Searching for mutations in pigs using the human genome

A.S. Leaflet R2024

Laura Grapes, postdoctoral research associate, Stephen Rudd, Centre for Biotechnology, Turku, Finland, Rohan L. Fernando, professor of animal science, Karine Megy, Sygen International, University of Cambridge, Dominique Rocha, Sygen International, University of Cambridge, and Max F. Rothschild, distinguished professor of animal science

Summary and Implications

In humans, mice, rats, primates and pigs, it has been observed that some genes contain a high number of mutations, while others contain no mutations at all. It has not been determined whether a relationship exists across species for the variability of mutation number (i.e. a gene with a large number of mutations in one species will also have a large number of mutations in a closely related species). Here, the number of mutations in regions of genes that code for proteins were compared across pig, human and mouse. For the comparison, the pig mutations were obtained using computational methods that sorted through gene sequences to find differences. The mutations in humans and mice were real mutations from studies of individuals/animals. A high human-pig correlation and a lesser human-mouse correlation were found after comparing mutation number in genes. This is the first evidence of similarity across species in the variability of mutation number in genes. It indicates that discovery of mutations in pigs could be increased by searching in pig genes that are similar to human genes previously shown to have a large number of mutations in their protein coding regions. Some of these newly discovered mutations will change the protein that is produced by the gene, which can affect the protein's function, and result in changes in economically important traits such as growth, meat quality and reproduction. These mutations can then be used to select animals that are genetically superior.

Introduction

The number of mutations in human genes is thought to be affected by variation in mutation rates as well as natural selection forces and has been found to be highly variable across the human genome, even when comparing regions of genes that code for proteins. Similar observations about variation in mutation number have been made in mouse and rat and, on a smaller scale, in chimpanzee and pig. Due to the high degree of similarity between the protein coding regions of human and pig genes, the factors affecting mutation number in humans may affect pigs similarly. Thus, the frequency of coding region mutations in humans and pigs may be similar when comparing similar genes. Conversely, human and mouse coding mutation frequencies should be less similar than that of human and pig, as humans and mice have a lower level of coding sequence similarity on average.

Materials and Methods

Identification of mutations from pig sequences

Possible pig mutations were identified using a computational method that scanned through a set of pig sequences for a given gene. In previous studies to determine the sequences of genes in pigs, multiple sequences for the same gene were determined in several breeds of pigs. This information is freely available in public databases. All such pig sequences (~150,000) were downloaded from public databases and grouped together by gene. The minimum number of sequences that were considered to be a group was 8. Then, when comparing across sequences for a group, the relative frequency of a sequence difference at one position had to be 0.3 or higher. So, when comparing the minimum of 8 sequences, at least 3 of them had to be different to call the position in the gene a mutation. These sequence groups were compared to human and other mammalian gene sequences to determine if they contained protein coding regions, and if so, how large those regions were.

Correlation of coding region mutation frequency

The frequency of mutations in the coding region of a pig gene was calculated as $F = N / L$, where F was the frequency for a pig sequence, N was the number of mutations identified by the computational method, and L was the length of the gene's protein coding sequence that was examined. Human and mouse mutation frequencies were calculated in a similar manner, except N equaled the number of coding sequence mutations as listed by a mutation database, and L was the length of the protein coding sequence of the gene. We assume that the entire coding sequence has been examined when considering the mutations deposited in database for humans and mice.

Results and Discussion

Identification of pig coding sequence mutations

Unlike humans and mice, pigs do not have a large database of identified mutations, and the availability of such information in the near future is unlikely. Sequencing the porcine genome would allow large-scale mutation detection, and there is a joint sequencing project between Denmark and China, however it is unclear when results from this project will be made public. So, from our computational approach to identifying potential mutations, a total of 452 sequence groups were found to contain 1,394 mutations. All of the sequence group information and the mutation data have been made publicly available at

http://sputnik.btk.fi/project?name=swine. This represents the first database of its kind for pigs.

Comparison of human and pig coding sequence mutation frequencies

The 452 pig sequence groups that contained mutations were compared to human and other mammalian gene sequences, and 231 mutations were found in the protein coding regions of 80 different genes. Validation studies were performed for a sample of 25 mutations, and 16 (64%) were experimentally validated, indicating that the stringent conditions of the computational methods produced reliable data.

Starting with the 80 pig genes identified as containing coding sequence mutations, genes were eliminated from the comparison based upon known characteristics of the gene that might lead to false results. With these conditions only 25 pig genes were useful for comparison. However, the correlation between the frequency of human and pig coding sequence mutations for these 25 genes was found to be high (0.77, $P < 0.00001$) (Figure 1). The average pig coding mutation frequency of these 25 genes was 1.9 mutations per 1,000 base pairs of sequence, with the average human coding mutation frequency being 1.6 mutations per 1,000 base pairs of sequence.

While the pig sequences sampled here do come from a diverse mixture of breeds, it is unlikely that the sample is representative of all porcine species worldwide. However the sample is likely to accurately represent the domestic pig species of America and Western Europe, and it serves as an example of porcine protein coding DNA sequence. For the human genes considered in this study, the coding sequence mutations listed in the database were derived from many different project types, ranging from computational searches of human gene sequences to projects comparing sequences from dozens of individuals from diverse populations. Although it is not feasible to obtain a mutation frequency that is representative of the entire human population for every gene, the mutations considered here were typically found in several populations described as including Caucasian, Russian, Japanese and African individuals. Thus, the mutation frequency in the genes compared here can serve as an example of mutation frequency in human coding sequence.

Comparison of human and mouse coding sequence mutation frequencies

If the high level of similarity between human and pig protein coding sequence leads to a strong correlation between their coding mutation frequencies, then the correlation between human and mouse coding sequence mutation densities should be lower because their sequence identity is generally less. The coding sequence mutation frequencies from a primarily random sample of 50 human and mouse genes were compared and showed a moderate correlation (0.48, $P < 0.0005$) (Figure 2). Unfortunately, only 7 of the 25 genes from the initial pig-human data set contained coding sequence mutations in the mouse according to the database. These 7 were included in the set of 50 genes used for the human-mouse comparison. Although a more extensive comparison of coding sequence mutation frequency across all three species was not possible, the correlation between the mouse and human coding sequence mutation frequency for these 7 genes was zero, while the pig and human correlation was high (0.73, $P < 0.07$) (data not shown). The decreased correlation between the frequencies of human and mouse coding sequence mutations, as compared to that of human and pig, supports the idea that factors regulating the number of mutations in coding regions will affect closely related species in a similar manner.

The apparent relationship between the number of mutations in human and pig protein coding regions is directly applicable to pig genomics research, as it will allow site-directed screening of porcine genes for coding sequence mutations, resulting in their expedited discovery. Future use of computationally based mutation detection methods in pigs is dependent upon the amount of available sequence data. Results from a large-scale porcine sequencing project are currently awaited to allow creation of large data sets for analyses. Validation of computationally derived mutations and those found from human comparative studies will contribute to the knowledge base of genetic differences among animals and promote genetic improvement in traits such as reproduction, disease resistance, production and longevity.

Figure 1. Correlation between porcine and human coding sequence mutation frequency (Number of mutations per base pair of protein coding sequence).

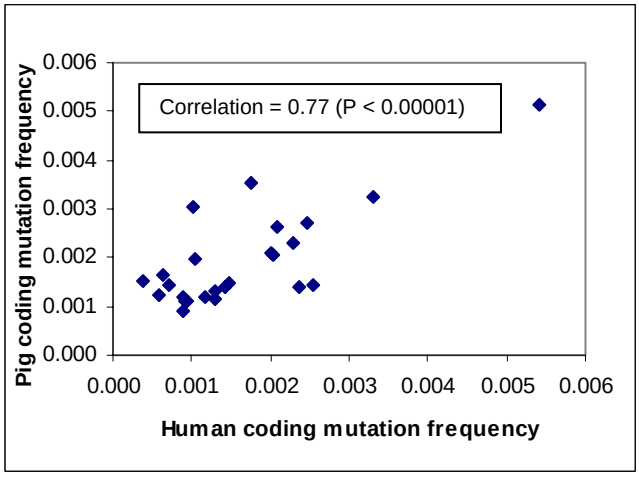
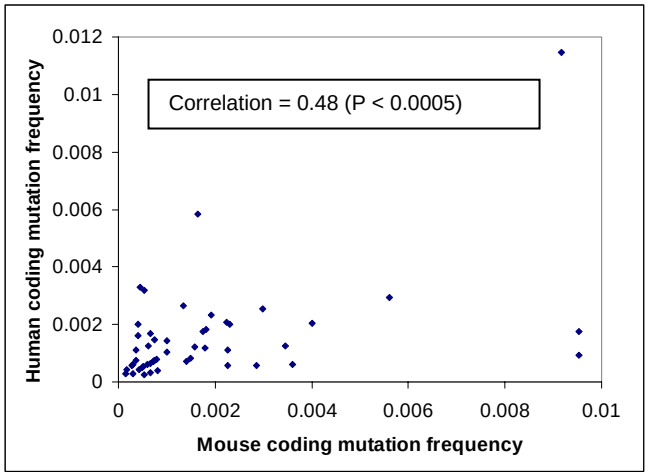


Figure 2. Correlation between human and mouse coding sequence mutation frequency (Number of mutations per base pair of protein coding sequence).



Acknowledgements

We thank Meena Bagga for her work in the validation of computationally derived porcine mutations. Thanks are given to Zhiliang Hu for his technical advice concerning database construction and programming. This work was supported by Sygen International, plc, and a United States Department of Agriculture National Research Initiative grant. Stephen Rudd is supported by the Academy of Finland and the German Genomanalyse im biologischen System Pflanze initiative (0312270/4) project. Karine Megy is supported by a European Union Marie Curie Industrial Host Fellowship.