Investigating pathogen-host interactions and adaptation with network biology approaches

Marton Laszlo Olbei

A thesis submitted for the degree of
Doctor of Philosophy

University of East Anglia

Earlham Institute

Quadram Institute BioScience

United Kingdom

March 2021

Abstract

Serovars of the genus *Salmonella* are widespread enteric pathogens, causing acute inflammatory gut infections. However, a subgroup of *Salmonella* adapted to a systemic lifestyle instead of a mucosal one. A systems-level understanding of how molecular level changes accompanying this adaptive process potentially modify the behaviour of these invasive strains is crucial for future intervention processes, and possible treatments.

In this thesis, I generated and analysed multi-layered interaction networks for 20 strains in the genus *Salmonella*. I collated protein-protein, transcriptional regulatory, and metabolic interaction data from low and high-throughput experiments and performed predictive measures to add further connections to the systems. The resulting networks culminated in the update to SalmoNet, the first integrated network database for *Salmonella* serovars. Through comparative network approaches, users can highlight elements under selection in these invasive serovars, increasing our understanding of the host adaptation process leading to their systemic lifestyle.

During the last year of my PhD, I redeployed for 6 months to work on COVID-19 related research. This effort led to a systematic literature curation highlighting different cytokine responses in patients caused by SARS-CoV-2 compared to other similar viruses. I also led the effort to establish a new network resource, CytokineLink, aimed at highlighting avenues of cell-to-cell communication mediated by cytokines, to better understand inflammatory and infectious diseases.

Overall, the work presented in this thesis has increased our understanding of the *Salmonella* host adaptation process, by highlighting specific elements under selection, while also exhibiting how network information can be created, and used for understanding such evolutionary processes.

Acknowledgements

I would like to thank everyone who contributed to my PhD research, either professionally or personally. While the work presented in this thesis was of course done by me, I could not have created it in a vacuum. I would like to thank Tamas, Rob, Isa, Paddy, David, Joska, Eszter, Dezso, Leila, Balazs, Aggie, Matthew, Martina, Amanda, Szandra, Falk, Ross, Denes for all their help during the years, and all members of the Korcsmaros and Kingsley groups for their assistance and input, as I tried to come to terms with *Salmonella* and networks.

I am thankful for my loving wife, for supporting and understanding me all through these years. I could not have done this without you.

I am thankful for my family, my parents, my sister and my friends, who supported me in moving abroad, even if it meant we would see each other less often. And of course, I am thankful for Mimi, who made sure I get out the door every day, even in the pandemic.

I am thankful for the NRP BBSRC DTP programme, the Earlham and Quadram Institutes, for taking a chance on me, and funding my PhD studies.

I am eternally grateful for Tamas, who is the best supervisor anyone could have. A real mentor in and outside of science.

Table of abbreviations

AP-MS	affinity purification mass spectrometry
ARDS	acute respiratory distress syndrome
c-di-GMP	cyclic-di-GMP; cyclic diguanylate
ChIP-seq	chromatin immunoprecipitation sequencing
COVID-19	coronavirus disease 2019
CRS	cytokine release syndrome
DBD	DNA binding domain
dRNA-Seq	differential RNA-Sequencing
DGC	diguanylate-cyclase
DNA	deoxyribonucleic acid
E. coli	Escherichia coli
EDTA	Ethylenediaminetetraacetic acid
EI	extraintestinal
FBA	flux balance analysis
GEM	genome-scale metabolic model
GI	gastrointestinal
GRN	gene regulatory network
HDC	hypothetically disrupted sequence
HGT	horizontal gene transfer
ID	identifier
IFN	interferon

IFN-I	type I interferon
IFN-II	type II interferon
IgM	immunoglobulin M
IL	interleukin
iNTS	invasive non-typhoidal Salmonella
KS-test	Kolmogorov-Smirnov test
LPS	lipopolysaccharide
MC3	metropolis coupling markov chain monte carlo
MERS-CoV	middle east respiratory syndrome coronavirus
miRNA	micro RNA
mRNA	messenger RNA
NTS	non-typhoidal Salmonella
PCR	polymerase chain reaction
PPI	protein-protein interaction
PSSM	position specific scoring matrix
RNA	ribonucleic acid
SARS-CoV	severe acute respiratory syndrome coronavirus
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SD	standard deviation
SPI	*Salmonella* Pathogenicity Island
ssRNA	single-stranded RNA
T3SS	type three secretion system
TF	transcription factor
TFBS	transcription factor binding site
TG	target gene

TSS	transcriptional start site
URT	upper respiratory tract
UTR	untranslated region

List of figures

List of tables

List of peer reviewed publications

These peer-reviewed articles were published during my PhD studies (2017-2021).

Chapter 2:

- Mehta, T. K., Koch, C., Nash, W., Knaack, S. A., Sudhakar, P., **Olbei, M.**, Bastkowski S., Penso-Dolfin, L., Korcsmaros, T., Haerty, W., Di-Palma, F. (2021). Evolution of regulatory networks associated with traits under selection in cichlids. *Genome Biology*, 22(1), 25. doi: https://doi.org/10.1186/s13059-020-02208-8

Chapter 3:

- Métris, A., Sudhakar, P., Fazekas, D., Demeter, A., Ari, E., **Olbei, M**., Branchu, P., Kingsley, RA., Baranyi, J., Korcsmáros, T. (2017). SalmoNet, an integrated network of ten Salmonella enterica strains reveals common and distinct pathways to host adaptation. *NPJ Systems Biology and Applications*, 3, 31. https://doi.org/doi:10.1038/s41540-017-0034-z
- **Olbei, M.**, Kingsley, RA., Korcsmaros, T., Sudhakar, P. (2019) Network Biology Approaches to Identify Molecular and Systems-Level Differences Between Salmonella Pathovars. *Methods Mol Biol*. 1918:265-273. doi: 10.1007/978-1-4939-9000-9_21.
- Csabai, L, **Ölbei, M,** Budd A, Korcsmáros, T, Fazekas, D. (2018). SignaLink: Multilayered Regulatory Networks. *Methods Mol Biol*. 1819:53-73. doi: https://doi.org/10.1007/978-1-4939-8618-7_3.

Chapter 5:

- **Olbei, M.,** Hautefort, I., Modos, D., Treveil, A., Poletti, M., Lejla, G., Shannon-Lowe, C.D., Korcsmaros, T. (2021). SARS-CoV-2 causes a different cytokine response compared to other cytokine storm-causing respiratory viruses in severely ill patients. *Frontiers in Immunology*. 12:381 doi: https://doi.org/10.3389/fimmu.2021.629193

Publications not presented in this thesis:

- Turei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivaova, O., **Olbei, M**., Gabor, A., Theis, F., Modos. D., Korcsmaros, T., Saez-Rodriguez, J. (2021). Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*. 17:e9923 doi: https://doi.org/10.15252/msb.20209923
- Demeter, A., Romero-Mulero, M. C., Csabai, L., **Olbei, M.**, Sudhakar, P., Haerty, W., Korcsmaros T., (2020) ULK1 and ULK2 are less redundant than previously thought: computational analysis uncovers distinct regulation and functions of these autophagy induction proteins. *Sci Rep*. 10(1):10940. doi: https://doi.org/10.1038/s41598-020-67780-2.
- Treveil, A., Bohar, B., Sudhakar, P., Gul, L., Csabai, L., **Olbei, M.**, Poletti, M., Madgwick, M., Andrighetti, T., Hautefort, I., Modos, D., & Korcsmaros, T. (2021). ViralLink: An integrated workflow to investigate the effect of SARS-CoV-2 on intracellular signalling and regulatory pathways. *PLoS Computational Biology*, 17(2), e1008685. https://doi.org/10.1371/journal.pcbi.1008685
- Treveil, A., Sudhakar, P., Matthews, ZJ., Wrzesiński, T., Jones, EJ., Brooks, J., **Ölbei, M**., Hautefort, I., Hall, L.J., Carding, S.R., Mayer, U., Powell, P.P.,

Wileman, T., Di Palma, F., Haerty, W., Korcsmáros, T. (2020) Regulatory network analysis of Paneth cell and goblet cell enriched gut organoids using transcriptomics approaches. *Mol Omics*. 16(1):39-58. doi: https://doi.org/10.1039/c9mo00130a

Table of contents

Structure of the thesis

This thesis is organised into the following chapters:

- **Chapter I: General Introduction.** An introductory chapter describing the necessary literature background of the topics covered in the thesis, and summarises the aims within.
- **Chapter II: Construction of a multi-layered network database for Salmonella research.** This chapter presents the steps necessary to construct a multi-layered network database for a non-model organism, and describes the steps taken to improve and expand on the original version of it.
- **Chapter III: Network biology methods to study adaptation and evolution.** Chapter III describes the approaches and network analysis

tools appropriate for the network level analysis of adaptation and evolution, through a study involving the adaptive radiation of cichlid fish species, and a comparison of typhoidal and gastrointestinal *Salmonella* strains.

- **Chapter IV: The role of cytokines in SARS-CoV-2 infection.** In Chapter V, I detail the results of my 6-month redeployment into COVID-19 research. It describes a study on the differences in cytokine responses from patients infected by various cytokine release syndrome causing viruses, and a network resource aimed at understanding how cell types communicate with each other using cytokines.
- **Chapter V: Final discussion, perspectives and future work.** This final chapter describes the impact, conclusions, and perspectives of the work I presented in the thesis.

1. General introduction

The *Salmonella* genus consists of Gram-negative facultative anaerobic pathogens belonging to the *Enterobacteriaceae* family, a member of the *Proteobacteria* phylum. The majority of the serovars have broad host range and cause a self-limiting intestinal inflammation (gastrointestinal serovars). The gastrointestinal serovars use this process to modify the intestinal environment to their advantage and facilitate their transmission. A small subset of the genus, however, evolved alternative strategies of transmission, by adapting to an invasive lifestyle instead of a mucosal one, restricting their host range in the process, and colonising alternative sites in the host (extraintestinal serovars). In this thesis, I generated and analysed multi-layered interaction networks of multiple *Salmonella* strains, including both broad and narrow host range serovars, to understand this process, known as host adaptation.

The host adapted serovars cannot be placed on a single monophyletic lineage when attempting to map the phylogenetic relationships of *Salmonella* serovars, because their emergence is a result of convergent evolution (Vázquez-Torres, 2018). The host adaptation process usually involves the degradation of key genes that are not used in the novel environments of the pathogen, or are detrimental to them, and involves a change in infection phenotype, from an acute inflammatory one, to a “stealth” phenotype that leads to bacteraemia and fever (de Jong, Parry, van der Poll, & Wiersinga, 2012; Klemm et al., 2016; Uzzau et al., 2000).

While host adaptation is a process often coupled with genome degradation and expansion, the over 700,000 SNP divergence that characterises the phenotypically varied subspecies I. of *Salmonella enterica* is not purely a comparative genomics problem (Desai et al., 2013). It could be further elucidated by considering the way absent or newly acquired polymorphisms modify the system that can eventually lead to changes that make *Salmonella* alter its behaviour in the host organism, through formation or loss of regulatory sequences affecting gene expression or metabolic pathways, or non-synonymous SNPs altering the function of proteins. As such, the integration of multiple levels of knowledge could provide insight into distinct and shared interaction patterns that characterize *Salmonella* virulence and pathogenicity (Métris et al., 2017). However, the availability of different levels of knowledge one would integrate to carry out systems level analyses is scarce, especially for non-model organisms like *Salmonella,* and the information present is scattered in various databases.

Motivated by the information above, I set out with the following hypothesis:

> *The difference in the host adaptation capabilities of gastrointestinal and extraintestinal Salmonella enterica serovars can be characterized by specific changes in the topology of their metabolic, regulatory, or protein-protein interaction networks.*

This thesis contributes to progress toward testing this hypothesis, through the following aims:

- Generation of multi-layered interaction networks for extraintestinal and gastrointestinal *Salmonella enterica* serovars.

- Applying appropriate workflows and approaches to analyse evolutionary processes such as host adaptation, by using interaction networks as a medium.

The aims make progress toward addressing thes hypothesis, by generating a high-quality network resource and knowledgebase of *Salmonella* interaction information as the subject of analysis, and by involving network comparison methods successfully applied in other similar studies published in the relevant literature.

The following introduction chapter includes the literature and theory necessary to understand *Salmonella* as a pathogen, the host adaptation process, and the fundamentals of systems biology research.

1.1. History of *Salmonella* research and its importance in public health

1.1.1. History of *Salmonella* research

Members of the *Salmonella* genus are motile enteric pathogens, capable of causing a variety of diseases, from gastroenteritis to systemic infections. A member of the genus was observed for the first time by Karl Joseph Eberth, in the spleens of typhoid patients, who suspected it might be the cause of typhoid fever (Eberth, 1880). The bacteria was isolated and grown into a culture just a few years later by Gaffky (Gaffky, 1884). Around the same time *Salmonella enterica* serovar Cholerasuis was first described by Theobald Smith during his work at the Bureau of Animal Industry in Washington, DC., who worked in the

group of Daniel Salmon. The name of the pathogen was later given after him (Meštrović, 2018; Schultz, 2008).

1.1.2. *Salmonella* and public health

The *Salmonella* genus causes somewhere between 90 million to 1.3 billion cases of foodborne gastroenteritis, and up to 3 million deaths each year. The gastroenteritis caused by these pathogens is one of the most common foodborne illnesses, the incidence of intestinal disease caused by non-typhoidal Salmonella species is the highest in the developing world, and is also considerable in developed countries (Coburn, Grassl, & Finlay, 2007). A subgroup of strains causing enteric fever affects 11.9 – 27.1 million patients globally, with over 100.000 of these infections leading to death (Coburn et al., 2007; GBD 2017 Typhoid and Paratyphoid Collaborators, 2019; Hohmann, 2001; Majowicz et al., 2010).

The burden of disease caused by *Salmonella* is not new – these pathogens were one of the most prevalent food poisoning organisms of the 20th century and were most likely a constant foodborne threat in the past as well. A study went as far to propose a *Salmonella enterica* subspecies enterica serovar Paratyphi C outbreak to be one of the strong candidates for the epidemic causing the population decline in the 16th century Aztec empire (Vågene et al., 2018). While microbiology progressed a lot to understand their structure, relationships and natural history, much of their qualities remain unclear (Hardy, 2004). As is the case with other commonly occurring pathogens, there is an increased prevalence of multidrug resistant Salmonella strains in recent years, further increasing the health risk and

public health cost associated with infections, including the health of livestock, as food animals often serve as reservoirs of the pathogen (Branchu, Bawn, & Kingsley, 2018; Hofer, 2019).

1.1.3. *Salmonella* nomenclature

Today, when we talk about *Salmonella*, we usually refer to *Salmonella enterica* and its subspecies, most often subspecies I., as the pathogens in this subspecies are the ones responsible for most infections in warm blooded animals (A. Bäumler & Fang, 2013). *Salmonella* nomenclature has not always been this clear cut. *Salmonella* was first recognised as a distinct group of organisms by 1900, and as research interest grew around them in North America and Europe, different laboratories and methodologies led to the same organisms receiving multiple names, and the same name given to multiple organisms.

As an example, *Salmonella enterica* subspecies *enterica* serovar Typhimurium, one of the most well studied serovars today, was once known under multiple aliases: Mutton type, Hatton strain, Breslau type, Freiburg type, *Salmonella aertrycke* and *Salmonella suipestifer*, to name a few (Hardy, 2004). *Salmonella* can be and was for a long time classified by its serotype. Serotypes are determined by the Kauffman-White classification scheme, that can distinguish subsets of microbes based on surface antigens they carry. In the case of *Salmonella*, this is based on the O and H antigens, the former a part of the lipopolysaccharide (LPS) coating, while the latter is a part of the flagellum. Based on the combination of the small scale differences in these markers the isolated bacteria can be assigned a serotype (Ibrahim & Morin, 2018).

Today, there are two species in the genus: *Salmonella bongori*, and *Salmonella enterica*. In 2005, the International Committee for Systematics of Prokaryotes designated the type species of the *Salmonella* genus to be *Salmonella enterica* (previously known as Salmonella choleraesuis) and its type strain to be LT2 *(Salmonella enterica* subsp. *enterica* serovar Typhimurium strain LT2) (Judicial Commission of the International Committee on Systematics of Prokaryotes, 2005). As mentioned above, the majority of diversity and public health burden comes from *Salmonella enterica* subspecies I., also known as subspecies enterica. There are six subspecies in total:

- I - *enterica*
- II - *salamae*
- IIIa - *arizona*
- IIIb - *diarizonae*
- IV - *houtenae*
- VI - *indica*

The gap in numbering between *houtenae* and *indica* is caused by the reclassification of *Salmonella bongori* into a separate species, formerly known as subspecies V (Brenner, Villar, Angulo, Tauxe, & Swaminathan, 2000; Desai et al., 2013). To shorten reports the names of serovars are often curtailed. For example, one can find *Salmonella enterica* subspecies *enterica* serovar Typhimurium shortened as *Salmonella* Typhimurium or *S.* Typhimurium (Brenner et al., 2000).

1.1.4. Evolution of *Salmonella enterica*

The common ancestor of the *Salmonella* genus existed about 25-40 million years ago. The two *Salmonella* species are closely related to *Escherichia coli*, a commensal and opportunistic pathogen bacterium commonly found in the lower intestine. The divergence from the *Escherichia coli* lineage happened in three to five major steps, depending on the model we apply (A. J. Bäumler, Tsolis, Ficht, & Adams, 1998; Winfield & Groisman, 2004), occurring approximately 100 to 160 million years ago (Doolittle, Feng, Tsang, Cho, & Little, 1996; Ochman & Wilson, 1987).

In the first phase of divergence from the common ancestor, one branch of speciation led to *Escherichia coli*, a commensal bacterium living in the gut of mammals. The other, pathogenic subset acquired a set of genes needed to infect the intestine, including the Salmonella Pathogenicity Island 1 (SPI-1), which eventually gave rise to *Salmonella bongori*. The SPI-1 island is a 40-kb long region encoding effector proteins, a type three secretion system (T3SS-1), and elements required to regulate these. The T3SS is an intricate protein structure that assembles into a syringe-like complex, a needle structure that can penetrate the host epithelial cells and translocate SPI-1 effectors into it. There are more SPI like genomic islands present in the genus, most of them acquired later on in the speciation process (Lou, Zhang, Piao, & Wang, 2019); (Winfield & Groisman, 2004).

In later evolutionary steps the ancestral pathogen accumulated genes that are required for the colonization of deeper tissues leading to *Salmonella enterica*

subspecies II, IIIa, IIIb, IV, VI, VII. The new set of genetic material includes another large island called SPI-2, encoding a similar T3SS required for intra-macrophage survival. In the last phase the *Salmonella enterica* subspecies I formed, expanding the host range to warm blooded vertebrates, which includes us, humans (A. J. Bäumler, Tsolis, et al., 1998), and as such a majority of human infections come from this subspecies. The acquisition of these genetic islands also means that *Salmonella* serovars cause disease by very similar mechanisms, utilising the same virulence genes (Tanner & Kingsley, 2018). Figure 1 depicts the three major steps as outlined by Bäumler et al in 1998.

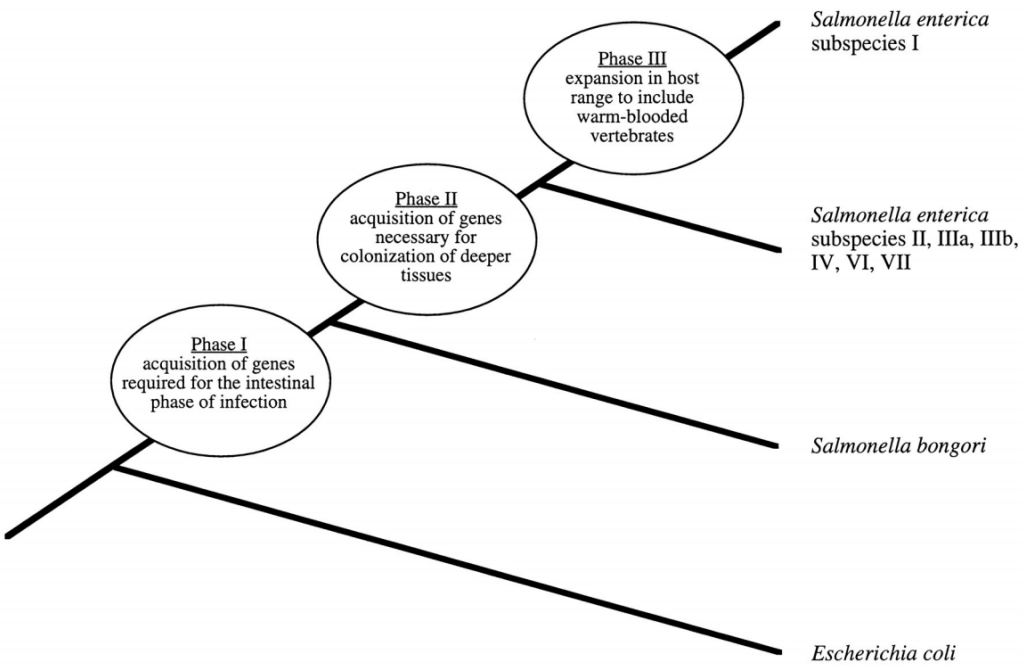

Figure 1. Phases of Salmonella evolution depicting the necessary steps which granted the pathogen the ability to infect humans, and the formation of the new species, subspecies. Source: (A. J. Bäumler, Tsolis, et al., 1998).

Infections of most of the *Salmonella* serovars cause a self-limiting gastroenteritis. The invasion induces an inflammatory event, that shapes the

intestinal niche to one that favours the pathogen by releasing metabolites it is suited to utilize better, and as such it can use to outcompete the local microbiome (Rivera-Chávez & Bäumler, 2015; Stecher et al., 2007; Tanner & Kingsley, 2018). *Salmonella enterica* subspecies I harbours a large number of serovars, many of them adapted to various host species. Figure 2 depicts the phylogenetic relationships of the major *Salmonella* serovars in subspecies I.

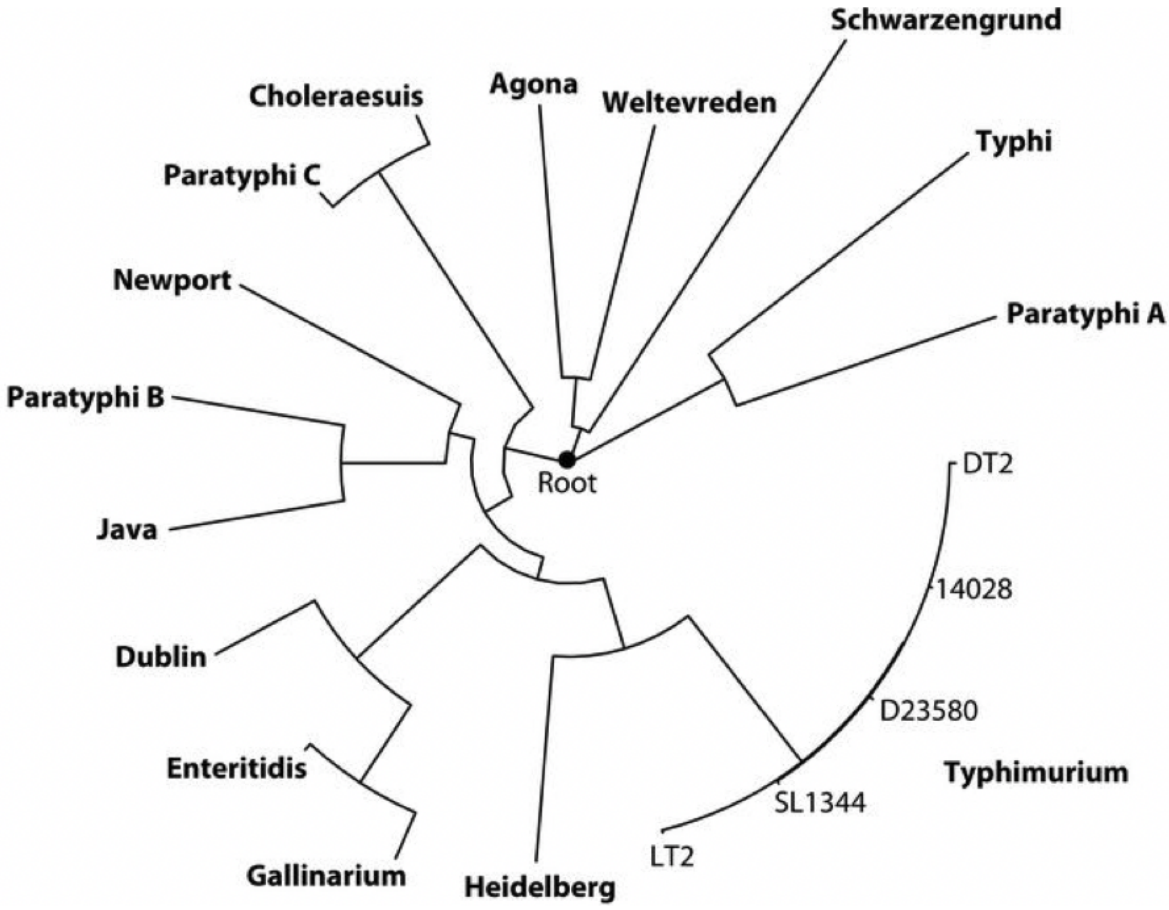

Figure 2. Phylogenetic relationships of the major Salmonella enterica subspecies I serovars. Image from (Branchu et al., 2018)*, licensed under CC-BY 4.0*

1.2. Host adaptation in Salmonella serovars

1.2.1. Defining host adaptation

The relationships between pathogens and the hosts they infect can vary based on the level of co-evolution. Some pathogens have a broad host range, capable of infecting many species, while others are more specialised, only focusing on one or a few specific host species. Most pathogenic bacteria fall into the first category, and this is the case with *Salmonella* as well. The typical Salmonella infection leads to a self-limiting gastroenteritis shaping the intestinal environment to one that favours the pathogen (Stecher et al., 2007).

Host adaptation is commonly assumed to be the ability of a serotype to cause disease only in the subset of animal species it is adapted to. The reality, however, is a bit more complicated when taking all available data into consideration (R A Kingsley & Bäumler, 2000). Defined in an article by (R A Kingsley & Bäumler, 2000) host adaptation is the 'ability of a pathogen to circulate and cause disease in a host population'. This ability is unrelated to its virulence for other host species. For example, *Salmonella enterica* subsp. *enterica* serotype Choleraesuis is not considered swine adapted because it causes a more serious disease in them as in humans, but because it is able to persist in pig populations by direct transmission (R A Kingsley & Bäumler, 2000).

Although the incidence of human infection by serovars that are host adapted to animals is rare, the infection can be quite invasive and cause a serious illness. The

behaviour of host adapted strains can also be context dependent - for example *Salmonella enterica* subspecies *enterica* serovar Typhimurium, which causes gastroenteritis in a human host, but bacteraemia in rodents. The main difference lies here - infection with a non-host adapted serovar usually leads to a self-limiting illness, while a person getting infected with *Salmonella enteric*a subspecies *enterica* serovar Typhi might end up transmitting the disease to others (Tanner & Kingsley, 2018). The self-limiting nature of a non-host adapted infection can also be interpreted as beneficial from a public health point of view, as multi-drug resistant serotypes become more prominent (Eng et al., 2015).

1.2.2. Changes accompanying host adaptation

Narrow host range serovars of *Salmonella* typically cause a systemic disease, beyond the intestine, and exhibit increased virulence. The exact mechanisms and reasons for specialization are still studied, but it has been implicated that the potential to expand into new niches might be a strong driving force, as the pathogen does not have to compete with as much local microbiota outside of the intestine (A. J. Bäumler, Tsolis, et al., 1998; Tanner & Kingsley, 2018). Other studies suggest that members of an ecological system equipped with specialists can increase resource exploitation within the system, which could also be potentially driving the process (A. Bäumler & Fang, 2013).

Host adapted *Salmonella* variants have emerged on multiple occasions, convergently (Hiyoshi et al., 2018; Vázquez-Torres, 2018). The pressure to exploit available resources can be one of the potential drivers behind it, one of the defining differences between pathogenic and commensal bacteria, is the

ability of utilising niches commensals cannot. Gene inactivation caused by genome degradation is one of the recurring features of host adaptation, even though the events initiating it are not completely understood yet (Klemm et al., 2016). One of the mechanisms thought behind this phenomenon, is that these pathogens' genes often degrade over time, when they affect pathways that are non-essential in their new niches within the host, and neutral mutations slowly accumulate in them. Biofilm formation is typically one of the functions that is less effective in host adapted serovars, but there are other major biological functions impacted as well, such as chemotaxis or anaerobic metabolism (Holt et al., 2009; MacKenzie, Palmer, Köster, & White, 2017; Nuccio & Bäumler, 2014). Another reason for the emergence of loss of function mutations in certain genes is antagonistic pleiotropy, that some of the pathways that were useful in one environment might be counterproductive in the new niche. Genome size can change quite dynamically in bacteria as they can utilize resources better by not transcribing genes they are not using, and can outcompete individuals by dividing faster than ones with larger genomes (Ilyas, Tsai, & Coombes, 2017; Nilsson et al., 2005).

Host adaptation is not only driven by gene inactivation, many of the host adapted serovars have also accumulated genes for which there are no orthologous proteins in broad range serovars, in the form of additional *Salmonella* pathogenicity islands, such as SPI-7, SPI-8 and SPI-10 (Winfield & Groisman, 2004)). The gain and loss of genes modulates the possible range of host microbe interactions. For example, the expression of the *S.* Typhimurium effector *gtgE* in *S.* Typhi allows it to survive and multiply in a mouse host by promoting survival

inside mouse macrophages, which it would be unable to do, it being a human restricted serovar normally (Spanò & Galán, 2012).

1.2.3. Degrees of host adaptation, terminology

Going through the processes outlined above, within the *Salmonella enterica* subspecies I some of the serovars became host adapted, and thus the group can be divided into different categories, in a multitude of ways. Most often we find two pairs of terms when talking about these pathogens: typhoidal - nontyphoidal categories, and extraintestinal - gastrointestinal categories. Although they mean similar things, the context in which they are used matters, as the first two refers to the human disease they cause (typhoid fever), while the latter refers to their relationship to the intestine as a niche. Table 1 highlights the differences between the terms.

Pathovar	**Gastrointestinal**	**Extraintestinal**	
Human disease	***Non-typhoidal***	***Non-typhoidal***	***Typhoidal***
Typical serovars	S. Typhimurium, S. Enteritidis, S. Heidelberg, S. Newport	S. Choleraesuis, S. Dublin	S. Typhi, S. Paratyphi (A,B,C)
Disease	Self-limiting gastroenteritis (intact immune system), bacteraemia (immunocompromised host)	Systemic infection	(Para)typhoid fever
Host range	Broad	Host adapted (porcine, bovine)	Host restricted (human)

Table 1: *Salmonella* pathovars, and their relationships to human disease and host range.

The extraintestinal pathogens are a small group of specialists adapted to new environments in their host. The most extensively studied member is *Salmonella enterica* serotype Typhi (Rivera-Chávez et al., 2016). The level of host adaptation in *Salmonella enterica* serotypes varies, with *S.* Typhi being generally considered one of the most specialised member of the group, while (from a human disease point of view) *S.* Typhimurium being a typically broad host range, gastrointestinal serovar. Host adaptation can sometimes progress into host restriction, where the pathogen limits itself to one single host species, and causes a more severe illness (Klemm et al., 2016). Figure 3 details the host range of host adapted serovars in *Salmonella*.

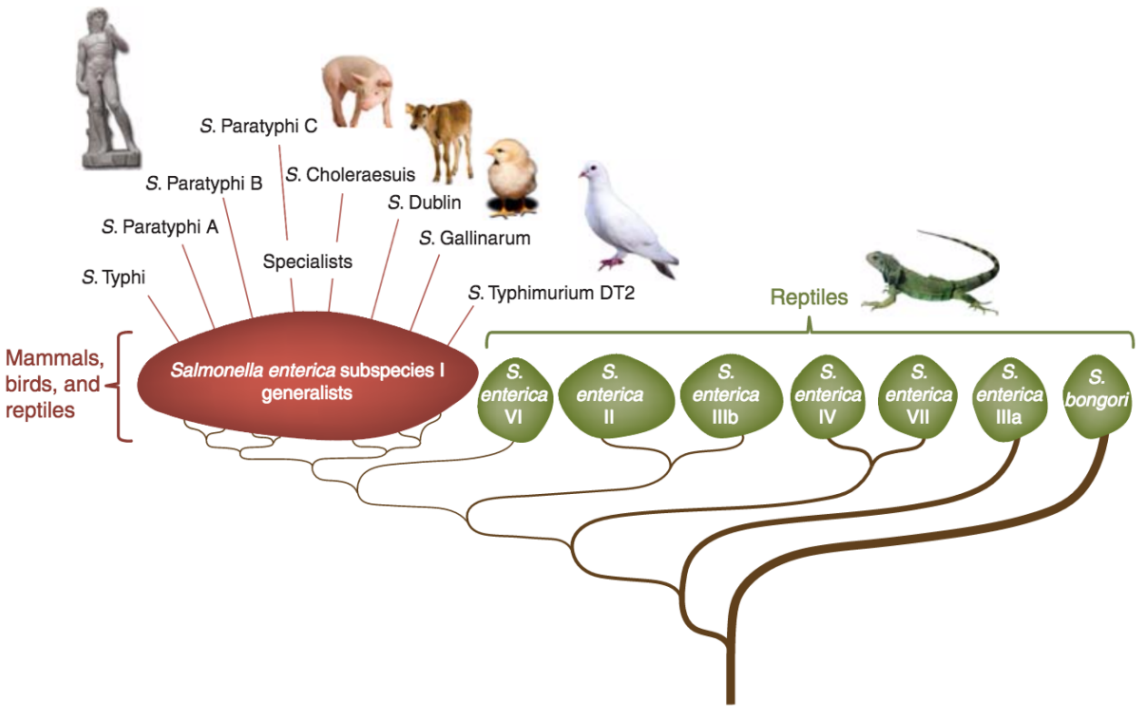

Figure 3. Host range of pathogens in the Salmonella enterica subspecies. Subspecies I mostly consist of generalists from which specialists emerge from time to time. Source: (A. Bäumler & Fang, 2013)*, with permission of the copyrights holder Cold Spring Harbor Laboratory Press.*

S. Typhi is the causative agent of typhoid fever, a dangerous disease manifesting as a high fever, with abdominal pain, and headaches. *S.* Paratyphi A can cause a very similar condition which is why these two are often referred to as typhoidal Salmonellae (J Parkhill et al., 2001). As outlined above, the host-adaptation of *S.* Typhi happened via genetic degradation and the recruitment of many genes associated with virulence (den Bakker et al., 2011; Klemm et al., 2016).

A prominent example of this is the Vi exopolysaccharide capsule of *S.* Typhi. Generally speaking, broad and narrow host range serovars approach the infection process from opposite ends of the spectrum - the former group evolved to elicit inflammation to reduce competition and free up metabolites it can use in the intestine (e.g. tetrathionate), while the latter evolved to avoid the immune system for as long as possible in order to disseminate to organs of the reticuloendothelial system (Vazquez-Torres et al., 1999). The Vi capsule, encoded by the *viaB* locus on SPI-7 does that, by preventing the activation of the complement system (Pickard et al., 2003; Wangdi et al., 2014). The pathogens often downregulate the gene *tviA* as well for similar reasons, which is responsible for the regulation of flagella expression (Winter, Raffatellu, Wilson, Rüssmann, & Bäumler, 2008). *S.* Typhi and Paratyphi are host restricted, they do not cause disease in nonhuman hosts (they can infect higher primates, but do not cause typhoid fever in them) (J Parkhill et al., 2001). They express the typhoid toxin, a cytolethal distending toxin, causing G2/M cell cycle arrest which leads to the apoptosis of the affected cells (Galán, 2016).

Even though *S.* Typhi and Paratyphi share many similarities, they can be quite different from other extraintestinal serovars which evolved alternative ways to

disseminate in the host system. Extraintestinal serovars underwent convergent evolution, reaching a similar systemic lifestyle through different mechanisms. This convergence has been previously observed in the patterns of genome degradation of extraintestinal Salmonella enterica serovars (Galán, 2016; Nuccio & Bäumler, 2014; J Parkhill et al., 2001).

1.2.4. Host adaptation is an ongoing process

Host adaptation is a constantly ongoing process, and recently *S.* Typhimurium pathovariants emerged, that follow a host adapted lifestyle, and share functional changes with other host adapted Salmonella. The *Salmonella enterica* serovar Typhimurium definitive type 2 (DT2) is host restricted to rock pigeons (*Columbia livia*). These isolates form a distinct cluster within *S.* Typhimurium but share a common ancestor with them in the recent past, and in some ways represent a microcosm of *Salmonella* evolution. These isolates adapted to the higher (42 °C) internal temperature of the avian host, and went through many of the functional changes, such as downregulation of flagella and motility one would see in other host restricted pathovariants such as *S.* Typhi (Bawn et al., 2020; Robert A Kingsley et al., 2013; Tanner & Kingsley, 2018; Winter et al., 2010).

Over the past decades, another group of host adapted *Salmonella* appeared as one of the most commonly isolated pathogens from the blood of patients (Feasey, Dougan, Kingsley, Heyderman, & Gordon, 2012). The invasive nontyphoidal *Salmonella* (iNTS) strains cause a similar systemic infection as human adapted extraintestinal serovars such as Typhi and Paratyphi, most often in

immunocompromised individuals, e.g. young children, AIDS patients. The infection often leads to bacteremia and meningitis, and multidrug-resistant variants have caused epidemics in several African countries. Sub-Saharan Africa is one of the worst impacted regions. The infection is most often caused by a sequence type of *Salmonella* Typhimurium and of *Salmonella* Enteritidis (Gilchrist & MacLennan, 2019). ST313, containing these iNTS strains, consists of three lineages, the third of which was described very recently (Pulford et al., 2021).

The variants causing this bacteraemia show similar molecular changes discussed above with other host adapted variants and seem to be distinctly adapting to infection in immunocompromised hosts. The iNTS strains can still cause intestinal inflammation, but the genome degradation alters functions required for survival in the intestine, for the environment outside the host, for serum resistance, and human-to-human mode of transmission (Robert A Kingsley et al., 2009; Okoro et al., 2015). Altogether, these sequence types evolved to have reduced capability of intestinal pathogenesis, but increased systemic dissemination (Carden et al., 2017; Okoro et al., 2012; Singletary et al., 2016).

The 10,000 (10k) *Salmonella* Genomes Project was launched specifically to address and understand invasive non-typhoidal *Salmonella* infections, collecting samples from Africa and South America (10K Salmonella Genomes Project, 2017).

In a well-documented case, there was an example of a bloodborne *S.* Enteritidis infection showing signs of host adaptation in an immunocompromised patient. Over the course of 15 years, the non-typhoidal

Salmonella infection recurred, always resulting from a relapse rather than reinfection, and it culminated in the pathogen slowly losing functionality in genes that are not necessary for a systemic lifestyle (Klemm et al., 2016).

The events following host adaptation - genome size reduction, formation of pseudogenes, acquisition of mobile/IS elements - are not unique to *Salmonella* and have been described in multiple other bacterial clades, e.g. *Shigella*, another group closely related to *Escherichia coli* (Hawkey, Monk, Billman-Jacobe, Palsson, & Holt, 2020), or other groups such as *Yersinia, Rickettsia, Bordetella* (Cole et al., 2001; Moran & Plague, 2004; Julian Parkhill et al., 2003).

1.2.5. Convergent evolution

A common result of natural selection is that sometimes similar pressures result in similar solutions from relatively distant - or at least not monophylatically related - organisms. With eukaryotes, especially animals and plants this is something where many examples exist describing this process, both including currently alive and fossil specimens.

The process is most obvious on a phenotypic level. Commonly used examples are the similar anatomical solutions fish and other vertebrates came up with that returned to water, or the flying apparatuses of bats and pterosaurs, the similarity of the hummingbird hawk-moths (*Macroglossum stellatarum*) and hummingbirds, the anatomy of the eye in humans and certain cephalopods.

Convergent evolution occurs on a molecular level as well. A very timely example is the identical single nucleotide polymorphism mutations collected by geographically distant lineages of the currently ongoing SARS-CoV-2 pandemic. There are select amino acid changes in the spike proteins of these variants of concern, that have emerged independently of each other, and are responsible for increased transmissibility.

1.2.5.1. Adapting to a host affects similar functions

In the case of the host adaptation process, *Salmonella* serovars go through similar changes, both on a molecular and a phenotypic level. The selection driven genome degradation in *S.* Typhi and *S.* Paratyphi associated with loss of function events affects genes known to be important in gastroenteritis, and effectors that are normally translocated into host cells (McClelland et al., 2004). It also affects chemotaxis, virulence, motility, biofilm formation, and resistance to antibiotics.

The affected functions are the same, but the way the individual serovars solve them can be different: for example, *S.* Typhi and *S.* Paratyphi A can both avoid the respiratory burst from phagocytes, but through different manners. The former does this by preventing the antibody-mediated complement activation utilising its Vi polysaccharide capsule, while *S.* Paratyphi A uses very long O-antigen chains containing the O2 antigen to avoid the binding of the antigen. These typhoidal *Salmonella* strains cause a very similar enteric fever, and both are human adapted, and because of the similar pressures they came up with solutions converging on the same problem – preventing complement activation

– but through different means. Figure 4 shows the comparison of the aforementioned structures.

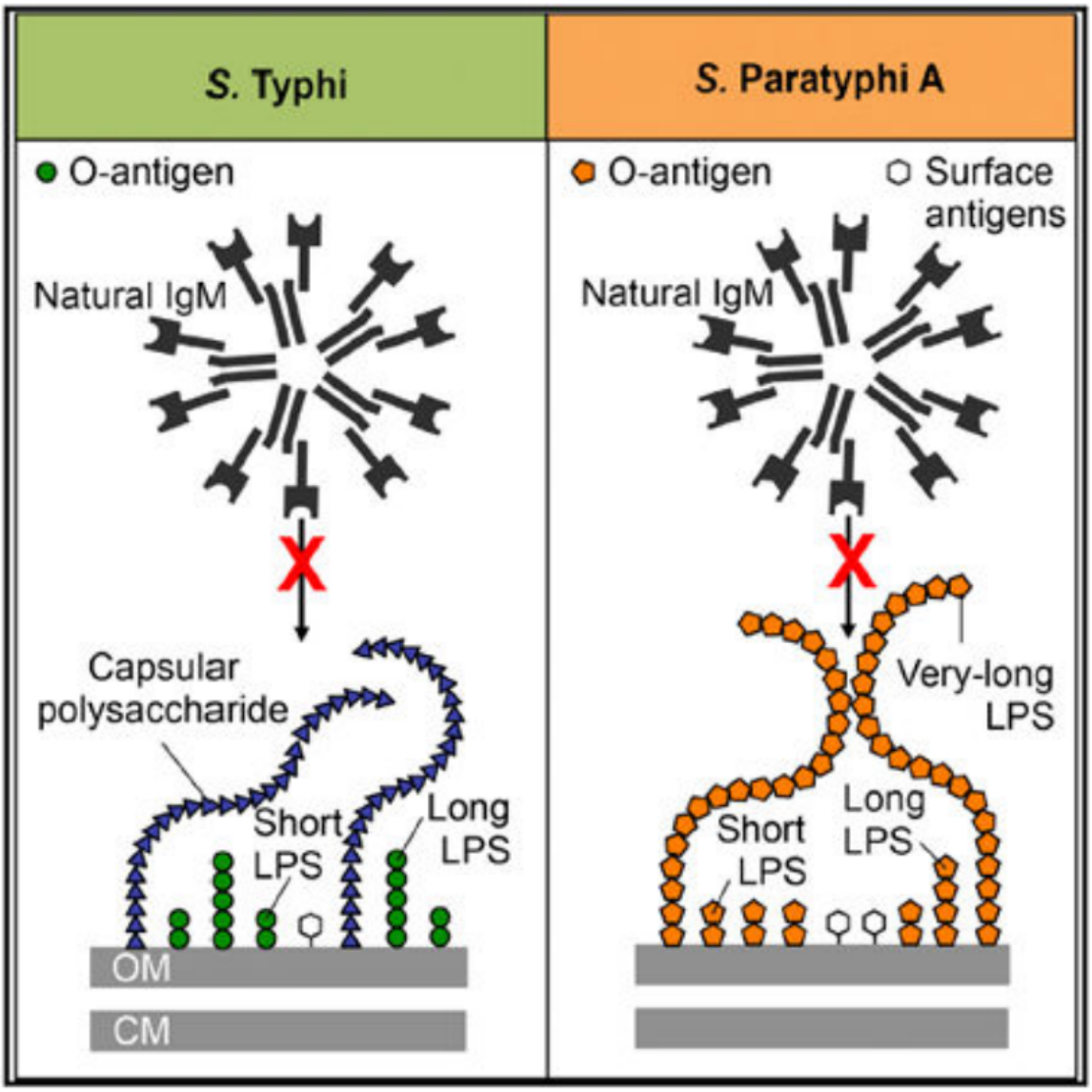

Figure 4. Typhoidal Salmonella serovars both avoid phagocyte respiratory burst, but arrived at the solution through different means. S. Typhi uses its polysaccharide capsule to prevent IgM binding, while S. Paratyphi A achieves the same result using very long LPS chains. OM: outer membrane, CM: plasma membrane, LPS: lipopolysaccharide. Image from (Hiyoshi et al., 2018) *licensed under CC BY-NC-ND 4.0*

In a similar way, the aforementioned phage type DT2 *S.* Typhimurium variants collated changes to their physiology to make them more fit to living in

an avian host, but did it through the rewiring of their transcriptional profile through accumulating point mutations in regulatory regions instead of the acquirement of novel genes, such as *tviA* in *S.* Typhi (Bawn et al., 2020; Robert A Kingsley et al., 2013; Winter et al., 2010).

1.2.5.2. Biofilm formation is hindered in host adapted *Salmonella* strains

Biofilms of all kinds are produced by a large number of bacteria, serving as a different mode of growth, usually on physical surfaces. They allow the bacteria to create multicellular communities, resist antibiotics, protect the cells from phagocytosis, and enhance their abilities to create slow growing persister populations (Tursi et al., 2020). *Salmonella* biofilms are commonly characterised by the so-called rdar phenotype, the red, dry, and rough appearance of their colonies grown on agar plates stained by Congo red dye.

The main structural components of *Salmonella* biofilms are curli fimbriae, a matrix of amyloid proteins, intermixed with cellulose. The two combined produce a resistant extracellular matrix for the pathogens. Biofilm formation is often one of the functions that degrades as *Salmonella* adapts to a systemic lifestyle, and has been noticed in the iNTS strains that formed relatively recently (MacKenzie et al., 2017).

1.2.5.3. Cyclic di-GMP

One of the most important signals regulating biofilm formation is the presence of cyclic di-GMP (c-di-GMP). C-di-GMP is a secondary messenger metabolite commonly found in bacteria, where it, amongst others, regulates motility, biofilm production and virulence. C-di-GMP is produced by diguanylate cyclases, and can be degraded by phosphodiesterases (Römling, Galperin, & Gomelsky, 2013). The activity of these two enzyme groups, the availability of precursor molecules, and extracellular signals control the levels of c-di-GMP. The decisions to reduce motility, and/or virulence and start producing biofilms are quite important and severe from the bacteria's point of view, and as such are under tight control (Jenal, Reinders, & Lori, 2017; Petersen, Mills, & Miller, 2019).

C-di-GMP is commonly summarised as a sessile-motile switch in many bacterial species, like *Vibrio*, where high intracellular c-di-GMP concentration leads to reduced motility and biofilm formation, while a low concentration of the metabolite promotes free swimming motile behaviour. The former instance typically occurs, when the pathogen is outside the host, and wants to persist until it can get taken up by another host organism, while the latter state is mostly descriptive of the within host state (Tamayo, Pratt, & Camilli, 2007). An increase in c-di-GMP in *Salmonella* leads to the same phenotype, regulated by CsgD, a transcriptional regulatory protein. CsgD, and the regulatory network it controls flows into the regulation of virulence as well further downstream, as the formation of biofilms downregulates virulent traits in return.

To sum up the process, environmental signals (e.g. L-arginine) promote the production of c-di-GMP and biofilm formation separately, and culminate on the CsgD master regulator. A low level of c-di-GMP leads to an inactive CsgD state, and virulence (motility, chemotaxis, active T3SS-1), while increasing c-di-GMP levels turns CsgD on, and suppresses virulence, while increasing curli and cellulose production (MacKenzie et al., 2017).

The regulation of biofilm production and virulence intersect in *Salmonella* in very interesting ways. While the process generally works as detailed above (turning off during infection, and on in between), in a study the authors have found, that during the intra-macrophage stage of infection of a mice model by *S.* Typhimurium, the pathogen actually induces c-di-GMP signalling and cellulose synthesis, deliberately suppressing its own virulence. They hypothesize that this way the bacteria can exploit host resources slowly, prolonging the infection and increasing the chance of transmission (Pontes, Lee, Choi, & Groisman, 2015).

1.2.6. *Salmonella* Pathogenesis

Salmonella strains apply a variety of strategies to infect their preferred host species, depending on their host range. Non-typhoidal and typhoidal *Salmonella* follows different strategies to ensure replication and transmission success, but one of the key aspects of all *Salmonella* infections is how the pathogen tries to hide itself from the immune system, and even hijack certain aspects of it (Gut, Vasiljevic, Yeager, & Donkor, 2018; Ohl & Miller, 2001).

Salmonella enters the host through contaminated food or water sources, which is what makes them one of the most common foodborne pathogens. The first goal of the bacterium is to reach the epithelial cells of the intestine, but there are of course physical, chemical and biological barriers in the way that it has to overcome first. For example, gastric acidity is one of the first lines of defence against enteric pathogens, considering the pH of an empty stomach (in the case of humans) can be as low as 2. The specific microenvironment, i.e. the surface of the foodstuff can be protective for *Salmonella*, by temporarily raising the pH of the stomach, and providing a source of amino acids for the pathogen to maintain its acid resistance genes (Garai, Gnanadhas, & Chakravortty, 2012; Waterman & Small, 1998).

Once *Salmonella* reaches the intestinal lining, it attaches to the epithelial cells, activates its endocytic pathway, and its uptake into the epithelial cells of the host. It specifically targets Peyer's patch, the microfold or M-cells found here, and the immune cells below. The pathogen accomplishes this by activating one of its type III secretion systems. These systems, as mentioned previously, are a needle-like protein structure that can pump effectors into the host cells, affecting its behaviour. The genes that encode these T3SSs sit on the aforementioned SPIs (Ehrbar & Hardt, 2005). The specific island required for this phase of the infection is SPI-1, meaning this is the genomic region that is shared across the entire genus, including *Salmonella bongori* (A. J. Bäumler, Norris, et al., 1998).

Following the uptake of *Salmonella* into the intestinal epithelial cell, it passes through towards the submucosa. The submucosa is rife with immune cells, which *Salmonella* uses to its advantage. Once it reaches this layer it will be shortly

engulfed by a macrophage in a phagocytosis-like process led by the genes of SPI-2 (and other Salmonella Pathogenicity Islands), encoding for a similar T3SS. The pathogen initiates the production of a special modified phagosome, the *Salmonella*-containing vacuole (SCV), which it uses as a protective niche within the macrophage (Dougan & Baker, 2014). Under normal circumstances there are multiple antibacterial mechanisms the host can employ, but the effectors of SPI-2 neutralises many of these, e.g. they block the fusion of acidifying lysosomes to the SCV (Giannella, 1996).

This is where the pathogenic process bifurcates for non-typhoidal and typhoidal *Salmonella,* from the viewpoint of an infected human. The former group are eventually eliminated by an inflammatory cascade (Mayuzumi, Inagaki-Ohara, Uyttenhove, Okamoto, & Matsuzaki, 2010). Typhoidal *Salmonella,* despite having the same SPI-1 and SPI-2 virulence factors, have gained additional tools that it can use to evade the innate immune system. *S.* Tpyhi can downregulate its flagella using the gene *tviA*, it's Vi capsule causes a lower inflammatory response, and this *Salmonella* can effectively utilise the macrophages as a safe niche within the host, where it can hide, replicate, and propagate to different sites in the host, like the liver or the spleen. *S.* Typhi also produces a toxin within host cells, called the typhoid toxin, which causes cytoplasmic detention and cell cycle arrest. What the specific role of the toxin is not currently known (Galán, 2016; Tanner & Kingsley, 2018). Figure 5 summarises the main differences in the pathogenesis process between non-typhoidal and typhoidal serovars.

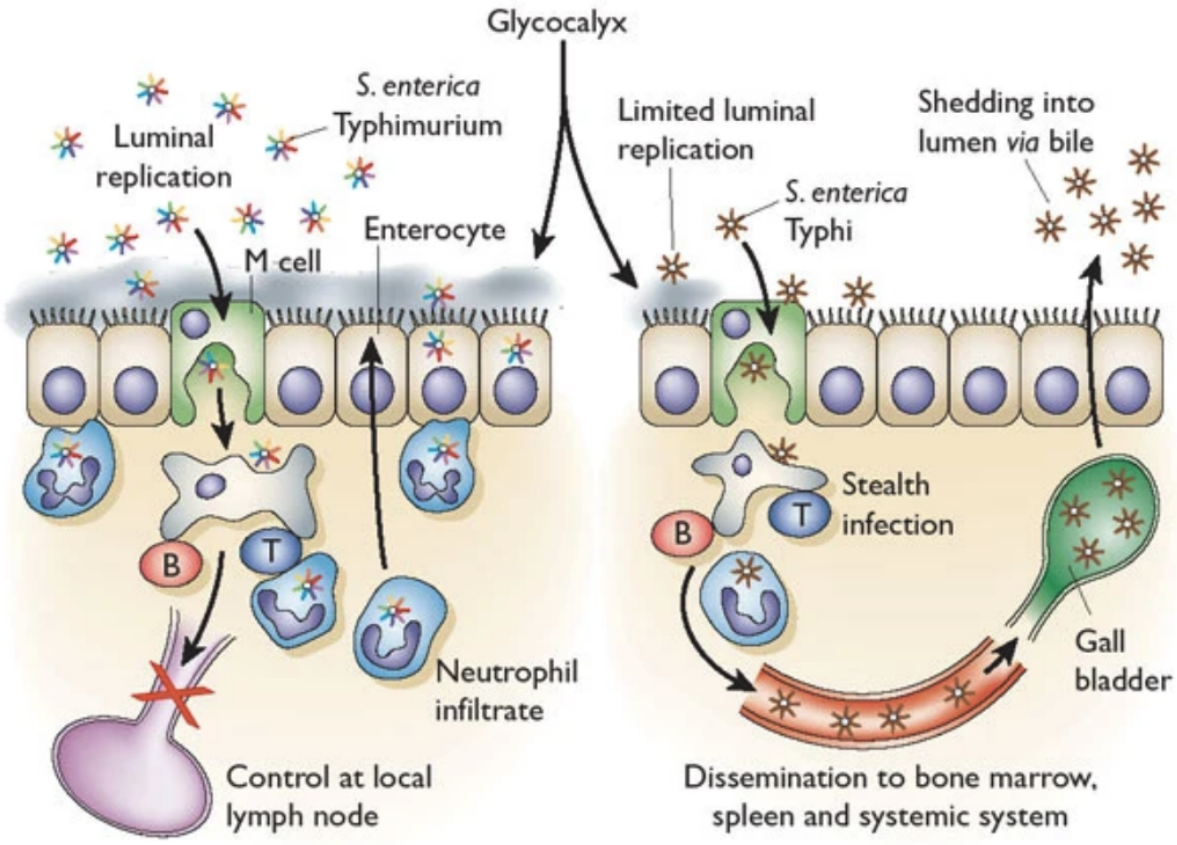

Figure 5. Differences in pathogenesis between nontyphoidal and typhoidal serovars S. Typhimurium and S. Typhi. The self-limiting non typhoidal infection is eventually stopped at the local lymph nodes, while the typhoidal pathogen can hijack macrophages and disseminate further into the host system. Image source: (Young et al. 2002) *with permission of the rights holder, Springer Nature.*

1.3. SARS-CoV-2 and COVID-19

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a positive-sense single-stranded RNA (ssRNA) virus, and the causative agent of the currently ongoing coronavirus disease 2019 (COVID-19) pandemic. This is the third major outbreak linked to the members of the *Coronaviridae* family. The family sits within the order of *Nidovirales*, and can be further divided into four genera: *alphacoronavirus*, *betacoronavirus*, *gammacoronavirus* and *deltacoronavirus*. While alpha-and betacoronaviruses only infect mammalian species, the other two genera have a broader host range, and can infect avian species as well. An epidemic of severe acute respiratory syndrome coronavirus (SARS-CoV) broke out in 2002, and multiple times for the middle east respiratory syndrome coronavirus (MERS-CoV). SARS-CoV-2 emerged in Wuhan, China, due to a spillover of an animal coronavirus to humans, similarly as it has happened in the cases of MERS-CoV and SARS-CoV (Andersen, Rambaut, Lipkin, Holmes, & Garry, 2020; Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020; Machhi et al., 2020).

SARS-CoV-2 consists of 29 proteins in total, 16 of which are non-structural proteins. It shares 79% of its genome with SARS-CoV and 50% with MERS-CoV. The observed outcomes following an infection differ between the three viruses, with SARS-CoV-2 being the most transmissible, but having a lower mortality rate (2.3% vs 9.6% of SARS-CoV and 35% of MERS-CoV) (Suryawanshi, Koganti, Agelidis, Patil, & Shukla, 2021).

Respiratory coronaviruses, including SARS-CoV-2, are transmitted primarily through respiratory droplets. Because of this, the virus enters the host most often through the respiratory tract, airway and alveolar epithelial cells (Harrison, Lin, & Wang, 2020).. Capable of infecting cells carrying the angiotensin-converting enzyme 2 (ACE2) and TMPRSS2 surface receptors, they enter the cells, and start replicating. It is important to mention, that there is a possibility of faecal-oral route of transmission as well. Human coronaviruses have been known to cause gastrointestinal infections, with varying degrees of severity (Harrison et al., 2020; Lamers et al., 2020).

Most common coronaviruses tend to cause mild upper respiratory tract (URT) illnesses, and occasionally attack the intestines. However, the highly pathogenic coronaviruses, such as SARS-COV-2 or SARS-CoV cause severe influenza-like symptoms that can progress to severe pathologies, such as acute respiratory distress syndrome (ARDS), pneumonia, renal failure, and death (Guan et al., 2020; Harrison et al., 2020).

The host response consists of aggressive inflammatory responses, in part responsible for the damage done to the airways. In a subset of patients, the inflammatory responses can progress to a hyper-inductive state, also known as cytokine release syndrome (CRS) or cytokine storm. This occurs when a large number of immune cells activate, and release inflammatory cytokines, activating more cells in return. Although the process can resolve on its own after the clearance of the virus, in severe cases it can persist for longer, and lead to tissue damage, and the pathologies listed above. This process is responsible for an increased level of mortality observed with COVID-19 for a subgroup of patients

(Costela-Ruiz, Illescas-Montes, Puerta-Puerta, Ruiz, & Melguizo-Rodríguez, 2020; Jung, Potapov, Chillara, & Del Sol, 2021; P. Mehta et al., 2020).

How SARS-CoV-2, and other CRS-causing viruses modulate immune responses is not completely understood. They have certain effector proteins they can use to influence or delay the type-I interferon response, one of the first lines of defence mounted against viral infections by the innate immune system (Channappanavar et al., 2019; Murira & Lamarre, 2016).

How these viruses, especially focusing on SARS-CoV-2 alter intracellular signalling and other networks in various tissues, is an area of active research (Bouhaddou et al., 2020; D. E. Gordon et al., 2020; Treveil et al., 2021; Zhou et al., 2020). Global collaborative efforts, such as the COVID-19 disease map consortium have been created to reconstruct the virus-host interactions aimed to combat the underlying causes of the currently ongoing pandemic (Ostaszewski et al., 2020).

1.4. Introduction to networks

1.4.1. Networks describe complex systems

Graph theory is a branch of mathematics that studies graph models used to describe relationships between certain objects. They are widely applicable and have found their way into many of the sciences, be they computer science, social sciences or biology. The first study of graph theory was written by Leonhard Euler, one of the most prominent mathematicians of all time. Euler tried to solve

something that sounds quite simple at first. A city sits on both sides of a river, and has two islands in the water, both connected to the riverbanks, and to each other by a bridge. The problem was: can one cross the city in a way that only involves crossing each bridge once? Euler's abstraction of the problem laid the fundamentals of graph theory, namely the establishment of what we know today as vertices or nodes, points and edges, or links. The names are often used interchangeably, but they are mostly used as vertices and edges when discussing graphs, and nodes and links when talking about networks.

In the last few decades network science has grown into its own discipline, dealing with complex problems from various fields. The true merit of this approach is its ability to make sense of systems in a way that cannot be done purely from knowledge of its constituents. Network science is interdisciplinary, data-driven and computational at its heart.

Most systems can be described and analysed with networks. Biological systems are no different. They can be described with networks where nodes represent the constituents of a biological system (e.g. genes), and where edges represent the relationships between them (e.g. inhibition). The specific types of nodes, and edges of different qualities can further nuance these systems.

1.4.1.1. Nodes

The nodes in a network are the members we are looking to connect, be they genes, proteins, metabolites or complete organisms. Their nature will determine the type of interactions we can use to connect them to each other. A phenomenon one can often encounter regarding nodes, is that often they can be put into different sets, on some qualitative trait - e.g. gender, whether an animal

is a predator or prey, or whether we are talking about a protein or RNA. Connecting the members of these sets grants us a multipartite graph, i.e. bipartite, tripartite etc. graphs.

1.4.1.2. Degree

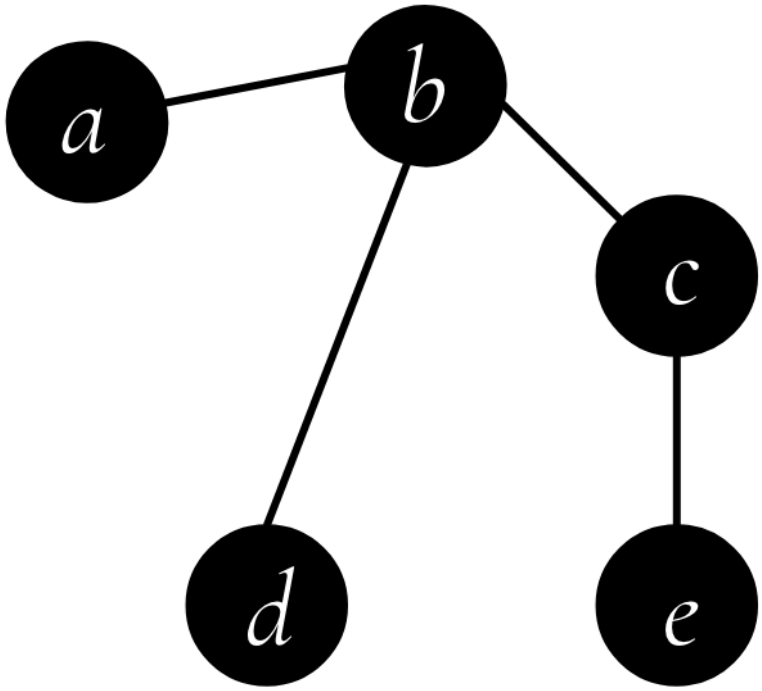

Figure 6. An example network. Node b has a degree of 3, as it has three immediate interactors.

One of the most commonly applied and important number used to describe a node is its degree. Let D_i be the degree of the i^{th} node in the network. It describes the number of direct links a node has to its neighbours, e.g. in the example above on Figure 6 node *b* has a degree of 3 ($D_b = 3$), node *c* has a degree of 2 ($D_c = 2$), while all other nodes have a degree of 1. To get the total amount of links in an undirected network, one can take the sum of all degrees, and divide it by two, as

to not count each interaction twice (e.g. between A and B, and B and A) (Barabási, 2016).

1.4.1.3. Hubs

Hubs are the highest degree nodes in the network, and have a degree larger than the average (Barabási, 2016). In biological networks they fulfil very important roles, their mutations often becoming lethal. Typical biological examples of hub nodes are chaperone proteins. They have a high degree, as they interact with many other proteins to help them fold into the correct shape, and a loss or mutation in a chaperone protein often leads to disorders, as the organism becomes prone to producing misfolded proteins (Macario, Grippo, & de Macario, 2005). In this thesis, regulatory hubs are discussed in Chapter III, in the context of correcting by degree during rewiring analysis.

1.4.1.4. Edges

Within network biology there are a few types of commonly used networks. On a molecular level most commonly one can find protein-protein interaction networks, where the links depict the physical interactions of proteins, or regulatory networks, that show the relationships of transcription factors and their regulated genes. There are supra individual level networks as well, depicting dynamics in ecology, like food webs (Dunne, Williams, & Martinez, 2002). Edges can be simple, showing an undirected interaction between two partners, but they can become more complex with the addition of more meta-data. They can be directed, determining the origin and target of the interaction, signed, signalling whether the interaction is stimulatory/inhibitory, weighted, conveying the importance or confidence in the interaction. Although we are only dealing with

simple graphs, where an edge can only connect two nodes, it would be remiss of me not to mention hypergraphs, where this is not the case. In these special constructs an edge can connect any number of nodes, which we then call hyperedges. They are less commonly used in biology, but there are certainly examples of it (Klamt, Haus, & Theis, 2009).

Biological networks, especially on a molecular level, are most of the time incomplete, they do not contain all interactions of the process they are attempting to depict. There are multiple reasons for this. First, simply not all interactions of all molecular constituents have been captured experimentally - either due to chance, or caused by technical limitations, interactions of certain proteins are harder to capture than others. Depending on the type of interaction, one can establish high-throughput experiments to gain as many interactors as possible, although these methods can have their blind spots as well, certain interactions they are unable to capture, rooted in the specific methodology used. A commonly used tactic to fill up these gaps, is to turn to the literature or network repositories, where one can collect missing interactions to complete their network, that were established by different experimental approaches (Türei, Korcsmáros, & Saez-Rodriguez, 2016).

Another major tool in our kit is the ability to predict/infer interactions between certain constituents, e.g. regulatory interactions where transcription factors can bind to putative target sites. These interactions are based on a set of heuristics or algorithms, and can fill in important gaps in the network, although they can add considerable noise as well (Bailey et al., 2009; Nguyen et al., 2018). As a rule of thumb, one should always strive to compile as complete of a network

as one can, as long as it fits the biological question, and the scope of the study. Having a more complete network can increase the predictive power of a network, and can lead to increased insight on our part (Santolini & Barabási, 2018). The nature of the field is, however, that new kinds of interactions can arise as long as there are meaningful relationships to analyse between things, and as such one can always find new ways of interpreting system level problems.

1.4.1.5. Paths

Paths are a sequence of nodes connected by a sequence of edges; they describe the steps needed to go from one node to another. An often-referenced special path is the shortest path, which describes the shortest paths by which nodes can be connected. This is an important attribute of the network, as it can be used to quantify certain properties of the network, such as information flow. Betweenness centrality is the measure of the shortest paths going through a given node or edge – the higher the value, the more information flowing through that specific node or edge. The diameter of the network is also quantified using the shortest path measure: it is the longest shortest path in the graph.

1.4.2. Protein-protein interaction networks

The quality of our queried interactors determines the kind of interactions one can distinguish. In biology, the most commonly used interactions are protein-protein interactions or PPIs.

Protein-protein interactions are the purposeful, non-random, physical interactions of proteins, occurring in or outside the cell. These interactions can happen between standalone proteins, or in complexes, and are responsible for,

amongst others, the signal flow in cells, and play a central role in the cellular systems of all living organisms. The breakdown of PPI signalling patterns can be indicative of a disease state (and thus, disease genes), play a fundamental role in drug discovery, and when comparing different organisms, can shine a light on the evolutionary path of those (Barabási, Gulbahce, & Loscalzo, 2011; De Las Rivas & Fontanillo, 2010; Kuzmanov & Emili, 2013).

Protein-protein interactions can be discovered by experimental techniques, or be inferred by computational approaches. One of the most widely used experimental approaches to describe interactions between proteins is the yeast two-hybrid screening method (Terentiev, Moldogazieva, & Shaitan, 2009). This approach permits the pairwise analysis of binding between two proteins, by expressing them in a *Saccharomyces cerevisiae* model. The binding of the proteins is inferred from the activation of the Gal4 reporter genes used. The yeast is grown on limiting media, and the proteins in question are fused with parts of the Gal4 transcription factor. If there is a close enough physical interaction between the queried proteins, the halves of the fused transcription factor combine, and Gal4 starts expressing, letting the microbe synthesize nutrients it needs to survive. The method has its limitations, especially when post-translational modifications have to be considered in the case of human proteins for example, but it has remained one of the mainstays of the methodology (Brückner, Polge, Lentze, Auerbach, & Schlattner, 2009); (Maple & Møller, 2007).

Other experimental approaches, like the affinity purification mass spectrometry (AP-MS) allow the detection of stable interactions, in a high-

throughput way. In essence, in these experiments the protein we would like to find partners for is tagged (thus becoming the bait) and is selectively purified along putative interaction partners (the preys) from the *in vivo* source (i.e. cells or tissue cultures). These purification steps are repeated for many sets of potential preys, and each of them is analysed with a mass spectrometer. From the results one can deduce the protein-protein interactions, and part of the underlying protein-protein interaction network (Gavin et al., 2002; Ho et al., 2002; Kim, Sabharwal, Vetta, & Blanchette, 2010; Rigaut et al., 1999; Tian, Zhao, Gu, & He, 2017).

There are multiple computational approaches used to predict protein-protein interactions (Obenauer & Yaffe, 2004). Methods using the amino acid sequence data typically utilise machine learning methods, such as random forest and support vector machines that attempt to predict interactions from pairs of protein sequences. Approaches using comparative genomic data are similar, but take sequence comparisons into account, and look at the conservation of gene neighborhoods, gene fusion, and gene co-occurrence (Kotlyar, Rossos, & Jurisica, 2017). Other approaches use protein domain information, or even the tertiary structure of proteins to infer interactions. Recently, approaches started integrating these data types, and basing interaction predictions on the combination of these (Q. C. Zhang, Petrey, Garzón, Deng, & Honig, 2013). Figure 7 shows an example protein-protein interaction network, from the seminal study of Gordon et al. 2020, mapping the protein-protein interactome of SARS-CoV-2 proteins in the host.

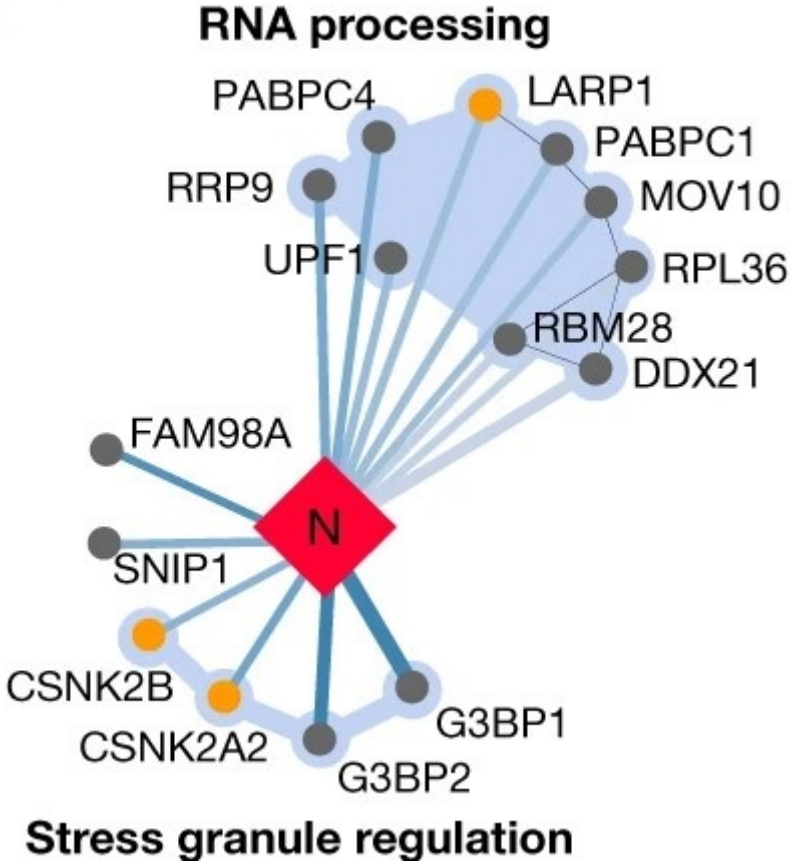

Figure 7. A protein-protein interaction network showing a SARS-CoV-2 protein (protein N, in red) interacting with human proteins. Putative drug targets are coloured in orange. Figure modified, from (D. E. Gordon et al., 2020)*, with permission of the rights holder Springer Nature.*

1.4.3. Gene regulatory networks

Gene regulatory networks (GRNs) describe the interactions of molecular regulatory elements that control the expression of RNA, and in turn, the levels of specific proteins. These regulators are most often proteins themselves, transcription factors, that can act alone, or in complexes with other proteins or nucleic acids. Gene regulatory networks control many of the cellular decisions, responses to stimuli, and their nature also makes them the most

dynamic of the commonly studied interaction times, from an evolutionary point of view. The edges of GRNs therefore contain interactions between transcription factors, and their regulated target genes.

Compared to eukaryotes, the genes of prokaryotes are organised in special, co-regulated transcriptional units called operons. The genes contained in operons are transcribed and controlled together, with the help of the regulatory sequences found next to the transcribed units, in the so-called untranslated regions (UTR), both up- and downstream from the structural genes (5`-UTR and 3`-UTR) (Mao, Dam, Chou, Olman, & Xu, 2009). The UTR contains the promoter region, a specific region of the DNA that can bind the RNA polymerase, to initiate the transcription of the genes into RNA (Kröger et al., 2012). The promoter region transcription factors bind to is specific and sensitive to changes, which makes this interaction layer dynamic, as even a small change, introduced by a point mutation can affect the binding affinity (Shou et al., 2011). Recently, novel techniques have been developed that can identify the first nucleotide of a transcript, termed the transcriptional start site (TSS). The approach called differential RNA-sequencing (dRNA-seq) can identify individual -10 and -35 promoter motifs (Sharma et al., 2010). This approach was applied with great success to *Salmonella*, identifying the TSS of major virulence regulators in *Salmonella*, such as *phoP, slyA,* and *invF* (Kröger et al., 2013, 2012).

Novel regulatory interactions of transcription factors and target genes can be uncovered in several ways. A commonly used experimental method is the chromatin immunoprecipitation sequencing (ChIP-Sequencing or ChIP-Seq) technology, which is a kind of sequencing method that looks for interactions of

proteins with DNA. In brief, following the binding of proteins to DNA by a binding agent, DNA is sheared, and the target proteins (the transcription factors of interest) are captured using specific antibodies against the protein. The bound DNA is recovered, and sequenced, highlighting the genomic regions the protein(s) bound to (Furey, 2012).

From a practical point of view, the binding sequences transcription factors recognise are usually stored as position-specific scoring matrices. These are matrices aligned from sequences representative of the binding region, and contain the probability of each base occurring at each position in the sequence. One way of visualizing them is by using sequence logos, where the size of the base letters are proportionate to the probability of them occurring in the binding region (Nguyen et al., 2018). Figure 8 shows the sequence logo for the binding site recognised by the Fur transcription factor in *Salmonella*.

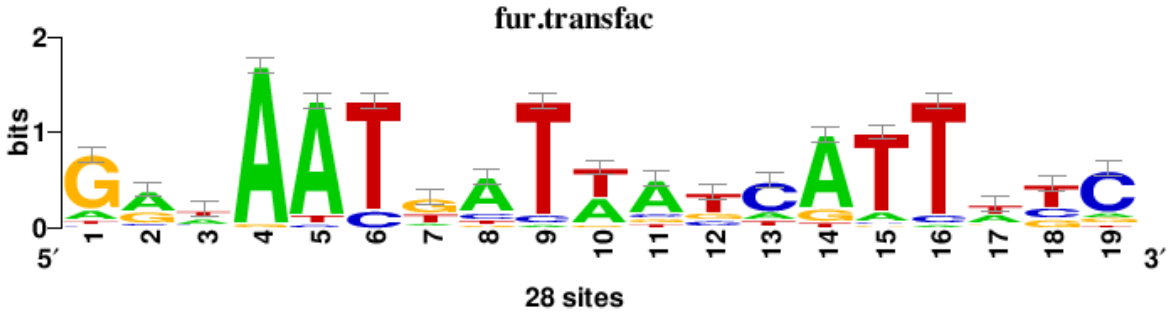

Figure 8. Sequence logo of the binding site recognised by the Salmonella transcription factor Fur, generated by the RSAT suite.

There are *in silico* ways of inferring or predicting regulatory interactions between select transcription factors and target genes, based on gene expression and other -omics data. For eukaryotes, there are a number of databases that contain high quality regulatory interaction data for *Homo sapiens*, such as

OmniPath, SignaLink from our group, and TRRUST or HTRI (Bovolenta, Acencio, & Lemke, 2012; Fazekas et al., 2013; Han et al., 2018; Türei et al., 2016). When studying bacteria, there are a few resources that are especially useful in this regard. RegulonDB and CollecTF collect transcription factor binding site data for *Escherichia coli* and other prokaryotes that can be used to predict and infer regulatory interactions (Kılıç et al., 2016; Santos-Zavaleta et al., 2019). Through the principle of regulogs, the homology-based conservation of transcription factors, target genes and transcription factor binding sites the inference of regulatory interactions is made possible, provided the interacting partners are well conserved, both on the level of proteins, and the interacting protein-DNA interface (Rodionov, 2007; H. Yu et al., 2004). The RSAT suite is a collection of on-line bioinformatic tools that make it possible to make the predictions based on the data from RegulonDB and CollecTF for example, by combining it with promoter data from the genomes of interest (Nguyen et al., 2018; Rodionov, 2007; H. Yu et al., 2004).

Gene regulation does not only exist in the form of transcription factor – target gene interactions, but other elements can also influence the expression of genes as well, for example on a posttranscriptional level. *Salmonella* small RNAs (sRNAs) have been identified previously and can alter the expression of a large number of genes in the pathogen. There are multiple interaction databases containing a posttranscriptional layer of regulation, and there is evidence of conservation of sRNAs within *Enterobacteriaceae* family (Kröger et al., 2012; Van Assche, Van Puyvelde, Vanderleyden, & Steenackers, 2015). Due to the amino-acid sequence based orthology of the database sRNAs were not added to this release of SalmoNet, but their addition will be an important upcoming step for the longevity of the database.

1.4.4. Metabolic networks

The metabolic networks discussed in this work are derived from genome-scale metabolic models (GEMs).

GEMs are computational models used to describe associations of genes and proteins to reactions for entire metabolic pathways in an organism. They collect existing knowledge of the metabolism for the organism, and most of the time they are assumed to be complete. They can be used to simulate and predict metabolic fluxes for various systems-level metabolic studies. The first GEM was created for *Haemophilus influenzae* in 1999, shortly after its genome was sequenced, and in the following years the number of GEMs for model and non-model organisms has grown considerably. As of February of 2019, there were more than 6000 models available, mostly for bacteria (Edwards & Palsson, 1999; Gu, Kim, Kim, Kim, & Lee, 2019; E. J. O'Brien, Monk, & Palsson, 2015). GEMs have many uses, and have found their way into many fields of biology, as they can be used to redesign aspects of the metabolism of a bacteria to enhance the production of certain desired metabolites, can be applied to study the essentiality of genes, to find oncogenes and biomarkers of cancer in systems medicine, and much more (Gu et al., 2019; C. Zhang & Hua, 2015). One can find GEMs in specific online repositories, such as BiGG or BioModels (King et al., 2016; Malik-Sheriff et al., 2020).

GEMs, from a practical point of view, are networks, where nodes constitute metabolites, and they are connected to each other by reactions, each associated

with the necessary enzymes. To represent stoichiometric coefficients, GEMs use a matrix (S matrix) in addition, to represent all the coefficients for all metabolic reactions. A positive coefficient means the metabolite is produced, while a negative means it is consumed (E. J. O'Brien et al., 2015; C. Zhang & Hua, 2015).

Flux Balance Analysis (FBA) is a type of constraint-based reconstruction and analysis (COBRA) method used to calculate the flow of metabolites through a genome scale metabolic network, from a network input to a network output. The output of the analysis is essentially a map showing that under certain parameters how the system must balance itself to achieve a homeostatic state. The results obtained from the analysis can be used to predict the growth rate of the organism as a whole, or a specific metabolite (E. J. O'Brien et al., 2015; Orth, Thiele, & Palsson, 2010).

The curated metabolic model for *Salmonella* published by Thiele et al. has been widely used since its release. Recently, a new set of metabolic reconstructions has been released, generating 410 metabolic models for 64 serovars. The authors used these results to show how different nutrient conditions showed the catabolic capacities of the studied strains, and what their optimal growth environments are (Seif et al., 2018; Thiele et al., 2011).

1.4.5. Multi-layered networks

Multi-layered biological networks can show connections between multiple networks belonging to the same system which can make them quite descriptive. The analysis of integrated networks (ones that combine multiple

levels of knowledge, e.g. regulation and protein-protein interactions) allows us to gain new insight into regulation, signal transduction on multiple levels. We can focus on specific processes without excluding complete levels of a biological system, e.g. to see whether a signalling pathway changes anything on a metabolic level with its downstream effectors. This is especially useful in this case, when we know of two similar but very differently behaving groups (Csabai, Ölbei, Budd, Korcsmáros, & Fazekas, 2018). Figure 9 shows the schematic representation of a multi-layered network.

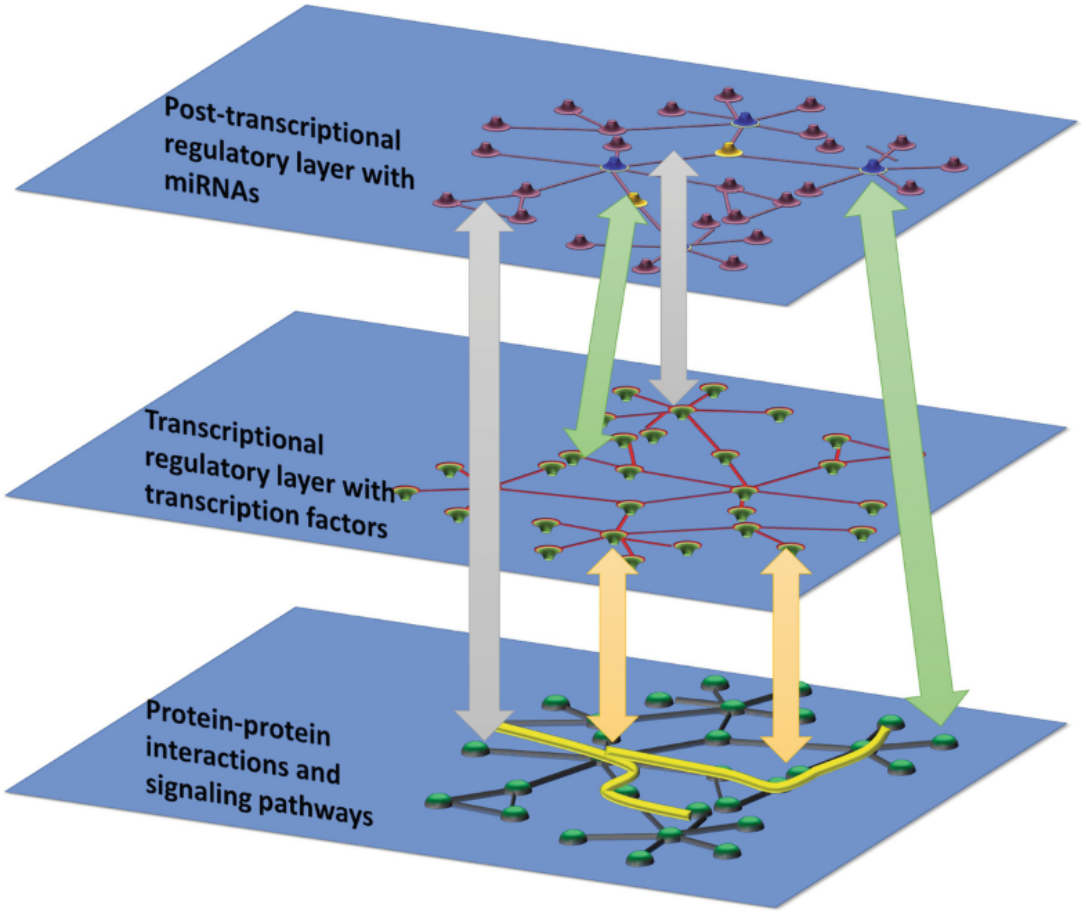

Figure 9. Schematic representation of a multi-layered network. Specific nodes connect different levels of information, leading to the multi-layered structure. Figure from Csabai et al., 2018.

1.4.6. Network properties

There are many ways of describing the properties of networks beyond its constituents. The degree is one of the most important characteristics of a node, and similarly, the degree distribution of a network can tell us a lot about the system we are attempting to model. Looking at the entire network, if the degree distribution of all nodes in the network follows a power law, we note these networks as scale-free. Most real networks - e.g. many biological networks, social networks, computational networks - fit or approach this distribution. Figure 9 shows the out-degree distribution of one of the networks from SalmoNet.

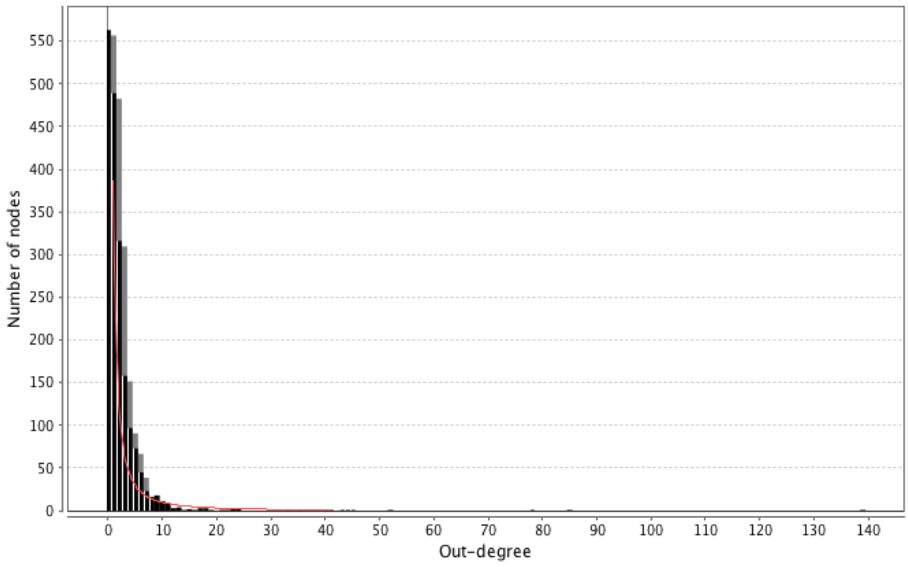

Figure 10. Out-degree distribution of the consensus network from SalmoNet 1.0. This network captures interactions shared amongst all included strains. The degree distribution approaches a power law (correlation: 0.96, R2: 0.82).

The largest interconnected, non-disjunct part of the network is called the giant component. This is where most connections lie, and most analysis takes place. Within these giant components one can often find modules. These are subgraphs, whose elements are more connected to each other, than to nodes

outside of it. They are often functional elements, and there are many approaches for finding them. One I employed in my PhD studies was the MCODE software developed for Cytoscape (Bader & Hogue, 2003).

1.5. Network resources

In computational biology, there are more and more options one can acquire network information from. There are various network repositories focusing on a specific types of interaction information, e.g. the Autophagy Regulatory Network (ARN) contains information related to the regulation of core autophagy proteins, ImmunoGlobe contains interactions occurring between various elements of the immune system, while other resources collate data from more specific databases like the ones mentioned previously (Atallah et al., 2020; Türei et al., 2015). An example for the latter is OmniPath or STRING, both of which contain multiple types of interaction data (e.g. protein-protein interactions, regulatory interactions, intercellular interactions) (Szklarczyk et al., 2019; Türei et al., 2016).

The structure and standardization of interaction data has made a lot of progress in recent years. To make sharing and processing information easier, more and more network resources utilize the PSI-MITAB system, a heavily standardized tabular format, where every field has a set function and values it can take.

1.5.1. Databases and network repositories

One of the most well-known interaction resources is the STRING database. It fills an important role in the ecosystem of network resources, as it contains many interactions, for >5000 species, from a diverse array of sources, ranging from very high quality, experimentally validated interactions, to interactions based on co-expression, co-occurrence and text mining. As such, it has a very large coverage of interactions, but importantly, it only focuses on protein-protein interaction data (Szklarczyk et al., 2019).

IntAct is a molecular biology interaction database, focusing on protein-protein interactions, curated from the literature, or directly from data depositions. The developers established a curation tool, allowing the users to fine-tune the quality of data they would like to work with, which is something I utilised in the development of SalmoNet 2 (see chapter 2) (Orchard et al., 2014) .

OmniPath is a database of literature-curated human signalling interactions. It was compiled of 34 resources, including both directed, signed and causal interactions. Released in 2016, the resource covers 39% of the human proteome, 61% of disease-gene associations, >80% of cancer related genes and druggable proteins.

OmniPath 2, its novel update now combines over 100 resources into a single database, covering inter-and intracellular signal transduction, as well as transcriptional and post-transcriptional regulation (miRNA-mRNA) (Türei et al.,

2016). This is a database I have used extensively in my work related to SARS-CoV-2 (see Chapter 5) and helped the publication of the updated manuscript by adding a degree of quality control and authored workflows for the R and python programming language access point libraries for the users of the resource.

1.5.2. Network analysis and visualization software

One of the most commonly used network analysis software in biology is Cytoscape (Shannon et al. 2003). It is an easy to use, free graphical user interface (GUI) application, capable of reading and writing multiple network file formats, and can be used both as a network analysis and visual exploration tool. One of the largest advantages of the software is its community, that develops additional functionality for it in the form of Cytoscape apps. I have used a number of these libraries during my PhD work, i.e. DyNet, MCODE, CHAT, ISMAGS (Bader & Hogue, 2003; Goenawan, Bryan, & Lynn, 2016; Muetze et al., 2016; Van Parys et al., 2017). The software also allows for automated analyses, through multiple libraries in commonly used programming languages, such as R or python (Otasek, Morris, Bouças, Pico, & Demchak, 2019).

More advanced network analysis and visualization tools exist as libraries in the R and Python languages, e.g. igraph, networkx, RCy3. These allow for automated and programmatic analysis of networks and are really important from the standpoint of reproducibility.

1.6. Primary research aims

My primary research aims in this thesis are the following:

1. Expanding the coverage of *Salmonella enterica* serovar strains found in the SalmoNet database and increasing the information content of the individual networks.
2. Validation of the SalmoNet approach: using experimental information to show the biological relevance of the included interactions.
3. Testing the scientific hypothesis set at the beginning of this thesis, that the difference in the host adaptation capabilities of gastrointestinal and extraintestinal *Salmonella* enterica serovars can be characterized by specific changes in the topology of their metabolic, regulatory, or protein-protein interaction networks, and highlighting how network comparison and network analysis workflows can be used to identify elements of molecular interaction networks under selection, stemming from their lifestyle or environment.

Added aims due to the COVID-19 pandemic:

1. Carry out a study comparing cytokine responses of patients infected by Cytokine Release Syndrome causing viruses, including SARS-CoV-2.
2. Generate an interaction network of cell-cell communication mediated by cytokines, aimed at uncovering leading intercellular interactions of CRS.

2. Construction of a multi-layered network database for *Salmonella* research

2.1. Introduction

Salmonella enterica is an important group of pathogens endangering the health of humans and other animal species alike. *Salmonella enterica* subspecies *enterica* houses over 2000 serovars, and is responsible for the majority of *Salmonella* infections in warm blooded animals (A. J. Bäumler, Tsolis, et al., 1998). Most of the subspecies I. serovars have a broad host range, and cause acute gastroenteritis in the host organism (Uzzau et al., 2000). Gastrointestinal *Salmonella* serovars induce this self-limiting gastroenteritis to engineer the gut luminal environment to one that benefits them, by inducing the release of metabolites these pathogens can use as terminal electron acceptors, and increasing the oxygen saturation of the gut lumen from the production of reactive oxygen species by the cellular elements of the immune system (Nuccio & Bäumler, 2014). The latter, although still harmful to the pathogen, paradoxically, enables the growth of gastrointestinal *Salmonella,* by reducing the anaerobic stress on the pathogen (Nuccio & Bäumler, 2014; Rogers, Tsolis, & Bäumler, 2021). A smaller subgroup of *Salmonella* serovars has adapted to a systemic

lifestyle, causing bacteremia and enteric fever. The adaptation to invasive disease markedly alters the pathogenesis process, symptoms, and immune responses to these Salmonella serovars. They are atypical bacteria, their virulence factors focusing on inhibiting the generation of normal antibacterial responses, leading to a "stealth" phenotype (Tsolis, Young, Solnick, & Bäumler, 2008).

To understand the changes the genus and these host adapted serovars went through, many studies have focused on genomic differences. Since one of the hallmarks of host adaptation is a level of genome degradation, and genetic content gain through horizontal gene transfer, these approaches have highlighted the genetic elements that underwent selection, and consequently functions that were lost or potentially gained through diversification (den Bakker et al., 2011; Robert A Kingsley et al., 2013; Klemm et al., 2016).

Studying *Salmonella* host adaptation also means studying convergent evolution, and there are a number of examples, detailed in Chapter I., where different serovars arrived at similar solutions, through different molecular level means. It stands to reason that to understand the underlying reasons and mechanisms of host adaptation, we should also analyse functional convergence, and compare these serovars on a systems level. Although the contraction and expansion of genomes is the process that gives rise to the functional changes, it is the way those absent or newly acquire genes modify and fit into the system, that makes *Salmonella* alter its behaviour within the host organism, through the emergence or loss of novel signalling, regulatory or metabolic pathways, or the combination of them. The combination of this information via multi-layered networks allows us to focus on specific processes important to the question at

hand, e.g. to see whether a regulator can affect a metabolic pathway further downstream (Csabai et al., 2018).

For *Salmonella*, the different levels of knowledge exist in separate data resources, which makes complex, integrated analysis difficult. SalmoNet was developed for *Salmonella* strains to circumvent this problem. SalmoNet is the first multi-layered network database for *Salmonella*, combining regulatory, metabolic and protein-protein interactions for 10 *Salmonella* serovars in the first version, and 20 strains in the second one. In addition to being a tool for a specific scientific purpose, the study of host adaptation in extraintestinal serovars of *Salmonella*, it also aims to be a gap-filling knowledgebase for this non-model organism. The networks contain manually curated *Salmonella* specific interactions, and inferred interactions from *Escherichia coli*. Data was collated from multiple sources: literature, primary and secondary databases, high-throughput experiments.

This chapter describes how the first version of the database was built, focusing on the workflow, principles and methods utilized to collate the networks, and details the steps how I updated it with the release of the new version. The work highlighted in the introduction of this chapter were carried out by: Aline Métris who designed much of the work, Padhmanand Sudhakar who carried out the construction of regulatory networks, David Fazekas, who created the protein-protein interaction and metabolic layers and set up the web resource (http://salmonet.org/). Amanda Demeter performed the manual curation of interaction data sources, Eszter Ari, who contributed by inferring the classification trees and dendrograms. Priscilla Branchu, Rob A. Kingsley, Tamas

Korcsmaros and Jozsef Baranyi contributed to framing the biological basis of the work, and supervised the project. My role in the first release was internal testing and quality control, as this project finished not too long after I started my PhD studies.

2.1.1. Construction of a multi-layered network for non-model organisms

SalmoNet 1 was the first multi-layered network resource for the pathogenic non-model organism *Salmonella*. To have a better chance of understanding how members of this phenotypically diverse group differ from each other, there was a need to combine various levels of information together, to make integrated analysis possible.

The genus holds a lot of diversity, and ten well-studied strains were selected to capture this, by including five host-adapted and five broad host range serovars. Since the majority of *Salmonella* information used to exist in separate repositories, the database fulfilled the important role of a *Salmonella* specific knowledgebase for these strains, beyond being an interaction resource. We achieved this through integrating various levels of knowledge from multiple data sources and approaches: protein-protein, regulatory and metabolic information, both predicted and experimental, from high-throughput and low throughput experiments, and the available literature (Métris et al., 2017).

2.1.1.1. Strains included in SalmoNet 1

In SalmoNet 1, five broad host range, gastrointestinal serovars, and five narrow host range, extraintestinal serovars were selected from subspecies

enterica. Table 2 lists the included representative strains, and information regarding their lifestyle.

Serovar	Strain	Taxonomy ID	Lifestyle
Typhi	CT18	90370	Extaintestinal, causes typhoid fever in humans
Paratyphi A	ATCC 9150	295319	Extaintestinal, causes paratyphoid fever in humans
Choleraesuis	SC-B67	321314	Extaintestinal, porcine adapted
Dublin	CT 02021853	439851	Extaintestinal, bovine adapted
Gallinarum	287/91	550538	Extaintestinal, avian adapted
Agona	SL483	454166	Gastrointestinal
Enteritidis	P125109	550537	Gastrointestinal
Heidelberg	SL476	454169	Gastrointestinal
Newport	SL254	423368	Gastrointestinal
Typhimurium	SL1344	216597	Gastrointestinal
Typhimurium	LT2	99287	Gastrointestinal

Table 2: list of serovars in the first version of SalmoNet.

2.1.1.2. Prediction of interactions across organisms, orthology mapping

One of the challenges many face doing comparative (micro)biological work is that despite our best efforts, name and various identifier usage can be inconsistent across strains and serovars. Orthology mapping can provide a common denominator, by homology-based clustering of the protein sequences that serve as nodes in the final networks (Altenhoff et al., 2016; Remm, Storm, &

Sonnhammer, 2001). Every node, regardless of which layer they belong to, is therefore treated as a protein.

For SalmoNet 1, the standalone software version of InParanoid was used to create orthologous relationships between the proteins of the *Salmonella* strains listed above, and the model organism *Escherichia coli* K12 (K. P. O'Brien, Remm, & Sonnhammer, 2005; Sonnhammer & Östlund, 2015). The latter is a close relative of *Salmonella*, and well-studied model organism. This makes it possible to include orthologous interaction data based on conserved proteins and the concept of interologs, the transfer of interaction annotation from one organism to another, and *E. coli* can as such act as an important link in transferring more established and well-studied information to a non-model organism, such as *Salmonella* (H. Yu et al., 2004).

To begin the orthology mapping, the complete protein sequences of all available genes belonging to the listed serovars were downloaded from the UniProt database, in January of 2015. To identify homologous protein sequences, InParanoid starts with an all-vs-all BLAST comparison of all protein sequences in two species and following that applies a clustering rules to build ortholog groups. As the authors of InParanoid summarize, "The purpose of the ortholog detection algorithm is to find non-overlapping groups of orthologous sequences using pairwise similarity scores. This is essentially a sequence clustering problem." (Remm et al., 2001). In brief, InParanoid first identifies the best scoring sequence pairs bi-directionally (since it is always comparing two proteomes at a time), and marks these as the main ortholog pair of a specific ortholog group. The detection of the subsequent orthologs follows independently

for each ortholog group, until the similarity scores reach the predetermined cutoff value. For SalmoNet 1, a strict sequence similarity cutoff of >= 95% was set to minimize false positives, as the comparison is made between strains of the same species. Previously >= 80% was used in other works when creating interologs between different species, e.g. *Caenorhabditis elegans* and *Drosophila melanogaster* (Remm et al., 2001; H. Yu et al., 2004).

2.1.2. Reconstructing the interaction networks

2.1.2.1. Protein-protein interactions

To create the protein-protein interaction layer, a guided literature curation protocol was used, originally developed for SignaLink (Fazekas et al., 2013) (Csabai et al., 2018). The workflow uses the tools iHop and ChiliBot in addition to direct PubMed searches to look for signalling interactions between *Salmonella* proteins (Chen & Sharp, 2004; Hoffmann & Valencia, 2005). In addition, experimentally verified *Salmonella* protein-protein interactions were included from the IntAct database (S Kerrien et al., 2007; Orchard et al., 2014). To further increase the coverage of the networks, interactions were transferred from a closely related model organism - *Escherichia coli* - based on the concept of interologs. Interolog mapping is the process of transferring annotation data, from one organism to another, based on orthologous relationships, established through InParanoid for SalmoNet 1.0 (Métris et al., 2017). The source of interologs were the Interactome 3D, IntAct, and BioGrid databases, and a high-throughput yeast-2-hybrid screen of the *Escherichia coli* interactome (Mosca, Céol, & Aloy, 2013; Oughtred et al., 2019; Sonnhammer & Östlund, 2015; Stark et al., 2011).

2.1.2.2. Metabolic interactions

Genome-scale metabolic networks describe the interactions (reactions) of metabolites, mediated by the various enzymes needed to process them. The metabolic networks included in this work are essentially an inversion of these genome-scale networks. They were defined as the following: if a metabolite is a product of a reaction, and a substrate in another, the two proteins catalysing the reactions are linked, with the exception for ones appearing in more than 10 reactions (Kreimer, Borenstein, Gophna, & Ruppin, 2008).

To construct these interactions the STM 1.0 model mentioned in Chapter I (Thiele et al., 2011), and automatically generated data from the BioModels database was used (BMID000000140711).

2.1.2.3. Regulatory interactions

SalmoNet 1 contains interactions based on both experimentally validated and predicted regulatory interactions, that represent the binding of transcription factors to promoter regions of specific target genes. As mentioned previously in Chapter I, the promoter region is a specific sequence of the DNA that can bind the RNA polymerase, to initiate the transcription of the genes into RNA. The promoter region transcription factors bind to is specific, and sensitive to changes,

To build the regulatory layer, first low throughput, experimentally validated data was collected on transcription factor binding sites. This was done similarly using ChiliBot and iHop as in the PPI layer, and relevant databases, such as CollecTF, RegulonDB or Prodoric (Gama-Castro et al., 2016; Kılıç et al., 2016;

Münch et al., 2003; Santos-Zavaleta et al., 2019). High-throughput experiment data can also be used to infer binding sites from specific genomic locations, in the case of SalmoNet 1 the results from Smith et al. 2016 were processed (Smith, Stringer, Mao, Palumbo, & Wade, 2016). To get statistically significant binding motifs from data like this, amongst others the MEME suite of tools can be used, specifically the MEME-ChIP that can extract binding data from ChIP-chip or ChIP-seq data (Bailey et al., 2009).

Taken together, the collected binding sites can be used to generate a Position Specific Scoring Matrix containing the consensus binding signature for that specific transcription factor. This was done with the *consensus* tool from the RSAT suite of tools (now deprecated, see below). Once formatted to transfac format (with RSAT *convert-matrix*), an input file format originally developed for the TRANScriptio FACtor database, one can start the genome wide scan of promoter regions (Wingender, 2008). The UTR regions were retrieved with RSAT's *retrieve-sequence* method, but can also be acquired using *bedtools* for example(Nguyen et al., 2018). Prokaryotic transcription factors typically bind to the noncoding regions starting upstream from the start codon of the fist gene located in the operon. For SalmoNet 1, the first 5000 base pairs upstream from the start codon were used, or smaller, should a gene sit in the overlapping region upstream. (Browning & Busby, 2004)

For each PSSM optimal P-value thresholds can be determined using the RSAT *matrix-quality* tool. This step, although not always necessary, can reduce the amount of false positive hits when PSSMs were constructed from a few sites, or have low information content otherwise. Finally, The RSAT *matrix-scan*

pattern matching tool synthesizes the results of the previous steps, and attempts to match the PSSMs to the promoter sequences, assigning p-values to each hit. Pattern matching is a computational process, during which predefined signatures (the PSSMs in this example) are used to find putative copies of the signatures in a target string (the promoter regions in this example) (Medina-Rivera et al., 2011; Olbei, Kingsley, Korcsmaros, & Sudhakar, 2019; Turatsinze, Thomas-Chollier, Defrance, & van Helden, 2008).

Orthologous information can also be used in the reconstruction of regulatory networks. Regulogs are sets of coregulated genes with conserved regulatory sequences across multiple organisms, which we can use to our advantage when generating networks for non-model organisms (Rodionov, 2007; H. Yu et al., 2004). With SalmoNet 1, experimentally verified *Escherichia coli* transcription factor - binding site pairs were downloaded from the RegulonDB database and checked for orthologous proteins - both on the side of the transcription factor and the target gene. If both are present, the downloaded binding site is matched against the regulatory region of the target gene, and the result is only included, if a hit is found. The presence of orthologous transcription factors, target genes, and the matching binding sites are the three conditions for regulog mapping, as introduced by H. Yu et al. (Cock et al., 2009; Rodionov, 2007; H. Yu et al., 2004).

2.1.2.4. Removal of pseudogenes

Salmonella strains when undergoing host-adaptation, tend to go through a degree of genome degradation, leading to a loss of function in numerous genes of importance (Bawn et al., 2020; Holt et al., 2009; Robert A

Kingsley et al., 2013). The precise annotation, and subsequent removal of these genes from our network is an important step, as otherwise we would keep false positives in the data, leading to interactions that should not be there. To remove all hypothetically disrupted coding DNA sequences (HDCs), the curation made by Nuccio & Bäumler was used to remove such entries (Nuccio & Bäumler, 2014).

2.1.2.5. Data formats & Website

To make the interaction data widely accessible, an interactive website was designed to showcase *Salmonella* interaction data: https://salmonet.org. The website allows the users to query information from proteins of interest, download these subgraphs directly, and look up orthologous proteins within SalmoNet 1, and other resources.

The interaction data was made available in a custom .csv and .cys formats, the latter being the input format for Cytoscape, a popular network visualization analysis and platform.

2.2. Aims

The aims of this project were the following:

- Assess the areas where the original SalmoNet database could be upgraded and extended.
- Identify the required changes in methodology for the new version.
- Develop the second version of the SalmoNet database.
- Compare the information content of the two releases.

2.3. Methods

The second half of this chapter describes the update resulting in the second version of the database. All work detailed, including network reconstruction, implementing changes in methodology, consequent analysis and interpretation of the results were carried out by myself.

2.3.1. SalmoNet 2

2.3.1.1. Motivations for the update

SalmoNet 1 contains a lot of information for *Salmonella* researchers, and the database aimed to cover the most prevalent strains studied in the field. By integrating regulatory, PPI, and enzyme-enzyme interaction information, the networks can provide a more exhaustive view of signal transduction in the system and could be used to highlight upstream regulators of genes involved in establishing infection and metabolic functions. However, there are a number of limitations to the resource, that I attempted to amend with the updated version.

First, the interaction database only contains information on proteins that have interaction partners in at least one of the layers. As such, understudied genes without any interactions captured can be left out of the study and bias the usability of the networks to more studied nodes. While this is still a limitation of the updated version, I reduced study bias in the networks to reflect the biological system more accurately by increasing the coverage of proteins involved in the networks through additional resources and quality control steps.

Secondly, the strains included in SalmoNet 1, while containing many well studied organisms, were restricted to just the *Salmonella enterica* species, and the *enterica* subspecies. With the update I wanted to make future studies possible, where the users could investigate the effects of greater evolutionary distance between the strains, made possible by the inclusion of more distantly related pathogens, such as the species *Salmonella bongori* and the *S. enterica* subspecies *arizonae*. This may provide further insight into the evolution of all Salmonellae, by highlighting conserved pathways or interactions (Fookes et al., 2011). Besides the interesting evolutionary questions made possible with the inclusion of these more distant relatives, one of the main goals of the SalmoNet project is to help researchers understand the very real problem of *Salmonella* infection in humans. The first release of the database contained only two human-adapted typhoidal strains, which I wanted to extend with other well understood pathogens causing disease in humans. The addition of further human pathogens could make more focussed research into host adaptation possible, as the most prevalent extraintestinal strains are phylogenetically more distantly related to each other, and the fact those included in SalmoNet 1 do not share host species adds another layer of complexity and noise to the question and analysis.

From the release of SalmoNet 1, strain of particular recent interest was omitted, *S.* Typhimurium D23580, associated with the invasive non-typhoidal *Salmonella* (iNTS) disease. This serovar currently causes significant mortality in many countries of sub-Saharan Africa, and as such is a subject of numerous studies (Canals et al., 2019; Carden et al., 2017; Robert A Kingsley et al., 2009; Owen et al., 2017; Singletary et al., 2016). The inclusion of D23580 could certainly inform much of the currently ongoing research and help inform studies on the specific differences arising in this novel pathogenic lineage.

Although the information in SalmoNet 1 can highlight elements under evolutionary pressure, or indicate important interactions between regulators and targets, especially the predicted interactions cannot be taken on face value alone and should be used as a list of potential targets for molecular biology testing. While this is still the case with the updated version of the database, I wanted to increase the information content underlying interactions where this was possible, to increase confidence in interactions as much as possible, and compare the obtained results with published information to assess their reliability.

While including metabolic interactions was very important step, the first release of this level of knowledge was incomplete. Due to a technical error, SalmoNet 1 only contained enzyme-enzyme interactions derived from irreversible reactions. I wanted to extend these to contain ones from reversible reactions as well, and more importantly, use a now updated background model as its basis. In their seminal work Seif et al. have developed strain specific metabolic models for all of the *Salmonella* strains I planned on including and more (Seif et al., 2018; Seif, Monk, Machado, Kavvas, & Palsson, 2019).

One of the key motivations for the new release of the database was the Uniprot Proteome Redundancy project (https://www.uniprot.org/help/proteome_redundancy) that affected the utility of SalmoNet 1. For various reasons, many primarily prokaryotic sequences in the Uniprot database became redundant, or were assigned new accession numbers. SalmoNet 1 was primarily UniProt based, and the Uniprot Proteome Redundancy Project made parts of the SalmoNet 1 dataset progressively less user friendly as time went by, as users had to look up deprecated IDs, and match them up with new ones, which became more difficult going forwards. I therefore wanted to improve the annotation data within the database, and add more strains, including

other widely used lab strains, human pathogens and iNTS strains which were requested from the members of the *Salmonella* community.

Although the main structure of the database remained the same, the underlying workflow changed. Figure 11. details the sources of information and layers contained in SalmoNet 2.

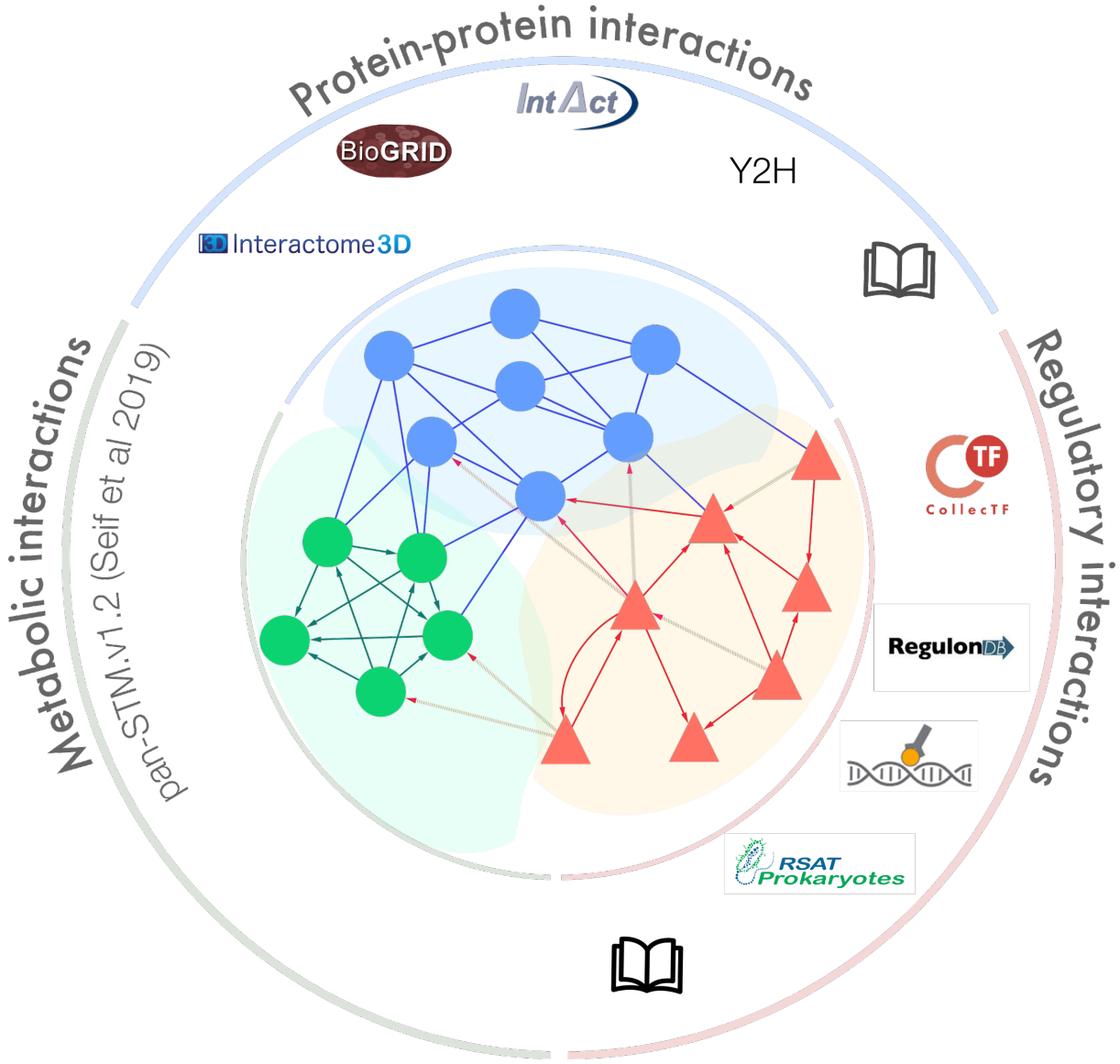

Fig 11. Interaction sources and layers in SalmoNet 2.

2.3.1.2. Orthology mapping using the OMA database

In SalmoNet 2 I used the OMA ("Orthologous Matrix") standalone software to construct the orthologous relationships between the available *Salmonella* strains from the OMA browser database. OMA is a large-scale orthology database and toolkit, containing much of the information we need for SalmoNet in one place, including the proteomes and genomes of the strains on request, and important annotation data (Altenhoff et al., 2018).

The reason for this change was the ease of use, extra information provided by the OMA database, and making it easier to generate minor and major releases for the future. Relying on an external database for the maintenance of a seperate database has advantages and disadvantages. The advantages are, an additional level of quality control, and an extra resource to refer to when wanting to compare or look for context on interaction data. Using a database that specialises on orthologous data is especially useful, as OMA contains data on 1688 bacterial and 153 archaeal genomes (as of the August 2020 release), and as such any studies wanting to look up how the results they obtained from SalmoNet compare to the rest of the prokaryotes have an easier time doing so. The disadvantages are, that databases are only as good as their maintenance, and if any of the pillars supporting them (such as another database) stops being updated, it can hinder the future of the resource (Merali & Giles, 2005). Seeing as OMA has gained multiple releases since I started using it, I was confident in its future.

The OMA algorithm is similar to the previously employed InParanoid method, as both use the pairwise amino acid sequence similarity to determine the orthologous status of proteins from the compared species. Once all putative

orthologs or stable pairs between all species or strains were found, OMA builds a network from the pairwise orthologs. In this network, the authors defined “OMA Groups” as cliques in the graph, where each node in that subgraph is connected to all other nodes in the same subgraph, the resulting part containing groups in which all the genes are connected to each other via pairwise orthologus relationships. These OMA groups, where all genes are orthologous to each other, were used as the template for orthologous relationships in SalmoNet 2. Although the clique approach is quite stringent, as just the loss of one edge in the subgraph can eject a gene from a group, the authors have implemented a tolerance parameter to combat this, the resulting structures termed “quasi-cliques” (Altenhoff et al., 2019; Train, Glover, Gonnet, Altenhoff, & Dessimoz, 2017; Zahn-Zabal, Dessimoz, & Glover, 2020).

Orthology mapping is the most computationally intensive step of the SalmoNet workflow. The OMA standalone software can save a lot of time and resources here. First, the all-against-all Smith-Waterman sequence alignments can be parallelised, both on single computers or high-performance clusters. Adding a new strain or species in the future is also made easier, as OMA Standalone does not require an all-against-all recomputing of the orthologous relationships in these cases, as the previously used pre-computed results can be submitted, in which case only the new genomes require computation time. What this means in practice, that a requested strain can be added, generated and compared much quicker and easier than in the previous version. In this iteration I ended up using the strains already in OMA, since they included all the requested strains, but from the perspective of database longevity it is an important step for the future.

It is important to note, that the outputs of the orthology prediction tools can be slightly different: according to a study comparing these methods the OMA standalone output OMA groups lead generally to more precise, but also more strict mapping, leading to less false positives (and true positives as well) (Altenhoff et al. 2016). I did however get very similar, and in cases better recall than with SalmoNet 1.0 (between 69-75% overlap with the 4140 proteins from *E. coli*) using InParanoid. The following table contains the list of included strains in SalmoNet 2, and the overlap of their respective orthologous proteins with *Escherichia coli*.

Strain	**Five letter code**	**Orthologous protein overlap with *E. coli***	**Percentage match**
Salmonella enterica subsp. *Enterica* serovar *Agona* str. *SL483*	SALA4	3016	72.8%
Salmonella enterica subsp. *Arizonae* serovar *62:z4,z23:-*	SALAR	2859	69.1%
Salmonella bongori NCTC 12419	SALBC	2961	71.5%
Salmonella enterica subsp. *Enterica* serovar *Choleraesuis* str. *SC-B67*	SALCH	2987	72.1%
Salmonella enterica subsp. *Enterica* serovar *Dublin str. CT 02021853*	SALDC	2983	72.1%
Salmonella enterica subsp. *Enterica* serovar *Enteritidis* str. *P125109*	SALEP	3092	74.7%
Salmonella enterica subsp. *Enterica* serovar *Gallinarum* str. *287/91*	SALG2	3075	74.3%
Salmonella enterica subsp. *Enterica* serovar *Heidelberg* str. *SL476*	SALHS	3044	73.5%
Salmonella enterica subsp. *Enterica* serovar *Newport* str. *SL254*	SALNS	3033	73.3%
Salmonella enterica subsp. *Enterica* serovar *Paratyphi A* str. *AKU 12601*	SALPK	3006	72.6%
Salmonella enterica subsp. *Enterica* serovar *Paratyphi A* str. *ATCC 9150*	SALPA	2960	71.5%

Salmonella enterica subsp. *Enterica* serovar *Paratyphi B* str. *SPB7*	SALPB	3077	74.3%
Salmonella enterica subsp. *Enterica* serovar *Paratyphi C str. RKS4594*	SALPC	2996	72.3%
Salmonella enterica subsp. *Enterica* serovar *Schwarzengrund* str. *CVM19633*	SALSV	2993	72.3%
Salmonella enterica subsp. *Enterica* serovar *Typhimurium* str. *14028S*	SALT1	3109	75.1%
Salmonella enterica subsp. *Enterica* serovar *Typhimurium* str. *ST4/74*	SALT4	3110	75.1%
Salmonella enterica subsp. *Enterica* serovar *Typhimurium* str. *SL1344*	SALTS	3107	75%
Salmonella enterica subsp. *Enterica* serovar *Typhimurium* str. *D23580*	SALTD	3095	74.8%
Salmonella enterica subsp. *Enterica* serovar *Typhi* str. *CT18*	SALTY	3013	72.8%

Table 3: List of strains in SalmoNet 2, and the overlap of the orthologous proteins with that of Escherichia coli, used as a measure of recall.

Using OMA is not only beneficial for the orthology mapping, it is also really helpful for the re-annotation of proteins. As mentioned before, the first version of SalmoNet was essentially UniProt based, with UniProt IDs serving as the primary identifiers of the database. Because of the UniProt Redundancy Project, we found ourselves in a state where not all proteins of all strains have a matching, active UniProt ID. This is where the OMA IDs come in as our new primary identifier.

2.3.1.3. Re-construction of network layers

Once I recreated the orthologous relationships of the *Salmonella* strains listed above, I began reconstructing the network layers. I followed the protocols described above, developed for SalmoNet 1, and as described in (Métris et al., 2017; Olbei et al., 2019). The workflow for the transcriptional regulatory and protein-protein interaction layers remained largely the same.

2.3.1.4. Protein-Protein Interaction layer

For the protein-protein interactions (PPI), when sourcing orthologous relationships based on *E. coli* information from IntAct, I have used the scoring IntAct has developed for their experimental interactions to filter the incoming data, to increase reliable coverage. This scoring contains the weighted cumulated value assigned to each interaction based on detection method (e.g. biochemical, biophysical, imaging, etc.) and interaction type (association, physical association, etc.). In the past, for the first version, only one specific kind of detection method `psi-mi:"MI:0096"(pull down)` was used to filter interactions. Most experimentally validated interactions are still captured by this method, but the novel scoring system allows us to select the higher quality ones based on additional information and use interactions from other methodologies that would have been left out otherwise, such as tandem affinity purification (`psi-mi:"MI:0676"`). Tandem affinity purification (TAP) is a molecular biology method for discovering physical interactions of proteins, through immunoprecipitation. In short, the proteins of interest are tagged, and the tagged fusion proteins are expressed as normal in the cell, where they may interact freely with their normal interacting partners. Following that the tagged proteins are

separated using beads coated with an antibody that specifically binds the tag antigen (Gavin et al., 2002; Gully, Moinier, Loiseau, & Bouveret, 2003).

The following paragraph is an excerpt from the intact website, describing how the IntAct scores are attained:

> *How is the intact-miscore calculated?*
>
> *The IntAct MI score is based on the manual annotation of every instance of a binary interaction (A-B) within the IntAct database. First all instances of the A-B interacting pair are clustered by accession number. Each entry has been annotated using the PSI-CVs and we use this information to score by the interaction detection method and by the interaction type. Additionally we count the number of publications the interaction has appeared in, up to a maximum of 8. Each of these variables is normalised between 0-1. The cumulative score is also normalised between 0-1 across the entire IntAct database, with 1 representing an interaction in which we have the highest confidence. From:* https://www.ebi.ac.uk/intact/pages/faq/faq.xhtml

This increased the amount of *Escherichia coli* interactions I could create interologs from, by using interactions that had an MI score > 0.50. Figure 12 shows the frequency distribution of PSI-Miscores of the *Escherichia coli* data in IntAct.

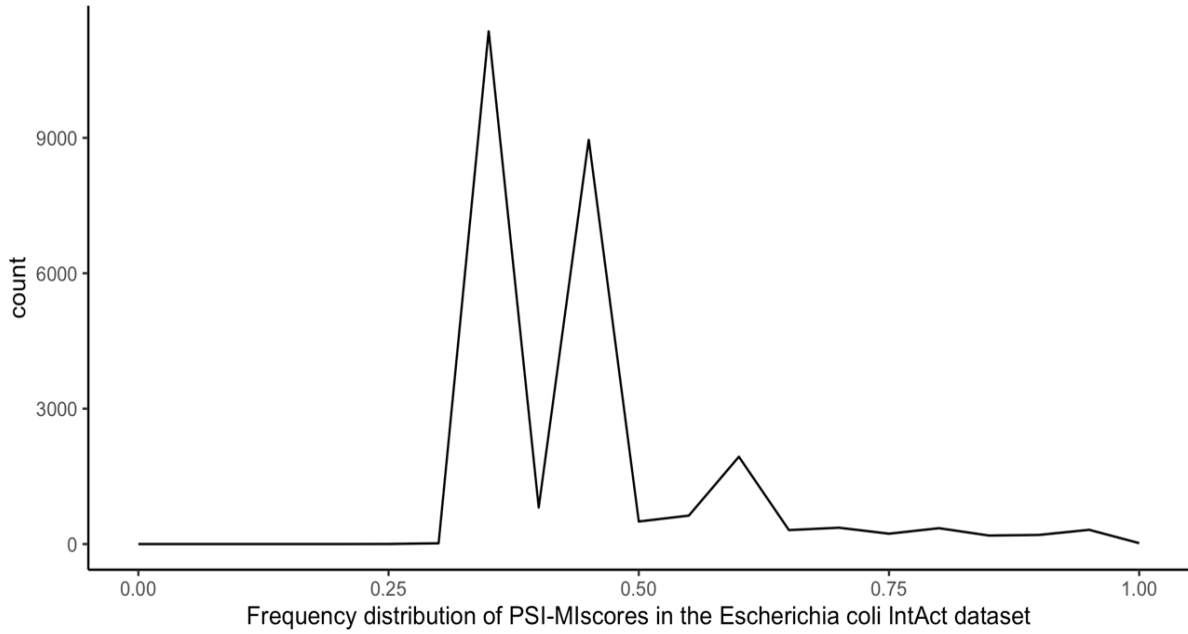

Figure 12. Frequency of PSI-Miscores in the Escherichia coli IntAct data.

2.3.1.5. Transcriptional regulatory layer

The establishment of the transcriptional regulatory networks was done in an identical way to SalmoNet 1 but with updated input information. Figure 13 shows the workflow for the construction of the regulatory layer.

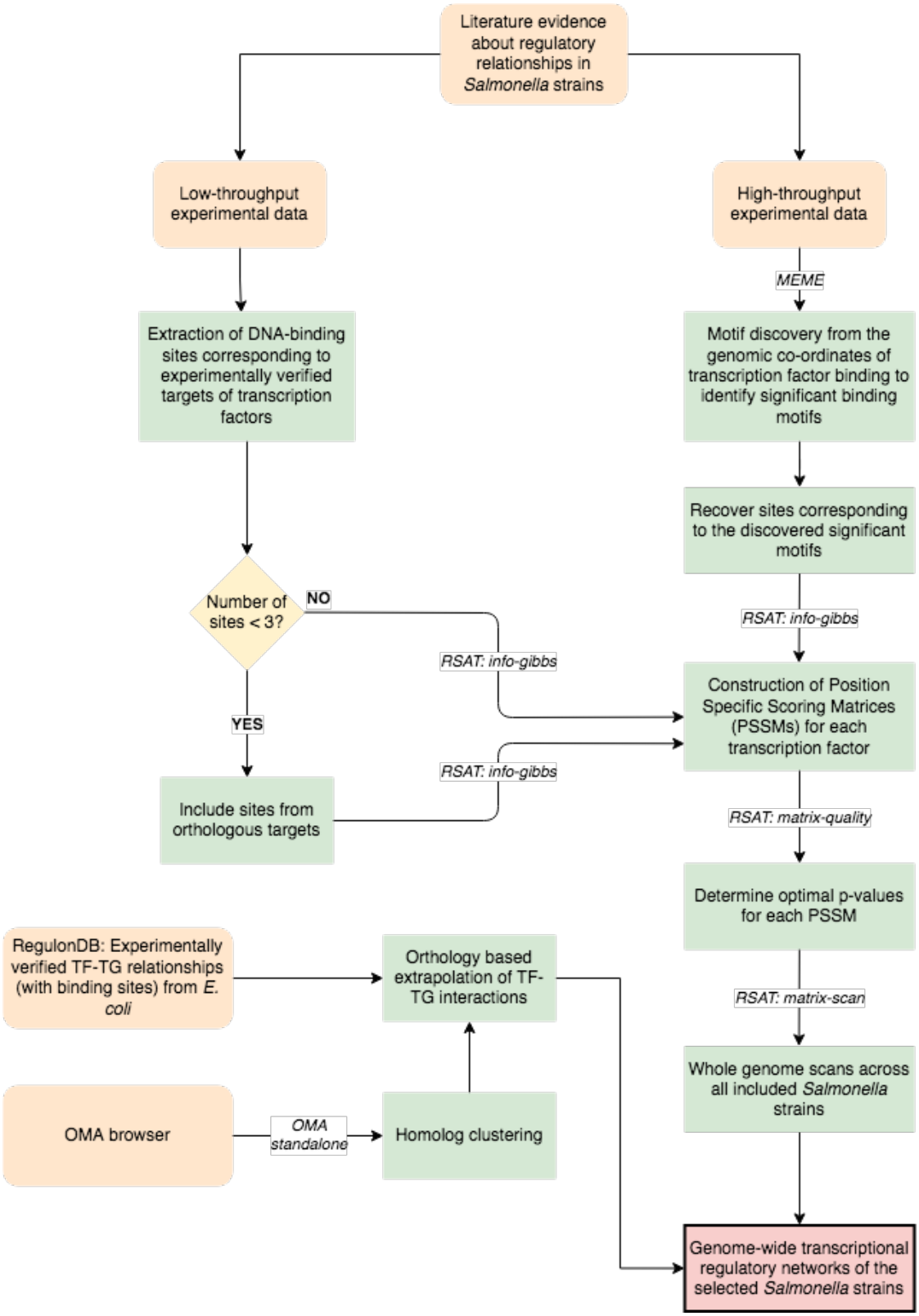

Figure 13. Workflow for the construction of the regulatory layer, updated for the second version of SalmoNet. Orange boxes contain information sources, green boxes contain actions, italic text refers to the necessary software.

The information content of PSSMs used to carry out the genome-wide scans was enhanced with novel binding sites published since the first version of the database, from the available literature and new data uploaded to the CollecTF repository. The sources of TFBS information can be seen in Table 4. RSAT's *consensus* is no longer available on the web server, *info-gibbs* took its place, which is the tool that was used to construct the matrices. Similarly, as previously, RSAT *retrieve-sequence* was used to gather the putative promoter regions for the genomes included in SalmoNet, and *matrix-scan* was used to establish putative transcription factor - target gene (promoter region) pairs.

TF	Strain	Method	PMID
AraC	*Salmonella enterica* serovar Typhimurium 14028s	High-throughput (HT)	24272778
ArcA	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	21144897
Crp	*Salmonella enterica* serovar Typhimurium LT2	Low-throughput (LT)	9068635
CueR	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	20807206
FimY	*Salmonella enterica* serovar Typhimurium LB5010	Low-throughput (LT)	24462182
Fis	*Salmonella enterica* serovar Typhimurium SF530	Low-throughput (LT)	16777370, 17483226
FruR	*Salmonella enterica* serovar Typhimurium LT2	Low-throughput (LT)	8230205
Fur	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	22017966
Fur	*Salmonella enterica* serovar Typhimurium SL1344	Low-throughput (LT)	21573071
GolS	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	20807206
HilA	*Salmonella enterica* serovar Typhimurium SL1344	High-throughput (HT)	17483226
HilC	*Salmonella enterica* serovar Typhimurium 14028s	High-throughput (HT)	27601571
HilD	*Salmonella enterica* serovar Typhimurium 14028s	High-throughput (HT)	27601571
HypT	*Salmonella enterica* serovar Typhimurium 4/74	Low-throughput (LT)	30733296

Ihf	*Salmonella enterica* serovar Typhimurium LT2	Low-throughput (LT)	1511875
InvF	*Salmonella enterica* serovar Typhimurium SL1344, *Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT), High-throughput (HT)	11296219, 27601571
LeuO	*Salmonella enterica* serovar Typhimurium LT2	Low-throughput (LT)	12871947
MetR	*Salmonella enterica* serovar Typhimurium SL1344	Low-throughput (LT)	2676984, 7896708, 1904437
MntR	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	15659669
OmpR	*Salmonella enterica* serovar Typhi Ty2 and Typhimurium	Low-throughput (LT)	23190111
PhoP	*Salmonella enterica* serovar Typhimurium LT2	Low-throughput (LT)	20661307, 15703297
PmrA	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	23690578
RamA	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	18577510
RamR	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	22123696
RcsB	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	20724387
RpoN	*Salmonella enterica* serovar Typhimurium LT2	Low-throughput (LT)	24007446
RstA	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	18790861
RtsA	*Salmonella enterica* serovar Typhimurium 14028s	High-throughput (HT)	27601571
RtsB	*Salmonella enterica* serovar Typhimurium 14028s	High-throughput (HT)	27601571
SlyA	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	11882648, 15208313, 15813739
SoxS	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	19460824
SprB	*Salmonella enterica* serovar Typhimurium 14028s	High-throughput (HT)	27601571
SsrAB	*Salmonella enterica* serovar Typhimurium 14028s, *Salmonella enterica* serovar Typhimurium SL1344	Low-throughput (LT), High-throughput (HT)	15491370, 20300643
YncC	*Salmonella enterica* serovar Typhimurium 14028s	Low-throughput (LT)	20713450

Table 4. List of transcription factors and literature sources with their binding site information.

2.3.1.6. Metabolic layer

The largest changes in terms of data sources occurred to the metabolic layer of the database. As mentioned previously, version one used the STM 1.0 model generated by Thiele et al., and an automatically generated model (BMID000000141143) to generate the enzyme-enzyme interactions. While this was a good starting point, a better resolution is available using novel data.

In two studies Seif et al. have generated genome-scale metabolic models for *Salmonella*, in a second work extending these to describe the metabolism of O-antigens (Seif et al., 2018, 2019). The models used the same STM 1.0 model as a starting point, but updated it with new genes and reactions, and were made strain specific, leading to the metabolic models 410 *Salmonella* strains belonging to 64 serovars.

I used these models as an input to the metabolic networks replacing STM1.0, as the collection contained the GEMs of each of my strains. Due to a technical problem during the development of SalmoNet 1, only enzymes connected by biochemically irreversible reactions were included in the database. By including enzyme-enzyme interactions from reversible reactions as well, I could include more interactions, despite the number of nodes not increasing significantly in the layer.

2.3.1.7. Removal of pseudogenes

To remove all hypothetically disrupted coding DNA sequences (HDCs), the curation made by Nuccio & Bäumler was used to remove such entries (Nuccio

and Bäumler 2014) and HDCs in *S.* Typhimurium D23580 were removed based on previously published analyses (Canals et al., 2019; Robert A Kingsley et al., 2009).

2.3.1.8. Phylogenetic trees, network dendrograms and validation of regulatory layer

Core genome SNPs were determined with *snippy* (version: 4.3.6), with the *snippy-multi* and *snippy-core* functions, ran on the Earlham Institute High Performance Cluster. MegaX was used to build a newick tree file from the resulting core genome SNP alignment. All trees were visualized using the *ggtree R* language package (Kumar, Stecher, Li, Knyaz, & Tamura, 2018; G. Yu, Smith, Zhu, Guan, & Lam, 2016).

The network dendrograms were generated using a Metropolis coupling Markov Chain Monte Carlo (MC^3) from the MrBayes (version: 3.2.4) software with 10 million generations, and 25% of the samples were discarded during the MrBayes run. To accommodate the binary data, the data type was set to restriction, and no substitution model was used (Huelsenbeck & Ronquist, 2001). This is identical to the approach that was used to generate network based dendrograms for the first version of SalmoNet.

For assessing the relevance of regulatory interactions, the overlap of differentially expressed genes for all applicable regulators from the Supplementary Table 3 of (Colgan et al., 2016) were compared with the targets of the same transcription factors in SalmoNet 2. Hypergeometric test was done with the phyper function of R, and the adjustment for multiple testing was carried out via the p.adjust function in R with the Benjamini-Hochberg method.

2.3.1.9. Data formats and website

Users can query data from the Browse menu. The pages of individual proteins show their interactors, the layer the interactions occur, and the orthologous proteins of the selected node. Links lead to the OMAbrowser website, where further phylogenomic analysis can be done.

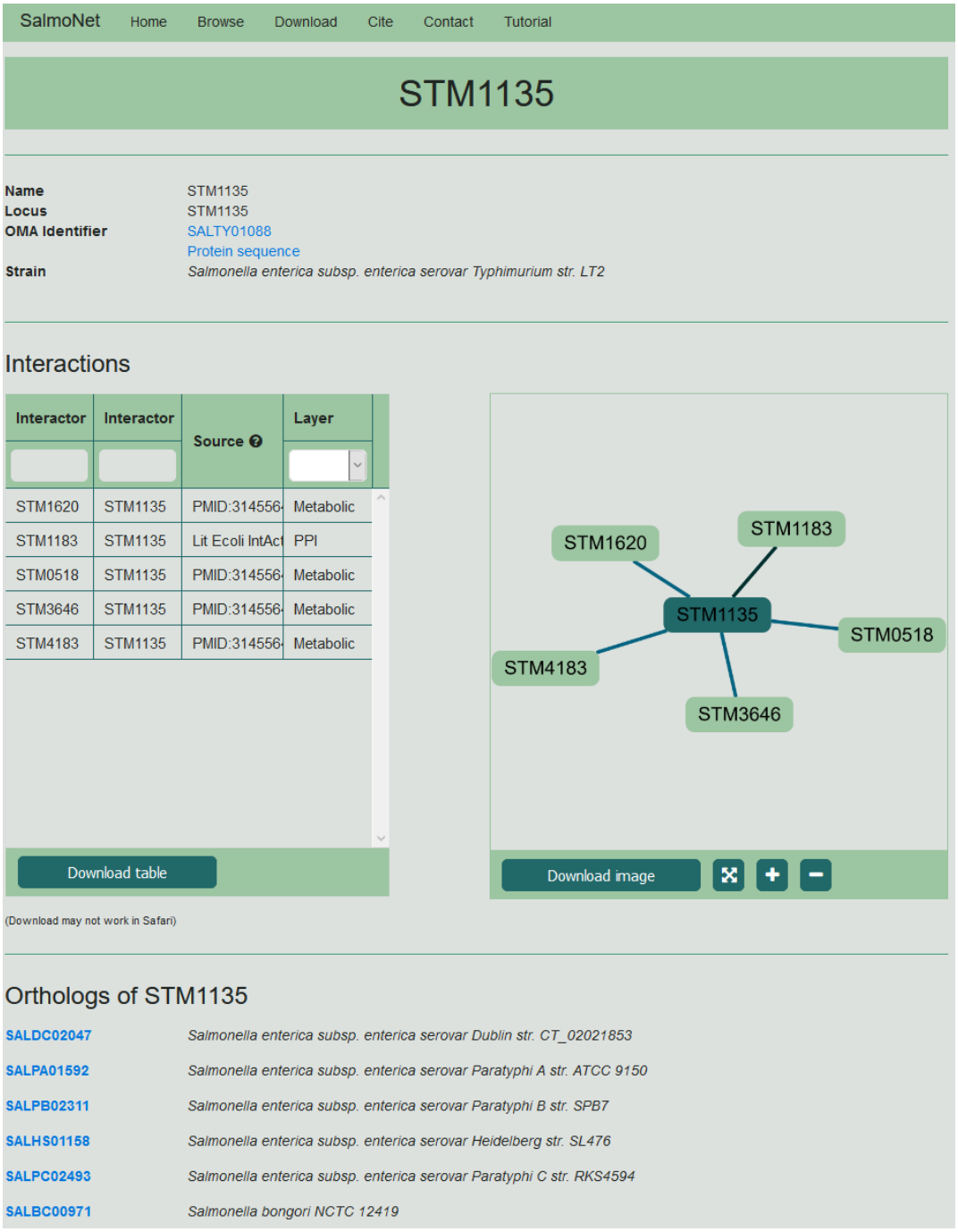

Figure 14. User interface of the SalmoNet website.

2.4. Results

2.4.1. Comparison of SalmoNet 1 and SalmoNet 2

SalmoNet 2 increases the amount of available *Salmonella* strains to 20, including more widely used lab-strains, such as *S.* Typhimurium 14028S or S. Typhimurium 4/74, and more distantly related strains, from other subspecies (*Salmonella enterica* subspecies arizonae (strain ATCC BAA-731 / CDC346-86 / RSK2980)) or other species in the genus (*Salmonella bongori* (strain ATCC 43975 / DSM 13772 / NCTC 12419)). Figure 15 shows the comparison in the size of networks between SalmoNet 1 and SalmoNet 2.

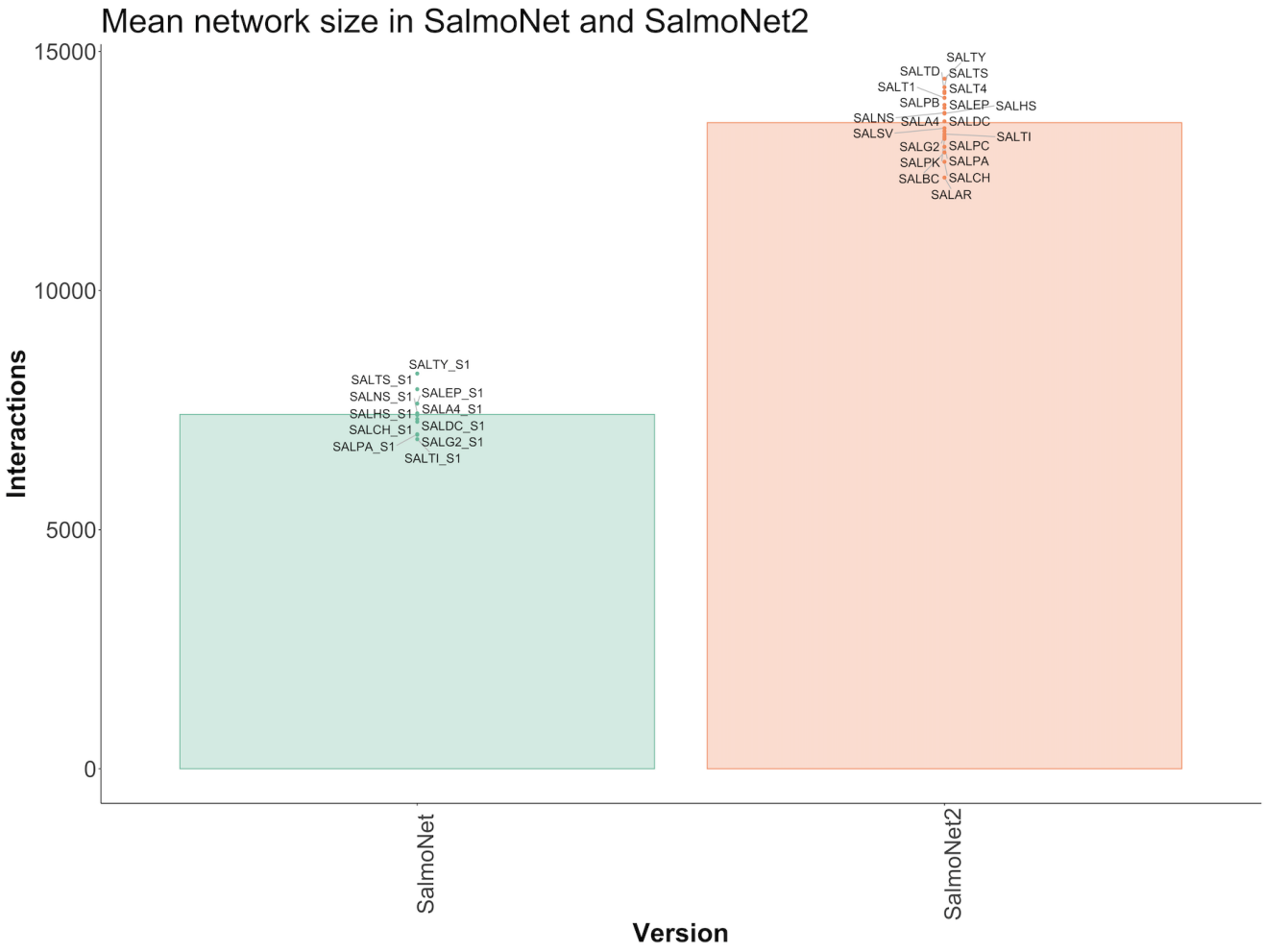

Figure 15. Comparison of SalmoNet 2 network sizes with the first version. The new version increases the information content of the networks, for all included strains.

SalmoNet 2 has increased the information content of all SalmoNet 1 and novel networks, especially that of the protein-protein interaction layer. By using the IntAct MI Score as the quality filter, instead of selecting just one experimental method, I could cover more of the interactome, without losing the quality of interactions. The regulatory layer increased in the number of nodes it connects, and the metabolic layer, at least the mean size of the layers became ever so slightly

smaller, as a result of the more specific metabolic interaction mapping, and the inclusion of strains with smaller metabolic capabilities captured by the metabolic models. Figure 16. shows the size of the layers by interactions, and the number of nodes contained in them.

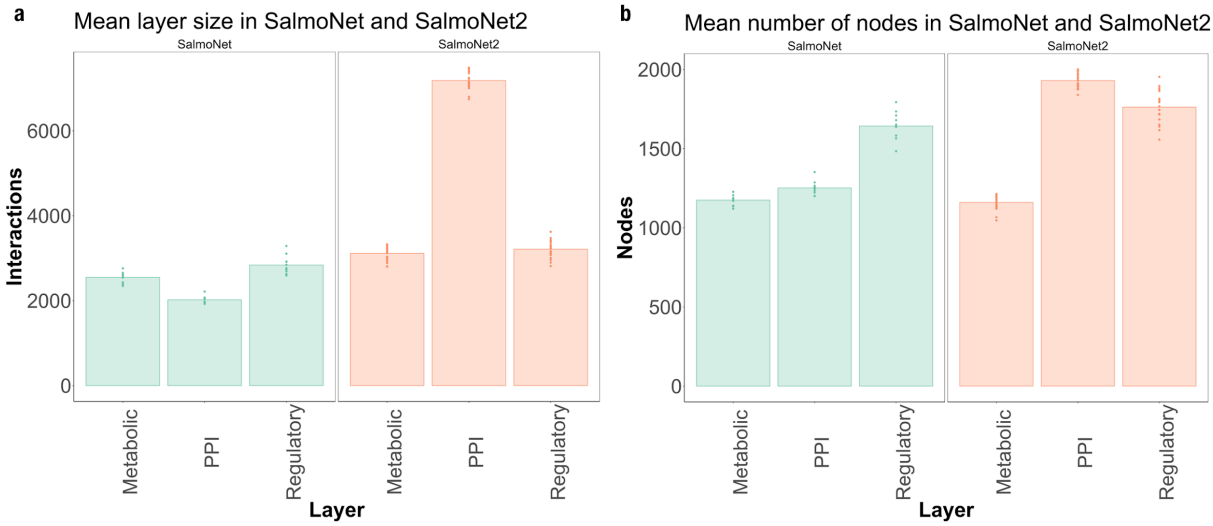

Figure 16. Comparing the number of interactions (a) and nodes (b) between SalmoNet 1 and SalmoNet 2. The inclusion of the new nodes increased the size of the protein-protein interaction layer the most.

The total number of interactions increased from 81,514 to 270215, due to the expansion of the PPI layer, and the increase in the number of strains included. The composition of the consensus network, comprised of shared interactions amongst all strains included in the database slightly changed from the first version of SalmoNet. 24.4% of regulatory interactions (up from 16%), 68.1% of PPI interactions (down from 72%), and 51.8% (down from 69%) of metabolic interactions were shared amongst all strains, forming the core network of *Salmonella* interactions.

I built a core genome SNP based tree to determine the phylogenetic relationships of the included *Salmonella* strains, under the assumption that the strains have an approximately equal rate of mutation. The results are in accordance with previously published phylogenies, such as the one in (Branchu et al. 2018), although the latter publication only concerns subspecies I serovars. There is no clear clustering of the pathovars in the phylogenetic tree. This is consistent with observations in previous works in the literature, the extraintestinal and gastrointestinal strains could not be distinguished based on genomic dendrograms (Timme et al., 2013). This observation is consistent with the hypothesis that extraintestinal, host adapted strains emerge independently from gastrointestinal serovars, through a convergent evolutionary process, accompanied by genome degradation in important functions (Nuccio & Bäumler, 2014; Timme et al., 2013). The structure of the PPI, regulatory and metabolic network dendrograms very closely resembles their phylogenetic relationships, with the notable exception of *S.* Dublin, that does not cluster together with *S.* Enteritidis and *S.* Gallinarum in the network structure-based strains. Figure 17 shows the phylogenetic tree and network structure based dendrograms.

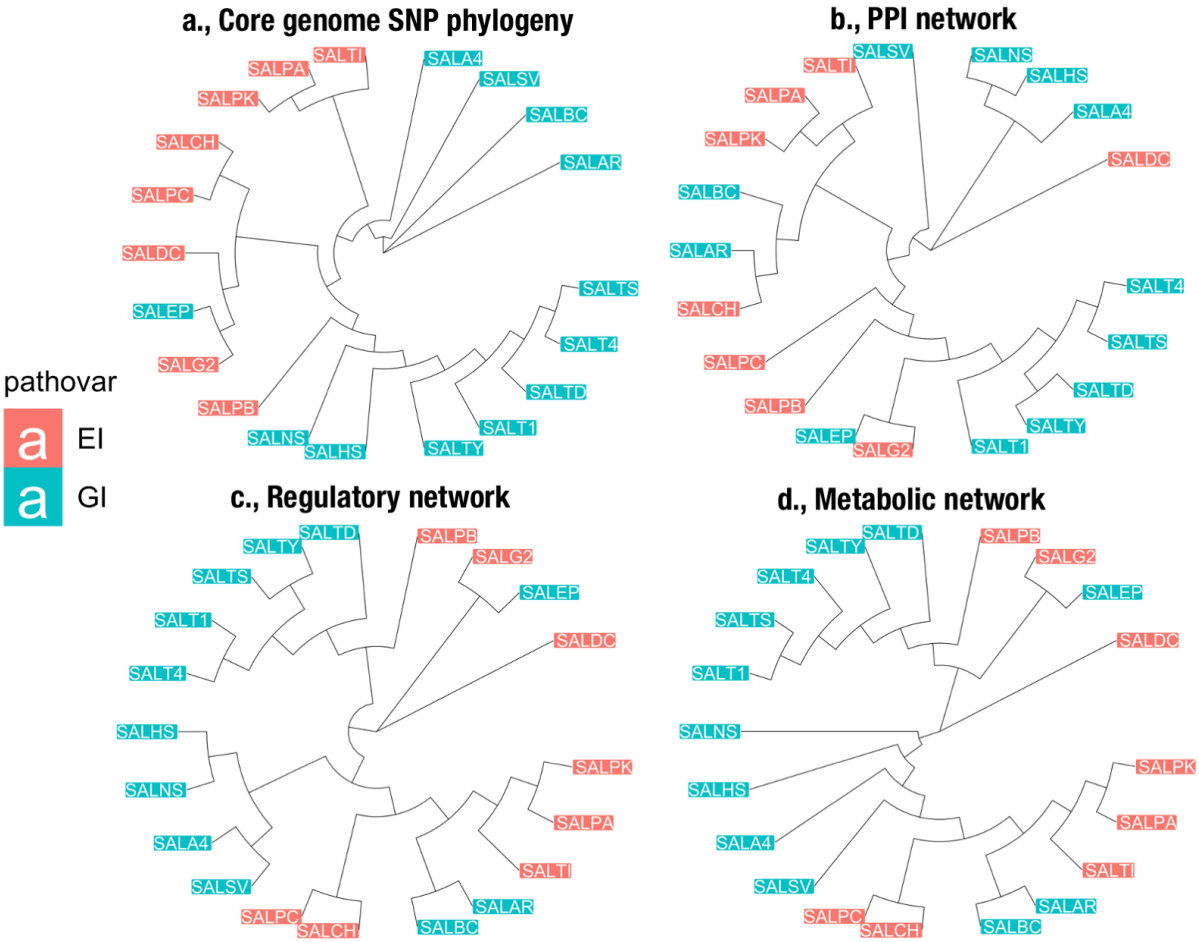

Figure 17 Core genome SNP based phylogenetic tree, and hierarchical classification of network layers. Extraintestinal serovars labelled with red, gastrointestinal serovars with blue labels. A., Neighbour-joining tree from core genome SNPs of the strains. B-D., Hierarchical classification trees based on matrix representation of protein-protein, regulatory and metabolic networks.

To resolve the largest bottleneck in the usability of the database caused by the unfamiliar nature of network resources to most molecular microbiologists, I have written detailed step-by-step tutorials on how to import and analyse network data using Cytoscape, available from the SalmoNet 2 website (https://salmonet.org).

To help computational biologists access network information from SalmoNet 2, we now provide networks in the community standard PSI-MITAB format as well,

which contains a strictly regulated vocabulary for interaction data, helping interoperability between network resources, and is a prerequisite for the resource to be included in the PSICQUIC ecosystem as well. We also plan on sharing the novel interaction resources in The Network Data Exchange (NDEx) repository (Pratt et al., 2015). The latter is an open-source community driven framework, where network information can be stored, shared and queried directly from Cytoscape, solving one of the larger bottlenecks in the accessibility of our data (Pillich, Chen, Rynkov, Welker, & Pratt, 2017; Pratt et al., 2017, 2015).

2.4.2. Assessing the reliability of SalmoNet 2 interactions using experimental information

To ascertain the validity of regulatory interactions included in SalmoNet 2 in an unsupervised analysis, I compared the overlap of TF-TG associations with systematic regulatory knockouts from the *SalComRegulon* database (Colgan et al., 2016). In this work, the authors have generated 18 regulatory knockouts of virulence-related global regulatory systems in *S.* Typhimurium 4/74. Since the *S.* Typhimurium 4/74 strain is one of the newly added strains included in SalmoNet 2, it is an appropriate candidate to test the relevance of the regulatory interactions predicted for it, using an experimental dataset that was not used to create the interaction networks.

Using the lists of differentially expressed genes provided for each knocked out regulator shared by the authors of *SalComRegulon* in the supplementary materials of their paper, I performed a hypergeometric test for every transcription factor, to see if there is a significant association between the

genes differentially expressed following a regulator knockout, and the genes targeted by the same transcription factor in SalmoNet 2. Figure 18 shows the rationale for this comparison.

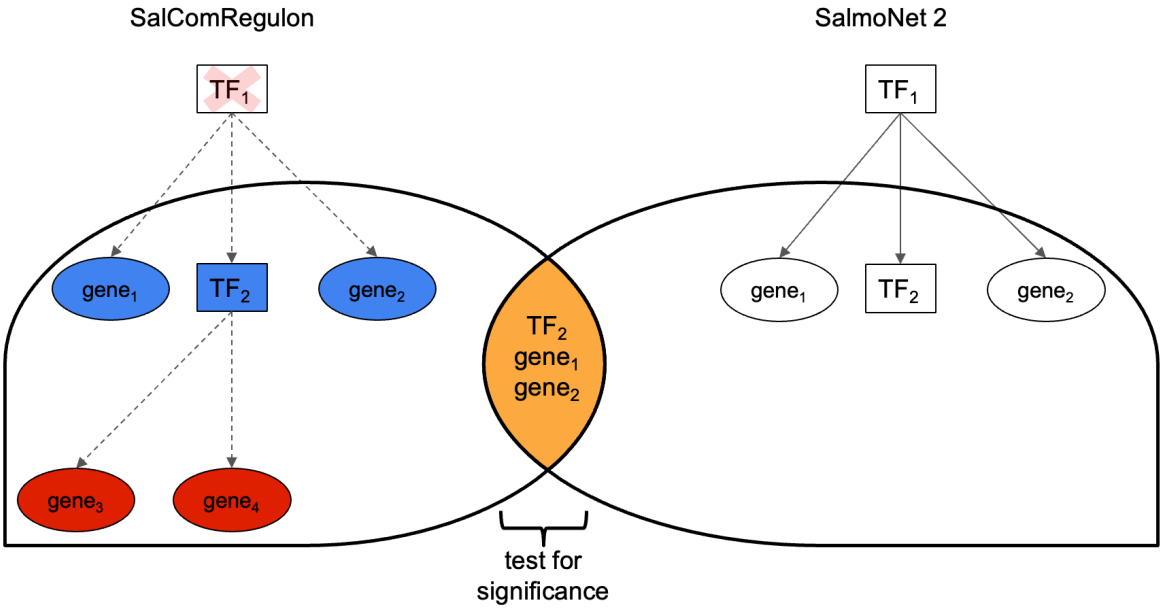

Figure 18. Schematic overlap of differentially expressed genes from regulatory knockouts and predicted SalmoNet 2 regulatory interactions. By comparing the sets of differentially expressed genes (as examples both significantly downregulated in blue and upregulated in red) following knockouts of infection relevant regulators with the predicted targets of each of the kocked out regulators in SalmoNet 2, I was able to measure whether the predicted interactions capture biologically relevant targets.

It is important to keep in mind that regulatory knockouts will not only affect the genes they regulate directly, in the case of multi-step regulatory mechanisms they might affect the expression of genes further downstream, as shown in the example on Figure 18. The implication of this is that the performed test does not test for the precision or coverage of TF-TG binding directly, but

rather tests whether there exists a statistically significant relationship between targets of regulatory interactions implemented in SalmoNet 2 and all downstream affected genes of the *SalComRegulon* regulatory knockouts, under the assumption that a portion of which the tested transcription factors can regulate directly.

Out of the 18 knocked out regulators in SalComRegulon, the majority were included in SalmoNet 2 (the HilD regulon was determined in three different culture conditions). The regulators Dam, HilE, RpoS, RpoE, Hfq were not tested, as binding information from these regulators was not included in SalmoNet 2. Dam and Hfq regulate genes through post-transcriptional control, and as such were outside the scope of SalmoNet 2 (López-Garrido & Casadesús, 2010; Vogel & Luisi, 2011). RpoS and RpoE are alternative sigma-factors, while HilE interacts with HilD to repress *hilA* transcription (Baxter, Fahlen, Wilson, & Jones, 2003; Fang et al., 1992; Humphreys, Stevenson, Bacon, Weinhardt, & Roberts, 1999). The addition of the necessary information regarding these regulators should be considered in the next release of the database.

Following p-value adjustment with the Benjamini-Hochberg method for multiple testing, the following associations were found, listed in table 5.

Regulator	Adjusted p-value
fliZ	1.16×10^{-01}
fur	3.09×10^{-02}
hilA	1.29×10^{-02}
hilC	1.92×10^{-07}
hilD (early stationary phase conditions)	3.89×10^{-03}
hilD (late exponential phase conditions)	3.98×10^{-04}
hilD (SPI-2 inducing conditions)	1.56×10^{-01}
ompR	5.60×10^{-01}
phoBR	6.00×10^{-02}
phoPQ	5.18×10^{-04}

slyA	1.15×10^{-01}
ssrA	2.94×10^{-06}
ssrB	1.26×10^{-06}
ssrAB	1.84×10^{-05}

Table 5. Significance of the overlap between targets of transcription factors as listed in SalmoNet 2, and differentially expressed genes following the knockout of each individual transcription factor. Significance was determined using a hypergeometric test, and the Benjamini-Hochberg method was used to correct for multiple testing.

Using a significance cut-off of adjusted p-value <= 0.05, 9 out of 14 tests have shown a significant relationship between the compared sets. This implies that despite some of the differentially expressed genes measured following a regulatory knockout might not be directly regulated by the knocked-out transcription factor, the regulatory layer in SalmoNet 2 is able to capture biologically relevant interactions using the regulatory prediction pipeline for most of the regulators. Where the significance cutoff was not reached, it was due to a lack or too small of an overlap between target genes and DEGs, indicating regulators where further binding site information should be incorporated in the future to enhance the effectiveness of the regulatory pipeline.

To assess the specificity of the individual transcription factors, I compared the overlaps of their individual target genes in regulatory interaction shared by all involved strains. In the SalmoNet 2 network model, the individual transcription factors regulate a distinct set of genes, with relatively small overlaps. Figure 19 shows the amount of overlapping target genes for the ten highest degree transcription factors, and the enriched biological processes of the shared target genes of the Fis and Crp regulons, containing the most shared

targets across all comparisons. Both Fis and Crp function as global regulators of transcription, involved in energy metabolism, amino acid and nucleotide biosynthesis, nutrient transport, and many other housekeeping functions, and the enriched terms of their shared target genes reflect this (El Mouali et al., 2018; Rosu et al., 2007; H. Wang, Liu, Wang, & Wang, 2013).

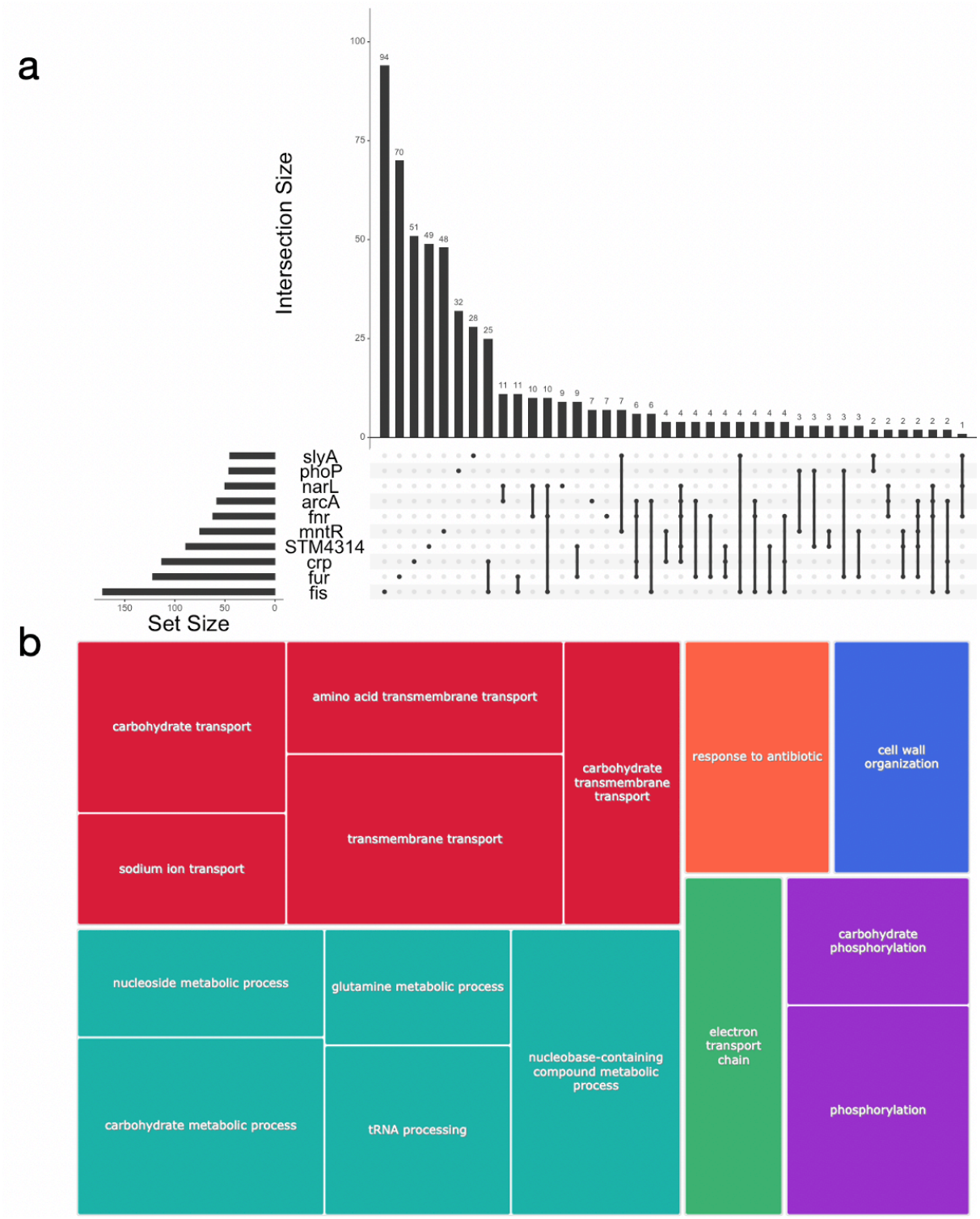

Figure 19: Evaluating the target specificity of transcription factors in SalmoNet 2. A: Overlap of target genes of the ten highest degree transcription factors in regulatory interactions shared by all strains in SalmoNet 2. The vertical bars on the top of the UpSet plot signify the size of the intersection between the

individual transcription factors, the horizontal bars show the size of the sets (i.e. the equivalent of a circle on a Venn-diagram), while the dots connected in the matrix show the specific subset. Transcription factor – target gene relationships are specific in SalmoNet 2, with only a few overlaps in the core regulatory mechanisms. B: Enriched terms in the shared target genes of the Fis and Crp transcription factors. The enriched Gene Ontology biological process terms capture many of the roles Fis and Crp are known to regulate, such as transport of sugars and catabolic functions. Figures generated with UpSetR and Revigo [Supek et al 2011, Conway et al 2017]

2.4.3. Key predictions of SalmoNet 2 in the literature

Since SalmoNet 2 contains predicted regulatory interactions based on genome-wide scans of putative regulatory regions, data published in the literature that was not used in the construction of the database can serve as an independent judge on the quality of some of these predictions. Since the generation of these interactions, a number of studies have been published, where regulatory interactions also predicted in SalmoNet 2 were confirmed experimentally.

For example, in their study, Romero-González and colleagues have studied the effects of regulators downstream from the main SPI-1 regulators such as HilD and HilA. One of these regulators is InvF, a transcription factor usually bound to SicA, a T3SS associated chaperone. The two proteins regulate their cognate genes as a complex. InvF can bind *in vitro* to the promoter region of *sopB*, independently of SicA, although the transcriptional activation of *sopB* still requires bonth InvF and SicA (Romero-González et al., 2020). The latter

interaction, the regulation of *sopB* by InvF has also been predicted by the SalmoNet 2 pipeline, in 18 out of the 20 strains, where the upstream regulatory regions of the gene *sopB* are essentially identical. In the two strains where the interaction has not been predicted, *S.* Typhimurium strain D23580 and *S.* Choleraesuis strain SC-B67 the upstream regulatory region is truncated and the upstream gene *pipD* sits much closer to the *sopB* start site in their genomes. Further work is required to elucidate what caused this rearrangement in these two invasive strains, and whether it affects InvF binding.

In another example Choi and Groisman investigated the effects of the horizontally acquired regulator SsrB, and its effects on the PhoP/PhoQ system and a virulence gene called *ugtL*. They established that SsrB is required to activate the ancestral regulatory PhoP/PhoQ system responsible for the regulation of a large proportion of the *Salmonella* genes, and that SsrB binding to *ugtL* is required for the activation of PhoP/PhoQ. Once again, the SalmoNet 2 pipeline predicted the SsrB – *ugtL* interaction in 18 out of 20 strains. The interaction is missing in *S.* Heidelberg strain SL476, as it is lacking the orthologous protein for UgtL, and it is missing from *Salmonella bongori*, which, as Choi and Groisman have also demonstrated, lacks the gene *ssrB* and the SsrB binding site in the *ugtL* promoter. This fascinating study also found, that the regulator SsrB promotes the transcription of *phoP* by binding to the *coding region* of the upstream gene (*purB*) of the *phoP* promoter (Choi & Groisman, 2020). As this was outside of the scope of the genome wide scans used to establish putative regulatory interactions in SalmoNet 2, the interaction is not present in the database.

In another example Lim et al. have been studying the *iroBCDEN* operon, and verified that a binding site of the transcription factor Fur lies within

the promoter region of the *iroBCDE* operon (Lim et al., 2020). The interaction is once again present in all but two SalmoNet 2 strains. The lack of the interactions can be explained by the fact that the orthologous proteins are missing from *Salmonella bongori*, and in the case of *S.* Typhi, a small gene of unknown function, STY2889 is present in the regulatory region of *iroBCDE*. However, SalmoNet 2 does predict a Fur binding site in the upstream regulatory region of the introgressing STY2889. More work is required to determine whether this small gene could be coregulated with *iroBCDE*.

2.5. Discussion

Multi-layered network databases collate information from various sources, and are useful knowledgebases of interaction information. With SalmoNet 2 including additional important human pathogenic *Salmonella* strains, both typhoidal and non-typhoidal, more targeted analysis is now possible focusing on human disease. Since most of the included extraintestinal serovars have adapted to different host species, eliminating the differences from the acclimation of these pathogens to their specific microenvironments could help specialists target the human-disease specific interactions and subgraphs.

By greatly increasing the number of available strains compared to SalmoNet 1, SalmoNet 2 now extends beyond subspecies I., and includes information on members of another subspecies (subspecies *arizonae*), or an entirely different species (*Salmonella bongori*). The larger evolutionary distance between this additional subspecies and species can further help *Salmonella*

researchers study the evolutionary history of the genus as as whole, and contrast the differences to the more studied human pathogenic strains (Fookes et al., 2011; Park & Andam, 2020).

The comparison of differentially expressed gens from *SalComRegulon* and transcription factor target genes from SalmoNet 2 highlights that the majority of the tested transcription factors capture biologically relevant target genes, and where they do not, clearly points out areas if improvement, where additional data should be included in the next iteration of the SalmoNet database for the affected regulators (Colgan et al., 2016). Additionally, I found experimental evidence for multiple regulatory interactions from the literature that have been published since the generation of the interaction data, such as the examples from the papers describing the interactions of InvF with *sopB* or the regulation of *ugtL* via SsrB (Choi & Groisman, 2020; Romero-González et al., 2020).

Developing a more tight-knit structure between SalmoNet and other available large-scale evolutionary genomics tools such as OMA, there is increased potential to generate interaction networks for specific Salmonella strains on request, or build similar data resources for other non-model organisms, similarly as to how it is described in the work above. With the change to OMA as the backbone of SalmoNet interactions, there is also a great untapped potential to study the evolutionary history of proteins, and potentially even interactions. Although it was largely outside the scope of this thesis, I did end up using OMA in a research article we published, where we mapped the interaction differences of two paralogous proteins affecting autophagy, and the OMA database provided the missing information, pinpointing the specific genome duplication event giving

rise to the studied proteins (Demeter et al., 2020). While this specific example is not directly applicable to prokaryotes, the on-demand availability of orthologous proteins from outside of our studied organism or clade could make larger scope comparisons possible. The programmatic access interfaces implemented into OMA make these integrated analyses reproducible, and scalable as well (Kaleb, Warwick Vesztrocy, Altenhoff, & Dessimoz, 2019).

The availability of strain specific metabolic models, and the increased specificity of PPI data, although still reliant on orthology mapping, increases the resolution of the resulting network models, and the more interwoven interaction layers get, the more valuable the information content of the database gets. Although there are other resources containing Salmonella interaction data, such as STRING for PPI interactions, RegPrecise for regulatory interactions, or BioCyc for metabolic interactions, no other resource combines the listed connection types besides SalmoNet (Caspi et al., 2019; Novichkov et al., 2013; Szklarczyk et al., 2019).

To increase the usability and interoperability of the generated interaction information, I have generated the data files in the PSI-MITAB format as well, quickly becoming a standard of biological network information (Samuel Kerrien et al., 2007; Perfetto et al., 2019). Beyond their raw information content, databases are as good as their usability and their availability, and the potential for SalmoNet data to be found and utilised in as many ways as possible is crucial for this effort to be useful for the scientific community (Merali & Giles, 2005).

2.5.1. Future research directions

The integration of SalmoNet into the OMA ecosystem makes the possibility of genome-to-network pipelines feasible, and meaningful to generate in the future. In the past, an addition of a novel strain would have meant the re-computation of the all-against-all orthology mapping, which took weeks on the Earlham Institute High Performance Cluster. The OMA standalone software cuts down on this large computational bottleneck, by only requiring us to compute the relationships of the novel genome. Although only those genomes were included in this update that were already carried by OMA, the potential to generate interaction networks on request, or to map *Salmonella* breakouts, not only through genomics, but comparative network studies, could be a useful tool in the future for *Salmonella* studies. While I have made every necessary precaution to remove false-positive interactions from the resources, another argument for a further increase in the amount of networks - to the scale of hundreds or thousands - would be, that it would possibly give more statistical, and lineage based backing to interactions.

In addition, the possibility of generating strain specific networks to characterize the samples of a specific outbreak or epidemic strain could give us further insights into the adaptation of *Salmonella* to specific environments and stressors. In the future, statistics based methods could be involved when trying to categorise *Salmonella* serovars into gastrointestinal or invasive phenotypes, such as the machine learning assisted DeltaBS applied by Wheeler et al., that was able to categorise the recently emerged iNTS strains as invasive purely based on sequence data, or the approach used by Langridge *et al*, that assigns invasiveness

to the individual serovars, by what percentage of the samples were isolated from the blood, compared to the total amount of samples (Langridge, Nair, & Wain, 2009; Wheeler, Barquist, Kingsley, & Gardner, 2016).

To increase the reach of the resource as much as possible, the goal of future minor and major updates should be to integrate the resource in the PSICQUIC web service. The implementation of the PSI-MITAB interaction format already serves this goal. The advantage of this integration would be, that potential users do not have to query and process SalmoNet data from the website directly, but it would be accessible directly from the PSICQUIC service as well, in combination with any other compliant data, similarly to NDEx, increasing accessibility (del-Toro et al., 2013; Perfetto et al., 2019).

As mentioned above, the potential value and information content of each interaction increases by their potential interconnections between layer types. As such, in the future involving other information as additional layers could be an important step to increase the specificity and usability of the interaction resource. One such layer is the addition of protein complex information, such as the one found in the Complex Portal (Meldal et al., 2015). Interactions between proteins often occur in complexes, and the potential to include this information could lead to novel modelling approaches, and insights into the studied systems as a whole.

To understand gene regulation in more detail, post-transcriptional regulatory interactions could be included. The presence of small RNAs in *Salmonella* has been described previously, and many contemporary interaction databases carry and model with this kind of data - albeit not yet in interaction databases involving

prokaryotes (Kröger et al., 2012) (Türei et al., 2016). Including this interaction layer is one of the most important tasks for the future releases of SalmoNet. Recently, there have been novel results of post-transcriptional modifications of proteins as well, which could be a fruitful avenue in the future (Macek et al., 2019).

3. Network biology methods to study evolution and adaptation

3.1. Network resources

The underlying assumption of network analysis is that by putting relationships to the individual interactors, we notice emergent patterns that might better explain their behaviour than studying them in a vacuum or different context. Hubs, for example, are such properties – without putting the interaction data to genes or proteins, we would not know of their promiscuous nature, and potentially heightened biological relevance. This is the case on any level we aim to analyse networks, be they molecular or supra-individual networks (Miele, Matias, Robin, & Dray, 2019). The availability of networks depends on the subject, but in molecular biology, there are more and more repositories at our disposal where we can query interactions from. These databases collect and curate interaction data, often from individual research articles or from other similar data resources.

Most things that can be represented as matrices can be represented as networks as well - which is why the most important step in network analysis is determining whether networks are an appropriate tool for the study of the particular question at hand, and one would not be more successful applying clustering methods, or principal component analysis for example. While networks can be, for the most part, intuitive, there are elements in their behaviour that are the opposite of that – all interactions between two nodes are not only considered in the context of this pair of interactors, but also the context of the global network as a whole (Miele et al. 2019; Barabási et al. 2011; Barabási and Oltvai 2004). This can become even more complex when one starts comparing networks, to study evolutionary processes, and has to assess whether an interaction is important on a local (i.e. between a pair of interactors), global (i.e. on the scale of the entire network) and an evolutionary (i.e. taking all compared networks into account) scale.

This chapter describes the theoretical background of network comparison methods used in this thesis, and their applications. The first subject is a large-scale study on the regulatory evolution of cichlid fish species, where I first developed and applied my network rewiring approaches. Although the subject of the analysis was different from what is the main topic of my PhD research, the approaches developed for this work laid the foundation of my *Salmonella* studies. The second half of the results section describes how network rewiring can be used to study SalmoNet 2 networks of typhoidal and gastrointestinal *Salmonella* strains for hypothesis generation.

3.2. Aims

The aims of this project were the following:

- Identify approaches in the literature applicable to the analysis of molecular interaction networks.
- Application of one of these approaches (DyNet rewiring) on a study involving the comparison of gene regulatory networks of East African cichlid fish species.
- Highlight how network rewiring can be applied to interaction networks of typhoidal and non-typhoidal *Salmonella* strains, and what downstream analyses can be applied to help understanding the results of rewiring.

3.3. Network comparisons

3.3.1. Network rewiring

The increasing availability of incredible amounts of biological data enables us to build interaction networks such as the ones detailed in Chapter 2 of this thesis. Using these resources, one can formulate questions that study how the parts of these systems work together or differ in areas. These interaction networks are difficult to compare naively due to their complexity. Thankfully, there are a group of methods aimed at solving this problem (Han and Goetz 2019). The methodology I used most extensively in my PhD research belongs to a group of approaches often categorised as "network rewiring" methods. Network rewiring is a broad term used to describe many approaches aimed at quantifying changes between interaction networks.

The specific tool I used most often is a third-party module for Cytoscape called "DyNet" (Goenawan et al., 2016). The reason I chose this was its ease of use thanks to its integration into the Cytoscape ecosystem, and since it was an appropriate tool for the interaction networks I analysed.

3.3.1.1. The Dynet tool

DyNet identifies the most dynamically changing, or most rewired neighbourhoods between the compared networks (Goenawan et al., 2016; Salamon, Goenawan, & Lynn, 2018). The tool does this, by assigning a score, the D_n rewiring score to each node, that effectively sums up the quantitative (i.e. how

many interactors does the node have) and qualitative (i.e. what nodes is it interacting with) differences between the same nodes across different networks, so even a node that has the same amount of interactors in two networks, but some of those interactors are different proteins entirely, will be assigned a rewiring score.

The rewiring score is modelled as a weighted node adjacency matrix, where instead of indicating the presence or absence of an edge with binary 0-1 values, one can supply any number, used to represent the weight or importance of that edge, through some predefined process, like categorising them based on interaction detection methods. In this model this weighted matrix is extended by a third dimension, *S*. This describes the state-space, where states represent the compared interaction networks, or in other words, every state has its own weighted adjacency matrix. The rewiring score is calculated by first calculating the mean of non-zero edge weights of all states (if using weighted data) and following that standardizing the data through dividing by the previously calculated means. Following that the centroids for each node over all states will be calculated, by taking the average of the sum of standardized weights. The Euclidean distance from the centroid is calculated for each node by taking the value for the standardizes weights minus centroids, for each interaction (i.e. the row in the adjacency matrix), and calculating the square root of the sum of their squares. The final rewiring value of a node is calculated by dividing the sum of distances with n-1, where n is the number of compared states.

Using these terms, the formula for the Dn-score is the following:

$$D_n - score = \frac{\sum_{i=1}^{n} \quad [distance(V_i, centroid)]^2}{n-1}$$

Where V_i is each vector representing a node in each network (the compared nodes), and n is the number of networks (states) being analysed. If using edge weights, normalization is performed before, by dividing by the average (of non-zero) values across all networks (Salamon et al. 2018).

Figure 20 shows a worked example included from the supplementary materials of (Goenawan et al., 2016).

A worked example of DyNet's dynamic node rewiring score(Dn) *for a single node (Node A) within a five node (A,B,C,D,E) undirected dynamic network over five network states. The dynamic network is represented as five weighted adjacency matrices, one per state. The absence (zero value) of edges and nodes across states is also represented. Entries are first standardized by the mean across states (blue matrix) and then the variance in the distance of the node compared to its centroid (green matrix) is calculated.*

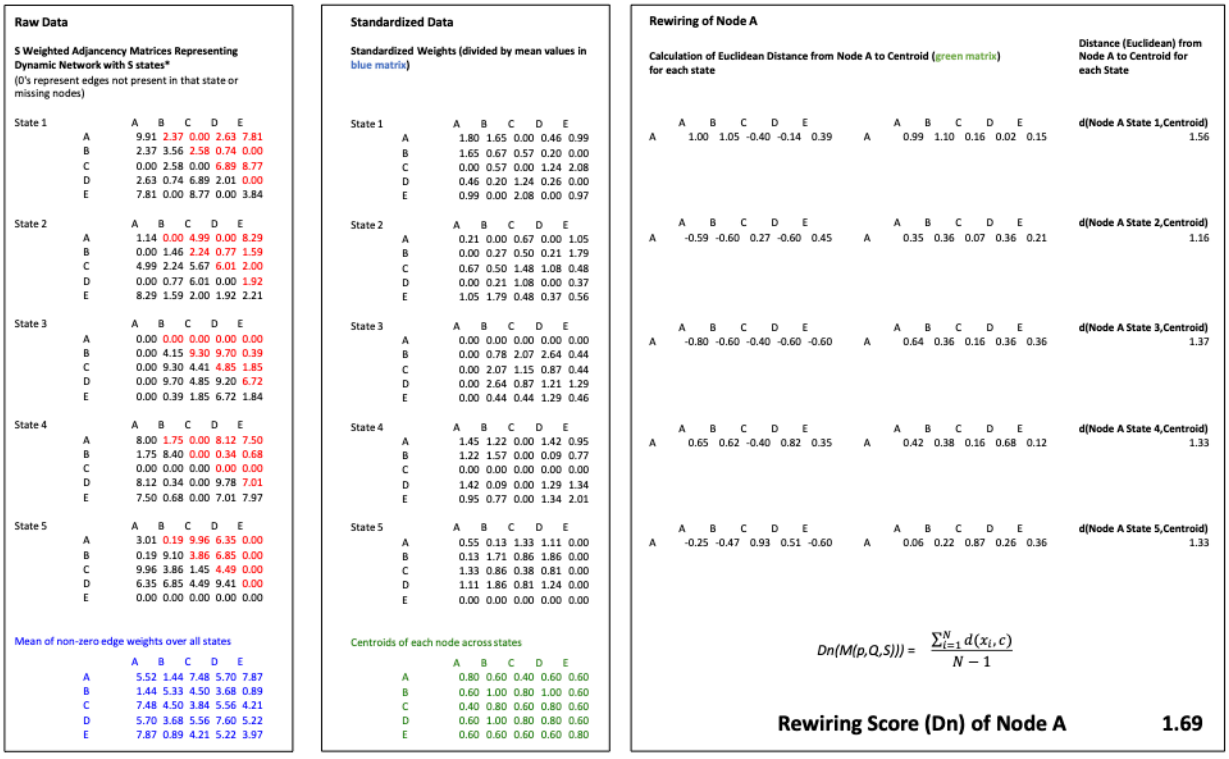

Figure 20. Worked example demonstrating the steps to calculate the rewiring value using DyNet from (Goenawan et al., 2016). With permission of the rights holder, Oxford University Press.

3.4. Evolution of regulatory networks associated with traits under selection in East African cichlid species

In this work, my colleagues and I were studying how gene regulatory changes can lead to changes in anatomy and phenotype of East African cichlid fish species. The resulting findings highlighted how network rewiring approaches can be used to study regulatory evolution in non-model organisms, one that I applied in my studies related to *Salmonella* (while not a vertebrate, similarly a non-model organism). My role in the work was to carry the out network rewiring analysis, and help in the interpretation of its results.

Below is the summary of the background, approaches and main results of the work, necessary to understand the importance and value of the network rewiring analysis I applied. This half of the chapter is based on the peer-reviewed article published in *Genome Biology,* which I am a co-author of (T. K. Mehta et al., 2021). The detailed description of the study can be found in the paper. While I cannot claim authorship over all parts of the project, as this has been an over five year long endeavour altogether, it was important to include in this thesis, as the parts I worked on fundamentally shaped my PhD research. Figure 21 shows the main steps and approaches used in this project. My role in this project was to measure and interpret regulatory network rewiring using DyNet in the generated networks.

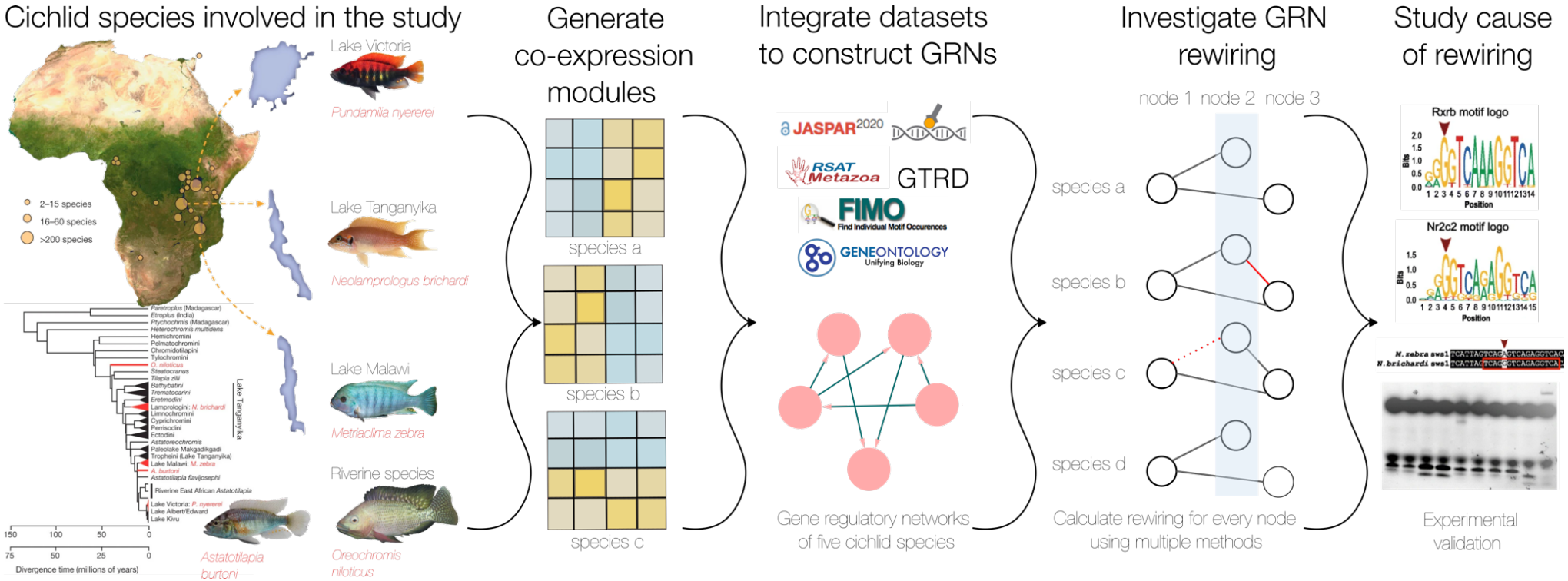

Figure 21. Graphical abstract of the work on the evolution of regulatory networks associated with traits under selection in East African cichlid species. The RNA-Seq data of five cichlid fish species was used to generate modules of co-expressing genes using the Arboretum software. Additional data was integrated into the resulting expression modules, including ChIP-Seq, and Gene Ontology data. Multiple approaches (RSAT, FIMO) were used to scan the UTR regions of coding genes for putative transcription factor binding sites, using binding signature data from multiple databases. The resulting gene regulatory networks were analysed with three distinct approaches aimed at quantifying network rewiring. Candidate targets containing rewired, modified transcription factor binding sites were tested experimentally. Image modified, from (Brawand et al. 2014)*, licensed under CC BY-NC-SA 3.0.*

3.4.1. Background

Among all vertebrate species, ray finned fishes are among the largest of any group, with East African cichlid species displaying one of the most striking examples of adaptive radiation (Brawand et al., 2014). Over 1500 species exist today in the Great Lakes of East Africa (Lake Tanganyika, Lake Victoria, and Lake Malawi), which evolved in a relatively short amount of time. 250-500 species formed in each lake, taking between 15,000-100,000 years in the case of Lake Victoria, less than 5 million years for Lake Malawi, and approximately 10-12 million years for Lake Tanganyika, from just a few ancestral lineages of cichlid fish species.

These novel species inhabit a wide range of ecotypes, exhibit a range of varying behaviours and morphology. Sexual selection, indicated by colourful phenotypes and elaborate bower building, and their ecological roles in terms of foraging behaviour both converge on the cichlid visual systems, that have trichromatic vision, with eight opsin genes. Altogether, the evolution of these species has been shaped by cycles of population expansions, and shrinkage, as their environment changed over the time (Brawand et al., 2014; T. K. Mehta et al., 2021). On an evolutionary timescale, the fastest rewiring interaction layer is the gene regulatory network (GRN) one (Shou et al., 2011). Mutations accumulating in the *cis*-regulatory elements of genes (transcription factor binding sites of promoters and enhancers), or *trans* regulatory changes leading to the levels of a regulator can lead to phenotypic differences, stemming from GRN rewiring events.

As such, the study aimed at researching if evolutionary regulatory changes on the level of whole gene regulatory networks can lead to phenotypic variation and

facilitate adaptation to a variety of ecological niches found in the great lakes of East Africa, and the surrounding rivers.

3.4.2. Methods

3.4.2.1. Co-expression modules

The Arboretum software generates gene regulatory modules for multiple species using expression data, and gene phylogenies. Modules of co-expressed genes were identified using the software in five cichlid species (*Pundamilia nyererei* (Pn), *Maylandia zebra* (Mz), *Astatotilapia burtoni* (Ab), *Neloamprologus brichardi* (Nb) and *Oreochromis niloticus* (On)) (Roy et al., 2013). Co-expression modules were generated for six tissues (brain, eye, heart, kidney, muscle, testis), from RNA isolated from adult animals. As a result, 18,799 orthogroups including 69,989 genes and 34,220 1-to-1 orthologs were identified.

3.4.2.2. Gene regulatory networks

To establish putative transcription factor – target gene interactions, transcription factor binding site information was retrieved from the JASPAR database, and other similar resources such as HOCOMOCO or UniPROBE (Hume, Barrera, Gisselbrecht, & Bulyk, 2015; Khan et al., 2018; Kulakovskiy et al., 2013). ChIP-seq peaks were called from experiments of human and mouse transcription factors, retrieved from GTRD (Yevshin, Sharipov, Valeev, Kel, & Kolpakov, 2017). Similarly, as in the case of SalmoNet, position specific scoring matrices (PSSMs) were generated using the *info-gibbs* module from the RSAT suite (see 3.3.2.3 for details), and scanned 20kb upstream of the starting sites of genes and conserved non-coding elements using RSAT's *matrix-scan* and FIMO

from the MEME suite of tools (Bailey et al., 2009; Nguyen et al., 2018). The optimal p-value cutoff for every putative TF-PSSM pair was calculated using RSAT's *matrix-quality* or a default value was used in cases where this could not be determined.

3.4.2.3. Network rewiring

To study transcription factor – target gene rewiring between the five species, three approaches were developed and used for this study. All three aim at identifying differently interconnected parts of the compared networks. My role was the implementation and interpretation of the DyNet approach.

The first method compares TF-TG edges to a selected species versus others in the context of gene expression module assignment (e.g. module changing transcription factors). In this metric a rewired interaction is present in when a unique transcription factor - target gene edge is present in only one "focal" species, but the transcription factor ortholog is state changed in module assignment, and is present as a node in other TF-TG edges in any of the other species.

The second approach collects TF rate of edge gain and loss in networks. This method uses a continuous-time Markov process parametrized by transcription factor - target gene gain and loss rates, and uses an expectation-maximization based algorithm to estimate gain and loss rates. Regulators that have a degree > 25 were used, as less than 25 edges would greatly hinder statistical analysis (T. K. Mehta et al., 2021).

Finally, the third approach utilised DyNet rewiring scores. In this approach, I used the DyNet (version: 1.0) package implemented in Cytoscape (version: 3.7.1.). I calculated and visualized the degree-corrected rewiring (D_n) score of orthologous nodes across the five species. Following this, the D_n score of each orthogroups rewiring score was ordered, and the mean calculated. To measure the significance of each orthogroups rewiring score against all others, the non-parametric Kolmogorov-Smirnov test (KS-test) was applied.

3.4.2.4. Summary of molecular biology approaches used in this study

The DNA-binding domains of two cichlid proteins, NR2C2 and RXRB were predicted using multiple sequence alignment and conversation with their mouse and human orthologues. *M. zebra* and *N. brichardi* specimens were sacrificed using triacine at the University of Hull, UK and at the University of Basel, Switzerland. RNA was extracted and first strand cDNA synthesis of the DNA-binding domain specific regions was done. The expression of the DNA-binding domain was resolved by SDS-PAGE. The EMSA assay was carried out using double-stranded DNA probes with in vitro expressed DNA-binding domains as described above. The double-stranded DNA probes were generated through annealing the sense and antisense oligonucleotides in an annealing buffer. Further detail of the experiments can be found in (T. K. Mehta et al., 2021).

3.4.3. Results

In this study five East-African cichlid fish species were used: *Pundamilia nyererei, Maylandia zebra, Astatotilapia burtoni, Neloamprologus brichardi* and *Oreochromis niloticus,* whose gene regulatory networks were established as described in the *Methods* section of this chapter. Using the Arboretum software, 10 modules of 12,051-14,735 coexpressed genes were determined, that were represented in 18,799 orthogroups.

During the analysis of the gene regulatory networks of the aforementioned species, using the Dynet degree corrected rewiring scores, we identified 60 candidate genes linked with phenotypic diversity based on previously published literature. These genes have a few standard deviations higher degree-corrected rewiring scores than the mean of all orthologs (0.23 ± 0.007 SD; KS-test p-value 6×10^{-4}). Figure 22 shows the violin plots detailing the distribution of degree-corrected D_n rewiring scores.

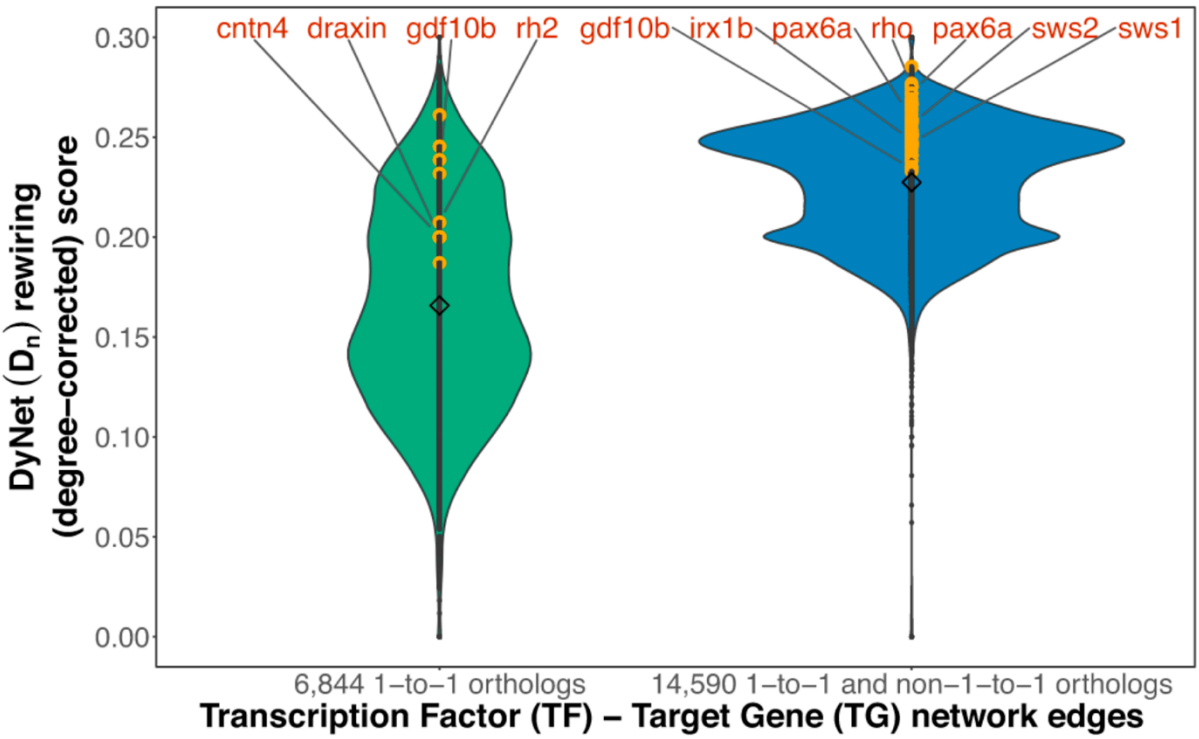

Figure 22. Distribution of DyNet degree-corrected rewiring scores between 1-to-1 (green) and 1-to-1 and many-to-many (blue) orthologs. Mean rewiring shown as white diamond in the center. Non-candidate genes shown with black dots through the center. Candidate genes linked with morphogenetic trait diversity have a few standard deviations higher score, highlighted in orange. Image modified from (Mehta et al. 2021)*, licensed under CC BY 4.0.*

These highlighted phenotypic diversity genes are involved multiple important functions, such as craniofacial development (*dlx1a, nkx2-5*), tooth morphogenesis (*notch1*) genes, telencephalon diversity (*foxg1*) and interestingly, most visual opsin genes and genes associated with photoreceptor cell differentiation and eye development (*rho, sws1, sws2, actr1b, pax6a*).

As a case study, we focused more in detail on the highly rewired nodes of the visual system highlighted above. The changes in regulation can lead to large shifts in the adaptive spectral sensitivity of adult cichlids, and as such we hypothesized

that the diversity in opsin expression could be the result of adaptive gene regulatory network evolution (Carleton, 2009).

Sws1 is an ultraviolet opsin, responsible for the short-wavelength section of the visual palette in *N. brichardi* and *M. zebra*, two of the lake representative species. The two species share many regulators for this gene, but there are multiple unique transcription factors associated with only one of them. Overall, the analysis identified more unique significant TF regulators of *sws1* in *M. zebra,* than in *N. brichardi* (38 vs 6). Interestingly, the rewiring analysis also highlighted that one of the causes of the rewiring is that *M. zebra* has a potentially broken interaction caused by a mutation in the binding sites of the NR2C2 and RXRB transcription factors, an interaction that is present in *N. brichardi*. Figure 23 shows the comparison of regulatory networks in the two species, the single-nucleotide polymorphism (SNP) responsible for the loss of interaction.

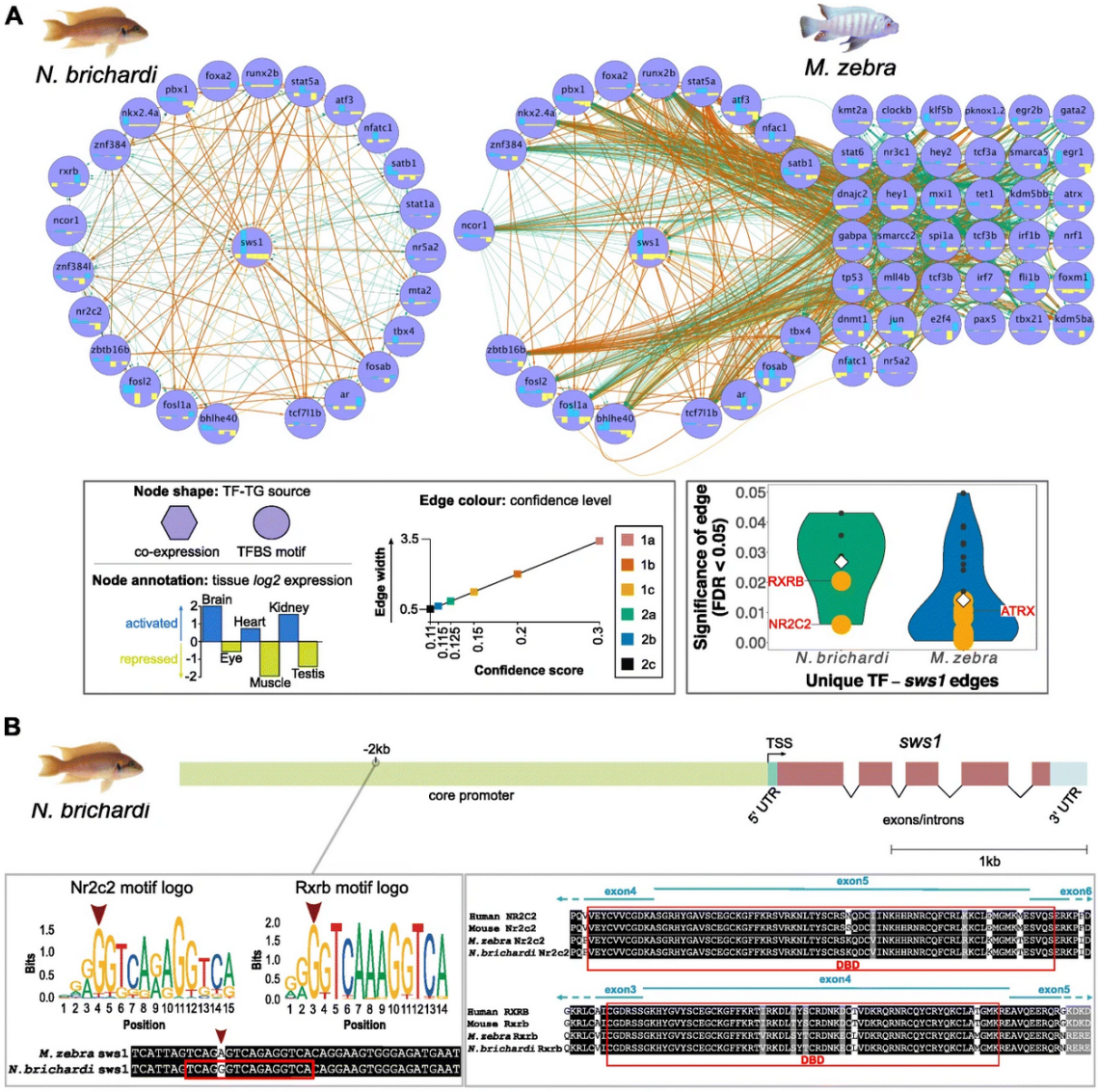

Figure 23. A: Regulatory networks of sws1 in N. brichardi and M. zebra. Sws1 sits in the middle, nodes organized in a circular layout are shared regulators, grid layout nodes are unique regulators in M. zebra. For node annotation please see legend on the bottom left. Bottom right shows a violin plot of edge significance, with highly significant edges highlighted in orange. B: A SNP in the promoter region of the sws1 gene leads to regulatory rewiring between two species. Top: the SNP found at approximately -2kb from the transcriptional start site of sws1. Bottom left: binding motif logos for the two TFs predicted to bind to the region in N. brichardi, the species with intact interactions. A G→ A mutation potentially disrupts TF binding. Bottom right: protein alignment of the DNA binding domain from the two predicted interacting transcription factors in the two cichlid species, and their corresponding orthologues in Homo sapiens and Mus musculus. Image modified from (Mehta et al. 2021)*, licensed under CC BY 4.0.*

My colleague's experimental validation using an EMSA assay confirmed that NR2C2, but not RXRB can bind to the *sws1* promoter in *N. brichardi,* and that the variant in *M. zebra* has disrupted binding, and potentially regulation of *sws1* in the latter species. Results are shown on Figure 24. These results are further supported by their better correlation with the expression values of the regulators, meaning NR2C2 is better associated with *sws1* than RXRB, especially in the eye tissue.

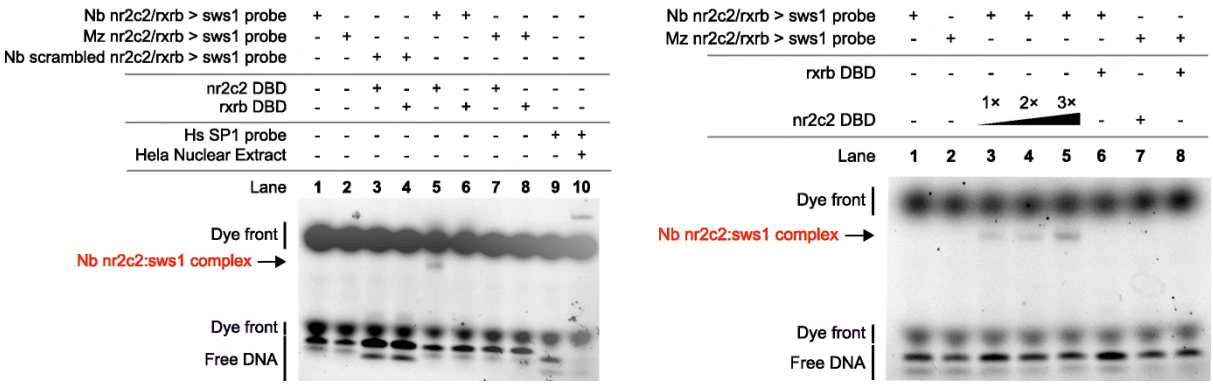

Figure 24. EMSA assay to screen for DNA binding from the NR2C2 and RXRB transcription factors. Left: Table on top contains the combinations of DNA probe and expressed DNA binding domain in EMSA reactions. Lanes 1-4: negative controls, 5-6 N. brichardi DNA binding assay, 7-8: M. zebra DNA binding assay, 9: kit negative, 10: binding positive control. Right: EMSA validation of increasing DNA binding domain concentrations and binding to predicted transcription factor binding site in N. brichardi sws1 promoter. Image modified from (Mehta et al. 2021)*, licensed under CC BY 4.0.*

The results show that the variations in nucleotides found in the binding sites of transcription factors can drive regulatory divergence through the observed GRN rewiring events. The elements identified by network rewiring highlighted traits under natural or sexual selection, such as the visual system, possibly shaping cichlid adaptation to a variety of ecological niches.

3.4.4. Discussion

In this work, a computational workflow was established to study gene regulatory networks of non-model organisms. We applied this workflow on five East African cichlid fish species to study examples of adaptive radiation through the evolution of GRNs. By putting putative predicted regulatory interactions to the genomes in a tissue specific manner, the approach has shown that network comparison methods, more specifically network rewiring methods can highlight regulatory hotspots in gene regulatory networks, in this case caused by selection pressure, arising from natural or sexual selection. The combination of tissue-specific expression data with reconstructed gene regulatory networks captures lineage specific changes in a well-studied trait in the species of this group, the visual system. Using the DyNet network rewiring approach I was able to highlight a key regulatory variation in transcription factor binding sites of genes involved in this system, that have been then shown experimentally to break specific transcription factor – gene interactions, and as such drive gene regulatory network evolution, and drive evolutionary innovations in the studied species, that can help them adapt to different ecological niches. The generation and subsequent comparison of regulatory networks, through the utilised workflows can add functionality to the observed differences (e.g., regulation by a specific transcription factor) as compared to multiple sequence alignments between the species, for example.

While the biological system studied here is far removed from my main topic of investigation, the network rewiring workflow and strategy developed and applied in this study was also implemented in my investigation to analyse changes in

Salmonella strains as well. Gene regulatory network rewiring has been successfully shown to drive phenotypic diversity in other kingdoms of life as well, from leaf shape to the emergence of pregnancy in mammals, showing how the approach itself is agnostic of a model system (Ichihashi et al., 2014; Lynch, Leclerc, May, & Wagner, 2011).

3.4.5. Future research directions

The approaches used in this study to construct regulatory networks from tissue-specific RNA-Seq data for multiple species can serve as a general guideline for other model organisms in the future. The majority of the results captured by the network rewiring and other methods applied in this study have not been verified experimentally, and further examination of these transcription factor – (opsin) target gene interactions could shed further light on the variances and sites under selection within the visual system in the studied cichlid species.

While this study focused finding divergent *cis*-regulatory elements involved in adaptation to certain conditions, other levels of regulation could also be used for this reason, with adequate changes to the methodology, e.g. post-transcriptional regulation, studies of enhancer regions, tracking gene duplication events.

3.5. Applications of SalmoNet 2 – Using network rewiring to identify functional differences in *Salmonella enterica*

To explore the utility of a multi-layered network resource such as SalmoNet, I compared the degree of interaction rewiring between the interactomes of host adapted typhoidal *Salmonella* strains and gastrointestinal *Salmonella* strains, captured the potential functional differences using Gene Ontology enrichment analysis, and compared the rewired subgraphs to find the causes of the rewiring. The approach outlined in this section could be used to provide insight into the functional differences caused by differing interactions between *Salmonella* pathovars, highlight key genes and proteins, and importantly provide targets for hypothesis generation and experimental validation.

SalmoNet 2 added three additional typhoidal *Salmonella* serotype strains to the interaction resource, now containing four typhoidal pathogens in total. Comparing the interactions patterns of these extraintestinal strains by contrasting them with those of gastrointestinal *Salmonella* could be utilised to show conserved or diverging subnetworks in these strains, related to their invasive lifestyle, and help researchers better understand these important human pathogens.

The results shown in this section indicate that SalmoNet 2 and the enrichment analysis captures adequate biological information, and the presence/absence of edges can pinpoint nodes under a level of selection pressure, that could be subjects of future work and experimental testing.

3.5.1. Methods

To calculate network rewiring I used the DyNet app in Cytoscape to calculate the rewiring value of the nodes in each group separately. Four typhoidal strains (*S.* Paratyphi A (AKU 1261), *S.* Paratyphi A (ATCC 9150), *S.* Paratyphi C (RKS4594), *S.* Typhi (Ty2) and four gastrointestinal strains (*S. A*gona (SL483), *S.* Newport (SL254), *S.* Heidelberg and *S.* Typhimurium (LT2)) were compared for interaction differences.

The level of rewiring was calculated across all strains, and the degree-corrected rewiring values were ordered in a descending list, where the top 50 hits were further analysed. Rewiring was also calculated within-group (i.e., between the selected typhoidal, and between the selected gastrointestinal strains separately), to allow for identifying which group is the source of variance comes when comparing all strains. To alleviate the bias towards hub nodes, the degree corrected D_n value was used for the cutoff.

To calculate the enrichment of Gene Ontology terms in the identified subgraphs I downloaded the up-to-date Gene Ontology annotation of the target genes using the topGO library in R, and following that the R library clusterProfiler was used to calculate Gene Ontology enrichment with the enricher() function, from Biological Process terms (Alexa & Rahnenfuhrer, 2021; Wu et al., 2021). P-

value adjustment for multiple testing was done via the Benjamini-Hochberg approach, using the *p.adjust* function in R.

The statistically significant enrichment results were compared side-by-side between the groups, and the differences in enrichment were further studied by comparing the sets of genes responsible for (underlying) the enriched terms, i.e., if one group was enriched in a specific term, the presence/absence of the orthologous genes responsible for the enrichment was analysed in the members of the other group. Terms were deemed respective of a group if that term was present in all members of a group, and simultaneously in one or none of the other.

To study the relationship of YreP and YjcS to the extraintestinal pathovar, network rewiring was calculated in an identical manner as above, but all extraintestinal and gastrointestinal strains from SalmoNet 2 were involved in the comparisons.

BLAST searches for the YreP and YjcS genes was done through the pubMLST website, with default parameters (Jolley, Bray, & Maiden, 2018). The entire genomic sequence of the genes and their shared regulatory region was queried, as taken from *S.* Gallinarum strain 287/91. The hits were filtered for above 95% sequence identity, and the top 10% of bitscores to make sure the compared sequences contain both the genes and the shared regulatory region.

3.5.2. Results

To evaluate whether the interaction networks in SalmoNet 2 can be used to the study the effects of host adaptation in invasive strains of *Salmonella*, I studied and compared the interaction networks of four typhoidal and four

gastrointestinal strains using the DyNet network rewiring tool. In total, I analysed the 50 highest D_n nodes, and I highlighted the potential functional relevance of these interaction differences using Gene Ontology enrichment analysis.

In general, many of the top hits or most rewired nodes are important global regulators, such as Crp, Fis and Fur, despite correcting for degree bias. The significantly enriched functions are similar between the compared strains, with a few key differences. For example, the gene Fur or ferric uptake regulator senses metal concentration and the redox state of cells, and regulates many operons and genes involved in these processes (Troxell, Fink, Porwollik, McClelland, & Hassan, 2011). Fur is enriched in the GO function "iron ion homeostasis" in all included gastrointestinal strains, while this enrichment is absent from the typhoidal strains. Upon further inspection of the genes responsible for the enrichment of the term and their orthologous status, Fur is missing interactions present in GI strains towards the genes *fhuA, fhuE,* caused by the disruption of coding sequences in these genes in the typhoidal serovars, as highlighted previously in the literature (Nuccio & Bäumler, 2014; Y. Wang et al., 2018).

Fur is similarly enriched in the term "cell adhesion" in all gastrointestinal strains, whereas this function is not enriched in typhoidal strains, except *S.* Paratyphi C. Once again, inspection of the genes underlying the enrichment result reveals that the culprit behind the mismatch in functional enrichment is the pseudogenization and subsequent missing interactions with the genes *stiH* and *stiA* in the rest of the typhoidal *Salmonella* strains, two genes responsible for the production of fimbriae, highlighted previously in the literature (Nuccio & Bäumler, 2014). Figure 25. shows the functional differences

in Gene Ontology enrichment between the compared groups through examples involving the Fur transcription factor.

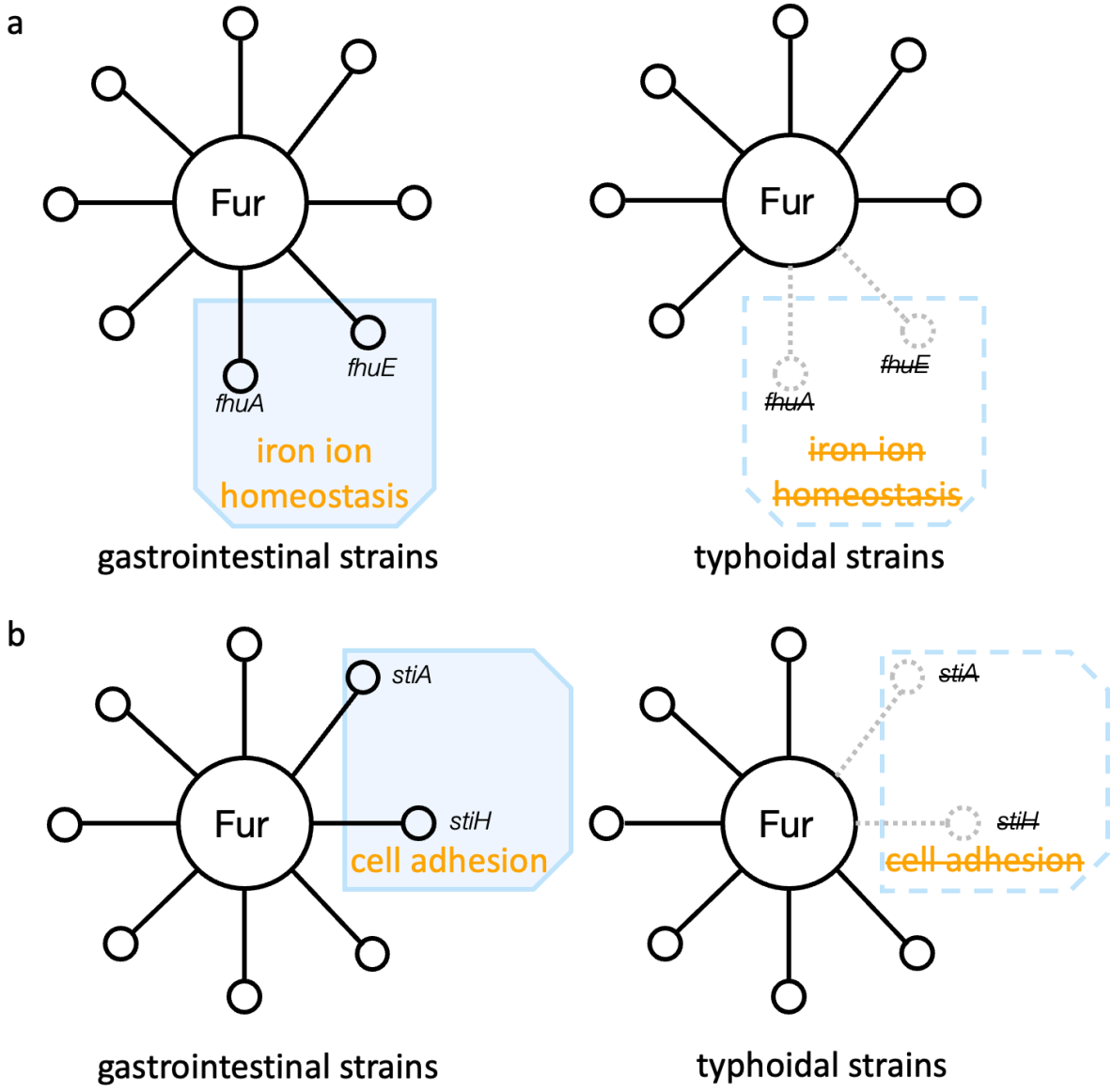

Figure 25. The causes of different Gene Ontology enrichment between gastrointestinal and typhoidal strains as highlighted by rewiring analysis. The Fur regulon controls many genes that undergo pseudogenisation in invasive strains of Salmonella. The rewiring analysis highlights two such examples for the regulon, where genes involved in metabolic processes, such as iron ion homeostasis (a), or fimbriae production (b) get disrupted, and the consequent loss of interactions causes a loss in the enrichment of functions.

The rewiring analysis has revealed many other examples like this, where the cause of rewiring can be led back to the disruption of coding sequences and thus loss of interactions as described by (Nuccio & Bäumler, 2014). For example, LyxK is enriched in "carbohydrate phosphorylation" and "C4-dicarboxylate transport" in GI strains but not in typhoidal ones, or 'nitrate assimilation' in the case of Fnr, showing the same split across groups. These examples highlight the functional, phenotypical effects of genome reduction on the interaction networks of extraintestinal strains as they adapt to their host, which has been well documented in the literature (Hiyoshi et al., 2018; Langridge et al., 2015; MacKenzie et al., 2019; Vázquez-Torres, 2018). As detailed in Chapter 1, gene inactivation caused by genome degradation is one of the recurring features of host adaptation. The specific changes captured by this analysis reflect two often changing processes in extraintestinal strains, those of anaerobic metabolism and cell adhesion (fimbrial genes) (Nuccio & Bäumler, 2014). The rewiring analysis highlights the magnitude of change genome reduction incurs on the interaction networks of typhoidal strains, as most of the analysed genes were rewired due to this phenomenon (Nuccio & Bäumler, 2014). From the top 50 most rewired nodes, on average 33 nodes had at least one pseudogene first neighbour in the typhoidal serovars, and on average 4% of the first neighbours of the top 50 most rewired nodes were pseudogenes. In the gastrointestinal strains on average 7 nodes had pseudogene first neighbours, and only 1% of their first neighbours were pseudogenes.

While a large part of the rewiring was due to gene loss in typhoidal and extraintestinal serovars, during my work I found examples where the cause of rewiring was due to the exclusivity of genes to the invasive group. The following

analysis is an example of how SalmoNet 2 can be used in conjunction with other computational tools to make inferences regarding the role of previously unknown or less known interactions and genes.

The two proteins, YreP and YjcS, are present in all extraintestinal serovars of *Salmonella* in SalmoNet 2 but are missing from almost all gastrointestinal strains. This interesting pathovar specific split led me to analyse the genes in more detail, with the assumption that they could be involved with the invasion process in a pathovar specific manner. The protein YjcS has an orthologue in *S.* Enteritidis, but the protein is otherwise missing from the gastrointestinal group. Since their presence is, except for the *S.* Enteritidis case, is restricted to the host adapted serovars in SalmoNet 2, they only receive interactions in these strains. The two genes get regulatory input from three transcription factors: HilC and RtsA involved in the regulation of SPI-1 genes amongst others, and the global regulator Fur. YjcS has an additional protein-protein interaction in all strains where it is included, as the protein can interact with itself. The source of the regulatory interactions are the genome wide scans of the *Salmonella* strains included in SalmoNet 2.

The two genes have first been described together previously in *Escherichia coli*, in two analysed strains: *E. coli* SMS-3-5, and environmental pathogenic isolate with multiple antibiotic resistances, and *E. coli* (NMEC) O7:K1 strain CE10, causing neonatal meningitis. The first gene, *yreP* (*dgcY* in *E. coli*), encodes a diguanylate cyclase, its suggested function based on it carrying the signature GGDEF domain. Diguanylate-cyclases facilitate the production of c-di-GMP, a ubiquitous secondary messenger metabolite in prokaryotes (Povolotsky & Hengge, 2016; Ryjenkov, Tarutina, Moskvin, & Gomelsky, 2005). The second gene, *yjcS* (EcSMS35_*1714 in E. coli)*, is an alkyl-sulfatase. This enzyme has been

first described in *Pseudomonas*, where a strain carrying this enzyme was able to grow on the surfactant sodium dodecyl sulfate (SDS), and the gene has been characterised in *E. coli* as well (Liang, Gao, Dong, & Liu, 2014; Williams & Payne, 1964). The orthologous proteins exist in multiple other genuses within the *Enterobacteriaceae* family, such as *Enterobacter,* or *Klebsiella* based on synteny information in the OMA database (Altenhoff et al., 2020). Since SalmoNet 2 is now based on orthologous information from the OMA orthology database, users can very quickly look up the phylogenetic spread of proteins of interest, like in this example with YjcS and YreP.

After noting their presence in the well-studied extraintestinal strains included in SalmoNet 2, I expanded the search into a more expansive data source, to see if this was representative of the serovars as a whole, not just the specific strains in SalmoNet 2. BLAST searches were executed within the available genomes of the pubMLST database (Jolley et al., 2018). PubMLST itself is a collection of databases, containing 18638 *Salmonella* genomes (accessed on 09/10/2021). The entire gene sequences of YreP and YjcS were input as the BLAST search query, including their shared regulatory region. Figure 26 shows the results of the BLAST searches in the pubMLST database.

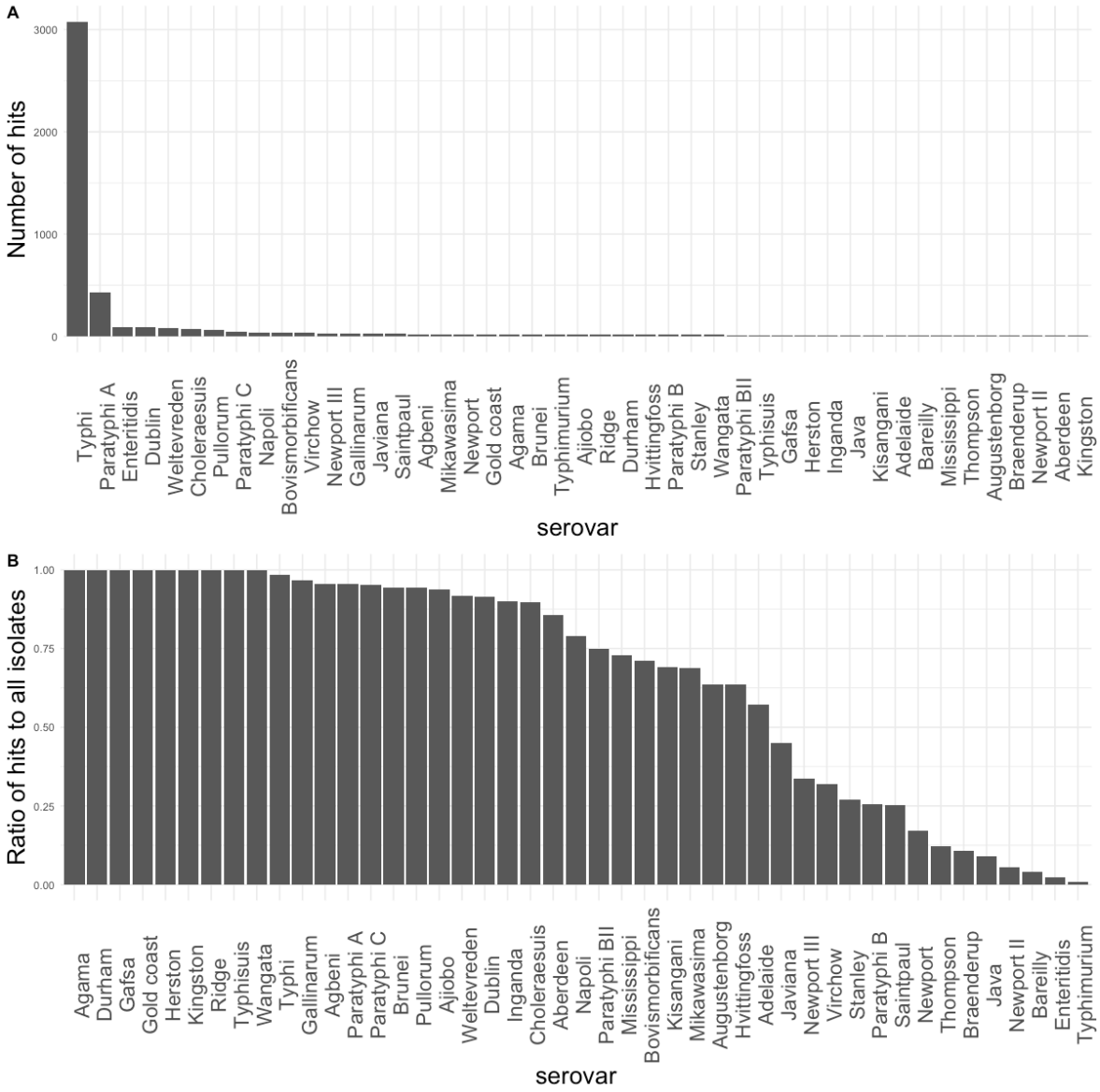

Figure 26. Prevalence of the yreP + promoter + yjcS segment in Salmonella serovars based on BLAST hits. Figure 26 A shows the total amount of hits in each serovar. Figure 26 B shows the ratio of hits to all isolates in the serovar. Serovars containing < 5 isolates were removed from this figure for clarity.

The distribution of BLAST hits from the genomes included in the pubMLST database is for the most part in accordance with what I have found in SalmoNet 2, the genomes of well-known extraintestinal serovars seem to contain

the sequence more often. In total 83% of hits come from well-known extraintestinal serovars, dominated by *S.* Typhi strains. The top 10 serovars in terms of number of hits are also mostly invasive serovars: *S.* Typhi, *S.* Paratyphi A, and *S.* Paratyphi C are notable typhoidal serovars adapted to humans, *S.* Dublin, *S.* Pullorum and *S.* Choleraesuis are well-known host adapted serovars of cattle, poultry and pigs (Métris et al., 2017; Tanner & Kingsley, 2018). *S.* Napoli is an emerging serovar in Europe, phylogenetically closely related to *S.* Paratyphi A, carrying an almost identical pattern of typhoid-associated genes, and capable of causing a form of invasive nontyphoidal disease (Gori et al., 2018; Huedo et al., 2017). The invasive behaviour is not as clear cut with the rest of the serovars, but there have been reports of it in the literature. *S.* Bovismorbificans is capable of causing bloodstream infections, and has recently been described as an emerging disease in Malawi, converging towards a phenotype resembling a human adapted iNTS variant (Bronowski et al., 2013). Although not strictly an extraintestinal serovar, *S.* Virchow has been known to cause invasive illness (Eckerle, Zimmermann, Kapaun, & Junghanss, 2010; Mani, Brennand, & Mandal, 1974; Messer, Warnock, Heazlewood, & Hanna, 1997; Todd & Murdoch, 1983). *S.* Weltevreden is an emerging cause of diarrheal and sometimes invasive disease in humans in tropical regions, and is hypothesized to be adapted or adapting to life in aquatic hosts (Hounmanou et al., 2020; Makendi et al., 2016). While large in total numbers in the database, *S.* Enteritidis only makes up 2% of the positive hits. Since *S.* Enteritidis is one of the most commonly isolated iNTS strains, there exists a possible link to invasive behaviour (Feasey et al., 2016; M. A. Gordon, 2011). However, more work is needed to uncover the cause and extent of this curious split between pathovars.

This brief analysis above highlights how the information contained in and linked with SalmoNet 2 can be used to form scientific questions relating the functionality of genes to the behaviour and phylogenetics of *Salmonella,* based on molecular interaction information. SalmoNet 2 contains example strains from the most prevalent serovars, and the information can further be extended using the easily accessible sequence data and homology information through OMA and other computational resources.

3.5.3. Discussion & Future research directions

In this work I demonstrated how large-scale network rewiring analysis can be applied to compare interaction networks of gastrointestinal and typhoidal *Salmonella* serovars. The results highlights the effects genome degradation has on host adapted *Salmonella*, as the loss of genes related to anaerobic metabolism, chemotaxis and related functions were present behind the rewiring for the majority of most rewired nodes in the analysed typhoidal strains (Holt et al., 2009; Nuccio & Bäumler, 2014).

In a second round of analysis, I demonstrated how downstream investigations can be followed through using information gained from SalmoNet 2 through the example of the YreP and YjcS proteins, that seem to associate to host adapted serovars, and do not have orthologous proteins in gastrointestinal serovars barring *S.* Enteritidis. Querying their genomic sequence from the linked OMA database and running BLAST searches against more than 18000 *Salmonella* genomes results in a similar picture.

However, the results shown here are inconclusive, and require further functional analysis of the studied genes, and whether they are beneficial to the

invasion process at any point. Based on available information we can only hypothesize what their roles could be. The potential role of YjcS is in the invasive lifestyle is difficult to assess, owing to the lack of relevant information of the alkyl-sulfatase domain structure in this setting (Liang et al., 2014). YreP on the other hand, a diguanylate-cyclase, can potentially affect host adaptation. The role of diguanylate cyclases and c-di-GMP in bacteria has been mostly understood as a sessile-motile switch, first shown in *Vibrio*, where low c-di-GMP levels correspond to the host environment and increased motility, while high c-di-GMP levels decrease motility in the aquatic environment, highlighting how the potentially increased production of c-di-GMP can influence cell fate decisions relating to virulence and sessility/biofilm formation (Ahmad et al., 2011; Jenal et al., 2017; Tamayo et al., 2007). The decisions to reduce motility, and/or virulence and start producing biofilms are quite important and severe from the bacteria's point of view, and as such are under tight spatial, temporal and multiple levels of regulatory control, and there are often direct protein-protein interactions occurring between the effectors and the signalling enzymes (Hengge, 2009). However, host adapted serovars are for the most part weaker biofilm formers than their gastrointestinal counterparts (MacKenzie et al., 2017). As parts of the cellulose synthase operon are pseudogenized in many of the host adapted serovars (Nuccio & Bäumler, 2014), I hypothesize the consequences of increased c-di-GMP levels may be different in terms of sessility-motility or virulence attenuation in extraintestinal and gastrointestinal *Salmonella* serovars.

However, more future work is required to confirm whether the two genes have any role in the invasion process, and to further solidify their link to extraintestinal serovars. We have started a series of RNA-Seq experiments to capture the differentially expressed genes between a wild-type and *ΔyjcS/ΔyreP*

knockout strain of *S.* Gallinarum, but unfortunately the experiments were halted by the COVID-19 pandemic. The transcriptomics readout could potentially answer what sets of genes are affected by the activity of the YreP and YjcS, and the mutants could be used in further future experiments involving other functions linked to c-di-GMP production, such as efflux pump activity (Holden & Webber, 2020). While the latter half of this analysis could have been done without network information, since SalmoNet 2 predicts multiple upstream regulatory interactions to potentially control *yreP* and *yjcS,* future experimental work could evaluate which one of these regulators, and under what circumstances regulate these genes, which is the added value of the network biology approach.

4. The role of cytokines in SARS-CoV-2 infection

4.1. Introduction

The currently ongoing pandemic has mobilised scientists and people around the world, to understand and resolve the infection caused by the novel pathogen. Globally, as of the writing of this text (16th March 2021), there have been 119,791,453 confirmed cases of COVID-19, including 2,652,966 deaths, reported to the WHO (https://covid19.who.int/). Major efforts now concentrate on how severe acute respiratory syndrome β-coronavirus 2 (SARS-CoV-2) changes the efficacy of normal antiviral immune responses, and why host antiviral immune responses are unable to resolve it in a subgroup of patients. The clinical symptoms of the disease range from asymptomatic, through mild (fever, persistent cough, loss of taste and smell, gastrointestinal problems) to severe pneumonia, organ failure, and even death (Pedersen & Ho, 2020). Although SARS-CoV-2 appears to alter host inflammatory defences, similar modifications have also been observed in the recent β-coronavirus epidemics caused by SARS-CoV and MERS-CoV, and the ones responsible for the H5N1 and H7N9 influenza A subtype outbreaks (Channappanavar et al., 2016, 2019).

Although these viruses can cause similar symptoms, the specifics of the pathogenesis may be caused by different factors. A shared trait is their effect on the pro-inflammatory host immune response. One of the key characteristics of these viruses, including SARS-CoV-2, is that they can lead to a Cytokine Release Syndrome (CRS), or "cytokine storm", which can increase the mortality observed for this illness in a subgroup of patients (P. Mehta et al., 2020). This phenomenon occurs when a large number of innate and adaptive immune cells, such as B-cells, T-cells, NK-cells, macrophages or dendritic cells activate, and start producing pro-inflammatory cytokines, establishing a feedback loop of inflammation. This process normally resolves after the antiviral response successfully clears the pathogen from the host, but it can persist in serious cases. In these situations, the inflammatory response can become so severe it damages organs and tissues, and can eventually lead to death (Del Valle et al., 2020).

One of the first lines of defence against viral infections is the type-I interferon response, carried out by the type-I interferons (IFN-α, -β, -κ, -ε, -τ, -ω and -ζ). Produced by a large number of cell types, as part of the innate immune system they are an ancient, very conserved evolutionary response against viruses. Their role is activating a cascade of signalling that results in the expression of a cluster of genes, called the Interferon Stimulated Genes. These cascades can attenuate the inflammation to avoid tissue damage, lead to the production of cytokines such as IL-12, and further carry the signal, eventually resulting in the activation of the adaptive immune response through IFN-γ (Betakova, Kostrabova, Lachova, & Turianova, 2017; Kang, Brown, & Hwang, 2018; Makris, Paulsen, & Johansson, 2017).

Due to the changed circumstances caused by the pandemic, I was re-deployed for 6 extra months during my PhD studies to work on COVID-19 related research, working on these topics in the Korcsmaros Group, and as a member of the COVID-19 Disease Map community (Ostaszewski et al., 2020). The results presented in this chapter build on the biological and methodological knowledge I acquired over the years as a postgraduate research student. While many studies focus on the intracellular effects of the virus, from its entry into the cell through TMPRSS2 and ACE2, to the downstream affected pathways, because of the potential danger of CRS I wanted to map the pathogenic process backwards. My goal was to trace and compare the cytokine responses caused by SARS-CoV-2 and similar viruses, and highlight the affected, differently behaving source cell types. The goal was to find conserved, and unique immune response patterns between CRS-causing viruses to help specialists identify interventions that can alleviate serious cases of COVID-19, and other illnesses that cause CRS, and potentially pinpoint immune cell populations that behave differently than expected in CRS.

In this chapter, I am going to detail the results of two projects I led during my redeployment. The first one, published in *Frontiers in Immunology*, is a systematic literature curation of cytokine responses to CRS-causing viruses from the relevant literature (Olbei et al., 2021). In the second project, I developed a novel network resource, CytokineLink, aimed at highlighting how cell types can communicate using cytokines, with the goal in mind that the established networks can pinpoint specific cytokines that mediate intercellular communication between important celltypes and tissues.

Although these were the main studies I worked on during this period, I was also involved in two other works, led by other members of our group. The first such

project was ViralLink, published in *PLOS Computational Biology*, a systems biology workflow which reconstructs and analyses networks representing the effect of viruses on intracellular signalling (Treveil et al., 2021).

The second such project, still in progress, has the working title of "Gut-COVID project", where we study the effects of the SARS-CoV-2 infection on the intestine, which was shown to be productively infected by the virus earlier last year (Lamers et al., 2020).

4.2. Aims

The aims of these projects were the following:

- Collect the available patient derived cytokine response data for five cytokine release syndrome causing viruses, including SARS-CoV-2, through a systematic curation process.
- Compare the acquired cytokine data, and study the observed differences.
- Generate a novel network resource (CytokineLink) to map cytokine mediated signalling using patient derived data in COVID-19 and other inflammatory and infectious diseases, and assign validity to its interactions through curation via systems immunology databases and the published literature.
- Connect the two works by analysing the COVID-19 cytokine response data we collected with CytokineLink.

4.3. Methods

The results and approaches in the first half of this chapter (comparing cytokine responses from five cytokine release syndrome causing viruses) were developed by me, with the help of Isabelle Hautefort, who did one half of the curation and worked on the interpretation of the data, as well as with support from Dezso Modos, who helped in the formal analysis, hierarchical clustering, and interpretation of the results. Claire Shannon-Lowe, Agatha Treveil, Martina Poletti and Leila Gul contributed to the paper that forms the foundation of this chapter.

The second half of the chapter details the construction of a novel network resource aimed at understanding cytokine-mediated intercellular communication. The resource was conceived of and developed by myself, with the help of Dezso Modos, Isabelle Hautefort and Tamas Korcsmaros, who advised me during the project.

4.3.1. Comparing cytokine responses from five cytokine release syndrome causing viruses

4.3.1.1. Literature curation

We performed a mass literature search of 98 well-studied cytokines in the PubMed resource using PubTator, and in the bioRxiv and medRxiv non-peer reviewed pre-publication repositories (Wei, Allot, Leaman, & Lu, 2019). The

targets include commonly studied interleukins, interferons, chemokines involved in both anti- and pro-inflammatory responses, in particular those that are involved in CRS. We only included studies where the direction of change in the level of a cytokine was included. We could not meaningfully collect the amplitude of change, only the presence or absence of change. We restricted the study to five CRS-causing viruses, all of them responsible for an epidemic in the past few decades. Three β-coronaviruses SARS-CoV, SARS-CoV-2 and MERS-CoV, and two influenza A subtypes, H5N1 and H7N9. The names of each pathogen, and each cytokine was used as search terms, e.g. "H5N1 CCL2". In ambiguous cases, multiple forms ("IFN-b", " IFN-beta", "IFN b", "IFN beta") were tried. If the search resulted in more than 50 hits "patient" was added to the search terms.

The resulting articles were then manually processed for cytokine data. We only considered results valid for curation, if the results came directly from studies including at least 10 patients, i.e. model-organism, or cell-line based results were excluded. From the main text of the resulting articles the direction of change of the listed cytokines was noted in a spreadsheet. We closed the curation on 03/06/2020. We estimated the size of the discarded literature using a shell script, available in a GitHub repository [https://github.com/korcsmarosgroup/CRS]. Figure 27 highlights the steps of the curation process.

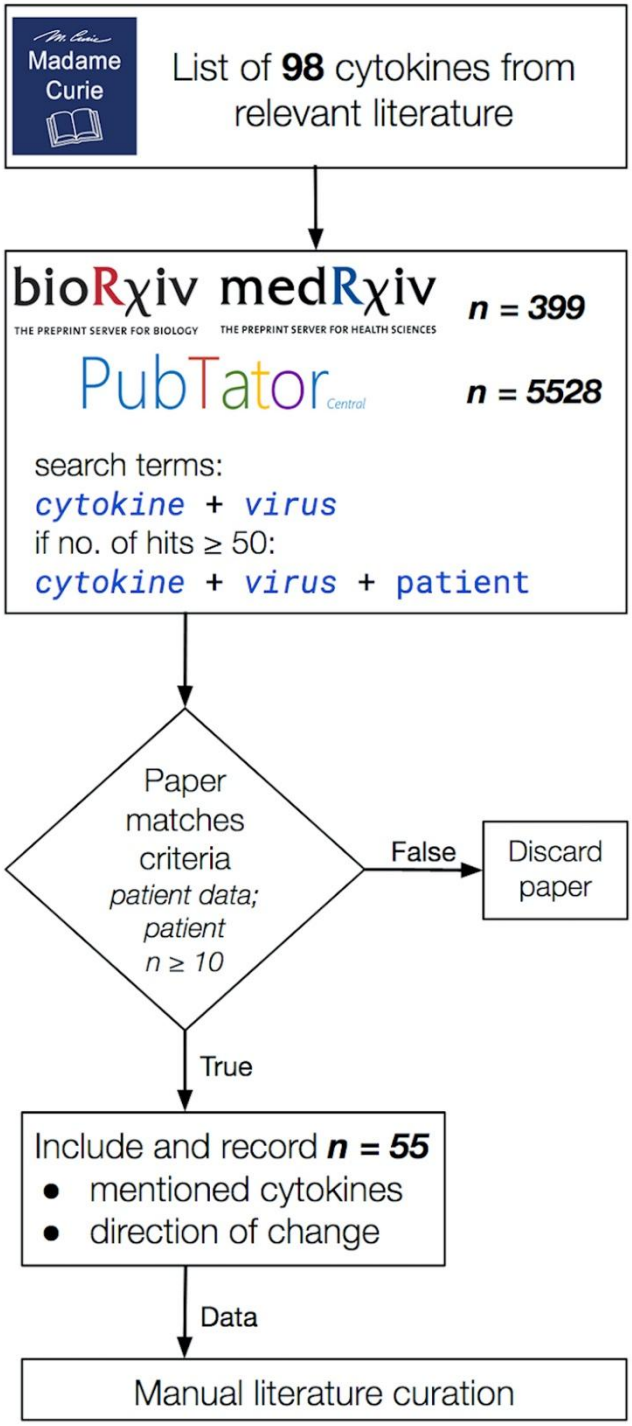

Figure 27. The literature curation workflow applied in this study. Publications were considered valid for inclusion into our data collection i) if they contained patient-derived data (model organisms and cell lines were excluded) and ii) the study data were collected from cohorts of at least 10 participants per group iii) if it included a directional change in cytokine levels. In the end 55 publications were selected that matched the criteria above.

4.3.1.2. Hierarchical clustering

The hierarchical clustering of the cytokine clusters was done using the *seaborn* python package, with Jaccard index as the similarity measure, and the complete linkage method. The former similarity measure is used to calculate the dissimilarity of two sets, while the latter linkage method calculates each cluster's distance from each other (the farthest point from each cluster). This approach is sensitive for the furthest elements, and it does not join the furthest elements, giving a clearer result. It performs well when applied to finding appropriate clusters in synthetic studies. The code for the clustering method is available at the GitHub repository. [https://github.com/korcsmarosgroup/CRS]

4.3.2. Construction of an intercellular cytokine-cytokine communication network resource, CytokineLink

The cytokine-cytokine interaction network resource was built between tissues and blood cell types available in the Human Protein Atlas (Uhlén et al., 2015).

The consensus RNA-Seq data for all cytokines and their receptors listed in ImmuneXpresso and ImmunoGlobe, and a relevant literature source was downloaded using a custom shell script, and processed using an R language script (Atallah et al., 2020; Kveler et al., 2018).

To establish potential interactions between tissues and cell types mediated by cytokines, I made the following abstraction. Using cytokine - receptor interactions received from the appropriate literature and the OmniPath database (Cameron & Kelvin, 2013; Türei et al., 2016), I created meta-edges between

tissues and blood cell types expressing these cytokines, creating the meta-network of tissue level cytokine communication. To uncover prospective cytokine - cytokine interactions, I inverted the same dataset. In these cases, the interactions signify what cell types can certain cytokines act on, and what cytokines do these cell types produce. An example interaction, with all of the subsequent steps involving IL-7 and IFN-γ can be seen on Fig 28.

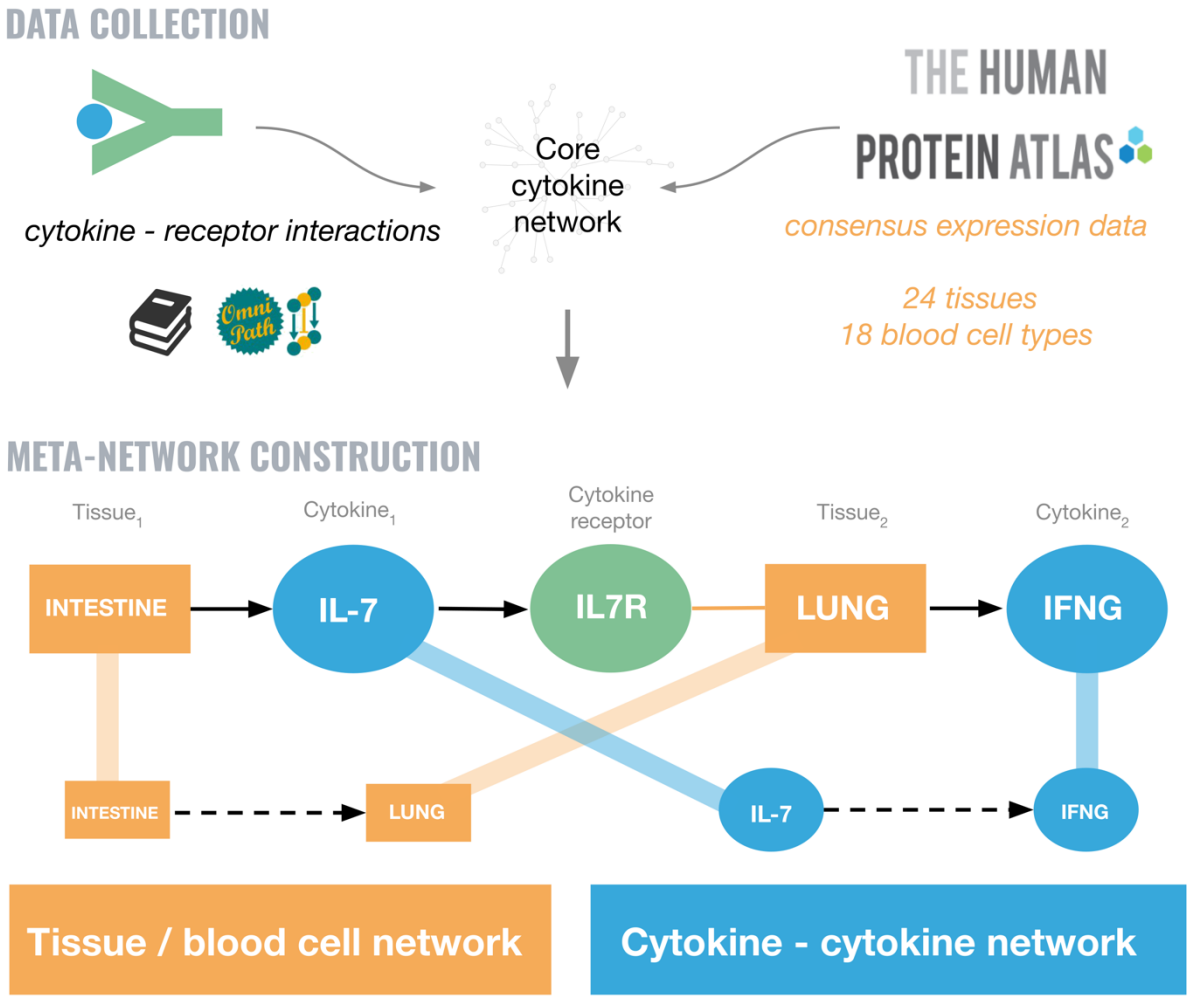

Figure 28. Construction of CytokineLink. Data was downloaded from the Human Protein Atlas, with cytokine-receptor data queried from OmniPath and the relevant literature. The base networks contain tissue - cytokine - receptor interactions, from which the abstracted meta-edges were created. These meta cytokine - cytokine edges symbol the potential ways which the production of a cytokine can alter the production of another, by binding to its receptor, carried by a cell type expressing the secondary cytokine.

These custom paths (i.e. $Tissue_1$, $Cytokine_1$, Receptor, $Tissue_2$, $Cytokine_2$) were generated using the tissue-cytokine interactions as an input, with the *get_simple_paths()* function of the *NetworkX* (version: 2.5) python library (python version: 3.8.6).

Beyond containing important cytokine data, ImmuneXpresso and ImmunoGlobe also curates interaction data collated from the literature, between cell types and cytokines, both as sources and sinks (Atallah et al., 2020; Kveler et al., 2018). These interaction annotations were included in the resource, assigning confidence to the underlying edges.

To add further information to the network, a layer of regulatory interactions was also integrated into the data resource. The interactions were included from a recent publication, utilising enhanced yeast-one hybrid assays to collect interactions between 265 transcription factors and 108 cytokines (Santoso et al. 2020).

The network can be instantiated using custom (e.g. single-cell RNA-Seq) datasets, in these cases the presence/absence of cytokines is based on that of the input data, instead of the Human Protein Atlas results.

For the COVID-19 use case, the elevated cytokine list contained: CCL2, CCL3, CCL4, CSF2, CXCL10, CXCL11, IFNG, IL1B, while the complemented cytokine list contained: IL2, IL4 and IL5. The networks were generated in Cytoscape

(Shannon et al., 2003), by filtering down the cell-cell networks to ones only containing interactions mediated by the above cytokines.

4.4. Results & Discussion

4.4.1. SARS-CoV-2 causes a different cytokine response compared to other cytokine storm-causing respiratory viruses in severely ill patients

In this work, we collected cytokine responses from patient-derived data published in the literature that met our curation criteria listed above. The curation protocol followed the steps shown on Figure 27.

We compared the amount of increased, decreased or mixed status cytokines from the collected literature. Figure 29. shows the number of cytokines measured in the studies for each of the five CRS-causing viruses.

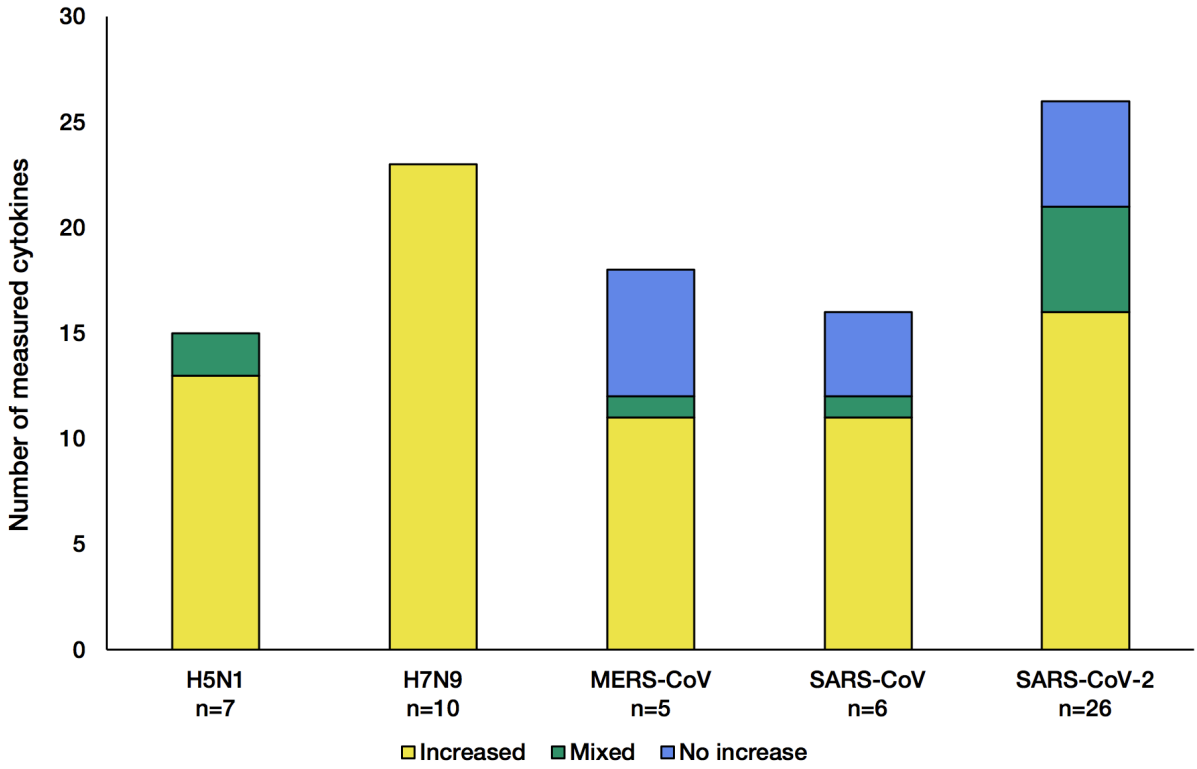

Figure 29. The number of cytokines measured in the included studies for each of the five CRS-causing viruses. Each bar of the stacked barcharts represents how many different cytokines were found increased (yellow), reported as both increasing and not increasing (green) or not increasing (blue). The n number on the bottom of the chart corresponds to the number of articles citing cytokine changes during infection.

Out of the 98 queried cytokines, we found 38 that were included in the studies matching the curation criteria. Only a small group of cytokines was measured for all viruses (CXCL8, IL-6, CXCL10, IL-2, FN-γ, TNF-α). Figure 29. shows how variable the number of different measured cytokines is across the different viruses. This variation can be most likely attributed to the increased interest in CRS-causing viruses over the recent years, in no small part due to the current pandemic, and the increased availability and sensitivity of the detection methods.

One of the notable differences between influenza A subtypes and β-coronaviruses is that the former group triggers an increase in almost all measured cytokine

levels, while in the latter case, some cytokines were detected at levels normally found in control groups (non-increased), or the data disagrees between different studies (mixed results). This highlights the potential differences in the underlying kinetics and pathogenesis process between the CRS-causing viruses.

Specificity	Cytokines elevated at least in one study (elevated & mixed)
Virus-specific	16
Shared between 2 viruses	5
Shared between 3 viruses	8
Shared between 4 viruses	2
Common to all five viruses	5

Table 6. Number of cytokines elevated in at least one study. Column 2 shows the number of elevated (or mixed) measurements, and their overlaps between viruses. Mixed observations occur when one or more studies show no change in the level of a cytokine, whereas others show an increase.

Table 6. details the number of cytokines whose levels are increasing in one, two, three, four or all viral infections, from the curated literature. Only five cytokines are shared across all of the conditions (CXCL8, IL-6, CXCL10, IL-2, IFN-γ, TNF-α), and 20 are shared to a lesser degree. 16 cytokine responses are unique to the selected viruses.

A limitation of our study is that the amplitude of change for the measured cytokines is not included, which can be different between the different diseases and disease states. To examine the presence and absence of cytokine responses

between the viruses more in detail, we constructed a heatmap of collected data, shown on figure 30.

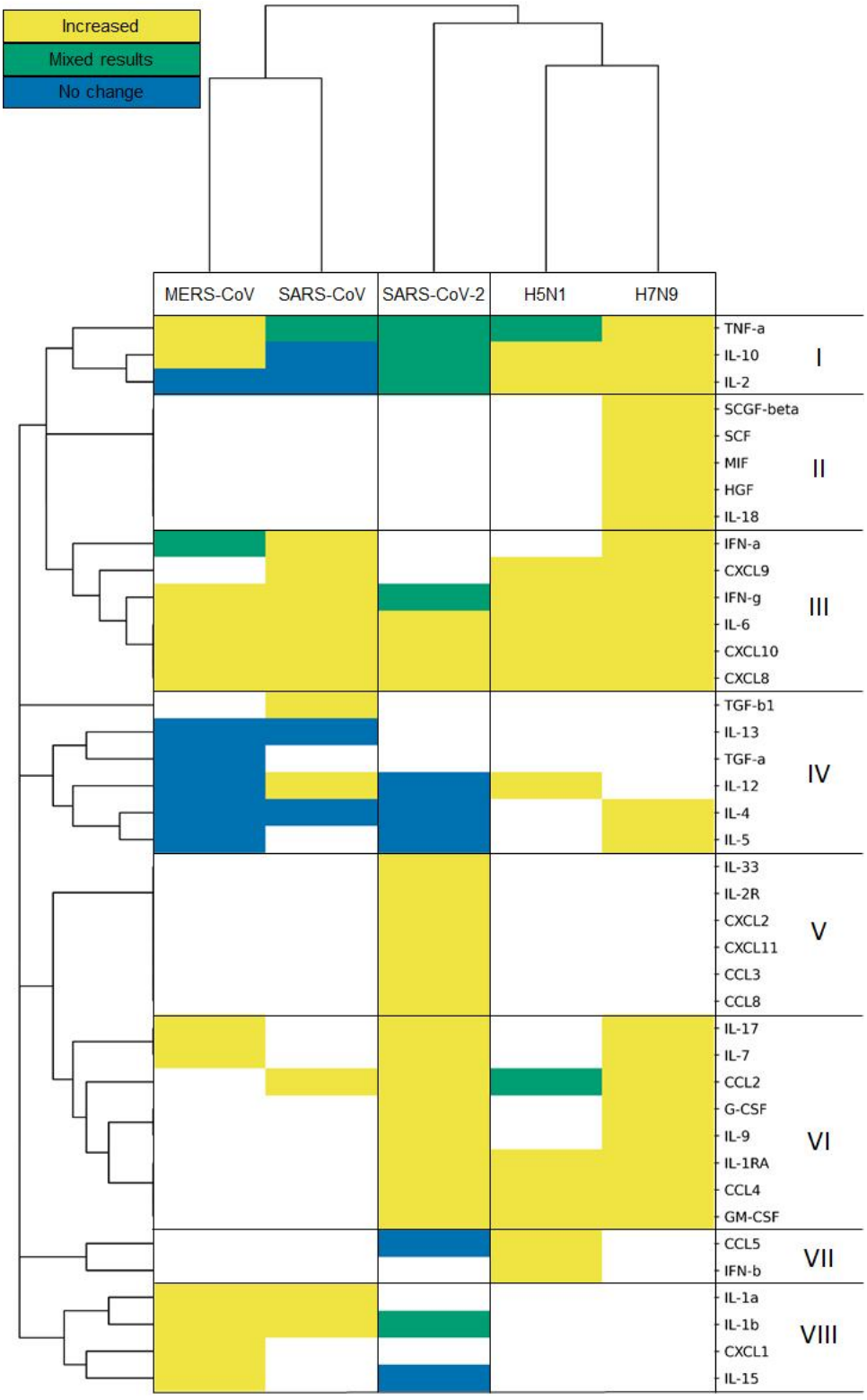

Figure 30. Hierarchical clustering of cytokine responses from influenza A subtype viruses and beta coronaviruses. The influenza viruses, SARS-CoV and

MERS-CoV, and SARS-CoV-2 form separate clusters. Hierarchical clustering results are based on Jaccard distance complete linkage.

Based on the results of the cluster analysis, eight clusters can be delineated. It is important to highlight that the resulting grouping can be biased by the missing information for certain cytokines.

With this in mind, cluster I. contains two anti-inflammatory cytokines IL-2 and IL-10, and the pro-inflammatory cytokine TNF-α. The literature reports mixed results for all three cytokines in the SARS-CoV-2 literature, but has all possible outcomes in the β-coronavirus cluster, while they are predominantly increased in the influenza viruses.

Clusters III and VI contain most of the increased pro-inflammatory cytokines, elevated for almost all viruses, but not measured in every case. Among them the cornerstones of the type I and type II interferon response, IFN-α and IFN- γ, and IL-6, one of the main pro-inflammatory cytokines, and target of many clinical interventions.

Cytokines from Cluster IV measured during coronavirus infections do not fluctuate, while most of them are elevated during influenza infection, e.g. IL-4 and IL-5 upon H7N9 infections. IL-4 is involved in Th2 differentiation, and the Th2 cells can produce IL-5 to mitigate eosinophil infiltration (X.-Z. J. Guo & Thomas, 2017). Such differences observed between virus-specific pathologies reflect the strong alterations caused by coronavirus infections, especially SARS-CoV-2 (Tan et al., 2020).

The cytokines in cluster VII and VIII do not always respond to SARS-CoV-2: IL-15 and CCL5 (also known as RANTES) are not elevated after SARS-CoV-2 infection. IL-15 is involved in natural killer cell differentiation as part of antiviral response (Y. Guo, Luan, Patil, & Sherwood, 2017). Meanwhile, CCL5 mediates eosinophil infiltration which is considered to be involved in the recovery after SARS-CoV infection alterations observed in coronavirus infections, particularly SARS-CoV-2 (Patterson et al., 2020).

Clusters II and V contain cytokines measured only in H7N9 and SARS-CoV2, respectively, whereas TGF-b1 was measured only in SARS-CoV studies in cluster IV.

To put the results more in context, we decided to focus on a small part of the infection process, and focus on the differences in cytokine responses involved in the type-I interferon response, and the cytokines involved downstream of it. Figure 31 shows the presence/absence of key cytokines in the analysed viruses in the process.

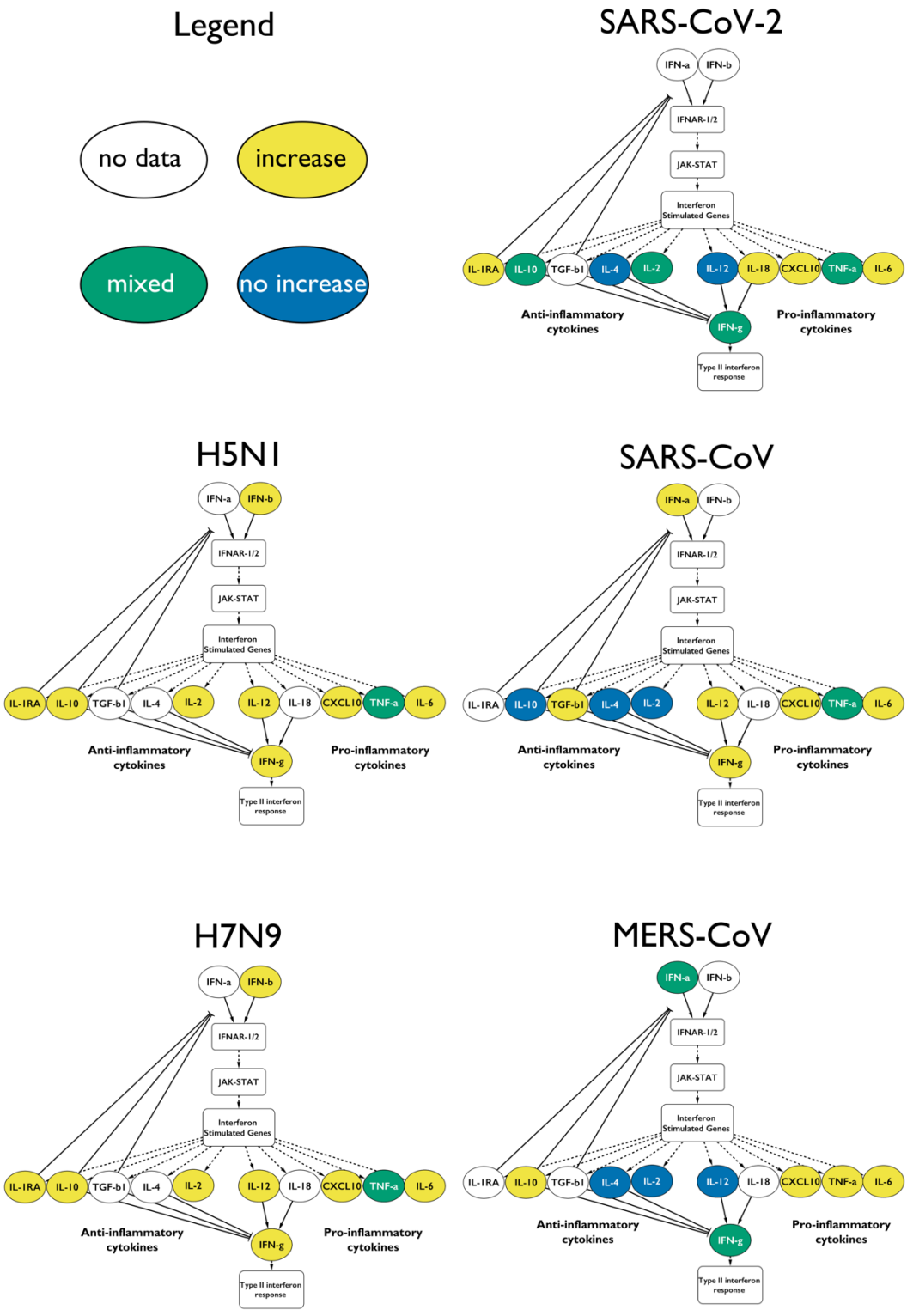

Figure 31. Type I interferon response upon infection with the different CRS-causing viruses. The measured cytokines in the influenza viruses are increased. In the case of the coronaviruses the response is mixed, not all of the anti-inflammatory cytokines are elevated. Only a fraction of cytokines shown for clarity.

Based on the data we collected, infection with either of the two influenza subtypes seems to increase the levels of measured type-I IFN-relevant cytokines, resulting in an antiviral immune response, with the appropriate cytokines showing elevated levels in all influenza A studies. However, the responses given to the coronaviruses show a more variable IFN-I response. In the case of SARS-CoV it is active, including the downstream activation of IL-12 and IFN-γ, which indicates the involvement of mature dendritic cells based on the former, and the activation of the type-II interferon response based on the latter signature.

For MERS-CoV, the type-I response is active, but there are some conflicts about parts of it in the literature. In certain studies IL-12 does not increase, in line with the inactivation of IFN-γ. Despite this, the mostly anti-inflammatory IL-10 is active, although caution should be taken as interpreting this cytokine as a solely anti-inflammatory, as there are more and more studies now confirming its role as a pro-inflammatory agent in certain scenarios (Mühl, 2013).

Based on the curated responses above we found that SARS-CoV-2 is characterized by an apparent dysregulation of the type-I IFN response, and consequently parts of the downstream cytokine machinery, involving IL-4, IL-12, IL-2, IL-10, and the type-II IFN response.

4.4.2. CytokineLink: an intercellular cytokine-cytokine communication network resource

To uncover how affected cytokines, such as the ones highlighted in the previous section might disrupt intercellular communication, I built a new network resource aimed at depicting all possible indirect interactions they can have, and contextualised it using Human Protein Atlas consensus expression data.

CytokineLink contains 24 tissues, 18 blood cell types, from which I generated two meta-networks (cell to cell, cytokine to cytokine), containing 6573 meta edges. In total, I included 115 cytokines, with 308 unique cytokine-receptor interactions. The latter has a large, 95% overlap with the published literature. To add more confidence to the interactions, I added annotation data from two systems immunology databases, immuneXpresso, and ImmunoGlobe, assigning literature references from research articles and textbooks, literature enrichment cores and signage (i.e. stimulatory / inhibitory) data to the interactions (Atallah et al. 2020; Kveler et al. 2018). 46% of interactions have at least one data point of annotation attached to them, indicating the degree to which the resource captures already known biology.

To demonstrate the applicability of the resource, I have selected the cytokines found to be increased in SARS-CoV-2 patients from the previous results section. While the optimal use case would be to apply it to single-cell RNA-Seq data involving many of the involved immune cell types, at the time of carrying out this project no such dataset was available. To help future studies utilising the resource

with single-cell data, an additional layer of cytokine-specific regulation was added, to add further context to up or downregulated cytokine responses. From the results of the CytokineLink resource based analysis, an interesting pattern emerges, as shown on Figure 31 A. The majority of cells that communicate using the elevated cytokines are involved in innate immune responses. From the T-cell elements MAIT T-cells generally show an innate-like behaviour, while gdT-cells generally bridge the response between the innate and adaptive immune system. The innate immune response, as recently shown, is the part of the COVID-19 disease process that might be the underlying cause of the large scale heterogeneity observed in outcomes (Hinks et al., 2020; Holtmeier & Kabelitz, 2005; Schultze & Aschenbrenner, 2021).

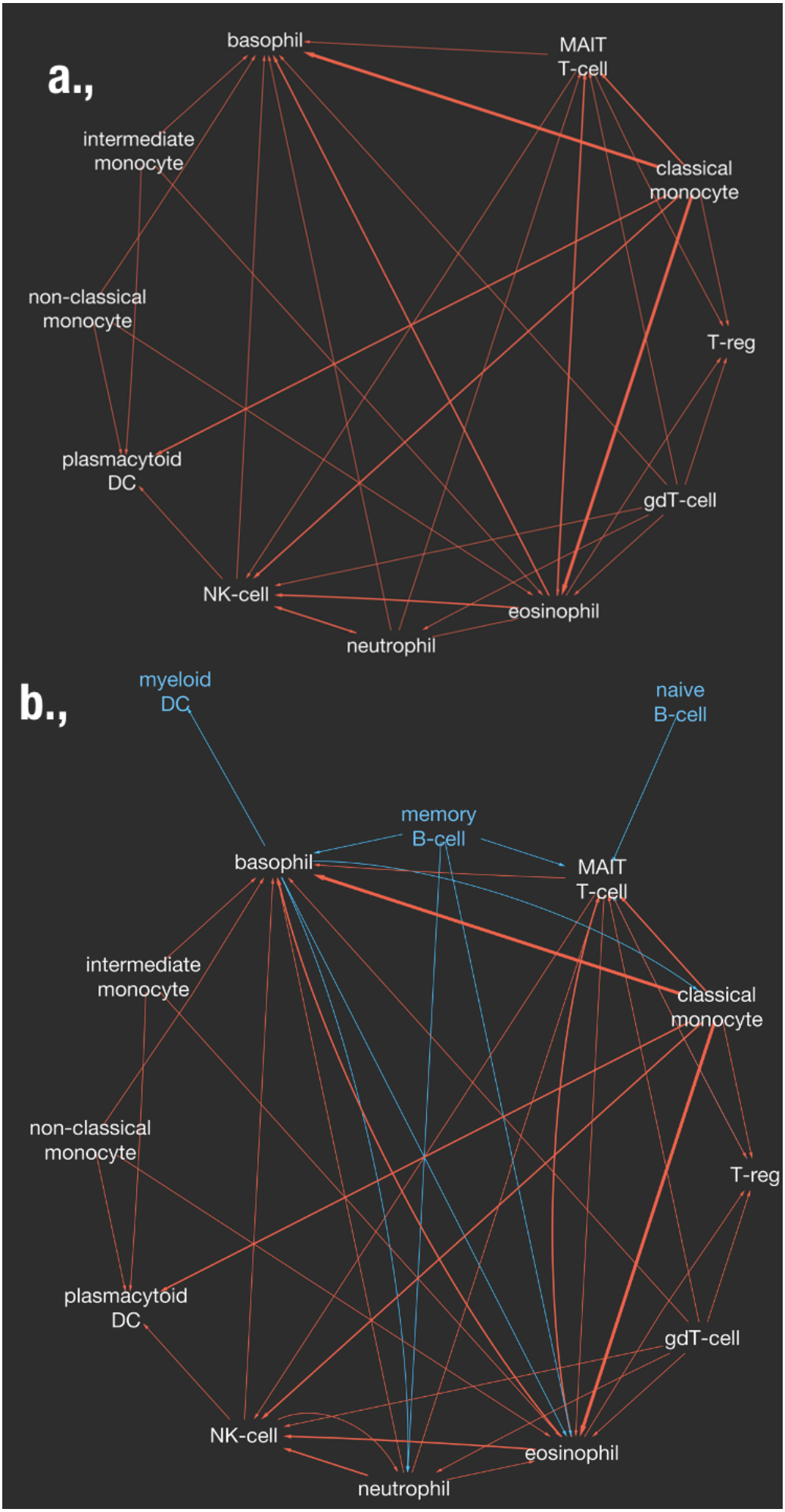
a.,
basophil
MAIT
T-cell
intermediate
monocyte
classical
monocyte
non-classical
monocyte
T-reg
plasmacytoid
DC
gdT-cell
NK-cell
eosinophil
neutrophil
b.,
myeloid
DC
naive
B-cell
memory
B-cell
basophil
MAIT
T-cell
intermediate
monocyte
classical
monocyte
non-classical
monocyte
T-reg
plasmacytoid
DC
gdT-cell
NK-cell
eosinophil
neutrophil

Figure 32. Cell-to-cell communication mediated by cytokines increased in COVID-19 patients. A: Interactions involving cytokines elevated in COVID-19 patients (red). B: Interactions involving elevated cytokines in COVID-19 patients (red), and interactions with cytokines that are missing following a SARS-CoV-2 infection, but are present in other CRS-causing viruses (blue). Edge width corresponds to the amount of cytokines mediating the specific interaction.

In part B of Figure 32, I attempted to model what the cell to cell communication would look like, if cytokines that are elevated in some other CRS-causing viruses were elevated. In other words, I "complemented" the system with these cytokines, to show what parts of the communication seem to be missing based on the literature data we curated. The complemented model shows three novel cell types: naïve-B-cells, memory B-cells, and the presence of myeloid DCs. The latter cell type is responsible for the secretion of multiple cytokines, including IL12, that also seems to be missing from the system, based on the literature curation analysis (see 5.4.1). The presence (or lack of) memory B-cells and naïve-B-cell can be explained by a variety of factors. An explanation arising from the bias of our methodology would be that much of the data we gathered from the curation contained samples from the time of hospital admission, where these responses could not form yet.

The CytokineLink based analysis gave a possible mechanistic link to the lack of one of the key anti-inflammatory cytokines. IL12 is missing from the system, potentially caused by an under-activation of myeloid DCs. This can be the result of the missing edge between myeloid DCs and basophils, an interaction mediated

by IL4. Although there are other cell types capable of producing IL12 in the model, its lack can be possibly caused by this missing cell type. The lack of IL12 can be seen in other patterns of the cytokine responses as well, as it is one of the main stimulators of IFN-γ (annotated as "mixed" in our curation) and the type II interferon response, which is the key in leading the innate immune response into the adaptive one, and eliminating the viral infection (Bhardwaj, Seder, Reddy, & Feldman, 1996; Kang et al., 2018; Lee & Ashkar, 2018).

4.5. Future research directions

COVID-19 research moves at an incredible speed, and seems to accelerate almost day by day. This makes following the literature an exciting and daunting task, and as such systematic reviews as the one we performed provide an important gap-filling function, summarising the state of the literature at the point of curation. The data and literature collected for them can be useful on its own, and can be further applied later down the line, systematically. I think systematic reviews like the one summarised in this chapter should be an iterative process, especially in a dynamically changing topic such as COVID-19 research. To enable this, we put together a robust semi-automated workflow, that allows periodic re-scanning of the literature, to fill gaps such as the ones seen in the hierarchical clustering figure in the first half of the chapter. Although there is always going to be a level of bias between the other compared viruses and SARS-CoV-2, caused by the sheer amount of literature being released on the latter, a functional comparison of these viruses can hopefully let us get a glimpse of the underlying pathomechanisms undiscovered so far.

The all-purpose version of CytokineLink, collated from Human Protein Atlas consensus expression data can give an overview of the cellular and molecular actors involved in certain infectious and inflammatory diseases. While it is well annotated, using two external systems immunology databases in the form of ImmunoGlobe and immuneXpresso, the trade-off is the relatively low level of resolution we can analyse the data under. Although this is appropriate for exploratory analyses and hypothesis generation, I think in the future, if applicable single-cell RNA-Seq data resources exist, a second round of CytokineLink analysis would provide much higher resolution results, especially combined with the integrated regulatory layer further informing on the expression status of the studied cytokines. Although I built the resource with the COVID-19 research and effort in mind, it can be utilised in all situations involving cytokines, such as autoimmune diseases, or even *Salmonella* infections. On the latter point, I think there exists a yet undiscovered niche of *Salmonella* research, wherein one could combine network data from both the host side (e.g. cytokine responses and their upstream signalling), a mechanistic intermediary layer established with tools such as MicrobioLink, and an intra-pathogen layer, such as the ones generated for SalmoNet 2 (Andrighetti, Bohar, Lemke, Sudhakar, & Korcsmaros, 2020). A complex model like this, generated from dual-RNA-Seq experiments for example, could give further insight into host-microbe interactions, by simultaneously uncovering the responses to the infection process, from both the host and the pathogen side, and potentially allow for previously unknown insight regarding the pathogenesis process and intervention therapies.

5. Final discussion

Network biology approaches are an appropriate tool to study infectious diseases and interactions of the host and the pathogen. By involving interactions between the genes or proteins of an organism, or between organisms, they provide a way to study these organisms on a systems level (Mulder, Akinola, Mazandu, & Rapanoel, 2014; Sudhakar et al., 2019).

Host adapted *Salmonella* serovars, or typhoidal *Salmonella* serovars when focusing on human disease, cause between 200,000 to 600,000 deaths every year (GBD 2017 Typhoid and Paratyphoid Collaborators 2019). Understanding how these invasive serovars form and behave is crucial to developing better intervention and surveillance strategies. In this thesis, I aimed to develop and update a network resources that enables us to study extraintestinal and gastrointestinal *Salmonella* serovars in this context.

Extraintestinal serovars are the products of convergent evolution, and are not monophyletically related to each other, in most cases their closest relative is a gastrointestinal serovar (Branchu et al., 2018). Because of this, many comparative genomic approaches have been applied to the problem, with great success: these led to the discovery of how *S.* Typhi and *S.* Paratyphi neutralise the phagocyte respiratory burst, how hundreds of genes, often belonging to the same functional categories degrade in these serovars, or how they collectively lose part of their ability to form biofilms (Hiyoshi et al. 2018; Nuccio and Bäumler 2014; MacKenzie et al. 2017).

Interaction information for non-model organisms are more difficult to acquire than better studied models. SalmoNet 2, described in Chapter 2, describes the need for molecular interaction network resources such as this, and the general logic of how they can be constructed for any organism of importance. The utilised data sources contain a lot of information for my Gram-negative bacterial species, and there exist Gram-positive alternatives for the involved layers (Sierro, Makita, de Hoon, & Nakai, 2008). The frameworks and workflows developed for SalmoNet 1 & SalmoNet 2 can help other scientific communities which lack integrated resources, and achieve the same goal of serving as a knowledge base for understudied organisms, and simultaneously drive research by predicting previously unknown interactions (Olbei et al., 2019).

The update to the database doubled the included serovars to cover more of the *Salmonella* genus and aimed to make the information content more precise than before. This was done through the involvement of novel data resources, such as strain specific metabolic models for all involved *Salmonella* strains, the usage of the IntAct scoring to involve high quality protein-protein interactions attained from multiple types of experiments, and the involvement of novel transcription binding sites used for genome wide scans. Through the inclusion of novel strains and data sources total number of interactions in the database grew from 81,514 to 270,196, more than tripling that of the first version of the SalmoNet database. Anecdotally, a second update in a database's lifecycle is an important steppingstone, signalling to the scientific community that there is still work going on, they can count on the data in it, and await further updates. This trust between user and developer is very important for a tool and resource like this to better integrate into the scientific community.

Through the comparison of differentially expressed genes following the knockout of infection relevant regulators and predicted regulatory targets of the same regulators in SalmoNet 2, I was able to assess the validity to my predicted regulatory interactions, insofar as they capture biologically relevant interactions for the majority of analysed global regulators (Colgan et al., 2016). In addition, the literature published since the generation of the interactions has independently confirmed, and added function to multiple predicted regulatory interactions in SalmoNet 2, such as the regulation of *sopB* through InvF, or the regulation of *ugtL* via SsrB, and its lack of regulation in *Salmonella bongori* (Choi & Groisman, 2020; Romero-González et al., 2020). However, future releases will have to strive for even greater precision regarding the quality of interactions, as the included number of included strains and genomes grows, and include other layers of information, such as posttranscriptional regulation mediated by small RNAs (Van Assche et al., 2015).

Integrating network information into comparative genomics studies can further highlight elements under selection and potentially even explain parts of the organism's behaviour. We showed an example of this, through a study involving the adaptive radiation of East African cichlid fish species (T. K. Mehta et al., 2021). The network rewiring analysis we applied highlighted the gene regulatory network rewiring of the visual opsin apparatus in a number of species, caused by a single nucleotide polymorphism in the 5` UTR transcription factor binding region of the genes in question. This finding fits into the hypothesis that the fish species adapted to different ecotypes and feeding behaviours utilise and require different parts of the visible light spectrum. Although the authors of the DyNet software used case-control experiments involving PPI networks and drug treated

cell line data as their use case in the original publication, we showed that the approach can be used to seek out and understand differences arising from the pressures of evolution (Goenawan et al., 2016). By adding regulatory interactions to the genes in the compared fish species, we were able to add functionality to the regulatory mutations observed (T. K. Mehta et al., 2021).

Applying the same rewiring approach to the interaction networks generated for SalmoNet 2 highlighted the effect genome degradation has on the interaction networks of typhoidal *Salmonella* strains, as this pseudogenization process was a large driving force behind the observed rewiring results of the most rewired genes between the pathovars, as described previously in the literature (Holt et al., 2009; Robert A Kingsley et al., 2013; Nuccio & Bäumler, 2014; Tanner & Kingsley, 2018). The large-scale comparison involving all serovars in SalmoNet 2 highlighted a pair of genes that seem to associate to invasive serovars. Due to the extraordinary challenges we face in our current times, we were not able to carry out our planned experiments to get a deeper understanding of the potential functions of these genes, but the computational approaches attempted to characterize the two proteins highlight how interaction information from SalmoNet 2 can be used for hypothesis generation, and how downstream analyses using the linked or external knowledgebases such as OMA or pubMLST can help in understanding and describing novel characteristics, such as the phylogenetic spread of the studied proteins (Altenhoff et al., 2018; Jolley et al., 2018).

The COVID-19 pandemic is still ongoing all over the planet. Although many countries have started vaccinating against the virus, hopefully indicating changes

to come, the past year has really highlighted the need to understand not just the virus itself, but how the immune system responds to it. In my 6-month redeployment I attempted to uncover parts of this mechanism, by comparing the cytokine responses from patients infected by functionally and phylogenetically related viruses. What we have found here is that there exists a group of cytokines that are not activating in response to a SARS-CoV-2 infection in the analysed datasets, but are present in other similar viral infections.

To give a mechanistic explanation as to why some cytokines activate, and why others do not, I built a novel network resource, called CytokineLink, aimed at highlighting how cells can communicate using cytokines as a medium, and vice versa, how cytokines can possibly affect each other's expression. I integrated other systems immunology databases into this network resource, giving more confidence to the individual links. When analysing the elevated and non-elevated cytokine levels from the previous study with CytokineLink, I identified a potential lacking cell signature from myeloid dendritic cells, that could explain why we noticed the differences when comparing the cytokine responses of viruses.

While these were interesting and novel research projects, adequate for a 6-month period, in the future I would like to see them iteratively refined, and scaled up, respectively. Systematic analyses and reviews of the literature of a fast-moving field such as COVID-19 research is important to sum up the current status of the field, and identify potential blind spots that need to be addressed.

Although the all-purpose version of CytokineLink is an adequate, high-level starting point for analyses involving cytokine responses, in my opinion the real

power of the approach would lie in applying it to large scale single-cell datasets, coming from infectious or inflammatory diseases, where both the regulatory layer, and the cytokine interactions could be analysed simultaneously. To make initial analyses easier, in the future I would like to prepare a simple web-based access point to the data as well, where interested researchers could quickly retrieve interactions certain cytokines or cells are involved in.

In this thesis I detailed the workflows and data resources I developed to create and compare molecular biological interaction networks, to answer specific biological questions. The methodologies I developed are, for the most part, agnostic of biological system, as shown by the publications I have co-authored involving other model (and non-model) organisms. The research presented in this thesis shows what gaps the pairwise or multiple comparison analysis of biological networks can fill.

In conclusion, this thesis has added to the understanding of *Salmonella* host adaptation, by generating multi-layered molecular interaction networks and a knowledgebase for multiple extra-and gastrointestinal *Salmonella* serovars important in human health. The methods I applied to the networks can be used in the future to identify rewiring hotspots in biological network comparisons, and the results obtained with them can lead to validatory analyses and future hypotheses. The results highlighted in this thesis lead to a framework that could be used to study not just host adaptation, but any other phenotypic split in a group of organisms in the future.

Bibliography

10K Salmonella Genomes Project. (2017). 10K Salmonella Genomes Project. Retrieved May 28, 2020, from http://10k-salmonella-genomes.com/update/2017/07/13/project-rationale.html

Ahmad, I., Lamprokostopoulou, A., Le Guyon, S., Streck, E., Barthel, M., Peters, V., … Römling, U. (2011). Complex c-di-GMP signaling networks mediate transition between virulence properties and biofilm formation in Salmonella enterica serovar Typhimurium. *Plos One*, *6*(12), e28351. https://doi.org/10.1371/journal.pone.0028351

Alexa, A., & Rahnenfuhrer, J. (2021). topGO: Enrichment Analysis for Gene Ontology. Retrieved October 28, 2021, from https://www.bioconductor.org/packages/release/bioc/html/topGO.html

Altenhoff, A. M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D. A., DeLuca, T., Forslund, K., … Dessimoz, C. (2016). Standardized benchmarking in the quest for orthologs. *Nature Methods*, *13*(5), 425–430. https://doi.org/10.1038/nmeth.3830

Altenhoff, A. M., Glover, N. M., Train, C.-M., Kaleb, K., Warwick Vesztrocy, A., Dylus, D., … Dessimoz, C. (2018). The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Research*, *46*(D1), D477–D485. https://doi.org/10.1093/nar/gkx1019

Altenhoff, A. M., Levy, J., Zarowiecki, M., Tomiczek, B., Warwick Vesztrocy, A., Dalquen, D. A., … Dessimoz, C. (2019). OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Research*, *29*(7), 1152–1163.

https://doi.org/10.1101/gr.243212.118

Altenhoff, A. M., Train, C.-M., Gilbert, K. J., Mediratta, I., Mendes de Farias, T., Moi, D., … Dessimoz, C. (2020). OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa1007

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nature Medicine*, *26*(4), 450–452. https://doi.org/10.1038/s41591-020-0820-9

Andrighetti, T., Bohar, B., Lemke, N., Sudhakar, P., & Korcsmaros, T. (2020). MicrobioLink: An Integrated Computational Pipeline to Infer Functional Effects of Microbiome-Host Interactions. *Cells*, *9*(5). https://doi.org/10.3390/cells9051278

Atallah, M. B., Tandon, V., Hiam, K. J., Boyce, H., Hori, M., Atallah, W., … Mallick, P. (2020). ImmunoGlobe: enabling systems immunology with a manually curated intercellular immune interaction network. *BMC Bioinformatics*, *21*(1), 346. https://doi.org/10.1186/s12859-020-03702-3

Bader, G. D., & Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, *4*, 2. https://doi.org/10.1186/1471-2105-4-2

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., … Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(Web Server issue), W202-8. https://doi.org/10.1093/nar/gkp335

Barabási, A.-L. (2016). *Network Science* (Illustrated, p. 475). Cambridge, United Kingdom: Cambridge University Press.

Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews. Genetics*, *12*(1), 56–68.

https://doi.org/10.1038/nrg2918

Bäumler, A., & Fang, F. C. (2013). Host specificity of bacterial pathogens. *Cold Spring Harbor Perspectives in Medicine*, *3*(12), a010041. https://doi.org/10.1101/cshperspect.a010041

Bäumler, A. J., Norris, T. L., Lasco, T., Voight, W., Reissbrodt, R., Rabsch, W., & Heffron, F. (1998). IroN, a novel outer membrane siderophore receptor characteristic of Salmonella enterica. *Journal of Bacteriology*, *180*(6), 1446–1453. https://doi.org/10.1128/JB.180.6.1446-1453.1998

Bäumler, A. J., Tsolis, R. M., Ficht, T. A., & Adams, L. G. (1998). Evolution of host adaptation in Salmonella enterica. *Infection and Immunity*, *66*(10), 4579–4587. https://doi.org/10.1128/IAI.66.10.4579-4587.1998

Bawn, M., Alikhan, N.-F., Thilliez, G., Kirkwood, M., Wheeler, N. E., Petrovska, L., … Kingsley, R. A. (2020). Evolution of Salmonella enterica serotype Typhimurium driven by anthropogenic selection and niche adaptation. *PLoS Genetics*, *16*(6), e1008850. https://doi.org/10.1371/journal.pgen.1008850

Baxter, M. A., Fahlen, T. F., Wilson, R. L., & Jones, B. D. (2003). HilE interacts with HilD and negatively regulates hilA transcription and expression of the Salmonella enterica serovar Typhimurium invasive phenotype. *Infection and Immunity*, *71*(3), 1295–1305. https://doi.org/10.1128/IAI.71.3.1295-1305.2003

Betakova, T., Kostrabova, A., Lachova, V., & Turianova, L. (2017). Cytokines induced during influenza virus infection. *Current Pharmaceutical Design*, *23*(18), 2616–2622. https://doi.org/10.2174/1381612823666170316123736

Bhardwaj, N., Seder, R. A., Reddy, A., & Feldman, M. V. (1996). IL-12 in conjunction with dendritic cells enhances antiviral CD8+ CTL responses in vitro. *The Journal of Clinical Investigation*, *98*(3), 715–722. https://doi.org/10.1172/JCI118843

Bouhaddou, M., Memon, D., Meyer, B., White, K. M., Rezelj, V. V., Correa Marrero, M., …

Krogan, N. J. (2020). The Global Phosphorylation Landscape of SARS-CoV-2 Infection. *Cell*, *182*(3), 685–712.e19. https://doi.org/10.1016/j.cell.2020.06.034

Bovolenta, L. A., Acencio, M. L., & Lemke, N. (2012). HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*, *13*, 405. https://doi.org/10.1186/1471-2164-13-405

Branchu, P., Bawn, M., & Kingsley, R. A. (2018). Genome Variation and Molecular Epidemiology of Salmonella enterica Serovar Typhimurium Pathovariants. *Infection and Immunity*, *86*(8). https://doi.org/10.1128/IAI.00079-18

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., … Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, *513*(7518), 375–381. https://doi.org/10.1038/nature13726

Brenner, F. W., Villar, R. G., Angulo, F. J., Tauxe, R., & Swaminathan, B. (2000). Salmonella nomenclature. *Journal of Clinical Microbiology*, *38*(7), 2465–2467. https://doi.org/10.1128/JCM.38.7.2465-2467.2000

Bronowski, C., Fookes, M. C., Gilderthorp, R., Ashelford, K. E., Harris, S. R., Phiri, A., … Winstanley, C. (2013). Genomic characterisation of invasive non-typhoidal Salmonella enterica Subspecies enterica Serovar Bovismorbificans isolates from Malawi. *PLoS Neglected Tropical Diseases*, *7*(11), e2557. https://doi.org/10.1371/journal.pntd.0002557

Browning, D. F., & Busby, S. J. (2004). The regulation of bacterial transcription initiation. *Nature Reviews. Microbiology*, *2*(1), 57–65. https://doi.org/10.1038/nrmicro787

Brückner, A., Polge, C., Lentze, N., Auerbach, D., & Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. *International Journal of Molecular Sciences*, *10*(6), 2763–2788. https://doi.org/10.3390/ijms10062763

Cameron, M. J., & Kelvin, D. J. (2013). Cytokines, Chemokines and Their Receptors - Madame

Curie Bioscience Database - NCBI Bookshelf.

Canals, R., Hammarlöf, D. L., Kröger, C., Owen, S. V., Fong, W. Y., Lacharme-Lora, L., … Hinton, J. C. D. (2019). Adding function to the genome of African Salmonella Typhimurium ST313 strain D23580. *PLoS Biology*, *17*(1), e3000059. https://doi.org/10.1371/journal.pbio.3000059

Carden, S. E., Walker, G. T., Honeycutt, J., Lugo, K., Pham, T., Jacobson, A., … Monack, D. (2017). Pseudogenization of the Secreted Effector Gene sseI Confers Rapid Systemic Dissemination of S. Typhimurium ST313 within Migratory Dendritic Cells. *Cell Host & Microbe*, *21*(2), 182–194. https://doi.org/10.1016/j.chom.2017.01.009

Carleton, K. (2009). Cichlid fish visual systems: mechanisms of spectral tuning. *Integrative Zoology*, *4*(1), 75–86. https://doi.org/10.1111/j.1749-4877.2008.00137.x

Caspi, R., Billington, R., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., … Karp, P. D. (2019). BioCyc: A Genomic and Metabolic Web Portal with Multiple Omics Analytical Tools. *The FASEB Journal*.

Channappanavar, R., Fehr, A. R., Vijay, R., Mack, M., Zhao, J., Meyerholz, D. K., & Perlman, S. (2016). Dysregulated Type I Interferon and Inflammatory Monocyte-Macrophage Responses Cause Lethal Pneumonia in SARS-CoV-Infected Mice. *Cell Host & Microbe*, *19*(2), 181–193. https://doi.org/10.1016/j.chom.2016.01.007

Channappanavar, R., Fehr, A. R., Zheng, J., Wohlford-Lenane, C., Abrahante, J. E., Mack, M., … Perlman, S. (2019). IFN-I response timing relative to virus replication determines MERS coronavirus infection outcomes. *The Journal of Clinical Investigation*.

Chen, H., & Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, *5*, 147. https://doi.org/10.1186/1471-2105-5-147

Choi, J., & Groisman, E. A. (2020). Horizontally acquired regulatory gene activates ancestral

regulatory system to promote Salmonella virulence. *Nucleic Acids Research*, *48*(19), 10832–10847. https://doi.org/10.1093/nar/gkaa813

Coburn, B., Grassl, G. A., & Finlay, B. B. (2007). Salmonella, the host and disease: a brief review. *Immunology and Cell Biology*, *85*(2), 112–118. https://doi.org/10.1038/sj.icb.7100007

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., … Barrell, B. G. (2001). Massive gene decay in the leprosy bacillus. *Nature*, *409*(6823), 1007–1011. https://doi.org/10.1038/35059006

Colgan, A. M., Kröger, C., Diard, M., Hardt, W.-D., Puente, J. L., Sivasankaran, S. K., … Hinton, J. C. D. (2016). The Impact of 18 Ancestral and Horizontally-Acquired Regulatory Proteins upon the Transcriptome and sRNA Landscape of Salmonella enterica serovar Typhimurium. *PLoS Genetics*, *12*(8), e1006258. https://doi.org/10.1371/journal.pgen.1006258

Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*, *5*(4), 536–544. https://doi.org/10.1038/s41564-020-0695-z

Costela-Ruiz, V. J., Illescas-Montes, R., Puerta-Puerta, J. M., Ruiz, C., & Melguizo-Rodríguez, L. (2020). SARS-CoV-2 infection: The role of cytokines in COVID-19 disease. *Cytokine & Growth Factor Reviews*, *54*, 62–75. https://doi.org/10.1016/j.cytogfr.2020.06.001

Csabai, L., Ölbei, M., Budd, A., Korcsmáros, T., & Fazekas, D. (2018). Signalink: multilayered regulatory networks. *Methods in Molecular Biology*, *1819*, 53–73. https://doi.org/10.1007/978-1-4939-8618-7_3

de Jong, H. K., Parry, C. M., van der Poll, T., & Wiersinga, W. J. (2012). Host-pathogen interaction in invasive Salmonellosis. *PLoS Pathogens*, *8*(10), e1002933. https://doi.org/10.1371/journal.ppat.1002933

De Las Rivas, J., & Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Computational Biology*, *6*(6), e1000807. https://doi.org/10.1371/journal.pcbi.1000807

del-Toro, N., Dumousseau, M., Orchard, S., Jimenez, R. C., Galeota, E., Launay, G., … Hermjakob, H. (2013). A new reference implementation of the PSICQUIC web service. *Nucleic Acids Research*, *41*(Web Server issue), W601-6. https://doi.org/10.1093/nar/gkt392

Del Valle, D. M., Kim-Schulze, S., Hsin-Hui, H., Beckmann, N. D., Nirenberg, S., Wang, B., … Gnjatic, S. (2020). An inflammatory cytokine signature helps predict COVID-19 severity and death. *MedRxiv*. https://doi.org/10.1101/2020.05.28.20115758

Demeter, A., Romero-Mulero, M. C., Csabai, L., Ölbei, M., Sudhakar, P., Haerty, W., & Korcsmáros, T. (2020). ULK1 and ULK2 are less redundant than previously thought: computational analysis uncovers distinct regulation and functions of these autophagy induction proteins. *Scientific Reports*, *10*(1), 10940. https://doi.org/10.1038/s41598-020-67780-2

den Bakker, H. C., Moreno Switt, A. I., Govoni, G., Cummings, C. A., Ranieri, M. L., Degoricija, L., … Wiedmann, M. (2011). Genome sequencing reveals diversification of virulence factor content and possible host adaptation in distinct subpopulations of Salmonella enterica. *BMC Genomics*, *12*, 425. https://doi.org/10.1186/1471-2164-12-

425

Desai, P. T., Porwollik, S., Long, F., Cheng, P., Wollam, A., Bhonagiri-Palsikar, V., … McClelland, M. (2013). Evolutionary Genomics of Salmonella enterica Subspecies. *MBio*, *4*(2). https://doi.org/10.1128/mBio.00579-12

Doolittle, R. F., Feng, D. F., Tsang, S., Cho, G., & Little, E. (1996). Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science*, *271*(5248), 470–477. https://doi.org/10.1126/science.271.5248.470

Dougan, G., & Baker, S. (2014). Salmonella enterica serovar Typhi and the pathogenesis of typhoid fever. *Annual Review of Microbiology*, *68*, 317–336. https://doi.org/10.1146/annurev-micro-091313-103739

Dunne, J. A., Williams, R. J., & Martinez, N. D. (2002). Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(20), 12917–12922. https://doi.org/10.1073/pnas.192407699

Eberth, K. J. (1880). Die Organismen in den Organen bei Typhus abdominalis. *Archiv Für Pathologische Anatomie Und Physiologie*, 58–74.

Eckerle, I., Zimmermann, S., Kapaun, A., & Junghanss, T. (2010). Salmonella enterica serovar Virchow bacteremia presenting as typhoid-like illness in an immunocompetent patient. *Journal of Clinical Microbiology*, *48*(7), 2643–2644. https://doi.org/10.1128/JCM.00217-10

Edwards, J. S., & Palsson, B. O. (1999). Systems properties of the Haemophilus influenzae Rd metabolic genotype. *The Journal of Biological Chemistry*, *274*(25), 17410–17416. https://doi.org/10.1074/jbc.274.25.17410

Ehrbar, K., & Hardt, W.-D. (2005). Bacteriophage-encoded type III effectors in Salmonella enterica subspecies 1 serovar Typhimurium. *Infection, Genetics and Evolution*, *5*(1), 1–

9. https://doi.org/10.1016/j.meegid.2004.07.004

El Mouali, Y., Gaviria-Cantin, T., Sánchez-Romero, M. A., Gibert, M., Westermann, A. J., Vogel, J., & Balsalobre, C. (2018). CRP-cAMP mediates silencing of Salmonella virulence at the post-transcriptional level. *PLoS Genetics*, *14*(6), e1007401. https://doi.org/10.1371/journal.pgen.1007401

Eng, S.-K., Pusparajah, P., Ab Mutalib, N.-S., Ser, H.-L., Chan, K.-G., & Lee, L.-H. (2015). *Salmonella* : A review on pathogenesis, epidemiology and antibiotic resistance. *Frontiers in Life Science*, *8*(3), 284–293. https://doi.org/10.1080/21553769.2015.1051243

Fang, F. C., Libby, S. J., Buchmeier, N. A., Loewen, P. C., Switala, J., Harwood, J., & Guiney, D. G. (1992). The alternative sigma factor katF (rpoS) regulates Salmonella virulence. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(24), 11978–11982. https://doi.org/10.1073/pnas.89.24.11978

Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálfy, M., Dúl, Z., … Korcsmáros, T. (2013). SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Systems Biology*, *7*, 7. https://doi.org/10.1186/1752-0509-7-7

Feasey, N. A., Dougan, G., Kingsley, R. A., Heyderman, R. S., & Gordon, M. A. (2012). Invasive non-typhoidal salmonella disease: an emerging and neglected tropical disease in Africa. *The Lancet*, *379*(9835), 2489–2499. https://doi.org/10.1016/S0140-6736(11)61752-2

Feasey, N. A., Hadfield, J., Keddy, K. H., Dallman, T. J., Jacobs, J., Deng, X., … Thomson, N. R. (2016). Distinct Salmonella Enteritidis lineages associated with enterocolitis in high-income settings and invasive disease in low-income settings. *Nature Genetics*, *48*(10), 1211–1217. https://doi.org/10.1038/ng.3644

Fookes, M., Schroeder, G. N., Langridge, G. C., Blondel, C. J., Mammina, C., Connor, T. R.,

… Thomson, N. R. (2011). Salmonella bongori provides insights into the evolution of the Salmonellae. *PLoS Pathogens*, *7*(8), e1002191. https://doi.org/10.1371/journal.ppat.1002191

Furey, T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews. Genetics*, *13*(12), 840–852. https://doi.org/10.1038/nrg3306

Gaffky, G. T. A. (1884). Zur Aetiology des Abdominaltyphus. In *Mittheilungen aus dem Kaiserlichen Gesundheitsamte* (Vol. 2, pp. 372–420).

Galán, J. E. (2016). Typhoid toxin provides a window into typhoid fever and the biology of Salmonella Typhi. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(23), 6338–6344. https://doi.org/10.1073/pnas.1606335113

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., … Collado-Vides, J. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, *44*(D1), D133-43. https://doi.org/10.1093/nar/gkv1156

Garai, P., Gnanadhas, D. P., & Chakravortty, D. (2012). Salmonella enterica serovars Typhimurium and Typhi as model organisms: revealing paradigm of host-pathogen interactions. *Virulence*, *3*(4), 377–388. https://doi.org/10.4161/viru.21087

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., … Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, *415*(6868), 141–147. https://doi.org/10.1038/415141a

GBD 2017 Typhoid and Paratyphoid Collaborators. (2019). The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet Infectious Diseases*, *19*(4), 369–381. https://doi.org/10.1016/S1473-3099(18)30685-6

Giannella, R. A. (1996). Salmonella. In S. Baron (Ed.), *Medical Microbiology* (4th ed.). Galveston (TX): University of Texas Medical Branch at Galveston.

Gilchrist, J. J., & MacLennan, C. A. (2019). Invasive nontyphoidal salmonella disease in africa. *EcoSal Plus*, *8*(2). https://doi.org/10.1128/ecosalplus.ESP-0007-2018

Goenawan, I. H., Bryan, K., & Lynn, D. J. (2016). DyNet: visualization and analysis of dynamic molecular interaction networks. *Bioinformatics*, *32*(17), 2713–2715. https://doi.org/10.1093/bioinformatics/btw187

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., … Krogan, N. J. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, *583*(7816), 459–468. https://doi.org/10.1038/s41586-020-2286-9

Gordon, M. A. (2011). Invasive nontyphoidal Salmonella disease: epidemiology, pathogenesis and diagnosis. *Current Opinion in Infectious Diseases*, *24*(5), 484–489. https://doi.org/10.1097/QCO.0b013e32834a9980

Gori, M., Ebranati, E., Scaltriti, E., Huedo, P., Ciceri, G., Tanzi, E., … Bolzoni, L. (2018). High-resolution diffusion pattern of human infections by Salmonella enterica serovar Napoli in Northern Italy explained through phylogeography. *Plos One*, *13*(8), e0202573. https://doi.org/10.1371/journal.pone.0202573

Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biology*, *20*(1), 121. https://doi.org/10.1186/s13059-019-1730-3

Guan, W.-J., Ni, Z.-Y., Hu, Y., Liang, W.-H., Ou, C.-Q., He, J.-X., … China Medical Treatment Expert Group for Covid-19. (2020). Clinical characteristics of coronavirus disease 2019 in china. *The New England Journal of Medicine*, *382*(18), 1708–1720. https://doi.org/10.1056/NEJMoa2002032

Gully, D., Moinier, D., Loiseau, L., & Bouveret, E. (2003). New partners of acyl carrier protein

detected in Escherichia coli by tandem affinity purification. *FEBS Letters*, *548*(1–3), 90–96. https://doi.org/10.1016/s0014-5793(03)00746-4

Guo, X.-Z. J., & Thomas, P. G. (2017). New fronts emerge in the influenza cytokine storm. *Seminars in Immunopathology*, *39*(5), 541–550. https://doi.org/10.1007/s00281-017-0636-y

Guo, Y., Luan, L., Patil, N. K., & Sherwood, E. R. (2017). Immunobiology of the IL-15/IL-15Rα complex as an antitumor and antiviral agent. *Cytokine & Growth Factor Reviews*, *38*, 10–21. https://doi.org/10.1016/j.cytogfr.2017.08.002

Gut, A. M., Vasiljevic, T., Yeager, T., & Donkor, O. N. (2018). Salmonella infection - prevention and treatment by antibiotics and probiotic yeasts: a review. *Microbiology*, *164*(11), 1327–1344. https://doi.org/10.1099/mic.0.000709

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., … Lee, I. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, *46*(D1), D380–D386. https://doi.org/10.1093/nar/gkx1013

Hardy, A. (2004). Salmonella: a continuing problem. *Postgraduate Medical Journal*, *80*(947), 541–545. https://doi.org/10.1136/pgmj.2003.016584

Harrison, A. G., Lin, T., & Wang, P. (2020). Mechanisms of SARS-CoV-2 Transmission and Pathogenesis. *Trends in Immunology*, *41*(12), 1100–1115. https://doi.org/10.1016/j.it.2020.10.004

Hawkey, J., Monk, J. M., Billman-Jacobe, H., Palsson, B., & Holt, K. E. (2020). Impact of insertion sequences on convergent evolution of Shigella species. *PLoS Genetics*, *16*(7), e1008931. https://doi.org/10.1371/journal.pgen.1008931

Hengge, R. (2009). Principles of c-di-GMP signalling in bacteria. *Nature Reviews. Microbiology*, *7*(4), 263–273. https://doi.org/10.1038/nrmicro2109

Hinks, T. S. C., van Wilgenburg, B., Wang, H., Loh, L., Koutsakos, M., Kedzierska, K., … Chen, Z. (2020). Study of MAIT cell activation in viral infections in vivo. *Methods in Molecular Biology*, *2098*, 261–281. https://doi.org/10.1007/978-1-0716-0207-2_17

Hiyoshi, H., Wangdi, T., Lock, G., Saechao, C., Raffatellu, M., Cobb, B. A., & Bäumler, A. J. (2018). Mechanisms to evade the phagocyte respiratory burst arose by convergent evolution in typhoidal salmonella serovars. *Cell Reports*, *22*(7), 1787–1797. https://doi.org/10.1016/j.celrep.2018.01.016

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S.-L., … Tyers, M. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, *415*(6868), 180–183. https://doi.org/10.1038/415180a

Hofer, U. (2019). The cost of antimicrobial resistance. *Nature Reviews. Microbiology*, *17*(1), 3. https://doi.org/10.1038/s41579-018-0125-x

Hoffmann, R., & Valencia, A. (2005). Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, *21 Suppl 2*, ii252-8. https://doi.org/10.1093/bioinformatics/bti1142

Hohmann, E. L. (2001). Nontyphoidal salmonellosis. *Clinical Infectious Diseases*, *32*(2), 263–269. https://doi.org/10.1086/318457

Holden, E. R., & Webber, M. A. (2020). Mara, rama, and soxs as mediators of the stress response: survival at a cost. *Frontiers in Microbiology*, *11*, 828. https://doi.org/10.3389/fmicb.2020.00828

Holt, K. E., Thomson, N. R., Wain, J., Langridge, G. C., Hasan, R., Bhutta, Z. A., … Parkhill, J. (2009). Pseudogene accumulation in the evolutionary histories of Salmonella enterica serovars Paratyphi A and Typhi. *BMC Genomics*, *10*, 36. https://doi.org/10.1186/1471-2164-10-36

Holtmeier, W., & Kabelitz, D. (2005). gammadelta T cells link innate and adaptive immune

responses. *Chemical Immunology and Allergy*, *86*, 151–183. https://doi.org/10.1159/000086659

Hounmanou, Y. M. G., Dalsgaard, A., Sopacua, T. F., Uddin, G. M. N., Leekitcharoenphon, P., Hendriksen, R. S., … Larsen, M. H. (2020). Molecular characteristics and zoonotic potential of salmonella weltevreden from cultured shrimp and tilapia in vietnam and china. *Frontiers in Microbiology*, *11*, 1985. https://doi.org/10.3389/fmicb.2020.01985

Huedo, P., Gori, M., Zolin, A., Amato, E., Ciceri, G., Bossi, A., & Pontello, M. (2017). Salmonella enterica Serotype Napoli is the First Cause of Invasive Nontyphoidal Salmonellosis in Lombardy, Italy (2010-2014), and Belongs to Typhi Subclade. *Foodborne Pathogens and Disease*, *14*(3), 148–151. https://doi.org/10.1089/fpd.2016.2206

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754–755. https://doi.org/10.1093/bioinformatics/17.8.754

Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., & Bulyk, M. L. (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, *43*(Database issue), D117-22. https://doi.org/10.1093/nar/gku1045

Humphreys, S., Stevenson, A., Bacon, A., Weinhardt, A. B., & Roberts, M. (1999). The alternative sigma factor, sigmaE, is critically important for the virulence of Salmonella typhimurium. *Infection and Immunity*, *67*(4), 1560–1568. https://doi.org/10.1128/IAI.67.4.1560-1568.1999

Ibrahim, G. M., & Morin, P. M. (2018). Salmonella serotyping using whole genome sequencing. *Frontiers in Microbiology*, *9*, 2993. https://doi.org/10.3389/fmicb.2018.02993

Ichihashi, Y., Aguilar-Martínez, J. A., Farhi, M., Chitwood, D. H., Kumar, R., Millon, L. V.,

… Sinha, N. R. (2014). Evolutionary developmental transcriptomics reveals a gene network module regulating interspecific diversity in plant leaf shape. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(25), E2616-21. https://doi.org/10.1073/pnas.1402835111

Ilyas, B., Tsai, C. N., & Coombes, B. K. (2017). Evolution of Salmonella-Host Cell Interactions through a Dynamic Bacterial Genome. *Frontiers in Cellular and Infection Microbiology*, *7*, 428. https://doi.org/10.3389/fcimb.2017.00428

Jenal, U., Reinders, A., & Lori, C. (2017). Cyclic di-GMP: second messenger extraordinaire. *Nature Reviews. Microbiology*, *15*(5), 271–284. https://doi.org/10.1038/nrmicro.2016.190

Jolley, K. A., Bray, J. E., & Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. [version 1; peer review: 2 approved]. *Wellcome Open Research*, *3*, 124. https://doi.org/10.12688/wellcomeopenres.14826.1

Judicial Commission of the International Committee on Systematics of Prokaryotes. (2005). The type species of the genus Salmonella Lignieres 1900 is Salmonella enterica (ex Kauffmann and Edwards 1952) Le Minor and Popoff 1987, with the type strain LT2T, and conservation of the epithet enterica in Salmonella enterica over all earlier epithets that may be applied to this species. Opinion 80. *International Journal of Systematic and Evolutionary Microbiology*, *55*(Pt 1), 519–520. https://doi.org/10.1099/ijs.0.63579-0

Jung, S., Potapov, I., Chillara, S., & Del Sol, A. (2021). Leveraging systems biology for predicting modulators of inflammation in patients with COVID-19. *Science Advances*, *7*(6). https://doi.org/10.1126/sciadv.abe5735

Kaleb, K., Warwick Vesztrocy, A., Altenhoff, A., & Dessimoz, C. (2019). Expanding the Orthologous Matrix (OMA) programmatic interfaces: REST API and the *OmaDB*

packages for R and Python [version 2; peer review: 2 approved]. *F1000Research*, *8*, 42. https://doi.org/10.12688/f1000research.17548.2

Kang, S., Brown, H. M., & Hwang, S. (2018). Direct Antiviral Mechanisms of Interferon-Gamma. *Immune Network*, *18*(5), e33. https://doi.org/10.4110/in.2018.18.e33

Kerrien, S, Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., … Hermjakob, H. (2007). IntAct--open source resource for molecular interaction data. *Nucleic Acids Research*, *35*(Database issue), D561-5. https://doi.org/10.1093/nar/gkl958

Kerrien, Samuel, Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod, N., … Hermjakob, H. (2007). Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biology*, *5*, 44. https://doi.org/10.1186/1741-7007-5-44

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., … Mathelier, A. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, *46*(D1), D260–D266. https://doi.org/10.1093/nar/gkx1126

Kim, E. D., Sabharwal, A., Vetta, A. R., & Blanchette, M. (2010). Predicting direct protein interactions from affinity purification mass spectrometry data. *Algorithms for Molecular Biology*, *5*, 34. https://doi.org/10.1186/1748-7188-5-34

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., … Lewis, N. E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, *44*(D1), D515-22. https://doi.org/10.1093/nar/gkv1049

Kingsley, R A, & Bäumler, A. J. (2000). Host adaptation and the emergence of infectious disease: the Salmonella paradigm. *Molecular Microbiology*, *36*(5), 1006–1014. https://doi.org/10.1046/j.1365-2958.2000.01907.x

Kingsley, Robert A, Kay, S., Connor, T., Barquist, L., Sait, L., Holt, K. E., … Dougan, G.

(2013). Genome and transcriptome adaptation accompanying emergence of the definitive type 2 host-restricted Salmonella enterica serovar Typhimurium pathovar. *MBio*, *4*(5), e00565-13. https://doi.org/10.1128/mBio.00565-13

Kingsley, Robert A, Msefula, C. L., Thomson, N. R., Kariuki, S., Holt, K. E., Gordon, M. A., … Dougan, G. (2009). Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. *Genome Research*, *19*(12), 2279–2287. https://doi.org/10.1101/gr.091017.109

Kılıç, S., Sagitova, D. M., Wolfish, S., Bely, B., Courtot, M., Ciufo, S., … Erill, I. (2016). From data repositories to submission portals: rethinking the role of domain-specific databases in CollecTF. *Database: The Journal of Biological Databases and Curation*, *2016*. https://doi.org/10.1093/database/baw055

Klamt, S., Haus, U.-U., & Theis, F. (2009). Hypergraphs and cellular networks. *PLoS Computational Biology*, *5*(5), e1000385. https://doi.org/10.1371/journal.pcbi.1000385

Klemm, E. J., Gkrania-Klotsas, E., Hadfield, J., Forbester, J. L., Harris, S. R., Hale, C., … Kingsley, R. A. (2016). Emergence of host-adapted Salmonella Enteritidis through rapid evolution in an immunocompromised host. *Nature Microbiology*, *1*(3). https://doi.org/10.1038/nmicrobiol.2015.23

Kotlyar, M., Rossos, A. E. M., & Jurisica, I. (2017). Prediction of Protein-Protein Interactions. *Current Protocols in Bioinformatics*, *60*, 8.2.1-8.2.14. https://doi.org/10.1002/cpbi.38

Kreimer, A., Borenstein, E., Gophna, U., & Ruppin, E. (2008). The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(19), 6976–6981. https://doi.org/10.1073/pnas.0712149105

Kröger, C., Colgan, A., Srikumar, S., Händler, K., Sivasankaran, S. K., Hammarlöf, D. L., … Hinton, J. C. D. (2013). An infection-relevant transcriptomic compendium for

Salmonella enterica Serovar Typhimurium. *Cell Host & Microbe*, *14*(6), 683–695. https://doi.org/10.1016/j.chom.2013.11.010

Kröger, C., Dillon, S. C., Cameron, A. D. S., Papenfort, K., Sivasankaran, S. K., Hokamp, K., … Hinton, J. C. D. (2012). The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(20), E1277-86. https://doi.org/10.1073/pnas.1201061109

Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., & Makeev, V. J. (2013). HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Research*, *41*(Database issue), D195-202. https://doi.org/10.1093/nar/gks1089

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Kuzmanov, U., & Emili, A. (2013). Protein-protein interaction networks: probing disease mechanisms using model systems. *Genome Medicine*, *5*(4), 37. https://doi.org/10.1186/gm441

Kveler, K., Starosvetsky, E., Ziv-Kenet, A., Kalugny, Y., Gorelik, Y., Shalev-Malul, G., … Shen-Orr, S. S. (2018). Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed. *Nature Biotechnology*, *36*(7), 651–659. https://doi.org/10.1038/nbt.4152

Lamers, M. M., Beumer, J., van der Vaart, J., Knoops, K., Puschhof, J., Breugem, T. I., … Clevers, H. (2020). SARS-CoV-2 productively infects human gut enterocytes. *Science*, *369*(6499), 50–54. https://doi.org/10.1126/science.abc1669

Langridge, G. C., Fookes, M., Connor, T. R., Feltwell, T., Feasey, N., Parsons, B. N., …

Thomson, N. R. (2015). Patterns of genome evolution that have accompanied host adaptation in Salmonella. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(3), 863–868. https://doi.org/10.1073/pnas.1416707112

Langridge, G. C., Nair, S., & Wain, J. (2009). Nontyphoidal*salmonella* serovars cause different degrees of invasive disease globally. *The Journal of Infectious Diseases*, *199*(4), 602–603. https://doi.org/10.1086/596208

Lee, A. J., & Ashkar, A. A. (2018). The dual nature of type I and type II interferons. *Frontiers in Immunology*, *9*, 2061. https://doi.org/10.3389/fimmu.2018.02061

Liang, Y., Gao, Z., Dong, Y., & Liu, Q. (2014). Structural and functional analysis show that the Escherichia coli uncharacterized protein YjcS is likely an alkylsulfatase. *Protein Science*, *23*(10), 1442–1450. https://doi.org/10.1002/pro.2528

Lim, D., Kim, K., Song, M., Jeong, J.-H., Chang, J. H., Kim, S. R., … Shin, M. (2020). Transcriptional regulation of Salmochelin glucosyltransferase by Fur in Salmonella. *Biochemical and Biophysical Research Communications*, *529*(1), 70–76. https://doi.org/10.1016/j.bbrc.2020.06.009

López-Garrido, J., & Casadesús, J. (2010). Regulation of Salmonella enterica pathogenicity island 1 by DNA adenine methylation. *Genetics*, *184*(3), 637–649. https://doi.org/10.1534/genetics.109.108985

Lou, L., Zhang, P., Piao, R., & Wang, Y. (2019). Salmonella Pathogenicity Island 1 (SPI-1) and Its Complex Regulatory Network. *Frontiers in Cellular and Infection Microbiology*, *9*, 270. https://doi.org/10.3389/fcimb.2019.00270

Lynch, V. J., Leclerc, R. D., May, G., & Wagner, G. P. (2011). Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics*, *43*(11), 1154–1159. https://doi.org/10.1038/ng.917

Macario, A. J. L., Grippo, T. M., & de Macario, E. C. (2005). Genetic disorders involving

molecular-chaperone genes: A perspective. *Genetics in Medicine*, *7*(1), 3–12. https://doi.org/10.1097/01.GIM.0000151351.11876.C3

Macek, B., Forchhammer, K., Hardouin, J., Weber-Ban, E., Grangeasse, C., & Mijakovic, I. (2019). Protein post-translational modifications in bacteria. *Nature Reviews. Microbiology*, *17*(11), 651–664. https://doi.org/10.1038/s41579-019-0243-0

Machhi, J., Herskovitz, J., Senan, A. M., Dutta, D., Nath, B., Oleynikov, M. D., … Kevadiya, B. D. (2020). The Natural History, Pathobiology, and Clinical Manifestations of SARS-CoV-2 Infections. *Journal of Neuroimmune Pharmacology*, *15*(3), 359–386. https://doi.org/10.1007/s11481-020-09944-5

MacKenzie, K. D., Palmer, M. B., Köster, W. L., & White, A. P. (2017). Examining the Link between Biofilm Formation and the Ability of Pathogenic Salmonella Strains to Colonize Multiple Host Species. *Frontiers in Veterinary Science*, *4*, 138. https://doi.org/10.3389/fvets.2017.00138

MacKenzie, K. D., Wang, Y., Musicha, P., Hansen, E. G., Palmer, M. B., Herman, D. J., … White, A. P. (2019). Parallel evolution leading to impaired biofilm formation in invasive Salmonella strains. *PLoS Genetics*, *15*(6), e1008233. https://doi.org/10.1371/journal.pgen.1008233

Majowicz, S. E., Musto, J., Scallan, E., Angulo, F. J., Kirk, M., O'Brien, S. J., … International Collaboration on Enteric Disease "Burden of Illness" Studies. (2010). The global burden of nontyphoidal Salmonella gastroenteritis. *Clinical Infectious Diseases*, *50*(6), 882–889. https://doi.org/10.1086/650733

Makendi, C., Page, A. J., Wren, B. W., Le Thi Phuong, T., Clare, S., Hale, C., … Dougan, G. (2016). A Phylogenetic and Phenotypic Analysis of Salmonella enterica Serovar Weltevreden, an Emerging Agent of Diarrheal Disease in Tropical Regions. *PLoS Neglected Tropical Diseases*, *10*(2), e0004446.

https://doi.org/10.1371/journal.pntd.0004446

Makris, S., Paulsen, M., & Johansson, C. (2017). Type I interferons as regulators of lung inflammation. *Frontiers in Immunology*, *8*, 259. https://doi.org/10.3389/fimmu.2017.00259

Malik-Sheriff, R. S., Glont, M., Nguyen, T. V. N., Tiwari, K., Roberts, M. G., Xavier, A., … Hermjakob, H. (2020). BioModels-15 years of sharing computational models in life science. *Nucleic Acids Research*, *48*(D1), D407–D415. https://doi.org/10.1093/nar/gkz1055

Mani, V., Brennand, J., & Mandal, B. K. (1974). Invasive illness with Salmonella virchow infection. *British Medical Journal*, *2*(5911), 143–144. https://doi.org/10.1136/bmj.2.5911.143

Mao, F., Dam, P., Chou, J., Olman, V., & Xu, Y. (2009). DOOR: a database for prokaryotic operons. *Nucleic Acids Research*, *37*(Database issue), D459-63. https://doi.org/10.1093/nar/gkn757

Maple, J., & Møller, S. G. (2007). Yeast two-hybrid screening. *Methods in Molecular Biology*, *362*, 207–223. https://doi.org/10.1007/978-1-59745-257-1_15

Mayuzumi, H., Inagaki-Ohara, K., Uyttenhove, C., Okamoto, Y., & Matsuzaki, G. (2010). Interleukin-17A is required to suppress invasion of Salmonella enterica serovar Typhimurium to enteric mucosa. *Immunology*, *131*(3), 377–385. https://doi.org/10.1111/j.1365-2567.2010.03310.x

McClelland, M., Sanderson, K. E., Clifton, S. W., Latreille, P., Porwollik, S., Sabo, A., … Wilson, R. K. (2004). Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid. *Nature Genetics*, *36*(12), 1268–1274. https://doi.org/10.1038/ng1470

Medina-Rivera, A., Abreu-Goodger, C., Thomas-Chollier, M., Salgado, H., Collado-Vides, J.,

& van Helden, J. (2011). Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Research*, *39*(3), 808–824. https://doi.org/10.1093/nar/gkq710

Mehta, P., McAuley, D. F., Brown, M., Sanchez, E., Tattersall, R. S., Manson, J. J., & HLH Across Speciality Collaboration, UK. (2020). COVID-19: consider cytokine storm syndromes and immunosuppression. *The Lancet*, *395*(10229), 1033–1034. https://doi.org/10.1016/S0140-6736(20)30628-0

Mehta, T. K., Koch, C., Nash, W., Knaack, S. A., Sudhakar, P., Olbei, M., … Di-Palma, F. (2021). Evolution of regulatory networks associated with traits under selection in cichlids. *Genome Biology*, *22*(1), 25. https://doi.org/10.1186/s13059-020-02208-8

Meldal, B. H. M., Forner-Martinez, O., Costanzo, M. C., Dana, J., Demeter, J., Dumousseau, M., … Orchard, S. (2015). The complex portal--an encyclopaedia of macromolecular complexes. *Nucleic Acids Research*, *43*(Database issue), D479-84. https://doi.org/10.1093/nar/gku975

Merali, Z., & Giles, J. (2005). Databases in peril. *Nature*, *435*(7045), 1010–1011. https://doi.org/10.1038/4351010a

Messer, R. D., Warnock, T. H., Heazlewood, R. J., & Hanna, J. N. (1997). Salmonella meningitis in children in far north Queensland. *Journal of Paediatrics and Child Health*, *33*(6), 535–538. https://doi.org/10.1111/j.1440-1754.1997.tb01666.x

Meštrović, T. (2018, August 23). Salmonella History. Retrieved May 9, 2020, from https://www.news-medical.net/health/Salmonella-History.aspx

Métris, A., Sudhakar, P., Fazekas, D., Demeter, A., Ari, E., Olbei, M., … Korcsmáros, T. (2017). SalmoNet, an integrated network of ten Salmonella enterica strains reveals common and distinct pathways to host adaptation. *NPJ Systems Biology and Applications*, *3*, 31. https://doi.org/10.1038/s41540-017-0034-z

Miele, V., Matias, C., Robin, S., & Dray, S. (2019). Nine quick tips for analyzing network data. *PLoS Computational Biology*, *15*(12), e1007434. https://doi.org/10.1371/journal.pcbi.1007434

Moran, N. A., & Plague, G. R. (2004). Genomic changes following host restriction in bacteria. *Current Opinion in Genetics & Development*, *14*(6), 627–633. https://doi.org/10.1016/j.gde.2004.09.003

Mosca, R., Céol, A., & Aloy, P. (2013). Interactome3D: adding structural details to protein networks. *Nature Methods*, *10*(1), 47–53. https://doi.org/10.1038/nmeth.2289

Muetze, T., Goenawan, I. H., Wiencko, H. L., Bernal-Llinares, M., Bryan, K., & Lynn, D. J. (2016). Contextual Hub Analysis Tool (CHAT): A Cytoscape app for identifying contextually relevant hubs in biological networks. [version 2; peer review: 2 approved]. *F1000Research*, *5*, 1745. https://doi.org/10.12688/f1000research.9118.2

Mühl, H. (2013). Pro-Inflammatory Signaling by IL-10 and IL-22: Bad Habit Stirred Up by Interferons? *Frontiers in Immunology*, *4*, 18. https://doi.org/10.3389/fimmu.2013.00018

Mulder, N. J., Akinola, R. O., Mazandu, G. K., & Rapanoel, H. (2014). Using biological networks to improve our understanding of infectious diseases. *Computational and Structural Biotechnology Journal*, *11*(18), 1–10. https://doi.org/10.1016/j.csbj.2014.08.006

Münch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E., & Jahn, D. (2003). PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Research*, *31*(1), 266–269. https://doi.org/10.1093/nar/gkg037

Murira, A., & Lamarre, A. (2016). Type-I Interferon Responses: From Friend to Foe in the Battle against Chronic Viral Infection. *Frontiers in Immunology*, *7*, 609. https://doi.org/10.3389/fimmu.2016.00609

Nguyen, N. T. T., Contreras-Moreira, B., Castro-Mondragon, J. A., Santana-Garcia, W., Ossio, R., Robles-Espinoza, C. D., … Thomas-Chollier, M. (2018). RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Research*, *46*(W1), W209–W214. https://doi.org/10.1093/nar/gky317

Nilsson, A. I., Koskiniemi, S., Eriksson, S., Kugelberg, E., Hinton, J. C. D., & Andersson, D. I. (2005). Bacterial genome size reduction by experimental evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(34), 12112–12116. https://doi.org/10.1073/pnas.0503654102

Novichkov, P. S., Kazakov, A. E., Ravcheev, D. A., Leyn, S. A., Kovaleva, G. Y., Sutormin, R. A., … Rodionov, D. A. (2013). RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics*, *14*, 745. https://doi.org/10.1186/1471-2164-14-745

Nuccio, S.-P., & Bäumler, A. J. (2014). Comparative analysis of Salmonella genomes identifies a metabolic network for escalating growth in the inflamed gut. *MBio*, *5*(2), e00929-14. https://doi.org/10.1128/mBio.00929-14

Obenauer, J. C., & Yaffe, M. B. (2004). Computational prediction of protein-protein interactions. *Methods in Molecular Biology*, *261*, 445–468. https://doi.org/10.1385/1-59259-762-9:445

O'Brien, E. J., Monk, J. M., & Palsson, B. O. (2015). Using genome-scale models to predict biological capabilities. *Cell*, *161*(5), 971–987. https://doi.org/10.1016/j.cell.2015.05.019

O'Brien, K. P., Remm, M., & Sonnhammer, E. L. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, *33*(Database issue), D476-80. https://doi.org/10.1093/nar/gki107

Ochman, H., & Wilson, A. C. (1987). Evolution in bacteria: evidence for a universal

substitution rate in cellular genomes. *Journal of Molecular Evolution*, *26*(1–2), 74–86. https://doi.org/10.1007/BF02111283

Ohl, M. E., & Miller, S. I. (2001). Salmonella: a model for bacterial pathogenesis. *Annual Review of Medicine*, *52*, 259–274. https://doi.org/10.1146/annurev.med.52.1.259

Okoro, C. K., Barquist, L., Connor, T. R., Harris, S. R., Clare, S., Stevens, M. P., … Kingsley, R. A. (2015). Signatures of adaptation in human invasive Salmonella Typhimurium ST313 populations from sub-Saharan Africa. *PLoS Neglected Tropical Diseases*, *9*(3), e0003611. https://doi.org/10.1371/journal.pntd.0003611

Okoro, C. K., Kingsley, R. A., Connor, T. R., Harris, S. R., Parry, C. M., Al-Mashhadani, M. N., … Dougan, G. (2012). Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. *Nature Genetics*, *44*(11), 1215–1221. https://doi.org/10.1038/ng.2423

Olbei, M., Hautefort, I., Modos, D., Treveil, A., Poletti, M., Gul, L., … Korcsmaros, T. (2021). SARS-CoV-2 Causes a Different Cytokine Response Compared to Other Cytokine Storm-Causing Respiratory Viruses in Severely Ill Patients. *Frontiers in Immunology*, *12*, 629193. https://doi.org/10.3389/fimmu.2021.629193

Olbei, M., Kingsley, R. A., Korcsmaros, T., & Sudhakar, P. (2019). Network Biology Approaches to Identify Molecular and Systems-Level Differences Between Salmonella Pathovars. *Methods in Molecular Biology*, *1918*, 265–273. https://doi.org/10.1007/978-1-4939-9000-9_21

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., … Hermjakob, H. (2014). The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, *42*(Database issue), D358-63. https://doi.org/10.1093/nar/gkt1115

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature*

Biotechnology, *28*(3), 245–248. https://doi.org/10.1038/nbt.1614

Ostaszewski, M., Mazein, A., Gillespie, M. E., Kuperstein, I., Niarakis, A., Hermjakob, H., … Schneider, R. (2020). COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Scientific Data*, *7*(1), 136. https://doi.org/10.1038/s41597-020-0477-8

Otasek, D., Morris, J. H., Bouças, J., Pico, A. R., & Demchak, B. (2019). Cytoscape Automation: empowering workflow-based network analysis. *Genome Biology*, *20*(1), 185. https://doi.org/10.1186/s13059-019-1758-4

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., … Tyers, M. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, *47*(D1), D529–D541. https://doi.org/10.1093/nar/gky1079

Owen, S. V., Wenner, N., Canals, R., Makumi, A., Hammarlöf, D. L., Gordon, M. A., … Hinton, J. C. D. (2017). Characterization of the prophage repertoire of african salmonella typhimurium ST313 reveals high levels of spontaneous induction of novel phage BTP1. *Frontiers in Microbiology*, *8*, 235. https://doi.org/10.3389/fmicb.2017.00235

Park, C. J., & Andam, C. P. (2020). Distinct but Intertwined Evolutionary Histories of Multiple Salmonella enterica Subspecies. *MSystems*, *5*(1). https://doi.org/10.1128/mSystems.00515-19

Parkhill, J, Dougan, G., James, K. D., Thomson, N. R., Pickard, D., Wain, J., … Barrell, B. G. (2001). Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. *Nature*, *413*(6858), 848–852. https://doi.org/10.1038/35101607

Parkhill, Julian, Sebaihia, M., Preston, A., Murphy, L. D., Thomson, N., Harris, D. E., … Maskell, D. J. (2003). Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. *Nature Genetics*,

35(1), 32–40. https://doi.org/10.1038/ng1227

Patterson, B. K., Seethamraju, H., Dhody, K., Corley, M. J., Kazempour, K., Lalezari, J. P., … Sacha, J. B. (2020). Disruption of the CCL5/RANTES-CCR5 Pathway Restores Immune Homeostasis and Reduces Plasma Viral Load in Critical COVID-19. *MedRxiv*. https://doi.org/10.1101/2020.05.02.20084673

Pedersen, S. F., & Ho, Y.-C. (2020). SARS-CoV-2: a storm is raging. *The Journal of Clinical Investigation*, *130*(5), 2202–2205. https://doi.org/10.1172/JCI137647

Perfetto, L., Acencio, M. L., Bradley, G., Cesareni, G., Del Toro, N., Fazekas, D., … Licata, L. (2019). CausalTAB: the PSI-MITAB 2.8 updated format for signalling data representation and dissemination. *Bioinformatics*, *35*(19), 3779–3785. https://doi.org/10.1093/bioinformatics/btz132

Petersen, E., Mills, E., & Miller, S. I. (2019). Cyclic-di-GMP regulation promotes survival of a slow-replicating subpopulation of intracellular Salmonella Typhimurium. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(13), 6335–6340. https://doi.org/10.1073/pnas.1901051116

Pickard, D., Wain, J., Baker, S., Line, A., Chohan, S., Fookes, M., … Dougan, G. (2003). Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding Salmonella enterica pathogenicity island SPI-7. *Journal of Bacteriology*, *185*(17), 5055–5065.

Pillich, R. T., Chen, J., Rynkov, V., Welker, D., & Pratt, D. (2017). Ndex: A community resource for sharing and publishing of biological networks. *Methods in Molecular Biology*, *1558*, 271–301. https://doi.org/10.1007/978-1-4939-6783-4_13

Pontes, M. H., Lee, E.-J., Choi, J., & Groisman, E. A. (2015). Salmonella promotes virulence by repressing cellulose production. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(16), 5183–5188.

https://doi.org/10.1073/pnas.1500989112

Povolotsky, T. L., & Hengge, R. (2016). Genome-Based Comparison of Cyclic Di-GMP Signaling in Pathogenic and Commensal Escherichia coli Strains. *Journal of Bacteriology*, *198*(1), 111–126. https://doi.org/10.1128/JB.00520-15

Pratt, D., Chen, J., Pillich, R., Rynkov, V., Gary, A., Demchak, B., & Ideker, T. (2017). Ndex 2.0: A clearinghouse for research on cancer pathways. *Cancer Research*, *77*(21), e58–e61. https://doi.org/10.1158/0008-5472.CAN-17-0606

Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., … Ideker, T. (2015). NDEx, the network data exchange. *Cell Systems*, *1*(4), 302–305. https://doi.org/10.1016/j.cels.2015.10.001

Pulford, C. V., Perez-Sepulveda, B. M., Canals, R., Bevington, J. A., Bengtsson, R. J., Wenner, N., … Hinton, J. C. D. (2021). Stepwise evolution of Salmonella Typhimurium ST313 causing bloodstream infection in Africa. *Nature Microbiology*, *6*(3), 327–338. https://doi.org/10.1038/s41564-020-00836-1

Remm, M., Storm, C. E., & Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, *314*(5), 1041–1052. https://doi.org/10.1006/jmbi.2000.5197

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., & Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, *17*(10), 1030–1032. https://doi.org/10.1038/13732

Rivera-Chávez, F., & Bäumler, A. J. (2015). The pyromaniac inside you: salmonella metabolism in the host gut. *Annual Review of Microbiology*, *69*, 31–48. https://doi.org/10.1146/annurev-micro-091014-104108

Rivera-Chávez, F., Lopez, C. A., Zhang, L. F., García-Pastor, L., Chávez-Arroyo, A., Lokken, K. L., … Bäumler, A. J. (2016). Energy Taxis toward Host-Derived Nitrate Supports a

Salmonella Pathogenicity Island 1-Independent Mechanism of Invasion. *MBio*, *7*(4). https://doi.org/10.1128/mBio.00960-16

Rodionov, D. A. (2007). Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chemical Reviews*, *107*(8), 3467–3497. https://doi.org/10.1021/cr068309+

Rogers, A. W. L., Tsolis, R. M., & Bäumler, A. J. (2021). Salmonella versus the Microbiome. *Microbiology and Molecular Biology Reviews*, *85*(1). https://doi.org/10.1128/MMBR.00027-19

Romero-González, L. E., Pérez-Morales, D., Cortés-Avalos, D., Vázquez-Guerrero, E., Paredes-Hernández, D. A., Estrada-de Los Santos, P., … Ibarra, J. A. (2020). The Salmonella Typhimurium InvF-SicA complex is necessary for the transcription of sopB in the absence of the repressor H-NS. *Plos One*, *15*(10), e0240617. https://doi.org/10.1371/journal.pone.0240617

Römling, U., Galperin, M. Y., & Gomelsky, M. (2013). Cyclic di-GMP: the first 25 years of a universal bacterial second messenger. *Microbiology and Molecular Biology Reviews*, *77*(1), 1–52. https://doi.org/10.1128/MMBR.00043-12

Rosu, V., Chadfield, M. S., Santona, A., Christensen, J. P., Thomsen, L. E., Rubino, S., & Olsen, J. E. (2007). Effects of crp deletion in Salmonella enterica serotype Gallinarum. *Acta Veterinaria Scandinavica*, *49*, 14. https://doi.org/10.1186/1751-0147-49-14

Roy, S., Wapinski, I., Pfiffner, J., French, C., Socha, A., Konieczka, J., … Regev, A. (2013). Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Research*, *23*(6), 1039–1050. https://doi.org/10.1101/gr.146233.112

Ryjenkov, D. A., Tarutina, M., Moskvin, O. V., & Gomelsky, M. (2005). Cyclic diguanylate is a ubiquitous signaling molecule in bacteria: insights into biochemistry of the GGDEF

protein domain. *Journal of Bacteriology*, *187*(5), 1792–1798. https://doi.org/10.1128/JB.187.5.1792-1798.2005

Salamon, J., Goenawan, I. H., & Lynn, D. J. (2018). Analysis and visualization of dynamic networks using the dynet app for cytoscape. *Current Protocols in Bioinformatics*, *63*(1), e55. https://doi.org/10.1002/cpbi.55

Santolini, M., & Barabási, A.-L. (2018). Predicting perturbation patterns from the topology of biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(27), E6375–E6383. https://doi.org/10.1073/pnas.1720589115

Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., … Collado-Vides, J. (2019). RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Research*, *47*(D1), D212–D220. https://doi.org/10.1093/nar/gky1077

Schultz, M. (2008). Theobald Smith. *Emerging Infectious Diseases*, *14*(12), 1940–1942. https://doi.org/10.3201/eid1412.081188

Schultze, J. L., & Aschenbrenner, A. C. (2021). COVID-19 and the human innate immune system. *Cell*, *184*(7), 1671–1692. https://doi.org/10.1016/j.cell.2021.02.029

Seif, Y., Kavvas, E., Lachance, J.-C., Yurkovich, J. T., Nuccio, S.-P., Fang, X., … Monk, J. M. (2018). Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. *Nature Communications*, *9*(1), 3771. https://doi.org/10.1038/s41467-018-06112-5

Seif, Y., Monk, J. M., Machado, H., Kavvas, E., & Palsson, B. O. (2019). Systems Biology and Pangenome of Salmonella O-Antigens. *MBio*, *10*(4). https://doi.org/10.1128/mBio.01247-19

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T.

(2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., … Vogel, J. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, *464*(7286), 250–255. https://doi.org/10.1038/nature08756

Shou, C., Bhardwaj, N., Lam, H. Y. K., Yan, K.-K., Kim, P. M., Snyder, M., & Gerstein, M. B. (2011). Measuring the evolutionary rewiring of biological networks. *PLoS Computational Biology*, *7*(1), e1001050. https://doi.org/10.1371/journal.pcbi.1001050

Sierro, N., Makita, Y., de Hoon, M., & Nakai, K. (2008). DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Research*, *36*(Database issue), D93-6. https://doi.org/10.1093/nar/gkm910

Singletary, L. A., Karlinsey, J. E., Libby, S. J., Mooney, J. P., Lokken, K. L., Tsolis, R. M., … Fang, F. C. (2016). Loss of Multicellular Behavior in Epidemic African Nontyphoidal Salmonella enterica Serovar Typhimurium ST313 Strain D23580. *MBio*, *7*(2), e02265. https://doi.org/10.1128/mBio.02265-15

Smith, C., Stringer, A. M., Mao, C., Palumbo, M. J., & Wade, J. T. (2016). Mapping the Regulatory Network for Salmonella enterica Serovar Typhimurium Invasion. *MBio*, *7*(5). https://doi.org/10.1128/mBio.01024-16

Sonnhammer, E. L. L., & Östlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, *43*(Database issue), D234-9. https://doi.org/10.1093/nar/gku1203

Spanò, S., & Galán, J. E. (2012). A Rab32-dependent pathway contributes to Salmonella typhi host restriction. *Science*, *338*(6109), 960–963. https://doi.org/10.1126/science.1229224

Stark, C., Breitkreutz, B.-J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., … Tyers, M. (2011). The BioGRID Interaction Database: 2011 update. *Nucleic Acids Research*, *39*(Database issue), D698-704. https://doi.org/10.1093/nar/gkq1116

Stecher, B., Robbiani, R., Walker, A. W., Westendorf, A. M., Barthel, M., Kremer, M., … Hardt, W.-D. (2007). Salmonella enterica serovar typhimurium exploits inflammation to compete with the intestinal microbiota. *PLoS Biology*, *5*(10), 2177–2189. https://doi.org/10.1371/journal.pbio.0050244

Sudhakar, P., Jacomin, A.-C., Hautefort, I., Samavedam, S., Fatemian, K., Ari, E., … Nezis, I. P. (2019). Targeted interplay between bacterial pathogens and host autophagy. *Autophagy*, *15*(9), 1620–1633. https://doi.org/10.1080/15548627.2019.1590519

Suryawanshi, R. K., Koganti, R., Agelidis, A., Patil, C. D., & Shukla, D. (2021). Dysregulation of Cell Signaling by SARS-CoV-2. *Trends in Microbiology*, *29*(3), 224–237. https://doi.org/10.1016/j.tim.2020.12.007

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., … Mering, C. von. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613. https://doi.org/10.1093/nar/gky1131

Tamayo, R., Pratt, J. T., & Camilli, A. (2007). Roles of cyclic diguanylate in the regulation of bacterial pathogenesis. *Annual Review of Microbiology*, *61*, 131–148. https://doi.org/10.1146/annurev.micro.61.080706.093426

Tan, L., Wang, Q., Zhang, D., Ding, J., Huang, Q., Tang, Y.-Q., … Miao, H. (2020). Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. *Signal Transduction and Targeted Therapy*, *5*, 33. https://doi.org/10.1038/s41392-020-0148-4

Tanner, J. R., & Kingsley, R. A. (2018). Evolution of Salmonella within Hosts. *Trends in*

Microbiology, *26*(12), 986–998. https://doi.org/10.1016/j.tim.2018.06.001

Terentiev, A. A., Moldogazieva, N. T., & Shaitan, K. V. (2009). Dynamic proteomics in modeling of the living cell. Protein-protein interactions. *Biochemistry (Moscow)*, *74*(13), 1586–1607. https://doi.org/10.1134/S0006297909130112

Thiele, I., Hyduke, D. R., Steeb, B., Fankam, G., Allen, D. K., Bazzani, S., … Bumann, D. (2011). A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. *BMC Systems Biology*, *5*, 8. https://doi.org/10.1186/1752-0509-5-8

Tian, B., Zhao, C., Gu, F., & He, Z. (2017). A two-step framework for inferring direct protein-protein interaction network from AP-MS data. *BMC Systems Biology*, *11*(Suppl 4), 82. https://doi.org/10.1186/s12918-017-0452-y

Timme, R. E., Pettengill, J. B., Allard, M. W., Strain, E., Barrangou, R., Wehnes, C., … Brown, E. W. (2013). Phylogenetic diversity of the enteric pathogen Salmonella enterica subsp. enterica inferred from genome-wide reference-free SNP characters. *Genome Biology and Evolution*, *5*(11), 2109–2123. https://doi.org/10.1093/gbe/evt159

Todd, W. T., & Murdoch, J. M. (1983). Salmonella virchow: a cause of significant bloodstream invasion. *Scottish Medical Journal*, *28*(2), 176–178. https://doi.org/10.1177/003693308302800217

Train, C.-M., Glover, N. M., Gonnet, G. H., Altenhoff, A. M., & Dessimoz, C. (2017). Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. *Bioinformatics*, *33*(14), i75–i82. https://doi.org/10.1093/bioinformatics/btx229

Treveil, A., Bohar, B., Sudhakar, P., Gul, L., Csabai, L., Olbei, M., … Korcsmaros, T. (2021). ViralLink: An integrated workflow to investigate the effect of SARS-CoV-2 on intracellular signalling and regulatory pathways. *PLoS Computational Biology*, *17*(2),

e1008685. https://doi.org/10.1371/journal.pcbi.1008685

Troxell, B., Fink, R. C., Porwollik, S., McClelland, M., & Hassan, H. M. (2011). The Fur regulon in anaerobically grown Salmonella enterica sv. Typhimurium: identification of new Fur targets. *BMC Microbiology*, *11*, 236. https://doi.org/10.1186/1471-2180-11-236

Tsolis, R. M., Young, G. M., Solnick, J. V., & Bäumler, A. J. (2008). From bench to bedside: stealth of enteroinvasive pathogens. *Nature Reviews. Microbiology*, *6*(12), 883–892. https://doi.org/10.1038/nrmicro2012

Turatsinze, J.-V., Thomas-Chollier, M., Defrance, M., & van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols*, *3*(10), 1578–1588. https://doi.org/10.1038/nprot.2008.97

Türei, D., Földvári-Nagy, L., Fazekas, D., Módos, D., Kubisch, J., Kadlecsik, T., … Korcsmáros, T. (2015). Autophagy Regulatory Network - a systems-level bioinformatics resource for studying the mechanism and regulation of autophagy. *Autophagy*, *11*(1), 155–165. https://doi.org/10.4161/15548627.2014.994346

Türei, D., Korcsmáros, T., & Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, *13*(12), 966–967. https://doi.org/10.1038/nmeth.4077

Tursi, S. A., Puligedda, R. D., Szabo, P., Nicastro, L. K., Miller, A. L., Qiu, C., … Tükel, Ç. (2020). Salmonella Typhimurium biofilm disruption by a human antibody that binds a pan-amyloid epitope on curli. *Nature Communications*, *11*(1), 1007. https://doi.org/10.1038/s41467-020-14685-3

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., … Pontén, F. (2015). Proteomics. Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419. https://doi.org/10.1126/science.1260419

Uzzau, S., Brown, D. J., Wallis, T., Rubino, S., Leori, G., Bernard, S., … Olsen, J. E. (2000). Host Adapted Serotypes of *Salmonella enterica*. *Epidemiology and Infection*, *125*(2), 229–255. https://doi.org/10.1017/s0950268899004379

Vågene, Å. J., Herbig, A., Campana, M. G., Robles García, N. M., Warinner, C., Sabin, S., … Krause, J. (2018). Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nature Ecology & Evolution*, *2*(3), 520–528. https://doi.org/10.1038/s41559-017-0446-6

Van Assche, E., Van Puyvelde, S., Vanderleyden, J., & Steenackers, H. P. (2015). RNA-binding proteins involved in post-transcriptional regulation in bacteria. *Frontiers in Microbiology*, *6*, 141. https://doi.org/10.3389/fmicb.2015.00141

Van Parys, T., Melckenbeeck, I., Houbraken, M., Audenaert, P., Colle, D., Pickavet, M., … Van de Peer, Y. (2017). A Cytoscape app for motif enumeration with ISMAGS. *Bioinformatics*, *33*(3), 461–463. https://doi.org/10.1093/bioinformatics/btw626

Vázquez-Torres, A. (2018). Less is best in the convergent evolution of typhoidal salmonella. *Cell Host & Microbe*, *23*(2), 151–153. https://doi.org/10.1016/j.chom.2018.01.009

Vazquez-Torres, A., Jones-Carson, J., Bäumler, A. J., Falkow, S., Valdivia, R., Brown, W., … Fang, F. C. (1999). Extraintestinal dissemination of Salmonella by CD18-expressing phagocytes. *Nature*, *401*(6755), 804–808. https://doi.org/10.1038/44593

Vogel, J., & Luisi, B. F. (2011). Hfq and its constellation of RNA. *Nature Reviews. Microbiology*, *9*(8), 578–589. https://doi.org/10.1038/nrmicro2615

Wang, H., Liu, B., Wang, Q., & Wang, L. (2013). Genome-wide analysis of the salmonella Fis regulon and its regulatory mechanism on pathogenicity islands. *Plos One*, *8*(5), e64688. https://doi.org/10.1371/journal.pone.0064688

Wang, Y., Chen, X., Hu, Y., Zhu, G., White, A. P., & Köster, W. (2018). Evolution and sequence diversity of fhua in salmonella and escherichia. *Infection and Immunity*,

86(11). https://doi.org/10.1128/IAI.00573-18

Wangdi, T., Lee, C.-Y., Spees, A. M., Yu, C., Kingsbury, D. D., Winter, S. E., … Bäumler, A. J. (2014). The Vi capsular polysaccharide enables Salmonella enterica serovar typhi to evade microbe-guided neutrophil chemotaxis. *PLoS Pathogens*, *10*(8), e1004306. https://doi.org/10.1371/journal.ppat.1004306

Waterman, S. R., & Small, P. L. (1998). Acid-sensitive enteric pathogens are protected from killing under extremely acidic conditions of pH 2.5 when they are inoculated onto certain solid food sources. *Applied and Environmental Microbiology*, *64*(10), 3882–3886.

Wei, C.-H., Allot, A., Leaman, R., & Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, *47*(W1), W587–W593. https://doi.org/10.1093/nar/gkz389

Wheeler, N. E., Barquist, L., Kingsley, R. A., & Gardner, P. P. (2016). A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. *Bioinformatics*, *32*(23), 3566–3574. https://doi.org/10.1093/bioinformatics/btw518

Williams, J., & Payne, W. J. (1964). Enzymes induced in a bacterium by growth on sodium dodecyl sulfate. *Applied Microbiology*, *12*, 360–362.

Winfield, M. D., & Groisman, E. A. (2004). Evolution and ecology of salmonella. *EcoSal Plus*, *1*(1). https://doi.org/10.1128/ecosalplus.6.4.6

Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in Bioinformatics*, *9*(4), 326–332. https://doi.org/10.1093/bib/bbn016

Winter, S. E., Raffatellu, M., Wilson, R. P., Rüssmann, H., & Bäumler, A. J. (2008). The Salmonella enterica serotype Typhi regulator TviA reduces interleukin-8 production in intestinal epithelial cells by repressing flagellin secretion. *Cellular Microbiology*, *10*(1),

247–261. https://doi.org/10.1111/j.1462-5822.2007.01037.x

Winter, S. E., Winter, M. G., Godinez, I., Yang, H.-J., Rüssmann, H., Andrews-Polymenis, H. L., & Bäumler, A. J. (2010). A rapid change in virulence gene expression during the transition from the intestinal lumen into tissue promotes systemic dissemination of Salmonella. *PLoS Pathogens*, *6*(8), e1001060. https://doi.org/10.1371/journal.ppat.1001060

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., … Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (New York, N.Y.)*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Yevshin, I., Sharipov, R., Valeev, T., Kel, A., & Kolpakov, F. (2017). GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Research*, *45*(D1), D61–D67. https://doi.org/10.1093/nar/gkw951

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2016). GGTREE: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.12628

Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J.-D. J., … Gerstein, M. (2004). Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Research*, *14*(6), 1107–1118. https://doi.org/10.1101/gr.1774904

Zahn-Zabal, M., Dessimoz, C., & Glover, N. M. (2020). Identifying orthologs with OMA: A primer. [version 1; peer review: 2 approved]. *F1000Research*, *9*, 27. https://doi.org/10.12688/f1000research.21508.1

Zhang, C., & Hua, Q. (2015). Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine. *Frontiers in Physiology*, *6*, 413. https://doi.org/10.3389/fphys.2015.00413

Zhang, Q. C., Petrey, D., Garzón, J. I., Deng, L., & Honig, B. (2013). PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Research*, *41*(Database issue), D828-33. https://doi.org/10.1093/nar/gks1231

Zhou, Z., Ren, L., Zhang, L., Zhong, J., Xiao, Y., Jia, Z., … Wang, J. (2020). Heightened Innate Immune Responses in the Respiratory Tract of COVID-19 Patients. *Cell Host & Microbe*, *27*(6), 883–890.e2. https://doi.org/10.1016/j.chom.2020.04.017

Peer-reviewed publications:

Mehta et al Genome Biology (2021) 22:25
https://doi.org/10.1186/s13059 020 02208 8

Genome Biology

RESEARCH Open Access

Evolution of regulatory networks associated with traits under selection in cichlids

Tarang K. Mehta[1], Christopher Koch[2], Will Nash[1], Sara A. Knaack[3], Padhmanand Sudhakar[1,4], Marton Olbei[1,4], Sarah Bastkowski[1,4], Luca Penso-Dolfin[1], Tamas Korcsmaros[1,4], Wilfried Haerty[1], Sushmita Roy[2,3,5*] and Federica Di-Palma[1,6,7*]

* Correspondence: sroy@biostat.wisc.edu; F.Di-Palma@uea.ac.uk
[2]Department of Biostatistics and Medical Informatics, UW Madison, Madison, USA
[1]Earlham Institute (EI), Norwich, UK
Full list of author information is available at the end of the article

Abstract

Background: Seminal studies of vertebrate protein evolution speculated that gene regulatory changes can drive anatomical innovations. However, very little is known about gene regulatory network (GRN) evolution associated with phenotypic effect across ecologically diverse species. Here we use a novel approach for comparative GRN analysis in vertebrate species to study GRN evolution in representative species of the most striking examples of adaptive radiations, the East African cichlids. We previously demonstrated how the explosive phenotypic diversification of East African cichlids can be attributed to diverse molecular mechanisms, including accelerated regulatory sequence evolution and gene expression divergence.

Results: To investigate these mechanisms across species at a genome-wide scale, we develop a novel computational pipeline that predicts regulators for co-extant and ancestral co-expression modules along a phylogeny, and candidate regulatory regions associated with traits under selection in cichlids. As a case study, we apply our approach to a well-studied adaptive trait—the visual system—for which we report striking cases of network rewiring for visual opsin genes, identify discrete regulatory variants, and investigate their association with cichlid visual system evolution. In regulatory regions of visual opsin genes, in vitro assays confirm that transcription factor binding site mutations disrupt regulatory edges across species and segregate according to lake species phylogeny and ecology, suggesting GRN rewiring in radiating cichlids.

Conclusions: Our approach reveals numerous novel potential candidate regulators and regulatory regions across cichlid genomes, including some novel and some previously reported associations to known adaptive evolutionary traits.

Keywords: Gene regulatory network, Co-expression, Cichlid, Opsin, Molecular evolution

Background

Seminal studies by King and Wilson [1] analyzing protein evolution in vertebrates speculated the importance of evolutionary changes in "regulatory processes" for morphological diversity [2, 3]. These ideas were soon expanded on by François Jacob [4],

who suggested that the molecular "tinkering" of pre-existing systems is a hallmark of evolution where, for example, regulatory processes can either be transformed or combined for functional gain [4]. These theories underlie many studies on the divergence of regulatory processes associated with morphological evolution, and broadly focus on changes in gene regulatory networks (GRNs) that determine the expression patterns of genes [5, 6]. Such changes can be mutations within transcription factor binding sites (TFBSs) located in *cis*-regulatory elements (promoters and enhancers) of genes or *trans* regulatory changes that are due to changes in the level of a regulator [6]. Alterations of GRNs can lead to phenotypic divergence [7], and these GRN changes between species, irrespective of direct and indirect functional consequence, are defined as GRN "rewiring" events. This is characterized by regulatory interactions present in one or more species but absent in another species, and potentially replaced by a new interaction between the orthologous TF and a target gene. Several comparative studies of GRNs underlying mechanisms of adaptation and evolution have been carried out in unicellular prokaryotes, *E. coli* [8] and several non-vertebrate eukaryotes, including yeast [9, 10], plants [11], fruit fly [12], and echinoderms [12, 13]. While there are efforts to collate and integrate several genomic datasets for vertebrates, including human and mouse [14], comparative analysis of regulatory networks from these data alone remains a major computational challenge and very little is known about the phenotypic effect of genome-wide regulatory network rewiring events in non-model vertebrates [15].

In vertebrates, ray-finned fishes are the largest radiation of any group, and the East African cichlids represent arguably the most speciose modern examples of adaptive radiations. In the great lakes of East Africa (Tanganyika, Victoria, and Malawi) and within the last few million years [16, 17], one or a few ancestral lineages of cichlid fish have independently radiated to collectively give rise to over 1500 species. These species occupy a large diversity of ecological niches and differ dramatically in phenotypic traits, including skeletal morphology, dentition, color patterning, and a range of behavioral traits. We have previously demonstrated that a number of molecular mechanisms have shaped East African cichlid genomes, e.g., rapid evolution of regulatory elements and gene expression divergence [18], and the "evolutionary tinkering" of these systems [19] has provided the necessary substrate for diversification [18]. This, coupled with the recent origin of cichlid species and ongoing gene flow [20], suggests that evolutionary regulatory changes have an important functional role in controlling gene expression and, ultimately, phenotypic variation. However, very little is known about the genome-wide evolution of regulatory networks that may underlie several traits of cichlid phenotypic diversity. Here we developed a novel computational framework to characterize the evolution of regulatory networks and analyze the plausibility of whether the "tinkering" of regulatory systems could contribute towards phenotypic diversity in closely related cichlids.

Results

Gene co-expression is tissue-specific and highlights functional evolutionary trajectories

We applied the Arboretum [9] algorithm to RNA-seq data of six tissues in five species and identified 10 modules of 12,051–14,735 co-expressed genes (1205–1474 genes per module per species) represented across 18,799 orthogroups (Fig. 1a). Modules of co-

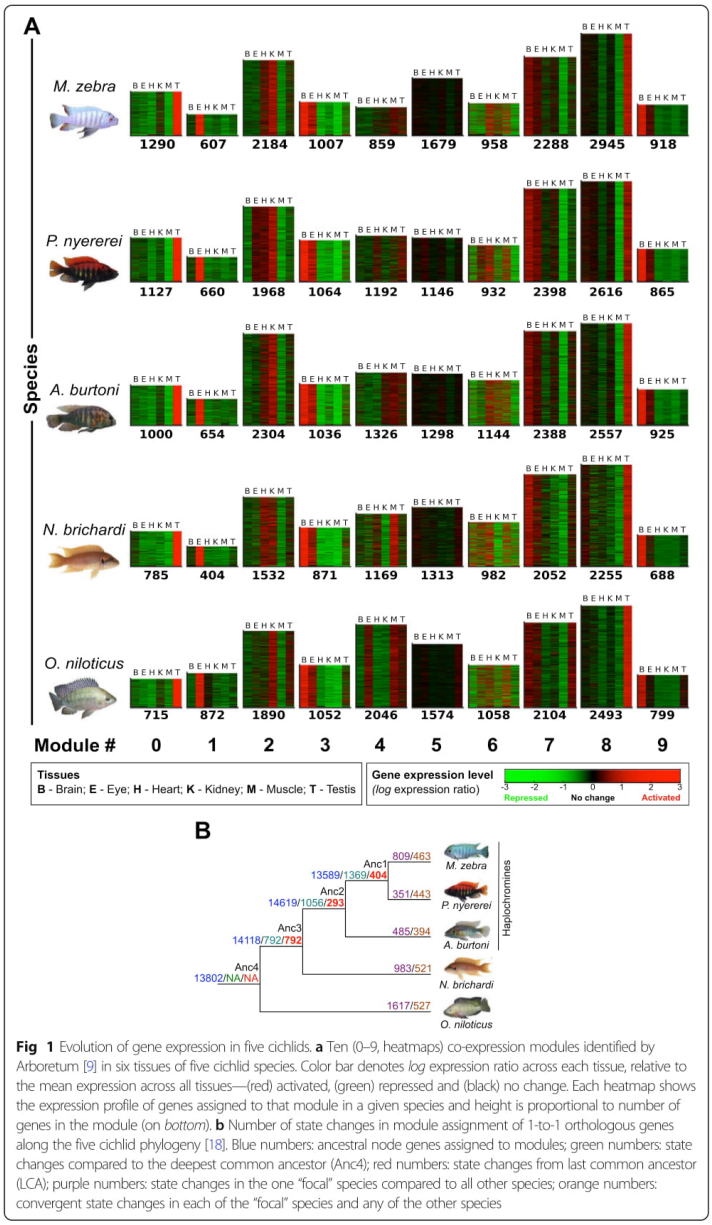

Fig 1 Evolution of gene expression in five cichlids. **a** Ten (0–9, heatmaps) co-expression modules identified by Arboretum [9] in six tissues of five cichlid species. Color bar denotes *log* expression ratio across each tissue, relative to the mean expression across all tissues—(red) activated, (green) repressed and (black) no change. Each heatmap shows the expression profile of genes assigned to that module in a given species and height is proportional to number of genes in the module (on *bottom*). **b** Number of state changes in module assignment of 1-to-1 orthologous genes along the five cichlid phylogeny [18]. Blue numbers: ancestral node genes assigned to modules; green numbers: state changes compared to the deepest common ancestor (Anc4); red numbers: state changes from last common ancestor (LCA); purple numbers: state changes in the one "focal" species compared to all other species; orange numbers: convergent state changes in each of the "focal" species and any of the other species

expressed genes across the five species show varying expression levels in specific tissues, e.g., module 1 is eye specific, while module 3 is heart, kidney, and muscle specific (Fig. 1a). Consistent with the phylogeny and divergence times, there are more (13,171/18,799) orthologous genes that are conserved in module assignment (orthologous modules) in the three closely related haplochromines (*Pundamilia nyererei*, *Maylandia zebra*, and *Astatotilapia burtoni*) and *Neolamprologus brichardi*, than with *Oreochromis niloticus* (11,212/18,799 orthologous genes). Examples of modules where orthologs are not conserved in module assignment (non-orthologous modules) include modules 2, 4, and 6 (Additional file 1: Fig. S1a, blue off-diagonal elements) and are representative of gene expression divergence across the species. Between the haplochromines alone, 4179/18,799 orthologous genes are distributed in either one of two modules, e.g., 0 or 8 (Additional file 1: Fig. S1a, blue off-diagonal elements in haplochromines), indicative of gene expression divergence along the phylogeny.

The assignment of co-expressed gene modules by Arboretum [9] is inferred using a probabilistic framework starting from the last common ancestor (LCA) in the phylogeny. This allows us to model the evolutionary trajectory of orthologous genes and their co-expression along the species tree [9]. Orthologous genes of each species can be assigned to non-orthologous modules (Fig. S-R1a), indicative of co-expression divergence and potential transcriptional rewiring from the LCA; this is referred to as "state changes" in module assignment. In total, 7587/18,799 (40%) orthologous genes exhibit state changes in module assignment across branches. To ensure orthologous genes of all branches are included in subsequent analysis, we focused on state changes of 6844 1-to-1 orthologous genes to assess convergent and unique state changes along the phylogeny (Fig. 1b). We identified convergent state changes of 732 genes along all ancestral nodes versus Anc4 (Additional file 1: Fig. S2). This is made up of 772 genes in Anc3 and Anc2, 734 genes in Anc3 and Anc1, and 996 genes in Anc2 and Anc1 (Additional file 1: Fig. S2), including a few TFs (46 TFs—Anc3-2-1; 49 TFs—Anc3-2; 46 TFs—Anc3-1; 66 TFs—Anc2-1) such as *tbx20*, *nkx3-1*, and *hoxd10*. We identified unique state changes and expression divergence of 655 genes along ancestral nodes (Fig. 1b), including several cellular and developmental TFs (51 TFs—Anc4/3; 20 TFs—Anc3/2; 34 TFs—Anc2/1) such as *foxo1*, *hoxa11* and *lbx1*. Several of these state changed regulatory TFs are also enriched (fold enrichment 1.1–1.7; false discovery rate, FDR < 0.05) in gene promoters of relevant tissue-specific modules; for example, promoters of module 1 genes (eye-specific expression) are significantly enriched (fold enrichment 1.1–1.6; FDR < 0.05) for TF motifs involved in retina- and lens-related development/functions, e.g., CRX, PITX3, and OTX1 [21] (Additional file 1: Fig. S3, Additional file 2: Fig. S2). Further examination identifies that there are differences in the levels of TF motif enrichment across species genes, including that of retina/lens-related TFs, e.g., RAR *α/β/γ* and RXR *α/β/γ* [22] of module 1 gene promoters in all species except *N. brichardi* (Additional file 1: Fig. S3, Additional file 2: Fig. S2). Such differences in motif enrichment could be associated with changes in the level of TF expression, where state changes (Fig. 1b) reflect shifted domains of tissue expression and imply differential regulatory control of target genes across tissues and along the phylogeny. We tested this by taking (1) the *log* expression ratio (as used for Arboretum input), for all 337 expressed TFs in each species tissue; (2) the corresponding 2064 TF motif enrichment scores (−*log q*-value, FDR < 0.05) calculated across 12,051–14,735

promoters regions of all species genes in the 10 modules; and (3) calculating the cross-species Pearson correlation coefficient (r) between the motif enrichment score and expression value of each TF and in each tissue (Additional file 2: Fig. S3-S8) using the $n = 5$ species. We note different patterns of correlation between cross-species TF motif enrichment and tissue-specific expression; in total, 102–119/337 TFs had no correlation (0 r 0.01, $n = 5$) and included many TFs that had large shifts in motif enrichment and/or expression in several species, representative of several phylogenetic state changes, e.g., Kidney-Module2-FOXO1 ($r = 0.01$, $n = 5$) (Extended Data S-R1F). On the other hand, there is positive correlation ranging from small (0.1 r 0.3, $n = 5$) for 161–197 TFs, medium ($0.3 < r$ 0.5) for 161–186 TFs, and large values ($0.5 < r$ 1) for 226–262 TFs. The largely correlated TFs ($0.5 < r$ 1) includes cases where there is comparable motif enrichment scores across species, as calculated by the variance distribution (see "Methods"), and either no shifts (no TF state changes), e.g., Brain-Module9-FOXA2 ($r = 0.97$, $n = 5$, p value < 0.05) or focused shifts (TF state change in one or subsets of species), e.g., Eye-Module2-CDX1 ($r = 0.98$, $n = 5$, p value < 0.05) in TF tissue expression (Additional file 1: Fig. S5, Additional file 2: Fig. S3-S8). Such patterns of focused shifts in expression are also observed in TFs of selected modules like, for example, module 1 which contains eye-expressed genes. We find that retinal TFs that are known to modulate opsin expression, e.g., CRX [23], have variable motif enrichment (fold enrichment 1.2–1.4) in eye-expressed genes, and are associated ($r = 0.85$, $n = 5$, p value < 0.1) with a concurrent change (increase in four species or decrease in *N. brichardi*) in TF eye expression along the phylogeny (Additional file 1: Fig. S6; see Additional file 1 *text*). For most TFs (226–262/337 TFs) and tissues, motif enrichment is largely correlated ($0.5 < r$ 1) with TF expression. After calculating the variance of each TF motif enrichment and categorizing the tails into either similar or dissimilar levels of TF motif enrichment (see "Methods"), we note that similar motif enrichment (across species) is associated with either expression conservation (across all species) or subtle expression changes (in one or subsets of species) and is more stable (in expression differences) than TFs with dissimilar/variable motif enrichment along the phylogeny (Additional file 2: Fig. S3-S8). Gene co-expression differences and convergence between species could therefore be driven by differences in TF motif levels in gene promoter regions.

Fine scale nucleotide variation at TF binding sites drives regulatory divergence in cichlids through GRN rewiring

Cis-regulatory elements, including promoters and enhancers, are central to gene expression regulation, largely acting through the binding of TFs to multiple transcription factor binding sites (TFBSs). Therefore, mutations within TFBSs can alter target gene transcription without affecting the expression pattern of other genes co-regulated by the same TF, thus driving GRN evolution. In the five cichlid genomes however, there is no significant increase in evolutionary rate at promoter regions compared to fourfold degenerate sites (Additional file 1: Fig. S7). However, we identify a few outlier genes with significantly higher evolutionary rate at promoter regions at ancestral nodes (12–351 genes, Additional file 1: Fig. S7b) and within species (29–352 genes, Additional file 1: Fig. S7d), indicative of small-scale changes in promoter regions (see Additional file 1

text). Concurrently, of all the identified pairwise species variation (8 to 32 million variants), a large proportion (13–28%) overlap predicted TFBSs in promoter regions, and this is higher than (8–9%) of variants that are present in flanking gene promoter regions of the same length (Additional file 1: Table S2, Additional file 1: Fig. S8). GO enrichment analysis of co-expressed genes with variation in their regulatory regions, against a background of all genes in each genome, highlights associations with key molecular processes, e.g., signal transduction-promoter TFBSs (Additional file 1: Fig. S9).

To further investigate patterns of divergent regulatory programs that could be associated with discrete nucleotide variation at regulatory binding sites, we developed and applied a computational framework (see "Methods," Additional file 1: Fig. S20) to comparatively study regulatory interactions/relationships across the five cichlids. This involved the reconstruction of species-specific GRNs through the integration of different genomic datasets (Additional file 1: Table S3). We focused on regulatory interactions/relationships of trans-acting factors (TFs) and DNA (gene promoter regions); this involved integrating an expression-based network with in silico predictions of TF binding to target gene (TG) promoters using our cichlid-specific and vertebrate-wide TF motif scanning pipeline (see "Methods," Additional file 1: Fig. S20). We first used species- and module-specific gene expression levels to infer an expression-based network [24] (see "Methods," Additional file 1: Fig. S20), generating 3180–4099 transcription factor-target gene (TF-TG) edges across the five species (FDR < 0.05, Additional file 1: Table S3). Next, based on our in silico TFBS motif prediction pipeline, we predicted TFBS motifs up to 20 kb upstream of a gene transcription start site (TSS), and using sliding window analysis of 100 nucleotides (nt), we retained TF motifs in the gene promoter region, defined as up to 5 kb upstream of a gene TSS (see "Methods," Additional file 1: Fig. S22). Each statistically significant TFBS motif (FDR < 0.05) was associated to its proximal target gene (TG) and represented as two nodes and one TF-TG edge. Based on the integrated approach (see "Methods," Additional file 1: Fig. S20), we predicted a total of 3,295,212–5,900,174 TF-TG edges (FDR < 0.05) across the five species that could be encoded into a matrix of 1,131,812 predicted TF-TG edges (FDR < 0.05), where each edge is present in at least two species. To ensure accurate analysis of GRN rewiring and to retain relevant TF-TG interactions, all collated edges were then further pruned to a total of 843,168 TF-TG edges (FDR < 0.05) where (1) the edge is present in at least two species; (2) edges are not absent in any species due to node loss or mis-annotation; and (3) edges are based on the presence of nodes in modules of co-expression genes (see "Methods").

We used three metrics to study large-scale TF-TG network rewiring between species that included: (1) state changes in module assignment; (2) DyNet [25] network rewiring scores; and (3) TF rate of edge gain and loss in networks. The first metric compares TF-TG edges of a single "focal" species versus the other species in the context of gene co-expression, while the second and third metric compute a likelihood score for the overall extent of edge changes (across all species) associated with single nodes of interest. We first focused on 6844 1-to-1 orthologous genes represented in 215,810 TF-TG interactions, termed "TF-TG 1-to-1 edges," along the five cichlid tree. Using a background set of all module genes (18,799 orthogroups), the TF-TG 1-to-1 edges are associated with morphogenesis and cichlid traits under selection, e.g., eye and brain development (FDR < 0.05, Additional file 1: Fig. S10a). There are 379 TFs represented

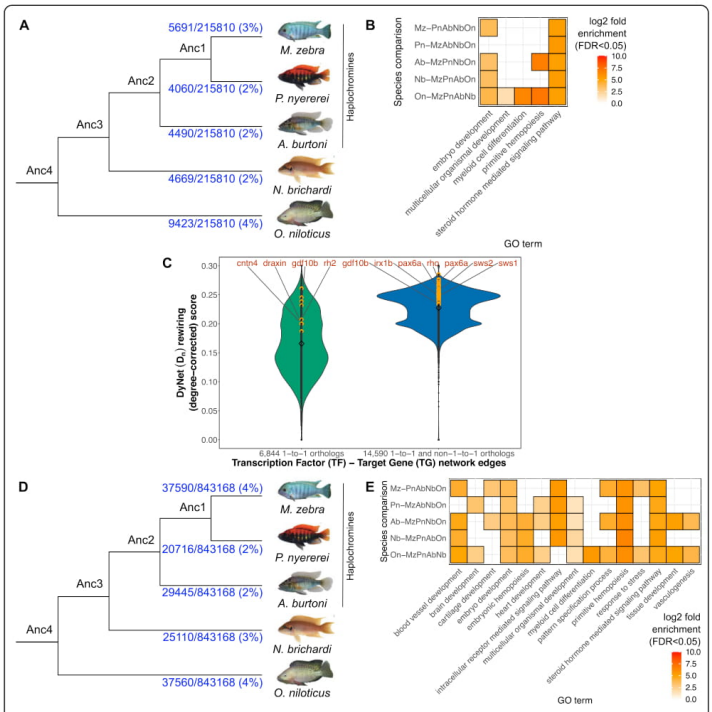

Fig 2 Network rewiring between species including TFs state changing co-expression module assignment and their targets (TGs). **a** Rewiring events linked to module assignment state changes of TFs in 215,810 TF-TG 1-to-1 edges (FDR < 0.05) of each species in the cichlid phylogeny compared to the other four species (see Additional file 1 for other FDR thresholds). **b** GO term enrichment of the 50–70 TFs that are rewired and state changed, and their associated TGs in one "focal" species vs the other four species (4060-9423/215,810 TF-TG 1-to-1 edges, **a**) against a background of all module genes, shown as grid heatmap of *log10* fold enrichment (legend on *right*, FDR < 0.05). **c** Violin plots of DyNet (D_n) rewiring score (degree-corrected) from 6844 1-to-1 orthologs in 215,810 TF-TG network edges (green, left violin) and 14,590 1-to-1 and many-to-many orthologs in 843,168 TF-TG network edges (blue, right violin). Mean rewiring score shown within each plot (white diamond). Degree-corrected rewiring score shown for non-candidate genes (black dots through center) and candidate morphogenetic trait genes (orange dots) with rewiring scores higher than the mean, and selected candidate examples are demarcated within. **d** Rewiring events linked to module assignment state changes of TFs in 843,168 TF-TG all edges of each species in the cichlid phylogeny against the other four species. **e** GO term enrichment of 100–140 rewired and state changed TFs and their associated TGs in each focal species vs the other 4 species (20,716-37,590/843,168 edges TF-TG all edges, **d**) against a background of all module genes, shown as grid heatmap of *log10* fold enrichment (legend on *right*, FDR < 0.05)

in the TF-TG 1-to-1 edges, and we focus on their interactions/relationships to determine whether TFs with (state) changes in module assignment have altered regulatory edges. In the first metric, rewiring is characterized as a unique TF-TG edge present in only one "focal" species, where the TF node is (1) state changed in module assignment and (2) present as a node in different TF-TG edges in any/all of the other species. Using this metric, 50–70 out of the 379 TFs (13–18%) are rewired (spanning 4060–9423/215,810 edges, FDR < 0.05, Fig. 2a; see Additional file 1 text) and change module

assignment across the five species (in one focal vs all four other species). The gene nodes connected by the rewired edges are associated with signalling pathways and processes such as cell differentiation and embryonic development (FDR < 0.05, background of all module genes, Fig. 2b). Further examination of rewiring rates in the networks of 6844 1-to-1 orthologous genes (in 215,810 TF-TG interactions) using the DyNet [25] degree-corrected rewiring (D_n) score (Fig. 2c, Additional file 3: Table S1) identifies rewired networks of nine teleost and cichlid trait genes associated with morphogenesis from previous studies (Fig. 2c, Additional file 3: Table S2). These genes have a few standard deviations higher degree-corrected rewiring (D_n) score than the mean (0.17 ± 0.03 SD), and their rewiring scores are comparatively higher (Kolmogorov–Smirnov KS-test p value = 6×10^{-4}) than all 1-to-1 orthologs (Fig. 2c, left violin plot, orange dots; Additional file 3: Table S3; see Additional file 1 text). Examples of these rewired 1-to-1 genes include *gdf10b* associated with axonal outgrowth and fast evolving in cichlids [18] and the visual opsin gene, *rh2* (Fig. 2c, left violin plot; Additional file 3: Table S3 S-R3C). To enable a genome-wide study of network rewiring, we extend our analyses beyond the 6844 1-to-1 orthologs only, by including an additional 7746 many-to-many orthogroups (see "Methods") resulting in a set of 843,168 "TF-TG all edges" across the five species. Using a background set of all module genes (18,799 orthogroups), the gene nodes in the 843,168 TF-TG all edges are associated with morphogenesis, e.g., retina development (FDR < 0.05, Fig. SR3aB). These edges include interactions of 783 TFs of which 13–18% (100–140 TFs) are predicted to be rewired (in 20,716-37,590/843,168 edges, FDR < 0.05, Fig. 2d) and change module assignment across the five species (in one focal vs all four other species), indicating their associated transcriptional programs (FDR < 0.05, background of all module genes) are also altered (Fig. 2e). By examining the network rewiring rates of 14,590 orthogroups (in 843,168 TF-TG interactions, Additional file 3: Table S4) using DyNet [25], we identify 60 candidate teleost and cichlid trait genes associated with phenotypic diversity from previous studies (Fig. 2c, *right* violin plot; Additional file 3: Table S5). These genes have a few standard deviations higher degree-corrected rewiring (D_n) score than the mean (0.23 ± 0.007 SD) of all orthologs, and their rewiring score is comparatively higher (KS-test p value = 6×10^{-14}) (Fig. 2c, *right* violin plot, orange dots; Additional file 3: Table S4). These genes include those associated with craniofacial development, e.g., *dlx1a* and *nkx2-5* [21], telencephalon diversity, e.g., *foxg1* [26], tooth morphogenesis, e.g., *notch1* [27], and strikingly, most visual opsins, e.g., *rho*, *sws2*, and *sws1*, as well as genes associated with photoreceptor cell differentiation, *actr1b* [28], and eye development, *pax6a* [21] (Fig. 2c, *right* violin plot; Additional file 3: Table S5). We then focus on the gain and loss rates of 186/783 TFs with > 25 TF-TG edges along the five cichlid tree (see "Methods"). Out of the 186 TFs, 133 (72%) are predicted to have a higher rate of edge gain than loss, e.g., DLX5 and NEUROD2, possibly acting as recruited regulators of gene expression in each branch from their last common ancestor (LCA) (Additional file 3: Table S6), whereas 53/186 TFs (28%) have a higher loss of edges than gains, e.g., OLIG2 and NR2C2, implying loss of gene expression regulatory activity from their LCA (Additional file 3: Table S6). In general, TFs and their binding sites are evolving towards gaining, rather than losing regulatory edges from their LCA.

To further characterize the role of the observed changes in *cis*-regulatory elements and their potential association with cichlid traits, we extended our analyses to include

several radiating cichlid species. We screened all predicted TFBS (see "Methods") variants between *M. zebra* (a Lake Malawi species) and the other four cichlids, with their corresponding positions in 73 phenotypically distinct Lake Malawi species [20], to identify between-species variation at regulatory sites along the phylogeny (Additional file 1: Fig. S11). As expected, the majority of variation at regulatory sites is identified between *M. zebra* and distantly related Lake Malawi species clades, e.g., NKX2.1 TFBS in *sws1* gene promoter, whereas shared ancestral sites are found with mainly same/closely related Lake Malawi clades, e.g., EGR2 TFBS in *cntn4* gene promoter. Genes that are associated with traits under selection, e.g., visual systems [29] (*sws1)* and morphogenesis [18] (*cntn4*), harbor between species regulatory variants that segregate according to phylogeny and ecology of radiating lake species.

is-regulatory changes lead to GRN alterations that segregate according to phylogeny and ecology of radiating cichlids

Through our comparative approach, we can examine the regulatory network topology of several genes that are important for cichlid diversification [30, 31] and represented by our six tissues. As a case study, we focus on the cichlid visual system; the evolution of cichlid GRNs and diverse palettes of co-expressed opsins can induce large shifts in adaptive spectral sensitivity of adult cichlids [29], and thus, we hypothesize that opsin expression diversity is the result of rapid adaptive GRN evolution in cichlids. Indeed, by focusing on species utilizing the same wavelength visual palette and opsin genes, we note that several visual opsin genes (*rh2b*, *sws1*, *sws2a*, and *rho*) have considerably rewired regulatory networks (Additional file 3: Table S6). Across the predicted transcriptional networks of cichlid visual opsins, there are several visual-system-associated regulators (TFs) of opsin genes (*sws2a*, *rh2b*, and *rho*) that are either common, e.g., STAT1A, CRX, and GATA2, or unique to each species, e.g., IRF1, MAFA, and GATA2A (Additional file 1: Fig. S12–14). These patterns of TF regulatory divergence could therefore contribute to differential opsin expression.

Sws1 (ultraviolet) opsin is utilized as part of the short-wavelength sensitive palette in *N. brichardi* and *M. zebra*. While there are common regulators associated with retinal ganglion cell patterning in both species networks, e.g., SATB1 [32], there are also several unique regulators associated with nuclear receptor signalling, e.g., RXRB and NR2C2 [33], and retinal neuron synaptic activity, e.g., ATRX [34] (Fig. 3a). Overall, using a significance threshold of FDR < 0.05 for predicted TF-TG edges, there are more predicted unique TF regulators of *sws1* in *M. zebra* (38 TFs) as compared to *N. brichardi* (6 TFs) (Fig. 3a, *bottom right*). Furthermore, we identify that a candidate regulatory variant has likely broken the *M. zebra* NR2C2/RXRB shared motif that is otherwise predicted 2 kb upstream of the *N. brichardi sws1* TSS (Fig. 3b). Functional validation via EMSA confirms that NR2C2 and not RXRB binds to the predicted motif in the *N. brichardi sws1* promoter, forming a complex, and the variant has likely disrupted binding, and possibly regulation of *M. zebra sws1* (Fig. 3c, d). This is further supported by correlating expression values of these regulators and *sws1*, where NR2C2 is better associated with *sws1* than RXRB, particularly when focusing on eye tissue (Additional file 1: Fig. S16a *on right*; Additional file 1: Fig. S16b; see Additional file 1 *text*).

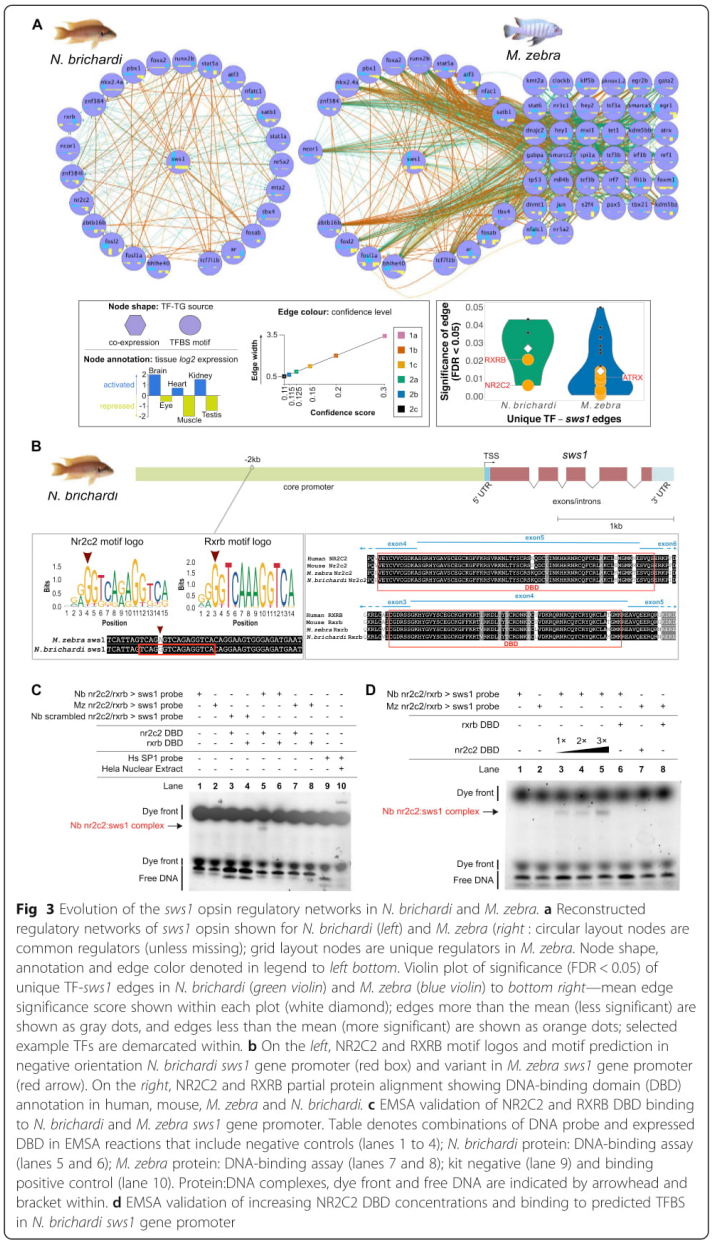

Fig 3 Evolution of the *sws1* opsin regulatory networks in *N. brichardi* and *M. zebra*. **a** Reconstructed regulatory networks of *sws1* opsin shown for *N. brichardi* (*left*) and *M. zebra* (*right* : circular layout nodes are common regulators (unless missing); grid layout nodes are unique regulators in *M. zebra*. Node shape, annotation and edge color denoted in legend to *left bottom*. Violin plot of significance (FDR < 0.05) of unique TF-*sws1* edges in *N. brichardi* (*green violin*) and *M. zebra* (*blue violin*) to *bottom right*—mean edge significance score shown within each plot (white diamond); edges more than the mean (less significant) are shown as gray dots, and edges less than the mean (more significant) are shown as orange dots; selected example TFs are demarcated within. **b** On the *left*, NR2C2 and RXRB motif logos and motif prediction in negative orientation *N. brichardi sws1* gene promoter (red box) and variant in *M. zebra sws1* gene promoter (red arrow). On the *right*, NR2C2 and RXRB partial protein alignment showing DNA-binding domain (DBD) annotation in human, mouse, *M. zebra* and *N. brichardi*. **c** EMSA validation of NR2C2 and RXRB DBD binding to *N. brichardi* and *M. zebra sws1* gene promoter. Table denotes combinations of DNA probe and expressed DBD in EMSA reactions that include negative controls (lanes 1 to 4); *N. brichardi* protein: DNA-binding assay (lanes 5 and 6); *M. zebra* protein: DNA-binding assay (lanes 7 and 8); kit negative (lane 9) and binding positive control (lane 10). Protein:DNA complexes, dye front and free DNA are indicated by arrowhead and bracket within. **d** EMSA validation of increasing NR2C2 DBD concentrations and binding to predicted TFBS in *N. brichardi sws1* gene promoter

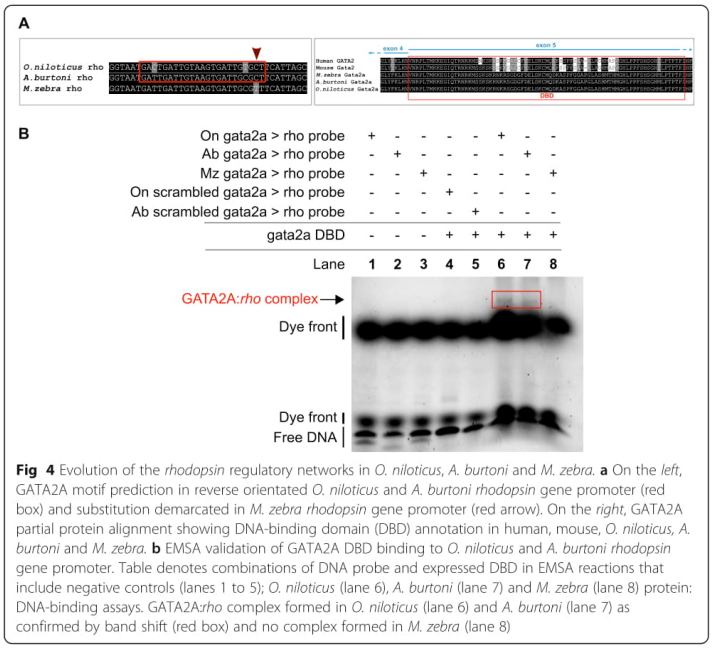

Fig 4 Evolution of the *rhodopsin* regulatory networks in *O. niloticus*, *A. burtoni* and *M. zebra*. **a** On the *left*, GATA2A motif prediction in reverse orientated *O. niloticus* and *A. burtoni rhodopsin* gene promoter (red box) and substitution demarcated in *M. zebra rhodopsin* gene promoter (red arrow). On the *right*, GATA2A partial protein alignment showing DNA-binding domain (DBD) annotation in human, mouse, *O. niloticus*, *A. burtoni* and *M. zebra*. **b** EMSA validation of GATA2A DBD binding to *O. niloticus* and *A. burtoni rhodopsin* gene promoter. Table denotes combinations of DNA probe and expressed DBD in EMSA reactions that include negative controls (lanes 1 to 5); *O. niloticus* (lane 6), *A. burtoni* (lane 7) and *M. zebra* (lane 8) protein: DNA-binding assays. GATA2A:*rho* complex formed in *O. niloticus* (lane 6) and *A. burtoni* (lane 7) as confirmed by band shift (red box) and no complex formed in *M. zebra* (lane 8)

In another example, *rhodopsin rho)*, associated with dim-light vision, is predicted to be regulated by GATA2 in *O. niloticus*, *A. burtoni*, and *M. zebra* but not its duplicate gene, GATA2A only in *M. zebra* (Additional file 1: Fig. S14). We identify a candidate variant (red arrow, Fig. 4a) that has likely broken the *M. zebra* GATA2A motif that is otherwise predicted 1.8 kb and 1.9 kb upstream of the *O. niloticus* and *A. burtoni rho* TSS (Fig. 4a). Functional validation via EMSA confirms that GATA2A binds to the predicted motif in the *O. niloticus* and *A. burtoni rho* promoter, and the variant is likely to have disrupted binding, and possibly regulation of *M. zebra rho* (Fig. 4b). Species-specific expression correlations with the *rho* target gene are supportive of GATA2's possible conserved role in all three species (*O. niloticus* $r = 0.89$; *A. burtoni* $r = 0.39$; *M. zebra* $r = 0.28$, $n = 6$ Additional file 1: Fig. S17c), while a more divergent role of GATA2A (*O. niloticus* $r = 0.79$ and *A. burtoni* $r = 0.21$, $n = 6$) and negative correlation in *M. zebra* ($r = -0.18$, $n = 6$) is supportive (Additional file 1: Fig. S17c) of the EMSA validation (Fig. 4). This further supports the notion that discrete point mutations in TFBSs could be driving GRN evolution and rewiring events in traits that are under selection in radiating cichlids.

Finally, we studied GRN rewiring as a result of between species TFBS variation in the context of phylogeny and ecology of lake species. Owing to the variability and importance of spectral tuning of visual systems to the foraging habits of all cichlid species, we focused on variants at regulatory sites of rewired visual opsin genes in the Lake Malawi species, *M. zebra*, as a reference to compare GRN rewiring (through TFBS variation)

that could be associated with the ecology of sequenced Lake Malawi species [20]. If indeed the TFBSs are likely functional, we hypothesize that radiating species with similar foraging habits would share conserved regulatory genotypes, to possibly regulate and tune similar spectral sensitivities, whereas distally related species with dissimilar foraging habits would segregate at the corresponding regulatory site. For this, we started with 157,232 sites that (1) have identified variation between the five cichlid species and (2) are located in TFBSs of *M. zebra* candidate gene promoters. We identified 5710/157,232 sites with between species variation across 73 Lake Malawi species (Additional file 1: Fig. S11) that also exhibited flanking sequence conservation, representative of shared ancestral variation. The homozygous variant (T|T) that breaks binding of NR2C2 to *M. zebra sws1* promoter (Fig. 3 and Fig. 5 *blue arrow*) is (1) conserved with the fellow algae eater, *Tropheops tropheops*, that also utilizes the same short-wavelength palette; (2) heterozygous segregating *Petrotilapia genalutea*—C|T and *Iodotropheus sprengerae*—T|C) in closely related Mbuna species; and (3) homozygous segregated (C|C) in distantly related Mbuna species (*Cynotilapia afra*, *Corydoras axelrodi*, and *Genyochromis mento*) and most other Lake Malawi species of which some utilize the same short-wavelength palette and are algae eaters, e.g., *Hemitilapia oxyrhynchus* (Fig. 5). This suggests that in species closely related to *M. zebra*, and with a similar diet and more importantly, habitat, *sws1* may not be regulated by NR2C2, whereas in other species it could be, similar to *N. brichardi* (Fig. 3 and Fig. 5 *red arrow*). In another example, regulation of *rho* by GATA2, and not its duplicate, GATA2A (Fig. 4), could be sufficient for regulating dim-light vision response in rock dweller species (*M. zebra* and possibly *Petrotilapia genulatea*, *Tropheops tropheops* and *Iodotropheus sprengerae*), but both *gata2* copies could be required to regulate *rho* in many other Lake Malawi species (79% with C|C genotype that otherwise predicts the GATA2A TFBS in *rho* gene promoter), as well as *A. burtoni* and *O. niloticus* (Additional file 1: Fig. S14–15). This highlights the potential differential usage of a duplicate TF in dim-light vision regulation. Phylogenetic independent contrast analysis [37] of the NR2C2-*sws1* (Additional file 1: Fig. S18a-f) and GATA2A-*rho* (Additional file 1: Fig. S19a-f) genotypes against visual traits and ecology of each of the 73 Lake Malawi species highlights very little change in correlation once the phylogeny is taken into account and a regression model fitted. Based on these examples of TFBS variants that segregate according to phylogeny and ecology of lake species, GRN rewiring through TFBS variation could be a key contributing mechanism of evolutionary innovation, especially visual systems, in East African cichlid radiations.

Discussion

The evolutionary "tinkering" of regulatory systems through GRN divergence can facilitate the evolution of phenotypic diversity and rapid adaptation [19]. Various mechanisms underlie these events, including horizontal gene transfer and regulatory reorganization in bacteria [38]; gene duplication in fungi [39]; *cis*-regulatory expression divergence in flies [40]; variable gene co-expression in worms [41]; dynamic rewiring of TFs in plant leaf shape [11]; coding and non-coding evolution in stickleback fish [42]; alternative splicing [43], and differential rate of gene expression evolution shaped by various selective pressures [44, 45] in mammals. However, since very little is known about the combined effect of some of these mechanisms; in-depth analyses of

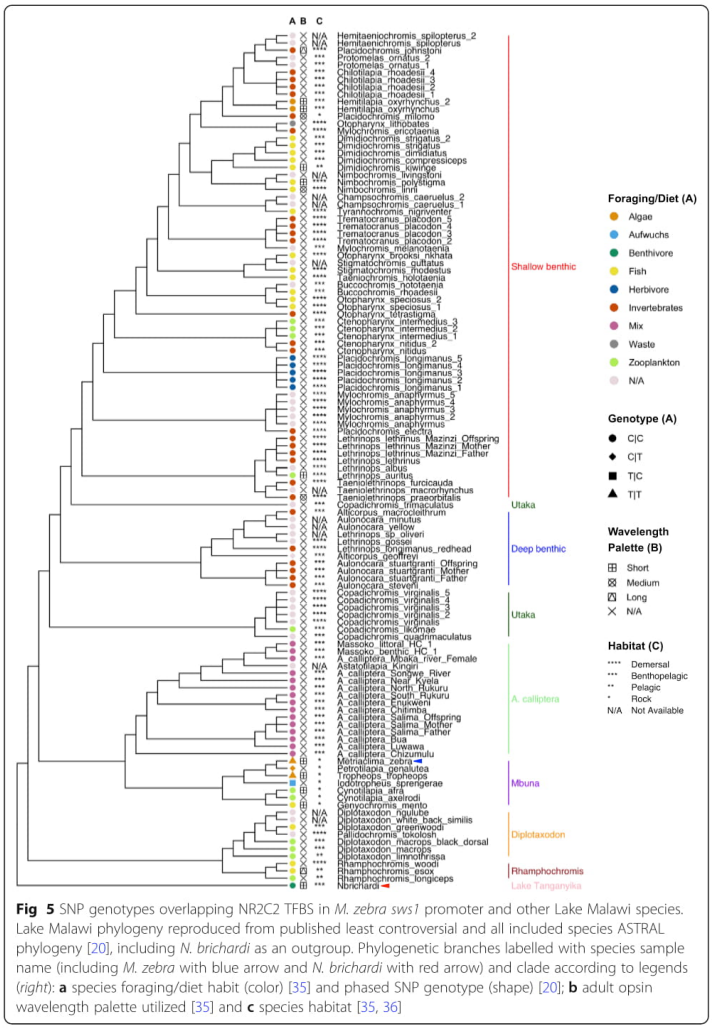

Fig 5 SNP genotypes overlapping NR2C2 TFBS in *M. zebra sws1* promoter and other Lake Malawi species. Lake Malawi phylogeny reproduced from published least controversial and all included species ASTRAL phylogeny [20], including *N. brichardi* as an outgroup. Phylogenetic branches labelled with species sample name (including *M. zebra* with blue arrow and *N. brichardi* with red arrow) and clade according to legends (*right*): **a** species foraging/diet habit (color) [35] and phased SNP genotype (shape) [20]; **b** adult opsin wavelength palette utilized [35] and **c** species habitat [35, 36]

regulatory network evolution can shed light on the key contributing mechanisms associated with phenotypic effect across ecologically diverse species in a phylogeny.

The three great lakes of East Africa (Tanganyika, Victoria, and Malawi) have independently experienced rapid radiations and explosive diversification of well over 1500 cichlid species. Alongside ecological opportunity [17], East African cichlid diversification has been shaped by complex evolutionary and genomic forces, including divergent selection acting upon regulatory regions [18] that is largely based on a canvas of low genetic diversity between species [20]. All of these findings imply the rapid evolution of

regulatory networks underlying traits under selection; however, little is known about the genome-wide evolution of regulatory networks that may underlie several traits of cichlid phenotypic diversity [46]. Here we developed a novel approach to identify and compare gene regulatory modules and networks across six tissues of five East African cichlid species.

Along the phylogeny, our analyses identified gene co-expression modules with tissue-specific patterns and differential trajectories across six tissues of five cichlids. Between the haplochromine species alone, nearly a quarter of all orthologous genes are distributed in either one of two modules. Considering the smaller divergence time of the three haplochromines (~ 6 MYA) and the three haplochromines vs *O. niloticus* (~ 19 MYA) [47], this indicates gene expression divergence over different evolutionary timescales and co-expression of different clusters of genes across species. Given that the volumes and, hence, representation of region-specific cell types of selected organ, e.g., brain regions can be different, even between closely related cichlids [48], it is plausible that the observed expression differences between species are driven by changes in cell type abundances. However, given that expression data was generated from the organs of multiple similarly sized adult individuals and the identification of conserved tissue-specific patterns across all tissues and species, e.g., module 1 is eye specific (Fig. 1a), we suspect that the majority of observed co-expression differences are connected to gene regulatory differences. Indeed, these genes are predicted to be regulated by divergent suites of regulators, including TFs that are state changed in co-expression module assignment. This suggests that gene co-expression differences and convergence between species could be driven by differences in TF motif levels in gene promoter regions and could be associated with gene regulatory changes underpinning traits under selection in cichlids, such as the visual system [29]. In the five cichlids, transcriptional rewiring events and differential gene expression could therefore contribute to phenotypic diversity of the six studied tissues.

Cis-regulatory elements (including promoters and enhancers) are central to cichlid gene expression regulation [18], and in this study, we show that discrete nucleotide variation at binding sites drives regulatory edge divergence through GRN rewiring events. Comparative analysis of GRNs across species identifies that TFs and their binding sites are evolving towards gaining, rather than losing regulatory edges, and possibly regulatory activity of genes from their LCA. Comparative GRN analysis also identified striking cases of rapid network rewiring for genes known to be involved in traits under natural and/or sexual selection, such as the visual system, possibly shaping cichlid adaptation to a variety of ecological niches. While there are common regulators of the *sws1* visual opsin in two species (*N. brichardi* and *M. zebra*) sharing the same short-wavelength palette, the *sws1* networks of these two species have substantially diverged. Such tight TF-based regulation of *N. brichardi sws1* could induce rapid shifts in expression and spectral shift sensitivities between a larger peak $_{max}$ of 417 nm in *N. brichardi* single cones [49] compared to 368 nm of *M. zebra* SWS1 [50]. Also, diverse regulation in *M. zebra* can increase *sws1* expression and, in turn, increase spectral sensitivity to UV light and the ability for *M. zebra* to detect/feed on UV-absorbing phytoplankton and algae, as previously shown for Lake Malawi cichlids [35]. In regulatory regions of *sws1*, in vitro assays confirm that variations in TFBSs NR2C2) have driven network structure rewiring between the two species *N. brichardi* and *M. zebra*) sharing the same visual palette.

Since the modulation of cichlid visual sensitivity occurs through heterochronic shifts in opsin expression [51], our results are consistent with recent findings that visual tuning differences between cichlid species require regulatory mutations that are constrained by mutational dynamics [52].

Gene duplications have also been implicated in cichlid evolutionary divergence, including differences in duplicate TF gene expression [18]. However, due to incomplete lineage sorting (ILS) and variability in duplicates identified by three separate methods (gene trees, read-depth analyses and array comparative genomic hybridization) [18], we instead focus on particular examples of gene duplication associated with network rewiring of visual system genes. We predict that the dim-light vision gene, *rho*, is regulated by GATA2 and potentially common to regulating dim-light vision in *M. zebra, A. burtoni*, and *O. niloticus* but a duplicate TF, GATA2A, is predicted to be a unique regulator of *rho* in *A. burtoni* and *O. niloticus* only, owing to a variant in the GATA2A TFBS of the *M. zebra rho* gene promoter. Furthermore, *M. zebra* variants overlapping TFBSs in gene promoter regions, e.g., *sws1* (NR2C2) and *rho* (GATA2A) segregate according to phylogeny and ecology of Lake Malawi species [20], suggesting ecotype-associated network rewiring events could be linked to traits under selection in East African cichlid radiations. This is consistent with the adaptive potential of visual system evolution in cichlid species, where changes in spectral tuning of visual signals are likely to lead to dramatic species evolution and possibly speciation events [53]. Given that single regulatory mutations of *Tbx2a* can cause heterochronic shifts in opsin expression and visual tuning diversity between two distinct cichlid species [52], it is likely that the regulatory variation at opsin gene promoter TFBSs that we have predicted and experimentally validated, is a contributing mechanism of evolutionary innovation across many cichlid species. Furthermore, the identification (in predicted TFBSs) of segregating sites across several Lake Malawi species, with conservation of flanking regions, is indicative of shared ancestral variation and functional evolutionary constraint. The differences we identify at opsin gene promoter TFBSs and their implications in visual tuning could correspond to species variation of habitat choice, foraging habits, diet, and male nuptial coloration. Phylogenetic independent contrast analysis [37] shows that fitting the Lake Malawi phylogeny has little effect on the correlation between regulatory genotypes, visual traits, and ecology, suggesting possible covariance between these genotypes and traits. However, given the weak correlation (low adjusted r^2 and p values), the impact of ecotype-associated network rewiring events requires further testing. This analysis would further benefit from (1) the addition of any missing data (wavelength palette, habitat, and/or foraging habit/diet) in the phylogeny; (2) the addition of further variables, e.g., average water depth measurements; (3) additional species data from lowly represented clades, e.g., Mbuna; and (4) further experimental testing, particularly in phenotypically divergent species pairs. Beyond the visual systems, we also identify network rewiring of genes associated with several cichlid adaptive traits like, for example, *runx2* associated with jaw morphology [54]; *ednrb1* in pigmentation and egg spots [18, 55]; and *egr1* implicated in behavioral phenotypes [56]. These also represent case studies that can be validated in species pairs that diverge for the trait of interest.

The regulatory networks generated here represent a rich scientific resource for the community, powering further molecular analysis of adaptive evolutionary traits in cichlids. As an example, further examination of the vast regulatory factors that we have

predicted for the visual systems that could both up- and downregulate opsin expression diversity and could further shed light on preliminary studies of SWS1 [57], LWS, and RH2 [52] in other cichlid species. This could involve further functional validation to define a definitive link to trait variation by (1) high-throughput protein-DNA assays to confirm binding of hundreds of sites; (2) reporter and/or cell-based TF-perturbation assays to show that the regulatory variants indeed affect transcription; and (3) genome editing, e.g., CRISPR mutations of TFBS variants followed by phenotyping to observe trait effect. Nonetheless, this study is the first genome-wide exploration of GRN evolution in cichlids, and the computational framework (Additional file 1: Fig. S20) is largely applicable to other phylogenies to study the evolution of GRNs. In this study, we largely focus on *cis*-regulatory mechanisms of GRN rewiring. However, given the potential impact of other genetic mechanisms (protein coding changes, small RNAs, and posttranslational modifications) towards cichlid phenotypic diversity [18, 46], our framework can be extended by the inclusion of relevant datasets to allow for studies on the regulatory effect of other mechanisms, e.g., miRNAs, enhancers, and gene duplications on network topology during cichlid evolution. While many of the predicted TF-TG interactions/relationships could be false positives, our integrative approach ensured that we could apply rigorous filtering at each step, including stringent statistical significance measures, co-expression-based pruning, and all while accounting for gene node loss and mis-annotations in selected species (see "Methods").

While it appears that cichlids utilize an array of regulatory mechanisms that are also shown to drive phenotypic diversity in other organisms [11, 39–42, 58], we provide experimental support of selected TF-TG rewiring events in regulatory regions of genes associated with adaptive traits in cichlids [18]. This is further supported by large-scale genotyping studies of the predicted sites in radiating cichlid species [20]. This potential link between GRN evolution and genes associated with adaptive trait variation in cichlids requires additional experimental verification and support by further studies on cichlid species that largely focus on large-scale genotyping [20]; whole-genome analysis and transgenesis assays [18]; behavioral and transcriptomic assays [59]; population studies and CRISPR mutant assays [60]; and transcriptomic/*cis*-regulatory assays [35, 49, 52, 57].

Conclusions

We present a novel computational framework to study the evolution of regulatory networks in representative species of the rapid adaptive radiations of East African cichlids. Using six tissues from five species, our approach identified tissue-specific gene expression divergence between the five cichlid species that is likely associated with gene regulatory changes. As a case study, we focus on a well-studied trait—the visual system—for which we identified regulatory variation at TFBSs and demonstrate how the functional disruption of TFBSs abrogates binding of key regulators and, thus, can drive GRN evolution. Our approach revealed hundreds of novel potential regulatory regions and regulators of the five cichlid genomes, many of which have been previously associated with evolutionary traits. In conclusion, we show that regulatory network evolution can be driven by discrete changes at regulatory binding sites, and network rewiring events are likely to be a contributing source to evolutionary innovations in radiating cichlid species. This approach, with further functional validations, has the potential to identify novel genes linked to other evolutionary traits in cichlids and other evolutionary systems.

Methods

A comparative framework to study the evolution of tissue-specific regulatory networks in cichlids

We developed a comparative framework (Additional file 1: Fig. S20) to infer gene regulatory networks across five representative East African cichlid species—*O. niloticus* (On), *N. brichardi* (Nb), *A. burtoni* (Ab), *P. nyererei* (Pn), and *M. zebra* (Mz). Our framework comprises (1) identifying modules of co-expressed genes from multi-tissue/multi-species and single-tissue/multi-species data; (2) integrating several datasets (gene expression and *cis* regulatory regions) to reconstruct gene regulatory networks (GRNs) to find fine-grained tissue-specific network modules; (3) examining factors driving evolutionary innovation in cichlids, i.e. nucleotide divergence within regulatory binding sites and determining their mechanistic roles towards regulatory network and module divergence; and (4) using an integration of the reconstructed networks, co-expression modules, and enrichment of curated biological processes to interpret GRN evolution of genes in the context of cichlid adaptive traits.

Inference of multi- and single-tissue transcriptional modules in five cichlids

We ran Arboretum [9], an algorithm for identifying modules of co-expressed genes on gene expression values of six tissues (brain, eye, heart, kidney, muscle, testis) from five cichlid species—*O. niloticus* (On), *N. brichardi* (Nb), *A. burtoni* (Ab), *P. nyererei* (Pn), and *M. zebra* (Mz) [18]. Tissues were isolated and RNA extracted from several adult individuals as described previously [18] and summarized here: *O. niloticus* tissues were isolated from Swansea stock individuals in the laboratory of Dr. Gideon Hulata (Volcani Center, Bet Dagan, Israel) and RNA extracted in the lab of Dr. Micha Ron (Volcani Center, Bet Dagan, Israel) using the mirVana miRNA Isolation Kit (Ambion); *N. brichardi* tissues were isolated from individuals inbred for ~ 10 generations in the laboratory of Prof. Walter Salzburger (University of Basel, Basel, Switzerland) and RNA extracted using TRIzol® (Invitrogen, USA); *A. burtoni* tissues were isolated from individuals inbred for ~ 60 generations in the laboratory of Dr. Hans Hoffman (University of Texas, Austin, TX, USA) and RNA extracted using TRIzol® (Invitrogen, USA); *P. nyererei* tissues were isolated from individuals inbred for ~ 5 generations in the lab of Prof. Ole Seehausen and RNA extracted using the QIAGEN RNeasy Plus Universal mini kit; *M. zebra* tissues were isolated from wild individuals in the laboratory of Dr. Karen Carleton (University of Maryland, College Park, MD, USA) and RNA extracted using the QIAGEN RNeasy Kit. In brief, the gene expression values used here were obtained from [18], and as described previously, this included (1) confirming RNA integrity on Agilent 2100 Bioanalyzer; (2) construction of RNA-seq libraries using a strand-specific dUTP protocol; (3) sequencing of RNA-seq libraries on HiSeq2000 (Illumina), yielding > 35 million 76 bp paired-end reads per tissue; (4) de novo transcriptome assembly using Trinity [61] and splice junction database from PASA gene models; (5) read alignment with TopHat2 [62]; and (6) calculating gene expression values (FPKM) with Cufflinks [63] using the protein-coding gene annotation as reference [18]. To ensure equality in *n*-fold change of expression, the gene expression values were log-transformed as: $log(x + 1)$, where x is the raw expression value [18], and "*log*" is the natural logarithm, and then expression was normalized across each gene to have mean zero to be used as input for Arboretum [9]. The *log* expression ratio shown

across modules is each gene expression relative to the mean expression across all tissues. Selection of the six tissues allowed us to study tissue-specific associated traits under natural and/or sexual selection in cichlids: brain (development, behavior and social interaction); eye (adaptive water depth/turbidity vision); heart (blood circulation and stress response); kidney (hematopoiesis and osmoregulation associated with water adaptation); muscle (size, shape, and movement associated with dimorphism and agility); and testis (sexual systems associated with behavior and dimorphism).

In total, 18,799 orthogroups, including 69,989 genes, and 34,220 1-to-1 orthologous genes (see "Cichlid gene trees"), and their associated expression data and gene tree information were inputted into Arboretum [9]. In total, this represents 59–68% of all protein-coding genes in the five cichlid genomes [18]. Certain annotated cichlid genes could not be included for a few reasons: (1) lack of tissue expression data for all five species; (2) no mapped reads for selected tissues; (3) Lack of co-expression with other genes; and (4) use of single development stage (adult). We selected the number of modules using a combination of strategies. First, we tried to identify the optimal number of multi-tissue modules *k)* automatically from the data by scoring the Arboretum learned model based on the penalized log likelihood and silhouette index for $k = 7–14$ modules in increments of 1 (Additional file 1: Fig. S21a). This gave us $k = 10$ and 12 as the settings were local maxima for silhouette index. Second, we manually inspected the modules to see if increases of k yield patterns of expression that we have not seen before or generate recurring patterns ($k = 12$ is shown in Additional file 1: Fig. S21b). Based on our strategy, we found $k = 10$ modules to be optimal. Finally, we devised a metric for the top three random initializations, based on a silhouette index, orthology overlap, and cross-species cluster mean dissimilarity, selecting the optimal k stable to the initialization. Using a similar approach, this time for single tissues clustering, we found $k = 5$ modules to be optimal. The single-tissue modules were only initially used to assess tissue-specific gene expression divergence.

Handling ILS in arboretum

The Arboretum algorithm internally tries to reconcile a tree that is not obeying the species tree by adding additional duplication and loss events. An alternate approach is to use a different species trees each representing the different ILS types and estimating the parameters of each such tree. However, there are many different cases of ILS, as identified previously [18], and the number of gene trees in each category varied significantly. Estimating the conditional distributions for each branch in each ILS type would not be feasible as there are not enough example trees.

Cichlid gene trees

By considering the gene tree of 18,799 orthologous groups (orthogroups), Arboretum [9] is able to generate module assignments reflecting many-to-many relationships between orthologs resulting from gene duplication and loss. To construct gene trees with different levels of duplication, we obtained the protein sequences of the longest transcripts from five cichlids as well as stickleback, spotted gar, and zebrafish as outgroups. Spotted gar was added as it predates the teleost-specific genome duplication event (3R) and zebrafish, as a model teleost to leverage known molecular interactions as an initial

prediction of functional relationships in cichlids based on orthology. We applied OrthoMCL-1.4.0 [64] followed by TreeFix-1.1.10 [65] to learn the reconciled gene trees. We noticed that several of the trees exhibited incomplete lineage sorting (ILS) for the cichlid-specific subtree but disappeared once the tree was relearned using the cichlid only species. We therefore relearned gene trees for the cichlid only species—in total, we reconstructed 17,858 gene families of which 108 had gene duplication events. A fraction of these (29 gene families) also exhibited ILS. We also observed ILS for gene groups without gene duplications: of the 17,756 gene families that had no duplication, 810 exhibited ILS.

Functional and transcription factor binding site (TFBS) enrichment in modules

We use the false discovery rate (FDR) corrected hypergeometric p value (q-value) test to assess enrichment of Gene Ontology (GO) terms and TFBSs (motifs) in a given gene set. In all cases, enrichment is tested using a set-based approach where a set of candidate genes is compared to a background (control set) of either all genes in species modules (18,799 orthogroups) or each genome (stated within figure legend for each test). We summarize the enrichment of terms/motifs with $q < 0.05$ statistical significance and conservation in all extant and ancestral species. GO terms for the five cichlids were from those published previously [18]. To study *cis*-regulatory elements likely driving tissue-specific expression patterns, we defined promoter regions for all genes in each of the five genomes. For this, we used the following published assemblies and associated gene annotations [18] for each species: *P. nyererei* PunNye1.0, NCBI BioProject: PRJNA60367; BROADPN2 annotation; *M. zebra* MetZeb1.1, NCBI BioProject: PRJNA60369; BROADMZ2 annotation; *A. burtoni* AstBur1.0, NCBI BioProject: PRJNA60363; BROADAB2 annotation; *N. brichardi* NeoBri1.0, NCBI BioProject: PRJNA60365; BROADNB2 annotation; *O. niloticus*—Orenil1.1 (NCBI BioProject: PRJNA59571; BROADON2 annotation. Gene promoter regions were defined as up to 5 kb upstream of the transcription start site (TSS) of each gene. This gene promoter region is based on analyzing the distribution of motifs in 100-nt window regions up to 20 kb upstream of each gene TSS, and observing a plateau of motifs (and distribution of CNEs) after ~ 5 kb in each species (Additional file 1: Fig. S22). Motif enrichment in *cis*-regulatory regions was carried out using TFBSs obtained by the method below, with a background (control set) of all motifs (FDR < 0.05) predicted within module gene promoters.

Transcription factor (TF) motif scanning

TFBSs of known vertebrate transcription factors (TFs) were obtained from the JASPAR vertebrate core motif (2018 release) [66]. Binding peak information from ChIP-seq experiments of various human and mouse TFs were retrieved from GTRD v17.04 [14] and associated to protein-coding genes within a vicinity of 10 kb. Using core motif sequences available from JASPAR [66] or alternative databases like UniPROBE [67] and HOCOMOCO [68], sequences matching these motifs were identified within the TF binding peaks. In cases where the core motifs were not available for specific TFs with ChIP-seq data, they were predicted de novo from the sequences under peaks themselves using MEME [69] with default settings. The aforementioned steps provided a list

of transcription factor-target gene (TF-TG) interactions with the exact coordinates of the corresponding binding site(s). Cichlid sites were extrapolated based on (1) gene-level orthology; (based on gene trees above), (2) minimum 70% sequence similarity [70, 71] between the vertebrate motif sequence and a sequence within the cichlid promoter, and (3) functional domain overlap as derived using *Interpro scan 5* [72] to both source organisms (human, mouse). Extrapolated sites from the promoters of each cichlid species were used to construct cichlid species-specific (CS) Position Specific Scoring Matrices (PSSMs) for each TF using the *info-gibbs* script from the RSAT tool suite [73]. In cases where the number of extrapolated sites per species was less than three, we aggregated the sites to construct generic cichlid-wide (CW) PSSMs. Using the PSSMs for each TF, we scanned up to 20 kb upstream of a genes TSS and conserved non-coding elements (CNEs) with FIMO [74] using either (1) an optimal calculated p value for each TF PSSM, calculated using the *matrix quality* script from the RSAT tool suite [73], with 1000 matrix permutations, or (2) FIMO [74] default p value (1e−4) for JASPAR [66] PSSMs and PSSMs for which an optimal p value could not be determined. Based on the distribution of motifs in 100-nt windows of up to 20 kb upstream of gene TSSs (Additional file 1: Fig. S22), we only retained motifs up to 5 kb upstream of a gene TSS as the gene promoter region (Additional file 1: Fig. S22). Statistically significant motifs were called using a q-value (FDR) < 0.05 and grouped in confidence levels and scores of (1a) overlap of mouse and human to cichlid extrapolated—0.3; (1b) mouse to cichlid extrapolated—0.2; (1c) human to cichlid extrapolated—0.15; (2a) FIMO [74] scans using extrapolated CS matrices—0.125; (2b) FIMO [74] scans using extrapolated CW matrices—0.110; and (2c) FIMO [74] scans using JASPAR [66] matrices—0.115. To assess whether motifs are predicted by chance, we also scanned randomized promoter sequences using the same PSSMs.

Calculating tissue specificity index tau)

As a measure for tissue specificity of gene expression, we calculated τ (Tau) [75] using log-transformed and normalized gene expression data (as inputted to run Arboretum):

$$\tau \quad \frac{\sum_{i\ 1}^{n}(1-\widehat{x_i})}{n-1} \quad \widehat{x_i} \quad \frac{x_i}{\max_{1\ i\ n}(x_i)}$$

Here, n is the number of tissues and x_i is the expression profile component normalized by the maximal component value [75]. The values of tau vary from 0 to 1: ubiquitous or broad expr (τ 0.5); intermediate expr ($0.5 < \tau < 0.9$); and tissue-specific or narrow expr ($\tau \geq 0.9$) [75]. Amongst existing methods, τ has been shown to be a reliable method for calculating tissue specificity [76]. Testes normally express far more genes than any other tissue, generally displaying a tissue-specific pattern of expression. As tau was used to assess genome-wide expression levels across all tissues, but between species, testis expression data was included for each species to obtain a true representation of variation in transcriptional programs.

Variation and evolutionary rate at coding and non-coding regions

We noticed several anomalous start site annotations of genes in *M. zebra*, *P. nyererei*, *A. burtoni*, and *N. brichardi* when compared to *O. niloticus*. Owing to these anomalies, we re-defined gene start sites to extract putative promoter regions. For each gene, we used the 1st exon (± 100 bp) of the longest protein-coding sequence in *O. niloticus* to identify, via BLAT-35 [77], corresponding orthologous start sites in the other four cichlid genomes. We filtered the output based on coherent overlap with original annotations [18] and orthogroups in cichlid gene trees. We re-annotated gene start sites (*M. zebra*—10,654/21,673; *P. nyererei*—10,030/20,611; *A. burtoni*—10,050/23,436; *N. brichardi*—8464/20119) based on BLAT orthology and end sites based on original annotations [18], which was otherwise used for annotating the remaining genes. Based on new annotations, for all 1:1 orthologs where gene expression data is available and there is no overlap of gene bodies, we extracted putative promoter regions, taken as up to 5 kb upstream of the transcription start site (TSS) as per methods above. Using *mafft-7.271* [78], we aligned 1:1 orthologous promoter, cds and protein sequences based on orthogrouping in gene trees (see "Cichlid gene trees"). We estimated the number of nonsynonymous substitutions per nonsynonymous site (d*N*) and synonymous substitutions per synonymous site (d*S*) in the 1:1 protein alignments using the *codeml* program in the PAML-4.9 package [79] for each branch and ancestral node in the species tree. Otherwise, we estimated evolutionary rate for each branch and ancestral node in the species tree at promoter regions and fourfold degenerate sites, using 1:1 promoter and cds alignments in *baseml* and *codeml* programs in the PAML-4.9 package [79], requiring that at least 10% of the alignment contains nucleotides and that at least 100 nucleotides are present for each species.

By using the published "*cichlid-5way.maf*" [18], we categorized pairwise substitutions for all species and intersected with annotated genomics regions (see Additional file 1: Table S2) using *bedtools-2.25.0* intersect [80].

Reconstructing regulatory networks

To infer essential drivers of tissue-specific expression in cichlids, we constructed regulatory and functional interaction/association networks through the integration of several datasets and approaches (Additional file 1: Fig. S20). This approach was largely centered on the integration of expression-based and in silico TFBS motif prediction-based networks.

We first used species- and module-specific gene expression levels to infer an expression-based network. For this, we merged the cichlid gene expression data into a single 30 (five species, six tissues) dimensional dataset to learn cichlid-specific transcription factor (TF)-target gene (TG) interactions using the Per Gene Greedy (PGG) approach, a prior expression-based network inference method [24]. We projected the network into species-specific networks by considering edges that would not be present due to gene loss. We then integrated in silico-predicted TF-TG edges (see "Transcription factor (TF) motif scanning") based on TFBS predictions in gene promoter regions. To ensure accurate analysis of GRN rewiring through an integrative approach, all collated edges were then pruned to ensure edges were (1) not absent in at least one species due to gene loss/poor annotation and (2) based on the presence of genes in co-expression modules.

To maintain a structured and connected network approach, we analyzed network topology using two methods; firstly, and to ensure suitable integration of co-expression data with all TF-TG predicted edges, one set of all gene nodes and their edges were constrained by Arboretum module assignments to correlate to their respective patterns of tissue-specific expression and co-expression module analysis. Secondly, since all included genes will not necessarily exhibit tissue-specific co-expression (and cluster accordingly) due to (1) differences in cell type abundance, (2) cell heterogeneity; and (3) small development stage differences, and as well as despite not being co-expressed, the fact that TFs are trans-acting factors able to regulate any gene, we also analyzed all network edges for selected candidate genes without constraining based on module assignment (co-expression). Accordingly, for candidate genes with rewired networks, we also analyzed network topology without constraining edges based on same module assignment (co-expression) and, instead, analyzed the Pearson correlation coefficient (r) between cross-species significant TF motif enrichment (FDR < 0.05), taken as $-log(q\text{-value})$, in all module genes and expression (zero-mean *log* expression ratio) in each tissue. Similar or dissimilar levels of TF motif enrichment were determined by calculating the variance over each TF motif enrichment, taken as $-log(q\text{-value})$ across the five species, and then by plotting the density distribution of the variance, categorizing TFs in each of the tails into similar or dissimilar fold enrichment (FE).

Functional landscape of reconstructed regulatory networks

We use the FDR-corrected hypergeometric p value to assess enrichment of GO terms for genes in reconstructed networks. We used GO terms for the published five cichlids [18] and carried out enrichment analysis as previously done for Arboretum module genes (see "Methods" above).

Regulatory rewiring analysis of gene sets

Regulatory rewiring of TF-TG interactions is based on predictions derived from TFBS scanning and TF-TG co-expression relationships inferred by the PGG method [24]. To ensure rewiring of TFs are correctly compared between species, and not based on gene loss/poor annotation, we only included edges for analysis where the TF had a 1-to-1 orthologous relationship in species where the TF-TG relationship or non-directed relationship exists. Also, we filtered out any TGs and their TF interaction/relationships if, based on orthologous gene *tblastx* [81], whether the gene was present in the genome but not annotated. Of the 18,799 orthogroups used for generating modules of co-expressed genes and network interactions, 4209 orthogroups had many-to-many genes actually present in the genome of at least one of the five species. These 4209 orthogroups were filtered out, retaining 843,168/1,131,812 predicted TF-TG edges across the five species; in summary, these represent edges that are (1) present in at least two species, (2) not absent in any species due to node loss or mis-annotation; and (3) based on the presence of nodes in modules of co-expression genes. The 843,168/1,131,812 predicted TF-TG edges across the five species were then used for network rewiring analysis.

Three metrics were used to study large-scale TF-TG network rewiring between species that included (1) state changes in module assignment, (2) DyNet [25] network rewiring scores and (3) TF rate of edge gain and loss in networks.

State changes in module assignment

In this metric, a rewired edge is where a unique TF-TG edge is present in only one "focal" species, but the TF ortholog is state changed in module assignment and is a node in other TF-TG edges in any of the other species.

DyNet network rewiring scores

The DyNet-2.0 package [25], implemented in Cytoscape-3.7.1 [82], was used for network visualization and calculation of a degree-corrected rewiring (D_n) score of TF-TG interactions in each orthogroup. The D_n score for each orthogroup was ordered and the mean calculated; the significance of difference of each orthogroups rewiring score against all orthogroups was compared by calculating differences in the standard deviation and applying the non-parametric Kolmogorov–Smirnov test (KS-test).

TF rate of edge gain and loss in networks

Gain and loss rate analyses were similar to that performed previously [10]. This approach uses a continuous-time Markov process parameterized by TF-TG edge gain and loss rates and uses an expectation-maximization (EM)-based algorithm to estimate rates [83, 84]. The input network comprised target genes of 783 individual regulator genes mapped across the five cichlid species based on gene orthology. Each species regulator required a minimum of 25 edges as < 25 edges greatly hinder statistical analysis in this context. This resulted in a total of 345 regulators with 25 to 23,935 edges, with an average of 2609. Gain and loss rate was estimated for each regulator using the EM-based algorithm on the edge gain and loss pattern across the five cichlid phylogeny. Rates were inferred using published five cichlid branch lengths [18] that described neutral sequence evolution across the species. Stability analysis of rate estimations were carried out as follows: (1) gain and loss rate input values were scanned from 0 to 400 in intervals of 5 for each regulator matrix, and (2) from each scan, rates with the greatest likelihood were chosen as the recommended gain and loss rate (< 100), defining a final set of inferred rates for 186/345 regulators.

Identification of segregating sites in TFBSs

Species pairwise variation was identified based on an *M. zebra* v1.1 assembly centered 8-way teleost *multiz* alignment [18]. Pairwise (single-nucleotide) variants were then overlapped with TFBS positions as determined by TF motif scanning using *bedtools-2.25.0* intersect [80]. Pairwise variants of *M. zebra* were overlapped with single-nucleotide polymorphisms (SNPs) in Lake Malawi species [20] using *bedtools-2.25.0* intersect [80]. Both sets of pairwise variants overlapping motifs and lake species SNPs were then filtered based on the presence of the same pairwise variant in orthologous promoter alignments. This ensured concordance of whole-genome alignment-derived variants with variation in orthologous promoter alignments and predicted motifs. At each step, reference and alternative allele complementation was accounted for to ensure correct overlap. This analysis was not to distinguish population differentiation due to genetic structure, but to instead map regulatory variants onto a number of radiating cichlid species to link to phylogenetic and ecological traits.

Phylogenetic independent contrasts

Phylogenetic independent contrasts (PICs) were carried out to statistically test the effect of fitting the least controversial and all included 73 Lake Malawi species phylogeny [20] on the covariance of segregating TFBSs, visual (wavelength palette) and ecological traits (habitat and foraging habit/diet). This involved (1) categorically coding segregating TFBS genotypes (of NR2C2 > *sws1* and GATA2A > *rho*), visual trait and ecological measurements for each of the 73 Lake Malawi species (119 individuals), and (2) using the *ape* package (v5.4.1) in R (v4.0.2) to apply the PICs test [37] on all correlations with the TFBS genotypes (genotype vs wavelength palette, genotype vs habitat, and genotype vs foraging habit/diet). PICs assume a linear relationship and process of Brownian motion between traits, and thus, for each combination of data, a scatterplot was first generated. To test any change in the correlation (due to phylogenetic signal), the regression model was compared between relationships excluding and including the published Lake Malawi phylogeny [20].

Expression of protein DNA-binding domains DBDs)

DNA-binding domains (DBDs) of cichlid proteins (NR2C2 and RXRB) were predicted based on alignment and conservation to annotated human and mouse orthologs. *M. zebra* and *N. brichardi* individuals were sacrificed according to schedule 1 killing using overdose of MS-222 (tricaine) at The University of Hull, UK and University of Basel, Switzerland. Tissues were stored in RNA later using a 1:5 ratio. RNA was extracted from brain, liver, and testis tissues of adult *M. zebra* and *N. brichardi* using the RNeasy Plus Mini Kit (Qiagen), achieving RNA integrity (RIN) in the range of 8–10 (Agilent Bioanalyzer Total RNA Pico Assay). First-strand cDNA synthesis of DBD-specific regions was carried out using RevertAid H Minus Reverse Transcriptase (Thermo Scientific) and DBDs amplified (2-step RT-PCR) using Platinum Taq DNA Polymerase (Invitrogen) and the primers listed in Additional file 1: Table S1. Resulting cDNA was concentrated using Minelute PCR purification (Qiagen) and 700 ng used for in vitro transcription/translation using TnT T7 Quick for PCR DNA (Promega) and the Fluorotect GreenLys tRNA (Promega) labelling system. DBD expression was resolved by SDS-PAGE and detection using the fluorescein filter in the ChemiDoc Touch (Bio-Rad) system.

Electrophoretic mobility shift assay EMSA) validation of predicted TF-TG interactions

EMSA was carried out using double-stranded Cy5 fluorophore 5′-modified (IDT) DNA probes, in vitro expressed DBDs (see above) and the Gel Shift Assay Core System (Promega). Double-stranded DNA probes were generated by annealing sense and antisense oligonucleotides (see Additional file 1: Table S1) in annealing buffer (10 mM Tris pH 7.5, 1 mM EDTA, 50 mM NaCl) for 3 min at 96 °C, 1 min at 90 °C, 1 min at 85 °C, 3 min at 72 °C, 1 min at 65 °C, 1 min at 57 °C, 1 min at 50 °C, 3 min at 42 °C, and 3 min at 25 °C in a PCR thermocycler. Binding reactions were carried out in a final volume of 9 l composed of Gel Shift Binding 5x Buffer (20% glycerol, 5 mM $MgCl_2$, 2.5 mM EDTA, 2.5 mM DTT, 250 mM NaCl, 50 mM Tris-HCl (pH 7.5), 0.25 mg/ml poly (dI-dC)•poly (dI-dC)); 0.01 M of Cy5-dsDNA probe covering the motif and flanking region (28 nt); and either 23 ng (RXRB, 10.42 kDa) or 27 ng (NR2C2, 10.73 kDa) of expressed DBD. For EMSA validation with increasing Nr2c2 DBD concentrations, 1× = 27 ng. For kit controls, 0.01 M of human SP1 DNA probe was combined with 10,000 ng HeLa nuclear extract. Binding reactions were incubated at room

temperature for 20 min. Protein-DNA complexes were resolved on 1 mm NuPAGE 4–12% Bis-Tris polyacrylamide gels (Invitrogen) in 0.5× TBE at 100 V for 60 min. Protein-DNA complexes were detected using the Cy5 filter on the ChemiDoc MP (Bio-Rad) system. Exposure settings were adjusted in Image Lab v6.0.1_build34 (Bio-Rad) with same high (5608), low (1152) and gamma (1.0) values set for all associated images.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-020-02208-8.

Additional file 1 Supplementary analysis notes, figures and tables. This file includes supplementary notes, along with supplementary figs. S1-S22 and Tables S1-S3 referenced in the main text. This file also includes legends for figs and tables in Additional files 2 and 3.

Additional file 2 Extended data figure S1-S8. This file includes extended figures that support the findings of this study, including Fig. S1. GO enrichment of module genes (FDR < 0.05); Fig. S2. Motif enrichment of module genes (FDR < 0.05); Fig. S3. Brain heatmap pearson-correlation; Fig. S4. Eye heatmap pearson-correlation; Fig. S5. Heart heatmap pearson-correlation; Fig. S6. Kidney heatmap pearson-correlation; Fig. S7. Muscle heatmap pearson-correlation; and Fig. S8. Testis heatmap pearson-correlation.

Additional file 3 Large data Tables S1-S6. This file includes extended data tables that support the findings of this study.

Additional file 4 Review history.

Acknowledgments

We thank the BROAD institute and the Cichlid Genome Consortium for providing full access to genomic data. We thank Dr. Domino Joyce (The University of Hull, UK) and Prof. Walter Salzburger (University of Basel, Switzerland) for providing cichlid species tissues for EMSA validations. We would also like to thank Prof. Ole Seehausen and members of his group for providing access to relevant SNP information from Lake Victoria cichlid species. While we did not ultimately publish this analysis, we found it useful for comparative analysis of segregating sites overlapping TFBSs.

Review history

The review history is available as Additional file 4.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors contributions

CK, SAK, and SR constructed gene trees and ran Arboretum and gene ontology (GO) enrichment; TKM, WN, and PS developed and ran transcription factor (TF) motif prediction and enrichment; TKM analyzed co-expression modules, enrichment and breadth of gene expression; WN, TKM, and WH calculated and analyzed evolutionary rates; SAK and SR generated co-expression edges; TKM reconstructed networks and carried out GO enrichment and analyses; SB and LPD analyzed network structure; MO, TKM, and TK analyzed network rewiring; TKM and WN analyzed variants overlapping TFBSs; TKM carried out EMSA; TKM, WH, SR, and FDP wrote the manuscript with input from SAK, WN, PS, SB, and TK. The author(s) read and approved the final manuscript.

Authors information

Twitter handles: @TK_mehta (Tarang K. Mehta); @nashalselection (Will Nash); @martonolbei (Marton Olbei); @BastkowskiSarah (Sarah Bastkowski); @LucaPensoDolfin (Luca Penso-Dolfin); @KorcsmarosLab (Tamas Korcsmaros); @WHaerty (Wilfried Haerty), @sroyyors (Sushmita Roy); @ScienceisGlobal (Federica Di-Palma).

Funding

TKM, WN, PS, LPD, WH, and FDP were supported and the project strategically funded by the Biotechnological and Biosciences Research Council (BBSRC), Institute Strategic Programme BB/J004669/1 and Core Strategic Programme Grants BB/P016774/1; BBS/E/T/000PR9817; and BB/CSP17270/1 at the Earlham Institute. TK was supported by a fellowship in computational biology at Earlham Institute (Norwich, UK) in partnership with the Quadram Institute (Norwich, UK), and strategically funded by BBSRC, UK (BB/J004529/1 and BB/P016774/1). MO was supported by the BBSRC Norwich Research Park Biosciences Doctoral Training Partnership (grant BB/M011216/1). SR was supported by a National Science Foundation (NSF) career award (DBI: 1350677) and with CK and SAK, the McDonnell foundation at The Wisconsin Institute for Discovery.

Availability of data and materials

Cichlid PWMs that support the findings of this study are available in a figshare repository [85].
Datasets relevant to network reconstruction and their outputs are also available in figshare [86–88].
Original, uncropped gel images of EMSA experiments that support the findings of this study are available in figshare [89].
Datasets that are otherwise absent from this published article are available from the corresponding authors upon request.

The source code to run motif prediction and network reconstruction from TFBS and TF-TG co-expression is freely available to all under the Creative Commons Attribution-ShareAlike licence (CC BY-SA) and under the standard GPL 3.0 licence from Github [90].
Otherwise, all other scripts relevant to this published article are available from the corresponding authors on request.

Ethics approval and consent to participate
All animal procedures were approved by the relevant university and carried out in accordance with approved guidelines. *M. zebra* individuals were sacrificed according to Home Office licence schedule 1 killing using overdose of MS-222 (tricaine) at Dr. Domino Joyce lab, The University of Hull. *N. brichardi* individuals were sacrificed according to cantonal veterinary permit nr. 2317 killing using overdose of MS-222 (tricaine) at Prof. Walter Salzburger lab, University of Basel, Switzerland.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Earlham Institute (EI), Norwich, UK. [2]Department of Biostatistics and Medical Informatics, UW Madison, Madison, USA. [3]Wisconsin Institute for Discovery (WID), Madison, USA. [4]Quadram Institute, Norwich, UK. [5]Department of Computer Sciences, UW Madison, Madison, USA. [6]Norwich Medical School, University of East Anglia, Norwich, UK. [7]School of Biological Sciences, University of East Anglia, Norwich, UK.

Received: 23 April 2020 Accepted: 18 November 2020
Published online: 08 January 2021

References
1. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975;188:107–16.
2. Wilson AC, Maxson LR, Sarich VM. Two types of molecular evolution: evidence from studies of interspecific hybridization. Proc Natl Acad Sci U S A. 1974;71:2843–7.
3. Prager EM, Wilson AC. Slow evolutionary loss of the potential for interspecific hybridization in birds: a manifestation of slow regulatory evolution. Proc Natl Acad Sci U S A. 1975;72:200–4.
4. Jacob F. Evolution and tinkering. Science. 1977;196:1161–6.
5. Carroll SB. Endless forms: the evolution of gene regulation and morphological diversity. Cell. 2000;101:577–80.
6. Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell. 2008;134:25–36.
7. Peter IS, Davidson EH. Evolution of gene regulatory networks controlling body plan development. Cell. 2011;144:970–85.
8. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet. 2002;31:64–8.
9. Roy S, Wapinski I, Pfiffner J, French C, Socha A, Konieczka J, et al. Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. Genome Res. 2013;23:1039–50.
10. Koch C, Konieczka J, Delorey T, Lyons A, Socha A, Davis K, et al. Inference and evolutionary analysis of genome-scale regulatory networks in large phylogenies. Cell Syst. 2017;4:543–58.
11. Ichihashi Y, Aguilar-Martinez JA, Farhi M, Chitwood DH, Kumar R, Millon LV, et al. Evolutionary developmental transcriptomics reveals a gene network module regulating interspecific diversity in plant leaf shape. Proc Natl Acad Sci. 2014;111:2616–21.
12. Levine M, Davidson E. Gene regulatory networks for development. Pnas. 2005;102:4936–42.
13. Israel JW, Martik ML, Byrne M, Raff EC, Raff RA, McClay DR, et al. Comparative developmental transcriptomics reveals rewiring of a highly conserved gene regulatory network during a major life history switch in the sea urchin genus Heliocidaris. Plos Biol. 2016;14:e1002391.
14. Yevshin I, Sharipov R, Valeev T, Kel A, Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids Res. 2017;45:D61–7.
15. Pfennig DW, Ehrenreich IM. Towards a gene regulatory network perspective on phenotypic plasticity, genetic accommodation and genetic assimilation. Mol Ecol. 2014;23:4438–40.
16. Genner MJ, Seehausen O, Lunt DH, Joyce DA, Shaw PW, Carvalho GR, et al. Age of cichlids: new dates for ancient lake fish radiations. Mol Biol Evol. 2007;24:1269–82.
17. Wagner CE, Harmon LJ, Seehausen O. Ecological opportunity and sexual selection together predict adaptive radiation. Nature. 2012;487:366–9.
18. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, et al. The genomic substrate for adaptive radiation in African cichlid fish. Nature. 2014;93:17–9.
19. Kratochwil CF, Meyer A. Evolution: tinkering within gene regulatory landscapes. Curr Biol. 2015;25:R285–8.
20. Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, et al. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat Ecol Evol. 2018;2:1940–55.
21. Bloomquist RF, Fowler TE, Sylvester JB, Miro RJ, Streelman JT. A compendium of developmental gene expression in Lake Malawi cichlid fishes. BMC Dev Biol. 2017;17:3.
22. Browman HI, Hawryshyn CW. Retinoic acid modulates retinal development in the juveniles of a teleost fish. J Exp Biol. 1994;193:191–207.
23. Takechi M, Seno S, Kawamura S. Identification of cis-acting elements repressing blue opsin expression in zebrafish UV cones and pineal cells. J Biol Chem. 2008;283:31625–32.

24. Siahpirani AF, Roy S. A prior-based integrative framework for functional transcriptional regulatory network inference. Nucleic Acids Res. 2017;45:2221.
25. Goenawan IH, Bryan K, Lynn DJ. DyNet: visualization and analysis of dynamic molecular interaction networks. Bioinformatics. 2016;32:2713–5.
26. Sylvester JB, Rich CA, Yi C, Peres JN, Houart C, Streelman JT. Competing signals drive telencephalon diversity. Nat Commun. 2013;4:1745.
27. Fraser GJ, Bloomquist RF, Streelman JT. Common developmental pathways link tooth shape to regeneration. Dev Biol. 2013;377:399–414.
28. Whited JL. Dynactin is required to maintain nuclear position within postmitotic Drosophila photoreceptor neurons. Development. 2004;131:4677–86.
29. Carleton K. Cichlid fish visual systems: mechanisms of spectral tuning. Integr Zool. 2009;4:75–86.
30. Kocher TD. Adaptive evolution and explosive speciation: the cichlid fish model. Nat Rev Genet. 2004;5:288–98.
31. Henning F, Meyer A. The evolutionary genomics of cichlid fishes: explosive speciation and adaptation in the postgenomic era. Annu Rev Genomics Hum Genet. 2014;15:417–41.
32. Peng YR, Tran NM, Krishnaswamy A, Kostadinov D, Martersteck EM, Sanes JR. Satb1 regulates Contactin 5 to pattern dendrites of a mammalian retinal ganglion cell. Neuron. 2017;95:869–83.
33. Evans RM, Mangelsdorf DJ. Nuclear receptors, RXR, and the big bang. Cell. 2014;157:255–66.
34. Medina CF, Mazerolle C, Wang Y, Bérubé NG, Coupland S, Gibbons RJ, et al. Altered visual function and interneuron survival in Atrx knockout mice: inference for the human syndrome. Hum Mol Genet. 2009;18:966–77.
35. Hofmann CM, O'Quin KE, Justin Marshall N, Cronin TW, Seehausen O, Carleton KL. The eyes have it: regulatory and structural changes both underlie cichlid visual pigment diversity. Plos Biol. 2009;7:e1000266.
36. Froese R, Pauly D. Fishbase. FishBase. 2017; Available from: www.fishbase.org. Accessed 28 Aug 2019.
37. Felsenstein J. Phylogenies and the comparative method. Am Nat. 1985;125:1–15.
38. McAdams HH, Srinivasan B, Arkin AP. The evolution of genetic regulatory systems in bacteria. Nat Rev Genet. 2004;5:169–78.
39. Thompson DA, Roy S, Chan M, Styczynski MP, Pfiffner J, French C, et al. Evolutionary principles of modular gene regulation in yeasts. Elife. 2013;2:e00603.
40. Wittkopp PJ, Haerum BK, Clark AG. Regulatory changes underlying expression differences within and between Drosophila species. Nat Genet. 2008;40:346–50.
41. Yanai I, Hunter CP. Comparison of diverse developmental transcriptomes reveals that coexpression of gene neighbors is not evolutionarily conserved. Genome Res. 2009;19:2214–20.
42. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484:55–61.
43. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;338:1587–93.
44. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478:343–8.
45. Chen J, Swofford R, Johnson J, Cummings BB, Rogel N, Lindblad-Toh K, et al. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. Genome Res. 2019;29:53–63.
46. Salzburger W. Understanding explosive diversification through cichlid fish genomics. Nat Rev Genet. 2018;19:705–17.
47. Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, et al. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc Natl Acad Sci U S A. 2018;115:6249–54.
48. York RA, Byrne A, Abdilleh K, Patil C, Streelman T, Finger TE, et al. Behavioral evolution contributes to hindbrain diversification among Lake Malawi cichlid fish. Sci Rep. 2019;9:19994.
49. O'Quin KE, Hofmann CM, Hofmann HA, Carleton KL. Parallel evolution of opsin gene expression in African cichlid fishes. Mol Biol Evol. 2010;27:2839–54.
50. Carleton KL, Hárosi FI, Kocher TD. Visual pigments of African cichlid fishes: evidence for ultraviolet vision from microspectrophotometry and DNA sequences. Vis Res. 2000;40:879–90.
51. Carleton KL, Spady TC, Streelman JT, Kidd MR, McFarland WN, Loew ER. Visual sensitivities tuned by heterochronic shifts in opsin gene expression. BMC Biol. 2008;6:22.
52. Sandkam BA, Campello L, O'Brien C, Nandamuri SP, Gammerdinger W, Conte M, et al. Tbx2a modulates switching of RH2 and LWS opsin gene expression. Mol Biol Evol. 2020;37:2002–14.
53. Seehausen O, Terai Y, Magalhaes IS, Carleton KL, Mrosso HDJ, Miyagi R, et al. Speciation through sensory drive in cichlid fish. Nature. 2008;455:620–6.
54. Fraser GJ, Hulsey CD, Bloomquist RF, Uyesugi K, Manley NR, Streelman JT. An ancient gene network is co-opted for teeth on old and new jaws. Plos Biol. 2009;7:0233–47.
55. Santos ME, Baldo L, Gu L, Boileau N, Musilova Z, Salzburger W. Comparative transcriptomics of anal fin pigmentation patterns in cichlid fishes. BMC Genomics. 2016;17:712.
56. Burmeister SS, Jarvis ED, Fernald RD. Rapid behavioral and genomic responses to social opportunity. PLoS Biol. 2005;3:1996–2004.
57. Nandamuri SP, Conte MA, Carleton KL. Multiple trans QTL and one cis-regulatory deletion are associated with the differential expression of cone opsins in African cichlids. BMC Genomics. 2018;19:945.
58. Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitxl enhancer. Science. 2010;327:302–5.
59. York RA, Patil C, Abdilleh K, Johnson ZV, Conte MA, Genner MJ, et al. Behavior-dependent cis regulation reveals genes and pathways associated with bower building in cichlid fishes. Proc Natl Acad Sci U S A. 2018;115:E11081–90.
60. Kratochwil CF, Liang Y, Gerwin J, Woltering JM, Urban S, Henning F, et al. Agouti-related peptide 2 facilitates convergent evolution of stripe patterns across cichlid fish radiations. Science. 2018;362:457–60.
61. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
62. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.

63. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011;12:R22.
64. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
65. Wu Y-C, Rasmussen MD, Bansal MS, Kellis M. TreeFix: statistically informed gene tree error correction using species trees. Syst Biol. 2013;62:110–20.
66. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2017;46:D260–6.
67. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. 2015;43:D117–22.
68. Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, et al. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Res. 2013;41:D195–202.
69. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43:W39–49.
70. Marshall H, Studer M, Popperl H, Aparicio S, Kuroiwa A, Brenner S, et al. A conserved retinoic acid response element required for early expression of the homeobox gene Hoxb-1. Nature. 1994;370:567–71.
71. Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, et al. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. Proc Natl Acad Sci U S A. 1995;92:1684–8.
72. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.
73. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, et al. RSAT 2015: regulatory sequence analysis tools. Nucleic Acids Res. 2015;43:W50–6.
74. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27:1017–8.
75. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 2005;21:650–9.
76. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Brief Bioinform. 2017;18:205–14.
77. Kent WJ. BLAT - the BLAST-like alignment tool. Genome Res. 2002;12:656–64.
78. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
79. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.
80. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
81. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
82. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.Js: a graph theory library for visualisation and analysis. Bioinformatics. 2015;32:309–11.
83. Hobolth A, Jensen JL. Statistical inference in evolutionary models of DNA sequences via the EM algorithm. Stat Appl Genet Mol Biol. 2005;4:1–22.
84. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25(12):i54–i62.
85. Mehta T, et al. Motifs. figshare; 2020. https://doi.org/10.6084/m9.figshare.7599293.v1.
86. Mehta, Tarang *et al.* Network Reconstruction files. figshare. 2020 doi: https://doi.org/10.6084/m9.figshare.7707437.v1.
87. Mehta, Tarang *et al.* Edge_attribute_file_RewiringAnalysis_file. figshare. 2020 doi: https://doi.org/10.6084/m9.figshare.7707455.v1.
88. Mehta, Tarang *et al.*TF motif scanning outputs. figshare. 2020 doi: https://doi.org/10.6084/m9.figshare.7712423.v1.
89. Mehta, Tarang *et al.* Original EMSA gel images. figshare. 2020. doi: https://doi.org/10.6084/m9.figshare.13221212.v1.
90. Mehta, Tarang *et al.* Gene regulatory network reconstruction of five cichlid species (*M. zebra*, P. nyererei, *A. burtoni*, *N. brichardi* and *O. niloticus*). GitHub. 2020. https://github.com/TGAC/Cichlid-GRNs. Accessed 13 Nov 2020.

Publisher s Note

npj | Systems Biology and Applications

www.nature.com/npjsba

TECHNOLOGY FE TURE OP N

SalmoNet an integrated network of ten *almonella enterica* strains reveals common and distinct pathways to host adaptation

Aline Métris[1,5], Padhmanand Sudhakar[1,2], David Fazekas[2,3], Amanda Demeter[1,2,3], Eszter Ari[3,4], Marton Olbei[1,2], Priscilla Branchu[1,6], Rob A. Kingsley[1], Jozsef Baranyi[1] and Tamas Korcsmáros[1,2]

Salmonella enterica is a prominent bacterial pathogen with implications on human and animal health. *Salmonella* serovars could be classi ed as gastro-intestinal or extra-intestinal. Genome-wide comparisons revealed that extra-intestinal strains are closer relatives of gastro-intestinal strains than to each other indicating a parallel evolution of this trait. Given the complexity of the differences, a systems-level comparison could reveal key mechanisms enabling extra-intestinal serovars to cause systemic infections. Accordingly, in this work, we introduce a unique resource, SalmoNet, which combines manual curation, high-throughput data and computational predictions to provide an integrated network for *Salmonella* at the metabolic, transcriptional regulatory and protein-protein interaction levels. SalmoNet provides the networks separately for ve gastro-intestinal and ve extra-intestinal strains. As a multi-layered, multi-strain database containing experimental data, SalmoNet is the rst dedicated network resource for *Salmonella*. It comprehensively contains interactions between proteins encoded in *Salmonella* pathogenicity islands, as well as regulatory mechanisms of metabolic processes with the option to zoom-in and analyze the interactions at speci c loci in more detail. Application of SalmoNet is not limited to strain comparisons as it also provides a *Salmonella* resource for biochemical network modeling, host-pathogen interaction studies, drug discovery, experimental validation of novel interactions, uncovering new pathological mechanisms from emergent properties and epidemiological studies. SalmoNet is available at http://salmonet.org.

npj Systems Biology and Applications (2017)3:31 ; doi:10.1038/s41540-017-0034-z

INTRODUCTION

The genus *Salmonella* includes pathogens associated with syndromes ranging from gastroenteritis to bacteraemia and enteric fever.[1] Gastroenteritis caused by *Salmonella* is one of the most common foodborne diseases, with nearly 100 million cases per year occurring worldwide.[2] While enteric fever is rare in developed countries, it is still associated with signi cant mortality and morbidity in low income countries with over 90,000 deaths worldwide in 2015.[3]*Salmonella* pathogenesis depends on a large number of virulence genes including those located on large pathogenicity islands encoding type III secretion systems that translocate effector proteins into the host cell cytoplasm.[4] *S. enterica* subspecies includes over 1500 different serovars and accounts for the vast majority of human infections.[5] Based on the epidemiological record, disease symptoms and observations from experimental infections has resulted in the classi cation of serovars into two pathovars, namely gastro-intestinal and extra-intestinal. Most serovars of subspecies I are of the former pathovar and most often associated with gastro-intestinal infections. Serovars of gastro-intestinal pathovars often exhibit a broad host range. However, a small number of serovars are host-adapted and are characterized by an extra-intestinal infection and dissemination beyond the intestinal mucosa followed by colonization of systemic sites of the reticuloendothelial system. As most serovars are of the gastro-intestinal pathovar type, the most parsimonious explanation for the extra intestinal serovars is that they evolved from a gastro-intestinal pathovar ancestor, most likely on multiple occasions. The molecular basis of host adaptation has been studied most extensively in *S. enterica* serovar Typhi (*S.* Typhi), the causative agent of typhoid fever. Adaptation of *S.* Typhi is characterized by the acquisition of a number of virulence associated genes and the loss of coding capacity affecting over 200 genes.[6,7]

Genes horizontally acquired by *S.* Typhi include a large pathogenicity island (SP-7) encoding biosynthesis genes for the Vi polysaccharide capsule and the TviA regulator protein,[6] and the typhoid toxin that is encoded outside SPI-7.[8] Dissemination beyond the intestinal mucosa is in part mediated by evasion of detection by the host innate immune system by expression of the Vi polysaccharide capsule,[9] and by the down-regulation of flagella expression, a pathogen associated molecular pattern (PAMP), mediated by TviA.[10] The function of the typhoid toxin in pathogenesis is not clear, however many of the symptoms of

[1]Quadram Institute Bioscience, Norwich Research Park, Norwich NR4 7UA, UK; [2]Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK; [3]Department of Genetics, Eötvös Loránd University, Pázmány P. s. 1C, H-1117 Budapest, Hungary and [4]Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, Szeged, Hungary
Correspondence: Tamas Korcsmáros (Tamas.Korcsmaros@earlham.ac.uk)
[5]Present address: Safety and Environmental Assurance Centre, Unilever, Colworth Science Park, Sharnbrook, Bedfordshire, UK
[6]Present address: IRSD, Université de Toulouse, INSERM, INRA, ENVT, UPS, Toulouse, France
Aline Métris, Padhmanand Sudhakar, and David Fazekas contributed equally to this work.

Received: 6 May 2016 Revised: 19 September 2017 Accepted: 22 September 2017
Published online: 18 October 2017

Published in partnership with the Systems Biology Institute

npj nature partner journals

Table 1 Information about the numbers corresponding to the data sources and the reconstructed networks for the reference strain *Salmonella Typhimurium* LT2

Network type	Data source		Number of interactions in *S. Typhimurium* LT2
Metabolic	Model validated by flux-balance analysis[82]		2312
	BioModel database[75]		754
Regulatory	Experimental evidence in *Salmonella*	Manual curation of low-throughput experiments	9
		Datasets containing high-throughput experiments	234
	Genome-wide predictions	Based on experimentally veri ed binding sites in *Salmonella*	1189
		Based on *E. coli* binding sites from RegulonDB[17]	1865
PPI	Experimental evidence in *Salmonella*	Manual curation	27
		IntAct database[72]	29
	Proteome-wide predictions	Structure based predictions using the Interactome 3D resource[96]	290
		Orthology based predictions using *E. coli* PPI data from,[98] IntAct[72] and BioGRID[97]	1846

typhoid fever were induced by injection of the typhoid toxin into mice.[11]

However, many of the extra-intestinal serovars of *Salmonella* do not encode SPI-7 or the typhoid toxin. Therefore, alternative mechanisms for the systemic dissemination are likely to have evolved in these serovars. This reflects the convergent evolution to an extra-intestinal lifestyle reflected in the phylogenetic relationship of these pathogens. Convergence in genome sequence polymorphisms of extra-intestinal serovars of *S. enterica* has previously been observed in the form of loss of coding capacity (genome degradation) due to deletions and pseudogene formation.[7] Degradation of coding sequences corresponding to genes associated with the intestinal phase of infection such as anaerobic metabolism, motility and chemotaxis, and enteropathogenesis was over-represented in these serovars. A similar pattern of genome degradation was also observed in a rapidly evolving hypermutator strain of *S. Enteritidis* that was restricted to a systemic site niche in an immunocompromised patient.[12]

Serovars of *S. enterica* subspecies I exhibit divergence in their nucleotide sequences that corresponds to approximately 737,000 SNPs.[13] In some cases, non-synonymous SNPs alter the function of proteins, and may alter the function of non-coding sequences when the SNPs are present in regulatory sequences or small RNAs. Serovars also encode hundreds of serovar-speci c genes, as well as contain varying degrees of genome degradation that result in considerable differences in coding capacities and gene expression regulation. In light of this complexity, there is a need to apply a systems biology approach to compile network information across the metabolic, transcriptional regulatory and protein-protein interaction (PPI) layers in order to address the hypothesis that extra-intestinal serovars exhibit convergence in molecular mechanisms of host adaptation. Integration of interaction information from multiple layers is expected to provide insights into the shared attributes that characterize *Salmonella* pathogenicity and virulence.

In order to gain further insight into how *Salmonella* host adaptation has evolved there is a need to integrate different levels of knowledge (e.g., metabolism and regulation) as current data resources store the different layers separately, making complex analysis dif cult. While a substantial amount of data on regulation, metabolism and protein-protein interactions is available, much of this is curated for model organisms, such as *Escherichia coli*. Integrating different types of information into a complex network remains a challenge for non-model organisms, like *Salmonella*.[14] For example, in the case of transcriptional regulation, widely used resources such as ORegAnno,[15] PAZAR[16], or RegulonDB[17] do not contain information on *Salmonella*. Other resources such as KEGG[18] provides information only on metabolic pathways, and even these reactions are not direct *Salmonella* connections but orthology based inferences using *E. coli*. Furthermore, there are no resources that combine curated and predicted interaction information for *Salmonella*. Thus existing resources are either not comprehensive enough to capture multi-layer information or do not contain *Salmonella* speci c interaction data.

We therefore compiled the metabolic, regulatory and PPI networks of 10 representative strains of *Salmonella* comprising 5 gastro-intestinal and 5 extra-intestinal strains. In addition to the interactions corresponding to the manually curated information speci c to *Salmonella* pathogenicity islands, the networks also contain regulatory interactions derived from high-throughput experiments and whole-genome motif scans apart from interactions inferred from *E. coli* by orthology.

RESULTS AND DISCUSSION

Networks, data representation, and quality control

In this study, we have established a workflow to collect and infer interaction information from three different network levels (metabolic, regulatory and PPI) based on various sources (Table 1). We used data derived from literature, online databases, as well as genome-wide predictions. To further enrich the dataset, we performed this on the whole genome sequence assemblies of a range of *Salmonella* strains (Table 2) representing two pathovars (gastro-intestinal and extra-intestinal) that exhibit distinct life styles. The resulting networks are available to the scienti c community for further analysis and enhancement at http://salmonet.org. The networks contain three layers (metabolic, regulatory and PPI) for 5 gastro-intestinal and 5 extra-intestinal strains (Supplementary Table 1).

The SalmoNet database consists of a total of 81,514 interactions involving 30,870 genes across the studied strains (see the strain speci c distribution in Table 3). Considering all the annotated genes for the strains (49,472 genes), SalmoNet therefore covers 62% of the coding capacity of all the strains. In terms of the number of individual ortholog groups, SalmoNet contains information on the interacting partners of 132 sets of transcription factors (TFs) in the regulatory network, 1282 sets of proteins in the PPI network, as well as information on 1196 sets of enzymes in the

Table 2 Strains included in the study and their life-style

Serovar	Strain	Lifestyle	N.p[a]	Genome assembly ID[b]
Typhi	CT18	Extra-intestinal, causes typhoid fever in humans	2	000195995.1
Paratyphi	ATCC 9150	Extra-intestinal, second most prevalent cause of typhoid fever	0	000011885.1
Choleraesuis	SC-B67	Extra-intestinal, causes swine paratyphoid	2	000008105.1
Dublin	CT 02021853	Extra-intestinal, bovine-adapted serovar	1	000020925.1
Gallinarum	287/91	Extra-intestinal, causative agent of fowl typhoid in poultry	0	000009525.1
Agona	SL483	Gastro-intestinal	1	000020885.1
Enteritidis PT4	P125109	Gastro-intestinal	1	000009505.1
Heidelberg	SL476	Gastro-intestinal	2	000020705.1
Newport	SL254	Gastro-intestinal	2	000016045.1
Typhimurium	SL1344	Gastro-intestinal	3	000210855.2
Typhimurium	LT2	Gastro-intestinal reference strain closely related to SL1344)	1	000006945.1

[a]*N.p.* number of plasmids
[b] GenBank database http://www.ncbi.nlm.nih.gov/genbank/)

Table 3 Number of genes/proteins and their interactions from the networks for the different *Salmonella* strains

Strain name	Number of proteins	Metabolic network		Regulatory network		PPI network	
		nodes	links	nodes	links	nodes	links
Typhi	4718	1121	2348	1710	2595	1235	1949
Choleraesuis	4607	1137	2390	1650	2913	1223	1953
Dublin	4606	1170	2542	1583	2735	1247	2036
Gallinarum	3943	1140	2432	1484	2628	1200	1924
Paratyphi	4083	1136	2380	1565	2692	1202	1923
Agona	4592	1182	2584	1652	2845	1235	1978
Enteritidis	4192	1206	2653	1680	2921	1266	2062
Heidelberg	4757	1187	2590	1638	2736	1266	2072
Newport	4784	1189	2611	1582	2766	1256	2055
Typhimurium SL1344	4657	1228	2762	1735	3107	1287	2068
Typhimurium LT2	4533	1227	2763	1794	3288	1352	2213
Average	4497	1175	2550	1643	2838	1251	2021

metabolic network. Of the total 6070 unique connections in the regulatory network, 16% were present in all the 10 strains (Supplementary Figure 1) spanning the gastro-intestinal and extra-intestinal pathovars, thus comprising a core subset of the *Salmonella* regulatory networks inferred from our workflow (Fig. 1). The edges in this core are those that represent higher con dence since they follow the principle of cross-strain conservation.[19] This methodology has previously been used as a basis for minimizing false positives. The ratio for PPIs and metabolic interactions present in all the 10 strains were higher: 72.6% and 69.2%, respectively. Variation in the fraction of each network represented by the core in regulatory and PPI/metabolic networks supports the idea that transcriptional regulation evolves at a faster rate than the PPI or metabolic levels,[20] although the noise arising from the heterogeneous sources used for the reconstruction of the regulatory network cannot be ignored. The use of the *matrix quality* tool[21] to determine customized *P*-values for every TF-strain combination for the transcriptional regulatory (binding site) predictions minimizes the high false positive rates which could otherwise arise from using generic *P*-value thresholds. Due to the low number of true positive sets, Precision-Recall calculations could not be inferred for most of the transcription factors analyzed in the study. However, to exemplify the reliability of the networks, we determined the target Recall rates (recovery of known targets) for one of the transcription factors SsrB. SsrB regulates the expression of multiple target genes including a number of virulence factors including members of the *Salmonella* pathogenicity islands.[22,23] 24 instances of the 18 bp SsrB binding motif in *S. Typhimurium* SL1344 have been reported,[23] By performing a random and equal bifurcation of the known binding sites into test and training sets, we were able to achieve recall rates of 75% in the reconstructed regulatory network for *S. Typhimurium SL1344* (Supplementary Table 2). Furthermore, swapping the test and training sets yielded a recall rate of 64% suggesting that the reconstructed networks are robust in terms of recovery of true positives. With this example, we show that the predicted regulatory interactions in SalmoNet recover previously reported binding sites due to the employed stringencies such as an informed *P*-value. Besides, users can further enhance the networks by choosing only those interactions which occur in multiple strains of each serovar or all the strains in the study depending on their use case.

The scienti c community can access the database via the aforementioned dedicated web resource in which the molecular entities can be searched by their gene names, UniProt accession IDs, and locus tags. The original source of the interactions that was used during the data integration workflow is also displayed. To enable the comparison of interactions across strains, the ortholog clustering IDs (generated during this study) are also provided for individual molecular entities. An easy-to-use option is provided for users to download selected interaction sets from particular layers of the network or for particular strains. The core *Salmonella* network (the set of interactions conserved across all the strains) can also be accessed similarly. The les are available for download both in CSV and Cytoscape formats allowing straightforward further ltering and visualization, respectively.

Network dendrograms for comparison among strains

To determine the evolutionary relatedness of the ten *Salmonella* strains in SalmoNet, we constructed a phylogenetic tree based on their nucleotide sequences. All the polymorphic sites found in the common ortholog genes of all the strains were used to build a Bayesian dendrogram (Fig. 2a). Four of the gastro-intestinal strains (*S. Typhimurium* LT2, *S. Typhimurium* SL1344, *S.* Heidelberg SL476 and *S.* Newport SL254) were monophyletic on the polymorphic site based phylogenetic tree but two of them were clustered together with extra-intestinal strains: *S.* Enteritidis P125109 with

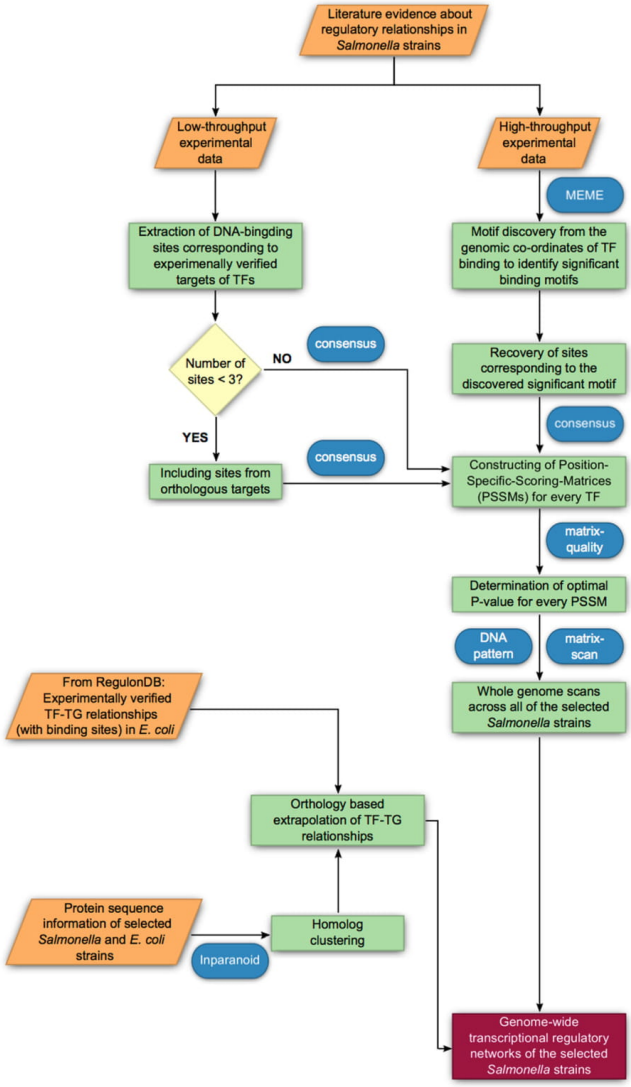

Fig 1 Workflow depicting the steps followed in the reconstruction of the transcriptional regulatory networks of the *Salmonella* strains

S. Gallinarum 287 91; and *S.* Agona SL483 with *S.* Typhi CT18 and *S.* Paratyphi A ATCC 9150. The tree was constructed assuming an approximately equal rate of mutation in each strain, and based on this assumption, the common ancestor of these strains is central within the genome based tree. Strains from extra-intestinal and gastro-intestinal serovars could not be distinguished based on the topology of the genome based dendrogram as observed in previous studies.[24] This is consistent with these pathovars independently emerging as extra-intestinal pathogens and that their attributes arise multiple times during evolution by a process of convergence in genome degradation in the anaerobic metabolism as also described by Nuccio et al.[7]

Next, we compared the phylogenetic relationship of the extra-intestinal and gastro-intestinal pathovars with their metabolic, regulatory and PPI networks to determine if network adaptations converge or if they reflect the evolutionary history of the strains. We used the matrix representation of the networks to infer Bayesian trees corresponding to hierarchical classi cations

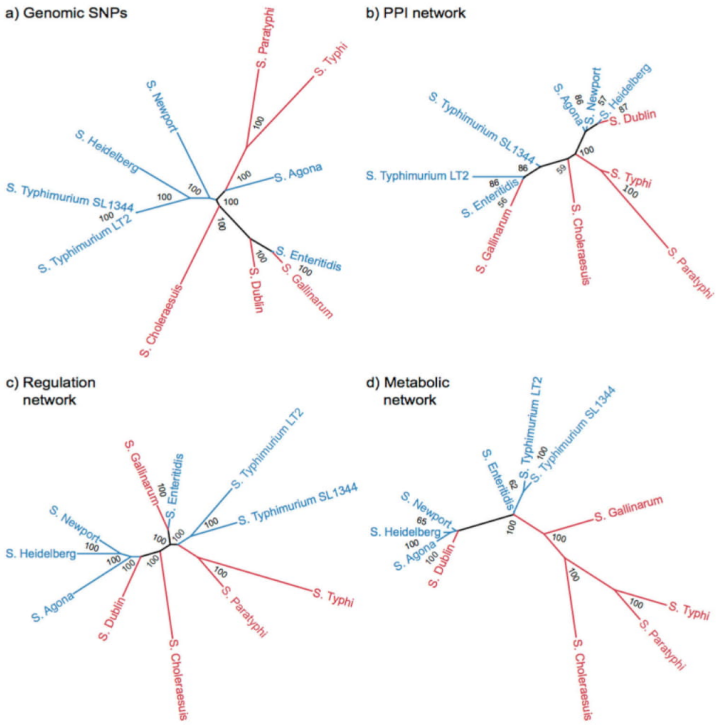

Fig 2 Genome-based phylogenetic tree and hierarchical classi cation of networks. To distinguish different serovar types, gastro-intestinal serovars were colored to blue and extra-intestinals to red. Posterior probability values as percentages) are shown on each node. **a** Bayesian phylogenetic tree based on the polymorphic sites of all common genes. **b d** Hierarchical classi cation trees based on the matrix representation of protein-protein interaction networks **b**, regulatory networks **c**, and metabolic networks **d**. We note that four strains Heidelberg, Agona, Newport and Dublin) form a cluster in all the three network based dendrograms due to technical reasons see details in the main text)

(Fig. 2b–d). The dendrograms for each network were in all cases well established with nearly all posterior probability percentages at the nodes greater than 85. However, none of the networks resulted in the clustering of the extra-intestinal and gastro-intestinal strains. The hierarchical classi cation based on the metabolic networks separate the two pathovar types most pronouncedly, with only the *S.* Dublin metabolic network exhibiting greater similarity to gastro-intestinal pathovars than extra-intestinal pathovars. This is consistent with the loss of shared metabolic pathways that can be dispensed with by all extra-intestinal pathovars that have in common the loss of intestinal colonization as a mode of pathogenesis. The loss of metabolic pathways associated with the intestinal phase of infection is relatively strongly indicated. There is no evidence for loss of PPIs and regulatory sub-networks in the extra-intestinal pathovars. This could reflect the absence of changes in these networks in response to the dispensing of functions required speci cally for the intestinal phase of infection or changes to these networks associated with adaptation to the extra-intestinal environment may be distinct in each extra-intestinal pathovar with weak or absent convergence of mechanisms.

We note that four strains (Heidelberg, Agona, Newport, and Dublin) form a cluster in all the three network based dendrograms (see details in the main text). This is most likely due to the absence of particular genes having interactions to some key genes not present in the data sources used in our pipeline. As SalmoNet only contains genes with interaction data, if one of the interacting pair is missing, and the other interactor has no other interactions, SalmoNet does not contain that particular gene. For these four strains, this methodological limitation resulted in leaving out 31 genes, and because of that, these strains were clustered together.

Functional enrichment analysis of regulons point to pathovar speci c transcription factor functions

Host adapted extra-intestinal pathovars are exposed to distinct host environments and conditions compared to the gastro-intestinal counterparts which are con ned to the environment of the intestinal lumen and mucosa. Evolution to this alternative lifestyle was likely accompanied by plasticity in the regulation of functions in the extra-intestinal pathovars. Extra-intestinal pathovars mediating systemic infections are associated with increased severity and distinct pathogenicity.[25–29] Enrichment analysis of the putative regulons revealed regulation of different functional processes in the two pathovar types by the same orthologous transcription factor (Fig. 3a–b). For instance, the virulence modulating regulator (CsgD) is known to control the expression of various pathogenicity related genes which are important for

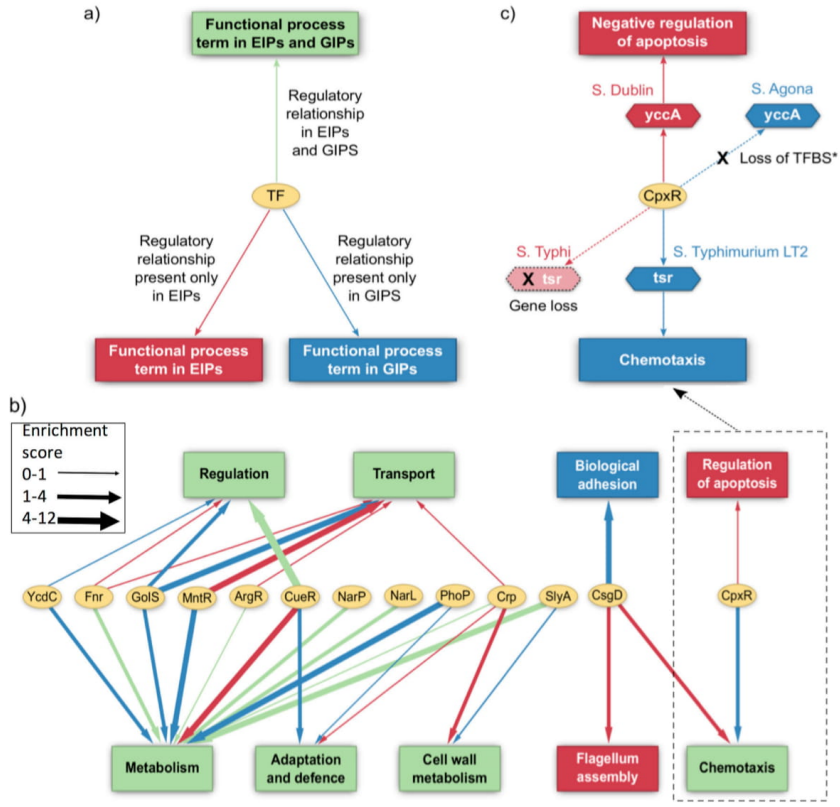

Fig 3 Network of pathovar speci c enriched functions and transcription factors. **a** Network legend for the gure. **b** Graphical representation of the functional processes predicted to be commonly and differentially modulated by orthologous transcription factors TFs) in extra-intestinal and gastro-intestinal pathovars. **c** A speci c example, enlarged from **b** demonstrating the loss in gastro-intestinal and extra-intestinal pathovars of regulatory relationships between cpxR and genes involved in the negative regulation of apoptosis and chemotaxis respectively. *TF* transcription factor, *TFBS* transcription factor binding site

virulence, persistence and bio lm formation.[30,31] In our analysis, the set of genes putatively regulated by CsgD were found to be enriched with the speci c functional process of 'Biological adhesion in gastro-intestinal pathovars. However, the functional process of 'Flagellum assembly was over-represented among the putative CsgD targets in extra-intestinal pathovars while 'Chemo-taxis was over-represented in both extra-intestinal and gastro-intestinal pathovars representing speci c differences and com-monalities in the role of CsgD between the two pathovars. Similarly, comparative analysis of the CpxR regulons revealed pathovar-speci c enrichment of functions within the regulons. CpxR encodes a cognate response regulator and forms part of the CpxAR two component system involved in the sensing of and response to various cell envelope stresses and stimuli.[32,33] In accordance with already known information that CpxR regulates motility and chemotaxis,[34] the set of putative targets of the CpxR regulon in gastro-intestinal pathovars was enriched with the functional process of chemotaxis. In the extra-intestinal pathovars, however, the functional process of regulation of apoptosis was found to be over-represented as a result of distinct cis-regulatory differences. For example, the gene encoding the YccA protein in the extra-intestinal serovar harbored a CpxR binding site in its promoter region while the YccA ortholog in the gastro-intestinal serovar was observed to have a complete loss of the CpxR binding site due to truncation of its promoter (Supplementary Figure 2, Fig. 3c). YccA is homologous to the human anti-apoptotic protein BI-1 which represses the activity of the tumor suppressor protein Bax.[35] Due to the high conservation between *E. coli* YccA and the human BI-1, the YccA protein was even able to protect yeast cells against apoptosis induced by ectopically expressed human Bax protein[36] thus suggesting that YccA is associated with the modulation of host apoptosis. Other apoptotic related factors regulated by CpxR include genes or their orthologs encoding proteins such as the periplasmic serine endoprotease DegP/HtrA,[37] Hemolysin expression-modulating protein Hha,[38] and the toxicity modulator TomB with which Hha forms a putative toxin–antitoxin pair.[39] CpxR also modulates the expression of members of two distinct classes of proteins namely porins (such OmpF) and chemotaxis related proteins (such as CheA, CheW) which are known to modulate apoptosis in the host cell upon infection.[40,41]

Apoptosis of macrophages is a common host response once Salmonella has established an infection systemically but this is tightly regulated as the over-induction of apoptosis is detrimental to Salmonella.[42] Hence, given that extra-intestinal pathovars cause systemic infections, it may be bene cial for the extra-intestinal pathovars to down-regulate apoptosis. This is one possible explanation for the over-representation of apoptosis regulation genes within the CpxR regulon in extra-intestinal pathovars and could indicate that CpxR plays a role in modulating apoptosis in extra-intestinal pathovars. The importance of CpxR in extra-intestinal pathovars is also demonstrated by studies which point out the use of CpxR as a safe and effective vaccine candidate against fowl typhoid caused by *Salmonella* Gallinarum, an extra-intestinal serovar. The results from the compositional analysis of the regulons indicate that the regulons of the two pathovars may have evolved to adapt to their respective pathogenic niches. The differences with respect to the functional processes regulated by *Salmonella* transcription factors could essentially be due to the expected adaptive evolution of extra-intestinal pathovars in contrast to the gastro-intestinal pathovars, which are mostly con ned to the gut.

Applications of the database

The molecular interactions forming the biological interface between *Salmonella* and its host play a signi cant role in the colonization and infection process. *Salmonella* pathogenesis depends on its ability to adhere to host epithelial cells and the Type III secretion system mediated injection of effector proteins, which then cause the re-arrangement of the host cell cytoskeleton and internalization.[43–46] *Salmonella* resides, survives and multiplies in specialized membrane bound vacuoles.[47,48] Various genes known to be involved in *Salmonella* virulence and pathogenesis have been implicated in the interactions of *Salmonella* with the host cells.[43–46] From the regulatory networks in SalmoNet, the transcription factors, which potentially regulate the virulence genes whose products are involved in the interactions with the host, can be identi ed. Moreover, by merging the regulatory networks with an increasing number of datasets such as the PPIs between *Salmonella* and the human host,[49] it is expected to enhance our understanding of the increasing number of mechanisms employed by *Salmonella* to infect and survive inside host cells.

Although the PPI and metabolic networks were inferred by orthological extrapolation, the original sources of data from which the extrapolation was performed are reliable due to their experimental basis even though the source data for the PPI networks were derived from high-throughput techniques. Hence, given the lack of PPI data for these *Salmonella* strains in this study, the inferred PPI and metabolic networks can be considered as a primary starting point for hypothesizing and uncovering new mechanisms. The PPIs are very interesting in this regard since previous studies have shown that *Salmonella* modulates many post-translational modi cations such as ubiquitination of host proteins in order to avoid host responses such as autophagy.[50] The multi-layered nature of the SalmoNet resource can also be exploited in order to uncover potential biological insights by which *Salmonella* subverts host mechanisms. Recent evidence points to the modulation of metabolism (both of the *Salmonella* and the host) as yet another mechanism employed by *Salmonella* to acquire nutrients, evade host defense and survive under harsh intracellular conditions.[51–59]

An integrative analysis of the regulatory and metabolic networks has the potential to reveal new insights into the transcriptional regulatory modulation of metabolic enzymes and could identify new metabolic drug targets as an intervention strategy. Integrating the PPI and regulatory networks not only provides a combined view of signal transduction and gene regulation but also help us to shortlist upstream regulators of genes involved in establishing infection and metabolic functions. The activities of such individual regulators and transcription factors can be taken up for testing and screened for inhibition by small molecules/antibiotics.[60] Besides, SalmoNet also provides information on binding sites which can be used to design transcription factor decoys (anti-sense nucleotides of the transcription factor binding motif)[61] that prevent the binding of transcription factors to their targets. Anti-sense oligonucleotides have been used to modulate gene expression in a wide variety of intracellular bacterial pathogens such as *E. coli*,[62,63] *Salmonella Typhimurium*,[62] *Enterococcus faecalis*,[64] and *Klebsiella pneumoniae*.[65] Clinical trials have also been performed using anti-sense oligonucleotides for the treatment of human diseases[66,67] thus suggesting that the potential of using anti-sense oligonucleotides against bacterial infections looks promising. In addition, the proteins and enzymes involved in critical PPIs and metabolic reactions respectively can also be subjected to the classical or systems-based drug-discovery pipelines.[68] The multi-layered network of SalmoNet allows discovering new molecular targets and strategies for therapeutic or prophylactic interventions based on the emergent properties of the networks. Modern drug and target discovery pipelines[69,70] advocate a systemic approach, which relies on the integration of various heterogeneous data such as expression pro les and other multiple prior knowledge networks. SalmoNet satis es this need by providing the prior knowledge networks for various strains of *Salmonella*.

As a source of both experimental and predicted interactions, SalmoNet contains information on the transcriptional regulatory targets of various transcription factors based on genome-wide motif scans. In addition, predicted targets of recently characterized transcription factors, such as RtsB,[70] which regulates the expression of invasion and flagellar genes, are also provided for future experimental veri cation and validation. Similar experimental testing and veri cation can also be performed on the predicted PPIs as well given the importance of PPIs in the survival of *Salmonella* inside the host cells.

From an epidemiological perspective, information on networks of multiple strains and strain-speci c interactions further enriches *Salmonella* epidemiology studies. Traditional epidemiological approaches rely on tracking genetic polymorphisms and loss/gain of virulence genes speci c to certain contexts and conditions. Hence, interaction networks could help to evaluate the effects of genetic polymorphisms in a systematic way, and thus, help in lling the gap between observed phenotypes of mutated strains and their genotypes. For example, SalmoNet can be used to further investigate the effect of cis-regulatory mutations on interactions, as well as the network level properties which determine the virulence characteristics of different strains of *Salmonella*.

Benchmarking *Salmonella* network data

There is no single resource storing *Salmonella*-speci c protein-protein interactions (PPIs) but they are captured in general databases such as STRING[71] and IntAct.[72] In STRING, PPIs are based on different types of data such as genomic context, co-occurrence, co-expression, data derived text-mining and imported data from IntAct and other PPI resources. STRING contains only 237 experimentally veri ed interactions in addition to 1014634 predicted interactions based on criteria such as neighborhood, gene-fusion, gene-co-occurence, co-expression, and text mining among *Salmonella* proteins. In IntAct, which contains manually curated and also imported PPI data from other databases, there are only 31 PPIs for *Salmonella* proteins.

In the case of transcriptional regulatory networks, RegulonDB[17] stands out as one of the most comprehensive repositories for prokaryotic gene regulation. However, RegulonDB is restricted to

E. coli. RegPrecise[73] contains information for multiple bacterial species using genome-wide predictions based on manually curated reconstruction of regulons (which are set of genes whose expression is regulated by a transcription factor). Unfortunately, RegPrecise does not provide the original source of data used for the predictions, making further application of the data dif cult. While well-known sources such as ORegAnno[15] and PAZAR[16] also capture regulatory interactions for multiple species, they do not contain any interactions for *Salmonella*.

As for the metabolic networks, there are numerous resources such as KEGG,[18] MetaCyc/BioCyc[74] and the BioModels databases[75] containing seemingly *Salmonella* speci c metabolic reactions. However, these databases are either not curated systematically or are not based on experimental results. KEGG for example contains information on pathway reactions and their entities for a large number of *Salmonella* strains but the *Salmonella* pathway annotations are based on computational predictions and not on experimental data. Similarly, coliBASE[76] captures comparative genomic data in terms of whole genome alignments and ortholog gene lists for *Salmonella*. Further information on bacterial metabolism can be found in specialized databases such as PATRIC for pathogens[77] or TRACTOR DB for Gamma proteobacteria.[78] However, most of the above mentioned resources contain limited interaction information for *Salmonella* and do not enable researchers for comparative network analysis or systems biological modeling of processes other than metabolism (e.g., they do not provide regulators of metabolic processes).

CONCLUSION

We present the rst public biological network resource for *Salmonella* research. SalmoNet contains network data (metabolic, regulatory, protein-protein interaction) for 10 representative pathogenic *Salmonella* strains. To elucidate the virulence program of *Salmonella* for either discovering knowledge on biological mechanisms or for therapeutic interventions, it is rewarding to integrate the different network layers that capture emergent properties of the system. SalmoNet represents a resource which contains information on interactions from multiple layers of biological organization that can be analyzed as such, or as a topological backbone to be integrated with new -omic datasets for analyzing the dynamics. SalmoNet opens the possibility for systems-level studies of the pathogen *Salmonella* with unprecedented details in a standardized and well documented format. The resource can be browsed and downloaded as a whole or in user-de ned interaction sets at http://salmonet.org. The analysis of SalmoNet could go far beyond fundamental biological and systems biology research. SalmoNet can be applied by medical microbiologists and epidemiologists to understand the strain speci c differences of *Salmonella* and can serve as a starting point for further experimental investigations and systems medicine based drug discovery.

METHODS

Strains and orthology

Five strains of serovars with a predominantly gastro-intestinal lifestyle of *Salmonella* and ve strains of serovars of extra-intestinal lifestyle were selected (Table 2). We included *Salmonella enterica subsp. enterica serovar Typhimurium* str. SL1344 as a sixth gastro-intestinal strain since it is widely used as a reference strain. We determined orthologous proteins among the selected strains, as well as for the model organism *E. coli* K12 with InParanoid.[79] We used a reciprocal best hit approach using BLAST to identify homologous protein sequences including those from plasmids corresponding to the selected strains. The protein sequences were downloaded from UniProt[80] as of January 2015. The results from the comparison of proteins one by one among any pair of strains were used to derive the ortholog clusters. Sequence similarity was set at > = 95% in order to minimize false positives given that the chosen strains belonged to the same species. Clustered groups contained both paralogs and orthologs (Supplementary Table 3). We removed the pseudo-genes listed in [7] from the resultant ortholog list.

Reconstruction of networks

We developed metabolic, regulatory and protein-protein interaction networks for *Salmonella*, using complementary approaches followed by merging the three layers into a uni ed *Salmonella* network. We performed this process for the 10 strains separately that resulted in 10 strain speci c molecular networks.

Metabolic networks: We de ned the metabolic network as follows: if a metabolite is a product of a reaction and substrate in another, the two proteins catalysing the different reactions were linked, as described in ref. 81 We did not consider the links for metabolites appearing in more than 10 reactions as outlined in ref. 81

We retrieved the metabolic reactions from two different sources with different levels of curation: the manually curated metabolic model of *S. Typhimurium* LT2 (referred to as STM),[82] and predictions from the BioModels database[75] containing Enzyme Commission (EC) numbers. In the latter, EC number assignments are automatic and include predictions for enzymes present in *Salmonella spp* and not necessarily present in *E. coli*. The STM model was derived from an *E. coli* metabolic model and con rmed by flux balance analysis. For the extrapolation of metabolic reactions from the above mentioned models, we assumed that the same reactions occur in the *Salmonella* strains when orthologous protein(s) of the enzyme(s) involved in the reactions were found to be present in the corresponding strains.

Regulatory networks: Regulatory interactions represent the binding of transcription factors to gene promoters. They consist of both predicted and experimentally veri ed interactions in our study. We collected experimentally veri ed DNA-binding sites of *Salmonella* transcription factors (Supplementary Table 4) from the literature with manual curation, as well as information from publicly available datasets. For the high-throughput datasets retrieved from,[83] peaks were identi ed as described elsewhere.[84,85] For all the other high-throughput datasets, the processed data (transcription factors, their targets and corresponding binding motifs when available) was retrieved from the cited sources (Supplementary Table 4). We then constructed Position Speci c Scoring Matrices (PSSMs) from the manually inferred binding sites and sites corresponding to the signi cant consensus motifs from the low- and high-throughput datasets respectively using the *consensus* tool[86] with default parameters. PSSMs constructed from too few binding sites have low predictive values. Hence, in instances where the number of binding sites (according to published data) corresponding to a transcription factor were less than three, we used corresponding sites from orthologous targets present in one or more of the other *Salmonella* strains under study for the PSSM construction. Since the predictive capacity and information content varies among PSSMs, we determined speci c optimal P-value thresholds for every PSSM-strain combination using the *matrix quality* tool[21] (Supplementary Table 5). We used the TRANSFAC-formatted PSSMs via the *matrix scan* tool[87] to scan the promoter regions of all the genes from the genomes of the selected *Salmonella* strains. We con ned the binding site search to 5000 bp upstream of the start codon of every protein coding gene to capture functionally active transcription factor binding sites in genomic regions including intergenic sequences between convergent genes.[88] However, sequences which overlap with upstream coding sequences were excluded. The promoters were retrieved

using the "retrieve sequence" function of the RSAT tool suite. For the background model, we used a Markov order of 1, and the model was estimated individually for every strain. Both the strands of the genome were scanned for the presence of putative binding sites and the optimal P-value determined for every TF-strain combination as described previously was used during the corresponding scans. Putative hits with a *P*-value lower than the corresponding optimal cut-off values were considered to be signi cant. Based on the principle of "regulogs",[89,90] we also inferred transcription factor-target gene relationships in *Salmonella* strains. Regulogs are regulatory interactions rst detected in one species (in this case in *E. coli*) and then predicted to be potentially present in another species (in this case in *Salmonella*) based on sequence homology of the transcription factor, the target gene and the transcription factor binding site. Accordingly, we used the *E. coli* transcription factor - target gene binding site information retrieved from RegulonDB[17] in conjunction with the homolog clustering results to extrapolate regulatory interactions from *E. coli* to the *Salmonella* strains. Operon information was retrieved from DOOR.[91] The workflow is presented in Fig. 1.

PPI networks: We performed manual curation to retrieve existing PPI information in *Salmonella* spp from the literature using a curation protocol we developed for the SignaLink eukaryotic signaling network resource as previously described.[92,93] Briefly, we collected signaling interactions involving *Salmonella* proteins from primary research articles identi ed by using iHOP[94] and ChilliBot[95] tools in addition to those articles directly found in PubMed searches. The main text and the abstracts of these articles were examined to retrieve the interactions between *Salmonella* proteins. Experimentally veri ed *Salmonella* PPIs were retrieved from IntAct.[72] Proteome-wide predictions to predict PPIs were carried out using 3D protein-based structure predictions of Interactome 3D[96] and using *E. coli* PPIs for interolog predictions. The interologs were inferred based on *E. coli* PPIs retrieved from IntAct,[72] BioGrid[97] and a high-throughput, yeast-2-hybrid screen of the *E. coli* interactome.[98]

Phylogenetic tree construction

Gene sequences corresponding to the *Salmonella* strains considered in this study were downloaded using the *retrieve sequence* tool from the Regulatory Sequence Analysis Tools.[87] Out of the 2912 common ortholog gene sets from the strains in this study, 457 ortholog sets were discarded due to discrepancies such as misconverted locus tags/IDs. Ortholog genes that had different lengths across strains were aligned by using ClustalOmega[99] implemented in the *msa* Bioconductor R package.[100] We identi ed 85 ortholog gene sets where one or more strains had more than one sequences (due to gene duplication or misannotation). We discarded these extra sequences after manual curation and the sequences that were more similar to the sequences of other strains were retained.

MrBayes v3.2.4[101]—which is a program for Bayesian inference based selection of evolutionary models—was used to analyze the phylogenetiic relationships of the strains using the polymorphic sites from genes in the orthologous gene sets. The parameters of the evolutionary model between the sequence sites were unlinked. Gaps were not considered as polymorphisms since the applied phylogenetic software treated them as missing data. Thus, gaps did not contribute to the phylogenetic information. Ortholog groups whose ratio of polymorphic sites to gene lengths was more than 0.1 (100 genes) were discarded and consequently polymorphic sites from potentially *false* orthologs that had low sequence similarity were excluded. After applying the above lter, 64,531 polymorphic sites from 2360 orthologous genes were used to infer a genome based phylogenetic tree. Metropolis coupling Markov Chain Monte Carlo (MC^3) analysis was performed for 10 million generations and 25% of the samples from the beginning of the chain were discarded when applying MrBayes.

Network based dendrograms

The networks (Supplementary Table 6) were represented by interaction matrices containing binary data, where "1"-s represented the presence of an interaction between the same pair of nodes inferred by orthology across the strains and "0"-s stood for missing interactions (Supplementary Table 6). In order to represent the hierarchical classi cation of strains from network data, we constructed dendrograms based on the metabolic, regulatory and PPI interaction matrices using MrBayes v3.2.4. To calculate network based dendrograms, the same MrBayes MC^3 analyses were performed as for the genome-based tree except that the datatype was set to "restriction" and no substitution model was applied.

Functional analysis of the transcriptional regulatory network

In order to understand the biological context within the regulons of the two serovars, functional enrichment analysis was performed to infer the over-represented Gene Ontology (GO) Biological Process Terms within the predicted regulons. Here, we considered only those interactions that were predicted to occur in at least two of the ten studied strains. This was performed to minimize the chances of including possible false positives in our analyses. In addition, we considered only the predicted regulons and GO terms containing at least 10 entities across all the strains within each pathovar. The background set comprised the entire collection of genes with annotated GO terms in the genomes. To determine the enriched GO Biological Process terms, the hypergeometric test with the Bonferroni correction was applied. The signi cance score for each enrichment event was calculated as the -log10 function of the corrected *P*-value. Enriched terms with a signi cance score greater than zero were considered as signi cant. We retrieved TF-GO relationships, which were exclusive to either of the two serovars. We restricted the analysis to TFs that were predicted to contain different enriched GO Biological Process terms within their putative regulons in extra-intestinal and gastro-intestinal pathovars. We also performed a manual assignment of functional processes derived from the Gene Ontology database for every GO term identi ed in the previous step. Subsequently, we replaced GO terms with their corresponding functional process(es) in order to simplify the graph without losing information.

Data availability statement

The datasets generated in the study are freely available at http://salmonet.org/. The source data as well as the generated datasets are provided as Supplementary tables which are freely available via *NPJ Systems Biology and Applications website*. The tools and resources such as the RSAT suite, Mr. Bayes, InParanoid, iHOP, ChilliBot, ClustaOmega, RegulonDB, Interactome 3D, IntAct, BioGrid, and UniProt which are used in this study are publicly available. Custom codes used in the study are available upon request.

ACKNOWLEDGEMENTS

The authors are grateful for the helpful discussions to the members and visitors of the Baranyi, Korcsmaros, and Kingsley groups, as well as for the gap- lling ChIP-Seq and RNA-seq datasets provided by Joseph Wade (Wadsworth Center, USA). This work was supported by a fellowship to T.K. in computational biology at the Earlham Institute (Norwich, UK) in partnership with the Quadram Institute (Norwich, UK), and strategically supported by the Biotechnological and Biosciences Research Council, UK grants (BB/J004529/1, BB/P016774/1 and BB/CSP17270/1).

AUTHOR CONTRIBUTIONS

A.M contributed to the design of the work and drafting the manuscript. P.S. carried out the network reconstructions and drafted and revised the manuscript. D.F. contributed to the network reconstructions and set up the web-resource. A.D. performed the curation and testing of the website. E.A. was involved in inferring the classi cation trees and dendrograms. M.O. contributed to internal testing and quality control. P.B. and R.K. contributed to framing the biological basis of the work and the discussions in the manuscript. J.B. and T.K. conceived and supervised the entire study. All the authors read and approved the nal version of the manuscript.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Systems Biology and Applications* website (https://doi.org/10.1038/s41540-017-0034-z).

Competing interests: The authors declare no competing nancial interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional af liations.

REFERENCES

1. Sánchez-Vargas, F. M., Abu-El-Haija, M. A. & Gómez-Duarte, O. G. *Salmonella* infections: an update on epidemiology, management, and prevention. *Travel. Med. Infect. Dis.* **9**, 263–277 (2011).
2. Majowicz, S. E. et al. The global burden of nontyphoidal *Salmonella gastro enteritis*. *Clin. Infect. Dis.* **50**, 882–889 (2010).
3. GBD. Global, regional, and national life expectancy, all-cause mortality, and cause-speci c mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388**, 1459–1544 (2016).
4. Mills, D. M., Bajaj, V. & Lee, C. A. A 40 kb chromosomal fragment encoding *Salmonella typhimurium* invasion genes is absent from the corresponding region of the *Escherichia coli* K-12 chromosome. *Mol. Microbiol.* **15**, 749–759 (1995).
5. Aleksic, S., Heinzerling, F. & Bockemuhl, J. Human infection caused by Salmonellae of subspecies II to VI in Germany, 1977-1992. *Zent. Bakteriol.* **283**, 391–398 (1996).
6. Parkhill, J. et al. Complete genome sequence of a multiple drug resistant *Sal monella enterica* serovar Typhi CT18. *Nature* **413**, 848–852 (2001).
7. Nuccio, S.-P. & Bäumler, A. J. Comparative analysis of *Salmonella* genomes identi es a metabolic network for escalating growth in the inflamed gut. *MBio* **5**, e00929–14 (2014).
8. Haghjoo, E. & Galán, J. E. *Salmonella* typhi encodes a functional cytolethal distending toxin that is delivered into host cells by a bacterial-internalization pathway. *Proc. Natl. Acad. Sci. USA* **101**, 4614–4619 (2004).
9. Wilson, J. W. et al. Space flight alters bacterial gene expression and virulence and reveals a role for global regulator Hfq. *Proc. Natl. Acad. Sci. USA* **104**, 16299–16304 (2007).
10. Winter, S. E., Raffatellu, M., Wilson, R. P., Rüssmann, H. & Bäumler, A. J. The *Salmonella enterica* serotype Typhi regulator TviA reduces interleukin-8 production in intestinal epithelial cells by repressing flagellin secretion. *Cell. Microbiol.* **10**, 247–261 (2008).
11. Song, J., Gao, X. & Galán, J. E. Structure and function of the *Salmonella* Typhi chimaeric A(2)B(5) typhoid toxin. *Nature* **499**, 350–354 (2013).
12. Klemm, E. J. et al. Emergence of host-adapted *Salmonella* Enteritidis through rapid evolution in an immunocompromised host. *Nat. Microbiol.* **1**, 15023 (2016).
13. Desai, P. T. et al. Evolutionary genomics of *Salmonella enterica* subspecies. *MBio* **4**, e00579–12 (2013).
14. Gonçalves, E. et al. Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models. *Mol. Biosyst.* **9**, 1576–1583 (2013).
15. Grif th, O. L. et al. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.* **36**, D107–D113 (2008).
16. Portales-Casamar, E. et al. The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res.* **37**, D54–D60 (2009).
17. Gama-Castro, S. et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* **44**, D133–D143 (2016).
18. Kanehisa, M. et al. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205 (2014).
19. Karimpour-Fard, A., Detweiler, C. S., Erickson, K. D., Hunter, L. & Gill, R. T. Cross-species cluster co-conservation: a new method for generating protein interaction networks. *Genome Biol.* **8**, R185 (2007).
20. Shou, C. et al. Measuring the evolutionary rewiring of biological networks. *PLoS Comput. Biol.* **7**, e1001050 (2011).
21. Medina-Rivera, A. et al. Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res.* **39**, 808–824 (2011).
22. Walthers, D. et al. The response regulator SsrB activates expression of diverse *Salmonella* pathogenicity island 2 promoters and counters silencing by the nucleoid-associated protein H-NS. *Mol. Microbiol.* **65**, 477–493 (2007).
23. Tomljenovic-Berube, A. M., Mulder, D. T., Whiteside, M. D., Brinkman, F. S. L. & Coombes, B. K. Identi cation of the regulatory logic controlling Salmonella pathoadaptation by the SsrA-SsrB two-component system. *PLoS Genet.* **6**, e1000875 (2010).
24. Timme, R. E. et al. Phylogenetic diversity of the enteric pathogen Salmonella enterica subsp. enterica inferred from genome-wide reference-free SNP characters. *Genome Biol. Evol.* **5**, 2109–2123 (2013).
25. Chiu, C.-H. et al. The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. *Nucleic Acids Res.* **33**, 1690–1698 (2005).
26. Abbott, S. L., Ni, F. C. Y. & Janda, J. M. Increase in extraintestinal infections caused by Salmonella enterica subspecies II-IV. *Emerg. Infect. Dis.* **18**, 637–639 (2012).
27. Wilkins, E. G. & Roberts, C. Extraintestinal salmonellosis. *Epidemiol. Infect.* **100**, 361–368 (1988).
28. Chen, P. L. et al. Extraintestinal focal infections in adults with nontyphoid Salmonella bacteraemia: predisposing factors and clinical outcome. *J. Intern. Med.* **261**, 91–100 (2007).
29. Huang, D. B. & DuPont, H. L. Problem pathogens: extra-intestinal complications of *Salmonella enterica* serotype Typhi infection. *Lancet Infect. Dis.* **5**, 341–348 (2005).
30. MacKenzie, K. D. et al. Bistable expression of CsgD in *Salmonella enterica* serovar Typhimurium connects virulence to persistence. *Infect. Immun.* **83**, 2312–2326 (2015).
31. Zakikhany, K., Harrington, C. R., Nimtz, M., Hinton, J. C. D. & Römling, U. Unphosphorylated CsgD controls bio lm formation in *Salmonella enterica* serovar Typhimurium. *Mol. Microbiol.* **77**, 771–786 (2010).
32. Raivio, T. L. & Silhavy, T. J. Transduction of envelope stress in *Escherichia coli* by the Cpx two-component system. *J. Bacteriol.* **179**, 7724–7733 (1997).
33. Pogliano, J., Lynch, A. S., Belin, D., Lin, E. C. & Beckwith, J. Regulation of *Escherichia coli* cell envelope proteins involved in protein folding and degradation by the Cpx two-component system. *Genes Dev.* **11**, 1169–1182 (1997).
34. Wolfe, A. J., Parikh, N., Lima, B. P. & Zemaitaitis, B. Signal integration by the two-component signal transduction response regulator CpxR. *J. Bacteriol.* **190**, 2314–2322 (2008).
35. Xu, Q. & Reed, J. C. Bax inhibitor-1, a mammalian apoptosis suppressor identi ed by functional screening in yeast. *Mol. Cell.* **1**, 337–346 (1998).
36. Chae, H.-J. et al. Evolutionarily conserved cytoprotection provided by Bax Inhibitor-1 homologs from animals, plants, and yeast. *Gene* **323**, 101–113 (2003).
37. Hegde, R. et al. Identi cation of Omi/HtrA2 as a mitochondrial apoptotic serine protease that disrupts inhibitor of apoptosis protein-caspase interaction. *J. Biol. Chem.* **277**, 432–438 (2002).
38. Hong, S. H., Lee, J. & Wood, T. K. Engineering global regulator Hha of *Escherichia coli* to control bio lm dispersal. *Microb. Biotechnol.* **3**, 717–728 (2010).
39. García-Contreras, R., Zhang, X.-S., Kim, Y. & Wood, T. K. Protein translation and cell death: the role of rare tRNAs in bio lm formation and in activating dormant phage killer genes. *PLoS One* **3**, e2394 (2008).
40. Gorga, F., Galdiero, M., Buommino, E. & Galdiero, E. Porins and lipopolysaccharide induce apoptosis in human spermatozoa. *Clin. Diagn. Lab. Immunol.* **8**, 206–208 (2001).
41. Rolig, A. S., Carter, J. E. & Ottemann, K. M. Bacterial chemotaxis modulates host cell apoptosis to establish a T-helper cell, type 17 (Th17)-dominant immune response in Helicobacter pylori infection. *Proc. Natl Acad. Sci. USA* **108**, 19749–19754 (2011).
42. Takaya, A. et al. Derepression of Salmonella pathogenicity island 1 genes within macrophages leads to rapid apoptosis via caspase-1- and caspase-3-dependent pathways. *Cell. Microbiol.* **7**, 79–90 (2005).
43. Lara-Tejero, M. & Galán, J. E. *Salmonella enterica* serovar Typhimurium pathogenicity island 1-encoded type III secretion system translocases mediate intimate attachment to nonphagocytic cells. *Infect. Immun.* **77**, 2635–2642 (2009).
44. Galán, J. E. Salmonella interactions with host cells: type III secretion at work. *Annu. Rev. Cell. Dev. Biol.* **17**, 53–86 (2001).
45. Kaur, J. & Jain, S. K. Role of antigens and virulence factors of Salmonella enterica serovar Typhi in its pathogenesis. *Microbiol. Res.* **167**, 199–210 (2012).
46. Zhang, S. et al. Molecular pathogenesis of *Salmonella enterica* serotype typhimurium-induced diarrhea. *Infect. Immun.* **71**, 1–12 (2003).
47. Bakowski, M. A., Braun, V. & Brumell, J. H. Salmonella-containing vacuoles: directing traf c and nesting to grow. *Traf c* **9**, 2022–2031 (2008).

48. Steele-Mortimer, O. The Salmonella-containing vacuole: moving with the times. *Curr. Opin. Microbiol.* **11**, 38–45 (2008).
49. Schleker, S. et al. The current Salmonella-host interactome. *Proteom. Clin. Appl.* **6**, 117–133 (2012).
50. Rytkönen, A. & Holden, D. W. Bacterial interference of ubiquitination and deubiquitination. *Cell. Host. Microbe* **1**, 13–22 (2007).
51. Kim, J. S. et al. Molecular characterization of the InvE regulator in the secretion of type III secretion translocases in *Salmonella enterica* serovar Typhimurium. *Microbiology* **159**, 446–461 (2013).
52. Wynosky-Dol , M. A. et al. Oxidative metabolism enables Salmonella evasion of the NLRP3 inflammasome. *J. Exp. Med.* **211**, 653–668 (2014).
53. Antunes, L. C. M. et al. Impact of salmonella infection on host hormone metabolism revealed by metabolomics. *Infect. Immun.* **79**, 1759–1769 (2011).
54. Hernandez, L. D., Hueffer, K., Wenk, M. R. & Galán, J. E. Salmonella modulates vesicular traf c by altering phosphoinositide metabolism. *Science* **304**, 1805–1807 (2004).
55. Dandekar, T. et al. Salmonella-how a metabolic generalist adopts an intracellular lifestyle during infection. *Front. Cell. Infect. Microbiol.* **4**, 191 (2014).
56. DeRubertis, F. R. & Woeber, K. A. Accelerated cellular uptake and metabolism of L-thyroxine during acute *Salmonella typhimurium* sepsis. *J. Clin. Invest.* **52**, 78–87 (1973).
57. Arsenault, R. J., Napper, S. & Kogut, M. H. *Salmonella enterica* Typhimurium infection causes metabolic changes in chicken muscle involving AMPK, fatty acid and insulin/mTOR signaling. *Vet. Res.* **44**, 35 (2013).
58. Mazé, A., Glatter, T. & Bumann, D. The central metabolism regulator EIIAGlc switches Salmonella from growth arrest to acute virulence through activation of virulence factor secretion. *Cell Rep.* **7**, 1426–1433 (2014).
59. Herzberg, M., Jawad, M. J. & Pratt, D. Succinate metabolism and virulence in *Salmonella typhimurium*. *Nature* **204**, 1285–1286 (1964).
60. Berg, T. Inhibition of transcription factors with small organic molecules. *Curr. Opin. Chem. Biol.* **12**, 464–471 (2008).
61. Mann, M. J. Transcription factor decoys: a new model for disease intervention. *Ann. NY Acad. Sci.* **1058**, 128–139 (2005).
62. McKinney, J. S., Zhang, H., Kubori, T., Galán, J. E. & Altman, S. Disruption of type III secretion in *Salmonella enterica* serovar Typhimurium by external guide sequences. *Nucleic Acids Res.* **32**, 848–854 (2004).
63. Tilley, L. D. et al. Gene-speci c effects of antisense phosphorodiamidate morpholino oligomer-peptide conjugates on *Escherichia coli* and *Salmonella enterica* serovar typhimurium in pure culture and in tissue culture. *Antimicrob. Agents Chemother.* **50**, 2789–2796 (2006).
64. Shen, N. et al. Inactivation of expression of several genes in a variety of bacterial species by EGS technology. *Proc. Natl Acad. Sci. USA* **106**, 8163–8168 (2009).
65. Kurupati, P., Tan, K. S. W., Kumarasinghe, G. & Poh, C. L. Inhibition of gene expression and growth by antisense peptide nucleic acids in a multiresistant beta-lactamase-producing *Klebsiella pneumoniae* strain. *Antimicrob. Agents Chemother.* **51**, 805–811 (2007).
66. Sharma, V. K., Sharma, R. K. & Singh, S. K. Antisense oligonucleotides: modi cations and clinical trials. *Med. Chem. Commun.* **5**, 1454–1471 (2014).
67. Koo, T. & Wood, M. J. Clinical trials using antisense oligonucleotides in duchenne muscular dystrophy. *Hum. Gene. Ther.* **24**, 479–488 (2013).
68. Wang, R.-S., Maron, B. A. & Loscalzo, J. Systems medicine: evolution of systems biology from bench to bedside. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **7**, 141–161 (2015).
69. Butcher, E. C., Berg, E. L. & Kunkel, E. J. Systems biology in drug discovery. *Nat. Biotechnol.* **22**, 1253–1259 (2004).
70. Ellermeier, C. D. & Slauch, J. M. RtsA and RtsB coordinately regulate expression of the invasion and flagellar genes in *Salmonella enterica* serovar Typhimurium. *J. Bacteriol.* **185**, 5096–5108 (2003).
71. Szklarczyk, D. et al. STRINGv10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
72. Orchard, S. et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).
73. Novichkov, P. S. et al. RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC. Genom.* **14**, 745 (2013).
74. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42**, D459–D471 (2014).
75. Juty, N. et al. BioModels: Content, Features, Functionality, and Use. *CPT Pharmacomet. Syst. Pharmacol.* **4**, e3 (2015).
76. Chaudhuri, R. R., Khan, A. M. & Pallen, M. J. coliBASE: an online database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics. *Nucleic Acids Res.* **32**, D296–D299 (2004).
77. Wattam, A. R. et al. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **42**, D581–D591 (2014).
78. González, A. D., Espinosa, V., Vasconcelos, A. T., Pérez-Rueda, E. & Collado-Vides, J. TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes. *Nucleic Acids Res.* **33**, D98–D102 (2005).
79. Sonnhammer, E. L. L. & Östlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* **43**, D234–D239 (2015).
80. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
81. Kreimer, A., Borenstein, E., Gophna, U. & Ruppin, E. The evolution of modularity in bacterial metabolic networks. *Proc. Natl. Acad. Sci. USA* **105**, 6976–6981 (2008).
82. Thiele, I. et al. A community effort towards a knowledge-base and mathematical model of the human pathogen *Salmonella typhimurium* LT2. *BMC. Syst. Biol.* **5**, 8 (2011).
83. Smith, C., Stringer, A. M., Mao, C., Palumbo, M. J. & Wade, J. T. Mapping the regulatory network for *Salmonella enterica* serovar Typhimurium invasion. *MBio* **7**, e01024–e01026 (2016).
84. Fitzgerald, D. M., Bonocora, R. P. & Wade, J. T. Comprehensive mapping of the *Escherichia coli* flagellar regulatory network. *PLoS Genet.* **10**, e1004649 (2014).
85. Singh, S. S. et al. Widespread suppression of intragenic transcription initiation by H-NS. *Genes Dev.* **28**, 214–219 (2014).
86. Thomas-Chollier, M. et al. RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res.* **39**, W86–W91 (2011).
87. Medina-Rivera, A. et al. RSAT 2015: regulatory sequence analysis tools. *Nucleic Acids Res.* **43**, W50–W56 (2015).
88. Haycocks, J. R. J. & Grainger, D. C. Unusually Situated Binding Sites for Bacterial Transcription Factors Can Have Hidden Functionality. *PLoS One* **11**, e0157016 (2016).
89. Rodionov, D. A. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem. Rev.* **107**, 3467–3497 (2007).
90. Yu, H. et al. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.* **14**, 1107–1118 (2004).
91. Mao, F., Dam, P., Chou, J., Olman, V. & Xu, Y. DOOR: a database for prokaryotic operons. *Nucleic Acids Res.* **37**, D459–D463 (2009).
92. Korcsmáros, T. et al. Uniformly curated signaling pathways reveal tissue-speci c cross-talks and support drug target discovery. *Bioinformatics* **26**, 2042–2050 (2010).
93. Fazekas, D. et al. SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC. Syst. Biol.* **7**, 7 (2013).
94. Hoffmann, R. Using the iHOP information resource to mine the biomedical literature on genes, proteins, and chemical compounds. *Curr. Protoc. Bioinform.* Chapter 1, Unit1.16 (2007). https://doi.org/10.1002/0471250953.bi0116s20
95. Chen, H. & Sharp, B. M. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinforma.* **5**, 147 (2004).
96. Mosca, R., Céol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat. Methods* **10**, 47–53 (2013).
97. Chatr-Aryamontri, A. et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* **43**, D470–D478 (2015).
98. Rajagopala, S. V. et al. The binary protein-protein interaction landscape of *Escherichia coli*. *Nat. Biotechnol.* **32**, 285–290 (2014).
99. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
100. Bodenhofer, U., Bonatesta, E., Horej -Kainrath, C. & Hochreiter, S. msa: an R package for multiple sequence alignment. *Bioinformatics* **31**, 3997–3999 (2015).
101. Ronquist, F. et al. MrBayes 3.2: ef cient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).

Chapter 21

Network Biology Approaches to Identify Molecular and Systems-Level Differences Between *Salmonella* Pathovars

Marton Olbei, Robert A. Kingsley, Tamas Korcsmaros, and Padhmanand Sudhakar

Abstract

The field of systems biology endeavors to map, study, and simulate cellular systems and their underlying mechanisms. The internal mechanisms of biological systems can be represented with networks comprising nodes and edges. Nodes denote the constituents of the biological system whereas edges indicate the relationships among them. Likewise, every layer of cellular organization can be represented by networks. Multilayered networks capture interactions between various network types, such as transcriptional regulatory networks, protein–protein interaction networks, and metabolic networks from the same biological system. This property makes multilayered networks representative of the system while its internal mechanisms are investigated. However, there are not many multilayered networks containing integrated data for nonmodel organisms including the bacterial pathogens *Salmonella*. Here, we outline the steps to create such an integrated network database, through the example of SalmoNet, the first integrated multilayered data resource for multiple strains belonging to distinct *Salmonella* serovars.

Key words Systems biology, Multilayered networks, Network reconstruction framework, Pathogen, *Salmonella*

1 Introduction

Salmonella enterica and its more than 1500 serovars is one of the most common foodborne pathogens affecting human health. The *Salmonella* genus consists of gram-negative bacteria belonging to the Enterobacteriaceae family. They are related to Escherichia coli, a species containing both commensal bacteria found in the gut and pathogenic variants. Most *Salmonella* enterica serovars cause gastroenteritis, one of the most common foodborne illnesses accounting for almost 100 million cases each year [1], or disseminated (extraintestinal) diseases such as typhoid fever and bacteremia [2, 3].

The outcome of infection with *Salmonella enterica* depends on the genotype of the pathogen and the host species and immune

Arnaud Bridier (ed.), *Foodborne Bacterial Pathogens: Methods and Protocols*, Methods in Molecular Biology, vol. 1918, https://doi.org/10.1007/978-1-4939-9000-9_21,

status. As most *Salmonella enterica* serovars are generalist intestinal pathogens, it is thought that this is the ancestral state. A handful of serovars evolved to become adapted to circulate in specific host species, or indeed a single species and cause a more severe disseminated disease. Thus, *S. enterica* can be further divided into two pathovars: gastrointestinal and extraintestinal. Extraintestinal pathovars are specialists adapted to new environments in their host. The level of host adaptation in *Salmonella enterica* serotypes varies, with *Salmonella enterica serovar Typhi* being a specialist member of the group, while *Salmonella enterica serovar Typhimurium* being a generalist serovar.

Host adaptation is a complex evolutionary process and the integration of different levels of information is needed. The analysis of integrated networks (ones that combine many levels of data, e.g., regulation and protein–protein interactions) allows us to gain new insights into regulation, signal transduction, and metabolism. We can focus on specific processes important to the question at hand, without excluding entire levels of a biological system, e.g., to see whether a signaling pathway can alter anything on a metabolic level with its downstream effectors.

The SalmoNet database (http://salmonet.org/) includes the multilayered interaction networks of five well-known gastrointestinal and extraintestinal serovars of *Salmonella enterica* [4]. These networks consist of three layers namely protein–protein interactions, transcriptional regulation, and the metabolic layer. Every layer requires a specific protocol to collate and evaluate external data, the steps of which we outline in this chapter.

We provide a template for future studies intending to develop similar network resources for pathogenic or other bacteria, and for nonmodel organisms in general. The frameworks and workflows in SalmoNet can help other scientific communities which lack integrated network resources. By collecting information from the studied organisms and inferring information from closely related model organisms, they can serve as a knowledge base for less known species, while at the same time driving research forward by predicting interactions which were previously unknown.

2 Reconstruction of Transcriptional Regulatory Networks

1. Retrieve low throughput, experimentally validated data on transcription factor binding sites. This information can be retrieved either from manual curation of literature or from databases. The most commonly used tools for text-mining based literature searches include ChiliBot [5] and iHop [6] (for more information on databases, *see* **Note 1**).

2. Datasets generated from high-throughput experiments can also be used to infer binding sites from the genomic locations of transcription factor binding peaks (*see* **Notes 1** and **2**).
3. From the corresponding resources (Collect TF [15], RegulonDB [16], and Prodoric [17]), recover DNA binding sites from the experimentally verified targets of transcription factors.
4. For every transcription factor, use the recovered sites (from step 1.3) to construct a binding signature in a matrix format—otherwise known as a Position Specific Scoring Matrix (PSSM) using the *consensus* tool (*see* **Notes 3–5**).
5. Convert the PSSMs into the transfac format (*see* **Note 6**) using the *convert-matrix* tool.
6. PSSMs constructed from too few binding sites can have a low predictive power owing to their reduced information content (*see* **Note 7**). In such cases, orthologous sites can be included from closely related strains for transcription factors that have less than three binding sites.
7. Calculate optimal *P*-value thresholds (*see* **Note 8**) for every PSSM (generated in **step 4**) with the *matrix-quality* tool (*see* **Notes 9** and **10**).
8. Retrieve promoter sequences in the bacterial genome(s) of interest. The length of the promoter sequences depends on various factors such as the type of transcription factors being investigated in addition to other aspects (*see* **Note 11**). Promoter sequences can be retrieved (*see* **Note 12**) using the *retrieve-sequence* tool (http://embnet.ccg.unam.mx/rsat/retrieve-seq_form.cgi) within the RSAT tool suite. Alternatively, if the genome is not supported by RSAT, use *bedtools* (*see* **Note 13**) to extract promoter sequences of interest from whole genome sequence.
9. Predicting transcription factor binding sites using the PSSMs generated in the previous **steps (4–6)**, optimal *P*-values determined in **step 7** and proomoter sequences retrieved in **step 8**.
   - Promoter sequences from previous **step 8** can be used as the input for the subsequent step.
   - With the pattern matching (*see* **Note 14**) *matrix-scan* tool (http://embnet.ccg.unam.mx/rsat/matrix-scan_form.cgi) (*see* **Notes 15** and **16**), scan the promoter sequences using the constructed PSSMs to detect the presence of putative transcription factor binding sites.
   - Select the appropriate background model before the scan is performed (*see* **Note 16**).
   - Hits with a *P*-value less than the predetermined optimal *P*-value (from step 1.7) are considered to be potentially true.

10. Inferring cis-regulatory element enriched regions (CRERs) (*see* **Note 17**).
    - Follow all the preceding steps till step 1.9.
    - Set the CRER window size. Default values are in the range between 30 and 500 bp.
    - Set the site *P*-value. Only those sites whose *P*-value is below the chosen threshold at this step will be considered for the prediction of the CRER. Other parameters such as the CRER significance score and the CRER *P*-value can also be used as options for filtering.
11. Extrapolate based on orthology the experimentally verified transcriptional regulatory interactions from model prokaryotic species like E. coli (in the case of gram-negative organisms) or B. subtilis (in the case of gram-positive organisms) (*see* **Notes 18** and **19**, Fig. 1).

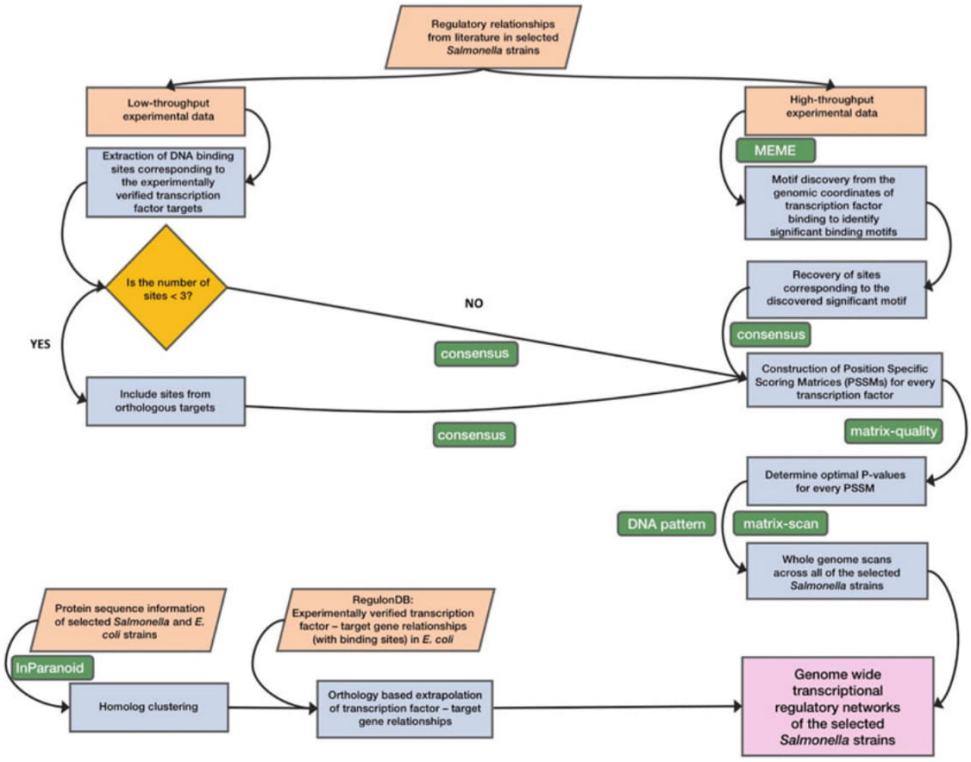

Fig. 1 Graphical description of the workflow used to reconstruct transcriptional regulatory networks in SalmoNet

3 Reconstructing Protein–Protein Interaction Networks

1. Use the text-mining based tools ChiliBot [5] and iHop [6], to extract organism specific protein–protein interaction information from literature sources and publications.
2. Experimentally verified organism specific protein–protein interactions from high-throughput experiments are added from the IntAct database [7] (*see* **Note 20**).
3. Retrieve predicted protein–protein interactions for your organism of interest from the Interactome 3D database [8].
4. Additional predictions are inferred by extrapolation based on orthology (*see* **Note 18**) from E. coli data obtained from the IntAct and BioGrid databases, and from yeast-2-hybrid screens of E. coli.

4 Reconstructing Metabolic Networks

1. Metabolic networks which are usually defined as a collection of enzyme metabolite reactions can be reconstructed into a graph. Generally, metabolites are represented as nodes and reactions as edges. Since the desired representation is for macromolecular components such as proteins and genes, a suitable approximation can be performed (*see* Subheading 3, **step 2**) by transforming metabolic reactions into networks as described in [9].
2. If a particular metabolite is a product of a reaction and at the same time a substrate in another, the two enzymatic proteins catalyzing the different reactions are connected to each other by an edge [9]. Metabolites appearing in more than ten reactions are not considered to avoid bias.
3. Collect metabolic networks from manually curated sources. For example, Flux Balance Analysis validated metabolic resources such as [10], which provide genome scale metabolic models for *Salmonella,* can be transformed.
4. For additional metabolic models, please refer to the BioModels database [11] which contains metabolic model predictions specific to the organism(s) in question.

5 Prediction of Interactions Across Organisms

1. Using InParanoid [12] or similar homology based clustering tools [13], create groups of orthologous genes encoding proteins (*see* **Note 21**).

2. Inference of the regulatory connections is based on the principle of regulogs [14] The principle of regulogs utilizes the homology based conservation of the transcription factor, the target gene as well as the transcription factor binding site on the target gene [14] to extrapolate interactions across organisms.
3. For the protein–protein interaction networks and the metabolic networks, only the sequence level orthology of the interacting components is used for the extrapolation of interactions.

6 Notes

1. Information on the binding sites of bacterial transcription factors can be retrieved from various databases such as Collect TF [15], RegulonDB [16], and Prodoric [17], among others.
2. Various applications such as MEME-ChIP belonging to the MEME suite of tools [18] can be used for extracting statistically significant motifs from high-throughput protein–DNA interaction profiling datasets (e.g., ChIP-chip, ChIP-seq).
3. PSSMs represent an easy way to capture the position wise frequency distribution of nucleotides which comprise the binding sites recognized by a particular transcription factor.
4. The RSAT (Regulatory Sequence Analysis Tools) suite [19] provides the users with a collection of different tools tailor-made for various kinds of analysis using regulatory sequences.
5. The *consensus* tool from within RSAT was used to construct PSSMs from sites. As of the date of writing this chapter, the *consensus* tool stands withdrawn from the RSAT tool suite. Users are advised to implement analogous tools such as *infogibbs* which can be found within the same tool suite.
6. For more information on matrix formats, please refer to the following address: http://floresta.eead.csic.es/rsat/help.convert-matrix.html.
7. The information content of a PSSM is described in detail in [20].
8. Since the predictive capacity of a PSSM is dependent on its information content, the statistical threshold for distinguishing a true positive from a false positive needs to be determined in a customized manner for every PSSM-strain combination.
9. *P*-values provide an indication of the false positive rate. For instance, a *P*-value threshold of 0.001 produces one false positive prediction for every kilobase. The optimal *P*-value thresholds depend on the information content of the PSSM. By determining over a range of *P*-values the weight-score distributions derived from both the original and permuted version of

the PSSMs, *matrix-quality* [21] identifies the point of divergence of the distributions which subsequently enables the identification of the optimal *P*-value. More details regarding the usage of the *matrix-quality* tool are described in [21].

10. The optimal *P*-value determination needs to be performed for every PSSM-genome combination. For instance, if there are *m* PSSMs which need to be scanned against *n* genomes, [$m \times n$] number of tests need to be performed with *matrix-quality*.
11. Typically, bacterial transcription factors bind to noncoding regions, which are immediately upstream to the start codon of the first gene of the regulated operon. In such cases, anywhere up to 5000 bp upstream from the regulated gene can be considered for scanning to detect the presence of a potential transcription factor binding site (TFBS). However, recent studies suggest that bacterial transcription factors especially those with repressor activity are known to bind even within the coding regions of the regulated genes in addition to the noncoding regions [22]. Therefore, ideally, depending on the type of transcription factor being studied, the promoter regions need to be retrieved on a case-by-case basis.
12. Please refer to http://embnet.ccg.unam.mx/rsat/supported-organisms.cgi for the complete list bacterial genomes supported by RSAT.
13. *bedtools* (http://bedtools.readthedocs.io/en/latest/) is a collection of simple easy to use command-line tools for handling large sequences especially nucleotides. It has multiple features which enable users to perform various operations such as manipulation and extraction of sequences based on annotations. *Bedtools* utilities are also available via the galaxy platform (https://test.galaxyproject.org/) which is a user-friendly interface targeted toward scientists without advanced level proficiency in bioinformatics especially on the command-line.
14. Pattern-matching is a process by which predefined signatures (in our case the PSSMs) are used to detect other potential copies of the signature in a query string object (promoter sequences).
15. A detailed protocol describing the use of the *matrix-scan* tool to detect putative TFBSs is outlined in [23].
16. In addition to the optimal *P*-value, there are other deterministic parameters which dictate the outcome of the pattern matching procedure. These include the background model whose Markov order and sequence specificity can be set accordingly. Furthermore, depending on the type of sequences being scanned (noncoding regions only or coding and noncoding regions), the sequence type can also be customized for determining the background model. Options also exist to exclude either of the strands or include both of them for the scan.

17. Cis-regulatory element enriched regions or CRERs [24] are short defined spans in the considered regulatory sequences and which are overrepresented with overlapping or nonoverlapping binding sites. They represent possible regulatory hot spots governing gene expression to a higher extent than other regions without a clustering of binding sites.
18. For most nonmodel organisms, experimental interaction information is sparse for almost all the network layers discussed herein. One of the strategies which have been suggested recently to overcome this limitation is to use the already available molecular level interaction data derived from model organisms [14]. Orthology information can be used to extrapolate the interactions from the model organism(s) to the species of interest.
19. Various resources such as RegulonDB [16] and DBTBS [25] contain experimentally verified transcriptional regulatory interactions including binding sites for E. coli and B. subtilis respectively.
20. The IntAct database (https://www.ebi.ac.uk/intact/) [7] is a data resource which contains a vast array of protein–protein interaction information from experiments.
21. It is generally recommended to use protein level information for homology based clustering.

Acknowledgments

The authors would like to acknowledge all the contributors of the SalmoNet resource as well as the helpful discussions from the members and visitors of the Baranyi, Korcsmaros, and Kingsley groups. This work was supported by a fellowship to T.K. in computational biology at the Earlham Institute (Norwich, UK) in partnership with the Quadram Institute (Norwich, UK), and strategically supported by the Biotechnological and Biosciences Research Council, UK grants (BB/J004529/1, BB/P016774/1, and BB/CSP17270/1). This work was also supported by the BBSRC Norwich Research Park Biosciences Doctoral Training Partnership grant number BB/M011216/1.

References

1. Majowicz SE, Musto J, Scallan E et al (2010) The global burden of nontyphoidal salmonella gastroenteritis. Clin Infect Dis 50:882–889. https://doi.org/10.1086/650733
2. Guirguis GF, Patel K, Gittens-Williams L et al (2017) Salmonella enterica serotype typhi bacteremia complicating pregnancy in the third trimester. Case Rep Obstet Gynecol 2017:4018096. https://doi.org/10.1155/2017/4018096
3. Mohanty S, Gaind R, Paglietti B et al (2010) Bacteraemia with pleural effusions complicating typhoid fever caused by high-level ciprofloxacin-resistant salmonella enterica serotype

Typhi. Ann Trop Paediatr 30:233–240. https://doi.org/10.1179/146532810X12786388978760

4. Métris A, Sudhakar P, Fazekas D et al (2017) SalmoNet, an integrated network of ten salmonella enterica strains reveals common and distinct pathways to host adaptation. NPJ Syst Biol Appl 3:31. https://doi.org/10.1038/s41540-017-0034-z
5. Chen H, Sharp BM (2004) Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics 5:147. https://doi.org/10.1186/1471-2105-5-147
6. Hoffmann R, Valencia A (2004) A gene network for navigating the literature. Nat Genet 36:664. https://doi.org/10.1038/ng0704-664
7. Kerrien S, Aranda B, Breuza L et al (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40:D841–D846. https://doi.org/10.1093/nar/gkr1088
8. Mosca R, Céol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. Nat Methods 10:47–53. https://doi.org/10.1038/nmeth.2289
9. Kreimer A, Borenstein E, Gophna U, Ruppin E (2008) The evolution of modularity in bacterial metabolic networks. Proc Natl Acad Sci U S A 105:6976–6981. https://doi.org/10.1073/pnas.0712149105
10. Thiele I, Hyduke DR, Steeb B et al (2011) A community effort towards a knowledge-base and mathematical model of the human pathogen salmonella Typhimurium LT2. BMC Syst Biol 5:8. https://doi.org/10.1186/1752-0509-5-8
11. Chelliah V, Juty N, Ajmera I et al (2015) BioModels: ten-year anniversary. Nucleic Acids Res 43:D542–D548. https://doi.org/10.1093/nar/gku1181
12. O'Brien KP, Remm M, Sonnhammer ELL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33:D476–D480. https://doi.org/10.1093/nar/gki107
13. Nichio BTL, Marchaukoski JN, Raittz RT (2017) New tools in orthology analysis: a brief review of promising perspectives. Front Genet 8:165. https://doi.org/10.3389/fgene.2017.00165
14. Yu H, Luscombe NM, Lu HX et al (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res 14:1107–1118. https://doi.org/10.1101/gr.1774904
15. Kiliç S, White ER, Sagitova DM et al (2014) CollecTF: a database of experimentally validated transcription factor-binding sites in bacteria. Nucleic Acids Res 42:D156–D160. https://doi.org/10.1093/nar/gkt1123
16. Gama-Castro S, Salgado H, Santos-Zavaleta A et al (2016) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res 44:D133–D143. https://doi.org/10.1093/nar/gkv1156
17. Grote A, Klein J, Retter I et al (2009) PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. Nucleic Acids Res 37:D61–D65. https://doi.org/10.1093/nar/gkn837
18. Bailey TL, Boden M, Buske FA et al (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202–W208. https://doi.org/10.1093/nar/gkp335
19. Medina-Rivera A, Defrance M, Sand O et al (2015) RSAT 2015: regulatory sequence analysis tools. Nucleic Acids Res 43:W50–W56. https://doi.org/10.1093/nar/gkv362
20. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15:563–577
21. Medina-Rivera A, Abreu-Goodger C, Thomas-Chollier M et al (2011) Theoretical and empirical quality assessment of transcription factor-binding motifs. Nucleic Acids Res 39:808–824. https://doi.org/10.1093/nar/gkq710
22. Haycocks JRJ, Grainger DC (2016) Unusually situated binding sites for bacterial transcription factors can have hidden functionality. PLoS One 11:e0157016. https://doi.org/10.1371/journal.pone.0157016
23. Turatsinze J-V, Thomas-Chollier M, Defrance M, van Helden J (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat Protoc 3:1578–1588. https://doi.org/10.1038/nprot.2008.97
24. Nelson AC, Wardle FC (2013) Conserved non-coding elements and cis regulation: actions speak louder than words. Development 140:1385–1395. https://doi.org/10.1242/dev.084459
25. Sierro N, Makita Y, de Hoon M, Nakai K (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. Nucleic Acids Res 36:D93–D96. https://doi.org/10.1093/nar/gkm910

frontiers in Immunology

SYSTEMATIC REVIEW
published: 01 March 2021
doi: 10.3389/fimmu.2021.629193

SARS-CoV-2 Causes a Different Cytokine Response Compared to Other Cytokine Storm-Causing Respiratory Viruses in Severely Ill Patients

Marton Olbei[1,2], *Isabelle Hautefort*[1], *Dezso Modos*[1,2], *Agatha Treveil*[1,2], *Martina Poletti*[1,2], *Lejla Gul*[1], *Claire D. Shannon-Lowe*[3] *and Tamas Korcsmaros*[1,2]*

[1] Earlham Institute, Norwich, United Kingdom, [2] Gut Microbes and Health Programme, Quadram Institute Bioscience, Norwich, United Kingdom, [3] Institute of Immunology and Immunotherapy, The University of Birmingham, Birmingham, United Kingdom

OPEN ACCESS

Edited by:
Kuldeep Dhama, Indian Veterinary Research Institute (IVRI), India

Reviewed by:
Chiranjib Chakraborty, Adamas University, India
Manojit Bhattacharya, Fakir Mohan University, India
Suliman Khan, Second Affiliated Hospital of Zhengzhou University, China

***Correspondence:**
Tamas Korcsmaros
tamas.korcsmaros@earlham.ac.uk

Specialty section:
This article was submitted to Viral Immunology, a section of the journal Frontiers in Immunology

Received: *20 November 2020*
Accepted: *29 January 2021*
Published: *01 March 2021*

Citation:
Olbei M, Hautefort I, Modos D, Treveil A, Poletti M, Gul L, Shannon-Lowe CD and Korcsmaros T (2021) SARS-CoV-2 Causes a Different Cytokine Response Compared to Other Cytokine Storm-Causing Respiratory Viruses in Severely Ill Patients. Front. Immunol. 12:629193. doi: 10.3389/fimmu.2021.629193

Hyper-induction of pro-inflammatory cytokines, also known as a cytokine storm or cytokine release syndrome (CRS), is one of the key aspects of the currently ongoing SARS-CoV-2 pandemic. This process occurs when a large number of innate and adaptive immune cells activate and start producing pro-inflammatory cytokines, establishing an exacerbated feedback loop of inflammation. It is one of the factors contributing to the mortality observed with coronavirus 2019 (COVID-19) for a subgroup of patients. CRS is not unique to the SARS-CoV-2 infection; it was prevalent in most of the major human coronavirus and influenza A subtype outbreaks of the past two decades (H5N1, SARS-CoV, MERS-CoV, and H7N9). With a comprehensive literature search, we collected changing the cytokine levels from patients upon infection with the viral pathogens mentioned above. We analyzed published patient data to highlight the conserved and unique cytokine responses caused by these viruses. Our curation indicates that the cytokine response induced by SARS-CoV-2 is different compared to other CRS-causing respiratory viruses, as SARS-CoV-2 does not always induce specific cytokines like other coronaviruses or influenza do, such as IL-2, IL-10, IL-4, or IL-5. Comparing the collated cytokine responses caused by the analyzed viruses highlights a SARS-CoV-2-specific dysregulation of the type-I interferon (IFN) response and its downstream cytokine signatures. The map of responses gathered in this study could help specialists identify interventions that alleviate CRS in different diseases and evaluate whether they could be used in the COVID-19 cases.

Keywords: SARS-CoV-2, cytokine response, influenza A, MERS- and SARS-CoV, literature analysis, systematic review

INTRODUCTION

The current coronavirus 2019 (COVID-19) pandemic has focused its attention on viral infectious diseases that the host antiviral immune response is unable to resolve. Major efforts are now concentrating on how severe acute respiratory syndrome β-coronavirus 2 (SARS-CoV-2) alters normal antiviral immune responses (1–3). SARS-CoV-2 causes a wide range of clinical symptoms

from asymptomatic, through mild fever, persistent cough, loss of taste and smell, to severe inflammation-driven pneumonia, resulting in multiple organ failure and ultimately death (4–6). SARS-CoV-2 induces an anti-inflammatory response attacking both the upper and lower respiratory tracts (7, 8). Although SARS-CoV-2 appears to modify host inflammatory defenses, similar modifications are also observed in other severe respiratory infections caused by viruses such as influenza A, β-coronaviruses SARS-CoV and MERS-CoV (9–11). These agents all constitute a global health threat with colossal economic consequences (12, 13).

Although these different viruses cause similar clinical symptoms, the pathogenesis may be driven by different triggers. Multiple studies have described an increase in the pro-inflammatory host immune response associated with severe forms of the diseases, including cytokine storms or cytokine release syndrome (CRS) (11, 14, 15). Although CRS usually resolves following completion of the antiviral response, it persists in severe cases (16). It can lead to tissue damage, multiple organ failure and death in critically-ill patients if the clinical intervention is not rapid (17, 18). In such cases, concentrations of both pro- and anti-inflammatory cytokines are significantly increased in blood and other tissues, including the type-I interferons (IFNs) (IFN-α, -β, -κ, -ε, -τ, -ω, and -ζ) (19–22). Type-I IFN signaling cascades also attenuate inflammation to avoid tissue damage during viral infection (23). The main effectors of the type-I IFN signaling are IFN-α and IFN-β, which activate other cytokines, such as IL-12 and the type-II IFN cytokine, IFN-γ (24, 25). However, cytokines such as IL-10 block the type-I IFN response. Certain pathogens, including SARS-CoV and MERS-CoV, encode proteins that can influence and delay the type-I IFN response leading to various pathologies (26–28). In the case of SARS-CoV, the build-up of activated macrophages in the lungs can cause tissue damage, while MERS-CoV can intensify engagement by neutrophils, leading to an increase in the production of pro-inflammatory cytokines (29–32). Furthermore, influenza A and coronavirus infections can trigger increased levels of type-I IFN-α and IFN-β, reflecting the normal initiation of this signaling pathway in response to viral infections (33–36). However, in severe infections with SARS-CoV-2, the type-I IFN signaling is impaired, culminating in an altered development of adaptive immunity (15, 37–39).

The similar clinical symptoms and the range of disease severity of different respiratory viral infections tend to blur the accuracy of the initial diagnosis (40, 41). Capturing a clear picture of the immune response triggered in each patient, early enough in infection remains challenging. It impairs the prevention of the severe form of the disease and, consequently, the potential onset of CRS. Defining the overlap and/or specificity in the patient immune cytokine signaling across CRS-causing viruses would help clinicians to develop a more tailored treatment strategy for future cases. Recent reviews have attempted to compare diseases caused by influenza A and β-coronaviruses (42–45). To provide mechanistic insight into the role of pro- and anti-inflammatory cytokines in the development of severe diseases caused by SARS-CoV, SARS-CoV-2, MERS-CoV, and influenza viruses, understanding the differences in cytokine responses between the different viruses is vital.

To identify the similarities and differences in the cytokine response, we collected and analyzed the patterns of cytokine changes caused by these CRS-causing respiratory viruses. By comparing available patient data from the literature, we were able to show (i) where similarities lie between the immune responses mounted against these pathogens, (ii) the differences between influenza A subtypes and coronaviruses and (iii) the unique aspects of the currently circulating SARS-CoV-2 virus.

METHODS

Literature Search

A mass literature search of 98 cytokines (46) was performed in PubMed using PubTator and in bioRxiv (https://www.biorxiv.org/) and medRxiv (https://www.medrxiv.org/) non-peer reviewed pre-publication repositories (47). This included the commonly studied interleukins, IFNs, tumor growth factors and chemokines involved in pro-inflammatory and anti-inflammatory responses, in particular, those associated with disease-associated CRS manifestations. Only studies indicating increase or no change in cytokine levels were included. The amplitude of change was not measured, only the presence or absence of it. We focused our study on five important CRS-causing viruses: two influenza A virus subtypes, H5N1 and H7N9, and three β-coronaviruses, SARS-CoV, MERS-CoV, and SARS-CoV-2 (**Figure 1**). We used the names of each virus and the cytokines and chemokines as search terms, e.g., "SARS-CoV-2 + CXCL10" (**Figure 1**). The collected studies were then screened to retain the studies using only patient-derived data, measured in at least 10 patients. A second pass was done adding "patient" to the search terms, e.g., "SARS-CoV-2 + CXCL10 + patient" in cases where the original search term yielded more than 50 hits. We only considered articles valid if they contained patient-derived data directly; the cell line or model organism-based results (and reviews) were excluded. From the main text of the resulting articles, we generated a table containing the presence of the queried cytokines and their direction of change in each disease. We closed the curation on March 06, 2020 (See **Supplementary Table 2** for the full list of queried cytokines). A script to generate the search URLs can be found in the publication of GitHub repository (https://github.com/korcsmarosgroup/CRS). The amount of discarded articles was estimated using custom python and shell scripts, also available in the publication repository.

Hierarchical Clustering

We clustered our data using the clustermap function from the python package seaborn with Jaccard distance and the complete linkage method (48). Jaccard distance calculates the distance between two sets of objects (49). Complete linkage clustering means that the distance from one cluster to another is calculated based on the furthest members of the cluster (50). The used clustering is sensitive for the furthest elements. Complete linkage does not join together with the furthest clusters, producing a

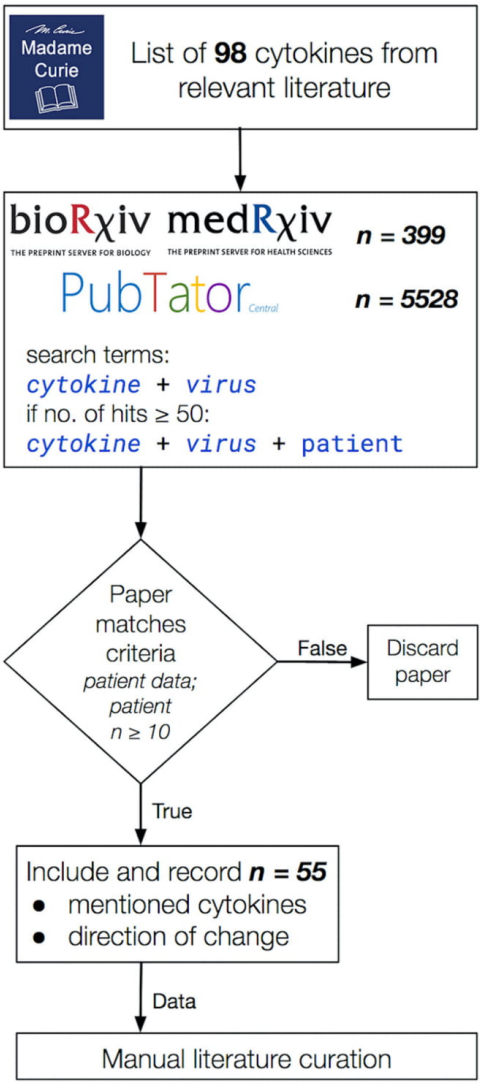

FIGURE 1 | The literature curation workflow applied in this study. Publications were considered valid for the inclusion into our data collection if (i) they contained patient-derived data (model organisms and cell lines were excluded), (ii) the study data were collected from cohorts of at least 10 participants per group and (iii) it included a directional change in cytokine levels. Total hits to queries in bioRxiv, medRxiv, and PubTator are shown separately in the second box from the top. In the end, 55 publications were selected that matched our curation criteria listed above.

clear picture. It performs well for finding the correct clusters in synthetic studies (51). We used all cytokine categories as input. The code is available at our GitHub repository (https://github.com/korcsmarosgroup/CRS).

RESULTS

In order to capture the breadth of the relevant published literature, we based our curation on a list of cytokines from the book chapter titled "Cytokines, Chemokines and Their Receptors" of the Madame Curie Bioscience Database (46) (**Figure 1**). We only used studies that reported the directional change of measured cytokines. Our curation approach allowed us to highlight shared and differing cytokine responses between influenza A and β-coronaviruses, contributing to further the understanding of why SARS-CoV-2 in particular differs so much not only from influenza A CRS-causing viruses but also from other β-coronaviruses, also capable of inducing a cytokine storm in severe cases.

β-coronaviruses and Influenza A Viruses Show Marked Differences in Some Cytokine Responses

Out of the nearly 100 cytokines measured across all initially-collected studies, only 38 were retained as they matched our criteria (See Methods section; **Supplementary Table 1**). Only a small group of cytokines was commonly measured for all viruses (CXCL8, IL-6, CXCL10, IL-2, IL-10, IFN-γ, and TNF-α). Across the 55 literature references used here (**Figure 1**), we first assessed how comparable the number of different cytokines measured in these studies was across the five CRS-causing viruses. **Figure 2** shows how variable this number is between virus-specific studies (e.g., 15 for H5N1 and 26 for SARS-CoV-2). This variation reflects (i) the increasing interest developed for CRS-causing pathologies over recent years (26 recent studies reported cytokine measurement for SARS-CoV-2 against only 10 H5N1-related studies) and (ii) the increased availability and sensitivity of the multiplex detection method.

The influenza A viruses trigger an increase in all cytokine levels measured (**Figure 2**, yellow). In contrast, during infection with each of the β-coronaviruses, some cytokines were detected at levels normally found in control groups (blue). This indicates that β-coronaviruses can subvert the immune response, reflecting different kinetics and pathogenesis between the influenza- and coronavirus-associated diseases. Of note, studies of H5N1 infections showed that a few cytokines were increased compared with control groups, and no change was observed in other studies (36, 52), illustrating the greater complexity of these diseases, probably due to the multifactorial nature of the mechanisms involved.

Table 1 shows the number of cytokines whose levels are increasing in one, two, three, four or all five virus-related infections from the interrogated literature. Only five cytokines were modulated regardless of the virus-associated disease concerned, with 20 other cytokines being shared to some degree. Increased levels observed in 16 cytokines were unique to a

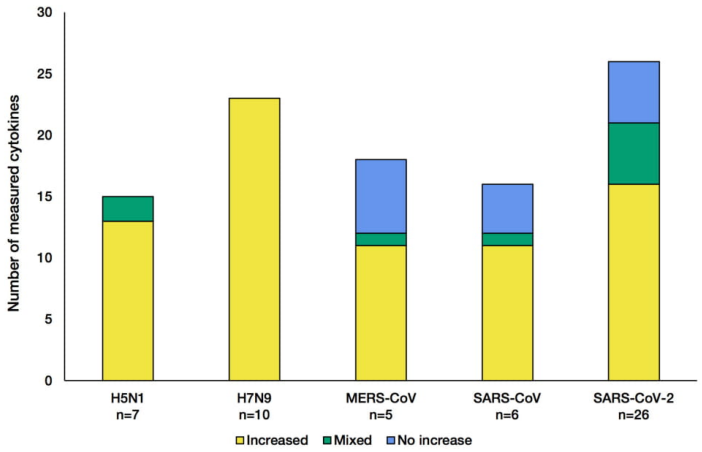

FIGURE 2 | Number of cytokines measured in the studies for each of the five CRS-causing viruses. Each stacked bar indicates how many cytokines were found at increased levels (yellow) in the blood/solid tissue of the patients, not changed (blue) or both increased and not changed across different studies of the same virus (green). The *n* number shown at the bottom of the bar charts corresponds to the number of articles citing cytokine changes during infection.

TABLE 1 | Number of cytokines which were elevated in at least one study.

	Cytokines elevated at least in one study (elevated and mixed)
Virus-specific	16
Shared between 2 viruses	5
Shared between 3 viruses	8
Shared between 4 viruses	2
Common to all 5 viruses	5

Cytokines measured in one or more of the virus-induced infections. Column 2 indicates the number of elevated or mixed measurements, and their overlap between viruses. Mixed observations occur when one or more studies show no change in a cytokine level upon infection, whereas others show an increase.

single virus at a time. It is important to keep in mind that the amplitude of change in the cytokines is not considered, which can be different between the different diseases, adding to the heterogeneity of those severe respiratory infectious diseases. This backs up the highly complex nature of the associated diseases as well as the past and current struggles to develop efficient vaccines and treatments.

To examine the presence of the measured cytokines and directionality of their change, we constructed a heatmap of the included viruses and cytokine responses.

The Cytokine Response to SARS-CoV-2 Sits in Between the Ones Given to Other β-coronaviruses and Influenza A Viruses

We used a hierarchic clustering algorithm on the viruses using Jaccard distance and complete linkage, clustering them based on the cytokine responses they cause. The method groups the pathogens in three clusters. SARS-CoV and MERS-CoV comprise the coronavirus cluster, and H5N1 and H7N9 form the influenza cluster, while SARS-CoV-2 sits in an individual cluster (**Figure 3**), slightly closer to the two influenza A viruses than to the two β-coronaviruses.

The cluster analysis of cytokines defines eight clusters, based on the direction of their modulation upon infection with each virus. It is important to note that the results of this cluster analysis are biased by the missing information for some cytokines. Bearing this in mind, it is worth looking into the detailed patterns of cytokine responses of the various CRS-inducing viruses. The cytokine cluster I includes the pro-inflammatory, TNF-α, and two anti-inflammatory cytokines, IL-2 and IL-10. All of them had mixed results in SARS-CoV-2, while encompassing all three categories of results for the other two coronavirus infections, which were predominantly increased during influenza infections. Unfortunately, cluster II seems to be restricted to cytokines measured only in H7N9-mediated infections, preventing us from comparing influenza A viruses vs. with β-coronaviruses. Clusters III and VI carry the generally increased pro-inflammatory cytokines, which are elevated for almost all of the viruses but not measured in all of the cases of cluster VI. Among those cytokines are IFN-α and IFN-γ, typical representatives of the general antiviral response (type-I and type-II IFNs), as well as IL-6, one of the most prominent pro-inflammatory cytokines, along various chemokines. Cytokines from Cluster IV measured during coronavirus infections do not fluctuate, while most of them are elevated during an influenza infection, e.g., IL-4 and IL-5 upon H7N9 infections. IL-4 is involved in Th2 differentiation, and the

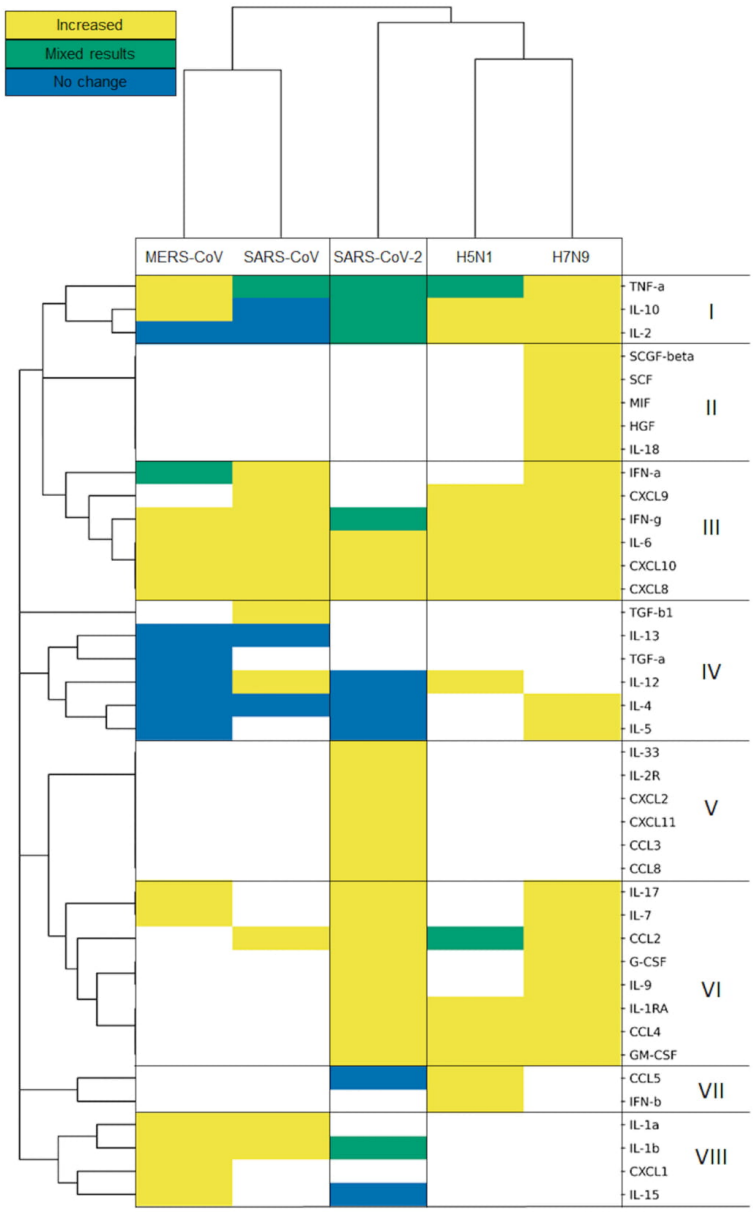

FIGURE 3 | Influenza viruses, SARS-CoV and MERS-CoV and SARS-CoV-2 form separated clusters (I–VIII) based on their cytokine response. Hierarchical clustering is based on Jaccard distance and complete linkage.

Th2 cells can produce IL-5 to mitigate eosinophil infiltration (53). Such differences observed between virus-specific pathologies reflect the strong alterations observed in coronavirus infections, particularly SARS-CoV-2 (54). The cytokines in Cluster VII and VIII do not always respond to SARS-CoV-2: IL-15 and CCL5 (RANTES) are not elevated after SARS-CoV-2 infection. IL-15 is involved in natural killer cell differentiation as part of an antiviral response (55). Meanwhile, CCL5 mediates eosinophil infiltration which is considered to be involved in the recovery after SARS-CoV infection (56). Clusters II and V contain cytokines measured only in H7N9 and SARS-CoV-2, respectively, whereas TGF-β1 was measured only in SARS-CoV studies in cluster IV.

Type-I IFN Signaling Can Be More Strongly Altered Upon Infection With SARS-CoV-2 Than in SARS-CoV- or MERS-CoV-infections

Both type-I and type-II IFNs play an instrumental role in the immune response to viral infection.

Our analysis indicates that early induction of type-I IFNs occurs upon H5N1 and H7N9 influenza A infection as well as upon the β-coronavirus SARS-CoV and MERS-CoV (21, 34, 57). However, type-I IFN response is only weakly elicited following a SARS-CoV-2 infection, if at all (37, 58).

Infection with either of the two influenza subtypes seems to increase the levels of measured type-I IFN-relevant cytokines, resulting in an antiviral immune response, with the appropriate cytokines showing elevated levels in all influenza A studies (**Figure 4**, **Supplementary Table 1**).

The β-coronavirus-mediated responses show a much more variable IFN response: with SARS-CoV, we see that the type-I IFN response is active, including the downstream-activated IL-12 that reflects the involvement of mature dendritic cells. IL-12 also indirectly activates IFN-γ further downstream. IL-10 is not elevated, which potentially prevents the downregulation of the type-I IFN response.

In MERS-CoV infections, the type-I IFN response is induced, but not in all cases (59). In some studies, the levels of IL-12 do not increase, in agreement with IFN-γ also staying at low levels. Yet, we see the involvement of the (mostly) anti-inflammatory IL-10. However, caution needs to be applied when looking at IL-10 in an inflammation context, as more and more clinical evidence suggests that this cytokine displays pro-inflammatory characteristics *in vivo* (60, 61).

We showed here that SARS-CoV-2-mediated infections are characterized by a clear dysregulation of type-I IFN response and, consequently, the downstream cytokine signatures, such as IL-4, IL-12, IL-2, and IL-10s, and the downstream type-II IFN response (**Figure 4**).

DISCUSSION

In this study, we analyzed relevant cytokine levels measured in patients, each infected with one of the five major respiratory viral pathogens, through a comprehensive literature curation of the published patient data. We generated a map of such responses to help specialists identify routes of interventions to successfully alleviate CRS in different diseases and evaluate whether they could be used in COVID-19 cases. Based on our literature curation, the five investigated viruses cause atypical cytokine responses in severely ill patients, reported here in **Figure 3**.

While most studies have focused on clinical or phylogenetic parameters (virus genome, patient age, transmissibility, fatality rate, creatinine, and coagulation among others), we aimed to add a mechanistic understanding to the host immune response. The cytokine response during viral infection is a dynamic process, with multiple changes in the cytokine levels during the course of the infection (62). During SARS and MERS infection, a slow initial innate immune response accompanied by the infection of alveolar macrophages leads to increased severity of these lower respiratory tract diseases (63–66). In contrast, SARS-CoV-2 seems to induce a number of cytokines at a very early stage, possibly explaining why the symptoms of severely ill patients deteriorate rapidly (67). A long-lasting pro-inflammatory cytokine production results in high mortality due to the development of severe conditions such as acute respiratory distress syndrome (ARDS) or acute lung injury [9.5% fatality rate for SARS and 34.4% for MERS compared to 2.3% for COVID-19 (43)].

Severe SARS patients show particularly low levels of the anti-inflammatory cytokine IL-10 (**Figures 3**, **4**) (68). During MERS infection, patients develop an expected increased production of IL-10, yet the low levels of IFN-γ-inhibiting IL-4 and IL-2 lead to elevated IFN-γ and the induction of type-II IFN response (**Figure 3**) (59, 69, 70). In contrast, during influenza A infection, the antiviral response activates without much delay with the presence of an intact negative feedback loop. Both viruses considered in our curation induce most of the pro- and anti-inflammatory cytokines downstream of type-I IFN response (**Figure 3**). Although influenza A viruses have effectors that dysregulate IFN-I (e.g., NS1, PB1-F2, polymerase proteins), the IFN-I response is nonetheless sustained, and its excessive activation during severe illness can lead to increased mortality. Furthermore, during H7N9 and H5N1 severe infections, TGF-β fails to be activated, contributing to increased pathogenicity (71–73). SARS-CoV-2 stands out from the other β-coronaviruses and influenza A viruses, with a highly perturbed response downstream of type-I IFN signaling, as reflected in the poor balance of measured pro-and anti-inflammatory cytokines (**Figures 3**, **4**). Of note, IFN-α was found to be increased (similar to the other viruses) only in one small ($n = 4$) patient study, which did not match our inclusion criteria. Type-II IFN-γ was also only increased in patients placed in intensive care units (ICUs), while it was within normal ranges in other studies (14, 74, 75).

Although the cytokine signaling enabling the reduction of the inflammatory environment is active (**Figures 3**, **4**), both influenza viruses H5N1 and H7N9 can cause CRS. In severe cases of infection, CRS could result from insufficient production of important cytokines such as TGF-β (73). Furthermore, the presence of impaired and less abundant effector CD4+ and CD8+ T cells was found to be a characteristic feature

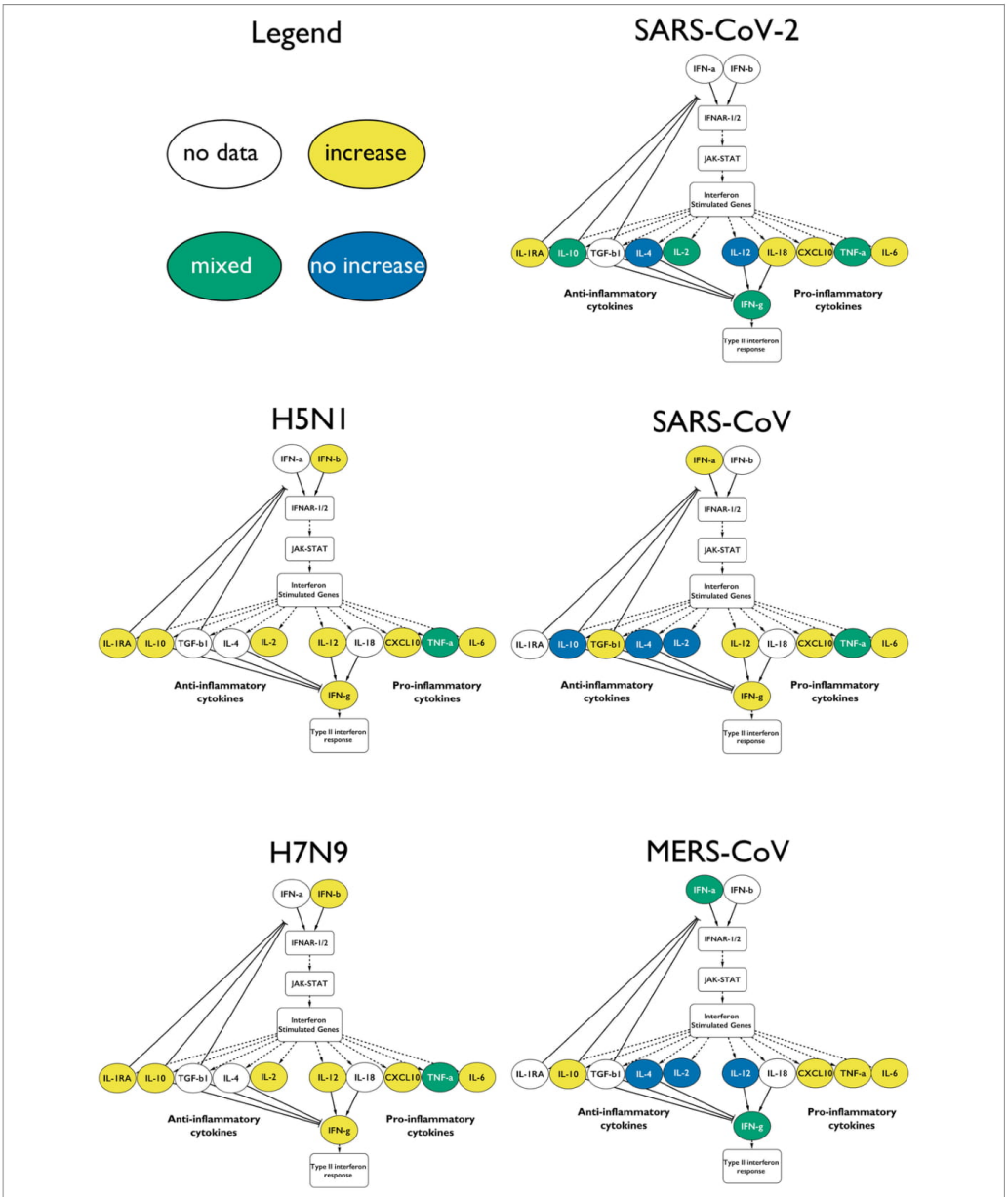

FIGURE 4 | Type-I IFN response upon infection with the different CRS-causing viruses. The measured cytokines in the influenza virus infected patients are increased. In the case of the coronaviruses, the responses are mixed, and not all of the anti-inflammatory cytokines are elevated. Only a fraction of the cytokines is depicted for clarity: yellow for increase in that virus, green for mixed results and blue for no change.

accompanying CRS in those diseases. Finally, monocytes that normally would differentiate from a pro-inflammatory state to an anti-inflammatory state with enhanced antigen presentation activity as the infection progresses remain in a chronic pro-inflammatory activation state, preventing the normal resolution of the host response (16, 76, 77). In future studies, patient-derived data including the size and activation status of innate and adaptive immune cell populations would help increase the understanding of CRS mechanisms in influenza-mediated diseases.

In our study, we found resolution of the pro-inflammatory immune response to be a key difference between coronaviruses (MERS-CoV and SARS-CoV) and influenza viruses (H5N1 and H7N9). Both MERS-CoV and SARS-CoV induce CRS, yet they also appear to impair the normal resolution of the antiviral immune response. In contrast, H5N1 and H7N9 induce high levels of pro- and anti-inflammatory cytokine levels in severe cases, leading to an inflammatory cytokine storm, yet leaving the immune system unimpeded to move toward a general resolution of the antiviral response appears in **Figures 3**, **4**) (36). However, SARS-CoV-2 induction of the CRS is eventually followed by a resolution of the pro-inflammatory responses in 80% of the cases.

One limitation of this study is the lack of anatomical and dynamic dimensions of the cytokine response. Firstly, the set of cytokines measured in the peripheral blood of each patient across the entire disease course or following recovery varied across the studies analyzed. Patients were sampled at different stages of the disease, which further add to the noise observed in the data. Finally, systematic patient-based studies matching our strict curation criteria could not be collected, leaving many gaps in our comparisons (**Figure 3**, white cells).

While confirming many already reported disease traits, our analysis has highlighted several new features that are shared or different between the viral diseases analyzed, contributing to filling the gap in the understanding of SARS-CoV-2 and other CRS-causing viruses. Blockage of the cytokine response in SARS-CoV-2 infection through IL-6 specific antibody has failed during Phase 3 randomized clinical trial (NCT04320615), even with promising results in earlier stages (78–80), suggesting that further mechanistic investigation of the cytokine storms during SARS-CoV-2 infection will be needed.

The ongoing accumulation of patient-derived large data sets will inform the research community and clinicians of the intricacy of host/virus interactions (81). Systematic reviews such as this study should be part of an iterative process, increasing the resolution of the comparisons listed above, by continuously integrating novel data. Recently published data and literature repositories, such as H2V and LitCovid, can further enhance the effectiveness of this iterative process (82, 83). In this study, we provided an example of this through a literature curation of patient-derived data and a comparative map across CRS-causing β-coronaviruses and influenza A viruses, linking shared or specific changing cytokines and interferon signaling alterations to those pathogens. In this study, we provided the methodology and scripts to perform this iterative analysis easier in the future.

CONCLUSIONS

Using our literature curation workflow, we showed that based on available patient data, SARS-CoV-2 generates a different cytokine response compared to other CRS causing respiratory viruses. SARS-CoV-2 does not elevate all of the expected cytokines in patients as the other studied respiratory viruses, e.g., the cytokines following an influenza infection such as IL-2, IL-10, IL-4, or IL-5. Although for a subset of pro-inflammatory cytokines, SARS-CoV-2 does induce a similar response to the compared viruses, the literature reports conflicting results for a few important cytokines such as IFN-γ and IL-1β. Applying the collected data to the type-I IFN cascade, the cytokine signature indicates a dysregulation of this process and that of the downstream type-II IFN responses, involving cytokines such as the aforementioned IL-10, IL-2, IL-4, or IL-12.

In our systematic analysis, we collated a map of patient-derived cytokine responses given to different CRS-causing viruses. Our goal is that such a resource of unique and conserved cytokine responses will aid specialists to identify interventions that can alleviate serious cases of COVID-19 and other illnesses that cause CRS.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article's **Supplementary Material** and available at https://github.com/korcsmarosgroup/CRS.

AUTHOR CONTRIBUTIONS

MO and IH collected and analyzed the literature data with DM, and wrote the manuscript together, with contributions from AT, MP, LG, and CS-L. DM performed the clustering analysis for **Figure 3**. TK supervised the project. All authors discussed the results and contributed to the final manuscript.

FUNDING

MO, AT, LG, and MP were supported by the UKRI Biotechnological and Biosciences Research Council (BBSRC) funded by Norwich Research Park Biosciences Doctoral Training Partnership (grant numbers BB/M011216/1 and BB/S50743X/1). The work of TK, DM, and IH were supported by the Earlham Institute (Norwich, UK) in partnership with the Quadram Institute (Norwich, UK) and strategically supported by the UKRI BBSRC UK grants (BB/J004529/1, BB/P016774/1, and BB/CSP17270/1). CS-L was supported by MRC MR/N023781/1 and the Histiocytosis Society, USA. TK and DM were also funded by a BBSRC ISP grant for Gut Microbes and Health BB/R012490/1 and its constituent projects, BBS/E/F/000PR10353 and BBS/E/F/000PR10355.

ACKNOWLEDGMENTS

We thank the current and past members of the Korcsmaros group and the COVID-19 Disease Map Community for their ideas and support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2021.629193/full#supplementary-material

REFERENCES

1. Luo W, Li Y-X, Jiang L-J, Chen Q, Wang T, Ye D-W. Targeting JAK-STAT signaling to control cytokine release syndrome in COVID-19. *Trends Pharmacol Sci.* (2020) 41:531–43. doi: 10.1016/j.tips.2020.06.007
2. Conti P, Ronconi G, Caraffa A, Gallenga C, Ross R, Frydas I, et al. Induction of pro-inflammatory cytokines (IL-1 and IL-6) and lung inflammation by Coronavirus-19 (COVI-19 or SARS-CoV-2): anti-inflammatory strategies. *J Biol Regul Homeost Agents.* (2020) 34:327–31. doi: 10.23812/CONTI-E
3. McGonagle D, Sharif K, O'Regan A, Bridgewood C. The role of cytokines including interleukin-6 in COVID-19 induced pneumonia and macrophage activation syndrome-like disease. *Autoimmun Rev.* (2020) 19:102537. doi: 10.1016/j.autrev.2020.102537
4. Omarjee L, Janin A, Perrot F, Laviolle B, Meilhac O, Mahe G. Targeting T-cell senescence and cytokine storm with rapamycin to prevent severe progression in COVID-19. *Clin Immunol.* (2020) 216:108464. doi: 10.1016/j.clim.2020.108464
5. Rahmati M, Moosavi MA. Cytokine-targeted therapy in severely ill COVID-19 patients: options and cautions. *EJMO.* (2020) 4:179–81. doi: 10.14744/ejmo.2020.72142
6. Bouhaddou M, Memon D, Meyer B, White KM, Rezelj VV, Correa Marrero M, et al. The global phosphorylation landscape of SARS-CoV-2 infection. *Cell.* (2020) 182:685–712.e19. doi: 10.1016/j.cell.2020.06.034
7. Neumann J, Prezzemolo T, Vanderbeke L, Roca CP, Gerbaux M, Janssens S, et al. An open resource for T cell phenotype changes in COVID-19 identifies IL-10-producing regulatory T cells as characteristic of severe cases. *medRxiv [Preprint].* (2020). doi: 10.1101/2020.05.31.20112979
8. Liu Y, Zhang C, Huang F, Yang Y, Wang F, Yuan J, et al. Elevated levels of plasma cytokines in COVID-19 reflect viral load and lung injury. *Natl Sci Rev.* (2020) 7:1003–11. doi: 10.1093/nsr/nwaa037
9. Totura AL, Baric RS. SARS coronavirus pathogenesis: host innate immune responses and viral antagonism of interferon. *Curr Opin Virol.* (2012) 2:264–75. doi: 10.1016/j.coviro.2012.04.004
10. Nelemans T, Kikkert M. Viral innate immune evasion and the pathogenesis of emerging RNA virus infections. *Viruses.* (2019) 11:961. doi: 10.3390/v11100961
11. Huang K-J, Su I-J, Theron M, Wu Y-C, Lai S-K, Liu C-C, et al. An interferon-gamma-related cytokine storm in SARS patients. *J Med Virol.* (2005) 75:185–94. doi: 10.1002/jmv.20255
12. Wei P, Cai Z, Hua J, Yu W, Chen J, Kang K, et al. Pains and gains from China's experiences with emerging epidemics: from SARS to H7N9. *Biomed Res Int.* (2016) 2016:5717108. doi: 10.1155/2016/5717108
13. Munster VJ, Koopmans M, van Doremalen N, van Riel D, de Wit E. A novel coronavirus emerging in China - key questions for impact assessment. *N Engl J Med.* (2020) 382:692–4. doi: 10.1056/NEJMp2000929
14. Pedersen SF, Ho Y-C. SARS-CoV-2: a storm is raging. *J Clin Invest.* (2020) 130:2202–5. doi: 10.1172/JCI137647
15. Ye Q, Wang B, Mao J. The pathogenesis and treatment of the 'cytokine storm' in COVID-19. *J Infect.* (2020) 80:607–13. doi: 10.1016/j.jinf.2020.03.037
16. Wauters E, Van Mol P, Garg AD, Jansen S, Van Herck Y, Vanderbeke L, et al. Discriminating mild from critical COVID-19 by innate and adaptive immune single-cell profiling of bronchoalveolar lavages. *Cell Res.* (2021). doi: 10.1038/s41422-020-00455-9. [Epub ahead of print].
17. Tisoncik JR, Korth MJ, Simmons CP, Farrar J, Martin TR, Katze MG. Into the eye of the cytokine storm. *Microbiol Mol Biol Rev.* (2012) 76:16–32. doi: 10.1128/MMBR.05015-11
18. Shimabukuro-Vornhagen A, Gödel P, Subklewe M, Stemmler HJ, Schlößer HA, Schlaak M, et al. Cytokine release syndrome. *J Immunother Cancer.* (2018) 6:56. doi: 10.1186/s40425-018-0343-9
19. Kalliolias GD, Ivashkiv LB. Overview of the biology of type I interferons. *Arthritis Res Ther.* (2010) 12(Suppl. 1):S1. doi: 10.1186/ar2881
20. Borden EC, Sen GC, Uze G, Silverman RH, Ransohoff RM, Foster GR, et al. Interferons at age 50: past, current and future impact on biomedicine. *Nat Rev Drug Discov.* (2007) 6:975–90. doi: 10.1038/nrd2422
21. Mi Z, Ma Y, Tong Y. Avian influenza virus H5N1 induces rapid interferon-beta production but shows more potent inhibition to retinoic acid-inducible gene I expression than H1N1 *in vitro*. *Virol J.* (2012) 9:145. doi: 10.1186/1743-422X-9-145
22. Brassard DL, Grace MJ, Bordens RW. Interferon-α as an immunotherapeutic protein. *J Leukoc Biol.* (2002) 71:565–81. doi: 10.1189/jlb.71.4.565
23. Lee AJ, Ashkar AA. The dual nature of type I and type II interferons. *Front Immunol.* (2018) 9:2061. doi: 10.3389/fimmu.2018.02061
24. Kang S, Brown HM, Hwang S. Direct antiviral mechanisms of interferon-gamma. *Immune Netw.* (2018) 18:e33. doi: 10.4110/in.2018.18.e33
25. Bhardwaj N, Seder RA, Reddy A, Feldman MV. IL-12 in conjunction with dendritic cells enhances antiviral CD8+ CTL responses *in vitro*. *J Clin Invest.* (1996) 98:715–22. doi: 10.1172/JCI118843
26. Guix S, Pérez-Bosque A, Miró L, Moretó M, Bosch A, Pintó RM. Type I interferon response is delayed in human astrovirus infections. *PLoS ONE.* (2015) 10:e0123087. doi: 10.1371/journal.pone.0123087
27. Murira A, Lamarre A. Type-I interferon responses: from friend to foe in the battle against chronic viral infection. *Front Immunol.* (2016) 7:609. doi: 10.3389/fimmu.2016.00609
28. To KKW, Hung IFN, Chan JFW, Yuen K-Y. From SARS coronavirus to novel animal and human coronaviruses. *J Thorac Dis.* (2013) 5(Suppl. 2):S103–8. doi: 10.3978/j.issn.2072-1439.2013.06.02
29. Channappanavar R, Fehr AR, Zheng J, Wohlford-Lenane C, Abrahante JE, Mack M, et al. IFN-I response timing relative to virus replication determines MERS coronavirus infection outcomes. *J Clin Investig.* (2019) 29:3625–39. doi: 10.1172/JCI126363
30. Channappanavar R, Fehr AR, Vijay R, Mack M, Zhao J, Meyerholz DK, et al. Dysregulated type I interferon and inflammatory monocyte-macrophage responses cause lethal pneumonia in SARS-CoV-infected mice. *Cell Host Microbe.* (2016) 19:181–93. doi: 10.1016/j.chom.2016.01.007
31. Okabayashi T, Kariwa H, Yokota S, Iki S, Indoh T, Yokosawa N, et al. Cytokine regulation in SARS coronavirus infection compared to other respiratory virus infections. *J Med Virol.* (2006) 78:417–24. doi: 10.1002/jmv.20556
32. Payne B, Bellamy R. Novel respiratory viruses: what should the clinician be alert for? *Clin Med.* (2014) 14(Suppl. 6):s12–6. doi: 10.7861/clinmedicine.14-6-s12
33. Califano D, Furuya Y, Roberts S, Avram D, McKenzie ANJ, Metzger DW. IFN-γ increases susceptibility to influenza A infection through suppression of group II innate lymphoid cells. *Mucosal Immunol.* (2018) 11:209–19. doi: 10.1038/mi.2017.41
34. Yao Z, Zheng Z, Wu K, Junhua Z. Immune environment modulation in pneumonia patients caused by coronavirus: SARS-CoV, MERS-CoV and SARS-CoV-2. *Aging.* (2020) 12:7639–51. doi: 10.18632/aging.103101
35. Zeng H, Belser JA, Goldsmith CS, Gustin KM, Veguilla V, Katz JM, et al. A(H7N9) virus results in early induction of proinflammatory cytokine responses in both human lung epithelial and endothelial cells and shows increased human adaptation compared with avian H5N1 virus. *J Virol.* (2015) 89:4655–67. doi: 10.1128/JVI.03095-14

36. de Jong MD, Simmons CP, Thanh TT, Hien VM, Smith GJD, Chau TNB, et al. Fatal outcome of human influenza A (H5N1) is associated with high viral load and hypercytokinemia. *Nat Med.* (2006) 12:1203–7. doi: 10.1038/nm1477
37. Acharya D, Liu G, Gack MU. Dysregulation of type I interferon responses in COVID-19. *Nat Rev Immunol.* (2020) 20:397–8. doi: 10.1038/s41577-020-0346-x
38. Ma Y, Wang C, Xue M, Fu F, Zhang X, Li L, et al. The coronavirus transmissible gastroenteritis virus evades the type i interferon response through IRE1α-mediated manipulation of the MicroRNA miR-30a-5p/SOCS1/3 Axis. *J Virol.* (2018) 92:e00728-18. doi: 10.1128/JVI.00728-18
39. Bost P, Giladi A, Liu Y, Bendjelal Y, Xu G, David E, et al. Host-viral infection maps reveal signatures of severe COVID-19 patients. *Cell.* (2020) 181:1475–88.e12. doi: 10.1016/j.cell.2020.05.006
40. Boncristiani HF, Criado MF, Arruda E. Respiratory viruses. In: Schaechter M, editor. *Encyclopedia of Microbiology*. Oxford: Elsevier (2009). p. 500–18.
41. Caini S, Kroneman M, Wiegers T, El Guerche-Séblain C, Paget J. Clinical characteristics and severity of influenza infections by virus type, subtype, and lineage: a systematic literature review. *Influenza Other Respir Viruses.* (2018) 12:780–92. doi: 10.1111/irv.12575
42. Jiang C, Yao X, Zhao Y, Wu J, Huang P, Pan C, et al. Comparative review of respiratory diseases caused by coronaviruses and influenza A viruses during epidemic season. *Microbes Infect.* (2020) 22:236–44. doi: 10.1016/j.micinf.2020.05.005
43. Petrosillo N, Viceconte G, Ergonul O, Ippolito G, Petersen E. COVID-19, SARS and MERS: are they closely related? *Clin Microbiol Infect.* (2020) 26:729–34. doi: 10.1016/j.cmi.2020.03.026
44. Chu H, Chan JF-W, Yuen TT-T, Shuai H, Yuan S, Wang Y, et al. Comparative tropism, replication kinetics, and cell damage profiling of SARS-CoV-2 and SARS-CoV with implications for clinical manifestations, transmissibility, and laboratory studies of COVID-19: an observational study. *Lancet Microbe.* (2020) 1:e14–23. doi: 10.1016/S2666-5247(20)30004-5
45. Ceccarelli M, Berretta M, Venanzi Rullo E, Nunnari G, Cacopardo B. Differences and similarities between Severe Acute Respiratory Syndrome (SARS)-CoronaVirus (CoV) and SARS-CoV-2. Would a rose by another name smell as sweet? *Eur Rev Med Pharmacol Sci.* (2020) 24:2781–3. doi: 10.26355/eurrev_202003_20551
46. Cameron MJ, Kelvin DJ. Cytokines and chemokines-their receptors and their genes: an overview. *Adv. Exp. Med. Biol.* (2003) 520:8–32. doi: 10.1007/978-1-4615-0171-8_2
47. Wei C-H, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.* (2019) 47:W587–93. doi: 10.1093/nar/gkz389
48. Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline DC, et al. *Mwaskom/Seaborn: V0.8.1 (September 2017)*. Zenodo (2017).
49. Jaccard P. The distribution of the flora in the alpine zone. *New Phytol.* (1912) 11:37–50. doi: 10.1111/j.1469-8137.1912.tb05611.x
50. Mcquitty LL. Comprehensive hierarchical analysis. *Educ Psychol Meas.* (1960) 20:805–16. doi: 10.1177/001316446002000418
51. Ferreira L, Hitchcock DB. A comparison of hierarchical methods for clustering functional data. *Commun Stat Simul Comput.* (2009) 38:1925–49. doi: 10.1080/03610910903168603
52. Wu C, Lu X, Wang X, Jin T, Cheng X, Fang S, et al. Clinical symptoms, immune factors, and molecular characteristics of an adult male in Shenzhen, China infected with influenza virus H5N1. *J Med Virol.* (2013) 85:760–8. doi: 10.1002/jmv.23492
53. Guo X-ZJ, Thomas PG. New fronts emerge in the influenza cytokine storm. *Semin Immunopathol.* (2017) 39:541–50. doi: 10.1007/s00281-017-0636-y
54. Tan L, Wang Q, Zhang D, Ding J, Huang Q, Tang Y-Q, et al. Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. *Signal Transduct Target Ther.* (2020) 5:33. doi: 10.1038/s41392-020-0148-4
55. Guo Y, Luan L, Patil NK, Sherwood ER. Immunobiology of the IL-15/IL-15Rα complex as an antitumor and antiviral agent. *Cytok Growth Factor Rev.* (2017) 38:10–21. doi: 10.1016/j.cytogfr.2017.08.002
56. Russell CD, Unger SA, Walton M, Schwarze J. The human immune response to respiratory syncytial virus infection. *Clin Microbiol Rev.* (2017) 30:481–502. doi: 10.1128/CMR.00090-16
57. Zhou J, Wang D, Gao R, Zhao B, Song J, Qi X, et al. Biological features of novel avian influenza A (H7N9) virus. *Nature.* (2013) 499:500–3. doi: 10.1038/nature12379
58. Wei L, Ming S, Zou B, Wu Y, Hong Z, Li Z, et al. Viral invasion and type I interferon response characterize the immunophenotypes during COVID-19 infection. *SSRN Electron J.* (2020). doi: 10.2139/ssrn.3564998. [Epub ahead of print].
59. Chan JFW, Lau SKP, To KKW, Cheng VCC, Woo PCY, Yuen K-Y. Middle East respiratory syndrome coronavirus: another zoonotic betacoronavirus causing SARS-like disease. *Clin Microbiol Rev.* (2015) 28:465–522. doi: 10.1128/CMR.00102-14
60. Mühl H. Pro-inflammatory signaling by IL-10 and IL-22: bad habit stirred up by interferons? *Front Immunol.* (2013) 4:18. doi: 10.3389/fimmu.2013.00018
61. Lauw FN, Pajkrt D, Hack CE, Kurimoto M, van Deventer SJ, van der Poll T. Proinflammatory effects of IL-10 during human endotoxemia. *J Immunol.* (2000) 165:2783–9. doi: 10.4049/jimmunol.165.5.2783
62. Channappanavar R, Perlman S. Pathogenic human coronavirus infections: causes and consequences of cytokine storm and immunopathology. *Semin Immunopathol.* (2017) 39:529–39. doi: 10.1007/s00281-017-0629-x
63. Drosten C, Günther S, Preiser W, van der Werf S, Brodt H-R, Becker S, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med.* (2003) 348:1967–76. doi: 10.1056/NEJMoa030747
64. Peiris JSM, Lai ST, Poon LLM, Guan Y, Yam LYC, Lim W, et al. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet.* (2003) 361:1319–25. doi: 10.1016/S0140-6736(03)13077-2
65. van Boheemen S, de Graaf M, Lauber C, Bestebroer TM, Raj VS, Zaki AM, et al. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *MBio.* (2012) 3:e00473-12. doi: 10.1128/mBio.00473-12
66. Kuiken T, Fouchier RAM, Schutten M, Rimmelzwaan GF, van Amerongen G, van Riel D, et al. Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. *Lancet.* (2003) 362:263–70. doi: 10.1016/S0140-6736(03)13967-0
67. Sun J, Ye F, Wu A, Yang R, Pan M, Sheng J, et al. Comparative transcriptome analysis reveals the intensive early stage responses of host cells to SARS-CoV-2 infection. *Front Microbiol.* (2020) 11:593857. doi: 10.3389/fmicb.2020.593857
68. Chien J-Y, Hsueh P-R, Cheng W-C, Yu C-J, Yang P-C. Temporal changes in cytokine/chemokine profiles and pulmonary involvement in severe acute respiratory syndrome. *Respirology.* (2006) 11:715–22. doi: 10.1111/j.1440-1843.2006.00942.x
69. Mahallawi WH, Khabour OF, Zhang Q, Makhdoum HM, Suliman BA. MERS-CoV infection in humans is associated with a pro-inflammatory Th1 and Th17 cytokine profile. *Cytokine.* (2018) 104:8–13. doi: 10.1016/j.cyto.2018.01.025
70. Kim ES, Choe PG, Park WB, Oh HS, Kim EJ, Nam EY, et al. Clinical progression and cytokine profiles of middle east respiratory syndrome coronavirus infection. *J Korean Med Sci.* (2016) 31:1717–25. doi: 10.3346/jkms.2016.31.11.1717
71. Neumann G. H5N1 influenza virulence, pathogenicity and transmissibility: what do we know? *Future Virol.* (2015) 10:971–80. doi: 10.2217/fvl.15.62
72. Ramos I, Fernandez-Sesma A. Innate immunity to H5N1 influenza viruses in humans. *Viruses.* (2012) 4:3363–88. doi: 10.3390/v4123363
73. Koutsakos M, Kedzierska K, Subbarao K. Immune responses to avian influenza viruses. *J Immunol.* (2019) 202:382–91. doi: 10.4049/jimmunol.1801070
74. Wei X-S, Wang X-R, Zhang J-C, Yang W-B, Ma W-L, Yang B-H, et al. A cluster of health care workers with COVID-19 pneumonia caused by SARS-CoV-2. *J Microbiol Immunol Infect.* (2020) doi: 10.1016/j.jmii.2020.04.013
75. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet.* (2020) 395:497–506. doi: 10.1016/S0140-6736(20)30183-5
76. Zhou Y, Fu B, Zheng X, Wang D, Zhao C, qi Y, et al. Pathogenic T cells and inflammatory monocytes incite inflammatory storm in severe COVID-19 patients. *Natl Sci Rev.* (2020) 7:998. doi: 10.1093/nsr/nwaa041
77. Diao B, Wang C, Tan Y, Chen X, Liu Y, Ning L, et al. Reduction and functional exhaustion of T cells in patients with coronavirus disease 2019 (COVID-19). *Front Immunol.* (2020) 11:827. doi: 10.3389/fimmu.2020.00827
78. Eimer J, Vesterbacka J, Svensson AK, Stojanovic B, Wagrell C, Sönnerborg A, et al. Tocilizumab shortens time on mechanical ventilation and length of

hospital stay in patients with severe COVID-19: a retrospective cohort study. *J Intern Med.* (2020). doi: 10.1111/joim.13162. [Epub ahead of print].
79. Saha A, Sharma AR, Bhattacharya M, Sharma G, Lee S-S, Chakraborty C. Tocilizumab: a therapeutic option for the treatment of cytokine storm syndrome in COVID-19. *Arch Med Res.* (2020) 51:595–7. doi: 10.1016/j.arcmed.2020.05.009
80. Chakraborty C, Sharma AR, Bhattacharya M, Sharma G, Lee S-S, Agoramoorthy G. COVID-19: consider IL-6 receptor antagonist for the therapy of cytokine storm syndrome in SARS-CoV-2 infected patients. *J Med Virol.* (2020) 92:2260–2. doi: 10.1002/jmv.26078
81. Treveil A, Bohar B, Sudhakar P, Gul L, Csabai L, Olbei M, et al. ViralLink: an integrated workflow to investigate the effect of SARS-CoV-2 on intracellular signalling and regulatory pathways. *PLoS Comput Biol.* (2021) 17:e1008685. doi: 10.1371/journal.pcbi.1008685
82. Zhou N, Bao J, Ning Y. H2V: a database of human genes and proteins that respond to SARS-CoV-2, SARS-CoV, and MERS-CoV infection. *BMC Bioinformatics.* (2021) 22:18. doi: 10.1186/s12859-020-03935-2
83. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res.* (2021) 49:D1534–40. doi: 10.1093/nar/gkaa952

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.