Metabolites: The Development of the Potentially Valuable Medical Diagnosing Tool

**Introduction**

The doctors (referencing my father and his colleagues) have been saying they loved their jobs because desire the synonyms being "similar," diagnosing the disease/disorder often is so biologically individualism that medically pinpointing things is not boring.

The new medical journal, “Cold Spring Harbor Molecular Case Studies,” with its first issue released to public on October 1st, 2015 stated that their mission is to strengthen the “precision medicine” approaches. Instead of the hit-or-miss innovation discoveries, this journal wants to focus on evaluating the data collecting and data analyzing technologies and methods.

Historically, since about the 1950s (e.g. test strips for diabetes), medical scientists were aware that evaluating a precision diagnose cannot simply be done by looking at RNA and/or protein expressions. Measuring the expression of RNA and proteins have been the foundations of tracking possible biological function regulations. But we also have essential biological molecules with little to no RNA or protein features, catalogizing them as “metabolites.” For example, lipids, fatty acids, bile acids, amines, cationic, and polar molecules. Processing the expression levels of most metabolites, however, is quite complicated and not feasible, implicitly before the 21st century. Worse, the biological specimens’ data are prone to environmental impressions, both externally and internally. When handling possibly confusing data, it has been helpful to think, “That is just biology.”

I would not be surprised if the "that is just biology" attitude would progressively become less "acceptable" with the growing availably of biological metadata sharing networks. The network specified for sharing measurements of metabolite expression is at the <https://metabolomicsworkbench.org/> site. (The Metabolomics Workbench= MW.) I didn’t read enough metabolites-type studies, but I have the impression that while we do have most protocols and standards in place for cross-referencing, but there are still the challenges of how to process and analyze the metabolites-type of datasets.

The metabolite dataset, containing metabolomic information from 30 knockout mice, was selected for this final project. I have ignorantly picked this dataset because I have experience working with various knockout mouse models. I now realized that I have opened a big can of "innovation" worms to play with. For my final project the "data descriptor" journal article is referenced using the metabolomic workbench ID number, ST001154.

The “Cold Spring Harbor Molecular Case Studies” physician journal series appears to yet notice the ST001154 study's advocating of a potential chemical compound classify benchmarking approach. ST001154 demonstrates how the parameter adjustable ChemRICH software can help sort out the meaning of metabolites expressions in correlation to known (and theoretically unknown) biological mechanisms.

Be noted that the authors collected a minimum n=3 per gender group from 30 knockout mice lines. I just want to say that in my firsthand experience, the n=3 sample size per group have been accepted as a publishable "suggestive conclusion" in most research journals. Would the growing availably of sharable metadata exchanges eventually make the n=3 sample size's data less valuable, maybe even an embarrassment 100 years later? Let's start calling n=3 the "preliminary exploration" quantity to save our faces.

In order to identify metabolites and their expression, the ST001154 study explained the process of cleaning up the raw data, cross-referencing filtering processes, then how their ChemRICH software can help with the chemical compound classifications. This ChemRICH process's R programmed files are available on the ST001154 author's GitHub, https://github.com/barupal/ChemRICH.

There are plenty of things we can do with ST001154's dataset. For this project, I am having things simple, by making my own subset of data from the ST001154 study's monstrous dataset. I want to try the phenotypes of three groups, two knockouts and one wildtype control.

**Methods**

ST001154's raw, processed, and exported datasets are available for anyone to download copies of. I used the ST001154's paper's supplement tables. This dataset also appears to be available on Kaggle and GitHub.

I suspect that the ST001154's Kaggle moderate grade of 7 out of 10 was because the datasets were not user-friendly. The headers could have been labeled better (e.g. are those mice ID numbers or chemical ID numbers?). And the series of data could be organized better (e.g. the tables with mice ID numbers for six animals from 30 knockout lines were columned in no particular order.) It did take me a while to figure out what I was looking at and locate individuals' data from random columns for my subset.

I initially was confused by the built-in metabolomic data analyze in the MW's [Statistics Toolbox for Study: ST001154](https://metabolomicsworkbench.org/data/stats_toolbox.php?STUDY_ID=ST001154&ANALYSIS_ID=AN001941) Suggesting that an artificial intelligent machine probably did some "standard" work. Some results were jamming data from all 30 knockout groups together as one group, but I now realized the pie charts are probably showing the chemical compounds groupings, which can be used to identify possible active mechanism pathways. As that was, the toolbox's settings are focusing mainly on the expression of chemical phenotypes. The 30 knockout mice lines' non-chemical-compound type of phenotypes and gender separations were not factored in at MW. At best, the site's "toolbox for study" could help with preliminary exploration. As of the early 21st century, we still need people's comprehensive thinking powers to provide better, definable results.

For my small Rscript analyze with the gender factored in, my datatable contains two knockout lines and wildtype groups of both genders. Information is sampled from ST001154's supplement excel file as the following:

|  |  |  |
| --- | --- | --- |
| Data source | Which data | Data subset picked |
| TableS15\_SampleMeta | MouseID, genotype, gender | Genotype="Null" for wildtype control. And two random knockout lines…Pebp1 and Idh1, simply as those are the first two groups on the list. |
| TableS1\_Pheno | PhenotypeID, phenotype | First 33 phenotypes out of so many, because most individual mouse has measured data for those top 33. |
| TableS2\_PhenoData | Selected individuals from TableS15 | Null, Pebp1, and Idh1 |

Which of the Metabolomics Workbench's toolbox approaches would be good choices for my subset of samples? (My subset file, "ST00114\_subset," is uploaded to myCourse with this document.)

The MW's built-in statistics of chemical compound results shows me that some statistic methods may be useless. Then there's this linear discriminant analysis (LDA) approach looking interesting enough (see Figure 1). Let's see how LDA measures up for non-chemical phenotypes (Figure 2). Then (maybe) add the gender separation factor (Figure 3)?

**Results**

(pending)

**Conclusion**

Today's acquitting and processing technologies has been becoming sensitive, now enough to detect the growing list of multiple chemical compound types and be able to measure their expressions. My understanding is that using the chemical classification process could improve the metabolites analyzing protocol for finding likehood biological answers.

The ST001154's article closed with this statement: "We foresee this dataset’s use in developing next generation bioinformatics as well as in teaching courses for metabolomics and as a test case for benchmarking software [ChemRICH]."