Regression Testing for Region-based Genetic Association under Genomic Partitioning adapted to Linkage Disequilibrium

Shelley B. Bull, Lunenfeld-Tanenbaum Research Institute
University of Toronto bull@lunenfeld.ca

Banff International Research Station
Genomics and Metagenomics in Human Health:
Recent Developments in Statistical and Computational Methods
2 February 2019

Lunenfeld-Tanenbaum Research Institute
Sinai Health System

UNIVERSITY OF TORONTO
DALLA LANA SCHOOL OF PUBLIC HEALTH

Two-Stage Approach to GW Association

Genomic Partitioning:

- For comprehensive genomic region definition
- Based on linkage disequilibrium (LD) structure
  at imputation density (6 -12 million variants)
- Designed to produce quasi-independent LD blocks
- Feasible computation for the entire autosome

Global Test Statistics for Regional Association:

- Multiple regression (linear or logistic) with covariates
- Dimension reduction adapted to LD within regions
- Testing is non-adaptive to trait data
- Asymptotic p-values

Aim

Bridge GWAS discovery *with* Region characterization

Region-based analysis

- Common and low frequency variants
- More powerful than single-variant analysis under plausible genetic architectures
- Robust to population differences & genetic heterogeneity
- Integrate intergenic variants with promoter, regulatory and/or coding functions
- Reduce multiple testing burden

Big question:

How to specify appropriate regional variant sets ?

Applications

Population-based GWAS cohorts:

- state-of-the-art genotyping platforms
- dense set of variants (imputed to 1000 Genomes)

1) **DCCT**: Baseline lipid levels in therapeutic RCT in type 1 diabetes * *quantitative traits*

   - 1340 participants, Illumina Human Core Exome Array
   - chromosomes 1-22: 6.61M variants (MAF > 5%)

2) **MSHPH**: Toronto site of the international lung cancer case-control consortium (ILCCO) * *categorical traits*

   - 1359 cases, 949 controls genotyped by OncoArray
   - chromosome 8: 335K variants (MAF > 5%)

Methods – Region-based Association

Analytic Objectives:

Sensitive to complex gene architecture
Feasible for genome-wide analysis
Incorporates local variant correlation (LD) structure, but
NOT sample-based knowledge of trait association

Yoo YJ et al, 2017. Multiple linear combination (MLC) regression tests for common variants adapted to linkage disequilibrium structure. *Genet Epidemiol*;41(2):108–21.

MLC is a constrained regression test statistic:

- adapts to complex LD structure to construct clusters of closely correlated variants, coded such that the majority of pairwise correlations are positive.
- asymptotically valid – nominal type I error in linear regression simulations under various architectures

Regression Testing – Dimension Reduction

Multi-variant joint regression model of *K* variants:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_i + \cdots + \beta_K X_K + \varepsilon$$

Under global H_0:

$$G_W = \hat{\beta}^T \Sigma^{-1} \hat{\beta}$$

is generalized Wald statistic (*K df*)

MLC constrained test statistic oriented to a restricted alternative:

$$G_M = n\,(\mathrm{C}^T\,\hat{\beta})^T\,[\mathrm{C}^T \textstyle\sum \mathrm{C}]^{-1}\,(\mathrm{C}^T\,\hat{\beta})$$

where

$$\mathrm{C} = (\Sigma^{-1} J)\,(J^T \Sigma^{-1} J)^{-1}$$

J = *K* by *L* matrix assigning variants to clusters

Under H_0 : G_M ~ asymptotically central chi-squared with *L* < *K df*

How to choose *J*:

Cluster variants into bins using within-region LD

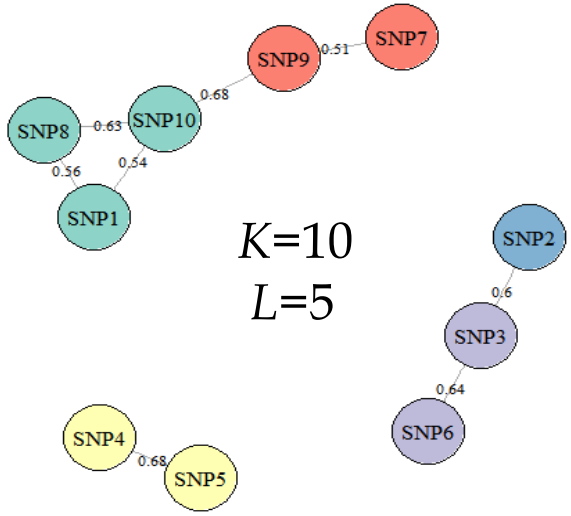
High, positive correlation of variants within clusters, low correlation between clusters

Regression Testing – Dimension Reduction

Clique-based clustering algorithm

- models variants as a graph
- clustering by LD measure of additively coded variant pairs
- size & # of clusters depends on choice of the correlation threshold
- maximizes positive correlation within a cluster

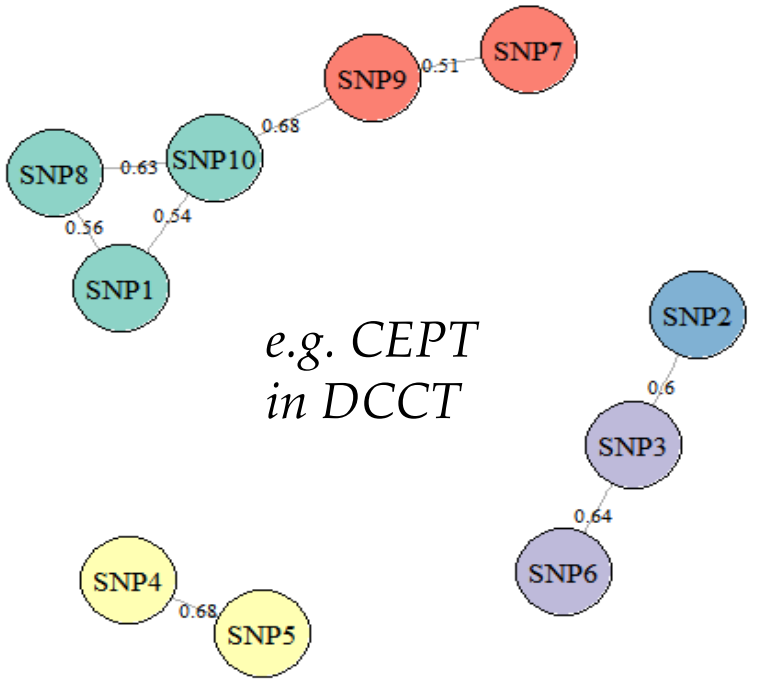
e.g. CEPT in DCCT



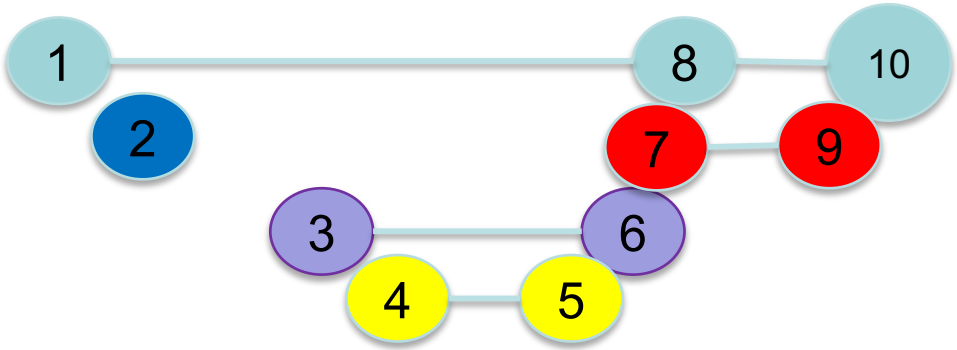
Clustering of SNPs by applying CLQ algorithm to linkage disequilibrium (r) pattern. Edges with |r| < 0.5 are removed.

Yoo *et al*, Clique-based clustering of correlated SNPs can improve performance of gene-based multi-bin linear combination test, *Biomedical Research International* 2015; 852341

Yoo YJ et al, 2017. Multiple linear combination (MLC) regression tests for common variants adapted to linkage disequilibrium structure. *Genet Epidemiol*;41(2):108–21.



e.g. CEPT in DCCT



- 10 SNPs clustered into 5 bins
- SNPs within a bin are not necessarily physically contiguous
- Bins can overlap according to bp position

- MLC takes a weighted linear combination of regression coefficients within each bin
- Bin-specific statistics are summed as squares and cross products

Methods – Genomic Partitioning

Interval graph modeling to cluster correlated variants

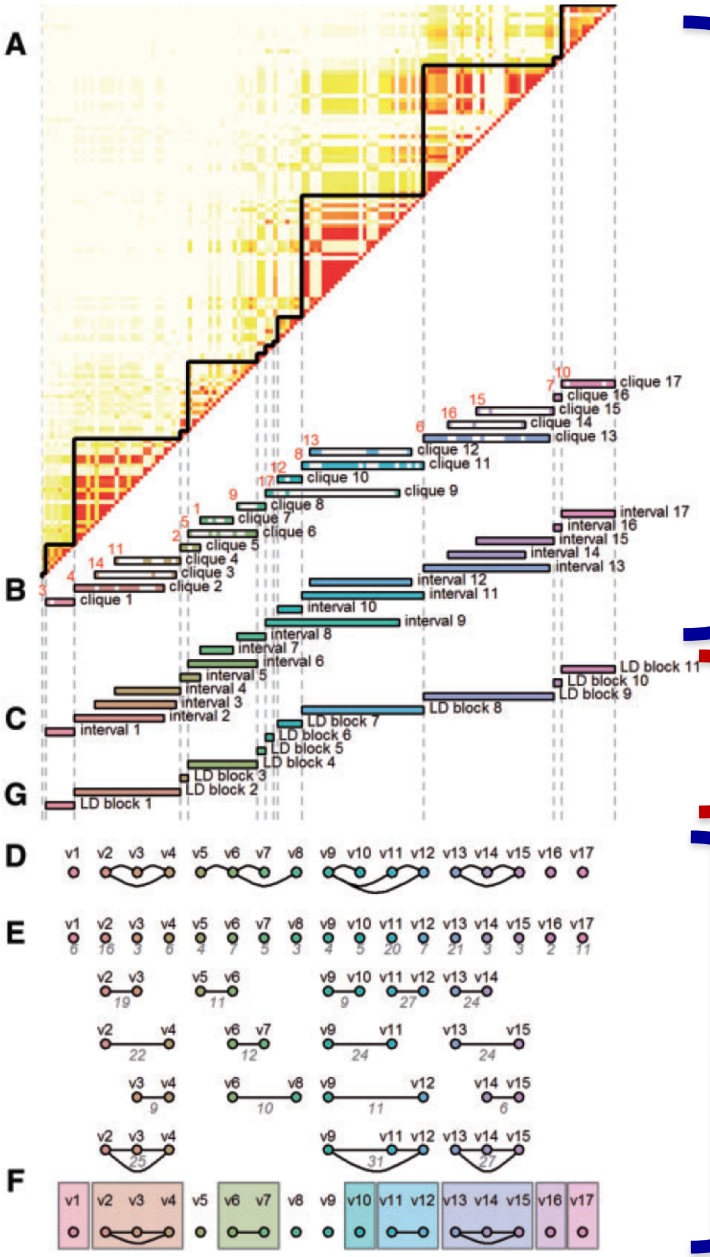
- *GPART* using "*Big-LD*" algorithm
- Agnostic to gene boundaries
- Produces a large number of non-overlapping & approximately independent LD-blocks

Kim S-A et al, 2018. A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics* 34(3):388-397 Software http://github.com/sunnyeesl/BigLD

Compared to existing methods, "Big-LD" approach

- Larger, more invariant LD blocks
- Better LD optimization within & across LD blocks
- Boundaries agree with known recombination hotspots

Genomic Partitioning – BigLD algorithm



1. SNP clustering based on correlation

A. Pairwise SNP correlation in a region

B. Identification of clusters of SNPs (cliques) with pairwise correlations > clustering parameter (CLQ)

C. Clusters are converted to genomic intervals defined by chromosome positions of the two most extreme SNPs

Quasi-independent blocks of consecutive SNPs obtained after 1& 2

2. Interval graph model of SNP clusters

D. Interval graph model, edges connect pairs of overlapping intervals (nodes)

E. Intervals merged successively to form consecutive non-overlapping intervals (F)

Applications

Population-based GWAS cohorts:

- state-of-the-art genotyping platforms
- dense set of variants (imputed to 1000 Genomes)

1) **DCCT**: Baseline lipid levels in therapeutic RCT in type 1 diabetes * *quantitative traits*

   - 1340 participants, Illumina Human Core Exome Array
   - chromosomes 1-22: 6.61M variants (MAF > 5%)

2) **MSHPH**: Toronto site of the international lung cancer case-control consortium (ILCCO) * *categorical traits*

   - 1359 cases, 949 controls genotyped by OncoArray
   - chromosome 8: 335K variants (MAF > 5%)

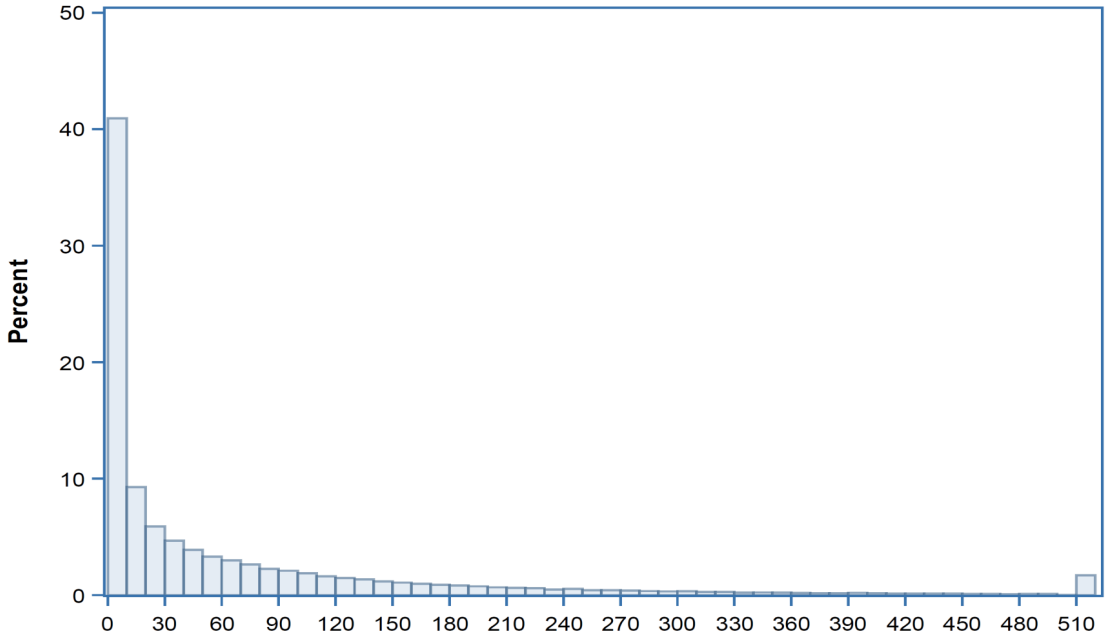
DCCT Results – Genomic Partitioning

In total: 6.61M variants (MAF > 5%) on 22 autosomes

6,551,457 variants in 91,052 LD blocks + 57,504 singletons

Mean: 69.49 variants per block

- Total # of blocks: 91,052
- Average # of SNPs per block: 70



# of SNPs per block

Region-based Test Statistics

Global generalized Wald statistic (*K df*)

- each regression coefficient enters the test statistic in squared and cross-product terms

MLC statistic (*L* < *K df*)

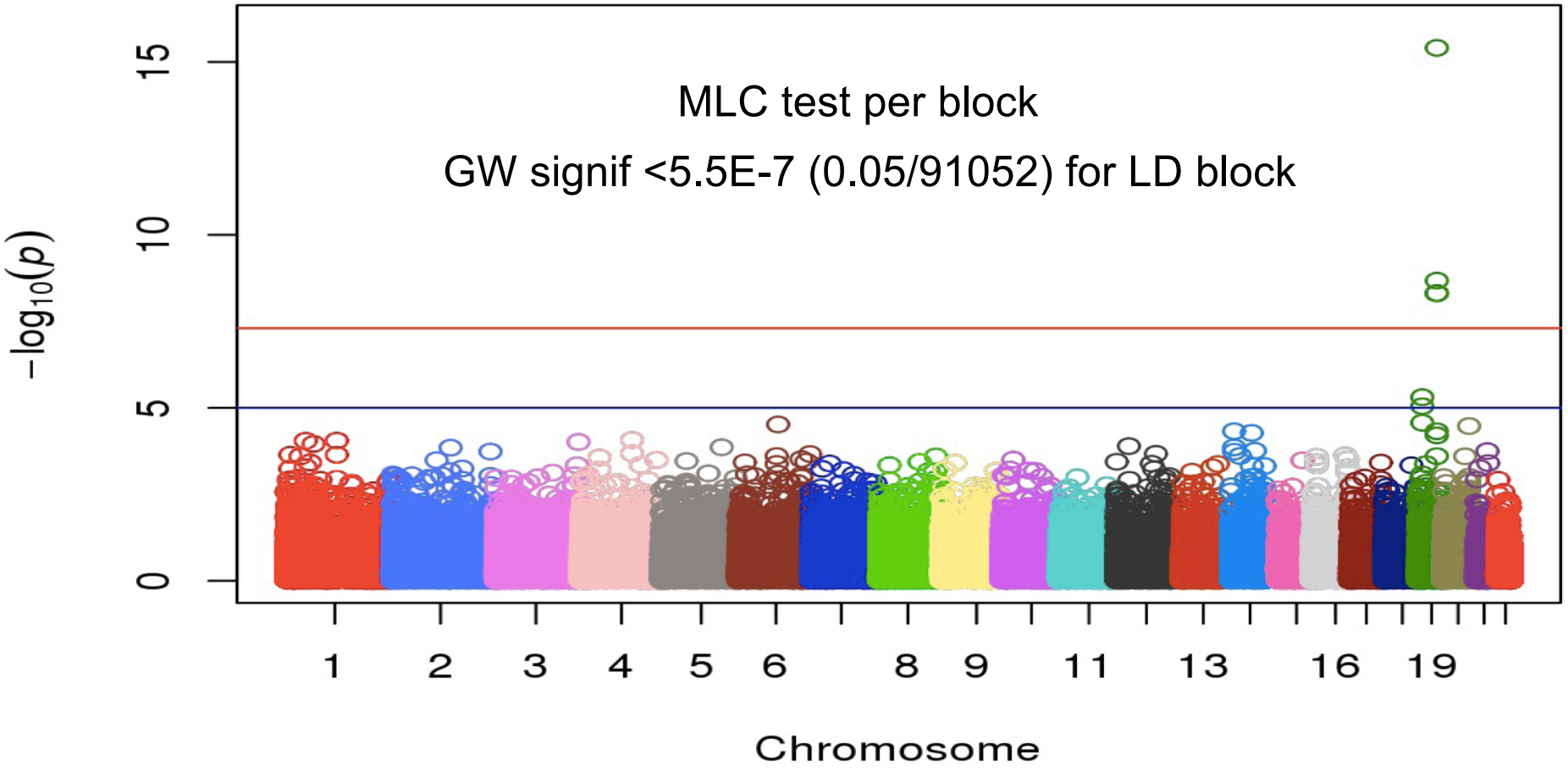
- reduced *df* equal to the number of clusters
- *within* cluster linear combination of regression coefficients
- cluster-specific terms are aggregated in a sum of squared and cross-product terms

PC80 (< *K df*):

- global test based on regression of minimum number of principal components capturing 80% of variance in regional variant set [Gauderman et al, 2007]
- reduces dimension prior to regression model fitting

DCCT Results – Region-based Association

Linear regression of quantitative baseline LDL-cholesterol (with age, sex, age by sex) in each LD block/singleton



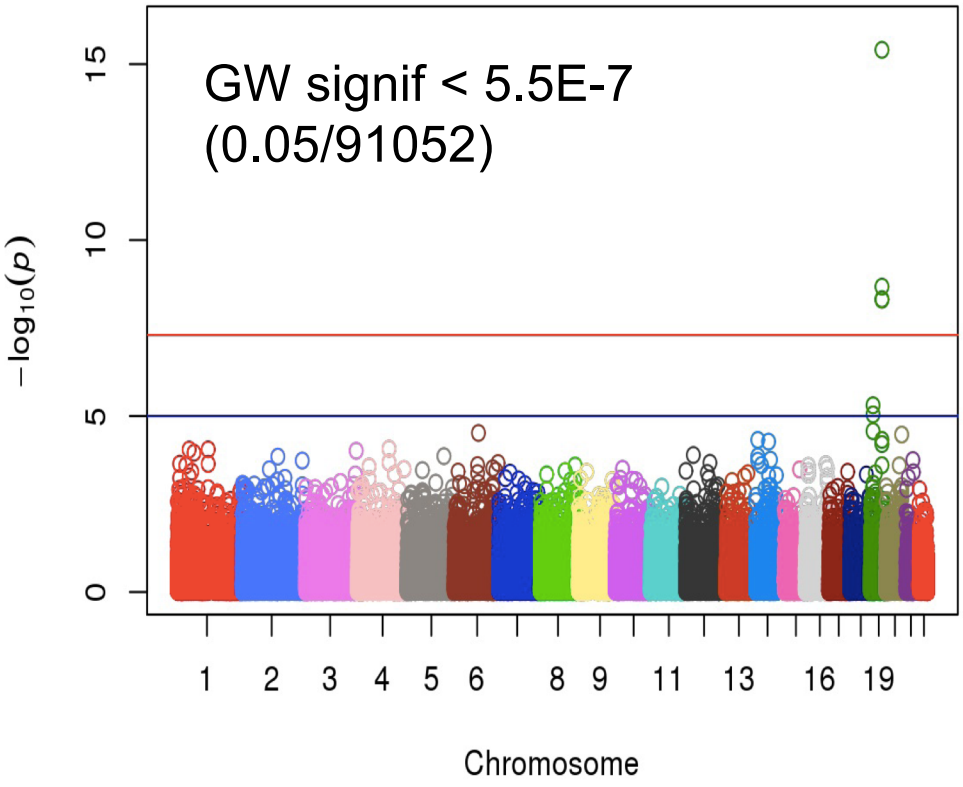
Top regions, with recapitulation of established associations (*LDLR, APOE*)

	CHR	LD block region	Generalized Wald df	Generalized Wald P value	MLC df	MLC P value	PC80 df	PC80 P value
	1	1703	59	6.24E-03	15	8.96E-05	5	7.08E-04
	1	3706	1	8.89E-05	1	8.89E-05	1	8.89E-05
	3	5996	14	6.01E-05	5	9.71E-05	3	3.80E-01
	4	3486	52	4.43E-02	12	8.50E-05	6	1.75E-04
	6	2823	6	1.13E-04	4	3.02E-05	3	NA
	14	434	3	1.69E-04	2	4.76E-05	2	4.62E-05
	14	1629	2	5.33E-05	2	5.33E-05	2	5.33E-05
	19	636	4	1.10E-04	2	8.99E-06	1	1.50E-03
LDLR	19	637	7	6.12E-07	3	4.94E-06	2	2.42E-06
	19	638	6	1.91E-04	3	2.69E-05	2	4.83E-06
	19	1742	50	1.73E-03	11	4.76E-05	4	1.52E-02
	19	1743	2	6.26E-05	2	6.26E-05	2	6.26E-05
APOE	19	1749	41	4.35E-12	14	3.94E-16	6	1.84E-10
	19	1750	6	4.47E-11	2	2.11E-09	2	9.69E-09
	19	1751	2	4.96E-09	2	4.96E-09	2	4.96E-09
	19	1752	2	4.78E-09	2	4.78E-09	2	4.78E-09
	20	2371	2	3.94E-05	1	3.36E-05	1	3.87E-05

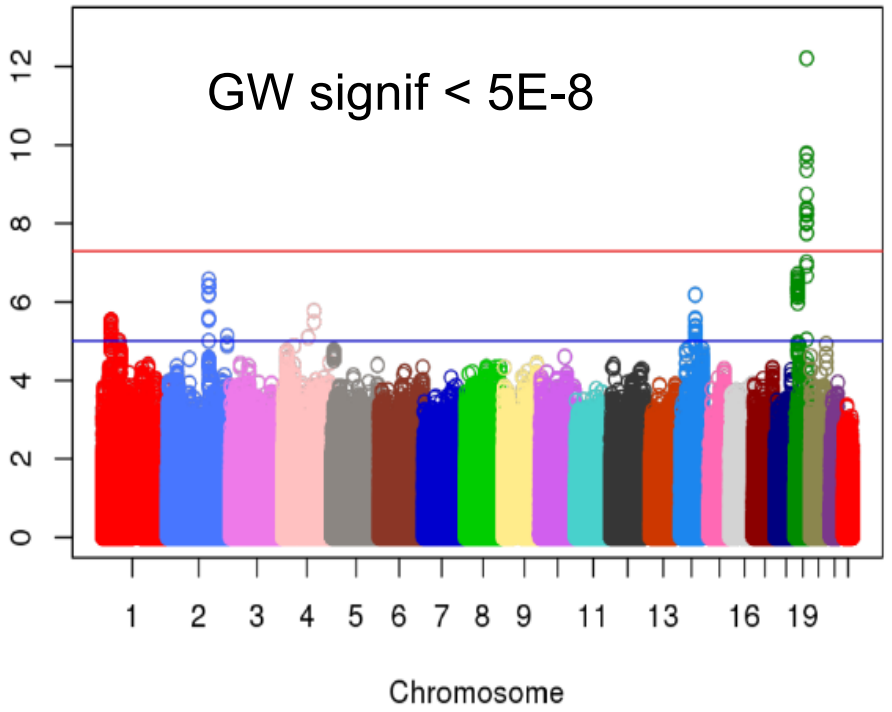
DCCT Results – Region-based vs Single SNP

Linear regression of quantitative baseline LDL-cholesterol (with age, sex, age by sex) in each LD block/singleton

MLC



Single SNP



DCCT Results – Two-Stage Approach

- *Big-LD*
  - Genome partitioning is feasible for genome-wide imputation-dense data.
  - Captures gene regions as well as inter-genic regions reasonably well since partitioning depends on genetic distance.
- *MLC*
  - Feasible for genome-wide studies of imputation-dense data.
  - Captures known GWAS loci
  - Significance threshold lowered due to reduction in multiple testing burden
  - P-value improved compared to GWAS

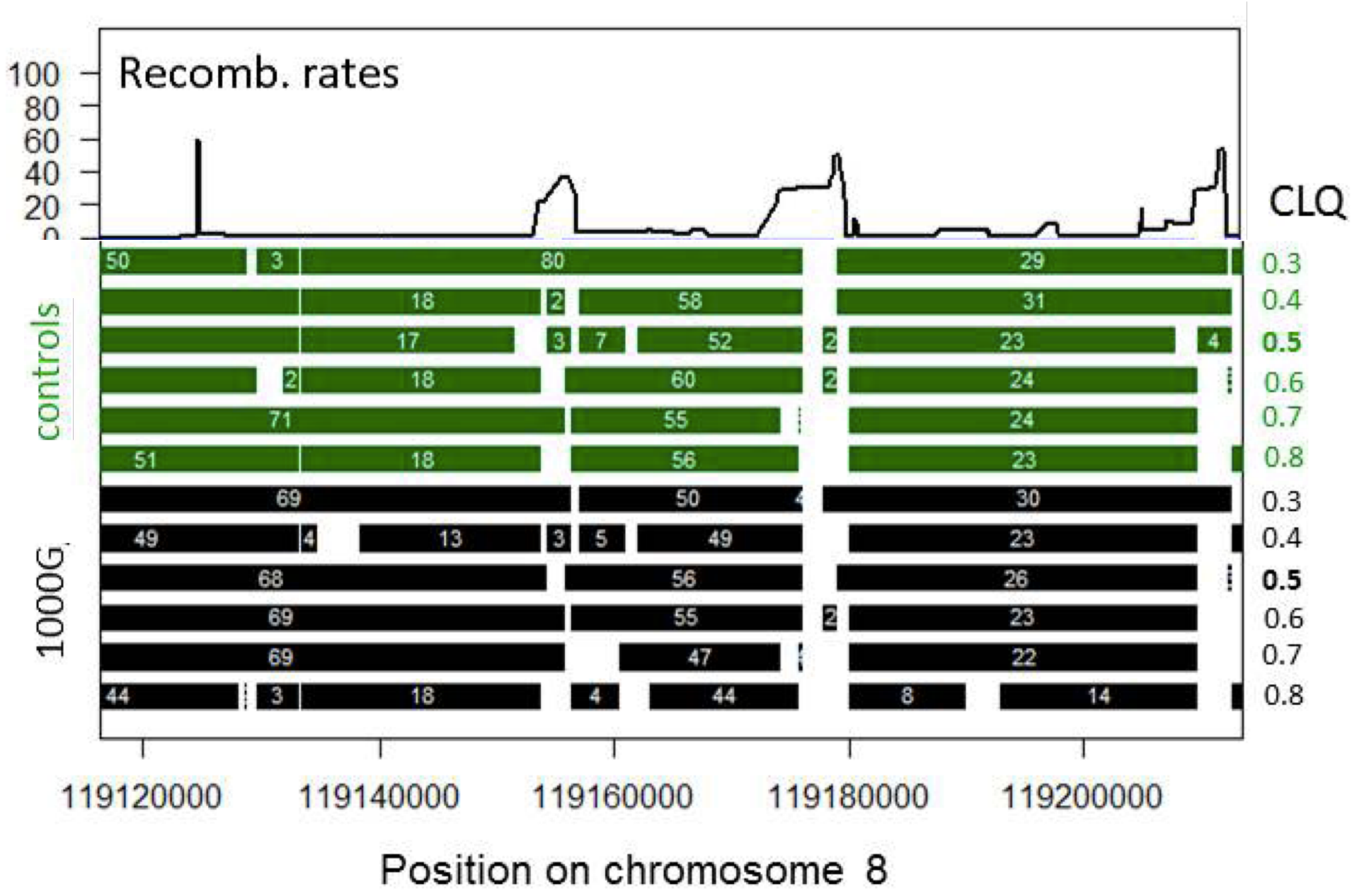
MSHPH Results – Genomic Partitioning

- Genome-wide SNP data :
  ~500K genotyped SNPs (Illumina OncoArray) &
  Imputed to 1000Genomes

  334,628 SNPs from chromosome 8
  (info≥ 0.4 & MAF in controls ≥5%)

  5,266 LD blocks in controls (99.2% SNPs in blocks) with:
  - right-skewed distribution of the number of SNPs per block (median=15, range 2-1071)
  - high within- & low between-block correlation (mean=0.66 & 0.33)

Results in a random region on chromosome 8 with comparison to 1000 Genomes European samples (sensitivity to CLQ parameter):

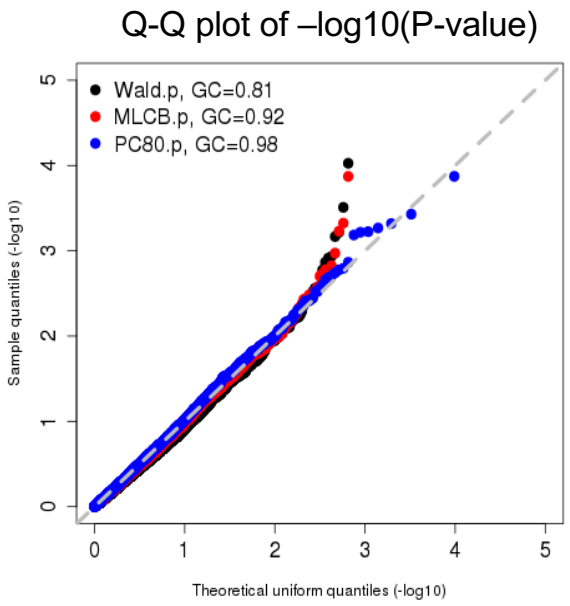


MSHPH Results – Region-based Association

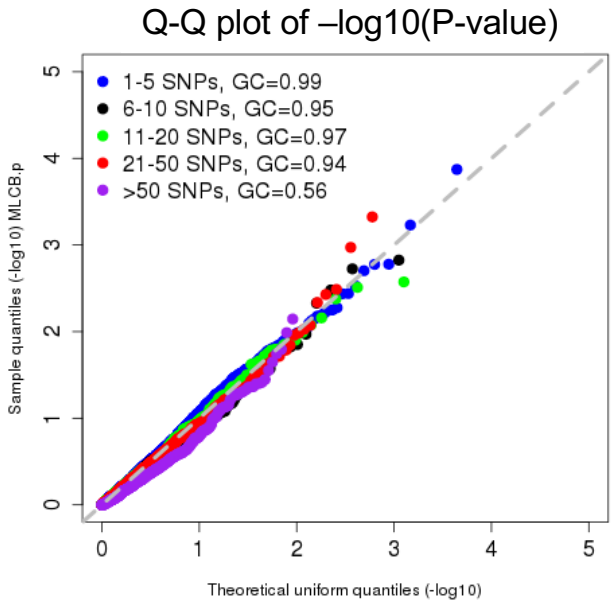
Logistic regression of case-control status (3 PCs, age & sex)
335K variants, 5K LD blocks (MAF > 5%) chromosome 8

# SNPs per block	1-5	6-10	11-20	21-50	>50
# blocks	2208	560	631	897	592

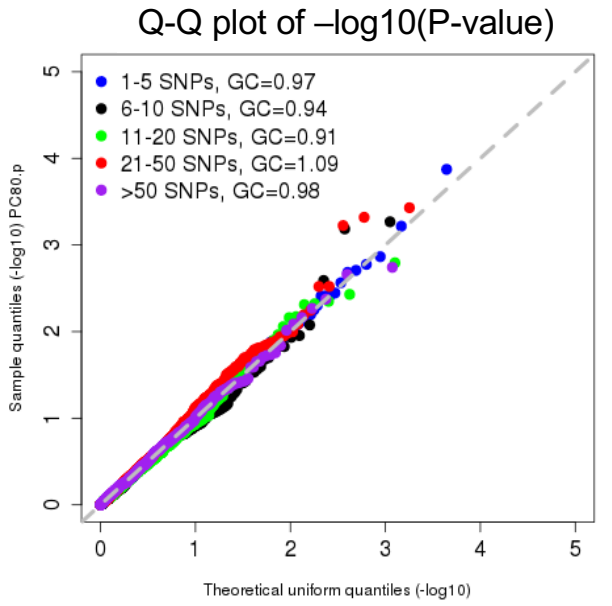
Wald, MLC, PC80



MLC by region size



PC80 by region size



Summary

Genomic region-based association discovery analysis

- Complementary to standard single-variant approach
- *Genomic partitioning* addresses variant set construction, and
- Facilitates comprehensive region-based testing
- Computationally feasible for imputation-dense data

Region-based test statistics

- Number of variants per LD-defined region is right skewed
- Very large regions produce conservative tests
- Strategies to deal with near linear dependencies
- *Dimension reduction* improves type 1 error control/power

Acknowledgements

Lunenfeld-Tanenbaum Research Institute

Myriam Brossard (PDF)
Yannick MacMillan (Summer student)

CIHR IRSC
Canadian Institutes of Health Research | Instituts de recherche en santé du Canada

Rayjean Hung & colleagues
International Lung Cancer Consortium (ILCCO)

NIH NATIONAL CANCER INSTITUTE

Seoul National University

SEOUL NATIONAL UNIVERSITY

Yun Joo Yoo & colleagues
Sun Ah Kim (PDF)

National Research Foundation of Korea

SickKids Research Institute

SickKids
THE HOSPITAL FOR SICK CHILDREN
Genetics & Genome Biology Program

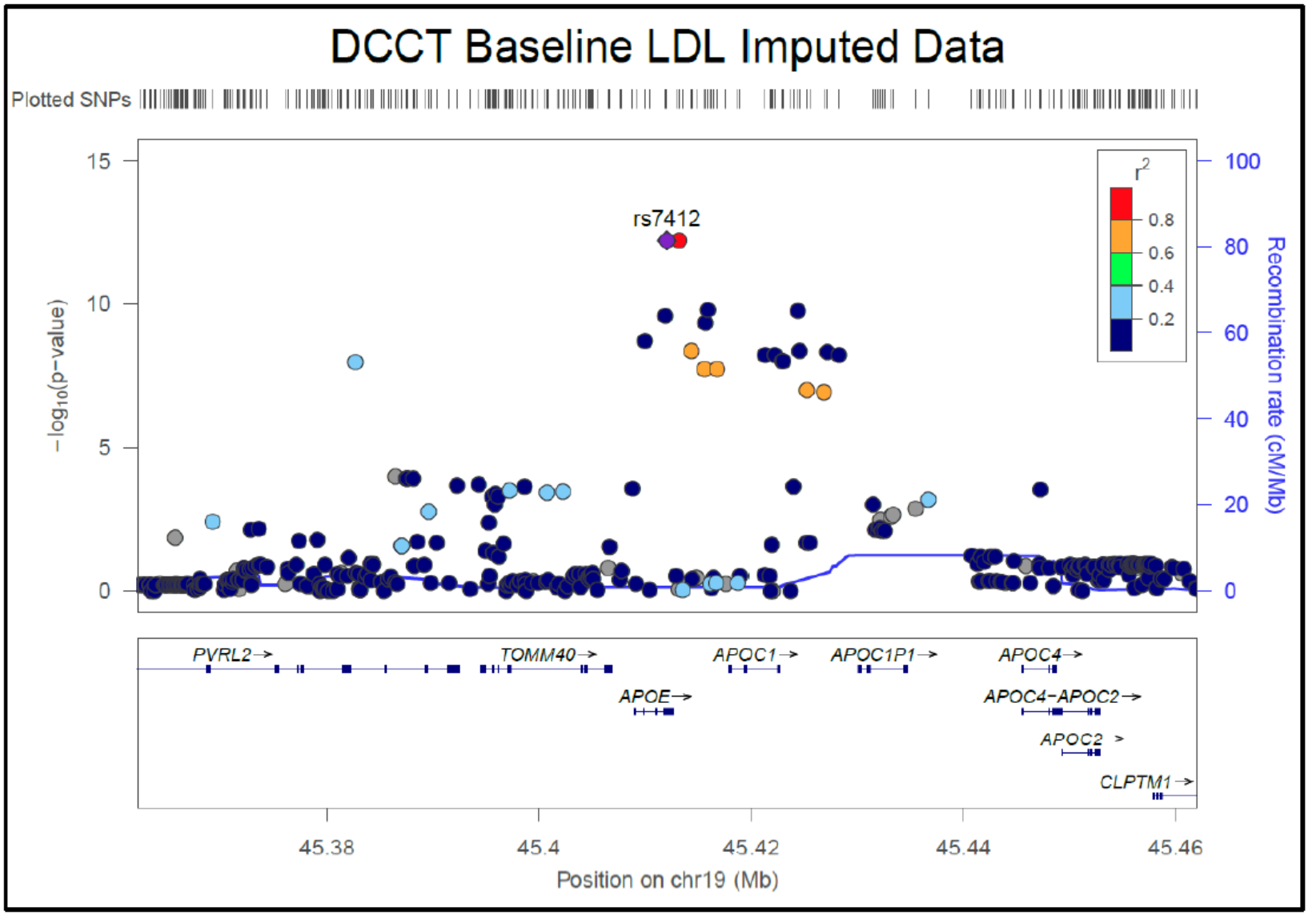
Andrew Paterson
Delnaz Roshandel (PDF)
DCCT/EDIC Genetics Study

JDRF IMPROVING LIVES. CURING TYPE 1 DIABETES.

CIHR STAGE
Strategic Training for Advanced Genetic Epidemiology

Response to questions

Region Plot rs7412 ± 50kb



Bioinformatics (2018) Supplemental (Simulations of genes from 1000Genomes)

Table S13. Empirical power of multi-SNP association tests corresponding to adjusted region-wide significance level by Bonferroni

Region	Method	N. of blocks	Wald	LC-B	LC-Z	MLC-B	MLC-Z	SKAT	SKAT-O
CRKL 21,017,148~213,153,84 (747 SNPs) σ=10	Big-LD	1	0.393	0.828	0.832	0.346	0.344	0.845	0.828
	S-MIG++	82	0.236	0.314	0.312	0.273	0.273	0.309	0.29
	MIG++	89	0.304	0.304	0.304	0.304	0.304	0.304	0.304
	Haploview(CI)	148	0.266	0.266	0.266	0.266	0.266	0.266	0.266
	Haploview(FGT)	129	0.263	0.284	0.277	0.263	0.263	0.293	0.263
	Haploview(SS)	60	0.23	0.342	0.331	0.264	0.263	0.362	0.321
MIF 24,211,980~24,238,079 (67 SNPs) σ=10	Big-LD	1	0.427	0.509	0.484	0.701	0.695	0.674	0.764
	S-MIG++	5	0.355	0.455	0.446	0.495	0.485	0.488	0.573
	MIG++	3	0.262	0.519	0.514	0.523	0.537	0.553	0.609
	Haploview(CI)	3	0.262	0.519	0.514	0.523	0.537	0.553	0.609
	Haploview(FGT)	8	0.353	0.486	0.447	0.478	0.484	0.441	0.503
	Haploview(SS)	2	0.399	0.585	0.476	0.665	0.68	0.615	0.659
GSTT1 24,344,926~24,385,697 (36 SNPs) σ=25	Big-LD	1	0.386	0.92	0.893	0.792	0.794	0.923	0.924
	S-MIG++	20	0.602	0.618	0.62	0.618	0.62	0.605	0.605
	MIG++	15	0.593	0.658	0.651	0.658	0.651	0.638	0.637
	Haploview(CI)	36	0.552	0.552	0.552	0.552	0.552	0.552	0.552
	Haploview(FGT)	36	0.552	0.552	0.552	0.552	0.552	0.552	0.552
	Haploview(SS)	17	0.553	0.649	0.615	0.553	0.553	0.635	0.627
ZNRF3 29,523,628~29,556,739 (44 SNPs) σ=14	Big-LD	1	0.564	0.936	0.952	0.829	0.825	0.95	0.944
	S-MIG++	10	0.72	0.754	0.791	0.667	0.667	0.778	0.793
	MIG++	11	0.719	0.82	0.82	0.82	0.82	0.808	0.807
	Haploview(CI)	11	0.719	0.82	0.82	0.82	0.82	0.808	0.807