

QTL analysis

Martin Farrall

Dept. Cardiovascular Medicine
University of Oxford
Wellcome Trust Centre for Human Genetics



- linear statistical models (ANOVA)
 - application to clustered data (sibships/twins)
 - SAS (www.sas.com)

- family-based models of linkage and association
 - QTDT (www.sph.umich.edu/csg/abecasis/QTDT/)

- multilocus models of linkage/association
 - measured genotype and haplotype analysis
 - PAP - Pedigree Analysis Package

 - (hasstedt.genetics.utah.edu/pap5/)

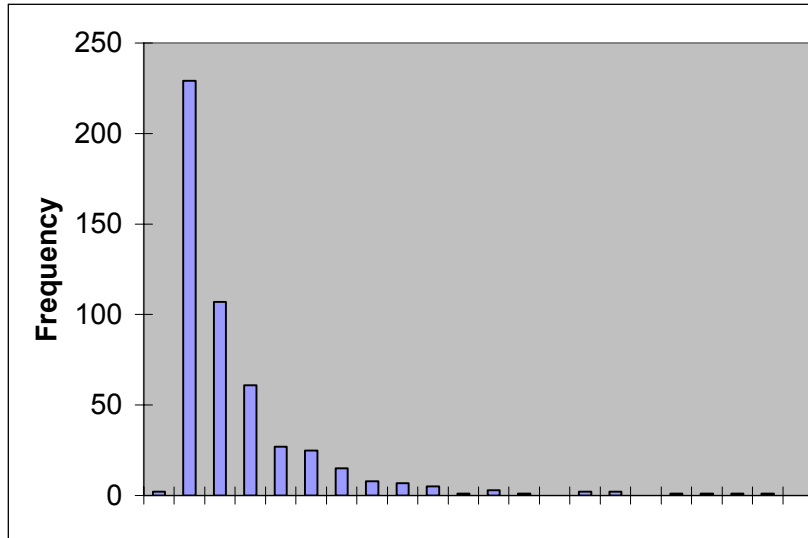
First things first, don't forget basic statistics!

check distribution of trait for approximate normality
and outliers

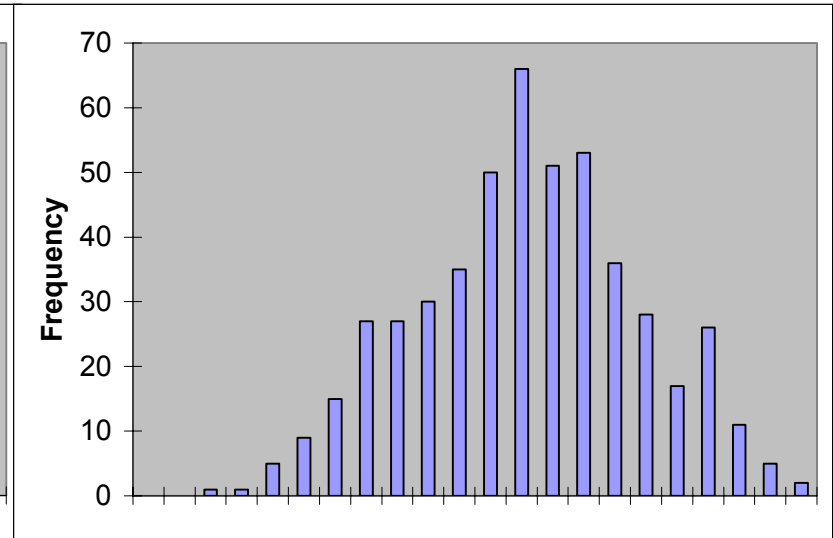
consider transformation, trimming, winsorization as
necessary

think about sources of covariation (sex, age etc.)
ensure model makes epidemiological sense

many biological traits show positive skewness

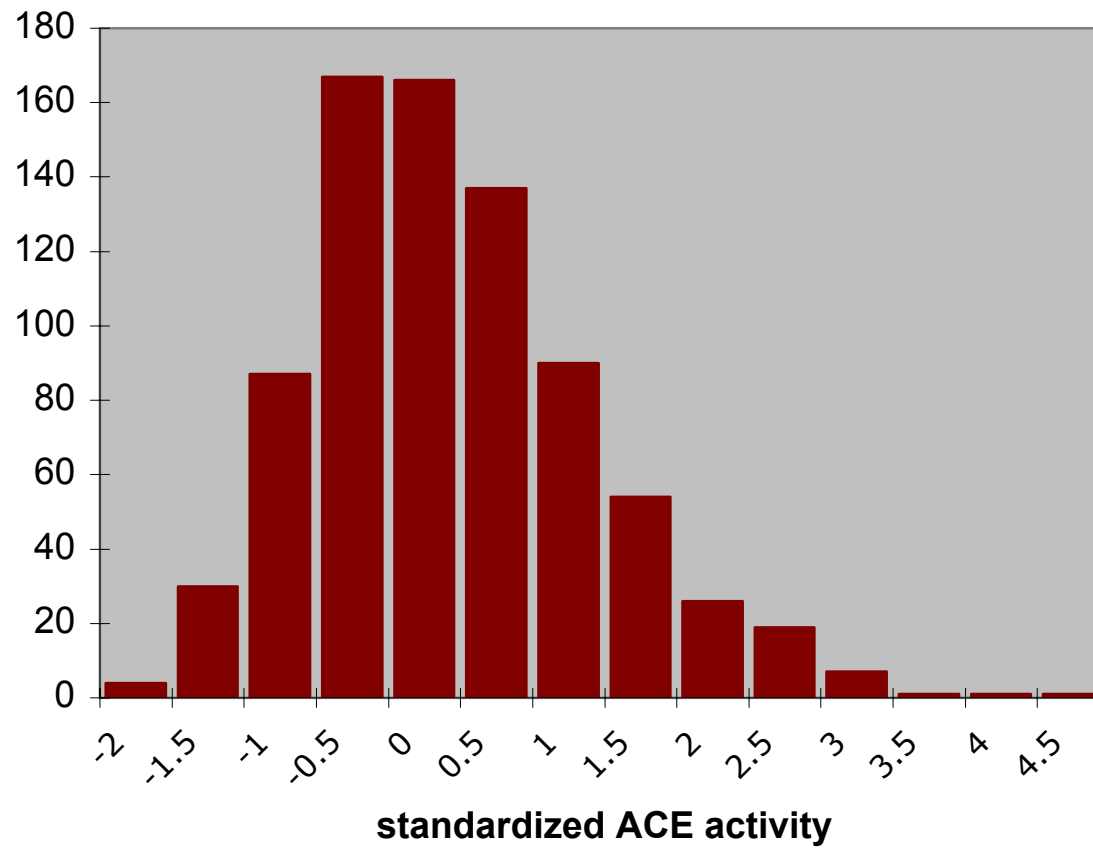


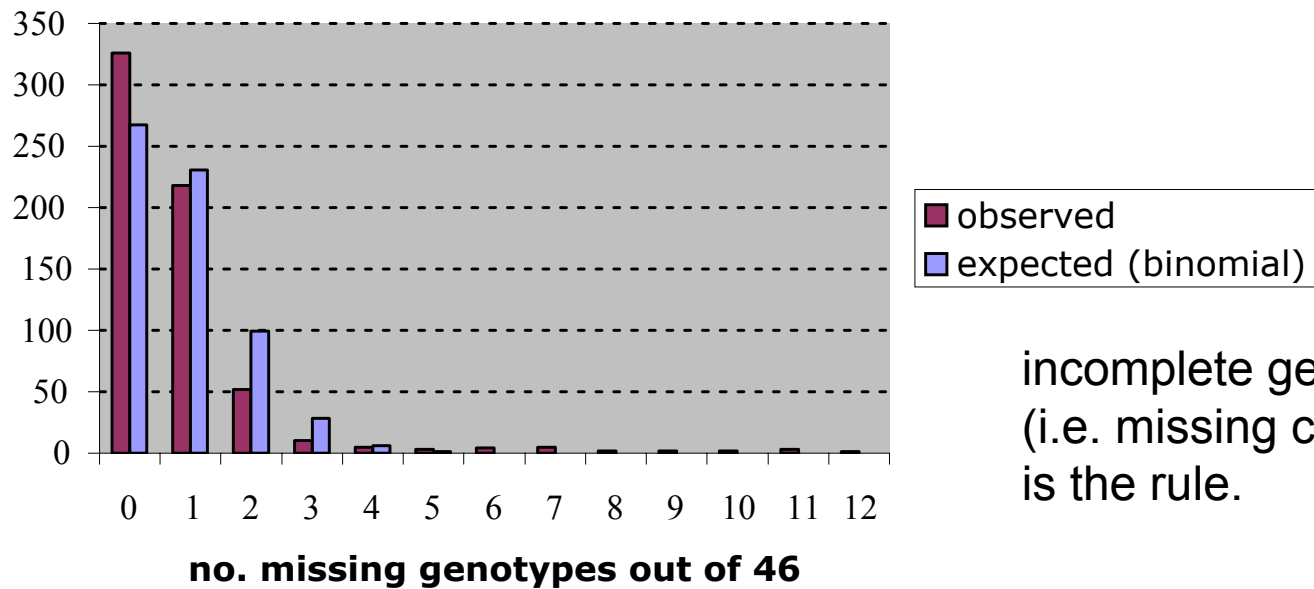
untransformed trait



log(trait)

Distribution of ACE activity in Nigerian families





incomplete genotyping
(i.e. missing covariates)
is the rule.

patterns of missingness
may not be random!

N	A6138C	A23495G	31839insC	A31958G
594	typed	typed	typed	typed
79	typed	typed	missing	missing
43	missing	missing	typed	typed
32	missing	typed	typed	typed
8	typed	typed	missing	typed
4	missing	typed	missing	missing
4	typed	typed	typed	missing
1	missing	missing	missing	typed
1	missing	missing	typed	missing
1	typed	missing	missing	missing
1	typed	missing	typed	typed

linear modeling of QTL

$$y = X\beta + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$,

X is a suitable vector of indicator variables and
 β is a vector of coefficients for genetic effects

modeling additive and dominance effects

	low/low	low/high	high/high
additive	0	1	2
dominance	0	1	0

linear QTL models can be devised to deal with clustered (familial) data (mixed linear model)

$$y = X\beta + Z\mu + \varepsilon$$

□ fixed and random effects

where $\varepsilon \sim N(0, \sigma^2 I)$ and $\mu \sim N(0, D)$ where D models the covariation (e.g. between sibs)

similar approach for variance-components linkage analysis

SAS mixed model analysis of single SNPs

```
proc mixed order=data METHOD=REML CL ALPHA=.05 COVTEST;
class familial_clustering;
model ace = a6138 / DDFM=SATTERTH SOLUTION;
random familial_clustering /;
run;
quit;
```

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr > Z	Alpha	Lower	Upper
familial_clustering	0.2539	0.04616	5.50	<.0001	0.05	0.1832	0.3757
Residual	0.6208	0.04128	15.04	<.0001	0.05	0.5472	0.7104

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
a6138	-0.4122	0.05389	656	-7.65	<.0001
a23495	0.7993	0.05870	691	13.62	<.0001
a31839	-0.1400	0.09012	654	-1.55	0.1208
a31958	0.6710	0.04974	641	13.49	<.0001

SAS mixed model analysis of multiple SNPs

```
proc mixed order=data METHOD=REML CL ALPHA=.05 COVTEST;
class familial_clustering;
model ace = a6138 a23495 a31839 a31958 / DDFM=SATTERTH SOLUTION;
random familial_clustering /;
run;
quit;
```

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
familial_clustering	0.2075	0.03919	5.29	<.0001	0.05	0.1479	0.3122
Residual	0.4566	0.03394	13.45	<.0001	0.05	0.3968	0.5312

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	-0.3996	0.07679	486	-5.20	<.0001
a6138	-0.2128	0.05258	571	-4.05	<.0001
a23495	0.2345	0.09054	572	2.59	0.0098
a31839	-0.5042	0.1049	573	-4.81	<.0001
a31958	0.6295	0.08400	573	7.49	<.0001

SAS mixed model analysis of multiple SNPs - twin data and dominance

```
proc mixed order=data METHOD=REML CL ALPHA=.05 COVTEST;
class family twin_type sex;
model log_fc_level=age sex xa xd m4a / DDFM=SATTERTH SOLUTION;
repeated /type=ar(1) subject=family group=twin_type;
run;
```

Covariance Parameter Estimates

Cov Parm	Subject	Group	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
Variance	family	DZ	0.3112	0.01905	16.34	<.0001	0.05	0.2770	0.3522
AR(1)	family	DZ	0.4129	0.04701	8.78	<.0001	0.05	0.3208	0.5050
Variance	family	MZ	0.3498	0.04600	7.60	<.0001	0.05	0.2746	0.4608
AR(1)	family	MZ	0.8188	0.03407	24.03	<.0001	0.05	0.7520	0.8856

Solution for Fixed Effects

Effect	sex	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		1.5532	0.1361	487	11.42	<.0001
age		-0.00607	0.001859	442	-3.27	0.0012
sex	female	0.3687	0.08671	358	4.25	<.0001
sex	male	0
xa		-0.2353	0.03496	720	-6.73	<.0001
xd		0.1156	0.04484	698	2.58	0.0101
m4a		-0.2503	0.03434	717	-7.29	<.0001

```

zeon<125>qtdt -d nigeria_4.dat -p nigeria_4.pre -i nigeria_4.ibd --snp -x-100.000 -a- -weg -vega
QTDT - Quantitative TDT 2.4.3
(c) 1998-2002 Goncalo Abecasis (goncalo@umich.edu)

```

The following parameters are in effect:

```

Association Model :          NONE (-a[a|d|f|m|o|p|r|t|w|-])
Full Model Variances :      NON SHARED (-v{e|c|g|n|t|a|d|-})
&                            POLYGENIC
&                            ADDITIVE
Null Model Variances :      NON SHARED (-w{e|c|g|n|t|a|d|-})
&                            POLYGENIC
Transmission Scoring :      FULL PEDIGREE (-t[n|p])

```

The following models will be evaluated...

```

NULL MODEL
Means = Mu
Variances = Ve + Vg

FULL MODEL
Means = Mu
Variances = Ve + Vg + Va

```

Testing trait: ace

Testing marker: A6138C

Allele	df(0)	-LnLk(0)	df(V)	-LnLk(V)	ChiSq	p
All	787	340.65	786	332.77	15.74	7e-05 (790 probands)

Testing marker: A23495G

Allele	df(0)	-LnLk(0)	df(V)	-LnLk(V)	ChiSq	p
All	787	340.65	786	330.84	19.60	1e-05 (790 probands)

Testing marker: 31839insC

Allele	df(0)	-LnLk(0)	df(V)	-LnLk(V)	ChiSq	p
All	787	340.65	786	329.40	22.49	2e-06 (790 probands)

Testing marker: A31958G

Allele	df(0)	-LnLk(0)	df(V)	-LnLk(V)	ChiSq	p
All	787	340.65	786	329.38	22.53	2e-06 (790 probands)

QTDT variance-components linkage analysis

$$\text{ChiSq}/2\ln(10) = \text{LOD}$$

$$\text{i.e. } 22.53/4.605 = 4.89$$

```
zeon<128>qtdt -d nigeria_4.dat -p nigeria_4.pre -i nigeria_4.ibd --snp -x-100.000
QTDT - Quantitative TDT 2.4.3
(c) 1998-2002 Goncalo Abecasis (goncalo@umich.edu)
```

The following parameters are in effect:

```
Association Model :      ORTHOGONAL (-a[a|d|f|m|o|p|r|t|w|-])
Full Model Variances :  NOT MODELLED (-v{e|c|g|n|t|a|d|-})
Null Model Variances :  NOT MODELLED (-w{e|c|g|n|t|a|d|-})
Transmission Scoring :  FULL PEDIGREE (-t[n|p])
```

The following models will be evaluated...

```
NULL MODEL
Means = Mu + B
```

```
FULL MODEL
Means = Mu + B + W
```

```
Testing trait:          ace
=====
```

```
Testing marker:        A6138C
-----
```

Allele	df(0)	Rsq(0)	df(T)	Rsq(T)	F	p	
1	662	0.05	661	0.07	15.01	0.0001	(186/664 probands)

```
Testing marker:        A23495G
-----
```

Allele	df(0)	Rsq(0)	df(T)	Rsq(T)	F	p	
1	694	0.13	693	0.20	61.00	2e-14	(168/696 probands)

```
Testing marker:        31839insC
-----
```

Allele	df(0)	Rsq(0)	df(T)	Rsq(T)	F	p	
1	654	0.00	653	0.00	1.16		(69/656 probands)

```
Testing marker:        A31958G
-----
```

Allele	df(0)	Rsq(0)	df(T)	Rsq(T)	F	p	
1	656	0.18	655	0.22	37.23	2e-09	(169/658 probands)

QTDT gene-association analysis (orthogonal model)

zeon<131>qtdt -d nigeria_4.dat -p nigeria_4.pre -i nigeria_4.ibd --snp -x-100.000 -ap -wega
 QTDT - Quantitative TDT 2.4.3
 (c) 1998-2002 Goncalo Abecasis (goncalo@umich.edu)

The following parameters are in effect:

Association Model : STRATIFICATION (-a[a|d|f|m|o|p|r|t|w|-])
 Full Model Variances : NOT MODELLED (-v{e|c|g|n|t|a|d|-})
 Null Model Variances : NON SHARED (-w{e|c|g|n|t|a|d|-})
 & POLYGENIC
 & ADDITIVE
 Transmission Scoring : FULL PEDIGREE (-t[n|p])

The following models will be evaluated...

NULL MODEL

Means = $\mu + X$

Variances = $V_e + V_g + V_a$

FULL MODEL

Means = $\mu + X + W$

Variances = $V_e + V_g + V_a$

QTDT testing for stratification

Testing trait: ace
 =====

Testing marker: A6138C

Allele	df(0)	-LnLk(0)	df(S)	-LnLk(S)	ChiSq	p
1	659	259.92	658	259.47	0.91	(186/664 probands)

Testing marker: A23495G

Allele	df(0)	-LnLk(0)	df(S)	-LnLk(S)	ChiSq	p
1	691	216.88	690	215.19	3.39	0.0656 (168/696 probands)

Testing marker: 31839insC

Allele	df(0)	-LnLk(0)	df(S)	-LnLk(S)	ChiSq	p
1	651	273.16	650	272.81	0.69	(69/656 probands)

Testing marker: A31958G

Allele	df(0)	-LnLk(0)	df(S)	-LnLk(S)	ChiSq	p
1	653	204.25	652	204.08	0.35	(169/658 probands)

```
zeon<132>qtdt -d nigeria_4.dat -p nigeria_4.pre -i nigeria_4.ibd --snp -x-100.000 -weg -vega
QTDT - Quantitative TDT 2.4.3
```

The following parameters are in effect:

```
Association Model :      ORTHOGONAL (-a[a|d|f|m|o|p|r|t|w|-])
Full Model Variances :  NON SHARED (-v{e|c|g|n|t|a|d|-})
                        &
                        POLYGENIC
                        &
                        ADDITIVE
Null Model Variances :  NON SHARED (-w{e|c|g|n|t|a|d|-})
                        &
                        POLYGENIC
Transmission Scoring :  FULL PEDIGREE (-t[n|p])
```

The following models will be evaluated...

```
NULL MODEL
Means = Mu + B + W
Variances = Ve + Vg
```

```
FULL MODEL
Means = Mu + B + W
Variances = Ve + Vg + Va
```

Likelihood ratio statistic from Abecasis et al (AJHG, 2000)

```
Testing trait:          ace
=====
```

```
Testing marker:        A6138C
-----
```

Allele	df(0)	-LnLk(0)	df(V)	-LnLk(V)	ChiSq	p	
1	659	264.06	658	259.47	9.20	0.0024	(186/664 probands)

```
Testing marker:        A23495G
-----
```

Allele	df(0)	-LnLk(0)	df(V)	-LnLk(V)	ChiSq	p	
1	691	221.38	690	215.19	12.39	0.0004	(168/696 probands)

```
Testing marker:        31839insC
-----
```

Allele	df(0)	-LnLk(0)	df(V)	-LnLk(V)	ChiSq	p	
1	651	279.46	650	272.81	13.30	0.0003	(69/656 probands)

```
Testing marker:        A31958G
-----
```

Allele	df(0)	-LnLk(0)	df(V)	-LnLk(V)	ChiSq	p	
1	653	206.54	652	204.08	4.93	0.0264	(169/658 probands)

QTDT Fulker's candidate gene test

measured genotype analysis

exact likelihood solution

- PAP - Pedigree Analysis Package

- overall μ and σ

- dominance $(\mu_2 - \mu_1)/(\mu_3 - \mu_1)$

- displacement $(\mu_3 - \mu_1)/\sigma$

- class D regressive model (Bonney) to model residual intra-familial correlations (e.g. shared polygenes, environment)

MODEL: test
 FREQUENCY: papfqh
 TRANSMISSION: paptcal
 MAJOR LOCUS DISCRETE PARAMETERS: dmlpr0
 MAJOR LOCUS QUANTITATIVE PARAMETERS: qmlprdd
 WITHIN GENOTYPE PARAMETERS: papwgfc
 PENETRANCE SUBROUTINE: papenqa
 CORRELATION SUBROUTINE: papcrqa

THE LIKELIHOODS ARE NOT CORRECTED FOR ASCERTAINMENT

817 INDIVIDUALS MEASURED IN 234 PEDIGREES

THE GENETIC MODEL CONTAINS 136 AUTOSOMAL GENOTYPES
 CONCOMITANT 1: Sex

FREQUENCIES OF HAPLOTYPES

0.2797314	0.0656904	0.0000000	0.1037531
0.0070694	0.1653816	0.0000000	0.0000000
0.2908793	0.0293307	0.0000000	0.0170760
0.0000000	0.0401329	0.0009552	0.0000000

RECOMBINATION AT THE AUTOSOMAL LOCI:

FEMALE	0.0000000	0.0000000	0.0000000
MALE	0.0000000	0.0000000	0.0000000

TOTAL MEAN -0.0682325

TOTAL STANDARD DEV 0.9751950

DOMINANCE 0.5000000 0.5000000 0.5000000 0.5000000

DISPLACEMENT

LOCUS	1	2	3	4
	-0.4099250 /1/	0.4399300 /2/	-1.3590780 /3/	1.5985130 /4/
SPOUSE CORRELATION	0.3996215 /5/			
P-O (MD,MS,FD,FS)	0.2862925 /6/	0.2862925 /6/	0.2862925 /6/	0.2862925 /6/
SIB CORR (SS,SB,BB)	0.2123224 /7/	0.2123224 /7/	0.2123224 /7/	

PARAMETER ESTIMATED VALUE STANDARD ERROR

1	-0.4099250E+00	0.1538053E+00
2	0.4399300E+00	0.2094933E+00
3	-0.1359078E+01	0.2205692E+00
4	0.1598513E+01	0.2094523E+00
5	0.3996215E+00	0.7012840E-01
6	0.2862925E+00	0.4315028E-01
7	0.2123224E+00	0.6331140E-01

LOG10 LIKELIHOOD = -0.1132857E+04 -2 LN LIKELIHOOD = 0.5216997E+04

*** STANDARD ERROR CALCULATION SUCCESSFUL ***

measured genotype analysis

exact likelihood solution (Elston and Stewart 1971)

computational complexity increases exponentially with number of loci

cope with missing data (genotypes) “well”

measured haplotype analysis

particularly relevant for *cis* models

exact solution for small problems (4-5 loci)

approximate solution for larger problems (>6 loci)

cope with missing data (genotypes) “quite well”

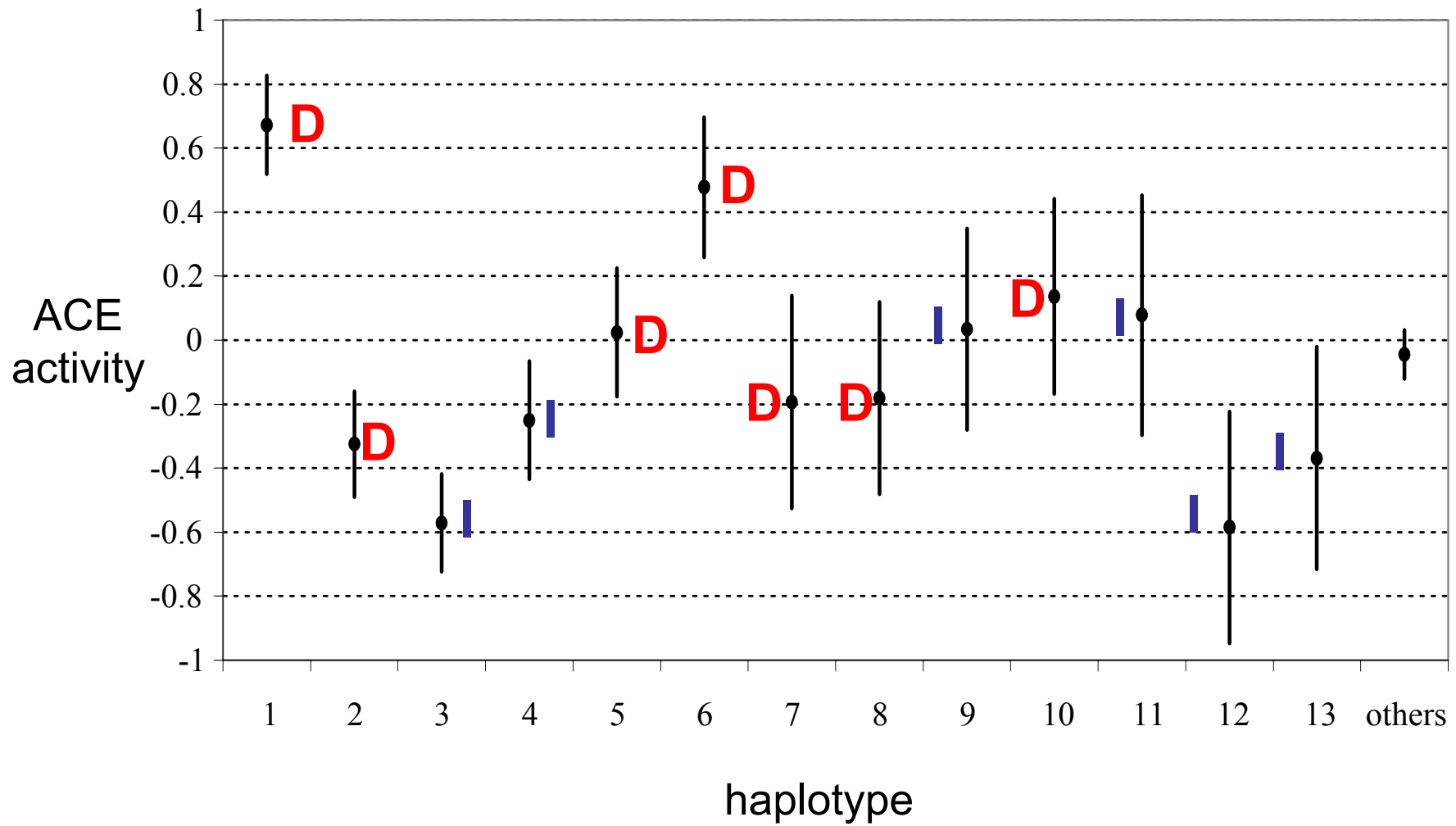
measured haplotype analysis

- assume no intragenic recombination
- compile list of alternative phase solutions
- analyse pairs of haplotypes as “genotypes”
- may be impractical to test for dominance
 - can test for epistasis

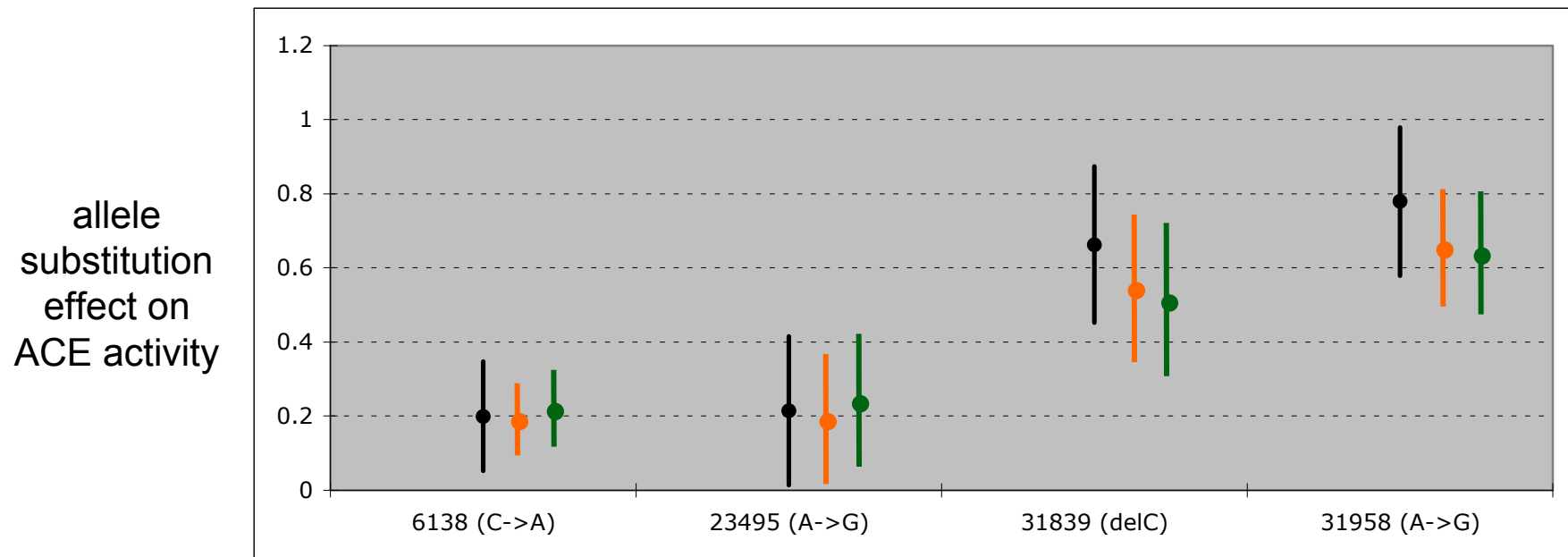
15 locus analysis of ACE

	haplotype	ace1	ace2	ace4	s1.2	I/D	ace8	19942	20120	20326	20373	20833	22982	23152	23202	23949	frequency
1	1567	1	1	1	1	D	2	1	1	1	1	2	2	2	2	1	0.10
2	1096	1	1	1	1	D	1	1	1	2	1	1	1	2	2	2	0.09
3	28804	2	2	2	1	I	1	1	2	1	1	1	1	1	2	2	0.08
4	28807	2	2	2	1	I	1	1	2	1	1	1	1	2	2	1	0.07
5	1311	1	1	1	1	D	1	2	1	1	1	2	2	2	2	1	0.06
6	1568	1	1	1	1	D	2	1	1	1	1	2	2	2	2	2	0.06
7	1031	1	1	1	1	D	1	1	1	1	1	1	1	2	2	1	0.03
8	1303	1	1	1	1	D	1	2	1	1	1	2	1	2	2	1	0.03
9	16519	2	1	1	1	I	1	1	2	1	1	1	1	2	2	1	0.02
10	3359	1	1	1	2	D	1	2	1	1	1	2	2	2	2	1	0.02
11	135	1	1	1	1	I	1	1	2	1	1	1	1	2	2	1	0.02
12	8199	1	2	1	1	I	1	1	1	1	1	1	1	2	2	1	0.02
13	30855	2	2	2	2	I	1	1	2	1	1	1	1	2	2	1	0.02

15 locus measured haplotype analysis of ACE



measured genotype vs. **measured haplotype** vs.
mixed linear model analysis



data show point estimates and 95% confidence intervals

- linear statistical models (ANOVA)
 - application to clustered data (sibships/twins)
 - SAS (www.sas.com)

- family-based models of linkage and association
 - QTDT (www.sph.umich.edu/csg/abecasis/QTDT/)

- multilocus models of linkage/association
 - measured genotype and haplotype analysis
 - PAP - Pedigree Analysis Package

 - (hasstedt.genetics.utah.edu/pap5/)