Published OnlineFirst March 3, 2010; DOI: 10.1158/1055-9965.EPI-09-0812

Cancer Epidemiology, Biomarkers & Prevention

Research Article

Saliva-Derived DNA Performs Well in Large-Scale, High-Density Single-Nucleotide Polymorphism Microarray Studies

Melanie Bahlo[1], Jim Stankovich[2], Patrick Danoy[3], Peter F. Hickey[1,4], Bruce V. Taylor[2], Sharon R. Browning[6], The Australian and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), Matthew A. Brown[3,7], and Justin P. Rubio[5,8]

Abstract

As of June 2009, 361 genome-wide association studies (GWAS) had been referenced by the HuGE database. GWAS require DNA from many thousands of individuals, relying on suitable DNA collections. We recently performed a multiple sclerosis (MS) GWAS where a substantial component of the cases (24%) had DNA derived from saliva. Genotyping was done on the Illumina genotyping platform using the Infinium Hap370CNV DUO microarray. Additionally, we genotyped 10 individuals in duplicate using both saliva- and blood-derived DNA. The performance of blood- versus saliva-derived DNA was compared using genotyping call rate, which reflects both the quantity and quality of genotyping per sample and the "GCScore," an Illumina genotyping quality score, which is a measure of DNA quality. We also compared genotype calls and GCScores for the 10 sample pairs. Call rates were assessed for each sample individually. For the GWAS samples, we compared data according to source of DNA and center of origin. We observed high concordance in genotyping quality and quantity between the paired samples and minimal loss of quality and quantity of DNA in the saliva samples in the large GWAS sample, with the blood samples showing greater variation between centers of origin. This large data set highlights the usefulness of saliva DNA for genotyping, especially in high-density single-nucleotide polymorphism microarray studies such as GWAS. *Cancer Epidemiol Biomarkers Prev; 19(3); 794–8. ©2010 AACR.*

Introduction

Genome-wide association studies (GWAS) require the collection of thousands of DNA samples to (1) attain sufficient power to map loci responsible for complex diseases.

DNA for such studies is usually extracted from lymphocytes, with blood collected from participants by phlebotomy. A recent study of DNA collection in a Danish nurse cohort (2) showed a markedly greater response rate in recruitment of DNA samples from saliva versus blood (72% versus 31% of invited participants returned samples, respectively). These differences in response stemmed from a variety of issues, including the commitment required to provide blood by attending a clinic and uneasiness, or inability, to undergo phlebotomy. In comparison, saliva collection can be done at home, without professional help, and is much less invasive than phlebotomy. However, saliva-derived DNA is often contaminated with large amounts of bacterial DNA and it is notoriously difficult to determine the relative proportion of human DNA. Moreover, different protocols are required for purification and quantitation of DNA derived from blood and saliva (3, 4). It is therefore entirely plausible that DNA derived from these different sources might show systematic differences in genotyping efficiency and accuracy.

We recently performed a multiple sclerosis (MS) GWAS with 1,618 MS cases and 3,413 controls under the auspices of the Australia and New Zealand MS Genetics Consortium (ANZgene; ref. 5). Twenty-four percent of MS cases had genotyping done on DNA derived from saliva and the remaining samples were derived from blood. The data set used in the current study consisted of genotypes from these MS cases, some

Authors' Affiliations: [1]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia; [2]Menzies Research Institute, University of Tasmania, Hobart, Tasmania, Australia; [3]Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Brisbane, Queensland, Australia; [4]Department of Mathematics and Statistics and [5]The Florey Neuroscience Institutes, University of Melbourne, Melbourne, Victoria, Australia; [6]Department of Statistics, The University of Auckland, Auckland, New Zealand; [7]Botnar Research Centre, Nuffield Department of Orthopaedic Surgery, University of Oxford, Oxford, United Kingdom; and [8]Genetics Division, Research and Development, GlaxoSmithKline, Harlow, Essex, United Kingdom. ANZGene Consortium authors and affiliations are listed in the Supplementary Data.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Corresponding Author: Melanie Bahlo, The Walter and Eliza Hall Institute of Medical Research, Parkville 3052, Victoria, Australia. Phone: 61-3-9345-2630; Fax: 61-3-9347-0852. E-mail: bahlo@wehi.edu.au

doi: 10.1158/1055-9965.EPI-09-0812

©2010 American Association for Cancer Research.

Downloaded from cebp.aacrjournals.org on May 12, 2021. © 2010 American Association for Cancer Research.

Published OnlineFirst March 3, 2010; DOI: 10.1158/1055-9965.EPI-09-0812

of the controls, as well as some samples that were not included in the published GWAS. In addition, a small data set consisting of 10 sample duplicates, where both saliva- and blood-derived DNAs were available, was also examined. These data were used to investigate the relative genotyping performance of blood- and saliva-derived genomic DNA using the Illumina single-nucleotide polymorphism (SNP) microarray platform.

Materials and Methods

Genomic DNA from Australian MS cases and controls was extracted from whole blood using a variety of standard laboratory approaches including phenol-chloroform extraction (6), salting out (7), and three different commercial kits from Qiagen, Nucleon, and Roche, as per manufacturers' instructions. Genomic DNA from MS cases from New Zealand (NZ) was isolated from saliva self-collected into Oragene DNA tubes according to the manufacturer's instructions (DNAgenotek). DNA concentrations were assessed using Pico Green fluorescence, UV ($A_{260\text{ nm}}$) spectrophotometry, and/or on an ethidium bromide–stained low-percentage agarose gel compared with a high molecular weight standard. Because of possible bacterial genomic DNA contamination and difficulty in obtaining reliable Pico Green and spectrophotometry measurements, all saliva DNA samples were estimated and assessed for their integrity by agarose gel electrophoresis and using at least one other method. Extraction methods, age of samples, and time to extraction are summarized by study center in Supplementary Table S1.

Genotyping was done on the Illumina genotyping platform with the Infinium Hap370CNV DUO microarrays for both saliva and blood DNAs using the same protocol, at the same facility (Diamantina Institute, Brisbane, Queensland, Australia), within a 6-mo time frame.

Each genotype call was associated with a genotyping quality score. For the Illumina platform, this is known as a GCScore or GenCall Score (see Supplementary Data) and ranges from 0 to 1. The GenCall Score is a summary measure consisting of three parts: (*a*) a SNP-specific score known as the GenTrain Score, which describes the clustering properties of the genotyped SNP; (*b*) the fit of the current sample to the clustering profile of the SNP; and (*c*) the DNAscore, which summarizes the overall DNA quality of the individual. The GenCall Score thus takes into account SNP properties, sample DNA properties, and SNP-specific sample properties. The score has greatest sensitivity in the range of 0.2 to 0.7, with scores >0.7 signaling high-quality genotypes. Illumina BeadStudio guidelines stipulate that the genotype calling software determines a call to be a “no-call” if the GCScore is <0.15; however, the GCScores are available for all SNPs, even if the GCScore is below this threshold.

In this analysis, we elected to keep all genotyping results regardless of GCScore rather than working with a censored distribution. Thus, all SNP markers were used for the call rate and GCScore analyses (n = 353,203 of a total of 370,405 probes on the array), which exclude the copy number variation (CNV) probes (n = 17,202) on the array.

The call rate of a sample was determined by the number of SNPs with a genotype call divided by the total number of SNPs considered. Illumina Beadstudio guidelines suggest that samples with a no-call rate of >2% are likely to be of poor quality, and thus samples that failed this threshold were removed from any further analysis.

Paired Sample Study

Venous blood and saliva were collected from 10 consecutively recruited MS patients from study center Aus9, and the extracted DNAs were genotyped (20 samples in total). The samples were assessed individually for call rate and in a paired comparison for genotype concordance and GCScore similarity.

GWAS Data Set

The original recruitment number of ANZgene samples for the MS study was 2,000 DNA samples. Of these, 75 failed preliminary QC requirements. Of the remaining 1,925 samples, 1,873 were cases and 52 controls. Stringent QC analysis to remove samples and SNPs using additional methods beyond call rate and GCScore thresholds, pertaining to the published MS GWAS association analysis, led to the inclusion of only 1,618 ANZgene cases and 41 controls for the published MS GWAS analysis (5). We chose to use the larger data set involving 1,925 samples for the genotyping quality analysis in the current study because samples and SNPs rejected for other QC reasons than genotyping quality and quantity were still informative for assessment of blood and saliva differences. Control data used in the GWAS data set (n = 3,370; ref. 5), which was provided by the Wellcome Trust Case Control Consortium and the Illumina iControl database, were not used in the current study because genotyping quality scores were not available. Case/control status was not taken into account in the analysis.

Ten centers from Australia, designated as Aus1,…, Aus10, contributed blood-derived genomic DNA to the GWAS (Supplementary Table S1). One center, in New Zealand (NZ), sent out Oragene saliva collection kits within NZ to MS cases self-identified to the investigators through a national prevalence survey. Saliva samples (95% of mailed kits) were returned to the NZ recruitment center over a 6-mo period and were sent to study center Aus9 where DNA extractions were done. All samples were derived from MS patients with ethnic background verbally verified to be Caucasian. All blood- and saliva-derived DNA samples were genotyped in several genotyping batches over a 6-mo period. No randomization for genotype centers or case/control status was done over genotype batches due to time constraints. We were unable to investigate genotype batch effects and study center–specific differences between the samples, such as age of sample, due to strong confounding between genotype batch and study center.

Downloaded from cebp.aacrjournals.org on May 12, 2021. © 2010 American Association for Cancer Research.

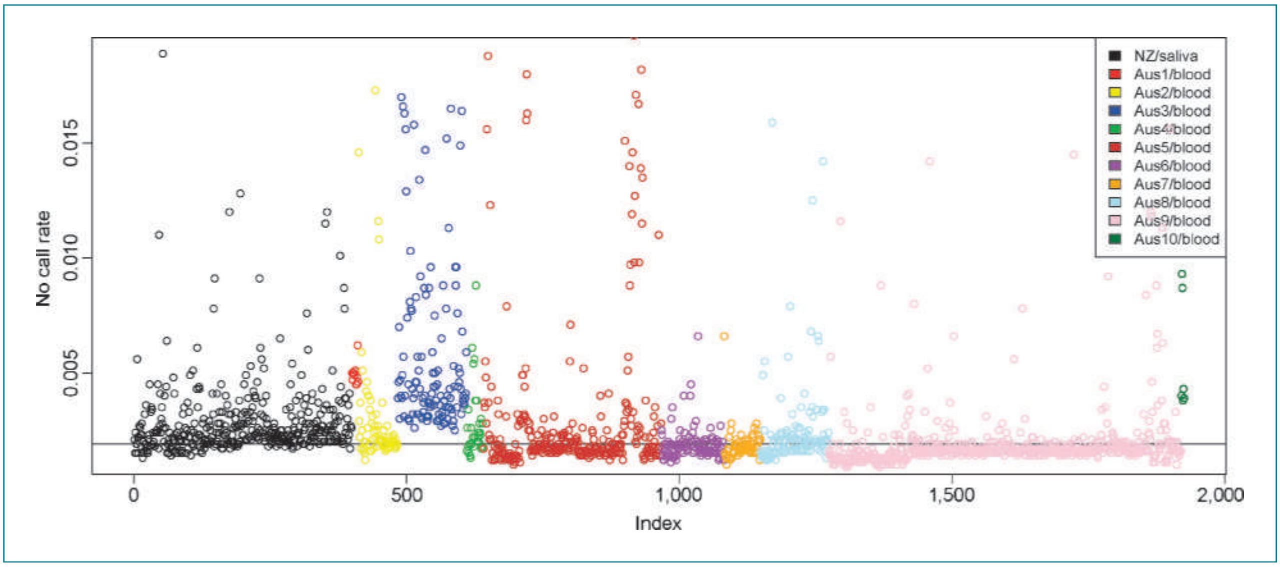

Figure 1. No-call rates for the GWAS by study centre. No-call rates across all Australian and New Zealand samples (N = 1925), which passed the initial 98%, call rate threshold. Call rates are coloured according to study centre. The New Zealand samples (black) are all saliva derived DNAs. All Australian samples are from blood derived DNA. The black line is the median overall no-call rate (0.0019). Samples are plotted (when possible) according to ID number. Not all study centres have numerical IDs.

The GWAS data set was assessed for genotype call rate across batches, study centers (10 Australian + 1 New Zealand), and DNA source (blood or saliva). The GCScore distribution for each individual was summarized using the three quartiles (25th, 50th, and 75th percentiles) of the GCScore distribution and these were averaged within relevant pools, either by study center or by DNA source. The 50th percentile, or second quartile, corresponds to the median. The interquartile range (IQR) was taken as the range of values delimited by the values at the 3rd (upper) and 1st (lower) quartiles, respectively. If the data were normally distributed, the first and third quartiles would correspond to approximately $\mu - 0.7s$ and $\mu + 0.7s$, where μ is the mean and s the SD.

The published MS GWAS analysis (5) included an analysis looking for sample stratification that could unduly influence the association analysis. The source of the sample stratification is usually ethnic stratification but can also be potentially caused by technical factors such as genotyping batches or DNA source. The sample stratification analysis was carried out only on the published MS GWAS data set with the software EIGENSTRAT (8), which identifies data clusters using the statistical technique of principal component analysis. There is a strong overlap between the set analyzed by EIGENSTRAT and the current set of data, as the EIGENSTRAT analysis included 1,659 of 1,925 (86%) samples in the current data set.

Results

Paired Sample Study

This study allowed a paired comparison of the blood- and saliva-derived DNA samples, thus removing biological variation and most of the technical variation. The blood DNA sample of one of the paired samples had very poor genotyping quality with a call rate of 0.59. Therefore, this paired sample was dropped entirely from the subsequent statistical analysis, leaving nine paired samples. In the remaining nine paired samples, there was no significant difference in call rates between blood and saliva DNA for the complete set of SNPs (353,203 SNPs; one-sided paired two-sample t test, T_8 = 0.58, P = 0.3). We also observed high correlation between GCScores between the samples with a median Spearman's rank correlation coefficient of 0.82 (range, 0.81-0.83) based on all SNPs, regardless of GCScore (Supplementary Figs. S1-S5). There was no difference when SNPs with a GCScore <0.15 were excluded. The number of discordant genotype calls in these paired samples was very low with a median of 0.0035% discordant genotypes, with a range of 0.0007% to 0.0142%, n = 9.

GWAS Data Analysis

The median SNP call rate (99.82%; n = 1,659) in the GWAS data set (5) was comparable to other published GWAS, and of the original 2,000 MS case DNA samples genotyped, 75 (3.8%) did not pass the call rate threshold of 98%. No saliva-derived samples failed the call rate threshold. The difference in blood and saliva call rate threshold failures is highly statistically significant (Z = 8.71, P < 2.2e−16), with saliva-derived DNAs much less likely than blood-derived DNAs to fail the QC threshold in this study.

The median no-call rate of all samples was 0.0019 (call rate of 99.81%; n = 1,925), for blood DNA it was 0.0018,

Downloaded from cebp.aacrjournals.org on May 12, 2021. © 2010 American Association for Cancer Research.

Published OnlineFirst March 3, 2010; DOI: 10.1158/1055-9965.EPI-09-0812

and for saliva DNA it was 0.0024 (Fig. 1; Supplementary Table S1). There was a significantly higher rate of no calls in the saliva samples when compared with the blood samples (including only samples with no-call rates <0.02; n_{saliva} = 399; n_{blood} = 1,526; two-sample Wilcoxon rank-sum test = 171,469; P < 2.2e−16).

Blood derived from three centers (Aus1, Aus3, and Aus10) performed poorly in comparison with blood- and saliva-derived DNA from other centers (Supplementary Table S2). Aus5 and Aus9 study center samples also showed additional, intrabatch effects that could not be explained by other factors, such as genotyping batch, because all samples in these cohorts were either genotyped in the same batch (Aus9) or only two batches (Aus5). Further investigation of these data revealed a relationship with sample ID, suggesting a link to time of collection and possibly systematic differences in sample storage and/or processing.

A comparison of GCScores across studies centers also revealed similar patterns of genotyping bias as evidenced by differences in no-call rate (Supplementary Table S3). The saliva samples had a higher median GCScore than the blood samples, but identical quartiles and IQR. The study centers whose samples had lower call rates (Aus1, Aus3, and Aus10) often had lower IQR and lower median GCScores, indicating that they now represent a biased, or much cleaner, sample. In general, the GCScore quartiles were very similar across all studies centers but this may reflect the nondiscriminatory nature of the GCScore.

The published GWAS data (5) were subjected to stringent QC including principal components analysis using EIGENSTRAT (8) software, but this did not identify any principal components that clustered with DNA source (data not shown), suggesting that saliva-derived DNA did not have its own SNP genotyping signature in comparison with blood-derived DNA.

Discussion

DNA samples genotyped in our GWAS were collected through 11 different study centers at different times using different extraction methods, and these factors seemed to affect genotype data quality more than the source of DNA, although we were unable to test this hypothesis specifically. The large size of our study permitted averaging over these confounders, and a comparison of SNP array–based genotyping performance between blood- and saliva-derived DNA was done. Our findings suggest that saliva collected using the Oragene kit provides good-quality genomic DNA, which is comparable to blood as a template for SNP genotyping on the Illumina platform. Some studies using different genotyping methods such as PCR for a small number of polymorphisms (9) support our findings, whereas others have had little success with saliva-derived DNA (10). No other studies have conducted such an extensive comparison of genotyping quality using saliva- versus blood-derived DNA.

Although both Affymetrix and Illumina claim that Oragene saliva DNA works well on their SNP arrays, there has been no external validation to date. It is plausible that bacterial genomic DNA, which contaminates saliva DNA but not blood DNA, may interfere with genotyping quality. However, our results suggest that the Illumina platform is robust to this potential confounder. This could be due to processing steps and/or the length of the probes on Illumina arrays (50 bp) that may help to overcome the effects of bacterial DNA contamination. If the length of the probes were important, one would expect the Affymetrix system to fare worse than that of Illumina because the probes are shorter (25 bp). The Illumina system seems to be quite sensitive to DNA concentration and it is recommended that template DNA be standardized to 50 ng/μL before genotyping. Here we standardized both saliva and blood DNA to the recommended concentration without taking account of the relative proportion of human versus bacterial DNA in the saliva DNA samples.

Researchers contemplating genetic studies where it is difficult to derive DNA from blood due to disability or aversion to phlebotomy will be reassured that data generated from saliva-derived DNA incur few losses in terms of either genotyping data quantity or quality in comparison with blood-derived DNA. It is unknown if our findings will translate to other high-throughput genotyping platforms such as Nimblegen and Affymetrix, although highly variable results were reported for the 500K Affymetrix platform using unquantitated saliva-derived DNA (4). Interestingly, DNA from whole genome amplified samples also seems to perform well on the Illumina platform (11), but genotyping quality and call rate were worse than we observed for either saliva- or blood-derived DNA.

Unlike the Affymetrix platform, the Illumina platform permits multiple samples to be genotyped per chip, depending on the chip design. Illumina fixed-content GWAS SNP chips vary from single sample designs to duo designs, like the chip used for this study, up to designs which can assay 12 samples simultaneously on a single chip. This study was not designed to investigate the influence of array design, and there is as yet no published evidence to indicate that this is likely to be a significant factor in genotyping quality.

Finally, including all consumables and labor costs, each saliva sample used in this study costs on average US$80 to collect and process. Here we have shown that saliva DNA is of high quality and suitable as a template for array-based SNP genotyping, at least for the Illumina Infinium genotyping platform. The Oragen kit also enables self-collection and therefore presents minimal inconvenience to the participant, resulting in high response rates. Further, we have shown that saliva samples can be sent in the mail to a central collection point, thereby reducing transportation costs and the risk of

Downloaded from cebp.aacrjournals.org on May 12, 2021. © 2010 American Association for Cancer Research.

Published OnlineFirst March 3, 2010; DOI: 10.1158/1055-9965.EPI-09-0812

duplication. These financial and logistical benefits will positively affect other genetic studies seeking to expand collections for future research.

Disclosure of Potential Conflicts of Interest

S.J. Foote: ownership interest, Murigen Therapeutics Pty. Ltd.; Consultant/Advisory Board, Genera Biosciences.

Acknowledgments

We thank individuals with MS in Australia and New Zealand for supporting this research, M. Tanner for database management, and J. Wright and C. Remediakis from Multiple Sclerosis Research Australia (MSRA) for expediting this research.

Grant Support

M. Bahlo is supported by Career Development Award from The National Health and Medical Research Council of Australia (NHMRC). M.A. Brown is an NHMRC Principal Research Fellow. H. Butzkueven (ANZgene Consortium), is an NHMRC Peter Doherty Postdoctoral Fellow. J. Field is an MSRA Postdoctoral Fellow. M.B. Cox is supported by a grant from the John Hunter Hospital Charitable Trust Fund and a special grant from Macquarie Bank. This work was supported by the Ian Potter Foundation, the Australian Cancer Research Foundation, MSRA, the John T. Reid Charitable Trusts, the Trish MS Research Foundation, and the Australian Research Council, under the Linkage Projects Scheme (LP0776744).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received 08/16/2009; revised 11/19/2009; accepted 01/05/2010; published OnlineFirst 03/02/2010.

References

1. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology. Nat Genet 2008;40:124–5.

2. Hansen TV, Simonsen MK, Nielsen FC, Hundrup YA. Collection of blood, saliva, and buccal cell samples in a pilot study on the Danish nurse cohort: comparison of the response rate and quality of genomic DNA. Cancer Epidemiol Biomarkers Prev 2007:2072–6.

3. Quinque D, Kittler R, Kayser M, Stoneking M, Nasidze I. Evaluation of saliva as a source of human DNA for population and association studies. Anal Biochem 2006:272–7.

4. Herráez DL, Stoneking M. High fractions of exogenous DNA in human buccal samples reduce the quality of large-scale genotyping. Anal Biochem 2008:329–31.

5. ANZgeneConsortium. Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet 2009;41:824–8.

6. Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 1987;162:156–9.

7. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 1988; 16:1215.

8. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–9.

9. Rylander-Rudqvist T, Håkansson N, Tybring G, Wolk A. Quality and quantity of saliva DNA obtained from the self-administrated oragene method—a pilot study on the cohort of Swedish men. Cancer Epidemiol Biomarkers Prev 2006:1742–5.

10. Philibert RA, Zadorozhnyaya O, Beach SR, Brody GH. Comparison of the genotyping results using DNA obtained from blood and saliva. Psychiatr Genet 2008:275–81.

11. Teo YY, Inouye M, Small KS, et al. Whole genome-amplified DNA: insights and imputation. Nat Methods 2008;5:279–80.

Downloaded from cebp.aacrjournals.org on May 12, 2021. © 2010 American Association for Cancer Research.

Published OnlineFirst March 3, 2010; DOI: 10.1158/1055-9965.EPI-09-0812

Cancer Epidemiology,
Biomarkers & Prevention

American Association
for Cancer Research

Saliva-Derived DNA Performs Well in Large-Scale, High-Density Single-Nucleotide Polymorphism Microarray Studies

Melanie Bahlo, Jim Stankovich, Patrick Danoy, et al.

Cancer Epidemiol Biomarkers Prev 2010;19:794-798. Published OnlineFirst March 3, 2010.

Updated version	Access the most recent version of this article at: doi:10.1158/1055-9965.EPI-09-0812
Supplementary Material	Access the most recent supplemental material at: http://cebp.aacrjournals.org/content/suppl/2010/03/02/1055-9965.EPI-09-0812.DC1

E-mail alerts	Sign up to receive free email-alerts related to this article or journal.
Reprints and Subscriptions	To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.
Permissions	To request permission to re-use all or part of this article, use this link http://cebp.aacrjournals.org/content/19/3/794. Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.

Downloaded from cebp.aacrjournals.org on May 12, 2021. © 2010 American Association for Cancer Research.