A general framework for the region-based analysis of rare variants data in family-based association studies

Julian Hecker, Christoph Lange

08/08/2018

Introduction

setting

- genomic region of p SNPs, either only rare variants or combination rare/common
- test region for association with phenotype

existing approaches

- population-based study designs:

  e.g. CMC, SKAT (Li and Leal 2008, Wu et al. 2011)
- family-based study designs:

  e.g. rare-variant GDT, rare-variant FBAT, FB-SKAT (He et al. 2017, De et al. 2013, Ionita-Laza et al. 2013)

Region-based family-based association testing

- advantage of family-based settings: allows to construct association tests that are robust against population stratification
- base of transmission-based approaches as TDT/FBAT
- multiple variants: empirical estimates of correlation, asymptotic theory problematic

Region-based family-based association testing

- advantage of family-based settings: allows to construct association tests that are robust against population stratification

- base of transmission-based approaches as TDT/FBAT

- multiple variants: empirical estimates of correlation, asymptotic theory problematic

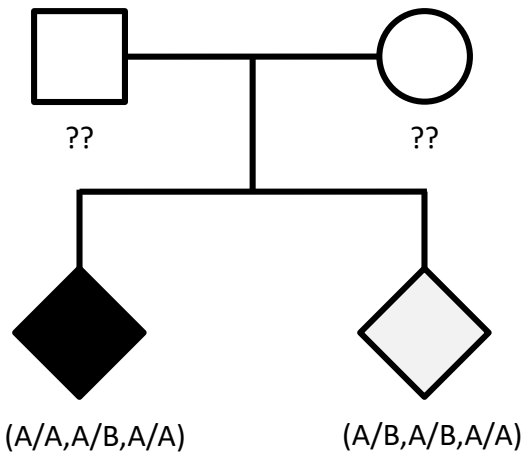
→ propose our general framework for region-based association analysis in family-based association studies

Framework

1. conditional offspring genotype distribution for nuclear family
2. construction of suitable region-based association test statistics
3. evaluation of significance

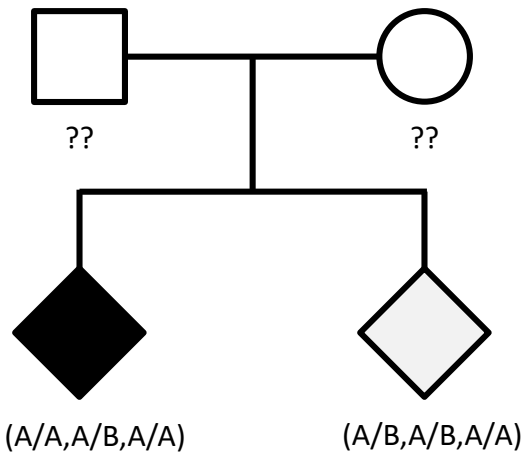
FBAT-haplotype algorithm

- genomic region with p tightly linked markers
- nuclear family i, parental genotypes may be missing, observed offspring genotypes X_i, phenotypes T_i
- FBAT-haplotype algorithm utilizes sufficient statistic approach (Laird and Rabinowitz 2000, Horvath et al. 2004)
- output: $X \mid S_i$, joint offspring genotype distribution given sufficient statistic S_i



FBAT-haplotype algorithm: details

- requires construction of all possible parental mating types for given offspring genotypes
- comparison of likelihood ratios along parental mating types
- number of potential phased parental mating types can be very large



FBAT-haplotype algorithm: improvement

- identify set h_{off} of all haplotypes that are compatible with obsered offspring genotypes
- instead of constructing all possible parental mating types, use only h_{off} haplotypes

Received: 22 August 2017 | Revised: 29 September 2017 | Accepted: 10 October 2017
DOI: 10.1002/gepi.22094

BRIEF COMMUNICATION

WILEY Genetic Epidemiology
OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

Family-based tests for associating haplotypes with general phenotype data

Improving the FBAT-haplotype algorithm

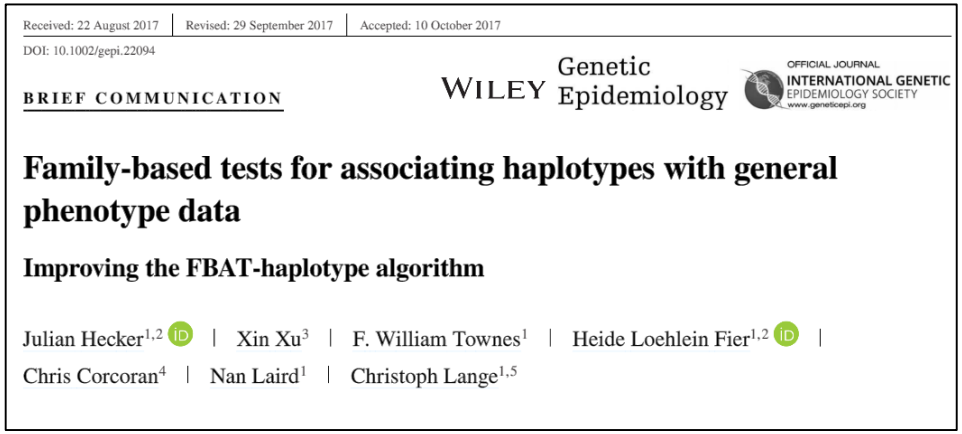
Julian Hecker[1,2] | Xin Xu[3] | F. William Townes[1] | Heide Loehlein Fier[1,2] | Chris Corcoran[4] | Nan Laird[1] | Christoph Lange[1,5]

FBAT-haplotype algorithm: improvement

- identify set h_{off} of all haplotypes that are compatible with obsered offspring genotypes

- instead of constructing all possible parental mating types, use only h_{off} haplotypes

→ output maintained



Received: 22 August 2017 | Revised: 29 September 2017 | Accepted: 10 October 2017
DOI: 10.1002/gepi.22094

BRIEF COMMUNICATION

WILEY Genetic Epidemiology
OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

Family-based tests for associating haplotypes with general phenotype data

Improving the FBAT-haplotype algorithm

Julian Hecker[1,2] | Xin Xu[3] | F. William Townes[1] | Heide Loehlein Fier[1,2] | Chris Corcoran[4] | Nan Laird[1] | Christoph Lange[1,5]

FBAT-haplotype algorithm: improvement

- identify set h_{off} of all haplotypes that are compatible with obsered offspring genotypes

- instead of constructing all possible parental mating types, use only h_{off} haplotypes

→ output maintained

→ speed up by several magnitudes

Received: 22 August 2017 | Revised: 29 September 2017 | Accepted: 10 October 2017
DOI: 10.1002/gepi.22094

BRIEF COMMUNICATION

WILEY Genetic Epidemiology
OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

Family-based tests for associating haplotypes with general phenotype data

Improving the FBAT-haplotype algorithm

Julian Hecker[1,2] | Xin Xu[3] | F. William Townes[1] | Heide Loehlein Fier[1,2] | Chris Corcoran[4] | Nan Laird[1] | Christoph Lange[1,5]

Advantage in WGS studies

- nuclear family, 4 offspring, no parental genotypes

8 markers, major/minor A/B, phase unknown

offspring 1: (A/A, A/A, A/A, A/A, A/A, A/B, A/A, A/A)=g_1
offspring 2: (A/A, A/A, A/B, A/A, A/A, A/A, A/A, A/A)=g_2
offspring 3: (A/A, A/A, A/A, A/A, A/A, A/B, A/A, A/A) =g_1
offspring 4: (A/A, A/A, A/B, A/A, A/A, A/A, A/A, A/A) =g_2

Advantage in WGS studies

- nuclear family, 4 offspring, no parental genotypes

8 markers, major/minor A/B, phase unknown

offspring 1: (A/A, A/A, A/A, A/A, A/A, A/B, A/A, A/A)=g_1
offspring 2: (A/A, A/A, A/B, A/A, A/A, A/A, A/A, A/A)=g_2
offspring 3: (A/A, A/A, A/A, A/A, A/A, A/B, A/A, A/A) =g_1
offspring 4: (A/A, A/A, A/B, A/A, A/A, A/A, A/A, A/A) =g_2

original version: number of potential parents : 257

Advantage in WGS studies

- nuclear family, 4 offspring, no parental genotypes

8 markers, major/minor A/B, phase unknown

offspring 1: (A/A, A/A, A/A, A/A, A/A, A/B, A/A, A/A)=g_1
offspring 2: (A/A, A/A, A/B, A/A, A/A, A/A, A/A, A/A)=g_2
offspring 3: (A/A, A/A, A/A, A/A, A/A, A/B, A/A, A/A) =g_1
offspring 4: (A/A, A/A, A/B, A/A, A/A, A/A, A/A, A/A) =g_2

original version: number of potential parents : 257
conditional distribution: $P[1xg_1, 3x\ g_2] = 0.286$, $P[2xg_1, 2xg_2] = 0.429$, $P[3xg_1, 1xg_2] = 0.286$

Advantage in WGS studies

- nuclear family, 4 offspring, no parental genotypes

8 markers, major/minor A/B, phase unknown

offspring 1: (A/A, A/A, A/A, A/A, A/A, A/B, A/A, A/A)=g_1
offspring 2: (A/A, A/A, A/B, A/A, A/A, A/A, A/A, A/A)=g_2
offspring 3: (A/A, A/A, A/A, A/A, A/A, A/B, A/A, A/A) =g_1
offspring 4: (A/A, A/A, A/B, A/A, A/A, A/A, A/A, A/A) =g_2

original version: number of potential parents : 257
conditional distribution: $P[1xg_1, 3x\ g_2] = 0.286$, $P[2xg_1, 2xg_2] = 0.429$, $P[3xg_1, 1xg_2] = 0.286$

haplotypes in h_{off} : $3 \ll 2^8 = 256$ (due to rare variants)

Advantage in WGS studies

- nuclear family, 4 offspring, no parental genotypes

8 markers, major/minor A/B, phase unknown

offspring 1: (A/A, A/A, A/A, A/A, A/A, A/B, A/A, A/A)=g_1
offspring 2: (A/A, A/A, A/B, A/A, A/A, A/A, A/A, A/A)=g_2
offspring 3: (A/A, A/A, A/A, A/A, A/A, A/B, A/A, A/A) =g_1
offspring 4: (A/A, A/A, A/B, A/A, A/A, A/A, A/A, A/A) =g_2

original version: number of potential parents : 257
conditional distribution: $P[1xg_1, 3x\ g_2] = 0.286$, $P[2xg_1, 2xg_2] = 0.429$, $P[3xg_1, 1xg_2] = 0.286$

haplotypes in h_{off} : $3 \ll 2^8 = 256$ (due to rare variants)

number of parental mating types considered: 4, same conditional distribution

Application to Alzheimer‘s Disease WGS study

WGS study with 441 nuclear families
→ 421 have no parental genotypes available!

Application to Alzheimer‘s Disease WGS study

WGS study with 441 nuclear families
→ 421 have no parental genotypes available!

Set	Number of variants	original version	modified version
1	5	8.18 sec	0.04 sec
2	5	9.89 sec	0.05 sec
3	6	230.76 sec	0.04 sec
4	6	191.15 sec	0.14 sec
5	7	43 min	0.06 sec
6	7	27 min	0.04 sec
7	8	**~ 21 hr**	**0.11 sec**

Construction of region-based association tests

Knowledge:

- observed offspring phenotypes T_i
- offspring genotypes X_i
- corresponding conditional distribution

$$T = T(X)$$

→ construct suitable association test statistics $T(X)$ to test the association between genotypes and phenotypes

Multivariate FBAT

- define p dimensional residual vector $\mathrm{U_i} = (X_i - E[X_i|S_i])T_i$
- corresponding $p \times p$ dimensional covariance matrix $V_i = Var(U_i|S_i)$
- both objects computed using the conditional distribution
- Similar to multimarker $FBAT_{MM}$ (Rakovski et al. 2007), but does not need empirical correlation matrix

$$FBAT_{MV} = [\sum_i U_i]^T [\sum_i V_i]^{-1} \; [\sum_i U_i]$$

Burden FBAT

- define p dimensional weight vector W
- collapse residual vector by setting $U_i^* = W^T U_i$
- compute corresponding $V_i^* = W^T V_i W$
- similar to $FBAT_{v0}/FBAT_{v1}$ (De et al. 2013)

$$FBAT_{burden} = \frac{(\sum_i U_i^*)^2}{\sum_i V_i^*}$$

FBAT-SKAT

- overall $\mathrm{N} := \sum_i n_i$ dimensional phenotype vector T
- overall $\mathrm{p} \times N$ dimensional genotype matrix X
- $p \times p$ weight matrix W

$$FBAT_{SKAT} = T^T X^T W X T - T^T E[X^T W X | S] T$$

Association p-values

based on conditional offspring genotype distribution, p-values can be computed by

- asymptotic theory (determine first two moments)
- simulations (draws from conditional distribution)
- exact calculation of p-value (Schneiter, Laird, Corcoran 2005)

$$P_{H_0}[\, T(X) \geq t_{observed}] = ?$$

Simulation study: type 1 error

- null hypothesis
- 400 trios using haplotypes from the EUR sample (1000 Genomes Project)
- 30k windows of 30 consecutive variants with at least one minor allele
- $FBAT_{MV}$ and $FBAT_{SKAT}$ based on simulated p-values (100k replicates)

*based on on 3912 observations

test statistic	0.01	0.05	0.1
$\mathrm{FBAT_{MV}}$	0.00981	0.05074	0.10008
$\mathrm{FBAT_{SKAT}}$	0.01011	0.05077	0.09854
$\mathrm{FBAT_{burden}}$	0.01036	0.04992	0.10047
$\mathrm{FBAT_{burden-w}}$	0.01087	0.04955	0.09881
$\mathrm{FBAT_{v0}}$	0.01035	0.05035	0.10032
$\mathrm{FBAT_{v1}}$	0.01069	0.04951	0.09900
$\mathrm{FBAT_{MM}}$ *	0.03064	0.09834	0.14631

Association p-values

based on conditional offspring genotype distribution, p-values can be computed by

- asymptotic theory (determine first two moments) → rare variants
- **simulations (draws from conditional distribution)**
- exact calculations → complexity

$$P_{H_0}[\, T(X) \geq t_{observed}] = ?$$

Simulation-based p-values and whole-genome scans

- significance levels of interest are very small → computational intensive

Simulation-based p-values and whole-genome scans

- significance levels of interest are very small → computational intensive

- adaptive strategies in existing genetic association analysis tools, e.g. PLINK (Chang et al. 2015), use heuristics to stop early if the p-value is obviously non-significant

Simulation-based p-values and whole-genome scans

- significance levels of interest are very small → computational intensive
- adaptive strategies in existing genetic association analysis tools, e.g. PLINK (Chang et al. 2015), use heuristics to stop early if the p-value is obviously non-significant
- existing sequential Monte Carlo methodology complicated

Simulation-based p-values and whole-genome scans

- significance levels of interest are very small → computational intensive

- adaptive strategies in existing genetic association analysis tools, e.g. PLINK (Chang et al. 2015), use heuristics to stop early if the p-value is obviously non-significant

- existing sequential Monte Carlo methodology complicated

→ sequential testing approach

Simulation-based p-values and whole-genome scans

- p true, unknown association p-value
- sequence $x_1, x_2, \ldots$ where $x_1 = 1$ iff simulated statistic more extreme, 0 otherwise
- we introduce a small indifference region and consider the hypotheses

$$H_1: p \leq p_1 \text{ vs. } H_2: p \geq p_2 = p_1 + d$$

(e.g. $p_1 = 4 * 10^{-8}$ and $d = 10^{-8}$)

Objects and decision rule

objects

pre-specified error probabilities $\alpha_1, \quad \alpha_2$ (e.g. $\alpha_1 = \alpha_2 = 10^{-10}$).

define (Pavlov 1991)

$$\tau_i(\alpha_i) \coloneqq \min\{n\colon \pi_n / \sup_{\theta \in D_i} p_n(\theta, x^n) \geq \alpha_i^{-1}\}$$

for $i = 1,2$, where $D_1 = [0, p_1], D_2 = [p_2, 1], x^n = (x_1, \dots, x_n), p_n(\theta, x^n) = \prod_{i=1}^n p(\theta, x_i)$, $\pi_n \coloneqq \prod_{i=1}^n p(\hat{\theta}_{i-1}, x_i)$ and $\boldsymbol{\widehat{\theta}_{i-1}} \coloneqq \frac{\sum_{k=1}^{i-1} x_k + \frac{1}{2}}{i}$.

$p(\theta, x)$ Bernoulli density with parameter θ.

Objects and decision rule

objects

pre-specified error probabilities α_1, α_2 (e.g. $\alpha_1 = \alpha_2 = 10^{-10}$).

define (Pavlov 1991)

$$\tau_i(\alpha_i) := \min\{n\colon \pi_n / \sup_{\theta \in D_i} p_n(\theta, x^n) \geq \alpha_i^{-1}\}$$

for $i = 1,2$, where $D_1 = [0, p_1], D_2 = [p_2, 1], x^n = (x_1, \dots, x_n), p_n(\theta, x^n) = \prod_{i=1}^n p(\theta, x_i)$, $\pi_n := \prod_{i=1}^n p(\hat{\theta}_{i-1}, x_i)$ and $\boldsymbol{\widehat{\theta}_{i-1}} := \frac{\sum_{k=1}^{i-1} x_k + \frac{1}{2}}{i}$.

$p(\theta, x)$ Bernoulli density with parameter θ.

decision procedure STr

If $\tau_1(\alpha_1) \leq \tau_2(\alpha_2)$, we set $\partial = 2$ and $N = \tau_1(\alpha_1)$. If $\tau_1(\alpha_1) > \tau_2(\alpha_2)$, we set $\partial = 1$ and $N = \tau_2(\alpha_2)$.

Theoretical result

Theorem (Pavlov 1991, Tartakovsky 2014)

1.) $P_\theta[\delta = 2] \leq \alpha_1$ for $\theta \in D_1$ and $P_\theta[\delta = 1] \leq \alpha_2$ for $\theta \in D_2$

2.) Let $K(t_1, t_2, \alpha)$ be the class of all decision rules (N', ∂') such that $P_\theta[\delta' = 2] \leq t_1\alpha$ for $\theta \in D_1$and $P_\theta[\delta' = 1] \leq t_2\alpha$ for $\theta \in D_2$, then

$$\frac{E_{\theta[N]}}{\inf\limits_{(N',\partial')\in K(t_1,t_2,\alpha)} E_\theta[N']} = 1 + o(1) \text{ as } \alpha \to 0 \text{ for all } \theta \in [0,1].$$

→ error probabilities are strictly controlled
→ approaches theoretical minimum number of expected simulations if error level goes to zero

Comparison with confidence interval based approach

- $\hat{p}$ empirical estimate of p-value after n simulations
- $(\hat{p} - c_\alpha SE, \hat{p} + c_\alpha\, SE)$ corresponding $1-\alpha$ confidence interval, based on asymptotic theory, c_α is $1 - \frac{\alpha}{2}$ quantile of standard normal distribution

CI-based rule (CIr)
choose $\partial = 1$ if $\hat{p} + c_\alpha SE \leq p_2$, set $\partial = 2$ if $\hat{p} - c_\alpha SE \geq p_1$. Similar to adaptive strategy implemented in PLINK (Chang et al. 2015)

- simulated 12,045,191 p-values for SNPs in LD, mimicked testing by Bernoulli draws where success parameter = p-value
- compared overall number of required draws for different choices for p_1 and p_2.

Comparison with confidence interval based approach

p_1/p_2	α_1/α_2	α	STr	CIr	ratio CIr/STr	error STr	error CIr
1e-09/2e-09	1e-10/1e-10	1e-10	6.04e08	7.62e09	**12.62**	0/0	0/0
5e-08/6e-08	1e-10/1e-10	1e-10	1.23e09	7.86e09	**6.39**	0/0	0/0
9e-04/1e-03	1e-10/1e-10	1e-10	2.41e10	1.85e10	**0.77**	0/0	**10/0**
9e-04/1e-03	1e-10/4e-03	1e-10	1.66e10	1.85e10	**1.11**	**0/0**	**10/0**

STr: total number of simulations for STr

CIr: total number of simulations for CIr

error: number of observed „type 1 / type 1“ errors

Comparison with confidence interval based approach

p_1/p_2	α_1/α_2	α	STr	CIr	ratio CIr/STr	error STr	error CIr
1e-09/2e-09	1e-10/1e-10	1e-10	6.04e08	7.62e09	**12.62**	0/0	0/0
5e-08/6e-08	1e-10/1e-10	1e-10	1.23e09	7.86e09	**6.39**	0/0	0/0
9e-04/1e-03	1e-10/1e-10	1e-10	2.41e10	1.85e10	**0.77**	0/0	**10/0**
9e-04/1e-03	1e-10/4e-03	1e-10	1.66e10	1.85e10	**1.11**	**0/0**	**10/0**

STr: total number of simulations for STr

CIr: total number of simulations for CIr

error: number of observed „type 1 / type 1“ errors

CIr: type 1 error at least 0.00425

Remarks

- STr: roughly 98% of simulations for 1% of SNPs
- can be applied to any association test statistic
- sequential Monte Carlo
  $\mathrm{H}_1\colon\ p \leq p_1$ vs. $\mathrm{H}_2\colon\ p > p_1$,
  worst-case $p \approx p_1$
- interesting scenario $d \to \varepsilon$

Supplementary materials for this article are available at http://pubs.amstat.org/toc/jasa/104/488.

Sequential Implementation of Monte Carlo Tests With Uniformly Bounded Resampling Risk

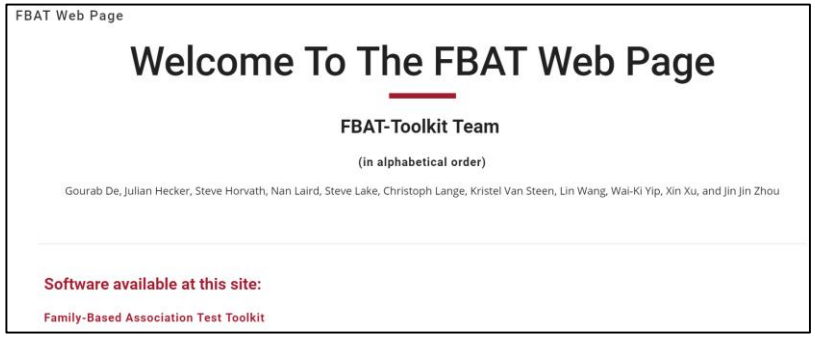
Axel GANDY

This paper introduces an open-ended sequential algorithm for computing the p-value of a test using Monte Carlo simulation. It guarantees that the resampling risk, the probability of a different decision than the one based on the theoretical p-value, is uniformly bounded by an arbitrarily small constant. Previously suggested sequential or nonsequential algorithms, using a bounded sample size, do not have this property. Although the algorithm is open-ended, the expected number of steps is finite, except when the p-value is on the threshold between rejecting and not rejecting. The algorithm is suitable as standard for implementing tests that require (re)sampling. It can also be used in other situations: to check whether a test is conservative, iteratively to implement double bootstrap tests, and to determine the sample size required for a certain power. An R-package implementing the sequential algorithm is available online.

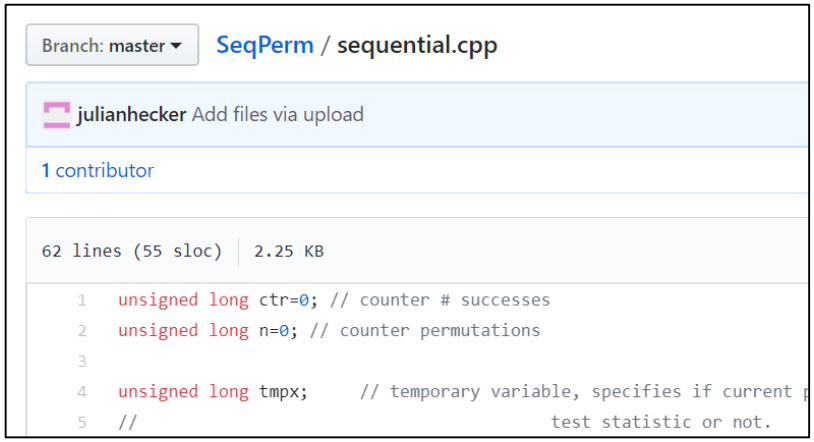
KEY WORDS: Monte Carlo testing; p-value; Sequential estimation; Sequential test; Significance test.

Discussion

- general framework for region-based association analysis in family-based studies
- robustness due to conditional genotype distribution
- multivariate, burden and SKAT association test statistics
- efficient and rigorous procedure to evaluate simulation-based p-value
- implementation available soon



https://sites.google.com/view/fbat-web-page



Acknowledgements

- Brent Coull (HSPH Boston)
- Nan Laird (HSPH Boston)
- Ingo Ruczinski (Johns Hopkins)
- F. William Townes (HSPH Boston)
- Edwin Silverman (Brigham and Women‘s Hospital)
- Michael Cho (Brigham and Women‘s Hospital)
- Chris Corcoran (Utah State)