

Significant Pattern Mining: Efficient Algorithms and Biomedical Applications

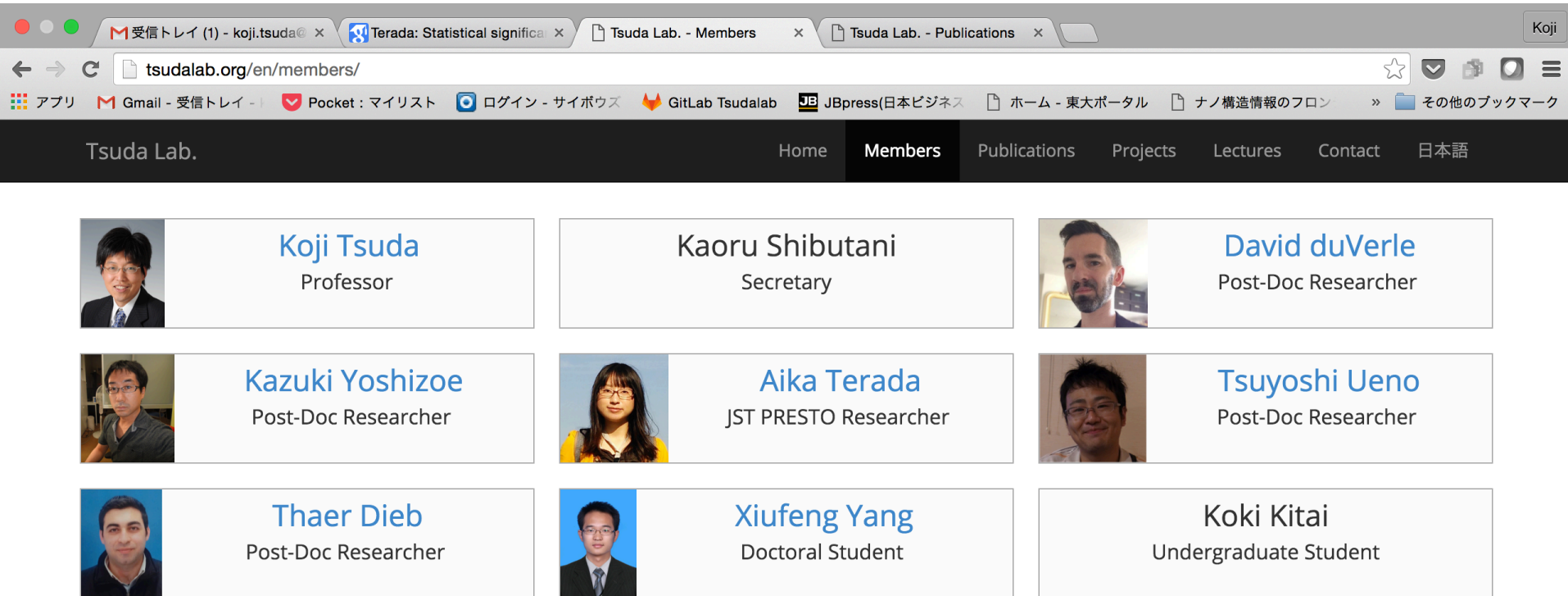
Koji Tsuda

Department of Computational Biology and
Medical Sciences
Graduate School of Frontier Sciences
University of Tokyo





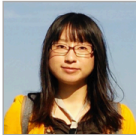




Kashiwa Campus, University of Tokyo



http://www.tsudalab.org/ (since 2014)



The screenshot shows a web browser window with the URL <http://www.tsudalab.org/en/members/>. The browser's address bar and tabs are visible at the top. The website's header includes the text "Tsuda Lab." and a navigation menu with links: "Home", "Members" (which is highlighted), "Publications", "Projects", "Lectures", "Contact", and "日本語". Below the header, the members are listed in a grid. Each member's entry consists of a small portrait photo, their name in blue text, and their title in black text.

 Koji Tsuda Professor	 Kaoru Shibutani Secretary	 David duVerle Post-Doc Researcher
 Kazuki Yoshizoe Post-Doc Researcher	 Aika Terada JST PRESTO Researcher	 Tsuyoshi Ueno Post-Doc Researcher
 Thaer Dieb Post-Doc Researcher	 Xiufeng Yang Doctoral Student	 Koki Kitai Undergraduate Student

FT Magazine

[Home](#)[World ▾](#)[Companies ▾](#)[Markets ▾](#)[Global Economy ▾](#)[Lex ▾](#)[Arts ▾](#)[Magazine](#)[Food & Drink ▾](#)[House & Home ▾](#)[Lunch with FT](#)[Style](#)[Books ▾](#)[Pursuits ▾](#)

March 28, 2014 11:38 am

Big data: are we making a big mistake?

By Tim Harford

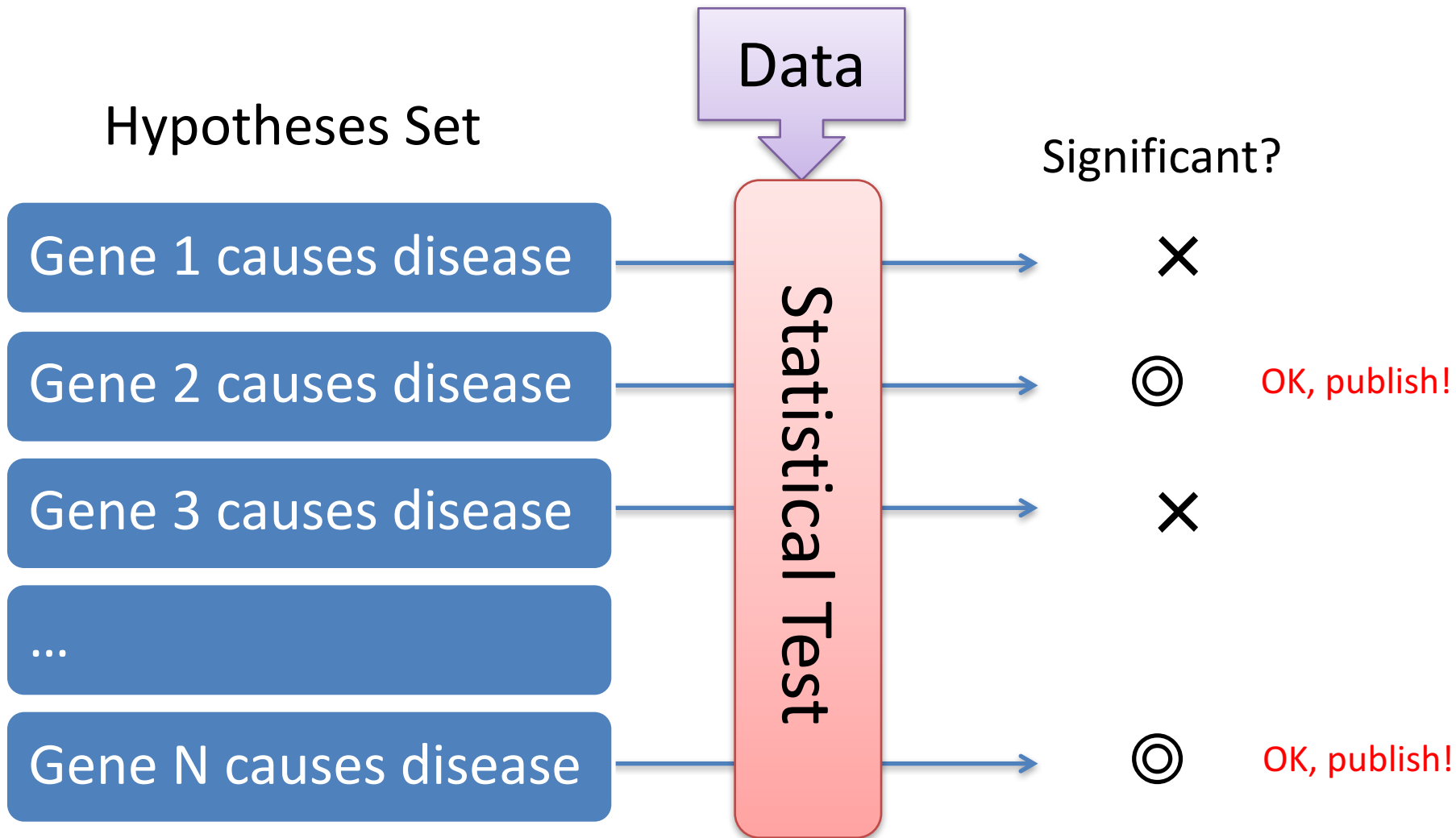
Big data is a vague term for a massive phenomenon that has rapidly become an obsession with entrepreneurs, scientists, governments and the media



Quotes

- In 2005, John Ioannidis, an epidemiologist, published a research paper with the self-explanatory title, “Why Most Published Research Findings Are False”. The paper became famous as a provocative diagnosis of a serious issue. One of the key ideas behind Ioannidis’s work is what statisticians call the “multiple-comparisons problem”.

Publishing system in life sciences



Reproducibility Crisis!

(and statistics is to blame)

- Biological results reported in journals cannot be reproduced
 - Bayer: could not reproduce 43 of 67 studies
 - Amgen: could not reproduce 47 of 53 studies



J. Ioannidis (Stanford)

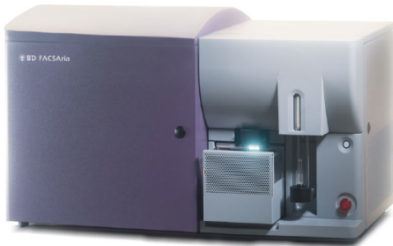
Data that are multidimensional (ie contain many features) are particularly at risk of **false positives** and overfitting, particularly when analyzed by inexperienced or untrained analysts.

(Lancet, 2014)

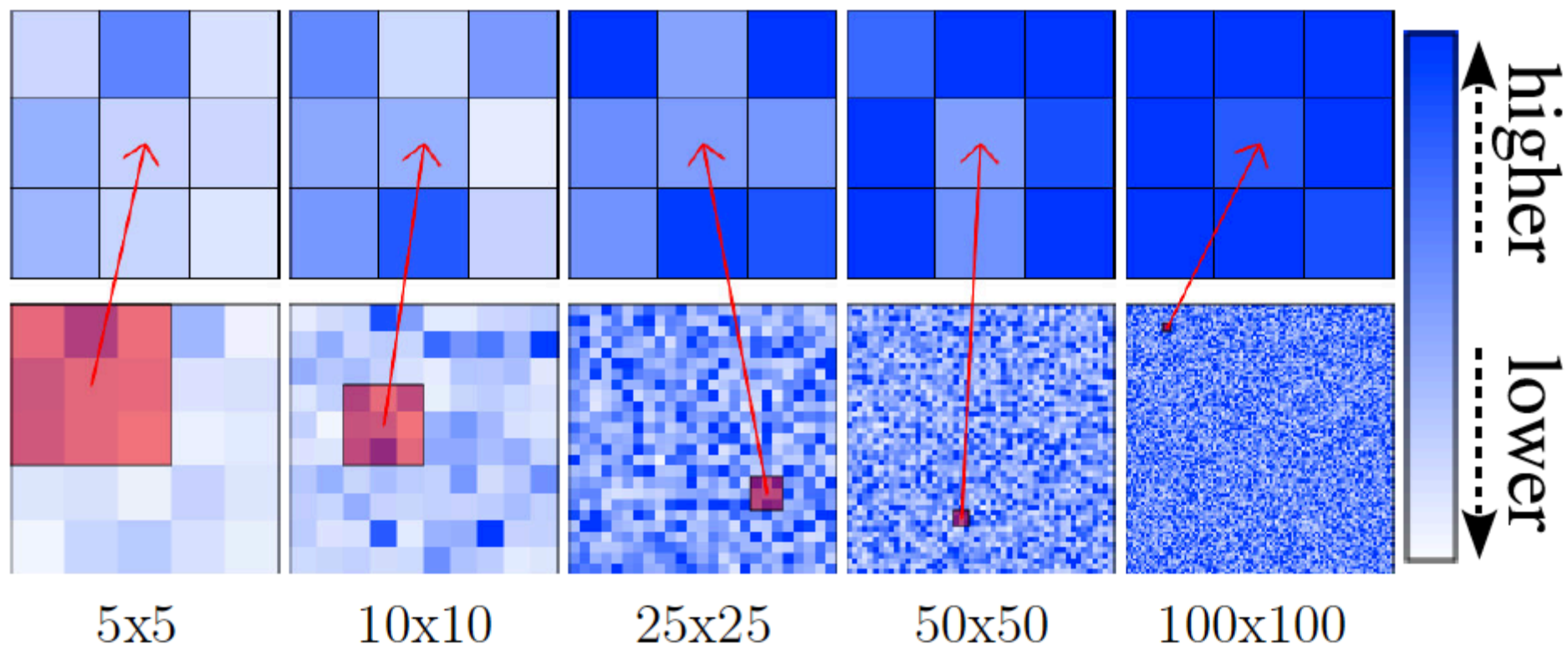
Curse of Dimensionality in Testing



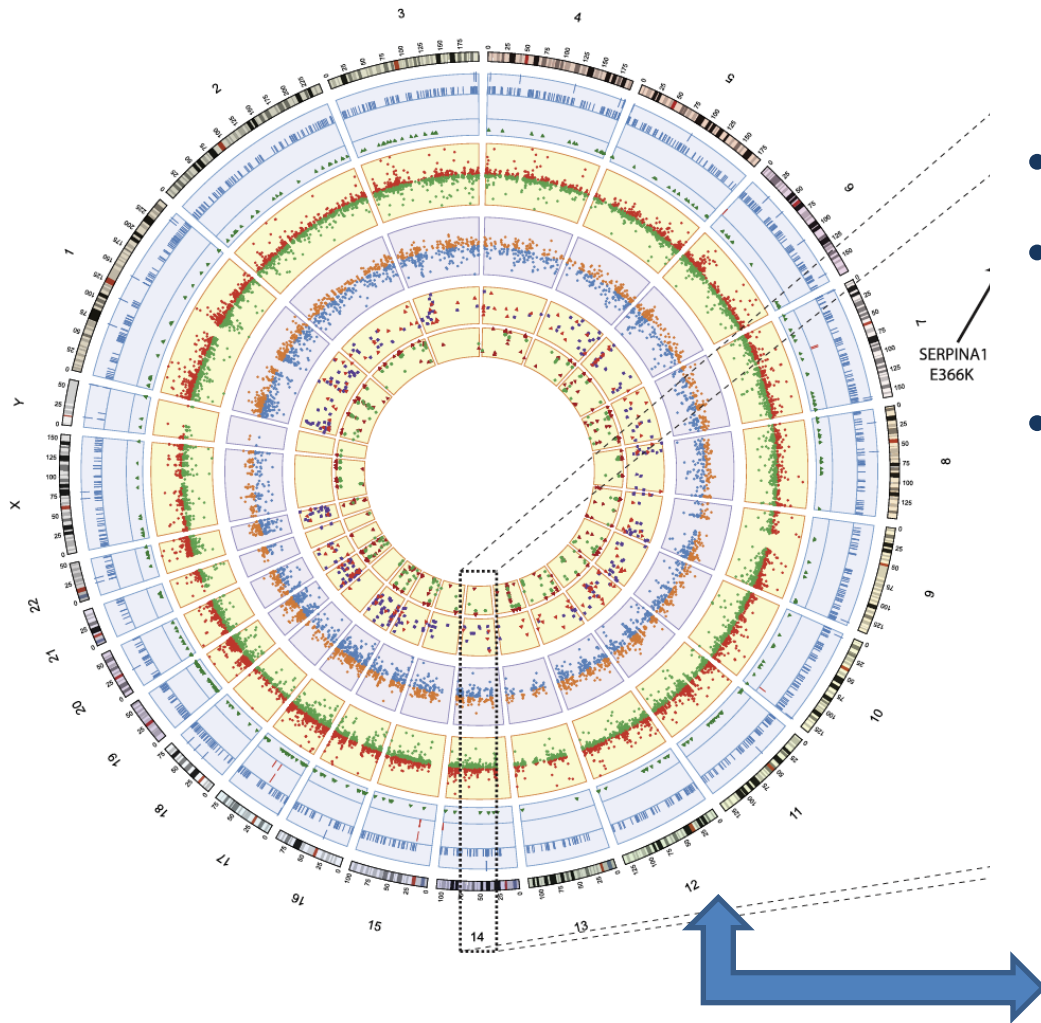
Huge increase in explanatory variables
No increase in examples



False positive more likely
→ Have to apply stricter criterion
→ Fewer discovery (!)



Trans-omics Data

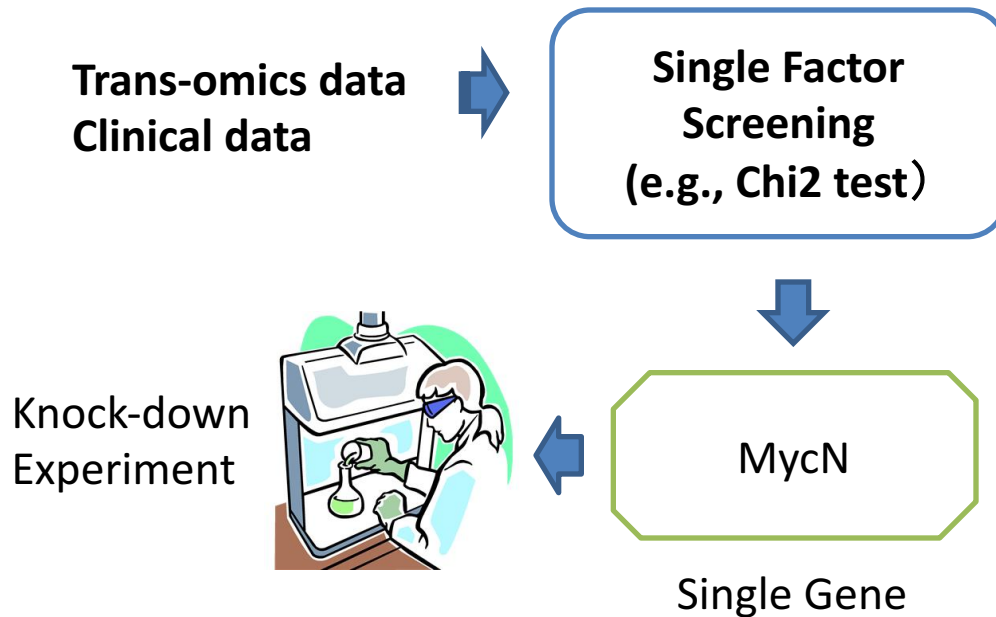


- **DNA** (mutation, insertion, deletion, CNV etc)
- **DNA methylation, Histon modification**
- **mRNA expression, ncRNA**
- **Protein expression, modification**
- **Metabolite** (Sugar, Amino acids, Nucleotides, lipids)

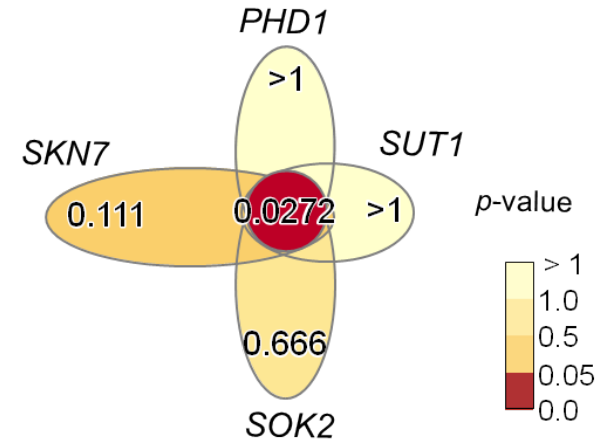
Clinical Data
Survival rate, Drug resistance, Relapse, Family history

Drawbacks of “Single Factor Screening”

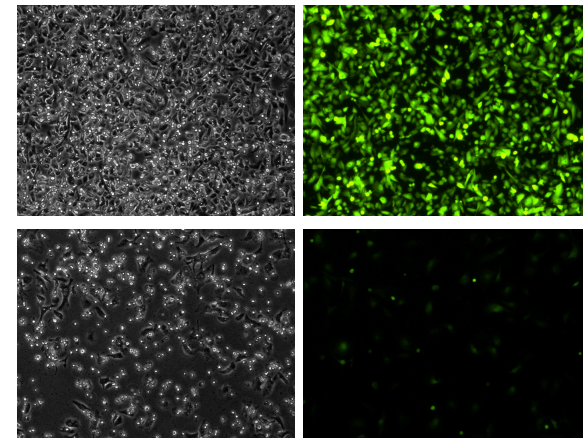
- Discover single factor causing phenotype (e.g., disease)
- BUT cellular processes are highly combinatorial



Single factor screening misses combinatorial causes



Knock down Experiments

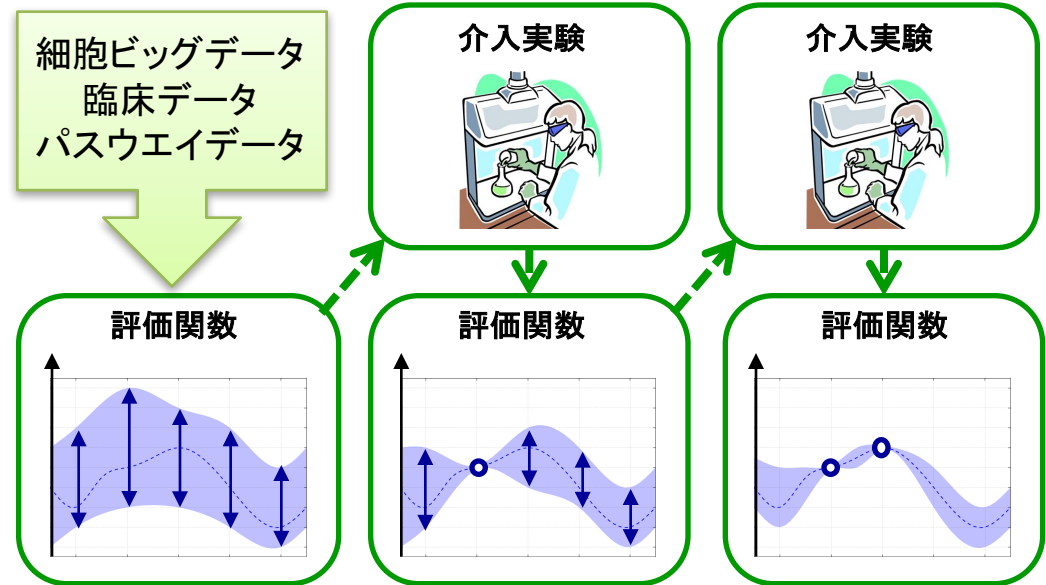
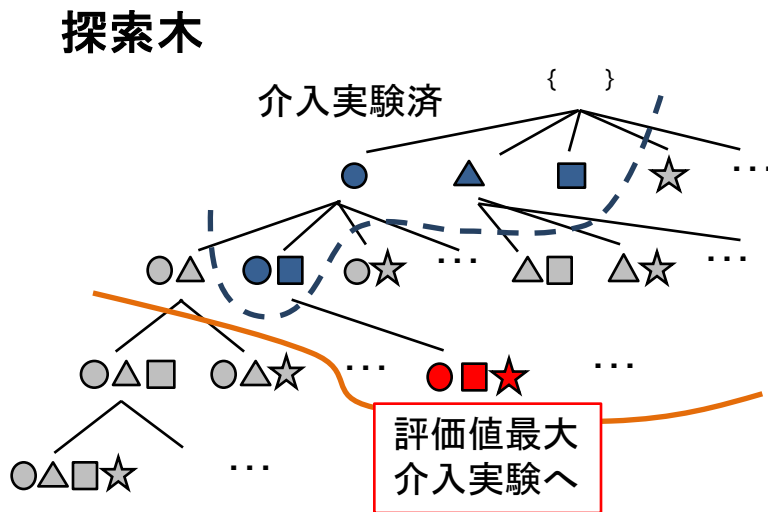


Challenge:

Discovering Combinatorial Factors Associated with Biological Phenomena

- **Combinatorial Explosion**

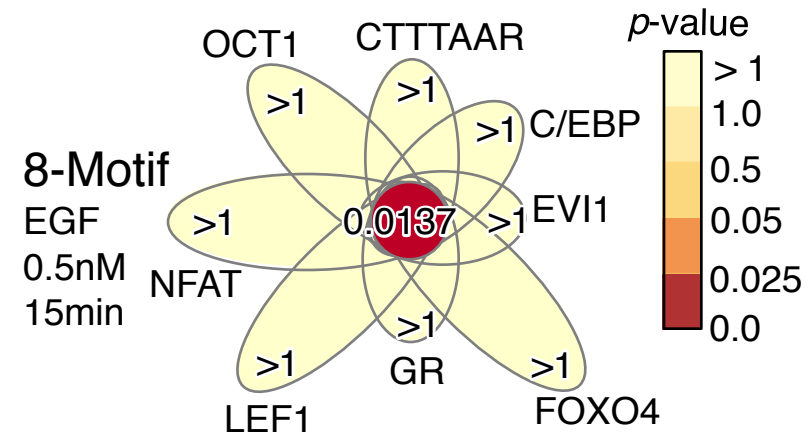
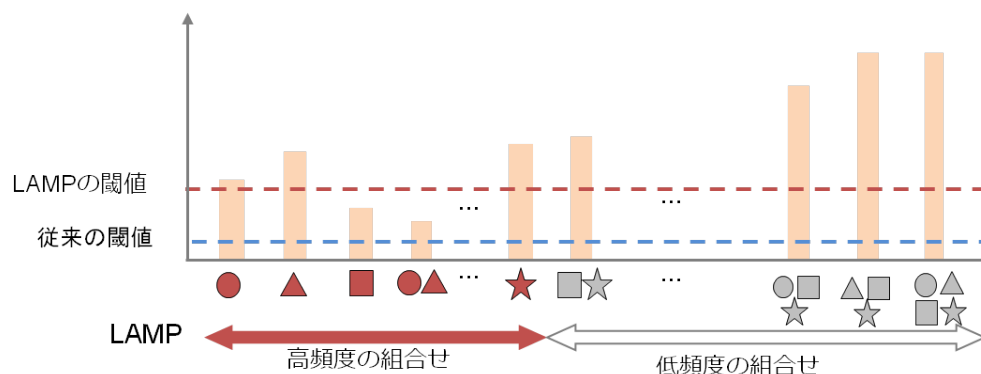
- 100m SNP x 10,000 Expression x 10,000 CNV = 100 trillion scores



Limitless Arity Multiple testing Procedure (LAMP)

Terada, Okada-Hatakeyama, Tsuda and Sese, PNAS, 2013

- Reliability of scientific discovery is assessed by P-values
- Multiple test (**Bonferroni**): If n candidates are available, use $0.05/n$ as significance level
- Number of combinatorial factors is huge: No chance of discovery
- Reduce the Bonferroni factor dramatically by itemset mining-based algorithm



Talk Agenda

- Theory of LAMP
 - Set Bonferroni factor to the number of “testable patterns” (Terada et al., PNAS 2013)
- Efficient Algorithms of LAMP
 - Support increase algorithm (Minato et al., ECMLPKDD 2014)
 - Parallel implementation for cloud platforms (NEW)
- Selective inference in pattern mining
 - P-value conditional on selection event (Taylor and Tibshirani, PNAS 2015)

Limitless Arity Multiple testing Procedure (LAMP)

Terada, Okada-Hatakeyama, Tsuda and Sese, PNAS, 2013

Statistical significance of combinatorial regulations

Aika Terada^{a,b,c}, Mariko Okada-Hatakeyama^d, Koji Tsuda^{c,e,1}, and Jun Sese^{a,b,1}

^aDepartment of Computer Science and ^bEducation Academy of Computational Life Sciences, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan; ^cMinato Discrete Structure Manipulation System Project, Exploratory Research for Advanced Technology, Japan Science and Technology Agency, Sapporo, Hokkaido 060-0814, Japan; ^dLaboratory for Integrated Cellular Systems, RIKEN Center for Integrated Medical Sciences (IMS-RCAI), Yokohama, Kanagawa 230-0045, Japan; and ^eComputational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Koto-ku, Tokyo 135-0064, Japan

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved July 3, 2013 (received for review February 4, 2013)

More than three transcription factors often work together to enable cells to respond to various signals. The detection of combinatorial regulation by multiple transcription factors, however, is not only computationally nontrivial but also extremely unlikely because of multiple testing correction. The exponential growth in the number of tests forces us to set a strict limit on the maximum arity. Here, we propose an efficient branch-and-bound algorithm called the “limitless arity multiple-testing procedure” (LAMP) to count the exact number of testable combinations and calibrate the Bonferroni factor to the smallest possible value. LAMP lists significant combinations without any limit, whereas the family-wise error rate is rigorously controlled under the threshold. In the human breast cancer transcriptome, LAMP discovered statistically significant combinations of as many as eight binding motifs. This method may contribute to uncover pathways regulated in a coordinated fashion and find hidden associations in heterogeneous data.

deliberately excluding such tests. Here, we propose an efficient branch-and-bound algorithm, called the “limitless arity multiple-testing procedure” (LAMP). LAMP counts the exact number of “testable” motif combinations and derives a tighter bound of FWER, which allows the calibration of the Bonferroni factor as the FWER is controlled rigorously under the threshold.

In comparison with existing methods that can find only two-motif combinations, our testing procedure may contribute to finding larger fractions of regulatory pathways and TF complexes, thus providing more concrete evidence for further investigation. In legacy yeast expression data (29), a four-motif combination corresponding to a known pathway was found using LAMP, whereas only two motifs in the combination had been predicted using the existing method. When applied to human breast cancer transcriptome data (30), combinations of up to eight motifs were found to be statistically significant.

Results

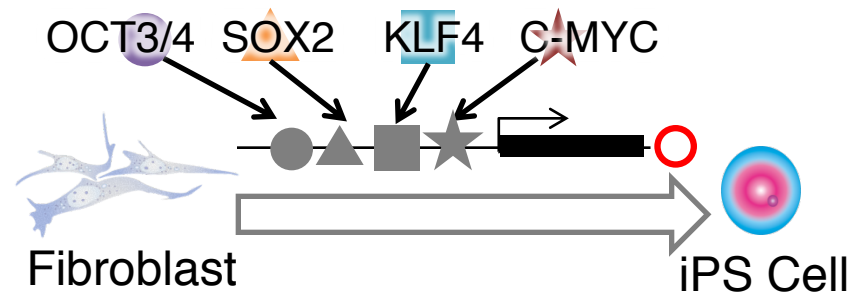
Method Overview. To present our strategy for combinatorial regu-

Transcription factors (TFs) work in combination

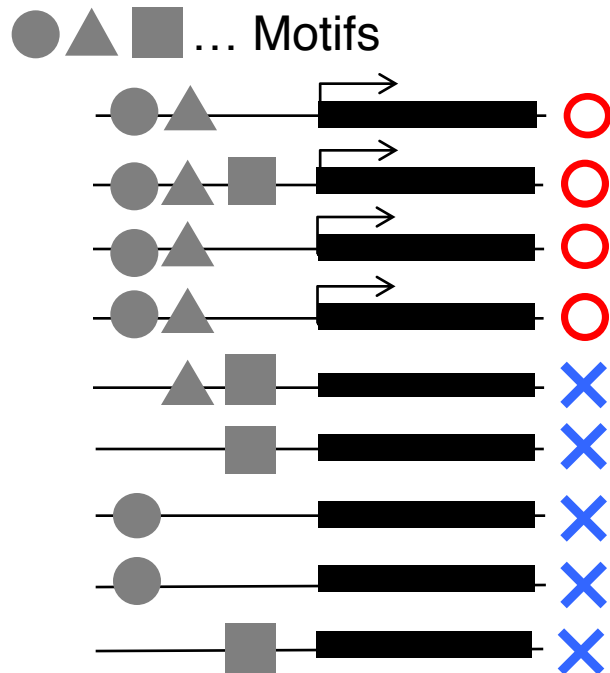
- Often several TFs are necessary to induce the expression of downstream genes



Example: Yamanaka Factor (K. Okita *et al.*, Nature, 2007)



Find statistically significant combinations of TF binding motifs



Contingency table for ●▲

	Up-regulated	No-regulated
With Motif Combination	4	0
Without	0	5

P-value by Fisher exact test
0.0079

Significant?







No – You have to apply multiple testing procedure

Bonferroni Correction

- Family-wise error rate(FWER)
 - At least one false discovery occurs
- P-value threshold δ is determined such that FWER is below α
- For m tests,

$$\delta = \frac{\alpha}{m}$$

- 100 motifs in total
- Number of tests

			...	100
			...	4,950
Total				5,050

- Corrected threshold
$$\delta = 0.05/5050$$
$$= 9.9 \times 10^{-6}$$
- Bonferroni is too conservative!

New Proposal:

Limitless Arity Multiple testing Procedure (LAMP)

- Count the exact number of “testable” combinations
 - Infrequent combinations do not affect family-wise error rate
 - Stepwise procedure involving itemset mining
- Calibrate the correction factor to the smallest possible value

Raw p-value

	Up regulated	No regulated
With Motif Combination	a	b
Without	c	d

- Null Hypothesis H
 - Two variables are independent
- P-value: $p(a,b,c,d)$
 - Probability of observing stronger table than observed
 - If smaller than α , reject H (discovery!)
- Type-I error: reject H when it is true
- Probability of type-I error must satisfy

$$P(p < \alpha \mid H) \leq \alpha$$

Multiple Tests

- m null hypotheses H_1, \dots, H_m
- V : Number of rejections in m tests
- Probability that more than one type-I error occurs: Family-wise error rate (FWER)

$$P(V > 0 \mid \bigcap_{i=1}^m H_i)$$


- Multiple testing procedures aim to control FWER under α

Bonferroni Correction

- Given threshold δ , FWER is bounded as

$$\begin{aligned} P(V > 0 \mid \bigcap_{i=1}^m H_i) &\leq \sum_{i=1}^m P(p_i \leq \delta \mid H_i) && \text{Union bound} \\ &\leq m\delta && \text{Definition of p-value} \end{aligned}$$

- Thus, setting $\delta = \alpha/m$ calibrate FWER bound to α

	Up-regulated	Not regulated	
With Motif Combination	a	b	x 
Without	c	d	N-x
	n_u	$N-n_u$	N

Occurrence Frequency (Support)

- P-value by Fisher exact test cannot be smaller than

$$f(x) = \frac{\binom{n_u}{x}}{\binom{N}{x}}$$

- No chance of false discovery, if $f(x) \geq \delta$

$$P(p < \delta \mid H) = 0$$

Tarone Correction (Biometrics, 1990)

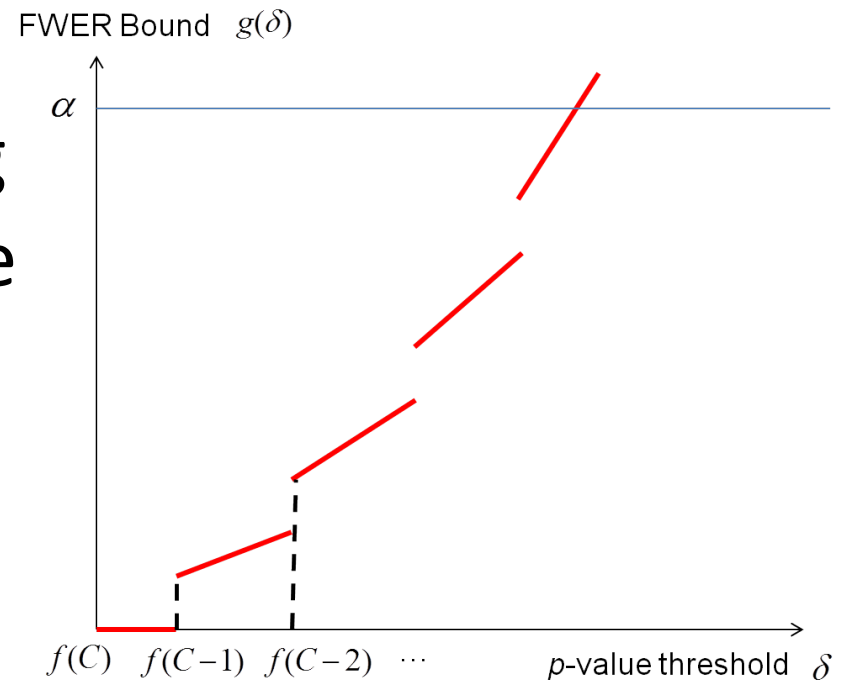
- Considering minimum p-value, FWER is bounded as follows

$$\begin{aligned} P(V > 0 \mid \bigcap_{i=1}^m H_i) &\leq \sum_{i=1}^m P(p_i \leq \delta \mid H_i) && \text{Union bound} \\ &= \sum_{\{i \mid f(x_i) \geq \delta\}} P(p_i \leq \delta \mid H_i) && \text{Use minimum p-value to remove hypotheses} \\ &\leq |\{i \mid f(x_i) \geq \delta\}| \delta && \text{Definition of p-value} \end{aligned}$$

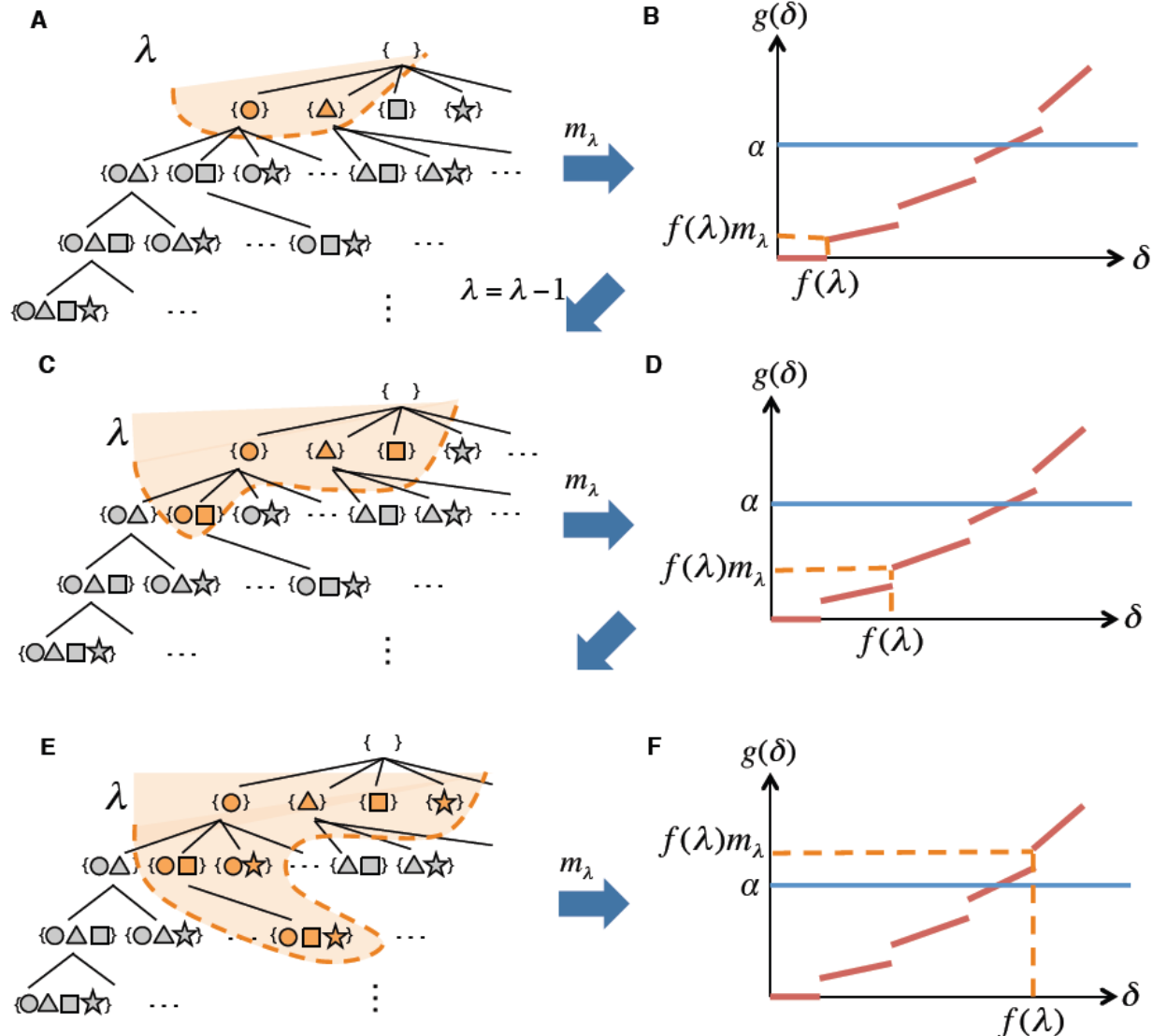
- Take maximum δ that keeps FWER bound below α

Finding optimal cut-off δ that calibrates FWER bound to α

- FWER bound is piecewise linear
- Repeat itemset mining with decrementing the frequency parameter
- A line segment drawn by a mining call
- Finish if line segment reaches α



Repeat itemset mining with decrementing support until all testable patterns are found



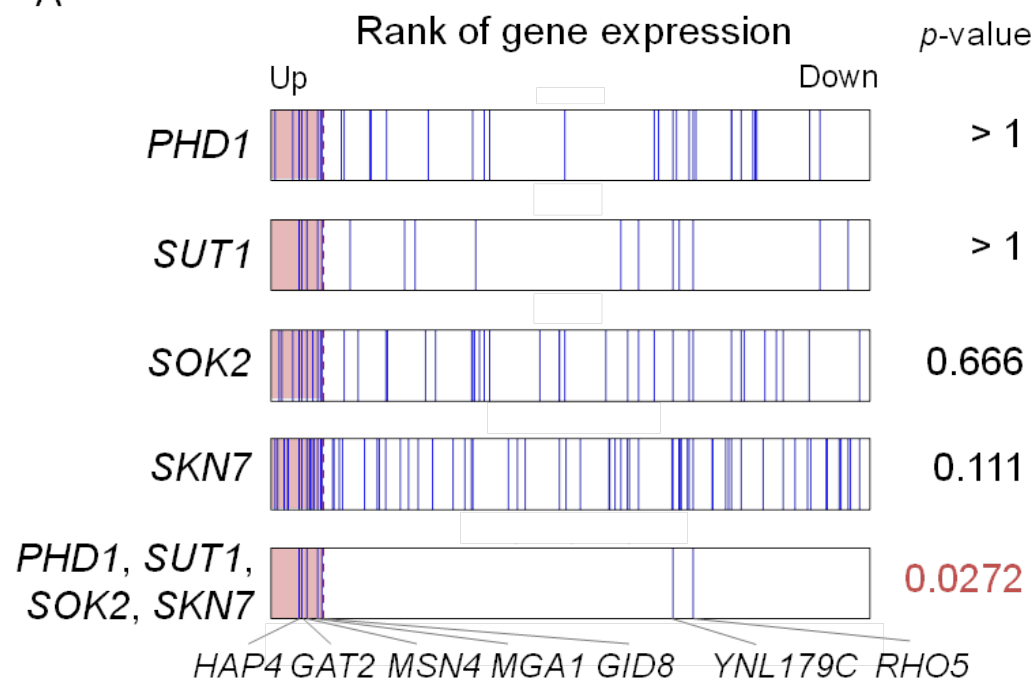
Statistically significant TF combinations under a heat shock condition (Yeast)

Corrected p-value (p-value*K)

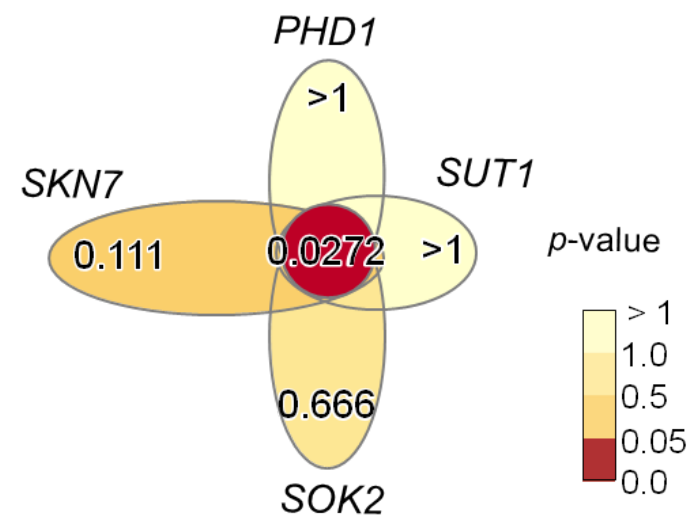
Combination	LAMP (≤ 102)	Bonferroni (≤ 4)
	K = 303	K = 4,426,528
HSF1	4.41E-24	6.44E-20
MSN2	3.73E-11	5.45E-07
MSN4	0.00053	> 1
SKO1	0.00839	> 1
SNT2	0.0192	> 1
PHD1, SUT1, SOK2, SKN7	0.0272	> 1

Red : significant

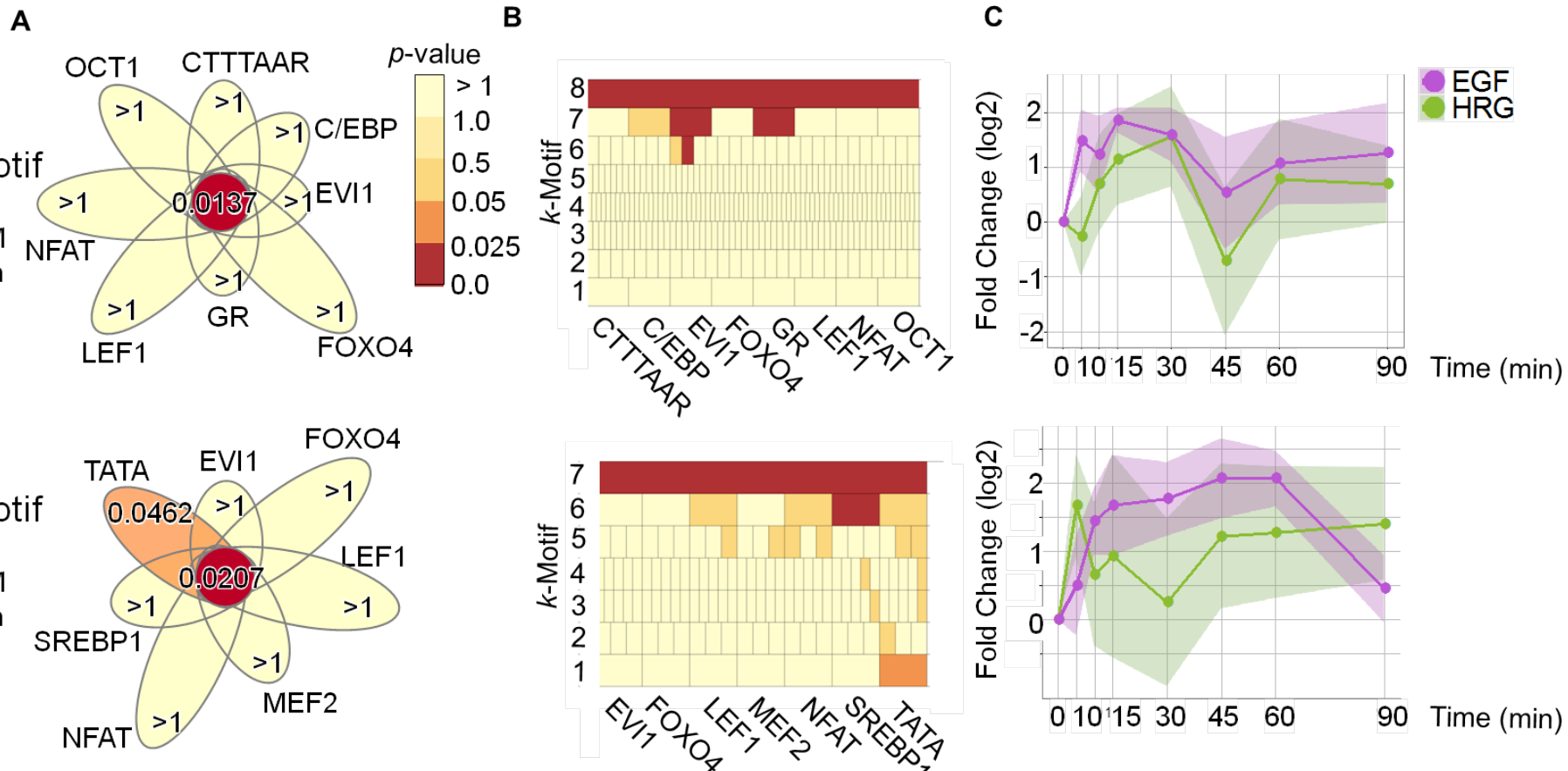
A



B



Application to MCF7 human breast cancer cells (GSE6462)



Fast Westfall-Young Methods

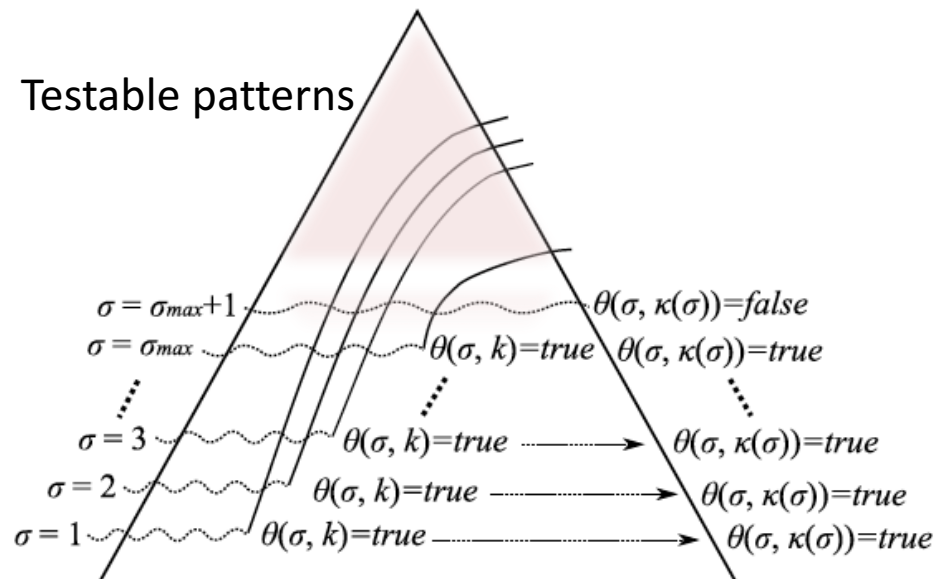
- A. Terada, K. Tsuda, and J. Sese. Fast Westfall-Young Permutation Procedure for Combinatorial Regulation Discovery, *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 153-158, 2013.
- Sugiyama, M., Llinares-López, F., Kasenburg, N., & Borgwardt, K. M. (2015). Significant subgraph mining with multiple testing correction. In *SIAM SDM* (pp. 37-45).
- Llinares-López, F., Sugiyama, M., Papaxanthos, L., & Borgwardt, K. (2015). Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 725-734). ACM.
- Llinares-López, F., Grimm, D. G., Bodenham, D. A., Gieraths, U., Sugiyama, M., Rowan, B., & Borgwardt, K. (2015). Genome-wide detection of intervals of genetic heterogeneity associated with complex traits. *Bioinformatics*, 31(12), i240-i249.

Development of LAMP implementations

- Stepwise search (Terada et al., PNAS 2013)
 - Repeats itemset mining with decrementing support
 - **SLOW**: up to several hundreds of variables
- Support increase algorithm (Minato et al., ECMLPKDD 2014)
 - Depth first search with incrementing support
 - 100x speed up
- Massively Parallel LAMP (**NEW**)
 - 1000 cores in cloud or on-premises
 - MPI: Work stealing and Reduce-Broadcast
 - <https://github.com/tsudalab/mp-lamp>

Support Increase Algorithm

- Depth first search of testable patterns
 - Deliberately overshoot by starting from a small support
 - Pruning based on **count table**: Number of closed patterns of different values of support



LAMPLINK (Terada et al., Bioinformatics 2016)

- LAMP implementation for genome-wide association study (GWAS) datasets
- Same input/output format as PLINK

Bioinformatics, 2016, 1–3

doi: 10.1093/bioinformatics/btw418

Advance Access Publication Date: 13 July 2016

Applications Note

OXFORD

Genetic and population analysis

LAMPLINK: detection of statistically significant SNP combinations from GWAS data

Aika Terada^{1,2,3,*}, Ryo Yamada⁴, Koji Tsuda^{2,3,5} and Jun Sese^{3,6,*}

¹PRESTO, Japan Science and Technology Agency, Saitama 332-0012, Japan, ²Department of Computational Biology and Medical Science, Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8561, Japan, ³Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan, ⁴Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, 606-8507 Japan, ⁵Center for Materials Research by Information Integration, NIMS, Ibaraki, 305-0047 Japan and ⁶Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

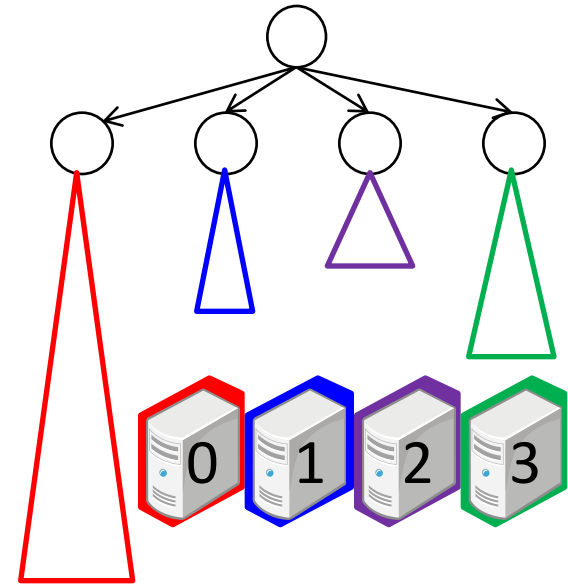
Received on February 22, 2016; revised on June 20, 2016; accepted on June 25, 2016

- Human exome data by 1000 Genomes Project
12758 SNPs and 697 individuals
- Japanese vs Non-Japanese
- 106 significant patterns (21 sec) by LAMPLINK

ID	SNP	Chr	Position (bp)	Gene	LAMPLINK	Adjusted p-value Bonferroni correction		
						≤ 3	≤ 4	≤ 5
1	rs34902660	6	25,850,874	<i>SLC17A3</i>	7.7695E-05	1	1	1
	rs2298091	6	26,158,211	<i>HIST1H2BD</i>				
	rs1150723	6	28,283,939	<i>PGBD1</i>				
2	rs2303080	5	7,878,311	<i>MTRR</i>	0.012638	NA	NA	1
	rs2287779	5	7,889,103	<i>MTRR</i>				
	rs2287780	5	7,889,191	<i>MTRR</i>				
	rs16879334	5	7,891,393	<i>MTRR</i>				
	rs3815990	12	121,253,285	<i>CAMKK2</i>				
3*	rs2472647	5	141,331,138	<i>PCDHGA1</i>	0.019122	1	1	1
	rs36012859	6	132,734,332	<i>VNN3</i>				
	rs17238245	15	61,951,918	<i>VPS13C</i>				
4*	rs79825658	3	57,508,536	<i>DNAH12</i>	0.019122	1	1	1
	rs2472647	5	141,331,138	<i>PCDHGA1</i>				
	rs17170011	7	34,827,570	<i>NPSR1</i>				

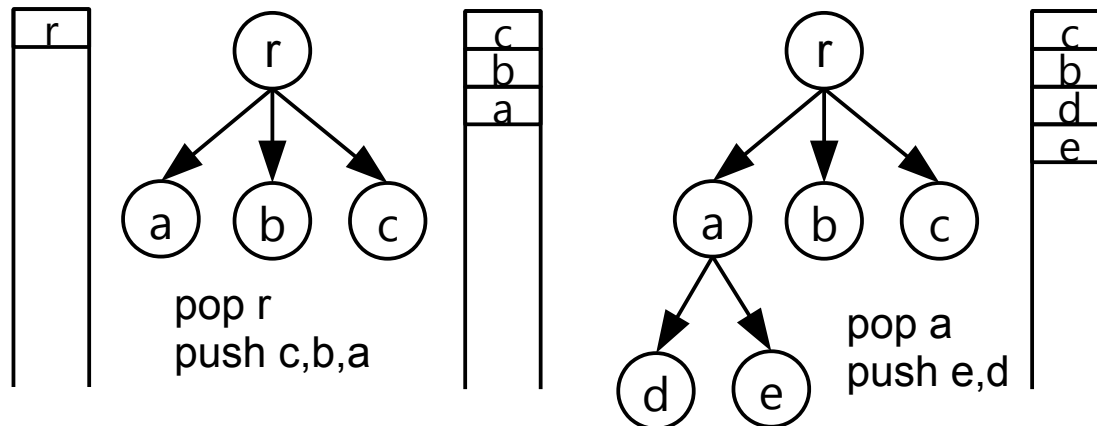
Parallel Implementation of LAMP

- Tree splitting is not good
- Dynamic load balancing is necessary
 - Passing tasks among computing nodes
- Count table must be shared



Stack-based depth first search

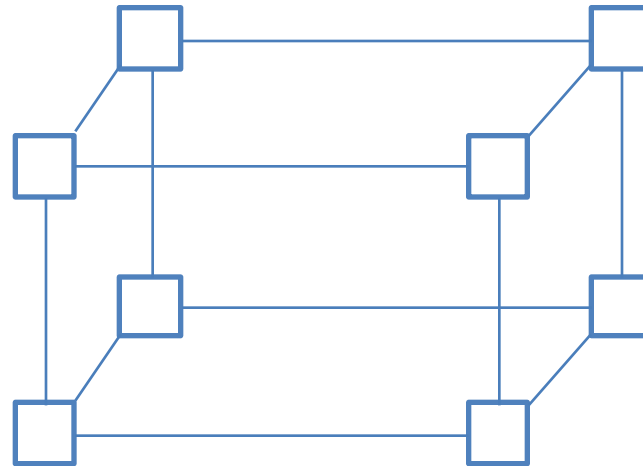
- Each worker starts exploring under an assigned first-level child
- **Each worker has a stack**
- Search by popping a node from stack and pushing its children back to the stack



Dynamic load balancing

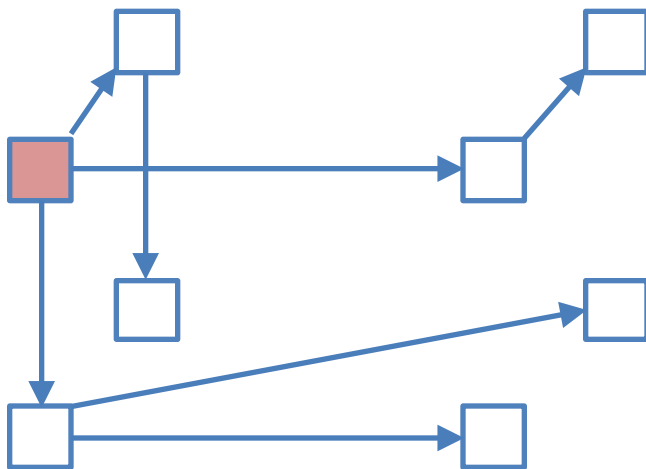
(Saraswat et al., 2011)

- Hypercube-like communication graph
- When the stack of a worker becomes empty, it sends requests to neighbors
- If a neighbor has nodes in stack, it gives a half
- If not, the worker quiesces after informing neighbors
- When neighbors get new tasks, the worker revives

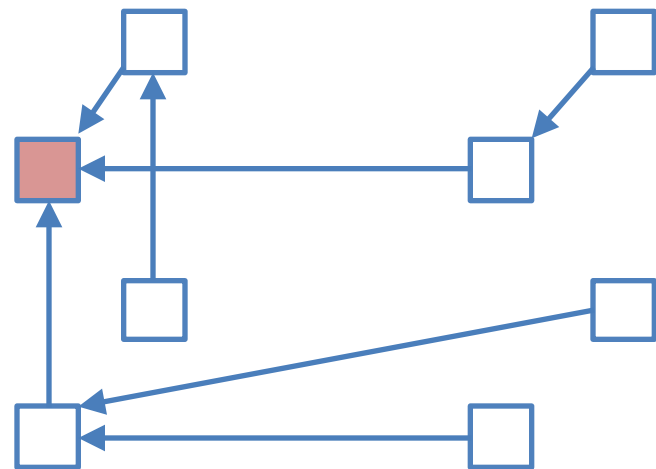


Sharing count table by reduce-broadcast

- Tree-shaped communication graph
- Root worker maintains the total count table
- Every 1ms, the root worker emits an *exploration message* including the total table. Every worker pass it to its children
- Leaf worker gives back an *echo message* including its count table to its parent.
- The parent sums up childrens' tables with its own and give it to its parent



Exploration messages



Echo messages

- eQTL study of Alzheimer disease: 114658 SNPs.
364 samples

■ **Combinations significantly associated with SOX10 expression level**

- SOX10 is an important regulator of neural crest and nervous system development [Kim 2003].

Combination	P-value	Freq.
rs5928731, rs12840385, rs12846200	4.69E-09	16
rs12840385, rs12846200	7.79E-09	18
rs5928731, rs12840385	3.17E-08	17
rs873275, rs6637756	3.36E-08	12

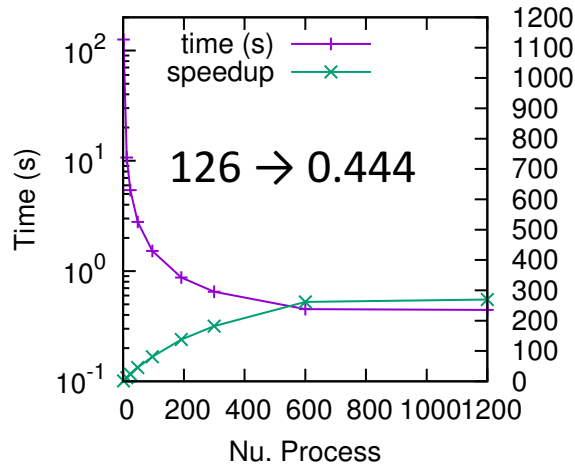
Combination	P-value	Freq.
rs2864894, rs17258147	3.36E-08	12
rs5928731, rs11795655 rs12840385, rs12846200	8.37E-08	14
rs11795655, rs12840385, rs12846200	1.27E-07	16

Adjusted significance level: 1.30E-07 (Correction factor: 384,708)

Speedup on a computer cluster (1,200 cores = 100 nodes)

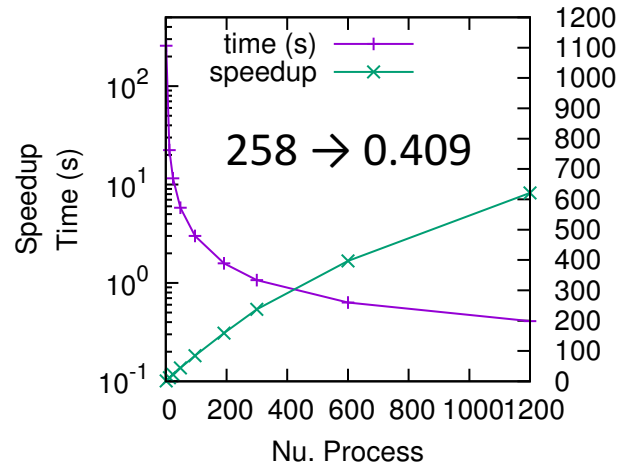
HapMap dom. 10%

item: 11253, trans:697, dens:1.0%



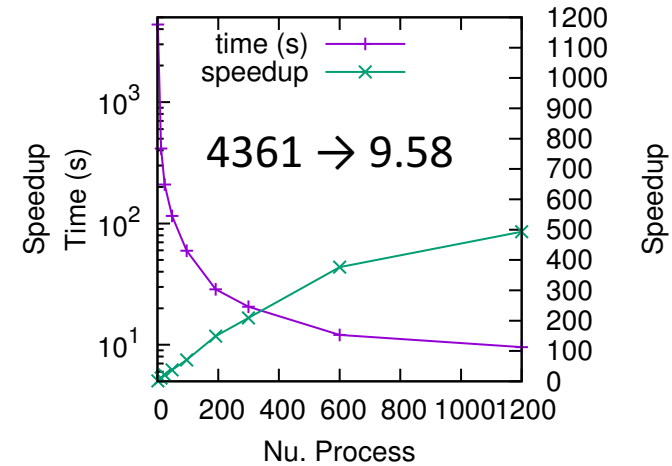
Alz. dom. 5%

item:44052, trans:364, dens:5.4%



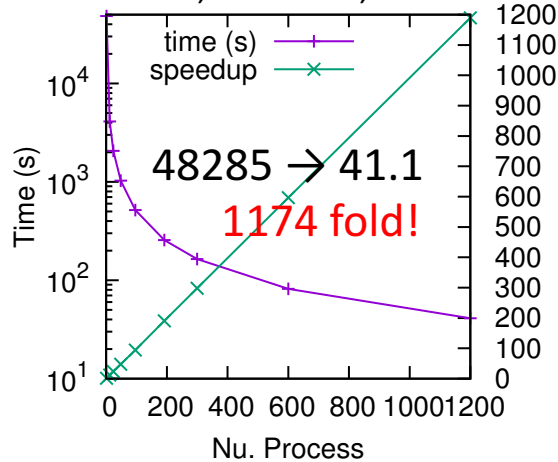
Alz. rec. 30%

item: 250,20, trans:364, dens:2.9%



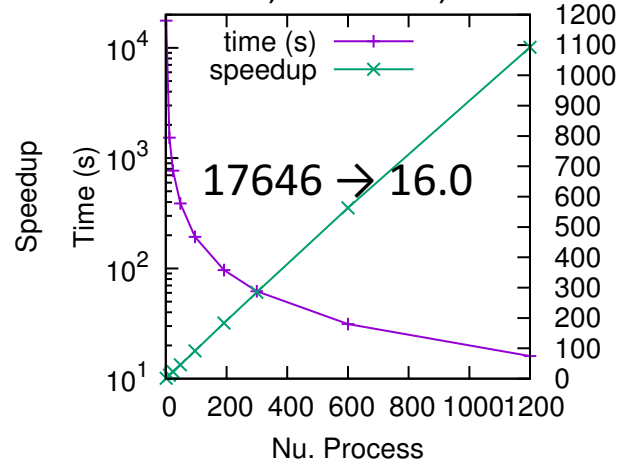
HapMap dom. 20%

item:11914, trans:697, dens:1.9%



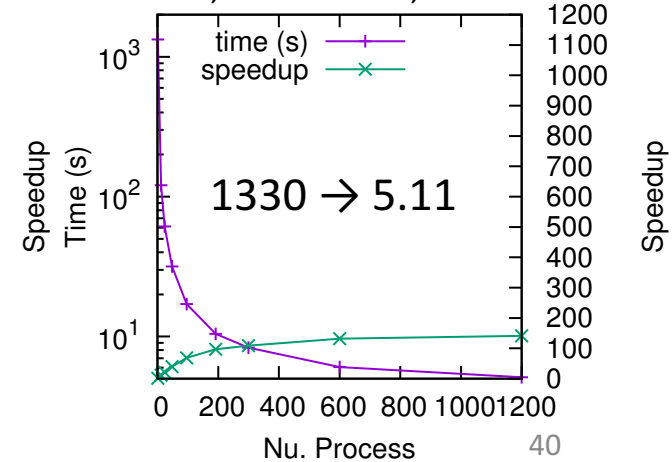
Alz. dom. 10%

item: 91126, trans: 364, dens:9.8%



MCF7

item:397, trans:12773, dens:2.9%



Comparison to Naïve Approach (Tree Splitting)

	HapMap dom.10%	HapMap dom.20%	Alz. dom.5%	Alz. dom.10%	Alz. rec.30%	MCF7
1 core	126	48285	258	17646	4361	1330
12 cores	10.7	4108	22.4	1535	415	121
48 cores	2.79	1029	5.8	387	115	31.7
Naive12	13.7	6559	24.1	3486	657	385
Naive48	7.26	3611	9.9	3480	398	387

Selective Inference



Statistical learning and selective inference

Jonathan Taylor^a and Robert J. Tibshirani^{b,1}

^aDepartment of Statistics, Stanford University, Stanford, CA 94305; and ^bDepartment of Health Research & Policy and Department of Statistics, Stanford University, Stanford, CA 94305

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2012.

Contributed by Robert J. Tibshirani, May 7, 2015 (sent for review March 2, 2015; reviewed by Rollin Brant and John D. Storey)

We describe the problem of “selective inference.” This addresses the following challenge: Having mined a set of data to find potential associations, how do we properly assess the strength of these associations? The fact that we have “cherry-picked”—searched for the strongest associations—means that we must set a higher bar for declaring significant the associations that we see. This challenge becomes more important in the era of big data and complex statistical modeling. The cherry tree (dataset) can be very large and the tools for cherry picking (statistical learning methods) are now very sophisticated. We describe some recent new developments in selective inference and illustrate their use in forward stepwise regression, the lasso, and principal components analysis.

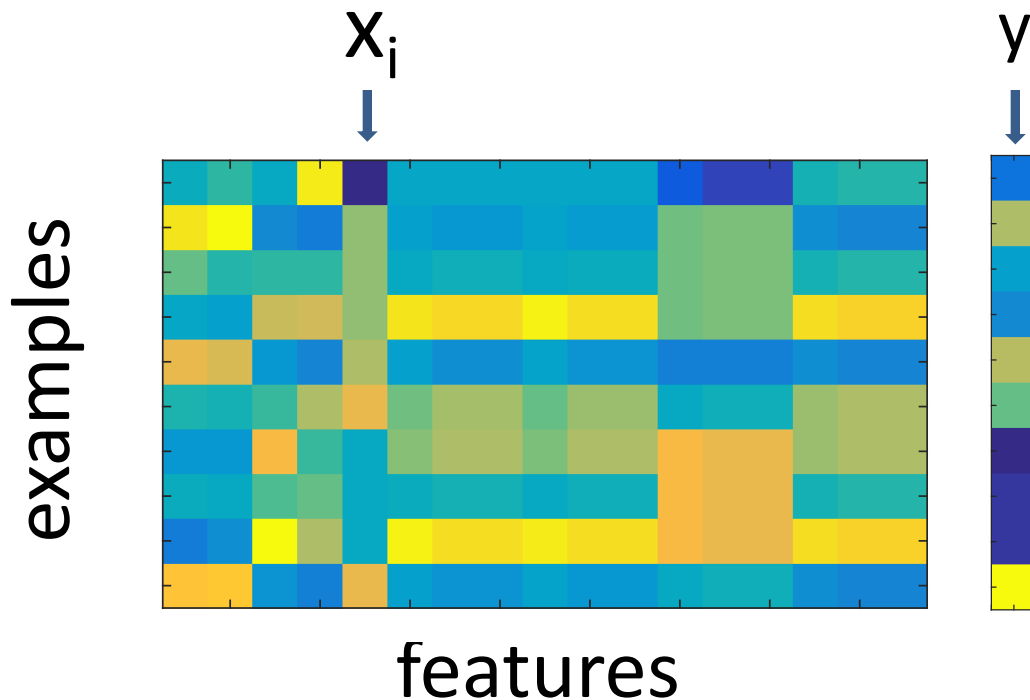
inference | *P* values | lasso

out many new treatments and reported to you only ones for which $|z| > 2$? Then a value of 2.5 is not nearly as surprising. Indeed, if the two treatments were equivalent, the conditional probability that $|z|$ exceeds 2.5, given that it is larger than 2, is about 27%. Armed with knowledge of the process that led to the value $z = 2.5$, the correct selective inference would assign a *P* value of 0.27 to the finding, rather than 0.01.

If not taken into account, the effects of selection can greatly exaggerate the apparent strengths of relationships. We feel that this is one of the causes of the current crisis in reproducibility in science (e.g., ref. 1). With increased competitiveness and pressure to publish, it is natural for researchers to exaggerate their claims, intentionally or otherwise. Journals are much more likely to publish studies with low *P* values, and we (the readers) never hear about the great

Finding significant features

- x_i : Vector of i -th feature
- y : Outcome vector subject to Gaussian
- Correlation: $s_i = \sum_{j=1}^d x_{ij} y_j$

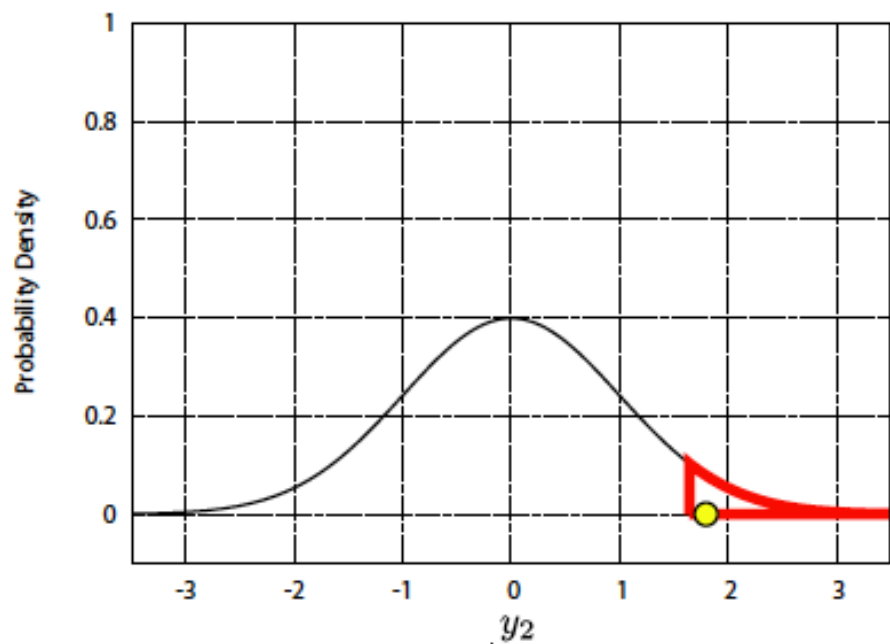


Selection event

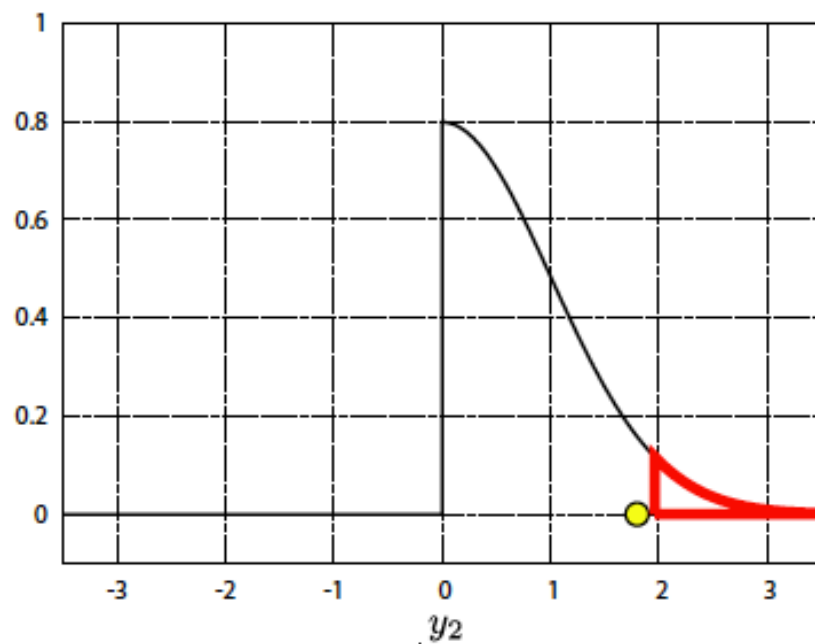
- Compute correlation s_i for all features, select top-k
- The set of all outcome vectors leading to the same selection is described as

$$Y = \{\mathbf{y} \mid A\mathbf{y} \leq \mathbf{b}\}$$

- Null distribution of s_i conditioned on the selection event is **truncated Gaussian** (Lee et al., 2013)



(a) naive inference



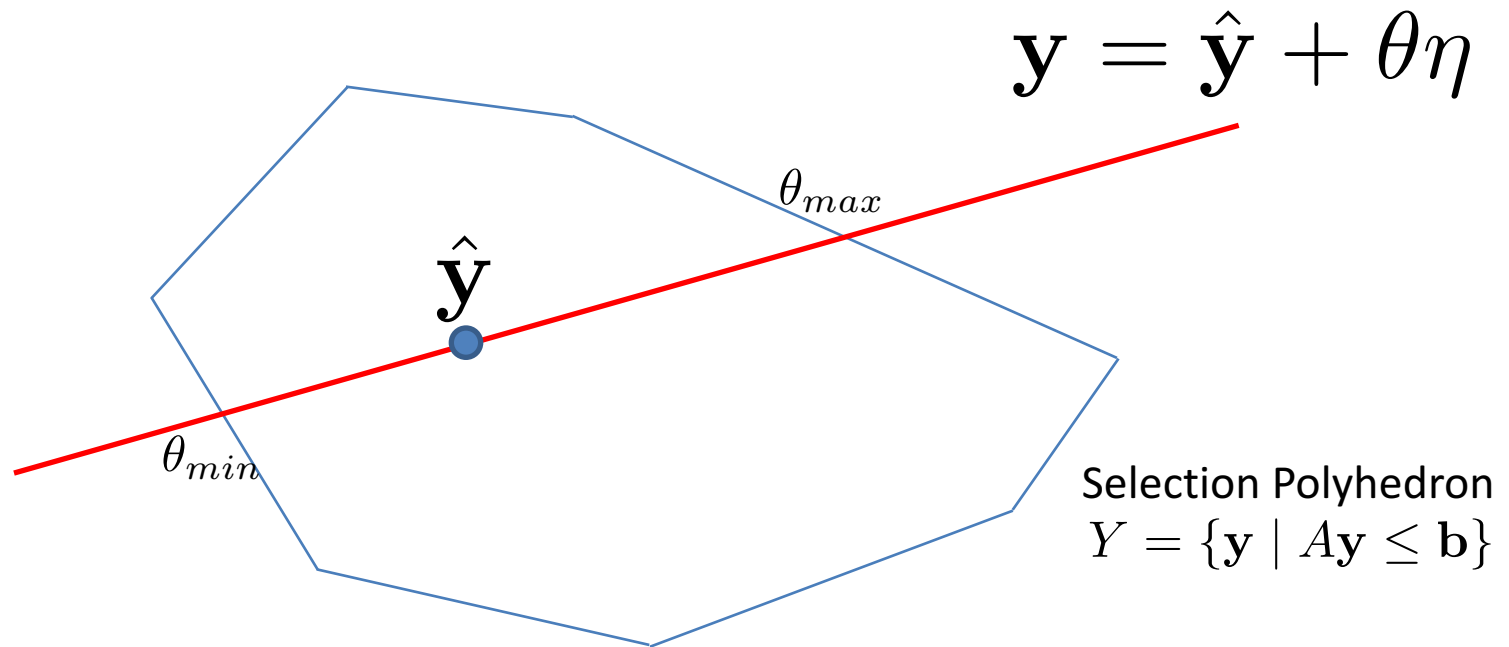
(b) selective inference

Our contribution: Applying selective inference to pattern mining

- Computation of selective null distribution needs reference to all features
- Impossible in combinatorial cases !
- New bound for disregarding large patterns
- Applicable to stepwise selection, LASSO selection
- Suzumura et al., Selective Inference Approach for Statistically Sound Discriminative Pattern Discovery, arXiv 2016.

Geometric View

- Pick a selected pattern. Occ. Vec.: η
- Need θ_{\min} and θ_{\max} to compute the null distribution



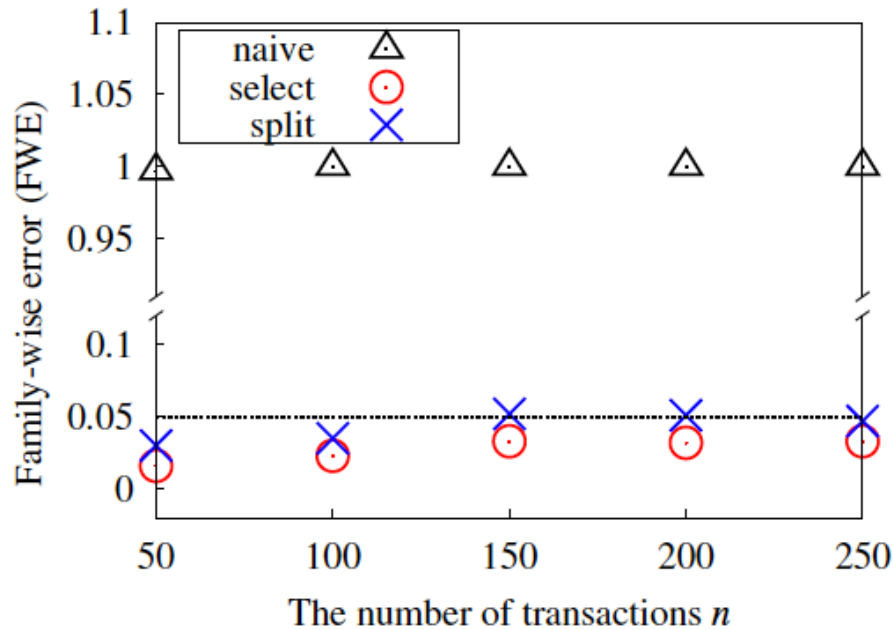
Identifying intersection points

- A pattern = A face of selection polyhedron
- Rerun pattern mining to identify the cross points θ_{\min} and θ_{\max}

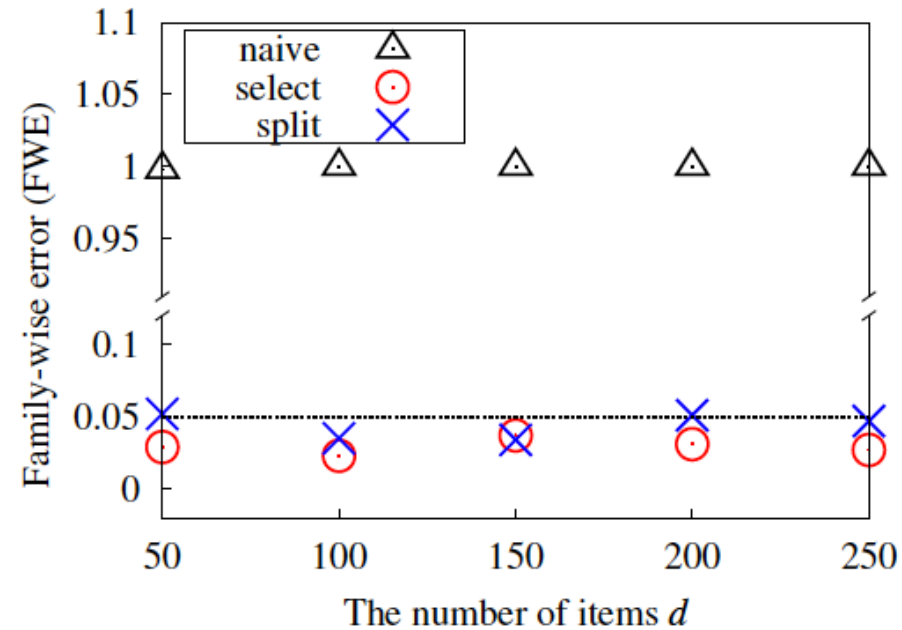
$$\theta_{\min} = \max_{\substack{(j,j') \in \mathcal{K} \times \{[J] \setminus \mathcal{K}\}, \\ (\tau_{j'} - \tau_j)^\top \eta < 0}} \frac{(\tau_j - \tau_{j'})^\top y}{(\tau_{j'} - \tau_j)^\top \eta},$$

$$\theta_{\max} = \min_{\substack{(j,j') \in \mathcal{K} \times \{[J] \setminus \mathcal{K}\}, \\ (\tau_{j'} - \tau_j)^\top \eta > 0}} \frac{(\tau_j - \tau_{j'})^\top y}{(\tau_{j'} - \tau_j)^\top \eta}.$$

Discriminative itemset mining (k=5): Family-wise error

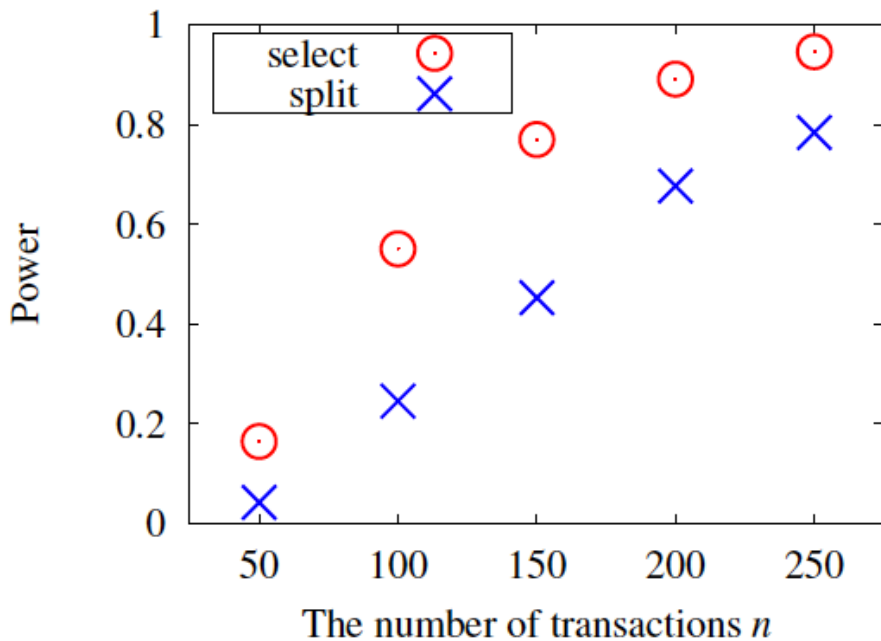


(a) $n \in \{50, \dots, 250\}$

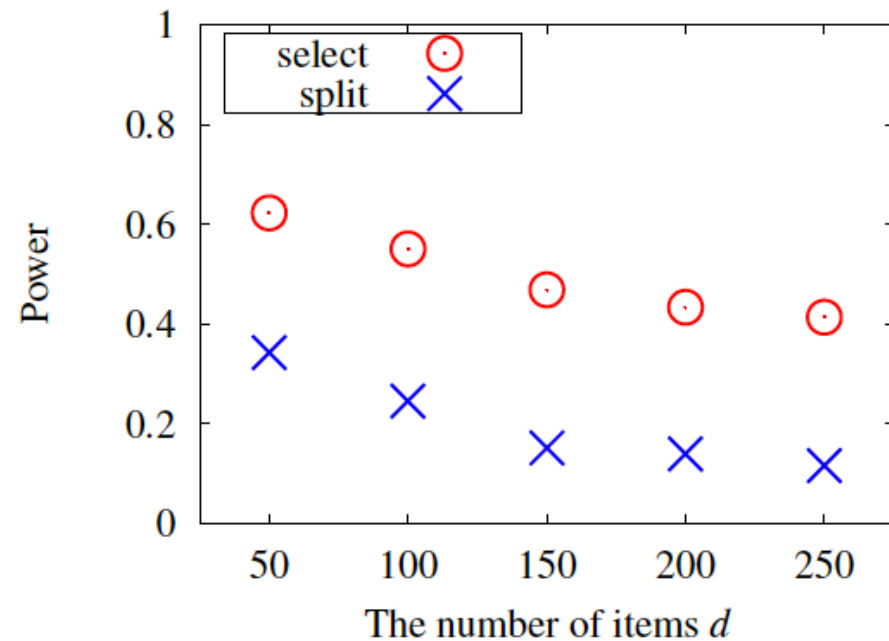


(b) $d \in \{50, \dots, 250\}$

Discriminative itemset mining (k=5): Power

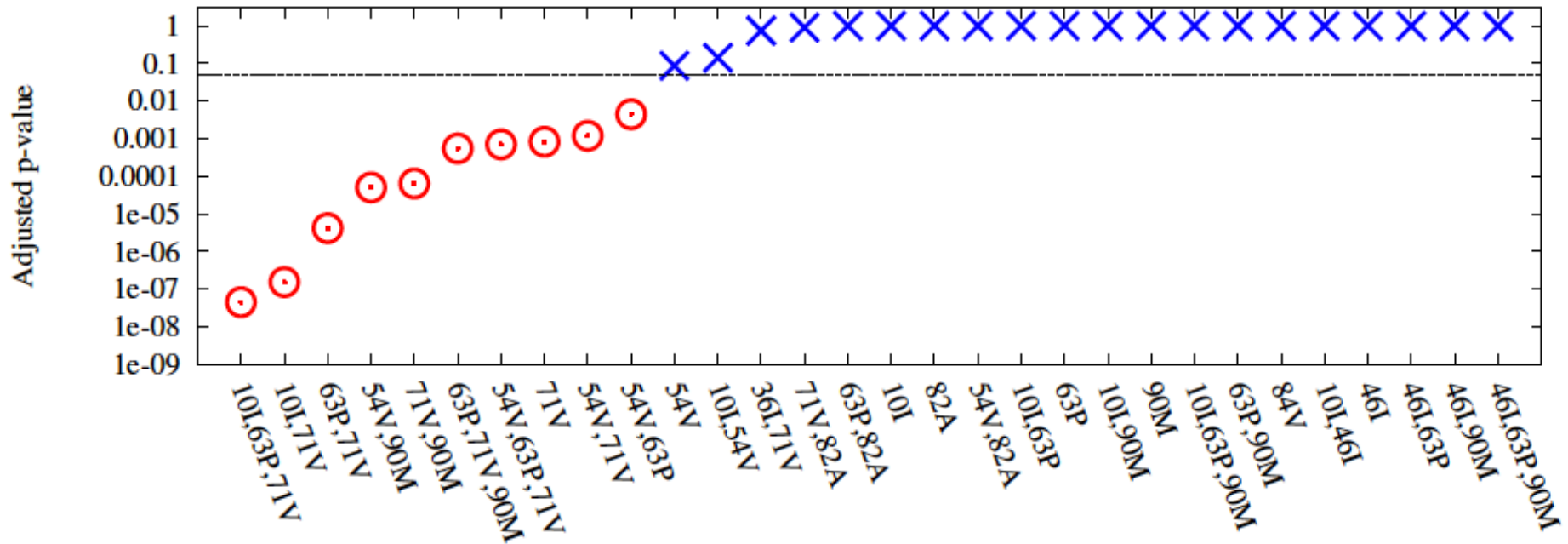


(a) $n \in \{50, \dots, 250\}$



(b) $d \in \{50, \dots, 250\}$

HIV drug resistance: significant combination of mutations



Conclusion

- False positive control is crucial in sciences
- Statistics need to **change** !
- Compromise without principles (p-hacking) can kill sciences
- LAMP and selective inference methods enhance chance of discovery

10th international conference on multiple comparison procedures (MCP 2017)

- Session: Data mining methods under multiplicity control
- June 20-23, 2017, UC Riverside
- Abstract deadline: January 31, 2017
- <http://www.mcp-conference.org/hp/2017/>