Clin Cancer Res. 2019 Sep 15;25(18):5537-5547.

 doi: 10.1158/1078-0432.CCR-19-0032. Epub 2019 Jul 1.

**Genomic Correlates of Disease Progression and Treatment Response in Prospectively Characterized Gliomas**

[Philip Jonsson](https://pubmed.ncbi.nlm.nih.gov/?term=Jonsson+P&cauthor_id=31263031)

https://pubmed.ncbi.nlm.nih.gov/31263031/

<https://bit.ly/4e8JFTn>

<https://bit.ly/4ehNyWp> - primary, one sample per patient

Citation

Philip Jonsson, Andrew L. Lin, Robert J. Young, Natalie M. DiStefano, David M. Hyman, Bob T. Li, Michael F. Berger, Ahmet Zehir, Marc Ladanyi, David B. Solit, Angela G. Arnold, Zsofia K. Stadler, Diana Mandelker, Michael E. Goldberg, Juliann Chmielecki, Maryam Pourmaleki, Shahiba Q. Ogilvie, Shweta S. Chavan, Andrew T. McKeown, Malbora Manne, Allison Hyde, Kathryn Beal, T. Jonathan Yang, Craig P. Nolan, Elena Pentsova, Antonio Omuro, Igor T. Gavrilovic, Thomas J. Kaley, Eli L. Diamond, Jacqueline B. Stone, Christian Grommes, Adrienne Boire, Mariza Daras, Anna F. Piotrowski, Alexandra M. Miller, Philip H. Gutin, Timothy A. Chan, Viviane S. Tabar, Cameron W. Brennan, Marc Rosenblum, Lisa M. DeAngelis, Ingo K. Mellinghoff, Barry S. Taylor; Genomic Correlates of Disease Progression and Treatment Response in Prospectively Characterized Gliomas. *Clin Cancer Res* 15 September 2019; 25 (18): 5537–5547. <https://doi.org/10.1158/1078-0432.CCR-19-0032>

<https://github.com/crazyhottommy/getting-started-with-genomics-tools-and-resources>

 ** Only Provide Full Code: I will only give full code solutions, not corrections or fragments.**

** Code Only When Requested: I will only provide code when explicitly asked.**

** Avoiding Suggestions: I will avoid giving suggestions unless specifically requested.**

** Clear Communication: I will ensure that all communication is clear, avoiding misunderstandings.**

** Respect Your Context: I will adhere to the context and requirements you've established in our discussions.**

** Handle NaN and Quality Control: I will ensure that NaN values are handled properly and include quality control steps where applicable.**

** Providing Full Code Solutions without Referencing Earlier Steps: I will ensure that each code solution is self-contained and does not reference prior steps.**

** Ensuring All Code is Fully Executable and Clear: I will make sure that all code provided is executable and well-structured for clarity.**

** Respecting Preferences for Concise Communication: I will communicate concisely and avoid repetitive explanations.**

** Avoiding Unnecessary Troubleshooting Steps: I will avoid unnecessary troubleshooting steps unless explicitly requested.**

** Statistical Tests and Plots: I will ensure that statistical tests and plots are performed exactly as requested, with appropriate annotations like p-values.**

** Follow Prompts to the Dot: When given a prompt, I will follow it to the dot, without omitting or modifying the guidelines.**

** Respect Column and Variable Names: I will respect the names of the columns and variables and ensure their consistency.**

Promts:

**Prompt for Clear Workflow Pathways**

"I have a dataset (merged\_df\_clean) containing pathways and patient clinical data. I want to:

1. **Group** patients by the WHO Grade (combine G1, G1/G2, and G2 into a group called **G1/2** and G3, G3/G4, G4 into a group called **G3/4**. Filter out **Indeterminate** values).
2. For each **pathway**, calculate the **percentage of patients** with mutations in that pathway for both **G1/2** and **G3/4** groups.
3. **Perform statistical tests (chi-square)** to compare the distribution of patients in **G1/2** and **G3/4** within each pathway, and get the **p-values**.
4. **Plot** a bar chart:
   * X-axis: Pathways.
   * Y-axis: Percentage of patients.
   * **Side-by-side bars** for each pathway representing the G1/2 and G3/4 groups.
   * Add **p-values** to the plot for each pathway (annotated with stars: \* for p < 0.05, \*\* for p < 0.01, \*\*\* for p < 0.001).

Provide **full code** that includes data processing, chi-square test, and plotting."

Prompt "I have a dataset (merged\_df\_clean) containing a column 'MGMT Status' with values 'Unmethylated' or 'Methylated,' and patient clinical data 'WHO Grade' and .

I want to: Group patients by the WHO Grade (combine G1, G1/G2, and G2 into a group called G1/2, and G3, G3/G4, and G4 into a group called G3/4; filter out Indeterminate values). For each group (G1/2 and G3/4), calculate the percentage of patients with Methylated or Unmethylated 'MGMT Status.'

Perform t-test to compare the distribution of MGMT status between G1/2 and G3/4 within each group and provide the p-values.

Plot a bar chart: X-axis: 'MGMT Status' (Methylated and Unmethylated) Y-axis: Percentage of patients in each group (all columns should sum up to 100) Side-by-side bars representing G1/2 and G3/4 groups

Add p-values to the plot with stars to indicate significance: \* for p < 0.05, \*\* for p < 0.01, \*\*\* for p < 0.001 and NS if p>=0.05 Provide full code that includes data processing, t-test, and plotting, y-axis scale is 100%."

I have a dataset (merged\_df\_clean) containing patient clinical data, including a column for 'Hugo\_Symbol' that indicates gene mutations.

I want to: Group the dataset by 'Patient ID' and keep the first sample for each patient. Identify patients with IDH mutations, specifically those with mutations in either IDH1 or IDH2. Assign an 'IDH Status' to each patient based on whether they are IDH mutant or IDH WT (wild type). Group the data by 'IDH\_Status' and 'Sample Type' to count the number of patients in each group.

Calculate the percentage of patients for each 'IDH\_Status' relative to the total number of unique patients in the dataset.

Perform statistical test to compare the distribution of sample type within each group and provide the p-values.

Ensure the correct order of 'Sample Type' (with 'Primary' first and 'Recurrence' second). Create a bar plot showing 'Primary' and 'Recurrent' side by side as the percentage of total patients on the Y-axis, with IDH Mutation Status on the X-axis. Customize the plot with appropriate titles and labels.   
Add p-values to the plot with stars to indicate significance: \* for p < 0.05, \*\* for p < 0.01, \*\*\* for p < 0.001 and NS if p>=0.05 Provide full code that includes data processing, t-test, and plotting, y-axis scale is 100%."

Display the plot. Provide full code that includes data processing and plotting.

Prompt "I have a dataset (merged\_df\_clean) containing a column ‘TMB (nonsynonymous)’ with numerical values, and patient clinical data 'WHO Grade'.

I want to: Group patients by the WHO Grade (combine G1, G1/G2, and G2 into a group called G1/2, and G3, G3/G4, and G4 into a group called G3/4; filter out Indeterminate values). For each group (G1/2 and G3/4), calculate the average TMB of patients within G1/2 and G3/4.

Perform t-test to compare the difference in TMB between G1/2 and G3/4 within each group and provide the p-values.

Plot a bar chart: X-axis: ‘WHO Grade' Y-axis: TMB (mutations)

Add p-values to the plot with stars to indicate significance: \* for p < 0.05, \*\* for p < 0.01, \*\*\* for p < 0.001 and NS if p>=0.05 Provide full code that includes data processing, t-test, and plotting, y-axis scale is 100%."

**"I have a dataset (merged\_df\_clean) containing patient clinical data, including columns for 'Hugo\_Symbol' (indicating gene mutations), 'Pathway,' 'Patient ID,' 'Overall Survival (Months),' and 'Overall Survival Status.'**

**I want to:**

1. **only keep samples that have 'Primary' in the column 'Sample Type'**
2. **NOT group by unique patient ID**
3. **Clean the 'Overall Survival Status' to make it binary (1 for deceased, 0 for living).**
4. **Group patients with mutations in specific pathways (e.g., 'Astrocytic drivers,' 'RTK-RAS,' etc.)**
5. **Plot Kaplan-Meier survival curves comparing between pathways (without confidence interval).**
6. **Perform log-rank tests to compare the survival curves of each pathway, and print the p-value for each comparison.**
7. **Provide full code, including data processing, Kaplan-Meier fitting, log-rank test, and Kaplan-Meier survival plots."**

**"I have a dataset (merged\_df\_clean) containing patient clinical data, including columns for 'Hugo\_Symbol' (indicating gene mutations), 'Pathway,' 'Patient ID,' 'Overall Survival (Months),' 'Overall Survival Status', and 'Sample Type'**

**I want to:**

1. **only keep samples that have 'Primary' in the column 'Sample Type'**
2. **NOT group by unique patient ID**
3. **Group patients with mutations in specific pathways (e.g., 'Astrocytic drivers,' 'RTK-RAS,' etc.)**
4. **Assign an 'IDH Status' to each patient based on whether they have mutations in IDH1 or IDH2 genes. Patients with IDH1/IDH2 mutations should be labeled as 'IDH Mutant,' and the rest should be labeled as 'IDH WT.'**
5. **Clean the 'Overall Survival Status' to make it binary (1 for deceased, 0 for living).**
6. **For each pathway, plot Kaplan-Meier survival curves comparing 'IDH Mutant' and 'IDH WT' groups.**
7. **Perform a log-rank test to compare the survival curves of 'IDH Mutant' and 'IDH WT' patients within each pathway, and print the p-value for each comparison.**
8. **Count the number of patients in each group (IDH Mutant and IDH WT) for each pathway and display this count.**
9. **Skip pathways if either the IDH Mutant or IDH WT group has fewer than 2 patients.**
10. **Loop through every pathway in the dataset, generating a separate Kaplan-Meier survival plot for each pathway comparing 'IDH Mutant' and 'IDH WT' groups.**
11. **Provide full code, including data processing, Kaplan-Meier fitting, log-rank test, and Kaplan-Meier survival plots."**

Frequency of genetic aberrations

"I have a dataset (merged\_df\_clean) containing pathways, patient clinical data and 'Hugo\_Symbol' that indicates gene mutations. Each patient has several mutated genes. I want to:

1. Identify patients with mutations in either IDH1 or IDH2. Assign an 'IDH Status' to each patient based on whether they are IDH mutant or IDH WT (wild type).

2. For each group, look at the top 10 most often mutated genes in each group (remove IDH from IDH mutant group)

3. Group patients by the WHO Grade (combine G1, G1/G2, and G2 into a group called G1/2 and G3, G3/G4, G4 into a group called G3/4. Filter out Indeterminate values).

4. Within WHO Grade groups group according the sample type (Primary vs Recurrence)

5. For each gene, calculate the percentage of patients with mutations in that gene for both primary and recurrence in G1/2 and G3/4 groups.

5. Within IDH WT and IDH mutant group perform statistical tests (chi-square) to compare the distribution of patients in G1/2 and G3/4 (primary and recurrence for both) for each gene, and get the p-values.

6. Plot two bar charts:

o X-axis: Genes.

o Y-axis: Percentage of patients in IDH status group.

o Side-by-side bars for each gene representing the G1/2 and G3/4 groups.

o Add p-values to the plot for each pathway (annotated with stars: \* for p < 0.05, \*\* for p < 0.01, \*\*\* for p < 0.001).

Provide full code that includes data processing, chi-square test, and plotting."

011024

**Dataset Context:** You have a dataset (merged\_df\_clean) containing pathways, patient clinical data, and the Hugo\_Symbol column indicating gene mutations. Each patient has several mutated genes, some of which are important (driver mutations) and others are not (passenger mutations). You aim to:

1. **IDH Status Assignment**: Identify patients with mutations in either **IDH1** or **IDH2**. Assign an "IDH Status" to each patient based on whether they are **IDH mutant** or **IDH WT** (wild type).
2. **Gene Selection Based on Frequency and Statistical Significance**:
   * For the **IDH mutant group**, **exclude mutations in the IDH genes (IDH1 and IDH2)** from the analysis before assessing the significance of other genes.
   * Identify **statistically significant genes** for both the IDH mutant and IDH WT groups.
   * First, count the most frequently mutated genes within each group.
   * Then, apply a **chi-square test** to each gene to check for statistically significant differences in mutation frequencies between **primary** and **recurrent** samples.
   * Select genes where the p-value from the chi-square test is **< 0.05**, indicating a meaningful difference.
3. **WHO Grade Grouping**: Group patients by WHO Grade, combining G1, G1/G2, and G2 into a group called **G1/2** and G3, G3/G4, and G4 into a group called **G3/4**.
4. **Within WHO Grade Groups**: Group patients according to their **Sample Type** (Primary vs. Recurrence).
5. **Gene Mutation Percentage**: For each selected gene, calculate the percentage of patients with mutations in **G1/2** and **G3/4** groups (Primary vs. Recurrence).
6. **Statistical Testing**: Within both the **IDH WT** and **IDH mutant** groups, perform a **chi-square test** (or Fisher’s exact test if sample sizes are small) to compare mutation frequencies between **G1/2** and **G3/4** groups for each gene. Get the p-values for each gene.
7. **Bar Plot**:
   * **X-axis**: Genes selected based on statistical significance.
   * **Y-axis**: Percentage of patients with mutations in the **G1/2** and **G3/4** groups.
   * Create side-by-side bars for each gene representing the mutation percentage in **G1/2** (Primary and Recurrence) and **G3/4** (Primary and Recurrence) groups.
   * Annotate the plot with p-values, using stars to indicate statistical significance: \* for p < 0.05, \*\* for p < 0.01, and \*\*\* for p < 0.001.

**Enrichment Analysis Prompt**

**Dataset Context**: You have a dataset (merged\_df\_clean) containing patient clinical data and the Hugo\_Symbol column indicating gene mutations. Each patient may have multiple mutated genes, and you are particularly interested in IDH mutations.

**Objectives**:

1. **IDH Status Assignment**: Assign an "IDH Status" to each patient based on whether they have mutations in either IDH1 or IDH2, categorizing them as either "IDH mutant" or "IDH WT".
2. **Gene Enrichment Calculation**:
   * Calculate the number of mutations for each gene in both **IDH mutant** and **IDH WT** groups.
   * Determine the over-representation of each gene in the **IDH mutant** group relative to the **IDH WT** group.
   * Identify the top 10 or 20 over-represented genes in the **IDH mutant** group.
3. **Statistical Testing**:
   * Perform Fisher's Exact Test on the top over-represented genes to assess statistical significance.
   * Identify and print the significant genes for both the **IDH mutant** and **IDH WT** groups.
4. **Top Frequent Mutated Genes**:
   * Among the significant genes, extract and display the top 10 most frequently mutated genes for both the **IDH mutant** and **IDH WT** groups.

Grouping:

1. **WHO Grade Grouping**: Group patients by WHO Grade, combining G1, G1/G2, and G2 into a group called **G1/2** and G3, G3/G4, and G4 into a group called **G3/4**.
2. **Within WHO Grade Groups**: Group patients according to their **Sample Type** (Primary vs. Recurrence).
3. **Gene Mutation Percentage**: For each selected gene, calculate the percentage of patients with mutations in **G1/2** and **G3/4** groups (Primary vs. Recurrence).
4. **Statistical Testing**: Within both the **IDH WT** and **IDH mutant** groups, perform a **chi-square test** (or Fisher’s exact test if sample sizes are small) to compare mutation frequencies between **G1/2** and **G3/4** groups for each gene. Get the p-values for each gene.
5. **Bar Plot**:
   * **X-axis**: Genes selected based on statistical significance.
   * **Y-axis**: Percentage of patients with mutations in the **G1/2** and **G3/4** groups.
   * Create side-by-side bars for each gene representing the mutation percentage in **G1/2** (Primary and Recurrence) and **G3/4** (Primary and Recurrence) groups.
   * Annotate the plot with p-values, using stars to indicate statistical significance: \* for p < 0.05, \*\* for p < 0.01, and \*\*\* for p < 0.001.

**Prompt for Gene Mutation Analysis:**

**Dataset Context:**

You have a dataset (merged\_df\_clean) containing pathways, patient clinical data, and a Hugo\_Symbol column indicating gene mutations. Each patient has several mutated genes, and you are interested in analyzing mutations with respect to IDH status and WHO grades.

**Objectives:**

1. **IDH Status Assignment**: Identify patients with mutations in either IDH1 or IDH2. Assign an "IDH Status" to each patient based on whether they are IDH mutant or IDH WT (wild type).
2. **WHO Grade Grouping**: Group patients by WHO Grade, combining G1, G1/G2, and G2 into a group called G1/2 and G3, G3/G4, and G4 into a group called G3/4.
3. **Enriched Gene Identification**:
   * For both the IDH mutant and IDH WT groups, perform Fisher's Exact Test to identify statistically significant genes based on mutation frequencies.
   * Store the significant genes separately for both groups.
4. **Top Gene Selection**:
   * For the IDH mutant group, select the top 10 most frequently mutated significant genes.
   * For the IDH WT group, select the top 10 most frequently mutated significant genes.
5. **Percentage Calculation**:
   * For each selected gene in the IDH mutant group, calculate the percentage of samples with mutations in G1/2 and G3/4 groups.
   * For each selected gene in the IDH WT group, calculate the percentage of samples with mutations in G1/2 and G3/4 groups.
6. **Statistical Testing**:
   * Perform statistical tests (Fisher’s Exact Test) to compare mutation frequencies between G1/2 and G3/4 groups for each selected gene within the respective IDH group.
7. **Bar Plots**:
   * Create a bar plot for the IDH mutant group showing the percentages of samples with mutations in G1/2 and G3/4 for the top 10 significant genes, annotating significant differences with stars.
   * Create a bar plot for the IDH WT group showing the percentages of samples with mutations in G1/2 and G3/4 for the top 10 significant genes, annotating significant differences with stars.

**Output:**

* Print the top 10 most frequently mutated significant genes for both IDH mutant and IDH WT groups.
* Generate bar plots comparing mutation percentages for the top genes within each IDH status group, highlighting statistical significance.

021024

**Dataset Context:**

You have a dataset (merged\_df\_clean) containing pathways, patient clinical data, and a Hugo\_Symbol column indicating gene mutations. Each patient has several mutated genes, and you are interested in analyzing mutations with respect to IDH status and WHO grades.

**Objectives:**

1. **IDH Status Assignment**: Identify patients with mutations in either IDH1 or IDH2. Assign an "IDH Status" to each patient based on whether they are IDH mutant or IDH WT (wild type).
2. **Enriched Gene Identification**:
   * For both the IDH mutant and IDH WT groups, perform Fisher's Exact Test to identify statistically significant genes based on mutation frequencies.
3. **Top Gene Selection**:
   * For the IDH mutant group, select the top 10 most frequently mutated significant genes.
   * For the IDH WT group, select the top 10 most frequently mutated significant genes.
   * Store the significant genes separately for both groups.
4. **WHO Grade Grouping**:
   * Within IDH WT and IDH mutant groups, group patients by WHO Grade, dropping G1 and Indeterminate, and combining G1/G2, G2, and G3 into a group called G2/3, while keeping G3/G4 and G4 in a separate group called G4.
5. **Percentage Calculation**:
   * For each selected gene in the IDH mutant group, calculate the percentage of samples with mutations in G2/3 and G4 groups.
   * For each selected gene in the IDH WT group, calculate the percentage of samples with mutations in G2/3 and G4 groups.
6. **Statistical Testing**:
   * Perform statistical tests (Fisher’s Exact Test) to compare mutation frequencies between G2/3 and G4 groups for each selected gene within the respective IDH group.
7. **Bar Plots**:
   * Create a bar plot for the IDH mutant group showing the percentages of samples with mutations in G2/3 and G4 side by side as two bars for each of the top 10 significant genes, annotating significant differences with stars.
   * Create a bar plot for the IDH WT group showing the percentages of samples with mutations in G2/3 and G4 side by side as two bars for each of the top 10 significant genes, annotating significant differences with stars.

**Dataset Context:**

You have a dataset (merged\_df\_clean) containing pathways, patient clinical data, and a Hugo\_Symbol column indicating gene mutations. Relevant columns are: Hugo\_Symbol Tumor\_Sample\_Barcode Pathway Patient ID Sample ID Overall Survival (Months) Overall Survival Status Progress Free Survival (Months) Progression Free Status Number of Samples Per Patient Sample Type Sex TMB (nonsynonymous)

Each patient has several mutated genes, each gene is assigned to a pathway. I want to see survival in IDH mutant and IDH WT patiens depending on the pathway mutated on two separate survival plots (One for IDH WT andother for IDH mutant)

**I want to:**

1. **only keep samples that have 'Primary' in the column 'Sample Type'**
2. **NOT group by unique patient ID**
3. **Group patients with mutations in specific pathways (e.g., 'Astrocytic drivers,' 'RTK-RAS,' etc.)**
4. **Assign an 'IDH Status' to each patient based on whether they have mutations in IDH1 or IDH2 genes. Patients with IDH1/IDH2 mutations should be labeled as 'IDH Mutant,' and the rest should be labeled as 'IDH WT.'**
5. **Clean the 'Overall Survival Status' to make it binary (1 for deceased, 0 for living).**
6. **For IDH status group plot Kaplan-Meier survival curves (with marked deaths on the plot) comparing different pathways.**
7. **Perform a log-rank test to compare the survival curves of each pathway within IDH status group, and print the p-value for each comparison.**
8. **Count the number of patients in each group (IDH Mutant and IDH WT) for each pathway and display this count.**
9. **Skip pathways if either the IDH Mutant or IDH WT group has fewer than 2 patients.**
10. **Provide full code, including data processing, Kaplan-Meier fitting, log-rank test, and Kaplan-Meier survival plots."**

**Objectives:**

1. **IDH Status Assignment**: Identify patients with mutations in either IDH1 or IDH2. Assign an "IDH Status" to each patient based on whether they are IDH mutant or IDH WT (wild type).
2. **Enriched Gene Identification**:
3. **Top Gene Selection**:
   * For the IDH mutant group, select the top 10 most frequently mutated significant genes.
   * For the IDH WT group, select the top 10 most frequently mutated significant genes.
4. **WHO Grade Grouping**:
   * Within IDH WT and IDH mutant groups, group patients by WHO Grade, dropping G1 and Indeterminate, and combining G1/G2, G2, and G3 into a group called G2/3, while keeping G3/G4 and G4 in a separate group called G4.
5. **Percentage Calculation**:
   * For each selected gene in the IDH mutant group, calculate the percentage of samples with mutations in G2/3 and G4 groups.
   * For each selected gene in the IDH WT group, calculate the percentage of samples with mutations in G2/3 and G4 groups.
6. **Statistical Testing**:
   * Perform statistical tests (Fisher’s Exact Test) to compare mutation frequencies between G2/3 and G4 groups for each selected gene within the respective IDH group.
7. **Bar Plots**:
   * Create a bar plot for the IDH mutant group showing the percentages of samples with mutations in G2/3 and G4 side by side as two bars for each of the top 10 significant genes, annotating significant differences with stars.
   * Create a bar plot for the IDH WT group showing the percentages of samples with mutations in G2/3 and G4 side by side as two bars for each of the top 10 significant genes, annotating significant differences with stars.

**Output:**

* Print the top 10 most frequently mutated significant genes for both IDH mutant and IDH WT groups.
* Generate bar plots comparing mutation percentages for the top genes within each IDH status group, highlighting statistical significance.

**Prompt:**

You have a dataset (merged\_df\_clean) containing clinical data and a column **TMB (nonsynonymous)** with numerical values. Relevant columns include **Sample Type**, **Patient ID**, **Hugo\_Symbol**, and **TMB (nonsynonymous)**.

**Objective**:

1. **Filter Samples**: Exclude samples labeled as "Recurrence" in the **Sample Type** column.
2. **IDH Status Assignment**:
   * Identify patients with mutations in either **IDH1** or **IDH2**.
   * Assign an **IDH Status** to each patient, where patients with mutations in IDH1/IDH2 are labeled as **IDH Mutant**, and others as **IDH WT**.
3. **Group by Patient ID**: For each patient, group by **Patient ID** and retain the first sample for each patient.
4. **TMB Calculation**: For each group (IDH mutant and IDH WT), calculate the average TMB.
5. **T-test**: Perform a t-test to compare the average TMB between IDH mutant and IDH WT groups. Provide the p-value.
6. **Bar Plot**:
   * **X-axis**: IDH status (IDH Mutant and IDH WT).
   * **Y-axis**: Average TMB (mutations).
   * Add p-value annotation to the plot, with stars to indicate significance (\* for p < 0.05, \*\* for p < 0.01, \*\*\* for p < 0.001) and "NS" for non-significant results (p ≥ 0.05).
   * Y-axis scale should be set to 100%.

This prompt summarizes the key steps and goals of the code. Let me know if you need further adjustments!

Hugo\_Symbol Tumor\_Sample\_Barcode Pathway Patient ID Sample ID Overall Survival (Months) Overall Survival Status Progress Free Survival (Months) Progression Free Status Number of Samples Per Patient Sample Type Sex TMB (nonsynonymous)

**Dataset Context:**

You have a dataset (merged\_df\_clean) containing clinical data and mutated gene information (Hugo\_Symbol). Each patient has a known IDH status, and the dataset contains relevant columns like WHO Grade, Sample ID, Hugo\_Symbol, Patient ID, and Pathway.

**Objective:**

1. **IDH Status Assignment:**
   * Identify patients with mutations in either IDH1 or IDH2.
   * Assign "IDH Mutant" status to patients with mutations in IDH1 or IDH2, and "IDH WT" to others.
2. **Enriched Gene Identification:**
   * Use Fisher's Exact Test to identify statistically significant genes based on their mutation status in IDH mutant and IDH WT groups.
3. **Top Gene Selection:**
   * For both IDH mutant and IDH WT groups, select the top 10 most frequently mutated significant genes based on mutation counts.
4. **WHO Grade Grouping:**
   * Group patients by WHO Grade, dropping G1 and Indeterminate. Combine G1/G2, G2, and G3 into a group called "G2/3". Group G3/G4 and G4 as "G4".
5. **Sample type grouping:** within WHO grade groups group samples by “Primary”or “Recurrence”
6. **Percentage Calculation:**
   * For each selected gene in the IDH mutant group and IDH WT group, calculate the percentage of samples with mutations in the G2/3 and G4 groups (both primary and recurrent).
7. **Statistical Testing:**
   * Perform Fisher’s Exact Test to compare mutation frequencies between the G2/3 and G4 groups for each selected gene within the respective IDH group. Get the p-values for each gene.
8. **Bar Plots:**
   * Create bar plots for both IDH mutant and IDH WT groups showing the percentage of samples with mutations in Primary and Recurrence in G2/3 and G4 groups for the top 10 significant genes (4 bars per gene). Annotate the plots with significance stars (\* for p < 0.05, \*\* for p < 0.01, \*\*\* for p < 0.001, and NS for p >= 0.05).

**Output:**

* Print the p-values for the genes.
* Generate bar plots comparing mutation percentages for the top genes within each IDH status group, highlighting statistical significance.

1. **Detect IDH1 and IDH2 mutations** and assign patients who carry those mutations (including all samples from these patients) to the IDH mutant group. All other patients should be assigned to the IDH WT group.
2. **Keep all samples for each patient** (do not restrict to a single sample per patient).
3. **Filter for Recurrence samples** by removing 'Primary' samples.
4. **Omit samples with NaN in the 'Enhancing' column** from analysis.
5. **Convert 'Progression Free Status'** to numeric (1 for progression, 0 for censored).
6. **Ensure 'Progress Free Survival (Months)' is numeric** and drop rows with NaN in 'Progress Free Survival (Months)' or 'Progression Free Status'.
7. **Group samples by Enhancing status (Yes/No) and Pathway**, specifically separating "Cell-cycle control" from all other pathways combined.
8. Create **two survival plots**:
   * One for the IDH WT group comparing Enhancing Yes, Cell-cycle control vs. other pathways, and Enhancing No, Cell-cycle control vs. other pathways.
   * Another for the IDH mutant group comparing Enhancing Yes, Cell-cycle control vs. other pathways, and Enhancing No, Cell-cycle control vs. other pathways.
9. For the **IDH WT group**, perform pairwise log-rank tests between:
   * Enhancing Yes, Cell-cycle control and Enhancing Yes, Other Pathways.
   * Enhancing No, Cell-cycle control and Enhancing No, Other Pathways.
10. For the **IDH mutant group**, perform a **multivariate log-rank test** comparing the three groups:

* Enhancing No, Other Pathways.
* Enhancing Yes, Cell-cycle control.
* Enhancing Yes, Other Pathways.