conference

October 6-8, 2021 | VIRTUAL

Contributed Presentation

UNIVERSITY OF LOUISVILLE
SCHOOL OF PUBLIC HEALTH & INFORMATION SCIENCES

Normalization Methods and Statistical Inference *to Identify Differentially Expressed MicroRNAs with an Application to a Residential Cohort Exposed to Environmental Toxins and Pollutants*

UNIVERSITY OF LOUISVILLE
SCHOOL OF PUBLIC HEALTH & INFORMATION SCIENCES




Authors

- **Christina Pinkston, MS - University of Louisville**
- Matthew Cave, PhD - University of Louisville
- Brian Chorley, PhD - USEPA
- ***Shesh Rai, PhD- University of Louisville***

Disclosures and Disclaimer

The authors declare they have no actual or potential competing financial interests. Findings presented here are those of the authors and do not necessarily represent the official position of USEPA, UofL, ATSDR, or NIH.

UNIVERSITY OF LOUISVILLE
SCHOOL OF PUBLIC HEALTH & INFORMATION SCIENCES

Acknowledgments

- **Linda Birnbaum (NIEHS/NIH)**
- **Marian Pavuk (ATSDR/CDC)**
- USEPA: Gleta Carswell and Gail Nelson
- UofL: Banrida Wahlang, Youssef Saad, and Kimberly Head
- Investigators who measured biomarkers, including Andreas Sjödin, Richard Jones, Frank Middleton, Paula Rosenbaum, Heather Clair.

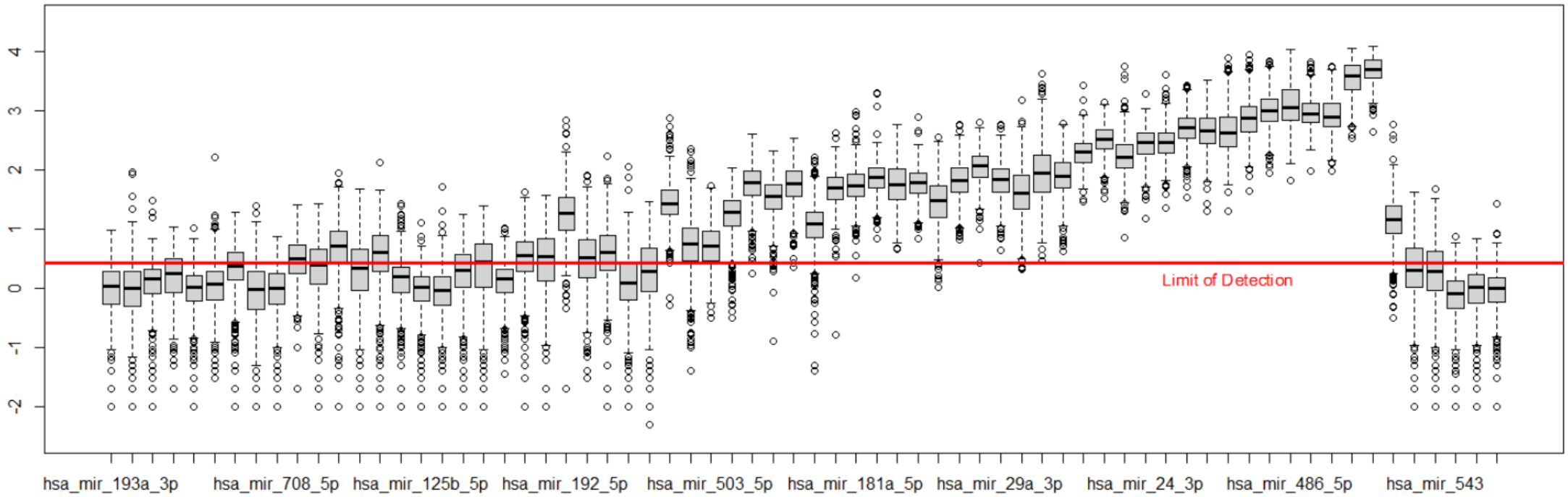
Funding

This research was supported, in part, by the National Institute of Environmental Health Sciences (R35ES028373, R01ES032189, T32ES011564, P42ES023716, P30ES030283 and R21ES031510 and the Intramural Research Program Z01-ES100475); the Agency for Toxic Substances and Disease Registry/Centers for Disease Control and Prevention (200-2013-M-57311);and the Wendell Cherry Endowed Chair. The Anniston Community Health Survey original data collection was funded by the Agency for Toxic Substances and Disease Registry (5U50TS473215).

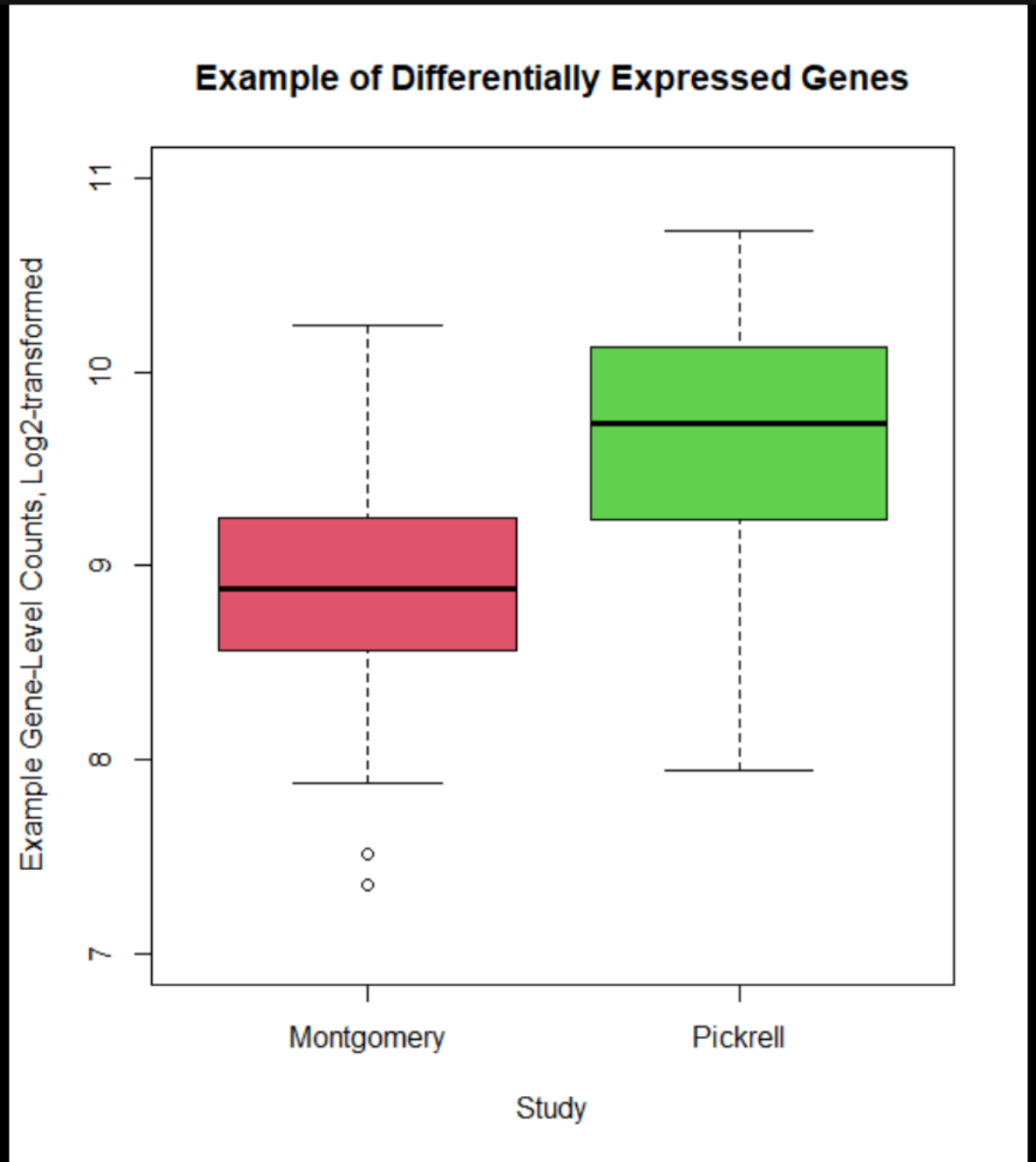
UNIVERSITY OF LOUISVILLE
SCHOOL OF PUBLIC HEALTH & INFORMATION SCIENCES

Study Aims

Aim 1: Determine the best normalization strategy for microRNAs (miRs) profiled using the Fireplex® platform technology by Abcam.



Study Aims



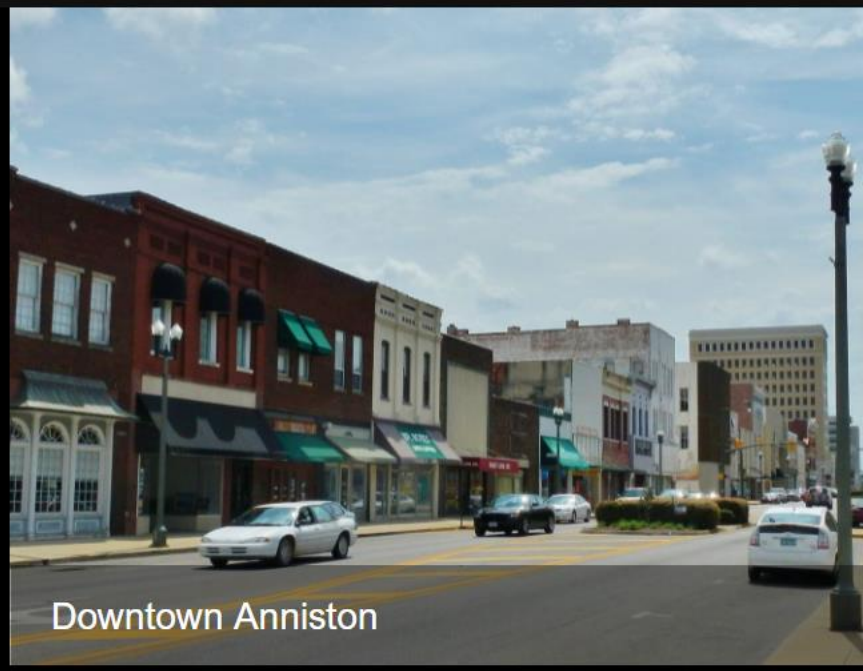
Aim 2: Identify differentially expressed hepatotoxicity miRs in a cohort exposed to polychlorinated biphenyls (PCBs).

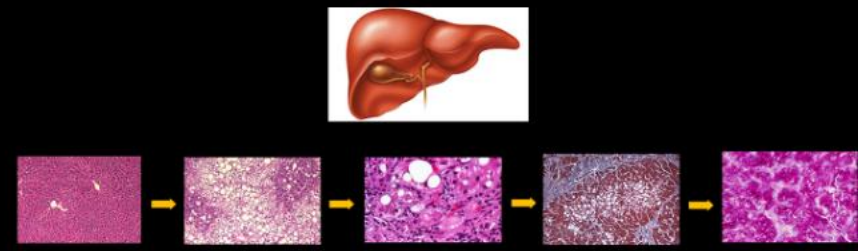
UNIVERSITY OF LOUISVILLE
SCHOOL OF PUBLIC HEALTH & INFORMATION SCIENCES

UNIVERSITY OF LOUISVILLE
SCHOOL OF PUBLIC HEALTH & INFORMATION SCIENCES

Anniston Community Health Study I (Anniston, Alabama; 2005-2007)

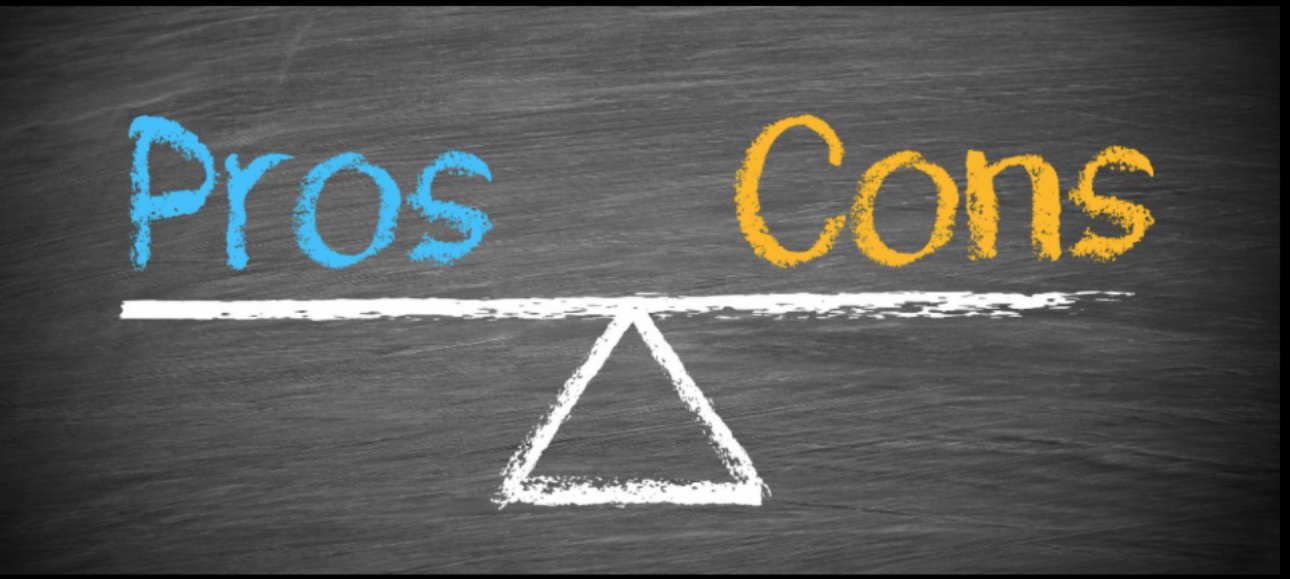
- Participants lived near a former Monsanto Chemical Corporation plant where PCBs were manufactured until 1971 [1]
- Fasting blood draws were collected on all participants [1]
- Liver biomarkers were measured by the Cave Laboratory [2]
- **60% show signs of liver disease** (>200 U/L Keratin 18-M30 (K18-M30) and/or >300 U/L K18-M65) [2]
- Primary covariates of interest: Age, Sex, Race, and BMI; most differ by liver disease status
- PCBs were whole weight, measured by NIEHS, and adjusted for total lipids in the models [1]
- IRB-approved (University of Louisville) [2]


Downtown Anniston



UNIVERSITY OF LOUISVILLE
SCHOOL OF PUBLIC HEALTH & INFORMATION SCIENCES

MicroRNAs (miRs)



Profiling Technology

Targeted miRs - small, single-stranded non-coding RNA [3] - were measured in serum by FirePlex® technology by Abcam [4]

Pros of the current profiling technology

- Allows profiling of limited serum samples not possible with other technologies
- Profile up to 68 miRs simultaneously with lower cost than other technologies

Cons of the current profiling technology

- Detection of low abundance miRs is reduced and has greater variability [5]
- There were no housekeeping miRs included to use for normalization

Normalization Procedures

For this talk, we compare fold change of differentially expressed miRs from the following data:

1. **Unnormalized** (raw, background subtracted) expression values

2. **geNorm**: Software provided [6, 7]

3. **Quantile Normalization: Overall** [8]

4. **Quantile Normalization: by liver disease class** [9]

Differential Expression

Modeling to find the least-square mean estimates by group:

- Log-10 transform the raw, background subtracted miR expressions
- Model includes: age, self-reported sex, self-reported race, BMI, total PCBs, total lipids

$$log_{10}(miR_{(i)}) = \alpha_{(i)} + \beta_{1(i)} * Liver\ Disease_{Necrotic} + \beta_{2(i)} * Liver\ Disease_{Other} +$$
$$\gamma_{1(i)} * Self - reported\ Race_{NHW} + \gamma_{2(i)} * Sex_{Female} + \gamma_{3(i)} * Age +$$
$$\gamma_{4(i)} * BMI + \gamma_{5(i)} * log_{10}(total\ lipids) + \gamma_{6(i)} * assayPlate +$$
$$\gamma_{7(i)} * log_{10}\left(\sum PCBs\right) + \varepsilon_{(i)} \qquad (Eq\ 1)$$

$$FC_{j(i)} = 10^{[\beta_{j(i)}]} \qquad (Eq\ 2)$$

$$SE(FC_{j(i)}) = 10^{[\beta_{j(i)}]} * ln(10) * SE(\beta_{j(i)}) \qquad (Eq\ 3)$$

Differential Expression

Fold change of Liver Disease vs. No Liver Disease

- Mean and variance are estimated as shown in [10]

$$\widehat{\theta_{(i)}} = e^{(ln(10)*LSM_{(1i)} + \frac{ln(10)^2*S^2_{LSM_{(1i)}}}{2} - ln(10)*LSM_{(0i)} - \frac{ln(10)^2*S^2_{LSM_{(0i)}}}{2})} \quad (Eq\ 4)$$

$$\widehat{log(\theta_{(i)})} = ln(10) * LSM_{(1i)} + \frac{ln(10)^2 * S^2_{LSM_{(1i)}}}{2} - ln(10) * LSM_{(0i)} - \frac{ln(10)^2 * S^2_{LSM_{(0i)}}}{2} \quad (Eq\ 5)$$

$$\widehat{Var}\left(\widehat{log(\theta_{(i)})}\right) = \frac{ln(10)^2 * S^2_{LSM_{(1i)}}}{n_1} + \frac{ln(10)^4 * S^4_{LSM_{(1i)}}}{2(n_1 - 1)} + \frac{ln(10)^2 * S^2_{LSM_{(0i)}}}{n_0} + \frac{ln(10)^4 * S^4_{LSM_{(0i)}}}{2(n_0 - 1)} \quad (Eq\ 6)$$

Differential Expression

Find confidence interval estimates [11]

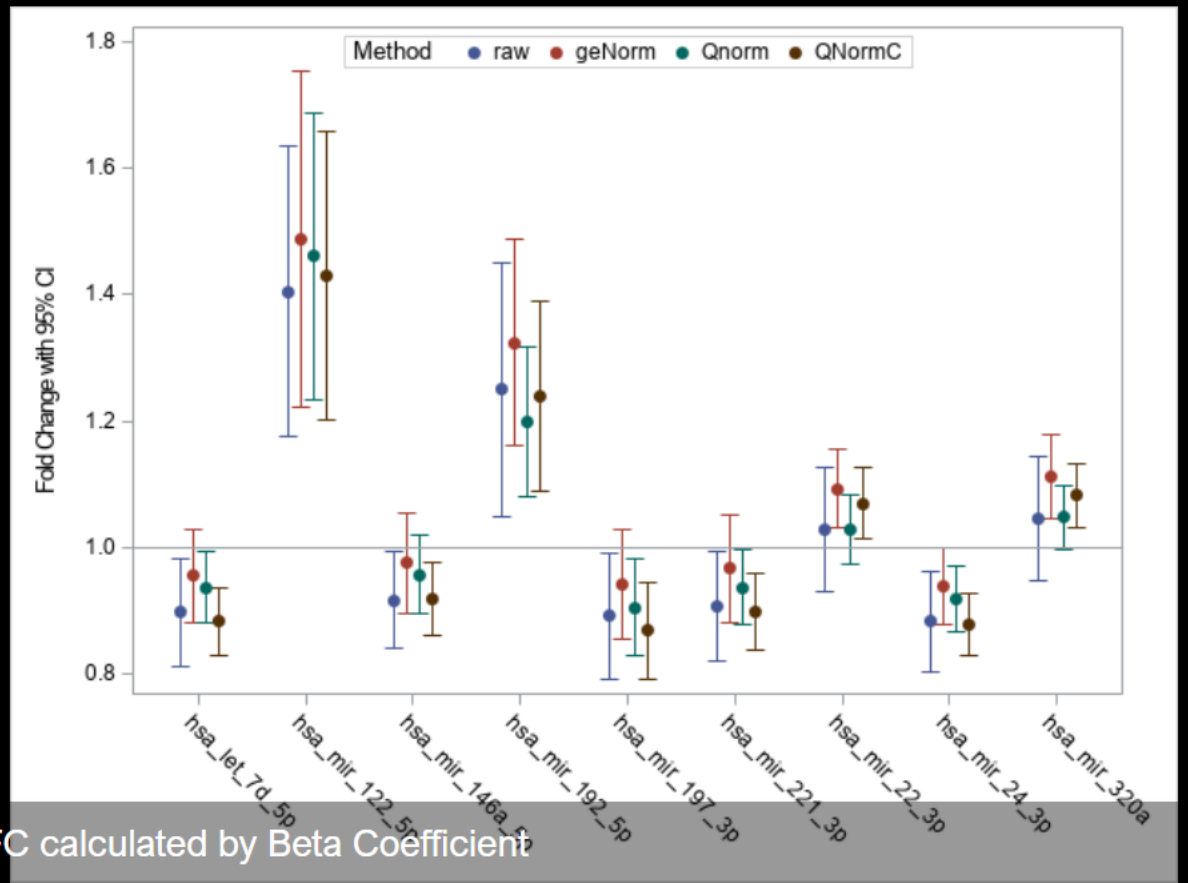
- We construct the 95% confidence interval for log(θ) using *t* instead of *z*
- Scale this to θ by taking the anti-logs

$$\widehat{log(\theta_{(i)})} \pm t_{0.95,n_0+n_1-2} * \sqrt{\widehat{Var}\left(\widehat{log(\theta_{(i)})}\right)} \quad (Eq\ 7)$$

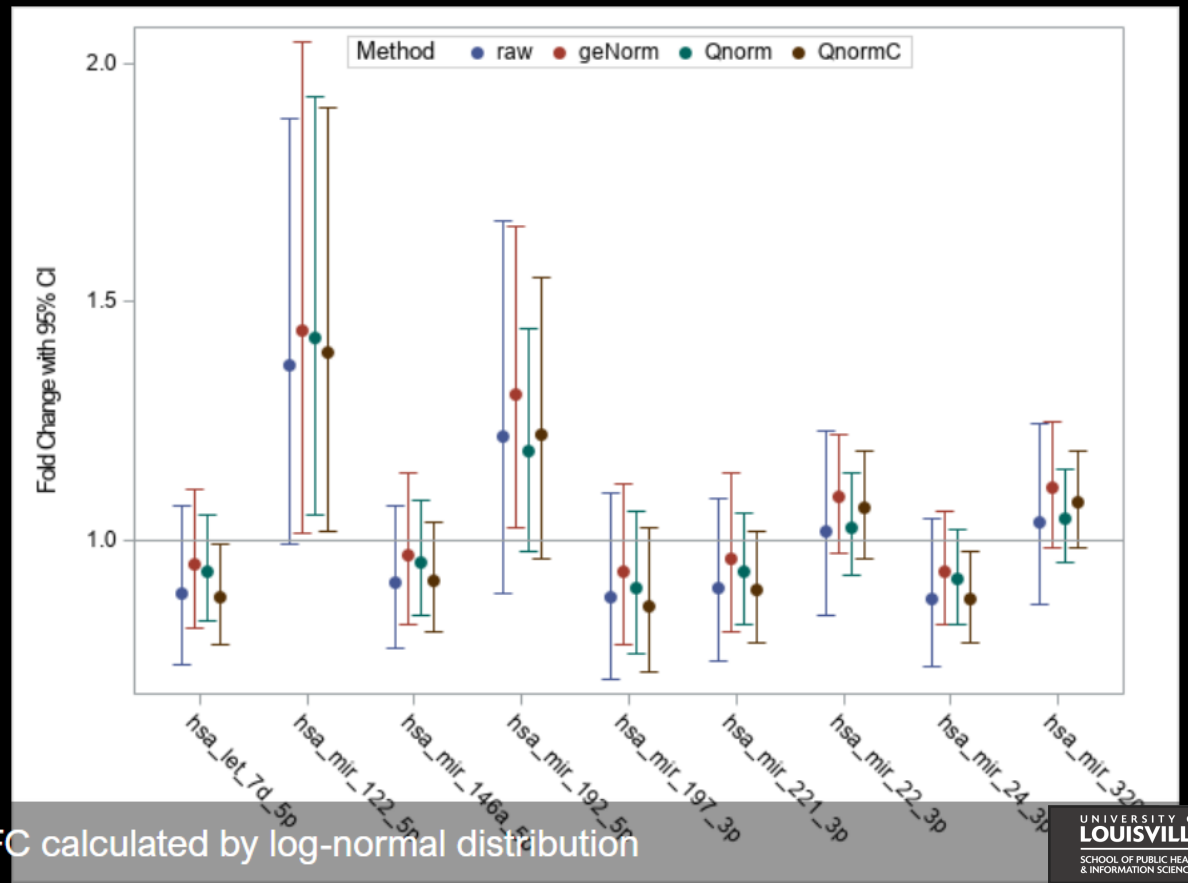
Findings

Performance by Normalization Technique and Fold Change Equation

We display selected miRs from our sample. We show the four methods of normalization: None, geNorm, Quantile Normalization Overall (QNorm), and Quantile Normalization by class (QNormC). Two methods of fold change (FC) calculation are shown: the anti-log of the beta coefficient, and the derived Fold change based on the log-normal distribution.



FC calculated by Beta Coefficient



FC calculated by log-normal distribution

Conclusions

Summary of Findings

- We recommend quantile normalization by class/group, especially when the groups are unbalanced
- It may be beneficial to compare different fold change calculations to compare the different underlying responses (i.e. mean vs. median)

References

1. Pavuk M, Olson JR, Sjödin A, et al. (2014) *Science of The Total Environment.* 473-474:286-297
2. Clair HB, Pinkston CM, Rai SN, et al. (2018) *Toxicological Sciences* 164(1):39-49
3. Lee RC, Feinbaum RL, Ambros V. (1993) *Cell* 75:843-854
4. Tackett MR, Diwan I. (2017) *Methods in Molecular Biology* 1654:209-219
5. Godoy PM, Barczak AJ, DeHoff P, et al. (2019) *Cell Reports* 29:4212-4222
6. FirePlex Analysis Workbench User Guide (2017)
7. Vandesompele J, De Preter K, Pattyn F, et al. (2002) *Genome Biology* 3(7):research0034.1–0034.11
8. Bolstad BM, Irzarry RA, Astrad M, et al. (2003) *Bioinformatics* 19(2):185-193
9. Zhao Y, Wong L, Goh WWB. (2020) *Scientific Reports* 10:15534
10. Olsson U. (2005) *Journal of Statistics Education* 12(1)
11. Zhou XH, Gao S. (1997) *Statistics in Medicine* 16:783-790

UNIVERSITY OF LOUISVILLE
SCHOOL OF PUBLIC HEALTH & INFORMATION SCIENCES