Feasibility of High-Throughput Genome-Wide Genotyping using DNA from Stored Buccal Cell Samples

Share Your Story

Biomarker Insights

METHODOLOGY

Feasibility of High-Throughput Genome-Wide Genotyping using DNA from Stored Buccal Cell Samples

Stephanie J. Loomis[1], Lana M. Olson[2], Louis R. Pasquale[1], Janey Wiggs[1], Daniel Mirel[3], Andrew Crenshaw[3], Melissa Parkin[3], Brandon Rahhal[3], Stephanie Tetreault[4], Peter Kraft[4], Shelley S. Tworoger[5], Jonathan L. Haines[2] and Jae H. Kang[5]

[1]Massachusetts Eye and Ear Infirmary, 243 Charles Street, Boston, MA 02114, USA. [2]Center for Human Genetics Research, Vanderbilt University Medical Center, 2215 Garland Avenue, Nashville, Tennessee, 37232, USA. [3]Center for Genotyping and Analysis, Broad Institute of Massachusetts Institute of Technology and Harvard University, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [4]Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115. [5]Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115. Corresponding author email: louis_pasquale@meei.harvard.edu

Abstract: It is unclear if buccal cell samples contain sufficient human DNA with adequately sized fragments for high throughput genetic bioassays. Yet buccal cell sample collection is an attractive alternative to gathering blood samples for genetic epidemiologists engaged in large-scale genetic biomarker studies. We assessed the genotyping efficiency (GE) and genotyping concordance (GC) of buccal cell DNA samples compared to corresponding blood DNA samples, from 32 Nurses' Health Study (NHS) participants using the Illumina Infinium 660W-Quad platform. We also assessed how GE and GC accuracy varied as a function of DNA concentration using serial dilutions of buccal DNA samples. Finally we determined the nature and genomic distribution of discordant genotypes in buccal DNA samples. The mean GE of undiluted buccal cell DNA samples was high (99.32%), as was the GC between the paired buccal and blood samples (99.29%). GC between the dilutions versus the undiluted buccal DNA was also very high (>97%), though both GE and GC notably declined at DNA concentrations less than 5 ng/μl. Most (>95%) genotype determinations in buccal cell samples were of the "missing call" variety (as opposed to the "alternative genotype call" variety) across the spectrum of buccal DNA concentrations studied. Finally, for buccal DNA concentration above 1.7 ng/ul, discordant genotyping calls did not cluster in any particular chromosome. Buccal cell-derived DNA represents a viable alternative to blood DNA for genotyping on a high-density platform.

Keywords: buccal cell DNA, saliva DNA, GWAS, genotyping, low concentration DNA

This article is available from http://www.la-press.com.

Introduction

The use of genome-wide association studies (GWAS) as a means of detecting genetic biomarkers (single nucleotide polymorphisms (SNPs)) associated with disease typically requires large sample sizes, and the collection of DNA specimens from buccal samples represents an opportunity to minimize the involvement of study personnel or ancillary services in the sample collection process. Furthermore, buccal cell sample collection is minimally invasive, allowing study subjects to collect their sample from home and send it back via the mail, which can maximize participation when the study population is geographically dispersed.[1–6] However, inadequate DNA yield,[3–6] the presence of bacterial DNA,[7] and the need to possibly store samples for long periods prior to use may compromise the utility of buccal cell samples for GWAS. Re-sampling in aging cohorts to augment DNA concentrations may not be an option because of limited resources, loss to follow up, or death of cohort members.

Researchers have demonstrated the efficacy of buccal cell-derived DNA samples for use in genetic biomarker studies in low[8–14] and moderately high density genotyping platforms (between 200,000 and 400,000 single nucleotide polymorphisms),[7,15–16] but no study evaluated whether buccal cell DNA produces acceptable genotype efficiency and accuracy on higher density genotyping platforms. Such information may be useful to researchers deciding how to best utilize non-replenishable buccal cell repositories that may be depleted in the genotyping process.

In this study, we assessed genotyping efficiency and accuracy of paired buccal cell- and blood-derived DNA samples from 32 Nurses' Health Study (NHS) participants genotyped on the Illumina Infinium 660W-Quad genotyping platform. We also performed serial dilutions on a subset of buccal cell DNA samples to assess whether a genome-wide scan was feasible at low DNA concentrations.

Materials and Methods

The Nurses' Health Study is a longitudinal cohort study of 121,700 nurses from across the United States, with data collected since 1976.[17] After IRB approval and informed consent from participants, we collected blood and buccal cells from a subset of NHS participants, and randomly chose 32 individuals with both blood and buccal DNA for this study.

We collected blood samples from May 1989 to September 1990, by mailing blood collection kits to participants; kits included mail back instructions and an icepack to facilitate return of samples via an overnight courier. We then separated buffy coat from red blood cells and plasma, and stored all fractions at the NHS Blood Lab in freezers at −130 °C. We later extracted DNA using QIAamp 96 DNA blood kit (Qiagen #51162).

We also collected buccal cells from participants by the mail from 2000 to 2002 and from 2004 to 2006. Participants used the "swish and spit" method, swishing with 10 ml Scope® mouthwash for 30 seconds and spitting into a Nalgene sample cup that they then mailed back to the investigators within 24 hours of collection. Before processing, we stored the samples in their original vial at 4 °C for up to one week. We centrifuged the samples and extracted DNA from 200 μl of the buccal pellet using QIAamp 96 DNA blood kit (Qiagen #51162). We then eluted DNA in 200 μl Tris EDTA and stored it at −80 °C.

The blood and buccal cell samples were stored for a mean of 19.47 (range, 18.67–19.67) years, and 4.89 (range, 4.33–5.25) years respectively. The mean age at blood draw was 53.64 (range, 43.33–68.33) years, and the mean age at buccal collection was 68.23 (range, 57.00–83.00) years.

We measured all DNA concentrations with the Quant-iT PicoGreen dsDNA Assay (Invitrogen P7589) at the Broad Institute, Cambridge, MA. The Broad Institute Genetic Analysis Platform (GAP) genotyped 32 blood and 32 paired buccal samples, four buccal cell duplicates, and one HapMap control using the Sequenom iPLEX platform for a 24-SNP Finger-Print panel (FP). We then loaded 4 μl of each sample for genotyping on the Illumina Infinium 660W-Quad genotyping platform and generated genotyping calls using the Illumina BeadStudio and Autocall software and the Illumina-provided genotype cluster definitions file (Human660W-Quad_v1_A.egt, generated using HapMap project DNA samples).[18]

We serially diluted a randomly selected subset (n = 4) of the original buccal cell samples at 1:2, 1:4, 1:8 and 1:16, and determined the concentrations by Picogreen flourometry. We genotyped the diluted samples, the undiluted samples and two HapMap

controls on the Illumina Infinium 660W-Quad. We arrayed samples on the production plate so as to minimize confounding by batch and processing effects.

To analyze the raw genotyping data, we used the programs PLATO[19] and PLINK.[20–21] For SNP quality filtering, we pooled data from the two rounds of genotyping, and checked samples' gender and relatedness. Of the 657,366 total SNPs, we removed all copy number variant loci (n = 64,527), SNPs that failed across all samples (n = 31,349), Y chromosome SNPs (n = 8) and SNPs with coding discrepancies (n = 1), leaving 561,481 SNPs for analysis. For each sample, we calculated the genotype efficiency (GE) as the number of uncalled SNPs divided by 561,481, all subtracted from one. After removing filtered SNPs (n = 95,885), we calculated genotype concordance (GC) as the percentage of calls that were identical between two samples. We used SAS (v 9.1, SAS Institute, Cary, NC) for statistical analysis, and as the data were not normally distributed (determined by the Kolmogorov-Smirnov test) we used non-parametric tests in analysis. We tested the data from the paired blood and buccal samples with Wilcoxon signed-rank tests, and the associations between dilutions or concentrations in relation to GE or GC with Spearman's rank correlation test. We considered P values < 0.05 statistically significant.

Results

As shown in Table 1, both sample types (blood and buccal) showed mean genotype efficiencies (GEs) of >99%. All samples had GEs > 95%, and most samples (100.00% of blood and 93.75% of buccal) had GEs > 98%. The mean GE was higher in blood (99.89%) compared to buccal samples (99.32%; $P < 0.0001$ for the difference in GE). The GE was not correlated with buccal DNA concentrations in the range 10–50 ng/μl ($P = 0.09$), nor was GE correlated with years of sample storage ($P = 0.93$ for blood, $P = 0.92$ for buccal) or age of participant at collection ($P = 0.16$ for blood, $P = 0.35$ for buccal). The mean genotype concordance (GC) between buccal DNA and the blood DNA was extremely high (mean GC = 99.29 ± 0.60%).

In a second round of genotyping, serial dilutions (1:2, 1:4, 1:8, and 1:16) were made from four of the original 32 buccal cell samples. The mean concentrations in these samples ranged from 1.10 ng/μl at the lowest dilution (1:16) to 31.63 ng/μl for the undiluted samples (Table 2). The mean GEs were still high, with only the lowest dilution showing values of <98%. Nonetheless, GEs did significantly decrease with either increasing DNA dilution ($P = 0.04$) or decreasing DNA concentration ($P = 0.02$). GC also significantly decreased with decreasing DNA concentration when either undiluted buccal cell DNA ($P = 0.01$) or the appropriate paired blood DNA sample was used as the reference ($P = 0.03$). However, even at the lowest dilutions, the GC remained high (>97%).

As there were variations in DNA concentrations at each dilution, we also examined GE and GC as a function of DNA concentration in more detail. We divided samples into concentration ranges with an equal number of samples in each range. The mean GE and the mean GC (using the undiluted buccal DNA results as a reference to assess GC) as a function of DNA concentration range is illustrated in Figure. 1. Even at the lowest concentration range (≤1.6 ng/μl), both mean

Table 1. Genotype efficiency and concordance in matched buccal and blood samples from 32 women in the Nurses' Health Study.

	DNA source		P value[*]
	Blood (n = 32)	**Buccal (n = 32)**	
Mean concentration (SD) (ng/μl)	45.96 (9.18)	43.04 (11.98)	0.13
Concentration range (ng/μl)	14.82–50.00	10.00–50.00	–
Mean genotype efficiency (SD) (%)[†]	99.89 (0.05)	99.32 (0.57)	<0.0001
Genotype efficiency range (%)[†]	99.78–99.95	97.79–99.89	–
Mean concordance (SD) (%)[‡]		99.29 (0.60)	–
Concordance range (%)[‡]		97.60–99.91	–

[*]P values from paired Wilcoxon tests of the differences between the paired blood and buccal samples.
[†]Genotype efficiency was calculated as 1-(number of called SNPs/561,481).
[‡]Concordance was calculated as the percentage of SNPs with identical calls between the two samples being compared.

Table 2. Concordance between serial dilutions of buccal DNA with paired undiluted buccal DNA and paired blood DNA among four women in the Nurses' Health Study.

Dilution	Mean concentration (SD) (ng/µl)	Mean GE (SD) (%)[*]	Mean concordance with undiluted buccal DNA (SD) (%)[†]	Mean concordance with blood DNA (SD) (%)[†]
1:1 (undiluted)	31.63 (18.52)	99.55 (0.23)	n/a	99.46 (0.23)
1:2	15.83 (7.28)	99.46 (0.32)	99.38 (0.21)	99.36 (0.30)
1:4	7.20 (3.32)	99.32 (0.24)	99.25 (0.13)	99.22 (0.24)
1:8	3.18 (1.65)	98.17 (2.45)	98.07 (2.58)	97.96 (2.67)
1:16	1.10 (0.70)	97.57 (2.78)	97.44 (2.79)	97.35 (2.94)
P value[‡] (by dilution)	n/a	0.04	0.02	0.03
P value[‡] (by concentration)	n/a	0.02	0.01	0.03

[*]4 µl of each sample was loaded onto the Infinium platform; GE was calculated as 1-(number of called SNPs/561,481).
[†]Concordance was calculated as the percentage of SNPs with identical calls between the two samples being compared.
[‡]P values were derived from Spearman correlation tests between dilution or concentration in relation to GE or concordances.

GE and mean GC were >97%; however, there was a notable decline for GE and GC below 4.6 ng/µl.

Although overall GCs were very high between diluted and undiluted samples, we explored the nature of genotyping discordances (GDs) that occurred at decreasing DNA concentrations. When compared to the genotyping calls from the undiluted buccal DNA sample, there were two general types of discordance: a "missing genotyping call", (e.g. no genotype call vs. A/T, or vice versa), and a "different genotyping call" (e.g. A/A vs. A/T or T/T). Table 3 shows a breakdown of the type of GDs in each sample dilution. Across all

samples and dilutions, there were many more "missing call" than "different call" discordances; overall, 96.95% of the GDs were of the "missing call" type, while only 3.05% of the GDs were of the "different call" type. The percentage of "different call" discordances increased with greater dilution ($P = 0.04$).

To further quantify the SNP discordances, we assigned concordance scores to each SNP corresponding to the number of individuals, out of four total, in which a SNP was discordant between the undiluted and diluted sample. We repeated this process for each concentration range. A score of 0 indicated that the SNP was concordant between the undiluted and diluted samples for none of the four individuals, and a score of 4 indicated that a SNP was concordant between diluted and undiluted samples for all four individuals. For samples with $\geq$10 ng/µl of DNA, 98.18% of SNPs were concordant with the undiluted sample in all four individuals, 1.35% of SNPs were concordant in three individuals, 0.38% of SNPs were concordant in two individuals, 0.08% of SNPs were concordant in only one individual, and 0.01% of SNPs were concordant to the undiluted sample in none of the individuals. The percentages of SNPs achieving concordance scores from 0 to 4 at each dilution are summarized in a bar graph (see Fig. 2). For all four concentration ranges, the majority of SNPs showed concordance between the diluted and undiluted sample in all four individuals, yet as concentration decreased, the percentage of SNPs concordant across the four individuals decreased.

To explore the origin of discordances, we found the chromosomal location of all discordant SNPs across

Mean genotype efficiency and concordance by concentration range
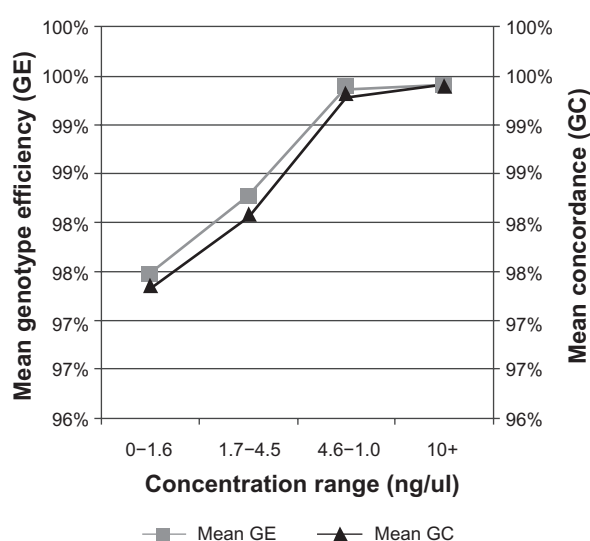
Figure 1. Mean genotype efficiency and concordance with undiluted buccal samples range from 97.35% to 99.40%, decreasing with concentrations less than 4.6 ng/µl.

Table 3. Detailed analysis of discordant genotype calls among buccal cell samples at various diluted concentrations.

Concentration (ng/µl)	Concordance (%)*	Mean number of discordant SNPs (%)		
		All discordances	Missing call	Different call
10+	99.40	3372	3306 (97.94)	66 (2.06)
4.6–10	99.31	3889	3809 (97.90)	80 (2.10)
1.7–4.5	98.08	10762	10129 (96.21)	634 (3.79)
0–1.6	97.35	14874	14196 (95.76)	679 (4.24)
P value†		–	–	0.04

*Concordance was calculated as the percentage of SNPs with identical calls between the dilutions and the undiluted samples.
†*P* values were derived from Spearman correlation tests between concentration and percentage of different calls. Explanations for missing calls and different calls are provided in the text.

the genome and discovered that the discordance did not significantly cluster in any chromosome for all but the lowest DNA concentration range ($P = 0.58$–0.68 for concentration ranges above 1.7 ng/µl). At lower DNA concentrations, however, there were slightly more discordant SNPs on chromosomes 4, 5, 6 and 18, which had discordant percentages $\geq 0.13\%$ compared to a mean discordant percentage of 0.10% across the genome overall.
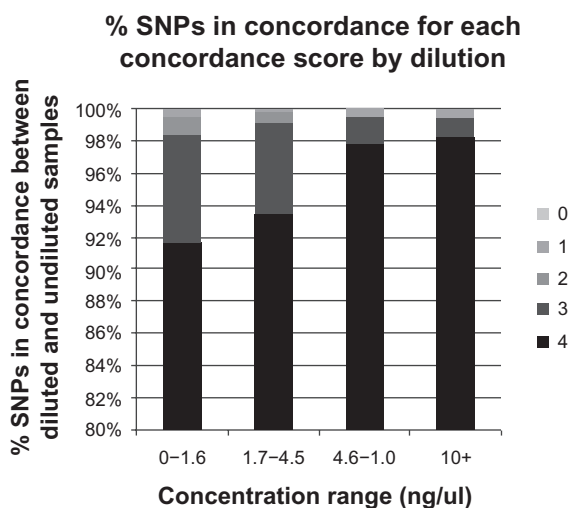
Discussion

With the burgeoning interest in GWAS (see www.genome.gov/gwastudies),[22] there is an increased need to minimize costs and maximize patient participation

% SNPs in concordance for each concordance score by dilution

Figure 2. Distribution of the percent of SNPs concordant between diluted and undiluted DNA from buccal cells for each dilution among four women. Each dilution is broken down into the percentage of SNPs that are concordant between diluted and undiluted samples in none, 1, 2, 3 or all (4) individuals (i.e. a concordance score of 0, 1, 2, 3 or 4). The lightest gray bar indicates the percent of SNPs concordant between diluted and undiluted samples in none of the individuals (barely visible), and the darkest gray bar indicates the percent of SNPs concordant between diluted and undiluted samples in all four of the individuals.

for DNA collection since large sample sizes are typically necessary to detect associations between genetic biomarkers and complex traits of interest. Our study demonstrates that it is possible to adequately perform genotyping on a high-density genotyping platform using stored DNA from buccal cells collected by mail. In this study, 561,481 SNPs were genotyped at mean GE and GC > 98%, even when DNA concentrations were as low as 5 ng/ul. Nonetheless, while mean GE was >99% for both undiluted buccal and blood samples (Table 1), GE was significantly higher with DNA from blood samples, though the difference was small (0.57%).

In addition to showing that buccal cell DNA can be used for genome-wide genotyping, we also provide evidence that very low concentrations of archived buccal cell DNA produces high quality genotyping results, with high rates of genotyping calls (e.g. mean GE at the 1:4 dilution, where the mean DNA concentration was 7.20 ± 3.22 ng/µl, was 99.32%). The high quality was also supported by the high concordance between the diluted and undiluted buccal samples and by the high concordance between diluted buccal and paired blood samples. The decline in GE and concordance below 5 ng/µl indicates that 5 ng/µl may be a reasonable lower limit for achieving at least 99% GE with the Illumina Infinium 660K genotyping platform. This has important implications for investigators with scarce or archived DNA samples, making it possible to maximize the usefulness of a limited sample (i.e. more information can be gained from a single sample with large-scale genotyping than from genotyping a finite panel of candidate SNPs).

Our in-depth analyses of the type and location of discordant SNPs showed that the discordances were predominantly of the "missing call" type as opposed

to the "different call" type. This most likely indicates that the Illumina platform has rigorous requirements for call quality, such that at lower concentrations, a SNP will not be called if there is doubt about the accuracy of the call. The small number of different calls relative to missing calls means that at low concentrations, data will more likely be lost rather than be false, thereby limiting misleading genotyping calls. Also, discordances did not cluster on any particular chromosome for concentration ranges >1.6 ng/μl, indicating that the decrease in genotype calling quality with decreased concentration occurred randomly across the genome, except at very low concentrations. Because concordance increases as concentration decreases, the clustering that occurred at very low concentrations is likely due to a greater number of discordant SNPs overall and chance that they occurred on specific chromosomes.

A limitation to our study is that the genotyping on buccal DNA was conducted only on the Illumina Infinium 660W-Quad platform. While other platforms may perform equally as well, further studies are needed before extending these data for use on other platforms. In addition, our sample size was relatively small, but despite this, we were still able to detect significant differences in GE between blood and buccal cell samples. Furthermore, we were able to detect trends in GE and concordance as a function of DNA concentration. Finally, buccal samples were stored for a mean of 4 years and we do not know what the effect of longer storage periods would be on GE and GC.

In conclusion, we have shown that buccal cell DNA is a viable alternative to blood DNA for genome-wide genotyping, and that low DNA concentrations can be used for genotype determination on high-throughput genome-wide scan platforms. Genotype efficiency and concordance does decline with DNA concentration on the Illumina platform, and genotyping on the Illumina platform may suffer at DNA concentrations below 5 ng/μl; therefore, we recommend using higher concentrations if available.

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest.

The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Le Marchand L, Lum-Jones A, Saltzman B, Visaya V, Nomura AM, Kolonel LN. Feasibility of collecting buccal cell DNA by mail in a cohort study. *Cancer Epidemiol Biomarkers Prev*. 2001 Jun;10(6):701–3.
2. Heath EM, Morken NW, Campbell KA, Tkach D, Boyd EA, Strom DA. Use of buccal cells collected in mouthwash as a source of DNA for clinical testing. *Arch Pathol Lab Med*. 2001 Jan;125(1):127–33.
3. Garcia-Closas M, Moore LE, Rabkin CS, et al. Quantitation of DNA in buccal cell samples collected in epidemiological studies. *Biomarkers*. 2006 Sep–Oct;11(5):472–9.
4. Rogers NL, Cole SA, Lan HC, Crossa A, Demerath EW. New saliva DNA collection method compared to buccal cell collection techniques for epidemiological studies. *Am J Hum Biol*. 2007 May–Jun;19(3):319–26.
5. King IB, Satia-Abouta J, Thornquist MD, et al. Buccal cell DNA yield, quality, and collection costs: comparison of methods for large-scale studies. *Cancer Epidemiol Biomarkers Prev*. 2002 Oct;11(10 Pt 1):1130–3.
6. Agalliu I, Schweitzer PA, Leanza SM, Burk RD, Rohan TE. Illumina DNA test panel-based genotyping of whole genome amplified-DNA extracted from hair samples: performance and agreement with genotyping results from genomic DNA from buccal cells. *Clin Chem Lab Med*. 2009;47(5):516–22.
7. Herraez DL, Stoneking M. High fractions of exogenous DNA in human buccal samples reduce the quality of large-scale genotyping. *Anal Biochem*. 2008 Dec 15;383(2):329–31.
8. Paynter RA, Skibola DR, Skibola CF, Buffler PA, Wiemels JL, Smith MT. Accuracy of multiplexed Illumina platform-based single-nucleotide polymorphism genotyping compared between genomic and whole genome amplified DNA collected from multiple sources. *Cancer Epidemiol Biomarkers Prev*. Dec 2006;15(12):2533–6.
9. Hong J, Leung E, Fraser A, Krissansen GW. Nucleic acid from saliva and salivary cells for noninvasive genotyping of Crohn's disease patients. *Genet Test*. Dec 2008;12(4):587–9.
10. Beckett SM, Laughton SJ, Pozza LD, et al. Buccal swabs and treated cards: methodological considerations for molecular epidemiologic studies examining pediatric populations. *Am J Epidemiol*. 2008 May 15;167(10):1260–7.
11. He H, Argiro L, Dessein H, Chevillard C. Improved technique that allows the performance of large-scale SNP genotyping on DNA immobilized by FTA technology. *Infect Genet Evol*. 2007 Jan;7(1):128–32.
12. Feigelson HS, Rodriguez C, Robertson AS, et al. Determinants of DNA yield and quality from buccal cell samples collected with mouthwash. *Cancer Epidemiol Biomarkers Prev*. 2001 Sep;10(9):1005–8.
13. Swanson JM, Moyzis RK, McGough JJ, et al. Effects of source of DNA on genotyping success rates and allele percentages in the Preschoolers with Attention-Deficit/Hyperactivity Disorder Treatment Study (PATS). *J Child Adolesc Psychopharmacol*. 2007 Oct;17(5):635–46.
14. Montgomery GW, Campbell MJ, Dickson P, et al. Estimation of the rate of SNP genotyping errors from DNA extracted from different tissues. *Twin Res Hum Genet*. 2005 Aug;8(4):346–52.

15. Feigelson HS, Rodriguez C, Welch R, et al. Successful genome-wide scan in paired blood and buccal samples. *Cancer Epidemiol Biomarkers Prev.* 2007 May;16(5):1023–5.

16. Woo JG, Sun G, Haverbusch M, et al. Quality assessment of buccal versus blood genomic DNA using the Affymetrix 500 K GeneChip. *BMC Genet.* 2007;8:79.

17. Belanger CF, Hennekens CH, Rosner B, Speizer FE. The nurses' health study. *Am J Nurs.* Jun 1978;78(6):1039–40.

18. Sabeti PC, Varilly P, Fry B, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature.* 2007 Oct 18; 449(7164):913–8.

19. Sexton DPM,JL, Giles JT, Bush WS, Bradford Y, Haines JL. The Whole-genome Association Study Pipeline (WASP): A Comprehensive Tool for Large-Scale Association Studies. *AJHG.* 57(413S).

20. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007 Sep;81(3):559–75.

21. Purcell S. PLINK v 1.07. 2009; http://pngu.mgh.harvard.edu/purcell/plink/

22. Catalog of Published Genome-Wide Association Studies. 2010; http://www.genome.gov/gwastudies/. Accessed 2/3/2010, 2009.