Association Analysis of Genotype Patterns With Digenic Traits

Advanced Gene Mapping Course, January 2022

Jurg Ott, Ph.D., Professor Emeritus

Rockefeller University, New York

http://lab.rockefeller.edu/ott/

ott@rockefeller.edu

PH +1 646 321 1013

Jurg Ott, PhD

Professor Emeritus and Director
Laboratory of Statistical Genetics
Rockefeller University
New York, NY 10065

https://lab.rockefeller.edu/ott/
EM: ott@rockefeller.edu
PH: +1 646 321 1013

Research Interests

Development of analysis methods for genetic data, genetic linkage and association analysis

Implementation in computer programs, dissemination on website

Collaboration with researchers world-wide on their data

Recent publications: [1-5] (#1 now freely available from https://www.jurgott.org/linkage/LinkageHandbook.pdf)

1. Terwilliger DJ, Ott J. Handbook of human genetic linkage. Johns Hopkins University Press. 1994.

2. Imai-Okazaki A, Li Y, Horpaopan S, Riazalhosseini Y, Garshasbi M, Mosse YP, et al. Heterozygosity mapping for human dominant trait variants. Hum Mutat. 2019 Apr 24;40(7):996-1004.

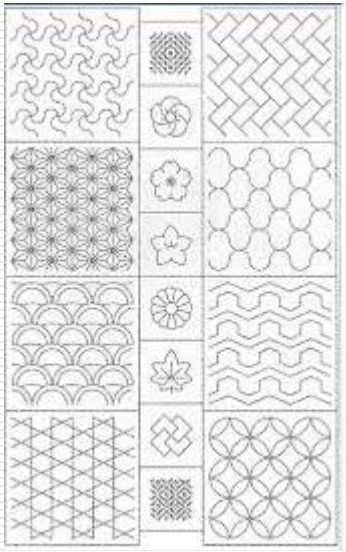
3. Horpaopan S, Fann CSJ, Lathrop M, Ott J. Shared genomic segment analysis with equivalence testing. Genet Epidemiol. 2020 Oct;44(7):741-47.

4. Okazaki A, Yamazaki S, Inoue I, Ott J. Population genetics: past, present, and future. Human genetics. 2020:1-10.

5. Okazaki A, Horpaopan S, Zhang Q, Randesi M, Ott J. Genotype pattern mining for pairs of interacting variants underlying digenic traits. Genes. 2021;12(8):1160.

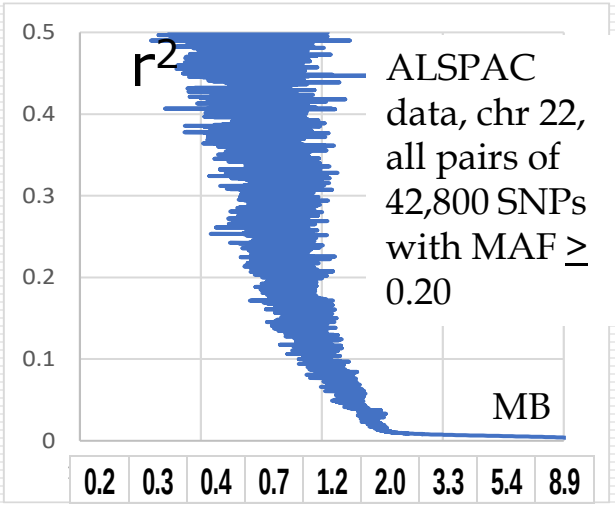
Topics

- Science develops independently in different fields:
  - Human gene mapping
  - Frequent Pattern Mining
- Case-control association analysis
  - Main effects in genetic association studies
  - Interaction effects in case-control data
- Mining consumer databases
  - The *Apriori* algorithm
  - Newer algorithms: *eclat*, *fpgrowth*
  - Analysis of AMD dataset



Main association effects

- Consider two DNA variants with minor alleles *A* and *T*. Even when on the same chromosome, the frequency of *A*-*T* chromosomes (haplotypes) is the product of allele frequencies, P(*A*-*T*) = P(*A*) P(*T*) → linkage equilibrium. Variants very close together: P(*A*-*T*) ≠ P(*A*) P(*T*) → LD, linkage disequilibrium.
- Disease variant vs marker variant: Different genotype frequencies in cases and controls → genetic association.
- **Recessive traits**: Variants close to disease tend to be homozygous (homozygosity mapping; Lander & Botstein, *Science* 1987;**236**:1567-70).
- **Dominant traits**: Variants close to disease tend to be heterozygous (Imai-Okazaki et al, *Hum Mutat* 2019;**40**:996-1004):
  P(het) > 1 – *f*, P(het, popul.) = 2*f*(1 – *f*), *f* = MAF



	AA	AG	GG
affected	0.18	0.70	0.12
unaffected	0.04	0.32	0.64
OR	~5	~5	~0.1

Effects of genotypes

- Case and control samples, test for different genotype frequencies with chi-square tests. Alternative: logistic regression. Let Y = 1 and 0 for a case and a control individual, respectively. Let x = 0, 1, 2 = respective genotypes AA, AT, TT. Define π(x) = P(Y = 1 | x). Then, the log odds, or logit, is expressed as

$$\text{logit}[\pi(x)] = \log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x.$$

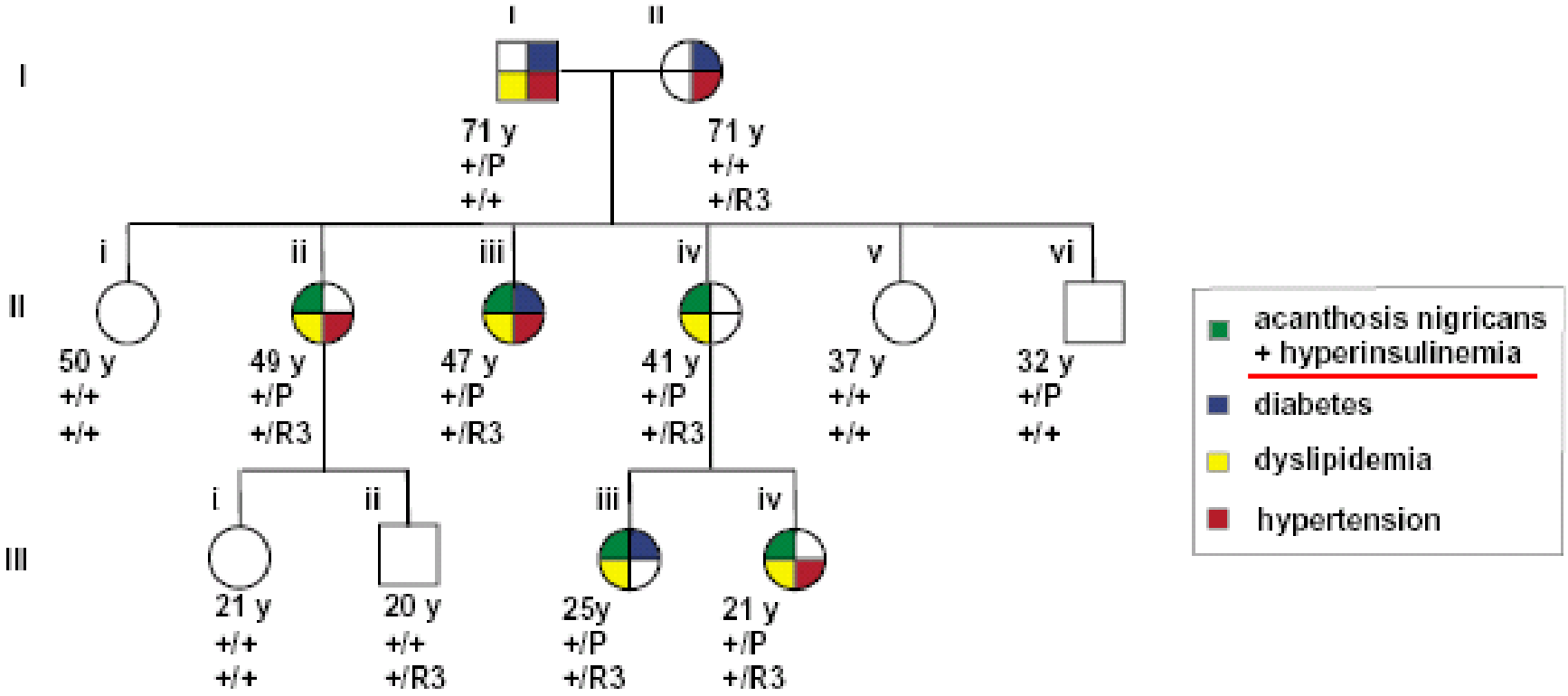
Agresti (2002) *Categorical Data Analysis*. Wiley

- For disease associations of multiple variants, $\text{logit} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots$

- In addition to these main effects of individual predictor variants, one may specify **interaction effects** by adding products like $\beta_{12}\, x_1\, x_2$. (Cordell, *Nat Rev Genet* 2009;**10**:392-404).
- Nowadays, we generally work with genotypes, not alleles.
- Multiple variants: Work with sequences of genotypes (diplotypes) rather than sequences of alleles (haplotypes) except for HLA genes.

Digenic Inheritance of Severe Insulin Resistance

Savage et al. (2002) *Nat Genet* **31**, 379



"... all five family members with severe insulin resistance, and no other family members, were compound heterozygous with respect to two frameshift mutations of these two unlinked genes."

Multiple Hits ... Digenic Diseases

Ming & Muenke (2002) *Am J Hum Genet* **71**, 1017 (review)
Schaffer A (2013) *J Med Genet* 50, 641-52 (review)

Effect and Phenotype	Gene 1		Gene 2	
	Mutation	Phenotype	Mutation	Phenotype
Synergistic:				
RP	*ROM1*$^{+/G80insG}$	Normal	*RDS*$^{+/L185P}$	Normal
RP	*ROM1*$^{+/L114insG}$	Normal	*RDS*$^{+/L185P}$	Normal
Bardet-Biedl	*BBS2*$^{Y24X/Q59X}$	Normal	*BBS6*$^{+/Q147X}$	Normal
Deafness	*GJB2*$^{+/35delG}$	Normal	*GJB6*$^{+/-}$	Normal
Deafness	*GJB2*$^{+/167delT}$	Normal	*GJB6*$^{+/-}$	Normal
Hirschsprung	*RET*$^{+/I6471I}$	Normal	*EDNRB*$^{+/S305N}$	Normal
Severe insulin resistance	*PPARG*$^{+/A553delAAAiT}$	Normal	*PPP1R3A*$^{+/C1984delAG}$	Normal
Modifier:				
Juvenile-onset glaucoma	*MYOC*$^{+/G399V}$	Adult-onset glaucoma	*CYP1B1*$^{+/R368H}$	Normal
Usher 1	*USH3*$^{mut/mut}$	Usher 3	*MYO7A*$^{+/delG\ (exon\ 25)}$	Normal
Congenital nonlethal JEB	*COL17A1*$^{R1226X/L855X}$	Juvenile JEB	*LAMB3*$^{+/R635X}$	Normal
More severe ADPKD	*PKD1*$^{+/mut}$	Less severe ADPKD	*PKD2*$^{+/2152delA}$	Less severe ADPKD
More severe hearing loss	*DFNA1*	Mild hearing loss	*DFNA2*	Mild hearing loss
WS2/OA	*MITF*$^{+/944delA}$	?WS2	*TYR*$^{+/R402Q}$	Normal
More severe WS2/OA	*MITF*$^{+/944delA}$	?WS2	*TYR*$^{R402Q/R402Q}$	Normal

How to analyze interaction effects?

- Hyperlipidemia data: 5 relevant genes, ~200 variants in each gene, look for interactions in each of the 10 pairs of genes. Work with LR chi-square!

CASES	Variant 1		
Var 2	GG	GT	TT
AA	...	...	...
AC	...	...	...
CC	...	...	...

CONTROLS	Variant 1		
Var 2	GG	GT	TT
AA	...	...	...
AC	...	...	...
CC	...	...	...

Data	chi-sq	df
cases	3.3591	4
controls	3.6658	4
both	1.4255	4
heterogeneity	5.5994	4

- $\chi^2_{\text{Heterogeneity}} = \chi^2_{\text{Cases}} + \chi^2_{\text{Controls}} - \chi^2_{\text{both}}$

Var 1 ->	GG			GT			TT		
Var 2 ->	AA	AC	CC	AA	AC	CC	AA	AC	CC
cases	...	...	...	...	...	...	...	...	...
controls	...	...	...	...	...	...	...	...	...

Source	chi-sq	df
Var 1 main	0.4196	2
Var 2 main	48.1979	2
Interaction	5.5994	4
Total table	54.2169	8

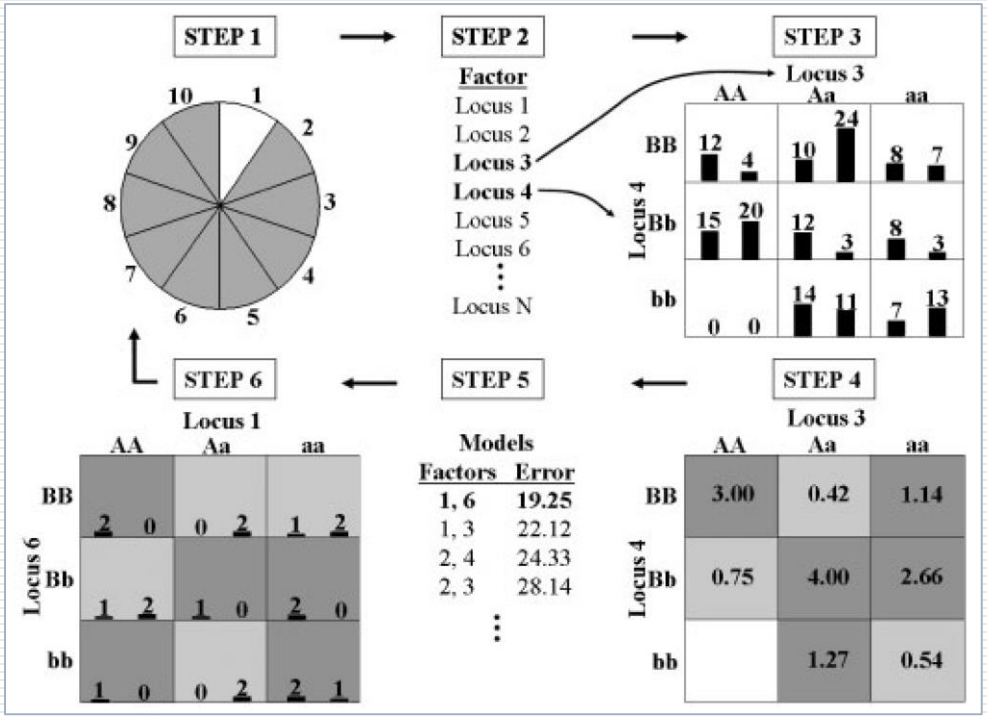
- $\chi^2_{\text{Interaction}} = \chi^2_{\text{Total}} - \chi^2_{\text{Var1}} - \chi^2_{\text{Var2}}$

Finding disease-associated pairs of genotypes

1. Multifactor Dimensionality Reduction (MDR)
   Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions ... Genet Epidemiol 2003;24:150–157
2. Exhaustive evaluation of all pairs of genotypes at all pairs of variants
3. Applying off-the-shelf pattern search algorithms
   Chee C-H, Jaafar J, Aziz IA, Hasan MH, Yeoh W. Algorithms for frequent itemset mining: a literature review. Artificial Intelligence Review. 2019;52(4):2603-21
4. Construction of Bayesian network
   Guo Y, Zhong Z, et al. Epi-GTBN: An approach of epistasis mining based on genetic Tabu algorithm and Bayesian network. BMC Bioinform 2019;20:444
5. Sophisticated computational approaches
   Titarenko SS, Titarenko VN, Aivaliotis G, Palczewski J. Fast implementation of pattern mining algorithms ... . Journal of Big Data. 2019; 13(6):37

1. MDR

- Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions ... Genet Epidemiol 2003;24:150–157
- Classify each of the 9 cells as high risk or low risk.
- Evaluate prediction error (case vs. control) by cross-validation.
- Find model that maximizes cross-validation consistency and minimizes prediction error.



2. Exhaustive search for interacting SNPs

- "Discovering Genetic Factors for psoriasis through exhaustively searching for significant second order SNP-SNP interactions"
- Kwan-Yeung Lee, Kwong-Sak Leung, Nelson L. S. Tang & Man-Hon Wong. *Sci Rep* 2018;**8**:15186

- Abstract: To deal with the enormous search space, our search algorithm is accelerated with eight **biological plausible interaction** patterns and a pre-computed look-up table. After our search, we have discovered several **SNPs having a stronger association to psoriasis when they are in combination with another SNP**...

3. Frequent Pattern Mining

- Thirty years ago, supermarkets started collecting huge amounts of consumer data at their cashiers. Consumer habits – if someone buys bread, how likely will they also buy milk and wine?
- **Apriori algorithm** (Agrawal et al, *ACM SIGMOD Conference on Management of Data* 1993; 207-216): Efficient search for frequent sets of items ("itemsets") purchased by one consumer ("transaction"). Development of **association rules**, that is, conditional probabilities P(Y|X), with Y and X being items or itemsets.
- Research published in conference proceedings, rarely in traditional journals.
- In the absence of strong main effects, we need to directly search for **genotype patterns** (at two or more variants) with different frequencies in cases and controls, without consulting main effects.
- Zhang Q, Long Q, Ott J. **AprioriGWAS**, a new pattern mining strategy for detecting genetic variants associated with disease through interaction effects. PLoS Comput Biol. 2014 Jun;10(6):e1003627
- Newer implementations of search algorithms, e.g. *fpgrowth* (https://borgelt.net/software.html). Huge memory demands: Using Linux desktop with 512 GB of memory.

4. Bayesian networks

- Guo Y et al. Epi-GTBN: An approach of epistasis mining based on ... Bayesian network. BMC Bioinform. 2019;20:444
- Like many other approaches, Epi-GTBN employs a Bayesian network, that is, a probabilistic model to represent actions and interactions among variants and phenotypes.
- Authors analyzed a well-known dataset on age-related macular degeneration (AMD), which has been investigated by various other researchers. For analysis by Epi-GTBN, to reduce the computational burden, only the 1,039 SNPs with **smallest p-values** ($p < 0.01$) out of the original 103,611 SNPs were retained.
- Results were comparable to those obtained elsewhere.
- Focusing on variants with strong main effects is fallacious! Frequencies of genotype patterns depend on main and interaction effects: Strong main effects are likely to lead to strong (significant) genotype patterns.

5. Sophisticated computational approaches

Titarenko SS et al. Fast implementation of pattern mining algorithms ... . Journal of Big Data. 2019; 13(6):37

- Common FPM implementations require huge memory resources (exception: AprioriGWAS) so that only a few thousand variants can be analyzed.
- Titarenko et al: Sophisticated approach, much more powerful than existing FPM methods.
- Dewan A et al. HTRA1 promoter polymorphism in wet age-related macular degeneration. Science. 2006;314(5801):989-92
- 81,930 variants, 96 cases, 127 controls, 43% cases.

- Single-variant trend test:
- 2 variants significant
- For FPM analysis:
- Disregard 4 variants

CHR	SNP	CHISQ	DF	pBon	EMP2
10	rs10490924	42.58	1	0.000006	0.00001
8	rs10504152	23.33	1	0.1057	0.0490
0	SNP_A-1706540	20.26	1	0.4259	0.2448
0	SNP_A-1702501	18.5	1	0.7514	0.5270
7	rs10499342	18.11	1	0.8187	0.6047
13	rs2011847	17.11	1	0.9447	0.8012
8	rs1377131	17.1	1	0.9450	0.8019

- https://www.jurgott.org/linkage/GPM.html

AMD data: Genotype pattern analysis

- Search for patterns (genotype pairs) with minimum support of 40. Perform 1000 random permutations for p-value estimation (corrected for multiple testing).
- Find m = 18,044,794 patterns.
- Two patterns are significant, p = 0.015, compared with best p = 0.60 in single-variant analysis.
- Expect many more significant genotype patterns than single-variant results.
- Different ways of establishing significance: Bonferroni correction and FDR depend on large number m of "null" results.
- Compare confidence of 90% with 43% of cases ("null" confidence) in data.

supp	conf	chisq	pPerm	OR	ch1	ch2
40	90.0	47.2604	0.015	18.5	3	4
40	90.0	47.2604	0.015	18.5	3	4
53	81.1	42.5124	0.098	9.5	1	5
56	78.6	39.481	0.315	8.1	6	7
56	78.6	39.481	0.315	8.1	6	7

alpha	Perm	Bonf	FDR-BY
0.001	0	2	2
0.01	0	9	3
0.02	2	11	11
0.03	2	16	13
0.04	2	18	61
0.05	2	19	11905

A Novel Mapping Strategy Utilizing Mouse Chromosome Substitution Strains Identifies Multiple Epistatic Interactions That Regulate Complex Traits

Anna K. Miller,* Anlu Chen,† Jacquelaine Bartlett,‡ Li Wang,* Scott M. Williams,*,‡,1 and David A. Buchner*,†,1,2

- *Genes, Genomes, Genetics* 2020;**10**:4553-44564
- "The SNPs in the epistatic QTL pairs that accounted for the largest variances were undetected in our single locus association analyses."
- For n = 57 (chrom. 4) + 50 (chrom. 6) = 107 variants, a two-locus genome scan was carried out for $n(n-1)/2$ = 5,671 variant pairs.
- Appropriate significance thresholds for bivariate lod scores obtained via permutation analysis.