A brief Introduction to Genetic Epidemiology using Stata

Neil Shephard

n.shephard@sheffield.ac.uk

Institute for Cancer Reasearch

University of Sheffield

Outline

- Brief Overview of Genetics

Outline

- Brief Overview of Genetics
- Data Formatting Issues

Outline

- Brief Overview of Genetics
- Data Formatting Issues
- Common Tests

Outline

- Brief Overview of Genetics
- Data Formatting Issues
- Common Tests
- User-written Commands

Outline

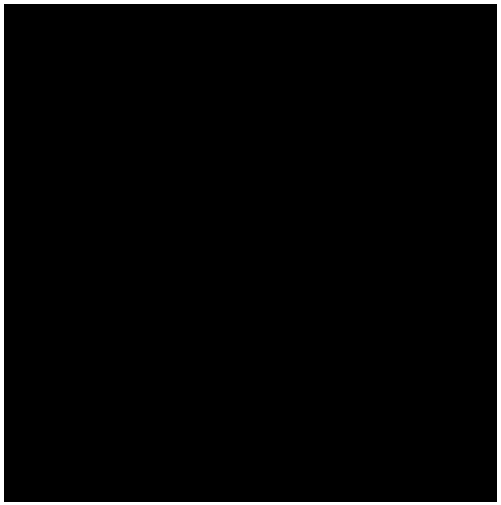
- Brief Overview of Genetics
- Data Formatting Issues
- Common Tests
- User-written Commands

What is Genetics?

- *Heritability* and *Variation*

A Brief History

- 1866 - Gregor Mendel founder of genetics [a]
- 1944 - DNA shown to be genetic material [b]
- 1953 - Watson and Crick publish structure of DNA [c]



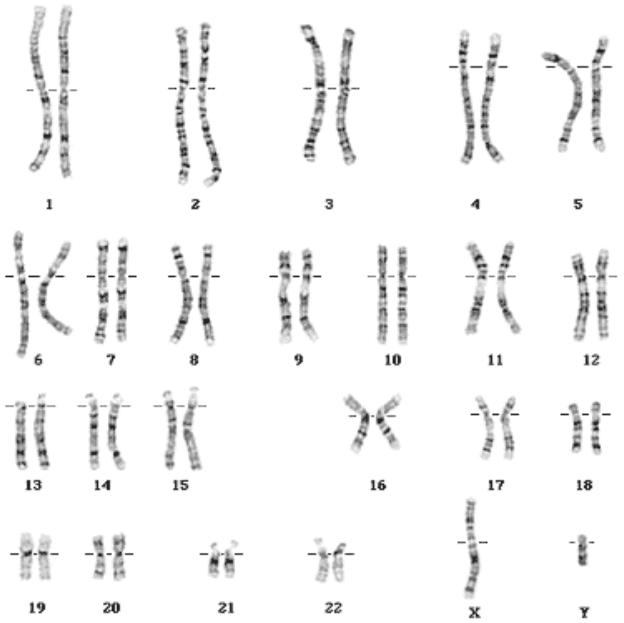
a Mendel (1866) *Verhandlungen des naturforschenden Vereines* 4:3-47

b Avery, MacLeod, McCarty (1944) *J Exp Med* 79: 137158

c Watson, Crick (1953) *Nature* 171:737-738

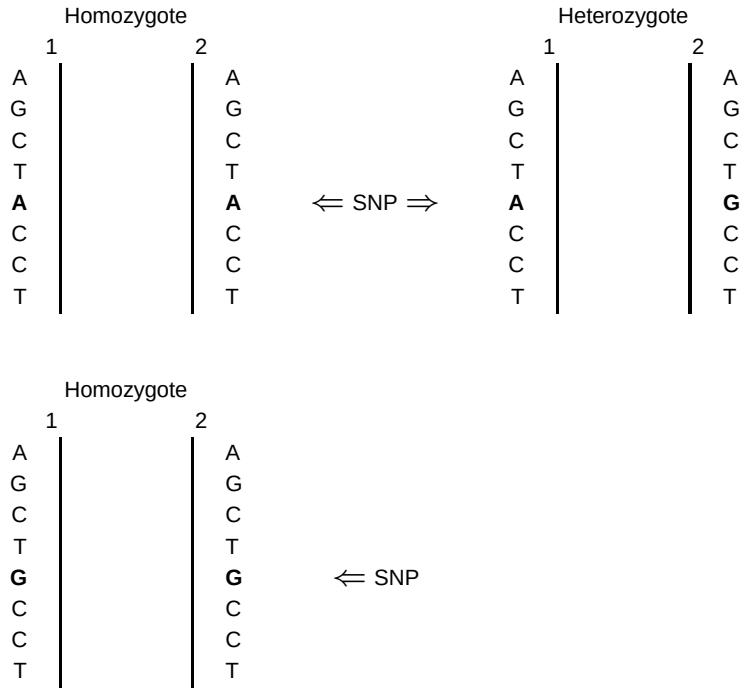
DNA

*What is Genetics? (*The Human Genome*)*



- 23 Chromosomes
- 3 billion nucleotides
- 20-25000 genes
- Humans are diploid

Genetic Variation



- Basic level of genetic variation is Single Nucelotide Polymorphism (SNP)
- Bi-alelic markers common throughout the genome (5.5 million validated SNPs)
- Cheap and easy to genotype ($\sim \$0.10$ cents per SNP)

Genetic Epidemiology

- Does genetic variation affect disease status?

Genetic Epidemiology

- Does genetic variation affect disease status?
- *Monogenic* : one gene e.g. Cystic Fibrosis, Huntingdons, Sickle Cell Anemia

Genetic Epidemiology

- Does genetic variation affect disease status?
- *Monogenic* : one gene e.g. Cystic Fibrosis, Huntingdons, Sickle Cell Anemia
- *Complex* : multiple genes e.g. Type II Diabetes, Autoimmune Diseases, Cancer, Heart Disease

Genetic Epidemiology

- Does genetic variation affect disease status?
- *Monogenic* : one gene e.g. Cystic Fibrosis, Huntingdons, Sickle Cell Anemia
- *Complex* : multiple genes e.g. Type II Diabetes, Autoimmune Diseases, Cancer, Heart Disease
- Environment can greatly influcence both

Genetic Epidemiology

- Does genetic variation affect disease status?
- *Monogenic* : one gene e.g. Cystic Fibrosis, Huntingdons, Sickle Cell Anemia
- *Complex* : multiple genes e.g. Type II Diabetes, Autoimmune Diseases, Cancer, Heart Disease
- Environment can greatly influcence both
- Family based studies (*monogenic*)

Genetic Epidemiology

- Does genetic variation affect disease status?
- *Monogenic* : one gene e.g. Cystic Fibrosis, Huntingdons, Sickle Cell Anemia
- *Complex* : multiple genes e.g. Type II Diabetes, Autoimmune Diseases, Cancer, Heart Disease
- Environment can greatly influcence both
- Family based studies (*monogenic*)
- **Population based studies (*complex*)**

Data Structure

Long format

ID	locus	_1	_2
ABC001	snp1	A	A
ABC001	snp2	G	T
ABC001	snp3	T	T
ABC001	snp4	C	C
ABC002	snp1	A	A
ABC002	snp2	G	T
ABC002	snp3	T	T
ABC002	snp4	C	C
ABC003	snp1	A	A
ABC003	snp2	G	T
ABC003	snp3	T	T
ABC003	snp4	C	C
.	.	.	.

Wide format

ID	snp1_1	snp1_2	snp2_1	snp2_2	snp3_1	snp3_2	snp4_1	snp4_2	...
ABC001	A	A	G	T	T	T	C	C	...
ABC002	A	T	G	G	T	T	G	G	...
ABC003	A	A	G	T	C	T	C	C	...
ABC004	A	A	T	T	C			C	...
ABC005	A	A	G	T	T	T	C	C	...
ABC006	T	T	G			C	C	G	...
ABC007			G	T	C	T	C	C	...
ABC008	A	T	T	T	T	T	G	G	...
.	.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	.	...

Data Management

- `odbc` connectivity makes extracting data straight-forward
- `reshape` the data from `long` to `wide`
- `encode` genotype data. Common allele 1; Rare allele 2
- Encode genotypes as dummy variables

Genotype	A A	A G	G G
Encoded	1 1	1 2	2 2
Dummy	0	1	2

Hardy-Weinberg equilibrium

- Proposed simultaneously by Hardy [a] and Weinberg [b]
- Prediction of genotype frequencies based on allele frequencies
- Various assumptions, but robust to deviations
- Useful in detecting genotyping errors

[a] Hardy (1908) *Science* 28:49-50

[b] Weinberg (1908) *Jahreshefte Verein f. vaterl. Naturk* 64:368-82

H-W eqm (cont.)

- Bi-allelic locus (e.g. SNP)
- Allele A with frequency p
- Allele G with frequency $1 - p$
- Expected Genotype frequencies follow $Binom(2, p)$

Genotype	AA	AG	GG
Expected	p^2	$2p(1-p)$	$(1-p)^2$

Calculating H-W equilibrium : genhw

- Use genhw written by Mario Cleves to test H-W equilibrium [a]

```
. genhw snp_1 snp_2 if(status == 0)
       Genotype |     Observed        Expected
    ------------+-----------------------------
             11 |          132          129.94
             12 |          206          210.12
             22 |           87           84.94
    ------------+-----------------------------
          total |          425          425.00

         Allele | Observed     Frequency      Std. Err.
    ------------+---------------------------------------
              1 |      470        0.5529         0.0172
              2 |      380        0.4471         0.0172
    ------------+---------------------------------------
          total |      850        1.0000

    Estimated disequilibrium coefficient (D) =   0.0048

    Hardy-Weinberg Equilibrium Test:
             Pearson chi2 (1) =    0.163  Pr= 0.6862
    likelihood-ratio chi2 (1) =    0.163  Pr= 0.6862
     Exact significance prob  =               0.6951
```

[a] Alternative command hwsnp by Mario Cleves

Trend Test for Association

- Trend Test for association [a]
- Robust to deviations from H-W eqm
- Use `nptrend` to perform test
- Use genotypes encoded as `0`, `1`, `2`

```
. nptrend snp1, by(status)

casestatus     score       obs      sum of ranks
         0         0       425      177115.5
         1         1       449      205259.5

          z  =  2.57
  Prob > |z| = 0.010
```

[a] Sasieni (1997) *Biometrics* 53:1253-1261

Logistic Regression

- Trend test demonstrate 'association'.
- Logistic regression used to estimate effect size and determine primary effects [a]
- Estimate Genotype Relative Risk (GRR)

Genotype	AA	AG	GG
Dummy	0	1	2
Risk	$-$	OR_1	OR_2

[a] Cordell & Clayton (2002) *Am J Hum Gen* 70:124-141

Logistic Regression (cont)

```
. xi: logistic casestatus i.snp1 i.snp2 i.snp3
i.snp1            _Isnp1_0-2          (naturally coded; _Isnp1_0 omitted)
i.snp2            _Isnp2_0-2          (naturally coded; _Isnp2_0 omitted)
i.snp3            _Isnp3_0-2          (naturally coded; _Isnp3_0 omitted)

note: _Isnp3_2 != 0 predicts success perfectly
      _Isnp3_2 dropped and 1 obs not used


Logistic regression                               Number of obs   =        865
                                                  LR chi2(5)      =      11.33
                                                  Prob > chi2     =     0.0452
Log likelihood = -593.54416                       Pseudo R2       =     0.0095

------------------------------------------------------------------------------
  casestatus | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Isnp1_1 |   1.255109   .2132321     1.34   0.181     .8996417    1.751028
    _Isnp1_2 |   1.521735   .3274461     1.95   0.051     .9981089    2.320065
    _Isnp2_1 |   .9863323   .1745972    -0.08   0.938     .6971824    1.395404
    _Isnp2_2 |   .9826968   .5031001    -0.03   0.973     .3602795    2.680399
    _Isnp3_1 |   .6158163   .1506146    -1.98   0.047     .3812999    .9945706
------------------------------------------------------------------------------
. swaic, model
Stepwise Model Selection by AIC
logistic regression.
number of obs = 865
------------------------------------------------------------------------------
casestatus          |  Df     Chi2     P>Chi2  -2*ll    Df Res.  AIC
--------------------+---------------------------------------------------------
Null Model          |                          1198.4   864      1200.4
Step 1:_Isnp3*      |  1      6.5723   .0104   1191.8   863      1195.8
Step 2:_Isnp1*      |  2      4.7548   .0928   1187.1   861      1195.1
Step 3:_Isnp2*      |  2      .00657   .9967   1187.1   859      1199.1
------------------------------------------------------------------------------
minimun AIC = 1195.095;  model: _Isnp3* _Isnp1*
```

Linkage Disequilibrium

- SNPs are not indepdent
- Non-random association between loci is **Linkage Disequilibrium**
- Number of different measures of LD [a] e.g. D', Δ and R^2
- David Clayton's `pwld` command can calculate a range of LD measures

[a] Devlin & Risch (1995) *Genomics* 29:311-322

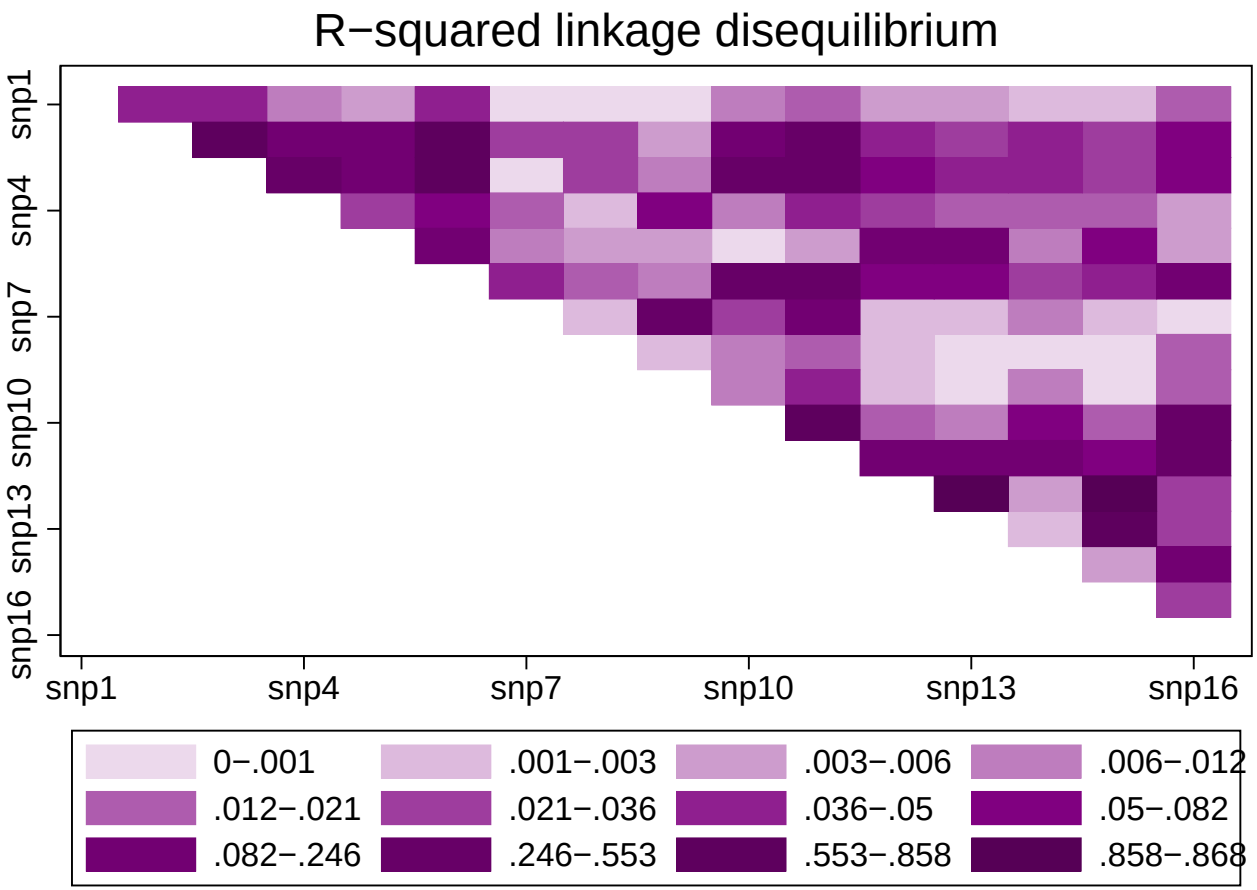
Linkage Disequilibrium (cont.)

```
. pwld snp*_* if(status == 0), me(R2) matrix(pwld_r2) replace

Off-diagonal elements are estimates of R-squared (assuming H-W equilibrium)
Diagonal elements are relative frequencies of allele 2

       snp1   snp2   snp3   snp4   snp5   snp6   snp7   snp8   snp9  snp10  snp11  snp12  snp13  snp14  snp15
snp1   0.06
snp2   0.05   0.47
snp3   0.04   0.73   0.45
snp4   0.01   0.17   0.25   0.21
snp5   0.00   0.11   0.12   0.02   0.08
snp6   0.04   0.55   0.56   0.08   0.13   0.42
snp7   0.00   0.03   0.00   0.02   0.01   0.05   0.06
.      .      .      .      .      .      .      .     .
```

- Results can be stored in a matrix for subsequent plotting
- Use Adrian Manders `plotmatrix` to generate “heatmap” of LD

```
. plotmatrix, mat(pwld) color(purple) upper nodiag title("R-squared Linkage Disequilibrium")
Percentiles are used to create legend
    purple*0.15  purple*0.88
```

Linkage Disequilibrium (cont)



Haplotype Estimation

- A haplotype is a combination of alleles at multiple linked loci that are transmitted together

		SNP 1		
		AA	AT	TT
	GG	AG AG	AG TG	GT GT
SNP 2	GC	AG AC	AG TC or AC TG	TG TC
	CC	AC AC	AC TC	TC TC

Haplotype Estimation (cont.)

- Association of haplotypes can be tested using Adrian Manders `hapipf` [a]

```
. hapipf snp1_* snp2_* snp3_*, ipf(l1*l2*l3*caco) mv nolog \\
    model(0)

Marker information
------------------
Alleles for l1 are (snp1_1 , snp1_2)
Alleles for l2 are (snp2_1 , snp2_2)
Alleles for l3 are (snp3_1 , snp3_2)

Haplotype Frequency Estimation by EM algorithm
----------------------------------------------
 Model             = l1*l2*l3*caco
 No. loci          = 3
 Log-Likelihood    = -2878.036717229983
 Df                = 0
 No. parameters    = 16
 No. cells         = 16
```

```
. hapipf snp1_* snp2_* snp3_*, ipf(l1*l2*l3+caco) mv nolog \\
    model(1) lrtest(0, 1)

Marker information
------------------
Alleles for l1 are (snp1_1 , snp1_2)
Alleles for l2 are (snp2_1 , snp2_2)
Alleles for l3 are (snp3_1 , snp3_2)

Haplotype Frequency Estimation by EM algorithm
----------------------------------------------
 Model             = l1*l2*l3+caco
 No. loci          = 3
 Log-Likelihood    = -2883.266498455095
 Df                = 7
 No. parameters    = 9
 No. cells         = 16

Likelihood Ratio Test Comparing Model l1*l2*l3+caco to l1*l2*l3*caco
--------------------------------------------------------------------
llhd2 (df2)           = -2883.2665 7
llhd1 (df1)           = -2878.0367 0
-2*(llhd2-llhd1)      = 10.459562
Change in df          =  7
p-value               =  .16399138
```

[a] Quantitative trait associations can be tested using `qhapipf`

Putting it all together

- Often have lots of loci genotyped (upto $500,000$)

Putting it all together

- Often have lots of loci genotyped (upto $500,000$)
- Efficent method of analysing and reporting results

Putting it all together

- Often have lots of loci genotyped (upto $500,000$)
- Efficent method of analysing and reporting results
- Use `qui foreach` loops to pass over all loci

Putting it all together

- Often have lots of loci genotyped (upto $500,000$)
- Efficent method of analysing and reporting results
- Use `qui foreach` loops to pass over all loci
- Write scalars to text-files using `file write`

Putting it all together

- Often have lots of loci genotyped (upto $500,000$)
- Efficent method of analysing and reporting results
- Use `qui foreach` loops to pass over all loci
- Write scalars to text-files using `file write`
- Use `parmest` or `estout` for saving and compiling regression results

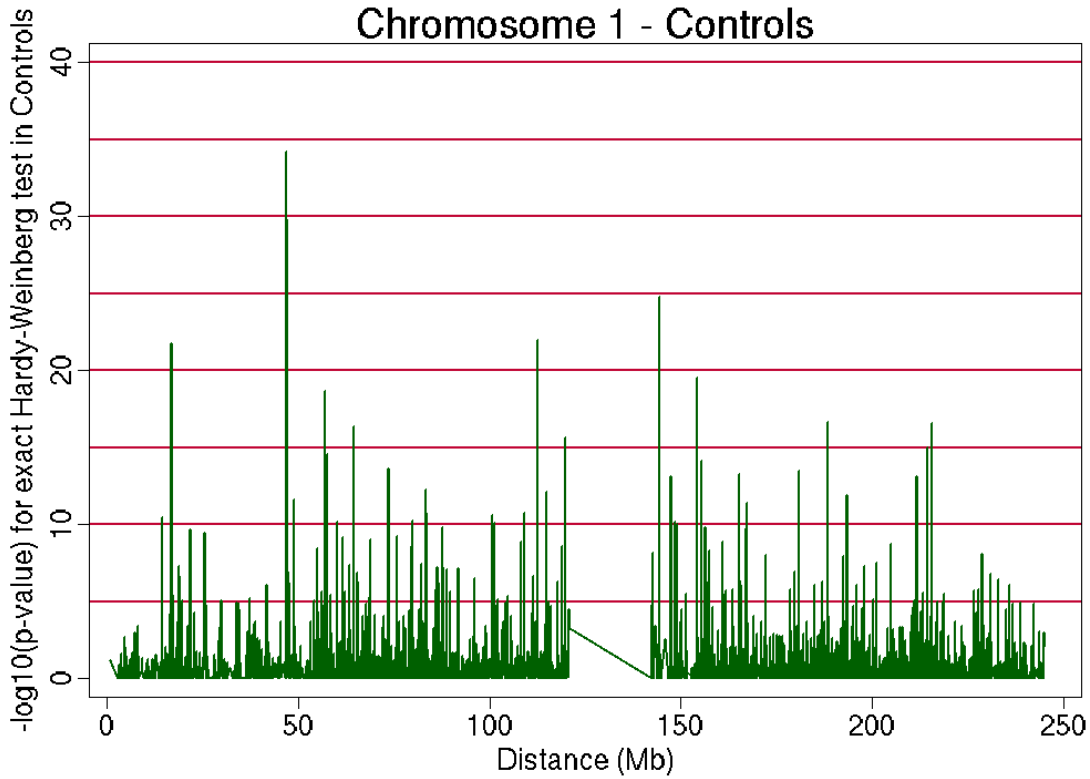
Putting it all together

- Often have lots of loci genotyped (upto $500,000$)
- Efficent method of analysing and reporting results
- Use `qui foreach` loops to pass over all loci
- Write scalars to text-files using `file write`
- Use `parmest` or `estout` for saving and compiling regression results
- Use `listtex` or `tabout` for generating tables

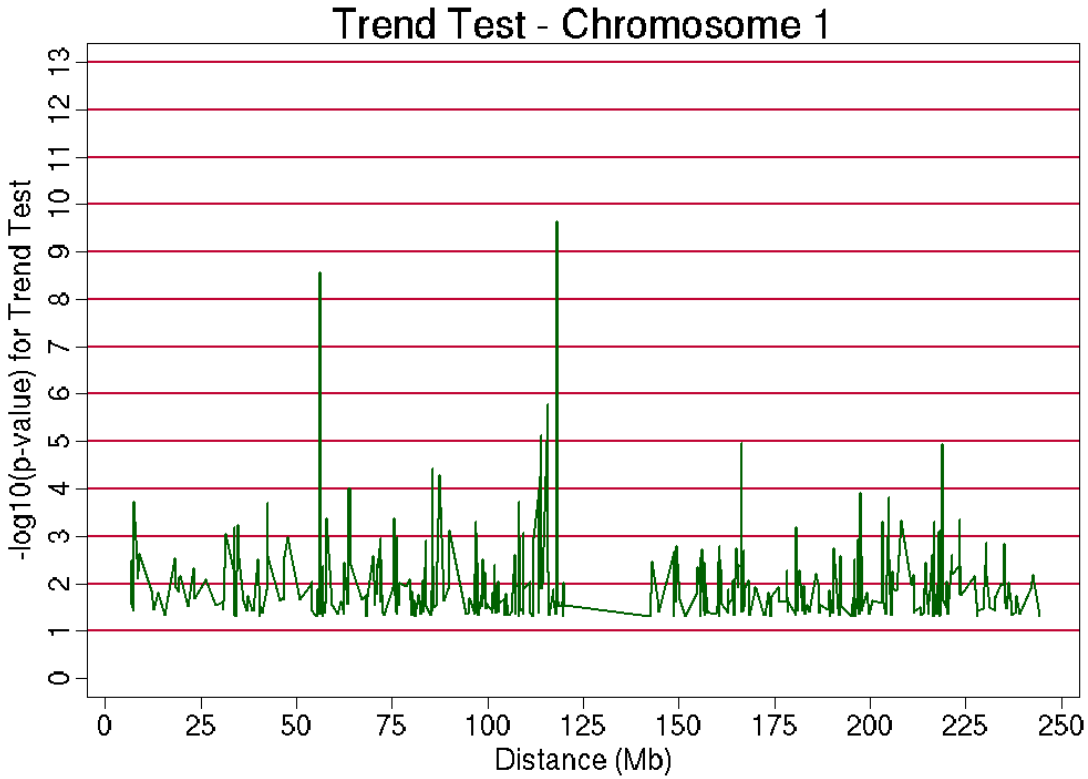
Putting it all together

- Often have lots of loci genotyped (upto $500,000$)
- Efficent method of analysing and reporting results
- Use `qui foreach` loops to pass over all loci
- Write scalars to text-files using `file write`
- Use `parmest` or `estout` for saving and compiling regression results
- Use `listtex` or `tabout` for generating tables
- Stata's excellent graph functions for plotting results

Whole Genome Association Study



Whole Genome Association Study



Summary

- Stata provides a number of general commands for analysis of genetic data

Summary

- Stata provides a number of general commands for analysis of genetic data
- A growing number of user written commands for specific genetic analysis

Summary

- Stata provides a number of general commands for analysis of genetic data
- A growing number of user written commands for specific genetic analysis
- Analysis of large number of loci facilitated by judicious programming

Summary

- Stata provides a number of general commands for analysis of genetic data
- A growing number of user written commands for specific genetic analysis
- Analysis of large number of loci facilitated by judicious programming
- Many useful commands for summarising and reporting

Summary

- Stata provides a number of general commands for analysis of genetic data
- A growing number of user written commands for specific genetic analysis
- Analysis of large number of loci facilitated by judicious programming
- Many useful commands for summarising and reporting