Hannah Watkins
Mentor: Dr. Bino John

Identifying Regions of High Conservation Potentially Related to Breast Cancer

Introduction

There are regions in the human genome that are nearly identical to regions in other organisms' genomes. [1] Web browsers, such as the UCSC Human Genome Browser, make it relatively simple to identify these areas of conservation. My research project will focus on using such browsers to find regions of high conservation in pre-identified sequences that may be involved in breast cancer. These sequences were collected using a tiling array and will be described in greater detail in the Design and Methods section. There is significance in locating highly conserved areas, because it suggests that those regions are under strong negative selection. Mutations to a conserved region are more likely to cause harm than mutations in a quickly evolving region. This is because if a region of the genome is conserved it indicates that the individuals with different sequences for the region were not able to as successfully propagate their changed sequences to the next generation, and thus the gene remained conserved. The well conserved sequences will be compared against the human genome to see if similar sequences exist and what those sequences code for.

Design and Methods

My research project will build off of data that was already collected by the Dr. Bino John lab at the University of Pittsburgh. Data was collected using a tiling array. The tiling array consisted of thousands of probes that spanned most of the human genome. RNA from healthy tissue and from breast cancer tissue was reverse transcribed into cDNA, labeled with fluoroform, and put onto the probes. By looking at the different color ratios and performing other computer analyses, 357 sequences were found to potentially be involved with breast cancer. Next, the approximate locations of the

sequences were determined and saved in an excel spreadsheet. My research will involve analyzing these sequences.

First, I will use the human genome browser to update the coordinate positions from the July 2003 assembly to the March 2006 assembly. Though the most recent human assembly was compiled in 2009, it is not yet ready to be used for comparative analysis. Next, I will use the UCSC genome browser's comparative analysis tools to individually view each sequence and asses the likelihood that it is part of a highly conserved region. There are 44 vertebrates that the browser uses to compare with the sequence.

Two different algorithms, phyloP and phastCons, are used to assign scores on how well the sequence matches similar regions in the other organisms. The phyloP score judges how well each individual base in the sequence matches the bases in the other vertebrates. If it is likely that the base is conserved, the score is positive; otherwise, it is negative. The phastCons score uses a hidden Markov method, so it takes into account preceding and proceeding bases when deciding when a region is conserved. PhastCons scores represent the probability that the region is under negative selection, so a score of .9 would indicate a 90% chance that the region is conserved. For my analysis, I will focus primarily on the phastCons scores. PhastCons scores are calculated using only the mammals and also using all 44 available vertebrate.

Additionally, graphs for individual organisms using a grayscale to indicate when bases are conserved (the darker the graph, the more bases are conserved) can be displayed. There is another score that I will use to judge whether a region is conserved called the log-odds score (LOD). This score divides the probability that a region is conserved by the probability it is not conserved and then takes the log of that number. The browser provides options to zoom in and out of regions, so in addition to looking at the sequence itself, I will look in surrounding regions to see if I detect any interesting regions there. When searching for conservation, I will look to make sure the regions do not correspond to an

exon, or to a region near an exon. This is because those regions have already been identified and I am searching for new information.

After completing the first analysis with the browser, I will select around 15 sequences that appear to be well conserved and label them in the fasta format. Next, I will use a computer program to format the sequences so that the Basic Local Alignment Search Tool (BLAST) is able to interpret them. These sequences are called "queries." Using ensembl, I will download the cDNA human genome, which will serve as the "target". I will then run BLAST to see if any of the queries have matches in the target. If there are matches, I will use ensembl to find what the function of those regions are, because it is likely that the query would have the same function. If time allows, I will BLAST the remaining 342 sequences after completing the first 15.

Predicted Results and Discussion

It is likely that some of the sequences will be better conserved than others; however, it is unknown whether the most conserved regions will result in the closet matches in BLAST. If BLAST does find matches, the sequences it aligns to should give hints about which types of RNA coding sequences give rise to cancer when mutated. The conserved areas will most likely vary in length, and it is questionable whether the longer or shorter conservation sequences will produce more interesting results. The analysis should show whether the conserved regions are in introns, or if they are in unknown areas of the genome. It will be interesting to see how regions well conserved in only non-mammalian vertebrate compare with the regions that are well conserved in only mammalian vertebrate. This study should help to increase knowledge about which areas of the genome are related to breast cancer.

Bibliography

[1.]Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Hauler D. "Ultraconserved Elements in the Human Genome." Science 304 (2004): 1321-325.

Karolchik D, Kuhn, RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res*. 2008 Jan;36:D773-9

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002 Jun;12(6):996-1006.