Gene-gene (and gene-environment) interactions

Heather J. Cordell

Institute of Genetic Medicine
Newcastle University, UK

(Pairwise) interaction

- Statistical interaction most easily described in terms a of (logistic) regression framework
  - Supppose x_1 and x_2 are binary factors whose presence/absence (coded 1/0) may be associated with a disease outcome
  - Logistic regression models their effect on the log odds of disease as:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

Marginal effect of factor 1

$$\log \frac{p}{1-p} = \beta_0 + \beta_2 x_2$$

Marginal effect of factor 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Main effects of factors 1 and 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

Main effects and interaction term

  - For quantitative traits, use linear regression (replace $\log \frac{p}{1-p}$ with y)

Gene-gene interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data

Gene-gene interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value

Gene-gene interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value
    - Having factor 2 adds β_2 to your trait value

Gene-gene interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value
    - Having factor 2 adds β_2 to your trait value
    - Having both factors adds an additional β_{12} to your trait value
      $\Rightarrow$ Implies that the overall effect of two variables is greater (or less) than the 'sum of the parts'

Gene-gene interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value
    - Having factor 2 adds β_2 to your trait value
    - Having both factors adds an additional β_{12} to your trait value
      $\Rightarrow$ Implies that the overall effect of two variables is greater (or less) than the 'sum of the parts'

- Suppose no main effects ($\beta_1 = \beta_2 = 0$)
  - Then we have

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_{12}$	β_0
0	β_0	β_0

  - Trait value only differs from baseline if both factors present

Gene-gene interaction

- However genetic predictors e.g. SNPs are not binary, but rather take 3 levels according to the number of copies (0,1,2) of the susceptibility allele possessed
- Most general 'saturated' (9 parameter) genotype model allows all 9 penetrances to take different values
  - Via modelling log odds in terms of:
    - A baseline effect (β_0)
    - Main effects of locus G (β_{G_1}, β_{G_2})
    - Main effects of locus H (β_{H_1}, β_{H_2})
    - 4 interaction terms

	Locus H		
Locus G	2	1	0
2	$\beta_0+\beta_{G_2}+\beta_{H_2}+\beta_{22}$	$\beta_0+\beta_{G_2}+\beta_{H_1}+\beta_{21}$	$\beta_0+\beta_{G_2}$
1	$\beta_0+\beta_{G_1}+\beta_{H_2}+\beta_{12}$	$\beta_0+\beta_{G_1}+\beta_{H_1}+\beta_{11}$	$\beta_0+\beta_{G_1}$
0	$\beta_0+\beta_{H_2}$	$\beta_0+\beta_{H_1}$	β_0

  - Corresponds in statistical analysis packages to coding x_1, x_2 (0,1,2) as a "factor"

Gene-gene interaction

- Alternatively we can assume additive effects of each allele at each locus:
  - Corresponds to fitting

$$\log \frac{p}{1-p} = \beta_0 + \beta_G x_1 + \beta_H x_2 + \beta_{GH} x_1 x_2$$

with x_1, x_2 coded (0,1,2)

	Locus H		
Locus G	2	1	0
2	$\beta_0 + 2\beta_G + 2\beta_H + 4\beta_{GH}$	$\beta_0 + 2\beta_G + \beta_H + 2\beta_{GH}$	$\beta_0 + 2\beta_G$
1	$\beta_0 + \beta_G + 2\beta_H + 2\beta_{GH}$	$\beta_0 + \beta_G + \beta_H + \beta_{GH}$	$\beta_0 + \beta_G$
0	$\beta_0 + 2\beta_H$	$\beta_0 + \beta_H$	β_0

Relationship to biological interaction

- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387
  - Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171
  - Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159
  - Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540
  - Wang, Elston and Zhu (2010) Hum Hered 70:269-277

Relationship to biological interaction

- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387
  - Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171
  - Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159
  - Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540
  - Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
  - In terms of whether, for example, gene products physically interact

Relationship to biological interaction

- Much discussion in the literature
    - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387
    - Thompson (1991) J Clin Epidemiol 44:221-232
    - Phillips (1998) Genetics 149:1167-1171
    - Cordell (2002) Hum Molec Genet 11:2463-2468
    - McClay and van den Oord (2006) J Theor Biol 240:149-159
    - Phillips (2008) Nat Rev Genet 9:855-867
    - Clayton DG (2009) PLoS Genet 5(7): e1000540
    - Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
    - In terms of whether, for example, gene products physically interact
- However, existence of statistical interaction does imply both loci are "involved" in disease in some way

Relationship to biological interaction

- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387
  - Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171
  - Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159
  - Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540
  - Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
  - In terms of whether, for example, gene products physically interact
- However, existence of statistical interaction does imply both loci are "involved" in disease in some way
  - Provides a good starting point for further investigation of their (joint) involvement
    - Can be informed by the estimated penetrance values
    - Best addressed through other types of experimental data

Some references

- For more details on gene-gene (G×G) interactions (epistasis) see
  - Cordell HJ (2009) Nat Rev Genet 10(6): 392-404
  - Wei, Hemani and Haley (2014) Nat Rev Genet 15(11):722-33
- For more details on gene-environment (G×E) interactions see
  - Thomas D (2010) Nat Rev Genet 11(4): 259-272
- Conceptually many similar issues in terms of definition and mathematical modelling
  - However, many practical issues rather different

G×G versus G×E

- For G×E, we generally have to decide which environment(s) to measure/test
  - For G×G, assuming we have GWAS data, we have already measured the genetic factors of interest
- Measurement error and confounding are a bigger issue for environmental factors?
  - e.g. diet, smoking, pollution levels
  - Issues of recall bias?
  - Current genotyping platforms (though not necessarily sequencing platforms) have relatively low error rates, less prone to biases

G×G versus G×E

- Typically GWAS measure thousands if not millions of genetic variants
  - But only a few (tens or at most 100s) of environmental factors
- Feasible to consider all G×E combinations
- All pairwise G×G combinations possible, but much more time consuming
  - And leads to greater multiplicity of tests
  - Also, why stop at 2-way interactions?
    - Could look at all 3 way, 4 way etc. combinations
    - Scale of problem quickly gets out of hand
    - Less obvious reason to do this for G×E...

G×G versus G×E

- Risk estimation more important for G×E (?)
  - Estimating genetic risks in particular environments
  - Estimating effect of environmental factor on particular genetic background
    - Important for treatment/screening strategies and public health interventions
- For G×G, focus of interest is more related to
  - Increasing power to detect an effect (by taking into account the effects of other genetic loci)
  - Modelling the biology, especially related to the joint action of the loci

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone

- Depending on circumstances, any of these tests may be a sensible option

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone
- Depending on circumstances, any of these tests may be a sensible option
- Most tests of interaction/joint action can be thought of as a version of one or other of these tests
  - Although different tests vary in their precise details
  - And their relationship to the logistic regression formulation not always clearly described

Testing for interaction

- Case/control studies
  - Measure risk factors (e.g. SNP genotypes) x_1 and x_2 in sample of affected individuals (cases) and unaffected individuals (controls)
  - At each SNP each person has one of 3 possible genotypes

$$1|1, \qquad 1|2 = 2|1, \qquad 2|2$$

    - Can code as 3 different levels, count alleles or make dominance/recessive assumptions ⇒ binary factor
  - Analyse using logistic regression (e.g. in R, SAS, PLINK)
    - If use binary coding, end up with 1 interaction term
    - If genotypes coded as 3 levels, get 4 interaction terms.

Testing for interaction

- Case/control studies
  - Measure risk factors (e.g. SNP genotypes) x_1 and x_2 in sample of affected individuals (cases) and unaffected individuals (controls)
  - At each SNP each person has one of 3 possible genotypes

$$1|1, \qquad 1|2 = 2|1, \qquad 2|2$$

    - Can code as 3 different levels, count alleles or make dominance/recessive assumptions ⇒ binary factor
  - Analyse using logistic regression (e.g. in R, SAS, PLINK)
    - If use binary coding, end up with 1 interaction term
    - If genotypes coded as 3 levels, get 4 interaction terms.
- For analysis of quantitative traits in unrelated individuals, use linear rather than logistic regression
- Family studies: use extension of case/pseudo-control approach (Cordell et al. (2004) Genet Epidemiol 26:186-205) or else use linear mixed models

Case-only analysis

- Piergorsh et al. 1994; Yang et al. 1999; Weinberg and Umbach 2000
- Several authors have shown that, for binary predictor variables, a test of the interaction term β_{12} in the logistic regresssion model

  $$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  can be obtained by testing for correlation (association) between the genotypes at two separate loci, within the sample of cases
- Gains power from making assumption that genotypes (alleles) at the two loci are uncorrelated in the population
  - So only really suitable for unlinked or loosely linked loci (since closely linked loci are likely to be in LD)
- Alternatively contrast the genotype correlations in cases with those in controls (`--fast-epistasis` in PLINK)

PLINK's *fast-epistasis* statistics

- PLINK takes unphased genotype data

	Locus H		
Locus G	2	1	0
2	a	b	c
1	d	e	f
0	g	h	i

  and expands it to 2×2 allelic table

	Locus H	
Locus G	H_1	H_2
G_1	$A = 4a + 2b + 2d + e$	$B = 4c + 2b + 2f + e$
G_2	$C = 4g + 2h + 2d + e$	$D = 4i + 2h + 2f + e$

- PLINK estimates the log OR (λ) for association/correlation between the loci as as $\log(AD/BC)$ with estimated variance (v):

$$\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$

- A z-test of whether correlation exists (case-only) or is different in cases and controls (case/control) is:
  - Case-only: $T_{\text{FE-co}} = \frac{\lambda_A^2}{\hat{v}_A}$
  - Case/control: $T_{\text{FE-cc}} = \frac{[\lambda_A - \lambda_N]^2}{\hat{v}_A + \hat{v}_N}$

Testing correlation between loci

- A similar idea is implemented in EPIBLASTER (Kam-Thong et al. 2011 EJHG 19:465-571)
- Wu et al. (2010) (PLoS Genet 6:e1001131) also proposed a similar approach
- All these methods test whether correlation exists (case-only) or is different in cases and controls (case/control) via testing a log OR for association between two loci
  - However, the log OR for association (λ) encapsulates a slightly different quantity between the different methods

Testing correlation between loci

- Unfortunately, both Wu et al. (2010) and PLINK calculate the variance of their log ORs incorrectly
  - Resulting in a severe inflation in type 1 error (false positive) rate for the Wu et al. method
  - PLINK's statistic remains approximately correct (and completely correct in PLINK 1.9)
- We demonstrated this problem, and used the results of Brown (1975) to calculate the correct variances
  - Resulting in adjusted versions of all four statistics
- We also proposed some alternative Joint Effects (JE) statistics that have some advantages over these previous methods
  - See Ueki and Cordell (2012) PLoS Genetics 8(4):e1002625
  - Implemented in CASSI
    - http://www.staff.ncl.ac.uk/richard.howey/cassi/

Screening for interactions

- So far we have considered how to test for interaction between two specific factors
- In GWAS we typically test for (marginal) association at between 500,000 and 1 million SNPs across the genome
- Simplest way to search for interactions is to perform an exhaustive search, considering all pairwise combinations
  - If testing G×E with 5 environmental variables (for example), we end up with 5×1 million $= 5\times10^6$ tests
  - If testing G×G, we end up with [1 million choose 2] $\approx 5\times10^{11}$ tests
    - Computationally possible, but time-consuming
    - And dramatically increases multiple testing burden
    - But may be outweighted by increased power (Marchini et al. (2005) Nat Genet 37:413-417)
- Also need to decide exactly which test to perform
  - Several 'methods' (programs) choose their test on the basis of convenience/speed, given their chosen search strategy

Exhaustive testing

- Several recent publications have focussed on trying to speed up exhaustive search procedure
- E.g. by making use of data compression techniques and parallelization
    - Steffens et al. (2010) Hum Hered 69:268-284
- Or by using Graphical Processing Units (GPUs)
    - Sinnott-Armstrong et al. (2009) BMC Res Notes 2:149
    - Greene et al. (2010) Bioinformatics 26:694-695
    - Hu et al. (2010) Cell Res 20:854-857
    - Hemani et al. (2011) Bioinformatics 27:1462-1465
    - Kam-Thong et al. (2012) Hum Hered 73:220-236

Exhaustive testing (cont.)

- Or by computing faster tests e.g. log linear models rather than logistic regression
  - INTERSNP (Herold et al. (2009) Bioinformatics 25:3275-3281)
  - BOOST (Wan et al. (2010) AJHG 87:325-340)
  - BiForce Toolbox (Gyenesei et al. (2012) PMID:22689639)
- Or by performing an ‘approximately’ complete search
  - SIXPAC (Prabhu and Pe’er (2012) Genome Res 22:2230-2240)

Problems with exhaustive testing

- Problem of interpretation/noise
  - Will the effect be strong enough to withstand the multiple testing problem/lower prior probability that any effect is real?
- Why stop at pairwise combinations (why not 3-way, 4-way etc.)?
  - Most methods do not scale up to all 3-way, 4-way etc. combinations
  - Even if they did, problem of interpretation/multiple testing would be even worse
- May need to use 'filtering' approach where only consider a subset of loci chosen based on loose single-locus significance or other (biological or statistical) considerations

Biological filtering

- Emily et al. (2009) reported four significant cases of epistasis between unlinked loci in the WTCCC (Crohn's, Bipolar, Hypertension and Rheumatoid Arthritis) data
    - When limiting search on the basis of experimental knowledge of biological networks
- Herold et al. (2009) used their INTERSNP program to identify two SNPs that predispose to male pattern baldness, lying in genes from a joint pathway
- Results require replication (as has become gold standard in GWAS)
    - Problematic for interactions, owing to larger sample size required to give sufficient power to detect interactions (in comparison to main-effects)
        - Gauderman (2002) Am J Epid 155:478-484
        - Zuk et al. (2012) PNAS 109:1193-1198

Biological filtering

- Test at the gene level rather than the SNP level?
- E.g. across genes (G x G)
    - Wang et al. (2009) Genet Epidemiol 33:6-15
    - He et al. (2011) EJHG 19:164-172
    - Li and Cui (2012) Annals of Applied Statistics 6:1134-2261
    - Rajapakse et al. (2012) Genet Epidemiol 36:622-630
    - Ma et al. (2013) PLoS Genet 9(2): e1003321
        - Compared 4 different tests that combine *P* values from pairwise (SNP x SNP) interaction tests: min *P*, extended Simes, truncated tail, truncated product
        - Showed that the truncated tests did best
        - Presented an application only considering gene pairs known to exhibit protein-protein interactions
- Or within genes
    - Dinu et al. (2012) PLOS ONE 7:e43035
    - Wei et al. (2013) PLOS ONE 8:e71203

Statistical filtering

- Only test for interactions between ‘significant’ loci from a single-locus scan
  - Strange et al. 2010 (Nat Genet 42:985-990) found interactions between *HLA-C* and *ERAP1* in psoriasis
  - Evans et al. 2011 (Nat Genet 43:761-767) found interactions between *HLA-B27* and *ERAP1* in ankylosing spondylitis
  - Castillejo-Lopez et al. (2012) (Ann Rheum Dis 71:136-142) found interactions between polymorphisms in *BANK1* and *BLK* in SLE
- Only test for interactions between top 20% (or similar) loci from a single-locus scan
  - Nothing found in WTCCC Crohn’s data (Cordell 2009)
- Test for interactions between ‘significant’ loci and all other loci

Statistical filtering

- Two-stage procedures
  - Test all pairwise combinations at screening stage
  - Follow up with independent test of all pairs passing some threshold
  - Reduces multiple testing problem at second stage by constructing tests that are independent of 1st stage
    - Murcray et al. (2009) Am J Epidemiol 169:219-226
    - Lewinger et al. (2013) Genet Epid 37:440-451
    - Jiao et al. (2013) Genet Epid 37:452-464
    - Fråanberg et al. (2015) PLOS Genetics 11(9):e1005502 "Discovering genetic interactions in large-scale association studies by stage-wise likelihood ratio tests"

Data mining approaches

- Clever computational algorithms for searching (fast) through plausible space of models
  - Including models that involve multi-way (not just pairwise) interactions
- Most methods cross-validation to avoid over-fitting
  - E.g. fit a model using 9/10 of data, use remaining 1/10 to assess performance, and repeat many times
  - Choose final model (set of predictors) that performs well
- Often use permutation approaches to assess final significance
  - And final model often re-fit by logistic regression to provide parameter estimates
- Replication in an independent data set is crucial

MDR

- Multifactor Dimensionality Reduction (Ritchie et al. (2001) AJHG 69:138-147)
  - Divide data into 10 equal parts
  - Fit model using $\frac{9}{10}$ of data, use remaining $\frac{1}{10}$ to assess performance, repeat for each $\frac{9}{10}/\frac{1}{10}$ partition
  - Pick best-fitting model from all considered partitions
- Within each partition, perform exhaustive search over all single-locus models, 2-locus models, 3-locus models...
  - Computationally prohibitive for large numbers of loci
  - Use in conjunction with initial filtering method e.g. TuRF
- Has been used to detect potential interacting loci in breast cancer, type 2 diabetes, rheumatoid arthritis and coronary artery disease
  - Require replication in an independent data set

MDR

- Model construction based on classifying genotype combinations at the n loci as 'high' or 'low' risk
  - Based on the number of cases and controls in each cell
- Equivalent to fitting saturated genotype model

	Locus H		
Locus G	2	1	0
2	$\beta_0+\beta_{G_2}+\beta_{H_2}+\beta_{22}$	$\beta_0+\beta_{G_2}+\beta_{H_1}+\beta_{21}$	$\beta_0+\beta_{G_2}$
1	$\beta_0+\beta_{G_1}+\beta_{H_2}+\beta_{12}$	$\beta_0+\beta_{G_1}+\beta_{H_1}+\beta_{11}$	$\beta_0+\beta_{G_1}$
0	$\beta_0+\beta_{H_2}$	$\beta_0+\beta_{H_1}$	β_0

- But then reduce resulting 3^n dimensional model to 2-dimensional model (high/low risk)

Other model-search based approaches

- Random forests
    - Based on classification and regression trees (CART)
- Penalized regression methods
    - E.g. Zhu et al. (2014) Genet Epid 38:353-368
- Entropy-based methods
    - e.g. MECPM (maximum entropy conditional probability modelling)
        - Miller et al. (2009) Bioinformatics 25:2478-2485
        - Performed extremely well in comparison to other approaches in comprehensive simulation study by Chen et al. (2011) BMC Genomics 12:344
- Bayesian model selection
    - Involves specifying prior distributions for the number of loci and their effect sizes (=regression coefficients) including interactions
    - Use MCMC techniques to search through space of possible models, find model that maximizes likelihood
        - See also a recent Bayesian network approach (LEAP) which uses a heuristic search algorthm; outperformed MECPM in the study by Jiang and Neapolitan (2015) Genet Epid 39:173-184.

BEAM

- Zhang and Liu (2007) Nat Genet 39:1167-1173
- Loci divided into 3 groups:
  - Not associated with disease
  - Contribute via main effects only
  - Contribute via saturated interaction model
- Use MCMC to jump through space of possible models (divisions of loci)
- Generates posterior probabilities for each SNP of being in each group
  - Or test via B-statistic
- Method has been recently extended/improved (BEAM2, BEAM3)
  - Zhang et al. (2011) Ann Appl Stat 5:2052-2077
  - Zhang (2012) Genet Epidemiol 36:36-47

Empirical evidence for epistasis

- Hypothesis-based studies
  - Several papers by Combarros et al., most notably those part of "The Epistasis Project"
    - An attempt to replicate previous findings of epistasis in Alzheimer's disease, or discover new findings through restricting to candidate genes
    - Some success, but replication evidence quite weak: recommend cautious interpretation

Empirical evidence for epistasis

- Hypothesis-based studies
    - Several papers by Combarros et al., most notably those part of “The Epistasis Project”
        - An attempt to replicate previous findings of epistasis in Alzheimer’s disease, or discover new findings through restricting to candidate genes
        - Some success, but replication evidence quite weak: recommend cautious interpretation
    - Epistasis among *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1* in multiple sclerosis (Lincoln et al. 2009 PNAS 106:7542-7547)
    - Epistasis between *BANK1* and *BLK* in SLE (Castillejo-Lopez et al. 2012)
    - Epistasis between *HLA-C* and *ERAP1* in psoriasis (Strange et al. 2010)
    - Epistasis between *HLA-B27* and *ERAP1* in ankylosing spondylitis (Evans et al. 2011)

Empirical evidence for epistasis (cont.)

- Hypothesis-free studies
  - Exhaustive searches in WTCCC (Wan et al. 2010; Lippert et al. 2013) generally find lots of interactions within MHC for type 1 diabetes and rheumatoid arthritis
    - Could represent haplotype effects?
    - Require replication?
  - Prabhu and Pe'er (2012) used SIXPAC to identify a pair of interacting SNPs in Bipolar disorder
    - Regions replicated, though actual discovery SNPs did not
  - Gusareva et al. (2014) "Genome-wide association interaction analysis for Alzheimer's disease" found a reasonably convincing (partially replicating) interaction between SNPs on chromosome 6 (*KHDRBS2*) and 13 (*CRYL1*)

Empirical evidence for epistasis (cont.)

- Hypothesis-free studies
  - Hemani et al. 2014 (Nature 508:249-253) found 501 instances of epistatic effects on gene expression, of which 30 could be replicated in two independent samples
    - Many SNPs are close together, could represent haplotype effects?
    - Or the effect of a single untyped variant?
    - See Wood et al. (2014) Nature 514(7520):E3-5. PMID:25279928
  - Brown et al. 2014 (eLIFE 3:e01381) found 508 'candidate' SNPs showing potential interactions (G×G or G×E) on gene expression, based on their effect on trait variance
    - Twin studies suggested G×E played a role in 70% of these findings (but we don't know what the relevant environmental factors are)
    - 57 G×G interactions (between specific SNP and gene) replicated in a smaller data set

Conclusions

- Gene-gene and gene-environment interactions can be modelled in genetic (including genome-wide) association studies
- Computationally intensive if considering large numbers of loci: may need to filter down
- May be worth doing in some situations to increase power to detect effects (but further work needed on optimal search strategies)
  - Utility depends heavily on true underlying genetic model
  - Potentially useful for *detection* of interacting loci
  - Biological interpretation complex...and perhaps better addressed via alternative experiments