J.R. Shaffer[1,2]*, E. Feingold[1,3], and M.L. Marazita[1,2,4]

[1]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, 130 DeSoto St., A300 Crabtree Hall, Pittsburgh, PA 15260, USA; [2]Center for Oral Health Research in Appalachia, University of Pittsburgh, Pittsburgh, PA, 15261 and West Virginia University, Morgantown, WV 26506; [3]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA; and [4]Center for Craniofacial & Dental Genetics, Department of Oral Biology, School of Dental Medicine, and Clinical and Translational Science Institute, and Department of Psychiatry, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA; *corresponding author, jrs51@pitt.edu

*J Dent Res* 91(7):637-641, 2012

# Genome-wide Association Studies: Prospects and Challenges for Oral Health

## ABSTRACT

The genomic era of biomedical research has given rise to the genome-wide association study (GWAS) approach, which attempts to discover novel genes affecting an outcome by testing a large number (*i.e.*, hundreds of thousands to millions) of genetic variants for association. This article discusses the issues surrounding the GWAS approach with emphasis on the prospects and challenges relevant to the oral health research community.

**KEY WORDS:** genomics, genetics, polymorphism(s), GWAS, Human Genome Project.

## INTRODUCTION

Oral health genetics has been an area of active research for over 80 years (Jackson, 1968), with evidence from twin and family studies overwhelmingly demonstrating the heritability of dental caries (Mansbridge, 1959; Wang *et al*., 2010), periodontal disease (Michalowicz *et al.,* 1991; Mucci *et al.,* 2005), tooth loss (Mucci *et al.,* 2005), and non-syndromic orofacial clefts (Marazita, 2002) and other dental traits (Liu *et al.,* 1998). More recently, candidate gene studies have investigated the associations between a limited number of genetic variants in and around candidate genes—chosen *a priori* based on known biological functions with plausible impact on disease—and many oral health outcomes. Examples include dental caries, for which genes involved in tooth development, salivary function, and diet/taste have been reported (Wright, 2010), and orofacial clefts, which are associated with genes involved in development (Dixon *et al.,* 2011). The genomic era of biomedical research has given rise to the genome-wide association study (GWAS) approach, which attempts to discover novel genes affecting an outcome by testing a large number (*i.e.,* hundreds of thousands to millions) of genetic variants for association. The GWAS design is now being applied to the study of many common human disorders, including oral health outcomes. This article discusses the issues surrounding the GWAS approach with emphasis on the prospects and challenges relevant to the oral health research community.

## UNDERSTANDING GWAS

The GWAS approach was made possible by biotechnological and bioinformatics advances over the past decade, thanks in large part to the Human Genome Project. These advances include: (1) databases such as dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), an archive of known Single Nucleotide Polymorphisms (SNPs, *i.e.,* simple genetic variants), and the International HapMap Project (The International HapMap Consortium, 2003), a catalog of genetic variation in individuals across several ancestral populations, which provides a wealth of reference information on the human genetic variation; (2) "SNP chip" microarray technologies, which have made affordable the simultaneous assessment of up to millions of SNPs in an individual DNA sample; and (3) computer resources, including sophisticated software for genotype calling, data cleaning, and genetic association analysis, and

DOI: 10.1177/0022034512446968

Received October 13, 2011; Last revision March 26, 2012; Accepted April 9, 2012

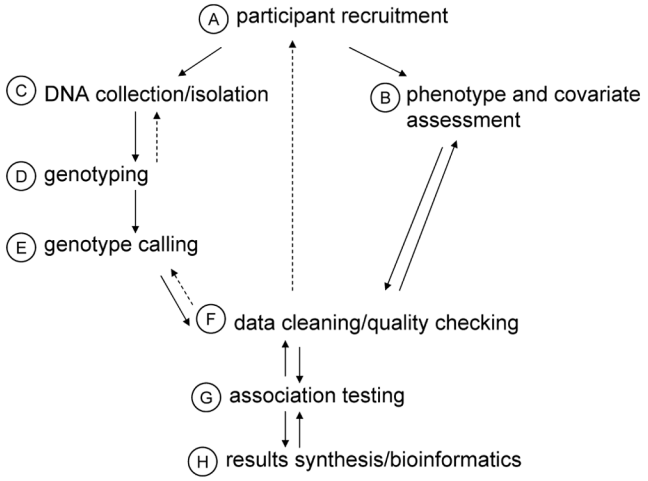


**Figure.** Flow chart of a GWAS study: Arrows indicate forward and backward progression through steps. Dashed arrows indicate re-visiting previous steps to troubleshoot possible problems. **(A)** Participants may be collected according to a variety of study designs: family-based, population-based, case/control, etc. **(B)** Accurate and complete collection of phenotype/covariate data is critical. Standardized assessment tools (see PhenX toolkit, www.phenxtoolkit.org) may facilitate harmonization of data across studies, meta-analysis, and discovery of gene-by-environment interactions. **(C)** Whole-blood and saliva samples are preferred sources of DNA. **(D,E)** High-throughput genotyping is available *via* several Illumina and Affymetrix "SNP chip" platforms. Appropriate genotype calling depends on the particular platform used. Poor-quality DNA samples may need to be recollected and/or replaced. **(F)** Substantial data cleaning, quality checking, and pre-processing are necessary, including rigorous investigations of the following items: SNP call rates to identify/exclude poorly genotyped variants, genotype batch effects to detect genotyping artifacts, Hardy-Weinberg equilibrium to identify poorly performing SNPs, relationship testing to verify known kinships and detect cryptic kinships, gender tests to help identify sample swaps, and tests for "connectivity" to identify sample contamination. Additional important data processing includes assessment of population structure, assessment of large and small chromosomal aneuploidy in individual samples, and phenotype/covariate data cleaning. **(G)** Appropriate tests for association based on sample type (family, population, case/control) and phenotype (binary, categorical, continuous, non-normal, etc.) are available. **(H)** Results of statistical tests for up to millions of genetic markers must be synthesized by graphical and bioinformatics software to identify top signals and pull relevant information for associated loci such as linkage disequilibrium, physically proximal genes, variant putative functions, gene functions, biological pathways, etc.

powerful hardware capable of handling the computational burden and storage of very large datasets. Ongoing improvements in all three of these areas have made possible GWAS projects of ever-increasing size and scope.

GWAS is usually hypothesis-generating rather than hypothesis-testing (Fig.). While GWAS discoveries can be directly useful for clinical risk prediction on rare occasions, the more typical scientific path is for GWAS to generate hypotheses about genes that may be associated with a disorder. Those genes are then studied further for better understanding of the biology of the disease, which in turn eventually leads to clinically useful knowledge that may or may not be directly related to the original genes. This exploratory perspective is reflected in the SNP chip design, which typically assays common variants (*i.e.*, those occurring at frequencies $\geq 5\%$) across the entire genome chosen for near-uniform coverage. Because only common variants are assayed, GWAS projects are therefore designed according to the "common disease-common variant" hypothesis, which predicts that common, complex human diseases are due to the cumulative effects of common risk alleles at many genetic loci, each with weakly detrimental effects. Accordingly, in a GWAS project, each SNP (of potentially millions) is individually tested for association with the outcome. The results (*i.e.*, effect estimates and p values) from these numerous tests are synthesized and explored by a variety of graphical and bioinformatics tools, to answer pertinent questions: What are the top-ranking associated SNPs? In or near what genes are these SNPs physically located? What other genetic variants are in linkage disequilibrium (*i.e.,* genetically correlated) with these SNPs? What are the putative functional roles of these SNPs? What are the biological roles of the nearby genes?

The answers to these questions help prioritize SNPs and/or genetic loci for replication in independent samples and/or follow-up functional studies. False-positives are prevented by the application of an arbitrary threshold for "genome-wide significance", often p value $< 5 \times 10^{-8}$, which corresponds to the Bonferroni adjustment for one million independent tests. However, given that the SNPs tested for genetic association are not independent, but rather highly correlated across regions of the genome, this adjustment is extraordinarily conservative, and runs counter to the hypothesis-generating nature of the GWAS approach. In many cases, GWAS projects contain insufficient sample sizes to detect realistic associations at such strict significance; thus, in practice, researchers may relax the statistical burden of evidence and place greater weight on the bioinformatics results of the top hits. This strategy is likely to be important for initial GWAS of oral health outcomes, because individual studies are unlikely to command the very large sample sizes on par with GWAS consortia (samples sizes of 10,000 or more) currently investigating other common, complex diseases. As the field matures, GWAS consortia may play a similarly important role in identifying and characterizing the genes involved in oral health conditions.

The GWAS approach has been hugely successful in implicating novel genetic variants in common disease. Thousands of genetic associations for complex disease from over 951 GWAS studies had been published by the second quarter of 2011 (see the Catalog of Published Genome-Wide Association Studies, www.genome.gov/gwastudies). This total includes a few recent GWAS papers on oral health outcomes such as oral clefting (Birnbaum *et al.,* 2009; Grant *et al.,* 2009; Mangold *et al.,* 2009; Beaty *et al*., 2010), childhood dental caries (Shaffer *et al*., 2011), and tooth eruption (Pillas *et al.,* 2010; Geller *et al*., 2011). Several more GWAS studies of these oral health outcomes, plus periodontal disease, dental fear, and saliva flow rate, are currently under way. Notable associations for oral health outcomes include the *PVT1*/*GSDMC* locus on chromosome 8q24, which was independently identified in three GWAS studies of non-syndromic cleft lip with or without cleft palate (CL/P; Birnbaum *et al*., 2009; Grant *et al*., 2009; Beaty *et al*.,

**Table.** Major Issues and Characteristics of the GWAS Study Design

| Issue or Characteristic | Effect on GWAS |
|---|---|
| Hypothesis-generating approach | GWAS is a discovery-based approach with the potential to lead to new hypotheses; GWAS is usually not well-suited for rigorous testing of existing hypotheses or clinical prediction. |
| Weak effects | Individual genetic variants require large sample sizes for detection and may individually have negligible clinical impact/predictive utility. |
| Genetic heterogeneity | Different causal alleles from different ancestral populations may prevent detection and/or replication. |
| Admixture | Recent interbreeding of historically genetically isolated populations affects patterns of linkage disequilibrium, requiring careful adjustment to prevent false-positive signals. |
| Population substructure | The presence of genetically distinct sub-samples requires careful adjustment to prevent false-positive signals. |
| Genotyped SNPs are not causal | Observed genetic variants are (usually) not themselves causal; instead, they are in linkage disequilibrium (*i.e.,* correlated) with unobserved causal variants, which require experimental work to identify. |
| Linkage disequilibrium | Genetic variants are correlated with other physically proximal variants, which makes appropriate adjustment for multiple comparisons difficult and hinders determination of causal alleles within an associated "linkage block" (*i.e.,* group of correlated variants). |
| Incomplete coverage | Not all genetic variants are adequately observed either directly or through proxy; therefore, true associations may be missed. |
| "Missing heritability" | The cumulative effects of all associated loci identified by GWAS can explain only a fraction of the estimated heritability from family studies. |
| Genome-wide significance | Very strict (and arbitrary) significance thresholds are necessary to prevent false-positives due to the large number of tests performed; this statistical burden of evidence is sometimes unachievable for realistic effect sizes and sample sizes. |
| Synthesizing results | Results of statistical tests are too vast for interpretation *via* inspection, and therefore must be processed by visualization and bioinformatics tools. |
| Poor clinical prediction | Because of weak individual effects and the problem of "missing heritability", associations identified through GWAS usually cannot be used for accurate prediction of disease. |
| "Deep phenotyping" | Clinical outcomes caused by the convergence of multiple pathophysiologies may be poor phenotypes for GWAS; instead, sub-clinical measures and/or "endo-phenotypes", which more closely represent individual pathophysiologies, may be better suited for gene discovery. |

2010), though its mechanism of action remains unknown. Among the other loci significantly associated with CL/P is the *IRF6* gene (Birnbaum *et al.,* 2009; Beaty *et al*., 2010), mutations in which also cause Van der Woude syndrome. No genome-wide significant associations have yet been reported for dental caries, although a recent GWAS scan reported suggestive associations with several biologically plausible genes including *ACTN2*, which regulates ameloblasts during tooth enamel formation, and *TFIP11*, an enamel gene that localizes to the extracellular enamel matrix (Shaffer *et al.,* 2011). Independent GWAS scans for both primary and permanent tooth eruption have implicated variants at chromosome 12q14 (near *HMGA2*) and chromosome 2q35 (near *TNP1*), though the biological roles of these loci in tooth eruption are unknown (Pillas *et al.,* 2010; Geller *et al.,* 2011). The success of GWAS in identifying and in some cases replicating genetic associations with oral health conditions has opened doors to new avenues of basic science that may ultimately lead to better understanding of disease. But much more work is needed before the fruits of GWAS may be fully realized. Additional studies exploring the cellular function of implicated genetic variants may lead to the discovery of biological pathways and mechanisms affecting oral health outcomes, which in turn may lead to biomarkers for predicting disease, early identification of at-risk patients, or novel therapeutic targets.

## LIMITATIONS OF GWAS

Despite the overall success of the GWAS approach in identifying novel genetic variants affecting complex disease, there are important limitations, complications, and failures of this study design (Table). Foremost is the issue of very weak effect sizes of individual associated SNPs (*i.e*., odds ratios typically less than 1.5 and often much less for the majority of associations), which has been observed for nearly all GWAS studies. Though weak individual effects are expected under the "common disease-common variant" hypothesis and are predicted by population genetics theory that high-risk alleles should be eliminated from a population through selection, the practical consequence is that huge sample sizes (thousands to tens of thousands) are necessary for adequate statistical power to detect genome-wide associations. Moreover, the weak effect sizes of associated loci mean that individual variants may not be clinically meaningful, even if statistically significant. To further complicate the interpretation of results, genetic heterogeneity (*i.e.*, different causal variants across populations), admixture (*i.e.*, interbreeding of historically genetically isolated populations), and population substructure (*i.e.*, genetically distinct sub-samples within a population or sample) can all have an adverse impact on the capacity of GWAS to identify true genetic associations and avoid false-positives. Therefore, great caution is necessary at each stage of the study (*e.g.*, sample recruitment,

genotyping, data cleaning, analysis, and interpretation of results) to control for these intricacies appropriately and guard against erroneous findings.

Another difficulty of GWAS is interpreting individual associated variants. Even though many variants are assayed, this represents only a small fraction of the total genetic variation. The assumption of GWAS is that assayed SNPs are not themselves causal variants, but instead are in linkage disequilibrium (LD)—that is, physically proximal and correlated due to population history—with the causal variants. A given gene will contain hundreds or thousands of variants, but any given panel of GWAS markers typically includes ten or fewer SNPs within each gene. GWAS leverages the LD structure of the genome by seeking genetic associations with observed SNP proxys for the unobserved causal variant(s). Leveraging LD makes GWAS realizable because genotyping all possible genomic variants in large study samples is currently cost-prohibitive. The downside of this strategy is that identifying the nearby causal variant can be very difficult, and exacerbated by the uncertainty of where a causal variant may be physically located in relation to the gene it affects. Coding variants within a gene (*i.e.*, variants that alter the protein product of a gene) represent only a fraction of genetic associations. Instead, non-coding causal genetic variants that somehow affect regulation of gene expression, alternative splicing, DNA methylation, DNA folding, RNA stability, or other regulatory systems likely represent a large portion of associated loci, and may be located at various physical distances from the gene they influence. Understanding the mechanisms of action in these cases is quite difficult. Therefore, the progression from statistical association observed through GWAS to inferred causality and/or functional consequences for disease can be arduous.

Another important issue for the GWAS approach is the problem of "missing heritability", that is, the gap between the amount of disease variance cumulatively accounted for by all associated variants and the estimated genetic variance calculated from twin and family studies. For nearly every outcome studied to date, associated variants explain only a fraction of the total trait heritability (Manolio *et al*., 2009). Several reasons for the "missing heritability" have been proposed, such as sparse genomic coverage or the existence of numerous common variants with such weak effects that they have yet to be discovered. Other explanations include the effects of rare alleles (occurring at frequencies < 1%) and copy number variations (*i.e.*, differences between individuals in the number of copies of a genomic region or gene, but not necessarily differences in the genetic sequence itself) that are not adequately assayed or tested under the GWAS approach. It is also possible that the "missing heritability" is partly due to gene-by-gene interactions and gene-by-environment interactions, the effects of which may not be observed or wholly appreciated under the typical GWAS approach of testing associations of individual genetic variants one at a time. Additionally, epigenetics—that is, inherited patterns of DNA methylation that can affect gene expression—may account for a portion of trait heritability not strictly due to genetic variation.

These issues may be very important for GWAS investigations of oral health outcomes, such as dental caries, which is clearly multifactorial. Numerous avenues may lead to genetic susceptibility to dental caries, such as host defense to cariogenic bacteria, tooth positional and morphological characteristics, enamel composition, dietary choices, oral hygiene behaviors, salivary composition and flow rate, and others. Potentially tens, hundreds, or thousands of individual genetic variants operating through numerous biological mechanisms may affect each of these converging risk factors. Unraveling this complexity will be difficult and may require clever or innovative analysis of genetic data. Similarly intricate scenarios are likely for other oral outcomes, including CL/P, periodontal disease, malocclusion, fluorosis, tooth agenesis, and others.

## CHALLENGES AND OPPORTUNITIES

Moving forward, the oral health research community has an invaluable opportunity to take advantage of the wealth of lessons learned from years of GWAS studies in other fields. While research on the genetics of oral health lags behind that on many other diseases—which is surprising and unfortunate, given the extent and impact of oral health problems—we are poised to make rapid progress into genomics territory. Success will require applying best practices in the pursuit of genetic discoveries, while carefully defining what "success" actually entails.

Among the most critical issues are careful phenotype assessment, including sub-clinical measures [such as subepithelial orbicularis oris muscle defects as an expanded cleft lip/cleft palate phenotype (Marazita, 2007)], and novel "endophenotypes" generated through statistical modeling [such as orospatial modeling of surface-level dental caries assessments (Shaffer *et al.,* 2012)]. This so-called "deep phenotyping", which may more closely reflect disease at the cellular level, holds potential for improving the ability of the GWAS approach to find novel genes. Psychiatric genetics has a long history of studying endophenotypes, which have proved to be a rich source of discoveries. Many of the approaches used in that work can be adapted to oral health phenotypes.

Another opportunity is to carefully design GWAS studies to adequately negotiate the issues uncovered by work in other fields. Oral health researchers should expect to find associations with regulatory variation, which will require additional functional work, and be prepared to thoughtfully model gene-by-environment interactions, gene-by-gene interactions, and copy number variations. Gene-by-environment interactions may be especially critical for diseases with important environmental risk factors, such as fluoride exposure for dental caries. Care must also be taken to deal appropriately with population substructure, admixture, and genetic heterogeneity, using methods and insight derived from previous work. Last, innovative methods that dispense with the purely hypothesis-free study design may be important for making the most of GWAS studies of oral health outcomes. Newer methods that model biological pathways, pulling together statistical results with bioinformatics from public databases, have the potential to identify not only genes of interest, but also entire functional pathways and biological mechanisms affecting disease. Judicious use of newer methodologies, to complement the classic GWAS approach, may yield important insights into the genetic nature of disease.

Ultimately, overcoming the many hurdles of GWAS may require very large sample sizes, which may be obtainable only by combining the efforts of multiple smaller studies into a "mega-consortium". To facilitate this, oral health researchers should adopt standard definitions and assessment methods for oral health conditions and important covariates (thus eliminating the problem of phenotype harmonization that has plagued mega-consortia of other diseases). The PhenX toolkit (www.phenx.org) is a resource that provides consensus assessments aimed at facilitating cross-study comparisons and combined analyses. Similarly, applying standardized quality-assurance and genotype imputation procedures may assist in merging large genetic datasets for combined analysis. Development of new meta-analysis methods that combine association results for a genetic locus, while allowing the exact associated SNP to differ among studies, may be necessary to improve statistical power and identify weak or heterogeneous effects.

In summary, the field of oral health research has already made its first steps into the realm of genomics. As more GWAS studies of oral health traits are planned and executed, we have the luxury to learn from pioneering work in other fields, design our studies accordingly, and utilize the wealth of analytical and bioinformatics tools at our disposal, to make the most of our GWAS investments and bring meaningful discoveries to light.

## ACKNOWLEDGMENT

The authors acknowledge funding from the NIH/NIDCR (R03-DE021425 and R01-DE014899). The authors declare no potential conflicts of interest with respect to the authorship and/or publication of this article.

## REFERENCES

Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, *et al.* (2010). A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nat Genet* 42:525-529.

Birnbaum S, Ludwig KU, Reutter H, Herms S, Steffens M, Rubini M, *et al.* (2009). Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat Genet* 41:473-477.

Dixon MJ, Marazita ML, Beaty TH, Murray JC (2011). Cleft lip and palate: understanding genetic and environmental influences. *Nat Rev Genet* 12:167-178.

Geller F, Feenstra B, Zhang H, Shaffer JR, Hansen T, Esserlind AL, *et al.* (2011). Genome-wide association study identifies four loci associated with eruption of permanent teeth. *PLoS Genet* 7:e1002275.

Grant SF, Wang K, Zhang H, Glaberson W, Annaiah K, Kim CE, *et al.* (2009). A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. *J Pediatr* 155:909-913.

The International HapMap Consortium (2003). The International HapMap Project. *Nature* 426:789-796.

Jackson D (1968). Genes and dental caries. *Proc R Soc Med* 61:265-269.

Liu H, Deng H, Cao CF, Ono H (1998). Genetic analysis of dental traits in 82 pairs of female-female twins. *Chin J Dent Res* 1:12-16.

Mangold E, Ludwig KU, Birnbaum S, Baluardo C, Ferrian M, Herms S, *et al.* (2009). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet* 42:24-26.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, *et al.* (2009). Finding the missing heritability of complex diseases. *Nature* 461:747-753.

Mansbridge JN (1959). Heredity and dental caries. *J Dent Res* 38:337-347.

Marazita ML (2002). Segregation analysis. In: Cleft lip and palate: from origin to treatment. Wysznyski D, editor. Oxford, UK: Oxford University Press, pp. 222-233.

Marazita ML (2007). Subclinical features in non-syndromic cleft lip with or without cleft palate (CL/P): review of the evidence that subepithelial orbicularis oris muscle defects are part of an expanded phenotype for CL/P. *Orthod Craniofac Res* 10:82-87.

Michalowicz BS, Aeppli D, Virag JG, Klump DG, Hinrichs JE, Segal NL, *et al.* (1991). Periodontal findings in adult twins. *J Periodontol* 62:293-299.

Mucci LA, Bjorkman L, Douglass CW, Pedersen NL (2005). Environmental and heritable factors in the etiology of oral diseases—a population-based study of Swedish twins. *J Dent Res* 84:800-805.

Pillas D, Hoggart CJ, Evans DM, O'Reilly PF, Sipila K, Lahdesmaki R, *et al.* (2010). Genome-wide association study reveals multiple loci associated with primary tooth development during infancy. *PLoS Genet* 6:e1000856.

Shaffer JR, Wang X, Feingold E, Lee M, Begum F, Weeks DE, *et al.* (2011). Genome-wide association scan for childhood caries implicates novel genes. *J Dent Res* 90:1457-1462.

Shaffer JR, Feingold E, Wang X, Cuenco KT, Weeks DE, DeSensi RS, *et al*. (2012). Heritable patterns of tooth decay in the permanent dentition: principal components and factor analyses *BMC Oral Health* 12:7.

Wang X, Shaffer JR, Weyant RJ, Cuenco KT, DeSensi RS, Crout R, *et al.* (2010). Genes and their effects on dental caries may differ between primary and permanent dentitions. *Caries Res* 44:277-284.

Wright JT (2010). Defining the contribution of genetics in the etiology of dental caries. *J Dent Res* 89:1173-1174.