GENEVA Marazita Guatemala Dental Caries Project
Quality Control Report
University of Washington

University of Washington
October 20, 2011

Contents
1 Summary and recommendations for dbGaP users 4
2 Project overview 4
3 Genotyping process 6
4 Quality control process and participants 6
5 Sample and participant number and composition 6
6 Gender identity 6
7 Chromosomal anomalies 7
8 Relatedness 8
9 Population structure 8
10 Missing call rates 10
11 Plate effects 10
12 Duplicate discordance 10
13 Mendelian Errors 11
14 Hardy-Weinberg equilibrium 11
15 Sample exclusion and filtering 12
16 SNP filters 12
17 Confounding covariates 12
18 Preliminary association test results 12
19 URLs 14
A Sample filters

B Project participants
List of Tables

1. Summary of recommended SNP filters ........................................... 18
2. Summary of DNA samples and scans .................................................. 18
3. IBD kinship coefficient expected values ........................................... 18
4. SNP duplicate discordance error rates and counts ............................... 19
5. Covariate test results for self-identified Guatemalan subjects .............. 19

List of Figures

1. Gender check intensity plots ........................................................ 20
2. Example of nondisjunction ............................................................ 21
3. “BAleleFreq” (BAF) example ........................................................... 22
4. Normal scans of LogRRatio and BAlleleFreq of chromosome 1 ........... 23
5. Abnormal scans of LogRRatio and BAlleleFreq of chromosome X ........ 24
6. Abnormal scans of LogRRatio and BAlleleFreq of chromosome 6 ........ 25
7. Estimated identity-by-descent coefficients using all subjects ................ 26
8. PCA of 544 unrelated subjects, with HapMap controls ........................ 27
9. PCA of 544 unrelated subjects, without HapMap controls .................... 28
10. PCA direct and indirect test computation, set 1 results ...................... 29
11. PCA direct and indirect test computation, set 2 results ...................... 30
12. PCA results for all study samples, direct and indirect computation ........ 31
13. PCA results for U.S. subjects from two Dental Caries studies .............. 32
14. Histogram of missing call rate per sample ....................................... 33
15. Boxplots of missing call rate per sample ......................................... 34
16. Median autosomal missing call rate versus number of samples per batch 35
17. Allele frequency difference between each genotyping batch ................. 36
18. Duplicate sample discordance, within study .................................... 37
19. Duplicate sample discordance, cross study ..................................... 38
20. QQ plots of HWE p-values for homogeneous Guatemalan subjects ........ 39
21. Venn diagram of SNP filters .......................................................... 40
22. QQ plots of association test p-values for Guatemalan subjects ............. 41
23. Manhattan plots of association test p-values for Guatemalan subjects .... 42
24. Allelic intensity cluster plots for self-identified Guatemalan subjects ...... 43
25. Allelic intensity cluster plots for unrelated Guatemalan subjects .......... 44
1 Summary and recommendations for dbGaP users

A total of 1,033 dental caries study subjects consisting of families and unrelated individuals were genotyped on the Illumina Human610 Quadv1_B array. The median call rate is 100% and the error rate estimated from 23 pairs of study sample duplicates, including 3 pairs of MZ twins, is $3.5 \times 10^{-6}$.

Genotypic data are provided for all subjects and SNPs. However, we recommend selective filtering of genotypic data prior to analysis to remove sample-chromosome combinations with chromosomal anomalies and/or missing call rate > 5% and to remove whole samples with an overall missing call rate > 2%. The recommended filters are provided (Appendix A) along with code to produce filtered PLINK data files. Preliminary association test results are provided as an example of how to apply the filters. All SNPs are included in the association test results file, but we recommend that these be filtered according to the criteria specified in Table 1. A composite SNP filter is provided, along with each of the component criteria so that the user may vary thresholds. Additional specific recommendations are highlighted in the following document in italics.

2 Project overview

Dental caries (also known as tooth decay) remains the most common chronic disease of childhood, five times more common than asthma and seven times more common than environmental allergies, with more than 40% of children exhibiting caries when they enter kindergarten. In 2005, it was estimated that dental health care costs were approximately $84 billion, of which 60% or about $50 billion were related to treatment of dental caries. The etiology of dental caries has been studied for many years. Multiple factors contribute to a person’s risk for caries, including: 1) environmental factors such as diet, oral hygiene, fluoride exposure and the level of colonization of cariogenic bacteria and 2) host factors such as salivary flow, salivary buffering capacity, position of teeth relative to each other, surface characteristics of tooth enamel and depth of occlusal fissures on posterior teeth. In spite of all that is known about this disease, there are still individuals who appear to be more susceptible to caries and those who are extremely resistant, regardless of the environmental risk factors to which they are exposed, implying that genetic factors also play an important role in caries etiology. This conclusion is supported by studies in both humans and animals, with the most compelling evidence coming from studies of twins reared apart in which investigators found significant resemblance within monozygotic (MZ) but not dizygotic (DZ) twin pairs for percentage of teeth and surfaces restored or carious and estimated the genetic contribution to caries as 40%. Other recent studies of twins reared together estimated the heritability for caries, adjusted for age and gender, as ranging from 45-64%. In our study populations of families, we also estimated caries heritability as approximately 54%-70% of variation in primary dentition caries scores and 35%-55% in the permanent dentition [1]. The first GWAS of dental caries in the primary dentition data (from our GENEVA Dental Caries GWAS parent project) has been recently published [2].

Orofacial clefts (OFCs), particularly cleft lip with or without cleft palate (CL/P) and isolated cleft palate (CP) are a major public health problem, affecting one in every 500-1000 births worldwide thus representing the most common facial birth defect and one of the most common of all congenital anomalies. CL/P is a major structural birth defect that is notable for significant lifelong morbidity and complex etiology. The extensive psychological, surgical, speech and dental involvement emphasize the importance of understanding the underlying causes of CL/P. Therefore, many research groups have attempted to elucidate the etiology of CL/P, with some recent success by our research group and others [3, 4, 5]. It is clear that CL/P can occur as part of Mendelian syndromes, that certain chromosomal abnormalities include CL/P in the phenotype, and that certain teratogens can increase the risk of having an offspring with CL/P. However, phenotypes of known etiology comprise only a small portion of all individuals with a CL/P or CP, and the major focus of research into OFCs is to develop an understanding of the etiology of nonsyndromic (NS) forms of clefting. A major focus of the University of Pittsburgh CCDG has been to study additional phenotypes within nonsyndromic OFC
families in order to identify sub-clinical expressions of OFC risk genes or risk variants, e.g. SNPs [6]. A
detailed oral exam is conducted as part of these extended phenotypic studies, including a dental caries
exam.

This study is part of the Gene Environment Association Studies initiative (GENEVA, http://www.
genevastudy.org), which was developed through the trans-NIH Genes, Environment, and Health Ini-
tiative (GEI). Furthermore this study brings together multiple research priorities of the University of
Pittsburgh Center for Craniofacial and Dental Genetics (www.ccdg.pitt.edu). A genome-wide panel of
610,000 SNPs was genotyped at the Broad Institute to be comparable to our other pertinent GENEVA
studies that are also part of dbGaP (dbGaP accession number phs000095.v1.p1, “Dental Caries: Whole
Genome Association and Gene x Environment Studies” and dbGaP accession number phs000094.v1.p1,
“International Consortium to Identify Genes and Interactions Controlling Oral Clefts”). The goal of
this study is to investigate genetic determinants to dental caries and to OFCs in a novel study pop-
ulation. To date, most genetic studies of dental caries have been conducted in Whites, and of OFCs
in Whites and Asians. The Guatemalan population under study is rural and ethnically mixed with a
high proportion of native-south-american.

The Guatemalan population under study here is part of a large international study conducted by the
University of Pittsburgh Center for Craniofacial and Dental Genetics (CCDG, www.ccdg.pitt.edu). This
study is known as the Pittsburgh Orofacial Cleft Study (POFC), and has data collection sites in several
countries (Guatemala, Hungary, Brazil, Argentina, China, the Philippines) the U.S.A. (Pittsburgh,
Texas, Iowa). The goal of POFC is to investigate etiologic factors in OFCs, particularly nonsyndromic
CL/P. The major focus of the study is to understand genetic components contributing to risk of
CL/P and to understand the underlying heterogeneity in this group of disorders. As a tool for better
understanding CL/P, POFC is conducting extensive phenotyping within families with one or more
individuals affected with CL/P, and in control families with no known family history of CL/P. The
phenotypes under investigation include a variety of sub-clinical phenotypes that are seen in the general
population (and in our study control families) but are seen at an increased frequency in relatives of
individuals with CL/P compared to controls [6, 7]. Among these sub-clinical phenotypes are several
oral measurements, including assessment of dental caries status. Thus, this GWAS allows studies of
CL/P and dental caries in the same study population.

The Guatemalan families under study have been investigated for some candidate genes for dental
caries [8] and for several candidate genes for CL/P (e.g. [7, 9]). Note that although there are some
reports in the literature of higher caries experience in individuals with clefts, the most recent meta-
analysis of those literature reports concluded that individuals born with these defects do not have a
higher frequency of caries [10]. Notably, we investigated the association of CL/P and caries in three
of our study populations (including part of the Guatemalan population in this GWAS study) and also
found no increase in caries rates of CL/P cases versus controls [3, 4].

Thirty-six of the Guatemalan subjects were also enrolled in an oral clefts study (dbGaP Study Ac-
cession: phs000094.v1.p1); they are annotated as such in “Sample_annotation.csv”. In addition to the
families ascertained in Guatemala, 98 individuals recruited to augment the data in the parent GENEVA
study (GWAS of dental caries, dbGaP accession number phs000095.v1.p1), in particular individuals
from the IOWA and PITT GENEVA study sites in the U.S., were also included. Since the Guatemalan
subjects form a new study focused on both dental caries and CL/P, while the U.S. subjects were re-
cruited to supplement the parent dental caries study (dbGaP accession phs000095.v1.p1), sample-level
data, such as phenotypes, raw genotype files, and sample annotations, are separated into two portions,
Guatemalan and U.S. The Guatemalan portion is posted as a separate study, and the US portion to
the parent study. SNP-level data, such as SNP annotations, are shared by all samples regardless of
U.S. or Guatemalan; therefore, they are posted to both studies. This QC report is also posted to both
studies.
3 Genotyping process

Tissue samples for DNA extraction were collected from each subject. Two sample types were used for the study subjects: 1.) Blood (38.3%); and 2.) Saliva (58.5%). DNA samples were sent to the Broad Institute Biological Samples Platform (Broad) for QC, and then plated in 96-well plates for genotyping by the Broad Institute Genetic Analysis Platform using the Illumina Human610_Quadv1_B array. Data were analyzed using the Illumina BeadStudio software (framework version 3.1.3.0) and GenTrain genotyping module (version 3.2.32).

The Broad Institute considered a SNP as a technical failure if any of the following criteria were met: 1) Call rate over scans $< 97\%$; 2) Number of replicate sample genotype discordances $> 2$; 3) Number of sample trio inheritance errors $> 1$; 4) BeadStudio GenTrain score $< 0.6$; and 5) Cluster separation $< 0.4$. Manual review of sex chromosomal and mitochondrial SNPs also took place.

4 Quality control process and participants

Genotypic data that passed initial quality control at the Broad were released to the GENEVA Coordinating Center (CC), the NCBI dbGaP team and the Dental Caries project team. These data were further analyzed by all four groups and the results were discussed in weekly conference calls, which also included NHGRI personnel. Key participants in this process and their institutional affiliations are given in Appendix B. Analysis tools varied by group, but include primarily PLINK [11] and the R statistical package [12]. If not otherwise noted, analyses described below were done using R and the “ncdf” library to access data stored in netCDF files (see URL below).

5 Sample and participant number and composition

A total of 1,116 DNA samples from study subjects were put into genotyping production, of which 1,064 passed the Broad’s QC process (Table 2). The subsequent QA process identified 8 scans that will not be included in the dataset posted on dbGaP because of sample contamination or questionable identity (unexpected relatedness and gender error).

The set of scans to be posted include 1,056 from the Dental Caries study and 13 HapMap controls. Among the 1,056 study samples, 955 derive from 940 Guatemalan subjects and 101 from 96 U.S. subjects.

Among the 940 Guatemalan subjects, 15 were genotyped twice (intended duplicates); there are also 3 pairs of MZ twins. The self-identified race is Guatemalan for all but one subject, who is self-identified as African. The 940 genotyped Guatemalan subjects occur as 138 singletons, and 802 in families consisting of 223 assorted duos, trios, other nuclear and extended families ranging in size from 2 to 14. The sex composition is 42% male and 58% female.

Among the 96 U.S. subjects, all of which are singletons, 5 were genotyped twice (intended duplicates). The sex composition is 58% male and 42% female. The self-identified race is primarily European American (91.67%), with the next largest group as African American (7.29%).

The 13 HapMap control scans derive from 6 subjects, of which all are replicated two or more times. The HapMap controls include one trio of CEU and one trio of YRI.

6 Gender identity

To check gender identity, we look at both X chromosome heterozygosity and the means of the intensities of SNP probes on the X and Y chromosomes. The expectation is that male and female samples will fall into distinct clusters that differ markedly in X and Y intensities. Figure 1a shows the results for all samples prior to exclusions based on questionable identity. There are two distinct clusters as expected, with four gender discrepancies. Two U.S. samples, one male and one female, appear to be
misidentified, and were dropped from the study due to unresolved identity. Two Guatemalan samples, one male and one female, were also mis-annotated. The error was caused by sample switch and was corrected accordingly.

Figures 1b, 1c, and 1d show the results for all samples after the four discrepancies were handled, and demonstrate gender identities that are consistent with the X and Y chromosome intensity and heterozygosity. There are several samples that appear to be outliers with respect to their annotated gender. One male shows a Y-chromosome intensity typical of a male and an X intensity typical of a female and are very likely to have an XXY karyotype. One female shows unusually low X intensity and appear to be XX/XO mosaic. Several females have higher than usual X intensity, but only one appears to be XXX. Karyotype is determined from BAlleleFreq plots discussed below.

7 Chromosomal anomalies

Gross chromosomal anomalies (such as aneuploidy and large insertion/deletion events) were detected by analyzing relative intensity (“LogRRatio”) and a measure of allelic imbalance (“BAlleleFreq”) [13]. BAlleleFreq (BAF) is a transformation of the polar coordinate angle of the intensities of the two SNP alleles. Figure 2 shows that the occurrence of trisomic cells (or a mixture of disomic and monosomic cells) results in different positions for heterozygotes at different loci [13] describe a transformation of the polar coordinate angle Θ, which they call “BAlleleFreq”. This value is one of the metrics output by the Illumina GenomeStudio software. Figure 3 shows that BAF is a transformation that standardizes the positions of the three diploid genotypic classes to 0 (AA), 0.5 (AB) and 1.0 (BB).

To identify aneuploid or mosaic samples systematically, we used two methods. For anomalies that split the intermediate BAF band into two components, we used Circular Binary Segmentation (CBS) [14] on BAF values for SNPs not called as homozygotes. For heterozygous deletions (with loss of the intermediate BAF band), we identified runs of homozygosity accompanied by a decrease in LRR. For anomalies greater than 10 Mb in length, we estimated the absolute value of the median BAF deviation from the expected value of 0.5 and selected those with a deviation greater than 0.1. Genotypes for SNPs in such anomalies are error-prone. Among the 1,036 study participants, three were found to have such an anomaly on the autosomes and two on the sex chromosomes.

Figure 4 shows a normal scan of BAF for chromosome 1 in sample A, where each point represents a different SNP locus in one sample. The scan has bands at approximately 0, 0.5 and 1. The bands at 0 and 1 represent homozygotes and the band at 0.5 represents heterozygotes. Figure 5 shows LRR and BAF scans of the X chromosome in sample A (i.e. the same sample as in Figure 4). In this case, the intermediate band is split for the entirety of the q arm as shown in the BAF scan, with an increase in intensity of probes (higher LRR value) on the q arm relative to the p arm and other chromosomes for sample A. In many cases it is not clear whether detected anomalies are constitutional or somatic mosaics, but in this case it is likely that the anomaly is constitutional because the two intermediate bands occur at the one-third and two-thirds positions expected for a constitutional trisomic.

Figure 6 shows an abnormal scan of chromosome 6 in sample B where only the distal part of the q arm is affected. This anomaly is accompanied by an increase in LogRRatio (intensity) in the same region, suggesting a duplication.

We recommend filtering out the genotypes for all SNPs in sample-chromosome combinations with chromosomal anomalies totaling greater than 10 Mb in length. The indicators are provided in the file “jcmat.anom.miss.05.csv” on dbGaP. See also Appendix A.

In addition to finding distinct anomalies detected by a split in the intermediate band of BAF scans, we also flag each chromosome in each sample with one or more windows having a BAF SD greater than 4 SD from the mean. Samples with all chromosomes flagged may be a mixture of DNAs from more than one participant and may have more than 3 bands in BAlleleFreq scans of all chromosomes. BAF plots for each of these samples are reviewed. Four apparent mixtures were detected for this project and were deleted from the dbGaP posting.
8 Relatedness

The relatedness between each pair of subjects was evaluated by estimation of three coefficients corresponding to the probability that zero \((k_0)\), one \((k_1)\) or two \((k_2)\) pairs of alleles are identical by descent (IBD). The kinship coefficient \((KC)\) for a pair of participants is

\[
KC = \frac{1}{2} k_2 + \frac{1}{4} k_1
\]  

(1)

Table 3 shows the expected coefficients for some common relationships. Any two alleles at a locus are either identical by descent or not and this gives rise to variation of actual identity around the expected values. When markers over the entire genome are used to estimate the kinship coefficient, there is a need to take into account the dependencies among markers due to linkage. Expressions for the variance in a summary measure of actual identity have been given in the past \([15, 16, 17, 18]\) and have been extended by the Coordinating Center to the three IBD coefficients.

For this study, the IBD coefficients were estimated using 118,625 autosomal SNPs and the method of moments procedure used by PLINK \([11]\), but implemented in the R package “SNPRelate”. The SNPs were selected by removing those with p-value < 0.01 for association with any of the first three principal components, which were obtained with Principal Component Analysis (PCA) performed using 562 samples annotated as unrelated. This SNP selection procedure was used to avoid confounding population structure with family structure in the self-identified Guatemalans.

The IBD coefficients estimates identified one subject with unexpected relatedness that could not be resolved, and the data for this subject have been removed from the dataset to be posted on dbGaP.

Furthermore, 127 changes were made to the pedigree, along with sample relabeling to correct genotypic and phenotypic data mismatch for 2 pairs of samples.

After these corrections, all of the IBD estimates become reasonably consistent with expectation, as shown in Figure 7, which displays only the pairs of participants with a \(KC > 1/32\), i.e., half the expected value for first cousins. A few notables include a half-sibling/avuncular/grandparent-grandchild (half-sibling-like) pair that falls close to the expected values for full siblings, and a few full-sibling and half-sibling-like pairs that are a little far away from their respective cluster. However, other relationships involving those individuals are consistent, so these discrepancies are most likely due to sampling error. Several unrelated pairs are in the half-sibling-like cluster. We were unable to specify a pedigree structure for subjects with such relationships. In such cases, we define superfamilies to include core families in the pedigree, which are connected by pairs of individuals with half-sibling-like relationships. See “MarazitaG_Pedigree_Guatemala.csv” and “IBD_estimates.csv”.

For an analysis that assumes all participants are unrelated, we recommend selecting the maximum unrelated subjects from each family or superfamily unit. See Appendix A.

9 Population structure

We used principal component analysis (PCA), essentially as described by Patterson et al. (2006) \([19]\), but implemented in R package “SNPRelate”. To select SNPs for PCA, we started from a pool of autosomal SNPs with missing call rate < 5% and minor allele frequency > 5% across all study subjects. There are 490,120 such SNPs. Then we performed linkage disequilibrium pruning (using R package “SNPRelate”) by recursively removing SNPs within a sliding window of 5Mb so that no pairs had a genotypic correlation \(r > 0.45\). The resulting 113,686 SNPs were used to generate the principal components.

Figure 8 shows a plot of the first two eigenvectors from an analysis of 544 unrelated study subjects, along with 1,201 unrelated HapMap III controls (CEU, YRI, CHB, JPT, CHD, MKK, ASW, GIH, MEX, TSI, and LWK). The first eigenvector, accounting for 8.12% of the variance, separates the self-identified Guatemalan and White subjects from the self-identified African and African American subjects. The second eigenvector, accounting for 3.67% of the variance, separates the self-identified Guatemalan from...
the White and Asian subjects. We also generated principal components from an analysis of the 544 unrelated study subjects without the HapMap controls (Figure 9). The resulting population structure is similar to what is in Figure 8.

From the PCA results in Figure 9, we defined a homogeneous set of samples with PCA-defined Guatemalan ancestry. Samples that fell outside two standard deviations from the mean for eigenvectors one and two were excluded. This resulted in a set of 418 unrelated study samples of PCA-defined Guatemalan ancestry. This sample set was used for the Hardy-Weinberg Equilibrium (HWE) tests as explained in Section 14. We ran HWE with this sample set in order to reduce or eliminate population structure so that the HWE test will detect mainly genotyping artifacts, because the HWE test can be affected by population structure and by inclusion of related individuals.

For the preliminary association test, we include the first few eigenvectors and those eigenvectors that are significantly associated with the outcome as part of the adjustment. When PCA is performed naively on a set of individuals that include many related subjects as are in this study, resulting eigenvectors may not reflect population structure in the study data accurately. One possible solution to this problem is described by Zhu et al. [20]. They suggested calculating both sample and SNP eigenvectors for a subset of unrelated subjects and then using the resulting SNP eigenvectors, along with the genotype calls for the remaining subjects, to estimate their sample eigenvectors. We have implemented this approach for this study. In the following, we use the term “direct estimation” to refer to the usual method of calculating sample eigenvectors directly from the genotype calls alone and we use the term “indirect estimation” to refer to calculating sample eigenvectors from genotype calls and SNP eigenvectors that were calculated from related subjects.

Before applying the Zhu method to the study data, we first conducted an exercise to validate our implementation of the method for two sets of individuals: (1) Set 1: all study subjects; and (2) Set 2: the self-identified Guatemalan subjects only. For each of the two sets, we took two members from each of the study families and randomly assigned one member to subset A and one member to subset B. We calculated PCA results directly for subset A (as described above), then used the Zhu method to indirectly calculate sample eigenvectors for the members of subset B. Then, we did the reverse: calculated PCA results directly for subset B and indirectly for subset A.

Plots comparing the first two direct and indirect eigenvectors for Set 1 are shown in Figure 10. Plots comparing the first two direct and indirect eigenvectors for Set 2 are shown in Figure 11.

For Set 1, both eigenvectors one and two show a high correlation between direct and indirect calculation, indicating that the inferred eigenvectors for related samples will be useful when adjusting for population structure in association tests. For Set 2, however, eigenvector 2 does not correlate well between direct and indirect calculation. This is not too surprising since eigenvector 2 only accounts for 0.37% of the total variance of this data set, and can be affected by minor sampling errors existing between subsets A and B. It has been observed that eigenvalues for the eigenvectors accounting for little variance can switch order. Therefore, it is safe to conclude that the inferred eigenvectors for related samples are useful when adjusting for population structure in association tests.

The preliminary association tests were performed on self-identified Guatemalan subjects only. For the set of 544 unrelated, self-identified Guatemalan samples, we estimated the sample eigenvectors directly; then we used the directly calculated results for the unrelated to infer the eigenvectors for the remaining family members. Figures 12 shows the combined direct and inferred results, which are very similar to those in Figure 9.

The first 8 eigenvectors for the self-identified Guatemalan subjects and unrelated self-identified Guatemalan subjects are provided as sample-by-eigenvector matrices on dbGaP and labeled as “Principal_components_gua.csv” and “Principal_components_unrel_gua.csv”, respectively. The sample sets are identified by the row labels in this file. The SNP set is identified by the variables in the SNP table (“SNP_analysis_results.csv”) with variable names “pca.set”.

Since the U.S. subjects in this study are going to be posted to the previous Dental Caries study, we carried out cross-study checks on the two sets to ensure that the genotyping is accurate, and the two sets have the same population structure as we had anticipated. This was achieved by: (1) Calculating
sample discordance rate for overlapping samples (see Section 12); and (2) Combining the two sets and performing PCA on the combined data, the results of which are described next.

Figure 13 shows a plot of the first two eigenvectors from an analysis of 97 unrelated subjects from the U.S. cohort in this study, along with 2,248 unrelated subjects from the previous study. As expected, they appear like two samples drawn from the same population.

10 Missing call rates

Two missing call rates were calculated for each sample and for each SNP in the following way (and provided in files “SNP_analysis_results.csv” and “Sample_analysis_results.csv” on dbGaP). (1) “missing.n1” is the missing call rate per SNP over all samples (including HapMap controls). (2) “missing.e1” is the missing call rate per sample for all SNPs with missing.n1 < 100%. (3) “missing.n2” is the missing call rate per SNP over all samples with “missing.e1” < 5%. In this project, there were no samples with missing.e1 5%, thus, there are neither missing.n2 nor missing:e2 values, as they are the same as “missing.n1” and “missing.e1”.

Figure 14 shows the distribution of missing.e1 for all study samples. The median value is 0.06%, the 95th percentile is 0.28% and the maximum missing.e1 value is 2.99%.

The Illumina Human610_Quadv1_B array contains assays for 592,532 SNPs (genotypes and intensity) and 28,369 probes (intensity only). Prior to data release, the Broad QC process failed 18,886 SNPs with call rate < 97% and other criteria (Section 3).

Among the remaining 573,646 SNPs, the median missing call rate is 0% and the 95th percentile is 1%.

Figure 15 shows the distribution of missing.e1, categorized by the type of biological specimen from which the DNA was extracted. All samples have a rather low missing call.

For analysis, we recommend filtering out samples with a missing call rate ≥ 2%.

11 Plate effects

The samples were processed together in batches consisting of complete or partial 96-well plates. Figure 16 shows the mean missing call rate per batch versus the number of passing samples per batch. The batch effect on missing call rate is not significant (p-value = 0.06); besides, all batches have a low mean missing rate.

Another way to detect genotyping plate effects is to assess the difference in allelic frequencies between each plate and a pool of the other plates. We calculated a 1 d.f. Chi-squared test statistic for each SNP and each plate and then averaged these statistics over SNPs. This statistic is a measure of how different each plate is from the other plates. It can be affected not only by laboratory processing, but also by the biological characteristics of the samples within the plate. The characteristic most likely to affect the distinctiveness of a plate is its ethnic composition relative to the mean composition across plates. Figure 17 shows a plot of genotyping plate composition versus the mean χ² statistic, where plate composition is measured as the fraction of samples in the plate that are unrelated, and self-identified as “Guatemalan”. No plates appear to be problematic with regard to the allelic frequency test.

12 Duplicate discordance

A total of 20 pairs of intended study duplicates and 3 pairs of MZ twins were genotyped and the median discordance is 7e−6.

Figure 18 shows the ranked duplicate discordance rates for the 20 pairs of study subjects, along with 3 pairs of MZ twins. The duplicate discordance rate is a very useful measure of genotyping accuracy because it can be utilized to estimate genotyping error rates. The genotype at any SNP may be called
correctly, or miscalled as either of the other two genotypes. If $\alpha$ and $\beta$ are the two error rates, the probability that duplicate genotyping instances of the same participant give a discordant genotype is $2(1-\alpha-\beta)(\alpha+\beta+\alpha\beta)$. When $\alpha$ and $\beta$ are very small, this is approximately $2(\alpha+\beta)$ or twice the total error rate. Potentially, each true genotype has different error rates (i.e. three $\alpha$ and three $\beta$ parameters), but here we assume they are the same. In this case, since the overall discordance rate is about $7e^{-6}$, a rough estimate of the mean error rate is $3.5e^{-6}$ errors per SNP per sample, indicating a very high level of reproducibility.

Duplicate discordance estimates for individual SNPs can be used as a SNP quality filter. The challenge here is to find a level of discordance that would eliminate a large fraction of SNPs with high error rates, while retaining a large fraction with low error rates. The probability of observing $>x$ discordant genotypes in a total of $n$ pairs of duplicates can be calculated using the binomial distribution. Table 4 shows these probabilities for $x = 0$ to 7 and $n = 23$. Here we chose $n = 23$ to correspond to the number of pairs of high quality duplicates in which both members of the pair have a missing call rate $<2\%$.

We recommend a filter threshold of $>0$ discordant calls because this retains nearly all SNPs with error rates $<3.5e^{-4}$, while eliminating 15% of those with an unacceptably high error rate of $3.5e^{-3}$. Although we prefer to eliminate a higher percentage of SNPs at the error rate of $3.5e^{-3}$, it is difficult to do so with such a small number of duplicated samples.

As mentioned in Section 9, we also calculated sample discordance rate for overlapping samples between the two Dental Caries studies. Three HapMap subjects were genotyped in both Dental Caries studies; each subject was genotyped at least twice within a study, yielding 21 pairs of cross-study scans. The median discordance rate is $3.7e^{-5}$, reflecting high genotyping reproducibility. Figure 19 shows the ranked duplicate discordance rates for the 21 pairs of samples.

As indicated in Section 2, thirty-six study subjects, along with six HapMap controls, were genotyped in both this Dental Caries study and the oral clefts study, yielding 92 pairs of cross-study scans. The median discordance rate is $6.08e^{-5}$, again confirming high quality of the data.

13 Mendelian Errors

One of the SNP filters is based upon Mendelian errors. The large number of parent-offspring pairs and trios in this study provides an unusual opportunity to detect Mendelian errors and to use these data for quality filtering of SNPs. A total of 2,124 SNPs have at least five errors, detected using 428 unique parent-offspring trios and duos. Cluster plots show that SNP assays with more than four Mendelian errors usually have either poorly separated clusters or more than three clusters. The number of Mendelian errors per SNP is a useful quality filter, although establishing a suitable threshold for the filter is difficult. The threshold recommended is rather subjective but is based upon viewing cluster plots for a large number of SNPs with varying numbers of errors. We recommend using a threshold of $>4$ Mendelian errors.

14 Hardy-Weinberg equilibrium

We calculated an exact test of Hardy-Weinberg equilibrium (HWE) using a set of PCA-defined unrelated Guatemalan subjects ($n = 417$) (see Section 9) with $missing.e1 < 2\%$. Guatemalan subjects were defined as those within 2 SD of the mean of eigenvector 1 and within 2 SD of the mean of eigenvector 2 for self-identified “Guatemalan” subjects. Figure 20 shows QQ plots for autosomal and X chromosome SNPs. Autosomal SNPs deviate from expectation at about 0.1, whereas X chromosome SNPs show deviation at slightly lower p-values. Although QQ plots show deviation of observed from expected p-values for autosomal SNPs at about 0.1, we suggest using a filter threshold of $p = 0.0001$ because examination of cluster plots reveals good plots for many assays with $p$-values $>0.0001$. The threshold is rather subjective, but we are reluctant to recommend a higher threshold that would eliminate many good SNP assays.
15 Sample exclusion and filtering

As noted previously, a total of 1,116 DNA samples from study subjects were put into genotyping production, of which 1,064 were successfully genotyped and passed the Broad’s QC process (Table 2). Our QA process identified 8 scans of questionable identity or sample quality that will not be included in the dataset posted on dbGaP. Therefore, a total of 1,069 scans of study and HapMap control samples will be posted on dbGaP.

For association testing, we recommend filtering out SNPs for each sample-by-chromosome combination with a chromosomal anomaly (as described in Section 7) and/or with a missing call rate per chromosome \( \geq 5\% \). We also generally recommend filtering out samples with an overall missing call rate \( \geq 2\% \). For specific analyses, such as Hardy-Weinberg testing, additional filters are suggested, such as including only one ancestry group. To facilitate the application of these filter suggestions, we provide one sample-by-chromosome matrix ("jcmat.anom.miss.05.csv") and whole sample filter vectors on dbGaP (see Appendix A).

16 SNP filters

Table 1 summarizes a sequence of SNP failure criteria applied by the Broad’s QC process prior to data release and a set of additional filters suggested for removing assays of low quality or informativeness.

The suggested composite quality filter is a logical vector in the file “SNP_analysis_results.csv”, which also has the individual quality metrics so that the user can apply alternative thresholds. The recommended filters remove 4.7% of the SNP assays attempted. Figure 21 shows the overlap of SNPs filtered out when using 6 of the recommended filters. There is relatively little overlap, indicating that different metrics detect different kinds of problems in the SNP assays.

In addition to the composite quality filter, we also suggest applying a minor allele frequency filter for viewing association test results. The frequency criterion depends upon power considerations, but is usually on the order of 1-5%. In Table 1 we choose a 5% MAF filter due to the small sample size of this study.

Regardless of what filters are applied to association test results, it is highly recommended to view SNP cluster plots for any SNPs of interest.

17 Confounding covariates

Associations between the outcome and experimental factors that affect the quality of genotyping may lead to spurious results. In this project, the response variable TOT_D2MFT is not significantly associated with log10 of autosomal missing call rate for the self-identified Guatemalan subset (p-value = 0.135).

For the complete set, however, the association between missing call rate and the outcome is significant (p-value = 0.004), suggesting that allelic frequency differences are potentially biased by non-random missingness. The association can be accounted for by age at time of examination (“age”) when TOT_D2MFT is regressed on both “age” and log10 of missing call rate. “Age” is significantly associated with log10 of autosomal missing call rate as well as TOT_D2MFT in two separate linear regressions. Therefore, we suggest that “age” be used for adjustment in the linear regression of TOT_D2MFT for the complete set.

18 Preliminary association test results

Linear regression models were used to obtain preliminary association test results. Only self-identified Guatemalan subjects were selected for the tests. We performed regression on two data sets, one consisting of the unrelated only, and the other all self-identified Guatemalan subjects. In the analysis
of the unrelated subjects, samples are independent and p-values should be unbiased, but the analysis is highly under-powered. In the analysis of the all-Guatemalan set, the inclusion of related subjects may lead to genomic inflation. Eigenvectors estimated directly were included for adjustment for the unrelated subjects as usual; for the remaining subjects, eigenvectors obtained indirectly with Zhu method were included for adjustment (see Section 9).

To determine which covariates to use in the analysis, we first performed linear regression of TOT_D2MFT on each of a set of potential covariates including sex, age, $age^2$, and principal components 1-8. The resulting p-values can be found in Table 5. The fourth principal component is the only one, among all eight, that is associated with TOT_D2MFT. Nevertheless, all first four principal components were included in the association tests. Only one linear regression model was used for the preliminary association tests:

$$TOT_D2MFT \sim age + age^2 + sex + ev1 + ev2 + ev3 + ev4 + SNP.$$  

SNP and covariate effects were assessed with likelihood ratio tests. Sample filtering was performed using the filters described in Section 15. TOT_D2MFT was analyzed with the given covariates and the genotype of one SNP coded as 0, 1 and 2 (additive genotypic model). In performing association tests for X-linked SNPs, male genotypes were coded as 0 and 2 (for BY and AY), whereas female genotypes were coded as 0, 1, and 2 (for BB, BA and AA). This coding seems appropriate to reflect the fact that, with X inactivation in females, the number of active alleles in homozygous females equals that in hemizygous males.

Figure 22 shows the QQ plots of the p-values. The plots on the upper row are for all the self-identified Guatemalan samples, and those on the lower row the “unrelated”. On each row, the plot on the left was made before SNPs were filtered for quality, the one in the middle had the composite SNP quality filter applied but not the MAF filter (see Table 1), while the right plot was made after both the composite SNP quality filter and the MAF filter were applied. As the MAF filter filters out SNPs with low MAF and high p-values, genomic inflation increases slightly. The tests for all subjects (including relatives) show slightly more genomic inflation that those for just the unrelated.

Manhattan plots of the p-values are shown in Figure 23, all created after the composite SNP quality filter (without the MAF filter) was applied.

None of the regressions and filtering produced genome-wide significant association between TOT_D2MFT and SNPs.

Allelic intensity cluster plots of the nine SNPs with the lowest p-values (after applying the recommended SNP quality filter) were generated for both sample sets. All 24 plots are provided in files “guat_model2_24_lowest_p_filt.pdf” and “guat_unrel_model2_24_lowest_p_filt.pdf”. All plots show well-separated clusters without obvious artifacts.
19 URLs

netCDF data files: http://www.unidata.ucar.edu/software/netcdf/

Illumina data files: http://www.illumina.com/

References


Appendices

A Sample filters

For Hardy-Weinberg and association testing, we recommend filtering genotypic data using a sample-chromosome filter matrix ("jcmat.anom.miss.05.csv") and whole-sample filters.

The file “jcmat.anom.miss.05.csv” is a sample-by-chromosome matrix with 1,039 samples and 26 chromosomes (1-22, X, pseudo-autosomal, Y and mitochondrial). Each element of the matrix is a TRUE or FALSE value indicating whether to filter out the SNP genotypes in a particular sample-chromosome combination. The elements of this matrix are FALSE when either the sample-chromosome combination has a chromosomal anomaly > 10 Mb and/or when it has a missing call rate > 5%. PLINK-formatted genotypic data files are provided before and after the application of this filter matrix. We recommend using the filtered version for all analyses (labeled as “zeroed-out”). In addition, we recommend whole-sample filters for each type of analysis. Each whole-sample filter is provided as a logical vector indicating inclusion (TRUE) or exclusion (FALSE) of the 1,039 subjects, including study subjects and HapMap controls. These logical vectors are described below and are given in the “Sample_analysis_results.csv” file on dbGaP. In addition, the “keep” series text files provide sample identifiers for use in selecting samples for analysis in PLINK.

The components used to construct these whole-sample vectors are “geno.cntl” (distinguishing study subjects from HapMap controls), “unrel.gua” (indicating maximal unrelated self-identified Guatemalan subjects), “missing.e1” (missing call rate), “pca.gua” (PCA-defined Guatemalan subjects), “guats” (self-identified Guatemalan subjects). They are included in the “Sample_analysis_results.csv” file.

The vector “hwe.gua” and the file “hwe.gua.keep.txt” select unrelated Guatemalan subjects that are PCA-defined with “geno.cntl”= 0, “pca.gua”= TRUE, and “missing.e1”< 0.02. This filter was used for the HWE testing of the homogeneous Guatemalan subjects.

The vector “assoc.gua.keep.txt” selects all self-identified Guatemalan subjects, using “subj.plink”= TRUE, “geno.cntl”= 0, “missing.e1”< 0.02, and “guats”= TRUE. This filter was used for the preliminary association testing of all the self-identified Guatemalans.

The vector “assoc.unrel.gua.keep.txt” selects all unrelated, self-identified Guatemalan subjects, using “geno.cntl”= 0, “missing.e1”< 0.02, and “unrel.gua”= TRUE. This filter was used for the preliminary association testing of the unrelated, self-identified Guatemalans.

B Project participants

University of Pittsburgh
Rebecca DeSensi, Eleanor Feingold, Mary Marazita, Toby McHenry, John Shaffer, Xiaojing Wang

Center for Inherited Disease Research, Johns Hopkins University
Andrew Crenshaw, Daniel Mirel

dbGaP, NCBI
Mike Feolo, Justin Paschall, Nataliya Sharopova, Stephanie Pretel

NIDCR
Emily Harris

GENEVA program, NHGRI
Teri Manolio, Anastasia Wise
Table 1: Summary of recommended SNP filters. “the Broad” refers to SNPs failed by the genotyping center and “CC” refers to filters recommended by the GENEVA Coordinating Center. The total number of SNPs attempted is 592,532, which is the total number of probes on the array (620,901) minus the number of intensity-only probes. The numbers of SNPs lost to each filter are given through sequential application of the filters. These numbers vary with the order of filter application. The percentage of SNP assays attempted is the number of “SNPs kept” in the line directly above, divided by 592,532.

<table>
<thead>
<tr>
<th>SNP kept</th>
<th>SNP lost</th>
<th>filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>620,901</td>
<td>0</td>
<td>Total probes attempted</td>
</tr>
<tr>
<td>592,532</td>
<td>28,369</td>
<td>Intensity Only</td>
</tr>
<tr>
<td>573,646</td>
<td>18,886</td>
<td>the Broad: Technical Failure</td>
</tr>
<tr>
<td>568,613</td>
<td>5,033</td>
<td>CC: MAF=0 for all samples</td>
</tr>
<tr>
<td>566,769</td>
<td>1,844</td>
<td>CC: call rate &lt; 98%</td>
</tr>
<tr>
<td>565,189</td>
<td>1,580</td>
<td>CC: &gt; 4 Mendelian errors</td>
</tr>
<tr>
<td>564,797</td>
<td>392</td>
<td>CC: HWE p-value &lt; 10^{-4}</td>
</tr>
<tr>
<td>564,718</td>
<td>79</td>
<td>CC: &gt; 0 discordant call in 23 dup pairs</td>
</tr>
<tr>
<td>95.3%</td>
<td>4.7%</td>
<td>Percentage of SNP assays attempted</td>
</tr>
<tr>
<td>475,623</td>
<td>89,095</td>
<td>MAF ≤ 5%</td>
</tr>
<tr>
<td>80.3%</td>
<td>19.7%</td>
<td>Percentage of SNP assays attempted</td>
</tr>
</tbody>
</table>

Table 2: Summary of DNA samples and scans.

<table>
<thead>
<tr>
<th>Study</th>
<th>HapMap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,116</td>
<td>DNA samples into genotyping production</td>
</tr>
<tr>
<td>-52</td>
<td>Failed genotyping</td>
</tr>
<tr>
<td>1,064</td>
<td>Scans released by genotyping center</td>
</tr>
<tr>
<td>-8</td>
<td>Sample identity issues</td>
</tr>
<tr>
<td>1,056</td>
<td>Scans to post on dbGaP</td>
</tr>
<tr>
<td>1,056</td>
<td>Filtered scans for analysis</td>
</tr>
</tbody>
</table>

Table 3: Expected identity-by-descent coefficients for some common relationships.

<table>
<thead>
<tr>
<th>k2</th>
<th>k1</th>
<th>k0</th>
<th>Kinship</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.5</td>
<td>MZ twin or duplicate</td>
</tr>
<tr>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.25</td>
<td>parent-offspring</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50</td>
<td>0.25</td>
<td>0.25</td>
<td>full siblings</td>
</tr>
<tr>
<td>0.00</td>
<td>0.50</td>
<td>0.50</td>
<td>0.125</td>
<td>half siblings</td>
</tr>
<tr>
<td>0.00</td>
<td>0.25</td>
<td>0.75</td>
<td>0.0625</td>
<td>cousins</td>
</tr>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>unrelated</td>
</tr>
</tbody>
</table>
Table 4: Probability of observing more than the given number of discordant calls (“# discordant”) in 23 pairs of duplicate samples, given an assumed error rate. The number of SNPs with a given number of discordant calls is also provided in the last column. The row in red (i.e. SNPs with > 0 discordant call) is the recommended threshold for SNP filtering.

<table>
<thead>
<tr>
<th>Assumed error rate</th>
<th># discordant</th>
<th>3.5e-6</th>
<th>3.5e-5</th>
<th>3.5e-4</th>
<th>3.5e-3</th>
<th># SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 0</td>
<td>1.610e-3</td>
<td>1.609e-2</td>
<td>1.597e-1</td>
<td>0.149</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>&gt; 1</td>
<td>1.240e-8</td>
<td>1.238e-6</td>
<td>1.227e-4</td>
<td>0.0112</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>&gt; 2</td>
<td>6.074e-13</td>
<td>6.068e-10</td>
<td>6.006e-7</td>
<td>0.0005</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>&gt; 3</td>
<td>0</td>
<td>2.124e-13</td>
<td>2.101e-9</td>
<td>1.892e-5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>&gt; 4</td>
<td>0</td>
<td>0</td>
<td>5.589e-12</td>
<td>5.026e-7</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>&gt; 5</td>
<td>0</td>
<td>0</td>
<td>1.177e-14</td>
<td>1.056e-8</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>&gt; 6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.797e-10</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>&gt; 7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.521e-12</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Covariate test results for self-identified Guatemalan subjects.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>0.002038</td>
</tr>
<tr>
<td>age</td>
<td>&lt; 2.2e-16</td>
</tr>
<tr>
<td>age²</td>
<td>&lt; 2.2e-16</td>
</tr>
<tr>
<td>DNA.Source</td>
<td>0.8601</td>
</tr>
<tr>
<td>ev1</td>
<td>0.4426453</td>
</tr>
<tr>
<td>ev2</td>
<td>0.1849762</td>
</tr>
<tr>
<td>ev3</td>
<td>0.3972563</td>
</tr>
<tr>
<td>ev4</td>
<td>0.03914631</td>
</tr>
<tr>
<td>ev5</td>
<td>0.9212742</td>
</tr>
<tr>
<td>ev6</td>
<td>0.2110072</td>
</tr>
<tr>
<td>ev7</td>
<td>0.2423526</td>
</tr>
<tr>
<td>ev8</td>
<td>0.5352246</td>
</tr>
</tbody>
</table>

Table 5: Covariate test results for self-identified Guatemalan subjects.
Figure 1: The X and Y intensities are calculated for each sample from the mean of the sum of the normalized intensities of the two alleles for each probe on those chromosomes. SNP sizes are 17,681 for the X chromosome and 2,160 for the Y chromosome. X heterozygosity is the fraction of heterozygote calls out of all non-missing genotype calls on the X chromosome for each sample.
Figure 2: Schematic diagram showing how trisomic cells (or a mixture of disomic and monosomic cells) can result in two different polar coordinate angle ($\theta$) positions for heterozygotes at different loci. “A” and “B” represent two alleles at one locus, where the former is tagged with Cy3 and the latter with Cy5. Loci with allele B on the duplicated chromosome have ABB heterozygotes and those with allele A on the duplicated chromosome have AAB heterozygotes.
Figure 3: "BAAlleleFreq" (BAF) is a transformation of the polar coordinate angle (Theta) to standardize the positions of the three diploid genotypes to 0, 0.5 and 1.0. The quantities tAA, tAB and tBB are the mean Theta values for each genotypic class. Theta values less than tAA are assigned BAF=0, while those greater than tBB are assigned BAF=1.0. Values in between tAA and tAB are positioned by linear interpolation between 0 and 0.5, and similarly for those between tAB and tBB. Red=AA, green=AB, blue=BB, black 'x' = missing genotype call.
Figure 4: Normal scans of LogRRatio and BAlleleFreq of chromosome 1 for sample A. Each point represents a SNP. In the BAllele Freq plot, the upper and lower bands represent homozygotes and the intermediate band represents heterozygotes.
Figure 5: Abnormal scans of LogRRatio and BAleleFreq of chromosome X for sample A. The intermediate band is split for the entirety, with an increase in intensity of probes (higher LRR value) relative other chromosomes for sample A. This subject is a female with XXX karyotype.
Figure 6: Abnormal scans of chromosome 6 for sample B. The intermediate band of BAlleleFreq is split into two components on the distal end of the q arm. The LogRRatio scan shows a corresponding increase in intensity, suggesting duplication.
Figure 7: This plot shows all study pairs of participants with an estimated $KC > 1/32$. Each point is a pair of subjects. The diagonal line is $k_0 + k_1 = 1$. The upper orange bar is the 95% prediction range for the $k_0$ and $k_1$ values of the half-sibling-like relationships; the lower is calculated in the same manner for the first cousins. The orange rotated ellipse covers the 95% prediction area of the bivariate normal distribution for the full sibling relationships. “PO” refers to parent-offspring; “FS” refers to full-sibling; “HS,Av,GpGe” refers to half-sib-like; “Dup” refers to duplicated; and “U” refers to less related relationship.
Figure 8: Principal component analysis of 544 unrelated subjects, with 1,201 unrelated HapMap III controls. CEU, YRI, CHB, JPT, CHD, MKK, ASW, GIH, MEX, TSI, and LWK indicate HapMap III samples external to controls genotyped with the study participants. The rest are the study samples. “EA” refers European American. “AA” refers to African American. “EA+AA” refers to admixture ancestry of EA and AA. Color-coding is according to self-identified race.
Figure 9: Principal component analysis of 544 unrelated subjects, without hapMap controls. “EA” refers European American. “AA” refers to African American. “EA+AA” refers to admixture ancestry of EA and AA. Color-coding is according to self-identified race.
Figure 10: Comparison of the first two eigenvectors for subset A and subset B, both selected from set 1, indirect versus direct estimates. Each of the two subsets contains 336 unrelated study subjects. The subjects can be related across subsets.
Figure 11: Comparison of the first two eigenvectors for subset A and subset B, both selected from set 2, indirect versus direct estimates. Each of the two subsets contains 283 unrelated, self-identified Guatemalan subjects. The subjects can be related across subsets.
Figure 12: Principal component analysis calculated directly for 544 unduplicated, unrelated study samples. Eigenvectors were inferred for the remaining 489 study samples. Plotted are the results of the direct and indirect analyses. “EA” refers to European American. “AA” refers to African American. “EA+AA” refers to admixture ancestry of EA and AA. Color-coding is according to self-identified race.
Figure 13: Principal component analysis calculated directly for 96 unrelated, U.S. subjects from this study (study 2), along with 2,248 unrelated subjects from the previous Dental Caries study (study 1).
Figure 14: Histograms of the missing call rate per sample ($missing.e1$).

(a) All Counts

(b) Truncated
Figure 15: Boxplots of the missing call rate per sample, classified by type of specimen from which the DNA was extracted.
Figure 16: Median autosomal missing call rate versus number of samples per batch.
Figure 17: A test of allele frequency difference between each genotyping batch and a pool of the other batches, plotted as a function of the ethnic composition of the samples in the batch. The vertical dashed line is the mean composition.
Figure 18: Increasing duplicate genotype discordance between duplicate samples of 20 study participants, along with 3 monozygotic twin pairs. The twin pairs are plotted in red. The median discordance is $7e^{-6}$. 
Figure 19: Increasing genotype discordance between 21 pairs of duplicated HapMap samples. The median discordance is $3.7e-5$. 
Figure 20: QQ plots showing the distribution of p-values from the exact test of Hardy-Weinberg equilibrium in unrelated, PCA-defined Guatemalan subjects.
Figure 21: Venn diagram showing the overlap of various SNP filters. In the diagram, “MCR” refers to SNPs filtered out with a missing call rate > 2%, “HWE” shows the SNPs filtered out with a HWE p-value < 10^{-4}, “DupDisc” refers to SNPs with more than zero discordant call in duplicate genotyping scans, “MendErr” refers to the SNPs filtered out with more than 4 mendelian error per SNP, and “SexDiff” refers to the difference of either heterozygosity or allele frequency between two sexes is greater than 0.3.
Figure 22: Quantile-quantile plots of p-values from the linear regression model described in Section 18. The p-values are for both unfiltered and filtered SNPs. "all" refers to all the self-identified Guatemalan subjects; "unrelated" refers to the unrelated only. When filtering is applied, "(no MAF)" means the composite quality filter was used without the 5% MAF filter being added, while "(with MAF)" means the latter was added. SNP filters are described in Table 1.
Figure 23: Manhattan plots for p-values from the linear regression model described in Section 18. The p-values are all filtered with the composite SNP quality filter. “all” refers to all the self-identified Guatemalan subjects; “unrelated” refers to the unrelated only. SNP filters are described in Table 1.
Figure 24: Allelic intensity cluster plots for the nine SNPs with the lowest association test p-values in the analysis of all the self-identified Guatemalan subjects (TOT_D2MFT ~ age + age² + sex + ev1 + ev2 + ev3 + ev4 + SNP).
Figure 25: Allelic intensity cluster plots for the nine SNPs with the lowest association test p-values in the analysis of the unrelated Guatemalan subjects (TOT_D2MFT ~ age + age^2 + sex + ev1 + ev2 + ev3 + ev4 + SNP).