Statistics and Applications ISSN 2454-7395 (online)
Volume 19, No. 1, 2021 (New Series), pp 209–221

Stratified Subsampling Based *p*-values for Hypothesis Tests in Genomics Research

Sudesh Pundir[1,2*], Yanrong Ji[1*], Arunima Shilpi[1] and Ramana V. Davuluri[3]
[1]*Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.*
[2]*Department of Statistics, Pondicherry University, Pondicherry, India*
[3]*Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA.*

Received: 14 August 2020; Revised: 30 January 2021; Accepted: 04 February 2021

Abstract

Multiple testing, which refers to testing of more than one hypothesis in an experiment, is routinely performed in statistical analysis of genome-wide data, such as testing the association of single-nucleotide polymorphisms (SNPs) with a particular phenotype. A common practice is application of multiple-testing correction methods to exclude candidate SNPs that could otherwise be spuriously marked as statistically significant. However, in many cases such methods are overly conservative and often result in no significant SNPs at all. In this paper, we summarize commonly used multiple-testing correction procedures and Monte Carlo simulation-based methods. We propose a simple modification to subsampling-based simulation method to estimate empirical *p*-values by borrowing the principles of stratified sampling. Using real datasets from the cancer genome atlas (TCGA) data repository, we demonstrate that the traditional multiple testing correction methods yielded almost none or very few significant risks associated SNPs, whereas the proposed stratified subsampling successfully resulted in appropriate number of significant candidate SNPs. We also show that the proposed modification has provided meaningful *p*-values and made the test more powerful as compared to simple subsampling without stratification.

Key words: Multiple comparison test; Subsampling; Stratified sampling; *p*-value.

1. Introduction

With the exponential growth of the omics data, computational analysis of large datasets has become commonplace in the study of human biology and disease. The sampled subjects, on which the data is collected, usually differ by sex, race, age and ethnicity, leading to heterogeneous data. The research presented here is motivated by statistical analyses of such genome-scale data, *e.g.*, The Cancer Genome Consortium Data (Ding *et. al.*, 2018), involving multiple comparisons of thousands of genomic features between heterogeneous populations. While human genome sequences are mostly identical between different individuals, a small number of genetic differences exist that result in the striking phenotypic variation observed among individuals. Studying the association between genetic and phenotypic variation and identification of disease associated genetic variants and their prevalence across different

Corresponding Author: Ramana V. Davuluri;
Email: Ramana.Davuluri@stonybrookmedicine.edu
*Equal contribution

populations have been the subjects of numerous genome projects since the publication of the human genome (Lander *et. al.*, 2001; Landrum *et. al.*, 2018).

The most common genetic variation is single nucleotide polymorphism (SNP), which roughly occur every 1200 base pairs in comparisons of a pair of human chromosomes. For example, dbSNP database provides a general catalog of SNPs that are characterized according to frequency, distribution among populations and functional genomic regions, potential functional consequences, inferred mutation pattern, linkage, and organization within each chromosome in each individual (haplotype) (Sherry *et. al.*, 2001; Neykov *et. al.*, 2019). These SNPs can help discern small differences both within a population and among different populations, leading to the identification of population based risk genetic variants for common and complex diseases.

Motivating Example: Cancer is a complex genetic disease with significant heterogeneity across patients. Molecular understanding of tumor heterogeneity is key to effective cancer treatment and personalized medicine. High-grade serous ovarian carcinoma (HGSOC) accounts for 70 to 80 percent of ovarian cancer deaths, with little improvement in overall survival in recent years (Siegel *et. al.*, 2016). The standard therapy for HGSOC includes maximal cytoreductive surgery followed by platinum and taxane chemotherapy. While the majority of HGSOC patients respond to initial treatment, most tumors recur and become increasingly resistant to chemotherapy, with an overall 5-year survival rate of approximately 30 percent (Reid *et. al.*, 2017). As a heterogeneous disease, understanding how genetic differences in individuals contribute to their cancer susceptibility and response to therapy can help guide medical practitioners to give the best advice to achieve a favorable outcome for the patient. As genome technologies evolve, genotyping of individuals could be available to all patients using a simple saliva test. Large-scale genome-wide association studies and meta analyses have provided powerful insights into SNPs that may be predictive of disease and an individual's length of survival (or response to therapy). For example, The Cancer Genome Atlas (TCGA) data portal (https://portal.gdc.cancer.gov) provides multiple layers of -omics data (*e.g.* gene expression, methylation, SNPs) along with clinical/phenotypic information (*e.g.* cancer stage, survival information, drugs/treatment information) for more than 1500 ovarian cancer patients (Cancer Genome Atlas Research, 2011; GTExProject, 2017). These data provide an unprecedented opportunity for exploratory data analysis to identify SNPs that are associated with cancer, survival status and response therapy. It is expected that the catalogue of such SNPs will provide the foundation for tailored detection, prevention and treatment of diseases leading to the era of personalized cancer medicine (Dayem Ullah *et. al.*, 2018). One common goal in large genome-wide experiments is to identify the genomic markers (*e.g.* genes or SNPs) that are significantly different between different populations or associated with a response or covariate of interest. The response could be censored survival time or other clinical outcomes, the covariates could be either categorical (*e.g.* treatment/control status, cancer subtype) or continuous (*e.g.* dose of a drug).

In the above example of ovarian cancer data, our main goal is to identify the SNPs that are associated with patient survival. Log rank test is the most widely used test for testing the equality of survival distributions between different patient populations. However, a major challenge in the analysis and interpretation of such large-scale genome studies is the simultaneous handling of multiple comparisons, where a large number of genes or SNPs (or null hypotheses) are simultaneously tested. For example, let us suppose that an experiment involves 100 SNPs to be tested, each with a Type 1 error probability of 0.05, assuming the null hypothesis is true for each SNP the expected number of false significant SNPs is equal to 5.

Moreover, if all tests are mutually independent, then the probability that at least one true null hypothesis will be rejected is given by $1–0.95^{100} = 0.994$. Therefore, in any large genome-wide study involving large number of SNPs (usually more than a million), any truly significant calls will be accompanied by correspondingly large number of false findings.

2. Multiple-Testing Correction Methods

Multiple-testing correction methods adjust the significance level for each test to a value α such that the overall type I error for the study (the probability of rejecting a correct null hypothesis in at least one of the tests) will not exceed a predetermined acceptable level, often set to 0.05. Widely accepted approaches to deal with the multiple-testing problem control either the family wise error rate (FWER), which is the probability of at least one false rejection (Hochberg and Tamhane, 1987), or the false discovery rate (FDR), which is the expected proportion of falsely rejected null hypotheses (Reiner *et. al.*, 2003; Benjamini and Yekutieli, 2005).

For example, Bonferroni correction, which controls FWER, for testing the SNPs that are associated with survival, is performed as

(i) Compute *p*-values using log rank test.
(ii) Reject the null hypothesis for $p_l \leq \frac{\alpha}{m}$.

where m is the total number of comparisons/tests we are performing, or total number of hypotheses.

Similarly, Benjamini-Hochberg (BH) correction, which controls FDR, is performed by following step-wise procedure.

(i) Sort the *p*-values in increasing order.
(ii) For a given α, find the largest l such that $p_l \leq \frac{l}{m}\alpha$, where m is again the total number of hypotheses to be tested, and l is the rank of SNPs.
(iii) Reject the null hypotheses for all $H_{(m)}$, m=1, 2, …, l.

Resampling-based multiple-testing correction methods: Resampling-based multiple testing procedures are widely used in genome data analysis, especially when the sample size is small or the distribution of test statistic does not follow normality assumption or is unknown. Resampling-based multiple testing procedures can account for dependent structures among *p*-values or test statistics, resulting in lower type II errors. The commonly used resampling techniques include permutation tests and bootstrap methods.

In permutation tests, the distribution of the test statistics is constructed by calculating all possible values or a sufficiently large number of test statistics (usually 1000 or above) from permuted sampling observations under the null hypothesis. Permutation tests are distribution-free, which can provide exact *p*-values even when sample size is small. Bootstrap method finds an approximate distribution of the test statistic by taking many repeated samples with replacement from one random sample (Efron and Tibshirani, 1994). The bootstrap method provides an asymptotically unbiased estimator for the variance of a sample median and for error rates in a linear discrimination problem outperforming cross-validation (Efron, 1979).

The *p*-values obtained by the bootstrap method are less exact than *p*-values computed from the permutation method, and the bootstrap estimated *p*-values are asymptotically convergent to the true *p*-values (Pollard and van der Laan, 2004). Please refer (Farcomeni, 2008) for a review of multiple hypothesis testing procedures and applications in the analysis of DNA microarray data.

Subsampling-based multiple-testing correction methods: Subsampling procedure is different from resampling technique. While resampling scheme generates multiple samples (of size equal to the original sample size) by choosing the observations from the sample with replacement, subsampling scheme selects the observations from the sample without replacement. Statistical inference based on the samples of fixed size in resampling but in case of subsampling scheme, the inference is drawn on the samples of smaller size than fixed sample size. The samples are drawn in resampling technique by using simple random sampling with replacement (SRSWR), whereas in subsampling the samples are drawn by simple random sampling without replacement (SRSWOR). Subsampling (or Two stage sampling) at few places in the literature should not be confused with subsampling defined by (Politis and Romano, 1993). Technically, while two-stage sampling is a two-stage-sampling scheme, subsampling is resampling method without replacement by selecting a smaller size subsamples from the original sample. For example, (Nigam and Rao, 1996) constructed second order balanced designs when sample size (n) is a composite and prime number, and extended the results to stratified multistage samples and provided inferential procedures on balanced bootstrap for stratified multistage samples.

The distribution of Studentized statistic was estimated by subsampling by (Politis and Romano, 1993). They constructed confidence regions by approximating the sampling distribution of a statistic based on the values of the statistic computed over small subsets of the data, and showed their method works well under weak assumptions (Politis and Romano, 1994). In the subsequent publications, they approximated the sampling distribution of a statistic based on the values of the statistic computed over small subsets of the data, and illustrated its application on time series data (Politis and Romano, 1996). Their book provides some of the foundation for subsampling methodology and related methods (Politis *et. al.*, 1999). Further, the asymptotic theory of subsampling was discussed in (Politis *et. al.*, 2001), and *K*-sample subsampling for iid observations and time series data were discussed by (Politis and Romano, 2008). In a later publication, they constructed the confidence intervals and *p*-values for the tests based on subsampling by shortening the number of iterations (Berg *et. al.*, 2010). They showed that the new *p*-values were asymptotically uniform under the null hypothesis and converged to zero under alternative hypothesis, leading to improved power of the test and meaningful *p*-values.

The application of subsampling methods for assessing the significance of observations in large-scale genome studies was discussed in (Bickel *et. al.*, 2010). Recently, a subsampling without replacement-based normalization scheme was employed for identification of differentially expression that accounted for the hierarchy and amplitude of effect sizes within samples (Mohorianu *et. al.*, 2017). Xavier *et al.* (Xavier *et. al.*, 2017) proposed the use of subsampling bootstrap Markov chain in genomic prediction. The proposed method consists of fitting whole-genome regression models by subsampling observations in each round of a Markov Chain Monte Carlo. Further, the subsampling based approach was effectively used for determining appropriate sequencing depth trough efficient read subsampling of RNA-seq data (Robinson and Storey, 2014).

In this paper, we propose a modification to the subsampling scheme, by performing stratified sampling without replacement. Because of the complex and heterogeneous nature of disease population (cancer patients in the current study), there is a need to account for the heterogeneity; such as race, living status, cancer type *etc*. Using a real example of TCGA ovarian and brain cancer data, we demonstrate that dividing the heterogeneous data into strata and then applying subsampling approach leads to more meaningful empirical *p*-values for log rank test. We also show that traditional multiple testing correction methods seem to be too strict for studies on a genomics scale, whereas the proposed stratified subsampling approach can successfully result in appropriate number of significant observations. In the following section, we begin by introducing the basic principle of stratified subsampling.

3. Stratified Subsampling

In stratified subsampling, instead of drawing a subsample of size $b<<n$, we first partition the sample into non-overlapping groups, and then subsamples without replacement are drawn within each stratum as explained below. Strata are non-overlapping and homogeneous with respect to the characteristic under study. For example, in a survival analysis study based on genome sequencing data from cancer patients, the sample usually consists of both living patients and diseased, usually with varying proportions. If subsamples are drawn without accounting for this heterogeneity, the subsamples may disproportionately consist of one group versus the other, therefore, leading to spurious *p*-values. Here, we propose an approach to statistical significance in the analysis of genome-wide data sets, based on the concept of stratified sub-sampling *p*-values.

Procedure of stratified subsampling:

1. Divide the sample of N units into k strata. Let the i^{th} stratum have n_i, i=1,2, …, k, number of units, such that $N = \sum_{i=1}^{k} n_i$.
2. Draw a subsample of size b_i from sample of size n_i from i^{th} stratum using SRSWOR.
3. All the subsampling units drawn from each stratum will constitute a stratified sample of size b.

Let us define the following symbols as
k: Number of strata
n_i: Numbers of sampling units to be drawn from i^{th} stratum
b_i: Number of subsampling units to be drawn from i^{th} stratum
$n = \sum_{i=1}^{k} n_i$: Total sample size
$b = \sum_{i=1}^{k} b_i$: Total subsample size.

Let $x_n = (X_1, X_2, \ldots, X_n)$ be a sample of n independent and identically distributed (iid) random variables taking values in an arbitrary sample space S with unknown probability distribution P. P belongs to a class of distributions H *w*hich may be parametric, nonparametric or semiparametric. The idea is to approximate the sampling distribution of a statistic based on the values of the statistic computed over smaller subsets of the data.

Let $t(P)$ be the parameter and its estimator (or statistic) is given by

$$t_n = f(X_1,\ X_2,\ \ldots,\ X_n).$$

Then the sampling distribution of the statistic is given by

$$J_n(x,P) = P\{\tau_n[(t_n - t(p)) \le x]\},$$

where τ_n is a normalizing sequence.

The fundamental idea behind subsampling is that $J_n(x,P)$ can be accurately approximated by the normalized distribution of the same estimator calculated on appropriately data chosen subsets of data of size b ($b<<n$).

Let the statistic calculated on the i^{th} subset of size b is denoted by

$$t_{n,b,i} = t_b(X_{i1}, X_{i2}, \ldots, X_{ib}).$$

Let $N_n = \binom{n}{b}$ be the total number of available subsets of the data of size b. In this case, the i^{th} subsample is constructed by sampling without replacement from iid data with purpose of forming a subsample of size b.

The subsampling estimator of $J_n(x,P)$ is defined as follows

$$L_{n,b}(x) = \frac{1}{N_n}\sum_{i=1}^{N_n} I[\tau_b(t_{n,b,i} - t_n) \le x]$$

where I is the indicator function. Under general conditions

$$N_n \to \infty, b \to \infty, \quad \frac{b^k}{n} \to 0$$

and for the appropriate values of k and τ_n is such that $\frac{\tau_b}{\tau_n} \to 0$ whenever $\frac{b}{n} \to 0$. Politis and Romano (1994) showed that

$$L_{n,b}(x) - J_n(x,P) \xrightarrow{P} 0.$$

Let the hypotheses for testing the parameter be

$$H_0: t(P) = \theta_0, \qquad P \in P_0$$
$$H_1: t(P) > \theta_0, \qquad P \in P_1$$

The sampling distribution of the statistic under null hypothesis is given by

$$J_n(x,P_0) = P[\tau_n(t_n - \theta_0) \le x]$$

and its subsampling estimator is given by

$$L_{n,b}(x,P_0) = \tfrac{1}{N_n}\textstyle\sum_{i=1}^{N_n} I[\tau_b(t_{n,b,i} - \theta_0) \le x].$$

Politis *et al.* (1999) gave the proof of the consistency of the test. The test rejects H_0 when

$$\frac{1}{N_n}\sum_{i=1}^{N_n} I[\tau_n(t_n - \theta_0) \ge \tau_b(t_{n,b,i} - \theta_0)] > 1 - \alpha$$

$$\frac{1}{N_n}\sum_{i=1}^{N_n} I[T_n \ge T_{n,b,i}] > 1 - \alpha$$

Under null hypothesis, the subsampled distribution of $T_{n,b,I}$ approximates the sampling distribution of T_n .

Stratified subsampling and Log rank test: Log rank test is the most widely used test for testing the equality of survival distributions. Let

Y: Time until an event occurs where event is death of person
T: Failure time with distribution function $F(x)$ and probability density function $f(x)$
C: Censoring time with distribution function $G(x)$ and probability density function

$$\Delta = min(T, C) = \begin{cases} 1, & T \leq C \\ 0, & T > C \end{cases}$$

Survival function is defined as the probability that a person will survive beyond a time *t*. It is defined as

$$S(t) = P(Y > t) = \int_t^{\infty} f(x)dx = 1 - F(t), \qquad 0 < t < \infty.$$

Consider the following $q \times 2$ table classifying those with and without the event of interest

Group	Event		Total
	Dead at time T_i	Alive at time T_i	
0	D_{0i}	$N_0(T_i) - D_{0i}$	$N_0(T_i)$
1	D_{1i}	$N_1(T_i) - D_{1i}$	$N_1(T_i)$
2	D_{2i}	$N_2(T_i) - D_{2i}$	$N_2(T_i)$
.	.	.	.
.	.	.	.
.	.	.	.
q	D_{qi}	$N_q(T_i) - D_{qi}$	$N_q(T_i)$
Total	D_i	$N(T_i) - D_i$	$N(T_i)$

where $T_1, \ldots, T_i, \ldots, T_k$ are distinct failure times
$N_g(T_i)$: Number of persons in group g at risk at T_i
D_{gi}: Number of persons in group g who fail at T_i, $g = 0,1, 2, \ldots, q$, $i = 1, 2, \ldots, k$.

D_{gi} follows hypergeometric distribution.

The hypotheses for testing the survival functions of different groups are given as

$$H_0: S_0(t) = S_1(t) = \cdots = S_q(t)$$

H_1: Two or more Survival functions are different from others.

The log rank test statistic for testing the above hypotheses is defined as

$$\chi = \frac{\sum_{i=1}^{k}(O_i - E_i)}{\sqrt{\sum_{i=1}^{k} V_i}}$$

where O_i: Observed number of failures, E_i: Expected number of failures, V_i: Variance of observed number of failures. Under H_0, χ (or χ^2) follows standard normal (or chi-square) distribution approximately. This approximation is generally used to obtain an approximate test

for H_0 by comparing the observed value of χ (or χ^2) to the tail area of the standard normal (or chi-square) distribution.

Empirical *p*-values based on Monte Carlo simulations: Monte Carlo simulations are routinely applied in permutation and resampling based methods to estimate the *p*-values. Suppose χ^2_{obs} is the observed Chi-squared test-statistic value for a given random sample from log rank test. In Monte Carlo simulations, independent random datasets are generated using pseudo-random number either by resampling or subsampling methods. Assuming *m* such data sets are simulated under the null hypothesis, each yielding a distinct test statistic χ^2_{sim}, the ideal *p*-value is $p_\infty = P(\chi^2_{sim} > \chi^2_{obj})$. However, p_∞ is unknown, because generating infinite number of datasets is not possible and only a finite number (*m*) of datasets are available. Let *B* be the number of times out of m that $\chi^2_{sim} > \chi^2_{obs}$. It was previously shown that the unbiased estimator $\hat{p}_\infty = B/m$ leads to an invalid test that does not correctly control the type I error rate at the required level (Phipson and Smyth, 2010), therefore, computing the tail probability directly for the Monte Carlo results was suggested as a valid approach. Therefore, in a randomization test, the test statistic is *B* rather than χ^2_{obs}, and the required tail probability is *P(B ≤ b)*. It was shown by (Phipson and Smyth, 2010) that, under the null hypothesis, the marginal distribution of *B* over all possible data sets is discrete uniform on the integers from 0, …, *m*, and the exact Monte Carlo *p*-value is estimated as

$$P_u = P(B \leq b) = \frac{b+1}{m+1}.$$

While this is not an unbiased estimator, the amount of positive bias is just enough to allow for the uncertainty of estimation and to produce a test with the correct size. For further details about this *p*-value calculation, please refer (Edgington and Onghena, 2007; Phipson and Smyth, 2010).

4. Application on real-life datasets

In order to compare our stratified subsampling scheme with other multiple testing correction methods, we have applied our method on two real-life datasets: SNP array data of TCGA ovarian cancer (OV, 570 patients, 580,886 SNPs) and low-grade glioma (LGG, 505 patients, 251,258 SNPs). Each patient has three potential genotypes: AA (reference), Aa (heterozygous) and aa (alternative), for each SNP. We associated their survival functions with the genotypes and used log rank test to determine the statistical significance of the overall survival difference between 3 genotypes. For each genotype, we further stratified on the vital status of the patient, and drew random subsamples with different number of subsampling percentage (60%, 70%, 80%) for n = 500 and 1,000 iterations. We then compared our stratified subsampling scheme with other methods for multiple testing correction, including Bonferroni and Benjamini-Hochberg procedures, Bootstrapping method, as well as subsampling scheme without stratification, by plotting an empirical distribution of χ^2-statistic from log rank test. Specifically, we compared the χ_1^2, …, χ_m^2, …, χ_n^2 with the χ_0^2 obtained using the original un-permuted sample, and computed the empirical *p*-value based on the Monte Carlo empirical p-value formula below (and introduced in previous section):

$$P_{emp} = \frac{r+1}{M+1},$$

where *r* is the total number of iterations that $\chi_m^2 > \chi_0^2$.

Table 1 shows the comparison of number of significant SNPs declared at different thresholds for OV and LGG respectively.

Table 1: Number of significant SNPs for different α

Method	Ovarian (OV)			Brain (LGG)		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Single Sample Logrank test	8,228	35,257	66,855	5,275	18,917	33,645
Multiple comparison-Bonferroni Correction	1	2	2	4	4	8
Multiple comparison-Benjamini Hochberg Correction	2	2	12	4	67	234
Bootstrapping	0	0	0	0	0	0
Stratified Subsampling with size 60% (500 iterations)	2	542	5,943	44	2,164	11,872
Stratified Subsampling with size 70% (500 iterations)	0	46	1,396	4	630	5,298
Stratified Subsampling with size 80% (500 iterations)	0	1	107	0	71	1,022

Table 1 shows that traditional multiple testing correction methods, including Bonferroni and Benjamini-Hochberg procedures, as well as resampling-based method (Bootstrapping), all did not control number of significant findings to an appropriate level as they appear to be too stringent on a genomics scale, where hundreds of thousands of tests are performed simultaneously. This fact is more apparent when we compare across ovarian cancer (total 580,886 SNPs) and glioma (total 251,258 SNPs), where less SNPs results in more significant candidates after multiple testing correction due to the less total number of tests performed in LGG. Compare with the methods above, stratified subsampling provided more candidates at different levels, across the two datasets. Moreover, decrease of subsampling percentage seems to be able to provide additional relaxation, allowing number of candidates to be controlled by adjusting the subsampling parameters.

It is also noteworthy that bootstrapping gives no significant candidates in our case no matter what cutoff we chose. To potentially elucidate why this happens, as well as why a larger subsample size results in smaller number of significant candidates, we plotted the estimated sampling distributions in these cases for the particular SNP with lowest *p*-value from single log rank test (rs10824799 for OV, rs7754576 for LGG), with increased number (10,000) of iterations (Figures 1 and 2).

Table 2 shows that for both OV and LGG random subsampling tends to give less significant candidates as compared to stratified subsampling, indicating that it may again be too strict. Moreover, the random subsampling returns similar number of candidates as Benjamini-Hochberg approach in both cases. Since there is much more computation associated with subsampling approach compared to traditional multiple testing correction methods, applying simple random subsampling does not seem to offer any advantage. We can see from the examples and comparison that stratification can best capture the heterogeneity within the sample while not being too stringent. In this example, the number of strata is 2, with stratification based on living status – dead or living. However, the stratification and number of strata can be modified depending on other attributes, such as, race, ethnicity, sex, *etc.*, provided such information is available and the sample size is large enough to yield desired power.

Table 2: Number of significant SNPs for different α

Subsampling percentage	Ovarian (OV)			Brain (LGG)		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
60% (random)	0	19	666	0	28	438
70% (random)	0	1	75	1	2	72
80% (random)	0	0	2	0	1	2
60% (stratified)	2	542	5,943	44	2,164	11,872
70% (stratified)	0	46	1,396	4	630	5,298
80% (stratified)	0	1	107	0	71	1,022

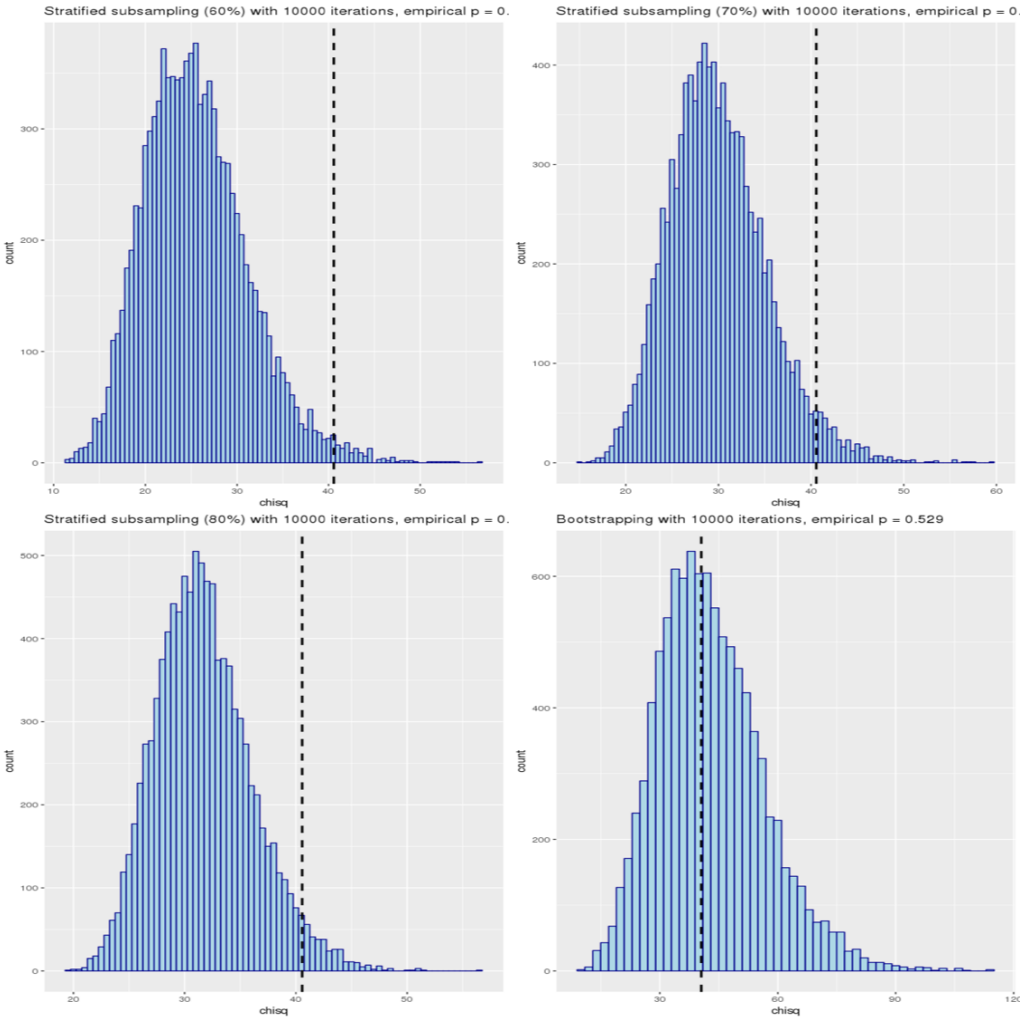

Figure 1: Distribution of simulated test-statistic (χ^2_{sim}) based on stratified subsampling with 60% (top left), 70% (top right), 80% (bottom left) and bootstrapping (bottom right) with 10,000 iterations for OV. Black line indicates the actual test statistic value on the overall sample (χ^2_{obs}).

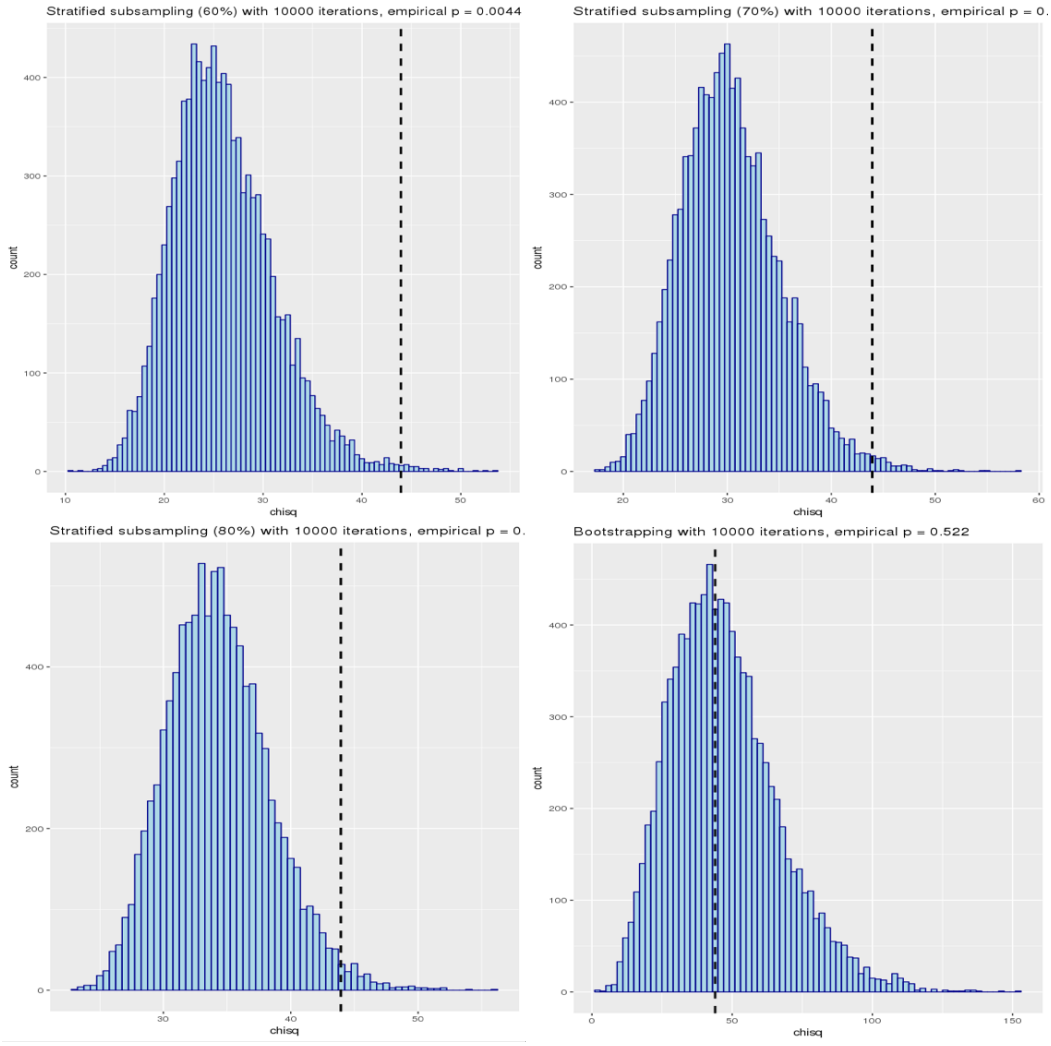

Figure 2: Distribution of simulated test-statistic (χ^2_{sim}) based on stratified subsampling with 60% (top left), 70% (top right), 80% (bottom left) and bootstrapping (bottom right) with 10,000 iterations for LGG. Black line indicates the actual test statistic value on the overall sample (χ^2_{obs}).

5. Conclusions

In this paper, we introduced the concept of stratified subsampling for constructing *p*-values for hypothesis tests in genomics research and showed that it can effectively handle the problem of multiple testing while not being too conservative. While the stratified subsampling based empirical *p*-values are proposed for the log rank test, the method can be generalized for any other statistical test. The proposed modification can be applied in case of heterogeneous data and when subsampling is performed to construct the *p*-values.

Based on the empirical evaluation, we found that the simple random subsampling returned much less significant SNPs than stratified subsampling, suggesting that the simple random subsampling is also too stringent, and considering the computational burden, subsampling based *p*-values (without stratification) do not have advantages over traditional multiple testing correction (*e.g.* BH, Bonferroni), as they similarly returned very few candidates. We are currently working on theoretical aspects of constructing the confidence

intervals and *p*-values based on the stratified subsampling procedure proposed here. In addition, further work is needed to derive and evaluate the asymptotic properties of the proposed test-statistic under the null and alternative hypotheses.

Acknowledgements

This work was supported by the National Library of Medicine of the NIH [R01LM011297 to RD]. We thank the reviewer and Dr. V.. Gupta for their suggestions and thoughtful comments, which substantially helped the revised version.

References

Benjamini, Y. and Yekutieli, D. (2005). Quantitative trait Loci analysis using the false discovery rate. *Genetics*, **171**, 783-790.

Berg, A., McMurry, T. L. and Politis, D. N. (2010). Subsampling *p* values. *Statistics and Probability Letters*, **80**, 1358-1364.

Bickel, P. J., Boley, N., Brown, J. B., Huang, H. and Zhang, N. R. (2010). Subsampling methods for genomic Inference. *The Annals of Applied Statistics*, **4**, 1660–1697.

Cancer Genome Atlas Research, N. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609-615.

Dayem Ullah, A. Z., Oscanoa, J., Wang, J., Nagano, A., Lemoine, N. R. and Chelala, C. (2018). SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Research*, **46**, W109-W113.

Ding, L., Bailey, M. H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D. L., Weerasinghe, A., Huang, K. L., Tokheim, C. *et al*. (2018). Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell*, **173**, 305-320 e310.

Edgington, E. S. and Onghena, P. (2007). *Randomization tests*. Chapman & Hall/CRC, Boca Raton, FL.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **8**, 1-26.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap.* CRC Press, New York.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, **17**, 347-388.

GTExProject. (2017). Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Natural Genetics*, **49**, 1664-1670.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple comparison procedures*. Wiley, New York.

Lander, E. S. Linton, L. M. Birren, B. Nusbaum, C. Zody, M. C. Baldwin, J. Devon, K. Dewar, K. Doyle, M. FitzHugh, W. et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W. *et al*. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, **46**, D1062-D1067.

Mohorianu, I., Bretman, A., Smith, D. T., Fowler, E. K., Dalmay, T. and Chapman, T. (2017). Comparison of alternative approaches for analysing multi-level RNA-seq data. *PLoS One*, **12**, e0182694.

Neykov, M., Lu, J. and Liu, H. (2019). Combinatorial Inference for Graphical Models. *The Annals of Statistics*, **47**, 795-827.

Nigam, A. K. and Rao, J. N. K. (1996). On balanced bootstrap for stratified multistage samples. *Statistica Sinica*, **6**, 199-214.

Phipson, B. and Smyth, G. K. (2010). Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, **9**, Article39.

Politis, D. N. and Romano, J. P. (1993). Estimating the distribution of studentized statistic by subsampling. *Bulletin of the International Statistical Institute*, **49**, 315-316.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, **22**, 2031-2050.

Politis, D. N. and Romano, J. P. (1996). Subsampling for econometrics models- Comments on bootstrapping time series models. *Econometric Review*, **15**,169-176.

Politis, D. N. and Romano, J. P. (2008). K-sample subsampling in general spaces: the case of independent time series. *Journal of Multivariate Analysis*, **101**, 316-326.

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling*. Springer-Verlag, New York.

Politis, D. N., Romano, J. P. and Wolf, M. (2001). On the asymptotic theory of subsampling. *Statistica Sinica*, **11**, 1105-1124.

Pollard, K. S. and van der Laan, M. K. (2004). Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference*, **125**, 85-100.

Reid, B. M., Permuth, J. B. and Sellers, T. A. (2017). Epidemiology of ovarian cancer: a review. *Cancer Biology and Medicine*, **14**, 9-32.

Reiner, A., Yekutieli, D. and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368-375.

Robinson, D. G. and Storey, J. D. (2014). subSeq: determining appropriate sequencing depth through efficient read subsampling. *Bioinformatics*, **30**, 3424-3426.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, **29**, 308-311.

Siegel, R. L., Miller, K. D. and Jemal, A. (2016). Cancer statistics, 2016. *CA - A Cancer Journal for Clinicians*, **66**, 7-30.

Xavier, A., Xu, S., Muir, W. and Rainey, K. M. (2017). Genomic prediction using subsampling. *BMC Bioinformatics*, **18**, 191.