Bayesian Nonparametric Clustering and Association Studies for Large-scale SNP Observations

Charlotte Wang

Department of Mathematics, Tamkang University

This is a joint work with Dr. Raffaele Argiento, Dr. Fabrizio Ruggeri (CNR IMATI - Milano, Italy) and Dr. Chuhsing Kate Hsiao (CPH, NTU)

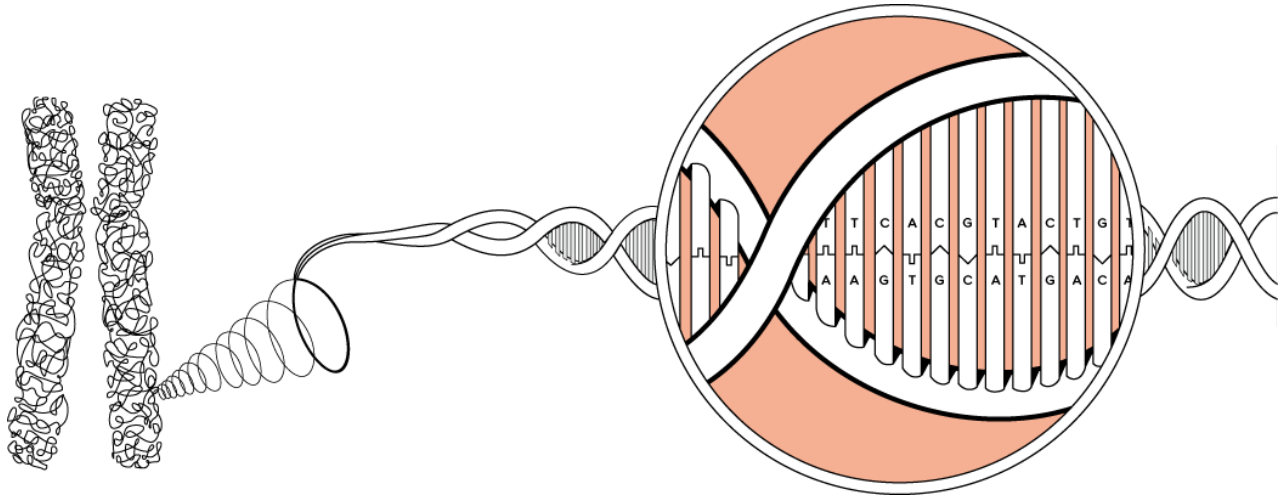
October 5, 2017

Outline

Outline

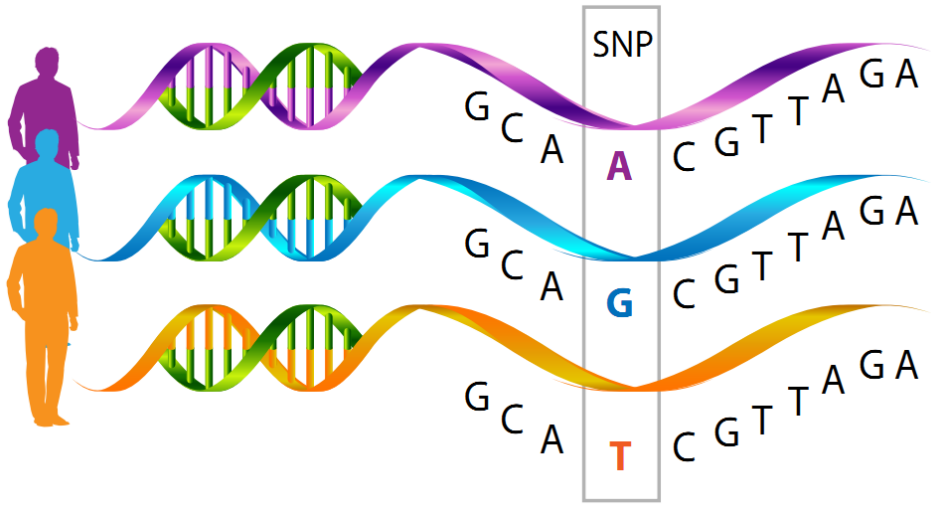
Human Genome

- The human genome is an ordered sequence of 22+1 pairs of chromosomes.
- Chromosome is made up of deoxyribonucleic acid (DNA) tightly coiled many times around proteins called histones that support its structure.
- DNA is double-stranded and composed of a phosphate group, a five carbon sugar and a nitrogenous base (A, T, C, G).



What are SNPs?

- Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people.
- Single nucleotide polymorphisms (SNP) is a single-nucleotide difference between two or more individuals at a particular genetic locus.
- 150,482,731 SNPs were fund in human genome (NCBI dbSNP Build 146, 2015/11/24). That is, in human genome there is in the average of 50 SNPs per 1kb.



SNPs Data Coding

- We use the term allele to identify which nucleotide is present at a SNP.
- There are three kinds of genotypes, e.g. AA/AG/GG.
- The SNP coding usually indicates the number of minor alleles of an individual carrying.

Alleles of SNP rs123556 (located on chromosome 1)

Michelangelo:	CTTAGATTCAT **G** TCACTAGCTAGG
	CTTAGATTCAT **G** TCACTAGCTAGG
Raffaello:	CTTAGATTCAT **G** TCACTAGCTAGG
	CTTAGATTCAT **A** TCACTAGCTAGG
Leonardo:	CTTAGATTCAT **G** TCACTAGCTAGG
	CTTAGATTCAT **G** TCACTAGCTAGG
Caravaggio:	CTTAGATTCAT **A** TCACTAGCTAGG
	CTTAGATTCAT **A** TCACTAGCTAGG

(Argiento et al., 2015)

SNPs Data Coding

SNP coding

Suppose A is the minor allele in the SNP rs123556, then

SNP		X
Michelangelo	$\xrightarrow{\text{genotype}}$	0
Raffaello	$\xrightarrow{\text{genotype}}$	1
Leonardo	$\xrightarrow{\text{genotype}}$	0
Caravaggio	$\xrightarrow{\text{genotype}}$	2

eotide is present at a

/AG/GG.

of minor alleles of

6 (located on chromosome 1)

Michelangelo:	CTTAGATTCAT **G** TCACTAGCTAGG CTTAGATTCAT **G** TCACTAGCTAGG
Raffaello:	CTTAGATTCAT **G** TCACTAGCTAGG CTTAGATTCAT **A** TCACTAGCTAGG
Leonardo:	CTTAGATTCAT **G** TCACTAGCTAGG CTTAGATTCAT **G** TCACTAGCTAGG
Caravaggio:	CTTAGATTCAT **A** TCACTAGCTAGG CTTAGATTCAT **A** TCACTAGCTAGG

(Argiento et al., 2015)

Outline

Genetic Association Study

Genetic variants vs. disease of interest

- Single-marker test: low power, multiple comparison
- Test a group of markers
  - Multiple-marker test: df, same effect, genotype/allele frequency
  - Haplotype analysis: phase, composition, computational burden
  - SNP-set/Gene-set/pathway analysis: joint effects, power

Unphased haplotypeg

Heterozygous genotypes at 3 sites

AC TG AT

The 4 possible consistent pairs of haplotypes

ATT	ATA	AGT	AGA
CGA	CGT	CTA	CTT

Genetic Association Study

Genetic variants vs. disease of interest

- Single-marker test: low power, multiple comparison
- Test a group of markers
  - Multiple-marker test: df, same effect, genotype/allele frequency
  - Haplotype analysis: phase, composition, computational burden
  - SNP-set/Gene-set/pathway analysis: joint effects, power

Unphased haplotypeg

Heterozygous genotypes at 3 sites

AC TG AT

The 4 possible consistent pairs of haplotypes

ATT	ATA	AGT	AGA
CGA	CGT	CTA	CTT

Association Studies with Multiple Makers

- More powerful, efficient, biologically meaningful Ackermann *et al.*, 2009; Asimit and Zeggini, 2010
- e.g. regularized regression models Chen*et. al.*, 2009; Malo *et. al.*, 2008; Chen *et. al.*, 2010; Zhou *et. al.*, 2010; Li *et. al.*, 2011, gene-set analysis Subramanian *at al.*, 2005; Efron and Tibshirani, 2007; Hu and Tzeng, 2014, pathway analysis Ramanan *et. al.* 2012, network analysis Lee *et. al.*, 2011
- Analyze genes in the same pathway or network from pre-specified genetic regions
- At the genome-wide level, the relation among a massive set of genes is unclear

Association Studies with Multiple Makers

- More powerful, efficient, biologically meaningful Ackermann *et al.*, 2009; Asimit and Zeggini, 2010
- e.g. regularized regression models Chen*et. al.*, 2009; Malo *et. al.*, 2008; Chen *et. al.*, 2010; Zhou *et. al.*, 2010; Li *et. al.*, 2011, gene-set analysis Subramanian *at al.*, 2005; Efron and Tibshirani, 2007; Hu and Tzeng, 2014, pathway analysis Ramanan *et. al.* 2012, network analysis Lee *et. al.*, 2011
- Analyze genes in the same pathway or network from pre-specified genetic regions
- At the genome-wide level, the relation among a massive set of genes is unclear

Association Studies with Multiple Makers

- More powerful, efficient, biologically meaningful Ackermann *et al.*, 2009; Asimit and Zeggini, 2010
- e.g. regularized regression models Chen*et. al.*, 2009; Malo *et. al.*, 2008; Chen *et. al.*, 2010; Zhou *et. al.*, 2010; Li *et. al.*, 2011, gene-set analysis Subramanian *at al.*, 2005; Efron and Tibshirani, 2007; Hu and Tzeng, 2014, pathway analysis Ramanan *et. al.* 2012, network analysis Lee *et. al.*, 2011
- Analyze genes in the same pathway or network from pre-specified genetic regions
- At the genome-wide level, the relation among a massive set of genes is unclear

Association Studies with Multiple Makers

- More powerful, efficient, biologically meaningful Ackermann *et al.*, 2009; Asimit and Zeggini, 2010
- e.g. regularized regression models Chen*et. al.*, 2009; Malo *et. al.*, 2008; Chen *et. al.*, 2010; Zhou *et. al.*, 2010; Li *et. al.*, 2011, gene-set analysis Subramanian *at al.*, 2005; Efron and Tibshirani, 2007; Hu and Tzeng, 2014, pathway analysis Ramanan *et. al.* 2012, network analysis Lee *et. al.*, 2011
- Analyze genes in the same pathway or network from pre-specified genetic regions
- At the genome-wide level, the relation among a massive set of genes is unclear

Association Studies based on Bayesian Analysis

- Case-control study
- Bayes factors for hypothesis testing
- Mixture model for every single markers
- No clustering procedure in advance
- The parameters of single-marker effects were assumed exchangeable while conditioning on hyper-prior parameter values

(Wakefield, 2007; Wakefield, 2009; Wei *et al.*, 2010; Wakefield, 2009)

Association Studies based on Bayesian Analysis

- Case-control study
- Bayes factors for hypothesis testing
- Mixture model for every single markers
- No clustering procedure in advance
- The parameters of single-marker effects were assumed exchangeable while conditioning on hyper-prior parameter values

(Wakefield, 2007; Wakefield, 2009; Wei *et al.*, 2010; Wakefield, 2009)

Association Studies based on Bayesian Analysis

- Case-control study
- Bayes factors for hypothesis testing
- Mixture model for every single markers
- No clustering procedure in advance
- The parameters of single-marker effects were assumed exchangeable while conditioning on hyper-prior parameter values

(Wakefield, 2007; Wakefield, 2009; Wei *et al.*, 2010; Wakefield, 2009)

Cluster Analysis for Genetic Sequencing Data

- An exploratory tool to integrate the information contained in SNPs
- Define analytic regions or to detect possible the relationship among genes if no prior knowledge
- Current clustering algorithms, e.g. k-means, hierarchical clustering methods
  - For quantitative measurements
  - Similarity metric: Euclidean distance and Mahalanobis distance
- Few clustering algorithms for discrete data
  - *k*-mode (Huang, 1997): need to pre-sepecify *k*
  - Zhang *et al.*(2006): computationally complex, low-demential dataset
- Few clustering algorithms for SNPs data
  - PCA (Paschou *et al.*, 2007)
  - Selinski and Ickstadt (2008)
  - HD-Cluster (Wang *et al.*, 2015)

Cluster Analysis for Genetic Sequencing Data

- An exploratory tool to integrate the information contained in SNPs
- Define analytic regions or to detect possible the relationship among genes if no prior knowledge
- Current clustering algorithms, e.g. k-means, hierarchical clustering methods
  - For quantitative measurements
  - Similarity metric: Euclidean distance and Mahalanobis distance
- Few clustering algorithms for discrete data
  - *k*-mode (Huang, 1997): need to pre-sepecify *k*
  - Zhang *et al.*(2006): computationally complex, low-demential dataset
- Few clustering algorithms for SNPs data
  - PCA (Paschou *et al.*, 2007)
  - Selinski and Ickstadt (2008)
  - HD-Cluster (Wang *et al.*, 2015)

Cluster Analysis for Genetic Sequencing Data

- An exploratory tool to integrate the information contained in SNPs
- Define analytic regions or to detect possible the relationship among genes if no prior knowledge
- Current clustering algorithms, e.g. k-means, hierarchical clustering methods
  - For quantitative measurements
  - Similarity metric: Euclidean distance and Mahalanobis distance
- Few clustering algorithms for discrete data
  - *k*-mode (Huang, 1997): need to pre-sepecify *k*
  - Zhang *et al.*(2006): computationally complex, low-demential dataset
- Few clustering algorithms for SNPs data
  - PCA (Paschou *et al.*, 2007)
  - Selinski and Ickstadt (2008)
  - HD-Cluster (Wang *et al.*, 2015)

Cluster Analysis for Genetic Sequencing Data

- An exploratory tool to integrate the information contained in SNPs
- Define analytic regions or to detect possible the relationship among genes if no prior knowledge
- Current clustering algorithms, e.g. k-means, hierarchical clustering methods
  - For quantitative measurements
  - Similarity metric: Euclidean distance and Mahalanobis distance
- Few clustering algorithms for discrete data
  - *k*-mode (Huang, 1997): need to pre-sepecify *k*
  - Zhang *et al.*(2006): computationally complex, low-demential dataset
- Few clustering algorithms for SNPs data
  - PCA (Paschou *et al.*, 2007)
  - Selinski and Ickstadt (2008)
  - HD-Cluster (Wang *et al.*, 2015)

Bayesian nonparametric clustering

- Handle the problems of clustering or partitioning through using Dirichlet process.
- The finite-dimensional prior distributions can be replaced with stochastic process.
- Do not need to determine the number of clusters in advance.

Notation

Data structure

Subject	Disease	Genotype
1	Y_1	$\mathbf{X}_1$
2	Y_2	$\mathbf{X}_2$
$\vdots$	$\vdots$	$\vdots$
n	Y_n	$\mathbf{X}_n$

$\Rightarrow$

SNP_1	$\cdots$	SNP_m
X_{11}	$\cdots$	X_{1m}
X_{21}	$\cdots$	X_{2m}
	$\ddots$	
X_{n1}	$\cdots$	x_{nm}
$\mathcal{S}_1$	$\cdots$	$\mathcal{S}_m$

- Assume the data set contains m SNP's and n subjects.
- $Y_i \in \{0, 1\}$ be the disease status of the $i-$th subject.
- $X_{ip} \in \{0, 1, 2\}$ be the genotype coding of the $p-$th SNP for $i-$th subject.

Outline

Multinomial Model for SNP Frequencies

Let

$$\mathcal{S}_p = (S_{p0}, S_{p1}, S_{p2})$$

- $S_{pj} = \sum_{i=1}^{n} I(X_{ip} = j)$ is the total number of subjects whose genotype on SNP p is coded j, where $j \in \{0, 1, 2\}$.
- Given $\underline{\theta}_p = (\theta_{p0}, \theta_{p1}, \theta_{p2}) \in \Theta$,

$$\mathcal{S}_p \sim Mult(n, \underline{\theta}_p)$$

- To allow clustering of SNPs within chromosome regions, we suppose that there are ties among $\underline{\theta}_p$, i.e. we will group SNPs that show similar frequencies of allele coding.

Multinomial Model for SNP Frequencies

Let

$$\mathcal{S}_p = (S_{p0}, S_{p1}, S_{p2})$$

- $S_{pj} = \sum_{i=1}^{n} I(X_{ip} = j)$ is the total number of subjects whose genotype on SNP p is coded j, where $j \in \{0, 1, 2\}$.
- Given $\underline{\theta}_p = (\theta_{p0}, \theta_{p1}, \theta_{p2}) \in \Theta$,

$$\mathcal{S}_p \sim Mult(n, \underline{\theta}_p)$$

- To allow clustering of SNPs within chromosome regions, we suppose that there are ties among $\underline{\theta}_p$, i.e. we will group SNPs that show similar frequencies of allele coding.

A Bayesian Model for Cluster Analysis I

Conditional likelihood

Dirichlet process mixture (DPM) model

$$\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m | C_1, \ldots, C_K, \underline{\psi}_1, \ldots, \underline{\psi}_K \sim \prod_{k=1}^{K} \prod_{p \in C_k} Mult(n, \underline{\psi}_k) \quad (1)$$

- $\rho = \{C_1, \ldots, C_K\}$ is a partition of the the data index set $\{1, 2, \ldots, m\}$.

A Bayesian Model for Cluster Analysis II

Prior specification

$$\begin{aligned} &\underline{\psi}_1, \underline{\psi}_2, \ldots, \underline{\psi}_k | K \sim P_0(\cdot) \\ &\rho \sim \textit{eppf}(n_1, \ldots, n_k, \alpha) \end{aligned} \tag{2}$$

- $P_0(\cdot)$ is a Dirichlet distribution on Θ.
- Exchangeable partition probability function (eppf)

$$\pi(\rho) = Pr(n_1, n_2, \ldots, n_K) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+n)} \alpha^{K-1} \prod_k K(n_k - 1), \tag{3}$$

where $n_k = \sharp C_k$ and $\alpha \in \Re^+$.

Dirichlet process

- Stochastic process
- Generate a discrete distribution point-by-point, sampling either from a given probability measure or from a previously generated point.
- Given a Dirichlet process $DP(P_0, \alpha)$,
  - P_0 is the base distribution or base probability measure.
  - α is a positive real number

  draw from DP will return a random distribution over some of the values that can be drawn from P_0.
- Simple illustration
  - Beta-binomial distribution
  - Dirichlet-multinomial distribution

Outline

Association study I

- Let $\rho = (C_1, \ldots, C_K)$ a partition of SNPs and $\psi = (\underline{\psi}_1, \ldots, \underline{\psi}_K)$
- For each SNP-set cluster C_k, let $G_{i,k}$ be a genetic score for ith subject

$$G_{i,k} = \ln \frac{Pr(\mathbf{X}_{i,C_k} | \underline{\psi}_k)}{\sharp C_k} \tag{4}$$

where $Pr(\mathbf{X}_{i,C_k} | \underline{\psi}_k) = \prod_{X_{ip} \in C_k} \psi_{k0}^{I(X_{ip}=0)} \psi_{k1}^{I(X_{ip}=1)} \psi_{k2}^{I(X_{ip}=2)}$, $k = 1, \ldots, K$ and $p = 1, \ldots, m$.

- Interpretation: $G_{i,k}$ is the probability of observing the SNPs configuration of patient i within cluster C_k, given the parameters ρ's and ψ's

Association study II

Bayesian logistic regression

$$\text{logit}(Pr(Y_i = 1|\rho, \mathbf{X}_i, \psi)) = \beta_0 + \sum_{k=1}^{K} \beta_k G_{i,k} + \gamma \boldsymbol{E}, \tag{5}$$

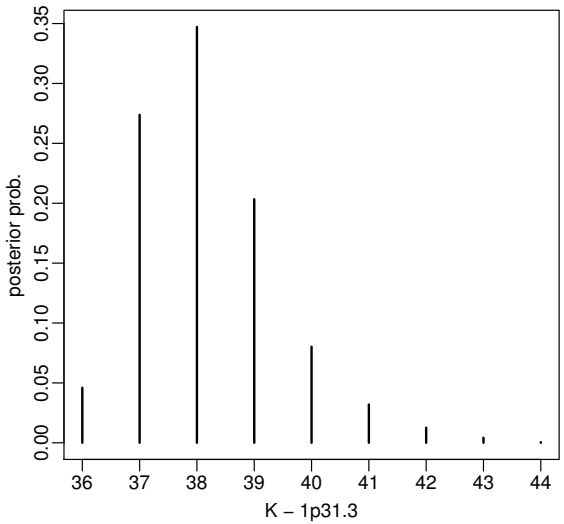
where **E** are environmental covariates.

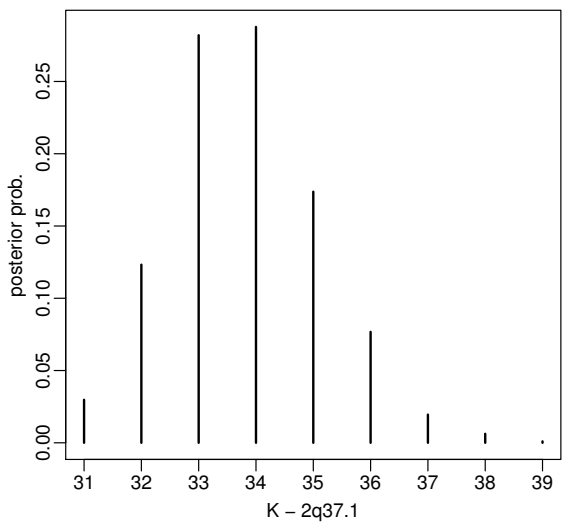
- Prior: $\beta_k \sim N(0, 100)$
- A SNP cluster C_k is declared significative if the posterior 90% credible interval of β_k do not contains zero.

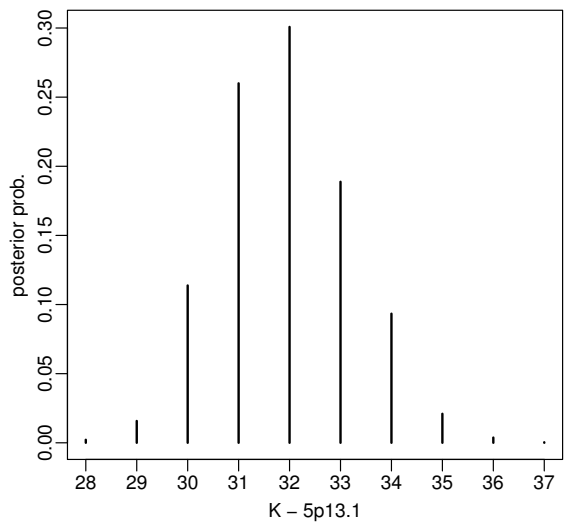
Real Data Application

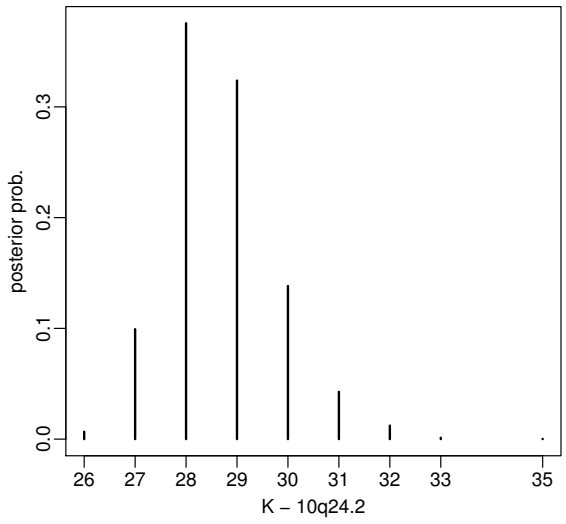
- The genotyping data of Crohn's disease (CD) study from the Wellcome Trust Case Control Consortium study (WTCCC)
- Five chromosome regions were selected for analyzing.
  - 1p31.3
  - 2q37.1
  - 5p13.1
  - 10q24.2
  - 16q12.1
- 1748 patients with Crohn's disease and 2938 shared controls were considered. **Total number of patients** $n = 4686$.
- After excluding SNPs with minor allele frequency (MAF) lower than 0.01 or in Hardy-Weinberg disequilibrium, $m = 3704$ **SNPs** were left for our analysis.

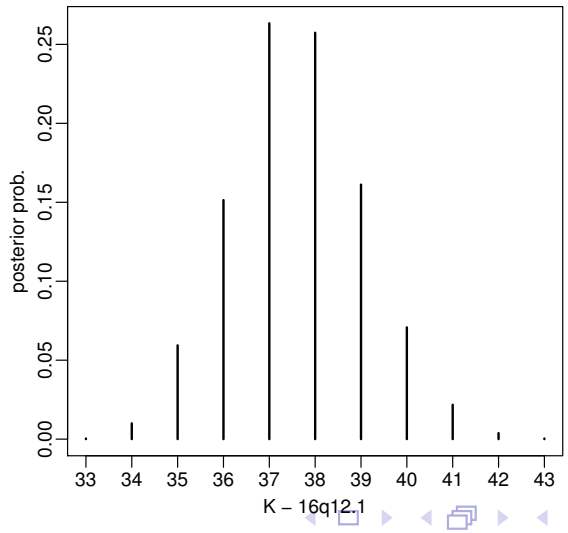
Posterior Histogram of K for Each Chromosome



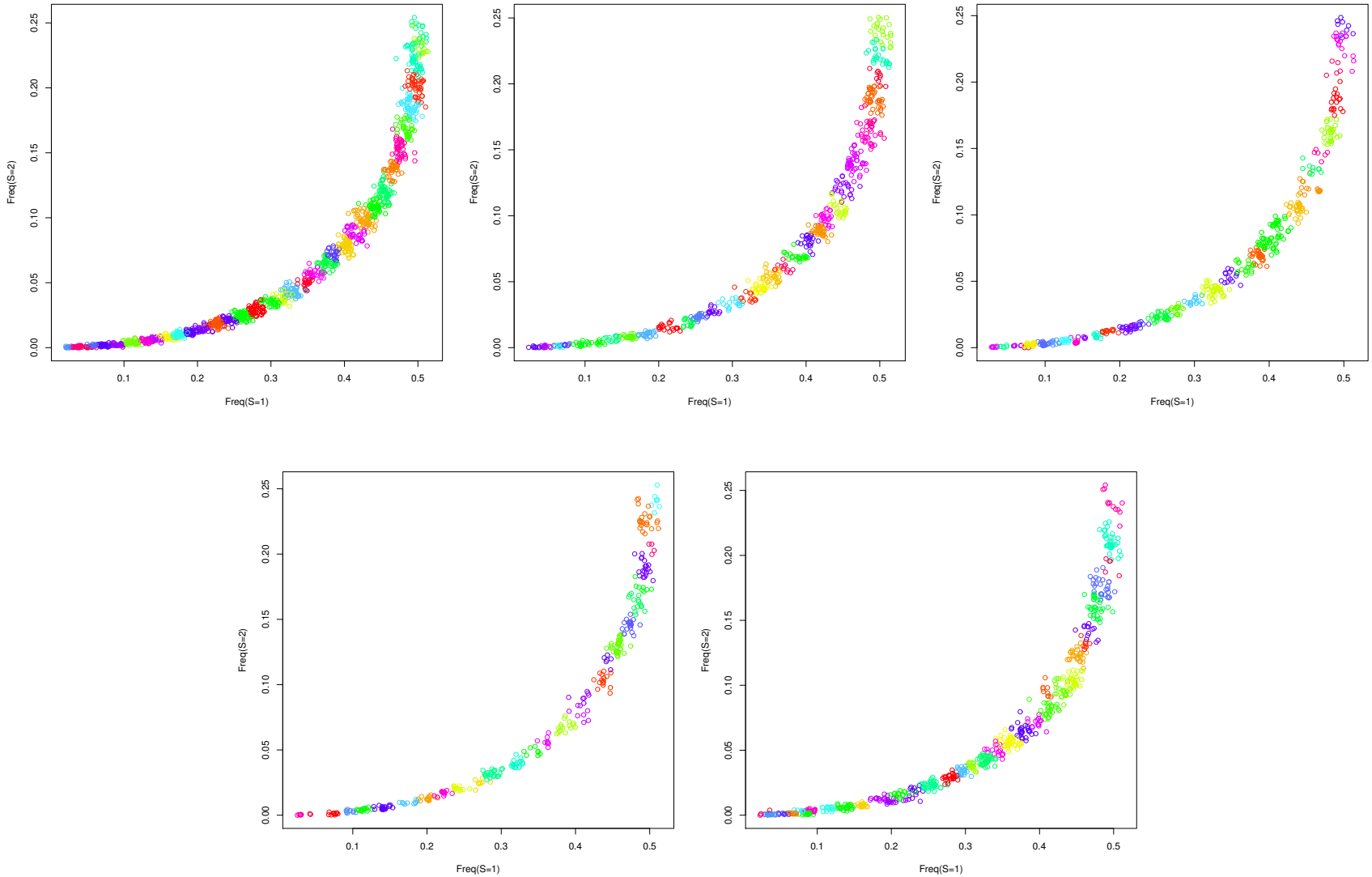




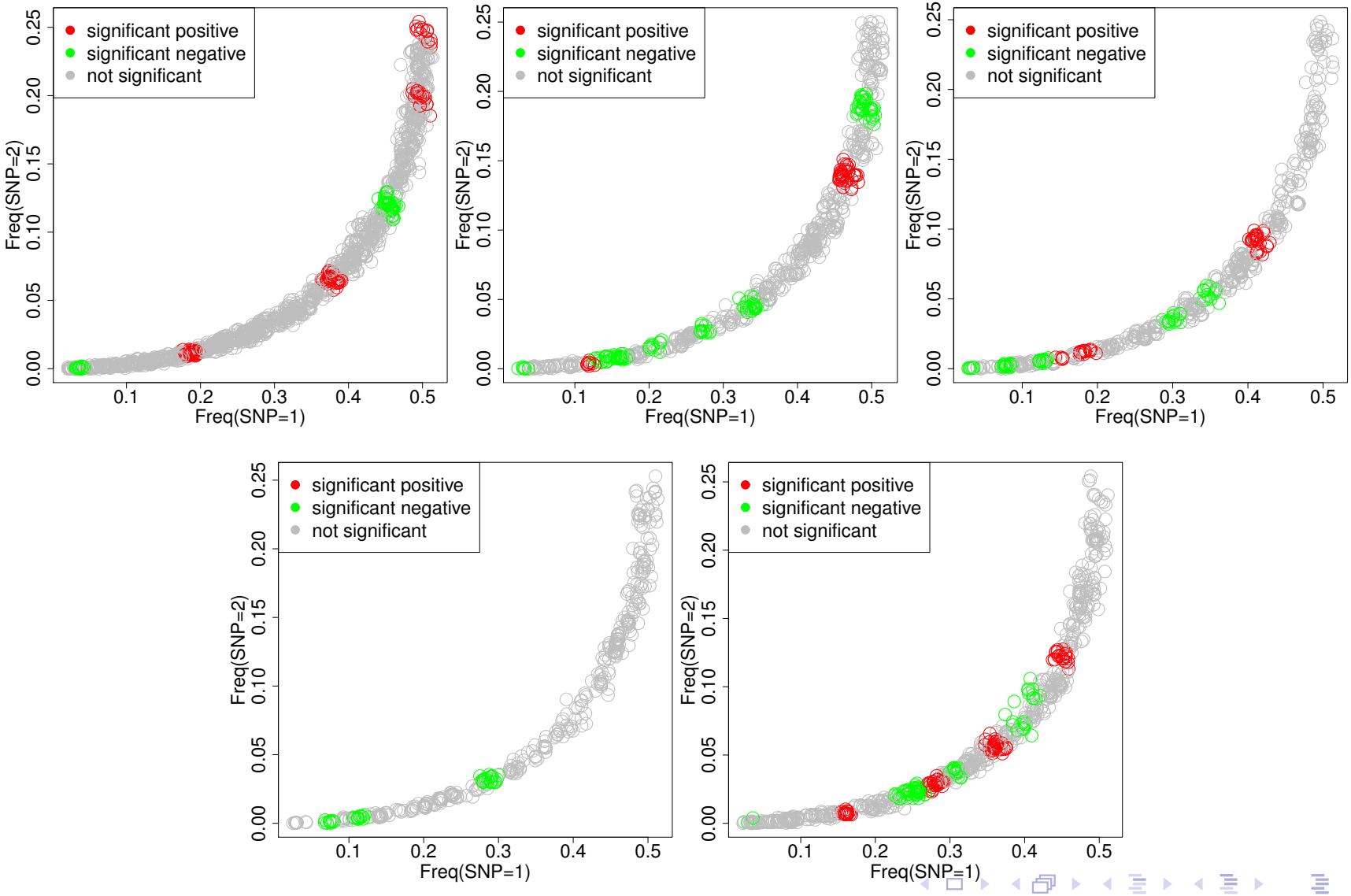




SNP Clusters



SNP Clusters



SNP Clusters

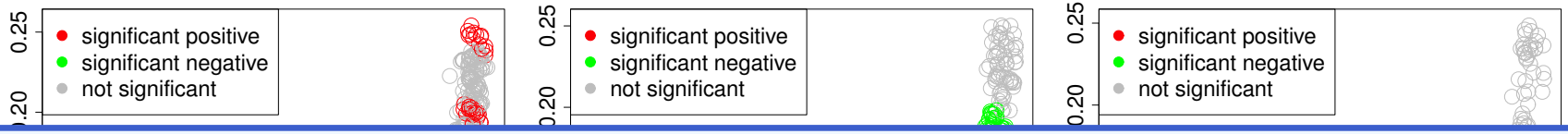
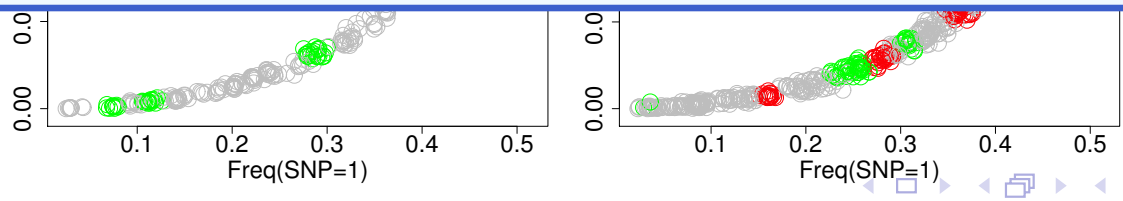
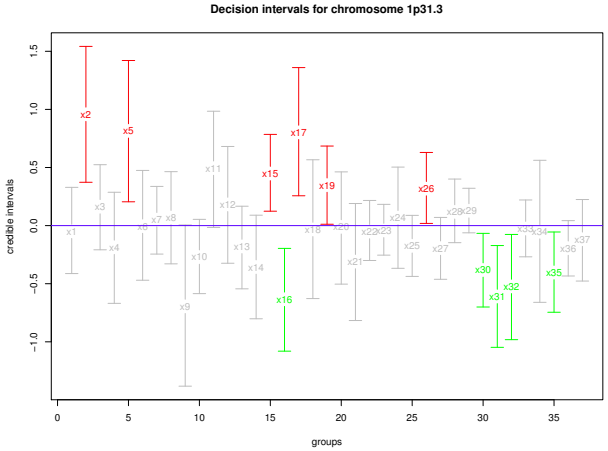


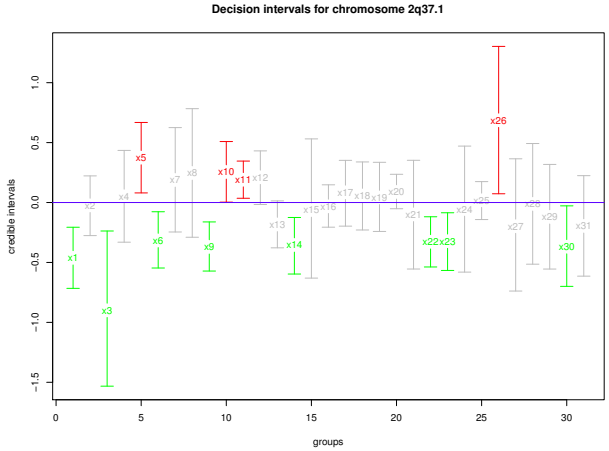
Table 1: Descriptive statistics for SNPs and clusters per chromosome region.

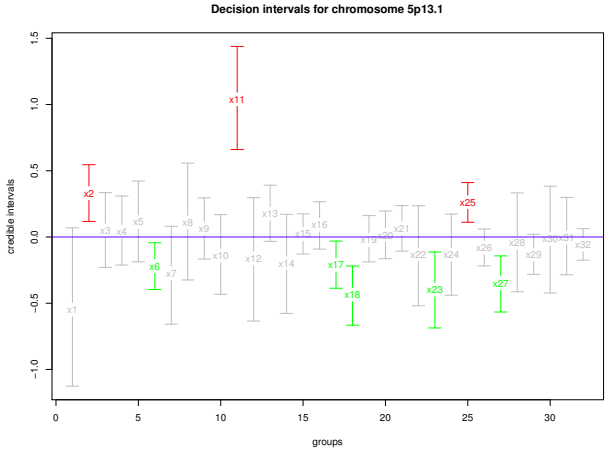
Region	No. of SNPs	$\hat{K}$	Cluster size			Association	
			min	median	max	positive	negative
1p31.3	1357	37	12	19	36	6	5
2q37.1	662	31	7	22	45	4	8
5p13.1	554	32	2	15	44	3	5
10q24.2	390	28	2	12.5	36	0	4
16q12.1	742	37	1	17	45	4	7

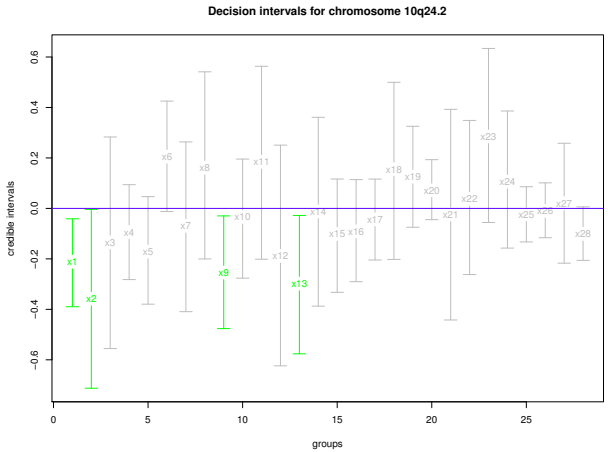


90% Credible Intervals of β_k









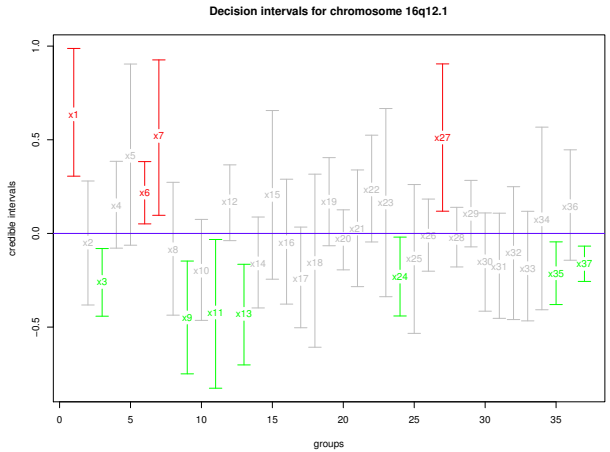
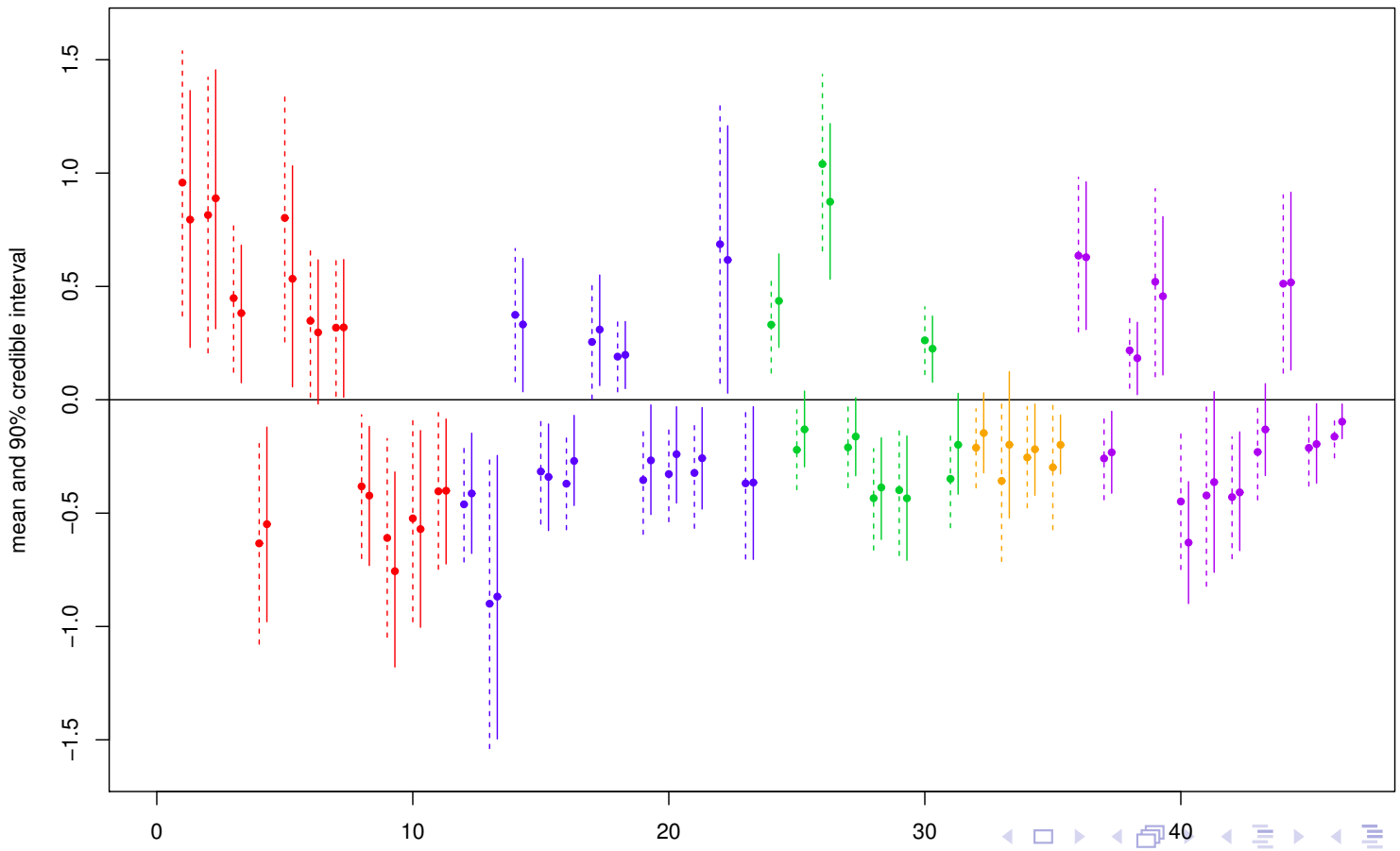


Table 2: Estimates of association for SNP clusters in 1p31.3.

Cluster ID	$\hat{\beta}$	90% CI for β	$\widehat{\text{OR}}$	No. of SNPs	Gene Symbol
1p31.3					
2	0.958	(0.370, 1.539)	2.606	42	*CACHD1, GADD45A, GNG12, IL12RB2, IL23R, INADL, KANK4, LEPR, NFIA, PDE4B, ROR1, SGIP1, TM2D1, USP1*
5	0.815	(0.207, 1.423)	2.259	54	*ALG6, C1orf141, EFCAB7, GADD45A, IL23R, INADL, INSL5, ITGB3BP, KANK4, LEPR, NFIA, PDE4B, PGM1, ROR1, RPE65, SGIP1, USP1, WDR78*
15	0.448	(0.122, 0.783)	1.566	48	*CACHD1, IL23R, INADL, MIER1, NFIA, RAVER2, ROR1, SGIP1, SLC35D1, TM2D1, WDR78*
16	-0.633	(-1.078, -0.193)	0.531	53	*ATG4C, DOCK7, IL12RB2, IL23R, ITGB3BP, KANK4, LEPR, NFIA, PDE4B, PGM1, RAVER2, ROR1, SGIP1, SLC35D1, WDR78*
17	0.802	(0.255, 1.358)	2.231	18	*ALG6, INADL, ITGB3BP, PDE4B, ROR1, SGIP1, UBE2U*
26	0.318	(0.016, 0.626)	1.374	39	*CACHD1, IL23R, NFIA, RAVER2, ROR1, SGIP1, UBE2U, USP1, WLS*
30	-0.382	(-0.700, -0.066)	0.682	28	*CACHD1, GADD45A, GNG12, L1TD1, NFIA, PDE4B, ROR1, UBE2U, USP1, WLS*
31	-0.609	(-1.047, -0.171)	0.544	39	*GNG12, INADL, LEPR, NFIA, PDE4B, ROR1, SGIP1, USP1*
32	-0.523	(-0.979, -0.074)	0.593	31	*C1orf141, INADL, NFIA, PDE4B, PGM1, RAVER2, ROR1, SGIP1, TM2D1, UBE2U, USP1*
35	-0.404	(-0.747, -0.057)	0.668	28	*INADL, KANK4, NFIA, PDE4B, ROR1, RPE65, USP1*

Comparison of 90% Credible Intervals

Comparison of 90% credible Intervals for the clusters fitted separately and clusters fitted collectively.



Conclusion

- We have introduced a Bayesian nonparametric model for cluster SNPs.
- A Bayesian DPM model was used.
- A Bayesian logistic regression with genetic score for clustered SNPs was used to perform association studies
- We found 46 significant SNP clusters including 106 genes of which 16 genes have been reported.

Discussion

- Advantages
  - no pre-specification of the number of clusters
  - statistical inference of this number
  - incorporation of the uncertainty in SNP allocations to clusters in analysis of associations
  - posterior inference of the susceptibility for each cluster and SNP
- Coding system can be changed, if the dominance or recessive inheritance model is assumed or if a certain allele is of major interest.
- This approach can accommodate continuous variables as response vectors.
- Performance for analyzing rare variants is not clear.