2016

Identification of Functional Single Nucleotide Polymorphisms Associated with Breast Cancer Based on Chromatin Modifications

Laura E. Hayward
Claremont McKenna College

Identification of Functional Single Nucleotide Polymorphisms Associated with Breast Cancer Based on Chromatin Modifications

A Thesis Presented

by

Laura E. Hayward

To the Keck Science Department

Of Claremont McKenna, Pitzer, and Scripps Colleges

In partial fulfillment of

The degree of Bachelor of Arts

Senior Thesis in Biology

April 24, 2016

Table of Contents

Abstract

Breast cancer affects 1 in 8 women and can be deadly; yet when detected early enough it is often treatable. Thus, early detection of breast cancer is imperative to save lives. The success of early detection depends, in part, on being able to stratify risk. A new approach to determining risk involves identifying genetic variants that alter an individual's risk for developing breast cancer. This thesis identified key functional candidates involved in breast cancer development, some of which have been verified by other studies. For a few of the functional candidates, further research needs to be done in order to determine the biological significance they play in the development of breast cancer. The functional candidates were identified by comparing SNPs in Linkage Disequilibrium with high risk SNPS—determined by GWAS—using histone modification markers to identify functional genomic elements in breast cell lines. The results yielded three top tier candidates and multiple second tier candidates. Further research should be done in order to assess the risk involved with these variants and the underlying biological mechanism. As genetic testing becomes more accessible to the public, the identification and understanding of these high risk variants will be an essential tool in preventing and treating breast cancer.

Introduction

According to the Centers for Disease Control and Prevention, approximately 8 million people die from cancer each year.[1] It is a disease that affects not only the patient diagnosed, but also the families of the patients. In addition to the emotional component, cancer often affects family members physically, as many cancers have a significant genetic and familial component. Individuals with a first degree relative who has been diagnosed with

breast cancer have a two fold increase in risk for developing breast cancer themselves.[1] Although these statistics are disheartening, one third of all cancer related deaths can be prevented.[1] Cancer prevention includes cancer screenings for individuals at a high risk for developing cancer.[1] Given the large genetic component to cancer and the importance of screening in prevention, it is crucial to determine the specific genetic markers associated with the disease.

Certain highly penetrant genes associated with breast cancer have been previously identified in a multitude of studies.[2] However, highly penetrant genes associated with breast cancer present in the following genes: BRCA1/2, TP53, ATM,CHEK2, BRIP1, PALB2 and PTEN, account for less than 25% of the genetic risk of cancer, suggesting that the remaining genetic component of cancer risk is due to the combination of many common genetic variants.[2] Currently, there has been a shift from looking for more of these highly penetrant genes towards identifying single nucleotide polymorphisms (SNP) in specific regions that increase risk for developing cancer.[2] Below are a few specific studies that illustrate the importance of examining individual SNPs in determining an individual's risk of developing cancer.

According to Pharaoh et al.'s (2007) study on 120 variations associated with breast cancer, multiple alleles contribute to familial risk of development of cancer, as is understood in the polygenic model of cancer. In this study, 710 SNPs were examined in 4,400 breast cancer cases and 4,400 controls.[3] The authors identified multiple SNPs that are associated with breast cancer; however, each candidate is likely responsible for a slight increase in risk.[3] Although risk associated with individual single nucleotide polymorphisms are small, an accumulation of these SNPs could significantly increase a patient's risk for cancer.[3] Chan et

al. (2012) reported that patients with greater than 6 or more high risk alleles had a 90% increased risk of developing breast cancer.[4] Their study consisted of 1291 patients with breast cancer of southern Chinese descent.[4] One of the high risk SNPs identified was rs2981579, a SNP examined in this thesis. Similarly, Easton et al. conducted a genome wide association study in 2007, identifying five novel loci associated with increased breast cancer susceptibility.[2] In this study, 4,398 breast cancer cases and 4,316 controls were tested in the first stage.[2] Discovering these high risk SNPS are critical in determining a patient's overall lifetime risk for developing breast cancer.

The studies mentioned above are examples of Genome Wide Association Studies (GWAS) where the purpose of these studies is to "identify genetic risk factors for diseases that are common in the population".[5] GWAS examine variations across the human genome and identify correlations between individual variants and disease. Eventually, the results of Genome Wide Association Studies may improve screening prevention and early detection. Most SNPs identified by GWAS as strongly associated with disease have no known biological effect.[5] Only a small number of these SNPs have been shown to directly change biological function.[5] These are the important SNPs to identify, also known as the functional SNPs. The mechanisms behind which these SNPs may change biological function include changes to mRNA, DNA-binding affinities, and modifications to protein sequences.[5] The reason for this discrepancy between a SNPs association with disease and its actual effect on biological function is due to a concept known as linkage disequilibrium. Linkage Disequilibrium (LD), as defined by Bush et al. (2012), is a "property of SNPs on a contiguous stretch of genomic sequence that describes the degree to which an allele of one SNP is inherited or correlated with an allele of another SNP within a population."[5] Meaning,

the marker or tag SNP that is initially associated with an increased risk for developing a disease might not be the functional SNP contributing to disease. A tag SNP represents a larger number of SNPs in high disequilibrium that is highly associated with a disease, such as cancer.[5] In cases where there is high linkage disequilibrium with the functional SNP, it is possible that the tag SNP may be genotyped and statistically associated with the SNP, but not responsible for change in biological function. In GWAS studies, linkage disequilibrium is used to screen many variants at once with a small number of markers capturing the variants in the region.[5] The presence of linkage disequilibrium limits GWAS studies because it is impossible to know whether the statistically significant SNP has an indirect association or whether it is the functional SNP, with a direct association.[5] In this thesis, the location of the breast-cancer causing functional SNPs will be determined by comparing the SNPs in LD with the tag SNP to histone modification data in breast cell lines.

Histone modification is a key component of epigenetic regulation, and it is epigenetic abnormalities coupled with genetic alterations that are responsible for disease.[6] Epigenetics defined as changes, not affecting the actual DNA sequence, in gene expression that are heritable.[6] Since histone modifications are important markers of function it is important aspect to examine when determining which SNP affects function.[7] If a SNP is associated with a region affecting histone modification or regulation it is possible that it is a functional SNP. Histone modifications, more specifically, H3K4me1 and H3K27ac have been previously discovered at nucleosomes near enhancer elements.[8] According to Heintzman et al. (2009), enhancers have histone modification configurations that are cell specific.[9] The actual role that histone modifications place in changing enhancer function has been debated and the conclusions are controversial. Thus, we cannot conclude that these SNPS located in

regions associated with histone modifications regions directly affect enhancer function; however, it can act as a marker for functional elements. According to Mcvicker et al., the exact role of non-coding regulatory sequences in affecting disease phenotypes is not wholly understood and is a topic of interest that should be further studied.[7] Although our knowledge on the mechanism behind histone modification's involvement in cancer development is limited, it is clear that histone modifications are markers that can be used to find functional elements. According to Jones and Baylin (2007) review on the epigenomics of cancer, an important characteristic of cancer is changes in patterns of gene expression.[6] Since alterations in enhancer elements, which impact gene expression, affect the development of cancer, histone modifications are good markers to use in looking at SNPs that may increase risk of developing cancer. By examining the relationship between potential functional SNP candidates and high density of histone modifications, the actual functional SNP can be identified. For example, if a SNP is located in the same region of a gene that is associated with a histone modifications it is likely that this SNP may be responsible for biological change in function.

A lot of genome wide association studies have used populations from European descent in their research. Due to the fact that in different populations there are differing Linkage Disequilibrium patterns, highly associated SNPs in one population may not correspond to a high associative risk in another population.[5] In the case of this thesis, the population examined is that of European descent. So in looking at the SNPs below, the findings will only be helpful or indicative of a potential increased risk in strictly European populations. Further studies will need to be done to determine the high risk SNPs in other populations.

In summary, there are a multitude of unknown genetic variants known as SNPs that are likely to increase an individual's risk of developing breast cancer. Given that cancer affects so many people and is often treatable when detected early, determining these high risk variants can save lives. However, determining the exact functional SNP responsible for affecting biological function is difficult due to linkage disequilibrium. Thus, this thesis will compare SNPs—that are in LD with the high risk SNPs associated with breast cancer—with histone modification markers in breast cell lines. Looking forward, if the functional SNPs involved in breast cancer can be identified, eventually medical professionals can use this information in order to determine an individual's risk for developing breast cancer and potential prevent late detection in patients.

Methodology

In order to determine the mechanism behind highly associated SNPs, previously collected data located in the UCSC genome browser were compared to breast cell tracks containing non-coding region information from the Epigenomics Roadmap.[10] Initially, the GWAS catalog was used to locate single nucleotide polymorphisms that have been previously associated with breast cancer. The catalog provided rs2981579 as the SNP with the lowest p value (2x10^-170).[11] The catalog also provides the SNPs with the highest odds ratio, which is important to look at as it shows the SNP that has the highest risk with breast cancer. SNP rs594206 had the highest odds ratio with a ratio of 36.3.[11] This means that individuals who have this SNP have a 36.3 higher chance of getting breast cancer than the average individual.[12] Odds ratio is important to look at as it gives us an idea of what are the

chances of an individual without the variant have of developing cancer in comparison to the odds of an individual with the variant of developing cancer.[12]

Once these SNPs were identified, the SNPs were entered into the web program known as rAggr and LD list generated questions regarding the SNPs population were answered.[13] In the two SNPs examined, the population consisted of individuals from European descent. The r squared value ranged from 0.5 and 1.0, meaning that only values that had a r-squared of 0.5 or above were included in the file. Once submitted, the bed files of all the SNPs in LD with the tag SNP were obtained. This step was repeated for both tag SNPS: rs2981579 and rs594206.

The resulting data sets were added as custom tracks on the genome browser at UCSC. Once the tracks were added the location of the SNPs were initially visually compared with the layered H3K27AC track containing regulatory elements from 7 cell lines from ENCODE. Observations were made in order to determine if there were any SNPs in LD that corresponded in a peak with the layered H3K27AC track containing regulatory elements from 7 cell lines from ENCODE. The genome browser Epigenomics Roadmap data was then accessed in order to add additional tracks indicating CpG and MRE methylation sites.[10] Once the cell tracks were loaded into the genome browser and zoomed in on the location, the peaks were detected in some of the tissues. Tracks containing information on specific breast cell lines, as can be seen in Table 1, were added to the genome browser at UCSC.

After adding the tracks from breast cells lines, the SNPs in LD were compared to the cell lines. Through visual observation it became clear which SNPs were located in a region of high density in the specific breast cell tracks. Once the visual comparisons were made,

further research on these SNPS, via PubMed, was completed in order to determine if previous research could shed light on the function of the identified SNP.

The process mentioned above was completed for both: rs2981579 and rs594206.

Table 1. Description of tracks based on cell line, type of histone modification marker, ID number, and location where information was obtained

Cell Line	Histone Modification Marker	Location	ID
Breast vHMEC	H3K4me3	REMC/UCSF	35
Breast vHMEC	H3k4me1	REMC/UCSF	35
Breast Myoepithelial	H3k4me1	REMC/UCSF	80
Breast Myoepithelial	H3K4me1	REMC/UCSF	66
Breast Myoepithelial	H3K4me3	REMC/UCSF-UB	80
Breast Myoepithelial	H3K4me3	REMC/UCSF	66
7 Cell Lines	Layered H3K27AC	Encode	NA

Results

The figures below will assist in explaining the results obtained in this study. The haploytpe blocks, as can be seen below, are measured with color, where black corresponds to an r-squared valued of 1 and grey corresponds to an r-squared value of 0.5.

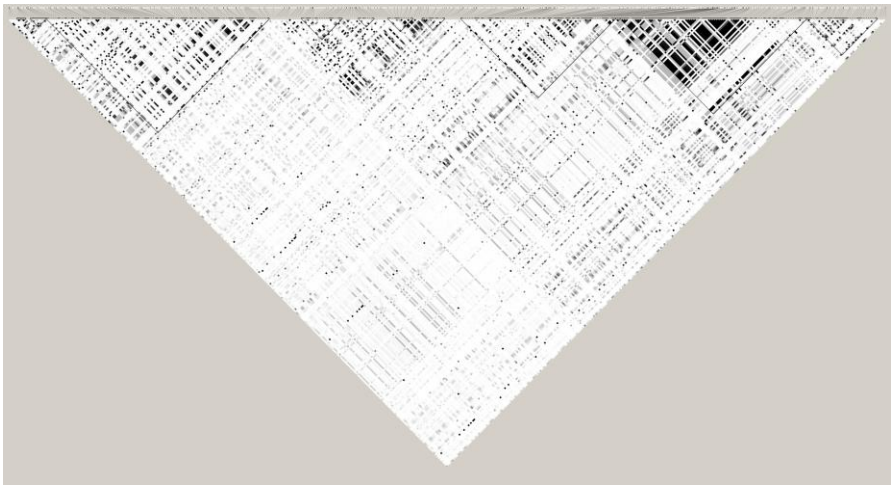

Fig 1. Haplotype blocks for rs2981579 containing all the SNPs in the area bounded by the two farthest away SNPs with an r2 value >0.2.

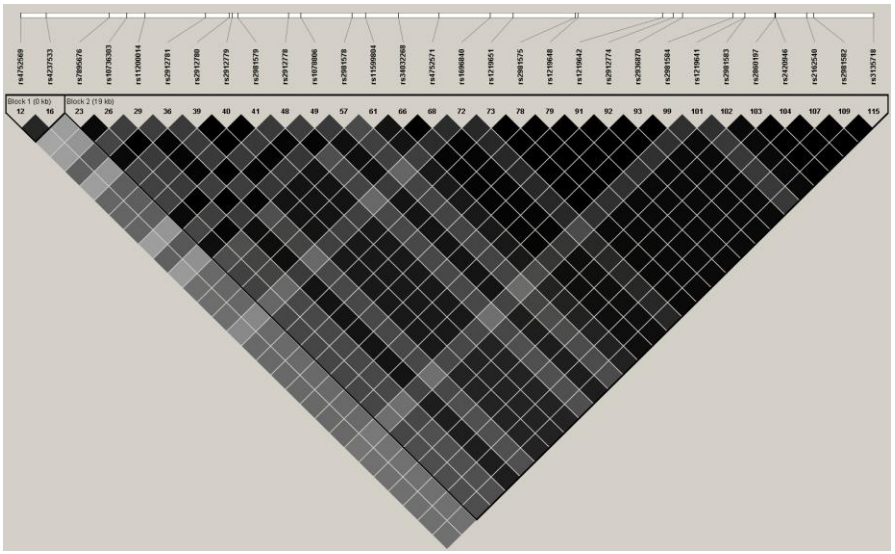

Fig 2. Haplotype blocks for rs2981579 containing SNPs in LD with and r2 >0.5.

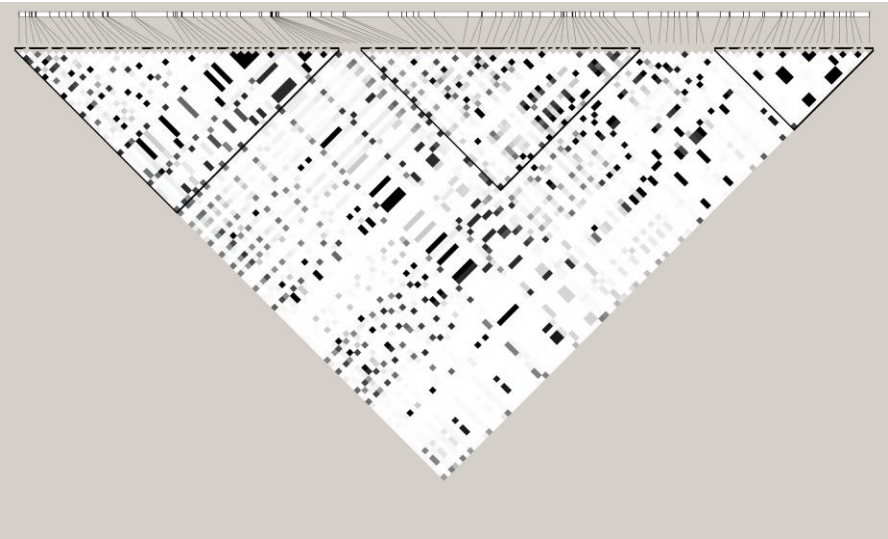

Fig 3. Haplotype blocks for rs594206 containing all the SNPs in the area bounded by the two farthest away SNPs with an r2 value >0.2.

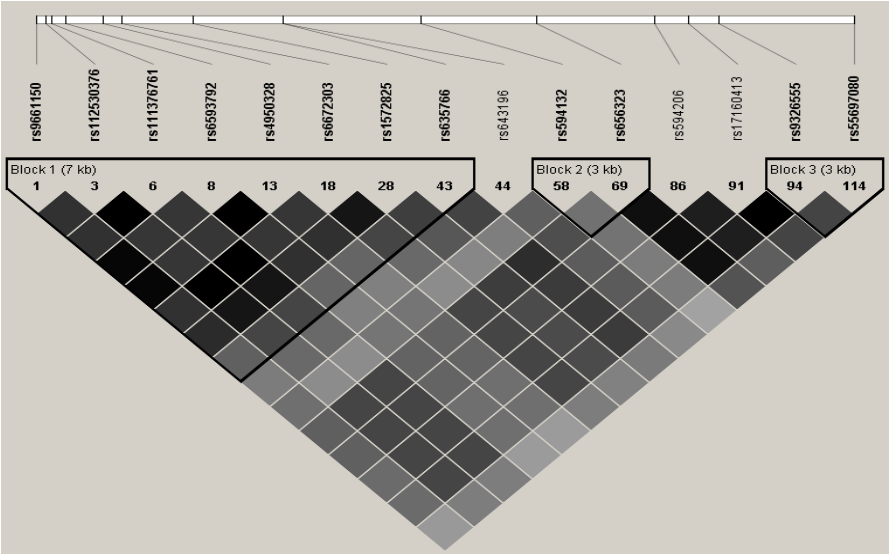

Fig 4. Haplotype blocks for rs594206 containing SNPs in LD with and r2 >0.5.

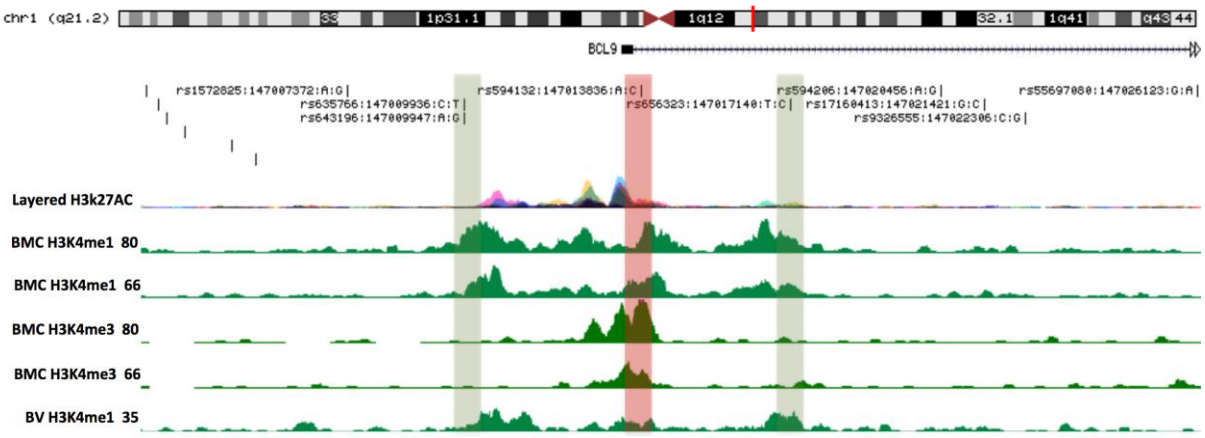

Fig 5. UCSC Genome Browser view of SNPs in LD with rs594206 indicated. ChIP-seq tracks for enhancer histone markers are Breast Myoepithelial Cells (BMC) H3K4me1 80, Breast Myoepithelial Cells (BMC) H3K4me1 66, Breast Myoepithelial Cells H3K4me1, and Breast Myoepithelial Cells H3K4me3 from the REMC/UCSF.

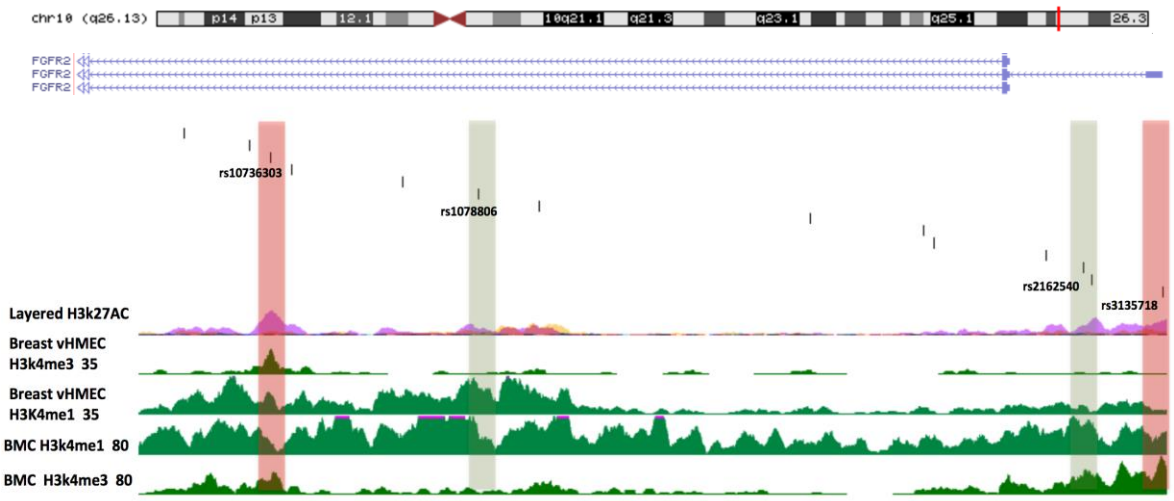

Fig 6. UCSC Genome Browser view of SNPs in LD with rs2981579 indicated. ChIP-seq tracks for enhancer histone markers are Breast vHMEC H3K4me1, Breast vHMEC H3k4me3, Breast Myoepthelial Cells H3k4me1, Breast Myoepithelial Cells H3k4me3 from REMC/UCSF.

After the visual comparison of the SNPs in LD with rs594206, rs594132 stood out as the primary candidate as the functional SNP contributing to high risk associated with rs594206 (Figure 5). SNP rs594132 mutation from A to C is located where there is a high density in the Chip-seq tracks for enhancer histone marks in the following cell lines: Breast Myoepithelial Cells (BMC) H3K4me1, Breast Myoepithelial Cells (BMC) H3K4me1, Breast Myoepithelial Cells (BMC) H3k4me1, and Breast Myoepithelial Cells (BMC) H3K4me3 80 from the REMC/UCSF. Both rs635766 and rs643196, SNPs associated with rs594206, align with a slightly dense area of the Chip-seq tracks for enhancer histone marks within Breast Myoepithelial Cells H3K4me1 track, suggesting that these are secondary functional candidates. SNP rs594206 may be a potential candidate has it is located where there are slightly high peaks in the track for enhancer histone markers in Breast vHMEC H3K4me1 and Breast Myoepithelial Cell H3K4me1 Signal From REMC/UCSF-BUC track (Figure 5). Lastly, SNP rs656323 aligns with a high density of H3K4me1 and H3K4me3 signaling and histone modifications in breast cell lines (Figure 5).

In comparing the SNPs in LD with rs2981579, SNP rs10736303 corresponded with a significantly high peak in the breast vHMEC H3K4me3 Histone Modification by ChIP-Seq Signal from REMC/UCSF track as well as with the layered H3k27AC marker often found near active regulatory elements on 7 cell lines from Encode track (Figure 6). Secondly, SNP rs3135718 is in the same location has a very strong peak in the Breast Myoepithelial Cells H3K4me3 Signal from REMC/UCSF-BUC track (Figure 6). Both SNPs 10736303 and rs3135718 are top tier functional candidates as they follow patterns of an enhancer that correspond with high peaks in regulatory regions of breast cell lines. These two SNPs are the top candidates for potential being the functional SNP in LD with rs2981579. Additionally,

there are two other SNPs that follow the pattern of second tier candidates. SNP rs1078806 is located in a region of the Breast vHMEC H3K4me1 Signal from REMC/UCSF-UBC track with a peak (Figure 6). SNP rs2981582 corresponds with a peak in the Breast Myoepithelial Cells H3K4me1 Histone Modification by ChIP-Seq Signal from REMC/UCSF track (Figure 6). Both rs1078806 and rs2981582 are second tier candidates based on the heights and number of peaks they are closely located near.

Discussion

As mentioned above, SNP rs594132 aligns with a high density of H3K4me3 histone modifications in breast cell lines. Unfortunately, previous studies reveal very little about the SNPs in LD with rs594132. Chung et al. (2013) found that rs594132 to be strongly associated with chemotherapy induced alopecia in 880 breast cancer patients.[14] However, this result is not significantly related to my topic of interest. This lack of information suggests that further studies should look at the functionality of these SNPs and the potential causal role they may play in the development of breast cancer.

A few of the candidates in LD with rs2981579, as mentioned above, have been associated with breast cancer risk in previous studies. In Kim et al. (2012) study identifying breast cancer risk variants in 6,222 Korean breast cancer patients and 5,897 Korean individuals acting as controls, rs10726303 was confirmed to have a strong association with breast cancer.[15] However, it was not the most heavily associated SNP identified to increase risk.[15] These results corroborate the findings mentioned above, confirming the possibility that rs10736303 is in fact a top candidate for being the functional SNP. In a Russian population, examined by Boyarskikh (2009), rs3135718 was found to have a stronger association with

breast cancer than a previously identified SNP, suggesting that this SNP could be a causative variant.[16] This study included 766 cases and 665 control woman from Siberia, Russian Federation to look at the possible risk variants causing breast cancer within the FGFR2 gene.[16] Similar to the last study, these results corroborate the findings we see in this thesis. Suggesting that rs3135718 is a potential functional variant playing a role in increasing risk of developing breast cancer. Even the second tier candidates are associated with previous findings. According to Anderson et al. (2013) study on breast cancer susceptibility and potential modifications by post-menopausal hormonal therapy, rs1219648 is associated with breast cancer and it is unclear whether these effects are modified by post-menopausal hormonal therapy use.[17] Similarly, SNP rs2981582 has been previously shown to be a candidate SNP for the development of breast cancer in multiple populations.[18] This association was mentioned in a study examining 81 normal breast tissues from Caucasian American, African American, and other unknown backgrounds.[18] According to this study, it is unclear whether this correlation is present in breast cancer cell tissues or simply normal breast cell tissues. The results of the study were inconclusive.

In summary, it is clear that there are a few functional candidates that could play a biological role and increase an individual's risk of developing cancer. Given the location of the SNP in correspondence with histone modification marks it seems probable that rs594206 is a top candidate involved with enhancers affecting gene regulation and biological function. Similarly, based on the results mentioned above, SNPs rs10736303 and rs3135718 in LD with rs298157 are top tier functional candidates, which has been confirmed by their association with breast cancer in previous studies. The second tier candidates, rs1219648 and rs2981582, in LD with rs298157 have been shown to have an association with breast cancer

in studies. The candidate functional SNPs in LD with rs298157, even the second tier candidates, have been found to be associated with breast cancer in previous studies. However, there is more research that needs to be done in examining the under studied functional candidates that are in LD with rs594206.

Ultimately, discoveries made through GWAS studies allow medical professionals to determine, which individuals are at a significantly higher risk than the general public for various disease, creating personalized and preventative care for patients.[5] This information gathered in GWAS studies is not only used in predicting who is at higher risk for developing breast cancer, but it can be utilized in creating new pharmacologic therapies.[5] Similarly, dosing for existing pharmacologic therapies is affected by genes.[5] In understanding a patient's genetic variations, the patient's ability to respond to certain drugs and the appropriate dosage can be determined.

When thinking about the implications of these Genome Wide Association Studies it is important to understand the common disease common variant hypothesis. The common disease common variant hypothesis is, "that common disorders are likely influenced by genetic variation that is also common in the population."[5] This hypothesis ordinates from the fact the penetrance for any single variant that may be involved in a common disease is small.[5] Combined with the fact that common diseases frequently show high heritability, disease susceptibility for these common diseases may therefore be determined by many variants.[5] In order to truly determine the risk of an individual an examination of multiple variants is required.

Due to the implications of this hypothesis, more studies are necessary to be able to improve risk estimates that are based on the contribution of many individual variants. Yet at

some point risk estimates need to help identify the point at which a patient's risk is high enough to take preventative measures. Although there are non-invasive screening measures (MRIs and mammograms) the decision to have a mastectomy to prevent breast cancer is an incredibly important choice. How strong must risk be to warrant the recommendation for mastectomy? The answer could be individualized; however, there needs to be data to help women even consider a prophylactic mastectomy. According to Janssen and Dujin (2008), the predictive value of common diseases based on multiple genetic variants is limited.[19] Not enough susceptibility variants have been identified. In order to improve the predictive value of the risk associated common variants, studies must further identify high-risk variants associated with common diseases, such as breast cancer.

Looking at an individual's genome, for high risk variants is an example of personalized medicine. In general, personalized medicine utilizes information from an individual patient's genetic background and biological features to treat a patient, ensuring that the patient's healthcare is fitted to their specific needs.[14] Quantifying high risk variants will further the field of personalized medicine and hopefully positively change the way that we understand complex multifactorial etiology.[14] However, universal personalized medicine is only a possibility if it is not available to the public. Given the implications of CDCV hypothesis, 500,000 to a million common SNPs would need to be genotyped for each individual, who is of European descent, in order to determine risk levels for a common disease for every individual.[5] For this to be feasible, genotyping technologies need to be very cost-effective. Fortunately, the current chip-based technologies are likely to be replaced by new inexpensive technologies.[5] With these decreases in cost, using genetic variants to assess risk of breast cancer in individuals can become a reality.

The top functional candidates identified earlier in this thesis may play a causal role in the development of breast cancer. Hopefully, with further research in this area, and specifically looking at these SNPS, the use of these variants in assessing breast cancer risk can be determined. Eventually this information should be utilized to accurately ascertain the individual's risk of developing breast cancer and assist in preventing this disease that negatively affects so many individuals and families across the world.

REFERENCES

1. CDCBreastCancer. World Cancer Day. *Centers for Disease Control and Prevention* (2016). Available at: http://www.cdc.gov/cancer/dcpc/resources/features/worldcancerday/index.htm. (Accessed: 1st April 2016)
2. Genome-wide association study identifies novel breast cancer susceptibility loci : Article : Nature. Available at: http://www.nature.com/nature/journal/v447/n7148/full/nature05887.html%3Freferer=www.clickfind.com.au7?message=remove&referer=www.clickfind.com.au7. (Accessed: 1st April 2016)
3. Pharoah, P. D. P. *et al.* Association between Common Variation in 120 Candidate Genes and Breast Cancer Risk. *PLOS Genet* **3,** e42 (2007).
4. Chan, M. *et al.* Association of common genetic variants with breast cancer risk and clinicopathological characteristics in a Chinese population. *Breast Cancer Res. Treat.* **136,** 209–220 (2012).
5. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLOS Comput Biol* **8,** e1002822 (2012).
6. Jones, P. A. & Baylin, S. B. The Epigenomics of Cancer. *Cell* **128,** 683–692 (2007).
7. McVicker, G. *et al.* Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science* **342,** 747–749 (2013).
8. Chromatin Modifications in Enhancers. Available at: http://epigenie.com/chromatin-modifications-in-enhancers/. (Accessed: 14th April 2016)

9. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
10. Roadmap Epigenomics Project - Home. Available at: http://www.roadmapepigenomics.org/. (Accessed: 14th April 2016)
11. Catalog of Published Genome-Wide Association Studies. Available at: https://www.genome.gov/26525384. (Accessed: 1st April 2016)
12. NCI Dictionary of Cancer Terms. *National Cancer Institute* Available at: http://www.cancer.gov/publications/dictionaries/cancer-terms. (Accessed: 1st April 2016)
13. rAggr | Home. Available at: http://raggr.usc.edu/. (Accessed: 14th April 2016)
14. Chung, S. *et al.* A genome-wide association study of chemotherapy-induced alopecia in breast cancer patients. *Breast Cancer Res. BCR* **15,** R81 (2013).
15. Kim, H. *et al.* A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. *Breast Cancer Res. BCR* **14,** R56 (2012).
16. Boyarskikh, U. A. *et al.* Association of FGFR2 gene polymorphisms with the risk of breast cancer in population of West Siberia. *Eur. J. Hum. Genet.* **17,** 1688–1691 (2009).
17. Andersen, S. W. *et al.* Breast cancer susceptibility associated with rs1219648 (FGFR2) and postmenopausal hormone therapy use in a population-based U.S. study. *Menopause N. Y. N* **20,** 354–358 (2013).
18. Sun, C., Olopade, O. I. & Di Rienzo, A. rs2981582 is associated with FGFR2 expression in normal breast. *Cancer Genet. Cytogenet.* **197,** 193–194 (2010).

19. Janssens, A. C. J. W. & Duijn, C. M. van. Genome-based prediction of common diseases: advances and prospects. *Hum. Mol. Genet.* **17,** R166–R173 (2008).