Paired Analysis of Lung Cancer Tumor Cells Using JIVE

Introduction

This analysis addresses challenges in studying squamous cell lung carcinoma (LUSC) using The Cancer Genome Atlas (TCGA) data. TCGA, a major cancer genomics program, offers comprehensive multi-omics data, including for LUSC, a leading cause of cancer-related deaths globally. LUSC, often linked to smoking, exhibits genomic alterations distinct from other lung cancers. Key analytical challenges include the non-independence of normal and tumor tissue samples from the same patient, the high-dimensionality of the data (RNA-sequencing data with 60,660 genes and methylation data with 422,357 sites), and the integration of multiple data sources. To address these, the analysis employs Principal Component Analysis (PCA) for dimension reduction, using an eigendecomposition approach ($\Sigma p \times p = Up \times pDp \times pUpT \times p$) to approximate the data matrix ($Xp \times n \approx Up \times rSr \times n$). PCA's orthogonality helps handle multicollinearity. Further, to manage multi-source data integration, Joint and Individual Variation Explained (JIVE) extends PCA, enabling simultaneous analysis of high-dimensional data from various sources ($Xk = UkS + WkSk + Rk$). JIVE distinguishes between joint and individual variation across data sources, ensuring unique variation capture. Our methodology differs from traditional differential gene expression analysis by focusing on the significance of principal components from paired data sources, rather than comparing independent groups for each gene.

Tools used

All of the analysis for this project was performed in R version 4.1.1 (R Core Team, 2021). The data was queried using the R package TCGAbiolinks (Colaprico et al., 2015) and the metadata was retrieved using the R package GenomicDataCommons (Morgan and Davis, 2021). The JIVE algorithm was performed with the R package r.jive (O'Connell and Lock, 2016). The analysis used the March 29, 2022 update to the GDC Data Portal.

Quantitative / Data Analytics

We first sought to describe the distributions of the RNA-seq and methylation data. Because there were a total of 60,660 genes and 19,315 methylation sites, we cannot look at the distribution of each gene and site individually, so instead we considered the distribution of their means. For the RNA-seq data, 2,788 genes had a mean of zero because they only contained counts of zero. We omitted these genes when considering the distribution of the means in Table 1 and Figure 1.

	Min	Q1	Median	Mean	Q3	Max
Normal Sample	0.0001	0.0262	0.2183	19.3237	3.7760	26191.4
Tumor Sample	0.0000	0.0286	0.2237	17.2879	3.0799	35999.5

Tab. 1: Summary statistics for the mean TPM of each gene for normal and tumor samples.

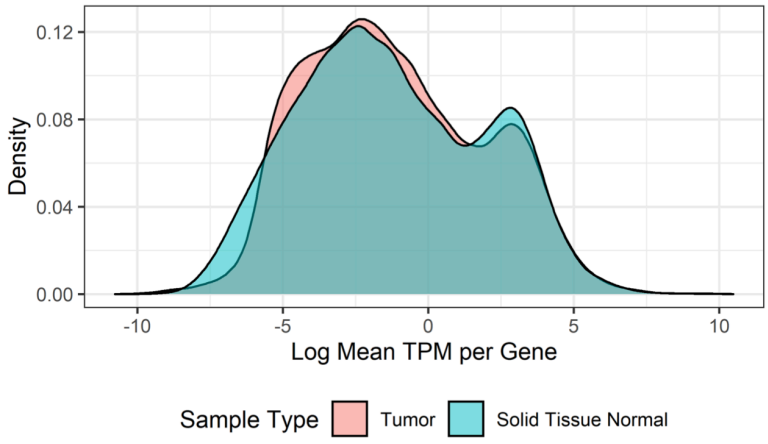

Fig. 1: Distribution of the log mean TPM of each gene for normal and tumor samples.

From Table 1, we see that the normal and tumor samples both have minimum values very close to zero. The first and third quartiles, as well as the median, are also very similar. However, the maximum is noticeably larger for the tumor samples. Looking at Figure 1, we see that the distributions for the two sample types are relatively similar.

Because a beta value represents the proportion of methylation at a particular locus, all beta values fall between 0 and 1. We see from Table 2 and Figure 2 that the distributions of normal and tumor samples are relatively similar. The median beta value of the normal samples is 0.06 and the median beta value of the tumor samples is 0.07, so for both sample types over half of the loci are less than 10% methylated.

	Min	Q1	Median	Mean	Q3	Max
Normal Sample	**0.0081**	**0.0334**	**0.0591**	**0.2293**	**0.3450**	**0.9909**
Tumor Sample	**0.0094**	**0.0405**	**0.0735**	**0.2317**	**0.3765**	**0.9840**

Tab. 2: Summary statistics for the mean beta value of each locus.

We also can observe from Figure 2 that the distribution of mean beta values is right-skewed for both normal and tumor samples. The normal samples do have slightly more loci with high beta values, but this is arguably the only observable difference between the two densities.

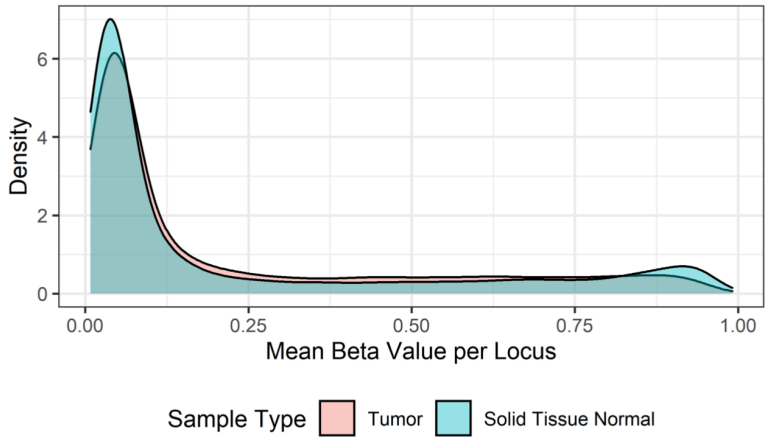

Fig. 2: Distribution of the mean beta value per locus for normal and tumor samples.

The JIVE algorithm was used to generate loadings and scores that could describe the variation in the RNA-seq and methylation data and determine which genes have the strongest influence on the difference between normal and tumor samples. Recall that JIVE selected 1 principal component for the joint structure, 1 principal component for the individual RNA-seq structure, and 20 components for the individual methylation structure.

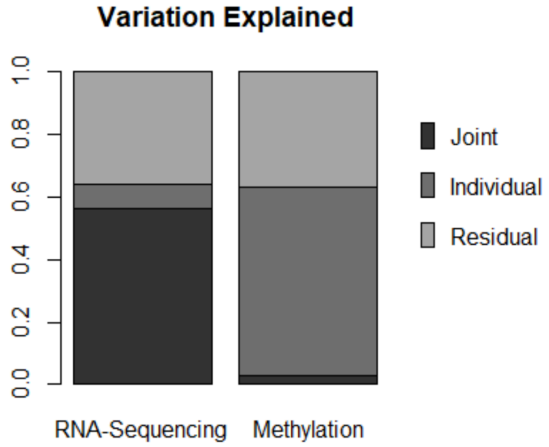

Fig. 3: Amount of variation explained by joint and individual structure of the data.

Looking at Figure 3, we can see how much variation is explained by the joint structure relative to the individual structure for each of the two data sources. For the RNA-sequencing data, about 55% of the variation is explained by the joint structure, 10% of the variation is explained by the individual structure, and the remaining 35% is residual noise. For the methylation data, about 5% of the variation is explained by the joint structure, 60% of the variation is explained by the individual structure, and the remaining 35% is residual noise. We can look more closely at the relationships among the untransformed scores of each type of variation using Figure 4. Because the JIVE algorithm centers the data sources by subtracting the mean of each row, all of the components are centered at zero. We see that all of the scores are relatively close to zero, probably because the matrices are scaled by their Frobenius norms.

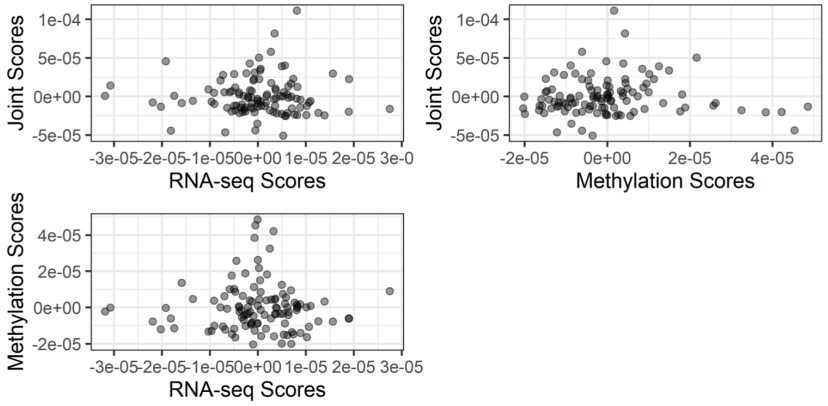

Fig. 4: Examine patterns in the scores. The axes represent the first principal component for each source of variation. Alpha shading is used to make masses of points easier to visualize.

Looking at all three plots, we see that most of the scores fall between -2e-05 and 2e-05, with a few scores falling further away from zero. While the joint component and RNA-seq scores have a relatively equal number of points trailing off in both directions, the methylation component has more outlying points in the positive direction. The left two plots both have a mass of points in the center, while the right graph of the methylation scores against the joint scores appears to be more spread out.

Results

Many differential omics studies with TCGA data have to choose between loss of information or ignoring the pairing of tumor and normal tissue samples. The approach in this paper aims to balance the loss of information while properly accounting for the paired observations. Taking the difference of the normal and tumor samples before performing the JIVE algorithm adjusts for the concern that samples taken from the same patient are most likely more correlated than samples from different patients. One complication of this approach is that it necessitates the imputation of missing normal samples to avoid information loss. We ended up using a naive imputation method in this paper, as the JIVE imputation performed slightly worse. Even this naive method produced results that were highly correlated with the true values in our cross-validation study. If a more informative imputation method could be found that effectively incorporates information from other sources, then this approach would be more accurate.

The purpose of running the JIVE algorithm was not only to determine what joint and individual structure exists within the RNA-seq and methylation data, but also to see which genes and loci contribute most strongly to explaining the variation in the differences among gene counts and methylation beta values. We did this by looking at the top ten loadings for the first principal component for each source of variation. Due to some overlap between which genes had the largest loadings in the joint and individual structure, there were 16 total

genes considered in the results. One interesting aspect of the results was that over half of the genes under consideration could be grouped into a few broad gene classifications.

Gene Name	Ensembl ID	Loading
IGKC	ENSG00000211592.8	0.9204
IGHG1	ENSG00000211896.7	0.1996
IGHA1	ENSG00000211895.5	0.1411
FTL	ENSG00000087086.15	-0.0627
MT-CO2	ENSG00000198712.1	-0.0546
IGLC2	ENSG00000211677.2	0.0523
IGKV3-20	ENSG00000239951.1	0.0461
MT-ND4	ENSG00000198886.2	-0.0457
IGKV4-1	ENSG00000211598.2	0.0419
MT-RNR2	ENSG00000210082.2	-0.0398

Gene Name	Ensembl ID	Loading
FTL	ENSG00000087086.15	0.6069
S100A9	ENSG00000163220.11	-0.5953
IGKC	ENSG00000211592.8	0.3320
S100A8	ENSG00000143546.10	-0.1685
IGHG1	ENSG00000211896.7	0.1255
PI3	ENSG00000124102.5	-0.1105
MT-CO1	ENSG00000198804.2	0.1003
KRT14	ENSG00000186847.6	-0.0879
KRT6A	ENSG00000205420.11	-0.0871
MT-ND4	ENSG00000198886.2	0.0839

Six of those 16 genes are immunoglobulins: IGKC, IGHG1, IGHA1, IGLC2, IGKV3-20, and IGKV4- 1. These were all positively correlated with the first joint component. The largest loading for the joint component was IGKC, which was also found in the top ten loadings for the first RNA-seq component. IGHG1 was also positively correlated with the first RNA-seq component. Four of the 16 genes are mitochondrial genes: MT-CO2, MT-ND4, MT-RNR2, and MT-CO1. Mutations found in mitochondrial genes can lead to an inability to produce enough cellular energy, to the extent that certain physical traits may be affected and a person may develop diseases (Chinnery and Hudson, 2013). MT-CO2, MT-ND4, and MT-RNR2 are negatively correlated with the first joint component, while MT-CO1 and MT-ND4 are positively correlated with the first RNA-seq component. FTL, the fourth gene that is found in the top ten loadings for both the first joint and first RNA-seq component, is the abbreviation for ferritin light chain. FTL functions as an iron storage protein and can affect the uptake and release of iron (The Human Protein Atlas, 2008). It is the largest loading for the first RNA-seq component. All of the other remaining genes are only found in the top ten loadings for the first RNA-seq component. KRT14 and KRT6A are keratin genes, S100A9 and S100A8 are calcium binding proteins, and PI3 stands for peptidase inhibitor 3. Meanwhile, the genes associated with the top ten loadings for the first methylation component are not part of any of the same classifications as the genes for the joint and individual structure. There also does not appear to be a clear pattern amongst these ten genes.

The JIVE algorithm on the differential expression and differential methylation matrices aims to maximize the variation explained in these matrices in as few components as possible. High variation in a gene does not necessarily indicate that the gene is differentially expressed or methylated. To verify whether this method is useful for identifying candidate genes for differential expression, we conducted post-hoc paired t-tests for the 16 genes identified in the previous section. A Bonferroni correction was used to account for multiple comparisons, so the significance level for this test was 0.05/16 = 0.003125. It is worth noting that as post-hoc tests, these tests are likely subject to bias since the tests are conditional on the genes contributing heavily to the overall variability in the differential expression or methylation matrices.

Of the 16 genes, there were significant differences between the normal and tumor samples in 13 of them. This suggests that our approach can be useful for identifying candidate genes for differential expression and differential accessibility. The three genes that were not significantly different were IGHA1, MT-CO2, and MT-RNR2, all of which were in the top ten loadings for the first joint component.

Additional Resource

For specific codes from this research project that are potentially permit for release, please reach out to: juyuan_li@hsph.harvard.edu / juyuanli@mit.edu / lij193@miamioh.edu