|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***Materials:***The materials that I need for this project are a computer, the internet, Cancer Genome Anatomy Project (CGAP) Website, and Microsoft Excel. The internet is used to research genes and drugs by using sites such as Locus Link and GenBank at the National Center for Biotechnology Information (NCBI). The CGAP website is where I obtain data on chromosome aberrations and the levels of expression for various genes in normal and cancerous tissues. I use Microsoft Excel to organize the data and apply a chi squared test to determine the statistical significance of whether each gene is different in its level of expression in normal versus cancer tissue. ***Step 1:*** Identify frequently recurring chromosome aberrations in breast cancer. Procedure: Go to CGAP home page and select Catalog of Resources.  Click on Chromosomes and go to the Recurrent Chromosome Aberrations in cancer Database Searcher  Under structural aberrations, for aberration type, enter "all". This gives all types of possible aberrations. For topography, enter "breast". This gives chromosome aberration locations for breast tissue. For morphology enter "adenocarcinoma". This is the type of cancer I will be studying. For gene, enter "all genes" and then hit retrieve.  A list of approximately 200 aberrations comes to the screen after a minute. Use file save on the browser to download the list in html format. Open the list in Excel and save in workbook format (xls).  In excel, sort the list in descending order of cases observed. This results in a list of the most frequently occurring chromosome aberrations in cancer.  From this list I selected the aberrations 11p15, 19p13, 19q13, 1p36, and 8q24. I selected these, not because they were the most frequently occurring, but because they were the most frequently occurring that were not at a centromeric location. I do not want to use aberrations at a centromeric location because they do not contain many genes. Further explanation on why I chose these aberration sites can be found in Results. ***Step 2:*** Find the list of genes that map to the identified cytogenetic location and download Virtual Northern data for each gene.Procedure: Go to the CGAP webpage. Go to Gene Finder.  Under General, for choose the organism, enter "homo sapiens". Under tissue type, enter "breast". For the Gene ontology term, leave it blank. In cytogenetic location, enter the location of the chromosome aberration being studied, such as "11p15". Leave the Keyword in gene name blank. Then submit the query.  A list of approximately 100 genes at the cytogenetic location is displayed. Using File Save from the browser, save the gene list in html format, open it up in Excel and resave it in xls format.  In the gene list spread sheet, enter additional columns for number of Expressed Sequence Tags (ESTs) found in normal tissue for the specific gene, number of ESTs found in cancer tissue for the specific gene, the total number of ESTs in the normal tissue for all genes expressed in the tissue, the total number of ESTs in the cancer tissue for all genes expressed in the cancer tissue.  Go back to the gene list on your browser. Click on the hyperlink for each gene. This leads you to the Gene Info page for that gene. Click on the hyperlink titled Virtual Northern. For mammary gland, enter the four EST values into the excel spread sheet. Go back to the Gene list and repeat for each gene on the list.  The result is an excel spread sheet with the list of genes and their relative expression levels in normal and cancer tissues.  ***Step 3:*** Statistical Analysis of the relative expression levels of normal and cancer tissues.  Procedure:  To make a statistical analysis of the data, I went to the Differential Gene Expression Displayer (DGED). This tool displays statistically significant differences in gene expression between two libraries or two pools of cDNA libraries chosen by the investigator.  At the CGAP Catalog of Resources, go to DGED. Click on the hyperlink, sequence odds ratio and the chi squared test. This page describes the procedure on how to apply the chi squared test to a two x two table of the following form:   |  |  |  | | --- | --- | --- | |  | Gene 1 | All other genes | | Normal tissue | ESTs assigned to Gene 1 | # of ESTs in normal that are not Gene 1 | | Cancer tissue | ESTs assigned to gene 1 | # of ESTs in normal that are not Gene 1 |  To determine the number of ESTs in normal that are not Gene 1, I entered a formula, subtracting the ESTs assigned to the specific gene from the total number of genes. I used the formula, G12-E12, where G is the column heading for ESTs assigned to Gene1, and E is the column heading for the total ESTs in normal tissue. 12 refers to the row of the first gene. I then highlighted the column beneath row 12, to the row of the last gene in the gene list. I pressed ctrl + D. This fill down command propagates the formula down the column with the proper row numbers for each gene. I repeated this procedure for the cancer tissue columns.  I then asked my dad for help on chi squared, who emailed the CGAP helpdesk to ask for the formula to calculate the chi squared test, and they sent back this response:  ***From: <Marty\_Bartholdi@BERLEX.COM>***  ***To: <marty-bartholdi@worldnet.att.net>***  ***Subject: chi sq and P value***  ***Date: Monday, November 26, 2001 1:50 PM***  ***----- Forwarded by Marty Bartholdi/RM/USR/SHG on 11/26/01 01:49 PM -----***  ***"Schaefer,***  ***Carl (NCI)"***  ***<schaefec@mai To: Marty Bartholdi/RM/USR/SHG@SCHERINGGROUP***  ***l.nih.gov> cc: "Greenhut, Susan (NCI)" <greens@osp.nci.nih.gov>***  ***Subject:***  ***11/26/01 chi sq and P value***  ***12:57 PM***  ***I think typos are getting in our way here.***  ***We compute the chi sq value using Perl code that does the following:***  ***[(a+b+c+d) x (a x d - b x c) x (a x d - b x c)] / [(a+b) x (c+d) x (a+c) x***  ***(b+d)]***  ***Note that (a x d - b x c) x (a x d - b x c) is just the square of (a x d -***  ***b x c); there was a typo in the previous mail we sent you: "b x d" should***  ***have read "b x c".***  ***We compute the P value as a continuous value rather than using a lookup***  ***table. We use some C code that we obtained from this source:***  ***http://www.fourmilab.ch/rpkp/experiments/distribution/rpkp.tar.gz***  ***-- Carl Schaefer***  ***National Cancer Institute***  ***Center for Bioinformatics***  ***301-435-1535***  In the excel spread sheet, I entered the chi squared calculation using the proper column and row designations for each value.  (E11+I11+F11+J11)\*(E11\*J11-I11\*F11)\*(E11\*J11-I11\*F11)/((E11+I11)\*(F11+J11)\*(E11+F11)\*(I11+J11))  This equation produces the value for the chi squared distribution, which I then entered into the formula CHIDIST(K12,1) which yields the p value. The smaller the p value, the more likely it is that the observed expression difference is not the result of random variation. For example, a p value of .05 means that there is a 95% chance that the difference seen is significant and a 5% chance that it is due to random variation.  I then used the fill down command to propagate this command through the whole list of genes.  My dad and I checked the correctness of the equation in Excel by comparing our results with the results from the DGED tool, at the website.  I then entered the equation to calculate the fold change of expression for each gene in normal tissue versus cancer:  (E14/I14)/(F14/J14) This calculates the ratio of the percentage of ESTs for gene 1 in normal versus cancer.  I then sorted the gene list spread sheet on ascending order of p value, to rank order the genes starting with the smallest p value.  ***Step 4:*** Identification of genes that are significantly differentially expressed at a recurrent chromosome aberration, which may have a function that contributes to cancer progression.  For the most significantly differentially expressed genes, I went to Locus Link to study the gene�s functions. I can try to understand if it may be a possible drug target. |
|  |