Bagging Nearest-Neighbor Prediction independence Test: an efficient method for nonlinear dependence of two continuous variables

Yi Wang1, Yi Li3, Xiaoyu Liu1, Weilin Pu1, Xiaofeng Wang2, Jiucun Wang2, Momiao Xiong1,3, Yin Yao Shugart1,4 & Li Jin2

Testing dependence/correlation of two variables is one of the fundamental tasks in statistics. In this work, we proposed an efficient method for nonlinear dependence of two continuous variables (X and Y). We addressed this research question by using BNNPT (Bagging Nearest-Neighbor Prediction independence Test, software available at https://sourceforge.net/projects/bnnpt/). In the BNNPT framework, we first used the value of X to construct a bagging neighborhood structure. We then obtained the out of bag estimator of Y based on the bagging neighborhood structure. The square error was calculated to measure how well Y is predicted by X. Finally, a permutation test was applied to determine the significance of the observed square error. To evaluate the strength of BNNPT compared to seven other methods, we performed extensive simulations to explore the relationship between various methods and compared the false positive rates and statistical power using both simulated and real datasets (Rugao longevity cohort mitochondrial DNA haplogroups and kidney cancer RNA-seq datasets). We concluded that BNNPT is an efficient computational approach to test nonlinear correlation in real world applications.

Dependence is any statistical relationships between two random variables and correlation describes any kind of the statistical relationships including dependence. In practice, correlations can be used to predict any potential relationships of interest. The Pearson correlation coefficient appears to be the most commonly used method for assessing correlation. However, the Pearson correlation is sensitive to linear correlations, while several other methods are more robust to detect the non-linear correlations1–3. Testing linear/nonlinear dependence of two variables is one of the fundamental tasks in statistics.

The Pearson correlation (or Pearson's r), first proposed by Karl Pearson and Francis Galton4–8, is a measure of the correlation between two random variables (X and Y). It assigns a value that varies from -1 to 1. The correlation between the two variables is defined as the product of their covariance divided by their standard deviation. Although the Pearson correlation coefficient is often used, the Pearson's r of sample statistic is not distributionally robust (non-normal distribution)9 and its values may be misleading when there are outliers10,11.

The Spearman correlation coefficient (or Spearman's rho) is a nonparametric statistical method of measuring the statistical association between two variables. It evaluates the process during which two variables can be described by monotonic functions. The Spearman correlation coefficient is defined as the Pearson correlation coefficient that is applied to the ranks of the data.

1Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China. 2State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China. 3Human Genetics Center, School of Public Health, University of Texas Houston Health Sciences Center, Houston, Texas, USA. 4Unit on Statistical Genomics, Division of Intramural Division Programs, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA. Yi Wang, Yi Li and Xiaoyu Liu contributed equally to this work. Correspondence and requests for materials should be addressed to Y.Y. (email: yin.yao@nih.gov) or L.J. (email: lijin@fudan.edu.cn)
coefficient of the rank variable\(^3\). The Kendall rank correlation coefficient (or Kendall’s tau coefficient), proposed by Maurice Kendall in 1938, is another nonparametric statistical method to test the correlations of two variables\(^4\). And there is no assumption made about the distribution of \(X\) and \(Y\) or \((X, Y)\).

Other commonly used statistical methods to evaluate the correlations of two random variables include distance correlation, Hoeffding’s independence test, maximal information coefficient (MIC), Hilbert-Schmidt Independence Criterion (HSIC) and so on. The distance correlation is a statistical method of measuring statistical dependence of two random variables. The distance correlation coefficient is zero if and only if the two random variables are statistically independent. The distance correlation coefficient was proposed by Gabor J Szekely (2005), which solves the deficiency of Pearson correlation coefficient (Pearson’s \(r\) can be zero for dependent variables). When the Pearson correlation coefficient is 0, it indicates linearly irrelevant but does not imply independence, whereas the distance correlation is 0 if and only if the random variables are statistically independent\(^5,6\). Hoeffding’s independence test, named after Wassily Hoeffding, is a measure of group deviation. Hoeffding derived an unbiased estimate of \(H\), which can be used to test the independence of the two variables. This test can only be applied to continuously distributed dataset. A sample-based version of this measure was discussed under the null distribution\(^7,8\). MIC is an established method to measure the linear or non-linear correlation between two random variables. MIC belongs to the nonparametric statistical method based on the maximal information theory\(^9\). MIC uses binning to apply mutual information to continuous random variables and MIC is an approach for selecting the number of bins and finding a maximum over possible grids. HSIC (Gretton et al. year) was an independence criterion based on the eigen-spectrum of covariance operators in reproducing kernel Hilbert spaces (RKHSs), consisting of an empirical estimate of the Hilbert-Schmidt Independence Criterion\(^10,11\). HHG (proposed by Heller et al.) is a powerful test that is applicable to all dimensions, consistent against all alternatives and is easy to implement\(^12\).

We had previously developed a new algorithm called continuous variance analysis (CANOVA)\(^13\), the idea came from the analysis of variance (ANOVA) of continuous response with a categorical factor\(^14\). In the CANOVA framework, we first define a neighborhood of each data point according to its \(X\) value, and then calculate the variance of the \(Y\) value within the neighborhood, and finally use a permutation test to assess the significance of the observed “within neighborhood variance”\(^13\). CANOVA is an efficient method in case of non-linear correlation, especially when the function is highly oscillating.

In the current study, we proposed a new nonlinear dependence measure method: Bagging Nearest-Neighbor Prediction Independence Test (BNNPT). BNNPT is based on a permutation test of the square error (SE) of bagging nearest neighbor estimator. In pattern recognition, the k-Nearest Neighbors algorithm (or simply k-NN) is a nonparametric approach for classification and regression\(^15\). The optimal choice of \(k\) depends on the distribution of the data. Typically, large \(k\) values may reduce the effect of noise on the classification\(^16\), but the boundaries between the classes are less distinct. The special case where the class is predicted to be the class of the closest training sample (\(k = 1\)) is called the nearest neighbor algorithm. Bagging, also known as “bootstrap aggregation”, is an algorithm for machine learning aggregators that aims to improve the stability and accuracy of the machine learning algorithms. Furthermore, it reduces the variance, decreasing the possibility of over-fitting. Bagging is a special case of the model averaging method. On the other hand, it can slightly reduce the performance of stabilization methods, such as K-nearest neighbors\(^17\).

In the BNNPT framework, we first used the value of \(X\) to construct a bagging neighborhood structure. And then, we got the out of bag estimator of \(Y\) based on the bagging neighborhood structure. The square error (SE) was calculated to measure how good \(Y\) is predicted by \(X\). Finally, a permutation test was applied to detect the significance of the observed square error. We compared the false positive ratio\(^18\) and statistical power\(^19\) of BNNPT with seven other common correlation coefficient algorithms in simulation study. Furthermore, we compared their performance in a real Rugao longevity cohort (mitochondrial DNA haplogroups)\(^20\) and a kidney cancer RNA-seq (transcriptome sequencing) data set\(^21,22\).

**Methods**

**Summary.** The main framework of BNNPT is based on a permutation test\(^23\) of the square error of a bagging nearest neighbor estimator. For two vectors \(X\) and \(Y\) of length \(N\), we first construct a bagging neighborhood structure based on \(X\) only. The neighborhood structure is an index matrix of \(N\) rows and \(K\) (number of bags) columns. The element \(X_{\text{Neighborhood}(i,j)}\) is defined as the \(j\)th bag’s nearest neighbor of \(X_i\), Neighborhood(\(i, j\)) = \(i\). The element of Neighborhood(\(i, j\)) is sampled as follow: we draw a bag of \(m\) \(Y\) values from \(X\), and choose the one \(X_{\text{Neighborhood}(i,j)}\) that is nearest. When neighborhood structure is available, we were able to construct a bagging nearest neighbor estimator of each \(Y_j\):

\[
H_i = \frac{\sum_{j \in \text{Neighborhood}(i,j)} Y_j}{\text{bags}}
\]

The square error \(SE = \|H - Y\|\) indicates how well \(Y\) is predicted by \(X\). To assess the statistical significance level, a permutation test is conducted using \(SE\) as the test statistics. We randomly shuffle \(Y\) many times and count the probability that \(SE_{\text{random}} < SE\), which is reported as the \(p\)-value.

Denote \(SE_{\text{BNNPT}}(X, Y, b, m)\) as the squared root of the residual of bagging nearest neighbor estimator using \(X\) as a predictor and \(Y\) as response and \(b\) (number of bagging), \(m\) (mtry) as parameters of BNNPT. Our null hypothesis is the following:

\[
SE_{\text{BNNPT}}(X, Y, b, m) = SE_{\text{BNNPT}}(X, Y', b, m), Y' \sim \text{Random shuffle}(Y)
\]
Pseudocode for BNNPT.

Input: two data vector X and Y, both are of length N.
Parameter: bags, mtry (default $= \sqrt{N}$), permutations.

Algorithm

1: Construct neighbourhood structure of X:
   for each $X_i$
      for (n = 1 to bags)
         Nearest = 0
         for (j = 1 to mtry)
            k = random_uniform(1, N), $k \neq i$
            if($|X_k - X_i| < |X_{Nearest} - X_i|$) Nearest = k
         Neighbourhood[i][n] = Nearest
   }

2: Calculate observed square error:
   for (i = 1 to N) $H_i =$ sum($Y_{Neighbourhood[i][j]}$, j)/bags
   obs_se = $\|H - Y\|_2$

3: Permutation test:
   less_equal = 0
   for (n = 1 to permutations)
      $Y'$ = Fisher–Yates shuffle(Y)
      for (i = 1 to N) $H_i =$ sum($Y'_{Neighbourhood[i][j]}$, j)/bags
      ran_se = $\|H - Y'\|_2$
      if(ran_se <= obs_se) less_equal = less_equal + 1;
   }

4: Return p-value:
   p-value = less_equal/permutations

---

| N = 50, X ~ U(−1, 1) | BNNPT | Pearson | Spearman | Kendall |
|-----------------------|-------|---------|----------|---------|
| $y = y + N(0, 1)$     | 0.050 | 0.058   | 0.053    | 0.055   |
| $y = x + N(0, 1)$     | 0.839 | 0.958   | 0.951    | 0.950   |
| $y = 0.5(x + 1)^2 + N(0, 1)$ | 0.861 | 0.961   | 0.949    | 0.946   |
| $y = \sin(\pi x) + N(0, 1)$ | 0.957 | 0.937   | 0.912    | 0.904   |
| $y = \sin(3\pi x) + N(0, 1)$ | 0.795 | 0.180   | 0.079    | 0.073   |
| $y = \cos(\pi x) + N(0, 1)$ | 0.947 | 0.066   | 0.065    | 0.151   |
| $y = \cos(2\pi x) + N(0, 1)$ | 0.888 | 0.060   | 0.072    | 0.109   |
| $y = \cos(3\pi x) + N(0, 1)$ | 0.707 | 0.064   | 0.072    | 0.109   |

Table 1. Simulation power in nine sample functions. The bold means the first place result of all methods compared.
Simulation study. Nine simple functions (including constant function, linear function, quadratic function, sine function and cosine function) were simulated. Additionally, we added Gaussian noise (mean $= 0$, Gaussian variance $= 1$) to $Y$ in these nine simple functions, as shown in Table 1. In the simulation data, we set different Gaussian noise levels (mean $= 0$, Gaussian variance $= 1/9, 1/4, 4$ and $9$) and reported the power across noise levels (shown in Supplemental Materials 1). We selected seven algorithms as the benchmarks: Pearson correlation coefficient, Spearman rank correlation coefficient, Kendall rank correlation coefficient, Distance correlation, Hoeffding’s independence test, MIC and CANOV A. One thousand sets of simulations were carried out to calculate the false positives rate and the statistical power. Two different sample sizes were selected ($N = 50$ and $760$), $x$ as the independent variable which was uniformly distributed in $(-1, 1)$ and $y$ as the dependent variable (shown in Supplemental Materials 1). Notably, MIC has a bias/variance parameter (the ‘alpha’ parameter in the minerva implementation): the maximal allowed resolution of any grid $^{17}$. Reshef et al. also reported that different parameter setting ($\alpha = 0.55, c = 5$) is faster than the default setting and does not significantly affect performance $^{31}$. For simplicity, the default parameters of the MIC ($\alpha = 0.6, c = 15$) was used in this work.

### Table 1

| mtDNA haplogroup | BNNPT | Pearson | Spearman | Kendall | Hoeffding* | Distance | Canova |
|------------------|-------|---------|----------|---------|------------|----------|--------|
| D                | 0.998 | 0.423   | 0.567    | 0.567   | 1.000      | 0.421    | 0.541  |
| D4               | 0.655 | 0.175   | 0.358    | 0.357   | 1.000      | 0.162    | 0.486  |
| D4a              | 0.519 | 0.809   | 0.888    | 0.888   | 1.000      | 0.951    | 0.485  |
| D4b              | 0.568 | 0.647   | 0.784    | 0.784   | 1.000      | 0.786    | 0.482  |
| D4b2             | 0.981 | 0.376   | 0.449    | 0.449   | 1.000      | 0.419    | 0.508  |
| D4b2b            | 0.799 | 0.580   | 0.548    | 0.548   | 1.000      | 0.728    | 0.426  |
| D5               | 0.188 | 0.568   | 0.694    | 0.694   | 1.000      | 0.782    | 0.502  |
| M12              | 0.907 | 0.739   | 0.605    | 0.605   | 1.000      | 0.888    | 0.527  |
| G                | 0.303 | 0.933   | 0.723    | 0.723   | 1.000      | 0.943    | 0.507  |
| G2               | 0.149 | 0.161   | 0.232    | 0.232   | 1.000      | 0.261    | 0.529  |
| M7               | 0.957 | 0.961   | 0.994    | 0.994   | 1.000      | 0.947    | 0.500  |
| M7b              | 0.619 | 0.705   | 0.992    | 0.992   | 1.000      | 0.806    | 0.512  |
| M8               | 0.963 | 0.863   | 0.851    | 0.851   | 1.000      | 0.368    | 0.528  |
| M8a              | 0.447 | 0.397   | 0.565    | 0.565   | 1.000      | 0.146    | 0.455  |
| C                | 0.246 | 0.513   | 0.583    | 0.583   | 1.000      | 0.713    | 0.501  |
| M9               | 0.541 | 0.030   | 0.054    | 0.054   | 1.000      | 0.016    | 0.433  |
| M10              | 0.347 | 0.793   | 0.963    | 0.963   | 1.000      | 0.866    | 0.503  |
| N9               | 0.313 | 0.024   | 0.060    | 0.060   | 1.000      | 0.037    | 0.435  |
| N9a              | 0.352 | 0.084   | 0.193    | 0.193   | 1.000      | 0.130    | 0.471  |
| A                | 0.029 | 0.371   | 0.530    | 0.530   | 1.000      | 0.532    | 0.484  |
| F                | 0.224 | 0.113   | 0.065    | 0.065   | 1.000      | 0.170    | 0.434  |
| F1               | 0.442 | 0.239   | 0.127    | 0.127   | 1.000      | 0.280    | 0.466  |
| B                | 0.180 | 0.388   | 0.368    | 0.368   | 1.000      | 0.451    | 0.544  |
| B5               | 0.656 | 0.201   | 0.524    | 0.524   | 1.000      | 0.188    | 0.501  |
| B5a              | 0.321 | 0.189   | 0.653    | 0.653   | 1.000      | 0.177    | 0.547  |
| B5b              | 0.709 | 0.654   | 0.740    | 0.740   | 1.000      | 0.479    | 0.508  |
| B4a              | 0.012 | 0.097   | 0.086    | 0.086   | 1.000      | 0.109    | 0.499  |
| B4b              | 0.746 | 0.540   | 0.833    | 0.833   | 1.000      | 0.544    | 0.391  |

Table 2. The p-value comparison of benchmarked methods in Rugao longevity cohort data. The significant (significance level $= 0.05$) p-values of methods were marked in bold. The genotype data X (28 mitochondrial haplogroups data) were drawn from a discontinuous distribution, Hoeffding’s independence test may have a defect for discontinuous distributions.

| Kidney cancer dataset | BNNPT | Pearson | Spearman | Kendall | Hoeffding* | Distance | CANOV A | MIC |
|-----------------------|-------|---------|----------|---------|------------|----------|---------|-----|
| The number of unique genes (reported in pubmed) | 387 (15) | 15 (1) | 41 (1) | 0 (0) | 0 (0) | 120 (1) | 8 (1) | 3 (0) |
| Significant number | 10617 | 8239 | 11629 | 11569 | 4953 | 10946 | 5901 | 8081 |
| Computing time (seconds) | 80* | 0.0023 | 0.0025 | 0.0082 | 1.8 | -5000 | 20 | 0.027 |

Table 3. Comparison of computing time and detected significant genes numbers of all methods in kidney cancer dataset (the significance level $\alpha = 2.435e-06$). In order to compare the computing time, the number of permutations of BNNPT is set to 10,000,000 times. If the number of permutations of BNNPT is set to 100,000 times, it only needs 1 seconds. The bold means the first place results of all methods compared. The Computing time was recorded between 1 gene and 604 samples.
Applications on Rugao longevity cohort dataset. We compared the BNNPT algorithm with the other seven algorithms using a real Rugao longevity cohort for mitochondrial DNA haplogroups, which included 1852 samples (463 exceptional longevity samples, 926 elder sampled, 463 middle-aged samples) and 28 major mitochondrial haplogroups\(^6\). The samples with missing values were omitted (remained 1835 samples).

The level of correlations between genotype data \(X\) (28 mitochondrial haplogroups data) and phenotype data \(Y\) (ages) were tested. For simplicity, the other algorithms were applied the default parameters (especially for MIC, \(\alpha = 0.6, c = 15\)). The \(p\) value results and comparisons are shown in Table 2. The significance level was preset to be 0.05.

Applications on kidney cancer dataset. We also compared the BNNPT algorithm with the other seven algorithms using a real RNA-seq dataset for kidney cancer, which included 604 samples (532 cancer samples, 72 normal samples) and 20531 genes\(^7,8\).

The level of correlations between genotype data \(X\) (20,531 gene expression data) and phenotype data \(Y\) (whether kidney cancer or not) were evaluated. The computing time of each algorithm was also compared. The significance level is preset to be 2.435e-06 (Bonferroni correction). For simplicity, the other algorithms were applied the default parameters (especially for MIC, \(\alpha = 0.6, c = 15\)). The results and comparisons are shown in Table 3.

Results

Results from simulation study. It can be seen that when the constant function \((y = 0)\) was used, we compared the false positive rate of the different methods at the significance level of 0.05 in Table 1. Pearson correlation coefficient, Spearman's rank correlation coefficient, Kendall's rank correlation coefficient, Distance correlation coefficient, CANOVA, MIC and BNNPT, all showed a false positive rate around 0.05. It does mean that the Type I error rate was adequately controlled. However, the false positive rate of the Hoefding's independence test was slightly higher than 0.05. Therefore, it is crucial to note that under settings similar to the simulation study, the Hoefding's method led to more false positives than the other methods.

For the comparison of the statistical power of other non-constant functions in the simulation data, we observed the following in Table 1: (1) In case of linear correlations, the Pearson correlation coefficient is the most powerful method, BNNPT is less powerful than Pearson correlation coefficient, but does not fail (power > 0.5); (2) In the case of non-linear correlation, BNNPT appeared to be most powerful, when the function is highly oscillatory/nonlinear, its power is higher than other methods. (3) BNNPT is more powerful than the MIC algorithm in all cases.

By comparing the non-constant correlations shown in Supplemental Materials 1: We concluded that: (1) When the Gaussian noise level is low (Gaussian variance \(= 1/9, 1/4\)), most of the methods have a high power, especially in simple linear relationships. But BNNPT has a higher power in most non-constant functions, especially in non-linear functions. (2) When the Gaussian noise level is high (e.g. Gaussian variance \(= 4, 9\), most methods had much lower power while BNNPT achieved better power than other methods in complex sine/cosine functions. (3) When the sample size is larger (N = 760), BNNPT still achieved better power than other methods in complex sine/cosine functions. However, Pearson's correlation coefficient is more powerful in the simple linear
functions. Therefore, when the relationship between the two random variables is linear, we recommend the use of the Pearson correlation coefficient to obtain higher statistical power. When the relationship is nonlinear or complicated, BNNPT is a good choice to explore the correlation structure of the data.

**Results from the Rugao longevity cohort dataset.** The p-value comparison for the Rugao longevity cohort is shown in Table 2. It indicated that BNNPT detected two mtDNA haplogroups, haplogroup A and haplogroup B4a (P value < 0.05). Pearson correlation coefficient detected two mtDNA haplogroups: haplogroup M9 and haplogroup N9 (P value < 0.05). Distance also detected two mtDNA haplogroups, the same two as Pearson. All BNNPT and CANOVA results were realized in the C++ environment and the other six benchmarks were calculated using the R packages 'energy', 'Hmisc' and 'minerva'. All BNNPT results were calculated in parallel (fully using all 8 CPU cores) on a desktop PC, equipped with an AMD FX-8320 CPU and 32 GB memory. In addition, all of the R code was computed in parallel through an R package named 'snow'.

Literature review for validation of each haplogroup was then performed in the pubmed database. In one Japanese population, the mitochondrial haplogroups A confers a significant risk for coronary atherosclerosis which is a kind of age-related disease. B4a was reported that has negatively correlated with ages in Rugao population. Haplogroup M9 and haplogroup N9 were reported to be related to longevity.

**Results from the kidney cancer study.** The comparison and computing time for kidney cancer dataset is shown in Table 3. In order to compare the computing time, the number of permutations of BNNPT is set to be 10,000,000 times (Table 3). In Supplemental Materials 2, we provided genes that were only detected by the BNNPT method (that was not detected by other methods). For comparison, we also listed genes that can only be detected by other methods in Supplemental Materials 3. All BNNPT and CANOVA analyses were conducted in the C++ environment and the other six benchmarks were calculated using the R packages.
We observed that the Spearman correlation coefficient can detect the most number of significant genes (11629 genes, $\alpha = 2.435 \times 10^{-06}$, in Table 3) in real kidney cancer RNA-seq data. The BNNPT method detects slightly less (10617 genes) than Spearman's correlation coefficient.

Figure 2. The scatterplot and probability density distribution of UGT1A9 gene expression (reported significant genes detected only by CANOVA) between kidney-cancer and normal groups.

Figure 3. The scatterplot and probability density distribution of HDAC1 gene expression (reported significant genes detected only by Pearson) between kidney-cancer and normal groups.

We observed that the Spearman correlation coefficient can detect the most number of significant genes (11629 genes, $\alpha = 2.435 \times 10^{-06}$, in Table 3) in real kidney cancer RNA-seq data. The BNNPT method detects slightly less (10617 genes) than Spearman's correlation coefficient. To explore the biological relevance of the detected genes
Figure 4. The scatterplot and probability density distribution of UPK3A gene expression (reported significant genes detected only by Spearman) between kidney-cancer and normal groups.

Figure 5. The scatterplot and probability density distribution of SLC26A9 gene expression (reported significant genes detected only by Distance) between kidney-cancer and normal groups.
CDCP1 expression has been correlated with poor prognosis in renal cancer\(^3^9,4^0\). Emerging evidences suggest (CDCP1) is an integral membrane protein whose expression is up-regulated in various cancer types. And high \(\alpha\) relevant to renal cancer or disease: CDCP1, GSTT1, E2F3, MAPK1, SALL4, SIRT1, ADAMTS13, Gfr1, ASPH, in Supplemental Materials 3 and the genes reported in pubmed are showed in Figs 2, 3, 4 and 5.

There is only one gene associated with polycystic kidney disease\(^5^4,5^5\). SALL4 is a zinc finger function, meaning that the variances of the two distributions are different. BNNPT is still capable of distinguishing between kidney cancer and normal groups under complex distributions, such as the bimodal distribution in the renal function\(^4^9\). The APOE gene has been reported to be indirectly associated with chronic kidney disease. Knockout of APOE causes hypercholesterolemia, which in turn leads to chronic kidney disease\(^4^6\). Mutations in the renal cancer or disease: CDCP1, GSTT1, E2F3, MAPK1, SALL4, SIRT1, ADAMTS13, Gfr1, ASPH, in Supplemental Materials 3, Fig. 4. It has been reported that Class I HDAC isoforms 1 and 2 are highly expressed in renal cell cancer\(^5^8\). The only unique gene detected by Distance (also reported in pubmed) is SLC26A9 (identified in Supplemental Materials 3, Fig. 5), which was reported to be associated with vesico-ureteral reflux (VUR), which resulted in 8.5% of end-stage renal disease in children\(^5^9\). The only unique gene detected by Spearman (also reported in pubmed) is UPK3A (identified in Supplemental Materials 3, Fig. 4). It has been reported to be associated with vesico-ureteral reflux (VUR), which resulted in 8.5% of end-stage renal disease in children\(^5^9\). The only unique gene detected by Distance (also reported in pubmed) is SLC26A9 (identified in Supplemental Materials 3, Fig. 5), which was reported to be associated with renal disease. SLC26A9 plays an important role in maintaining acid-base balance in renal tubules and nephrons as a chloride ion exchanger\(^6^0\). MIC didn't find unique genes that were previously reported. Hoeffding's independence test and Kendall's rank correlation coefficients did not detect any unique significant genes.

**Discussion**

Longevity is a multifactorial trait with a genetic contribution and mitochondrial DNA (mtDNA) polymorphisms were found to be involved in the phenomenon of longevity. In an autopsy study of 1,536 patients in Japanese elderly, haplogroups A and M7a were significantly associated with coronary atherosclerosis, with odds ratios (95% confidence intervals) of 1.80 (1.09-2.97; \(p = 0.023\)) and 1.92 (1.23-3.01; \(p = 0.004\)) respectively\(^2^7\). In the study of a population-based case-control study in a Chinese Han population residing in Rugao, Jiangsu Province, a significantly decreasing trend of B4a frequency was observed from middle-aged subjects (4.2%), elderly subjects (3.8%) and longevity subjects (1.7%) in females (\(p = 0.045\)). What's more, significant reduction of M9 haplogroups was observed in longevity subjects (0.2%) when compared with both elderly subjects (2.2%) and middle-aged subjects (1.7%). Linear-by-linear association test revealed a significant decreasing trend of N9 frequency from middle-aged subjects (8.6%), elderly subjects (7.2%) and longevity subjects (4.8%) (\(p = 0.018\))\(^2^7\).

Among all the benchmarked methods, BNNPT detected a unique set of genes (15 genes) related to renal cancer or renal diseases in Pubmed database. It was reported that CDCP1 is a unique HIF-2α target gene involved in the regulation of cancer metastasis and suggest that CDCP1 is a biomarker and potential therapeutic target for metastatic cancers\(^6^2\). GSTT1 null genotype is a risk factor for patients with more primitive urologic malignancies.
(bladder, prostate and kidney) and it is more frequent in patients with multiple urologic tumors. Clinical trials have shown that E2F3 is overexpressed in advanced clear cell renal cell carcinoma (ccRCC), and there are multiple E2F3 binding sites in the promoter of HIF-2a. Thus, targeting E2F3-HIF-2a interactions may be a promising treatment procedure for ccRCC. The SNP rs743409 in MAPK1 is a variant of miRNA binding site single nucleotide polymorphisms (SNPs). Under the additive model, the variants were reduced with a 10% risk, indicating that there is a correlation between the miRNA binding site SNP and the RCC risk in the VHL-HIF1 alpha pathway. SALL4 is a zinc finger transcription factor that plays an important role in kidney development, and SALL4 mutation causes kidney deformity. SIRT1 was identified as a direct target for miR-22, and miR-22 might act as a tumor suppressor in RCC and blocks RCC growth and metastasis by direct targeting of SIRT1, indicating a potential new therapeutic effect in RCC therapy. The ADAMTS13 mRNA encodes the von Willebrand factor cleavage protease, which has been detected in a variety of tissues including the kidney. Human renal tubular epithelial cells synthesize ADAMTS13 with biological activity that regulates local microenvironment after release from tubules. Gfrα1 regulates renal development and ureteral maturation in the interaction with the tyrosine kinase Ret and the ligand glial cell-line derived neurotrophic factor (Gdnf). Other unique genes (ASPH, MIR17HG, APOE, BMP4, RCO1, NUMB and SEC63) detected by BNNPT are associated with renal diseases.

Theoretically, any machine learning algorithm that predicts Y using X may become the kernel function of our permutation test. Previously, CANOVA can be viewed as a permutations test of a simple moving average machine learning algorithm. We also tested a random forest as the kernel, however, both are not as powerful as BNNPT. We speculate that the reason for BNNPT’s superiority is that kNN is the most powerful method in one dimensional machine learning case. Further, we can take advantage of machine learning methods to solve correlation analysis problems.

One important advantage of BNNPT over CANOVA is that the bandwidth parameter can be left as default in most cases. The experiment demonstrates that mtry = sqrt(N) is robust. Thus our test can be viewed as “tuning free”. Also setting mtry = sqrt(N) instead of N (the conventional one nearest neighbor rule) is not only faster but also more powerful due to regularization effect (decorrelation among bags). BNNPT is also robust with the other parameter, number of bags (default is 256 for computing efficiency).

In this study, we can only test independence between two continuous variables. We can’t directly make covariate adjustments. However, we can further take covariate adjustments incorporate into account by first regressing response variable on covariates and then test the independence between the residual error and the response variable Y using BNNPT.

Typically when there exist non-linear correlation between two variables, the appropriate data transformation can efficiently bring the non-linearity to linear. We have compared the power of different methods by transforming the data first (including quadratic function, sine function and cosine function in Supplemental Materials). And Pearson correlation coefficient is the most powerful method using this strategy. In practice, the true relationship is typically not complex. A two-dimensional scatter plot can help us to reveal the relationship between two variables followed by appropriately chosen transformation model such as log, square or square root transformation. Furthermore, automatically finding the optimal data transformation model is a promising research direction which we will work on in the near future. However, BNNPT is still an efficient method to explore the non-linear relationships between two continuous variables without specific domain knowledge. According to the null hypothesis, if X really have prediction ability for Y, this dependence could be detected by BNNPT. Moreover, we will also develop a multivariable test which may be important in complex traits area.

While each method has its own advantages, the results of different methods can often be correlated with each other. Our simulation results indicate that the use of both linear correlation algorithms (Pearson, Spearman or Kendall) and non-linear correlation algorithms (BNNPT, CANOVA, MIC, Hoeffding or Distance) could increase the probabilities of detecting real biological signals.

To sum, we developed a robust algorithm to detect independence between two random variables especially in non-linear situations. To conclude, our BNNPT method appears to be efficient in testing nonlinear correlation in real data applications.

References
1. Sinha, H., Croxton, F. E. & Cowden, D. J. Applied General Statistics. Sankhya 5, 453–454 (1941).
2. Dietrich, C. F., Uncertainty, Calibration and Probability: The Statistics of Scientific and Industrial Measurement. (Taylor & Francis, 1991).
3. Atkine, A. C. Statistical Mathematics. (Read Books, 2012).
4. Galton, F. Typical Laws of Heredity. (publisher not identified, 1877).
5. Lockyer, N. Nature. (Macmillan Journals Limited, 1885).
6. Galton, F., Okamoto, S. & Eugenics, G. L. f. N. Regression Towards Mediocrity in Hereditary Stature. (Harrison and Sons, 1885).
7. Pearson, K. Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London 58, 240–242 (1895).
8. Stigler, S. M. Francis Galton’s account of the invention of correlation. Statistical Science, 73–79 (1989).
9. Wilcox, R. R., Introduction to robust estimation and hypothesis testing. (Academic Press, 2011).
10. Devlin, S. J., Gnanadesikan, R. & Kettenring, J. R. Robust Estimation and Outlier Detection with Correlation-Coefficients. Biometrika 62, 531–545, https://doi.org/10.1093/biomet/62.3.531 (1975).
11. Lorch, J., M. International Journal of Statistical Science. (Springer, 2011).
12. Myers, J., Well, A. & Lorch, R. F. Research design and statistical analysis. (Routledge, 2010).
13. Kendall, M. G. A new measure of rank correlation. Biometrika 30, 81–93, https://doi.org/10.1093/biomet/30.1-2.81 (1938).
14. Szekely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. Ann Stat 35, 2769–2794, https://doi.org/10.1214/009053607000000505 (2007).
15. Kosorok, M. R. On Brownian Distance Covariance and High Dimensional Data. Ann Appl Stat 3, 1266–1269, https://doi.org/10.1214/09-AOS312 (2009).
16. Wilding, G. E. & Mudholkar, G. S. Empirical approximations for Hoeffding’s test of bivariate independence using two Weibull extensions. Statistical Methodology 5, 160–170, https://doi.org/10.1016/j.stamet.2007.07.002 (2008).
57. Margaillan, G.
56. Grosse, L.
55. Porath, B.
58. Fritzsche, F. R.
54. Bergmann, C. & Weiskirchen, R. It’s not all in the cilium, but on the road to it: Genetic interaction network in polycystic kidney and liver disease. *Am J Hum Genet* 98, 1193–1207, https://doi.org/10.1038/ajhg.2016.504 (2016).

53. Adanea, M., Tari, R., Karlsson, J., Bekassy, Z. D. & Karpman, D. Biologically active ADAMTS13 is expressed in renal tubular epithelial cells. *Pediatr Nephrol* 25, 87–96, https://doi.org/10.1007/s00467-010-1622-y (2010).

51. Jing, J., Lin, N., Guo, S., Chen, J. & Xiong, M. Methods for joint imaging and RNA-seq data analysis. *arXiv preprint arXiv:1409.3899* (2014).

50. Pei, Z.
39. Li, L.
38. Li, L.
36. Tierney, L., Rossini, A. & Li, N. Snow: A Parallel Computing Framework for the R System. *Int J Parallel Prog* 31, 78–90, https://doi.org/10.1007/s10766-008-0077-2 (2009).

48. Vivante, A. Biologically active ADAMTS13 is expressed in renal tubular epithelial cells. *Pediatr Nephrol* 25, 87–96, https://doi.org/10.1007/s00467-010-1622-y (2010).

47. Jain, S. The many faces of RET dysfunction in kidney. *Oncogene* 5, 177–190, https://doi.org/10.1038/s41598-017-12783-9 (2017).

46. Manea, M., Tati, R., Karlsson, J., Bekassy, Z. D. & Karpman, D. Biologically active ADAMTS13 is expressed in renal tubular epithelial cells. *Pediatr Nephrol* 25, 87–96, https://doi.org/10.1007/s00467-010-1622-y (2010).

45. Porath, B. et al. Exome Sequencing Discerns Syndromes in Patients from Consanguineous Families with Congenital Anomalies of Kidney and Urinary Tract. *J Am Soc Nephrol* 28, 69–75, https://doi.org/10.1681/ASN.2015100962 (2016).

44. Marrone, A. K. et al. MicroRNA-17 similar to 92 Is Required for Nephrogenesis and Renal Function. *J Am Soc Nephrol* 25, 1440–1452, https://doi.org/10.1681/ASN.2014034390 (2014).

43. Wei, H. et al. MicroRNA target site polymorphisms in the VHL-HIF1alpha pathway predict renal cell carcinoma risk. *Mol Carcinog* 53, 1–7, https://doi.org/10.1002/mc.21917 (2014).

42. Stroustrup, B. *The C++ Programming Language*. (Pearson Education, 2013).

41. Huang, W., Shi, H., Hou, Q., Mo, Z. & Xie, X. GSTM1 and GSTT1 polymorphisms contribute to renal cell carcinoma risk: evidence from an updated meta-analysis. *Scientific reports* 5, 17971, https://doi.org/10.1038/srep17971 (2015).

40. Gao, Y. et al. E2F3 upregulation promotes tumor malignancy through the transcriptional activation of HIF-2α in clear cell renal cell carcinoma. *Oncotarget* (2016).

39. Li, L. et al. Mitochondrial genome and exceptional longevity in a Chinese population: the Rugao longevity study. *Age (Dordr)* 37, 9750, https://doi.org/10.1007/s11357-015-9750-8 (2015).

38. Li, L. et al. Antibody mediated CDCP1 degradation as mode of action for cancer targeted therapy. *Mol Oncol* 7, 1142–1151, https://doi.org/10.1016/j.molonc.2013.08.009 (2013).

37. Jiang, J., Lin, N., Guo, S., Chen, J. & Xiong, M. Methods for joint imaging and RNA-seq data analysis. *arXiv preprint arXiv:1409.3899* (2014).

36. Tierney, L., Rossini, A. & Li, N. Snow: A Parallel Computing Framework for the R System. *Int J Parallel Prog* 37, 78–90, https://doi.org/10.1007/s10766-008-0077-2 (2009).

35. Sawabe, M. et al. Mitochondrial haplogroups A and M7a confer a genetic risk for coronary atherosclerosis in the Japanese elderly: an autopsy study of 1,536 patients. *J Atheroscler Thromb* 18, 166–175 (2011).

34. Harrell, F. E. Jr & Harrell, M. F. E. Jr. Package ‘Hmisc’. (2017).

33. Szekely, G. J. & Rizzo, M. L. Energy statistics: A class of statistics based on distances.

32. Stroustrup, B. *The C++ Programming Language*. (Pearson Education, 2013).

31. Reshef, D. N. et al. Detecting novel associations in large data sets. *Science* 334, 1518–1524, https://doi.org/10.1126/science.1205438 (2011).

30. Good, P. Two Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. (Springer New York, 2013).

29. Cohen, J. Statistical power analysis for the behavioral sciences Lawrence Erlbaum Associates.

28. Jiang, J., Lin, N., Guo, S., Chen, J. & Xiong, M. Methods for joint imaging and RNA-seq data analysis. *arXiv preprint arXiv:1409.3899* (2014).

27. Cai, X. Y. Snow: A Parallel Computing Framework for the R System. *Int J Parallel Prog* 37, 78–90, https://doi.org/10.1007/s10766-008-0077-2 (2009).

26. Cohen, J. Statistical power analysis for the behavioral sciences Lawrence Erlbaum Associates.

25. Jiang, J., Lin, N., Guo, S., Chen, J. & Xiong, M. Methods for joint imaging and RNA-seq data analysis. *arXiv preprint arXiv:1409.3899* (2014).

24. Breiman, L. Bagging predictors. *Machine Learning* 24, 123–140, https://doi.org/10.1023/A:1010854314450 (1996).

23. Everitt, B. S., Landau, S., Leese, M. & Stahl, D. *Cluster Analysis*. (Wiley, 2011).

22. Altmann, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am Stat* 46, 173–185, https://doi.org/10.2307/2685209 (1992).

21. Scheffé, H. *The Analysis of Variance*. (Wiley, 1999).

20. Wang, Y. *Einfluss der Patientenausscheidung auf die Pharmakokinetik von Omeprazol*. (Springer, 2011).

19. Reshef, D. et al. Detecting novel associations in large data sets. *Science* 334, 1518–1524, https://doi.org/10.1126/science.1205438 (2011).

18. Gretton, A., Bousquet, O., Smola, A. & Schölkopf, B. In *International conference on algorithmic learning theory*, 63–77 (Springer).

17. Reshef, D. N. et al. Detecting novel associations in large data sets. *Science* 334, 1518–1524, https://doi.org/10.1126/science.1205438 (2011).
60. Soleimani, M. SLC26 Cl-/HCO3- exchangers in the kidney: roles in health and disease. *Kidney Int* **84**, 657–666, https://doi.org/10.1038/ki.2013.138 (2013).
61. Bishop, C. M. *Pattern recognition and machine learning*. (Springer, 2006).
62. Hofmann, T., Scholkopf, B. & Smola, A. J. Kernel methods in machine learning. *Ann Stat* **36**, 1171–1220, https://doi.org/10.1214/009053607000000677 (2008).
63. Breiman, L. *Random forests*. *Machine Learning* **45**, 5–32, https://doi.org/10.1023/A:1010933404324 (2001).
64. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186, https://doi.org/10.1016/j.cell.2017.05.038 (2017).

**Acknowledgements**

This research was supported by National Science Foundation of China (31330038, 31521003), Ministry of Science and Technology (2015FY111700, 2011BAI09B00), and the 111 Project (B13016) from Ministry of Education (MOE). The computations involved in this study were supported by the Fudan University High-End Computing Center. The views expressed in this presentation do not necessarily represent the views of the NIMH, NIH, HHS or the United States Government.

**Author Contributions**

Y.W., Y.L. and L.J. conceived the idea, proposed the BNNPT method. Y.W., Y.L., X.Y.L. and L.J. contributed to writing of the paper. Y.W., Y.L., Y.Y.S. and L.J. contributed the theoretical analysis. Y.W. also contributed to the development of BNNPT software using C++. Y.L. helped maintain BNNPT software and used R to generate tables and figures for all simulated and real datasets. W.L.P. and Y.L. used the R package ‘ggplot2’ to plot figures. X.F.W. and J.C.W. support the Rugao longevity cohort dataset. M.M.X. helped support the kidney RNA-seq dataset. Y.W., Y.L., X.Y.L., M.M.X., Y.Y.S. and L.J. contributed to scientific discussion and manuscript writing. All authors contributed to final revision of the paper.

**Additional Information**

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-12783-9.

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

---

© The Author(s) 2017