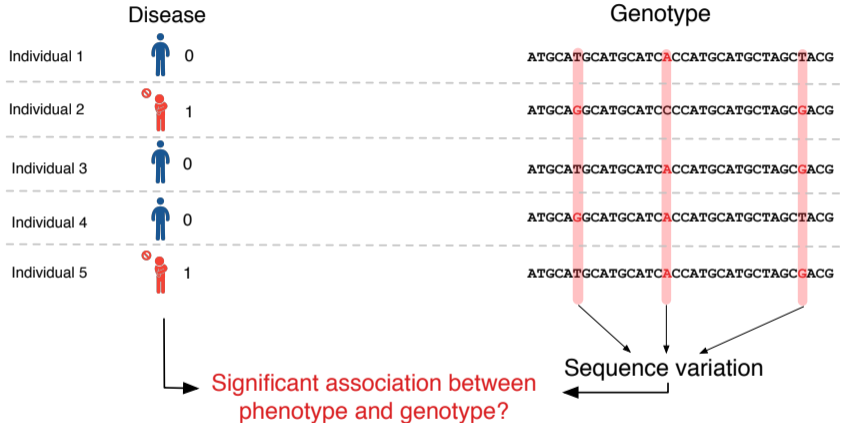
# Combinatorial Association Mapping

Karsten Borgwardt

ETH Zürich, Department Biosystems

Saarbrücken, May 10, 2017

# Mapping Phenotypes to the Genome



A **genome-wide association study (GWAS)** examines whether variation in the genome (in form of single nucleotide polymorphisms, SNPs) correlates with variation in the phenotype.

# Missing Heritability

- Since 2001: More than 2000 new disease loci due to GWAS
- Problem: Phenotypic variance explained still disappointingly low

Vol 461|8 October 2009|doi:10.1038/nature08494

nature

## REVIEWS

---

### Finding the missing heritability of complex diseases

Teri A. Manolio<sup>1</sup>, Francis S. Collins<sup>2</sup>, Nancy J. Cox<sup>3</sup>, David B. Goldstein<sup>4</sup>, Lucia A. Hindorf<sup>5</sup>, David J. Hunter<sup>6</sup>, Mark I. McCarthy<sup>7</sup>, Erin M. Ramos<sup>5</sup>, Lon R. Cardon<sup>8</sup>, Aravinda Chakravarti<sup>9</sup>, Judy H. Cho<sup>10</sup>, Alan E. Guttmacher<sup>1</sup>, Augustine Kong<sup>11</sup>, Leonid Kruglyak<sup>12</sup>, Elaine Mardis<sup>13</sup>, Charles N. Rotimi<sup>14</sup>, Montgomery Slatkin<sup>15</sup>, David Valle<sup>9</sup>, Alice S. Whittemore<sup>16</sup>, Michael Boehnke<sup>17</sup>, Andrew G. Clark<sup>18</sup>, Evan E. Eichler<sup>19</sup>, Greg Gibson<sup>20</sup>, Jonathan L. Haines<sup>21</sup>, Trudy F. C. Mackay<sup>22</sup>, Steven A. McCarroll<sup>23</sup> & Peter M. Visscher<sup>24</sup>

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable insights into their genetic architecture. Most variants identified so far confer relatively

# Missing Heritability

## Epistasis as a Potential Reason

- Most current analyses neglect interactive effects between loci
- Need for approaches for **combinatorial association mapping**

Mackay and Moore *Genome Medicine* 2014, 6:42  
<http://genomemedicine.com/content/6/6/42>



### COMMENT

## Why epistasis is important for tackling complex human disease genetics

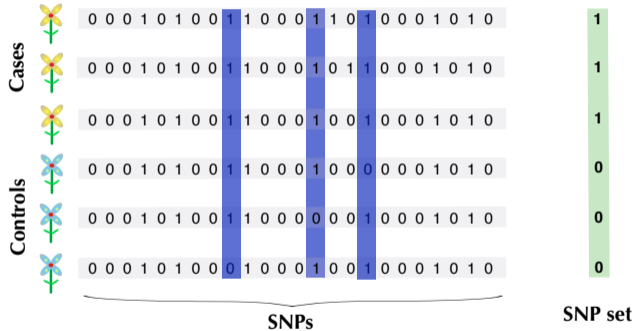
Trudy FC Mackay<sup>1\*</sup> and Jason H Moore<sup>2</sup>

#### Editorial summary

Epistasis has been dismissed by some as having little role in the genetic architecture of complex human disease. The authors argue that this view is the result

and the effects of alleles at these loci are highly sensitive to the environmental circumstances to which the individuals are exposed. Quantitative variation in phenotypes and disease risk must result in part from the perturbation of highly dynamic, interconnected and non-linear net-

# Combinatorial Association Mapping



- Computational challenge: Combinatorial explosion of the number of candidate sets
- Statistical challenge: Combinatorial explosion of the number of association tests

# Combinatorial Association Mapping

## Multiple Hypothesis Testing Problem

- What if we consider associations of groups of  $c$  SNPs with the phenotype?
- This leads to an enormous multiple testing problem: Any of the  $k$  SNP sets would correspond to a hypothesis that is tested ( $k \in O(d^c)$ ).
- If unaccounted for,  $\alpha$  per cent of all SNP sets might be considered significantly associated by random chance.
- It is imperative to control for multiple testing, e.g. the family-wise error rate!
- If accounted for, e.g. by Bonferroni correction ( $\frac{\alpha}{k}$ ), we might lose all statistical power.

# Combinatorial Association Mapping

## Multiple Hypothesis Testing Problem

- What if we consider associations of groups of  $c$  SNPs with the phenotype?
- This leads to an enormous multiple testing problem: Any of the  $k$  SNP sets would correspond to a hypothesis that is tested ( $k \in O(d^c)$ ).
- If unaccounted for,  $\alpha$  per cent of all SNP sets might be considered significantly associated by random chance.
- It is imperative to control for multiple testing, e.g. the family-wise error rate!
- If accounted for, e.g. by Bonferroni correction ( $\frac{\alpha}{k}$ ), we might lose all statistical power.
- **Long considered unsolvable dilemma**





# Significant Pattern Mining

## Tarone's trick

- Contingency table for testing enrichment of a pattern in a class

|         |         |                   |           |
|---------|---------|-------------------|-----------|
|         | $S = 1$ | $S = 0$           |           |
| $y = 1$ | $a$     | $n_1 - a$         | $n_1$     |
| $y = 2$ | $x - a$ | $n - n_1 - x + a$ | $n - n_1$ |
|         | $x$     | $n - x$           | $n$       |

- A popular choice is Fisher's exact test to test whether  $S$  is overrepresented in one of the two classes.
- The common way to compute  $p$ -values for Fisher's exact test is based on the hypergeometric distribution and assumes fixed total marginals  $(x, n_1, n)$ .

# Significant Pattern Mining

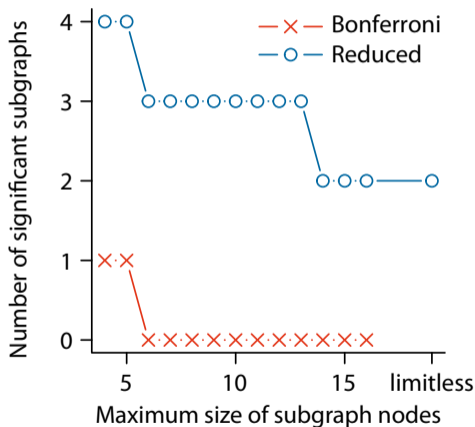
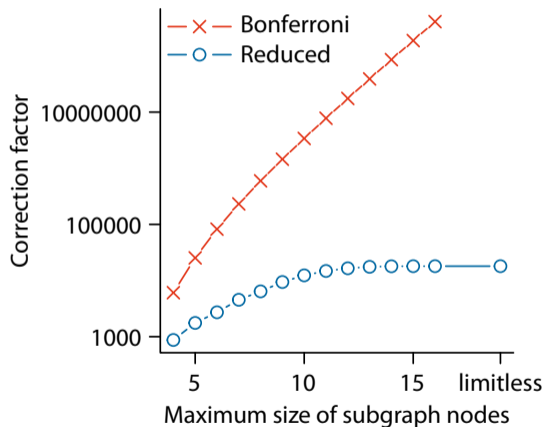
## Tarone's trick

- Contingency table for testing enrichment of a pattern in a class

|         |         |                   |           |
|---------|---------|-------------------|-----------|
|         | $S = 1$ | $S = 0$           |           |
| $y = 1$ | $a$     | $n_1 - a$         | $n_1$     |
| $y = 2$ | $x - a$ | $n - n_1 - x + a$ | $n - n_1$ |
|         | $x$     | $n - x$           | $n$       |

- Tarone (1990) noted that when working with discrete test statistics, e.g. Fisher's exact test, there is a **minimum  $p$ -value** that a pattern can achieve.
- There are many **untestable hypotheses** whose minimum  $p$ -value is not smaller than  $\frac{\alpha}{k}$ .
- Only the remaining  $m(k)$  **testable hypotheses** can reach significance at all.
- One can **correct for  $m(k)$  instead of  $k$** . As often  $m(k) \ll k$ , this greatly improves statistical power.

## Example: PTC dataset (Helma et al., 2001)



# Significant Pattern Mining

## Tarone's approach (1990)

- Assume  $k$  is the number of tests that we correct for.
- $m(k)$  is the number of testable hypotheses at significance level  $\frac{\alpha}{k}$ .
- $m(k)$  is a function of  $k$  and we require  $k \geq m(k)$  to correct for all testable hypotheses.
- Then the optimization problem is

$$\begin{aligned} \min k \\ \text{s. t. } k \geq m(k) \end{aligned}$$

# Significant Pattern Mining

## Tarone's approach (1990)

- Assume  $k$  is the number of tests that we correct for.
- $m(k)$  is the number of testable hypotheses at significance level  $\frac{\alpha}{k}$ .

**procedure** Tarone( $D, \alpha$ )

$k := 1;$

**while**  $k < m(k)$  **do**

$k := k + 1;$

**return**  $k$

# Significant Pattern Mining

## Tarone's approach (1990)

- Assume  $k$  is the number of tests that we correct for.
- $m(k)$  is the number of testable hypotheses at significance level  $\frac{\alpha}{k}$ .

**procedure** Tarone( $D, \alpha$ )

$k := 1;$

**while**  $k < m(k)$  **do**

$k := k + 1;$

**return**  $k$

- How to efficiently compute  $m(k)$  without running through all  $O(d^c)$  possible hypotheses?

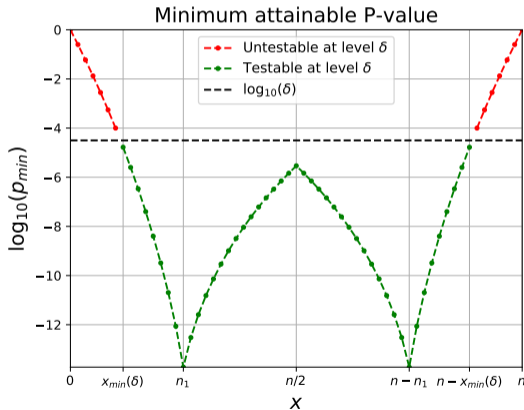
# Significant Pattern Mining

## Data mining challenge

- How to efficiently find  $m(k)$  without running through all  $O(d^c)$  possible hypotheses?
- Solution: Minimum  $p$ -value is determined by the frequency of a pattern.
- One can use frequent pattern mining algorithms from Data Mining to enumerate all patterns that pass a certain  $p$ -value threshold (Terada et al., PNAS 2013).

# Significant Pattern Mining

- Frequency versus minimum  $p$ -value





# Significant Pattern Mining

## Tarone's approach with frequent itemset mining

- Assume  $k$  is the number of tests that we correct for.
- $m(k)$  is the number of testable hypotheses at significance level  $\frac{\alpha}{k}$ .

**procedure** Tarone( $D, \alpha$ )

$k := 1$ ;

**while**  $k < m(k)$  **do**

$k := k + 1$ ;

$m(k) :=$  frequent itemset mining( $D, \theta(\frac{\alpha}{k})$ );

**return**  $k$

# Significant Pattern Mining

## Tarone's approach with frequent itemset mining

- Assume  $k$  is the number of tests that we correct for.
- $m(k)$  is the number of testable hypotheses at significance level  $\frac{\alpha}{k}$ .

**procedure** Tarone( $D, \alpha$ )

$k := 1$ ;

**while**  $k < m(k)$  **do**

$k := k + 1$ ;

$m(k) :=$  frequent itemset mining( $D, \theta(\frac{\alpha}{k})$ );

**return**  $k$

- For small  $k$ ,  $\theta(\frac{\alpha}{k})$  is small. Frequent itemset mining will be extremely expensive!

# Significant Pattern Mining

## Our contributions

- How to *efficiently* find the optimal  $k$ ? (SDM 2015)
- Patterns are in subset/superset relationships. How to account for this dependence between tests? (KDD 2015)
- Can we retain efficiency and statistical power when accounting for categorical covariates such as age and gender? (NIPS 2016)
- Can we develop new association mapping approaches based on Tarone's trick? (ISMB 2015, OUP Bioinformatics 2017)

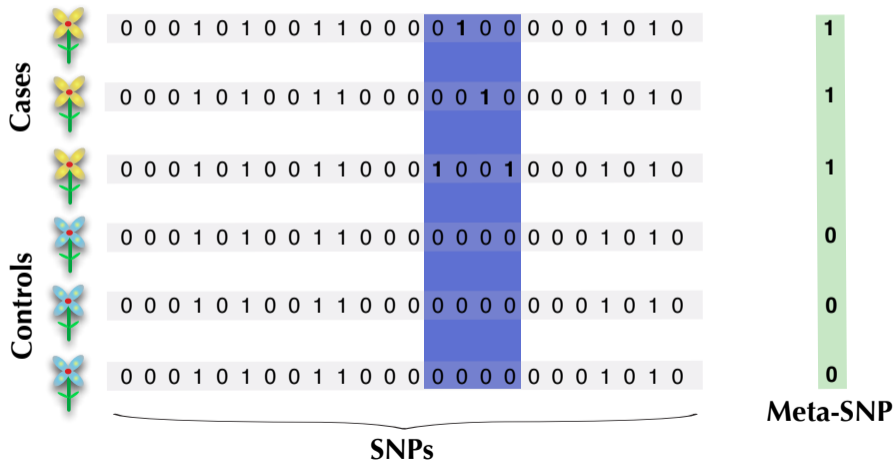
## Genetic Heterogeneity Discovery

## FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

### Genetic heterogeneity

- Genetic heterogeneity refers to the phenomenon that several different genes or sequence variants may give rise to the same phenotype.
- The correlation between each individual gene or variant and the phenotype may be too weak to be detected, but the group may have a strong correlation.
- The only current way to consider genetic heterogeneity is to consider fixed groups of variants. Genome-wide scans cause tremendous computational and statistical problems.

## FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



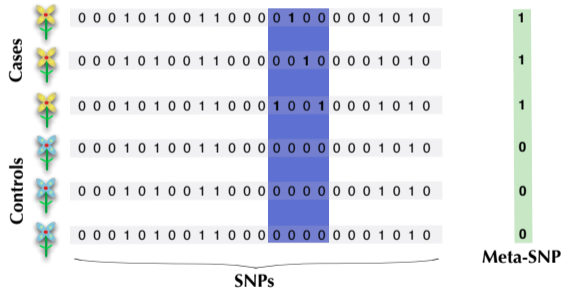
# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

## Fast Automatic Interval Search (Llinares-Lopez et al., ISMB 2015)

- Our goal is to **search for intervals that may exhibit genetic heterogeneity**, while
  - allowing for arbitrary start and end points of the intervals,
  - properly correcting for the inherent multiple testing problem, and
  - retaining statistical power and computational efficiency.
- We model the search as a **pattern mining problem**: Given an interval, an individual contains a pattern, if it has at least one minor allele in this interval.

## FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

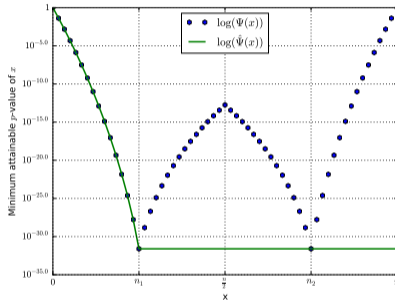
Finding trait-associated genome **segments** with at least one minor allele



- An interval is represented by its maximum value. The longer an interval, the more likely it is that this maximum is 1.

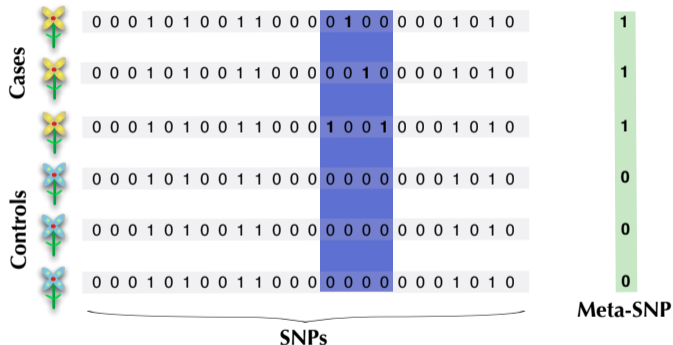


# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



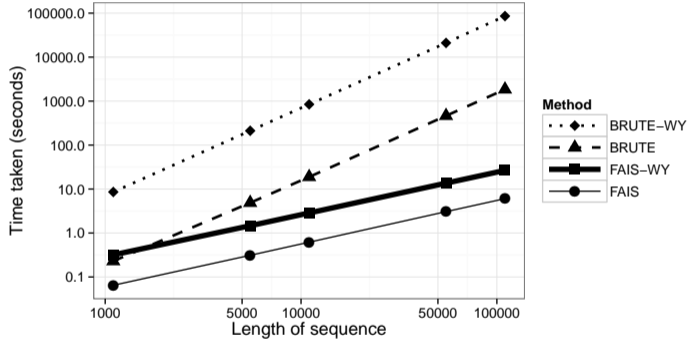
- **Pruning criterion 1:** If too many individuals have a particular pattern, the corresponding interval is not testable.

## FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



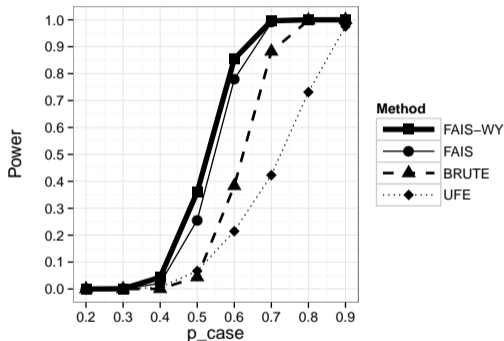
- **Pruning criterion 2:** If a pattern is too frequent to be testable, then none of the superintervals of the corresponding interval is testable.

# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



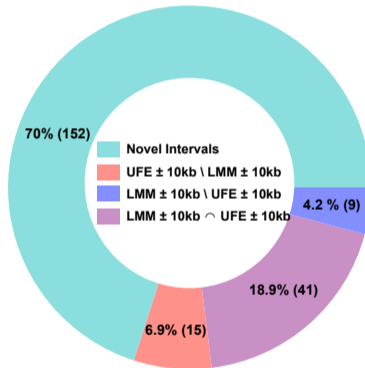
- Our method FAIS (Fast Automatic Interval Search) improves over the brute-force interval search in terms of runtime in simulations.

## FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



- Our method FAIS (Fast Automatic Interval Search) improves over brute-force interval search and univariate approaches in terms of power in simulations.

## FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



- Most significant intervals would have been missed by univariate approaches (UFE and LMM) on 21 binary phenotypes from *Arabidopsis thaliana* (Atwell et al., Nature 2010).

## FAIS: Conclusions and Outlook

### Conclusions

- We can search for intervals that may exhibit genetic heterogeneity
  - efficiently,
  - without pre-defining the boundaries of intervals,
  - while properly correcting for multiple testing.

## FAIS: Conclusions and Outlook

### Conclusions

- We can search for intervals that may exhibit genetic heterogeneity
  - efficiently,
  - without pre-defining the boundaries of intervals,
  - while properly correcting for multiple testing.

### Outlook: Genetic heterogeneity discovery

- How to account for covariates like age and gender? → Solution for categorial covariates  
(NIPS 2016, Bioinformatics 2017)
- How to extend our approach to networks of SNPs or genes? → current work

# Combinatorial Association Mapping

## Summary

- **Combinatorial Association Mapping** allows to study epistasis, one important potential reason for missing heritability.
- The high dimensionality of the problem leads to an enormous **computational and statistical challenge**.
- **Solving both problems** at the same time is **largely unachieved**.
- We have developed several **Significant Pattern Mining** approaches that achieve both.

[www.significant-patterns.org](http://www.significant-patterns.org)



## Some pointers

## easyGWAS

- We have been developing [easygwas.org](http://easygwas.org) (Grimm et al., 2017), a Machine Learning platform for Geneticists (819 users as of May 9, 2017):



# Software

## Graph Kernels

- Data and Code for graph and network comparison via [graph-kernels.org](http://graph-kernels.org)






# Thank you



- Alfried-Krupp-Award for Young Professors
- Starting Grant (ERC-Backup Scheme of the SNSF)
- Horizon2020 Research and Innovation Action

<http://www.bsse.ethz.ch/mlcb>

## References

-  D. G. Grimm, *et al.*, *The Plant Cell* **29**, 5 (2017).
-  F. Llinares-López, *et al.* (ACM Press, 2015), pp. 725–734.
-  F. Llinares-López, *et al.*, *Bioinformatics* **31**, i240 (2015).
-  F. Llinares-López, *et al.*, *Bioinformatics (Oxford, England)* (2017).
-  L. Papaxanthos, *et al.*, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, *et al.*, eds. (2016), pp. 2271–2279.
-  M. Sugiyama, *et al.*, *SIAM Data Mining*, S. Venkatasubramanian, J. Ye, eds. (SIAM, 2015), pp. 37–45.

Icon source: Icons made by Freepik from [www.flaticon.com](http://www.flaticon.com) is licensed under CC BY 3.0