Statistics and Its Interface Volume 13 (2020) 119–125

GPU Accelerated Liquid Association GALA

Guani Wu, Yu-Cheng Li, Yi-Chang Lu, Ker-Chau Li, and Shinsheng Yuan*

High throughput biological assays have provided numerous data sources for studying complex interactions between multiple variables in a biological system. Many computational tools for exploring the voluminous biological data are based on pair-wise correlation between variables. Liquid Association (LA) is a novel statistical concept for inferring higher order of association between variables in a system. While LA was originally introduced to study gene-gene interaction involving three genes at a time, it can be applied for correlating biological measurements with clinical variables such as drug sensitivity profiling and patient survival time. It is computationally expensive to compute LA scores for all possible triplets in very large datasets. Here we show how to take advantage of Graphic Processing Units (GPUs) for speeding up the LA computing. Our GPU-accelerated version of LA computation (GALA) achieved nearly 200-fold improvement over the traditional CPU-alone version. A companion package in R was developed for facilitating follow-up analysis and improving user experience. An example on Global Health Observatory data is provided to showcase how LA analysis can be applied in other data intensive fields.

Keywords and phrases: Liquid Association, Correlation coefficient, GPU, Gene expression.

1. INTRODUCTION

Correlation is a simple yet powerful concept in analyzing gene expression data. Two genes with positively correlated expression profiles are likely to be functionally associated and they may participate in the same or related biological process. However, functionally associated genes may not have correlation in expression. For instance, they may not be regulated at the transcription level and they have multiple functions. Co-expressed genes may become uncorrelated or even turn into contra-expressed when the underlying cellular state changes. Liquid association (LA), as opposed to "steady" association, is designed to quantify the size and the direction of the change of correlation between two genes. LA describes the ternary relationship between variables in a system [6, 8, 7, 16, 13, 14]. In gene expression study, the total computing complexity of LA is O(n^3) where n is the number of the genes. For integrated studies, it is time-consuming to compute all possible combinations from whole genome gene expression, SNP, or copy number variation data. To mitigate this problem, we developed a program via Compute Unified Device Architecture (CUDA) language for Graphic Processing Unit (GPU) platforms to accelerate the performance of LA score computation. A 200 times speed-up over the CPU version was obtained. A companion R package was also developed. The users can use it for visualizing the correlation changes and for conducting further analyses. We expect LA to have wide application in the big data era other than bioinformatics. An example concerning government expenditure and health outcome indicators is provided.

2. METHOD

2.1 Liquid Association

In the context of gene expression, LA conceptualizes the mediation of the change in the co-expression pattern of two genes (X, Y) by a third gene Z. A positive LA score indicates that the correlation between gene X and gene Y is likely to change from being negative to positive. Conversely, a negative LA score indicates the change from positive to negative correlation. The standard procedure to obtain LA score $LA(X, Y|Z)$ requires two steps [6]:

1. Normal score transformation. To standardize each gene-expression profile with normal score transformation, the m values in the profile are compared with each other and their ranks $R_1, ..., R_m$ are recorded. The ranks are then used to obtain the transformed profile, $\Phi^{-1}(R_1/(m+1)), \Phi^{-1}(R_2/(m+1)), ..., \Phi^{-1}(R_m/(m+1))$, where $\Phi(.)$ is the cumulative normal distribution.
2. LA score computation. Compute the average product of the three transformed profiles, $(X_1Y_1Z_1 + \ldots + X_mY_mZ_m)/m$. This gives the LA score $LA(X, Y|Z)$.

It is computer intensive to obtain LA scores because the number of combinations in choosing three from N genes or probes under study grows rapidly as N increases. It is typical for N to exceed 50K in commercial human gene expression chips and the number gets 10 times higher in SNP,

*Corresponding author.

DNA copy number, or methylation arrays. To improve user experience, we also compare the computed LA scores and save the top positive LA scores and bottom negative LA scores. This helps speed up the response time for on-line queries.

2.2 GPU Accelerated Liquid Association

GPUs were first introduced to accelerate computing speeds in computer graphics. General Purpose Computing on GPU (GPGPU) is a technique of using GPUs, which generally requires a set of stream processors and a hierarchical memory structure, to execute the computing tasks in parallel. We chose the popular CUDA language for reprograming the LA computation. The speeds of GALA running on two different GPUs will be compared to the C version running on the CPU machine in this article.

Since GPU executes in SIMT (Single Instruction Multiple Thread) mode, we must design the instruction set for each thread, the GPU kernel function, to perform LA computation for the three normal-score transformed profiles. In general, an optimized GPU kernel function consists of several steps such as utilization of shared memory for computation, effective usage of global memory bandwidth, efficient coordination of multiple threads. Our kernel function was constructed with these performance considerations.

Shared memory is the key to the reduction of global memory traffic. In order to fully utilize the shared memory, GALA partitions data into subsets so that each subset matches the size of shared memory. Coordinated by the GPU scheduler, the GPU processing elements execute a fixed number of the threads at a time and within the grouped threads, *warp*, the executed instructions must be the same at any time point. Because the size of the warp is limited, we constructed our GPU kernel function to tailor the dimensions of the matrices of the three transformed profiles declared in the shared memory. As GPU transfers data by moving one block of consecutive memory bits at a time, our input data are arranged with the memory coalescing technique to minimize the transfer counts. The GPU scheduler also determines when and which warp to be executed or placed on hold. A barrier synchronization function is employed to coordinate the parallel activities of multiple warps, thus enabling the more efficient parallel execution of threads (Algorithm 1).

Initially, GALA dynamically declares the feasible number of threads according to the size of input. When the input is too large to be computed, GALA will split the input into smaller pieces so that each of them fits in the allowable number of threads for the kernel function. In addition, if the input size is too small, GALA will launch the kernel function with an adjusted number of threads to prevent the kernel function from running the extra threads.

Algorithm 1: The kernel function of GALA

```
kernelOfGALA $(X, Y, Z)$
    inputs : $X$ and $Y \in \Re^{k \times m}$, $Z \in \Re^{v \times m}$
    output: $LA(X, Y, Z) \in \Re^{k \times k \times v}$
    foreach $t \in v$ do
        foreach $i \in m$ by $Block\_Size$ do
            $\_\_shared\_\_\ x_i \leftarrow X[Block\_Size][Block\_Size]$;
            $\_\_shared\_\_\ y_i \leftarrow Y[Block\_Size][Block\_Size]$;
            $\_\_shared\_\_\ z_{t,i} \leftarrow Z_t[Block\_Size]$;
            $\_\_syncthreads()$;
            $LA(X, Y|Z_t) \leftarrow LA(X, Y|Z_t) + LA(x_i, y_i|z_{t,i})$;
            $\_\_syncthreads()$;
    return $LA(X, Y, Z)$;
```

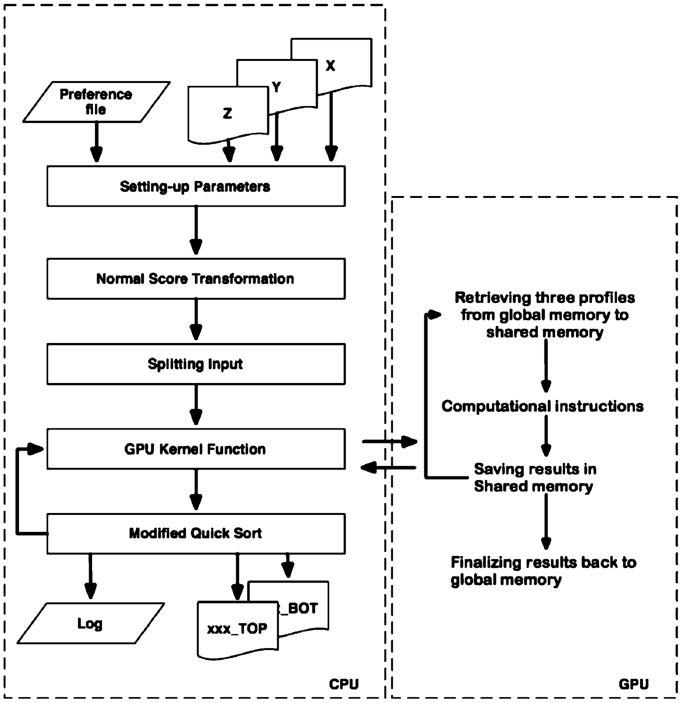
X
Y
Z
Preference file
Setting-up Parameters
Normal Score Transformation
Splitting Input
GPU Kernel Function
Modified Quick Sort
Log
xxx_TOP
_BOT
CPU
Retrieving three profiles from global memory to shared memory
Computational instructions
Saving results in Shared memory
Finalizing results back to global memory
GPU

Figure 1. The flowchart of GALA. The normal score transformation and sorting of computed LA scores are performed by CPU as shown on the left panel. Computation of LA scores, the most time-consuming part, is executed by GPU as shown on the right panel.

The output of the kernel function is an array identifier and the LA scores with allocated consecutively in the global memory. Once the kernel function was executed, GALA will perform a modified version of Quick Sort. This sorting function is used to sort the outputs from the kernel function and to filter LA scores according to the parameters of the preference file. Iterations between the kernel function and the sorting function will be continued until all LA scores are computed (Figure 1).

Table 1. Seven Gene Expression Datasets

ID	Sources
S1, S2	NCI-60 cancer cell line [10].
S3	Lung adenocarcinoma [11].
S4	High-grade lung neuroendocrine tumors of the lung [4].
S5	Various human and mouse tissues [12].
S6	Frozen tissue of primary lung tumors [1].
S7	Normal human tissues from selected samples [9].

2.3 Performance

We demonstrated the improvement of GALA over the original LA program with seven public available gene expression datasets as Table 1 shows. We used two different types of GPU cards to implement GALA, Tesla M2050 which contains 448 sets of 1.3 GHz processors with 3 GB dedicated memory and Tesla M2090 which contains 512 sets of 1.3 GHz processors with 6 GB dedicated memory. On the other hand, the CPU version of LA is performed on an Intel Core i7 965 model with the clock-speed at 3.2 GHz and 6 GB main memory. Comparing the cost of three devices, at this writing, the CPU was around $998, and the GPU cards were $149 for M2050 and $165 for M2090. (April 25, 2019 https://www.amazon.com/)

Since the loading ratio between the LA-score computation and LA-score sorting was around 10:1, the speed comparison for GALA would be focused on the LA-score computation only. We used the most time-demanding on-line query, i.e. finding the top LA scores of $(X, Y|Z)$ over all possible pairs of (X, Y) from an input of Z, as the submitted job and recorded the elapsed time of computing in each of the aforementioned test datasets. In addition, the elapsed time also involved the data transportation between the main memory and the global memory. In Table 2, the time listed under Tesla M2050 and Tesla M2090 is the elapsed time for GPU kernel function. For fair comparison, the column under CPU, only recorded the time on computing LA scores. We found that GALA outperformed CPU version and the improvement generally ranged from 40-fold to 190-fold. Moreover, the result shows that our implementation takes full advantage of GPU card upgrade. Compared to Tesla M2050, Tesla M2090 has 64 more computational cores and 17% higher memory bandwidth. Our implementation had better performance on Tesla M2090 than that on Tesla M2050 with a 17% speedup in average. In Figure 2, the strong linear relationship was also observed between elapsed time and complexity. The relationship signals that GALA have the same performance regardless of the complexity of data.

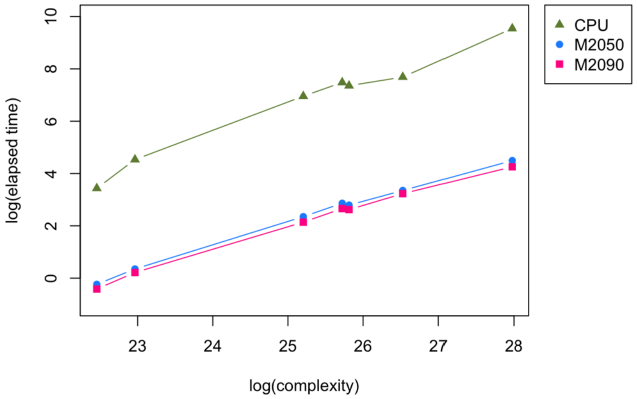

Figure 2. Complexity versus Elapsed Time. The x-axis is the log(Complexity) and y-axis is the log(Elapsed Time) in log scale.

2.4 LA Package in R

For encouraging the routine use of LA analysis, we also developed LA package in R to calculate LA scores and draw LA plots for further inspection of correlation patterns. We may select one triplet from the outcomes of GALA, and employ `drawla` to exam the relationship among three variables. The package contains `drawla` function and a dataset for the demonstration of LA. `drawla` has the following arguments:

```
drawla(x, y, z, ename, xyzLabels, switch = 2,...)
```

Three vectors `X`, `Y`, and `Z` are taken as input variables, and the order is also arranged as $LA(X, Y|Z)$. We can change the order of three vectors to observe the changes of LA plots such as $LA(Y, Z|X)$ or $LA(X, Z|Y)$. `drawla` aids the

Table 2. LA Computation Performance Comparison

Dataset	M2090(sec.)	M2050(sec.)	CPU(sec.)	Complexity(log)	Subjects	Genes
S1	0.66	0.79	31	9.75	60	9,706
S2	1.24	1.42	93.01	9.97	59	12,625
S3	8.5	10.5	1049	10.95	179	22,215
S4	14.3	17.52	1774	11.17	91	40,368
S5	13.74	16.33	1566.11	11.21	143	33,689
S6	25.29	28.57	2182.37	11.51	111	54,683
S7	70.61	89.59	13695.81	12.15	473	54,675

The column, Complexity, is defined as the number of subjects multiplied by the square of the number of genes in log scale.

Algorithm 2: Finding cut points of LA

```
findCutsOfLA $(X, Y, Z)$
    inputs : $X, Y, Z \in \Re^{1\times m}$
    output: $cut_1, cut_2 \in \Re^{1\times 1}$
    Sort $\{X, Y, Z\}$ by $Z$
    foreach Try $cut_1 \in \{1, cut_2\}$ do
        $b \leftarrow cov(Y_{1:cut_1}, X_{1:cut_1})/\sigma^2_{X_{1:cut_1}}$;
        $a \leftarrow \bar{Y}_{1:cut_1} - b\bar{X}_{1:cut_1}$;
        $RSS_1 \leftarrow (Y_{1:cut_1} - a - bX_{1:cut_1})^2$;
        foreach $cut_2 \in \{cut_1 + 1, n\}$ do
            $RSS_2 \leftarrow (Y_{cut_1+1:cut_2-1} - \bar{Y}_{cut_1+1:cut_2-1})^2$;
            $c \leftarrow cov(Y_{cut_2:n}, X_{cut_2:n})/\sigma^2_{X_{cut_2:n}}$;
            $d \leftarrow \bar{Y}_{cut_2:n} - b\bar{X}_{cut_2:n}$;
            $RSS_3 \leftarrow (Y_{cut_2:n} - c - dX_{cut_2:n})^2$;
            $RSS \leftarrow RSS_1 + RSS_2 + RSS_3$;
            $l \leftarrow -\frac{m}{2}\log(2\pi RSS) + \frac{1}{2}(m-1)$;
            If $Max(l)$ return $cut_1, cut_2$;
```

visualization of correlation between X and Y given different status of Z, where Z are split into three *status* (low, median, high). Cut points used to split Z were optimized by Algorithm 2, which maximizes log-likelihood function $l(\mu, \sigma^2; X^*, Y^*)$

$$RSS = \sum_{i=1}^{cut_1}(Y_i^* - \hat{\alpha_0} - \hat{\alpha}X_i^*)^2 + \sum_{i=cut_1+1}^{cut_2-1}(Y_i^* - \hat{\beta_0})^2 + \sum_{i=cut_2}^{n}(Y_i^* - \hat{\gamma_0} - \hat{\gamma}X_i^*)^2 \tag{1}$$

$$\sum_{i=1}^{cut_1}(Y_i^* - \hat{\alpha_0} - \hat{\alpha}X_i^*)^2 = \sigma^2_{Y^*_{1:cut_1}}(1 - \rho(Y^*_{1:cut_1}, X^*_{1:cut_1})) \tag{2}$$

$$\sum_{i=cut_1+1}^{cut_2-1}(Y_i^* - \hat{\beta_0})^2 = \sum_{i=cut_1+1}^{cut_2-1}(Y_i^* - \bar{Y}_{cut_1+1:cut_2-1})^2 \tag{3}$$

$$\sum_{i=cut_2}^{n}(Y_i^* - \hat{\gamma_0} - \hat{\gamma}X_i^*)^2 = \sigma^2_{Y^*_{cut_2:n}}(1 - \rho(Y^*_{cut_2:n}, X^*_{cut_2:n})), \tag{4}$$

where (X^*, Y^*) denotes (X, Y) sorted by Z, and ρ is the function of correlation coefficient.

3. APPLICATION

This example concerns the cross-nation comparison of public health expenditure and efficiency using the Global Health Observatory (GHO) data released by [15]. Cost/efficiency evaluation between different health care systems in different nations is a complicate issue requiring deep analysis from many perspectives and by different models [2, 5, 3]. The GHO data website is WHO's gateway to health-related statistics in its 194 Member States. There are over a thousand indicators including overall health status indicators, the indicators for the specific health and health-related targets of the Sustainable Development Goals. We downloaded the Year 2012 data, containing 306 female-related indicators, 306 male-related indicators and 379 gender-irrelevant indicators. One of our key findings (Figure 3) is the triplet, X= "Number of people (females) dying between ages 75 and 79[1]" (FD75), Y = "Age-standardized (female) mortality rate by all causes" (FMR) Z = "General government expenditure on health as a percentage of total government expenditure" (GGE) as shown in Figure 3.

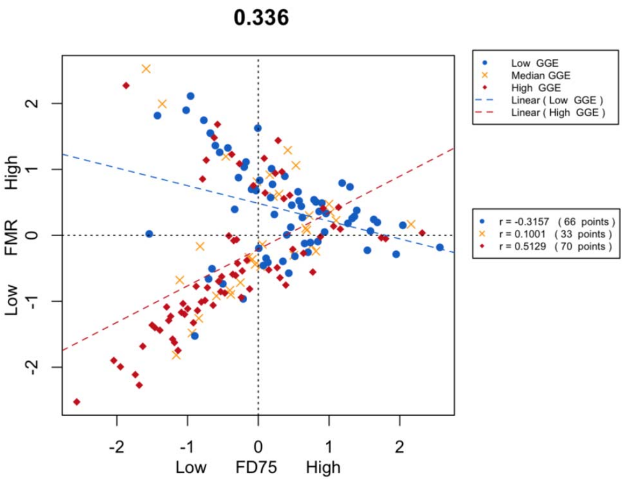

Figure 3. LA plot for (FD75, FMR, GGE). The correlation between FD75 and FMR is shown to change from negative to positive between low GGE nations and high GGE nations.

The correlation between FD75 and FMR is shown to change from negative for nations with lower GGE to positive for nations with higher GGE. This triplet showed the highest LA score when we set Z = GGE to explore the association between the set (as Y) of 131 indicators of age-standardized female mortality rates by different causes and the set (as X) of all 306 female-related indicators. Further investigation on how FMR correlates with female mortality rate for other age intervals, showed an interesting dynamic pattern of LA (Figure 4). Concerning the statistics in male population, similar pattern is also observed for X = "Number of people (males) dying between ages 70 and 75" (MD75), Y = "Age-standardized (male) mortality rate by all causes"

[1] Number of people dying between the beginning of the age group x and the beginning of the next age group $x + n$, n being the interval of the age group, given the hypothetical birth l0 = 100,000 [15].

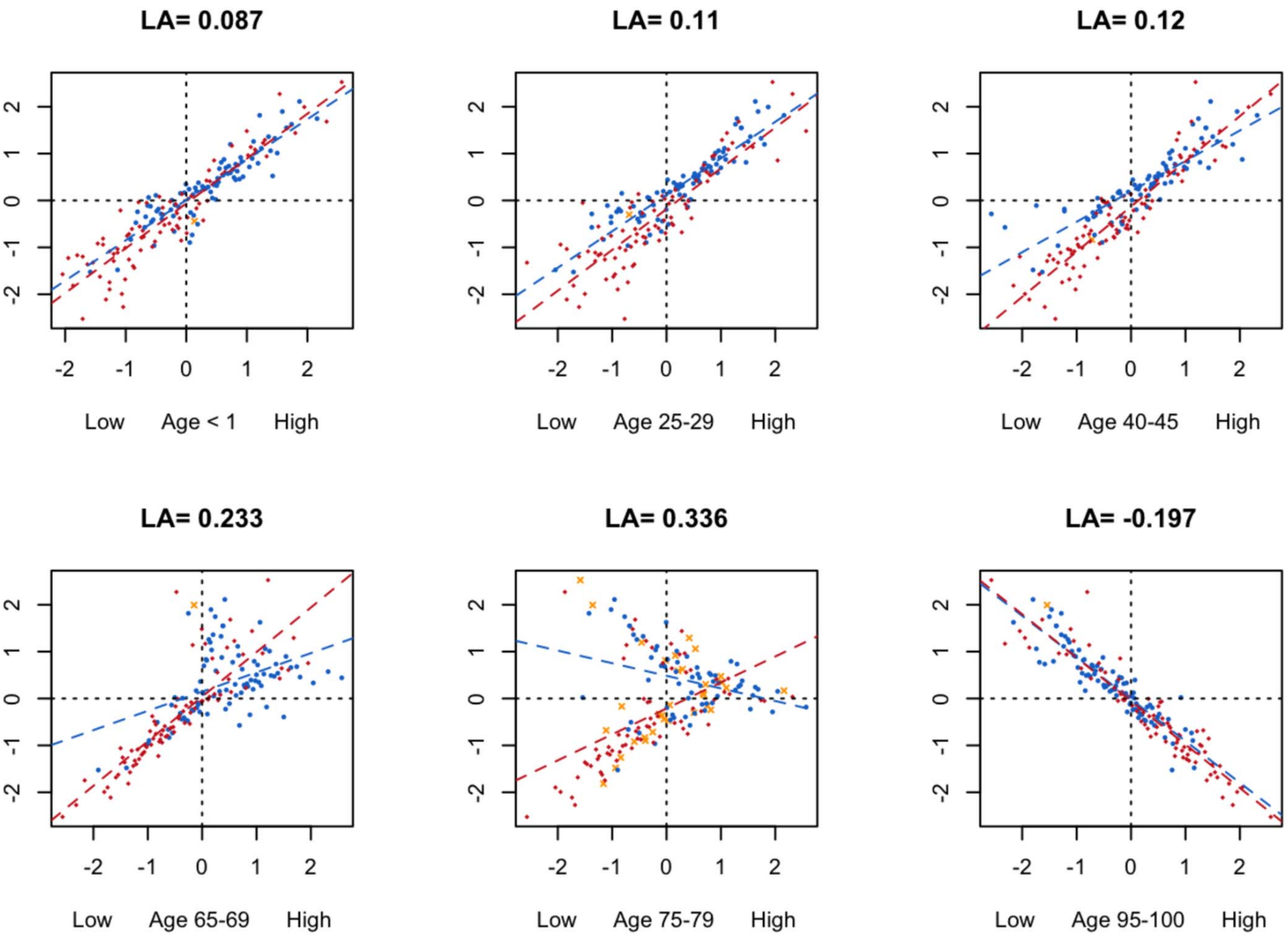

Figure 4. The changes of LA scores, where X axis is number of females dying between ages (x, x + 5),and Y axis is, FMR, the age-standardized mortality rate by all causes.

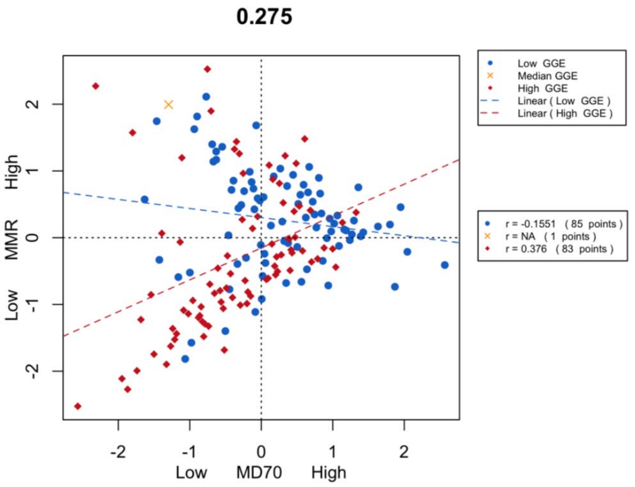

Figure 5. LA plot for (MD70, MMR, GGE). The correlation between MD70 and MMR is shown to change from negative to positive between low GGE nations and high GGE nations.

(MMR), Z = GGE (Figure 5), and the corresponding dynamic pattern of LA is presented in Figure 6. The interval shift from 75-79 to 70-75 may reflect that males typically have shorter life span than women.

4. CONCLUSION

In this article, we demonstrate a hybrid CPU/GPU program to obtain LA scores. The input data were arranged in a certain order for the efficient access from GPUs, and the configuration took the advantage of multiple cores of GPUs to speed up the LA scores computation. We recorded the elapsed time in testing eight real datasets, and compared GALA with the original LA program. GALA was much faster at executional speed regardless of the complexity of data. The use of the companion R code for visualizing the dynamic change of association between variable is illustrated. Our package can be widely applied in analyzing complex data from various scientific areas.

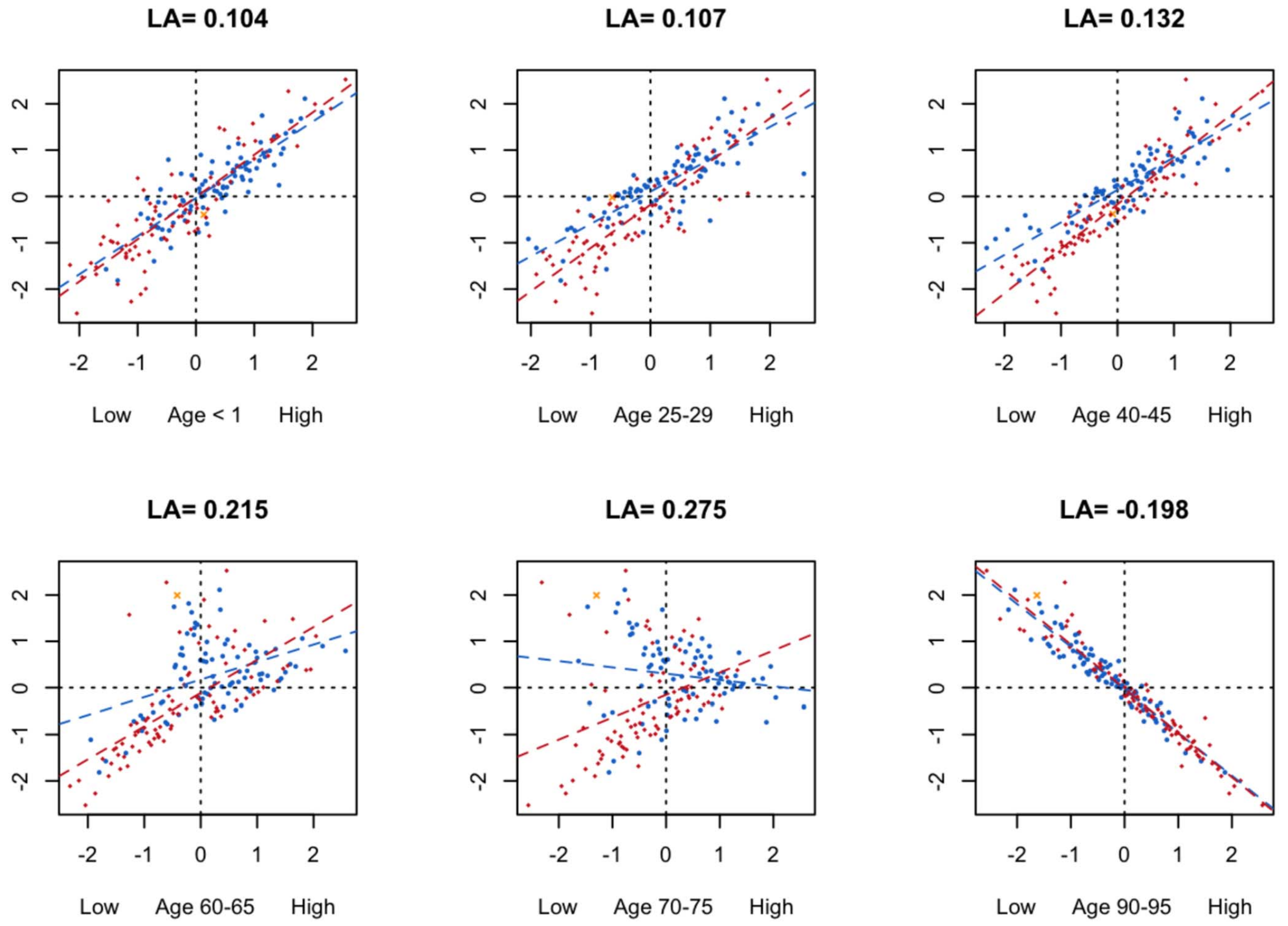

Figure 6. The changes of LA scores, where X axis is number of males dying between ages (x, x + 5), and Y axis is, Age-standardized (male) mortality rate by all causes.

ACKNOWLEDGEMENTS

This research was supported by Academia Sinica, Taiwan (Mathematics in Biology & AS-104-TP-A07), Ministry of Science and Technology, Taiwan, MOST106-2314-B-001-005 and National Science Foundation, USA, NSF, DMS-1513622. The content is solely the responsibility of the authors and does not necessarily represent the official views of NSF.

Received 15 January 2019

REFERENCES

[1] Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., Joshi, M.-B., Harpole, D., Lancaster, J. M., Berchuck, A., Olson Jr, J. A., Marks, J. R., Dressman, H. K., West, M., and Nevins, J. R. (2005). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature 439*, 353.

[2] Elola, J., Daponte, A., and Navarro, V. (1995). Health indicators and the organization of health care systems in western Europe. *American Journal of Public Health* **85**, 10 (Oct.), 1397–1401.

[3] Frogner, B. K., Frech, H., and Parente, S. T. (2015). Comparing efficiency of health systems across industrialized countries: a panel analysis. *BMC Health Services Research* **15**, 1 (Sept.), 415.

[4] Jones, M. H., Virtanen, C., Honjoh, D., Miyoshi, T., Satoh, Y., Okumura, S., Nakagawa, K., Nomura, H., and Ishikawa, Y. (2004). Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *The Lancet* **363**, 9411 (Mar.), 775–781.

[5] Kim, T. K. and Lane, S. R. (2013). Government Health Expenditure and Public Health Outcomes: A Comparative Study among 17 Countries and Implications for US Health Care Reform. *American International Journal of Contemporary Research 3(9)*, 8–13.

[6] Li, K.-C. (2002). Genome-wide coexpression dynamics: Theory and application. *Proceedings of the National Academy of Sciences* **99**, 26, 16875–16880.

[7] Li, K.-C., Palotie, A., Yuan, S., Bronnikov, D., Chen, D., Wei, X., Choi, O.-W., Saarela, J., and Peltonen, L. (2007). Finding disease candidate genes by liquid association. *Genome Biology* **8**, 10, R205–R205.

[8] Li, K.-C. and Yuan, S. (2004). A functional genomic study on NCI's anticancer drug screen. *The Pharmacogenomics Journal 4*, 127.

[9] Roth R. (2007). Human body index - transcriptional profiling.

[10] Shankavaram, U. T., Varma, S., Kane, D., Sunshine, M., Chary, K. K., Reinhold, W. C., Pommier, Y., and Weinstein, J. N. (2009). CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics 10*, 277–277.

[11] Shedden, K., Taylor, J. M., Enkemann, S. A., Tsao, M. S., Yeatman, T. J., Gerald, W. L., Eschrich, S., Jurisica, I., Venkatraman, S. E., Meyerson, M., Kuick, R., Dobbin, K. K., Lively, T., Jacobson, J. W., Beer, D. G., Giordano, T. J., Misek, D. E., Chang, A. C., Zhu, C. Q., Strumpf, D., Hanash, S., Shepherd, F. A., Ding, K., Seymour, L., Naoki, K., Pennell, N., Weir, B., Verhaak, R., Ladd-Acosta, C., Golub, T., Gruidl, M., Szoke, J., Zakowski, M., Rusch, V., Kris, M., Viale, A., Motoi, N., Travis, W., and Sharma, A. (2008). Gene Expression-Based Survival Prediction in Lung Adenocarcinoma: A Multi-Site, Blinded Validation Study: Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma. *Nature medicine* **14**, 8 (Aug.), 822–827. MR2756942

[12] Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences* **101**, 16, 6062–6067.

[13] Sun, W., Yuan, S., and Li, K.-C. (2008). Trait-trait dynamic interaction: 2d-trait eQTL mapping for genetic variation study. *BMC Genomics* **9**, 1 (May), 242.

[14] Tai, S.-K., Wu, G., Yuan, S., and Li, K.-C. (2010). Genome-wide expression links the electron transfer pathway of Shewanella oneidensis to chemotaxis. *BMC Genomics* **11**, 1 (May), 319.

[15] World Health Organization. (2017). Global Health Observatory (GHO) data.

[16] Wu, T., Sun, W., Yuan, S., Chen, C.-H., and Li, K.-C. (2008). A method for analyzing censored survival phenotype with gene expression data. *BMC Bioinformatics* **9**, 1 (Oct.), 417. MR2562471

Guani Wu
Department of Statistics University of California, Los Angeles
USA
E-mail address: guani@ucla.edu

Yu-Cheng Li
Department of Electrical Engineering, National Taiwan University
Taiwan
E-mail address: yuchengli@stat.sinica.edu.tw

Yi-Chang Lu
Department of Electrical Engineering, National Taiwan University
Taiwan
E-mail address: yiclu@cc.ee.ntu.edu.tw

Ker-Chau Li
Institute of Statistical Science, Academia Sinica
Department of Statistics University of California, Los Angeles
USA
E-mail address: kcli@stat.sinica.edu.tw

Shinsheng Yuan
Institute of Statistical Science, Academia Sinica
Taiwan
E-mail address: syuan@stat.sinica.edu.tw