Statistically significant SNP combinations for GWAS data

Xavier Duran
GCAT Genomes for Life
Institut de Recerca Germans Trias i Pujol (IGTP)

Bioinfo Talks
February 15th 2017

Missing heritability problem on GWAS

GWAS studies have discovered many loci associated with various complex traits but they only explain a very small part of the human heritability

Screening of individual SNPs using statistical tests to assess the association of each SNP with a phenotype

Statistical epistatic interactions are hard to find

This unexplained variation because GWAS tries to solve common variant, common disease hypothesis

Most complex traits are due to the effect of interactions of different SNP

Combinatorial effects of multiple SNPs

Epistasis can be part of the explanation

Why?

Apply data mining and machine learning techniques to biological data analysis

Hard computational problem

GCAT

Limitless arity multi-testing procedure (LAMP)

Computational complexity

Statistical significance

Outline

The complexity of combinatorial variant discovery

How does LAMP approaches a solution

Results on a lung cancer dataset

Finding combinations of features

Computational problem

Exploring all combinations is computationally prohibitive

Exponential growth: for M binary variables, 2^M tests computed

Finding combinations of features

Statistical problem

Discovered combinations are statistically unlikely due to multiple testing correction

For M binary variables, Bonferroni correction sets significance below $\frac{\alpha}{2^M}$

Finding combinations of features

We cannot test all combinations

Most machine learning methods (RF) don't evaluate statistical significance of the reported results

Too much false positives, very costly to futher explore hypothesis

Not comprehensive

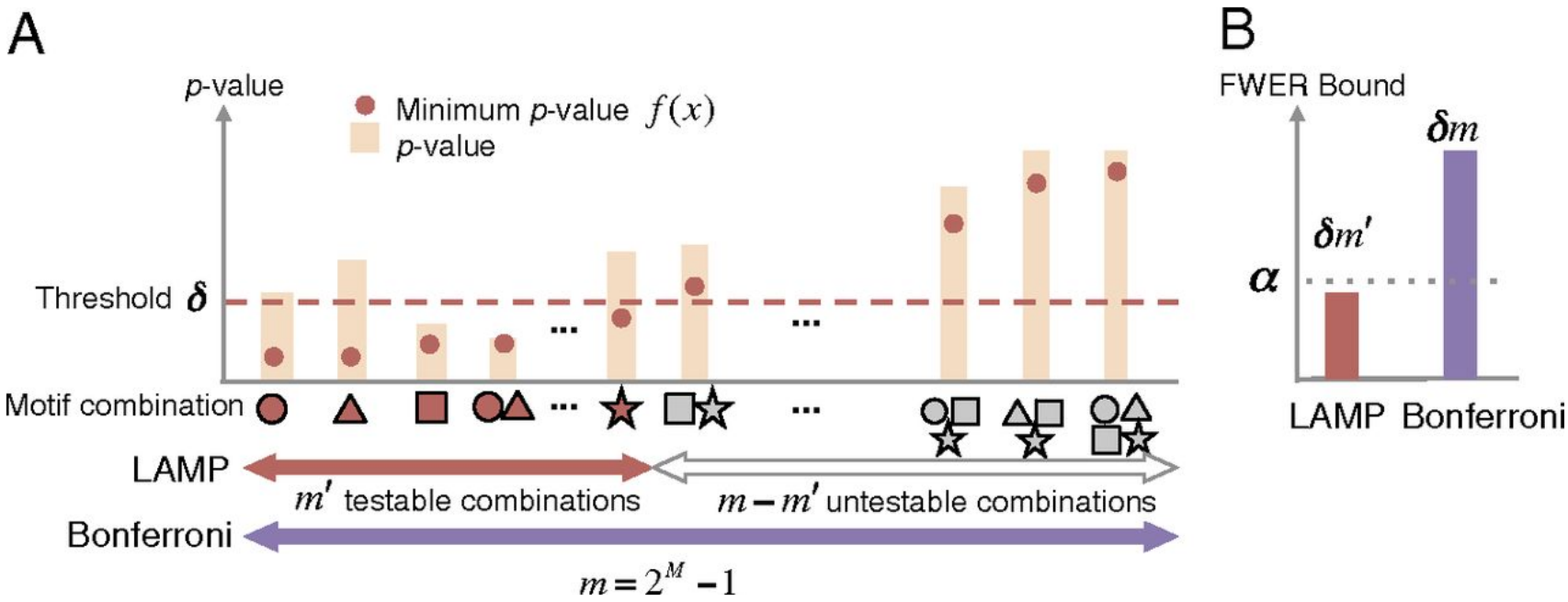
Limitless arity multi-testing procedure (LAMP)

Multiple testing procedure for listing ALL statistically significant high order interactions

Limitless arity multi-testing procedure (LAMP)

Testable combinations

Upper bound of Family Wise Error Ratio (FWER)



[Terada et al. 2013]

Limitless arity multi-testing procedure (LAMP)

Fisher's exact test

Not all combinations are frequent enough to become frequent in any case/control setting

Each combination has a maximum p-value, independent of its distribution on the two classes

Test only what is relevant to test

	Case	Control	Total
Has S_i	171	206	377
Hasn't S_i	29	94	123
total	200	300	500

Limitless arity multi-testing procedure (LAMP)

Fisher's exact test

Not all combinations are frequent enough to become frequent in any case/control setting

Each combination has a maximum p-value, independent of its distribution on the two classes

Test only what is relevant to test

	Case	Control	Total
Has S_i	k	K-k	K
Hasn't S_i	n-k	N-K-n+k	N-K
total	n	N-n	N