Journal of Biometrics & Biostatistics

ISSN: 2155-6180

Du et al., J Biomet Biostat 2014, 5:4
DOI: 10.472/2155-6180.1000204

Research Article

Open Access

A Permutation Test of Genetic Association between Salmonella Isolated On-farm and At-abattoir using Amplified Fragment Length Polymorphism

Yingzhou Du[1], Chong Wang[1,2]* and Peng Liu[1]

[1]*Department of Statistics, Iowa State University, Ames, Iowa 50011, USA*
[2]*Department of Veterinary Diagnostic and Production Animal Medicine, Iowa State University, Ames, Iowa 50011, USA*

Abstract

Pork and pork products have been identified as a significant source of Salmonella infection, which is a major public health concern. The contamination of Salmonella on pork can happen both on farms (before slaughter) and at abattoirs (after slaughter). Salmonella isolates were collected from both feces on farms and lymph nodes in the abattoir to determine if contamination at abattoirs can be linked back to the farms of origin. Molecular subtyping of the isolated Salmonella was performed using amplified fragment length polymorphism (AFLP), a Polymerase chain reaction-based, high-throughput, relatively inexpensive method. In this paper, we develop a permutation test for the genetic association of Salmonella isolated on-farm and at-abattoir using the AFLP data. Simulation studies show that the proposed permutation test controls the type I error rate appropriately as well as possesses high power. An application of the proposed permutation test to the real Salmonella ALFP data results in a p-value of 0.038 which shows strong evidence of association between Salmonella isolated on-farm and at-abattoir.

Keywords: ALFP; Genetic association; Permutation test; Salmonella

Introduction

Salmonella is a kind of rod-shaped, predominantly motile enterobacteria whose dimension is 0.7-1.5 μm in diameter and 2-5 μm in length. It can be found in cold-blooded animals, warm blooded animals and human beings. In the US, about 40,000 Salmonella infection cases are reported each year. From 1990 to 2006, 1,316 deaths in the US were identified to be related to Salmonella infection [1]. It has been discussed that up to 30% Salmonella infections to human beings are related to the consumption of pork or pork products [2,3]. Therefore, it is important to track the propagation of Salmonella in pork products. Lots of efforts have been made on studying the Salmonella contamination of pork products. The contamination of Salmonella could happen both on-farm and at-abattoir [4,5]. A risk assessment model showed that Salmonella-contaminated pigs were able to spread the contamination during processing at abattoirs [3]. Another study of the origin of Salmonella contamination on pig carcasses in two commercial slaughter houses showed that carcass contaminations did not come only from the corresponding infected pigs [6]. Instead, the majority of positive carcasses in both slaughter houses were contaminated by the pigs slaughtered earlier or from dispersed sources in the environment.

Based on the above-mentioned researches, the Salmonella contaminations of pig carcasses have two origins: the infected pigs on farms and the contamination during processing at abattoirs. It is important to test whether the genetic information of Salmonella collected at abattoir is associated with that collected on the farms of origin. Such test of genetic association is crucial in establishing the validity of tracing bacteria back to the farms of origin. For this purpose, Salmonella isolates were collected from both feces on farms (before slaughter) and lymph nodes at the abattoir (after slaughter) [7]. Molecular subtyping of the isolated Salmonella was performed using amplified fragment length polymorphism (AFLP), a Polymerase chain reaction (PCR)-based, high-throughput, relatively inexpensive method. We have reviewed the statistical literature but found no methods that can be applied directly to the AFLP data for the research question of interest. Thus, our objective of this paper is to develop statistical methodologies for testing the association between the isolates collected on-farm and those collected at-abattoir using data generated by the AFLP technology.

Materials and Method

Data structure

The AFLP data studied in this research were collected from 9 farms in the United States [7]. For each farm, 10 grams feces were collected from the rectum of 30 pigs one-to-three days before slaughter. The pigs were randomly chosen from the identified pigs that would be slaughtered, i.e., the harvest cohort. At the abattoir, the pig cohort was put in a pen until they were slaughtered at the same day. After slaughter, 30 mesenteric lymph nodes were collected randomly when the intestine moved on the production belt. The data after slaughter were collected from these lymph nodes. The intestine could be identified only at the cohort level but not at the pig level. Thus outcomes from the farm and abattoir can only be linked to the same farm, but not to the same pig. All samples collected both from feces and lymph nodes were frozen by wet ice and shipped to the lab. Samples were processed on the next day of delivery. The number of isolates identified at the farm level before or after slaughter varies from 1 to 26 among nine farms in the real data being analyzed.

Table 1 shows part of the data to illustrate the data structure. Each row of the dataset corresponds to one Salmonella isolate and includes the following information: 1) the VDL number (the sample ID); 2) the

***Corresponding author:** Chong Wang, Department of Statistics, Iowa State University, Ames, Iowa 50011, USA, Tel: 515-294-8047; Fax: 515-294-4040; E-mail: chwang@iastate.edu

Received July 22, 2014; **Accepted** August 05, 2014; **Published** August 11, 2014

Citation: Du Y, Wang C, Liu P (2014) A Permutation Test of Genetic Association between Salmonella Isolated On-farm and At-abattoir using Amplified Fragment Length Polymorphism. J Biomet Biostat 5: 204. doi:10.4172/2155-6180.1000204

VDL#	Farm	Type	Allele.1	Allele.2	Allele.3	…	Allele.440
828441	3	FF	0	0	1	…	0
828451	3	FF	0	0	1	…	0
822974	9	FF	0	0	1	…	0
…	…	…	…	…	…	…	…
832587	21	GAL	0	1	0	…	0
832588	21	GAL	1	1	0	…	0
…	…	…	…	…	…	…	…

Table 1: Part of the original AFLP data for illustration. Each row of the dataset contains genetic information for one Salmonella isolate. The columns are: the VDL number (sample ID); the farm number; the source of data: data were either collected from feces before slaughter (FF) or were collected from lymph nodes after slaughter (GAL); the genetic information for each of 440 alleles obtained by the AFLP method for the Salmonella isolate, where "1" means that the corresponding allele was detected in the corresponding isolate and "0" otherwise.

i	1	2	3	4	5	6	7	8	9
n_i	2	2	1	1	1	4	1	3	3
n'_i	5	26	3	14	21	13	1	6	1

Table 2: The number of isolates for different farms in the two parts of the AFLP dataset. The number of isolates before slaughter is denoted by n_i and the number of isolates after slaughter is denoted by n'_i for the i-th farm.

farm number; 3) the source of data: data were either collected from feces before slaughter (FF) or were collected from lymph nodes after slaughter (GAL); 4) genetic information obtained by the AFLP method for the Salmonella isolate. The genetic information is stored in a row vector consisting of 440 entries (0 or 1) corresponding to 440 alleles, where "1" means that the corresponding allele was detected in the isolate and "0" otherwise.

For convenience, we divide the whole dataset into two parts: the first part includes the data before slaughter and the second part includes the data after slaughter. We use binary responses, $\{X_{ijk}\}$ and $\{Y_{ij'k}\}$, to denote genetic data for the first part (before slaughter) and the second part (after slaughter), respectively, where i (=1, 2,…, G) indicates farm; j (=1, 2,…, n_i) indicates isolate in the first part of data, j' (=1, 2,…, n'_i) indicates isolate in the second part of data; and k (=1, 2,…, V) indicates the allele. Every farm (i=1, 2,…, G) has both data collected before slaughter and after slaughter. However, the number of Salmonella isolates, n_i and n'_i, on the same i'-th farm may not be the same, and they vary for different farms. The statistical question is to test for the association between the binary responses $\{X_{ijk}\}$ and $\{Y_{ij'k}\}$ at the farm level.

The total number of farms, G, is 9, and the number of alleles, V, is 440 in the Salmonella AFLP data. The numbers of isolates for each farm in the AFLP data are shown in Table 2.

A Permutation test for association

Our null hypothesis (H_0) is that the isolates collected on a farm before slaughtering pigs are not associated with those collected from the same farm after slaughter. To test such a hypothesis, we need to first quantify the association.

Consider two specific isolates from the same i'-th farm, $\{X_{ijk}$, k=1, 2,…, V$\}$ for an isolate before slaughter and $\{Y_{ij'k}$, k=1, 2,…, V$\}$ for an isolate after slaughter. The squared distance (d^2) between the two corresponding vectors could be used to quantify the distance between them:

$$d_{ijj'}^2 = \sum\nolimits_k (X_{ijk} - Y_{ij'k})^2 \qquad (1)$$

With the definition above, if the two isolates have similar genetic information, the squared distance between X_{ijk} and $Y_{ij'k}$ (k=1, 2,…, V) tends to be small. For example, the distance between (0, 0, 0) and (0, 0, 0) (d^2=0) is smaller than that between (0, 0, 0) and (0, 1, 0) (d^2=1).

The average of d^2 within a farm can be calculated by

$$d_i^2 = \frac{\sum_{jj'} d_{ijj'}^2}{n_i \times n'_i} = \frac{\sum_{jj'}[\sum_k (X_{ijk} - Y_{ij'k})^2]}{n_i \times n'_i} \qquad (2)$$

To get the total squared distance for all farms, say D^2, we take the sum of all the average squared distances from G farms, i.e.,

$$D^2 = \sum\nolimits_i d_i^2 = \sum\nolimits_i \left(\frac{\sum_{jj'}[\sum_k (X_{ijk} - Y_{ij'k})^2]}{n_i \times n'_i} \right) \qquad (3)$$

A permutation test is a type of non-parametric statistical test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels (farms) that are related to the research focal points in the observed data, i.e., under exchangeability when the null hypothesis is true. The p-value of the permutation test is calculated as the probability of the test statistic to be at least as extreme as the observed value based on the null distribution obtained from rearranged samples. The permutation test is useful when the distribution of test statistic is unknown or hard to derive.

To perform the permutation test, the second part (after slaughter) of data was randomly permuted at the farm level, and the D^2's are calculated for the permuted data following the same procedure using equations (1)-(3). We do not permute the first part of data (before slaughter) because permuting the farms for the second part of data is sufficient to achieve all the combinations. If the isolates from the same farm between the two parts of data are not associated, the squared distance calculated from original data, say D_0^2, should not be a rare value compared to the distribution of D^2 from the permuted data. If the alternative hypothesis is true, then the observed D_0^2 in data is expected to be small compared to the permuted distribution of D^2. Theoretically, there are $G!$ different permutations. If an exact test is performed, which means all permutations are used, we will obtain $G!$ test statistic $\{D_m^2, \text{where } m = 1, 2, \ldots, G!\}$ from the G! permuted datasets. A p-value can be obtained by calculating the proportion of values as extreme as or more extreme than the observed one $(D_m^2 \le D_0^2)$, i.e.,

$$p = \sum\nolimits_m I(D_m^2 \le D_0^2) / G! \qquad (4)$$

When the value of $G!$ is too large, e.g. $G!$=9!=362880 in the Salmonella AFLP data, a random set of N (N<$G!$) permutations could be used to save computation time.

Simulation Studies

We first use simulation studies to evaluate the performance of the proposed permutation test. We simulate data according to the structure of the AFLP data discussed previously to mimic the real data. More specifically, we simulate two parts that correspond to data before and after slaughter, respectively. Within each part, there are G=9 groups (corresponding to 9 farms). Each group has a number of row vectors the same as the number of isolates in the AFLP data. Because 145 alleles in the AFLP data contain either all 0's or all 1's before and after slaughter, which have no contribution of information for the test of association, these alleles are removed in the simulation study. Hence, there are 295 columns (alleles) in the simulated data compared to 440 alleles in the AFLP data.

Two matrices, one with dimension of $\left(\sum_i n_i = 18\right)$) x 295 and another with dimension of $\left(\sum_i n'_i = 90\right)$ x 295, are created to store the simulated data for the two parts, respectively. Each of the rows in the matrix stores a full set of alleles. The genetic data are simulated from Bernoulli distributions:

X_{ijk} ~ iid Bernoulli($\pi_{i,k}$),

$Y_{ij'k}$ ~ iid Bernoulli($\pi'_{i,k}$),

where $\pi_{i,k}$ and $\pi'_{i,k}$ are the probabilities for the outcomes to be 1 in the *i*'-th group (farm) and the *k*-th column (allele) for the first and second part of data, respectively. Investigation of the Salmonella AFLP data shows that the parameters π's vary among alleles. To mimic the real data, we identify 6 clusters of alleles by using the K-means clustering algorithm. Within each cluster, the probabilities for outcomes to be 1 are assumed follow a beta distribution, i.e., $\pi \sim$ beta(α, β). The parameters in the beta distributions, α and β, are estimated by fitting the real data within the cluster with beta-binomial model. The estimated parameters are shown in Table 3, and the corresponding probability density functions (pdf) for the beta distribution are shown in Figure 1.

During the process of simulation, we first divide all alleles into six clusters as identified by K-means clustering analyses of the AFLP data. Then for each *k*th allele and *i*th farm, the proportion of outcome 1, $\pi_{i,k}$ and $\pi'_{i,k}$, are simulated from the beta distribution with parameters estimated from the corresponding cluster of real data, which are shown in Table 3. In simulation of data under the null hypothesis, the proportions $\pi_{i,k}$'s and $\pi'_{i,k}$'s are simulated independently for the data before and after slaughter, thus leaving {X_{ijk}, $j=1,\ldots,n_i$} and {$Y_{ij'k}$, $j'=1,\ldots,n'_i$} to be independent as well. On the other hand, the data under alternative hypothesis should be simulated with some degrees of association between {X_{ijk}, $j=1,\ldots,n_i$} and {$Y_{ij'k}$, $j'=1,\ldots,n'_i$} at the farm level. This degree of association is controlled by controlling the proportion of matched columns in data. A matched column is a column within which the data of same group (farm) share the same random parameter of π for both parts of data (before and after slaughter), i.e., $\pi_{i,k}=\pi'_{i,k}$. By sharing the same value of the random proportion π, association is introduced between the binary outcomes in the matched column. If the percent of matched columns is 0%, the data within same farm for the two parts of data are completely uncorrelated, which is the scenario under the null hypothesis. As the percentage of matched columns increases, the association between the two parts of data increases as well. In our simulations, we study scenarios with the percent of matching columns ranging from 0% to 50% by an increment of 5%. For each percentage, 3000 data sets are simulated.

cluster	1	2	3	4	5	6
α	1.78	1.05	1.11	8.16	0.84	3.07
β	3.05	4.27	1.83	1.78	2.29	58.51

Table 3: Estimated parameters in the beta distributions, beta(*α*, *β*), for the 6 clusters identified from the Salmonella AFLP data.

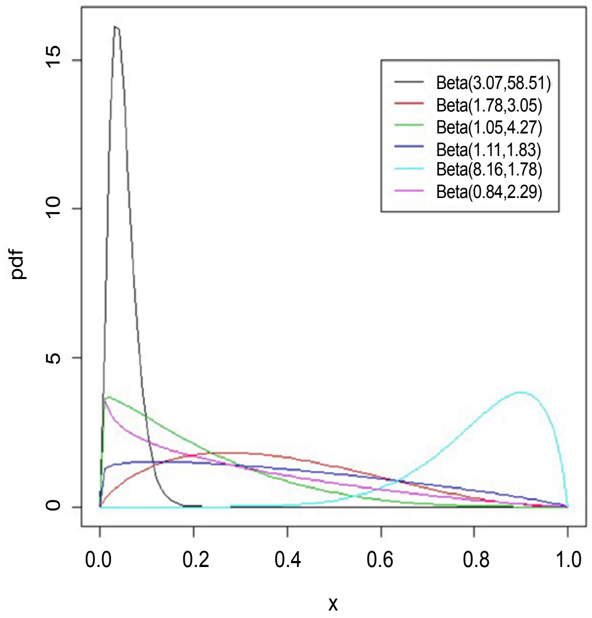

Figure 1: Probability density functions (pdf) for the beta distributions of the 6 clusters identified from the analysis of the Salmonella AFLP data.

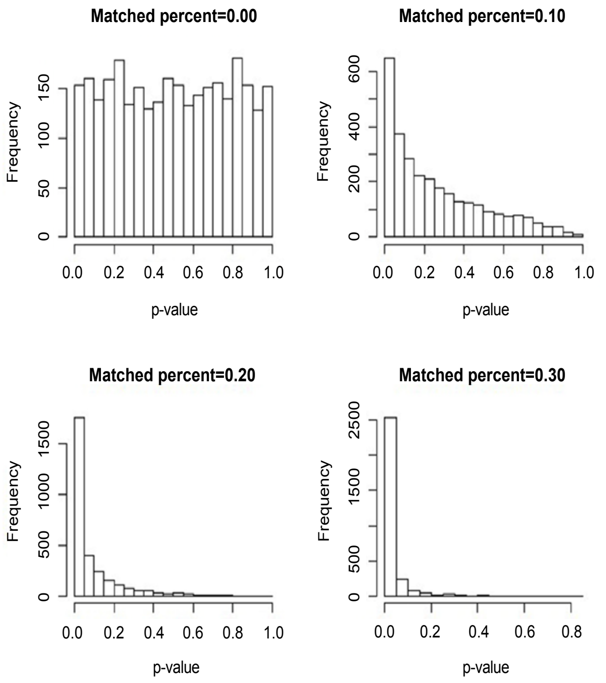

Figure 2: Histograms of p-values in simulation studies for various levels of association which are controlled by the percentage of matched columns.

Once the datasets are simulated, the permutation test is applied to test the association of the two parts of data. Theoretically there are 9!=362,880 permutations when we re-arrange the labels of the 9 farms. However, exploring all possible permutations requires a lot of computation time. Instead, we randomly choose 2000 permutations which show reasonable performance (as shown below) with acceptable computation time. In this case, formula (4) is changed to be

$$p = \sum_m I(D_m^2 \le D_0^2) / 2000 \qquad (5)$$

For each value of the matched column percentage ranging from 0% to 50%, 3000 datasets are simulated which result in 3000 p-values. The p-value histograms of the first four simulation settings (percent of matched columns=0%, 5%, 10%, and 15%) are shown in Figure 2. It can be observed that p-values obtained under the setting with 0% matching percentage roughly follows the distribution of Uniform(0,1), which is expected from a legitimate hypotheses testing procedure when

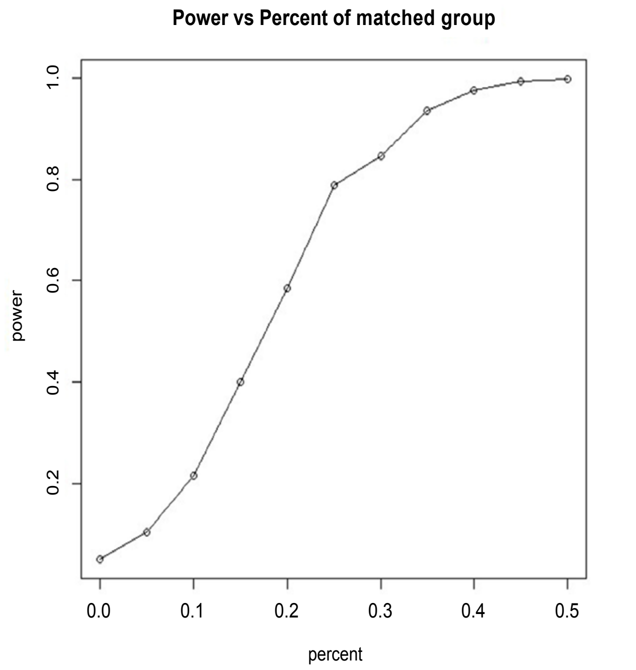

Figure 3: Probability of rejecting the null hypothesis vs the percentage of matched columns, which controls the level of association in simulated data. P-value ≤ 5% is considered significant.

the null hypothesis is true. It can also be observed from Figure 2 that, as the matching percentage increases, the p-value distribution decreases in stochastic order. The p-value histograms with none-zero matching percentage value all show decreasing density functions with a mode at 0. All these features are as expected.

Figure 3 presents the probability of rejecting the null hypothesis with the type I error rate controlled at 5% level at various values of the matching percentage based on our simulation studies. When the percentage is 0, i.e., the null hypothesis is true, the estimated probability of rejecting the null hypothesis is 4.57%. This shows that the type I error rate is successfully controlled using our permutation test procedure. As the level of association increases when the matching percentage increases from 0% to 50%, the chance of rejecting the null hypothesis also increases rapidly to 99.9%. This shows that our proposed hypothesis testing procedure is powerful when there is reasonable level of association in data.

Application to the Salmonella AFLP Data

We hereby apply the proposed permutation testing procedure to the Salmonella AFLP data collected from 9 farms in the United States [7]. The numbers of isolates identified before and after slaughter for each farm are shown in Table 2. A total of V=440 alleles were analyzed for each isolate using the AFLP technology.

Applying our permutation test to the Salmonella AFLP data results in a p-value of 0.038, which indicates strong evidence to reject the null hypothesis. Therefore, the data from feces and the data from lymph nodes within same farms are significantly associated at the farm level. This suggests that farm is an origin of Salmonella infection for the pork products. The current treatment procedures applied when pigs enter the abattoir are not adequate to remove Salmonella completely.

Conclusion

In this manuscript, we propose a novel permutation procedure to test the genetic association between Salmonella collected before and after slaughter using the AFLP data. Simulation studies show that the proposed method possesses high statistical power and controls the type I error rate well. A real data analysis results in a p-value of 0.038, which shows strong evidence of association between Salmonella isolated on-farm and at-abattoir at the farm level. The squared distance is used as a measure of genetic difference between isolates in our proposed permutation test procedure. Modification to the procedure could be considered by replacing the squared distance with alternative measures, such as the Euclidean distance. One may also consider putting weights on groups based on the number of isolates in them.

Acknowledgement

We thank Dr. Annette O'Connor and Dr. Bing Wang for discussion of the motivating question and sharing their Salmonella ALFP data.

References

1. Cummings PL, Sorvillo F, Kuo T (2010) Salmonellosis-related mortality in the United States, 1990-2006. See comment in PubMed Commons below Foodborne Pathog Dis 7: 1393-1399.
2. Guo C, Hoekstra RM, Schroeder CM, Pires SM, Ong KL, et al. (2011) Application of Bayesian techniques to model the burden of human salmonellosis attributable to U.S. food commodities at the point of processing: adaptation of a Danish model. See comment in PubMed Commons below Foodborne Pathog Dis 8: 509-516.
3. Barron UG, Soumpasis I, Butler F, Prendergast D, Duggan S, et al. (2009) Estimation of prevalence of Salmonella on pig carcasses and pork joints, using a quantitative risk assessment model aided by meta-analysis. See comment in PubMed Commons below J Food Prot 72: 274-285.
4. Alban L, Stärk KD (2005) Where should the effort be put to reduce the Salmonella prevalence in the slaughtered swine carcass effectively? See comment in PubMed Commons below Prev Vet Med 68: 63-79.
5. Botteldoorn N, Heyndrickx M, Rijpens N, Grijspeerdt K, Herman L (2003) Salmonella on pig carcasses: positive pigs and cross contamination in the slaughterhouse. See comment in PubMed Commons below J Appl Microbiol 95: 891-903.
6. Botteldoorn N, Herman L, Rijpens N, Heyndrickx M (2004) Phenotypic and molecular typing of Salmonella strains reveals different contamination sources in two commercial pig slaughterhouses. See comment in PubMed Commons below Appl Environ Microbiol 70: 5305-5314.
7. Wang B, Wang C, McKean JD, Logue CM, Gebreyes WA, et al. (2011) Salmonella enterica in swine production: assessing the association between amplified fragment length polymorphism and epidemiological units of concern. See comment in PubMed Commons below Appl Environ Microbiol 77: 8080-8087.