Is Saliva a Good Alternative to Blood for High Density Genotyping Studies: SNP and CNV Comparisons?

Alban Fabre1, Emilie Thomas1, Sylvain Baulande1, Emile Sohier2, Lyan Hoang1, Pascal Soulauré1, Stéphane Ragusa1, Françoise Clavel-Chapelon1,2* and David G. Cox1

1INSERM, Centre for Research in Epidemiology and Population Health, Gustave-Roussy Institute Villejuif, France
2Paris South University, UMR 1018, France
3Fondation Synergie-Lyon Cancer, Lyon, France
4PartnerChip, Evry, France
5Statlife, Institut Gustave Roussy, Villejuif, France
6INSERM U1052, Lyon, France

Introduction

The current trend of genomic methods used in molecular genetic epidemiology studies is towards hypothesis-free tests, such as genome wide association studies (GWAS) and exome or even whole genome sequencing. Consequently, very large data sets are required to counter the loss of statistical power due to the potentially millions of statistical tests that may be carried out in any given study. Obtaining biological samples suitable for DNA extraction from sufficiently large populations is challenging. This is particularly true in geographically dispersed prospective cohorts, where biological samples not only need to be collected, but stored for long periods of time while phenotypes are developed in the population.

For decades, blood collection has been carried out in many studies. However, the logistics of collecting blood samples can prove daunting. In studies not based in medical settings, access to phlebotomists and transporting blood samples may be difficult and/or expensive. In addition, while drawing blood is minimally invasive, study participants may still not be willing to provide a sample due to the inconvenience and possible adverse effects (pain, bruising, etc.) Recently, the collection of buccal samples has provided the possibility to obtain cells suitable for DNA extraction, and methods of both saliva collection and DNA extraction have improved to the point where saliva is now used regularly in molecular genetic epidemiology studies [1-3].

It is generally accepted that DNA obtained from saliva, while generally of slightly lower quality than that obtained from blood samples, is suitable for use in high throughput genotyping assays, including GWAS platforms for SNP genotyping [3]. This is likely due to the quantitative nature of SNP genotyping (presence or absence of one allele or the other) as opposed to the quantitative nature of CNV calling. However, to our knowledge no studies of the suitability of saliva DNA for use in copy number variation (CNV) genotyping from GWAS platforms have been carried out in humans, and one recent study has examined this question in canines [4]. CNVs occur throughout the genome, and may influence gene expression and therefore disease risk. Genotyping CNVs has shown to be relatively robust both with respect to analyzing the same DNA sample multiple times on one platform as well as across platforms [5,6]. The majority of association studies using CNV data have mainly been carried out using blood, tumor, or cell line DNA.

The suitability of DNA obtained from saliva for CNV analysis is relevant as the DNA is often of lower molecular weight and fragmented. This is mainly attributed to the fact that cells present in saliva are generally dead, often through apoptosis, and may even have severely damaged or altered genomes. While these alterations have little to no effect on SNP genotypes, they may have measurable effects on CNVs. In this study, we have compared both SNP and CNV genotyping in DNA extracted from blood and saliva samples from the same individual, particularly for shorter CNV regions.

Keywords:
- ATP; Portable; Bioluminescence sensor system; Trichloroacetic acid; Luminescence intensity

Abstract

Modern molecular genetic epidemiology is scaled towards large-scale analyses, including genome wide association studies (GWAS) containing hundreds of thousands to millions of single nucleotide polymorphisms. In addition to generating information on alleles at each SNP, GWAS can also be used to evaluate copy number variation (CNVs) across the genome. Traditionally, these studies have been carried out using DNA extracted from lymphocytes in blood samples. More recently, the use of DNA extracted from less invasive methods has become attractive in epidemiological studies. Here, we examine the feasibility of using DNA from saliva to assess CNVs in a pangenome study. We have compared SNP and CNV genotypes among 30 individuals genotyped with the Affymetrix GeneChip NspI genotyping array using DNA from blood and saliva samples of the same individual. In general, while we find that the DNA extracted from these cells is of sufficient quantity and quality to genotype SNPs in a GWAS setting, the results of CNV analyses differed between blood and saliva samples from the same individual, particularly for shorter CNV regions.

Materials and Methods

Participants and sample collection

The E3N is a prospective cohort initiated in 1990 that includes 98,995 French women born between 1925 and 1950, who subscribe to a health insurance plan mainly covering teachers. Blood samples...
were collected between 1995 and 1998 from 24,505 E3N participants, and were aliquoted into plasma, serum, buffy coat, and erythrocytes, then stored in liquid nitrogen [7]. Among these, we randomly selected 30 women who were participants, as controls, in an on-going genome wide association study. We first contacted them in 2010 by telephone to request their participation in a pilot test of exfoliated buccal cell collection, and all 30 responded positively. Each participant received a letter with information on the study, and informed consent form, and detailed instructions on how to deliver the saliva sample using the enclosed kit (Oragene, DNA Genotek Inc., Ottawa, Ontario, Canada). Additionally, the subjects were asked not to drink nor eat for 30 minutes prior to collection, and to provide approximately 0.5mL of saliva.

**DNA extraction and quantification**

DNA was extracted from 200 µl lymphocyte samples using Mini-Ext-Blood-PK kits adapted from the Nucleospin® 96 Blood Quick Pure Kit (Macherey Nagel) according to the manufacturers instructions, under contract with GenoScreen (Lille, France). Purification of genomic DNA from saliva was performed according to the manufacturer’s instructions. Briefly, 500 µl of saliva in Oragene conservative buffer were heated for 2 hours at 50°C to release DNA and inactive nucleases. A 1/25th volume of Oragene DNA Purifier (OG-L2P) was added before incubating the sample on ice for 10 minutes, followed by 5 minutes of centrifugation at 15,000 x g. The supernatant was then carefully transferred into a new microcentrifuge tube and room temperature absolute ethanol was added. After mixing gently, samples were allowed to stand at room temperature for 10 minutes to allow the DNA to fully precipitate. Tubes were then centrifuged for 2 minutes at 15,000 x g. The resulting DNA pellet was washed with 70% ethanol, then resuspended in TE buffer.

Picogreen dye (Quant-IT Pico Green, Invitrogen) was used in order to quantify double-stranded DNA in order to avoid possible contaminating RNA in samples according to the manufacturer's protocol. DNA purity was measured using spectroscopy by Nanodrop (ThermoScientific), comparing the ratio of absorbtion between 260 and 280 nm.

**Genotyping**

Sample genotyping was performed on the Affymetrix GeneChip platform using the NspI mapping array, with approximately 260,000 SNPs, according to the manufacturer's instructions. Briefly, 250 ng of dsDNA was digested with NspI and ligated to adapters using T4 DNA ligase. Samples were then amplified using Titanium taq polymerase (Clonetech) and purified on Nucleofast 96-well plates (Clonetech). 90 µg of PCR products were then subjected to fragmentation with NspI and ligated to adapters using T4 DNA ligase. Samples were then injected into cartridges, hybridized, washed and stained. Array images were obtained using the GeneChip Scanner 3000 and raw signals were deduced and compiled into CEL files. SNP genotypes were called using both the Affymetrix DM algorithm and the BRLMM basecalling algorithm analyzing all CEL files in one batch.

CNV genotypes were called based on the Copy-number estimation using Robust Multichip Analysis (CRMA) version 1 from the R-package aroma.affymetrix with a corrected Gain and Loss Analysis of DNA (GLAD) model [8]. This model was corrected by subtracting the signal at a given probe from the mean signal of all probes in the same region.

---

**Figure 1:**

![Graphs showing cumulative effect of the number of regions according to the length of probes](image-url)
Signals within one standard deviation (SD) of the median value for a region were considered neutral, while those greater or less than one SD away from the median were considered as gains or losses, respectively. The GLAD model was fit independently for each subject and each DNA type (blood or saliva). Regions were variation occurred (either a gain or loss in either DNA sample) were included in analyses. Events were considered equivalent in blood and saliva if a gain or loss was present in a given region (defined by > 80% overlap measured in base pairs or with n common probes in a region of no less than n + 2 overlapping probes) in both DNA samples. The distribution of the lengths of each region is shown in Figure 1.

Statistical analyses

Wilcoxon signed-rank tests were used to evaluate differences of amount and purity between the two methods of collection. We compared reproducibility of SNP results using a Kappa score. Area under the Receiver Operating Characteristic curve, assuming the blood value was true in order to estimate true positives (same event in blood and saliva), false positives (event only in saliva), false negatives (events only in blood), and true negatives (neutral events in both saliva and blood), as well as chi-squared and Kappa statistics were used to evaluate the differences between CNV genotyping in blood and saliva samples.

Results

Blood samples yielded a median concentration of 114.3 ng/µl of DNA, with a median 260/280 ratio of 1.85. Saliva samples were similar, with a median concentration of 133.6 ng/µl and median 260/280 ratio of 1.90. No difference between concentration or purity was observed in comparisons of blood and saliva samples using the Wilcoxon signed-rank test (Table 1).

The median genotyping call rates between blood and saliva were also similar. On average, 97.8% of the SNPs were called across the blood samples, and 96.8% in the buccal samples. However, four samples from saliva failed to genotype for > 93% of the SNPs. Concordance between blood and saliva samples for SNP genotyping (AA, AB, and BB genotypes) was 98.1%, with a Kappa statistic of 0.98. Upon excluding the poorly performing saliva samples (with call rates < 93%), the concordance between blood and saliva was 98.5%, with a Kappa of 0.99 (Table 2).

Figure 2 shows that the larger a region of overlapping events between blood and saliva, the greater the proportion of same events is. Table 3 shows the concordance statistics between blood and saliva CNV calling. Concordance as measured by chi-squared testing was greatest when regions overlapping at 65 or greater probes were used (N regions total). Kappa and ROC analyses show that regions overlapping at 25 or more probes were most similar between blood and saliva. The kappa statistic in this instance was 0.67, and 87% of the regions evaluated were concordant. While concordance increased to 97% with larger overlapping regions, kappa declined to 0.59. Similarly, ROC testing showed that this analysis had 75% sensitivity and 97% specificity. While specificity increased to > 99% with larger overlapping regions, sensitivity declined to 50%. These observations are likely due to the lower number of regions in these analyses (see Table 3).

Discussion

In these analyses, we confirm prior evidence that using DNA ex-
As we move into the post-GWAS era, projects combining not only high throughput genotyping as described here, but also high-throughput sequencing will be the state of the art. While we have shown that in some situations DNA extracted from buccal cells in saliva samples is of sufficient quantity and quality to carry out large scale SNP genotyping and CNV analyses, further work needs to be carried out to evaluate the suitability of such samples for next generation sequencing.

Acknowledgements

The authors are indebted to all participants for providing data and to practitioners for providing pathology reports. The authors are grateful to R. Chalt, M. Fangon, M. Niravong, Gwenaelle Guillas, Lyan Hoang and C. Kernaleguen for managing the data.

The E3N study is being carried out with financial support from the French League Against Cancer, the Mutuelle Generale de l’Education Nationale, the Institut Gustave Roussy and the Institut National de la Santé et de la Recherche Medicale. The present study was financially supported by INCa-Lilly and the Agence Nationale de la Recherche. Financial support for the collection of saliva samples was given by the French League Against Cancer and the French Public Health Research Institute (through the Large Research Infrastructure Action Plan). Alban Fabre was supported by the Canéropôle Région Ile de France. The authors are grateful to DNA Genetek for providing the saliva kits and 250K Affymetrix chips.

References

1. Paynter RA, Skibola DR, Skibola CF, Buffler PA, Wiemels JL, et al. (2006) Is Saliva a Good Alternative to Blood for High Density Genotyping Studies: SNP and CNV Comparisons? J Biotechnol Biomaterial 1:119. doi: 10.4172/2155-952X.1000119

2. Bengtsson H, Irizarry R, Carvalho B, Speed TP (2008) Estimation and assessment of SNP genotyping errors. Bioinformatics 24: 759-767.

3. Bengtsson H, Bischof C, Heyer W, Siedersleben D, Schleicher EM, et al. (2006) Systematic comparison of DNA quality and genotyping in blood and saliva samples. J Med Genet 43: 201-206.

4. Fabre A, Thomas E, Baulande S, Sohier E, Hoang L, et al. (2011) Is Saliva a Good Alternative to Blood for High Density Genotyping Studies: SNP and CNV Comparisons? J Biotechnol Biomaterial 1:119. doi: 10.4172/2155-952X.1000119

5. Fabre A, Thomas E, Baulande S, Sohier E, Hoang L, et al. (2011) Is Saliva a Good Alternative to Blood for High Density Genotyping Studies: SNP and CNV Comparisons? J Biotechnol Biomaterial 1:119. doi: 10.4172/2155-952X.1000119

6. Bengtsson H, Ray A, Spellman P, Speed TP (2009) A single-sample method for SNP genotyping using Illumina BeadChip technology. Bioinformatics 25: 1413-1419.

Table 1: DNA quantity and quality in blood and saliva samples.

| Collection     | Blood (ng/µL; range) | Saliva (ng/µL; range) | p-value |
|----------------|----------------------|-----------------------|---------|
| DNA concentration (median) | 114.3 (17.7 - 478.2) | 133.6 (24.5 - 475.7) | 0.15    |
| 260/280 ratio (median; range) | 1.85 (1.42 - 2.15) | 1.90 (1.71 - 2.04) | 0.18    |

Table 2: SNP genotyping comparisons between blood and saliva DNA samples.

| Chip Pairs | ROC | Chi-Squared | K | % concordance |
|------------|-----|-------------|---|---------------|
| Global     | 0.15 | 0.43 | 1.59 x 10^-14 | -0.43 | 50.9 |
| Global SE25 | 0.75 | 0.97 | 4.76 x 10^-14 | 0.67 | 87.1 |
| Global SE45 | 0.72 | 0.97 | 5.05 x 10^-14 | 0.64 | 88.2 |
| Global SE65 | 0.50 | 0.99 | 1.95 x 10^-14 | 0.59 | 96.6 |

Table 3: CNV genotyping statistics between pairs of blood and saliva samples.