BioTechniques

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ibtn20

A Simple *in Silico* Approach to Generate Gene-Expression Profiles from Subsets of Cancer Genomics Data

Mohammed Khurshed, Remco J Molenaar & Cornelis JF van Noorden

To cite this article: Mohammed Khurshed, Remco J Molenaar & Cornelis JF van Noorden (2019) A Simple *in Silico* Approach to Generate Gene-Expression Profiles from Subsets of Cancer Genomics Data, BioTechniques, 67:4, 172-176, DOI: 10.2144/btn-2018-0179

To link to this article: https://doi.org/10.2144/btn-2018-0179

Reports

BioTechniques

For reprint orders, please contact: reprints@futuremedicine.com

A simple *in silico* approach to generate gene-expression profiles from subsets of cancer genomics data

Mohammed Khurshed*[,1,2], Remco J Molenaar[2] & Cornelis JF van Noorden[1,3]

ABSTRACT

In biomedical research, large-scale profiling of gene expression has become routine and offers a valuable means to evaluate changes in onset and progression of diseases, in particular cancer. An overwhelming amount of cancer genomics data has become publicly available, and the complexity of these data makes it a challenge to perform *in silico* data exploration, integration and analysis, in particular for scientists lacking a background in computational programming or informatics. Many web interface tools make these large datasets accessible but are limited to process large datasets. To accelerate the translation of genomic data into new insights, we provide a simple method to explore and select data from cancer genomic datasets to generate gene expression profiles of subsets that are of specific genetic, biological or clinical interest.

METHOD SUMMARY

A simple *in silico* method to import and integrate subsets of data samples with specific genomic, biological and/or clinical interest in order to generate gene expression profiles and crosslink these profiles with DNA methylation and protein expression, which can be integrated to research hypotheses for specific subtypes of cancer.

KEYWORDS

cancer genomics • cBioPortal • data mining • epigenetics • gene expression • *in silico*

*[1]Cancer Center Amsterdam, Department of Medical Biology, Amsterdam UMC at the Academic Medical Center, Amsterdam, The Netherlands; [2]Cancer Center Amsterdam, Department of Medical Oncology, Amsterdam UMC at the Academic Medical Center, Amsterdam, The Netherlands; [3]Department of Genetic Toxicology & Cancer Biology, National Institute of Biology, Ljubljana, Slovenia; *Author for correspondence: m.khurshed@amsterdamumc.nl*

BioTechniques 67: 172-176 (October 2019) 10.2144/btn-2018-0179

In the past decade, advances in genome technologies have enabled the identification of molecular mechanisms of biological processes and diseases, impacting all areas of clinical research, cancer in particular. Intratumoral heterogeneity, dynamic changes in the genome of cancer cells and genetic aberrations are unique fingerprints for each type of cancer [1]. These features of cancer, in combination with prognostic subtype classifications and risk stratification, have revealed that gene expression profiling allows for a better understanding of molecular backgrounds of, for example, prognosis and therapy sensitivity in cancer. Moreover, gene-expression profiling is a powerful molecular approach to predict drug sensitivity [2,3].

In order to generate catalogs of genomic alterations in different cancer types, coordinated large-scale cancer genomic projects are being developed. The two main projects are the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) [4], including many centers utilizing different platforms to provide cancer genomics information such as gene expression, DNA mutations, DNA methylation, protein expression and clinical data. These projects provide large amounts of genomic data to assist researchers in generating or testing novel hypotheses that may ultimately aid in the development of novel cancer therapies, diagnostic methods and preventive strategies [5]. However, exploration, integration and analysis of the large amounts of complicated data is challenging, especially for scientists lacking a background in computational programming or informatics.

The effective use of the large amounts of cancer genome data remains a challenge due to the limitations of computational methodologies and insufficient guidance. Data visualization is very helpful for efficient data analysis and advanced tools have been developed to facilitate data visualization, such as the open-access portals cBioPortal, UCSC Cancer Browser and canEvolve (Table 1). However, open-access portals mainly facilitate investigations of large datasets and are sometimes limited when exploring the datasets in more depth. Here, we describe a simple but effective method to investigate subsets of samples or patients with a specific genetic, biological or clinical interest. We focus on profiling of gene expression and present a method for the analysis of gene expression data in relation to DNA methylation and protein expression (Table 2), which can be integrated to test research hypotheses for specific types of cancer.

MATERIALS & METHODS

Protocol for *in silico* gene expression profiling

Gene-expression profiling is a powerful technique for studying biological processes at the molecular level. Gene activity, or expression, can be assessed by protein identification but gene expression is usually investigated by examining the RNA message or transcript. Two high-throughput methods that are commonly used for comprehensive gene-expression profiling are RNA sequencing with next-generation sequencing (NGS) and DNA microarrays [6].

In general terms, there are two types of gene-expression approaches in cancer: the differential and the relative analysis. In the differential approach, tumor-expression profiles relative to the patient-matched or unmatched normal tissue samples are elucidated, whereas the relative approach compares transcript levels across tumor types or cell and tissue samples. Depending on the specific approach, gene-expression profiling of samples and specimens ▶

Table 1. Overview of open-access portals with analytical tools for visualization of cancer genomics data.

Advanced visualization tool	Source
The cBioPortal for Cancer Genomics	http://cbioportal.org/
UCSC Xena	https://xenabrowser.net/
Genomic Data Commons Data Portal	https://portal.gdc.cancer.gov/
Cancer Genome Workbench (CGWB)	https://cgwb.nci.nih.gov/
CanEvolve	https://canevolve.org/
The Broad GDAC Firehose	https://gdac.broadinstitute.org/

Table 2. Overview of different cancer genomics data and type for profiling.

Genomic type	Data
Gene expression	RNA-seqTumor RNA (microarray)
DNA methylation	Methylation (HM27)
Protein expression	Reverse-phase protein array (RPPA)

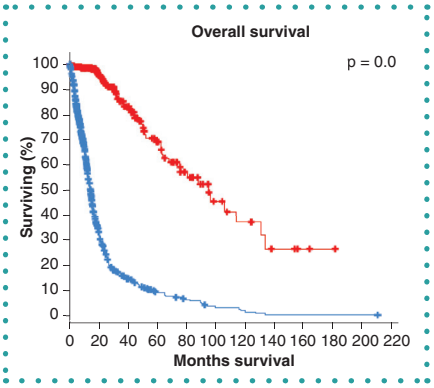

Figure 1. Representative analysis of overall survival curves comparing *IDH1*^{MUT} and *IDH1*^{WT} glioma patients in the TCGA database. For analysis, the merged cohort of low-grade glioma and glioblastoma multiforme (TCGA, Cell 2016) study was analyzed, including 411 *IDH1*^{MUT} versus 401 *IDH1*^{WT} glioma patients. Overall survival Kaplan–Meier plot shows approximately sixfold prolonged survival of *IDH1*^{MUT} glioma patients (red) compared to *IDH1*^{WT} glioma patients (blue).

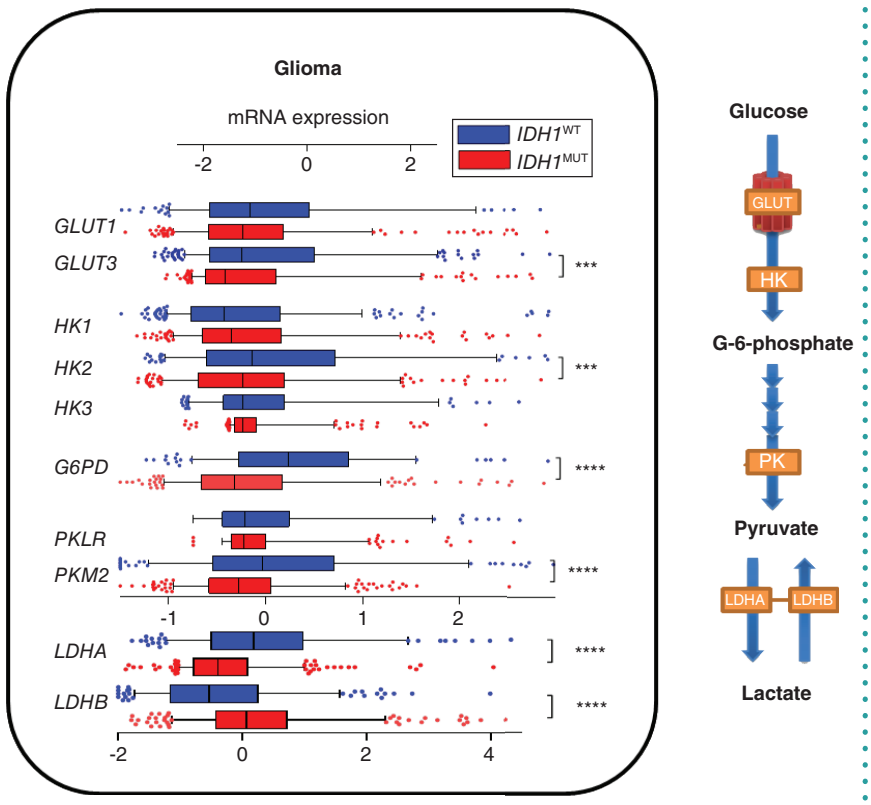

Figure 2. Representative analysis of mRNA expression levels of enzymes involved in glucose metabolism in *IDH1*^{WT} versus *IDH1*^{MUT} glioma. Analysis of *IDH1*^{WT} (n = 112) and *IDH1*^{MUT} (n = 399) low-grade glioma and *IDH1*^{WT} (n = 157) and *IDH1*^{MUT} (n = 9) glioblastoma samples, obtained from the cBioPortal using the TCGA datasets Brain Lower Grade Glioma (provisional) and Glioblastoma Multiforme (provisional). Merged data of relative mRNA expression levels are shown for I*DH1*^{WT} (blue) and *IDH1*^{MUT} (red).
***p < 0.001; ****p < 0.0001.
G6PD: Glucose-6-phosphate dehydrogenase; GLUT: Glucose transporter; HK: Hexokinase; LDH: Lactate dehydrogenase; PK: Pyruvate kinase.

can provide insights not only in biology but also provide details of structure, alterations and variations of transcripts [7,8]. Many open-access portals facilitate tools for exploration of gene-expression data. Our protocol is illustrated with the tool provided by cBioPortal [9,10]. The other open-access portals such as UCSC Cancer Browser and canEvolve can likewise be used for exploration of genomic data. We provide a step-by-step protocol with the next chapters (Supplemental Protocol):

1. Overview and selection of cancer dataset of interest (cBioPortal);

2. Creation of case sets/subsets of interest in a single study;

3. Integrative analysis of genes in a single study. After defining the cancer study of interest in section 1 and creating subsets of samples/patients with clinical or genetic data of interest in section 2. This section classifies each gene in each sample and is used for all genomic data analysis and visualization;

4. Collection of gene-expression and protein-expression data;

5. Collection of methylation data;

6. Correlation analysis. In order to investigate the correlation between gene expression and either methylation status or protein level, this section provides a tool to plot the relationship;

7. Graphical visualization and statistical analysis. Visualization and analysis of case sets of mRNA expression, methylation or protein-expression data collected in section 5, or data of correlation analysis of section 6.

RESULTS & DISCUSSION

Representative results

Mutations in the *IDH1* gene are ancestral events in the formation of low-grade glioma and secondary glioblastoma [11–13]. The presence of an *IDH1* mutation (*IDH1*^{MUT}) is associated with prolonged survival of glioma patients compared with *IDH1* wild-type (*IDH1*^{WT}) patients [11–13]. Utilizing clinical outcome possibilities of the cBioPortal, survival is illustrated in an overall survival plot with approximately sixfold prolonged survival of *IDH1*^{MUT} glioma patients compared with *IDH1*^{WT} glioma patients (Figure 1).

IDH1^{MUT} induces metabolic rewiring that is not fully understood [14,15] but exploration of differences in expression levels of metabolic enzymes is a promising investigational approach. The effects of *IDH1*^{MUT} on the expression of genes that encode for metabolic enzymes offer an opportunity to demonstrate the possibilities of the cBioPortal to perform data integration, exploration and analysis. TCGA offers data of 112 *IDH1*^{WT} versus 399 *IDH1*^{MUT} low-grade glioma (LGG) samples and 157 *IDH1*^{WT} versus nine *IDH1*^{MUT} glioblastoma samples to investigate and integrate for analysis.

In glucose metabolism, genes that encode for rate-limiting metabolic enzymes were selected: *GLUT1/3*, *HK1*, *HK2*, *HK3*, *PKLR*, *PKM2*, *LDHA* and *LDHB*. In *IDH1*^{WT} versus *IDH1*^{MUT} LGG and glioblastoma patient samples, higher levels of gene expression were observed for *GLUT3*, *HK2*, *PKM2* and *LDHA* (Figure 2), suggesting that *IDH1*^{WT} glioma depend more on glycolysis for ATP production than *IDH1*^{MUT} glioma.

As mutations in *IDH1/2* also occur in 20% of patients with myeloid neoplasms, including AML, an example of mRNA expression analysis of the three groups, *IDH*^{WT}, *IDH1*^{MUT} and *IDH2* ^{MUT} is presented in Figure 3. The study of acute myeloid leukemia (AML; TCGA, Provisional) offers 136 *IDH*^{WT}, 16 *IDH1*^{MUT} and 16 *IDH2*^{MUT} AML samples to investigate gene-expression profiles. In Figure 3, mRNA expression levels of the *ATM* gene, a DNA damage-response protein [16], in *IDH*^{WT}, *IDH1*^{MUT} and *IDH2*^{MUT} AML samples indicate that *ATM* mRNA expression is severely decreased in *IDH1*^{MUT} AML.

Another example is illustrated in Figure 4, which is a plot of gene expression versus DNA methylation of the *LDHA* gene in LGG. Lower expression levels of *LDHA* as observed in *IDH1*^{MUT} glioma were associated with hypermethylation of its promoter (Figure 4A), but lower expression levels of *LDHB* gene in *IDH1*^{WT} did not correlate with methylation (Figure 4B).

To investigate whether gene expression levels correlate with protein abundance, an illustrating example is demonstrated in Figure 5. In *IDH1*^{MUT} glioma, lower gene expression levels of G6PD were observed compared with *IDH1*^{WT} glioma (Figure 5A), whereas protein levels of G6PD were equal in *IDH1*^{MUT} and *IDH1*^{WT} LGG (Figure 5B), suggesting additional post-translational mechanisms at work [17].

Constant innovation has greatly aided the expansion of our understanding of cancer but has also transformed cancer research into one of the most data-intensive fields of biology. Well-structured and organized cancer genomics projects are offering researchers huge amounts of tumor samples that are similarly prepared, normalized and processed for computational analysis to extend our understanding of cancer genetics. The protocol that is listed here in combination with open-access tools lowers the barriers of access to these complex data and offers data mining in more depth to accelerate the translation of genomic data into novel biological and clinical insights.

The cancer genomics project of glioma was one of the first projects of TCGA that provided well-structured data of tumor samples from multiple platforms. Genomic ▶

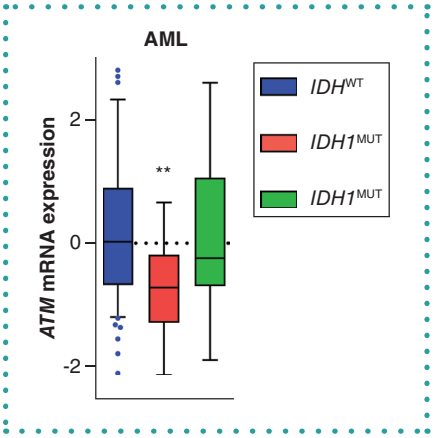

Figure 3. Representative analysis of mRNA expression levels of the ATM gene in *IDH*^{WT}, *IDH1*^{MUT} and *IDH2*^{MUT} AMLsamples. Analysis of *IDH*^{WT} (n = 138), *IDH1*^{MUT} (n = 16) and *IDH2*^{MUT} (n = 16) AML samples, obtained from the cBioPortal using the TCGA datasets Acute Myeloid Leukemia (provisional). Data of relative mRNA expression levels are shown for *IDH*^{WT} (blue), *IDH1*^{MUT} (red) and *IDH2*^{MUT} (green).
**p < 0.01.
AML: Acute myeloid leukemia.

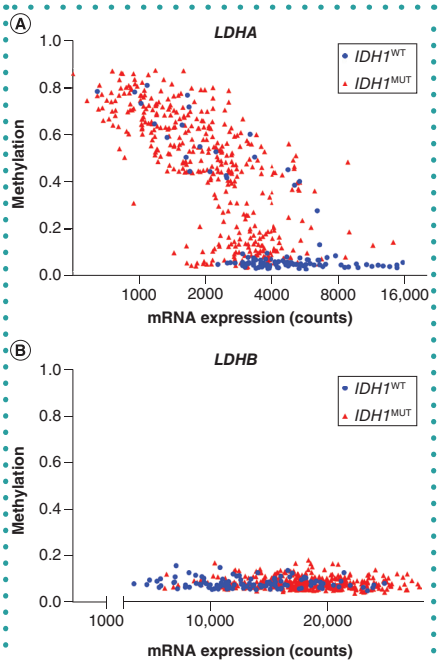

Figure 4. Representative analysis of the correlation between mRNA expression and methylation. Plot of correlation of gene expression and DNA methylation of the (A) *LDHA* gene and (B) *LDHB* gene in low-grade glioma (Brain Lower Grade Glioma, provisional) according to *IDH1*^{MUT} status (blue: *IDH1* ^{WT}, red: *IDH1* ^{MUT}).

▶ analysis of these data identified clinically relevant subtypes of glioblastoma [18] and delineated three different molecular classes in low-grade glioma, including the class with the *IDH* mutation [19]. Open-access portals facilitate access to these datasets but are limited in investigating specific groups. The protocol addressed in this paper describes a simple method to investigate subsets of samples or patients with a specific genetic, biological or clinical interest, such as the tumor samples with an *IDH* mutation. Secondly, the protocol describes how to generate expression profiles of genes involved in a particular pathway or process, such as metabolism, in this particular subset of samples. This allows selection of individual genes of interest instead of exploring all genes, and classifies each gene in each sample that is used for analysis and visualization. Finally, multidimensional analysis is provided to investigate gene expression in relation to DNA methylation and protein expression.

Comparable to other tools available, this protocol utilizes web interface tools that do not require additional software. A critical step in the protocol is the selection of the correct cancer genomics study or project that contains the data of interest. Currently, many portals store data from datasets from the literature and the TCGA portal. As an example, cBioPortal currently provides 76 cancer genomics projects of gene expression (RNAseq and microarray) in combination with 21 methylation and 41 protein expression projects. The validity of the comparison of genomics data is dependent on how well a sample is matched to the reference in terms of technical (e.g., type of data processing) and biological (e.g., molecular subtype) biases. Therefore, using portals that provide genomics data from well-structured cancer genomic projects require no advanced normalization techniques and batch corrections.

In summary, our method allows the import and integration of a selective subset of samples with specific genomic, biological or clinical interest, such as genomic alteration, mutation, cancer subtypes or survival properties. This method contains a unique concept to generate gene-expression profiles and to crosslink these profiles with DNA methylation and protein expression, which can be integrated to test research hypotheses in specific subtypes of cancer.

FUTURE PERSPECTIVE

Cancer research has evolved into one of the most data-intensive disciplines in biology. With the Genomics Evidence Neoplasia Information Exchange (GENIE) project among the largest fully public cancer genomic data sets released to date. Easy manageable portals, such as cBioPortal, will play an increasingly essential role in this discipline.

AUTHOR CONTRIBUTIONS

MK designed and performed the research, RJM and CJFvN supervised the study, MK and CJFvN wrote the manuscript, all authors read and approved the final version of the manuscript

FINANCIAL & COMPETING INTERESTS DISCLOSURE

This research was supported by the Dutch Cancer Society (KWF grants UVA 2014-6839 and AMC 2016.1-10460). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

OPEN ACCESS

REFERENCES

1. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 100(1), 57–70 (2000).
2. Chang JC, Wooten EC, Tsimelzon A *et al.* Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* 362(9381), 362–369 (2003).
3. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* 17(5), 257–271 (2016).
4. International Cancer Genome C, Hudson TJ, Anderson W *et al.* International network of cancer genome projects. *Nature* 464(7291), 993–998 (2010).
5. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nat. Med.* 17(3), 297–303 (2011).
6. Weeraratna AT, Nagel JE, De Mello-Coelho V, Taub DD. Gene expression profiling: from microarrays to medicine. *J. Clin. Immunol.* 24(3), 213–224 (2004).
7. Johnson JM, Castle J, Garrett-Engele P *et al.* Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302(5653), 2141–2144 (2003).
8. Trapnell C, Williams BA, Pertea G *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28(5), 511–515 (2010).

9. Gao J, Aksoy BA, Dogrusoz U *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal* 6(269), pl1 (2013).
10. Cerami E, Gao J, Dogrusoz U *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2(5), 401–404 (2012).
11. Parsons DW, Jones S, Zhang X *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* 321(5897), 1807–1812 (2008).
12. Balss J, Meyer J, Mueller W, Korshunov A, Hartmann C, Von Deimling A. Analysis of the ***IDH1*** codon 132 mutation in brain tumors. *Acta Neuropathol* 116(6), 597–602 (2008).
13. Bleeker FE, Atai NA, Lamba S *et al.* The prognostic IDH1(R132) mutation is associated with reduced NADP+-dependent IDH activity in glioblastoma. *Acta Neuropathol.* 119(4), 487–494 (2010).
14. Khurshed M, Molenaar RJ, Lenting K, Leenders WP, Van Noorden CJF. ***In silico*** gene expression analysis reveals glycolysis and acetate anaplerosis in ***IDH1*** wild-type glioma and lactate and glutamate anaplerosis in ***IDH1***-mutated glioma. *Oncotarget* 8(30), 49165–49177 (2017).
15. Khurshed M, Aarnoudse N, Hulsbos R *et al.* IDH1-mutant cancer cells are sensitive to cisplatin and an ***IDH1***-mutant inhibitor counteracts this sensitivity. *FASEB J.* 32, 6344–6352 (2018).
16. Molenaar RJ, Radivoyevitch T, Nagata Y *et al.* ***IDH1/2*** mutations sensitize acute myeloid leukemia to PARP inhibition and this is reversed by ***IDH1/2***-mutant inhibitors. *Clin. Cancer Res.* 24(7), 1705–1715 (2018).
17. Frederiks WM, Bosch KS, De Jong JS, Van Noorden CJ. Post-translational regulation of glucose-6-phosphate dehydrogenase activity in (pre)neoplastic lesions in rat liver. *J. Histochem. Cytochem.* 51(1), 105–112 (2003).
18. Verhaak RG, Hoadley KA, Purdom E *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in ***PDGFRA, IDH1, EGFR,*** and ***NF1***. *Cancer Cell* 17(1), 98–110 (2010).
19. Cancer Genome Atlas Research N, Brat DJ, Verhaak RG *et al.* Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* 372(26), 2481–2498 (2015).

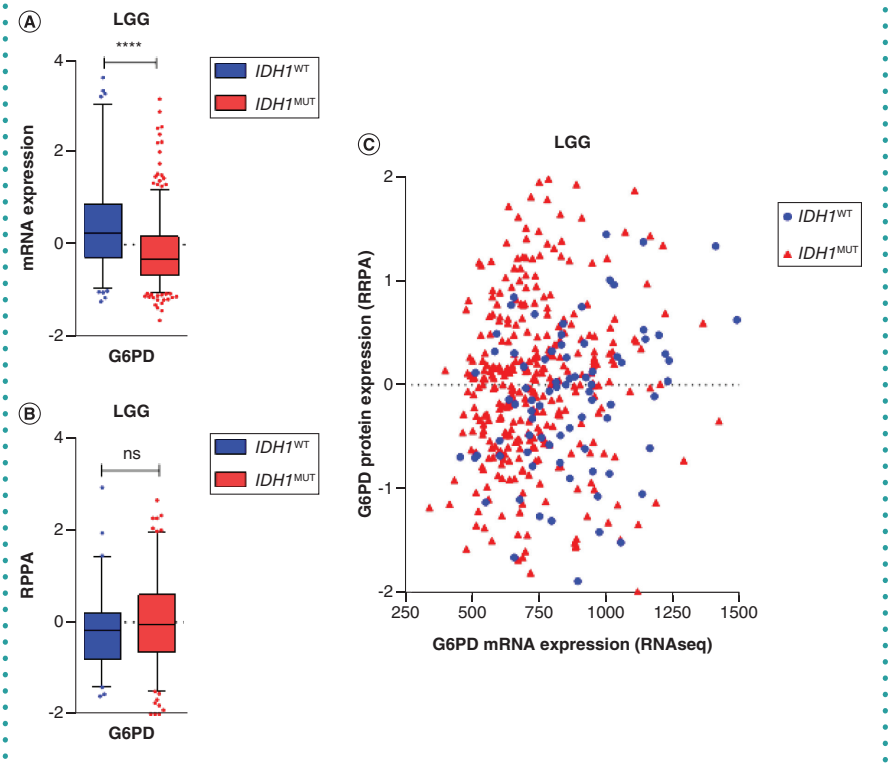

Figure 5. Representative analysis of the correlation between mRNA expression and protein abundance. (A) Analysis of gene expression levels of G6PD in low-grade glioma (LGG) in correlation with (B) protein abundance of G6PD according to IDH1^{MUT} status (blue: IDH1 ^{WT}, red: IDH1 ^{MUT}). (C) Plot of correlation of gene expression and protein abundance.
****$p < 0.0001$.
LGG: Low-grade glioma; ns: Not significant.