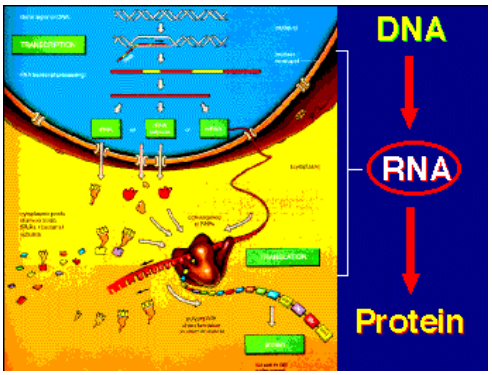
Pharmacogenomics

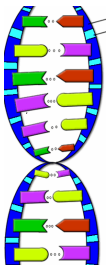
Yossi Levy

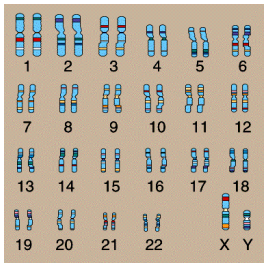
Central Dogma of Molecular Biology



Genome and DNA

- **Genome** – contains all biological information
- Biological information is encoded in **DNA**
- DNA is divided to discrete units called **Genes**
- Genes are packed into **Chromosomes**
- DNA is made of four bases: A, G, C and T





Alleles and expression

- Each gene is represented by two copies, called **Alleles**
- **Genotype –** Combination of alleles
- **Homozygous** gene – both alleles are the same
- **Heterozygous** gene – alleles are different
- **Phenotype-** expression of genotype
- **A dominant allele** is almost always expressed
- **A recessive allele** is expressed only if there are two copies of that allele

Polymorphism

- Some expressed traits are attributed to variation in DNA sequence
- When two individuals display different phenotypes in the same trait, they have two different alleles in the same gene.
- That gene is therefore said to be **polymorphic**.

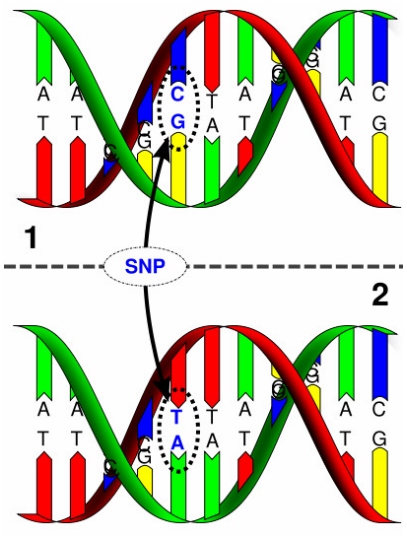
The Human Genome

- 46 chromosomes – 23 pairs
- 2 meters of DNA
- 3 billion DNA bases
- 25000 genes
- 10 million SNPs

"*Genome*, Henderson! We're working on human *genome*!"

SNPs

- **SNP** - Single Nucleotide Polymorphism
- Most common type of genetic variation
- Each SNP represent a difference in a single DNA base
- The SNP in the picture is CT or AC or AG or GT – they are all the same
- SNP can have 3 possible values: AA, Aa or aa



Types of genetics studies

Studies to investigate genotype-trait association within a population of **unrelated individuals**:

- Candidate polymorphism studies
- Candidate gene studies
- Fine mapping studies
- Gnome-wide association studies (GWAS)

Candidate polymorphism studies

- Consider polymorphism(s) within a gene
- There is an a priori hypothesis about functionality
- Primary hypothesis: the variable site under investigation is **functional**.
- That is, the given SNP (or set of SNPS) influence the disease trait directly

Candidate gene studies

- Consider multiple SNPs within a gene
- SNPs are not assumed to be functional
- However, the selected SNPs may be associated to a functional SNP within the gene
- This association is called Linkage Disequilibrium

Fine mapping studies

- Set to identify with a high level of accuracy the location of a disease-causing variant

Gnome Wide Association Studies

- Similar to candidate gene approach
- Aim to identify association between SNPs and trait
- Less hypothesis driven
- Involves the characterization of a much larger number of SNPs

Hardy-Weinberg Equilibrium

- A theoretical description of the relationship between genotype and allele frequencies
- HWE denotes independence of the alleles at a single site between two homologous chromosomes
- Let p be the frequency of the dominant allele A and q and let be the frequency of the recessive allele a (p+q=1).
- The expected genotype frequencies are:

$$p_{AA} = p^2$$
$$p_{Aa} = 2pq = 2p(1-p)$$
$$p_{aa} = q^2 = (1-p)^2$$

Testing HWE

		Homolog 2		
		A	a	
Homolog 1	A	n_{11}	n_{12}	$n_{1.}$
	a	n_{21}	n_{12}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	n

- n_{12} and n_{21} are not observed. Only $n^*_{12} = n_{12}+n_{21}$ is known
- p_A is estimated by $(2n_{11} + n^*_{12})/2n$
- Using the estimate for p_A we can calculate the expected counts E_{11}, E^*_{12} and E_{22} corresponding to n_{11}, n^*_{12} and n_{22} and construct a goodness of fit Chi-square test
- Another option is using Fisher's Exact test

Example

Genotype	AA	AC	CC
Count (n_i)	48	291	724
Expected (O_i)	35.22	316.55	711.22

$$p_A = \frac{2 \cdot 48 + 291}{2 \cdot 1063} = 0.182$$ MAF – Minor Allele Frequency

$$O_{AA} = 1063 \cdot 0.182^2 = 35.22$$

$$O_{AC} = 1063 \cdot 2 \cdot 0.182 \cdot (1 - 0.182) = 316.55$$

$$O_{CC} = 1063 \cdot (1 - 0.182)^2 = 711.22$$

$$\chi^2 = \frac{(48 - 35.22)^2}{35.22} + \ldots = 6.927 > 3.84 = \chi^2_{1,0.05}$$

HWE implications

- HWE implies constant alleles frequencies over generations
- HWE is violated in the presence of **population admixture** – a situation in which mating occurs between two populations for which the allele frequencies differ
- HWE is violated in the presence of **population stratification** – combination of populations in which breeding occurs within but not between subpopulations
- HWE is violated when mating occurs between relatives

Deviation from HWE

- Check if population admixture or stratification is present
  - Approaches: covariates, PCA, MDS
- May indicate genotyping error

Linkage Disequilibrium

- Recall that in candidate gene studies and GWAS, studied SNPs may not be functional
- However, it is hoped that they are associated with the trait under consideration
- LD: an association in the alleles present at each of two sites present on a genome

Linkage Disequilibrium

Expected allele distributions under independence

		Site 2 B	Site 2 b	
Site 1	A	$n_{11} = Np_Ap_B$	$n_{12} = Np_Ap_b$	$n_{1.} = Np_A$
	a	$n_{21} = Np_ap_B$	$n_{22} = Np_ap_b$	$n_{2.} = Np_a$
		$n_{.1} = Np_B$	$n_{.2} = Np_b$	$N = 2n$

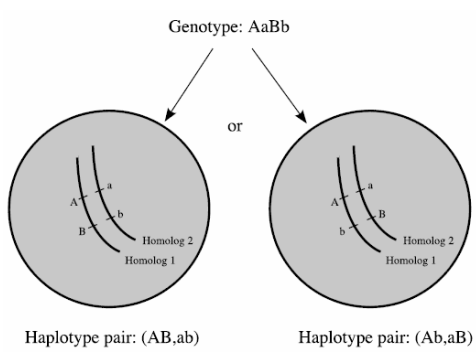
Observed allele distributions under LD

		Site 2 B	Site 2 b	
Site 1	A	$n_{11} = N(p_Ap_B + D)$	$n_{12} = N(p_Ap_b - D)$	$n_{1.}$
	a	$n_{21} = N(p_ap_B - D)$	$n_{22} = N(p_ap_b + D)$	$n_{2.}$
		$n_{.1}$	$n_{.2}$	$N = 2n$

Estimation of D

$$\hat{p}_A = n_{1.} / N \quad \hat{p}_B = n_{.1} / N \quad \hat{p}_{AB} = ???$$

The number of individuals with A and B on the same allele is not observed



Estimation of p_{AB}

Genotype counts for two biallelic loci

		Site 2		
		BB	Bb	bb
Site 1	AA	n_{11}	n_{12}	n_{13}
	Aa	n_{21}	n_{22}	n_{23}
	aa	n_{31}	n_{32}	n_{33}

$$\theta = (p_{AB}, p_{Ab}, p_{aB}, p_{ab})$$

$$\begin{aligned}\log L(\theta|n_{11},\ldots,n_{33}) \propto\ & (2n_{11}+n_{12}+n_{21})\log p_{AB} \\ & + (2n_{13}+n_{12}+n_{23})\log p_{Ab} + (2n_{31}+n_{21}+n_{32})\log p_{aB} \\ & + (2n_{33}+n_{32}+n_{23})\log p_{ab} + n_{22}\log(p_{AB}p_{ab}+p_{Ab}p_{aB})\end{aligned}$$

$p_{Ab} = p_A - p_{AB}$, $p_{aB} = p_B - p_{AB}$ and $p_{ab} = 1 - p_A - p_B - p_{AB}$.

Definition of D'

$$D' = \frac{|D|}{D_{\max}}$$

$$D_{\max} = \begin{cases} \min(p_A p_b, p_a p_B) & D > 0 \\ \min(p_A p_B, p_a p_b) & D < 0 \end{cases}$$

Another approach for LD

		Site 2		
		B	b	
Site 1	A	$n_{11} = N(p_A p_B + D)$	$n_{12} = N(p_A p_b - D)$	$n_{1.}$
	a	$n_{21} = N(p_a p_B - D)$	$n_{22} = N(p_a p_b + D)$	$n_{2.}$
		$n_{.1}$	$n_{.2}$	$N = 2n$

- Calculate "Pearson's chi-square statistic" for this table
- Define

$$r^2 = \chi^2 / N$$

- However, be aware that the "p-value" associated with the chi-square statistic is not valid

Relationship between r^2 and D

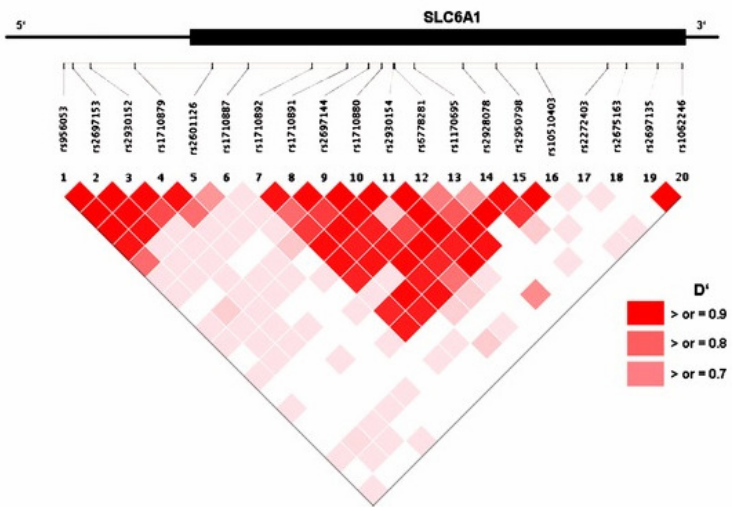
$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{il}} = \sum_{i,j} \frac{(N \cdot D)^2}{E_{il}} =$$

$$= (ND)^2 \cdot \left(\frac{1}{Np_A p_B} + \frac{1}{Np_A p_b} + \frac{1}{Np_a p_B} + \frac{1}{Np_a p_b} \right) =$$

$$= \frac{ND^2}{p_A p_B p_a p_b}$$

$$r^2 = \frac{\chi^2}{N} = \frac{D^2}{p_A p_B p_a p_b}$$

LD graphical presentation



Trait-genotype relationship

- Ultimate goal: identify SNP or set of SNPs that predict the phenotypic trait
- In pharmaceutical industry – the interesting trait is response to treatment

"My physician prescribed a customized formulation for me. Here's my DNA sequence."

Logistic regression

- Goal: relate explanatory variables x to a binary response variable y
- Let y^* be a continuous variable. It is not part of the data, only part of the model
- Model relationship between y^* and x using simple linear regression: $y^*=\beta_0+\beta_1 x+\varepsilon$
- Model the relationship between y and y^* as a function of the sign of y^*: y=1 if $y^*>0$, =0 otherwise
- Assume that the errors ε follow a logistic distribution:

$$F(t)=\frac{\exp(t)}{1+\exp(t)}$$

Logistic regression

$$P(y=1\mid x)=P(y^*>0\mid x)=$$
$$=P(\beta_0+\beta_1 x+\varepsilon>0\mid x)=$$
$$=P(\varepsilon>-(\beta_0+\beta_1 x))$$
$$=P(\varepsilon<\beta_0+\beta_1 x)=$$
$$=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$
$$\Rightarrow$$
$$\log\frac{P(y=1\mid x)}{P(y=0\mid x)}=\beta_0+\beta_1 x$$

MLE for logistic regression

$$L(\theta) = \prod [P(y_i = 1 \mid x_i)]^{y_i} [1 - P(y_i = 1 \mid x_i)]^{1-y_i}$$

$$\text{Denote } \pi_{\mathrm{i}} = P(y_i = 1 \mid x_i) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$l(\theta) = \sum \left(y_i \log \pi_{\mathrm{i}} + (1 - y_i) \log(1 - \pi_{\mathrm{i}}) \right)$$

Comparing logistic models

- Let M and M' be two logistic regression models

$$\mathrm{M}: \log \frac{P(y=1 \mid x_1, \ldots, x_p)}{P(y=0 \mid x_1, \ldots, x_p)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

$$\mathrm{M}': \log \frac{P(y=1 \mid x_1, \ldots, x_p, x_{p+1}, \ldots, x_{p'})}{P(y=0 \mid x_1, \ldots, x_p, x_{p+1}, \ldots, x_{p'})} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \beta_{p+1} x_{p+1} + \ldots + \beta_{p'} x_{p'}$$

- Let |M| and |M'| be the dimensions of the models
- Let l*(M) be the maximum value of the log-likelihood function of model M
- Let the deviance of model M be D(M)=-2l*(M)
- Since l*(M)≤l*(M') then D(M)≥D(M')
- Note that this result holds because the models are nested

Comparing logistic models

To test the hypothesis

$$H_0 : \beta_{p+1} = \ldots = \beta_{p'} = 0$$

one can use the likelihood ration statistic:

$$G^2(M \mid M') = D(M) - D(M') \xrightarrow{D} \chi^2_{p'-p}$$

Comparing logistic models

If the models are non nested, one can use:

$$AIC(M) = D(M) + 2|M|$$

$$BIC(M) = D(M) + \log(n)|M|$$

Example

Data on ~100 subjects, ~250 SNPs

	WEIGHT	HEIGHT	bsev	age	disdur	response	snp	pat	drug	sex	snpid
1	77.0	168.0	2.0	43	3.5	0	GG	378	A	F	rs10001
2	68.0	175.0	1.0	50	7.5	1	GG	379	P	M	rs10001
3	50.0	159.0	1.5	46	7.6	1	GG	380	P	F	rs10001
4	55.0	162.0	3.5	46	3.4	0	GG	381	A	F	rs10001
5	55.0	163.0	2.0	24	3.4	1	GG	383	P	F	rs10001
6	69.0	164.0	4.0	45	5.5	1	GG	384	P	F	rs10001
7	63.0	168.0	3.0	42	2.4	0	GG	385	A	F	rs10001
8	62.0	165.0	1.5	26	1.2	1	GG	386	P	M	rs10001
9	67.0	174.0	1.0	27	0.7	0	GG	387	P	F	rs10001
10	47.0	154.0	2.0	28	3.7	1	GG	388	P	F	rs10001
11	72.0	173.0	3.5	27	3.8	0	GG	389	A	F	rs10001
12	75.0	180.0	2.0	30	10.3	0	GG	397	A	F	rs10001
13	70.0	171.0	3.0	28	3.3	1	GG	398	P	M	rs10001
14	71.0	171.0	2.5	35	5.3	0	GG	399	A	F	rs10001
15	62.0	167.0	3.0	46	10.3	0	AG	390	P	F	rs10001
16	69.0	178.0	3.0	29	1.2	1	GG	391	A	M	rs10001
17	90.0	163.0	2.0	24	1.2	1	GG	392	P	F	rs10001
18	87.0	162.0	4.0	37	2.3	0	GG	394	P	F	rs10001
19	85.0	175.0	2.5	26	0.4	1	GG	396	P	F	rs10001
20	104.0	163.0	5.0	33	5.3	0	GG	454	P	F	rs10001
21	95.0	169.0	2.0	32	6.6	0	GG	455	A	F	rs10001
22	63.0	166.0	5.0	42	2.6	0	GG	456	A	F	rs10001
23	79.0	168.0	5.0	45	1.3	0	GG	458	P	F	rs10001
24	60.6	164.0	5.0	37	10.8	0	GG	459	P	F	rs10001

Data for SNP rs10014

Table 1 of snp by drug			
Controlling for resp=0			
snp	drug		
Frequency	A	P	Total
AC	6	2	8
CC	5	19	24
Total	11	21	32

Table 2 of snp by drug			
Controlling for resp=1			
snp	drug		
Frequency	A	P	Total
AC	8	11	19
CC	29	19	48
Total	37	30	67

Model 1: SNP only

The only explanatory variable is SNP

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	126.598	128.473
SC	129.193	133.664
-2 Log L	124.598	124.473

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.1243	1	0.7244
Score	0.1231	1	0.7257
Wald	0.1230	1	0.7258

Model 2: SNP and drug

Explanatory variables are SNP and drug

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	126.598	126.701
SC	129.193	134.486
-2 Log L	124.598	120.701

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3.8971	2	0.1425
Score	3.8424	2	0.1464
Wald	3.7550	2	0.1530

Model 3: add interaction and covariates

Explanatory variables are SNP, drug, SNP*drug and all covariates

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	126.598	118.902
SC	129.193	139.663
-2 Log L	124.598	102.902

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	21.6955	7	0.0029
Score	19.5362	7	0.0067
Wald	16.2817	7	0.0227

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
drug	1	0.0030	0.9564
snp	1	0.0141	0.9056
drug*snp	1	10.1119	0.0015
disdur	1	2.9776	0.0844
bsev	1	3.1186	0.0774
sex	1	1.6254	0.2023
age	1	1.1926	0.2748

Model 4: remove non-contributing covariates

Explanatory variables are SNP, drug, SNP*drug and bsev

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	126.598	117.937
SC	129.193	130.913
-2 Log L	124.598	107.937

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.6608	4	0.0022
Score	15.3764	4	0.0040
Wald	13.4176	4	0.0094

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
drug	1	0.1534	0.6953
snp	1	0.0039	0.9499
drug*snp	1	8.6785	0.0032
bsev	1	3.1997	0.0737

Typical GWAS study approach

- Data QC
  - Remove SNPs with >5% missing data and or nonrandom missingness
  - Remove SNPs with low Minor Allele Frequency
  - Remove SNPs that depart from HWE
  - Remove individuals with high percent of missing data
- Run logistic regression model for each of the SNPs
- Identify top SNPs with significant drug and SNP interaction
- Try to model interactions between top SNPs (later)
- Identify SNPs for candidate gene study

Log linear models

- Alternative approach to model association between categorical variables
- Instead of modeling the response probability, expected cell counts are modeled: $\log(m_{ij})=\ldots$

		X_2		
		1	2	
X_1	1	n_{11}	n_{12}	$n_{1.}$
	2	n_{21}	n_{12}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	n

Independence model

$$\hat{m}_{ij} = \frac{n_{i.} n_{.j}}{n}$$

$$\Rightarrow \log \hat{m}_{ij} = -\log n + \log n_{i.} + \log_{n.j}$$

$$\Rightarrow \text{Model:}$$

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} \quad i = 1,2 \;\; j = 1,2$$

$$u_{1(1)} + u_{1(2)} = 0, \quad u_{2(1)} + u_{2(2)} = 0$$

General model for 2x2 table

Saturated model:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \quad i = 1,2 \;\; j = 1,2$$

$$u_{1(1)} + u_{1(2)} = 0, \quad u_{2(1)} + u_{2(2)} = 0$$

$$u_{12(1j)} + u_{12(2j)} = 0 \text{ for } j = 1,2$$

$$u_{12(i1)} + u_{12(i2)} = 0 \text{ for } i = 1,2$$

Interpretation of parameters:

$$u = \frac{1}{4} \sum_{ij} \log m_{ij}$$

$$u_{1(1)} = \frac{1}{4} \log \frac{m_{11} m_{12}}{m_{21} m_{22}}$$

$$u_{12(11)} = \frac{1}{4} \log \frac{m_{11} m_{22}}{m_{21} m_{12}}$$

etc.

The hypothesis of independence between X_1 and X_2 is equivalent to

H0: $u_{12(11)}=0$

Simple example

		SNP		
		BB (1)	Bb or bb (2)	
Response	No (1)	2037	958	2995
	Yes (2)	1757	218	1975
		3794	1176	4970

Analysis

"Usual" chi-square analysis

Table of disease by snp			
disease	snp		
Frequency Expected	BB	Bb	Total
N	2037 2286.3	958 708.68	2995
Y	1757 1507.7	218 467.32	1975
Total	3794	1176	4970

Statistic	DF	Value	Prob
Chi-Square	1	289.1536	<.0001
Likelihood Ratio Chi-Square	1	312.4785	<.0001
Continuity Adj. Chi-Square	1	287.9950	<.0001
Mantel-Haenszel Chi-Square	1	289.0954	<.0001
Phi Coefficient		-0.2412	
Contingency Coefficient		0.2345	
Cramer's V		-0.2412	

Log-linear analysis – saturated model

Analysis of Maximum Likelihood Estimates					
Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
disease	N	0.4071	0.0204	396.27	<.0001
snp	BB	0.7103	0.0204	1206.66	<.0001
disease*snp	N BB	-0.3331	0.0204	265.39	<.0001

Connection between log-linear models and logistic regression

- Assuming independence:

$$\log\frac{P(X_1=2\mid X_2=j)}{P(X_1=1\mid X_2=j)}=\log\frac{P(X_1=2,X_2=j)}{P(X_1=1,X_2=j)}=$$
$$=\log\frac{p_{2j}}{p_{1j}}=\log\frac{m_{2j}}{m_{1j}}=\log m_{2j}-\log m_{1j}=$$
$$=(u+u_{1(2)}+u_{2(j)})-(u+u_{1(1)}+u_{2(j)})=$$
$$=u_{1(2)}-u_{1(1)}$$

- This is the intercept only logistic regression

$$\log\frac{P(X_1=2\mid X_2)}{P(X_1=2\mid X_2)}=\beta_0$$

What about a saturated model?

- Similarly we receive

$$\log\frac{P(X_1=2\mid X_2=1)}{P(X_1=1\mid X_2=1)}=\left(u_{1(2)}-u_{1(1)}\right)+\left(u_{12(21)}-u_{12(11)}\right)$$
$$\log\frac{P(X_1=2\mid X_2=2)}{P(X_1=1\mid X_2=2)}=\left(u_{1(2)}-u_{1(1)}\right)+\left(u_{12(22)}-u_{12(12)}\right)$$

- Which is actually a logistic regression model, with intercept an a term that depends on X_2

$$\log\frac{P(X_1=2\mid X_2)}{P(X_1=2\mid X_2)}=\beta_0+\beta_1X_2$$

3-way table

Table 1 of snp1 by snp2				
Controlling for disease=No				
snp1	snp2			
Frequency Expected	BB	Bb	bb	Total
AA	1167 1176.6	377 364.48	186 188.88	1730
Aa	763 760.39	225 235.55	130 122.07	1118
aa	107 99.98	29 30.971	11 16.05	147
Total	2037	631	327	2995

Table 2 of snp1 by snp2				
Controlling for disease=Yes				
snp1	snp2			
Frequency Expected	BB	Bb	bb	Total
AA	1509 1515.1	16 16.385	179 172.47	1704
Aa	234 226.74	2 2.4519	19 25.81	255
aa	14 15.116	1 0.1635	2 1.7206	17
Total	1757	19	200	1976

The saturated model [123]

$$\log m_{ujk} = u +$$
$$+ u_{1(i)} + u_{2(j)} + u_{3(k)} +$$
$$+ u_{12(ij)} + u_{13(ik)} + u_{23(jk)} +$$
$$+ u_{123(ijk)}$$

Saturated model results

Analysis of Maximum Likelihood Estimates					
Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
disease	No	0.9930	0.0852	135.74	<.0001
snp1	AA	1.5119	0.0894	286.33	<.0001
	Aa	0.2643	0.1120	5.57	0.0183
disease*snp1	No AA	-0.5029	0.0894	31.68	<.0001
	No Aa	0.3116	0.1120	7.74	0.0054
snp2	BB	1.5597	0.0899	301.13	<.0001
	Bb	-1.0411	0.1487	49.01	<.0001
disease*snp2	No BB	-0.4999	0.0899	30.94	<.0001
	No Bb	0.8819	0.1487	35.17	<.0001
snp1*snp2	AA BB	0.0478	0.0940	0.26	0.6112
	AA Bb	-0.1897	0.1570	1.46	0.2268
	Aa BB	0.1510	0.1163	1.69	0.1940
	Aa Bb	-0.2399	0.2035	1.39	0.2383
disease*snp1*snp2	No AA BB	-0.1187	0.0940	1.60	0.2066
	No AA Bb	0.2077	0.1570	1.75	0.1858
	No Aa BB	-0.2138	0.1163	3.38	0.0659
	No Aa Bb	0.1749	0.2035	0.74	0.3901

Independence model [1][2][3]

$$\log m_{ujk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$

Maximum Likelihood Analysis of Variance			
Source	DF	Chi-Square	Pr > ChiSq
disease	1	205.90	<.0001
snp1	2	2049.73	<.0001
snp2	2	3114.38	<.0001
Likelihood Ratio	12	1072.57	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
disease	No	0.2079	0.0145	205.90	<.0001
snp1	AA	1.3194	0.0298	1960.84	<.0001
	Aa	0.4027	0.0321	156.92	<.0001
snp2	BB	1.2461	0.0223	3112.04	<.0001
	Bb	-0.5181	0.0304	290.59	<.0001

Conditional independence model

[12][13]: conditional independence of X_2 and X_3 given X_1:

$$\log m_{ujk} = u + \\ + u_{1(i)} + u_{2(j)} + u_{3(k)} + \\ + u_{12(ij)} + u_{13(ik)}$$

Conditional independence model

Maximum Likelihood Analysis of Variance			
Source	DF	Chi-Square	Pr > ChiSq
disease	1	336.40	<.0001
snp1	2	1570.67	<.0001
snp2	2	2253.48	<.0001
disease*snp1	2	409.65	<.0001
disease*snp2	2	212.25	<.0001
Likelihood Ratio	8	8.59	0.3778

Analysis of Maximum Likelihood Estimates					
Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
disease	No	1.0909	0.0595	336.40	<.0001
snp1	AA	1.5682	0.0457	1179.22	<.0001
	Aa	0.4001	0.0489	67.00	<.0001
snp2	BB	1.6169	0.0430	1415.12	<.0001
	Bb	-1.2326	0.0792	242.06	<.0001
disease*snp1	No AA	-0.6008	0.0457	173.11	<.0001
	No Aa	0.1306	0.0489	7.14	0.0075
disease*snp2	No BB	-0.6165	0.0430	205.71	<.0001
	No Bb	1.0610	0.0792	179.37	<.0001

Two other possible models

One variable independent of two others [1][23]:

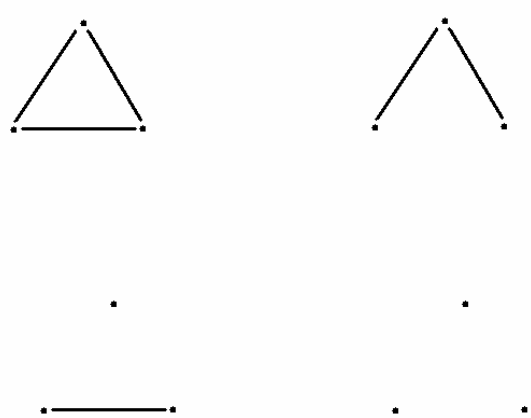
X1 is independent of {X2 , X3}

$$\log m_{ujk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}$$

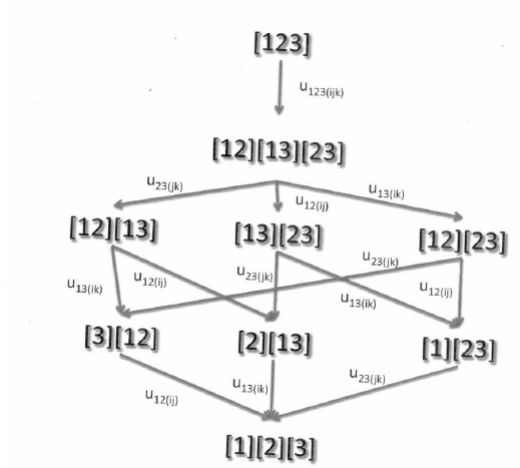
No second order interaction [12][13][23]: no clear interpretation

$$\begin{aligned}\log m_{ujk} = u &+ \\ &+ u_{1(i)} + u_{2(j)} + u_{3(k)} + \\ &+ u_{12(ij)} + u_{13(ik)} + u_{23(jk)}\end{aligned}$$

Association molecule



Model hierarchy



Bayesian approach

- The log-linear models fail when one (or more) of the cells in the contingency table has a frequency of zero
- A common fix for that is to replace the zero by 0.5 or by 1
- This approach is criticized since the data is perturbed
- A possible approach is the Bayesian approach
- The count data is multinomial, but what if we assume that the multinomial distribution parameters are also random variables?

Model setup

- Let D be the observed cell count for a 2x2 contingency table: D={n_{11}, n_{12}, n_{21}, n_{22}}
- The data D could have arisen under two hypotheses
  - H_1: X_1 and X_2 are independent
  - H_2: X_1 and X_2 are not independent
- Before seeing the observed data, we assume *a priori that both hypotheses are equally likely:*

$$P(H_1) = P(H_2) = 0.5$$

Applying Bayes theorem

$$P(H_i \mid D) = \frac{P(D \mid H_i)P(H_i)}{P(D)}$$

$$\Rightarrow \frac{P(H_2 \mid D)}{P(H_1 \mid D)} = \frac{P(D \mid H_2)P(H_2)}{P(D \mid H_1)P(H_1)} = B_{21} \cdot \frac{P(H_2)}{P(H_1)}$$

where B_{21} is the Bayes Factor

$$B_{21} = \frac{P(D \mid H_2)}{P(D \mid H_1)}$$

- The Bayes Factor represent the ratio of the posterior odds of H_1 to its prior odds

Integrated likelihood

- $P(D|H_i)$ is the integrated likelihood of D, obtained by averaging the likelihood over all possible values of the parameters under H_i.
- What are the parameters?

Modeling the prior distribution

		SNP		
		BB (1)	Bb or bb (2)	
Response	No (1)	α	α	2α
	Yes (2)	α	α	2α
		2α	2α	4α

- Before seeing the data, we have no knowledge about which combination of categories are more or less likely
- The natural way to model the distribution of the multinomial parameters is the Dirichlet distribution – an extension of the Beta distribution, as it is conjugate the Multinomial distribution

The Dirichlet Distribution

$$X = (X_1, \cdots, X_k) \sim Dirichlet(\alpha_1, \cdots, \alpha_k) \sim Dirichlet(\alpha):$$

$$f_{X_1,\cdots X_k}(x_1, \cdots, x_k \mid \alpha_1, \cdots, \alpha_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdot \ldots \cdot \Gamma(\alpha_k)} \cdot x_1^{\alpha_1 - 1} \cdot \ldots \cdot x_k^{\alpha_k - 1}$$

if $\beta \mid X \sim Multinomial(X)$

and $X \sim Dirichlet(\alpha)$

then $X \mid \beta \sim Dirichlet(\alpha + \beta)$

Assuming H_2 - interaction

$$P(D \mid p) = M \cdot p_{11}^{n_{11}} \cdot p_{12}^{n_{12}} \cdot p_{21}^{n_{21}} \cdot p_{22}^{n_{22}}$$

$$P(p_{11}, p_{12}, p_{21}, p_{22} \mid \alpha) = \frac{\Gamma(4\alpha)}{\Gamma(\alpha)^4} p_{11}^{\alpha-1} \cdot p_{12}^{\alpha-1} \cdot p_{21}^{\alpha-1} \cdot p_{22}^{\alpha-1}$$

$$P(p_{11}, p_{12}, p_{21}, p_{22} \mid D, \alpha) =$$

$$\frac{\Gamma(n + 4\alpha)}{\Gamma(n_{11} + \alpha) \cdot \Gamma(n_{12} + \alpha) \cdot \Gamma(n_{21} + \alpha) \cdot \Gamma(n_{22} + \alpha)} \cdot p_{11}^{n_{11}+\alpha-1} \cdot p_{12}^{n_{12}+\alpha-1} \cdot p_{21}^{n_{12}+\alpha-1} \cdot p_{22}^{n_{22}+\alpha-1}$$

Integrated likelihood under H_2

$$P(D \mid H_2) =$$

$$= \int p_{11}^{n_{11}} p_{12}^{n_{22}} p_{21}^{n_{21}} p_{22}^{n_{22}} P(p_{11}, p_{12}, p_{21}, p_{22} \mid \alpha) dp_{11} dp_{12} dp_{21} dp_{22} =$$

$$= \frac{\Gamma(n+4\alpha)}{\Gamma(n_{11}+\alpha)\cdot\Gamma(n_{12}+\alpha)\cdot\Gamma(n_{21}+\alpha)\cdot\Gamma(n_{22}+\alpha)} \cdot \frac{\Gamma(\alpha)^4}{\Gamma(4\alpha)}$$

Assuming H_1 - independence

- $P_{ij}=p_{i.}\cdot p_{.j}$, therefore:

$$P(D \mid p) = M \cdot p_{1.}^{n_{1.}} \cdot p_{2.}^{n_{2.}} \cdot p_{.1}^{n_{.1}} \cdot p_{.2}^{n_{.2}}$$

- Assume independent Dirichlet prior for raw and columns marginal probabilities:

$$P(p_{1.}, p_{2.} \mid \alpha) = \frac{\Gamma(4\alpha)}{\Gamma(2\alpha)^2} p_{1.}^{2\alpha-1} \cdot p_{2.}^{2\alpha-1}$$

$$P(p_{.1}, p_{.2} \alpha) = \frac{\Gamma(4\alpha)}{\Gamma(2\alpha)^2} p_{.1}^{2\alpha-1} \cdot p_{.2}^{2\alpha-1}$$

Integrated likelihood under H_2

$$P(p_{1.}, p_{2.}, p_{.1}, p_{.2} \mid D, \alpha) =$$

$$\frac{\Gamma(n+4\alpha)}{\Gamma(n_{1.}+2\alpha)\Gamma(n_{2.}+2\alpha)\Gamma(n_{.1}+2\alpha)\Gamma(n_{.2}+2\alpha)} \cdot p_{1.}^{n_{1.}+2\alpha-1} p_{2.}^{n_{2}+2\alpha-1} p_{.1}^{n_{.1}+2\alpha-1} p_{.2}^{n_{22}+2\alpha-1}$$

$$P(D \mid H_1) =$$

$$\frac{\Gamma(n+4\alpha)}{\Gamma(n_{1.}+2\alpha)\cdot\Gamma(n_{2.}+2\alpha)\cdot\Gamma(n_{.1}+2\alpha)\cdot\Gamma(n_{.2}+2\alpha)} \cdot \frac{\Gamma(2\alpha)^4}{\Gamma(4\alpha)^2}$$