Genomics & Informatics Vol. 5(4) 168-173, December 2007

Application of Random Forests to Association Studies Using Mitochondrial Single Nucleotide Polymorphisms

Yoonhee Kim[1,2] and Ho Kim[1]*

[1]Department of Biostatistics and Epidemiology, School of Public Health, Seoul National University, Seoul, 151-742, Republic of Korea, [2]Inherited Disease Research Branch, NHGRI/NIH, Baltimore, MD 20892, USA

Summary

In previous nuclear genomic association studies, Random Forests (RF), one of several up-to-date machine learning methods, has been used successfully to generate evidence of association of genetic polymorphisms with diseases or other phenotypes. Compared with traditional statistical analytic methods, such as chi-square tests or logistic regression models, the RF method has advantages in handling large numbers of predictor variables and examining gene-gene interactions without a specific model. Here, we applied the RF method to find the association between mitochondrial single nucleotide polymorphisms (mtSNPs) and diabetes risk. The results from a chi-square test validated the usage of RF for association studies using mtDNA. Indexes of important variables such as the Gini index and mean decrease in accuracy index performed well compared with chi-square tests in favor of finding mtSNPs associated with a real disease example, type 2 diabetes.

Keywords: association, mtSNPs, Random Forests

Introduction

Mitochondria are double-membrane organelles present in most cells, and play a central role in energy transduction processes of eukaryotic cells including ion homeostasis, intermediary metabolism, and apoptosis (Burger *et al.*, 2003). Mitochondria have their own genetic system called mitochondrial DNA (mtDNA); meanwhile, other extranuclear organelles (nuclear genome) in the cell do not have their own genome. The architecture of mtDNA varies depending on the organism, especially human mtDNA, which has a circular shape of 16.6 kb. Mitochondrial DNA consists of 13 protein-coding, 2 rRNA, and 22 tRNA genes that are involved in five processes: respiration and oxidative phosphorylation, translation, transcription, RNA maturation, and protein import (Burger *et al.*, 2003, Park and Lee, 2004). There are particular features of mitochondrial DNA that are different from nuclear DNA, including haploid number, high copy number, apparent lack of recombination, high substitution rate, and maternal mode of inheritance (Ingman *et al.*, 2000).

Due to the aforementioned features of mtDNA, mtDNA has been used in evolutionary studies (Ladoukakis & Eyre-Walker, 2004) and in association studies of complex diseases such as MELAS syndrome (Mukae *et al.*, 2003; Niemi *et al.*, 2003; Nigou *et al.*, 1998), non-insulin-dependent diabetes mellitus (NIDDM , type 2) (Guo *et al.*, 2005; Kahn *et al.*, 1996; Matsunaga *et al.*, 2001; Ohkubo *et al.*, 2001; Poulton *et al.*, 2002; Suzuki, 2004; Suzuki *et al.*, 2003), and aging (Kato *et al.*, 2002). Association studies of diabetes and mitochondrial DNA variants (mtSNPs) have been widely performed since mitochondria have a central role in ATP production, which is related to insulin production and release, and because type 2 diabetes has a high burden of disease (Cho *et al.*, 2004; Guo *et al.*, 2005; Ohkubo *et al.*, 2001; Poulton *et al.*, 2002; Rosenbloom *et al.*, 1999; Suzuki, 2004). In general, most association studies have been conducted using chi-square tests or logistic regression models in case-control designs.

However, most common complex human diseases have been identified such that multilocus genes under complicated biological mechanisms as well as gene-gene or gene-environment interactions are the causative factors rather than a single gene (Bureau *et al.*, 2005). Traditional association study methods, such as chi-square tests and logistic regression methods in the context of parametric approaches, need a prespecified model using a relatively small number of predictors. But, they are confronted by limitations to deal with detecting such complex diseases using a large number of predictors efficiently. Because of this, in previous genomic and proteomics studies with nuclear DNA, one of several machine learning methods, Random Forests (RF), in the context of the nonparametric approach has been used successfully to generate evidence of association of genetic polymorphisms with diseases or other phenotypes, especially in the presence of gene-gene interactions with a large number

*Corresponding author: E-mail hokim@snu.ac.kr
Tel +1-410-550-7125, Fax +1-410-550-751
Accepted 15 October 2007

of predictor variables (Bureau *et al.*, 2005; Bureau *et al.*, 2003; Diaz-Uriarte & Alvarez de Andres, 2006; McKinney *et al.*, 2006; Shi *et al.*, 2005).

RF (http://www.stat.berkely.edu/~breiman/RandomForests) is a fairly new ensemble method that combines trees grown on bootstrap samples of data and random subset bagging of predictor variables (Breiman, 2001). During randomization of features, RF can provide an importance index of independent variables by calculating accuracy and the Gini index. Furthermore, the importance index has captured the interactions between predictors by randomizations of predictors, and its performance to rank risk SNPs was better than that of univariate tests such as Fisher's exact test when interactions are present (Lunetta *et al.*, 2004). In terms of robustness to outliers and noise, and calculation time, RF is superior to other machine learning methods such as bagging or boosting (Lee *et al.*, 2005).

In previous nuclear genomic studies, we hypothesized that RF is appropriate for association studies using mtSNPs with unique characteristics such as haploid number and lack of recombination, unlike nuclear SNPs. To our knowledge, this paper is the first application of RF to investigate the association of mtSNPs with disease. We validate the usage of RF compared with the results from the chi-square test using example data searching the association between mtDNA and type 2 diabetes.

Example Data

Demographic information of example data is not consented to publish. One hundred thirty unrelated patients with type 2 diabetes and 65 well-matched, unaffected control subjects were used in the analyses. Type 2 diabetes was diagnosed according to World Health Organization criteria. Selection of non-diabetes was based on the following criteria: matched age with diabetes cases, no past history of diabetes, no diabetes in first-degree relatives. One hundred thirty-two mitochondrial biallelic SNPs among whole mtDNA were selected for subsequent analyses using Restriction Fragment Length Polymorphism (RFLP). We used the revised Cambridge Reference Sequence (rCRS, (Chinnery *et al.*, 1999)) to denote which allele is the "common" allele at each mtSNP locus. Interestingly, one of mtSNPs, bp4985, has all variant alleles in type 2 diabetes patient cases and all common alleles in unaffected controls. Even though this finding is inconsistent with previous studies, we included bp4985 in the analysis to see an extreme condition. Data providers do not provide details or data quality information on genotypes and phenotypes. Thus, any biological interpretation and conclusion from our example data are not included for the purpose of this study.

Methods

Since mtDNA is haploid, we do not need to assume any genetic model (e.g., dominant, recessive, or environmental) to produce genotype data. Thus, we can generate a 2 x 2 table of mtSNP versus disease status at the test locus directly. Fisher's exact chi-square tests and univariate logistic regression analyses for calculating odds ratios were performed using SAS v9.1 for each mtSNP site to detect mtSNPs that show different allele frequency distributions between cases and controls.

For RF analysis, we used RF classification tree methods (number of trees =10,000 [computing time; approximately 5 minutes using Windows version]; number of random features at each node = $\sqrt{132}$ = 11) implemented in the Random Forests package v4.5 in R v2.3.1. In RF, trees were grown to the deepest possible level using random 2/3 subsets of the cases and controls and were not pruned. After each tree was grown, the remaining OOB (out-of-bag) cases and controls (remaining 1/3 of the data) were used to estimate the classification error rate of that tree. Once all trees were grown, all data were classified using each tree. The following algorithm demonstrates construction of the "Forests" consisting of classification trees.

1. Draw a bootstrap sample T* consisting of n cases with replacement from the original training data T with n cases (about 2/3 of the data). Remaining data (about 1/3 of the data) are left out for the OOB data and used for the estimation of prediction error.
2. When a classification tree is grown using T* sample,
   2.1 Choose a small number r, which is randomly selected without replacement among R predictor variables (= random subsets of features), and the default of r usually is the square root of the available number of variables (r is the constant over all trees in a "forest").
   2.2 At each node, choose a best predictor (=independent) variable that splits the training sample at that node among the subset of predictor variables selected in previous step 2-1.
3. Iterate steps 2-1 and 2-2 until the tree is fully grown (no pruning).
4. Repeat steps 1 through 3 to construct a tree to yield a forest of pre-determined size.

For each pair of individuals in the data, count the number of trees in which the pair of individuals is in the same terminal node, and divide by the number of trees. This is the "proximity" for this pair of individuals averaged across all trees. Dissimilarity is calculated as 1- proximity for each pair of individuals. Metric scaling projects the dissimilarity from a Euclidean space in a high-dimensional space onto a low-dimensional space. In metric scaling, the first and

second scaling coordinates give useful information about the data. Thus, Multi Dimensional Scaling (MDS) plot, for which the dissimilarity is used as input, is depicted as the graph of the second versus the first coordinates (Shi *et al.*, 2005). Through MDS plots, we can have informative views of the data and evaluate the results of classification intuitively (Fig. 1).

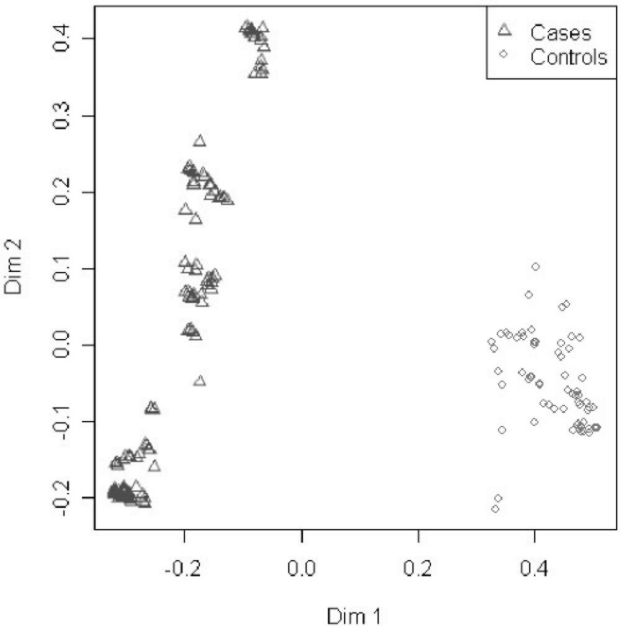

Fig. 1. Multidimensional Scale Plot To check the Random Forest classification results intuitively, the dissimilarities= 1-proximities of each observation are drawn in a two-dimensional Euclidean space. Red triangles: type 2 Diabetes patients (n=130), green circles: unaffected control subjects (n=65). Dimension 1 gave good separation of the observations into two groups.

To measure the importance of predictor variables, the mean decrease in accuracy and Gini index at each node were used. Fig. 2 illustrates the 20 most important variables of each measure. Mean Decrease in Accuracy exploits the margin, defined as the average of (% of votes for true class in the untouched OOB data) - (% of votes for the correct class in the variable-permuted OOB data) over all trees. In other words, the larger the size of the margin, the more important the predictor is. Gini importance is calculated for each variable using the Gini impurity criterion of the resulting subsets of the data at each decision node where the variable was used. Gini impurity is based on the squared probabilities of cases and controls in the two resultant subsets after a split is made using a variable. By definition, the impurity in the resulting subsets must be less than in the parent subset. The Gini index for a given variable is the sum over all trees of the decrease in Gini impurity after each split that involved that variable. We validated

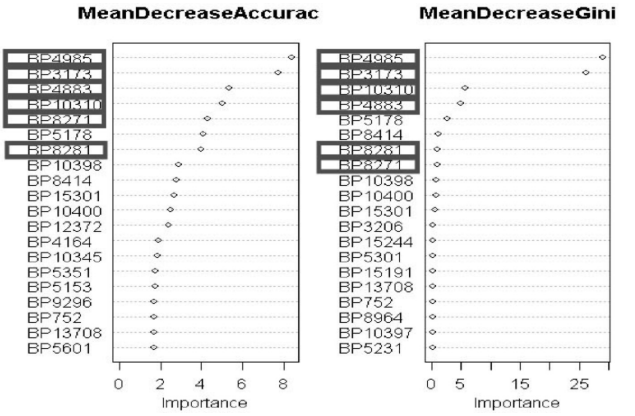

Fig. 2. Variable Importance Plots with Top 20 Variables Left panel contains the 20 most important variables for predicting case-control status descending by Mean Decrease Accuracy (average of (% of votes for true class in the untouched OOB data) - (% of votes for the correct class in the variable-permuted OOB data) over all trees). Right panel contains the 20 most important variables descending by Mean Decrease Gini Index (adding up the Gini decrease for each individual variable over all trees). Y axis: top 20 variables lists, red rectangular mtSNPs: 6 representative mtSNP sites that have p-values less than 0.0001 using Fisher's exact test, X axis: importance indexes (left: mean decrease accuracy, right: mean decrease Gini index).

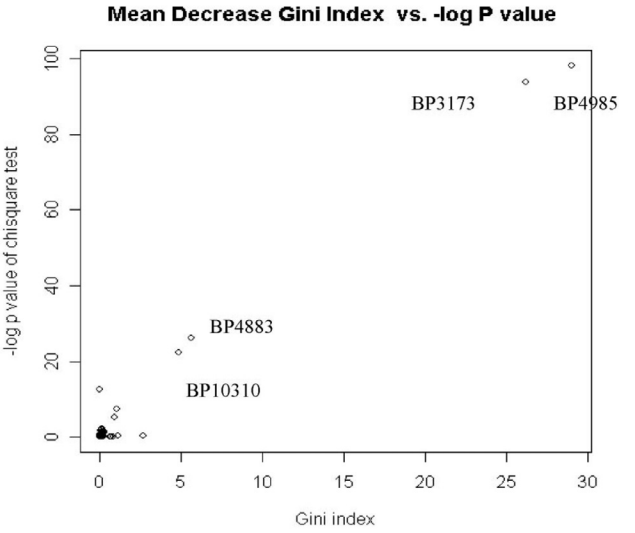

Fig. 3. Validation of results from Fisher's exact test and Random Forests Points might be on the y = x line if both methods (chi-square test and Random Forests classifications) show concordance of results concerning association of mtSNP sites with phenotype. Bp4985 and bp3173 have strong concordance with the Gini index from Random Forests. Bp4883 and bp10310 have moderate concordance with the Gini index from Random Forests. Y-axis indicates the "-log p value" of Fisher's exact test of 132 mtSNP sites. X-axis indicates Gini index as a variable importance measure.

these measures with p-values from the Fisher's exact chi-square tests. We have plotted the negative log p values and the Gini index values (Fig. 3).

Results

To give an intuitive view of the data, the MDS plot shows two distinct point clouds indicated by red triangles as type 2 diabetes patients and green circles as unaffected control subjects in Fig. 1. The first dimension favors splitting 195 subjects into two clusters; positives for controls versus negatives for cases.

To detect which mtSNP associates with disease, Table 1 shows the 6 significant mtSNPs (p values < 0.0001) among 132 mtSNPs using Fisher's exact test and odds ratio estimates from logistic regression. Due to zero cells in the contingency table except for bp10310, odds ratios were not calculated for 5 mtSNPs. Multivariate logistic regression analyses with interaction terms had no noticeable results (data not shown).

In RF, we calculated importance indexes for predicting case-control status both in Mean Decrease Accuracy (left panel in Fig. 2, average of (% of votes for true class in the untouched OOB data) - (% of votes for the correct class in the variable-permuted OOB data) over all trees) and in Mean Decrease Gini index (right panel in Fig. 2, adding up the Gini decrease for each individual variable over all trees). Both the accuracy measure and the Gini index detected the 6 mtSNPs (red boxes), which had significant p-values less than 0.0001 for the Fisher's exact test within the 20 most important variables.

Under the goal of this paper, we validated the RF results compared with Fisher's exact test. We plotted the scatter plot between Gini index from RF and negative p-values from chi-square tests of 132 mtSNP sites in Fig. 3. Points might be on the y = x (diagonal) line if both methods (chi-square test and RF) show concordance of results with respect to association of mtSNP sites with phenotype. Bp4985 and bp3173 were strongly detected as important variables using both methods. Additionally, bp10310 and bp4883 showed concordant results in terms of the strength of significance.

Discussion

The advantage of RF over traditional statistical methods

Table 1. Significant mtSNPs (p value < 0.0001) using Fisher's exact tests

Coding Region	Location	MtSNPs (Variant allele) (Common allele)	Type 2 diabetes (N=130) n (%)	Non-diabetes controls (N=65) n (%)	OR (95% C.I)	* p-values
rRNA	3173	C (+)∫	0 (0%)	63 (96.22%)	N/A+ (N/A)	*1.799E-49
		C	130 (100%)	2 (3.78%)	1.00 (reference)	
ND2	4883	T	31 (23.9%)	0 (0%)	N/A+ (N/A)	*1.312E-06
		C	99 (76.1%)	65 (100%)	1.00 (reference)	
ND2	4985	G	130 (100%)	0 (0%)	N/A+ (N/A)	*2.058E-53
		A	0 (100%)	65 (100%)	1.00 (reference)	
NC	8271	C (-) ∬	17 (13.1%)	0 (0%)	N/A+ (N/A)	*8.771E-04
		C	113 (86.9%)	65 (100%)	1.00 (reference)	
NC	8281	A (-) ∬	0 (0%)	7 (10.77%)	N/A+ (N/A)	*3.649E-04
		A	130 (100%)	58 (89.23%)	1.00 (reference)	
ND3	10310	A	72 (55.4%)	2 (3.1%)	39.10 (9.176-166.63)	*1.198E-14
		G	58 (44.6%)	63 (96.9%)	1.00 (reference)	

+ N/A: not available for calculating OR due to zero in denominators
* p-values from Fisher's exact tests
∫ : Insertion allele at the locus
∬ : Deletion allele at the locus

for association studies—e.g., chi-square test or logistic regression—is the possibility to handle a large number of predictor variables simultaneously, and to examine gene-gene interactions without a specific model. Very recently, in nuclear genomic association studies with hundreds of thousands of predictors, RF has been useful in reducing large amounts of predictors to practical numbers for subsequent analytic steps in the presence of interactions.

In this paper, we investigated the feasibility of RF as a tool for detection of association using the mitochondrial genome. Because mtDNA does not undergo recombination, which may lead to a lack of independence between mtSNP sites, RF methods that do not require strong independence assumptions among predictor variables are particularly applicable to mtDNA markers. As we expected, RF methods performed as well as chi-square tests in terms of consistency of detecting risk mtSNPs, and somewhat superior to logistic regression in terms of ease of modeling complex relationships between the predictors flexibly.

We found bp3173, bp4985, bp5178, bp8414, bp10310, and bp4883 as important determinants in top 20 ranks for classifying 130 diabetes and 65 non-diabetes patients according to the Gini index and the mean decrease in accuracy, similar to what we had observed using the Fisher's exact tests ($p<0.0001$). Separation of the cases from the controls based on the models derived from the RF procedure was quite good. Concerning gene x gene interaction, we chose the simplest interaction model, the two-locus interaction model, for multivariate logistic regression analyses ($_{132}C_2$= 8,646; number of two-locus interaction terms among 132 mtSNPs, data not shown). However, we cannot be sure about only the efficiency of this analysis using never-ending terms in a model, but also the interpretation of results. In contrast, RF gave us the rank of risk SNPs elicited from the complex relationships between them without requesting a model to researchers.

We also applied CART (Grajski *et al.*, 1986) and a logistic regression method to our example data (data not shown); thus, only bp4985 was detected as a predictor because of the perfect contrast of observations between cases and controls. When we performed multivariate logistic regression model including bp4985 to assess the relative effect of each SNP, we could not estimate the effects of other mtSNPs except for bp4985 owing to too much information of only this mtSNP in a regression model. In this case, information on other mtSNPs, which had a relatively small effect to big effect from the only perfect one mtSNP, cannot be provided to researchers. Such minute effects should also be detected since such SNPs might have interactions with other SNPs as a causal factor. In this point of view, RF can prevent the conclusion that only this one mtSNP is an important predictor of diabetes risk, because it ranks the set of important predictors by allocating an importance index to each mtSNP regardless of one big effect. Several of the most important predictor variables can then be studied in independent datasets to further evaluate their importance.

Consequently, we are convinced that synthesizing and summarizing the results from RF machine learning methods can be useful analysis tools for detecting evidence of association of disease risk with mitochondrial DNA variants.

Acknowledgments

This work was partly supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD)(KRF 2005-213-C00007), in part by the Korean Science and Engineering Foundation, and in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

References

Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P., and Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol.* 28, 171-82.

Bureau, A., Dupuis, J., Hayward, B., Falls, K., and Van Eerdewegh, P. (2003). Mapping complex traits using Random Forests. *BMC Genet.* 4 Suppl 1, S64.

Burger, G., Gray, M.W., and Lang, B.F. (2003). Mitochondrial genomes: anything goes. *Trends Genet.* 19, 709-16.

Chinnery, P.F., Howell, N., Andrews, R.M., and Turnbull, D.M. (1999). Clinical mitochondrial genetics. *J Med Genet.* 36, 425-36.

Cho, Y.M., Ritchie, M.D., Moore, J.H., Park, J.Y., Lee, K.U., Shin, H.D., Lee, H.K., and Park, K.S. (2004). Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia.* 47, 549-54.

Diaz-Uriarte, R., and Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics.* 7, 3.

Grajski, K. A., Breiman, L., Viana Di Prisco, G., and Freeman, W.J. (1986). Classification of EEG spatial patterns with a tree-structured methodology: CART. *IEEE Trans Biomed Eng.* 33, 1076-86.

Guo, L.J., Oshida, Y., Fuku, N., Takeyasu, T., Fujita, Y., Kurata, M., Sato, Y., Ito, M., and Tanaka, M. (2005). Mitochondrial genome polymorphisms associated with type-2 diabetes or obesity. *Mitochondrion.* 5, 15-33.

Ingman, M., Kaessmann, H., Paabo, S., and Gyllensten, U.

(2000). Mitochondrial genome variation and the origin of modern humans. *Nature* 408, 708-13.

Kahn, C.R., Vicent, D., and Doria, A. (1996). Genetics of non-insulin-dependent (type-II) diabetes mellitus. *Annu Rev Med.* 47, 509-31.

Kato, Y., Miura, Y., Inagaki, A., Itatsu, T., and Oiso, Y. (2002). Age of onset possibly associated with the degree of heteroplasmy in two male siblings with diabetes mellitus having an A to G transition at 3243 of mitochondrial DNA. *Diabet Med.* 19, 784-6.

Ladoukakis, E.D., and Eyre-Walker, A. (2004). Evolutionary genetics: direct evidence of recombination in human mitochondrial DNA. *Heredity.* 93, 321.

Lee, J.W., Lee, J.B., Park, M., and Song, S.H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Comp Stat & Data Analysis.* 48, 869-885.

Lunetta, K.L., Hayward, L.B., Segal, J., and Van Eerdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.* 5, 32.

Matsunaga, H., Tanaka, Y., Tanaka, M., Gong, J.S., Zhang, J., Nomiyama, T., Ogawa, O., Ogihara, T., Yamada, Y., Yagi, K., and Kawamori, R. (2001). Antiatherogenic mitochondrial genotype in patients with type 2 diabetes. *Diabetes Care.* 24, 500-3.

McKinney, B.A., Reif, D.M., Ritchie, M.D., and Moore, J.H. (2006). Machine learning for detecting gene-gene interactions: a review. *Appl Bioinformatics.* 5, 77-88.

Mukae, S., Aoki, S., Itoh, S., Sato, R., Nishio, K., Iwata, T., and Katagiri, T. (2003). Mitochondrial 5178A/C genotype is associated with acute myocardial infarction. *Circ J.* 67, 16-20.

Niemi, A.K., Hervonen, A., Hurme, M., Karhunen, P.J., Jylha, M., and Majamaa, K. (2003). Mitochondrial DNA polymorphisms associated with longevity in a Finnish population. *Hum Genet.* 112, 29-33.

Nigou, M., Parfait, B., Clauser, E., and Olivier, J.L. (1998). Detection and quantification of the A3243G mutation of mitochondrial DNA by ligation detection reaction. *Mol Cell Probes.* 12, 273-82.

Ohkubo, K., Yamano, A., Nagashima, M., Mori, Y., Anzai, K., Akehi, Y., Nomiyama, R., Asano, T., Urae, A., and Ono, J. (2001). Mitochondrial gene mutations in the tRNA(Leu(UUR)) region and diabetes: prevalence and clinical phenotypes in Japan. *Clin Chem.* 47, 1641-8.

Park, H.S., and Lee, S.U. (2004). MitGEN: Single Nucleotide Polymorphism DB Browser for Human Mitochondrial Genome. *Genomics & Informatics* 2(3), 147-148.

Poulton, J., Luan, J., Macaulay, V., Hennings, S., Mitchell, J., and Wareham, N.J. (2002). Type 2 diabetes is associated with a common mitochondrial variant: evidence from a population-based case-control study. *Hum Mol Genet.* 11, 1581-3.

Rosenbloom, A.L., Joe, J.R., Young, R.S., and Winter, W.E. (1999). Emerging epidemic of type 2 diabetes in youth. *Diabetes Care.* 22, 345-54.

Shi, T., Seligson, D., Belldegrun, A.S., Palotie, A., and Horvath, S. (2005). Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Mod Pathol,* 18, 547-57.

Suzuki, S. (2004). Diabetes mellitus with mitochondrial gene mutations in Japan. *Ann N Y Acad Sci.* 1011, 185-92.

Suzuki, S., Oka, Y., Kadowaki, T., Kanatsuka, A., Kuzuya, T., Kobayashi, M., Sanke, T., Seino, Y., and Nanjo, K. (2003). Clinical features of diabetes mellitus with the mitochondrial DNA 3243 (A-G) mutation in Japanese: maternal inheritance and mitochondria-related complications. *Diabetes Res Clin Pract.* 59, 207-17.