


Multilevel Model Application to Genetic Analysis Workshop Data



Jian Wang
September 4, 2012

Genetic Analysis Workshop

- ❑ The Genetic Analysis Workshops (GAW), which began in 1982, were initially motivated by the development and publication of several new algorithms for statistical genetic analysis, using different methods of analysis, had reached contradictory conclusions
- ❑ More than a year before each Genetic Analysis Workshop, suggestions for topic and appropriate data sets are solicited from people on the GAW mailing list. Topics are chosen and a small group of organizers is selected by the GAW Advisory Committee. Data sets are assembled, and six or seven months before each GAW, data are distributed.
- ❑ Investigators who wish to participate in GAW submit written contributions approximately 6-8 weeks before the Workshop.
- ❑ The proceedings of each GAW are published.
- ❑ <http://www.gaworkshop.org/>

GAW 18 Data

- Whole sequence data in a pedigree-based sample, longitudinal phenotype data for hypertension and related traits:
 - Data are obtained from two studies: the San Antonio Family Heart Study and the San Antonio Family Diabetes/Gallbladder Study
 - 1043 individuals from 20 Mexican American families
 - Systolic blood pressure, diastolic blood pressure, hypertension, smoking and antihypertensive medication
 - Four time points

	Exam 1	Exam 2	Exam 3	Exam 4
N*	855	605	622	233
Year	1981# - 1996	1997 - 2000	1998 - 2006	2009 - 2011
Age	39.6 (16 - 94)	42.9 (17 - 97)	46.3 (18 - 95)	50.9 (30 - 81)
SBP	122 (80 - 216)	125 (90 - 211)	125 (76 - 220)	128 (93 - 233)
DBP	71 (40 - 123)	72 (43 - 115)	71 (32 - 108)	78 (46 - 126)
Meds (%)	9.79	18.97	28.75	43.29
Hypertension (%)	18.13	28.38	34.77	51.93
Smoking (%)	22.90	18.25	20.00	11.16

* Number with blood pressure measurements.

GAW 18 Data

- 200 replicates of simulated longitudinal phenotype data that utilizes the real genotypes, pedigree structures and trait distributions.
 - Based on real pedigrees and the cleaned imputed sequence data
 - Gene expression levels were used to select 'functional' genes for the phenotype simulation.
 - PolyPhen was used to identify potentially deleterious coding variants
 - There are 1243 variants in 245 genes influencing DBP and 1040 variants in 205 genes influencing SBP.
 - Smoking is not associated with SBP or DBP
 - $SBP > 140$ or $DBP > 90$ -> hypertension=1
 - A proportion of hypertensive individuals were then chosen to be 'treated', and their SBP and DBP were decreased by 6.2 and 7.9 respectively. (So I did not adjust them anymore)

GWAS Data

- I focused on GWAS data
 - Based on 1457 variants used for simulation (causal variants), extracted the SNPs available in GWAS data: 149 SNPs
 - Based on 849 individuals in simulation data, extracted individuals with same parents as “siblings” and removed the parents: 741 individuals and 310 sibling groups

- For the purpose of demonstration, I only consider one replicate
 - 3 follow-ups
 - Consider SBP only
 - Additive genetic model

Read File

```
. clear
. insheet using simu1.csv
(160 vars, 2223 obs)
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	2223	371	213.9562	1	741
timeid	2223	2	.8166803	1	3
newid	2223	153.4993	92.20628	1	310
age	2223	41.40644	16.53089	11.1	98.96
sex	2223	1.581646	.4933998	1	2
dbp	2223	72.16345	9.169359	34.22552	102.1596
sbp	2223	125.4371	15.33099	71.67409	185.636
htn	2223	.2379667	.4259345	0	1
bpmed	2223	.1907332	.3929676	0	1
smoke	2223	.2226721	.4161335	0	1
rs2246732	2214	.8441734	.7092571	0	2
rs4654736	2109	.0014225	.0376978	0	1
rs2902667	2214	.5420054	.619605	0	2
rs520713	2109	.0256046	.1579899	0	1
rs35659744	2109	.3271693	.5264598	0	2
rs2231863	2211	.1573948	.3894692	0	2
rs11247653	2211	.5983718	.6350808	0	2
rs4313386	2214	.7113821	.6724093	0	2
rs926830	2109	.4039829	.5786388	0	2
rs17522918	2103	.4664765	.6002642	0	2
rs5174	2214	.3672087	.5674	0	2
rs1137100	2187	.515775	.5997124	0	2
rs1137101	2211	.7598372	.6778571	0	2
rs8179183	2109	.2645804	.4958864	0	2
rs1166698	2103	.2482168	.4639378	0	2
rs3754131	2100	.9185714	.6808745	0	2
rs186724	2211	.2320217	.4679661	0	2
rs580183	2214	.3116531	.5106013	0	2
rs16833336	2106	.0683761	.2524501	0	1
rs12038198	2109	.3328592	.5201503	0	2

Intraclass Correlation Coefficient (ICC)

```
. xtreg sbp, i(indid) mle
```

```
Iteration 0: log likelihood = -8600.2958  
Iteration 1: log likelihood = -8600.2936
```

```
Random-effects ML regression      Number of obs   =    2223  
Group variable: indid            Number of groups =     741
```

```
Random effects u_i ~ Gaussian    Obs per group:  min =     3  
                                avg =    3.0  
                                max =     3
```

```
Log likelihood = -8600.2936      Wald chi2(0)    =     0.00  
                                Prob > chi2      =     .
```

sbp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_cons	125.4371	.5086195	246.62	0.000	124.4402	126.4339
/sigma_u	13.04114	.3830374			12.3116	13.81391
/sigma_e	8.053716	.1479289			7.768938	8.348934
rho	.723912	.0143336			.6951394	.7512835

```
Likelihood-ratio test of sigma_u=0: chi bar2(01)= 1244.04 Prob>=chi bar2 = 0.000
```

Proportion Third-Level Variance (PTLV)

. xtmixed sbp||newid:||indid:

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0: log restricted-likelihood = -8560.8167
 Iteration 1: log restricted-likelihood = -8560.816

Computing standard errors:

Mixed-effects REML regression Number of obs = 2223

Group Variable	No. of Groups	Observations per Group		
		Minimum	Average	Maximum
newid	310	3	7.2	33
indid	741	3	3.0	3

sbp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
_cons	124.9235	.6802464	183.64	0.000	123.5902 126.2568

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]
newid: Identity sd(_cons)	8.852324	.7239131	7.541347 10.3912
indid: Identity sd(_cons)	9.958309	.42213	9.164386 10.82101
sd(Residual)	8.053721	.1479305	7.768939 8.348942

LR test vs. linear regression: chi2(2) = 1323.41 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

```
. mat list e(V)
symmetric e(V)[4,4]
      sbp:      lns1_1_1:      lns2_1_1:      lnsig_e:
      _cons      _cons      _cons      _cons
      sbp:_cons      .46273522
lns1_1_1:_cons      0      .00668741
lns2_1_1:_cons      0      -.00134029      .00179689
lnsig_e:_cons      0      4.033e-10      -.00007356      .00033738

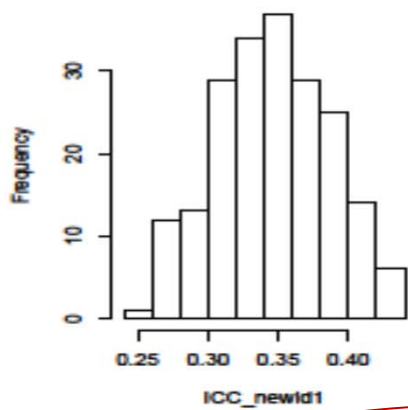
. local var_newid exp([lns1_1_1]_b[_cons])^2
. local var_id exp([lns2_1_1]_b[_cons])^2
. local var_e exp([lnsig_e]_b[_cons])^2
. nlcom (PTLV: `var_newid' / (`var_e' + `var_id' + `var_newid'))
      PTLV: exp([lns1_1_1]_b[_cons])^2 / (exp([lnsig_e]_b[_cons])^2 + exp([lns2_1_1]_b[_cons])^2 + exp([lns1_1_1]_b[_cons])^2)
```

sbp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
PTLV	.3232904	.0414687	7.80	0.000	.2420132 .4045676

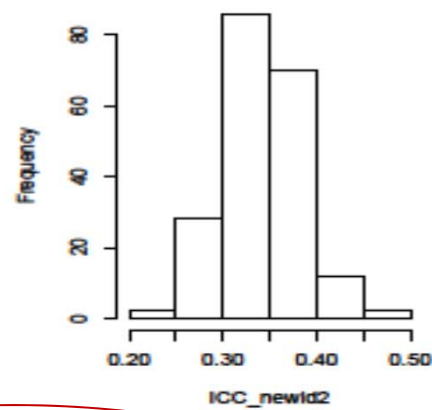
Histograms of ICCs and PTLV



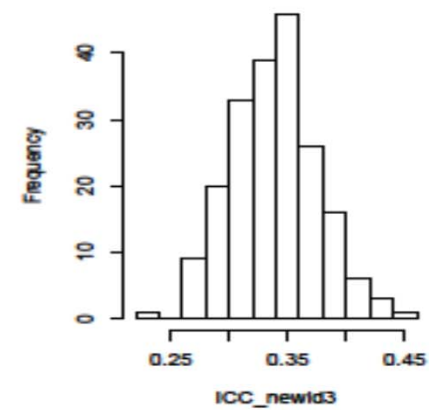
Histogram of ICC_newid1



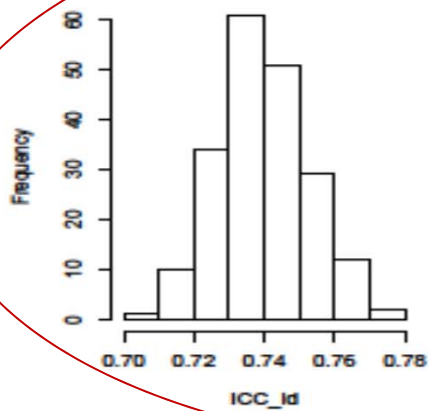
Histogram of ICC_newid2



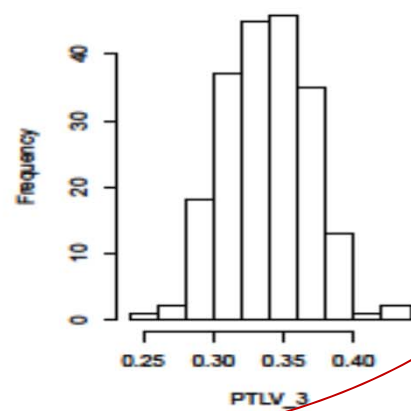
Histogram of ICC_newid3



Histogram of ICC_id



Histogram of PTLV_3



Comparison of 2 and 3 level models

- R: lme4 package
- Likelihood ratio test
- The 3 level model is better fit

- `> model1<-lmer(sbp~1+(1|indid),data=data)`
- `> model2<-lmer(sbp~1+(1|indid)+(1|newid),data=data)`
- `> anova(model1,model2) # the model 2 is better`

- Data: data
- Models:
- model1: `sbp ~ 1 + (1 | indid)`
- model2: `sbp ~ 1 + (1 | indid) + (1 | newid)`
- | | Df | AIC | BIC | logLik | Chisq | Chi Df | Pr(>Chisq) |
|--------|----|-------|-------|---------|--------|--------|---------------|
| model1 | 3 | 17207 | 17224 | -8600.3 | | | |
| model2 | 4 | 17131 | 17154 | -8561.3 | 77.889 | 1 | < 2.2e-16 *** |
- ---
- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Add Covariates in the Model

- STATA
- Using the proportional reduction in error variance (PREV)

$$\begin{aligned} \text{EV}(\text{null model}) &= 8.85^2 + 9.96^2 + 8.05^2 \\ &= 242.33 \end{aligned}$$

$$\begin{aligned} \text{EV}(\text{full model}) &= 8.7^2 + 10^2 + 8.06^2 \\ &= 240.65 \end{aligned}$$

$$\begin{aligned} \text{PREV} &= (242.33 - 240.65) / 242.33 \\ &= 0.7\% \end{aligned}$$

0.7% variance reduction is Achieved when including the SNP of interest.

```
. xtmixed sbp rs3006475 ||newid: ||indid:
Performing EM optimization:
Performing gradient-based optimization:
Iteration 0: log restricted-likelihood = -8523.9021
Iteration 1: log restricted-likelihood = -8523.9013
Computing standard errors:
Mixed-effects REML regression              Number of obs   =   2214
```

Group Variable	No. of Groups	Observations per Group Minimum	Average	Group Maximum
newid	309	3	7.2	33
indid	738	3	3.0	3

```
Log restricted-likelihood = -8523.9013          Wald chi2(1) = 4.59
                                                Prob > chi2 = 0.0322
```

sbp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
rs3006475	-2.819183	1.31603	-2.14	0.032	-5.398555 - .2398102
_cons	125.4396	.7117887	176.23	0.000	124.0446 126.8347

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]
newid: Identity			
sd(_cons)	8.679961	.7308727	7.359438 10.23743
indid: Identity			
sd(_cons)	10.0007	.4253132	9.200899 10.87003
sd(Residual)	8.057324	.1482972	7.771847 8.353288

```
LR test vs. linear regression: chi2(2) = 1299.31 Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

Add Covariates in the Model

- ❑ `> model2.1<-lmer(sbp~snp+(1|indid)+(1|newid),data=data)`
- ❑ `> model2.2<-lmer(sbp~snp+age+(1|indid)+(1|newid),data=data)`
- ❑ `> model2.3<-lmer(sbp~snp+age+sex+(1|indid)+(1|newid),data=data)`
- ❑ `> model2.4<-lmer(sbp~snp+age+sex+smoke+(1|indid)+(1|newid),data=data)`
- ❑ `> anova(model2,model2.1,model2.2,model2.3,model2.4)`

❑ Data: data

❑ Models:

❑ model2: `sbp ~ 1 + (1 | indid) + (1 | newid)`

❑ model2.1: `sbp ~ snp + (1 | indid) + (1 | newid)`

❑ model2.2: `sbp ~ snp + age + (1 | indid) + (1 | newid)`

❑ model2.3: `sbp ~ snp + age + sex + (1 | indid) + (1 | newid)`

❑ model2.4: `sbp ~ snp + age + sex + smoke + (1 | indid) + (1 | newid)`

	Df	AIC	BIC	logLik	Chisq	Chi	Df	Pr(>Chisq)
model2	4	17131	17154	-8561.3				
model2.1	5	17061	17090	-8525.6	71.4593	1		< 2.2e-16 ***
model2.2	6	16696	16730	-8342.0	367.1482	1		< 2.2e-16 ***
model2.3	7	16680	16720	-8332.9	18.2250	1		1.963e-05 ***
model2.4	8	16682	16728	-8332.9	0.0013	1		0.9711

❑ ---

❑ Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model Adequacy

- Model adequacy can be evaluated using residuals, much like in conventional regression analysis.
- In two-level modeling:
 - Total: $\hat{e}_{ij} = y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 X_{1,ij} - \dots - \hat{\beta}_p X_{p,ij}$
 - Level 1: $\hat{\varepsilon}_{ij} = \hat{e}_{ij} - \hat{\zeta}_j$
 - Level 2: $\hat{\zeta}_j$
- The level 2 residuals can be obtained as the so-called empirical Bayes estimates by Stata internally and automatically.
- The level 1 residuals were obtained by subtraction
- If the model is adequate, we will expect the standardized level 1 and level 2 residuals to follow approximately a normal distribution.
- Our example have three level.

Model Adequacy

□ STATA: gllamm

```
. gllamm sbp rs3006475 age sex, i(indid newid) adapt
```

```
Running adaptive quadrature
Iteration 0: log likelihood = -8839.9869
Iteration 1: log likelihood = -8527.6759
Iteration 2: log likelihood = -8340.3592
Iteration 3: log likelihood = -8333.2968
Iteration 4: log likelihood = -8332.9295
Iteration 5: log likelihood = -8332.9295
```

```
Adaptive quadrature has converged, running Newton-Raphson
Iteration 0: log likelihood = -8332.9295
Iteration 1: log likelihood = -8332.9295 (backed up)
Iteration 2: log likelihood = -8332.9288
```

```
number of level 1 units = 2214
number of level 2 units = 738
number of level 3 units = 309
```

```
Condition Number = 173.12045
```

```
gllamm model
```

```
log likelihood = -8332.9288
```

	sbp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
	rs3006475	-3.173901	1.124418	-2.82	0.005	-5.377721 -.9700815
	age	.5000678	.0241469	20.71	0.000	.4527408 .5473949
	sex	-3.677559	.8571906	-4.29	0.000	-5.357621 -1.997496
	_cons	111.1393	1.750141	63.50	0.000	107.7091 114.5695

```
Variance at level 1
```

```
-----
57.147965 (2.1058327)
```

```
Variances and covariances of random effects
```

```
***|level 2 (indid)
```

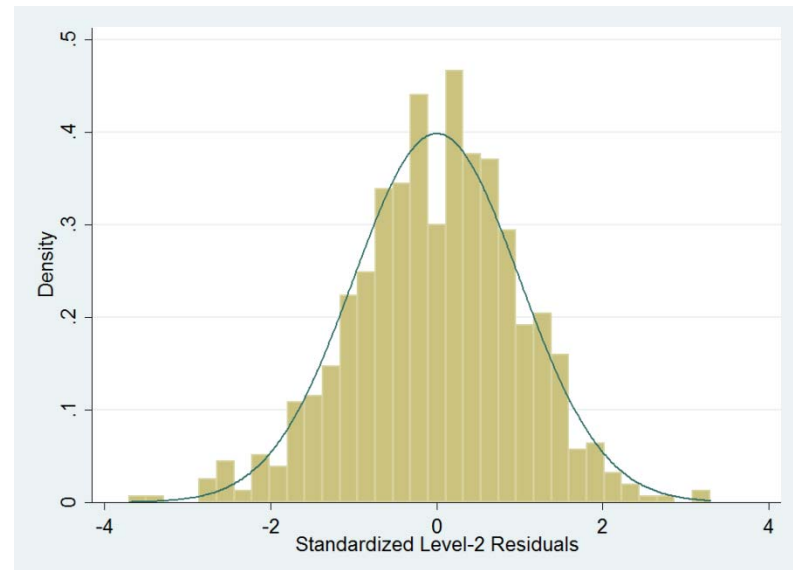
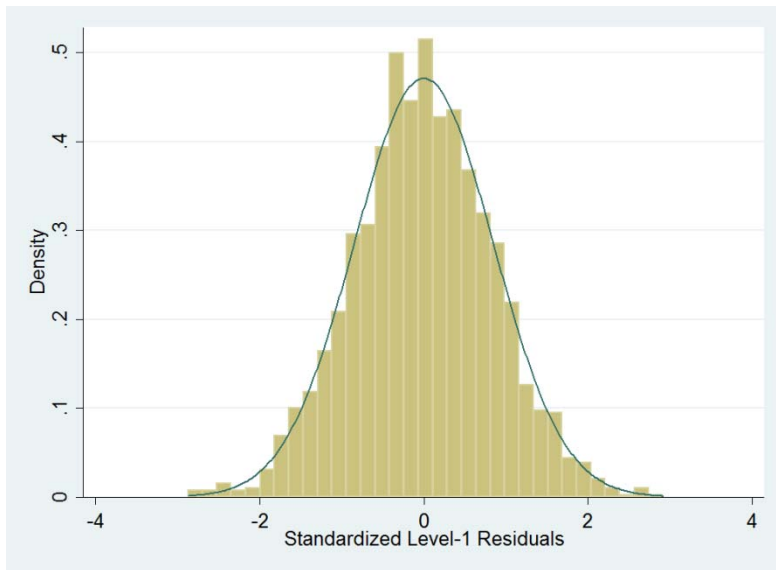
```
var(1): 94.039091 (7.3078168)
```

```
***|level 3 (newid)
```

```
var(1): 22.128565 (6.2301664)
-----
```

Model Adequacy

- STATA: gllamm



GWAS Data Analysis: Power

- Test each SNP at a time; 3 level model; use age and sex as covariates

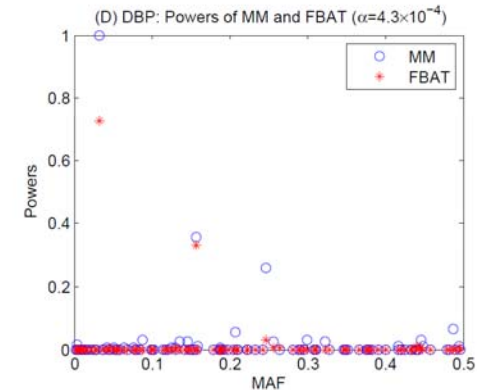
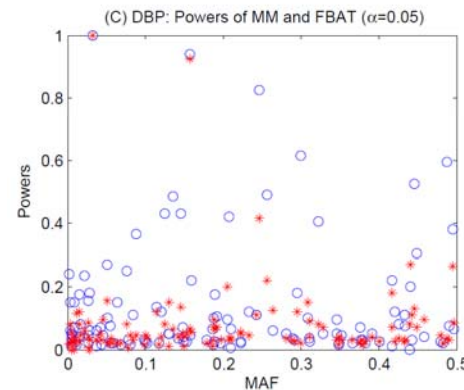
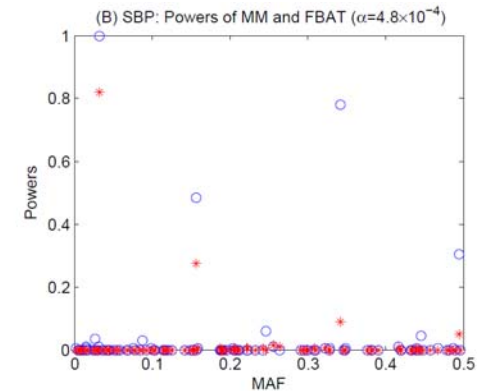
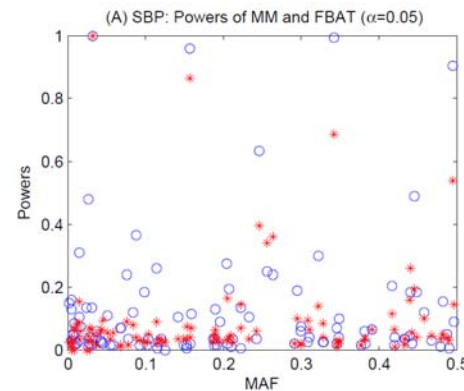
chr	pos	rs	genes	maf	true_b	b_mean	se_mean	p_mean	b_median	se_median	p_median	<0.05	<0.05/149
3	48040283	rs11711953	MAP4	0.0318	-9.9107	-17.0780	1.8947	0.0000	-17.0856	1.8923	0.0000	1	1
3	47958037	rs1060407	MAP4	0.342	-0.0002	-2.9272	0.7053	0.0014	-2.9751	0.7072	0.0000	0.995	0.755
1	66075952	rs8179183	LEPR	0.1567	3.8730	3.2252	0.9360	0.0080	3.2475	0.9374	0.0005	0.96	0.45
3	58109162	rs1131356	FLNB	0.4947	1.0007	-2.0033	0.6509	0.0165	-1.9847	0.6511	0.0021	0.905	0.255
3	141162185	rs16851435	ZBTB38	0.246	-0.0083	-1.8198	0.8183	0.0724	-1.7940	0.8195	0.0306	0.635	0.055
19	55598724	rs1054940	EPS8L1	0.4456	0.5530	1.2477	0.6565	0.1378	1.2800	0.6559	0.0516	0.49	0.025
9	15459821	rs3087653	SNAPC3	0.0264	-0.6029	-4.6878	2.3529	0.1353	-4.5732	2.3511	0.0558	0.48	0.025
1	175105996	rs2072036	TNN	0.1147	0.0000	1.8761	1.0516	0.1669	1.8277	1.0515	0.0784	0.385	0.005
1	153344636	rs3006475	S100A12	0.0877	-0.3090	-1.8832	1.1269	0.1950	-1.8799	1.1269	0.0867	0.365	0.03
17	39674641	rs1050784	KRT15	0.2036	0.0000	-1.2243	0.7933	0.2238	-1.2902	0.7931	0.1016	0.32	0.01
7	75442723	rs11465293	CCL24	0.0148	-1.1227	-5.1634	3.2626	0.2032	-5.3044	3.2612	0.1045	0.31	0.005
5	118675901	rs1355124	TNFAIP8	0.3221	-0.0602	-1.0474	0.6751	0.2127	-1.0664	0.6753	0.1179	0.3	0.005
9	124065224	rs2230287	GSN	0.2041	-0.6470	-1.2104	0.8236	0.2370	-1.2344	0.8236	0.1343	0.275	0.005
1	151496718	rs16833336	CGN	0.0216	0.0000	2.3795	1.8307	0.2952	2.3380	1.8324	0.2050	0.265	0.005
19	10088271	rs2161468	COL5A3	0.2304	0.0000	1.1279	0.7907	0.2520	1.1411	0.7904	0.1601	0.26	0.015
19	19017862	rs10330	COPE	0.114	0.0005	-1.5509	1.0533	0.2316	-1.5412	1.0519	0.1447	0.26	0
3	141058687	rs4683602	ZBTB38	0.256	0.1075	-0.9783	0.6962	0.2364	-0.9794	0.6945	0.1597	0.25	0.005
1	27951127	rs2231863	FGR	0.076	0.0812	1.5170	1.1820	0.2780	1.5400	1.1810	0.1839	0.24	0.005
3	58192585	rs9815775	DNASE1L3	0.2637	0.0496	-1.0612	0.7842	0.2694	-1.0816	0.7837	0.1693	0.24	0
1	186316488	rs3753565	TPR	0.0506	0.0000	-1.8495	1.4406	0.2921	-1.8308	1.4389	0.2071	0.235	0
1	78392446	rs1166698	NEXN	0.1298	0.0000	1.2910	1.0047	0.2982	1.3355	1.0056	0.1803	0.23	0
13	28016521	rs7988222	MTIF3	0.2887	0.0000	-0.9407	0.7045	0.2703	-0.9285	0.7048	0.1841	0.23	0
3	56771251	rs3772219	ARHGEF3	0.4891	0.0000	-0.8724	0.6455	0.2674	-0.8595	0.6453	0.1857	0.22	0.01
13	28624294	rs1933437	FLT3	0.4167	1.7874	0.9219	0.6804	0.2665	0.9184	0.6796	0.1822	0.205	0.01
1	45987574	rs17522918	PRDX1	0.2071	0.0911	-1.0460	0.7953	0.2815	-1.0278	0.7955	0.1899	0.195	0
1	66036441	rs1137100	LEPR	0.295	0.0030	-0.9509	0.7528	0.2883	-1.0065	0.7521	0.1733	0.19	0
1	204379452	rs2089891	PPP1R15B	0.0264	0.0000	-2.4949	2.1257	0.3311	-2.4261	2.1289	0.2492	0.19	0
1	66058513	rs1137101	LEPR	0.4402	0.0470	-0.7463	0.6617	0.3310	-0.7527	0.6615	0.2471	0.185	0.005
11	72946204	rs1626154	P2RY2	0.0985	-0.5943	-1.2623	1.0226	0.3111	-1.2131	1.0245	0.2355	0.185	0.005
17	17409560	rs7946	PEMT	0.4484	0.0032	-0.8246	0.6487	0.2908	-0.8251	0.6480	0.1920	0.185	0
1	153507176	rs2228293	S100A6	0.0049	-0.0368	-5.4328	4.3902	0.3009	-5.3627	4.3962	0.2200	0.16	0
17	1673276	rs1136287	SERPINF1	0.4821	0.0237	-0.7249	0.6414	0.3445	-0.7411	0.6411	0.2260	0.155	0
1	24077987	rs520713	TCEB3	0.0018	0.0001	-3.7187	3.1273	0.3126	-3.6358	3.1284	0.2378	0.15	0.005

Comparison with FBAT

- FBAT: family-based association test
- The test statistic U is based on a linear combination of offspring genotypes and traits:
 - $U = S - E[S], \quad S = \sum_{ij} T_{ij} X_{ij},$
 - X_{ij} is some function of the genotypes of the j th offspring in family i at the locus being tested.
 - T_{ij} is the coded trait. In general, the coding for T_{ij} is specified as $Y_{ij} - u_{ij}$. Here, Y_{ij} denotes the observed trait and u_{ij} is seen as an offset value.
 - $E(U) = 0$
 - $V = \text{Var}(U) = \text{Var}(S)$ can be calculated under the null.
 - If X_{ij} is a scalar summary of an individual's genotype, then $Z = U / \sqrt{V}$ is approximately $N(0, 1)$
 - If X_{ij} is a vector, then $\chi^2 = U'V^{-1}U$ has an approximate χ^2 distribution with df equal to the rank of V .

Power Comparisons

- 105 SNPs causal for SBP (panels (A) and (B))
- 117 SNPs causal for DBP (panels (C) and (D))
- Two significance levels
 - $\alpha_1=0.05$
 - $\alpha_2=0.05/\text{number of SNPs tested}$
- Multilevel model has relatively higher or comparable powers for most of the causal SNPs.
- For most of the causal SNPs, both approaches have poor powers.
- When the bonferroni-corrected significance levels were used, both approaches have no to very little powers.
- The MAFs did not have a big impact

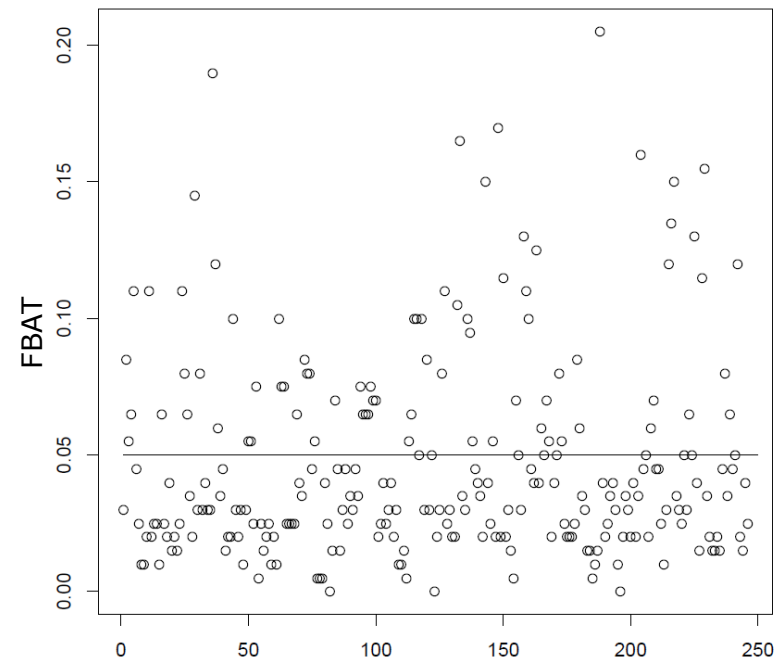
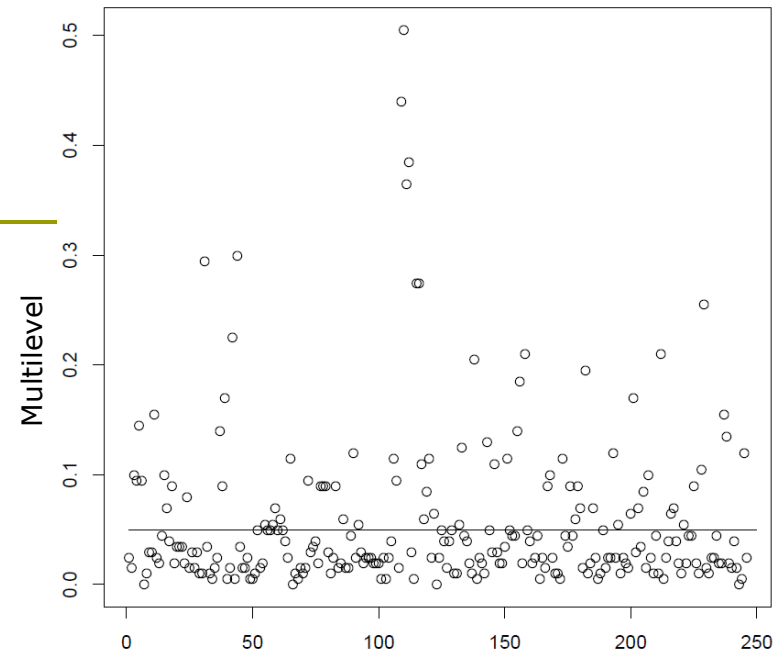


GWAS Data Analysis: Type I Error

- In the GWAS data, I selected null SNPs which are in linkage equilibrium with 149 causal SNPs
 - $r^2 < 0.01$
- **Linkage disequilibrium**
 - The non-random association of alleles at two or more loci, that may or may not be on the same chromosome.
 - In other words, linkage disequilibrium is the occurrence of some combinations of alleles or genetic markers in a population more often or less often than would be expected from a random formation of haplotypes from alleles based on their frequencies.
 - The amount of linkage disequilibrium depends on the difference between observed and expected (assuming random distributions) allelic frequencies.
 - Populations where combinations of alleles or genotypes can be found in the expected proportions are said to be in linkage equilibrium.

Type I Errors

- 1506 SNPs with $r^2 < 0.01$ for all 149 causal SNPs
- 246 SNPs with $MAFs > 0.05$
- Using SBP as phenotype
- Significance level: 0.05
 - Multilevel model: 76 SNPs have inflated type I errors
 - FBAT: 77 SNPs have inflated type I errors
- Bonferroni-corrected sig levels:
 - Multilevel model: 11 SNPs have inflated type I errors
 - FBAT: 8 SNPs have inflated type I errors
- The type I errors for both approaches are comparable





□ Thanks!