The 15th
Korea-Japan-China Bioinformatics Symposium 2017

A Few Examples of Association Rule Mining in Bioinformatics

Sangsoo Kim

based on

Dr. Sung Hee Park's works

Dept. of Bioinformatics & Life Sciences
Soongsil University

- Introduction to Association Rule Mining
- Classification of PPI Types
- Ascertainment of a Multivariate Phenotype
- Combinatorial Chromatin Modification Patterns

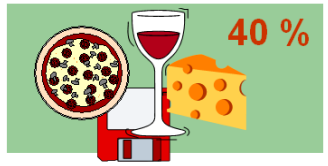
ARM: Association Rule Mining

Which product are bought together?

- If- then rules show relationships

TID	Items
1	Bread, Wine
2	Bread, Diaper, Beer, Eggs
3	Wine, Pizza, Floppy, Cheese
4	Wine, Pizza, Floppy, Cheese, Coke
5	Bread, Wine, Pizza, Coke

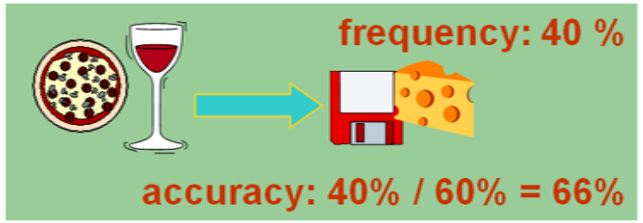
item sets



Wine and Pizza

Wine, Pizza , Floppy and Cheese

association-rule



IF Wine and Pizza THEN Floppy and Cheese

Head Body

Wine and Pizza → Floppy and Cheese(0.4, 0.66, 1.66)

Support

$$s = \frac{\sigma(Wine, Pizza, Floppy, Chesse)}{|T|} = \frac{2}{5} = 0.4$$

Confidence

$$c = \frac{\sigma(Wine, Pizza, Floppy, Cheese)}{\sigma(Wine, Pizza)} = \frac{2}{3} = 0.66$$

Lift

$$l = \frac{p(Wine, Pizza, Floppy, Cheese)}{p(Wine, Pizza) \times p(Floopy, Cheese)} = \frac{0.4}{0.6 \times 0.4} = 1.66$$

BMC Bioinformatics

BioMed Central

Research article

Open Access

Prediction of protein-protein interaction types using association rule based classification

Sung Hee Park[1], José A Reyes[2,3], David R Gilbert[2], Ji Woong Kim[1,4] and Sangsoo Kim*[1]

Address: [1]Department of Bioinformatics & Life Science, Soongsil University, Seoul, 156-743, Korea, [2]School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, UB8 3PH, UK, [3]Facultad de Ingeniería, Universidad de Talca, Talca, Chile and [4]Equispharm Co., Ltd, Seoul, 443-766, Korea

Email: Sung Hee Park - shpark@ssu.ac.kr; José A Reyes - jareyes@dcs.gla.ac.uk; David R Gilbert - david.gilbert@brunel.ac.uk; Ji Woong Kim - phosphoros@ssu.ac.kr; Sangsoo Kim* - sskimb@ssu.ac.kr

* Corresponding author

Published: 28 January 2009

BMC Bioinformatics 2009, **10**:36 doi:10.1186/1471-2105-10-36

Received: 19 May 2008
Accepted: 28 January 2009

Park *et al. BMC Systems Biology* 2011, **5**(Suppl 2):S13
http://www.biomedcentral.com/1752-0509/5/S2/S13

BMC Systems Biology

PROCEEDINGS Open Access

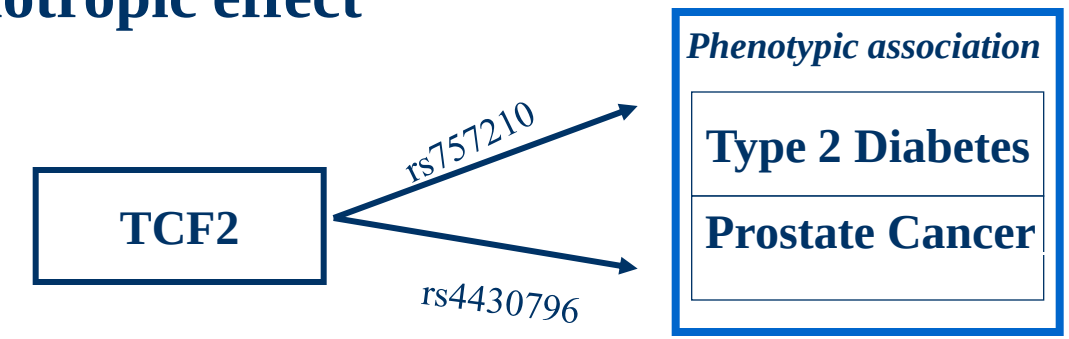
A methodology for multivariate phenotype-based genome-wide association studies to mine pleiotropic genes

Sung Hee Park, Ji Young Lee, Sangsoo Kim*

From 22nd International Conference on Genome Informatics
Busan, Korea. 5-7 December 2011

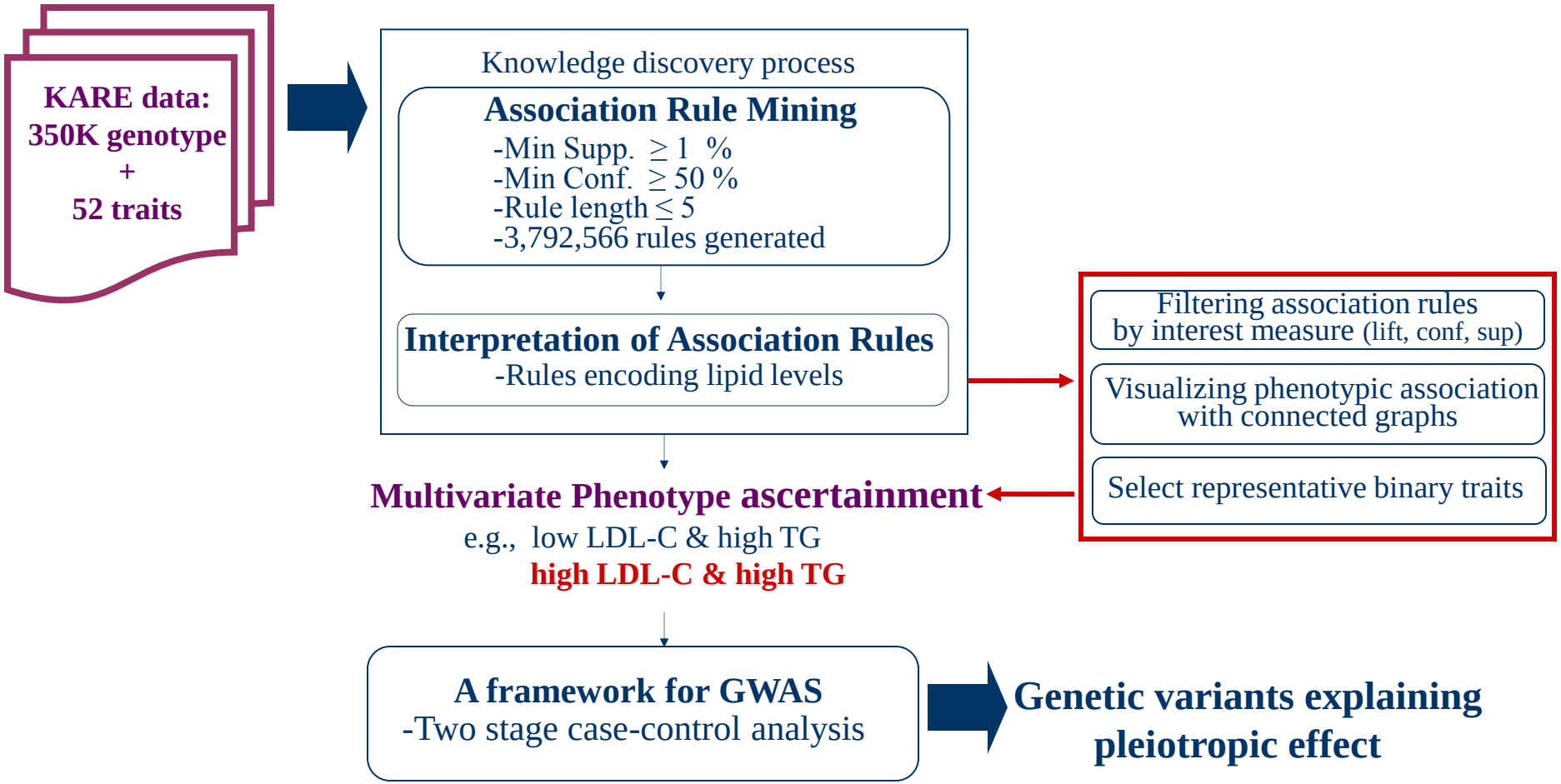
Heterogeneity Issues & Phenotype Definition

Pleiotropic effect



- **Examples**
  - some of the genes that affect eye color can also influence the overall height.
- **GWAS observe pleiotropic effects linking to diseases [Frayling08, Weedon08, Elliott10]**
  - variants of TCF2 increases risk of prostate cancer but decreases risk of type 2 diabetes [Gudmundsson07, Elliott10, Stevens10]
  - the same allele of GDF5 that associates with greater height also associates with reduced risk of osteoarthritis [Southam07]
  - Alleles at ApoE associated with total serum cholesterol and also influence other lipid related phenotypes (triglyceride and lipoprotein) [Templeton06]

Our Approach for GWAS of Multivariate Phenotypes



Park et.al, 2009, BIBM workshop
Park et al, 2012, IJDBMB

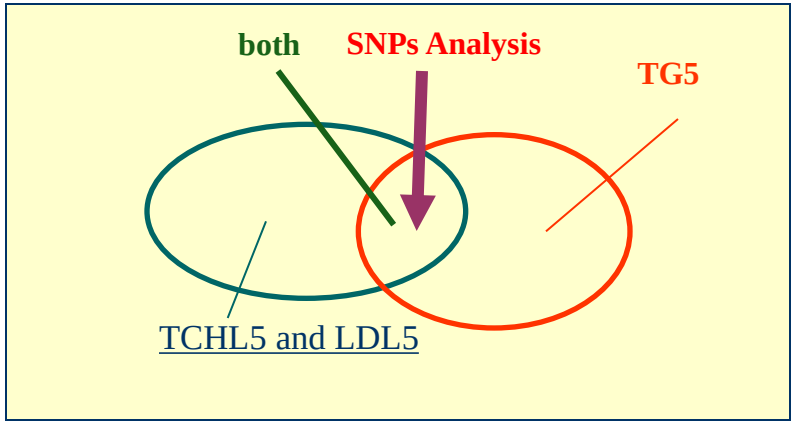
Application of ARM to Discovery of Phenotypes

- Items (floppy, chesses,…., etc.) ➔ 52 biochemical traits (age, gender, waist, …., etc.) ➔ equal binning (228 items)
- Transactions () ➔ individuals (6989 transactions)

TG1 < 91.5
91.5 ≤ TG2 < 116.5
116.5 ≤ TG3 < 146.5
146.5 ≤ TG4 < 196.5
TG5 ≥ 195.6

Examples of Association Rules

- **Rule form: IF *P1* and *P2* and *P3* … and *Pn* ➔ *Pk* (*con*, *sup*)**
- TCHL5 ➔ TG5 (0.75, 0.5)
- **TCHL5 and LDL5 ➔ TG5 (0.67, 0.4)**
- 65 < age < 70 and 80 < pulse < 90 ➔ high=Tchl (0.90, 0.03)



TID	Items
1	Female, Waist3, TCHL5,LDL5, TG5
2	Female, Waist4, TCHL5, LD1, TG5
3	Male, Waist2, TCHL5, LDL5, TG3
4	Male, Waist5, TCHL5, LDL5, TG5
5	Male, Waist1, TCHL4, LDL2, TG5

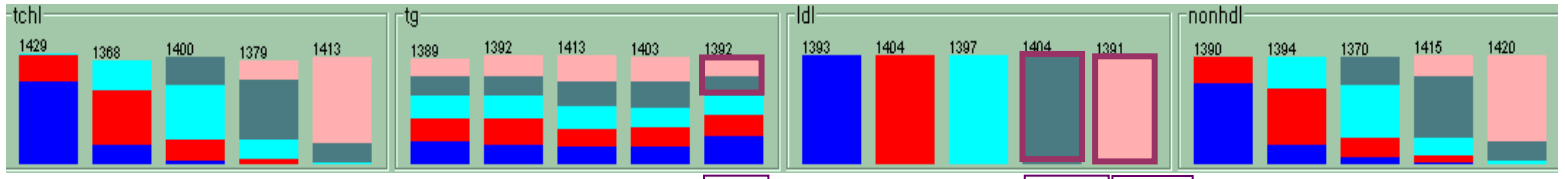
Association Rules Discovered

- 10, 162 rules contains high levels of TG (TG5) with reliable confidence (conf ≥0.7)
- 359 rules are associated with high level of TG(TG5) and high levels of LDL(LDL4-5)

#	Rule Body		Rule Head	Supp	Conf	Lift
	Rules encoding highTG levels					
1	LDL5, BMI5, **TG5**, TCHL5	→	NONHDL5	0.0157	1.0000	5.1732
2	**TG5,** TCHL5, WHR5, LDL5	→	NONHDL5	0.0136	1.0000	5.1732
3	**TG5**, PLAT5, LDL5, NONHDL5		TCHL5	0.0127	1.0000	5.1465
4	**TG5**, LDL5, SBP4, TCHL5	→	NONHDL5	0.0126	1.0000	5.1732
5	DBP5, **TG5**, LDL5, TCHL5	→	NONHDL5	0.0122	1.0000	5.1465
6	LDL5, **TG5**, TCHL5, GLU05	→	NONHDL5	0.0122	1.0000	5.1465
7	GLU605, **TG5**, TCHL5, LDL5	→	NONHDL5	0.0120	1.0000	5.1732
8	GLU1205, TCHL5, LDL5, **TG5**	→	NONHDL5	0.0119	1.0000	5.1732
9	**TG5**, TCHL5, INS05, LDL5	→	NONHDL5	0.0119	1.0000	5.1732
10	INS605, TCHL5, **TG5**, LDL5	→	NONHDL5	0.0116	1.0000	5.1732
11	LDL5, INS1205, TCHL5, **TG5**	→	NONHDL5	0.0107	1.0000	5.1732
12	**TG5**, LDL5, TCHL5, DS1	→	NONHDL5	0.0107	1.0000	5.1732
13	T_HDL5, NONHDL5, LDL4, HDL2	→	**TG5**	0.0102	0.8875	4.1105
14	TCHL2, LDL1 NONHDL2, PH1	→	**TG5**	0.0100	0.8333	3.8594
	Rules encoding highLDL levels					
15	TCHL5, NONHDL5, GLU605	→	**LDL5**	0.0405	0.8324	4.2651
16	DS1, NONHDL5, GLU605	→	**LDL5**	0.0243	0.8333	4.2697
17	NONHDL5, SONA4, TCHL5	→	**LDL5**	0.0242	0.8450	4.3297
18	TG3, BUN5, NONHDL5	→	**LDL5**	0.0114	1.0000	5.1239

TG: Triglyceride
LDL: low density lipoprotein cholesterol

TG1 < 91.5
91.5 ≤ TG2 < 116.5
116.5 ≤ TG3 < 146.5
146.5 ≤ TG4 < 196.5
TG5 ≥ 195.6
LDL1 < 89.3
89.3 ≤ LDL2 < 106.9
106.9≤ LDL3 < 122.9
122.9≤ LDL4 < 143.5
LDL5 ≥ 143.5



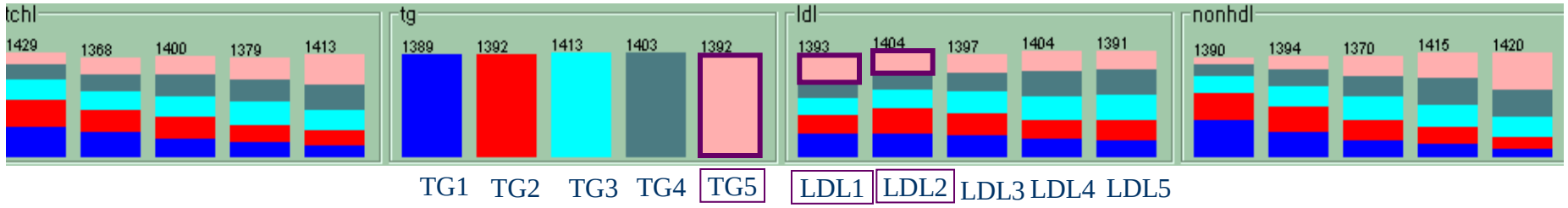
Interesting Association Rules

- **49 rules are associated with low LDL and high TG**

rule #	rule boby		rule head	supp.	conf.
1	NONHDL4, LDL2	→	TG5	0.014	1.000
2	NONHDL3, LDL2, HDL1	→	TG5	0.013	0.841
3	LDL2, T_HDL5	→	TG5	0.019	0.838
4	TCHL2, LDL1, NONHDL2, PH1	→	TG5	0.010	0.833
5	HDL1, T_HDL5, LDL2	→	TG5	0.017	0.818
6	NONHDL2, LDL1, TCHL2	→	TG5	0.018	0.805
7	NONHDL3, LDL2, CHL_CI2	→	TG5	0.012	0.783
8	T_HDL4, LDL1	→	TG5	0.013	0.759
9	NONHDL3, LDL2	→	TG5	0.027	0.756
10	NONHDL2, SG1, LDL1	→	TG5	0.013	0.746
11	T_HDL4, HDL1, LDL1	→	TG5	0.011	0.728

TG1 < 91.5
91.5 ≤ TG2 < 116.5
116.5 ≤ TG3 < 146.5
146.5 ≤ TG4 < 196.5
TG5 ≥ 195.6

LDL1 < 89.3
89.3 ≤ LDL2 < 106.9
106.9≤ LDL3 < 122.9
122.9≤ LDL4 < 143.5
LDL5 ≥ 143.5



low LDL → high TG

2017-06-26 Sung H. Park 9

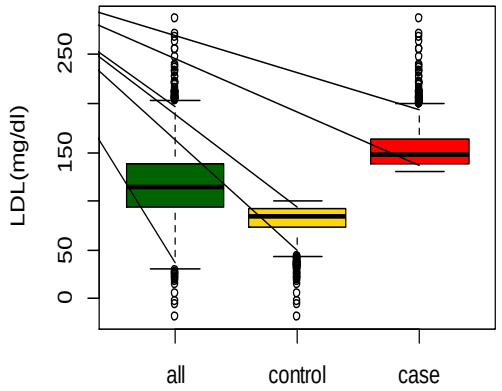
Association Analysis for Multivariate phenotype

Study1: Find SNPs associated with *highLDLhighTG*

❑ Stage1 : GWA for highLDLhighTG

(mg/dl)

	control	case
condition	LDL-C ≤ 100 and TG ≤ 100	**LDL-C ≥ 130 and TG ≥ 180**
phenotyped ind.	681(9.74 %)	545 (7.80%)

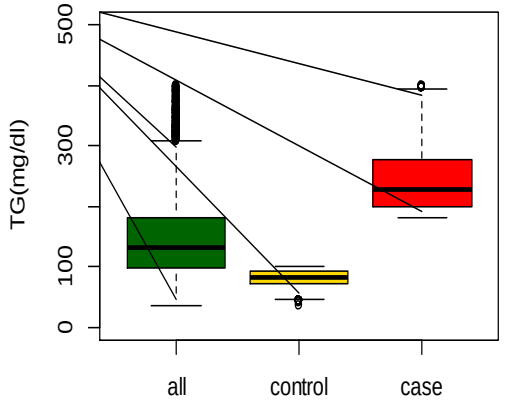


❑ Stage2: GWA for highLDL and GWA for high TG

(mg/dl)

	control	case
condition	LDL-C ≤ 100	LDL-C ≥ 130
phenotyped ind.	2215 (31.69%)	2271 (32.49%)

	control	case
condition	TG ≤ 100	TG ≥ 180
phenotyped ind.	1914 (27.39%)	1779 (25.45%)



2017-06-26 Sung H. Park 10

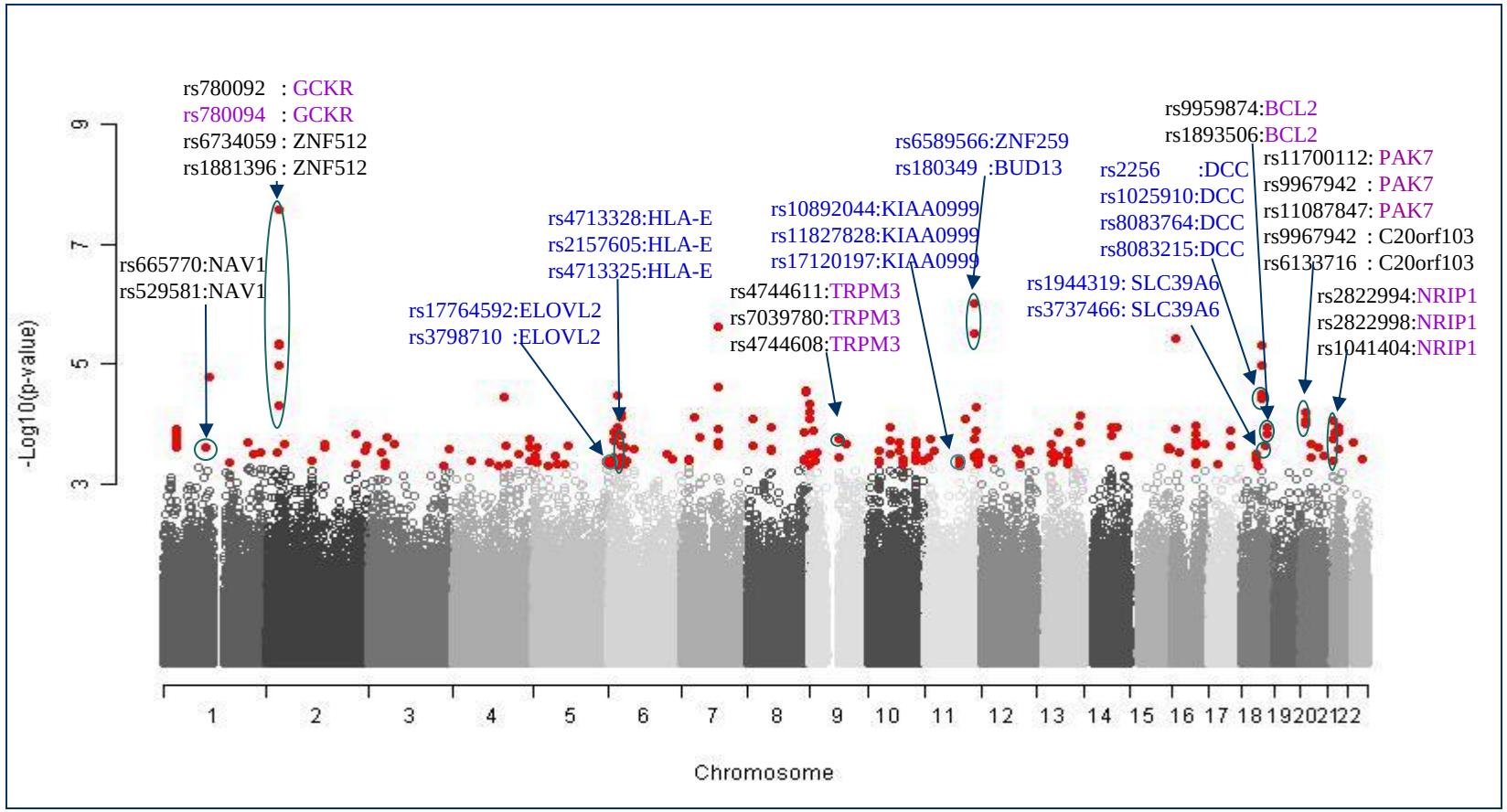
Association Analysis: Filtering SNPs

SNPs are filtered by the following conditions:

- P value for association test of Multivariate phenotype $\leq 5 \times 10^{-4}$
- -Log odd ratio (OR_m) ≥ 1
  - $OR_m = -\text{Log}_{10}\left(\frac{\text{P value of Stage1}}{\text{P value of Stage2}}\right)$
  - $-\text{Log}_{10}\left(\frac{\text{P value (highLDLhighTG)}}{\text{P value (highLDL} \cup \text{highTG)}}\right) \geq 1$
- LD clumping:
  - p value of an indexed SNP $< 10^{-5}$
  - p value of clumped SNPs $< 10^{-4}$

2017-06-26 Sung H. Park 11

highLDLhighTG GWAS results

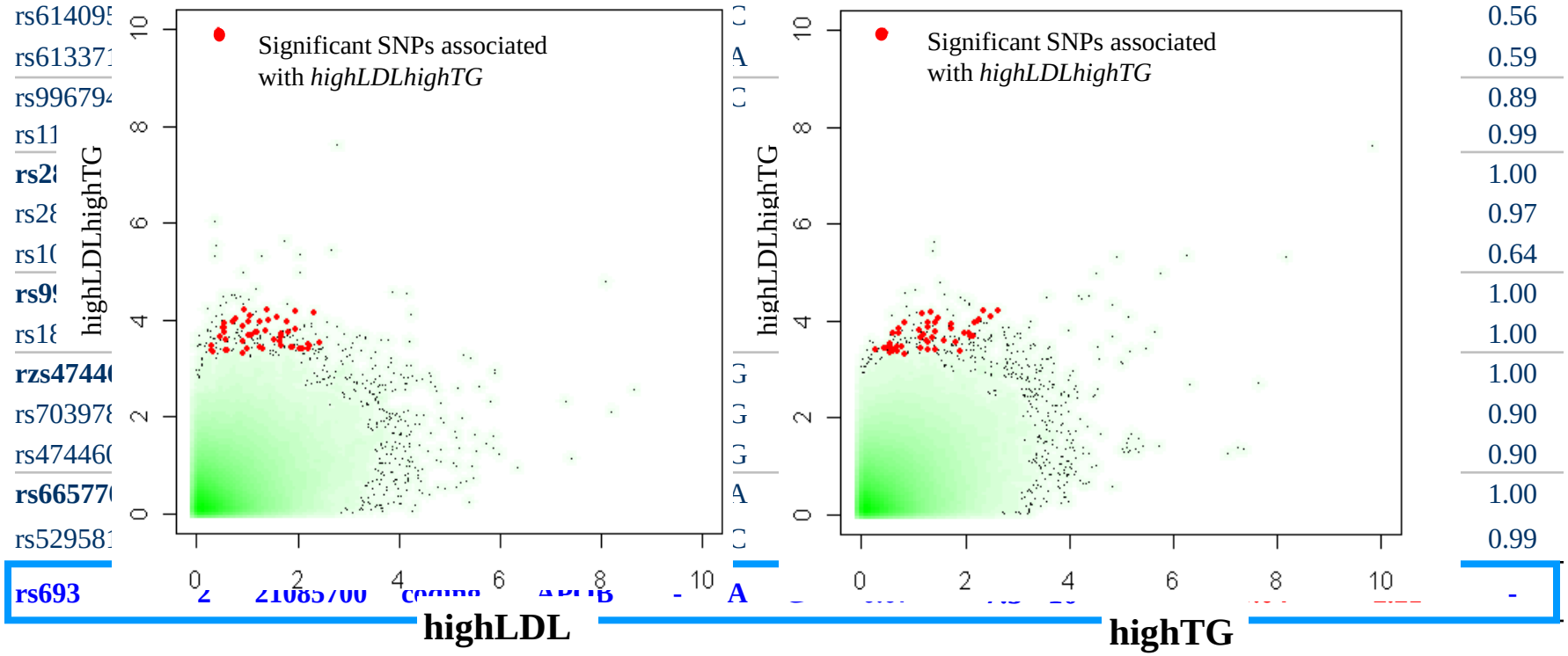


rs numbers in blue were pruned

Gene name in purple was identified in previous GWAS of lipids related traits (Kathiresan 2008, Lu 2008, Noron06) .

Genetic variants associated with *highLDLhighTG*

SNP	Chr	BP	SNP type	gene	Str-and	Allele (+/-)		Freq (+)	P-value hLDLhTG	OR	mOR_p^{LDL}	mOR_p^{TG}	r^2
rs11700112	20	9495018	**nonsyn**	PAK7	-	**G**	**C**	0.31	**6.2×10^{-5}**	1.44	3.25	1.58	1.00
rs61409													0.56
rs61337							A						0.59
rs99679													0.89
rs11													0.99
rs2													1.00
rs2													0.97
rs1													0.64
rs9													1.00
rs1													1.00
rzs4744													1.00
rs70397													0.90
rs47446													0.90
rs66577							A						1.00
rs52958													0.99
rs693	2	21085700	c6qinp	ABHB	-	**A**							-



highLDL **highTG**

SNP rs693 reported in a previous study (Kathiresan, et al, 2008) for associations between high TG and high LDL.
an indexed SNP for LD was **in bold**

In silico Replication

		Ansung			Ansan			Combined	Male			Female			Combined
gene	SNP	LDLTG n=205	LDL-C n=919	TG n=936	LDLTG n=340	LDL-C n=1352	TG n=843	LDLTG n=545	LDLTG n=288	LDL-C n=1044	TG n=969	LDLTG n=257	LDL-C n=1277	TG n=810	LDLTG n=545

PA

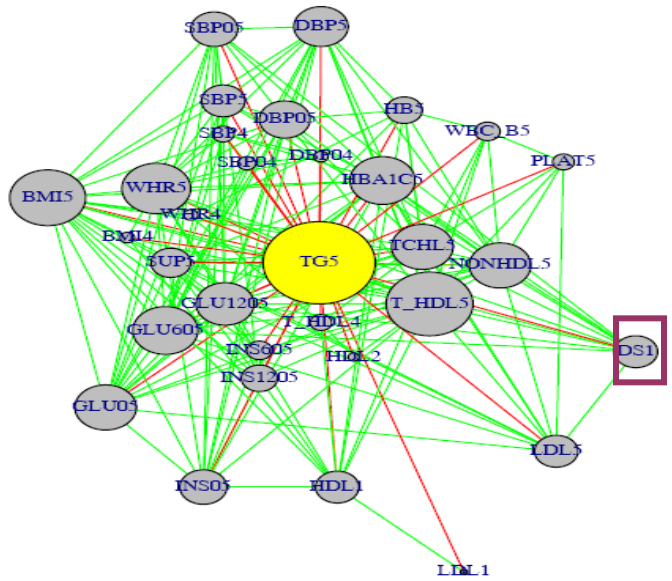
Bone. 2006 Jul;39(1):213-21. Epub 2006 Mar 10.

C20o **Multilocus analysis of estrogen-related genes in Spanish postmenopausal women suggests an interactive role of ESR1, ESR2 and NRIP1 genes in the pathogenesis of osteoporosis.**

PA Morón FJ, Mendoza N, Vázquez F, Molero E, Quereda F, Salinas A, Fontes J, Martínez-Astorquiza T, Sánchez-Borrego R, Ruiz A.

Departame… Hospital Un… que Tecnológico Isla de la Cartuja, 41092-Sevilla, and Servicio de Ginecología y Obsetricia,

NR **Abstrac**



Osteoporo… estrogen-… Looking fo… Follicle St… Receptor … nominal P… test corre… (P=0.045)… osteoporo… Replicatio… model inc… the geneti…

netic risk factors involved. Using a marker-by-marker approach, the role of different nost of these studies ignore the complex multigenic nature of human osteoporosis. e nucleotide polymorphisms located in genes related to the estrogen pathway, ase (CYP19A1) gene, the Estrogen Receptor alpha (ESR1) gene, the Estrogen 1 (NRIP1) gene in 265 unrelated postmenopausal women. We have obtained =0.013 and P=0.02 respectively), but no gene seems to be associated after multiple men confirmed our results and only detect marginal effects for ESR2 marker is between ESR1, ESR2 and NRIP1 loci and its involvement in postmenopausal SR2-NRIP1 and ESR2-ESR1 genes strongly associated with osteoporosis (P=0.007). ted interactions (P<0.01). We proposed a non-additive non-multiplicative oligogenic genotypes involved in osteoporosis. Our results reaffirm the polygenic nature and e (NRIP1) for association studies of bone-related traits.

NA

AP

Significant SNPs (p < 0.05)

The Author(s) *BMC Bioinformatics* 2016, **17**(Suppl 16):452
DOI 10.1186/s12859-016-1307-z

BMC Bioinformatics

RESEARCH Open Access

ChARM: Discovery of combinatorial chromatin modification patterns in hepatitis B virus X-transformed mouse liver cancer using association rule mining

Sung Hee Park[1], Sun-Min Lee[2], Young-Joon Kim[2,3*] and Sangsoo Kim[1*]

From The 10th International Workshop on Machine Learing in Systems Biology (MLSB)
Den Haag, The Netherlands. 3-4 September 2016

Data: Epigenetic modifications

Livers of C57BL/6(norm) 3 month old mice (10 samples)
Livers of Hepatitis B virus X protein(HBx) transgenic 3 month old mice (10 samples)

ChIP-seq: Chromatin immunoprecipitation & SOLEXA sequencing

H3K4me3 – active mark in promoter

H3K27me3 – repressive mark in promoter

H3K36me3 – active mark in exon

DNA methylation – repressive mark in promoter, active or repressive mark in genebody

RNA Pol2 serine5phosphate – initiation complex in promoter

RNA Pol2 serine2phosphate – elongation complex in gene-body

Number of uniquely matched reads in each experiment

Experiment \ # Reads	Normal	HBx
Mnase-seq	20,232,106	12,814,254
FAIRE-seq	6,429,571	9,409,055
H3K4me3, ChIP-seq	8,961,493	10,949,790
H3K27me3, ChIP-seq	9,206,684	9,456,550
DNA methylation, MIRA-seq	19,758,627	18,558,297
Pol2s5p, ChIP-seq	11,733,916	12,071,214

Research Problems to address

- Are there cancer specific aberrant chromatin modification signatures in contrast to normal ones ?
- What are the promoter or gene-body specific patterns on genome-wide scale ?
- What are the shapes of the patterns, expressions and functions?

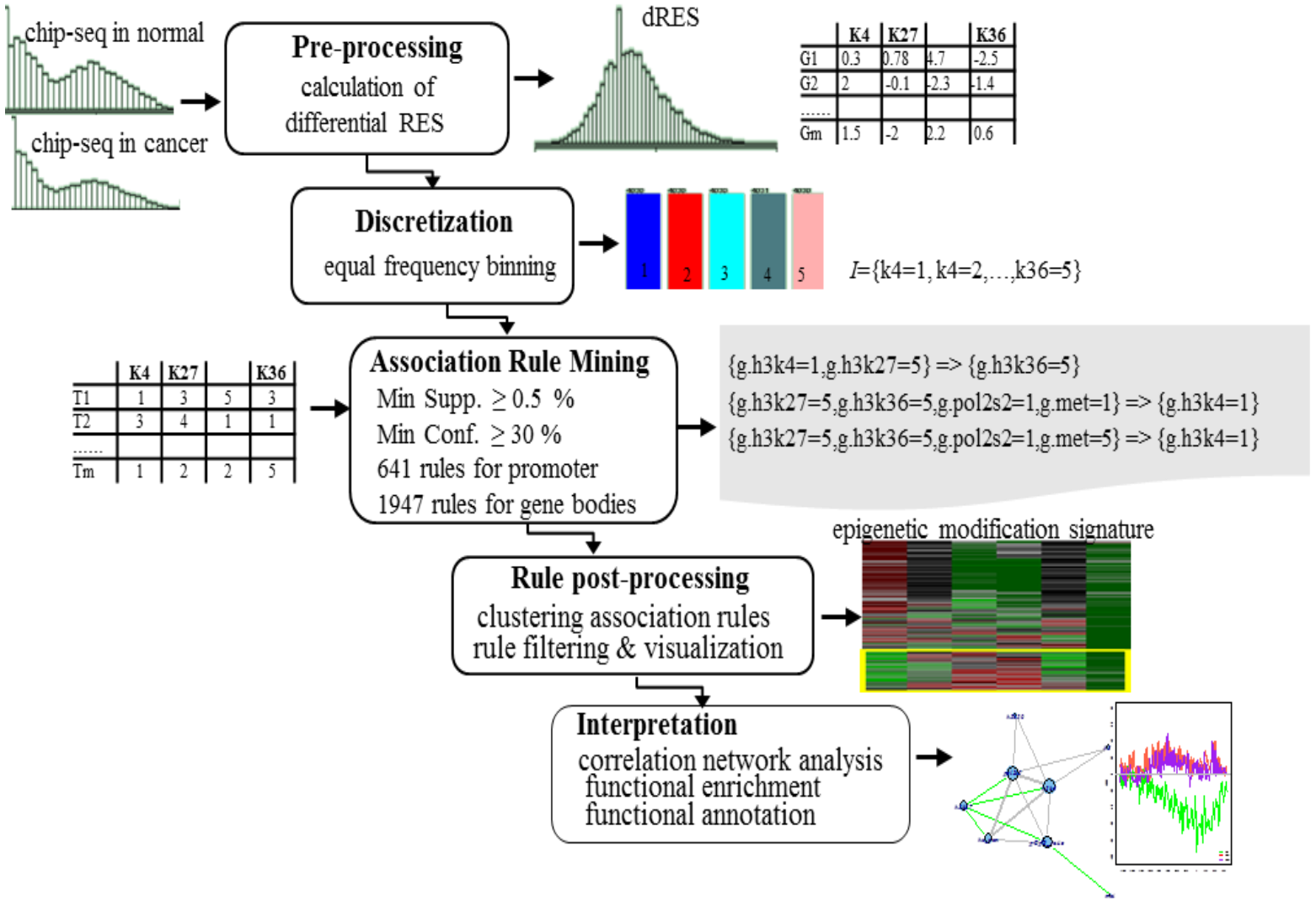
- **Discover global combinatorial patterns of chromatin modifications(GCPCM) characterizing cancer cells**
  - global combinatorial patterns → the most frequently occurring combination of chromatin modifications across different loci, e.g., promoters, gene bodies and tiles
  - Differentially modified patterns → different conditions, e.g. cancer vs. normal cells

  → A global combination of chromatin modifications differentially occurring in two different cell types .

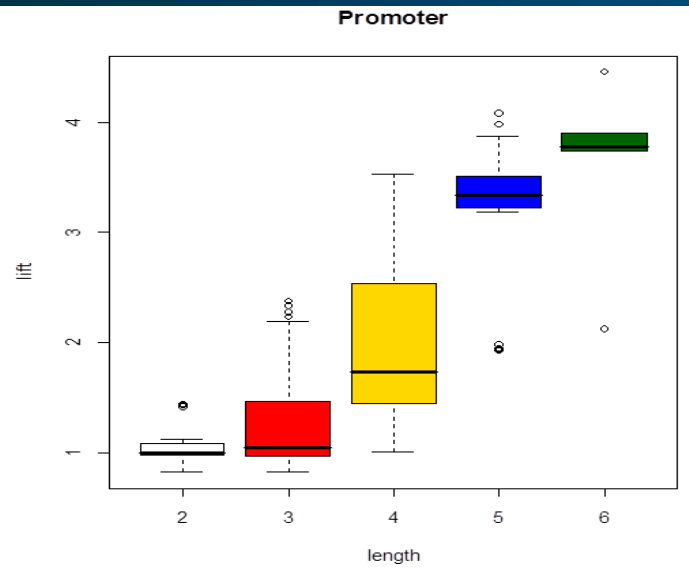
A workflow of *ChARM* method

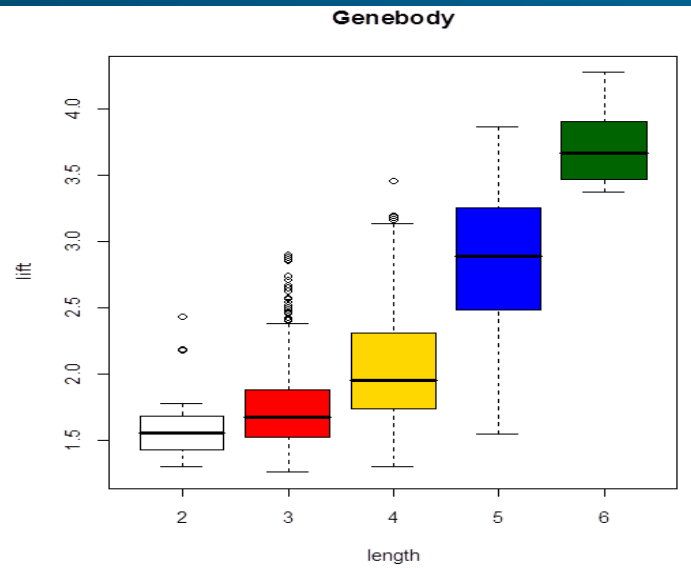
RES: ((target read count/target size)/(total read count/genome size))

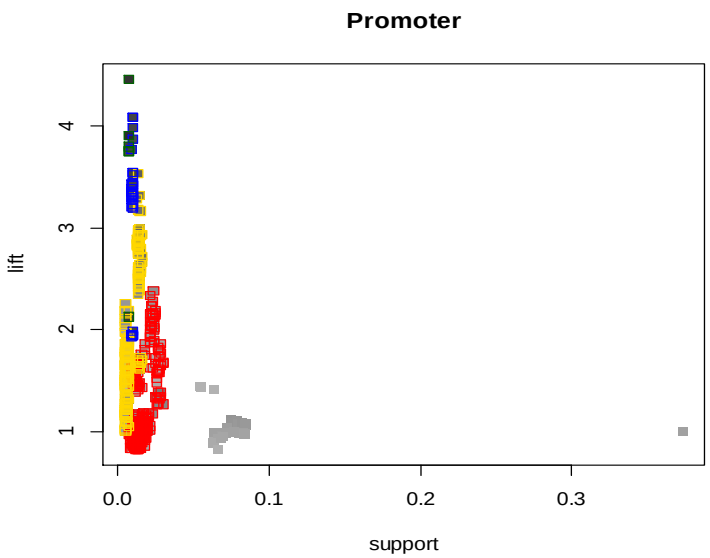
H3K4me3

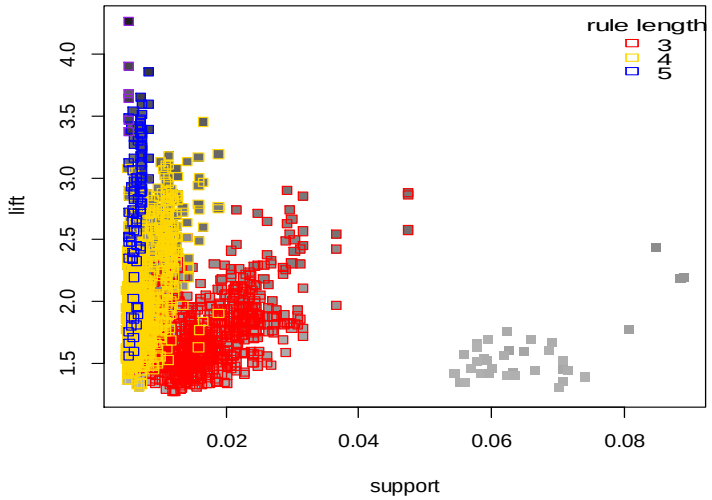


rule length & lift









Representative Association rules

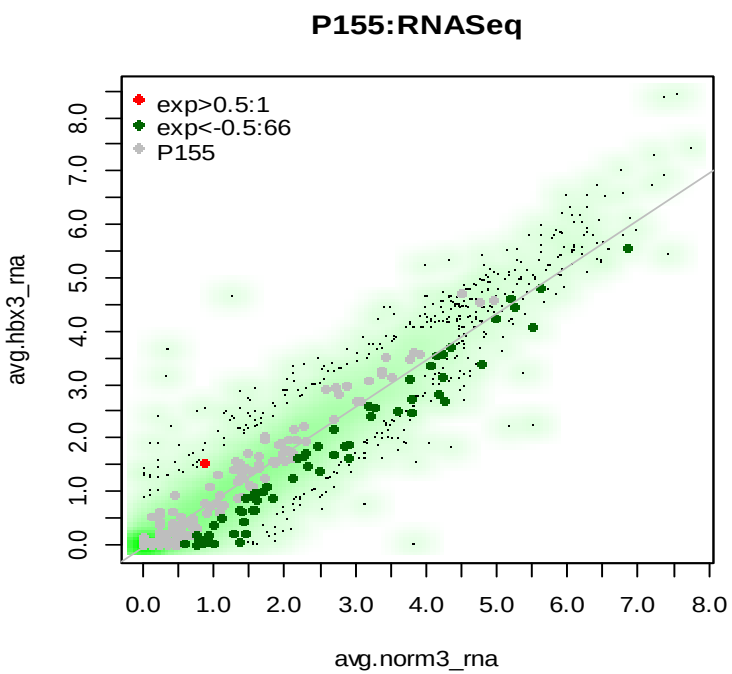
No	Rule Description for promter	support	conf	lift	Annotation
1	p.h3k27=5 p.h3k36=5 ==> p.h3k4=1	0.018	0.350	1.865	P155
2	p.h3k4=1 p.h3k36=5 ==> p.h3k27=5	0.018	0.340	1.744	P155
3	p.h3k4=1 p.h3k27=5 ==> p.h3k36=5	0.018	0.330	1.805	P155
4	p.h3k27=5 p.h3k36=5 p.pol2s5=1 ==> p.h3k4=1	0.005	0.410	2.172	Super set & highest lift
5	p.h3k4=1,p.h3k27=5,p.h3k36=5 => p.met=2	0.007	0.392	1.052	Super set & Lowest lift
6	p.h3k27=5 p.h3k36=5 p.met=2 ==> p.h3k4=1	0.007	0.360	1.888	Super set & Top 5 lift
7	p.h3k4=1 p.h3k27=5 p.met=2 ==> p.h3k36=5	0.007	0.330	1.798	Super set & Top 5 lift
8	p.h3k4=1 p.h3k36=5 p.pol2s5=1 ==> p.h3k27=5	0.005	0.340	1.736	Super set & Top 10 lift
9	p.h3k4=1 p.h3k27=5 p.pol2s5=1 ==> p.h3k36=5	0.005	0.330	1.774	Super set & Top 10 lift
10	p.h3k4=4,p.h3=3,p.h3k27=2,p.h3k36=2,p.met=2 => p.pol2s5=3	0.008	0.801	4.459	Top 5 lift
11	p.h3k4=4,p.h3=3,p.h3k27=2,p.h3k36=2 => p.pol2s5=3	0.010	0.734	4.085	Top 5 lift
12	p.met=2	0.373	0.373	1.000	Top 5 support
13	p.pol2s5=1 => p.met=2	0.084	0.398	1.068	Top 5 support
14	p.h3k27=3 => p.met=2	0.083	0.364	0.976	Top 5 support
	Rule Description for genebody				
15	g.h3k27=5 g.h3k36=5 ==> g.h3k4=1	0.048	0.561	2.576	G155
16	g.h3k4=1 g.h3k27=5 ==> g.h3k36=5	0.048	0.540	2.878	G155
17	g.h3k4=1 g.h3k36=5 ==> g.h3k27=5	0.048	0.530	2.863	G155
18	g.h3k4=1,g.h3k27=5,g.pol2s2=1,g.met=5 => g.h3k36=5	0.006	0.661	3.543	Super set & Top 5 lift
19	g.h3k4=1,g.h3=5,g.h3k36=5,g.met=1 => g.h3k27=5	0.005	0.650	3.486	Super set & Top 5 lift
20	g.h3k4=1,g.h3k36=5,g.met=1 => g.h3k27=5	0.017	0.644	3.453	Super set & Top 5 lift
21	g.h3k4=1,g.h3k36=5,g.pol2s2=1,g.met=1 => g.h3k27=5	0.007	0.638	3.423	Super set & Top 5 lift
22	g.h3k4=1,g.h3=1,g.h3k36=5,g.met=1 => g.h3k27=5	0.006	0.633	3.395	Super set & Top 5 lift
23	g.h3k4=1,g.h3k27=5,g.h3k36=5,g.met=1 => g.h3=5	0.0053	0.318	1.55282	Super set & lowest lift
24	g.h3=3,g.h3k27=3,g.pol2s2=4,g.met=2 => g.h3k36=4	0.008	0.793	3.859	Top 5 lift
25	g.h3=3,g.h3k27=3,g.h3k36=4,g.pol2s2=4,g.met=2 => g.h3k4=4	0.005	0.639	3.682	Top 5 lift
26	g.h3k36=5 => g.h3k4=1	0.089	0.477	2.190	Top 5 support
27	g.h3k4=1 => g.h3k36=5	0.089	0.409	2.190	Top 5 support
28	g.h3k27=5 => g.h3k4=1	0.088	0.474	2.179	Top 5 support
29	g.h3k4=1 => g.h3k27=5	0.088	0.406	2.179	Top 5 support
30	g.h3k27=5 => g.h3k36=5	0.085	0.454	2.434	Top 5 support
31	g.h3k36=5 => g.h3k27=5	0.085	0.454	2.434	Top 5 support

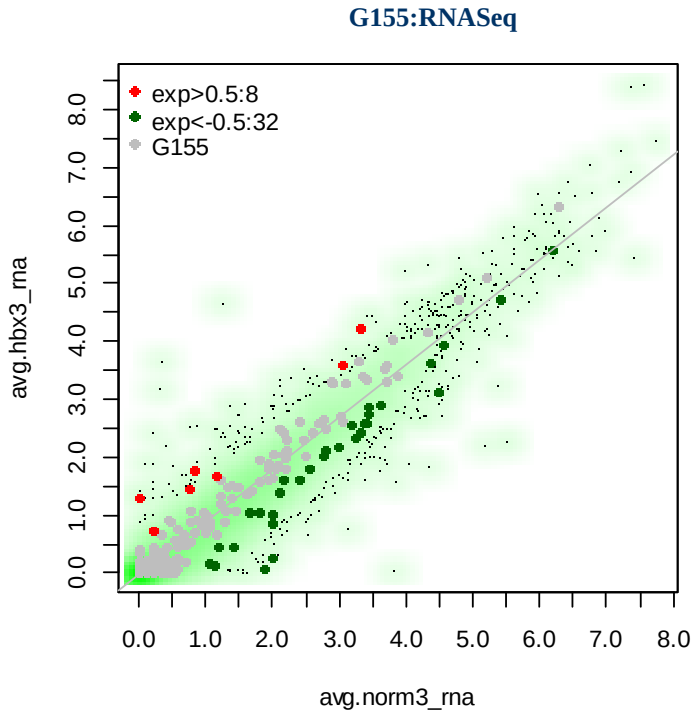
p.h3k27=5 p.h3k36=5 ==> p.h3k4=1
p.h3k4=1 p.h3k36=5 ==> p.h3k27=5
p.h3k4=1 p.h3k27=5 ==> p.h3k36=5

g.h3k27=5 g.h3k36=5 ==> g.h3k4=1
g.h3k4=1 g.h3k27=5 ==> g.h3k36=5
g.h3k4=1 g.h3k36=5 ==> g.h3k27=5

Expression changes in the patterns

In RNA-Seq, 86 (78.9%) of 109 differential expressed genes were down-regulated in the promoter pattern.





Functional enrichment analysis using IPA & DAVID

	Category	Term or pathway	*P*-value
Promoter			
	SP_PIR_KEYWORDS	transcription regulation	2.00E-08
	GOTERM_MF	transcription regulator activity	6.72E-06
	GOTERM_BP	regulation of transcription from RNA polymerase II promoter	8.83E-06
	GOTERM_MF	transcription factor activity	2.56E-05
	SP_PIR_KEYWORDS	phosphoprotein	5.31E-05
	SP_PIR_KEYWORDS	dna-binding	9.99E-05
	GOTERM_BP	regulation of RNA metabolic process	1.31E-04
	GOTERM_BP	positive regulation of transcription	1.43E-04
	SP_PIR_KEYWORDS	developmental protein	1.56E-04
	Canonical pathway	**Role of NFAT in cardiac Hypertrophy**	**4.36E-06**
		Wnt/β-catenin signaling	**2.42E-04**
		Molecular Mechanisms of Cancer	**3.91E-04**
		cAMP-mediated signaling	**5.60E-04**
		Dopamine-DARPP32 Feedback in cAMP signaling	**6.05E-04**
Genebody	GOTERM_MF	DNA binding	1.82E-07
	INTERPRO		3.34E-06
	GOTERM_MF	sequence-specific DNA binding	9.11E-06
	SP_PIR_KEYWORDS	developmental protein	9.66E-06
	GOTERM_MF	transcription regulator activity	4.65E-05
	GOTERM_MF	transcription factor activity	9.25E-05
	SP_PIR_KEYWORDS	transcription regulation	1.62E-04
	Canonical pathway	**Notch Signaling**	**4.89E-05**

only annotations that are with a corrected p-value <0.02 after Benjamini-Hochberg correction for multiple hypothesis testing.
Full lists and more details are provided in supplemental Table .

Disscussion and Conclusion

- ***ChARM*, an unsupervised method for discovering de novo combinatorial chromatin modification patterns occurring globally and differntialy between two different conditions**
  - applied *ChARM* to investigate an HBx-transformed mouse liver tumour model and discovered an aberrant histone modification pattern
    - a combination of a loss of H3K4Me3 and gains of H3K27Me3 and H3K36Me3
    - The pattern characterised with CpG content of underlying DNA sequences
    - H3K27Me3 and H3K36Me3 hypermethylation in HBx are occurred in intermediate promoter regions where CpG ratio is low..
    - The significant canonical pathways enriched in the pattern accounted for the pathogenesis of HBx and were linked to a general cancer pathway.
      - E.g., Notch signalling, and Wnt/β-catenin, cAMP mediated, and Ras pathway
  - identify combinatorial singures of differentialy modified regions without prior knowledge
  - provides a scalable framework to find various levels of combination patterns, which should reflect a range of globally common to locally rare chromatin modifications.

감사합니다!!!

Thank you for your attention!

多谢 晒

2017-06-26 Sung H. Park