More New Sib-pair Stuff (2008)

David Duffy

Genetic Epidemiology Laboratory

Introduction

- Overview of development of Sib-pair
- Recently added procedures
- Sib-pair Monte-Carlo methods
  - Simple gene dropping
  - Conditional gene dropping
  - MCMC genotype simulation
  - MCMC GLMMs
- Sib-pair for GWAS

Overview of Sib-pair

An extensible platform for genetic data manipulation and analysis

A platform for methodological experimentation

First code written in 1995. Now all standard Fortran 95, compiles using multiple compilers.

Creeping featurism has continued to today (55000 lines of code + 7000 lines of comments; 9000 LOC in last 12 months)

The Language

- Simple interpreted language, over 200 commands
- Commands for linkage, association, variance components …
- Offers the usual record-wise operations on data – algebra, logical conditions
- Family-centric data operations – subsetting, pruning etc
- Some elementary databasing type operations – merging, editing
- Flexible data export and scripting to use other programs

Recent Additions 2006-7

Analysis	Sib-pair command
Penrose sib-pair linkage	*penrose*
Univariate twin analysis	*twin*
Genetic survival analysis	*reg wei sim*
Multicategory trait association	*assoc cat*
IBD conditioned gene-dropping	*assoc ibd*
Segregation ratios	*seg*
Predicted genotype distrib.	*ito*
GLMMs/Segregation analysis	*fpm*
Ranking of test results	*summary*
Postscript plotting	*plot*
Language extensions	*eval*

Recent Additions 2007-8

Analysis	**Sib-pair command**
Multiple imputation (genotypes)	*regress rep*
non-MCMC ML genotypic probabilities	*gpe*
Twopoint lod score linkage	*lod*
Multilocus haplotype inference	*dis*
Multilocus haplotype association	*dis <trait>*
Multipoint IBD with markers in LD	*qtl/var*
Haploid marker association analysis	*mit/yhap*
Conditional logistic regression	*sdt/clr*
List pedigree loops	*loo*
Moskvina *et al* N_{eff} markers	*nef*
Yazdi *et al* Weibull heritability	*fpm*

Recent Additions 2007-8

Facility	**Sib-pair command**
Read PLINK *.bed* etc files	*read plink*
Write MENDEL 8.0 files	*wri men new*
Read/write compressed “binary” files	*rea/wri bin*
Some updating/merging of data	*update/hash*
Internals for handling large datasets	
Large regression testing suite	*testsuite.in*
Increased flexibility of output	*set pri*
Full incorporation of MZ twins in analysis	*ass* etc

Multiple imputation for missing genotypes

- Popular approach for complex datasets in nongenetic contexts (Rubin 1987)
- Impute missing data probabilistically so generate multiple (different) versions of data (usually only 5-10)
- Calculate a test statistic (S) and error variance (U) for each dataset as if it was fully observed
- Estimate between-replicate and within-replicate variances
- Calculate corrected Wald test

$$\overline{S} = m^{-1}\Sigma S^{(i)}$$

$$V(\overline{S}) = (1 + m^{-1})[(m-1)^{-1}\Sigma(S^{(i)} - S)^2] + m^{-1}\Sigma U^{(i)}$$

A simple approach to multipoint IBD

- If a set of markers are in complete linkage, each is giving an estimate of IBD for the entire set
- If IBD sharing at any single marker is perfectly known for a relative pair, this is true for the set
- Otherwise we can combine single-point estimates as a variance weighted average for each relative pair
- Approach is computationally cheap, and unaffected by within-set LD
- Automatically carried out by the Sib-pair **qtl** and **ibd** commands if markers are closer than a given threshold

Lod score linkage analysis

Sib-pair uses the iterative peeling approach described by Wang et al [1996] (following Janss et al [1992]).

Briefly, the iterative approach peels up and down simultaneously by calculating anterior and posterior values for each individual, where the anterior and posterior values represent the scaled likelihood contributions for ancestors and siblings, and mates and descendants respectively.

The values for any individual rely on those for the other relatives, so these are reciprocally updated over multiple iterations until they converge. Usually a maximum of 10-20 iterations suffices,

If an ideal ordering of evaluations is used, a single iteration in unlooped pedigrees is sufficient ie it is equivalent to the usual pedigree traversal algorithms

Some quickies: Utilities

```
>> help grr

grr <prev> <pA> <GRR> [<add|dom|rec>] {recurrence risks}
grr <prev> <pCa> <pCo> cas {recurrence risks case-control data}

>> grr 0.01 0.05 0.03 case
```

Single Major Locus Recurrence Risk Calculation

Frequency(A): 0.030200; Pen(AA): 0.027; Pen(AB): 0.016; Pen(BB): 0.010
Trait Prev : 0.010000; Pop AR: 4.0%; Var(Add): 0.000003; Var(Dom): 0.0

Measure	MZ Twin	Sib-Sib	Par-Off	Second
Rec risk	0.010	0.010	0.010	0.010
Rel risk	1.027	1.014	1.014	1.007
Odds rat	1.028	1.014	1.014	1.007
PRR	1.027	1.013	1.013	1.007
ibd|A-A	1.000	0.503	0.500	0.252
ibd|A-U	1.000	0.500	0.500	0.250

Freq of A if Affected: 0.050000 (0.003,0.095,0.902)
Freq of A if Unaffctd: 0.030000 (0.001,0.058,0.941)

Mating	Proportion	Risk to offspring
UnA x UnA	0.980	0.010
Aff x UnA	0.020	0.010
Aff x Aff	0.000	0.010

Some quickies: Scripting

```
>> out zzz

Writing output to “zzz”.

>> grr { 0.01 0.02 0.05 0.10 0.15 } 0.05 0.03 case

>> out

Ending output to “zzz”.

>> file print /Pen/ zzz
 Frequency(A): 0.030200; Pen(AA): 0.027; Pen(AB): 0.016; Pen(BB): 0.010
 Frequency(A): 0.030400; Pen(AA): 0.054; Pen(AB): 0.032; Pen(BB): 0.019
 Frequency(A): 0.031000; Pen(AA): 0.130; Pen(AB): 0.079; Pen(BB): 0.048
 Frequency(A): 0.032000; Pen(AA): 0.244; Pen(AB): 0.153; Pen(BB): 0.096
 Frequency(A): 0.033000; Pen(AA): 0.344; Pen(AB): 0.223; Pen(BB): 0.145
```

Some quickies: database type queries

```
>>  print where agedx2 < 10

Print where “agedx2 < 10”:

!                    s c
!                    e m
! pedigree      id    x m    DOB        yob      dodiag    agedx2
!
002845      020760    f y  19770727 1977.0000  19860411    8.7064
007415      023364    f y     x     1904.0000  19610802   -8.4162
014033      014033    m y     x     1923.0000  19720101    1.9986
050583      050583    m y  19800317 1980.0000  19891213    9.7413
052078      052078    m y  19830316 1983.0000  19900906    7.4771
```

Some quickies: testing DOB

```
>>  test dob DOB greg

Checking for DOB inconsistencies using variable “DOB”.
Threshold for inconsistencies =        4380 days (12.0 years)

Pedigree   Parent ID  Parental DOB Child ID     Child DOB   Diff (yrs)
--------- ---------- ------------ ---------- ------------ ------------
000178    016898       1961-07-26       063611   1961-07-26      +0.00
000369    015727       1904-01-01       000369   1912-02-27      +8.16
000369    015726       1901-01-01       000369   1912-02-27     +11.15
000369    015727       1904-01-01       063359   1913-11-20      +9.89
000369    015727       1904-01-01       063360   1915-10-06     +11.76
000906    023383       1956-09-06       029111   1950-05-21      -6.30
001746    015533       1913-11-23       006821   1924-10-22     +10.91
001758    018324       1925-06-27       001758   1936-09-27     +11.25
…
```

Some quickies: testing sex

```
>> test sex

Pedigree     Individual  Sex  Post.Pr(M)   X-marker hets

04620        0462030      f    1.000000       0/ 19
09901        0990151      f    1.000000       0/ 19
10157        1015704      m    0.000000      14/ 19
23204        2320403      m    0.000000      16/ 18
25122        2512203      m    0.000000      13/ 19
25122        2512204      f    1.000000       0/ 19
29344        2934401      f    1.000000       1/ 19
29344        2934450      m    0.000000      12/ 19
32802        3280203      m    0.000000      14/ 19
32802        3280204      f    1.000000       0/ 19

Designated     Sex inferred via sex-linked markers
Sex          Likely Male    Uncertain   Likely Female
----------   -----------    ---------   -------------
 Male             721          1806             5
 Unknown            0           145             0
 Female             5          2136           811
```

Some quickies: lod score linkage analysis

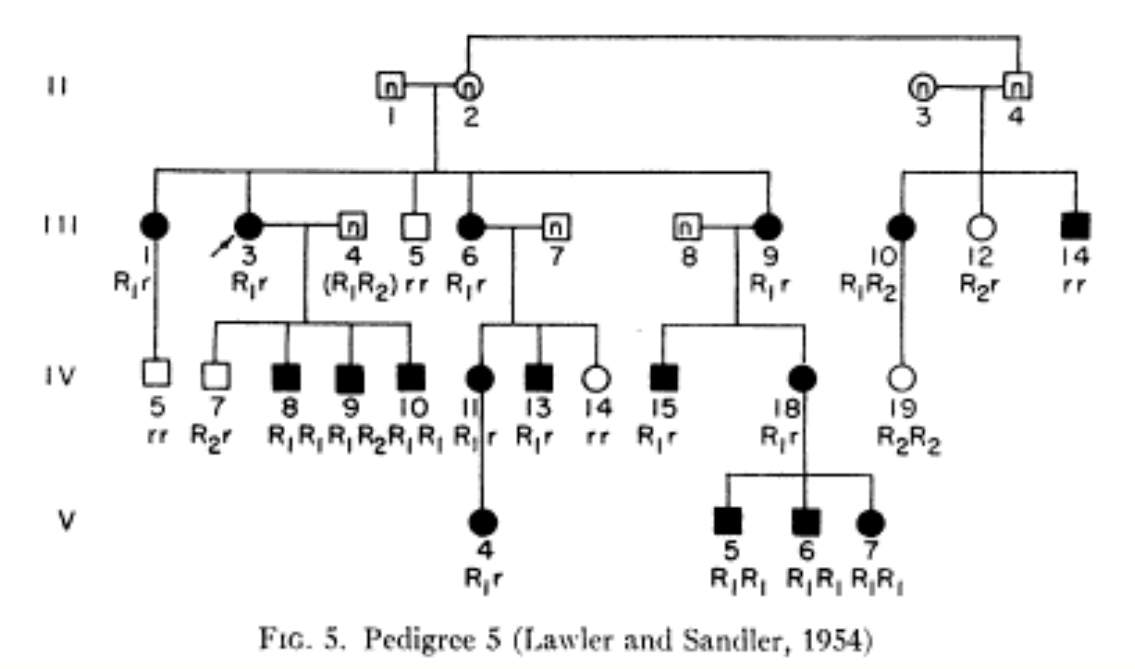


FIG. 5. Pedigree 5 (Lawler and Sandler, 1954)

```
>> set locus traitlocus marker
>> if (ellipto) then traitlocus="1/2"
>> if (not ellipto) then traitlocus="1/1"
>> set freq traitlocus 0.9999 0.0001

NOTE:  The marker "traitlocus" has prespecified allele frequencies:
       1=0.9999 2=0.0001

>> lod traitlocus rhesus
```

```
------------------------------------
Two-point lod score linkage analysis
------------------------------------

NOTE:  Population allele frequencies for “traitlocus” are prespecified as:
       1=0.9999 2=0.0001

“traitlocus” (2 alleles) v. “rhesus” (3 alleles).

 LogLikelihood   LOD        Theta
--------------  ---------  -------
    -59.3092        0.000   0.5000
    -56.2217        1.341   0.0001
    -52.5057        2.955   0.0100
    -51.9104        3.213   0.0250
    -51.7533        3.282   0.0500
    -51.8939        3.220   0.0750
    -52.1640        3.103   0.1000
    -52.9107        2.779   0.1500
    -53.8264        2.381   0.2000
    -55.9736        1.449   0.3000
    -58.2181        0.474   0.4000
```

Some quickies: testing haplotype association

```
>> dis G2215A ACE_ID G2350A hiace

Inter-marker allelic association analysis

Trait:   hiace(2)
Markers: G2215A(2) ACE_ID(2) G2350A(2)

Haplotype  n          y
---------  ---------   ---------
1 1 1      0.0000      0.0000
2 1 1      0.6711      0.2975
1 2 1      0.0000      0.0000
2 2 1      0.0000      0.0000
1 1 2      0.0000      0.0000
2 1 2      0.0000      0.0000
1 2 2      0.3263      0.6994
2 2 2      0.0026      0.0032

hiace         190         158

            Number of loci =         3
 No. genotyped individuals =       348
 No. obs. unique genotypes =         8
  Stratified LD Chi-square =     18.26 (df=  48, P=1.0000)
    Association Chi-square =     98.91 (df=   2, P=0.0000)

NOTE:  Degrees of freedom calculation for association test assumes only
          3 haplotypes to be present in the population.
```

Sib-pair Monte-Carlo procedures

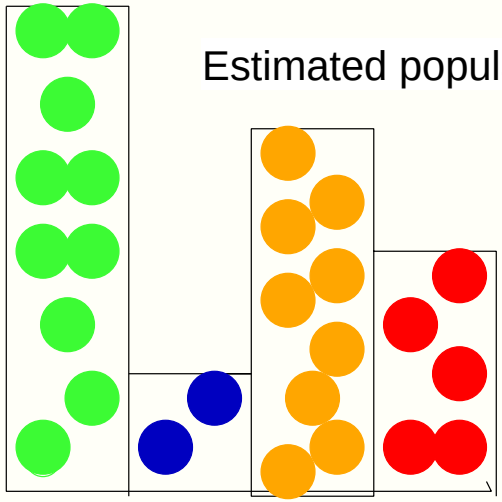
If some key variables are unobserved, analysis is complicated.

If all key variables are observed, analysis is often simple

Simulation allows us to approximate the distributions of unobserved variables by averaging over repeated simple analyses of data that has been “filled in” using an appropriate mechanism (random or quasirandom)

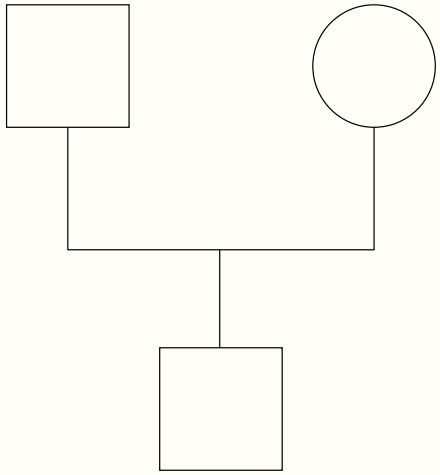
- Estimation of model parameters
- Assessment of statistical significance
- Assessment of statistical power

Gene dropping 1

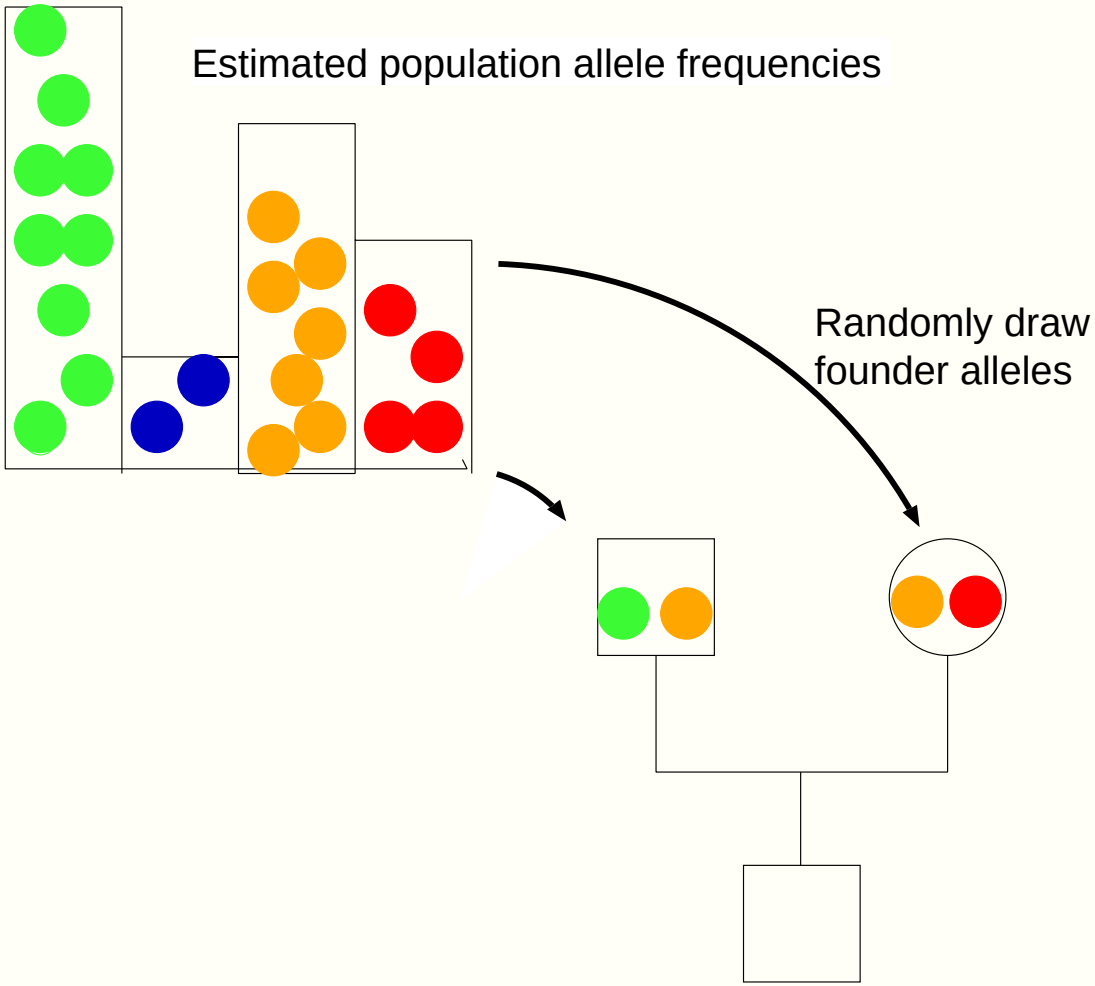


Estimated population allele frequencies

Pedigree for simulation

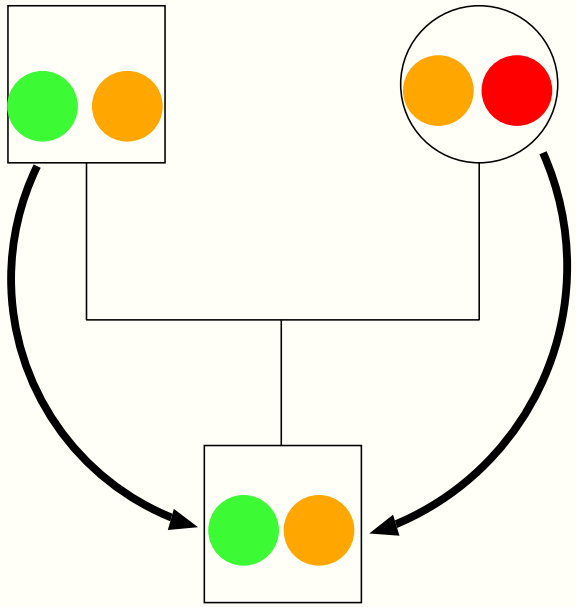


Gene dropping 2

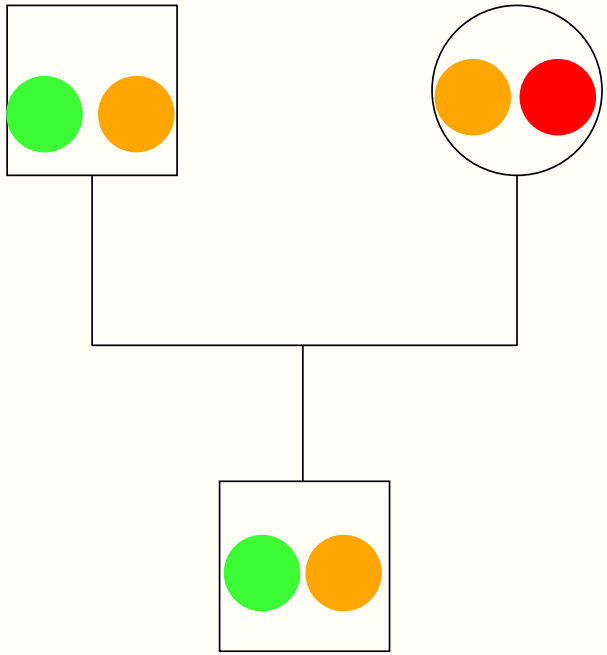


Gene dropping 3

Randomly draw one allele from each parent



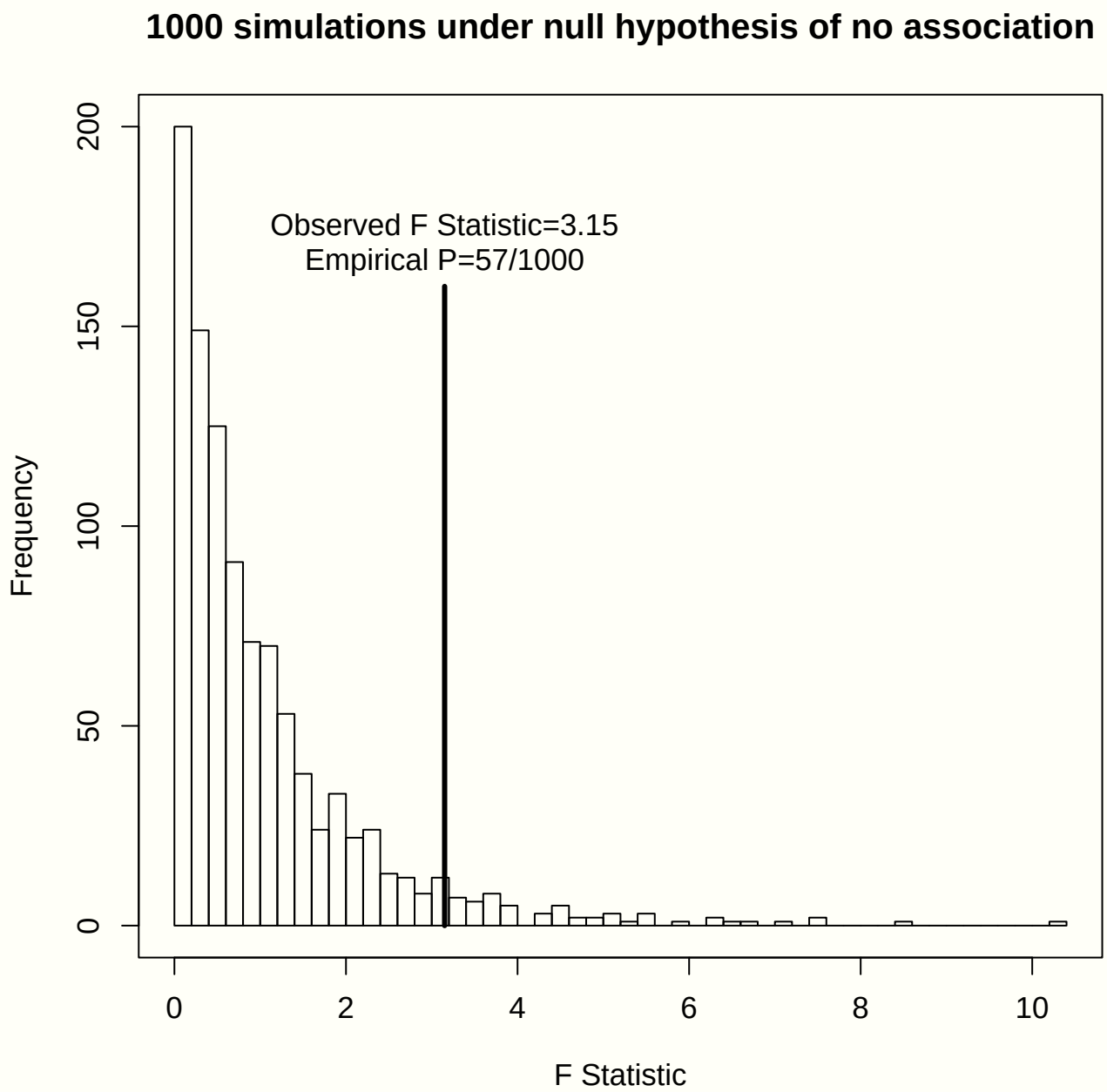
Gene dropping 4



Y=14.5

Calculate test statistic given the simulated genotypes

$$Y = \beta_G G_i + \beta_{PG} PG_i$$


1000 simulations under null hypothesis of no association
Observed F Statistic=3.15
Empirical P=57/1000
Frequency
0
50
100
150
200
0
2
4
6
8
10
F Statistic

Gene-Dropping: Sib-pair applications

Gene-dropping based P-values are calculated for:

Analysis	**Sib-pair command**
Marker HWE test	*hwe*
Marker homozygosity test	*homoz*
Binary trait association	*assoc*
Binary trait FBAT	*assoc*
Binary trait TDT	*tdt*
Binary trait linkage	*apm*
Quantitative trait association	*assoc*
Quantitative trait TDT	*assoc* / *tdt*
Quantitative trait linkage	*sib simulate*

MCMC methods in Sib-pair

Markov Chain Monte-Carlo is the most commonly used MC algorithm for complex genetic simulation.

- MCMC for exact contingency table analysis in Sib-pair
- MCMC simulation of unobserved marker genotypes in Sib-pair
- MCMC GLMMs in Sib-pair

A Segue into Monte-Carlo Markov Chains

Simple Monte Carlo simulation augmented by rejection of inappropriate samples (**rejection sampling**) becomes inefficient if there are too many side conditions. For a pedigree with many missing genotypes, but also many nonmissing genotypes, the efficiency can be worse than 1 accepted sample per 10^7.

Markov Chain Monte-Carlo is considerably more efficient, if it can be implemented for the problem.

Sampler	**Proposal mechanism**	**Filter**	**Samples**
Rejection	Gene-drop *every* possibility	Rejection	Independent
MCMC	Proposal based on last *sample*	Surrogate LR	Correlated

MCMC for exact contingency table analysis in Sib-pair

$X_1 \leftrightarrow X_2 \leftrightarrow X_3$

	2	3
3	**0**	3
2	2	0

	2	3
1		2
1		1

	2	3
2		1
0		2

A Monte Carlo Markov Chain can be constructed that moves one step at a time between all the legal tables, and the proportion of samples of each table represents the probability of that table.

Proposal: choose two rows (*xorigin* and *xdestination*) and two columns (*yorigin* and *ydestination*). Change the counts (a,b,c,d) by {+1,-1,-1,+1}.

Filter: Calculate the ratio of probabilities of present table to proposed table ($\frac{ad}{(b+1)(c+1)}$). This ratio is the Metropolis criterion (q).

If $q \geq 1$, always accept the new sample proposal; otherwise accept the new proposal q proportion of times, or keep the old table as the new sample.

MCMC simulation of unobserved marker genotypes

We can try and perform a similar trick to visit every legal table of the missing genotypes for a pedigree. The Metropolis criterion is easy to calculate, being the ratio of the likelihoods for the *changed* genotypes:

- founders: the frequency of the genotype in the population
- nonfounders: the probability that genotype would be transmitted from her parents.

The difficult part is defining a proposal mechanism that will eventually visit **every** legal possible genotype, without having a large number of rejections (non-Mendelian proposals).

For diallelic markers, Lange and Matthysse (1989) described a correct method that is fairly efficient.

For multiallelic markers, there is no simple proposal method, but a variety of work-arounds are in use (in programs such as SIMWALK2, LOKI, and MORGAN).

MCMC simulation of unobserved marker genotypes in Sib-pair

Sib-pair uses a mixture of methods to try and guarantee that all legal constellations of genotypes will be sampled. It successfully simulates the genotypes for the "bad" pedigrees of Jensen and Sheehan (1998).

Analysis	**Sib-pair command**
Marker allele frequencies	*mcf mcmc*
Binary trait linkage	*apm ibd*
Quantitative trait linkage	*qtl full*
GLMM variance components	*fpm*

The MCMC estimation of single-marker *IBD* by Sib-pair is considerably faster than MERLIN for large pedigrees.

The MCMC ML estimates of marker allele frequencies by Sib-pair compare well to those estimated by MENDEL.

Association analysis using MCMC for missing genotyped

So we can infer the distributions of the values of unobserved genotypes using genotyped relatives. If trait values were available but the genotype was missing, we have increased the usable data.

MENDEL carries this out using the usual ML methods, and Sib-pair can now do this using MCMC multiple imputation.

```
>> set analysis imputed
>> regress onset = marker1 weibull trait rep 20
```

MCMC GLMMs in Sib-pair: introduction

A number of often-used likelihood models can be represented in a single simple form, as exponential families. These **generalized linear models** take the form of a linear regression, where the usual *y* variable is replaced by a linear predictor, a function of the observed dependent variable. So for a binary outcome, for example, the **link function** might be the **logistic** or **probit.**

Just as the usual linear model can be extended to give a variance components model or a mixed model, the GLM can be extended to a **generalized linear mixed model**.

The multifactorial threshold model is very close to being a probit link GLMM, and points out one difficulty of these models, that integration over the unobserved random effects must be performed.

It is possible to design a Markov chain that simulates these integrated likelihoods.

MCMC GLMMs in Sib-pair: latent variables

In the Sib-pair MCMC GLMM, the simulated (unobserved) variables include:

- Diallelic QTL genotypes
- Gaussian breeding values
- Maternal effect values
- Family environmental effect values
- A single QTL allele frequency (shared by all QTLs in the major gene or finite polygenic model)
- Up to three genotypic means (shared by all QTLs in the FPM)
- V_A, V_C, V_M

MCMC GLMMs in Sib-pair: link functions

The trait model can be:

- Gaussian
- Binomial with identity, probit or logit link
- Multifactorial threshold model
- Poisson (log link)
- Weibull

MCMC GLMMs in Sib-pair: proposal and transition

The proposal mechanisms for each variable:

- Diallelic QTL genotypes: Lange and Matthysse (1989)
- Breeding values etc: random normal deviates

The likelihood contribution from the i^{th} individual to the Metropolis criterion for these models is (Guo and Thompson 1993):

$$LL = F\Sigma(\log(P(G_j))) + F * \log(f(a|V_A)) + (1-F) * \log(f(a|a_{FA}, a_{MO})) + \log(c|V_C) + I * \log(f(y|G_1, \ldots G_j, a, c, V_E))$$

MCMC GLMMs in Sib-pair: pieces of the likelihood

$P(x)$ denotes the probability of x,

$f(x)$ denotes the density of x,

y is the trait value,

a is the breeding value,

c is the pedigree-specific intercept,

G_j is the genotype at the j^{th} QTL,

V_A is the additive polygenic variance,

V_C is the familial environmental variance,

V_E is the error variance,

$F = 1$ when a founder, 0 when a nonfounder

$I = 1$ when phenotype observed, 0 when unobserved.

The conditional density for the breeding values of offspring includes the correction for inbreeding (the segregation variance being $1 - \frac{1}{2}(F_{FA} + F_{MO})$ (cf Guo and Thompson).

Data augmentation in Sib-pair: replication of pedigrees

There are often several latent variables being estimated for each individual in the dataset.

I have found that a simple replication of the data, so global parameters are shared but random effects are estimated by parallel correlated chains, significantly reduces upward bias in variance component estimates. This method is superior to even very long MCMC runs on the original pedigree data.

MCMC GLMMs in Sib-pair: summarizing the simulations

As opposed to the case of rejection sampling, the simulated samples from a MCMC are not independent. Therefore the effective number of samples is a lot smaller than the observed number.

Batching is one method the Monte Carlo error variances for summary statistics from correlated samples. Sib-pair summarizes parameters as means of the entire chain of simulated values, and the associated MC standard errors as the standard deviation of $\sqrt{B}$ subsample means (Jones et al 2005).

An alternative approach is **thinning**, where one retains only every N^{th} sample. The contingency table P-value estimator in Sib-pair sets N to the number of observations in the table.

The modes for the parameter values simulations are nonparametric MLEs derived from thinned samples using the algorithm of Meyer (2001).

MCMC GLMMs in Sib-pair: commands

Contingency table analysis	*table*
GLMM polygenic models	*fpm nqtl 0*
GLMM segregation models	*fpm nqtl 1*
Finite polygenic models	*fpm nqtl N*

MCMC GLMMs in Sib-pair: a simple genetic example

Binomial GLMM analysis of rat toxicology dataset of Weil et al (1970) using different approaches. PQL1 is the penalised quasilikelihood approach implemented as **glmmPQL()** in the MASS package [Venables and Ripley 2002], while PQL2, Laplace are results from **lmer()** in the lme4 package of Bates and Sarkar [2005] using penalized quasilikelihood, Laplace approximation respectively. The SAS results are from Wang and Louis [2002].

	Parameter Estimate (SE)				
Method	Sib-pair	Laplace	PQL1	PQL2	SAS NLMIXED
SD Litter RE	1.40 (0.33)	1.30	1.27	1.49	1.34 (0.33)
Intercept	2.64 (0.50)	2.63 (0.45)	2.37 (0.41)	2.37 (0.48)	2.62 (0.48)
Treatment	-1.10 (0.64)	-1.09 (0.60)	-0.96 (0.56)	-0.96 (0.66)	-1.07 (0.62)

MCMC GLMMs in Sib-pair: a nongenetic example

Poisson GLMM analysis of European male melanoma death rate dataset of Langford et al (1998) using different approaches. PQL1 is the penalised quasilikelihood approach implemented as **glmmPQL()** by Ripley and Venables [2002] in the MASS package, while PQL2, AGQ are results from **lmer()** in the lme4 package of Bates and Sarkar (2005). The STATA result used the **gllamm** command, and comes from the review article at *http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-software/gllamm.shtml.*

	Parameter Estimate (SE)				
Method	Sib-pair	AGQ	PQL1	PQL2	GLLAMM
Region variance	0.172 (0.032)	0.170 (-)	0.161	0.125	0.170 (0.031)
Intercept	-0.156 (0.057)	-0.139 (0.049)	-0.129 (0.049)	-0.129 (0.043)	-0.139 (0.049)
UVB insolation	-0.037 (0.010)	-0.034 (0.010)	-0.038 (0.010)	-0.038 (0.009)	-0.034 (0.010)

MCMC GLMMs in Sib-pair: a twin survival analysis

Using Sib-pair to analyse a GWAS

I have spent a bit of time over the last 1-2 years doing some optimization of the code so that it is not too onerous to use Sib-pair in the analysis of large datasets.

SNP genotype data can be stored internally as 4 bits per genotype, so that large datasets fit into memory. Even the default format for storing genotype is now 4 times smaller than it was.

A binary image of a dataset can be saved and reread from disk. This is much quicker than reading in the original locus and pedigree files. The image is compressed (*gzip*).

The **summary** command allows one to rank and subset out only the results of interest from a large set of tests. This can also generate a Postscript plot, or a *.WIG* file for the UCSC browser. The **keep** and **drop** commands allow one to select loci based on Hardy-Weinberg disequilibrium or allele frequencies.

Sib-pair compared to PLINK: Making a binary file

Creating a binary file for subsequent analysis:

PLINK (35 seconds):

```
plink –noweb –file wgas1 –make-bed –out wgas2
```

Sib-pair (2 minutes 28 seconds):

```
read loc plink wgas1.map
read ped wgas1.ped
set che off
set imp -1
set ple -1
run
write bin wgas1.bin compress
```

The space taken by the resulting files:

```
-rw-r–r– 1 davidD davidD 2.2k Oct 24 13:46 wgas2.fam
-rw-r–r– 1 davidD davidD 7.8M Oct 24 13:46 wgas2.bim
-rw-r–r– 1 davidD davidD 5.3M Oct 24 13:46 wgas2.bed
-rw-r–r– 1 davidD davidD 9.5M Oct 24 13:50 wgas1.bin.gz
```

Sib-pair compared to PLINK: Allele frequencies

Estimating allele frequencies:

PLINK (9 seconds):

```
plink –noweb –bfile wgas2 –freq –out freq1
```

CHR	SNP	A1	A2	MAF	NCHROBS
1	rs3094315	G	A	0.1236	178
1	rs6672353	A	G	0.005618	178
1	rs4040617	G	A	0.1167	180
1	rs2905036	0	T	0	180
1	rs4245756	0	C	0	180
1	rs4075116	C	T	0.05556	180
1	rs9442385	T	G	0.3933	178
1	rs6603781	0	G	0	178

…

Sib-pair compared to PLINK: Allele frequencies

Estimating allele frequencies:

Sib-pair (14 seconds):

```
read bin wgas1.bin; fre snp
```

OR (15 seconds):

```
read plink wgas2; fre snp
```

```
Marker     NAll  Allele(s)  Freq   Het    Ntyped
rs3094315     2     G (A)   0.1236 0.2179     89 792429 (chr 1)
rs6672353     2     A (G)   0.0056 0.0112     89 817376 (chr 1)
rs4040617     2     G (A)   0.1167 0.2073     90 819185 (chr 1)
rs2905036     1     T       1.0000  -         90 832343 (chr 1)
rs4245756     1     C       1.0000  -         90 839326 (chr 1)
rs4075116     2     C (T)   0.0556 0.1055     90 1043552 (chr 1)
rs9442385     2     T (G)   0.3933 0.4799     89 1137258 (chr 1)
rs6603781     1     G       1.0000  -         89 1198554 (chr 1)
…
```

Sib-pair compared to PLINK: HWE

Testing Hardy-Weinberg equilibrium

PLINK (22 seconds):

```
plink –noweb –bfile wgas2 –hardy –out hwe1
```

CHR	SNP	TEST	A1	A2	GENO	O(HET)	E(HET)	P
1	rs3094315	G	G	ALL	0/22/67	0.2472	0.2166	0.3476
1	rs3094315	G	G	AFF	0/15/33	0.3125	0.2637	0.5771
1	rs3094315	G	G	UNAFF	0/7/34	0.1707	0.1562	1
1	rs6672353	A	A	ALL	0/1/88	0.01124	0.01117	1
1	rs6672353	A	A	AFF	0/1/48	0.02041	0.0202	1
1	rs6672353	A	A	UNAFF	0/0/40	0	0	1
1	rs4040617	G	G	ALL	0/21/69	0.2333	0.2061	0.5994
1	rs4040617	G	G	AFF	0/14/35	0.2857	0.2449	0.5714
1	rs4040617	G	G	UNAFF	0/7/34	0.1707	0.1562	1
1	rs2905036	0	0	ALL	0/0/90	0	0	1

…

Sib-pair compared to PLINK: HWE

Testing Hardy-Weinberg equilibrium. Note that Sib-pair calculates two tests of HWE for each SNP (Chi-square and exact test), but only prints the exact P-value here. The usual Sib-pair HWE Chi-square test uses founders and nonfounders, and gene-drops a correct P-value.

Sib-pair (56 seconds):

```
read bin wgas1.bin
set iter 0
hwe
select trait
hwe
unselect
select not trait
hwe
```

Marker	Typed	Genos	Chi-square	Asy P	Emp P	Iters	
rs3094315	89	3	3.1	0.3476	1.0000	0	HWE .
rs6672353	89	3	0.0	1.0000	1.0000	0	HWE .
rs4040617	90	3	2.8	0.5994	1.0000	0	HWE .
rs2905036	90	1	0.0	1.0000	1.0000	0	HWE .
rs4245756	90	1	0.0	1.0000	1.0000	0	HWE .
rs4075116	90	3	0.6	1.0000	1.0000	0	HWE .
rs9442385	89	3	2.1	0.1815	1.0000	0	HWE .
rs6603781	89	1	0.0	1.0000	1.0000	0	HWE .
rs11260562	88	3	0.1	1.0000	1.0000	0	HWE .

…

Sib-pair compared to PLINK: filtering

Filtering individuals and markers:

PLINK (16 seconds):

```
plink –bfile wgas2 –maf 0.01 –geno 0.05 –mind 0.05 –hwe 1e-3 –make-bed
–out wgas3
```

```
Before frequency and genotyping pruning, there are 228694 SNPs
90 founders and 0 non-founders found
Writing list of removed individuals to [ wgas3.irem ]
1 of 90 individuals removed for low genotyping ( MIND > 0.05 )
74 markers to be excluded based on HWE test ( p <= 0.001 )
        65 markers failed HWE test in cases
        74 markers failed HWE test in controls
Total genotyping rate in remaining individuals is 0.995473
2728 SNPs failed missingness test ( GENO > 0.05 )
46834 SNPs failed frequency test ( MAF < 0.01 )
After frequency and genotyping pruning, there are 179493 SNPs
After filtering, 48 cases, 41 controls and 0 missing
After filtering, 44 males, 45 females, and 0 of unspecified sex
```

Sib-pair compared to PLINK: filtering

Filtering individuals and markers. In Sib-pair, **select** and **unselect** are for individuals, and **keep** and **drop** are for loci. The PLINK filters are applied independently of each other, while the Sib-pair filtering acts sequentially.

Sib-pair (35 seconds):

```
read bin wgas1.bin
select where numtyp <= 217259
show ped
unselect
select where numtyp > 217259
select not trait
set iter 0
drop where hwe 0.001
unselect
drop where max 0.99
keep where num 85
write bin wgas3.bin compress
```

```
Permanently deleted 48459 loci.

Reread   89 pedigrees, 89 individuals (5.06 s).
Dataset occupies 64.170 Mb.
```

Sib-pair compared to PLINK: association

Simple association analysis

PLINK (13 seconds):

```
plink –noweb –bfile wgas3 –assoc –adjust –out assoc1
...
Writing main association results to [ assoc1.assoc ]
Computing corrected significance values (FDR, Sidak, etc)
Genomic inflation factor (based on median chi-squared) is 1.25937
Mean chi-squared statistic is 1.22972
Correcting for 179493 tests
Writing multiple-test corrected significance values to [
assoc1.assoc.adjusted
]
```

OR (55 seconds)

```
plink –noweb –bfile wgas3 –logistic logistic1
```

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR
1	rs3094315	792429	G	0.1489	0.08537	A	1.684	0.1944	1.875
1	rs4040617	819185	G	0.1354	0.08537	A	1.111	0.2919	1.678
1	rs4075116	1043552	C	0.04167	0.07317	T	0.8278	0.3629	0.5507
1	rs9442385	1137258	T	0.3723	0.4268	G	0.5428	0.4613	0.7966
1	rs11260562	1205233	A	0.02174	0.03659	G	0.3424	0.5585	0.5852
1	rs6685064	1251215	C	0.3854	0.439	T	0.5253	0.4686	0.8013
...									

CHR	SNP	BP	A1	TEST	NMISS	ODDS	STAT	P
1	rs3094315	792429	G	ADD	88	2.061	1.381	0.1672
1	rs4040617	819185	G	ADD	89	1.804	1.12	0.2629
1	rs4075116	1043552	C	ADD	89	0.5303	-0.9272	0.3538
1	rs9442385	1137258	T	ADD	88	0.8197	-0.687	0.4921
1	rs11260562	1205233	A	ADD	87	0.5758	-0.5877	0.5567
1	rs6685064	1251215	C	ADD	89	0.8409	-0.6398	0.5223
...								

Sib-pair compared to PLINK: association

Simple association analysis

Sib-pair (35 seconds):

```
read bin wgas3.bin
set iter 0
ass trait
summary 20
```

OR (2 minutes 35 seconds):

```
read bin wgas3.bin
set iter 0
ass trait snp
summary 20
```

```
Marker       Typed  Allels Chi-square Asy P  Emp P  Iters
----------   ------ ------ ---------- ------ ------ ------
rs3094315        88      2        1.7 0.1944 1.0000      0 AssX2-HWE .
rs4040617        89      2        1.1 0.2919 1.0000      0 AssX2-HWE .
rs4075116        89      2        0.8 0.3629 1.0000      0 AssX2-HWE .
rs9442385        88      2        0.5 0.4613 1.0000      0 AssX2-HWE .
rs11260562       87      2        0.3 0.5585 1.0000      0 AssX2-HWE .
rs6685064        89      2        0.5 0.4686 1.0000      0 AssX2-HWE .
...
```

```
Total number of tests = 180235

Locus      Position   P-value  -log10(P)
---------- ---------- ------- ----------
rs2513514       75.92  0.0000      6.329  75922141 (chr 11)
rs6110115       13.91  0.0000      6.149  13911728 (chr 20)
rs2508756       75.92  0.0000      5.677  75921549 (chr 11)
rs16976702      54.12  0.0000      5.661  54120691 (chr 15)
rs11204005      12.90  0.0000      5.103  12895576 (chr 8)
rs16910850      94.48  0.0000      4.915  94478347 (chr 9)
rs1195747      129.97  0.0000      4.846  129970575 (chr 12)
rs7207095       77.93  0.0000      4.774  77933018 (chr 17)
rs16971118      77.67  0.0000      4.720  77672467 (chr 15)
rs6074704       14.12  0.0000      4.696  14115283 (chr 20)
rs1570484       14.14  0.0000      4.696  14139687 (chr 20)
rs9944528       77.89  0.0000      4.664  77894039 (chr 17)
rs636006        32.43  0.0000      4.642  32426349 (chr 3)
...
```

Marker	OR	95% CI	P-value	
rs3094315	2.061	0.739 – 5.749	0.8361E-01	792429 (chr 1)
rs4040617	1.804	0.642 – 5.068	0.1314	819185 (chr 1)
rs4075116	1.886	0.493 – 7.207	0.1769	1043552 (chr 1)
rs9442385	1.220	0.692 – 2.151	0.2460	1137258 (chr 1)
rs11260562	1.737	0.276 – 10.948	0.2784	1205233 (chr 1)
rs6685064	1.189	0.699 – 2.022	0.2612	1251215 (chr 1)
...				

```
Total number of tests = 180235

Locus       Position   P-value  -log10(P)
----------- ---------- ------- ----------
rs1548299         3.64  0.0000      5.263  3640174 (chr 9)
rs2513514        75.92  0.0000      5.187  75922141 (chr 11)
rs6110115        13.91  0.0000      4.841  13911728 (chr 20)
rs2508756        75.92  0.0000      4.753  75921549 (chr 11)
rs16976702       54.12  0.0000      4.711  54120691 (chr 15)
rs11204005       12.90  0.0000      4.474  12895576 (chr 8)
rs9302779         3.87  0.0000      4.387  3873346 (chr 16)
rs17534370       70.30  0.0000      4.341  70297172 (chr 9)
rs11781505      142.00  0.0001      4.266  142002911 (chr 8)
rs11785430      142.08  0.0001      4.266  142078709 (chr 8)
rs1195747       129.97  0.0001      4.146  129970575 (chr 12)
rs16910850       94.48  0.0001      4.126  94478347 (chr 9)
rs1570484        14.14  0.0001      4.124  14139687 (chr 20)
rs6074704        14.12  0.0001      4.124  14115283 (chr 20)
...
```

Sib-pair compared to PLINK: association

Stratified association analysis

PLINK (14 seconds):

```
plink –noweb –bfile wgas3 –mh –within pop.cov –adjust –out cmh1


...
Cochran-Mantel-Haenszel 2x2xK test, K = 2
Writing results to [ cmh1.cmh ]
Computing corrected significance values (FDR, Sidak, etc)
Genomic inflation factor (based on median chi-squared) is 1.0147
Mean chi-squared statistic is 0.998899
Correcting for 179493 tests
Writing multiple-test corrected significance values to [ cmh1.cmh.adjusted
]
```

CHR	SNP	A1	A2	BP	CHISQ	P	OR	L95	U95
1	rs3094315	G	A	792429	1.047	0.3062	1.887	0.5668	6.285
1	rs4040617	G	A	819185	0.8534	0.3556	1.777	0.5288	5.974
1	rs4075116	C	T	1043552	1.339	0.2472	0.2929	0.04513	1.901
1	rs9442385	T	G	1137258	0.4069	0.5236	0.7773	0.363	1.665
1	rs11260562	A	G	1205233	0.04184	0.8379	1.27	0.1455	11.08
1	rs6685064	C	T	1251215	1.994	0.1579	1.804	0.7988	4.074

...

Sib-pair compared to PLINK: association

Stratified association analysis

Sib-pair (2 minutes 35 seconds):

```
read bin wgas3.bin
set locus group qua
update popcov.dat
set iter 0
ass trait cov group snp
summary 20
```

OR

```
read bin wgas3.bin
set locus japan aff
japan=n
select ped JA*
japan=y
unselect
set iter 0
ass trait cov japan snp
summary 20
```

Marker	OR	95% CI	P-value	
rs3094315	2.061	0.739 – 5.749	0.8361E-01	792429 (chr 1)
rs4040617	1.804	0.642 – 5.068	0.1314	819185 (chr 1)
rs4075116	1.886	0.493 – 7.207	0.1769	1043552 (chr 1)
rs9442385	1.220	0.692 – 2.151	0.2460	1137258 (chr 1)
rs11260562	1.737	0.276 – 10.948	0.2784	1205233 (chr 1)
rs6685064	1.189	0.699 – 2.022	0.2612	1251215 (chr 1)
...				

```
Total number of tests = 180235

Locus       Position   P-value   -log10(P)
----------- ---------- -------  ----------
rs1548299         3.64  0.0000       5.263  3640174 (chr 9)
rs2513514        75.92  0.0000       5.187  75922141 (chr 11)
rs6110115        13.91  0.0000       4.841  13911728 (chr 20)
rs2508756        75.92  0.0000       4.753  75921549 (chr 11)
rs16976702       54.12  0.0000       4.711  54120691 (chr 15)
rs11204005       12.90  0.0000       4.474  12895576 (chr 8)
rs9302779         3.87  0.0000       4.387  3873346 (chr 16)
rs17534370       70.30  0.0000       4.341  70297172 (chr 9)
rs11781505      142.00  0.0001       4.266  142002911 (chr 8)
rs11785430      142.08  0.0001       4.266  142078709 (chr 8)
rs1195747       129.97  0.0001       4.146  129970575 (chr 12)
rs16910850       94.48  0.0001       4.126  94478347 (chr 9)
rs1570484        14.14  0.0001       4.124  14139687 (chr 20)
rs6074704        14.12  0.0001       4.124  14115283 (chr 20)
...
```

Sib-pair compared to PLINK: filtering

Extracting a single SNP

PLINK (3 seconds):

```
plink –noweb –bfile wgas3 –recode –snp rs11204005 –out tophit
```

Sib-pair (5 seconds):

```
read bin wgas3.bin
keep trait rs11204005
write sib rs11204005.ped
write locus sib rs11204005.in rs11204005.ped
```

Sib-pair compared to PLINK: logistic regression

Logistic regression

PLINK (0.007 seconds):

```
plink –noweb –file tophit –logistic –covar pop.cov
```

```
 CHR          SNP       BP   A1   TEST   NMISS    ODDS    STAT          P
   8   rs11204005 12895576    A    ADD      89 0.06667   -4.33  1.489e-05
   8   rs11204005 12895576    A   COV1      89   79.15    4.68  2.871e-06
```

Sib-pair compared to PLINK: logistic regression

Logistic regression

Sib-pair (0.17 seconds):

```
set locus trait                 affection          .
set locus rs11204005            marker          12.895600 12895576 (chr 8)
read pedigree rs11204005.ped
run
set loc japan aff
japan = n
sel ped J*
japan=y
uns
reg trait = rs11204005 japan
ass trait snp cov japan
```

```
-----------------------------------------------
Binomial regression analysis of trait “trait”
-----------------------------------------------
    Variable          Beta    Stand Error         t-Value
  ---------------------------------------------------------
  Intercept        -4.5510         0.9950          4.5737 ***
  rs11204005        2.7080         0.6254          4.3303 ***
  japan             4.3713         0.9341          4.6799 ***

No. usable observations =     89      (100.0%)
Number of affecteds     =     48

Null deviance           =    122.8291
Number of iterations    =      7
Model LR Chi-square     =     64.5816 (df=   2)
Akaike Inf. Criterion   =     64.2474

-----------------------------------------------
Allelic association testing for trait “trait”
-----------------------------------------------
NOTE:  Covariates are: “japan”.

Marker             OR       95% CI               P-value
-------------- -------- -------------------- ----------
rs11204005       15.000    7.264 –   30.976 0.4975E-02 12895576 (chr 8)
```

Sib-pair compared to PLINK: logistic regression

Interaction within logistic regression

PLINK

```
plink –file tophit –logistic –covar pop.cov –interaction
```

```
...
 CHR         SNP         BP A1       TEST   NMISS     ODDS      STAT         P
   8 rs11204005 12895576  A        ADD      89   0.2918    -0.645    0.5189
   8 rs11204005 12895576  A       COV1      89    319.1     2.655  0.007937
   8 rs11204005 12895576  A  ADDxCOV1      89   0.3366   -0.7811    0.4348
```

Sib-pair compared to PLINK: logistic regression

Interaction within logistic regression

Sib-pair. This is not a canned procedure in Sib-pair, but is easily automated.

```
# Genotypic analysis interaction
macro interaction
 llm %1 %2 %3 %1*%2 %1*%3 %2*%3
;;;;
```

```
interaction trait rs11204005 japan
...
Model: Intercept  trait(2)  rs11204005(3)  japan(2)  trait(2)*rs11204005(3)
       trait(2)*japan(2) japan(2)*rs11204005(3)

    Term                          Beta  Stand Error     Exp(Beta)     t-Value
-------------------------------------------------------------------------------
    Intercept                   2.3920       0.3016        10.935       7.931 ***
    trait(2)                   -5.1292       1.2687         0.006       4.043 ***
    rs11204005(2)               0.5595       0.3774         1.750       1.483 +
    rs11204005(3)              -1.0235       0.5841         0.359       1.752 +
    japan(2)                   -0.5895       0.5025         0.555       1.173 .
    trait(2)*rs11204005(2)      3.5295       1.1854        34.108       2.978 *
    trait(2)*rs11204005(3)      5.7167       1.3889       303.896       4.116 ***
    trait(2)*japan(2)           4.6967       1.1258       109.584       4.172 ***
    japan(2)*rs11204005(2)     -2.5075       1.1266         0.081       2.226 +
    japan(2)*rs11204005(3)     -3.4291       1.2858         0.032       2.667 *

 No. of complete observations =    89 ( -0.0%)

                   Model LRTS =     0.30
           Degrees of freedom =     2
              Nominal P-value =     0.8598
```

```
# Allelic additive analysis interaction macro
macro interaction
 macro p <- ple
 set ple -2 silent
 macro a <- alleles %2
 out %% silent
 doanalysis %1 %2 %a %3
 out
 echo
 echo Interaction analysis trait=%1 marker=%2 covariate=%3
 file print /LRT/ %%
 file delete %%
 set ple %p silent
;;;;
macro doanalysis
 set loc dose qua
 dose=x
 if (%2 == "%3/%3") then dose=0
 if (%2 == "%3/%4") then dose=1
 if (%2 == "%4/%4") then dose=2
 set loc inter qua
 inter=x
 inter = dose*%5
 reg %1 = dose %5 inter
 reg %1 = dose %5
 lrt
;;;;
```

```
interaction trait rs11204005 japan
Interaction analysis trait=trait marker=rs11204005 covariate=japan
LRTS        0.6601    1  0.4165
```

Sib-pair compared to PLINK: logistic regression

More subsetting examples

PLINK:

```
plink –file tophit –filter-males –logistic
plink –file tophit –filter-females –logistic
plink –file tophit –logistic –sex –interaction
```

Sib-pair:

```
read bin gwas3.bin
select male
ass trait snp
unselect
select female
ass trait snp
unselect
set loc fem aff
fem=female
interaction trait rs11204005 fem
```

Sib-pair compared to PLINK: filtering on interSNP LD

Subsetting to obtain SNPs in LD, and plotting clusters

PLINK: (1 minute 17 seconds)

```
plink –bfile wgas3 –indep-pairwise 50 10 0.2 –out prune1
```

Sib-pair (3 minutes for one chromosome)

```
read bin wgas3.bin
drop where r2 0.2
```

Sib-pair compared to PLINK

Individual MDS plot

PLINK

```
plink –bfile wgas3 –extract prune1.prune.in –genome –out ibs1
plink –bfile wgas3 –read-genome ibs1.genome –cluster –ppc 1e-3 –cc –mds-plot
2 –out strat1
```

Sib-pair compared to PLINK: merging in additional data

Merge in additional “fine mapping” genotyping

PLINK

```
plink –bfile wgas3 –snp rs11204005 –window 100 –merge extra.ped extra.map \
      –make-bed –out followup
plink –bfile followup –mh –within pop.cov –out followup-cmh
```

Sib-pair

```
read bin wgas3.bin
which rs11204005
keep trait 108321 -- 108421
include extras.in
update extras.dat
set loc japan aff
japan=n
sel ped J*
japan=y
unselect
# reorder data by map position
order trait japan $mm
pack
ass trait cov japan
write bin followup.bin compress
```

Sib-pair v. PLINK and haplo.stats: Haplotype association

SNP haplotype association analysis

PLINK (0.1 seconds)

```
plink –bfile followup2 –chap –hap-snps rs2460915-rs2460338
–each-versus-others
```

```
HAPLO      FREQ     OR(A)    SPEC(A)    OR(N)
  -------    ------   -------  ---------  -------
    TGTAG    0.1712   (-ref-)  1.247e-07  (-ref-)
    AGTAG    0.0111 1.117e-09    0.07573     |
    TATAG   0.06471     32.39   0.004304     |
    AGGAG    0.1441     2.201   0.004375     |
    AAGAG   0.04592     21.65    0.04734     |
    AAGGG   0.03485     823.6    0.00586     |
    AGGAC   0.02286     1.726     0.2204     |
    AGGGC   0.04373    0.7162    0.01523     |
    AAGGC    0.4382     23.52   1.83e-07     |
  -------    ------   -------    ------- --------

Model comparison test statistics:

                           Alternate       Null
                  -2LL :       63.33      122.8

  Likelihood ratio test: chi-square = 59.5
                         df = 8
                         p = 5.835e-10

[ haplo.glm() LRTS=60.24186 ]
```

Sib-pair (85 seconds)

```
dis rs2460915 rs7835221 rs2460911 rs11204005 rs2460338 trait
```

Poor old Sib-pair! In its defence, I should add it fits a “full” model testing for extragametic association, as well as the association model testing for differences in haplotype frequency. The former model does not scale up well.

```
Haplotype   n           y           [haplo.group() results]
---------   ---------   ---------
A A G A C   0.0000      0.0112      [ -       0.01133 ]
A G G A C   0.0366      0.0109      [ 0.01322 0.06798 ]
A G T A C   0.0122      0.0000      [ 0.01220 -       ]
A A G G C   0.2537      0.5993      [ 0.25375 0.59611 ]
A G G G C   0.0754      0.0140      [ 0.07537 0.01389 ]
T G G G C   0.0123      0.0000      [ 0.00566 0.01234 ]
A A G A G   0.0132      0.0672      [ 0.01322 0.06798 ]
T A G A G   0.0000      0.0104      [ -       0.01053 ]
A G G A G   0.2307      0.0773      [ 0.23069 0.07817 ]
T A T A G   0.0135      0.1126      [ 0.01352 0.11388 ]
A G T A G   0.0245      0.0000      [ 0.02454 -       ]
T G T A G   0.3278      0.0332      [ 0.32779 0.03359 ]
A A G G G   0.0000      0.0638      [ -       0.06346 ]

trait           41          48

            Number of loci =         5
 No. genotyped individuals =        89
 No. obs. unique genotypes =        30
  Stratified LD Chi-square =        44.30 (df= 470, P=1.0000)
    Association Chi-square =        76.92 (df=   7, P=0.0000)

NOTE:  Degrees of freedom calculation for association test assumes only
          8 haplotypes to be present in the population.

[ the LRTS based on haplo.em() is 76.66 ]
```

Sib-pair v. PLINK: Conclusions

- Similar range of simpler association analysis
- Sib-pair is still slower, but not annoyingly so (some exceptions)
- Sib-pair offers greater interactivity/programmability

Sib-pair: Things to do

- Multithreading via OMP: probably 3-4 fold speed improvement
- GLMM VC linkage analysis