Doing research separated by distance and time with blinded duplication: Building an Epigenetic Database

Scott Hines[1,a], Namita Biala[1,a], Daniel Hutchinson[1,a], Jacob Suazo[2,a], Sade Akinkuotu[3,b], Sage Arbor[1,c]

[1]Marian University College of Osteopathic Medicine - Indianapolis, Indiana, [2]St. George's University School of Medicine - True Blue, Grenada, [3] Howard University Department of Psychology - Washington DC

[a]Medical school 3rd year student , [b]PhD candidate , [c]Faculty P.I. Biochemistry PhD

Abstract

The design, administration, input, and evolution of an epigenetic database containing metadata across studies was created using cloud-based tools with permission and blinding capacities. The database was designed to store the odds ratio (OR) of various interventions such as exercise, environmental exposure, and food consumption on health outcomes, many of which work mechanistically via epigenetic modifications. To this end, existing published research on the epigenetic influences on disease and overall health promotion were evaluated and compared through a variety of normalized metrics. Research is often hampered by aggregating qualified researchers in one place to ensure continuity and interaction. Cloud based informatic tools allow for the electronic formation of such teams, while asynchronous data input creates a shared library of inputs and outputs essential for standardized database creation. Zotero was used for literature sharing and Google Docs for selective data sharing and visualization. The manual curation of primary literature to build databases with metadata is a fruitful avenue of research for those pursuing terminal degrees that sequester too much of their time for more classical contiguous lab bench work. This database is currently being created by third year medical school students at two medical schools, as well as a PhD student at a third school. A future use of this database is the creation of a smartphone application that would allow users to query certain behaviors or outcomes (e.g. running or cancer) and obtain a rank-ordered report on how various interventions or therapeutics would combat a disease based on their odds ratios. The emerging role of epigenetics in modifying one's own phenotype can be leveraged with a database such as this to allow for patient-executed lifestyle interventions, which could increase health while lowering healthcare costs.

Introduction

Epigenetics, the turning on or off of gene expression without permanent modification of the genetic code, continues to gain attention as a targetable aspect of a patient's biology to increase quality and quantity of life. A plethora of compounds and activities have been found to affect human epigenetics such as alcohol consumption, exercise, and sleep. There is currently a lack of metadata to compare the efficacy between various lifestyle interventions (both medicinal and patient executed). We are currently codifying primary literature on epigenetics and developing an epigenetic database to compare actions and therapies that patients can choose to undergo.

The inputting of standardized data among a spatially and temporally disparate group of researchers presents difficulties which can be alleviated through the use of cloud based platforms. Google docs (recently rebranded as "G Suite") is a free platform with significant functionality to help in the normalization and blinding of data. The function "importrange" allows for real-time import of data between worksheets. This function created a web of interconnected Google sheets, allowing data duplication using a shared input vocabulary without one researcher seeing another's data.

Figure 1 : Google Doc connections and access

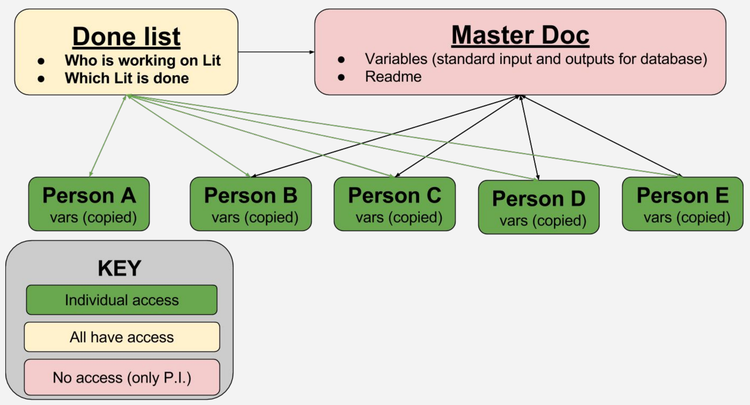

Figure 1: Flowchart showing how data was able to be input into individual Google Sheets, while a "Master Doc" visible only to the P.I. included data from all Google Sheets.

Figure 2 : MeSH term tree structure

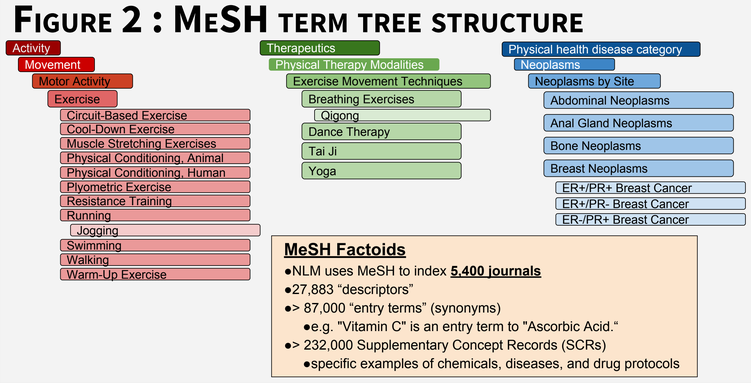

Figure 2: MeSH terms (Medical Subject Headings) is a hierarchical and controlled vocabulary used by the National Library of Medicine to organize all scientific literature. The examples above show how tagging a primary literature's data with a MeSH term allows its inclusion in a database with the nested MeSH hierarchy.

Abbreviations

OR = Odds Ratio
CI = Confidence Interval
DB = Database
MeSH = Medical Subject Headings

Results

The database currently contains 92 rows of data from 26 papers. Most actionable interventions could be binned into the following: sleep, exercise, food/nutrition consumption, or exposure (eg: pollution). Of these, sleep was the most influential, with 25 significant odds ratios related to health ranging from hypertension, anemia, and depression to weight gain. Exercise was next in magnitude with 11 significant odds ratios affecting illnesses such as cold incidence as well as improving survival rates of breast cancer and cardiovascular disease. Exercise was theorized to act through the promotion of a so-called "healthy methylation profile", meaning protective genes showed increased expression while deleterious protooncogenes were silenced via methylation. Food/nutrition consumption research was significant in 6 papers, which mostly revealed the importance of vitamins such as Vitamin D for health promotion. Vitamin D intake was found to be significantly protective against vascular dementia among the elderly and played a role in controlling conditions of atopy such as wheezing, food hypersensitivity, and dermatitis in populations of all ages. Exposure effects, such as smoking were found significant in a single study on maternal smoking and its effect on infant neurobehavior. The effect is mediated by epigenetic regulation of the placental glucocorticoid receptor gene. Further research on the epigenetic effects of first and second hand smoking is forthcoming.The ability for researchers to input data into their own individual Google Sheets has resulted in excellent blinding, ensuring that data is duplicated properly. Zotero has also proven to be an excellent tool to gather articles into one library accessible by multiple parties.

Figure 3a: Example of conflicting data entry

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
entered by	variable description	input	input type	input min value	input max value	input unit	output	output type	output min value	output max value	output unit	Risk	measure of Risk	CI low	CI high	p value	organism	Lit Ref	notes
Scott Hines	Af. Am. vigorous physical activity and breast cancer	total physical activity (time)	range	2	166	hours/week	Breast Neoplasms	number	-0.64	-0.64	percent	0.36	odds Ratio (rel)	0.17	0.75	0.01	human	Sheppard, V.B., Makambi, K., Taylor, T., Wallington, S.F., Sween, J., and Adams-Campbell, L. (2011). PHYSICAL ACTIVITY REDUCES BREAST CANCER RISK IN AFRICAN AMERICAN WOMEN. Ethn Dis 21, 406–411.	African American women who engaged in vigorous physical activity (> 2 hours/week in the past year) had a 64% reduced risk of breast cancer compared to those who did not participate in any vigorous activity
Sage Arbor	physical activity	high exercise intensity (>70% max heart rate)	range	2	66.66	hours/week	Breast Neoplasms	number	0	1	yes/no - binary (1/0)	0.36	odds Ratio (rel)	0.17	0.75	0.01	human	Sheppard, V.B., Makambi, K., Taylor, T., Wallington, S.F., Sween, J., and Adams-Campbell, L. (2011). PHYSICAL ACTIVITY REDUCES BREAST CANCER RISK IN AFRICAN AMERICAN WOMEN. Ethn Dis 21, 406–411.	African American women who engaged in vigorous physical activity (> 2 hours/week in the past year) had a 64% reduced risk of breast cancer compared to those who did not participate in any vigorous activity (odds ratio, OR = 0.36; 95% confidence interval, CI = 0.17–0.75). There was a OR of 0.83 and 0.36 of breast cancer for the 2nd most active and 3rd most active tertile compared to the least, but this data could not be input in the DB.

Figure 3a: Example showing the importance of duplicating data. Output data from the same paper is portrayed differently. Notice that one researcher focused on the percent decrease in breast cancer, whereas the second researcher focused on the presence or absence of cancer.

Figure 3b: Example of accurate data duplication

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
entered by	variable description	input	input type	input min value	input max value	input unit	output	output type	output min value	output max value	output unit	Risk	measure of Risk	CI low	CI high	p value	organism	Lit Ref	notes
Scott Hines	LINE-1 hypermethylation after deployment	War Exposure	number	1	1	yes/no - binary (1/0)	Alu Methylation	number	1	1	yes/no - binary (1/0)	1.46	odds ratio (rel)	1.08	1.97	0.05	human	Rusiecki, J.A., Chen, L., Srikantan, V., Zhang, L., Yan, L., Polin, M.L., and Baccarelli, A. (2012). DNA methylation in repetitive elements and post-traumatic stress disorder: a case–control study of US military service members. Epigenomics 4.	For the pre-deployment comparisons of cases and controls... Alu, predeployment comparisons of cases and controls showed a positive association for the total population
Jacob Suazo	Methylation of Alu repetitive element in soldiers predeployment to Afganistan that would later be diagnosed with PTSD	Alu Methylation	number	1	1	yes/no - binary (1/0)	Alu Methylation	number	1	1	yes/no - binary (1/0)	1.46	odds ratio (rel)	1.08	1.97	<0.05	human	Rusiecki, J.A., Chen, L., Srikantan, V., Zhang, L., Yan, L., Polin, M.L., and Baccarelli, A. (2012). DNA methylation in repetitive elements and post-traumatic stress disorder: a case–control study of US military service members. Epigenomics 4.	The Alu sequence family (named for the restriction endonuclease cleavage enzyme Alu I) is the most highly repeated interspersed repeat element in humans (over a million copies). It is derived from the 7SL RNA component of the SIGNAL RECOGNITION PARTICLE and contains an RNA polymerase III promoter. Transposition of this element into coding and regulatory regions of genes is responsible for many heritable diseases.

Figure 3b: Example of duplicated data. Data from the paper is similar in both lines, however the variable description and notes are different.

Materials & Methods

Google Docs - Cloud based password protected interconnected tables

Google Sheets is an online, excel-like functionality that allows for control of who can edit by user account. Data is interconnected and used for everything from dropdowns to conditional formatting over multiple sheets. (Tbl1)

Tbl1 - Google functions useful for remote collaborative work

Purpose	Google Function	Example
Import data from other google sheet	=importRange	`=importrange("uniqueKeyInUrlOfGoogleSheetWithData","rangeToCopy")` `=importrange("SuchAsIKAXSurhjXxllK795zdygmwylNRHulWidNunsjCiquCU","doneList!A1:O500")`
Filter data	=filter	`=filter(rangeToFilter, condition1, condition2)` `=filter(D10:D13, D10:D13 >= 2,F10:F13 >= 6)`
Embed image in cell	=image	`=image("https://docs.google.com/drawings/d/1yBIHATnr2w_DoXu4cRda6FPQ4T8XS/pub?w=960&h=720")` NOTE - google images are also collaborative cloud based documents , which can then be embedded in google sheet cells or other cloud documents.
Order data	=sort	`=sort(rangeToCopyAndSort,columnToSortBy,sortAscendingOrDescending)` `=sort(C10:D13,2,1)`
Dropdowns	Data→Validation	cellRange=cellsToHaveDropdowns Criteria = listFromRange vars!AB3:AB11 Criteria = Number between -66666 and 66666

Flag by color	Format→ Conditional Formatting → Custom formula is	`=if columnC says "number" make sure D & E are equal, otherwise flag red`
		`=if(C2="number",if(D2<>E2,1,0))`
		`=if columnC says "range" make sure D is less than E, otherwise flag red`
		`=if(C2="range",if(D2>=E2,1,0))`
		`=Format cells if… "cell is empty" - to highlight required fields.`

Link to more info	=hyperlink	`=hyperlink("https://websiteOfInterest.com","text you want to see in google sheet cell")`

Finding relevant papers in PubMed

Researchers used the public database PubMed to find full-length primary literature texts. Search terms frequently included "epigenetic", "odd ratio" and a type of lifestyle modification (i.e. "running", "vitamin D", "in vitro fertilization"). Once an appropriate paper was found, it was saved into a Zotero library accessible by all researchers.

Reading papers for relevent data

Data including odds ratios, p-values, confidence intervals, and explanations of tables and figures were highlighted in the PDF version of the paper; highlights were then visible to other researchers through Zotero. Ideal data showed a clear and statistically significant association between a lifestyle modification and a disease process that is known to be greatly influenced by epigenetics.

Inputing data into Google Sheets

Data was put into a Google Sheet only visible to the researcher who found the data in order to insure blinding (see Figure 1). One paper may have several rows of relevant data. As data is input cells turned from red to white in order for the researcher to know that a row was completed once all cells were white. Once all the data from a paper was properly input, the researcher put their name into another Google Sheet visible by all researchers titled "doneList" next to the bibliography of the paper; this allows other researchers to know when a paper is completed and is ready to be duplicated.

Duplication of papers

After one researcher finished extracting data from a paper and inputting it into their personal Google Sheet, a second researcher then opened the paper from Zotero and input data into their own Google Sheet. Duplication of papers ensured a minimal number of mistakes. The second researcher indicated that they were done duplicating a paper by putting their name in the doneList Google Sheet.

Figure-4 : Mock-up of smartphone app

Figure 4: Potential layout of a smartphone app utilizing data from the database. With this app, users would be able to conveniently and efficiently see a rank list of which lifestyle modifications are the best for their disorder, how strong of a recommendation these modifications are, and other useful data. Having this data in an easy-to-use app on a device a large number of people own can help patients take control of their health, potentially improving health while decreasing healthcare costs.

Conclusion

Our database of compiled epigenetic research can provide a groundwork for a smartphone app that would deliver large amounts of data to the user, allowing patients to take more control of their health and possibly decrease healthcare costs through lifestyle modifications. This app would clearly show a ranked list of which modifications are most effective for the patient's specific disease or genetic makeup.

Future Direction: Once input into a relational database the epigenetic database can have more functionality. Multiple unique inputs could be input with AND logic, for example if you both run three times a week and consume less than 1,500 calories what is the odd ratio of developing cancer. That could be compared to papers which have each of those inputs alone to decipher any synergistic or maxed out mechanisms. In addition allowance of user account creation will allow a wider audience to participate in database creation. Lastly the corresponding authors of papers can be invited to input their papers or proof data others have input to act as an authorized standard of quality control.

References

Kobayashi, L.C., Janssen, I., Richardson, H., Lai, A.S., Spinelli, J.J., and Aronson, K.J. (2013). Moderate-to-vigorous intensity physical activity across the life course and risk of pre- and post-menopausal breast cancer. Breast Cancer Research And Treatment 139, 851–861.

Junge, K.M., Bauer, T., Geissler, S., Hirche, F., Thürmann, L., Bauer, M., Trump, S., Bieg, M., Weichenhan, D., Gu, L., et al. (2016). Increased vitamin D levels at birth and in early infancy increase offspring allergy risk—evidence for involvement of epigenetic mechanisms. Journal of Allergy and Clinical Immunology 137, 610–613.

Panchenko, P.E., Voisin, S., Jouin, M., Jouneau, L., Prézelin, A., Lecoutre, S., Breton, C., Jammes, H., Junien, C., and Gabory, A. (2016). Expression of epigenetic machinery genes is sensitive to maternal obesity and weight loss in relation to fetal growth in mice. Clin Epigenetics 8.

Sheppard, V.B., Makambi, K., Taylor, T., Wallington, S.F., Sween, J., and Adams-Campbell, L. (2011). PHYSICAL ACTIVITY REDUCES BREAST CANCER RISK IN AFRICAN AMERICAN WOMEN. Ethn Dis 21, 406–411.

Rusiecki, J.A., Chen, L., Srikantan, V., Zhang, L., Yan, L., Polin, M.L., and Baccarelli, A. (2012). DNA methylation in repetitive elements and post-traumatic stress disorder: a case–control study of US military service members. Epigenomics 4.

More papers were input in the epigenetic database than are shown for data on this poster.

Contact Info

Sage Arbor
Assistant Professor of Biochemistry
Marian University College of Osteopathic Medicine
3200 Cold Spring Rd, Indianapolis, IN 46222
www.marian.edu/research/sage
Sarbor@Marian.edu
October 21st Marian University - Research Day