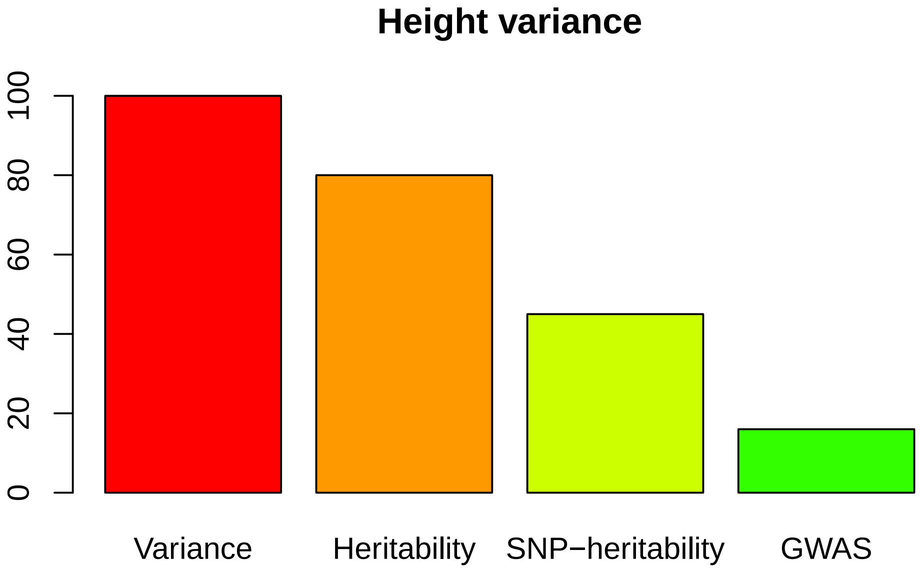
Significant pattern mining for GWAS data

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

Bioinfo Talks
February 15th 2017

Missing heritability problem on GWAS



Limitless arity multi-testing procedure (LAMP)

Significant pattern mining techniques can help to find high-order interactions on GWAS data (and other biological data)

Limitless arity multi-testing procedure (LAMP)

Significant pattern mining techniques can help to find high-order interactions on GWAS data (and other biological data)

Outline

The complexity of combinatorial variant discovery

Limitless arity multi-testing procedure (LAMP)

Significant pattern mining techniques can help to find high-order interactions on GWAS data (and other biological data)

Outline

The complexity of combinatorial variant discovery

How does LAMP approaches a solution

Limitless arity multi-testing procedure (LAMP)

Significant pattern mining techniques can help to find high-order interactions on GWAS data (and other biological data)

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

Finding combinations of features

Computational problem

Exploring all combinations is computationally prohibitive

M^2 second order possible interactions

Finding combinations of features

Computational problem

Exploring all combinations is computationally prohibitive

M^2 second order possible interactions

2^M limitless order interactions

Finding combinations of features

Computational problem

Exploring all combinations is computationally prohibitive

M^2 second order possible interactions

2^M limitless order interactions

Statistical problem

Discovered combinations are statistically unlikely due to multiple testing correction

Finding combinations of features

Computational problem

Exploring all combinations is computationally prohibitive

M^2 second order possible interactions

2^M limitless order interactions

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

Machine learning approaches

Random Forests, Suport Vector Machines, Multifactor Dimensionality Reduction

Finding combinations of features

Machine learning approaches

Random Forests, Suport Vector Machines, Multifactor Dimensionality Reduction

Variable rankings

Finding combinations of features

Machine learning approaches

Random Forests, Suport Vector Machines, Multifactor Dimensionality Reduction

Variable rankings

Too much false positives

Finding combinations of features

Machine learning approaches

Random Forests, Suport Vector Machines, Multifactor Dimensionality Reduction

Variable rankings

Too much false positives

Very costly to futher explore hypothesis

Limitless arity multi-testing procedure (LAMP)

Fisher's exact test

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

Limitless arity multi-testing procedure (LAMP)

Fisher's exact test

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

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

Limitless arity multi-testing procedure (LAMP)

Fisher's exact test

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

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

	Case	Control	Total
Has S_i			13
Hasn't S_i			357
total	184	186	370

Limitless arity multi-testing procedure (LAMP)

Fisher's exact test

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

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

	Case	Control	Total
Has S_i	13	0	13
Hasn't S_i	171	186	357
total	184	186	370

Limitless arity multi-testing procedure (LAMP)

Fisher's exact test

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

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

	Case	Control	Total
Has S_i	13	0	13
Hasn't S_i	171	186	357
total	184	186	370

raw p-value $= 9.1 * 10^{-5}$

Limitless arity multi-testing procedure (LAMP)

Fisher's exact test

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

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

	Case	Control	Total
Has S_i	13	0	13
Hasn't S_i	171	186	357
total	184	186	370

raw p-value $= 9.1 * 10^{-5}$

FWER threshold $\delta = \alpha/1000 = 0.05/1000 = 5 * 10^{-5}$

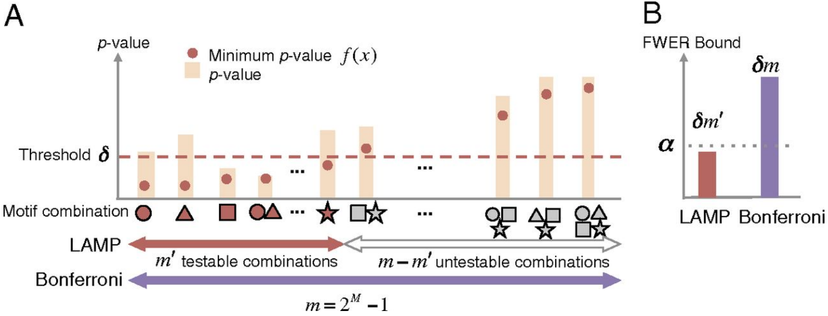
Limitless arity multi-testing procedure (LAMP)

Multiple testing procedure for listing ALL statistically significant high order interactions

Limitless arity multi-testing procedure (LAMP)

Multiple testing procedure for listing ALL statistically significant high order interactions

Upper bound of Family Wise Error Ratio (FWER)



[Terada et al. 2013]

LAMPLINK

LAMPLINK is implemented as additional features to PLINK

LAMPLINK

LAMPLINK is implemented as additional features to PLINK

Model dominant/recessive for the risk class for the minor allele

LAMPLINK

LAMPLINK is implemented as additional features to PLINK

Model dominant/recessive for the risk class for the minor allele

- Find all significant combinations
- Remove combinations with SNPs in linkage desequilibrium

LAMPLINK

LAMP in a lung cancer dataset

GWAS data of lung cancer progression

GWAS threshold	p-value $< 10^{-4}$
SNPs	695
Individuals	178
Statistical test	Fisher's exact test
Adjusted significance level	$5.8 * 10^{-9}$
Correction factor	8619336
Significant combinations	5019
r^2 for LD	0.2
Significant combinations after LD pruning	145
Significant SNPs	25
Maximum arity	7

LAMPLINK

LAMP in a lung cancer dataset

COMBID	Raw_P	Adjusted_P	COMB	arity
COMB7	0.00000000	0.00001538	rs438228:161484124:A:C,rs35684:10326686:A:G,rs1565656:188922545:A:G,rs4545589,rs139996291:17192744:G:A	5
COMB10	0.00000000	0.00002144	rs2271545:16095316:C:T,rs438228:161484124:A:C,rs35684:10326686:A:G,rs1565656:188922545:A:G	4
COMB39	0.00000000	0.00004028	rs438228:161484124:A:C,rs35684:10326686:A:G,rs1565656:188922545:A:G,rs4545589,rs9788969,rs139996291:17192744:G:A	6
COMB42	0.00000000	0.00008586	rs2271545:16095316:C:T,rs35684:10326686:A:G,rs1565656:188922545:A:G,rs139996291:17192744:G:A	4
COMB47	0.00000000	0.00009664	rs35684:10326686:A:G,rs1565656:188922545:A:G,rs4545589,rs9788969,rs139996291:17192744:G:A	5
COMB62	0.00000000	0.00011584	rs35684:10326686:A:G,rs1565656:188922545:A:G,rs4545589,rs139996291:17192744:G:A	4
COMB85	0.00000000	0.00013264	rs2271545:16095316:C:T,rs438228:161484124:A:C,rs35684:10326686:A:G,rs1565656:188922545:A:G,rs139996291:17192744:G:A	5
COMB159	0.00000000	0.00025099	rs2937667:117246037:C:A,rs10985542:124887090:G:A,12:48798429:T:C,rs139996291:17192744:G:A	4
COMB192	0.00000000	0.00050371	rs35684:10326686:A:G,rs2937667:117246037:C:A,rs1565656:188922545:A:G,rs139996291:17192744:G:A	4
COMB274	0.00000000	0.00058472	rs438228:161484124:A:C,rs35684:10326686:A:G,rs1565656:188922545:A:G,rs4545589,rs9788969	5
COMB278	0.00000000	0.00058472	rs438228:161484124:A:C,rs35684:10326686:A:G,rs6822954:35695840:A:G,rs1565656:188922545:A:G,rs4545589	5
COMB287	0.00000000	0.00067780	rs1565656:188922545:A:G,rs7111257:9930813:A:G,rs4545589,rs139996291:17192744:G:A	4
COMB328	0.00000000	0.00078732	rs2271545:16095316:C:T,rs438228:161484124:A:C,rs35684:10326686:A:G,rs1565656:188922545:A:G,rs9788969	5
COMB368	0.00000000	0.00078732	rs35684:10326686:A:G,rs2937667:117246037:C:A,rs1565656:188922545:A:G,rs4545589,rs139996291:17192744:G:A	5
COMB374	0.00000000	0.00078732	rs35684:10326686:A:G,rs2937667:117246037:C:A,rs1565656:188922545:A:G,rs71317450:27405120:A:T,rs139996291:17192744:G:A	5
COMB376	0.00000000	0.00078732	rs35684:10326686:A:G,rs6822954:35695840:A:G,rs1565656:188922545:A:G,rs4545589	4
COMB423	0.00000000	0.00079983	rs2271545:16095316:C:T,rs35684:10326686:A:G,rs1565656:188922545:A:G,rs9788969,rs139996291:17192744:G:A	5
COMB425	0.00000000	0.00079983	rs35684:10326686:A:G,rs6822954:35695840:A:G,rs11740157:10041128:A:G,12:51088287:AATACATAC:A	4
COMB447	0.00000000	0.00117950	rs438228:161484124:A:C,rs1565656:188922545:A:G,rs4545589,rs139996291:17192744:G:A	4
COMB610	0.00000000	0.00151520	rs2937667:117246037:C:A,rs10985542:124887090:G:A,12:48798429:T:C,rs9788969,rs139996291:17192744:G:A	5

Table 4: Statistically significant variant combinations

LAMPLINK

LAMP in a lung cancer dataset

CHR	SNP	A1	A2	TEST	AFF	UNAFF	P	OR	COMB
22	rs139996291:17192744:G:A	A	G	DOM	34/7	74/62	0.00094253	4.06950	106
4	rs1565656:188922545:A:G	G	A	DOM	33/8	74/62	0.00327766	3.45608	92
3	rs35684:10326686:A:G	G	A	DOM	30/11	56/80	0.00035202	3.89610	88
16	rs9788969	C	T	DOM	34/7	72/64	0.00051405	4.31746	56
1	rs438228:161484124:A:C	C	A	DOM	32/9	77/59	0.01679720	2.72439	49
1	rs2271545:16095316:C:T	C	T	DOM	32/9	64/72	0.00058287	4.00000	41
11	rs4545589	G	A	DOM	28/13	57/79	0.00409010	2.98516	41
12	12:51088287:AATACATAC:A	AATACATAC	A	DOM	33/8	79/57	0.00967982	2.97627	36
3	rs2937667:117246037:C:A	C	A	DOM	32/9	77/59	0.01679720	2.72439	32
5	rs11740157:10041128:A:G	G	A	DOM	27/14	42/94	0.00009612	4.31633	31
4	rs6822954:35695840:A:G	G	A	DOM	33/8	68/68	0.00055543	4.12500	15
9	rs10985542:124887090:G:A	G	A	DOM	26/15	49/87	0.00224055	3.07755	13
12	12:48798429:T:C	T	C	DOM	21/20	33/103	0.00174931	3.27727	12
21	rs71317450:27405120:A:T	T	A	DOM	30/11	82/54	0.14438900	1.79601	9
12	12:48792747:A:G	A	G	DOM	21/20	33/103	0.00174931	3.27727	7
5	rs11744968:10054699:T:C	C	T	DOM	23/18	36/100	0.00064061	3.54938	5
11	rs7111257:9930813:A:G	A	G	DOM	29/12	56/80	0.00120037	3.45238	5
16	rs59689196:78692994:A:C	C	A	DOM	18/23	27/109	0.00366244	3.15942	5
4	rs28657552:161256788:G:A	A	G	DOM	31/10	69/67	0.00661204	3.01014	3
13	rs41286971:41026812:G:A	A	G	DOM	30/11	74/62	0.04572730	2.28501	3
17	rs8065393:12974799:T:C	C	T	DOM	32/9	65/71	0.00063784	3.88376	3
11	rs61400460:8176765:TA:T	T	TA	DOM	18/23	11/125	0.00000071	8.89328	2
21	rs2242720	G	A	DOM	30/11	59/77	0.00118010	3.55932	2
11	rs9943610:86367530:C:T	C	T	DOM	24/17	47/89	0.01033820	2.67334	1
13	rs1464811:108513537:A:G	G	A	DOM	25/16	50/86	0.00706479	2.68750	1

Table 5: Variants statistically significant in any combination

Summary

SNP interactions may explain a part of the missing heritability but is a computationally and statistically challenging problem

Significant pattern mining can help finding statistically significative combinations of SNPs

Teh methodology is valid for other types of biomedical data