The Dissertation Committee for Cory David DuPai certifies that this is the approved version of the following dissertation:

Applications of large, heterogeneous datasets in understanding and treating pathogenic microbes

Committee:

Claus O. Wilke, Supervisor

Bryan W. Davies, Co-Supervisor

Jennifer A. Maynard

Jeffrey E. Barrick

William H. Press

Applications of large, heterogeneous datasets in understanding and treating pathogenic microbes

by

Cory David DuPai

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2020

Dedication

To Aditi, the storm in my calm, and our son, who will know me as doctor dad.

Acknowledgements

I thank both of my advisors, Claus Wilke and Bryan Davies, for their mentorship and guidance. Few graduate students are privileged to work with one world class adviser, yet I was lucky enough to have the counsel of two. Dr. Davies taught me, among many things, to think big first and worry about the finer details later. One of Dr. Wilke's core lessons was that it is often the finer details that truly matter. These perspectives, ambivalent as they are, have laid the foundation for my own research outlook.

I would further like to thank my colleagues in both the Wilke and Davies labs as it truly takes a village to earn a doctorate. In particular, thank you Adam Hockenberry for your advice and friendship, Tanvi Ingle for excelling despite my guidance, Benjamin Jack & Dariya Sydykova for being the older, cooler graduate students with (almost) all of the answers, and Xun Wang for encouraging me through your own curiosity.

I would also like thank all the teachers and supervisors whose mentorship has helped to make me the scientist I am today. In particular, thank you Patricia Husband & Magistra Baker for encouraging a big fish in a little pond to look for deeper waters, Karina Chmielewski for giving a naïve college student his first shot at bioinformatics, Nada Kalaany & Tamara Zaytouni for showing me the ins and outs of real scientific research, and Elizabeth Skippington for the chance to test my skills in the real world.

Lastly, I would like to thank my friends and family for their love and encouragement. Mom & dad, thank you for always being there and for giving me all the tools I needed to succeed. Lauren, Ariel, & Kelsey, I could not have asked for more loving, protective, and supportive big sisters and role models. Suhas & Aparna, you have shown me a lifetime of love and kindness in the ten years I have known you.

Abstract

Applications of large, heterogeneous datasets in understanding and treating pathogenic microbes

Cory David DuPai, Ph.D.

The University of Texas at Austin, 2020

Supervisor: Claus O. Wilke

Co-Supervisor: Bryan W. Davies

Major advances in a myriad of technologies over the past two decades have led to a remarkable increase in the generation of biological data. In response to this increase, researchers have developed methods to pool and analyze large, heterogeneous datasets for novel insights. Here I do just that, leveraging existing data to expand our understanding of therapeutic proteins and pathogenic microbes. In Chapter 2 I outline major shortcomings in existing viral annotation standards using metadata from all influenza A sequences submitted to the GISAID database between 2005 and 2018. I further establish updated nomenclature standards to improve annotation accuracy moving forward. In Chapter 3 I use published *Vibrio cholerae* sequencing data to derive a comprehensive gene coexpression network. This network provides direct insights into genes influencing pathogenicity, metabolism, and transcriptional regulation, further clarifies results from previous sequencing experiments in *V. cholerae*, and expands upon micro-array based findings in related gram-negative bacteria. In Chapter 4 I systematically probe all 49,000 unique beta hairpin substructures contained within the Protein Data Bank to uncover key characteristics correlated with stable beta hairpin structure, including amino acid biases and enriched inter-strand contacts. I also establish a set of broad design principles that can be applied to the generation of libraries encoding bioactive proteins. These findings highlight the untapped potential, promise, and power of pooled analyses using large, heterogeneous datasets.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Microbiology, like most scientific fields, has benefited immensely from the advent of the information age and ongoing advancements in technology, data storage, and computer processing power over the past twenty years. Alongside steady improvements in technologies like nuclear magnetic resonance (NMR) imaging [1–3], mass spectrometry [4,5], and various methods of biological microscopy [6–8], the rapid introduction of reliable, cost effective high-throughput sequencing platforms [9,10] has created a massive shift in the amount of data generated by and available to scientists.

1.1 THE -OMICS REVOLUTION

While it took a global team of scientists nearly fifteen years and three billion dollars to compile the first full human genome sequence in 2004 [11,12], the same feat can now be accomplished for under $3000 [13] in roughly one day of machine time [14]. Individual sequencing instruments are able to generate petabytes of data per annum [15], the equivalent of roughly a few gigabytes of data for a million bacterial samples per sequencing instrument per year. The unprecedented production and availability of high-quality sequencing and other biological data have proven a great boon for biologists across all disciplines.

In parallel with the massive growth in microbial -omics data generation, researchers have established global efforts to collect and catalog such data in a way that makes them practically accessible. Initiatives and databases such as the Global Initiative on Sharing All Influenza Data (GISAID) [16], the Sequence Read Archive (SRA) [17], the Protein Data Bank (PDB) [18], and GenBank [19] have aggregated hundreds of thousands of samples each. Platforms such as these not only archive records but also extend convenient data access to almost any researcher with an internet connection. This broad accessibility encourages transparency and reproducibility in published findings and provides further fuel for novel pooled analyses using an amalgam of heterogeneous datasets.

1.2 WHAT DO WE DO WITH ALL THESE DATA?

Sizable collections of all flavors of microbial -omics data have been successfully leveraged for a variety of research efforts. For instance, microbial genetic information encoded in Genbank and GISAID is particularly amenable to the large-scale phylogenetic analyses which form the backbone of research, surveillance, prevention, and treatment efforts for a wide array of pathogens. Complete genome sequences and the phylogenies constructed from them power research into microbes ranging from *Vibrio cholerae* [20] and *Enterobacteriaceae* [21] to influenza viruses [22] and even novel SARS-CoV-2 [23]. For community and microbiome analyses, high quality reference genomes, transcriptomes, and proteomes are essential for taxonomic assignment of sequences [24], gene cluster identification [25–27], and transcript and protein domain classification [28–31]. Transcriptomic and genomic sequencing data from the SRA have also been pooled and analyzed to resolve a multitude of research questions that may be difficult to answer with single, smaller experiments. From datasets of hundreds to hundreds of thousands of samples, aggregated SRA data has been applied to the identification of common genetic variants [32], recognition of microbiome-based disease signatures [33], and the strengthening differential gene expression findings across multiple comparable experiments [34].

In concert with the growing abundance of protein-related sequencing data, structural databases such as PDB represent a vital resource for researchers interested in modelling, predicting, and engineering protein structures. The unprecedented availability of structural data in PDB and elsewhere has fueled an ongoing revolution in protein structural prediction as accurate protein structures are an essential aspect of any prediction pipeline. Whether used simply to train a larger model [35] (ADD MORE) or as an integral part of the modelling process itself [36,37], protein modeling approaches rely on high quality known structures. Large, diverse libraries of structural data further underly many other protein-focused computational resources. From structural alignment and search programs [38–40] to frameworks aimed at designing novel protein structures [41,42], databases like PDB provide the foundation for essential tools to both study and engineer protein molecules.

In addition to the biological data of interest, most -omics and structural databases also collect some amount of metadata tied to each sample or experiment. Such metadata are particularly beneficial in tying together large, pooled datasets as they can often provide scientifically relevant information that can be used to stratify or classify sample data into meaningful groups. This is the case with passaging annotations for viral sequencing data [43–45], geographic and date/time data for pathogen surveillance [46–49], and growth condition information for various microbial samples [50]. As microbial researchers increasingly recognize the importance and value of sample metadata, metadata specific databases such as Bac*Dive* [51] have been developed to collate relevant information from a vast array of sources into an freely accessible repository.

1.3 SUMMARY

Here, I leverage large, heterogeneous datasets to provide novel insights into pathogenic microbes and the therapeutics used to treat them clinically. Utilizing metadata from influenza A sequences in the GISAID database, I highlight key concerns with existing nomenclature procedures and outline guidelines to improve and clarify the sequence annotation process (Chapter 2). Next, I develop a comprehensive gene coexpression network by employing all published V. cholerae RNA sequencing data contained in the

SRA in combination with select transposon insertion and chromatin immunoprecipitation sequencing data (Chapter 3). This network ultimately provides insight into the role of functionally unannotated genes and identifies novel gene-gene relationships. Finally, I exploit published PDB structures to characterize the properties of a protein structural motif, the beta hairpin, common to antimicrobial peptides and other bioactive compounds (Chapter 4). I use this data to establish a set of principles applicable to the design of protein-therapeutics and other molecules containing the beta hairpin motif.

Chapter 2: Influenza passaging annotations: What they tell us and why we should listen

2.1 INTRODUCTION

This work has been previously published in the journal *Virus Evolution*.[1]

Thousands of influenza viruses are sequenced annually in a global public health endeavor aimed at understanding and combating seasonal epidemics. The constant, steady proliferation of sequenced viruses from year to year has both led to more information available for vaccine development [52] and allowed researchers to create progressively more detailed viral phylogenies in an effort to identify regions under selection [53,54]. Increasingly sophisticated analyses using sequences from various collaborative influenza databases, such as the Influenza Research Database Flu database (IRD) [55] and the Global initiative on sharing all influenza data Epiflu (GISAID) database (https://www.gisaid.org), have helped identify long-term evolutionary trends in influenza viruses [56–58]. While these efforts have greatly expanded our understanding of influenza virus evolution and have led to more informed vaccine development, they have also highlighted a major stumbling block in influenza research as a whole: spurious adaptation signals introduced by cell passaging [43,59–61].

Although it has long been known that high levels of passaging and cultivation in certain cell types can alter influenza strain phenotype and sequence [62–64], recently it has been shown that even low levels of passaging in a wide range of cell types can introduce false adaptation signals. Spurious adaptation signals were first identified in egg passaged

[1] DuPai, C. D. *et al.* Influenza passaging annotations: what they tell us and why we should listen. *Virus Evol.* **5**, (2019). C.D.D. and C.O.W. conceived and designed the analysis pipeline. C.D.D. & C.D.M. collected and analyzed the data. C.B.M., R.G., & S.M. provided input on data interpretations and technical expertise. C.D.D. wrote the manuscript and designed all figures. All authors edited and revised the manuscript.

influenza sequences some 25 years ago [63]. Since then, similar signals have been shown to originate in samples cultivated in a myriad of cell types derived from diverse species and tissues, including canine [43,65,66], monkey [43], and hamster [67] cell lines.

The recent identification of a spurious Zanamivir (influenza neuraminidase inhibitor) resistant mutation in MDCK (Mardin Darby Canine Kidney) passaged sequences [68] highlights that such false signals represent more than just a theoretical concern for the influenza research and the larger medical communities. While the Zanamivir example is concerning, the impact of such erroneous information on seasonal vaccine development is of a potential greater medical threat. False signals complicate downstream analysis and can lead to poorly inferred evolutionary trends, which may ultimately result in improper strain selection and the development of less effective vaccines. Indeed, recent sub-optimal vaccine strains may have passaged isolates to blame, as highlighted by structural and biochemical analyses linking the poor performance of vaccines developed from egg-passaged sequences directly to mutations caused by passaging [69–71].

While research efforts into other human viruses may also encounter false adaptation signals related to cell passaging, the issue is particularly pronounced in influenza virus because of the diversity of cell lines used to culture the virus and the seasonal vaccine development efforts. In terms of cell line diversity, most other human viruses are solely cultured in primate cell lines (e.g., Zika virus or Ebola virus) or in human cell lines (HIV) [72–74], and these cell lines are likely to produce less significant adaptation signals than the broad collection of cell lines in which influenza samples are cultivated. Influenza is also unique in that global influenza surveillance efforts are aimed at producing yearly vaccines that are likely more influenced by false adaptation signals compared to vaccines or treatments developed over longer periods for slower evolving and non-seasonal disease agents.

With growing focus on the effects of cell passaging on influenza sequencing data, it is becoming increasingly important for researchers to clearly understand the nomenclature used to annotate passaged sequences. To facilitate this understanding, we provide below a clear outline of existing annotation strategies for common sequences in the IRD, GISAID, and SIB OpenFluDB (OpenFlu, http://OpenFlu.vital-it.ch/) databases, and we propose a standardized approach to annotating isolates. We hope that this perspective will catalyze a more systematic approach to creating, storing, and analyzing passaging information in the influenza research community, and that this effort will ultimately strengthen research efforts that lead to refinements in the seasonal vaccine strain selection process.

2.2 PASSAGING: WE DO IT BECAUSE WE HAVE TO

Given the growing body of research illuminating issues with cell-passaged influenza sequences and the increasing availability of sequences from original clinical specimens, one has to ask why passage influenza samples at all? Most influenza research, whether the investigation involves vaccine development or not, needs to propagate viruses in vitro in order to analyze host characteristics. In an ideal world, clinical specimens would be sequenced directly, analyzed, and used to create an accurate model of influenza adaption that would inform strain inclusion for vaccine development with little to no bias. This best-case scenario, however, is impeded by the techniques needed to characterize viruses in vitro.

While it is true that many influenza clinical specimens are directly sequenced without passaging, clinical specimens do not typically provide the amount of virologic material necessary to perform the standard antigenic assays: the hemagglutination-inhibition (HI) test, which is essential in strain selection during vaccine development

(Figure 2.1) or animal experiments [75,76]. Indeed, the HI assay requires a minimum of approximately 7 logs of virus per 50uL [77] (8 hemagglutination units/50uL [78]), which is acquired through one or more rounds of passaging. Additionally, two types of vaccines require strains which must be passaged either in eggs or in a qualified MDCK cell line [79,80]. Thus, while the use of sequences derived from original clinical material represent a research ideal, the current reality is that passaged isolates are a necessary step in obtaining sufficient antigenic and genetic information for vaccine development; therefore, it is important to have a clear understanding of passaging and its effects on viral sequences.

2.3 PASSAGING NOMENCLATURE

Currently, the vast majority of influenza sequences are passaged isolates from a menagerie of various cell types. This passaging information is indicated via a patchwork of non-standard nomenclature methods that vary wildly across and even within databases. Indeed, the passaging information associated with A (H3N2) samples collected between 2005-2018 and stored in OpenFlu, IRD, and GISAID illustrates the haphazard naming and numbering strategies for various cell types used to passage isolates (Table 1). In the GISAID database alone, MDCK passaged samples are indicated with at least 15 variable naming schemes from different institutions. The absence of clear labeling patterns combined with the extreme variability in naming conventions across cell types and databases create a Gordian knot for researchers seeking to disentangle the effects of passaging on influenza sequences. Despite ongoing work to develop tools to parse passage history abbreviations any such tool will require constant manual updates to keep pace with novel abbreviations introduced by new entries.

Notwithstanding substantial heterogeneity in approach, all passaging annotations aim to provide similar relevant information about the history of the cultivation of an influenza isolate sequence: typically, the type of cell(s) used in passaging, number of passages, and cell handling data (movement between laboratories and/or change in substrate)**.** This information is then used both to identify factors responsible for false adaptation signals and to help distinguish which sequences should be excluded from downstream analyses.

2.3.1 Cell type

Annotated influenza samples are typically passaged in one or more of only a handful of cell types. Annotations generally begin with an indicator of the cell line used, such as SIAT for MDCK SIAT cells, E for egg, and PMK for primary monkey kidney (Table 2). These indicators are in no way standardized and substantial variation exists for each cell type. Figure 2.2 illustrates the frequency of unique labels used to identify sequences passaged at least once in a monkey cell line. While these labels are likely similar enough to allow a researcher to manually distinguish samples that have been passaged in monkey cell lines from those that have not, they are dissimilar enough to make it difficult for an automated script to efficiently do the same. In addition, many indicators are vague or ambiguous (i.e. X1 or C1 for a sample that has been passaged once in "an unspecified cell line"), and a significant portion of influenza A(H3N2) and A(H1N1) samples lack any cell type indicator or passaging annotation whatsoever (Figure 2.3A). Although the proportion of such unclearly annotated isolates has decreased in recent years, they still accounted for roughly one third of all recorded A(H3N2) isolates in 2017 (Figure 2.3B).

Even though many influenza virus isolates have missing passaging annotations or are ambiguously labeled, several studies have successfully identified the impact of passaging on sequence fidelity in the most commonly used cell types [43,65–67]. For influenza A(H3N2) viruses, these studies have shown that MDCK cells expressing human SIAT1 produce sequences that differ the least from sequences derived from original clinical samples and are now (from 2015 onwards) the predominate cell type used in North America to passage isolates as per GISAID records [43,61,66].

2.3.2 Number of passages

Passage number, the most uniformly recorded annotation aspect, is consistently indicated by a number succeeding a cell type indicator with or without a space between (e.g., MDCK2 or MDCK 2). This convention allows researchers to easily parse annotations for passage number information, although some confusion arises when cell line names include numbers (e.g., MV1LU cells) and when passaging annotations lack clear indicators for all cell types (e.g., C 3 + 1). As with cell type, many samples exclude information about passage number. This lack of information is represented either explicitly with an X following a cell type indicator or implicitly with lack of a number indicating no information.

Due to the reasonably clear and consistent nomenclature currently in use, passage number is perhaps the easiest factor to study when focusing on influenza adaptation to cell culture. As such, several groups have been able to show that each additional passage has a consistent, additive impact on the presence of false adaptation signals in sequenced samples. Across the most commonly used cell types, sequences diverge more from original clinical specimens as passage number increases [43,62]. Since most annotated sequences plainly indicate the number of passages it is also easy to consider this trend when inferring influenza virus phylogenies or selecting vaccine strains. Researchers can simply favor sequences which have been passaged less, although the additive effect of passaging makes it difficult to provide an absolute limit on the number of passages acceptable for any given cell type.

2.3.3 Heterogeneous passaging and cell handling data

Samples are also often passaged in multiple cell lines and/or cell types. Such passaging is commonly indicated by a wide array of symbols, including “and”, blank spaces, “,”, “-”, “_”, “;”, and “+”. Cell lines can also be listed with no separation, e.g. M3C3 to indicate three passages in MDCK and three in some other cell line, and certain symbols are used by some researchers to indicate more information than just passaging, such as “/” indicating the transfer of a strain between labs or institutions. This diversity adds another layer of difficulty to parsing sample annotations, and it has made it particularly difficult to investigate the impact of heterogeneous passaging on sequence fidelity. Consequently, most studies lump such samples together and either analyze them as a heterogeneous group [61] or exclude them altogether [43]. Such lumped analyses mean that little can be determined about the effects of heterogeneous passaging and freeze–thaw cycles in specific cell lines.

2.3.4 Database differences

In an effort to strengthen influenza surveillance efforts, several databases of influenza sequences are maintained, the three largest of which are GISAID, IRD, and OpenFlu. Each of these databases collects sequences from different sources, although there is a fair amount of overlap (see Figure 2.3A) as all include publicly available samples from the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org), stored in the National Center for Biotechnology Information’s GenBank repository [19]. While each database supplements these INSDC samples with user-uploaded sequences, IRD and OpenFlu have far fewer user-uploaded data than GISAID, which includes nearly all available influenza A(H3N2) and (H1N1) data (Fig. 2.3A). Despite the large amount of shared isolate data across databases, each database parses annotations into differently named fields. For example, the sequence and metadata for the influenza A(H3N2) isolate

A/Zhuhai/964/2008 was uploaded to Genbank on July 24, 2016 with passage indicated as "MDCK" under the "lab_host" field of the structured comments. In the corresponding GISAID record the passage information is listed correctly under "passage details/history" as "MDCK" while in IRD and OpenFlu the same sample is listed as "N/A" and "no information" under the fields "Passage History" and "passage", respectively. The correct annotation is present in IRD under "lab host", a field that cannot be included when downloading sample information. The correct annotation is wholly absent from the OpenFlu record, even though other records from the same submission did properly import the passaging information from GenBank to OpenFlu under the "passage" field. Consequently, each database will not only contain some degree of unique samples but also divergent nomenclature standards and metadata that can produce conflicting information for the same sample. These variations make it difficult for researchers to easily integrate and investigate sequences from multiple databases, and they may bring a well annotated isolate's passaging status into question.

2.3.5 Towards a standard nomenclature

Because of the great diversity in annotation strategies, it is difficult to effectively establish exclusion criteria for passaged influenza sequences. Until the effects of cell passaging are better understood it will remain unclear which, if any, cell types produce influenza sequences that are truly free from false adaptation signals. We therefore propose the use of new standard names for common cell lines (Table 2) and a new universal passage annotation convention for all influenza samples (Table 3). These new standards use elements from common existing annotations. They were selected for ease of both human and machine parsability, while staying as close as possible to existing annotation practices, to minimize potential confusion and cost of switching over. The vast majority of existing isolates are passaged in one of the cell lines indicated in Table 2 but adding additional names for uncommon or novel cell lines should prove relatively straightforward once initial guidelines are established.

Additionally, we strongly suggest that influenza databases incorporate changes to encourage accurate passaging annotation and facilitate passage-focused research. These include requiring a passage history field for all sequence submissions, validating that passage history entries either match existing standards (for common cell lines) or are further explained in another field (uncommon cell lines), and making passage history searchable by discrete categories such as Egg, Cell, and Original Clinical Specimen.

2.4 CONCLUSION

Making sense of influenza passaging annotations is a daunting task. However, it is becoming increasingly important for epidemiologists and vaccine developers to consider passage history of isolates when selecting sequences for inclusion in phylodynamic analyses or in vaccines. While a clear and definitive understanding of the effects of viral passaging in all cell types is a distant end point of current research efforts, awareness within the influenza community of the negative impact of cell passaging on sequence fidelity is easily and currently attainable and can only improve epidemiological and clinical research efforts. This highlights the need for a more standard approach to passage nomenclature across influenza researchers and producers of sequence data. As the influenza community is looked upon as a model of data sharing for other epidemic viruses [16], we encourage this highly collaborative community to work together to enact a new global naming convention that further evinces the power and effectiveness of open research.

2.5 METHODS

Annotations were obtained for all (i.e. global) unique, non-laboratory influenza A(H3N2) and A(H1N1) isolates collected from humans between Jan 1st 2005 and Nov 8th 2018 and uploaded to GISAID, OpenFlu, or IRD by Nov 8th 2018 for A(H3N2) isolates and Nov 12th 2018 for A(H1N1) isolates. All annotations were first converted to uppercase characters and then occurrences of each unique isolate and annotation were counted and manually sorted by cell type and passage number. These data were used to generate Figures 2.2 and 2.3 and all Tables.

2.6 Figures

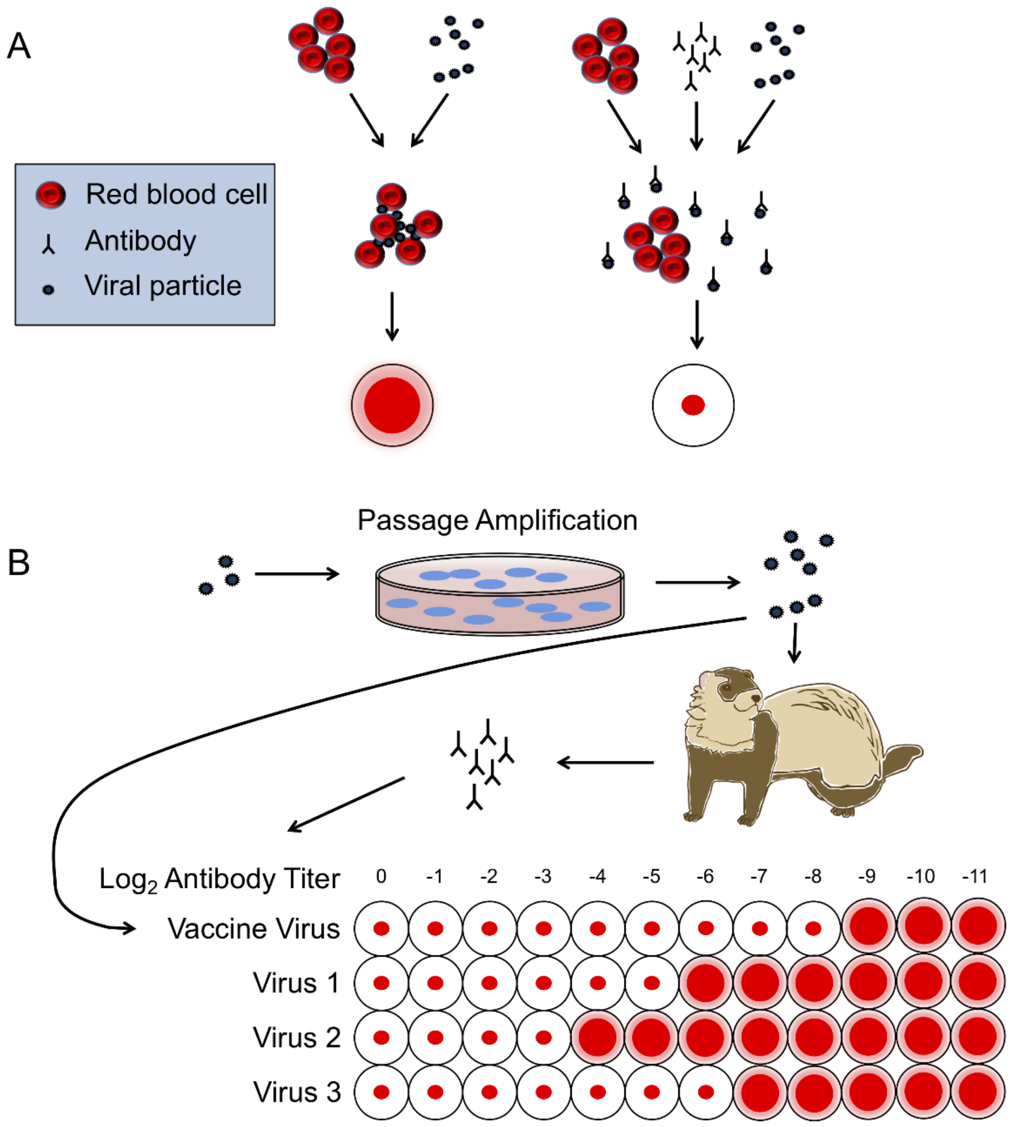

Figure 2.1: (Continued on next page)

Figure 2.1: Hemagglutination inhibition (HI) assay for vaccine development.

(A) Overview of the HI assay. Hemagglutinin on the surface of viral particles bind red blood cells, creating a lattice of blood cells that show up as a diffuse layer at the top of a microtiter plate well. When enough antibodies with strong affinity for the viral hemagglutinin are present, viral particles are bound and the red blood cells sink, forming a small dot at the bottom of the microtiter plate well. (B) Journey of a viral particle, from isolation to HI assay. Viral particles are isolated, cell passaged (often multiple times), and then either tested for hemagglutination activity or used to produce antibodies via infection of animals with naïve immune systems.

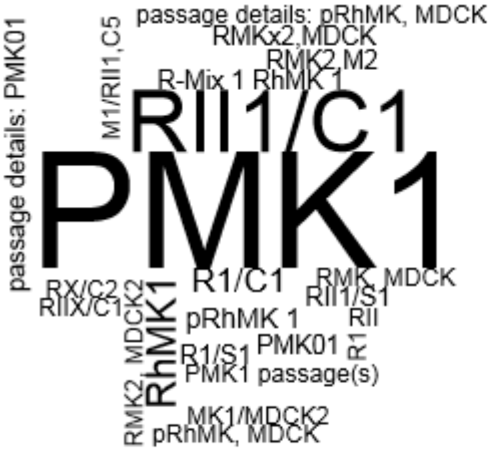

Figure 2.2: Word cloud of influenza A(H3N2) sequence annotations.

This word cloud indicates passaging in one or more cell lines where at least one is a primary monkey kidney cell line. Word height corresponds to number of sequences exhibiting a given pattern.

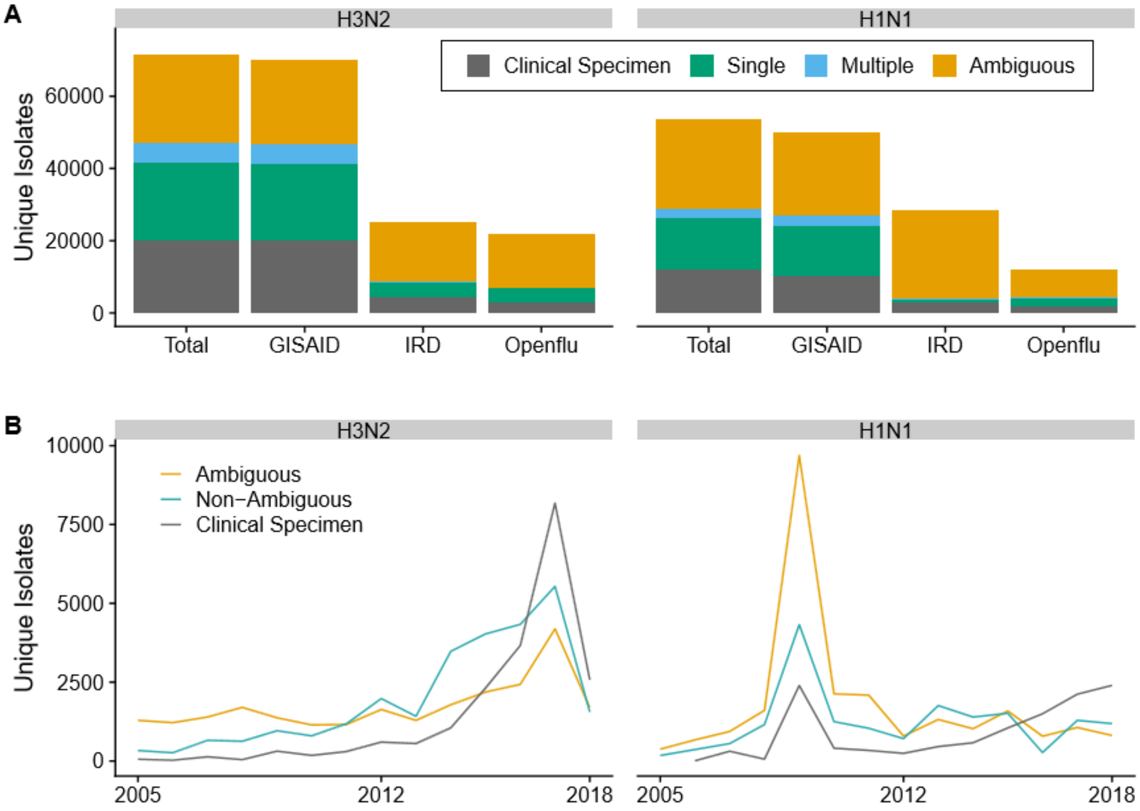

Figure 2.3: Influenza A sequence isolate counts across three databases, 2005–2018.

(A) Aggregate isolate data by type. GISAID accounts for the majority of unique isolates in both strains, about half of which either lack clear passage information or have been passaged in multiple cell lines. Isolate types are defined as follows: "clinical specimen": any unpassaged direct clinical specimen; "single": passaged in single identified cell line; "multiple": passaged in multiple identified cell lines; "ambiguous": passaging information unclear (may be single unidentified cell line or multiple cell lines with at least one line unclear). (B) Yearly isolate data by type. In both analyzed strains there is an increase in direct clinical specimen sequences relative to other samples in more recent years. Includes unique records across all three databases (GISAID, IRD, Openflu). Isolate types are as under (A) except "non-ambiguous" which refers to isolates passaged in one or multiple identified cell lines.

2.7 TABLES

Table 2.1: Common influenza A(H3N2) passaging annotation patterns across three databases.

Base Pattern	No. Cell Lines	Example
CLINICAL SPECIMEN	0	CLINICAL SPECIMEN
DIRECT	0	DIRECT
OR	0	OR
ORIGINAL	0	ORIGINAL
ORIGINAL SAMPLE	0	ORIGINAL SAMPLE
ORIGINAL SPECIMEN	0	ORIGINAL SPECIMEN
PI	1	PI
T	1	MDCK
T#	1	MDCK1
T CELLS	1	MDCK CELLS
P-#	1	P-1
P#	1	P1
PASSAGE DETAILS: T#	1	PASSAGE DETAILS: PMK01
PASSAGE DETAILS: T	1	PASSAGE DETAILS: MDCK
T# (MM/DD/YYYY)	1	S1 (09/30/2008)
T# (YYYY-MM-DD)	1	S2 (2008-09-30)
T # +#	1	MDCK 1 +1
TT#	1	MDCKMDCK1
TT#	2	HEPGMDCK1
T/T#	2	X/C1
T # +T#	2	MDCK 2 +SIAT1
T#/T#	2	C1/C2
T# T#	2	MDCK2 Siat1
T#/T# (MM/DD/YYYY)	2	C1/S1 (01/04/2015)
	No Information	
-N/A-	No Information	-N/A-

T represents type of cell, # represents a single digit number.

Table 2.2: Existing naming conventions and suggested standardized names for common cell lines used to passage influenza viruses.

<table>
<tr><th>Cell Type</th><th>Common Existing Annotations</th><th>Suggested Name</th></tr>
<tr><td><b>Egg</b></td><td>E# | Egg# | Embryonated Eggs | AM</td><td>EGG</td></tr>
<tr><td><b>Madin-Darby Canine Kidney</b></td><td>MDCK# | M# | MDCK CELLS</td><td>MDCK</td></tr>
<tr><td><b>Rhesus Monkey Kidney</b></td><td>RMK# | RHMK# | RII | PMK# | PRHMK#</td><td>RhMK</td></tr>
<tr><td><b>Madin-Darby Canine Kidney - SIAT</b></td><td>MDCK-SIAT# | S# | SIATMDCK# | SIAT#</td><td>SIAT</td></tr>
<tr><td><b>sss</b></td><td>Original | OR | Clinical Specimen | No Passage | Primary | Direct | Nasal Swab | CS</td><td>Original Specimen</td></tr>
<tr><td><b>Unknown</b></td><td>None | | -N/A-</td><td>N/A</td></tr>
<tr><td><b>Unknown Cell</b></td><td>C# | P# | X#</td><td>Unknown Cell</td></tr>
</table>

The symbol “#” represents a single digit number other than 0. Note, this table does not attempt to provide a complete list of all possible cell lines but rather focuses on the most common cell types across three databases.

Table 2.3: Suggested passaging annotation scheme for influenza isolates.

Suggested annotation changes	**Example**
One standardized name per cell line	MDCK
Cell line names should not end in numbers or X	SIAT
Passage number indicated via a number immediately following the cell line	SIAT1
Unknown passage number indicated with an X	SIATX
Intra-lab passaging in multiple cell lines is denoted with +	SIAT1+EGG1
Passaging in multiple cell lines with transfer between labs is denoted with /	SIAT1/EGG1

Cell lines should be represented by a standardized name not ending in X or a number. This includes SIAT cells which previously were also designated as SIAT1 and should be referenced only as SIAT for consistency with the new scheme. Multiple passages in the same cell line should be represented by a number (if number is known) or an X (if number is not known) immediately following the cell line name. Passages in different cell lines should be separated by a plus (+) to indicate intra-lab passaging or a slash (/) to indicate transfer between labs.

Chapter 3: A comprehensive co-expression network analysis in *Vibrio cholerae*

3.1 Introduction

This work has been previously published in the journal *mSystem*.[1]

Since the completion of the first *Vibrio cholerae* genome sequence in 2000, over a thousand *V. cholerae* isolates have been sequenced [81,82]. These sequences has allowed for the development of sophisticated phylogeographic models, which emphasize the importance of controlling the spread of virulent and antibiotic resistant *V. cholerae* strains to lower disease burden, in addition to fighting endemic local strains [82–86]. The integration of hundreds of genomes paired with temporal and geographic information into ever growing phylogenies enables analyses using selection models to predict future population trends and derive biologically meaningful insights into *V. cholerae* evolution [87,88]. By developing treatment and vaccination strategies based on phylogenetic models [89], organizations and governments can more efficiently leverage limited resources and more effectively prevent disease spread in line with the World Health Organization's goal of eradicating cholera by 2030 [90].

Alongside advances in genomics research, the *V. cholerae* and broader bacterial biology communities have benefited greatly from other next generation sequencing (NGS) technologies. Targeted sequencing experiments have been essential in mapping complex virulence pathways, illuminating a novel interbacterial defense system, and expanding our knowledge of the role of non-coding RNA (ncRNA) in the vibrio life cycle [91–97]. Further

[1] DuPai, C. D., Wilke, C. O. & Davies, B. W. A Comprehensive Coexpression Network Analysis in *Vibrio cholerae*. *mSystems* **5**, e00550-20 (2020). C.D.D. and C.O.W. conceived and designed the analysis pipeline. C.D.D. collected and analyzed the data. B.W.D. provided input on data interpretations and expertise pertaining to *V. Cholerae* genetics. C.D.D. wrote the manuscript and designed all figures. All authors edited and revised the manuscript.

discoveries such as transcription factor mediated transposon insertion bias [98] and the role of cAMP receptor protein in host colonization [99] have benefited from composite research strategies utilizing multiple technologies. Similarly, meta-analyses utilizing pooled data from multiple experiments are empowered by the increasing availability of high quality bacterial NGS datasets. Expression data is particularly amenable to such pooling and can be used to accurately group genes into functional modules based on their co-expression [100]. In bacteria, weighted gene co-expression network analysis (WGCNA) [101] has been successfully used to underscore biologically important genes and gene-gene relationships via “guilt-by-association” approaches [102,103]. These studies have taken advantage of larger and larger heterogeneous microarray datasets to provide novel biological insights via existing data.

Despite major advances in sequencing technologies and research strategies, most of the over two dozen existing RNA-seq experiments in *V. cholerae* have been limited to targeted approaches that involve quantifying the differential abundance of genetic material across a handful of conditions. Via these approaches, any change in expression observed in one experiment is nearly impossible to generalize to other treatment conditions and analyses are limited to a few pathways or genes of interest. In contrast, meta-analyses such as WGCNA can uncover much broader relationships throughout the entire genome by combining information from multiple datasets. As there is no existing co-expression analysis in *V. cholerae* to date, the accumulation of over 300 publicly available RNA-seq samples from targeted RNA-seq experiments represents a heretofore untapped resource for the cholera community.

Motivated by the success of pooled genetic sequencing analyses, our current work utilizes all publicly available *V. cholerae* RNA-seq based expression-level data to generate a co-expression network. We expand upon existing bacterial WGCNA approaches by integrating broader sequencing data (including ChIP-seq and Tn-seq) and multiple annotation platforms into our analysis. Our network ultimately contributes information on connections across all *V. cholerae* genes, including the roughly 1500 predicted but functionally un-annotated genetic elements that account for some 37% of the genome. More specifically, we implicate new loci in virulence regulation and clearly demonstrate a powerful and accurate approach to hypothesis generation via our described network.

3.2 RESULTS

To generate our co-expression analysis in *V. cholerae,* we applied our WGCNA pipeline to analyze twenty-seven *V. cholerae* RNA sequencing experiments deposited in NCBI's Sequence Read Archive (SRA) in addition to two novel experiments. The RNA sequencing samples are derived from experiments exploring a range of important *V. cholerae* processes including intestinal colonization, quorum sensing, and stress response. In total, our network includes 300 individual RNA-seq samples (Supp. Table S1). All samples were mapped to a recently inferred *V. cholerae* transcriptome derived from the N16961 reference genome [81,93]. This reference was chosen because the majority (293) of samples were collected from strains N16961 or the closely related C6706 and A1552.

Figure 3.1 outlines the process used to generate our co-expression network with a small subset of genes. The five included loci are known to be involved in cysteine metabolism with VC0384–VC0386 and VC0539–VC0540 falling within two separate operons. Following normalization of mapped transcripts (Fig. 3.1A), a weighted gene co-expression network analysis was performed using WGCNA [101]. First, a Pearson correlation matrix is calculated for expression levels of all genes (Fig. 3.1B). This correlation matrix clearly captures strong relationships between co-expressing genes but can produce background noise from un-related gene pairs and underlying gene structures (i.e. operons).

We limit this noise by calculating a topological overlap matrix (TOM) [104] that weights pairwise co-expression data based on each gene's interactions with all other genes (Fig. 3.1C). In this way, the relationships between genes that fall within the same subnetwork are favored while signals from less tightly co-regulated genes are abated, This TOM, after filtering for normalized values greater than 0.1, is used to construct an accurate co-expression network that captures biologically meaningful relationships while minimizing background noise (Fig. 3.1D).

In addition to co-expression data, our network and analyses incorporate information from multiple other sources. Our network includes predicted pathway annotations and gene functional knowledge from the NCBI Biosystems database as well as the DAVID, Panther, and KEGG databases [105–108]. Operon structure was inferred using Operon-mapper [109]. Additionally, importance labels were applied to genes with no known function which have been implicated as playing a role in intestinal colonization or *in vitro* growth via Tn-seq based essentiality experiments [94,110]. Information from ChIP-seq binding assays and microarray results were incorporated in downstream analyses to substantiate network derived relationships. By combining all of these data sources we were able to develop and analyze an informative network of co-expressing genes that provides both qualitative and quantitative information about relationships between transcripts across forty-nine gene-clusters covering the entire *V. cholerae* genome (Supp. Data S1, S2).

3.2.1 A network of novel, unexpected, and informative interactions

As many functionally related bacterial genes are co-expressed in operons such as VC0384–VC0386 above, we sought to uncover if operon structure was a contributing factor to our network or specific subnetworks. Indeed, gene pairs predicted to fall within the same operon did show significantly higher average normalized co-expression than their nonoperon counterparts (0.186 vs. 0.147, $p < 0.001$), and some subnetworks, such as the Ribosome Related subnetwork (Fig. 3.2A), contain a high proportion of intra-operon gene pairs (Supp. Fig. 1). However, across our full network only 0.2% of all co-expressing gene pairs fall within the same operon and no subnetwork has a majority of such pairs (Supp. Fig. 1). Moreover, our overall network captures information on relationships with the roughly one third of unannotated *V. cholerae* genes (Supp. Fig. 2), providing insight into functional roles that are not obvious based on gene homology or known operon structure.

3.2.2 Genes in known pathways cluster together and contextualize genes of unknown function

As a demonstration of the accuracy of our approach, we have highlighted several clusters that recapitulate known interactions between transcripts involved in highly conserved, well studied cellular processes (Fig. 3.2). The correct grouping of transcripts encoding ribosomal proteins, tRNAs, and amino acid synthesis proteins into significantly co-expressing subnetworks provides a positive control for our overall network (Fig. 3.2A–C). Importantly, our analysis clustered together genes known to be involved in more specialized processes such as motility and biofilm formation (Fig. 3.2D, E), with corresponding gene ontology (GO) [111] and KEGG [107] pathway terms enriched for genes within these subnetworks (Fig. 3.3 and Supp. Table S2).

In addition to capturing relationships between genes involved in specific pathways, our approach can also accurately group genes involved in interconnected processes that share overlapping regulation, as seen in the environmental sensing subnetwork (Fig. 3.2F). This subnetwork includes high level transcriptional regulators, such as AphA, TfoS, and TfoY, with known roles mediating the complex interplay between quorum sensing, natural competence, type VI secretion, and other related pathways [112–117]. As each of these transcription factors is involved in a multitude of cellular processes and significantly co-expresses with hundreds of other genes, our analysis describes their closest connections under parameters designed to find meaningful relationships that are also manageable to interpret. By altering these parameters (significance cut-offs, minimum number of genes per cluster, clustering algorithm, etc.) analysis of the overall network can be fine-tuned to focus in on specific biological processes or explore the nodes that drive connections between processes that are necessary for *V. cholerae* to adapt and survive in diverse environments.

The subnetworks outlined in Fig. 3.2 support the utility of our analysis in powering guilt-by-association based inference of gene function [118]. Because each of these gene clusters contain co-expressing genes that are involved in the same biological process, it can be assumed that unannotated genes in the same cluster are likely involved in the same process. Such links, while not definitive on their own, can be used with other data to hint at gene functions. For example, genes with known function in Fig. 3.2E are primarily involved in biofilm formation [119,120]. This clustering of biofilm genes suggests that the few genes with no known function in this subnetwork may be involved in the same process. Two of these unannotated transcripts, VC1937 and VC2388, are, per GO cellular component location labels, “integral membrane components.” Further, the VC2388 locus is directly upstream of a Vcr084, a short RNA involved in quorum sensing which is essential for biofilm formation [121]. Taken together, this evidence suggests that VC1937 and VC2388 may play a role in some of the complex membrane restructuring necessary for biofilm formation. In facilitating such guilt-by-association approaches to novel hypothesis generation, our co-expression network serves as a highly efficient substitute for more traditional screening assays.

3.2.3 A virulence subnetwork suggests novel gene functions

While the biofilm associated subnetwork (Fig. 3.2E) presents a relatively simple example of the functional insights our co-expression data can yield, the virulence-related subnetwork (Fig. 3.4A) represents a more complex case in which genes of known function provide clues to the role of unannotated genes. The majority of transcripts in this module originate from within the virulence-related ToxR regulon that consists principally of genes on the *V. cholerae* pathogenicity island 1 (VC0809–VC0848) and cholera toxin sub-units A and B (*ctxAB*, VC1456 and VC1457) [122]. Other genes in this subnetwork, such as *vpsJ*, VC1806, VC1810, and chitinase, are predominately localized to virulence islands and other areas of the genome under tight control of the known virulence regulators ToxR, ToxT, or H-NS as determined via ChIP and/or RNA-seq [123–125]. Genes in this subnetwork are also enriched for virulence related GO and KEGG terms, such as "pathogenesis" and "*Vibrio cholerae* infection" (Fig. 3.3). The clustering of such genes with well-characterized interactions into a cohesive subnetwork is further validation of our ability to generate accurate co-expression maps of related genes. The association of uncharacterized genes in these clusters suggests they may also play a role in *V. cholerae* virulence and generates hypotheses about the function of unknown genes within this module.

Many of the important transcripts with unknown function are expected to co-express with known virulence genes because they fall within vibrio pathogenicity island (VPI)-1 (VC0810, VC0821–VC0823, VC0842) or VPI-2 (VC1806, VC1810), or are proximal to other virulence genes (VC1945) [126,127]. However, our analysis also identified genes such as VCA0094–VCA0096 which are on a completely different chromosome than the rest of the subnetwork and do not neighbor any known virulence elements.

A major benefit of our approach is that we incorporate additional regulatory data such as ChIP and Tn-seq into our co-expression analysis, allowing us to verify the association between VCA0094–VCA0096 and virulence pathways using existing experimental data. Tn-seq analysis has previously identified VCA0094 and VCA0095 as essential for infection of a rabbit intestine [94], suggesting that these loci play a role in virulence. Because transcripts for these genes co-express with genes regulated by ToxT, ToxR, and H-NS, we also probed existing ChIP-seq binding datasets [92,99,124] to see if any of these well-studied transcription factors bind near the VCA0094–96 loci. While ToxT binding was not observed near this site (data not shown), our analysis identified significant peaks in the promoter region of VCA0094 for both ToxR and H-NS as calculated via re-analysis of existing binding data from [124]. Both peaks showed a large and significant increase in binding affinity ($\log_2$ fold change in average occupancy) when compared against input controls (Fig. 3.4B). H-NS showed a clear binding peak in the region of the VCA0094 promoter that extended in a diffuse manner to the VCA0095 transcription start site while ToxR binding covered a similar region but was more diffuse throughout (data not shown). Collectively these results indicate virulence related functions for the products of the VCA0094–VCA0096 transcripts. Although the exact mechanistic role of these genes remains elusive, we have nevertheless demonstrated the ability of our pipeline to generate meaningful hypotheses by incorporating existing data from a multitude of sources.

3.2.4 Co-expression data provides an accurate *in silico* complement to RNA-seq

In addition to the guilt-by-association inference described above, co-expression analysis can provide a partial substitute or complement to RNA-seq experiments. Novel, meaningful genetic relationships can be found in a co-expression network by focusing on the transcripts that are co-regulated with a gene of interest.

We can apply a network-based approach in lieu of new RNA-seq based experiments to identify genes which co-express with *rpoS* (VC0534) and are similarly involved in bacterial stress response. As our network utilizes only RNA-seq based transcriptomics studies and none of these studies involves direct manipulation of *rpoS*, we can compare existing microarray data involving an *rpoS* (VC0534) deletion mutant [128] to determine how accurate our approach is. When applying an absolute co-expression cutoff of 0.1, 272 genes are identified as having a relationship with *rpoS* expression in both our network analysis and the *rpoS* mutant microarray data (Fig. 3.5A). This represents nearly two-thirds of genes identified as differentially expressed in the original microarray study. While our network links far more genes with *rpoS* than the microarray approach, this is in line with recent RNA-seq based work that found that 23% of the E. coli genome is regulated by RpoS [129]. Additionally, all of the flagella and chemotaxis related proteins highlighted as particularly informative in the original study are identified by our analysis (Fig. 3.5B) and relevant values (i.e. network co-expression and microarray-derived log fold change in expression) for the 273 shared transcripts have a Spearman correlation of -0.516. This accuracy is achieved without any direct genetic manipulation of the *rpoS* locus in the RNA-seq datasets used to generate our co-expression network and serves as a testament to the potential utility and versatility of our approach.

Our approach to isolating genetic interactions also has advantages over transcriptomics-focused sequencing. As seen in Fig. 3.5A, our network-based analysis identifies far more genes associated with *rpoS*. This is likely because RNA-seq-based approaches are can identify a broader range of gene transcripts as they are not limited by restrictive microarray probes [130]. Separate from differences in underlying technology, co-expression networks are also more likely to detect genes regulating a target's expression than traditional transcriptomics experiments which largely capture downstream responses to changes in a target's expression [131,132]. Thus, a co-expression network can provide an alternative perspective to complement or clarify transcriptomics data.

3.3 Discussion

We have successfully constructed the first *V. cholerae* co-expression network through a computationally inexpensive process that is simple, easily expanded upon, and straightforward to implement in other organisms. Our network effectively identifies canonical gene clusters related to specific molecular pathways or functions, such as those corresponding to tRNAs or biofilm proteins. We have also outlined two use-cases for the data provided and have shown the accuracy of both approaches using existing data. Additionally, we have included relevant network files as well as raw read counts across RNA-seq conditions (Supp. Data S1, S2 & Supp. Table S3) alongside all code used in our analysis (see Materials and Methods) to encourage broad usage of this data.

Our results have proven both the utility and accuracy of our approach despite in-depth analysis limited to a handful of genes across five of the forty-nine observed gene clusters. Furthermore, our work with the virulence subnetwork supports previously published research loosely implicating genes VCA0094–VCA0096 in virulence and virulence related functions. All three transcripts have shown up in screens focusing on biofilm development [133], and SOS response [93]. From a mechanistic perspective, protein homology analysis via NCBI's Conserved Domain Database [134] indicates that VCA0094 possesses a DNA-binding transcriptional regulator domain while VCA0096 contains domains that implicate it in protein activation via proteolysis. These data combined with our novel findings hint at the potential biological importance of this genomic locus.

When viewed through the lens of a specific gene of interest, co-expression data is in large part analogous to the differential expression data produced by RNA-seq experiments. While RNA-seq offers finer assay control and can be tailored more exactly to suit a specific research question, there are both technical and practical limitations that may make such an approach impractical. Whether an experimenter is interested in examining the role of an essential locus or is limited by available resources, our co-expression analysis presents a fast, free, and faithful alternative for probing genetic interactions as outlined in our analysis of *rpoS* above.

Major motivations for this work include the successful implementation of bacterial-focus, microarray-based co-expression networks and the lack of clear functional knowledge for a large portion of *V. cholerae* genes. Besides more simple guilt-by-association studies [102,103], co-expression networks have helped to elucidate relationships in diverse microbial communities [135–138] and enable comparisons across strains and species [139–141]. These works as well as the relative dearth of knowledge about the *V. cholerae* genome (roughly two third of genes are annotated compared to around 86% percent of all *E. coli* genes [142]) and the growing abundance of *V. cholerae* focused NGS data served as the impetus for this research.

The calculated co-expression network, though accurate, could be improved via the inclusion of more experiments and more extensive SRA annotations. Our somewhat limited pooled dataset consisting of three hundred samples is an order of magnitude off from the few thousand samples necessary to derive the most faithful co-expression estimates [143]. Though sample size will improve as more *V. cholerae* RNA-seq experiments are published, more samples may also increase the risk posed by batch effects which cause spurious correlations among genes through technical variation [144,145]. The diverse structure of our current data helps to minimize the impact of batch effects but this would be offset by the future inclusion of larger datasets from single experiments. While automated sample clustering methods [146–148] can effectively group overly correlated samples, there is no way to know if the correlation is biological (i.e. meaningful) or technical (i.e. noise) in origin. Likewise, manual curation of batch annotations is also difficult since few SRA records are extensively annotated with detailed experimental conditions (e.g. bacterial growth stage, exact medium used). Thus, careful consideration may be necessary when expanding and generalizing this analysis to include future data.

The mapping of raw reads to a transcriptome derived from a single reference genome presents a limitation to our current work. While this approach is reasonable given the similarity of the vast majority of included strains to our reference, a more elaborate comparative transcriptomic strategy [149,150] would be ideal if more diverse samples are included in future analyses. This is especially true when considering the inclusion of expression data from clinical samples which are likely to have much more genomic variability than the closely related lab cultured strains used to construct our network. On the other hand, because comparative transcriptomics requires defining homologous alleles across all strains analyzed [151], such an approach would greatly increase the difficulty of incorporating strains without an assembled genome.

In summary, our co-expression network can drive functional hypotheses for unannotated genes in *V. cholerae*. As the Vibrio community steadily adds high quality data from increasingly sophisticated sequencing experiments to public databases our imputed network can only improve, providing ever deeper insights into the *V. cholerae* genome. At the same time, highly annotated transcript-based co-expression networks can empower research with related technologies (e.g. single cell transcriptomics and dual RNA-seq) and research into a host of other clinically relevant bacteria, such as *Pseudomonas aeruginosa* or *Staphylococcus aureus* which have over 2000 and 1400 RNA-seq experiments in SRA respectively.

3.4 Materials and Methods

3.4.1 Data collection and processing

All RNA and ChIP sequencing data were downloaded from the Sequence Read Archive (SRA)[17] and converted to compressed fastq files using the SRA toolkit (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/) (see Table S1 for details on included experiments)**.** RNA-seq samples were selected by searching the SRA on Sept 10th, 2019 for the Organism and Strategy terms "vibrio cholerae" and "rna seq" respectively, resulting in 326 initial samples including the 34 novel samples from this publication (PRJNA601792). Samples were mapped to a recently inferred *V. cholerae* transcriptome derived from the N16961 reference genome [81,93] using Kallisto version 0.45.1 [152]. This reference was chosen because the majority (293) of samples were collected from strains N16961 or the closely related C6706 and A1552. 26 low quality samples with < 50% of reads mapping to the reference transcriptome were discarded before further analysis, leaving 300 samples used for further analysis.

For ChIP-seq analysis, accession numbers were identified via the relevant publications [92,99,124] and sequences were downloaded from SRA and converted to fastq files as above. Raw reads were mapped to the same N16961 reference genome using Bowtie 2 version 2.3.5.1 [153]. From this mapping, peaks were identified using MACS2 version 2.1.2 with an extsize of 225 (various sizes from 150 to 500 were tested with little observable difference in peaks identified) [154] and differential binding and significance were calculated using DiffBind version 2.12.0 [155].

Processed Tn-seq data were collected directly from published datasets. *In vitro* essentiality and semi-essentiality labels were derived from Chao et al. 2013 Table S1 [110], with the original labels of domain essential and sick genes replaced with essential and semi-essential respectively. We used Table S2 from Fu, Waldor, and Mekalanos 2013 [94] to label genes involved in host infection, with any gene exhibiting a $\log_2$ fold change less than negative three deemed essential and any gene with a $\log_2$ fold change between negative one and negative three deemed semi-essential.

3.4.2 Network construction

Figure 3.1 highlights the process used to generate our co-expression network. Kallisto derived reads were first imported into R via tximport [156], then normalized using DESeq2 version 1.24.0 [157], resulting in values that are comparable across conditions and experiments. Following normalization, a weighted gene co-expression network analysis was performed using WGCNA [101]. This process is highlighted with a subset of data in Figure 3.1 and consists of the sequential calculation of a Pearson correlation matrix, adjacency matrix with power ß=6, and, ultimately, topological overlap matrix (TOM) [104] from normalized gene expression counts across conditions. We further filtered this TOM to exclude samples with weighted co-expression <0.1 for all analysis included in the Results section.

Predicted pathway annotations and gene functional knowledge are derived from the NCBI Biosystems database as well as DAVID, Panther, and KEGG databases [105–108]. Genes lacking functional knowledge which are identified as essential or semi-essential in either Tn-seq dataset are labeled in network visualizations as “Important Unknown.” Operon predictions were inferred using Operon-mapper [109].

3.4.3 Data availability

SRA accession numbers and information on included samples can be found in Supplementary Table S1. A full, unfiltered network graph is provided in Supplementary

File S1 with the corresponding node labels in Supplementary File S2. Raw, un-normalized read counts are also provided in Supplementary Table S3. All data analysis and figure generation were done using the R programming language, with code available at DOI: 10.5281/zenodo.3572870. Supplementary Files are available alongside the original publication [158].

3.5 FIGURES

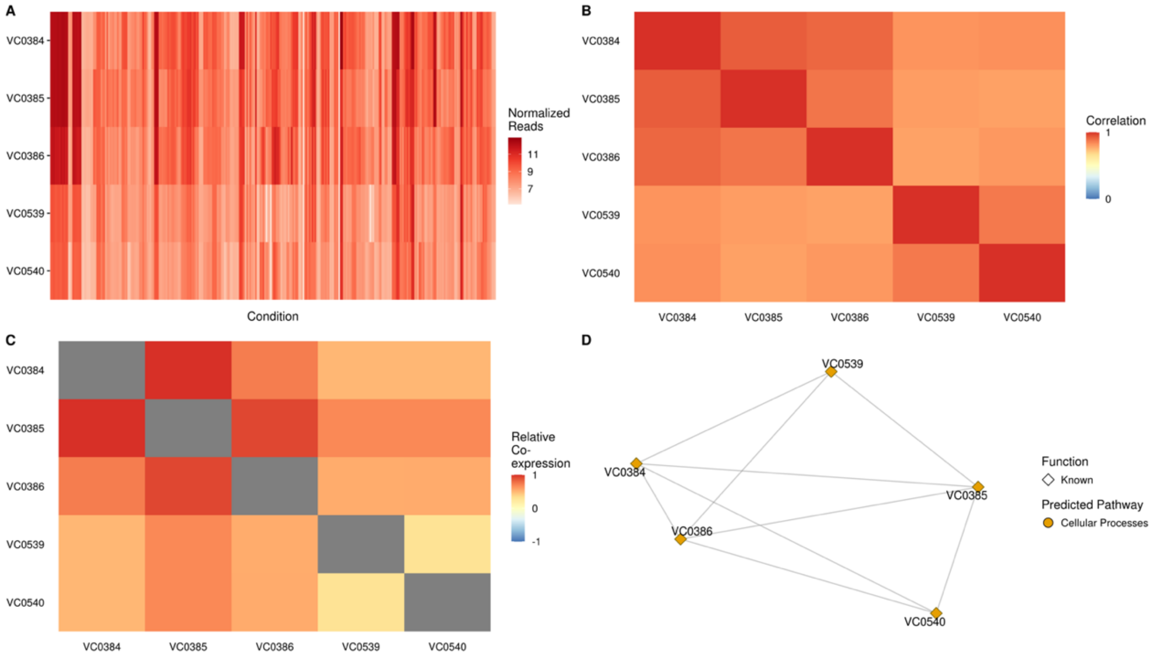

Figure 3.1: General outline of network construction.

To explain the overall WGCNA process we have chosen a subset of genes that are involved in the same core process, cysteine metabolism. Loci VC0394 – VC0386 are predicted to fall within one operon while loci VC0539 and VC0540 in are predicted to be in another. (A) Normalized (log2) expression reads for the same genes across multiple conditions supply the basis for our co-expression analysis. (B) Correlations are calculated from the normalized counts in part A for every pair of genes. (C) An adjacency matrix (not shown) is calculated from the correlations in part B and ultimately used to produce a topological overlap matrix (TOM) that supplies network edge weights with less noise than the raw correlation matrix. While the signal of co-expressing pairs is dampened slightly, this step greatly decreases spurious relationships as it favors transcripts which co-express with similar sets of genes rather than potentially noisy direct correlations. (D) The final network groups transcripts that tightly co-express while indicating what pathway they are involved in. In this example, all genes significantly co-express except VC0539 and VC0540 despite their co-localization within the same operon. After network construction, information is added to label genes based on their function and essentiality under virulence and growth conditions.

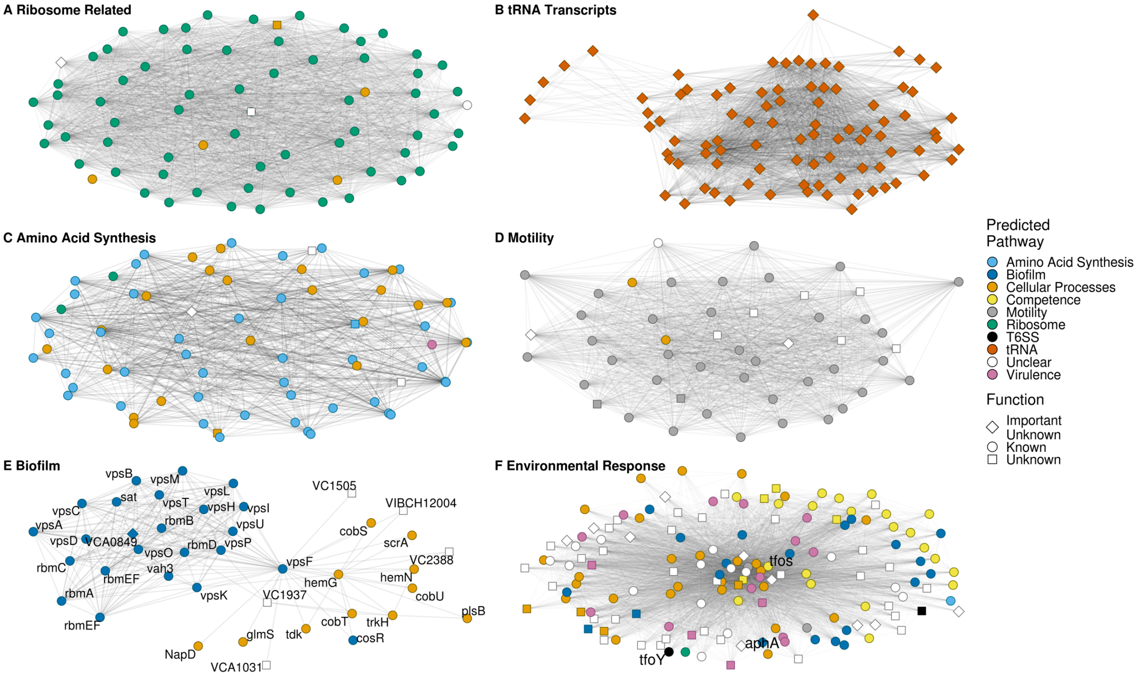

Figure 3.2: Subnetworks recapitulating known results.

The depicted subnetworks each contain transcripts that are known to be largely involved in one or more related biological process(es). For each subnetwork, the nodes represent transcripts while the edges represent a co-expression relationship of at least 0.1 between transcripts. (A¬–F) Subnetworks involved in the following core processes: ribosome related, tRNA transcripts, amino acid synthesis, motility, biofilm, and environmental response. For the environmental response subnetwork, nodes corresponding to labelled genes of interest are enlarged and outlined in red.

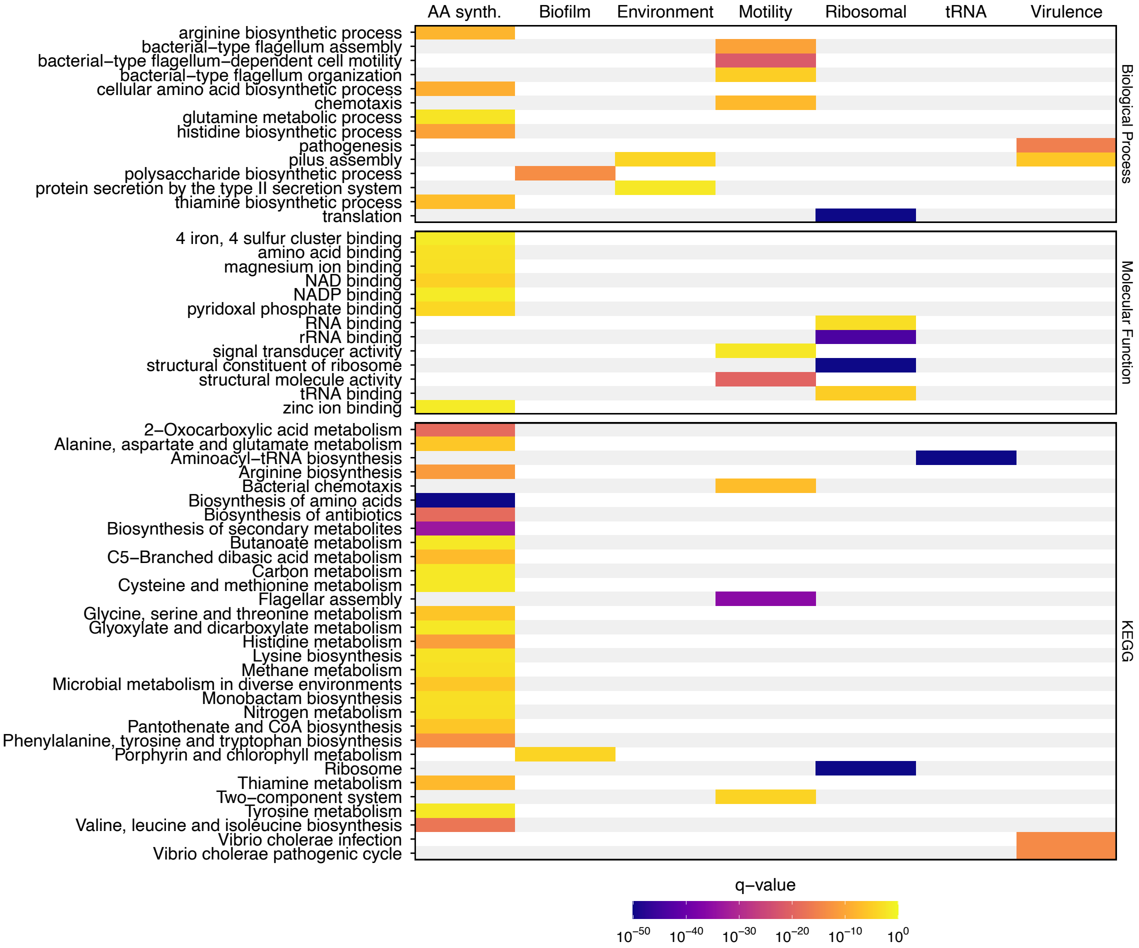

Figure 3.3: Significantly enriched GO and KEGG terms for specific subnetworks.

The indicated terms are significantly enriched within highlighted pathways, with the color indicating the significance of said enrichment as determined via the FDR adjusted p-value (q-value). The terms are divided by database and, for Gene Ontology (GO) terms, GO domain as indicated to the right.

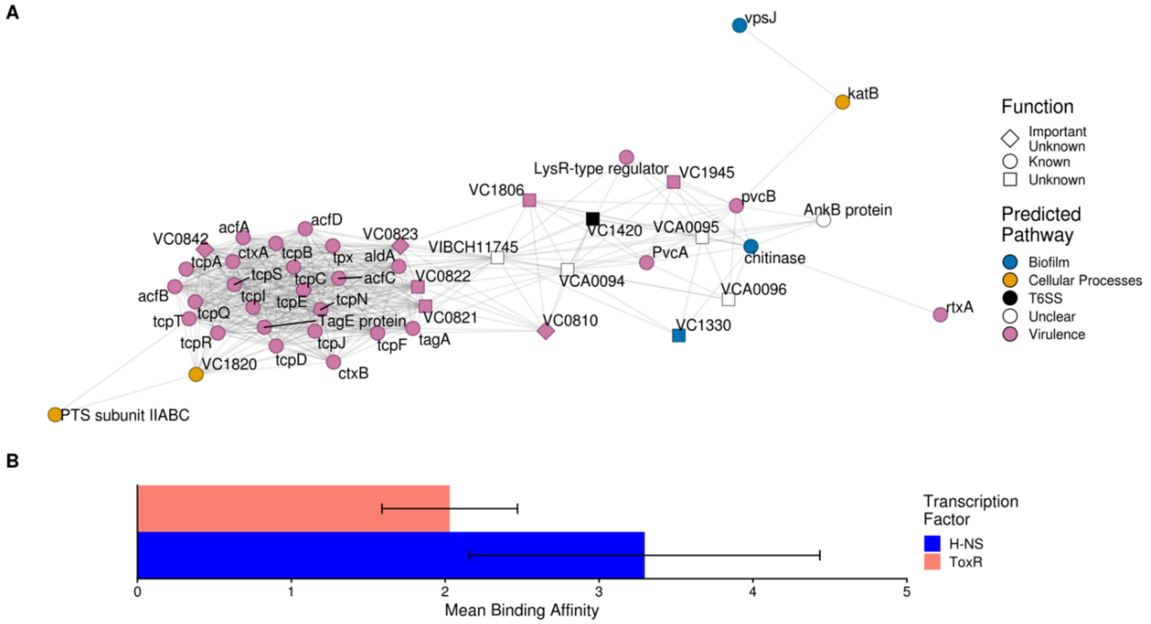

Figure 3.4: Virulence related subnetwork.

(A) This subnetwork contains a majority of genes that are predicted to be involved in virulence related pathways, providing clues to the genes with no known functions such as those at locus VCA0094–VCA0096. The label PTS subunit IIABC stands for PTS system fructose-specific transporter subunit IIABC. (B) Mean binding affinity (log2 fold change in occupancy compared to loading control) for different virulence-associated transcription factors near the VCA0094–VCA0096 locus. Both H-NS and ToxR show a significant binding preference for this region. Error bars indicate standard deviation from the mean.

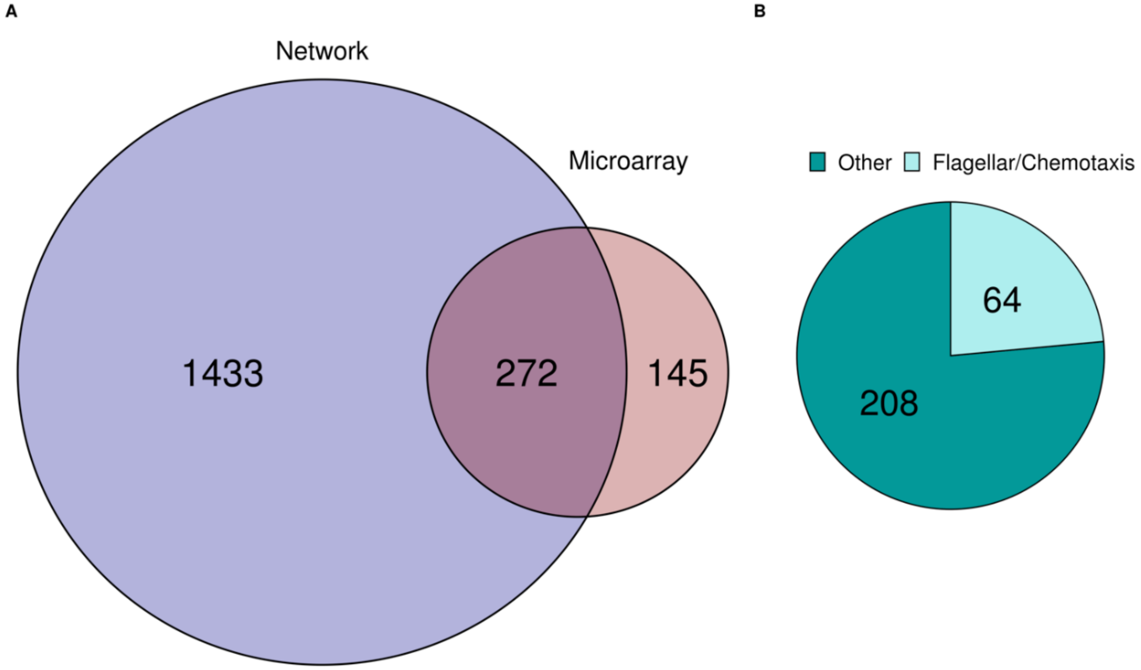

Figure 3.5: Comparing RpoS microarray data to co-expressing genes in our WGCNA.

(A) Overlap of genes with expression pattern related to rpoS expression as identified via our network analysis (blue) and existing microarray data (red). The overlapping region identifies 272 genes that are common between the two analyses. (B) Breakdown of shared genes (overlapping region in A). All of the flagellar and chemotaxis genes highlighted as particularly important in the microarray dataset are identified by both methods.

Chapter 4: A systematic analysis of the beta hairpin motif in the Protein Data Bank

4.1 INTRODUCTION

This work has been submitted for publication.[1]

Beta hairpins, one of the simplest stable protein structural elements, consist of two antiparallel beta-sheets joined by a short loop region. Despite their simplicity in form, beta hairpins are highly adaptable in function. Beta strands are known to participate in protein–protein interactions that are often facilitated by specific amino acid orientations[159] and beta hairpin motifs are no different.[160–162] Indeed, these motifs are a core feature in a diverse array of bioactive molecules, from large beta barrel proteins that transport cargo through cellular membranes[163–165] to substantially smaller antimicrobial peptides and peptide derivatives.[166–168] Whether through self-aggregation,[169,170] target binding,[171] or amphipathic structure formation,[164,172] beta hairpin motifs facilitate a range of different biological functions.

In addition to its prevalence in nature, the beta hairpin motif is stable in even small structures and extensively adaptable to specific functions, making it a popular choice in engineered protein structures. Efforts to design such structures have benefited from several decades of research aimed at identifying how beta hairpins form[173–175] and what factors influence their stability and specific activity.[160,176–179] Examples of synthetic proteins that have successfully adapted the beta hairpin motif for specialized functions include

[1] DuPai, C. D., Davies, B. W & Wilke, C. O. A systematic analysis of the beta hairpin motif in the Protein Data Bank. Manuscript submitted (2020). C.D.D. and C.O.W. conceived and designed the analysis pipeline. C.D.D. collected and analyzed the data. B.W.D. provided input on data interpretations and expertise pertaining to bioactive peptides. C.D.D. wrote the manuscript and designed all figures. All authors edited and revised the manuscript.

hydrogels,[167] antimicrobial peptides,[180] and various molecules with material science applications.[166]

Although largely successful, beta hairpin engineering efforts are typically limited to testing relatively small libraries involving derivatives of a stable scaffold structure or existing protein via peptidomimetics.[7,160,162,177,181] With the increasing availability of high throughput screening platforms to test for activity in large libraries of de novo sequences[182–184] there is an obvious need for broader design principles that can be applied to the generation of libraries with millions of diverse beta hairpin containing proteins. Knowledge of amino acid propensities throughout known beta hairpin sub-structures could inform such design principles but existing catalogs are too broadly focused on beta sheets, outdated, or limited in scope.[174,178,185–188] An up-to-date characterization of amino acid distributions at specific positions within beta hairpins does not exist.

Using a systematic analysis of sequence and structural data from all beta hairpin containing proteins in the Protein Data Bank (PDB), we derived key sequence factors and patterns common to beta hairpins. Important features include amphipathic faces created by the periodic alternation of hydrophilic and hydrophobic amino acids within beta strands, the high prevalence of aspartic acid/asparagine caps at the N-terminal end of beta strands, and specific residue contacts that are over (e.g. cysteine-cysteine, salt bridges) and under (e.g. proline-lysine) represented. These findings give us a broader understanding of naturally occurring beta hairpins and will aid future efforts in the design of bioactive molecules containing the beta hairpin motif.

4.2 RESULTS

To identify and classify motifs we used the following process (see Materials & Methods for further detail). We first collected all PDB structures[18] and their corresponding amino acid sequences filtered to 90 percent similarity. We then used DSSP-derived secondary structure annotations[189] to identify potential beta hairpin substructures consisting of two antiparallel beta-sheets joined by a short loop region (Fig. 4.1). After determining contacting residues between beta strands, we excluded any structures with less than four contacts from further analysis. This process identified nearly 50,000 unique beta hairpin motifs from some 24,000 independent protein structures. Using these structures, we calculated average amino acid frequencies within structural regions and observed amino acid contacts between hairpin beta strands. We then classified and divided motif structures based on turn length and orientation of beta strand faces. Using these groupings, we determined average amino acid frequencies at each position of the beta hairpin motif.

4.2.1 Secondary structure explains average amino acid frequencies

It has long been known that different secondary structural elements tend to favor the inclusion of certain amino acids over others.[186,187,190,191] This is exactly what we see with our analysis of beta hairpin motifs (Fig. 4.2), with a clear difference in average amino acid frequencies between beta strands, the turn region, and background levels across all included protein structures. Our analysis agrees with previous work illustrating a strong preference for glycine, asparagine, and aspartic acid in flexible turn regions.[186,187] While proline is also more common in the turn region than in either beta strand, we see no difference in turn region prevalence when compared to background levels. This is in contrast to previous findings that saw significant enrichment of proline in turn regions.[166,192] This lack of proline enrichment and the relatively low average proline abundance in the turn region is particularly surprising given the known role of such residues in stabilizing beta turns.[192,193]

When looking at amino acid levels in the beta strands, there appears to be little to no difference in prevalence between strands. Both strands show an increased occurrence of isoleucine, valine, and several other chiefly hydrophobic residues in beta sheet structures, supporting previous research.[194] Additionally, both strands show a greater tolerance for positively charged residues as is commonly observed with anti-parallel beta strands as opposed to their parallel counterparts.[165,168,195] We further probed for differences across domains of life but saw no strong trends in individual amino acids (Supp. Fig. 1A). There were, however, taxa specific differences in turn region preference for polar and negatively charged amino acids (Supp. Fig. 1B).

4.2.2 Residue positional biases are linked to flexibility, stability, and hydrophobicity

Beta hairpins, especially those in membrane interacting structures such as beta barrels and some antimicrobial peptides, are known to incorporate amphipathic beta sheets that periodically alternate between hydrophilic and hydrophobic amino acids, creating two distinct faces[196,197] (Fig. 4.1). To account for these faces in our analysis, we divided our dataset based on the presentation of an initial polar or hydrophobic face for both the N and C terminal beta strands (see Materials and Methods). After accounting for these amphipathic faces as well as differences in turn region length, clear patterns emerged in all regions of the beta hairpin motif (Fig. 4.3). The most obvious pattern observed was the alternating preference for charged/polar and hydrophobic residues in both beta strands (Fig. 4.3A-B). While hydrophobic residues appear to be more favorable in either beta strand on average (Fig. 4.2), polar and charged residues are well tolerated when oriented correctly.

On a more granular level, we further surveyed for differences in amino acid frequencies at specific locations within the larger hairpin motif. In contrast to their average beta strand frequencies, hydrophobic amino acids are also less tolerated at the C-terminal edge of either beta strand regardless of orientation. In their place, aspartic acid and (to a lesser extent) asparagine are over-represented at these loci, with this effect being particularly strong for the N-terminal beta strand where the last residue is one of these two amino acids in nearly 20% of observed hairpins. This frequency is roughly that observed for these two amino acids, on average, in the turn region (Fig. 4.2, Fig. 4.3C), although other common turn and cap-associated residues, namely glycine and proline, do not show an over-representation at these positions. Interestingly, aspartic acid residues at the C-terminal end of either beta strand also correlate with increased frequencies of bulky aromatic amino acids (i.e. tyrosine, tryptophan, and phenylalanine) at the N-terminally adjacent position and a preference for glycine at the first N-terminal strand residue (Supp. Fig. 4.2A).

Although proline showed no enrichment in the average turn region compared to background levels (Fig. 4.2), proline frequencies are slightly higher than background in the first residue of turns with three to four amino acids and substantially higher than background in the second residue of turns with five amino acids (Fig. 4.3C). These findings largely agree with existing evidence on the prevalence and importance of prolines in the beginning of turn regions[198–200] but the nearly four-fold enrichment for residue two prolines in hairpin structures with five amino acid long turn regions when compared to background levels is particularly surprising. In combination with the fact that over half of all fourth residues in five amino acid long turn regions are glycines, these findings suggest that beta hairpins with longer turn regions may have very specific physiochemical requirements that limit amino acid diversity.

4.2.3 Amino acid contacts between strands favor stabilizing interactions

As the overall beta hairpin structure is stabilized by interactions between the two beta strands, we sought to identify enriched amino acid pairings between strands to see if certain interactions were more common than expected. Pairings between residues with similar electrostatic properties, that is two hydrophobic residues or a polar residue and a polar/charged residue, were largely more common than expected (Fig. 4.4, Supp. Fig. 3). This data agrees with our previous findings regarding the grouping amino acids into beta strand faces based on similar physiochemical properties. In a similar vein to the pairing of electrochemically similar residues, oppositely charged residues tended to pair together in electrostatically favorable salt bridges that are known to stabilize protein structures.[201–204] Such salt bridges represented some of the most enriched amino acid pairings.

The most enriched amino acid pairing between beta strands is that of cysteine with itself to create a structurally stabilizing di-sulfide bond. Such pairings are often used to stabilize engineered peptide structures[205,206] and cysteine coupling is so preferential in nature that many organisms possess a proteome-wide bias towards even numbers of cysteine residues.[207]

In contrast to enriched contact pairings, several classes of interactions, typically those between electrochemically dissimilar residues, were observed much less than expected. The low observance of inter-strand contacts between polar/charged and hydrophobic amino acids (Fig. 4.4) is intuitive given the strong repulsive nature between such residues which could destabilize overall protein structure.

4.2.4 Design principles

Taken altogether, our work provides a strong foundation of general principles that can be applied to the design of functionally diverse high throughput beta hairpin libraries (Table

1). First, libraries should seek to incorporate beta strands with amphipathic faces as seen in our analysis of beta strand positional biases (Fig. 4.3A-B). Second, aspartic acid and asparagine should be favored at C-terminal beta strand residues, especially in the beta strand preceding the turn region. Next, proline and glycine should be utilized in residues two and four of five residue turn regions given their overwhelming enrichment in these positions (Fig. 4.3C). Fourth, average secondary-structure amino acid preferences should inform design choices, especially within the turn region. While residues in both hairpin beta strands show positionally specific frequency deviations from secondary structure averages (Fig. 4.2, Fig. 4.3A-B), there is much less deviation within the turn region (Fig. 4.3C). Lastly, stabilizing interactions should be favored between beta strands. Such interactions include salt bridges, disulfide bonds, and the pairing of certain biochemically similar residues (i.e. hydrophobic-hydrophobic and polar-polar pairings) (Fig. 4.4). These simple guidelines are specific enough to inform design choices while flexible enough to allow for applications across broad research areas.

4.3 DISCUSSION

By analyzing the composition of beta hairpin motifs across all proteins within the PDB we have identified key characteristics of this versatile structure. Expanding on existing knowledge of secondary structure biases, we outline the preference for the amphipathic orientation of amino acids within beta strands to create two faces with different physiochemical properties. We further identify key positional preferences for specific amino acids in all regions of the hairpin motif. Lastly, we highlight the importance of stabilizing interactions between residues in the N and C terminal beta strands of the hairpin.

Our results integrate and expand upon existing knowledge of protein amino acid biases and intra-protein interactions to provide a systematic framework and novel insights to describe the beta hairpin motif. We find that stable beta hairpin structures tend to possess site-specific amino acid preferences and to incorporate amphipathic character in both hairpin beta strands. While existing secondary-structure-specific amino acid distributions [186,187] are accurate and informative, such averages prove inadequate to capture the inherent nuances of the beta hairpin motif. For instance, while our analysis finds that an average hairpin beta strand would consist of only hydrophobic residues (Fig. 4.2), a beta hairpin containing two such average strands without any amphipathic character would be statistically improbable (Fig. 4.3A-B) and highly unlikely to fold correctly [179], let alone function biologically [168].

Position-specific amino acid biases need to be considered to help form stable protein structures. Our observation that prolines are less enriched in turn regions (Fig. 4.2) than previously observed [166,192] is perhaps best explained by the extreme position-specific preference of proline residues in turn regions of a given length (Fig. 4.3C). Thus, certain proline residues are enriched within and likely to stabilize hairpin turn regions even though there is no strong trend when averaged across all turn residues. Outside of the turn region, hairpin beta strands also exhibit amino acid biases at key loci as well as a strong proclivity to incorporate stabilizing inter-strand contacts. We find that asparagine and aspartic acid residues are much more common at the C-terminal end of either hairpin beta strand (Fig. 4.3A-B, Supp. Fig. 2). These residues may participate in a beta capping phenomenon to block the continuation of beta structure into a turn region [174]. A beta capping role may also explain our observation of an increased prevalence of bulky aromatic residues preceding terminal aspartic acids (Supp. Fig. 2) as aromatic residues are known to stabilize beta hairpin structures [176,177]. Lastly, appropriate contacts between hairpin beta strands are imperative to provide structural stability. As an example, we identified cysteine pairings as being particularly enriched in beta hairpin substructures (Fig. 4.4). Such pairings have long been used to stabilize engineered peptide structures [205,206], are so preferential in nature that many organisms possess a proteome-wide bias towards even numbers of cysteines [207].

While our analysis of amino acid preferences within beta hairpin secondary structures across the domains of life showed no strong differences (Supp. Fig. 1A) there were some interesting minor trends as well as a notable difference in turn region composition between taxa (Supp. Fig. 1B). Cysteines, which are fairly uncommon across proteins in general, appear twice as often in Eukaryotic beta hairpins than in Prokaryotic or Archaeaotic beta hairpins. This observation agrees with previous data showing the same trend of increasing cysteine occurrence in proteomes of more complex organisms [208–211]. Of greater note is the inverse relationship between polar and negative amino acid propensities within beta hairpin turn regions across taxa. Frequencies for negatively charged amino acids within the turn region decrease from Archaea to Bacteria, Eukarya, and finally Viruses while polar amino acids show the opposite trend. This difference is likely explained by protein adaptations to harsh environments in Archaea/Bacteria [212] that are less commonly encountered by Eukaryotic or viral proteins. This trend is not seen in either beta strand of the hairpin as turn structures are some of the most accessible protein regions [213] and would likely experience more selective pressure in harsh environments than less exposed beta strands.

One major limitation of our approach is that we were only able to establish broad general properties of beta hairpins that might influence overall structure or function. This is in contrast to prior work that has focused on identifying key design factors for specific beta hairpin scaffolds [7,181,198,214] or grouping beta hairpins and related structures into increasingly detailed classifications [178,188,200]. While the PDB dataset that we analyzed could be used to expand upon these highly focused areas of research, the broad applicability of our results would be compromised.

In combination with prior research efforts, our simple design guidelines (Table 1) can be adapted to the creation of large-scale protein or peptide libraries aimed at almost any functional purpose, from anticancer drugs to biosensors. For example, beta hairpin antimicrobial peptides are known to incorporate multiple disulfide bonds and favor an overall net positive charge while still maintaining amphipathic character [168,171]. Adapting our design principles with these properties in mind would facilitate the construction of a library of positively charged, disulfide stabilized peptides with presumptive beta hairpin structure to test for antimicrobial activity.

In summary, our findings are broadly adaptable to creating large libraries of beta hairpin containing molecules skewed towards a specific functionality and will help engineering efforts keep pace with the ever-expanding capacity of screening assays.

4.4 MATERIALS AND METHODS

4.4.1 Identification of beta hairpin substructures

We defined the beta hairpin motif as an amino acid sequence containing two sets of four to fourteen extended beta strand residues joined by one to five turn, bend, or unannotated residues. A maximum beta strand length of fourteen was selected based on the typical length of beta strands in monomeric beta barrel proteins [215] while the range of turn lengths was selected based on prior research into beta hairpins [175]. We searched DSSP [189] derived secondary structure annotations of all PDB proteins (downloaded from https://cdn.rcsb.org/etl/kabschSander/ss.txt.gz on July 22nd 2020) for this motif. We further filtered our dataset to include only IDs for representative structures clustered to within 90% sequence identity. Clusters were obtained from PDB on July 22nd 2020 using the RESTful Web Service Interface (https://www.rcsb.org/pdb/software/rest.do). Further manual filtering was applied to exclude redundant and overly similar hairpin sequences, largely from structures of nanobodies, antibodies, and their derivatives.

4.4.2 Identification of contacting residues

To ensure that our analyzed motifs possessed the correct beta hairpin 3D structure, we filtered our dataset to only include structures in which at least four amino acid side chain pairs formed contacts between the N and C terminal beta strands. We defined contacts as any pair of residues in which side-chain beta carbons were within 8 Angstroms of one another. Determining contacts via the presence of backbone hydrogen bonds produced similar results (data not included). To calculate expected contact frequencies, individual amino acid frequencies were derived using the relative occurrence of each amino acid across all contact pairs. Values for amino acids in a pairing were then multiplied together to establish an expected frequency for every possible pairing of amino acids.

4.4.3 Grouping of beta hairpin substructures

To characterize the amphipathic faces of each beta strand, solvent accessibility was averaged across odd and even numbered amino acid residues with the first amino acid being the residue closest to the turn region. Strands in which the odd amino acid residues have a higher mean accessibility were categorized as polar while strands with the opposite phenotype were categorized as hydrophobic. Solvent accessibility was chosen in lieu of hydrophobicity or other metrics as PDB structures contain accessibility information and solvent accessibility is known to correlate with hydrophobicity [213].

4.4.4 Data and figures

All data was analyzed in R using the tidyverse family of packages [216] in combination with the data.table [217] and seqinr [218] packages. All figures were created using ggplot2 [219] and cowplot [220]. Supplementary Figure 3 additionally utilized the ggseqlogo package [221]. All processed data, scripts, and Supplementary Files are available at https://doi.org/10.5281/zenodo.4069580.

4.5 FIGURES

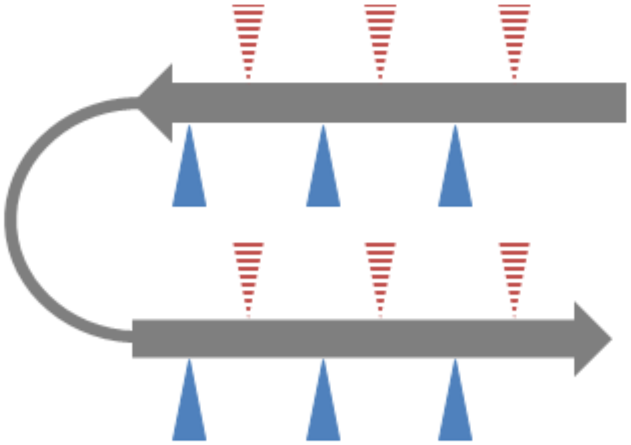

Figure 4.1: General beta hairpin structure.

Beta hairpins consist of two anti-parallel beta strands (grey arrows) linked with a flexible turn region (grey line). Beta strands typically have amphipathic characteristics conferred by alternating hydrophobic and hydrophilic residues. Triangles represent amino acid side chains for the beta strands, with red indicating hydrophobic and blue indicating hydrophilic residues. Solid triangles indicate side chains that are oriented towards the viewer while dashed lines indicate side chains with the opposite orientation.

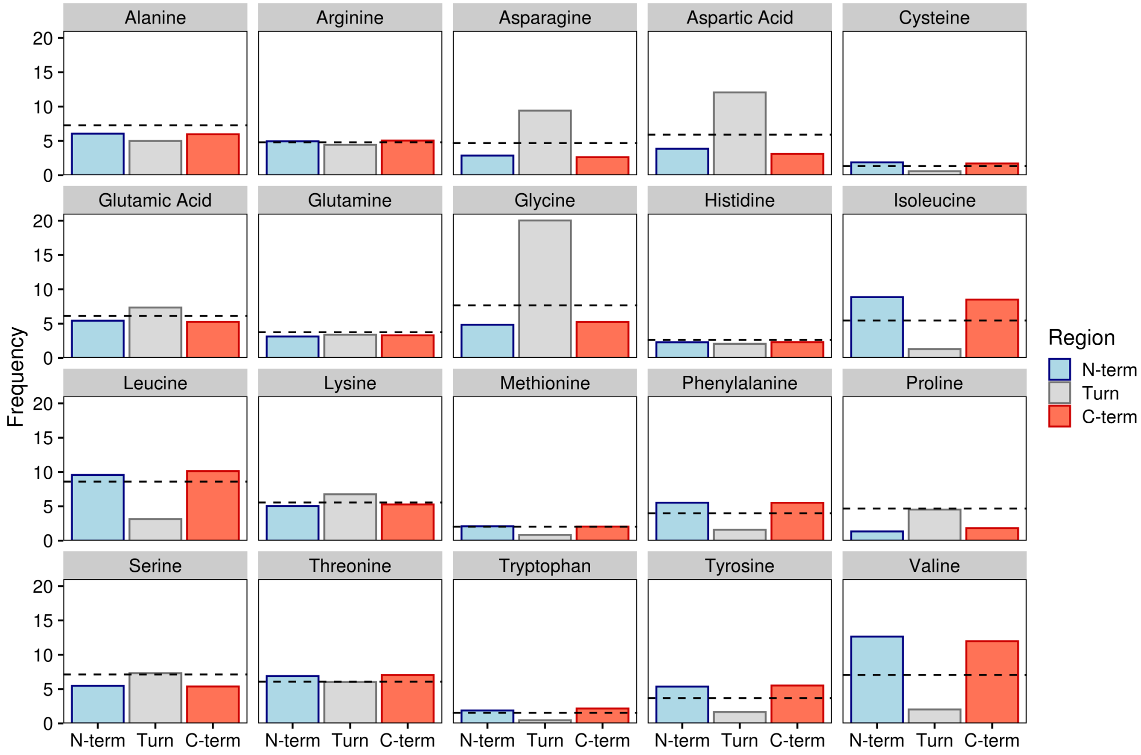

Figure 4.2: Amino acid frequencies by beta hairpin secondary structure region.

Bars indicate average amino acid frequencies for each amino acid within a given region of all beta hairpins. The black dashed line indicates background amino acid frequencies for all sites in all proteins containing the beta hairpin motif. N-term and C-term refer to the N- and C-terminal beta strands while turn denotes the turn region.

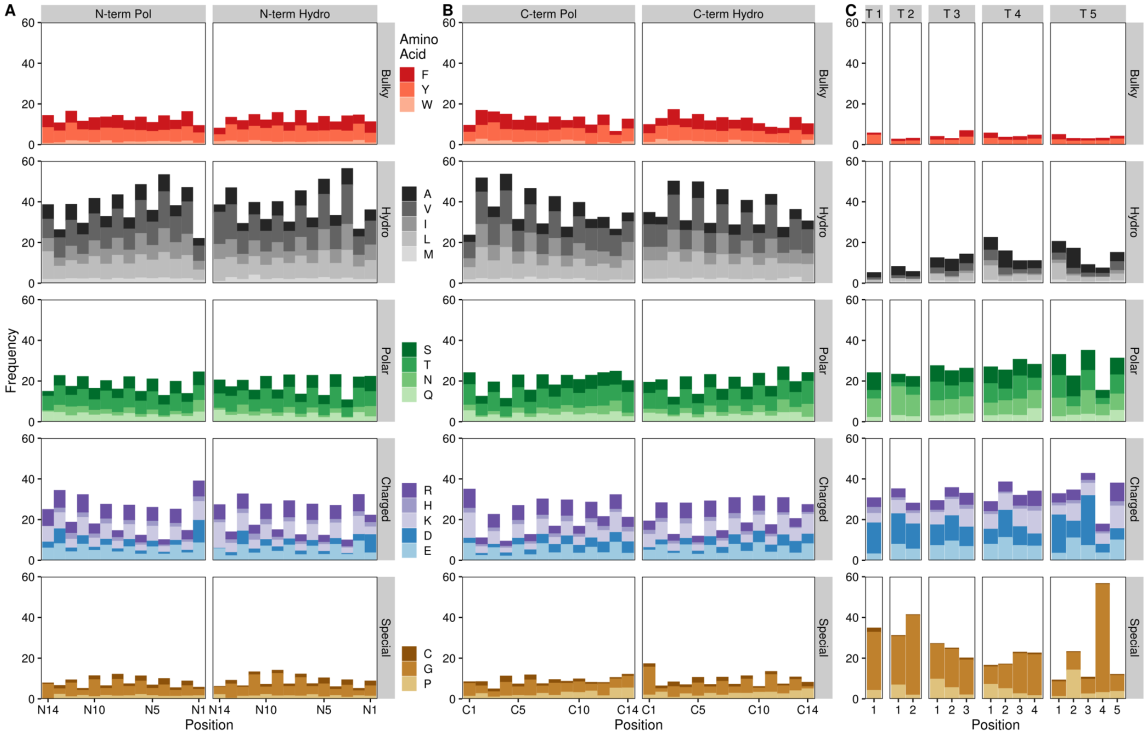

Figure 4.3: Amino acid frequencies by beta hairpin residue position.

Bars indicate average amino acid frequencies for each amino acid at a given position across all beta hairpin structures. N-term and C-term refer to the N- and C-terminal beta strands while T # denotes a turn region of a given length (e.g. T 3 indicates a three residue turn region). Pol refers to beta strands containing a polar face adjacent to the turn region, Hydro denotes a hydrophobic face at this position. Beta strand residues are numbered from the turn region, with residue 1 representing the residue closest to the turn. Turn residues are numbered from N-terminal (residue 1) to C-terminal.

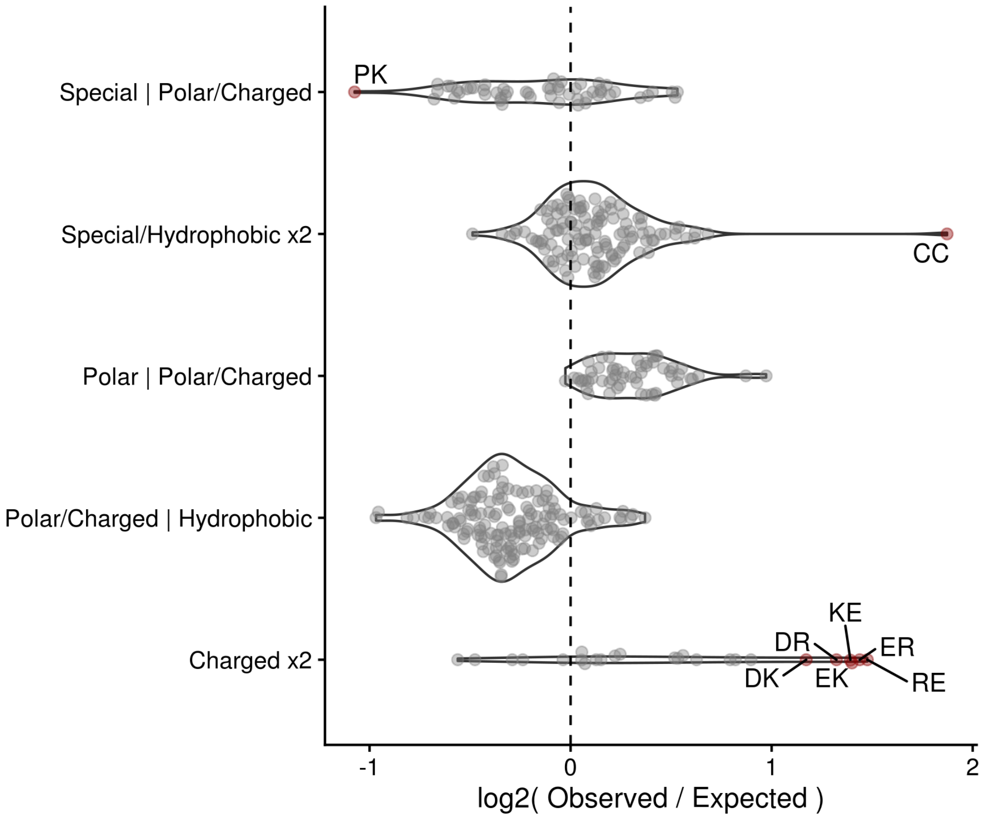

Figure 4.4: Grouped differences in observed vs. expected residue contacts.

Dots represent individual contacting pairs with red, labelled dots indicating contacts that are enriched or depleted at least two-fold vs. expected values. Residues are grouped as follows: Special refers to cysteine, proline, glycine; Hydrophobic refers to valine, leucine, isoleucine, methionine, alanine, tryptophan, tyrosine, phenylalanine; Polar refers to glutamine, threonine, serine, asparagine; Charged refers to arginine, histidine, lysine, aspartic acid, and glutamic acid.

4.6 Tables

Table 4.1: Beta hairpin design principles

Design principles
1. Incorporate amphipathic beta strand faces
2. Favor aspartic acid/asparagine at C-terminal beta strand residues
3. Incorporate T2 proline and T4 glycine in five residue turns
4. Account for secondary structure biases, especially in the turn region
5. Favor salt bridges and di-cysteine interactions to provide stability

Chapter 5: Conclusion

Here I have described applications of large, publicly available datasets to the analysis of microbes and the characterization of a protein structural motif relevant to microbe-focused therapeutics. I began with a thorough evaluation of nomenclature standards and sequence passaging annotation practices for influenza samples across three databases. This work highlighted major shortcomings in existing procedures and outline several improvements that should be implemented to improve the validity of influenza sample metadata moving forward. I next developed a gene coexpression network that provides data on interactions across the entire genome of *Vibrio cholerae*. The resulting network is an invaluable tool that not only contextualizes genes of unknown function but also provides fuel for novel hypothesis generation and presents a cost-effective analogue to sequencing-based transcriptomics approaches. Lastly, by analyzing all beta hairpin-containing protein structures contained in the Protein Data Bank I successfully identified key components of a biologically common and therapeutically relevant structural motif. I ultimately established a set of clear and simple design principles applicable to the development of high throughput protein libraries that incorporate beta hairpin substructures. Taken altogether, this work underscores the value of biological big data.

5.1 Future Directions

My work characterizing influenza annotation best practices is an important step towards better metadata collection efforts in the future. This is of particular importance as databases such as GISAID are expanded to host novel virus data such as that from the ongoing COVID-19 pandemic. The accurate and consistent nomenclature practices I outlined in Chapter 2 are just as applicable to novel viruses as they are to influenza.

Ongoing work within our own group is focused on replicating my coexpression network analysis on a dozen other bacteria and collecting the results in an accessible webserver. My ultimate hope is that this dataset can be regularly updated as transcriptomics data continues to accumulate for more and more bacterial species.

My analysis of the beta hairpin motif provides an outline for the in-depth characterization of other protein structural motifs as well as the foundation for the design of a multitude of high throughput protein libraries. Regarding the former, my analysis pipeline could be easily altered to identify other motifs of interest based on secondary structure and/or amino acid contacts. Concerning the design of protein libraries, our group is currently testing the antimicrobial efficacy of two such beta hairpin inspired libraries and I foresee other researchers adopting the design principles outlined above as well.

Bibliography

1. Ziarek, J. J., Baptista, D. & Wagner, G. Recent developments in solution nuclear magnetic resonance (NMR)-based molecular biology. *J. Mol. Med.* **96**, 1–8 (2018).
2. Nagana Gowda, G. A. & Raftery, D. Recent Advances in NMR-Based Metabolomics. *Anal. Chem.* **89**, 490–510 (2017).
3. Blümich, B. Introduction to compact NMR: A review of methods. *TrAC Trends Anal. Chem.* **83**, 2–11 (2016).
4. Lössl, P., van de Waterbeemd, M. & Heck, A. J. The diverse and expanding role of mass spectrometry in structural and molecular biology. *EMBO J.* **35**, 2634–2657 (2016).
5. Couvillion, S. P. *et al.* New mass spectrometry technologies contributing towards comprehensive and high throughput omics analyses of single cells. *Analyst* **144**, 794–807 (2019).
6. Swedlow, J. R. Innovation in biological microscopy: current status and future directions. *Bioessays* **34**, 333–340 (2012).
7. Di Natale, C. *et al.* Engineered β-hairpin scaffolds from human prion protein regions: Structural and functional investigations of aggregates. *Bioorg. Chem.* **96**, (2020).
8. Wu, Y. & Shroff, H. Faster, sharper, and deeper: structured illumination microscopy for biological imaging. *Nat. Methods* **15**, 1011–1019 (2018).
9. Tripathi, R., Sharma, P., Chakraborty, P. & Varadwaj, P. K. Next-generation sequencing revolution through big data analytics. *Front. Life Sci.* **9**, 119–149 (2016).
10. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27–38 (2013).
11. Hood, L. & Rowen, L. The Human Genome Project: big science transforms biology and medicine. *Genome Med.* **5**, 79 (2013).
12. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
13. Schwarze, K., Buchanan, J., Taylor, J. C. & Wordsworth, S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet. Med.* **20**, 1122–1130 (2018).
14. Kulski, J. K. K. E.-J. K. Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. in Ch. 1 (IntechOpen, 2016). doi:10.5772/61964.
15. Goldfeder, R. L., Wall, D. P., Khoury, M. J., Ioannidis, J. P. A. & Ashley, E. A. Human Genome Sequencing at the Population Scale: A Primer on High-Throughput DNA Sequencing and Analysis. *Am. J. Epidemiol.* **186**, 1000–1009 (2017).
16. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017).

17. Leinonen, R., Sugawara, H., Shumway, M. & Collaboration, I. N. S. D. The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
18. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980 (2003).
19. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **46**, D41–D47 (2018).
20. Mutreja, A. & Dougan, G. Molecular epidemiology and intercontinental spread of cholera. *Vaccine* **38**, A46–A51 (2020).
21. Piña-Iturbe, A. *et al.* Comparative and phylogenetic analysis of a novel family of Enterobacteriaceae-associated genomic islands that share a conserved excision/integration module. *Sci. Rep.* **8**, 10292 (2018).
22. Zhuang, Q. *et al.* Diversity and distribution of type A influenza viruses: an updated panorama analysis based on protein sequences. *Virol. J.* **16**, 85 (2019).
23. Eaaswarkhanth, M., Al Madhoun, A. & Al-Mulla, F. Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? *Int. J. Infect. Dis.* **96**, 459–460 (2020).
24. Almeida, A., Mitchell, A. L., Tarkowska, A. & Finn, R. D. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *Gigascience* **7**, giy054 (2018).
25. Aleti, G. *et al.* Identification of the Bacterial Biosynthetic Gene Clusters of the Oral Microbiome Illuminates the Unexplored Social Language of Bacteria during Health and Disease. *MBio* **10**, e00321-19 (2019).
26. Donia, M. S. *et al.* A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **158**, 1402–1414 (2014).
27. Sugimoto, Y. *et al.* A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. *Science (80-.).* **366**, eaax9176 (2019).
28. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
29. Ugarte, A., Vicedomini, R., Bernardes, J. & Carbone, A. A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. *Microbiome* **6**, 149 (2018).
30. Martinez, X. *et al.* MetaTrans: an open-source pipeline for metatranscriptomics. *Sci. Rep.* **6**, 26447 (2016).
31. Petersen, T. N. *et al.* MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads. *PLoS One* **12**, e0176469–e0176469 (2017).
32. Tsui, B., Dow, M., Skola, D. & Carter, H. Extracting allelic read counts from 250,000 human sequencing runs in Sequence Read Archive. *Pac. Symp. Biocomput.* **24**, 196–207 (2019).
33. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
34. Sweeney, T. E., Haynes, W. A., Vallania, F., Ioannidis, J. P. & Khatri, P. Methods to increase reproducibility in differential gene expression via meta-analysis.

Nucleic Acids Res. **45**, e1–e1 (2017).

35. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
36. Croll, T. I., Sammito, M. D., Kryshtafovych, A. & Read, R. J. Evaluation of template-based modeling in CASP13. *Proteins Struct. Funct. Bioinforma.* **87**, 1113–1127 (2019).
37. Trevizani, R., Dio, F. L. C., Santos, K. B. Dos & Dardenne, L. E. Critical features of fragment libraries for protein structure prediction. *PLoS One* **12**, 1–22 (2017).
38. Guzenko, D., Burley, S. K. & Duarte, J. M. Real time structural search of the Protein Data Bank. *PLOS Comput. Biol.* **16**, e1007970 (2020).
39. Deng, L., Zhong, G., Liu, C., Luo, J. & Liu, H. MADOKA: an ultra-fast approach for large-scale protein structure similarity searching. *BMC Bioinformatics* **20**, 662 (2019).
40. Ayoub, R. & Lee, Y. RUPEE: A fast and accurate purely geometric protein structure search. *PLoS One* **14**, e0213712 (2019).
41. Zhou, J., Panaitiu, A. E. & Grigoryan, G. A general-purpose protein design framework based on mining sequence–structure relationships in known protein structures. *Proc. Natl. Acad. Sci.* **117**, 1059 LP – 1068 (2020).
42. Pearce, R., Huang, X., Setiawan, D. & Zhang, Y. EvoDesign: Designing Protein-Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function. *J. Mol. Biol.* **431**, 2467–2476 (2019).
43. McWhite, C. D., Meyer, A. G. & Wilke, C. O. Sequence amplification via cell passaging creates spurious signals of positive adaptation in influenza virus H3N2 hemagglutinin. *Virus Evol.* **2**, vew026 (2016).
44. Ozenne, V. *et al.* Mapping the Potential Energy Landscape of Intrinsically Disordered Proteins at Amino Acid Resolution. *J. Am. Chem. Soc.* **134**, 15138–15148 (2012).
45. Leber, M. F. *et al.* Sequencing of serially passaged measles virus affirms its genomic stability and reveals a nonrandom distribution of consensus mutations. *J. Gen. Virol.* **101**, 399–409 (2020).
46. Choi, S. Y. *et al.* Phylogenetic Diversity of Vibrio cholerae Associated with Endemic Cholera in Mexico from 1991 to 2008. *MBio* **7**, e02160-15 (2016).
47. Wang, H. *et al.* Genomic epidemiology of Vibrio cholerae reveals the regional and global spread of two epidemic non-toxigenic lineages. *PLoS Negl. Trop. Dis.* **14**, e0008046 (2020).
48. Lu, L., Lycett, S. J. & Leigh Brown, A. J. Determining the Phylogenetic and Phylogeographic Origin of Highly Pathogenic Avian Influenza (H7N3) in Mexico. *PLoS One* **9**, e107330 (2014).
49. Mokrousov, I. *et al.* Latin-American-Mediterranean lineage of Mycobacterium tuberculosis: Human traces across pathogen’s phylogeography. *Mol. Phylogenet. Evol.* **99**, 133–143 (2016).
50. Weber Zendrera, A., Sokolovska, N. & Soula, H. A. Robust structure measures of

metabolic networks that predict prokaryotic optimal growth temperature. *BMC Bioinformatics* **20**, 499 (2019).

51. Reimer, L. C. *et al.* BacDive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Res.* **47**, D631–D636 (2019).
52. Ampofo, W. K. *et al.* Improving influenza vaccine virus selection: Report of a WHO informal consultation held at WHO headquarters, Geneva, Switzerland, 14-16 June 2010. *Influenza Other Respi. Viruses* **7**, 52–53 (2013).
53. Timofeeva, T. A. *et al.* Predicting the Evolutionary Variability of the Influenza A Virus. *Acta Naturae* **9**, 48–54 (2017).
54. Anderson, T. K. *et al.* A Phylogeny-Based Global Nomenclature System and Automated Annotation Tool for H1 Hemagglutinin Genes from Swine Influenza A Viruses. *mSphere* **1**, e00275-16 (2016).
55. Zhang, Y. *et al.* Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Res.* **45**, D466–D474 (2017).
56. Belanov, S. S. *et al.* Genome-wide analysis of evolutionary markers of human influenza A(H1N1)pdm09 and A(H3N2) viruses may guide selection of vaccine strain candidates. *Genome Biol. Evol.* **7**, 3472–3483 (2015).
57. Du, X., King, A. A., Woods, R. J. & Pascual, M. Evolution-informed forecasting of seasonal influenza A (H3N2). *Sci. Transl. Med.* **9**, eaan5325 (2017).
58. Moncla, L. H., Florek, N. W. & Friedrich, T. C. Influenza Evolution: New Insights into an Old Foe. *Trends Microbiol.* **25**, 432–434 (2017).
59. Gatherer, D. Passage in egg culture is a major cause of apparent positive selection in influenza B hemagglutinin. *J. Med. Virol.* **82**, 123–127 (2010).
60. Lee, H. K. *et al.* Comparison of mutation patterns in full-genome a/H3N2 influenza sequences obtained directly from clinical samples and the same samples after a single MDCK passage. *PLoS One* **8**, 1–9 (2013).
61. Chen, H. *et al.* Dynamic convergent evolution drives the passage adaptation across 48 years' history of H3N2 influenza evolution. *Mol. Biol. Evol.* **33**, 3133–3143 (2016).
62. Wyde, P. R., Couch, R. B., Mackler, B. F., Cate, T. R. & Levy, B. M. Effects of low- and high-passage influenza virus infection in normal and nude mice. *Infect. Immun.* **15**, 221–229 (1977).
63. Robertson, J. S., Nicolson, C., Major, D., Robertson, E. W. & Wood, J. M. The role of amniotic passage in the egg-adaptation of human influenza virus is revealed by haemagglutinin sequence analyses. *J. Gen. Virol.* **74**, 2047–2051 (1993).
64. Bush, R. M., Smith, C. B., Cox, N. J. & Fitch, W. M. Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 6974–80 (2000).
65. Lin, Y. *et al.* The characteristics and antigenic properties of recently emerged subclade 3C . 3a and 3C . 2a human influenza A (H3N2) viruses passaged in MDCK cells. *Influenza Other Respi. Viruses* 263–274 (2017) doi:10.1111/irv.12447.
66. Li, D. *et al.* In Vivo and In Vitro Alterations in Influenza A / H3N2 Virus M2 and

Hemagglutinin Genes : Effect of Passage in MDCK-SIAT1 Cells and Conventional MDCK Cells. *J. Clin. Microbiol.* **47**, 466–468 (2009).

67. Govorkova, E. A. *et al.* Selection of Receptor-Binding Variants of Human Influenza A and B Viruses in Baby Hamster Kidney Cells . *Virology* **38**, 31–38 (1999).

68. Little, K. *et al.* Zanamivir-resistant influenza viruses with Q136K or Q136R neuraminidase residue mutations can arise during MDCK cell culture creating challenges for antiviral susceptibility monitoring. *Euro Surveill.* **20**, 1–9 (2015).

69. Wu, N. C. *et al.* A structural explanation for the low effectiveness of the seasonal influenza H3N2 vaccine. *PLoS Pathog.* **13**, 1–17 (2017).

70. Zost, S. J. *et al.* Contemporary H3N2 influenza viruses have a glycosylation site that alters binding of antibodies elicited by egg-adapted vaccine strains. *Proc. Natl. Acad. Sci.* **114**, 201712377 (2017).

71. Chen, H., Alvarez, J. J. S., Ng, S. H., Nielsen, R. & Zhai, W. Passage adaptation correlates with the reduced efficacy of the influenza vaccine. *Clin. Infect. Dis.* **67**, ciy1065–ciy1065 (2018).

72. Himmelsbach, K. & Hildt, E. Identification of various cell culture models for the study of Zika virus. *World J. Virol.* **7**, 10–20 (2018).

73. Broadhurst, M. J., Brooks, T. J. G. & Pollock, N. R. Diagnosis of Ebola Virus Disease: Past, Present, and Future. *Clin. Microbiol. Rev.* **29**, 773 LP – 793 (2016).

74. Krowicka, H. *et al.* Use of tissue culture cell lines to evaluate HIV antiviral resistance. *AIDS Res. Hum. Retroviruses* **24**, 957–967 (2008).

75. Eisfeld, A. J., Neumann, G. & Kawaoka, Y. Influenza A virus isolation, culture and identification. *Nat. Protoc.* **9**, 2663–2681 (2014).

76. Krauss, S., Walker, D. & Webster, R. G. Influenza Virus Isolation. in *Influenza Virus: Methods and Protocols* (eds. Kawaoka, Y. & Neumann, G.) 11–24 (Humana Press, 2012). doi:10.1007/978-1-61779-621-0_2.

77. Hierholzer, J. C. & Killington, R. A. Virus isolation and quantitation. in *Virology Methods Manual* 25–46 (1996). doi:10.1016/B978-012465330-6/50003-8.

78. Pedersen, J. C. Hemagglutination-inhibition assay for influenza virus subtype identification and the detection and quantitation of serum antibodies to influenza virus. *Methods Mol. Biol.* **1161**, 11–25 (2014).

79. Weir, J. P. & Gruber, M. F. An overview of the regulation of influenza vaccines in the United States. *Influenza Other Respi. Viruses* **10**, 354–360 (2016).

80. Grohskopf, L. A. *et al.* Prevention and Control of Seasonal Influenza with Vaccines: Recommendations of the Advisory Committee on Immunization Practices—United States, 2018–19 Influenza Season. *MMWR. Recomm. Reports* **67**, 1–20 (2018).

81. Heidelberg, J. F. *et al.* DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. *Nature* **406**, 477–483 (2000).

82. Weill, F.-X. *et al.* Genomic insights into the 2016–2017 cholera epidemic in Yemen. *Nature* **565**, 230–233 (2019).

83. Greig, D. R. *et al.* Evaluation of Whole-Genome Sequencing for Identification and

Typing of Vibrio cholerae. *J. Clin. Microbiol.* **56**, e00831-18 (2018).
84. Domman, D. *et al.* Defining endemic cholera at three levels of spatiotemporal resolution within Bangladesh. *Nat. Genet.* **50**, 951–955 (2018).
85. Weill, F.-X. *et al.* Genomic history of the seventh pandemic of cholera in Africa. *Science (80-.).* **358**, 785–789 (2017).
86. Domman, D. *et al.* Integrated view of Vibrio cholerae in the Americas. *Science* **358**, 789–793 (2017).
87. Li, Z. *et al.* Expanding dynamics of the virulence-related gene variations in the toxigenic Vibrio cholerae serogroup O1. *BMC Genomics* **20**, 360 (2019).
88. Rahman, M. H. *et al.* Distribution of genes for virulence and ecological fitness among diverse Vibrio cholerae population in a cholera endemic area: tracking the evolution of pathogenic strains. *DNA Cell Biol.* **27**, 347–355 (2008).
89. Lessler, J. *et al.* Mapping the burden of cholera in sub-Saharan Africa and implications for control: an analysis of data across geographical scales. *Lancet (London, England)* **391**, 1908–1915 (2018).
90. WHO | Ending Cholera. *WHO* (2017).
91. Herzog, R., Peschek, N., Fröhlich, K. S., Schumacher, K. & Papenfort, K. Three autoinducer molecules act in concert to control virulence gene expression in Vibrio cholerae. *Nucleic Acids Res.* **47**, 3171–3183 (2019).
92. Davies, B. W., Bogard, R. W., Young, T. S. & Mekalanos, J. J. Coordinated Regulation of Accessory Genetic Elements Produces Cyclic Di-Nucleotides for V. cholerae Virulence. *Cell* **149**, 358–370 (2012).
93. Krin, E. *et al.* Expansion of the SOS regulon of Vibrio cholerae through extensive transcriptome analysis and experimental validation. *BMC Genomics* **19**, 373 (2018).
94. Fu, Y., Waldor, M. K. & Mekalanos, J. J. Tn-Seq Analysis of Vibrio cholerae Intestinal Colonization Reveals a Role for T6SS-Mediated Antibacterial Activity in the Host. *Cell Host Microbe* **14**, 652–663 (2013).
95. Mandlik, A. *et al.* RNA-Seq-based monitoring of infection-linked changes in Vibrio cholerae gene expression. *Cell Host Microbe* **10**, 165–174 (2011).
96. Kamp, H. D., Patimalla-Dipali, B., Lazinski, D. W., Wallace-Gadsden, F. & Camilli, A. Gene Fitness Landscapes of Vibrio cholerae at Important Stages of Its Life Cycle. *PLoS Pathog.* **9**, e1003800 (2013).
97. Pukatzki, S. *et al.* Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system. *Proc. Natl. Acad. Sci.* **103**, 1528 LP – 1533 (2006).
98. Kimura, S., Hubbard, T. P., Davis, B. M. & Waldor, M. K. The Nucleoid Binding Protein H-NS Biases Genome-Wide Transposon Insertion Landscapes. *MBio* **7**, e01351-16 (2016).
99. Manneh-Roussel, J. *et al.* cAMP Receptor Protein Controls Vibrio cholerae Gene Expression in Response to Host Colonization. *MBio* **9**, e00966-18 (2018).
100. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* **9**, 1090 (2018).

101. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

102. Jiang, J. *et al.* Construction and application of a co-expression network in Mycobacterium tuberculosis. *Sci. Rep.* **6**, 28422 (2016).

103. Liu, W. *et al.* Construction and Analysis of Gene Co-Expression Networks in Escherichia coli. *Cells* **7**, 19 (2018).

104. Li, A. & Horvath, S. Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics* **23**, 222–231 (2006).

105. Geer, L. Y. *et al.* The NCBI BioSystems database. *Nucleic Acids Res.* **38**, D492–D496 (2009).

106. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44 (2008).

107. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

108. Mi, H. *et al.* PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.* **38**, D204–D210 (2009).

109. Taboada, B., Estrada, K., Ciria, R. & Merino, E. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. *Bioinformatics* **34**, 4118–4120 (2018).

110. Chao, M. C. *et al.* High-resolution definition of the Vibrio cholerae essential gene set with hidden Markov model-based analyses of transposon-insertion sequencing data. *Nucleic Acids Res.* **41**, 9033–9048 (2013).

111. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

112. Rutherford, S. T., van Kessel, J. C., Shao, Y. & Bassler, B. L. AphA and LuxR/HapR reciprocally control quorum sensing in vibrios. *Genes Dev.* **25**, 397–408 (2011).

113. Haycocks, J. R. J. *et al.* The quorum sensing transcription factor AphA directly regulates natural competence in Vibrio cholerae. *PLOS Genet.* **15**, e1008362 (2019).

114. Zhang, Y. *et al.* Transcriptional Regulation of the Type VI Secretion System 1 Genes by Quorum Sensing and ToxR in Vibrio parahaemolyticus. *Front. Microbiol.* **8**, 2005 (2017).

115. Yamamoto, S. *et al.* Regulation of natural competence by the orphan two-component system sensor kinase ChiS involves a non-canonical transmembrane regulator in Vibrio cholerae. *Mol. Microbiol.* **91**, 326–347 (2014).

116. Dalia, A. B., Lazinski, D. W. & Camilli, A. Identification of a membrane-bound transcriptional regulator that links chitin and natural competence in Vibrio cholerae. *MBio* **5**, e01028 (2014).

117. Metzger, L. C. *et al.* Independent Regulation of Type VI Secretion in Vibrio cholerae by TfoX and TfoY. *Cell Rep.* **15**, 951–958 (2016).

118. van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.* **19**, 575–592 (2017).

119. Silva, A. J. & Benitez, J. A. Vibrio cholerae Biofilms and Cholera Pathogenesis. *PLoS Negl. Trop. Dis.* **10**, e0004330 (2016).

120. Teschler, J. K. *et al.* Living in the matrix: assembly and control of Vibrio cholerae biofilms. *Nat. Rev. Microbiol.* **13**, 255–68 (2015).

121. Papenfort, K., Förstner, K. U., Cong, J.-P., Sharma, C. M. & Bassler, B. L. Differential RNA-seq of Vibrio cholerae identifies the VqmR small RNA as a regulator of biofilm formation. *Proc. Natl. Acad. Sci.* **112**, E766–E775 (2015).

122. Weber, G. G., Klose, K. E. & Klose. The complexity of ToxT-dependent transcription in Vibrio cholerae. *Indian J. Med. Res.* **133**, 201–6 (2011).

123. Dorman, M. J. & Dorman, C. J. Regulatory Hierarchies Controlling Virulence Gene Expression in Shigella flexneri and Vibrio cholerae. *Frontiers in Microbiology* vol. 9 2686 (2018).

124. Kazi, M. I., Conrado, A. R., Mey, A. R., Payne, S. M. & Davies, B. W. ToxR Antagonizes H-NS Regulation of Horizontally Acquired Genes to Drive Host Colonization. *PLOS Pathog.* **12**, e1005570 (2016).

125. Ayala, J. C., Wang, H., Silva, A. J. & Benitez, J. A. Repression by H-NS of genes required for the biosynthesis of the Vibrio cholerae biofilm matrix is modulated by the second messenger cyclic diguanylic acid. *Mol. Microbiol.* **97**, 630–645 (2015).

126. Boyd, E. F., Jermyn, W. S. & Boyd, E. F. *Characterization of a novel Vibrio pathogenicity island (VPI-2) encoding neuraminidase (nanH) among toxigenic Vibrio cholerae isolates*. *Microbiology* vol. 148 3681–3693 (Microbiology Society, 2002).

127. Karaolis, D. K. R. *et al.* A Vibrio cholerae pathogenicity island associated with epidemic and pandemic strains. *Proc. Natl. Acad. Sci.* **95**, 3134 LP – 3139 (1998).

128. Nielsen, A. T. *et al.* RpoS controls the Vibrio cholerae mucosal escape response. *PLoS Pathog.* **2**, e109–e109 (2006).

129. Wong, G. T. *et al.* Genome-Wide Transcriptional Response to Varying RpoS Levels in Escherichia coli K-12. *J. Bacteriol.* **199**, e00755-16 (2017).

130. Russo, G., Zegar, C. & Giordano, A. Advantages and limitations of microarray technology in human cancer. *Oncogene* **22**, 6497–6507 (2003).

131. Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M. & Ligterink, W. Learning from Co-expression Networks: Possibilities and Challenges. *Front. Plant Sci.* **7**, 444 (2016).

132. Koschmann, J., Bhar, A., Stegmaier, P., Kel, A. E. & Wingender, E. 'Upstream Analysis': An Integrated Promoter-Pathway Analysis Approach to Causal Interpretation of Microarray Data. *Microarrays (Basel, Switzerland)* **4**, 270–286 (2015).

133. Mueller, R. S. *et al.* Vibrio cholerae Strains Possess Multiple Strategies for Abiotic and Biotic Surface Colonization. *J. Bacteriol.* **189**, 5348–5360 (2007).

134. Marchler-Bauer, A. *et al.* CDD/SPARCLE: Functional classification of proteins

via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203 (2017).

135. Duran-Pinedo, A. E., Paster, B., Teles, R. & Frias-Lopez, J. Correlation Network Analysis Applied to Complex Biofilm Communities. *PLoS One* **6**, e28438 (2011).

136. Geng, H., Tran-Gyamfi, M. B., Lane, T. W., Sale, K. L. & Yu, E. T. Changes in the Structure of the Microbial Community Associated with Nannochloropsis salina following Treatments with Antibiotics and Bioactive Compounds. *Front. Microbiol.* **7**, 1155 (2016).

137. Meisel, J. S. *et al.* Commensal microbiota modulate gene expression in the skin. *Microbiome* **6**, 20 (2018).

138. Jackson, M. A. *et al.* Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. *PeerJ* **6**, e4303 (2018).

139. Hosseinkhan, N., Mousavian, Z. & Masoudi-Nejad, A. Comparison of gene co-expression networks in Pseudomonas aeruginosa and Staphylococcus aureus reveals conservation in some aspects of virulence. *Gene* **639**, 1–10 (2018).

140. Peña-Castillo, L. *et al.* Gene co-expression network analysis in Rhodobacter capsulatus and application to comparative expression analysis of Rhodobacter sphaeroides. *BMC Genomics* **15**, 730 (2014).

141. Wang, J., Wu, G., Chen, L. & Zhang, W. Cross-species transcriptional network analysis reveals conservation and variation in response to metal stress in cyanobacteria. *BMC Genomics* **14**, 112 (2013).

142. Keseler, I. M. *et al.* The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res.* **45**, D543–D550 (2017).

143. Ballouz, S., Verleyen, W. & Gillis, J. Guidance for RNA-seq co-expression network construction and analysis: Safety in numbers. *Bioinformatics* **31**, 2123–2130 (2015).

144. Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.* **32**, 915–925 (2014).

145. Goh, W. W. Bin, Wang, W. & Wong, L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology* vol. 35 498–507 (2017).

146. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).

147. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-Wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 10101–10106 (2000).

148. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* vol. 11 733–739 (2010).

149. Cohen, O. *et al.* Comparative transcriptomics across the prokaryotic tree of life. *Nucleic Acids Res.* **44**, W46–W53 (2016).

150. Chang, Y. M. *et al.* Comparative transcriptomics method to infer gene coexpression networks and its applications to maize and rice leaf transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 3091–3099 (2019).

151. Rodríguez-García, A., Sola-Landa, A. & Barreiro, C. RNA-Seq-Based Comparative Transcriptomics: RNA Preparation and Bioinformatics BT - Microbial Steroids: Methods and Protocols. in (eds. Barredo, J.-L. & Herráiz, I.) 59–72 (Springer New York, 2017). doi:10.1007/978-1-4939-7183-1_5.

152. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

153. Langmean, B. & Salzberg, S. L. Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

154. Gaspar, J. M. Improved peak-calling with MACS2. *bioRxiv* 496521 (2018) doi:10.1101/496521.

155. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389 (2012).

156. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2015).

157. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

158. DuPai, C. D., Wilke, C. O. & Davies, B. W. A Comprehensive Coexpression Network Analysis in Vibrio cholerae. *mSystems* **5**, e00550-20 (2020).

159. Watkins, A. M. & Arora, P. S. Anatomy of β-strands at protein-protein interfaces. *ACS Chem. Biol.* **9**, 1747–1754 (2014).

160. Robinson, J. A. β-Hairpin Peptidomimetics: Design, Structures and Biological Activities. *Acc. Chem. Res.* **41**, 1278–1288 (2008).

161. Chakraborty, K., Shivakumar, P., Raghothama, S. & Varadarajan, R. NMR structural analysis of a peptide mimic of the bridging sheet of HIV-1 gp120 in methanol and water. *Biochem. J.* **390**, 573–581 (2005).

162. Batalha, I. L., Lychko, I., Branco, R. J. F., Iranzo, O. & Roque, A. C. A. β-Hairpins as peptidomimetics of human phosphoprotein-binding domains. *Org. Biomol. Chem.* **17**, 3996–4004 (2019).

163. Remmert, M., Biegert, A., Linke, D., Lupas, A. N. & Söding, J. Evolution of Outer Membrane β-Barrels from an Ancestral ββ Hairpin. *Mol. Biol. Evol.* **27**, 1348–1358 (2010).

164. Chaturvedi, D. & Mahalakshmi, R. Transmembrane β-barrels: Evolution, folding and energetics. *Biochim. Biophys. Acta - Biomembr.* **1859**, 2467–2482 (2017).

165. Gupta, K. *et al.* Mechanism of membrane permeation induced by synthetic β-hairpin peptides. *Biophys. J.* **105**, 2093–2103 (2013).

166. Marcelino, A. M. C., Gierasch, L. M. & Craik, C. Roles of beta-turns in protein folding: from peptide models to protein engineering. *Biopolymers* **89**, 380–391 (2008).

167. Dong, N., Ma, Q., Shan, A. & Cao, Y. Design and biological activity of beta-hairpin-like antimicrobial peptide. *Sheng Wu Gong Cheng Xue Bao* **28**, 243—250 (2012).

168. Edwards, I. A. *et al.* Contribution of amphipathicity and hydrophobicity to the antimicrobial activity and cytotoxicity of β-hairpin peptides. *ACS Infect. Dis.* **2**,

442–450 (2016).

169. Mirecka, E. A. *et al.* Sequestration of a β-Hairpin for Control of α-Synuclein Aggregation. *Angew. Chemie Int. Ed.* **53**, 4227–4230 (2014).

170. Larini, L. & Shea, J. E. Role of β-hairpin formation in aggregation: The self-assembly of the amyloid-β(25-35) peptide. *Biophys. J.* **103**, 576–586 (2012).

171. Panteleev, P. V, Bolosov, I. A., Balandin, S. V & Ovchinnikova, T. V. Structure and Biological Functions of β-Hairpin Antimicrobial Peptides. *Acta Naturae* **7**, 37–47 (2015).

172. Worthington, P., Langhans, S. & Pochan, D. β-hairpin peptide hydrogels for package delivery. *Adv. Drug Deliv. Rev.* **110–111**, 127–136 (2017).

173. Lewandowska, A., Ołdziej, S., Liwo, A. & Scheraga, H. A. β-hairpin-forming peptides; models of early stages of protein folding. *Biophys. Chem.* **151**, 1–9 (2010).

174. FarzadFard, F., Gharaei, N., Pezeshk, H. & Marashi, S. A. β-Sheet capping: Signals that initiate and terminate β-sheet formation. *J. Struct. Biol.* **161**, 101–110 (2008).

175. Gunasekaran, K., Ramakrishnan, C. & Balaram, P. β-Hairpins in proteins revisited: Lessons for de novo design. *Protein Eng.* **10**, 1131–1141 (1997).

176. Santiveri, C. M. & Jiménez, M. A. Tryptophan residues: Scarce in proteins but strong stabilizers of β-hairpin peptides. *Pept. Sci.* **94**, 779–790 (2010).

177. Mahalakshmi, R. Aromatic interactions in β-hairpin scaffold stability: A historical perspective. *Archives of Biochemistry and Biophysics* vol. 661 39–49 (2019).

178. Milner-White, E. J. & Poet, R. Four classes of beta-hairpins in proteins. *Biochem. J.* **240**, 289–292 (1986).

179. Popp, A., Wu, L., Keiderling, T. A. & Hauser, K. Effect of Hydrophobic Interactions on the Folding Mechanism of β-Hairpins. *J. Phys. Chem. B* 118–14234 (2014) doi:10.1021/jp506658x.

180. Rughani, R. V & Schneider, J. P. Molecular Design of beta-Hairpin Peptides for Material Construction. *MRS Bull.* **33**, 530–535 (2008).

181. Cochran, A. G. *et al.* A minimal peptide scaffold for β-turn display: Optimizing a strand position in disulfide-cyclized β-hairpins. *J. Am. Chem. Soc.* **123**, 625–632 (2001).

182. Wu, C.-H., Liu, I.-J., Lu, R.-M. & Wu, H.-C. Advancement and applications of peptide phage display technology in biomedical science. *J. Biomed. Sci.* **23**, 8 (2016).

183. Tucker, A. T. *et al.* Discovery of Next-Generation Antimicrobials through Bacterial Self-Screening of Surface-Displayed Peptide Libraries. *Cell* **172**, 618-621.e13 (2018).

184. Wójcik, M., Telzerow, A., Quax, W. J. & Boersma, Y. L. High-Throughput Screening in Protein Engineering: Recent Advances and Future Perspectives. *Int. J. Mol. Sci.* **16**, 24918–24945 (2015).

185. Bhattacharjee, N. & Biswas, P. Position-specific propensities of amino acids in the -strand. *BMC Struct. Biol.* **10**, 1–10 (2010).

186. Fujiwara, K., Toda, H. & Ikeguchi, M. Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. *BMC Struct. Biol.* **12**, 18 (2012).

187. Otaki, J. M., Tsutsumi, M., Gotoh, T. & Yamamoto, H. Secondary Structure Characterization Based on Amino Acid Composition and Availability in Proteins. *J. Chem. Inf. Model.* **50**, 690–700 (2010).

188. Sibanda, B. L., Blundell, T. L. & Thornton, J. M. Conformation of β-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J. Mol. Biol.* **206**, 759–777 (1989).

189. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

190. Koehl, P. & Levitt, M. Structure-based conformational preferences of amino acids. *Proc. Natl. Acad. Sci.* **96**, 12524 LP – 12529 (1999).

191. Geisow, M. J. & Roberts, R. D. B. Amino acid preferences for secondary structure vary with protein class. *Int. J. Biol. Macromol.* **2**, 387–389 (1980).

192. Trevino, S. R., Schaefer, S., Scholtz, J. M. & Pace, C. N. Increasing protein conformational stability by optimizing beta-turn sequence. *J. Mol. Biol.* **373**, 211–218 (2007).

193. Melnikov, S. *et al.* Molecular insights into protein synthesis with proline residues. *EMBO Rep.* **17**, 1776–1784 (2016).

194. Tsutsumi, M. & Otaki, J. M. Parallel and antiparallel beta-strands differ in amino acid composition and availability of short constituent sequences. *J Chem Inf Model* **51**, (2011).

195. Bowerman, C. J. & Nilsson, B. L. Self-assembly of amphipathic β-sheet peptides: insights and applications. *Biopolymers* **98**, 169–184 (2012).

196. Sivanesam, K., Kier, B. L., Whedon, S. D., Chatterjee, C. & Andersen, N. H. Hairpin structure stability plays a role in the activity of two antimicrobial peptides. *FEBS Lett.* **590**, 4480–4488 (2016).

197. Zhang, X. C. & Han, L. How does a β-barrel integral membrane protein insert into the membrane? *Protein Cell* **7**, 471–477 (2016).

198. Blandl, T., Cochran, A. G. & Skelton, N. J. Turn stability in beta-hairpin peptides: Investigation of peptides containing 3:5 type I G1 bulge turns. *Protein Sci.* **12**, 237–247 (2003).

199. Wang, H., Varady, J., Ng, L. & Sung, S.-S. Molecular dynamics simulations of β-hairpin folding. *Proteins Struct. Funct. Bioinforma.* **37**, 325–333 (1999).

200. Shapovalov, M. I., Vucetic, S. & Dunbrack, R. L. A new clustering and nomenclature for beta turns derived from high-resolution protein structures. (2019) doi:10.1371/journal.pcbi.1006844.

201. Pylaeva, S., Brehm, M. & Sebastiani, D. Salt Bridge in Aqueous Solution: Strong Structural Motifs but Weak Enthalpic Effect. *Sci. Rep.* **8**, 13626 (2018).

202. Ciani, B., Jourdan, M. & Searle, M. S. Stabilization of β-Hairpin Peptides by Salt Bridges: Role of Preorganization in the Energetic Contribution of Weak

Interactions. *J. Am. Chem. Soc.* **125**, 9038–9047 (2003).

203. Donald, J. E., Kulp, D. W. & DeGrado, W. F. Salt bridges: geometrically specific, designable interactions. *Proteins* **79**, 898–915 (2011).

204. Bosshard, H. R., Marti, D. N. & Jelesarov, I. Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. *J. Mol. Recognit.* **17**, 1–16 (2004).

205. Sussman, D. *et al.* Engineered cysteine antibodies: an improved antibody-drug conjugate platform with a novel mechanism of drug-linker stability. *Protein Eng. Des. Sel.* **31**, 47–54 (2018).

206. Dombkowski, A. A., Sultana, K. Z. & Craig, D. B. Protein disulfide engineering. *FEBS Lett.* **588**, 206–212 (2014).

207. Dutton, R. J., Boyd, D., Berkmen, M. & Beckwith, J. Bacterial species exhibit diversity in their mechanisms and capacity for protein disulfide bond formation. *Proc. Natl. Acad. Sci.* **105**, 11933–11938 (2008).

208. Brooks, D. J. & Fresco, J. R. Increased Frequency of Cysteine, Tyrosine, and Phenylalanine Residues Since the Last Universal Ancestor. *Mol. &amp;amp; Cell. Proteomics* **1**, 125 LP – 131 (2002).

209. Wiedemann, C., Kumar, A., Lang, A. & Ohlenschläger, O. Cysteines and Disulfide Bonds as Structure-Forming Units: Insights From Different Domains of Life and the Potential for Characterization by NMR . *Frontiers in Chemistry* vol. 8 280 (2020).

210. Miseta, A. & Csutora, P. Relationship Between the Occurrence of Cysteine in Proteins and the Complexity of Organisms. *Mol. Biol. Evol.* **17**, 1232–1239 (2000).

211. Tsuji, J. *et al.* The frequencies of amino acids encoded by genomes that utilize standard and nonstandard genetic codes. *Bios* **81**, 22–31 (2010).

212. Reed, C. J., Lewis, H., Trejo, E., Winston, V. & Evilia, C. Protein adaptations in archaeal extremophiles. *Archaea* vol. 2013 (2013).

213. Lins, L., Thomas, A. & Brasseur, R. Analysis of accessible surface of residues in proteins. *Protein Sci.* **12**, 1406–1417 (2003).

214. Pastor, M. T., López de la Paz, M., Lacroix, E., Serrano, L. & Pérez-Payá, E. Combinatorial approaches: A new tool to search for highly structured β-hairpin peptides. *Proc. Natl. Acad. Sci.* **99**, 614 LP – 619 (2002).

215. Tamm, L. K., Hong, H. & Liang, B. Folding and assembly of β-barrel membrane proteins. *Biochim. Biophys. Acta - Biomembr.* **1666**, 250–263 (2004).

216. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

217. Dowle, M. & Srinivasan, A. data.table: Extension of `data.frame`. (2020).

218. Charif, D. & Lobry, J. R. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis BT - Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. in (eds. Bastolla, U., Porto, M., Roman, H. E. & Vendruscolo, M.) 207–232 (Springer Berlin Heidelberg, 2007). doi:10.1007/978-3-540-35306-5_10.

219. Wilkinson, L. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics* **67**, 678–679 (2011).
220. Wilke, C. O. cowplot: Streamlined Plot Theme and Plot Annotations for ‘ggplot2’. (2019).
221. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).

Vita

Cory DuPai received his Bachelor of Arts in Biology from Boston University in January 2013. He then spent 2014-2016 working for Dr. Nada Kalaany at Boston Children's Hospital, studying cancer metabolism. In the summer of 2016, he joined the labs of Dr. Wilke and Dr. Davies through the Cell and Molecular Biology graduate program at the University of Texas at Austin.

Permanent address: Corydupai@gmail.com

This manuscript was typed by the author