Designing Software to Locate Differences in the Shrimp Genome

A.S. Leaflet R2353

Danielle M. Bowen, graduate assistant;
Zhi-Liang Hu, associate scientist;
Zhi-Qiang Du, postdoctoral research associate;
Max F. Rothschild, distinguished professor

Summary and Implications

In order to determine where important differences in the genomic sequence of Pacific White Shrimp occur, many copies each of multiple regions of DNA sequence are needed. Then similar sequences can be aligned so that almost all of the bases are identical between the sequences and differences are easy to notice. One of the major issues with predicting single base position differences (SNPs) in this manner is that DNA sequencing techniques are not 100% consistent in most cases. Consequently, it needs to be determined whether a particular base is different because the true genetic sequence is variable at that position or because the sequencing process resulted in the base position being incorrectly called. SNPidentifier is a newly developed computer program that takes into account the unreliability of sequence data and tries to use only the more reliable sequences to predict where true SNPs are located. The goal of locating SNPs in Pacific White Shrimp is to identify base positions that can possibly be used in the future as molecular markers for traits of interest to shrimp breeders.

Introduction

Pacific White Shrimp have been farmed in the Americas for many years and are becoming more popular in Asia. Little is known about the genomic sequence of this species, so selective breeding based on genetics cannot be used extensively to improve production. In order to use differences in the genome to decide which animals to mate, knowledge is needed of where those differences are located.

There are about 26,000 small DNA sequences from Pacific White Shrimp that are publicly available. Sequences that are extremely similar to one another are assumed to be based on the same region of DNA being sequenced multiple times (either from the same individual or from different individuals). Therefore, we can align these similar sequences using available software and try to identify differences based on the few base positions that differ between the mostly identical sequences.

When DNA is sequenced, a computer program is used to predict whether the base at each position along the sequence is an A, C, G, or T, and it sometimes is confused and predicts the wrong base (for more information on DNA sequencing procedures see http://en.wikipedia.org/wiki/DNA_sequencing). Consequently, the publicly available sequences cannot be trusted to be 100% correct. This knowledge has to be taken into account when designing a computer program to predict the location of differences in the DNA sequence, since an observed difference may be the result of a previous computer predicting the wrong base at a given position, rather than a true difference in the shrimp genome. SNPidentifier, a computer program, was developed to locate single base differences (also known as single nucleotide polymorphisms or SNPs) in the shrimp genome given a set of these similar sequences that have been aligned.

Materials and Methods

The 25,937 DNA sequences publicly available for Pacific White Shrimp were manually examined for obvious problems. Then they had regions with the same sequence of bases repeated over and over again removed. This is done, since the sequence in these regions is significantly less reliable than in regions without repeats, making the observed differences in these regions more likely to be due to computer error and less likely to be helpful for making breeding decisions. Next, similar sequences were aligned using the computer program, CAP3. This program produced 3,532 groups of aligned sequences.

SNPidentifier was designed to take the aligned sequences from the CAP3 computer program and: 1) Trim off the first 10 bases of each sequence, since sequence data is less reliable close to the start of a sequence. 2) Count the number of bases in the sequence that the previous computer deemed undeterminable and exclude the sequence if more than 10% of the bases could not be determined. 3) Compare the remaining sequences to look for differences between aligned sequences. If a difference is detected, the program first checks the 15 bases on each side of the base of interest to ensure that they match the other sequences to increase the probability that the correct base position was predicted. Next, SNPidentifier computes how many of the sequences contain an A, how many contain a C, and so on with G and T at that particular base. If over 90% of the sequences contain the same base, then the program decides the differences were probably due to the unreliability of the sequences used. Also, if a less common base is observed fewer than 4 times, it is also assumed to be an artifact of unreliable sequence data. Using these criteria, 504 SNPs were predicted from 141 groups of similar DNA sequences.

To check the accuracy of these SNP predictions, DNA was isolated from 18 different Pacific White Shrimp. Many copies of the DNA fragments around 39 of the predicted SNPs were made using a process called polymerase chain reaction (PCR). These fragments of DNA were then sequenced and the computer program Sequencher was used to align the matching fragments from different individuals and look for base positions that differed within or between individuals.

Results and Discussion

Out of 39 predicted SNPs that were tested, 17 (44%) were confirmed to differ between the 18 individuals examined. This result is fairly comparable to some other methods that require more information about sequence reliability to predict SNPs. Hopefully some of the predicted SNPs will be significantly linked to traits of interest to shrimp breeders and can be used as molecular markers for determining which animals to mate.

Acknowledgements

Funding was provided by the USDA NRSP8 National Animal Genome Program, the Iowa Agricultural Home Economics Experiment Stations, State of Iowa and Hatch funding. D.M.Bowen is supported by a USDA-CSREES National Needs fellowship under Grant no. 2007-38420-17767.