



Published in final edited form as:

Epidemiology. 2011 November ; 22(6): 845–847. doi:10.1097/EDE.0b013e31822ffbe7.

How much are we missing in SNP-by-SNP analyses of GWAS?

Min Shi and Clarice R. Weinberg

Abstract

Genome-wide association studies have discovered common genetic variants associated with susceptibility for several complex diseases; but they have been unfruitful for many others. Typically analysis is done by “agnostically” considering one single nucleotide polymorphism (SNP) at a time, controlling the overall Type I error rate by correcting for multiple testing. In this short report we use oral clefting as a disease model to develop a range of toy example scenarios: risk might only involve genes, might involve both genes and exposure and might involve genes, exposure and their super-multiplicative interaction. These examples illustrate that important genetic variants can be obscured by using a one-SNP-at-a-time analysis when in fact multiple biological pathways and multiple genes jointly influence etiology. These examples highlight the need for better methods for gene-by-environment and gene-by-gene analyses.

The decoding of the human genome and the availability of affordable and accurate high-throughput genotyping have raised great expectations for genomic medicine. While genome-wide association studies (GWAS) have identified some important genetic variants associated with certain common diseases, GWAS results have not fully lived up to our high expectations.^{1,2} In addition, variants that have been identified appear to explain a small proportion of genetic heritability, even for diseases that clearly have a major genetic component.³ Various explanations have been offered. For some outcomes, the genetic contribution could be due to the action of diverse but rare susceptibility variants, which can result in apparent modest associations with any one common SNP.⁴ This short paper will provide examples to illustrate another sort of mechanism, where gene-by-gene interplay of common variants can result in very modest associations when the etiology requires particular genotypes at multiple loci. We will here focus on the birth defect, oral clefting, but similar examples can readily be constructed involving common outcomes, like Type II diabetes.

The prevalence at birth of oral clefting ranges between 1/600 to 1/1200, depending on ethnicity, geographic region and socio-economic status. The recurrence rate in siblings of affected babies is 40 times that in the general population and the concordance rate for monozygotic twins (MZ) is 25–40%, much higher than that in dizygotic (DZ) twins (3–6%),⁵ which is quite close to the risk in a later non-twin full sibling of an affected baby. Genetics clearly plays an important role in clefting. However, even with the enormous efforts devoted to studying this condition^{6–9} the majority of its etiology has remained elusive.

Rare variants, structural variations, epigenetic effects, gene-gene and gene-environment interactions have all been proposed as phenomena that may contribute to the low yield of GWAS.¹⁰ The methods of analysis employed can also impose limitations. Typically analysis is done, either for case-parent triad designs or for case-control studies, by “agnostically” considering one SNP at a time, controlling the overall Type I error rate by correcting for multiple testing. In this short report we develop toy examples to illustrate that one can overlook important genetic variants by using a one-SNP-at-a-time analysis when in fact multiple biological pathways, multiple genes and gene-environment jointly play an important role in etiology.

Replete with genes having similar roles, the human genome is well known for “genetic redundancy,” and disruption of a single gene may often be selectively neutral.¹¹ Consequently it may take the malfunction of several genes in a pathway to produce a particular phenotype. At the same time genetic heterogeneity is common for complex diseases and many pathways may be involved in a biological process: disruption of any one of these pathways may produce the same phenotype. Palate formation, for example, requires coordination of several signaling pathways: fibroblast growth factor, transforming growth factor beta, bone morphogenetic protein and sonic hedgehog pathways.¹² Dysfunction of any of these pathways could potentially lead to clefting.

Consider a toy example. Suppose two independent pathways can cause clefting and in each pathway simultaneous disruption of four unlinked genes is needed to trigger the event (Figure 1). In other words, a fetus carrying at least one risk allele at each of four particular loci in the pathway will develop oral cleft. Suppose there are two such sets of loci, all 8 (for simplicity) risk alleles are unlinked, and each di-allelic locus has a carrier frequency of 0.11 for the susceptibility allele. Suppose a fetus with no susceptibility variant sets has a risk (the baseline risk) of 1/1500 and the risk associated with having all four susceptibility variants for either causative set is 1.0, so the relative risk for either high-risk sufficient multi-locus genotype compared to the baseline is 1500. This scenario is summarized in the first column of Table 1.

We performed simulations under the above assumptions (R code will be available at <http://www.niehs.nih.gov/research/atniehs/labs/bb/staff/weinberg/index.cfm#downloads>). We first simulated parental genotypes assuming Hardy-Weinberg equilibrium and generated two offspring per family assuming Mendelian transmission. The clefting status of the offspring was then assigned probabilistically through a Bernoulli trial based on the fetal genotype alone, applying the assumed risk model. We continued simulation until 100,000 families were obtained with the first offspring affected.

Simulation results allowed us to select a scenario for which the population prevalence at birth of clefting was approximately 1/1000 and the recurrence relative risk in siblings was 30. As shown in the first column of Table 1, for this scenario the estimated concordance rate in monozygotic twins was 0.305 while that in dizygotic twins was 0.029, corresponding to the epidemiologic data. The fraction of cases of clefting attributable to genotype (either set) was 0.305. Despite the huge relative risk (1,500) corresponding to co-occurrence of either set of four susceptibility alleles, the marginal relative risk for each SNP, when considered

alone, was only 2.6. The second column of Table 1 shows a modified scenario. In this scenario the carrier frequency at each of the 8 loci was set at 0.15 and the risk in a fetus with no susceptibility variant sets was 1/2475. The relative risk for either high-risk sufficient multi-locus genotype was 1100. The population prevalence at birth of clefting was approximately 1/1200 and the recurrence relative risk in a sibling of a case was 28. The concordance rates in MZ and DZ twins were 0.23 and 0.024 respectively. All of these numbers approximate the known epidemiology of oral clefting while the marginal relative risk for each of the 8 SNPs was only 3.4. In both toy scenarios, SNP-wise analysis would overlook most of the information carried by the four interacting loci and the effects of genotype could easily be missed in a GWAS analysis.

The etiology of clefting apparently also involves both environmental effects and gene-environment interactions. We next considered a more complicated scenario where in addition to the genetic risk factors, an environmental factor, say maternal smoking, can also increase risk of clefting. Again, we assumed two genetic pathways cause clefting, each consisting (for simplicity) of four unlinked SNPs in a high-risk sufficient set. We assume the maternal exposure status remains the same across pregnancies and consequently the concordance rate for DZ twins is the same as the sibling recurrence risk. The concordance rate for MZ twins, $\mathcal{E}(\text{risk}^2)/\mathcal{E}(\text{risk})$ (\mathcal{E} means expectation), where the averaging is over possible exposure (E) by genotype (G) categories, is also computed, by applying the known risk model and the known prevalences for E and G. One such scenario is described in Table 1, column 3. We assumed an exposure prevalence of 0.25, which does not change between the two births for a given mother, i.e. smokers continue to smoke. Except for an exposure relative risk of 1.5, all the other parameters remained the same as in scenario 2 (Table 1, column 2). As there is no GxE interaction in this scenario, the relative risk for either causative gene set is 1100 in both exposed and unexposed. We again used simulations to evaluate this scenario. A dichotomous maternal exposure was simulated via Bernoulli trials with outcome probability equal to the designated exposure prevalence. Our simulations showed that under this scenario the population prevalence at birth of clefting and the recurrence relative risk in a later sibling were approximately 1/1000 and 31 respectively, again agreeing well with epidemiologic data. The concordance rate in monozygotic twins became 0.27 while that in dizygotic twins was 0.03. The marginal relative risk for each SNP, which is the parameter estimated when testing SNP-by-SNP diminished to 3.5.

Our fourth scenario (column 4 of Table 1) now allows also a super-multiplicative interaction between one of the multi-locus genotypes (the first causative gene set) and the exposure, but not the other. The relative risk for the first causative gene set was 1100 in an unexposed maternal/fetal pair and 1650 in an exposed maternal/fetal pair (an interaction of 1.5), while the relative risk for the second causative gene set was 1100 in both exposed and unexposed pairs (no interaction). The exposure prevalence, exposure relative risk, variant allele carrier frequency and baseline disease risk remained the same as those in scenarios 2 and 3. Again the marginal effect for a single SNP was markedly reduced (to 3.3 in the unexposed).

These examples illustrate how single SNP analyses can miss important genetic effects, and future methods development should aim at better identification of GxE and GxG joint effects. Pathway libraries exist to identify genes known to be functionally involved in

particular pathways, and analyses can be undertaken to assess enrichment of those pathways among the “hits” from GWAS. We refer the interested reader to the review by Wang, et al.¹³ However, regulation of pathways is complex and poorly understood, and many genes can be involved in multiple pathways, so methods that do not rely on this prior knowledge are also needed. Another problem is that, unlike in our toy examples, the investigator is typically not lucky enough to have typed the causative SNPs, but must rely on “marker” SNPs (SNPs in linkage disequilibrium with the actual causative SNPs) to “tag” the real ones. This unavoidable use of proxies produces considerable measurement error, adding a further level of complication to inference related to multi-locus and GxE analyses of GWAS. Future methods development will also need to take this into consideration.

Our message here has been discouraging, but there are some positive implications. If we do a careful GWAS or a candidate gene study of a complex condition and fail to find evidence for any risk-related SNPs that can withstand a Bonferroni correction, perhaps it isn't that the genetic effects are all small or we have selected the wrong candidate genes. Perhaps we simply do not yet own the statistical tools we need to connect the dots and tease out the complex interactions that lead to this condition.

Acknowledgments

Financial Support: This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES040007; Z01-ES45002).

We thank Drs Douglas Bell and Dmitri Zaykin for their careful review and valuable comments.

Reference

1. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet.* 2010; 11(6):446–450. [PubMed: 20479774]
2. Chee-Seng K, En Yun L, Yudi P, Kee-Seng C. Genome-wide Association Studies: The Success, Failure and Future. *Encyclopedia of Life Sciences.* 2009
3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature.* 2009; 461(7265):747–753. [PubMed: 19812666]
4. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol.* 2010; 8(1):e1000294. [PubMed: 20126254]
5. Mitchell LE, Risch N. Mode of inheritance of nonsyndromic cleft lip with or without cleft palate: a reanalysis. *Am J Hum Genet.* 1992; 51(2):323–332. [PubMed: 1642234]
6. Birnbaum S, Ludwig KU, Reutter H, Herms S, Steffens M, Rubini M, Baluardo C, Ferrian M, Almeida de Assis N, Alblas MA, Barth S, Freudenberg J, Lauster C, Schmidt G, Scheer M, Braumann B, Berge SJ, Reich RH, Schiefke F, Hemprich A, Potzsch S, Steegers-Theunissen RP, Potzsch B, Moebus S, Horsthemke B, Kramer FJ, Wienker TF, Mossey PA, Propping P, Cichon S, Hoffmann P, Knapp M, Nothen MM, Mangold E. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat Genet.* 2009; 41(4):473–477. [PubMed: 19270707]
7. Mangold E, Ludwig KU, Birnbaum S, Baluardo C, Ferrian M, Herms S, Reutter H, de Assis NA, Chawa TA, Mattheisen M, Steffens M, Barth S, Kluck N, Paul A, Becker J, Lauster C, Schmidt G, Braumann B, Scheer M, Reich RH, Hemprich A, Potzsch S, Blaumeiser B, Moebus S, Krawczak M, Schreiber S, Meitinger T, Wichmann HE, Steegers-Theunissen RP, Kramer FJ, Cichon S,

- Propping P, Wienker TF, Knapp M, Rubini M, Mossey PA, Hoffmann P, Nothen MM. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet.* 2010; 42(1):24–26. [PubMed: 20023658]
8. Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, Liang KY, Wu T, Murray T, Fallin MD, Redett RA, Raymond G, Schwender H, Jin SC, Cooper ME, Dunnwald M, Mansilla MA, Leslie E, Bullard S, Lidral AC, Moreno LM, Menezes R, Vieira AR, Petrin A, Wilcox AJ, Lie RT, Jabs EW, Wu-Chou YH, Chen PK, Wang H, Ye X, Huang S, Yeow V, Chong SS, Jee SH, Shi B, Christensen K, Melbye M, Doheny KF, Pugh EW, Ling H, Castilla EE, Czeizel AE, Ma L, Field LL, Brody L, Pangilinan F, Mills JL, Molloy AM, Kirke PN, Scott JM, Arcos-Burgos M, Scott AF. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nat Genet.* 2010; 42(6):525–529. [PubMed: 20436469]
9. Grant SF, Wang K, Zhang H, Glaberson W, Annaiah K, Kim CE, Bradfield JP, Glessner JT, Thomas KA, Garris M, Frackelton EC, Otieno FG, Chiavacci RM, Nah HD, Kirschner RE, Hakonarson H. A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. *J Pediatr.* 2009; 155(6):909–913. [PubMed: 19656524]
10. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics.* 2008; 9(5):356–369.
11. Kimura, M. *The neutral theory of molecular evolution.* Cambridge University Press; 1985.
12. Murray JC, Schutte BC. Cleft palate: players, pathways, and pursuits. *J Clin Invest.* 2004; 113(12): 1676–1678. [PubMed: 15199400]
13. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet.* 2010; 11(12):843–854. [PubMed: 21085203]

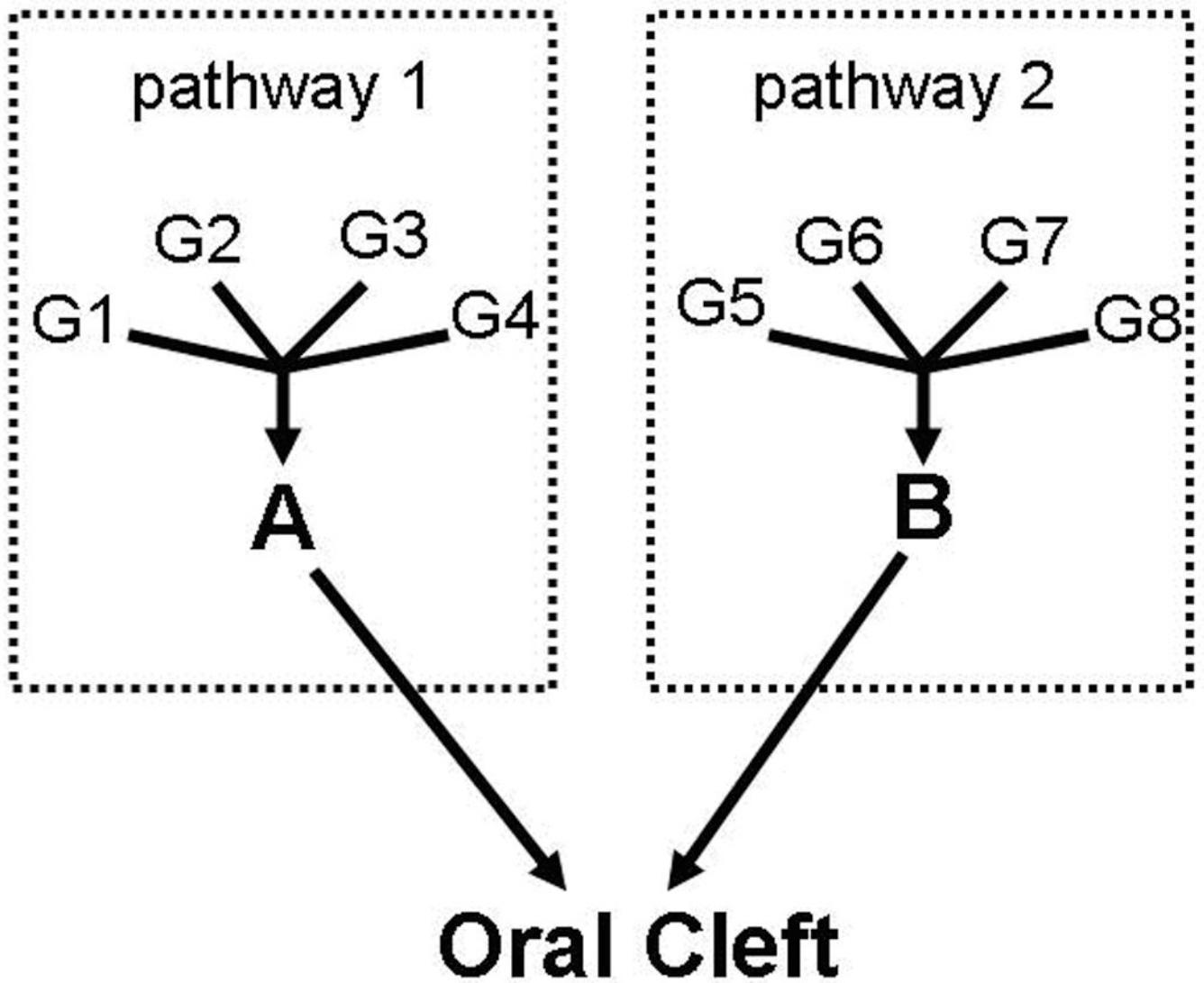


Figure 1.

Schematic drawing of the hypothetical etiology for clefting. Two independent pathways can cause oral clefting. In each pathway the combined dysfunction of four unlinked genes (either G1 through G4 or G5 through G8) is needed to break down the pathway, and breaking down either pathway results in oral clefting.

Table 1

Oral cleft toy examples. Relative risks shown are based on simulating a population of 100,000 families where the first child had oral cleft.

Scenarios	scenario_1 ¹	scenario_2 ¹	scenario_3	scenario_4
Carrier Frequency	0.11	0.15	0.15	0.15
Baseline Risk	1/1500	1/2475	1/2475	1/2475
Genotype RR	1500	1100	1100	1100
Exposure RR	1	1	1.5	1.5
Interaction RR	1	1	1	1.5
Exposure Prevalence	NA	NA	0.25	0.25
Population Prevalence at birth (per 1000 birth)	0.95	0.85	0.96	1.00
Sibling Recurrence Relative Risk	30.0	28.4	31.4	34.4
MZ twin Concordance Rate	0.31	0.23	0.27	0.33
DZ twin Concordance Rate	0.029	0.024	0.030	0.034
RR of SNP in Set 1 in Unexposed	2.6	3.4	3.5	3.3
RR of SNP in Set 2 in Unexposed	2.6	3.4	3.3	3.4
RR of SNP in Set 1 in Exposed	NA	NA	3.4	4.6
RR of SNP in Set 2 in Exposed	NA	NA	3.5	3.0

Abbreviations: RR: relative risk; MZ: monozygotic; DZ dizygotic

¹The marginal relative risk of a SNP in scenarios without risk related to exposure refers to the relative risk in the overall samples (without exposure stratification).