Continuation/Research Progression Projects Form (7)

Required for projects that are a continuation/progression in the same field of study as a previous project.

This form must be accompanied by the previous year's abstract and Research Plan/Project Summary.

Student's Name(s) Suchir Misra

To be completed by Student Researcher: List all components of the current project that make it new and different from previous research. The information must be on the form; use an additional form for previous year and earlier projects.

Components	Current Research Project	Previous Research Project: Year: 2018-2019
1. Title	Abstract text mining to create an exhaustive disease-disease correlation database	Meta-Analysis of cancer-related gene sets: Linking craniosynostosis and endometrial cancer
2. Change in goal/ purpose/objective	To elucidate the accuracy of the use of abstract text mining to create a database of disease-disease correlations for a list of 1857 diseases To elucidate disease-disease correlations for craniosynostosis as a proof of concept	To elucidate the role of craniosynostosis genes as predictive markers of cancer To elucidate a specific cancer via association of gene mutations to craniosynostosis
3. Changes in methodology	Creation of new database by: 1. Compilation of comprehensive list of diseases and conditions (N = 1857) 2. Collection of abstracts from PubMed for all genetic papers for each disease 3. Extraction of genes mutated in each disease from each abstract using text mining 4. Determination of significance of each disease-disease correlation through calculations of significance of gene overlap between diseases 5. Creation of a functional user interface to display all correlations from the database 6. Creation of disease networks to elucidate disease relationships based on their similarity to each other and to other diseases	1. Creation of craniosynostosis-mutated gene list using literature review of 15 sources encompassing 1514 craniosynostosis cases 2. Comparison of biologically curated gene sets to craniosynostosis-mutated gene list 3. Rankings of each cancer by mutation percentage from The Cancer Gene Atlas for each craniosynostosis-mutated gene
4. Variable studied	Presence of significant (p-value < 3.68*10^-8, 0.05/1,357,128, where 1,357,128 is the total number of disease-disease correlations) disease-disease correlations in literature to validate the accuracy of the newly created database Genetic distance ((sqrt(n1*n2))/m), where n1 is the number of genes mutated in the first disease, n2 is the number of genes mutated in the second disease, and m is the number of genes mutated in both to create the disease networks.	Efficacy of the use of craniosynostosis as a predictive marker for cancers only (no other diseases studied) Significance of correlation (p-value) between 33 different cancers and craniosynostosis
5. Additional changes	Includes disease-disease correlations for a large range of diseases (Ex: Alzheimer's Disease, Horner-Bernard Syndrome, Colorectal Cancer, etc.) Creates first disease-disease correlation database to use solely abstract text mining Creates disease networks that identify clusters of closely-related diseases for which similar treatments can be administered	Scope was limited to correlating craniosynostosis and cancer Existing databases were used to find a correlation, no new databases created

Attached are:

☑ Abstract and Research Plan/Project Summary, Year 2018-2019

I hereby certify that the above information is correct and that the current year Abstract & Certification and project display board properly reflect work done only in the current year.

Suchir Misra — Student's Printed Name(s)

Signature

11/30/19 — Date of Signature (mm/dd/yy)

OFFICIAL ABSTRACT and CERTIFICATION

Meta-Analysis of cancer-related gene sets: Linking craniosynostosis and endometrial cancer

Suchir Misra (2018-2019)
Jericho Senior High School, Jericho, NY, USA

Annually, over 22 billion dollars is devoted to congenital defect repair. Craniosynostosis, a congenital disease, shares genetic mutations with cancers but has not been studied as a potential predictive biomarker (i.e. craniosynostosis genetic mutations have been linked to cancer promoting pathways). This study investigated if craniosynostosis could be a marker of future cancer diagnosis. Molecular signatures in craniosynostosis-mutated gene lists were identified using the Molecular Signatures Database (MSigDB). Craniosynostosis may be a marker for cancer diagnosis based upon MSigDB' s identification of Kaposi' s sarcoma-associated herpesvirus (Kshv) Infection Angiogenic Markers Up signatures ($p < 1.96*10-9$) and Pathways in Cancer signatures ($p < 3.71*10-9$) prevalence in craniosynostosis-mutated genes. The Cancer Gene Atlas (TCGA) was then used to rank cancers (N = 33) by mutation percentage for each craniosynostosis-mutated gene. Craniosynostosis was highly correlated to endometrial cancer with an average rank of 1.12 while lower grade glioma produced the lowest rank of 22.5, indicating that craniosynostosis could be a predictive biomarker for endometrial cancer. Using TCGA, forty-five craniosynostosis-mutated genes had higher mutation percentages than half of the genes mutated in endometrial cancer (N = 22,162; $p<2.94*10-5$). This is the first study to elucidate a link between craniosynostosis and endometrial cancer for a large gene set; this connection calls for extensive screening for endometrial cancer in women born with craniosynostosis. Future investigations will elucidate the nature of the connection by investigating single-nucleotide polymorphisms to further clarify predictive markers for other congenital defects and cancer.

Category
Pick one only—
mark an "X" in box
at right

Animal Sciences	☐
Behavioral & Social Sciences	☐
Biochemistry	☐
Biomedical & Health Sciences	☐
Biomedical Engineering	☐
Cellular & Molecular Biology	☐
Chemistry	☐
Computational Biology & Bioinformatics	■
Earth & Environmental Sciences	☐
Embedded Systems	☐
Energy: Chemical	☐
Energy: Physical	☐
Engineering Mechanics	☐
Environmental Engineering	☐
Materials Science	☐
Mathematics	☐
Microbiology	☐
Physics & Astronomy	☐
Plant Sciences	☐
Robotics & Intelligent Machines	☐
Systems Software	☐
Translational Medical Sciences	☐

1. As a part of this research project, the student directly handled, manipulated, or interacted with (check ALL that apply):

   ☐ human participants ☐ potentially hazardous biological agents

   ☐ vertebrate animals ☐ microorganisms ☐ rDNA ☐ tissue

2. I/we worked or used equipment in a regulated research institution or industrial setting: ■ Yes ☐ No

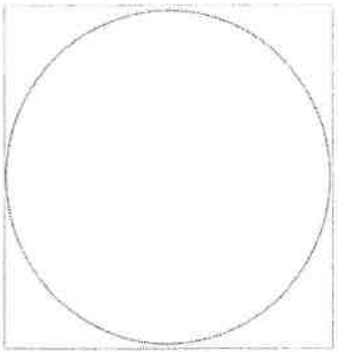
3. This project is a continuation of previous research. ☐ Yes ■ No

4. My display board includes non-published photographs/visual depictions of humans (other than myself): ☐ Yes ■ No

5. This abstract describes only procedures performed by me/us, reflects my/our own independent research, and represents one year's work only ■ Yes ☐ No

6. I/we hereby certify that the abstract and responses to the above statements are correct and properly reflect my/our own work. ■ Yes ☐ No

This stamp or embossed seal attests that this project is in compliance with all federal and state laws and regulations and that all appropriate reviews and approvals have been obtained including the final clearance by the Scientific Review Committee.



Suchir Misra

2018-2019

Meta-Analysis of cancer-related gene sets: Linking craniosynostosis and endometrial cancer

A. Rationale

Craniosynostosis is a facial disorder that occurs when cranial sutures, fibrous tissues composed of undifferentiated stem cells that lie between the major bones of the skull, prematurely fuse. Craniosynostosis affects 1 in 2100 to 1 in 2500 newborns[1-2]. A few major genes are mutated in craniosynostosis phenotypes. Fibroblast Growth Factor Receptor 2(FGFR2) is mutated in Crouzon Syndrome, which is characterized by a marked nose and a cleft palate in addition to craniosynostosis, and Apert Syndrome, which is characterized by shallow orbits, wide-set eyes and cardiac anomalies[3-5]. Fibroblast Growth Factor Receptor 3(FGFR3) is mutated in Muenke Syndrome, which is characterized by hearing loss and irregularly shaped bones. Finally, Twist Basic Helix-Loop-Helix Transcription Factor 1(TWIST1) is mutated in Saethre-Chotzen Syndrome, which is characterized by a low hairline, blocked tear ducts, and small ears[5]. FGFR2 and FGFR3 are essential genes in the Extracellular Signal-Regulated Kinase(ERK)-Mitogen Activated Protein Kinase(MAPK) pathway, which is implicated in different cancers, including melanoma[5,6]. TWIST1 is an important gene for the process of epithelial-mesenchymal transition, which is an important process for cancer progression and metastasis, making TWIST1 an important gene in a variety of cancers[7]. It is unknown if kids diagnosed with craniosynostosis are at a greater risk of being diagnosed with cancers when they get older as there has been no connection between the two pathologies themselves. This study will aim to find a connection between cancer and craniosynostosis by creating a comprehensive list of genes mutated in craniosynostosis and comparing the list to biologically curated gene sets to determine if craniosynostosis genes are overexpressed in cancer-related gene sets. If a significant overlap is found between the two lists, then it can be concluded that craniosynostosis is connected to cancer. A second aim of the study will be to link craniosynostosis to particular cancers using The Cancer Gene Atlas(TCGA) data.

B. Research Questions, Hypotheses, and Expected Outcomes

Hypotheses

- To determine if genes mutated in craniosynostosis are overrepresented in cancer-related gene sets
    - If a significant overlap ($p < 0.05/X$, where X is the number of gene sets in the collection) is found, craniosynostosis-mutated genes are connected to cancer.
    - Expected that there will be a significant overlap between cancer-related gene sets and genes mutated in craniosynostosis
- To establish correlations between craniosynostosis and particular cancers.
    - If there are a significant ($p < 0.05$) number of craniosynostosis-mutated genes above the median rank of genes mutated in that cancer, craniosynostosis is connected to that cancer.
    - Expected that there will be more than one cancer connected to craniosynostosis.

C

Procedures

Materials

- NIH DNA Sample Data
- NIH Gene Expression Omnibus(GEO) [8]
- iPython Version 6.4 [9]
- Molecular Signatures Database(MSigDB) Version 6.2 [10]
- The Cancer Gene Atlas(TCGA) Genomic Data Commons (GDC) Data Portal Version 12.0 [11]
- 2012 13-inch MacBook Pro

Procedure

Part 1 - Creating a Gene List

- Search "craniosynostosis" into the NIH DNA Sample Data and record each unique gene reported to be mutated in craniosynostosis.
- Search "craniosynostosis" into the NIH GEO and record each unique gene reported to be mutated in craniosynostosis.

- For each of three research articles, record each unique gene reported to be mutated in craniosynostosis [12-14].
- For each of ten review articles, record each unique gene reported to be mutated in craniosynostosis [2, 7, 15-22].
- Count number of sources each gene was reported to be mutated in craniosynostosis in , and count number of genes shown to be mutated in craniosynostosis across the studies.

Part II - Comparing the Gene List to Cancer-Related Gene Sets

- Download each collection of gene sets from MSigDB, except for the C1 Collection as those gene sets are only linked to chromosomal position.
- Write a program using the iPython interface to return the overlap between the gene list and each gene set in a given collection.
    - Open the gene sets from that collection and separate each gene set in the collection into a string.
    - Open the craniosynostosis gene list and make each gene a string, then create a list of the gene strings.
    - Create a for loop such that for each gene set, the genes within the gene sets get split into individual strings.
    - The strings from the gene set and the craniosynostosis gene list will then be matched to determine the overlap between the two.
- Count the number of genes that overlapped between the gene list and the gene sets.
- Write a program to perform a Fisher's Exact Test on all the gene sets in a given collection using the iPython interface.
    - Input all the necessary values for a Fisher's Exact Test into a text file.
    - Import necessary Python statistics libraries and open the input file.
    - Make each line of inputs its own separate string, then separate that string and make each number within that line its own integer.
    - Put the integer values into a contingency table, and calculate the p-value using the Fisher's Exact Test function.
    - Write the p-values into an output file.

- Use the Bonferroni correction on each p-value to filter down the number of significant gene sets.
- Repeat the above steps for each gene set collection (Hallmark Collection, C2-C7 Collections)

Part III - Determining Which Cancer Craniosynostosis Genes Have the Greatest Connection To

- Open the TCGA GDC Data Portal Website.
- Search for each craniosynostosis gene on the website.
- Each of the thirty-three cancers is ranked by percentage of cases in which a mutation of that gene was found - Record the ranking of each of the cancers for each craniosynostosis gene.
- Determine the average rank for each cancer in the database.
- For the cancer with the highest average rank, and cancers with lower average ranks, use the TCGA GDC Data Portal information on each of those cancers to rank the craniosynostosis genes out of all the genes mutated in those cancers.
- Perform Mood's Median Test to determine the significance of the number of the genes above the median rank as opposed to the number of genes below the median rank for both the cancers tested.

Risk and Safety

The only material to be used for this study is a computer, thus this study complies with minimal safety regulations. The only risk to the study is improper temperature modulation leading to overheating or malfunctioning of the computer, which would not pose a threat to the other occupants of the lab.

Data Analysis

Once the data is obtained and the number of genes that overlap between each gene set is recorded, then the significance of the overlap will be analyzed using a Fisher's Exact Test. Since there are thousands of gene sets in some collections of MSigDB, there is a chance that there will be hundreds of significant gene sets if the threshold of significance is set at $p < 0.05$. Thus, the Bonferroni correction will be performed on each p-value to filter down the number of significant gene sets. Once the rank of each craniosynostosis gene out of all the genes mutated in the cancer

with the highest average rank and a lower-average rank, a Mood's median test will be performed to determine the significance of the number of genes above the median rank for each cancer.

NO ADDENDUMS EXIST

D. Bibliography

[1] D. Johnson and A.O.M. Wilkie, "Craniosynostosis," *European Journal of Human Genetics*, vol. 19, no. 4, pp. 369-376, April 2011.

[2] M.A. Katsianou, *et al.*, "Signaling mechanisms implicated in cranial sutures pathophysiology: Craniosynostosis," *BBA Clinical*, vol.6, pp. 165-176, April 2016.

[3] F. Ursitti, *et al.*, "Evaluation and management of nonsyndromic craniosynostosis," *Acta Paediatrica*, vol. 100, pp. 1185-1194, April 2011.

[4] M.J. Pfaff, *et al.*, "FGFR2c-mediated ERK-MAPK activity regulates suture development," *Developmental Biology*, vol. 415, pp. 242-250, March 2016.

[5] S.R.F. Twigg and A.O.M Wilkie, "A genetic-pathophysiological framework for craniosynostosis," *The American Journal of Human Genetics*, vol. 97, pp. 359-377, Sep. 2015.

[6] C. Welbrock and I. Arozarena, "The complexity of the ERK/MAP-Kinase pathway and the treatment of melanoma skin cancer," *Frontiers in Cell and Developmental Biology*, vol. 4, no. 6, pp. 1-9, April 2016.

[7] E.C. Je, B.S. Lca, and G.A. Ga, "The role of transcription factor TWIST in cancer cells," *Genetic Syndromes and Gene Therapy*, vol. 4, no. 1, pp. 1-7, Jan. 2013.

[9] R. Edgar, M. Domrachev, and A.E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207-210, Jan. 2002.

[9] F. Pérez, and B.E. Granger, "IPython: A system for interactive scientific computing," *Computing in Science and Engineering*, vol. 9, no. 3, pp. 21-29,

May 2007.

[10] A. Subramanian, *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of America*, vol. 102, no. 43, pp. 15545-15550, Sep. 2005.

[11] R. Grossman, *et al.,* "Toward a shared vision for cancer genomic data," *The New England Journal of Medicine*, vol. 375, no. 12, pp. 1109-1112, Sep. 2016

[12] V.P. Sharma, *et al.*, "Mutations of TCF12, encoding a basic-helix-loop-helix partner of TWIST1, are a frequent cause of coronal craniosynostosis," *Nature Genetics*, vol. 45, no. 3, pp. 304-307, March 2013.

[13] Y. Xu, *et al.*, "Identification and analysis of the genetic causes in nine unrelated probands with syndromic craniosynostosis," *Gene*, vol. 641, pp. 144-150, Jan. 2018.

[14] K.A. Miller, *et al.*, "Diagnostic value of exome and whole genome sequencing in Craniosynostosis," *Journal of Medical Genetics*, vol. 54, no. 4, pp. 1-9, Oct. 2016.

[15] W. Lattanzi, *et al.*, "Genetic advances in craniosynostosis," *American Journal of Medical Genetics Part A*, vol. 173, no. 5, pp. 1406-1429, May 2017.

[16] A.V. Ciruea and C. Toader, "Genetics of craniosynostosis: review of the literature," *Journal of Medicine and Life*, vol. 2, no. 1, pp. 5-17, Jan. 2009.

[17] A. Jezela-Stanek and M Krajewska-Walasek, "Genetic causes of syndromic craniosynostosis," *European Journal of Pediatric Neurology*, vol. 17, no. 3, pp. 221-224, May 2013.

[18] A.O.M Wilkie, D. Johnson, and S.A. Wall, "Clinical genetics of craniosynostosis," *Current Opinion in Pediatrics*, vol. 29, no. 6, pp. 622-628, Dec. 2017.

[19] A.J. Yoon, B.N. Pham, and K.M. Dipple, "Genetic screening in patients with craniofacial malformations," *Journal of Pediatric Genetics*, vol. 5, no. 4, pp. 220-224, Dec. 2016.

[20] K. Senarath-Yapa, *et al.*, "Craniosynostosis: molecular pathways and future pharmacological therapy," *Organogenesis*, vol. 8, no. 4, pp. 103-113, Oct. 2012.

[21] H. Melville, *et al.*, "Genetic basis of potential therapeutic strategies for craniosynostosis," *American Journal of Medical Genetics Part A*, vol. 152A, no. 12, pp. 3007-3015, Dec. 2010.

[22] H. Chim, *et al.*, "Molecular signaling in pathogenesis of craniosynostosis: the role of Fibroblast growth factor and transforming growth factor-β," *Neurosurgical Focus*, vol. 31, no. 2, Aug. 2011.