A comparative study into the most frequently occurring mutations, causing human female breast and ovarian cancer, in the top 20 genes, as recorded in COSMIC, in particular whether they occur in both ovarian and breast cancer.

Personal Engagement/Aim:
I have decided to do research into the gene mutations that cause breast cancer. In my initial research I found that the BRCA1 gene mutations are the largest cause of hereditary breast cancer. I am personally very intrigued by this as my mother suffered from breast cancer a few years ago. At this point I was only about 9 years old so I had trouble understanding what exactly was happening, but after we had looked into genetics during class time I was automatically interested in the gene mutations that cause or trigger certain types of cancer. Although only around 5% to 10% of breast cancer is hereditary[1], it is something to be aware of and something very important especially if relatives in your family have suffered from breast cancer. Another thing that interested me is that the abnormal BRCA1 gene is not necessarily the only thing that causes breast cancer in that certain patient. "Researchers are learning that other mutations in pieces of chromosomes -- called SNPs (single nucleotide polymorphisms) – "[2], this gives evidence for a 'butterfly effect', the BRCA1 malfunction causing certain other malfunctions, that then culminate causing the breast cancer. Since the beginning of my research my question/aim has changed a little. As the COSMIC database uses somatic mutations, hereditary gene mutations aren't specifically taken into account. Throughout my research I became more interested between the connections of ovarian and breast cancers and whether similar genes cause these types of cancers. Therefore, I have decided to do more research into comparing the gene mutations causing breast and ovarian cancer, instead of focusing on the connection of BRCA1 inheritance and the occurrence of breast cancer. However, the BRCA1 (and also sometimes the BRCA2 gene) plays an important role in both ovarian and breast cancer.

Introduction:
Through the research process explained above I have come to the research question/statement *A comparative study into the most frequently occurring mutations causing breast and ovarian cancer, in the top 3 genes and whether they occur in both ovarian and breast cancer.* My hypothesis for this is that there will be similar genes in the top 20 genes and maybe one same gene in the top 3 genes. The connection between breast and ovarian cancer[i] that I have researched mainly talked about the occurrence between the cancers, due to hereditary factors. This factor is not really taken into account in the COSMIC database; therefore it is difficult to hypothesize the exact connection between the occurrences of mutations causing the two different cancers.

Breast Cancer
Breast cancer is caused by an uncontrolled and abnormal growth of cells in the breast usually in the glands/milk ducts of the breast. It can then go on to affect other tissue within the breast or move to other parts of the body (metastasise to other regions).

Ovarian Cancer
There are three main types of tumours caused by different cells in the ovaries. Tumours that begin at the cells that produce the eggs are called germ cell tumours. Stromal

---

[1] http://www.breastcancer.org/risk/factors/genetics
Accessed 14 April 2015

[2] http://www.breastcancer.org/risk/factors/genetics
Accessed 14 April 2015

tumours start at the cells that hold the actual ovary together and also produce progesterone and oestrogen. The epithelial tumour is the one that is found in most cases of ovarian cancer. It affects the cells that cover the outside surface of the ovary.

BRCA1/BRCA2

Both the BRCA1 and BRCA2 gene are so called tumour suppressors. They fix any 'broken' DNA to ensure the sequence stays intact. Women with mutations in their BRCA genes are more likely to develop either ovarian/breast cancer.

Loss of cell cycle control

The uncontrolled growth of cells is what essentially leads to the growth of a tumor, which then causes cancer. Although there are other factors that can cause a predisposition to cancer such as hereditary factors, an unhealthy lifestyle or exposure to harmful chemicals, the uncontrolled cell growth is what forms the tumor.

Cyclins are proteins that ensure the cell cycle occurs in the correct order. They also make sure that the cell only moves to the next stage of the cell cycle once it has successfully completed the previous stage of the cycle. If the stage of the cycle is not completed correctly by the time it has met the specific cyclin the cell 'self-destructs'. There are four main Cyclins in the body: Cyclin D, Cyclin E, Cyclin A and Cyclin B.

The cyclin bonds to a type of enzyme called cyclin-dependent kinases. The kinases are activated in this process and attaches phosphate groups to various proteins within the cell. This bond (between the phosphate and the protein) prompts the activation of other proteins and the tasks for a specific phase of the cell cycle are completed.

The four different types of cyclins previously mentioned take place after each other at certain parts of interphase and mitosis. Cyclin D "triggers cells to move from G0 to G1 and from G1 to S phase"[3]. Cyclin E "prepares the cell for DNA replication in S phase"[4] (part of interphase). Cyclin A "activates DNA replication inside the nucleus in S phase"[5] and Cyclin B "promotes the assembly of the mitotic spindle and other tasks in the cytoplasm to prepare for mitosis". An example of how the cyclins may affect the creation of a cancer is portrayed in a the paper *Cyclins and Cell Cycle Control in Cancer and Diseases.* It states that "Cyclin D1 overexpression is found in more than 50% of human breast cancers"[6]. The overexpression of the cyclin causes the loss of the normal cell cycle subsequently causing an abnormal growth of cancerous cells.

Cosmic Database

The COSMIC (Catalogue Of Somatic Mutations In Cancer) Database is a database including mutations acquired somatically that are found in human tumors. The database uses information from scientific papers and from the Cancer Genome Project. Since its creation the database has grown immensely including more and more mutations for a variety of different cancers (Database published June 2004)[7].

---

[3] IB Biology book page 56
[4] IB Biology book page 56
[5] IB Biology book page 56
[6] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3636749/ Accessed 16 June 2015
[7] http://www.nature.com/bjc/journal/v91/n2/pdf/6601894a.pdf Accessed 1 July 2015

To explore my research question I will be using the Cosmic Cancer Browser Database[8] to find the top 20 mutations for both ovarian and breast cancer. I have not controlled the type of tissue, sub-tissue, histology and subHistology affected, as leaving these variables unspecified would result in the largest number of samples and give the general overview in the mutations caused. Furthermore, it enables a more fair comparison between the two types of cancers. The data that I will receive will not be 100% fair as the sample sizes for the different type of cancers and the different mutations differs between ovarian and breast cancer. I will do all my data collection within two days maximum to ensure that none of the data is changed by the database.

Variables

Dependent Variable: The occurrence of different genetic mutations that cause either breast or ovarian cancer

Controlled Variable: I will always be keeping the sub-sections that are used to specify the genetic mutations the same. Always keeping them non-specified to ensure I get the largest sample size. I will also make sure that I collect all my data in a span of maximum two days to ensure that no new information is added to the database that could change the comparison between the two sets of genetic mutations.

Independent Variable: The two different types of cancer (breast and ovarian) are the independent variable.

Uncontrolled variables

It is hard to have controlled variables as the lifestyle, age and other environmental factors may not be the same for each sample of gene. These are all factors that are unable to be controlled using the database I have chosen. Although, the researchers of the database may have this information it is not made public. From a moral/ethical perspective this is right as it ensures the privacy of the patient. One could assume that many of the people with mutations may have acquired their cancers through unhealthy habits such as smoking, excessive consumption of alcohol, exposure to certain chemicals or unhealthy eating habits. This information could help to make further connections between the gene mutations. For example, a certain type of behaviour could be causing the majority of a certain mutation.

Safety/Ethical/Environmental Issues

A safety issue could be the use of the Internet and the ethical implications that come along with storing information on a database that anyone is able to access. However, in this case it isn't really a problem as there is basically no patient information included therefore no privacy issues occur. The authenticity of the database could also be considered as a problem. As I am not collecting the data by myself I cannot be 100% sure that the information on the database is 100% correct. However, the database provides some information on their website in response to this 'issue'. Here they describe in which ways they obtain their data etc., suggesting that the data portrayed in the database is very accurate. The use of a computer could be seen as an environmental issue due to the use of electricity and the implications of that.

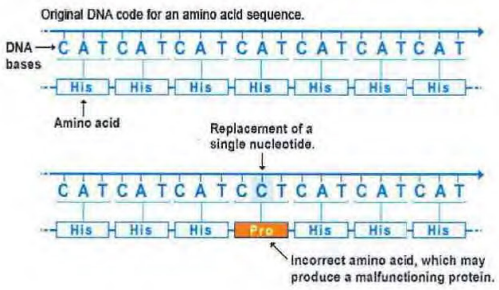
Different types of mutations

There are many different reasons for mutations that occur in the DNA during replication. These include substitution missense and nonsense, synonymous substitution, insertion inframe and frameshift and deletion inframe and frameshift. Substitution missense is the change of a single nucleotide base. This results in a different codon (to the original one), which then codes for a different amino acid. Substitution

[8] http://cancer.sanger.ac.uk/cosmic Accessed 25 May 2015

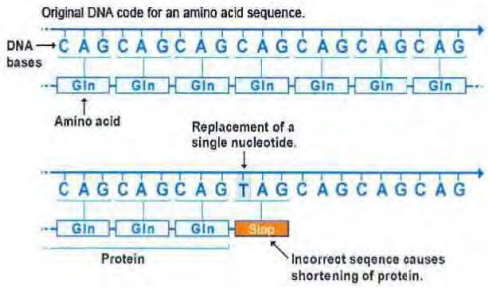
nonsense is when a single nucleotide base is substituted causing a premature stop codon, this stops the translation of the DNA sequence into an amino acid. An insertion inframe is when nucleotides are inserted in groups of 3 bases (keeping the frame the same, except for an increase in amino acids that can be coded). It therefore doesn't directly affect the process of translation unless the group of bases codes for a stop codon. This type of mutation generates a change in protein depending on the quantity of nucleotides inserted. An insertion frameshift is when the number of nucleotides inserted is not divisible by 3 (therefore changing the frame of the sequence) it will alter the process of translation (of the code into amino acids) from the point of insertion and onwards. This could create non-functioning proteins. A deletion inframe and frameshift is the same concept as the insertion inframe and frameshift, however instead of nucleotides being inserted, nucleotides are 'deleted'. The images below show two types of mutations[9]

Missense mutation



U.S. National Library of Medicine

Nonsense mutation



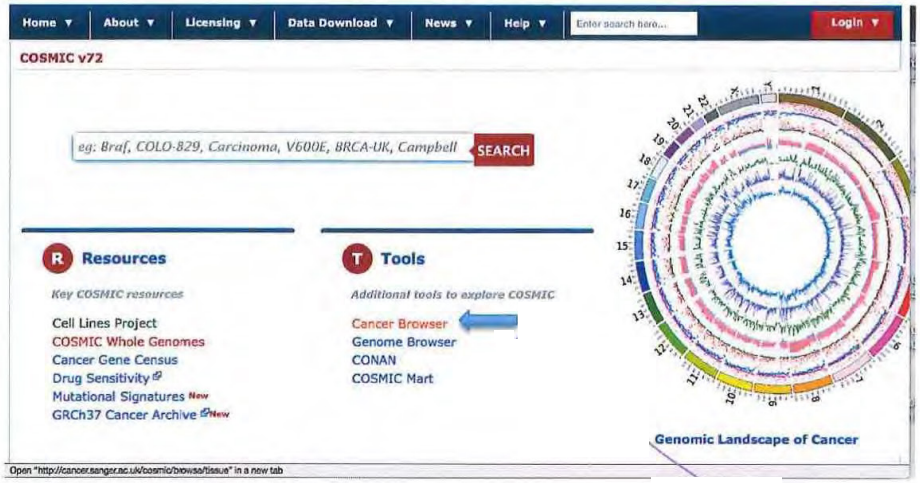
U.S. National Library of Medicine

[9] http://ghr.nlm.nih.gov/handbook/mutationsanddisorders/possiblemutations Accessed 18 June 2015

Method:

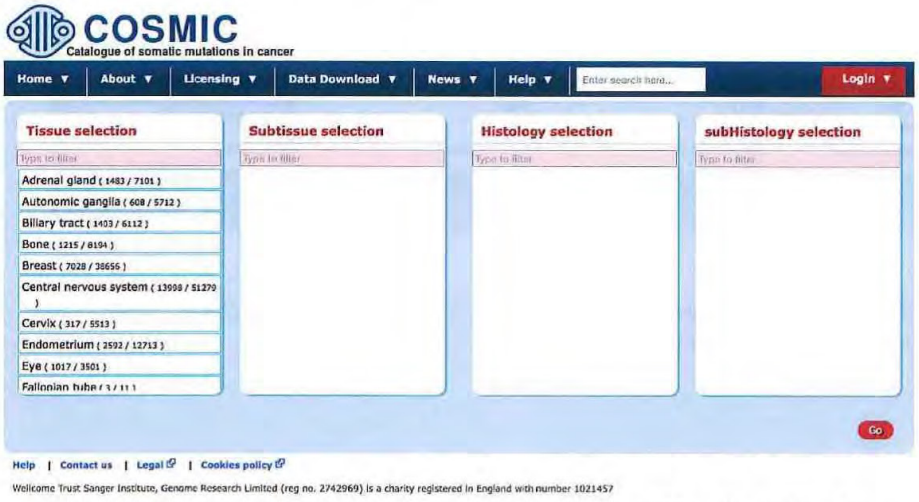
To show the different steps I have taken to find my data I have provided screenshots below with short explanations of what I have done and why I have chosen to do that certain step.

1)



This is the homepage of the cosmic database (http://cancer.sanger.ac.uk/cosmic) Accessed 25 May 2015; to find the necessary data for my study you click cancer browser (which is found under the tools sub-heading).

2)

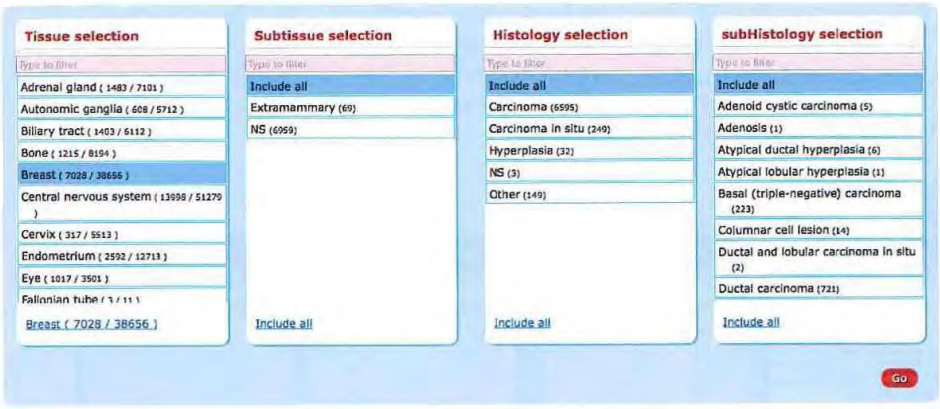


One is then able to select the different tissue that would be affected by the various types of cancer. For the data found in this comparative study I selected breast and ovaries.

3)

As previously mentioned I chose 'Include all' for the more specific selection of data. This is to ensure that I have the largest spread of data possible. The image above shows the steps and buttons I pressed to find my results for ovarian cancer.

4)



I repeated the same steps, however selecting breast as my initial tissue, to find the data for breast cancer.

Results:

5)

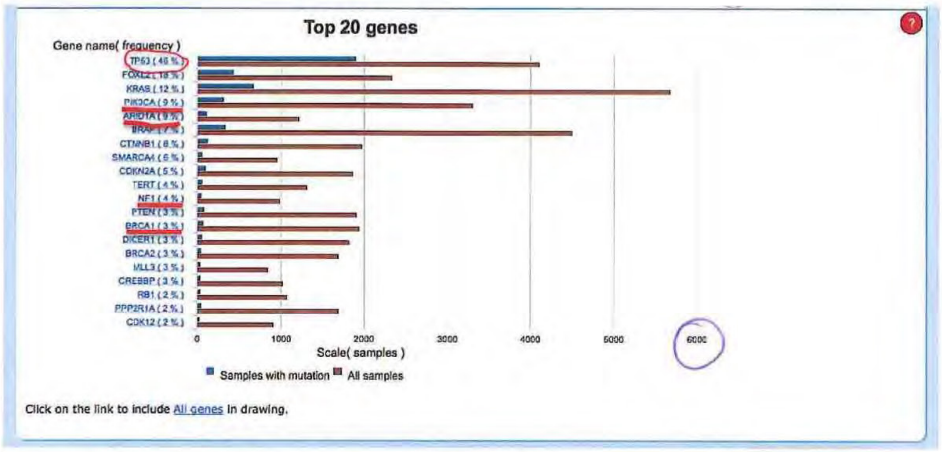


Figure 5 shows the top 20 genes that caused ovarian cancer. The blue line represents the samples with a mutation, the red line is the total quantity of samples tested for that certain mutation. The two values (of the red and blue line) are used to calculate the percentage (frequency) of that mutation

6)

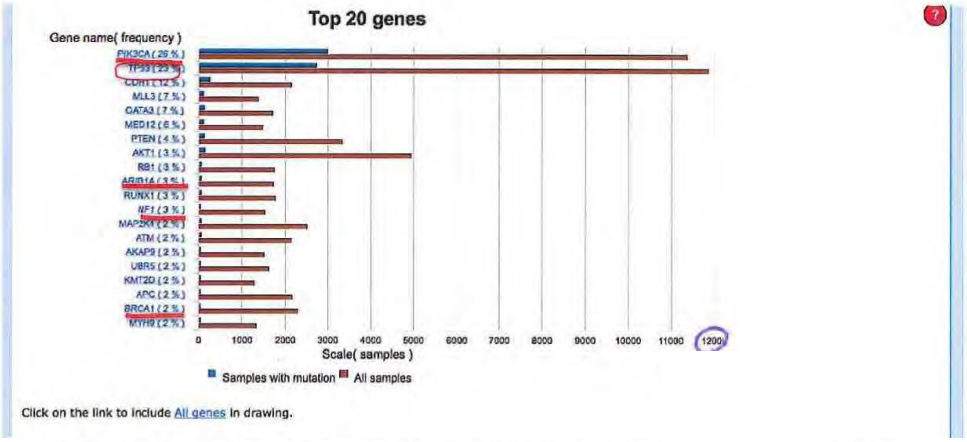


Figure 6 shows the top 20 gene's causing breast cancer: The data above shows the top 20 genes causing breast cancer. As one can see from the scale at the bottom the number of samples for breast cancer (12,000 samples) is nearly twice as high as the number of samples for ovarian cancer (6000 samples).

The genes that are circled in the top 20 genes for breast and ovarian cancer are the ones that occur in the top 3 mutations in both cancers. The ones underlined in red occur in the total top 20 mutations in both of the cancers.

Breast Cancer: Results

According to the COSMIC Database the top three genes mutating and causing breast cancer are: *PIK3CA, TP53* and *CDH1*. The *PIK3CA* gene mutation results in 26% of breast cancer cases of the tested samples. Of the 11,348 samples tested 2981 had mutations. The *TP53* mutation was responsible for causing 23% of tested cases (11869 total samples were taken of which 2726 samples were mutated). The final gene, *CDH1*, was mutated 12% of the time, clearly less than the top two genes. Of 2135 samples, 248 samples were mutated.

Calculations for the mutated samples

$$Mutation\ frequency = \frac{mutated\ samples\ for\ that\ certain\ gene}{total\ number\ of\ samples\ for\ that\ gene}$$

Ovarian Cancer: Results

Although the COSMIC database doesn't differentiate between the different types of tumours, it could be hypothesised that the most frequently occurring gene mutations correspond with the epithelial tumour, as it is the tumour that is found to cause the cancer in most women. The top 3 genes causing ovarian cancer are *TP53, FOXL2* and *KRAS*. The *TP53* gene was mutated in 46% of samples (1893 samples with mutation of a total 4095 samples tested). The second gene mutation (*FOXL2*) was found in 18% of the tested samples (416 mutated out of 2328 total samples). The third most frequent gene mutation, *KRAS*, was mutated 659 out of 5664 times resulting to 12%.

TP53 gene[ii]

This gene is the one that occurs in the top 3 gene mutations causing breast and ovarian cancer. It occurs at first place in ovarian cancer with 1893 out of 4095 samples being mutated (46%). In breast cancer 2726 samples of the 11869 were mutated accounting for 23% of all cancer cases. Although there are more mutated samples for breast cancer the total number of samples is also a lot larger. There is a possibility that the *TP53* wouldn't be the top gene mutation causing ovarian cancer if the total sample size would be increased.

The *TP53* (tumour protein p53) gene is a type of tumour suppressor that helps create the protein (tumour protein p53). It is located on the 17th chromosome at position 13.1. The job of this tumour suppressor is to control any 'unusual' division or growth of cells in an uncontrolled manner. If DNA becomes damaged due to certain factors such as radiation, or chemical exposure the *TP53* activates other genes to fix the damage (if it is able to). Otherwise, the cell is 'killed' to prevent the growth of a tumour.

Inherited mutations of the gene are rather rare in comparison to the somatically acquired ones. The somatically mutated genes reduce or completely eliminate the function of the tumour suppressor, therefore causing the abnormal growth of tissue causing the cancer.

PIK3CA gene[iii]

The *PIK3CA* gene is officially called phosphatidylinositol -4, 5-bisphosphate 3-kinase, catalytic subunit alpha. PI3K performs the actual actions whereas the other subunit regulates the enzymes activity. The PI3K adds O_2 and a phosphate group to other

proteins through phosphorylation. The enzyme phosphorylates molecules that trigger different reactions, which then emit chemical signals in the cell. This is vital for cell growth and movement, transportation of things within the cell and division. This transmitting of the special chemical signals could potentially also be involved in hormone regulation, according to certain studies.

The mutated *PIK3CA* is somatically acquired. The 2 most common mutations change glutamate to lysine (amino acids). The mutation causes the PI3K to emit unregulated chemical signals, leading to the abnormal growth of cells.

BRCA1/BRCA2 gene

Although the *BRCA1* only had a mutation rate of 3% for ovarian cancer (place 13 of the top 20 genes) and a 2% mutation rate (place 19 of the top 20 genes) in breast cancer, it is an important factor for the hereditary aspect of both ovarian and breast cancer. *BRCA2* is found at 15th place also with 3% of samples mutated in ovarian cancer, it is not found under the top 20 gene mutations causing breast cancer. This may seem rather unusual because the BRCA genes are always the genes that are automatically associated with breast cancer.

The *BRCA1* gene codes for a tumour suppressing protein, this means that it helps to fix damaged DNA. There are over 1,800 different mutations that can occur on the BRCA1 gene. Because the mutations become apparent in all of the bodys cells, they are often passed down to children, therefore being a common gene for hereditary breast cancer.

Essentially the *BRCA2* gene is very similar to the *BRCA1* gene as it also acts as a tumour suppressor. However, there are some small differences between the two genes and what their mutated gene can affect. For example women with *BRCA1* breast cancers don't respond as well to hormone therapies, making the cancer more difficult to cure. iv

BRCA Genes and Ovarian Cancer

Women that have inherited (or somatically acquired) the *BRCA1* mutation have a 54% chance of developing ovarian cancer, whereas patients with the *BRCA2* mutation only have a 27% of getting ovarian cancer.[10]

BRCA Genes and Breast Cancer

BRCA1 gene mutations cause an unusually short version of the BRCA1 to be produced. Therefore the damaged DNA is not fixed as efficiently causing other mutations within the DNA. These then lead to the development of cancerous cells. Women inheriting a mutated *BRCA1* gene have a 55-65% chance of getting breast cancer, women inheriting a mutated *BRCA2* gene only have an around 45% chance of developing breast cancer (by the age of 70)[11].

In both types of cancers, the *BRCA1* gene causes a higher percentage of cancer cases, demonstrating it may have a larger impact on the reduced function of the tumour suppressor.

Similarities and differences between ovarian and breast cancer gene mutations

---

[10] http://www.thebreastcaresite.com/before-surgery/brca1-brca2-genetic-mutations-alike-different/ Accessed 18 June 2015

[11] http://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet Accessed 18 June 2015

The gene that is found in both top 3 gene mutations is the *TP53* gene. In breast cancer, the mutation accounts for 23% (2726 samples out of 11869 were mutated) of breast cancer cases. In ovarian cancer the gene causes 46% (1893 mutated out of 4095 total samples) of cancer cases. The percentage of mutated samples in ovarian cancer is significantly higher, however the actual number of mutated samples from the breast cancer data is higher.

Results (continued):

7)

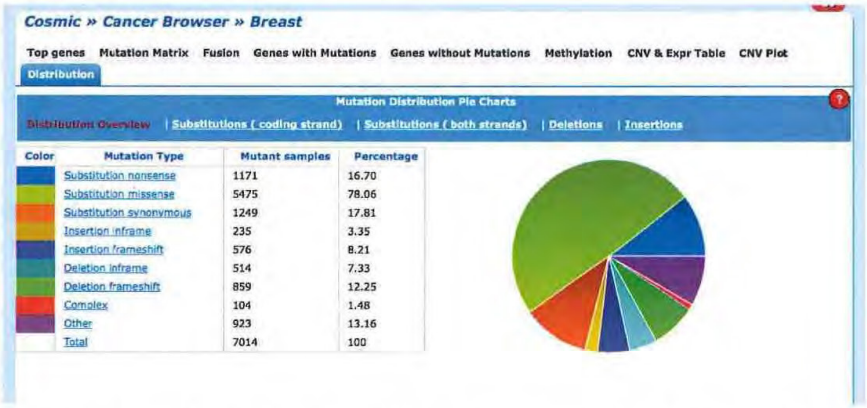


Figure 7 shows the distribution of the different mutations explained above for breast cancer that are present in 7014 samples from the database.

As the pie chart above shows the substitution missense occurred most often. The least occurring mutation of the ones I explained above was the deletion inframe. Although the substitution missense accounts for 78% of mutations in the chart it looks like it's only around 50%. This is due to the fact that in some genes more than one mutation occurred. This is also the reason why the percentages don't add up to 100.

8)

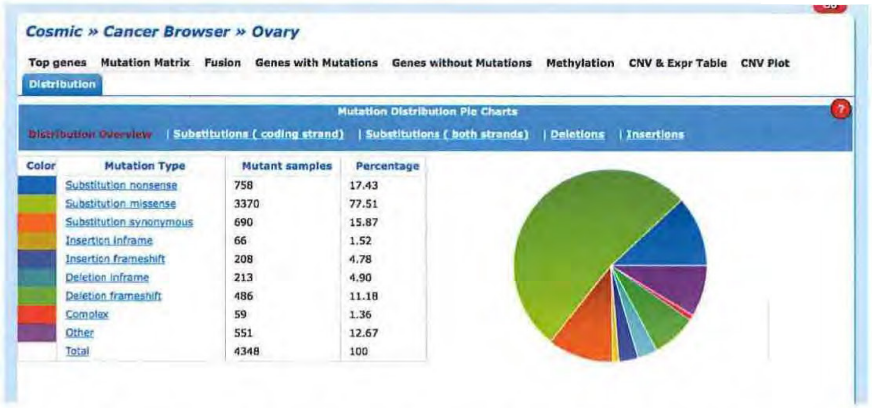
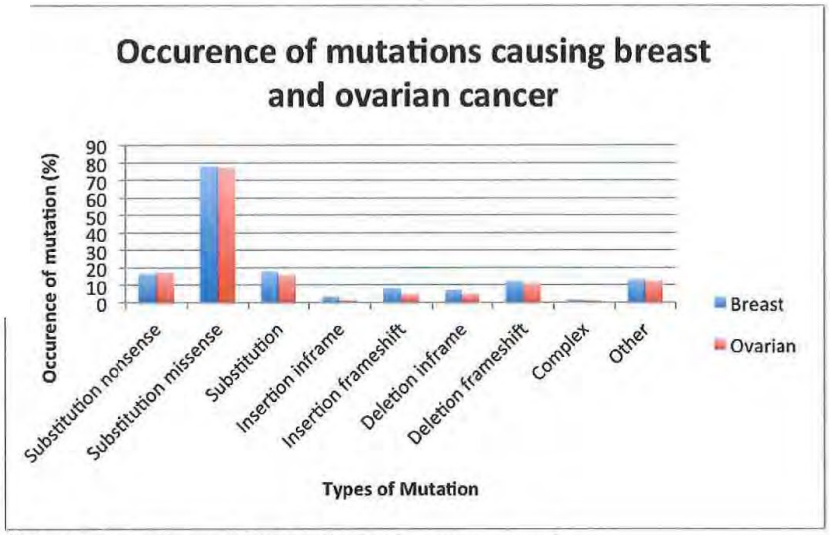


Figure 8 shows the mutations for ovarian cancer. As in breast cancer, the substitution missense accounts for most of the mutations that occurred in the samples for ovarian cancer. Insertion inframe causes the least mutations in

In both ovarian and breast cancer substitution missense causes the most genetic mutations, suggesting that it occurs most frequently in the body. This would make sense, as it is 'only' one mistake in a single nucleotide, which is more likely than an insertion/deletion of 3 nucleotides.

When looking for the specific types of mutations that are found in the top 3 genes a problem occurred. The numbers of mutated samples for e.g. substitution missense didn't add up (i.e. there were more mutated samples for the top 2 genes when added together, more meaning larger than the total mutated samples for that type of mutation). This may be due to a mutation on a variety of genes or different types of mutations that affect the total number of samples.



(Substitution is supposed to be substitution synonymous)

The graph above shows the comparison between the different mutations. It shows that the occurrence of the mutations is quite similar between the two types of cancer, suggesting that certain mutations are more likely to occur than others (e.g. substitution missense).

Analysis and Evaluation

There is definitely a connection between the occurrence of gene mutations (dependent variable) that cause breast cancer and those that cause ovarian cancer (independent variable). The TP53 gene that occurs in both top three gene mutations proves this. Furthermore, there are four other gene mutations in the overall top 20 mutations that occur in both breast and ovarian cancer, portraying some sort of similarity in the genes that cause the two cancers. Although this connection may not be extremely strong 25% of the gene mutations occur in both of the cancers. This suggest similar findings sources that have been referenced at the end and that have been mentioned at the beginning.

The numbers of samples for breast cancer (12000 samples) are nearly twice as high as the number of samples for ovarian cancer (6000 samples). This suggests that breast cancer is more common than ovarian cancer.

A reason why this data may be slightly unreliable is because the total number of samples aren't the same for each gene mutation. For example the top gene (PIK3CA) was tested a total of 11,348 times. However, the gene on the fourth position (MLL3) was only tested 1354 times. Maybe if more or less samples of a certain gene were taken the order for the 'top 20 genes' would change.

Except for that minor problem, the database is rather reliable as many people use it for professional research projects. In addition, the database gives information on how they collect their data, which is also very reliable.[12]

Furthermore, as no detailed patient information is given, it is impossible to deduct the cause of a person's cancer. It is unlikely to be hereditary as the database only includes somatically acquired gene mutations. However, it would be interesting to know about the patients' lifestyle to see what the mutagens were. If this information were to be made public (yet, still having the anonymity of name and place of residence) there would be the possibility to deduce, which mutagens cause the different types of gene mutations. There may not necessarily be a certain correlation, but it is something that could be carried out for further research, to understand the cause of certain mutations in more depth.

Some limitations to this data analysis may be only using one database. If a variety of databases were available the information would be more reliable, as the sample size would have increased and their would be a larger spectrum of samples (different nationalities, gender, age etc.) Another limitation would be the previously mentioned uncontrolled variables being the lifestyle of the patients that the samples were taken from. Anomalous results are not present in this data analysis. The samples that are a minority aren't counted as anomalous results as they are correct, just less frequently occurring.

Bibliography:

http://www.cancer.org/cancer/ovariancancer/detailedguide/ovarian-cancer-what-is-ovarian-cancer Accessed 29 May 2015
http://www.ovariancancer.jhmi.edu/hereditary.cfm#breast Accessed June 2nd 2015
http://ghr.nlm.nih.gov/gene/*TP53* Accessed 8 June 2015
http://cancer.sanger.ac.uk/cosmic Accessed 25 May 2015
Allott, Andrew, and David Mindorff. "Chapter 1: Cell Biology." *Biology*. N.p.: n.p., 2014. N. pag. Print.

---

[i] http://www.cancer.gov/types/breast/hp/breast-ovarian-genetics-

[ii] http://ghr.nlm.nih.gov/gene/TP53 TP53 gene information Accessed 20 June 2015

[iii] http://ghr.nlm.nih.gov/gene/PIK3CA PIK3CA gene information Accessed 20 June 2015

[iv] http://ghr.nlm.nih.gov/gene/BRCA1 BRCA1 gene information Accessed 20 June 2015

---

[12] http://cancer.sanger.ac.uk/cosmic/about Accessed 1 July 2015