Biased urn permutation: Establish the significance of rare-variant association tests in the presence of confounding factors

Yuan Lin
Post-doc Research Associate
@ Dr. Kirk Wilhelmsen's Lab
Department of Genetics, School of Medicine
UNC at Chapel Hill

Permutation test for case-control GWAS

- Commonly used to establish statistical significance of GWAS results (e.g., disease-associated SNPs), while accounting for the correlation between multiple tests (e.g., LD among SNPs)
- The key is a **resampling-without-replacement** procedure in which the disease labels (case/control) are rearranged to disrupt genotype-phenotype association in the original data
- This procedure is repeated certain times to generate multiple permuted data sets.

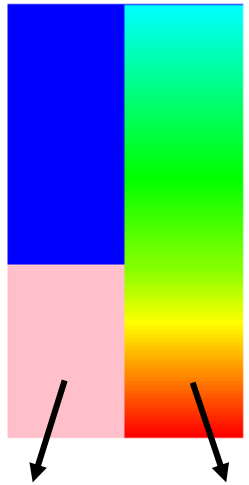
Permutation test for case-control GWAS

- Given an association test statistic *T*, each GWAS result (e.g., a SNP) has a *T_origin* calculated from the original data and a set of *T_perm* each calculated from a permuted data set.
- The set of *T_perm* is supposed to form an empirical distribution of *T* under the null hypothesis (i.e., no association).
- The position of *T_origin* in this null distribution gives us a P value indicating how significant the original GWAS result is.
  - We need at least *K* permuted data sets to have a significance level of 1/*K*

Permutation test for case-control GWAS

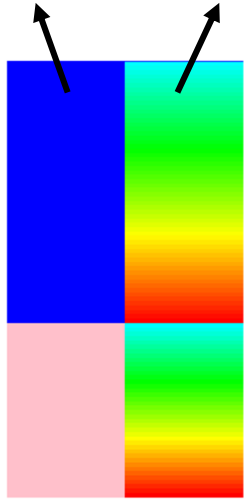
- We want the permutation procedure to maintain “other correlations” in the original data, especially confounding factors such as population stratification.
- Not accounting for confounders could lead to spurious or distorted associations, which is a more serious problem for rare-variant association tests
- The rearrangement of disease labels is completely random in the often used **random permutations**, which destroy confounding effects in the original data

Data set 1
(with some
confounding
effect)

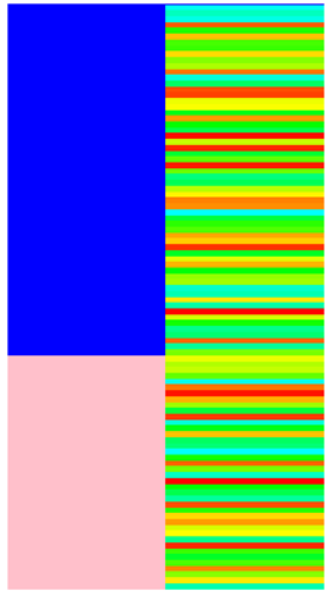


Confounder *Outcome*

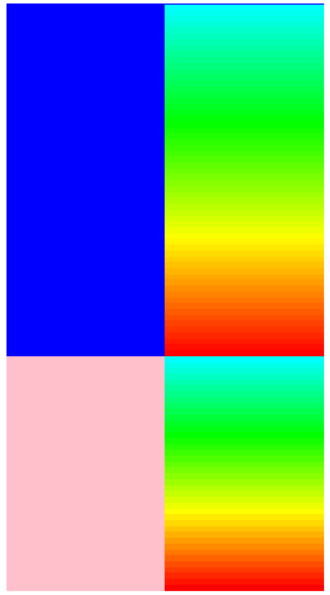
Data set 2
(no such
confounding
effect)



A permuted data set obtained by randomly shuffling the outcomes



The same permuted data set with reordered outcomes



Biased urn permutation

- (Epstein et al., 2012)
- Maintain the confounding structure in the original data
- Allow for an arbitrary number of categorical and continuous covariates
- Among all covariates, only actual confounders would be controlled for
- Preserve the numbers of cases and controls in permuted data sets

Biased urn permutation

- The major part is resampling from a multivariate Fisher's noncentral hypergeometric distribution (referred to as **biased urn sampling**)
- The mfnchypg distribution is parameterized on subject-specific odds ratio of disease (θ_j) given covariates C_j
- The value of θ_j is estimated from a logistic regression model

$$\log\left(\frac{P[D_j = 1 \mid \boldsymbol{C}_j]}{P[D_j = 0 \mid \boldsymbol{C}_j]}\right) \equiv \log(\theta_j) = \alpha + \boldsymbol{\gamma}^{\mathrm{T}} \cdot \boldsymbol{C}_j \qquad \widehat{\theta}_j = \exp(\widehat{\alpha} + \widehat{\boldsymbol{\gamma}}^{\mathrm{T}} \cdot \boldsymbol{C}_j)$$

Biased urn sampling

- Imagine we randomly draw some marbles (without replacement) from an urn that contains marbles of different colors.
- Hypergeometric distribution describes the probability of observing a specific color permutation, given that marbles have equal chances of being drawn, no matter what their colors are (an **unbiased** urn).
- Noncentral hypergeometric distribution describes the same probability, but assuming marbles have unequal chances of being drawn, (only) due to their colors (a **biased** urn).

Biased urn sampling

- Given the number of different colors in the urn, the nchypg distribution is either **univariate** (2 colors) or **multivariate** (> 2 colors)
- If the marbles are drawn one by one, the observed color permutation is governed by **Wallenius'** nchypg
- If the draws are independent of one another, the observed color permutation is governed by **Fisher's** nchypg

Biased urn sampling

- N: total number of marbles in the urn
- c : total number of colors in the urn
- m_i : the number of marbles with color $i \in \{1, \cdots, c\}$ in the urn
- n : the number of extracted marbles
- x_i : the number of extracted marbles with color i
- ω_i : the odds ratio of drawing a marble with color i

Univariate Fisher's nchypg

$$m_1, m_2 \in \mathbb{N}$$
$$N = m_1 + m_2$$
$$n \in [0, N)$$
$$\omega \in \mathbb{R}_+$$

$$x \in [x_{\min}, x_{\max}]$$
$$x_{\min} = \max(0, n - m_2)$$
$$x_{\max} = \min(n, m_1)$$

$$\frac{\binom{m_1}{x}\binom{m_2}{n-x}\omega^x}{P_0}$$

where $P_0 = \sum_{y=x_{\min}}^{x_{\max}} \binom{m_1}{y}\binom{m_2}{n-y}\omega^y$

Multivariate Fisher's nchypg

$$c \in \mathbb{N}$$
$$\mathbf{m} = (m_1, \ldots, m_c) \in \mathbb{N}^c$$
$$N = \sum_{i=1}^{c} m_i$$
$$n \in [0, N)$$
$$\boldsymbol{\omega} = (\omega_1, \ldots, \omega_c) \in \mathbb{R}_+^c$$

$$\mathrm{S} = \left\{ \mathbf{x} \in \mathbb{Z}_{0+}^c \ : \ \sum_{i=1}^{c} x_i = n \right\}$$

$$\frac{1}{P_0} \prod_{i=1}^{c} \binom{m_i}{x_i} \omega_i^{x_i}$$

where $P_0 = \sum_{(y_0, \ldots, y_c) \in \mathrm{S}} \prod_{i=1}^{c} \binom{m_i}{y_i} \omega_i^{y_i}$

https://en.wikipedia.org/wiki/Fisher's_noncentral_hypergeometric_distribution

Biased-urn permutation

- Permuted data sets are generated by sampling from a biased urn with the following features (i.e., a **mfnchypg** distribution with the following parameter settings)
- Each marble in the urn has a unique color and each color only has one marble, i.e., $m_j = 1$ for all $j \in \{1, \cdots, c\}$
- Each marble/color j represents Subject j. The total number of marbles/colors c equals to sample size N

Biased-urn permutation

- One drawing experiment yields a permuted data set $k \in \{1, \cdots, K\}$
- Each marble j is either selected or not, i.e., $r_{kj} \in \{0,1\}$. Accordingly, Subject j is assigned as case or control in the data set being generated
- The odds ratio of Subject j being assigned as a case against a control (θ_j) amounts to the odds ratio of drawing a specific marble/color j
- The number of marbles to be drawn, n, equals to the number of cases N_1 in the original data set

Result 1

- **Biased urn permutations can preserve confounding structure in the original data while random permutations cannot**
- Test on a real GWAS data set of African American subjects with schizophrenia (907 cases and 937 controls)
- Confounder variables are 8 top eigenvectors ($\gamma_1, \cdots, \gamma_8$) obtained from PCA on the data set using 41,182 SNPs in approximate linkage equilibrium

Table 2. Regression Coefficient Estimates Under Biased Urn and Random Permutation Schemes

		Permutation Scheme	
		Biased Urn	Random
	Original Data	Mean (SD)	Mean (SD)
γ_1	−8.39	−8.47 (2.02)	−0.06 (2.03)
γ_2	1.41	1.44 (2.14)	−0.08 (2.03)
γ_3	−2.13	−2.28 (2.00)	−0.11 (2.04)
γ_4	−4.86	−4.96 (2.05)	−0.09 (2.03)
γ_5	−0.88	−0.93 (2.02)	−0.08 (2.02)
γ_6	0.69	0.80 (2.09)	0.01 (2.03)
γ_7	−1.22	−1.24 (2.01)	0.00 (2.04)
γ_8	−0.76	−0.80 (1.99)	0.03 (1.96)

The results for each permutation scheme are based on 1,000 permutations of the data set. The following abbreviation is used: SD, standard deviation.

- Conduct 1000 random and 1000 biased urn permutations on the original data
- Fit a logistic regression model of disease status given the covariates to the original data set and each permuted data set
- Compare regression coefficient estimates of the original-data model with those of the permuted-data model (average over 1000)

Result 2

- **Biased urn permutations reduced type I errors without jeopardizing power when correcting for confounders**
- Applied to three rare-variant association tests that use permutations to establish significance of test statistics
  - CMAT (Zawistowski et al., 2010): burden test; fixed weights
  - RBT (Ionita-Laza et al., 2011): burden test; adaptive weights
  - C-alpha (Neale et al., 2011): variance-component test
- Test on simulated resequencing data sets subjected to confounding from population stratification and real sequencing data

Simulated data for type-I error comparison

- Use *cosi* to simulate a large number of haplotypes of a 10~100kb sequence with European or African ancestry.
- Randomly pairing these haplotypes (or further obtained admixture haplotypes) to simulate diplotypes (i.e., individual subjects)
- Determine the disease status of each simulated subject based on its average percentage of African ancestry across the simulated region
  - Intentionally incur inflation of type I error due to population stratification
- Create 10,000 GWAS data sets each with 300 cases and 300 controls
- Use top (?) PCA eigenvectors as covariates

Simulated data for type-I error comparison

As *cosi-generated* haplotypes are too short and do not contain enough SNPs, additional genotypes were simulated for each subject for PCA

1. Select ≥ 10,000 SNPs from HapMap that show marked allele-frequency differences between HapMap YRI and CEU samples
2. Filter out SNPs in strong LD ($r^2 \geq 0.5$) using PLINK LD pruning
3. Simulate genotypes at the remaining SNPs
4. Given a simulated GWAS data set, randomly select simulated genotypes for each subject based on her ancestry

Calculation of type-I error rate

- For each simulated data set, apply all three rare-variant association tests and establish the significance of association via 5000 random and 5000 biased urn permutations
- Type-I error rate was calculated as the proportion of all 10,000 simulated data sets that showed a significant association ($P<.05$ or .005).

Table 3. Type-I Error Results Under Confounding for 10 kb Regions

Test	Odds Ratio of Disease (YRI versus CEU)	$\alpha = 0.05$		$\alpha = 0.005$	
		Biased Urn	Random	Biased Urn	Random
CMAT	1	0.0521	0.0511	0.0046	0.0045
	2	0.0450	0.0850	0.0047	0.0123
	4	0.0485	0.1607	0.0053	0.0503
	8	0.0551	0.2366	0.0058	0.1004
RBT	1	0.0469	0.0468	0.0043	0.0042
	2	0.0487	0.0591	0.0045	0.0066
	4	0.0501	0.0962	0.0055	0.0169
	8	0.0546	0.1994	0.0055	0.0463
C-alpha	1	0.0491	0.0542	0.0043	0.0051
	2	0.0460	0.1712	0.0049	0.0364
	4	0.0453	0.4890	0.0042	0.2251
	8	0.0527	0.7603	0.0055	0.5011

Table 4. Type-I Error Results Under Confounding for 100 kb Regions

Test	Odds Ratio of Disease (YRI versus CEU)	$\alpha = 0.05$		$\alpha = 0.005$	
		Biased Urn	Random	Biased Urn	Random
CMAT	1	0.0466	0.0482	0.0045	0.0048
	2	0.0474	0.1040	0.0048	0.0192
	4	0.0480	0.2308	0.0050	0.0868
	8	0.0544	0.3035	0.0052	0.1439
RBT	1	0.0477	0.0497	0.0048	0.0053
	2	0.0445	0.0691	0.0046	0.0080
	4	0.0461	0.1463	0.0042	0.0308
	8	0.0515	0.3986	0.0057	0.1406
C-alpha	1	0.0440	0.0501	0.0044	0.0049
	2	0.0402	0.2834	0.0040	0.0727
	4	0.0410	0.7962	0.0050	0.5049
	8	0.0422	0.9771	0.0038	0.8729

A more fair comparison is between (a) biased Urn permutation which use logistic regression to get the odds ratio parameters and (b) first use linear regression to regress out some confounding effects and then use random permutation (after linear regression the original binary phenotype will become a continuous variable but it is OK, just permute the “adjusted” phenotype values among individual subjects

Simulated data for power comparison

- Simulate individual subjects the same way as mentioned earlier
- For a simulated region (10kb?), randomly select 10% of variants within the region that have MAF < 0.01 as causal variants
- Determine the disease status of each subject based on its simulated genotypes on these causal variants
- Create 1000 GWAS data sets each with 300 cases and 300 controls
- Make sure no confounding effect from population stratification

Simulated data for power comparison

- "Causal variants had identical relative risk and independently increased disease risk under a log-additive model" (?)
- For each simulated subject j , its odds ratio of disease is given by

$$log\big(P\big(D_j = 1\big)/P\big(D_j = 0\big)\big) = \alpha + \beta^T \boldsymbol{G}_j$$

- $D_j \in \{0,1\}$: disease status
- $\boldsymbol{G}_j \in \{0,1,2\}^m$: genotypes on m disease-causal variants
- $\alpha = \log\big(0.01/(1-0.01)\big)$, with 0.01 being disease prevalence
- $\beta = \{\log(RR)\}^m$, with RR being the relative risk of every causal variant

Table 5. Power Results for 10 kb Regions

Test	Permutation Scheme	Relative Risk of Rare Variant 1.5	2.0	2.5
CMAT	Biased urn	0.135	0.244	0.290
	Random	0.141	0.241	0.289
RBT	Biased urn	0.144	0.263	0.373
	Random	0.144	0.273	0.383
C-alpha	Biased urn	0.267	0.552	0.735
	Random	0.279	0.572	0.754

- For each simulated data set, apply all three rare-variant association tests and establish significance via 5000 random and 5000 biased urn permutations respectively
- P value threshold set to 0.05 (after all adjustment?)
- Results were averaged over 1000 replicated GWAS data sets

Real data: Dallas Heart Study

- Select subjects in the top and bottom 20% of the outcome distribution to mimic case-control study design
  - 500+ cases and 500+ controls for each phenotype
- Use the same three rare-variant association tests as mentioned
- Establish statistical significance via 10,000 random and 10,000 biased urn permutations respectively
- Biased urn permutations adjusted for potential confounding effects of age, gender, and ethnicity

Table 6. CMAT Analysis of Sequence Data from the Dallas Heart Study

		p Value of CMAT	
Trait	**Gene**	**Random Permutations**	**Biased Urn Permutations**
Triglycerides	*ANGPTL3*	<0.0001	0.0141
	ANGPTL4	<0.0001	0.0015
	ANGPTL5	0.0201	0.0974
BMI	*ANGPTL3*	0.5930	0.7418
	ANGPTL4	0.6984	0.7058
	ANGPTL5	0.0077	0.0301

Table 7. RBT Analysis of Sequence Data from the Dallas Heart Study

		p Value of RBT	
Trait	**Gene**	**Random Permutations**	**Biased Urn Permutations**
Triglycerides	*ANGPTL3*	0.0006	0.0126
	ANGPTL4	<0.0001	0.0034
	ANGPTL5	0.0231	0.1102
BMI	*ANGPTL3*	0.5174	0.6890
	ANGPTL4	0.9180	0.9348
	ANGPTL5	0.0046	0.0170

Table 8. C-Alpha Analysis of Sequence Data from the Dallas Heart Study

		p Value of C-Alpha Test	
Trait	**Gene**	**Random Permutations**	**Biased Urn Permutations**
Triglycerides	*ANGPTL3*	<0.0001	0.0010
	ANGPTL4	0.0001	0.0363
	ANGPTL5	0.1572	0.2043
BMI	*ANGPTL3*	0.9168	0.9814
	ANGPTL4	0.8314	0.8472
	ANGPTL5	0.2310	0.2872

Result 3

- Biased urn resampling provides a way to construct confidence intervals for the estimate of rare-variant risk effect
- **The 95% confidence intervals constructed using the resampling-based method had appropriate coverage and were smaller in magnitude than the corresponding asymptotic 95% confidence intervals**
- Test on 5000 GWAS sets each with 300 cases and 300 controls

Simulated data for CI comparison

- In each simulated data set, a subject j's odds ratio of disease is given by (?)

$$log\big(P\big(D_j = 1\big)/P\big(D_j = 0\big)\big) = \alpha + \beta g_j + \gamma \cdot I_j(African)$$

- D_j : disease status
- $g_j \in \{0,1,2\}$: minor allele counts at a predefined, single, uncommon risk locus (MAF = 2%)
- $I_j(African)$: whether the subject has African ancestry
- $\beta = 1$: effect size of the risk variant
- $\alpha = log(0.01/(1 - 0.01))$, with 0.01 as disease prevalence
- $\gamma = 4$: confounding effect of population structure

Calculation of confidence intervals

1. For each simulated GWAS data set, fit a logistic regression model

$$\log\left(\frac{P\left[D_j = 1 \mid G_j, \boldsymbol{C}_j\right]}{P\left[D_j = 0 \mid G_j, \boldsymbol{C}_j\right]}\right) \equiv \log(\theta_j) = \alpha + \beta \cdot G_j + \boldsymbol{\gamma}^{\mathrm{T}} \cdot \boldsymbol{C}_j$$

2. Get the maximum-likelihood estimates of coefficient β and θ_j (denoted $\hat{\beta}$ and $\widehat{\theta_j}$) and calculate an asymptotic 95% confidence interval for $\hat{\beta}$ based on the fitted model
3. Generate K = 10,000 permuted data sets using biased urn sampling parameterized on $\widehat{\theta_j}$
4. Refit the above model in every permuted data set and collect the permuted-data-based estimate of β, denoted $\widehat{\beta_r}$ where $r \in \{1, \cdots, K\}$
5. Calculate the 95% confidence interval for $\hat{\beta}$ using the distribution of $\widehat{\beta_r}$

Application to CHAT

- In the current version of CHAT, the significance of clusters is established by random permutations. There is no mechanism to adjust for covariates.
- A possible solution is to convert to a regression framework by coding cluster membership as pseudo genotype and then applying a regression-based rare-variant association test such as SKAT
  - However, SKAT requires weighting different variants in the same region. In our case, estimating the weights of clusters is difficult
  - In addition, by projecting association results back to the genome scale we may lose the edge of CHAT – sensitivity to local haplotype similarity
- Alternatively, we could replace the random permutation test in CHAT with the biased urn permutation test

Thank you!
Questions?