•Research corner•

Use of diplotypes – matched haplotype pairs from homologous chromosomes – in gene-disease association studies

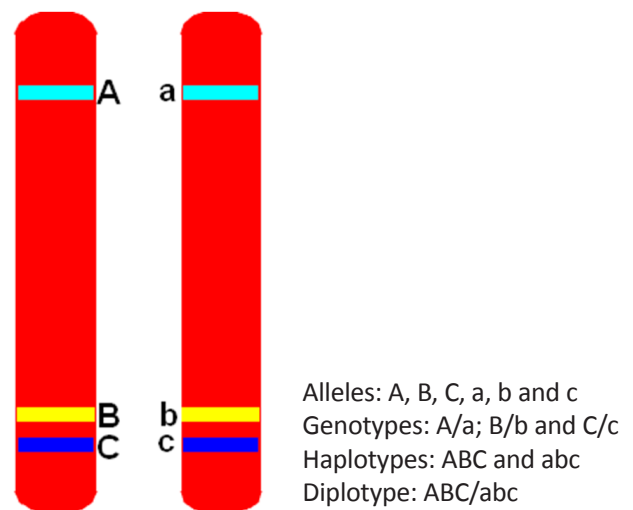Lingjun ZUO[1,2], Kesheng WANG[3], Xingguang LUO[1,2]*

Summary: Alleles, genotypes and haplotypes (combinations of alleles) have been widely used in gene-disease association studies. More recently, association studies using diplotypes (haplotype pairs on homologous chromosomes) have become increasingly common. This article reviews the rationale of the four types of association analyses and discusses the situations in which diplotype-based analyses are more powerful than the other types of association analyses. Haplotype-based association analyses are more powerful than allele-based association analyses, and diplotype-based association analyses are more powerful than genotype-based analyses. In circumstances where there are no interaction effects between markers and where the criteria for Hardy-Weinberg Equilibrium (HWE) are met, the larger sample size and smaller degrees of freedom of allele-based and haplotype-based association analyses make them more powerful than genotype-based and diplotype-based association analyses, respectively. However, under certain circumstances diplotype-based analyses are more powerful than haplotype-based analysis.

Key words: diplotype, haplotype, association analysis, genotypes, interaction effects, Hardy-Weinberg equilibrium

1. Introduction: definition and composition of diplotypes

Humans are diploid organisms; they have paired homologous chromosomes in their somatic cells, which contain two copies of each gene. An allele is one member of a pair of genes occupying a specific spot on a chromosome (called locus). Two alleles at the same locus on homologous chromosomes make up the individual's genotype. A haplotype (a contraction of the term 'haploid genotype') is a combination of alleles at multiple loci that are transmitted together on the same chromosome. Haplotype may refer to as few as two loci or to an entire chromosome depending on the number of recombination events that have occurred between a given set of loci. Genewise haplotypes are established with markers within a gene; familywise haplotypes are established with markers within members of a gene family; and regionwise haplotypes are established within different genes in a region at the same chromosome. Finally, a diplotype is a matched pair of haplotypes on homologous chromosomes.[1] (see Figure 1).

Figure 1. Model of alleles, genotypes, haplotypes and diplotypes on a pair of chromosomes

Alleles: A, B, C, a, b and c
Genotypes: A/a; B/b and C/c
Haplotypes: ABC and abc
Diplotype: ABC/abc

Traditionally, the expectation-maximum (EM) algorithm has been used to estimate haplotype frequencies.[2,3] This algorithm assumes Hardy-Weinberg Equilibrium (HWE).[4] However, if the genotype frequency distributions of individual markers are not in HWE, the assumption of the EM algorithm will be violated. The magnitude of the error of the EM estimates is greater when the HWE violation (the so-called Hardy-Weinberg Disequilibrium [HWD]) is attributable to a greater expected heterozygote frequency than the observed heterozygote frequency.[4]

Several programs can be used to construct both haplotypes and diplotypes. The HelixTree program[5] is based on the EM algorithm. New-generation programs such as the PHASE program are based on the Bayesian approach and the Partition Ligation algorithm; their proponents claim that they are more accurate in constructing haplotypes than the traditional programs based on the EM algorithm.[6-8] Both HelixTree and PHASE can estimate the diplotype frequency distributions among a population and estimate the diplotype probabilities for each individual. The probabilities of unambiguously observed diplotypes for each individual estimated by these programs should be 1.0; the probabilities of inferred diplotypes for each subject will be between 0.0 and 1.0.

2. Diplotype-based association analysis: application and interpretation

Haplotype-based and diplotype-based association analyses are more powerful than allele-based and genotype-based analyses.[9-11] Under certain circumstances (reviewed below), diplotype-based analysis is more powerful than haplotype-based analysis. Under these specific circumstances, diplotype-based association analysis is the most powerful of the four types of association analyses, a finding that has been confirmed in about 200 studies since 2002.[12,13] For example, Lee and colleagues[14] found that the 111 haplotype of the *Calpain-10* gene was associated with an increased risk of polycystic ovary syndrome (PCOS) (OR=2.4; 95% CI 1.8–3.3), the 112 haplotype was associated with a decreased risk of PCOS (OR=0.6; 95% CI 0.4–0.8), and the 121 haplotype was not associated with PCOS; however, the 111/121 diplotype was more strongly associated with increased susceptibility to PCOS than any of the haplotypes (OR=3.4; 95% CI 2.2–5.2). Luo and colleagues[15-22] reported that the diplotypes at *ADH1A, 1B, 1C, 4* and *7, CHRM2, OPRM1, OPRD1* and *OPRK1* were much more strongly associated with alcohol dependence, drug dependence and personality factors than the alleles, genotypes and haplotypes at these sites. And Li and colleagues[23] found that specific growth traits were significantly associated with the diplotypes of four individual SNPs at *IGF-II* but not with the haplotypes of these SNPs. Similar findings have been reported in other studies.[24,25]

There are several possible interpretations of these findings:

2.1 Haplotypes and diplotypes contain more information than alleles and genotypes

As shown in Figure 1, a haplotype is a combination of alleles from multiple loci on a single chromosome, a genotype is composed of two alleles on homologous chromosomes, and a diplotype is composed of two haplotypes (i.e., multiple genotypes) on homologous chromosomes. Theoretically, the information contained in a multi-locus haplotype is greater than that in a single-locus allele and the information contained in a multi-locus diplotype is greater than that contained in a single-locus genotype. Similarly, haplotypes with more alleles contain more information than those with less alleles and diplotypes with more genotypes contain more information than those with less genotypes.

A multi-locus haplotype is a specific variant of all possible combinations of single-locus alleles on the chromosome; both alleles and haplotypes reflect the features of chromosomes in the population. A diplotype is a specific variant of all possible combinations of single-locus genotypes on the paired chromosomes; both genotypes and diplotypes represent the types of chromosome pairs in each individual (see Table 1). A diplotype can also be conceptualized as a specific variant of all possible combinations of haplotypes from the two participating chromosomes. So haplotype-based analyses are equivalent to a stratified analysis of all alleles (at all loci), and diplotype-based analyses are equivalent to both stratified analysis of all genotypes at all loci, and to stratified analysis of all haplotypes. Thus, when the sample size is sufficiently large, haplotype- and diplotype-based analyses should be more powerful than allele-based and genotype-based analyses. Similarly, the analysis of an individual diplotype should be more informative than analysis of the corresponding individual haplotype.

Two alleles at one biallelic marker can divide the chromosomes in a population into two categories; these two alleles would result in three genotypes at the specified marker on homologous chromosomes and, thus, could be used to divide the individuals in a population into three categories. Assuming n independent biallelic markers, up to 2^n haplotypes constructed by these n markers can divide the chromosomes in a population into 2^n categories. At the same time, n independent biallelic markers would result in up to $2^n(2^n+1)/2$ diplotypes on the paired chromosomes, dividing the individuals in a population into $2^n(2^n+1)/2$ categories. (Note: each of these $2^n(2^n+1)/2$ diplotype categories is a subset of one of the 2^n haplotype categories.) When the sample size is large enough, dividing a sample into more categories increases the ability to identify meaningful variance between different subgroups in the sample, so haplotype-based and diplotype-based analyses are more powerful than allele-based and genotype-based analyses and an individual's diplotype is more informative than an individual's haplotype. However, the overall diplotype-based analysis may not be more powerful than the corresponding haplotype-based

Table 1. Comparison of haplotype-based and diplotype-based association analyses

	Haplotype-based association analysis	Diplotype-based association analysis
Composition	A haplotype is a subset of all alleles on specific chromosomes in the population.	A diplotype is a subset of all genotypes on homologous chromosome pairs in the population. A specific diplotype is one variant of all possible combinations of the haplotypes that exist in the population.
Feature	Both alleles and haplotypes reflect the components of chromosomes in individuals and in the population.	Both genotypes and diplotypes reflect the components of chromosome pairs in individuals and in the population.
n independent single-nucleotide polymorphisms (SNPs)	At most 2^n haplotypes	At most $2^n(2^n+1)/2$ diplotypes
Degrees of freedom in analysis	2^n-1	$[2^n(2^n+1)/2]-1$
Markers not in Hardy-Weinberg Equilibrium (HWE)	Less powerful predictor of disease status	More powerful predictor of disease status
Recessive genetic model	Less powerful predictor of disease status	More powerful predictor of disease status
With interaction	Less powerful predictor of disease status	More powerful predictor of disease status
Without interaction	More powerful predictor of disease status	Less powerful predictor of disease status
Sample size (n individuals)	2n	n
Frequency of rare categories	Less common	More common (decrease power)

analysis because in some situations the much greater degrees of freedom in a diplotype-based analysis than in the corresponding haplotype-based analysis weakens the strength of the identified associations.

The multi-locus haplotype and diplotype are composed of multiple markers that are in linkage disequilibrium (LD). They contain information from all of these individual markers and from several unknown flanking markers on the same chromosome. They are, therefore, usually more informative and closer to representing a 'whole gene' than single-marker alleles and genotypes. This is particularly the case when several of the known and unknown markers are etiologically related to the disease(s) of interest.[9-11]

2.2 Genotype-based and diplotype-based analyses remain valid in the presence of Hardy-Weinberg Disequilibrium

When the genotype frequency distributions of some markers are not in Hardy-Weinberg Equilibrium the allele-based and haplotype-based analyses become less powerful and may be invalid, but the genotype-based and diplotype-based analyses are still valid. When there is Hardy-Weinberg Disequilibrium the marker alleles and haplotypes are not independent of each other so the effects of disease predisposing alleles and haplotypes may be 'masked' by other non-

disease predisposing alleles and haplotypes[25] or, in the case of a recessive condition, by the presence of a dominant allele on the homologous chromosome. This weakens or invalidates the strength of the association between the allele or haplotype and the disease(s) of interest. However, genotype-based and diplotype-based association analyses remain valid even in the presence of strong Hardy-Weinberg disequilibrium. This has been demonstrated in several studies.[15-18,27-30]

2.3 Haplotype and diplotype analyses incorporate interaction effects and, thus, are more informative when interaction between assessed markers is present

The haplotypes or diplotypes incorporate information on linkage disequilibrium among markers; so information on the multivariate interaction effects between markers are incorporated into haplotype-based and diplotype-based analyses.[31] In most cases[18,20-22] reported interaction effects between alleles and between genotypes are similar to those seen with corresponding multi-locus haplotype-based and diplotype-based analyses; this supports the contention that diplotype-based analyses incorporate information on the interactions between different markers and between different haplotypes. The interaction effect is often a more powerful predictor of disease status than

the main effect,[32] especially when the main effects are marginal,[33] so when interaction effects occur diplotype-based association analyses would likely be more informative than association analyses based on haplotypes, genotypes or alleles.

2.4 Using quantitative measures instead of categorical measures makes diplotype-based analysis more powerful

Programs implementing the Bayesian approach can estimate the probabilities of all possible pairs of haplotypes (i.e., a 'full model' in which the probabilities of all diplotype categories are assessed) or the probabilities of the most relevant subset of diplotype categories (i.e., a "reduced" model) for each individual. The estimated diplotype probabilities are quantitative measures so they usually preserve more information than the original categorical list of the different diplotype categories. Thus the analyses are more powerful if they employ diplotype probabilities instead of diplotype categories.[17]

2.5 Avoiding multiple testing preserves the power of haplotype-based and diplotype-based analyses

When testing the association between single markers and a phenotype, multiple independent tests are required so the analysis needs to be adjusted for multiple testing, which reduces the power of the analysis to identify significant differences between groups. But there is no need to adjust for multiple testing when incorporating multiple markers into haplotype-based or diplotype-based analyses, preserving the power of the analysis.[34] This is another reason that haplotype-based and diplotype-based association analyses are more powerful than single-locus analyses.

3. Discussion: conclusion and future aspects

This review shows that haplotype-based association analyses are more powerful than allele-based association analyses and that diplotype-based association analyses are more powerful than genotype-based analyses. Moreover, under certain circumstances, diplotype-based analyses are more powerful than haplotype-based analysis. Thus, in circumstances where very large sample sizes are available, diplotype-based

association analysis is the most powerful of the four potential analytic strategies.

The sample sizes of association analyses based on alleles and haplotypes are twice those of the corresponding association analyses based on genotypes and diplotypes. And the degrees of freedom in allele-based and haplotype-based analyses are much less than the degrees of freedom of the corresponding genotype-based and diplotype-based analyses. Thus in circumstances where there are no interaction effects between markers and where the criteria for Hardy-Weinberg Equilibrium are met, allele-based association analyses are more powerful than genotype-based analyses and haplotype-based association analyses are more powerful than diplotype-based analyses.[9,33] However, in several other circumstances the diplotype-based analysis is more powerful than haplotype-based analyses: (a) when there are interaction effects between haplotypes, (b) when there is Hardy-Weinberg Disequilibrium, and (c) when considering a recessive model of inheritance.[33]

One disadvantage of diplotype-based analysis compared to haplotype-based analysis is that there are typically a greater number of rare diplotype categories (i.e., categories with few individuals) than the number of rare haplotype categories. For each category, no matter how small, an additional degree of freedom needs to be included in the analysis, so this results in a greater decrease in the power of diplotype-based association tests compared to haplotype-based association tests. Strategies to deal with rare observations include excluding such categories or merging them with other categories.[29,33]

双体型 - 同源染色体中配对的单体型对 – 在基因 - 疾病中的关联分析中的应用

左玲郡，王克胜，罗星光

概述：等位基因，基因型和单体型（等位基因组合）已被广泛应用于基因 - 疾病的关联研究。最近，使用双体型（同源染色体单体型对）的关联研究已经越来越普遍。本文综述了四种关联分析类型的基本原理，并探讨了为什么以双体型为基础的关联分析比其他类型的关联分析更高效。单体型关联分析比基于等位基因的关联分析更高效，以双体型为基础的关联分析比基于基因型的关联分析更高效。在标记之间没有交互作用并且符合 Hardy-Weinberg 平衡（HWE）标准的情况下，以等位基因和单体型为基础的关联分析样本量较大、自由度较小，使它们分别比基因型和双体型为基础的关联分析更高效。然而，在某些情况下以双体型为基础的关联分析比单体型关联分析更高效。

关键词：双体型，单体型，关联分析，基因型，交互作用，Hardy-Weinberg 平衡

References

1. Lu Q, Cui Y, Wu R. A multilocus likelihood approach to joint modeling of linkage, parental diplotype and gene order in a full-sib family. *BMC Genet.* 2004; **5**(1): 20. doi: http://dx.doi.org/10.1186/1471-2156-5-20

2. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol.* 1995; **12**(5): 921-927. doi: http://dx.doi.org/10.1002/gepi.10323

3. Fang M. A fast expectation-maximum algorithm for fine-scale QTL mapping. *Theor Appl Genet.* 2012; **125**(8): 1727-1734. doi: http://dx.doi.org/10.1007/s00122-012-1949-9

4. Fallin D, Schork NJ. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet*. 2000; **67**(4): 947-959. doi: http://dx.doi.org/10.1086/303069

5. Available at: http://www.goldenhelix.com/News/pressrelease20050914_affymetrix.html

6. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*. 2003; **73**(5): 1162-1169. doi: http://dx.doi.org/10.1086/379378

7. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*. 2001; **68**(4): 978-989. doi: http://dx.doi.org/10.1086/319501

8. Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet*. 2002; **70**(1): 157-169. doi: http://dx.doi.org/10.1086/338446

9. Akey J, Jin L, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet*. 2001; **9**(4): 291-300. doi: http://dx.doi.org/10.1038/sj.ejhg.5200619

10. Mao WG, He HQ, Xu Y, Chen PY, Zhou JY. Powerful haplotype-based Hardy-Weinberg equilibrium tests for tightly linked loci. *PLoS One*. 2013; **8**(10): e77399. doi: http://dx.doi.org/10.1371/journal.pone.0077399

11. Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered*. 2003; **56**(1-3): 18-31. doi: http://dx.doi.org/10.1159/000073729

12. Yang CM, Chen HC, Hou YY, Lee MC, Liou HH, Huang SJ, et al. A high IL-4 production diplotype is associated with an increased risk but better prognosis of oral and pharyngeal carcinomas. *Arch Oral Biol*. 2014; **59**(1): 35-46. doi: http://dx.doi.org/10.1016/j.archoralbio.2013.09.010

13. Cusinato DA, Lacchini R, Romao EA, Moysés-Neto M, Coelho EB. Relationship of Cyp3a5 genotype and Abcb1 diplotype to Tacrolimus disposition in Brazilian kidney transplant patients. *Br J Clin Pharmacol*. 2014; doi: http://dx.doi.org/10.1111/bcp.12345 (in press)

14. Lee JY, Lee WJ, Hur SE, Lee CM, Sung YA, Chung HW. 111/121 diplotype of Calpain-10 is associated with the risk of polycystic ovary syndrome in Korean women. *Fertil Steril*. 2009; **92**(2): 830-833. doi: http://dx.doi.org/10.1016/j.fertnstert.2008.06.023

15. Luo X, Kranzler HR, Zuo L, Lappalainen J, Yang BZ, Gelernter J. ADH4 gene variation is associated with alcohol dependence and drug dependence in European Americans: results from HWD tests and case-control association studies. *Neuropsychopharmacology*. 2006; **31**(5): 1085-1095. doi: http://dx.doi.org/10.1038/sj.npp.1300925

16. Luo X, Kranzler HR, Zuo L, Lappalainen J, Yang BZ, Gelernter J. CHRM2 gene predisposes to alcohol dependence, drug dependence and affective disorders: results from an extended case-control structured association study. *Hum Mol Genet*. 2005; **14**(16): 2421-2434. doi: http://dx.doi.org/10.1038/sj.npp.1300925

17. Luo X, Kranzler HR, Zuo L, Wang S, Schork NJ, Gelernter J. Diplotype trend regression analysis of the ADH gene cluster and the ALDH2 gene: multiple significant associations with alcohol dependence. *Am J Hum Genet*. 2006; **78**(6): 973-987. doi: http://dx.doi.org/10.1086/504113

18. Luo X, Kranzler HR, Zuo L, Wang S, Schork NJ, Gelernter J. Multiple ADH genes modulate risk for drug dependence in both African- and European-Americans. *Hum Mol Genet*. 2007; **16**(4): 380-390. doi: http://dx.doi.org/10.1093/hmg/ddl460

19. Luo X, Kranzler HR, Zuo L, Zhang H, Wang S, Gelernter J. CHRM2 variation predisposes to personality traits of agreeableness and conscientiousness. *Hum Mol Genet*. 2007; **16**(13): 1557-1568. doi: http://dx.doi.org/10.1093/hmg/ddm104

20. Luo X, Kranzler HR, Zuo L, Zhang H, Wang S, Gelernter J. ADH7 variation modulates extraversion and conscientiousness in substance-dependent subjects. *Am J Med Genet B Neuropsychiatr Genet*. 2008; **147B**(2): 179-186. doi: http://dx.doi.org/10.1002/ajmg.b.30589

21. Luo X, Zuo L, Kranzler H, Zhang H, Wang S, Gelernter J. Multiple OPR genes influence personality traits in substance dependent and healthy subjects in two American populations. *Am J Med Genet B Neuropsychiatr Genet*. 2008; **147B**(7): 1028-1039. doi: http://dx.doi.org/10.1002/ajmg.b.30701

22. Zuo L, Gelernter J, Kranzler HR, Stein MB, Zhang H, Wei F, et al. ADH1A variation predisposes to personality traits and substance dependence. *Am J Med Genet B Neuropsychiatr Genet*. 2009; **153B**(2): 376-386. doi: http://dx.doi.org/10.1002/ajmg.b.30990

23. Li X, Bai J, Hu Y, Ye X, Li S, Yu L. Genotypes, haplotypes and diplotypes of IGF-II SNPs and their association with growth traits in largemouth bass (Micropterus salmoides). *Mol Biol Rep.* 2012; **39**(4): 4359-4365. doi: http://dx.doi.org/10.1007/s11033-011-1223-2

24. Tou J, Wang L, Liu L, Wang Y, Zhong R, Duan S, et al. Genetic variants in RET and risk of Hirschsprung's disease in Southeastern Chinese: a haplotype-based analysis. *BMC Med Genet*. 2011; **12**: 32. doi: http://dx.doi.org/10.1186/1471-2350-12-32

25. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*. 2002; **11**(20): 2463-2468. doi: http://dx.doi.org/10.1093/hmg/11.20.2463

26. Chen Y, Li X, Li J. A novel approach for haplotype-based association analysis using family data. *BMC Bioinformatics*. 2010; 11 Suppl 1: S45. doi: http://dx.doi.org/10.1186/1471-2105-11-S1-S45

27. Nielsen DM, Ehm MG, Weir BS. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am J Hum Genet*. 1998; **63**(5): 1531-1540. doi: http://dx.doi.org/10.1086/302114

28. Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics*. 1997; **53**(4): 1253-1261. doi: http://dx.doi.org/10.2307/2533494

29. Jannot AS, Essioux L, Clerget-Darpoux F. Association in multifactorial traits: how to deal with rare observations? *Hum Hered*. 2004; **58**(2): 73-81. doi: http://dx.doi.org/10.1159/000083028

30. Lin WY, Schaid DJ. Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. *Genet Epidemiol*. 2009; **33**(3): 183-197. doi: http://dx.doi.org/10.1002/gepi.20364

31. Hu Y, Jason S, Wang Q, Pan Y, Zhang X, Zhao H, et al. Regression-based approach for testing the association between multi-region haplotype configuration and complex trait. *BMC Genet*. 2009; **10**(1): 56. doi: http://dx.doi.org/10.1186/1471-2156-10-56

32. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*. 2005; **37**(4): 413-417. doi: http://dx.doi.org/10.1038/ng1537

33. Sha Q, Dong J, Jiang R, Zhang S. Tests of association between quantitative traits and haplotypes in a reduced-dimensional space. *Ann Hum Genet*. 2005; **69**(Pt 6): 715-732. doi: http://dx.doi.org/10.1111/j.1529-8817.2005.00216.x

34. Bardel C, Danjean V, Hugot JP, Darlu P, Génin E. On the use of haplotype phylogeny to detect disease susceptibility loci. *BMC Genet*. 2005; **6**(1): 24. doi: http://dx.doi.org/10.1186/1471-2156-6-24

Dr. Zuo graduated from Shanghai Medical University in 1991 and obtained her PhD from Fudan University School of Medicine in 2001. She is currently an assistant professor and the Director of the Psychiatric Genetics Lab (Zuo) at the Department of Psychiatry, Yale University School of Medicine. Her research interests are the genetics and epigenetics of psychiatric disorders and related behaviors.