Machine Learning in Medicine: Combinatorial Association Mapping

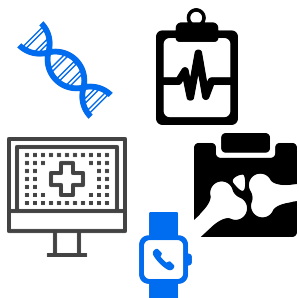
Karsten Borgwardt

ETH Zürich, Department Biosystems

Google Research Zürich, February 27, 2018

Machine Learning in Medicine

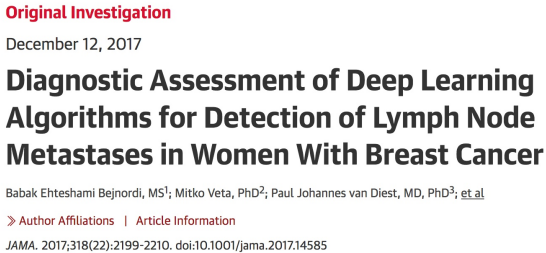
Key Topics



Machine Learning in Medicine

Key Topics

- Automation of diagnoses



Original Investigation

December 12, 2017

Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer

Babak Ehteshami Bejnordi, MS[1]; Mitko Veta, PhD[2]; Paul Johannes van Diest, MD, PhD[3]; et al

» Author Affiliations | Article Information

JAMA. 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585

Machine Learning in Medicine

Key Topics

- Automation of diagnoses
- Biomarker discovery

Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth

Marcel Adam Just, Lisa Pan, Vladimir L. Cherkassky, Dana L. McMakin, Christine Cha, Matthew K. Nock & David Brent

Nature Human Behaviour **1**, 911–919 (2017)
doi:10.1038/s41562-017-0234-y
Download Citation

Received: 06 February 2017
Accepted: 04 October 2017
Published online: 30 October 2017

Machine Learning in Medicine

Key Topics

- Automation of diagnoses
- Biomarker discovery

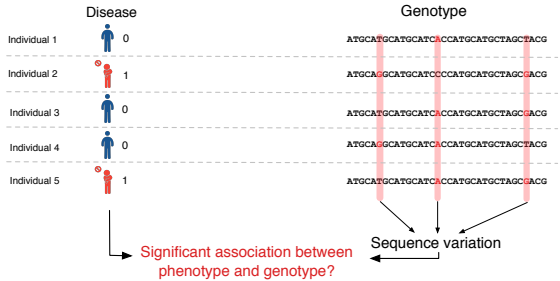
Roche to **buy Flatiron** Health for $1.9 billion to expand cancer care ...
Reuters - 15.02.2018
Roche to **buy Flatiron** Health for $1.9 billion to expand cancer care portfolio ... S) said on Thursday it would **buy** the rest of U.S. cancer data company **Flatiron** Health for $1.9 billion to speed development of cancer medicines and support its efforts to ... Privately held **Flatiron**, backed by Alphabet Inc (GOOGL.

- Biomedical data management

Machine Learning in Medicine

Key Topics

- Automation of diagnoses
- Biomarker discovery
  1. Algorithm development
  2. Applications with Biomedical Researchers
- Biomedical data management



Machine Learning in Medicine

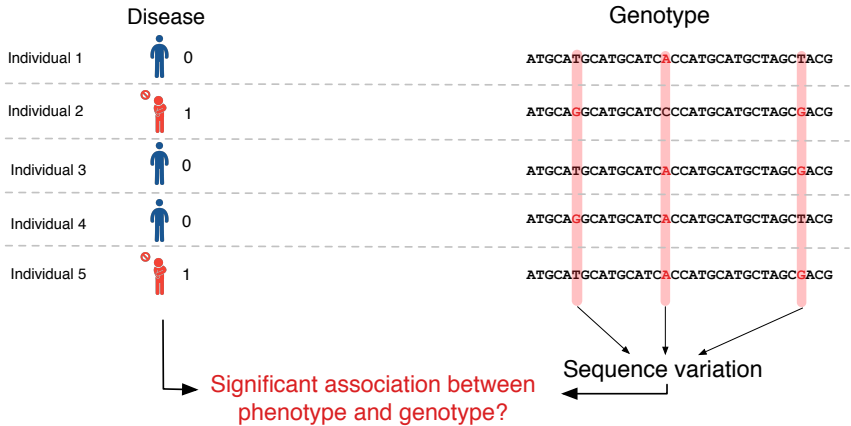
Key Topics

- Automation of diagnoses

- Biomarker discovery
  1. Algorithm development
  2. Applications with Biomedical Researchers
- Biomedical data management
  3. Software development for Biomedical Researchers

Algorithm Development for Biomarker Discovery

Association Mapping: 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.

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

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

Association Mapping: 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

Genome Medicine

COMMENT

Why epistasis is important for tackling complex human disease genetics

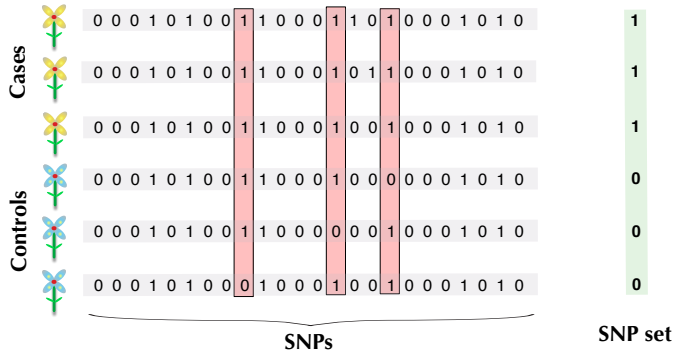
Trudy FC Mackay[1*] and Jason H Moore[2]

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

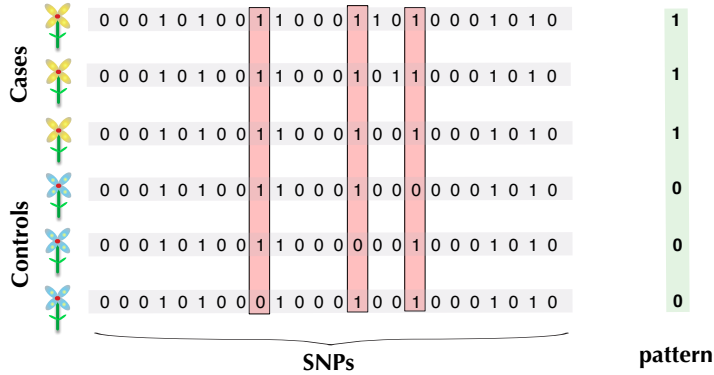
- What if we consider associations of groups of s SNPs with the phenotype?
- This leads to an enormous multiple testing problem: Any of the k SNP sets would correspond to a hypothesis that is tested ($k \in O(f^s)$), where f is the number of SNPs.
- If unaccounted for, α per cent of all SNP sets might 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 s 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 ($k \in O(f^s)$), where f is the number of SNPs.
- If unaccounted for, α 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**

Combinatorial Association Mapping as a Data Mining Problem

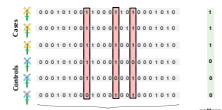


- Feature Selection: Find features that distinguish classes of objects
- Pattern Mining: Find higher-order **combinations of binary features**, so-called *patterns*, to distinguish one class from another

Combinatorial Association Mapping as a Data Mining Problem

Pattern

- D is a dataset of n patients. The i-th patient is represented by a binary vector $\mathbf{d}^{(i)} \in \{0,1\}^f$ and a class label $y_i \in \{0,1\}$.
- We choose a subset $\mathcal{S}$ of all features $\mathcal{F}$ in a dataset: $\mathcal{S} \subseteq \mathcal{F}$.
- Then an object $\mathbf{d}^{(i)}$ includes the pattern $\mathcal{S}$ if $\prod_{t \in \mathcal{S}} d^{(i)}(t) = 1$, otherwise not.



Problem Statement: Significant Pattern Mining

- We want to find all subsets $\mathcal{S}$ such that there is a statistically significant association between $\prod_{t \in \mathcal{S}} d^{(i)}(t)$ and y_i for $i \in \{1, \ldots, n\}$, while controlling the family-wise error rate at level α.

Significant Pattern Mining

Tarone's trick

- Contingency table for testing enrichment of a pattern in one of two classes

	Pattern present	Pattern absent	
y=0	a	$n_1 - a$	n_1
y=1	$x - a$	$n - n_1 - x + a$	$n - n_1$
	x	$n - x$	n

- A popular choice is Fisher's exact test to test whether the pattern 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

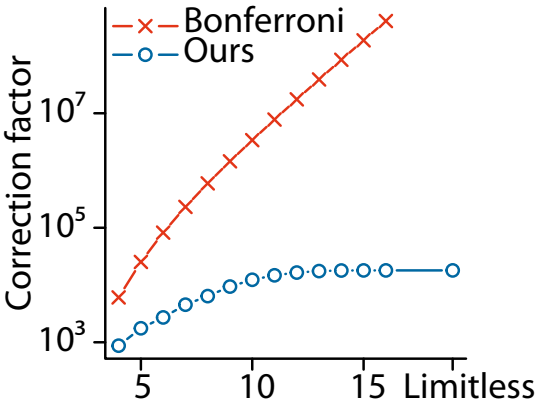
Tarone's trick

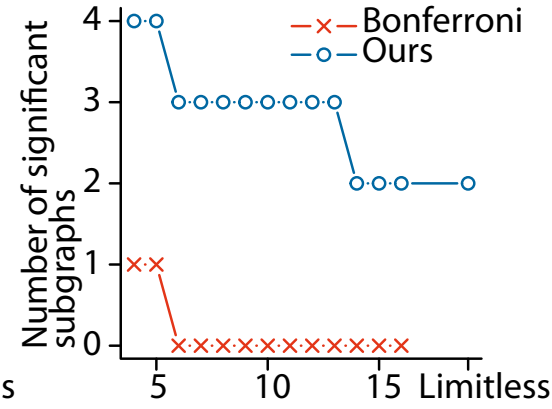
- Contingency table for testing enrichment of a pattern in a class

	Pattern present	Pattern absent	
y=0	a	$n_1 - a$	n_1
y=1	$x - a$	$n - n_1 - x + a$	$n - n_1$
	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 $m(k) << k$, 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 level $\frac{\alpha}{k}$.

procedure `Tarone`
  $k := 1$;
  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 level $\frac{\alpha}{k}$.

```
procedure Tarone
    k := 1;
    while k < m(k) do
        k := k + 1;
    return k
```

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

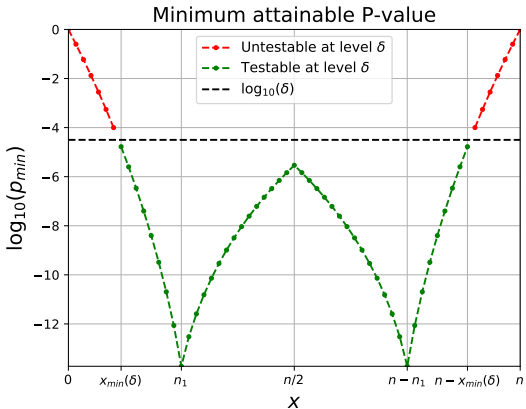
Significant Pattern Mining

Data mining challenge

- How to efficiently find $m(k)$ without running through all $O(f^s)$ 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):
  - `frequent itemset mining`(D, θ) enumerates all patterns in a dataset D of frequency at least θ.

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, α)
  $k := 1$;
  while $k < m(k)$ **do**
    $k := k + 1$;
    $m(k) :=$ frequent itemset mining$(D, \phi(\frac{\alpha}{k}))$;
  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, α)
  $k := 1$;
  while $k < m(k)$ **do**
    $k := k + 1$;
    $m(k) :=$ frequent itemset mining$(D, \phi(\frac{\alpha}{k}))$;
  return k

- Note: $\phi(\frac{\alpha}{k})$ is the minimum frequency of a pattern that is testable at level $\frac{\alpha}{k}$.

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, α)
  $k := 1$;
  while $k < m(k)$ **do**
    $k := k + 1$;
    $m(k) :=$ frequent itemset mining$(D, \phi(\frac{\alpha}{k}))$;
  return k

- Note: $\phi(\frac{\alpha}{k})$ is the minimum frequency of a pattern that is testable at level $\frac{\alpha}{k}$.
- For small k, $\phi(\frac{\alpha}{k})$ is small. Frequent itemset mining will be extremely expensive!

Starting Grant: Significant Pattern Mining

Contributions

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

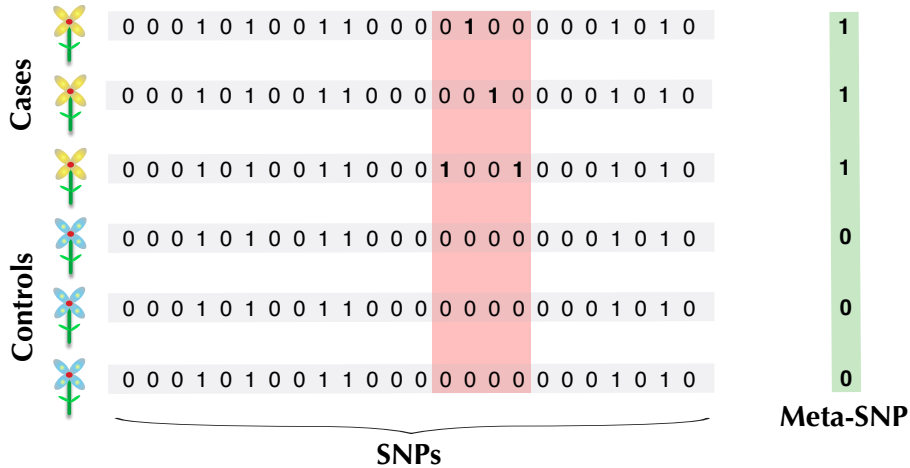
Starting Grant: Significant Pattern Mining

Contributions

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

Applications of Combinatorial Association Mapping: Genetic Heterogeneity Discovery

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

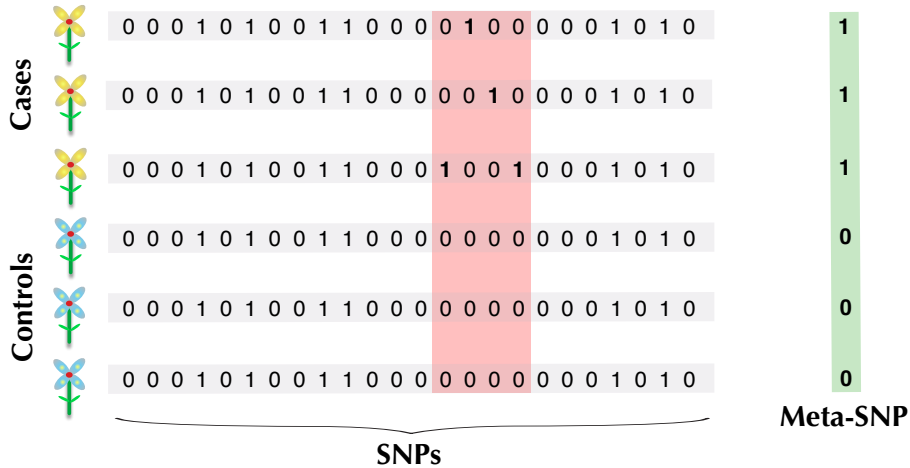


FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

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 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



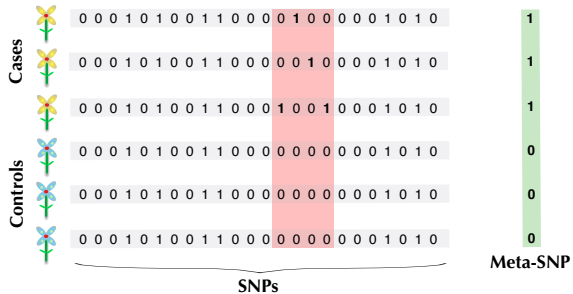
FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

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.

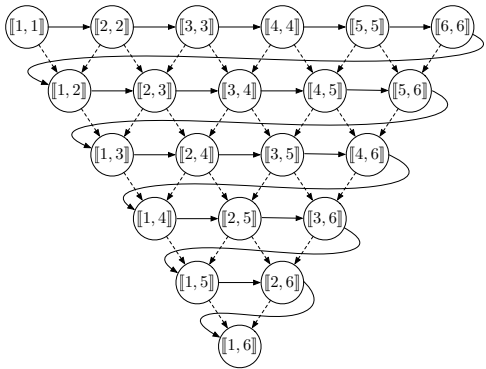
FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

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



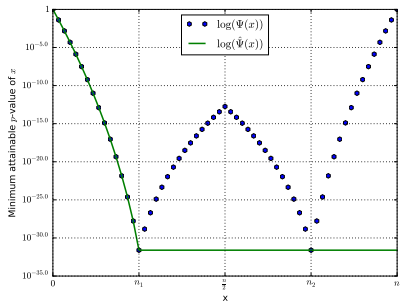
- An interval is represented by its maximum value. The longer an interval, the more likely it is that this maximum is 1.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



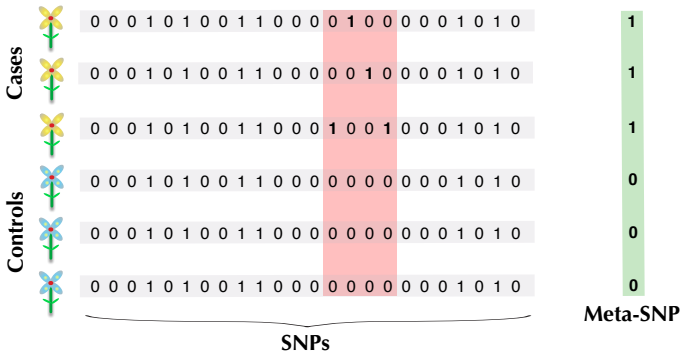
- **Search strategy:** We search intervals of increasing length and prune untestable superintervals.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



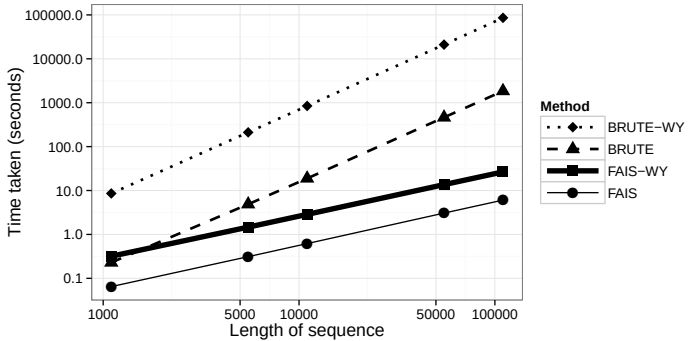
- **Pruning criterion 1:** If too many individuals have a particular pattern, the corresponding interval is not testable.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



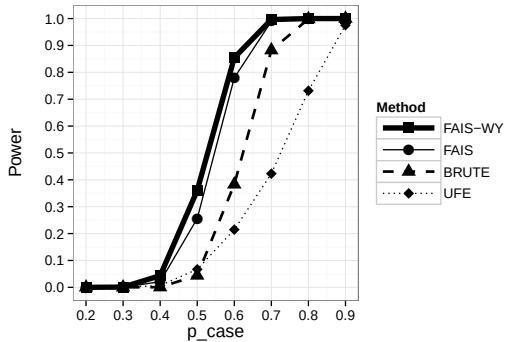
- **Pruning criterion 2:** If a pattern is too frequent to be testable, then none of the superintervals of the corresponding interval is testable.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



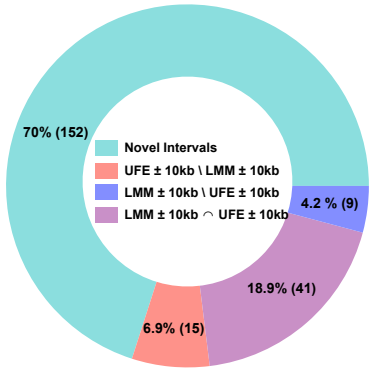
- Our method FAIS (Fast Automatic Interval Search) improves over the brute-force interval search in terms of runtime in simulations.

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.

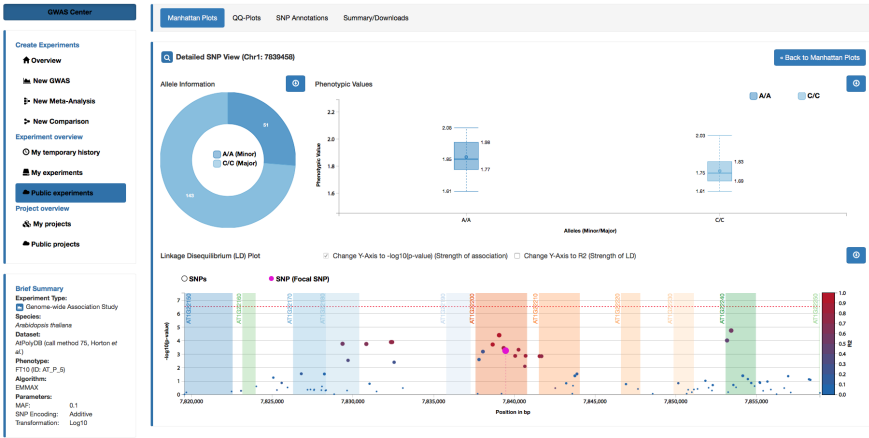
Next steps: Genetic heterogeneity discovery

- How to account for covariates like age and gender?
  - Solution for categorial covariates (NIPS 2016, Bioinformatics 2017)
  - Study in collaboration with the COPDGene Consortium to detect intervals associated with Chronic Obstructive Pulmonary Disease (Bioinformatics 2017).

Biomedical Software Development

easyGWAS

- We have been developing easygwas.org (Grimm et al., 2017), a cloud platform for genome-wide association studies (1248 users as of February 23, 2018):



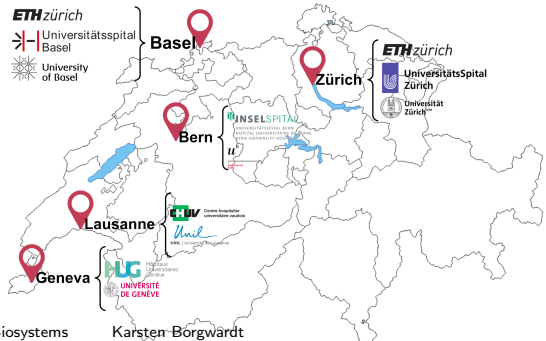
What's next?

Personalized Swiss Sepsis Study

- Consortium of 22 research labs and 5 university hospitals in Switzerland
- Goal: Predict sepsis and sepsis-related mortality
- Approach: Integrate clinical data and molecular data for joint biomarker discovery

Adrian Egli
PI SPHN
Clinical Microbiology, University Hospital Basel

Karsten Borgwardt
PI PHRT
MLCB, D-BSSE, ETH Zürich



- Duration:
  3 years
  (2018-2021)
- Total funding:
  5.3 Million CHF

Machine Learning in Medicine

Summary

- At the heart of Machine Learning in Medicine is the development of novel algorithms for biomarker discovery, such as the Combinatorial Association Mapping approaches presented here.
- The high dimensionality of the problem leads to an enormous computational and statistical challenge.
- Solving both problems at the same time was largely unachieved.
- We have developed several Significant Pattern Mining approaches that achieve both.

`www.significant-patterns.org`

Machine Learning in Medicine

Outlook

- Medicine and the Life Sciences promise to become a central application domain for Machine Learning.
- The number of possible topics is vast, reaches from
  - algorithm development,
  - collaborations with biomedical researchers to
  - software development.

www.mlpm.eu

Thank you

- Marie-Curie-Initial Training Network for 'Machine Learning for Personalized Medicine' (mlpm.eu, 2013-2016)
- Starting Grant (ERC-Backup Scheme of the SNSF)
- Alfried-Krupp-Award for Young Professors
- SPHN-PHRT Driver Project 'Personalized Swiss Sepsis Study'

`http://www.bsse.ethz.ch/mlcb`

References

 F. Llinares-López, *et al.*, *KDD* (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.*, *NIPS*, 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.