Online ISSN : 2249-4626
Print ISSN : 0975-5896
DOI : 10.17406/GJSFR

Global Journal

of Science Frontier Research: G

Bio-Tech & Genetics

Genes Conferring Resistance

Multivariate Regression Analysis

Highlights

Studies in Molecular Medicine

Test Statistic for Model Selection

Discovering Thoughts, Inventing Future

VOLUME 21 ISSUE 1 VERSION 1.0

Global Journal of Science Frontier Research: G
Bio-Tech & Genetics

Global Journal of Science Frontier Research: G

Bio-Tech & Genetics

Volume 21 Issue 1 (Ver. 1.0)

Open Association of Research Society

The opinions and statements made in this book are those of the authors concerned. Ultraculture has not verified and neither confirms nor denies any of the foregoing and no warranty or fitness is implied.

Engage with the contents herein at your own risk.

The use of this journal, and the terms and conditions for our providing information, is governed by our Disclaimer, Terms and Conditions and Privacy Policy given on our website http://globaljournals.us/terms-and-condition/menu-id-1463/

By referring / using / reading / any type of association / referencing this journal, this signifies and you acknowledge that you have read them and that you accept and will be bound by the terms thereof.

All information, journals, this journal, activities undertaken, materials, services and our website, terms and conditions, privacy policy, and this journal is subject to change anytime without any prior notice.

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027
Employer Identification Number (EIN):
USA Tax ID: 98-0673427

Global Journals Inc.

*(A Delaware USA Incorporation with "Good Standing"; **Reg. Number: 0423089**)*
Sponsors: *Open Association of Research Society*
Open Scientific Standards

Publisher's Headquarters office

Global Journals® Headquarters
945th Concord Streets,
Framingham Massachusetts Pin: 01701,
United States of America
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Journals Incorporated
2nd, Lansdowne, Lansdowne Rd., Croydon-Surrey,
Pin: CR9 2ER, United Kingdom

Packaging & Continental Dispatching

Global Journals Pvt Ltd
E-3130 Sudama Nagar, Near Gopur Square,
Indore, M.P., Pin:452009, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org*
Investor Inquiries: *investors@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Excluding Air Parcel Charges):

Yearly Subscription (Personal & Institutional)
250 USD (B/W) & 350 USD (Color)

Editorial Board

Global Journal of Science Frontier Research

Contents of the Issue

Global Journal of Science Frontier Research: G
Bio-Tech & Genetics
Volume 21 Issue 1 Version 1.0 Year 2021
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

Genome Wide Association Studies in Molecular Medicine

By Dr. Fahmida Khatoon
Jinnah University

Chapter 1: Introduction- The achievements in Human Genome Project and subsequent advancements observed in Genotyping techniques have led an influx of exciting new era of discovering human genome and genetics. These technologies have provided the scientists with a comprehensive data on human genomes and have introduced state-of-art methods of investigating the fundamentals of forensic medicine (Moore 2009). The Human genome is now capable to incite in depth and precise data information and allows access into processes that extract detailed sequences for genomics in order to analyze and answer not only theoretical questions but also identifies the practically feasible genetic characteristics (Moore 2009).

GJSFR-G Classification: FOR Code: 060199

GENOMEWIDEASSOCIATIONSTUDIESINMOLECULARMEDICINE

Strictly as per the compliance and regulations of :

Genome Wide Association Studies in Molecular Medicine

Dr. Fahmida Khatoon

I. Chapter 1: Introduction

The achievements in Human Genome Project and subsequent advancements observed in Genotyping techniques have led an influx of exciting new era of discovering human genome and genetics. These technologies have provided the scientists with a comprehensive data on human genomes and have introduced state-of-art methods of investigating the fundamentals of forensic medicine (Moore 2009). The Human genome is now capable to incite in depth and precise data information and allows access into processes that extract detailed sequences for genomics in order to analyze and answer not only theoretical questions but also identifies the practically feasible genetic characteristics (Moore 2009).

Genomic sequences aid in appreciating the origination and nature of the particular mutations in genetic data and thus, incorporate functional genetics, biological sciences and bioinformatics to extract the cause of a variation in genetic framework. In the evolutionary field of forensics, the combination of information and the understanding of genetic architecture lead to dissect the unidentified entities and translate for making connections in solving ongoing genotype and phenotype puzzles of unknown nature (Moore 2009).

The last five years in the field of genomic studies have been an era of ensurgement of discoveries for epidemiological studies associated with the genetic and genomic technological advancements. High-Throughput genotyping and sequencing centers are containing torrents of data and information related to surveying genetic variants across the genome for association of complex diseases, human phenotypical characteristics and quantitative traits (Park et al. 2010). The Genome Wide Association Studies has been the work force for these efforts and now is considered as a standard technique for disease gene mapping.

There are two fundamental and statistical challenges that Genome Wide studies currently faces, the first one is the majority of identified variants have minor impacts and so collectively only explains a small proportion of the entire genetic variances (Gibson, 2010; Park et al. 2010). Secondly, the total number of Single Nucleotide Polymorphisms (SNPs) being analyzed are usually over half a million, and is often larger in magnitude as compared with the number of genetic observations recorded (Gibson, 2010). Referring to current situation, even a minute amount of error in the information about genetic data can have wide adverse effects that may greatly cost the power and credibility of Genome Wide Association Studies.

According to GWAS, the inductive spurious associations between a phenotype for the pathology and a variance or in case when there is no association between a true variant and the disease phenotype, subsequently assists in evaluating true relationships between genetic variables (Gibson, 2010). Therefore, it has been crucial for the success of genomic association studies to execute careful Quality Control (QC) and appropriate data cleaning before using the data for testing for genetic association purposes (Gibson, 2010).

a) The Empirical Hallmarks for Misidentification in Forensic Genomic Sampling

In recent times, Laurie et al (2010) and Anderson et al (2010) signified that the best guidelines for successful genomic association studies is the inclusion of per individual and per-SNP steps in sequencing. One essential error that appears during genomic association studies is the misidentification of the samples which result from pipetting errors in the laboratory during analyzing the samples, labeling mistakes and other data or information handling mistakes. These types of misidentification errors can lead to considerable loss of accuracy specifically in diseases with low prevalence (Edwards et al. 2005; Zheng and Tian, 2005).

Misidentification is an error that can be challenging to detect since it is a fundamental feature for many Genome Wide Association Studies that all the samples are unrelated to each other. Thus, the use of already known Mandelian relationships between pairs of individuals included in the study sample is for purpose to infer the tendency to make mistakes which is not exclusively possible to elicit successful outcomes. Instead, misidentification of samples can be detected by following comparisons between a). Sex characteristics associated with X and Y chromosomes data for genome with not established identity of sex b). The inferred data for constructing ancestry details based on principle component analysis with self proclaimed ancestry

Author: *Professor, College of medicine, University of Ha'il, KSA, Biochemistry Department, United Medical College, Jinnah University Pakistan, Current Affiliation College of Medicine University of Ha'il, Kingdom of Saudi Arabia. e-mail: f.khatoon@uoh.edu.sa*

information (Laurie et al. 2010; Anderson et al. 2010). Most of the Genome Wide Association Studies are designed in ways so as to include the single individual from the same population in order to minimize errors and biasness caused by population stratification.

The utilization of Genome Wide Association Studies for detecting many diseases of forensic interest and traits for sex distributions are due to sex associated variations in prevalence (Laurie et al. 2010). The sex and ancestry related data and genetic information are important components for Quantity Check (QC) in GWAS and can be implemented in practice to remove misidentifications of samples and stratifications in population samples. The approach to identify the misidentification in processes where strong associations are detected between genotype and phenotype characteristics can be inferred if the observed phenotype is expected to be caused by the observed genotype for every subject within the study (Hindorff et al. 2009). This observed phenotype if detected to be too extreme in accordance with the genotype then the person is flagged as a possible subject for misidentification (Laurie et al. 2010).

b) *Categorization of Expressive Traits for Genotype and Phenotype in GWAS*

The Genome Wide Association Studies enable the discrimination for genotype and phenotype relationships and also evaluates the techniques for application to real GWAS data sets (Hindorff et al. 2009). The genotype and phenotype associations can be detected in two ways; Firstly, it is always desirable to use previously identified genomic researches with subsequent publications on genotype and phenotype relationships, where the phenotypical characteristics are represented by single or multiple genetic variants and is largely variable in the population under study provided that no one in the sample possessed any trait for Mendelian disease even if a strong affiliation between phenotype and genotype is available, hence it will provide little information for misidentifying samples (Hindorff et al. 2009).

The phenotype traits which are useful tools for anthropometric measurements for forensic purposes include the color of eyes and hair, freckling, hair structure; elicit response to better tastes, urinary and creatinine excretion and metabolic rates (Laurie et al. 2010). Previous research by Hindorff et al. (2009) suggested that the known relationships between genotype and phenotype can be accessed through online data bases such as NHGRI GWAS catalog. In many of these studies, the frequency tables for genotype and phenotypes are being published as a component of descriptive statistics for extracting mutual information when genomic association studies is underway.

On the other hand, the datasets for genomic information can be used to estimate genotype and phenotype associations but presents a demerit that probabilities can be calculated from the same genomic dataset where the researcher is trying to search for misidentifications, hence it presents the challenge in flagging the individuals as sources of potential misidentifications (Hindorff et al. 2009). The discrepancies between sampling criteria and study populations and in matching the associations in identifying the individuals need to be kept under higher considerations while studying genomic association characteristics between populations.

The study of genomic and associated genetic markers requires the step wise following of techniques like firstly, a set of information about genotype and phenotype relationship need to be established and identified (Hindorff et al. 2009). Secondly, phenotypes need to be ascertained at additional cost before the initiation of study, possibly for reasons of extracting usual phenotype data during the collection method. Thirdly, the modeling for the mixture sample that is under consideration for estimating phenotype and genotype relationship needs to be optimized to deliver highest set of information. Finally, the combination of all the information assessed from the relationships should be higher enough to give better sensitivity and specificity. It can also be possible to assess Single Nucleotide Polymorphisms (SNPs) in determining genotype for each sample, before the execution of costly Genome Associated Studies or Genome Sequencing into practice (Hindorff et al. 2009).

c) *Aim and Objectives of the Study*

1. To explore the phenomenon of Epitasis and its relevance with Genome Wide Association Studies Genomic data and privacy factors for public access of population genomic information and access to genomic datasets.
2. To critically evaluate the Genome Wide Association Studies in accordance with population based forensic investigations.
3. To examine the population stratification with respect to Genome Wide Association Studies and its related challenges in illustrating error-free forensic determination.
4. To evaluate the challenge of development of statistical techniques and implementing computational efficacy in evaluating genetic variance and their interactions with genomic expression in Logistic Regression (LR), Multifactorial Dimensionality Reduction (MDR), Random Forests (RF) and Evaporative Cooling (EC) Methods.

d) *Research Questions*

1. What are the factors that are responsible of causing stratifications within population samples destined for

Genome Wide Association Studies for forensic determination?

2. What are the statistical methods followed in GWAS affecting the phenomenon of Epitasis that produces randomization in phenotype resulting in variances in allelic expression?

e) Outline of the Study

The Chapter 1 of this dissertation provides a general idea of this study from the perspective of the researcher will develop from, clearly introducing the question of this study and aim and objectives adopted for the research process.

The chapter 2 of this dissertation provides a brief overview of the topic concerned following a description over the aspects of Genome Wide Association Studies incorporated into the field of forensics and respective literature evidences.

The chapter 3 of the study is the methodology of the research used for describing the study adequately and how it is carried out together with the detailed procedure undertaken in relation to the critical review and the studies included in the study. The ethical considerations recognized essential for this dissertation will also be presented along with the justification demonstrating that this perspective is unbiased, by documenting the literature searching and critical appraisal process.

The chapter 4 of the study presents the analysis and discussion of the data determining the benefits of the studying this problem, assessing and comparing the studies related to our topic along with its discussion, and interventions.

The chapter 5 of the study opens up with the results shortened in the form of a conclusion to the dissertation along with the implications for research and practice.

II. Chapter 2: Literature Review

The section will open up with the review of literature of the published work related to the topic. It will also highlight the theories close to the subject of research along with the general review of the studies selected. The aim of this review is to study the application of Genome Wide Association Studies incorporated into Forensic Medicine.

a) Ancestral Sampling in Genome Wide Association Studies

The alternative approach used to amplify genetic causes of complex traits is population-based Genome studies. These studies focus on the correlation between genetic markers and relevant traits among unrelated population samples (Cardon et al. 2001). The reasons for ancestral mutation of genes that occurred hundreds of years ago and of which lineages of all descendents carry the casual variants, are thus considered susceptible for polygenic mutations. If this casual variant is assayed into Genome Wide studies' perspectives, persons carrying the derived allele will disperse it in different phenotypes as compared to those who carry ancestral alleles (Cardon et al. 2001).

In human demographic history, the markers are in Linkage Disequilibrium with one another within the same population. The genotype segment can be in Linkage Disequilibrium with the casual variants efficiently tracking ancestral chromosomal mutations that have been long aroused but decayed due to recombination. This genotype marker will appear in association with the phenotype and allows the localization of genetic loci (Cardon et al. 2001). In accordance to this, several methods have been proposed for confounding effects of population substructure. Family based association tests (Chen & Abecasis, 2007), used to assess the transmission of alleles from the parents to the offspring and serve as useful tools for assessing ancestral characteristics within sub groups of population. The family-based test analysis is considered immune to potential genetic heterogeneity that lies between generations and among persons from similar families (Chen & Abecasis, 2007).

Global studies (Jakobsson et al. 2008) reveal that genome wide population structure on various levels can be assessed using statistical approaches. Individuals in different continents are differentiated in genetic framework through genome-wide SNP data but however, some level of over-lapping do exists between continental regions (Jakobsson, et al. 2008). This genetic overlapping increases with decreasing distances between continental regions. For example, there is a strong relationship between geographic correlation and genetics similarity of an individual who resides in Europe, but there is a considerable overlap of genetic similarity among neighboring European subpopulations which determines accurate determinations for genetic variability (Jakobsson, et al. 2008).

It is assumed that Genome Wide sets of Autosomal Single Nucleotide Polymorphisms (SNPs) that has become available through Microassays, helps in studying a wide range of geographical genetic variability. Furthermore, Genome wide population substructure is large enough to determine ancestry with large number of Autosomal SNPs at the level of continental resolutions (Mangold E. et al. 2010). Furthermore, the Y-chromosomal informative sets and mt DNA markers needed to be combined along with those from autosomes for precise DNA based inferences of biogeographic ancestry (Mangold E. et al. 2010). Genome Wide Association Studies provide researchers to identify relevant genes involved in creating a specific personality trait or genetic variations. This method scrutinizes the whole genome and investigates the hundreds and thousands of SNPs, simultaneously.

Scientists can use data from this technique and search for genes that are responsible for predisposing a person at risk of developing disease or a trait important in terms of forensic interest (Lueders et al. 2003). Selective genotyping is an application used to identify different ethnic groups or a mix population and can be beneficial in identifying specific traits in a homogenous ancestry distribution. Similarly, by comparing relative allele frequencies between different phenotypic groups, Genome Wide Studies can detect SNPs that are associated specifically to a disease process or an ethnic trait. GWAS are preferred by many researchers because they provide more accurate localization of casual genes and also they provide unbiased detection of whole genome for genetic association (Lueders et al. 2003).

b) Genetic Mapping Scheme in Analyzing Familial Linkages

In order to demonstrate the heritable characteristics of an unidentified person and to detect mutations at the loci implies that Genetic Mapping Scheme is based on linkages in familial studies of randomly illustrated phenotypical characteristics. However, it is difficult to ascertain that as to how polymorphic loci are adhered to establish linkage relationship into practice. To address such problems, to adapt linkages to a polymorphic marker is better than the availability of no linkage (Lueders et al. 2003). For these purposes, tighter the linkage more accurate will be the predicted genotype of the individual and higher will be the degree of polymorphism. The construction of genetic linkage map requires lineages spanning multiple generations and can be selected from large pedigree. The application of restriction fragment fingerprinting that result from each probe, from each person can be recorded and analyzed for familial linkages (Lueders et al. 2003).

These can be made possible by distant translocations; however, can reveal unusual inheritance since the probe may be detecting such sequences that are genetically non-linking. The pedigree analysis can show linkage relationships between modifications of genes by depicting co dominant phenotype. The genes that are dominant or recessive or are structurally modified for a regulatory enzyme can bring about activity and expression. The application of such a system may affect genome detection with different probes already identified through mapping to reflect sequences imbedded in linkage groups. The mentioned characteristics can be distinguished by their dominancy state, rearrangements illustrating co dominance and modifying enzymes that can be dominant or recessive (Marsh et al. 2000).

The analysis for identifying the linkages in familial expressions allows the execution of frequency of polymorphism among 71 proteins, and is studied in European populations (Osbon et al. 2000), and the inferences deduced state that 0.28 base changes per 1000 base pairs suggesting that DNA Sequence Polymorphism depends on the size of the average gene. The visible Electromorphs represents about 1/3 rd of Nucleic acid sequence changes and out these third base changes are not reflected in corresponding amino acids. This depicts that the DNA sequence polymorphism is about 0.001 for each of the base pairs among various protein coding sequences (Osbon et al. 2000). The repetitive DNA segments are exclusively found in humans and serves as a probe for specific loci susceptible to show polymorphism. This mobile sequence can be replicated, cloned and used to screen the genome library of humans. Forensic illustrations in genomic studies can be made possible by making possible the identification of unknown loci and polymorphism that is particular to a family and thus can assist in tracing the desired individual among entire set of population (Osbon et al. 2000).

c) Phenotypical Variance Incorporated into Genomic Studies and SNPs

There are numerous forensic cases in which DNA-based inferences of biogeographic ancestry information is vital to suggest police investigations to find unknown individuals or unidentified victims (Mangold E. et al. 2010). However, it is important to acknowledge that when and how biogeographic ancestry is applied in answering forensic queries and where DNA testing techniques can be introduced to determine the level of ethnic disparity using useful DNA marker sets in association studies. It is evident to keep in mind that the genetic diversity allows the involvements of various appearance traits likely to be error prone as no phenotype is restricted to a certain geographic region. In similar terms, appearances of the unidentified individuals can be known by utilizing markers from genes incorporated into data sets for genomic studies that are functionally active and strongly associated with person's appearances and skin color (Mangold E. et al. 2010).

The genetic effect of Single Nucleotide Polymorphisms (SNPs) on the phenotype of an individual depend on the number of contributing SNPs and non-genetic influences like environmental effects, in determining accurately an identity of a person. Phenotype studies illustrate that eye color is the most successfully predictable phenotype essential in accurately identifying basic physical factors in an individual (Mangold E. et al. 2010). The challenges undertaken in applying DNA prediction into categorizing appearance traits and eye color is its expected variability in conceptual understanding of trait information. For example, people assign same eye color to various color categories and therefore look different from others using an eye color provided by DNA predictions. In order to minimize this problem, studies investigated about the genetic basis of variation in eye color utilizing SNP data analysis (Mangold E. et al. 2010).

Prediction for hair color can also be illustrated by using SNPs and more precisely for red hair (Mangold E. et al. 2010). A recent systemic study investigated 46 SNPs from 13 genes of physical characteristics of various hair colors and discovered models for differentiating between red and blond-red, and blond and dark blond. The lowest prediction accuracy for blond hair can be due to the fact that over the years this color undergoes under extensive changes and thus depicts high variations (Mangold E. et al. 2010). Previous studies in forensic sciences also reported that three SNPs are responsible to express 76 % of total variability in hair melanin (Mangold E. et al. 2010).

Two of the SNPs are believed to be highly specific to hair color while the third relates to reflect biogeographic ancestry rather than emphasizing on the hair color (Mangold E. et al. 2010). Another trait that expresses a physically visible characteristic can be useful in successfully predicting appearance of an individual specially age. Two DNA- based approaches for age predictions are based on mt DNA deletions and age-dependent telomere shrinkage have been suggested for forensic assistance but further research is needed to diversify their effects (Mangold E. et al. 2010). Genome wide studies, on age related features for analyzing DNA methylation pattern may provide extensive benefits for establishing suitable age-predictive biomarkers (Mangold E. et al. 2010). By targeting the Genome data that is relevant and informative can be obtained and is generally applicable. This is because of the fact that the specific information contained within the Genome is generally correlated with all the normal cells in the human body as all cells contain similar DNA sequences, same mutations and polymorphisms.

d) Genome Wide Association Studies Serves as a Tool to Understand the Complexity of Genomic Framework

With the embankment of the technologies in genetic sequencing and phenotyping, the execution of genome wide studies to evaluate relationships between phenotypical and genotypic variants can discover reliable inferences. Genome Wide Association Studies typically utilizes the high tech and thorough input platform to gather large quantity of valid data from millions of genetic variations (Manolio, 2008). The incorporation of Single Nucleotide Polymorphisms (SNPs) in genomic studies with larger DNA or RNA sequencing and variants from unknown individuals assist the scientists to interpret the genetic variability and polymorphism (Manolio, 2008). Genome Wide Association Studies have been quite helpful in investigating the infliction of diseases, various types of lymphomas, carcinomas and tumors including chronic pathologies like Diabetes Mellitus and Insipidus (Gudmundsson, 2007).

The advancements of Genome Wide Studies in establishing etiologic, pathodiagnostic and physiochemical basis for the diseases has been diverse and its current use in forensic medicine is under study. Genome studies is not associated with objectives that are without any imperative challenges of which some are fatal in consequences as they are unable to replicate the findings of the initial studies and also are limited in availability of sample population (Chanock, 2007). This failure in replication has been referred to various reasons identified as higher potency of false positive results and lower power of true positive inferences which, causes poor replication in a follow up sampling (Chanock, 2007).

According to Tian (2008), the mixed population sampling without stratification yields the results for common ancestry but not the causation of the mutations. The problem of insufficient matching of replication samples with the initial population causes the depiction of risk variations that have not been found in the initial sampling. However, the false positive genetic variations in Genomic studies lead to significant disproportionate patterns, variability in allelic expressions and phenotypical differentiation between initial and replicated samples. The application of genomics into forensic investigations assists in finding genes unique to certain genotypic framework which was previously limited due to limitations in technology in genetic studies. Today, the Genome wide association studies help in useful data extraction from genetic information in predicting technical details like the heritability data found in complex traits (Monolio, 2009).

The complexities in the genetic architecture demand application of complex methods in analyzing data for explaining all causes of variations in complex ancestry (Monolio, 2009). The current advancements in GWAS are now increasingly formulating a tool that helps to elevate interest in understanding other types of structural variants that have not been considered for genome studies (Frazer, 2009). However, there has also been other reasons for which genomic studies has not yielded successes many thought they would; still a lot of effort is being put forward in finding better analytical methods in addressing problems associated with GWAS (Monolio, 2009).

e) Genetic Ancestry and Association Study Analysis

The aim and objectives for Genetic Ancestry approach was to adjust funding and utilized resources on designing a panel of Single Nucleotide Polymorphisms (SNPs) that magnifies the chances of observing common variations in a population sample (Need et al. 2009). The genomic studies has also been carried within the European population for achieving better inferences in observing a common group of individuals for genomic variance in order to reduce the expenses and efforts for collecting samples and genotyping analysis (Rosenberg et al., 2010). Another reason for European population based studies was the availability of large and homogeneous sample for populations, such as in cohorts for Finland and Sardina derived population samples (Rosenberg et al., 2010). These cohorts are believed to contain extensive collaborations and the prior genetic and phenotypical information to help in analysis and interpretation of Genome Wide Association Studies (Rosenberg et al., 2010).

The genetic framework for the European derived genomes further helped in designing the first episode for GWA studies (Jakkula et. al. 2008). The comparative lower levels of sub structures within the population within European continent allowed the scientists to properly remodel the HapMap project for genomic determination (Conrad et. al. 2006). The implications for Single Nucleotide Polymorphisms (SNPs) into genomic association studies were initially designed only to observe the common variances among different populations within European territory using tag SNP approach (Gunderson et al., 2006). The relative higher level of similarity in genetic framework in European populations enable the tag SNPs approaches to be transferred to other regions within European geographical limits (Conrad et. al. 2006). The level of Linkage Disequilibrium (LD) has been extended over patterns in European populations, thus it would have been convenient to cover the genome by using fewer tag for Single Nucleotide Polymorphisms (SNPs) (Conrad et. al. 2006).

The implementation of filtering in Multifactor Dimensionality Reduction technique is still at a preliminary stage of evaluation, there is still more improvements need to be implied in MDR's utility to act as a filter for genome Wide Association Studies. Future aims to solve problems that have not been addressed in this study for reason attributable to the limitations of the research grounds and resources, for example lack of induction of significant testing of models before applying their unique characteristics into genomic studies. An analysis of more diversified approach is needed to better determine errors in classification for thresholds for more acceptable and precise effect sizes that can easily be tested in actual datasets.

f) *Structured Association and Genetic markers in GWAS*

The technique of Structured Association works on the principle of Hardy and Weinberg's Equilibrium within populations with linkages to genetic markers that are also responsible for establishing equilibrium to detect and correct stratification in a homogenous sub group cluster based on genotypic characteristics. Test for detecting familial or hereditary pathological traits in the generations can be achieved by studying association of each genetic marker to disease expression phenotype (Alexander, 2009).

The current limitations of methods of optimization of power to detect associations at dual stages of analysis in which the selected models in the first stage can be further tested for main effects (Wang, 2006; Kwak, 2009). It can also be beneficial to incorporate biological information as a component of genomic strategy as it is predicted in current genomic association studies that the results from these analysis can be proved to be true for biological experimentations. As study by Moskvina, 2011 shows that the study of genetic association between subjects improves the chances of finding true associations among products with unrelated functionality.

Further, the same research revealed that filters like MDR also suffers from limitations from previous methodologies such as increased chances of false positive associations. It is therefore, essential to explore the in depth understanding of how unspecified genetic linkages are responsible for depicting interactions in genomics and phenotype analysis. In the future, the application of filter approach in real datasets will be needed to comprehend with important points to be considered. Firstly, in the simulation studies the correct number of SNPs needs to be identified whether mutations are a result of single locus effect or is due to interactions. It is desired for this study to serve as an initiating point for testing for single locus as well as interactional models.

In the coming years, the computational advancements will take place, high order interactions can be made feasible. Secondly, the implication of *'Post Hoc'* analysis needs to be involved to study and understand the actual models. Post Hoc analysis can be applied to investigate the dual loci models through the involvement of MDR to understand the interactive effects. A higher rank of a locus does not necessarily signifies an interactive effect and the development of state of art techniques will help in the future to assist in detecting the underlying abnormalities of complicated genetic models as it can serve as an active research area in molecular genetics (Hu, 2010; Kooperberg, 2008).

The software named 'Structure' was formerly developed which works by considering structured approach to detect associations. Since, this structured process provides better correctional opportunities for removing stratifications, but however, the computational intensity for this technique allows the detection of partial cluster membership, hence limiting its usefulness for incorporating it into Genome Wide Association Studies. Another new package for software that determines the stratifications is 'Admixture' which is also based on clustering cryptic relationship in family structures (Alexander, 2009).

g) *The Use of Principal Component Analytical Method in Genome Association Studies*

The method of Principal Components Analysis (PCA) has been in use to detect and control stratification in the population data and also been utilized in the form of multidimensional approaches to identify a consistent axis of deviation of data for genetic expression. In unknown, forensic samples, the genotype and phenotype are later adjusted by variables attributed to pedigree of ancestry along each of the base axis, forming sets for matched cases and their control samples (Price, 2006). The application of genomic association studies is then made possible by calculating the adjusted phenotypes and genotypes.

According to Price, (2006) 'Eigenstrat' is an another software package has been implemented in genomic studies to execute this method and is shown to be higher in power at detecting casual genetic markers, better at handling highly differing Single Nucleotide polymorphisms (SNPs) than is evident in genomic control (Price, 2006). This software induction also has enabled the effectiveness in tracing computational datasets as compare to Structure technique in handling large sample populations (Price, 2006).

In addition to the above mentioned advantages, Eigenstrat has been able to draw attention towards the impact of differential biasness which may illustrate quality control possible. It needs also be noted that the top two phenotype and genotype adjustment factors may not necessarily due to structure for population but also depends on familial relations and linkage disequilibrium (Tian, 2008). The application of Principal Components Analysis can detect familial structure effectively but on the other hand it is unable to identify correctional models that are individuals with defective phenotypes.

h) *The Application of Mix Methodology Models in GWA Studies*

These Mixed Methods can be implied within Genomic association studies and can model various stratification within the population, that is it studies the effects of structure of a sample population, structure of familial predisposition and cryptic relations between case and control sample size (Visscher, 2008). The primary principles on which these methods are based are the model phenotypes that uses the combination of fixed effects like candidate Single Nucleotide Polymorphisms, sex, biological age of the subject; and also requires input of data from random effects for example, phenotypical variants in matrices (Kang, 2010). When considering the implementation of these techniques into Genome Wide Association Studies, there are still queries as to whether structure for the population sampling should be modeled as a fixed or random entity (Zhang, 2010).

While every of these mixed methods can sufficiently alter and detect stratifications in the genome, it has been assumpted that stratification modeling as taken a random effect can leads to successful discovery of spurious associations specifically for unique SNPs (Price, 2010), as these all are newly introduced methods, more assessment studies need to be done on the best approach for modeling of stratification. When stratification modeling is termed as fixed effect it requires additional induction of previous Principal Components Analysis as a reservoir for accounting stratification of populations to investigate genetic ancestry of each person in the sample (Zhang, 2010).

III. Chapter 3: Methodology

The main objective of the literature review is to draw conclusions for Genomics and its relevance in applying techniques in forensic investigations. This chapter will highlight the definition of literature review and its significance as a method for this study. Then, it will outline the way to carry out this study including the searching strategy process, limitations, and the qualitative design employed in this study. Finally, it will conclude with ethical issues along with the inclusion and exclusion criteria.

a) *Literature Review*

The search of literature is defined as a systemic approach for retrieving, identifying and bibliographic material to manage the studies which are self governing. The main purpose of the study was to locate the information or knowledge on any particular topic for recognizing the study parts for future and forming the clinical practice guidelines. For establishing any search of literature, it is important to comprehend any research and its role in informing the clinical practice as well as questioning (Parahoo 2006). This systemic process was defined by the researcher as utilizing the definite methods for cracking the problem and answering to the research question (Parahoo 2006). Parahoo suggest and argues that the findings of research were not the problem solutions however, the study was more likely offering the current information which assist in the decision making (Parahoo, 2006). Therefore, the eventual the purpose of the study is to expand, develop and refine the knowledge of the body (Parahoo 2006).

b) *Qualitative Design*

The main argument for this particular research study is to carry out a review to examine the molecular genetics and the application of genome and DNA-amplification into forensic investigations. To conduct a literature review a qualitative research design is the most suitable research design. The qualitative research design is the most appropriate research design to examine the human behavior and it assists to legalize and authenticate the data which is selectively collected from the secondary sources. It helps in refining the research and additionally it adds a primary hand worth to it. For many of the researchers the secondary data is crucial in the health care and medicine field as it suggest the previous researches. Creswell (2009), stated that a qualitative research comprise of using exclusive steps of strategies and analysis of inquiry with researchers interpreting what they hear, see, and understand. The method used in this study is qualitative. As compared to quantitative research, Qualitative research is more subjective, and it is based on unlikely methods of gathering information. This research is more or less based on the literature review and the conclusions are drawn on the basis of actual resources (mainly primary researches).

c) *Critical Analysis*

The critical analysis is central process of any research study. It involves critical thinking which applies logical and rational thinking while deconstructing the text. It is a complex, intellectual activity involving analysis and critical comment on the material formerly gathered. Browne and Keeley (2001) defined the critical thinking as a set of awareness which interrelates with the research question under study which is expected to be critically analyzed. It is an ability to answer and ask the critical question at a specific time and it desires to use the critical questions actively). Consequently, it is not just a descriptive list of the set of accessible summary and material. Without a critical appraisal and systemic literature any academic study cannot be applied to a methodology or in any way put in to the knowledge (Hart, 2005). Critical appraisal of the literature also offers a room for the comparison of the results of diverse researches.

d) *Search Strategy*

For research evaluation on the molecular DNA and forensics, the inclusive review of most of the studies available was important. The databases used and accessed for this purpose included: ProQuest, PubMed, Cochrane Library, Science Direct and CINAHL. For the data bases searches following keywords were utilized including Molecular genetics, Genome-wide studies, Polymerase Chain Reaction, DNA Phenotyping, DNA-amplification and sources for DNA-extraction. The bibliographic references of studies selected for the literature review were also searched from these databases. All studies meeting up with the explicit inclusion & exclusion criteria for each literature review section were retained at the end of the search.

e) *Inclusion criteria*

This search strategy was based on evidence available within a range of past ten years and was basically primary researches. The search of literature only found out the material published in English and for this reason it created no issues to reject any possible study for the reason of the translating complexities of the papers. This search of literature found researches including the literature reviews, scientific documents as well as the prospective studies fitting the remit of the search strategy.

f) *Exclusion Criteria*

Exclusion criteria included the articles not published as full manuscripts or not in the peer review literature. The studies including controversial ethical or legislative materials are not included. The studies which were conference papers or studies in progress, or government reports will be excluded. The searches will be restricted to retrieve the literature available in English.

g) *Limitation*

In this review there was a likelihood of publication bias. By rejecting those studies with negative outcomes and were unpublished it was probable to overrate the effects of techniques used, nevertheless, for the published literature the comprehensive search for potentially appropriate articles was undertaken by means of a strategy for systemic review for the purpose of avoiding bias. It was preceded by the contacts with the first and corresponding authors. As the nature of the study was provided (for instance it was difficult to fund, it was complex and expensive), as these diminutive negative trials were not likely and they were expected to manipulate the consequences.

h) *Ethical Considerations*

Most ethical issues in numerous research studies falls within the following criteria: protection from informed consent, honesty, harm, and right to privacy. The most thoughtful and prominent moral trouble that most authors cannot avoid is the utilization of knowledge that was produced by some other author while conducting a same inquiry. No falsification will be occurring while citing the literature. This study lies on primarily on the original and primary sources as it is known that the unethical, fraudulent and dishonest researchers can circumvent the scientific method. It is important to properly cite the person whose work is being used. This is because the person who has conducted an actual research may have served a lot of effort and time to extract the outcomes of the study and hence it is one's responsibility to give credit legally to the individual who has conducted that particular work.

IV. Chapter 4: Analysis and Discussion

This section will include the thematic analysis and critique of the selected articles. This chapter will also discuss the selected researches in detail and their findings, based on the findings of the reviewed literature.

a) Critical Analysis

In 1999, Pritchard and Rosenberg proposed a technique which was based on the concept of differences in the genetic framework between cases and their controls that could be tested. The variations in the genetic details of cases and controls can be measured by utilizing markers which are unrelated with the mutations detected (Weir, 2006). This method has been successful in evaluating the specific structural characteristics in the population; however, it is unable to correct the detected abnormalities in the provided data information about genetic mutations. In the same year, Devlin and Roeder presented Genomic Control method which was also based on the concept of biological markers for genetic detections (Alexander, 2009). Genomic Control also detects significant stratification but effectively compensates for it to be corrected. Their technique was based on the assumption of population structure with respect to case and control data that can be alter by implying multiplicative factorization in proportion to structure of population in unidentified genetic markers.

The inflation factor can later be incorporated into analysis for detecting associations and for correcting stratifications. One benefit for this method is that it can easily be implemented for analyzing the various categories for DNA data information and sequestering. Further, it also allows the implication of more accuracy in results as the number of genetic markers used to calculate the inflation factor, which makes it an efficacious technique to be followed in Genome Wide Association Studies (Weir, 2006). But on the downhill, it cannot on consistent basis detect and correct for relations with the family structure when they are responsible for causing stratification and can lead to corresponding loss of power when markers are strictly differing across the ancestry of subsequent population framework (Weir, 2006).

The serious issue in Genomic Association studies is studying the influence of genomic ancestry on the design for Genome Wide Association Studies. The confounding effects of genetic ancestry in association studies play a significant effect in understanding the population based characteristics that differentiates one subset from the other regarding genotype and phenotype information (Need et al. 2009). Most of the Genome Wide Association Studies are been conducted on European derived populations (Need et al. 2009). In the initiation stages of GWAS, extensive researches were under taken on Europeans mostly due to logistical and research design reasons. In logistical terms, the financial burden for the conduction of GWAS lead to many early studies to be conducted on same general populations with extensive variance in physical characteristics (Need et al. 2009).

b) Relevant Themes

Some of the common themes extracted from the studies above are described as follows:

Theme No. 1: Epitasis and its Relevance with Genome Wide Association Studies

The first definition of epitasis was provided by William Bateson in 1909, and was described in terms of deviation from Mendelian inheritance. Currently, epitasis is commonly defined as interaction among genes at various loci where the impact of allele at one locus is inhibited by an allele on the other locus (Moore, 2005). The effects of epitasis can be favorable in producing no random phenotype and can be harmful in various ways. In the current era, epitasis is now becoming widely accepted as a tool in determining the relations between phenotype and genotype (Moore, 2005). Biologists and forensic experts study gene to gene interactions by incorporating epitasis as physical and biochemical reactions which occur between genes and regulatory networks. Biological component of epitasis can be expressed in several ways; by interactions between factors responsible for transcriptions and by enzymes involved in metabolic pathways (Moore, 2005). For example, blood borne or metabolic abnormalities like Maple Syrup Urine Disease in forensic sampling can aid to separate the individual profile from the majority of the population, thus identity of the unknown can be made possible. Similarly, in Sickle Cell Disease there is a biomolecular interactions between globulin proteins, hence helps in differentiating sick from healthy individuals (Moore, 2005).

c) Epidemiology and Significance of Epitasis

Epitasis has been increasingly significant factor in understanding the genotype for complex abnormality (Shao, 2008). These inferences depict that epitasis can be more prevalent not only in failed replications for Genome Wide Studies but also have main effects in detecting heritability in samples with unique genetic characteristics (Greene, 2009). It has also been shown that simple Mendelian traits like identification of metabolic disorders in body secretions samples, Cystic Fibrosis in blood samples etc. can be made easy to detect when interactions between other genes and the mutant variations that serve to produce phenotypes for a particular genetic expression are assessed.

d) Challenges in Identifying Epitasis

The correct identification of a particular genotype in varying phenotypical expressions can be convoluted that can be one of the reasons epitasis can be challenging in giving success. This is mostly applicable on human subjects which limit the benefits of methods such as genetic manipulation on organisms that serves as models for humans, hence making the execution of techniques even more difficult. In order to overcome this problem, characterization of epitasis on genome wide scale allows numerous biological, computational as well as statistical challenges not only in analyzing genomic data but also to prove functional validity of such inferences. This method presents even more challenges when human populations are assessed as researchers can not benefit from all tools that could be tested on model organisms (Lewontin, 2006).

The conventional methods for applying association theory comprised of using linear models, while these models work better in finding main effects, but use to fall short in finding interactive effects (Lewontin, 2006). The detailed and diverse data framework that needs to be analysed also presents analytical challenge and demand the development of better statistical techniques to cater non-handle components for these interactions which are influenced on by environmental factors. This technique requires the filtration and priotization of biological relevance along with maintaining computational efficacy in evaluating genetic variance and their interactions with genetic environment. Several methods have been developed in order to study the epitasis with regards to Genome association studies, including logistic regression methods, penalized methods and data mining approaches such as Random Forests, Recursive Elimination of Feature-F, Grammatical Evolution Neural Networks, Multifactor Dimensionality Reduction and Combinational Partitioning (Motsinger-Reif, 2008; Wang, 2011, Moore, 2007).

e) Logistic Regression Method in Determining Epitasis

Regression based techniques have been deployed for analyzing epitasis interactions in genomic association studies. They use the statistical approaches to determine the relationships between linear values for predictor variables for example, genetic variants and the possible phenotypical expression. Logistic regression is also been in use when the outcome for the population is binary in nature that is either case or control. This method can also be applied for variable selections for example, in step wise regression, where in variables which elicit main effects are selected first followed by interactive factors that lie between them. The utilization of regression method, however, eliminates the possibility of finding interactions between variations with no profound effects. Logistic regression also has limitations when data are greatly multidimensional as is common when information is analyzed for association studies thus, can lead to higher rates for false positive results. The availability of software packages that help in performing statistical correlation for logistic regression on genomic population samples, one of such tools is named as Genome Analysis Tool (PLINK), which provides the option of logistic regression in association analysis (Purcell, 2007).

f) Multifactor Dimensionality Reduction (MDR)

It is a non-parametric estimation of model free genetic model to detect interactions in the absence of significant effects and has proved itself as an impressive technique to identify statistical differentiations in assessing interactions in GWAS (He, 2009). This technique comprised of incorporating the data into divisions according to total number of samples like 'N' across validity intervals for testing in order to avoid model over fitting such that the final result can predict all of the total data information. In the subsequent step of the method, the list of loci combinations is placed in tabular k-dimensions that contain all possible n-loci combinations (Velez, 2007). The ratio with the number of cases to control sampling for each genetic susceptibility is then calculated and compare with the existing baseline measurements. The entities below the threshold are termed as low risk and above are labeled as high risk. This process efficiently reduces the highly diverse information into two categories high and low risk factors.

The balance between unequal number of cases to control sampling and clarification accuracy needs to be calculated. This accuracy prediction gives a value for mutation risk in the testing sample. This method is repeated for all possible loci combinations and for every cross validation interval. There have been several amendments to Multifactor Dimensional Method including Log linear Based, Odd ratio Based, and Generalized Linear Model (Lou, 2007; Chung, 2007; Lee, 2007). All MDR sub-extensions have the ability to reduce highly diverse data input into high and low risk without illustrating any assumptions about the genetic framework. It serves as a powerful tool in identifying genotypic errors and data that has been gone missing (Richie, 2003). Conversely, MDR is limited in execution of heterogeneity as it losses its ability to detect best possible differences.

g) Random Forests and Evaporative Cooling Methods in GWAS

Random Forests and Evaporative Cooling Techniques are being used as learning machines that utilizes combinations of recursions, iterative logarithms to select subjects with maximum potentials for interactions. According to Breiman (2001), Random Forests Method allows the construction of classification tree to detect specific patterns in the data by using a bootstrap sample with variants at the tree nodes having being taken as randomly from the initial information. This tree is built in a fashion of node by node formation of genetic variations from the entire data and is sampled with replacement from the total sample size. The individuals are classified on the basis of size of forests of trees that have been included. Errors in predictability are later calculated from the persons that were not initially selected for the building of the tree (Moore, 2007).

An approach that uses the calculation for finding the distance that lies between subjects based on the variability of similarity in genetic information despite the differences in phenotypes by applying Genome wide association studies. On the other hand, Evaporation Cooling technique is featured with the selection algorithm and combines them with thermodynamic principles incorporated in cooling mechanism of atoms. The main aim of this technique is to increase the density by evaporation of variants within the featured space, which contains the elements with highest potential of interaction (Breiman, 2001). Due to the fact that large quantity of variables are involved in genome wide studies and thus large amount of interactions need to be probed during the investigation of epitasis in order to find statistically essential interactions in filtering the inferences.

The concept of non-parametric and genetic data free approach, to detect gene to gene and gene to environment interactions with additional benefits of detecting higher interactions makes Multifactor approach successful for analysis of data. Further, this dissertation will analyze Genome Wide Association Studies' data and utilize it as a tool for detecting and exploring in depth explanations for problems associated with GWAS in accordance with epitasis (Ritchie, 2005). The minimum number of epitasis associations that can be investigated in Genomic Association Studies is tested for all possible paired interactions between genes and variants. For such types of studies, statistically essential information can be found that is not specific to simple interactions, then the implications for additional methods which can filter the inferences and select what is the most useful and helpful results.

MDR algorithm can also accomplish the testing of all interactive factors in a database and not only it selects single best outcome based on prediction accuracy but its unique framework also allows for retaining all the models for future evaluation since it goes for extensive search. MDR also follows the filtering technique that involves population stratification, a confounding source for Genome association studies which identifies this problem and to stratify the data into proper number of sub populations before analyzing it with Multifactor Dimensional Reduction technique (Breiman, 2001). The sub domains comprised of racial or ethnic lineages through which population clustering takes place. This technique resulted in separately analyzing each population and reduces the important sample size for the entire study.

h) Theme No. 2: Population Stratification with Respect to Genome Wide Association Studies

Population stratification can be defined as the differences in allele frequencies between controls and cases that are related to commonalities in genetic ancestry but are unrelated to mutations in a heterogeneous population and serves as a confounder for Genome Wide Association Studies. These variations in allelic expressions between different populations like for example, geographical migratory patterns, genetic mutations, mating behaviors, drift and selections for sub-populations in the sample can be responsible for creating differences in the frequencies for the alleles over a period of time (Freedman, 2004).

Stratification acts like a tool for confounding subgroups for various populations within the sample that have different occurrences for a particular trait. These differences in the prevalence for the unique characteristics in a sub group of a population for most effected individuals can be studied coupled with a genetic variant of various allele frequencies that are over represented in association studies. When samples with diverse ethnic data and variations are analyzed for stratification, this can lead to depiction of false positive association. This population stratification can also yield results for samples that possess distant relations but similar ethnic backgrounds with no familial affiliations (Freedman, 2004). However, population stratification in some cases is the reason for failed attempts for replications of previous association studies as higher level of false positive inferences are found in these studies.

It is due to the fact that some of these association studies contains subjects from diverse heterogeneity in terms of genetic frameworks while the others contain information that are only true for a specific sub group of a entire population sample (Freedman, 2004). Therefore, population stratification may only minimizes the confounding effects of diversity in genetic determination in which casual genes differ in populations, along with elevated power for searching associations which differ among various population subgroups (Freedman, 2004). There have several been association studies which signify on characterizing the extend of stratification problems and also the impact that it can create on final association analysis (Helgason, 2005, Freedman, 2004).

Conversely, the genetic attribution can be the presence of occurrences within the population with genuine ancestry that contains differences in genetic encoding (Freedman, 2004). The problem of stratification is also enhanced with the increase in the sample size of the population (Marchini, 2004). The prevalence of Single Nucleotide Polymorphism (SNP) in a population, for example in European Americans, Afro-Americans or Asians is quantified and its impacts are studied in association tests. These results illustrate that the constant incidence of certain abnormalities or pathologies across populations can be made possible through stratifications analysis. The problem of not achieving proper stratification of population samples is minimized when homogenous samples with small ratios for intermingled data are observed (Helgason, 2005). According to Helgason (2005) the mild concentrations of stratification can be a significant confounding factor in genomic association studies if not adequately measured and subsequently corrected.

One of the methods to remove errors from final results of Genome Wide Association Studies is the nature of study design adopted; the utilization of adequately synchronized control and case samples will eliminate variations in genetic backgrounds between the two groups under addressed. This prevention of stratification in Genome Wide Association Studies, works by assuming SNPs that are not directly part of the candidate genes there by in saves association signals that can be removed through the effect of stratification process. However, this matching criterion is not always successful in eliminating all forms of stratifications that are part of the initial data of the Genomic Association as several techniques have been developed to identify and correct the discrepancies in stratification of population; some of them are mentioned below.

i) Transmission Disequilibrium Test (TDT) for Mendelian Law in GWA Studies

The induction of designs for family relatedness like when relatives of the cases act as controls, have also been proposed as a means of minimizing the effects of stratification of population and this test is one of those effective methods to be employed. The parents of the affected individuals act as controls and this principle is based on Mendelian law that each of the parents transfers a single allele to its child. The knowledge of which of the alleles are transferred is afterwards used to form a placebo case control sample where the cases are defective alleles transferred from the parents to the offspring's; and untransmitted alleles are termed as the controls. This technique of case to control matching minimizes the impacts of stratifications by deleting bias from differences in allele frequency.

Whereas this method is able to eliminate the problem of population stratification, since the familial information is reduced in biasness the cost for this collection of data process is much higher as compared to other techniques and the design itself is not applicable specifically in cases when the mutation under study are late in expression. This method has been found to be less efficient as it needs a genotype of three persons to obtain to form a single case control match. This method also requires the parents of the subject to be heterozygous in order to know the information of allele that is passed from them to their offspring. The irrelevant data information that is not useful for the process of analysis that is some unconcerned genotypic information about distant relatives, can also result in loss of power for the reliability of the method.

j) Theme No. 3: Genomic Association Data and its Discloser for Public Access, Barrier to Human Privacy and Protection

Church et al. (2009) demonstrated that the given out of information pertaining to Genome Wide Genotypic Data from a person for the purpose of utilizing it for forensic sampling in Genome Wide Association Studies. This research further evaluates the actions taken by the authorities in USA to inhibit public access to Genome Wide Data dispersion but however, limitations have been observed to restrict sensitive information and also the use of genotype frequency data from each of the previous study which have been published using funding from public services agencies like National Institute of Health (NIH) and Welcome Trust. Reactions to this implication in the public sector has been not so encouraging when authorities consider it too late and too little that has been done in maintaining privacy for individuals. However, the response from responsible persons in genetic field of experimentation models declared the breaching of public data as a high level beaurocratic response to a minimum risk which can unnecessarily inhibit the futuristic scientific research.

Other scientific concerns were also been highlighted concerning to situations when an individual's identity is accurately determined by applying Genome Wide Association Studies datasets. Several other researchers have also signified that misuse of an identification of an individual along with familial and medical history records and if genetics will not held these sensitive data information it will backfire to their own capabilities to carry further sequence of research network. Church et al. (2010) notified that appropriateness of measures in restricting access to information is likely to exclude the scientists who are specialist in handling sophisticated data. The active researchers in the field of genomic wide studies further argue that consent for complete disclosure for releasing genomic information should be taken from the respondents rather than making this procedure difficult to access for promises for anonymity.

In similar context, Martin and Bobrow compared and contrasted the related potential risks and advantages in execution, hence proposed four stepped representation; a). For current scenarios limited access to personal genetic information need to be implemented; b). Declaration of any activity that comes under the category of illegalarity need to be banned or be termed as malicious practice; c). Increase in the level of knowledge for the general public to identify the nature of their personal information that can be misused for purposes harmful for them in physical, social or financial frameworks; d). Encouragement for appropriate recognition need to be stimulated in order to understand the professional relationship between the genetic researchers and the study participants.

k) Population Genomics and Access to Genomic Datasets

Homer, Szelinger, Redman, Duggan, Tembe, et al. (2008) illustrated that Single Nucleotide Polymorphisms data from respondents inducted in Genome Association Studies can depict that individuals' genetic mixture consists of up to 1000 respondents whose DNA profiling have been disclosed for research proposes. As the mentioned case implies the basic statistical theory, the condition illustrated that this needs to elevate the frequency of subsequent discussions that can minimize the privacy concerns and trigger anticipating scientific possibilities (Human Genome Project Information, 2003).

Homer further entails that scientists who are continually using the sensitive data for research purposes, need to adhere to a code of conduct which is internationally known and accepted to provide the proof of research credibility as a bona fide researcher (Human Genome Project Information, 2003). A researcher then be awarded a permit and can be placed on a registry of users who are have refined access to genomic information in data bases (Braun et al. 2009). This can avoid the need to repeatedly applying to proof bona fide status to various regulatory bodies. Any misuse or misconduct by the researcher will be termed as denied for access to genomic data bases according to the ethical code of conduct. IN other terms, interim models should also be adopted which are consistent with the principles of permit mechanism (Braun et al. 2009).

The determinants of a bona fide researcher may be internet oriented but also require previous affiliation to an authentic institution (Braun et al. 2009). The permit also needs to link to a known and registered individual by using a universally accepted 'Researcher ID System' such as being promoted by GEN2-PHEN project (Human Genome Project Information, 2003). The individuals who are responsible for handling excessive secondary information need to be expected to follow the same standards and practices as primary data providers when disclosing information. To implement this work, global strategies like that of Burmuda Principles (Human Genome Project Information, 2003) is required to be followed. In order to get the researchers and their administrators sign a legal document, and if they do so it does not guarantee compliance. For example, during a forensic investigations access to high security data need psychosocial assessments of relatives, and personality evaluation that is highly personal and sensitive information. However, sometimes regulatory personnel possess some classified data outside the security zone.

Human genomics are observed to be extremely useful and is also considered to be capable of analyzing numerous numbers of angles. Another perspective has been illustrated in a Consortium (2003) over Celera in the map of human genome project (Wellcome Trust Sanger Centre Press Releases, 2003). The individuals who have decided to display their genomic data and personal biological information to be freely accessible on the web have played a great part in introducing sharing of information with mutual consents. This approach of data sharing is termed as a norm in genomics research (Wellcome Trust Sanger Centre Press Releases, 2003) and is also a prerequisite for suitable funding for the project. However, genomic research is not limited to the research community.

There are now private institutions that are liable to collect sequence data and sensitive information (Jupp 2001). The real challenge for genomic studies is not the availability of the information within the realm of research community, the availability of the information about genomic sequence to people outside the research community, as these are the concerns safeguarding the rights, over sighting the confidentiality and evaluating the professional codes for conduct (Huang et al. 2009). There has been many factors identified that serve as risks for privacy for studying genomic sequence data but also breech to information can also takes place through availability of information on other resources (Huang et al. 2009).

The possibility of the datasets to get combined with relevant sources increases the risk for breeching of sensitive information (Greenbaum et al. 2008). For example, in a study by Gitschier it has been detailed that iterary compare and contrast of subsequent surnames administered in genealogical data bases and information regarding Y chromosomes from HapMap project, genetic information about the concerned individuals can be detected by increased accuracy (Jupp 2001). Furthermore, Nyholt and colleagues depicted the difficulty of securing the sensitive genomic information by using Linkage Disequilibrium and data about polymorphisms to access information that has not been previously released (Nyholt et al. 2008). There information can be used by legal and health insurances in order to take decisions about the individuals (Jupp 2001).

The accessibility to web based approaches and face book utilization to extract birth records of individuals while these evolving data processing technologies allow effective collection of data records (Greenbaum et al. 2008). I order to raise these issues Greenbaum et al. also recently suggested that anonymity has already be a thing of the past due to the presence of infer data and the quantity, nature and variety of data that has been freely available (Greenbaum et al. 2008). The ethical consideration for the collected data has led to balance the privacy of the respondents using the principles of anonymity and informed consent. It is termed as unrealistic to promise respondents of complete confidentiality when genomic data is analyzed and assessed (Taylor 2008). The evident increment in the number of data sources that are available in commercial and public sectors together with the ease in evaluating the genetic information suggests the content of data being released, has altered over couple of years (Taylor 2008).

If the researchers are not inclined to take better confidentiality measures towards the support of respondents and also to the public indulged into genomic research then it can have damaging consequences for futuristic scientific researches. In genuine terms ethical questions that arise from Genome Wide Association Studies or cohort design of research are actually derived mainly from funders or insurances to encourage far and wide sharing of basic research data and from the power of using internet tools for sharing of data (Greenbaum et al. 2008).

One contrasting likelihood, is the availability of liable genomic data to many intelligent minds which maximizes the chances of dispensing results of the researches to societal and community based groups but on the same time increases the breeching opportunities of privacy for research respondents. One such solution entails to the fact that to make only aggregated genetic data available publicly accessible on internet but with restrictions to accessing individuals' genotypic data (Greenbaum et al. 2008).

The presence of a person's DNA can be detected even if it is in minor concentrations in a mixture, and it is therefore possible that under very unique circumstances an irrelevant individual can deduce that particular individual is a part of a group of patients who has been suffering from a particular disease state, thus causing the data to be de-identified and the privacy of the person thus be invaded. If the researcher abandon large scale genomic studies and minimize the pressure for data sharing methods, it can be disproportionate response to the level of threat as is observed. The need to get consent from the respondents is a possible solution to the problem but the detailed, obfuscate and technical documents can only make respondents confused and can only be effective for shifting the legal burden that can be a part of the ethical consideration during the research process.

Braun et al. (2009), in his study depicted that is it possible for the individual with a genotype which is known and he belongs to a sample of respondents for whom only allele frequencies are identified? These allele frequencies can be identified through reference sampling and this query can be de phrased as; does the member of this group of individuals in a particular sample belong to a reference or test sample? Homer at al. discussed the wide subject of genomic wide association studies in a forensic context and stated that there has always an interest in acknowledging that a specific person belongs to a particular population and is a contributor to a mixed sample of genetic information like DNA from more than one individual (Braun et al. 2009). The implication of genome wide association studies for determining allele frequencies but not for individual genotypes, are being made publicly accessible and available; also the interest of the individuals in knowing whether a specific person is a part of a genomic study should be restricted from public access (Braun et al. 2009).

V. Chapter 5: Conclusion

Over the span of last 10 years, the fields of genomics have under taken interesting and dramatic shifts in how to comprehend forensic detailing with respect to investigations about unknown pedigree, unidentified complex diseases and ethnogeographic ancestry of subjects (Scott, 2007). The wealth of information which can be extracted from immense map of human genome along with advancements in genotyping and related technologies has presented many opportunities for the scientists to test assumptions about the complex genomics for human. This has led a new era of genetic investigation like Genomic Association studies that incorporates to solve analytical questions about heritage and population based studies (Scott, 2007).

The genomic association approaches have enabled researchers to study in detail the effectiveness of candidate genes analysis which largely depends on previous information about the unique characteristics in a familial pedigree that have been based on inferences from previous assumptions (Scott, 2007). The Genome Wide Association Studies allows an effective scanning of the entire genomic data base at a single time, without the need of assumptions about genetic details for the missing encoding information. The implication of candidate analysis for gene expression has also added profound advantages of uncovering novel forensics.

In this study, the analysis for Genomics yield conclusive evidences for cases when common traits have been detected in genomic samples like detection of familial tendency of pathologies, mutation at fixed loci and phenotypical similarities (Fernando, 2008). For example, some of the phenotypic appearances are specific to an ethnic group, for example the hair color, skin color and morphological features. It has now been suggested that the prevalence of the minor allele frequencies have higher penetrance and may be responsible for familial diseases (Schork, 2009).

In 2005, Genome Wide Association Studies has evoked several other successful studies searching for answers for genetic variants related to complexed hereditary information (Scott, 2007; Hunter, 2007). However, other studies also illustrated that results cannot be replicated in other independent sampling and these scenarios demand deeper investigations in order to cater the replication failures. The popularity and the utilization of Genome Wide Association Studies have increased over the years, along with the challenges like identification of stratifications in a population has been made feasible for forensic investigations.

Future implications for minimizing the failures in replicative studies pertaining to Genome Association Studies can led to development of methods to define and evaluate the genomics and to prevent errors. The robust application of Genomics and associated studies can also have the confounding impacts as factors that can improve the reliability of the inferences extracted. On the other hand, the novelty of the technique has limitations to the amount of genetic variations detailed by the risk variants found to be related to the genetic factors under investigation.

The Genome Wide Association Studies has also suggested the limited approach of this method as it may not serve as a powerful tool to discover the complete genetic predisposition of the mutations unique to a familial trait, or even can predict a disease risk which has not been noted due to insufficient knowledge about genetic framework. This limitation can be attributed to the held assumption that the effects of main factors that act independently but chiefly contribute to the genotype for causation of mutations. However, this notion has now been challenged as more empirical studies have depicted that epistatic impact can occur despite of the presence of any main effect and thus, can essentially effect the phenotype, as has been proved solvable in Genome Wide Association Studies.

As concerning to epitasis in association studies, it sometimes becomes a burden on computational analysis for reasons because detecting and searching for epitasis will benefit the filtering techniques for prioritizing variants (Fernando, 2008). The utilization of next generation for sequencing data for genomic association studies also provides the challenge for implementing stratification techniques to the current datasets. Most of the Genome Wide Association Studies include a lot of rare variants as it signifies that the correction of the possessed stratification in the population samples are more important to be deleted (Fernando, 2008). There is an increased chance for the variants to be exclusive to ethnic, cultural and regional based variations, thus is responsible for causing an elevated number of stratifications.

It has also been a chance for these stratifications to deviate from the initial assumptions possessed by many currently in use methods for stratifications, as sequencing of the data is likely to be done on persons who have unique phenotype for the variants being studied, therefore depicting deviations from expected adaption of rare variants (Price, 2010). There are no available stratification techniques that can accept data for sequencing, hence if this has been possible then the genotyping from Single Nucleotide Polymorphisms associated with the rare variants as well that are scatter haphazardly across genome can be adjusted for the confounding and stratification detection (Fernando, 2008).

Due to this realization about Genome Wide Association studies, 3 million or more variants are detected for interactive effects which presents as a computational challenge for GWAS analysis. These large number of variations examined for interactions also elevated the tendency for finding many essential combinations of risk variants, therefore genomic analysis methods need to be developed to filter and evaluate the most viable interactions. The influence of environmental factors on the variations in phenotype in the genetic analysis has been under much debate, for example the impact of toxins on human genome also accounts many contributions by GWAS (Thomas 2010).

The implications for statistical models for determining casual relationships between genetic variants and their biological interpretations can also limit success of functional validity in GWAS findings (Cordell, 2009).The predictions of genetics involving in complex phenotypical observations have limitations for appreciating single effects and thus wider possibility of gene to gene interaction (epitasis) is ignored. The use of technique of epitasis in genomic determination rests on the identification of variants that effect profoundly on final genetic expressions for alleles, gene to gene interactions and gene to environmental interactions, which are observed to significantly affect the nature of phenotype (Greene, 2009).

The genomic genotype sequencing based approach to discover variants in a sub group of a population require a defined location of a casual variant. This method subsequently allows the scientists and researchers to extract maximum advantage by reducing costs since the technique used for sequencing data require large sample size and thus are quiet expensive. New Genome processing and approaches are implemented into practice to study the interactive phenomenon between genes and environments in order to acknowledge the better understanding for biological mechanisms involve in unique characteristics for a sub population (Fernando, 2008).

Major study datasets have already contain genomic data that is available on the web and internet sources, these subsequent sites can be blocked to only gain access for the researchers on terms that the content of the data bases will be used appropriately and purposefully for research processing. If such obligations are made compulsory and are reinforced into genomic research field there would not be major obstacles for the utilization of the data and could be retained for indefinite time periods (Gill et al. 2009). Secondly, the declaration that the misuse of data to establish the identity of the individual is a legally punishable offense can also aid in maintaining confidentiality of respondents. To limit the deliberate invasion into data sets for breeching purposes is a fundamental principle for protection of data but clear statements of punishments and the desire to reinforce these implications can be extremely helpful.

The induction of samples for population allele frequencies in genome wide association studies require the inductive process to be based on simulated and real data process to show original analysis. The susceptibility to Linkage Disequilibrium along with the markers for genetic determination rests on the identification of allele frequency between testing and referencing samples. With the elevating occurrence of the forensic DNA Profiling, that resulted from the determination of how the template amplifications and increasing number of samples for forensic investigations contain DNA from numerous contributors and the inferences drawn from such samples have progressed significantly (Gill et al. 2009).

As final point, Genome Wide Association Studies associated approaches are been incorporated into validation remains a standard, by which the acceptance of most of the hypothesis and techniques can be appraised. It will also be an important hallmark for the Genome Wide Association Studies to evaluate the same impacts in multiple populations for forensic determination in varying ancestries to extract a better outcome of population specific variants in order to depict effect on human identification in cases of mass disasters or for analyzing ancient samples for familial characteristic determination (Gill et al. 2009).

a) *The Road Ahead in Genomic Association Studies for Research and Practice*

Looking towards the future, it seems evident that the current enhancements in Genome Wide Association Studies is going to transform through changes for finding genetic susceptibilities that are responsible for causing characteristics unique to a familial trait (Bird, 2005). In 2010, Durbin provided a series of projects in the '1000 Genomic Project' that has provided an avenue for testing assumptions by utilizing sequencing techniques to identify rare variations in human genome. The use of sequential data for detecting rare genes for variant determination has also been observed possible through association analysis.

One of the barriers for effective execution of association studies in the samples containing higher amounts of heterogeneity that would be contaminated with variants in determining phenotype for each of the individual in the familial data (McClellan, 2010). Another feature for rare variants is their lower frequency of detections as the stratification techniques are not designed to consider this variation and is difficult to decide which stratification technique would be more beneficial. It is also thought that the implementation of new techniques is the need for current field of molecular genetics in order to identify rare variants for obtaining population based discreet data (Gill et al. 2009).

If individuals are inducted into distinct jurisdictions, then it can restrict the reach to their respective privacy by police, insurers, and state agencies and there would be a little risk of harm (Gill et al. 2009). This harm to a research participant is not the result of their participation into study process but from other activities, for example gain of access for data collection from hospital records and from private agencies keeping data sets for genomic studies, thus allows some genomic information which is related to identifying the basic phenotypical characteristics of the individual to become accessible to the data miner (Gill et al. 2009). Public discussion to alarm people and authorities to take legal actions can be made effective by placing genomic data on places that are secured rather than placing the information on potential places from where they are easily accessible. Further, individuals mostly do have confidential relationships with professionals like lawyers, bankers, doctors on whom they possess trust as they are bound by professional codes of conduct along with subsequent liable penalties (Fernando, 2008).

Researchers and the institutions that hire them to work for these agencies, probably require making their own statements in this regard as there need to be clear codes of conduct and punishments for breaching the records and information within their possession. Future clarifications for re identification of the individuals are mostly relevant in relation to the reference population , thus is hoped that future will entails greater sense of clarity and time to form a proportionate long term response (Gill et al. 2009).

According to Visscher and Hill, the need for further progress in genomic sensitivity for forensic genotyping means allelic drop is common and demand the utilization of sophisticated techniques to be induced in order to calculate the ratios. Secondly, during a forensic investigation, the scientists abandon the use of 13 to 20 microsatellite markers in favor of very large number of Single Nucleotide Polymorphisms (SNPs) as the sample under study contains a large data base for the offender profiling (Gill et al. 2009). In June 2009, there have been over more than million profiles available in US data bases in determining the forensic analyses for finding whether or not all the alleles are observed within the mixture of profiles and then calculating the ratios would be effective (Gill et al. 2009). The technology into genetics has been on the incline but at the same instance there lies a genuine fear of scientific abuse in medical and gene technology and a great potential of great harm to both research processes and health related implementations in identifying issues that have not been addressed sensitively (Gill et al. 2009). This has been termed essential to get the balance of anonymity protection and privacy restrictions, honest and uniform consent rights and implication for encouraging greater participations in the debate and subjective education surrounding the issues (Gill et al. 2009).

Acknowledgment

In the name of Allah the most compassionate and beneficent.

I am most grateful to all mighty Allah for giving me the insight and patience to complete this work.

Firstly, I would like to thank my elder son Mohammad Harmain for providing me with excellent guidance and support throughout the duration of his study. His enthusiasm and drive was a constant source of inspiration and encouragement.

I would also like to express my most sincere gratitude to my colleagues and my husband Zahid Balouch . I would also like to thank Publisher and all team members to not only consider this book for publication but also publish in the shortest possible time, despite COVID-19 pandemic-associated international restrictions, lockdown, work from home, and several other limiting factors. All the reports referred to in this book are given proper citation. However, any missed information is just unintentional and explicable.

I am also thankful to University Of Ha'il for the support during this work

My special thanks to Prof Awdah ALHazimi, Dr Hend AlKawai, Prof Mohammad Khuddus, for his initial encouragement and I would like to thanks to Prof. Dr Bilal Faiz Khan Dean Academic and Professor Opthalmology UMDC, PK for initiation and useful information and vital feedback during this study.

My warmest and deepest thanks to my parent my sibling, my research fellows and students are central to all my research and academic work. My father Late Gulam Mohammad Shaheed is the motivating force behind anything constructive I do. They are truly brilliant and have a bright future. I would like to express my special thanks of gratitude to my mentors, teachers, and students who gave me the strength to accomplish this. Also, I would like to thank my colleagues, family, and friends who provided a lot of encouragement and support during the work on this book.

VII. Preface

The achievements in Human Genome Project and subsequent advancements observed in Genotyping techniques have led an influx of exciting new era of discovering human genome and genetics. These technologies have provided the scientists with a comprehensive data on human genomes as human genome is now capable to incite in depth and precise data information and allows access into processes that extract detailed sequences for genomics in order to analyze theoretical questions but also identifies the practically feasible genetic characteristics. According to GWAS, the inductive spurious associations between a phenotype for the pathology and a variance subsequently assist in evaluating true relationships between genetic variables. Most of the Genome Wide Association Studies are designed in ways so as to include the single individual from the same population in order to minimize errors and biasness caused by population stratification. The phenotype traits which are useful tools for anthropometric measurements for forensic purposes include the color of eyes and hair, freckling, hair structure; elicit response to better tastes, urinary and creatinine excretion and metabolic rates. The discrepancies between sampling criteria and study populations and in matching the associations in identifying the individuals need to be kept under higher considerations while studying genomic association characteristics between populations.

VIII. Methodology

The methods for research for this particular study is to demonstrate a review of relevant literature to examine application of Genome Wide Association Studies in accordance with population based forensic investigations. To conduct a literature review a qualitative research design is the most suitable research design. It provides the rationale for assessing the human behavior and assists to legalize and authenticate the data which is selectively collected from the secondary sources. It is important to properly cite the person whose work is being used. This is because the person who has conducted an actual research may have served a lot of effort and time to extract the outcomes of the study and hence it is one's responsibility to give credit legally to the individual who has conducted that particular work. This search strategy was based on evidence available within a range of past ten years and was basically primary researches.

IX. Analysis

The effects of epitasis can be favorable in producing no random phenotype and can be harmful in various ways. In the current era, epitasis is now becoming widely accepted as a tool in determining the relations between phenotype and genotype. It has also been shown that simple Mendelian traits like identification of metabolic disorders in body secretions samples can be made easy to detect when interactions between other genes and the mutant variations that serve to produce phenotypes for a particular genetic expression are assessed. The correct identification of a particular genotype in varying phenotypical expressions can be convoluted that can be one of the reasons epitasis can be challenging in giving success. Several methods have been developed in order to study the epitasis with regards to Genome association studies, including logistic regression methods, penalized methods and data mining approaches such as Random

Forests, Recursive Elimination of Feature-F, Grammatical Evolution Neural Networks, Multifactor Dimensionality Reduction and Combinational Partitioning. Stratification acts like a tool for confounding subgroups for various populations within the sample that have different occurrences for a particular trait. These differences in the prevalence for the unique characteristics in a sub group of a population for most effected individuals can be studied coupled with a genetic variant of various allele frequencies that are over represented in association studies. The accessibility to web based approaches and facebook utilization to extract birth records of individuals while these evolving data processing technologies allow effective collection of data records.

X. Conclusion

It has been a chance for the stratifications to deviate from the initial assumptions possessed by many currently in use methods for stratifications, as sequencing of the data is likely to be done on persons who have unique phenotype for the variants being studied, therefore depicting deviations from expected adaption of rare variants. The induction of samples for population allele frequencies in genome wide association studies require the inductive process to be based on simulated and real data process to show original analysis. Genome Wide Association Studies associated approaches are been incorporated into validation remains a standard, by which the acceptance of most of the hypothesis and techniques can be appraised.

References Références Referencias

1. Al Safar, H. S., Abidi, F. H., Khazanehdari, K. A., Dadour, I. R., & Tay, G. K. (2011). Evaluation of different sources of DNA for use in genome wide studies and forensic application. *Applied Microbiology & Biotechnology*, *89*(3), 807-815. doi: 10.1007/s00253-010-2926-3
2. Alexander,D. H., Novembre, J., and Lange,K. (2009). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Res. 19 (9)* 1655-1664.
3. Anderson, C., F. Pettersson, G. Clarke, L. Cardon, A. Morris, and K. Zondervan (2010): "Data quality control in genetic case-control association studies," *Nature Protocols*, 5, 1564–1573.
4. Bird,T. D. (2005). Genetic Factors in Alzheimer's Disease. *N. Engl. J. Med. 352 (9)* 862-864.
5. Braun R, Rowe W, Schaefer C, Zhang J. Buetow K (2009) Needles in the haystack: identifying individuals present in pooled genomic data. *PLoS Genet 5(9):* e1000668. doi:10.1371/journal.pgen. 1000668.
6. Breiman,L. (2001). Random Forests. *Machine Learning 45 (1)* 5-32.
7. Browne, M & Keeley, S (2001), *Asking the right questions: a guide to critical thinking,* 6th edn, Prentice-Hall, Upper Saddle River, N.J.
8. Cardon, L.R. and Bell, J.I., 2001 Association study designs for complex diseases. *Nat Rev Genet*; 2(2):p.91-9.
9. Chanock, S. J., Manolio, T. A., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., Hirschhorn, J. N., and et al. (2007). Replicating Genotype-Phenotype Associations. *Nature*, 447 (7145) 655-660.
10. Chen, W. M., & Abecasis, G. R. (2007). Family-based association tests for genomewide association scans. *The American Journal of Human Genetics*, *81*(5), 913-926.
11. Chung, Y., Lee, S. Y., Elston, R. C., and Park, T. (2007). Odds Ratio Based Multifactor Dimensionality Reduction Method for Detecting gene–gene Interactions. *Bioinformatics 23 (1)* 71-76.
12. Church, G., Heeney, C., Hawkins, N., de Vries, J., Boddington, P., Kaye, J., & ... Weir, B. (2009). Public Access to Genome-Wide Data: Five Views on Balancing Research with Privacy and Protection. *Plos Genetics*, *5*(10), 1-4. doi: 10.1371/journal. pgen.1000665
13. Conrad, D.F., etal., 2006. A worldwide survey of Haplotype variation and linkage disequilibrium in the human genome. *Nat Genet, 38(11):*p.1251-60.
14. Cordell,H. J. (2009). Detecting Gene-Gene Interactions that Underlie Human Diseases. *Nat.Rev.Genet. 10 (6)* 392-404.
15. Creswell, J.W. (2009) Research Design Qualitative, Quantitative, and Mixed Methods Approaches Third Edition, *SAGE Publications, Inc,* pp. 296 retrieved from http://www.sagepub.com/books/Book232401.
16. Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton A., Brooks L.D., et al. (2010). A Map of Human Genome Variation from Population-Scale Sequencing. *Nature 467 (7319)* 1061-1073.
17. Edwards, B., C. Haynes, M. Levenstien, S. Finch, and D. Gordon (2005): "Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies," *BMC Genetics*, 6, 18.
18. Ekstrøm, C. T., & Feenstra, B. (2012). Detecting Sample Misidentifications in Genetic Association Studies. *Statistical Applications In Genetics & Molecular Biology*, *11*(3), 1-17.
19. Eskin,E. (2010). Variance Component Model to Account for Sample Structure in Genome Wide Association Studies. *Nat. Genet. 42 (4)* 348-354.
20. Estrada, K., Krawczak, M., Schreiber, S., van Duijn, K., Stolk, L., van Meurs, J. B., ... & Kayser, M. (2009). A genome-wide association study of northwestern Europeans involves the C-type

natriuretic peptide signaling pathway in the etiology of human height variation. *Human molecular genetics*, *18*(18), 3516-3524.

21. Fan, L., van der Lijn, F., Schurmann, C., Gu, Z., Chakravarty, M., Hysi, P. G., & ... Daboul, A. (2012). A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. *Plos Genetics*, *8*(9), 1-13. doi: 10.1371/journal.pgen.1002932
22. Fernando, M. M. A., Stevens, C. R., Walsh, E. C., De Jager, P. L., Goyette, P., Plenge, R. M., Vyse, T. J., and Rioux, J. D. (2008). Defining the Role of the MHC in Autoimmunity: A Review and Pooled Analysis. *PLoS Genet 4 (4)* e1000024.
23. Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human Genetic Variation and its Contribution to Complex Traits. *Nat. Rev. Genet. 10 (4)* 241-251.
24. Freedman, M. L., Reich, D., Penney, K. L., McDonald, G.J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato,C. N., Pato, M. T., Petryshen, T. L., Kolonel, L. N., Lander, E. S., Sklar, P., Henderson, B., Hirschhorn, J. N., and Altshuler, D. (2004). Assessing the Impact of Population Stratification on Genetic Association Studies. *Nat.Genet. 36 (4)* 388-393.
25. Genome-Wide Association Scans. *Genet. Epidemiol. 30 (4)* 356-368.
26. Gibson, G. (2010): "Hints of hidden heritability in GWAS." Nature Genetics, 42,558–60.Park, J.-H., S. Wacholder, M. H. Gail, U. Peters, K. B. Jacobs, S. J. Chanock, and N. Chatterjee (2010): "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries." *Nature Genetics*, 42.
27. Gill P, Puch-Solis R, Curran J (2009) The lowtemplate-DNA (stochastic) threshold – its determination relative to risk analysis for national DNA databases. *Forensic Sci Int Genet* 3:104–111.
28. Greenbaum D, Du J, Gerstein M (2008) Genomic anonymity: have we already lost it? *Am J Bioeth 8*: 71–74.
29. Greene, C. S., Penrod, N. M., Williams, S. M., and Moore, J. H. (2009). Failure to Replicate a Genetic Association may Provide Important Clues about Genetic Architecture. *PLoS ONE 4 (6)* e5639.
30. Gudmundsson, J., Sulem, P., Manolescu, A., Amundadottir, L. T., Gudbjartsson, D., Helgason, A., Rafnar, T., Bergthorsson, J. T., Agnarsson, B. A., Baker, A., Sigurdsson, A., Benediktsdottir, K. R., Jakobsdottir, M., Xu, J., Blondal, T., Kostic,J., Sun, J., Ghosh, S., Stacey, S. N., Mouy, M., Saemundsdottir, J., Backman, V. M., Kristjansson, K., Tres, A., Partin, A. W., Albers-Akkers, M., Godino-Ivan Marcos, J., Walsh,P. C., Swinkels, D. W., Navarrete, S., Isaacs, S. D., Aben, K. K., Graif, T., Cashy, J., Ruiz-Echarri, M., Wiley, K. E., Suarez, B. K., Witjes, J. A., Frigge, M., Ober, C., Jonsson, E., Einarsson, G. V., Mayordomo, J. I., Kiemeney, L. A., Isaacs, W. B., Catalona, W. J., Barkardottir, R. B., Gulcher, J. R., Thorsteinsdottir, U., Kong, A., and Stefansson, K. (2007). Genome-Wide Association Study Identifies a Second Prostate Cancer Susceptibility Variant at 8q24. *Nat. Genet. 39 (5)* 631-637.
31. Gunderson, K. L., et al., (2006). Whole-genome genotyping of Haplotype tag single nucleotide polymorphisms. *Pharmacogenomics, 7(4):*p.641-8.
32. Hart C (2005). Doing your master dissertation. *London: Sage Publications.*
33. He, H., Oetting, W., Brott, M., and Basu, S. (2009). Power of Multifactor Dimensionality Reduction and Penalized Logistic Regression for Detecting Gene-Gene Interaction in CaseControl Study. *BMC Med. Genet. 10 (1)* 127.
34. Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J., and Stefansson, K. (2005). An Icelandic Example of the Impact of Population Structure on Association Studies. *Nat. Genet. 37 (1)* 90-95.
35. Hindorff, L., H. Junkins, P. Hall, J. Mehta, and T. Manolio (2009): "A catalog of published genome-wide association studies. Available at: http://www.genome.gov/gwastudies/.
36. Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., ... & Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, *4*(8), e1000167.
37. Hu, X., Liu, Q., Zhang, Z., Li, Z., Wang, S., He, L., and Shi, Y. (2010). SHEsisEpi, a GPU Enhanced Genome-Wide SNP-SNP Interaction Scanning Algorithm, Efficiently Reveals the Risk Genetic Epistasis in Bipolar Disorder. *Cell Res. 20 (7)* 854-857.
38. Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, et al. (2009) Genotype imputation accuracy across worldwide human populations. *Am J Hum Genet 84*: 235–250
39. Human Genome Project Information (2003). Policies on the release of human genomic sequence data: bermuda-quality sequence. Available:www.ornl.gov/hgmis/research/bermuda.html.
40. Jacobs, K. B., Yeager, M., Wacholder, S., Craig, D., Kraft, P., Hunter, D. J., & ... Chatterjee, N. (2009). A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature Genetics*, *41*(11), 1253-1257.
41. Jakkula, E., etal., (2008). The genome-wide patterns of variation expose significant substructure in a founder population. *Am J Hum Genet. 83 (6):* p.787-94.

42. Jakobsson, M et al. (2008). Genotype, haplotype and copy number variation in worldwide human populations. *Nature 451*, 998-1003.
43. Johnson, A. D., Leslie, R., & O'Donnell, C. J. (2011). Temporal Trends in Results Availability from Genome-Wide Association Studies. *Plos Genetics*, *7*(9), 1-3. doi:10.1371/journal.pgen.1002269
44. Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., Sabatti, C., and Kayser, M., & de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, *12*(3), 179-192.
45. Kayser, M., & Schneider, P. M. (2009). DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations. *Forensic Science International: Genetics*, *3*(3), 154-161.
46. Kayser, M., Liu, F., Janssens, A. C. J. W., Rivadeneira, F., Lao, O., van Duijn, K., ... & van Duijn, C. M. (2008). Three Genome-wide Association Studies and a Linkage Analysis Identify<i> HERC2</i> as a Human Iris Color Gene. *The American Journal of Human Genetics*, *82*(2), 411-423.
47. Kooperberg, C. and Leblanc, M. (2008). Increasing the Power of Identifying Gene x Gene Interactions in Genome-Wide Association Studies Genet. *Epidemiol. 32 (3)* 255-263.
48. Kwak, M., Joo, J., and Zheng, G. (2009). A Robust Testfor Two-Stage Design in Genome Wide Association Studies. *Biometrics 65 (4)* 1288-1295.
49. Laurie, C., K. Doheny, D. Mirel, E. Pugh, L. Bierut, T. Bhangale, F. Boehm, N. Caporaso, M.Cornelis, H. Edenberg, et al. (2010): “Quality control and quality assurance in genotypic data for genome-wide association studies,” *Genetic Epidemiology, 34*, 591–602.
50. Lee, S.Y., Chung, Y., Elston, R. C., Kim, Y., and Park, T. (2007). Log-Linear Model-Base Multifactor Dimensionality Reduction Method to Detect gene–gene Interactions. *Bioinformatics 23 (19)* 2589-2595.
51. Lewontin, R. C. (2006). Commentary: Statistical Analysis or Biological Analysis as Tools for Understanding Biological Causes. *International Journal of Epidemiology 35 (3)* 536-537.
52. Liu, F., Struchalin, M. V., Duijn, K., Hofman, A., Uitterlinden, A. G., van Duijn, C., & ... Kayser, M. (2011). Detecting Low Frequent Loss-of-Function Alleles in Genome Wide Association Studies with Red Hair Color as Example. *Plos ONE*, *6*(11), 1-12. doi:10.1371/journal.pone.0028145
53. Lou, Xiang-Yang, Guo-Bo Chen, Lei Yan, Jennie Z. Ma, Jun Zhu, Robert C. Elston, and Ming D. Li. 2007. A Generalized Combinatorial Approach for Detecting Gene-by-Gene and Gene-by-Environment Interactions with Application to Nicotine Dependence. *Am.J.Hum.Genet. 80 (6)* 1125-1137.
54. Lueders, T., & Friedrich, M. W. (2003). Evaluation of PCR amplification bias by terminal restriction fragment length polymorphism analysis of small-subunit rRNA and mcrA genes by using defined template mixtures of methanogenic pure cultures and soil DNA extracts. *Applied and Environmental Microbiology*, *69*(1), 320-326.
55. Mangold E. et al. 2010. Genome wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft plate. *Nature Genet. 42;* 24-26.
56. Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867-2873.
57. Manolio, T. A., Brooks, L. D., and Collins, F. S. (2008). A Hap Map Harvest of Insights into the Genetics of Common Disease. *J.Clin.Invest. 118 (5)* 1590-1605.
58. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The Effects of Human Population Structure on Large Genetic Association Studies. *Nat.Genet. 36 (5)* 512-517.
59. Marigorta, U. M., Lao, O., Casals, F., Calafell, F., Morcillo-Suárez, C., Faria, R., & ... Navarro, A. (2011). Recent human evolution has shaped geographical differences in susceptibility to disease. *BMC Genomics*, *12*(1), 55-68. doi:10.1186/1471-2164-12-55
60. Marsh, T. L., P. Saxman, J. Cole, and J. Tiedje. 2000. Terminal restriction fragment length polymorphism analysis program, a web-based research tool for microbial community analysis. *Appl. Environ. Microbiol.66:* 3616–3620.
61. Masca, N., Burton, P., & Sheehan, N. (2011). Participant identification in genetic association studies: improved methods and practical implications. *International Journal of Epidemiology*, *40*(6), 1629-1642. doi:10.1093/ije/dyr149
62. McCarroll, S. A., and Visscher, P. M. (2009). Findingthe Missing Heritability of Complex Diseases. *Nature 461 (7265)* 747-753.
63. Miyagawa, T., Nishida, N., Ohashi, J., Kimura, R., Fujimoto, A., Kawashima, M., & ... Tokunaga, K. (2008). Appropriate data cleaning methods for genome-wide association study. *Journal Of Human Genetics*, *53*(10), 886-893. doi:10.1007/s10038-008-0322-y

64. Moore, Jason and Bill White. (2007). Tuning ReliefF for Genome-Wide Genetic Analysis. In Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, edited by Elena Marchiori, *Jason Moore and Jagath Rajapakse. Vol. 4447,* 166-175: Springer Berlin / Heidelberg. doi:10.1007/978-3-540-71783-6_16
65. Moore, J. H. and Williams, S. M. (2005). Traversing the Conceptual Divide between Biological and Statistical Epistasis: Systems Biology and a More Modern Synthesis. *Bioessays 27 (6)* 637-646.
66. Moore, J. H. and Williams, S. M. (2009). Epistasis and its Implications for Personal Genetics.
67. Moskvina, V., Craddock, N., Müller-Myhsok, B., Kam-Thong, T., Green, E., Holmans, P., Owen, M. J., and O'Donovan, M. C. (2011). An Examination of Single Nucleotide Polymorphism Selection Prioritization Strategies for Tests of Gene–Gene Interaction. Biol. *Psychiatry 70 (2)* 198-203.
68. Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W., and Ritchie, M. D. (2008). Comparison of Approaches for Machine-Learning Optimization of Neural Networks for Detecting Gene Gene Interactions in Genetic Epidemiology. *Genet. Epidemiol. 32 (4)* 325-340.
69. Need, A.C. and Goldstein, D.B., 2009.Next generation disparities in human genomics: concerns and remedies. *Trends Genet. 25(11):* p .489-94.
70. Nyholt DR, Yu CE, Visscher PM (2008) On JimWatson's APOE status: genetic information is hard to hide. *Eur J Hum Genet 17*: 147–149.
71. Osborn, A. M., E. R. B. Moore, and K. N. Timmis. 2000. An evaluation of terminal restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. *Environ. Micro-biol. 2:* 39–50.
72. Jupp P, B (2001) Divided by information? The ''digital divide'' that really matters and the implications of the new meritocracy. London: *Demos*.
73. Parahoo K (2006) Nursing Research: Principles, Process and Issues, 2nd edn. *Palgrave Macmillan, Basingstoke*
74. Price, A. L., Butler, J., Patterson, N., Capelli, C., Pascali, V. L., Scarnicci, F., & ... Hirschhom, J. N. (2008). Discerning the Ancestry of European Americans in Genetic Association Studies. *Plos Genetics*, *4*(1), e236. doi: 10.1371/journal.pgen. 003023
75. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies *Nat.Genet. 38 (8)* 904-909.
76. Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New Approaches to Population Stratification in Genome-Wide Association Studies. *Nat.Rev.Genet. 11 (7)* 459-463.McClellan, J., and King,M. (2010). Genetic Heterogeneity in Human Disease. *Cell 141 (2)* 210-217.
77. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). *PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. 81 (3)* 559-575.
78. Risch, N., Herrell, R., Lehner, T., Liang, K. Y., Eaves, L., Hoh, J.,... & Merikangas, K. R. (2009). Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. *Jama*, *301*(23), 2462-2471.
79. Ritchie, M. D. and Motsinger, A. A. (2005). Multifactor Dimensionality Reduction for Detecting Gene-Gene and Gene-Environment Interactions in Pharma-cogenomics Studies. *Pharmacogenomics. 6 (8)* 823-834
80. Ritchie, M. D., Hahn, L. W., and Moore, J. H. (2003). Power of Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions in the Presence of Genotyping Error, Missing Data, Phenocopy, and Genetic Heterogeneity.*Genet.Epidemiol. 24 (2)* 150-157.
81. Rosenberg, N.A., et al., 2010. Genome-wide association studies in diverse populations. *Nat Rev Genet. 11(5):* p.356-66.
82. Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common Vs. Rare Allele Hypotheses for Complex Diseases. Curr.Opin.*Genet.Dev. 19 (3)* 212-219.
83. Shao, H., Burrage, L. C., Sinasac, D. S., Hill, A. E., Ernest, S. R., O'Brien, W., Courtl and H., Jepsen, K. J., Kirby, A., Kulbokas, E. J., Daly, M. J., Broman, K. W., Lander, E. S., and Nadeau, J. H. (2008). Genetic Architecture of Complex Traits: Large Phenotypic Effects and Pervasive Epistasis. Proc.*Nat.Acad.Sci. 105 (50)* 19910-19914.
84. Spielman RS, Mc Ginnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the Insulin gene region and insulin-dependent diabetes mellitus(IDDM). *Am J Hum Genet 52:*506-513.
85. Taylor P (2008) When consent gets in the way. *Nature 456*: 32–33.
86. Thomas, D. (2010). Gene-Environment-Wide Association Studies: Emerging Approaches. *Nat. Rev. Genet. 11 (4)* 259-272.
87. Tian, C., Gregersen, P. K., and Seldin, M. F. (2008). Accounting for Ancestry: Population Substructure and Genome-Wide Association Studies. *Hum.Mol. Genet. 17 (R2)* R143-R150.
88. Tian, C., Plenge, R. M., Ransom, M., Lee, A., Villoslada, P., Selmi, C., Klareskog, L., Pulver, A. E., Qi, L., Gregersen, P. K., and Seldin, M. F. (2008). Analysis and Application of European Genetic

Substructure using 300 K SNP Information. *PLoS Genet 4 (1) e4.*

89. Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Williams, S. M., and Moore, J. H. (2007). A Balanced Accuracy Function for Epistasis Modeling in Imbalanced Datasets using Multifactor Dimensionality Reduction. *Genet.Epidemiol. 31 (4)* 306-315.
90. Visscher, P. M. (2008). Sizing up human height variation. *Nature genetics*, *40*(5), 489-490.
91. Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the Genomics Era [Mdash] Concepts and Misconceptions. *Nat.Rev.Genet.9 (4)* 255-266.
92. Wang, H., Thomas, D. C., Pe'er, I., and Stram, D. O. (2006). Optimal Two-Stage Designs for Genome-Wide Association Scans. *Genet.Epidemiol. 30 (4)* 356-368.
93. Wang, Y., Liu, G., Feng, M., and Wong, L. (2011). An Empirical Comparison of several Recent Epistatic Interaction Detection Methods. *Bioinformatics 27 (21)* 2936-2943
94. Weir, B. S., Anderson, A. D., and Hepler, A. B. (2006). Genetic Relatedness Analysis: Modern Data and New Challenges. *Nat.Rev.Genet. 7 (10)* 771-780.
95. Welcome Trust Sanger Center Press Releases (2003) The finished human genome Wellcome to the genomic age. *Am.J.Hum.Genet*. 85 (3) 309-320.Available:http://www.sanger.ac.uk/Info/Press/2003/030414.shtml.
96. Zhang, Z., Ersoz, E., Lai, C., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., and Buckler, E. S. (2010). Mixed Linear Model Approach Adapted for Genome-Wide Association Studies. *Nat.Genet*. 42 (4) 355-360.
97. Zheng, G. and X. Tian (2005): "The impact of diagnostic error on testing genetic association in case-control studies*," Statistics in Medicine*, 24, 869–882.

Global Journal of Science Frontier Research: G
Bio-Tech & Genetics
Volume 21 Issue 1 Version 1.0 Year 2021
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

Null Distribution of the Test Statistic for Model Selection Via Marginal Screening: Implications for Multivariate Regression Analysis

By A. V. Rubanovich & V. A. Saenko
Russian Academy of Sciences

Summary- Marginal screening (MS) is the computationally simple and commonly used for the dimension reduction procedures. In it, a linear model is constructed for several top predictors, chosen according to the absolute value of marginal correlations with the dependent variable. Importantly, when k predictors out of *m* primary covariates are selected, the standard regression analysis may yield false-positive results if $m >> k$ (Freedman's paradox). In this work, we provide analytical expressions describing null distribution of the test statistics for model selection via MS. Using the theory of order statistics, we show that under MS, the common F-statistic is distributed as a mean of *k* top variables out of m independent random variables having a χ_1^2 distribution. Based on this finding, we estimated critical p-values for multiple regression models after MS, comparisons with which of those obtained in real studies will help researchers to avoid false-positive result. Analytical solutions obtained in the work are implemented in a free Excel spreadsheet program.

Keywords: linear models, multiple comparisons, freedman's paradox, extreme value theory, order statistics, genetic risk score.

GJSFR-G Classification: FOR Code: 060499

Null Distribution of the Test Statistic for Model Selection Via Marginal Screening: Implications for Multivariate Regression Analysis

A. V. Rubanovich [α] & V. A. Saenko [σ]

Summary- Marginal screening (MS) is the computationally simple and commonly used for the dimension reduction procedures. In it, a linear model is constructed for several top predictors, chosen according to the absolute value of marginal correlations with the dependent variable. Importantly, when *k* predictors out of *m* primary covariates are selected, the standard regression analysis may yield false-positive results if *m* >> *k* (Freedman's paradox). In this work, we provide analytical expressions describing null distribution of the test statistics for model selection via MS. Using the theory of order statistics, we show that under MS, the common *F*-statistic is distributed as a mean of *k* top variables out of *m* independent random variables having a χ_1^2 distribution. Based on this finding, we estimated critical *p*-values for multiple regression models after MS, comparisons with which of those obtained in real studies will help researchers to avoid false-positive result. Analytical solutions obtained in the work are implemented in a free Excel spreadsheet program.

Keywords: linear models, multiple comparisons, freedman's paradox, extreme value theory, order statistics, genetic risk score.

I. Introduction

Marginal screening (MS) is the simplest and most commonly used method of variable selection (Hastie, Tibshirani, 2003; Genovese et al, 2009, 2012; Leek, 2012). Its selection power is comparable to that of the Lasso, Stepwise, Compressed Sensing and other, but it is faster and easier than regularization approaches. In fact, MS is intuitively preferred by the researchers, when the number of objects (e.g. participants of a study, samples or outcomes) is much smaller than the number of explanatory variables (the so-called "n << p problem"). This problem is particularly acute in modern genetic studies such as GWAS, RNA-seq, microarray analysis, motif activity response analysis, etc. (Windle M., 2016). For all of them, a typical situation is when the number of objects is several orders of magnitude less than the number of covariates from which a statistically significant combination of predictors is derived. In fact, this is another side of an old problem of multiple comparisons, which is somewhat masked in regression analysis. Indeed, when a multiple regression model incorporating all investigated covariates is being established, no correction for multiple comparisons is required. For the standardized model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon, \qquad (1)$$

the mean of *p*-values (the *F*-test) is 0.5 under any *k* if all X_i are random and independent of *Y*. False-positives results are impossible even if *k* is large.

The situation changes dramatically under the model selection, e.g. via MS. If the predictors for a model (1) are pre-selected as top effects from a (very) large set of random variables, a false-positive result is practically inevitable. For example, if one selects top 5 predictors from 100 random variables and constructs a model (1) for 100 points, then, on average, the squared multiple correlation coefficient $R^2 = 0.23$, $p = 0.00015$, and a predictor with maximum effect with $p_1 = 0.01$ will be obtained. This phenomenon, known as Freedman's paradox (Freedman, 1983; Lukacs et al., 2010), is usually ignored in the textbooks and statistical software packages. In the literature, only analytical and empirical estimates of $E(R^2 \mid H_0)$ for MS are available (Alam, 1979; Rencher, Pun, 1980; Wray, 2013).

The purpose of this work was to explicitly address null-distribution of the *F*-statistic, which is used for testing the significance of a model (1), when model selection is performed with MS. In addition, we present related formulae for the case when the reduction of the number of independent variables is performed by calculating the so-called «risk score» (such as genetic risk score (GRS) or polygenic risk score (PGRS) in genetic studies). The derived analytical solutions are implemented in a free Excel program, which evaluates statistical significance of the results of a regression analysis after MS.

II. Notations and Abbreviations

MS - marginal (or correlation) screening;

RS - risk score; GRS – genetic risk score; PGRS – polygenic risk score;

Author α: Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia. e-mail: rubanovich@vigg.ru

Author σ: Department of Radiation Molecular Epidemiology, Atomic Bomb Disease Institute, Nagasaki University, Nagasaki, Japan.

CDF - cumulative distribution function;

PDF - probability density function;

iid - independently and identically distributed;

r.v. - random variable;

X – random variable X;

$E(\mathbf{X}), D(\mathbf{X})$ - the mean value and the variance of X;

X ~ Y - X and Y are identically distributed;

$\mathbf{X} \,|\, H_0$ - X under null hypothesis;

X $\sim \mathbf{F}_{k,n}$ - X has F-distribution with parameters *k* and *n* (degrees of freedom);

X $\sim \chi_k^2$ - X has chi-squared distribution with parameter *k* (degrees of freedom);

X $\sim \mathbf{N}(\mu, \sigma^2)$ - X has normal distribution with parameters μ and σ^2 (mean and variance, respectively);

$\Phi(x)$ - CDF of the standard normal distribution $\mathbf{N}(0,1)$;

$\Phi^{-1}(x)$ - the inverse CDF (quantile function) of the standard normal distribution;

n - sample size;

m - total number of independent variables (predictors);

k - the number of top predictors selected for a model construction.

III. Method

The distributions of top extreme (hereafter – top) values of r.v.s are examined in frame of the order statistics theory (Ahsanullah et al., 2013; Arnold et al., 2008).

Let r.v.s $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_m$ are independent and identically distributed, and $\mathbf{Z}_{1:m} \le \mathbf{Z}_{2:m} \le ... \le \mathbf{Z}_{m:m}$ are the same r.v.s in ascending order. Then r.v. $\mathbf{Z}_{i:m}$ is called the *i*th order statistic and $\mathbf{Z}_{m:m} = \max\{\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_m\}$. The arithmetic mean of top *k* order statistics is called the «selection differential»:

$$D_{k,m}(\mathbf{Z}) \equiv k^{-1}(\mathbf{Z}_{m:m} + \mathbf{Z}_{m-1:m} + ... + \mathbf{Z}_{m-k+1:m}) = k^{-1} \sum_{i=m-k+1}^{m} \mathbf{Z}_{i:m}.$$

The properties of the r.v. $D_{k,m}(\mathbf{Z})$ has been studied in detail (Nagaraja, 1980, 1982, 2006; Arnold et al., 2008). Further reasoning involves the formalism published earlier (Stigler, 1973). Let $m \to \infty$ and *p* = *k/m* = const. Then r.v. $D_{k,m}(\mathbf{Z})$ is an asymptotically normally distributed r.v. with the following parameters:

$$D_{k,m}(\mathbf{Z}) \sim \mathbf{N}\left(\mu_p, \frac{1}{k}\left(\sigma_p^2 + (1-p)(\mu_p - z_p)^2\right)\right), \quad (2)$$

$$\mu_p = E(\mathbf{Z} \,|\, \mathbf{Z} > z_p) = \frac{1}{p}\int_{z_p}^{\infty} z f(z)dz,$$

$$\sigma_p^2 = D(\mathbf{Z} \,|\, \mathbf{Z} > z_p) = \frac{1}{p}\int_{z_p}^{\infty} (\mu_p - z)^2 f(z)dz,$$

where $F(z)$ and $f(z)$ are the CDF and PDF of the r.v. $\mathbf{z}$; $z_p = F^{-1}(1-p)$; μ_p and σ_p^2 are the mean and variance of the distribution obtained by *F* trimming to below z_p.

IV. Results

a) Additivity of Cohen's f^2 effect size for independent predictors

Our analytical results described below are essentially based on the Proposition 1. We believe this proposition may be of independent interest.

Proposition 1: Let correlations between independent variables X_i in a regression model (1) are

$$Cor(X_i, X_j) = r_i r_j, \; i \ne j,$$

where $r_i = Cor(Y, X_i)$ are marginal correlations between the dependent variable and predictors. Then the squared multiple correlation (R^2) for model (1) satisfies

$$\frac{R^2}{1-R^2} = \sum_{i=1}^{k} \frac{r_i^2}{1-r_i^2}. \quad (3)$$

Proof. It is known that $r = C\beta$ and $R^2 = r^T\beta$, where *r* is a column vector whose *i*th entry is r_i; β is a column vector whose *i*th entry is β_i; *C* is the correlation matrix for the predictor variables. By the condition of Proposition 1, $C = rr^T + D$, where *D* is a diagonal matrix with the elements $D_{ii} = 1 - r_i^2$ on the diagonal. Then, $r = C\beta = r(r^T\beta) + D\beta = R^2 r + D\beta$. Further, $\beta = (1-R^2)D^{-1}r$ and hence, $r^T\beta = (1-R^2)(r^T D^{-1} r)$. Therefore,

$$\frac{R^2}{1-R^2} = r^T D^{-1} r = \sum_{i=1}^{k} \frac{r_i^2}{1-r_i^2}.$$

Comments

1. Index $f^2 = R^2/(1-R^2)$ is known as Cohen's f^2 effect size (Cohen, 1988; 1992).
2. The condition $Cor(X_i, X_j) = r_{ij} = r_i r_j$ assumes that all predictors are statistically independent and their mutual correlations are fully due to the correlation with *Y*. In this case, the partial

correlations between the predictor variables (with fixed *Y*) are:

$$r_{ij.Y} = \frac{r_{ij} - r_i r_j}{\sqrt{(1-r_i^2)(1-r_j^2)}} = 0$$

Thus, according to the Proposition 1 proved above, Cohen's indices *f*² are additive when the predictors are independent.

b) Null distribution of the F-statistic for marginal screening

When testing the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ for the model (1), Fisher's *F*-statistic is calculated as

$$\mathbf{F} = \frac{n-k-1}{k}\frac{\mathbf{R}^2}{1-\mathbf{R}^2} \,,$$

where R² is the squared multiple correlation coefficient for the sample size *n*. If the model is initially constructed for variables {*X*$_i$}, *i* = 1, …, *k*, then the null distribution of F is: $\mathbf{F} \,|\, H_0 \sim \mathbf{F}_{k,n-k-1}$. However, this is incorrect if the predictors were preliminarily selected as top from a large number of variables {*X*$_i$}, *i* = 1, …, *m*, where *m* >> *k*.

Proposition 2: Let the model (1) is constructed for *k* predictors which were chosen as the top $|\,Cor(Y, X_i)\,|$ out of *m* variables. Then, when *n* >> *k*, the *F*-statistic is approximately distributed as

$$\mathbf{F} \,|\, H_0 \sim \frac{1}{k}(\mathbf{Z}_{m:m} + \mathbf{Z}_{m-1:m} + ... + \mathbf{Z}_{m-k+1:m}) \equiv D_{k,m}(\chi_1^2) \,,$$

$$\mathbf{Z} \sim \chi_1^2 \,.$$

Proof. Let in a regression model all *Y*, *X*$_1$, *X*$_2$ …, *X*$_m$ are independent and identical r.v.s having standard normal distribution *N*(0,1). Consider r.v.s $\mathbf{f}_i^2 = \mathbf{r}_i^2 / (1-\mathbf{r}_i^2)$, where $\mathbf{r}_i^2$ are sample correlations. If the number of points in the model is *n*, then $(n-2)\mathbf{f}_i^2 \sim \mathbf{F}_{1,n-2} \to \chi_1^2$ under $n \to \infty$. Consequently, the "top" predictors correspond to *k* greatest values among the *m* values of independent r.v.s $\mathbf{Z} \sim \chi_1^2$. Then, when $n >> k$, according to Proposition 1:

$$\mathbf{F} = \frac{n-k-1}{k}\frac{\mathbf{R}^2}{1-\mathbf{R}^2} = \frac{n-k-1}{k}\sum_{i=1}^{k}\mathbf{f}_i^2 \sim \frac{n-k-1}{k(n-2)}(\mathbf{Z}_{m:m} + \mathbf{Z}_{m-1:m} + ... + \mathbf{Z}_{m-k+1:m}) \approx D_{k,m}(\chi_1^2) \,.$$

Corollaries

1. Model selection leads to the substantial inflation of $E(\mathbf{F} \,|\, H_0)$. Without model selection, $k = m$ and, consequently, $E(\mathbf{F} \,|\, H_0) = E(D_{m,m}(\chi_1^2)) = E(\chi_1^2) = 1$. In case $k << m$ (model selection by MS), $E(\mathbf{F} \,|\, H_0) = E(D_{k \cdot m}(\chi_1^2)) \to \infty$ when $m \to \infty$. In any case, the distribution of F| *H*$_0$ is weakly dependent on the sample size *n*.
2. As $n \to \infty$, $m \to \infty$ and $p = k / m = \text{const}$, the null distribution of the *F*-statistic is approximately $\mathbf{F} \,|\, H_0 \sim \mathbf{N}(\mu_\mathbf{F}, \sigma_\mathbf{F}^2)$, where

$$\mu_\mathbf{F} = 1 + \frac{1}{p}\sqrt{\frac{2z_p}{\pi}}e^{-\frac{z_p}{2}} \,,$$

$$\sigma_\mathbf{F}^2 = \frac{1}{k}\left(3 + \frac{1}{p}\sqrt{\frac{2z_p}{\pi}}(3 + z_p)e^{-\frac{z_p}{2}} - \mu_p^2 + (1-p)(\mu_p - z_p)^2\right) \quad (4)$$

and $z_p = F^{-1}(1-p) = \Phi^{-1}(1-p/2)^2$. Here F^{-1} and Φ^{-1} are the quantile functions of χ_1^2 and normally distributed r.v., respectively. Expressions (4) are obtained through the PDF of χ_1^2 substitution into formula (2). Series expansion of the expression (4) under $z_p \to \infty$ leads to an approximation for the distribution of r.v. F for the case $k << m$:

$$\mathbf{F} \,|\, H_0 \sim \mathbf{N}\left(2 + z_p,\ \frac{2}{k}\left(1 + \sqrt{\frac{2z_p}{\pi}}\right)\right). \quad (5)$$

3. Let $\mathbf{p} = 1 - F_{k,n-k-1}(\mathbf{F})$ is a r.v. corresponding to the spurious *p*-value for the model (1) under MS. Equation (4) allows to estimate the lower 5% quantile of its null-distribution: $Q_{5\%}(\mathbf{p} \,|\, H_0) \approx 1 - F_{k,n-k-1}(\mu_\mathbf{F} + 1.65\sigma_\mathbf{F})$
4. When *m* >> *k*, expectation of the squared multiple correlation coefficient under *H*$_0$ approximately equals to

$$E(\mathbf{R}^2 \,|\, H_0) = \frac{k\mu_\mathbf{F}}{n-k-1+k\mu_\mathbf{F}} \,. \quad (6)$$

c) Significance of the regression coefficient β$_1$ for a predictor based on the maximum effect

Usually the researcher desires to control not only significance of the whole model, but also of the predictors in the model (1). Here, we focus on the statistical significance of a predictor with the maximum effect, denoted as X_1: $|\beta_1| \equiv \max_i |\beta_i|$. The standard procedure for testing H_0: $\beta_1 = 0$ for the model (1) involves the calculation of a statistic

$$\mathbf{F}_1 = (n-k-1)\frac{\mathbf{R}^2 - \mathbf{R}^2_{2,\ldots,k}}{1-\mathbf{R}^2},$$

where $\mathbf{R}^2_{2,\ldots,k}$ are the squared multiple correlations for predictors X_2, X_3 …, X_m. It is known that without model selection, $\mathbf{F}_1 \sim \mathbf{F}_{1,n-k-1} \to \chi^2_1$ under $n \to \infty$. It turns out that with MS, r.v. $\mathbf{F}_1$ approximately equals to the maximum value of m independent realizations of r.v. $\mathbf{Z} \sim \chi^2_1$. Thus, we arrive to

Proposition 3. Let the model (1) is constructed for k predictors, which were selected as top of $|\mathbf{r}_i| = |Cor(\mathbf{Y}, \mathbf{X}_i)|$ from m variables. Then $\mathbf{F}_1 \,|\, H_0 \sim \mathbf{Z}_{m:m}$, where $\mathbf{Z} \sim \chi^2_1$.

Proof. Let r.v.'s Y, X$_1$, X$_2$ …, X$_m$ are independent r.v.s, which have normal distribution *N*(0,1). Let $\mathbf{f}_i^2 = \mathbf{r}_i^2/(1-\mathbf{r}_i^2)$ и $\mathbf{f}^2 = \mathbf{R}^2/(1-\mathbf{R}^2)$. Then according to Proposition 1,

$$\frac{\mathbf{R}^2_{2,\ldots,k}}{1-\mathbf{R}^2_{2,\ldots,k}} = \sum_{i=2}^{k}\mathbf{f}_i^2 = \mathbf{f}^2 - \mathbf{f}_1^2$$

Hence, under $n \to \infty$

$$\mathbf{F}_1 \,|\, H_0 = (n-k-1)\left(\frac{\mathbf{f}^2}{1+\mathbf{f}^2} - \frac{\mathbf{f}^2-\mathbf{f}_1^2}{1+\mathbf{f}^2-\mathbf{f}_1^2}\right)(1+\mathbf{f}^2) = \frac{(n-k-1)\mathbf{f}_1^2}{1+\mathbf{f}^2-\mathbf{f}_1^2} \sim \frac{(n-k-1)\mathbf{Z}_{m:m}}{n-2+\sum\limits_{i=m-k+1}^{m-1}\mathbf{Z}_{i:m}} \approx \mathbf{Z}_{m:m},\ \mathbf{Z} \sim \chi^2_1 .$$

Corollaries.

1. Marginal screening does not substantially change the $\mathbf{F}_1$ statistic, which only slightly depends on k and n provided $n > m$ (see Supplement 1, Fig. 2).
2. Let $\mathbf{p}_1$ is a r.v. related to the spurious p-value for the predictor with the maximum effect under H_0. Then, approximately, $\mathbf{p}_1 \,|\, H_0 \sim 1 - \mathbf{U}_{m:m}$ under $n >> 1$, where U has the uniform distribution on [0,1]. Indeed, under $n >> 1$

$$\mathbf{p}_1 \,|\, H_0 \sim 1 - F_{1,n-k-1}(\mathbf{F}_1) \approx 1 - F_{\chi^2_1}(\mathbf{Z}_{m:m}) = 1 - F_{\chi^2_1}(F^{-1}_{\chi^2_1}(\mathbf{U}_{m:m})) = 1 - \mathbf{U}_{m:m}$$

Here, the relation: $\mathbf{X}_{i:m} = F_X^{-1}(\mathbf{U}_{i:m})$ is used, where $F_X^{-1}$is the quantile function of a r.v. X.

Hence, $E(\mathbf{p}_1 \,|\, H_0) \approx 1 - E(\mathbf{U}_{m:m}) = 1/(m+1)$ (Ahsanullah et al., 2013), and the lower α-quantile for $\mathbf{p}_1$ equals to $Q_\alpha(\mathbf{p}_1 \,|\, H_0) \approx 1-(1-\alpha)^{1/m} \approx \alpha/m$.

d) Risk summation

A special method for reducing the number of predictors is widely used in modern genetic studies, termed "risk summation" (RS). For example, in «case - control» studies, the so-called genetic risk scores (GRS) are often calculated as the total number of "risk alleles", which are more frequent in patients or affected participants. Then, a comparison of GRS between samples of healthy controls and affected individuals is performed using e.g. Student's *t*-test. In the case of quantitative trait, its correlation with the number of alleles is sometimes calculated, resulting in the increase of the dependent variable.

Let us calculate the null distribution of the *t*-statistic for RS comparison. For the model (1) we define the new variable «risk score» X_{RS} as

$$X_{RS} = \mathrm{Sign}(\beta_1)X_1 + \mathrm{Sign}(\beta_2)X_2 + \ldots + \mathrm{Sign}(\beta_k)X_k,$$

and consider a simple regression $Y = \beta X_{RS} + \varepsilon$. Let $R_{RS} \equiv Cor(Y, X_{RS}) = \beta$. Corresponding *F*-statistic for this model (F$_{RS}$) is calculated in case of a continuous dependent variable, and Student's *t*-test is applied if the dependent variable is binary. The *t*-statistic is $\mathbf{T}_{RS} = \sqrt{\mathbf{F}_{RS}}$. Then, if H_0 is true, the following proposition takes place:

Proposition 4.

$$\mathbf{T}_{RS} \mid H_0 \sim \frac{1}{\sqrt{k}}(\mathbf{Z}_{m:m} + \mathbf{Z}_{m-1:m} + \ldots + \mathbf{Z}_{m-k+1:m}) = \sqrt{k} D_{k,m}(\boldsymbol{\chi}_1)\,,\ \mathbf{Z} \sim \boldsymbol{\chi}_1\,.$$

Proof. Indeed, for a standardized model $\sigma_{X_i} = \sigma_Y = 1$, $i = 1,\ldots,k$. Then, according to Proposition 1 under H_0:

$$\mathbf{R}_{RS} = Cor(\mathbf{Y}, \mathbf{X}_{RS}) = \left(\left(\sum_{i=1}^{k} |\mathbf{r}_i|\right)^2 + \sum_{i=1}^{k}(1-\mathbf{r}_i^2)\right)^{-1/2} \sum_{i=1}^{k} |\mathbf{r}_i|$$

from which under $n \to \infty$

$$\mathbf{F}_{RS} = (n-2)\frac{\mathbf{R}_{RS}^2}{1-\mathbf{R}_{RS}^2} = (n-2)\frac{\left(\sum_{i=1}^{k} |\mathbf{r}_i|\right)^2}{\sum_{i=1}^{k}(1-\mathbf{r}_i^2)} \approx (n-2)\left(\sum_{i=1}^{k} \frac{\mathbf{f}_i}{\sqrt{k}}\right)^2,$$

and $\mathbf{f}_i^2 = \mathbf{r}_i^2/(1-\mathbf{r}_i^2)$. Taking into account that $\mathbf{T}_{RS} = \sqrt{\mathbf{F}_{RS}}$, and $\sqrt{(n-2)}\mathbf{f}_i \sim \boldsymbol{\chi}_1$, we arrive to Proposition 4.

Corollaries.

1. When $n \to \infty$, $m \to \infty$ и $p = k/m = \mathrm{const}$, null-distribution of the *t*-statistic is:

$\mathbf{T} \mid H_0 \sim \mathbf{N}\left(\mu_{\mathrm{T}}, \sigma_{\mathbf{T}}^2\right)$, where

$$\mu_{\mathbf{T}} = \frac{1}{p}\sqrt{\frac{2k}{\pi}} e^{-\frac{z_p}{2}},$$

$$\sigma_{\mathbf{T}}^2 = 1 - \frac{2}{\pi p} e^{-z_p} - \sqrt{\frac{2z_p}{\pi}}\frac{(2p-1)}{p} e^{-\frac{z_p}{2}} + (1-p)z_p \qquad (7)$$

and $z_p = \Phi^{-1}(1-p/2)^2$. Expressions (7) are obtained through the PDF of χ_1, which is $f(z) = \sqrt{2/\pi} e^{-z^2/2}$, substitution into the formula (2). For $p < 0.1$, the following approximation can be used:

$$\mathbf{T} \mid H_0 \sim \mathbf{N}\left(\sqrt{k}\Phi^{-1}\left(1-\frac{k}{2m}\right),\ 2\Phi^{-1}\left(1-\frac{k}{2m}\right)^{-1}\right),$$

where Φ^{-1} is the quantile of the normal distribution. When $k > 0.2m$,

$$\mathbf{T} \mid H_0 \sim \mathbf{N}\left(\sqrt{2k/\pi},\ 1-2/\pi\right). \qquad (8)$$

2. $\mathbf{R}_{RS}^2 \le \mathbf{R}^2$. This inequality follows from Proposition 1 and the convexity of the $f(x) = x(1-x)^{-1}$ function. When $k << m$, an approximate equality takes place: $\mathbf{R}_{RS}^2 \approx \mathbf{R}^2$, and $\mathbf{F}_{RS} \approx k\mathbf{F}$, where F and R^2 are defined in Proposition 2.

V. Related Works

Analytical approximation for the distribution *F* | *H*$_0$ under MS, is described here for the first time to the best of the authors' knowledge. Previously, the analytical and empirical expressions for $E(\mathbf{R}^2 \mid H_0)$ (Rencher, Pun, 1980; Wray, 2013) and the top quintile $Q_{1-\alpha}(\mathbf{R}^2 \mid H_0)$ (Diehr, Hoflin, 1974; Salt, 2007) were proposed. The empirical formula for $Q_\alpha(\mathbf{p} \mid H_0)$ was obtained by Salt (2007). A possible link between the null-distribution of $\mathbf{R}^2$ and order statistics for χ_1^2 under MS was first noted by Alam and Wallenius (Alam, Wallenius, 1979).

Our results are closest to the recent work of J. Fan that addressed "spurious correlations" (Fan et. al., 2017). Applicably to our work, those results are as follows. Assume that the *Y* and *X$_i$* in (1) are independent and identical r.v.s having the standard normal distribution *N*(0,1). The maximum spurious correlation is defined as $\hat{\mathbf{R}}_n^2(k,m) = \max\limits_{\beta} Cor(Y, \sum_{i=1}^{m} \beta_i X_i)^2$ subject to $\|\beta\|_0 = k$ (the number of nonzero *β$_i$*'s equals *k*). If *n* >> (*k* ln *m*)^7, then the asymptotic distributions of the maximum spurious correlation is:

$$n\hat{\mathbf{R}}_n^2(k,m) \sim \mathbf{Z}_{m:m} + \mathbf{Z}_{m-1:m} + \ldots + \mathbf{Z}_{m-k+1:m}, \mathbf{Z} \sim \chi_1^2\,. \qquad (9)$$

The condition *n* >> (*k* ln *m*)^7 means that the practical use of this formula is possible only with very large *n*. Even at *m* = 100 the required sample size is *n* >> 44000. Otherwise, the formula (9) leads to inadequate estimates: $E(\hat{\mathbf{R}}_n^2(k,m)) > 1$. Our result (Proposition 2) is meaningfully similar to (9) yet, in contrast, requires only *n* >> *k* and therefore it is devoid of such a disadvantage.

Recently, J. Taylor and colleagues developed an innovative approach to the problem of statistical inference after model selection (Lee, Taylor, 2014; Tibshirani, Taylor, 2016). For a regression model with Gaussian errors, they presented a method for exact inference, conditional on a polyhedral constraint on the observations *Y*. This approach is applicable to different model selection procedures including marginal screening, forward stepwise regression, lasso, least angle regression and other sequential regression techniques being an exciting new advance. Note, however, that it is a conditional test in which any estimator used for inference is influenced by correlations with predictors that are not selected. As an example, consider the simplest case $k = 1$ (one top predictor). Than the "polyhedral lemma" (Lee, Taylor, 2014) would yield the following distribution of the regression coefficient β_1 under H_0:

$$\beta_1 \mid H_0 \sim TN[|r_2|; \infty].$$

Here r_2 is the second top marginal correlation between the dependent variable and predictors; $TN[a; b]$ is the normal distribution truncated to the interval $[a; b]$. Naturally, the p-value obtained will depend heavily on the unselected predictor X_2. In contrast, our unconditional test (Proposition 2) depends only on the selected first top predictor X_1:

$$(n-2)\frac{r_1^2}{1-r_1^2} \mid H_0 \sim Z_{m:m},\ Z \sim \chi_1^2\ .$$

We believe that the interpretation of the conditional "exact" p-value is still unclear since the expected value is also affected by the unselected predictors. It is also clear that a conditional post-selection test has lower power compared to the unconditional test, since for any X

$$\alpha > P(\text{Type 1 error} \mid \text{X}) > P(\text{Type 1 error, X}).$$

Post-selection inference is a very promising approach to inference that appears to work very well in case $n > m$. How well it performs when $m > n$ needs to be demonstrated in the future, since the method remains very new at this point.

VI. Practical Relevance

Our theoretical considerations have immediate application since the derived formulas enable quick and efficient evaluation of statistical significance of a multiple regression model, predictors in which are selected via MS. The accuracy of approximations was tested using computer simulations for a wide range of parameters (Supplement 1). More accurate estimates can be obtained by applying the $F_{1,n-2}$ instead of χ_1^2.

For calculations according to the formulas from Propositions 2-4, we suggest using the H0_Model_Selection.xlsx spreadsheet (Supplement 2). One needs to enter three numbers $\{n, m, k\}$, where n - sample size, m - the initial number of covariates (independent variables), and k - the number of top predictors selected to construct the model (1). The program instantly estimates the mean value and 95% quantiles under H_0 for the F-statistic, squared multiple correlation coefficient, and critical p-values for the model and for the top predictor. The program also includes an option of adjustment of the spurious p-values for the whole model and for a top predictor obtained in the experiment taking into account MS. In case of RS comparison, the program calculates mean values of the Student's t-statistic and of corresponding p-value under H_0 and recalculates the actual p-value based on the distribution of $\mathbf{T}|H_0$.

The major limitation of our theoretical formalism and, consequently, of the program stems from the $n \to \infty$ assumption. If sample size is very small ($n < 30$), formula 4 yields the F-statistic underestimated for 30-40%, depending on k, according to our computer simulations (data not shown). The problem of the F underestimation is negligible, however, if sample size exceeds 50.

VII. Discussion

Our work demonstrates that when MS is performed, the F-statistic under H_0 is distributed as a mean of k top variables selected out of m independent random variables having a χ_1^2 distribution. It is an intuitively expected and a clear finding. If MS is not performed (k=m), $E(F \mid H_0) = E(\chi_1^2) = 1$, precluding false-positive result. However, if the $m >> k$ condition takes place, $E(F \mid H_0)$ may become arbitrarily large that brings about a misleading result.

Besides of theoretical interest, our work resulted in a practical implementation in a form of software solution, which may be useful for a wide range of scientific investigations, including but not limited to genetic association studies and those employing regression analysis with stepwise selection (common in e.g. psychology, ecology and economics). For example, a regression model for 100 points with 3 predictors selected from the initial set of 15 covariates returns a p-value of 0.005 for the model. Having entered three values {100, 15, 3} to the program, one would find that under the null hypothesis the expected mean p-value for such a model is 0.025 and the 5%-quantile is 0.0019. This means that in fact the obtained model cannot be considered statistically significant. Entering the p-value of 0.005 to the program will result in the adjusted p-value of 0.155 for the model. Note that, among 10 independent covariates, there will be, on average, 2 top predictors for which the model (1) is formally "significant": $p \approx 0.04$ if $n = 100$.

Chance of obtaining a false-positive result is also high in the studies addressing risk scores. Equation (8) shows that when two samples are compared by RS, five factors are sufficient to obtain a "significant" difference using Student's t-test. Indeed, when $k = m = 5$,

$$\mathbf{T} \mid H_0 \sim \mathbf{N}\left(\sqrt{2 \cdot 5/\pi},\ 1-2/\pi\right) \approx N(1.78, 0.36),$$

which corresponds to the mean p-value of 0.037. Numerous examples of works with strongly overestimated significance of the effects based on GRS are given in our earlier work (Rubanovich, Khromov-Borisov, 2016).

It should be acknowledged that the formulas (4) only partly solve the problem of null distribution of the test statistic under model selection. More sophisticated algorithms may lead to the higher values of R^2 | H_0 than MS. This is especially noticeable for the methods based on sparsity solutions such as Lasso or Compressed Sensing (CS). Our computer simulations for $k>$ 3 and $m>$ 200 demonstrate that usually $E(\mathbf{R}^2_{MS} \mid H_0) < E(\mathbf{R}^2_{CS} \mid H_0)$. This hints that if model selection is preformed using the advanced methods, actual critical p-values can be even lower than those obtained with MS. However, theoretical and analytical basis for these situations is not available so far.

In conclusion, our work provides formal explanation of the distribution of the test statistic for multiple regression models which utilize a subset of top predictors derived from a large set of explanatory variables through the preliminary marginal screening. Although some approximations are used, the approach has sufficient accuracy. Software implementation of the analytical expressions is available and can be conveniently used even by an unexperienced researcher willing to evaluate the validity of own finding. We believe this implementation is a useful means of avoiding false-positive results in the studies, which might have been flawed by no fault of the investigators but rather by insufficiency of theoretical and practical solutions for the problem of "multiple testing" in regression analysis.

Supplementary Materials

Supplement 1. Computer simulations.

Supplement 2. H0_Model_Selection.xlsx spreadsheet

Acknowledgments

Conflict of Interest: None declared.

Funding

This work was supported in part by KAKENHI Grant Number 15K09438 from the Japan Society for the Promotion of Science (JSPS).

References Références Referencias

1. Ahsanullah, M., Nevzorov, V.N., and Shakil M. (2013). An Introduction to Order Statistics. Amsterdam – Paris – Beijing, Atlantis Press.
2. Alam, K., and Wallenius, K. (1979). Distribution of a sum of order statistics. *Scandinavian Journal of Statistics* 6, 123-126.
3. Arnold, B.C., Balakrishnan, N., and Nagaraja, H.N. (2008). A First Course in Order Statistics. SIAM-Society for Industrial and Applied Mathematics.
4. Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. 2nd ed, Routledge. Cohen, J. (1992). A Power Primer. *Psychological Bulletin* 112, 155-159.
5. Diehr, G., and Hoflin, D.R. (1974). Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics* 16, 317-320.
6. Fan, J., Shao, Q., and Zhou, W. (2017). Are Discoveries Spurious? Distributions of Maximum Spurious Correlations and Their Applications. arXiv:1502.04237 [math.ST]
7. Foster, D.P., and Stine, R.A. (2006). Honest confidence intervals for the error variance in stepwise regression. *Journal of Economic and Social Measurement* 31, 89-102.
8. Freedman, D.A. (1983). A Note on Screening Regression Equations. *The American Statistician*, 37, 152-155.
9. Genovese, C.R., Jin, J., and Wasserman, L. (2009). Revisiting marginal regression. *arXiv*:0911.4080v1 [math.ST].
10. Genovese, C.R., Jin, J., Wasserman, L., and Yao, Z.A. (2012). Comparison of the lasso and marginal regression. *Journal of Machine Learning Research* 13, 2107-2143.
11. Hastie, T., and Tibshirani, R. (2003). Expression arrays and the $p >> n$ problem. Available at https://web.stanford.edu/~hastie/Papers/pgtn.pdf
12. Lee, J.D., Taylor, J.D. (2014). Exact Post Model Selection Inference for Marginal Screening. arXiv:1402.5596 [stat.ME].
13. Leek, J. (2016). Prediction: the lasso vs just using the top 10 predictors. *Blog "Simply Statistics"*. Available at http://simplystatistics.tumblr.com/post/18132467723/prediction-the-lasso-vs-just-using-the-top-10.
14. Lukacs, P.M., Burnham, K.P., and Anderson, D.R. (2010). Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics* 62, 117–125.
15. Nagaraja, H.N. (1980). Contributions to the theory of the selection differential and to order Statistics. Dissertation. Digital Repository @ Iowa State University. Available at http://lib.dr.iastate.edu/rtd/6746/
16. Nagaraja, H.N. (1982). Some Nondegenerate Limit Laws for the Selection Differential," *Annals of Statistics* 10, 1306-1310.
17. Nagaraja, H.N. (2006). Order Statistics from Independent Exponential Random Variables and the Sum of the Top Order Statistics. In "Advances in Distribution Theory, Order Statistics, and Inference", eds. N. Balakrishnan, J.M. Sarabia, and E. Castillo.
18. Rencher, A.C., and Pun, F.C. (1980). Inflation of R^2 in best subset regression. *Technometrics* 22, 49-53.
19. Rubanovich, A.V., and Khromov-Borisov N.N. (2016). Genetic risk assessment of the joint effect

of several genes: critical appraisal. *Russian Journal of Genetics* 52, 757–769.

20. Salt, D.W., Ajmani, S., Crichton, R., and Livingstone, D.J. (2007). An improved approximation to the estimation of the critical F values in best subset regression. *Journal of Chemical Information and Modelingl* 47, 143-149.
21. Stigler, S.M. (1973). The asymptotic distribution of trimmed mean. *Annals of Statistics* 1, 472-477.
22. Tibshirani, R.J., Taylor, J., Richard Lockhart, R., Tibshirani, R. (2016). Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association*, 111:514, 600-620.
23. Windle, M. (ed.) (2016). Statistical Approaches to Gene x Environment Interactions for Complex Phenotypes. MIT Press.
24. Wray, N.R., Yang, J., Hayes, B.J., Wray, N.R., Yang, J., Hayes, B.J., Price. A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* 14, 507-515.

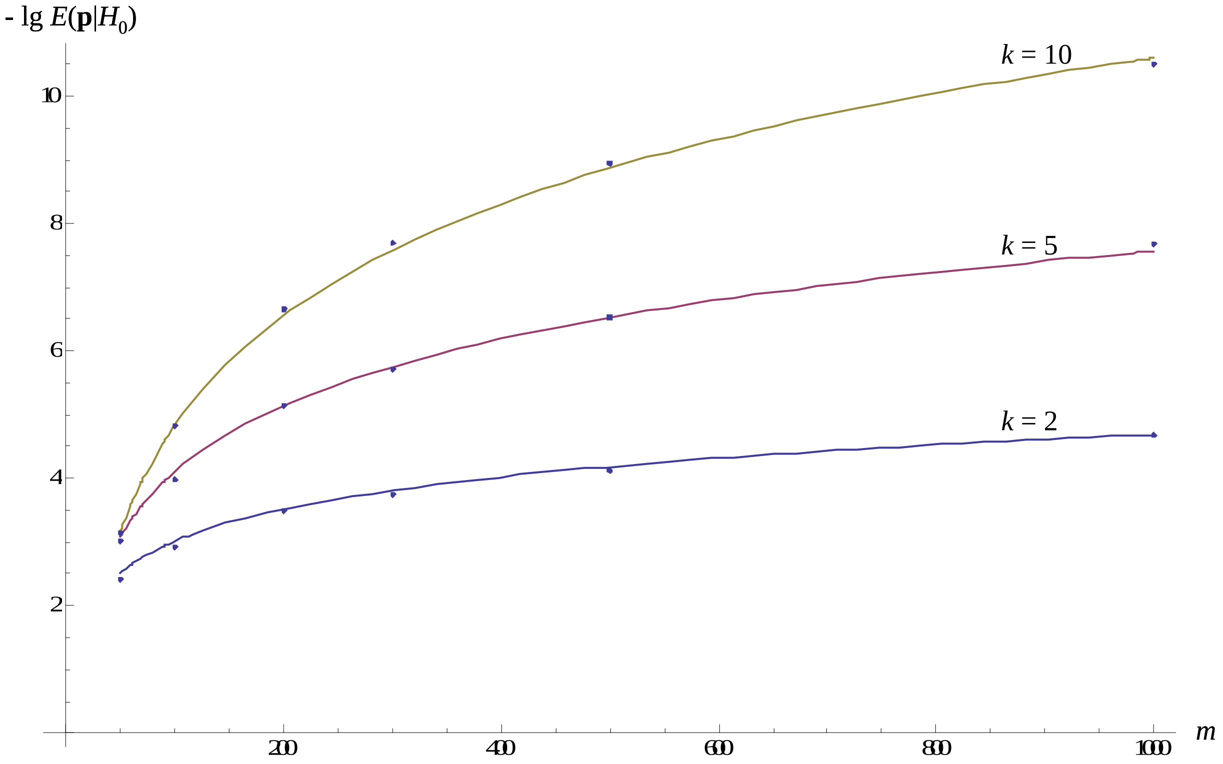

Fig. 1: Expected spurious *p*-value for the regression model (1) with regard to the total number of independent variables (*m*) and the number of top of them (*k*) used in the model under H_0 (n = 200). Plotted points are the result of numerical simulations (averaged for 1000 random repeats)

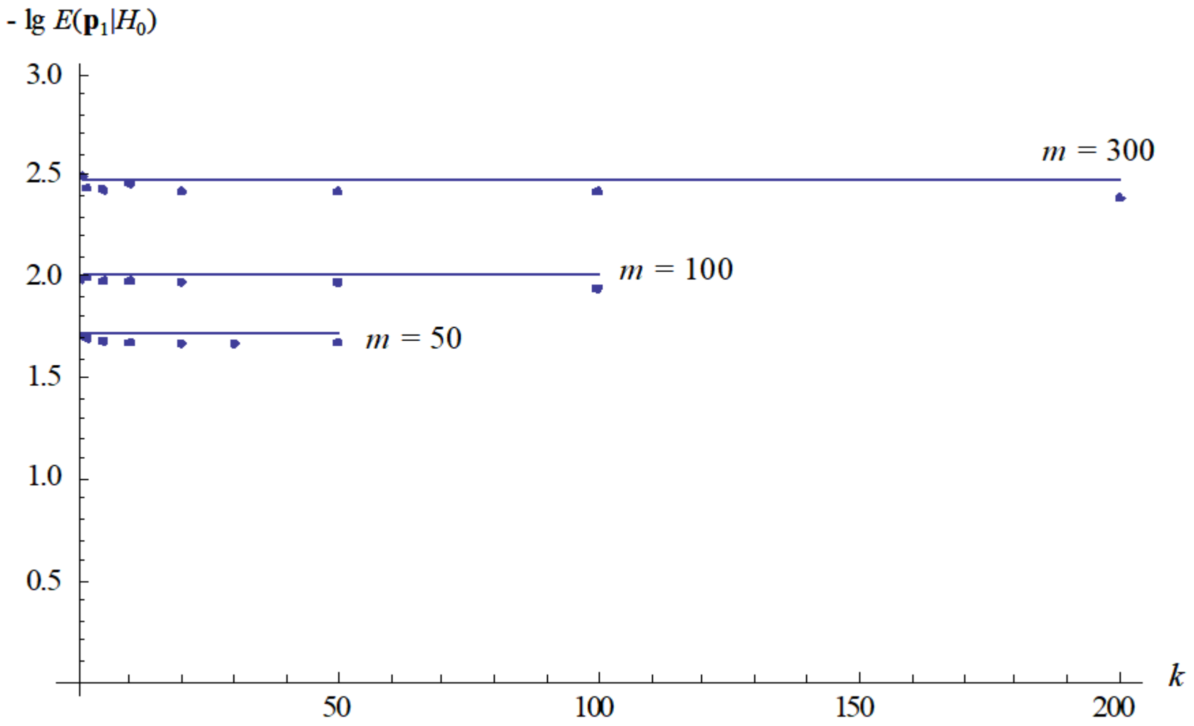

Fig. 2: Relationship between the spurious p-value for the top predictor (p_1) and the number of top predictors (k) used in the model (1) under H_0 ($n = 500$). Plotted points are the result of numerical simulations (averaged for 1000 random repeats). Horizontal lines represent the result described in Corollary 3.2: $E(p_1|H_0) \approx 1/(m + 1)$. When $n > m$, the mean $E(p_1|H_0)$ only weakly depends on k

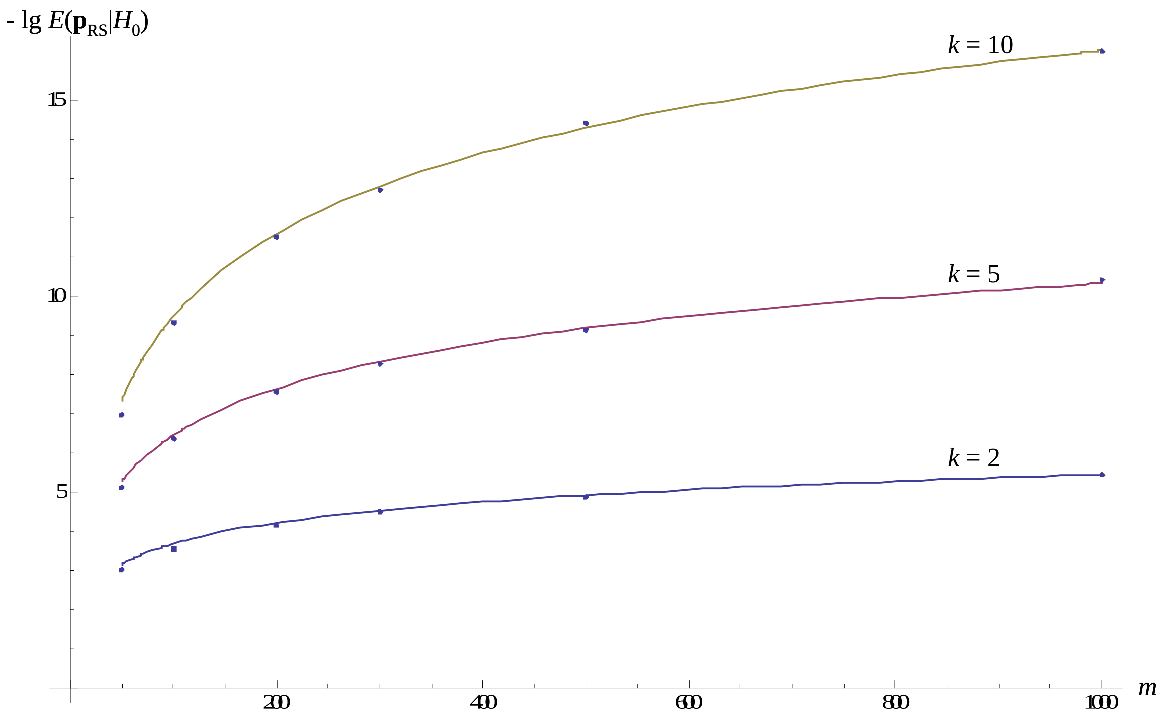

Fig. 3: Expected spurious p-value for the risk score (X_{RS}) comparison using Student's t-test with regard to the total number of independent variables (m) and the number of top of them (k) used to calculate the X_{RS} under H_0 ($n = 200$). Plotted points are the result of numerical simulations (averaged for 1000 random repeats)

Supplement 2

File H0_Model_Selection.xlsx

Global Journals Guidelines Handbook 2021

www.GlobalJournals.org

Memberships

FELLOWS/ASSOCIATES OF SCIENCE FRONTIER RESEARCH COUNCIL

FSFRC/ASFRC MEMBERSHIPS

Introduction

FSFRC/ASFRC is the most prestigious membership of Global Journals accredited by Open Association of Research Society, U.S.A (OARS). The credentials of Fellow and Associate designations signify that the researcher has gained the knowledge of the fundamental and high-level concepts, and is a subject matter expert, proficient in an expertise course covering the professional code of conduct, and follows recognized standards of practice. The credentials are designated only to the researchers, scientists, and professionals that have been selected by a rigorous process by our Editorial Board and Management Board.

> Associates of FSFRC/ASFRC are scientists and researchers from around the world are working on projects/researches that have huge potentials. Members support Global Journals' mission to advance technology for humanity and the profession.

FSFRC

FELLOW OF SCIENCE FRONTIER RESEARCH COUNCIL

FELLOW OF SCIENCE FRONTIER RESEARCH COUNCIL is the most prestigious membership of Global Journals. It is an award and membership granted to individuals that the Open Association of Research Society judges to have made a 'substantial contribution to the improvement of computer science, technology, and electronics engineering.

> The primary objective is to recognize the leaders in research and scientific fields of the current era with a global perspective and to create a channel between them and other researchers for better exposure and knowledge sharing. Members are most eminent scientists, engineers, and technologists from all across the world. Fellows are elected for life through a peer review process on the basis of excellence in the respective domain. There is no limit on the number of new nominations made in any year. Each year, the Open Association of Research Society elect up to 12 new Fellow Members.

Benefit

To the institution

Get letter of appreciation

Global Journals sends a letter of appreciation of author to the Dean or CEO of the University or Company of which author is a part, signed by editor in chief or chief author.

Exclusive Network

Get access to a closed network

A FSFRC member gets access to a closed network of Tier 1 researchers and scientists with direct communication channel through our website. Fellows can reach out to other members or researchers directly. They should also be open to reaching out by other.

Career | Credibility | Exclusive | Reputation

Certificate

Receive a print ed copy of a certificate

Fellows receive a printed copy of a certificate signed by our Chief Author that may be used for academic purposes and a personal recommendation letter to the dean of member's university.

Career | Credibility | Exclusive | Reputation

Designation

Get honored title of membership

Fellows can use the honored title of membership. The “FSFRC” is an honored title which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., FSFRC or William Walldroff, M.S., FSFRC.

Career | Credibility | Exclusive | Reputation

Recognition on the Platform

Better visibility and citation

All the Fellow members of FSFRC get a badge of "Leading Member of Global Journals" on the Research Community that distinguishes them from others. Additionally, the profile is also partially maintained by our team for better visibility and citation. All fellows get a dedicated page on the website with their biography.

Career | Credibility | Reputation

Future Work

Get discounts on the future publications

Fellows receive discounts on future publications with Global Journals up to 60%. Through our recommendation programs, members also receive discounts on publications made with OARS affiliated organizations.

Career Financial

GJ Internal Account

Unlimited forward of Emails

Fellows get secure and fast GJ work emails with unlimited forward of emails that they may use them as their primary email. For example,
john [AT] globaljournals [DOT] org.

Career Credibility Reputation

Premium Tools

Access to all the Premium Tools

To take future researches to the zenith, fellows and associates receive access to all the premium tools that Global Journals have to offer along with the partnership with some of the best marketing leading tools out there.

Financial

Conferences & Events

Organize seminar/conference

Fellows are authorized to organize symposium/seminar/conference on behalf of Global Journal Incorporation (USA). They can also participate in the same organized by another institution as representative of Global Journal. In both the cases, it is mandatory for him to discuss with us and obtain our consent. Additionally, they get free research conferences (and others) alerts.

Career Credibility Financial

Early Invitations

Early invitations to all the symposiums, seminars, conferences

All fellows receive the early invitations to all the symposiums, seminars, conferences and webinars hosted by Global Journals in their subject.

Exclusive

Publishing Articles & Books

Earn 60% of sales proceeds

Fellows can publish articles (limited) without any fees. Also, they can earn up to 60% of sales proceeds from the sale of reference/review books/literature/ publishing of research paper. The FSFRC member can decide its price and we can help in making the right decision.

Exclusive Financial

Reviewers

Get a remuneration of 15% of author fees

Fellow members are eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get a remuneration of 15% of author fees, taken from the author of a respective paper.

Financial

Access to Editorial Board

Become a member of the Editorial Board

Fellows may join as a member of the Editorial Board of Global Journals Incorporation (USA) after successful completion of three years as Fellow and as Peer Reviewer. Additionally, Fellows get a chance to nominate other members for Editorial Board.

Career Credibility Exclusive Reputation

And Much More

Get access to scientific museums and observatories across the globe

All members get access to 5 selected scientific museums and observatories across the globe. All researches published with Global Journals will be kept under deep archival facilities across regions for future protections and disaster recovery. They get 10 GB free secure cloud access for storing research files.

ASFRC

ASSOCIATE OF SCIENCE FRONTIER RESEARCH COUNCIL

ASSOCIATE OF SCIENCE FRONTIER RESEARCH COUNCIL is the membership of Global Journals awarded to individuals that the Open Association of Research Society judges to have made a 'substantial contribution to the improvement of computer science, technology, and electronics engineering.

> The primary objective is to recognize the leaders in research and scientific fields of the current era with a global perspective and to create a channel between them and other researchers for better exposure and knowledge sharing. Members are most eminent scientists, engineers, and technologists from all across the world. Associate membership can later be promoted to Fellow Membership. Associates are elected for life through a peer review process on the basis of excellence in the respective domain. There is no limit on the number of new nominations made in any year. Each year, the Open Association of Research Society elect up to 12 new Associate Members.

Benefit

To the institution

Get letter of appreciation

Global Journals sends a letter of appreciation of author to the Dean or CEO of the University or Company of which author is a part, signed by editor in chief or chief author.

Exclusive Network

Get access to a closed network

A ASFRC member gets access to a closed network of Tier 1 researchers and scientists with direct communication channel through our website. Associates can reach out to other members or researchers directly. They should also be open to reaching out by other.

Career Credibility Exclusive Reputation

Certificate

Receive a print ed copy of a certificate

Associates receive a printed copy of a certificate signed by our Chief Author that may be used for academic purposes and a personal recommendation letter to the dean of member's university.

Career Credibility Exclusive Reputation

Designation

Get honored title of membership

Associates can use the honored title of membership. The “ASFRC” is an honored title which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., ASFRC or William Walldroff, M.S., ASFRC.

Career Credibility Exclusive Reputation

Recognition on the Platform

Better visibility and citation

All the Associate members of ASFRC get a badge of "Leading Member of Global Journals" on the Research Community that distinguishes them from others. Additionally, the profile is also partially maintained by our team for better visibility and citation. All associates get a dedicated page on the website with their biography.

Career Credibility Reputation

Future Work

Get discounts on the future publications

Associates receive discounts on the future publications with Global Journals up to 60%. Through our recommendation programs, members also receive discounts on publications made with OARS affiliated organizations.

Career Financial

GJ Internal Account

Unlimited forward of Emails

Associates get secure and fast GJ work emails with unlimited forward of emails that they may use them as their primary email. For example, john [AT] globaljournals [DOT] org.

Career Credibility Reputation

Premium Tools

Access to all the premium tools

To take future researches to the zenith, fellows receive access to almost all the premium tools that Global Journals have to offer along with the partnership with some of the best marketing leading tools out there.

Financial

Conferences & Events

Organize seminar/conference

Associates are authorized to organize symposium/seminar/conference on behalf of Global Journal Incorporation (USA). They can also participate in the same organized by another institution as representative of Global Journal. In both the cases, it is mandatory for him to discuss with us and obtain our consent. Additionally, they get free research conferences (and others) alerts.

Career Credibility Financial

Early Invitations

Early invitations to all the symposiums, seminars, conferences

All associates receive the early invitations to all the symposiums, seminars, conferences and webinars hosted by Global Journals in their subject.

Exclusive

Publishing Articles & Books

Earn 30-40% of sales proceeds

Associates can publish articles (limited) without any fees. Also, they can earn up to 30-40% of sales proceeds from the sale of reference/review books/literature/publishing of research paper.

Exclusive Financial

Reviewers

Get a remuneration of 15% of author fees

Associate members are eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get a remuneration of 15% of author fees, taken from the author of a respective paper.

Financial

And Much More

Get access to scientific museums and observatories across the globe

All members get access to 2 selected scientific museums and observatories across the globe. All researches published with Global Journals will be kept under deep archival facilities across regions for future protections and disaster recovery. They get 5 GB free secure cloud access for storing research files.

ASSOCIATE	FELLOW	RESEARCH GROUP	BASIC
$4800	$6800	$12500.00	APC
lifetime designation	**lifetime designation**	**organizational**	**per article**
Certificate, LoR and Momento 2 discounted publishing/year **Gradation** of Research 10 research contacts/day 1 **GB** Cloud Storage **GJ** Community Access	**Certificate,** LoR and Momento **Unlimited** discounted publishing/year **Gradation** of Research **Unlimited** research contacts/day 5 **GB** Cloud Storage **Online Presense** Assistance **GJ** Community Access	**Certificates,** LoRs and Momentos **Unlimited** free publishing/year **Gradation** of Research **Unlimited** research contacts/day **Unlimited** Cloud Storage **Online Presense** Assistance **GJ** Community Access	**GJ** Community Access

Preferred Author Guidelines

We accept the manuscript submissions in any standard (generic) format.

We typeset manuscripts using advanced typesetting tools like Adobe In Design, CorelDraw, TeXnicCenter, and TeXStudio. We usually recommend authors submit their research using any standard format they are comfortable with, and let Global Journals do the rest.

Alternatively, you can download our basic template from https://globaljournals.org/Template.zip

Authors should submit their complete paper/article, including text illustrations, graphics, conclusions, artwork, and tables. Authors who are not able to submit manuscript using the form above can email the manuscript department at submit@globaljournals.org or get in touch with chiefeditor@globaljournals.org if they wish to send the abstract before submission.

Before and during Submission

Authors must ensure the information provided during the submission of a paper is authentic. Please go through the following checklist before submitting:

1. Authors must go through the complete author guideline and understand and *agree to Global Journals' ethics and code of conduct,* along with author responsibilities.
2. Authors must accept the privacy policy, terms, and conditions of Global Journals.
3. Ensure corresponding author's email address and postal address are accurate and reachable.
4. Manuscript to be submitted must include keywords, an abstract, a paper title, co-author(s') names and details (email address, name, phone number, and institution), figures and illustrations in vector format including appropriate captions, tables, including titles and footnotes, a conclusion, results, acknowledgments and references.
5. Authors should submit paper in a ZIP archive if any supplementary files are required along with the paper.
6. Proper permissions must be acquired for the use of any copyrighted material.
7. Manuscript submitted *must not have been submitted or published elsewhere* and all authors must be aware of the submission.

Declaration of Conflicts of Interest

It is required for authors to declare all financial, institutional, and personal relationships with other individuals and organizations that could influence (bias) their research.

Policy on Plagiarism

Plagiarism is not acceptable in Global Journals submissions at all.

Plagiarized content will not be considered for publication. We reserve the right to inform authors' institutions about plagiarism detected either before or after publication. If plagiarism is identified, we will follow COPE guidelines:

Authors are solely responsible for all the plagiarism that is found. The author must not fabricate, falsify or plagiarize existing research data. The following, if copied, will be considered plagiarism:

- Words (language)
- Ideas
- Findings
- Writings
- Diagrams
- Graphs
- Illustrations
- Lectures

- Printed material
- Graphic representations
- Computer programs
- Electronic material
- Any other original work

Authorship Policies

Global Journals follows the definition of authorship set up by the Open Association of Research Society, USA. According to its guidelines, authorship criteria must be based on:

1. Substantial contributions to the conception and acquisition of data, analysis, and interpretation of findings.
2. Drafting the paper and revising it critically regarding important academic content.
3. Final approval of the version of the paper to be published.

Changes in Authorship

The corresponding author should mention the name and complete details of all co-authors during submission and in manuscript. We support addition, rearrangement, manipulation, and deletions in authors list till the early view publication of the journal. We expect that corresponding author will notify all co-authors of submission. We follow COPE guidelines for changes in authorship.

Copyright

During submission of the manuscript, the author is confirming an exclusive license agreement with Global Journals which gives Global Journals the authority to reproduce, reuse, and republish authors' research. We also believe in flexible copyright terms where copyright may remain with authors/employers/institutions as well. Contact your editor after acceptance to choose your copyright policy. You may follow this form for copyright transfers.

Appealing Decisions

Unless specified in the notification, the Editorial Board's decision on publication of the paper is final and cannot be appealed before making the major change in the manuscript.

Acknowledgments

Contributors to the research other than authors credited should be mentioned in Acknowledgments. The source of funding for the research can be included. Suppliers of resources may be mentioned along with their addresses.

Declaration of funding sources

Global Journals is in partnership with various universities, laboratories, and other institutions worldwide in the research domain. Authors are requested to disclose their source of funding during every stage of their research, such as making analysis, performing laboratory operations, computing data, and using institutional resources, from writing an article to its submission. This will also help authors to get reimbursements by requesting an open access publication letter from Global Journals and submitting to the respective funding source.

Preparing your Manuscript

Authors can submit papers and articles in an acceptable file format: MS Word (doc, docx), LaTeX (.tex, .zip or .rar including all of your files), Adobe PDF (.pdf), rich text format (.rtf), simple text document (.txt), Open Document Text (.odt), and Apple Pages (.pages). Our professional layout editors will format the entire paper according to our official guidelines. This is one of the highlights of publishing with Global Journals—authors should not be concerned about the formatting of their paper. Global Journals accepts articles and manuscripts in every major language, be it Spanish, Chinese, Japanese, Portuguese, Russian, French, German, Dutch, Italian, Greek, or any other national language, but the title, subtitle, and abstract should be in English. This will facilitate indexing and the pre-peer review process.

The following is the official style and template developed for publication of a research paper. Authors are not required to follow this style during the submission of the paper. It is just for reference purposes.

Manuscript Style Instruction (Optional)

- Microsoft Word Document Setting Instructions.
- Font type of all text should be Swis721 Lt BT.
- Page size: 8.27" x 11'", left margin: 0.65, right margin: 0.65, bottom margin: 0.75.
- Paper title should be in one column of font size 24.
- Author name in font size of 11 in one column.
- Abstract: font size 9 with the word "Abstract" in bold italics.
- Main text: font size 10 with two justified columns.
- Two columns with equal column width of 3.38 and spacing of 0.2.
- First character must be three lines drop-capped.
- The paragraph before spacing of 1 pt and after of 0 pt.
- Line spacing of 1 pt.
- Large images must be in one column.
- The names of first main headings (Heading 1) must be in Roman font, capital letters, and font size of 10.
- The names of second main headings (Heading 2) must not include numbers and must be in italics with a font size of 10.

Structure and Format of Manuscript

The recommended size of an original research paper is under 15,000 words and review papers under 7,000 words. Research articles should be less than 10,000 words. Research papers are usually longer than review papers. Review papers are reports of significant research (typically less than 7,000 words, including tables, figures, and references)

A research paper must include:

a) A title which should be relevant to the theme of the paper.
b) A summary, known as an abstract (less than 150 words), containing the major results and conclusions.
c) Up to 10 keywords that precisely identify the paper's subject, purpose, and focus.
d) An introduction, giving fundamental background objectives.
e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition, sources of information must be given, and numerical methods must be specified by reference.
f) Results which should be presented concisely by well-designed tables and figures.
g) Suitable statistical data should also be given.
h) All data must have been gathered with attention to numerical detail in the planning stage.

Design has been recognized to be essential to experiments for a considerable time, and the editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned unrefereed.

i) Discussion should cover implications and consequences and not just recapitulate the results; conclusions should also be summarized.
j) There should be brief acknowledgments.
k) There ought to be references in the conventional format. Global Journals recommends APA format.

Authors should carefully consider the preparation of papers to ensure that they communicate effectively. Papers are much more likely to be accepted if they are carefully designed and laid out, contain few or no errors, are summarizing, and follow instructions. They will also be published with much fewer delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and suggestions to improve brevity.

Format Structure

It is necessary that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

All manuscripts submitted to Global Journals should include:

Title

The title page must carry an informative title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) where the work was carried out.

Author details

The full postal address of any related author(s) must be specified.

Abstract

The abstract is the foundation of the research paper. It should be clear and concise and must contain the objective of the paper and inferences drawn. It is advised to not include big mathematical equations or complicated jargon.

Many researchers searching for information online will use search engines such as Google, Yahoo or others. By optimizing your paper for search engines, you will amplify the chance of someone finding it. In turn, this will make it more likely to be viewed and cited in further works. Global Journals has compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Keywords

A major lynchpin of research work for the writing of research papers is the keyword search, which one will employ to find both library and internet resources. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining, and indexing.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy: planning of a list of possible keywords and phrases to try.

Choice of the main keywords is the first tool of writing a research paper. Research paper writing is an art. Keyword search should be as strategic as possible.

One should start brainstorming lists of potential keywords before even beginning searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in a research paper?" Then consider synonyms for the important words.

It may take the discovery of only one important paper to steer in the right keyword direction because, in most databases, the keywords under which a research paper is abstracted are listed with the paper.

Numerical Methods

Numerical methods used should be transparent and, where appropriate, supported by references.

Abbreviations

Authors must list all the abbreviations used in the paper at the end of the paper or in a separate table before using them.

Formulas and equations

Authors are advised to submit any mathematical equation using either MathJax, KaTeX, or LaTeX, or in a very high-quality image.

Tables, Figures, and Figure Legends

Tables: Tables should be cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g., Table 4, a self-explanatory caption, and be on a separate sheet. Authors must submit tables in an editable format and not as images. References to these tables (if any) must be mentioned accurately.

Figures

Figures are supposed to be submitted as separate files. Always include a citation in the text for each figure using Arabic numbers, e.g., Fig. 4. Artwork must be submitted online in vector electronic form or by emailing it.

Preparation of Eletronic Figures for Publication

Although low-quality images are sufficient for review purposes, print publication requires high-quality images to prevent the final product being blurred or fuzzy. Submit (possibly by e-mail) EPS (line art) or TIFF (halftone/ photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Avoid using pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings). Please give the data for figures in black and white or submit a Color Work Agreement form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution at final image size ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs): >350 dpi; figures containing both halftone and line images: >650 dpi.

Color charges: Authors are advised to pay the full cost for the reproduction of their color artwork. Hence, please note that if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a Color Work Agreement form before your paper can be published. Also, you can email your editor to remove the color fee after acceptance of the paper.

Tips for Writing a Good Quality Science Frontier Research Paper

Techniques for writing a good quality Science Frontier Research paper:

1. Choosing the topic: In most cases, the topic is selected by the interests of the author, but it can also be suggested by the guides. You can have several topics, and then judge which you are most comfortable with. This may be done by asking several questions of yourself, like "Will I be able to carry out a search in this area? Will I find all necessary resources to accomplish the search? Will I be able to find all information in this field area?" If the answer to this type of question is "yes," then you ought to choose that topic. In most cases, you may have to conduct surveys and visit several places. Also, you might have to do a lot of work to find all the rises and falls of the various data on that subject. Sometimes, detailed information plays a vital role, instead of short information. Evaluators are human: The first thing to remember is that evaluators are also human beings. They are not only meant for rejecting a paper. They are here to evaluate your paper. So present your best aspect.

2. Think like evaluators: If you are in confusion or getting demotivated because your paper may not be accepted by the evaluators, then think, and try to evaluate your paper like an evaluator. Try to understand what an evaluator wants in your research paper, and you will automatically have your answer. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

3. Ask your guides: If you are having any difficulty with your research, then do not hesitate to share your difficulty with your guide (if you have one). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work, then ask your supervisor to help you with an alternative. He or she might also provide you with a list of essential readings.

4. Use of computer is recommended: As you are doing research in the field of science frontier then this point is quite obvious. Use right software: Always use good quality software packages. If you are not capable of judging good software, then you can lose the quality of your paper unknowingly. There are various programs available to help you which you can get through the internet.

5. Use the internet for help: An excellent start for your paper is using Google. It is a wondrous search engine, where you can have your doubts resolved. You may also read some answers for the frequent question of how to write your research paper or find a model research paper. You can download books from the internet. If you have all the required books, place importance on reading, selecting, and analyzing the specified information. Then sketch out your research paper. Use big pictures: You may use encyclopedias like Wikipedia to get pictures with the best resolution. At Global Journals, you should strictly follow here.

6. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right? It is a good habit which helps to not lose your continuity. You should always use bookmarks while searching on the internet also, which will make your search easier.

7. Revise what you wrote: When you write anything, always read it, summarize it, and then finalize it.

8. Make every effort: Make every effort to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in the introduction—what is the need for a particular research paper. Polish your work with good writing skills and always give an evaluator what he wants. Make backups: When you are going to do any important thing like making a research paper, you should always have backup copies of it either on your computer or on paper. This protects you from losing any portion of your important data.

9. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several unnecessary diagrams will degrade the quality of your paper by creating a hodgepodge. So always try to include diagrams which were made by you to improve the readability of your paper. Use of direct quotes: When you do research relevant to literature, history, or current affairs, then use of quotes becomes essential, but if the study is relevant to science, use of quotes is not preferable.

10. Use proper verb tense: Use proper verb tenses in your paper. Use past tense to present those events that have happened. Use present tense to indicate events that are going on. Use future tense to indicate events that will happen in the future. Use of wrong tenses will confuse the evaluator. Avoid sentences that are incomplete.

11. Pick a good study spot: Always try to pick a spot for your research which is quiet. Not every spot is good for studying.

12. Know what you know: Always try to know what you know by making objectives, otherwise you will be confused and unable to achieve your target.

13. Use good grammar: Always use good grammar and words that will have a positive impact on the evaluator; use of good vocabulary does not mean using tough words which the evaluator has to find in a dictionary. Do not fragment sentences. Eliminate one-word sentences. Do not ever use a big word when a smaller one would suffice.

Verbs have to be in agreement with their subjects. In a research paper, do not start sentences with conjunctions or finish them with prepositions. When writing formally, it is advisable to never split an infinitive because someone will (wrongly) complain. Avoid clichés like a disease. Always shun irritating alliteration. Use language which is simple and straightforward. Put together a neat summary.

14. Arrangement of information: Each section of the main body should start with an opening sentence, and there should be a changeover at the end of the section. Give only valid and powerful arguments for your topic. You may also maintain your arguments with records.

15. Never start at the last minute: Always allow enough time for research work. Leaving everything to the last minute will degrade your paper and spoil your work.

16. Multitasking in research is not good: Doing several things at the same time is a bad habit in the case of research activity. Research is an area where everything has a particular time slot. Divide your research work into parts, and do a particular part in a particular time slot.

17. Never copy others' work: Never copy others' work and give it your name because if the evaluator has seen it anywhere, you will be in trouble. Take proper rest and food: No matter how many hours you spend on your research activity, if you are not taking care of your health, then all your efforts will have been in vain. For quality research, take proper rest and food.

18. Go to seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

19. Refresh your mind after intervals: Try to give your mind a rest by listening to soft music or sleeping in intervals. This will also improve your memory. Acquire colleagues: Always try to acquire colleagues. No matter how sharp you are, if you acquire colleagues, they can give you ideas which will be helpful to your research.

20. Think technically: Always think technically. If anything happens, search for its reasons, benefits, and demerits. Think and then print: When you go to print your paper, check that tables are not split, headings are not detached from their descriptions, and page sequence is maintained.

21. Adding unnecessary information: Do not add unnecessary information like "I have used MS Excel to draw graphs." Irrelevant and inappropriate material is superfluous. Foreign terminology and phrases are not apropos. One should never take a broad view. Analogy is like feathers on a snake. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Never oversimplify: When adding material to your research paper, never go for oversimplification; this will definitely irritate the evaluator. Be specific. Never use rhythmic redundancies. Contractions shouldn't be used in a research paper. Comparisons are as terrible as clichés. Give up ampersands, abbreviations, and so on. Remove commas that are not necessary. Parenthetical words should be between brackets or commas. Understatement is always the best way to put forward earth-shaking thoughts. Give a detailed literary review.

22. Report concluded results: Use concluded results. From raw data, filter the results, and then conclude your studies based on measurements and observations taken. An appropriate number of decimal places should be used. Parenthetical remarks are prohibited here. Proofread carefully at the final stage. At the end, give an outline to your arguments. Spot perspectives of further study of the subject. Justify your conclusion at the bottom sufficiently, which will probably include examples.

23. Upon conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print for the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects of your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form which is presented in the guidelines using the template.
- Please note the criteria peer reviewers will use for grading the final paper.

Final points:

One purpose of organizing a research paper is to let people interpret your efforts selectively. The journal requires the following sections, submitted in the order listed, with each section starting on a new page:

The introduction: This will be compiled from reference matter and reflect the design processes or outline of basis that directed you to make a study. As you carry out the process of study, the method and process section will be constructed like that. The results segment will show related statistics in nearly sequential order and direct reviewers to similar intellectual paths throughout the data that you gathered to carry out your study.

The discussion section:

This will provide understanding of the data and projections as to the implications of the results. The use of good quality references throughout the paper will give the effort trustworthiness by representing an alertness to prior workings.

Writing a research paper is not an easy job, no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record-keeping are the only means to make straightforward progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear: Adhere to recommended page limits.

Mistakes to avoid:

- Insertion of a title at the foot of a page with subsequent text on the next page.
- Separating a table, chart, or figure—confine each to a single page.
- Submitting a manuscript with pages out of sequence.
- In every section of your document, use standard writing style, including articles ("a" and "the").
- Keep paying attention to the topic of the paper.
- Use paragraphs to split each significant point (excluding the abstract).
- Align the primary line of each section.
- Present your points in sound order.
- Use present tense to report well-accepted matters.
- Use past tense to describe specific results.
- Do not use familiar wording; don't address the reviewer directly. Don't use slang or superlatives.
- Avoid use of extra pictures—include only those figures essential to presenting results.

Title page:

Choose a revealing title. It should be short and include the name(s) and address(es) of all authors. It should not have acronyms or abbreviations or exceed two printed lines.

Abstract: This summary should be two hundred words or less. It should clearly and briefly explain the key findings reported in the manuscript and must have precise statistics. It should not have acronyms or abbreviations. It should be logical in itself. Do not cite references at this point.

An abstract is a brief, distinct paragraph summary of finished work or work in development. In a minute or less, a reviewer can be taught the foundation behind the study, common approaches to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Use comprehensive sentences, and do not sacrifice readability for brevity; you can maintain it succinctly by phrasing sentences so that they provide more than a lone rationale. The author can at this moment go straight to shortening the outcome. Sum up the study with the subsequent elements in any summary. Try to limit the initial two items to no more than one line each.

Reason for writing the article—theory, overall issue, purpose.

- Fundamental goal.
- To-the-point depiction of the research.
- Consequences, including definite statistics—if the consequences are quantitative in nature, account for this; results of any numerical analysis should be reported. Significant conclusions or questions that emerge from the research.

Approach:

- o Single section and succinct.
- o An outline of the job done is always written in past tense.
- o Concentrate on shortening results—limit background information to a verdict or two.
- o Exact spelling, clarity of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else.

Introduction:

The introduction should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable of comprehending and calculating the purpose of your study without having to refer to other works. The basis for the study should be offered. Give the most important references, but avoid making a comprehensive appraisal of the topic. Describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will give no attention to your results. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here.

The following approach can create a valuable beginning:

- Explain the value (significance) of the study.
- Defend the model—why did you employ this particular system or method? What is its compensation? Remark upon its appropriateness from an abstract point of view as well as pointing out sensible reasons for using it.
- Present a justification. State your particular theory(-ies) or aim(s), and describe the logic that led you to choose them.
- Briefly explain the study's tentative purpose and how it meets the declared objectives.

Approach:

Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done. Sort out your thoughts; manufacture one key point for every section. If you make the four points listed above, you will need at least four paragraphs. Present surrounding information only when it is necessary to support a situation. The reviewer does not desire to read everything you know about a topic. Shape the theory specifically—do not take a broad view.

As always, give awareness to spelling, simplicity, and correctness of sentences and phrases.

Procedures (methods and materials):

This part is supposed to be the easiest to carve if you have good skills. A soundly written procedures segment allows a capable scientist to replicate your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order, but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt to give the least amount of information that would permit another capable scientist to replicate your outcome, but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section.

When a technique is used that has been well-described in another section, mention the specific item describing the way, but draw the basic principle while stating the situation. The purpose is to show all particular resources and broad procedures so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step-by-step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

Materials may be reported in part of a section or else they may be recognized along with your measures.

Methods:

- Report the method and not the particulars of each process that engaged the same methodology.
- Describe the method entirely.
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures.
- Simplify—detail how procedures were completed, not how they were performed on a particular day.
- If well-known procedures were used, account for the procedure by name, possibly with a reference, and that's all.

Approach:

It is embarrassing to use vigorous voice when documenting methods without using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result, when writing up the methods, most authors use third person passive voice.

Use standard style in this and every other part of the paper—avoid familiar lists, and use full sentences.

What to keep away from:

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings—save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part as entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Use statistics and tables, if suitable, to present consequences most efficiently.

You must clearly differentiate material which would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matters should not be submitted at all except if requested by the instructor.

Content:

- Sum up your conclusions in text and demonstrate them, if suitable, with figures and tables.
- In the manuscript, explain each of your consequences, and point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation of an exacting study.
- Explain results of control experiments and give remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or manuscript.

What to stay away from:

- Do not discuss or infer your outcome, report surrounding information, or try to explain anything.
- Do not include raw data or intermediate calculations in a research manuscript.
- Do not present similar data more than once.
- A manuscript should complement any figures or tables, not duplicate information.
- Never confuse figures with tables—there is a difference.

Approach:

As always, use past tense when you submit your results, and put the whole thing in a reasonable order.

Put figures and tables, appropriately numbered, in order at the end of the report.

If you desire, you may place your figures and tables properly within the text of your results section.

Figures and tables:

If you put figures and tables at the end of some details, make certain that they are visibly distinguished from any attached appendix materials, such as raw facts. Whatever the position, each table must be titled, numbered one after the other, and include a heading. All figures and tables must be divided from the text.

Discussion:

The discussion is expected to be the trickiest segment to write. A lot of papers submitted to the journal are discarded based on problems with the discussion. There is no rule for how long an argument should be.

Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implications of the study. The purpose here is to offer an understanding of your results and support all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of results should be fully described.

Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact, you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved the prospect, and let it drop at that. Make a decision as to whether each premise is supported or discarded or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."

Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work.

- You may propose future guidelines, such as how an experiment might be personalized to accomplish a new idea.
- Give details of all of your remarks as much as possible, focusing on mechanisms.
- Make a decision as to whether the tentative design sufficiently addressed the theory and whether or not it was correctly restricted. Try to present substitute explanations if they are sensible alternatives.
- One piece of research will not counter an overall question, so maintain the large picture in mind. Where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

When you refer to information, differentiate data generated by your own studies from other available information. Present work done by specific persons (including you) in past tense.

Describe generally acknowledged facts and main beliefs in present tense.

The Administration Rules

Administration Rules to Be Strictly Followed before Submitting Your Research Paper to Global Journals Inc.

Please read the following rules and regulations carefully before submitting your research paper to Global Journals Inc. to avoid rejection.

Segment draft and final research paper: You have to strictly follow the template of a research paper, failing which your paper may get rejected. You are expected to write each part of the paper wholly on your own. The peer reviewers need to identify your own perspective of the concepts in your own terms. Please do not extract straight from any other source, and do not rephrase someone else's analysis. Do not allow anyone else to proofread your manuscript.

Written material: You may discuss this with your guides and key sources. Do not copy anyone else's paper, even if this is only imitation, otherwise it will be rejected on the grounds of plagiarism, which is illegal. Various methods to avoid plagiarism are strictly applied by us to every paper, and, if found guilty, you may be blacklisted, which could affect your career adversely. To guard yourself and others from possible illegal use, please do not permit anyone to use or even read your paper and file.

Criterion for Grading a Research Paper (Compilation)
by Global Journals

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals.

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

Index

Global Journal of Science Frontier Research

Visit us on the Web at www.GlobalJournals.org | www.JournalofScience.org
or email us at helpdesk@globaljournals.org

ISSN 9755896

9 70116 58698 2 61427>