BRIDGE NEWSLETTER

BIOSTATISTICS **R**ESEARCH & **I**NVESTIGATION **DIGE**ST

SPRING 2021

THE UNIVERSITY OF TENNESSEE
HEALTH SCIENCE CENTER.
BIOSTATISTICS

INSIDE THIS ISSUE

FACULTY/STAFF

Chi-Yang Chiu
Assistant Professor

Hyo Young Choi
Assistant Professor

Gregory Farage
Postdoc

Trish Goedecke
Staff Statistician

Tristan Hayes
Consulting Manager

Tamekia Jones
Associate Professor

Mehmet Kocak
Associate Professor

Xioayu Liang
Assistant Professor

Śaunak Sen
Division Chief, Professor

Fridtjof Thomas
Professor

Elizabeth Tolley
Professor

Jim Wan
Professor

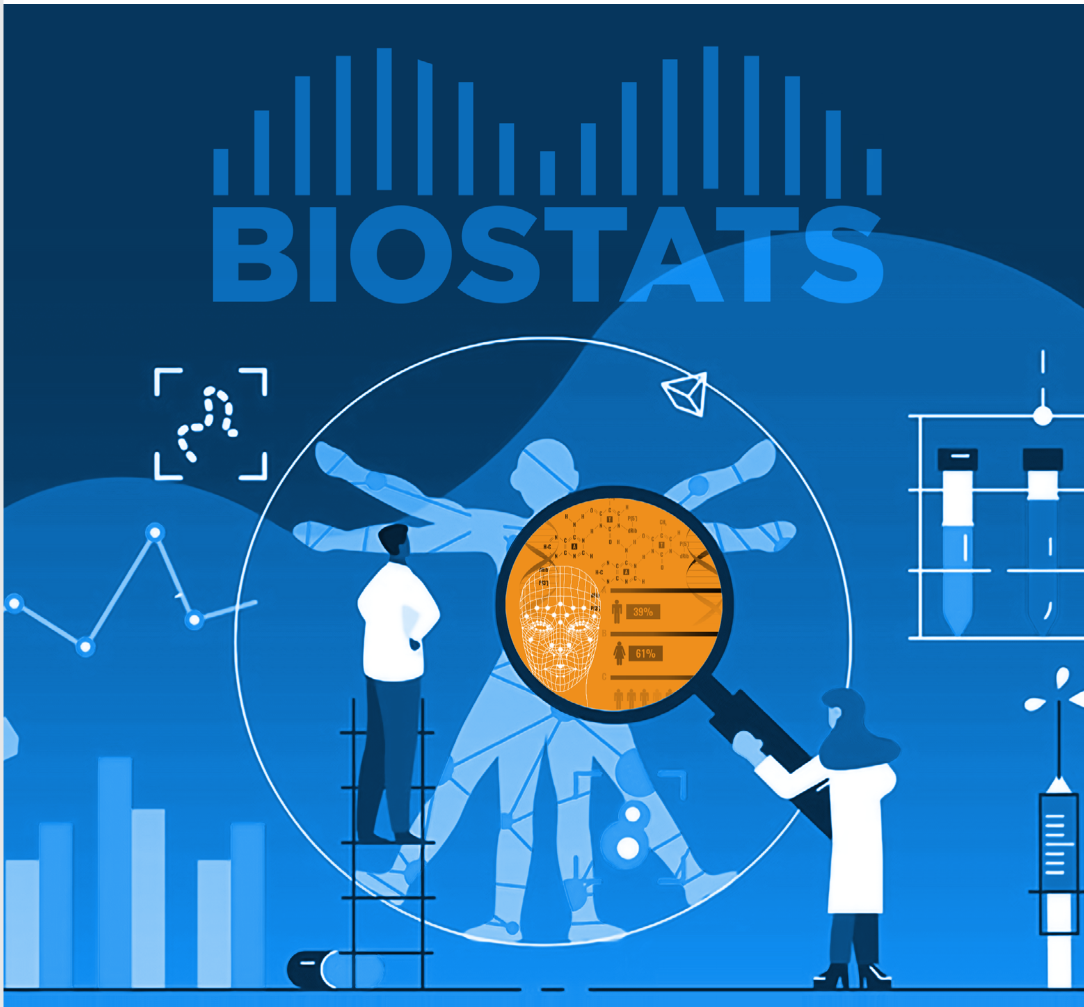
Rongshun Zhu
Staff Statistician

We are pleased to present the first Biostatistics Newsletter: BRIDGE (Biostatistics Research & Investigation DiGEst). We hope that this newsletter will help build a bridge between us and other researchers, contributing to improved mutual understanding and collaborations. We will try to give you, the reader, a flavor of what we do, what our field has to offer, and the people who make up the Division of Biostatistics.

In this issue, we feature two interviews, one with our newest faculty member *(Welcoming our new faculty member Xiaoyu Liang)* and another with our seniormost faculty member *(Faculty spotlight: Tea time with Dr. Tolley)*. On the research side, there is a note on a newly developed software package based on a new method for genetic analysis of high-dimensional traits developed by one of our postdoctoral scholars *(Introducing Julia package FlxQTL.jl written by Hyeonju Kim)*. For R users and those dealing with large data, there is a nice article on the role of data orientation (whether your spreadsheet is very long or very wide) on how quickly you can read the data, even though theoretically you can go from one format to the other by rotating the spreadsheet (The impact of data orientation and data type on I/O speed).

I thank the newsletter committee (Hyo Young Choi, Tristan Hayes, Chelsea Trotter, and Patricia Goedecke) for putting together this inaugural version. Their dedication, insight, and vision shine through in the product. We welcome your feedback and suggestions for the future. Do you have a question about our division, how we work, or how we can improve your research? Please drop us a note!

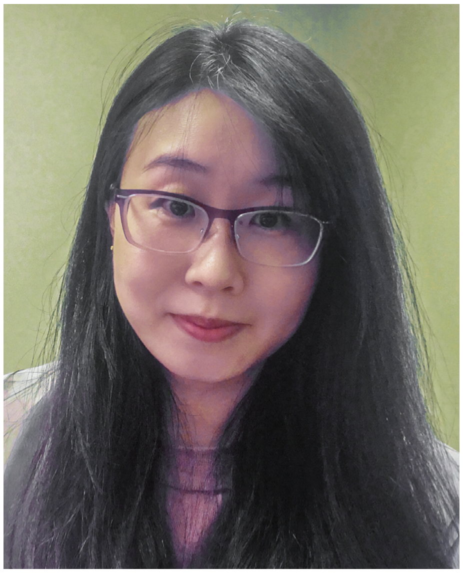
Śaunak Sen
Chief, Division of Biostatistics



INTERVIEWEE:
DR. XIAOYU LIANG

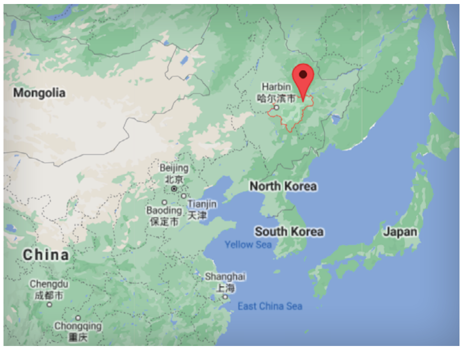
Interviewer: Tristan Hayes

TELL US A LITTLE BIT ABOUT YOURSELF?



I am from Harbin, China. I got my bachelor's degree in statistics in China (2013) and came to the US to study biostatistics. At first I did not plan to become a faculty member, but then I did a teaching assistantship and loved it. I also wanted to do analysis with applied data. Then I found a post-doc position at Yale where I could focus on my research interest in DNA methylation.

WHAT SHOULD WE KNOW ABOUT YOUR HOMETOWN (HARBIN)?



My hometown is in NE China, the coldest place in China. When I moved to the US at first it was to Michigan which was almost the same weather! I am not good at driving in snow. Harbin is close to Russia and we have lots of Russian food and Russian beer. Harbin is also famous for an ice sculpture park in winter. I like Memphis' weather a lot better, but people told me summer here is terrible so I will see.

WHY MEMPHIS?

When I was on my job search, I received several offers, but I felt best with my UT interviews. When I asked questions of people, even the chair, they responded quickly. I also liked the fact that Memphis has no traffic. When I lived in New Haven, CT, the traffic was awful! It is a lot better here.

WHAT GOT YOU INTO BIOSTATISTICS?

My bachelor was in statistics but I wanted to do something applied. I connected with a researcher in Michigan who was looking for a research assistant. For me biostatistics is a kind of applied statistics. This is why I came to the US for a degree in biostatistics.

WHAT DO YOU FIND MOST CHALLENGING ABOUT BEING A STATISTICIAN?

I find that for me, especially in BERD Clinics, the statistical problem is not hard. What is hard is understanding the background of a client or collaborator and what they are really looking for.

SPEAKING OF COLLABORATORS, WHAT ARE YOU LOOKING FOR IN A RESEARCH COLLABORATION?

In our department, my research interests are in two main areas: one is single cell RNA-seq analysis; and the other is DNA-methylation. Currently I am looking for more collaborators in DNA-methylation.

WHAT'S ON YOUR WISHLIST FOR THE NEXT 5 YEARS AT UTHSC?

For my research, I want to publish more papers and complete more collaborations, especially within our university. I want to collaborate with different people. For teaching I plan to develop a new course. I also want to be provide good service to the community through the BERD Clinic. Lastly, I want to organize a seminar.

MOVIES, COOKING, PETS

You mentioned cooking. What's your favorite lockdown meal? Least favorite Food?

Traditional Chinese food, like dumplings and Kung-Pao Chicken. I like spicy food so all my recipes are spicy. I also like Mexican food- those restaurants have spicy food and good portions. I don't like donuts. As a young grad student in the US, I didn't have time to cook. Donuts are so tasty and quick so I ate them almost every day – until I couldn't anymore.

LAST QUESTION: IF YOU COULD HAVE ANOTHER JOB FOR JUST ONE DAY, WHAT WOULD IT BE?

Dog walker, without a doubt.

TEA WITH ELIZABETH TOLLEY

QUESTION: COULD YOU TELL US WHAT IT WAS LIKE ARRIVING IN MEMPHIS AND ON THE UT CAMPUS FOR THE FIRST TIME? WHERE WERE YOU COMING FROM AND WHAT WERE YOUR IMPRESSIONS?

It is very difficult to explain what UTHSC looked like when I arrived. It has been more than 35 years. There are buildings here now that weren't then, and buildings that have gone; and others that have changed dramatically.

When I came to UTHSC, it was called the "University of Tennessee, Memphis," then the "University of Tennessee, Center for the Health Sciences" and finally its current name – though it always has been the same place. However, the department that I was recruited into no longer exists: the Department of Biostatistics and Epidemiology. My title therefore was and still is, "Professor of Biostatistics and Epidemiology." It was a stand-alone department, separate from all of the colleges; there were four stand-alone departments at that time. Like IT, the original department was meant to serve all of the colleges. How it's set up today is very different and our departmental mission has also evolved.

At first, Baptist Hospital Central was located on the south side of Madison Avenue across from "the 9s," which were also owned by Baptist. Now, during recruitment we house candidates in hotels downtown; when I came for my interview, I was hosted in the 910 building, which housed residents' rooms: a twin bed, desk, shower and bath. My big impression during that interview trip was the shower, which ran alternately either freezing cold or scalding hot! It was an old building at that time, too. Elements of "the 9s" are still the same, but the renovations to the interior have transformed them completely.

This interview visit was not my first trip to Memphis. From a post-doc position in Raleigh at North Carolina State University, I went to Battle Creek Michigan, to work for the Kellogg Corporation as a corporate statistician. I had been in Memphis on a start-up for a new cereal at the Kellogg plant on Democrat. I spent a week on that project; so I knew something about Memphis, but with a limited perspective. I was there with Kellogg not quite a year, and during that time realized being a corporate statistician was not my goal for my life; I preferred academia.

During my interview visit, I met a lot of people, most of whom are no longer here. Only Bob Burns is still a member of department; he was doing a fellowship in Preventive Medicine at the time.

QUESTION: ARE THERE PEOPLE WHO STOOD OUT TO YOU AS PARTICULARLY HELPFUL DURING YOUR EARLY YEARS, WHILE YOU WERE GETTING YOUR FOOTING?

People I met during my interview visit led me to take this job. Grant Somes was chair of the department of biostatistics and epidemiology, and had been so for about one year. He and his wife Brenda made me feel very welcome, and introduced me after I joined the faculty to many, many people. I felt comfortable to come here because of the two of them.

The work was similar to what it is now: right after arriving, I was assigned clients to whom I offered statistical consulting. As new clients came, they were referred to me. I quickly began establishing relationships all over campus, with all kinds of projects. It was an exciting job, and a great way to meet people on campus.

When I joined the faculty, our department was housed in four rooms on the mezzanine of the library. Offices opened to the mezzanine; everyone in the mezzanine and on the main floor of the library below could hear our conversations. The rest of the offices on the mezzanine were occupied by staff and faculty in the Department of Computer Science.

While offices have not changed very much, computers and computing power definitely have. I was asked, what type of computer do I want? Originally, I had a state-of-the-art Dec terminal; I asked for stand-alone MacIntosh, the first computer ever called a MacIntosh. It had 512K of RAM. SAS for desktop at that time didn't exist. SAS was mounted on a server in Knoxville. We sent programs to Knoxville via telephone modem. At first, the jobs didn't come back to my Dec or my Mac; they came to a mainframe computer, in a now non-existent building on Dunlap, near the GEB. A computer called a PDPA could mount reels of tape; it was also the only place on campus to get printouts, on rolls of sheets we called "green-bar paper." I'd have to walk from the library to that computer location to get printouts for clients.

QUESTION: IN WHAT WAYS HAVE YOU SEEN THE BIOSTATISTICS DIVISION DEVELOP, AND SUPPORT THE CAMPUS COMMUNITY?

My job description is still the same; but how I execute the job has evolved so impressively that my original position is almost unrecognizable.

Shortly after I arrived on faculty, we began adding people, both biostatisticians and epidemiologists. One of Dr. Somes' main goals was to hire faculty with PhD training in epidemiology. He began the dual track within the small department. From four little offices we grew to use more than half of the mezzanine of the library. Back in '97 or '98, the opportunity came for our department to merge with Department of Preventive Medicine. At that point, Karen Johnson was a junior faculty member in Preventive Medicine. When we merged, we merged a very well-funded and renowned Department of Preventive Medicine with our very strong and active Department of Biostatistics and Epidemiology; we then moved together to the fifth and sixth floors of the Pauline building.

Before that merger, we developed a graduate program leading to a master of science degree in epidemiology. Before this program started, I had unofficial "students" – I actually trained colleagues, faculty members and residents, who were interested in biostatistics or epidemiology, so that they could perform their own analyses. Dr. Martin Croce was one of my first "students". When he joined the UTHSC faculty, he asked me to teach him how to run logistic regression and chi-square tests for 2x2 tables, t-tests, and other statistics to analyze his own data. Over the course of about a year, he would bring projects and we would go over them together. Soon, he was on his own doing routine analyses for multiple projects in the Department of Surgery. When he recognized a problem beyond his skill level, he always came back for advice and some advanced training.

Dr. Elizabeth Tolley greets a tall friend.

Once we had our master of science program, over the years I have helped formally train countless numbers of residents, faculty members, and students from various colleges within the university. Now we even offer courses online for distance students who can complete their entire degree in epidemiology without having to come to campus.

QUESTION: YOU'VE BEEN A PIONEER ON OUR CAMPUS OF PROVIDING ONLINE AND HYBRID LEARNING. WHAT ADVICE WOULD YOU LIKE TO SHARE WITH THOSE WHO ARE TEACHING ONLINE FOR THE FIRST TIME?

Get your feet wet now. When I started doing this, in 2008 or 2009, I said yes. When I was first approached, the College of Nursing had initiated a new aspect of their PhD in Nursing including distance students. Five distance students enrolled in my first class in my basic course in biostatistics for the health sciences I and II.

I was very concerned: would I be able to make statistics understandable to students via an online program? The first distance students were also working students – full- or part-time working students.

Would I be willing to make my course available by making a recording of the class? I said yes, then thought, "Nobody should take statistics online." I've since changed my mind. The way I have it set up, students from anywhere in the world can come in synchronously. I am doing every lecture completely from my desktop. Doing this requires me to run a number of software programs simultaneously. In order to use a whiteboard that also projects my desktop in a way that I can annotate documents in real time, I have a tablet (iPad) with specialized software on it too. The tablet makes a huge difference – all the difference in the world really. The tablet mirrors what's on screen; when you activate the tablet, it becomes the screen. I can draw in different colors: purple, blue, dark pink, and so on; I try to be sensitive of people with red-green color blindness. I haven't had complaints. And of course, I have an eraser. When I write on the iPad, I'll fill it up, then remind students: "Take a screenshot: it's going to go away."

I also have SAS running on my laptop and can show students how to program in real time. In addition, I have a high-quality piece of software to make high quality recordings, which I convert to mp4 files, and upload for students to view. Students can look at these recordings at their leisure, for asynchronous learning or for review. They access the videos through Blackboard.

Now I must say that it has taken me about ten years to get to the point that I am now satisfied with my online lectures. Why? Because… now, there is no difference, whether you're sitting in the classroom in the 4th floor of the Pauline building, or sitting in your office in Knoxville, or working at night in your home office in Colorado. You can have the same experience at night in your home that you would have in the classroom. The only difference is that you must send me an email or text if you have a question. It is not as congenial as being in a classroom; it's somewhat reassuring to be in the room with a lecturer. However, everyone is used to Zoom now and working and communicating remotely. Students' technical capabilities have grown as well.

I wouldn't want to go back to teaching the way it was back when I started as a young faculty member. Over the years I have become much better at explaining biostatistical concepts and methods. Much of that improvement has come from teaching remotely and online.

QUESTION: WHAT ASPECTS OF RESEARCH COLLABORATION DO YOU MOST ENJOY? WHAT DO YOU LOOK FOR IN A POTENTIAL COLLABORATOR?

This is a harder question. I've collaborated with a lot of different people, and personalities. One characteristic that stands out is a client's ability to listen. When clients come to me or another statistician, they are asking for advice. Those who are willing to listen will benefit more from interacting with me or another statistician. We are trying to help them do better research. When they are set in their ways and just want to insist on a certain approach, that can limit what can be gained from our expertise. When clients value what we have to say as colleagues, we can work together to strengthen their research.

I've had many clients show their appreciation by including me as a coauthor on papers or by mentioning my assistance in an acknowledgement.

Generally, when a client involves a statistician in their project, they are asking for a statistician to contribute their ideas, interpretations, and help them craft a manuscript to appear in a more favorable light for reviewers. I value people who want to work with me, who want to improve their own research by collaborating with a statistician.

Finally, I learn a lot from our collaborators. I enjoy working with collaborators who teach me about their work.

QUESTION: WOULD YOU LIKE TO DESCRIBE A STUDY YOU'VE WORKED ON THAT STANDS OUT AS HAVING HAD A PROFOUND IMPACT?

Kudsk et al. 1992, “Enteral versus parenteral feeding. Effects on septic morbidity after blunt and penetrating abdominal trauma.”

It turns out the latter reduces morbidities and mortality dramatically. This was the first study to make this direct comparison. If you don’t put nutrition in the gut, the gut becomes static rather than active; activity apparently contributes to the healing. This paper has 1400+ citations, some of which are recent – it’s that important of a paper!

QUESTION: WHAT OTHER CAREER HIGHLIGHTS WOULD YOU LIKE TO SHARE WITH US?

There are several things I’ve done that brought very positive light on our department. I was elected to the Faculty Senate and later became president of the Senate in the 90s. We worked on the first of several major revisions of the Faculty Handbook. My contributions made our department visible.

Other highlights are more intrinsic: I’ve watched students over the years, as their careers developed. On this campus, they may have been fellows or residents or PhD students. They’ve gone on to other places and developed truly outstanding careers. One former student in our program in epidemiology is on the faculty at the U of Pittsburgh; he is doing outstanding research at an institution known for research. He has come back to give grand rounds lectures here at UTHSC.

Another student has a position with the Tennessee Board of Regents system; she works closely with THEC for higher education in the state of Tennessee, influencing what happens on multiple campuses.

One of the first surgery residents who completed a master of science in epidemiology has come back to UTHSC as a full professor. There are other graduates in the Department of Pediatrics and on the faculty of St. Jude Children’s Research Hospital.

We also have some former students on the faculty in the colleges of Pharmacy and Nursing – they shed a positive light on our department.

[1] Kudsk, K.A., Croce, M.A., Fabian, T.C., Minard, G.A.Y.L.E., Tolley, E.A., Poret, H.A., Kuhl, M.R. and Brown, R.O., 1992. Enteral versus parenteral feeding. Effects on septic morbidity after blunt and penetrating abdominal trauma. Annals of surgery, 215(5), p.503.

FlxQTL:

FLEXIBLE MULTIVARIATE LINEAR MIXED MODELS FOR STRUCTURED MULTIPLE TRAITS

By Hyo Young Choi

Dr. Hyeonju Kim, a postdoctoral fellow in the Division of Biostatistics at the Department of Preventive Medicine, UTHSC, has recently published an open source software called FlxQTL.

FlxQTL is a Julia package for quantitative trait loci (QTL) mapping that can test associations between genetic markers and multivariate traits. One emerging interest in QTL analysis is the investigation of structured multi-dimensional traits – such as measured in multiple environments, at multiple timepoints, under different treatments, or more intricately structured traits. Although multivariate traits provide rich information and are of great interest, implementation of multivariate traits analysis is challenging. This new software, FlxQTL, addresses this challenge while also offering faster and more flexible genetic mapping compared with other existing software.

FlxQTL is a multivariate linear mixed model-based QTL analysis tool that supports incorporating information from trait covariates, such as time or different environments. In other words, if phenotype (or trait) data have been measured in continuous time or latitudinal gradient, users can prudently (but optionally) select from base functions such as B-splines, wavelet, or Fourier to capture a trend by elapsed time or by continuous geographical locations. Dimension reduction in estimating parameters is consequently achieved when dealing with higher-dimensional phenotype data.

The package supports computation of one-dimensional and two-dimensional multivariate genome scans, visualization of genome scans, LOCO (leave-one-chromosome-out), computation of genetic relatedness (kinship) matrices, and distributed computing. A one-dimensional multivariate genome scan regresses multivariate phenotypes on one genetic marker. Likewise, a two-dimensional scan, or pair scan, regresses them on two genetic markers within one chromosome.

These scans can be implemented in parallel across multiple cores in one computer as well as in a high-performance computing (HPC) system to achieve significant reduction in computation time. Additionally, a genome scan with millions of markers for moderately high-dimensional traits is also possible. A genome scan can be performed if genotype probability is available; so FlxQTL is very flexible for multi-parental populations such as four-way outbred and heterogeneous stocks.

Another nice feature of the software is its beautifully written documentation (**senresearch.github.io/FlxQTL.jl/stable/**). The documentation provides a step-by-step guide for QTL analysis with concise sample code using an example dataset. This greatly helps readers quickly understand and easily reuse the examples – even for users who are unfamiliar with Julia. In addition, the documentation contains solid description on every single function, providing details on inputs, outputs, and mathematical or statistical concepts where needed.

Dr. Kim, the lead developer of FlxQTL, works in the lab of Dr. Saunak Sen, professor in the department of Preventive Medicine and chief of the Biostatistics division. Before working with Dr. Sen, Kim received her doctorate in statistics at the University of Arizona with a dissertation on ruin problems in economics and insurance. Dr. Kim’s research interest is in high-dimensional data analysis for large-scale data using optimization techniques, stochastic processes, and extreme value theory. During her postdoctoral work, she is focusing on developing a fast and flexible method in the context of multivariable linear mixed models for QTL mapping, in application to gene by environment detection and functional data. After she completes her fellowship, Dr. Kim will join Columbia University Center for Statistical Genetics as a Postdoctoral Research Scientist starting July, 2021.

THE IMPACT OF DATA ORIENTATION AND DATA TYPE ON I/O SPEED

By Chelsea Trotter

This article is adapted from an internal talk at Department of Preventive Medicine, Biostatistics Division.

THE MOTIVATION.

The first step to your data analysis project, before you do any analysis or exploration, is loading the data into computer memory, using a programming language. This step is usually straight forward, but if you have a large dataset, this process can be lengthy. People focus on algorithmic changes, to achieve performance gain, and often over look the impact of such a simple task. I ran into this issue in a project, where we were developing the backend of web service. When using web services, every second feels longer than necessary, (if you have stared at a spinning circle, you know what I mean). So, this pushes us to use the fastest methods for every step of our algorithm.

In this article, I will share with you what I have found. The code chunks in this article are all in R. Results might differ using different software and/or hardware. TL, DR:

- If there is only one type of data (as in, all columns have the same kind of data) in your whole dataset, it is much faster and memory efficient to read data in as matrix, rather than dataframe.
- If there are a lot of columns, and less rows, your IO speed is at the slowest. To mitigate this, you may pay a one-time cost of transposing your data and write it out. Then, for all future IO, read data in using the new file, and then transpose it. It is faster to read and transpose datathan it is to read original data in directly.
- A Dataframe takes up a lot more space than a matrix, with the same data. Use matrices whenever you can.

For our project, reading and writing the BXD data set, takes up majority of the time in the data cleaning process. We use r/qtl package to read in genotype and phenotype data. The shape of the data is short and wide, meaning we have lots of columns and few rows. We found that, **given the same data, if we pass them in as is, compared to passing them in as transposed, the latter version is much faster.** This is very interesting. Perhaps this is already known to you that wide data takes longer to read than long and slender data. I wrote a micro benchmark to discover what is the reason for long IO time.

THE MICROBENCHMARK SETUP:

“A microbenchmark is either a program or routine to measure and test the performance of a single component or task.” [1]

There will be 8192 (2^13) elements.

- Dimension for tall matrix is 8192 x 1,
- Dimension for wide matrix is 1 x 8192.

And we want to time how long the functions **read.csv** and write.csv take for tall matrix vs wide matrix.

```
# Comparing the reading and writing time of wide and tall matrix

tall = matrix(rnorm(8192),nrow=8192)
wide = matrix(rnorm(8192),nrow=1)

write.csv(tall, “tall.csv”, row.names=FALSE)
write.csv(wide, “wide.csv”, row.names=FALSE)

print(“Reading time:”)
system.time({tallr = read.csv(“tall.csv”)})
system.time({wider = read.csv(“wide.csv”)})

print(“Writing time:”)
system.time(write.csv(tallr, “tallr.csv”, row.names=FALSE))
system.time(write.csv(wider, “wider.csv”, row.names=FALSE))

print(“Reading transposed data, then transposing it.”)
system.time({
   wide_transpose = read.csv(“tall.csv”)
   wider_wt = t(wide_transpose)

})
```

Continued

```
[1] “Reading time:”
    user  system elapsed
    0.005   0.000   0.006
    user  system elapsed
    1.405   0.017   1.423
[1] “Writing time:”
    user  system elapsed
    0.007   0.000   0.007
    user  system elapsed
    0.437   0.092   0.529
[1] “Reading transposed data, then transposing it.”
    user  system elapsed
    0.005   0.001   0.005
```

There is a lot going on in this code block.

First, we created our tall and wide matrix. The tall matrix is 8192x1, and the wide matrix is 1x8192. Then we write out the matrix to file.

Now comes the time to benchmark read.csv and write.csv (code line 9 to 21).

As the output shows, reading time for the tall matrix (code line 10) takes 0.005 second, while the wide matrix takes 1.268 seconds(code line 11). (Your output may vary because of differences in hardware and randomness). The performance gap is jaw dropping.

For the case of write time, the timing tells the same story (code lines 14 to 15), a wide dataset takes longer than a tall dataset, only a little less dramatic.

However, if we read in the tall matrix, then transpose it (the same effect as reading in a wide matrix), it takes about the same time as reading in a tall dataset. (see code lines 18 to 21).

IN READ.CSV DOCUMENTATION:

“read.table is not the right tool for reading large matrices, especially those with many columns: it is designed to read data frames which may have columns of very different classes. Use scan instead for matrices.” [2]

```
# As suggested by the documentation, we try to time scan.
# system.time({wides <- scan(“wide.csv”, sep=”,”,skip = 1, quiet = FALSE, nlines=1)})
system.time({wides <- scan(“wide.csv”, sep=”,”,skip = 1, quiet = FALSE)})
```

```
user  system elapsed
0.003   0.000   0.003
```

WHY THE DIFFERENCE?

Dataframes are ideal when you have different data types in different columns. The convenient data manipulation functions, and syntax sugar that allows you to write concise and prettified code, does not come free. All this convenience comes with theoverhead cost of the dataframe itself. If the dataset is tall (lots of rows and not a lot columns), then the overhead diminishes.

```
# Size of memory taken by 8192 double precision floating point numbers.
8*8192

print(“Tall:”)
object.size(tall) # size of a 8192x1 matrix
object.size(tallr) # size of a 8192x1 dataframe

print(“Wide:”)
object.size(wide) # size of a 1x8192 matrix
object.size(wider) # size of a 1x8192 dataframe
object.size(wider_wt) # side of a 1x8192 numeric
object.size(wides) # side of a 1x8192 numeric
```

Continued

```
65536
   [1] “Tall:”
   65752 bytes
   66264 bytes
   [1] “Wide:”
   65752 bytes
   1049184 bytes
   66040 bytes
   65584 bytes
```

Knowing there are tradeoffs with memory vs performance and convenience, we can use this information to our advantage. One double precision floating point number (which is the default) takes up 8 bytes in memory. Theoretically speaking, if there is no overhead, 8192 doubles takes 65538 bytes. Looking at the output of above code cell, one thing that grasps my attention is how much memory is taken up by object wider (1x8192 dataframe): 1049184 bytes, much more than our theoretical estimate! Memory consumption is not always the cause of bad performance, but they are usually related: bigger memory consumption means it will take more time to retrieve data, has bad data locality, and also is unfriendly to cache.

USING READ_CSV()

Tidyverse also has a reading function: read_csv (notice the underscore, rather than .) It will read data into tibble, while read.csv will load data as a dataframe. The following code block output shows, yes, it is much faster than read.csv, but what we observed that wider datasetstake longer to read and write than tall datasets still stands using read_csv or write_csv.

```
library(tidyverse)
system.time({tall_r = read_csv(“tall.csv”)})
system.time({wide_r = read_csv(“wide.csv”)})
system.time({write_csv(tall_r, “tall_tb.csv”)})
system.time({write_csv(wide_r, “wide_tb.csv”)})

print(“Tall tibble:”)
object.size(tall_r) # size of a 8192x1 dataframe

print(“Wide tibble:”)
object.size(wide_r) # size of a 1x8192 dataframe

print(“Reading transposed data, then transposing it.”)
system.time({
   wide_transpose = read_csv(“tall.csv”)
   wide_r_wt = t(wide_transpose)

})
object.size(wide_r_wt) # size of a 8192x1 dataframe
```

```
Parsed with column specification:
cols(
   V1 = col_double()
)

user  system elapsed
0.003   0.001   0.004

Parsed with column specification:
cols(
   .default = col_double()
)
See spec(...) for full column specifications.

   user  system elapsed
   1.411   0.053   1.463
   user  system elapsed
   0.002   0.001   0.002
   user  system elapsed
   0.401   0.074   0.474
```

Continued

```
[1] “Tall tibble:”
68304 bytes
[1] “Wide tibble:”
4655264 bytes
[1] “Reading transposed data, then transposing it.”

Parsed with column specification:
cols(
   V1 = col_double()
)

user  system elapsed
0.004   0.001   0.005
66040 bytes
```

WANT FAST READING FOR A DATA.FRAME?

fread is a lot faster than read.csv. This has to do with memory use when reading the file. Essentially, fread memory maps the file into memory and then iterates through the file using pointers. Whereas read.csv reads the file into a buffer via a connection.

```
library(data.table)
# many people claim fread is faster (use verbose=TRUE, if interested in full log)
system.time({tallr = fread(“tall.csv”)})
system.time({wider = fread(“wide.csv”)})

object.size(tallr) # size of a 8192x1 dataframe
object.size(wider) # size of a 1x8192 dataframe
```

```
Attaching package: ‘data.table’

The following objects are masked from ‘package:dplyr’:
      between, first, last

 The following object is masked from ‘package:purrr’:
      transpose

user  system elapsed
0.002   0.001   0.004
user  system elapsed
0.009   0.000   0.009

66688 bytes
1573832 bytes
```

CONCLUSION

The wide data orientation will take much longer for I/O speed for a data frame, particularly, reading speed. **Even if you have a wide data set, it is faster to have the data written out as a transposed version, because later on, reading in + transposing the data will take less time than reading in the wide data, as is.**

If you have a matrix, consider using the scan function for speed reasons.

If you have a dataframe (columns of data may have different type of data), then the your choice depends on what you have. If you only consider performance, then the rank of these functions in terms of performance- fread > read_csv > read.csv

References
[1] Poggi N. (2019) Microbenchmark. In: Sakr S., Zomaya A.Y. (eds) Encyclopedia of Big Data Technologies. Springer, Cham. **doi.org/10.1007/978-3-319-77525-8_111**
[2] R Documentation, read.table: Data Input

For more information, please contact:
Department of Preventive Medicine
Division of Biostatistics
66 N. Pauline Street, Suite 633
Memphis, TN 38163 | t 901.448.5900
Biostat-newsletter@uthsc.edu

THE UNIVERSITY OF TENNESSEE HEALTH SCIENCE CENTER.
BIOSTATISTICS