Testing for Disease-Rare Variant Association with Sequence Data

Wei Pan[1]

[1]Division of Biostatistics, School of Public Health

University of Minnesota

Oct 28, 2011

Outline

- Introduction: problem

  No data preprocessing; genotypes called.

- Review some existing and new methods

  Pooled association tests, e.g., Sum test;

  Newer ones: aSum, SSU tests.

- Example data: 1000 Genome Project

- Main refs:

  Pan (2009, *Genet Epi*), Han and Pan (2010, *Hum Hered*), Basu and Pan (2011, *Genet Epi*), Pan and Shen (2011, *Genet Epi*), ...

Introduction

- Single Nucleotide Polymorphism (SNP) or Variant (SNV)

  DNA seq 1 – AAGC**C**TA

  DNA seq 2 – AAGC**T**TA

  two alleles, C and T; 3 genotypes: CC, TT, CT;

  SNP: a minor allele freq (MAF) $\geq$ 5% (or 1%).

  SNV: less frequent variant or rare variant (RV) with MAF $<$ 1%.

- Genome-wide *association* studies (GWAS):

  Genome-wide tag SNPs (1 M) are measured as markers for each subject;

  Target: common disease–common variant (CD-CV) association;

  Ultimate goal: to detect *causal* CVs.

- GWAS: a success!?

  As of **10/5/11** (or 01/19/11 or 9/24/09 or 11/24/08), the NIH

Catalog of Published Genome-Wide Association Studies "includes **1030** (or 791 or 396 or 202) publications and **5108** (or 3939 or 1760 or 435) SNPs" that are associated with some phenotypes, such as prostate cancer, diabetes, bipolar disorder...

- But ... explain only a **small** proportion of heritability! Willer et al (2009): BMI; $n = 3287$ and 45018 for stages 1 and 2; identified 8 loci, explaining 0.84% of phenotype variance; genetic heritability 40-70%.
- Possibilities: polygenic (small) effects; G-G and G-E interactions; other variants (e.g. CNV); RVs; ...
- PCSK9 gene (Kotowski et al 2006):
  some RVs associated with **lower** plasma levels of LDL-C;
  some RVs associated with **higher** plasma levels of LDL-C;
- Next-generation sequencing (NGS):

Sequence (SNVs) of whole exome or genome for each subject;
Target: common disease–RV association

- Most common study design: case-control;
  n in hundreds, then thousands, then ?
- Analysis unit
  GWAS: single SNPs; more multi-SNP analyses?
  NGS: multiple RVs, e.g. in a candidate gene or region;

- Data:

Obs	Y	SNP1	SNP2	SNP3	...	SNPk
1	1	CT	AG	CG	...	AC
2	1	TT	AG	GG	...	AA
3	1	CT	AA	CG	...	CC
......						
1001	0	CT	AG	CC	...	AC
1002	0	TT	GG	CC	...	AC
1003	0	CC	GG	CC	...	CC
......						

- A binary response: $Y = 0$ or 1; each SNP j is coded as $X_j = 0$, 1 or 2, # copies of minor alleles;
- Statistical question: any SNP associated with Y?
- Most popular test in GWAS: univariate or single SNP-based

- Should it be multivariate?
  e.g., $k > 1$ SNPs inside a **given** LD block or sliding window.
  Selection of LD block or window size: relevant, not trivial.

- For RVs: small MAF $\Longrightarrow$ univariate tests ...
  $n = 1000$, MAF=1% $\Longrightarrow$ #(minor alleles) ≈ 20;
  $n = 1000$, MAF=0.1% $\Longrightarrow$ #(minor alleles) ≈ 2;
  Design matrix X: almost all 0's!

- RVs: small MAF $\Longrightarrow$ aggregation!
  combine multiple RVs!

Existing methods

- Single-locus (or SNP-by-SNP or univariate) analysis: GWAS
  - Model: $Y \sim SNP_j$

$$\text{Logit } \Pr(Y_i = 1) = \beta_{M,0j} + X_{ij}\beta_{M,j}, \qquad (1)$$

  - $H_{0,j}$: $\beta_{M,j} = 0$ for each $j = 1, ..., k$
    $\Longrightarrow p_j$.
  - Combining: $UminP = \min(p_1, p_2, ..., p_k)$ or ...
    Need to do multiple test adjustment!
  - Model (1): as a 2×3 table; Cochran-Armitage trend test.

- Multivariate (or global or joint) analysis:
  - Model: $Y \sim SNP_1 + ... + SNP_k$

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j, \tag{2}$$

  - H_0: $\beta_1 = ... = \beta_k = 0$
  - Use the score, Wald or LR test:
    $T_W = \hat{\beta}' V^{-1} \hat{\beta}$, $T_S = U' V_U^{-1} U \sim \chi^2_k$ under H_0;
    $V = Cov(\hat{\beta})$, $V_U = Cov(U)$;
    Possibly large $DF = k$.
  - Hotelling's T^2 test: closely related to the score test.

- Pooled association tests: aggregation; Sum test
  - *Working* (and *incorrect*) assumption: $\beta_1 = ... = \beta_k \equiv \beta_c$.
  - Model:

$$\text{Logit Pr}(Y_i = 1) = \beta_{0,c} + \sum_{j=1}^{k} X_{ij}\beta_c = \beta_{0,c} + X_{i,c}\beta_c, \qquad (3)$$

  - $H_{0,c}$: $\beta_c = 0$
  - Apply the score, Wald or LR test
    $T_W = \hat{\beta}_c^2/V_c \sim \chi_1^2$ under $H_{0,c}$.
  - Feature: DF=1; no multiple testing!
  - Correct test size:
    $H_0 \Longrightarrow H_{0,c}$!
  - Closely related to CMC (Li and Leal 2008), weighted sum (Madsen and Browning 2009) tests:
    $\vee_{j=1}^{k} X_{ij} \approx \sum_{j=1}^{k} X_{ij}$

Power: OR=(2, 2, 2, 2, 2, 2, 2, 2); No LD; $n = 500 + 500$; MAFs $\sim U(.001, .01)$ for controls;

Test	# of neutral RVs					
	0	4	8	16	32	64
UminP	.441	.336	.296	.222	.175	.117
Score	.746	.632	.595	.471	.332	.245
CMC	.938	.853	.777	.616	.399	.211
wSum	.940	.846	.782	.618	.424	.267
Sum	**.951**	**.875**	**.808**	**.673**	.484	.313
aSum	.933	**.858**	.780	.669	.499	.313
SSU	.756	.702	.694	.626	.499	**.423**
KMR(Linear)	.762	.711	.699	.631	**.509**	**.438**
C-alpha	.771	.712	.688	.627	.484	.378

Power: $OR = (3, 3, 2, 2, 2, 1/2, 1/2, 1/2)$; No LD; $n = 500 + 500$; MAFs $\sim U(.001, .01)$ for controls;

Test	# of neutral RVs				
	0	4	8	16	32
UminP	.607	.532	.481	.417	.346
Score	.869	.772	.721	.632	.483
CMC	.661	.544	.456	.336	.204
wSum	.659	.548	.459	.335	.228
Sum	.682	.566	.465	.365	.258
aSum	.854	.745	.684	.574	.430
SSU	.895	**.835**	**.815**	**.774**	**.696**
KMR	**.897**	**.842**	**.824**	**.783**	**.707**
C-alpha	**.906**	**.844**	**.823**	**.775**	**.674**

Newer methods

- Summary: 1) pooled association tests (Sum, CMC, wSum) do not perform well if there are opposite association directions!
- A strategy: decide the association directions first!
  An adaptive Sum (aSum) test: Han and Pan (2010);
  More works:... But ...
- Equally (or more?) importantly, pooled association tests (Sum, CMC, wSum) do not perform well if there are many non-asscoiated RVs.
  Presence of non-asscoiated RVs: expected!
- A strategy: SSU test!

- Recall $LRT \approx Wald's \approx Score = U'V^{-1}U$,
  $U = \sum_{i=1}^{m} X_i(Y_i - \bar{Y})$,
  $V = Cov(U) = I_F = \bar{Y}(1-\bar{Y})(X-\bar{X})'(X-\bar{X})$.

- New tests:

$$SSU = U'U \approx SSB = \sum_{j=1}^{k} \hat{\beta}_{M,j}^2,$$

- Null distributions for $Q = U'W^{-1}U$:
  1) $W = I$ and $W = \text{Diag}(V_M)$ in the above;
  2) $Q \sim \sum_{j=1}^{k} c_j \chi_1^2$, where c_j's are the eigen values of $V_M W^{-1}$;
  3) Zhang (2005, JASA): approximate by $a\chi_d^2 + b$ with

$$a = \frac{\sum_{j=1}^{k} c_j^3}{\sum_{j=1}^{k} c_j^2}, \quad b = \sum_{j=1}^{k} c_j - \frac{\left(\sum_{j=1}^{k} c_j^2\right)^2}{\sum_{j=1}^{k} c_j^3}, \quad d = \frac{\left(\sum_{j=1}^{k} c_j^2\right)^3}{\left(\sum_{j=1}^{k} c_j^3\right)^2}.$$

  4) $Pr(SSU > s|H_0) \approx Pr\left(\chi_d^2 > (s-b)/a\right)$.

- A weighted version of SSU: $SSUw = U'diag(V)^{-1}U$.

- Result 1: SSU = Goeman's EB test for high-dim data:
- Goeman's test:
  - Set-up: "large k, small n" as for microarray data;
  - Assume $\beta = (\beta_1, ..., \beta_k)'$ random:
    $E(\beta) = 0$, $Cov(\beta) = \tau^2 I$.
  - Test H_{0,τ^2}: $\tau^2 = 0$ by a score test.
  - For logistic regression:
    $T_{Go} = \frac{1}{2}(U'U - \mathrm{Trace}(I_F))$, where $U = X'(Y - \bar{Y})$,
    and $I_f = Cov(U) = \bar{Y}(1 - \bar{Y})(X - \bar{X})'(X - \bar{X})$.

$$\begin{aligned} T_{Go} &= \frac{1}{2}(Y - \bar{Y})'XX'(Y - \bar{Y}) - \\ & \frac{1}{2}\bar{Y}(1 - \bar{Y})\mathrm{Trace}((X - \bar{X})'(X - \bar{X})), \end{aligned}$$

Conditional on Y the second term is fixed (i.e. non-random) and can be dropped:

$$T_{Go} = \frac{1}{2}U'U + c_0 = \frac{1}{2}U'_M U_M + c_0 \propto SSU.$$

- Why do SSU/SSUw work?
  How could they beat "optimal" score, Wald and LR tests???
- Cox and Hinkley, *Theoretical Statistics*, 1974:
  - ∗ Optimality of the score, Wald and LR tests:
    locally most powerful, but only for ...;
    o/w, no uniformly most power (unbiased) (UMPU) test!
  - ∗ If we knew β, then
    $T_{MP} = \beta'U$, **but** ...
  - ∗ Try $\max_b b'U$ s.t. $Var(b'U) = b'I_F b = 1$?
- We estimate T_{MP} by
  $T_{EMP} = \hat{\beta}'_M U$.

– $T_{EMP} \approx SSUw = U'\mathrm{Diag}(I_F)^{-1}U$ because

$$\hat{\beta}_M = I_{M,d}^{-1} U_M + O_p(m^{-1}), \qquad U = U_M. \tag{4}$$

– How about estimating β by $\hat{\beta}$?
$T_{EMP,J} = \hat{\beta}' U \approx U' I_F^{-1} U$, which is ...

- Result 2: SSU = kernel machine regression(KMR) (Wu et al 2010, 2011, *AJHG*) if a suitable kernel (or design matrix) is used.
  - KMR (Kwee et al 2008, *AJHG*; Wu et al 2010, *AJHG*): use a semi-parametric regression model

    $$\text{Logit Pr}(Y_i = 1) = \beta_0 + h(X_{i1}, ..., X_{ik}), \tag{5}$$

    $h(.)$ is an unknown function to be estimated. The form of $h(.)$ is determined by a user-specified positive and semi-definite (psd) kernel function $K(.,.)$: by the representer theorem (Kimeldorf and Wahba 1971), $h_i = h(X_i) = \sum_{j=1}^{n} \gamma_j K(X_i, X_j)$ with some γ_1,...,γ_n.
  - To test H_0: $h = (h_1(X_1), ..., h_n(X_n))' = 0$.
    let $K = (K(X_i, X_j))$, $\gamma = (\gamma_1, ..., \gamma_n)'$, then $h = K\gamma$.
    Assume h as subject-specific random effects:
    $E(h) = 0$, $Cov(h) = \tau K$.

$H_0 = H_0'$: $\tau = 0$.
Score test for H_0':

$$Q = (Y - \bar{Y}1)' K (Y - \bar{Y}1) = SSU$$

for H_0'': $b = 0$ in

$$\text{Logit} \Pr(Y = 1) = b_0 + Zb$$

with $K = ZZ'$.

- Result 3: SSU = genomic distance based regression (GDBR) (Wessel and Schork 2006, *AJHG*) if a suitable distance metric (or design matrix) is used.

$$\begin{aligned} F &= \frac{tr(\hat{Y}'\hat{Y})}{tr(R'R)} = \frac{tr(\hat{Y}\hat{Y}')}{tr(RR')} = \frac{tr(HYY'H)}{tr((I-H)YY'(I-H))} \\ &= \frac{tr(HGH)}{tr((I-H)G(I-H))} \propto SSU \end{aligned}$$

for H_0'': $b = 0$ in

$$\text{Logit Pr}(Y=1) = b_0 + Zb$$

with $G = ZZ'$.

- A side-product: KMR=GDBR=SSU if $K = G = ZZ'$.
- Result 4: SSU $\approx$ C-alpha test (Neale et al 2011, *PLoS Genet*)
  Recall: SSU = Goeman's EB test;
  Assume $\beta = (\beta_1, ..., \beta_k)' \sim N(0, \tau^I)$, test H_0: $\tau^2 = 0$.
  Both Goeman's and C-alpha tests: a homogeneity test!
- Remark: weighting can be used,
  1) as in wSum, weight $\propto$ 1/MAF;
  2) functional prediction, e.g. by SIFT,...

Power: $OR = (3, 1/3, 2, 2, 2, 1/2, 1/2, 1/2)$; with LD.

	# of neutral RVs				
Tests	0	4	8	16	32
UminP	.489	.479	.452	.365	.318
Score	.599	.538	.491	.380	.276
CMC	.365	.296	.283	.189	.182
wSum	.369	.297	.287	.191	.200
Sum	.342	.312	.315	.258	.239
aSum	.350	.323	.325	.258	.243
SSU	.603	**.624**	**.635**	**.581**	**.574**
KMR	.611	**.630**	**.644**	**.597**	**.590**
C-alpha-P	**.629**	**.650**	**.668**	**.607**	**.598**

Power: only one causal RV with OR=5:

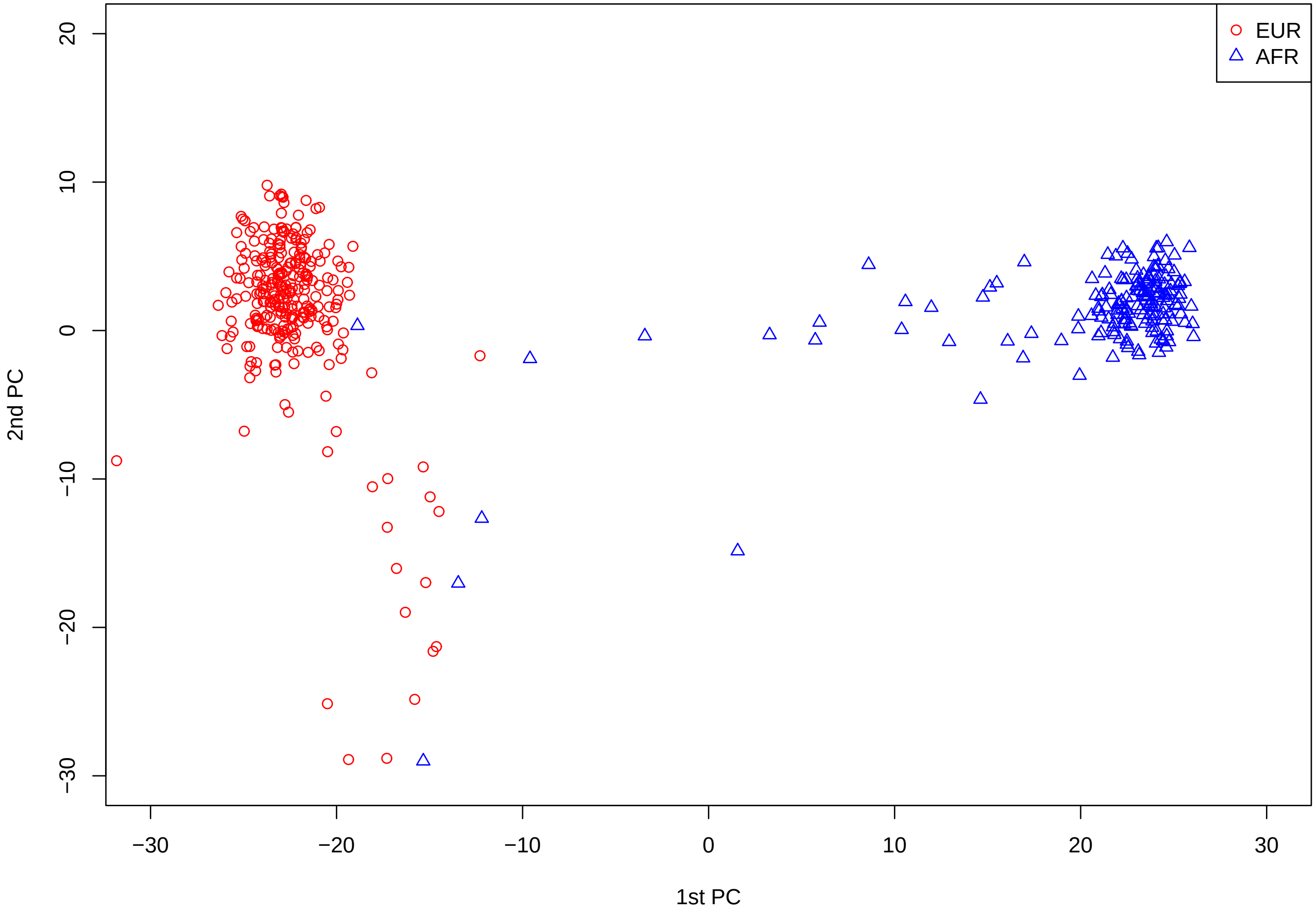
Test	# of neutral RVs					
	8	16	32	64	96	128
UminP	.696	.629	.556	.496	.479	.461
Sum	.365	.263	.160	.096	.088	.086
aSum	.447	.314	.215	.152	.130	.126
KBAC	.629	.483	.330	.193	.128	.103
PWST	.665	.533	.405	.280	.211	.174
EREC	.685	.545	.424	.272	.197	.184
SSU	.710	.664	.580	.520	.470	.427
aSSU	**.736**	**.685**	.628	.561	.518	.481
aSPU	.707	.683	**.645**	**.615**	**.592**	**.571**

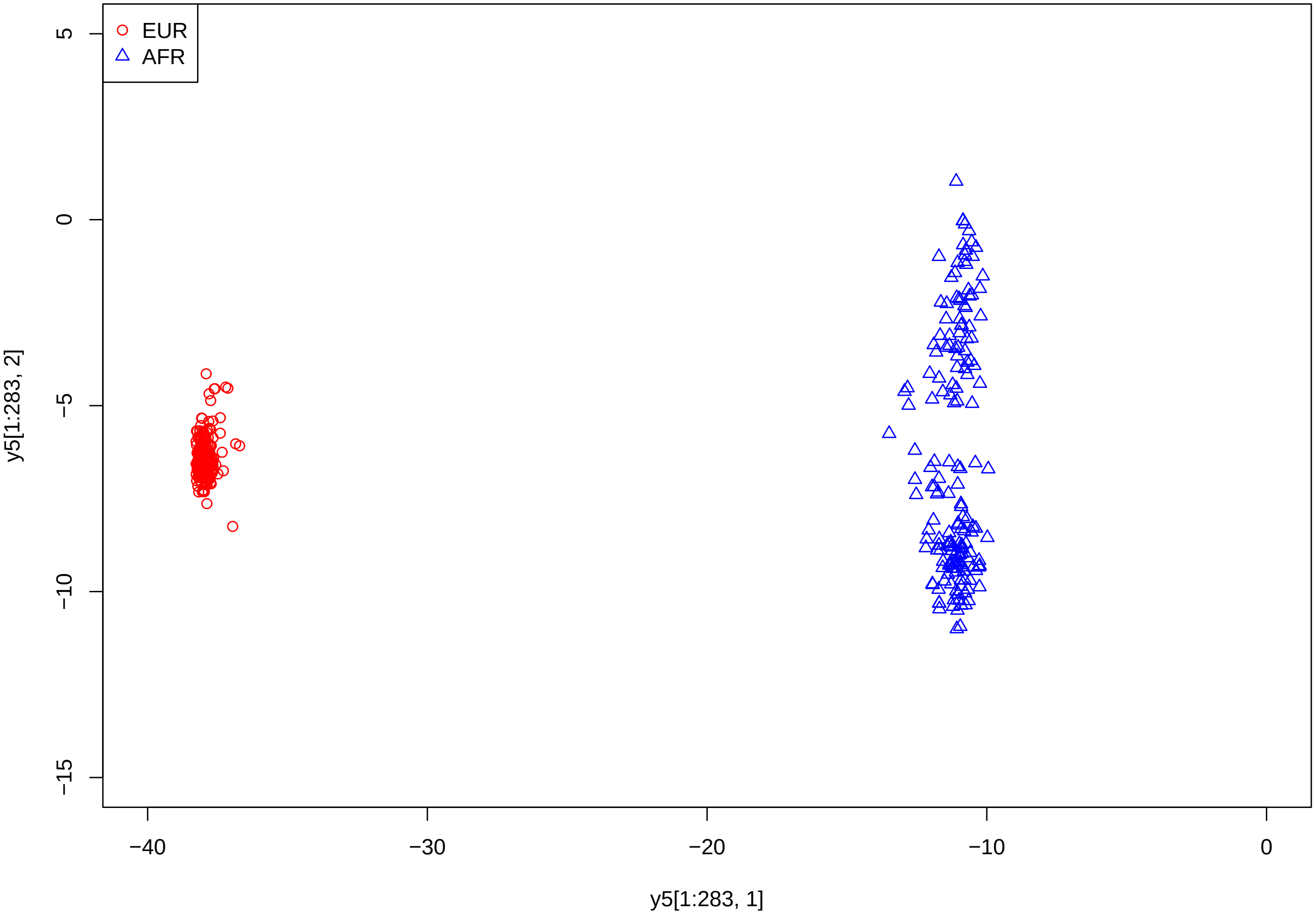
Example

- 1000 Genome Project, http://www.1000genomes.org/
  "The genomes of about 2500 unidentified people from about 25 populations around the world will be sequenced using next-generation sequencing technologies. The results of the study will be freely and publicly accessible to researchers worldwide."
  June 2011 Data Release: "Genotypes for 1094 individuals for the May 2011 snp calls from the 2010-11-23 sequence and alignment release of the 1000 genomes project has now been made."

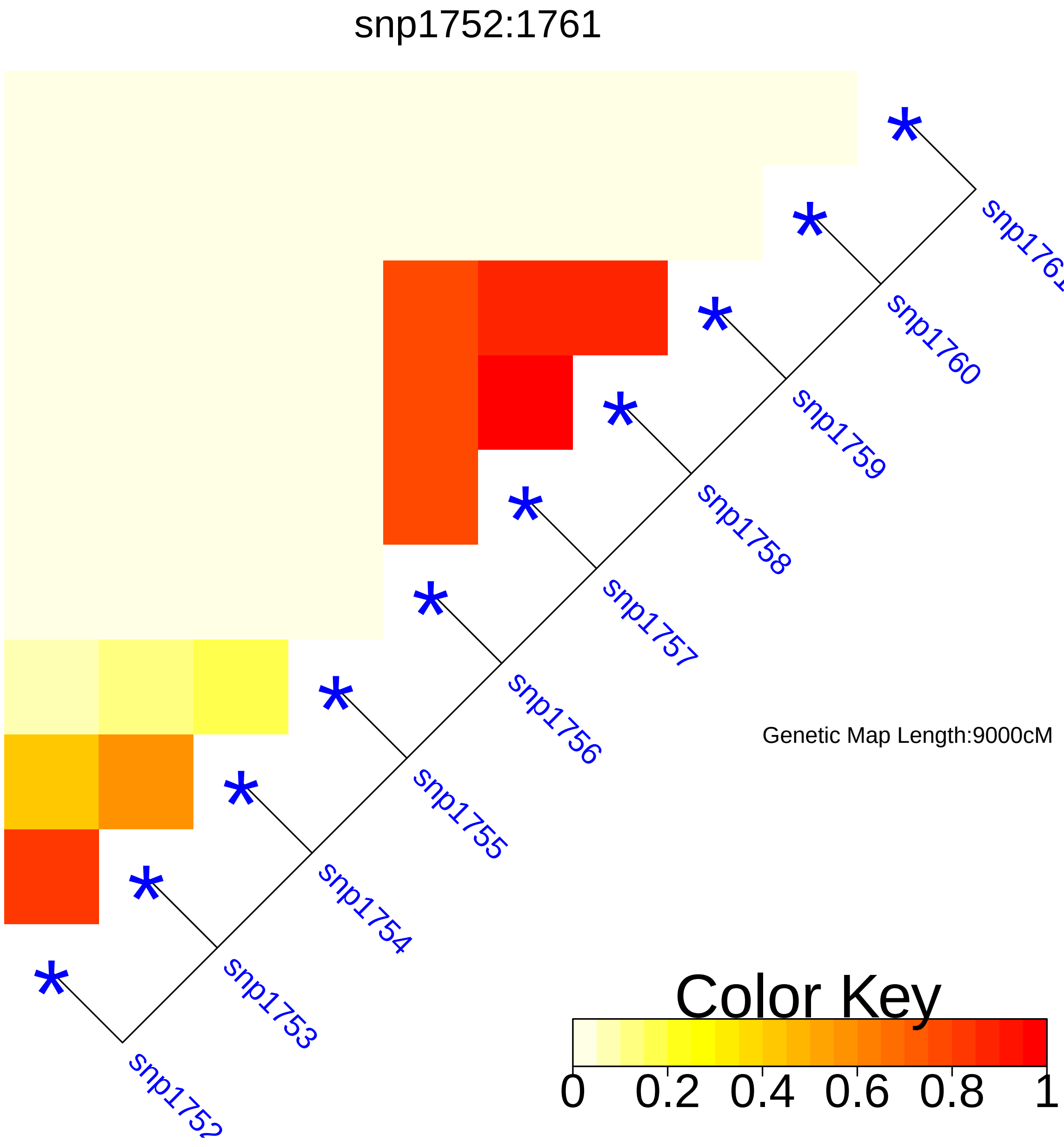
- The 1000 Genomes Project (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061-73.

- Data (08/2010): 283 Europeans; 174 Africans (AFR)
- RVs: defined here with MAF 1-5%;
  CVs: defined here with MAF $> 5\%$
- Chr 1:
  EUR: 894,828 SNVs; AFR: 1,279,571 SNVs;
  Common: 694,329 SNVs; 146,378 RVs; 478,241 CVs; 69,710 others
- MAF distributions:
  EUR: (Q1, Q2, Q3)=(.0053, .0424, .2014)
  AFR: (Q1, Q2, Q3)=(.0115, .0431, .1609)
  PCs based on CVs or RVs:


○ EUR
△ AFR
2nd PC
1st PC
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30



EUR
AFR
5
0
−5
−10
−15
y5[1:283, 2]
−40
−30
−20
−10
0
y5[1:283, 1]


snp1752:1761
snp1752
snp1753
snp1754
snp1755
snp1756
snp1757
snp1758
snp1759
snp1760
snp1761
Genetic Map Length:9000cM
Color Key
0
0.2
0.4
0.6
0.8
1

- Population stratification:
  Spurious disease-RV association due to race/ethnic groups as confounders;
- Many methods proposed for GWAS.
  Use PC's to adjust;
- Example: randomly drawn from the sample data,
  "Cases": 90% Europeans + 10% Africans;
  Controls: 10% Europeans + 90% Africans;

Type I errors at $\alpha = 0.05$:

Tests	No PC	1 PC	5 PCs	10 PCs
UminP	.417	.069	.069	.075
Score	.812	.089	.079	.081
Sum	.899	.046	.044	.052
SSU	.057	.057	.054	.061

Power at $\alpha = 0.05$: randomly chose 4 causal SNPs

Tests	No PC	1 PC	5 PCs	10 PCs
	$\log \text{OR} \sim U(-\log 4, \log 4)$			
UminP	.377	.381	.380	.389
Score	.359	.357	.357	.362
Sum	.295	.289	.291	.300
SSU	**.421**	.422	.422	.431
	$\log \text{OR} \sim U(0, \log 4)$			
UminP	**.719**	.717	.721	.725
Score	.678	.665	.667	.666
Sum	.659	.652	.654	.657
SSU	.686	.683	.684	.687

Power at $\alpha = 0.05$: 10 causal SNPs

Tests	No PC	1 PC	5 PCs	10 PCs
	$\log$ OR $\sim U(-\log 3, \log 3)$			
UminP	.582	.581	.578	.582
Score	.629	.623	.623	.634
Sum	.380	.383	.385	.388
SSU	**.633**	.638	.639	.651
	$\log$ OR $\sim U(0, \log 1.5)$			
UminP	.462	.460	.464	.466
Score	.408	.405	.412	.413
Sum	**.617**	.612	.619	.616
SSU	.536	.530	.533	.525

Discussion

- Pooled association (burden) tests perform well only if 1) no opposite association directions and 2) no or few non-associated RVs.
  Not likely!
- SSU test in general is powerful.
  But may lose power with too many non-associated RVs.
- No test is uniformly most powerful!
  The identity (or construction) of a more powerful test depends on the unknown truth (of the association pattern).
- Adaptive tests are needed!
- An **exciting** topic!

- Penalized regression?
  Disease-CV asscoiation testing: Basu et al (2011, *Genet Epi*);
  Phenotype prediction: high-dim; but sparse models?

Acknowledgement: This research was supported by NIH and a UM-IHI Seed Grant.

My collaborators: Dr Xiaotong Shen, Dr Saonli Basu, Dr Weihua Guan, Yiwei Zhang and other students.

You can download our papers from

http://www.sph.umn.edu/biostatistics/research/reports.asp

Thank you!