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:
  and Pan (2011, *Genet Epi*), Pan and Shen (2011, *Genet Epi*), ...
Introduction

• Single Nucleotide Polymorphism (SNP) or Variant (SNV)
  DNA seq 1 – AAGC\textbf{C}TA
  DNA seq 2 – AAGC\textbf{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 \textit{causal} CVs.

• GWAS: a success!?
  As of \textbf{10/5/11} (or 01/19/11 or 09/24/09 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:

<table>
<thead>
<tr>
<th>Obs</th>
<th>Y</th>
<th>SNP1</th>
<th>SNP2</th>
<th>SNP3</th>
<th>...</th>
<th>SNPk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>CT</td>
<td>AG</td>
<td>CG</td>
<td>...</td>
<td>AC</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>TT</td>
<td>AG</td>
<td>GG</td>
<td>...</td>
<td>AA</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>CT</td>
<td>AA</td>
<td>CG</td>
<td>...</td>
<td>CC</td>
</tr>
</tbody>
</table>

......

| 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 \( \implies \) univariate tests ...
  \( n = 1000, \text{MAF}=1\% \implies \#(\text{minor alleles}) \approx 20; \)
  \( n = 1000, \text{MAF}=0.1\% \implies \#(\text{minor alleles}) \approx 2; \)
  Design matrix \( X \): almost all 0’s!

• RVs: small MAF \( \implies \) aggregation!
  combine multiple RVs!
Existing methods

• Single-locus (or SNP-by-SNP or univariate) analysis: GWAS
  - Model: \( Y \sim SNP_j \)
    \[
    \text{Logit } \Pr(Y_i = 1) = \beta_{M,0j} + X_{ij}\beta_{M,j}, \quad (1)
    \]
  - \( H_{0,j}: \beta_{M,j} = 0 \) for each \( j = 1, ..., k \)
    \[\Rightarrow p_j.\]
  - Combining: \( \text{U}\text{min}P = \min(p_1, p_2, ..., p_k) \) or ... 
    Need to do multiple test adjustment!
  - Model (1): as a \( 2 \times 3 \) table; Cochran-Armitage trend test.
• Multivariate (or global or joint) analysis:
  – Model: $Y \sim SNP_1 + ... + SNP_k$

    $\text{Logit Pr}(Y_i = 1) = \beta_0 + \sum_{j=1}^{k} X_{ij} \beta_j,$ \hspace{1cm} (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: \( \beta_1 = \ldots = \beta_k = \beta_c \).
  – Model:
    \[
    \text{Logit } \Pr(Y_i = 1) = \beta_{0,c} + \sum_{j=1}^{k} X_{ij} \beta_c = \beta_{0,c} + X_{i,c} \beta_c, \tag{3}
    \]
  – \( H_{0,c} : \beta_c = 0 \)
  – Apply the score, Wald or LR test
    \( T_W = \frac{\hat{\beta}_c^2}{V_c} \sim \chi^2_1 \) under \( H_{0,c} \).
  – Feature: DF=1; no multiple testing!
  – Correct test size:
    \( H_0 \implies H_{0,c}! \)
  – Closely related to CMC (Li and Leal 2008), weighted sum
    (Madsen and Browning 2009) tests:
    \( \bigvee_{j=1}^{k} X_{ij} \approx \sum_{j=1}^{k} X_{ij} \)
**Power:** OR=(2, 2, 2, 2, 2, 2, 2, 2); No LD; \( n = 500 + 500; \) MAFs ~ \( U(.001, .01) \) for controls;

<table>
<thead>
<tr>
<th>Test</th>
<th># of neutral RVs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>UminP</td>
<td>.441</td>
</tr>
<tr>
<td>Score</td>
<td>.746</td>
</tr>
<tr>
<td>CMC</td>
<td>.938</td>
</tr>
<tr>
<td>wSum</td>
<td>.940</td>
</tr>
<tr>
<td>Sum</td>
<td>.951</td>
</tr>
<tr>
<td>aSum</td>
<td>.933</td>
</tr>
<tr>
<td>SSU</td>
<td>.756</td>
</tr>
<tr>
<td>KMR(Linear)</td>
<td>.762</td>
</tr>
<tr>
<td>C-alpha</td>
<td>.771</td>
</tr>
</tbody>
</table>
**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;

<table>
<thead>
<tr>
<th>Test</th>
<th># of neutral RVs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>UminP</td>
<td>.607</td>
</tr>
<tr>
<td>Score</td>
<td>.869</td>
</tr>
<tr>
<td>CMC</td>
<td>.661</td>
</tr>
<tr>
<td>wSum</td>
<td>.659</td>
</tr>
<tr>
<td>Sum</td>
<td>.682</td>
</tr>
<tr>
<td>aSum</td>
<td>.854</td>
</tr>
<tr>
<td>SSU</td>
<td>.895</td>
</tr>
<tr>
<td>KMR</td>
<td>.897</td>
</tr>
<tr>
<td>C-alpha</td>
<td>.906</td>
</tr>
</tbody>
</table>
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

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

  4) $Pr(SSU > s | H_0) \approx Pr(\chi_d^2 > (s - b)/a)$. 
• 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 $\beta = (\beta_1, \ldots, \beta_k)'$ random:
    $E(\beta) = 0$, $Cov(\beta) = \tau^2 I$.
  
  – Test $H_{0,\tau^2}$: $\tau^2 = 0$ by a score test.
  
  – For logistic regression:
    
    $T_{Go} = \frac{1}{2}(U'U - \text{Trace}(I_F))$, where $U = X'(Y - \bar{Y})$,
    
    and $I_f = Cov(U) = \bar{Y}(1 - \bar{Y})(X - \bar{X})'(X - \bar{X})$.

    
    $T_{Go} = \frac{1}{2}(Y - \bar{Y})'XX'(Y - \bar{Y}) -$

    
    $\frac{1}{2}\bar{Y}(1 - \bar{Y})\text{Trace}((X - \bar{X})'(X - \bar{X}))$,

    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 \( \beta \), then
    \[ T_{MP} = \beta'U, \text{ but } ... \]
  * Try \( \max_b b'U \) s.t. \( Var(b'U) = b'\mathbf{I}_F b = 1 \)?

- We estimate \( T_{MP} \) by
  \[ T_{EMP} = \hat{\beta}_M' U. \]
- $T_{EMP} \approx SSUw = U' \text{Diag}(I_F)^{-1}U$ because

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

- How about estimating $\beta$ 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

\[
\text{Logit Pr}(Y_i = 1) = \beta_0 + h(X_{i1}, ..., X_{ik}),
\]

(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 $\gamma_1, ..., \gamma_n$.

– 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.

\[ F = \frac{tr(\hat{Y}'\hat{Y})}{tr(R'R)} = \frac{tr(\hat{Y}\hat{Y}')} {tr(\hat{R}R')} = \frac{tr(HYY'H)}{tr((I-H)YY'(I-H))} \]

\[ = \frac{tr(HGH)}{tr((I-H)G(I-H))} \propto SSU \]

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 $\propto 1/\text{MAF}$; 
  2) functional prediction, e.g. by SIFT,...
### Power: $OR = (3, 1/3, 2, 2, 2, 1/2, 1/2, 1/2); \text{ with LD.}$

<table>
<thead>
<tr>
<th>Tests</th>
<th># of neutral RVs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>UminP</td>
<td>.489</td>
</tr>
<tr>
<td>Score</td>
<td>.599</td>
</tr>
<tr>
<td>CMC</td>
<td>.365</td>
</tr>
<tr>
<td>wSum</td>
<td>.369</td>
</tr>
<tr>
<td>Sum</td>
<td>.342</td>
</tr>
<tr>
<td>aSum</td>
<td>.350</td>
</tr>
<tr>
<td>SSU</td>
<td>.603</td>
</tr>
<tr>
<td>KMR</td>
<td>.611</td>
</tr>
<tr>
<td>C-alpha-P</td>
<td><strong>.629</strong></td>
</tr>
</tbody>
</table>
**Power**: only one causal RV with OR=5:

<table>
<thead>
<tr>
<th>Test</th>
<th># of neutral RVs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
</tr>
<tr>
<td>UminP</td>
<td>.696</td>
</tr>
<tr>
<td>Sum</td>
<td>.365</td>
</tr>
<tr>
<td>aSum</td>
<td>.447</td>
</tr>
<tr>
<td>KBAC</td>
<td>.629</td>
</tr>
<tr>
<td>PWST</td>
<td>.665</td>
</tr>
<tr>
<td>EREC</td>
<td>.685</td>
</tr>
<tr>
<td>SSU</td>
<td>.710</td>
</tr>
<tr>
<td>aSSU</td>
<td><strong>.736</strong></td>
</tr>
<tr>
<td>aSPU</td>
<td>.707</td>
</tr>
</tbody>
</table>
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.”

• 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:
26-1
• 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 $\alpha = 0.05$:

<table>
<thead>
<tr>
<th>Tests</th>
<th>No PC</th>
<th>1 PC</th>
<th>5 PCs</th>
<th>10 PCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>UminP</td>
<td>.417</td>
<td>.069</td>
<td>.069</td>
<td>.075</td>
</tr>
<tr>
<td>Score</td>
<td>.812</td>
<td>.089</td>
<td>.079</td>
<td>.081</td>
</tr>
<tr>
<td>Sum</td>
<td>.899</td>
<td>.046</td>
<td>.044</td>
<td>.052</td>
</tr>
<tr>
<td>SSU</td>
<td>.057</td>
<td>.057</td>
<td>.054</td>
<td>.061</td>
</tr>
</tbody>
</table>
Power at $\alpha = 0.05$: randomly chose 4 causal SNPs

<table>
<thead>
<tr>
<th>Tests</th>
<th>No PC</th>
<th>1 PC</th>
<th>5 PCs</th>
<th>10 PCs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>log $\text{OR} \sim U(-\log 4, \log 4)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U_{\text{minP}}$</td>
<td>.377</td>
<td>.381</td>
<td>.380</td>
<td>.389</td>
</tr>
<tr>
<td>Score</td>
<td>.359</td>
<td>.357</td>
<td>.357</td>
<td>.362</td>
</tr>
<tr>
<td>Sum</td>
<td>.295</td>
<td>.289</td>
<td>.291</td>
<td>.300</td>
</tr>
<tr>
<td>SSU</td>
<td><strong>.421</strong></td>
<td>.422</td>
<td>.422</td>
<td>.431</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>log $\text{OR} \sim U(0, \log 4)$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_{\text{minP}}$</td>
<td><strong>.719</strong></td>
<td>.717</td>
<td>.721</td>
<td>.725</td>
</tr>
<tr>
<td>Score</td>
<td>.678</td>
<td>.665</td>
<td>.667</td>
<td>.666</td>
</tr>
<tr>
<td>Sum</td>
<td>.659</td>
<td>.652</td>
<td>.654</td>
<td>.657</td>
</tr>
<tr>
<td>SSU</td>
<td>.686</td>
<td>.683</td>
<td>.684</td>
<td>.687</td>
</tr>
</tbody>
</table>
Power at $\alpha = 0.05$: 10 causal SNPs

<table>
<thead>
<tr>
<th>Tests</th>
<th>No PC</th>
<th>1 PC</th>
<th>5 PCs</th>
<th>10 PCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>log OR $\sim U(-\log 3, \log 3)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UminP</td>
<td>.582</td>
<td>.581</td>
<td>.578</td>
<td>.582</td>
</tr>
<tr>
<td>Score</td>
<td>.629</td>
<td>.623</td>
<td>.623</td>
<td>.634</td>
</tr>
<tr>
<td>Sum</td>
<td>.380</td>
<td>.383</td>
<td>.385</td>
<td>.388</td>
</tr>
<tr>
<td>SSU</td>
<td><strong>.633</strong></td>
<td>.638</td>
<td>.639</td>
<td>.651</td>
</tr>
</tbody>
</table>

| 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 association 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!