A LINKAGE DISEQUILIBRIUM-BASED APPROACH TO SELECTING DISEASE-ASSOCIATED RARE VARIANTS

Sanjay Shete

Department of Biostatistics

UT MD Anderson Cancer Center

Importance of Rare Variants

- Complex disorders

  -extreme allelic heterogeneity

  -caused by multiple rare variants with moderate to high penetrance

- Evolution theory suggests that allelic heterogeneity might be extensive with multiple susceptible alleles of independent origin (Pritchard, J.K., and Cox, N.J. (2002). The allelic architecture of human disease genes: Common disease-common variant or not? Hum. Mol. Genet. 11, 2417–2423).
- Rare variants are more likely to be disease predisposing

Importance of Rare Variants-read examples

- Multiple rare variants identified to be associated with common complex diseases

Low plasma levels of HDL cholesterol (Cohen et al. 2004, Science, Cohen et al. 2006 PNAS, Romeo et al. 2007 Nature Genetics)

Obesity (Ahituv et al. 2007 AJHG)

Colorectal adenomas (Azzopardi et al. 2008 Cancer Research)

Schizophrenia (Walsh 2008 Science)

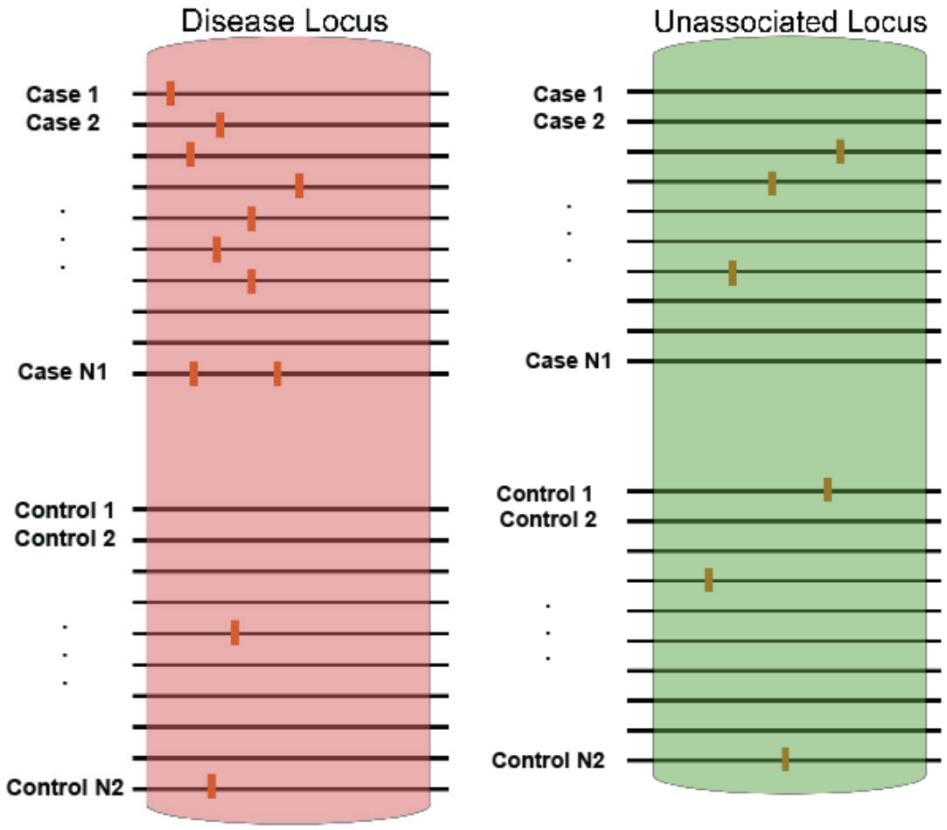
Rare Variant Analysis

- Rare Variants MAF<0.5%
- Rare variants are cited as measure contributor to disease etiology
- Several statistical methods are developed to analyze rare variants.

Rare Variant Analysis

- Association Testing

Goal– Determining if rare variants at a particular locus are associated with the disease.



Some of the Existing Methods

- Morris and Zeggini have shown that tests based on single variants have limited power compared with tests based on summing or collapsing rare variants.

- Cohort Allelic Sum Test (CAST) based on the difference in the number of variant alleles in cases and controls. It collapses information on all rare variants within a region into a single binary variable for each individual-whether or not an individual has ANY rare variants within the region and then perform regression with this binary variable (Morgenthaler and Thilly).

Some of the Existing Methods

- Morris and Zeggini: collapse by counting the number of rare variants within a region per subject, then apply the standard regression approach.

- The Combined Multivariate and Collapsing (CMC) test extends CAST by collapsing all variants within a region into subgroups based on a minor allele frequency threshold then collapse all rare variants within a subgroup as in CAST, and then applies a multiple regression model (significance is tested using Hotelling's T^2 statistic.

Some of the Existing Methods

- In both the CMC and CAST, all variants are assumed to have an equal effect on the phenotype.
- Therefore, Madsen and Browning proposed a Weighted Sum Statistic (WSS), which weighs the variants based on the inverse of the estimated standard deviation of the total number of rare variants in the sample. (assumes rarer variants have higher impact on the disease). For each variant i and individual j, a genetic score is calculated.

$$\gamma_j = \sum_{i=1}^{L} \frac{I_{ij}}{\hat{w}_i},$$

Some of the Existing Methods

- Where I_{ij} number of rare variants at location i for individual j. The weights w's are the estimated standard deviation of the number rare variants in the sample.
- Next all individuals (cases and controls) are ranked according to γ_j and then sum ranks for all cases is the test statistics.

$$x = \sum_{j \in A} \text{rank}(\gamma_j),$$

- Price et al. proposed a variable allele-frequency threshold method (VT) for selecting rare variants based on the assumption that variants with minor allele frequency below an unknown allele frequency threshold are more likely to be functional.

Some of the Existing Methods

- Hoffmann et al. used the general regression framework and model the weights as the product of three variables. The multiplicative model allows for weighting, direction of effect, and variant selection to be incorporated into the framework. They also proposed selecting variants based on functional significance and a data-driven method of variable selection called “step-up,” which is based on the standard forward selection algorithm.
- Wu et al. proposed a regression approach, sequence kernel association test (SKAT), based on the score-based variance components test.

Some of the Existing Methods

- SKAT

$$y_i = \alpha_0 + \boldsymbol{\alpha}' \mathbf{X}_i + \boldsymbol{\beta}' \mathbf{G}_i + \varepsilon_i,$$

$$\mathrm{H}_0\text{: } \boldsymbol{\beta} = \mathbf{0}, \text{ that is, } \beta_1 = \beta_2 = \ldots = \beta_p = 0.$$

- To increase power, SKAT assumes βs follow a distribution with mean zero and variance ώτ and equivalently tests for τ=0 using variance component score statistic

$$Q = (\mathbf{y} - \widehat{\boldsymbol{\mu}})' \mathbf{K} (\mathbf{y} - \widehat{\boldsymbol{\mu}}),$$

The Step-up method

- The step-up method is a data-driven method that tries to find the best possible set of rare variants by minimizing the p-value or maximizing a particular test statistic using the standard forward selection algorithm.
- The forward selection algorithm starts with no variants in the model and, at each iteration, adds variants to the model to maximize the Wald test statistic. The process stops when adding a variant to the model no longer increases the value of the test statistic.

Some of the Existing Methods

- Summing/ Weighted Summing/selection

  - SSU Test [Han and Pan(2010)]
  - C-alpha Test [Neale et al.(2011)]
  - SKAT [Wu et al.(2011)]
  - Combined Multivariate and Collapsing [Li and Leal(2008)]
  - Weighted Sum statistic [Madsen and Browning(2009)]
  - Step-up Method [Hoffman et al.(2010)]

Linkage Disequilibrium Patterns

- Our approach is to consider information about linkage disequilibrium (LD) among rare variants.
- Several measures of LD: r^2 and D' are most commonly used.
- We propose an approach, based on the LD among rare variants, for selecting a subset of variants to include in the analysis.

Linkage Disequilibrium

- **linkage disequilibrium** is the nonrandom association of alleles at two loci.

$$D_{AB} = p_{AB} - p_A p_B$$

- D' is D_{AB} divided by maximum possible value of D_{AB}.

$$r^2 = \frac{D^2}{p_A\,(1 - p_A) p_B (1 - p_B)}$$

Variable Selection- Improving Power

- Subset selection plays an important role.

- $Y = f(X_1, X_2, \ldots, X_n)$

  Goal– Select the best subset of predictors that will lead to an increase in power for association.

- Existing Method
  - Step Up (Hoffman et al. (2010))

Association Test

- Case Control Data for multiple rare variants.

$$g(Y_i) = \beta_0 + \beta_1 \left[\sum_{k=1}^{n} w_k X_k \right]$$

  - w_k can be modeled as $w_k = a_k s_k v_k$
  - a_k is used for up weighing variants based on their minor allele frequencies.
  - s_k is -1 or 1 based on whether the rare variant is protective or deleterious.
  - v_k is used to determines whether the rare variant is to be included in the model.

- Here we are only concerned with the variant selection hence we only use v_k in the formulation, a_k and s_k can be modeled using the best possible way.

- This model can include common variants as

$$g(Y_i) = \beta_0 + \beta_1 \left[\sum_{k=1}^{n} v_k X_k \right] + \sum_{j=1}^{p} \alpha_j Z_j$$

  - Z_j are the common variants.

Preliminary Observations

- We performed preliminary simulations to identify rare variants using these step-based approaches and found that

-(a) when rare variants were in LD, the step-based variable selection procedures resulted in a loss of power compared with the model that simply collapses all rare variants (named hereafter the full model) and

-(b) when rare variants were independent of each other (i.e., no LD among rare variants), step-based variable selection had higher power than the full model.

- Can we do better than selecting all variants in presence of LD?

Subset selection based on LD

- Motivation: Subset selection methods lose power because of the correlation among rare variants.
- We need to include all correlated variants in the model.
- New proposed subset selection algorithm: Remove a subset of noisy variants that don't contribute.

Algorithm for Subset Selection

The proposed algorithm that accounts for LD, called LDSEL, is as follows:

- 1. Use the step-down-up (backward-forward) selection method to select associated variants in the model.

  - Starts with all rare k variants in the model. Calculate the model p-value and R^2 (or equivalently likelihood ratio test statistic)
  - Calculate model p-value for all models in which only k-1 rare variants are included. Then pick the (k-1) variant model which has the smallest p-value and non-significant reduction in the R^2 value.

Algorithm for Subset Selection

- Forward component of the algorithm

  - When s variables are selected: calculate likelihood . Calculate the model p-value and R^2 value
  - Calculate model p-value for all models in which only s+1 rare variants are included. Then pick the (s+1) variant model which has the smallest p-value and non-significant reduction in the R^2 value.

Algorithm for Subset Selection

The proposed algorithm that accounts for LD, called LDSEL, is as follows:

- 1. Use the step-down-up (backward-forward) selection method to select associated variants in the model.
- 2. For each selected variant in step 1, identify all other variants that are in LD with the selected variants in the cases.
- 3. The union of variants identified in step 1 and step 2 forms the final selected subset of rare variants.
- The measure used for LD was r^2.

LD among rare variants

- Our algorithm collapses to the regular step-based methods when the rare variants are not in LD.
- It also collapses to the full model on the other extreme when all the rare variants are in LD with each other.
- Real data will have a block structure in which some variants are in LD and some are independent.

Comparing Methods

- Existing methods
  - Step-Up → Forward selection
  - VT→ Variable threshold
  - SKAT → Sequence Kernel Association Test

- We also considered variations of Step-Up method.
  - Step-Down → Backward selection
  - Step-Up-Down → Forward-backward selection.
  - Step-Down-Up → Backward-forward selection.

- We compared the results of these methods to the full model in which all the variants were included.

Simulation for observed LD in HapMap Data

- In this scenario, we simulated rare variants using the LD structure of the DRD2 gene from the HapMap3 data (HapMap3 Genome Browser, release #2 [Phase 3 - genotypes, frequencies, & LD]).
- The DRD2 gene spans 112,785,528 bp to 112,851,091 bp on chromosome 11.

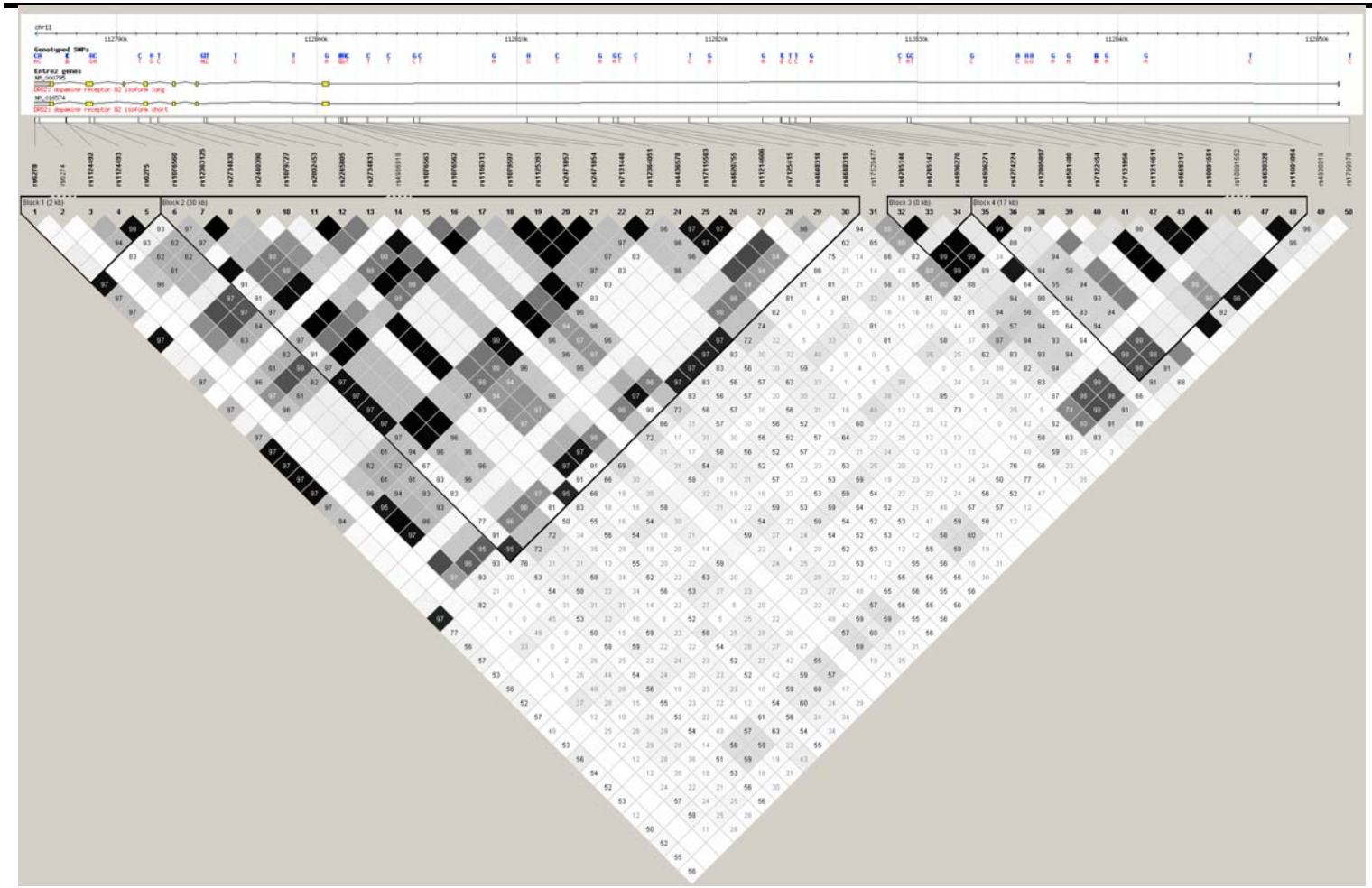
- We also considered CHRNA3/A5/B4, a gene cluster encompassing multiple genes.
- The CHRNA3/A5/B4 cluster spans 76,490,686 bp to 76,899,993 bp on chromosome 15.

Simulation for realistic LD

- DRD2 gene
- 50 rare variants were simulated with 0.25% < MAF < 0.5%
- The linkage disequilibrium between variants was simulated from Hap-map data of DRD2 gene.
- Out of 50 rare variants 5 variants were randomly designated to be causal variants.
- The disease model was a logistic regression model with OR =(2,2,2,2,2) for the 5 causal variants.

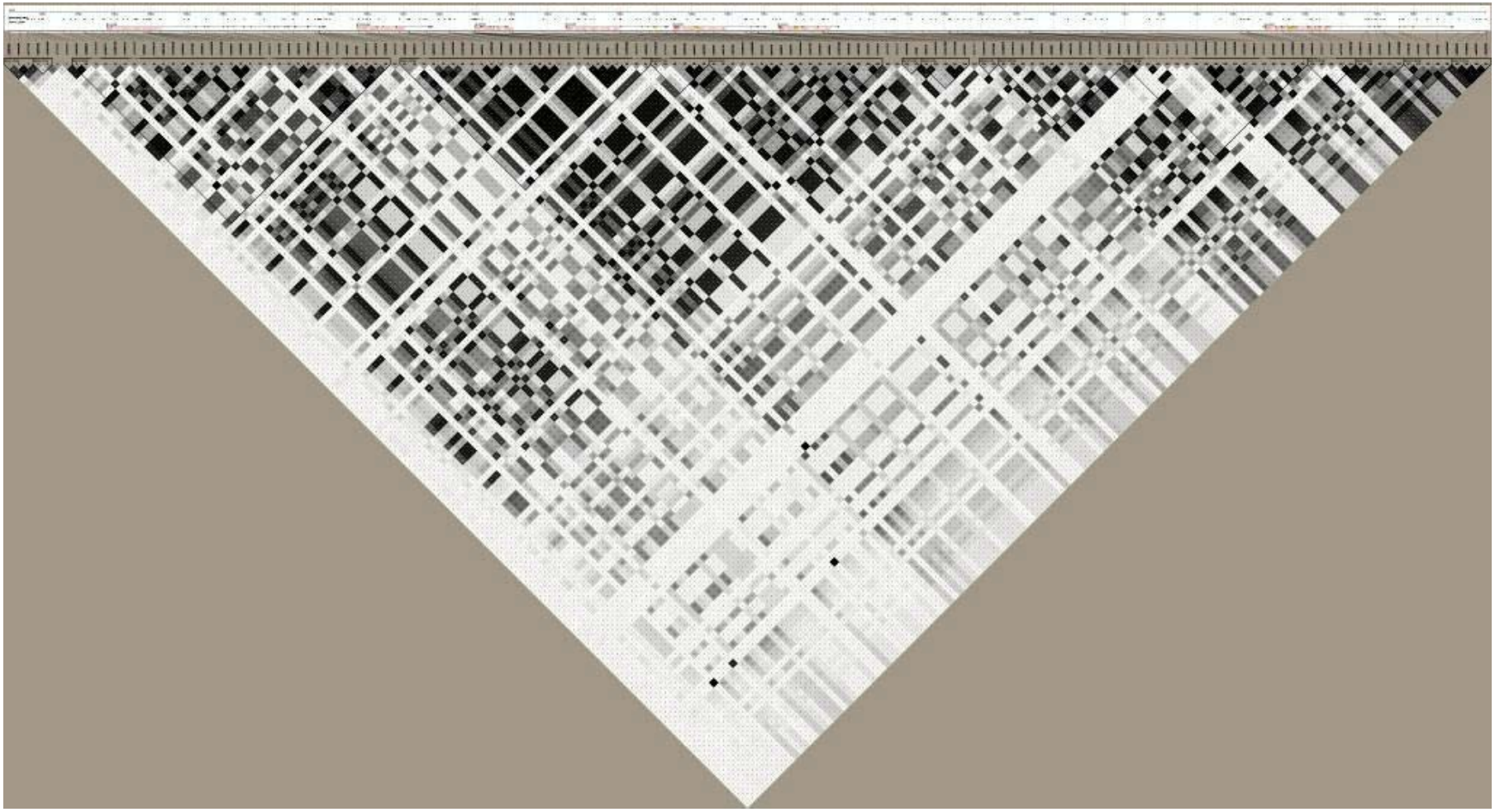
- CHRNA3/A5/B4: Of the 154 rare variants, 7 variants were randomly designated to be causal variants.

DRD2 Linkage Disequilibrium



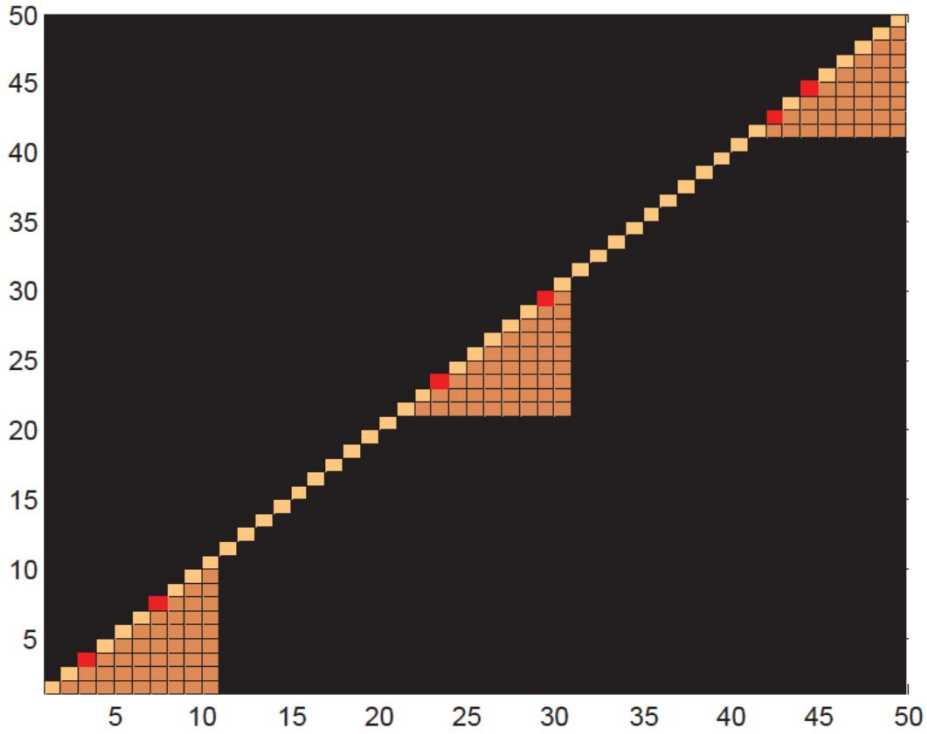
- Courtesy@ Hapmap data , Haploview

CHRNA3/A5/B4 Linkage Disequilibrium



Simulation for Block LD

- Block LD-all SNPs are in LD within a block
- Three LD blocks of 10 rare variants each were simulated.
- Two variants randomly designated as causal from each LD block.



Simulation for Block LD

- In addition to the associated variants within the three blocks, we simulated 20, 70, or 170 independent variants (i.e., not in LD with any other variants) outside the three LD blocks.

- These three different numbers (20, 70 and 170) of non-causal and non-associated variants were simulated to assess the performances of the different methods over a range of signal-to-noise ratios.

Simulation for variants in linkage equilibrium

- All 50 variants were independently simulated for the linkage equilibrium scenario with 0.25% < MAF < 0.5%
- 5 variants were randomly designated as causal.
- 1000 cases and 1000 controls and results are based on 1000 replicates.

Type 1 error Comparison

- Type 1 error rate for 1000 replicates generated from a null model

Method	Type 1 error			
	Uncorrected		Corrected	
	5%	1%	5%	1%
Step-up	0.996	0.946	0.052	0.009
Full model	0.049	0.012	0.049	0.012
Step-down	1	0.966	0.051	0.010
Step-up-down	0.996	0.946	0.052	0.009
Step-down-up	1	0.966	0.051	0.010
LDSEL	0.353	0.183	0.049	0.010
SKAT	0.045	0.011	0.048	0.009
VT	0.071	0.016	0.053	0.012

doi:10.1371/journal.pone.0069226.t001

- Use permutations to control Type 1 Error.

Power Comparison in presence of block LD

- The power of the full model was higher than that of the step-based methods.
- For low number of non associated variants the proposed LDSEL method had higher power than all the step-based methods, VT, SKAT and also slightly better power than the full model
- For higher number of non associated variants, the power of our proposed LDSEL method was higher than the full model, SKAT, and VT.

Method	Power					
	20 NAV		70 NAV		170 NAV	
	5%	1%	5%	1%	5%	1%
Step-up	0.794	0.560	0.508	0.290	0.319	0.134
Full model	0.872	0.710	0.733	0.502	0.575	0.313
Step-down	0.797	0.566	0.511	0.292	0.322	0.135
Step-up-down	0.794	0.560	0.508	0.290	0.319	0.134
Step-down-up	0.797	0.566	0.511	0.292	0.322	0.135
LDSEL	0.882	0.715	0.794	0.590	0.629	0.361
SKAT	0.838	0.604	0.734	0.528	0.608	0.338
VT	0.824	0.596	0.635	0.373	0.504	0.226

The three panels correspond to simulation scenario 2 having 20, 70, and 170 non-associated variants (NAV) respectively along with three LD blocks of 10 variants, with 2 causal variants in each block.
doi:10.1371/journal.pone.0069226.t003

Power Comparison with LD as in DRD2 and CHRNA3/A5/B4

- For the DRD2 gene, the power of LDSEL approach was higher than SKAT, VT, and the step-based methods.
- For the CHRNA3/A5/B4 cluster, the LDSEL approach had higher power than the step-based methods and VT, but had similar power as the full model and SKAT.
- The VT method had slightly lower power than the step-based methods in the first panel whereas in the second panel it had higher power than the step-based methods.

Method	Power			
	DRD2		CHRNA3/A5/B4	
	5%	1%	5%	1%
Step-up	0.526	0.297	0.352	0.148
Full model	0.538	0.302	0.565	0.32
Step-down	0.528	0.298	0.359	0.152
Step-up-down	0.526	0.297	0.352	0.148
Step-down-up	0.528	0.298	0.359	0.152
LDSEL	0.552	0.308	0.569	0.329
SKAT	0.468	0.228	0.579	0.323
VT	0.512	0.239	0.486	0.253

The first panel corresponds to simulation scenario 1 using the LD structure of the DRD2 gene and the second panel corresponds to simulation scenario 1 using the LD structure of the CHRNA3/A5/B4 gene cluster.
doi:10.1371/journal.pone.0069226.t002

Power Comparison without LD

- The step-based methods and outperformed all other methods when the variants were in linkage equilibrium.
- The step down up method performed marginally better compared to the other step based methods
- The LDSEL method had power which was significantly higher than the full model but marginally lower than the step-based methods and SKAT.

Method	Power	
	5%	1%
Step-up	0.463	0.232
Full model	0.331	0.128
Step-down	0.469	0.237
Step-up-down	0.463	0.232
Step-down-up	0.469	0.237
LDSEL	0.442	0.216
SKAT	0.45	0.226
VT	0.172	0.074

doi:10.1371/journal.pone.0069226.t004

Discussion

- LDSEL is a flexible method and, depending on the structure of LD between the variants, it converts to a full model when all of the rare variants are in LD or to a step-based approach when all the variants are in linkage equilibrium.

- As the number of non-associated variants being pooled increased, substantial power was gained by the LDSEL method compared with the full model or the step-based selection methods.

Total 51 Genes Sequenced

	Gene_Symbol	Chr	Gene_start	Gene_end	Gene_size
1	CHRNB2	1	154,540,257	154,552,502	12,245
2	ADCY3	2	25,042,038	25,142,708	100,670
3	ADRA2B	2	96,778,623	96,781,984	3,361
4	CHRNA1	2	175,612,320	175,629,200	16,880
5	CHRND	2	233,390,703	233,401,377	10,674
6	CHRNG	2	233,404,437	233,412,546	8,109
7	CREB1	2	208,394,461	208,470,284	75,823
8	ADCY5	3	123,001,143	123,168,605	167,462
9	DRD3	3	113,847,499	113,918,254	70,755
10	ADRA2C	4	3,768,075	3,770,253	2,178
11	CHRNA9	4	40,337,346	40,357,234	19,888
12	ADCY2	5	7,396,321	7,830,194	433,873
13	ADRA1B	5	159,343,740	159,400,017	56,277
14	ADRB2	5	148,206,156	148,208,197	2,041
15	DRD1	5	174,867,675	174,871,163	3,488
16	SLC6A3	5	1,392,905	1,445,545	52,640
17	SLC22A2	6	160,592,093	160,698,670	106,577
18	ADCY1	7	45,613,739	45,762,715	148,976
19	ADCY8	8	131,792,547	132,054,672	262,125
20	ADRA1A	8	26,605,667	26,724,790	119,123
21	CHRNA2	8	27,317,278	27,337,400	20,122
22	CHRNA6	8	42,607,763	42,651,535	43,772
23	CHRNB3	8	42,552,519	42,592,550	40,031
24	DBH	9	136,501,482	136,524,466	22,984
25	SH2D3C	9	130,500,596	130,541,048	40,452
26	ADRA2A	10	112,836,790	112,840,665	3,875

	Gene_Symbol	Chr	Gene_start	Gene_end	Gene_size
27	ADRB1	10	115,803,806	115,806,667	2,861
28	CHAT	10	50,817,141	50,901,925	84,784
29	CYP2E1	10	135,333,910	135,374,724	40,814
30	DRD2	11	113,280,317	113,346,413	66,096
31	DRD4	11	637,293	640,706	3,413
32	HTR3A	11	113,845,603	113,861,035	15,432
33	HTR3B	11	113,775,399	113,817,287	41,888
34	ADCY6	12	49,159,975	49,182,820	22,845
35	ADCY4	14	24,787,555	24,804,299	16,744
36	AGPHD1	15	78,799,906	78,829,715	29,809
37	CHRFAM7A	15	30,653,443	30,686,052	32,609
38	CHRNA3	15	78,885,394	78,913,637	28,243
39	CHRNA5	15	78,857,862	78,887,611	29,749
40	CHRNA7	15	32,322,691	32,464,722	142,031
41	CHRNB4	15	78,916,461	79,012,628	96,167
42	IREB2	15	78,729,773	78,793,798	64,025
43	PSMA4	15	78,832,747	78,841,604	8,857
44	ADCY7	16	50,280,048	50,352,046	71,998
45	SLC6A2	16	55,689,516	55,740,104	50,588
46	ARRB2	17	4,613,784	4,624,795	11,011
47	CDK5R1	17	30,813,637	30,818,274	4,637
48	CYP2A6	19	41,349,443	41,356,352	6,909
49	CYP2B6	19	41,497,204	41,524,301	27,097
50	ADRA1D	20	4,201,278	4,229,721	28,443
51	CHRNA4	20	61,974,665	62,009,753	35,088

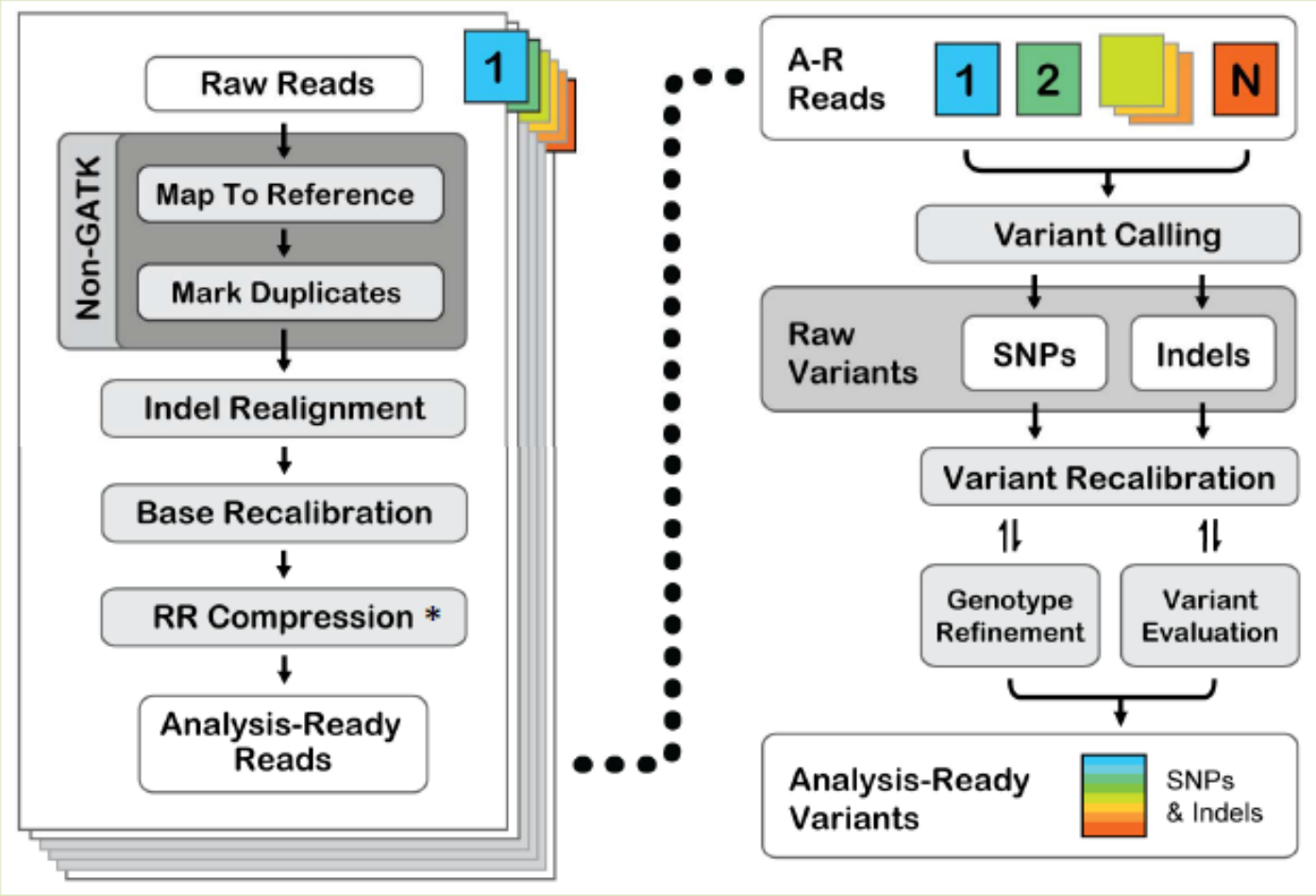
Number of Variant Calls from 431 VCF Files

chr	count
1	202,337
2	211,542
3	167,048
4	168,914
5	152,950
6	150,173
7	142,040
8	140,727
9	106,717
10	129,697
11	138,567
12	117,206
13	
14	79,310
15	124,057
16	93,881
17	74,315
18	
19	87,633
20	63,240
21	
22	
23	
24	
26	
sum	2,350,354

~Total 2.3 millions

Number of Variant Calls from 431 INDEL VCF Files and SNP VCF Files

chr	INDEL vcf count	SNP vcf count	INDEL vcf count	SNP vcf count
1	163,971	202,337	163,971	202,337
2	171,894	211,542	171,894	211,542
3	129,142	167,048	129,142	167,048
4	125,906	168,914	125,906	168,914
5	122,901	152,950	122,901	152,950
6	116,341	150,173	116,341	150,173
7	116,807	142,040	116,807	142,040
8	107,231	140,727	107,231	140,727
9	84,871	106,717	84,871	106,717
10	103,556	129,697	103,556	129,697
11	107,183	138,567	107,183	138,567
12	97,450	117,206	97,450	117,206
13	62,956	80,973		
14	62,909	79,310	62,909	79,310
15	111,175	124,057	111,175	124,057
16	74,194	93,881	74,194	93,881
17	65,217	74,315	65,217	74,315
18	51,188	65,836		
19	83,432	87,633	83,432	87,633
20	50,979	63,240	50,979	63,240
21	29,673	38,265		
22	30,336	34,203		
23	70,603	67,337		
24	6,612	7,523		
26	44	289		
sum	**2,146,571**	2,644,780	**1,895,159**	2,350,354



* RR Compression – compress file size through reducing reads (remove redundant info)

Ref: http://www.broadinstitute.org/gatk/guide/

Thoughts

- Our initial goal was to reduce the number of SNV's (rare) to be short listed (for biological follow-up) using the LD approach.
- We also tried to reduce the number of variants by comparing LD pattern of selected variants in cases and controls.
- Will the case control study design better to identify rare variants or will the family study design be better?
- Answer depends upon many factors:
- Single variant segregating in multiple families
- Different variant segregating in different families

Acknowledgement

- Rajesh Talluri, UT MD Anderson Cancer Center

- Talluri R, Shete S (2013) A Linkage Disequilibrium-Based Approach to Selecting Disease-Associated Rare Variants. PLoS ONE 8(7): e69226. doi:10.1371/journal.pone.0069226

- National Institutes of Health grants R01CA131324, R01DE022891, and R25DA026120.