**Report**

**Background:**

Metabolite analysis involves multiple approaches, but none is as powerful as Mass spectrometry(MS). Molecules due to their inherent mass(m) and charge(z) have a unique mass to charge (m/z) signature to each molecule. It has a wide base of applications such as find and compare various molecules in a sample, check interaction of proteins, check isotopes in samples, to elucidate the unknown samples, etc. MS can be done based exploiting various parameters and methods used to detect the m/z ratios like various ionization techniques such as MALDI, Electron Spray Ionization, etc. Mass variations can be detected using techniques such as Time of flight, Orbitrap etc. These techniques used with further different detectors can give m/z ratios. Pre-processing of samples can help in purification and separation of samples and reducing noise and is usually some form of chromatography such as Liquid Chromatography(LC), High Performance Liquid Chromatography, Capillary electrophoresis etc.

The given samples in the question were of LC-MS, meaning samples were first purified using Liquid Chromatography and then MS was done on them. The data was analysed using EI MAVEN, which is an open source mass spec data processing software with many automated features for analysis. The processed file which has the sample details has the following information:

rT- retention time --- gives the time the sample was in the mass spec, retention changes according to the m/z ratio thus gives the unique identifying signature. This rT is cross checked with known compounds and the sample identity are then assigned.

m/z – mass to charge ratio of the molecule another unique signature.

Other quality parameters

The given experiment was a pulse chase experiment, where a labelled C13 probe(glucose) was given and chased for 2 hours, with collection times being- 0, 1, 3, 5, 15, 45, 120 and 240 mins. The samples analysed via LC MS and compounds results for Glycolysis and TCA cycle were provided.

Glycolysis is a metabolic pathway that metabolizes Glucose to give two pyruvate molecules(in aerobic conditions), which can then be fed to other pathways. It breaks down glucose to produce 2ATP molecules and 2 NADH molecules. NADH2 is further converted to ATP by electron transport chain. ATP is considered energy currency for the cell.

TCA or Tricarboxylic acid pathway, is a series of reactions that convert tricaboxylic chains to other forms, and releases ATP and NADH and FADH2 in the process. It takes AcetylCoA (One pyruvate gives 2 AcetylCoA; AcetylCoA is also produced in other reactions) (and many other) molecules as its input and processes the tricarboxylic carbon chains. It is a major part of energy retrieval via aerobic respiration. Like NADH and FADH2 too is further processed in the electron transport chain and produce ATP. Each cycle of TCA cycle produces 12 ATP (considering NADH and FADH2 get converted into ATP eventually).

Thus from Glucose to TCA 36 ATP are produced. Both these pathways form a very important backbone for the energy production and hence survival of the cell.

The problems given were solved in **programming language R (file= Analysis.R)**. In R, majorly “dplyr”, “tidyr”, and “reshape2” packages were used for data wrangling. And “ggplot2” and “ggbiplot” packages were used for graph visualization. For natural abundance correction package “accucor” was used for isoptopolgues abundance correction. The usage details are given in the following sections.

**Global Analysis**

The first step was a global analysis. To do this it was suggested to perform a PCA for pool totals of all given samples. PCA would broadly tell us if there is a difference between the samples, if so, then which samples would exhibit these differences.

In order to find the **pool totals and fraction enrichment**,

* The files were read with the following names:
  + metadata - sample\_metadata.csv
  + maven\_data - Maven\_processed.csv
* the unnecessary columns were dropped, and only columns with sample names and compound names were retained.
* The rows were grouped according to the compound name, and rows of each of the column (according to the group) was added together to get the total and data frame called pool\_total was made.
* To calculate fractional enrichment columns, compound, note, and samples, were selected.
* This wide format data was converted into long format data, using “melt” function.
* The samples were given their Phenotype and Time by joining this dataframe with metadata dataframe. ---1
* Also, pool\_total dataframe was converted from wide format to long format
* The metadata was assigned to pool\_total converted dataframe ---2
* Dataframes 1 and 2 were then joined to have all values and metadata. Meaning, all information including measurements, phenotype, time and totals were now present in one dataframe—fraction\_enrichment
* Next to calculate fraction enrichment—column with measurements (value) was divided by Total column and multiplied by 100 and put in a new column called fraction\_enrich.

Two dataframes – pool\_total and fraction\_enrichment, were used for all analyses.

To plot the PCA pool\_total dataframe was used. In order to perform PCA the data has to normally distributed and more importantly, the variation amongst all the variables (samples in our case) need to same. If the variation is not taken care of a variable with huge values will lead to huge variations, where as a variable with low values will cause no variation. These variations can skew the Principal components being calculated. Thus we need to check if the data is normalized or not.

I checked the data if it is normal and in the same scale for all columns in pool\_total (graphs not saved but code present in the file). **None of the data were normal or to scale**. I then tried normalizing the data from 0-1. This normalized the variation, however, the scales for each variable were still different. Moreover, there was a clear skew of data towards the extremes, indicating a non-normal distribution. Thus I went for **log transformation of the data**. Many of the variables now showed normal distribution (qualitatively, not checked with statistics). I then calculated the PCA based on the log transformed data.

The PCA showed that most of the most of the variation can be explained by the first **two principle components(PC) (96.6%).** It was interesting to observe that **Mock vs. Viral treated samples segregated into two separate groups** along the second PC in a **PC1 vs. PC2 plot (PCAlogPC1PC2.pdf)**, suggesting clear differences in the two samples. To probe further, where these differences may lie, **PC2 vs. PC3 (PCAlogPC2PC3.pdf)** were plotted. This showed that samples at **time points 120 and 240 minutes aggregated together**, whereas earlier time points were clubbed together. Moreover, **the segregation of Mock vs. Viral was stark** in these axis, suggesting strong effects start to be seen at later time points. Since it is a pulse chase experiment, changes at later time points indicate cumulative total pool changes in the two conditions. This needs to further analysed.

To check if there was any **effect on any specific metabolite** the **pool\_total dataframe was transposed** and PCA was performed. It **did not show any significant enrichment (PCAlogTPC1PC2.pdf)** of any pathway specific measurements. The only metabolites showing strong vectors were either the end products, the products in the initial stages of the pathways or rate limiting steps.

**Specific analysis**

In order to perform the specific analysis, line graphs of **pool totals vs Time (totalgraph.pdf)** and **fractional enrichments vs. Time(fractiongraph.pdf)** of each of the metabolites were plotted for Mock vs. Virus. This was achieved by grouping the metabolite with infection (compound + Phenotype – for pool totals, note + Phenotype – for fractional enrichments) by using grouping parameter in ggplot2. Furthermore, to separate out each plot from other, as per the metabolite (compound), the graphs were facetted according the metabolite(compound).

**The virus infection seems to affect the Glycolysis more as compared to TCA cycle. The viral infection seems to inhibit glycolysis. These cells further show compensatory increase in TCA cycle.**

Since Glycolysis usually takes a few minutes as compared to TCA which takes a few hours, thus whatever measurements are observed in the **initial time points (0-45 minutes) can be interpreted due to glycolysis** and **later times points of (120-240 minutes) can be interpreted due to TCA cycle**.

We observe that most of the products of glycolysis such as **glucose 6 phosphate, fructose-1,6-bisphosphate and pyruvate have lower pool total** at the initial time points. This indicates that glucose uptake is lower in the virus infected cells. These level changes are **also reflected in fraction enrichment** of glucose-6-phosphate and D- gluconate. Only ~75% of unlabelled glucose-6-phosphate and ~60% gets converted to C13 labelled glucose-6-phosphate by 15 minutes whereas more than 90% of glucose-6-phosphate and more than 70% of D gluconate is converted in Mock conditions.

The effect of reduced glycolysis is also translated to **reduced energy** in the cell which can seem by the **reduced levels of ATP and NADH**. Ideally it would be best to take NAD+/(NADH|NADPH) and ADP/ATP ratios to comment on the energy levels, but qualitatively the ratio seems higher in virus infected cells. Indicating lower energy status at the initial time points.

At later time points TCA cycle seems to play a compensatory role in the viral infected cells as we see increase in TCA cycle metabolites. The major metabolite of the TCA cycle starting from **Acetyl CoA, succinate, fumarate, malate, all increase their total pools and their fractional pools of C13**. An increase in aspartate works as a readout for oxaloacetate. Aspartate also feeds into the TCA cycle thus increasing the input for TCA Cycle. The increased total pool and fractional pool of labelled C13 again indicate an increase in TCA cycle at later points. Other read of TCA cycle such as, Moreover, these TCA specific substrates also show a slight spike during the initial time points, indicating cells behaviour to compensate for the loss in glycolysis by increasing TCA cycle.

To compensate for reduced pyruvate, the viral infected cells seem to **use the other metabolites to feed into the TCA cycle more**, which can be observed by increased levels of **isoketovalerate (ketogenesis), 4-aminobutyrate (GABA shunt), glutamate, glutathione, glutathione disulphide,** but not via Folic acid (folate), tryptophan.

The cells seem to synthesizing more nucleotide UDP, UTP, UDP-D-glucose. We can speculate that the cell may be synthesising more RNA, probably for virus production.

Moreover, energy status of the cell is also majorly maintained by the TCA cycle as the levels of ATP and NADH increase at later time points which can be seen by increased levels of pool totals and fraction pools at later time points.

**Conclusion**

To conclude, Viral infection in the given cells leads to a decrease in glycolysis and a compensatory increase in TCA cycle. The inputs to the TCA cycle are increased via other pathways to compensate for the loss seen by the glycolysis. In the long run the cellular energy levels are maintained by the TCA cycle.

**Bonus Question**

There exists a naturally occurring isotopes of C, N, H etc. These are comparatively rare and are thus rarely found in the organisms. However, when performing pulse chase experiments, with labelled isotopes there can be certain metabolites these naturally occurring isotopologues can skew the data. This can specially be true in cases where the readouts are small thus even few molecules can disrupt the signal to noise ratio and thus lead to wrong interpretations. To overcome this problem, the probabilities of naturally occurring isotopologues are measured and the recorded samples are normalized to them. This correction is called as natural abundance correction.

To implement natural abundance correction, **I used Accucor package in R**. The package takes maven file as input and gives corrected data as another file as an output, which can then be used for further processing.

I used one of the metabolite (**AcetylCoA**) to demonstrate the effect of natural abundance correction. The csv file was manually made by pasting values from original as well as corrected sheets and was then analysed **(file= corrected\_data\_analysis.R)**.

The trends in the pool totals remained similar (**correctedtotalgraph.pdf**), however, many of the C13 labelled metabolites were now corrected for after natural abundance correction (**correctedfractiongraph.pdf**). It was observed that **20 Carbon (in Mock) and 3 Carbon (in Virus infected) labelled AcetylCoA could have been possibly due to naturally occurring isotopologues.** And thus the fraction enrichment of these metabolite changes. These changes can lead to change in the interpretation of the data.