Pan-Cancer Multiomics Landscape: Summary of Key *Preliminary* Findings

1. Mutation Frequency Landscape

1. The distribution of mutation frequencies across genes follows the characteristic long-tail pattern typical of cancer genomics - most genes exhibit very low mutation frequencies, while a small number show very high mutation frequencies. This reflects the biological reality that while most genes are rarely mutated, a core set of driver genes is recurrently altered across many tumors.
2. Notably, genes such as TP53, TTN, APC, MUC16, KRAS, and PIK3CA appear as frequently mutated across multiple cancer types, consistent with their roles as canonical drivers of tumor initiation and progression. These recurrent mutations make them strong candidates for therapeutic targeting and biomarker development.
3. Tumors with higher mutation burden (e.g., LUAD, LUSC) may also be more amenable to immunotherapy, as high mutation load is linked to increased neoantigen formation and immune recognition.

2. CNV Patterns Across Tumors

2.1 CNV Frequency Distribution

1. Most genes show low CNV frequencies, with median values near zero for deep deletion (−2) and amplification (+2) events, indicating that high-level CNV changes are relatively rare and localized to specific driver loci. In contrast, shallow gains (+1) and shallow losses (−1) are far more common, with median frequencies around 0.13–0.14, reflecting widespread chromosomal instability.
2. Tumor types such as LUAD, LUSC, HNSC, and ESCA exhibit a high burden of shallow CNV events, with long-tailed frequency distributions and maximum values exceeding 0.8, consistent with their known chromosomal instability.
3. Conversely, THCA stands out as a CNV-poor tumor type with narrow distributions centered near zero.

2.2 Top Amplified and Deleted Genes

1. Heatmap analyses highlight amplification as the dominant CNV signal, especially in ESCA, LUSC, BRCA, and HNSC, where oncogenic loci like SOX2 and DCUN1D1 on 3q26 show recurrent high-frequency amplifications.
2. Deletions are more focal but consistently observed at tumor suppressor loci such as CDKN2A/CDKN2B, particularly in ESCA, LUSC, BLCA, and HNSC.
3. UpSet analysis reveals 31 amplified and 34 deleted genes shared across multiple cancers, indicating pan-cancer CNV hotspots.
4. There are also unique signatures, such as TNFRSF10B and RHOBTB2 amplification in UCEC-BRCA and a deletion signature spanning 9 genes uniquely altered in UCEC, LUAD, and PRAD. These may point to subtype- or lineage-specific vulnerabilities.

2.3 Variability Analysis

1. Selecting top genes by variance in CNV frequency across cancers highlights sets of genes that best differentiate tumor types. Distinct grouping patterns (e.g., LUAD, BLCA, HNSC vs. ESCA, KIRC, LUSC, CHOL) emerge visually in heatmaps, suggesting shared CNV profiles among certain cancers.
2. However, these groupings are exploratory and should be validated with formal clustering analyses before biological interpretation.

3. Expression and Methylation Changes

1. Z-score transformation of |logFC| values allowed direct comparison between methylation and expression layers, which otherwise differ in scale (expression logFC up to ~10 vs methylation logFC up to ~0.6).
2. Most genes cluster near zero, with expression distributions broader and more skewed, indicating a subset of genes with large transcriptional changes.
3. Methylation changes are tightly centered around zero, reflecting more subtle alterations.
4. CHOL, LUAD, LUSC, COAD, and UCEC show the strongest expression perturbations, whereas methylation remains comparatively stable across cancers.

4. Volcano Plot Analysis

1. Volcano plots revealed a large number of significantly differentially regulated genes in both methylation and expression datasets.
2. These plots provide a systematic way to extract gene lists for downstream integration and cross-cancer comparisons.
3. Key genes showing both significant methylation and expression alterations were identified (n = 822), forming a candidate set for further mechanistic investigation.

5. CNV Amplification and Deletion Impact on Expression

1. Using the Wilcoxon rank-sum test to compare |logFC| distributions between CNV-altered and unaltered groups revealed:
2. Amplified genes show a statistically significant difference in expression magnitude. Even though the shift is small, the large sample size makes this difference detectable.
3. Deleted genes show no significant shift, indicating that CNV deletions in this dataset do not systematically affect expression magnitude.
4. This result underscores that statistical significance does not necessarily imply a large biological effect, but it can still point to subtle yet consistent regulatory influences.

6. CNV Burden and Expression Correlation

1. The Spearman correlation between CNV burden and expression perturbation is very weak, both pan-cancer and within individual tumor types (max rho ~ 0.07).
2. This indicates that global CNV burden alone does not explain variation in expression changes across tumors.
3. Similarly, methylation–expression correlations are close to zero, suggesting that while these layers may interact locally at specific genes, there is no strong global monotonic relationship.

7. Integrated Observations and Biological Implications

1. Tumor genomes display a heterogeneous landscape with distinct mutation, CNV, expression, and methylation signatures across cancer types.
2. Certain driver genes (e.g., TP53, KRAS, SOX2, CDKN2A) show recurrent alterations across multiple layers, making them high-priority candidates for targeted studies.

3. CNV amplification has a measurable but modest effect on gene expression, whereas CNV deletions show limited impact globally.
4. Pan-cancer CNV hotspots and expression outliers can serve as a starting point for biomarker discovery or therapeutic target prioritization.
5. Integration with external driver gene databases (e.g., COSMIC) or pathway-level annotation will be essential to distinguish driver alterations from passenger events.

8. Next Steps / Recommendations

- Perform unsupervised clustering (e.g., hierarchical clustering, PCA/UMAP) to validate cancer-type groupings based on CNV profiles.
- Integrate with driver gene databases and pathway analyses to refine gene prioritization.
- Explore co-occurrence patterns of CNV, methylation, and expression changes for potential multi-omic biomarkers.
- Investigate key recurrent genes in the context of clinical data (e.g., survival, therapy response) to assess translational relevance.