Haplotype analysis of case-control data

Yulia Marchenko

Senior Statistician
StataCorp LP

2010 UK Stata Users Group Meeting

Outline

1. Haplotype-based disease association studies
   - Genetic markers
   - Lung-cancer example
   - The `haplologit` command
   - New capabilities
2. Genome-wide association studies (GWAS)
   - Sliding windows
   - GWAS of lung-cancer data
3. Future work

Genetic population-based disease association studies

- Main goal: determine genetic variants influencing complex diseases
- Genetic information is available through genetic markers such as biallelic SNPs (International SNP Map Working Group 2001, International Hapmap Consortium 2003, 2005, 2007)
- Genetics effects are often small and thus difficult to detect
- Genetic effects often interact with environmental factors
- Efficient analysis of genetic effects and their interactions with environment is of great importance

Genetic markers - SNPs

- *Single nucleotide polymorphism* (SNP, pronounced as "snip") is a single nucleotide (A, T, C, or G) variation of the DNA sequence that occurs in at least 1% of the population.
- Example: C-T SNP
  DNA fragment of subject 1: AAGC**C**TA
  DNA fragment of subject 2: AAGC**T**TA
- C and T are *alleles*, alternative forms of a DNA segment at a single locus. One of these alleles is common, another one is rare
- Subjects' genetic information is described by SNP genotypes, e.g. CC, CT, or TT
- Standard categorical methods can be used to test for association between a disease and a SNP genotype under various genetic models (additive, dominant, recessive, etc.)

Lung-cancer example

- Consider a subset of case-control lung-cancer data of current and former smokers from Amos et al. (2008)
- 9 SNPs, variables `snp1-snp9`, spanning the interval between rs8034191 and rs8192475
- Other characteristics: `cancer`, `female`, `smkformer`, `packyrs`
- Two SNPs, rs8034191 (`snp1`) and rs1051730 (`snp8`), in a region of 15q25.1 containing nicotinic acetylcholine receptors genes are significantly associated with risk of lung cancer
- Data summary:

Characteristic	Cases	Controls
Sex (% female)	42.98	43.36
Former smokers (%)	52.25	57.78
Pack years (s.d.)	51.49 (31.41)	44.57 (30.16)
Total	1154	1137

- For example, we can use `tabodds` to obtain genotypic odds ratios separately for each SNP of interest:

```
. tabodds cancer snp1, or

-----------------------------------------------------------------------------
        snp1 |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+---------------------------------------------------------------
          AA |    1.000000          .           .             .          .
          AG |    1.188315       3.65      0.0561      0.995320   1.418732
          GG |    1.811803      20.08      0.0000      1.391670   2.358770
-----------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(2)  =    20.16
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    18.34
                                  Pr>chi2  =   0.0000


. tabodds cancer snp8, or

-----------------------------------------------------------------------------
        snp8 |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+---------------------------------------------------------------
          GG |    1.000000          .           .             .          .
          AG |    1.250974       6.15      0.0132      1.047655   1.493752
          AA |    1.777132      18.92      0.0000      1.366588   2.311010
-----------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(2)  =    19.83
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    19.37
                                  Pr>chi2  =   0.0000
```

Haplotypes and diplotypes

- Single SNP analysis may have low power to detect genetic effects (Akey et al. 2001, de Bakker et al. 2005)
- Alternative: analyze multiple SNPs simultaneously via haplotypes
- Humans' genetic information is comprised of diplotypes
- In practice, we usually observe genotypes (the sums of two haplotypes) rather than diplotypes
- Example: 2 SNPs (binary notation: 0 is common allele, 1 is rare allele)

  4 possible haplotypes: 00, 01, 10, 11
  16 possible diplotypes: (00,00), (00,01), . . . , (11,10), (11,11)
  9 possible genotypes: 00, 01, 02, 10, 11, 12, 20, 21, 22

Lung-cancer data, haplotype analysis

- Let's now analyze two SNPs of interest simultaneously using `haplologit` (Marchenko et al. 2008)
- Major (reference) and minor alleles are coded as 0 and 1, respectively
- A is a reference allele for `snp1`, G is a reference allele for `snp8`

```
. haplologit cancer, snp(snp1 snp8)
Handling missing SNPs:
Building consistent haplotype pairs:
Obtaining initial haplotype frequency estimates from the control sample:
Haplotype frequency EM estimation under HWE
Number of iterations   =          8
Sample log-likelihood = -1329.3903
```

haplotype	frequency*
00	.652003
01	.011145
10	.013344
11	.323507

```
* frequencies > .001
```

(*Continued on next page*)

```
Performing gradient-based optimization:
note: using the most frequent haplotype from the control sample as a risk haplotype
Haplotype-effects logistic regression
Mode of inheritance: additive                        Number of obs      =      2291
Genetic distribution: Hardy-Weinberg equilib.        Number phased      =      1289
Genotype: snp1 snp8                                  Number unphased    =      1000
                                                     Number missing     =         2
                                                     Wald chi2(1)       =     18.47
Retrospective log likelihood = -2746.8085            Prob > chi2        =    0.0000
```

cancer	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hap_00	-0.263	0.061	-4.30	0.000	-0.382	-0.143

Haplotype Frequencies	Estimate	Std. Err.	[95% Conf.	Interval]
hap_00	.652029	.0099915	.632446	.671612
hap_01	.0105619	.0014741	.0076727	.0134512
hap_10	.011765	.0015559	.0087154	.0148146
hap_11	.325644	.0095724	.3068825	.3444055

- Let's use the most frequent haplotype 00 as a reference and include effects of all other haplotypes:

```
. haplologit cancer, snp(snp1 snp8) riskhap1("11") riskhap2("10") riskhap3("01") noemshow
Handling missing SNPs:
Building consistent haplotype pairs:
Obtaining initial haplotype frequency estimates from the control sample:
Performing gradient-based optimization:
Haplotype-effects logistic regression
Mode of inheritance: additive                    Number of obs       =      2291
Genetic distribution: Hardy-Weinberg equilib.    Number phased       =      1289
Genotype: snp1 snp8                              Number unphased     =      1000
                                                 Number missing      =         2
                                                 Wald chi2(3)        =     19.51
Retrospective log likelihood = -2746.2814        Prob > chi2         =    0.0002
```

cancer	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hap_11	0.275	0.062	4.40	0.000	0.152	0.397
hap_10	0.017	0.266	0.06	0.949	-0.503	0.537
hap_01	0.161	0.280	0.58	0.565	-0.388	0.710

Haplotype Frequencies	Estimate	Std. Err.	[95% Conf.	Interval]
hap_00	.6520033	.0099923	.6324187	.6715878
hap_01	.0111454	.002217	.0068002	.0154905
hap_10	.0133441	.0024204	.0086003	.018088
hap_11	.3235072	.0098137	.3042727	.3427417

Why use haplologit?

- haplologit allows joint estimation of multiple SNPs via haplotypes and, thus, can be more powerful in detecting genetic associations
- haplologit accounts for retrospective sampling design and, thus, is more appropriate for the analysis of case-control data
- haplologit can be more efficient than standard prospective logistic regression under the assumptions of Hardy-Weinberg equilibrium (HWE) and independence between haplotypes and environmental factors
- haplologit handles unphased and missing genotypes

What does `haplologit` do?

`haplologit` fits haplotype-based logistic regression to case-control data and estimates the effects of haplotypes of interest on the disease and, optionally, their interactions with environmental factors using efficient semiparametric method of Spinka et al. (2005) and Lin and Zeng (2006) which

- accounts for retrospective sampling design
- incorporates phase uncertainty
- handles missing genotypes

Haplotype-based logistic model

$$\begin{aligned}\text{logit}\left\{\Pr(D=1|\mathbf{Z},\mathbf{G})\right\} &= \alpha_0 + \beta_1 I_{H_1^\star} + \beta_2 I_{H_2^\star} + \dots \\ &\quad + \gamma_1 I_{H_1^\star} Z_1 + \gamma_2 I_{H_1^\star} Z_2 + \dots\end{aligned}$$

- βs are haplotype main effects, γs are haplotype-environment interaction effects
- $\mathbf{Z}$ are environmental covariates, $\mathbf{G}$ are observed genotypes
- $I_{H_i^\star}$s are genetic covariates, which are determined by a chosen genetic model and depend on the number of copies of a risk haplotype $H_i^\star$ in observed genotypes $\mathbf{G}$ (or, more specifically, corresponding diplotypes).

Retrospective sampling

- Select cases ($D = 1$) and sample from them to obtain values of genotypes $\mathbf{G}$ and covariates $\mathbf{Z}$
- Select controls ($D = 0$) and sample from them to obtain values of genotypes $\mathbf{G}$ and covariates $\mathbf{Z}$
- Samples are obtained conditional on the disease status D:

$$f(\mathbf{Z}, \mathbf{G}|D) = \frac{\Pr(D|\mathbf{Z}, \mathbf{G}) f(\mathbf{Z}, \mathbf{G})}{\Pr(D)}$$

- Standard logistic regression (ignoring retrospective design) is semiparametric-efficient when covariate distribution $f(\mathbf{Z}, \mathbf{G})$ is unrestricted (Breslow et al. 2000)

- To increase efficiency, we can utilize information about $f(\mathbf{Z},\mathbf{G})$ often associated with genetic data:

  a) population in Hardy-Weinberg equilibrium

$$
\begin{aligned}
q\{(H_k,H_l);\boldsymbol{\theta}\} &= \theta_k^2 & \text{if } H_k = H_l \\
&= 2\theta_k\theta_l & \text{if } H_k \neq H_l
\end{aligned}
$$

  θ_k denotes the frequency for haplotype H_k.

  b) gene-environment independence – $f(\mathbf{Z},\mathbf{G}) = g(\mathbf{Z})q(\mathbf{G})$

- To handle unphased and missing genotypes, we need to impose restrictions on the genetic distribution (such as HWE or certain deviations from it)

Missing genotypes

- Genotypes **G** are assumed to be missing at random
- Keeping in mind binary notation, missing components of **G** may be any value from $\{0, 1, 2\}$ resulting in multiple plausible diplotypes for a subject with incomplete genetic information
- Missing genotypes are handled by "averaging" the likelihood over all such constituent diplotypes for each subject
- Accommodation of missing genotypes requires distributional assumptions (e.g., HWE) for the genetic data

Unphased genotypes

- Consider 2 SNP genotypes AG and CT of a subject
- Two diplotypes are consistent with the observed genotype: (AC, GT) and (AT, GC)
- Thus, phase is indeterminant (ambiguous) for this subject
- More generally, phase ambiguity arises for heterozygous subjects who carry different alleles at two or more SNP loci
- Phase ambiguity can be viewed as a missing-data problem and is handled similarly

haplologit's capabilities

Marchenko et al. (2008) presented the `haplologit` command for haplotype analysis of case-control genetic data in the important special case of

- a rare disease
- a single candidate gene in HWE
- gene-environment independence

The command also supported a number of genetic models, such as additive, recessive, and dominant.

New capabilities include:

- relaxing the assumption of HWE
- extending the catalogue of genetic models to include codominant models
- genome-wide association analysis

New capabilities

- relaxing the assumption of HWE:

$$
\begin{aligned}
q\{(H_k, H_l);\boldsymbol{\theta}\} &= \theta_k^2 + \rho\theta_k(1-\theta_k) && \text{if } H_k = H_l \\
&= (1-\rho)\theta_k\theta_l && \text{if } H_k \neq H_l
\end{aligned}
$$

  where ρ denotes the inbreeding coefficient.
- codominant models:
  - homozygous/heterozygous model — the effect of having two copies of a rare haplotype is allowed to be different from the effect of having only one copy
  - additive/recessive model — the effect of a rare haplotype is decomposed into two separate components, additive and recessive, allowing to test if the effects are additive, recessive, or dominant

Hardy-Weinberg disequilibrium

```
. haplologit cancer, snp(snp1 snp8) riskhap1("11") hwd
Handling missing SNPs:
Building consistent haplotype pairs:
Obtaining initial haplotype frequency estimates from the control sample:
Haplotype frequency EM estimation under HWD
Number of iterations  =        175
Sample log-likelihood = -1329.3914
```

haplotype	frequency*
00	.652003
01	.011145
10	.013344
11	.323507

```
* frequencies > .001
Inbreeding rho =   .000023
```

(*Continued on next page*)

```
Performing gradient-based optimization:
Iteration 0:   Retrospective log likelihood = -2766.2715
Iteration 1:   Retrospective log likelihood = -2746.4871
Iteration 2:   Retrospective log likelihood = -2746.4482
Iteration 3:   Retrospective log likelihood = -2746.4482
Haplotype-effects logistic regression
Mode of inheritance: additive                  Number of obs      =       2291
Genetic distribution: Hardy-Weinberg disequil. Number phased      =       1289
Genotype: snp1 snp8                            Number unphased    =       1000
                                               Number missing     =          2
                                               Wald chi2(1)       =      19.17
Retrospective log likelihood = -2746.4482      Prob > chi2        =     0.0000

------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hap_11 |   .2713723   .0619759     4.38   0.000     .1499017    .3928429
------------------------------------------------------------------------------

------------------------------------------------------------------------------
   Haplotype Frequencies |   Estimate   Std. Err.         [95% Conf. Interval]
-------------------------+----------------------------------------------------
                  hap_00 |   .6510113   .0097365          .6319281    .6700946
                  hap_01 |   .0120607   .0016671          .0087932    .0153282
                  hap_10 |   .0134345   .0017577          .0099896    .0168795
                  hap_11 |   .3234934   .0098139          .3042586    .3427282
                     rho |   4.02e-08          .                 .           .
------------------------------------------------------------------------------
```

Codominant model: hetero/homo-zygous effects

```
. haplologit cancer, snp(snp1 snp8) riskhap1("11") inheritance(codominant) or
Haplotype-effects logistic regression
Mode of inheritance: type I codominant          Number of obs      =      2291
Genetic distribution: Hardy-Weinberg equilib.   Number phased      =      1289
Genotype: snp1 snp8                             Number unphased    =      1000
                                                Number missing     =         2
                                                Wald chi2(2)       =     20.97
Retrospective log likelihood =   -2745.75       Prob > chi2        =    0.0000
```

cancer	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hap_11						
heteroz.	1.239025	.0972226	2.73	0.006	1.062402	1.445011
homoz.	1.777553	.223547	4.57	0.000	1.389231	2.27442

Haplotype Frequencies	Estimate	Std. Err.	[95% Conf.	Interval]
hap_00	.6510032	.0097367	.6319196	.6700867
hap_01	.0120649	.0016677	.0087963	.0153334
hap_10	.0134386	.0017582	.0099927	.0168846
hap_11	.3234933	.0098139	.3042585	.3427281

- Adjust for packyrs and consider haplotype-packyrs interaction:

```
. haplologit cancer packyrs, snp(snp1 snp8) riskhap1("11", inter(packyrs))
> inheritance(codominant) or

Haplotype-effects logistic regression
Mode of inheritance: type I codominant         Number of obs      =       2291
Genetic distribution: Hardy-Weinberg equilib.  Number phased      =       1289
Genotype: snp1 snp8                            Number unphased    =       1000
                                               Number missing     =          2
                                               Wald chi2(5)       =      52.42
Retrosp. profile log likelihood = -4318.1426   Prob > chi2        =     0.0000
```

cancer	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
packyrs	1.006844	.0018279	3.76	0.000	1.003268	1.010433
hap_11						
heteroz.	1.235895	.1580349	1.66	0.098	.9619177	1.587909
homoz.	1.478571	.2756675	2.10	0.036	1.025989	2.130796
hap_11Xpac~s						
heteroz.	1.00005	.0019853	0.03	0.980	.9961662	1.003948
homoz.	1.003496	.002579	1.36	0.175	.9984536	1.008563

Note: _cons = b0 + ln(N1/N0) - ln{Pr(D=1)/Pr(D=0)}

Haplotype Frequencies	Estimate	Std. Err.	[95% Conf.	Interval]
hap_00	.6510032	.0097367	.6319196	.6700867
hap_01	.0120649	.0016677	.0087963	.0153334
hap_10	.0134386	.0017582	.0099927	.0168846
hap_11	.3234933	.0098139	.3042585	.3427281

- Consider all 9 SNPs:

```
. haplologit cancer, snp(snp1-snp9) riskhap1(158) riskhap2(161) riskhap3(320)
> riskhap4(448)
Haplotype frequency EM estimation under HWE
Number of iterations  =          52
Sample log-likelihood = -3457.3456
```

haplotype	frequency*
010000000	.002378
010000001	.357418
010011101	.020671
010011111	.002505
010100000	.044521
010100001	.012574
010110001	.003078
010111101	.006391
010111111	.003492
011100000	.001865
011100001	.007798
011111101	.193263
011111111	.002383
100000001	.001764
100111101	.00108
100111111	.097734
110100001	.005431
110111101	.003251
110111111	.225815
111111101	.001352

* frequencies > .001

```
Performing gradient-based optimization:
note: removing 27 observations; constituent haplotype frequencies are
      smaller than .001
Iteration 0:   Retrospective log likelihood = -6690.1467
Iteration 1:   Retrospective log likelihood = -6658.5547
Iteration 2:   Retrospective log likelihood = -6658.1273
Iteration 3:   Retrospective log likelihood = -6658.1259
Iteration 4:   Retrospective log likelihood = -6658.1259
Haplotype-effects logistic regression
Mode of inheritance: additive                    Number of obs      =      2264
Genetic distribution: Hardy-Weinberg equilib.    Number phased      =       687
Genotype: snp1 snp2 snp3 snp4 snp5               Number unphased    =      1546
          snp6 snp7 snp8 snp9                    Number missing     =        31
                                                 Wald chi2(4)       =     28.60
Retrospective log likelihood = -6658.1259        Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ha~010011101 |     -0.470      0.249    -1.89   0.059      -0.958       0.018
ha~010100000 |      0.267      0.141     1.89   0.058      -0.009       0.542
ha~100111111 |      0.196      0.101     1.95   0.051      -0.001       0.394
ha~110111111 |      0.323      0.071     4.54   0.000       0.184       0.463
------------------------------------------------------------------------------
```

(*output omitted*)

Genome-wide data

- Our earlier example included 9 SNPs comprising a small DNA region, variations in which were statistically associated with the increased risk of lung cancer
- There are about 10 million common SNPs which make up about 90% of variations in human genome
- The International HapMap Consortium (2007) provides over 3.1 million SNPs accounting for about 35% of common SNP variation in human genome
- Can't we somehow use the information available in the whole genome to identify various regions of DNA which could be associated with a disease?
- One way is to perform genome-wide association analysis (e.g., Risch and Merikangas 1996)

Genome-wide association analysis

- Objective: find genetic variations across the whole genome associated with a disease
- Challenge: computationally infeasible to analyze even hundreds of SNPs simultaneously
- Solution: use sliding window approach (e.g., de Bakker et al. 2005)

Sliding windows

- Arrange all SNPs of interest into blocks of a particular size
- Each block of SNPs determines a “window” and the number of SNPs in each block determines the window size
- Test for association within each window to obtain multiple observed significance levels
- Adjust observed significance levels for multiple tests
- Test statistics from adjacent windows are often correlated because of overlapping windows or LD of the constituent SNPs

Adjustments for multiple testing

- Commonly used Bonferroni correction
- Permutation method
- k-FWER (family-wise error rate) method to control the probability of k (≥ 1) or more false positives
- In GWAS, test statistics from adjacent windows are often correlated because of overlapping windows or linkage disequlibrium of the constituent SNPs
- A more powerful alternative for GWAS is a Monte Carlo (MC) method of Huang et al. (2007)
- The MC method is implemented in `gwhaplologit`, currently under development

GWAS of lung-cancer data

- Recall our lung-cancer example
- We consider a version of the data containing 41 SNPs surrounding the region containing two SNPs of interest: rs8034191 (`snp21`) and rs1051730 (`snp28`)
- We use `gwhaplologit` to investigate regions of associations with lung cancer among these 41 SNPs

- Consider single-SNP GWAS first (windows of size 1):

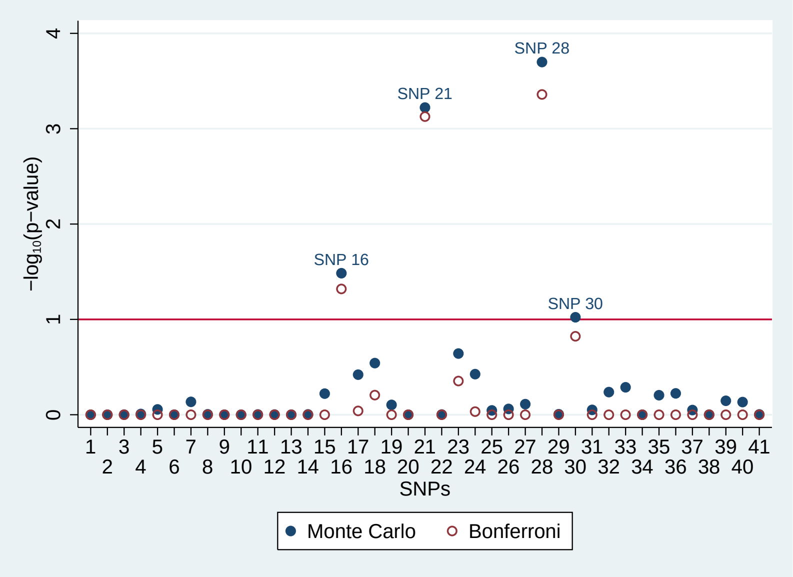
```
. gwhaplologit cancer, snp(snp1-snp41) wsize(1)
Windows (41):
  .........10.........20.........30.........40.
Genomewide association analysis                   Number of windows  =       41
Haplotype-effects logistic regression                      overlap  =        0
Mode of inheritance: additive                     Alpha (FWER)       =       .1
Genetic distribution: Hardy-Weinberg equil.       Number of SNPs     =       41
Haplotype model: main effects                     Number of obs      =     2291
                                                           cases    =     1154
                                                           controls =     1137
```

Windows (1)	P-value, (k=1) Unadjusted	k-FWER	k-FWER-MC	Null model DF	N	LogL
1-1	0.6099	1.0000	0.9996	1	2291	-2223.7770
2-2	0.6103	1.0000	0.9994	1	2291	-2225.1633
3-3	0.5001	1.0000	0.9980	1	2291	-1644.2568
4-4	0.8618	1.0000	0.9820	1	2291	-2163.2535
5-5	0.8739	1.0000	0.8790	1	2291	-2346.2864
6-6	0.4828	1.0000	0.9988	1	2291	-1798.4522
7-7	0.0765	1.0000	0.7324	1	2291	-2145.5205
8-8	0.2867	1.0000	0.9904	1	2291	-2364.8668
9-9	0.6808	1.0000	0.9992	1	2291	-2243.6853
10-10	0.6667	1.0000	0.9996	1	2291	-2159.3543
11-11	0.8296	1.0000	0.9944	1	2291	-2326.8001
12-12	0.5014	1.0000	0.9964	1	2291	-2339.4497
13-13	0.7450	1.0000	0.9988	1	2291	-1777.9610
14-14	0.2801	1.0000	0.9926	1	2291	-2309.4833

(*Continued on next page*)

15-15	0.0487	1.0000	0.6008	1	2291	-1709.3345
16-16*	0.0012	0.0479	0.0328	1	2291	-2148.8787
17-17	0.0222	0.9116	0.3800	1	2291	-2080.2937
18-18	0.0152	0.6223	0.2874	1	2291	-2367.9991
19-19	0.0929	1.0000	0.7880	1	2291	-2235.6978
20-20	0.6062	1.0000	0.9998	1	2291	-1583.0288
21-21*	0.0000	0.0007	0.0006	1	2291	-2278.9731
22-22	0.3541	1.0000	0.9954	1	2291	-1248.6997
23-23	0.0108	0.4429	0.2282	1	2291	-1753.2560
24-24	0.0226	0.9273	0.3752	1	2291	-2291.1795
25-25	0.1446	1.0000	0.9012	1	2291	-2339.4240
26-26	0.1211	1.0000	0.8686	1	2291	-2341.3457
27-27	0.0889	1.0000	0.7746	1	2291	-2337.5105
28-28*	0.0000	0.0004	0.0002	1	2291	-2279.8622
29-29	0.2888	1.0000	0.9878	1	2291	-788.1882
30-30*	0.0037	0.1504	0.0950	1	2291	-1742.0743
31-31	0.1362	1.0000	0.8892	1	2291	-2212.3007
32-32	0.0453	1.0000	0.5788	1	2291	-2238.4966
33-33	0.0363	1.0000	0.5154	1	2291	-1474.4632
34-34	0.4966	1.0000	0.9990	1	2291	-959.7251
35-35	0.0545	1.0000	0.6240	1	2291	-2353.6201
36-36	0.0503	1.0000	0.5970	1	2291	-2349.5156
37-37	0.1344	1.0000	0.8930	1	2291	-1581.0391
38-38	0.7942	1.0000	0.9978	1	2291	-2255.4285
39-39	0.0703	1.0000	0.7140	1	2291	-2347.9133
40-40	0.0756	1.0000	0.7366	1	2291	-2346.1990
41-41	0.3717	1.0000	0.9924	1	2291	-1934.6021

(obs. with constituent haplotypes with frequencies smaller than .001 omitted)
(haplotypes with freq. smaller than .002182 plus most frequent used as reference)
(*) means candidate window according to k-FWER-MC p-value


SNP 28
SNP 21
SNP 16
SNP 30
−log10(p−value)
0
1
2
3
4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
SNPs
Monte Carlo
Bonferroni

- Consider 2-SNP GWAS (windows of size 2) overlapping by one SNP:

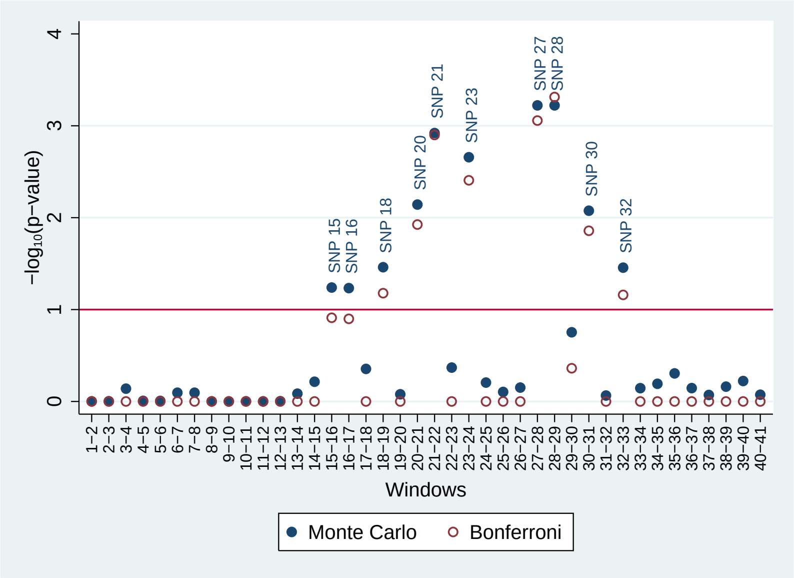
```
. gwhaplologit cancer, snp(snp1-snp41) wsize(2) overlap(1) significant
Windows (40):
  .........10.........20.........30.........40

Genomewide association analysis                   Number of windows  =       40
Haplotype-effects logistic regression                      overlap  =        1
Mode of inheritance: additive                     Alpha (FWER)       =       .1

Genetic distribution: Hardy-Weinberg equil.       Number of SNPs     =       41
Haplotype model: main effects                     Number of obs      =     2291
                                                           cases     =     1154
                                                           controls  =     1137
```

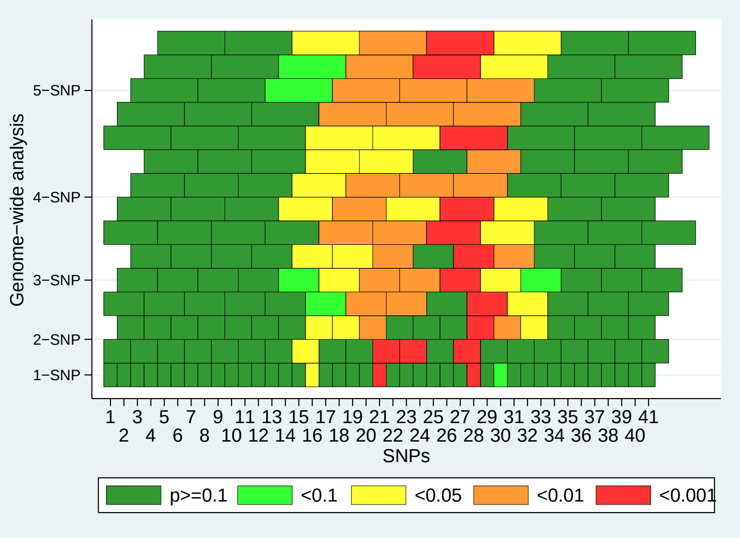
Windows (2)	P-value, (k=1) Unadjusted	k-FWER	k-FWER-MC	Null model DF	N	LogL
15-16*	0.0031	0.1228	0.0576	2	2289	-3691.7850
16-17*	0.0032	0.1261	0.0584	2	2289	-3904.8767
18-19*	0.0017	0.0663	0.0346	3	2291	-4603.6833
20-21*	0.0003	0.0119	0.0072	3	2291	-3794.7572
21-22*	0.0000	0.0013	0.0012	2	2287	-3175.5475
23-24*	0.0001	0.0039	0.0022	2	2289	-3794.9488
27-28*	0.0000	0.0009	0.0006	2	2291	-3860.3080
28-29*	0.0000	0.0005	0.0006	2	2291	-3021.2687
30-31*	0.0003	0.0139	0.0084	2	2290	-3748.7077
32-33*	0.0017	0.0692	0.0350	3	2291	-3627.4546

```
(obs. with constituent haplotypes with frequencies smaller than .001 omitted)
(haplotypes with freq. smaller than .002182 plus most frequent used as reference)
(*) means candidate window according to k-FWER-MC p-value
```



−log10(p−value)
0
1
2
3
4
SNP 15
SNP 16
SNP 18
SNP 20
SNP 21
SNP 23
SNP 27
SNP 28
SNP 30
SNP 32
1−2
2−3
3−4
4−5
5−6
6−7
7−8
8−9
9−10
10−11
11−12
12−13
13−14
14−15
15−16
16−17
17−18
18−19
19−20
20−21
21−22
22−23
23−24
24−25
25−26
26−27
27−28
28−29
29−30
30−31
31−32
32−33
33−34
34−35
35−36
36−37
37−38
38−39
39−40
40−41
Windows
Monte Carlo
Bonferroni

- We can collect MC p-values of sliding window haplotype tests of association for lung-cancer data from `gwhaplologit` for varying window sizes and plot them following the approach of Mathias et al. (2006)



Future work

- Relax gene-environment independence assumption
- Allow multiple genes and gene-gene interactions
- Handle untyped SNPs
- Accommodate population stratification
- Accommodate association tests including interaction effects in GWAS

Acknowledgment

Grant.
This work was supported by the NIH SBIR grant "Statistical Software for Genetic Association Studies" to StataCorp LP.

Consultants.
Christopher I. Amos is a professor of epidemiology at the M. D. Anderson Cancer Research Center.

Raymond J. Carroll is a distinguished professor of statistics, nutrition, and toxicology at Texas A&M University.

Danyu Lin is a Dennis Gillings distinguished professor of biostatistics at the University of North Carolina.

Donglin Zeng is an associate professor of biostatistics at the University of North Carolina.

References

Akey, J., L. Jin, and M. Xiong. 2001. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics* 9: 291–300.

Amos, C. I., X. Wu, P. Broderick, et al. 2008. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics* 40: 616–622.

De Bakker, P. I. W., R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, and D. Altshuler. 2005. Efficiency and power in genetic association studies. *Nature Genetics* 37: 1217–1223.

Breslow, N. E., J. M. Robins, and J. A. Wellner. 2000. On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* 6: 447–455.

Huang, B. E., C. I. Amos, and D. Y. Lin. 2007. Detecting haplotype effects in genomewide association studies. *Genetic Epidemiology* 31: 603–812.

International Hapmap Consortium. 2003. The international HapMap project. *Nature* 426: 789–796.

International Hapmap Consortium. 2005. A haplotype map of the human genome. *Nature* 437: 1299–1320.

International Hapmap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851–862.

International SNP Map Working Group. 2001. A map of human genome sequence variation containing 14.2 million single nucleotide polymorphisms. *Nature* 409: 928–933.

Lake, S., H. Lyon, E. Silverman, S. Weiss, N. Laird, and D. Schaid. 2003. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Human Heredity* 55:56–65.

Lin, D. Y. and D. Zeng. 2006. Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association* 101: 89–118.

Marchenko, Y. V., R. J. Carroll, D. Y. Lin, C. I. Amos, and R. G. Gutierrez. 2008. Semiparametric analysis of case-control genetic data in the presence of environmental factors. *The Stata Journal* 8(3): 305–333.

Mathias, R. A., P. Gao, J. L. Goldstein, A. F. Wilson, E. W. Pugh, P. Furbert-Harris, G. M. Dunson, F. J. Malveaux, A. Togias, K. C. Barnes, T. H. Beaty, and S.-K. Huang. 2006. A graphical assessment of P-values from sliding window haplotype tests of association to identify asthma susceptibility loci on chromosome 11q. *BMC Genetics* 7:38.

Risch, N. and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science* 273:1616–1617.

Schaid, D. J., C. M. Rowland, D. E. Tines, R. M. Jacobson, and G. A. Poland. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* 70:425–434.

Spinka, C., R. J. Carroll, and N. Chatterjee. 2005. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology* 29: 108–127.