**Meta-analysis of differential expressed genes in melanoma and glioblastoma**

Guangyuan(Frank) Li

**Abstract**

*Background*:

Understanding the molecular mechanisms underpinning tumor development is the focus of intensive research. High-throughput next-generation sequencing enables researchers to systematically measure thousands of critical genes’ expression level and identify oncogenic targets to design better treatment. At present, thousands of tumor patients’ samples have been deep-sequenced in different studies, a holistic and quantitative integration of these sequencing data would be valuable to determine robust up- or -down regulated genes with high confidence and provide a much more stable molecular insight to guide front-line treatment.

*Method*:

We collected 1,039 melanoma patients and 578 gliblastoma patients’ bulk RNA sequencing data from 8 independent studies in The Cancer Genome Atlas (TCGA) database, along with 1,811 healthy skin tissues and 2,651 normal brain tissue sequencing data from Genotype-Tissue Expression(GTEx) consortium. Random-effect meta-analyis was performed upon 17,535 commonly-detected genes in melanoma and 19,493 commonly-detected genes in glioblastoma. We further performed gene enrichment analysis and visualized gene-gene interaction network to better understand the functional implication of up-regulated genes in cancer.

*Result*:

We observed a high degree of heterogeneity among different sequencing studies (I2 > 90% in some cases) and we provided detailed discussion of the possible reason behind that. We found 1,764 significantly up-regulated genes(p < 0.05) in melanoma patients and 1,611 out of them are also significantly up-regulated(p < 0.05) in glioblastoma patients. Gene enrichment analysis suggests the oncogenic roles of these highly-expressed genes in tumors and contributes to tumor progression.

*Contribution:*

Our contributions can be described below, (1) We systematically analyzed currently available sequencing data among melanoma and glioblastoma and provided discussion toward why such a high heterogeneity was observed. (2) We proposed a transferable workflow to analyze every other cancer type. The customized Python3 and R code are freely available from (https://github.com/frankligy/Meta-analysis-TCGA).

**Introduction**

Cancer is the leading cause of death worldwide and is now responsible for two times more death rate than cardiovascular disease in high income countries[[1]](https://paperpile.com/c/nDNKD0/K0Hw). As a malignant disease featured as its high heterogeneity, traditional targeted therapies often suffer from low response rates among a large patient cohort and tumor relapse is nearly inevitable in most situations. In light of its heterogeneous characteristics, precision medicine, especially cancer immunotherapy has emerged as a novel paradigm to fight against tumors. Different from targeted therapy where usually patients are administered small molecules that can target and inhibit a certain oncogenic protein, therefore blocking the succeeding effect, immunotherapy aims to harness our own immune system by reinvigorating deactivated immune cells, then the re-activated immune components can exert their function and kill cancerous cells. Cancer immunotherapy can take varied forms on administration of Interleukin-2(IL-2), a main growth factor for T cells, and tumor vaccine, Autologous T cell therapy (ACT) [[2]](https://paperpile.com/c/nDNKD0/wvz1). It has shown great clinical efficacy in melanoma, non-small cell lung cancer (NSCLC) and colon cancers[2]. While promising, a large amount of cancer patients still do not benefit from immunotherapy. As an example, the most common brain primary malignant glioblastoma, only observing less than 10% response rate for immunotherapy in clinical trials[[3]](https://paperpile.com/c/nDNKD0/RjlZ). Hence, understanding the cause of this drastic difference as to response rate toward immunotherapy is of significant interests. To this end, we attempted to explore the potential molecular differences between two categories of cancer, namely, “good prognosis” cancer -- Melanoma against “poor prognosis” cancer -- glioblastoma.

Meta-analysis has been widely used in evidence-based medicine and serves as a foundational tool for combining statistical power from different studies and mitigating ill-posed effects from individual covariates[[4]](https://paperpile.com/c/nDNKD0/rgob). It has also been adapted in the bioinformatics field to quantitatively integrate measurable effect sizes from various researches, Ibanez et al. utilized meta-analysis proposing a molecular explanation of observed comorbidity between cancer and central nervous system(CNS) disorders[[5]](https://paperpile.com/c/nDNKD0/mD5a). Yarchoan et al. revealed a pan-cancer correlation between tumor neoantigen burden with objective response[[6]](https://paperpile.com/c/nDNKD0/HFIe). Here, we adopted the widely-used random effect model to seek concordant highly-expressed genes between melanoma and glioblastoma.

To obtain a holistic view of gene expression level across different studies and derive a robust inference, we download all TCGA melanoma and glioblastoma tumor tissues with bulk RNA sequencing data available. We then performed a quality control procedure to only retain genes being detected across all studies, selected genes followed by Log2 transformation of their normalized Transcript Per Million (TPM) and Fragment Per Kilobase Million(FPKM) value (Detail in Methods). Hedges’ g estimator was calculated to be the effect size of each gene in an individual study. Random effect model was performed to derive combined effect size of each gene across diverse study settings and identified up-regulated genes in tumor will undergo enrichment analysis (Detail in Methods) and network visualizations.

**Methods**

*Data Collection and Eligibility Criteria*

We downloaded the bulk RNA sequencing data from The Cancer Genome Atlas(TCGA) cBioPortal API (<https://www.cbioportal.org/>). We selected study using metadata information corresponding to (1)melanoma and glioblastoma and with (2)bulk RNA sequencing data stored on the database. To be specific, we collected 8 different studies and their specific information is shown in Table 1.

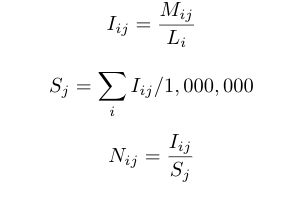
Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| First Author | Study | Number of samples | Reference |
| Liang et al. | Acral melanooma | 36 | [[7]](https://paperpile.com/c/nDNKD0/zwu7) |
| Snyder et al. | Melonoma | 21 | [[8]](https://paperpile.com/c/nDNKD0/dmh4) |
| Van Allen et al. | Metastatic melanoma | 40 | [[8,9]](https://paperpile.com/c/nDNKD0/dmh4+n8eO) |
| TCGA legacy | Skin cutaneous melanoma | 472 | TCGA firehorse |
| Hoadley et al. | Skin cutaneous melanoma | 443 | [[10]](https://paperpile.com/c/nDNKD0/sLWu) |
| Brennan et al. | glioblastoma | 152 | [[11]](https://paperpile.com/c/nDNKD0/FDcL) |
| TCGA legacy | Glioblastoma multiforme | 166 | TCGA firehorse |
| Hoadley et al. | Glioblastoma multiforme | 160 | [[10]](https://paperpile.com/c/nDNKD0/sLWu) |

We then downloaded the healthy control sequencing data from Genotype-Tissue Expression (GTEx) consortium (<https://gtexportal.org/home/datasets>). Latest version VIII and Gene TPMs value matrix was selected. We will explain TPM value in the next paragraph.

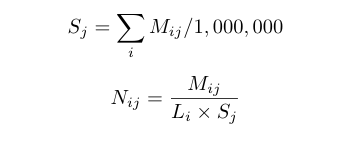
*Data Preprocessing*

The downloaded data from GTEx are in the form of Transcript Per Million(TPM), this metric is the standard for normalizing the length difference of each gene and different sequencing depth across tissue samples. For sample, Gene A of length 10K base pair will have a higher chance to be sequenced than Gene B of length 100 base pair. Likewise, one sample will hold more reads if sequencing depth is relatively higher in this run. We normalize the gene length and sequencing depth effect using TPM:



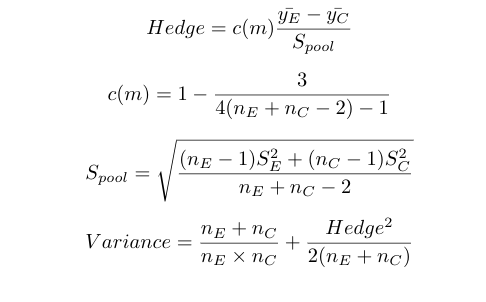
Mij denotes the original gene count matrix where ith row and jth column means the ith gene and jth sample, the value corresponds to the read counts detected in the sequencing process. Sj means the scaling factor that normalizes the sequencing depth per sample. Li signifies the length of gene i. Iij denotes an intermediate matrix after normalizing gene length, Nij denotes the post-transformed TPM matrix.

The downloaded data from TCGA are in the form of Fragment Per Kilobase Per Million (FPKM), this metric is an alternative way to normalize the length difference and sequencing depth but inverse the computation order,



Where Sj, Li, Mij, Nij have the same meaning as TPM specification.

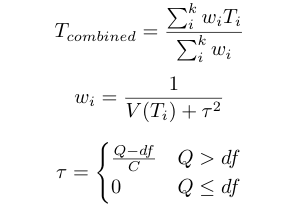
In order to fairly compare TPM value and FPKM value from two different sources and also make their distribution normal-like, we perform log2 transformation upon TPM and FPKM expression matrix. 0 was smoothed to 0.005 to avoid invalid argument in log2 transformation. Further, Hedges’ g estimator was utilized as the recommended metric to represent effect size of gene expression[[12]](https://paperpile.com/c/nDNKD0/zE8z).

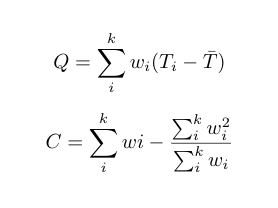


Where nE is the number of studies in the experimental group(TCGA), nC is the number of studies in the control group(GTEx). c(m) is a normalization factor to adjust positive bias. Hedge estimator and Variance will serve as the individual effect size and within-study variance.

*Random Effect Model*

Random Effect model was performed through function “rma” in R package “metafor”, to be specific, the mathematical computation can be summarized as below:





Where wi denotes the total variance for each individual study, including the between-study variance tau. Between-study variance tau is determined by total variance Q and normalization term C. Heterogeneity was represented as I2, I2 > 75% were considered highly heterogeneous. Genes with reported p-value < 0.05 were retained for up-regualted genes in melanoma and glioblastoma.

*Functional genomic analysis*

Gene enrichment analysis was performed using the online interactive tool Enrichr (<https://maayanlab.cloud/Enrichr/>), pathway and gene ontology with statistical significance are kept for further interpretation. Gene-gene interaction visualization was performed on online tool GeneMania (<https://genemania.org/>).

Pathway enrichment analysis is underpinned by a hypergeometric distribution. Given a list of genes X = {x1,x2,x3,x5,...xn} and pre-curated pathway information P = {p1,p2,p3,....pm}, each pathway Pi is composed of dozens of genes that serve as components in this pathway and orchestrate the flow of information. To assess whether a pathway Pi is enriched in a given list of queried genes, we tried to solve, while randomly drawing npi samples from a finite population containing NX genes, the likelihood of obtaining ki genes from our interested pathway Pi. According to the definition of hypergenometric distribution,



Where N is equal to number of genes in the queried list, K denotes number of genes in pathway Pi, n is the number of genes we randomly drew from population N. Lower case i denotes the index of pathways.

**Result**

Up-regulated genes in melanoma

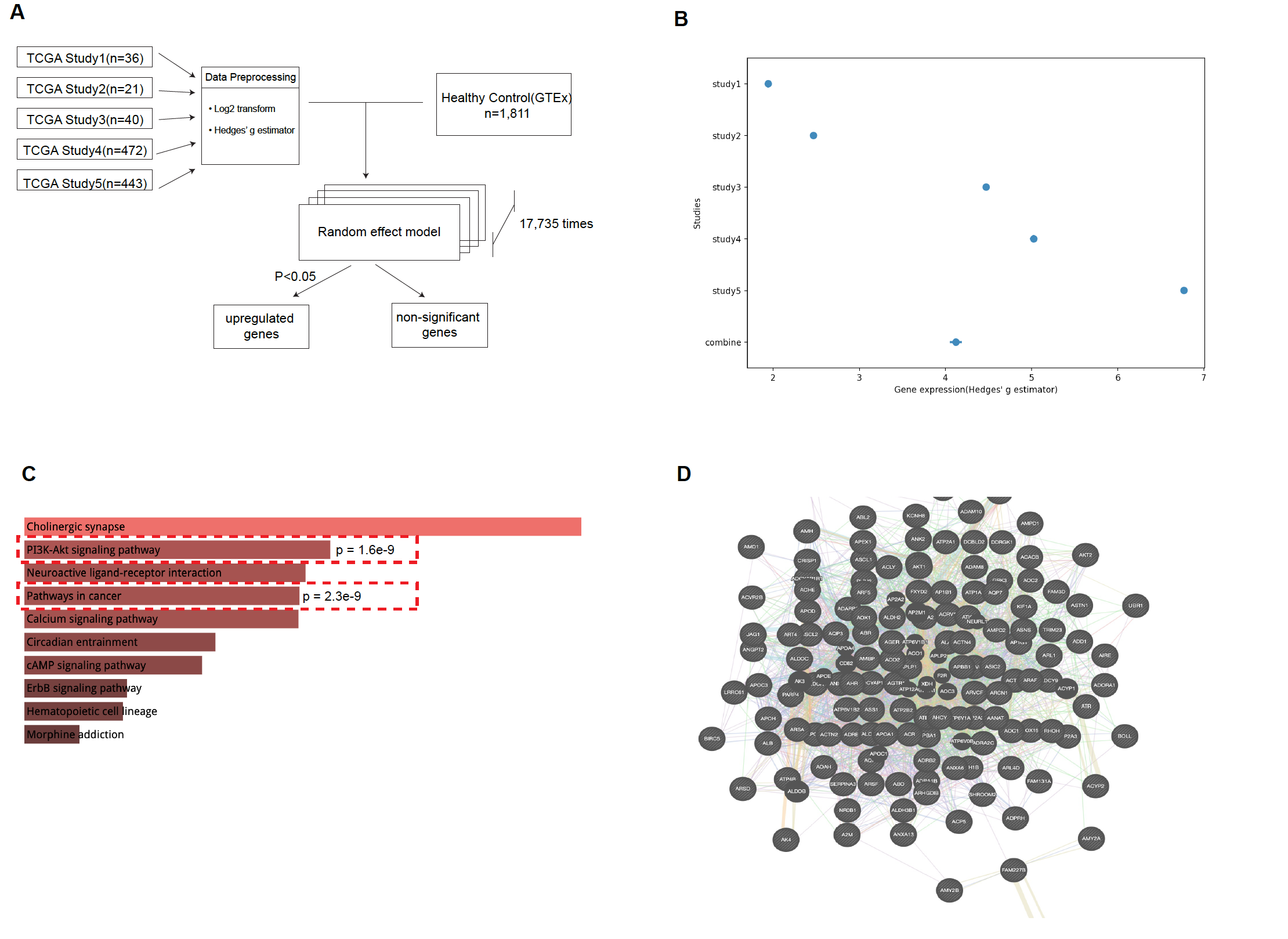


Figure 1 upregulated genes in 5 TCGA melanoma studies. (A) Overall workflow of analysis. (B) Forest plot for a representative gene ACTN2, its individual study effect size was shown in study1-5, combined effect size was shown at the bottom, Filled dot means the average of effect size, the range of flanking horizontal line represents its 95% confidence interval(under normal distribution). © Gene enrichment analysis, highly enriched pathways were selected and relevant ones were marked by their associated p-value. (D) Genemanine gene-gene interaction visualization among up-regulated genes in melanoma.

The overall workflow was shown in Figure1A, details are described in the Method section. In brief, 5 TCGA studies were downloaded and we converted their gene expression level from FPKM to log2 transformed Hedge’s estimator to approximate a normal distribution. Same pre-processing steps were conducted to healthy tissue sequencing data from GTEx. A random effect meta-analysis model was run on the R “metafor” package for each gene (17,535 genes in total). A gene was defined as up-regulated if the reported p value in the random effect model is less than 0.05. In total, 1,764 genes are up-regulated in melanoma compared to its counterpart healthy tissues.

To better understand the between-study variance, we selected one example from 17,535 analyzed genes -- ACTN2. It plays an important role in cytoskeletal proteins and is associated with cardiomyopathy and congenital myopathy disease. We plotted its forest plot as Figure 1B. It is obvious that the within-study variance is relatively small, however, their between-study variation is huge regarding gene expression effect size. The heterogeneity measure I2 is above 99%, indicating a drastic discrepancies among these 5 studies. The possible reasons for the heterogeneity will be discussed in the first section in Discussion. As a result, the 95% confidence interval of combined effect size becomes extended a bit, but the combined effect is by large sitting around the numeric average of each individual hedges’ g estimator.

We further set out to exploit the functional impact of these up-regualted genes and try to find out the relatedness of them to cancer etiology. A pathway enrichment analysis was performed upon all the up-regualted genes in melanoma(details in Method). PI3K-Akt signalling pathway was among the top hit with an associated p-value = 1.6e-9, as shown in Figure 1C, PI3K-Akt pathway is an intracellular pathway that promotes cell proliferation, growth and angiogenesis, dysfunction of PI3K-Akt pathway can lead to altered cell cycle and further result in tumor development[[12,13]](https://paperpile.com/c/nDNKD0/zE8z+KrbW). We visualized a subset of up-regulated genes and their gene-gene interaction relationship, as shown in Figure 1D, over 60% of them hold co-expression pattern and physical interaction evidence, suggesting that these up-regulated genes co-operate to exert shared functions.

Up-regulated genes in glioblastoma

As an example of “poor prognosis” cancer type, we chose glioblastoma, the most common form of primary brain malignancy. The overall workflow is shown in Figure 2A, 3 TCGA independent individual studies were chosen and underwent data preprocessing steps including log2 transformation and hedges’ g computation. Random effect model was performed using R “metafor” packages and reported combined effect and variance were retrieved for each queried gene. Likewise, we showed a forest plot for a representative gene -- ABCA4, as shown in Figure 2B, which is a membrane-associated protein exerting its function via transporting molecules across extracellular membranes. It is again obvious that, the between-study variance is huge as to the measure of gene expression level (I square > 99%), suggesting a highly heterogeneous nature of different study settings that we need to take cautions when interpreting the result.

KEGG enrichment analysis demonstrated that several cancer-related pathway, for example, pathway in cancer (p = 7.5e-92), MAPK pathway (p = 1.04e-55), PI3k-Akt pathway (p = 1.8e-52) and proteoglycan in cancer (p = 7.35e-49) are among the top hits. Uncontrolled MAPK pathway is a necessary step for the development of cancer and proteoglycan also plays an important role in cell-cell communication and prevents cancerous cells from developing contact inhibition. Gene-gene network analysis again revealed the cooperative nature of these up-regulated genes in glioblastoma compared to normal brain tissue.

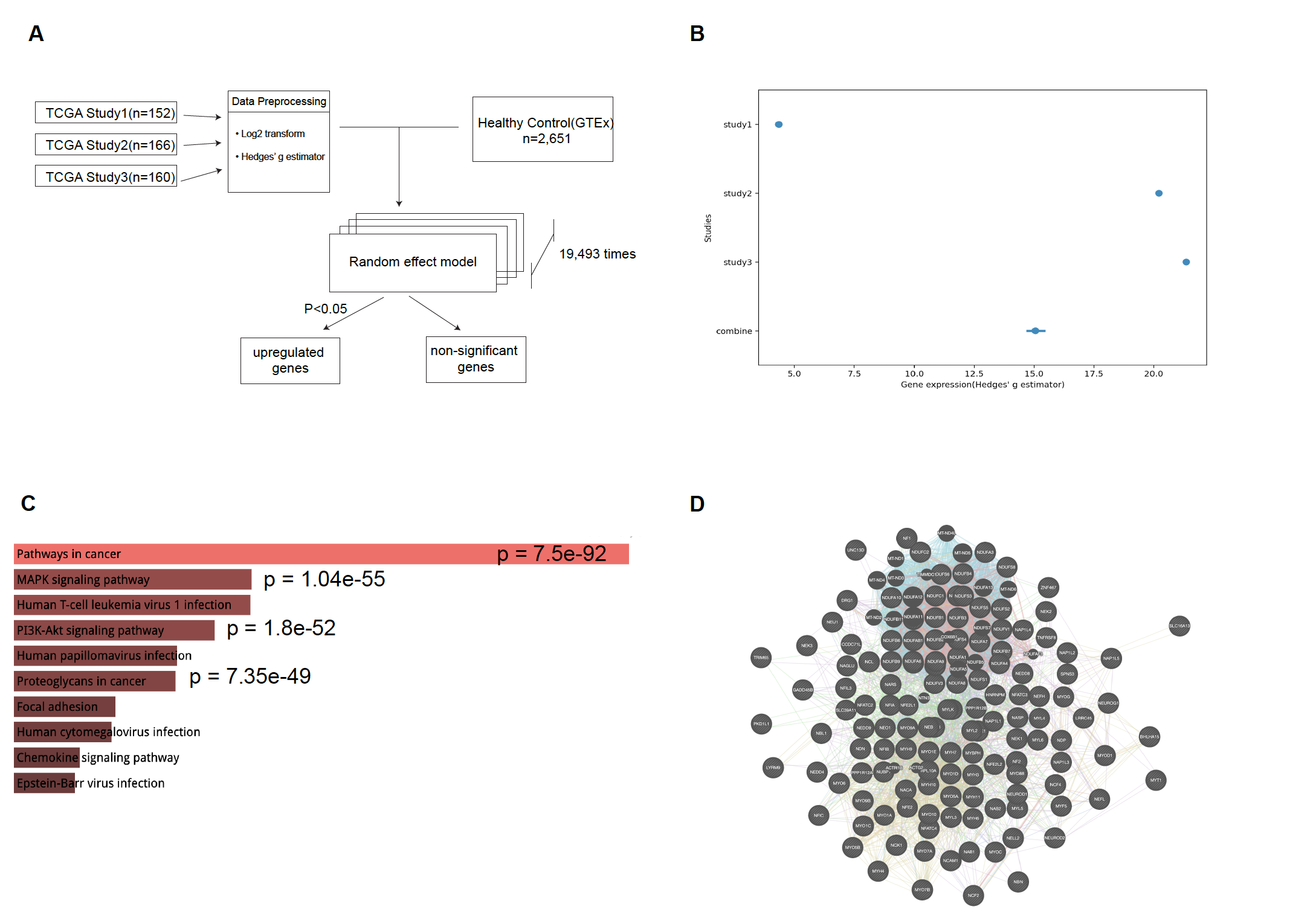


Figure 2 upregulated gene in glioblastoma. (A) Overall workflow of glioblastoma analysis. (B) the forest plot of a representative gene -- ABCA4, its individual expression level are shown in each study and the combined effects are shown at the bottom. Horizontal line signifies 95% confidence interval. © Pathway enrichment analysis and associated p-value. (D) GeneMania network visualization of gene-gene interaction and co-expression pattern.

Shared up-regulated genes across melanoma and glioblastoma

Next, we sought to explore if there exists any genes that could explain the drastic outcome response toward immunotherapy between melanoma(“good prognosis”) and glioblastoma(“poor prognosis”). Among all the up-regulated genes in melanoma and glioblastoma, respectively, we identified 1,611 shared up-reguated genes between these two types of cancers. In order to ascertain their relative expression level, we plotted their combined effects from a random effect model(REM), as shown in Figure 3. We circled out two areas where the genes falling into these two areas are exclusively high-expressed in one cancer versus the other, These candidate genes can be of great interest since they may account for the clinical differences between their response to immunotherapy.

Specifically, TLX2, a member of the homeobox-containing transcription factor family, is highly-expressed in melanoma, as shown in lower right corner in Figure 3, might regulate T cell activity and hence maintain the active state of the immune system. Conversely, EDARADD, a gene that occurs in the upper left corner, holds a high expression level in glioblastoma but not in melanoma. Although the causal relationship between EDARADD and poor response is still elusive, the approaches could shed light on the potential molecular targets worth intensive research in the future.

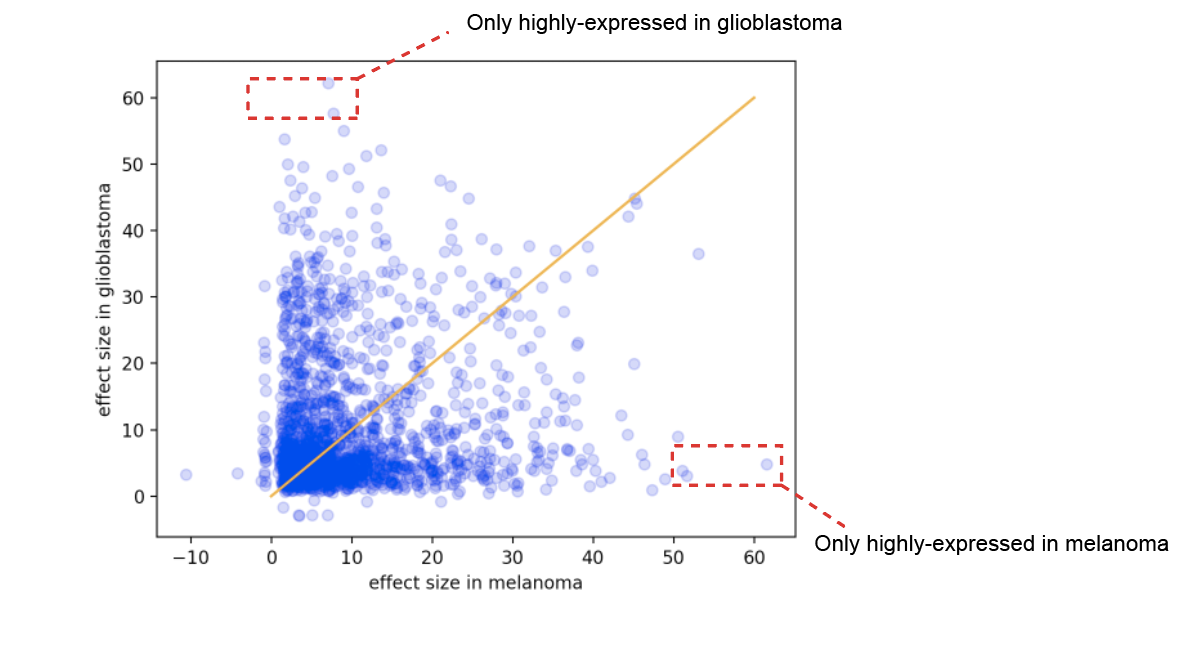


Figure 3 Correlated and anti-correlated expressed genes between glioblastoma and melanoma

**Discussion**

High heterogeneity in gene-centric meta-analysis in bioinformatics (Risk of bias)

Meta analysis has been adopted in evidence-based medicine for a long time and has shown its promising power to assess between-study variance and strengthen the estimate from multiple small-sample-size studies. Conventionally, the measure of heterogeneity I2 is in the range of [0,100%]. An I2 value greater than 50%-75% is considered large[[4]](https://paperpile.com/c/nDNKD0/rgob), under which specific care should be taken when interpreting the result since it indicates true effect size of variable of interest would vary a lot among selected studies. Meta-analysis in bioinformatics is theoretically similar but is different in many ways practically, which could contribute to the high heterogeneity observed in this study.

First, different from clinical trials where several experiments aiming to address the same question would be performed in a multi-centre fashion. It is not common in the field of bioinformatics to repeatedly conduct any sequencing experiment if the same questions have been previously proposed and the raw data are already available in the public repository. While possible to reproduce others' conclusions, the high demand of tumor tissue and the costly sequencing experiment makes exactly identical sequencing data aiming to answer the same questions very rarely happen. When systematically assessing the transcriptomic diversity and dynamic changes, researchers tend to knock-out(remove) or knock-down(decrese) certain critical genes, therefore, the changes reflected in transcriptome sequencing can be a direct evidence to prove its function and can corroborate their initial hypothesis. However, the introduction of this perturbation can lead to very different biological context, and makes the meta-analysis based on them become highly heterogeneous. Likewise, at present, a popular research topic would be the effect of drugs and delineate the mechanisms of drug mechanisms, the addition of drug effect can greatly obscure the original condition of tissues and make the resultant meta-analysis integration become unstable.

Second, cancer is a disease featured as its high level of heterogeneity, it even manifests even within the same patients where gene expression profile could differ a lot between tumor lesion and adjacent tissue. It is unrealistic and shouldn’t be expected to obtain tissues with the same transcriptomic profiles. Furthermore, tumor stage may largely impact its internal gene expression level. Hence, abovementioned factors makes meta-analysis conducted upon tumor samples become challenging. Although meta-regression can be used to regress out the effect of covariates, it is not easy to pinpoint correct covariates in each study. Incorrect assignment of covariate weight could lead to the loss of genuine information, therefore making the analysis less tractable, if not impossible. Finally, the privacy of information imposes another layer of barrier to easily assess all possible covariates across different studies.

Third, RNA sequencing, along with other high-throughput techniques, including but not limited to DNA sequencing, Mass Spectrometry, etc, all suffer from batch effect issue by which here meaning same samples prepared by different vendors or sequenced by different sequencers may result in different joint distributions. Dedicated batch effect removal software has been developed and may be useful in the field of meta-analysis.

Taken together, the observed high heterogeneity between studies is not desirable but it can be attributed to various aspects ranging from sample preparation, tumor heterogeneity and the complex interactions between different covarites. Meta-analysis is not able to resolve all of the issues by the method itself and we need to take caution when attempting to interpret the results with varied effect sizes observed.

Comments on meta-analysis

Simply speaking, meta-analysis, especially the Fixed Effect Model (FEM) is essentially a weighted average of each individual study. However, the usage of meta-analysis goes far beyond its simple use case but in general serves as a way of scientific thinking to (1) understand the variation of your measurement, (2) Ascertain the statistical power of each individual study of small sample size. In addition to the FEM and REM method for combining effect size from diverse studies, we can also integrate several reported significance levels (i.e. p-value) from each single study. Fisher’s method has been used in lots of bioinformatics settings including deriving differentially expressed genes from its descendent isoform significance levels.

Zhang et al. described a probabilistic approach[[14]](https://paperpile.com/c/nDNKD0/burh) to infer mutation burden and utilized Fisher’s method to combine p-value from each individual to conclude the disease-level significance. In the era of increasingly exploded data, it is inevitable that multiple sources of studies would be conducted several times to assess the same concept. In this context, mata-analysis would be extremely useful for quantitatively measuring their statistical power, taking their covariates, sample size and publication bias into consideration.

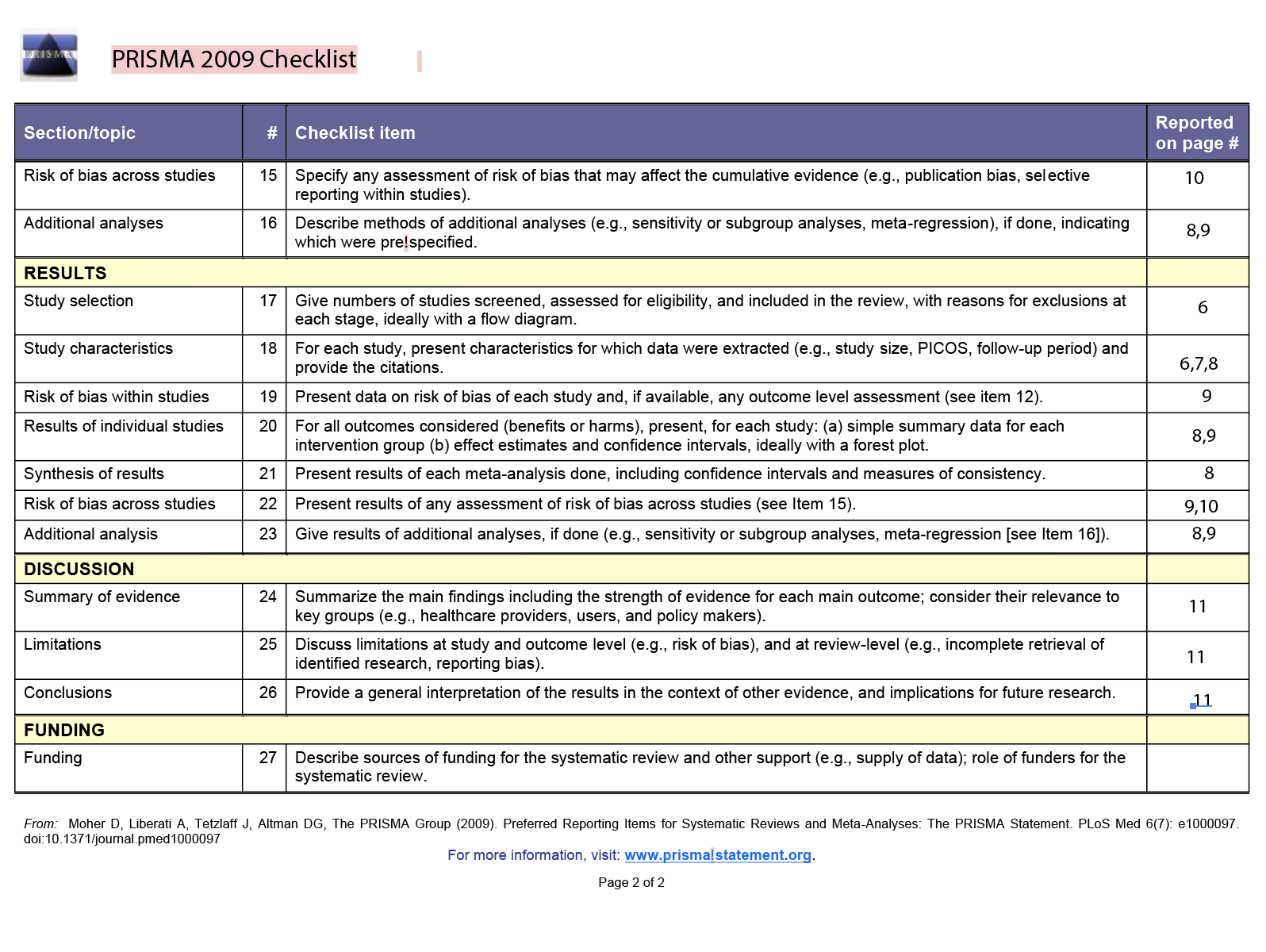
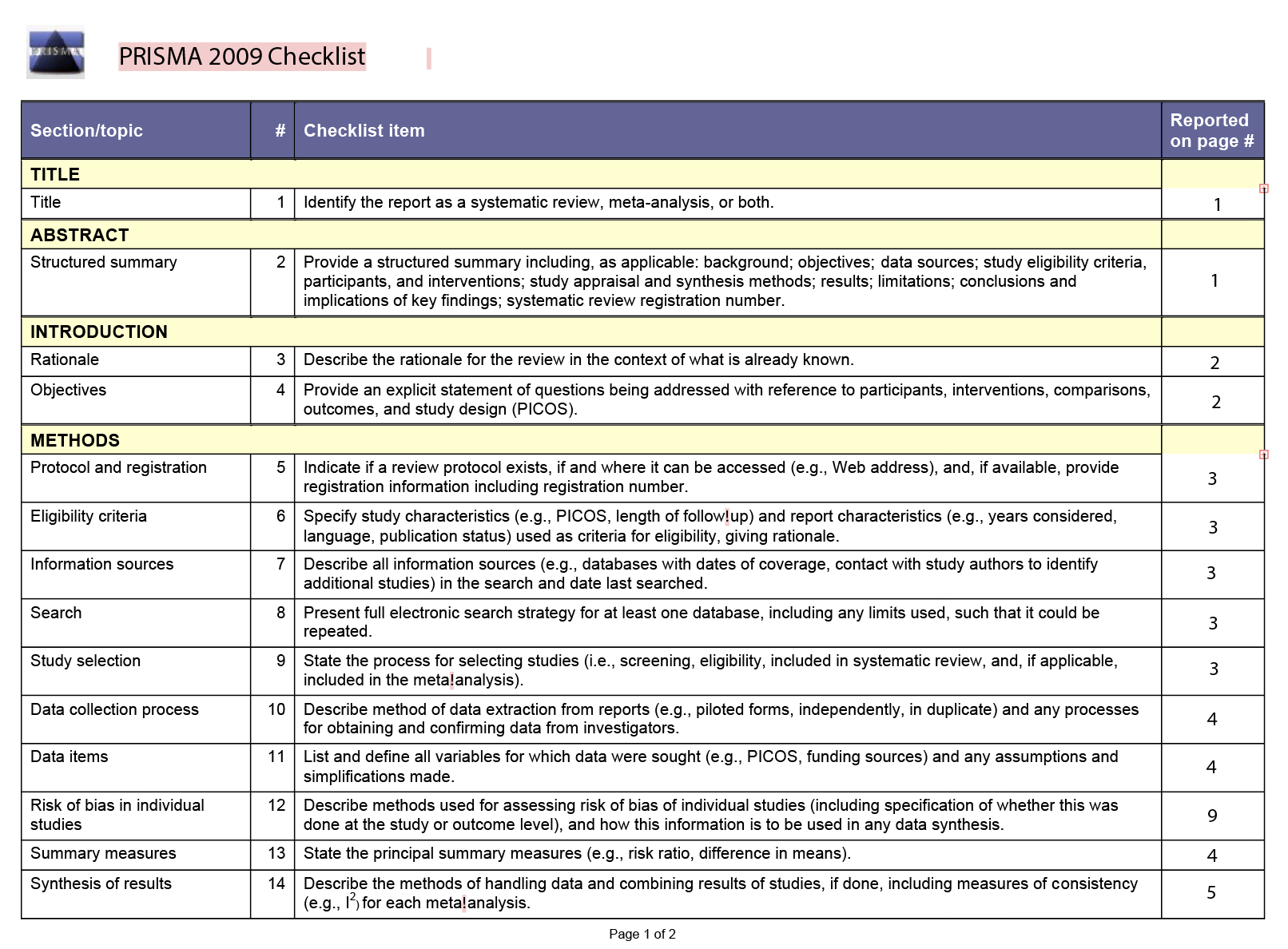
Despite the promising power meta-analysis could offer, there are also several caveats that we might want to think carefully and make sure the assumption of each model has been met. For instance, in gene-centric meta-analysis, we assume the effect size will follow a normal distribution, however, the commonly-used gene expression measurement(FPKM, TPM, raw count) are usually non-negative, which would make the data distribution to be skewed and put all downstream analysis in risk. Under this scenario, log transformation should be a desirable way to mitigate the effect and approximate a Gaussian-like distribution. Next, we should look very carefully at the method each individual study used and rationally gauge how it will impact the way we integrate it at the following step. In this study, some of TCGA data is represented in TPM but some of them are in FPKM format, ideally it is not recommended to directly compare them but the accessibility issue prevented us from obtaining raw count. Hence, we utilized log2 transformation and compute Hedges’ g estimator to alleviate the effect resulting from metric differences.

The major critics of meta-analysis lie on the mis-interpretation of combined results, it is worsened by the factor that meta-analysis is relatively easy to implement with the aid of a plethora of off-the-shelf R packages and online web tools. Meta-analysis itself is just a method, which allows us to explore and derive more robust conclusions from large amounts of available data, however, the misuse of meta-analysis and especially, interpreting the result without taking any cautions could be quite dangerous and often lead us to the opposite direction. Under any circumstances, it is crucial to carefully read and figure out the scope, aim of individual studies and justify why we can use a certain meta-analysis model compared to others.

Conclusions and Limitations

In this study, we attempted to utilize Random Effect Model(REM) to identify differentially expressed genes across two representative cancer types, melanoma and glioblastoma, corresponding to good and poor prognosis respectively toward cancer immunotherapy , seeking to shed light on the transcriptomic differences contributing to different clinical outcomes . Our work serves as a transferable workflow to explore any gene-centric meta-analysis questions and we provided detailed explanation of the observed heterogeneity between studies. Despite the advantages, our study can be better extended to include, (1) meta-regression model to further eliminate potential covariates across studies. (2) performing individual level meta-analysis to allow finer control over different combinations of moderators.

**Prisma checklist**



**Reference**

1. [Mahase E. Cancer overtakes CVD to become leading cause of death in high income countries. BMJ. 2019;366: l5368.](http://paperpile.com/b/nDNKD0/K0Hw)

2. [Leko V, Rosenberg SA. Identifying and Targeting Human Tumor Antigens for T Cell-Based Immunotherapy of Solid Tumors. Cancer Cell. 2020. pp. 454–472. doi:](http://paperpile.com/b/nDNKD0/wvz1)[10.1016/j.ccell.2020.07.013](http://dx.doi.org/10.1016/j.ccell.2020.07.013)

3. [Zhao J, Chen AX, Gartrell RD, Silverman AM, Aparicio L, Chu T, et al. Immune and genomic correlates of response to anti-PD-1 immunotherapy in glioblastoma. Nat Med. 2019;25: 462–469.](http://paperpile.com/b/nDNKD0/RjlZ)

4. [Serghiou S, Goodman SN. Random-Effects Meta-analysis: Summarizing Evidence With Caveats. JAMA. 2019;321: 301–302.](http://paperpile.com/b/nDNKD0/rgob)

5. [Ibáñez K, Boullosa C, Tabarés-Seisdedos R, Baudot A, Valencia A. Molecular evidence for the inverse comorbidity between central nervous system disorders and cancers detected by transcriptomic meta-analyses. PLoS Genet. 2014;10: e1004173.](http://paperpile.com/b/nDNKD0/mD5a)

6. [Yarchoan M, Hopkins A, Jaffee EM. Tumor Mutational Burden and Response Rate to PD-1 Inhibition. N Engl J Med. 2017;377: 2500–2501.](http://paperpile.com/b/nDNKD0/HFIe)

7. [Liang WS, Hendricks W, Kiefer J, Schmidt J, Sekar S, Carpten J, et al. Integrated genomic analyses reveal frequent aberrations in acral melanoma. Genome Res. 2017;27: 524–532.](http://paperpile.com/b/nDNKD0/zwu7)

8. [Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. N Engl J Med. 2014;371: 2189–2199.](http://paperpile.com/b/nDNKD0/dmh4)

9. [Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science. 2015;350: 207–211.](http://paperpile.com/b/nDNKD0/n8eO)

10. [Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018;173: 291–304.e6.](http://paperpile.com/b/nDNKD0/sLWu)

11. [Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma. Cell. 2013;155: 462–477.](http://paperpile.com/b/nDNKD0/FDcL)

12. [Toro-Domínguez D, Villatoro-García JA, Martorell-Marugán J, Román-Montoya Y, Alarcón-Riquelme ME, Carmona-Sáez P. A survey of gene expression meta-analysis: methods and applications. Brief Bioinform. 2020. doi:](http://paperpile.com/b/nDNKD0/zE8z)[10.1093/bib/bbaa019](http://dx.doi.org/10.1093/bib/bbaa019)

13. [Brazil DP, Hemmings BA. Ten years of protein kinase B signalling: a hard Akt to follow. Trends in Biochemical Sciences. 2001. pp. 657–664. doi:](http://paperpile.com/b/nDNKD0/KrbW)[10.1016/s0968-0004(01)01958-2](http://dx.doi.org/10.1016/s0968-0004(01)01958-2)

14. [Zhang J, Liu J, McGillivray P, Yi C, Lochovsky L, Lee D, et al. NIMBus: a negative binomial regression based Integrative Method for mutation Burden Analysis. BMC Bioinformatics. 2020;21: 474.](http://paperpile.com/b/nDNKD0/burh)