Published in final edited form as:

Cancer Epidemiol Biomarkers Prev. 2008 June ; 17(6): 1368–1373. doi:10.1158/1055-9965.EPI-07-2830.

Pathway Analysis of Single-Nucleotide Polymorphisms Potentially Associated with Glioblastoma Multiforme Susceptibility Using Random Forests

Jeffrey S. Chang[1], **Ru-Fang Yeh**[1], **John K. Wiencke**[2], **Joseph L. Wiemels**[1], **Ivan Smirnov**[2], **Alexander R. Pico**[3], **Tarik Tihan**[4], **Joe Patoka**[2], **Rei Miike**[2], **Jennette D. Sison**[2], **Terri Rice**[2], **Margaret R. Wrensch**[2]

[1]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California

[2]Department of Neurological Surgery, Division of Neuroepidemiology, University of California, San Francisco, San Francisco, California

[3]Gladstone Institute of Cardiovascular Disease, University of California, San Francisco, San Francisco, California

[4]Department of Pathology, University of California, San Francisco, San Francisco, California

Abstract

Glioma is a complex disease that is unlikely to result from the effect of a single gene. Genetic analysis at the pathway level involving multiple genes may be more likely to capture gene-disease associations than analyzing genes one at a time. The current pilot study included 112 Caucasians with glioblastoma multiforme and 112 Caucasian healthy controls frequency matched to cases by age and gender. Subjects were genotyped using a commercially available (ParAllele/Affymetrix) assay panel of 10,177 nonsynonymous coding single-nucleotide polymorphisms (SNP) spanning the genome known at the time the panel was constructed. For this analysis, we selected 10 pathways potentially involved in gliomagenesis that had SNPs represented on the panel. We performed random forests (RF) analyses of SNPs within each pathway group and logistic regression to assess interaction among genes in the one pathway for which the RF prediction error was better than chance and the permutation $P < 0.10$. Only the DNA repair pathway had a better than chance classification of case-control status with a prediction error of 45.5% and $P = 0.09$. Three SNPs (rs1047840 of *EXO1*, rs12450550 of *EME1*, and rs799917 of *BRCA1*) of the DNA repair pathway were identified as promising candidates for further replication. In addition, statistically significant interactions ($P < 0.05$) between rs1047840 of *EXO1* and rs799917 or

Jeffrey S. Chang, Department of Epidemiology and Biostatistics, University of California, San Francisco, 44 Page Street, Suite 503, San Francisco, CA 94143-1215. Phone: 510-642-6299; Fax: 510-643-1735. jeffrey.chang@ucsf.edu.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

[5]http://www.snplogic.org

[6]http://cran.r-project.org/

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

rs1799966 of *BRCA1* were observed. Despite less than complete inclusion of genes and SNPs relevant to glioma and a small sample size, RF analysis identified one important biological pathway and several SNPs potentially associated with the development of glioblastoma. (Cancer Epidemiol Biomarkers Prev 2008;17(6):1368–73)

Introduction

Studying one or a few genes may not be sufficient to understand complex diseases such as cancer because they are unlikely to result from the effect of only one or a few genes. Analyses focusing on single genes or individual single-nucleotide polymorphisms (SNP) may have several limitations: (*a*) A large number of comparisons increases the probability of false positives. (*b*) Interactions are ignored; each SNP or gene alone may have little or no effect on risk of disease, but together (gene-gene interaction) may increase the risk substantially. (*c*) Most SNPs are either unannotated or of unknown biological function, making it difficult to link putative associations with disease in a biological context. (*d*) Low replication rate, such that failure for replication may indicate a true null association, or alternatively, may be due to different linkage disequilibrium patterns with the true causal SNP(s) among different study populations. These problems may be alleviated if one examines gene effects in the context of biological pathways. Although limitations of single gene (or SNP) analysis and strengths of pathway analysis have long been recognized, the biostatistical, bioinformatic, and genotyping tools necessary for such pathway analysis have only recently become widely available.

Random forests (RF) is a tree-based classification method developed by Breiman (1, 2). It has several key features ideal for analyzing multiple SNPs: (*a*) the ability to analyze a data set that has a high ratio of number of predictor variables to observations; (*b*) the ability to detect a SNP that has a weak main effect but has significant interaction with other SNPs (3); (*c*) the importance measure from RF, which gives a natural ranking of SNPs; (*d*) having no requirement to specify the mode of inheritance (dominant, codominant, or recessive); and (*e*) the effect of a risk allele not canceling the effect of a protective allele in a RF analysis, as the RF algorithm does not assume any directionality for the risk association for each allele. Diaz-Uriarte et al. (4) showed that classification error rates of RF are equivalent to those produced by other classification methods including support vector machines and *K*-nearest neighbor. In addition, Diaz-Uriarte developed a RF-based gene selection procedure that matches or outperforms alternative methods by selecting fewer genes with equal or lower prediction errors (4). A pathway approach analysis using RF has been applied successfully to gene expression data by Pang et al. (5). The current analysis applies a similar strategy to analyze associations between SNPs and glioblastoma status in a pilot case-control study to evaluate the feasibility of using RF as an analytic tool to study SNP-disease association.

Materials and Methods

Study Subjects.

Subject recruitment has been detailed previously (6, 7). Briefly, eligible case subjects included incident adult (age >20 y) glioma cases diagnosed between August 1991 and April

1994 (series 1) and between May 1997 and August 1999 (series 2), who resided in six San Francisco Bay Area counties at the time of diagnosis. Newly diagnosed adults with glioma were identified through a rapid case ascertainment program from the San Francisco Bay Area population-based cancer registry, the Northern California Cancer Center. Controls identified using random digit dialing were frequency matched to cases by age, gender, and ethnicity. One hundred twelve Caucasians with glioblastoma (20 from series 1 and 92 from series 2) and 112 Caucasian controls were included in this pilot genotyping study.

This study was approved by the University of California, San Francisco Committee on Human Research. All participants signed detailed consent forms in accordance with the Helsinki Doctrine.

Genotyping.

Genotyping was done using the commercially available ParAllele (now part of Affymetrix) SNP panel, which contains 10,177 nonsynonymous coding SNPs; this represented all such known SNPs spanning the genome that could be accurately genotyped with the ParAllele genotyping method at the time of the study (8). A list of the 10,177 SNPs on the assay panel is provided in Supplementary Table S1.

Statistical and Bioinformatic Methods

Choosing SNPs within Pathways.—Because this analysis is a feasibility study for using RF to analyze SNPs, we only selected a few pathways that have some biological plausibility for being involved in gliomagenesis. The 10 pathways chosen included several commonly involved in cancers [phase I and II carcinogen metabolism (9, 10), DNA repair (11), cell cycle (12), apoptosis (13), cell adhesion (14, 15), and calcium signaling (16)], two pathways that are dysregulated in brain tumors [mitogen-activated protein kinase signaling (17) and WNT signaling (17)], and two immune function pathways potentially involved in gliomagenesis [transforming growth factor-β (18) and interleukin-6 (19)]. We then selected genes and SNPs from the panel belonging to those pathways using the review article by Wood et al. (11) for the DNA repair pathway, and pathway websites including KEGG (20), BioCarta (21), and GenMAPP (22) for the other pathways. We also used an interactive web resource, called SNPLogic,[5] which we developed to help identify and categorize genes and SNPs potentially related to glioma (23). SNPLogic provides an integrated view across multiple pathway resources together with a variety of SNP annotations, haplotype information, and functional predictions.

Identification of Important Pathways.—Within each pathway, RF analyses were done with SNPs using random-Forest R package version 4.5–18 (by Liaw and Wiener), available through the Comprehensive R Archive Network (CRAN) website.[6] Of the 227 SNPs examined in this analysis (33 SNPs belong to more than one pathway), 149 SNPs had some missing values. Among these 149 SNPs, 134 (90%) had <5% with missing values. Missing values of SNPs were imputed based on the proximity measure (2, 24) and this procedure was carried out using the rfImpute function in randomForest R package.

RF is a tree-based classification algorithm similar to Classification and Regression Tree (CART; refs. 1, 2, 24). In contrast to CART, which builds only one classification tree, RF builds a collection of trees to produce a more stable prediction error (1, 2, 24); 20,000 trees were built for each pathway in the current analysis. RF builds each individual tree by taking a bootstrap sample (sampling with replacement) of the original data, and on average, about one third of the original data are not sampled [out of bag (OOB)]. Those sampled are used as the training set to grow the trees, and the OOB data are used as the test set. At each node of the tree, a random sample of m of the total M variables is chosen and the best split is found among the m variables. The default value for m in the randomForest R package is the square root of M. In the current analysis, we tested a range of m from half of the square root of M to two times the square root of M and used the m that gave the lowest prediction error. Each tree in the RF analysis is grown fully without pruning. Each classification tree of the forest gets one vote for each OOB observation, and for each observation the class (case versus control status in our study) with the most votes is the RF prediction. The OOB error rate is the percentage of time the RF prediction is incorrect. In addition to OOB error rate, RF also produces importance scores that can be used to rank variables. The importance scores are determined by permuting the values of each predictor variable in the testing set; the more important a variable is, the larger the increase in the OOB error rate will be due to the permutation.

For each pathway with a prediction error <50%, 100 data sets were generated, which randomly shuffled the relationship between the case-control status and the SNPs. RF analysis was done with each of the 100 data sets to generate a null distribution of the prediction error; the P value was then determined by the percentage of prediction errors equal to or lower than the prediction error of the original data set.

Identification of Important SNPs.—For the pathways with a prediction error rate better than chance (prediction error <50% and $P \leq 0.10$), further analyses were done using varSelRF R package (4) to reduce the number of SNPs down to the “best” set. Because this is a pilot study with a small sample size mainly used to evaluate the feasibility of using RF to analyze SNPs by pathways, it was felt that $P \leq 0.10$ was a reasonable cutoff for further investigating a pathway. The best set of SNPs was determined using an iterative process of fitting RF and dropping the lowest ranked SNP. The smallest set of SNPs with the lowest OOB error rate was considered the best set. Although OOB error rate was used to select the best set of SNPs, it cannot be used as the prediction error rate when the iterative fitting of RF is done because this process leads to overfitting, causing downward bias of the prediction error rate (4). The prediction error rate was estimated by the .632+ bootstrap method using 200 bootstrap samples to produce an unbiased estimate of the prediction error (4, 25). In addition, the stability of the selected SNPs was measured by the frequency of their inclusion in the best set of SNPs by each of the 200 bootstrap samples (4). We also conducted logistic regression of case-control status with SNPs using a logadditive model, adjusting for age and gender, to determine P values and compared ranking of SNPs by logistic regression and RF. SNPs ranked higher by RF than logistic regression may suggest interaction. Such interaction between SNPs was subsequently tested by logistic regression. The P value for interaction

was derived by log-likelihood ratio test comparing the full model with interaction terms with the submodel without the interaction terms.

Results

Seventy-one of the 112 (63.4%) cases and 72 of the 112 (64.3%) controls were male. The mean ages for cases and controls were 57.7 and 57.2 years, respectively. A detailed description of subject participation and characteristics has been provided previously (26).

Among the 10 pathways examined, four pathways (DNA repair, mitogen-activated protein kinase signaling, calcium signaling, and transforming growth factor-β pathways) had a prediction error <50% in classifying case-control status (Table 1); however, only the prediction error of the DNA repair pathway had a *P* value of <0.10 (*P* = 0.09).

Of the 57 SNPs in the DNA repair pathway, 17 SNPs were selected as the best set using the selection algorithm described above with a similar prediction error of 45.3, compared with the prediction error of 45.5 when all 57 SNPs were used (Table 2). These 17 SNPs included the top three ranking SNPs from the original RF analysis (rs1047840 of *EXO1*, rs12450550 of *EME1*, and rs799917 of *BRCA1*). These three SNPs were relatively stable with >50% probability of being repeatedly included as part of the best set of SNPs.

Although the logistic regression *P* value for rs799917 of *BRCA1* ranked 34th of the 57 DNA repair SNPs, RF ranked this SNP 2nd, suggesting that rs799917 may be important through its interaction with another SNP. Among those with a GG genotype of rs1047840, those with a TT genotype of rs799917 had a reduced risk of glioblastoma compared with those with a CC genotype (odds ratio, 0.02; 95% confidence interval, 0.001–0.226; Table 3). The *P* value for interaction was 0.0005. Although the nonsynonymous allele variant of rs799917 may have functional effect, it is in linkage disequilibrium with rs1799966 of *BRCA1*, which is located in a region of *BRCA1* known as BRCT. The BRCT homologue in yeast, Brc1, has been shown to interact with *EXO1* and *EME1* in homologous recombination (27). We therefore evaluated the interaction between rs1799966 of BRCA1 (ranked 4th by RF) and rs1047840 of EXO1 and the result was significant with $P_{\text{interaction}} = 0.01$ (Table 3). The interaction between rs1799966 and rs12450550 of *EME1* was also assessed and the result was suggestive of an interaction but not statistically significant (*P* = 0.10).

Discussion

The current RF analysis identified the DNA repair pathway as an important pathway for identifying glioblastoma cases-control status. Three SNPs (rs1047840 of *EXO1*, rs12450550 of *EME1*, and rs799917 of *BRCA1*) of the DNA repair pathway were identified as promising candidates for further replication. In addition, statistically significant interactions between rs1047840 in *EXO1* and rs799917 or rs1799966 in *BRCA1* were observed.

Rs1047840 is a nonsynonymous SNP located in the coding region of EXO1, a double-stranded DNA exonuclease (28). The polymorphism results in a dramatic amino acid change from a negatively charged glutamate to a positively charged lysine residue. This change could potentially have an effect on internal structure or a protein-protein binding interface.

Rs12450550 is a nonsynonymous SNP located in the coding region of EME1 within the crossover junction endonuclease domain. EME1 interacts with MUS81 to form DNA structure-specific endonuclease implicated in DNA repair (29). The resulting amino acid change is, on its own, relatively conservative from an aliphatic isoleucine to a polar threonine; however, this SNP disrupts a potential binding site for the transcription factor MYB according to Delta-MATCH (dif-z score = 0.2; dif-z score predicts the effect on transcription factor site binding due to allele substitution), a computer program that predicts the importance of SNPs in the transcription factor binding sites (30). It is also in linkage disequilibrium with rs3744526 (pairwise Tagger with $r^2 > 0.8$), which enhances a potential binding site for the MSX1 transcription factor (dif-z score = —0.3148; ref. 30).

An attractive feature of the RF method is its ability to account for interaction between SNPs. Lunetta et al. (3) showed that RF has a greater power in detecting important SNPs, when there are SNP-SNP interactions, compared with Fisher's exact test. Among the SNPs we analyzed for the DNA repair pathway (Table 2), both rs799917 and rs1799966 of *BRCA1* did not have a strong main effect but received a high ranking (second and fourth among all 57 DNA repair SNPs, respectively) by the RF. The tests for interaction between rs799917 and rs1047840 and between rs1799966 and rs1047840 by unconditional logistic regression were statistically significant. Rs799917 is a nonsynonymous SNP located in the coding region of *BRCA1*, another critical DNA repair gene. The polymorphism leads to an amino acid change from proline to leucine at position 871 in the BRCA1 protein. This is a nonconservative change as proline conveys unique structural properties to the polypeptide. Furthermore, this polymorphism lies in the middle of a strongly conserved region of the gene as measured by phastCons analysis of >28 species (31). Rs799917 is also in linkage disequilibrium with rs1799966. Notably, rs1799966 is a nonsynonymous SNP in the coding region of the COOH-terminal domain of BRCA1, referred to as BRCT. A recent study has shown that the BRCT homologue in yeast, Brc1, mediates suppression of the *Smc6–74* allele in concert with *Exo1* and *Eme1* (27). This suppression is essential for homologous recombination in processes such as repairing double-stranded breaks in DNA. The interaction between rs1799966 and rs12450550 of *EME1* was suggestive, but not statistically significant, in the current analysis.

In the current analysis, rs9352 of *CHAF1A* was repeatedly included (78% of the 200 bootstrap samples) in the best set of SNPs. A recent study showed that another SNP (rs243356) of *CHAF1A* was associated with glioma risk (32). This suggests that *CHAF1A* may contribute to gliomagenesis and warrants further investigation.

In this study, genes were grouped by pathways for analysis. Analyzing SNP data at a pathway level may have several advantages over analysis of single SNPs or multiple SNPs within a single gene. Because the SNPs were grouped by pathways, the number of comparisons was greatly reduced (227 SNPs versus 10 pathways in the current analysis), decreasing the probability of false positives. In addition, grouping the SNPs by biological pathways allows for a biologically meaningful interpretation of the results. Finally, it is often difficult to replicate the findings of individual SNPs or haplotypes due to different linkage disequilibrium patterns or different allele frequencies (important when the SNPs being studied are not causal) among different study populations. Thus, it may be more feasible and

meaningful to perform replication at the level of biological pathways, although there have been too few studies using this type of pathway analysis to show this.

One must be aware of the several limitations associated with this study: (*a*) Because this was a pilot study, the genotyped SNPs were from a commercially available SNP chip that was not specifically designed for detecting important SNPs associated with glioma. Furthermore, the limited inclusion of genes on the panel precluded us from analyzing some important pathways associated with glioma (e.g., pathways of allergic disorders such as *IL-4* and *IL-13*). In addition, SNPs on the assay panel only included nonsynonymous SNPs, and thus the coverage for variation in each gene was far from complete. Furthermore, even the number of nonsynonymous SNPs in the assay was lower than those on more recent panels. The results of our analysis do depend on the completeness of the genes and SNPs included in the pathway because the function of a gene may depend on its interaction with other genes in the same pathway; therefore, the null findings with some of the pathways examined by this study do not preclude their importance in gliomagenesis. (*b*) The small sample size may have limited the statistical power. (*c*) We did not adjust for genetic ancestry to account for the potential population stratification, although the inclusion of only the Caucasian subjects makes the effect of population stratification less likely (33).

Despite less than complete inclusion of genes and SNPs relevant to glioma and a small sample size for this pilot study, RF analysis was able to identify a potentially important biological pathway that distinguished glioblastoma cases and controls better than chance. By incorporating information on biological pathways and using statistical methods that can account for interaction between genes or SNPs, the power for detecting gene-disease association can be increased.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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.

Grant support: Accelerate Brain Cancer Cure; the National Brain Tumor Foundation; NIH grants CA52689, CA097257, CA89032, ES06717, and ES04705; and Robert J. and Helen H. Glaser Family Foundation. Dr. Jeffrey S. Chang is also supported by the National Cancer Institute R25 CA112355.

References

1. Breiman L Random forests. Machine Learning 2001;45:5–32.
2. Breiman L, Cutler A. 2007 Available from: http://www.math.usu.edu/~adele/forests/.
3. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 2004;5:32. [PubMed: 15588316]
4. Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7:3. [PubMed: 16398926]
5. Pang H, Lin A, Holford M, et al. Pathway analysis using random forests classification and regression. Bioinformatics 2006;22:2028–36. [PubMed: 16809386]

6. Wiemels JL, Wiencke JK, Kelsey KT, et al. Allergy-related polymorphisms influence glioma status and serum IgE levels. Cancer Epidemiol Biomarkers Prev 2007;16:1229–35. [PubMed: 17548690]
7. Wrensch M, McMillan A, Wiencke J, et al. Nonsynonymous coding single-nucleotide polymorphisms spanning the genome in relation to glioblastoma survival and age at diagnosis. Clin Cancer Res 2007;13: 197–205. [PubMed: 17200355]
8. Hardenbol P, Yu F, Belmont J, et al. Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. Genome Res 2005;15:269–75. [PubMed: 15687290]
9. Nebert DW, Dalton TP. The role of cytochrome P450 enzymes in endogenous signalling pathways and environmental carcinogenesis. Nat Rev Cancer 2006;6:947–60. [PubMed: 17128211]
10. McIlwain CC, Townsend DM, Tew KD. Glutathione *S*-transferase polymorphisms: cancer incidence and therapy. Oncogene 2006;25: 1639–48. [PubMed: 16550164]
11. Wood RD, Mitchell M, Lindahl T. Human DNA repair genes, 2005. Mutat Res 2005;577:275–83. [PubMed: 15922366]
12. Lobrich M, Jeggo PA. The impact of a negligent G2/M checkpoint on genomic instability and cancer induction. Nat Rev Cancer 2007;7: 861–9. [PubMed: 17943134]
13. Jana S, Paliwal J. Apoptosis: potential therapeutic targets for new drug discovery. Curr Med Chem 2007;14:2369–79. [PubMed: 17896985]
14. Hood JD, Cheresh DA. Role of integrins in cell invasion and migration. Nat Rev Cancer 2002;2:91–100. [PubMed: 12635172]
15. Witz IP. The selectin-selectin ligand axis in tumor progression. Cancer Metastasis Rev 2008;27:19–30. [PubMed: 18180878]
16. Monteith GR, McAndrew D, Faddy HM, Roberts-Thomson SJ. Calcium and cancer: targeting Ca2+ transport. Nat Rev Cancer 2007;7:519–30. [PubMed: 17585332]
17. Wrensch M, Fisher JL, Schwartzbaum JA, Bondy M, Berger M, Aldape KD. The molecular epidemiology of gliomas in adults. Neurosurg Focus 2005;19:E5.
18. Golestaneh N, Mishra B. TGF-β, neuronal stem cells and glioblastoma. Oncogene 2005;24:5722–30. [PubMed: 16123805]
19. Suzuki T, Maruno M, Wada K, et al. Genetic analysis of human glioblastomas using a genomic microarray system. Brain Tumor Pathol 2004;21:27–34. [PubMed: 15696966]
20. Kanehisa M, Goto S, Hattori M, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 2006;34: D354–7. [PubMed: 16381885]
21. Biocarta Pathway Collections. Available from: http://www.biocarta.com/genes/allPathways.asp.
22. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 2002;31:19–20. [PubMed: 11984561]
23. Pico A, Smirnov I, Chang JS, et al. SNPLogic: an interactive web resource for the pathway-based selection and prioritization of SNPs for genotyping studies. American Association for Cancer Research, ‘‘Approaches to Complex Pathways in Molecular Epidemiology’’ Santa Ana Pueblo, New Mexico, May 30-June 2, 2007.
24. Cutler A, Stevens JR. Random forests for microarrays. Methods Enzymol 2006;411:422–32. [PubMed: 16939804]
25. Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc 1997;92:548–60.
26. Wrensch M, Kelsey KT, Liu M, et al. Glutathione-S-transferase variants and adult glioma. Cancer Epidemiol Biomarkers Prev 2004; 13:461–7. [PubMed: 15006924]
27. Lee KM, Nizza S, Hayes T, et al. Brcl-mediated rescue of Smc5/6 deficiency: requirement for multiple nucleases and a novel Rad18 function. Genetics 2007;175:1585–95. [PubMed: 17277362]
28. Tran PT, Erdeniz N, Symington LS, Liskay RM. EXO1-A multitasking eukaryotic nuclease. DNA Repair (Amst) 2004;3:1549–59. [PubMed: 15474417]
29. Osman F, Whitby MC. Exploring the roles of Mus81–1/Mms4 at perturbed replication forks. DNA Repair (Amst) 2007;6:1004–17. [PubMed: 17409028]

30. Williamson DW, Pico AR, Zambon AC, Conklin BR, Mahley RW. DeltaMATCH: predicting which single nucleotide polymorphisms may create allele-specific transcription factor binding sites (in progress).
31. Available from: http://genome.ucsc.edu/.
32. Bethke L, Webb E, Murray A, et al. Comprehensive analysis of the role of DNA repair gene polymorphisms on risk of glioma. Hum Mol Genet 2008;17:800–5. [PubMed: 18048407]
33. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 2000;92:1151–8 [PubMed: 10904088]

Table 1.

Prediction error rates from RF analyses of glioblastoma cases and controls for SNPs in 10 biological pathways; the San Francisco Bay Area Adult Glioma Study

Pathway name	No. genes	No. SNPs*	Prediction error (%)	*P* †
DNA repair	34	57	45.54	0.09
MAPK signaling	20	24	45.98	0.17
Calcium signaling	22	27	48.21	0.38
TGF-β	18	24	49.11	0.38
WNT signaling	17	17	50.00	NC
Cell cycle	15	24	50.45	NC
Apoptosis	20	27	50.89	NC
Carcinogen metabolism	6	9	54.91	NC
Cell adhesion	31	47	55.36	NC
Interleukin-6	5	6	55.80	NC

Abbreviations: MAPK, mitogen-activated protein kinase; TGF-β, transforming growth factor β. NC, not calculated, because the prediction error was >50%.

*All SNPs were nonsynonymous coding SNPs present on a noncustom ParAllele assay panel.

†RF analyses were done on 100 data sets with permutated disease-SNP association randomly generated from the original data set. *P* value was derived by examining the distribution of prediction errors of these 100 data sets.

Table 2.

The best set of 17 SNPs selected by the varSelRF from 57 SNPs belonging to the DNA repair pathway; the San Francisco Bay Area Adult Glioma Study

Gene symbol	SNPs	Alleles*	No. cases HW/HT/HV	No. controls HW/HT/HV	Ranking in the original RF†	Frequency selected as part of the best SNPs in 200 bootstrap samples	Log-additive model OR (95% CI)‡	Ranking by *P* value from logistic regression
EXO1	rs1047840	A/**G**	55/42/15	29/59/22	1	0.99	0.54 (0.37–0.80)	1
CHAF1A	rs9352	C/**T**	20/57/35	32/48/32	15	0.78	1.29 (0.90–1.87)	10
EME1	rs12450550	C/**T**	51/52/9	68/35/9	3	0.71	1.46 (0.96–2.21)	3
BRCA1	rs799917	C/**T**	51/51/10	55/38/19	2	0.59	0.90 (0.62–1.32)	34
EXO1	rs1776148	A/**G**	44/57/11	46/47/18	10	0.55	0.90 (0.61–1.33)	35
BRCA1	rs16941	A/**G**	51/48/11	55/36/18	6	0.47	0.93 (0.64–1.37)	45
ERCC6	rs2228526	A/**G**	63/43/4	70/28/6	7	0.38	1.26 (0.79–2.01)	9
ATM	rs1801516	A/**G**	92/18/2	76/31/5	5	0.34	0.51 (0.30–0.88)	2
ERCC6	rs2228527	A/**G**	60/48/4	72/33/6	14	0.29	1.32 (0.84–2.09)	15
CCNH	rs2266690	C/**T**	71/34/7	63/45/4	17	0.27	0.87 (0.56–1.37)	31
MSH4	rs5745325	A/**G**	63/38/11	62/43/6	12	0.25	1.11 (0.73–1.68)	37
BRCA1	rs1799966	A/**G**	53/49/9	56/39/17	4	0.24	0.89 (0.60–1.31)	32
POLE	rs5744934	A/**G**	71/36/5	81/28/3	9	0.24	1.43 (0.88–2.33)	6
ERCC6	rs4253211	C/**G**	82/18/0	77/26/0	13	0.19	0.62 (0.31–1.23)	9
POLQ	rs3218651	A/**G**	72/24/2	75/20/2	11	0.16	1.16 (0.65–2.09)	36
POLQ	rs3218634	C/**G**	94/15/2	101/11/0	8	0.005	1.75 (0.82–3.71)	5
POLQ	rs532411	C/**T**	95/15/2	101/11/0	16	0.00	1.73 (0.82–3.67)	7
				Bootstrap (.632+) estimate of prediction error: 45.33%				

NOTE: The best set of SNPs was determined using an iterative process of fitting RF and dropping the lowest ranked SNP. The smallest set of SNPs with the lowest OOB error rate was considered the best set.

Abbreviations: HW, homozygous wild type; HT, heterozygous; HV, homozygous variant. OR, odds ratio; 95% CI, 95% confidence interval.

*Bold letters indicate minor alleles.

†The RF built with all SNPs in the DNA repair pathway.

‡Adjusted for sex and age using unconditional logistic regression, with the major allele as the reference group.

Table 3.

Interaction between SNPs of *BRCA1, EXO1*, and *EME1* on risk of glioblastoma; the San Francisco Bay Area Adult Glioma Study

	No. cases/controls	OR (95% CI)*	No. cases/controls	OR (95% CI)*
BRCA1 rs799917	*EXO1* rs1047840 G/G		*EXO1* rs1047840 A/A or A/G	
CC	30/13	reference	21/40	reference
CT	24/8	1.07 (0.36–3.17)	27/30	1.72 (0.82–3.62)
TT	1/8	0.02 (0.001–0.226)	9/11	1.57 (0.56–4.40)
		$P_{interaction}$ = 0.0005		
BRCA1 rs1799966	*EXO1* rs1047840 G/G		*EXO1* rs1047840 A/A or A/G	
AA	30/13	reference	23/41	reference
AG	22/9	0.88 (0.31–2.54)	27/30	1.63 (0.78–3.39)
GG	2/7	0.06 (0.01–0.41)	7/10	1.25 (0.42–3.75)
		$P_{interaction}$ = 0.01		
BRCA1 rs1799966	*EME1* rs12450550 T/T		*EME1* rs12450550 C/C or C/T	
AA	25/33	reference	28/23	reference
AG	24/22	1.46 (0.66–3.21)	25/17	1.08 (0.46–2.53)
GG	2/13	0.20 (0.04–0.99)	7/4	1.35 (0.34–5.30)
		$P_{interaction}$ = 0.10		

*Adjusted for sex and age using unconditional logistic regression.