

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Utilization and regulation of carbon sources and their metabolism in the halophilic archaeon Halogeometricum borinquense

Permalink

<https://escholarship.org/uc/item/6c95q6bs>

Author

Kim, Hyunsoo Gloria

Publication Date

2022

Peer reviewed|Thesis/dissertation

Utilization and regulation of carbon sources and their metabolism
in the halophilic archaeon *Halogeometricum borinquense*

By

HYUNSOO GLORIA KIM

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Microbiology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Marc T. Facciotti, Chair

Anthony D. Albano

Sean R. Collins

Dominik R. Haudenschild

Committee in Charge

2022

Copyright © 2022 by

Hyunsoo Gloria Kim

All rights reserved.

Proprietary Rights

The work presented in this dissertation has been previously published in peer-reviewed journals. Chapters 1, 2, and 6 are original contributions, with Chapters 2 and 6 serving as works in progress, in review with co-authors before publication at the time of writing. The references for the work associated with the remaining chapters are listed below. All articles are published under Creative Commons Attribution licenses, which allows reuse of the material with proper attribution.

Chapter 3

- Furrow, R. E., Kim, H. G., Abdelrazak, S. M., Dahlhausen, K., Yao, A. I., Eisen, J. A., Goldman, M. S., Albeck, J. G., & Facciotti, M. T. (2020). Combining Microbial Culturing With Mathematical Modeling in an Introductory Course-Based Undergraduate Research Experience. *Frontiers in microbiology*, 11. doi: [10.3389/fmicb.2020.581903](https://doi.org/10.3389/fmicb.2020.581903)
- Rodriguez-Medina, J., Kim, H. G., Castro, J., Contreras, C. M., Glon, C. L., Goyal, A., Guo, B. Y., Knowles, S., Lin, J. C., McGuinness, C. L., Sorkin, E., Stefani, J., Yegireddi, S. J., Chaganti, S., Cui, D., Deck, S. L., Deokule, Y., Douglas, H., Kenaston, M., O'Brien, A., Patterson, E., Schoppa, N., Vo, D. T., Tran, K., Tran, T., Pérez-Irizarry, V., Carrasquillo-Nieves, K., Montalvo-Rodriguez, R., Yao, A. I., Albeck, J. G., Facciotti, M. T., Nord, A. S., & Furrow, R. E. (2020). Draft genome sequences of 16 halophilic prokaryotes isolated from diverse environments. *Microbiology resource announcements*, 9(8), e01540-19. doi: [10.1128/MRA.01540-19](https://doi.org/10.1128/MRA.01540-19)

Chapter 4

- Shusterman, E., Kim, H. G., Facciotti, M., Igo, M., Sripathi, K., Karger, D., Segal, A. & Gal, K. (2021, April). Seeding Course Forums using the Teacher-in-the-Loop. In LAK21: 11th International Learning Analytics and Knowledge Conference (pp. 22-31). doi: [10.1145/3448139.3448142](https://doi.org/10.1145/3448139.3448142)

Chapter 5

- Geller, S. A., Hoernle, N., Gal, K., Segal, A., Kim, H. G., Karger, D., Facciotti, M. T., Sripathi, K. S., & Igo, M. (2021). New Methods for Confusion Detection in Course Forums: Student, Teacher and Machine. *IEEE Transactions on Learning Technologies*. doi: [10.1109/TLT.2021.3123266](https://doi.org/10.1109/TLT.2021.3123266)

Dedication

This dissertation is dedicated to those who have filled my life with love —
most of all: my mother, my grandparents, and Martin!

Land Acknowledgement

It has been a great privilege to learn and work at the University of California, Davis. At the same time, there is some complexity with regard to knowing the history of the land and understanding our place within that context. The UC Davis land acknowledgement was developed originally by Yocha Dehe Wintun Nation and approved by all three Patwin tribes – it reads as follows:

We should take a moment to acknowledge the land on which we are gathered. For thousands of years, this land has been the home of Patwin people. Today, there are three federally recognized Patwin tribes: Cachil DeHe Band of Wintun Indians of the Colusa Indian Community, Kletsel Dehe Wintun Nation, and Yocha Dehe Wintun Nation.

The Patwin people have remained committed to the stewardship of this land over many centuries. It has been cherished and protected, as elders have instructed the young through generations. We are honored and grateful to be here today on their traditional lands.

Land acknowledgements act only as starting points for acknowledging the historical and current, modern day systemic oppression and violence towards Indigenous communities. More vital than a repeated passage is learning about history and other cultures, reflecting on personal values and goals, and identifying ways to materially support vulnerable communities, as well as our environment.

The Native Governance Center has created a resource titled, “Beyond Land Acknowledgment: A Guide” (accessible at: nativegov.org/news/beyond-land-acknowledgment-guide). This guide provides an overview on why land acknowledgments can be problematic and what it means to go beyond land acknowledgments, outlining self-assessment questions and concrete next steps (e.g., conducting resource assessments, setting up monthly recurring donations, etc.), with explanations, to help one create an action plan. The NDN Collective’s website also provides many educational resources, in addition to information about ongoing climate justice and land back movements (ndncollective.org). Robin Wall Kimmerer (robinwallkimmerer.com) and Nick Estes (nickestes.blog/about) have also written beautiful and informative works that have helped provide structure to my personal framework. In addition to these people and organizations, I am thankful to the ancestors and those among us who have spoken out against injustices and prioritized learning, educating, and building community and kinship with others.

Personal Acknowledgements

This work was only possible through the support of a long list of people. Though I only name a fraction below, I am forever grateful to everyone who has shown me kindness, patience, respect, and kinship, prior to and throughout my graduate education.

I first extend my sincerest gratitude to my advisor, Marc Facciotti, for his support and mentorship. His contagious excitement about science, his thoughtful approaches towards teaching in the classroom and in the lab, and his encouragement to pursue research-external goals made a tremendous impact on my development, and is deeply appreciated. It has been a privilege and pleasure to learn from and work alongside such a genuine, dedicated, kind, and humorous person.

The other members of the Facciotti Lab deserve my wholehearted gratitude as well. I am thankful to Andrew Yao and Daniel Facciotti for their scientific insight, morale boosting, life advice, and the *plethora* of good stories. Andrew is deserving of extra acknowledgment for his contributions to the halophile work, and the kindness and support he has shown to me and countless other students. I am grateful also to Kamali Sripathi, who in addition to providing thoughtful research and writing feedback, kept me grounded with her compassion and encouragement. Additionally, a heartfelt thanks goes to Haven Rittershofer, Rachel “P.” Tanczos, and Sienna Wu, for their immense contributions and help in the lab, for everything they also taught me, and all the ways they brightened my day in and out of the lab.

I would also like to express much appreciation to my research collaborators near and far for their insight, contributions, and support otherwise. I am lucky to have had the opportunity to work with so many clever and sincere people. For the *H. borinquense* work, I thank Mara Cuebas-Irizarry and Rafael Montalvo-Rodríguez. For the CURE work, I thank John Albeck, David Coil, Jonathan Eisen, David Furlow, Robert Furrow, Mark Goldman, and Ashley Vater. For the Nota Bene work, I thank Anthony Albano, Jumana Almahmoud, Ariel Blobstein, Kobi Gal, Shay Geller, Michele Igo, David Karger, and Einat Shusterman.

Teaching was one of the most joyful and fulfilling experiences of graduate school for me largely because I had the opportunity to meet so many bright and intentional students, whom I learned so much from as well, and hold deep appreciation for. I also hold great appreciation for the Bis2A teaching team for supporting my growth as an educator. I am especially thankful for Erin Easlon for her mentorship,

camaraderie, and thoughtfulness. Additionally, I am so delighted to have been paired with Eduardo Ruiz, who was everything you could wish for in a Learning Assistant and more. I am immensely thankful for his support and friendship over the years.

I am grateful to have received so much support from people I met outside of lab and teaching spaces as well. Thank you to the California Initiative to Advance Precision Medicine team, especially my very supportive policy fellowship supervisors, Julianne McCall and Shannon Muir. Thank you to the *also* very supportive STEM for Girls team, especially Tina Jeoh and Sara Blair-Medeiros. Thank you to the University of California Graduate and Professional Council, especially Gwen Chodur. Thank you to the folks I started and ended my pre-pandemic days with on the bus to/from Davis, and thank you to the Across the Causeway folks from whom I learned so much about local advocacy alongside. Thank you to other students, faculty, and staff at UC Davis who contributed to my education and growth, especially my course professors and dissertation committee.

My graduate experiences have affirmed the incredible power of friendship. In addition to friends already named above, I will limit thanks to just two more. For her exceptional kindheartedness, the bottomless laughter and all of the incredible memories we have shared, I thank Jessica Lumian. For always being their full, intentional self, and for their earnest mind that centers love, I thank James Salanga. To have more people to thank than space to thank them is a blessing and a curse. Thank you to my gracious unnamed friends for nevertheless knowing my deep appreciation for them.

Finally, I am extremely grateful for all of the love and support my family has shown me (in-laws very much so included). For fortifying a knack for iteration and patience (useful traits for research!) through hours and hours of Overcooked, and for being my first friend in life, I thank my sister. For their generous spirits and sacrifices that allowed me to pursue graduate school, for their wisdom and advice, for their unconditional love and care, and for always believing in me and my goals and encouraging me, I am profoundly grateful to both my grandparents and my mother.

Last but not least, for his unwavering support and above and beyond generosity over the better half of a decade, I thank the love of my life, Martin, from the very bottom of my heart. I am so, so extremely lucky to have been loved and supported by such a compassionate, patient, clever, goofy, and thoughtful soul. Thank you for letting me be a part of your journey and for being the sunshine in mine.

**Utilization and regulation of carbon sources and their metabolism
in the halophilic archaeon *Halogeometricum borinquense***

Abstract

Gene regulatory systems contribute to an organism's fitness by providing a way to dynamically tune cellular responses in response to fluctuating environmental conditions. In this work, we compare the regulatory tuning of genes involved in the catabolism of glucose, maltose, and sucrose in the metabolically versatile halophilic archaeon *Halogeometricum borinquense* through functional transcriptomic profiling, targeted qRT-PCR, and sugar utilization assays. This work provides the first transcriptomic study for this organism, preliminarily characterizes the unique and shared responses of *H. borinquense* to single- and multiple-sugar growth conditions, identifies key metabolic regulatory points in the organism, and advances the discussion of archaeal carbon catabolite repression (CCR) by identifying putative targets of CCR that can be probed in future work to further explore underlying mechanisms.

Halophilic organisms are also suitable for use in educational settings. Here, I present the development and study examining the impact of a quantitative biology-focused Course Based Undergraduate Research Experience (CURE) on the development of student interest in and self-efficacy for quantitative biology. Students in the course isolate halophilic organisms from environmental samples, quantitatively characterize their growth, and thereafter sequence the genomes of the isolates. We found that although students initially reported high math-biology values, we saw improvements with regard to students seeing higher value in using quantitative calculations in biology and increased interest in pursuing future quantitative experiences (e.g., coursework, occupation).

Lastly, I present collaborative research that explores how students express affect when they engage with one another in the context of an online course textbook-based annotation and discussion forum. This work includes efforts to use machine learning to predict student affect and better understand student confusion, to increase student engagement with the course content by seeding forums, and to create and validate a limited emoji-hashtag pair set that enables more reliable student-driven affect tagging of forum posts.

Table of Contents

Proprietary Rights.....	ii
Dedication.....	iii
Land Acknowledgement.....	iv
Personal Acknowledgements.....	v
Abstract.....	vii
Table of Contents.....	viii
Introduction	1
Chapter 1. Introduction To Metabolic Regulation In Haloarchaea	5
1.1 Introduction	5
1.2 Introduction To Archaea	7
1.3 Introduction To Halophilic Archaea	9
1.4 The Haloarchaea As Model Organisms	10
1.5 Introduction To Haloarchaeal Metabolism.....	12
1.6 CCR / Metabolic Regulation	15
1.6.1 Overview Of CCR In Bacteria And Eukaryotes	16
1.6.1.1 The Textbook Mechanisms: <i>E. coli</i> And <i>B. subtilis</i>	16
1.6.1.2 CCR When Glucose Is Not the Primary Carbohydrate And In Other Bacteria	18
1.6.1.3 CCR In Eukaryotes	19
1.6.1.4 Summary Of CCR In Bacteria And Eukaryotes	20
1.6.2 Introduction To CCR In Archaea	20
1.6.2.1 CCR In <i>Sulfolobus solfataricus</i>	21
1.6.2.2 CCR In Other Archaea.....	24
1.6.2.3 Evidence Of CCR In Haloarchaea	24
1.7 References	27
Chapter 2. Carbon Utilization And Regulation In The Halophilic Archaeon	
<i>Halogeometricum borinquense</i>.....	33
2.1 Publication Status.....	33
2.2 Author Contributions.....	33
2.3 Introduction	34
2.4 Materials & Methodology	35
2.4.1 Strain And Growth Conditions	35
2.4.2 RNA-Seq.....	36
2.4.3 Bioinformatics	39
2.4.4 Phenotyping, Sample Collection, And Metabolite Analysis	39
2.4.5 Quantitative PCR	41
2.5 Results.....	42
2.5.1 Culture Phenotypes	42
2.5.2 OD ₆₀₀ , CFU Trajectory Phenotypes, And Sugar Utilization	42
2.5.2.1 OD ₆₀₀ And CFU Trajectory Phenotype	43
2.5.2.2 Sugar Utilization.....	46
2.5.3 RNA-Seq-Based Phenotyping	47
2.5.3.1 ABC-type Transporter Proteins.....	50
2.5.3.2 Glycosyl Hydrolases And Glycosyl Transferases	52
2.5.3.3 Central Carbon Metabolism And Gluconeogenesis	54
2.5.3.4 TCA Cycle And Glyoxylate Pathway	56
2.5.3.5 Nitrogen Metabolism.....	57
2.5.3.6 Other Functional Categories	58
2.5.3.7 Functional Coordination Of Co-Regulated Gene Expression.....	61
2.5.3.8 RNA-Seq Summary	62

2.5.4	Identifying Putative Glucose-Influenced Transcriptional Changes.....	63
2.5.4.1	Glucose-Induced Activation Of Gene Expression	64
2.5.4.2	Glucose-Induced Repression Of Gene Expression.....	65
2.5.4.3	Alternative Sugar- And Growth-Dependent Effects.....	67
2.5.4.4	Summary Of Targeted qRT-PCR Data.....	68
2.6	Discussion	71
2.7	Funding.....	73
2.8	References	74
Chapter 3.	Halophiles In Course-based Undergraduate Research Experiences (CUREs)	78
3.1	Publication Status.....	78
3.2	Author Contributions.....	79
3.3	Introduction	80
3.4	The Importance Of Undergraduate Research Experiences In STEM Curriculums	81
3.5	Overview Of Course-based Undergraduate Research Experiences (CUREs).....	85
3.6	“Genome Hunters” Course Framework And Curriculum Development Goals.....	89
3.6.1	Determining A Research Work Flow	89
3.6.2	Halophiles As Study Subjects In CUREs.....	96
3.6.3	Course Overview And Learning Objectives.....	97
3.6.4	Detailed Methodology	103
3.7	“Genome Hunters” Fall 2018 Course Outcomes.....	104
3.7.1	Draft Genome Sequences Of 16 Halophilic Prokaryotes Isolated From Diverse Environments	104
3.7.2	Student Attitudinal Outcomes	106
3.8	Future Directions For “Genome Hunters” CURE	107
3.9	Future Directions For CURE Course Work	110
3.10	Funding.....	111
3.11	Acknowledgements	111
3.12	References	113
Chapter 4.	Encouraging Collaborative Learning While Reading The Online Course	
Textbook With An Integrated Discussion Forum		117
4.1	Publication Status.....	117
4.2	Author Contributions.....	118
4.3	Introduction	118
4.4	The Rise Of Online Learning Modalities And Tools	119
4.5	The Development Of Online Textbooks	120
4.6	The Significance Of Collaborative Or Social Learning	123
4.7	Introduction To Nota Bene	125
4.8	Using Collaborative Nota Bene Features As Motivation For Students To Engage With The Course Text Before Class	127
4.9	Encouraging Deeper Engagement With Course Readings By Seeding Discussion Forums With Pre-Selected Posts	128
4.9.1	Related Work	130
4.9.1.1	Interventions For Improving Learning In Online Forums	130
4.9.1.2	Modeling Student Discussions.....	130
4.9.2	Using Nota Bene In The Context Of Seeding Experiments	132
4.9.3	Selecting Threads For Seeding	134
4.9.4	Computational Models	136
4.9.5	Online Seeding Experiment.....	140
4.9.5.1	Experimental Methodology	140
4.9.5.2	Hypothesis And Results.....	143
4.9.5.3	Discussion And Future Work	145
4.10	Funding.....	147
4.11	References	148

Chapter 5. Leveraging Affective Reporting In Discussion Forum Posts To Understand Student Confusion While Reading The Course Text	151
5.1 Publication Status	151
5.2 Author Contributions	152
5.3 Introduction	152
5.4 Affective Reporting Using Emojis And Hashtags As A Way To Tag And Filter Categories Of Student Discussion Forum Posts	154
5.4.1 Affective Reporting Via Emojis In Educational Settings	154
5.4.2 Affective Reporting Using Nota Bene Emojis And Hashtags	156
5.5 New Methods For Confusion Detection In Course Forums	158
5.5.1 Introduction	158
5.5.2 Overview Of The Nota Bene Dataset	162
5.5.3 Inter-Labeler Agreement When Using Labeling Tree	165
5.5.4 Agreement Between Students' Self-Reported Hashtags And Experts	168
5.5.5 Automatic Classification Of Confused Posts	171
5.5.6 Discussion	173
5.5.6.1 Labeling Confusion In Forum Posts	173
5.5.6.2 Automatically Detecting Confusion In Forum Posts	174
5.5.6.3 Pedagogical Implications	175
5.5.6.4 Theoretical Implications	176
5.5.7 Conclusion And Future Work	177
5.6 Funding	178
5.7 References	179
Chapter 6. Supporting Students' Self-Reporting Of Affective States In Online Course Forums And Text Annotation Platforms With Emoji Tags	181
6.1 Publication Status	181
6.2 Author Contributions	182
6.3 Introduction	182
6.4 Methodology & Data Collection	184
6.4.1 The NB Discussion Interface	184
6.4.2 Course Format And Study Participants	185
6.4.3 Survey Design, Administration, And Quality Control	187
6.5 Results	190
6.5.1 Original Emoji Evaluation (Survey 1)	190
6.5.2 Modified Emoji Set Evaluations (Surveys 2 And 3)	196
6.5.3 Final Emoji Set Evaluations (Survey 4)	202
6.6 Discussion	206
6.7 Funding	209
6.8 Accessing Material	209
6.9. References	210
Table Of Contents: Appendix Items In Chapter 2	211
Table Of Contents: Appendix Items In Chapter 6	212

Introduction

This dissertation examines several diverse questions, including those surrounding carbon catabolite repression (CCR), *Halogeometricum borinquense* sugar utilization and gene regulation, using halophiles as model organisms for study in innovative lab-based curriculums, such as Course-based Undergraduate Research Experiences (CUREs), and leveraging and improving online technologies to enhance student learning and affective expression in a general biology course. The main chapters of this dissertation are summarized below.

Chapter 1: Introduction To Metabolic Regulation In Haloarchaea

Regulation of gene expression is an important biological process that organisms carry out ubiquitously across the bacteria, eukarya, and archaea. CCR is one of the most well-studied metabolic regulatory paradigms, however, the vast majority of the work has been conducted on bacteria and eukaryotes. Much less is known about archaeal regulation, despite their unique biology and prolific distribution throughout the world. Haloarchaea are a suitable model archaeal group to study CCR, with their general ease of growth requirements and handling, expansive experimental biochemical toolkit, and increasing sequencing performed on haloarchaeal strains. This chapter provides a review on the background and literature surrounding archaeal metabolism and CCR.

Chapter 2: Carbon Utilization And Regulation In The Halophilic Archaeon *Halogeometricum borinquense*

The metabolically versatile, carbohydrate-utilizing *Halogeometricum borinquense* provides an opportunity to further explore carbon catabolite repression in halophilic archaea. This chapter details a study that explores the regulation of genes involved in glucose, maltose, and sucrose catabolism in *H. borinquense* and serves as the first transcriptomic study for this organism. A comparative analysis of carbohydrate metabolism is carried out through culturing

H. borinquense on each of the sugars (and combinations thereof), functional transcriptomic profiling, targeted qRT-PCR, and sugar utilization assays. This work provides a preliminary characterization of the unique and shared responses of *H. borinquense* to single- and multiple-sugar growth conditions, identifies key metabolic regulatory points, and advances the discussion of archaeal CCR by identifying putative targets of CCR that can be probed in future work to further explore underlying mechanisms.

Chapter 3: Halophiles In Course-based Undergraduate Research Experiences (CUREs)

In addition to their use in research settings, halophiles are also suitable organisms for study in educational settings, such as Course-based Undergraduate Research Experiences, or CUREs. CUREs provide an opportunity for hands-on laboratory learning and represent one avenue to make research experiences more accessible to a larger number of students. CUREs allow for students to participate in short-term research projects (as opposed to plug-and-chug lab curriculums), with the goal of giving students an increased sense of ownership of their learning. This may in turn improve student attitudes towards learning various skills related to science, including quantitative analyses, a topic known to be challenging or disproportionately avoided by undergraduate biology students.

This chapter details one example of a CURE developed at UC Davis, where halophilic organisms are isolated, cultured, quantitatively analyzed, and thereafter sequenced. Chapter 3 represents the introduction to my educational research work. Other education-based research entailed back-end course tool and analytics development, the focus Chapters 4-6.

Chapter 4: Encouraging Collaborative Learning While Reading The Online Course Textbook With An Integrated Discussion Forum

The rise and development of online learning technologies, further expedited by the COVID-19 pandemic, is motivating educators to reflect on using online learning tools effectively in courses. Textbooks and discussion forums are ubiquitous tools used in educational settings

as mechanisms of knowledge transfer and feedback provision. New technologies allow for some of these tools to be integrated – for example, Nota Bene (NB) hosts a discussion forum in the margins of an online, open-access textbook. Social, collaborative learning during traditionally individual processes like reading the textbook can be beneficial for improved engagement with the material and learning gains for the students. Initializing the course forum with suitable seed candidates consisting of posts from previous terms can encourage higher quality discussion.

Chapter 5: Leveraging Affect Reporting In Discussion Forum Posts To Understand Student Confusion While Reading The Course Text

In addition to the ability to interact with peers and instructors while anchoring comments to specific portions of the course text, common social affect reporting mechanisms (e.g., emojis and hashtags) can be integrated into educational platforms like NB to encourage students to express their feelings while learning, which can in turn help instructors gauge student sentiment and needs and guide course flow and potential revisions to the text. Emojis are used in NB to provide this function, and can simultaneously allow instructors to simplify analysis of large volumes of comments from discussion forums, which might otherwise be too cumbersome to manually read through and prioritize responses or interventions to otherwise. In the study discussed in this chapter, emojis contribute to the training of a machine learning classifier to automatically detect student confusion in untagged posts.

Chapter 6: Supporting Students' Self-Reporting Of Affective States In Online Course Forums And Text Annotation Platforms With Emoji Tags

In providing affect reporting tools and using them in discussion forum data analyses as in Chapter 5, it is important to validate that the emoji is a reliable indicator of the student state, given the multiple interpretations that may exist for many emoji. NB addresses this by pairing emojis with hashtags that are used to tag the discussion post with a more constrained meaning. This chapter details the development of a unique set with goals of highly associable emoji-

hashtag pairings with clear, non-redundant functions that accurately report sentiments that general biology students at UC Davis find meaningful to express while reading the course text and interacting with others in the course using NB.

Chapter 1. Introduction To Metabolic Regulation In Haloarchaea

1.1 Introduction

Archaea are a group of unicellular organisms found abundantly in diverse extreme environments but also as more minor constituents of many other microbiomes (Moissl-Eichinger, 2018; Bang, 2018). The Archaea were discovered to be a phylogenetically distinct clade of microbes in the 1970's by Woese et al. (1977) through an analysis of 16S ribosomal RNA (rRNA) sequences. At the time, Woese et al. suggested that Archaea form a third domain of life, alongside Eukarya and Bacteria (Woese, 1977; Woese, 1990). This three domain view of life was dominant for nearly 45 years. However, more recent phylogenetic analyses have instead pointed to a two-domain model that includes Bacteria and Archaea, with the Eukaryotes emerging from “advanced” members of the Archaea (Doolittle, 2020).

Due to the relatively recent discovery of the Archaea as a unique evolutionary clade, their extreme growth requirements, and the relative paucity of genetic tools for experimentally probing the Archaea, little work has been done to understand archaeal biology when compared to that done in traditional bacterial model systems like *Escherichia coli* and *Staphylococcus aureus*, and eukaryotic systems like *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and others. Given their evolutionary relationships to other forms of life, their contributions to the physiology of most microbial communities, and their unique repertoire of biological functions, studying the Archaea can help us close gaps in our functional and evolutionary understanding of life on Earth.

Our current understanding of metabolic regulation in Archaea remains comparatively limited, particularly the understanding of how archaeal organisms regulate gene expression in response to one or more nutrient sources (e.g., carbohydrates, amino acids) (Görke, 2008;

Bertrand, 1975). Rather, most of our understanding of such regulatory principles come from work in bacterial species.

Pioneering work by Jacob and Monod with *E. coli* in the early 1960's provided one early model of catabolic regulation. After growing *E. coli* on different carbohydrates and pairwise combinations of carbohydrates as carbon sources, they observed a growth pattern characterized by two distinct phases called diauxie (Jacob, 1961). In the first growth phase, an initial burst of cellular division is driven by the utilization of a primary/preferred sugar, most famously glucose, as a carbon and energy source. This initial growth spurt is followed by a short lag (adaptation) phase as the cell extinguishes the preferred sugar and converts its metabolism to consume the secondary sugar in a new growth phase.

This early work demonstrated that *E. coli* could actively choose to metabolize one food source and not metabolize another when both were present, only making the decision to metabolize the second (less favored) food after the more favored one was depleted. Further work determined that a critical mechanism used by *E. coli* to make these types of decisions involved the preferred carbon source(s) actively repressing the catabolic activities associated with the utilization of any secondary sugar(s) present. This type of regulatory relationship has become known as carbon catabolite repression, or CCR. While this phenomenon has been seen in diverse organisms, the complex molecular mechanisms of CCR and related metabolic regulation have been most extensively characterized in model bacteria like *E. coli* or eukaryotes like *S. cerevisiae*. Comparatively far less is known about CCR in archaea.

This chapter provides more detailed background and perspectives about archaea, CCR, and what is known about metabolic regulation and CCR in the archaeal domain. The background provides important context for an original study presented in Chapter 2 on the carbohydrate-dependent regulation of gene expression in the haloarchaeon *Halogeometricum borinquense* where evidence for CCR is observed.

1.2 Introduction To Archaea

“Archaea” is a term coined in the late 1970’s by Woese et al. after analyses of signatures of rRNA gene sequences from this group of organisms showed distinctiveness from bacterial and eukaryal organisms (Woese, 1977; Woese et al, 1990).¹ Archaea are most famously known to be abundant in extreme environments (e.g., high or low temperature, pH, salt, etc.) and/or capable of performing unique or specialized metabolic processes (e.g., methanogens, which reduce carbon dioxide to methane). However, they are also regularly present as constituents of other, less extreme microbial communities.

Initially, archaeal rRNA genes used to classify archaeal organisms were organized into two main phyla, Euryarchaeota and Crenarchaeota. More recently, data from whole genomes (4,141 archaeal genomes have been sequenced as of August 2021 and can be found at Genome Online Database: <http://genomesonline.org/>) and environmental sequencing efforts have led to the creation of 27 phyla within the Archaea (Adam, 2017). While only six of those 27 phyla have cultured representatives (Baker, 2020), improvements in sequencing capabilities over the last two decades have allowed researchers to peer into the physiological, metabolic and ecological diversity of the uncultivated archaea.

Structurally and genetically, the Archaea share some properties with both bacteria and eukaryotes, while also maintaining some unique features. Like bacteria, they are unicellular, lack lipid-based membrane-bound organelles, and they have a similar cell size and shape, and DNA structure (e.g., one circular chromosome, operon structures, plasmids). Archaea have seemingly mosaic genomes, possessing elements that show significant homology to eukaryotic information processing machinery, and bacterial-like features encoding operational, or

¹ In bacteria and archaea, the gene for the small subunit ribosomal RNA is known as the 16S rRNA gene, whereas in eukaryotes, it is known as the 18S rRNA gene. Comparing the sequences of the 16S/18S rRNA gene is useful for studying evolutionary relationships due to the existence of regions with both low and high conservation. Furthermore, the number of available genome sequences to make comparisons with are growing increasingly larger in size from both cultured and environmental organisms (Lang, 2013).

housekeeping functions (Leigh, 2011). Most archaeal mechanisms involved in information processing (e.g., replication, transcription, repair, and translation) are generally regarded as less complex versions of their eukaryotic counterparts. This observation underpins some of the evolutionary hypothesis that suggests that bacteria evolved from archaea (Doolittle, 2020; Bräsen, 2014; Leigh, 2011; Bell, 2001).

Gene regulatory systems provide a great example of the chimeric nature of the genome. Archaeal basal transcription machinery are generally more complex than those of their bacterial counterparts and more closely resemble the components used in eukaryotic transcription (Bell, 2001). Zillig et al. (1979) purified archaeal RNA polymerase (RNAP) and showed that RNAP subunit gene products were highly homologous to subunits previously thought to be uniquely eukaryal. Also similar to the more complex eukaryotic nuclear RNAPs, archaeal RNAPs require a subset of additional basal factors in order to reconstitute transcription initiation on most archaeal promoters *in vitro*. These proteins have been shown to be homologous to the eukaryal general transcription factors, TATA-binding proteins (TBP) and TFIIB, both of which are necessary for promoter recognition and specific transcription (Geiduschek, 2005; Baliga, 2000; Facciotti, 2007; Darcy, 1999; Hausner, 1997; Qureshi, 1997). By contrast, archaeal genomes have been found to contain many putative transcriptional regulators that are more closely related to bacterial counterparts than to eukaryotic regulators.

Physical properties that so far seem unique to the Archaea include having archaeal membrane lipids composed of isoprenoid chains linked to sn-glycerol 1-phosphate head groups rather than sn-glycerol 3-phosphate, and no murein (the peptidoglycan of bacterial cell wall), among others. In addition, evidence for the presence of many “classical” biochemical pathways have not been found in either biochemical or genomic data. Rather, “modified” versions of the classic sugar degradation pathways within central carbohydrate metabolism (e.g., Embden-Meyerhof-Parnas [EMP], Entner-Doudoroff [ED] pathways) have been identified with novel biocatalysts that can catalyze related, but slightly different, reactions in parallel pathways.

Additionally, the pentose phosphate pathway (PPP) is only partially present, if at all, with great differences in pentose degradation from bacterial model organisms (Bräsen, 2014). These metabolic differences will be discussed in greater detail in upcoming sections of this chapter.

Although genome-based metabolic reconstruction in archaea is sometimes hampered by the absence of identifiable orthologs to bacterial or eukaryotic genes encoding enzymes for known steps in metabolic pathways, gaps in knowledge can and have been filled biochemically with discoveries of non-orthologous enzymes (e.g., the de novo synthesis of mevalonate and purines) (Barkley, 2004; Grochowski, 2006; Graupner, 2001), as well as the discovery of novel enzymes and precursors in archaea that circumvent absent enzymes in classical pathways (e.g., pentose formation and aromatic amino acid synthesis) (Grochowski, 2005; Porat, 2006).

1.3 Introduction To Halophilic Archaea

Haloarchaea specifically refers to the halophilic and halotolerant groups of archaea within the Euryarchaeota and Crenarchaeota phyla (Oren, 2002; Anderson, 2011). Halophilic and halotolerant microorganisms are subcategories of extremophiles found broadly distributed in hypersaline environments in waters and sediments of salt lakes, salterns, and saline soils. Halophiles generally require at least 1.5 M salt, and grow optimally at 2-4.5 M NaCl (Torregrosa, 2017). Eukaryotic and bacterial halophiles and halotolerant organisms have been identified, though most extreme halophiles (>20% salinity) are classified as archaea (Andrei, 2012). Several extreme halophilic species were identified before the archaeal “domain” of life was suggested (Leigh, 2011). One of the best known, *Halobacterium salinarum*, became of great interest to the scientific community for producing the light-driven ion pump bacteriorhodopsin (Oesterhelt, 1971).

To survive in extreme conditions such as hypersalinity, haloarchaea and other halophiles have adapted unique mechanisms. Examples of these adaptations include the ability to accumulate high intracellular molar concentrations of potassium chloride and other compatible,

neutral or zwitterionic solutes, in the cytoplasm to increase the internal osmolarity of the cell. Furthermore, halophiles have evolved proteins with a higher density of negative charges on their surfaces compared to their homologs in mesophilic organisms (Anderson, 2011; Becker, 2014). Due to the high acidity of the proteome, many halophilic proteins are known to denature at low salt concentration.

Becker et al. (2014) analyzed the genomes of 80 sequenced haloarchaea and identified 304 core protein families, thus representing the unique haloarchaeal signature. Included among the core haloarchaeal proteins were ribosomal proteins, RNA polymerase subunits, and known general transcription factors, as well as other transcription factors predicted to have regulatory functions.

1.4 The Haloarchaea As Model Organisms

Features that make haloarchaea particularly interesting in research, education, and industrial settings include, but are not limited to, their ability to produce molecules of commercial interest (i.e., antimicrobials, bioplastics, biofuel, extremozymes, extracellular polysaccharides, retinal proteins, coloured pigments, and compatible solutes), as well as their broad environmental phenotype, prolific distribution throughout the world, roles in biogeochemical cycles, and unique metabolic profiles (Desai, 2019; Atomi, 2012; Leigh, 2011; Torregrosa, 2017; DeLong, 2001; Haque, 2020; see Chapters 2 and 3).

Furthermore, haloarchaea have many characteristics that Leigh et al. (2011) identify for a suitable model organism: they are easy to grow, have a relatively short generation time, and are amenable to experimental manipulation, with a diverse genetic toolbox. Due to the relative accessibility of their “extreme” environmental conditions (e.g., adding salt to water), Shand (1999) suggested that haloarchaea may actually be among the easiest archaea to culture. The fact that they can be used for numerous educational activities lends credence to this argument (See Chapter 3; Furrow, 2020).

Many species of haloarchaea grow in aerobic conditions, and are capable of growing in both defined and complex mediums (DasSarma, 1995; Shand, 1999). Further, haloarchaea have a common optimal temperature range between 37° and 42° C, and are resistant to contamination in typical laboratory settings due to their preference to grow in conditions that are naturally growth-inhibiting to non-halophilic and non-halotolerant organisms. Doubling times vary across members of the haloarchaea: some haloarchaea are slow growing (with a doubling time of at least two days), while others have doubling times of 1.5-3 hours at optimal growing temperatures and media conditions (e.g., the doubling times of *Haloferax volcanii* DS70 and *H. salinarum* NRC-1 are 1.83 and 1.86 hours, respectively, in their optimal growing temperatures) (Robinson, 2005).

Lastly, building upon the vast studies conducted on bacteria and eukaryotes, there is also a growing toolbox for genetic exploration and manipulation in archaea and haloarchaea, specifically (Leigh, 2011). Transformation systems, such as inducible expression systems and systematic gene knockout techniques, exist for several haloarchaeal species, which also may involve vector plasmids conferring resistance to mevinolin or novobiocin, and selectable markers (Lam, 1989; Holmes, 1991; Peck, 2000; Born, 2021). In fact, halophiles were among the first archaea to successfully undergo genetic transformation, and be used in a gene knockout study using a counter selectable marker (Charlebois, 1987).

With the development and improvement of new and existing tools and datasets (e.g., annotated genome sequences) in archaea, researchers have been able to learn more about archaeal physiology and simultaneously, learn ways to improve the means by which to investigate them (Becker, 2014). Berquist (2006) summarizes major genetic tools used in haloarchaeal research and Leigh (2011) provides specific examples of genetic tools available in *H. salinarum* and *H. volcanii*, two organisms which have been common subjects of many haloarchaeal studies to date. For example, promoter activity can be studied via reporters beta-galactosidase, dihydrofolate reductase, and green fluorescent protein (GFP) (Danner, 1996;

Gregor, 2001; Born, 2019). Other biochemical laboratory techniques such as enzyme assays, protein purifications, -omics analyses, and reverse transcription-quantitative PCR (qRT-PCR), among others, have also allowed for improved metabolic pathway reconstruction and characterization in haloarchaea as well (Bräsen, 2014).

1.5 Introduction To Haloarchaeal Metabolism

Haloarchaea are known to be aerobic or facultatively anaerobic, and mostly utilize heterotrophic metabolism. Some organisms are metabolically “flexible,” able to grow on simple media (e.g. *H. volcanii*, *Natronomonas pharaonis*) consisting of very few substrates, or able to utilize a wide selection of substrates. Meanwhile, others species may be nutritionally less flexible, having complex nutritional demands that require the presence of multiple (e.g., *H. salinarum*) or few, but specific, sets of substrates (e.g., pyruvate and dihydroxyacetone for *Haloquadratum walsbyi*) (Bardavid, 2008).

Both sugar utilizing (e.g., *Haloferax mediterranei*, *Haloarcula marismortui*, *H. borinquense*, and *Halococcus saccharolyticus*) and non-sugar utilizing (e.g., *H. salinarum* and *H. volcanii*) haloarchaea have been identified. The non-sugar utilizing species instead thrive on amino acids and other products often found in hypersaline environments, such as glycerol and organic acids that serve as important sources of carbon and nitrogen for halophilic organisms. According to Andrei (2012), most halophilic archaea preferentially consume amino acids over carbohydrates for carbon and energy. In support of this, Becker et al. (2014) in their comparative analysis of 80 haloarchaeal genomes showed universal conservation of ABC transporters predicted for peptides, amino acids, and/or metals as predicted substrates. Furthermore, the distribution of catabolic pathways, and the protein sequence conservation of associated enzymes shown in a separate analysis of the genomes of ten haloarchaeal species, provides evidence that haloarchaeal ancestors were capable of degrading many amino acids and fewer carbohydrates (Anderson, 2011).

Haloarchaeal sugar utilization has been studied most extensively in *Halobacterium* and *Haloarcula* genera. With metabolic pathway reconstructions in hand (see above), carbon labeling experiments and comparative -omics analyses have allowed researchers to experimentally observe the existence and utilization of specific metabolic pathways. These studies, along with others, showed that like other archaea, the haloarchaea generally use a modified ED and/or EM pathway to metabolize carbohydrates (Figure 1.1) (Falb, 2008). Johnsen (2001), for example, shows that *H. saccharolyticus* degrades glucose and fructose through a modified ED pathway, whereas fructose was degraded by both modified EM and ED pathways in a 96% to 4% split, respectively.

In another study, Williams (2019) combined genomics, proteomics, and metagenomics analyses to investigate the specific pathways used in sucrose metabolism in haloarchaea. They found that a number of the haloarchaeal genomes studied were missing enzymes (ketohexokinase and 1-phosphofructokinase) in the modified EM pathway, which suggests that some haloarchaea use novel hypothetical alternative pathways (e.g., semi-phosphorylative ED pathways involving GFO domain proteins, or XI/TM domain proteins) to metabolize sucrose and fructose (Williams, 2019).

Genomic analyses also provide evidence that haloarchaeal genomes encode a complete set of TCA cycle genes (Falb, 2008; Anderson, 2011). In addition, some haloarchaeal species are also able to use the glyoxylate and/or methylaspartate cycle, and respiratory electron transport chains using a variety of terminal electron acceptors (Falb, 2008; Khomyakova, 2011; Kornberg, 1957). The ability to use both the glyoxylate and methylaspartate cycles for acetyl-CoA assimilation was likely acquired via lateral gene transfer of key genes from the bacterial branch. Khomyakova (2011) suggests that the methylaspartate cycle could be a third mechanism (in addition to the glyoxylate cycle and the ethylmalonyl-CoA pathway) to balance intracellular glutamate levels. This is particularly relevant to halophiles as glutamate is a

common osmolyte, necessary for survival in hypersaline environments, as well as the universal donor of alpha-amino groups of amino acids.

Finally, unlike the rest of the archaea, halophiles are hypothesized to use a modified oxidative pentose phosphate pathway (Falb, 2008). This pathway is important for maintaining carbon homeostasis, generating precursor molecules for nucleotide and amino acid biosynthesis, generating NADPH for anabolic pathways, and mitigating oxidative stress (Stincone, 2015).

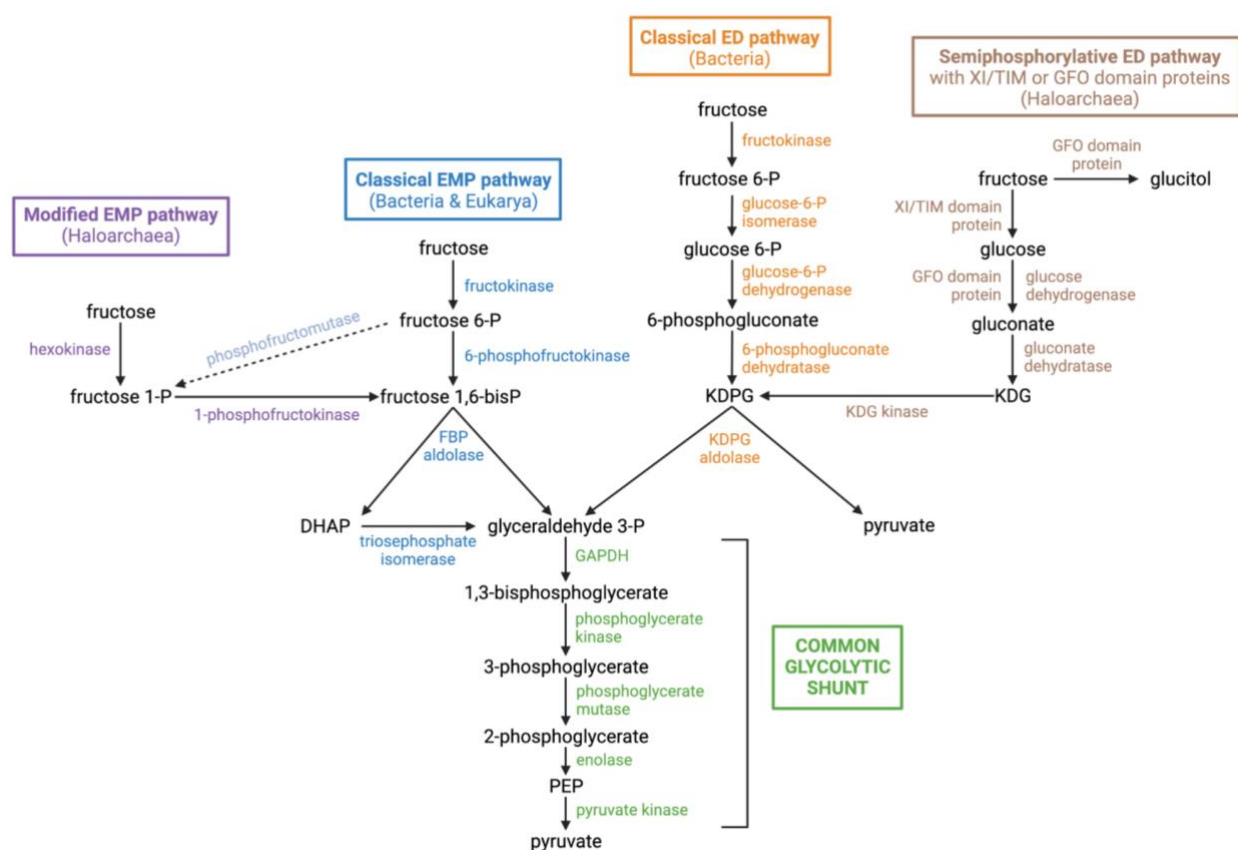


Figure 1.1. Modified Embden-Meyerhof-Parnas (EMP) and Semiphosphorylative Entner-Doudoroff (ED) pathways used by haloarchaea alongside classical EMP and ED pathways used by bacteria and eukarya. For simplicity, the conversion of gluconate to glucose is represented by a single arrow. Acronyms: DHAP – dihydroxyacetone phosphate; FBP – Fructose 1,6-bisphosphate; GAPDH – glyceraldehyde-3-phosphate dehydrogenase; GFO – glucose-fructose oxidoreductase; KDG – 3-deoxy-2-oxo-d-gluconate; KDPG – 2-dehydro-3-deoxy-phosphogluconate; PEP – phosphoenolpyruvate; XI/TIM – xylose isomerase-like, triosephosphate isomerase barrel (figure adapted from [Williams, 2019]).

1.6 CCR / Metabolic Regulation

Jacob and Monod's (1961) landmark studies in *E. coli* were among the first documented observations of diauxic growth behavior, also referred to as diauxie. Their work provided an example of preferential sugar utilization. In one now famous experiment, they showed that *E. coli* preferred glucose over lactose when both carbon sources were made simultaneously available to the organism. This experiment led to the discovery of a mechanistic framework for the expression and co-regulation of the *lac* operon by glucose and lactose.

The *lac* operon is structurally composed of three inducible genes, *lacXYZ*, and the *lacI* gene, encoding the transcriptional regulator known as the lac repressor. In this system, the product of *lacI* represses the expression of *lacXYZ* in the absence of lactose through the binding of the lac repressor to operator sites both near the promoter and within the operon. This binding induces the formation of a DNA loop that sequesters the promoter leading to the repression of the operon. In the presence of lactose, allosteric binding of the sugar to the lac repressor leads to the release of the protein from its operator sites and depression of the operon. In this way, the lactose acts as an "on"/"off" switch for the operon. However, the *lac* operon genes are only expressed strongly in the absence of glucose, whether lactose is present or not.

The activation of the *lac* operon in the absence of glucose is mediated by cAMP levels through the cAMP receptor protein, or CRP. When both glucose and lactose are available, the *lac* operon genes for metabolizing lactose are induced only after a lag phase following the depletion of glucose (Jacob, 1961; Oehler, 1990). This detailed mechanism explains, in part, how the decision to prioritize one carbon source over another can be made, in the specific case of glucose-lactose diauxie.

While the phenomenon of preferential carbon utilization was originally characterized in *E. coli*, it has since been seen in other organisms, including other bacteria, eukaryotes, and archaea (Amoozegar, 2017; Altekari, 1992; Bräsen, 2004; Bini, 2001). Since glucose was often found to be the preferred substrate for many organisms, this phenomenon was often called the

“glucose effect.” However, Jacob and Monod, among others, showed that this type of decision making also happens for other carbon sources and pairs of carbohydrates as well, and thus this process is now more commonly referred to as carbon catabolite repression (CCR) (Brückner, 2002; Blaiseau, 2021) (Figure 1.2).

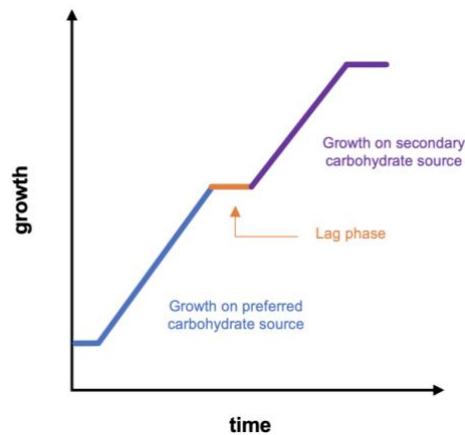


Figure 1.2. Model of diauxic growth.

1.6.1 Overview Of CCR In Bacteria And Eukaryotes

CCR mechanisms are observed in most free-living heterotrophic bacteria and are thought to regulate as many as 5-10% of their genes (Görke, 2008). Studies suggest that CCR-mediated gene regulation contributes to an organism's fitness by allowing it to make the most efficient use of available substrates.

1.6.1.1 The Textbook Mechanisms: *E. coli* And *B. subtilis*

In bacteria, the most research has been carried out on glucose-lactose diauxie, using model organisms *E. coli* and *Bacillus subtilis*. In these systems, some CCR-associated regulators were shown to affect a broad range of genes, while others targeted a single gene or operon only. Regulation in both *E. coli* and *B. subtilis* entails a phosphorylation cascade

involving the phosphoenolpyruvate–carbohydrate phosphotransferase system (PTS), specifically the IIA component (EIIA) (Deutscher, 2006).

Other key proteins in *E. coli* CCR regulation include cAMP, CRP, and adenylate cyclase. Adenylate cyclase activity is activated by the phosphorylated form of EIIA^{Glc} in the absence of glucose, to synthesize cAMP. Thereafter, cAMP is able to form a complex with CRP, and activate the promoters of numerous catabolic genes and operons, including the lac operon, as well as a glucose transporter gene (ptsG) (Kimata, 1997). When EIIA^{Glc} is dephosphorylated, such as during rapid growth on readily usable carbon sources like glucose, cAMP is not synthesized, and the transcriptional activation of secondary catabolic genes is prevented (Görke, 2008).

Regulation can also occur through control of sugar transporters and transcription factors among other signaling molecules (Titgemeyer, 1994; Görke, 2008). EIIA^{Glc} also is involved in regulating the transport of secondary carbon sources, such as lactose, maltose, melibiose, raffinose, and galactose. Lactose permease, LacY, responsible for transporting lactose into the cell, is inactive when glucose is available. When lactose or other secondary sugars are present alongside glucose, non-phosphorylated EIIA^{Glc} can bind and inactivate the lactose (and other secondary substrate) permease(s), and other metabolic enzymes and transporters (Inada, 1996; Deutscher, 2008). This mechanism of regulation is termed “inducer exclusion” or “inducer prevention” due to the prevention of intracellular formation of the inducer of the secondary catabolic system (Görke, 2008; Bini, 2001).

In *B. subtilis*, the key proteins involved in regulation include pleiotropic transcription factor catabolite control protein A (CcpA), PTS component histidine protein (HPr), bifunctional HPr kinase/phosphorylase (HPrK), and glycolytic intermediates fructose-1,6-bisphosphate (FBP) and glucose-6-phosphate (Glu-6-P). During conditions of high intracellular FBP and ATP levels – which indicates the availability of preferred carbon sources – HPr is phosphorylated by HPrK at its regulatory sites, Ser46 and His15. Serving as one of the effectors for CcpA, HPr-

(Ser-P) complexes with and enables CcpA's binding to operator sites, which results in the repression of transcription of catabolic genes. FBP and Glu-6-P enhance the binding interaction. On the other hand, when concentrations of ATP and FBP are low, with levels of high inorganic phosphate instead (reflecting nutrient limitation), HPrK dephosphorylates HPr(Ser-P).

The phosphorylation of the His15 site is significant for regulation as well. One of the actions of HPr(His-P) is to phosphorylate glycerol kinase GlpK and transcriptional regulators of PRDs (phosphoenolpyruvate-carbohydrate phosphotransferase system-regulatory domains) when glucose is not present, which results in the activation of catabolic operons (e.g., the *bglPH* operon used for aryl-beta-glucoside metabolism, which is under the control of the LicT antiterminator). When glucose is available, PRD regulator activation is prevented, another example of induction exclusion (Görke, 2008).

1.6.1.2 CCR When Glucose Is Not the Primary Carbohydrate And In Other Bacteria

A host of studies show that glucose is not always the preferred or primary carbohydrate for all organisms. In *Streptococcus thermophilus* (Van den Bogaard, 2000), *Bifidobacterium longum* (Parche, 2006), and *Pseudomonas aeruginosa* (Collier, 1996), glucose utilization genes are expressed only after the primary carbohydrate source is consumed (a process that some have referred to as reverse CCR). In *Rhodospirillaceae*, there are also examples of reverse CCR, with TCA cycle intermediates repressing enzymes involved in sugar metabolism. Rather than sugars, succinate and malate are instead preferred (Inui, 1996).

Research in other groups of bacteria suggest that each have evolved their own mechanisms of CCR (e.g., Firmicutes use "transcription repression," Actinobacteria use an unknown CCR mechanism triggered by glucose kinase, and pseudomonads use translation repression by an RNA-binding protein) (Moreno, 2001; van Wezel, 2007; Collier, 1996; Görke,

2008). Examples of bacteria in which CCR has not been observed include *Chlamydia trachomatis* and *Mycoplasma pneumoniae*, two pathogenic organisms that are highly adapted to a limited number of nutrient-rich host environments and have small genomes (Görke, 2008).

1.6.1.3 CCR In Eukaryotes

In eukaryotes, research has focused primarily on *S. cerevisiae*, also known as budding yeast. Notably, *S. cerevisiae* is unique in that it has evolved to preferentially ferment glucose and metabolize it to pyruvate via the EMP pathway. When presented with non-fermentable carbon sources, the carbon will enter the TCA cycle and the organism uses respiration to generate ATP while gluconeogenesis is simultaneously activated (Bini, 2001). CCR in this organism refers to the phenomenon of glucose repressing pathways related to respiration, thus resulting in ethanol formation. Repression at the level of transcription involves two types of regulation, one cAMP-independent mechanism and the other cAMP-dependent, and involves a phosphorylation cascade that ultimately inhibits a transcription repressor (Ronne, 1995; Trumbly, 1992). A study in filamentous fungi *Aspergillus nidulans* showed markedly different mechanisms of CCR from that of *S. cerevisiae*, involving deubiquitination via the creB protein, found to have high sequence similarity with analogous proteins in higher eukaryotes (Lockington, 2001).

In humans, cAMP is involved with mediating signal transduction pathways that target a cAMP response element binding protein (CREB), which in turn mediates gene expression at conserved enhancer elements (CRE) (Roesler, 2000; Park, 1999). A CCR-like system has been detected in plants as well, however, cAMP has not been evidenced to be involved in CCR regulation (Bini, 2001). A transduction cascade similar to that which occurs in yeast has been observed, initiated instead by hexokinase, glucokinase, or fructokinase generating a signal that leads to a transduction cascade and action at the transcriptional level (Farrar, 2000).

1.6.1.4 Summary Of CCR In Bacteria And Eukaryotes

Overall, the investigations of CCR in bacteria and eukaryotes, particularly in *E. coli* and *B. subtilis*, show that the underlying mechanisms are both complex and diverse, involving numerous interacting catabolic proteins and glycolytic intermediates. While *E. coli* and *B. subtilis* both utilize PTS for regulation, the mechanisms that drive CCR differ greatly between the two organisms. In *E. coli*, CCR is accomplished through the prevention of transcriptional activation of secondary catabolic genes. On the other hand, in *B. subtilis*, the primary driver of regulation is the transcriptional repression of secondary catabolic genes. Studies conducted on other bacteria and eukaryotes reveal heterogeneity in their underlying CCR mechanisms (or lack thereof), with some CCR systems showing a “reverse glucose effect,” or otherwise not utilizing glucose as the preferred substrate.

1.6.2 Introduction To CCR In Archaea

Since the biochemical pathways for carbohydrate metabolism and regulatory mechanisms used in archaea (e.g., modified EMP and ED pathways) are often analogous to those used in bacteria and eukaryotes, one might reasonably expect CCR to also exist in archaea. However, as previously described, the mechanisms for CCR and the key players involved have been shown to vary between organisms. In archaea, CCR-implicated transport systems like the bacterial PTS are not known to exist and thus cannot be a part of CCR in archaea. However, the targets of regulation, such as membrane protein transport systems and other key enzymes that are important in carbohydrate utilization regulation, may be involved in potential alternative archaeal CCR mechanisms (Bini, 2001).

Unfortunately, mechanisms of CCR in the archaea have not yet been elucidated. This section provides an overview of studies in archaea that have investigated carbohydrate catabolism and/or differential carbon source utilization that provide some relevant insight into

the discussion of CCR. These studies provide evidence of differential expression of key types of target genes (carbohydrate transporters, enzymes required for carbohydrate utilization) and observations suggestive of underlying CCR mechanisms. This review starts with *Sulfolobus solfataricus*, where numerous studies have provided strong clues of CCR, and conclude with a brief summary of findings from work on other archaea, including haloarchaea.

1.6.2.1 CCR In *Sulfolobus solfataricus*

S. solfataricus is a thermophilic member of the Crenarchaea found in acidic hot springs. This organism is able to use numerous carbohydrates as a sole carbon source, including glucose, xylose, sucrose, maltose, among others (De Rosa, 1975, 1984). Snijders (2000) used a two-dimensional electrophoresis, stable isotope labeling, and DNA microarray to reconstruct central carbon metabolism in *S. solfataricus*. Through the monitoring of changes at the transcriptional level in response to D-glucose versus yeast extract and tryptone, they found 184 genes that were differentially expressed, with the most significant differences associated with amino acid metabolism and transport (Snijders, 2000). This study sheds insight into the regulatory influences that glucose and other substrates play in determining substrate-specific transcriptional response of metabolic pathway genes, providing initial hints into genes or gene groups that may be targets of regulatory mechanisms like CCR. Other studies provide evidence of a glucose effect and CCR-like system in *S. solfataricus* that controls the expression of three specific glycosyl hydrolases: alpha-glucosidase or maltase (malA), alpha-amylase (amyA), and beta-glycosidase (lacS) (Haseltine, 1996, 1998, 1999a, 1999b; Rolfsmeier, 1995, 1998).

Differential expression of glycosyl hydrolases under different carbon sources:

Alpha-glucosidase is essential for hydrolysis of alpha-1,4-linked polysaccharides. After growing *S. solfataricus* using different carbon sources, analysis of crude cell extracts showed markedly different enzyme levels of alpha-glucosidase (Rolfsmeier, 1995). Levels of beta-

glycosidase, which hydrolyzes beta-linked oligosaccharides, were even more strongly affected by growth on different carbon sources than alpha-glucosidase (Haseltine, 1998). In a separate study, alpha-amylase was shown to hydrolyze starch, dextrin, and alpha-cyclodextrin (Haseltine, 1996) at different capacities when grown under different carbon sources. Gene expression of the three glycosyl hydrolases was highest when grown on sucrose, with minimal expression when grown on sucrose supplemented with yeast extract. Haseltine (1996) showed that this effect is driven by aspartate and asparagine activating the alpha-amylase. Meanwhile, glutamate was shown to have the most de-repressing effect on alpha-amylase.

Preliminary evidence of CCR via glycosyl hydrolases in S. solfataricus:

Additions of exogenous glucose in the media repressed the production of alpha-amylase, supporting the hypothesized existence of a glucose effect in this organism (Haseltine, 1996). Induction of alpha-amylase secretion occurred within one cell division following the relief of glucose repression. In their follow up study, Haseltine et al. (1999) showed that the transcription of alpha-glucosidase and beta-glycosidase was also repressed by exogenous additions of glucose. Their kinetic experiments revealed that full induction for beta-glycosidase occurred in 6 hours (less than one generation), and full repression occurred at a rate slower than cell division, requiring 200 hours of elapsed time, or nearly 33 generations (Haseltine, 1999). These results support a model of increased transcript synthesis (rather than degradation) under the CCR response.

Experiments utilizing enzyme assays, Southern blots, polyclonal antibodies, and RNA strand-specific probes (riboprobes) provide evidence that the three aforementioned glycosyl hydrolases are coordinately expressed (Haseltine, 1996,1999a, 1999b; Roflsmeier, 1998). Given that the genes are separated by a minimum of 25 Mb and are thus physically unlinked, the regulation of these catabolite repression genes are hypothesized to employ a trans-acting diffusible factor. In one investigation of this hypothesis, Haseltine et al. (1999b) conducted

experiments on spontaneous glycosyl hydrolase mutants with reduced expression of beta-glycosidase, and subjected them to reversion analyses. One group of isolated and purified mutants displayed a pleiotropic defect in glycosyl hydrolase expression, with revertants' activity restored either partially or completely to the level of the wildtype. These results are consistent with previous experiments conducted on *E. coli* catabolite repression (CR) mutants; mutants exhibited a pleiotropic defect in CR gene expression with reduced or non-inducible expression of the affected genes (adenylate cyclase and catabolite repression protein) (Haseltine, 1999b). The observations made by Haseltine et al. support the hypothesis of coordinated glycosyl hydrolase gene expression and the existence of a trans-acting factor in the potential CCR regulation of *S. solfataricus* (Haseltine, 1999b).

Regulation of carbohydrate uptake in *S. solfataricus*:

Lubelska et al. (2006) focused their investigations instead on sugar uptake in *S. solfataricus*, revealing the involvement of high affinity binding of ABC transporter proteins in regulation. They studied two sugar transport operons encoding the glucose and arabinose transporters, by growing cells on various carbon sources. The arabinose transport gene cluster was shown to be heavily regulated, with expression dependent on the presence of arabinose. Further, the inclusion of glucose and alternative sources of carbon, such as amino acids, in the media was shown to repress expression of arabinose transport genes (Lubelska, 2006). On the other hand, Elferink et al. (2001) showed that other binding proteins, such as those for glucose and trehalose, are constitutively expressed under a large variety of growth conditions.

While these observations suggest the presence of CCR, the specifics of the underlying mechanisms are yet to be discovered. The presence of cAMP, which is involved in CCR in other organisms) has been reported in *Sulfolobus solfataricus*, and two other archaea, *Methanobacterium thermoautotrophicum* and *H. volcanii* (Leichtling, 1986). However, the enzyme responsible for forming cAMP, adenylate cyclase, does not have obvious archaeal

orthologues (Bini, 2001). A unique adenyl cyclase orthologue may be present, the study of which could clarify the role of cAMP in *S. solfataricus* and other organisms.

1.6.2.2 CCR In Other Archaea

CCR-associated behavior has also been evidenced in other archaeal systems (e.g., *Thermococcus*, *Pyrococcus*) as well. Studies on carbohydrate catabolism in several archaeal organisms reveal hallmarks of CCR, such as carbon source preference and differential gene expression of catabolic proteins under different growth conditions (Xavier, 2000; van der Oost, 1998). However, as is the case with *S. solfataricus*, no specific CCR mechanisms have yet been elucidated. Thermococci are primarily marine organisms that ferment glucose via a modified EMP pathway. One non-marine species, *Thermococcus zilligi*, uses a second glycolytic pathway (likely a phosphoketolase pathway [PPK]) in addition to EMP, thus splitting carbon flux at the level of glucose-6-phosphate (Xavier, 2000). Xavier (2000) showed that the two branches are inversely regulated by growth conditions, where the EMP pathway is activated by glucose and repressed by tryptone, while the PPK pathway is activated by tryptone and repressed by glucose. *Pyrococcus furiosus* also appears to show carbon source preference, showing higher activity of two modified EMP kinases and glyceraldehyde-3-phosphate ferredoxin oxidoreductase (GAPOR) when grown on maltose instead of pyruvate (van der Oost, 1998). These metabolic patterns seen in *T. zilligi* and *P. furiosus* are indicative of hierarchical usage of carbon sources, and the existence of an underlying CCR-like system.

1.6.2.3 Evidence Of CCR In Haloarchaea

Several studies offer insight into carbohydrate metabolism and regulation in haloarchaea – however, there is a lack of continuity between the studies or a coherent narrative. I provide a brief summary of the studies and key results in this section.

- ***H. saccharolyticus***: Johnsen (2001) grew *H. saccharolyticus* on different media conditions and used carbon labeling, comparative enzyme activity measurements, and growth experiments to explore which pathways were preferentially utilized in the organism and signals of CCR regulation, as previously described. *H. saccharolyticus* was also grown on media with both glucose and fructose, and diauxic growth was shown where glucose was completely consumed first, then fructose was degraded following a lag phase. In suspension experiments, glucose-grown cells and sucrose-grown cells degraded their respective sugars preferentially, but upon longer incubation times, they also metabolized the alternate hexose (Johnsen, 2001).
- ***H. vallismortis***: Altekar (1992) investigated sucrose and mannitol catabolism in *Haloarcula vallismortis* and was unable to detect involvement of a PEP-sucrose-PTS system. They also saw that sucrase and NAD-specific mannitol dehydrogenase was induced only when sucrose and mannitol were supplied. Furthermore, they conducted growth assays which revealed that *H. vallismortis* growth on sucrose was biphasic in nature, indicating that either the glucose or fructose that comprises the disaccharide is used in a preferential manner.
- ***H. volcanii***: Altekar (1992) also studied carbohydrate catabolism in *H. volcanii*, and characterized a glucose-specific ABC transporter to be active and inducible during anaerobic growth. In a later study, Martin (2018) identified a DeoR-type transcriptional factor GlpR, as a negative regulator of fructose degradation during growth on other carbon sources like glucose and glycerol. Sherwood (2020) additionally found that the presence of glycerol can repress glucose metabolism.
- ***H. halobium***: Bonete (1996) studied NAD-specific glutamate dehydrogenase in *Halobacterium halobium* and found that it was strongly inhibited by TCA intermediates (e.g., fumarate, oxaloacetate, succinate, and malate), analogues such as NADP⁺, D-glutamate, and glutarate, and dicarboxylic compounds like adipate. Conversely, all

amino acids tested were activators of the enzyme, aside from the D-isomer of L-glutamate, which was shown to act as an inhibitor (Bonete, 1996).

- ***H. marismortui***: Bräsen and Schonheit (2004) investigated metabolism in *H. marismortui*, and showed that acetate is formed while the organism grows on glucose, and is used thereafter as a growth substrate. When the organism is grown with glucose and acetate together, glucose is preferentially utilized and a diauxic growth pattern is manifested. Enzyme upregulation and downregulation is inverted when grown in glucose versus acetate.

The diauxic growth phenotypes and carbon source preferences that have been determined through a host of biochemical techniques provide strong evidence of CCR in haloarchaea. However, there is currently no single best system with which to investigate sugar metabolism and related regulatory mechanisms, as we are currently still examining how different organisms utilize specific carbohydrates and combinations of carbohydrates.

Following greater characterization of carbohydrate usage and diauxie, future studies can further identify and explain regulatory mechanisms like the targeted repression or activation of catabolic genes by transcription factors (e.g., GlpR repressing fructose metabolism genes in *H. volcanii*) (Martin, 2018; Todor, 2013). With the increasing accessibility of sequencing technologies, it is becoming more and more feasible to investigate activation or repression of specific genes in haloarchaeal pathways under different nutritional conditions on a larger-scale (i.e., transcriptome-wide). Further validation and exploration of the underlying mechanisms will thereafter require experimentation, like localization or knock-in/-out studies, with the regulation-implicated genes.

1.7 References

- Adam, P. S., Borrel, G., Brochier-Armanet, C., & Gribaldo, S. (2017). The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *The ISME journal*, 11(11), 2407-2425.
- Altekar, W., & Rangaswamy, V. (1992). Degradation of endogenous fructose during catabolism of sucrose and mannitol in halophilic archaeobacteria. *Archives of microbiology*, 158(5), 356-363.
- Amoozegar, M. A., Siroosi, M., Atashgahi, S., Smidt, H., & Ventosa, A. (2017). Systematics of haloarchaea and biotechnological potential of their hydrolytic enzymes. *Microbiology*, 163(5), 623-645.
- Anderson, I., Scheuner, C., Göker, M., Mavromatis, K., Hooper, S. D., Porat, I., ... & Kyrpides, N. (2011). Novel insights into the diversity of catabolic metabolism from ten haloarchaeal genomes. *PLoS One*, 6(5), e20237.
- Andrei, A. Ş., Banciu, H. L., & Oren, A. (2012). Living with salt: metabolic and phylogenetic diversity of archaea inhabiting saline ecosystems. *FEMS microbiology letters*, 330(1), 1-9.
- Atomi, H., Imanaka, T., & Fukui, T. (2012). Overview of the genetic tools in the Archaea. *Frontiers in microbiology*, 3, 337.
- Baker, B. J., De Anda, V., Seitz, K. W., Dombrowski, N., Santoro, A. E., & Lloyd, K. G. (2020). Diversity, ecology and evolution of Archaea. *Nature microbiology*, 5(7), 887-900.
- Baliga, N. S., Goo, Y. A., Ng, W. V., Hood, L., Daniels, C. J., & DasSarma, S. (2000). Is gene expression in *Halobacterium* NRC-1 regulated by multiple TBP and TFB transcription factors? *MicroCorrespondence. Molecular microbiology*, 36(5), 1184-1185.
- Bang, C., & Schmitz, R. A. (2018). Archaea: forgotten players in the microbiome. *Emerging Topics in Life Sciences*, 2(4), 459-468.
- Bardavid, R. E., & Oren, A. (2008). Dihydroxyacetone metabolism in *Salinibacter ruber* and in *Haloquadratum walsbyi*. *Extremophiles*, 12(1), 125-131.
- Barkley, S. J., Cornish, R. M., & Poulter, C. D. (2004). Identification of an archaeal type II isopentenyl diphosphate isomerase in *Methanothermobacter thermautotrophicus*. *Journal of bacteriology*, 186(6), 1811-1817.
- Becker, E. A., Seitzer, P. M., Tritt, A., Larsen, D., Krusor, M., Yao, A. I., ... & Facciotti, M. T. (2014). Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. *PLoS genetics*, 10(11), e1004784.
- Bell, S. D., & Jackson, S. P. (2001). Mechanism and regulation of transcription in archaea. *Current opinion in microbiology*, 4(2), 208-213.
- Bertrand, K., Korn, L., Lee, F., Platt, T., Squires, C. L., Squires, C., & Yanofsky, C. (1975). New features of the regulation of the tryptophan operon. *Science*, 189(4196), 22-26.
- Berquist, B. R., Müller, J. A., & DasSarma, S. (2006). 27 genetic systems for halophilic archaea. *Methods in microbiology*, 35, 649-680.
- Bini, E., & Blum, P. (2001). Archaeal catabolite repression: a gene regulatory paradigm. *Advances in applied microbiology*, 50, 339-366.
- Blaiseau, P. L., & Holmes, A. M. (2021). Diauxic inhibition: Jacques Monod's Ignored Work. *Journal of the History of Biology*, 1-21.
- Bonete, M. J., Perez-Pomares, F., Ferrer, J., & Camacho, M. L. (1996). NAD-glutamate dehydrogenase from *Halobacterium halobium*: inhibition and activation by TCA intermediates and amino acids. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1289(1), 14-24.
- Born, J., & Pfeifer, F. (2019). Improved GFP variants to study gene expression in haloarchaea. *Frontiers in microbiology*, 10, 1200.

- Born, J., Weitzel, K., Suess, B., & Pfeifer, F. (2021). A Synthetic Riboswitch to Regulate Haloarchaeal Gene Expression. *Frontiers in Microbiology*, 12, 1566.
- Bräsen, C., & Schönheit, P. (2004). Regulation of acetate and acetyl-CoA converting enzymes during growth on acetate and/or glucose in the halophilic archaeon *Haloarcu-
la marismortui*. *FEMS microbiology letters*, 241(1), 21-26.
- Bräsen, C., Esser, D., Rauch, B., & Siebers, B. (2014). Carbohydrate metabolism in Archaea: current insights into unusual enzymes and pathways and their regulation. *Microbiology and Molecular Biology Reviews*, 78(1), 89-175.
- Brückner, R., & Titgemeyer, F. (2002). Carbon catabolite repression in bacteria: choice of the carbon source and autoregulatory limitation of sugar utilization. *FEMS microbiology letters*, 209(2), 141-148.
- Brunner, N. A., Brinkmann, H., Siebers, B., & Hensel, R. (1998). NAD⁺-dependent glyceraldehyde-3-phosphate dehydrogenase from *Thermoproteus tenax*: the first identified archaeal member of the aldehyde dehydrogenase superfamily is a glycolytic enzyme with unusual regulatory properties. *Journal of Biological Chemistry*, 273(11), 6149-6156.
- Charlebois, R. L., Lam, W. L., Cline, S. W., & Doolittle, W. F. (1987). Characterization of pHV2 from *Halobacterium volcanii* and its use in demonstrating transformation of an archaeobacterium. *Proceedings of the National Academy of Sciences*, 84(23), 8530-8534.
- Collier, D. N., Hager, P. W., & Phibbs Jr, P. V. (1996). Catabolite repression control in the *Pseudomonads*. *Research in microbiology*, 147(6-7), 551-561.
- De Rosa, M., Gambacorta, A., & Bu'Lock, J. D. (1975). Extremely thermophilic acidophilic bacteria convergent with *Sulfolobus acidocaldarius*. *Microbiology*, 86(1), 156-164.
- De Rosa, M., Gambacorta, A., Nicolaus, B., Giardina, P., Poerio, E., & Buonocore, V. (1984). Glucose metabolism in the extreme thermoacidophilic archaeobacterium *Sulfolobus solfataricus*. *Biochemical Journal*, 224(2), 407-414.
- Danner, S., & Soppa, J. (1996). Characterization of the distal promoter element of halobacteria in vivo using saturation mutagenesis and selection. *Molecular microbiology*, 19(6), 1265-1276.
- Darcy, T. J., Hausner, W., Awery, D. E., Edwards, A. M., Thomm, M., & Reeve, J. N. (1999). *Methanobacterium thermoautotrophicum* RNA polymerase and transcription in vitro. *Journal of bacteriology*, 181(14), 4424-4429.
- DasSarma, S., & Robb, F. T. (Eds.). (1995). *Archaea: A laboratory manual*. Halophiles. Cold Spring Harbor Laboratory Press.
- DeLong, E. F., & Pace, N. R. (2001). Environmental diversity of bacteria and archaea. *Systematic biology*, 50(4), 470-478.
- Desai, C., Patel, P., Markande, A. R., Kamala, K., & Sivaperumal, P. (2020). Exploration of haloarchaea for their potential applications in food industry. *International Journal of Environmental Science and Technology*, 17, 4455-4464.
- Deutscher, J., Francke, C., & Postma, P. W. (2006). How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. *Microbiology and molecular biology Reviews*, 70(4), 939-1031.
- Deutscher, J. (2008). The mechanisms of carbon catabolite repression in bacteria. *Current opinion in microbiology*, 11(2), 87-93.
- Doolittle, W. F. (2020). Evolution: two domains of life or three?. *Current Biology*, 30(4), R177-R179.
- Elferink, M. G., Albers, S. V., Konings, W. N., & Driessen, A. J. (2001). Sugar transport in *Sulfolobus solfataricus* is mediated by two families of binding protein-dependent ABC transporters. *Molecular microbiology*, 39(6), 1494-1503.

- Facciotti, M. T., Reiss, D. J., Pan, M., Kaur, A., Vuthoori, M., Bonneau, R., ... & Baliga, N. S. (2007). General transcription factor specified global gene regulation in archaea. *Proceedings of the National Academy of Sciences*, 104(11), 4630-4635.
- Falb, M., Müller, K., Königsmaier, L., Oberwinkler, T., Horn, P., von Gronau, S., ... & Oesterhelt, D. (2008). Metabolism of halophilic archaea. *Extremophiles*, 12(2), 177-196.
- Farrar, J., Pollock, C., & Gallagher, J. (2000). Sucrose and the integration of metabolism in vascular plants. *Plant Science*, 154(1), 1-11.
- Furrow, R. E., Kim, H. G., Abdelrazek, S. M., Dahlhausen, K., Yao, A. I., Eisen, J. A., ... & Facciotti, M. T. (2020). Combining Microbial Culturing With Mathematical Modeling in an Introductory Course-Based Undergraduate Research Experience. *Frontiers in microbiology*, 11.
- Geiduschek, E. P., & Ouhammouch, M. (2005). Archaeal transcription and its regulators. *Molecular microbiology*, 56(6), 1397-1407.
- Görke, B., & Stülke, J. (2008). Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nature Reviews Microbiology*, 6(8), 613-624.
- Graupner, M., & White, R. H. (2001). *Methanococcus jannaschii* Generates L-Proline by Cyclization of L-Ornithine. *Journal of bacteriology*, 183(17), 5203-5205.
- Gregor, D., & Pfeifer, F. (2001). Use of a halobacterial bgaH reporter gene to analyse the regulation of gene expression in halophilic archaea. *Microbiology*, 147(7), 1745-1754.
- Grochowski, L. L., Xu, H., & White, R. H. (2005). Ribose-5-phosphate biosynthesis in *Methanocaldococcus jannaschii* occurs in the absence of a pentose-phosphate pathway. *Journal of Bacteriology*, 187(21), 7382-7389.
- Grochowski, L. L., Xu, H., & White, R. H. (2006). *Methanocaldococcus jannaschii* uses a modified mevalonate pathway for biosynthesis of isopentenyl diphosphate. *Journal of bacteriology*, 188(9), 3192-3198.
- Haque, R. U., Paradisi, F., & Allers, T. (2020). *Haloferax volcanii* for biotechnology applications: challenges, current state and perspectives. *Applied microbiology and biotechnology*, 104(4), 1371-1382.
- Haseltine, C., Rolsmeier, M., & Blum, P. (1996). The glucose effect and regulation of alpha-amylase synthesis in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Journal of bacteriology*, 178(4), 945-950.
- Haseltine, C., Montalvo-Rodriguez, R., Bini, E., Carl, A., & Blum, P. (1999a). Coordinate transcriptional control in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Journal of bacteriology*, 181(13), 3920-3927.
- Haseltine, C., Montalvo-Rodriguez, R., Carl, A., Bini, E., & Blum, P. (1999b). Extragenic pleiotropic mutations that repress glycosyl hydrolase expression in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Genetics*, 152(4), 1353-1361.
- Hausner, W., Wettach, J., Hethke, C., & Thomm, M. (1996). Two transcription factors related with the eucaryal transcription factors TATA-binding protein and transcription factor IIB direct promoter recognition by an archaeal RNA polymerase. *Journal of Biological Chemistry*, 271(47), 30144-30148.
- Holmes, M. L., & Dyal-Smith, M. L. (1991). Mutations in DNA gyrase result in novobiocin resistance in halophilic archaeobacteria. *Journal of bacteriology*, 173(2), 642-648.
- Inada, T., Kimata, K., & Aiba, H. (1996). Mechanism responsible for glucose-lactose diauxie in *Escherichia coli*: challenge to the cAMP model. *Genes to Cells*, 1(3), 293-301.
- Inui, M., Vertès, A. A., & Yukawa, H. (1996). Reverse catabolite repression and the regulation of CO₂ fixation in *Rhodobacter* and related bacteria. *Research in microbiology*, 147(6-7), 562-566.
- Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3), 318-356.

- Johnsen, U., Selig, M., Xavier, K. B., Santos, H., & Schönheit, P. (2001). Different glycolytic pathways for glucose and fructose in the halophilic archaeon *Halococcus saccharolyticus*. *Archives of microbiology*, 175(1), 52-61.
- Khomyakova, M., Bükmez, Ö., Thomas, L. K., Erb, T. J., & Berg, I. A. (2011). A methylaspartate cycle in haloarchaea. *Science*, 331(6015), 334-337.
- Kimata, K., Takahashi, H., Inada, T., Postma, P., & Aiba, H. (1997). cAMP receptor protein–cAMP plays a crucial role in glucose–lactose diauxie by activating the major glucose transporter gene in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 94(24), 12914-12919.
- Kornberg, H. L., & Krebs, H. A. (1957). Synthesis of cell constituents from C 2-units by a modified tricarboxylic acid cycle. *Nature*, 179(4568), 988-991.
- Lam, W. L., & Doolittle, W. F. (1989). Shuttle vectors for the archaeobacterium *Halobacterium volcanii*. *Proceedings of the National Academy of Sciences*, 86(14), 5478-5482.
- Lang, J. M., Darling, A. E., & Eisen, J. A. (2013). Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PloS one*, 8(4), e62510.
- Leichtling, B. H., Rickenberg, H. V., Seely, R. J., Fahrney, D. E., & Pace, N. R. (1986). The occurrence of cyclic AMP in archaeobacteria. *Biochemical and biophysical research communications*, 136(3), 1078-1082.
- Leigh, J. A., Albers, S. V., Atomi, H., & Allers, T. (2011). Model organisms for genetics in the domain Archaea: methanogens, halophiles, Thermococcales and Sulfolobales. *FEMS microbiology reviews*, 35(4), 577-608.
- Lockington, R. A., & Kelly, J. M. (2001). Carbon catabolite repression in *Aspergillus nidulans* involves deubiquitination. *Molecular microbiology*, 40(6), 1311-1321.
- Lubelska, J. M., Jonuscheit, M., Schleper, C., Albers, S. V., & Driessen, A. J. (2006). Regulation of expression of the arabinose and glucose transporter genes in the thermophilic archaeon *Sulfolobus solfataricus*. *Extremophiles*, 10(5), 383-391.
- Martin, J. H., Sherwood Rawls, K., Chan, J. C., Hwang, S., Martinez-Pastor, M., McMillan, L. J., ... & Maupin-Furlow, J. A. (2018). GlpR is a direct transcriptional repressor of fructose metabolic genes in *Haloferax volcanii*. *Journal of bacteriology*, 200(17), e00244-18.
- Moissl-Eichinger, C., Pausan, M., Taffner, J., Berg, G., Bang, C., & Schmitz, R. A. (2018). Archaea are interactive components of complex microbiomes. *Trends in microbiology*, 26(1), 70-85.
- Moreno, M. S., Schneider, B. L., Maile, R. R., Weyler, W., & Saier Jr, M. H. (2001). Catabolite repression mediated by the CcpA protein in *Bacillus subtilis*: novel modes of regulation revealed by whole-genome analyses. *Molecular microbiology*, 39(5), 1366-1381.
- Oehler, S., Eismann, E. R., Krämer, H., & Müller-Hill, B. (1990). The three operators of the lac operon cooperate in repression. *The EMBO journal*, 9(4), 973-979.
- Oesterhelt, D., & Stoekenius, W. (1971). Rhodopsin-like protein from the purple membrane of *Halobacterium halobium*. *Nature new biology*, 233(39), 149-152.
- Oren, A. (2002). Diversity of halophilic microorganisms: environments, phylogeny, physiology, and applications. *Journal of Industrial Microbiology and Biotechnology*, 28(1), 56-63.
- Parche, S., Beleut, M., Rezzonico, E., Jacobs, D., Arigoni, F., Titgemeyer, F., & Jankovic, I. (2006). Lactose-over-glucose preference in *Bifidobacterium longum* NCC2705: glcP, encoding a glucose transporter, is subject to lactose repression. *Journal of bacteriology*, 188(4), 1260-1265.
- Park, E. A., Song, S., Vinson, C., & Roesler, W. J. (1999). Role of CCAAT enhancer-binding protein β in the thyroid hormone and cAMP induction of phosphoenolpyruvate carboxykinase gene transcription. *Journal of Biological Chemistry*, 274(1), 211-217.
- Peck, R. F., DasSarma, S., & Krebs, M. P. (2000). Homologous gene knockout in the archaeon *Halobacterium salinarum* with *ura3* as a counterselectable marker. *Molecular microbiology*, 35(3), 667-676.

- Porat, I., Sieprawska-Lupa, M., Teng, Q., Bohanon, F. J., White, R. H., & Whitman, W. B. (2006). Biochemical and genetic characterization of an early step in a novel pathway for the biosynthesis of aromatic amino acids and p-aminobenzoic acid in the archaeon *Methanococcus maripaludis*. *Molecular microbiology*, 62(4), 1117-1131.
- Qureshi, S. A., Bell, S. D., & Jackson, S. P. (1997). Factor requirements for transcription in the Archaeon *Sulfolobus shibatae*. *The EMBO journal*, 16(10), 2927-2936.
- Robinson, J. L., Pyzyrna, B., Atrasz, R. G., Henderson, C. A., Morrill, K. L., Burd, A. M., ... & Shand, R. F. (2005). Growth kinetics of extremely halophilic Archaea (family Halobacteriaceae) as revealed by Arrhenius plots. *Journal of bacteriology*, 187(3), 923-929.
- Roesler, W. J. (2000). At the Cutting Edge What is a cAMP response unit?. *Molecular and cellular endocrinology*, 162(1-2), 1-7.
- Rolfsmeier, M., & Blum, P. (1995). Purification and characterization of a maltase from the extremely thermophilic crenarchaeote *Sulfolobus solfataricus*. *Journal of bacteriology*, 177(2), 482-485.
- Rolfsmeier, M., Haseltine, C., Bini, E., Clark, A., & Blum, P. (1998). Molecular characterization of the α -glucosidase gene (malA) from the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Journal of bacteriology*, 180(5), 1287-1295.
- Ronne, H. (1995). Glucose repression in fungi. *Trends in Genetics*, 11(1), 12-17.
- Saier Jr, M. H. (1996). Cyclic AMP-independent catabolite repression in bacteria. *FEMS microbiology letters*, 138(2-3), 97-103.
- Schramm, A., Siebers, B., Tjaden, B., Brinkmann, H., & Hensel, R. (2000). Pyruvate kinase of the hyperthermophilic crenarchaeote *Thermoproteus tenax*: physiological role and phylogenetic aspects. *Journal of bacteriology*, 182(7), 2001-2009.
- Shand, R. F., & Perez, A. M. (1999). Haloarchaeal growth physiology. In *Enigmatic microorganisms and life in extreme environments* (pp. 411-424). Springer, Dordrecht.
- Sherwood, K. E., Cano, D. J., & Maupin-Furlow, J. A. (2009). Glycerol-mediated repression of glucose metabolism and glycerol kinase as the sole route of glycerol catabolism in the haloarchaeon *Haloferax volcanii*. *Journal of bacteriology*, 191(13), 4307-4315.
- Siebers, B., Wendisch, V. F., & Hensel, R. (1997). Carbohydrate metabolism in *Thermoproteus tenax*: in vivo utilization of the non-phosphorylative Entner-Doudoroff pathway and characterization of its first enzyme, glucose dehydrogenase. *Archives of microbiology*, 168(2), 120-127.
- Siebers, B., Klenk, H. P., & Hensel, R. (1998). PPi-dependent phosphofructokinase from *Thermoproteus tenax*, an archaeal descendant of an ancient line in phosphofructokinase evolution. *Journal of Bacteriology*, 180(8), 2137-2143.
- Snijders, A. P., Walther, J., Peter, S., Kinnman, I., De Vos, M. G., Van de Werken, H. J., ... & Wright, P. C. (2006). Reconstruction of central carbon metabolism in *Sulfolobus solfataricus* using a two-dimensional gel electrophoresis map, stable isotope labelling and DNA microarray analysis. *Proteomics*, 6(5), 1518-1529.
- Stincone, A., Prigione, A., Cramer, T., Wamelink, M. M., Campbell, K., Cheung, E., ... & Ralser, M. (2015). The return of metabolism: biochemistry and physiology of the pentose phosphate pathway. *Biological Reviews*, 90(3), 927-963.
- Titgemeyer, F., Mason, R. E., & Saier Jr, M. H. (1994). Regulation of the raffinose permease of *Escherichia coli* by the glucose-specific enzyme IIA of the phosphoenolpyruvate: sugar phosphotransferase system. *Journal of bacteriology*, 176(2), 543-546.
- Todor, H., Sharma, K., Pittman, A. M., & Schmid, A. K. (2013). Protein–DNA binding dynamics predict transcriptional response to nutrients in archaea. *Nucleic acids research*, 41(18), 8546-8558.

- Torregrosa-Crespo, J., Pire Galiana, C., & Martínez-Espinosa, R. M. (2017). Biocompounds from Haloarchaea and their uses in biotechnology. Archaea-New Biocatalysts, Novel Pharmaceuticals and Various Biotechnological Applications. InTech, 63-82.
- Trumbly, R. J. (1992). Glucose repression in the yeast *Saccharomyces cerevisiae*. Molecular microbiology, 6(1), 15-21.
- van den Bogaard, P. T., Kleerebezem, M., Kuipers, O. P., & de Vos, W. M. (2000). Control of lactose transport, β -galactosidase activity, and glycolysis by CcpA in *Streptococcus thermophilus*: evidence for carbon catabolite repression by a non-phosphoenolpyruvate-dependent phosphotransferase system sugar. Journal of bacteriology, 182(21), 5982-5989.
- van der Oost, J., Schut, G., Kengen, S. M., Hagen, W. R., Thomm, M., & de Vos, W. M. (1998). The Ferredoxin-dependent Conversion of Glyceraldehyde-3-phosphate in the Hyperthermophilic Archaeon *Pyrococcus furiosus* Represents a Novel Site of Glycolytic Regulation. Journal of Biological Chemistry, 273(43), 28149-28154.
- Van Wezel, G. P., König, M., Mahr, K., Nothaft, H., Thomae, A. W., Bibb, M., & Titgemeyer, F. (2007). A new piece of an old jigsaw: glucose kinase is activated posttranslationally in a glucose transport-dependent manner in *Streptomyces coelicolor* A3 (2). Journal of molecular microbiology and biotechnology, 12(1-2), 67-74.
- Williams, T. J., Allen, M. A., Liao, Y., Raftery, M. J., & Cavicchioli, R. (2019). Sucrose metabolism in haloarchaea: reassessment using genomics, proteomics, and metagenomics. Applied and environmental microbiology, 85(6), e02935-18.
- Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences, 74(11), 5088-5090.
- Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences, 87(12), 4576-4579.
- Xavier, K. B., da Costa, M. S., & Santos, H. (2000). Demonstration of a Novel Glycolytic Pathway in the Hyperthermophilic Archaeon *Thermococcus zilligii* by ^{13}C -Labeling Experiments and Nuclear Magnetic Resonance Analysis. Journal of bacteriology, 182(16), 4632-4636.
- Zillig, W., Stetter, K. O., & Janeković, D. (1979). DNA-dependent RNA Polymerase From the Archaeobacterium *Sulfolobus acidocaldarius*. European journal of biochemistry, 96(3), 597-604.

Chapter 2. Carbon Utilization And Regulation In The Halophilic Archaeon

Halogeometricum borinquense

2.1 Publication Status

Reviewing with co-authors before submitting for publication.

Abstract: This study combines growth studies and gene expression profiling through RNA-Seq and qRT-PCR to investigate carbohydrate metabolism and its regulation in the model haloarchaeon *Halogeometricum borinquense*. We specifically characterize both the unique and shared gene sets involved in the utilization of glucose, sucrose, maltose, and combinations thereof, and test for evidence of carbon catabolite repression (CCR) in this organism. Broadly, the sugar conditions show higher expression of genes for glycolysis, the glyoxylate pathway, and amino acid transport and synthesis relative to the rich media condition. The strongest sugar-specific differences in transcriptional profiles were observed in transport processes and key nodes in central carbon and nitrogen metabolism. Glucose showed an activating effect on the transcription of a key signal transduction molecule, nitrogen regulatory protein PII, and a repressive effect on metabolic genes citrate synthase, isocitrate lyase, and acetolactate synthase. The observation of glucose-induced repression provides preliminary evidence of CCR in this organism. This study represents the first transcriptome-level characterization of *H. borinquense* and contributes to our understanding of carbon and nitrogen utilization and regulation in archaea that live in hypersaline environments.

2.2 Author Contributions

HGK, AIY, MTF, RMR, and MC contributed to the conception and design of the experiments and performed the experiments. HGK, AIY, RMR, and MTF performed the data and statistical

analyses. RMR and MTF contributed reagents, materials, and analysis tools. HGK, AIY, RMR, and MTF wrote the manuscript.

2.3 Introduction

Mechanisms of gene regulation contribute to organismal fitness by dynamically tuning cellular responses to meet the challenge posed by diverse, simultaneous, and fluctuating stressors present in an organism's environment. A class of mechanisms known generally as carbon catabolite repression (CCR) helps some cells control metabolism and programs of cell division in response to carbon source availability. While diverse mechanisms for CCR exist in Nature (Bini, 2001; Görke, 2008; Brückner, 2002), a common phenotype exhibited by most CCR-based regulatory schemes is the activation of metabolic pathways and cellular functions associated with the utilization of preferred carbon source(s) and the concomitant repression of those associated with the utilization of less favored substrates.

Studies have revealed that at least some archaea exhibit phenotypic hallmarks of CCR. Archaea such as *Sulfolobus solfataricus* (De Rosa, 1975, 1984; Snijders, 2000; Haseltine, 1996, 1999a, 1999b), *Thermococcus zilligi* (Xavier, 2000), and *Pyrococcus furiosus* (van der Oost, 1998) have been shown to preferentially utilize certain nutrients when placed in conditions with multiple potential nutrient sources. In some cases, the phenotype of diauxic growth, often associated with CCR in bacteria and eukaryotes (Bini, 2001), has been associated with the differential utilization of food sources. Other evidence suggestive of CCR includes differential expression of proteins needed for carbohydrate uptake and catabolism, such as ABC transporters and glycosyl hydrolases (e.g., alpha-glucosidase or maltase, alpha-amylase, and beta-glycosidase) (Lubelska, 2006; Haseltine, 1996, 1999a, 1999b; Rolfsmeier, 1995, 1998). Despite all of this circumstantial evidence, no mechanisms of CCR have yet to be positively described in the archaea.

The haloarchaea, a distinct phylogenetic clade of archaea that thrive in hypersaline environments (2-4.5 M NaCl) (Torregrosa-Crespo, 2017), have qualities that make them suitable for investigating CCR and other mechanisms of metabolically-associated gene regulation (Leigh, 2011; Berquist, 2006; Anderson, 2011; Becker, 2014; Born, 2019). Observations of preferential carbon source utilization, diauxic growth, and differential expression of select catabolic genes under different sugar conditions in *Halococcus saccharolyticus*, *Haloferax volcanii*, *Haloarcula vallismortis*, and *Halobacterium halobium* (Johnsen, 2001; Altekar, 1992; Bonete, 1996; Bräsen, 2004) all hint that CCR may be present.

In this study, we characterize the regulation of genes involved in the catabolism of glucose, maltose, and sucrose in the metabolically versatile haloarchaeon *Halogeometricum borinquense*, known for its ability to utilize numerous carbohydrates, including glucose, mannose, fructose, sucrose, xylose, maltose, trehalose, cellobiose, raffinose and glycerol as sole carbon sources (Montalvo-Rodriguez, 1998; Williams, 2019). We integrate quantitative phenotypes through culturing *H. borinquense* on each sugar (and combinations thereof), broad transcriptomic profiling, targeted qRT-PCR, and sugar utilization assays to begin delineating the unique and shared responses of *H. borinquense* to these carbohydrates, identify key regulatory points in metabolism, and potential target genes subject to regulation by CCR. This work is also the first transcriptomic study in *H. borinquense* and lays the foundation for comparative analyses of carbohydrate metabolism in the haloarchaea.

2.4 Materials & Methodology

2.4.1 Strain And Growth Conditions

Halogeometricum borinquense (DSM 11551) was obtained from Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ; Braunschweig, Germany). Cells were cultured in aerobic conditions at 40°C in a New Brunswick Innova 44R incubator shaker set to 250 rpm

(Edison, New Jersey, USA); the culture flasks had a foam stopper and were loosely covered by aluminum foil.

Cells were cultured in a 250 mL Erlenmeyer flask containing 100 mL of media using the following formulations. Rich media, also known as Sehgal Gibbon's media (per 1 L): 3 g trisodium citrate, 2 g KCl, 20 g $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$, 200 g NaCl, 23 μg FeCl_3 , 10 g yeast extract, and 7.5 g casamino acids (Sehgal, 1960). The pH of the media was adjusted to 7.2-7.8 with 1 M NaOH, and the media was then autoclaved at 121°C for 15 min. Minimal media, a modification of the rich media that consists of (per 1 L): 2 g KCl, 20 g $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$, 200 g NaCl, 23 μg FeCl_3 , and 1 g yeast extract. For standard growth assays, individual carbohydrates were added to minimal media in the following concentrations: maltose (Pfanstiehl; Waukegan, IL, USA) 0.6% (w/v), sucrose (Pfanstiehl; Waukegan, IL, USA) 0.4 (w/v), and glucose (Fisher Scientific; Waltham, MA, USA) 0.4% (w/v) (Figure 2.1). Carbohydrate concentrations were determined by repeated in-house experiments that investigated optical density measurements at 600 nm (OD_{600}) at 0.2%, 0.4%, 0.6%, and 0.8% (w/v) of each sugar, alongside rich and minimal media conditions (Appendix 2.1). The pH of all media was adjusted to ~7.5 with 1 M NaOH. Minimal media and media with sugars added were filter sterilized with 0.2- μm membrane filters.

Media were inoculated with cultures adapted to minimal media at the logarithmic phase. Pre-cultures for inoculation were centrifuged at 4,000 x g for 5 min, the pellets were resuspended in the downstream media to a final OD_{600} of 0.05. Growth was monitored by measuring OD_{600} of 100 μL aliquots using an Eppendorf BioPhotometer (Eppendorf; Hamburg, Germany).

2.4.2 RNA-Seq

RNA extractions were performed with cells collected at a point during the late logarithmic phase where differential OD_{600} phenotypes between experimental conditions could be observed. Two biological replicates and two technical replicates were harvested for each sugar condition.

Total RNA was extracted using TRIzol reagent (15596-026, Invitrogen; Carlsbad, CA, USA). Briefly, cells were harvested by centrifuging 1 mL of approximately 1×10^8 cells at $10,000 \times g$ for 1 min at 37°C . The supernatant was aspirated and the remaining cells were flash frozen using liquid nitrogen and stored at -80°C for RNA extraction. Cell pellets were lysed using 1 mL of TRIzol reagent by pipetting the frozen pellet up and down and incubating for 5 min at room temperature. 0.2 mL of chloroform was added, and the tube was vigorously mixed by inversion and incubated for 2-3 min at room temperature. The TRIzol/chloroform mixture was then centrifuged at $12,000 \times g$ for 15 min at 4°C . The aqueous phase was removed and transferred into a new tube. Nucleic acid was precipitated from the aqueous phase by adding 0.5 mL of 100% isopropanol. This mixture was incubated at room temperature for 10 min and then centrifuged at $12,000 \times g$ for 10 min at 4°C . The supernatant was aspirated, retaining the pellet, and washed with 1 mL of 75% ethanol. The sample was briefly vortexed, and then centrifuged at $7,500 \times g$ for 5 min at 4°C . The supernatant was discarded and the pellet was air dried for 5-10 min. The pellet was then suspended in 8 μL of RNase-free water, followed by an incubation in a water bath or heat block at $55\text{-}60^\circ\text{C}$ for 10-15 min.

A DNase treatment to remove gDNA was performed as follows. To the suspended RNA, 1 μL of DNase I and 1 μL of 10X Reaction Buffer were added and gently mixed (Thermo Fisher; Waltham, MA, USA). The mixture was incubated for 15 min at room temperature. Then 1 μL of 50 mM EDTA stop solution was added, followed by an incubation at 70°C for 10 min. A second TRIzol extraction was performed, repeating the procedure described above, modifying the use of TRIzol reagent to 0.5 mL instead of 1 mL and reducing the other reagents to scale. The samples were suspended in 22.5 μL H_2O ; 1.5 μL was aliquoted for use in assessing RNA integrity on the Agilent 2100 Bioanalyzer (Agilent Technologies; Palo Alto, CA, USA) and another 1.5 μL was aliquoted for 16S rRNA PCR to confirm the absence of genomic DNA in the samples.

To test for gDNA contamination, 10 ng of Nanodrop (ThermoFisher, Waltham, MA, USA) quantified RNA was added to 2 μ L 10X buffer, 4 μ L Q solution, 1 μ L 10 μ M archaeal 16S rDNA forward primer (TCCGGTTGATCCTGCCRG; Archaeal_F_4Fa), 1 μ L 10 μ M archaeal 16S rDNA reverse primer (ACGGHTACCTTGTTACGACTT; Archaeal_and_Halo_R), 0.5 μ L 10 mM dNTPs, 0.125 μ L Taq DNA polymerase (Qiagen; Hillden, Germany), in a 20 μ L reaction volume. The PCR was run with the following settings on the thermocycler: initial denaturation of 10 min at 98°C, 35 cycles of 45 sec at 98°C, 1 min at 55°C, 90 sec at 72°C, and a final polymerization for 5 min at 72°C and hold at 4°C. Samples were considered gDNA-free if no observable product appeared after PCR on an 1% (w/v) agarose gel. Samples with persistent gDNA contamination were redigested with DNase I as described above until no 16S PCR products were observed.

To prepare total RNA for library construction, the ribosomal RNA (rRNA) was fragmented from total RNA samples using a protocol adapted from Morlan (2012) and Sharma (2021). For each sample, approximately 2 μ g total RNA was combined with 0.6 μ L 10X hybridization buffer (1M Tris-HCl [pH 7.0], 200 mM NaCl, RNase-free H₂O), 2 μ L synthetic DNA oligonucleotide probe mixture (96 oligos at 0.5 μ M/oligo) corresponding to 5S, 16S, and 23S subunits, and RNase-free H₂O to a total volume of 6 μ L (Appendix 2.2). Hybridization was performed in a thermal cycler (Bio-Rad; Hercules, CA, USA) programmed as follows: 95°C for 2 min, ramp down to 45°C at -0.1°C/sec, 45°C for 5 min, and hold at 45°C. To the hybridized total RNA, 1 μ L of pre-heated 10X Hybridase buffer (500 mM Tris-HCl pH 7.4, 1 M NaCl, and 200 mM MgCl₂), 1 μ L pre-heated RNase-free H₂O, and 2 μ L pre-heated thermostable Hybridase RNase H (5 U/ μ L; Epicentre Biotechnologies; Madison, WI, USA) was added and incubated at 45°C for 40 min. Hybridization oligos were removed using RNase-free DNase I (New England Biolabs; Ipswich, MA, USA), and 10 μ L 10X DNase I buffer, and 2 μ L DNase I were added to the RNase H digestion reaction (final volume 100 μ L) and incubated at 37°C for 40 min.

Agencourt RNAClean XP beads (Beckman Coulter; Brea, CA, USA) were used to purify the rRNA depleted RNA using a 1.8X volume of beads to sample ratio, as per manufacturer's protocol. The purified RNA was eluted off the Agencourt beads with 13 μ L RNase-free H₂O and used to prepare sequencing libraries using the KAPA Stranded RNA-Seq Library Preparation Kit for Illumina Platforms (KR0934 – v2.16) from Kapa Biosystems (Wilmington, MA, USA). A total of 23 barcoded libraries were constructed. Quality of the libraries were assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies; Palo Alto, CA, USA) and quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher; Waltham, MA) on the Qubit 2 Fluorometer (Thermo Fisher; Waltham, MA, USA) before being pooled and sequenced (Appendix 2.3).

2.4.3 Bioinformatics

The Illumina NextSeq 500 platform (SR50) was used to sequence the pooled library. Libraries were demultiplexed after sequencing and quality assessment using FastQC (Andrews, 2010). Adapter sequences and low quality reads were removed using Trimmomatic (Bolger, 2014). Trimmed reads were mapped to the *H. borinquense* genome (NCBI ID: 469382) (Malfatti, 2009) using Bowtie (Langmead, 2009). Counts of aligned reads mapping to NCBI predicted genes were calculated using HTSeqcount (Anders, 2015). Differential gene expression (DE) and statistical analyses were performed in R Studio using the package *DESeq2* (Love, 2014). The heatmap was constructed with vst transformed data, using the 'complete' clustering algorithm for significantly differentially expressed genes with +/- 2 log₂-fold change (794 genes total).

2.4.4 Phenotyping, Sample Collection, And Metabolite Analysis

Phenotyping assays were conducted by monitoring OD₆₀₀ of biological duplicates during growth of *H. borinquense* cultured in the following media: (1) Minimal media (Min); (2) 0.4%

(w/v) Glucose (G); (3) 0.6% (w/v) Maltose (M); (4) 0.6% (w/v) Maltose with 0.4% (w/v) glucose added at t=44 hrs (MaddG); (5) 0.4% (w/v) Sucrose; (6) 0.4% (w/v) Sucrose with 0.4% (w/v) glucose (from t=0 hr) (SG); and (7) 0.4% (w/v) Sucrose with 0.4% (w/v) glucose added at t=44 hrs (SaddG). The 44 hour timepoint was determined by collecting preliminary phenotyping curves and identifying the timepoint at which cultures grown in the presence of sucrose started exhibiting a distinct OD₆₀₀ phenotype to those grown in minimal and maltose conditions.

Throughout the duration of the culturing, we measured OD₆₀₀ every 3-5 hours with 100 μ L aliquots of the culture, using minimal media as a blank. Growth curve spline fitting and statistical analyses were performed in R Studio using the package *dpseg* (Machné, 2020). During 1-2 daily timepoints per day of culturing, 200 μ L aliquots were collected for dilution-based colony forming unit (CFU) analysis to assess growth by cell viability. During certain timepoints (see Appendix 2.4 for specific details regarding timepoints and collection strategy), up to 1 mL of culture was removed from the flasks and the cells harvested by centrifugation at 10,000 x g for 1 min at 37°C. The supernatants were transferred to a microcentrifuge tube and stored at -20°C for subsequent metabolite analysis, whereas the pellet was snap frozen in a dry ice-ethanol bath and then stored at -80°C. The Maltose-Sucrose-D-Glucose Assay Kit (K-MASUG, Megazyme; Bray, Ireland) was used for metabolite analysis according to the manufacturer's instructions, using a dilution factor of 10 to account for salt and carbohydrate concentrations in the media. Reaction volumes were also scaled down by a factor of 13.1 to accommodate usage of 96-well plates (Corning; Corning, NY, USA) in the place of cuvettes. Sugar assay curves were manually linearly segmented (Appendix 2.5). Absorbance measurements were taken with the Tecan Infinite F200 Pro microplate reader (Tecan; Mannedorf, Switzerland). Cell pellets were used for qRT-PCR, as detailed below.

2.4.5 Quantitative PCR

The protocol used for RNA extractions for qRT-PCR experiments mirrored the aforementioned methods, with slight modifications. Namely, the total RNA was isolated from frozen cell pellets using 1.0 mL of the TRIzol Reagent (one round of TRIzol extraction only). The DNase treatment was performed with the DNA-free DNA removal kit (Thermo Fisher; Waltham, MA, USA) using 3 μ L Inactivation Reagent. To remove impurities, a secondary wash procedure was performed, which consisted of the ethanol wash steps from the TRIzol protocol. Aliquots of samples were taken, as previously described, to check for RNA quality and genomic DNA contamination (Agilent Technologies; Palo Alto, CA, USA). Reverse transcription was performed according to the SuperScript IV Reverse Transcriptase kit (Invitrogen; Carlsbad, CA, USA) manual, with the inclusion of the optional RNaseH RNA removal step.

Quantitative PCR with 10 μ L reaction volumes was performed using KAPA SYBR FAST qPCR Master Mix (2X) kit (Kapa Biosystems; Wilmington, MA, USA) on a Roche Cobas Z 480 analyzer (Roche; Basel, Switzerland), using white Roche LightCycler 480 Multiwell 96-well plates and sealing film (Cat #: 04729692001, Roche; Basel, Switzerland). Internal controls included “Hbor_06420,” a highly expressed “Cell Division Protein” (Forward primer sequence: GGAAGTGGATCGAGGAGCAG, Reverse primer sequence: CGTCGATGTTTTCTTGCGCT), “Hbor_26990,” an annotated glyceraldehyde-3-phosphate dehydrogenase (Forward primer sequence: GTGACCTCTGGGAGAACTGC, Reverse primer sequence: ATCCGTCGTCACCATGCTTT), and, “Hbor_22340,” a transcription initiation factor IIB 2 (Forward primer sequence: AGAGTGACCGAGAGGCTGAA, Reverse primer sequence: TCTTTTCCCTCGCGTCGAA) (Integrated DNA Technologies; Skokie, IL). The thermal cycler was programmed as follows: enzyme activation at 95°C for 3 min, 40 cycles of denaturation at 95° for 1 sec and annealing at 65°C for 20 sec, and finally cool down to 37°C for 10 sec and then to 25°C for 10 sec.

2.5 Results

2.5.1 Culture Phenotypes

In-flask cultures of cells in different media exhibited different visible phenotypes, particularly as end-of-experiment cultures were allowed to sit on the benchtop. First, cultures of *H. borinquense* grown in minimal media lacked the typical pink coloration of cultures grown in rich media and other sugar-containing media. Moreover, even in conditions in which pigmentation was clearly evident, different shades of pink were observed. Of the sugar conditions, cells grown in maltose were the least pigmented, while those grown in glucose only appeared to have the most intense pigmentation. As cell cultures aged on the benchtop, coloration slowly faded from cells grown in glucose- and sucrose-containing media, with the condition to which glucose was added late (SaddG), “bleaching” fastest.

In addition to differences in pigmentation, cell cultures also exhibited differential accumulation of gas vesicles as judged by a “milky” appearance and bright birefringent “spots” under phase-contrast. Cells grown in the presence of a carbohydrate source produced large amounts of gas vesicles while cultures of those grown in minimal media did not appear milky. Finally, cells grown in the presence of sucrose, clumped and/or formed biofilm sheets, presumably an extracellular polymeric substance (EPS), after multiple undisturbed days on the benchtop. Cells grown in other conditions did not display a clumping/EPS phenotype.

2.5.2 OD₆₀₀, CFU Trajectory Phenotypes, And Sugar Utilization

When cultured in the media described in Methods, *H. borinquense* exhibits phenotypic behavior characterized by several distinct, condition-specific, transitions of slope when measured over time using OD₆₀₀. Rather than using the term “growth curve” we refer to this type of data as phenotype trajectories or OD₆₀₀ trajectories to acknowledge that cell division (growth) is likely not the sole factor contributing to light scattering. In the case of *H. borinquense*, both

growth and intra- or extracellular gas vesicle accumulation will contribute to the phenotype (Yao, 2011; Darnell, 2017; Facciotti, 2010). Other unknown factors, such as cell shape, free gas vesicles, pigment production, may also contribute.

2.5.2.1 OD₆₀₀ And CFU Trajectory Phenotype

Figure 2.1 shows the average of two replicate OD₆₀₀ trajectories sampled from each of the experimental conditions over a 191 hour culture period. To more quantitatively characterize these phenotypes, each individual curve was linearly segmented. Appendix 2.5 reports the breakpoints and slopes for each linear segment.

For the first 15 hours of the assay, all conditions show a similar increase in OD₆₀₀. The OD₆₀₀ trajectory of the Min condition is characterized by a rapid increase in OD₆₀₀ from t=0-24 hrs (0.1576 OD₆₀₀/hr) that is comparable to increases seen in non-glucose-containing media. Thereafter, the OD₆₀₀ trajectory of the Min condition plateaus for the remainder of the experiment. The OD₆₀₀ trajectory of the sucrose-containing (S and SaddG) and M conditions diverge from that of the Min condition at t=24 hrs (at this point, glucose has not yet been added to SaddG), and t=56 hrs, respectively, suggesting the beginning of sugar utilization at those timepoints. The OD₆₀₀ trajectory of cells grown in conditions that contain glucose from the onset of the experiment (G and SG) show an earlier deviation from the Min condition, increasing in OD₆₀₀ between t=15-30 hrs before each plateauing by t=78 hrs (~2.5 OD₆₀₀ for G and ~3.0 OD₆₀₀ for SG).

MaddG and SaddG show different behaviors when glucose is added at t=44 hrs. In MaddG, the OD₆₀₀/hr rate immediately increases from -0.0133 to 0.0382 OD₆₀₀/hr. SaddG, on the other hand, experiences a lag of about 12 hours (-0.0062 OD₆₀₀/hr), potentially pointing to a metabolic transition phase. Afterwards, the SaddG OD₆₀₀ increases, though at a slower rate than the MaddG condition (0.0184 OD₆₀₀/hr). By t=190 hrs, all conditions except for M and Min have surpassed G's OD₆₀₀ value.

A comparison of CFU measurements reveals that most of the growth in the cultures may be driven largely by the amount of yeast extract in the media. The Min condition supports a maximum culture carrying capacity that is either equivalent to or greater than those of the sugar containing conditions. Moreover, cultures appear to plateau faster in the Min condition with sugar-containing conditions, showing what appears to be a short period of adaptation and slower growth rate. These observations seem counterintuitive given that: (a) the carbohydrates are consumed (see Sugar Utilization below) and (b) the OD₆₀₀ trajectories for all cultures grown in sugar-containing media ultimately reach significantly higher values than those grown in Min. A possible explanation, explored more below and consistent with the available data, is that the catabolism of exogenous carbohydrate may be largely invested in the production of intracellular structures like gas vesicles (particularly in glucose containing media) and the extracellular polymeric substance that appears primarily in sucrose containing media. Under these experimental conditions, we suspect that cell division is limited, not by carbon availability, but rather another key nutrient like nitrogen. Our analysis of other data is guided at least in part by this physiological context hypothesis.

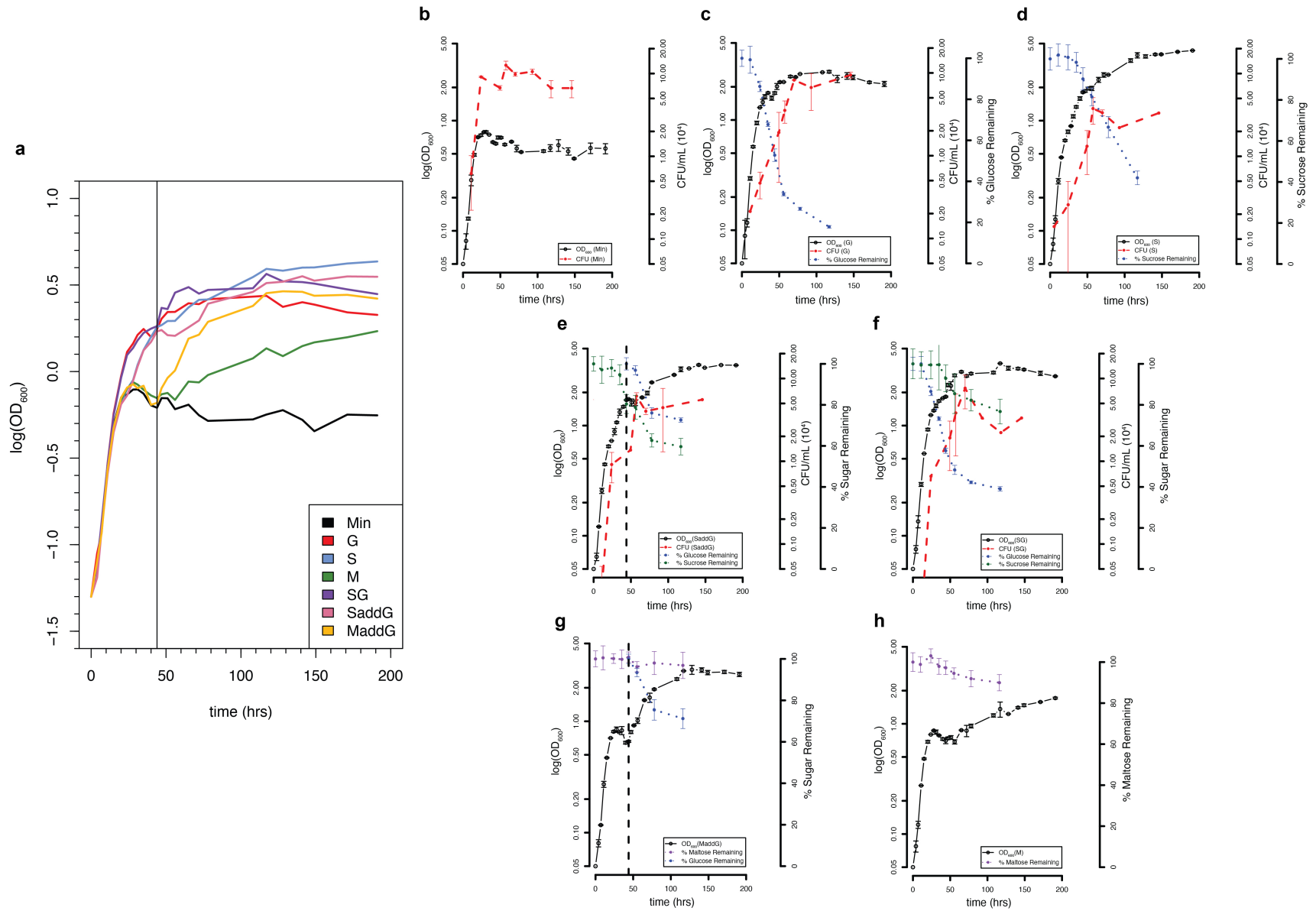


Figure 2.1 a. Absorbance measurements ($\log(\text{OD}_{600})$) of *H. borinquense* grown on various carbon sources over time (duplicates averaged). The dashed line indicates the timepoint at which glucose was added to SaddG and MaddG conditions ($t=44$ hrs).
b-h. *H. borinquense* absorbance measurements (OD_{600}), CFU, and sugar utilization over time (duplicates averaged).

2.5.2.2 Sugar Utilization

To determine the rate of sugar utilization in the cultures, the concentrations of glucose, sucrose, and maltose in spent culture media were determined at various times during culture. The dotted lines in Figure 2.1c-h shows the utilization of each carbohydrate during each respective OD₆₀₀ trajectory experiment and Appendix 2.5 shows utilization rates for all applicable conditions over segments of time.

Single-sugar conditions

In condition G, glucose is initially utilized at a rate of 2.3 mg/hr. At t=10.7 hrs, the rate of glucose utilization increases to 35.7 mg/hr. Glucose utilization peaks at 58.7 mg/hr between hours 23.9 and 55.7, before sharply declining to 9.2 mg/hr thereafter. In condition S, sucrose utilization rates lag behind those of glucose in the G condition, with heavier consumption beginning only at t=34.7 hrs (increasing ten-fold from 2.3 mg/hr to 29.4 mg/hr). The sharp decrease in sugar utilization rate seen at the end of the G condition's trajectory is not seen in the S condition by t=116.7 hrs. At this same timepoint (the last timepoint in the sugar utilization assay, t=116.7 hrs), ~20% and ~40% of the glucose and sucrose remain unused, respectively. Maltose utilization is minimal compared to sucrose and glucose, with the highest rate of utilization throughout the experiment being 3.9 mg/hr.

Multi-sugar conditions

In the SG condition, we observed early lags in utilization of each sugar, similar to those seen in the S and G conditions. By t=116.7 hrs, approximately twice as much glucose (~40% of initial) and sucrose (~80% of initial) remain in the SG condition, relative to their levels in each single sugar condition. At this point (t=116.7 hrs), the rate of glucose and sucrose consumption combined is 11.5 mg/hr in SG, whereas glucose is utilized at 9.2 mg/hr in the G condition, and sucrose is utilized at 29.4 mg/hr in the S condition.

In the SaddG condition, sucrose utilization is comparable to the S condition, and the utilization rate of glucose first starts at 7.7 mg/hr before increasing at $t=55.7$ hrs to a rate of 28.2 mg/hr. Between $t=78.4$ to 116.7 hrs, the utilization rate drops to 2.6 mg/hr. In MaddG, the addition of glucose results in an immediate utilization rate of 16.5 mg/hr, increasing to 24.6 mg/hr before slowing to 3.2 mg/hr by $t=78.4$ hrs. By $t=116.7$ hrs, there is ~60% sucrose remaining in the SaddG condition, compared to the 40% remaining in the S condition.

In the G, SG, SaddG, and MaddG conditions, there is a point at which sugar utilization rate sharply decreases despite availability of each of the carbohydrates. In the conditions where glucose is present from the start (G, SG), this timepoint is $t=55.7$ hrs, and in the conditions where glucose is added at $t=44$ hrs (SaddG, MaddG), this occurs later at $t=78.4$ hrs. In the glucose-containing media (including SaddG), this transition occurs as total sugar utilization reaches ~2 g (with the range being -1.8-2.2 g) (Appendix 2.5). In the maltose condition, the rapid decrease in utilization rate occurs at a lower total sugar utilization of approximately 0.7 g. In each case, the available sugar is far from depleted, indicating that other limiting factors are likely responsible for the slowed sugar consumption. This observation helps shape the hypothesis noted earlier that growth of *H. borinquense* is limited by a nutrient other than carbon.

The single-sugar utilization phenotypes observed together with the OD₆₀₀ phenotypes suggest that *Hbor* utilizes glucose most readily, followed by sucrose, and then maltose. With the addition of exogenous glucose at $t=44$ hrs in SaddG, glucose utilization rapidly occurs and sucrose utilization is shown to be abated, potentially pointing to a transition in metabolism. This may be related to how the later addition of glucose in the SaddG condition results in a brief lag phase before OD₆₀₀ increases thereafter.

2.5.3 RNA-Seq-Based Phenotyping

Phenotypic characterization of growth on different substrates provides a high-level context from which to begin studying the underlying patterns of gene expression that ultimately

contribute to the physiological responses and phenotypic differences we observe from cells grown on different carbohydrate sources. To begin characterizing the transcriptional response to different carbon sources, we selected a single timepoint at a point in late logarithmic phase where differential OD₆₀₀ phenotypes between experimental conditions could be observed. We harvested cells for RNA-Seq based phenotyping and constructed sequencing libraries as described in Methods. The resulting transcriptome consisted of 477M quality filtered reads, of which 6% is PhiX with a Q30 of 96.2%. A principal component analysis (PCA) was performed (Figure 2.2), with replicates (n=5 for G; n=6 for all other conditions) clustering tightly together, indicating that condition replicates are each more similar to themselves than to samples from other conditions.

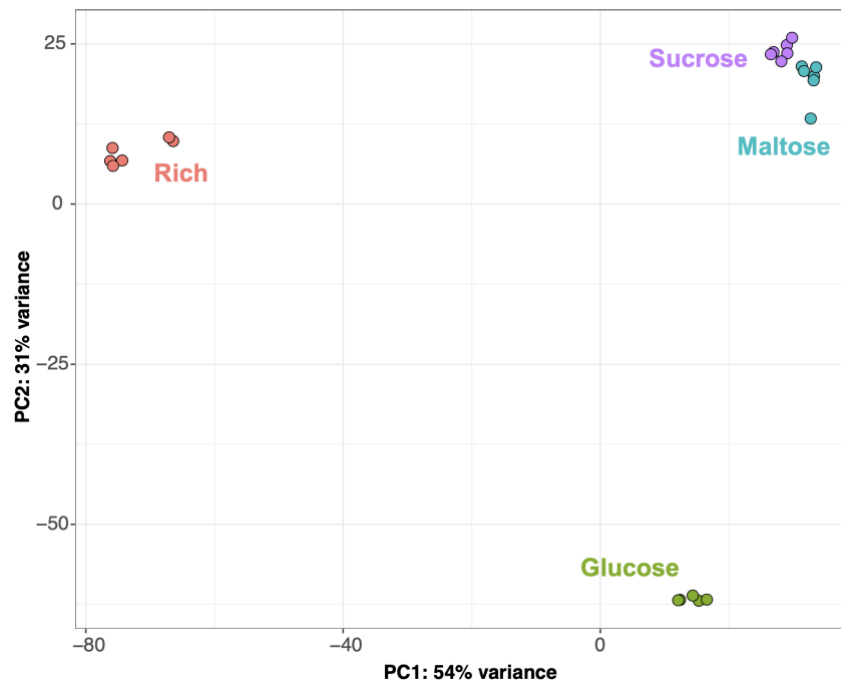


Figure 2.2. Principal components analysis (PCA) of the *H. borinquense* RNA-Seq dataset (n=5 for the glucose condition [green], n=6 for the sucrose [purple], maltose [blue], and rich [pink] conditions).

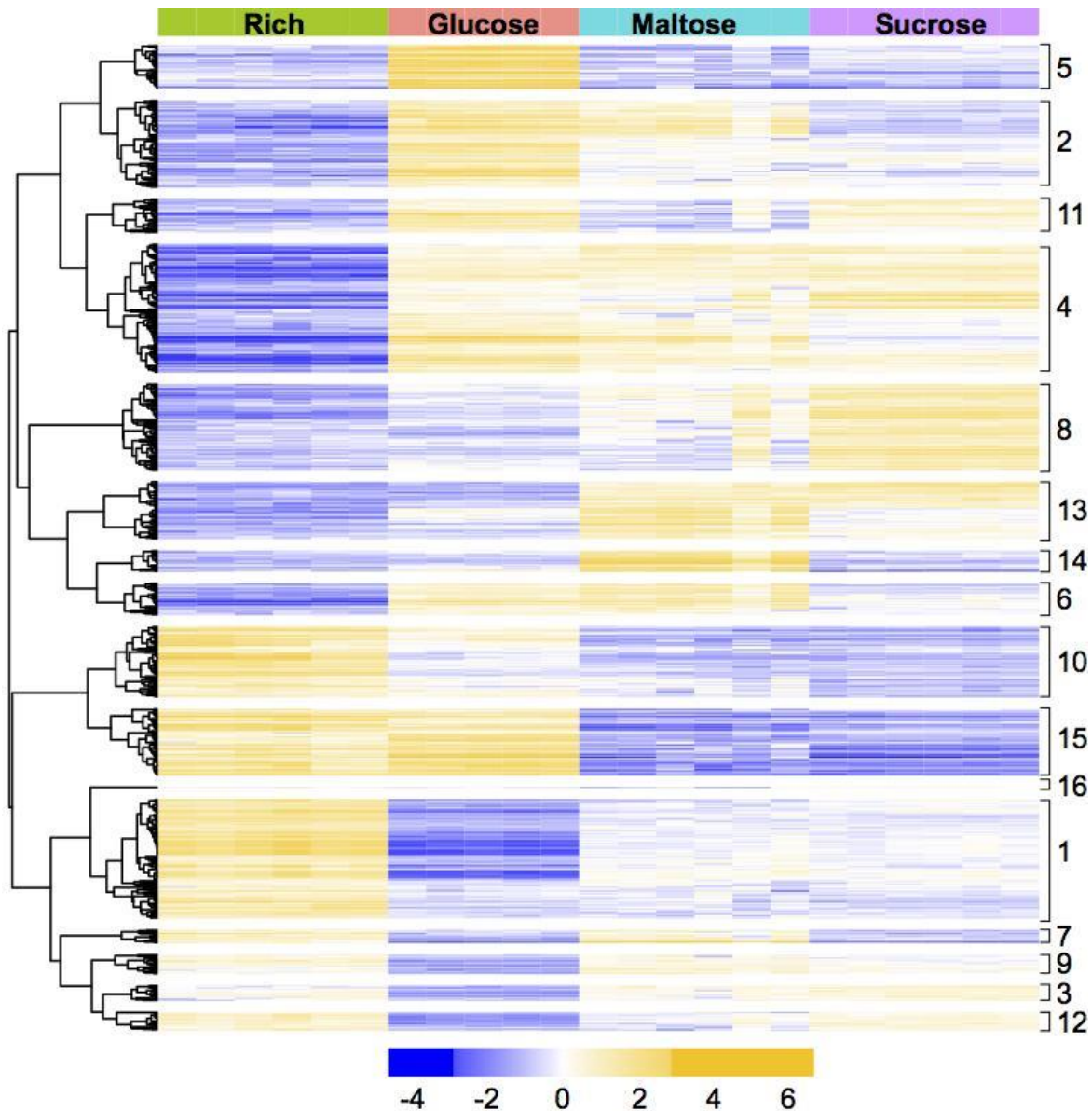


Figure 2.3. Heatmap of the *H. borinquense* RNA-Seq datasets for rich, glucose, maltose, and sucrose conditions, filtered to include significantly expressed genes with a ± 2 log₂ fold change, using the “complete” clustering algorithm to form 16 clusters (n=794 genes total).

The PCA plot also indicates that the patterns of mRNA abundance from cells growing on the disaccharides (S and M) are more similar to each other than they are to the R and M transcriptomes. Despite clustering tightly together in the PCA plot, the S and M conditions have at least 275 out of 3,898 total annotated genes whose mRNA abundance are significantly different between them (Figure 2.3, Appendix 2.6). The following sections summarize a

functional analysis of the differentially expressed genes, focusing primarily on those most directly involved in carbohydrate uptake and utilization, such as transporter proteins, glycosyl hydrolases & transferases, those encoding enzymes of glycolysis, gluconeogenesis, the TCA cycle and the glyoxylate pathway, nitrogen metabolism, and other functional categories (Figure 2.4).

2.5.3.1 ABC-type Transporter Proteins

In haloarchaea, carbohydrate uptake is known to occur through the involvement of ABC-type transporter systems (Koning, 2002). Table 2.1 lists the genes identified as ABC-type sugar transporters in *H. borinquense* with their KEGG annotations and RNA-Seq log2 fold change values normalized to the rich media condition. The mRNA abundance of the ABC-type sugar transport system encoded by genes *Hbor_11470-11500* are elevated in each of the three sugar conditions (G, M, and S) when compared to growth in rich media. While this operon was previously implicated in sucrose transport by Williams (2019), our data suggest that its gene products may also be involved in the transport of other sugars. By contrast, other ABC-type sugar transport operons seem to show a preference for specific sugars. For example, the operon *Hbor_30950-30980* appears to be uniquely differentially activated by glucose. The operon containing genes *Hbor_14210-12240*, and a cluster of genes including operon *Hbor_26470-26490* and gene *Hbor_26460* are most activated in maltose. Finally, the operon encoded by *Hbor_33370-33410* shows slight activation in the S condition, with an overall repression seen in the G condition.

Table 2.1. *H. borinquense* genes identified as ABC-type sugar transporters with their KEGG annotations and RNA-Seq log2 fold change values normalized to the rich media condition.

Name (Hbor)	Annotation	G/R	S/R	M/R
11470	AgIK / MalK / MsiK / MsmK / MsmX / SmoK, multiple sugar transport system ATP-binding protein	4.80	6.45	4.33
11480	ThuG, trehalose/maltose transport system permease protein	6.08	7.69	5.51
11490	ThuF, trehalose/maltose transport system permease protein	8.76	10.41	8.28
11500	ThuE, trehalose/maltose transport system substrate-binding protein	5.69	7.45	5.51
14210	AgIK / MalK / MsiK / MsmK / MsmX / SmoK, multiple sugar transport system ATP-binding protein	0.45	-0.24	2.07
14220	GtsC, glucose/mannose transport system permease protein	0.55	-0.18	2.46
14230	GtsB, glucose/mannose transport system permease protein	0.77	0.19	2.79
14240	GtsA, glucose/mannose transport system substrate-binding protein	0.67	0.82	2.95
26460	AgIK / MalK / MsiK / MsmK / MsmX / SmoK, multiple sugar transport system ATP-binding protein	1.50	2.58	7.70
26470	GanO, arabinogalactan oligomer / maltooligosaccharide transport system substrate-binding protein	-0.05	0.04	8.57
26480	GanP, arabinogalactan oligomer / maltooligosaccharide transport system permease protein	-0.19	-0.66	7.86
26490	GanQ, arabinogalactan oligomer / maltooligosaccharide transport system permease protein	0.04	-0.84	7.51
30950	UgpB, sn-glycerol 3-phosphate transport system substrate-binding protein	5.35	-0.92	-1.06
30960	UgpA, sn-glycerol 3-phosphate transport system permease protein	6.40	-0.05	0.05
30970	UgpE, sn-glycerol 3-phosphate transport system permease protein	6.52	0.10	-0.13
30980	UgpC, sn-glycerol 3-phosphate transport system ATP-binding protein	3.91	-0.24	-0.29
33370	hypothetical protein	-1.68	0.02	-0.33
33380	ABC-type sugar transport system, permease component	-2.69	0.22	-0.05
33390	permease component of ABC-type sugar transporter	-1.68	1.65	0.60
33400	ABC-type sugar transport system, periplasmic component	-2.24	1.70	1.10
33410	AgIK / MalK / MsiK / MsmK / MsmX / SmoK, multiple sugar transport system ATP-binding protein	0.83	3.73	2.70

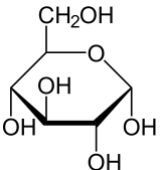
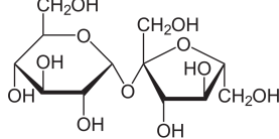
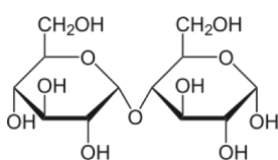
Appendix 2.7 lists the log2 fold change values for ABC-type transporter proteins unrelated to carbohydrates, including those associated with the transport of phosphonate,

phosphate, branched chain amino acids (BCAA), tungstate, alkanesulfonate, iron (III), and spermidine/putrescine, as well as proteins unrelated to transport involved in copper processing for nitrite reductase. The proteins annotated for BCAA and spermidine/putrescine transport all showed increased mRNA abundance in all sugar conditions when compared to their values in rich media. In the G condition, phosphate transport proteins showed increased mRNA abundance. Lower mRNA abundance was seen for other genes in the G, S and M conditions when compared to R. This includes copper processing genes, and phosphonate, tungstate, alkanesulfonate, iron (III) transport genes. Glucose had a more distinct transcriptional phenotype when compared to the other two sugars, which tended to have similar patterns of activation and repression for the non-carbohydrate ABC-type transporter proteins.

2.5.3.2 Glycosyl Hydrolases And Glycosyl Transferases

Glycosyl hydrolases (GH) are a broad family of proteins that catalyze the hydrolysis of glycosidic linkages. Glycosyl transferases, by contrast, form new glycosidic linkages by catalyzing a reaction in which a sugar moiety is added to an acceptor molecule. These classes of enzymes are thought to be generally important in the first steps of sugar utilization. In this study, we investigated the disaccharides sucrose, in which an α -1,2-glycosidic bond connects a glucose and fructose unit, and maltose, a disaccharide of glucose units connected by an α -1,4-glycosidic bond (Table 2.2).

Table 2.2. Chemical structures of the carbohydrates used in this study: glucose, sucrose, and maltose.

Carbohydrate	Glucose	Sucrose	Maltose
Molecular structure			
Bond	N/A	α -1,2-glycosidic bond	α -1,4-glycosidic bond
Subunits	N/A	Glucose + Fructose	Glucose + Glucose

Hydrolases with broad activity, and sucrose- and maltose-specific hydrolases have been identified in other organisms (Kim, 2004; Nichols, 1998). Table 2.3 lists *H. borinquense* genes with predicted glycosyl hydrolase or transferase activity and log2 fold change values for each sugar condition normalized to the rich media condition with a minimum threshold value of 1.5.

Table 2.3. *H. borinquense* genes identified with glycosyl hydrolase or transferase activity with their annotations and RNA-Seq log2 fold change values normalized to the rich media condition, with a minimum threshold value of +/- 1.5.

Name (Hbor)	Annotation	G/R	S/R	M/R
18930	(1->4)-alpha-D-glucan synthase (ADP-glucose)	2.80	1.32	3.10
33610	beta-glucosidase-like glycosyl hydrolase	3.25	4.24	3.85
35560	chitinase	0.81	4.49	4.21
31570	colanic acid/amylovoran biosynthesis glycosyltransferase	0.55	1.98	1.36
26450	glucoamylase	0.79	0.44	5.25
26440	glycosidase	0.84	0.63	4.23
14970	glycosyl hydrolase, glucoamylase	1.53	1.52	0.89
04610	glycosyl transferase 2 family protein	-2.17	-1.76	-0.54
05690	glycosyl transferase	-1.12	-1.35	-1.44
08250	glycosyl transferase	-1.96	-1.61	-0.97
17060	glycosyl transferase	-0.68	-0.21	1.00
23930	glycosyl transferase	-1.06	-1.63	-1.36
31440	glycosyl transferase	2.36	1.61	-0.03
31600	glycosyl transferase	1.67	0.49	0.31
36240	glycosyl transferase	-0.13	0.12	-1.60
36680	glycosyl transferase	0.83	2.22	1.66
15070	Reversibly glycosylated polypeptide	-2.81	-0.07	-0.30
35590	uncharacterized protein contain chitin-binding domain type 3	0.25	3.47	5.89

While several hydrolytic enzymes respond more specifically to the presence of maltose (*Hbor_26440-26450*), there do not appear to be any specifically responsive to sucrose at the transcript level. Through comparative genomics, Williams (2019) identified *Hbor_26450* as a potential sucrose hydrolase. However, its mRNA abundance profile (high in maltose media and low in other conditions, including those with sucrose) suggest that the production of this enzyme

may not be directly stimulated by sucrose and that its functional role may be more closely linked to its annotation as a glucoamylase. Though glucose is a monomeric saccharide, and sucrose and maltose are disaccharides with alpha linkages, a beta-glucosidase-like glycosyl hydrolase (*Hbor_33610*) and chitinase (*Hbor_35560*) were upregulated in all sugar conditions. Beta-glucosidase activity was also previously reported in cell lysates of *H. borinquense* grown in alpha-linked disaccharides (Cuebas-Irizarry, 2017). A similar observation has been made in cultures of *Sulfolobus solfataricus* (Haseltine, 1999a).

2.5.3.3 Central Carbon Metabolism And Gluconeogenesis

Similar to other haloarchaea, *H. borinquense* has been characterized to utilize the modified semi-phosphorylative Entner-Doudoroff (ED) and Embden-Meyerhof-Parnas (EMP) pathways of glycolysis to metabolize carbohydrates. However, in their analysis of sucrose metabolism in haloarchaea, Williams (2019) noted the absence of EMP genes, ketohexokinase and 1-phosphofructokinase, pointing to alternative pathways for fructose metabolism. We identify one fructokinase gene that may be involved in haloarchaeal fructose metabolism (*Hbor_20400*).

Table 2.4 lists the log2 fold change expression levels between the three sugar conditions relative to rich media for genes involved in glycolysis and gluconeogenesis. Genes encoding glycolytic enzymes are generally upregulated while accordingly, genes encoding gluconeogenic enzymes were generally downregulated in the glucose and sucrose conditions, relative to the maltose and rich conditions. Despite seeing strong activation of ABC transporter proteins in the maltose condition, most of the glycolysis-related genes are not as strongly upregulated when compared to their mRNA abundance in glucose and sucrose conditions.

Table 2.4. *H. borinquense* genes for glycolysis and gluconeogenesis with their annotations and RNA-Seq log2 fold change values normalized to the rich media condition.

Glycolysis					
Name (Hbor)	Annotation from Williams (2019)	Reaction	G/R	S/R	M/R
11690	GFO domain (halTADL1905 protein homolog)	glucose ↔ fructose	1.29	1.63	0.70
11700	XI/TM domain (halTADL1905 protein homolog)	glucose ↔ fructose	3.56	4.15	3.43
20400	fructokinase	ATP + D-fructose ↔ ADP + D-fructose 6-phosphate	0.89	1.55	0.80
22720	glucose dehydrogenase	glucose → gluconolactone → gluconate	1.60	1.96	0.18
17420	gluconate dehydratase	gluconate → KDG	-0.20	1.06	-0.58
26610	KDG kinase (3-deoxy-2-oxo-D-gluconate kinase)	KDG → KDPG	0.58	1.99	0.83
21830	KDPG aldolase (2-dehydro-3-deoxyaldolase phosphogluconate)	KDPG → GAP + pyruvate	0.26	1.70	0.31
17380	fructose-bisphosphate aldolase, class II	fructose-1,6P ₂ ↔ glyceraldehyde-3P or glycerone-P	-0.90	0.23	-1.41
19250	triosephosphate isomerase (TIM)	glycerone-P ↔ glyceraldehyde-3P	1.36	0.45	-0.13
27020	glyceraldehyde 3-phosphate dehydrogenase (phosphorylating)	glyceraldehyde-3 P → glycerate-1,3P ₂	2.53	1.41	-0.10
27010	phosphoglycerate kinase	glycerate-1,3P ₂ ↔ glycerate-3P	0.67	0.64	-0.78
08990	2,3-bisphosphoglycerate-independent phosphoglycerate mutase	glycerate-3P ↔ glycerate-2P	0.95	1.87	0.96
12470	enolase	glycerate-2P ↔ phosphoenolpyruvate	-2.37	-1.18	-1.38
24110	pyruvate kinase	phosphoenolpyruvate → pyruvate	0.22	0.28	-0.26
Gluconeogenesis					
Name (Hbor)	Annotation from Williams (2019)	Reaction	G/R	S/R	M/R
24040	phosphoenolpyruvate synthase	phosphoenolpyruvate ↔ pyruvate	-1.22	-1.75	-0.27
34530	phosphoenolpyruvate synthase	phosphoenolpyruvate ↔ pyruvate	2.55	0.90	1.82
17930	fructose-1,6-bisphosphatase I	fructose-6P ← fructose-1,6P ₂	-0.51	-1.33	-1.07
26990	glyceraldehyde-3-phosphate dehydrogenase (NAD(P))	glyceraldehyde-3 P ← glycerate-1,3P ₂	-2.53	-2.28	-0.07

2.5.3.4 TCA Cycle And Glyoxylate Pathway

The TCA cycle is conserved in haloarchaea and is used to produce reductants, anaplerotic and gluconeogenic intermediates (e.g., oxaloacetate, 2-oxoglutarate), and to harvest energy in conjunction with the respiratory chain. In haloarchaea, two anaplerotic pathways have been identified, the glyoxylate and the methylaspartate pathways (Borjian, 2016). Unlike the glyoxylate pathway, genes encoding the enzymes of the methylaspartate pathway, (e.g., glutamate mutase), have not been identified in *H. borinquense*.

Table 2.5 shows the genes encoding enzymes within the TCA Cycle and glyoxylate pathway with log2 fold change values greater than 1.5 between the sugar and rich conditions; many of the TCA cycle pathway genes did not exhibit strong increases or decreases of mRNA abundance (Appendix 2.8). Genes in the glyoxylate pathway, however, increased their mRNA abundance in all three sugar conditions relative to rich media, with the highest activation observed in the sucrose condition. As Beevers (1980) notes, this may be done to balance the carbohydrate-derived carbon units between biosynthesis and oxidation for energy harvesting.

Table 2.5. *H. borinquense* genes identified with TCA cycle or glyoxylate pathway activity with their annotations and RNA-Seq log2 fold change values normalized to the rich media condition, with a minimum threshold value of +/- 1.5.

Name (Hbor)	Pathway	Annotation	Reaction	G/R	S/R	M/R
27300	TCA Cycle	citrate synthase	acetyl-CoA + oxaloacetate ↔ citrate	-1.39	1.79	0.54
34280		citrate synthase	acetyl-CoA + oxaloacetate ↔ citrate	-1.45	2.72	1.50
32160		aconitate hydratase	citrate ↔ cis-aconitate; cis-aconitate ↔ isocitrate	-1.03	-4.23	-5.17
23330		2-oxoglutarate/2-oxoacid ferredoxin oxidoreductase subunit alpha	2-oxoglutarate ↔ succinyl-CoA	-1.94	0.28	0.26
05790	Glyoxylate Pathway	isocitrate lyase	isocitrate → glyoxylate and succinate	3.16	-0.17	4.32
23460		methylglyoxal synthase	glyoxylate and acetyl coa → malate	2.74	3.17	3.10
30530		bifunctional (S)-methyl-CoA lyase/thioesterase	glyoxylate and acetyl coa → malate	1.52	3.11	1.29

2.5.3.5 Nitrogen Metabolism

The nitrogen cycle allows many organisms, including haloarchaea, to use nitrogen compounds as a source for growth and to generate ammonium that is incorporated into the carbon skeleton of amino acids. Table 2.6 lists genes related to nitrogen metabolism with log2 fold change values greater than 1.5 between the sugar and rich conditions. The sugar conditions, especially sucrose and maltose, show repression of denitrification genes (e.g., NirK, NorBC) and activation of the gene expression patterns that favor ammonium and glutamate production (e.g., NirA and aspartate/tyrosine/aromatic aminotransferases). Several of the genes related to nitrogen metabolism had the highest log2 fold change values among the dataset of differentially expressed genes, including genes in an operon encoding ammonia permease (*amtB*) and nitrogen regulatory PII protein (*Hbor_01360* and *Hbor_01370*, respectively), both highly expressed in the glucose condition.

The nitrogen regulatory PII proteins are putative signal transduction molecules with binding sites for ATP/ADP and 2-oxoglutarate, classified into three major subgroups: glnB genes targeting glnA (the glutamine synthetase, or GS, structural gene) or nadE (NAD synthetase); glnK linked to *amtB*; or nifl genes known to be associated with genes annotated as structural components of nitrogenase (*nifH*, *nifD*, *nifK*). PII proteins are conserved in haloarchaea and it is thought that the primary mechanism, one of the most highly conserved features of nitrogen control, is through glnK linked to *amtB*, to control the flux of ammonia and control the intracellular nitrogen status through the ammonia channel transporter protein (Cheah, 1994; Sant'Anna, 2009; Thomas, 2000; Pedro-Roig, 2011, 2013). Previous studies show also that the GS of *Haloferax mediterranei* is activated *in vitro* by association with GlnK (Pedro-Roig, 2011, 2013; Arcondéguy, 2001).

When ammonia is present in excess, glutamate dehydrogenase (GDH; *Hbor_17850-17860*) can play a pivotal role in amino acid biosynthesis and carbon/nitrogen flux by catalyzing the reductive amination of 2-oxoglutarate to synthesize one glutamate. By contrast, under

conditions of abundant carbon availability or nitrogen limitation, the high activity glutamate synthase (GOGAT) offers an alternative glutamate biosynthesis reaction to allow for ammonia to assimilate into two glutamates. Finally, GS converts glutamate to glutamine. In this way, the GS/GOGAT pathway helps control the extent to which ammonia is assimilated and contributes to cell growth (Martín-Figueroa, 2000). Both glutamate and glutamine thereafter serve as nitrogen donors to produce other intracellular nitrogen-containing compounds (Huergo, 2015).

The results from this study show downregulation of GDH (*Hbor_17850-17860*) with upregulation of the GS (*Hbor_29510*) and GOGAT (*Hbor_23510*) in sugar conditions, relative to the rich media condition, indicating a likely state of nitrogen limitation and the possibility of active amino acid biosynthesis (Cheah, 1994). While this pattern of gene regulation is consistent with typically observed patterns seen under ammonia limitation in other systems, it is worth noting that to date *H. borinquense* has yet to be successfully cultured using inorganic nitrogen sources, raising the question of what might be responsible for the apparently similarity in gene expression behavior. Appendix 2.9 lists RNA-Seq log2 fold change values for genes involved in the biosynthesis of other amino acids. As expected, given the relative scarcity of amino acids in the sugar conditions, the mRNA profiles of these genes are generally elevated in all three sugars compared to rich media.

2.5.3.6 Other Functional Categories

The RNA-Seq log2 fold change values for other functional categories observed can be found in Appendix 2.10. In summary, relative to the rich media condition: Pentose phosphate genes are upregulated in the S condition, genes associated with chemotaxis genes are upregulated in the S and M conditions, gas vesicle protein genes are upregulated in the G and M conditions, biotin synthesis genes are upregulated in all sugar conditions, and carotenoid production genes are downregulated in the S condition.

Table 2.6. *H. borinquense* genes identified with nitrogen metabolism-related activity with their annotations and RNA-Seq log2 fold change values normalized to the rich media condition, with a minimum threshold value of +/- 1.5.

Name (Hbor)	Annotation	Function / Reaction	G/R	S/R	M/R
04420	NirA, ferredoxin-nitrite reductase	nitrite → ammonia	0.31	4.07	4.20
07880	carbamoyl-phosphate synthase large subunit	arginine and pyrimidine metabolism	3.49	2.14	2.01
09060	carbamoyl-phosphate synthase small subunit	arginine and pyrimidine metabolism	4.71	3.94	4.10
01360	AmtB, ammonia permease	ammonium transport	10.75	-0.40	1.53
01370	Nitrogen regulatory protein PII	regulates the activity of glutamine synthetase	11.07	0.21	0.31
11880	amidophosphoribosyltransferase	purine metabolism	1.47	1.13	2.08
16150	aspartate/tyrosine/aromatic aminotransferase	2-oxoglutarate → glutamate	2.43	3.41	2.41
17850	glutamate dehydrogenase (GDH)	L-glutamate + H ₂ O + NAD(P) ⁺ ↔ 2-oxoglutarate + NH ₃ + NAD(P)H + H ⁺	-1.37	-1.64	-1.18
17860	glutamate dehydrogenase (GDH)	L-glutamate + H ₂ O + NAD(P) ⁺ ↔ 2-oxoglutarate + NH ₃ + NAD(P)H + H ⁺	-2.26	-3.03	-1.62
21520	aspartate/tyrosine/aromatic aminotransferase	2-oxoglutarate → glutamate	0.41	2.07	0.89
23510	glutamate synthase family protein (GOGAT)	2 L-glutamate + 2 oxidized ferredoxin ↔ L-glutamine + 2-oxoglutarate + 2 reduced ferredoxin + 2 H ⁺	7.63	6.01	6.20
29510	glutamine synthetase (GS)	ATP + L-glutamate + NH ₃ ↔ ADP + phosphate + L-glutamine	3.73	2.67	2.02
30430	NosZ, nitrous oxide reductase apoprotein	nitrous oxide → nitrogen	-4.69	-8.16	-8.42
34290	NirK, dissimilatory nitrite reductase (NO-forming), copper type apoprotein	nitrite → nitric oxide	-1.00	-1.95	-2.74
34300	NorBC, nitric oxide reductase, NorZ apoprotein	nitric oxide → nitrous oxide	-2.28	-5.24	-6.22
35040	NarGHI or NxrAB, respiratory nitrate reductase alpha subunit apoprotein	nitrate ↔ nitrite	0.72	-5.27	-4.69
35050	NarGHI or NxrAB, respiratory nitrate reductase beta subunit	nitrate ↔ nitrite	0.80	-4.61	-4.30
35280	aspartate/tyrosine/aromatic aminotransferase	2-oxoglutarate → glutamate	-0.26	2.56	0.01
36540	glutamine---fructose-6-phosphate transaminase	amino acid and nucleotide sugar metabolism	5.58	3.94	3.34
39920	NirA, ferredoxin-nitrite reductase	nitrite → ammonia	2.86	4.37	5.59

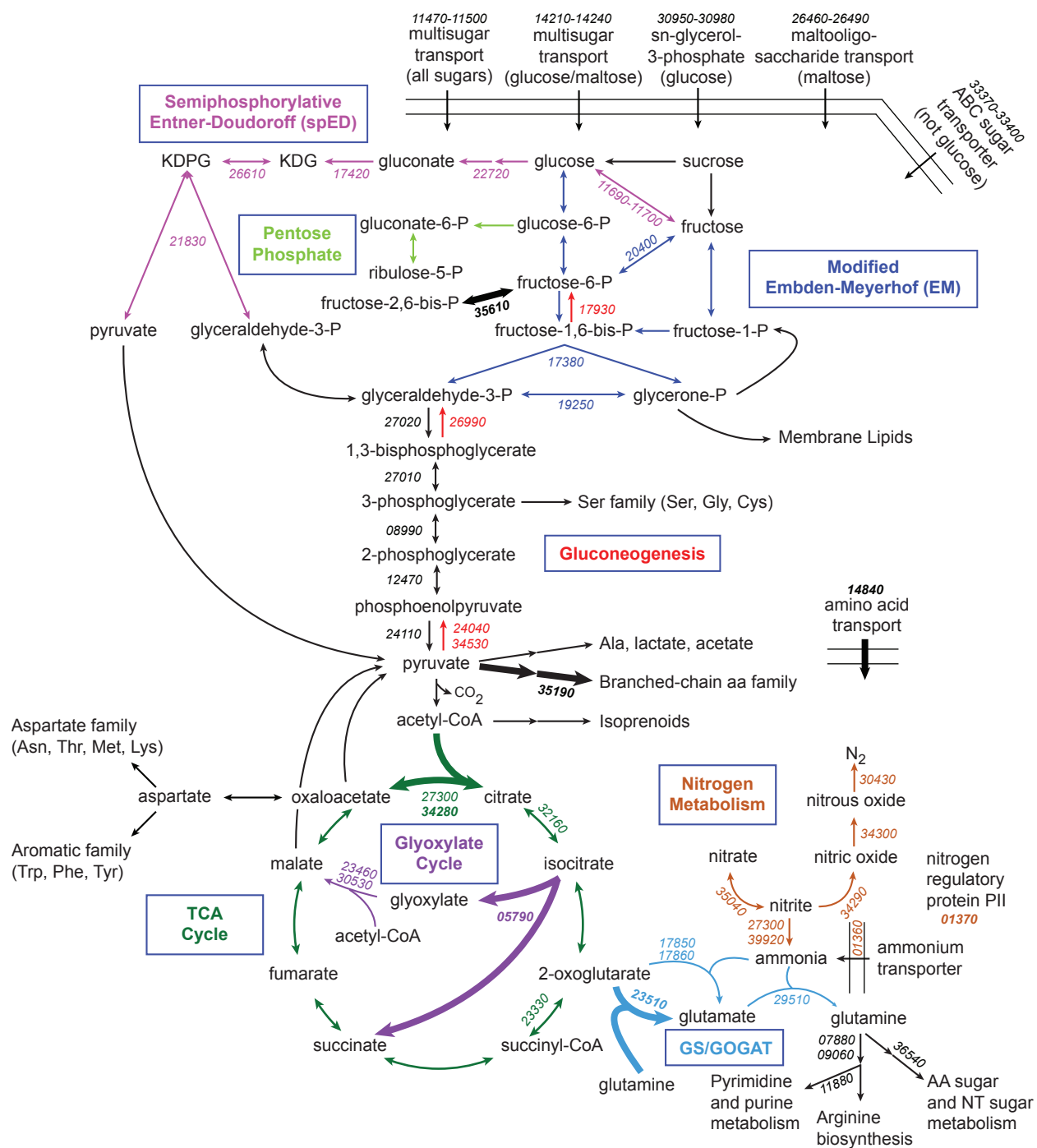


Figure 2.4. Central carbon and nitrogen metabolic pathways in *H. borinquense*. Genes identified as differentially regulated in the RNA-Seq data (tables above) are identified with Hbor numbers. Bolded text and thick arrows indicate genes probed as targets in the qRT-PCR section (2.5.3.).

2.5.3.7 Functional Coordination Of Co-Regulated Gene Expression

We examined co-regulation of expression of 794 *H. borinquense* genes that showed a minimum value of +/- 2-fold change, categorized within 16 clusters of phenotypes, as described in Methods (Figure 2.3; data in Appendix 2.11). From this analysis, we found that the rich condition is associated with higher abundances of genes related to protein synthesis: translation initiation factors (Cluster 1), RNA polymerase (Cluster 12), and other ribosomal proteins (Clusters 1 and 12). On the other hand, and as can be expected in a nitrogen-rich source like rich media, amino acid synthesis related proteins showed reduced expression, relative to the sugar conditions. Among the sugar conditions, Clusters 2 and 8 showed carbohydrate-specific expression of sets of amino acid transport and synthesis proteins. Cluster 2, which was most highly expressed in the glucose condition (then maltose, sucrose, and rich, in decreasing order), included genes for ABC-type BCAA transport systems, glutamine synthetase, and tryptophan synthase proteins. Cluster 8, most highly expressed in the sucrose condition, followed by maltose, glucose, and rich, included genes for ABC-type dipeptide/oligopeptide/nickel transport systems, acetolactate synthase, and aspartate/tyrosine/aromatic aminotransferase, which were co-regulated with fructose 2,6-bisphosphatase and TCA cycle genes (e.g., citrate synthase, isocitrate dehydrogenase). Finally, we observed 11 transcription factors and regulators distributed among the 16 clusters, several which show differential expression between the sugar conditions (e.g., *Hbor_24220*, *Hbor_37620*, *Hbor_35420*) (Table 2.7, 2.8).

Table 2.7. Transcription factors from the clusters in Figure 2.3, the *H. borinquense* RNA-Seq dataset filtered to include significantly expressed genes with a +/- 2 log2-fold change, using the “complete” clustering algorithm (n=794 genes).

Cluster	Name (Hbor)	Annotation	G/R	S/R	M/R
3	24220	transcription initiation factor TFIIIB	-3.21	0.07	-0.21
4	03590	Lrp/AsnC family transcriptional regulator, leucine-responsive regulatory protein	3.80	2.43	2.98
4	30180	transcriptional regulator	4.76	6.32	5.35
8	13700	Lrp/AsnC family transcriptional regulator, regulator for <i>asnA</i> , <i>asnC</i> and <i>gidA</i>	2.02	3.41	2.05
8	37620	transcriptional regulator	-0.73	2.94	1.06
9	19580	HTH-type transcriptional regulator, sugar sensing transcriptional regulator	-3.35	-1.52	-0.54
10	07500	transcriptional regulator	-1.04	-2.97	-2.19
10	34750	transcriptional repressor NrdR	-1.81	-5.81	-5.39
11	11430	HTH-type transcriptional regulator, sugar sensing transcriptional regulator	-3.94	0.33	-1.43
13	28460	transcriptional regulator of a riboflavin/FAD biosynthetic operon	0.69	1.82	2.68
13	35420	transcriptional regulator	0.26	4.20	3.53

2.5.3.8 RNA-Seq Summary

Altogether, the RNA-Seq results provide a snapshot of the transcriptional activity of *H. borinquense* at a point in late log phase where the phenotypes of different media conditions diverged. These data paint a broad picture of cells that in all sugar conditions are responding physiologically to similar stresses (e.g. limited nitrogen and carbon source availability) by activating high-affinity nitrogen scavenging systems (GS/GOGAT) and the anaplerotic glyoxylate anaplerotic cycle to shuffle carbon molecules into biosynthetic pathways. Despite common responses to environmental pressures, *H. borinquense* also mounts sugar-specific (either unique to a single sugar or shared between pairs of sugars) responses at the transcriptional level, involving numerous different functional categories, starting with sugar-

specific transport and processing systems to “accessory” functions involved in motility and pigment production.

2.5.4 Identifying Putative Glucose-Influenced Transcriptional Changes

The discovery of sugar-specific responses in the transcriptomic data suggested that it might be possible to identify genes whose regulation is influenced directly or indirectly by specific sugars and provides hints into regulatory cross-talk between the carbohydrates (Figure 2.4). In this work, we were particularly interested in finding putative targets of CCR. We used the data above to select specific genes that could be putative targets of glucose mediated regulation and sought to find additional evidence of glucose’s regulatory influence by qRT-PCR. Targets that showed differential sugar-specific expression in the RNA-Seq dataset and that had potential key roles in carbon and/or nitrogen metabolism were selected. The seven qRT-PCR targets included: nitrogen regulatory protein PII (*Hbor_01370*), isocitrate lyase (*Hbor_05790*), amino acid/amide ABC transporter (*Hbor_14840*), glutamate synthase (*Hbor_23510*), citrate synthase (*Hbor_34280*), thiamine pyrophosphate-dependent acetolactate synthase (*Hbor_35190*), and a predicted fructose 2,6-bisphosphatase (*Hbor_35610*).

In order to study the regulatory effects of glucose on these targets, cultures of *H. borinquense* were grown as described in Methods. At t=44 hrs, a spike of glucose was added to S and M cultures, creating SaddG and MaddG. Control flasks containing glucose, maltose and sucrose only were also included, as was an SG condition flask. Aliquots were harvested from all flasks at timepoints before and after glucose addition. mRNA was harvested and qRT-PCR was performed as described in Methods at t=20 hrs (during early growth), t=40 hrs (slightly before the addition of glucose), and t=72 hrs (following the glucose addition) (Figure 2.5).

2.5.4.1 Glucose-Induced Activation Of Gene Expression

Nitrogen regulatory protein PII (predicted to be GlnK) (*Hbor_01370*)

The signal transduction molecule, nitrogen regulatory protein PII GlnK (*Hbor_01370*), is found in a highly conserved operon with ammonia permease, *amtB* (*Hbor_01360*). The two proteins act as nitrogen sensors and interact to control the flux of ammonia in and out of the cell. The regulatory activity of GlnK is known to be mediated by a number of factors and to target a variety of proteins related to nitrogen flux including key metabolic and regulatory enzymes, transcription factors, and nutrient transporters (Javelle, 2004; Radchenko, 2011; Huergo, 2013). In *Haloflex mediterranei*, the *amtB-GlnK* gene pair is minimally expressed in glucose-containing ammonia-rich conditions and more highly expressed during ammonia starvation (Pedro-Roig, 2011).

In our RNA-Seq dataset (Table 2.5), the transcript of *Hbor_01370* was highly abundant in the G condition, relative to its levels in S, M, and rich media conditions, suggesting either that glucose is an activating influence or that the other sugars are negative influences on the expression of *Hbor_01370*. The qRT-PCR data also shows a stark contrast between glucose-containing conditions and the non-glucose-containing conditions by t=40 hrs, with sucrose-only conditions showing a very large decrease in mRNA abundance between the first and second timepoints relative to levels in the glucose-only condition. The SG condition behaves like G, suggesting that the decreased expression in sucrose-only conditions is not likely due to repression by sucrose. The addition of glucose to a sucrose-only flask at t=44 hrs (SaddG) results in a dramatic increase in mRNA abundance by t=72 hrs when compared to the S condition, suggesting that glucose has activated the expression of *Hbor_01370*.

2.5.4.2 Glucose-Induced Repression Of Gene Expression

Citrate synthase (*Hbor_34280*)

Citrate synthase (*Hbor_34280*) catalyzes the first reaction of the TCA cycle, the condensation reaction of acetyl-CoA and oxaloacetate to form citrate. This enzyme is known to be inhibited by high ATP:ADP and NADH:NAD ratios, glutamate, and catabolic intermediates. Moreover, in *B. subtilis*, Jin (1994) reported that transcription of citrate synthase genes *citA* and *citZ* were induced as cells enter stationary phase and were repressed by excess glucose.

In our RNA-Seq dataset, expression of *Hbor_34280* was highest in the sucrose condition, followed by maltose, rich, and lastly, glucose, suggesting possible repression by glucose or activation by other sugars. Our qRT-PCR data confirms overexpression of *Hbor_34280* in sucrose-containing conditions, relative to the glucose-containing conditions in a growth phase-specific manner. Expression of *Hbor_34280* in the sucrose-containing flasks increases between the first and second timepoints while staying relatively low in the SG flask throughout the experiment. The addition of glucose at t=44 hrs to a sucrose-containing flask (SaddG) decreased expression of *Hbor_34280* by t=72 hrs compared to the S flask, suggesting a repressive influence of glucose which may be associated with CCR.

Isocitrate lyase (*Hbor_05790*)

Isocitrate lyase (ICL) is the first enzyme in the glyoxylate shunt, which bypasses two decarboxylation steps in the TCA cycle in favor of producing biosynthetic compounds in carbohydrate-limited conditions. Observations in several other organisms suggest that, at least in those systems, ICL is likely regulated by CCR-like mechanisms (Kretzschmar, 2008; Graham 1994). In the RNA-Seq dataset, mRNA abundance levels of *Hbor_05790* in the M and G conditions were relatively high compared to those in the S and R conditions. By contrast, the qRT-PCR data shows high relative mRNA abundance of *Hbor_05790* in sucrose-only containing flasks compared to glucose-containing conditions (G and SG) at t=20 hrs. There is a dramatic

decrease in mRNA abundance of *Hbor_05790* transcript in sucrose-only conditions by the next timepoint, and little change in the glucose-containing conditions. However, the spike of glucose in a sucrose-only flask (SaddG) triggers the largest decrease in *Hbor_05790* mRNA abundance over the rest of the experiment. Therefore, while the expression of *Hbor_05790* seems to have a strong dependence on growth phase, the abundance of glucose clearly also negatively influences the expression of the gene irrespective of growth phase, again suggesting that a possible underlying CCR-like component of regulation may contribute to the control of *Hbor_05790* expression.

Table 2.8. log2 fold change values normalized to the glucose (G) condition for the three timepoint series qRT-PCR (t1=20 hrs, t2 = 40 hrs, t3=72 hrs) in *H. borinquense*.

Name (<i>Hbor</i>)	Annotation	Timepoint	S to G	SaddG to G	SG to G
01370	nitrogen regulatory protein PII	1 (20 hrs)	-2.96	-5.51	1.28
		2 (40 hrs)	-1086.88	-1261.55	0.36
		3 (72 hrs)	-692.24	2.76	2.50
05790	isocitrate lyase	1	106.30	132.26	0.02
		2	1.52	1.41	-1.87
		3	-4.83	-14.71	-5.83
14840	amino acid/amide ABC transporter	1	3.68	2.82	-1.25
		2	1.18	0.14	0.45
		3	1.57	-0.04	-2.13
23510	glutamate synthase	1	-9.77	-9.45	-1.13
		2	-1.46	-1.54	-0.16
		3	-1.48	-0.06	1.33
34280	citrate synthase	1	2.42	2.05	1.15
		2	18.47	17.90	-0.11
		3	9.26	-1.65	0.01
35190	thiamine pyrophosphate-dependent acetolactate synthase	1	1.32	1.11	1.10
		2	19.36	17.51	0.30
		3	0.01	-4.95	-1.19
35610	predicted fructose 2,6-bisphosphatase	1	-9.16	-9.69	-1.12
		2	3.08	2.98	1.50
		3	6.47	2.64	2.92

Thiamine pyrophosphate-dependent acetolactate synthase (Hbor_35190)

Thiamine pyrophosphate-dependent enzyme acetolactate synthase (ALS) (Hbor_35190) catalyzes the first committed step in branched chain amino acid (BCAA) synthesis. Previous studies in numerous organisms have shown that the BCAA synthesis pathway is feedback inhibited by the amino acid end products (Singh, 1992) and we assume, due to the highly conserved nature of this regulatory scheme, that similar mechanisms must be active in *H. borinquense*. However, the role of sugars in the regulation of this key metabolic branchpoint appear to not be as well understood. In the RNA-Seq data, ALS expression appeared enhanced in the S condition, and to some smaller degree M, when compared to expression in G or R. In the qRT-PCR data, a clear growth-phase dependence of Hbor_35190 mRNA abundance is observed. At t=20 hrs, all conditions show similar expression. By t=40 hrs, sucrose-only conditions have dramatically increased mRNA abundance while the abundance of Hbor_35190 transcript seems unchanged in the SG condition. While abundance of Hbor_35190 transcript declines to near its initial measured levels by the final timepoint in S, the glucose spike in the SaddG condition further lowers abundance relative to all conditions, again indicating a potential for a CCR-related repressive role for glucose in the control of this gene.

2.5.4.3 Alternative Sugar- And Growth-Dependent Effects

Predicted fructose 2,6-bisphosphatase (Hbor_35610) and glutamate synthase (Hbor_23510)

The homodimeric, bifunctional enzyme, fructose-2,6-bisphosphatase (FBPase-2)/PFK-2 (6-phosphofructo-2-kinase), catalyzes the synthesis and degradation of signal molecule fructose 2,6-bisphosphate (Fru-2,6-P₂) from fructose-6-phosphate and ATP. Fructose-2,6-bisphosphate contributes to the regulation of central carbon metabolism by activating phosphofructokinase, the enzyme controlling the rate-limiting step of glycolysis and inhibiting the gluconeogenic enzyme fructose bisphosphatase. In the RNA-Seq dataset, this enzyme is most highly expressed in the sucrose condition, followed by glucose, then maltose, and lastly rich – this is

consistent with the observation that the other glycolytic genes generally are also the most highly expressed in the sucrose condition. The qRT-PCR data reveals that initially (at t=20 hrs), FBPase-2 mRNA abundance is low compared to glucose containing conditions suggesting either repression by sucrose or glucose activation early in the growth trajectory of the organism. However, after timepoint 1, expression in the sucrose-only conditions increased whereas the expression of *Hbor_35610* in the glucose-containing condition SG remained unchanged relative to its level in G. While mRNA abundance continues to increase in all conditions through the last timepoint, the SaddG and SG conditions have the lowest abundance. These data suggest that both sucrose and glucose may have condition-specific repressive effects on *Hbor_35610* expression.

As discussed previously, glutamate synthase (GOGAT) (*Hbor_23510*) functions together with glutamine synthase (GS) to fix inorganic nitrogen, typically in nitrogen starved conditions. This enzyme is an important metabolic link between carbon and nitrogen metabolism in the cell. In the RNA-Seq data, the expression of *Hbor_23510*, the gene encoding GOGAT was found to have elevated expression in all sugar conditions when compared to levels in rich medium, with the sucrose condition showing greatest abundance. Nevertheless, when sugar conditions are compared to one another alone in our targeted qRT-PCR experiment, the behavior of this gene shows similar trends to that displayed by FBPase-2 (*Hbor_35610*), suggesting growth- and sugar-dependent regulatory inputs that may likewise suggest a possible repressive role for glucose under appropriate environmental constraints.

2.5.4.4 Summary Of Targeted qRT-PCR Data

In sum, the targeted qRT-PCR experiments reveal two key observations. First, we note that, not unexpectedly, patterns of mRNA abundance for the selected genes can show significant growth-phase dependent changes in behavior. Progress through the various stages of growth in a flask is subject to numerous changes in nutrient availability and accumulation of

metabolic byproducts that create a dynamic and multidimensional landscape of environmental signals that the cell needs to simultaneously respond to. Second, these data clearly demonstrate sugar-specific influences on the growth phase-dependent profiles of expression of these key genes. More specifically, the data demonstrate that glucose can have both an activating and repressing influence on different target genes, with the latter suggesting that CCR mechanisms may be active in *H. borinquense*.

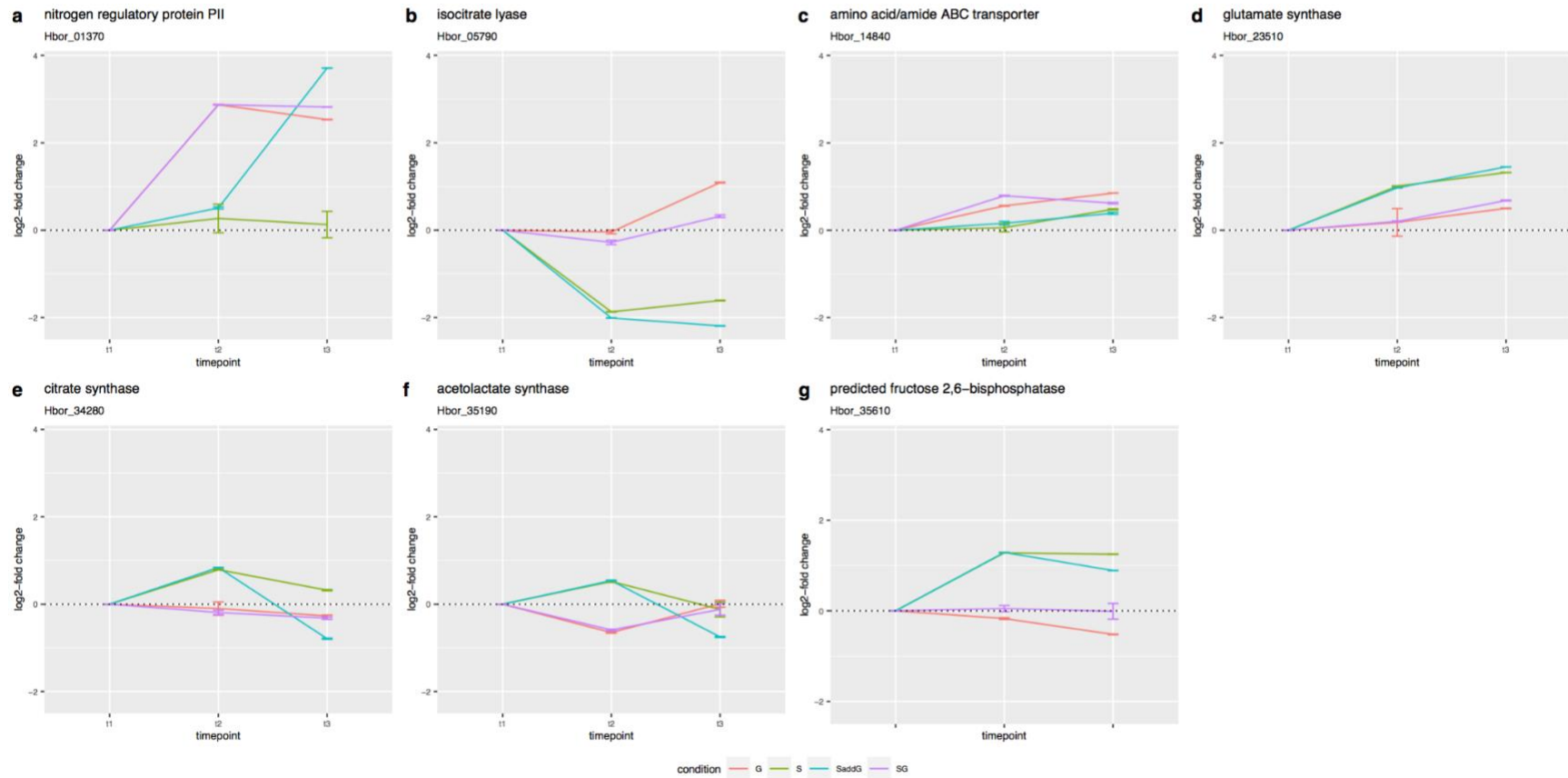


Figure 2.5. log₂ fold change values normalized to the first timepoint for the three timepoint series qRT-PCR (t1=20 hrs, t2=40 hrs, t3=72 hrs) for seven qRT-PCR targets: **a.** nitrogen regulatory protein PII (*Hbor_01370*), **b.** isocitrate lyase (*Hbor_05790*), **c.** amino acid/amide ABC transporter (*Hbor_14840*), **d.** glutamate synthase (*Hbor_23510*), **e.** citrate synthase (*Hbor_34280*), **f.** thiamine pyrophosphate-dependent enzyme (*Hbor_35190*), **g.** predicted fructose 2,6-bisphosphatase (*Hbor_35610*).

2.6 Discussion

This chapter details an exploration of differential carbohydrate utilization and CCR-like regulation in the haloarchaeon *Halogeometricum borinquense* through growth experiments on glucose, sucrose, maltose, and combinations thereof. We examined the behavior of these cultures using optical density, cell counting, & sugar utilization assays, RNA-Seq, and targeted qRT-PCR. Collectively, we observed unique, sugar-specific growth and gene expression phenotypes between the experimental conditions of Rich, Min, G, SG, S, SaddG, M, and MaddG (Figures 2.1-3).

Sugar utilization data reveals that the sugars are not consumed with the same kinetics, with glucose consumption beginning at an earlier timepoint and at a more rapid rate than the disaccharides sucrose and maltose, which are utilized in that order later during the growth experiment. Unsurprisingly, the Rich media condition seems to support the most robust cultures. Cell counts and sugar utilization data suggest that the organism transforms each sugar into growth with the following ranked preference/ability: glucose, sucrose, and maltose. Though sugars are utilized, growth in our experimental conditions seems largely driven and capped by the abundance of yeast extract. This observation is consistent with literature that suggests that most haloarchaea preferentially consume amino acids over carbohydrates (Andrei, 2012; Anderson, 2011). We suspect that sugar metabolism which is not directly linked to cell division nevertheless supports metabolic activity and the biosynthesis of cellular pigments, gas vesicles (Hechler, 2007), and, in the case of sucrose-containing media, an extracellular polymeric substance.

The RNA-Seq data provided a window of the transcriptome of *H. borinquense* grown in rich media, along with single sugar conditions G, S, and M, at late log phase, allowing for the identification of functional groups of genes that are both responsive to the general growth conditions of the carbohydrate-containing conditions and those responsive to specific sugars. First, relative to the rich condition, all three sugar conditions show lower expression of genes for

gluconeogenesis, and higher expression of genes for glycolysis, the glyoxylate pathway, and amino acid synthesis. Transport processes and key nodes in central carbon and nitrogen metabolism show some of the most dramatic sugar-specific differences in transcriptional profiles. Meanwhile, analysis of co-expression clusters also provides insight into patterns of potential co-regulation of differentially expressed genes, including the identification of a number of transcriptional regulators that may play a role in determining the observed patterns of transcriptional partitioning (Table 2.7, Appendix 2.11).

Transcriptomic data allowed us to identify genes that were influenced by specific sugars and select putative targets of glucose-mediated regulation for further study. To better understand how the behavior of specific genes centrally involved in carbon and nitrogen flux was influenced by the presence of glucose, we carried out a three-timepoint series targeted qRT-PCR experiments of cells grown in the G, S, SG, and SaddG conditions. These experiments revealed that the expression of target genes is typically growth phase-dependent and that glucose can act both as a positive and negative influence of target genes. Glucose showed an activating effect on expression for nitrogen regulatory protein PII (*Hbor_01370*) and a strong repressive influence on the abundance of citrate synthase (*Hbor_34280*), isocitrate lyase (*Hbor_05790*), and acetolactate synthase (*Hbor_35190*). Likely weaker, and more cell state-specific glucose repression was also observed for the genes encoding glutamate synthase (GOGAT) (*Hbor_23510*) and a predicted fructose 2,6-bisphosphatase (*Hbor_35610*). The protein products of each of these genes are either involved directly at key points in the control of carbon flux and nitrogen metabolism, or help in maintaining a balance between carbon and nitrogen recruitment and utilization. The observation of glucose-induced repression, while not revealing underlying mechanisms, is strongly suggestive of CCR for the identified targets. Further work will be required to unravel the mechanistic links between glucose and the expression of these genes.

This work, therefore, serves as a precursor for ultimately unraveling the mechanistic underpinnings of regulatory schemes involved in the control of carbon and nitrogen flux in the metabolically versatile *H. borinquense*. Future work could involve further delineation of the specific factors that, together with specific carbohydrate availability, contribute to the differential gene regulation observed in this study. This could include decomposing the currently complex organic nitrogen source into individual amino acids to isolate other key regulatory factors. The core goal of these studies would be to refine our understanding of the specific factors and conditions that drive reproducible and controllable patterns of gene expression. With conditions and factors defined, follow up studies to identify the regulators for the target genes in defined conditions could be carried out to elucidate their mechanisms of action. Finally, we expect that understanding the roles for multiple layers of regulation will be important for forming a complete picture of how *H. borinquense* regulates gene expression during growth on different carbohydrates. Future work could therefore also explore the potential regulatory roles of post-translational modification of key proteins and non-coding RNAs, the latter of which have been known to be involved in regulation for starvation, stress response, and other biological processes.

2.7 Funding

This work was supported by a grant by Howard Hughes Medical Institute (#52007566), and by the National Science Foundation (#1517797).

2.8 References

- Altekar, W., & Rangaswamy, V. (1992). Degradation of endogenous fructose during catabolism of sucrose and mannitol in halophilic archaebacteria. *Archives of microbiology*, 158(5), 356-363.
- Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *bioinformatics*, 31(2), 166-169.
- Anderson, I., Scheuner, C., Göker, M., Mavromatis, K., Hooper, S. D., Porat, I., ... & Kyrpides, N. (2011). Novel insights into the diversity of catabolic metabolism from ten haloarchaeal genomes. *PLoS One*, 6(5), e20237.
- Andrei, A. Ş., Banciu, H. L., & Oren, A. (2012). Living with salt: metabolic and phylogenetic diversity of archaea inhabiting saline ecosystems. *FEMS microbiology letters*, 330(1), 1-9.
- Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online. Retrieved May, 17, 2018.
- Arcondéguy, T., Jack, R., & Merrick, M. (2001). PII signal transduction proteins, pivotal players in microbial nitrogen control. *Microbiology and Molecular Biology Reviews*, 65(1), 80-105.
- Becker, E. A., Seitzer, P. M., Tritt, A., Larsen, D., Krusor, M., Yao, A. I., ... & Facciotti, M. T. (2014). Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. *PLoS genetics*, 10(11), e1004784.
- Beevers, H. (1980). The role of the glyoxylate cycle. In *Lipids: Structure and Function* (pp. 117-130). Academic Press.
- Berquist, B. R., Müller, J. A., & DasSarma, S. (2006). 27 genetic systems for halophilic archaea. *Methods in microbiology*, 35, 649-680.
- Bini, E., & Blum, P. (2001). Archaeal catabolite repression: a gene regulatory paradigm. *Advances in applied microbiology*, 50, 339-366.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120.
- Bonete, M. J., Perez-Pomares, F., Ferrer, J., & Camacho, M. L. (1996). NAD-glutamate dehydrogenase from *Halobacterium halobium*: inhibition and activation by TCA intermediates and amino acids. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1289(1), 14-24.
- Borjian, F., Han, J., Hou, J., Xiang, H., & Berg, I. A. (2016). The methylaspartate cycle in haloarchaea and its possible role in carbon metabolism. *The ISME journal*, 10(3), 546-557.
- Born, J., & Pfeifer, F. (2019). Improved GFP variants to study gene expression in haloarchaea. *Frontiers in microbiology*, 10, 1200.
- Bräsen, C., & Schönheit, P. (2004). Regulation of acetate and acetyl-CoA converting enzymes during growth on acetate and/or glucose in the halophilic archaeon *Haloarcula marismortui*. *FEMS microbiology letters*, 241(1), 21-26.
- Brückner, R., & Titgemeyer, F. (2002). Carbon catabolite repression in bacteria: choice of the carbon source and autoregulatory limitation of sugar utilization. *FEMS microbiology letters*, 209(2), 141-148.
- Cheah, E., Carr, P. D., Suffolk, P. M., Vasudevan, S. G., Dixon, N. E., & Ollis, D. L. (1994). Structure of the *Escherichia coli* signal transducing protein PII. *Structure*, 2(10), 981-990.
- Cuebas-Irizarry, M. F. (2017). Physiological and transcriptome-based profile of glycosyl hydrolases from the Halophilic archaeon *Halogeometricum borinquense* (Doctoral dissertation).

- Darnell, C. L., Tonner, P. D., Gulli, J. G., Schmidler, S. C., & Schmid, A. K. (2017). Systematic discovery of archaeal transcription factor functions in regulatory networks through quantitative phenotyping analysis. *MSystems*, 2(5), e00032-17.
- De Rosa, M., Gambacorta, A., & Bu'Lock, J. D. (1975). Extremely thermophilic acidophilic bacteria convergent with *Sulfolobus acidocaldarius*. *Microbiology*, 86(1), 156-164.
- De Rosa, M., Gambacorta, A., Nicolaus, B., Giardina, P., Poerio, E., & Buonocore, V. (1984). Glucose metabolism in the extreme thermoacidophilic archaebacterium *Sulfolobus solfataricus*. *Biochemical Journal*, 224(2), 407-414.
- Facciotti, M. T., Pang, W. L., Lo, F. Y., Whitehead, K., Koide, T., Masumura, K. I., ... & Baliga, N. S. (2010). Large scale physiological readjustment during growth enables rapid, comprehensive and inexpensive systems analysis. *BMC systems biology*, 4(1), 1-19.
- Görke, B., & Stülke, J. (2008). Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nature Reviews Microbiology*, 6(8), 613-624.
- Graham, I. A., Denby, K. J., & Leaver, C. J. (1994). Carbon catabolite repression regulates glyoxylate cycle gene expression in cucumber. *The Plant Cell*, 6(5), 761-772.
- Haseltine, C., Rolfsmeier, M., & Blum, P. (1996). The glucose effect and regulation of alpha-amylase synthesis in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Journal of bacteriology*, 178(4), 945-950.
- Haseltine, C., Montalvo-Rodriguez, R., Bini, E., Carl, A., & Blum, P. (1999a). Coordinate transcriptional control in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Journal of bacteriology*, 181(13), 3920-3927.
- Haseltine, C., Montalvo-Rodriguez, R., Carl, A., Bini, E., & Blum, P. (1999b). Extragenic pleiotropic mutations that repress glycosyl hydrolase expression in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Genetics*, 152(4), 1353-1361.
- Hechler, T., Frech, M., & Pfeifer, F. (2008). Glucose inhibits the formation of gas vesicles in *Haloferax volcanii* transformants. *Environmental microbiology*, 10(1), 20-30.
- Huergo, L. F., & Dixon, R. (2015). The emergence of 2-oxoglutarate as a master regulator metabolite. *Microbiology and Molecular Biology Reviews*, 79(4), 419-435.
- Javelle, A., Severi, E., Thornton, J., & Merrick, M. (2004). Ammonium sensing in *Escherichia coli*: role of the ammonium transporter AmtB and AmtB-GlnK complex formation. *Journal of Biological Chemistry*, 279(10), 8530-8538.
- Jin, S., & Sonenshein, A. L. (1994). Transcriptional regulation of *Bacillus subtilis* citrate synthase genes. *Journal of bacteriology*, 176(15), 4680-4690.
- Johnsen, U., Selig, M., Xavier, K. B., Santos, H., & Schönheit, P. (2001). Different glycolytic pathways for glucose and fructose in the halophilic archaeon *Halococcus saccharolyticus*. *Archives of microbiology*, 175(1), 52-61.
- Kim, H. S., Park, H. J., Heu, S., & Jung, J. (2004). Molecular and functional characterization of a unique sucrose hydrolase from *Xanthomonas axonopodis* pv. *glycines*. *Journal of bacteriology*, 186(2), 411-418.
- Koning, S. M., Albers, S. V., Konings, W. N., & Driessen, A. J. (2002). Sugar transport in (hyper) thermophilic archaea. *Research in microbiology*, 153(2), 61-67.
- Kretzschmar, U., Khodaverdi, V., Jeoung, J. H., & Görisch, H. (2008). Function and transcriptional regulation of the isocitrate lyase in *Pseudomonas aeruginosa*. *Archives of microbiology*, 190(2), 151-158.
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), 1-10.
- Leigh, J. A., Albers, S. V., Atomi, H., & Allers, T. (2011). Model organisms for genetics in the domain Archaea: methanogens, halophiles, Thermococcales and Sulfolobales. *FEMS microbiology reviews*, 35(4), 577-608.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12), 1-21.

- Lubelska, J. M., Jonuscheit, M., Schleper, C., Albers, S. V., & Driessen, A. J. (2006). Regulation of expression of the arabinose and glucose transporter genes in the thermophilic archaeon *Sulfolobus solfataricus*. *Extremophiles*, 10(5), 383-391.
- Machné, R., & Stadler, P. F. (2020). dptseg: Piecewise Linear Segmentation by Dynamic Programming. R Package Version 0.1, 2.
- Malfatti, S., Tindall, B. J., Schneider, S., Fährnich, R., Lapidus, A., LaButti, K., ... & Chain, P. (2009). Complete genome sequence of *Halogeometricum borinquense* type strain (PR3 T). *Standards in Genomic Sciences*, 1(2), 150-158.
- Martín-Figueroa, E., Navarro, F., & Florencio, F. J. (2000). The GS-GOGAT pathway is not operative in the heterocysts. Cloning and expression of glsF gene from the cyanobacterium *Anabaena* sp. PCC 7120. *FEBS letters*, 476(3), 282-286.
- Montalvo-Rodriguez, R., Vreeland, R. H., Oren, A., Kessel, M., Betancourt, C., & López-Garriga, J. U. A. N. (1998). *Halogeometricum borinquense* gen. nov., sp. nov., a novel halophilic archaeon from Puerto Rico. *International Journal of Systematic and Evolutionary Microbiology*, 48(4), 1305-1312.
- Morlan, J. D., Qu, K., & Sinicropi, D. V. (2012). Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue.
- Nichols, B. L., Eldering, J., Avery, S., Hahn, D., Quaroni, A., & Sterchi, E. (1998). Human small intestinal maltase-glucoamylase cDNA cloning: homology to sucrase-isomaltase. *Journal of Biological Chemistry*, 273(5), 3076-3081.
- Pedro-Roig, L., Camacho, M., & Bonete, M. J. (2011). In vitro proof of direct regulation of glutamine synthetase by GlnK proteins in the extreme halophilic archaeon *Haloferax mediterranei*. *Biochem Soc Trans*, 39(1):259-62.
- Pedro-Roig, L., Camacho, M., & Bonete, M. J. (2013). Regulation of ammonium assimilation in *Haloferax mediterranei*: interaction between glutamine synthetase and two GlnK proteins. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1834(1), 16-23.
- Radchenko, M. V., Thornton, J., & Merrick, M. (2014). Association and dissociation of the GlnK–AmtB complex in response to cellular nitrogen status can occur in the absence of GlnK post-translational modification. *Frontiers in Microbiology*, 5, 731.
- Rolfsmeier, M., & Blum, P. (1995). Purification and characterization of a maltase from the extremely thermophilic crenarchaeote *Sulfolobus solfataricus*. *Journal of bacteriology*, 177(2), 482-485.
- Rolfsmeier, M., Haseltine, C., Bini, E., Clark, A., & Blum, P. (1998). Molecular characterization of the α -glucosidase gene (malA) from the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Journal of bacteriology*, 180(5), 1287-1295.
- Sant'Anna, F. H., Trentini, D. B., de Souto Weber, S., Cecagno, R., Da Silva, S. C., & Schrank, I. S. (2009). The PII superfamily revised: a novel group and evolutionary insights. *Journal of molecular evolution*, 68(4), 322-336.
- Sehgal, S. N., & Gibbons, N. E. (1960). Effect of some metal ions on the growth of *Halobacterium cutirubrum*. *Canadian Journal of Microbiology*, 6(2), 165-169.
- Sharma, G., Yao, A. I., Smaldone, G. T., Liang, J., Long, M., Facciotti, M. T., & Singer, M. (2021). Global gene expression analysis of the *Myxococcus xanthus* developmental time course. *Genomics*, 113(1), 120-134.
- Singh, B., Szamosi, I., Hand, J. M., & Misra, R. (1992). *Arabidopsis acetohydroxyacid* synthase expressed in *Escherichia coli* is insensitive to the feedback inhibitors. *Plant physiology*, 99(3), 812-816.
- Snijders, A. P., Walther, J., Peter, S., Kinnman, I., De Vos, M. G., Van de Werken, H. J., ... & Wright, P. C. (2006). Reconstruction of central carbon metabolism in *Sulfolobus solfataricus* using a two-dimensional gel electrophoresis map, stable isotope labelling and DNA microarray analysis. *Proteomics*, 6(5), 1518-1529.

- Thomas, G., Coutts, G., & Merrick, M. (2000). The *glnKamtB* operon: a conserved gene pair in prokaryotes. *Trends in Genetics*, 16(1), 11-14.
- Torregrosa-Crespo, J., Pire Galiana, C., & Martínez-Espinosa, R. M. (2017). Biocompounds from Haloarchaea and their uses in biotechnology. *Archaea-New Biocatalysts, Novel Pharmaceuticals and Various Biotechnological Applications*. InTech, 63-82.
- van der Oost, J., Schut, G., Kengen, S. M., Hagen, W. R., Thomm, M., & de Vos, W. M. (1998). The Ferredoxin-dependent Conversion of Glyceraldehyde-3-phosphate in the Hyperthermophilic Archaeon *Pyrococcus furiosus* Represents a Novel Site of Glycolytic Regulation. *Journal of Biological Chemistry*, 273(43), 28149-28154.
- Williams, T. J., Allen, M. A., Liao, Y., Raftery, M. J., & Cavicchioli, R. (2019). Sucrose metabolism in haloarchaea: reassessment using genomics, proteomics, and metagenomics. *Applied and environmental microbiology*, 85(6), e02935-18.
- Yao, A. I., & Facciotti, M. T. (2011). Regulatory multidimensionality of gas vesicle biogenesis in *Halobacterium salinarum* NRC-1. *Archaea*, 2011.

Chapter 3. Halophiles In Course-based Undergraduate Research

Experiences (CUREs)

3.1 Publication Status

1. Furrow, R. E., **Kim, H. G.**, Abdelrazak, S. M., Dahlhausen, K., Yao, A. I., Eisen, J. A., Goldman, M. S., Albeck, J. G., & Facciotti, M. T. (2020). Combining Microbial Culturing With Mathematical Modeling in an Introductory Course-Based Undergraduate Research Experience. *Frontiers in microbiology*, 11. doi: [10.3389/fmicb.2020.581903](https://doi.org/10.3389/fmicb.2020.581903)

Abstract: Quantitative techniques are a critical part of contemporary biology research, but students interested in biology enter college with widely varying quantitative skills and attitudes toward mathematics. Course-based undergraduate research experiences (CUREs) may be an early way to build student competency and positive attitudes. Here we describe the design, implementation, and assessment of an introductory quantitative CURE focused on halophilic microbes. In this CURE, students culture and isolate halophilic microbes from environmental and food samples, perform growth assays, then use mathematical modeling to quantify the growth rate of strains in different salinities. To assess how the course may impact students' future academic plans and attitudes toward the use of math in biology, we used pre- and post-quarter surveys. Students who completed the course showed more positive attitudes toward science learning and an increased interest in pursuing additional quantitative biology experiences. We argue that the classroom application of microbiology methods, combined with mathematical modeling using student-generated data, provides a degree of student ownership, collaboration, iteration, and discovery that makes quantitative learning both relevant and exciting to students.

2. Rodriguez-Medina, J., **Kim, H. G. (co-first author)**, Castro, J., Contreras, C. M., Glon, C. L., Goyal, A., Guo, B. Y., Knowles, S., Lin, J. C., McGuinness, C. L., Sorkin, E., Stefani, J., Yegireddi, S. J., Chaganti, S., Cui, D., Deck, S. L., Deokule, Y., Douglas, H., Kenaston, M., O'Brien, A., Patterson, E., Schoppa, N., Vo, D. T., Tran, K., Tran, T., Pérez-Irizarry, V., Carrasquillo-Nieves, K., Montalvo-Rodriguez, R., Yao, A. I., Albeck, J. G., Facciotti, M. T., Nord, A. S., & Furrow, R. E. (2020). Draft genome sequences of 16 halophilic prokaryotes isolated from diverse environments. *Microbiology resource announcements*, 9(8), e01540-19. doi: [10.1128/MRA.01540-19](https://doi.org/10.1128/MRA.01540-19)

Abstract: Halophile-specific enzymes have wide-ranging industrial and commercial applications. Despite their importance, there is a paucity of available halophile whole-genome sequences. Here, we report the draft genome sequences of 16 diverse salt-tolerant strains of bacteria and archaea isolated from a variety of high-salt environments.

Selections from these two works have been incorporated in this chapter, indicated by italicized text.

3.2 Author Contributions

(Furrow, 2020; Sections 3.1-3.7.2 and 3.8.) AY, KD, HK, JE, JA, MG, MF, RF, and SA contributed to the conception and design of the course. MG, MF, and RF contributed to the conception and design of the educational assessment. JA, MF, and RF each served as co-instructors for at least one iteration of the course. HK and SA each served as teaching assistants for one iteration of the course. AY provided technical laboratory support. RF performed the qualitative and statistical analysis. KD, HK, MF, RF, and SA wrote the initial draft of the manuscript. HK, MF, MG, and RF reviewed and edited the final draft.

(Rodriguez-Medina, 2020; Sections 3.6.4. and 3.7.1.) JR, HK, AY, JG, MF, AN, and RF contributed to the conception and design of the two courses that led to this publication. JA, MF, and RF served as the co-instructors for the Fall course, and AN and RF served as the co-instructors for the Spring course. HK each served as teaching assistant for the Fall course and JR served as teaching assistant for the Spring course. AY provided technical laboratory support

for the Fall course. VP, KC, and RM provided environmental samples from Puerto Rico for the Fall course. HK, JC, CC, CG, AG, BG, SK, JL, CM, ES, JS, SY, SC, DC, SD, YD, HD, MK, AO, EP, NS, DV, KT, and TT collected and analyzed genomes from the environmental samples provided. HK performed the final steps of culturing, genomic preparations, and quality control. JR, HK, MF, and RF wrote the initial draft of the manuscript, reviewed, and edited the final draft.

3.3 Introduction

Chapter 1 reviewed metabolic regulation, with a focus on carbon catabolite repression, and provided justifications for using haloarchaea as a model group for research into the comparatively under-studied archaeal domain. An experimental study of carbon catabolite repression in the haloarchaeon *Halogeometricum borinquense* was presented in Chapter 2.

This chapter details the course design, implementation, and pedagogical study of a quantitative biology Course-based Undergraduate Research Experience (CURE) that engages students in the environmental isolation, DNA sequencing, and quantitative characterization of halophilic microorganisms. CUREs are pedagogical tools designed to increase opportunities for student access to research experiences. Various studies (described in more detail below) suggest that engaging students in hands-on research experiences early on in the undergraduate curriculum can positively impact inclusion, retention, and scientific identity building. These previously described outcomes provide part of the rationale for our own work on CUREs.

We hypothesize that engaging first-year-level undergraduate students in a quantitative biology CURE can have positive impacts on student attitudes towards the application of quantitative methods in biology, and test this hypothesis through the design and implementation of a CURE, titled “Genome Hunters.” This work served as my introduction into educational research and resulted in two published manuscripts. In both cases, I made significant

contributions to the experimental design, implementation, data analysis, and manuscript preparation as part of a greater collaborative effort.

3.4 The Importance Of Undergraduate Research Experiences In STEM Curriculums

Undergraduate research experiences, or UREs, provide important formative opportunities for undergraduate students to work collaboratively across educational levels on an intellectual question of interest. UREs can allow students to actively participate in the process of scientific inquiry by providing opportunities to plan and conduct research, perform data analysis, take ownership of a project, and develop technical skills and foundational knowledge.

Numerous efforts have tried to evaluate the impacts that UREs have on STEM students' development of personal confidence, scientific identity, conceptual knowledge, technical skills, interest and retention in the field, and vocational goals (Seymour, 2004; Linn, 2015; Eagan, 2013; Chang, 2014; Lopatto, 2004; Bauer, 2003; Zydney, 2002; Hathaway, 2002; Russell, 2007; Petrella, 2008; Cartrette, 2012; Junge, 2010). Table 3.1 summarizes the major identified potential benefits of UREs. Note that when interpreting the conclusions drawn from these studies, it is important to appreciate that many of these studies involve self-reported data or interviews on small samples, and may lack comparison groups.

Some of the most salient benefits observed relate to students' personal development, such as increased confidence in their scientific abilities, and increased development of scientific identity (Seymour, 2004). For example, in one large-scale study, 83% of respondents reported an increased sense of confidence in their research skills, and 88% of students reported that their understanding of how to conduct research projects increased "by a fair amount or a great deal" (Russell, 2007). Another key finding from longitudinal surveys is that UREs can also be important for improving the retention of students in science, and increasing student interest in

pursuing graduate school and STEM careers (Eagan, 2013; Hernandez, 2013; Russell, 2007; Seymour, 2004; Slovacek, 2012; Zydney, 2002). In one survey, participation in research programs increased the likelihood for students to intend to pursue a graduate or professional program in a STEM field by approximately 14-17% (Eagan, 2013).

Table 3.1. Overview of potential benefits provided by Undergraduate Research Experiences (UREs).

Personal and professional gains with regard to engaging in science	• Increased confidence in ability “to do” research and to contribute to science • Increased confidence to work creatively and independently • Increased confidence in addressing methodological or technical issues and approaching new techniques and lab instruments • Development of identity, often described as “feeling like a scientist” • Increased professional confidence in ability to plan presentations, explain work to others, and answer questions • Opportunities to work one-on-one with faculty, often times a student’s first chance to work with faculty as partners • Opportunities to work with peers and other mentors, and establish a sense of scientific community • Increased sense of excitement, accomplishment, and responsibility with regard to conducting scientific work • Increased intrinsic motivation and interest in learning
Developing and applying scientific knowledge and skills & gaining familiarity with the nature and culture of scientific research	• Increased critical thinking and problem-solving; approaching research questions by applying learnings from previous literature and developing and using methodology to address the questions • Increased ability to link theory and practice and understand motivations for learning material in course curriculums • Deepening and solidifying knowledge within the field and understanding cross-disciplinary concepts and connections • Increased understanding of the nature of scientific knowledge and “facts,” and how knowledge is built • Increased understanding the nature of conducting scientific work (e.g., scientific culture, funding logistics, etc.) • Development of scientific “temperamental suitability,” including patience and resilience encountering setbacks, handling uncertainties in their own understandings, and gaining bearings on the requisite amount of time and effort for conducting experiments and analyses
Gains in technical and work management skills	• Improved communication (presentation, explanation, discussion, giving and receiving critiques, defense of ideas, scientific and professional writing) • Opportunities to learn lab/field techniques (instrumentation, measurement, technical skills) • Increased work organization (time management, record keeping, data management, laboratory management) • Opportunities to learn and develop computer and data analysis skills • Gains in retrieving, comprehending, analyzing, critiquing academic literature and other sources of information • Opportunities to work collaboratively across various levels of education

Clarification or validation of education and vocational paths	• Clarification or validation of disciplinary interests • Increased awareness of and interest in graduate school • Refinement or clarification of vocational goals • Increased persistence in pursuit of undergraduate degree • Increased levels of pursuit of graduate degree
Enhanced graduate school and vocational preparation	• Opportunities to navigate “real world work experiences” • Opportunities for collaboration and networking across educational levels • Enhancement of resume and opportunities to receive letters of recommendation for graduate school and future jobs; • Exposure to new job-related experiences and opportunities for networking (e.g., participation in professional conference settings)
Other benefits	• Potential opportunity for course credit or employment, in some cases

Despite the well-studied benefits of UREs, the number of available research opportunities is in many cases outnumbered by student demand. This is largely a result of the dearth of resources that limit the availability of and access to one of the most traditional forms of UREs, student research positions in faculty laboratories. Due to the nature of the work involved, these types of UREs are typically thought to be resource-intensive, requiring time, effort, and funding from students, educators, and institutions alike.

Compounding the issue of the scarcity of UREs is the fact that these positions are typically offered as co-curricular experiences through which students can earn units. They are typically unpaid. This is particularly burdensome for students who have competing priorities, such as those with jobs or caretaking duties, or those requiring accommodations due to physical or mental health needs. Access to UREs can be further limited by highly competitive selection processes, which may favor students who have already achieved high grades and test scores, or those who have otherwise shown motivation to further pursue science (i.e., are not at risk of leaving STEM) (Linn, 2015).

Alongside the aforementioned external barriers, Bangera and Brownell (2014) identify several internal barriers that can reduce participation in research opportunities: (1) lack of awareness of existing research opportunities; (2) lack of awareness of the possible benefits of

research experiences; (3) lack of awareness of cultural norms associated with scientific research; and (4) perceived barriers to interactions with faculty. The cause of these internal barriers is often rooted in inequities in backgrounds (e.g., race, gender identity, socioeconomic status, first-generation status). Together, all of the factors discussed negatively impact students, especially those from historically excluded communities.

Outcomes of these barriers cascade into long-term impacts on subsequent opportunities available to individuals in the workforce. The impacts start “early,” as UREs are in many cases viewed as an unofficial prerequisite or expectation for graduate school admissions or securing employment post-graduation, particularly in STEM-heavy disciplines. Students that have decreased access to UREs are therefore less likely to gain access to graduate education or are at a disadvantage for securing STEM-related vocational opportunities. The impact from these barriers is heaviest on individuals from historically excluded communities, which, along with numerous other sociocultural and economic factors, has contributed to low racial diversity within the STEM workforce. The percentage of Black and Latin American workers in the STEM fields has trailed the percentages in the general workforce. Black and Latin American people composed 11% and 17% of the United States workforce as of 2019, yet only made up 9% and 8% of the STEM workforce, respectively. Some gains have been made in the life sciences field, where the Black and Latin American percentage grew from 4% to 6% and 7% to 8%, respectively, between 2016 and 2019. In the field of mathematics, the Latin American percentage increased to 8% from 7% in that same timespan, but was unchanged for Black Americans at 9% (Woolston, 2021).

Intemann (2009) identifies three main reasons for why diversity is important in the STEM workforce: (1) science practice and knowledge should be equitable to all and not simply for the privileged; (2) the social, political, and/or economic barriers may erroneously prohibit talented individuals from participating in science; and (3) it is beneficial in the practice of science to identify and address problematic biases and faulty assumptions, which can best be done with

the inclusion of diverse perspectives and lived experiences. In addition, they list some specific ways that diversity of researchers can help increase the “objectivity” (i.e., mitigate the effects of personal biases) of research, including but not limited to generating new research questions, identifying limitations with existing models and proposing new models, proposing alternative hypotheses and interpretations of data, assessing accurate and complete data from subjects, opening up new lines of evidence, revealing loaded language, and identifying a fuller range of risks (Intemann, 2009). Campbell (2013) conducted an empirical study which suggested that promoting diversity may also lead to higher quality science, showing that gender-heterogeneous research groups received 34% more citations on their work than gender-uniform groups. In another study, Antonio (2004) experimentally demonstrated that racially diverse compositions of college discussion groups were associated with higher perception of novelty from the contributors, and higher complexity thinking.

One proposed solution to address barriers to research opportunities is to restructure the entry point into research (Desai, 2008). CUREs offer students the opportunity to simply enroll in a course that provides them with short-term (often 1-2 academic terms), yet highly structured research experiences that focus on providing the elements of a UREs that have significant impact on retention and building scientific identity. Rather than enrolling in lab courses with known outcomes, “plug and chug” problems throughout the term, or relying on rote memorization, students in CUREs engage in exploring research questions that could lead to new knowledge, allowing for a greater feeling of relevance and purpose for completing the work compared to traditional laboratory courses (Shortlidge, 2017).

3.5 Overview Of Course-based Undergraduate Research Experiences (CUREs)

CUREs typically engage students in both didactic lessons, in which students develop conceptual knowledge from structured lectures, readings, and/or other content sources and active research, in which students engage directly in the work of formulating questions and the

processes that can help answer them (Shaffer, 2014; Szteinberg, 2013). The core elements of CUREs that separate them from traditional laboratory courses include exploring questions with unknown answers, engaging students in scientific practices and scientific discovery, emphasizing collaboration, and integrating iteration into the course (Auchincloss, 2014; Corwin, 2018). CUREs combine educational and research goals, and allow for structured, shorter-term research projects where students can work in teams or in close proximity with others on a shared goal. In a UC Davis CURE led by Justin Siegel and Ashley Vater, students work together to design, build, and test novel beta-galactosidase mutants while learning about protein modeling and contributing to a student-built database of mutants and their phenotypes (Vater, 2020). In another UC Davis CURE led by Gerald Quon, students explore RNA sequence data to explore how ancestry, *Helicobacter pylori*, and social determinants of health impact stomach biology. CUREs are practical and convenient, as their design allows them to be scalable (Brownell et al. designed a CURE for 250 students) – and are also relatively easily transferable and implementable at teaching-centered institutions with more limited research opportunities (Brownell, 2013; Vater 2019).

Several studies have shown that although CUREs are often shorter-term experiences, they can provide overlapping benefits to that of UREs, including students' increased feelings of project ownership, pride in contributing to a relevant research question, and improved ability to interpret data, in addition to providing more inclusive options for students (Williams, 2018; Brownell, 2015; Corwin, 2018). Further, for reasons noted earlier, UREs are often pursued by college students in their junior or senior level year; by offering the experiences as part of the course curriculum, CUREs make it possible to become engaged in research experiences earlier on (Szteinberg, 2013; Rodenbusch, 2016). Providing research experiences earlier in the undergraduate timeline has been shown to significantly improve the likelihood of students graduating with a STEM degree within 6 years and is beneficial for helping students define academic and other future goals earlier on (Corwin, 2018; Rodenbusch, 2016; Harrison, 2011).

Of note, CUREs are limited by the number of continuous terms that exist, usually spanning only 1-2 academic terms, and thus cannot exactly replicate some of the potential experiences and benefits of longer-term UREs. For example, several studies suggest that some of the benefits of UREs are more pronounced only after the first year of participation in the research experience, which often entails a great amount of time spent on setting up and conducting experiments. One study showed that the percentage of students who expected to obtain a Ph.D. stood at 8% for those with no research experience, 13% for those with 1-3 months of research experience, and 30% for those with over one year of research experience (Russell, 2007). Generally, shorter-term experiences (less than one year in duration), whether URE or CURE, can be limiting in terms of the amount of data analysis and iteration that students undertake, as well as the depth of understanding the student develops of the research question. These are important aspects of scientific identity building and are fundamental aspects of conducting research work, which students should be exposed to in order to more wholly understand the nature of science work and make more informed decisions about future occupations (Linn, 2015; Adedokun, 2014). CUREs can partially mitigate this by intentionally structuring the course to include built-in time for iteration, and to implement lecture sessions into the design of the course.

CUREs are not meant to serve as replacements for faculty lab-based UREs, rather they are a promising alternative way to envision providing hands-on research experiences in traditional curriculums. They represent additional avenues for gaining research experiences and are becoming more commonplace as undergraduate curriculums evolve to meet goals like those described in President's Council of Advisors on Science and Technology (PCAST) report (Olson, 2012; pp. 16, 25, and 38), to “advocate and provide support for replacing standard laboratory courses with discovery-based research courses.” Table 3.2 outlines the major comparisons between UREs, CUREs, and traditional laboratory courses.

Table 3.2. Comparisons of traditional lab courses, CUREs, and UREs.

	Traditional lab course	CUREs	UREs
Duration	1 term	Often 1 term, but can extend to 2+ terms	Can range from 1 term or less, to multiple years
Time commitment per week	Confined to set times for lectures, labs, office hours (hours per week should correspond to course units)	Confined to set times for lectures, labs, office hours (hours per week should correspond to course units)	Outside of class hours (3-12 hours per week, on average)
Nature of work being done	Laboratory exercises with “plug-and-chug” or known outcomes	Hypothesis-based research question broadly relevant to a subject matter expert; students are making new discoveries, collecting and analyzing novel data	Hypothesis-based research question pertaining to professor’s laboratory
Mentoring	Limited interactions with a small number of professors and TAs, shared between a class	Limited interactions with a small number of professors and TAs, shared between a class	Richer interactions with one professor, graduate and other undergraduate students, and lab staff
Accessibility	Open to majors, minors, then other students, limited by course enrollment limits	Open to majors, minors, then other students; sometimes specifically recruits students from historically excluded communities; limited by course enrollment limits	Some labs recruit students from historically excluded communities; others have space for many students, and others for just a few, leading to potentially selective processes
Advantages	Already the most widely-applied option (i.e., no new development needed); can be potentially more cost-effective with many students taking the course and not as much specialization as CUREs or UREs; still provides students with a basic understanding of what laboratory work entails in terms of technical skills and procedures	Research experiences can be scaled up so that more people can participate in and potentially gain benefits from research experiences than possible in traditional UREs; students receive course credit for their work and do not have to invest too much time outside of the class; can be packaged and implemented across institutions; instructors could theoretically receive teaching credits for mentorship work	Students have the opportunity to be involved in a longer-term experience and are likely to receive deeper levels of mentorship than possible during the duration of a CURE (many students spend more hours per week participating in UREs); with longer-term and less highly-structured experiences, there may be more opportunities for independent project development and iteration
Limitations	Does not allow for students to participate in new discoveries; shorter lab courses translate poorly to research experiences	Mentor time and energy investment needed to supervise research projects; cost is said to be higher than traditional lab courses	Significant investment of mentor time and energy; barriers to accessing due to limited space in the lab

3.6 “Genome Hunters” Course Framework And Curriculum Development Goals

“Genome Hunters” (Bis23A/B) is a two-course series that was designed for the new Quantitative Biology undergraduate curriculum at UC Davis, first piloted in Fall 2018 and Spring 2019. One of the main objectives of this course was to promote quantitative learning and cultivate positive attitudes towards math and quantitative analysis (which has historically been poor) in early-stage (freshmen and sophomore) biology-interested students (Wachsmuth 2017; Matthews, 2009). The CURE format was chosen to help meet this objective.

Bis23A/B was designed to introduce students to quantitative concepts in biology in both the wet lab and computational settings. In the Fall 2018 course, students conducted wet lab work learning to isolate pure microbial cultures from environmental samples, characterizing their growth and isolating their genomes. Students enrolled in the Fall had the option to extend their research experience by enrolling in the Spring 2019 course, a computer-based research experience in which students learned to computationally analyze the sequences of the genomes isolated in the Fall. The Fall Quarter team consisted of John Albeck, Marc Facciotti, Robert Furrow, Mark Goldman, Andrew Yao, and myself. The following sections of the thesis describe the development of Bis23A in Fall 2018 and a reporting/analysis of the course outcomes. The teaching staff defined the following goals for the course:

- 1) *Introduce students to and build basic proficiencies for success in biological research in the wet lab, including experimental design, core laboratory fundamentals, and safety,*
- 2) *Introduce students to and build basic proficiencies in the use of computational tools used in the analysis of data in biology (using the R programming language),*
- 3) *Introduce students to elements of quantitative analysis in biological research, and*
- 4) *Provide students an opportunity to engage closely with faculty, graduate students, and staff.*

3.6.1 Determining A Research Work Flow

Prior to Fall Quarter 2018, the Bis23 teaching team convened a series of discussions to define the course objectives and to identify the student experiences and pedagogical tools that could help meet the aforementioned goals. Instructors needed to weigh their objectives against

many critical factors, but particularly against time and resource constraints. Instructors also needed to take into consideration the relative inexperience of the students and incorporate prior knowledge from the literature. For instance, Shaffer (2014) advocated for increased investment in instructional time with combined lecture, discussion, and lab work in order to build confidence in students performing genomics work.

The teaching team sought to incorporate widely applicable skills, such as those commonly used in academic labs or industrial settings, into the course. This included wet lab skills like measurement, microbial culturing and characterization, microscopy, and spectrophotometry. Each of these technical skills could also be approached through the lens of quantitative reasoning and analysis. For example, when training students to pipet, they could also be asked to perform calculations of dilution factors and to assess the sources and magnitude of error in their measurements. Formal discussions of mathematical modeling could also complement the development of a student's technical ability to collect conditional growth assays of microorganisms by enhancing their ability to interpret the gathered data.

The teaching team also held discussions regarding mentor investment and interactions with students, methods to increase student buy-in, the ease or difficulty of learning techniques, how conducive techniques are to iterative learning within a university schedule (a 10-week term with lab courses occurring once a week during set hours, and considerations of office hours as well), as well as logistical constraints, such as the availability and cost of supplies, sample storage and labeling, and aligning experimental biology timelines (e.g., incubations, growth assays, etc.) with in-person instruction time. Table 3.3 provides an overview of the major aspects that went into the course design.

With these considerations in mind, we decided to use a modified microbial isolation workflow that has been used previously in other CUREs (Vater, 2016; Dunitz, 2015). In our modified workflow, students isolate microorganisms from environmental samples of interest and conduct growth assays to quantitatively characterize the organism's growth behavior. The

experimental work culminates with the isolation and sequencing of the genomes of individual isolates (Figure 3.1). At the end of the course term, instructors conducted quality control checks on the student-isolated genomic DNA and repeated the genomic isolation steps for DNA samples that did not pass quality control thresholds required for sequencing. The resulting data was analyzed in the Spring 2019 follow-up course.

The structure of (the) 10-week quantitative biology CURE include(d) 1 weekly 3-hour wet lab session, offered in the Molecular Prototyping and BioInnovation Lab at UC Davis, as well as a 1-hour weekly lecture held in a traditional classroom. This laboratory is a multi-use, collaborative biomaker space that hosts hands-on courses, student projects, and entrepreneurial ventures.² The lectures, which were hosted in a traditional classroom space, cover the quantitative theory associated with the hands-on research experience and connect this theory to the lab applications. Students also complete a weekly homework that combines a lab write-up with quantitative problem-solving. As the course transitions into student projects, student learning is assessed with an initial project proposal, a final written report, and an oral group presentation.

² <https://team.ucdavis.edu/>

Table 3.3. Overview of the “Genome Hunters” course design aspects.

Aspects of course design	Course elements and skills involved	Targets for success (Learning objectives)	Tools to evaluate success
Technical skill development	- Lab safety and best practices - Fluid and basic laboratory equipment handling - Microbial isolation and culture - Microscopy - Conditional growth assays - Spectrophotometry	- Demonstrate safe practices in the laboratory. - Demonstrate ability to select the correct measurement tool for routine laboratory fluid handling tasks. - Demonstrate ability to use a properly calibrated micropipette to transfer a target volume of low, medium and high viscosity solutions to within 5% of target value. - Identify, find and safely use vortex, micro centrifuge, balance and precision balance, and micro pipet. - Demonstrate the ability to independently set-up a liquid enrichment culture using clean/sterile technique. - Demonstrate ability to spread liquid samples on agar plates using beads, inoculation loops and bent bar spreader, doing little damage to the agar substrate and identifying when each method is most appropriate. - Make observations about microbial growth on agar plates and reflect on conditions needed for successful growth. - Demonstrate the ability to independently set-up a liquid enrichment culture and/or new streak plate of isolated colonies using clean/sterile technique. - Demonstrate the ability to independently outline the procedures involved with obtaining pure cultures of microbial isolates from environmental samples. - Demonstrate the ability to create new cultures from preexisting cultures. - Given some data regarding new environmental samples, create an appropriate procedural outline for a microbial isolation. - Demonstrate the ability to prepare microscope slides and make observations on microbial cell shape using a properly set-up compound light-microscope. - Propose one or more experimental protocol(s)/procedure(s) for quantitatively measuring and comparing growth rate of a microorganism(s) in one or more liquid culture media. The description of the experiment should include a basic protocol and rationale behind each step. - Create a sketch and accompanying explanation for how a very basic spectrophotometer functions, distinguishing at a basic level between absorbance and light scattering and where each of these phenomena apply in the experiments for this class (i.e., how is the data that you are collecting being generated).	Completion of lab tasks, Lab reports

	- Genomic DNA Purification - 16S rRNA PCR Amplification - Gel electrophoresis - Conditional growth assays	- Gain a basic understanding of how genomic DNA can be isolated from pure liquid cultures (be able to outline the basic steps of this protocol). - Be able to outline the basic steps involved in polymerase chain reaction (PCR). - Be able to explain the processes involved in gel electrophoresis. - Demonstrate the ability to design a conditional growth curve experiment.	
Conceptual skill development	- Measurement error - Searching for and reading scientific literature - Organizing data and metadata - Experimental design and set up - Interpreting cell growth data and using calculations for dilution factors - Environmental sampling and sampling from distributions	- Define the concepts of accuracy, precision, error, reproducibility, and repeatability and be able to select which of these terms can be associated with a part of an experimental setup or results. - Use at least two methods to find papers relevant to a question you are interested in. - Summarize the primary differences in purpose and typical content between primary literature and review articles. - Define the term “metadata”, describe its importance for scientific data, and give examples of metadata. - Explain how a relational database helps to organize data, and describe what “fields” and “records” refer to within such a database - List three challenges in using computers to store scientific data and discuss solutions to them. - Draw on a rudimentary understanding of microbial diversity to explain why it is sometimes desirable to perform isolation experiments in different isolation media. - Demonstrate the ability to design a conditional growth curve experiment. - Given an approximate cellular density in solution and the surface-area of a petri dish, calculate the dilutions of the cell culture that would be required to isolate single colonies on the plate assuming all cells, or some fraction thereof, in suspension are viable. - Demonstrate the ability to interpret data from growth curve experiments. - Understand the concept of a distribution, and describe 3 different types of distribution. - Be able to use 2 methods to graphically display a distribution of measurements. - Explain the concept of sampling and how it relates to knowing what a distribution looks like. - Be able to define key terms - environmental sample, enrichment culture, cell, colony - and create statements showing the relationship between the terms.	Lab reports, Assessments, Homework assignments, Presentations

	- Models in biology and modeling in R - Understanding -omics concepts - Understanding and using common laboratory vocabulary - Oral and written presentation skills	- Define the concept of a model and explain the importance of models in experimental science. - List key steps in developing a model. - Describe two simple types of population growth models. - List 2 well-known models in biology and explain what they represent and how they have been used. - Explain the difference between simulations and solutions for models. - Explain how binding constants are defined and how they relate to mass action models of binding. - Define the following terms and describe their use in modeling data: regression, residual, least squares, coefficient of determination. - Gain a basic understanding of why and how the 16S rRNA gene is used in the field of life sciences. - Explain how first and second generation DNA sequencing technologies work. - Describe uses for genome sequences. - Define the following terms: phenotype, genotype, lag phase, exponential phase, stationary phase, death phase, and colony-forming units (CFU). Be able to understand and use these vocabulary terms in both verbal and written communication and in an appropriate context. - Be able to define key terms: bacteria, archaea, genome, DNA, pellet, lysis, protein precipitation, primer, template, target gene, amplification. - Be able to define key terms: gel, well, fragment, electrode, buffer, charge. - Define the term genome, and list the elements of a functional genome. - Present scientific results in powerpoint, written, and verbal formats, both individually and within group settings.	
Mentor responsibilities (adapted from Linn et al. (2015))	- Identifying and communicating course expectations - Exploring and challenging undergraduate assumptions about research experiences - Monitoring student progress and encouraging regular reflection - Emotionally and intellectually supporting students		Reflections in weekly journals, Student evaluations, Office hour visits, Synthesis of research reports (e.g., Rodriguez-Medina, 2020)

Constraints	- Supply budget - Equipment availability - Growth rate of organisms - 10-week time frame with set hours for labs and lectures		
--------------------	--	--	--

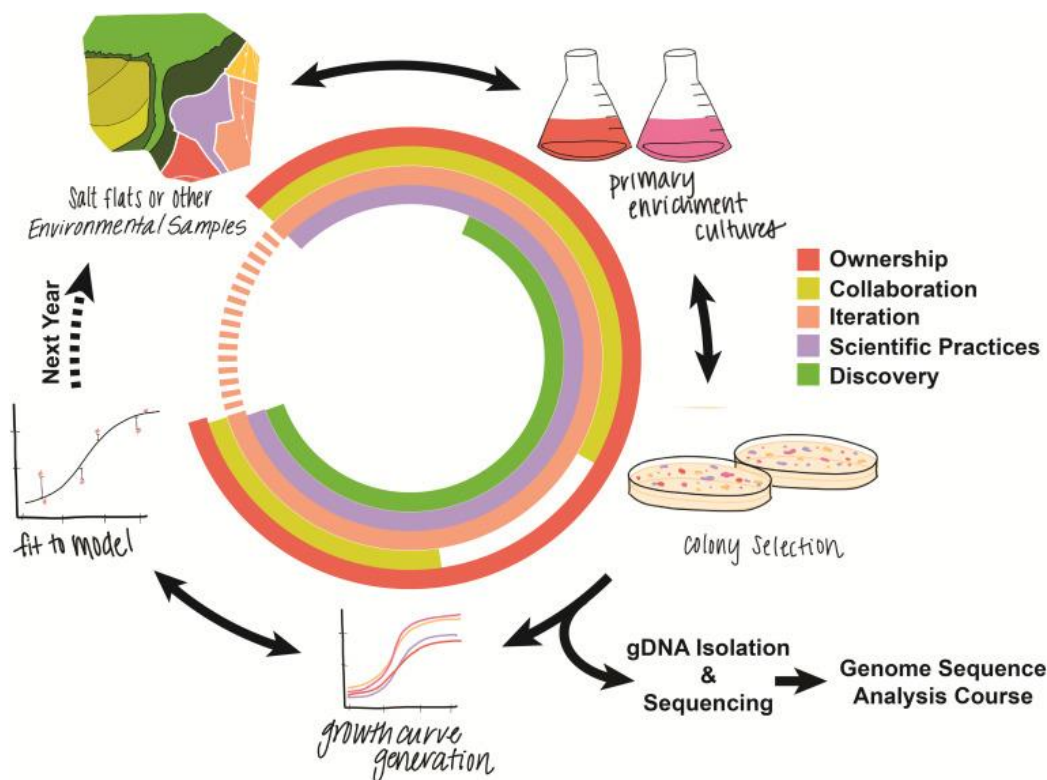


Figure 3.1. Graphical overview of the “Genome Hunters” course structure (Furrow, 2020).

3.6.2 Halophiles As Study Subjects In CUREs

The Fall 2018 iteration of this CURE focuses on culturing, isolating, and quantitatively characterizing the features of halophiles (Rodriguez-Medina, 2020). Chapter 1 reviews key features of halophiles that make them interesting and suitable for study in both independent lab and classroom research settings. In summary, halophiles are a group of organisms that span the major domains of life and are found broadly across the globe in hypersaline environments. Halophiles have interesting and unique biological mechanisms that allow them to thrive in harsh conditions (e.g., high pH, low water or nutrient availability) and produce products of interest, such as: *polyhydroxyalkanoates* (plastics industry), *amylases* (biofuel industry), *proteases* (laundry detergent), *beta-carotene* (food additive), and *glycerol* (cosmetic industry) (Yin, 2015; Amoozegar, 2017; Becker, 2014; Torregrosa-Crespo, 2017).

The fact that halophiles are non-pathogenic also makes them practical to use in the classroom. Because hypersaline growth conditions are inherently inhibitory to non-halophiles, halophilic culturing is more forgiving to small lapses in sterile technique, allowing students to successfully isolate pure cultures even if they are inexperienced in the laboratory. This is well-suited for introducing lab work and microbial culturing to first-year students. Although halophiles have a relatively long doubling time, a single weekly laboratory allows enough time for culture growth between each session. With advances in sequencing, the number of available -omics analysis tools and halophile genomes are increasing, allowing students the opportunity to contribute in meaningful ways while also having a foundation of available resources to work with.

3.6.3 Course Overview And Learning Objectives

The lab starts with an introduction to formal campus and site-specific lab safety training. We require all students to successfully complete the University of California “Fundamentals of Laboratory Safety Training,” an online course required for everyone who works in labs on campus. We explain to the students that this training certifies them to work in faculty research labs on campus. The site-specific training highlights that the workspace is used outside of class hours to host active student research projects (i.e., they are working in a “real” research space and not just a classroom and thus that we expect them to be aware of those activities and associated hazards as they work). We emphasize this training to provide a solid foundation in safety, but also to establish a classroom environment in which students feel like they are doing authentic research.

Following the safety training, an introductory activity on pipetting, mixing, and measurement teaches techniques and orients students to the various instruments and supplies in the lab required for the course. These hands-on exercises are complemented with lecture and homework in which students develop their understanding of measurement and sources of error.

For example, in one activity, students repeatedly pipette and weigh an identical volume of solution and compete to have the least error between their measurements for a small restaurant gift card. *During lecture students had previously learned about accuracy and precision and how these can be quantified, and on the homework students use the programming language R (R Core Team, 2019) to calculate these quantities for their self-collected measurements. This pattern of linked lecture, lab, and homework continues throughout the quarter as each student collects (or selects previously collected) environmental samples, progresses through microbial isolation, and phenotypically characterizes the isolates.*

Students self-collected local samples from soil, water, or salty foods. Students then practice microbial culturing and isolation, using their samples to start cultures. For this course, we use Halobacterium medium 372 (DSMZ, 2007) as the base medium and vary the concentrations of NaCl. By using media with multiple salinities, we can potentially culture a broader diversity of microbes from the samples. All cultures are grown in either a shaking (liquid culture) or static (agar plate) incubator at 37°C. Plates that grow colonies within 7 days are transferred to a 4°C refrigerator to pause growth. To prepare pure cultures, students make phenotypic observations of a single colony of interest, select cells from the colony to inoculate a new liquid culture, and prepare microscope slides to make observations on cell size and shape using a standard compound microscope equipped with phase contrast optics.

With pure liquid cultures in hand, students focus on measuring and modeling microbial growth. Using a spectrophotometer, students measure the growth of each isolate at multiple salinities to quantitatively characterize salt tolerance. Students then use the programming language R (R Core Team, 2019) to fit logistic growth models to their data, explore variants of the models, and decide on the model that best explains the patterns they observe. The course culminates with students performing genomic DNA isolation and shipping off the gDNA for whole genome sequencing. These novel data are then used in a spring quarter follow-up CURE

in which students continue to build quantitative skills as they apply bioinformatic and statistical techniques for comparative genomic analyses (Figure 3.1).

Table 3.4 outlines connections between the planning and implementation phases of the course, and shows alignments between the lecture and the lab sessions. Figure 3.1 shows an overview of the Fall 2018 Bis23A course flow. Table 3.5 outlines the course design, showing the main topics and learning objectives associated with each lecture and lab session. Select curricular materials used in the course, including lab handouts, lab report instructions, an R notebook and associated RData file for a computational lab, final presentation guidelines, and final lab report instructions, can be found in (Furrow, 2020).

Table 3.4. Overview of course planning vs. implementation, and alignment between lecture/lab.

Week	Lecture topic (Tues)	Lab topic (Thurs)	Goals for alignment
1	No class due to beginning of quarter timing.	Introduction to course, lab safety, experimental outline, sources and types of literature	
2	Measurement	Pipetting competition, lab tour, inoculate cultures	Students learned about measurement (error, accuracy, and precision) in lecture and applied their learnings in the lab, practicing pipetting specific amounts of volumes and using calculations to determine error, accuracy, and precision.
3	Data: Introduction to forms/structures of data and data management	Introduction to microbial isolation and culture	Students were introduced to data and metadata structures and organization. In the lab session, students were collecting data with metadata to track as they began microbial isolation and culturing from environmental samples.
4	Distributions and sampling	Continuation of microbial isolation with additional samples	When trying to isolate microbes from a population, there is a selective filter based on media and growth conditions. An exercise on sampling featuring an everyday life, UC Davis-specific example (counting red bikes) was used to help students understand the concept of distributions and sampling.

5	Models in biology	Preparation of pure cultures	Students were introduced to models and how they can be used to help to create a more formal description of the distributions. Different types of models and model parameters exist for different scenarios, with different distributions associated. This was not directly linked to the preparation of pure cultures and served instead as a preview, and built theory and conceptual knowledge helpful for the analysis of differential condition set up and analysis of kinetic growth data in Week 6.
6	Building and using quantitative models	Differential condition setup and analysis of kinetic data	Students were introduced to R and prepared to fit models to kinetic growth data collected in the lab session using R packages.
7	No classes were held due to the Paradise Fire; originally intended for buffer weeks for scientific iteration.		
8	No classes were held due to schedule conflict with the lecture instructor and Thanksgiving holiday; originally intended for buffer weeks for scientific iteration.		
9	Fitting models	Genomic DNA purification and 16S rRNA PCR amplification	Students continued R work and applied previous learnings to use R packages to fit models and quantitatively analyze kinetic growth curves data collected from Week 6 lab. The lab work did not directly relate to the lecture topic.
10	<i>Midterm assessment</i>	Gel electrophoresis & Conditional growth assay; Final presentation preparations and data analysis consultation with instructors	The midterm was used to assess student understanding of the conceptual knowledge up to that point.
11	Genomes	Final presentations	Learning about genomes and sequencing helps students understand the motivation of the work behind DNA purification and 16S rRNA PCR amplification and contextualizes the results obtained from Week 9 lab.

Table 3.5. Overview of main topics and specific learning objectives for each lab and lecture.

Session	Main Topic(s)	Associated Learning Objective(s)
Lab 1 (Week 1)	Introduction to course, lab safety, experimental outline, sources and types of literature	Demonstrate safe practices in the laboratory.

Lecture 1 (Week 2)	Measurement	- Define the concepts of accuracy, precision, error, reproducibility, and repeatability and be able to select which of these terms can be associated with a part of an experimental setup or results. - Use at least two methods to find papers relevant to a question you are interested in. - Summarize the primary differences in purpose and typical content between primary literature and review articles.
Lab 2 (Week 2)	Pipetting competition, lab tour, inoculate cultures	- Demonstrate ability to select correct measurement tools for routine laboratory fluid handling tasks. - Demonstrate ability to use a properly calibrated micropipette to transfer a target volume of low, medium and high viscosity solutions to within 5% of target value. - Identify, find and safely use vortex, micro centrifuge, balance and precision balance, and micro pipet. - Define the concepts of precision and accuracy and, either by collecting or with provided data, apply these concepts to describe the performance of a measurement tool.
Lecture 2 (Week 3)	Data: Introduction to forms/structures of data and data management	- Define the term “metadata”, describe its importance for scientific data, and give examples of metadata. - Explain how a relational database helps to organize data, and describe what “fields” and “records” refer to within such a database - List three challenges in using computers to store scientific data and discuss solutions to them.
Lab 3 (Week 3)	Introduction to microbial isolation and culture	- Demonstrate the ability to independently set-up a liquid enrichment culture using clean/sterile technique. - Draw on a rudimentary understanding of microbial diversity to explain why it is sometimes desirable to perform isolation experiments in different isolation media. - Demonstrate ability to spread liquid samples on agar plates using beads, inoculation loops and bent bar spreader, doing little damage to the agar substrate and identifying when each method is most appropriate. - Given an approximate cellular density in solution and the surface-area of a petri dish, calculate the dilutions of the cell culture that would be required to isolate single colonies on the plate assuming all cells, or some fraction thereof, in suspension are viable.
Lecture 3 (Week 4)	Distributions and sampling	- Understand the concept of a distribution, and describe 3 different types of distribution. - Be able to use 2 methods to graphically display a distribution of measurements. - Explain the concept of sampling and how it relates to knowing what a distribution looks like.
Lab 4 (Week 4)	Continuation of microbial isolation with additional samples	- Make observations about microbial growth on agar plates and reflect on conditions needed for successful growth. - Be able to define key terms - environmental sample, enrichment culture, cell, colony - and create statements showing the relationship between the terms. - Demonstrate the ability to independently set-up a liquid enrichment culture and/or new streak plate of isolated colonies using clean/sterile technique. - Given some cellular density in solution and some basic assumptions, calculate appropriate dilutions and plating volumes that will yield single colonies on agar plates. - Given some data regarding new environmental samples, create an appropriate procedural outline for a microbial isolation.

Lecture 4 (Week 5)	Models in biology	- Define the concept of a model and explain the importance of models in experimental science. - List key steps in developing a model. - Describe two simple types of population growth models.
Lab 5 (Week 5)	Preparation of pure cultures	- Demonstrate the ability to independently outline the procedures involved with obtaining pure cultures of microbial isolates from environmental samples. - Demonstrate the ability to prepare microscope slides and make observations on microbial cell shape using a properly set-up compound light-microscope.
Lecture 5 (Week 6)	Building and using quantitative models	- List 2 well-known models in biology and explain what they represent and how they have been used. - Explain the difference between simulations and solutions for models.
Lab 6 (Week 6)	Differential condition setup and analysis of kinetic data	- Propose one or more experimental protocol(s)/procedure(s) for quantitatively measuring and comparing growth rate of a microorganism(s) in one or more liquid culture media. The description of the experiment should include a basic protocol and rationale behind each step. - Define the following terms: phenotype, genotype, lag phase, exponential phase, stationary phase, death phase, and colony-forming units (CFU). Be able to understand and use these vocabulary terms in both verbal and written communication and in an appropriate context. - Create a sketch and accompanying explanation for how a very basic spectrophotometer functions, distinguishing at a basic level between absorbance and light scattering and where each of these phenomena apply in the experiments for this class (i.e., how is the data that you are collecting being generated).
~ 2 weeks of no class due to Paradise Fire, instructor scheduling conflict, and Thanksgiving holiday; originally intended for buffer weeks for scientific iteration.		
Lecture 6 (Week 9)	Fitting models	- Explain how binding constants are defined and how they relate to mass action models of binding. - Define the following terms and describe their use in modeling data: regression, residual, least squares, coefficient of determination. - Explain the difference between simulations and solutions for models.
Lab 7 (Week 9)	Genomic DNA purification and 16S rRNA PCR amplification	- Demonstrate the ability to interpret data from growth curve experiments. - Gain a basic understanding of how genomic DNA can be isolated from pure liquid cultures (be able to outline the basic steps of this protocol). - Gain a basic understanding of why and how the 16S rRNA gene is used in the field of life sciences. - Be able to outline the basic steps involved in polymerase chain reaction (PCR). - Be able to define key terms: bacteria, archaea, genome, DNA, pellet, lysis, protein precipitation, primer, template, target gene, amplification.
Lecture 7 (Week 10)	Midterm assessment	-

Lab 8 (Week 10)	Gel electrophoresis & Conditional growth assay; Final presentation preparations and data analysis consultation with instructors	• Be able to explain the processes involved in gel electrophoresis. • Demonstrate the ability to design a conditional growth curve experiment. • Demonstrate the ability to create new cultures from preexisting cultures. • Be able to define key terms: gel, well, fragment, electrode, buffer, charge. • Present scientific results in powerpoint, written, and verbal formats, both individually and within group settings.
Lecture 8 (Week 11)	Genomes	• Explain how first and second generation DNA sequencing technologies work. • Define the term genome, and list the elements of a functional genome. • Describe uses for genome sequences.
Lab 9 (Week 11)	Final presentation	• Present scientific results in powerpoint, written, and verbal formats, both individually and within group settings.

3.6.4 Detailed Methodology

To isolate a wider variety of strains, samples were plated in medium 372 at multiple salinity levels, either 100, 150, 200, or 250 g NaCl per 1,000 ml of medium. To grow a diversity of strains, we used both direct-plated samples and samples plated from liquid enrichment cultures. For solid samples, direct plating required hydrating the sample with H₂O at a density of 200 g of sample per liter. To create enrichment cultures, we combined 100 L of sample with 3 mL of liquid medium and incubated it at 37°C until turbid in 7 days. Colonies were grown by spread plating of the prepared samples. Pure liquid cultures were created from isolated plate colonies and grown at 37°C until turbid in 7 days. Purity was assessed using microscopy and visual analysis of streak-plated pure cultures.

Genomic DNA was isolated from the liquid cultures using a Wizard genomic DNA purification kit (Promega, Madison, WI, USA). MicrobesNG (Birmingham, UK) performed library preparation, sequencing, adapter trimming, and assembly (full protocol: microbesng.com/documents/5/MicrobesNG_Methods_Document_-_PDF.pdf). DNA was quantified using the Quant-iT double-stranded DNA (dsDNA) high-sensitivity (HS) assay (Thermo Fisher Scientific, Waltham, MA, USA). Libraries were prepared using a Nextera XT

library prep kit (Illumina, San Diego, CA, USA) with the following two protocol modifications: using 2 ng of DNA instead of 1 ng and adjusting the PCR elongation time to 1 minute. Libraries were sequenced using an Illumina HiSeq instrument with a 250-bp paired-end protocol. A mean of 1,122,706 reads per strain were generated by the sequencing. Reads were trimmed using Trimmomatic 0.30 with a sliding window quality cutoff of Q15 (Bolger, 2014). SPAdes 3.7 (Bankevich, 2012) was used for de novo assembly with recommended settings for 250-bp paired-end reads. The mean coverage ranged from 35 to 289. The mean N50 value was 442,887 bp with a mean L50 of 14.1. The taxonomy was estimated with BLAST (Altschul, 1990) using the 16S rRNA gene sequences of each strain.

3.7 “Genome Hunters” Fall 2018 Course Outcomes

3.7.1 Draft Genome Sequences Of 16 Halophilic Prokaryotes Isolated From Diverse Environments

We identified 14 bacteria (genera Halobacillus, Pontibacillus, Virgibacillus, Bacillus, and Halomonas) and 2 archaea (genus Halorubrum) (Table 3.6). The top query results all had above a 97% identity (99.4% mean identity). All 16 whole-genome sequences and SRA files have been deposited at DDBJ/ENA/GenBank as a BioProject under accession number PRJNA587497. A 17th sample is also included in the BioSample data for the BioProject, but this sample appeared to be a mix of two species and could not be assembled into a high-quality genome. This work was featured by American Society for Microbiology as a Microbiology Resource of the Month (asm.org/Articles/2020/February/Microbiology-Resource-of-the-Month-Student-Discover).

Table 3.6. Summary of the draft genome sequences for 16 microbes isolated from high-salt environments and culinary salt.

Species (% identity)	Isolate source (medium salinity [g/liter NaCl])	No. of reads	Median insert size (bp)	No. of contigs of ≥ 20 bp	Total length (bp)	G+C %	L_{50}	N_{50} (bp)
<i>Bacillus hwajinpoensis</i> (99.79)	French fleur de sel salt from grocery store (100)	560,183	433	87	4,387,231	40.22	3	797,473
<i>Halobacillus halophilus</i> (98.21)	Solar salt harvested from saltern, PR (100)	505,256	448	51	3,841,848	47.38	3	575,568
<i>Halobacillus litoralis</i> (99.23)	Sand harvested near saltern crystallizers, PR (100)	1,228,604	197	258	4,087,490	43.59	2	1,038,528
<i>Halobacillus litoralis</i> (99.36)	Sand harvested near saltern crystallizers, PR (100)	1,488,864	215	92	3,984,713	43.83	3	396,878
<i>Halobacillus litoralis</i> (99.04)	Filtered 34% NaCl saltern water, PR (100)	495,194	484	65	3,850,954	47.23	5	237,386
<i>Halobacillus litoralis</i> (99.1)	Filtered 34% NaCl saltern water, PR (100)	1,866,405	211	577	4,062,334	47.67	2	690,789
<i>Pontibacillus yanchengensis</i> (99.39)	French fleur de sel salt from grocery store (200)	454,979	275	75	4,476,898	38.35	7	200,610
<i>Pontibacillus yanchengensis</i> (99.39)	Filtered 34% NaCl saltern water, PR (200)	1,370,382	221	93	4,441,339	38.43	3	575,184
<i>Virgibacillus halodenitrificans</i> (100)	Sand harvested near saltern crystallizers, PR (100)	1,634,098	238	2,151	5,319,423	37.39	155	6,747
<i>Virgibacillus halodenitrificans</i> (99.87)	Solar salt harvested from saltern, PR (100)	399,182	482	109	4,012,173	37.36	6	215,037
<i>Virgibacillus massiliensis</i> (99.74)	Sand harvested near saltern crystallizers, PR (100)	1,323,006	239	166	4,260,595	36.96	2	1,149,539
<i>Halomonas alkaliantarctica</i> (97.5)	Filtered 34% NaCl saltern water, PR (200)	1,230,594	293	91	3,324,735	58.54	4	316,952
<i>Halomonas utahensis</i> (99.87)	French fleur de sel salt from grocery store (100)	287,001	378	57	3,351,671	62.12	6	219,081
<i>Halomonas utahensis</i> (99.87)	French fleur de sel salt from grocery store (150)	2,659,128	227	64	3,361,279	62.12	4	422,380
<i>Halorubrum terrestre</i> (100)	Solar salt harvested from saltern, PR (250)	999,388	393	90	3,385,208	67.81	11	97,966
<i>Halorubrum terrestre</i> (99.86)	Solar salt harvested from saltern, PR (150)	1,461,031	315	885	3,678,879	65.29	9	146,067

3.7.2 Student Attitudinal Outcomes

Assessments for student attitudinal gains were implemented in Fall 2018, and again in the next iteration of the course in Fall 2019. An online pre-survey was administered during the first week of the course, and a post-survey was given during the final weeks of the quarter; both were taken in class. A comparison group was used: *a non-overlapping group of survey respondents in the University Honors Program who had a biology-related major, both because many of our students were in the University Honors Program and because we hypothesized that other University Honors students may also be motivated to pursue research-oriented laboratory and quantitative experience.* The comparison group of students completed their surveys on their own. Student math-biology values were assessed using a previously created assessment instrument, The Math-Biology Values Instrument, which measures *interest in, utility of, and cost of* taking biology courses that incorporate math (Andrews, 2017). Additionally, a subsection of questions were taken from the CURE survey section: “Your opinions about yourself and about science” (Lopatto, 2008; Shaffer, 2010).

Although math-biology values had minimal change (in part, due to high student enthusiasm at the start of the course), students in the CURE group experienced significant gains for the Learning Science construct, and potentially see higher value to using quantitative calculations in biology. In the Fall 2019 course iteration, students were asked an additional short-answer question: To assess students’ future course and vocational plans related to quantitative biology, we asked additional qualitative questions in the survey for the 2019 iteration of the CURE. In the initial survey we asked: *“Do you have any plans to pursue future courses and/or a career in quantitative biology? Please briefly explain why or why not.”* In the end-of-quarter survey, we asked a matched short-answer question: *“Has this course changed your plans for pursuing future courses and/or a career in quantitative biology? If so, how?”*

Six of seventeen students indicated increased interest in future quantitative experiences, such as enrolling in additional quantitative courses or pursuing future quantitative work

opportunities. Among these six students, two explicitly discussed increased interest as a factor shaping their future plans (e.g., “The course has definitely made me more interested in quantitative biology.”) and two others mentioned positive feelings about doing quantitative biology (e.g., “I definitely found the passion that I was expecting.”). Some responses also revealed how students might be weighing the relative utility and cost of learning quantitative biology (e.g., “Though math is a challenging and often daunting subject for me, I do believe that it is essential to understanding and applying a bit of it into the science world...”).

3.8 Future Directions For “Genome Hunters” CURE

By design, the course’s focus on student research implicitly places students in the role of research scientists. However, student scientific identity could potentially be developed more explicitly by diversifying the structure of meetings and assignments to promote the kind of informal critical thinking, curiosity, and collaboration that occurs in research labs. Examples might include journal clubs, science coffee chats, poster making and presenting, and academic writing practice. To expand collaboration, one might convert part of each lab into a “lab meeting” in which students discuss with peers and take turns summarizing primary literature or sharing updates on their research (e.g., in the CURE presented by Oufiero, 2019). In addition, we would like to help students understand that failure, mistakes, and repeated iteration of data gathering and analysis are normal parts of scientific inquiry. Although the instructors in this CURE discussed these themes in passing during lecture and lab (Seidel, 2015), these goals were not explicit in our lesson planning or assessments.

One logistical challenge for this course was the organization of isolate metadata. Student project ownership may be a foundational source of student benefits from CUREs (Corwin, 2018), so we asked students to own their data from sample to final isolate to quantitative growth data. After students struggled to maintain a complete chain of metadata for their samples in 2018, we implemented a Google Sheets system in which all new data were

added as new columns to a continually growing classroom document. This became cumbersome by the end of the quarter, but also provided a shared workspace in which students could note each other's discoveries and feel like a part of a team effort.

Although we focused on salt-dependent growth rates for halophiles, this course could be adapted to other CUREs based around alternative research questions. Other phenotypic investigations might include color production, sugar utilization, halophilic gas vesicle production, or the production of easy-to-spot products like polyhydroxyalkanoates—all of which require a separate set of research techniques scalable to a course timeframe (e.g., colorimetric assays, visible light microscopy, etc.). One could expand on this to culture strains in various growth conditions such as shaking, oxygenation, salinity, pH, nutrient availability, and/or temperature. Alternatively, students could pick an interesting environment or sample to isolate organisms from and thereafter research and use the appropriate growth materials to study their organism. This may increase student interest and buy-in even further, though logistical organization (e.g., media management, instructor expertise of the organism(s), etc.) must be taken into consideration as well. The most promising alternatives are likely to be those which use a highly selective growth medium, preventing the cultures from being swamped by local contaminants. Our high-salinity media helped reduce contamination problems; extreme temperature, pH, or unusual food sources might be similarly effective.

The skills developed and data created from microbial culturing provide a productive way to engage undergraduate students in course-based research. By combining laboratory skills with growth rate modeling, students learn quantitative skills in a low-stakes environment in which they have ownership over the data they are generating and analyzing. The current version of this quantitative biology CURE emphasizes the growth rates of halophilic microbes, but we expect that this model for course design and implementation can be readily applied to a broad range of organisms and phenotypes. With extra time in the course, it would be worthwhile

to include other common lab techniques, such as enzymatic assays, Western or Southern blot, or qPCR, which are commonly used in research labs or taught in biology lab courses.

Student evaluations at the end of the course were useful for understanding their perspectives of the course and identifying the strengths and weaknesses of the course. The following are selections of student evaluations for the Fall 2018 course iteration:

- “Most interesting lab I have taken so far. It was very useful to have 3-5 instructors in lab with us to help. I liked the lab write ups but wish they were released sooner.”
- “Everything is taught ties into the work performed in the lab.”
- “Overall, great course. Some labs felt a bit rushed, but since it is the first time this course was taught it was understandable and did not retract from the educational quality of the class.”
- “I understand that this quarter was the first time this course was taught, but I was thoroughly confused for a good portion of the class. I wanted to ask questions, but I was not sure where to start or what to ask questions about.”
- “It was too fast paced, and not well put together. The lectures should have been used to explain what we were going to do in the lab. Very little guidance for labs”
- “The lab portion of the [c]lass was a great experience, and I got the chance to experience biological research without having to find an internship in a lab. However, the short explanations we were provided at the beginning of labs were often not quite enough to fully understand our procedures, and I think it would have been more helpful to use lecture periods to go over what we were doing in lab. The lectures often seemed pretty unrelated to what we were doing.”

The comments from the student evaluations indicate that the course itself seemed interesting and the lab sessions in particular were a good experience for students. However, there seemed to be issues with pacing (i.e., too fast-paced), and the alignment of lectures and labs could be improved or made more clear. It may be beneficial to intentionally use examples specific to the model organism when introducing quantitative theories and concepts, rather than using generic examples during lecture sessions. It may also be helpful to use more time during the lecture session to preview laboratory procedures and the theory behind and motivations for using certain techniques. Finally, holding interviews with students would be beneficial to better understand the student perspective towards quantitative learning and CUREs, their sense of belonging, and ways to improve for future course iterations for students with specific backgrounds or goals.

3.9 Future Directions For CURE Course Work

CUREs are becoming an increasingly popular pedagogical approach in higher education STEM fields to increase opportunities and access to research experiences, which have been identified as important, formative learning experiences for scientists-in-training. The benefits of CUREs have been well-documented, and the course and study described in this chapter further suggests that CUREs can potentially make a significant impact on a student's academic and vocational trajectory, including attitudinal learning gains in quantitative biology. Of note, many studies detail the importance of longer-term (as much as multi-year) research experiences for building scientific identities and reaping benefits of the research experiences (Thiry, 2012; Fechheimer, 2011; Thiry, 2011; Bauer, 2003; Hernandez, 2018). Designing CUREs to have continuity over the course of multiple terms can allow for students to participate for an extended amount of time, while also enabling them to “sample” a research experience and conclude their experience with the project at the end of the term.

The assessment of the benefits of CUREs is an area that needs improvement, as well. Further studies on the longitudinal effects of CUREs will help educators and policymakers understand the long-term impact of CUREs and areas that could be improved. Linn (2015) asserts that self-reports and interviews are flawed assessment measures, and should be validated with other outcomes to be agreed-upon, such as improved abilities to conduct scientific processes, interpret original sources, and identify flaws in research designs. Linn (2015) further encourages disaggregating populations of students to advance understandings of which communities of students are most benefited, which will help with the allocation of scarce resources. This may include disaggregating students by their demographic backgrounds and lived experiences, as well as by their pre-existing future goals and their stages of development in their academic careers (Linn, 2015; Hernandez, 2013). For example, Pacifici (2011) found that pre-med students viewed research as a stepping stone to becoming doctors and were not motivated to learn more about science, whereas non-premed students sought to learn more

about specific science topics and use research experiences to gain skills necessary for graduate school. It could be beneficial to design research experiences with these different goals in mind, and consider this in the analysis of student outcomes, while also keeping in mind that research experiences can also profoundly influence student motivations as well. Intentionally analyzing the specific aspects of course design or processes and directly relating them to student outcomes will be useful for improving course design to maximize student gains (Corwin, 2018; Adedokun, 2012). In studies that compare UREs and CUREs, appropriate controls must be included that account for differences among the students (Linn, 2015).

Finally, while it is an important goal of CURE courses to improve issues of diversity, equity, and inclusion in STEM, it is important to recognize that greater access to research experiences via CURE courses represent a potential solution that is only part of a bigger whole. Rather than simply opening access to greater numbers of students, it is also important to be creating safe spaces for students to learn, and having cultural competence and awareness of the structural barriers that continue to exist for many communities. Having a diversity of CURE courses to choose from, as well, will be beneficial for attracting a diversity of students, including research projects that are rooted in local problems, or community-based inquiries and learning opportunities.

3.10 Funding

This work was supported by a Howard Hughes Medical Institute Professor grant awarded to MG (#52008137).

3.11 Acknowledgements

We thank Jonathan Eisen, Katherine Dahlhausen, Mark Goldman, Ashley Vater, and David Coil for help with course design and planning; Elizabeth Moore for creating the artwork

used in Figure 3.1; and Sean ConroyDey, the BioInnovation Group, and the Molecular Prototyping and BioInnovation Lab student staff for technical help preparing and running the laboratory sessions. We are grateful to Rafael Montalvo-Rodríguez, Valeria Pérez-Irizarry, and Krismarie Carrasquillo-Nieves for collecting and sharing environmental samples from Cabo Rojo, Puerto Rico.

3.12 References

- Adedokun, O. A., Zhang, D., Parker, L. C., Bessenbacher, A., Childress, A., & Burgess, W. D. (2012). Understanding how undergraduate research experiences influence student aspirations for research careers and graduate education. *Journal of College Science Teaching*, 42(1), 82.
- Adedokun, O. A., Parker, L. C., Childress, A., Burgess, W., Adams, R., Agnew, C. R., ... & Teegarden, D. (2014). Effect of time on perceived gains from an undergraduate research program. *CBE—Life Sciences Education*, 13(1), 139-148.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403-410.
- Amoozegar, M. A., Siroosi, M., Atashgahi, S., Smidt, H., & Ventosa, A. (2017). Systematics of haloarchaea and biotechnological potential of their hydrolytic enzymes. *Microbiology*, 163(5), 623-645.
- Andrews, S. E., Runyon, C., & Aikens, M. L. (2017). The math–biology values instrument: Development of a tool to measure life science majors’ task values of using math in the context of biology. *CBE—Life Sciences Education*, 16(3), ar45.
- Antonio, A. L., Chang, M. J., Hakuta, K., Kenny, D. A., Levin, S., & Milem, J. F. (2004). Effects of racial diversity on complex thinking in college students. *Psychological science*, 15(8), 507-510.
- Auchincloss, L. C., Laursen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., ... & Dolan, E. L. (2014). Assessment of course-based undergraduate research experiences: a meeting report.
- Bangera, G., & Brownell, S. E. (2014). Course-based undergraduate research experiences can make scientific research more inclusive. *CBE—Life Sciences Education*, 13(4), 602-606.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5), 455-477.
- Bauer, K. W., & Bennett, J. S. (2003). Alumni perceptions used to assess undergraduate research experience. *The Journal of Higher Education*, 74(2), 210-230.
- Becker, E. A., Seitzer, P. M., Tritt, A., Larsen, D., Krusor, M., Yao, A. I., ... & Facciotti, M. T. (2014). Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. *PLoS genetics*, 10(11), e1004784.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120.
- Brownell, S. E., Kloser, M. J., Fukami, T., & Shavelson, R. J. (2013). Context matters: volunteer bias, small sample size, and the value of comparison groups in the assessment of research-based undergraduate introductory biology lab courses. *Journal of Microbiology & Biology Education*, 14(2), 176-182.
- Brownell, S. E., Hekmat-Scafe, D. S., Singla, V., Chandler Seawell, P., Conklin Imam, J. F., Eddy, S. L., ... & Cyert, M. S. (2015). A high-enrollment course-based undergraduate research experience improves student conceptions of scientific thinking and ability to interpret data. *CBE—Life Sciences Education*, 14(2), ar21.
- Campbell, L. G., Mehtani, S., Dozier, M. E., & Rinehart, J. (2013). Gender-heterogeneous working groups produce higher quality science. *PloS one*, 8(10), e79147.
- Cartrette, D. P., & Melroe-Lehrman, B. M. (2012). Describing changes in undergraduate students’ preconceptions of research activities. *Research in Science Education*, 42(6), 1073-1100.
- Chang, M. J., Sharkness, J., Hurtado, S., & Newman, C. B. (2014). What matters in college for retaining aspiring scientists and engineers from underrepresented racial groups. *Journal of Research in Science Teaching*, 51(5), 555-580.

- Corwin, L. A., Runyon, C. R., Ghanem, E., Sandy, M., Clark, G., Palmer, G. C., ... & Dolan, E. L. (2018). Effects of discovery, iteration, and collaboration in laboratory courses on undergraduates' research career intentions fully mediated by student ownership. *CBE—Life Sciences Education*, 17(2), ar20.
- Desai, K. V., Gatson, S. N., Stiles, T. W., Stewart, R. H., Laine, G. A., & Quick, C. M. (2008). Integrating research and education at research-extensive universities with research-intensive communities. *Advances in physiology education*, 32(2), 136-141.
- Dunitz, M. I., Lang, J. M., Jospin, G., Darling, A. E., Eisen, J. A., & Coil, D. A. (2015). Swabs to genomes: a comprehensive workflow. *PeerJ*, 3, e960.
- Eagan Jr, M. K., Hurtado, S., Chang, M. J., Garcia, G. A., Herrera, F. A., & Garibay, J. C. (2013). Making a difference in science education: the impact of undergraduate research programs. *American educational research journal*, 50(4), 683-713.
- Fechheimer, M., Webber, K., & Kleiber, P. B. (2011). How well do undergraduate research programs promote engagement and success of students?. *CBE—Life Sciences Education*, 10(2), 156-163.
- Furrow, R. E., Kim, H. G., Abdulrazak, S. M., Dahlhausen, K., Yao, A. I., Eisen, J. A., ... & Facciotti, M. T. (2020). Combining Microbial Culturing With Mathematical Modeling in an Introductory Course-Based Undergraduate Research Experience. *Frontiers in microbiology*, 11.
- Harrison, M., Dunbar, D., Ratmansk, L., Boyd, K., & Lopatto, D. (2011). Classroom-based science research at the introductory level: changes in career choices and attitude. *CBE—Life Sciences Education*, 10(3), 279-286.
- Hathaway, R. S., Nagda, B. A., & Gregerman, S. R. (2002). The relationship of undergraduate research participation to graduate and professional education pursuit: An empirical study. *Journal of College Student Development*, 43(5), 614-631.
- Hernandez, P. R., Schultz, P., Estrada, M., Woodcock, A., & Chance, R. C. (2013). Sustaining optimal motivation: A longitudinal analysis of interventions to broaden participation of underrepresented students in STEM. *Journal of educational psychology*, 105(1), 89.
- Hernandez, P. R., Woodcock, A., Estrada, M., & Schultz, P. W. (2018). Undergraduate research experiences broaden diversity in the scientific workforce. *BioScience*, 68(3), 204-211.
- Intemann, K. (2009). Why diversity matters: Understanding and applying the diversity component of the National Science Foundation's broader impacts criterion. *Social Epistemology*, 23(3-4), 249-266.
- Junge, B., Quinones, C., Kakietek, J., Teodorescu, D., & Marsteller, P. (2010). Promoting undergraduate interest, preparedness, and professional pursuit in the sciences: An outcomes evaluation of the SURE program at Emory University. *CBE—Life Sciences Education*, 9(2), 119-132.
- Linn, M. C., Palmer, E., Baranger, A., Gerard, E., & Stone, E. (2015). Undergraduate research experiences: Impacts and opportunities. *Science*, 347(6222).
- Lopatto, D. (2004). Survey of undergraduate research experiences (SURE): First findings. *Cell biology education*, 3(4), 270-277.
- Lopatto, D., Alvarez, C., Barnard, D., Chandrasekaran, C., Chung, H. M., Du, C., ... & Elgin, S. C. (2008). Genomics education partnership. *Science*, 322(5902), 684-685.
- Matthews, K. E., Adams, P., & Goos, M. (2009). Putting it into perspective: mathematics in the undergraduate science curriculum. *International Journal of Mathematical Education in Science and Technology*, 40(7), 891-902.
- Olson, S., & Riordan, D. G. (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Report to the President. Executive Office of the President.

- Oufiero, C. E. (2019). The organismal form and function lab-course: a new CURE for a lack of authentic research experiences in organismal biology. *Integrative Organismal Biology*, 1(1), obz021.
- Pacifici, L. B., & Thomson, N. (2011). Undergraduate science research: A comparison of influences and experiences between premed and non-premed students. *CBE—Life Sciences Education*, 10(2), 199-208.
- Petrella, J. K., & Jung, A. P. (2008). Undergraduate research: Importance, benefits, and challenges. *International journal of exercise science*, 1(3), 91.
- R Core Team, (2019). R: A language and environment for statistical computing.
- Rodenbusch, S. E., Hernandez, P. R., Simmons, S. L., & Dolan, E. L. (2016). Early engagement in course-based research increases graduation rates and completion of science, engineering, and mathematics degrees. *CBE—Life Sciences Education*, 15(2), ar20.
- Rodriguez-Medina, J., Kim, H. G., Castro, J., Contreras, C. M., Glon, C. L., Goyal, A., ... & Furrow, R. E. (2020). Draft genome sequences of 16 halophilic prokaryotes isolated from diverse environments. *Microbiology resource announcements*, 9(8), e01540-19.
- Russell, S. H., Hancock, M. P., & McCullough, J. (2007). Benefits of undergraduate research experiences.
- Seidel, S. B., Reggi, A. L., Schinske, J. N., Burrus, L. W., & Tanner, K. D. (2015). Beyond the biology: A systematic investigation of noncontent instructor talk in an introductory biology course. *CBE—Life Sciences Education*, 14(4), ar43.
- Seymour, E., Hunter, A. B., Laursen, S. L., & DeAntoni, T. (2004). Establishing the benefits of research experiences for undergraduates in the sciences: First findings from a three-year study. *Science education*, 88(4), 493-534.
- Shaffer, C. D., Alvarez, C., Bailey, C., Barnard, D., Bhalla, S., Chandrasekaran, C., ... & Elgin, S. C. (2010). The genomics education partnership: successful integration of research into laboratory classes at a diverse group of undergraduate institutions. *CBE—Life Sciences Education*, 9(1), 55-69.
- Shaffer, C. D., Alvarez, C. J., Bednarski, A. E., Dunbar, D., Goodman, A. L., Reinke, C., ... & Elgin, S. C. (2014). A course-based research experience: how benefits change with increased investment in instructional time. *CBE—Life Sciences Education*, 13(1), 111-130.
- Shortlidge, E. E., Banger, G., & Brownell, S. E. (2017). Each to their own CURE: Faculty who teach course-based undergraduate research experiences report why you too should teach a CURE. *Journal of microbiology & biology education*, 18(2), 18-2.
- Slovacek, S., Whittinghill, J., Flenoury, L., & Wiseman, D. (2012). Promoting minority success in the sciences: The minority opportunities in research programs at CSULA. *Journal of Research in Science Teaching*, 49(2), 199-217.
- Szteinberg, G. A., & Weaver, G. C. (2013). Participants' reflections two and three years after an introductory chemistry course-embedded research experience. *Chemistry Education Research and Practice*, 14(1), 23-35.
- Thiry, H., Laursen, S. L., & Hunter, A. B. (2011). What experiences help students become scientists? A comparative study of research and other sources of personal and professional gains for STEM undergraduates. *The Journal of Higher Education*, 82(4), 357-388.
- Thiry, H., Weston, T. J., Laursen, S. L., & Hunter, A. B. (2012). The benefits of multi-year research experiences: differences in novice and experienced students' reported gains from undergraduate research. *CBE—Life Sciences Education*, 11(3), 260-272.
- Torregrosa-Crespo, J., Pire Galiana, C., & Martínez-Espinosa, R. M. (2017). Biocompounds from Haloarchaea and their uses in biotechnology. *Archaea-New Biocatalysts, Novel Pharmaceuticals and Various Biotechnological Applications. InTech*, 63-82.

- Vater, A., Agbonavbare, V., Carlin, D. A., Carruthers, G. M., Chac, A., Doroud, L., ... & Coil, D. A. (2016). Draft genome sequences of *Shewanella* sp. strain UCD-FRSP16_17 and nine *Vibrio* strains isolated from Abalone feces. *Genome announcements*, 4(5), e00977-16.
- Vater, A., Dahlhausen, K., Coil, D. A., Anderton, B. N., Wirawan, C. S., Caporale, N., & Furlow, J. D. (2019). First-Year Seminars as a Venue for Course-Based Undergraduate Research Experiences: A Preliminary Report. *Bioscene: Journal of College Biology Teaching*, 45(2), 3-12.
- Vater, A., Mayoral, J., Nunez-Castilla, J., Labonte, J. W., Briggs, L. A., Gray, J. J., ... & Siegel, J. B. (2020). Development of a Broadly Accessible, Computationally Guided Biochemistry Course-Based Undergraduate Research Experience. *Journal of Chemical Education*, 98(2), 400-409.
- Wachsmuth, L. P., Runyon, C. R., Drake, J. M., & Dolan, E. L. (2017). Do biology students really hate math? Empirical insights into undergraduate life science majors' emotions about mathematics. *CBE—Life Sciences Education*, 16(3), ar49.
- Williams, L. C., & Reddish, M. J. (2018). Integrating primary research into the teaching lab: benefits and impacts of a one-semester CURE for physical chemistry. *Journal of Chemical Education*, 95(6), 928-938.
- Woolston, C. (2021). US SCIENCE STATIC ON DIVERSITY. *Nature* 592, 805-806.
- Yin, J., Chen, J. C., Wu, Q., & Chen, G. Q. (2015). Halophiles, coming stars for industrial biotechnology. *Biotechnology advances*, 33(7), 1433-1442.
- Zydney, A. L., Bennett, J. S., Shahid, A., & Bauer, K. W. (2002). Impact of undergraduate research experience in engineering. *Journal of Engineering Education*, 91(2), 151-157.

Chapter 4. Encouraging Collaborative Learning While Reading The Online Course Textbook With An Integrated Discussion Forum

4.1 Publication Status

Shusterman, E., **Kim, H. G.**, Facciotti, M., Igo, M., Sripathi, K., Karger, D., Segal, A. & Gal, K. (2021, April). Seeding Course Forums using the Teacher-in-the-Loop. In LAK21: 11th International Learning Analytics and Knowledge Conference (pp. 22-31). doi: [10.1145/3448139.3448142](https://doi.org/10.1145/3448139.3448142)

Abstract: Online forums are an integral part of modern day courses, but motivating students to participate in educationally beneficial discussions can be challenging. Our proposed solution is to initialize (or “seed”) a new course forum with comments from past instances of the same course that are intended to trigger discussion that is beneficial to learning. In this work, we develop methods for selecting high-quality seeds and evaluate their impact over one course instance of a 186-student biology class. We designed a scale for measuring the “seeding suitability” score of a given thread (an opening comment and its ensuing discussion). We then constructed a supervised machine learning (ML) model for predicting the seeding suitability score of a given thread. This model was evaluated in two ways: first, by comparing its performance to the expert opinion of the course instructors on test/holdout data; and second, by embedding it in a live course, where it was actively used to facilitate seeding by the course instructors. For each reading assignment in the course, we presented a ranked list of seeding recommendations to the course instructors, who could review the list and filter out seeds with inconsistent or malformed content. We then ran a randomized controlled study, in which one group of students was shown seeds that were recommended by the ML model, and another group was shown seeds that were recommended by an alternative model that ranked seeds purely by the length of discussion that was generated in previous course instances. We found that the group of students that received posts from

either seeding model generated more discussion than a control group in the course that did not get seeded posts. Furthermore, students who received seeds selected by the ML-based model showed higher levels of engagement, as well as greater learning gains, than those who received seeds ranked by length of discussion.

Selections from this work have been incorporated in this chapter, indicated by italicized text.

4.2 Author Contributions

ES, KG, HK, KS, MF, MI, and DK contributed to the conception and design of the study. ES, AS, KG, and DK contributed to the computational analyses. HK, MF, and MI provided labels for cognitive engagement. HK, MF, and MI provided preliminary input on seeding suitability. HK and KS provided technical support with seeding. MF served as the course instructor. ES performed statistical analyses. ES and KG wrote the initial draft of the manuscript. ES, AS, KG, HK, KS, MF, MI, and DK reviewed and edited the final draft.

4.3 Introduction

The CURE courses discussed in Chapter 3 represent just one way that education is reshaping at the curriculum level. Higher education is, in fact, continuously undergoing transformations as new technologies emerge and teaching strategies develop over time. In 2020, the COVID-19 pandemic forced higher education to re-assess the role of online technology in teaching and learning. Throughout the transition into, and in many cases, out of, remote and hybrid learning, educators have been reflecting on specific student needs that can potentially be satisfied in physical, remote, and hybrid educational contexts.

Textbooks are used ubiquitously as a medium for knowledge transfer in educational settings. Traditionally, textbooks took physical forms, but many publishers and independent authors, including instructors, have created or transitioned to online versions. One of the

benefits of the online text format is the possibility to create avenues for collaborative discussions, and add other features to improve engagement while reading the course text. Our research group seeks to make reading a more active learning process by integrating an annotation-based discussion forum tool called Nota Bene (NB) into the margins of the course material.

This chapter provides details about the motivations for and the potential benefits of the development of online collaborative learning technologies combining textbooks and discussion forums. Also discussed in this chapter is an educational research study that examines the effects of “seeding” a discussion forum with posts from previous course instances predicted to generate discourse or increase engagement.

4.4 The Rise Of Online Learning Modalities And Tools

Advances in technology have led to many changes in the way course content is managed and delivered and the ways students and instructors interact with each other. One of the earliest technologies associated with “online learning” was the Learning Management System (LMS), WebCT, which later became known as Blackboard (Singh, 2019). The Blackboard LMS serves as a structured virtual space that centralizes where and how instructors deliver course content, announcements, and/or information about assignments. Additionally, Blackboard supports online channels of communication between students and instructors by hosting discussion boards and private messaging.

Since the inception of Blackboard, various online learning tools and resources have been incorporated into LMSs. LMSs can serve various functions, including content management and course organization (Canvas, Blackboard), and assignment management and grading (Canvas, Crowdmark, Gradescope). Numerous other online resources that live outside of the LMS have also proliferated, many of which are offered at no cost to students as Open Educational Resources (OER). These may include open-access textbooks, along with other

instructional materials and tools, such as Ted Talks, YouTube, Khan Academy, NB, and LibreTexts, among many others (D'Antoni, 2009; Hilton, 2011). Common instructor motivations for using OERs are to expand access by reducing cost, while also potentially enhancing content and content delivery (e.g., Ted Talks, YouTube, LibreTexts), and enhancing student-instructor and student-student communication (e.g., NB, other discussion forums). The studies discussed in this chapter, as well as Chapter 5 and 6, focus on using NB to enhance student engagement and encourage interaction with content delivered through the OER LibreTexts.

4.5 The Development Of Online Textbooks

The textbook has been widely used in higher education since the availability of the printing press in the 1800s (Martin, 2017). A central function of the textbook is to organize relevant concepts and knowledge for a topic or discipline within a single catalog. Theoretically, textbooks can allow for wide access to disciplinary content, and can serve as a common reference, or even as a representation of standard knowledge in a field (Väljataga, 2014).

In addition to the book itself, textbook publishers have also created and packaged many instructional resources that can be used by the instructor to organize lectures, discussions, assignments, and other course materials. In this way, the textbook can sometimes act as an “intellectual core” around which courses are built. These functions and features make textbooks especially suitable for early-stage students, who may require more structure and may have expectations about the centrality of the book and its associated resources as defining the course of study.

With online technology becoming more widely adopted in both in-person and remote instructional curriculums over time, course textbooks have been undergoing digitization by commercial publishers. Since online technologies provide easier avenues to publishing broadly, there has also been rapid growth in open-access publications in an effort to provide free and open alternatives to students. This improvement in accessibility is important for decreasing

costs to financially burdened students, many of whom come from marginalized backgrounds. Martin et al. (2017) highlights the potential impacts of open, online textbooks in a survey of 676 students and 573 faculty at a large, private institution, where 66% of students had not purchased a textbook because of cost, and 91% of faculty were willing to use OER alternatives.

Other surveys and interviews have been conducted to explore how students and instructors used and perceived online textbooks to better understand their benefits, as well as their limitations. For instance, Bliss et al. (2013) surveyed students about which textbook format they preferred between online and print, 52% of students indicated a preference for the online text, 17% preferred the printed text, and 31% indicated no preference. In another survey conducted by Jung et al. (2017), faculty reported that extra time they may spend preparing to teach a course using open textbooks (i.e., tailoring the textbook for optimal use in their specific course) was acceptable, given the many advantages they provide, which include but are not limited to: 1) the immediate and ubiquitous access to textbooks for students, 2) the ability to keep content up to date, 3) interactivity potential (e.g., hyperlinked definitions for vocabulary terms, guided exercises), 4) better alignment with class activities, 5) higher textbook quality, 6) convenience for faculty, and 7) the ability to customize and modify content. Wiley (2014) additionally notes that the growing numbers of OERs translates to an increasing number of options of OERs to “retain, reuse, revise, remix, and redistribute” materials as needed. Aside from providing potential benefits to the general student body, the ease of configuring digital texts can be especially beneficial to instructors who wish to make adjustments for students who need various learning accommodations.

Despite the many advantages that open-access online textbooks provide to both students and instructors, there are still known limitations (Table 4.1). Studies have shown that students across many disciplines choose to not read the textbook or do so only sparingly, for both physical and online formats. For instance, Podolefsky et al. (2006) showed that less than 40% of students read their physical introductory physics textbook assignments at one institution,

and that only 55% of students at another institution engaged with the required submitted exercises associated with the reading. Gyllen et al. (2018) studied how students engage with an online textbook in an engineering course, and found that students spent minimal time on the instructional text and worked out examples, instead focusing more on practicing test problems. Furthermore, the act of reading, whether a physical or open/online text, can still be a solitary and isolating experience that could be improved. To mitigate this and promote greater engagement and active learning, attempts have been made to integrate learning technologies (e.g., dictionary tools, interactive exercises, collaborative spaces) into online course readings.

Table 4.1. Summary of the individual and shared limitations of traditional and digital textbooks.

Limitations of traditional textbooks	Limitations of digital textbooks	Shared limitations between physical and digital textbooks
- Costly for students - Not easily modifiable - Not as flexible or adaptable to fit certain course structures - Less potential for disabilities accommodations	- Requires technology and technology skills, which not all students have access to and not all teachers have - May require significant upfront instructor time and energy investment - Concerns about or difficulties finding suitable content	- Studies have shown that many students do not read the course text - Reading the course text is often a solitary experience, with discussions taking place outside of the reading process

The evolution to online textbooks has provided opportunities to bring in new ways of interacting with the course content. Unlike the static textbook, content online can potentially be enhanced with interactive and “smart” technologies. Capota and Severin (2011), for example, describe a Korean initiative to replace traditional textbooks with interactive and customizable online textbooks integrated with additional elements of interaction, such as animations and virtual reality.

One of the technologies explored in this thesis are tools that promote collaborative and social learning through discussion. Social annotation software, such as NB,³ Hypothesis,⁴

³ <https://nb.mit.edu/welcome>

Perusall,⁵ have been developed, allowing students to highlight sections of course text hosted on HTML web pages, PDFs, and EPUBs and make notes and reply to other students' notes in the margins within a separate panel of the web browser. The fact that technology can help facilitate discussions is important, as feedback and discourse can help resolve confusion around a topic, build knowledge, and enhance student learning.

4.6 The Significance Of Collaborative Or Social Learning

One of the core components of education is an exchange of ideas and feedback between students and instructors – a function of instruction that provides an opportunity for important, and potentially higher-level, assessment of and adaptation to student needs. Technological developments are allowing for more interactions to occur in online learning spaces, also known as Computer-Mediated Communication (CMC). A popular feature of many student-focused learning technologies are online discussions or forums. Online discussion forums are a prominent feature in asynchronous learning environments, such as MOOCs, and are also used abundantly in traditional, in-person courses. In many cases student participation in discussion is encouraged by the assignment of points, but this need not always be the case. In general, discussion forums serve as class-external opportunities for students and instructors to hold conversation on course topics in the form of questions, replies, reflections, or summaries.

Numerous studies provide evidence that collaborative work is beneficial for students, with regard to achievement, effort, persistence, and motivation (Johnson, 2007). Taylor (1997) compared a socially-assisted learning group to a control group, and found higher learning achievements resulting from peer collaboration, reflective questioning, shared ownership, and feedback. Atwater (1996) and Barak (2017) further found that discourse among learners and

⁴ <https://web.hypothes.is/>

⁵ <https://perusall.com/>

collaborative assignments led to improved scientific literacy and “higher order thinking.” Vellukunnel (2017) reports that in their computer science course’s discussion forum, “the largest portion of questions reflect some level of constructive problem-solving activity, and are positively correlated with course grades.” *Discussion forums have been used successfully as tools to facilitate interactions and exchanges of knowledge between learners and between learners and instructors (Poole, 2000). Promoting student discussion in online forums has been linked to increased learning gains (Cormier, 2010, DeSanctis, 2003). Online discussion forums have also been shown to promote an increased exchange of ideas, as well as improve students’ ability to make connections between concepts and to apply the course material to diverse contexts (Breslow, 2013).*

Several theoretical frameworks also speak to the potential importance of discussion. The social constructivist theory supports the idea that interactions between students and their peers help to co-create identities and norms, and develop a shared sense of goals and purpose (Barak, 2017). These ideas were influenced by Vygotsky’s (1978) sociocultural theory, which suggests that a person cannot independently activate or achieve the higher mental functions that arise from social interactions (both spoken, and written word). Bell (1985) builds upon this theory further, suggesting that cognitive conflict that occurs in social settings is beneficial as it allows students to gain awareness of the relativity and potential weaknesses of their ideas, as well as the practice of communicating and negotiating during the knowledge acquisition process. *However, not all forms of discussion are beneficial for learning (Romeo, 2001). Online discussions that do not promote higher levels of thinking are ineffectual in providing increased learning (Wang, 2015). Studies have also shown that pedagogical benefit arises only when discussion encourages students to share different interpretations and perspectives of the course (Light, 2000).*

As mentioned previously, various social annotation software have been developed that facilitate discussions amongst students and instructors during the process of reading the course

text and other educational content (e.g., videos, external readings and webpages), such as NB, Perusall, and Hypothesis. The next sections detail educational research efforts conducted in collaboration between UC Davis, Massachusetts Institute of Technology, and Ben Gurion University of the Negev, using NB along with the open-access online textbook, LibreTexts,⁶ in a general biology course. Taking into account that not all discussions may be beneficial for students, the NB team proposes *to initialize (or “seed”) a new course forum with comments from past instances of the same course that are intended to trigger discussion that is beneficial to learning.”*

4.7 Introduction To Nota Bene

Nota Bene is an open-access, collaborative annotation software developed by the Haystack Group at MIT Computer Science & Artificial Intelligence Lab in 2009.⁷ The course materials (PDF, HTML, video file) are uploaded onto the NB website, which can be accessed by students and thereafter annotated. LibreTexts, founded at UC Davis in 2008, is an open-access educational platform which hosts roughly 400 texts in over 150 courses. When NB and LibreTexts are used together, the NB annotation interface appears as a panel on the right side of the screen, with the course text hosted on LibreTexts on the left side. When a portion of the text is selected and highlighted in the left-panel, a text field appears in the right-side NB margin, where the student can type their annotation (Figure 4.1). The student can view their and their peers’ publicized annotations above the text field on the right-side panel as well. When a student responds to another, the discussion is organized into a thread, and the number of replies within a thread is displayed in the thread preview.

This annotation software serves as a centralized repository where students can post original annotations as they read, as well as collaboratively annotate and help solve questions

⁶ <https://libretexts.org/>

⁷ <https://www.csail.mit.edu/>

or hold discussions related to the course with their peers and instructors. This allows for students to engage in social and collaborative learning while helping to eliminate redundant questions, freeing instructor bandwidth for other course needs or improvement.

Figure 4.1. Nota Bene used in combination with LibreTexts.

The NB team studies three types of engagement to better understand how students interact with the course text using NB: 1) emotional engagement, how students feel about what they are interacting with online, 2) cognitive engagement, how deeply students are thinking about the material at hand, and 3) temporal engagement, how much time students are spending interacting with different parts of online content. The data that student comments provide is important for course instructors' understanding of the student state and types of engagement while learning – and with NB being used with LibreTexts, the student reflections can be mapped to distinct parts of the course text. With the relevant analytic tools or skills, instructors can take this information and use it to improve the course text, help shape class design and curriculum, and inform interventions to improve student learning and engagement.

4.8 Using Collaborative Nota Bene Features As Motivation For Students To Engage With The Course Text Before Class

There are known benefits of reviewing course materials (e.g., reading the course text and completing pre-lecture assignments, such as chapter quizzes) before class, including: being better able to follow material in class (Schwartz, 1998), asking more meaningful questions in class (Marcell, 2008), and higher performance on assessments (Dobson, 2008; Johnson, 2009; Miller, 2018). Despite this, studies such as the ones mentioned previously have shown that students across many disciplines opt to not or only sparingly read the textbook.

“Just in time teaching,” or JITT, has been proposed as a method to increase motivation for reading the textbook via grade incentives. With JITT, students are required to answer open-ended questions about the assigned reading before class, including questions that are solely dedicated to understanding where students had the most confusion (Novak, 1999). However, studies with traditional, physical textbooks show mixed results: Stelzer (2009) showed that 70% of students still never or rarely read the textbook for class with JITT structures, meanwhile, Heiner (2014) showed that JITT-like implementations led to 79% and 85% of students in their classes reading the assigned sections and completing the associated pre-class reading assignments which accounted for a small proportion of their total grade. Sappington (2002) notes that self-reported compliance ratings from students can be distorted.

Proponents of online, social textbook annotation suggest that by enabling, or even using points to incentivize discussion as in JITT, students may be more inclined to engage with the course text prior to class. Students may perceive more benefit from interactive discussions rather than individual exercises if the benefits of discussion are made clear. First, students may ask questions and receive responses from either their peers or instructors. Second, instructors may make it evident that the student discussions are helping to guide course pace, giving the student buy-in and initiative to participate.

In addition, we hypothesize that seeding discussion forums with suitable “stimulating” posts from previous academic terms can improve students’ learning gains by encouraging engagement in the forum. The next section summarizes the study that tests this hypothesis.

4.9 Encouraging Deeper Engagement With Course Readings By Seeding Discussion Forums With Pre-Selected Posts

In our work, we study how to improve the quality of online discussions by initializing (or “seeding”) course materials with comments from previous iterations of the course. This idea provides a practical intervention for instructors to incorporate prior to posting reading assignments, which can amplify the value and quality of student interactions in online forums. To scale this approach to large classes, we developed machine learning tools to select posts likely to stimulate discussion in future classes.

Here, we show how the instructor can leverage these tools to seed online discussions. We assess the influence of this approach on students’ learning outcomes and compare students’ interactions with the seeds. We evaluate this process in the context of a 186-student introductory biology course, using NB. Specifically, we wish to address the following questions:

- 1. Can seeding change students’ forum behavior, in terms of the number of participants and quality of discussion in the given thread, as well as students’ engagement with course materials?*
- 2. Is the changed behavior linked to better learning as defined by performance on course assessments?*
- 3. Are some seeding methods better than others at eliciting behavior change that leads to improved learning?*

Our methodology consisted of three steps: First, we designed a scale for determining the “seeding suitability” score of a given thread (an opening comment and its ensuing discussion). The scale takes into account the length and quality of the discussion, its relevance to the course material and learning objectives, as well as the level of engagement exhibited by students in the discussion.

Second, we constructed a supervised machine learning (ML) model for predicting the seeding suitability score of a given thread. The model combines lexical, emotive, and cognitive features that were extracted from students' forum interactions. We demonstrate the efficacy of the model by comparing its performance to the expert opinion of the course instructors on test/holdout data.

Third, we conducted an experiment in a live course. The ML model was used to select threads from a previous course instance. For each reading assignment in the live course, we presented a ranked list of seeding recommendations from previous courses to the live course's instructors, who could then filter the list for inconsistency and misinformation. The chosen list of seeds was subsequently seeded in the forum at the onset of the reading assignment.

We ran a randomized controlled study with three groups of students: a) students not exposed to seeds, b) students shown seeds recommended by the ML model, and c) students shown seeds recommended by an alternative model that ranked seeds purely by the length of discussion that they generated in previous course instances. For each group of students, we measured the impact of seeding on students' forum behavior, as well as on their learning gains in the course.

We found that discussion prompts from the seeding models generated more discussions than random discussions within the Control group. Furthermore, students who received seeds selected by the ML-based model exhibited higher levels of Cognitive Engagement in their replies to seeds than those who received seeds ranked by thread length of previous discussion. In particular, we found a causal relationship between receiving seeds in the ML-based approach and improvement in learning gains. Our results highlight the benefit of using an ML approach to augment instructors' abilities to improve students' learning when using online course forums. The ML scales up the selection process of candidate seeds in both numbers and speed, thereby providing the instructor with a manageable list of seeds for further evaluation and subsequent inclusion in the course readings.

4.9.1 Related Work

4.9.1.1 Interventions For Improving Learning In Online Forums

We were inspired by Miller et al. (2014), who studied the effect of seeding prior-semester comments in a physics class which contained several annotated discussion readings. They manually assessed students' comments based on a 3-point "quality" scale. Annotations lacking meaningful physics concepts received a quality score of 0 while factual, definition type annotations received a quality score of 1. Annotations that justified questions or explanations with substantiated physics concepts received the maximum quality score of 2. As a seeding schema, for each reading, the authors randomly selected 10 first-posts that obtained a quality score of 2 as seeds. In their experiment design, different groups of students were exposed to the seeding treatment in different readings, such that all the students were exposed to the treatment by the end of the experiment. Miller et al. (2014) compared annotations from students in seeded sections versus unseeded sections by their quality measure, and by an adaptation of (Hogan, 1999)'s scheme developed to examine discourse patterns and collaborative scientific reasoning in peer discussions. The researchers found that students in seeded sections produce longer threads, higher quality annotations, and a greater proportion of generative threads than unseeded sections. We extend this work by building on an ML based approach for seed selection (as opposed to the manual approach used by the authors) and by measuring the seeding impact on learning gains. This is in contrast to previous work that focused only on student discourse quality. In particular, our computational model identifies seeds that lead to perceived educational benefit.

4.9.1.2 Modeling Student Discussions

Previous studies have used a variety of models to analyze students' discussion behavior in forums and inspect the structure and quality of the discussion. In general, these works have

focused on analytics rather than intervention. Weimer and Gurevych (2007) proposed a domain-independent system for automatic quality assessment of forum posts that learns from human ratings. They constructed a computational model that uses several families of features, including surface-level (e.g., length, question or exclamation frequency), lexical (e.g., frequency of spelling errors), and similarity to the topic of the forum. They applied their approach to different web forums, highlighting situations that challenge the model, such as very short posts. Kim et al. (2006) modeled online student discussion as a series of speech acts and investigated dependencies among the messages using a set of relational dialogue rules. They identified topics discussed in threaded discussions and assessed whether the topics shift or remain focused within the threads. They found that students who participate more and elicit more messages tend to receive better grades in the course.

Feng et al. (2006) described a system for detecting the conversation focus of threaded discussions and finding the most authoritative answer in a thread. Their model included heterogeneous evidence from different sources, such as lexical similarity, the trustworthiness of the student writing the post, and speech act analysis of human conversations. They represented discussion threads as graphs and made use of the speech act relations to generate the links. They used this model to detect which message in a thread contains the most important information (i.e., the focus of the conversation). Several works have analyzed online forum discussions in terms of Cognitive Engagement, which is a measure to the interaction depth between the student and the course material, and has been shown to correlate with learning gains (Wang et al., 2016, Yogev et al., 2018). These works developed classification models for determining the level of Cognitive Engagement in a post. Yogev et al. (2018) showed that a visualization of Cognitive Engagement anchored in the text can give teachers valuable insight into their students' thinking and can help guide modifications of lectures and course readings to improve learning. Geller et al. (2020) designed a computational model for detecting confusion based on rules inferred from students' hashtags (which are used to convey emotions in student

posts). They showed that students' self-reported hashtags may not agree with experts' judgement about what constitutes a confused post. The authors designed computational models for automatically detecting confusion in posts that combine the perspectives of both students and experts (See Chapter 5). All of the works above combine discourse analysis of discussion threads by experts with a supervised learning model to achieve different tasks. Although none of these works use seeding as an approach, they provide important information about the aspects of online discussions that appear to have benefited learning outcomes. These works also informed our feature design process, in particular our decision to include cognitive and emotive factors in our analysis.

4.9.2 Using Nota Bene In The Context Of Seeding Experiments

The reading materials for each lecture are created by the course instructors and are hosted on LibreTexts, an online repository of open educational resources (libretexts.org). Students in the course receive reading assignments on the material uploaded to NB. The students are required to make at least three meaningful posts to the forum for each reading assignment before each class lecture. Students can satisfy this requirement by either opening a new thread or responding to another post in an existing thread. Students are also given credit for conveying their subjective opinions about the reading material by "tagging" one or more of their comments with a predefined hashtag(s) (specifically, #useful, #confused, #curious, #frustrated, #idea, #question, #help, #interested). Students receive points for this term on NB participation that count roughly 10% towards their final course grade.

We obtained data of students' forum interactions in four instances of the course: 1) Winter 2018, 2) Summer 2018, 3) Summer 2019 and 4) Winter 2020 terms. For all four course instances, both lecture and discussion sections were taught in-person. Data from these course instances were used to train our developed models. The Summer 2020 session of the course, in which we carried out the online seeding experiment, had 186 students and was taught online,

due to COVID-19. Table 4.2 shows the number of students and posts collected in each course instance. Each reading lecture in the course contains a set of instruction-defined learning objectives that the students are expected to master and the reading material contains passages relating to these learning objectives. We used pretrained Bidirectional Encoder Representations from Transformers (BERT) (Devlin, 2018) — based on BooksCorpus (800M words) and on the English Wikipedia (2,500M words) — to learn contextual relations between students’ comments and the course learning objectives. We consider a comment to relate to a given learning objective if the respective cosine similarity metric between their BERT embedding (Devlin, 2018) is at least 0.75. This threshold was determined in agreement with the course instructors, after observing 100 examples from the datasets of Table 4.2.

Table 4.2. Course and forum statistics.

	Winter 2018	Summer 2018	Summer 2019	Winter 2020	Summer 2020
Lectures count	25	25	15	25	26
Num. Students	714	781	118	940	186
Num. Threads	60,469	46,315	7,278	48,228	12,746

Table 4.3 shows examples of the learning objectives for reading lecture 2 of the Summer 2020 term and the numbers of posts in the lecture that relate to the learning objectives (some posts may match more than one learning objective). As shown by the table, some learning objectives are popular and dominate discussions, while others receive little attention from posts. Many posts (over a third in this lecture) do not relate to any of the learning goals, which is not captured in Table 4.3. If discussions were motivated to be more focused on learning objectives, this may improve students’ learning in the course.

Table 4.3. Learning objectives for a selected reading assignment in the Summer 2020 term and counts of related NB posts.

Learning Objective	Num. of posts matching learning obj.
<i>Identify the locations of the three core atomic elements (electrons, neutrons, and protons) from a basic model model for an atom and describe their basic properties</i>	183
<i>Recognize the symbols and names of the six core elements in biomolecules: C, H, N, O, P, S.</i>	68
<i>Interpret (identity elements and bond types) chemical and structural formulas (including 2 and 3-dimensional, of condensed, and line-angle) biomolecules</i>	26
<i>Define electronegativity and explain how this concept can be used to predict the types of bonds that may be formed by two atoms.</i>	490
<i>Identify ionic, covalent, polar covalent bonds, hydrogen bonds, and Van der Waals interactions in different types of molecular models.</i>	208

4.9.3 Selecting Threads For Seeding

Our work is predicated on the hypothesis that it is possible to choose discussion seeds from previous semesters that can stimulate discussion and learning among current students. We therefore developed a process to help faculty train and then use a machine learning system to select good seeds from a large repository of past discussions. When reviewing students' posts, course instructors consider the following questions when deciding if a first-post may lead to a beneficial discussion for learning:

1. To what degree does the post relate to the learning objectives of the given lecture?
2. How much apparent effort went into writing the post, beyond the minimum requirement to get credit?
3. Does the post raise a new idea/question that goes beyond the content presented in the reading material?
4. How many responses did the post generate and how do the responses relate to the reading material, course objectives and others' comments?

Based on the above, we defined the seeding score of a thread in NB as the degree to which seeding the first post in the thread would benefit learning in the course. The seeding score is measured on an ordinal scale between 0 (not suitable for seeding) and 3 (most suitable for

seeding). Two of the course instructors composed a list of general guidelines for determining the seeding score for different first-posts.

An example of these guidelines are shown in the right-most column in Table 4.4. The two course instructors established agreement over the labeling process by labeling 240 threads sampled from the 2018, 2019 and Winter 2020 course instances. The Kappa Interrater reliability was 0.44 (McHugh, 2012), and the instructors proceeded to resolve all of the disagreements. An additional 400 threads were subsequently labeled by one of the course instructors. The distribution of the seeding suitability score over the 640 threads is shown in Table 4.4. The average seeding suitability for the threads was 1.56, the median score was 1, and the standard deviation was 0.8. The low suitability reflects the need to improve the levels of discussion in the course forum. Table 4.5 shows a few examples of first-posts and their associated seeding suitability score, alongside the number of responses the posts generated. As shown by Table 4.5, there is a weak relationship between the length of the discussion and the associated seeding score for different threads.

Table 4.4. Distribution of seeding scores of first-posts sampled from the 2018, 2019 and Winter 2020 course instances, with examples of classification guidelines used by instructors.

Seeding score	Amount	Percent	Example of classification guidelines for score
3	98	.153	<i>The first-post encourages deeper understanding and critical thinking surrounding a learning objective or contains unique question/idea that brings a new perspective/discussion</i>
2	193	.301	<i>The first-post is related to a learning objective or the thread contain an evidence that the first-post generated productive discussion</i>
1	320	.500	<i>The first-post has a weak connection to a learning objective or contains a misconception/oversimplification</i>
0	29	.045	<i>The first-post answers a question presented in the text meant for students to answer, or detail about topics beyond the scope of class</i>

Table 4.5. Examples and justification for seeding scores.

First-post	Num. Responses	Seeding score	Justification
<i>"If energy is not made or produced, then is food the only way for our bodies to get energy? Are there any organisms that can make their own energy, if that is even possible? #question"</i>	1	3	<i>This question forces students to think about the inputs and outputs of the system, which is good for building mental models about energy.</i>
<i>"The larger the electronegativity of an atom allows the atom to attract other atoms. The most electronegative elements include Chlorine, Fluorine, Oxygen, and Nitrogen."</i>	4	1	<i>Very basic concept without anyone in the thread even asking questions.</i>
<i>"My past courses focused on ecological biology, it is refreshing to finally learn molecular biology in this course. #useful"</i>	15	0	<i>The discussions are about course logistics.</i>

4.9.4 Computational Models

In (Shusterman, 2021), we describe and compare different computational models for determining seeding suitability for the first-post in a thread. The input to a model is a given reading lecture in the course, and a set of candidate threads from past course instances whose posts relate to the reading materials for the lecture. The readings in each course instance are mostly identical so we can take comments from different semesters within the same reading. The output of the model is an ordered ranking of the threads in decreasing order of their predicted seeding score. The models used several different families of features.

Model 1 – Lexical and Similarity-based Features:

This family included features relating to the counts of lexical elements in the first-post, such as the number of words in the post, presence of question mark or #question hashtag, and whether it contained interrogative words (e.g., how, why) and comparative adverbs (e.g., worse, less). We also included a feature capturing the relationship between the course's learning objectives and the first-post. The feature value was the highest similarity between a first-post and any learning objective for the reading lecture. The similarity was computed as the cosine

similarity between the BERT embedded representations of the learning objectives and the first-post of a candidate threads, similar to Geller et al. (2020).

Model 2 – Engagement Features:

This family used features that build on the Cognitive Engagement (CE) taxonomy, which was developed by Chi (2009) to assess how deeply a student is interacting with the course material. Chi’s work showed that higher cognitive engagement in students leads to learning gains. In addition, other works showed that posts with high CE levels are correlated with increased learning gains (Miller, 2016, Wang, 2015, Yogev, 2018). Thus, we posited that discussions demonstrating higher cognitive engagement might make more effective seeds that could lead to other high-engagement discussions and learning gains. We introduced some features with this goal in mind. One feature directly measured the Cognitive Engagement (CE) of posts within the candidate thread.

Posts were labeled using three levels from the CE hierarchy, in increasing levels of engagement, following the taxonomy used by Yogev et al. (2018). A post is labeled Active (A) if it refers to specific course materials in the annotation by paraphrasing or repeating, but does not provide new insight (e.g., “What is an isoelectric point of an amino acid?”). A post is labeled Constructive (C) if it displays reasoning, includes a new idea, or refers to external sources — and is not a response within a series of interactions (e.g., “The answer is c. The amino acid sequence, peptide bonds and length of the protein are derived from the primary structure...”) A post is labeled Interactive (I) if it displays Constructive reasoning, but also builds upon, or challenges another post’s ideas (e.g., “Yes, I agree with you, if the interactions between the sheets change, it may cause the protein’s overall structure to change thus the function [...]”).⁸ The CE level of a given post must be inferred from the text of the post and other posts in the thread. To this end, we designed a two-stage binary classification model, first distinguishing

⁸ Chi’s taxonomy includes a fourth, Passive category, but comments on the course material cannot be passive by definition.

between A and C/I level posts and then distinguishing between C and I level posts. Each of the binary classification models used a Long Short Term Memory (LSTM) network structure to capture the temporal sequence exhibited by posts in the thread. The model was trained on 5,206 instances from the data sets of Table 4.2, tagged for CE by course experts. We evaluated the model using a cross validation process that respected the temporal relationships in the data. The model achieved Precision of 0.77, Recall of 0.76, and F1 score of 0.76, which was higher than the state-of-the-art model (Yogev, 2018) which did not use deep learning approaches.

We included a feature in the seeding model that used the CE predictions to compute an average CE score for a thread. We converted a categorical scale (I, C, A) to a numeric scale by assigning points of 5, 3, and 1 respectively to each value. We computed a simple average of these points according to the thread discussion. An example of threads with a high number of I level posts that achieved a weighted average CE score of 0.8 is presented in Table 4.6. We note that there were extremely few A level responses in the dataset as responses generally relate to the previous comment. Other features in the seeding model that related to CE included the number of posts in the thread exhibiting I level CE, as well as identifying that the maximum number of consecutive posts exhibiting the highest I level CE.

Table 4.6. Example of a thread with a weighted average CE score of 0.8.

Post Position	Post Text	CE Label
1 st post	"Why do certain cells stay in G0 permanently? How does this affect us, since such vital cells in our body are never dividing? #question"	A
Response 1 to 1 st post	"I believe this is because the cells that stay here, such as mature cardiac muscle and nerve cells, are not actively preparing to divide. Their goal is not to continuously divide and make new daughter cells immediately, but rather to hold off until a signal triggers their division. #useful"	I
Response 2 to 1 st post	"It makes sense that nerve cells do not divide. All the nerve cells we will ever have are usually in place by birth. This is because nerve cell development and brain development depends on the connections that nerve cells make with other nerve cells. #interested"	I
Response to Response 2	"If it were possible to turn this phase off for nerve cells and let them replicate on specific comment would it be possible to fix nerve damage in people? #question"	I

Another feature for describing engagement measured the likelihood of the first-post to trigger a future discussion. To this end, we trained a logistic regression classifier using BERT embedded representations of first-posts and responses from the datasets in Table 4.2.

Model 3 – Emotive Features:

This family of features reflects subjective emotive information in the discussion that is conveyed by the students in the form of hashtags (see section 3) and supported by the NB GUI.

We hypothesized that expressing emotions may invite other students to participate in the discussion and can indicate good threads for learning. We explored the correlation between different hashtags occurrences count in a thread discussion and the target variable (seeding score) and found that only specific hashtags have a significant correlation: #curious and #frustrated. Therefore, we focused on those hashtags alone as well as whether the post expresses confusion. We note that the educational literature shows that students' posts may reflect confusion without self-reporting this hashtag, for example, when they are stating something incorrect that they seem to think is correct (Plaut, 2006). To this end we used the classifier by Geller et al. (2020), which extracted features from the post and the selected text in the reading material to infer instances of student confusion that go beyond the use of #confused hashtag.

The most important features with positive contribution to seeding score were whether the first-post was a question, the response probability, and whether or not the longest sequential CE level in the thread discussion is interactive (as in Table 4.6). The most important features with negative contributions to seeding were whether the first-post expressed an incorrect statement without asking a question, the percentage of posts in the thread with fewer than 4 words, and the number of #frustrated hashtags in the thread discussion.

4.9.5 Online Seeding Experiment

In this section, we apply the computational model described in the previous section in an active course. We focused on the summer 2020 instance of Bis2A, which was conducted online due to COVID19. The course included 26 reading lectures, two midterm exams (after reading lecture 10 and after reading lecture 18) and a final (after reading lecture 26).

Students enrolled into a unique recitation/discussion section (20-24 students), each of which is moderated by a graduate student teaching assistant (TA). Each section used its own instance of NB and did not see the comments of the other sections. In the first half of the course (readings 1-12) we did not publish any seeds. The first midterm (hence referred to as Midterm 1) following reading lecture 10 was used as a baseline for comparing course performance between the different groups. Seeding commenced in the second half of the course (from reading lecture 13 to reading lecture 26). This permitted a difference in differences analysis, in which we compared the impact of a change at midpoint from the control condition to seeding.

The study was reviewed by the IRB of the hosting institution and deemed exempt. All students who participated in the online study in summer 2020 filled a consent form and were requested to opt-in the study.

4.9.5.1 Experimental Methodology

Our seeding methodology consisted of four steps: (1) For each reading lecture in which seeding is performed, select a set of candidate seeds from the first-posts of the relevant lecture in past instances of the course. All of the candidate seeds were selected from past instances of the course in 2018, 2019, and Winter 2020 from Table 4.2. (2) Rank the candidate seeds according to a scoring function. (3) Present a ranked list of the top scoring 15 seeds to the course instructors and allow the course instructors to review the ranked list to select a subset of

ten seeds for publishing. (4) Publish the seeds in the course forum in the relevant reading lecture. We expand on each of these steps in turn.

Step 1: The instructors selected a subset of learning objectives that fulfilled the following criteria: Each learning objective was related to readings in the first and second half of the course; as well as to questions in midterm 1 and the final exam. For each of these learning objectives, we selected the 50 first-posts with the highest similarity to the learning objective, computed using BERT embedding, and added these posts to the candidate seed set. As a preprocessing step, we removed threads whose first-post met one or more of the following: 1) selected text in the readings that was not in the given lecture reading, 2) did not contain any text, or 3) only contained a hashtag. In practice, there were between 200 and 250 candidates for each lecture.

Step 2: We used two different approaches to rank the candidate set of seeds from the previous step. The ML-supported approach used the regression model described in the previous section to rank candidate seeds according to their predicted seeding score. The Long approach ranked seeds according to the length of the discussion that they generated, in terms of number of responses. We selected this baseline approach following previous studies that measured thread quality by the number of responses (Kim, 2006). In addition, The Long approach does not require any AI, and is simple to compute. Both approaches broke ties by ordering with respect to the semantic similarity between the seed and the learning objectives of the reading lecture. To ensure that the selected seeds for the lecture were all consistent with a single poster, both approaches removed seeds whose marked text in the reading material overlapped with a higher ranking seed.

Step 3: The course instructors inspected the ranked list of 15 seeds for each seeding group using their expertise. For example, instructors identified seeds with inconsistencies or contradictions with another seed, in which case the lower ranked seed was removed. Instructors also removed seeds that explicitly referenced previous course instances. Grammar correction

was purposely not done. In practice, 38 of 180 seeds were rejected by the instructor in the ML condition and 34 of 180 seeds were rejected in the Long condition. The ten highest scoring seeds that passed the instructors' inspection were subsequently planted in the NB forum. An example of contradicting seeds is presented in Table 4.7.

Table 4.7. Examples of contradicting seeds.

Seed Text	Predicted Seeding Score
"It is clear how the membrane comes to be charged, but not what purpose the polarized membrane serves. does the polarization help to move the electrons along the membrane? this however wouldn't make sense if it is the electrons that cause this polarization. #confused"	2.15 (accepted)
"Basically, the proton motive force works as a tiny battery. it's energy can either be stored for later use in atp or can be used immediately to do work like powering the movement of flagella. this proton motive force occurs when the embedded electron carriers in the cell membrane become energized as a result of the electron transport reactions."	1.65 (rejected)

Step 4: The chosen ten seeds were manually inserted in NB to coincide with the time that the reading materials for the following lecture was published. Informing the students that the seeds were selected by the course staff could bias their response. There is evidence showing that students respond more positively to questions posed by fellow students (Mazzolini, 2003). Therefore, all published seeds were tagged with a fictitious student account assigned with the gender, race, and ethnically ambiguous name "Lee Jordan." The distribution of the location of the seeds in the reading material for the lecture was similar in both seeded conditions, in that most of the threads were positioned in the first 50% of the lecture. This trend was also exhibited by students in past course instances.

Our methodology was evaluated using a randomized control study. Discussion sections in the course were assigned to different experimental conditions. Students in the ML-seeded condition (59 students) and the Long-seeded condition (59 students) were assigned seeds

ranked by the ML-supported and Long approaches, respectively. Students in the Control condition (68 students) were not exposed to seeds.

4.9.5.2 Hypothesis And Results

In this section, we evaluate the effect of seeding on students' learning gains in the course. Before stating our main hypothesis, we define key terms. First, we use the term "seeded learning objectives" to refer to those learning objectives that were selected by the course staff for seeding in both ML- and Long-seeding conditions. All of the seeds planted in the study related to at least one of the seeded learning objectives. Second, we define the set of "relevant" questions in an exam as those questions that are related to the seeded learning objectives. There were 7 relevant questions in midterm 1 and 26 relevant questions in the final exam. Our hypotheses were as follows:

Learning gains hypothesis: *Students in the ML-seeded condition will incur higher learning gains than students in the Control condition, as measured by their performance in relevant questions in the final exam. The null hypothesis states that the learning gains in the ML-seeded condition will not be higher than in the control. To address this hypothesis, we first compared the success rate (the ratio of correct answers) on relevant exam questions between students in both ML- and Long-seeding conditions. We did not find a significant difference between the ML-seeded condition and the control in midterm 1 which occurred prior to the beginning of treatment (Repeated Measure ANOVA test ($F(2, 12) = 2.240$, $p = 0.149$)). This implies there was no difference in students' knowledge between the different conditions prior to beginning the seeding intervention.*

We found a significant difference between the condition groups (ML, Long, and Control) in the final exam (Repeated Measure ANOVA test ($F(2, 50) = 3.818$, $p < 0.03$)) after the treatment. A post hoc test revealed that the average success rate for the ML-seeded condition (0.69, $STD = 0.21$) was higher than that of the Control condition (0.64, $STD = 0.20$) ($p < .03$).

Thus, we can reject the null hypothesis in favor of the learning gains hypothesis. We do note that the average success rate of the ML-seeded condition was not significantly higher than that of the Long-seeded condition (0.66, STD=0.21) ($p = .09$)

*We also wished to determine whether there exists a causal effect of seeding for both seeding conditions. To this end, we used the difference-in-differences approach (Lechner et al., 2011). The independent variables for the model were as follows: Treatment (whether seeding or control), Time (whether midterm 1 or final exam), Treatment*Time (1 when Treatment is seeding and Time is the final exam; 0 otherwise). The dependent variable was the students' success rate in the relevant questions in midterm 1 or the final exam, depending on the value of Time. We controlled for the effect of a confounding variable (the identity of the Teaching Assistant) by clustering the standard errors (Bertrand et al., 2004, Wooldridge, 2003).*

*The difference-in-differences analysis between all the treatment groups (ML-seeded and Long-seeded) to the Control group revealed that both seeding treatments had a significant effect on students' learning gains, with a coefficient value of 5.517 for the Treatment*Time ($p < 0.001$). While focusing on the ML-seeded group, the difference-in-differences analysis between the ML-seeded and Control conditions showed a coefficient value of 5.28 for the Treatment*Time ($p = 0.07$), which implies that a causal relationship is possible between the ML-recommended seeds and students' learning gains. However, when focusing on the Long-seeded group the difference-in-differences analysis between the Long-seeded and Control conditions we did not find a meaningful effect on learning gains.*

We also provide the following results of an exploratory analysis of the seeding effects on several aspects of forum behavior of students in the different conditions.

Amount of discussion: *We found that the number of responses to seeds in the ML condition (177 responses) was 16% higher than the number of responses to the seeds in the Long condition (153 responses, proportion test $z = 1.795$, $p = 0.073$) and 22% higher than the number of responses in the Control threads (145 responses, proportion test $z = 2.398$, $p <$*

0.016). The results in the Control condition are based on a set of 10 randomly sampled threads for each lecture, the same number of seeds that were planted in each of the seeding conditions.

Length of discussions: The average length of responses (measured in the number of words) to seeds in the ML-seeded condition (52.75, STD= 26.10) was found to be higher than that of the Long-seeded condition (43.02, STD= 25.82) and the Control condition (43.02, STD= 22.27). There was a significant difference between the condition groups (Kruskal-Wallis test ($X^2(2) = 15.588, p < 0.001$)). Post hoc analysis showed that the average response length in the ML-seeded condition is higher than that of Long-seeded ($p < 0.001$) and Control ($p < 0.024$), and that the average word length was higher in the Control condition than in the Long-seeded condition ($p < 0.025$).

Quality of discussion: The CE level of posts made in reply to seeds was found to be highest in the ML-seeded condition, in comparison to the other two condition groups. We found that the number of posts that exhibited the highest CE level (interactive) was higher in the ML-seeded condition (112) than in the Long-seeded condition (54) (Chi-Square ($X^2(1) = 24.597, p < 0.001$)) and in the Control condition (49) (Chi-Square ($X^2(1) = 26.547, p < 0.001$)). In the end, ML seeds generated twice as many interactive responses compared to the baseline approach and compared to randomly sampled threads in the forum.

4.9.5.3 Discussion And Future Work

The goal of this study was to develop automatic support for course instructors who wish to engage students in meaningful online discussions on course material. Using an AI approach to generate candidate seeds from prior students allows for the scaling up of the selection process of candidate seeds in both numbers and speed and provides a ranked list of recommended seeds for the instructor. Rather than attempting to be fully automatic, the approach described in this study relies heavily on the involvement of human experts (e.g., the course instructors) for all steps. Instructors are engaged in designing a measure for scoring

seeds according to educational impact, using the measure to label candidate posts for seeding, inspecting the proposed list from the algorithm, and selecting the chosen set of candidate seeds to plant in the forum.

In this respect, our approach occupies a “sweet spot” in the continuum between providing no support and fully automated support for seeding, in which the algorithm and the human expert share the problem solving, each performing tasks for which they are best suited. Here we are demonstrating a prototype for Human-AI collaboration that can be experimented with other forms of task allocation, such as allowing the instructor to override the machine generated recommendations with their own choices, or to use active learning methods to improve the selection algorithm over time. Our results suggest that seeding can have a positive impact on commenting behavior by students, that the method used for selecting seeds makes a difference, and that some of these changes may even be associated with the potential for learning gains. Both ML-seeded and Long-seeded approaches changed students’ forum behavior in ways that led to more posts, longer threads, higher CE, and better learning gains.

Also of note is the apparent advantage of the ML-seeded approach over the Long-seeded approach with regard to the quality of the ensuing discussion and students’ learning gains. We found that students in the ML-seeded condition exhibited greater cognitive engagement in their posts than the students in the Long-seeded condition. We also found evidence for a possible causal association (using the difference-in-difference analysis) between the ML-seeded approach and student learning gains, when constraining our analysis to exam questions related to the specific learning objectives for which we planted seeds. This means that it is worthwhile to invest effort in understanding the reasoning and criteria used by instructors to choose good seeds, and to use that knowledge to inform the design of automatic models for supporting their work. We did not find a difference in performance between students who actively responded to seeds and other students in the same group. This implies that part of the effect may be on account of students reading, but not participating in the discussion in the

forum. This is encouraging, as students are significantly more likely to read comments than they are to write comments.

In the end, the ML classifiers developed in this study were able to take a large comment corpus and deliver a subset of suitable candidate seeds using lexical, engagement and emotive based features of posts, properties which are both informative and generalizable to other contexts and disciplines. Aside from re-posting comments as seeds, these comments may serve valuable purposes in instructional settings, for instance, as discussion questions or as questions to be embedded within the course text itself.

Finally, we mention several limitations of our work. First, although the benefits of the seeding approaches for student engagement and performance are clear cut, the increase in number of responses was not large. This can be traced back to the way NB is used in the course. Students generally choose to commence new threads (e.g., asking a question about or paraphrasing the material) rather than responding to existing threads, which requires more effort. Therefore, the signal from the intervention study is not strong. It may be that providing more feedback from course instructors, or requiring students to reply to other threads, will motivate students to be more responsive to others. Second, our work was done using the NB platform, which differs from many forums in that the discussions happen in the margins of the course material rather than a separate forum. We can offer no evidence that our approach will be effective in these more traditional platforms, although we believe that it would. For future work, we wish to use our model outputs (recommended seeds) to promote discussions in other contexts, such as for discussion topics in small groups, during class, to be put into the text itself, etc. We also wish to employ an active learning approach in which feedback from the instructors regarding good and bad seeds is used to update the model after each lecture.

4.10 Funding

This work was supported by a grant by the National Science Foundation (#1915724).

4.11 References

- Atwater, M. M. (1996). Social constructivism: Infusion into the multicultural science education research agenda. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 33(8), 821-837.
- Barak, M. (2017). Science teacher education in the twenty-first century: A pedagogical framework for technology-integrated social constructivism. *Research in Science Education*, 47(2), 283-303.
- Bell, N., Perret-Clermont, A. N., & Grossen, M. (1985). Sociocognitive conflict and intellectual growth. In *Peer conflict and psychological growth. New directions for child development* (Vol. 29, pp. 41-54). Jossey-Bass.
- Bliss, T. J., Hilton III, J., Wiley, D., & Thanos, K. (2013). The cost and quality of open textbooks: Perceptions of community college faculty and students.
- Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom research into edX's first MOOC. *Research & Practice in Assessment*, 8, 13-25.
- Capota, C., & Severin, E. (2011). IDB Briefly Noted: No. 10: August, 2011: The Use of Technology in Education: Lessons from South Korea Title: Aportes: No. 10: Agosto 2011: El uso de la tecnología en la educación: lecciones desde Corea del Sur (No. 79959). Inter-American Development Bank.
- Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in cognitive science*, 1(1), 73-105.
- Cormier, D., & Siemens, G. (2010). The open course: Through the open door--open courses as research, learning, and engagement. *EDUCAUSE review*, 45(4).
- D'Antoni, S. (2009). Open educational resources: Reviewing initiatives and issues.
- DeSanctis, G., Fayard, A. L., Roach, M., & Jiang, L. (2003). Learning in online forums. *European Management Journal*, 21(5), 565-577.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dobson, J. L. (2008). The use of formative online quizzes to enhance class preparation and scores on summative exams. *Advances in Physiology Education*, 32(4), 297-302.
- Feng, D., Shaw, E., Kim, J., & Hovy, E. (2006, June). Learning to detect conversation focus of threaded discussions. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference* (pp. 208-215).
- Geller, S. A., Hoernle, N., Gal, K., Segal, A., Zhang, A. X., Karger, D., ... & Igo, M. (2020, March). # Confused and beyond: detecting confusion in course forums using students' hashtags. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 589-594).
- Gyllen, J., Stahovich, T., & Mayer, R. (2018). How students read an e-textbook in an engineering course. *Journal of Computer Assisted Learning*, 34(6), 701-712.
- Heiner, C. E., Banet, A. I., & Wieman, C. (2014). Preparing students for class: How to get 80% of students reading the textbook before class. *American Journal of Physics*, 82(10), 989-996.
- Hilton III, J. L., & Wiley, D. (2011). Open access textbooks and financial sustainability: A case study on Flat World Knowledge. *The International Review of Research in Open and Distributed Learning*, 12(5), 18-26.
- Hogan, K., Nastasi, B. K., & Pressley, M. (1999). Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions. *Cognition and instruction*, 17(4), 379-432.
- Johnson, D. W., Johnson, R. T., & Smith, K. (2007). The state of cooperative learning in postsecondary and professional settings. *Educational Psychology Review*, 19(1), 15-29.

- Johnson, B. C., & Kiviniemi, M. T. (2009). The effect of online chapter quizzes on exam performance in an undergraduate social psychology course. *Teaching of Psychology*, 36(1), 33-37.
- Jung, E., Bauer, C., & Heaps, A. (2017). Higher education faculty perceptions of open textbook adoption. *The International Review of Research in Open and Distributed Learning*, 18(4).
- Kim, J., Shaw, E., Feng, D., Beal, C., & Hovy, E. (2006). Modeling and assessing student activities in on-line discussions. In *Proc. of the AAAI Workshop on Educational Data Mining* (pp. 16-17).
- Li, C., Lalani, F. (2020). The COVID-19 pandemic has changed education forever. This is how. *World Economic Forum*.
- Light, V., Nesbitt, E., Light, P., & Burns, J. R. (2000). 'Let's You and Me Have a Little Discussion': Computer mediated communication in support of campus-based university courses. *Studies in Higher Education*, 25(1), 85-96.
- Marcell, M. (2008). Effectiveness of regular online quizzing in increasing class participation and preparation. *International Journal for the Scholarship of Teaching and Learning*, 2(1), n1.
- Martin, M. T., Belikov, O. M., Hilton, J., Wiley, D., & Fischer, L. (2017). Analysis of student and faculty perceptions of textbook costs in higher education. *Open Praxis*, 9(1), 79-91.
- Mazzolini, M., & Maddison, S. (2003). Sage, guide or ghost? The effect of instructor intervention on student participation in online discussion forums. *Computers & Education*, 40(3), 237-253.
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), 276-282.
- Miller, K., Zyto, S., Karger, D., & Mazur, E. (2014, March). Improving online class forums by seeding discussions and managing section size. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 173-174).
- Miller, K., Lukoff, B., King, G., & Mazur, E. (2018, March). Use of a social annotation Platform for Pre-class reading assignments in a Flipped introductory Physics class. In *Frontiers in education* (Vol. 3, p. 8). *Frontiers*.
- Novak, G. M., Patterson, E. T., Gavrin, A. D., & Christian, W. (1999). Just in time teaching.
- Plaut, S. (2006). "I Just Don't Get It": Teachers' and Students' Conceptions of Confusion and Implications for Teaching and Learning in the High School English Classroom. *Curriculum Inquiry*, 36(4), 391-421.
- Podolefsky, N., & Finkelstein, N. (2006). The perceived value of college physics textbooks: Students and instructors may not see eye to eye. *The Physics Teacher*, 44(6), 338-342.
- Poole, D. M. (2000). Student participation in a discussion-oriented online course: A case study. *Journal of research on computing in education*, 33(2), 162-177.
- Romeo, L. (2001). Asynchronous environment for teaching and learning: Literacy trends and issues online. *Delta Kappa Gamma Bulletin*, 67(3), 24-28.
- Sappington, J., Kinsey, K., & Munsayac, K. (2002). Two studies of reading compliance among college students. *Teaching of psychology*, 29(4), 272-274.
- Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and instruction*, 16(4), 475-5223.
- Singh, V., & Thurman, A. (2019). How many ways can we define online learning? A systematic literature review of definitions of online learning (1988-2018). *American Journal of Distance Education*, 33(4), 289-306.
- Stelzer, T., Gladding, G., Mestre, J. P., & Brookes, D. T. (2009). Comparing the efficacy of multimedia modules with traditional textbooks for learning introductory physics content. *American Journal of Physics*, 77(2), 184-190.
- Taylor, J., & Cox, B. D. (1997). Microgenetic analysis of group-based solution of complex two-step mathematical world problems by fourth graders. *The Journal of the Learning Sciences*, 6(2), 183-226.

- Väljataga, T., & Fiedler, S. H. (2014, June). Going digital: literature review on E-textbooks. In International conference on learning and collaboration technologies (pp. 138-148). Springer, Cham.
- Vellukunnel, M., Buffum, P., Boyer, K. E., Forbes, J., Heckman, S., & Mayer-Patel, K. (2017, March). Deconstructing the discussion forum: Student questions and computer science learning. In Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education (pp. 603-608).
- Vygotsky, L. (1978). Interaction between learning and development. Readings on the development of children, 23(3), 34-41.
- Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating How Student's Cognitive Behavior in MOOC Discussion Forums Affect Learning Gains. International Educational Data Mining Society.
- Wang, X., Wen, M., & Rosé, C. P. (2016, April). Towards triggering higher-order thinking behaviors in MOOCs. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (pp. 398-407).
- Weimer, M., & Gurevych, I. (2007, September). Predicting the perceived quality of web forum posts. In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP) (pp. 643-648).
- Wiley, D. (2014). David Wiley, "The Access Compromise and the 5th R". An Open Education Reader.
- Yogev, E., Gal, K., Karger, D., Facciotti, M. T., & Igo, M. (2018, June). Classifying and visualizing students' cognitive engagement in course readings. In Proceedings of the Fifth Annual ACM Conference on Learning at Scale (pp. 1-10).
- Zhang, A. X., Igo, M., Facciotti, M., & Karger, D. (2017, April). Using student annotated hashtags and emojis to collect nuanced affective states. In Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale (pp. 319-322).

Chapter 5. Leveraging Affective Reporting In Discussion Forum Posts To Understand Student Confusion While Reading The Course Text

5.1 Publication Status

Geller, S. A., Hoernle, N., Gal, K., Segal, A., **Kim, H. G.**, Karger, D., Facciotti, M. T., Sripathi, K. S., & Igo, M. (2021). New Methods for Confusion Detection in Course Forums: Student, Teacher and Machine. IEEE Transactions on Learning Technologies. doi: [10.1109/TLT.2021.3123266](https://doi.org/10.1109/TLT.2021.3123266)

Abstract: Students' confusion can be a barrier for learning, contributing to loss of motivation and to disengagement with course materials. However, detecting students' confusion in large-scale courses is both time consuming and resource intensive. This paper provides three new approaches for confusion detection in online forums that combine input from both students and teachers. The first approach describes a labeling tree that facilitates manual labeling of posts exhibiting confusion by the course staff and significantly improves the inter reliability measure between labelers. The second method classifies confusion based on rules that were inferred from students' hashtags that are used to convey emotions. The third approach uses machine learning models to detect confusion in students' posts. We demonstrate how these three approaches can be applied in an online forum of a large-scale Biology course. We combine the benefits of the three approaches by 1) selecting posts for training the machine learning models based on the rules that were inferred from students' use of hashtags in their posts; 2) using the labeling tree to generate labels for evaluating the models. We show that a pre-trained language model (BERT) that classifies confusion in posts based on raw text, was able to outperform traditional machine learning models that rely on feature engineering. This model was also able to generalize across different terms of the same course. This work empowers teachers with better technologies for detecting and alleviating confusion in online discussion forums that combine input from both teachers and students.

Selections from this work have been incorporated in this chapter, indicated by italicized text.

5.2 Author Contributions

SG, KG, HK, KS, MF, MI, and DK contributed to the conception and design of the study. SG, AS, KG, and DK contributed to the computational analyses. SG and HK constructed the confusion decision tree. HK, MF, and MI provided labels for cognitive engagement and confusion in student posts. MF and MI served as course instructors. SG performed statistical analyses. SG and KG wrote the initial draft of the manuscript. SG, NH, AS, KG, HK, KS, MF, MI, and DK reviewed and edited the final draft.

5.3 Introduction

Chapter 4 discussed the potential benefits of using online learning technologies and introduced Nota Bene (NB) as a tool for integrating social annotation and discussion into the reading of the course text. Chapter 4 also reviewed a study in which we hypothesized that initializing, or “seeding” an online course forum with comments from previous course instances would encourage discussion and improve the quality of discussions in NB. These posts from previous course instances were predicted to trigger discussions, and considered features such as length of discussions in previous terms, level of cognitive engagement, and relatedness to course learning objectives. Machine learning (ML) classifiers were leveraged to select suitable comment candidates with which to seed the forum. Students exposed to the seeded comments were shown to generate more discussion than those who were not exposed. Further, students exposed to the ML-selected seeds (in contrast to seeded comments determined solely by the length of discussion generated in previous course instances) showed higher levels of engagement and increased learning gains (as measured by performance on exam questions associated with learning objectives related to the seeded posts) (Shusterman, 2021).

This chapter details how comments in the forums can be analyzed to provide a better understanding of the students’ cognitive and affective states while reading the course text.

Alongside the benefits of collaborative learning discussed in the previous chapter, datasets from student interactions in discussion forums provide a rich profile of student experiences and reactions for instructors to mine, for the purpose of understanding how students engage with the learning material and the extent to which they are progressing in the course. For example, several groups have mined MOOC or other forum data and developed algorithms to predict students at risk of dropping out of the course (Tang, 2015; Onah, 2014; Wen, 2014). Another type of analysis involves automatic clustering or coding, where student posts are categorized into different genres or behaviors (e.g., help or clarification seeking, peer assessing, asserting, interpreting) (Lin, 2009), allowing for automation of an otherwise laborious manual content analysis.

Information gathered from student discussions, such as those from the aforementioned studies, can be important for instructors to better understand where students feel most engaged and challenged, to know at which point to intervene, and to have a general understanding of the discourse characteristics that accompany the course topics and learning objectives. Instructor interventions may occur in the discussion forum itself to help clarify confusion, respond to curiosities, or to encourage students throughout their discussions. Interventions may also occur external to the discussion forum, such as in cases where instructors provide extra support (e.g., more office hours, review sessions, one-on-one communications) with students who are at risk of dropping out of the course. Otherwise, the learnings from discussion forum datasets can help guide class pace and revisions made to the course materials (e.g., in identified hotspots of student confusion).

We propose that students can use affective reporting tools like hashtags and emojis in their posts to express themselves while simultaneously self-tagging their posts for filtering purposes. Instructors can further use computational techniques, like machine learning, to detect affects, such as confusion, in untagged posts.

5.4 Affective Reporting Using Emojis And Hashtags As A Way To Tag And Filter Categories Of Student Discussion Forum Posts

Affective reporting refers to the act of conveying information regarding one's experience of feeling or emotion. Buissink-Smith et al. (2011) also include values, attitudes, dispositions, and behaviors under the umbrella of affective attributes. Affective reporting in educational environments can be accomplished through verbal and non-verbal communications, through written word, or assessed through surveys and questionnaires. With interactions increasingly taking place digitally, such as in NB, body language and other forms of non-verbal emotional expression are not always available, or as easy to sense, however. Additionally, as mentioned previously, one limitation with online writing exercises (e.g., discussion forums or reflections in larger class sizes), is that large volumes of comments may be difficult to mine and analyze for affective states and therefore may result in missed opportunities to respond to student needs or course improvements otherwise. One way to mitigate this could be to have students self-tag their affective state using popular social tools such as emojis and hashtags.

5.4.1 Affective Reporting Via Emojis In Educational Settings

During the late 1990s, researchers started suggesting the usage of emoticon faces as a way to bolster social presence in online courses (Rourke, 1999). Over time, emoticons have transformed into miniature pictures, or icons called “emojis,” which are managed by the Unicode Consortium. According to Emojipedia, the latest Emoji set release was Emoji 14.0 in September 2021, with over 3,600 emojis. Based on an analysis of over 6.7 billion tweets from ten years of global Twitter data, emoji use has been growing on a yearly basis and is at an all-time high. In

2014, one in ten tweets contained an emoji, and as of July 2021, more than one in five tweets contains an emoji.^{9,10}

With the widespread usage of social media, the majority of students have most likely already been exposed to or have had experience with affective reporting via emojis, within social media contexts (Facebook, Twitter, LinkedIn, etc.). Emoji are not as prevalent in educational settings as they are in social environments, though acceptance and usage of emojis in formal, educational or professional settings is becoming more common. For instance, emoji reactions were included in Zoom, a widely used video conferencing application that became popular for remote teaching during the onset of the COVID-19 pandemic. There are also cases where students will use social media to interact with other students in the course. Facebook is a common platform used for this purpose, which facilitates discussions and provides its own set of animated emojis, emoticons, and stickers to “react” to posts with (e.g., “like,” “love,” “sad,” etc.).

Dunlap (2016) reviews emoticon usage in online learning and suggests that students use emojis to improve communication, enhance social presence, and build community in an online educational context. Vandergriff (2013) found that college students in an educational chat room setting who did not want to openly disagree with other students often attempted to clarify non-malicious intent by including emoticons, alongside non-standard or multiple punctuation. These studies highlight how emojis can be useful for students in their communications with each other, in addition to being useful tools for tagging posts.

However, Doiron (2018) asserts that while some emoji may help clarify messages or have a widely agreed upon meaning or function, other emoji can be more complex due to cultural variations in interpretations, context-dependent usage, and variations between operating systems. They suggest that emojis can be used in education in a manner similar to how they are used in business contexts for branding and marketing – that is, to use a custom

⁹ <https://blog.emojipedia.org/emoji-use-at-all-time-high/>

¹⁰ <https://blog.emojipedia.org/whats-new-in-unicode-14-0/>

set of emojis that are easily recognized and used (Doiron, 2016). They proposed using emojis in peer assessment in online forums that were conceptually symbolized, for example, exploding fireworks emoji for exceptional critical thinking, bright light bulb for commendable critical thinking, open book emoji for bibliographic reference, communication tower emoji for website hyperlink, and a thumbs up emoji for collegiality. This builds off Romig's (2015) suggestion to use emojis as a "quick and vivid way" to connect to course concepts, and increase student engagement.

With the aforementioned considerations in mind, the NB team has created (and refined, as described in Chapter 6) a custom set of emojis with pre-defined associated hashtags. These emojis were created with the dual purpose of adding socially engaging features to assignments and allowing capabilities for discussion filtering and downstream analyses.

5.4.2 Affective Reporting Using Nota Bene Emojis And Hashtags

An early iteration of NB included 8 emojis (interested, question, curious, useful, confused, idea, help, and frustrated), with the two non-face emojis being question and idea (represented by a question mark and lightbulb, respectively) (Figure 5.1). These emojis appear in their pictorial form in the emoji bar integrated in the NB annotation text field, and when they are selected, the comment is tagged with the pre-set corresponding hashtag text. For example, if a student clicked on the lightbulb emoji, their comment would be tagged with "#idea" (Figure 5.2a-b).

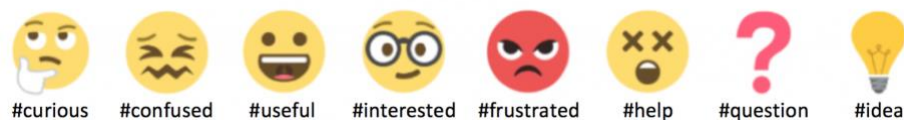


Figure 5.1. The original Nota Bene emoji and hashtag set.

Since discussion forum data can be large, unstructured, and complex, the emoji-hashtag system represents a simple and quick way to tag the post with an emotional reaction (e.g., curious, frustrated) or functional request (e.g., help). Zhang (2017), in agreement with previous literature, notes that emojis are particularly useful because they are a commonly used method to express the affective state, and that students are the best sources of their own affective state. Zhang also asserts that students may be eager to annotate as a form of self-expression or if they know other students or instructors will use the information to go through posts (2017).

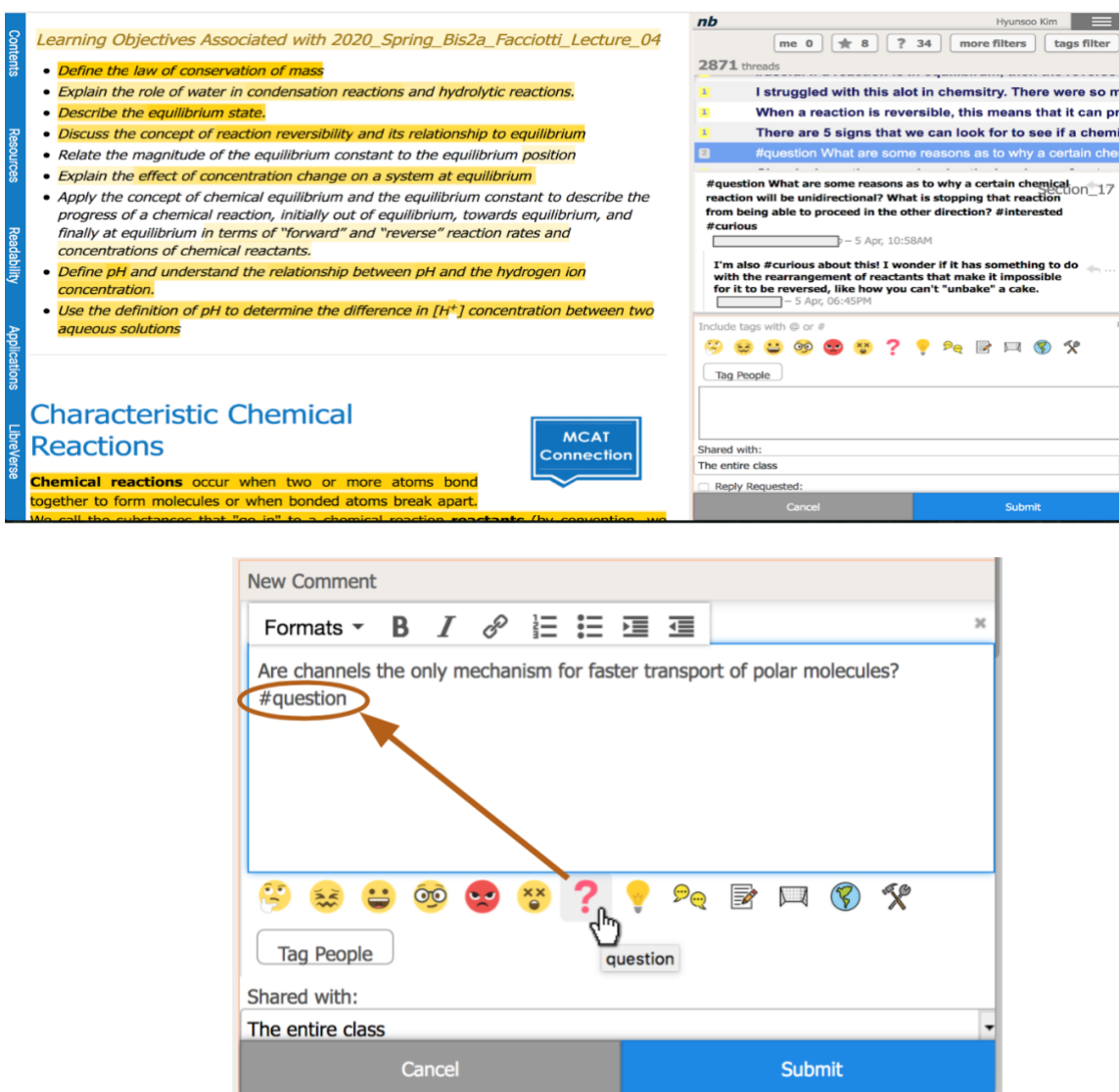


Figure 5.2 a (top). The Nota Bene interface overlaid onto the general biology online course textbook, **b (bottom).** A demonstration of how the emoji-hashtag tool within Nota Bene functions.

As previously stated, emojis are complicated as their usage can be context-dependent and they may have numerous interpretations due to cultural differences (e.g., symbolism, language). Thus, it is necessary to validate emojis and their interpretations for any downstream analyses (e.g., emoji heatmaps, machine learning) (See Chapter 6).

A preliminary study by Zhang (2017) explored emoji usage over 25,000 posts in NB in Bis2A. They recognized that two of the emojis within the set in particular, curiosity and confusion, were nuanced affects. The two affects were suggested to be related yet distinct, as students who express one versus the other are likely directed to different resources (Zhang, 2017). Classification algorithms that considered associated linguistic features for curiosity and confusion were developed, achieving 83% accuracy when categorizing a post for confusion or curiosity.

In a follow up study (Geller, 2020), the NB team endeavored to improve classifier performance in detecting confusion, as described in the following sections. The outcomes of this study provide further evidence that the affect of confusion is multifaceted, and that machine learning classifiers can still encounter difficulty with differentiating between specific kinds of confusion. Understanding the specific types of confusion that students are facing has important implications for intervention strategies by instructors.

5.5 New Methods For Confusion Detection In Course Forums

5.5.1 Introduction

Although confusion is often *viewed as a negative emotion by instructors and students alike* (Neary, 2009), *recent work suggests that confusion can both discourage and encourage student learning based on the other epistemic emotions that may accompany it* (D'Mello, 2014; Rodrigo, 2010). *For instance, experiencing confusion simultaneously with other negative emotions (i.e., frustration and boredom), may impede student progress. When left unresolved,*

these emotions may lead to lower achievement, dropout, or academic dishonesty (Yang, 2015; Alexandron, 2017). However, when teachers recognize and respond to students who are experiencing negative emotions in time, students often return to an engaged state, with an enhanced learning of the new material (D'Mello, 2014; Rodrigo, 2010; Yang, 2015; Grawemeyer, 2015; D'Mello, 2012). Thus, monitoring a student's emotional reaction to new material can enable faculty to provide interventions or other support to turn confusion into a positive learning experience.

The opportunity for instructor intervention with student confusion during online learning is a powerful tool for individualizing learning within the context of high-enrollment classrooms. However, key steps for effective instructor intervention are the accurate detection of student confusion and the ability to link this confusion to a particular location within the course material. Instructors can then use this information to guide their interactions with the students, develop online materials to help students with particularly problematic topics and inform better course design. We propose that the students' self-reported affective states in the form of hashtags can be used to identify when, and on what content, the students are expressing confusion. Furthermore, when hashtags are not present, as observed in most online student forums, we still aim to provide useful insights about confusion to the course instructors by harnessing the power of machine learning to automatically classify the text in students' posts.

In the absence of hashtags, a natural way to identify confused posts in a forum is to rely on course staff or trained experts for the manual labeling of the posts. For example, the Stanford MOOCPosts dataset contains thousands of posts that were manually labeled on a Likert-type scale of confusion by nine colleagues (Agrawal, 2014). However, as we worked with our course experts, we discovered that there was not a universally agreed upon definition of confusion. Based on these discussions, we decided to introduce more rigor into the labeling process by designing a labeling tree that leads experts toward more consistent label assignments. Such manually generated labels can be used to train machine learning classifiers,

which can identify confused posts in a dataset based on linguistic cues and the vocabulary used in the students' posts (Agrawal, 2015; Zeng, 2017). However, this approach imposes a significant burden on the course staff, who in many cases are required to do the hand-labeling of the posts.

We propose an alternative approach that infers confusion with course material directly from the students' use of hashtags. In this paper, we investigate the following research questions:

- Research Question 1 (RQ1): Does the introduction of a structured labeling tree for guiding experts result in higher agreement, compared to using a Likert-type scale measure?*
- Research Question 2 (RQ2): Do students' hashtags agree with experts' judgments about what constitutes confused posts?*
- Research Question 3 (RQ3): Can students' hashtags be used to inform automatic classification of confused posts for situations in which hashtags do not exist? Can these models generalize to other course terms?*

To answer these research questions, we suggest three new approaches for identifying confusion in this setting.

The first approach relies on manual labeling using experts (course teaching staff). We designed a labeling tree that helps the course teaching staff distinguish between the different types of confusion found in student posts. The labels in the tree reflect a taxonomy of different types of confusion, such as seeking information when students are aware they lack it, and making incorrect statements about course material without being aware they are incorrect. We found that using the labeling tree significantly improved inter-reliability agreement between experts when compared to traditional methods of collecting opinions such as the Likert-type scale (Yang, 2015; Agrawal, 2015). Although this tree-based method is optimal with respect to accuracy in detecting "confused posts," it is resource intensive in terms of both time and money. This led us to explore other approaches that might serve as a good proxy for teachers' judgments about the type of confusion exhibited by students in posts. We use the decision-tree approach to generate hundreds of ground truth labels with regard to whether students' posts

express confusion. We then use these labels to evaluate the reliability and reproducibility of the second and third less resource intensive approaches.

The second approach classifies posts exhibiting confusion based on students' use of hashtags in the posts. In many forums, students are given the option to select from a set of predefined hashtags in their posts to convey opinions, ideas, and emotions in a similar way to many social media platforms (Mohammad, 2010). By comparing student labeling using hashtags to labeling by experts using the decision tree approach, we show that a naive labeling rule that checks for the presence of the predefined #confused hashtag is a sufficient, but not necessary condition for inferring confusion in a student's post as defined by the course teaching staff. This discovery motivated a new labeling rule that considers additional hashtags that convey questioning and help-seeking behaviors, which may also be indicative of confusion. We show that this new labeling rule is more aligned with teachers' ground truth judgment of confusion in the students' posts. We use this rule to generate thousands of proxy labels of confusion in students' posts in a fast and efficient way.

The third approach uses machine learning classifiers for detecting confusion in students' posts that can be applied in the absence of student self-reported hashtags. We leveraged the proxy labels from the vast number of posts with students' self-reported hashtags to create a model trained on a dataset of posts labeled using the second approach. Experts then evaluated the validity of the ground truth labels generated by this model on posts without hashtags using the labeling tree we described in the first approach. We showed that a deep learning-based model pre-trained on general text corpora and used only on the raw text of the post as input, was able to outperform traditional machine learning models that rely on feature engineering. Moreover, this model was able to generalize between different offerings of the course, making it a good candidate model to be used in practice.

Our work shows that by making use of the students' self-annotated posts, we can augment existing models for confusion detection and can inform the development of automatic

confusion detection systems to support teachers' understanding of how students comprehend course readings. It also demonstrates the value of including the "human-in-the-loop" in the design of analytical tools for supporting online learning.

Providing students with a natural and familiar way to convey affective states improved the performance of machine learning models for detecting confusion, even for cases where hashtags were not available. Moreover, by directly involving teachers in the process of validating our rules, we were able to arrive at non-trivial conclusions for what constitutes confusion in online forums.

5.5.2 Overview Of The Nota Bene Dataset

Our dataset consists of two terms of Bis2A, summer and winter of 2018. Each term consisted of 25 lectures. Reading assignments for each lecture were uploaded to NB and students were required to provide "three meaningful posts" for each reading assignment in NB before the lecture. The purpose was to encourage active participation in forum discussions and to provide a way for students to obtain answers to their questions before lecture. The NB interface also allows students to generate pictorial representations (emojis) of their affective state using hashtags and text, (e.g., #confused, #useful, #frustrated) as shown in Figure 5.2. Table 5.1 shows the eight possible hashtags, including the frequency of their usage by the students in the summer term. The distribution of hashtags in the winter term was similar, hence for brevity we display results for summer term only in this section. Students received additional credit for including at least one hashtag in at least one of their posts per lecture. Another feature of NB is that posts can be filtered by hashtags, allowing instructors to navigate to posts that display specific emotions.

Table 5.1. Hashtags, their associated emojis, and the counts of their use in the summer term.

Hashtag	Count	Emoji	Hashtag	Count	Emoji
#confused	1228		#help	420	
#curious	6595		#question	7574	
#interested	9237		#useful	10 866	
#idea	9673		#frustrated	65	

In total, out of 58,811 posts in the summer term, 40,842 posts (70%) contained at least one hashtag. Similarly, out of 70,360 posts in the winter term, 35,016 (50%) contained at least one hashtag. These proportions are well above the minimal requirement required from students. This suggests that students may perceive intrinsic value from associating hashtags with their comments, supporting calls for providing students with opportunities for self-assessment (Ross, 2006).

Figure 5.3 displays the number of posts displaying #confused and #question hashtags across the different lectures in the summer term. As shown by the figure, students' use of those two hashtags varied across the different lectures. The information derived from their use of hashtags can contain useful insights for the course staff. For example, lecture three, which a priori was considered to be easy by the course staff, had the highest number of posts containing the hashtag #confused (105 posts) and #question (425 posts), suggesting that students found this material difficult. In contrast, lecture ten, which a priori was considered to be difficult by the teaching staff, had relatively few numbers of posts containing the hashtag #confused (28 posts) and #question (271 posts).

The number of unique hashtags used by students also varied widely, reflecting prior results showing the varied use of hashtags in popular social media platforms (Davidov, 2010). Figure 5.4 shows a histogram of the number of hashtags used at least once by students across all of the lectures in the summer term. As shown by the figure, the majority of the students (more than 85%) made use of four or more of the eight available hashtags, with a small minority

(2.2%) using all of the hashtags. In general, students contributed well above the minimal number of required hashtags for each lecture. Only a small minority of the students did not include any hashtag in their posts (2.5%).

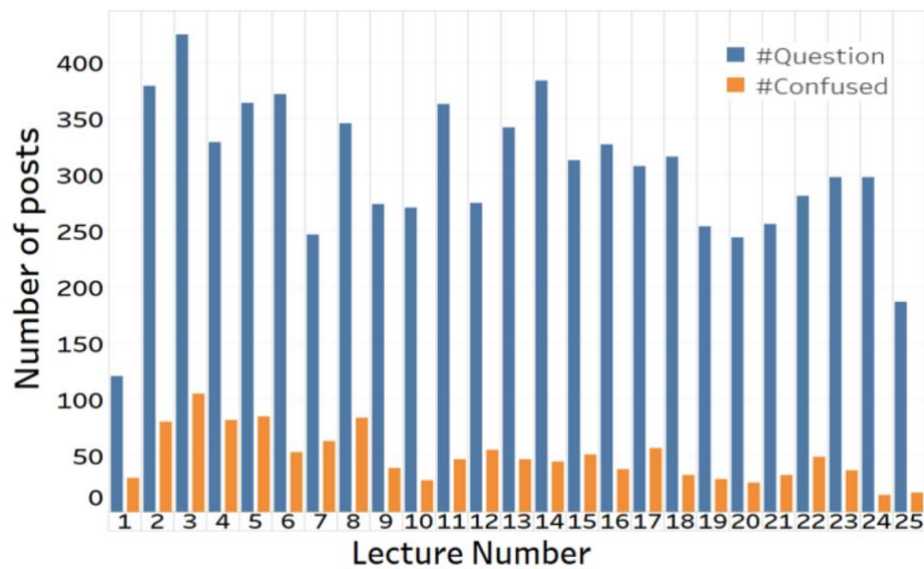


Figure 5.3. Distribution of #confused and #question hashtags in the summer term.

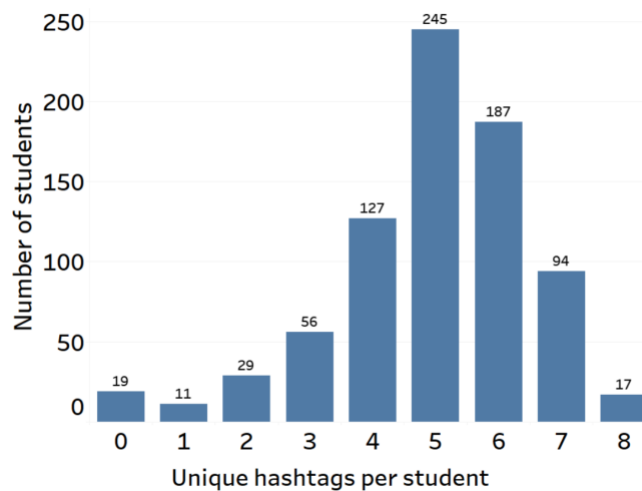


Figure 5.4. Histogram of unique hashtag use by students in the summer term.

5.5.3 Inter-Labeler Agreement When Using Labeling Tree

Previous works (Agrawal, 2015; Geller, 2020) used a Likert-type scale to obtain confusion labels from experts. We asked two experts from the Bis2A course staff to label over 200 posts using a Likert-type scale. Comparisons of the label assignments revealed low agreement ratios between the two experts (Kappa measure of 0.39 over 200 posts). This identified the need for a more structured approach to labeling confusion that reflects the taxonomy of different types of confusion from the pedagogical literature (Plaut, 2006). To this end, we developed a method that takes into account the different aspects of confusion by using a labeling tree, presented in Figure 5.5.

Other works have demonstrated the benefits of using labeling trees to facilitate analysis of students' forum behavior (Yogev, 2018; Wang, 2016). The labeling tree is composed of a series of questions to be answered by the experts (course staff members) and it directs them to choose a label that reflects the type of confusion exhibited by the post. Based on discussions with the course staff and manual analysis of student posts, we identified several types of confusion that are commonly expressed by students: feeling lost about course material, drawing an incorrect conclusion, asserting there are contradictions in the text, and seeking information about course material or logistics.

We also distinguished questions about material the students are expected to learn from questions that go beyond the course learning goals. Table 5.2 presents a complete list of label categories and an example of a post for each category of confusion.

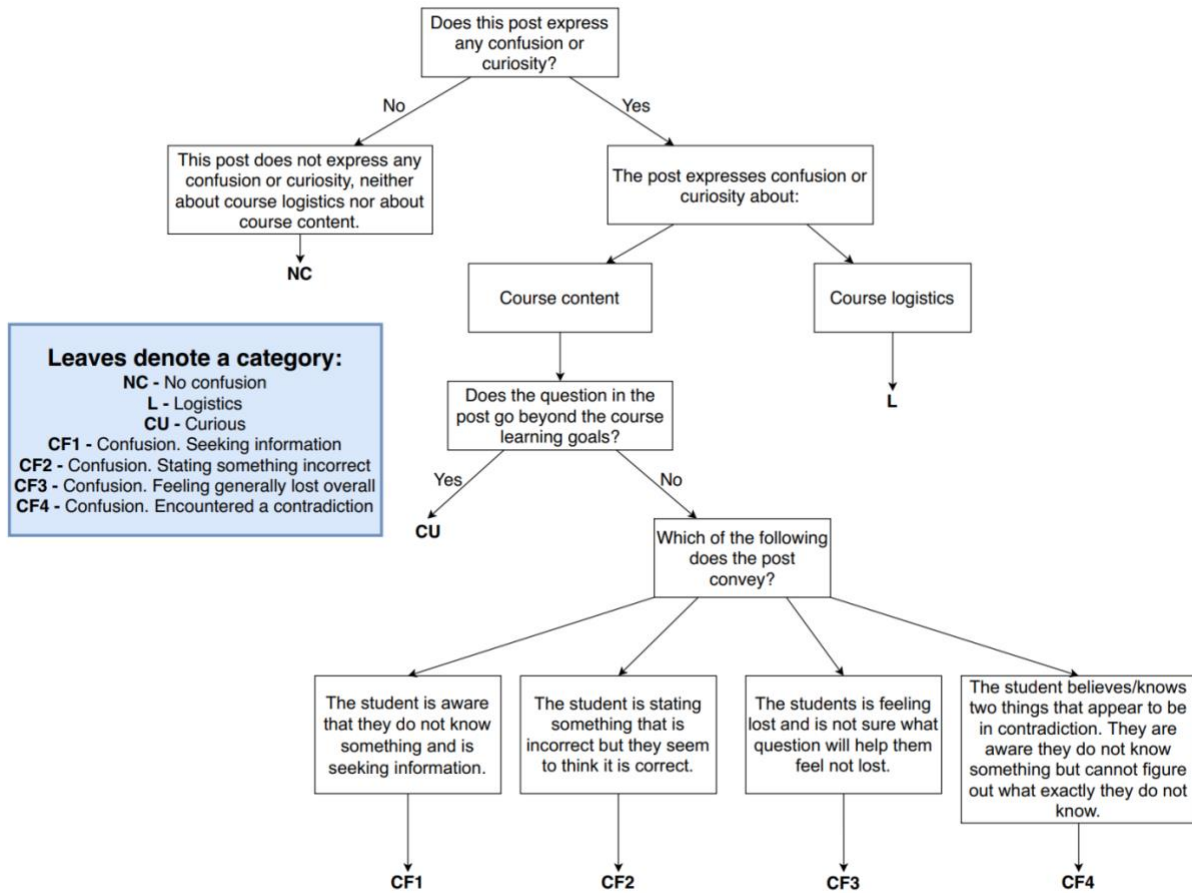


Figure 5.5. Labeling tree for confusion.

Table 5.2. Definition and example of each leaf label of the labeling tree.

Tree Label	Definition	Example
NC	No confusion	<i>"This can be remembered easily because the prefix 'mei' means to reduce!"</i>
CU - Curiosity	<i>Posts contain statements or questions that go beyond the course learning goals. The same question can be labeled as confusion or curious, depending on whether the student was expected to know the concept at that particular stage in the course or not.</i>	<i>"I wonder what other types of exposure can affect formation of different mistakes in DNA. Is it just radiation?"</i>
L - Course Logistics	<i>Posts relate to course technicalities like assignments, exams, or course requirements. For example, questions such as: "Do we need to know ____ for the exam?" Our experts seldom label these types of questions as confusion. Though this indeed is a type of confusion related to expectations for the students, we would like to separate them from cases where confusion is more directly related to misunderstanding the course content.</i>	<i>"If the main focus is to give a general view on how energy and matter are relevant in biological systems, will specific cases of energy such as in mitochondrial interactions and glycolysis be less important to know? Or will those still be required in future lessons?"</i>
CF1 - Confusion - Seeking Info	<i>The student is aware that they do not know something and is seeking information. This usually involves asking questions or wondering about specific portions of the text, or soliciting other students' opinions. This label captures most of the confused posts.</i>	<i>"Is this saying that there is no specific start and end of the ETC?"</i>
CF2 - Confusion - Wrong	<i>The student is stating something that is incorrect, but they seem to think it is correct.</i>	<i>"Yes, unlike the bonds in ATP, hydrogen bonds require energy to be broken."</i>
CF3 - Confusion - Lost	<i>The student is feeling lost and is not sure what question will help them feel not lost.</i>	<i>"I searched some info but I can't be sure that is correct. It says that in higher plants, the cyclic photophosphorylation helps with the condition that ATP is not sufficient."</i>
CF4 - Confusion - Contradiction	<i>The student encounters two things that appear to be in contradiction. They are aware they do not know something, but cannot figure out what exactly they do not know.</i>	<i>"How would you change the nucleotide sequence? I thought the strength of the promoter does not get altered? I thought activators and repressors in some sense adjust to 'regulate' the strong/weak promoters."</i>

To evaluate the labeling tree, we randomly picked 300 posts with no hashtags from each term of the Bis2A course, making 600 posts in total. We denote posts without hashtags as raw

posts. Three experts from the course staff labeled the 300 posts from the summer course term independently using the labeling tree of Figure 5.5. Experts labeled all of their posts before comparing their work. They achieved an intra-class correlation coefficient of 0.86, indicating high agreement among the experts. We further presented the 300 posts from the winter course to two of the three experts. They independently used the labeling tree to classify the confusion in the posts and achieved a Kappa score of 0.90. We can thus answer RQ1 affirmatively, in that the Kappa score obtained with the labeling tree (0.90) was significantly higher than that obtained with the Likert-type scale approach (0.39). All of the experts stated that using the labeling tree significantly facilitated the labeling process in terms of coherence.

5.5.4 Agreement Between Students' Self-Reported Hashtags And Experts

Interestingly, the use of the #confused hashtag makes up only 3% (1,228) and 11% (4,178) of the total number of the 40,842 and 35,016 posts with hashtags in the summer and winter terms, respectively. Thus, it is either the case that (1) the set of all confused posts is a superset that includes posts labeled with a #confused hashtag, or that (2) the amount of confusion displayed in this dataset is significantly lower than the amount noted by previous work (Zeng, 2017; Zhang, 2017). The above examples clearly show that the naive labeling rule, which only uses the #confused hashtag, is inadequate for detecting all of the confused posts, leading us to conclude that the first explanation is the correct one. We, therefore, need a new rule that goes beyond the #confused hashtag to identify all confused posts in student forums. To find such a labeling rule for confusion in this dataset, we use a computational method to investigate whether the hashtags are used in similar contexts (with similar vocabulary in their posts). We follow a method similar to that introduced by Eisner et al. (2016), which uses word embeddings to create vector representations for Twitter emojis. Similar to Wei et al. (2017), who trained their word embedding model on the posts from the Stanford MOOCPost dataset, we trained word embeddings using the Nota Bene posts. We then calculated the cosine similarity

between the embedding vector of the #confused and the other hashtags to define a new labeling rule that takes into consideration other hashtags rather than just #confused.

Table 5.3 shows the cosine similarity between the vector embedding of each hashtag to the vector embedding of the #confused hashtag. As shown by the table, both the #help and #question hashtags show the highest values for cosine similarity to the #confused hashtag. This suggests that our new labeling rule should consider #help and #question hashtags in addition to the #confused hashtag. Based on the cosine similarity scores, we grouped students' hashtags into two groups: one group containing hashtags that reflect a need for assistance (#question, #help, #confused) and another group containing hashtags that reflect interest in the material (#idea, #useful, #interested, #curious). We ignored the appearance of #frustrated due to its low occurrence in the data (accounting for 0.02% of all posts).

Table 5.3. Cosine similarity between the embedding for the #confused hashtag and the other hashtag embeddings (the closer the value is to 1, the more similar it is to #confused).

Hashtag	Cosine similarity
#help	.987
#question	.982
#curious	.973
#frustrated	.907
#interested	.89
#idea	.885
#useful	.881

This analysis led us to consider and compare the validity of the following two rules for detecting confused posts, the naive labeling rule (R1) and our data-informed rule (R2):

- R1: This rule declares a post to be confused if and only if it contains a #confused hashtag.
- R2: This rule declares a post to be confused if and only if it contains at least one hashtag from the assistance hashtag group, and none of the hashtags from the interest group. Thus, R2 is a super-set of R1.

We hypothesize that rule R1 is a sufficient condition for defining confused posts, but not a necessary condition. That is, a human expert could identify confused posts that satisfy rule R2, but not rule R1.

To test the agreement between expert labels and our two labeling rules that are based on students' self-reported hashtags, we formed three groups of posts from the set of posts in the summer term of Bis2A. Each group contained 150 posts with every post containing at least one hashtag. The first group included posts satisfying rule R1. The second group included posts satisfying rule R2, but not R1. The third group was a control that included posts that satisfy neither R1 nor R2. Due to the high inter-reliability agreement between experts on the 300 posts, the labels for these 150 posts were generated by a single expert. We say that the expert agrees with a rule for a given post if the expert labels that post as exhibiting a degree of confusion (CF1, CF2, CF3, or CF4 labels of the tree). Figure 5.6 shows the number of agreed instances between the expert and the posts in each group (orange bars). We refer to the posts in the groups above using the rule that generated them: group R1, group $R2 \setminus R1$, and control group.

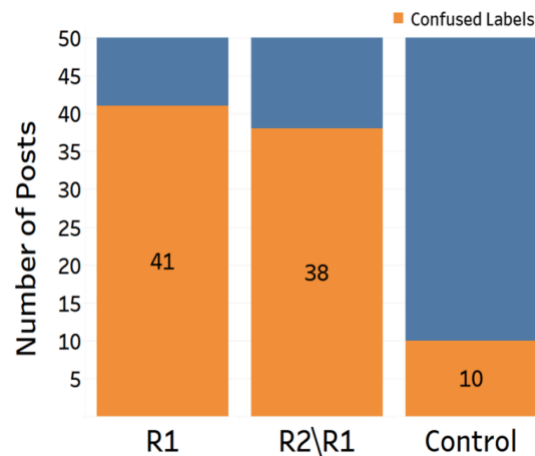


Figure 5.6. Agreement between expert and posts in groups R1, $R2 \setminus R1$, and control group.

Statistical significance was obtained using Z tests with a Bonferroni adjusted α level of 0.05. We used the Normal approximation to the Binomial distribution, as the sample size was large in all three cases. From the figure, we can conclude that 1) there was significantly more

agreement between the expert and the posts in both groups $R1$ and $R2 \setminus R1$ than the posts in the control group; and 2) there was more agreement between the expert and the posts in group $R1$ than in the group $R2 \setminus R1$, but this difference was not significant. The Kappa agreement between the expert and the posts in the $R1$ group was 0.28. However, the Kappa agreement between the expert and posts in the $R2$ group (which includes both the $R1$ group and $R2 \setminus R1$ group) was 0.55. This supports our premise that the labeling by experts is more likely to agree when the posts are labeled using the $R2$ rule. Moreover, it is important to note that the 38 posts labeled correctly as confused under the rule $R2$ would have been missed by the rule $R1$.

These results confirm our hypothesis that while $R1$ may be the most important indication for confused posts, it is incomplete, and adding posts that satisfy rule $R2$ will improve the detection of confusion in this dataset. Thus, we have answered research question 2 in the affirmative.

5.5.5 Automatic Classification Of Confused Posts

The training set for the models consisted of posts from summer and winter terms of Bis2A that were classified as confused using the $R2$ rule (hence referred to as $R2$ -labeled datasets). The test set for the models consisted of raw posts from summer and winter terms of Bis2A that had been labeled by the experts (hence referred to as expert-labeled datasets). The labels $CF1$, $CF2$, $CF3$, or $CF4$ in the tree were considered to be “confused,” and the other labels (NC , CU , L) were considered to be “non-confused.” For the summer term where we used three experts, we discovered that the majority of the posts were not labeled as “confused” by the experts. Overall, 64 posts were labeled as “confused,” while 236 posts were labeled as “non-confused.” For the winter term, where we used two experts, we kept only the instances where the two experts agreed on the post label. This resulted with 288 posts, with 52 labeled “confused,” and 236 labeled “non-confused.”

Details regarding the feature engineering and testing of six different machine learning models can be found in the publication associated with this chapter (Geller, 2021).

We study in detail the predictions of the BERT+D model (BERT model with downsampling rebalancing applied), which was the best performing model over both terms. We compare the predictions made by this model to the expert labels collected in the within-term setting. Table 5.4 presents results of the number of misclassified posts for summer and winter terms. The table shows that BERT-D performed best on the two classes for most labels, achieving an almost perfect prediction performance for the “No Confusion” (NC) and “Confusion— Seeking Info” (CF1) labels. However, the BERT-D model does not perform as well with the other labels, namely “Confusion—Wrong” (CF2), “Confusion—Lost” (CF3), “Confusion—Contradiction” (CF4), and “Curious” (CU), which occur infrequently in the datasets. Although the sample size is low, we see a trend that the misclassification rate on these posts is high, with a vast majority of these posts getting misclassified in the summer and winter terms, respectively.

Table 5.4. Error analysis of BERT+D model over expert-labeled posts for the summer and winter terms.

	Summer			Winter		
	Total	misclassified	%	Total	misclassified	%
CF1	51	3	6	45	0	0
CF2	7	7	1	7	7	1
CF3	3	3	1	0	0	—
CF4	3	0	0	0	0	—
NC	221	5	2	219	1	.4
L	2	2	1	3	3	1
CU	13	11	85	14	14	1

We highlight two categories that the model performed particularly poorly on, the CF2 and CU labels. The “Confusion—Wrong” (CF2) label denotes a post where the student makes a factual error without realizing it. Since the students are unaware of their misconceptions, they do not use the #confused, #question, or #help labels that would be used to train the classifier under R2. Moreover, any classifier hoping to correctly identify this form of confusion would need a

conceptual understanding of the content. This understanding, which is embodied by the experts during their labeling, goes beyond the ability of the natural language techniques that we applied. It is also important to note that the only labeling system to correctly identify these posts is that which depends on the expert labelers. Even the R2 labeling rule, which depends on the students' self-reported confusion, cannot identify this form of confusion in the dataset.

Second, the “Curious” (CU) label was used by the experts to identify statements that go beyond the course learning goals. As with the CF2 label, we can neither expect the students nor the classifier to identify this label, because it requires a thorough understanding of the Biology course curriculum.

5.5.6 Discussion

5.5.6.1 Labeling Confusion In Forum Posts

We found that using a labeling tree helped teachers align their understanding of the degree of confusion in students' posts. The adoption of the tree resulted in better agreement among the teachers who completed the labeling. We consider the teachers to represent the ground truth for what posts are truly confused, as they are able to identify, in particular, the CF2 (Confusion—Wrong) and CU (Curiosity) labels. It is important to note that even the students cannot identify these two labels themselves as they (1) do not always realize their misconception, and (2) are not always confident about the bounds of the scope of the course.

While we motivate the practical use of the labeling tree, we emphasize that it was not meant to provide a holistic definition for confusion. Rather, the tree was designed with the input of the Bis2A teaching staff, with the goal of reducing space for varied interpretations. For example, a student might be confused about course logistics. However, one teacher might not deem this as useful to the goal of understanding confusion in the reading material. This teacher might not label that post as confused, whereas a different teacher might rely more on the mere

presence of confusion and thus label it differently. By separating these various labels out, we have reduced the possibility that these multiple interpretations might occur and have thereby increased the alignment among the teachers' labels.

Due to a lack of precise definitions for the labels, there exists varied use of self-reported annotation among students. We show that the trivial rule for labeling confusion (one which merely considers the use of #confused) is insufficient for identifying a large number of posts that truly do reflect confusion. We, therefore, present an improved rule, one which considers more than just the #confused label, and we show that this new rule better captures the set of all confused posts. We argue that students' self-reported hashtags should be used to label confusion in the posts, as we cannot rely on the staff to hand-annotate the thousands of posts on a forum. While the students have difficulty identifying CF2 (Confusion—Wrong) and CU (Curiosity) posts, we show that the scalability of our approach outweighs this limitation.

5.5.6.2 Automatically Detecting Confusion In Forum Posts

We have discussed the value of using students' self-reported labels for identifying confusion in their posts. However, many posts do not contain such self-reported labels. For example, in our datasets, approximately $\frac{1}{3}$ of all posts did not contain a hashtag, and this was the case even though the students were explicitly asked to include them. It is clear that relying solely on students for self-reported hashtags is insufficient for labeling all the posts in a dataset. However, we show that an automated classifier can be trained to predict confusion on posts that do not contain hashtags. This classifier was trained on labels derived from the students' self-reported hashtags, but it was evaluated on the labels that were provided by the teaching staff who used the labeling tree. This automated classification is helpful, as it will identify the presence of confusion, even when students do not label their own posts with a hashtag.

The overall performance of the BERT-D model (F1 measure of 0.77) makes it a good candidate model to be used in practice. Despite this, we identified that certain confusion types

could not be reliably detected. In particular, the CF2 and CU labels were missed by the classifiers. However, this was not unexpected. Two reasons for the poor performance on these labels are: (1) the classifiers were trained on the data from students self-reporting, who would also be expected to miss these labels, and (2) the classifiers do not understand the semantics of the course material or the syllabus and, therefore, cannot possibly identify these two labels.

5.5.6.3 Pedagogical Implications

A few implications come from this work that address the way we should label student posts in online discussion for confusion, and highlight the limitations of the current machine learning models for this task. Our work highlights the importance of building machine learning models for detecting confusion that are informed by the pedagogical literature.

The ability to separate the multiple types of confusion using the tree structure is an important expressive power, which is lacking in other labeling methods that only rely on Likert-type scale methods. Therefore, we conclude that given a student's post, an expert course staff member who uses the labeling tree is better qualified to label the post for confusion than the student. Even though we showed a good mapping from students' self-reported hashtags to experts' labels, some types of confusion that may be rare but are expected, seem to be only identifiable by experts. Despite this, we show that the students' use of self-reported hashtags was very useful in training the Machine Learning models.

Finally, the course staff stated that labeling posts using the tree helped them to better understand how the students interact and understand the reading material. For example, the teachers were interested in understanding what specific parts in the reading material or course instruction may have led students to write incorrect statements (CF2 labels). Because these comments are associated with highlighted regions in the text, it would be relatively easy for the teachers to add clarifications or examples to paragraphs containing the highlighted regions, which could help resolve these misconceptions. Thus, the labeling tree, combined with students'

posts in the discussion forums, can be used as a pedagogical tool for teachers to obtain insights and resolve issues in the reading material.

5.5.6.4 Theoretical Implications

As summarized in (D'Mello, 2014), confusion has many definitions and implications in different schools of thought. Historical definitions characterize it as negative or undesirable (Perkins, 1992; Piaget, 2003), whereas conceptual change literature views it as a necessary step in students' corrections of their misconceptions (Chi, 2009). Recent work on confusion's epistemic properties highlight possibilities for both negative (Yang, 2015; Alexandron, 2017) and positive (D'Mello, 2014; Rodrigo, 2010) consequences. We adopted this more nuanced view in the current work: the easier it is for instructors to detect student confusion, especially in novel online learning environments, the greater the chance for instructors to guide students toward positive confusion-related outcomes. To that end, we applied Plaut's (2006) framework to identify common types of confusion in students' comments, allowing us to define the four types of confusion (CF1–CF4) in our decision tree.

Our results, specifically our between-terms predictions, also highlight the multi-faceted nature of confusion before any academic consequences (good or bad) can be observed. The strongest classifier, the BERT model, performed best on “Confusion—Seeking Information” (CF1). This category is the most straightforward depiction of a student's awareness of their own cognitive disequilibrium, because students appear to be asking very straightforward questions. The second straightforward example of disequilibrium is “Confusion—Contradiction,” which occurs very infrequently. This infrequency may be due to the amount and sophistication of a student's knowledge. In order to fall in the CF4 category, students must both identify what they do and do not know, and compare these two domains to specify their questions. As previously stated, this category may be outside the scope of classifier prediction because of the in-depth content knowledge required.

Our remaining Confusion categories may provide an expansion of the epistemic definition of confusion. “Confusion—Wrong” (CF2) and “Confusion—Lost” (CF3) both do not contain explicit questions (which is likely why students typically do not use hashtags for these comments). Both of these categories represent “timepoints” in student understanding before the students are aware of their own cognitive disequilibrium.

Conversely, the instructors are aware of this mismatch in student thought, as evidenced by the instructor creation of the CF2 and CF3 categories. Thus, although their identification is practically challenging, these two categories would likely be the most useful for instructors, because it would allow them to intervene before students ignore or forget these gaps in their own understanding. Additionally, these two categories emphasize the role of the instructor as a resource in helping students identify, and eventually resolve, mistakes in new knowledge incorporation.

5.5.7 Conclusion And Future Work

In this work, we study three research questions related to detecting confusion in student discussion forums. The first research question discusses methods for labeling confusion by experts. We show that confusion is prone to multiple interpretations, even by course staff members, and simply using Likert-type methods for labeling confusion can lead to poor agreement between the experts. We propose a labeling tree that consists of a series of questions to be answered by experts (e.g., course staff members), and it directs them to choose a label that reflects the type of confusion exhibited by the post. This tree was shown to produce consistent labels between experts by facilitating the labeling process in terms of coherence.

The second research question asks whether students’ self-reported hashtags agree with experts’ judgment about what constitutes a confused post. We show that the #confused hashtag is sufficient, but not a necessary condition for inferring confusion in a student’s post. We learned of embedding vectors for each hashtag, and show that the #help and #question are semantically

similar to the #confused hashtag. This led to the design of two labeling rules, a naive rule that uses only #confused to signify a confused post, and a data-informed rule that also incorporates the #help and #question. We show that the latter is more aligned with the experts' judgment.

In the third research question, we ask whether we can harness students' self-reported hashtags to inform automatic classifiers for detecting students' confusion in posts in which hashtags do not exist. We experiment with models that rely on hand-designed features, as well as with state of the art pretrained language models (BERT) that accept only the raw text of the post as an input. We show that classifiers trained using the hashtag-labeled posts perform well on expert-labeled posts and that the BERT model outperformed other models on the task. We also show that the models can generalize between terms, training on data from one term and predicting on another. Finally, we present an error analysis of the BERT model, which shows that some types of confusion are harder to detect than others. This illustrates the importance of the hashtags as valuable training data, and the fine-grained confusion labels obtained when experts used the labeling tree as a way to understand the models' true blind spots.

We also note the limitations of our approach in capturing only some aspects of students' confusions and not capturing other forms of confusion (Chi, 2009). In future work, we wish to extend the machine learning models to improve prediction of the less obvious confusion types such as CF2 (a wrong statement that the student is not aware of). We also seek to design a confusion "heatmap" of students' posts [that] will provide teachers with a high-level picture of the level of confusion throughout the reading material. Finally, we wish to study how feedback from this heatmap can guide teachers in their design of the course and the downstream effects on student learning gains and attitudes towards using tools such as NB.

5.6 Funding

This work was supported by a grant by the National Science Foundation (#1915724).

5.7 References

- Agrawal, A., & Paepcke, A. (2014). The Stanford MOOCposts dataset. Accessed: Dec, 15, 2020.
- Agrawal, A., Venkatraman, J., Leonard, S., & Paepcke, A. (2015). YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips.
- Alexandron, G., Ruipérez-Valiente, J. A., Chen, Z., Muñoz-Merino, P. J., & Pritchard, D. E. (2017). Copying@ Scale: Using harvesting accounts for collecting correct answers in a MOOC. *Computers & Education*, 108, 96-114.
- Buissink-Smith, N., Mann, S., & Shephard, K. (2011). How do we measure affective learning in higher education?. *Journal of Education for Sustainable Development*, 5(1), 101-114.
- Chi, M. T. (2009). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In *International handbook of research on conceptual change* (pp. 89-110). Routledge.
- D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2), 145-157.
- D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, 29, 153-170.
- Davidov, D., Tsur, O., & Rappoport, A. (2010, August). Enhanced sentiment learning using twitter hashtags and smileys. In *Coling 2010: Posters* (pp. 241-249).
- Doiron, J. G. (2016). Visual communication in peer assessment of online collaboration: Exploring the use of emojis for grading teamwork. *Proceedings of the Technology-Enhanced Collaborative Learning*, 71-74.
- Doiron, J. G. (2018). Emojis: Visual communication in higher education. *PUPIL: International Journal of Teaching, Education and Learning*, 2(2), 1-11.
- Dunlap, J. C., Bose, D., Lowenthal, P. R., York, C. S., Atkinson, M., & Murtagh, J. (2016). What sunshine is to flowers: A literature review on the use of emoticons to support online learning. *Emotions, technology, design, and learning*, 163-182.
- Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., & Riedel, S. (2016). emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*.
- Geller, S. A., Hoernle, N., Gal, K., Segal, A., Zhang, A. X., Karger, D., ... & Igo, M. (2020, March). # Confused and beyond: detecting confusion in course forums using students' hashtags. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 589-594).
- Grawemeyer, B., Mavrikis, M., Holmes, W., & Gutierrez-Santos, S. (2015, June). Adapting feedback types according to students' affective states. In *International Conference on Artificial Intelligence in Education* (pp. 586-590). Springer, Cham.
- Lin, F. R., Hsieh, L. S., & Chuang, F. T. (2009). Discovering genres of online discussion threads via text mining. *Computers & Education*, 52(2), 481-495.
- Mohammad, S., & Turney, P. (2010, June). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26-34).
- Neary, J. P. (2009). Putting the "new" into new trade theory: Paul Krugman's Nobel Memorial Prize in economics. *Scandinavian Journal of Economics*, 111(2), 217-250.
- Onah, D. F., Sinclair, J., & Boyatt, R. (2014). Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 proceedings*, 1, 5825-5834.
- Perkins, D. N. (1992). *Smart schools: Better thinking and learning for every child* (No. Sirsi) i9780028740188).
- Piaget, J. (2003). *The psychology of intelligence*. Routledge.

- Plaut, S. (2006). "I Just Don't Get It": Teachers' and Students' Conceptions of Confusion and Implications for Teaching and Learning in the High School English Classroom. *Curriculum Inquiry*, 36(4), 391-421.
- Rodrigo, M. M. T., Baker, R. S., & Nabos, J. Q. (2010). The relationships between sequences of affective states and learner achievement. In *Proceedings of the 18th international conference on computers in education* (pp. 56-60).
- Romig, J. (2015). Commenting by Emoji: A Tentative Glossary for Legal Writing Professors. *Bepress blog*.
- Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research, and Evaluation*, 11(1), 10.
- Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (1999). Assessing social presence in asynchronous text-based computer conferencing. *The Journal of Distance Education/Revue de l'education Distance*, 14(2), 50-71.
- Shusterman, E., Kim, H. G., Facciotti, M., Igo, M., Sripathi, K., Karger, D., ... & Gal, K. (2021, April). Seeding Course Forums using the Teacher-in-the-Loop. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 22-31).
- Tang, J. K., Xie, H., & Wong, T. L. (2015, July). A big data framework for early identification of dropout students in MOOC. In *International Conference on Technology in Education* (pp. 127-132). Springer, Berlin, Heidelberg.
- Vandergriff, I. (2013). Emotive communication online: A contextual analysis of computer-mediated communication (CMC) cues. *Journal of Pragmatics*, 51, 1-12.
- Wang, X., Wen, M., & Rosé, C. P. (2016, April). Towards triggering higher-order thinking behaviors in MOOCs. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 398-407).
- Wei, X., Lin, H., Yang, L., & Yu, Y. (2017). A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. *Information*, 8(3), 92.
- Wen, M., Yang, D., & Rose, C. (2014, July). Sentiment Analysis in MOOC Discussion Forums: What does it tell us?. In *Educational data mining 2014*.
- Yang, D., Wen, M., Howley, I., Kraut, R., & Rose, C. (2015, March). Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the second (2015) ACM conference on learning@ scale* (pp. 121-130).
- Yogev, E., Gal, K., Karger, D., Facciotti, M. T., & Igo, M. (2018, June). Classifying and visualizing students' cognitive engagement in course readings. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (pp. 1-10).
- Zeng, Z., Chaturvedi, S., & Bhat, S. (2017). Learner Affect through the Looking Glass: Characterization and Detection of Confusion in Online Courses. *International Educational Data Mining Society*.
- Zhang, A. X., Igo, M., Facciotti, M., & Karger, D. (2017, April). Using student annotated hashtags and emojis to collect nuanced affective states. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale* (pp. 319-322).

Chapter 6. Supporting Students' Self-Reporting Of Affective States In Online Course Forums And Text Annotation Platforms With Emoji Tags

6.1 Publication Status

Reviewing with co-authors before submitting for publication. Selections from this work have been incorporated in this chapter, indicated by italicized text.

Abstract: Emojis and hashtags are widely used tools that can contribute to the analyses of large volumes of student posts in educational forums. However, previous research shows that many emoji and hashtags can be semantically ambiguous and have context-dependent interpretations. One solution is to provide students with a custom set of emoji paired with a tag for one specific affect or function. In the social annotation platform Nota Bene (NB), forum participants are presented with a limited set of emoji (symbol) that are paired with a specific hashtag (label) to help precisely define the meaning of the emoji. In this study, we seek to validate and improve the set of emojis that correspond with hashtags available to tag student comments in NB. We distributed four surveys assessing (a) students' perception of the strength of the associations between the emojis and their instructor-assigned corresponding hashtags in the emoji set, and (b) the extent to which the emoji set captured the breadth of affects that students wanted to communicate while annotating the course text and discussing material with students and instructors. We collected responses from 648, 799, 175, and 418 students across four terms of a general biology course at a large public university. Student agreement on emoji meaning or function measured in a select-all-that-apply question showed improvement between the initial and final set of emojis and their corresponding hashtags (Survey 1 average degree-of-association: 74%, Kappa: 0.41; Survey 4 average degree-of-association: 91%, Kappa: 0.56). The surveys provided rich information for guiding student-driven revisions to the emoji set, expanding from 8 to 11 emoji-hashtag pairs, and allowed us to methodically study the range of

sentiments that students desired to express while posting in an online educational discussion forum.

6.2 Author Contributions

HGK, AB, ES, KG, JA, DK, MI, KNS, AA, and MTF contributed to the conception and design of the study. MTF and MI served as course instructors. HGK performed statistical analyses. HGK and MTF wrote the initial draft of the manuscript. HGK, AB, KG, JA, DK, MI, KNS, and MTF are in the review process for the final draft at the time of writing.

6.3 Introduction

Nota Bene (NB) was first introduced in Chapter 4 as a social annotation tool that facilitated student discussion and notetaking within the context of the course reading. Chapter 4 also described a study in which we sought to encourage quality discussions pertaining to course learning goals in NB via “seeding” the course forum with relevant, engaging posts from prior course instances. The following chapter detailed one approach to analyzing comments posted in NB to better understand students’ cognitive and affective states while they are reading the course text and interacting with others in the forum. Using the students’ self-reported affective states in the form of custom NB emoji-linked hashtags to train machine learning classifiers, we sought to identify instances where students expressed different sub-categories of confusion in untagged posts. In this chapter, we seek to validate and improve the set of emoji and hashtag pairings available to students in NB to tag their posts with in order to clarify the functions, and downstream analyses and interpretations, by students and instructors.

While potentially powerful, the use of emojis and/or hashtags in computational analyses is not always straightforward. Previous studies have revealed that the intended meanings of some emojis and hashtags can be semantically ambiguous, with lexical context, discrepancies

in cultural symbolism and linguistic rules, or imperfect alignments between vendor or operating system renderings all contributing to their interpretations (Guntuku, 2019; Shurick, 2020; Tigwell 2016; Budnik, 2019; Liu, 2012). Therefore, it is beneficial to be selective in terms of the options provided, and verify the alignment between the meaning and the image or text representation.

One approach to address these issues is to provide a limited set of well-defined and validated emoji and/or hashtags to forum participants. While restricting the affect-reporting dictionary constrains the possible spectrum of “tags,” it potentially improves the accuracy in the interpretation of what is reported. The annotation platform NB¹¹ uses a specific instance of this core idea by presenting forum participants with a limited set of emoji, each of which are associated with a specific hashtag label (Zyto, 2012; Zhang, 2017). This mechanism creates a unique set of emoji-hashtag pairs that link a symbolic (emoji) representation with a label (hashtag) that helps to precisely define the meaning of the associated symbol. Zhang et al. (2017) showed that this kind of system in NB could help classification algorithms achieve 83% accuracy in distinguishing between untagged student posts that express confusion and those that express curiosity. In a follow up study that also incorporated emoji-hashtag data, a classifier distinguished between four different subtypes of confusion with high (90.67%) agreement with instructor-tagged comments (n=600 comments) (Geller, 2020; See Chapter 5).

For this system to work best, however, the meaning of the image and hashtags text should be strongly associated with one another. We hypothesize that gathering data on how students interpret emoji can enable us to develop a set of emoji/hashtag that readers find useful and have a high unique association between hashtag and emoji pairings. Here, we discuss the student-informed development of a new set of emoji-hashtag pairs that we propose increases the reliability of affect reporting in the context of an educational discussion forum. To this end, we used student surveys to iteratively evaluate and improve sets of emoji-hashtags pairs with

¹¹ <https://www.csail.mit.edu/>

regard to: (a) the extent to which students feel a strong and unique association between the emojis and their assigned hashtags, and (b) the extent to which the emoji set adequately captures the breadth of affects that students wish to be represented. Overall, the new emoji-hashtag set is intended to provide more accurate information about the student affective state and improve practicality in machine learning functions or other analyses designed to uncover how students interact with online reading materials.

6.4 Methodology & Data Collection

6.4.1 The NB Discussion Interface

Students enrolled in a general biology course at a large public institution accessed course readings, authored and edited by course staff and former students, using the online Open Educational Resource, LibreTexts (<https://libretexts.org>). As shown in Figure 6.1a, students used the annotation tool NB (Zyto, 2012) to collaboratively annotate and discuss online course content in the margins of the readings, either as original comments or as a reply to another comment in a threaded discussion. A course reading assignment with various portions of text annotated with student comments anchored to highlighted text is illustrated. To explore comments and discussion threads, a reader can either click on the highlighted text or scroll through the list of student comments on the right and click to expand. To make a new annotation, one selects text of interest and uses the comment box (in the bottom right corner) to write a comment.

The buttons below the text box, as seen in Figure 6.1b, allow for students to quickly select a hashtag affect to express and “tag” their comment with; clicking on any of the emoji will automatically insert the paired hashtag as text. As an example, clicking on the question mark emoji will automatically insert “#question” in the text box. All surveyed emoji-hashtag sets can be found in Appendix 6.1.

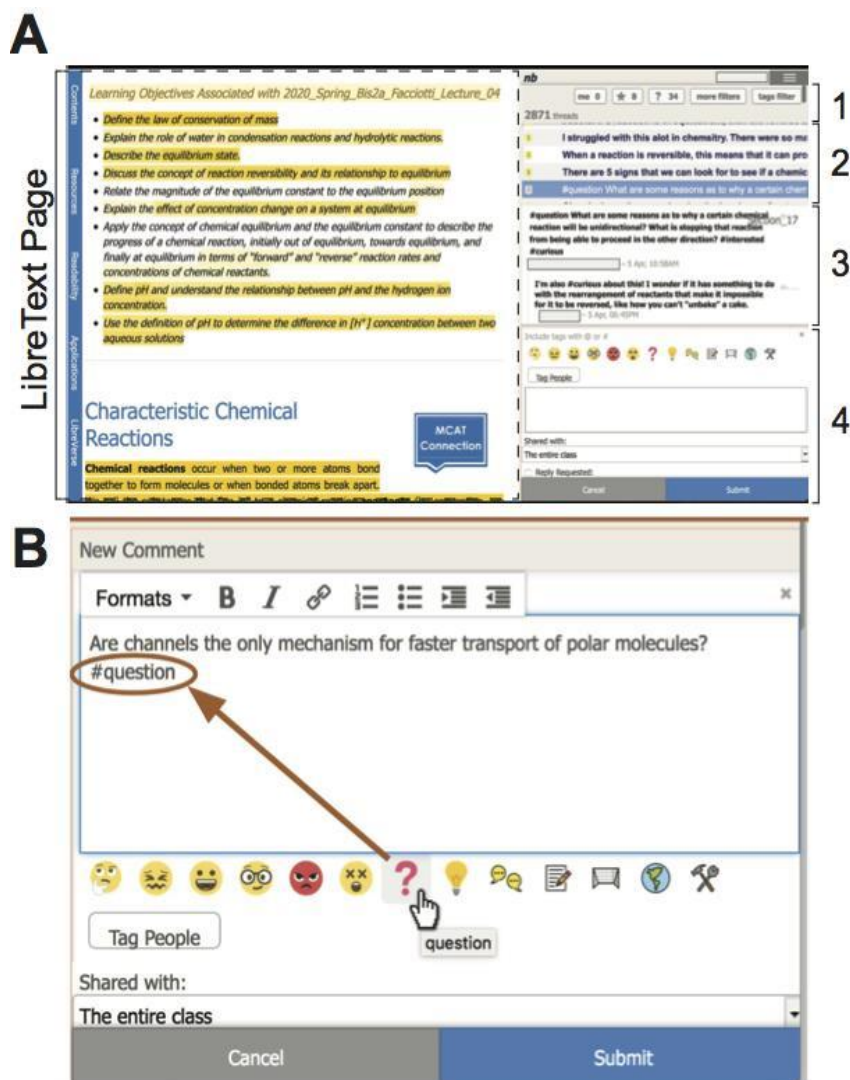


Figure 6.1 a. The NB interface (right) overlaid onto the general biology online course textbook (left, outlined in a dashed line box), showing (1) filtering options (e.g., “me” for self posts, emojis, thread length), (2) a list of one-line previews of student comments in the specified reading, (3) a full-thread view of a selected student comment or thread, and (4) a text box for student comments, with a clickable emoji toolbar; **b.** A demonstration of how the emoji-hashtag tool within NB functions.

6.4.2 Course Format And Study Participants

In this study, we analyzed survey responses submitted by 648, 799, 175, and 418 students enrolled, respectively, in four separate course instances of a general biology course: Winter Quarter 2020 (W20), Spring Quarter 2020 (S20), Summer Session 1 2020 (SS20), and Winter Quarter 2021 (W21), respectively (Table 6.1). The Spring, Summer Session, and Winter

2021 course instances were remote courses due to the COVID-19 pandemic, while the Winter 2020 course was taught in person.

Table 6.1. Course and survey information and statistics.

Course Term	W20	S20	SS20	W21
Emoji set surveyed	1	2	3	4
Number of enrolled at time of survey administration	876	990	216	798
Number of survey respondents	777	917	191	521
Number of respondents after QC and opt in filtering	648	799	175	418
Emoji Survey Question Style	Select-all-that-apply among a list of hashtag affects	Students self-generate the hashtag affect	Select-all-that-apply among a list of hashtag affects	Select-all-that-apply among a list of hashtag affects
Number of posts	64,033	81,696	19,718	29,247
Number of posts with emoji-hashtag (% of total posts)	44,044 (68.78%)	51,600 (63.16%)	12,081 (61.27%)	20,608 (70.46%)

In all course instances, students could earn up to 4 points for completing one of 26 NB assignments. On each assignment, students earned up to 3 points for making three substantive comments (one point per comment) and an additional point for using an emoji in at least one of their posts. Substantive comments could be either original posts or replies in a thread. Since each reading assignment accounted for 4 out of 1,000 total course points, credit earned using NB contributed to ~10% of a students' total course grade. In all but W20, students were given weekly reading assignments (2-3 readings per week, each corresponding to a lecture) and comments for all two/three readings were due on Friday (23:59). In W20, students were assigned 2-3 readings per week (with each reading corresponding to a lecture) and comments for each reading were due the morning of and before the respective lecture.

Students were asked to opt-in to NB studies at the beginning of each term, and the study data and student comments were used for ML classifier training in accordance with institutional review #1456274-1.

6.4.3 Survey Design, Administration, And Quality Control

The primary objective of this study was to improve upon the original emoji-hashtag pair set used in NB. We test the hypothesis that we could meet the following two design criteria: First, an improved emoji-hashtag pair set should demonstrate that students perceive more unique associations between their symbolic representation and their associated hashtag. Second, the emoji set should adequately capture the breadth of affects that students report they wish to be represented. Our strategy was to use student surveys to guide iterative assessment and refinement of label and emoji pairs that could work well together (Figure 6.2).

In W20, we conducted a survey to assess how closely students associate the original emoji with their paired hashtags (Zhang, 2017). The results were used to guide refinements to the original emoji set. In S20, we incorporated an alternative open-ended style format to solicit broader, student-derived hashtags that could be paired with the emoji. Finally, in SS20 and W21, we resumed refining and testing modifications to emoji-hashtag pairs using aggregated knowledge derived from the preceding surveys. We compare measures of student consensus on emoji-hashtag pairing fit and open-ended feedback across the four surveys to drive and evaluate improvements of the reliability of our affect reporting tool. In each course iteration, a small number of extra credit points were given to each student who submitted a survey response, regardless of opt-in status.

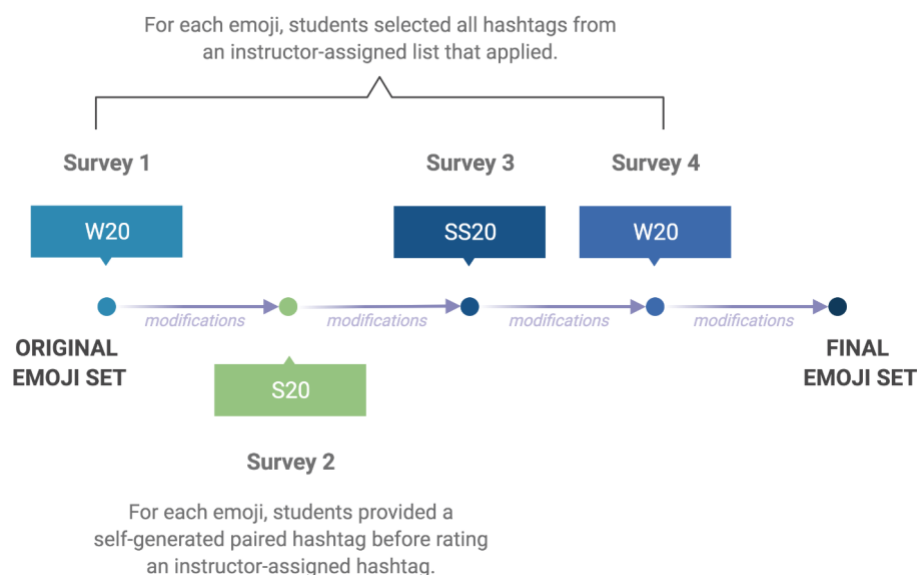


Figure 6.2. Overview of iterative surveying and emoji modification process.

The core of each of the four surveys involved presenting individual emojis in a randomized order, and investigating student consensus surrounding the “best fit” for the hashtag pairing. Each item of Surveys 1, 2, 3, and 4 can be found in Appendices 6.2-5, respectively. Survey 1 examined the original 8 emoji-hashtag pairings on NB: interested, question, curious, frustrated, useful, help, idea, and confused (Zhang, 2017). Surveys 2-4 iteratively assessed modified versions of the emoji sets (Appendix 6.1), which were strongly motivated by student feedback from the surveys.

Surveys 1, 3, and 4

For Surveys 1, 3, and 4, student perceptions were gathered via a select-all-that-apply style question (“What feelings or impressions does the emoji convey to you? Select one or more terms from the list below.”); students were shown one emoji at a time (i.e., the light bulb icon) and were provided a list of options to select if applicable, consisting of a randomized-order list of the NB paired-hashtag affect terms, as well as an “Other: (text input)” option. In Survey 1, there was an additional hashtag affect (flustered). This was included as a “control” term to see how many students would select a term that did not belong to the set of terms they were familiar with

for the various emojis, as well as to measure the extent to which each emoji was associated with that affect, recognizing the implications for learning.

Survey 2

In Survey 2, students were instead required to provide their own hashtag affect in short text format, so as to provide a response that aligns closer to their self-derived perception of the emoji – thereafter, they were asked to assess the given instructor-generated emoji-hashtag pairing suggestion.

In all surveys, students were also asked questions about perceived usefulness of the emoji (e.g., do they want the option to express the emoji, how frequently do they want to express the emoji). Finally, the students were solicited for open-ended feedback on the emoji set, at which point they could offer revision suggestions or thoughts otherwise. All surveys were administered using Qualtrics <qualtrics.com>.

Quality Control

After downloading each set of survey data and removing identifying information, keeping only the Qualtrics-assigned unique identifier, incomplete surveys were removed from the dataset, as well as any responses completed in under 3 minutes. In Surveys 3 and 4, an additional quality control question was inserted, asking students to select “Agree” as their answer choice – those who selected otherwise had their responses removed from the dataset.

Calculation of Interrater Reliability

Interrater reliability was used to indicate the level of similarity and dissimilarity in students’ perception of the emoji-hashtags. For the select-all-that-apply questions, each student’s response was disaggregated by possible term into a new dataframe; a “1” would denote that the student selected that term (whether or not other terms were selected) while a “0” would denote that the student did not select the term. If a student selected “Other:” and wrote in

a short-text response, the score was marked as a “1” for “Other” – the short-text response was not considered during statistical analyses and was instead qualitatively analyzed. Due to the fully-crossed design of this study that uses nominal variables (i.e., every student rated every item with categories as selections), we used the irr package in R (Gamer, 2019) to obtain the Light Kappa score (Hallgren, 2012), an arithmetic average of the Cohen’s Kappa score calculated across all pairwise raters (students) for each emoji to account for agreement that occurs by chance.

According to Landis & Koch (1997), a Kappa score of <0 corresponds to no agreement, 0-0.20 to slight agreement, 0.21-0.40 to fair agreement, 0.41-0.60 to moderate agreement, 0.61-0.80 to substantial agreement, and 0.81-1.00 to perfect agreement. Using these guidelines, we set a baseline of 0.4 as slight agreement, 0.6 as good agreement, and 0.8 as substantial agreement (Appendix 6.6).

6.5 Results

6.5.1 Original Emoji Evaluation (Survey 1)

Table 6.2 reports the percentage of total respondents of Survey 1 who selected each available term (rows) for each emoji (columns) presented (e.g.,when presented with the light bulb emoji, 99% of respondents selected the term “idea,” 11% of respondents selected the term “interested,” etc.). Values in cells along the diagonal indicate the frequency with which respondents selected the instructor-assigned hashtag. We refer to these on-diagonal associations as the “degree-of-association” (DA). Off-diagonal selections indicate that students associated the emoji with one or more alternate labels and helps reveal the lexical ambiguity of each emoji. Here, lexical ambiguity is defined as the ability to associate each emoji with two or more terms. In the context of our design objectives, emoji-hashtag pairs evaluated in this manner should strive to maximize on-diagonal scores and minimize off-diagonal elements. Light

Kappa scores, calculated as described in Methods, are also reported. This metric quantifies the degree to which pairs of students selected the same set of hashtag label(s) for each emoji and provides an estimate of the reliability of agreement amongst students for the primary and alternate emoji-hashtag associations. Finally, we report the percentage of total hashtags corresponding to each emoji found in the readings for the academic term surveyed. We interpret this term as an indirect measure of emoji-hashtag pair utility.

Degree-of-Association

Students reported high DA between the following emoji and their respective instructor-assigned hashtags: #idea (99% DA), #question (94% DA), and #frustrated (91% DA). By contrast, emoji showing lower pairing agreement with their instructor-assigned hashtag were: #help (43% DA), #useful (57% DA), and #confused (62% DA). Meanwhile, the two emoji/hashtag pairs #useful and #help stimulated respondents to provide alternative self-reported hashtag labels 19% and 14% of the time, respectively, by selecting the “Other” option.

Lexical Ambiguity

Since students were given the option to select more than one hashtag term, we calculated the distribution of students who selected one or more terms for each emoji (Appendix 6.7). This analysis provided a perspective on how strongly or weakly students uniquely associated each emoji with a specific hashtag. The lightbulb emoji was uniquely associated with a single term by 73% of respondents and with either 1 or 2 terms by 93% of students. The instructor-assigned #frustrated emoji had the next highest single association between emoji and one or two terms, 50% and 84%, respectively. By contrast, the instructor-assigned #question, #confused, and #curious emoji had the greatest percent of respondents associating the emoji with three or more terms, 42%, 37%, and 34%, respectively.

Reliability

Using the aforementioned Kappa score boundaries, five of the eight emoji-hashtag pairs show slight agreement (#help, #curious, #useful, #interested, #confused) in the combination of primary and alternate terms to associate with the target emoji, two have good agreement (#question, #frustrated), and just one of the emojis (#idea) has substantial agreement.

Table 6.2. Frequency of responses from Survey 1 (% of total respondents), Light Kappa scores, and proportional emoji use.

								
	#interested	#question	#curious	#frustrated	#useful	#help	#idea	#confused
interested	0.77	0.02	0.23	0.00	0.55	0.01	0.11	0.00
question	0.04	0.94	0.26	0.01	0.00	0.04	0.03	0.03
curious	0.36	0.21	0.71	0.00	0.09	0.02	0.05	0.01
frustrated	0.01	0.08	0.06	0.91	0.00	0.29	0.00	0.47
useful	0.41	0.01	0.02	0.00	0.57	0.00	0.15	0.00
help	0.04	0.46	0.09	0.10	0.02	0.43	0.02	0.33
idea	0.28	0.02	0.25	0.00	0.34	0.01	0.99	0.00
confused	0.02	0.51	0.51	0.15	0.00	0.56	0.00	0.62
flustered	0.01	0.10	0.05	0.42	0.01	0.37	0.00	0.60
Other:	0.06	0.00	0.04	0.09	0.19	0.14	0.02	0.05
Light Kappa	0.36	0.47	0.27	0.53	0.32	0.26	0.69	0.37
Emoji Use	0.255	0.151	0.179	0.004	0.175	0.017	0.183	0.036

Utility

Through the survey, participants also reported differences in the perceived utility of each instructor-assigned emoji/hashtag pairs (Appendix 6.8). Those reported by students to be most useful include: #question, #idea, #curious, and #interested (87-96% of respondents reported

these emoji/hashtag pairs somewhat or very useful). The instructor-assigned emoji/hashtag pairs #useful, #confused, and #help were reported as being somewhat or very useful by 61-78% of respondents. Meanwhile, only 48% of respondents reported finding the instructor-assigned emojis/hashtag pair #frustrated somewhat or very useful. While infrequent use of an emoji may not necessarily imply a lack of value, we note that the three lowest ranked emoji in terms of usefulness (#frustrated, #help, and #confused) were also the least frequently used (Table 6.2).

Unpacking confusion

From our previous work (Geller, 2021; See Chapter 5), we observed that students might be trying to communicate one or more subcategories of confusion using the single #confused emoji-hashtag pair. Survey 1 data (Table 6.2) confirmed this observation, showing that over 50% of students associated the confusion affect with the emojis for #question, #curious, and #help, in addition to that for #confused. Here, we sought to understand whether DA and reliability of the emoji-hashtag set could be improved by disambiguating different types of confusion and assigning one or more subtypes of confusion with their own emoji-hashtag pair(s).

To this end, we first confirmed whether various confusion-related statements accurately described how students experience confusion. Example statements included: “I am feeling lost,” and, “I am worried about knowing this information for an exam.” Over 75% of the students indicated that all eight of the presented confusion-related statements were somewhat or very accurate descriptors of how they experienced confusion in the course (Appendix 6.9). Then, students were prompted to select the most appropriate emoji within the set to pair with each statement. For all but two statements, at least two or more emoji were identified as the best-fit by 20% or more of the survey respondents (Appendix 6.9). The two emoji most often associated with multiple statements were those initially assigned the #confused and #help labels. This revealed that even when a complex, multifaceted sentiment like confusion is disaggregated into

subcategories, it is still challenging to unambiguously associate one or more of the initial limited palette of emoji to subcategories of complex affects like confusion. We concluded that in this case adding new emoji and hashtags to our constrained dictionary would be necessary to help resolve some of the existing ambiguity.

Contribution of symbol versus hashtag on tag selection

We sought to understand the extent to which the emoji itself, versus the associated hashtag label, informed the participant's decision of which tag to select. Data reported in Appendix 6.10 demonstrate that while both emoji and hashtag contribute in informing the participants' choices, the hashtag label has greater influence on their choice of emoji-hashtag pairs choice than the emoji; 35% of students indicated that their selection was heavily informed by the hashtag, while only 23% of students said they were heavily informed by the emoji. Furthermore, 14% of students indicated that the image does not inform their selection as opposed to the 5% for hashtags. While it is not surprising that the hashtag label primarily motivates the selection of emoji-hashtag pairs (i.e., the sentiment/meaning of the pair is explicitly stated by the label), it is nevertheless important to appreciate that 86% of study participants still reported being informed by the emoji in some capacity. These data suggest that optimizing the degree of perceived association between emoji and their associated hashtags might serve to improve the usability and interpretability of student intent in comments tagged with emoji-hashtag.

Open-ended feedback

We explored student opinions about the emoji-hashtag pairs as NB features and specific emoji-hashtags within the original set. The feedback also contained valuable information about student rationale for emoji usage, or lack thereof. Roughly half of the class also suggested that the set of hashtag labels could benefit from modifications and many suggestions were made. Examples of the most common suggestions include: #connection, #important, #overwhelmed,

#surprised, #agree, #exam-question, #learning-goal, #requestTA, and #clarify. The full set of label suggestions and feedback can be accessed through Dryad (see Accessing Material). The following are selected quotes from the open-ended feedback section of Survey 1:

- *“I think the #help icon could be changed. On social media, that kind of icon is more associated with “I’m dead,” meaning something is super funny.”*
- *“Too much emoji, hard to memorize their meaning sometimes”*
- *“After a while you get used to them and know which one to use at the right time. You start getting used to using them to express how you feel. And you know which one stands for which hashtag”*
- *“I personally really liked using the emojis because not only did it help other people understand where I was coming from, but it also helped me evaluate why I was making the comment in the first place. It organized feelings of having an idea, needing help, and being confused.”*
- *“keep the hashtag that corresponds to the emoji because the meanings of the emojis can be a little ambiguous.”*

As evidenced by the quotes above, and the full feedback provided, students have devised strategies for emoji-hashtag usage and have provided rationale for why they use certain emojis over others, stating reasons such as redundancy or poorly matched emoji and hashtag pairs (e.g., “On social media, the [#help] icon is more associated with “I’m dead”). In addition, some students reacted positively towards the concept and implementation of the emoji-hashtag pairs, stating associated benefits such as fun, usefulness, “helpfulness when studying,” and that they are a good mechanism to enable quick communication and filtering of student comments, among others. On the other hand, other students expressed dislike for the emoji-hashtag system, citing disapproval of the requirement to use them, and stating that their usefulness was unclear.

Lessons learned

The NB emoji-hashtag tool, by design, attempts to constrain the meaning of emoji by associating their images to pre-defined hashtag labels. However, Survey 1 reveals that despite our attempt to constrain the meaning of each emoji some innate ambiguity remains around which of the available hashtag labels best associates with each emoji. This is particularly

important given that participants report selecting emoji using both the emoji and hashtag labels. The ambiguity, therefore, contributes to noise in the reporting of affect and other sentiments that impacts the power of the instructor or computational tools that may try to use emoji-hashtag pairs to sort, interpret and learn from tagged comments. Together with our discovery that the existing emoji hashtags were insufficient for successfully disambiguating a complex affect like “confusion” with unique emoji-hashtag pairs this reinforced the idea that new emoji and hashtag could be usefully paired into emoji-hashtag set with the properties outlined in our design criteria.

Collectively, Survey 1 revealed areas for improving DA, reliability and for reducing ambiguity, provided a baseline from which to evaluate the influence of modifications to the emoji-hashtag pairs on our core metrics, and helped identify a strategy for improvement. Survey 1 data also helped directly inform possible modifications to the emoji-hashtag set. Some specific modifications coming directly from student suggestions, included creating emoji-hashtag pairs centered on communicating learning states (surprised, overwhelmed) or course-pertinent information (examable, learning goal, request TA).

6.5.2 Modified Emoji Set Evaluations (Surveys 2 And 3)

Survey 2 examined the first modified set of emoji-hashtag (Figure 6.3), with revisions made to the original emoji set based on results from Survey 1 and instructor discussions. For example, #confused, an emoji-hashtag pairing with lower agreement, became newly associated with #flustered hashtag in an attempt to begin disambiguating the functions and usages of the confusion-related emojis. The full rationale behind all emoji revision decisions can be found in Appendix 6.11.

Between the original and modified set, some emoji stayed the same (#frustrated, #question), some emoji changed hashtags (#curious to #justcurious; #confused to #flustered; #useful to #engaged; #interested to #learning; #help to #whoa; #idea to #explanation), and additional emoji were added (#learninggoal, #connection, #texterror, #pleaseclarifyinclass,

#willthisbeontheexam, #requestTAattention, #disagree). These emojis and hashtags were the subject of Survey 2. First, we sought to further broaden our understanding about how students associated emojis and hashtag affects by soliciting students' self-generated definitions or hashtag affects for the emojis in isolation. Thereafter, we evaluated student perceptions on hashtag alignment usefulness, and expected use of the revised instructor-generated emoji-hashtag set.



Figure 6.3. *The second emoji-hashtag set; tested in Survey 2.*

Student-generated tags from free response prompting in Survey 2

Free response prompting revealed the diversity of ways in how students interpret and assign the meaning of the emojis. Appendix 6.12 contains a list of the hashtag affect terms offered for each emoji that had a count of 100 or more students. The count values for the top hashtag affect suggested for each emoji vary between emojis, the highest count being 720 for question (90.11% of survey respondents wrote “question”) and the lowest being 156 for learning (19.52% wrote “smart”), with an average top term count of 354 students, or 44%. This means that on average, each emoji elicited the same, or an extremely similar hashtag affect (e.g., surprising, surprised, and shocked), and that over half of the survey respondents diverged from the most popular affect, though many of the offered terms are still thematically adjacent. For example, for the new flustered emoji students responded with “confused” (35%) and “frustrated” (17%). Of note, the students were already acclimated to the original emoji set since the revised set was not implemented at this point, only tested on in Survey 2. Overall, the variation within

the responses makes sense for some of the more symbolic emoji (e.g., globe for #connection), and shows that emoji interpretation is context- and environment-dependent. The vast diversity within the responses in general also tells us that there may be a maximal level of “singularity” in meaning for the emoji-hashtag fit, or in other words, a baseline level of noise in emoji meaning interpretation.

Utility in Survey 2

Afterwards, students were shown the instructor-suggested hashtag and were asked to rate three sentiments on a three point-Likert scale: (1) whether the provided hashtag text was appropriate for the emoji, (2) whether they wanted to express the hashtag affect, and (3) how frequently they would want to express it (frequently versus rarely) (Appendix 6.13). In general, there is a positive correlation between having an appropriate emoji-hashtag fit, wanting to express the proposed hashtag, and frequently using the emoji-hashtag in NB. The several emoji that students seem to want to express but could benefit from improving the paired hashtag affect are #pleaseclarifyinclass and #explanation.

Next, we wanted to get a sense of the student’s perceived frequency of current emoji usage (Appendix 6.14). This would allow us to contextualize the results from Appendix 6.13 in the context of current emoji usage perceptions and see which of the original emoji ended up being used infrequently, keeping in mind that low usage frequency does not necessarily mean that the emoji does not have good utility; some emoji might be used less but are important when they are used. Emoji that tended to be on the extreme end of low usage were #help and #frustrated. The majority of the emojis were used at least once, however, with the lowest rated emojis #help and #frustrated, as many as 21% and 31% of the survey respondents, respectively, indicated that they did not use the emoji at all. Notably, #frustrated and #flustered (previously #help) were the 5th and 7th (respectively) lowest emoji-hashtags predicted in terms of frequency of usage in the previous question (Appendix 6.13).

Open-ended feedback in Survey 2

Lastly, the open-ended feedback in Survey 2 revealed that while some students were excited about adding new emojis to the set, other students were less enthusiastic about having excessive additions, with a number of students commenting that too many emojis would be too confusing and/or counterproductive. Nonetheless, many students offered suggestions of other emojis or hashtags to include, indicating that the emoji set may still benefit from adding more emoji-hashtags. Common suggestions include: #connection, #important, #surprised, #agree, #exam-question, #learning-goal, #requestTA, #clarify. Emoji redundancy and/or ambiguity is still a common theme found within the comments. The following are selected quotes from Survey 2:

- “The emoji needs to somehow impact the way that the instructor, TA, or classmate responds to a comment. There are too many emoji's that simply translate into a question mark. We need a way to distinguish questions that need some clarification from questions that need a whole description of the concept and everything in between.”
- “I don't think that the frustrated emoji is needed, or it just didn't match the hashtag well. there needed to be a wider range of emoji's because I felt like I use the same to almost every single time. I really like how in the new site there are emojis to ask the TAs or professor.”
- “I would not use the emojis if they were not required. However, I did not find them to be a burden and I am actually glad that we were pushed to use them. They would help me frame a comment”

Again using the data collected from this survey, Survey 1, and course instructor insight, a further modified emoji set was created, for which we sought student input in the following survey.

Modifications informed by Surveys 1 and 2, tested in Survey 3

Survey 3 examined a second modified set of emoji-hashtag; revisions were made to the previously modified emoji set tested in Survey 2.

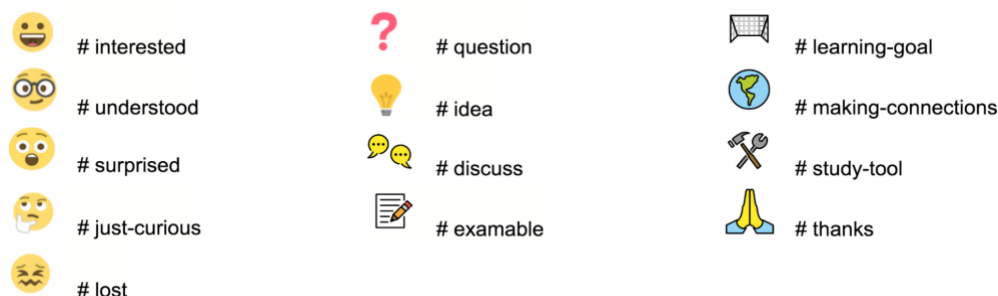


Figure 6.4. The third emoji-hashtag set; tested in Survey 3.

Some emoji stayed the same (e.g., #question, #just-curious) and some changed affect terms (e.g., #engaged to #interested; #learning to #understood; #explanation back to #idea). Previously #flustered, the newly paired hashtag became #lost, to provide a potentially more common way to express extreme confusion, which students indicated wanting to express in the previous survey. One emoji, #surprised (previously #whoa), had a slight modification of facial expression. There were also several additions of pairings (#discuss, #examine, #learning-goal, #making-connections, #study-tool, and #thanks), as well as some removals (#frustrated, #help). Survey 3 was designed to include elements of Surveys 1 and 2, and included both select-all-that-apply questions from Survey 1 and the Likert style utility questions from Survey 2.

Emoji-Hashtag DA Score & Lexical Ambiguity in Survey 3

The data for the Survey 3 select-all-that-apply question is available in Appendix 6.15. When compared with Survey 1, with a few exceptions (#understood, #interested, #just-curious, and #examine), the new emoji set seemed to have fewer off-diagonal selections, as indicated by the decrease in the average percentage of students who selected 3 or more affect terms in the select-all-that-apply question across the emoji set, from 26% in Survey 1 to 16% in Survey 3 (See Appendix 6.16 for the full data on distribution of the number of affect selections for Survey 3). In addition, the average DA across is higher in Survey 3 (81% DA) than in Survey 1 (74% DA), broadly indicating stronger associations with the modified emoji-hashtag pairings (Figure

6.4). As one example, the #lost emoji, which was originally #confused, increased from 62% DA in Survey 1 (Table 6.2) to 90% DA in Survey 3.

Reliability in Survey 3

The average Light Kappa of the whole emoji-hashtag sets were comparable between Survey 1 (0.41) and Survey 3 (0.43), indicating that there was no improvement in interrater reliability across the entire emoji set. The emoji with the lower ranking Light Kappa scores, which in the methodology section, were classified as poor agreement, were associated with: #understood (Kappa: 0.23), #interested (Kappa: 0.28), and #examable (Kappa: 0.25). On the other end, the emoji-hashtag pairings with the highest Light Kappa scores were #lost (0.62), #question (0.57), and #surprised (0.54). Notably, all three of these emoji-hashtag pairings had questionable agreement in previous surveys.

Utility in Survey 3

Appendix 6.17 shows data for the appropriateness of the emoji-hashtag pairing fit, desire to express, and predicted frequency of expression of each emoji. The two emoji with low values in all categories were #connection and #texterror, despite students in previous survey iterations recommending these emoji. The fact that the six lowest values for the frequency category were new emoji may indicate that students are reluctant to, or underestimate, their usage of novel emoji.

Open-ended feedback in Survey 3

In the open-ended feedback, students were largely split into three groups: students who preferred the original emoji set (which they had been using all quarter), those who preferred the emoji set being surveyed on, and lastly, those who advocated for a mix between the two sets. Students belonging in the first group frequently commented on there being too many new, distracting emoji-hashtags, similar to the sentiment in Survey 2 open-ended feedback. Other

students commented on ill-fitting emoji-hashtags, citing specific issues for certain emojis. Students who preferred the modified emoji-hashtag set that was surveyed in Survey 2 stated appreciation for a greater diversity of options that were useful for learning. A large number of students urged a blending of the original emoji-hashtag set with the new emoji set, stating that many of the changes were beneficial and indeed improvements, but not all. Suggestions from this group include adding many of the new emoji-hashtags, but not all, like #disagree, or changing the hashtag pairing to make it more intuitive and less symbolic. The following are selected quotes from Survey 3:

- “I feel like the top ones would be seen on like Facebook reactions and are kind of limiting. Below there are way more varying options”
- “I always got confused between the #useful and #interested emojis and would frequently pick the wrong one, but overall I think the emojis were useful to categorize and sort through comments.”
- “I believe Emoji Set 2 is useful and helpful but can be overwhelming with too many options, and only select, necessary emojis should contribute to Nota Bene.”

6.5.3 Final Emoji Set Evaluations (Survey 4)

The data from Survey 3, along with the previous surveys, provide a rich source of student input which was heavily considered in the construction of a third modified emoji set, which was tested in Survey 4 (Figure 6.5, Appendices 6.11, 6.18-21).

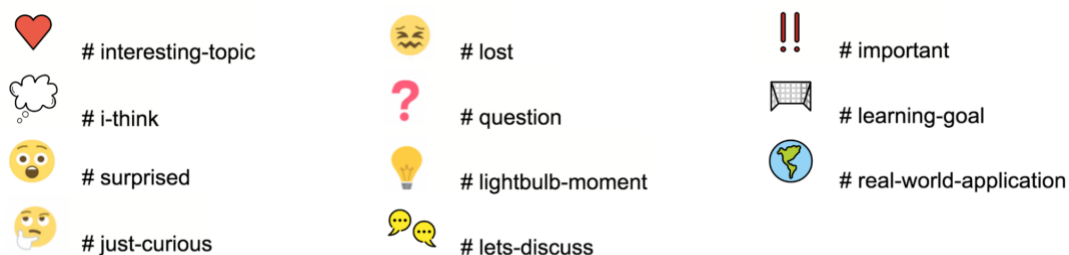


Figure 6.5. The fourth and final emoji-hashtag set; tested in Survey 4.

Emoji-Hashtag Association Score

As done in the select-all-that-apply question analyses from Surveys 1 and 3, we observed the percentage of students who selected the original affect term associated with the emojis provided in Survey 4 (the diagonal in the matrix in Table 6.3). Although there are still emojis with “off-diagonal,” or competing and/or overlapping affect terms in Survey 4 (e.g., #just-curious and #question), we see that between Survey 1 and 4, the average DA improved from 74% to 91%, with the Light Kappa score improving from 0.41 to 0.56 (Table 6.4).

Table 6.3. Frequency of responses from Survey 4 (% of total respondents)

											
	#interesting-topic	#surprised	#just-curious	#lost	#lets-discuss	#i-think	#question	#important	#learning-goal	#real-world-application	#lightbulb-moment
interesting-topic	0.71	0.19	0.22	0.00	0.12	0.07	0.00	0.16	0.01	0.06	0.10
surprised	0.01	0.97	0.01	0.00	0.00	0.01	0.02	0.36	0.00	0.00	0.04
just-curious	0.02	0.02	0.70	0.01	0.04	0.26	0.17	0.00	0.00	0.00	0.01
lost	0.00	0.07	0.13	0.96	0.00	0.04	0.41	0.04	0.03	0.01	0.00
lets-discuss	0.02	0.01	0.14	0.01	0.98	0.28	0.06	0.03	0.02	0.03	0.04
i-think	0.01	0.01	0.63	0.04	0.08	0.87	0.16	0.01	0.00	0.00	0.18
question	0.00	0.01	0.25	0.06	0.04	0.12	0.99	0.02	0.00	0.00	0.00
important	0.22	0.05	0.01	0.00	0.01	0.01	0.01	0.93	0.05	0.07	0.14
learning-goal	0.06	0.00	0.00	0.00	0.02	0.02	0.00	0.12	0.89	0.04	0.11
real-world-application	0.04	0.00	0.00	0.00	0.02	0.01	0.00	0.02	0.05	0.99	0.02
lightbulb-moment	0.01	0.23	0.01	0.00	0.01	0.01	0.00	0.07	0.01	0.01	0.98
Other: _____	0.29	0.04	0.07	0.24	0.02	0.05	0.02	0.04	0.07	0.03	0.06
Light Kappa Score	0.38	0.58	0.36	0.69	0.69	0.44	0.58	0.50	0.65	0.76	0.54

Returning to the trajectory of the #lost emoji – originally, the emoji was paired with the #confused tag, and in Survey 1 had 62% DA (Light Kappa: 0.37). In Survey 2, we assessed its fit with #flustered and found that students associated it more closely with “confused” and “frustrated.” In an effort to disambiguate the emoji functions within the set, we changed the tag to #lost, where in Survey 3 it had 90% DA (Light Kappa: 0.62). In light of the high association, the emoji-hashtag pairing was kept the same, and in Survey 4 reached 96% DA (Light Kappa: 0.69). This data indicates that the emoji-hashtag pairing has grown stronger over iteration, and potentially that the distinctions between the confusion-related emojis are less ambiguous.

Reliability

Figure 6.6 and Table 6.10 summarize the DA and Light Kappa score across the three surveys that utilized the select-all-that-apply question. In Survey 1, five of the eight emoji (62.5%) had a Kappa score of less than 0.4, our baseline of good reliability, while in Survey 4, that number decreased to 2 of 11 emojis (18.2%). Using these metrics, these trends suggest that the fit, or the strength of the association, between the emoji and the paired hashtag have generally improved over emoji set modifications and survey iterations.

Open-ended feedback

The suggested modifications and other general open-ended feedback showed that some students were enthusiastic about the modified emoji-hashtag set and the changes, while others thought it could benefit from further improvement. The feedback in this survey tended to be focused more on the emoji-hashtag fit, rather than the quantity of emoji within the set, compared to previous surveys. The following are selected quotes from Survey 4:

- “For the most part, all the emojis and their corresponding hashtags mesh very well. The only emoji I don’t necessarily ‘belong’ with the other emojis is the heart emoji with the #interesting-topic.”*
- “I felt that all the emojis other than question, lost, and important were irrelevant. Get rid of the requirement of emojis.”*
- “I think we are very used to putting emojis in our texts, so adding a few more will always be beneficial”*

- “I think Nota Bene is great! It's a great way for students to ask questions and help each other better understand topics. The hashtags were also helpful because they allowed me to better understand where students may have needed help. For example, I think when students used the "question" hashtag, I would get a notification. I think the least used emoji were #real-world application and #learning-goal. Although this was the case, I don't think they should be removed because they have the potential to provide interesting/important information when used.”
- “The tags are helpful to be clear about what point [I] am trying to make.”

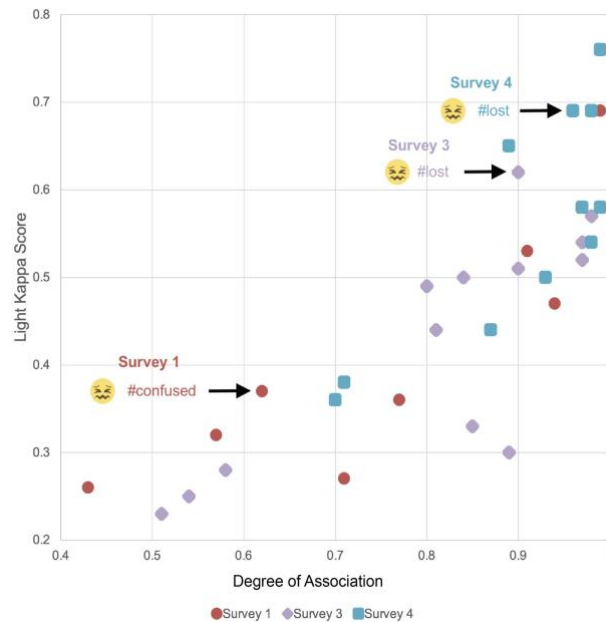


Figure 6.6. DA and Light Kappa Score for emoji surveyed in Surveys 1, 3, and 4. Each point represents one emoji within the set, with the trajectory of the original #confused emoji tracked across the survey sets.

Table 6.4. Summary data of DA and Light Kappa scores for Surveys 1, 3, and 4.

	Survey 1	Survey 3	Survey 4
Average DA	0.74	0.81	0.91
Average Light Kappa	0.41	0.43	0.56
# of emoji with Light Kappa: ≤ 0.30	2 of 8 (25%)	4 of 13 (30.8%)	0 of 11 (0%)
# of emoji with Light Kappa: 0.31-0.40	3 of 8 (37.5%)	1 of 13 (7.7%)	2 of 11 (18.2%)
# of emoji with Light Kappa: 0.41-0.50	1 of 8 (12.5%)	3 of 13 (23.1)	1 of 11 (9.1%)
# of emoji with Light Kappa: 0.51-0.60	1 of 8 (12.5%)	4 of 13 (30.8%)	3 of 11 (27.3%)
# of emoji with Light Kappa: $0.60 \leq$	1 of 8 (12.5%)	1 of 13 (7.7%)	4 of 11 (36.4%)

6.6 Discussion

From four iterations of student surveying efforts, we were able to learn of student perceptions and preferences for expressing affects or other information via preset emoji-hashtag pairs on a discussion forum used in a large general biology course. Our main objectives were to (a) strengthen the unique association between the emojis and their assigned hashtags, and (b) adequately capture the breadth of affects that students wish to be represented. To this end, our survey focus was on soliciting student opinions on the interpretations of the emoji, including the fit and usefulness of the emoji and instructor-generated hashtag pairings in the original and modified NB emoji sets, as well as suggestions on missing emojis-hashtags to include in future emoji sets.

The survey results provide evidence that emojis vary in terms of how singular of an interpretation or meaning people will prescribe to it – some emojis were associated with fewer definitions and others had a wider net of terms associated with it. This is demonstrated particularly in Survey 2, where we displayed an emoji by itself first and solicited students' self-generated responses, resulting in a large variety of responses for some emoji (e.g., emojis for #idea, #question) and very few for others (e.g., emojis for #lost, #useful). We also saw this in the select-all-that-apply style of questions in the other surveys that resulted in some emoji-hashtags associated with multiple selections as opposed to a more singular interpretation. Overall, these results are consistent with previously published literature regarding variable emoji interpretation (Guntuku, 2019; Shurick, 2020; Tigwell 2016; Budnik, 2019; Liu, 2012), and provides a basis for revising the original emoji set, while including students in the revision process.

One strategy to mitigate the issue of multiple interpretations is to subcategorize broader emojis in an attempt to disambiguate the plural meanings and functions of the emoji-hashtags, thereby increasing the utility of the emoji and the discussion forum for both students and instructors. For example, in Survey 1, the original confusion-related emoji-hashtag were

#question, #curious, #help, and #confused. In Survey 2, we created an emoji-hashtag set designed to subcategorize confusion functions. This emoji-hashtag set incorporated suggestions from students in the survey, as well as the instructional team: #question, #just-curious, #flustered, #pleaseclarifyinclass, #willthisbeontheexam, and #requestTAattention. In Surveys 3 and 4, we made revisions once again, based on student input that there were too many emojis (and also had to work off of pandemic limitations that barred typical activities). This emoji set included #question, #just-curious, #lost, and #discuss. In the final version of emojis, the only change was #discuss to #lets-discuss.

From our surveys, we also learned that there are a range of preferences with regard to an ideal number of emojis to include within a set. While there were abundant suggestions of new emojis and hashtags, students also said in their feedback that too many emojis (e.g., 13 emoji-hashtags) would be confusing for them to keep track of. The earlier suggestion of displaying the emojis and hashtags side-by-side would mitigate this but having too many emoji-hashtags could still lead to overwhelm. However, the fact that many students may already be accustomed to navigating a far larger breadth of emojis on their mobile devices or other digital systems could indicate that students have the capacity to learn many symbols, develop preferences for the ones they use often, and learn to search for those they need when they need them (e.g., search bar in iMessage) (Ramaswamy, 2019). One possible solution is to present the emoji-hashtags in an alternative fashion (e.g., drop down list, or buttons), where the emoji and accompanying affect are side-by-side (i.e., no memorization necessary).

Finally, our initial emoji set, which was derived from instructor intuition during the initial phase of development, grew from 8 to 11 emoji-hashtag pairs. The additional emojis and/or hashtags that students offered in their feedback was particularly useful in inspiring the incorporation of many new emojis, some of which remain part of the current set, including: #important, #examable, #learning-goal, and #real-life-applications. Only one of the eleven pairs in the current working set has the same emoji-hashtag pairing as the original set (question mark

emoji and #question). Using the metric of average student DA with the instructor-generated emoji-hashtag pairing, and Light Kappa scores from the core survey questions presented in Surveys 1 and 4, the emoji set improved appreciably (Average DA: 74% in Survey 1 to 91% in Survey 4; Average Light Kappa scores: 0.41 in Survey 1 to 0.56 in Survey 4) (Table 6.4).

Future emoji-hashtag revisions might be centered around the emoji with the lowest DA and Kappa scores. In Survey 4, the lowest Kappa score was 0.36 (70% DA) for #just-curious, a modest increase from Survey 1, which had a Kappa score of 0.27 (71% DA). Other low Kappa scores from this emoji-hashtag set include #interesting-topic (71% DA, Kappa: 0.38) and #i-think (87% DA, Kappa: 0.44). Furthermore, a common theme in the student feedback was to restore former emoji that students were acclimated to using throughout the course term but was changed in the new set (e.g., #confused, #useful). While it is beneficial to gain the perspective of students who have used the emojis, it would also be informative to survey students who have not been exposed to emojis at all in order to eliminate the possibility of bias for old emojis.

Notably, the emoji set may continue to evolve over time. With fluctuations in usage and meaning of certain emojis/hashtags, it is important to consistently evaluate each emoji's function and relevance. Further, it is imperative to assess the student demographic and make sure that diverse cultural backgrounds are accounted for, and that different levels of digital experience are considered as well. The emoji should be as universal as possible, while also considering accessibility and inclusivity.

This study has implications for usage by instructors in applications such as filtering for heat-map density analyses or machine learning (Geller, 2021; Shusterman, 2021). For example, areas of text with high occurrences of #lost or #question may signal a need for text revision, with each emoji signaling a different degree of confusion. Additionally, usage of other emoji, such as the lightbulb-moment emoji can be analyzed in order to better understand the characteristics or features that result in them, which may be useful for intervention or text revision efforts. One

future tool based on the emoji tags could be a responsive system that automatically directs students to other student comments or to other course or external resources.

Instructors using emojis for analyses may benefit from intentionally discussing the functions and roles of the emoji in guiding teaching pace and strategies so that students can use the emojis accordingly (Dunlap, 2016). Future studies may measure how incorporating this style of student feedback influences students' learning and success in the course. Future efforts may also entail a study of the discrepancies between emoji usage among students from various backgrounds, allowing instructors to identify communication preferences and better address the needs of different student groups.

6.7 Funding

This work was supported by a grant by the National Science Foundation (#1915724).

6.8 Accessing Material

- A current version of Nota bene is freely available at: <https://nb2.csail.mit.edu>.
- The version of Nota bene used in this study is available at: <https://nb.mit.edu/>.
- A recent version of reading materials used in this study are available at:
[https://bio.libretexts.org/Courses/University_of_California_Davis/BIS_2A%3A_Introductory Biology - Molecules to Cell/SS1_2021 Bis2A Facciotti](https://bio.libretexts.org/Courses/University_of_California_Davis/BIS_2A%3A_Introductory_Biology_-_Molecules_to_Cell/SS1_2021/Bis2A_Facciotti).
- Survey data can be accessed at:
<https://datadryad.org/stash/share/qNOp1Wh-Ye4BYMA8VGWYOFZsF6ku70AR62dqz06wfA8>

6.9. References

- Budnik, E., Gaputina, V., & Boguslavskaya, V. (2019, October). Dynamic of hashtag functions development in new media: Hashtag as an identificational mark of digital communication in social networks. In Proceedings of the XI International Scientific Conference Communicative Strategies of the Information Society (pp. 1-5).
- Dunlap, J. C., Bose, D., Lowenthal, P. R., York, C. S., Atkinson, M., & Murtagh, J. (2016). What sunshine is to flowers: A literature review on the use of emoticons to support online learning. *Emotions, technology, design, and learning*, 163-182.
- Emerson, P. (2013). The original Borda count and partial voting. *Social Choice and Welfare*, 40(2), 353-358.
- Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). Package 'irr'. Various coefficients of interrater reliability and agreement, 22.
- Geller, S. A., Hoernle, N., Gal, K., Segal, A., Kim, H. G., Karger, D., ... & Igo, M. (2021). New Methods for Confusion Detection in Course Forums: Student, Teacher and Machine. *IEEE Transactions on Learning Technologies*.
- Guntuku, S. C., Li, M., Tay, L., & Ungar, L. H. (2019, July). Studying cultural differences in emoji usage across the east and the west. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 13, pp. 226-235).
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1), 23.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.
- Liu, K. L., Li, W. J., & Guo, M. (2012, July). Emoticon smoothed language models for twitter sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 26, No. 1).
- Ramaswamy, S., Mathews, R., Rao, K., & Beaufays, F. (2019). Federated learning for emoji prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329*.
- Shurick, A. A., & Daniel, J. (2020). What's behind those smiling eyes: Examining emoji sentiment across vendors.
- Shusterman, E., Kim, H. G., Facciotti, M., Igo, M., Sripathi, K., Karger, D., ... & Gal, K. (2021, April). Seeding Course Forums using the Teacher-in-the-Loop. In LAK21: 11th International Learning Analytics and Knowledge Conference (pp. 22-31).
- Tigwell, G. W., & Flatla, D. R. (2016, September). Oh that's what you meant! reducing emoji misunderstanding. In Proceedings of the 18th international conference on human-computer interaction with mobile devices and services adjunct (pp. 859-866).
- Zhang, A. X., Igo, M., Facciotti, M., & Karger, D. (2017, April). Using student annotated hashtags and emojis to collect nuanced affective states. In Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale (pp. 319-322).
- Zyto, S., Karger, D., Ackerman, M., & Mahajan, S. (2012, May). Successful classroom deployment of a social document annotation system. In Proceedings of the sigchi conference on human factors in computing systems (pp. 1883-1892).

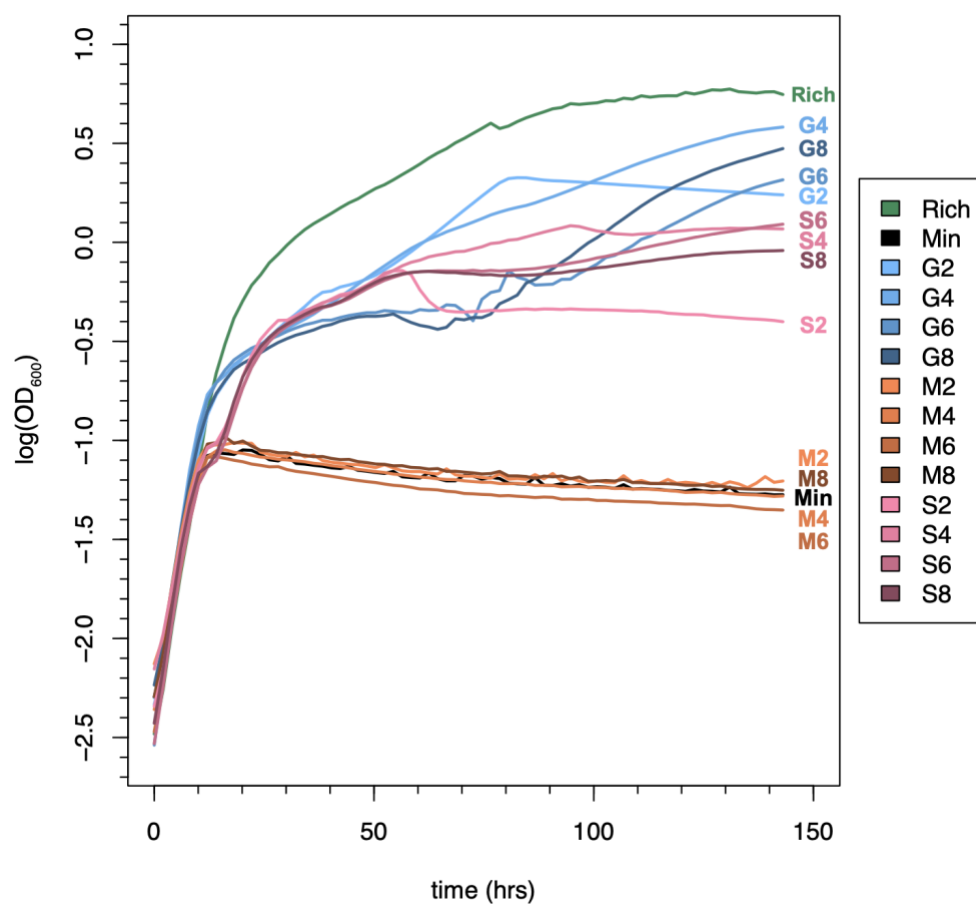
Table Of Contents: Appendix Items In Chapter 2

Appendix 2.1. <i>H. borinquense</i> optical density at 0.2%, 0.4%, 0.6%, and 0.8% (w/v) glucose, maltose, and sucrose, alongside rich and minimal media conditions.	213
Appendix 2.2. RNA-Seq hybridization oligo sequences for 5S, 16S, and 23S subunits.	214
Appendix 2.3. RNA-Seq samples sequencing library quality control assessment using the Agilent 2100 Bioanalyzer.	217
Appendix 2.4. Growth assay timepoints/strategy and OD ₆₀₀	219
Appendix 2.5. Segmented absorbance measurements and sugar utilization rates.	220
Appendix 2.6. Proportion of significantly up- and down-regulated genes in the RNA-Seq dataset across growth condition comparisons ($p < 0.05$).	223
Appendix 2.7. <i>H. borinquense</i> genes identified as non-carbohydrate-related ABC-type transport system proteins with their annotations and RNA-Seq log ₂ fold change values normalized to the rich media condition.	224
Appendix 2.8. <i>H. borinquense</i> genes identified as TCA Cycle proteins with their annotations, associated reactions, and RNA-Seq log ₂ fold change values normalized to the rich media condition.	226
Appendix 2.9. <i>H. borinquense</i> genes identified as amino acid biosynthesis proteins with their associated annotations, modules, reactions, and RNA-Seq log ₂ fold change values normalized to the rich media condition.	227
Appendix 2.10. <i>H. borinquense</i> genes for functional clusters with their annotations and RNA-Seq log ₂ fold change values normalized to the rich media condition.	228
Appendix 2.11. List of genes in RNA-Seq heatmap clusters (Figure 3.3).	231

Table Of Contents: Appendix Items In Chapter 6

Appendix 6.1. Iterative NB emoji-hashtag set development.....	254
Appendix 6.2. Copy of Survey 1 – Winter Quarter 2020 (W20).	255
Appendix 6.3. Copy of Survey 2 – Spring Quarter 2020 (S20).	259
Appendix 6.4. Copy of Survey 3 – Summer Session 1 2020 (SS20).	263
Appendix 6.5. Copy of Survey 4 – Winter Quarter 2021 (W21).	267
Appendix 6.6. Light Kappa score simulations for select-all-that-apply questions in Surveys 1, 3, and 4.	270
Appendix 6.7. Distribution of the number of terms selected (singular vs. multiple responses) for each emoji in the select-all-that-apply question of Survey 1.	271
Appendix 6.8. Emoji utility for the original emoji in Survey 1.	272
Appendix 6.9. Students' experience of confusion and emoji-hashtags deemed appropriate by students for various instances of confusion in Survey 1.	273
Appendix 6.10. The extent to which decisions for selecting tags to include posts are informed by emoji picture or hashtag text in Surveys 1 and 3.....	274
Appendix 6.11. Rationale for emoji-hashtag modifications throughout survey iterations.	275
Appendix 6.12. Student-generated terms with over 100 counts for emojis in Survey 2.....	279
Appendix 6.13. Student responses for questions pertaining to the appropriateness of emoji-hashtag pairing fit, desire to express the tag, and predicted frequency of expression of each emoji in Survey 2.	280
Appendix 6.14. Student perceived usage frequency of their current (the original) emoji set in Survey 2.	281
Appendix 6.15. Select-all-that-apply question data from Survey 3.	282
Appendix 6.16. Distribution of the number of terms selected (singular vs. multiple responses) for each emoji in the select-all-that-apply question of Survey 3.	283
Appendix 6.17. Distribution of student responses for questions pertaining to the appropriateness of emoji-hashtag pairing fit, desire to express the tag, and predicted frequency of expression for each emoji in Survey 3.....	284
Appendix 6.18. Distribution of the number of terms selected (singular vs. multiple responses) for each emoji in the select-all-that-apply question of Survey 4.	285
Appendix 6.19. Appropriateness of the emoji-hashtag pairings in Survey 4.....	286
Appendix 6.20. Student perceived usage frequency of their current (the original) emoji set in Survey 4.	287
Appendix 6.21. Distribution of student responses regarding the quantity of emoji in Survey 4.	288

Appendix 2.1. *H. borinquense* optical density at 0.2%, 0.4%, 0.6%, and 0.8% (w/v) glucose, maltose, and sucrose, alongside rich and minimal media conditions.



Appendix 2.2. RNA-Seq hybridization oligo sequences for 5S, 16S, and 23S subunits.

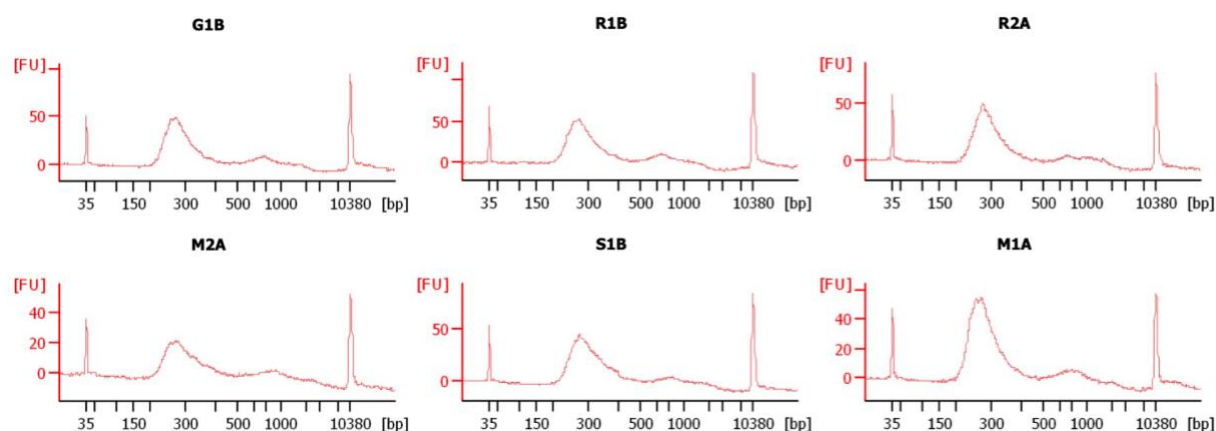
Sequence Name	Sequence	Tm (°C)
Hal_16S_01_dual	GGC TAA ATC GGA CTC CGA TAG CAA TGG CCT CCG GCA GGA TCA ACC GGA AT	72
Hal_16S_02_bor	CAC GTG TTA CTG AGC TAT ATG CTA CGA GTG TGA ACT CGT GCA ACT AGC AT	67
Hal_16S_02_sal	CAC GTG TTA CTG AGC TTT CCG CTG CGG GTC TAA ACC CGC ACA ACT AGC AT	71
Halo_16S_03_bor	GGA TTA GCC TCA GTT TCC CGA GGT TGT CCG TGT CTG CAG GGT AGC TTG AC	71
Halo_16S_03_sal	GGA TTA GCC TCA GTT TCC CGA GAG TAT TCC CGT CCA CAG GGT AGC TTG GC	71
Halo_16S_04_bor	GCG CCG GAG CGT TTG TGG ATC CCG GCA TTC CAG CGT GGG ATC GGT ATC GC	75
Halo_16S_04_sal	GCG CCG GAG CGT TTC CGG CTT TGC CCC TTC CAG GAG CAA AGC GTT ATC GG	75
Halo_16S_05_bor	GTG GGC CGT TAC CCC ACC GTC TAC CTA ATC GGC CGC AGC CGC ATC CTG CA	76
Halo_16S_05_sal	GTG GGC CGT TAC CCC ACC GTC TAC CTA ATC GAC CGC AGC CGC ATC CTG TG	74
Halo_16S_06_dual	ATT CCG TCT CCG GGC TCT TGC TCT CAC AAC CCG TAC CGA TTA TGG GCA CGG	73
Halo_16S_07_dual	TAA AGG TTT CGC GCC TGC TGC CCG TAG GGC CCG GAA TCT TGT CTC AG	74
Halo_16S_08_bor	GAC TAT ATG CCC TCG CAC TCG GAG TCC CCT TAT CGC ACT TGC GTG CAG TG	72
Halo_16S_08_sal	GGC TCT ATG CCT TCA CAC TCG GAG TCC CCT TAT CGC ACT TTC GTA CAG TG	70
Halo_16S_09_bor	GTC TTG CCC AGC TCT TAT TCG AGA ACC GCC TTA CGG TTC TCA AAA GCG AG	69
Halo_16S_09_bor	AAG ATT GGC CAT CAC TTG AGC TGC CGG TAT TAC CGC GGC GGC TGG CAC CG	75
Halo_16S_09_sal	GTC TTG CCC AGT CCT TAT TCG TGC ACC ACC TTA CGG TGT ACA AAA GTG AA	69
Halo_16S_09_sal	AAG ATC GGC CAT CAC TCG GAC TGC CGG TAT TAC CGC GGC GGC TGG CAC CG	76
Halo_16S_10_bor	GGA TTT CCC GAC GGA CTT TCG TGG CCA GCT ACG GAC GCT TTA GGC CCA AT	73
Halo_16S_10_sal	AGA TTT CCC AAC GGA CTT GTT CAG CCA GCT ACG GAC GCT TTA GGC CCA AT	71
Halo_16S_11_bor	CCG GTC CCA AGC CAC GAA GTT TCC GCC GGA CGC CCA TCT GTT GGG CAG AT	75
Halo_16S_11_sal	CCG GTC CCA AGC TGA ACA GTT TCC GCT GGA CGC CTG CCC GTT AAG CGG AC	75
Halo_16S_12_bor	TCC AGG ATT ACG GGA TTT CAC TCC TAC CCC GGA CGT ACC CCT CGA GCC TT	72
Halo_16S_12_sal	TCC AGG ATT ACA GGA TTT CAC TCC TAC CCC AGA CGT ACC CCT CAG GTC TT	70
Halo_16S_13_bor	CCT CAC TGT CGG ATC CGT CTT CTC GAG GTG CTT TCG CCA TCG GCG GTC CG	74
Halo_16S_13_sal	CCT CAC TGT CGG ATC CGT TCT CCT GAG GCG CTT TCG CCA CCG GCG GTC CG	76
Halo_16S_14_dual	AGC TAG GAC TAC CCG GGT ATC TAA TCC GGT TCG AGA CCC TAG CTT TCG TC	70
Halo_16S_15_bor	GGC ACA GCA CAG GCT CGT AGC CTG TGC CAC ACC TAG TGA ACA TTG TTT AC	71
Halo_16S_15_sal	GGC ACA GCG CAG GCT CGT AGC CTG CGC CAC ACC TAG CGG ACA TCG TTT AC	74
Halo_16S_16_bor	TCA TCC TTG CGG ACG TAC TTC CCA GGC GGT TCA CTT CTC GGC TTC CCT AC	72
Halo_16S_16_sal	TCA TCC TTG CAG ACG TAC TTC CCA GGC GGT CCG CTT CTC GGC TTC CCT AC	73
Halo_16S_17_dual	AAC CGC AGG CTC CTC CGG TTG TAG TGC TCC CCC GCC AAT TCC TTT AAG TT	73
Halo_16S_18_bor	GTC ATT ACT GTA GTC GGA GCT GGT GAG ATG TCC GGC GTT GAG TCC AAT TA	69
Halo_16S_18_sal	GTC ATT ACT ACT GTC GGG GCT GGT GAG ATG TCC GGC GTT GAG TCC AAT TA	70
Halo_16S_19_bor	CGG CCA TGC ACC TCC TCT CTA CAG CTC GTG ATA AGG TCA TCA ACC TGA TC	70
Halo_16S_19_sal	CGG CCA TGC ACC TCC TCT CAG TAG CCT CGG GTA AGG TCA TCA ACC TGA CC	72
Halo_16S_20_dual	TCT CGC TCG TTG CCT GAC TTA ACA GGC CTC ACG GTC CAG GCT GAC GG	73
Halo_16S_21_bor	CTA ATG TAC CCA GCT ACC CAA AGG GTA CTG CTG GCA ATT AGG AGT GCG GG	70
Halo_16S_21_sal	CTA ATG TAC CCA GCT ACC CAA AGG GTA CTG CTG GCA ATT AGG AGT GCG GG	70
Halo_16S_22_bor	CTG ACC TAC CGT TGC CCA TTC CTT CCT CCA GTT TGG CAC TGG CAG TCC AT	72
Halo_16S_22_sal	CTG ACC TAC CGT TGC CCG TTC CTT CCT CCG CTT TGG CAG CGG CAG TCC AC	74
Halo_16S_23_bor	ATT GTT CTG GCC ATT GTA GCC CGC GTG TAG CCC AGC TCA TTC GGG GCA TA	73
Halo_16S_23_sal	ATT GTC TCG ACC ATT GTA GCC CGC GTG TTG CCC AGC CCA TTC GGG GCA TA	74
Halo_16S_24_bor	AAC TAC GAC CAG ATT TAG GAG ATT AGC GGT CTC TTT CGA GAT TGC AAC CC	67
Halo_16S_24_sal	AAC TAC GAT CGA GTT TAG GAG ATT AGC GCC TCC TCT CGG AGT GGC TTC CC	70
Halo_16S_25_dual	GAT TAC TAC CGA ATC CAG CTT CAT GAG GGC GAG TTT CAG CCC TCA ATC CG	69
Halo_16S_26_bor	GCG GTG TGT GCA AGG AGC AGG GAC GTA TTC ACC GCA CTC TTG TGA AAT GC	72
Halo_16S_26_sal	GCG GTG TGT GCA AGG AGC AGG GAC GTA TTC ACC GCG CGC TGC TGA CAC GC	75
Halo_16S_27_bor	CGA CTA TCT TGC GAT AGC CTC ATC CGG ACC TCA CTC GGG TGC TTT GAC GG	71
Halo_16S_27_sal	TGA CCA GCG CAT GCC GGC CTC ATC CGA ACC CCA CTC GGG TGA TTT GAC GG	75
Halo_16S_28_bor	CCC CTA CGG CTA CCT TGT TAC GAC TTA AGC CCC CTT GCG AAG CCC AGA TT	72
Halo_16S_28_sal	CCC CTA CGG CTA CCT TGT TAC GAC TTA ATC CCC CTT GCG GAG CCC AGA TT	71
Halo_16S_29_dual	CCG ATC GTT AGG AGG TGA TCC AGC CGC AGA TT	67
Halo_23S_01_bor	CAG CTT GGC ACG TCC TTC GTC GGC TCT CGA GCC GAG CCA TTC ACC AG	74
Halo_23S_01_sal	CAG CTT GGC ACG TCC TTC GTC GGC ATC CGA GCC GAG CTA TCC ACC AGG TG	74
Halo_23S_02_bor	TTC GGA GAT CCT GAG TTC TTA GCC TCC GTG CGG CTC CCT CAG GCT TTT CG	72
Halo_23S_02_sal	TTA GGA GAT CCT GAG TTC TTA GCC TCC GTG CGG CTC CCT CAG GCT TAT CG	71
Halo_23S_03_bor	AAT TCG GAA CGT TCC CTA TTG CGC GTT GGC AAT TGT TGT AGG GAT TCC CA	70
Halo_23S_03_sal	AAT TCC GGC CGT TCC CCA TTG CGC AAG GCA ATT GTT ATA GGG ATT CCC A	71
Halo_23S_04_bor	CTA ACG ACA TCG CGG ATT GCT TTT TGT TCC TGT TCC TAC TGA GAT GTT TC	67
Halo_23S_04_sal	CTA ACG GCG TCT CAT TCG ATT TCT CTT CCT GGC CGT ACT GAG ATG CTT C	68

Halo_23S_05_bor	ACA TTT CCT TGC GGA TTC GGT TTG GAC TGT GTC GCG TTT ACT CGC CGT TA	70
Halo_23S_05_sal	ACA TTG CCC GAA GGC TTC GGT TTG GAC TGT GCC GCG TTC ATT CGC CAT TA	73
Halo_23S_06_bor	CCA GGA GAC TTC TCG TGT GTT GGT GAT CCG TTG TAG ATT GTC CGT ACA CC	69
Halo_23S_06_sal	TCA GGA GAC TTC ACA CGA ACG GTC AAA CGA TGA AAG TCA GTC CGA ACA CC	69
Halo_23S_07_bor	GGA TCG CGT TCA CAC GAT ACG GGG TTG TCA CCC TGT ATC ACG CTC CGT T	72
Halo_23S_07_sal	CGT CGT ACT GGT CCG TGA TAC GGG GCT GTC ACC CTG TAT CGC GCT CCG TT	73
Halo_23S_08_bor	TGC GAG ATT CAC GAG GGA TAT CCA ACC CCC GCT ACT CTG GAG CAA TCT CG	71
Halo_23S_08_sal	TGC CAC ATT CAC GAG GGA TTT CCA ACC CCC GCT ACT CTG GAG CTG ACG CA	73
Halo_23S_09_bor	ACT TGT TCG CTA TCG GTC TCG AGG AGT GTT TAG TCT TCG CAG TCG ATG CC	70
Halo_23S_09_sal	ACT TGT TCA CTA TCG GTC TCG GAG AGT ACT TAG TCT TGG CAG TCG ATG CC	68
Halo_23S_10_bor	GCC CTA TTG CAC CGC CCT TTT CGG GGT GCT TTT CAG CGT TCG CTC GCG CT	75
Halo_23S_10_sal	GCC CTA TTT CAC CCC CCT TTG TGG GGT GCT TTT CAG CGT TCG TTC ACA CT	72
Halo_23S_11_bor	TTT TCC GAG GAC CTT GTA TGC CCT GTC GCT CTA TCA CTG ACT GAT TTC AG	68
Halo_23S_11_sal	TCC CAG AGA GGC CTT TTA TGC CCC GTC GCT CCA TCG CCA CCT GAT TTC AA	73
Halo_23S_12_bor	CAG AAC ACC GGC TTC CTC GGA TTC CTA CTA GTG GGT CGC CCC ACG GTC GT	74
Halo_23S_12_sal	CGG AAC ATC GGC TCC TCT CAG GTC CTA CCA TGA GTT TGC ACT CAA GTC GT	71
Halo_23S_13_bor	TAG ACT CGC CAA GTA AGT GCA CTC CCT GGT TCG TTT TTC AAA ACG TAC GA	68
Halo_23S_13_sal	TAG ACT CGT CAA ACA GGC ACA CTC CCT GGC TCG TTT TTC AAA ACG TAC GA	69
Halo_23S_14_bor	GCA ATG CTG CGG CCT TGT CGG TTT CCC TGC GCC TTC CCG GAT GAT CCG GT	76
Halo_23S_14_sal	GCA ATG CTG CGG CCA TAT TGG TTT CCC TAC GCC TTC CCG GAT ACT CCG GT	73
Halo_23S_15_bor	GTT TCG GGT CGC ACC CAC TTG ACT CCC CGC GCT TGA ACA CGG CGG CCC TC	76
Halo_23S_15_sal	GTT TCG GGT CGT GCC CGT TTG ACT CCC CGC GCT TGA ACA CGG TGG CCC TC	75
Halo_23S_16_bor	CTC CAC GTG GCT TTC GCC ACG CTT CGT CTT GCC CTC GTG TAG ATC ATC TG	72
Halo_23S_16_sal	CTC CAC ATG GCT TTC GCC ATG CTT CAC CTT GCC CAC GCG TAG ATC ACC GG	73
Halo_23S_17_bor	CCT ACA CAC GGG TCA CGG GAG GGT ATT GTA AGA CAC CAA CCC TAA CAG AC	70
Halo_23S_17_sal	CCT ACA CGT AGG TCA CGG GAG GGT GTT GAA AGA CAC CAA CCC TAA CAG GC	71
Halo_23S_18_bor	ACA TGT TTC GGC TGG AAC CAG CTG TTG CCA GAT TCG ATG GGC CTT TCA CC	72
Halo_23S_18_sal	ACA TGT TTC GGT TGG AAC CAG CTG TTG CCC GGT TCG ATG GGC CTT TCA CC	73
Halo_23S_19_bor	CCA ATC GGT CGC TCT ACC TCG CAG ACT ATC TCG GCT GAG GTC ATG CTT CG	71
Halo_23S_19_sal	CCA ATC GGT CGC TCT ACC CCA CGA ACT ACC TCG GCA GAG GTC ATG CTT CG	72
Halo_23S_20_bor	GCA AGT TTG GAG TTC GAC AGG AAG GCC GAC TCC TCT CGG AGG CGG GTC TT	73
Halo_23S_20_sal	GTA GGT TTG GAG TTT GAC AGG GGA GAC AAC TCC TCT CGG AGT TGA ACT CC	69
Halo_23S_21_bor	CGG ACC ACA AGC TTA CCC CGC AGT CCG GAT TCC CAG CGT CTA CGA CGT CT	74
Halo_23S_21_sal	CGG TGC ACA GGC TTA CCC CGC ACA CCG GAT TCC CTG CGT CGA CGG CGT CC	77
Halo_23S_22_bor	ACT TAA TCC ACA CTT GGG GAC CTT AAC CCG GCT CTG GGT TGT CTC CCT CA	71
Halo_23S_22_sal	ACT TAG CTC GCA CTT TGG GAG CTT AAC CTA ACT CTG GGT TGT CTC CCT CA	70
Halo_23S_23_bor	TTC TAA GCT CAC CTC CCG GCT GTT TAG GGC TCG AGA CAA TCT TCG ATC GC	70
Halo_23S_23_sal	TTC TAA GCT CAC CTC CCG GCT GTC TAG GGC TCG AGA CCA CCT TCA ATC GC	72
Halo_23S_24_dual	CGG CCG GTA AGC TGT TAC GCT TTT CTT AGA GGG TAG CTG CTT CTA AGC TC	69
Halo_23S_25_bor	AGG TCT CGG TGG TGG ATT TGA GCC CCG ATC ATT TTC GAC GCC TCG AAC CT	72
Halo_23S_25_sal	AGG TCT CGG TGG CGA ACT TGA GCC CCG ATC ATT TTG GGC GCC TCG AAC CT	74
Halo_23S_26_bor	CGC ACA GTA TGC CAG CCT ACT AGG TCA GGC GTA TGG TGT CGC CTG CTG CC	73
Halo_23S_26_sal	CGA ACA GAA TGC CAA CCT ACT GGA TCA CGG TGT TAC CCG TGC CGT T	71
Halo_23S_27_bor	GGA TTT TCG TTA CTG AGC GGT CCA CGC GAG CTC TCG CCC GAG CTT CCA CC	73
Halo_23S_27_sal	GGA TTT TCT TTT CCA CTC GGT CCA CTC GAG TTC TCA CCC GAG CTT CCA CC	70
Halo_23S_28_bor	CTC GTA AGG TCG TCG GGG TTC GCA CCC GAC TCT CGC TGC TAC TGT GAC CA	73
Halo_23S_28_sal	CTC TTT CGG CCA TCG GGA TTC TTA CCC GAC TAA CGC TGC TAC TAT GGC CA	71
Halo_23S_29_bor	ACT TAG GAT CGA CTA ACC CTC GGC TGA TTG GCA GTG CCG AGG AAC CCT TA	71
Halo_23S_29_sal	CCT TAG GAT CGG CTA ACC CTC GGC TGA TAA GCA TTG CCG AGG AAC CCT TG	71
Halo_23S_30_bor	TAC TGG AAT ATT AAC CAG TTT CCC GTT TGT CTC ATT CGA GTT ACG GTG AG	66
Halo_23S_30_sal	CAC TGG AAT ATT AAC CAG TTT CCC TTT TGA CCC GCT CGA ATT ACG ACG GG	68
Halo_23S_31_bor	GAA GGC CCG GCT TGA AGT TTC CCA AAG CGT CGG CTT TGA GTG GTG ATG G	72
Halo_23S_31_sal	GAA TGC CCG GCT CAA GCT GCT CCG AGG CGT CGA CTT TCA ATG TAC GGT GG	73
Halo_23S_32_bor	TTT GCT TCG TGC CGT TAC GGC TTC CAC GAG CTT CCG TGA TTC GAC CGG GC	74
Halo_23S_32_sal	TTT GCT TCC TGC CAT TAC GGC TTC CAC GAA CTT CGA CGG TTC GAC CGG GC	73
Halo_23S_33_bor	GCC TTT TCA CGG GCT CTA GGT ACC TTC AAC TTG CGC TGT CTC ACG CTT CG	71
Halo_23S_33_sal	GCC TTT TCA CGG GCC CCA GAT TGA CTT GAG TTT CCC TGT TCC GAC GTT CG	71
Halo_23S_34_bor	CTT TCT CCG CCA GAG TAC CTG TGT CGG ATC TCG GTA CGA TCA TCA CGC TT	70
Halo_23S_34_sal	CTT CCT CTG CCA GAG CAC CTG TGT CGG ATC TCG GTA CGA ACA CCG TAC TC	71
Halo_23S_35_bor	GGA CTA ACT TGC CGA ATT CCC TAA CGT CGG TTG ATC CCG ACA GAC CTT GG	70
Halo_23S_35_sal	GGA CTA ACT TGC CGA ATT CCC TAA CGT CGG TTA TTC CCG ACA GGC CTT GG	70
Halo_23S_36_bor	ACT GCG ACC TGC TCG GTC AGG AGC AGG CAT CCC TTA TCG CGA ACT TAC G	73
Halo_23S_36_sal	ACT GCG ACC TGC CTC ATT GCG AGG CAG GCA TCC CTT ATC GCG AAC TTA CG	73
Halo_23S_37_bor	CGG ATT TGC GGT CAC CTA TGT TGT TAC TAG ACA GTT GGA GCG CCC GAG TC	70

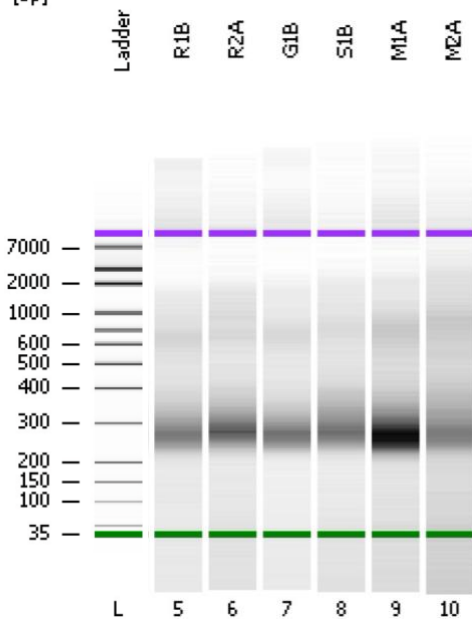
Halo_23S_37_sal	CGG ATT TGC GGT CAC CTA TGT TGT TAT TAG ACA GTC GGA GCC TCC GAG TC	69
Halo_23S_38_bor	GTG TTC AGA TAC CTG CAC TGG GCA GGA TTC AGT GAC CGT ACG AGT CCT TG	70
Halo_23S_38_sal	GTG TTC AGA TAC CCG CAC TGG GCA GGA TTC AGT GAC CGT ACG CGT CCT TG	72
Halo_23S_39_bor	GGG TCA TAG TTA CCC CCG CCG TTG ACA GGT CCT TCG TCC TCT TGT ACG AG	71
Halo_23S_39_sal	GGG TCA TAG TTA CCC CCG CCG TTA ACG GGT CCT TCG CCC CAT TGT ACT GG	72
Halo_23S_40_dual	CAT TCA TGC AAG CCG CTA CTG AAG CGG CAA GGT ACT ACG CTA CCT TAA GA	69
Halo_23S_41_bor	ACG TAC AGT TCA CCG GGC CCA ACG TTG GGA CAG TGG CGC TCT CAT TGA TC	73
Halo_23S_41_sal	ACG TAC AGT TCA CCG GGC CCA ACG TTG GGA CAG TGA GGC TCT CGT TGA TC	72
Halo_23S_42_dual	CTC TAT AGG GTC TTC GCT TCC CCT TGG GGG TCT CCA GAC TCC GCA CTG GA	72
Halo_23S_43_bor	ACC TAT GCT GCA CAT CAG CGA CCA CGT CCC AGC GAC AGC CTG CAG TAA AG	73
Halo_23S_43_sal	ACC TAC TCT GCA CAT CAG CGA CCG TGT CCC AGC GAC AGC CTG CAG TAA AG	72
Halo_23S_44_bor	CAA TGA TGA CTC GGT GGT CCG CTA GCG CGG ATA CCT GTG TAG TGT CTC CT	71
Halo_23S_44_sal	CAT GTG TGA CTC GGT GGC GTG CTA GCG CAC GTA CCT GTG TAA CGT CTC CT	72
Halo_23S_45_bor	CGG TGT CCT CCT CCC GGA GTG AGG GTC GCA GTC ACT GAC GGG TAG TAT TT	72
Halo_23S_45_sal	CGG TGT CCT CCT CCC GGA GTG AGG GTC ACA GTC ACT GAC GGG TAG TGT TT	73
Halo_23S_46_dual	GCT CGA TAT CTT TTC AAG CGC GTA CCG CCC CAG TCA AAC TGC CCG GCT AC	72
Halo_23S_47_bor	CTT GCG CTC TTC CAC GAG TTT CCG ACT CGC GTG AGG AAA TCT TTG GGC GC	72
Halo_23S_47_sal	CTT GCA CTC TTC GCC GAG TCT CTG ACT CGG CTG AGA TAG GCT TAG GGC GC	72
Halo_23S_48_bor	TCG CGT CAG CGT CAC TCG TTG GGA ATG ACA CTG TCA GAC TTC CGT GTG CT	72
Halo_23S_48_sal	TCG CGT CAG CGT CCC TCG TTG TGT AGG ACA CTG TCA GCC TAT CTT ATG CT	71
Halo_23S_49_bor	TCA TCT ACG GGT CCC ATG AAT GAA CCT CGT AGG TTC GCT AGA CCA CGC TT	70
Halo_23S_49_sal	TCA GCA ATT GGC CGC ATC AAG CAG CCT AAT TGG TTC GCT AGA CCA CGC TT	72
Halo_23S_50_bor	TAT GTG CTC TTG CCG GTG ACG ACT CTG TTA TCC CTA AGG TAG CTT TTC TG	68
Halo_23S_50_sal	TAT GTG CTC TTG CGA GTG ACG ACT CTG TTA TCC CTA AGG TAG CTT TTC TG	67
Halo_23S_51_bor	ACA GGC AGG ATG GAG GGA ACC GAC ATC GAG GTA GCA AGC CAC CGG GTC GA	75
Halo_23S_51_sal	ACG GGC AGG ATG GAG GGA ACC GAC ATC GAG GTA GCA AGC CAC TCG GTC GA	74
Halo_23S_52_dual	CTC ACG ACC TCC TTT AAT AGG CGA ACA ACC TCA CCC TTG CCC GCT TCT GC	71
Halo_23S_53_dual	ACA CCC CCA ATA GAT AGC AGC CGA CCT GTC TCA CGA CGG TCT AAA CCC AG	71
Halo_23S_54_bor	CCA ATC GTA GTT CCT CTC GTA CTA TAC GAT CGT TCC CGT CAA GTA CCG TA	67
Halo_23S_54_sal	CCG TTC GTA GTT CCT CTC GTA CTA TAC GAA CGT TCC TGT CAA GCA CCA AA	68
Halo_23S_55_bor	CGT AGC TAC TCG GCA TCT GCC CTT TCG GAC AAC CGA TAC ACC AGC GGC GA	73
Halo_23S_55_sal	CGT AGC TGC TCG GCA CAT GCC CTC TCG GAC AAC CGA TAC ACC AGT GGC AC	73
Halo_23S_56_bor	TTT CCA GGT GGG TTT CGA GCT TAG ATG CGT TCA GCT CTT ACC CCG TGT CG	71
Halo_23S_56_sal	TTT CCA GGT GGG TTT CGA GCT TAG ATG CAT TCA GCT CTT ACC CCG TGG TG	70
Halo_23S_57_bor	TCC CAA GTC TAT CGA ACT TGT CTT CTA CAA GTG TCC TCG GTG GTA CTT CT	67
Halo_23S_57_sal	CCC CAA GTC TAT CGA ACT CGT CTT CTA CGA GCG GTC TCA GTG GTG CTT CT	70
Halo_23S_58_sal	TGA CCT GTT AGT ACT CGC GGG CTA AAC GTC TCG TTG CCT CGA CGC GTA CA	72
Halo_5S_01_bor	TCC GTG TTC GGG ATG GGT ACG GGT GTT GCC TCA CCG CTG TGG CCG CCT TA	76
Halo_5S_01_sal	TCC GTG TTC GGG ATG GGT ACG GGA GTA ACC CCA CCG CTA TGG CCG CCT TA	74
Halo_5S_02_bor	CCA GAG GTT CGC ACA CTC CAG TAC TTG TTA TCA CGC TGG CGA GCT TAA CT	70
Halo_5S_02_sal	CCA GAG GCT CGC GCA CTC CAG TAC TGA CCG GAA CGC AGG CGG GCT TAT CT	75
Halo_5S_03_bor	AGG AGG CGG CGA GAT AAT GTT T	59
Halo_5S_03_sal	AGT AGG CGG CGA ACC GGA TTT C	62

Appendix 2.3. RNA-Seq samples sequencing library quality control assessment using the Agilent 2100 Bioanalyzer.

Select representative samples:

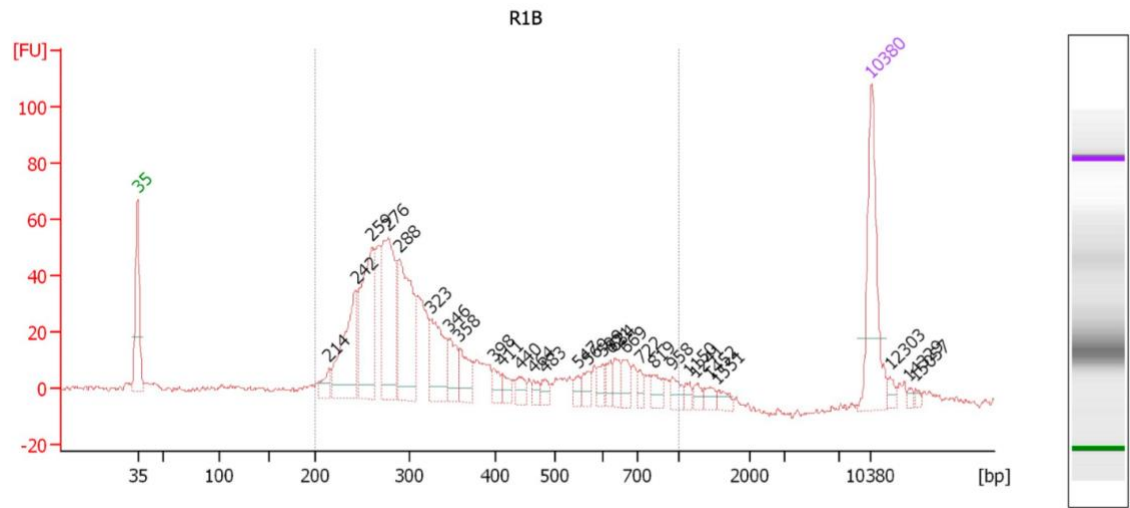


[bp]



Example of electropherogram summary and peak table for one sample (R1B):

Electropherogram Summary Continued ...



Overall Results for sample 5 : R1B

Number of peaks found: 29 Corr. Area 1: 814.1
Noise: 0.9

Peak table for sample 5 : R1B

Peak	Size [bp]	Conc. [pg/μl]	Molarity [pmol/l]	Observations	Aligned Migration Time [s]
1	35	125.00	5,411.3	Lower Marker	43.00
2	214	16.44	116.3		61.27
3	242	104.16	650.9		63.82
4	259	137.07	801.1		65.34
5	276	137.19	752.9		66.85
6	288	128.54	675.9		67.94
7	323	63.62	298.0		70.92
8	346	29.81	130.5		72.77
9	358	29.18	123.6		73.72
10	398	12.39	47.2		76.98
11	411	11.90	43.9		77.79
12	440	12.10	41.7		79.44
13	464	7.28	23.8		80.82
14	483	9.46	29.7		81.90
15	547	8.44	23.4		85.03
16	569	12.53	33.4		86.02
17	599	10.61	26.8		87.39
18	621	10.66	26.0		88.10
19	634	12.99	31.0		88.53
20	669	15.82	35.8		89.67
21	722	7.98	16.7		90.94
22	819	12.79	23.7		92.22
23	958	11.49	18.2		94.07
24	1,150	6.00	7.9		95.63
25	1,241	7.16	8.7		96.25

Appendix 2.4. Growth assay timepoints/strategy and OD₆₀₀.

Time (hr)*	DF	OD ₆₀₀													
		Min-1	Min-2	G-1	G-2	SG-1	SG-2	S-1	S-2	SaddG -1	SaddG -2	M-1	M-2	MaddG -1	Madd G-2
0.0	1	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
3.8	1	0.072	0.090	0.113	0.065	0.080	0.071	0.083	0.069	0.061	0.068	0.084	0.071	0.085	0.076
6.7	1	0.126	0.132	0.124	0.110	0.123	0.147	0.134	0.120	0.120	0.122	0.115	0.128	0.116	0.118
10.7	1	0.266	0.311	0.303	0.287	0.284	0.301	0.271	0.293	0.246	0.266	0.276	0.275	0.260	0.284
14.7	1	0.479	0.499	0.581	0.563	0.558	0.556	0.467	0.460	0.438	0.452	0.487	0.476	0.468	0.473
19.7	2	0.351	0.357	0.483	0.458	0.469	0.452	0.326	0.335	0.330	0.318	0.350	0.337	0.353	0.354
23.9	2	0.362	0.375	0.651	0.647	0.622	0.624	0.386	0.406	0.364	0.364	0.399	0.400	0.410	0.399
27.7	2	0.395	0.393	0.694	0.758	0.677	0.692	0.450	0.443	0.422	0.462	0.437	0.431	0.408	0.438
31.0	3	0.267	0.258	0.559	0.526	0.490	0.519	0.3549	0.373	0.361	0.352	0.272	0.289	0.269	0.266
34.7	3	0.248	0.249	0.580	0.594	0.548	0.566	0.4361	0.452	0.422	0.462	0.260	0.264	0.260	0.292
39.7	4	0.158	0.161	0.383	0.408	0.441	0.444	0.4078	0.386	0.365	0.377	0.185	0.179	0.157	0.164
43.7	4	0.157	0.153	0.431	0.448	0.453	0.462	0.457	0.447	0.424	0.430	0.183	0.168	0.163	0.167
Added exogenous glucose to SaddG and MaddG conditions at t=44 hours.															
47.2	4	0.172	0.180	0.527	0.482	0.569	0.598	0.4549	0.472	0.437	0.432	0.189	0.181	0.205	0.196
50.7	4	0.173	0.178	0.546	0.558	0.538	0.610	0.4801	0.499	0.427	0.385	0.183	0.193	0.231	0.228
55.7	4	0.149	0.154	0.554	0.552	0.729	0.700	0.5058	0.475	0.418	0.386	0.177	0.166	0.243	0.264
65.2	4	0.160	0.162	0.630	0.610	0.779	0.758	0.6086	0.564	0.446	0.454	0.220	0.218	0.384	0.390
71.8	4	0.146	0.133	0.607	0.619	0.716	0.694	0.6711	0.628	0.507	0.480	0.199	0.235	0.436	0.381
78.4	4	0.131	0.129	0.654	0.655	0.752	0.729	0.6537	0.650	0.612	0.621	0.232	0.245	0.490	0.479
99.3	4	0.065	0.063	0.589	0.594	0.636	0.662	0.820	0.781	0.668	0.644	0.207	0.193	0.551	0.571
107.7	4	0.130	0.1348	0.676	0.679	0.742	0.772	0.902	0.854	0.726	0.717	0.306	0.291	0.610	0.587
116.7	5	0.107	0.119	0.559	0.537	0.730	0.736	0.814	0.753	0.671	0.627	0.303	0.242	0.569	0.566
128.2	6	0.091	0.108	0.415	0.373	0.537	0.570	0.653	0.622	0.550	0.551	0.205	0.204	0.516	0.453
140.7	6	0.083	0.092	0.440	0.399	0.557	0.539	0.672	0.657	0.585	0.599	0.237	0.231	0.497	0.465
148.7	6	0.075	0.076	0.420	0.393	0.531	0.543	0.672	0.660	0.559	0.558	0.241	0.251	0.471	0.442
171.2	6	0.087	0.101	0.361	0.372	0.513	0.482	0.706	0.697	0.590	0.591	0.262	0.264	0.473	0.452
190.7	6	0.086	0.100	0.368	0.341	0.464	0.471	0.719	0.723	0.591	0.585	0.281	0.289	0.453	0.426

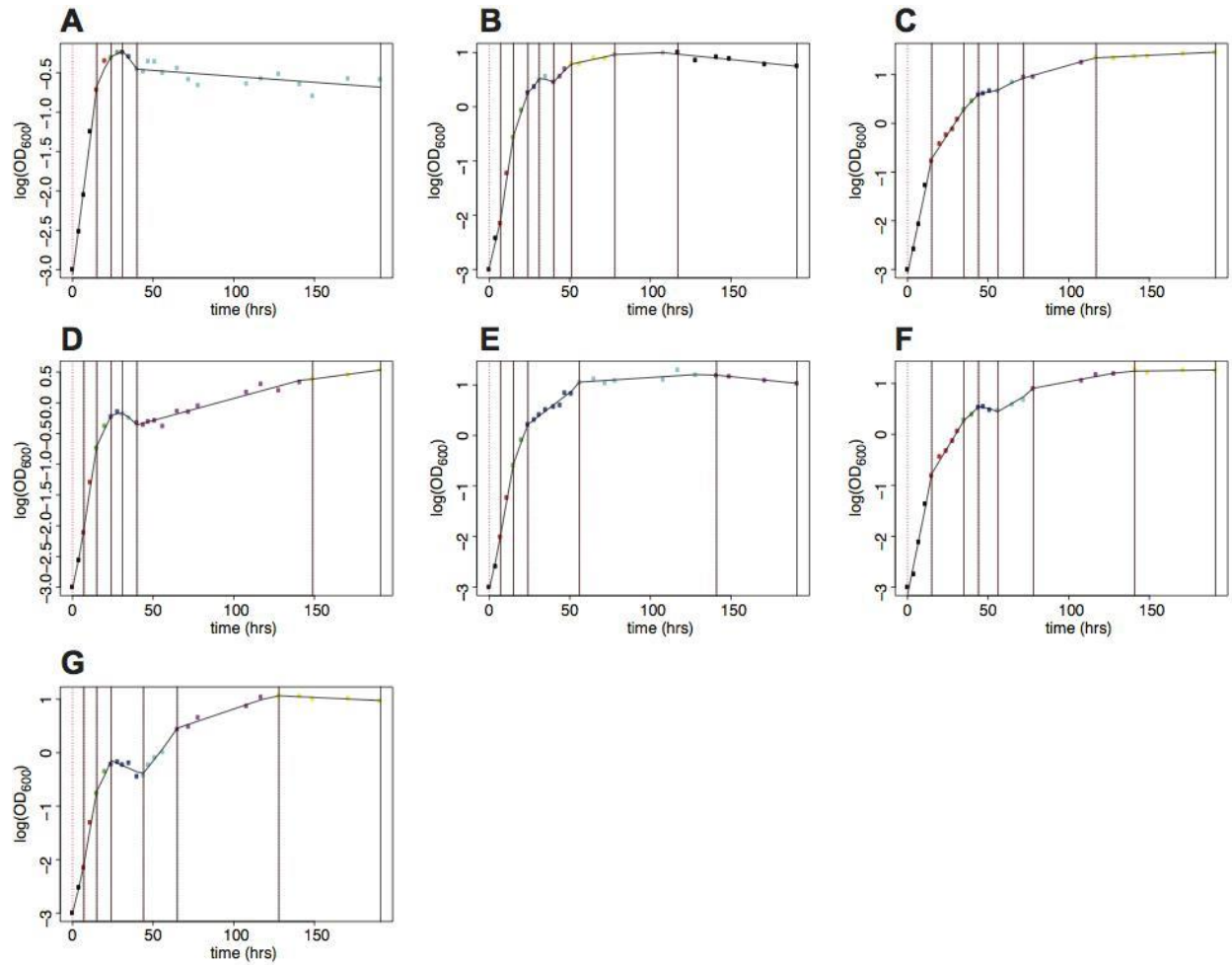
* Green-shaded timepoints represent samples used for sugar utilization assays; purple-shaded timepoints represent supernatant samples collected but not used for sugar utilization assays.

Appendix 2.5. Segmented absorbance measurements and sugar utilization rates.

Breakpoints and slopes (OD_{600}/hr) for *H. boryguense* in each manually segmented growth condition:

time (hr)	G	SG	S	SaddG	M	MaddG	Min
0							
3.8	0.12267733	0.1398			0.1259	0.1213	
6.7			0.1556	0.1551			0.1576
10.7	0.1984	0.1772			0.1721	0.1740	
14.7							
19.7	0.0914	0.0899			0.0569	0.0610	0.0467
23.9			0.0506	0.0523			
27.7	0.0321				0.0080		0.0100
31							
34.7	-0.0039				-0.0161	-0.0133	-0.0239
39.7		0.0244	0.0340	0.0278			
43.7							
47.2	0.0314						
50.7			0.0072	-0.0062		0.0382	
55.7							
65.2	0.00641844		0.0176	0.0184			
71.8							
78.4					0.0072		
99.3		0.0021	0.0094			0.0102	-0.0015126
107.7	0.0012			0.0059			
116.7							
128.2							
140.7							
148.7	-0.0030		0.0016				
171.2		-0.0032		0.0004		-0.0014	
190.7					0.0035		

Segmented absorbance measurements ($\log(OD_{600})$) of *H. borinquense* over time (duplicates averaged) for A) Minimal media, B) Glucose, C) Sucrose, D) Maltose, E) Sucrose with glucose simultaneously available at $t=0$ hrs (SG), F) Sucrose with glucose added at $t=44$ hrs (SaddG), and G) Maltose with glucose added at $t=44$ hrs (MaddG):



Segmented sugar utilization rates (mg/hr) by *H. borinquense* for each growth condition (triplicates averaged), and total sugar utilized calculated using those rates (g) (purple timepoints indicate sampling timepoints; dotted lines provide a visual reference for a sugar utilization transition point):

time (hr)	G		SG			S		SaddG			M		MaddG		
	Gluc util rate (mg/hr)	total sugar utilized (g)	Suc util rate (mg/hr)	Gluc util rate (mg/hr)	total sugar utilized (g)	Suc util rate (mg/hr)	total sugar utilized (g)	Suc util rate (mg/hr)	Gluc util rate (mg/hr)	total sugar utilized (g)	Malt util rate (mg/hr)	total sugar utilized (g)	Malt util rate (mg/hr)	Gluc util rate (mg/hr)	total sugar utilized (g)
0		0.000			0.000		0.000			0.000		0.000			0.000
3.8	-2.3	-0.009		-8.00E-02	0.000		-0.009			-0.021		-0.008			-0.005
6.7		-0.015			0.000		-0.015			-0.037		-0.015			-0.009
10.7		-0.025			0.000		-0.025			-0.059		-0.024			-0.015
14.7	-35.7	-0.167	-0.4		-0.180	-2.3	-0.034	-5.5		-0.081		-0.032			-0.021
19.7		-0.346			-0.405		-0.045			-0.108	-2.2	-0.043			-0.028
23.9		-0.496			-0.594		-0.055			-0.131		-0.053			-0.033
27.7		-0.719			-0.765		-0.064			-0.152		-0.061			-0.039
31		-0.913		-45.0	-0.914		-0.071			-0.171		-0.068			-0.043
34.7	-58.7	-1.130			-1.019		-0.080			-0.191		-0.076			-0.049
39.7		-1.423			-1.387		-0.227			-0.339		-0.087	-1.4		-0.056
43.7		-1.658	-28.6		-1.682		-0.344			-0.457		-0.096			-0.061
47.2		-1.864			-1.939		-0.447	-29.6	-7.7	-0.588		-0.110		-16.4	-0.123
50.7		-2.069			-2.197		-0.550			-0.718		-0.123			-0.186
55.7		-2.363			-2.565	-29.4	-0.697			-0.905		-0.143			-0.275
65.2		-2.450			-2.674		-0.977		-28.2	-1.454	-3.9	-0.180		-24.6	-0.522
71.8		-2.511			-2.750		-1.171			-1.835		-0.206			-0.693
78.4	-9.2	-2.571	-6.1	-5.4	-2.826		-1.365			-2.217		-0.231			-0.865
99.3		-2.764			-3.066		-1.979			-2.342		-0.313			-0.961
107.7		-2.841			-3.163		-2.226	-3.4	-2.6	-2.393		-0.346		-3.2	-1.000
116.7		-2.924			-3.266		-2.491			-2.447		-0.381			-1.041
...															
190.7															

Fraction of sugar remaining in each growth condition (duplicates averaged), relative to sugar available at t=0 hrs:

Condition	t=0 hrs	t=11 hrs	t=24 hrs	t=35 hrs	t=44 hrs	t=56 hrs	t=78 hrs	t=117 hrs
Glucose	1.000	.993	.863	.677	.528	.337	.266	.178
Sucrose	1.000	1.019	1.008	.983	.902	.815	.667	.420
SG: Sucrose	1.000	.994	.995	.996	.929	.855	.823	.768
SG: Glucose	1.000	1.000	.866	.732	.576	.483	.422	.390
SaddG: Sucrose	1.000	.971	.980	.947	.804	.788	.627	.596
SaddG: Glucose					1.000	.969	.760	.726
Maltose	1.000	.989	1.031	.979	.972	.945	.919	.900
MaddG: Maltose	1.000	1.005	1.002	.998	1.009	.962	.980	.969
MaddG: Glucose					1.000	.934	.754	.712

Appendix 2.6. Proportion of significantly up- and down-regulated genes in the RNA-Seq dataset across growth condition comparisons ($p < 0.05$).

Growth condition comparison	Proportion of genes significantly expressed	Number of up-regulated genes	Number of down-regulated genes
Maltose/Glucose	692 / 3,994	361	331
Maltose/Rich	828 / 3,964	496	332
Sucrose/Glucose	870 / 3,994	412	458
Sucrose/Rich	1,021 / 3,964	575	446
Glucose/Rich	1,025 / 3,464	599	426
Maltose/Sucrose	275 / 3,964	169	106

Appendix 2.7. *H. borinquense* genes identified as non-carbohydrate-related ABC-type transport system proteins with their annotations and RNA-Seq log2 fold change values normalized to the rich media condition.

Name (Hbor)	KEGG Orthology (KO)-assigned functional category	KO-assigned gene	G/R	S/R	M/R
23310	Cu2+-processing for NO ₂ reductase	NosY	-0.85	-1.54	-1.33
30470		NosF	-4.04	-6.79	-8.68
27480	phosphonate transport	PhnD	-0.68	-0.75	0.00
27490		PhnE	-2.36	-1.14	-0.27
27500		PhnC	-1.43	-1.11	-0.49
08060	phosphate transport	PstS	5.75	0.28	-0.06
08070		PstS	1.10	0.15	-0.07
08080		PstC	2.46	1.26	1.18
08090		PstA	2.88	1.33	1.22
08100		PstB	2.29	0.63	0.97
14810	branched-chain amino acid transport	LivK	5.93	4.07	4.19
20510		LivK	3.72	3.86	4.28
36390		LivK	2.14	3.98	4.64
12740		LivH	4.25	3.86	4.90
14840		LivH	8.65	5.64	5.58
20500		LivH	3.07	3.15	3.29
36380		LivH/LivM	3.22	3.88	3.86
12730		LivM	5.66	4.97	5.75
14850		LivM	5.37	2.18	3.41
20490		LivM	2.70	2.65	3.10
12720		LivG	5.12	3.87	5.20
14820		LivG	5.99	4.22	4.11
20480		LivG	2.09	1.72	2.68

36370		LivG	1.91	2.08	2.12
12710		LivF	4.09	2.77	3.73
14830		LivF	6.11	3.47	3.79
20470		LivF	2.15	1.47	2.18
36360		LivF	2.47	2.73	1.80
19310	tungstate transport	TupA	-1.70	-5.90	-5.20
04960		TupB	0.62	-0.33	0.21
04950		TupC	0.58	-0.67	-0.63
32110	alkanesulfonate transport	SsuA	-2.33	-0.29	-0.67
31960		SsuA	0.47	1.08	0.47
31950		SsuC	0.35	0.37	-0.22
32120		SsuC	-1.96	-0.38	-0.22
31940		SsuB	0.53	0.69	0.92
32130		SsuB	-0.73	-0.09	0.25
15060	iron (III) transport	AfuA	-3.61	-1.23	-1.47
15050		AfuB	-2.65	-0.56	-0.84
15040		AfuC	-1.16	-0.13	-0.23
16350	spermidine / putrescine transport	PotD	6.31	5.52	5.49
38130		PotD	3.71	3.11	3.81
16370		PotC	4.75	3.64	3.31
38100		PotC	2.06	1.39	1.58
16360		PotB	6.53	5.55	5.01
38110		PotB	3.63	2.71	3.23
16340		PotA	4.12	4.19	4.09

Appendix 2.8. *H. borinquense* genes identified as TCA Cycle proteins with their annotations, associated reactions, and RNA-Seq log2 fold change values normalized to the rich media condition.

Name (Hbor)	Annotation	Reaction	G/R	S/R	M/R
27300	citrate synthase	acetyl-CoA + oxaloacetate ↔ citrate	-1.39	1.79	0.54
34280	citrate synthase	acetyl-CoA + oxaloacetate ↔ citrate	-1.45	2.72	1.50
05490	aconitate hydratase	citrate ↔ cis-aconitate; cis-aconitate ↔ isocitrate	-0.11	-0.41	-0.06
32160	aconitate hydratase	citrate ↔ cis-aconitate; cis-aconitate ↔ isocitrate	-1.03	-4.23	-5.17
09930	isocitrate dehydrogenase (NADP)	isocitrate ↔ oxalosuccinate ↔ 2-oxoglutarate	0.23	-0.34	-1.00
23330	2-oxoglutarate/2-oxoacid ferredoxin oxidoreductase subunit alpha	2-oxoglutarate ↔ succinyl-CoA	-1.94	0.28	0.26
23340	2-oxoglutarate/2-oxoacid ferredoxin oxidoreductase subunit beta	2-oxoglutarate ↔ succinyl-CoA	-0.98	0.08	0.13
10970	succinyl-CoA synthetase (ADP-forming) alpha subunit	succinyl-CoA ↔ succinate	-0.53	1.00	0.31
10980	succinyl-CoA synthetase (ADP-forming) beta subunit	succinyl-CoA ↔ succinate	-0.78	0.84	0.20
35380	succinyl-CoA:acetate CoA-transferase	succinyl-CoA ↔ succinate	-1.12	0.95	-0.32
12870	succinate dehydrogenase / fumarate reductase, flavoprotein subunit	succinate ↔ fumarate	0.71	0.39	0.83
12880	succinate dehydrogenase / fumarate reductase, iron-sulfur subunit	succinate ↔ fumarate	0.78	0.96	1.22
12890	succinate dehydrogenase / fumarate reductase, membrane anchor subunit	succinate ↔ fumarate	0.56	0.49	0.68
12900	succinate dehydrogenase / fumarate reductase, cytochrome b subunit	succinate ↔ fumarate	0.01	-0.13	0.04
14250	fumarate hydratase, class II	fumarate ↔ (S)-malate	-0.61	0.48	-0.54
02300	malate dehydrogenase (NAD)	(S)-malate ↔ oxaloacetate	0.45	0.38	0.52

Appendix 2.9. *H. borinquense* genes identified as amino acid biosynthesis proteins with their associated annotations, modules, reactions, and RNA-Seq log2 fold change values normalized to the rich media condition.

Name (Hbor)	Annotation	Pathway	G/R	S/R	M/R
01700	argininosuccinate synthase	arginine biosynthesis	3.93	4.82	4.73
01710	argininosuccinate lyase	arginine biosynthesis	4.78	4.99	4.70
01720	LysW, alpha-aminoadipate/glutamate carrier protein	lysine biosynthesis	5.66	6.25	6.29
01730	LysW, L-2-aminoadipate ligase	lysine biosynthesis	5.42	5.81	5.37
01740	LysW, gamma-L-alpha-aminoadipyl-6-phosphate/L-glutamyl-5-phosphate reductase	lysine biosynthesis	5.67	5.94	5.49
01750	ArgB, N2-acetyl-L-aminoadipate kinase	lysine biosynthesis	5.85	6.67	6.50
01760	ArgD, LysW-gamma-L-lysine/LysW-L-ornithine aminotransferase	lysine biosynthesis	4.71	5.36	5.05
01770	ArgE, [amino group carrier protein]-lysine/ornithine hydrolase	lysine biosynthesis	6.03	6.21	5.70
01780	ornithine carbamoyltransferase	arginine biosynthesis	4.98	4.94	4.58
04460	acetylornithine deacetylase	arginine biosynthesis	0.99	1.18	0.70
09290	dihydroxy-acid dehydratase	valine, leucine, and isoleucine biosynthesis	2.03	3.13	1.91
16510	arginase	arginine biosynthesis	0.30	0.76	0.38
17280	2-isopropylmalate synthase	leucine biosynthesis	5.43	7.86	6.67
17300	acetolactate synthase I/II/III large subunit	valine, leucine, and isoleucine biosynthesis	1.85	4.27	3.07
17310	acetolactate synthase I/II/III small subunit	valine, leucine, and isoleucine biosynthesis	1.54	3.94	2.87
17320	ketol-acid reductoisomerase	valine, leucine, and isoleucine biosynthesis	1.40	4.13	2.81
17340	3-isopropylmalate/(R)-2-methylmalate dehydratase large subunit	leucine and isoleucine biosynthesis	1.83	4.63	3.33
17350	3-isopropylmalate/(R)-2-methylmalate dehydratase small subunit	leucine and isoleucine biosynthesis	1.99	4.92	3.28
17360	3-isopropylmalate dehydrogenase	leucine and isoleucine biosynthesis	2.08	5.07	3.46
22850	acetylornithine deacetylase	arginine biosynthesis	1.22	4.64	3.54
25080	branched chain amino acid aminotransferase	valine, leucine, and isoleucine biosynthesis	0.27	0.18	0.03
25660	(R)-citramalate synthase	isoleucine biosynthesis	5.56	5.51	5.24
27740	acetylornithine deacetylase	arginine biosynthesis	1.33	2.30	1.81
28430	branched chain amino acid aminotransferase	valine, leucine, and isoleucine biosynthesis	0.17	0.82	0.29
29460	threonine aldolase	glycine, serine, and threonine biosynthesis	-0.46	-2.73	-3.97
35190	acetolactate synthase I/II/III large subunit	valine, leucine, and isoleucine biosynthesis	-0.06	4.22	2.14

Appendix 2.10. *H. borinquense* genes for functional clusters with their annotations and RNA-Seq log2 fold change values normalized to the rich media condition.

Carotenoid Biosynthesis Genes				
Name (Hbor)	Annotation	G/R	S/R	M/R
10650	1-hydroxy-2-isopentenylcarotenoid 3,4-desaturase	-1.17	-2.66	-1.82
10660	lycopene elongase/hydratase (dihydrobisanhydrobacterioruberin-forming)	-0.99	-1.63	-1.13
10670	bisanhydrobacterioruberin hydratase	-0.43	-1.07	-0.68
10680	15-cis-phytoene synthase	0.94	-0.56	0.78
Chemotaxis Genes				
Name (Hbor)	Annotation	G/R	S/R	M/R
20350	response regulator containing CheY-like receiver domain and AraC-type DNA-binding domain	-0.83	1.56	1.30
20360	chemotaxis protein CheC, inhibitor of MCP methylation	-0.35	2.88	2.30
20370	chemotaxis protein CheD	0.56	3.88	3.01
20160	purine-binding chemotaxis protein CheW	-0.35	1.58	0.80
20170	CheB, chemotaxis response regulator containing a CheY-like receiver domain and a methylesterase domain	-0.84	2.14	1.86
20180	two-component system, chemotaxis family, sensor kinase CheA, histidine kinase like	0.39	3.39	2.85
20190	chemotaxis protein methyltransferase CheR	0.78	3.28	2.64
06000	methyl-accepting chemotaxis protein	-1.76	2.05	2.03
06520	methyl-accepting chemotaxis protein	0.67	3.89	3.59
06750	methyl-accepting chemotaxis protein	1.14	0.37	0.18
26550	methyl-accepting chemotaxis protein	-2.22	1.07	0.86
27730	methyl-accepting chemotaxis protein	-2.77	-4.49	-5.17
30220	methyl-accepting chemotaxis protein	1.29	2.85	2.34
37090	methyl-accepting chemotaxis protein	-0.41	1.79	1.61

Gas Vesicle Genes				
Name (Hbor)	Annotation	G/R	S/R	M/R
32490	gas vesicle protein	4.33	1.28	3.94
32500	gas vesicle synthesis protein GvpL/GvpF	3.80	0.70	3.51
32510	gas vesicle protein K	2.86	0.14	2.20
32520	gas vesicle protein	4.08	1.22	4.17
32560	gas vesicle synthesis protein GvpL/GvpF	4.51	1.36	4.32
32570	PadR family transcriptional regulator, activator of gas vesicle formation (GenBank) transcriptional regulator, PadR family	5.72	2.44	5.28
32580	gas vesicle operon repressor (GenBank) RecA-superfamily ATPase possibly involved in signal transduction	4.78	1.65	3.80
32590	gas vesicle structural protein (GenBank) Gas vesicle protein	4.67	1.88	3.62
32600	gas vesicle minor protein (GenBank) gas vesicle protein C (GVPC) family protein	5.39	2.08	4.65
32620	gas vesicle protein GvpN	5.37	2.14	4.54
32630	gas vesicle synthesis protein GvpO	3.78	0.73	3.13
Biotin Biosynthesis				
Name (Hbor)	Annotation	G/R	S/R	M/R
10510	3-oxoacyl-[acyl-carrier-protein] synthase II	1.80	1.67	1.54
10480	3-oxoacyl-[acyl-carrier protein] reductase	2.02	1.31	1.69
10640	3-oxoacyl-[acyl-carrier protein] reductase	0.28	0.50	-0.09
22400	3-oxoacyl-[acyl-carrier protein] reductase	-0.13	-0.61	0.64
27900	3-oxoacyl-[acyl-carrier protein] reductase	1.63	1.90	1.49
30020	3-oxoacyl-[acyl-carrier protein] reductase	5.56	1.56	0.60
34640	3-oxoacyl-[acyl-carrier protein] reductase	0.55	1.17	1.65
37100	3-oxoacyl-[acyl-carrier protein] reductase	1.43	1.18	1.16
37300	3-oxoacyl-[acyl-carrier protein] reductase	1.67	2.97	1.53
32750	biotin synthase	3.31	3.49	2.95
32760	8-amino-7-oxononanoate carboxylating dehydrogenase	4.07	4.13	3.69
32770	8-amino-7-oxononanoate synthase	3.73	3.61	3.21
32780	dethiobiotin synthetase	3.06	2.74	2.42

Pentose Phosphate Pathway					
Name (Hbor)	Annotation	Reaction	G/R	S/R	M/R
03760	6-phosphogluconate dehydrogenase (decarboxylating)	D-gluconate-6P → D-ribulose-5P	-0.74	-0.62	-0.14
00450	ribose-5-phosphate isomerase	D-ribulose-5P ↔ D-ribose-5P	-0.32	-0.23	0.14
29570	ribose-phosphate pyrophosphokinase	D-ribose-5P ↔ PRPP	0.12	-0.32	-0.52
18320	phosphopentomutase	D-ribose-1P ↔ D-ribose-5P	-2.05	-0.81	-1.00
02570	phosphomannomutase / phosphoglucomutase	D-ribose-1P ↔ D-ribose-5P	-2.14	-1.27	-0.70
09670	ribokinase	D-ribose-5P ↔ D-ribose; 2-deoxy-D-ribose ↔ 2-deoxy-D-ribose-5P	0.88	1.55	1.34
34590	ribokinase	D-ribose-5P ↔ D-ribose; 2-deoxy-d-ribose ↔ 2-deoxy-D-ribose-5P	-0.28	-0.47	-0.50
09700	deoxyribose-phosphate aldolase	2-deoxy-D-ribose-5p ↔ D-glyceraldehyde-3P	-0.96	-1.00	-1.04
06240	aldehyde:ferredoxin oxidoreductase	D-glyceraldehyde → glycerate	1.10	1.48	2.42
26060	aldehyde:ferredoxin oxidoreductase	D-glyceraldehyde → glycerate	0.10	-0.27	-0.95
31050	aldehyde:ferredoxin oxidoreductase	D-glyceraldehyde → glycerate	-4.23	-3.41	-0.38
16760	glycerate 2-kinase	glycerate → glycerate-2P	1.93	0.90	0.68
17420	gluconate/galactonate dehydratase	D-gluconate → 2-dehydro-3-deoxy-D-gluconate	-0.20	1.06	-0.58
26610	2-dehydro-3-deoxygluconokinase	2-dehydro-3-deoxy-D-gluconate → 2-dehydro-3-deoxy-D-gluconate-6P	0.58	1.99	0.83
21830	2-dehydro-3-deoxyphosphogluconate aldolase / (4S)-4-hydroxy-2-oxoglutarate aldolase	2-dehydro-3-deoxy-D-gluconate-6P → pyruvate or D-glyceraldehyde-3P	0.26	1.70	0.31
33540	D-glucosamine-6-phosphate ammonia-lyase	D-glucosamine-6P → 2-dehydro-3-deoxy-D-gluconate-6P	1.00	3.07	1.29

Appendix 2.11. List of genes in RNA-Seq heatmap clusters (Figure 3.3).

Genes in each of the 16 clusters from the RNA-Seq datasets for rich, glucose, maltose, and sucrose conditions filtered to include significantly expressed genes with a ± 2 log₂ fold change, using the “complete” clustering algorithm (n=794 genes total).

Cluster 1			
Name (Hbor)	COG	Family	Annotation
00390-400	COG805	U	Sec-independent protein secretion pathway component TatC
00470			
01060	COG1976	J	Translation initiation factor 6 (eIF-6)
01070	COG2097	J	Ribosomal protein L31E
01080			GO:0006412, GO:0005840, GO:0003735
01480	COG1131	V	ABC-type multidrug transport system, ATPase component
02380	COG1943	L	Transposase and inactivated derivatives
02970	COG1071	C	Pyruvate/2-oxoglutarate dehydrogenase complex, dehydrogenase (E1) component, eukaryotic type, alpha subunit
03790	COG2125	J	Ribosomal protein S6E (S10)
04740	COG3253	S	Uncharacterized conserved protein
05050	COG1412	R	Uncharacterized proteins of PilT N-term./Vapc superfamily
05060	COG5257	J	Translation initiation factor 2, gamma subunit (eIF-2gamma; GTPase)
05710	COG341	U	Preprotein translocase subunit SecF
05720	COG342	U	Preprotein translocase subunit SecD
05840	COG405	E	Gamma-glutamyltransferase
06550			
06610	COG547	E	Anthranilate phosphoribosyltransferase
07300			
10220	COG2106	S	Uncharacterized conserved protein
10230	COG87	J	Ribosomal protein L3
10240	COG88	J	Ribosomal protein L4
10250	COG89	J	Ribosomal protein L23
10260	COG90	J	Ribosomal protein L2
10270	COG185	J	Ribosomal protein S19
10280	COG91	J	Ribosomal protein L22
10290	COG92	J	Ribosomal protein S3
10300	COG255	J	Ribosomal protein L29
10310	COG1588	J	RNase P/RNase MRP subunit p29
10320	COG186	J	Ribosomal protein S17
10330	COG93	J	Ribosomal protein L14
10340	COG198	J	Ribosomal protein L24

10350	COG1471	J	Ribosomal protein S4E
10360	COG94	J	Ribosomal protein L5
10370	COG199	J	Ribosomal protein S14
10380	COG96	J	Ribosomal protein S8
10390	COG97	J	Ribosomal protein L6P/L9E
10400	COG1717	J	Ribosomal protein L32E
10410	COG2147	J	Ribosomal protein L19E
10420	COG256	J	Ribosomal protein L18
10430	COG98	J	Ribosomal protein S5
10440	COG1841	J	Ribosomal protein L30/L7E
10450	COG200	J	Ribosomal protein L15
10460	COG201	U	Preprotein translocase subunit SecY
10890			
10900			GO:0005622, GO:0003676, GO:0008270
10910			
11290	COG1278	K	Cold shock proteins
11560			
11790	COG1960	I	Acyl-CoA dehydrogenases
12060	COG1358	J	Ribosomal protein HS6-type (S12/L30/L7a)
12070	COG2053	J	Ribosomal protein S28E/S33
12080	COG2075	J	Ribosomal protein L24E
12090	COG105	F	Nucleoside diphosphate kinase
12210	COG2888	J	Predicted Zn-ribbon RNA-binding protein with a function in translation
12230	COG2058	J	Ribosomal protein L12E/L44/L45/RPP1/RPP2
12240	COG244	J	Ribosomal protein L10
12250	COG81	J	Ribosomal protein L1
12260	COG80	J	Ribosomal protein L11
12460	COG52	J	Ribosomal protein S2
12470	COG148	G	Enolase TIGR01060 phosphopyruvate hydratase [eno]
12480	COG1758	K	DNA-directed RNA polymerase, subunit K/omega
12500	COG103	J	Ribosomal protein S9
14400	COG605	P	Superoxide dismutase
14600	COG12	J	Predicted GTPase, probable translation factor
15060	COG1840	P	ABC-type Fe3+ transport system, periplasmic component
15250	COG697	GER	Permeases of the drug/metabolite transporter (DMT) superfamily
17920	COG4799	I	Acetyl-CoA carboxylase, carboxyltransferase component (subunits alpha and beta)
18370			
19460			

19910	COG1999	R	Uncharacterized protein SCO1/SenC/PrrC, involved in biogenesis of respiratory and photosynthetic systems
19920			
20930	COG184	J	Ribosomal protein S15P/S13E
20940-50			
20960	COG1890	J	Ribosomal protein S3AE
21290	COG6	E	Xaa-Pro aminopeptidase
21630			GO:0006355, GO:0003677
21840	COG842	V	ABC-type multidrug transport system, permease component
21940	COG615	MI	Cytidylyltransferase
22110			
22420	COG569	P	K ⁺ transport systems, NAD-binding component
22840			
23050			
23720			
23940	COG2322	S	Predicted membrane protein
25160	COG1631	J	Ribosomal protein L44E
25170	COG2051	J	Ribosomal protein S27E
25180	COG1093	J	Translation initiation factor 2, alpha subunit (eIF-2alpha)
25550	COG1997	J	Ribosomal protein L37AE/L43A
26510	COG1632	J	Ribosomal protein L15E
27100	COG1278	K	Cold shock proteins
28550	COG1156	C	Archaeal/vacuolar-type H ⁺ -ATPase subunit B
29280			
32170			
32200	COG1757	C	Na ⁺ /H ⁺ antiporter
32410	COG3473	Q	Maleate cis-trans isomerase
32420	COG146	EQ	N-methylhydantoinase B/acetone carboxylase, alpha subunit
32430	COG145	EQ	N-methylhydantoinase A/acetone carboxylase, beta subunit
32440	COG1914	P	Mn ²⁺ and Fe ²⁺ transporters of the NRAMP family
32460			
33960	COG2834	M	Outer membrane lipoprotein-sorting protein
33970	COG616	OU	Periplasmic serine proteases (ClpP class)
33980			
34380	COG3794	C	Plastocyanin
34610			GO:0006118, GO:0016491
34870			
36280	COG3396	S	Uncharacterized conserved protein
36290			

36300	COG3396	S	Uncharacterized conserved protein
36310	COG2151	R	Predicted metal-sulfur cluster biosynthetic enzyme
36320			
36800			
37490	COG444	EP	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component
37500	COG444	EP	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component
38580	COG614	P	ABC-type Fe ³⁺ -hydroxamate transport system, periplasmic component

Cluster 2			
Name (Hbor)	COG	Family	Annotation
00640			
01200	COG1926	R	Predicted phosphoribosyltransferases
01260	COG702	MG	Predicted nucleoside-diphosphate-sugar epimerases
01300			
02360			
03650	COG694	O	Thioredoxin-like proteins and domains
04810			
05380	COG714	R	MoxR-like ATPases
05400	COG1305	E	Transglutaminase-like enzymes, putative cysteine proteases
06200			
06210	COG659	P	Sulfate permease and related transporters (MFS superfamily)
06470			
06840			
07880	COG458	EF	Carbamoylphosphate synthase large subunit (split gene in MJ)
07900	COG702	MG	Predicted nucleoside-diphosphate-sugar epimerases
08080	COG573	P	ABC-type phosphate transport system, permease component
08270			
09340			
09350			GO:0006415, GO:0005737, GO:0016149
10200	COG3794	C	Plastocyanin
10760	COG512	EH	Anthranilate/para-aminobenzoate synthases component II
10770	COG147	EH	Anthranilate/para-aminobenzoate synthases component I
10780	COG135	E	Phosphoribosylanthranilate isomerase
10790	COG547	E	Anthranilate phosphoribosyltransferase
11170	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins
11340	COG316	S	Uncharacterized conserved protein
11980	COG1146	C	Ferredoxin

14010			
14800	COG3413	R	Predicted DNA binding protein
14810	COG683	E	ABC-type branched-chain amino acid transport systems, periplasmic component
14820	COG411	E	ABC-type branched-chain amino acid transport systems, ATPase component
14830	COG410	E	ABC-type branched-chain amino acid transport systems, ATPase component
14840	COG559	E	Branched-chain amino acid ABC-type transport system, permease components
14850	COG4177	E	ABC-type branched-chain amino acid transport system, permease component
14860	COG332	I	3-oxoacyl-[acyl-carrier-protein] synthase III
14930			GO:0016020
15230			
15420			
15690	COG855	P	Polyphosphate kinase
15790			GO:0005622, GO:0003676, GO:0008270
15800			
15960	COG180	J	Tryptophanyl-tRNA synthetase
16360	COG1176	E	ABC-type spermidine/putrescine transport system, permease component I
16670			
18200			
18720	COG784	T	FOG: CheY-like receiver
18950	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins
19200	COG674	C	Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, alpha subunit
19210	COG1013	C	Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, beta subunit
19610			
20600	COG704	P	Phosphate uptake regulator
21760			
23350			
23620			
24280	COG159	E	Tryptophan synthase alpha chain
24290	COG133	E	Tryptophan synthase beta chain
24300	COG134	E	Indole-3-glycerol phosphate synthase
25600			
28190			GO:0006950
29510	COG174	E	Glutamine synthetase
30320	COG3413	R	Predicted DNA binding protein
30940	COG412	Q	Dienelactone hydrolase and related enzymes
31350			
32490			GO:0012506, GO:0005198

32500			GO:0031412, GO:0031411
32510			GO:0031412
32530-50			
32560			GO:0031412, GO:0031411
32570			
32580	COG467	T	RecA-superfamily ATPases implicated in signal transduction
32590			GO:0012506, GO:0005198
32600			GO:0031412, GO:0031411
32610			GO:0045449, GO:0003677, GO:0008270
32620	COG714	R	MoxR-like ATPases
32630			GO:0031412
32660	COG2202	T	FOG: PAS/PAC domain
33240			
34530	COG3848	T	Phosphohistidine swiveling domain
34880	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins
35330			
37770			
39200	COG3592	S	Uncharacterized conserved protein
39480			

Cluster 3			
Name (Hbor)	COG	Family	Annotation
00920	COG459	O	Chaperonin GroEL (HSP60 family)
04330	COG1681	N	Archaeal flagellins
14180	COG840	NT	Methyl-accepting chemotaxis protein
14940	COG642	T	Signal transduction histidine kinase
20430	COG1960	I	Acyl-CoA dehydrogenases
24220	COG1405	K	Transcription initiation factor TFIIIB, Brf1 subunit/Transcription initiation factor TFIIB
25280	COG614	P	ABC-type Fe ³⁺ -hydroxamate transport system, periplasmic component
28100	COG51	J	Ribosomal protein S10
28110	COG5256	J	Translation elongation factor EF-1alpha (GTPase)
30100-10	COG444	EP	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component
30120	COG1173	EP	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components GO:0006810, GO:0016020, GO:0005215
33120	COG614	P	ABC-type Fe ³⁺ -hydroxamate transport system, periplasmic component
33130	COG492	O	Thioredoxin reductase
36250	COG2159	R	Predicted metal-dependent hydrolase of the TIM-barrel fold

Cluster 4			
Name (Hbor)	COG	Family	Annotation
01280			
01700	COG137	E	Argininosuccinate synthase
01710	COG165	E	Argininosuccinate lyase
01720			TIGR01206 lysine biosynthesis protein LysW [lysW]
01730	COG189	HJ	Glutathione synthase/Ribosomal protein S6 modification enzyme (glutaminyl transferase)
01740	COG2	E	Acetylglutamate semialdehyde dehydrogenase
01750	COG548	E	Acetylglutamate kinase
01760	COG4992	E	Ornithine/acetylornithine aminotransferase
01770	COG624	E	Acetylornithine deacetylase/Succinyl-diaminopimelate desuccinylase and related deacylases
01780	COG78	E	Ornithine carbamoyltransferase
02140			
02760	COG1966	T	Carbon starvation protein, predicted membrane protein
03280			GO:0006810, GO:0005215
03290-300	COG601	EP	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
03310-20	COG444	EP	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component
03330	COG312	R	Predicted Zn-dependent proteases and their inactivated homologs
03590	COG1522	K	Transcriptional regulators
04390			
04510	COG2234	R	Predicted aminopeptidases
04920	COG569	P	K ⁺ transport systems, NAD-binding component
05390	COG1721	R	Uncharacterized conserved protein (some members contain a von Willebrand factor type A (vWA) domain)
05760			
06480			
06670	COG2146	PR	Ferredoxin subunits of nitrite reductase and ring-hydroxylating dioxygenases
07210			GO:0000160, GO:0006355, GO:0000156
07410			
07950			
08090	COG581	P	ABC-type phosphate transport system, permease component
08450			
08690	COG765	E	ABC-type amino acid transport system, permease component
09050	COG3852	T	Signal transduction histidine kinase, nitrogen specific
09060	COG505	EF	Carbamoylphosphate synthase small subunit
09110	COG3844	E	Kynureninase
09290	COG129	EG	Dihydroxyacid dehydratase/phosphogluconate dehydratase
09540	COG2352	C	Phosphoenolpyruvate carboxylase

09640			
10090	COG5258	R	GTPase
10590			GO:0006950
11010	COG1028	IQR	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)
11360			
11440			
11450	COG673	R	Predicted dehydrogenases and related proteins
11460			
11470	COG3839	G	ABC-type sugar transport systems, ATPase components
11480-500	COG395	G	ABC-type sugar transport system, permease component
11630			
11700	COG1082	G	Sugar phosphate isomerases/epimerases
11710			
11800			GO:0003676
12710	COG410	E	ABC-type branched-chain amino acid transport systems, ATPase component
12730-40	COG4177	E	ABC-type branched-chain amino acid transport system, permease component
13570-80	COG2719	S	Uncharacterized conserved protein
13590	COG2766	T	Putative Ser protein kinase
13940			
14980	COG296	G	1,4-alpha-glucan branching enzyme
14990			GO:0005975, GO:0003824, GO:0043169
15370			
16120			
16150	COG436	E	Aspartate/tyrosine/aromatic aminotransferase
16320	COG1063	ER	Threonine dehydrogenase and related Zn-dependent dehydrogenases
16340	COG3842	E	ABC-type spermidine/putrescine transport systems, ATPase components
16350	COG687	E	Spermidine/putrescine-binding periplasmic protein
16370	COG1177	E	ABC-type spermidine/putrescine transport system, permease component II
16390	COG2031	I	Short chain fatty acids transporter
16820			
17280	COG119	E	Isopropylmalate/homocitrate/citramalate synthases
17570			
18960			
19380			
19390	COG693	R	Putative intracellular protease/amidase
20490	COG4177	E	ABC-type branched-chain amino acid transport system, permease component
20500	COG559	E	Branched-chain amino acid ABC-type transport system, permease components
20510	COG683	E	ABC-type branched-chain amino acid transport systems, periplasmic component

21360			
21880			
23230	COG456	R	Acetyltransferases
23260	COG365	I	Acyl-coenzyme A synthetases/AMP-(fatty) acid ligases
23460	COG1803	G	Methylglyoxal synthase
23510	COG69	E	Glutamate synthase domain 2
23670-80			
24270	COG1830	G	DhnA-type fructose-1,6-bisphosphate aldolase and related enzymes
24410			
25350			
25660	COG119	E	Isopropylmalate/homocitrate/citramalate synthases
25940			
26380			
26630	COG1249	C	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide dehydrogenase (E3) component, and related enzymes
27080			GO:0010181
27140			GO:0008152, GO:0008080
27390			
27460	COG77	E	Prephenate dehydratase
28150			
30090	COG747	E	ABC-type dipeptide transport system, periplasmic component
30180	COG1414	K	Transcriptional regulator
30510	COG277	C	FAD/FMN-containing dehydrogenases
32240	COG1249	C	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide dehydrogenase (E3) component, and related enzymes GO:0045454, GO:0005737, GO:0016491
32750	COG502	H	Biotin synthase and related enzymes
32760			
32770	COG156	H	7-keto-8-aminopelargonate synthetase and related enzymes
32780	COG132	H	Dethiobiotin synthetase
33610	COG1472	G	Beta-glucosidase-related glycosidases
34770			
34990	COG3413	R	Predicted DNA binding protein
36380	COG4177	E	ABC-type branched-chain amino acid transport system, permease component
36500			
36530			GO:0006355, GO:0005622, GO:0003700
36540	COG449	M	Glucosamine 6-phosphate synthetase, contains amidotransferase and phosphosugar isomerase domains
36560			
38110	COG1176	E	ABC-type spermidine/putrescine transport system, permease component I
38130	COG687	E	Spermidine/putrescine-binding periplasmic protein

38140	COG1454	C	Alcohol dehydrogenase, class IV
38170	COG174	E	Glutamine synthetase
38270	COG306	P	Phosphate/sulphate permeases
38310	COG1012	C	NAD-dependent aldehyde dehydrogenases
39240	COG346	E	Lactoylglutathione lyase and related lyases
39470	COG592	L	DNA polymerase sliding clamp subunit (PCNA homolog)
39540	COG3153	R	Predicted acetyltransferase
39630	COG2259	S	Predicted membrane protein

Cluster 5			
Name (Hbor)	COG	Family	Annotation
01360	COG4	P	Ammonia permease
01370	COG347	E	Nitrogen regulatory protein PII
01380			
04530	COG1011	R	Predicted hydrolase (HAD superfamily)
04860	COG171	H	NAD synthase
05650	COG398	S	Uncharacterized conserved protein
05660	COG1183	I	Phosphatidylserine synthase
07610			GO:0006352, GO:0006355, GO:0005622, GO:0003700, GO:0016987
08060	COG226	P	ABC-type phosphate transport system, periplasmic component
08430	COG697	GER	Permeases of the drug/metabolite transporter (DMT) superfamily
08600	COG3467	R	Predicted flavin-nucleotide-binding protein
09960			
14340	COG745	TK	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain
15660			
16970			GO:0006118, GO:0009055, GO:0046872
17480			
18780			
19770			
21770	COG1409	R	Predicted phosphohydrolases
21780			
22020	COG783	P	DNA-binding ferritin-like protein (oxidative damage protectant)
24400			TIGR03618 PPOX class probable F420-dependent enzyme
25230			
29780	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins
29790			

29930-50			
29960	COG3243	I	Poly(3-hydroxyalkanoate) synthetase
29970			
29980	COG2030	I	Acyl dehydratase
30020	COG1028	IQR	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)
30030	COG4799	I	Acetyl-CoA carboxylase, carboxyltransferase component (subunits alpha and beta)
30040			
30950	COG1653	G	ABC-type sugar transport system, periplasmic component
30960-70	COG1175	G	ABC-type sugar transport system, permease components
30980	COG3839	G	ABC-type sugar transport systems, ATPase components
31060	COG1840	P	ABC-type Fe ³⁺ transport system, periplasmic component
31070	COG1175	G	ABC-type sugar transport system, permease components
31080	COG1177	E	ABC-type spermidine/putrescine transport system, permease component II
31090	COG3842	E	ABC-type spermidine/putrescine transport systems, ATPase components
36170			

Cluster 6			
Name (Hbor)	COG	Family	Annotation
01590			GO:0006118, GO:0009055, GO:0046872
03560	COG3413	R	Predicted DNA binding protein
03600			
05700			
05790	COG2513	G	PEP phosphonmutase and related enzymes
06490	COG569	P	K ⁺ transport systems, NAD-binding component
06500			
06780			
08700	COG834	ET	ABC-type amino acid transport/signal transduction systems, periplasmic component/domain
11390	COG3360	S	Uncharacterized conserved protein
12030	COG1247	M	Sortase and related acyltransferases
12690			TIGR01409 Tat (twin-arginine translocation) pathway signal sequence
12720	COG411	E	ABC-type branched-chain amino acid transport systems, ATPase component
13600	COG2766	T	Putative Ser protein kinase
15720			
15810			
18930	COG438	M	Glycosyltransferase
19600			

20480	COG411	E	ABC-type branched-chain amino acid transport systems, ATPase component
21320-30			
21740	COG3361	S	Uncharacterized conserved protein
23490			
26870			
28770	COG3707	T	Response regulator with putative antiterminator output domain
28870			
32520			GO:0012506,GO:0005198
35520	COG2015	Q	Alkyl sulfatase and related hydrolases
36790			
39190	COG1335	Q	Amidases related to nicotinamidase
39270	COG11	S	Uncharacterized conserved protein

Cluster 7			
Name (Hbor)	COG	Family	Annotation
01900			
02390	COG675	L	Transposase and inactivated derivatives
03110	COG626	E	Cystathionine beta-lyases/cystathionine gamma-synthases
04710			
07770			
09200			GO:0006118, GO:0009055, GO:0046872
13670			
14660			
25960	COG5485	R	Predicted ester cyclase
26990	COG57	G	Glyceraldehyde-3-phosphate dehydrogenase/erythrose-4-phosphate dehydrogenase
27200			
31050	COG2414	C	Aldehyde:ferredoxin oxidoreductase
36930			

Cluster 8			
Name (Hbor)	COG	Family	Annotation
02440	COG2873	E	O-acetylhomoserine sulfhydrylase
02450	COG2021	E	Homoserine acetyltransferase
02470	COG2873	E	O-acetylhomoserine sulfhydrylase
02480	COG306	P	Phosphate/sulphate permeases
02560			GO:0008152, GO:0008080
02880	COG498	E	Threonine synthase
06300	COG788	F	Formyltetrahydrofolate hydrolase
07820	COG530	P	Ca ²⁺ /Na ⁺ antiporter
07830	COG4552	R	Predicted acetyltransferase involved in intracellular survival and related acetyltransferases
08580	COG1670	J	Acetyltransferases, including N-acetylases of ribosomal proteins
09360	COG3367	S	Uncharacterized conserved protein
09370	COG4948	MR	L-alanine-DL-glutamate epimerase and related enzymes of enolase superfamily
09470	COG1907	R	Predicted archaeal sugar kinases
10550	COG517	R	FOG: CBS domain
11210	COG136	E	Aspartate-semialdehyde dehydrogenase
12110	COG620	E	Methionine synthase II (cobalamin-independent)
13700	COG1522	K	Transcriptional regulator
13750	COG112	E	Glycine/serine hydroxymethyltransferase
15180			GO:0009396, GO:0005524, GO:0030272
15190	COG1139	C	Uncharacterized conserved protein containing a ferredoxin-like domain
15920	COG2220	R	Predicted Zn-dependent hydrolases of the beta-lactamase fold
17290			
17300	COG28	EH	Thiamine pyrophosphate-requiring enzymes [acetolactate synthase, pyruvate dehydrogenase (cytochrome), glyoxylate carboligase, phosphonopyruvate decarboxylase]
17310	COG440	E	Acetolactate synthase, small (regulatory) subunit
17320	COG59	EH	Ketol-acid reductoisomerase
17330			
17340	COG65	E	3-isopropylmalate dehydratase large subunit
17350	COG66	E	3-isopropylmalate dehydratase small subunit
17360	COG473	CE	Isocitrate/isopropylmalate dehydrogenase
18680-90	COG2391	R	Predicted transporter component
18800			
19650	COG436	E	Aspartate/tyrosine/aromatic aminotransferase
21820	COG288	P	Carbonic anhydrase
22860	COG253	E	Diaminopimelate epimerase
22870	COG19	E	Diaminopimelate decarboxylase

22880	COG2171	E	Tetrahydrodipicolinate N-succinyltransferase
22890	COG289	E	Dihydrodipicolinate reductase
22900	COG329	EM	Dihydrodipicolinate synthase/N-acetylneuraminate lyase
24920	COG436	E	Aspartate/tyrosine/aromatic aminotransferase
26260	COG3794	C	Plastocyanin
28120	COG460	E	Homoserine dehydrogenase
28130	COG2061	E	ACT-domain-containing protein, predicted allosteric regulator of homoserine dehydrogenase
29650	COG384	R	Predicted epimerase, PhzC/PhzF homolog
30070-80	COG1173	EP	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
30130	COG601	EP	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
30140	COG747	E	ABC-type dipeptide transport system, periplasmic component
30160	COG318	IQ	Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II
30170	COG1960	I	Acyl-CoA dehydrogenases
30530	COG2301	G	Citrate lyase beta subunit
31460	COG1051	F	ADP-ribose pyrophosphatase
33300	COG208	F	Ribonucleotide reductase, beta subunit
33410	COG3839	G	ABC-type sugar transport systems, ATPase components
33470	COG1063	ER	Threonine dehydrogenase and related Zn-dependent dehydrogenases
33480	COG2120	S	Uncharacterized proteins, LmbE homologs
33540	COG1921	E	Selenocysteine synthase [seryl-tRNA ^{Ser} selenium transferase]
33550	COG1082	G	Sugar phosphate isomerases/epimerases
34280	COG372	C	Citrate synthase
35150	COG488	R	ATPase components of ABC transporters with duplicated ATPase domains
35180	COG665	E	Glycine/D-amino acid oxidases (deaminating)
35190	COG28	EH	Thiamine pyrophosphate-requiring enzymes [acetolactate synthase, pyruvate dehydrogenase (cytochrome), glyoxylate carboligase, phosphonopyruvate decarboxylase]
35200	COG451	MG	Nucleoside-diphosphate-sugar epimerases
35210			TIGR00277 uncharacterized domain HDIG
35220			TIGR02118 conserved hypothetical protein
35230	COG6	E	Xaa-Pro aminopeptidase
35250	COG22	C	Pyruvate/2-oxoglutarate dehydrogenase complex, dehydrogenase (E1) component, eukaryotic type, beta subunit
35270	COG733	R	Na ⁺ -dependent transporters of the SNF family
35280	COG436	E	Aspartate/tyrosine/aromatic aminotransferase
35300	COG1012	C	NAD-dependent aldehyde dehydrogenases
35360	COG579	R	Predicted dehydrogenase
35370	COG1960	I	Acyl-CoA dehydrogenases
35390			
35400	COG4666	R	TRAP-type uncharacterized transport system, fused permease components

35610	COG406	G	Fructose-2,6-bisphosphatase
36350	COG1541	H	Coenzyme F390 synthetase
36420	COG1541	H	Coenzyme F390 synthetase
37300	COG1028	IQR	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)
37620	COG1414	K	Transcriptional regulator
39760			
39780	COG668	M	Small-conductance mechanosensitive channel
39870	COG1171	E	Threonine dehydratase
39890	COG1249	C	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide dehydrogenase (E3) component, and related enzymes
39910	COG251	J	Putative translation initiation inhibitor, yjgF family

Cluster 9			
Name (Hbor)	COG	Family	Annotation
04260	COG609	P	ABC-type Fe3+-siderophore transport system, permease component
06010			
06460	COG1071	C	Pyruvate/2-oxoglutarate dehydrogenase complex, dehydrogenase (E1) component, eukaryotic type, alpha subunit
08530	COG3425	I	3-hydroxy-3-methylglutaryl CoA synthase
15990			
16020	COG1033	R	Predicted exporters of the RND superfamily
16450	COG1252	C	NADH dehydrogenase, FAD-containing subunit
17880	COG2030	I	Acyl dehydratase
19580	COG1378	K	Predicted transcriptional regulators
19640	COG614	P	ABC-type Fe3+-hydroxamate transport system, periplasmic component
20010			
22230	COG491	R	Zn-dependent hydrolases, including glyoxylases
23830	COG183	I	Acetyl-CoA acetyltransferase
28570	COG1436	C	Archaeal/vacuolar-type H ⁺ -ATPase subunit F
28600			GO:0015986, GO:0045259, GO:0003936, GO:0046933, GO:0046961
29360			
33040			
37640			
39740	COG614	P	ABC-type Fe3+-hydroxamate transport system, periplasmic component

Cluster 10			
Name (Hbor)	COG	Family	Annotation
02070	COG534	V	Na ⁺ -driven multidrug efflux pump
02240	COG1141	C	Ferredoxin
02250			
03270			GO:0008152, GO:0008168
04880			
05220			GO:0006810, GO:0005622
05640			
06570			
07350	COG358	L	DNA primase (bacterial type)
07380	COG3403	S	Uncharacterized conserved protein
07390	COG1335	Q	Amidases related to nicotinamidase
07500	COG1309	K	Transcriptional regulator
10650	COG1233	Q	Phytoene dehydrogenase and related proteins
11540			
17860	COG334	E	Glutamate dehydrogenase/leucine dehydrogenase
18020	COG31	E	Cysteine synthase
19060			
19310	COG2998	H	ABC-type tungstate transport system, permease component
19330			TIGR01409 Tat (twin-arginine translocation) pathway signal sequence
19440			TIGR03618 PPOX class probable F420-dependent enzyme
19890-900			
19980	COG1278	K	Cold shock proteins
20970	COG3794	C	Plastocyanin
21130	COG2835	S	Uncharacterized conserved protein
21850	COG1131	V	ABC-type multidrug transport system, ATPase component
21920	COG1622	C	Heme/copper-type cytochrome/quinol oxidases, subunit 2
21950	COG1310	R	Predicted metal-dependent protease of the PAD1/JAB1 superfamily
22220			
23140	COG675	L	Transposase and inactivated derivatives
23730			
23910	COG1363	G	Cellulase M and related proteins
24180			
24870			
25540	COG4031	R	Predicted metal-binding protein
25970	COG591	ER	Na ⁺ /proline symporter
26290			

27730	COG840	NT	Methyl-accepting chemotaxis protein
28720			
29340	COG633	C	Ferredoxin
30410			GO:0016481, GO:0003677, GO:0016566
30420	COG3794	C	Plastocyanin
30430	COG4263	C	Nitrous oxide reductase
30440			
30450	COG3420	P	Nitrous oxidase accessory protein
30460	COG1131	V	ABC-type multidrug transport system, ATPase component
30470	COG1277	R	ABC-type transport system involved in multi-copper enzyme maturation, permease component
30870			
34300	COG3256	P	Nitric oxide reductase large subunit
34330	COG3413	R	Predicted DNA binding protein
34350			GO:0008033, GO:0005737, GO:0005515, GO:0016783
34360	COG535	R	Predicted Fe-S oxidoreductases
34370			GO:0033014, GO:0008883, GO:0050661
34390			TIGR03995 putative rSAM target protein, CGCGG family
34420	COG2132	Q	Putative multicopper oxidases
34430	COG3794	C	Plastocyanin
34540	COG1680	V	Beta-lactamase class C and other penicillin binding proteins
34750	COG1327	K	Predicted transcriptional regulator, consists of a Zn-ribbon and ATP-cone domains
34900			GO:0006118, GO:0009060, GO:0016021, GO:0005506, GO:0020037
34920	COG1917	S	Uncharacterized conserved protein, contains double-stranded beta-helix domain
34930			
34970			
36000-10			
37480	COG1173	EP	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
37510	COG492	O	Thioredoxin reductase
37740			
39350	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins
39750	COG3794	C	Plastocyanin

Cluster 11			
Name (Hbor)	COG	Family	Annotation
06080			
08590	COG1670	J	Acetyltransferases, including N-acetylases of ribosomal proteins
13330-40			
16140			
16950			
17490			
19560			GO:0008430
20640	COG584	C	Glycerophosphoryl diester phosphodiesterase
21300			
21670	COG1506	E	Dipeptidyl aminopeptidases/acylaminoacyl-peptidases
26190	COG122	L	3-methyladenine DNA glycosylase/8-oxoguanine DNA glycosylase
27020	COG57	G	Glyceraldehyde-3-phosphate dehydrogenase/erythrose-4-phosphate dehydrogenase
27890	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins
30000	COG1545	R	Predicted nucleic-acid-binding protein containing a Zn-ribbon
30010	COG183	I	Acetyl-CoA acetyltransferase
30520	COG346	E	Lactoylglutathione lyase and related lyases
31440	COG463	M	Glycosyltransferases involved in cell wall biogenesis
31490			
31500	COG1209	M	dTDP-glucose pyrophosphorylase
34020	COG2124	Q	Cytochrome P450
35630	COG2079	R	Uncharacterized protein involved in propionate catabolism
36780			
37670			
38150	COG5476	S	Uncharacterized conserved protein
38230-40			
39510			
39570	COG365	I	Acyl-coenzyme A synthetases/AMP-(fatty) acid ligases
39580	COG4327	S	Predicted membrane protein
39590	COG4147	R	Predicted symporter
39600	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins
39610			

Cluster 12			
Name (Hbor)	COG	Family	Annotation
07470	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins
07960	COG778	C	Nitroreductase
09630	COG2820	F	Uridine phosphorylase
11430	COG1378	K	Predicted transcriptional regulators
12490	COG1644	K	DNA-directed RNA polymerase, subunit N (RpoN/RPB10)
12510	COG102	J	Ribosomal protein L13
12520	COG1727	J	Ribosomal protein L18E
12530			
12540	COG202	K	DNA-directed RNA polymerase, alpha subunit/40 kD subunit
12550	COG100	J	Ribosomal protein S11
12560	COG522	J	Ribosomal protein S4 and related proteins
12570	COG99	J	Ribosomal protein S13 TIGR03629 archaeal ribosomal protein S13P
15050	COG1178	P	ABC-type Fe ³⁺ transport system, permease component
15070			GO:0030244, GO:0005794, GO:0009505, GO:0047210
17770			
17950	COG1250	I	3-hydroxyacyl-CoA dehydrogenase
28060			GO:0008152, GO:0003824
33450	COG251	J	Putative translation initiation inhibitor, yjgF family

Cluster 13			
Name (Hbor)	COG	Family	Annotation
02150	COG527	E	Aspartokinases
03910	COG747	E	ABC-type dipeptide transport system, periplasmic component
04420	COG155	P	Sulfite reductase, beta subunit (hemoprotein)
04440			
05450-60			
06520	COG840	NT	Methyl-accepting chemotaxis protein
07580			
08460	COG5256	J	Translation elongation factor EF-1alpha (GTPase)
08880	COG3430	S	Uncharacterized protein conserved in archaea
08890			TIGR02537 archaeal flagellin N-terminal-like domain
11020	COG733	R	Na ⁺ -dependent transporters of the SNF family
11990-200	COG2086	C	Electron transfer flavoprotein, beta subunit (11990) and alpha subunit (12000)
12010	COG644	C	Dehydrogenases (flavoproteins)

12020			
12100	COG620	E	Methionine synthase II (cobalamin-independent)
14140	COG654	HC	2-polyprenyl-6-methoxyphenol hydroxylase and related FAD-dependent oxidoreductases
14240	COG1653	G	ABC-type sugar transport system, periplasmic component
16830	COG554	C	Glycerol kinase
17100			
18640	COG318	IQ	Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II
19950	COG365	I	Acyl-coenzyme A synthetases/AMP-(fatty) acid ligases
20180	COG643	NT	Chemotaxis protein histidine kinase and related kinases
20190	COG1352	NT	Methylase of chemotaxis methyl-accepting proteins
20230	COG630	NU	Type IV secretory pathway, VirB11 components, and related ATPases involved in archaeal flagella biosynthesis
20240	COG2874	NU	Predicted ATPases involved in biogenesis of archaeal flagella
20250-60	COG3354	N	Putative archaeal flagellar protein G
20320	COG1681	N	Archaeal flagellins
20360	COG1776	NT	Chemotaxis protein CheC, inhibitor of MCP methylation
20370	COG1871	NT	Chemotaxis protein; stimulates methylation of MCP proteins
22850	COG624	E	Acetylornithine deacetylase/Succinyl-diaminopimelate desuccinylase, related deacylases
25900			GO:0006508, GO:0004252
26500			GO:0006355
26770			
27170			
28450	COG108	H	3,4-dihydroxy-2-butanone 4-phosphate synthase
28460	COG1339	KH	Transcriptional regulator of a riboflavin/FAD biosynthetic operon
29490			
30220	COG840	NT	Methyl-accepting chemotaxis protein
30910	COG840	NT	Methyl-accepting chemotaxis protein
32480	COG277	C	FAD/FMN-containing dehydrogenases
35410	COG2358	R	TRAP-type uncharacterized transport system, periplasmic component
35420	COG1414	K	Transcriptional regulator
35560-70	COG3325	G	Chitinase
35580-90	COG3979	R	Uncharacterized protein contain chitin-binding domain type 3
36390	COG683	E	ABC-type branched-chain amino acid transport systems, periplasmic component
37530			GO:0006950
37580	COG747	E	ABC-type dipeptide transport system, periplasmic component
37600			GO:0005488
37630			GO:0008152, GO:0008080
39920	COG155	P	Sulfite reductase, beta subunit (hemoprotein)
39930	COG1035	C	Coenzyme F420-reducing hydrogenase, beta subunit

Cluster 14			
Name (Hbor)	COG	Family	Annotation
05290			
05800			
06190			GO:0006118, GO:0051539, GO:0010181, GO:0016651, GO:0051287
14220-30	COG395	G	ABC-type sugar transport system, permease component
15730	COG1622	C	Heme/copper-type cytochrome/quinol oxidases, subunit 2
15740	COG843	C	Heme/copper-type cytochrome/quinol oxidases, subunit 1
15750			
26440	COG296	G	1,4-alpha-glucan branching enzyme
26450			GO:0003824
26460	COG3839	G	ABC-type sugar transport systems, ATPase components
26470	COG2182	G	Maltose-binding periplasmic proteins/domains
26480-90	COG1175	G	ABC-type sugar transport systems, permease components
27190			
28780	COG784	T	FOG: CheY-like receiver
33670			
37540-50	COG444	EP	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component
37560-70	COG1173	EP	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components

Cluster 15			
Name (Hbor)	COG	Family	Annotation
00110	COG3398	S	Uncharacterized protein conserved in archaea
01220			
02340			
02930			GO:0006950
07150			
08720	COG1612	O	Uncharacterized protein required for cytochrome oxidase assembly
11850	COG2132	Q	Putative multicopper oxidases
11870	COG3794	C	Plastocyanin
14080			
14090	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins
16730			
17810			
18160	COG843	C	Heme/copper-type cytochrome/quinol oxidases, subunit 1
20440	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins

21120	COG2346	R	Truncated hemoglobins
22620	COG444	EP	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component
26300	COG3478	R	Predicted nucleic-acid-binding protein containing a Zn-ribbon domain
26750	COG2132	Q	Putative multicopper oxidases
26760	COG3794	C	Plastocyanin
27330	COG1271	C	Cytochrome bd-type quinol oxidase, subunit 1
27340	COG1294	C	Cytochrome bd-type quinol oxidase, subunit 2
29460	COG2008	E	Threonine aldolase
29470			GO:0016020
29860			
30350	COG4716	S	Myosin-crossreactive antigen
30640	COG517	R	FOG: CBS domain
30660	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins
30670			GO:0006355, GO:0005622, GO:0003700
30700			
30750	COG517	R	FOG: CBS domain
30790	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins
30810	COG2217	P	Cation transport ATPase
30820	COG2836	S	Uncharacterized conserved protein
30830-40			
30850			GO:0006118, GO:0016020, GO:0004129, GO:0009482, GO:0005507, GO:0009483, GO:0009485, GO:0009481 1.9.3.1
30860	COG843	C	Heme/copper-type cytochrome/quinol oxidases, subunit 1
30880-90			
30900	COG398	S	Uncharacterized conserved protein
31370			
31530			
31700			
32160	COG1048	C	Aconitase A
32290	COG589	T	Universal stress protein UspA and related nucleotide-binding proteins
34730	COG208	F	Ribonucleotide reductase, beta subunit
34740	COG209	F	Ribonucleotide reductase, alpha subunit
34940			TIGR04031 rSAM-partnered protein, Htur_1727 family
34950	COG535	R	Predicted Fe-S oxidoreductases
34960	COG4309	S	Uncharacterized conserved protein
34980			
35000			
35010	COG723	C	Rieske Fe-S protein
35020	COG1290	C	Cytochrome b subunit of the bc complex

35030			
35040	COG5013	C	Nitrate reductase alpha subunit
35050	COG1140	C	Nitrate reductase beta subunit
35060			TIGR03477 DMSO reductase family type II enzyme, heme b subunit Energy metabolism:Electron transport
35070	COG3381	R	Uncharacterized component of anaerobic dehydrogenases
35080			
35090			GO:0005488
35100	COG489	D	ATPases involved in chromosome partitioning
35910			
37720	COG451	MG	Nucleoside-diphosphate-sugar epimerases
39290			

Cluster 16			
Name (Hbor)	COG	Family	Annotation
18570			hypothetical protein

Appendix 6.1. Iterative NB emoji-hashtag set development.

Four sets of iteratively developed emojis were centered in the surveys described in this work:

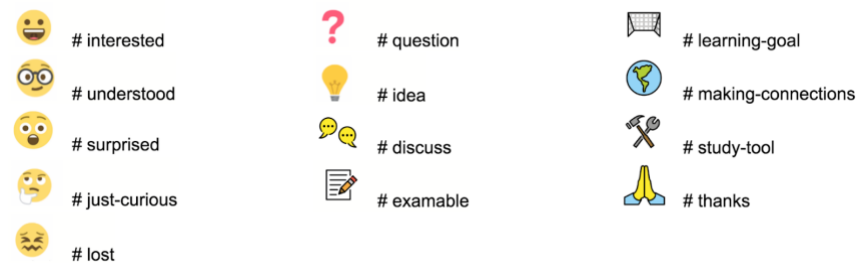
Survey 1 (Winter 2020)



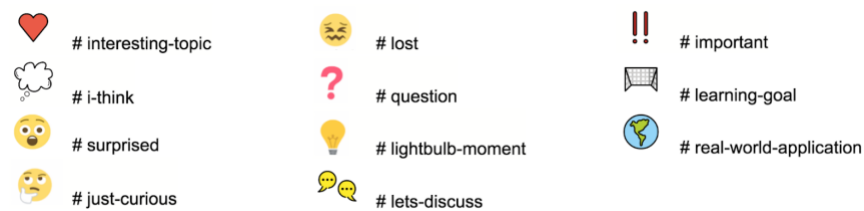
Survey 2 (Spring 2020)



Survey 3 (Summer Session 2020)



Survey 4 (Winter 2021)



Appendix 6.2. Copy of Survey 1 – Winter Quarter 2020 (W20).

INSTRUCTIONS

The purpose of this survey is to evaluate your experience using emoji when tagging comments in Nota Bene. The emoji and tags help an instructor understand the sentiment of students as they read and process information from the text. Your responses will be used to improve the commenting and tagging process.

* Please reflect on your emoji usage and answer the following questions to the best of your ability.

* Please refrain from referring to the NB interface or your previous comments to answer these questions -- we would like to better understand your working perception as is.

Please enter your first name in the box below, as it appears on your student ID.

Please enter your last name in the box below, as it appears on your student ID.

Please enter your student ID # in the box below, as it appears on your student ID.

Please answer the following questions regarding the emoji below:



What feelings or impressions does the emoji convey to you?

Select one or more terms from the list below. Feel free to use the same term for different emoji if needed.

- ☐ question
- ☐ help
- ☐ interested
- ☐ flustered
- ☐ frustrated

- ☐ idea
- ☐ curious
- ☐ useful
- ☐ confused
- ☐ Other:

How useful is this emoji in terms of your personal NB usage?

- ☐ Very useful
- ☐ Somewhat useful
- ☐ Not useful

In some of your posts using Nota Bene, you should have included an emoji/hashtag. As a reminder, these are your current options of emojis/hashtags to select from:



#curious



#confused



#useful



#interested



#frustrated



#help



#question



#idea

NOTE: Your answers to the previous set of questions need not align with the hashtags shown below the emojis presented in the thumbnails above.

When adding an emoji/hashtag to your post, how much does the emoji image or hashtag or both inform your choice of emoji to use?

	does not inform	moderately informs	heavily informs
The hashtag TEXT is...	☐	☐	☐
The PICTURE of the emoji...	☐	☐	☐

We would like to look at differences between the emojis used when students express a lack of understanding or uncertainty.

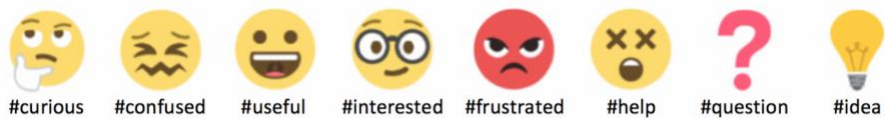
For the situations listed below, which of the following emoji do you think most accurately captures the description?

	#help	#confused	#question	#curious	#flustered
I am worried about knowing this information for an exam	☐	☐	☐	☐	☐
I am not sure what questions to ask to help me not feel lost	☐	☐	☐	☐	☐
I am seeking information	☐	☐	☐	☐	☐
I can't figure out the answer	☐	☐	☐	☐	☐
I am feeling lost	☐	☐	☐	☐	☐
I am aware that I do not know something	☐	☐	☐	☐	☐
I've encountered contradictory information	☐	☐	☐	☐	☐
I am seeking information because I would like to know more	☐	☐	☐	☐	☐

Confusion is one of the emoji that we are interested in exploring further. We know that it is a complex sentiment and there are many ways to be confused. How accurately do the statements below describe how you experience confusion?

	Not accurate	Somewhat accurate	Very accurate
I am aware that I do not know something	☐	☐	☐
I am not sure what questions to ask to help me not feel lost	☐	☐	☐
I am seeking information	☐	☐	☐
I am seeking information because I would like to know more	☐	☐	☐
I am feeling lost	☐	☐	☐
I've encountered contradictory information	☐	☐	☐
I am worried about knowing this information for an exam	☐	☐	☐
I can't figure out the answer	☐	☐	☐

For reference, here are your current set of options for emojis/hashtags.



Please describe any additional emoji that could be helpful in Nota Bene.

Would you be willing to be considered for participation in a short (10–15 minute) one-on-one interview conducted over Skype or in person? Small, monetary compensation will be provided.

- ☐ Yes
☐ No

Please provide any other comments regarding the Nota Bene emoji.

Appendix 6.3. Copy of Survey 2 – Spring Quarter 2020 (S20).

INSTRUCTIONS

The purpose of this survey is to evaluate your experience using emoji when tagging comments in Nota Bene. The emoji and tags help an instructor understand the sentiment of students as they read and process information from the text. This important, real-time feedback helps guide instruction and course improvement. Your responses will be used to improve NB and increase the quality of information we can use for course improvement and teaching.

- Please reflect on your emoji usage and answer the following questions to the best of your ability.
- Please refrain from referring to the NB interface or your previous comments to answer these questions -- we would like to better understand your working perception as is.

Please enter your first name in the box below, as it appears on your student ID.

Please enter your last name in the box below, as it appears on your student ID.

Please enter your student ID # in the box below, as it appears on your student ID. **PLEASE DOUBLE CHECK YOUR ID # AND MAKE SURE IT IS INPUTTED CORRECTLY -- YOU WILL NOT GET CREDIT IF WE CANNOT MATCH YOUR STUDENT ID WITH THE GRADEBOOK.**

In previous terms, we asked students about the types of sentiments they would like to express using emojis in Nota bene.

Using this feedback, we have created a modified set of emoji and their corresponding affects. In this survey, we invite you to validate and provide feedback otherwise on the modified emoji set.

You will be shown 15 emojis and will be asked to answer a few questions about each of the emojis and their corresponding affect, with just a few questions to answer following that.

Please evaluate the following emoji:



PART A: To me, the text (affect/sentiment) that this emoji represents is...

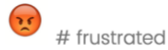
(describe the sentiment/affect that comes to your mind below)

PART B: Would you prefer to have the option to express **[text you wrote in PART A]** in NB?

☐ Yes, I would like the option to express [text you wrote in PART A] when I use NB.

☐ No, I don't feel the need to express [text you wrote in PART A] when I use NB.

PART C: We propose the following corresponding affect text for the emoji above:



Please evaluate the proposed emoji-text pair above and indicate your answer choice for the following statements on the spectrum below.

The proposed text does not appropriately describe the paired emoji.	☐ ☐ ☐	The proposed text appropriately describes the paired emoji.
I do not want to express this proposed hashtag in NB.	☐ ☐ ☐	I would like to express this proposed hashtag in NB.
I would rarely use this emoji/text pair in future NB settings.	☐ ☐ ☐	I would frequently use this emoji/text pair in future NB settings.

PART D: Do you have any other ideas to share for hashtag text that might better describe the emoji? Do you have any other comments to provide regarding the proposed emoji + text combinations?

Here is the proposed set of emojis for reference:

 # engaged	 # frustrated	 # texterror
 # learning	 # question	 # pleaseclarifyinclass
 # whoa	 # explanation	 # willthisbeontheexam
 # justcurious	 # learning goal	 # requestTAattention
 # flustered	 # connection	 # disagree

Are there any other HASHTAGS missing from the above set that you think could be helpful to include in Nota Bene? Please describe below or write "none" otherwise.

Are there any other EMOJIS missing from the above set that you think could be helpful to include in Nota Bene? Please describe below or write "none" otherwise.

Please rate your predicted usage of each emoji from this quarter's NB emoji set on a scale of 1-10. Putting 1 means you rarely used the emoji and putting 10 means you most frequently used the emoji -- please try to use the whole scale. Mark 0 if you never used the emoji.

No need to be precise or look at your post history, just try to answer based what comes to mind when you reflect on your usage.

For reference, here are your current set of options for emojis/hashtags:



Emoji Usage (0 = zero usage; 1 = very low usage; 10 = very frequent usage)

0 1 2 3 4 5 6 7 8 9 10

curious



Please share any other thoughts you have surrounding (current set of emojis) vs. (new set of emojis) as well as any other comments regarding the Nota Bene emoji. (OPTIONAL)

Thank you for providing your honest and thoughtful feedback! Would you be willing to be considered for participation in a short (10-15 minute) one-on-one interview conducted over Zoom? Small, monetary compensation will be provided.

☐ Yes

☐ No

Appendix 6.4. Copy of Survey 3 – Summer Session 1 2020 (SS20).

INSTRUCTIONS

The purpose of this survey is to evaluate your experience using emoji when tagging comments in Nota Bene. The emoji and tags help an instructor understand the sentiment of students as they read and process information from the text. This important, real-time feedback helps guide instruction and course improvement. Your responses will be used to improve NB and increase the quality of information we can use for course improvement and teaching.

- Please reflect on your emoji usage and answer the following questions to the best of your ability.
- Please refrain from referring to the NB interface or your previous comments to answer these questions -- we would like to better understand your working perception as is.

Please enter your first name in the box below, as it appears on your student ID.

Please enter your last name in the box below, as it appears on your student ID.

Please enter your student ID # in the box below, as it appears on your student ID. **PLEASE DOUBLE CHECK YOUR ID # AND MAKE SURE IT IS INPUTTED CORRECTLY -- YOU WILL NOT GET CREDIT IF WE CANNOT MATCH YOUR STUDENT ID WITH THE GRADEBOOK.**

In previous terms, we asked students about the types of sentiments they would like to express using emojis in Nota bene. Using this feedback, we have created a modified set of emoji and their corresponding affects. In this survey, we invite you to validate and provide feedback otherwise on the modified emoji set.

You will be shown 13 emojis and will be asked to answer a few questions about each of the emojis and their corresponding affect, with a short set of questions to answer following that.

Please answer the following questions regarding the emoji below:



What feelings or impressions does the emoji convey to you?

Select one or more terms from the list below. Feel free to use the same term for different emoji if needed.

- | | | |
|---------------------------------------|--|--|
| ☐ understood | ☐ discuss | ☐ making-connections |
| ☐ surprised | ☐ idea | ☐ study-tool |
| ☐ lost | ☐ question | ☐ thanks |
| ☐ just-curious | ☐ examable | ☐ Other: <input type="text"/> |
| ☐ interested | ☐ learning-goal | |

How useful is this emoji in terms of your personal NB usage?

- ☐ Very useful
☐ Somewhat useful
☐ Not useful

Would you like to express this emoji in NB?

- ☐ Yes
☐ Neutral
☐ No

Would you frequently use this emoji in future NB settings?

- ☐ Yes (frequent usage)
☐ Somewhat (occasional usage)
☐ No (infrequent usage)

When adding an emoji/hashtag to your post, how much does the emoji image or hashtag or both inform your choice of emoji to use?

- | | does not inform | moderately informs | heavily informs |
|---------------------|-----------------------|-----------------------|-----------------------|
| The hashtag TEXT... | ☐ | ☐ | ☐ |
| The emoji IMAGE... | ☐ | ☐ | ☐ |

We use this statement to discard the survey of people who are not reading the questions. Please select agree (not strongly agree) for this question to preserve your answers.

- ☐ Strongly Agree
- ☐ Agree
- ☐ Neutral
- ☐ Disagree
- ☐ Strongly Disagree

When reflecting on your ***personal*** NB habits, how useful is NB for facilitating each of the following? Drag and drop to the appropriate box.

Items	Very useful to me	Moderately useful to me	Not useful to me at all	NA / no opinion
Motivating me to review the course content before lectures				
Encouraging note-taking while reading				
Motivating me to reflect while reading				
Allowing me to share questions/ideas with classmates				
Allowing me to share questions/ideas with instructors				
Allowing me to have classwide interactions directly within the context of the course textbook				
Exposing me to different viewpoints or perspectives regarding course content				
Helping me "keep up" with the class pace				
Providing information to instructors who use it as a tool to understand student learning				

For reference, here are your latest set of options for emojis/hashtags.

	# interested		# question		# learning-goal
	# understood		# idea		# making-connections
	# surprised		# discuss		# study-tool
	# just-curious		# examable		# thanks
	# lost				

Are there any other **HASHTAGS** missing from the above set that you think could be helpful to include in Nota Bene? Please describe below or type "none" otherwise.

Are there any other **EMOJIS** missing from the above set that you think could be helpful to include in Nota Bene? Please describe below or type "none" otherwise.

Appendix 6.5. Copy of Survey 4 – Winter Quarter 2021 (W21).

2021 Bis2A Winter Extra Credit Survey

INTRODUCTION AND INSTRUCTIONS for PART 1 of 2

Greetings and thank you for taking this survey. Currently, a UC Davis teaching team is working alongside collaborators towards improving Nota bene. Presently, we are trying to evaluate the set of **emoji and hashtags** available for use in NB. Emoji and hashtags used in student comments can help an instructor understand the sentiment of students as they read and process information from the text. This important, real-time feedback can help guide instruction and course improvement.

In Part 1 of this survey, you will be solicited for your opinions regarding NB's emoji-hashtag set (11 emoji-hashtag in total). Your invaluable feedback will be used to evaluate and improve the annotation tool and increase the quality of information instructors can use for course improvement and teaching.

No prior knowledge is needed; please answer the survey questions to the best of your ability.

Please enter your first name in the box below, as it appears on your student ID.

Please enter your last name in the box below, as it appears on your student ID.

Please enter your student ID # in the box below, as it appears on your student ID. **PLEASE DOUBLE CHECK YOUR ID # AND MAKE SURE IT IS INPUTTED CORRECTLY -- YOU WILL NOT GET CREDIT IF WE CANNOT MATCH YOUR STUDENT ID WITH THE GRADEBOOK.**

Please answer the following questions regarding the emoji below:



What feelings or impressions does the emoji convey to you?

Select one or more terms from the list below. Feel free to use the same term for different emoji if needed.

- ☐ surprised
- ☐ learning-goal
- ☐ interesting-topic
- ☐ question
- ☐ important
- ☐ just-curious

- ☐ i-think
- ☐ lets-discuss
- ☐ lost
- ☐ real-world-application
- ☐ lightbulb-moment
- ☐ Other:

Please assess the appropriateness of the fit/pairing between each emoji and their corresponding hashtag.

	strong no (inappropriate)	no	yes	strong yes (appropriate)
 i-think	☐	☐	☐	☐
 just-curious	☐	☐	☐	☐
 surprised	☐	☐	☐	☐
 learning-goal	☐	☐	☐	☐
 lets-discuss	☐	☐	☐	☐
 question	☐	☐	☐	☐
 real-world application	☐	☐	☐	☐
 lost	☐	☐	☐	☐
 interesting-topic	☐	☐	☐	☐
 lightbulb-moment	☐	☐	☐	☐
 important	☐	☐	☐	☐

Which of the emoji-hashtag are the most/least important to be able to communicate to your instructor and peers?

Please rank the emoji-hashtag from most (1) to least important (11).

For reference, here is your current set of available emojis and hashtags.

	# interesting-topic		# question		# learning-goal
	# i-think		# lightbulb-moment		# real-world-application
	# surprised		# lets-discuss		
	# just-curious		# important		
	# lost				

How do you feel about the number of emoji-hashtags in the set?

- ☐ There are too many emoji
- ☐ The amount of emoji are just right
- ☐ There are too few emoji

How can we improve our emoji set? Could we make modifications to the wording of the hashtag? Could we find a new emoji for the hashtag text? Could we remove or add any new emoji-hashtag?

Please provide any other comments regarding the Nota Bene emoji or how you use NB.

Appendix 6.6. Light Kappa score simulations for select-all-that-apply questions in Surveys 1, 3, and 4.

One factor that contributes to the Light Kappa score is the number of available terms to select in the select-all-that-apply survey questions in Surveys 1, 3, and 4. We sought to understand how differences in the number of available terms (10 vs. 12 or 14 available terms) and number of terms selected (Scenarios A-M) would affect the Light Kappa score using simulations of different scenarios of (dis)agreement to validate desirable baseline Kappa values. The average spread across all scenarios between a 10-term and 14-term select-all-that-apply Light Kappa score for 12 hypothetical simulation scenarios is 0.048.

Light Kappa - 10 terms	Light Kappa - 12 terms	Light Kappa - 14 terms	Descriptions of Scenarios
1.00	1.00	1.00	Scenario A: complete agreement, no matter the term
0.74	0.75	0.76	Scenario F: agreement, one person chose 2 terms and one person chose 3 terms (extra term)
0.62	0.63	0.63	Scenario B: agreement, one person chose 1 term and one chose 2 terms
0.41	0.43	0.44	Scenario C: agreement, one person chose 1 term and one chose 3 terms
0.38	0.40	0.42	Scenario E: two people partially agree, on 1 of 2 selections - they each chose 2 terms
0.29	0.31	0.32	Scenario D: agreement, but one person chose 1 term and one chose 4 terms
0.21	0.25	0.28	Scenario G: partial agreement on 1 of 2 terms, one person chose 2 terms and one person chose 3 terms
-0.11	-0.09	-0.08	Scenario H: disagreement, each person selected 1 term; kappa score doesn't tell you which
-0.15	-0.13	-0.11	Scenario I: disagreement, one person chose 1 term and the other person chose 2 terms that don't match
-0.18	-0.14	-0.12	Scenario J: disagreement, one person chose 1 term and the other person chose 3 terms that don't match
-0.19	-0.15	-0.13	Scenario K: disagreement, one person chose 1 term and the other person chose 4 terms that don't match
-0.25	-0.20	-0.17	Scenario L: disagreement, each chose 2 terms
-0.32	-0.25	-0.21	Scenario M: disagreement, one person chose 2 terms and one person chose 3 terms

Survey 1

Survey 4

Survey 3

Appendix 6.7. Distribution of the number of terms selected (singular vs. multiple responses) for each emoji in the select-all-that-apply question of Survey 1.

The percentage of students who selected 1, 2, 3, or 4 or more terms for each presented emoji in the select-all-that-apply question featured in Survey 1 (n=648 students).



Appendix 6.8. Emoji utility for the original emoji in Survey 1.

The percentage of students who indicated that the emoji presented was “very useful” or “somewhat useful,” as opposed to “not useful,” in response to the question: “How useful is this emoji in terms of your personal NB usage?” (n=648 students).

Emoji-hashtag Pairing		% of students who found the emoji somewhat or very useful
	question	0.96
	idea	0.92
	curious	0.88
	interested	0.87
	useful	0.78
	confused	0.71
	help	0.61
	frustrated	0.48

Appendix 6.9. Students' experience of confusion and emoji-hashtags deemed appropriate by students for various instances of confusion in Survey 1.

Student response distribution for the Survey 1 question: "How accurately do the statements below describe how you experience confusion?" (n=648 students).

	I am aware that I do not know something	I am seeking information	I am worried about knowing this information for an exam	I am seeking information because I would like to know more	I am feeling lost	I am not sure what questions to ask to help me not feel lost	I can't figure out the answer	I've encountered contradictory information
not accurate	0.06	0.10	0.18	0.21	0.05	0.12	0.05	0.20
somewhat accurate	0.35	0.37	0.39	0.42	0.29	0.40	0.35	0.43
very accurate	0.59	0.53	0.43	0.36	0.66	0.48	0.60	0.38
accurate sum	0.94	0.90	0.82	0.79	0.95	0.88	0.95	0.80

Student response distribution for the Survey 1 question: "For the situations listed below, which of the following emoji do you think most adequately captures the description?" (n=648 students).

	confused	curious	flustered	help	question
I can't figure out the answer	0.20	0.01	0.10	0.52	0.17
I am aware that I do not know something	0.23	0.07	0.08	0.34	0.28
I am feeling lost	0.37	0.01	0.14	0.45	0.03
I am not sure what questions to ask to help me not feel lost	0.37	0.01	0.21	0.32	0.08
I've encountered contradictory information	0.64	0.05	0.09	0.09	0.13
I am seeking information	0.04	0.33	0.00	0.22	0.40
I am seeking information because I would like to know more	0.03	0.72	0.00	0.10	0.15
I am worried about knowing this information for an exam	0.09	0.08	0.23	0.37	0.22

Appendix 6.10. The extent to which decisions for selecting tags to include posts are informed by emoji picture or hashtag text in Surveys 1 and 3.

Student response distribution for the Survey 1, 3, and 4 question: “When adding an emoji/hashtag to your post, how much does the emoji image or hashtag or both inform your choice of emoji to use?” In Survey 1, the largest proportion of students (44%, n=648 students) indicated that both the hashtag text and the emoji picture moderately inform their tag selection, whereas the largest proportion of students in Survey 3 (41%, n=648 students) stated that the hashtag text heavily informed, and the emoji picture moderately informed their tag selection. The proportion of students who responded that the hashtag text heavily informed their decision increased from 35% to 74% between Survey 1 to Survey 3. On the other hand, for the emoji picture, this same percentage only increased from 23% to 34%, suggesting that the hashtag text is a more important factor for students’ tag selections.

Survey 1 (n=648)		does not inform	moderately informs	heavily informs	emoji picture
		0.14	0.64	0.23	sums
does not inform	0.05	0.03	0.01	0.01	
moderately informs	0.60	0.03	0.44	0.13	
heavily informs	0.35	0.08	0.18	0.08	
hashtag text	sums				

Survey 3 (n=175)		does not inform	moderately informs	heavily informs	emoji picture
		0.13	0.53	0.34	sums
does not inform	0.00	0.00	0.00	0.00	
moderately informs	0.26	0.01	0.13	0.13	
heavily informs	0.74	0.12	0.41	0.21	
hashtag text	sums				

Appendix 6.11. Rationale for emoji-hashtag modifications throughout survey iterations.

Survey 1 (WQ20)	Survey 2 (SQ20)	Survey 3 (SS20)	Survey 4 (WQ21)
 curious	 just-curious Changed to make separations between the #question, #confused, and #curious emoji functions less ambiguous; Added “just” to emphasize the curiosity aspect.	 just-curious Unchanged.	 just-curious Unchanged.
 confused	 flustered Part of the strategy to manage separation between #question, #confused, and #curious sentiments; Renamed to make the hashtag affect represent the emoji better.	 lost In open ended feedback, students expressed wanting to see a confusion sentiment; We still want to separate between #question, #confused, and #curious sentiments; “lost” is meant to serve as an emoji for students who experience this specific type of confusion.	 lost Unchanged.
 useful	 engaged #useful did not have a strong emoji-hashtag association; We thought engaged might capture both #useful and #interested and fit the face well.	 interested In Survey 2, most students wrote “happy,” then “understand,” “interested,” “useful.”; We did not feel that “happy” was going to be used often; and we thought the glasses face emoji was better for #understood, so we went with #interested for this emoji.	 interesting-topic It is a consistent trend across the surveys that the smiley face emoji evoked many different affects (i.e., is ambiguous); We believe we can venture away from the smiley face and use a heart to represent topics that students enjoy.
 interested	 learning Many students put “nerd” or “smart” in open response portions. We thought learning might fit better and association might be tighter.	 understood Students wrote in “interested,” “nerd,” “smartypants,” “understood/understanding.”; We thought that this face represented #understood well.	 i-think With #interesting-topic already in place, and students expressing liking the option to express #idea, this emoji serves as an option to express thinking.

 frustrated	 frustrated Despite strong associations between the emoji and hashtag, students expressed that this emoji was too extreme and it had low usage.	Removed – infrequent usage.	
 help	 whoa Students expressed that this emoji was extreme. We tried to repurpose it to a shocked type of affect, which was a common sentiment found in NB comments historically. This emoji was also noisy for the #confusion, #question, #curious set of affects.	 surprised The survey results indicated that the more immediate thought/affect that arises is surprise/shock, so the hashtag was changed, and the emoji slightly modified to be less ambiguous.	 surprised Unchanged.
 question	 question Unchanged.	 question Unchanged.	 question Unchanged.
 idea	 explanation Even though #idea had a strong emoji-hashtag association, its function or purpose was unclear; changed to #explanation to clarify its purpose.	 idea The majority of students expressed that #idea had a better fit with the lightbulb than #explanation. Thus, we reverted back to #idea.	 lightbulb-moment Suggested by the NB team to see where in the text this phenomenon was happening. #idea was converted to #i-think
	 learning-goal This emoji came from discussions with the NB team to identify specific parts of the text where students make connections with learning goals.	 learning-goal Students associated the target with correctness or accuracy; the goal emoji is a second attempt at assigning an emoji to the hashtag.	 learning-goal Unchanged.

	 connection Students often made external connections in their comments and also seem to be very engaged with concepts that can be applied to the “real world.”	 making-connections Changed to be more clear in function.	 real-life-application Changing the wording to real-life-application might be more clear and more exciting, and perhaps more familiar.
	 texterror Added to provide a way for students to address errors in the text	 study-tool Ideally, the text does not actually contain that many errors and text errors can be identified by other means. Changed to #study-tool, in light of the fact that many student comments contain study tools/tricks, the categorization of which may be useful for students.	Removed – Study tool had good association but scored lower on usefulness. This function could be captured in part by #lightbulb-moment or #important.
	 pleaseclarifyinclass The NB team sought to provide functional tools that make NB more compelling to use for students.	Removed – Emoji too ambiguous.	
	 willthisbeontheexam The NB team sought to provide functional tools that make NB more compelling to use for students. This emoji-hashtag was suggested by students in open ended feedback as well.	 examable Changed to be more succinct.	 important The emoji-hashtag association for #examable was poor. Also, in multiple surveys' open ended feedback, students expressed wanting an exclamation for important. We thought #examable and #learning-goal covered “important” but #examable may be too unfamiliar and #important may be preferred.

	 requestTAattention The NB team sought to provide functional tools that make NB more compelling to use for students.	Removed – This emoji was not viable because TA involvement was variable between terms.	
	 disagree The NB team sought to provide a way for students to express disagreement.	Removed – Students said that this emoji looked more like “#stop” or “raise hand emoji” and some students said that they would feel too combative expressing disagreement this way.	
		 discuss The NB team decided to try to encourage discussions within NB in light of removing #requestTAattention and #pleaseclarifyinclass, allowing students to call for discussion open to other students as well.	 lets-discuss Added “lets” to make the emoji sound more friendly and inviting.
		 thanks Suggested by many students.	Removed – Replaced by a thumbs up button feature. Many students also stated that the emoji looked like #please, in addition to #thanks.

Appendix 6.12. Student-generated terms with over 100 counts for emojis in Survey 2.

In Survey 2, students were presented with emoji without an accompanying instructor-generated tag and were asked to input their own text. The list below shows the terms that were suggested by at least 100 students for each emoji (n=799 students).

Instructor-generated pairings, for reference:

	# engaged		# frustrated		# texterror
	# learning		# question		# pleaseclarifyinclass
	# whoa		# explanation		# willthisbeontheexam
	# justcurious		# learning goal		# requestTAattention
	# flustered		# connection		# disagree

Student-generated pairings with counts of over 100:

	happy (378 student responses), understood (168)
	smart (156), interested (139), nerd (137), learning (114)
	shock (153), dead/dying (143), confused (132)
	curious (329), thinking (269), confused (121)
	confused (281), frustrated (139)
	frustrated (329), angry (318)
	question (720), confused (137)
	idea (716)
	target (286)
	world (300), globe (144), connect (136), earth (100)
	work (246), fix (226), tool (166)
	professor/teacher (297), apple (208)
	note (498), important (168), writing (157)
	attention (253), help (193), bell (101)
	stop (301), question (226), hand (raise) (162)

Appendix 6.13. Student responses for questions pertaining to the appropriateness of emoji-hashtag pairing fit, desire to express the tag, and predicted frequency of expression of each emoji in Survey 2.

The table below shows the percentage of students who indicated a “3” for the highest level of agreement or affirmation on a 3-point Likert scale, subtracted by the percentage of students who indicated a “1” for the lowest level of agreement for each of the listed statements in Survey 2 (n=799 students).

Emoji-hashtag pair	The proposed text appropriately describes the paired emoji.	I would like to express this proposed hashtag in NB.	I would frequently use this emoji/text pair in future NB settings.
 # question	0.97	0.94	0.90
 # willthisbeontheexam	0.41	0.74	0.60
 # justcurious	0.74	0.73	0.66
 # learning goal	0.48	0.49	0.30
 # requestTAattention	0.59	0.43	0.17
 # explanation	-0.09	0.41	0.31
 # pleaseclarifyinclass	-0.25	0.36	0.17
 # connection	0.19	0.32	0.12
 # frustrated	0.72	0.30	0.03
 # learning	0.47	0.26	0.11
 # flustered	0.51	0.24	0.09
 # texterror	0.29	0.22	-0.13
 # engaged	0.35	0.21	0.07
 # whoa	0.14	0.19	0.04
 # disagree	-0.23	-0.04	-0.33

Appendix 6.14. Student perceived usage frequency of their current (the original) emoji set in Survey 2.

The average frequency ranking provided by students for the Survey 2 question, “Please rate your predicted usage of each emoji from this quarter’s NB emoji set on a scale of 1-10,” (1 representing the lowest usage, and 10 representing the highest usage) along with the percentage of survey respondents who indicated not using the tag at all (n=799 students).

Emoji-hashtag pair		Average frequency ranking (0: did not use, 1: lowest, 10: highest)	% of survey respondents who did not use the emoji
	#interested	6.6	0.01
	#useful	6.3	0.02
	#question	6.2	0.03
	#curious	6.1	0.01
	#idea	5.5	0.05
	#confused	4.6	0.08
	#help	3.0	0.21
	#frustrated	1.9	0.31

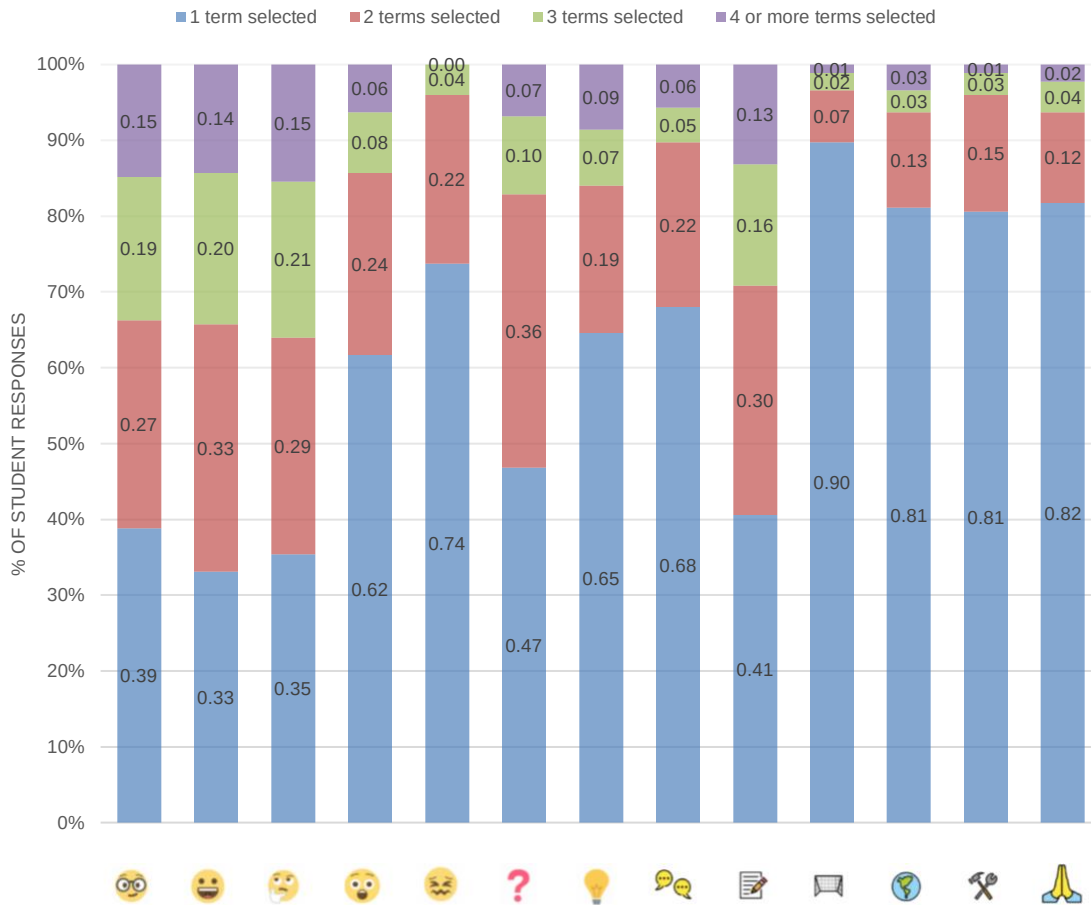
Appendix 6.15. Select-all-that-apply question data from Survey 3.

Students were shown each of the emoji in the iteratively modified NB emoji set, one-by-one in randomized order, and were asked to choose one or more terms conveyed by the emoji. This table shows the proportion of students who selected each of the terms for all emoji (n=175 students).

													
	#under- stood	#interested	#just- curious	#surprised	#lost	#question	#idea	#discuss	#examable	#learning- goal	#making- connection	#study- tool	#thanks
understood	0.51	0.60	0.01	0.06	0.00	0.01	0.10	0.01	0.04	0.07	0.02	0.01	0.10
interested	0.59	0.58	0.31	0.18	0.01	0.02	0.07	0.03	0.03	0.01	0.02	0.02	0.02
just-curious	0.32	0.05	0.85	0.05	0.01	0.22	0.03	0.04	0.02	0.02	0.05	0.00	0.01
surprised	0.03	0.24	0.03	0.97	0.02	0.01	0.06	0.01	0.01	0.01	0.01	0.01	0.01
lost	0.02	0.00	0.15	0.09	0.90	0.39	0.01	0.01	0.01	0.02	0.03	0.00	0.05
question	0.01	0.00	0.32	0.02	0.04	0.98	0.03	0.08	0.05	0.01	0.02	0.03	0.02
idea	0.22	0.14	0.22	0.09	0.00	0.06	0.97	0.06	0.11	0.02	0.04	0.03	0.01
discuss	0.05	0.02	0.10	0.01	0.01	0.05	0.02	0.90	0.13	0.01	0.04	0.03	0.02
examable	0.06	0.02	0.05	0.00	0.01	0.01	0.02	0.02	0.54	0.03	0.02	0.05	0.01
learning- goal	0.06	0.03	0.02	0.02	0.00	0.02	0.10	0.01	0.33	0.80	0.08	0.05	0.01
making- connections	0.14	0.11	0.08	0.09	0.01	0.01	0.15	0.25	0.22	0.05	0.81	0.08	0.05
study-tool	0.10	0.04	0.01	0.00	0.00	0.01	0.06	0.05	0.54	0.04	0.03	0.89	0.02
thanks	0.06	0.33	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.01	0.84
other	0.03	0.06	0.11	0.03	0.29	0.02	0.02	0.03	0.07	0.09	0.15	0.06	0.11
Light Kappa Score	0.23	0.28	0.33	0.54	0.62	0.57	0.52	0.51	0.25	0.49	0.44	0.3	0.5

Appendix 6.16. Distribution of the number of terms selected (singular vs. multiple responses) for each emoji in the select-all-that-apply question of Survey 3.

The percentage of students who selected 1, 2, 3, or 4 or more terms for each presented emoji in the select-all-that-apply question featured in Survey 3 (n=175 students).



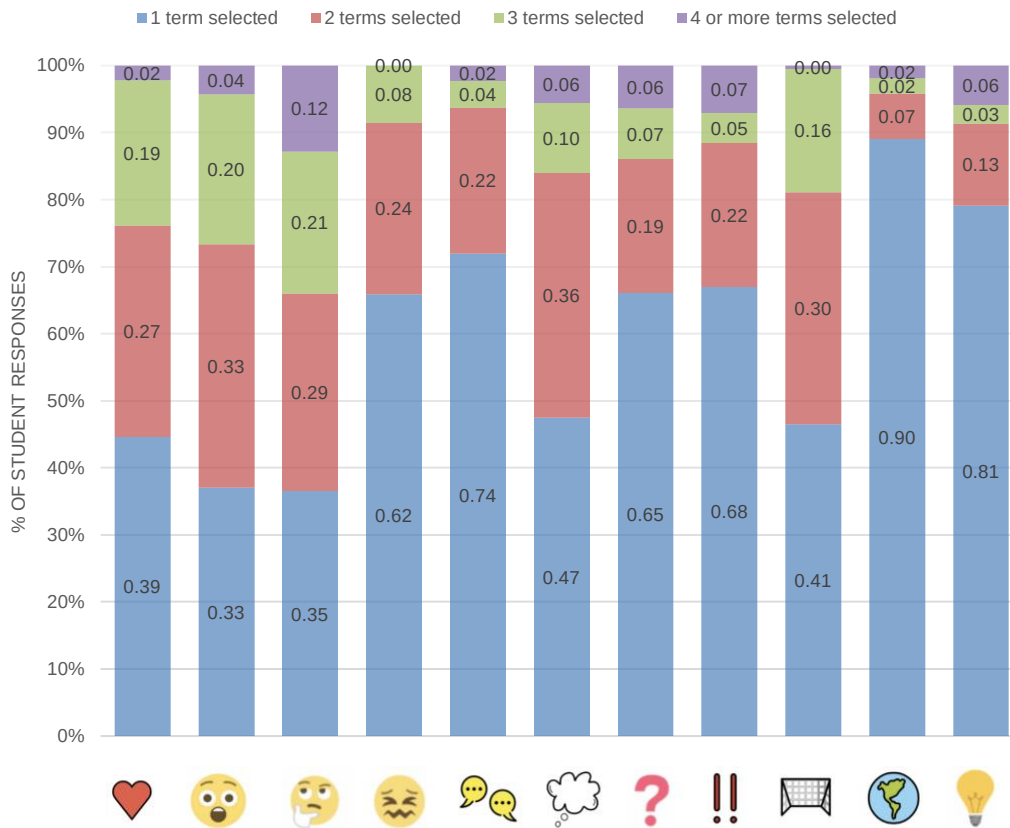
Appendix 6.17. Distribution of student responses for questions pertaining to the appropriateness of emoji-hashtag pairing fit, desire to express the tag, and predicted frequency of expression for each emoji in Survey 3.

The table below shows the distribution of student responses to each of the listed questions for each emoji presented in Survey 3 (n=175 students).

Emoji-hashtag pairing	How useful is this emoji in terms of your personal NB usage?			Would you like to express this emoji in NB?			Would you frequently use this emoji in future NB settings?		
	Not useful	Some-what useful	Very useful	No	Neutral	Yes	No (infrequent usage)	Some-what (occasional usage)	Yes (frequent usage)
 # understood	0.06	0.38	0.57	0.06	0.22	0.71	0.09	0.34	0.57
 # interested	0.07	0.43	0.49	0.04	0.30	0.66	0.11	0.37	0.53
 # just-curious	0.03	0.32	0.65	0.03	0.18	0.79	0.06	0.30	0.64
 # surprised	0.10	0.57	0.34	0.10	0.39	0.51	0.17	0.53	0.30
 # lost	0.12	0.42	0.46	0.10	0.30	0.59	0.25	0.39	0.37
 # question	0.01	0.18	0.81	0.01	0.12	0.87	0.05	0.17	0.78
 # idea	0.02	0.25	0.73	0.01	0.13	0.86	0.04	0.24	0.72
 # discuss	0.19	0.50	0.31	0.16	0.38	0.46	0.37	0.41	0.22
 # examable	0.20	0.46	0.34	0.18	0.41	0.42	0.32	0.42	0.26
 # learning-goal	0.36	0.44	0.20	0.41	0.34	0.25	0.55	0.29	0.16
 # making-connections	0.37	0.41	0.22	0.35	0.34	0.31	0.41	0.37	0.22
 # study-tool	0.22	0.49	0.29	0.18	0.42	0.41	0.33	0.46	0.21
 # thanks	0.30	0.46	0.25	0.24	0.35	0.41	0.42	0.34	0.23

Appendix 6.18. Distribution of the number of terms selected (singular vs. multiple responses) for each emoji in the select-all-that-apply question of Survey 4.

The percentage of students who selected 1, 2, 3, or 4 or more terms for each presented emoji in the select-all-that-apply question featured in Survey 4 (n=418 students).



Appendix 6.19. Appropriateness of the emoji-hashtag pairings in Survey 4.

The percentage of students who indicated the extent to which they agreed that the emojis and hashtags did or did not have an appropriate fit in Survey 4 (n=418 students).

Emoji-hashtag pairing		strong no (inappropriate)	no	yes	strong yes (appropriate)
	interesting-topic	0.06	0.21	0.42	0.31
	surprised	0.00	0.01	0.19	0.79
	just-curious	0.02	0.08	0.38	0.53
	lost	0.01	0.06	0.26	0.67
	lets-discuss	0.00	0.01	0.14	0.85
	i-think	0.01	0.06	0.30	0.63
	question	0.00	0.00	0.06	0.93
	important	0.01	0.02	0.16	0.82
	learning-goal	0.04	0.08	0.18	0.70
	real-world-application	0.00	0.01	0.09	0.90
	lightbulb-moment	0.00	0.00	0.07	0.92

Appendix 6.20. Student perceived usage frequency of their current (the original) emoji set in Survey 4.

Students in Survey 4 were asked to rank the 11 emoji-hashtag pairings in terms of most (1) to least (11) important to communicate to instructors and peers (n=418 students). The data has been summarized in the table below, using the original Borda counting method (Emerson, 2013):

Emoji-hashtag pairing	Borda count
 lets-discuss	9728
 lightbulb-moment	9540
 i-think	9200
 lost	9122
 important	8954
 just-curious	8906
 learning-goal	8324
 real-world-application	8164
 interesting-topic	6862
 question	6756
 surprised	6404

Appendix 6.21. Distribution of student responses regarding the quantity of emoji in Survey 4.

The percentage of students who indicated that there were too few emoji, the amount of emoji were just right, and there were too many emoji in Survey 4 (n=418 students).

Survey statement	Percent
There are too few emoji	0.07
The amount of emoji are just right	0.75
There are too many emoji	0.18