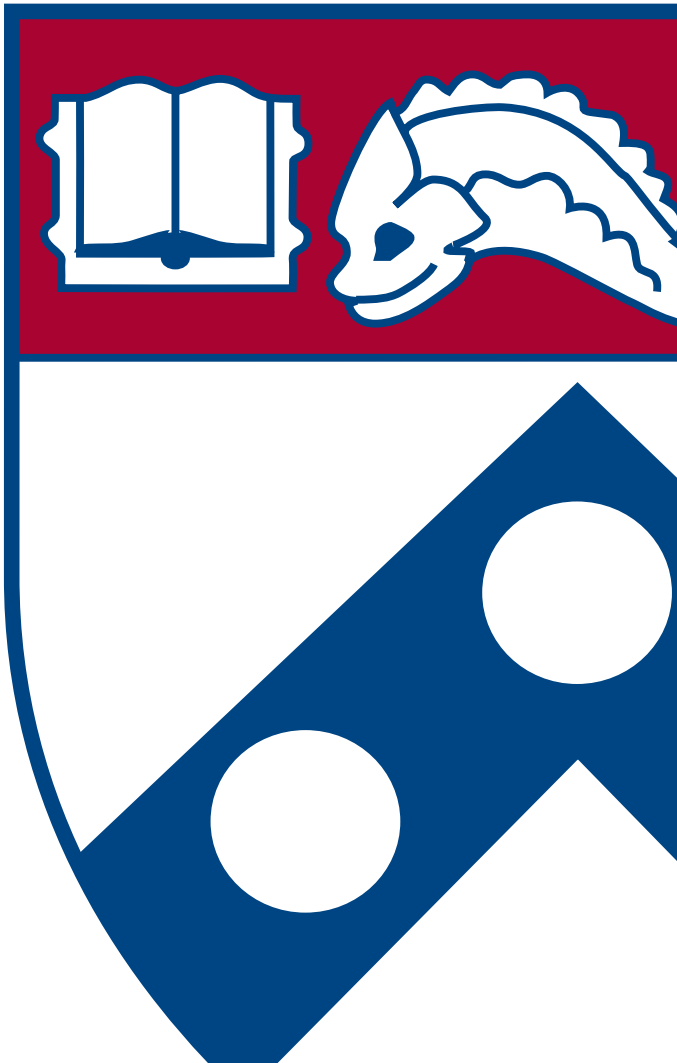
Simple haplotype analyses in R

Benjamin French, PhD
Department of Biostatistics and Epidemiology
University of Pennsylvania
bcfrench@upenn.edu



useR! 2011
University of Warwick
18 August 2011

Goals

- To integrate haplotypes into large association studies such that haplotype imputation is done once as a data-processing step
  - Case-control studies (binary outcome)
  - Prospective studies (censored survival outcome)
- To allow haplotype associations to be estimated in general-purpose statistical software (eg R) by researchers expert in the subject matter

"In world historical terms there is a lot to be said
for keeping data analysis out of the hands of statisticians"

— Thomas Lumley

Goals

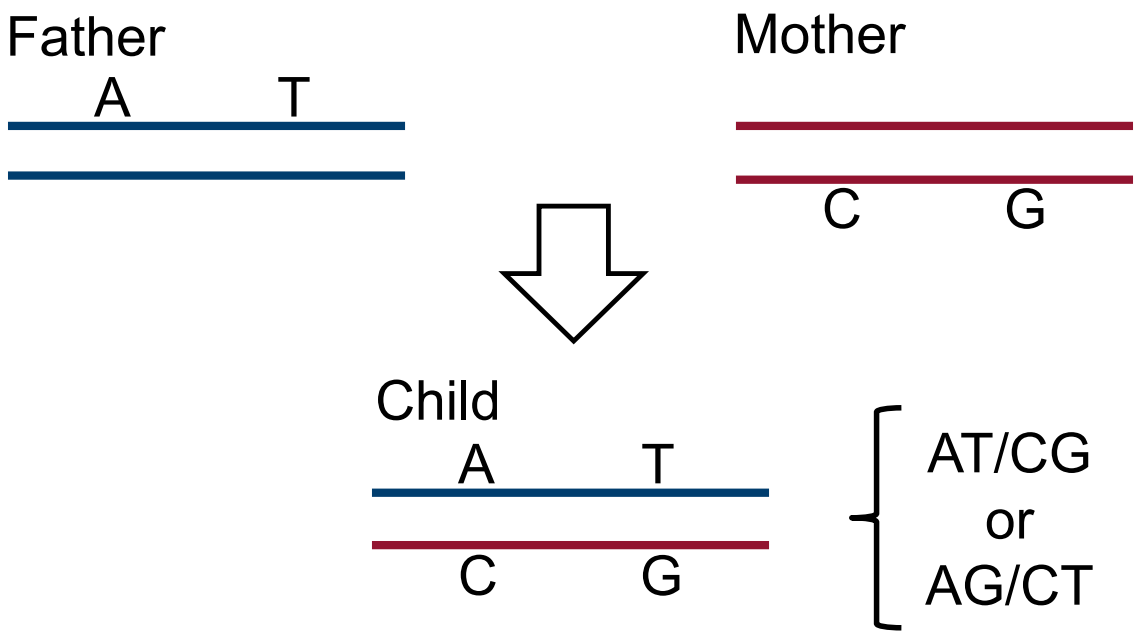
- To integrate haplotypes into large association studies such that haplotype imputation is done once as a data-processing step
  - Case-control studies (binary outcome)
  - Prospective studies (censored survival outcome)
- To allow haplotype associations to be estimated in general-purpose statistical software (eg R) by researchers expert in the subject matter

"In world historical terms there is a lot to be said
for keeping data analysis out of the hands of statisticians"

— Thomas Lumley

Phase ambiguity

- Observed data is composed of a set of unphased genotypes
- Diplotype (pair of haplotypes) may be ambiguous; may not know which allele was transmitted from maternal or paternal chromosome
- Missing data problem; impute the unobserved diplotype



Haplotype imputation

Expectation-maximization (EM) algorithm

- E: calculate expected phase given haplotype frequencies
- M: calculate MLEs for haplotype frequencies given phase
- Software: `haplo.stats` [Sinnwell and Schaid, 2009]

Bayesian inference

- Observed genotype data combined with expected haplotype patterns
- Haplotypes estimated from posterior distribution
- Software: PHASE [Stephens and Donnelly, 2003]

Diplotype uncertainty

Angiotensin II receptor type 1 (*AGTR1*)

Label	Haplotype	Haplotype frequency	Diplotype probability
D	TCCACGCATCTT	0.139	0.81
F	TCTGTGCATCTC	0.290	
C	TCCACGCATCTC	0.034	0.19
G	TCTGTGCATCTT	0.272	
Rare	TCCGCGCATCTC	< 0.001	< 0.01
Rare	TCTATGCATCTT	< 0.001	

Estimating associations

Non-iterative weighted estimation [French et al., 2006]

1. Impute haplotypes and estimate population haplotype frequencies
2. Create multi-record data for each individual
   - Design matrix: set of diplotypes consistent with observed genotype, possibly including environmental exposures
   - Weights equal to conditional probability of each diplotype

Weight	A	B	C	D	E	F	G	H	I	Rare
0.81	0	0	0	1	0	1	0	0	0	0
0.19	0	0	1	0	0	0	1	0	0	0
< 0.01	0	0	0	0	0	0	0	0	0	2

3. Estimate associations using a weighted regression model
   - Logistic regression for binary outcomes
   - Cox regression for censored survival outcomes
   - Robust or sandwich standard error estimator
   - Account for uncertainty in phase

Estimating associations

Non-iterative weighted estimation [French et al., 2006]

1. Impute haplotypes and estimate population haplotype frequencies
2. Create multi-record data for each individual
   - Design matrix: set of diplotypes consistent with observed genotype, possibly including environmental exposures
   - Weights equal to conditional probability of each diplotype

Weight	A	B	C	D	E	F	G	H	I	Rare
0.81	0	0	0	1	0	1	0	0	0	0
0.19	0	0	1	0	0	0	1	0	0	0
< 0.01	0	0	0	0	0	0	0	0	0	2

3. Estimate associations using a weighted regression model
   - Logistic regression for binary outcomes
   - Cox regression for censored survival outcomes
   - Robust or sandwich standard error estimator
   - Account for uncertainty in phase

Simulation study

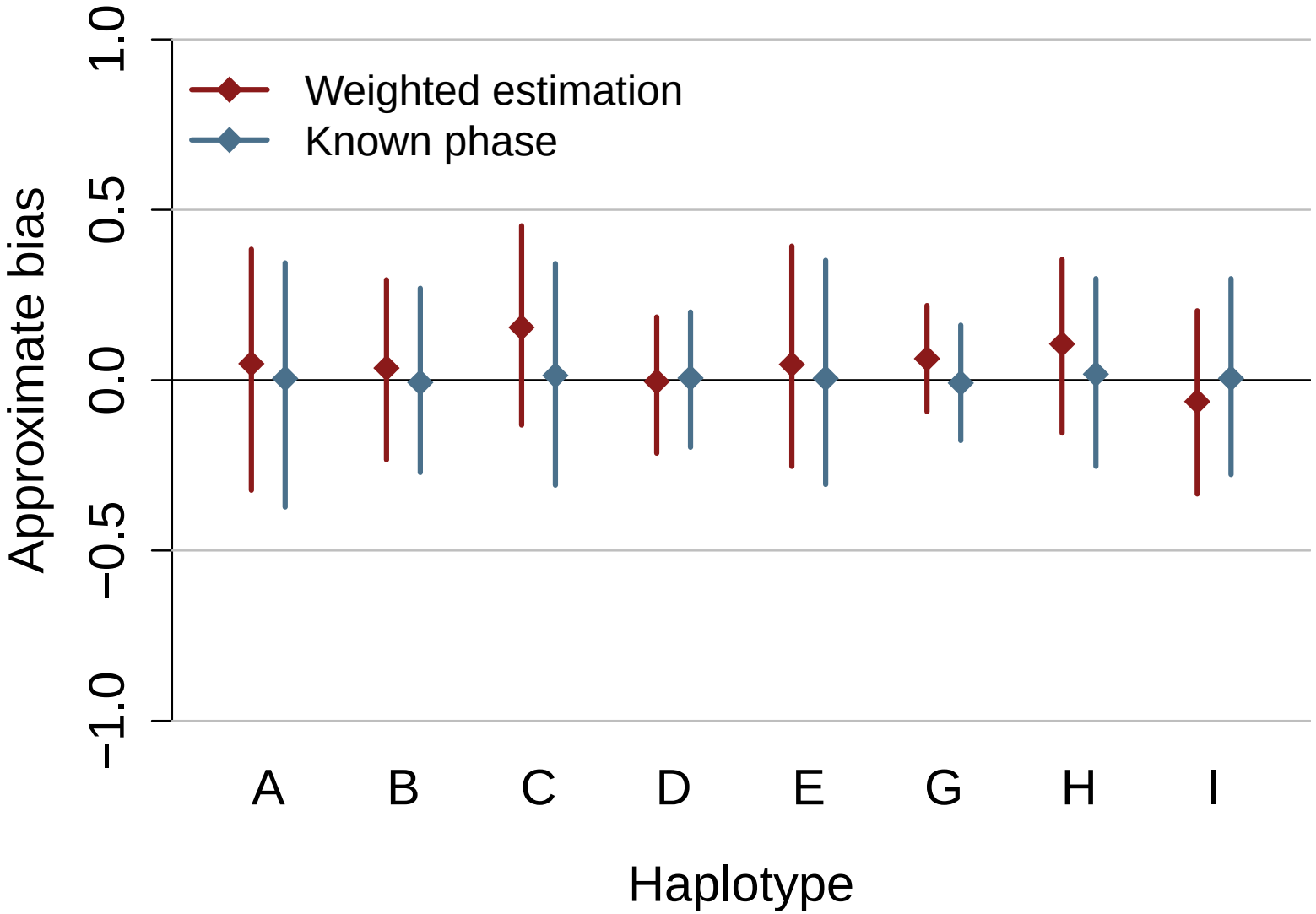
Angiotensin II receptor type 1 (*AGTR1*)

Label	Haplotype	Frequency	log HR
A	ATTATGCATCTC	0.029	$-\log 2.0$
B	ATTATGTGATCC	0.051	$-\log 1.75$
C	TCCACGCATCTC	0.027	$\log 1.25$
D	TCCACGCATCTT	0.090	$\log 1.5$
E	TCTGTGCAACTT	0.029	$-\log 1.25$
F*	TCTGTGCATCTC	0.223	—
G	TCTGTGCATCTT	0.188	$-\log 1.5$
H	TTTACACATCTC	0.038	$\log 1.75$
I	TTTACACATCTT	0.032	$\log 2.0$

* Referent

$n = 500$, 25% censoring

Simulation results



Application

CLCNKA haplotypes and adverse events in chronic heart failure

- Regulate renal potassium channels to control blood pressure
- SNP associated with heart failure in a large case-control study [Cappola et al., 2011]
- Genotypes available for 1150 genetically inferred Caucasians with heart failure enrolled in a prospective study
- 70% male; median age at study entry, 58 years
- 14 pre-selected SNPs
  Inferred 10 common haplotypes (frequency > 0.02)
- 65% had an unambiguous diplotype
  90% had a highest posterior probability > 0.765

Application

CLCNKA haplotypes and adverse events in chronic heart failure

- Outcome: time to all-cause mortality or cardiac transplantation
  - Median follow-up, 3 years; maximum, 5 years
  - 22% experienced an adverse event
- Non-iterative weighted estimation with Cox regression
  - Included all diplotypes consistent with observed genotype
  - Weighted by conditional probability of each diplotype
  - Stratified by 4-level classification for disease severity
  - Adjusted for gender, age, heart failure etiology, clinical site
  - Time-varying covariate for age (exhibited non-proportional hazards)
  - Robust variance estimator for standard error estimation

Application results

Label	Haplotype	Frequency	HR (95% CI)	*P*
Q	AGAGCGAGACGAGG	0.036	1.19 (0.80, 1.77)	0.39
R	AGAGCGAGGGAAGG	0.160	1.04 (0.80, 1.35)	0.79
S	AGAGCGGAGCAAGA	0.036	1.20 (0.80, 1.77)	0.38
T	AGCGAGAGGCAAGA	0.066	0.55 (0.34, 0.88)	0.01
U	GACGCGGAGCGCGG	0.063	0.80 (0.53, 1.20)	0.28
V	GGAACAAGGGAAGG	0.037	0.49 (0.26, 0.92)	0.03
W	GGAACAGAGCAAGA	0.299	Referent	
X	GGAACAGAGCAAGG	0.048	1.36 (0.95, 1.95)	0.09
Y	GGAGCAAGGCAAGG	0.050	1.15 (0.78, 1.69)	0.49
Z	GGCGCGGAGCAAGG	0.031	1.09 (0.62, 1.92)	0.76
Overall				0.02

Application results

Label	Haplotype	Frequency	HR (95% CI)	*P*
Q	AGAGCGAGACGAGG	0.036	1.19 (0.80, 1.77)	0.39
R	AGAGCGAGGGAAGG	0.160	1.04 (0.80, 1.35)	0.79
S	AGAGCGGAGCAAGA	0.036	1.20 (0.80, 1.77)	0.38
T	AGCGAGAGGCAAGA	0.066	0.55 (0.34, 0.88)	0.01
U	GACGCGGAGCGCGG	0.063	0.80 (0.53, 1.20)	0.28
V	GGAACAAGGGAAGG	0.037	0.49 (0.26, 0.92)	0.03
W	GGAACAGAGCAAGA	0.299	Referent	
X	GGAACAGAGCAAGG	0.048	1.36 (0.95, 1.95)	0.09
Y	GGAGCAAGGCAAGG	0.050	1.15 (0.78, 1.69)	0.49
Z	GGCGCGGAGCAAGG	0.031	1.09 (0.62, 1.92)	0.76
Overall				0.02

R packages: Non-iterative weighted estimation

`haplo.ccs` [French and Lumley, 2007]

- Weighted logistic regression for binary outcomes
- Depends on `haplo.stats` package to impute haplotypes
- Calls `glm(..., family=quasibinomial(link=logit))`
- Includes GEE-type sandwich standard error estimator

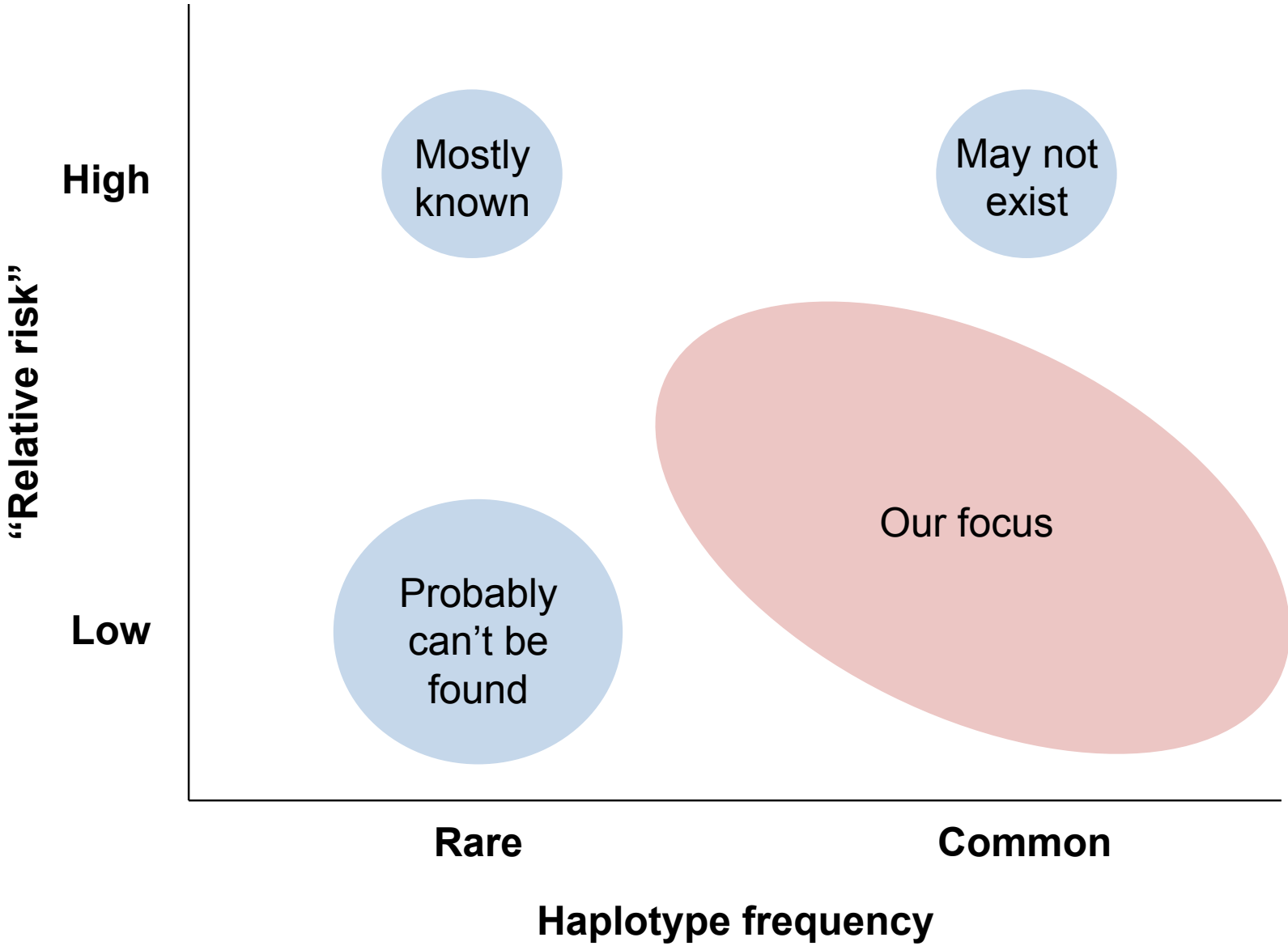
`haplo.cph` (in process)

- Weighted Cox regression for censored survival outcomes
- Will depend on `haplo.stats` package to impute haplotypes
- Will call `cph(..., robust=TRUE)` from `Design` package
- Allow stratification and time-varying exposures

Conclusions

- Non-iterative weighted estimation
  - Valid tests for genetic associations
  - Reliable estimates of modest genetic effects of common haplotypes
- Regression-based framework
  - Adjustment for or interaction with environmental exposures
  - Stratification and time-varying exposures in Cox regression
- Straightforward to implement in R
  - `haplo.ccs` for binary outcomes
  - `haplo.cph` for censored survival outcomes

Target


"Relative risk"
High
Low
Mostly known
May not exist
Our focus
Probably can't be found
Rare
Common
Haplotype frequency

Limitations

Our method and/or software may not be applicable to

- Related individuals
- Rare haplotypes
- Small studies

Collaborators

Nandita Mitra, PhD
Department of Biostatistics and Epidemiology
University of Pennsylvania

Thomas P Cappola, MD
Penn Cardiovascular Institute
University of Pennsylvania

Thomas Lumley, PhD
Department of Statistics
University of Auckland

References

1. Cappola TP, Matkovich SJ, et al. 2011. Loss-of-function DNA sequence variant in the CLCNKA chloride channel implicates the cardio-renal axis in interindividual heart failure risk variation. Proc Natl Acad Sci U S A 108:2456–61.
2. French B, Lumley T. 2007. haplo.ccs: Estimate haplotype relative risks in case-control data. R package 1.3.
3. French B, Lumley T, et al. 2006. Simple estimates of haplotype relative risks in case-control data. Genet Epidemiol 30:485–94.
4. Sinnwell JP, Schaid DJ. 2009. haplo.stats: Statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous. R package version 1.4.4.
5. Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–69.