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 imagine that 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. 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 teach myself about. For my final project the "data descriptor" journal article is referenced using the metabolomic workbench ID number, ST00114.

The “Cold Spring Harbor Molecular Case Studies” physician journal series appears to yet notice the ST00114 study's advocating of a potential chemical compound classifiy benchmarking approach. ST00114 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. I can imagine that the growing availably of sharable metadata exchanges could eventually make the n=3 sample size's data less valuable, logically being seen as "preliminary exploration" quantities.

In order to identify metabolites and their expression, the ST00114 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 ST001144 author's GitHub, https://github.com/barupal/ChemRICH.

I want to better understand the developments of metabolite measuring method, but the process is quite complicated with so much to learn about. That would take me at least a few weeks with my current R programming skills. Thus, let's keep things simple by familiarizing myself with the end-result basics by analyzing a subset of knockout phenotype differences from the ST00114 study's monstrous dataset.

Methods

ST00114's raw, processed, and exported datasets are available for anyone to download copies of. I used the ST00114's paper's supplement tables. This dataset also appears to be available on Kaggle and GitHub.

I think the ST00114'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' being columned in no particular order.) It did take me a while to figure out what I was looking at and locate the subsets I wanted.

The Metabolomics Workbench's [Statistics Toolbox for Study: ST00114](https://metabolomicsworkbench.org/data/stats_toolbox.php?STUDY_ID=ST001154&ANALYSIS_ID=AN001941) has some seemingly useless analyze results, suggesting that an artificial intelligent machine did some kind of general standard work. Some results were jamming data from all 30 knockout groups together as one group. Other results look at the knockout lines in somewhat confusing ways. And this current toolbox's settings are generally ignoring specific factors, such as genders. At best, the site's "toolbox for study" could help with preliminary exploration. This shows that as of the early 21st century, we still need people's comprehensive thinking powers to provide 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 ST00114's supplement excel file as the following:

|  |  |  |
| --- | --- | --- |
| Data source | Which data | Data subset picked |
| TableS15\_SampleMeta | MouseID, genotype, gender | Genotype="Null" for wildtype. 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 because those generally have all measurements for most individuals. |
| TableS2\_PhenoData | Selected individuals from TableS15 |  |

Which of the Metabolomics Workbench's toolbox approaches would be good choices for my subset of samples?

Results

Conclusion

In the days before the data acquitting and processing technologies became sensitive enough to detain differences of multiple molecule types and expressions of the individualism samples, the testing for metabolomic expressions had have been limiting. Now, with today's technology and resources, the current presenting challenge is improving the metabolites analyzing protocol for finding likehood biological answers.

The ST00114'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]."