Interpreting the Prevalence of Regulatory SNPs in Cancers and Protein-Coding SNPs among Non-Cancer Diseases Using GWAS Association Studies

Zoya Khalid[1§], Osman Ugur Sezerman[1]

***Abstract*— Biological mechanisms underlying diseases are quite challenging to understand, as there exists a complex relationship between human genetics and disease traits. Genome-wide association studies are potent methods in identifying single nucleotide polymorphisms (SNPs) which are linked with a large number of phenotypes. Even though GWAS list down the statistically significant SNPs which are found to be associated with phenotype of interest, still there is a need to look for the direct evidence regarding biological processes to better understand the disease development mechanism since it may differ in different complex diseases depending on the nature of the disease. From previous few years, there are above 11 million SNPs that have been labelled in databases specifically in dbSNP. Among them, the SNPs can be categorized as coding or non-coding SNPs depending on their location in the genome. Lethal changes in the coding parts of the genes might play role in development mechanism of complex diseases by directly affecting the functionality of the protein. Similarly those SNPs which are present at the regulatory regions splice regions, micro RNA binding sites and epigenetic sites might affect the level of gene expression and ultimately contributes in complex disease formation. We designed our study to understand the biological mechanism underlying complex diseases by performing statistical analysis on GWAS dataset. Seventeen different cancer types with the non-cancer diseases (autoimmune, Neurodegenerative and metabolic diseases) has been selected in order to understand the major factors involved in disease progression and the impact of SNPs in disease development. The statistical analysis includes the chi squared hypothesis testing with the null hypotheses H_0: SNPs in coding and non-coding areas are not significantly different and with the alternate hypothesis as H_1: They are significantly different.The results revealed that Complex diseases like Cancer are mostly caused by mutations occurring at non coding regulatory sites thus causing changes at expression levels of the genes involved, such as over expression of oncogenes and under expression of tumor suppressor genes as expected, whereas in other non-cancer diseases, mutations occurring at coding regions of the genes play more determinative role. These mutations change the functionality of the protein product thus having a direct impact on the autoimmune response. This study in future can be taken as a reference study for analysing coding and non-coding parts of the genome regarding divulging biological mechanisms involving complex diseases.**

I. INTRODUCTION

The first drift of large scale genome wide association studies (GWAS) has contributed in understanding the mechanism underlying complex diseases. For some diseases including Breast cancer, Prostate, Asthma and type II diabetes there has been speedy growth of loci involved in predisposition. From the studies it has been found that GWAS studies are suitable for the identification of SNP variants with effects on phenotype. [1] The major portion of medical research has inclined towards the protein-coding mutations owing to the fact that the mechanisms underlying the regulatory SNPs are quite complicated and still not fully understood. SNPs whether in the coding or non-coding areas may be detrimental and may contribute in the development of complex diseases. Genome Wide Association studies which includes hundreds and thousands of SNPs which are tested concurrently in large number of cases and control samples in order to associate them with the complex disease have developed the hunt for genetic basis of these diseases. The success of GWAS can be seen from the fact that it identified novel SNPs and have been hypothesized as valuable tools for finding complex disease genes with the help of association studies and, afterwards to be used as markers for further genetic analysis. [2, 3] Understanding a disease etiology is the challenging task for biologists as it is the first step for disease diagnosis and treatment. Among the complex diseases Asthma is one of the types. Asthma is an inflammatory disease of the airways which is characterized by intrusion and activation of inflammatory cells followed by structural changes. These changes are supposed to associate with the severity of asthma and moderately with the development of progressive lung function weakening. The principal mechanism including airway angiogenesis in asthma and its detailed clinical significance has not yet been fully revealed [4]. Asthma is a common, complex disease that is affecting more than 300 million people worldwide. [5] Similarly rheumatoid arthritis is also a heterogeneous autoimmune disease followed by caustic inflammation particularly in joints. The pervasiveness in the common population is about 0.5% to 1%, and among them women are at more risk for developing

the disease.[6,7] Among the cancer types, Prostate cancer is the second most common cancer among men population. The prostate cancer shares some common features with other cancer types like breast cancer and the colon cancer. GWAS studies have been contributing successfully in identifying common variants (SNPs) which are significantly associated with the Prostate cancer. [8] Similarly for the other cancer types GWAS appeared to be as a potent tool for identifying diseased loci in cancers [9].

In this study we selected seventeen different types of diseases including both cancer and non-cancer phenotypes namely Asthma, Arthritis, Parkinson Disease, Crohn's Disease, Hypertension, Obesity, Schizophrenia, Alzheimer, Multiple Sclerosis and for cancer types Breast cancer, Prostate cancer, Pancreatic, Colorectal, Liver, Esophageal and Bladder. We compared the ratio of SNPs distribution among all the disease types selected depending on their location in the genome by applying chi-squared test statistics.

II. METHODOLOGY

The methodology has been designed in order to find the significant SNPs and to analyse the ratio of these top selected significant SNPs in coding and non-coding regions of genome among different diseases and to further distinguish the driver mutations from the neutral ones in the coding regions of cancer. The SNP's were collected from GWAS. GWAS usually emphasizes on associations regarding single-nucleotide polymorphisms (SNPs) and complex diseases. We collected GWAS data for different cancer and non-cancer diseases. The significant SNP's were downloaded from https://www.gwascentral.org/

The significant SNPs are functionalized into coding and non-coding genomic regions by using SNPnexus Server. It is a functional annotation tool to evaluate the likely significance of candidate variants which is then linked to the gene/protein isoforms that might be phenotypically important. It can be accessed through http://www.snp-nexus.org/

For statistical inference, Chi-squared Test analysis was performed to examine that whether the mutations observed in coding regions are significantly different form the non-coding ones.

The following parameters are followed.

- Alpha value : 0.05 that means 95 % confidence interval.
- If the test statistics value is higher and the P-value is less than alpha ($p < 0.05$) the hypothesis is rejected that means two variables are statistically different.
- If the test statistics value is lower and the P-value is higher than alpha ($p < 0.05$) the hypothesis is accepted that means two variables are not statistically significant.

And we took hypothesis as follows

H_0: SNPs in coding and non-coding areas are not significantly different. (Null Hypothesis)

H_1: They are significantly different. (Alternate hypothesis)

III. RESULTS

Starting with the most significant SNPs obtained, polymorphisms belonging to selected cancer and non-cancer types have been analyzed. It has been observed that SNPs are occurring at different rates depending on their location in the genome i-e coding and non-coding areas. Table 1 and 2 summarized the distribution of SNPs in coding and non-coding areas per disease type and Figure 1 further depicts the distribution of non-coding parts intro introns and downstream regions. The first part of our study is to analyze the ratio of SNPs among cancer and non-cancer types depending on their genome location. For this the chi squared test analysis was performed. The two hypotheses selected are mentioned in the methods part. From the results it has been observed that for cancer diseases the value of test statistic obtained was 185.0561 with the p-value 0.000045, while for non-cancer ones it was 30.69 with p-value 3.6315e-08. On the basis of these values obtained it has been clear that our null hypothesis has been rejected and the alternative hypothesis has been accepted. The alternative hypothesis states that the SNPs are occurring at significantly different rates in the coding and non-coding regions. The results are mentioned in Table 3. Furthermore the ratio of coding and non-coding SNPs among cancer and non-cancer diseases can also be graphically viewed in Figure 2. From the figure it has been visualized that in cancer like diseases the SNPs are more prevalent in non-coding parts in comparison to the coding ones, while the scenario is quite opposite in non-cancer diseases where the coding SNPs are in higher proportion than the non-coding ones.

IV. DISCUSSION

From previous few years, there are above 11 million SNPs that have been labelled in databases specifically in dbSNP. Among them, the SNPs can be categorized into coding and non-coding SNP depending on their location in the genome. Lethal changes in the coding parts of the genes might play role in development mechanism of complex diseases by directly affecting the functionality of the protein. Similarly those SNPs which are present at the regulatory regions splice regions, micro RNA binding sites and epigenetic sites might affect the level of gene expression and ultimately contributes in complex disease formation. Biological mechanisms underlying diseases are quite challenging to

understand, as there exists a complex relationship between human genetics and disease traits. A Mendalian disease can occur as a result of a single mutation on a gene that explains all the disease cases, however, complex diseases do not associate strongly with a single gene mutation, and rather they are caused by variant forms of several genes which can be the direct result of several single nucleotide polymorphisms targeting these genes. Among various diseases, cancer is one of the multifactorial diseases because of the fact that they are likely associated with the effects of multiple genes in permutation with additional factors including lifestyle and environmental factors. To understand the biology behind complex disorders is very challenging, since the root cause for most of these disorders have not been identified so far. Genome Wide Association studies includes hundreds and thousands of SNPs which are tested concurrently in large number of cases and control samples in order to associate them with the complex disease have developed the hunt for genetic basis of these diseases. The success of GWAS can be seen from the fact that it identified novel common genetic risk factors involved with the significance of earlier recognized generic variants. Hence it is known that concentrating on few SNP and genes showing strong association with the disease is not sufficient to understand the underlying disease mechanism, because there exists a chance that those biologically important genetic variants that have a small disease risk are might get overlooked.

Bearing in mind all these factors we designed our study to understand the biological mechanism underlying complex diseases with GWAS studies. We analysed seventeen different types of cancer and non-cancer disease (autoimmune, neurodegenerative and metabolic). The top selected most significant SNPs which are ranked according to their p-values are filtered out. The threshold was set to p-value <0.05. We compared the frequency distribution for SNPs present in the coding and non-coding regions for all the disease types tested in this study. The analysis revealed that for cancer diseases more crucial part of research is the non-coding or the junk DNA. Mutation lying in this region is the regulatory mutation, which does not translate into protein but still has multiple impacts on disease. The regulatory mutations alter the ability of a transcription factor to bind to DNA which further affects the gene expression level and ultimately leading to disease. These regulatory mutations are likely the drug targets, so it's certainly important to look for these mutations in order to fully understand the biological process behind this and furthermore to target correct drugs for treatment and therapies.[10, 11]

.

TABLE I DISTRIBUTION OF CODING AND NON-CODING SNPS IN CANCER

*Corresponding Author (e-mail: zoyakhalid@sabanciuniv.edu).

Author Two (e-mails: ugur@sabanciuniv.edu).

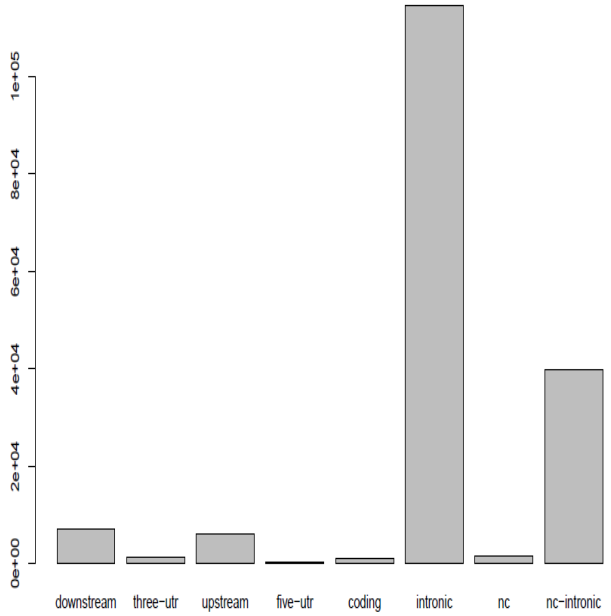

Figure 1 Distribution of non-coding SNPs across the genome.

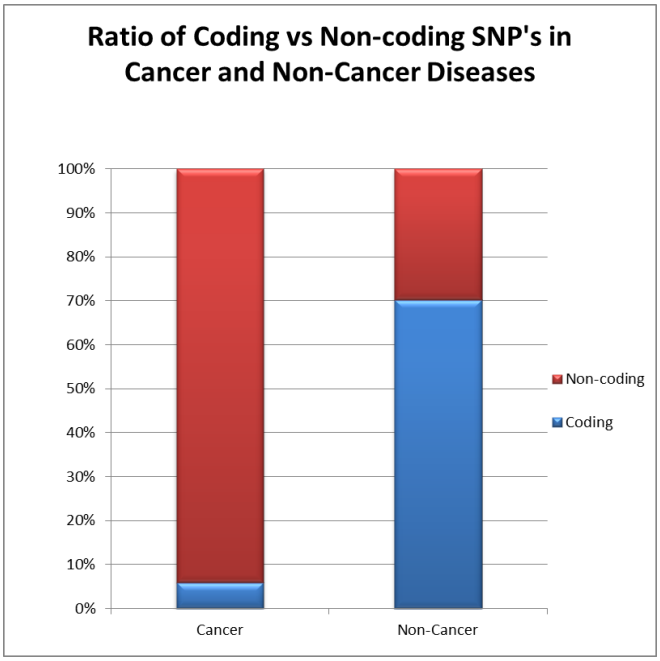

Figure 2 SNP frequencies in coding and non-coding regions of the Cancer and Non-Cancer Diseases

Cancer Name	Coding SNPs (Percentage)	Non-coding SNPs (Percentage)
Breast Cancer	10%	90%
Prostate Cancer	20%	80%
Liver Cancer	8%	92%
Pancreatic Cancer	15%	85%
Colorectal Cancer	18%	82%
Leukemia	12%	88%
Bladder Cancer	13%	87%
Esophageal Cancer	16%	84%
Total no of SNPs	**1818**	**29025**

TABLE II DISTRIBUTION OF CODING AND NON-CODING SNPS IN NON-CANCER DISEASES

Non-Cancer Disease	Coding SNPs (Percentage)	Non-coding SNPs (Percentage)
Asthma	90%	10%
Arthritis	85%	10%
Crohn's Disease	70%	30%
Multiple Sclerosis	80%	20%
Parkinson Disease	78%	22%
Obesity	75%	25%
Hypertension	90%	10%
Alzheimer	85%	10%
Schizophrenia	80%	20%
Total Number of SNPs	**2025**	**750**

TABLE III SUMMARY OF CHI SQUARE TEST STATISTIC

Mutations	Chi-squared Test Statistic	P-value
Cancer	185.0561	0.000045
Non-Cancer	30.369	3.6315e-08

V. CONCLUSION

The analysis revealed that Complex diseases like Cancer are mostly caused by mutations occurring at non coding regulatory sites thus causing changes at expression levels of the genes involved, such as over expression of oncogenes and under expression of tumor suppressor genes as expected. Whereas in other non-cancer diseases (autoimmune diseases, metabolic and Neurodegenerative diseases), mutations occurring at coding regions of the genes play a more determinative role. These mutations change the functionality of the protein product thus having a direct impact on the autoimmune response. This study in future can be taken as a reference study for analysing coding and non-coding parts of the genome regarding divulging biological mechanisms involving complex diseases

REFERENCES

[1] McCarthy MI, Abecasis RJ, Cardon RL, Goldstein BD, Little J,Ioannidis and Hirschonn,"Genome-wide association studies for complex traits consensus, uncertainty and challenges"**,** *Nat Rev Genet*, vol 9,pp, 356-69,2008.

[2] Martin, ER, Lai HE, Gilbert JR, Rogala AR, Afshari AJ, Riley J, Finch KL, Stevens JF, Livak KJ and Vence JM," SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease",*Am J Hum Genet*,vol 67,pp. 383-94,2000.

[3] Erichsen HC and Chanock SJ, SNPs in cancer research and treatment, *Br J Cancer*, Vol. 90 pp.747-51,2004.

[4] Ribatti, D, et al, "Angiogenesis in asthma",*Clin Exp Allergy*, Vol.39, pp. 1815-21,2009.

[5] Torgerson DG, et al, "Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations", *Nat Genet*,, Vol.43 pp.887-92,2011.

[6] Hughes LB, Renolds RJ, Brown EE, Keley JM, Thomson B, Conn DL, Jonas BL, Westfall AO, Padilla MA, Callahan LF, Smith EA, Brasington RD, Edberg JC, Kimberly RP, Moreland LW, Plenge RM and Bridgeqes SL, "Most common single-nucleotide polymorphisms associated with rheumatoid arthritis in persons of European ancestry confer risk of rheumatoid arthritis in African Americans", *Arthritis Rheum*, Vol.62 pp.3547-53, 2010.

[7] Olsson LM, Lindqvist AK, Kallberg H,Padyukov L, Burkhardt H, Alfreddson L, Klareskog L, Holmdahi R,"A case-control study of rheumatoid arthritis identifies an **a**ssociated single nucleotide polymorphism in the NCF4 gene, supporting a role for the NADPH-oxidase complex in autoimmunity", *Arthritis Res Ther*, Vol.9, R98,2007.

[8] Chen R, Ren S and Sun Y, "Genome-wide association studies on prostate cancer: the end or the beginning?", *Protein Cell* 2013 .

[9] Garcia-Closas M, et al, "**G**enome-wide association studies identify four ER negative-specific breast cancer risk loci", *Nat Genet* ,Vol.14,pp. 392-8, 398,2013.

[10] Chen, Y, Hao J,Jiang W, He T, Zhang X, Jiang T and Jiang R : "Identifying potential cancer driver genes by genomic data integration", *Sci Rep*, Vol.3 2013,pp. 3538,2013.

[11] Khurana E, et al, "Integrative annotation of variants from 1092 humans: application to cancer genomics", S*cience*, Vol.342,pp.1235587,2011.