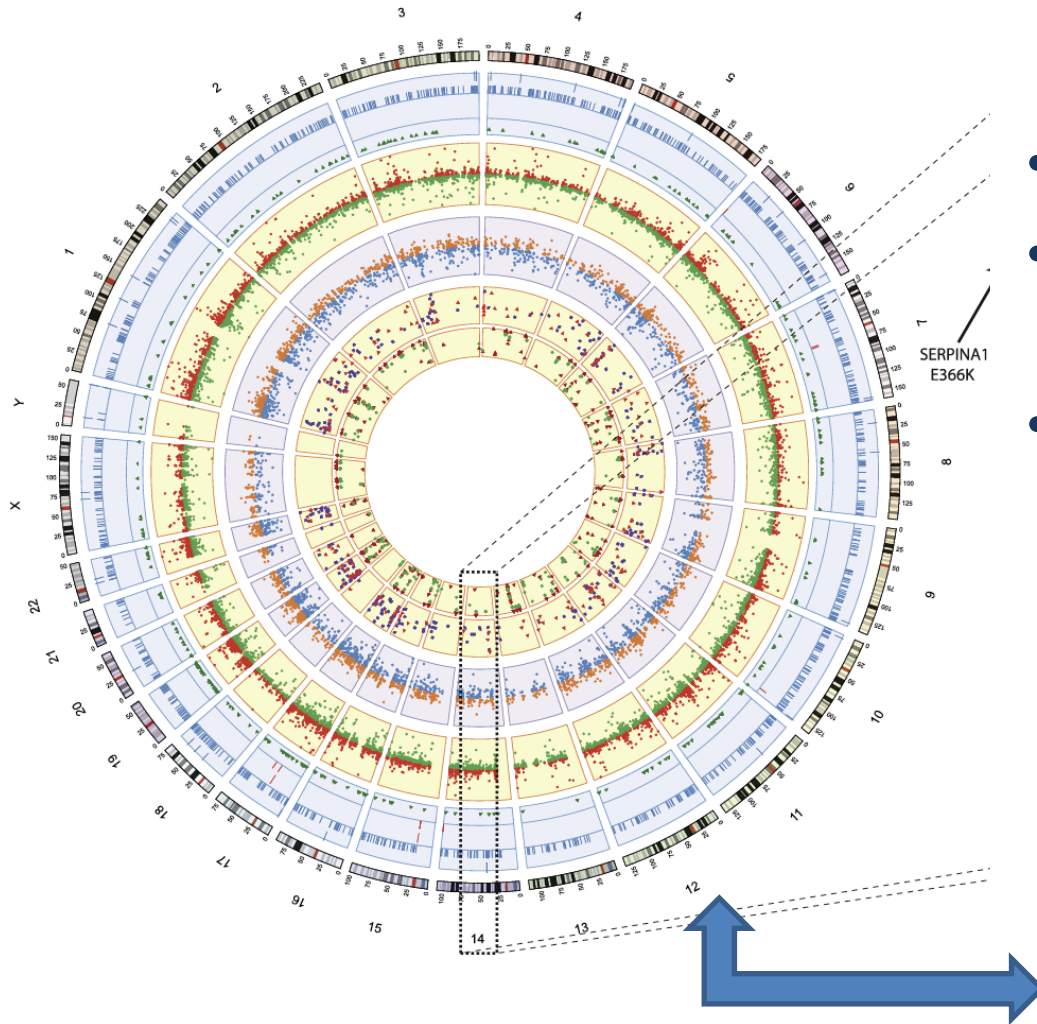


# Significant Pattern Mining for Biomedical Applications

Koji Tsuda

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

# Multi-omics Data



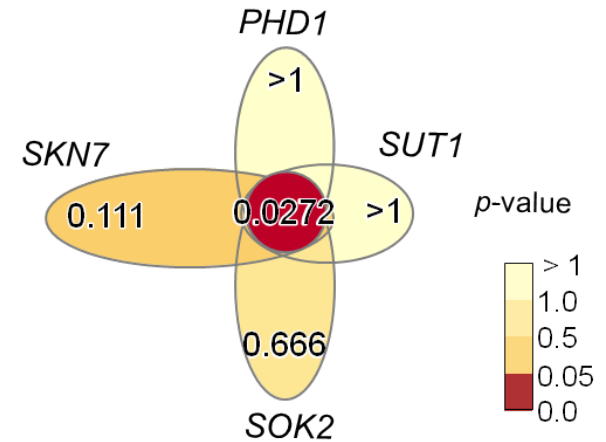
- **DNA** (mutation, insertion, deletion, CNV etc)
- **DNA methylation, Histon modification**
- **mRNA expression, ncRNA**
- **Protein expression, modification**
- **Metabolite** (Sugar, Amino acids, Nucleotides, lipids)

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

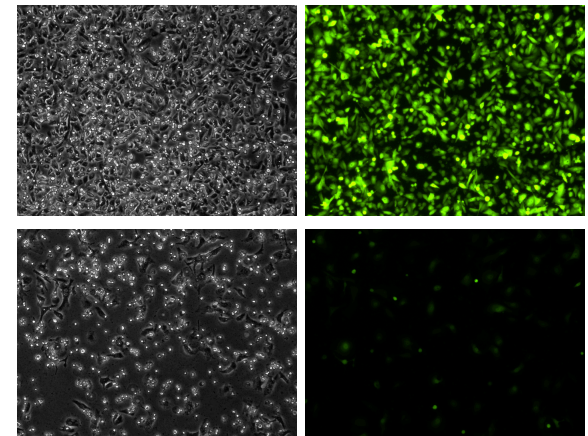
# Drawbacks of “Single Factor Screening”

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

Single factor screening misses combinatorial causes



Knock down Experiments



Trans-omics data  
Clinical data



Single Factor  
Screening  
(e.g., Chi2 test)



MycN

Single Gene

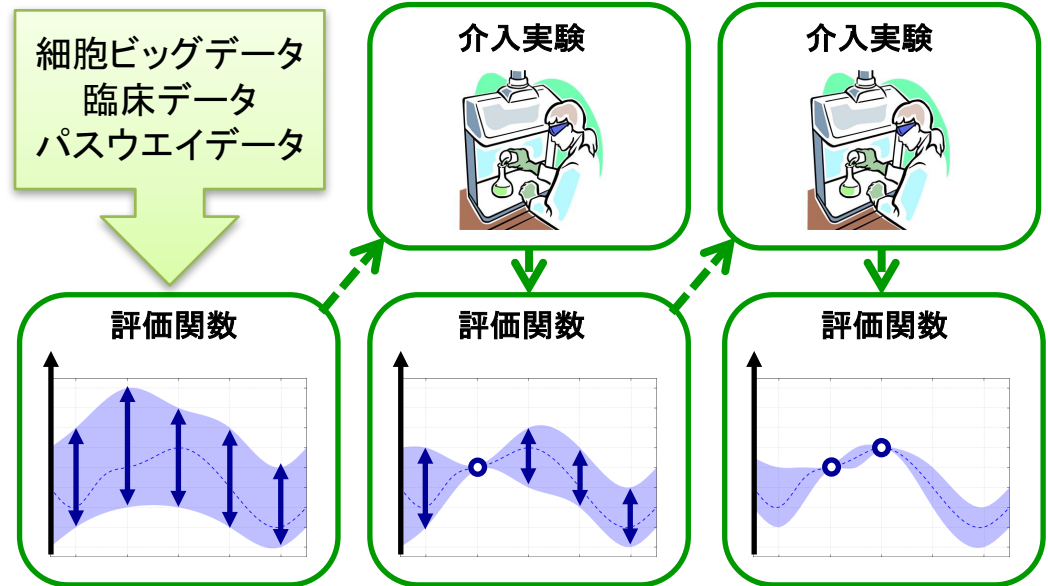
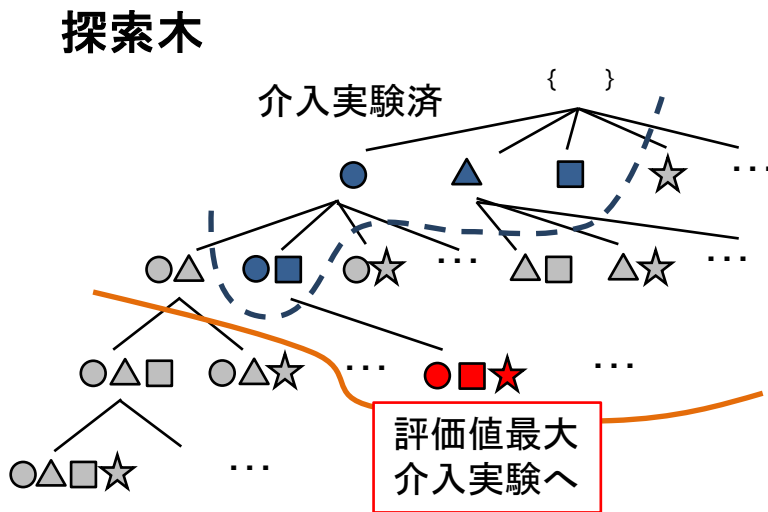
Knock-down  
Experiment



# Challenge:

## Discovering Combinatorial Factors Associated with Biological Phenomena

- Itemset mining, sequence mining, graph mining
- How to assess statistical significance?
- Multiple testing correction?



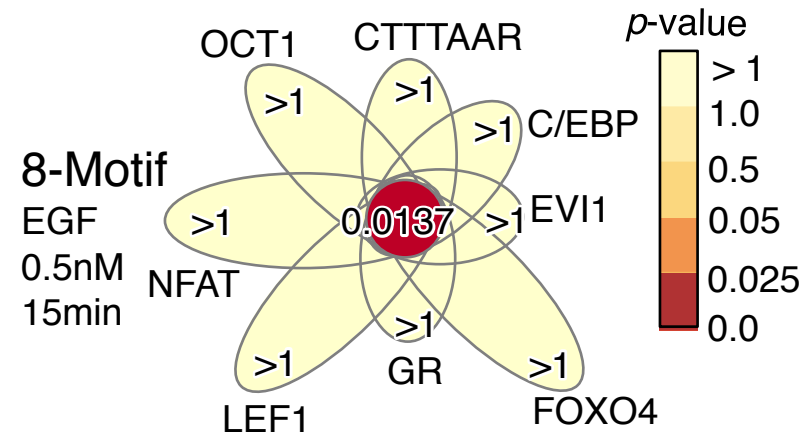
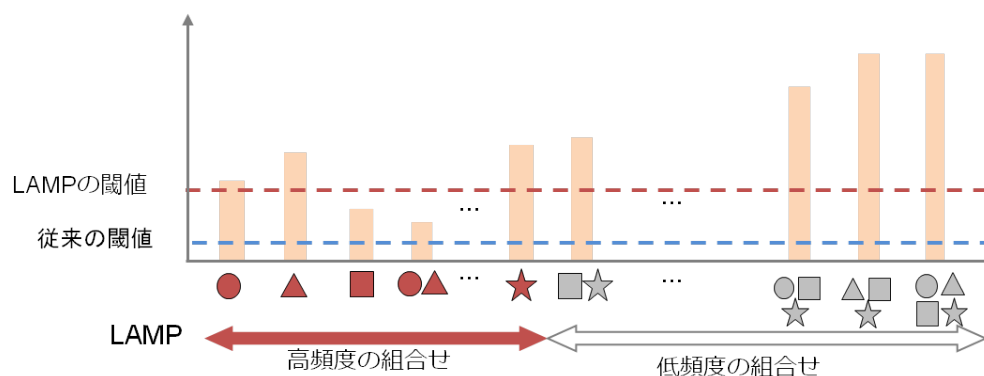
# Short History of Significant Pattern Mining

No multiple testing correction	Tarone-Bonferroni correction	Fast Westfall-Young methods
Webb, KDD 2007	Terada et al., PNAS 2013	Terada et al., BIBM 2013
Hamalainen, Know Inf Syst, 2012	Minato et al., ECMLPKDD 2014	Llinares-López et al., KDD 2015
	Sugiyama et al., SDM 2015	Llinares-López et al., Bioinformatics 2015
	Terada et al., PAKDD 2016	
	Terada et al., Bioinformatics, 2016	

# 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)
  - Applications to GWAS data

# Limitless Arity Multiple testing Procedure (LAMP)

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

## Statistical significance of combinatorial regulations

Aika Terada<sup>a,b,c</sup>, Mariko Okada-Hatakeyama<sup>d</sup>, Koji Tsuda<sup>c,e,1</sup>, and Jun Sese<sup>a,b,1</sup>

<sup>a</sup>Department of Computer Science and <sup>b</sup>Education Academy of Computational Life Sciences, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan; <sup>c</sup>Minato Discrete Structure Manipulation System Project, Exploratory Research for Advanced Technology, Japan Science and Technology Agency, Sapporo, Hokkaido 060-0814, Japan; <sup>d</sup>Laboratory for Integrated Cellular Systems, RIKEN Center for Integrated Medical Sciences (IMS-RCAI), Yokohama, Kanagawa 230-0045, Japan; and <sup>e</sup>Computational 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.

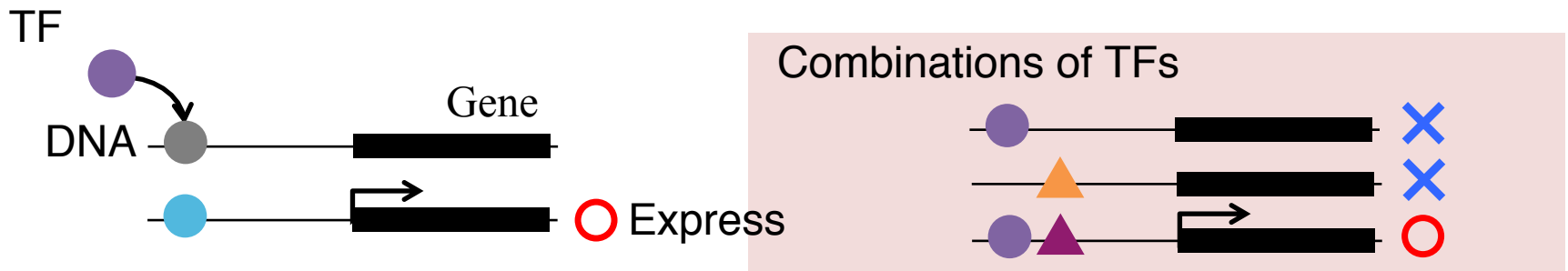
### Results

**Method Overview.** To present our strategy for combinatorial regu-

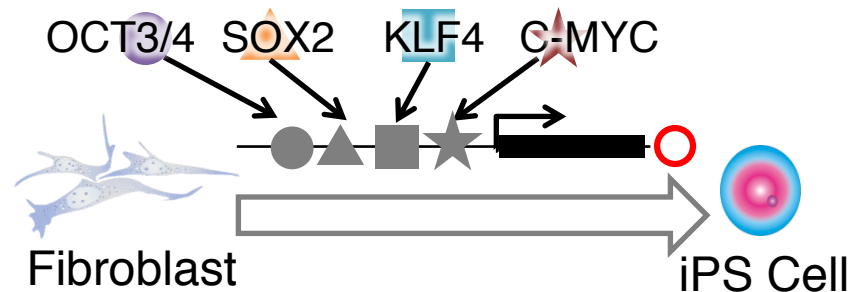


# Transcription factors (TFs) work in combination

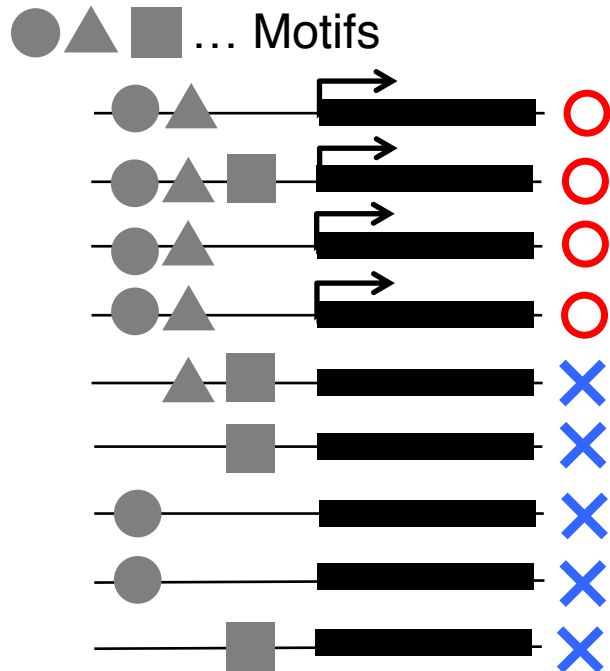
- Often several TFs are necessary to induce the expression of downstream genes



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  $\delta$  is determined such that FWER is below  $\alpha$
- 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  $\alpha$ , 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  $\alpha$

# Bonferroni Correction

- Given threshold  $\delta$ , FWER is bounded as

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

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

	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

$$P(V > 0 \mid \bigcap_{i=1}^m H_i) \leq \sum_{i=1}^m P(p_i \leq \delta \mid H_i) \quad \text{Union bound}$$

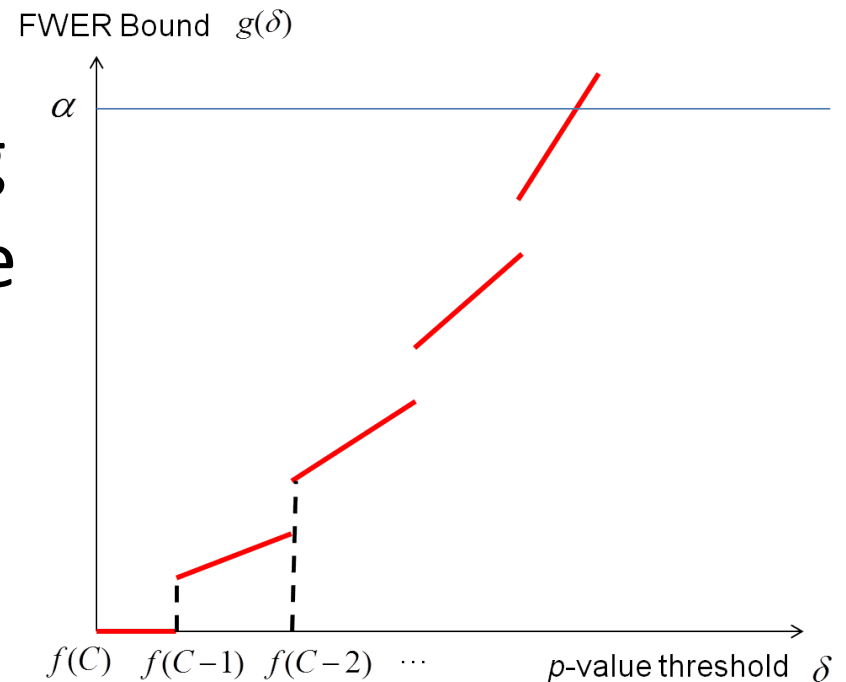
$$= \sum_{\{i \mid f(x_i) \geq \delta\}} P(p_i \leq \delta \mid H_i) \quad \text{Use minimum p-value to remove hypotheses}$$

$$\leq |\{i \mid f(x_i) \geq \delta\}| \delta \quad \text{Definition of p-value}$$

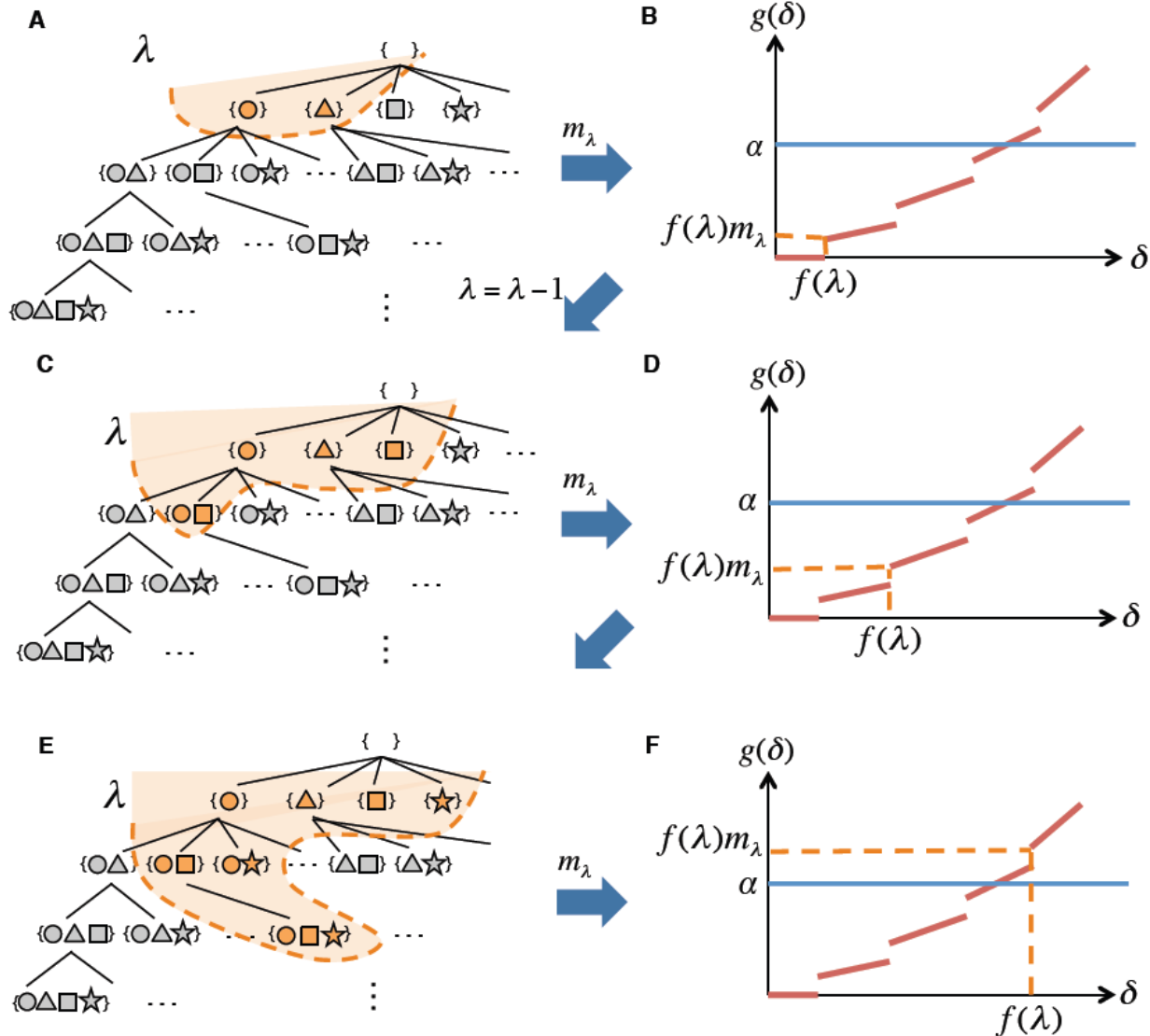
- Take maximum  $\delta$  that keeps FWER bound below  $\alpha$

# Finding optimal cut-off $\delta$ that calibrates FWER bound to $\alpha$

- 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  $\alpha$



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



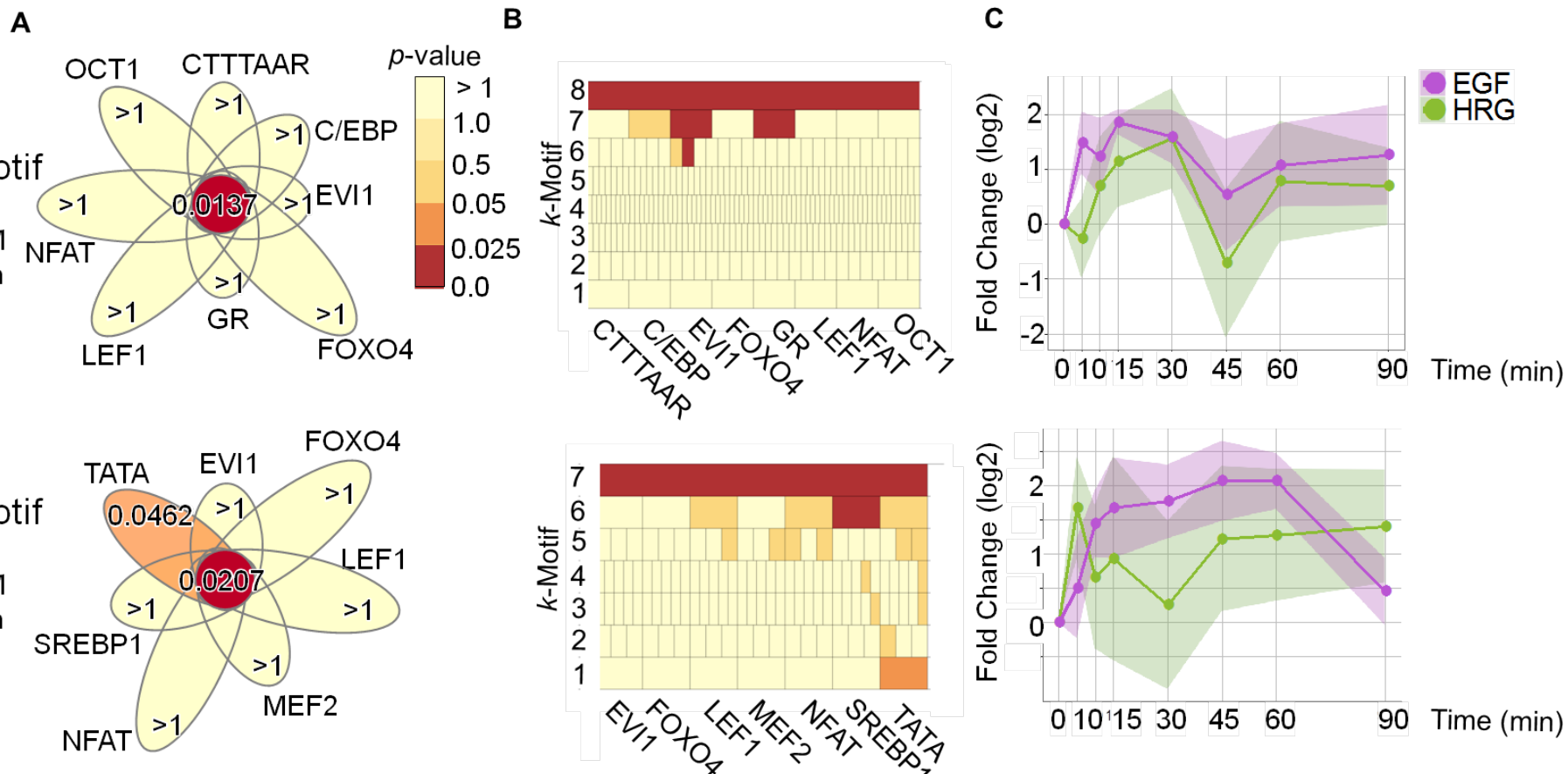
# Statistically significant TF combinations under a heat shock condition (Yeast)

Corrected p-value (p-value\*K)

Combination	LAMP ( $\leq 102$ )	Bonferroni ( $\leq 4$ )
	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

# Application to MCF7 human breast cancer cells (GSE6462)

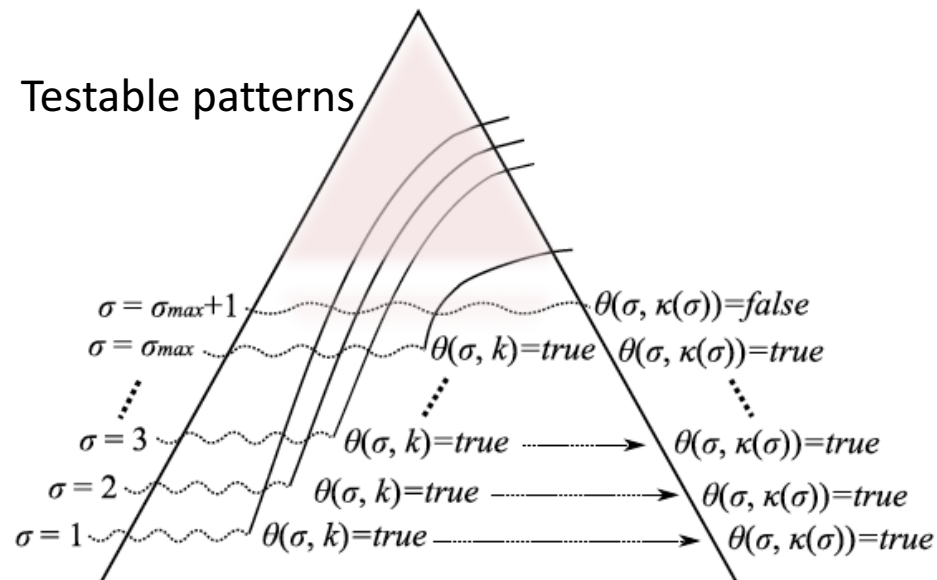


# Development of LAMP implementations

- 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

# 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<sup>1,2,3,\*</sup>, Ryo Yamada<sup>4</sup>, Koji Tsuda<sup>2,3,5</sup> and Jun Sese<sup>3,6,\*</sup>**

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

# MP-LAMP

<https://github.com/tsudalab/mp-lamp>

- Massively parallel implementation of LAMP based on MPI
- Runs on Amazon Web Service or on-premises computer cluster

The screenshot shows the GitHub repository page for `tsudalab/mp-lamp`. The repository has 4 watchers, 1 star, and 0 forks. It contains 5 commits, 1 branch, 1 release, and 1 contributor. The latest commit by `yoshizoe` is dated 17 days ago. The repository also has a `main` branch and an `aws` branch, both with their first commit 18 days ago.

Repository: `tsudalab / mp-lamp`

Unwatch 4 | Unstar 1 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Pulse | Graphs | Settings

No description or website provided. — Edit

5 commits | 1 branch | 1 release | 1 contributor

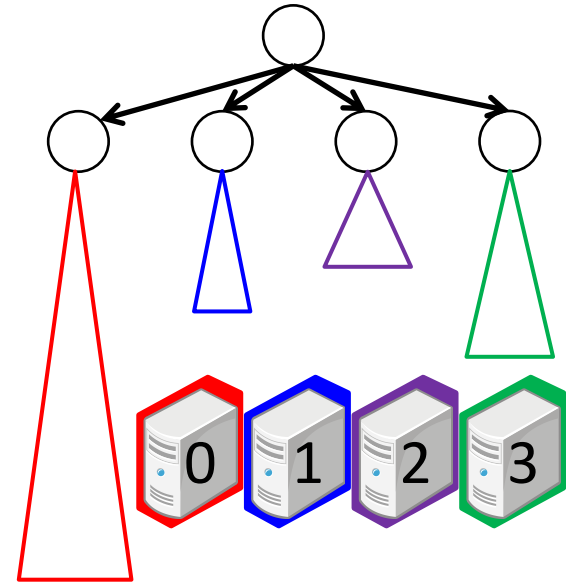
Branch: `master` | New pull request | Create new file | Upload files | Find file | Clone or download

`yoshizoe` fixed a type in readme. added a sample comment in local.sample.cfg | Latest commit `ac7770a` 17 days ago

<code>aws</code>	first commit	18 days ago
<code>main</code>	first commit	18 days ago

# Parallel Implementation of LAMP with Message Passing Interface

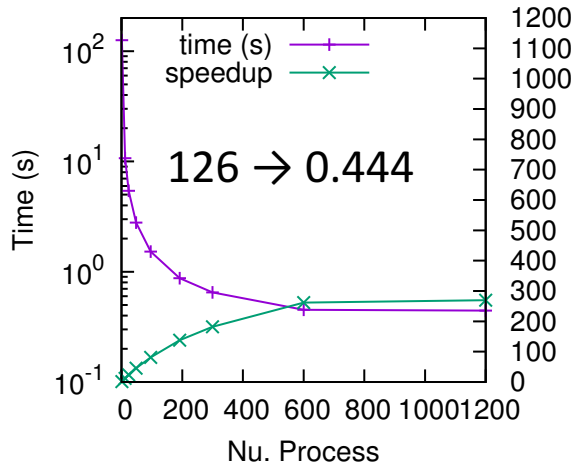
- Tree splitting is not good
- Dynamic load balancing
  - Passing tasks among computing nodes
- Count table shared by reduce-broadcast



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

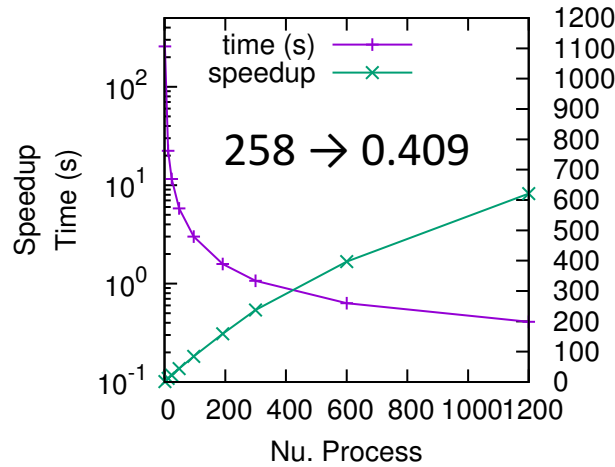
## HapMap dom. 10%

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



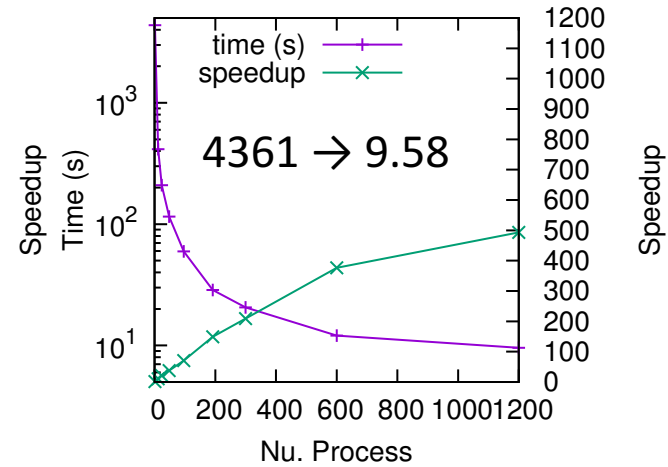
## Alz. dom. 5%

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



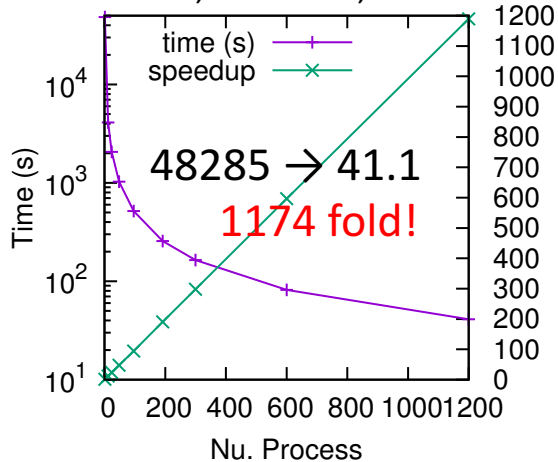
## Alz. rec. 30%

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



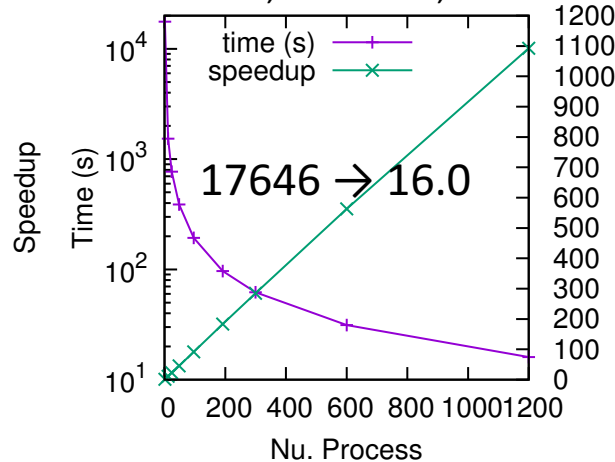
## HapMap dom. 20%

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



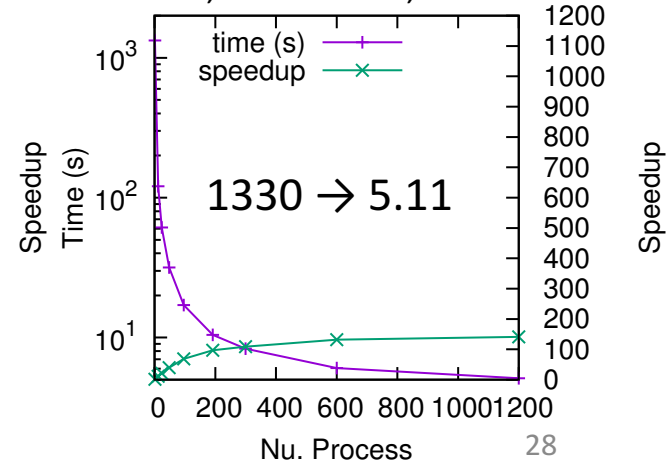
## Alz. dom. 10%

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



## MCF7

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



- 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)

# Conclusion

- False positive control is crucial in sciences
- Compromise without principles (p-hacking) can harm sciences
- Significant pattern mining is a promising way to discover combinatorial effects with statistical guarantees

# 10<sup>th</sup> international conference on multiple comparison procedures (MCP 2017)

- Statisticians' conference: **Tibshirani, Benjamini, Tarone etc..**
- 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/>

**MCP Conference 2017**  
*10th International Conference on Multiple Comparison Procedures in Riverside, California from June 20th - 23th, 2017*

UNIVERSITY OF CALIFORNIA  
RIVERSIDE

[Home](#) [Important Dates](#) [Speakers/Sessions](#) [Org. Committee](#) [Short Course](#)

**10th International Conference on Multiple Comparison Procedures**  
**Event Dates:** Tuesday June 20 - Friday June 23, 2017

**Newsletter**  
• [Subscribe](#)  
[Travel Information](#)