ABSTRACT

Title of dissertation: SOME STATISTICAL AND DYNAMICAL MODELS FOR THE ANALYSIS OF MICROBIAL ECOSYSTEMS AND THEIR GENOMIC DATA

Senthilkumar Muthiah
Doctor of Philosophy, 2019

Dissertation directed by: Professor Héctor corrada Bravo
Department of Computer Science

Embedded within their genetic makeup and ecology, microbes harbor unparalleled stories on natural selection, evolution and biomedicine. In modern biology, such stories are elucidated through rigorous interrogation of microbial ecosystems with a variety of theoretic and experimental techniques. These range from abstract, isolated mathematical models to high-resolution sequencing technologies that probe every single nucleotide of a cell's DNA. It is clear that inferences thus obtained are markedly sensitive to the unforeseen technical variability introduced during an experiment, and are limited by the tractability and robustness of the models in generating sound hypotheses. We have developed statistical and computational tools to advance statistical inference for microbial genomics by overcoming a subset of technical biases, and have explored certain interesting cases of microbial interactions and their evolution by developing tractable mathematical models.

Compositional bias induced by the sequencing machine. A DNA sequencing machine produces only percentage measurements (fraction molecules of a given type) of the

DNA molecules in its input. When contrasting measurements from different inputs, one therefore obtains confounded inferences on absolute concentrations (molecules per unit volume). We theoretically analyze this *compositional bias* problem with significant generality, and exploit it to develop an empirical Bayes approach to solve it under certain assumptions with particular emphasis on microbial sequencing technologies.

Suicidal attributes of prokaryotic adaptive immunity. The recently discovered CRISPR systems provide the first examples of bacterial and archaeal adaptive immune systems operating against invading viruses over ecological time scales. Equally surprising as their adaptive nature, is their ability to induce high rates of host autoimmunity. We theoretically analyze the ecological and evolutionary dynamics of such a costly defense mechanism in simplified models of prokaryote-phage coevolution. We show that by allowing for regulated post-infection activation, CRISPRs can function by exploiting a dual defense strategy of abortive infection and anti-viral resistance.

Additional statistical and analytic extensions for some related questions on clustering and multi-resolution analysis also appear.

SOME STATISTICAL AND DYNAMICAL MODELS FOR THE ANALYSIS OF MICROBIAL ECOSYSTEMS AND THEIR GENOMIC DATA

by

Senthilkumar Muthiah

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Héctor Corrada Bravo, Chair/Advisor
Professor Eric V. Slud
Professor Sridhar Hannehalli
Professor Mihai Pop
Professor Doron Levy

Foreword

Material from parts I and II of this thesis, respectively appear in the following research publications.

1. Kumar, M.S., Joshua B. Plotkin, and Sridhar Hannenhalli. "Regulated CRISPR Modules Exploit a Dual Defense Strategy of Restriction and Abortive Infection in a Model of Prokaryote-Phage Coevolution." PLoS Computational Biology (2015).

   † *Authors' contributions: Conceived and designed study: MSK SH. Performed the experiments: MSK. Analyzed data: MSK. Contributed analytical tools: MSK JBP. Wrote the paper: MSK JBP SH.*

2. Kumar, M.S., Eric V. Slud, Kwame Okrah, Stephanie C. Hicks, Sridhar Hannenhalli, and Héctor Corrada Bravo. "Analysis and Correction of Compositional Bias in Sparse Sequencing Count Data." BMC Genomics (2018).

   † *Authors' contributions: Conceived and designed study: MSK HCB. Performed the experiments: MSK. Contributed analytical tools: MSK EVS KO HCB. Data analysis and interpretation: MSK EVS KO HCB. Wrote the paper: MSK SH SCH EVS HCB.*

Acknowledgments

Deeply satisfiying scientific research often starts off with a well defined question and a fragment of an idea to address it, and ultimately burgeons to a full fledged scientific paper with rigorous logical discipline. As fascinating as that process sounds, working through it, a graduate student can experience sky-highs and hellish-lows that cast severe self-doubts. Looking back now, as I witness the winter migration of a giant flock of ca. 8000 arctic snow geese along the Delaware Bay, I know, for sure, I am able to navigate through them because of a loving family, encouraging friends and wisdomatic mentors.

I wish to thank my mentors Mukund Thattai and Russell S. Schwartz for inspiring confidence in scientific research; for me, it all started with them. Profs. Kevin Chen, Eduardo Sontag, and P.S. Thiagarajan made sure that that enthusiasm did not plateau off. I am indebted to Profs. Markus Deserno, Sridhar Hannenhalli, Doron Levy, Joshua Plotkin, Niranjan Nagarajan, and Eric V. Slud for their insightful lessons, thorough brainstorming sessions, and fantastic collaborations that altogether elevated the quality of my research experience beyond what I had wished for. Finally, Prof. Mihai Pop and my otherworldly gift of a PhD advisor, Prof. Héctor Corrada Bravo taught and patiently guided me the rest of the way ensuring that I graduate happily, all the while helping me lead a very comfortable life with enriching advice and uninterrupted funding.

Thanks also to all my friends Mathieu Almeida, Rajat Anand, Lokheshvar Balakumar, Claudia Bancila, Tristan Bereau, Denis Bertrand, Manikandan Chandran, Faezeh Dorri, Mohamed Geunady, Luis Guardado, Broc Gullet, Sarika Hegde, Joyce Hsiao, Keith Hughitt, Asif Javed, Debashis & Subhra Kar, Aparna Lakshmanan, Byoungkoo

Lee, Rajarajan Loganathan, Piötr Mardziel, Lee Mendelowitz, Vasanth P. Murari, Vinutha Nagaraj, Sanjanaa Nagarajan, Satyajeet Ojha, Kwame Okrah, Nathaniel Olson, Joseph N. Paulson, Constantine Pop, Elisabet Pujadas, Navneet Rai, Anand Babu Rajendran, Vigneshwar Ramakrishnan, Sathish Babu Shanmugam, Prasanth Selvarajan, John Smith, Gautam Singh, Gao Song, Arvind & Sharanya Suresh, Hisham Talukder, Kun Wang, Andreas Wilm, Chengxi Ye and others for their critical comments, encouragement, and memorable life-events.

No words can express the gratitude I feel toward my beloved parents, wife, sister, and grandmother as the love they have shown, and the sacrifice they have endured over the years for my livelihood, betterment, and happiness, will forever be unmatched.

It is only fitting that I dedicate this thesis to my family, friends and mentors.

M. Senthil Kumar
College Park, MD
2018


that
I know Jesse day
at Bruno Mook!

Table of Contents

List of Tables

List of Figures

Part I

Sequencing technology induced systematic biases

Chapter 1

On the fundamental role of DNA sequencing in modern biology, and its troubling output characteristic.

That phenotypic variance in biological traits is a consequence of underlying genetic changes was suggested concretely in the early 1900s by the work of G. Mendel, T.H. Morgan, R.A. Fisher and others [10–14]. Much of this phenotypic manifestation of genetic information is attributed to the central dogma of molecular biology [15–18], a foundational principle based on three key players: cellular genes in deoxy-ribonucleic acid (DNA) forms are first transcribed to their corresponding ribonucleic-acid (RNA) forms, which are subsequently translated to protein products. Molecular biologists have continued to disentagle the mechanistic basis of the central dogma, and in doing so, have not only specified new roles to existing players, but have also added new players to the story [19–22]. RNA and protein mediated regulation are examples of the former, while the epic epigenetic machinery and their growing list of potential consequences are examples of the latter [23–27]. The players and the interactions among them are, then, well poised to generate variability and stabilize organismal phenotypes over ecological and evolutionary time scales [14, 28–31].

While measurements on phenotypes are more easily obtained, the inverse problem of identifying the underlying genotypic determinants have continued to be challenging to this day [32]. Key experimental techniques and technologies have been developed along the way to help researchers address their questions on the cental dogma, the genes and their interactions efficiently. Knockout experiments are perhaps the most revealing, in making the first steps toward causal characterizations [33–37]. By carefully generating mutant organisms that are deficient in a target gene, and contrasting their behavior against wild-type controls, significant progress can be made in isolating the key functional roles played by the gene. This approach has been very effective in prokaryotes (bacteria and archaea), flies and small animals with shorter generation times, and with phenotypes that are largely determined by a single gene/locus in the genome [38–40]. In larger organisms like humans, although derived cell lines from specific tissues still allow for effective implementation of knockout designs, a more general approach to identifying multi-locus traits can be envisioned if one can access the underlying genomic sequence accurately in its entirety and measure the corresponding phenotypes. If such a procedure can be established, it can be viewed as a natural multi-variate knockout experiment that exploits the observed stochastic genetic variation in extant populations. Such is the utility, offered by the remarkable Nobel-prize winning *DNA sequencing* technology [41–46].

Briefly, accessing genotypes with DNA sequencing works as follows. The input genomic DNA is broken into short random pieces. Each piece is amplified at an average gain, save some technical artifacts, and its nucleotides read off. Most laboratory machines produce a few million such short sequences, each around three hundred base pairs length. However, what we seek is a full description of the genomes present in our input. So one

resorts to algorithmically stiching the output sequences based on their overlaps, as the set of overlaps of a given output sequence signifies its possible local neighborhoods in the input. In this way, one obtains more relevant larger genomic segments like genes or even entire genomes. More recent, expensive, pocketable machines can produce roughly full length bacterial genomes, in the order of of mega base pairs of DNA.

Interestingly, a sequencer's output is not only useful for identifying distinct sequences in the input, but it also allows quantification of their relative frequencies. Regardless of their output statistics and costs, the fundamental input-output characteristic of high throughput DNA sequencing machines remain the same: they produce distinct sequences in abundances proportional to their input relative frequencies, in the sense that if T total short sequences are generated by the machine, an input DNA sequence with a relative frequency q is represented qT times in the output, on average [3, 47–55]. It is this fundamental fact of sequencing based quantification that allows for much mischief from derived technologies that exploit DNA sequencing protocols. We revisit this point on quantification after introducing a few derived technologies next.

Remarkably, the ability to sequence DNA allows creative opportunities to identify not only a cell's genomic sequence, but also get a snapshot of other macromolecules and their behavior within a cell. This is illustrated in **Fig. 1.1**. All the biochemist has to do is encode the entity of interest in a DNA form so that the signal can be read off by the DNA sequencer. For instance, the enzyme reverse transcriptase catalyzes the conversion of RNA to DNA. If we can manage to transform the entire set of expressed RNAs in a cell to their corresponding DNA forms with such an enzymatic reaction, then with the aid of a DNA sequencer, one can expect to sequence and quantify the entire set of expressed

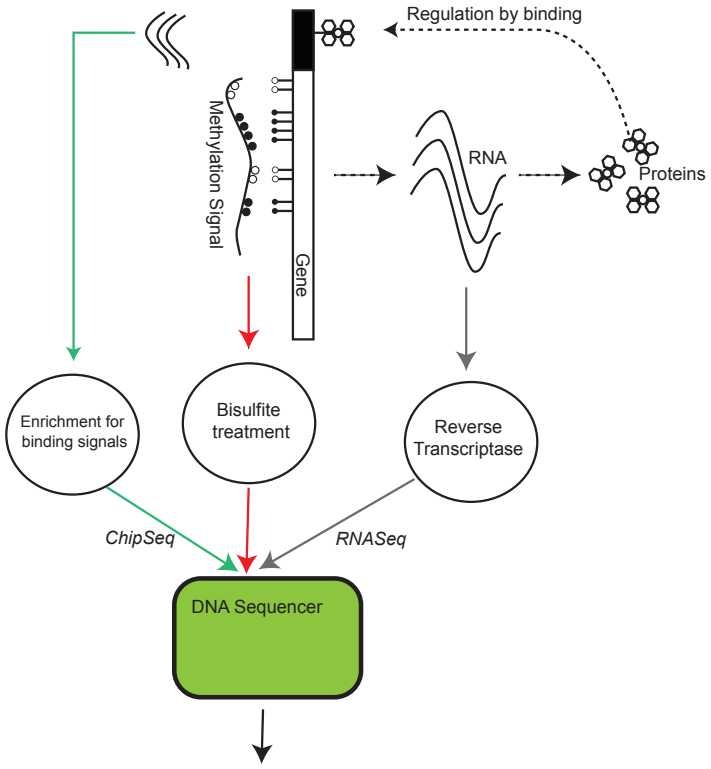

Figure 1.1: Probing the central dogma with a DNA sequencer. *The dashed lines map a gene expression program regulated with a positive feedback motif. A few genomic techniques that allow researchers to investigate the pathway's distinct steps are illustrated. Methylated cytosine residues along the gene body is indicated with tiny filled circles, and unmethylated residues with an open circle.*

RNAs in a single cell, measuring only a couple of microns. This is the idea that drives RNAseq technologies that aim to quantify gene expression, perhaps the most exploited technology built around DNA sequencing [3, 56–62]. Similarly, if one manages to enrich an input sample with only segments in human DNA that are bound by a particular protein molecule, sequencing the enriched sample with a DNA sequencer then yields information on the identity of the protein's binding sites. This is the idea behind the ChipSeq technology [63–67]. In fact, one can go further and aim to identify signals at a single nucleotide

level in the DNA! Millions of cytosine residues in the human DNA are found harboring an extra methyl group. A particular treatment allows transformation of such methylated residues to uracils, while unmethylated cytosines are retained as cytosines, leaving the rest of the DNA material more or less intact [68]. The resulting position specific methylation information is then read off by processing/assembling the output sequences from the DNA sequencer. This is the engineering feat behind the bisulfite sequencing technology [68–70]. Needless to say, other variations of such derived technolgies exist, each with its own target measurements of interest. At the time of this writing, many consortia like *ENCODE* [71, 72], *TCGA* [73–76], *GTex* [77–82], *HMP* [83–86], and *MetaHit* [87–91] with several million dollars in public funds have been established with contributors from all over the world. Their purpose, for the most part, is to exploit sequencing based technologies to produce and analyze associated (genotypic) data for diverse phenotypes of public health interest. The data produced is publicly available for the world's researchers to use. We hope it is clear to the reader that sequencing in various disguises have and will play a fundamental role in helping researchers identify and measure signals of diverse biological origins.

We indicated in the paragraph before last that sequencing technology allows only relative frequency, and not absolute abundance/concentration[1], quantifications of the input molecules[2]. While there are questions in biology where relative frequency measurements are useful (e.g., quantitative geneticists have traditionally tracked allele frequencies in characterizing their long term evolution and fixation in a target population [92, 93], al-

[1]Henceforth, the term *concentration* is used to mean the *absolute abundance* of a molecule in the units of number of molecules per unit volume of the input. We contrast this with *relative frequency/relative abundance*, which is used to mean the fraction molecules of a given type in the input.

[2]More generally, input molecules that are measured in an experiment will be termed *features*.

though one could argue that effective population size measures are still needed to track the fates of low relative frequency mutants), there are at least three fundamental reasons for why feature-wise concentration measurements are attractive and should be sought.

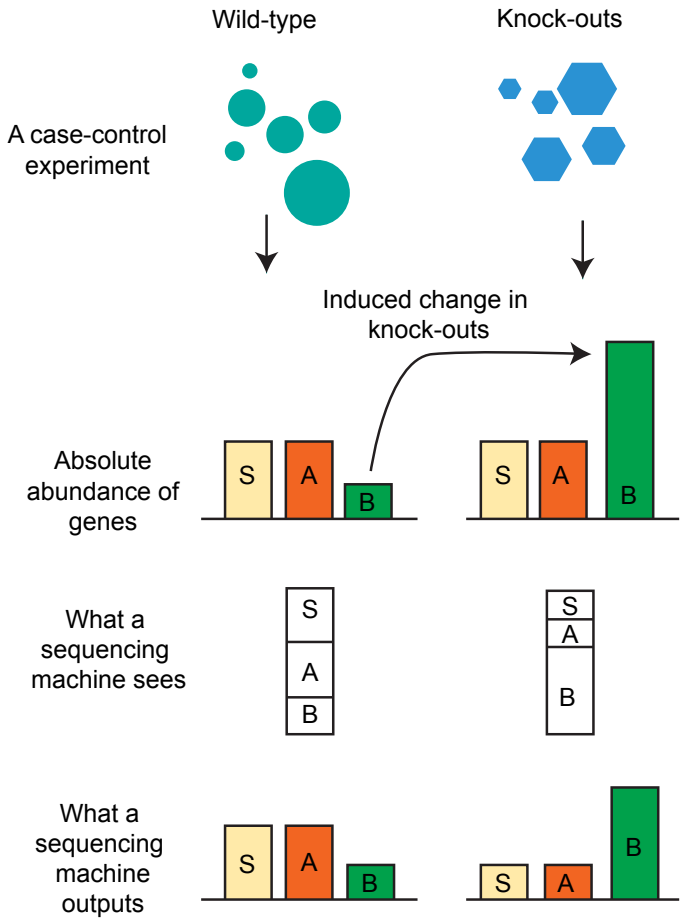

Figure 1.2: Compositional bias: Contrasting relative frequencies lead to confounded concentration inferences. *Suppose we want to compare gene expression measurements from wild-type and knock-out genotypes of a particular cell type. Suppose genes S and A have similar expressed RNA concentrations (number of molecules per unit volume of the cell) in wild-types and knock-out cells, while B has increased in its concentration in the knock-outs due to biological reasons. Because a sequencing machine's output allows relative frequency quantification only, an increase in B leads to reduced abundance measurements from other genes. An analyst might reason A and S to be significantly reduced in abundance, while, in reality they did not.*

First, it must be irrevocably emphasized that concentrations introduce far less ambiguities both in generating sound hypotheses, and in deriving sound scientific conclusions. We shall illustrate this with a few examples. Consider the thought experiment in **Fig. 1.2**, where knock-out genotypes experience an increased concentration of gene B's RNAs alone. If we run the derived RNAseq samples through a sequencer (which only provides relative frequency measurements), we will find that the output relative frequencies correctly indicate that gene B's expression has increased. But the output would also suggest that genes S and A have decreased in their expression. The latter conclusion is false, irrelevant, and is purely caused because of the bias (hereafter, referred to as *compositional bias*) induced by the relative frequency based quantification system. Compositional bias is caused solely because relative frequencies by definition are constrained to sum to 1, and are therefore anti-correlated. Had the experimentalist tracked concentrations, such confounded inferences would not have arisen in the first place. Some well known microbial markers of Crohn's disease based on host intestine associated microbial abundance markers turned out to be artifacts of relative frequency based quantification, and had no immediate relevance to the underlying biology [94]. In an RNASeq experiment contrasting genes' expression values in mice liver and kidney, a decreased expression of house keeping genes were attributable to the increased concentration of a few dominant genes in the liver tissue samples [52]. In fact, in any RNAseq experiment, genes with shorter lengths can appear to be lowly expressed simply because their longer length counterparts contribute more to the sequencing machine's output [50]! In the era of modern biology, where we attempt to base hypotheses and conclusions on millions of molecular features that are quantified using a DNA sequencer, how can we attribute any measured relative

frequency change to an underlying biologically relevant concentration change? In appendices A and C, we outline a couple of our own research programs where such concerns limit serious progress.

Second, key biological phenomena exist in which the absolute concentrations of the players have more meaningful roles, than their relative frequency descriptions. For instance, intracellular gene expression kinetics and their noise characteristics are largely driven by absolute RNA and protein numbers, not their relative frequencies [95–99]. Similarly, cystic fibrosis patients can exhibit very stably associated microbial relative frequencies in the same way as healthy controls, yet suffer from increased absolute total microbial loads [100, 101].

Finally, we must acknowledge that absolute measures are attractive simply because they are more general, allowing immediate access to relative frequency measures if needed. This generality is bound to be favorable in questions involving as yet unexplored biological systems and mechanisms.

Given the major impact sequencing and its derived technologies have in modern biology, sequencing machine induced compositional bias in molecular quantification is certainly a major cause for concern in designing experiments. In chapters I.2, I.3 and I.4, we aim to analyze, and estimate the sequencing machine induced compositional bias under certain assumptions.

Chapter 2

An analysis of sequencing technology induced compositional bias in generating confounded concentration inferences.

In the previous chapter, we mentioned that sequencing technology has been instrumental in measuring diverse biological signals. We also stressed that this remarkable flexibility comes with atleast one tradeoff: output from a DNA sequencer only retain relative frequencies of the input molecular features, and not their absolute concentrations. When contrasting feature-wise relative frequencies across distinct biological sample sources[1], truly null features can exhibit non-zero apparent contrasts. This artifact is shared by all relative frequency quantification systems, and the DNA sequencer is no exception. We illustrated the artifact in **Fig. 1.2**, and gave it the name *compositional bias*.

In this chapter, we will analyze compositional bias in significant generality. In particular, our interest would be in deriving the conditions under which a relative frequency measurement system like the DNA sequencer would yield unbiased concentration inferences. We will also identify the *compositional correction factor*, which when estimated correctly would remedy the compositional bias problem. It is not surprising that this

[1] an analysis known as differential abundance analysis or differential expression analysis in the biology literature.

quantity is a measure of the total feature load in the input sample.

2.1 A sequencing experiment

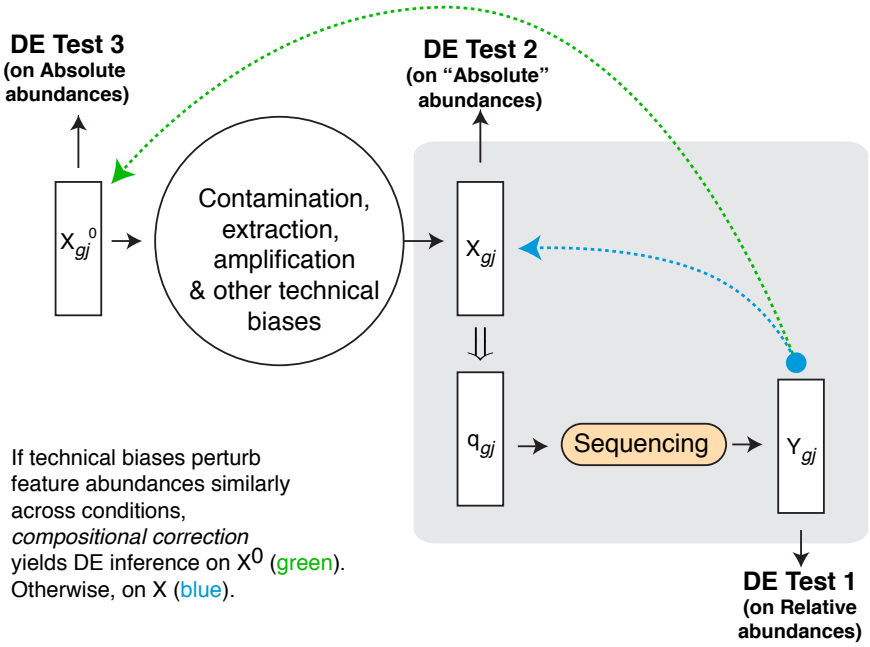

Figure 2.1: Compositional bias introduced by sequencing technology. *As a sample j from group g of interest is prepared for sequencing, its true internal feature concentrations (organized as a vector) X^0_{gj} is transformed by various technical biases to X_{gj}. A sequencing machine introduces compositional bias by generating counts Y_{gj} proportional to the input absolute abundances in X_{gj} according to proportions $q_{gj} = [X_{gji}/(\sum_k X_{gjk})]$, i and k indexing features. Directly performing a differential abundance test on Y (DE Test 1), by using normalization factors (discussed in text) proportional to that of total sequencing output (e.g., R/FPKM/subsampling in metagenomics [3–6]) amounts to testing for changes in* relative *abundances (frequencies) of features in X, in general (*not X^0*). For inferring differences in absolute abundances (concentrations), we need to reconstruct X^0 from Y to perform our inference (DE Test 3). For compositional bias correction in particular, we care about reconstructing X_j from Y (DE Test 2). We show more formally later that compositional correction can reconstruct X^0 if technical biases (including contamination) are comparable across treatment and control groups.*

Fig. 2.1 illustrates a general sequencing experiment and sets up the problem of com-

positional bias correction. We imagine a set of samples/observations $j = 1 \ldots n_g$ arising from biological conditions $g = 1 \ldots G$ (e.g., cases and controls). The true concentrations of features in every input sample is organized as a vector $X^0_{gj\cdot}$, are perturbed by various technical sources of variation as the sample is prepared for sequencing. These technical sources of variation include any unforeseen contaminants, and/or specific biases introduced in a measurement pipeline [51, 60, 61, 102–107]. For instance, when surveying microbial taxa by sequencing 16S ribosomal RNAs, taxonomy specific biases in the relative frequencies can arise by variation in the ribosomal RNA extraction efficiencies [108, 109], binding preferences of DNA amplification agents and even the target ribosomal RNA's Guanine-Cytosine content [107]. All these cause systematic, differential amplification across the surveyed microbial taxa. The end result after all such nuisance perturbations is a transformed concentration vector $X_{gj\cdot}$, the net total concentration of which is denoted by $T_{gj} = \sum_i X_{gji} = X_{gj+}$, where the $+$ indicates summing over that subscript. This is the input to the sequencer, which introduces compositional bias by producing sequencing reads *proportional* to the *absolute* feature abundances represented in $X_{gj\cdot}$. The output short sequences are processed and organized as counts for each feature in a vector $Y_{gj\cdot}$, which now retain only *relative* abundance/relative frequencies of features in $X_{gj\cdot}$ as $\hat{q}_{gji} = Y_{gji}/Y_{gj+} = Y_{gji}/\tau_{gj}$. Here $\tau_{gj} = Y_{gj+}$ is the total number of sequences produced by the machine (*sample depth*) for sample gj.

We discuss the question of recovering $X^0_{gj\cdot}$ for all g and j later in the text. For now, we shall restrict our attention to reconstructing X from Y, as it is in this step, the sequencing machine induces the compositional bias we are interested in. Because we are ignoring all other technical biases inherent to the experiment/technology (i.e., the

process from $X^0 \rightarrow X$), our discussions apply to all derived technologies based on DNA sequencing.

2.1.1 Analysis

Given only the feature-wise relative frequencies output by a sequencer (Y), our goal is to identify the conditions under which we can achieve both (a) unbiased *estimates* of true underlying concentration fold changes (contrasts), and (b) unbiased *inferences* on the estimated concentration contrasts for all features i, in **Fig. 2.1**), when using classical general linear models often exploited in genomics. We briefly summarize the steps in our analysis below.

- Lemma 2.1 provides the condition for obtaining unbiased concentration fold change (or contrast) estimates from relative frequencies. It serves to define the *compositional correction factor*.
- Conditions for achieving unbiased inferences with independent feature-wise general linear models are derived in two steps as follows:
  - Lemma 2.2 uses the idea that for any given feature i in the input, contrasting its frequencies with a linear model (between two experimental conditions, say) would yield accurate concentration inferences for the feature, when the rest of the features do not undergo any concentration change. This fact is reflected as a linear constraint that relates the feature-wise proportions to the compositional correction factor.
  - Theorem 2.3 generalizes Lemma 2.2, and asks when the constraint derived in

Lemma 2.2 would apply to all features in the input. We thus recover conditions to obtain accurate inferences for all feature-wise concentration contrasts with relative frequencies.

- Finally, Therorem 2.4 combines Lemma 2.1 and Theorem 2.3 to recover the conditions for achieving *both* unbiased contrast estimates and their inferences with relative frequencies. Theorem 2.5 generalizes the model dealt with in the aforementioned lemmas and theorems in a straightforward fashion.

Model For simplicity, we shall first consider the generative process in eqn. 2.1, and derive some consequences.

$$\begin{aligned} X_{gj\cdot} &\sim Multinomial(T_{gj}, q_{g\cdot}) \\ Y_{gj\cdot}|X_{gj\cdot}, \tau_{gj} &\sim Multinomial\left(\tau_{gj}, \frac{X_{gj\cdot}}{X_{gj+}}\right) \end{aligned} \tag{2.1}$$

We will note later that the conclusions also hold when the assumption on a fixed proportions vector $q_{g\cdot}$ for all samples at sage X is relaxed by requiring very general moment conditions. The Multinomial assumption on X follows for example from a Poisson assumption on the expression of features X_{gji} [47, 99, 110].

For our analysis, we only consider features truly expressed in the control group ($g = 1$, regardless of them being observed or not in a sequencing experiment) as we can only estimate fold changes for features occurring in the control group, and index them with $i = 1 \ldots p$. Let ϕ_g be the summed proportion of features internally expressed only in group g but not in the control group (regardless of whether they are observed or not). For

interestingness, we assume $p > 1$. Clearly, $0 < q_{1i} < 1$ for all i. Fold changes are defined as ratios of marginal expectations. Define feature-wise concentration fold changes at stage X, $\nu_{gi} = \frac{E[X_{g1i}]}{E[X_{11i}]}$. The corresponding apparent contrasts ξ_{gi} from relative abundances at stage Y is defined as: $\xi_{gi} = \frac{E[\widehat{q_{g+i}}]}{E[\widehat{q_{1+i}}]}$. Denoting $E[T_{g1}]$ as the marginal average of the total abundances T_{gj}, from model 2.1, we have $E[X_{gji}] = E[T_{g1}] \cdot q_{gi}$ for all $j = 1 \dots n_g$. Under model 2.1, the fold changes can be re-written as: $\nu_{gi} = E[X_{g1i}]/E[X_{11i}]$, and $\xi_{gi} = q_{gi}/q_{1i}$.

In the entire process, we only get to observe $Y_{gj\cdot}$ for all $j = 1 \dots n$ and $g = 1 \dots G$.

Lemma 2.1. *Under assumptions 2.1, for all features i,* $\nu_{gi} = \xi_{gi}$*, if and only if* $\Lambda_g = \frac{E[T_{g1}]}{E[T_{11}]} = 1$. Λ_g^{-1} *will be termed as the compositional correction factor.*

Proof. The proof follows directly from the definition of fold changes ν_{gi} associated with the i^{th} feature's concentrations.

$$\nu_{gi} = \frac{E[X_{g1i}]}{E[X_{11i}]} = \frac{E[T_{g1}]q_{gi}}{E[T_{11}]q_{1i}} \equiv \Lambda_g \cdot \frac{q_{gi}}{q_{i1}} = \Lambda_g \frac{E[\widehat{q_{g+i}}]}{E[\widehat{q_{1+i}}]} = \Lambda_g \xi_{gi} \tag{2.2}$$

which is equal to ν_{gi} iff $\Lambda_g = 1$. □

Lemma 2.2. *Under assumptions 2.1, when applying the standard log-linear mean model on the total sum normalized data independently for each feature i,* $\log E[Y_{gji}/\tau_{gj}] = \mu_i + \alpha_{gi}$ *with* μ_i *quantifying logged control group proportions,* $\log q_{1i}$*, and* α_{gi} *quantifying the log-fold change of relative abundances,* $\log \xi_{gi}$*, there is a necessary and sufficient condition under which* $\alpha_{gi} = 0 \iff \log \nu_{gi} = 0$*, the log- fold change associated with*

concentrations. Furthermore, this condition is given as:

$$\frac{1}{1-q_{1i}}\left[\Lambda_g\phi_g+\sum_{k,k\neq i}\nu_{gk}q_{1k}\right]=1$$

Proof. Following Lemma 2.1, re-write the proportion in group g as:

$$q_{gi}=\Lambda_g^{-1}\nu_{gi}q_{1i}=\frac{\Lambda_g^{-1}\nu_{gi}q_{1i}}{1}=\frac{\Lambda_g^{-1}\nu_{gi}q_{1i}}{\phi_g+\sum_k q_{gk}}=\frac{\nu_{gi}q_{1i}}{\Lambda_g\phi_g+\nu_{gi}q_{1i}+\sum_{k,k\neq i}\nu_{gk}q_{1k}}$$
$$\equiv\frac{1}{1+\frac{\nu_{g\setminus i}}{\nu_{gi}}\frac{(1-q_{1i})}{q_{1i}}} \tag{2.3}$$

where we have set:

$$\nu_{g\setminus i}=\frac{1}{1-q_{1i}}\left[\Lambda_g\phi_g+\sum_{k,k\neq i}\nu_{gk}q_{1k}\right] \tag{2.4}$$

Substituting eqn. 2.3 in the assumed linear model: $\log E[Y_{gji}|\tau_{gj}]=\log q_{gji}+\log\tau_{gj}=\mu_i+\alpha_{gi}+\log\tau_{gj}$, and noting $\mu_i=\log q_{1i}$, $\alpha_{gi}=\log\frac{q_{gi}}{q_{1i}}$, it is clear that $\alpha_{gi}=0\iff\frac{\nu_{g\setminus i}}{\nu_{gi}}=1$. It is thus seen that $\nu_{g\setminus i}=1$, is a necessary and sufficient condition for the statement $\alpha_{gi}=0\iff\nu_{gi}=1$ to hold. □

Theorem 2.3. *Under the model above, there exists a unique vector of fold changes $\nu^*_{g\cdot}$ under which $\forall i=1\ldots p,\ \alpha_{gi}=0\iff\nu_{gi}=1$. Furthermore, each $i=1\ldots p$ entry of $\nu^*_{g\cdot}$ is given as:*

$$\nu^*_{gi}(\Lambda_g,\phi_g,q_{1\cdot})=\left[\left(\frac{1-q_{1i}}{q_{1i}}\right)\left(\frac{\eta_g}{1-q_{1i}}-1\right)+\left(\frac{1}{1-p}\right)\sum_k\left(\frac{1-q_{1k}}{q_{1k}}\right)\left(\frac{\eta_g}{1-q_{1i}}-1\right)\right] \tag{2.5}$$

with $\eta_g = \Lambda_g \cdot \phi_g$.

Proof. We want to study the conditions under which $\nu_{gi} = 1 \forall i = 1 \dots p$. Substituting this in equation 2.4 from lemma 2.2, and stacking the constraints for all i, we get a linear system:

$$Q\nu = \gamma$$

where, Q is a $p \times p$ matrix with $Q(i,j) = \frac{q_{1j}}{1-q_{1i}}$ if $j \neq i$ and 0 otherwise. $\nu = [|\nu_{gi}|]_{i=1}^{p}$, a $p \times 1$ vector, and $\gamma = [|\gamma_{gi}|]_{i=1}^{p}$, a $p \times 1$ column vector with $\gamma_{gi} = 1 - \frac{\eta_g}{1-q_{1i}}$, where $\eta_g = \Lambda_g \phi_g$, a non-dimensional parameter. A unique solution for this equation is obtained directly as $\nu^* = [|\nu^*_{gi}|]_{i=1}^{p} = Q^{-1}\gamma$ if Q is invertible.

We now show that Q is invertible by observing that the column vectors of the $p \times p$ square matrix Q are linearly independent. If we denote the columns from left to right as $K_1, \dots K_p$, for linear dependence, we want the statement $\sum_{j=1}^{p} \alpha_j K_j = 0 \implies \alpha_j = 0$ to hold. Identifying each of the column's projections on all p dimensional unit vectors e_j, and noting all $q_{1i} \in (0,1)$, we can write:

$$\begin{aligned} K_j &= q_{1j}\left\{\frac{1}{1-q_{11}}e_1 + \dots + \frac{1}{1-q_{1,j-1}}e_{j-1} + \frac{1}{1-q_{1,j+1}}e_{j+1} + \dots + \frac{1}{1-q_{1,p}}e_p\right\} \\ &= q_{1j}\left\{\sum_{i=1}^{p}\frac{1}{1-q_{1i}}e_i - \frac{1}{1-q_{1j}}e_j\right\} \end{aligned}$$

Generating the required linear combination of these column vectors K_i, we find:

$$\sum_{j=1}^{p} \alpha_j K_j = 0 \implies \sum_{j=1}^{p} \alpha_j q_{1j} \sum_{i=1}^{p} \frac{1}{1-q_{1i}} e_i = \sum_{j=1}^{p} \frac{1}{1-q_{1j}} \alpha_j q_{1j} e_j$$

$$\implies \forall\, i,\ \frac{1}{1-q_{1i}} \sum_{j=1}^{p} \alpha_j q_{1j} = \alpha_i q_{1i} \frac{1}{1-q_{1i}}$$

$$\implies \forall\, i,\ \sum_{j=1}^{p} \alpha_j q_{1j} = \alpha_i q_{1i}$$

Summing the last equation over all $i = 1 \ldots p$, we get:

$$2 \sum_{i=1}^{p} \alpha_i q_{1i} = \sum_{i=1}^{p} \alpha_i q_{1i}$$

Because all $q_{1i} \in (0,1)$, the above equation can only be true if $\alpha_i = 0\ \forall i$. Hence, the vectors are lineary independent; Q is full rank, and invertible.

Indeed, we can go further and derive the solution analytically. Notice that $Q = rq_{1\cdot}^T - D$ where r is a $p \times 1$ vector with the i^{th} component equal to $\frac{1}{1-q_{1i}}$ and $q_{1\cdot}$ is a $p \times 1$ vector of control proportions. Notice all $0 < q_{1i} < 1$. D is a $p \times p$ diagonal matrix with diagonal entries given by $\frac{1-q_{1j}}{q_{1j}}\ \forall j = 1 \ldots p$. If we set $F = D - rq_{1\cdot}^T$, we then want $Q^{-1} = -F^{-1}$. Denoting $U = -r$, and $V = q_{1\cdot}^T$, we can write, $F^{-1} = (D + UV)^{-1}$. Woodbury identity then yields $F^{-1} = D^{-1} - D^{-1}U(I + VD^{-1}U)^{-1}VD^{-1}$, a $p \times p$ matrix with $F^{-1}(i,j) = \frac{1-q_{1j}}{q_{1i}} \left(\frac{1}{1-p}\right)$ if $i \neq j$, and $\frac{1-q_{1i}}{q_{1i}} \left(1 + \frac{1}{1-p}\right)$ if $i = j$. The exact solution for the fold changes satisfying the linear constrains are then given by $v_{g\cdot}^* = -F^{-1}\gamma$, with v_{gi}^* given by eqn. 2.5 above. □

Theorem 2.4 (Validity of Total Sum Normalization in Reconstructing X)**.** *Under the model above, the vector of feature-wise fold changes under which relative frequencies*

(total sum normalized data) can yield unbiased inferences (correct fold changes and non-zero significance) of concentrations of all $i = 1 \ldots p$ features in group g at stage X is given by $\nu_g^*(1, \phi_g, q_1)$, where $\nu_g^*(\Lambda_g, \phi_g, q_1)$ is defined in Theorem 2.3.

Proof. Proof follows directly from Lemma 2.1 and Theorem 2.3. □

Theorem 2.5 (Relaxing the fixed group-specific proportions assumption)**.** *The results derived in lemmas 2.1, 2.2, and theorems 2.3 and 2.4 hold under the following more general model as well. In this model, q_{gi} and ϕ_g are defined as marginal expectations of sample-wise relative frequencies, which are themselves assumed to be independent of T_{gj}.*

$$\begin{aligned}
&\{\tilde{q}_{gj\cdot}, \tilde{\phi}_{gj}\} \sim f(\cdot) \text{ such that } \tilde{\phi}_{gj} \in (0,1),\ \tilde{q}_{gji} \in (0,1),\ \tilde{\phi}_{gj} + \sum_i \tilde{q}_{gji} = 1, \\
&\text{with } E[\tilde{q}_{qji}] := q_{gi},\ E[\tilde{\phi}_{gj}] := \phi_g,\ T_{gj} \text{ independent of } \tilde{q}_{gji},\ \tilde{\phi}_{gj}. \\
&X_{gj\cdot} | T_{gj}, \tilde{q}_{gji} \sim Multinomial(T_{gj}, \tilde{q}_{gj\cdot}) \\
&Y_{gj\cdot} | X_{gj\cdot}, \tau_{gj} \sim Multinomial\left(\tau_{gj}, \frac{X_{gj\cdot}}{X_{gj+}}\right)
\end{aligned} \tag{2.6}$$

Here f is some unspecified distribution function (e.g., Dirichlet) that allows constrained sampling of observation-specific relative frequencies such that they sum to 1*, with finite feature-wise marginal expectations.*

Proof. One only needs to note that with $E[\tilde{q}_{gji}] = q_{gi}$, and $E[\tilde{\phi}_{gj}] = \phi_g$ forall $j = 1 \ldots n_g$

samples in group g, we obtain:

$$E[\widehat{\hat{q}_{g+i}}] = q_{gi}. \tag{2.7}$$

$$E[X_{g1i}] = E[E[X_{gji}|T_{gj}, \tilde{q}_{gji}]] = E[T_{gj}\ \tilde{q}_{gji}] = E[T_{gj}]q_{gi}. \tag{2.8}$$

$$\phi_g + \sum_{i=1}^{p} q_{gi} = E[\tilde{\phi}_{g1}] + \sum_{i=1}^{p} E[\tilde{q}_{g1i}] = E[\tilde{\phi}_{g1} + \sum_{i=1}^{p} \tilde{q}_{g1i}] = 1. \tag{2.9}$$

Equations 2.7 and 2.8 above are needed for lemma 2.1, and equation 2.9 is needed for the lemma 2.2 and theorem 2.4 to go through. □

The result in theorem 2.4 was also verified numerically. As an example, suppose $q_{1\cdot} = [0.25, 0.25, 0.1, 0.1, 0.3]^T$. For $\Lambda_g = 1$, and $\phi_g = 0.05$, the fold changes that need to be achieved for unbiased inference is given by: $v^*_{g\cdot} = [0.95, 0.95, 0.88, 0.88, 0.96]^T$ implying that downregulation across features can be detected well as the unique features will compete for sequencing output. For $\Lambda_g = 1$, and $\phi_g = 0.4$, no feasible solution exists. For the case $\phi_g = 0$, the optimal solution is trivial: $v_{gi} = 1$ for all i i.e., no perturbation in any of the features. Providing additional constraints by fixing at least one of the fold changes yields the single, constrained solution on the rest of the fold changes: the solution vector v^* is obtained by replacing $\eta_g = \Lambda_g \phi_g$ in the above equation with $\eta_g = \sum_{k \in F} v^*_{gk} q_{kg}$ where F is the set of features for which the fold changes are fixed apriori to v^*_{gk}, and restricting i to the rest of the features (other than those in F) present in the control group in the above derivation. Notice that there is an uncountable number of values (non-negative real values) the fold changes of features in the constraint set F can take. They will impose a particular value of η_g, and conditioned on this value, the fold changes the rest of the features can take in group g so that the linear model above achieves unbiased contrast

estimates and inferences are unique.

The conclusion of theorem 2.4/its generalized version in 2.5 is an unfortunate result as it says that to obtain unbiased concentration inference across all features with relative frequencies alone, the feature-wise concentration fold changes must behave in a unique fashion, and therefore appears unlikely to occur in practice. Notice that fold-change ν^*_{gi} can never be < 0. Thus, a feasible solution need not exist for arbitrary parameter values of $\eta_g = \Lambda_g \phi_g$ implying that unbiased inference may not always be possible. It is also interesting to note that unless the fold change of total feature content in group g (Λ_g) is somehow maintained the same across conditions despite contaminants present at proportion ϕ_g, achieving unbiased inference with normalization techniques based on the total sum is not possible. Group-specific expression of features are a major source of compositional bias and their sufficiently high expression can effectively wash out the signal. In metagenomic surveys, it is often the case that a large number of features are observed with a positive count in very few samples. Although this does not necessarily mean they are actually present in only a few observations, we can expect this to be the case with samples arising from diverse ecosystems.

In summary, strict unbiased inference with a DNA sequencer's output relative frequencies may or may not be possible depending on the underlying value of η_g; when possible, it can only occur under a unique set of fold changes. In practice, RNAseq experiments are performed across diverse tissues of various origins, and metagenomic surveys are constantly carried out across ecosystems. Thus, unbiased-ness in inference need not hold.

2.2 When can we hope to reconstruct X^0 from Y with compositional bias correction?

So far, we have concerned ourselves only with characterizing the conditions that enable accurate characterization of feature-wise concentrations in the sequencing input (X in Fig. 2.1), given only feature-wise frequencies. We can now ask when compositional correction is guaranteed to recover the true concentrations X^0 in the original source before it was marred by technical variation. In the next chapter, simple algebraic derivations reveal that as long as a feature i's unwanted technical variation in group g is comparable, on average, to that of the control group, compositional bias correction reconstructs its true concentration, i.e., $E[X^0_{g1i}]$ (eqn. 3.1).

With this result, we recover a slightly more general condition than the often cited assumption on some familiar genomic data normalization techniques [4, 6]. We not only want the technical biases to affect all the features the same way within a sample, but if any contamination is introduced we want those biases to also behave appropriately according to the above condition. We also emphasize that in-silico post-processing of sequencing count data for contaminants (for example, by excluding sequencing reads mapping to potential cotaminant reference sequences) does not help in compositional bias correction because they have already caused information loss by competing with other native features for being sequenced.

Chapter 3

On the generality of compositional correction factors, and some strategies to estimate them.

In the previous chapter, we addressed the question of inferring feature-wise concentrations with feature-wise relative frequencies output from a DNA sequencing machine. We found that in such a problem, a single, linear bias term denoted as Λ^{-1}, called the *compositional correction factor*, underlies all the confounded feature-wise concentration inferences. Among others, we arrived at two important conclusions:

1. That by transforming concentrations to frequencies, sequencing machine introduces one of the many unforeseen technical biases in our truly intended experiment to measure concentrations.

2. That by appropriately measuring or estimating the compositional correction factor, one can arrive at accurate feature-wise concentration inferences.

3.1 The generality of compositional correction factors in explaining technical variation

We now note that compositional factors are far more general in their utility than merely serving to describe compositional bias induced by the sequncer. Indeed, they can very well account for other sources of technical variation as well. To see this, we refer the reader back to Fig. 2.1 and its notation, and notice that the process $X^0_{gj\cdot} \to X_{gj\cdot}$ succinctly accounts for all unwanted, technical perturbations in the concentration of any feature i, in sample gj. For each feature i, on average, these perturbations are described by the corresponding fold changes defined relative to the true concentrations in the original sample source: $\mu_{gi} = \frac{E[X_{g1i}]}{E[X^0_{g1i}]}$.

We already saw that the concentration fold change at the time of input to sequencer (stage X in Fig. 2.1) is given as: $\nu_{gi} = \frac{E[X_{g1i}]}{E[X_{11i}]}$. With some algebra below, we observe how this apparent fold change is correlated with all technical perturbations abstracted away in the process $X^0_{gj\cdot} \to X_{gj\cdot}$:

$$\begin{aligned}
\nu_{gi} &= \frac{E[X_{g1i}]}{E[X^0_{g1i}]} \cdot \frac{E[X^0_{g1i}]}{E[X_{11i}]} \\
&= \mu_{gi} \cdot \frac{E[X^0_{g1i}]}{\mu_{1i} \cdot E[X^0_{11i}]} \\
&= \frac{\mu_{gi}}{\mu_{1i}} \cdot \frac{E[X^0_{g1i}]}{E[X^0_{11i}]} \\
&= \frac{\mu_{gi}}{\mu_{1i}} \times \nu^0_{gi} \quad =: \underbrace{\zeta_{gi}}_{\text{true technical fold change}} \times \underbrace{\nu^0_{gi}}_{\text{true biological fold change.}},
\end{aligned} \tag{3.1}$$

It is correctly observed above that if technical biases are comparable across cases

and control conditions, the first factor is one, and the apparent fold change measured at stage X equals the true biological fold change at stage X^0. Our formula for compositional correction factors is then altered correspondingly as $\Lambda_g = \nu_{g\cdot}^T q_{g\cdot} = (\zeta_{g\cdot} \circ \nu_{g\cdot})^T q_{g\cdot}$. Here $\circ$ denotes element-wise product.

Thus we see that compositional correction factors can account for more general technical biases introduced in a sequencing experiment. Given this significance, it is only fitting that we consider their estimation in detail.

3.2 Estimation strategies

Strategy 1: Measure total feature load We had indicated in the previous chapter that the compositional correction factor for each experimental group g, Λ_g^{-1}, is inversely related to the group's average total feature content. So a clear strategy is to measure, if possible, the total DNA in the input sample could serve to estimate Λ_g^{-1}; subsequently multiplying the estimates to the DNA sequencer's output relative frequencies should then restore feature-wise concentrations. However, this strategy only reconstructs concentrations at the time of input to the sequencer (the vector $X_{gj\cdot}$ in Fig. 2.1), and completely ignores all other technical variation introduced in the data. So, unless techinical biases are comparable across conditions (as described in the previous subsection), it will be limited in its practical utility.

Here is a more concrete approach. Suppose we know that some feature k is unchanged in its concentration across conditions. From eqn. 2.2, we see that for any feature i, $q_{gi} = \Lambda_g^{-1} \nu_{gi} q_{1i}$. Because the fold change for the unperturbed feature (at stage X)

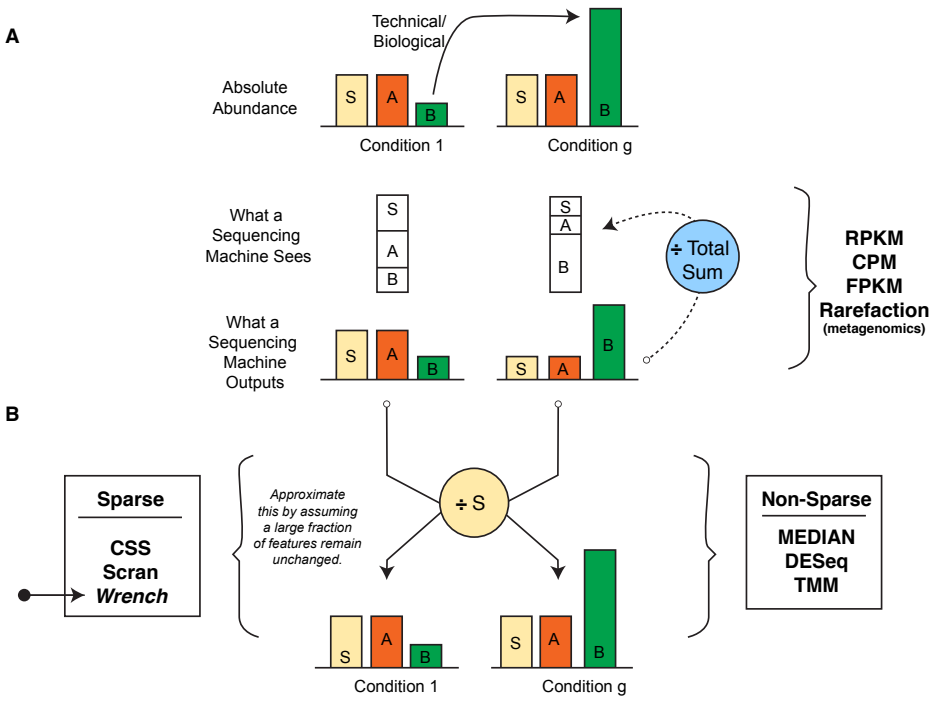

Figure 3.1: Scaling normalization techniques in genomics from the perspective of compositional bias correction. *(**A**) Features S and A have similar absolute abundances in two experimental conditions, while B has increased in its absolute abundance in condition g due to technical/biological reasons. Because of the proportional nature of sequencing, increase in B leads to reduced read generation from others (compositional bias). An analyst would reason A and S to be significantly reduced in abundance, while, in reality they did not. (**B**) Knowing S is expressed at the same concentration in both conditions allows us to scale by its abundance, resolving the problem. DESeq and TMM, by exploiting rerefence strategies across feature count data (described below), approximate such a procedure, while techniques that are based only on library size alone like RPKM and rarefication/subsampling can lead to unbiased inference only under very restrictive conditions. Currently available approaches for sparse data settings are indicated.* Wrench *is the proposed technique in the next chapter.*

$\nu_{gk} = 1$ for all groups $g = 1 \dots G$, we obtain $q_{gk} = \Lambda_g^{-1} q_{1k}$, which immediately suggests Λ_g^{-1} can be computed as ratios of proportions of this internal control feature. Furthermore, if we calculate the transformation $\log \frac{E[Y_{gji}]}{E[Y_{gjk}]} = \log \frac{q_{gi}}{q_{gk}} = \log \frac{\Lambda_g^{-1} \nu_{gi} q_{1i}}{\Lambda_g^{-1} q_{1k}} = \mu_i + \alpha_{gi}$, where with appropriate side conditions on the contrasts, the intercept estimates $\mu_i =$

$(\log q_{1i} - \log q_{1k})$. Our contrast variable then estimates $\alpha_{gi} = \log \nu_{gi}$, which is 0 only under the null $\nu_{gi} = 1$. Thus, the traditional data normalization idea of "dividing by a feature that does not change across conditions" automatically corrects for compositional bias induced through sequencing technology [4]. This is discussed further below. Notice that we do not necessarily need the internal control feature to have the same internal concentration across conditions. As long as we know their sample-wise absolute concentrations, their fold changes across conditions are also known, and these simply enter the above formulation as known constants that simply offset the linear models. (That is, we can write: $q_{gk} = \Lambda_g^{-1} \hat{\nu}_{gk} q_{1k}$, where $\hat{\nu}_{gk}$ is the now known fold change associated with the feature in group g.) These insights lead us to the following two estimation strategies:

Strategy 2: Introduce spike-in control features If all we need is a feature that is expressed at known abundances across conditions, why not inject it ourselves at the time of sequencing? Two potential techniques exist in the experimental literaure, one of which cannot protect us against compositional bias. In the ERCC spike-in protocol [111], widely used in various bulk tissue and some single cell RNAseq studies [112], a fixed amount of total RNA extract is obtained, and subsequently suspended in solution along with known concentrations of a chosen control feature (spike-ins). Because this procedure adds the spike-ins to the extract, an already compositional source, our inferences are limited to questions on relative abundances; a statement about differences in absolute abundances cannot be made unless the samples themselves behave according to the narrow conditions established in the previous chapter. An alternative, more effective strategy is to add known concentrations of barcodes/spike-in to the entire sample's suspension [113] (**Fig. 3.1B**).

This problem has also been noted by Stegle et al., in the context of designing scRNAseq experiments [114].

Strategy 3: Post-process abundance data with reference normalization strategies In the absence of internal control features like the spike-ins, effective correction for compositional bias can still be hoped for [52]. Here, is the central idea, which is so significant that it will appear repeatedly in our discussions: If most features do not change in their absolute abundances relative to the control group, the fundamental eqn. 2.2 should hold true for most features with $\nu_{gi} = 1$. Thus, an appropriate summary statistic of these ratios of proportions could serve as an estimate of Λ_g^{-1}.

With this idea in place, a normalization procedure for deriving sample-specific compositional scale factors Λ_{gj}^{-1} can be devised. One only needs to carry out the above procedure by pretending that every sample arises from its own experimental group. Indeed, as illustrated in **Table 3.1**, scale normalization methods in genomics can be viewed in this light, where some control set of proportions ("reference") is defined, and the Λ_{gj}^{-1} estimate is derived for every sample j based on the ratio of its proportions to that of the reference. This central idea being the same, the robustness of these methods are dependent on how well the assumptions hold with respect to the chosen reference, and the choice of the estimation algorithm.

To illustrate this idea further, we present the following derivation of a *DESeq-like* normalization strategy (refer Table 3.1 and Fig. 3.1). We use the same notation as in the last chapter and Table 3.1. Because each sample is considered to arise from its own group, the index j does not play any role here. We can fix $j = 1$, and let $g = 1 \ldots n$ index the

samples. Let $i = 1 \dots p$ index the features. For a given sample gj then, Y_{gji} indicates the measured sequencing count of the i^{th} feature in the sample; ν_{gji}, the feature's true concentration fold change in the sample; $\tau_{gj} = Y_{gj+}$ the sample's sequencing depth; q_{gji} the feature's proportion in the sample. Finally, let q_{1ji} denote the proportion of feature i in a control sample indexed with $g = 1$.

If one assumes that feature-wise count distributions follow a log-normal distribution, we obtain a DESeq-like estimator for compositional correction factors as below. Alter eqn. 2.2 with a multiplicative log-normal error term, and write for feature i in sample gj,

$$\begin{aligned} Y_{gji} &\sim \Lambda_{gj}^{-1} \cdot \tau_{gj} \cdot \nu_{gji} q_{1ji} \cdot LN(0, \sigma_i^2), \;\; i = 1 \dots p, \;\; j = 1 \dots n \\ &\equiv \mu_{gji} \cdot LN(0, \sigma_i^2) \end{aligned}$$

where, $LN(0, \sigma_i^2)$ refers to a log-normal random variate that when logged has a mean of 0 and a variance of σ_i^2. Then:

$$\begin{aligned} \prod_g Y_{gji} &\sim \left(\prod_g \mu_{gji} \right) LN(0, n\sigma_i^2) \\ \left[\prod_g Y_{gji} \right]^{\frac{1}{n}} &\sim \left(\prod_g \mu_{gji} \right)^{1/n} LN(0, \sigma_i^2) \\ \implies d_{gji} &= \frac{Y_{gji}}{\left(\prod_g Y_{gji} \right)^{\frac{1}{n}}} \sim \frac{\mu_{gji}}{\left(\prod_g \mu_{gji} \right)^{\frac{1}{n}}} \cdot LN(0, \sigma_i^2) \\ &= \frac{\Lambda_{gj}^{-1}}{\left(\prod_g \Lambda_{gj}^{-1} \right)^{\frac{1}{n}}} \cdot \frac{\tau_{gj}}{\left(\prod_g \tau_{gj} \right)^{\frac{1}{n}}} \cdot \frac{\nu_{gji}}{\left(\prod_g \nu_{gji} \right)^{\frac{1}{n}}} \cdot LN(0, \sigma_i^2) \\ &= k_\Lambda \Lambda_{gj}^{-1} \cdot k_\tau \tau_{gj} \cdot k_\nu \nu_{gji} \cdot LN(0, \sigma_i^2) \end{aligned}$$

in which we have collected the constant denominator (independent of gj) terms separately into three k terms with corresponding subscripts. Now, $\tilde{d}_{gji} = \frac{d_{gji}}{k_\tau \tau_{gj}} \sim k_\Lambda \Lambda_{gj}^{-1} \cdot k_\nu \nu_{gji} \cdot LN(0, \sigma_i^2)$, with expectation given by $k_\Lambda \Lambda_{gj}^{-1} \cdot k_\nu \nu_{gji} \cdot e^{\sigma_i^2/2} \propto \Lambda_{gj}^{-1} \cdot \nu_{gji} \cdot e^{\sigma_i^2/2}$. So if atleast a median fraction of features do not change on average relative to the reference sample, setting $\nu_{gji} = 1$ should hold for those features. We then arrive at:

$$\tilde{s}_{gj} = median_i \ \frac{\tilde{d}_{gji}}{e^{\sigma_i^2/2}} \propto \Lambda_{gj}^{-1} \tag{3.2}$$

and so $\tilde{s}_{gj}$ serves as an estimator of Λ_{gj}^{-1}. This is simply DESeq normalization factors presented in table 3.1, altered only by feature-wise variances.

In summary, the fact that compositional factors are linear technical biases shared by all measured features, makes it possible to take advantage of the class of scale normalization techniques in the genomics literature to estimate them [4, 52, 115, 116]. All of these approximate the aforementioned spike-in strategy by assuming that most features do not change on average across samples/conditions (Fig. 3.1). For the same reason, we have given such an interpretation to approaches like centered logarithmic transforms (CLR) from the theory of compositional data, which many analysts favor when working with relative abundances [117–123]. We must note that scaling normalization techniques have the same limitation as strategy 1 described above.

Reconstructing X^0 from Y It is worth emphasizing again that the aforementioned reference normalization strategies do not restrict compositional factors to only reflect biology-induced global abundance changes; in reality, if feature-wise perturbations (ν_{gi})

Technique	Proposed Abundance Measure, Scale factor	Signal for Compositional Scale in
Total Sum	$\frac{y_{gji}}{\tau_{gj}\cdot\Lambda_{gj}^{-1}}$, $\Lambda_{gj}^{-1}=1$	
TMM	$\frac{y_{gji}}{\tau_{gj}\cdot\Lambda_{gj}^{-1}}$, $\Lambda_{gj}^{-1}=e^{\left[\sum_{i:y_{ij}>0\ \cap\ i\in\text{trimmed set for j}} w_{ij}\log\left(\frac{q_{gji}}{q_{1ji}}\right)\right]}$	$\frac{q_{gji}}{q_{1ji}}$, ratio of proportions
DESeq	$\frac{y_{gji}}{C\cdot\tau_{gj}\cdot\Lambda_{gj}^{-1}}\propto\frac{y_{gji}}{\tau_{gj}\cdot\Lambda_{gj}^{-1}}$, $\Lambda_{gj}^{-1}=\text{median}_i\ \frac{q_{gji}}{[\prod_k q_{ik}]^{\frac{1}{n}}}$	$\frac{q_{gji}}{[\prod_k q_{ik}]^{\frac{1}{n}}}$, ratio of proportions
Median	$\frac{y_{gji}}{\tau_{gj}\cdot\Lambda_{gj}^{-1}}$, $\Lambda_{gj}^{-1}=\text{median}_i\ q_{gji}\propto\text{median}_i\ \frac{q_{gji}}{1/p}$	$\frac{q_{gji}}{1/p}$, ratio of proportions
Upper quartile	$\frac{y_{gji}}{\tau_{gj}\cdot\Lambda_{gj}^{-1}}$, $\Lambda_{gj}^{-1}=\text{upper quartile}_i\ q_{gji}\propto\text{upper quartile}_i\ \frac{q_{gji}}{1/p}$	$\frac{q_{gji}}{1/p}$, ratio of proportions
CLR Transformation	$\log\left(\frac{y_{gji}}{[\prod_i y_{gji}]^{\frac{1}{p}}}\right)\equiv\log\left(\frac{q_{gji}}{[\prod_i q_{gji}]^{\frac{1}{p}}}\right)\equiv\log\left(\frac{y_{gji}}{\tau_{gj}\cdot\Lambda_{gj}^{-1}}\right)$, with $\Lambda_{gj}^{-1}=\left[\prod_i q_{gji}\right]^{\frac{1}{p}}\propto\left[\prod_i\frac{q_{gji}}{1/p}\right]^{\frac{1}{p}}$	$\frac{q_{gji}}{1/p}$, closely tracks Median factors above; ratio of proportions
Scran	$\frac{y_{gji}}{\tau_{gj}\cdot\Lambda_{gj}^{-1}}$, $\Lambda_{gj}^{-1}=\text{fit linear models to }\left\{\frac{q_{1ji}}{\overline{q_{++i}}},\ldots,\frac{q_{nji}}{\overline{q_{++i}}}\right\}_{i=1}^{p}$	$\frac{q_{gji}}{\overline{q_{++i}}}$, ratio of proportions
Wrench	$\frac{y_{gji}}{\tau_{gj}\cdot\Lambda_{gj}^{-1}}$, $\Lambda_{gj}^{-1}=\frac{1}{p}\sum_i w_{ij}\frac{q_{gji}}{\overline{q_{++i}}}$	$\frac{q_{gji}}{\overline{q_{++i}}}$, ratio of proportions

Table 3.1: Scaling normalization approaches derive their technical bias estimates from ratio of proportions. *For each scaling normalization technique, we present the transformation they apply to the raw sequencing count data (second column) to produce normalized counts. The third column shows how all techniques use statistics based on ratio of proportions to derive their scale factors. $i=1\ldots p$ indexes features, each sample is considered to arise from its own singleton group: $g=1\ldots n$ and $j=1$, τ_{gj} the sample depth of sample j, q_{gji} the proportion of feature i in sample j, w_{ij} represents a weight specific to each technique, and $\overline{q_{++i}}$ is the average proportion of feature i across the dataset. In the second column, the first row in each cell represents the transformation applied on the raw count data by the respective normalization approach. They all adjust a sample's counts based on sample depth (τ_{gj}) and a compositional scale factor Λ_{gj}^{-1}. Continued on the following page.*

Table 3.1: *Continued from previous page. As noted in the third column, the estimation of Λ_{gj}^{-1} is based on the ratio of sample-wise relative abundances/proportions (q_{gji}) to a reference that are all some robust measures of central tendency in the count data. The logarithmic transform accompanying CLR should not worry the reader about its relevance here, in the following sense: the log-transformation often makes it possible to apply statistical tests based on normal distributions for the rescaled data; this is in-line with applying log-normal assumptions on the rescaled data obtained with the rest of the techniques. $C = \left[\prod_j \tau_{gj}\right]^{-1/n}$ is a constant factor independent of sample, and its presence does not matter. For the same reason, Median and Upper Quartile scalings and CLR transforms, can be thought to base their estimates on a reference that assigns equal mass to all the features or if the reader wishes, a more complicated reference that behaves proportionally. When most features are zero, values arising from classical scale factors can be severely biased or undefined as we shall illustrate in the next chapter. Wrench is the scale normalization strategy we propose to overcome this problem.*

are also of technical origin, they can well be correlated with other sources of technical variation, and can be seen to estimate technical variation beyond what is accounted for by sample depth adjustments. This was described below eqn.3.1. Thus, it is interesting to ask under what conditions compositional factors arising from scaling techniques (including our proposed technique in this work) can reconstruct X^0, the true concentrations in the source samples. From eqn. 3.1, it is clear that accurate compositional correction techniques can reconstruct true average concentration for any feature i when technical biases perturbing the feature is comparable between the treatment and control groups.

3.3 Simulation analyses

In this section, we naturally ask how several genomic normalization techniques fare in estimating compositional correction factors. Our analysis below is limited to methods that provide interpretable estimates of fold-changes. We therefore do not consider differ-

ential abundance inferences arising from rank-based methods. We also leave the analysis of non-linear normalization techniques for future work.

We note that traditional genomic normalization techniques [3, 4] like library size scaling, total sum/total count, reads per kilobase of transcript, per million mapped reads (RPKM), fragments Per kilobase of transcript per million mapped reads (FPKM), Counts per million (CPM), subsampling, rarefication based approaches are simple arbitrarily rescalings of relative frequencies, and for the purpose of this part of the thesis, one and the same. References for other genomic normalization techniques discussed will appear as appropriate.

Figure 3.2 illustrates our simulation strategy. Given the set of control proportions q_{1i} for features $i = 1 \dots p$, and the fraction of features that are perturbed across the two conditions $(1-\pi)$, we sample the set of true log fold changes ($\log \nu_{gi}$) from a fold change distribution for the random $(1-\pi)$ fraction of features that have been chosen to be perturbed. The fold change distribution (FCD) is a two-parameter distribution chosen either as a two-parameter Uniform or a Gaussian. Based on the expressions from eqn. 2.3, the target proportions were then obtained as $q_{gi} = \frac{\nu_{gi} q_{1i}}{\sum_k \nu_{gk} q_{1k}}$. Conditioned on the total number of sequencing reads τ, the sequencing output Y_{gj} for all i were obtained as a multinomial with proportions vector $q_{g\cdot} = [|q_{gi}|]_{i=1}^{p}$. We set the control proportions $q_{1\cdot}$ from various experimental sequencing datasets. With this setup, we can vary π, and the two parameters of the FCD, and ask, how various normalization and testing procedures compare in terms of their performance. Performance was quantified based on the sensitivity and specificity values in detecting truly perturbed features at a Benjamini-Hoschberg false discovery rate of .1.

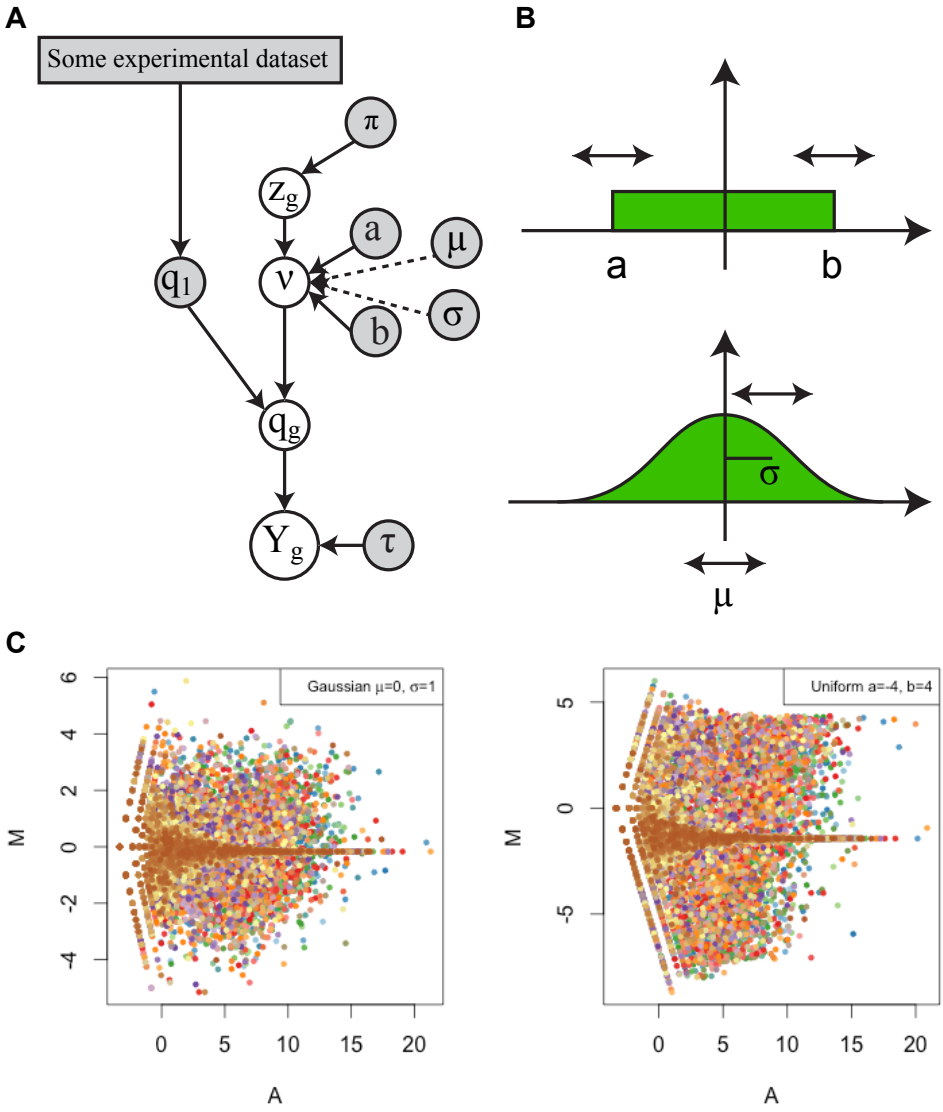

Figure 3.2: Simulation strategy for evaluating current normalization and differential expression analysis toolkits for compositional correction. *(A) Simulation set up.* $q_{1\cdot}, q_{g\cdot}$ *represent the control and case proportion vector of all the features.* $q_{1\cdot}$ *is obtained from a given experimental dataset.* π *represents the fraction of features that do not change across conditions.* $Z_{gi} \sim Bernoulli(\pi)$ *for all i represents the set of indicator variables that denote if a feature is not differentially expressed. Conditioned on* $Z_{g\cdot}$*, the logged vector of fold changes* $\log \nu$ *is sampled from a two-parameter fold change distribution, with* ν_{gi} *set to* 1 *whenever* Z_{gi} *is* 1*. Here i indexes the individual entries of the vector.*

Figure 3.2: *Continued from previous page. The sampled fold changes and control proportions are normalized to yield the case proportions. A multinomial draw for a fixed sample depth τ (20M reads) then yields the desired simulated sequencing output. The two fold change distributions, $Unif(a_v, b_v)$ and a $N(\mu_v, \sigma_v^2)$, considered in our study are shown in (B). Example M-A plots resulting from simulations when 75% (i.e., $\pi = 0.75$) of the features are fixed across conditions, with the rest perturbed according to log fold changes sampled from $Normal(0,1)$ and $Unif(-4,4)$ fold change distributions respectively are shown in (C). Each point in the M-A plot corresponds to a feature, and plots its grand average (A axis), against their empirical fold changes. Both are in* $\log_2$*-scale.*

With the above setup, we do not strictly enforce constant average total feature abundance across simulated cases and controls. We would like to keep the parameter variations sufficiently general that this condition roughly holds under some settings, while letting us appreciate the relative merits of reference normalization strategies under others.

In summary, for a given set of control proportions, we vary i) the fraction of features that change across conditions, ii) the shape, iii) mean and iv) variance of the fold change distribution that underlies the perturbation of features in the case-group, v) normalization approach and vi) testing technique. We also varied the control proportions themselves from various experimental datasets, and our results were similar. Our simulations are fairly general and should allow us to robustly characterize the performance of the current normalization and differential expression analysis practices in genomics.

Library size/Subsampling based approaches **Figure 3.3** plots the performance measures for a negative binomial based testing suite (edgeR software [124]) for a uniform fold change distribution after total sum normalization. Sensitivity values in detecting true underlying concentration changes never go beyond 65%, and heavy false positive rates are incurred even when 95% of the features remain unchanged across conditions. **Figure 3.4**

shows the performance under the Gaussian fold change distribution. In contrast to the uniform case above, we find sensitivities go up to 85%, but false positives are also accrued at higher rates. It would appear that higher variances and means lead to better performance, but as Figure 3.5 shows, many of these truly significant features were called significant for the wrong reason: wrong signs of fold changes. Higher means and variances of fold change distributions are therefore conditions that lead to heavily confounded inference under proportion based normalization strategies. These results were similar across testing platforms, and across testing techniques.

It is useful to also summarize the relevant results from previous chapter here. Total count/library size normalized data is equivalent to relative frequencies. We devoted the previous chapter to ask under what conditions, inferences made with relative frequencies alone would continue to reflect concentration changes in an unbiased manner. We formally analyzed its influence within the framework of linear models, a widely used statistical framework within several count data packages commonly used in genomics. Under the most natural adjustments based on the total count (e.g., unaltered reads per kilobase of transcript, per million mapped reads (RPKM)/ fragments Per kilobase of transcript per million mapped reads (FPKM)/ Counts per million (CPM)/subsampling/rarefication based approaches), we found that these conditions can be precisely characterized mathematically and are extremely limited in their applicability in general experimental settings. It may be tempting to argue that one can resort to total count-based normalization if total feature content is the same across conditions. However, as was shown in the last chapter, it is easy to see that this assumption is only valid when strict constraints on the levels of technical perturbation of feature abundances and sequence-able contaminants are respected,

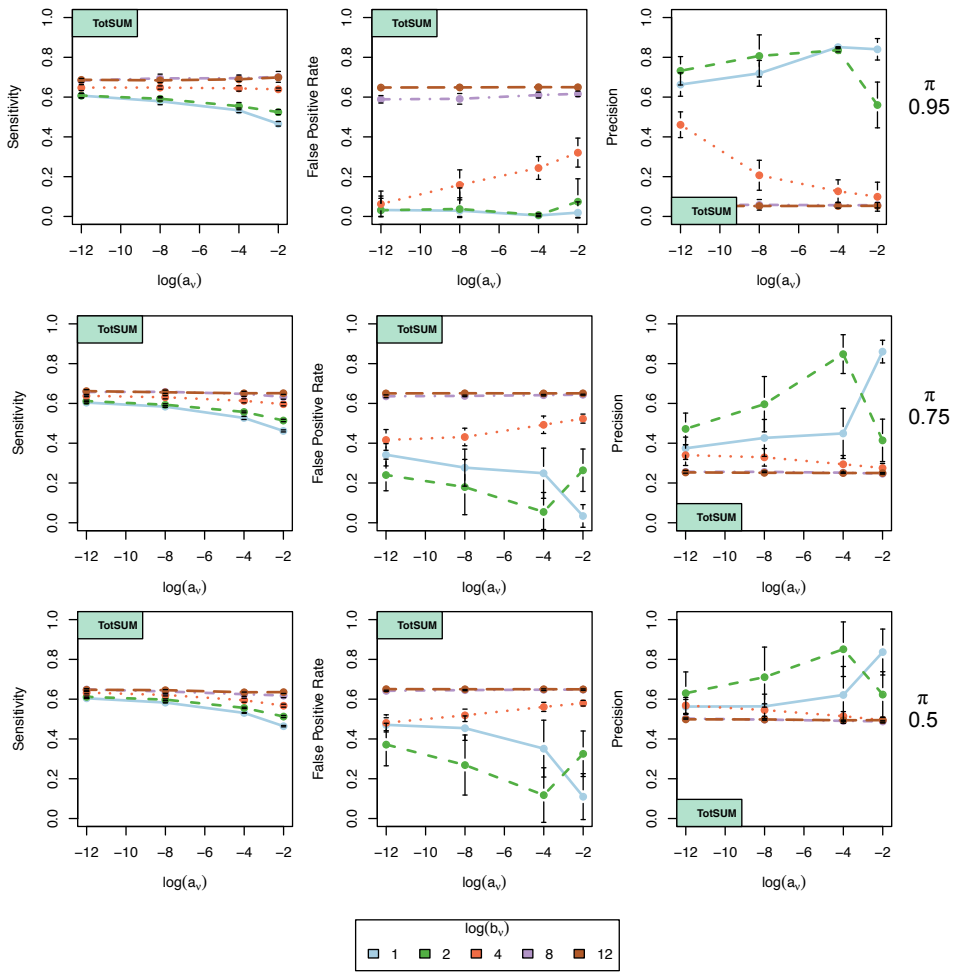

Figure 3.3: Total sum based normalization, like RPKM/Rarefication, under a Uniform fold change distribution. *The figure plots various performance metrics of the edgeR package as a function of the fraction of features that remain unchanged across conditions (π), and the lower (a_v) and upper bounds (b_v) of a Uniform fold change distribution. Control proportions ($q_{1.}$) were obtained from rat liver tissue of the rat bodymap [7]. Extremely high false positive rates result with higher variance and asymmetrically located fold change distributions (i.e., with positive or negative means) due to compositional bias. The results were similar across commonly used differential abundance testing platforms, and for the Gaussian fold change distribution (Fig. 3.4).*

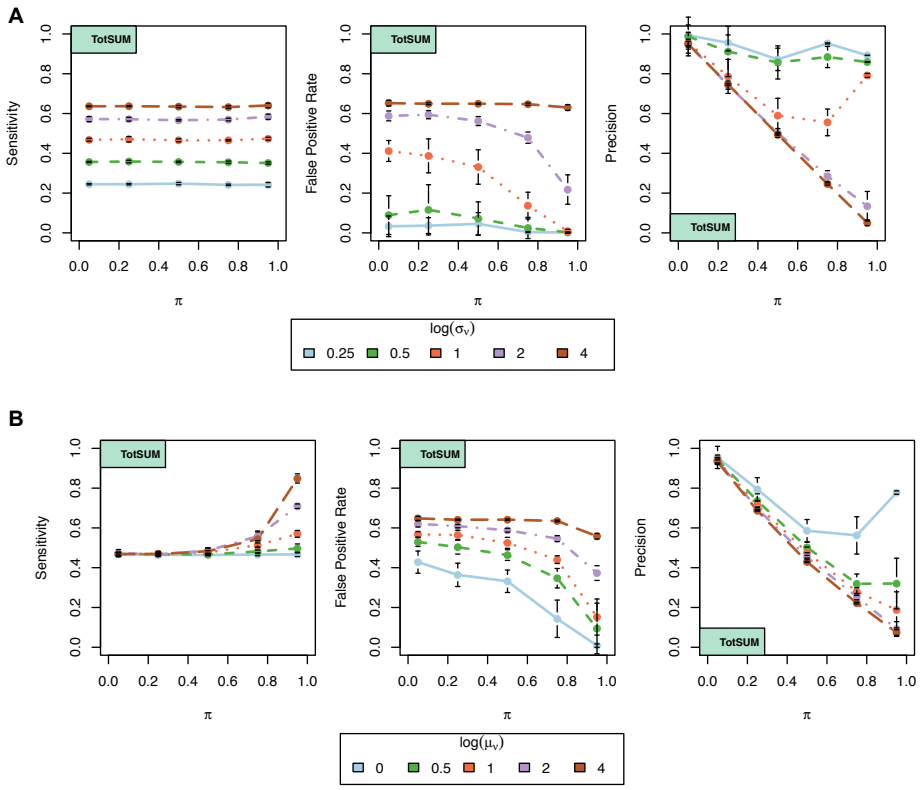

Figure 3.4: Total sum based normalization, like RPKM/Rarefication, under a Gaussian fold change distribution. *The figure plots various performance metrics of the edgeR package as a function of the fraction of features that remain unchanged across conditions (π), and the mean (μ_v) and standard deviation (σ_v) of the Gaussian fold change distribution for the same control proportions ($q_{1.}$) as in figure 5. (A) (σ_v, π) variations at $\mu_v = 0$. (B) (μ_v, π) variations at $\sigma_v = 1$. It would appear that higher fold change distribution variances and means lead to better performance, but these are also associated with higher false positive rates and as Fig. 3.5 shows, large fraction of these calls had wrong signed fold changes. Higher means and variances of fold change distributions are therefore instances that lead to heavily confounded inference. The results were similar across commonly used differential abundance testing pipelines.*

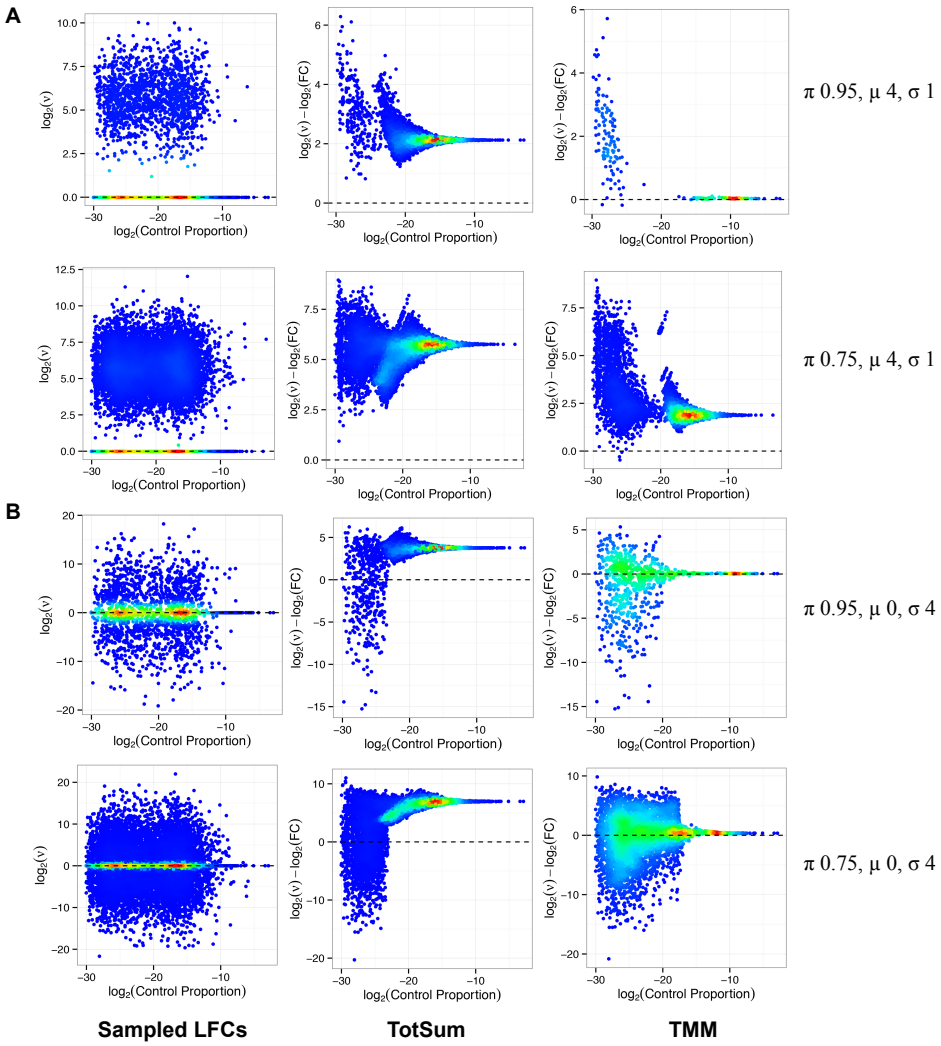

Figure 3.5: Confounded inference with total sum and reference normalization strategies. *For all features whose reconstructed fold changes had wrong signs when called significant, together with false negatives, we plot the sampled fold changes (first column) and deviations in the edgeR reconstructed fold changes from those of the true values after total sum (second column) and TMM (third column) normalizations. The corresponding parameter values for the simulations are shown alongside the plots. Larger deviations from the horizontal line at 0 imply higher confounding in inference. Asymmetric FCDs, which give rise to feature specific fold changes biased to be more positive or negative, can easily trick inference based on total sum based normalization approaches. TMM and other voting based strategies behave in a more robust fashion. However, when larger fraction of features (25%) varies across conditions, their performance becomes highly sensitive to the underlying FCDs. The color indicates the density of points, with blue, green and red indicating low, medium and high densities respectively.*

an assumption that can be very easily violated in metagenomic experiments [125–127], which usually feature high intra- and inter-group feature diversity.

Reference normalization and robust fold-change estimation techniques We now compare and contrast the aforementioned total count/library size adjustments (i.e., relative frequency measurements) with a few reference based techniques (reviewed in Table. 3.1) in overcoming compositional bias at *high* sample depths. Also, many widely used genomic differential abundance testing toolkits enforce prior assumptions on reconstructed fold changes, and moderate their estimation. This made us wonder about the robustness of these testing techniques in overcoming the false positives that would otherwise be created without compositional bias correction. With an exhaustive set of simulations at high coverage sample depths (similar to bulk RNAseq) with $20M$ reads per sample, by and large, we found that all testing packages behaved the same way, and the key ingredient to overcome compositional bias always was an appropriate normalization technique . We also found that reference based normalization procedures outperformed library size based techniques significantly, re-emphasizing the analytic insights we mentioned previously.

Figures 3.6 and **3.7** demonstrate the performance of TMM normalization, a reference based normalization strategy. In contrast to the above total sum-based normalization procedure, the false positive rates with TMM were maintained low, if not at zero, for a variety of parameter settings. At higher FCD means and variances, they also lead to wrong reconstruction of fold change signs but with a highly desirable twist: as long as the fraction of perturbed features across conditions is small, the fold change distribution is correctly centered throughout the abundance distribution except for those features with

very low abundances leading to very low false positive rates Fig. 3.5. For all normalization techniques, as the amount of features that change across conditions increases, false positive rates increase.

In the next chapter, we consider applying these techniques to the problem of compositional bias correction in metagenomic survey data. The data pose an interesting challenge, as the microbial abundance measurements resulting from them can be extremely sparse.

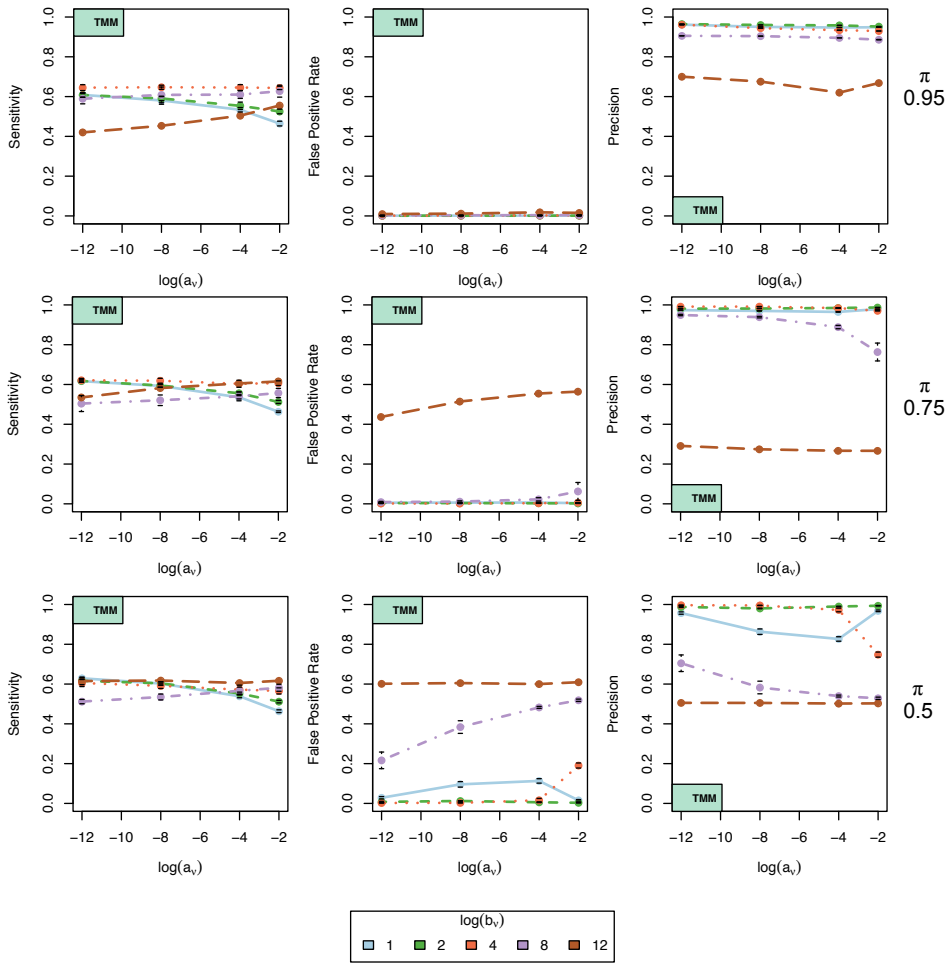

Figure 3.6: Reference normalization (TMM/DESeq/Median) under a Uniform fold change distribution. *The figure plots various performance metrics of the edgeR package with TMM normalization as a function of the fraction of features that remain unchanged across conditions (π), and the lower (a) and upper bounds (b) of a Uniform fold change distribution. Control proportions ($q_{1\cdot}$) were obtained from rat liver tissue of the rat bodymap [7]. In contrast to what was observed with total sum approaches, the false positive rates are maintained at low levels for a larger range of parameters. Sensitivity values still remained low. High false positive rates result with higher variance and asymmetrically located (with respect to 0) fold change distributions. The results were similar across testing platforms, for median based normalization techniques like DESeq/Median scaling, and for the Gaussian fold change distribution.*

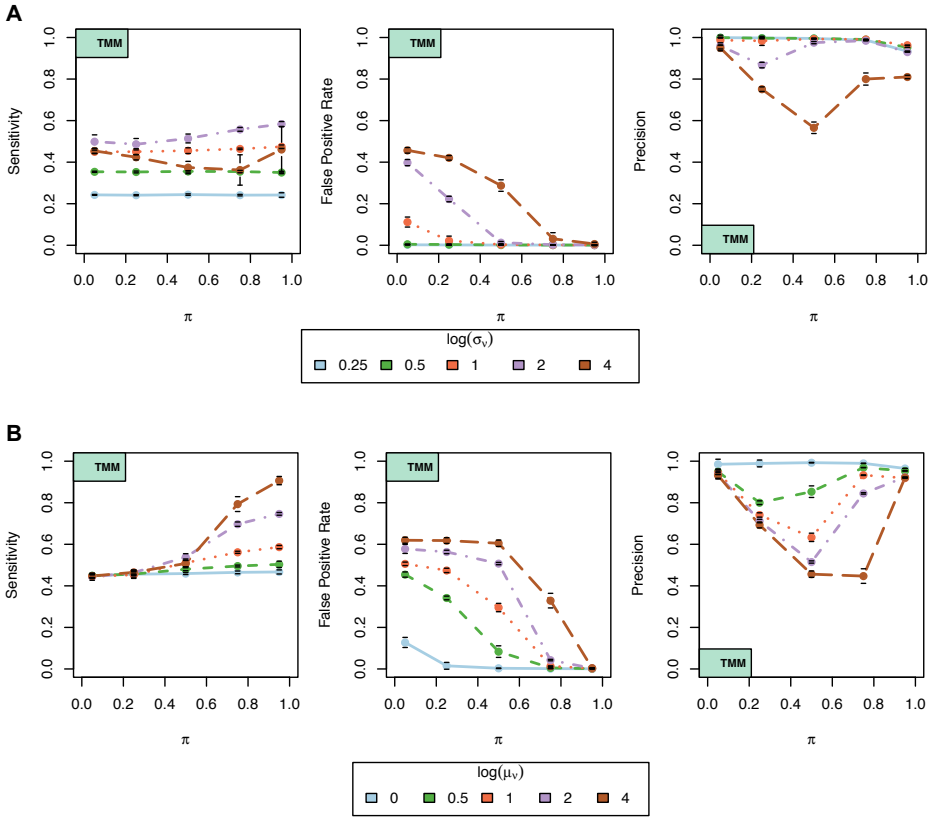

Figure 3.7: Reference normalization (TMM/DESeq/Median) under a Gaussian fold change distribution. *The figure plots various performance metrics of the edgeR package as a function of the fraction of features that remain unchanged across conditions (π), and the mean (μ_v) and standard deviation (σ_v) of the Gaussian fold change distribution for the same control proportions ($q_{1.}$) as in figure 3.6. (A) (σ_v, π) variations at $\mu_v = 0$. (B) (μ_v, π) variations at a constant $\sigma_v = 1$. When the fraction of unperturbed features is large, in contrast to what was observed with total sum approaches, higher fold change distribution variances and means lead to better performance. Figure 3.5 shows, many of these calls had wrong signed fold changes. Higher means and variances of fold change distributions are therefore cases that lead to heavily confounded inference. The results were qualitatively similar across commonly used differential abundance testing pipelines.*

Chapter 4

A scaling normalization technique for estimating compositional bias from sparse relative frequency data.

From previous chapters, we recall that:

1. The output of a sequencing machine only retain relative frequencies, and not the concentrations of the measured features. We call this unwanted technical bias introduced in our experiment as compositional bias. Compositional bias is present in the output of all derived technologies in genomics like RNAseq, ChipSeq etc., which exploit a DNA sequencing machine for quantification purposes.

2. Compositional bias can be corrected by estimating compositional correction factors. These factors are more general in that they correlate with other unwanted technical variation infused in the data, beyond compositional bias, as well.

3. Compositional correction factor is a linear technical bias shared by all features measured in a sequencing experiment. This fact makes it possible to exploit various scale normalization techniques in genomics to estimate them.

In this chapter, we consider the problem of estimating compositional correction factors for metagenomic 16s surveys, another derived technology based on sequencing.

Recognizing that 16s ribosomal RNAs (rRNA) are relatively specific for every prokaryotic Genus, Carl Woese and George Fox suggested that a simple strategy for identifying prokaryotic genera in a microbial sample is to sequence and catalogue the 16s rRNA sequences in it [128–133]. In fact, Woese & Fox demonstrated the promise of such a technology, rather dramatically, by adding a whole new domain of life – the *archaea* – to the phylogenetic tree of life! [129, 134]

The number of times a given prokaryotic genera's 16S sequence is found in the sequencing output serves as a measure of its frequency. This is the idea behind 16s marker gene surveys [135–138], which have now found widespread utility in biomedical research [139] and natural history studies involving large-scale oceanic microbial ecosystems [8, 138]. Like with other derived technologies, compositional effects are observable in the count data from the large-scale Tara oceans metagenomics project [8], (**Fig. 4.1**), in which a few dominant taxa are attributable to global differences in the between-oceans fold-change distributions.

We demonstrate that our strategy of adapting traditional genomic normalization techniques (discussed in the previous chapter) for estimating compositional bias fail with 16S survey data. This is mainly because a large fraction of features (the distinct 16s sequences) in 16S count data are very sparsely observed in the output. Given that all reference based normalization techniques base their compositional scale factor estimates on ratios of proportions, the large fraction of zeroes in the 16s survey data lead to mostly zero valued compositional correction factor estimates: DESeq failed to provide a solution for all the samples in a 16s survey of our interest, and TMM based its estimation of scale factors on very few features per sample (as low as 1). The median approach simply re-

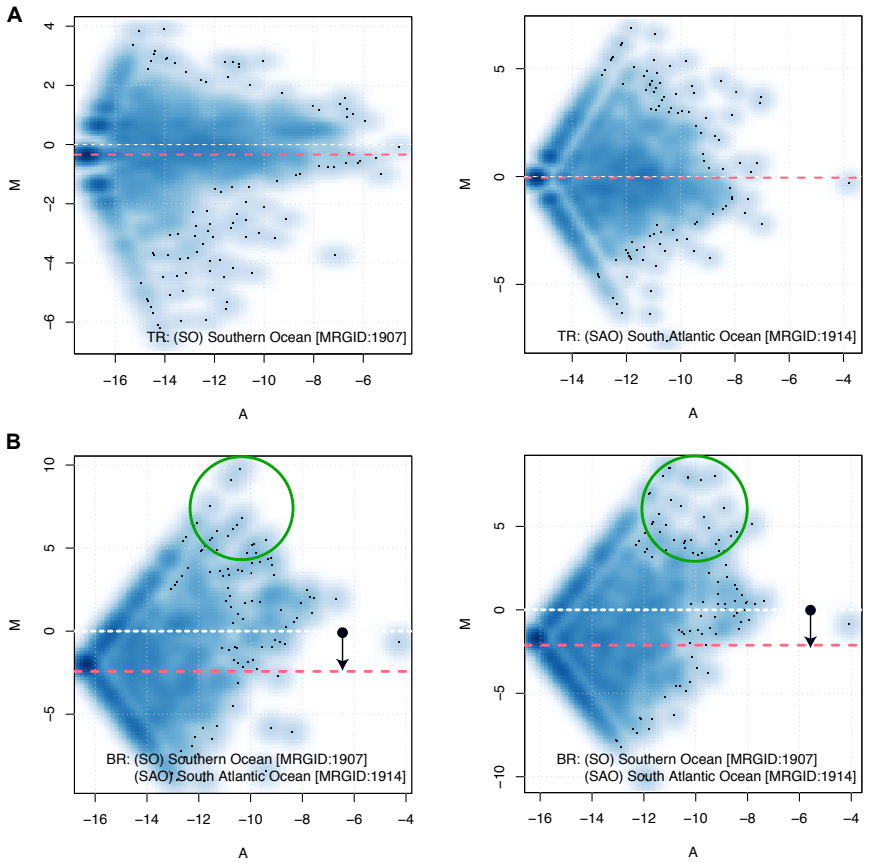

Figure 4.1: Importance of compositional bias correction in sparse metagenomic data. *(A) M-A plots of 16S reconstructions (from high sequencing depth, whole metagenome shotgun sequencing experiments) from two technical replicates each from the Tara oceans project [8] generated for the Southern and South Atlantic Oceans. In all subplots, x-axis plots for each feature, its average of the logged proportions in the two compared samples; y-axis plots the corresponding differences. The red dashed line indicates the median log fold change, which is* 0 *across the technical replicates. (B) M-A plots of the same replicates but plotted across the two oceans. The median of the log-fold change distribution is clearly shifted. A few dominant taxa in the South Atlantic Ocean (circled) are attributable for driving this overall apparent differences in the observed fold changes. The Tara 16s dataset, reconstructed from very deep whole metagenome shotgun experiments of oceanic samples, albeit boasting of an average 100,000 16S contributing reads per sample, still encourages a median 88% feature absence per sample.*

turned zero values. CLR transforms behaved similarly. When one proceeds to avoid this problem by adding pseudo-counts, owing to heavy sparsity underlying these datasets, the

transformations these techniques imposed mostly reflected the value of pseudocount and the number of features observed in a sample. A recently established scaling normalization technique, Scran [6], tried to overcome this sparsity issue in the context of single cell ribonucleic acid sequencing (scRNAseq) count data – which also entertains a large fraction of zeroes – by decomposing simulated pooled counts from multiple samples. That approach, developed for relatively high coverage single cell RNAseq, also failed to provide solutions for a significant fraction of samples in our datasets (as high as 74%). Furthermore, as we illustrate later, compositional bias affects data sparsity, and normalization techniques that ignore zeroes when estimating normalization scales (like CSS [5], and TMM) can be severely biased. The relatively low sequencing depth per sample (as low as 2000 reads per sample), large number of features and their diversity across samples thus pose a serious challenge to existing normalization techniques.

In this chapter, we develop a compositional bias correction technique (*Wrench*) for sparse count data based on an empirical Bayes approach that borrows information across features and samples. We demonstrate its improved performance in metagenomic 16S survey data. Based on the distribution of compositional scale factor estimates arising from several publicly available large scale 16S count datasets, we argue that detailed experiments specifically addressing the influence of compositional bias in metagenomics are needed.

4.1 Classic scale normalization techniques suffer with sparse 16s count data

In **Fig. 4.2**, we plot the feature-wise compositional scale estimates (i.e., ratio of sample proportion to that of the reference; third column entries in Table. 3.1), obtained from TMM and DESeq for a sample in two different 16S microbiome datasets. TMM computes a weighted average over these feature-wise estimates, while DESeq proposes the median. The first column corresponds to a bulk RNAseq study of the rat body map [7]; the second corresponds to those from a 16S metagenomic dataset [139]. Strikingly, while a large number of features agree on their scale factors for a sample arising from bulk RNAseq for both TMM and DESeq strategies, the sparse nature of metagenomic count data makes robust estimation of their scale factors extremely difficult. Furthermore, large variance is also observed across the scale factors suggested by the individual features. Clearly, a moderated estimation procedure is warranted.

One might wonder if adding pseudocounts to the orginal count data (a common procedure in metagenomic data analysis [118, 140]) effectively deals away with the problem. However, as shown in **Fig. 4.3**, with large number of features absent per sample, these scale factors roughly reflect the value of the pseudocount, and are systematically scaled down in value as sequencing depth, which is strongly correlated with feature presence, increases. This result suggests that addition of pseudocounts to data need not be the right strategy for deriving normalization scales based on CLR [141] or other similar methods, especially when the data is sparse. The alternate idea of only deriving scale factors based on positive values alone, are also associated with problems as we will see later in the text.

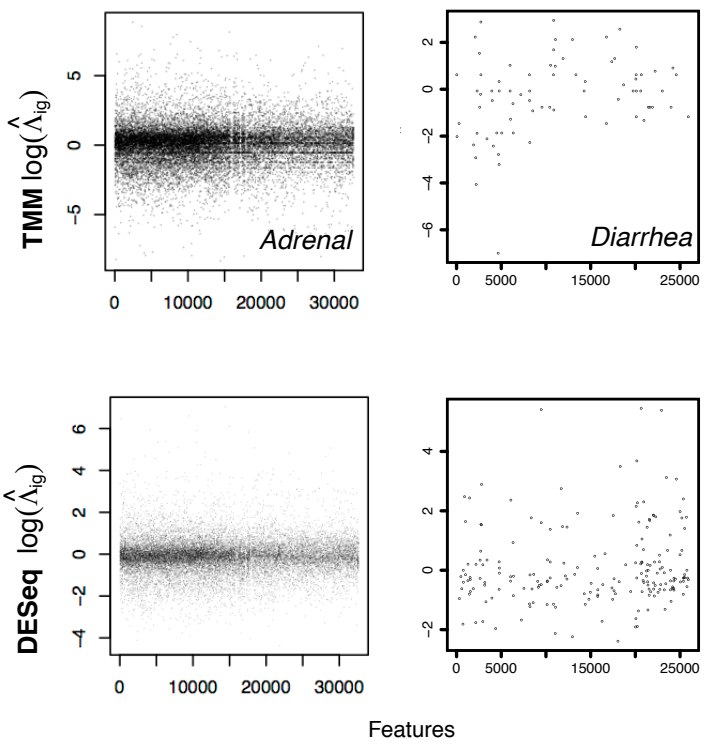

Figure 4.2: Estimation of compositional correction scales from sparse count data. *On the left column, we plot the feature-wise ratio (Λ_{gji}) estimates adjusted for sample depth from each feature i in one of the samples from the Adrenal tissue of the rat body map dataset (bulk RNAseq), and on the right column, we plot the same values arising from a sample in the Diarrheal dataset (16S metagenomics). The top and bottom rows correspond to the scales estimated using TMM and DESeq respectively. In the case of bulk RNAseq data, large numbers of individual feature estimates agree on a compositional scale factor. Simple averaging, or some robust averaging would help us obtain the scale factor exactly.* *Continued on next page.*

4.2 The proposed technique (Wrench) reconstructs precise group-wise compositional factor estimates

To overcome the issues faced by existing techniques, we devised an approach based on the following observations and assumptions. First, aggregated group/condition-wise feature count distributions are less noisy than sample-wise feature count distributions,

Figure 4.2: *Continued from previous page. A similar robust behavior is observed with all the tissues available in the bodymapRat dataset (considered later in text). On the second column, we plot the feature-wise ratio values from a metagenomic 16S marker gene survey of infant gut microbiota. There is no general agreement among the features on the scale factors, and simple averaging will not work. We note that what we have shown are fairly good cases. Several samples entertain only a few tens of shared species with an arbitrary reference sample within the dataset. In this work, we aimed to model this variability and estimate the scale factors robustly by borrowing information across features and samples.*

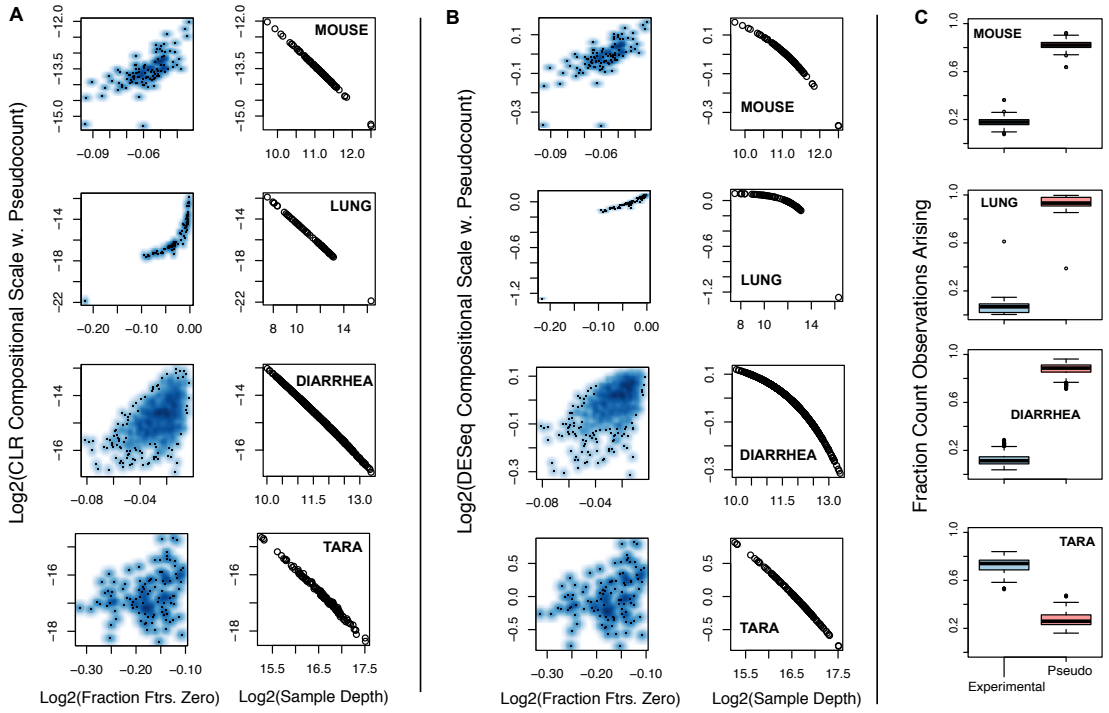

Figure 4.3: Adding pseudocounts leads to biased normalization. *For each of the four microbiome count datasets (rows: Mouse, Lung, Diarrheal and Tara Oceans), we plot (**A**) CLR and (**B**) DESeq compositional scales obtained after adding a pseudo count value of* 1*, as a function of fraction of features that are zero in the samples (first column) and the sample depth (second column). The observed behavior was not sensitive to the value of pseudocount used. A similar plot was alos generated for a pseudocount value of* 10^{-7}*. (**C**) shows the total number of pseudocounts added, which is essentially the number of features observed in a dataset, and the total actual counts observed in the dataset divided by their sum i.e., the total implied sequencing depth after pseudocounts addition. Continued on next page.*

Figure 4.3: *A large fraction of sequencing depth in the new pseudocounted dataset is now arising from pseudocounts than the true experimental counts, when the data is excessively sparse. Indeed, if the pseudocount value is altered to a very low positive fraction value, the boxplots will reflect reversed locations, but this plot is only used to stress the level of alteration made to a dataset. Only in the Tara Oceans project, where the sample depth is 100K reads, do the boxplots shift. However, at a roughly median 90% features absent, that data when altered by pseudocounts, also leads to biased scaling factors as seen in (A) and (B).*

and it may be useful to Bayes-shrink sample-wise estimators towards that of group-wise global estimates. Second, zero abundance values in metagenomic samples are predominantly caused by competition effects induced by sequencing technology (illustrated in Fig. 3.1), and therefore can be indicative of large changes in underlying compositions[1] with respect to a chosen reference. Indeed, ignoring sterile/control samples, the median fraction of features recording a zero count across samples in the mouse, lung, diarrheal, human microbiome project [142] and (the very high coverage) Tara oceans [8] datasets were: .96, .98, .98, .98 and .88. These respectively had median sample depths of roughly $2.2K$, $4.5K$, $3.3K$, $4.4K$ and $100K$ reads. In direct contrast, this value for the high coverage bulkRNAseq rat body map across 11 organs at a median sample depth of $9.7M$ reads, is .33. Large number of features, extreme diversity, and time-dependent dynamic fluctuations in microbial abundances can result in such high sparsity levels in metagenomic datasets. When working within the fundamental assumption that *most features do not change across conditions*, such extraordinary sparsity levels can then be attributed, by and large, to competition among features for being sequenced. As we illustrate in **Fig. 4.4**, zero observations in a sample are correlated with compositional changes, and truncated

[1] the idea being that in the limit $\Lambda_g \to \infty$, feature-wise relative frequency ratios that reflect Λ_g^{-1}, $\to 0$. Ref table 3.1 for discussions.

analyses that ignore them (as is done with TMM / DESeq / metagenomic CSS normalization techniques) effectively leads to loss of information and results that are opposite to what is expected.

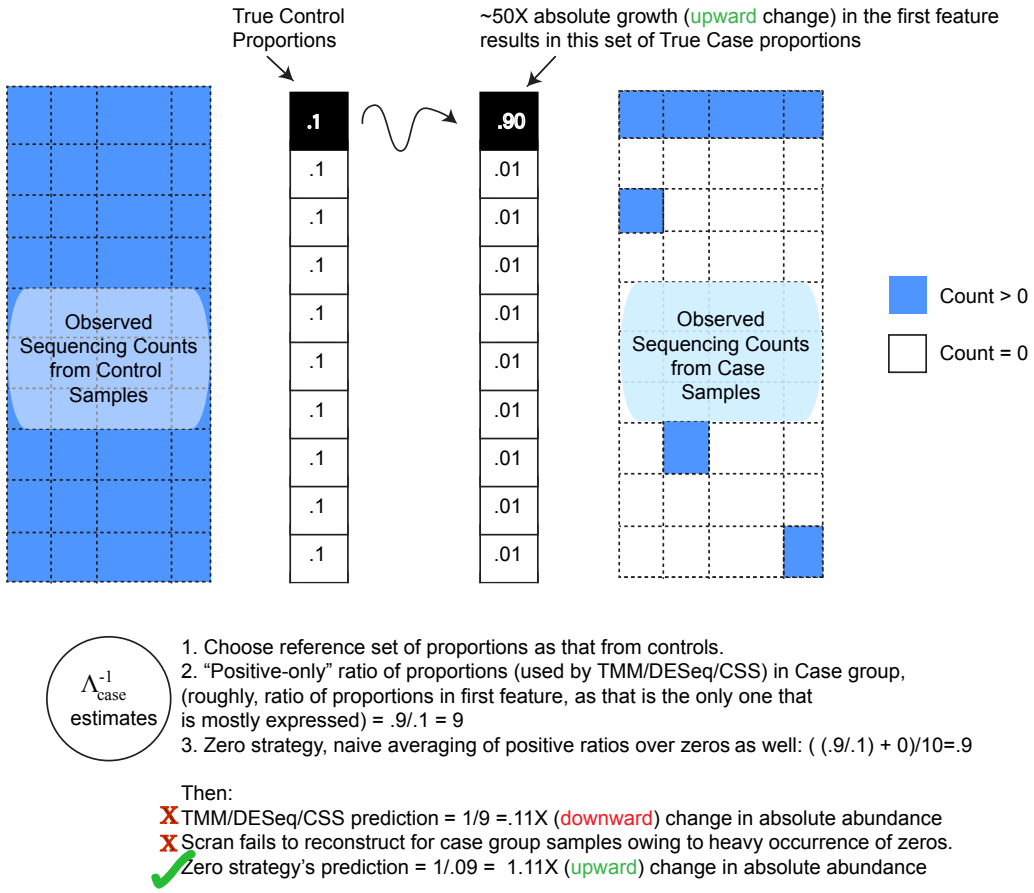

Figure 4.4: Ignoring zeroes can introduce bias in normalization, when zeroes predominantly arise from under-sampling. *An artificial example with* 10 *features and two groups ("controls" and "cases"), when one of the features undergoes a roughly 50X expansion (a* $\log_2$ *fold change of* 5.64*) in cases compared to controls. This drives the relative frequencies of the rest of the* 9 *features relatively low in the case group. As a result features that are largely present in the controls are not observed in the case group at moderate sequencing depths. Scaling normalization strategies that derive scales based only on the positive count values, can underestimate compositional changes as shown.*

We now give a brief overview of the technique (Wrench) proposed in this work.

More details are presented in the Methods section at the end of this chapter. With average proportions across a dataset as our reference, we model our feature-wise proportion ratios as a hurdle log-normal model[2], with feature-specific zero-generation probabilities, means and variances. For the purpose of metagenomic applications, and analytic convenience, we slighty relax the standard assumption that most features do not change across conditions by assuming that the feature-wise log-fold changes arise from a zero mean Gaussian distribution, a common assumption in differential abundance analysis [5, 143, 144]. The analytical tractability of the model allows us to standardize the feature-wise values within and across samples, and derive the compositional scale estimates by basing heavy weights on less variable features that are more likely to occur across samples in a dataset. In addition, to make the computed factors robust to low sequencing depths and low abundant features, we employ an empirical Bayes strategy that smooths the feature-wise estimates across samples before deriving the sample-wise factors. Such situations are rather common in metagenomics, and some robustness to overcome heavy sampling variations is desirable.

Table. 4.1 succinctly illustrates where current state of the art fails, while more comprehensive simulations illustrating the effectiveness of the proposed approach presented in **Fig. 4.5**. To generate table 4.1, roughly, we simulated two experimental groups, with $54K$ features whose proportions were chosen from the lung microbiome data, and let 35% of features change across conditions (see Methods for details on simulations). The net true compositional change resulting from each simulation, and their corresponding

[2]the random variable assumes a value of zero with probability π and a positive value based on its specific log-normal distribution with probability $(1-\pi)$

reconstructions by the various techniques when the count data are generated at different sequencing depths are shown. The following observations form the theme of these, and the more elaborate simulations summarized in Fig. 4.5: 1) TMM/CSS, because they focus on positive-valued observations only, are restricted in the range of scales they can reconstruct. 2) Scran can yield accurate estimators at very large sequencing depths when high feature-wise coverages are achieved. Unfortunately, this behavior is highly dependent on the underlying feature proportions and their diversity. 3) Wrench estimators offer better alternatives for under-sampled data, and as we shall observe below in their empirical performances, they can still offer robust protection against compositional bias at higher coverages. Similar results were obtained when Wrench was compared to pseudocounted CLR. In addition, **Figs. 4.6**, and **4.7** explore simulation performance as a function of group-wise sample size in balanced and unbalanced designs, where we find the performance to stabilize between roughly $10-20$ samples, depending on the fraction of features that change across conditions.

We briefly note a key ingredient about our simulation procedure. Simulating sequencing count data as *independent* Poissons / Negative Binomials – as is commonly done in benchmarking pipelines – does not inject compositional bias into simulated data. From the perspective of performance comparisons for compositional correction, doing so is therefore inappropriate. A renormalization procedure after assigning feature-wise fold-changes is necessary. Alternatively, if absolute concentrations are generated, subsampling to a desired sample depth needs to be performed.

Net Compositional Change (Λ_g)	Average Sample Depth	CLR	TMM	CSS	Scran	W_0	W_1	W_2	W_3
36.86X	1M	1.36	1.45	5.41	22.57	19.32	31.44	30.65	32.01
7.75X	10K	.95	3.05	1.47	12.08 (14/40 samples failed)	5.30	6.32	6.31	6.70

Table 4.1: Example simulations illustrate the limitations of current techniques. *Shown are the group-wise true and reconstructed compositional scales from the methods compared on two simulated examples, each at different sequencing depths and at different total true concentration changes for a roughly* $54K$ *features with control group proportions derived from the Lung microbiome. Low-coverage and/or high compositional changes are problematic for current techniques due to the sparsity they cause in the count data.* $W_1, \ldots W_3$ *are Wrench estimators proposed in the Methods section that adjust the base estimator* W_0 *for feature-wise zero-generation properties. All are presented here for comparison purposes. Our default estimator is* W_2.

4.2.1 Wrench has better normalization accuracy in experimental data

Below, we show five different results illustrating the improvements Wrench offers over existing techniques in experimental data. The first two show that Wrench leads to reduced false positive calls in differential abundance inference, while the other three demonstrate the improved quality of positive associations.

Reduction of false positives We used two approaches to compare the performance of Wrench in reducing false positive calls in differential abundance inference. Each of these analyses was performed across all biological groups with atleast 15 samples in the mouse (2 diet types), Diarrheal (2 groups), Tara (5 oceans), HMP (JCVI, 16 body sites), and HMP (BCM, 16 body sites) and averaged the results across these 41 experimental groups.

We ignored the lung microbiome for these analyses as Scran had particular difficulty making direct comparisons hard. Owing to the heavy sparsity in these datasets, Scran failed to provide scales for 53 out of 72 samples of the lung microbiome, 10 out of 132

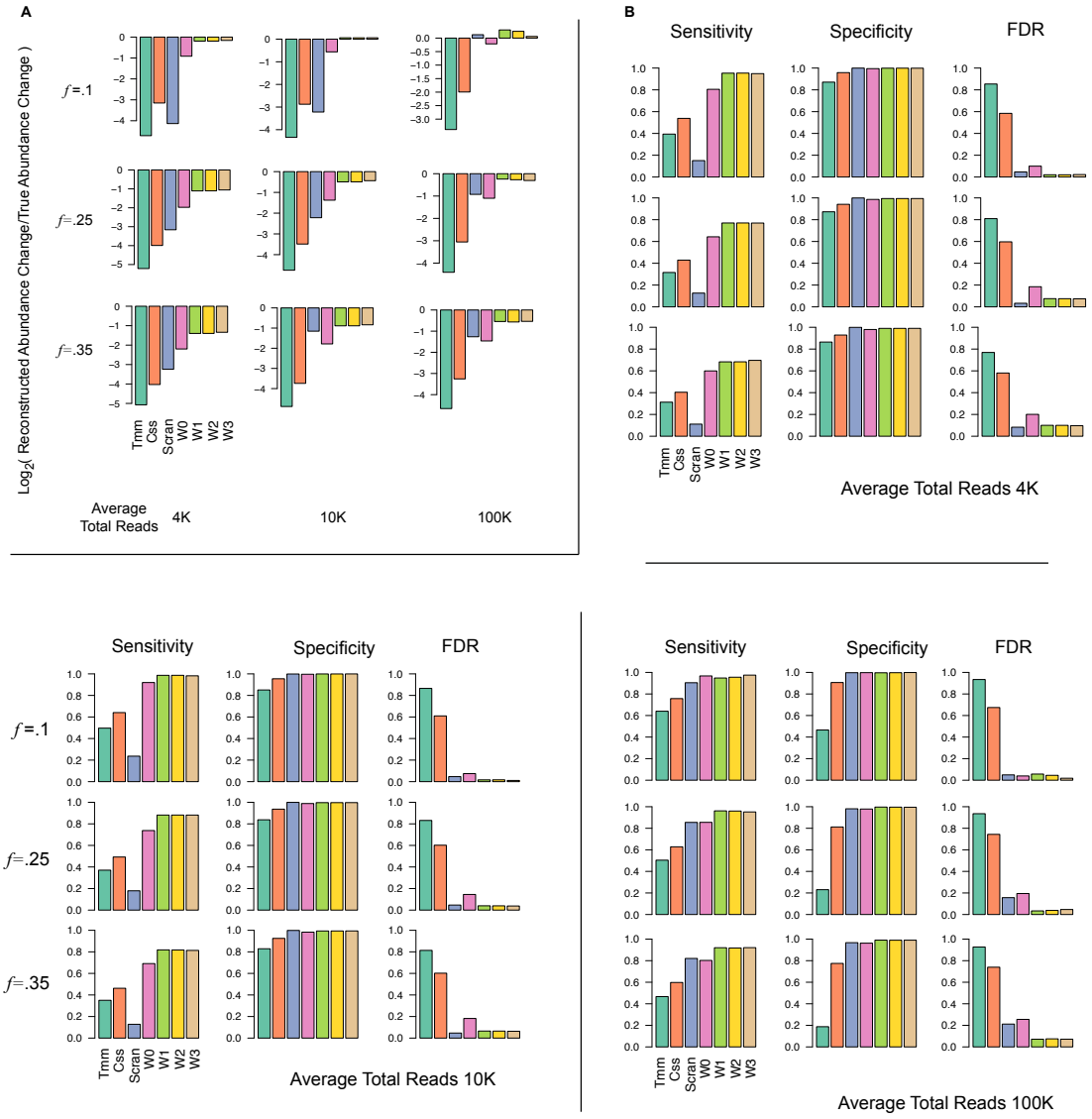

Figure 4.5: Wrench scales outperform competing approaches in reconstructing compositional changes and in differential abundance testing. *Multiple iterations of two group simulations are simulated with various fractions of features perturbed across conditions (rows, f in figures), total number of reads. Their average accuracy metrics in reconstruction and differential abundance testing are plotted. The control proportions were set to those obtained from the mouse microbiome dataset. Continued on next page.*

observations of the mouse microbiome, 6 out of 992 samples of the diarrheal dataset. Notice that Wrench not only recovers compositional scales for these samples, but also at

Figure 4.5: *Continued from previous page. (A) Average log ratios of reconstructed to true concentration changes. Each row corresponds to a particular setting of f, and each column a particular setting of average sequencing depth. Scran also suffered from being unable to provide scales for samples in each simulation set (sometimes as high as 60% of the samples at 4K and 10K average reads). (B) Average sensitivity, specificity and false discoveries at FDR .1 of detecting true differential concentration abundances. W_0 is the regularized Wrench estimator without sparsity adjustments and $W_1, ..W_3$ are various adjusted estimators compared here. For details on this and simulations, see Methods. Behavior was similar for other parameteric variations (variances of global and sample-wise fold change distributions, number of samples) of simulations.*

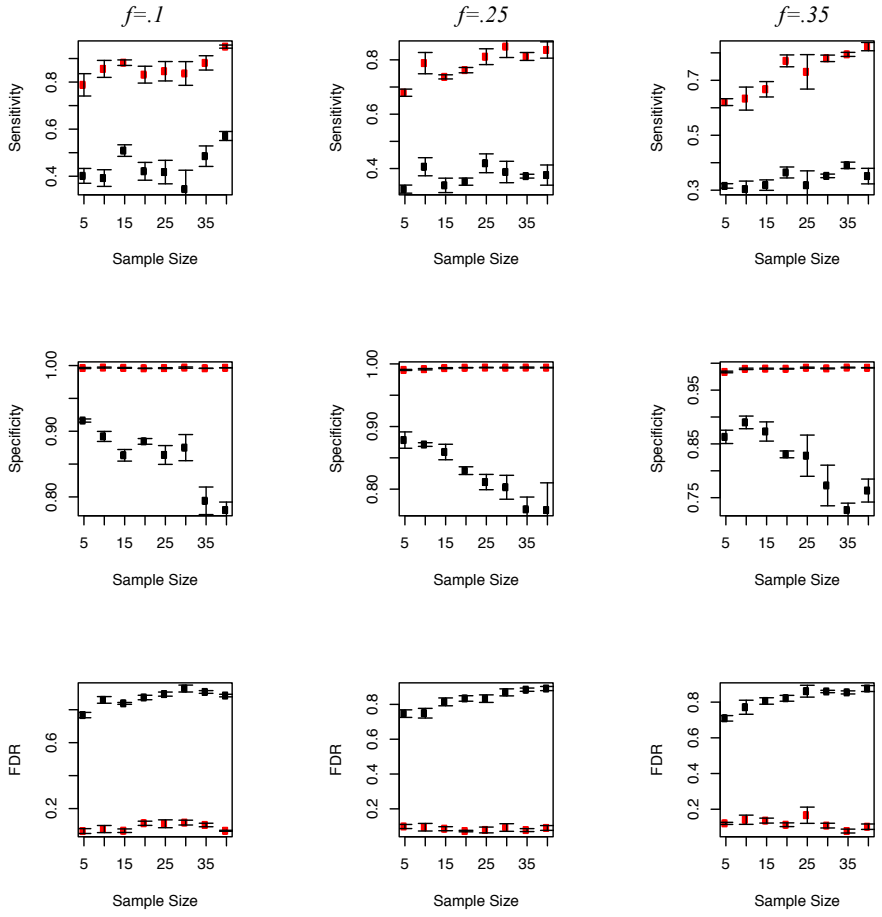

Figure 4.6: Simulation performance in a balanced design. *We plot the performance metrics as a function of sample size and fraction of features f that are perturbed in cases. The sample depth was fixed to* $10K$ *reads on average per sample. TMM is provided for reference. Legend: Red, Wrench; Black: TMM.*

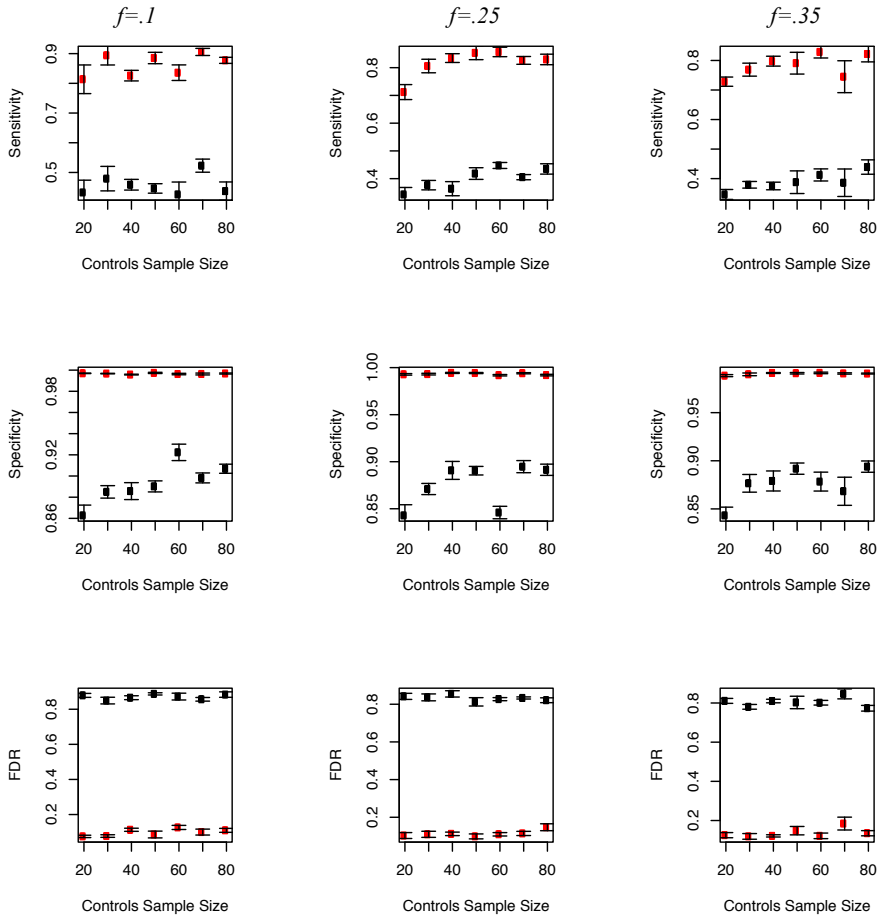

Figure 4.7: Simulation performance in an unbalanced design. *We plot the performance metrics as a function of sample size and fraction of features f that are perturbed in cases. The total number of case samples were fixed to 20, and the number of control samples were varied to simulate unbalanced designs. So in the plot, a sample size of 20 corresponds to a sample size of 20 for the case sample, and therefore reflects a balanced design. The rest represent unbalanced designs. The sample depth was fixed to* $10K$ *reads on average per sample. TMM is provided for reference. Legend: Red, Wrench; Black: TMM.*

magnitudes that were coherent with other samples from similar experimental groups (see next subsection) indicating some validity for the computed normalization factors.

First, a standard resampling analysis was performed. For every given experimental group, two artificial groups are repeatedly constructed via resampling (without replacement), and the total number of significant calls made during differential abundance analysis is recorded in each repetition. For each iterate, we compute the $\log_2(F_{Other}/F_{Wrench})$ ratio, where F_{Other} is the total number of significant calls made by a competing method (Total Sum / TMM / Scran / CSS) and F_{Wrench} is the total number of significant calls made by Wrench. If Wrench is superior these logged ratios should be > 0. The average of these ratios across all the experimental groups mentioned above is plotted in **Fig. 4.8A**, and we find Wrench meeting the goal. Although total sum does not show a significant difference in this analysis, as illustrated next, it is insufficient in capturing the null variation in the data.

We next exploited the offset-covariate approach introduced in [6]. For every feature/OTU within a homogenous experimental group, two generalized linear models are fitted: in model (a) Wrench normalization factors as offset, and those of a competing method as covariate. In model (b), normalization factors from a competing method as offset, and those of Wrench as covariate. The number of features for which the covariate term was called significant is recorded in both (a) and (b). We will denote them respectively as C_{Wrench} and C_{Other}. If Wrench sufficiently captures the variation in data, the number of times the covariate term from a competing method is called significant will be low. That is: the logged ratio $\log_2(C_{Other}/C_{Wrench})$ must be > 0. The average of these values across all the experimental groups mentioned above is plotted in **Fig. 4.8B**, and we

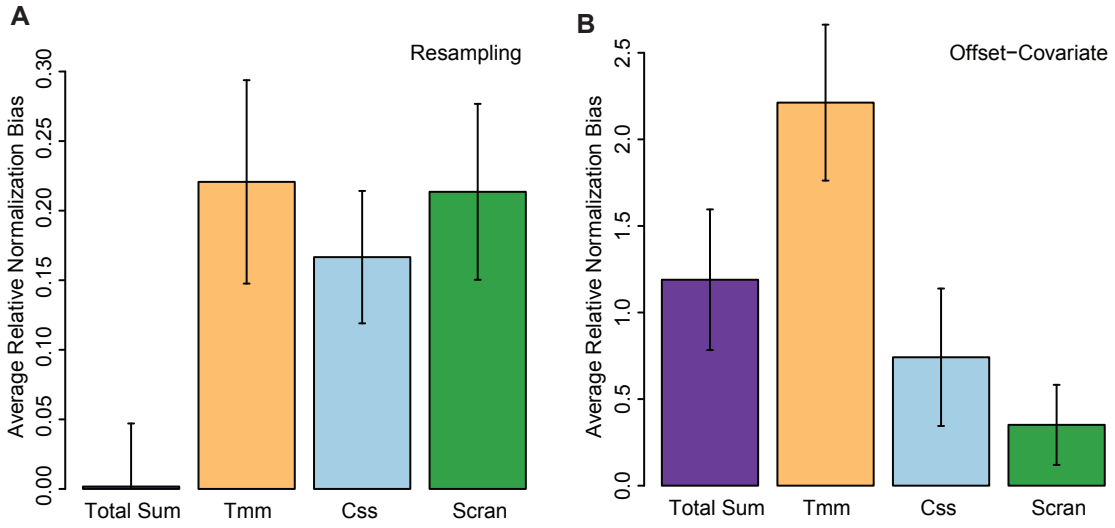

Figure 4.8: ***Wrench scales lead to reduced false positive calls.*** *(A) The average of* $\log_2(F_{Other}/F_{Wrench})$ *values obtained over artificial two group splits of homogeneous experimental group data is shown and (B) the average of* $\log_2(C_{Other}/C_{Wrench})$ *values across* 41 *metagenomic experimental groups are shown. Standard error bars are shown. In both plots, positive values for a method imply reduced accuracy relative to Wrench.* F_{Other}*: total number of diffferentially abundant features found by a competing method (total sum, TMM, CSS or Scran).* F_{Wrench}*: total number of differentially abundant features found by Wrench.* C_{Other}*: total number of features where the covariate term for Wrench normalization factors were found to be significant when competing method is used as offset.* C_{Wrench}*: total number of features where the covariate term for a competing method's normalization factors were found to be significant, when Wrench is used as covariate.*

find Wrench to improve upon other techniques.

Improved association discoveries To compare the quality of associations achieved with the various normalization methods, we re-analyzed the Tara Oceans 16S microbiome dataset.

Even though the contribution of true compositional changes and other technical biases are not identifiable from the compositional scales without extra information, we asked if the reconstructed scales correlate with orthogonal information on absolute abundances, and other measures of technical biases. The results are summarized in **Table 4.2**.

Dataset	Type	CLR	TMM	CSS	Scran	W_0	W_1	W_2	W_3
Tara Oceans [8]	16s (from Whole Metagenome)	0 (-2.65×10^{-6})	0.26	0.15	0.52	.58	.54	.53	.53
Rat BodyMap [7]	Bulk RNAseq	-0.36	0.22	0.16	0.18	.20	.19	.20	.26
Embryonic Stem Cells [62]	UMI/scRNAseq	-0.70	.70	.67	.67	.71	.70	.70	.68

Table 4.2: Correlations of compositional scales with orthogonal measurements on concentrations/technical biases. *Correlations of logged reconstructed abundance factors (1/compositional correction factor) with logged total flow cytometry cell counts is shown for the Tara project. Correlations of logged normalization factors with logged total ERCC counts are shown in the case of the rat body map and embryonic stem cells datasets. Given the high sparsity in these datsets, CLR factors computed by adding pseudocounts, essentially had no information on technical biases.* $W_1, \ldots W_3$ *are estimators proposed in the Methods section that adjust the base estimator* W_0 *for feature-wise zero-generation properties. All are presented here for comparison purposes. The default Wrench estimator (*W_2*) compares well at low and high coverage settings. For more details on these and the distinction in terminology between compositional correction factors and normalization factors, refer Materials and Methods.*

Interestingly, in the very high coverage Tara Oceans metagenomics project, Wrench and Scran estimators achieve comparable correlations (>50%) with absolute flow cytometry measurements of microbial counts from the Tara Oceans project. Scran failed to reconstruct the scales for 3 samples. TMM and CSS had substantially poor correlations. Similarly, Wrench normalization factors had comparable/slightly better correlations to the total ERCC spike-in counts in bulk and single cell RNAseq datasets. In direct contrast, CLR scale factors (the geometric means of proportions) computed with pseudocounts were either uncorrelated or highly anti-correlated with the aforementioned measurements reflecting technical biases. These results reaffirm that there are advantages to exploiting specialized compositional correction tools even with microbiome datasets teeming with microbes of extraordinary diversity.

We next analyzed the quality of differential abundance inference arising from competing normalization techniques, by performing two sets of enrichment analyses.

Detailed tables presenting the following results are provided as a supplement *additional file 2* in our related publication [145]. In the first procedure, we extracted broad genus-level functional annotations from the Faprotax database [146], and tested for their enrichment in positively associated genera in the deep chlorophyll (DCM) and the mesopelagic layer (MES) samples of the oceans relative to the surface layer. The total number of significantly differentially abundant OTU calls were widely different across techniques: Wrench and Scran made roughly 30% fewer calls compared to total sum, TMM, and CSS. Given the relatively general nature of the annotations, all methods yielded expected annotations in the DCM and MES layers based on previous studies, although there were a few differences (additional file 2). Nitrite respiration/reduction/anoxygenic phototropy, oil bioremediation were found enriched in mesopelagic layer by all methods, while methanogenesis, a function that is usually associated with mesopelagic and deep sea microbes [8, 146–149] was not found enriched in MES by total sum. Both Wrench and Scran did not find xylanolysis to be enriched in the mesopelagic layer, while other methods did. We were unable to find literature evidence supporting this call, and the result could potentially be due to the higher number of OTUs called differentially abundant by the other methods. Aerobic ammonia/nitrite oxidation and fixation were found to be enriched in DCM by all methods. Total sum and TMM found a methanogenesis related module enriched in DCM, while other methods did not.

To evaluate the methods in a more fine-grained setting, we devised the following validation approach. The design of the Tara oceans experiments - where 16S reconstructions are obtained from whole metagenome shotgun sequencing data - makes the following analysis feasible. Because the Tara project's functional (gene content summarized as

Kegg Modules, KMs) and 16S data arise from the same input DNA samples, the same compositional factors should apply for both datatypes. We therefore estimated compositional factors from 16S data using the different normalization methods and applied the resulting estimates to the KM abundance data from the corresponding matched samples. Next, we computed Spearman rank correlation between OTU and KM normalized abundances and annotated OTUs with those KMs which showed correlation of at least 0.75. Finally, we identified OTUs that were positively associated with each layer using differential abundance analysis. With the KM annotations in place, we performed Fisher exact tests to compute the enrichment scores in the identified OTUs. In mesopelagic samples, Scran finds enrichment in only 30 KMs, while other methods recovered at least 100 KMs. Specifically, ureolysis, motility, several denitrification/methanogenesis processes and aminoacid biosynthetic/transport mechanisms (functions that have been attributed to microbes in the mesopelagic layer and deep sea) [8, 146, 150, 151], were missed by Scran, while Wrench finds them. On the other hand, Total sum, TMM and CSS found more varied and general processes including various ribosomal, transcription/translation components to be enriched in both MES and DCM layers.

Notice that the first analysis gives a broad sense of the genera identified by the competing methods in light of existing annotations, while the second gives a sense of the quality of annotations one might confer on the OTUs based on the normalized expression levels of OTUs and the measured functional content themselves. In both cases, Wrench is shown to retain relevant information, and the relatively more specific nature of the latter analysis reveals that Wrench demonstrably improves upon other methods.

4.3 Inferences following compositional correction show improved coherence with experimental data

We further demonstrate the impact of compositional bias in downstream inference below. The experimental cell density measurements in the Tara Oceans project show a highly significant overall reduction in the mesopelagic samples when compared the surface layer (see Fig. 3 in ref [8]). Thus, we expect an overall negative change in the reconstructed fold changes, when performing a differential abundance analysis of the OTUs across these two ocean layers.

Summing the log-fold changes of significantly associated OTUs (both positive and negative) serves as a measure of a net change experienced by a community. If a given method produces fold change inferences that track the above mentioned empirical cell density measurements, we expect it to yield an overall negative net change value for the significantly differentially abundant OTUs in the mesopelagic community. As illustrated in **Fig. 4.9A**, this value for total sum normalized data is $+10577.99$, while that for Wrench is -8919.65, showing that differential abundances arising from Wrench agrees more appropriately with the underlying community change. **Fig. 4.9B** and **C**, show how these values distribute across the major phyla focussed in the Tara oceans article. These plots demonstrate that the two approaches lead to markedly different conclusions on the net change experienced by a phylum. In particular, Proteobacteria, Actinobacteria, Euryarchaeota were predicted to have drastically high positive changes by total sum (while Wrench predicts a marked decrease in the negative direction), and sizable differences were apparent in the values obtained with the rest of the phyla.

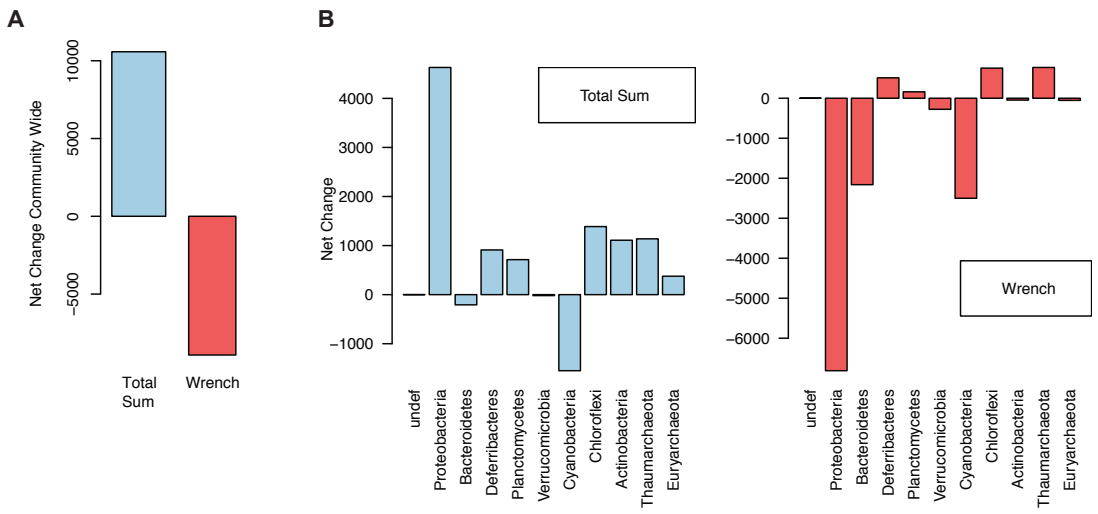

Figure 4.9: Wrench normalized data lead to better downstream inferences. *(A) The sum of log-fold changes of differentially abundant OTUs is used as a measure of net change experienced by a community. This value is plotted for the differentially abundant OTUs in the mesopelagic ocean layer relative to the surface layer in the Tara oceans 16S data, for Total Sum and Wrench normalization. (B) The same metric plotted for various major phyla of interest in the Tara oceans project.*

4.4 Compositional scale factor estimates imply substantial technical biases, indicating importance of further experimental studies

We next analyzed the phenotypic integrity of the compositional scales reconstructed by the various methods. In the absence of technical biases, following our discussion in the previous subsection, compositional factors should hover around 1 (upto some arbitrary scaling). This is *not* what we observe in samples from metagenomic datasets. All scale normalization techniques resulted in group-wise integrity in the scales they reconstructed within and across related phenotypic categories, potentially indicating the general importance of correcting for confounding induced by compositional bias in general practice. Total sum normalization is oblivious to these biases, making further experimental stud-

ies on compositional bias important. For instance, in the microbiome samples arising from the Human Microbiome Project [142], as shown in **Fig. 4.11A**, we noted systematic body site-specific global deviations in the fold change distributions. This is similar to what was illustrated with the Tara project in Fig. 4.1. We found the reconstructed compositional scales to largely organize by body sites, across normalization techniques (**Fig. 4.11B**), behind-ear and stool samples were distinctly located in terms of their compositional scales from the oral and vaginal microbiomes (notice the log scale in these plots). This behavior was also recapitulated in scales reconstructed from other centers. Similar results were obtained for samples arising from the J. Craig Venter Institute. In the case of the mouse microbiome samples, most normalization techniques predicted a mild change in differential feature content across the two diet groups (**Fig. 4.11C**, and). In the lung microbiome, the lung and oral cavities had roughly similar scales across smokers and non-smokers , while scales from the probing instruments had relatively higher variability, which we found to directly correlate with the high variability of feature presence in the count data arising from these samples. In the diarrheal datasets of children, however, no significant compositional differences were found across the various country/health-status populations (**Fig. 4.11D**).

For completeness, we also attach similar results from all the 11 organs of the rat body bulk RNASeq dataset in **Fig. 4.10**. We noted that the rat body map samples also showed systematic tissue-specific global deviations in the expressed features' fold change distribution. Fig. 4.10 shows this result and the general behavior of compositional scales across various methods compared and a few related statistics of the dataset. Given that these samples arise from a well designed series of experiments, the similarity in the scales

within and across related tissues, and across normalization methods, is striking; the observed trend in the reconstructed scales could indeed reflect underlying true compositional differences for the most part. TMM and CSS ascribe substantially deviated scales to muscle, heart and liver tissues, when compared to Scran and Wrench estimators. This effect may be due to the truncated estimation strategy which biases the scales for a relatively fewer but highly expressed genes in these tissues. Nevertheless, these results indicate potentially heavy compositional bias injected into downstream differential abundance analysis that compare tissues of different types. Compositional bias can be costly not only in metagenomics, but even in common bulk-RNAseq studies.

4.5 Methods

4.5.1 An approach (Wrench) for compositional correction of sparse, genomic count data

Briefly, our normalization strategy can be described as follows. Based on eqn. 2.2, for a chosen reference vector $q_{0\cdot}$, accounting for sample depth τ_{gj}, the mean model for the observed positive count of the i^{th} feature can be written as: $\log E[Y_{gji}|Y_{gji} > 0] = \log\left[q_{gji}\tau_{gj}\right] = \log\left[\frac{q_{gji}}{q_{0i}}q_{0i}\tau_{gj}\right] \equiv \log\left(\theta_{gji}q_{0i}\tau_{gj}\right)$, where $\theta_{gji} = \Lambda_{gj}^{-1}\nu_{gji}$. Thus the true ratio of proportions θ_{gji} encapsulate both the constant Λ_{gj}^{-1} and the concentration fold changes ν_{gji}, and can be viewed as the *net* fold change experienced by feature i in sample j from group g. For the purpose of metagenomic applications, and analytic convenience, we slighty relax the standard assumption that most features do not change across conditions by assuming that the feature-wise log-fold changes $\log\nu_{gji}$ arise independently from

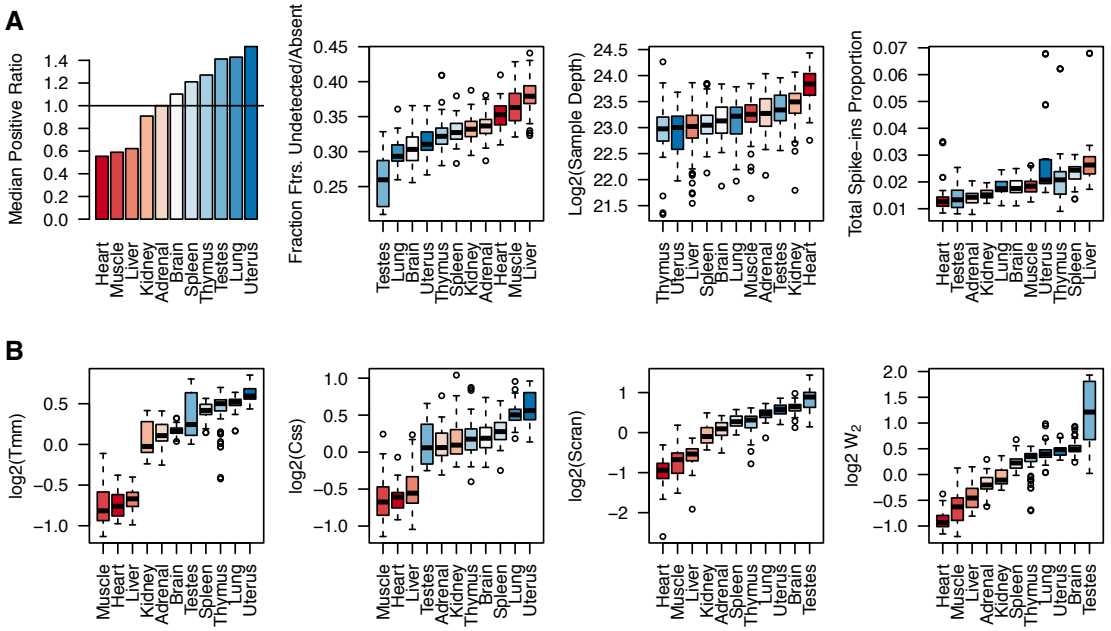

Figure 4.10: Importance of compositional correction in common bulk RNAseq studies. *(**A**) Application of scaling techniques to the rat body map data across tissues. Median positive ratio: median of the positive ratios of group-averaged proportions to that of Adrenal chosen as the reference. Subsequent figures in the top row indicate higher sparsity levels in the heart, muscle and liver samples, although at sequencing depths that are comparable/slightly higher to those from other tissue groups. (**B**) Reconstructed scales from several normalization techniques. If one were to perform a differential expression analysis between Testes and Heart, the fold changes are roughly 4X (ratio of medians) inflated as predicted by Scran/Wrench, which can lead to high false positive rates especially if most features are not changed across the two tissues. Notice the similarity in scales for closely related tissues, across techniques; for these tissues, the influence of compositional bias in the related differential abundance tests will be low.*

a zero mean Gaussian distribution, a common assumption in differential abundance analysis [5, 143, 144]. Assuming independence across features i, it then follows that $\log \theta_{gji}$ follows a Gaussian distribution with a mean parameter $\log \Lambda_{gj}^{-1}$. Thus, a robust location estimate of θ_{gji} for every sample leads us to the desired compositional scale estimate $\hat{\Lambda}_{gji}$. Below, we first illustrate how the θ_{gji} are estimated, and subsequently discuss the robust averaging procedure.

Model We assume the following hurdle log-normal model for the counts Y_{gji}:

$$
\begin{aligned}
Y_{gji} &\sim \begin{cases} 0 & \text{with probability } \pi_{gji} \\ e^{Z_{gji}} & \text{with probability } (1-\pi_{gji}) \end{cases}, \\
Z_{gji} &= \underbrace{\log q_{0i}}_{\text{log-reference}} + \underbrace{\log \tau_{gj}}_{\text{log-sample depth}} + \underbrace{\log \zeta_{0g} + \mu_{gj} + a_{gji}}_{=\log \theta_{gji},\ \text{log net fold change relative to reference}} + \varepsilon_{gji}, \\
a_{gji} &\sim N(0, \eta^2_{0g}),\ g = 1 \ldots G, \\
\varepsilon_{gji} &\sim N(0, \sigma^2_{0i}),\ i = 1 \ldots p, \\
\log\left(\frac{\pi_{gji}}{1-\pi_{gji}}\right) &= \beta_{i1} + \beta_{i2} \log \tau_{gj} + \text{possibly other covariates}
\end{aligned}
\tag{4.1}
$$

The model assumes the following. For each sample j from group g, the i^{th} feature's count value is sampled from a hurdle log-normal distribution, in which with probability π_{gji}, a value of 0 is realized; and with probability $1-\pi_{gji}$ a positive count is observed. The probabilities π_{gji} are determined by sample covariates, including the total sequencing depth. The positive count value is realized as an exponential of a Gaussian random variable Z_{gji} the mean of which is determined (in accordance with the eqn. 2.2) by the chosen reference value q_{0i}, sample-depth τ_{gj}, and the *net* fold change $\theta_{gji} = \nu_{gji} * \Lambda^{-1}_{gj}$, the log of which has been modeled in the above equation as a sum of group-wise effect ($\log \zeta_{0g}$), two-way group-sample interaction (μ_{gj}), a three-way group-sample-feature interaction random effect a_{gji} and a noise term.

Estimation of regularized ratios $\hat{\theta}_{gji}$: In the model, the 0 subscripted parameters are considered known, and are determined the following way. $\tau_{gj} = Y_{gj+}$ is the total count of sample gj. The reference value for each feature i, q_{0i}, is set to the average proportion value $\overline{\hat{q}_{++i}}$, where $\hat{q}_{gji}$ is the observed proportion of feature i in sample gj, i.e., $\hat{q}_{gji} = Y_{gji}/Y_{gj+} = Y_{gji}/\tau_{gj}$. The raw ratio of proportions are then given as: $r_{gji} = \frac{q_{gji}}{q_{0i}}$. The mean and variance parameters $\log \zeta_{0g}$ and η^2_{0g} of the Gaussian random effects distribution on the $\log \theta_{gji}$ are determined based on the moments of the corresponding empirical distribution of the group-wise pooled raw ratios of proportions. Specifically, we fix the group-wise compositional scale $\zeta_{0g} = \overline{r_{g+i}}$ i.e., as the average of the raw ratios including the zero values (following discussions in Fig. 4.4). We set the variance parameter $\eta^2_{0g} = \frac{1}{\sum_i I_{[Y_{gji}>0]}} \sum_{i:Y_{gji}>0} (\log r_{gji} - \overline{\log r_{g+i}})$ i.e., as the empirical variance of the logged-ratios. Finally, the feature-specific expression variances σ^2_{0i} are fixed with values obtained from Limma/Voom. With the above fixed, the unknown parameters μ_{gj} and a_{gji} are estimated/predicted using standard random effects estimators: $\hat{\mu}_{gj} = \sum_i w_{gji} \left(\log r_{gji} - \log \zeta_{0g}\right)$ with $w_{gji} \propto \frac{1}{\sigma^2_{0i}+\eta^2_{0g}}$, and $\hat{a}_{gji} = \frac{\sigma^2_{0i}}{\sigma^2_{0i}+\eta^2_{0g}} \left(\log r_{gji} - \log \zeta_{0g} - \hat{\mu}_{gj}\right)$. The identifiability of these terms is ensured as the other variance components are fixed. The $\hat{\pi}_{gji}$ are estimated with logistic regression. The regularized ratios are then calculated as: $\hat{\theta}_{gji} = exp(\log \zeta_{0g} + \hat{\mu}_{gj} + \hat{a}_{gji})$.

Robust averaging of the $\hat{\theta}_{gji}$: While averaging over the regularized ratios $W_0 =: \frac{1}{p}\sum_i \hat{\theta}_{gji}$ would be one estimation route to Λ^{-1}_{gj}, better control can be achieved by taking the variation in the feature-wise zero generation into account. We shall notice that $E[r_{gji}|r_{gji} > 0] = \theta_{gji} \cdot e^{\sigma^2_{0i}/2}$, and so a robust averaging over $\hat{\theta}_{gji}/e^{\sigma^2_{0i}/2}$, can serve as an

estimator of Λ_{gj}^{-1}. One might choose the weights for averaging to be proportional to that of the inverse hurdle/inclusion probabilities (as is done in survey analysis) $\propto 1/(1-\hat{\pi}_{gji})$ or on the inverse marginal variances ascribed by our model above $\propto \frac{1}{(1-\hat{\pi}_{gji})(\hat{\pi}_{gji}+e^{\sigma_{0i}^2+\eta_{0g}^2}-1)}$. An estimator that we also found to work well empirically is a weighted average of $\frac{\hat{\theta}_{gji}/e^{\sigma_{0i}^2/2}}{1-\hat{\pi}_{gji}}$ with weights proportional to $\frac{1}{\sigma_{0i}^2}$. The next subsection sketches the derivations for the weights.

An advantage of these weights (and hence the model) is that the weighting strategies proceed smoothly for features with zero expression values as well, unlike the binomial weights employed in the TMM procedure. Furthermore, when constructing averages, the weights have a favorable property of downweighting zeroes at higher sample depths relative to those in samples at lower sample depths.

In summary, we explored the performance of the following estimators for sample-wise compositional factors:

$$
\begin{aligned}
W_{0gj} &=: \frac{1}{p}\sum_i \hat{\theta}_{gji} = \overline{\hat{\theta}_{g+j}}, \\
W_{1gj} &=: \frac{1}{p}\sum_i w_{gji}\hat{\theta}_{gji}, \text{ with } w_{gji} \propto 1/(1-\hat{\pi}_{gji}) \\
W_{2gj} &=: \frac{1}{p}\sum_i w_{gji}\hat{\theta}_{gji}, \text{ with } w_{gji} \propto \frac{1}{\left(1-\hat{\pi}_{gji}\right)\left(\hat{\pi}_{gji}+e^{\sigma_{0i}^2+\eta_{0g}^2}-1\right)} \\
W_{3gj} &=: \frac{1}{p}\sum_i w_{gji}\frac{\hat{\theta}_{gji}}{1-\hat{\pi}_{gji}}, \text{ with } w_{gji} \propto \frac{1}{\sigma_{0i}^2}
\end{aligned}
\tag{4.2}
$$

The compositional bias corrected data is then obtained by dividing each sample's count data with its corresponding estimated compositional correction factor. For instance, if W_2 is the choice of estimator, the bias corrected data for sample gj is Y_{gj}/W_{2gj}.

We have found W_1, W_2 and W_3 to work comparably well in simulations and empirical comparisons, and W_0 slightly less so at high sparsity levels at low sample depths. We prefer W_2 as it systematically integrates both the hurdle and positive component variations. In our software implementation, users have the option for other weighted variants, and whether weighted averaging over zeroes is necessary as they see fit. Software documentation for Wrench embarks on further discussions on these ideas.

Derivation of marginal variance weights Setting $\phi_{0i} = e^{\sigma_{0i}^2/2}$, and $\gamma_{0g} = e^{\eta_{0g}^2/2}$, we have:

$$\begin{aligned} Var_\theta(E(Y_{gji}|\theta_{gji})) &= Var_\theta((1-\pi_{gji})\theta_{gji}\tau_{gj}q_{0i}\phi_{0i}) \\ &= \left((1-\pi_{gji})\tau_{gj}q_{0i}\phi_{0i}\right)^2 \underbrace{(\gamma_{0g}^2-1)\gamma_{0g}^2\zeta_{0g}^2}_{\text{group specific contribution}} \end{aligned} \tag{4.3}$$

Now, if we let Z to be an indicator random variable denoting whether a feature was zero or positive:

$$\begin{aligned} Var(Y_{gji}|\theta_{gji}) &= E_Z(Var(Y_{gji}|\theta_{gji},Z)) + Var_Z(E(Y_{gji}|\theta_{gji},Z)) \\ &= (1-\pi_{gji})\left(\theta_{gji}\tau_{gj}\phi_{0i}q_{0i}\right)^2\left[\pi_{gji}+(\phi_{0i}^2-1)\right] \end{aligned} \tag{4.4}$$

Similarly,

$$\begin{aligned} E(\theta_{gji}^2) &= Var(\theta_{gji}) + E(\theta_{gji}^2) \\ &= \left(\gamma_{0g}^2-1\right)\gamma_{0g}^2\zeta_{0g}^2 + \left(\zeta_{0g}\gamma_{0g}\right)^2 \\ &= \left(\zeta_{0g}\gamma_{0g}\right)^2\gamma_{0g}^2 \end{aligned} \tag{4.5}$$

Together, eqns. 4.3 and 4.4 lead to:

$$E(Var(Y_{gji}|\theta_{gji})) = (1-\pi_{gji})\left[\pi_{gji} + (\phi_{0i}^2 - 1)\right](q_{0i}\tau_{gj}\phi_{0i})^2\left(\gamma_{0g}^2\zeta_{0g}\right)^2 \tag{4.6}$$

Eqns. 4.3 and 4.6 then imply:

$$\begin{aligned} Var(Y_{gji}) &= (1-\pi_{gji})(q_{0i}\tau_{gj}\phi_{0i})^2\left[\pi_{gji} + \phi_{0i}^2\gamma_{0g}^2 - 1\right]\gamma_{0_g}^2\zeta_{0g}^2 \\ &\propto (1-\pi_{gji})(q_{0i}\tau_{gj}\phi_{0i})^2\left[\pi_{gji} + \phi_{0i}^2\gamma_{0g}^2 - 1\right] \end{aligned} \tag{4.7}$$

The variances for the adjusted ratios then follows from straightforward calculations, the inverse of which take the weight forms shown in in the previous subsection.

Data We principally demonstrate our results with five datasets from metagenomic surveys. A smoking study ($n = 72$) where the lung microbiome of smokers and non-smokers were surveyed (along with the instruments that were used to sample the individual). A diet study in which the gut microbiomes ($n = 139$) of carefully controlled laboratory mice fed plant-based or western diets were sequenced [152]. A large scale study of human gut microbiomes ($n = 992$) from diarrhea-afflicted and healthy children from various developing countries [139]. 16S metagenomic count data corresponding to all these studies were obtained from the R/Bioconductor package `metagenomeSeq` [5]. The Tara Oceans project's 16S reconstructions from whole metagenome shotgun sequencing ($n = 139$) deposited in http://ocean-microbiome.embl.de/data/ was obtained from file miTAG.taxonomic.profiles.release.tsv.gz. The flow cytometry counts for autotrophs, bacteria, heterotrophs, picoeukaryotes were obtained from TaraSampleInfo_OM.CompanionTables.txt

from the same website and summed to serve as a rough measure of total cell count that correlates with sequence-able DNA material. The Human Microbiome Project count data in file HMQCP/otu_table_psn_v35.txt.gz was downloaded from http://downloads.hmpdacc.org/data/, and the associated metadata are from v35_map_uniquebyPSN.txt.bz2 under the same website.

The processed bulk-RNAseq data corresponding to the rat body map from [7] was obtained from [153].

The Unique Molecular Identifier (UMI) single cell RNAseq data from Islam et al., [62] was downladed from GEO under accession GSE46980.

Implementation of normalization and differential abundance techniques All analysis and computations were implemented with the R 3.3.0 statistical platform. EdgeR's `compNormFactors` for TMM, DESeq's `estimateSizeFactors`, Scran's `computeSumFactors` (with `positive=TRUE` in sparse datasets) and `metagenomeSeq`'s `calcNormFactors` for CSS were used to compute the respective scales. Implementation of CLR factors used a pseudo-count of 1 following [140], and were computed as the denominator of column 3 in table 3.1. Limma's eBayes in combination with lmFit, edgeR's `estimateDisp`, `glmFit` and `glmLRT`, DESeq2's `estimateDispersionsGeneEst` and `nbinomLRT` were used to perform differential abundance testing [144]. Welch's t-test results were obtained with `t.test`.

Simulations Given a set of control proportions q_{1i} for features $i = 1p$, and the fraction of features that are perturbed across the two conditions f, we sample the set of true

log fold changes ($\log \nu_{gi}$) from a fold change distribution (fold change distribution) for those randomly chosen features that do change. The fold change distribution is a two-parameter distribution chosen either as a two-parameter Uniform or a Gaussian. Based on the expressions from the first subsection of the results section, the target proportions were then obtained as $q_{gi} = \frac{\nu_{gi} q_{1i}}{\sum_k \nu_{gk} q_{1k}}$. Conditioned on the total number of sequencing reads τ, the sequencing output $Y_{gi\cdot}$ for all i were obtained as a multinomial with proportions vector $q_{g\cdot} = [|q_{gi}|]_{i=1}^{p}$. We set the control proportions from various experimental datasets (specifically, mouse, lung and the diarrheal microbiomes). With this setup, we can vary f, and the two parameters of the fold change distribution, and ask, how various normalization and testing procedures compare in terms of their performance. For bulk RNAseq data, as illustrated in the previous chapter, we simulated $20M$ reads per sample.

For comparison of Wrench scales with other normalization approaches, we altered the above procedure slightly to allow for variations in internal abundances of features in observations arising from a group g. We used $\overline{\nu_{gi}}$ (where the bar indicates this value will now assume the role of an average) generated above as a prior fold change for observation-wise fold change generation. That is, for all samples $j \in 1 \ldots n_g$ for all g, where n_g represents the number of samples in group g, for all i (including the truly null features), sample ν_{gji} from $LN(\log \overline{\nu_{gi}}, \tilde{\sigma}_\nu^2)$ for a small value of $\tilde{\sigma}_\nu^2 = .01$. This induces sample specific variations in the proportions within groups. Notice that this makes the problem harder and more realistic, as feature marginal count distributions now arise from a mixture of distributions. Based on empirically observed MA plots for our metagenomic datasets, we set the mean and standard deviation of prior log-fold change distribution to 0 and 3 respectively. For generating 16S metagenomic-like datasets, logged sample depths were

sampled from a log-normal distribution with logged-standard deviation of .25 and logged-means corresponding to $\log(4K)$, $\log(10K)$ and $\log(100K)$ reads. These parameters were chosen based on comparisons with M-A plots, the sparsity levels and total sample depths observed in current experimental datasets. We repeated simulations for 20 iterations.

In both versions of simulations, the total induced abundance change relative to that of the control is $\Lambda_{gj} = \nu_{gj\cdot}^T q_{1\cdot}$, where $\nu_{gj\cdot}$ is the vector of fold changes for sample j in group g, and $q_{1\cdot}$ is the average vector of feature-wise control proportions. As it can be seen from the expression for Λ_{gj}, notice that perturbing features with very low relative frequencies do not demonstrably induce compositional bias at low sample depth settings (unless perturbed by very high fold changes). So for every simulation iteration, the fraction f of features that were perturbed in cases were chosen randomly according to their control proportions. We apply the term *compositional correction factor* for Λ_{gj}^{-1} and the term *normalization factor* for a sample as the product of its compositional correction factor with something that is proportional to that of its sample depth. Thus, all technical artifacts like total abundance changes, but sample depth, are incorporated into the definition of compositional factors.

Performance comparisons For simulations, we used edgeR as the workhorse fitting toolkit. The compositional scale factors provided by all normalization methods were provided to edgeR as offset factors. We define detectable differential abundance in our simulated count data as follows. For each simulation, as we know the true compositional factors, we input them as normalization factors in edgeR, and the detectable differences in abundances are recorded. All the performance metrics are then defined based on this

ground truth. Because we are interested in fold changes and their directions, the performance metrics we report are redefined as follows: Sensitivity as the ratio of the number of detectable true-positives with true sign over the total number of positives, False discovery as the ratio of the number of detectable true positives with false sign and false positives, over the total number of significant calls made.

The offset-covariate analysis followed the procedure in [6]. For resampling analysis, samples from each experimental group (with atleast 15 samples) were split in half randomly to construct two artificial groups. Normalization factors from each method were then used to perform differential abundance analysis, and the total number of differentially abundant calls were recorded. The procedure was repeated for ten iterations for each group, and the results were averaged across 41 experimental groups. Those samples for which Scran fails to reconstruct normalization scales were discarded from differential abundance analyses to avoid any power differences while testing. The normalization scales however, were obtained with all data for each method.

Fisher exact tests were used to perform functional enrichment analyses for positively associated OTUs. A Genus level functional enrichment analysis was first performed by aggregating annotations from the FAPROTAX1.1 database [146] at the Genus level. A more specific OTU level functional enrichment analysis was devised as follows. Because the Tara Oceans Kegg module (KM) abundance data (downloaded from TARA243.KO-module.profile.release.gz, under http://ocean-microbiome.embl.de/data/) and the 16S reconstructions are obtained from the same input DNA through whole metagenome shotgun, the same compositional factors apply to both datatypes. Each normalization approach's compositional factors for 16S data was used to rescale the KM relative frequen-

cies. This normalized KM data was used to annotate each OTU by (normalized) KMs that Spearman correlate at a value of atleast .75.

4.6 Discussions

For some researchers, statistical inference of differential abundance is a question of differences in relative frequencies; for others, it is a matter of characterizing differences in absolute abundances/concentrations of features expressed in samples across conditions [54, 154]. In this work, we took the latter view and aimed to characterize the compositional bias injected by sequencing technology on downstream statistical inference of concentrations of genomic features.

It is clear that the probability of sequencing a particular feature (ex: mRNA from a given gene or 16S RNA of an unknown microbe) in a sample of interest is not just a function of its own fold change relative to another sample, but inextricably linked to the fold changes of the other features present in the sample in a systematic, statistically non-identifiable manner. Irrevocably, this translates to severely confounding the fold change estimate and the inference thereof resulting from generalized linear models. Because the onus for correcting for compositional bias is transferred to the normalization and testing procedures, we reviewed existing spike-in protocols from the perspective of compositional correction, and analyzed several widely used normalization approaches and differential abundance analysis tools in the context of reasonable simulation settings. In doing so, we also identified problems associated with existing techniques in their applicability to sparse genomic count data like that arising from metagenomics and single cell RNAseq,

which lead us to develop a reference based compositional correction tool (Wrench) to achieve the same. Wrench can be broadly viewed as a generalization of TMM [52] for zero-inflated data. We showed that this procedure, by modeling feature-wise zero generation, reduces the estimation bias associated with other normalization procedures like TMM/CSS/DESeq that ignore zeroes while computing normalization scales. In addition, by recovering appropriate normalization scales for samples even where current state of the art techniques fail, the method avoids data wastage and potential loss of power during differential expression and other downstream analyses

Some practically relevant notes on the application of proposed method to metagenomic datasets follow. First, our choice of methodology and simplifying assumptions were principally determined by the scale and sparsity of the 16s metagenomic datasets and estimation robustness. While fully joint parameter inference algorithms will certainly be more accurate, they can be unwieldy and computationally intensive with large scale datasets boasting a large number of features with high sparsity. A case in point is the neat GAMLSS methodology [9], which improved over the proposed pipeline (Wrench normalization coupled with edgeR differential abundance analysis) in a small scale equimolar miRNA benchmarking dataset, but could not run to completion in the simplest of our metagenomic datasets, the mouse gut microbiome. In **Fig. 4.12**, we present the same benchmarking analysis as in Fig. 7 of Argyropolous et al., [9] for DeSeq2, GAMLSS, Wrench normalization + EdgeR and Scran normalizaiton + EdgeR pipelines for differential abundance.

Second, our simulation results indicate that the performance of Wrench stabilizes by $10-20$ samples per group depending on sample depth and the fraction of features that

change across conditions. While in our experience, this is very well within the limits of practically realized sample sizes in metagenomic experiments, at very low sample sizes and very low sample depths (less than a few thousand reads per sample), some care might be necessary. For instance, coherence of the reconstructed sample-wise compositional scales within groups relative to the experimental design can be checked and deviations from expectations analyzed/corrected. Third, our current implementation exploits categorical group information/factors alone (e.g., cases and controls), and extension to continuous covariates (e.g., age, time) underlying the sampling design are planned for future work. If a continuos covariate is present, converting it to factors by discretizing its range in to non-overlapping windows is an option that the analyst can entertain. Furthermore, because group information is exploited during normalization, our proposed methodology is not immediately applicable for classification purposes. In such applications, immediate extensions of the proposed empirical Bayes formalism by assuming priors on the unknown-sample's group membership (based vaguely, for example, on clustering distances) can be done, and is planned for future work.

A few important insights on compositional bias emerge from our theory, simulation and experimental data analyses. In our simulations, we found reference based normalization approaches to be far superior in correcting for sequencing technology-induced compositional bias than library size based approaches. From a more practically relevant perspective, we found that in all the tissues from the rat body map bulk RNAseq dataset, the scale factors can be robustly identified. We expect that in other bulk RNAseq datasets, the assumptions underlying compositional correction techniques to hold well. These results reinforce trust in exploiting such scaling practices for other downstream analyses of

sequencing count data apart from differential abundance analysis; for example, in estimating pairwise feature correlations. In the regimes where assumptions underlying these techniques are met, an analyst need not be restricted to scientific questions pertaining to relative frequencies alone. The fundamental assumption behind all the aforementioned techniques is that most features do not change across conditions (or the closely related assumption that the log-fold change distribution is centered at 0). As we illustrated, these assumptions appear to hold rather well in bulk RNAseq. Do we expect these to hold in arbitrary microbiome datasets as well? This question is not easy to address without more experiments, but the relatively high correlations obtained with orthogonal measurements of technical biases, the similarity in the compositional scales obtained within samples arising from biological groups, and their sometimes highly significant shifts preserved across normalization techniques and across sequencing centers in large scale studies certainly reinforce the critical importance of characterizing compositional biases, if any, in metagenomic analyses by establishing carefully designed spike-in protocols. In particular, given the inverse dependence of compositional correction factors on the total feature content in the absence of technical biases, the large compositional scale estimates obtained for stool samples (across all normalization techniques) is suspect. Compositional effects can amplify even when a few features experience adverse technical perturbations, and only carefully designed experiments can isolate these effects to inform further normalization approaches. Finally, our results also emphasize the tremendous care one needs to exercise before applying the most natural normalizations based on total sequencing depth or by applying pseudocounts when the data is excessively sparse (CLR, RPKM, CPM, rarefication are a few examples).

This brings us to the question of how effective spike-in strategies are in enabling us to overcome compositional bias. It is immediately clear that the widely used ERCC recommended spike-in procedure for RNAseq cannot help us in overcoming confounded inference due to compositional bias for the simple reason that it already starts with an extract, a compositional data source. If one is able to add the spike-in quantities at a prior stage during feature extraction, we would have some hope. Lovén et al., [113] demonstrate a procedure for RNAseq that precisely does this, in which the spike-ins are added at the time when the cells are lysed and suspended in solution [114]. One can perhaps extend these solutions to metagenomics, where we may expect confounding due to compositionality to be heavy by adding barcoded 16S RNAs during feature extraction. We expect similar problems to arise in other genomic and epigenetic measurement techniques that exploit sequencing technology, and the need for the development of appropriate spike-in procedures should be addressed.

Finally, it is imperative that we enforce new tools and techniques for normalization and differential abundance analysis of sequencing count data be benchmarked for compositional bias at least in the simulation pipelines. Data analyses based on large-scale integrations of different data types for predicting clinical phenotypes is increasingly common, and care should be taken to include effective normalization techniques to overcome compositional bias. We hope the results and ideas presented and summarized in this work enables a researcher to do just that.

4.7 Conclusions

Compositional bias, a linear technical bias, underlying sequencing count data is induced by the sequencing machine. It makes the observed counts reflect relative and not absolute abundances. Normalization based on library size/subsampling techniques cannot resolve this or any other practically relevant technical biases that are uncorrelated with total library size. Reference based techniques developed for normalizing genomic count data thus far, can be viewed to overcome such linear technical biases under reasonable assumptions. However, high resolution surveys like 16S metagenomics are largely undersampled and lead to count data that are filled with zeroes, making existing reference based techniques, with or without pseudocounts, result in biased normalization. This warrants the development of normalization techniques that are robust to heavy sparsity. We have proposed a reference based normalization technique (Wrench) that estimates the overall influence of linear technical biases with significantly improved accuracies by sharing information across samples arising from the same experimental group, and by exploiting statistics based on occurrence and variability of features. Such ideas can also be exploited in projects that integrate data from diverse sources. Results obtained with our and other techniques, suggest that substantial compositional differences can arise in (meta)genomic experiments. Detailed experimental studies that specifically address the influence of compositional bias and other technical sources of variation in metagenomics are needed, and must be encouraged.

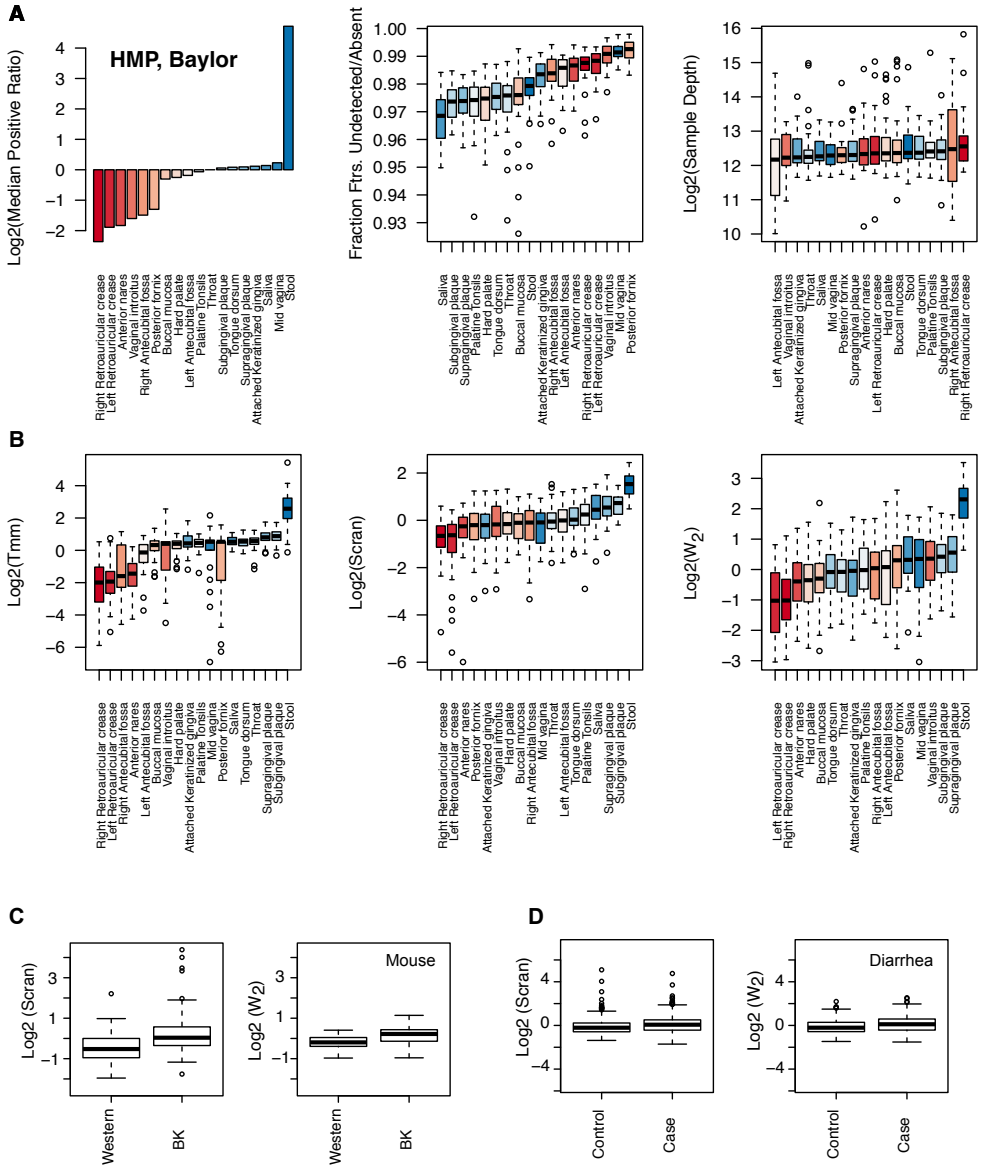

Figure 4.11: Wrench retains potential biological information, and indicates importance of compositional correction in general practice. *We plot some statistical summaries and the compositional scale factors reconstructed by a few techniques for various Human Microbiome Project samples, sequenced at the Baylor College of Medicine. (**A**) On the top-left, we plot the logged median of the positive ratios of group-averaged proportions to that of* Throat *chosen as the reference group. Stool samples show considerable deviation from the rest of the samples despite having comparable fraction of features detected and sample depths to other body sites. Notice the log scale. (**B**) The similarity in the reconstructed scales across techniques (second row) for closely related body sites are striking; although minor variations in the relative placements were observed across centers potentially due to technical sources of variation, the overall behavior of highly significant differences in the scales of behind-ear and stool samples were similar across sequencing centers and normalization methods.* *Continued on next page.*

Figure 4.11: *Continued from previous page. These techniques predict a roughly 4X-8X (ratio of medians)inflation in the Log2-fold changes when comparing abundances across these two body sites. (**C**) Wrench and scran compositional scale factors across the plant-based diet (BK) and Western diet (Western) mice gut microbiome samples. (**D**) Compositional scale factors for healthy (Control) and diarrhea afflicted (Case) children. Slight differences in the compositional scales are predicted in the diet comparisons with t-test p-values < 1e-3 for all methods except TMM, but not as much in the diarrheal samples.*

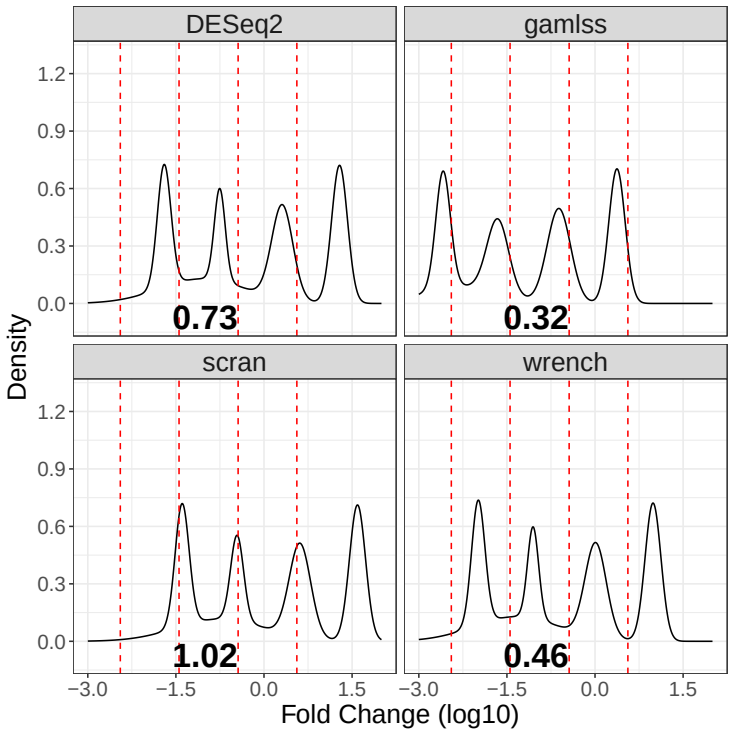

Figure 4.12: ***Benchmarking analysis of the small scale, high coverage Argyropolous et al., miRNA dataset for deviatioin from expected fold changes in the clustered symmetric DE without global changes in expression ratiometric A versus B.*** *Same as Fig. 7 in [9]. The shown numbers measure deviation of the reconstructed fold changes from the true expected fold changes by experimental design, for the pipeline. Lower is better. Refer [9], Fig. 7 for details on experimental design. The data was downloaded from authors' repository: https://bitbucket.org/chrisarg/rnaseqgamlss.*

Part II

Adaptive immunity in prokaryotes

Chapter 5

The curious case of prokaryotic adaptive immunity.

In vertebrates, adaptive immunity against infectious agents is special [155]. It provides hosts with immediate adaptations to counteract infections, over ecological time scales. This is in direct contrast to any innate defense novelties that may arise through natural selection over much longer evolutionary time scales. To fully appreciate the utility of adaptive immunity, one only needs to look back to the pathological complications of *Measles morbillivirus* or *Varicella zoster* viral infections, which respectively cause measles [156, 157] and chickenpox [158]. These examples are often quoted and perhaps, the most relatable.

In general with higher organisms, adaptive immunity operates in several steps. The first phase involves a rapid combinatorial production and selection of specialized cells that synthesize proteins (antibodies) with high specificity to bind an appropriately processed target infectious agent. The resulting stably bound complexes activate a series of host innate responses, ultimately clearing host infection. Perhaps owing to the complex nature of the system and the variety of components involved [155, 159, 160], until about a decade ago, it was generally thought only higher organisms like vertebrates entertain adaptive immunity. As far as prokaryotes were concerned, two distinct classes of rel-

atively rudimentary immune systems were known to operate, especially against viruses (phages) that invade them. The first makes the host resistant to phage infection. Examples of this type include mutations in the cell surface receptors that prevent phage adsorption (envelope resistance), and the variety of restriction-modification enzymes that recognize and cleave the intracellular phage DNA introduced during a phage infection. The second class induce altruistic cellular suicide of infected hosts, thus limiting the spread of infections to other concomitant hosts. Toxin/anti-Toxin and abortive infection (Abi) systems are examples of the latter.

In the late 2000s, a very simple yet highly effective adaptive immune system was documented in natural prokaryotic populations. Owing to their genomic architecture in bacterial DNA, the system was named CRISPR – clustered regularly interspersed palindromic repeats [161]. It is instructive to summarize a decade long story behind this discovery [162, 163]. The history also serves to emphasize the pivotal role sequencing and bioinformatics played in generating effective, testable hypotheses. The first steps towards the identification of CRISPRs were laid down by a series of papers that documented the presence of roughly equally spaced repeats, interleaved with some spacer DNA in the genomes of a few archaea and bacteria [164–171]. Continued public funding allowed the growth and maintenance of sequencing databases, along with effective algorithms to search them. These tools revealed that the aforementioned spacer DNA that interleave the repeat segments had extensive similarities to seemingly random segments of the genomic DNA of invading phages, among other things [161, 172–175]. Putative genes and promoters upstream of the CRISPR locus were also identified, and relating these findings back to the nobel prize winning RNA interference mechanism of gene expression

inhibition, several authors speculated that spacer DNA when expressed could function as anti-sense RNAs [176, 177]; when bound to their complementary phage targets, they could induce an RNA interference like pathway resulting ultimately in the destruction of target. By 2010, these speculations turned to biological facts through careful experiments [178–183]. But what makes CRISPRs *adaptive*? Sequencing CRISPR loci from laboratory cultures of CRISPR hosts before and after coevolution with invading phages showed that CRISPR spacer DNA reflect new segments in phage DNA over ecological time scales [178, 180, 184]. Furthermore, host resistance was correlated with the fraction of spacers that match segments in the invading phage genomic DNA [178, 180]. Thus, a very fundamental and highly significant advancement in microbiology was made in the last decade: prokaryotic CRISPR is an effective *adaptive* resistance mechanism.

Our interest in CRISPRs is piqued by another set of observations. With powerful experiments and bioinformatics analyses, various authors have demonstrated that autoimmunity is a major side-effect experienced by CRISPR hosts [185–188]. This is caused because the CRISPR machinery could, with very high error rates, exploit DNA segments derived from the host genome itself as spacer DNA. As a result, CRISPRs are configured to consider host DNA as foreign, adversely affecting host health. We analyze the influence of autoimmunity in CRISPR mediated prokaryote-phage coevolution over ecological time scales, and discuss some evolutionary implications. This is the subject of chapter 6.

Chapter 6

Ecological dynamics of autoimmune CRISPR induced prokaryote-phage coevolution.

Prokaryotes have evolved diverse molecular defense systems over billions of years of co-evolution with phages [189, 190]. Clustered Regularly Interspersed Palindromic Repeats (CRISPRs), found in roughly 40% of sequenced bacteria and 90% of archaea, are peculiar in that they confer adaptive immunity against invading phages [183, 191–194]. CRISPR, as a defense mechanism, works via targeted acquisition of 26-72bp fragments (called protospacers) from the target DNA, and subsequently use of acquired fragments (spacers) for target restriction through an RNAi-like mechanism [176, 178]. Acquisition events appear to concentrate around short 2 – 5bp motifs (protospacer adjacent motifs, or PAMs) in the target DNA [180, 183, 195]. CRISPR loci are organized as cassettes in which short repeats interleave spacers, and are located adjacent to highly diverse genes that code for the CRISPR associated protein machinery [183] [187].

Intriguingly, in addition to acquiring phage fragments, CRISPR systems can also acquire spacers from the host genome. This has been experimentally demonstrated in two model systems: first, selective induction of the acquisition machinery (in the ab-

sence of interference) in laboratory strains of Escherichia coli resulted in the accumulation of a large number of self-targeting spacers [185]; second, abolition of interference activity (and not the acquisition machinery) in wild type *Streptococcus thermophilus* resulted in unbiased acquisitions of self-targeting spacers alongside phage-targeting spacers [186]. However, a large-scale survey of CRISPR cassettes in microbial genomes identified that only about 0.4% of the spacers are self-targeting, which, considering the relative size of prokaryotic genomes over phages, suggests some mechanism of selection against self-targeting spacers, perhaps to avoid autoimmunity [186, 187, 196]. Indeed, directed experiments have conclusively shown that self-targeting can result in severe lethality [180, 197–201].

We therefore face a conundrum: how do prokaryotes maintain functional CRISPR systems [202]? Despite the conceptual similarities with restriction-modification systems that avoids autoimmunity by methylating the host genomes' target restriction sites [203], no analogous genome wide self- vs. non-self-discrimination (SND) mechanism is known for CRISPR systems. In fact, as noted above, the evidence thus far suggests that an efficient SND may not exist (The SND mechanism described by Marrafini and Sontheimer explains the evasion of self-destruction of CRISPR locus only and does not confer genome wide protection [204]). But there are other routes to avoiding autoimmunity. Toxin/antitoxin or abortive infection systems restrict the scope of autoimmunity to infected populations via infection-induced activation [205]. Indeed, upregulation of CRISPRs upon phage infection has been demonstrated experimentally [206–208]. This makes it possible that the accumulated self-targeting spacers may function as "toxins", which can be activated upon infection. We therefore address the following two questions in this study:

1. Does infection-induced activation allow CRISPRs to function as an abortive infection (ABI) system? If so, what is the relative contribution of ABI in determining coevolving host and phage densities?

2. If CRISPR suppression in uninfected host populations is required to avoid host extinction, how strong should this suppression be?

Clearly, the answers to these questions depend on key ecological and CRISPR kinetic parameters. For instance, while CRISPRs are highly active against phages in wild type *S. thermophilus* (a lactic acid bacteria widely used in industrial production of cheese) [180], artificial induction is essential to activate the system in *E.coli* [209]. To this end, we develop and analyze a dynamical model that integrates prokaryote-phage coevolutionary dynamics, with regulated, infection-induced CRISPR acquisition and interference activity. Several models of CRISPR-mediated prokaryote-phage coevolutionary dynamics have been previously reported [1, 2, 200, 210–213]. While refs. [211–213] account for an abstract CRISPR-associated cost, they do not include the specifics of autoimmunity kinetics/the regulatory aspect of CRISPRs. The model we develop here is detailed enough to incorporate the adaptive aspects of CRISPR, and general enough to allow intuitive (analytic) interpretations of the resulting qualitatively distinct steady states. We interrogate the model using simulations and bifurcation analyses, and we find that as a function of key host, ecological, and CRISPR evolutionary parameters, the operational behavior of CRISPRs (and the resulting host densities) decomposes into four qualitatively distinct regimes. In those regimes where CRISPR is advantageous to the host, both restriction and abortive infection operate; the latter dominates restriction in SND absence.

Crucially, CRISPR maintenance is determined by an upper bound on the activation level of CRISPRs in uninfected populations. This critical limit of activation – beyond which host extinction is inevitable – is determined by a simple dimensionless combination of parameters. We compare the current experimental data on CRISPR kinetics with these qualitative observations, which helps to explain the spacer deletion mechanism and absence of CRISPR activity in highly virulent and multi-drug resistant clinical isolates.

6.1 Results

6.1.1 Behavior of a simple prokaryotic immune system with regulated autoimmunity

Before proceeding to model the complexity of CRISPR dynamics in general, we start by considering the case of a simple prokaryotic immune system with regulated autoimmunity. The goal here is to analyze the influences of the regulation, immunity and autoimmunity on the resulting coevolutionary dynamics.

Fig 6.1 illustrates a simple coevolutionary model in which the immune system, apart from conferring immunity, also induces autoimmunity that is regulated in a cell state (infected / uninfected) specific manner. Dynamic variables are denoted with Roman letters, and parameters are denoted with Greek symbols. Any parameter associated with production of an item i is denoted as α_i and that with its degradation is denoted by γ_i. Free cells (p), grow exponentially at a rate of α_p, under a carrying capacity constraint of Φ_p. Phages (v) infect free cells to produce infected cells at a rate of α_q. Infected cells can lyse to release phages at a rate of $\gamma_{q\to v}$ or undergo immunity to become a free cell at a rate

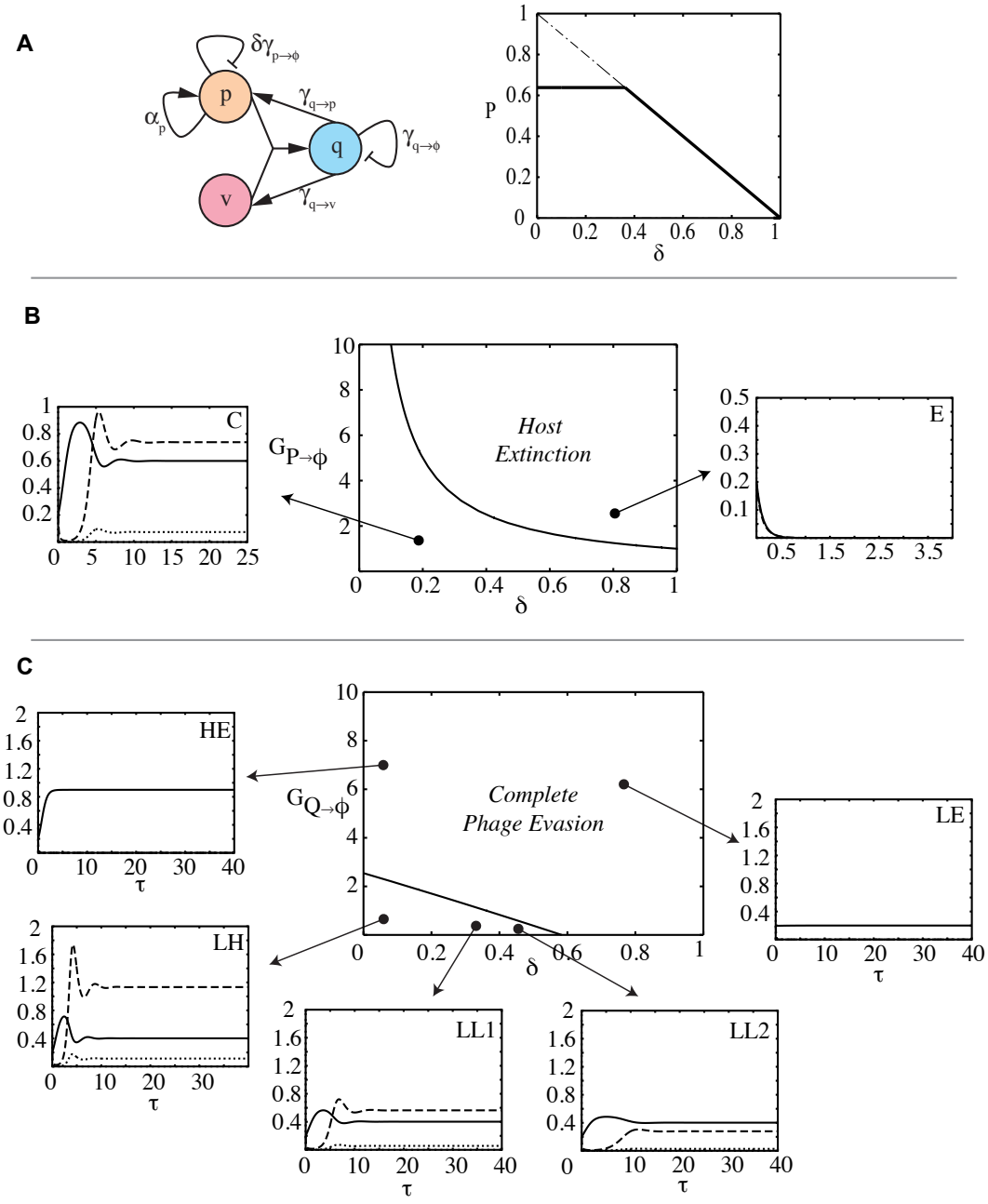

Figure 6.1: Bifurcation analysis of a simple model of a prokaryotic immune system with regulated autoimmunity side effect. *(A) p, q and v denote densities of uninfected, infected cells, and phage respectively. q undergoes autoimmunity at a rate of $\gamma_{q\to\phi}$, while p undergoes autoimmunity at a suppressed rate determined by $\delta\gamma_{p\to\phi}$. The second figure shows the bifurcation behavior of the free cell densities with respect to the control parameter δ, beyond a certain critical value of which one of the steady states vanishes.* *Continued on next page.*

Figure 6.1: *Continued from previous page. (B,C) Two-parameter bifurcation diagram revealing coexistence (C) and host extinction (E). Each plot instance is denoted by a tuple <AB> where A and B can indicate low (L) or high (H) values or extinction (E) of prokaryote (free–solid, infected–dotted) and phage (dashed) respectively.* $G_{P\to\phi}$ *and* $G_{Q\to\phi}$ *denote the rescaled free cell autoimmunity rate and infected cell autoimmunity (abortive infection) rates respectively. High values of* $G_{Q\to\phi}$ *lead to complete phage evasion. Parameter values* $\alpha_p = 1\ hr^{-1}$, $\Phi = 10^8\ cells\ ml^{-1}$, $\gamma_v = 5\ hr^{-1}$, $\alpha_v = 50$, $\alpha_q = 5\times10^{-9}\ ml\ phage^{-1}\ hr^{-1}$.

Variable	Description	Value, Units
p,q	Cell densities	$cells\ ml^{-1}$
v	Phage density	$phages\ ml^{-1}$
α_p	Free cell replication rate	hr^{-1}
α_q	Phage adsorption rate	$ml\ phage^{-1}\ hr^{-1}$
$\gamma_{p\to\phi}$, $\gamma_{q\to\phi}$, $\gamma_{q\to p}$, $\gamma_{q\to v}$	Autoimmunity, Immunity, and Lysis rates	hr^{-1}
γ_v	Phage death rate	hr^{-1}
α_v	Phage burst size	$phages$
Φ_p	Environmental carrying capacity	$cells\ ml^{-1}$
δ	Scale factor ($0 \le \delta \le 1$), determines CRISPR activity in free cells.	
μ_v	Phage mutation rate per protospacer	$protospacers^{-1}$

Table 6.1: Descriptions of variables and parameters in model 1. *Dynamic variables are denoted with Roman letters, and parameters are denoted with Greek symbols. Any parameter associated with production of an item i is denoted as* α_i *and that with its degradation is denoted as* γ_i*. Steady state value of an item i will be donted by* i^*.

of $\gamma_{q\to p}$, or undergo autoimmunity at a rate of $\gamma_{q\to\phi}$. Free cells undergo autoimmunity at a suppressed rate of $\delta\gamma_{p\to\phi}$, $(0 \le \delta \le 1)$. Note $\gamma_{p\to\phi}$ need not necessarily equal $\gamma_{q\to\phi}$, for reasons that will become clear later when we discuss the detailed CRISPR model. The condition $\delta = 0$ implies complete repression of autoimmunity in free cells, whereas $\delta = 1$ indicates no difference in repression across the two cell states. The burst size of phages is α_v. Phages also die at a rate of γ_v. **Table 6.1** describes the variables and model parameters.

The dynamical equations for this model can be written as:

$$\begin{aligned}
\dot{p} &= \alpha_p p\left(1-\frac{p+q}{\Phi_p}\right) - \delta\gamma_{p\to\phi}p - \alpha_q pv + \gamma_{q\to p}q \\
\dot{q} &= \alpha_q pv - (\gamma_{q\to\phi} + \gamma_{q\to p} + \gamma_{q\to v})q \\
\dot{v} &= \alpha_v\gamma_{q\to v}q - \gamma_v v - \alpha_q pv
\end{aligned} \tag{6.1}$$

Measuring all the state variables in units of Φ_p, and time in units of $\tau = \left[\alpha_q\Phi_p\right]^{-1}t$, and denoting all the transformed variables and parameters with their corresponding Roman alphabets, we obtain:

$$\begin{aligned}
\dot{P} &= A_p p\,(1-P-Q) - \delta G_{P\to\phi}P - PV + G_{Q\to P}Q \\
\dot{Q} &= pv - (G_{Q\to\phi} + G_{Q\to P} + G_{Q\to V})Q \\
\dot{V} &= \alpha_v G_{Q\to V}Q - G_V V - PV
\end{aligned} \tag{6.2}$$

We can study the influence of regulation (determined by the parameter δ), immunity and autoimmunity rates ($G_{Q\to V}, G_{Q\to\phi}$, and $G_{P\to\phi}$) on the above dynamical system using a bifurcation analysis. These results are summarized in **Fig. 6.1**. Fig. 6.1A shows that, as a function of δ, two fixed points collide at a critical value of δ (which we denote by δ_1), beyond which one of them ceases to exist. Fig. 6.1B shows that in the $(\delta, G_{P\to\phi})$ space, beyond a critical curve that falls roughly as $G_{P\to\phi}^{-1}$, hosts go extinct. Fig. 6.1C reveals in the $(\delta, G_{Q\to\phi})$ space, beyond a line of critical points, phages go extinct. Behavior in the $(\delta, G_{Q\to P})$ space is similar. We provide an analytical treatment below.

Bifurcations occur when the number of fixed points or their stability properties change in response to a dynamical parameter. Our system can approach three qual-

itatively distinct steady states: the first corresponds to host extinction, which we denote by $E^* = (P_e^*, Q_e^*, V_e^*) = (0,0,0)$. The second corresponds to a phage free system, which occurs with pure cultures where phages have not been introduced, or when hosts completely evade phage infection, which we denote by $F^* = (P_f^*, Q_f^*, V_f^*) = (P_f^*, 0, 0)$. The third corresponds to the case of prokaryote-phage coexistence, which we denote by $C^* = (P_c^*, Q_c^*, V_c^*)$.

In the phage free situation, the system evolves along the curve $\dot{P} = A_P(1-P) - \delta G_{P\to\phi} P$, towards the fixed point $P_f^* = 1 - \frac{\delta G_{P\to\phi}}{A_P}$. Non-extinction/positivity condition on this expression reveals a criticality condition on δ for maintenance of hosts carrying our simple immune system in the phage free case: $\delta < \frac{A_P}{G_{P\to\phi}} = \delta_1$. This is precisely the curve mapped out in Fig. 6.1B beyond which the hosts go extinct; when $\delta = \delta_1$, $F^* = E^*$, and when $\delta > \delta_1$, F^* is infeasible. Hence, as long as the immune system (with an autoimmunity side effect) is suppressed below a critical nondimensional ratio of the free cell reproduction rate to that of its autoimmunity potential, the phage free steady state is feasible.

The non-trivial fixed point for the case of coexistence, C^*, is given by:

$$
\begin{aligned}
P_c^* &= \frac{G_V}{\dfrac{\alpha_v}{\left(1+\underbrace{\dfrac{G_{Q\to\phi}+G_{Q\to P}}{G_{Q\to V}}}_{\text{immune advantage}}\right)}} - 1 \\
V_c^* &= \frac{A_P(1-P_c^*) - \delta G_{P\to\phi}}{A_P P_c^* + \frac{G_{Q\to\phi}+G_{Q\to P}}{G_{Q\to V}}} \\
Q_c^* &= \frac{P_c^* V_c^*}{G_Q}
\end{aligned}
\tag{6.3}
$$

Here $G_Q = (G_{Q\to\phi} + G_{Q\to P} + G_{Q\to V})$ denotes the overall removal rate of infected cells. In this coexistence regime, the steady state expression for P_c^* decomposes into the two parts: steady-state value when the dynamics is phage limiting and the advantage offered by the immune system in overcoming phage lysis. This advantage is given by the ratio of the sum of immunity and autoimmunity rates conferred by the immune system in infected cells to that of the phage specific lysis rates. Thus inducing autoimmunity, alongside immunity, in infected cells (abortive infection) is beneficial to the prokaryotic population when coevolving with phages. As is the case with predator-prey models, P_c^* is independent of the cell's own growth rate [214], and is completely determined by the immunity and autoimmunity parameters, along with the phage specific parameters. Furthermore, positivity conditions on the steady state values yields the feasibility conditions for the existence of this steady state: $(0 < P_c^* < 1)$, and $(0 \leq \delta < \delta_2)$ with $\delta_2 = \frac{A_P(1-P_c^*)}{G_{P\to\phi}}$ (as $V_c^* \leq 0$ otherwise), giving us a tighter constraint on δ for coexistence. Notice that $\delta_2 < \delta_1$. So regardless of the presence or absence of phages, a free cell autoimmunity suppression level of $\delta < \delta_1$ is required for the population to avoid losing the immune system altogether.

When free cells completely repress the immune system ($\delta = 0$), or when there is no autoimmunity ($G_{Q\to\phi} = 0$), V_c^* and Q_c^* achieve their maximum values. As $\delta \to \delta_2$, the values of V_c^* and Q_c^* are reduced progressively. The form of these equillibria implies that by increasing the net autoimmunity rate in free cells, lower net viral abundance is achieved. However, by doing so the range of δ that supports coexistence is narrowed. When $\delta > \delta_2$, the coexistence steady state C^* is infeasible, and the system operates in the phage free regime, at which point, the condition $\delta < \delta_1$ has to be satisfied to avoid host

extinction. The bifurcation diagram in Fig. 6.1A maps this behavior: C^* continues to be stable until $\delta < \delta_2$, whereas beyond δ_2 the otherwise unstable F^* becomes stable. The stability of the steady states ascertained by the Routh-Horwitz criteria [214].

To analyze the influence of *abortive infection* on coevolution, we produced a two-parameter bifurcation diagram for the $(\delta, G_{Q\to\phi})$ space (Fig. 6.1C). Two distinct regimes are clear: a coexistence regime, and a regime where hosts evade phages. A third regime corresponding to host extinction also occurs for autoimmunity suppression exceeding the value δ_1 (for the parameters in this figure, it occurs along the line $\delta = 1$). The bifurcation diagrams are similar for a variety of other parameter combinations tested. Coexistence occurs for low values of $G_{Q\to\phi}$, and are progressively lost as δ is increased. We can trace the line of critical points analytically as follows. Recall that the switch from coexistence to phage evasion is principally determined by the equality $\delta = \delta_2 = \frac{A_P(1-P_c^*)}{G_{P\to\phi}}$. If we let $G_Q = (G_{Q\to V} + G_{Q\to P} + G_{Q\to\phi})$ and substituting for P_c^*, we obtain $1 - \delta\frac{G_{P\to\phi}}{A_P} = \frac{G_V}{\frac{\alpha_v G_{Q\to V}}{G_Q} - 1}$. When $\frac{\alpha_v G_{Q\to V}}{G_Q} >> 1$, as a function of $G_{Q\to\phi}$ and δ, this condition spans the line:

$$\frac{\delta}{K_1} + \frac{G_{Q\to\phi}}{K_2} = 1 \tag{6.4}$$

where the intercepts are given by $K_1 = \frac{A_P}{G_{P\to\phi}}\left[1 - \frac{G_V}{\alpha_v}\left(1 + \frac{G_{Q\to P}}{G_{Q\to V}}\right)\right]$, and $K_2 = G_{Q\to V}\left[\frac{\alpha_v}{G_V} - \left(1 + \frac{G_{Q\to P}}{G_{Q\to V}}\right)\right]$. For the parameters in Fig. 6.1C, Routh-Horwitz criteria [214] reveals that the achieved C^* values are stable. Beyond this boundary, coexistence is infeasible, and cells assume a density determined completely by δ, and independent of $G_{Q\to\phi}$: $P_f^* = 1 - \frac{\delta A_P}{G_{P\to\phi}}$. Clearly, both K_1 and K_2 are reduced with increasing values of $G_{Q\to P}$ (immune rate), the net effect being reduction of the area under the line resulting in loss

of coexistence. To map the influence of immunity, one can similarly establish the critical line determining the boundary of coexistence explicitly as a function of $(\delta, G_{Q\to P})$.

In summary, our bifurcation analysis of this simple model (i) reveals the precise regimes for the three possible fates of a prokaryotic immune system with regulated autoimmunity (complete evasion of phages, coexistence with phages, or extinction) (ii) shows that infected cell autoimmunity (alongside restriction) is beneficial to the prokaryotic population, and (iii) reveals a strict limit on the free cell autoimmunity levels above which host extinction occurs.

Perhaps the most characteristic feature of CRISPRs is their adaptive ability for continued novelty resulting from spacer acquisitions and deletions. The model above does not incorporate spacer turnover kinetics or its regulation. Neither does it allow us to explicitly determine the influence of host protospacer levels on the interval of autoimmunity regulation $0 \leq \delta < \delta_1$; the larger this window, the higher the cellular tolerance for CRISPRs.

We will therefore proceed to incorporate CRISPR specific reactions into the simple model described above. We will show that (i) the simple model arises as a particular limit of a more general model, and (ii) by thwarting the accumulation of self-targeting spacers through an SND (whose existence/absence is hard to ascertain from existing data), and/or through a highly active spacer deletion mechanism, the range of free-cell CRISPR activity levels, δ, is widened. Furthermore, the general model will reveal other idiosyncratic features of CRISPR and its maintenance in populations over ecological time scales.

6.1.2 A detailed model for CRISPRs incorporating their adaptive ability and regulation

In this section we develop a more detailed model of CRISPR dynamics, which generalizes the simple model discussed above. Our modeling strategy in this section (**Fig 6.2**) is intermediate to models that fix a constant rate of immunity (as in [1]) and agent-based models that describe strain-specific immunity (as in [2]). Briefly, we track spacer accumulations over time and use linear mass action kinetics to model the CRISPR reactions and the resulting ecological dynamics due to immunity and autoimmunity. Such an approach offers the computational advantage to model growing populations while simultaneously accounting for the underlying regulatory dynamics of CRISPR and its kinetics. While this model cannot capture strain-specific behavior, we can nonetheless make qualitative and even quantitative predictions for the average spacer accumulation kinetics resulting from the adaptive nature of CRISPR dynamics. The key variables in this detailed model are described in **Table 6.2** and discussed below.

We let π_v denote the total number of phage protospacers per phage genome. The amount of self-targeting spacers per prokaryotic genome is defined relative to the phage protospacer amount as $\beta\pi_v$. Thus $\beta = 0$ implies no self-targeting protospacers per prokaryotic genome, which can also be interpreted as the absence of self-targeting protospacers due to the presence of an SND. At any time, both the free and infected cell populations (denoted as p and q respectively) have an associated CRISPR spacer content, the "per-cell" quotas of which are completely specified by y_{pA}, y_{pI}, y_{pS} and y_{qA}, y_{qI}, y_{qS} respectively (table 6.2). Here $y_{\cdot A}$ denotes the active spacer quota per cell (i.e., phage reactive), $y_{\cdot I}$ de-

Variable	Description	Value, Units
p	Free cell density	$cells\ ml^{-1}$
q	Infected cell density	$cells\ ml^{-1}$
v	Phage density	$phages\ ml^{-1}$
(y_{pA}, y_{pI}, y_{pS})	Average Active, Inactive, and Self-targeting spacer quota per free cell	$spacers\ cell^{-1}$
(y_{qA}, y_{qI}, y_{qS})	Average Active, Inactive, and Self-targeting spacer quota per infected cell	$spacers\ cell^{-1}$
x_A	Average Active phage protospacer quotea per infected cell	$protospacers\ cell^{-1}$
α_p	Free cell replication rate	$1\ hr^{-1}$
α_q	Phage adsorption rate	$5 \times 10^{-9}\ ml\ phage^{-1}\ hr^{-1}$
α_v	Phage burst size	$50\ phages$
γ_v	Phage death rate	$5\ hr^{-1}$
Φ_p	Environmental carrying capacity	$10^8\ cells\ ml^{-1}$
α_c	Acquisition rate of new spacers	$10^{-6}\ cells\ hr^{-1}$
γ_c	Deletion rate of new spacers	$varied\ hr^{-1}$
$\gamma_{q\to p}$	Immune rate per active spacer per infected cell	$1 - 10^{-6}(\frac{spacers}{cell})^{-1}\ hr^{-1}$
$\gamma_{q\to\phi}$	Autoimmunity rate per self-targeting spacer per infected cell	$varied\ (\frac{spacers}{cell})^{-1}\ hr^{-1}$
$\gamma_{q\to v}$	Lysis rate of infected cells	$1\ hr^{-1}$
π_v	Total number of protospacers per phage genome	$phage^{-1}$
$\beta \times \pi_v$	Total number of self-targeting protospacers per prokaryotic cell.	$varied\ protospacers\ cell^{-1}$
δ	Scale factor $(0 \leq \delta \leq 1)$, determines CRISPR activity in free cells.	
μ_v	Phage mutation rate per protospacer	$30 \times 10^{-8} protospacers^{-1}$

Table 6.2: Description of the different variables used in the detailed model. *Dynamic variables are denoted with Roman letters, and parameters are denoted with Greek symbols. Any parameter associated with production of an item i is denoted as α_i and that with its degradation is denoted as γ_i. Steady state value of an item i will be donted by i^*. Parameter values were obtained from [1, 2].*

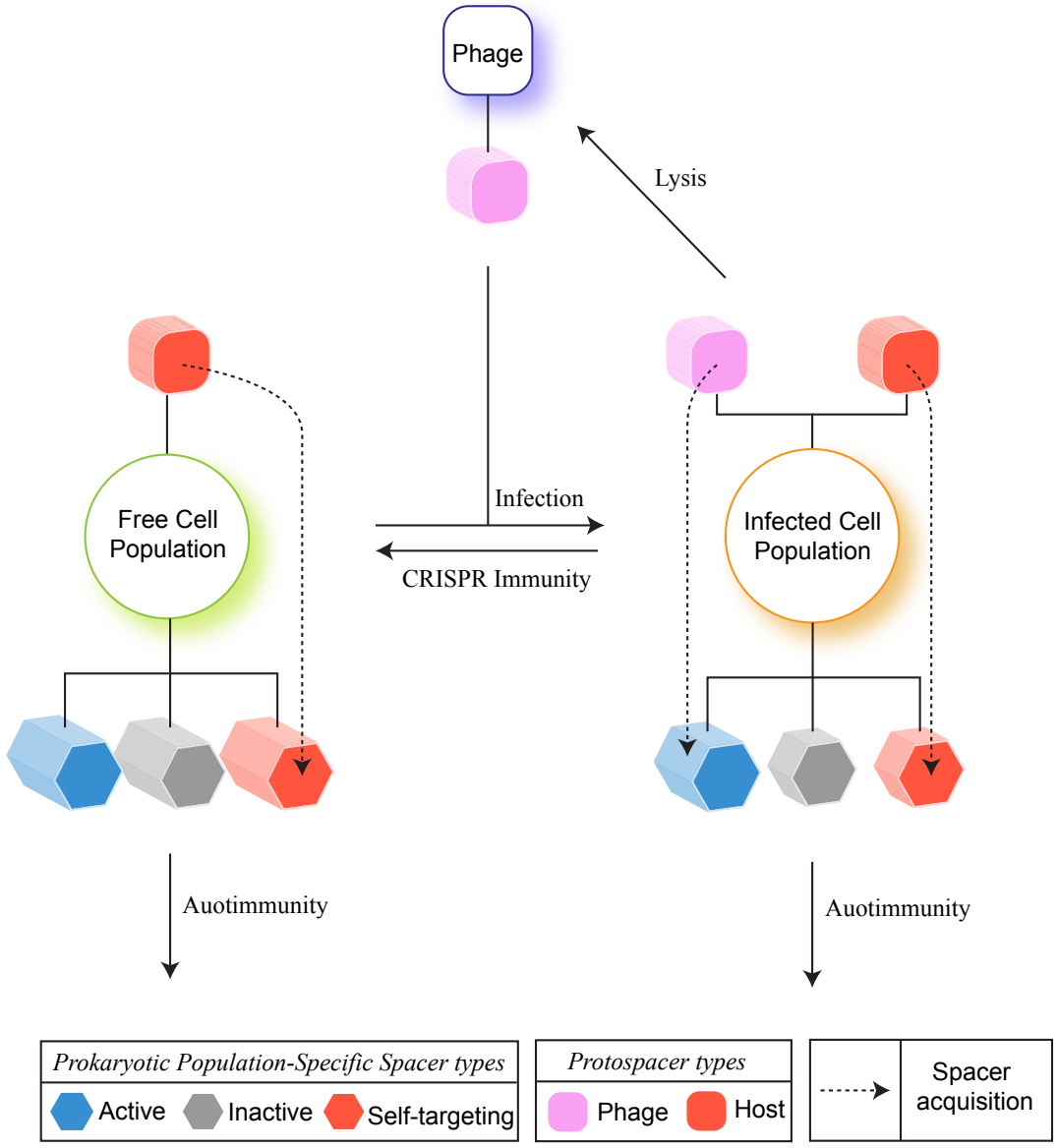

Figure 6.2: A detailed model of CRISPR dynamics. *The infected cell population (and its associated CRISPR spacer content) is created from the growing free cell population (and its corresponding CRISPR content) through phage infections. The overall CRISPR spacer content in each cell population is abstractly partitioned into* active, inactive *and* self-targeting. *Active spacers elicit phage restriction, while self-targeting spacers cause cell death (autoimmunity). While both the free and infected cell populations have genomic protospacers that contribute to the creation of self-targeting spacers, only the infected cell population has access to the released phage protospacers for the creation of active spacer content.* *Continued on next page.*

Figure 6.2: *At any given time, the CRISPR induced rate of immunity for an infected cell is proportional to its per capita quota of active spacer content associated with the population at that time. Similarly, we use the corresponding self-targeting spacer content to define the rates of autoimmunity for both the infected and free cell populations. In our equations, we directly model these per capita quotas. Thus the rates of CRISPR induced immunity and autoimmunity for a cell population are reflective of its associated spacer content at any given time, which in turn is determined by the kinetics of CRISPR and prokaryote-phage interaction.*

notes the inactive spacer quota per cell (i.e., phage inactive, due to mutations in the corresponding PAMs in phages) and $y_{\cdot S}$ denotes the self-targeting spacer quota per cell. The average phage protospacer quota per infected cell available for its new spacer acquisitions is denoted by x_A.

The per capita quotas of the various types of CRISPR spacer content are used to model the rates of acquisition and interference reactions in each subpopulation. Let $\gamma_{q\to p}$ be the rate of immunity conferred per active spacer; then at any given time the immunity rate per infected cell is assumed to be $\gamma_{q\to p} y_{qA}$. Similarly, if $\gamma_{q\to}$ denotes the rate of autoimmunity conferred per self-targeting spacer, the autoimmunity rate per infected cell is then $\gamma_{q\to\phi} y_{qS}$. To obtain the corresponding term for the free cell population we will first need to model infection-mediated CRISPR activation.

As the operonic structure of CRISPR/Cas genes lends itself to regulation based on free/infected cell states [208, 209, 215–218], we simply scale the rates of all the CRISPR reactions (acquisition, deletion and interference) by δ $(0 \leq \delta \leq 1)$, in the free cell population relative to that of the infected population. So $\delta = 0$ implies that all CRISPR reactions in free cells are switched off whereas $\delta = 1$ implies that there is no differential CRISPR expression between the free and infected cell populations. Note that, only infected cells

can acquire novel phage protospacers, while both infected and free cell populations can acquire self-targeting protospacers. The latter events occur when $\delta > 0$. Under these modeling assumptions, the corresponding autoimmunity rate per self-targeting spacer is given by $\delta\gamma_{q\to\phi}$; this is scaled by the per capita free cell quota of self-targeting spacers to calculate the autoimmunity rate per free cell, $\delta\gamma_{q\to\phi}y_{pS}$.

6.2 Population dynamics.

We now describe how the above reactions are coupled with prokaryote-phage co-evolution. Free cells (p) replicate at a rate α_p under the constraint imposed by the carrying capacity Φ_p. Free cells are also produced from infected cells (q) due to immune evasions of phage lysis at a rate of $\gamma_{q\to p}y_{qA}$ (as described above). Thus the net rate of infected cells that undergo immunity is given by $\gamma_{q\to p}y_{qA}q$. Phages (v) infect free cells with an adsorption rate constant α_q to produce q. In addition, free cells undergo autoimmunity at a rate of $\gamma_{q\to\phi}y_{pS}$, which is determined by the amount of self-targeting spacers (y_{pS}) in free cells and the degree of CRISPR activity in free versus infected cells (δ). Phages with a burst size of α_v are produced from lysis of infected cells at rate $\gamma_{q\to v}$ and removed a t a rate of γ_v. q can undergo autoimmunity at a rate of $\gamma_{q\to\phi}y_{qS}$, or switch to free cells with rate $\gamma_{q\to p}y_{qA}$. The differential equations are then given as:

$$
\begin{aligned}
\dot{p} &= \alpha_p p\left(1-\frac{p+q}{\Phi_p}\right) - \underbrace{\delta\gamma_{p\to\phi}y_{pS}p}_{\text{autoimmunity}} - \alpha_q pv + \underbrace{\gamma_{q\to p}y_{qA}q}_{\text{immunity}} \\
\dot{q} &= \underbrace{\alpha_q pv}_{\text{infections}} - (\gamma_{q\to\phi}y_{qS} + \gamma_{q\to p}y_{qA} + \gamma_{q\to v})q \\
\dot{v} &= \underbrace{\alpha_v\gamma_{q\to v}q}_{\text{lysis}} - \gamma_v v - \alpha_q pv
\end{aligned}
\tag{6.5}
$$

For convenience in exposition below, we will let $\Gamma_p = (\alpha_q v + \delta\gamma_{q\to\phi}y_{pS})$ and $\Gamma_q = (\gamma_{q\to p}y_{qA} + \gamma_{q\to\phi}y_{qS} + \gamma_{q\to v})$, which denote the overall removal rates of cells in the free and infected populations respectively.

6.3 Spacer and protospacer concents in free and infected cells

Fig. 6.3 presents the set of reactions influencing the total spacer and protospacer contents of different types. These give rise to the following derivatives when $q(t) \neq 0$ and $p(t) \neq 0$.

$$
\begin{aligned}
\dot{x}_A &= \alpha_q\frac{pv}{q}\left[(1-\mu_v)\pi_v - x_A\right] \\
\dot{y}_{qA} &= \alpha_c x_A + \alpha_q\frac{pv}{q}\left[y_{pA} - y_{qA}\right] - \gamma_c y_{qA} \\
\dot{y}_{qI} &= \alpha_q\frac{pv}{q}\left[y_{pI} - y_{qI}\right] - \gamma_c y_{qI} \\
\dot{y}_{qS} &= \alpha_c\beta\pi_v + \alpha_q\frac{pv}{q}\left[y_{pS} - y_{qS}\right] - \gamma_c y_{qS} \\
\dot{y}_{pA} &= \mu_v\left[y_{pI} - y_{pA}\right] + \gamma_{q\to p}\frac{y_{qA}q}{p}\left[y_{qA} - y_{pA}\right] - \delta\gamma_c y_{pA} \\
\dot{y}_{pI} &= \mu_v\left[y_{pA} - y_{pI}\right] + \gamma_{q\to p}\frac{y_{qA}q}{p}\left[y_{qI} - y_{pI}\right] - \delta\gamma_c y_{pI} \\
\dot{y}_{pS} &= \delta\alpha_c\beta\pi_v + \gamma_{q\to p}\frac{y_{qA}q}{p}\left[y_{qS} - y_{pS}\right] - \delta\gamma_c y_{pS}
\end{aligned}
\tag{6.6}
$$

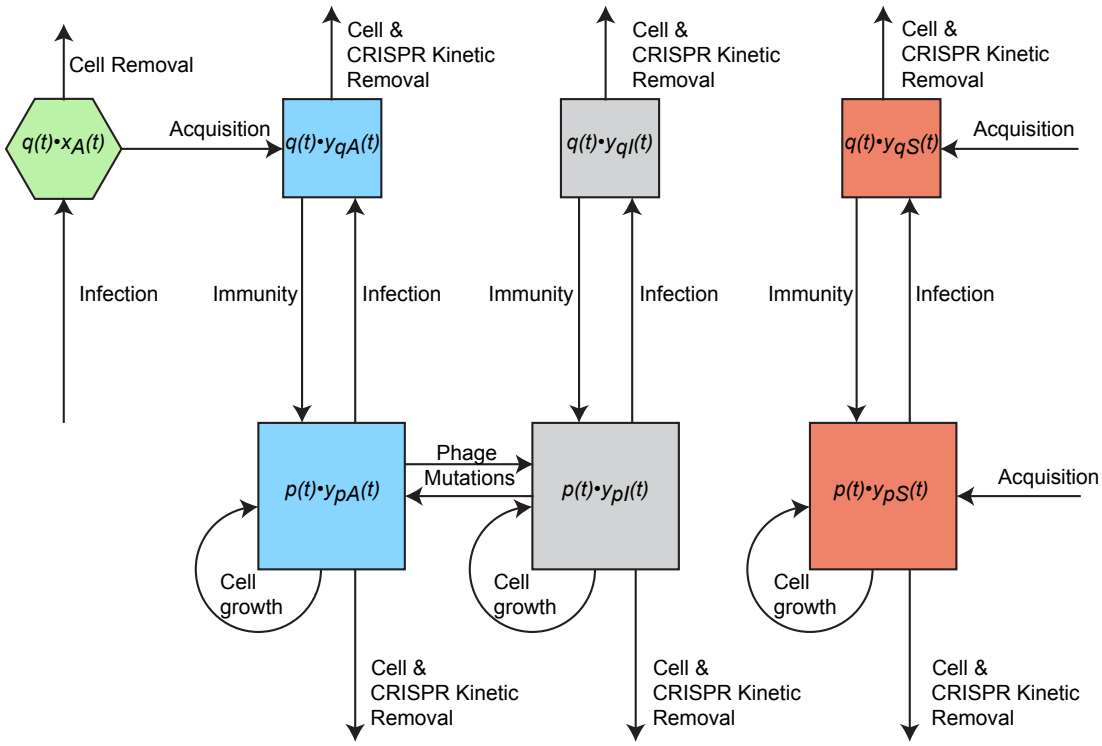

Figure 6.3: Reactions influencing total spacer and protospacer densities. *The inflow and outflow of different species are indicated. The figure shows the reactions influencing the total spacer and protospacer contents at any given time in the population. We use this reaction set to derive the rates of average spacer quota change over time. Squares in the top row correspond to the total protospacer and spacer content in the infected cell population; those in the bottom correspond to those in the free cell population. Note that while we model average spacer quotas this figure illustrate all the reactions that influence* total *spacer contents.*

We derive the aforementioned per cell quotas of protospacer (x_A), and various spacer contents as follows.

Active phage protospacer quota per infected cell Because we track the per capita quotas of protospacer contents per infected cell, any expression for its derivative has to account for the current spacer density in the infected cell population, the influx due to the newly infected cells, weighted by their corresponding population sizes (refer Fig. 6.3). At any time instant t, the total amount (density) of phage protospacers associated with

the entire infected population is $x_A(t)q(t)$. The total amount of newly released phage protospacers is given by the product of total amount of infections and the expected amount of native phage protospacers per phage as $\alpha_q p(t)v(t) \times (1-\mu_v)\pi_v$. The total amount of protospacers leaving the infected pool is proportional to the removal rate of infected cells and is equal to $\Gamma_q(t)q(t)x_A(t)$. For any small time interval Δt then, we can write $x_A(t+\Delta t)$ as:

$$x_A(t+\Delta t) = \frac{x_A(t)q(t) + \Delta t\alpha_q(1-\mu_v)\pi_v p(t)v(t) - \Delta t\Gamma_q(t)q(t)x_A(t)}{q(t) + \Delta t\alpha_q p(t)v(t) - \Delta t\Gamma_q(t)q(t)} \tag{6.7}$$

where the denominator is the expected infected cell density at time $t+\Delta t$. Thus $x_A(t+\Delta t)$ is precisely the average protospacer content per infected cell at time $t+\Delta t$. In a straightforward fashion, when $q(t) \neq 0$, we can compute the limit $\frac{dx_A(t)}{dt} = \lim_{\Delta t \to 0} \frac{x_A(t+\Delta t) - x_A(t)}{\Delta t}$ to obtain our derivative:

$$\dot{x}_A = \frac{\alpha_q p(t)v(t)\left[(1-\mu_v)\pi_v - x(t)\right]}{q(t)} \tag{6.8}$$

We now follow a similar procedure to calculate the average spacer contents in the free and infected cell populations.

CRISPR spacer quota per infected cell New additions to the active and self-targeting spacer content associated with infected cell population can occur upon infection due to acquisition reactions. In addition, they are also inherited from free cells that are infected. Inactive spacers in the infected cell population, however, can only be inherited. Furthermore, we will also account for the removal of spacers due to CRISPR kinetics through a

spacer deletion parameter γ_c .

Given the current phage protospacer levels available per infected cell (x_A) for spacer acquisitions and the acquisition rate of α_c, the total rate of new active spacer acquisitions is computed as $\alpha_c x_A$. Similarly, given the current genomic protospacer density of $\beta\pi_v q$, the rate of newly acquired self-targeting spacer content is given by $\alpha_c\beta\pi_v q$. For a given spacer type, the inflow due to inheritance is determined by the amount of infections ($\alpha_q pv$) and the spacer density of that type in the free cell population (e.g., $\alpha_q vp \times y_p A$ for active spacers). Finally, all three spacer types within an infected cell are also removed at a rate proportional to the removal rate of infected cells and spacer deletion. Taken together this results in the following equations for the different spacer contents at time $t+\Delta t$:

$$
\begin{aligned}
y_{qA}(t+\Delta t) &= \frac{q(t)y_{qA}(t) + \Delta t\,\alpha_c x_A(t)q(t) + \Delta t\,\alpha_q p(t)v(t)y_{pA}(t) - \Delta t\left(\Gamma_q+\gamma_c\right)q(t)y_{qA}(t)}{q(t)+\Delta t\,\alpha_q p(t)v(t) - \Delta t\Gamma_q(t)q(t)} \\
y_{qI}(t+\Delta t) &= \frac{q(t)y_{qI}(t) + \Delta t\,\alpha_q p(t)v(t)y_{pI}(t) - \Delta t\left(\Gamma_q+\gamma_c\right)q(t)y_{qI}(t)}{q(t)+\Delta t\,\alpha_q p(t)v(t) - \Delta t\Gamma_q(t)q(t)} \\
y_{qS}(t+\Delta t) &= \frac{q(t)y_{qS}(t) + \Delta t\,\alpha_c\beta\pi_v q(t) + \Delta t\,\alpha_q p(t)v(t)y_{pS}(t) - \Delta t\left(\Gamma_q+\gamma_c\right)q(t)y_{qS}(t)}{q(t)+\Delta t\,\alpha_q p(t)v(t) - \Delta t\Gamma_q(t)q(t)}
\end{aligned}
\tag{6.9}
$$

When $q(t) \neq 0$, we obtain the corresponding derivatives for the variables by computing the limits $\lim_{\Delta t\to 0}\frac{y_{qA}(t+\Delta t)-y_{qA}(t)}{\Delta t}$, $\lim_{\Delta t\to 0}\frac{y_{qI}(t+\Delta t)-y_{qI}(t)}{\Delta t}$ and $\lim_{\Delta t\to 0}\frac{y_{qS}(t+\Delta t)-y_{qS}(t)}{\Delta t}$.

CRISPR spacer quota per free cell New additions to the self-targeting spacer content in free cells is determined by the differential activation rate of acquisition in free cells ($\delta\alpha_c$), and the current available pool of genomic protospacers $\beta\pi_v p$. All three spacer types are inherited from the infected cells at a rate proportional to the amount of infected cells undergoing immunity. Further, at a rate determined by per protospacer spacer mu-

tation rate μ_v , mutated phage protospacers can switch to being native (and vice versa); at this rate then, this effect is also reflected in the corresponding CRISPR content as the transition of inactive spacers to active states (and vice versa). For simplicity, we do not consider the difference in the rates of forward and backward mutation rates. Finally, all three spacer types within a free cell replicate at a rate proportional to the effective free cell duplication rate, and are removed at a rate proportional to the removal rate of free cells and spacer deletion (which, as mentioned before, is scaled by the CRISPR activation rate δ). Taken together this results in the following equations for average spacer contents at time $t+\Delta t$ (for clarity, we ignore mentioning time dependence explicitly):

$$
\begin{aligned}
y_{pA}(t+\Delta t) &= \frac{y_{pA}p + \Delta t\mu_v y_{pI}p - \Delta t\mu_v y_{pA}p + \Delta t y_{qA}^2 q - \Delta t\Gamma_p p y_{pA} - \Delta t\delta\gamma_c p y_{pA} + \Delta t\alpha_p p\left(1-\frac{p+q}{\Phi}\right)y_{pA}}{p+\Delta t\gamma_{q\to p}y_{qA}q - \Delta t\Gamma_p p + \Delta t\alpha_p p\left(1-\frac{p+q}{\Phi}\right)} \\
y_{pI}(t+\Delta t) &= \frac{y_{pA}p + \Delta t\mu_v y_{pA}p - \Delta t\mu_v y_{pI}p + \Delta t y_{qA}q y_{qI} - \Delta t\Gamma_p p y_{pI} - \Delta t\delta\gamma_c p y_{pI} + \Delta t\alpha_p p\left(1-\frac{p+q}{\Phi}\right)y_{pI}}{p+\Delta t\gamma_{q\to p}y_{qA}q - \Delta t\Gamma_p p + \Delta t\alpha_p p\left(1-\frac{p+q}{\Phi}\right)} \\
y_{pS}(t+\Delta t) &= \frac{y_{pS}p + \Delta t\alpha_c\beta\pi_v p + \Delta t\gamma_{q\to p}y_{qA}q y_{qS} - \Delta t\Gamma_p p y_{pS} - \Delta t\delta\gamma_c p y_{pS} + \Delta t\alpha_p p\left(1-\frac{p+q}{\Phi}\right)y_{pI}}{p+\Delta t\gamma_{q\to p}y_{qA}q - \Delta t\Gamma_p p + \Delta t\alpha_p p\left(1-\frac{p+q}{\Phi}\right)}
\end{aligned}
\tag{6.10}
$$

When $p(t)\neq 0$, we obtain the derivatives as $\lim_{\Delta t\to 0}\frac{y_{pA}(t+\Delta t)-y_{pA}(t)}{\Delta t}$, $\lim_{\Delta t\to 0}\frac{y_{pI}(t+\Delta t)-y_{pI}(t)}{\Delta t}$ and $\lim_{\Delta t\to 0}\frac{y_{pS}(t+\Delta t)-y_{pS}(t)}{\Delta t}$.

We non-dimensionalize our equations by choosing to measure our cell density variables in units of the carrying capacity Φ_p, and phage density in units of $\alpha_v\Phi_p$, spacer and protospacer variables in units of the number of native phage protospacers π_v, and time in the non-dimensional units of $\tau=\alpha_c^{-1}t$ (CRISPR evolutionary time scales). This leads to the following set of equations, with effective parameters $A_V=\frac{\alpha_v\alpha_q}{\alpha_c}\Phi_p$, $G_{Q\to P}=\frac{\gamma_{q\to p}}{\alpha_c}\pi_v$, and $G_{Q\to\phi}=\frac{\gamma_{q\to\phi}}{\alpha_c}\pi_v$, while the rest of the rate parameters get scaled by α_c^{-1}. Non-

dimensionalization, apart from reducing the number of parameters in the model, also simplifies analysis of relative parameter sizes.

$$
\begin{aligned}
\dot{P} &= A_P P(1-(P+Q)) + G_{Q\to P} Y_{QA} Q - \delta G_{Q\to\phi} Y_{PS} P - A_V PV \\
\dot{Q} &= A_V PV - (G_{Q\to\phi} Y_{QS} + G_{Q\to P} Y_{QA} + G_{Q\to V}) Q \\
\dot{V} &= G_{Q\to V} Q - G_V V - \frac{A_V}{\alpha_v} PV \\
\dot{X}_A &= A_V \frac{PV}{Q} [(1-\mu_v) - X_A] \\
\dot{Y}_{QA} &= X_A + A_V \frac{PV}{Q} [Y_{PA} - Y_{QA}] - G_C Y_{QA} \\
\dot{Y}_{QI} &= A_V \frac{PV}{Q} [Y_{PI} - Y_{QI}] - G_C Y_{QI} \\
\dot{Y}_{QS} &= \beta + A_V \frac{PV}{Q} [Y_{PS} - Y_{QS}] - G_C Y_{QS} \\
\dot{Y}_{PA} &= M_V [Y_{PI} - Y_{PA}] + G_{Q\to P} \frac{Y_{QA} Q}{P} [Y_{QA} - Y_{PA}] - \delta G_C Y_{PA} \\
\dot{Y}_{PI} &= M_V [Y_{PA} - Y_{PI}] + G_{Q\to P} \frac{Y_{QA} Q}{P} [Y_{QI} - Y_{PI}] - \delta G_C Y_{PI} \\
\dot{Y}_{PS} &= \delta\beta + G_{Q\to P} \frac{Y_{QA} Q}{P} [Y_{QS} - Y_{PS}] - \delta G_C Y_{PS}
\end{aligned}
\tag{6.11}
$$

6.3.1 Simulations and bifurcation analysis.

All numerical simulations were performed with Matlab 2013b. Numerical bifurcation analyses were performed with XPPAUT (AUTO) [219].

6.3.2 SND absence is extremely lethal in the absence of regulation

In the absence of SND, given the large host genome size relative to that of phage (e.g. E.coli genome is roughly 100x the length of phage λ)and short PAM demarcating

protospacers, we expect an abundant host protospacer pool. In our model, this would imply a large host to phage protospacer ratio ($\beta > 1$). On the other hand, if SND is present, then its efficiency determines the β value, with higher efficiencies implying lower β values and vice versa. Similarly, the parameter δ determines the activation level of CRISPRs in free cells relative to that of infected cells; thus $\delta = 0$ represents complete repression, and $\delta = 1$ signifies no difference in CRISPR activation between free and infected cell populations.

To study the influence of host protospacers levels and regulation on prokaryotic densities, we vary δ and β across a large range of biologically feasible values (**Fig. 6.4**). Remarkably, as we observed in the case of our simple model, the steady state prokaryotic densities show a sharp, threshold-like behavior as a function of the degree of CRISPR regulation δ: hosts switch from maximal densities to complete extinction as the degree of free-cell CRISPR activity, δ, increases (Fig. 6.4A). Even in the case of comparable levels of host and phage protospacer ($\beta = 1$), greatly reduced levels of activation in free versus infected cells ($\delta < 0.01$) are required to guarantee host existence. While this tight window of prokaryotic existence is relaxed slightly at lower host protospacer levels, these results indicate that tight regulatory control is necessary for a wide range of host protospacer levels. It is therefore clear that the presence or absence of an SND is a crucial determinant of CRISPR maintenance in populations.

Fig. 6.4B shows the time course of several typical simulations for various (β, δ) combinations, to illustrate the effects of these two key parameters on intracellular spacer contents. For a wide range of parameters and initial conditions we find that the system approaches a steady state.

6.3.3 A simple constraint determines CRISPR maintenance in the model

We now work to derive an analytical understanding of the critical limit on δ (denoted by δ_1) that permits population survival. As in the simplified model, exact conditions for the threshold-like behavior of the system in the δ and β space can be obtained by considering the phage free system, in which case, the full system reduces to:

$$\begin{aligned}
\dot{P} &= A_P P(1-P) - \delta G_{Q\to\phi} Y_{PS} P \\
\dot{Y}_{PA} &= M_V [Y_{PI} - Y_{PA}] - \delta G_C Y_{PA} \\
\dot{Y}_{PI} &= M_V [Y_{PA} - Y_{PI}] - \delta G_C Y_{PI} \\
\dot{Y}_{PS} &= \delta\beta - \delta G_C Y_{PS}
\end{aligned} \tag{6.12}$$

These values give rise to the following fixed point: $\{P^* = 1 - \frac{\delta G_{Q\to\phi} Y_{PS}}{A_P}, Y^*_{PA} = 0, Y^*_{PI} = 0, Y^*_{PS} = \frac{\beta}{G_C}\}$. In the absence of any feedback from infections, and in the presence of an active spacer deletion mechanism, the active and inactive spacer contents are progressively lost from the population. The influence of CRISPR induced autoimmunity on free cell density is manifest in the steady state expression for free cells. For a population to not completely lose their CRISPR activity, the condition $P^* > 0$ must be satisfied. This leads us to the condition required for sufficient suppression of CRISPR in free cells:

$$\delta < \frac{A_P}{G_{Q\to\phi} Y^*_{PS}} = \frac{A_P G_C}{G_{Q\to\phi}\beta}, \tag{6.13}$$

For values of δ exceeding this upper bound, the system goes extinct. The same constraint holds for a system with phage, as non-negativity of the net cellular growth rate

is essential to avoid the only steady state of extinction. Note that, in the presence of a perfect SND, $\beta = 1$ and so the constraint on δ is effectively removed altogether. But in the absence of such a mechanism ($\beta > 0$), the internal steady state level of self-targeting spacers determines an upper limit on the free-cell CRISPR activity, δ.

The role of another crucial parameter is also apparent from this analysis: the spacer deletion rate. High spacer deletions can effectively remove self-targeting spacer accumulations, thus suppressing autoimmunity. So in addition to CRISPR regulation, the spacer deletion rate can also be increased to maintain CRISPR+ hosts in a population with larger host protospacer levels. We will use simulations below to determine how large this rate should be relative to the spacer acquisition rate.

6.3.4 Coevolutionary dynamics under the assumption of equilibrated spacer levels over CRISPR evolutionary time scales

For a wide variety of parameters and initial conditions tested, we found that the system converged to steady states (see Fig. 6.4B for an example). Let $(Y^*_{QA}, Y^*_{QS}, Y^*_{PS})$ denote the resulting steady state levels of intracellular spacer contents over CRISPR evolutionary time scales. These can then determine fixed rates of immunity $(G_{Q\to P}Y^*_{QA})$ and autoimmunity $(G_{Q\to\phi}Y^*_{QS}, G_{P\to\phi}Y^*_{PS})$. To do so, we use the simplified model shown in Fig. 6.1, which replaces all immunity and autoimmunity rates (which were originally functions of the spacer variables) by fixed rate constants. In such a limit, a thorough analysis of the coevolutionary dynamics is feasible. These results indicate that as long as the constraint on δ is met and the steady state intracellular levels of self-targeting spacers in infected cells

is non-zero, CRISPRs can exploit the abortive infection strategy alongside restriction. In the absence of SND, by contrast, the levels of self-targeting spacers will be much higher than phage reactive spacers. Under these conditions, the model predicts that CRISPRs will function principally as an abortive infection system.

We stress that we are not considering the situation that individual spacer sequences themselves are fixed in the population, but rather, the total number of them.

6.3.5 Four characteristic regimes of CRISPR activity

Given the importance of the dimensionless parameters $\{\delta, \beta, G_C\}$ in determining the evolutionary maintenance of CRISPR+ hosts, we now focus on understanding the influence of these parameters on the general model.

Free cell densities in the $\{\beta, G_C\}$ space for a given value of δ reveal a characteristic four-regime pattern. **Fig. 6.5** shows the free cell densities achieved (first column) and phage densities (second column) for various values of (β, G_C) values under two cases of δ: $\delta = 10^{-2}$ and $\delta = 10^{-4}$. Regime I occurs at low β and very high G_C values. Here both free cells and phages coexist; while the former assume significantly low levels (but never extinct), the latter achieve their highest densities. Regime II occurs at low β and low G_C values. Here hosts achieve their highest densities driving phage densities to very low values, if not extinction. In regime III, which occurs at high but still plausible β values, host extinction occurs. Regime IV is an extension of regime II's behavior, but at high G_C and high β values.

Hints to explain the existence of these four qualitative regimes, and their boundaries, are provided by the corresponding intracellular steady state spacer levels and the

constraint on δ we derived in the previous section. As we proceed to higher β values, the active spacer levels decrease and self-targeting spacer levels increase (see for example Fig. 6.4B). Higher β values lead to larger steady state levels of self-targeting spacers, effectively increasing the autoimmunity rate of infected cells. This inhibits immune mediated feedback of active spacers to the free cell population (through inheritance) and causes a reduction in the overall active spacer levels. Self-targeting spacers, on the other hand, can be independently acquired in free cells at a rate determined by δ. According to this basic intuition, we can now derive rough conditions for falling in each of the four qualitative regimes.

(Regime I) At high G_C values ($G_C \to \infty$) CRISPR cassettes are empty and the immunity and autoimmunity reactions are overwhelmed by phage lysis. Under these conditions, both the steady state spacer levels and their derivatives become zero, making the factor $\frac{G_{Q\to P}Y^*_{QA}+G_{Q\to\phi}Y^*_{QS}}{G_{Q\to V}} = 0$, resulting in no net growth advantage to CRISPR hosts (compare to steady state of the simple model). In this regime, the coevolutionary dynamics is phage limiting, resulting in steady state free cell levels of $\frac{G_V}{\alpha_v - 1}$ in terms of the simple model. (Regime II) At lower G_C values, and when the existence condition on δ is satisfied, both immunity and autoimmunity operate, allowing prokaryotes to evade phage lysis at significant rates. In this regime, phages are driven to very low densities or extinction. (Regime III) At lower GC values, progressing to higher β values increases steady-state levels of self-targeting spacers, thereby increasing the risk of not satisfying the constraint on δ. In such cases, regime III operates for all higher values of β, and extinction is inevitable. (Regime IV) This regime operates in the region where high levels of β are matched by corresponding high G_C values that are sufficient to reduce self-targeting

spacer levels so as to satisfy the δ constraint. In this regime, host extinction occurs. Here no active spacer mediated immunity occurs, but CRISPRs transform to a full-fledged abortive infection system. When $\delta = 0$, regime III does not occur, and regime IV extends into regime III. Thus the boundaries between regimes I and II, IV can be mapped by $\frac{G_{Q\to P}Y^*_{QA}+G_{Q\to\phi}Y^*_{QS}}{G_{Q\to V}} = 0$, and that between II, IV and III can be mapped by the critical condition on δ.

6.3.6 Elimination of abortive infection improves coexistence of phages

To study how ABI influences the coevolutionary dynamics in the general model, we remove the autoimmunity term from the model and compare the resulting prokaryotic and phage densities across several host protospacer and CRISPR activation levels (**Fig. 6.6**). We find that while removing ABI in infected cells increases the size of the coexistence regime and allows for improved phage densities. Indeed, this is the same effect predicted by our bifurcation analysis of the simplified model, where lower abortive infection rates lead to increased coexistence owing to higher phage turnover.

6.4 Discussion

A handful of prokaryote-phage experimental systems for studying CRISPR dynamics have been established. However, the extreme diversity of CRISPRs [190] makes it difficult to draw broad conclusions from any one biological model system. Computational models, which allow exploration over a wide range of feasible parameters, provide an attractive alternative.

In this work, we analyzed the influence of infection-induced activation of CRISPRs and their autoimmunity side effect on prokaryote-phage coevolutionary dynamics. Our model integrates the classical ingredients of the prokaryotic CRISPR immune system, along with aspects of regulation and autoimmunity. Our analysis suggests that CRISPRs exploit both restriction and abortive infection. Moreover, we identified a key constraint that determines the growth advantage associated with CRISPRs as a prokaryotic immune system. As summarized in **Fig. 6.7**, our model reveals a characteristic four-regime pattern determined principally by three effective parameters: the activation level of CRISPRs in uninfected population, the host to phage protospacer ratio, and spacer deletion to acquisition rate ratio in CRISPRs. In the presence of SND, the host to phage protospacer ratio is close to zero, and CRISPRs operate exclusively by exploiting restriction, while in the absence of SND, they tend to principally exploit the abortive infection route.

Several previous models have also studied CRISPR associated fitness costs, although as abstract functions. Nevertheless, these models reproduce and help to explain some of the key experimental and comparative genomics findings on CRISPRs. Levin and colleagues exploited classical density dependent ecological models to numerically analyze the invasion of costly CRISPR genotypes in the presence of innate (envelope) resistance and conjugative plasmids [1, 200, 220], and showed that selection due to continuous phage exposure and absence of less costly resistance mechanisms improve CRISPR maintenance in the population. Similar in spirit, Gandon and Vale make general discussions based on their analysis of general epidemiological models on the evolution of a CRISPR-like resistance mechanism, when the side effect associated is that of beneficial horizontal gene transfer impedance [213]. Childs et al., established a multiscale agent-

based simulation model to characterize CRISPR spacer and viral diversity during coevolution, and conclude that population dynamics is more sensitive to spacer acquisition rates than interference rates [2]. Weinberger et al., derive a critical threshold on CRISPR associated cost as a function of coevolving viral diversity, innate resistance and spacer acquisition rate and conclude that high viral diversity selects against CRISPRs [212]. Iranzo et al., used numerical simulations of a general agent based simulation model that additionally accounted for CRISPR loss and horizontal transfer, to exhaustively study CRISPR maintenance as a function of various kinetic parameters in their model [211]. They also concluded that CRISPR loss is encouraged at high prokaryote/phage population sizes.

Our analyses complement these studies summarized above, and they advance our understanding of CRISPR mechanisms in general. We have delineated the precise conditions under which CRISPRs can be lost even at low viral diversities. The level of complexity in our model, intermediate to previous simulations of agent-based models and models requiring radical simplifications and that do not account for the adaptive nature of CRISPR kinetics, provides an opportunity for mathematical analysis and intuitive understanding of the results. We have presented an analytical treatment of a particular limit of our model (which empirically hold for wide parameter regimes), summarizing qualitative behavior of the CRISPR system as a function of the underlying parameters.

It is also worthwhile to re-examine previous experimental and bioinformatic studies of CRISPRs, in light of the insights gained from our modeling analyses. We found that for CRISPRs to be maintained in a population, free-cell CRISPR activity must be sufficiently suppressed. This upper bound on free-cell activity is determined by a nondimensional ratio of free cell growth rate to that of its autoimmunity potential due to the accumulated

self-targeting spacers. An immediate consequence is that CRISPRs are likely to be lost from populations or cell types with reduced growth rates. This result helps to explain well-known empirical trends. For example, in general it is known that drug resistance or virulence is associated with moderate to high fitness costs; under these conditions cells often assume low growth rates [221]. According to our model, then, such strains should lack functional CRISPR elements, as has been confirmed for multi-drug resistant *Escherichia coli* [222]and for highly virulent *Francisella sp.* [223]. Furthermore, clinical isolates of *Pseudomonas aeruginosa* lack CRISPR resistance despite crRNA expression, and several virulent clinical isolates of pathogenic *Vibrio parahaemolyticus* [224], *Shigella* [225], pathogenic *Clostridium jejuni* [226] and *Mycoplasma gallisepticum* [227] seem to lack CRISPR resistance. While these studies have suggested a causal role played by CRISPR inactivity in the gain of virulence of clinical isolates, we propose an alternative mechanism: reduced growth rate in virulent strains induces selection for reduced CRISPR activity.

Under the assumptions of our model we can make approximate quantitative statements about the kinetic parameters underlying CRISPR function. In the absence of SND, our results suggest that CRISPRs can be maintained in a prokaryotic population only under high repression in free cells and/or high deletion rates ($> 10^2$ times the spacer acquisition rate in the absence of complete repression, as obtained in Fig. 6.5). But while high repression is possible through crosstalk with specialized pathways that detect phage invasion or foreign DNA element, as is often the case with toxin/anti-toxin or abortive infection systems [208,215,216,216–218,228], how can such high deletion to acquisition ratio be achieved? One possibility is a spacer deletion mechanism [180–182,229,230] but

we still lack sufficient biochemical characterization of this process. Our model assumed that the spacer deletion system is coupled with the rest of the CRISPR machinery, because it is likely that such a system must be expressed from the same operon as the rest of the CRISPR genes. We tested two hypothetical deletion systems that relax the requirement for high spacer deletion rates (**Fig 6.8**). The first is constitutively expressed regardless of the cell state. The second is regulated in a direction opposite to that of the rest of the CRISPR machinery – it is repressed when infected, and fully activated when uninfected. The reason these strategies work is because of the fundamental reduction they produce in the steady state expressions of the self-targeting spacers. Notice however that neither of these alterations guarantee CRISPR maintenance for arbitrarily large host protospacer levels. They still must respect the required constraint of reduced CRISPR activity in free cells.

A thorough biochemical characterization of the spacer deletion mechanism is required for advancing our understanding of CRISPRs. Stern et al. [187], in their large scale survey of CRISPR cassettes in microbial genomes, remarked that deactivated self-targeting spacers are found throughout the CRISPR array. This is in contrast to experimental conclusions that, in most systems, more recent acquisitions appear in the leader proximal end [181, 182, 229, 230]. In fact, Stern et al. found that self-targeting spacers with no signs of deactivation were limited to the leader proximal end, indicating that their acquisition followed immediate lethality. It is therefore tempting to suggest that the spacer deletion machinery was likely impaired, resulting in continued acquisitions alongside advantageous coevolving phage targeting spacers; and the continued selection pressure to evade self-targeting activity but retain phage targeting activity persisted and selected for

loss-of-function mutations in the self-targeting spacers. While this manuscript was in review, Levy et al., demonstrated that artificially induced CRISPR systems in laboratory populations of E.coli tend to exploit degradation products from the enzyme RecBCD, which processes double strand breaks resulting from replicating DNA and through the processing of exposed linear phage genomes after infection [231]. Because this bias reduces the effective number of self-targeting spacer acquisitions, this can be seen as a potential self- vs. non-self detection mechanism resulting in a relaxed constraint on CRISPR regulation. It is however crucial that the spacer deletion system is still in check so as to avoid the loss of effective antiviral spacers, thereby encouraging CRISPR maintenance in the population.

The rapidly growing empirical literature on CRISPR molecular and cellular biology will surely suggest further refinements to our model. Several avenues for model improvement are already apparent. First, the impact of the most commonly occurring alternative resistance mechanisms (such as envelope resistance) in laboratory populations was neglected. Second, our activation model where all CRISPR reactions are scaled uniformly in free cells is simplistic, as differences in activation levels among the acquisition and interference genes may occur. Third, assignment of equal autoimmunity rate constants for all the genomic protospacers is a rough approximation and it is known that the genetic sequences vary in their essentiality. Fourth, the current analytic cannot describe multiple CRISPR genotypes with diverse spacer configurations, in contrast to agent-based models [2, 212]. While we have presented some theory for explaining the maintenance of altruistic CRISPR hosts over ecological time scales, a clear characterization of factors determining their long-term evolutionary stability in well-mixed conditions continues to

be an open question (ref. Appendix B). Nevertheless, despite these simplifications, our analysis clarifies the effects of CRISPR autoimmunity in a general setting – a problem that is difficult to address experimentally, due to the lethality of self-targeting.

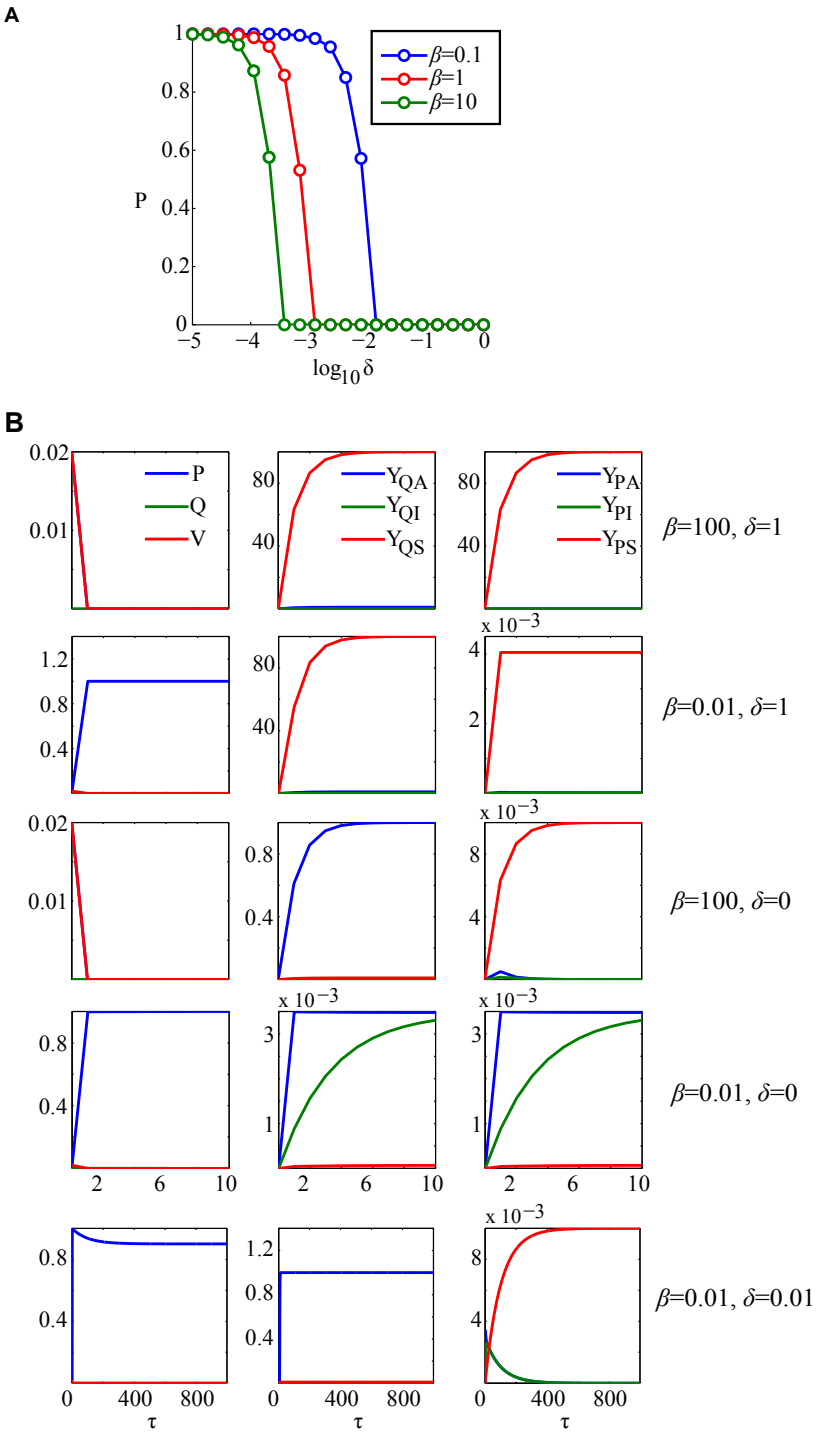

Figure 6.4: SND absence is lethal due to accumulation of self-targeting spacers. *(A) A sharp threshold-like behavior is observed with steady state prokaryotic densities in the (δ, β)space. Without a sufficient amount of CRISPR suppression in free cells, determined by δ, cells go extinct. (B) Time course trajectories of the species and spacer variables for several parameter settings. In the absence of strong regulation of auto-immunity, high host protospacer levels are extremely toxic and cause population extinction.*

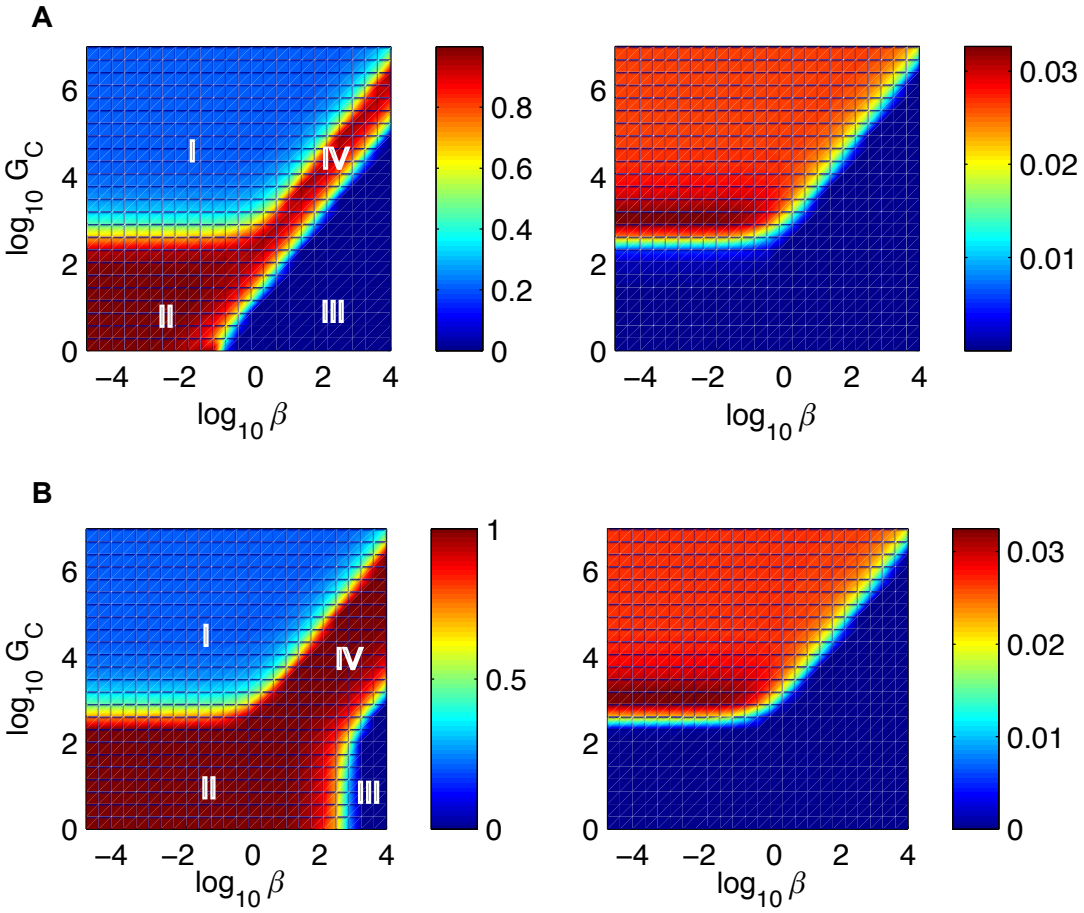

Figure 6.5: The (δ, β, G_C) ***space.*** *We plot steady state free cell densities (the first column) and the phage densities (the second column) for various values of* (β, G_C) *values under two cases: (A) moderately suppressed free-cell CRISPR activity,* $\delta = 10^{-2}$ *and (B) strongly suppressed free-cell CRISPR activity,* $\delta = 10^{-4}$*.* G_C *is the dimensionless parameter indicating the ratio of spacer deletion rate to spacer acquisition rate.* β *is a dimensionless parameter indicating the ratio of host to phage protospacer levels. (Regime I) Very high* G_C *values effectively reduce CRISPR content to very low levels (phage lysis rates are relatively overwhelming) offering no immune advantage to the hosts, resulting in free cell levels of* $\frac{G_{Q\to V}}{A_V}$*. (Regime II) Both abortive infection and immunity operate with the available intracellular steady state levels of active and self-targeting spacers. (Regime III) The constraint on* δ *is not satisfied and the hosts are extinct. (Regime IV) CRISPRs behave as full-fledged abortive infection systems exploiting only the accumulated self-targeting spacers, with phage reactive spacers eliminated due to high* G_C *values.*

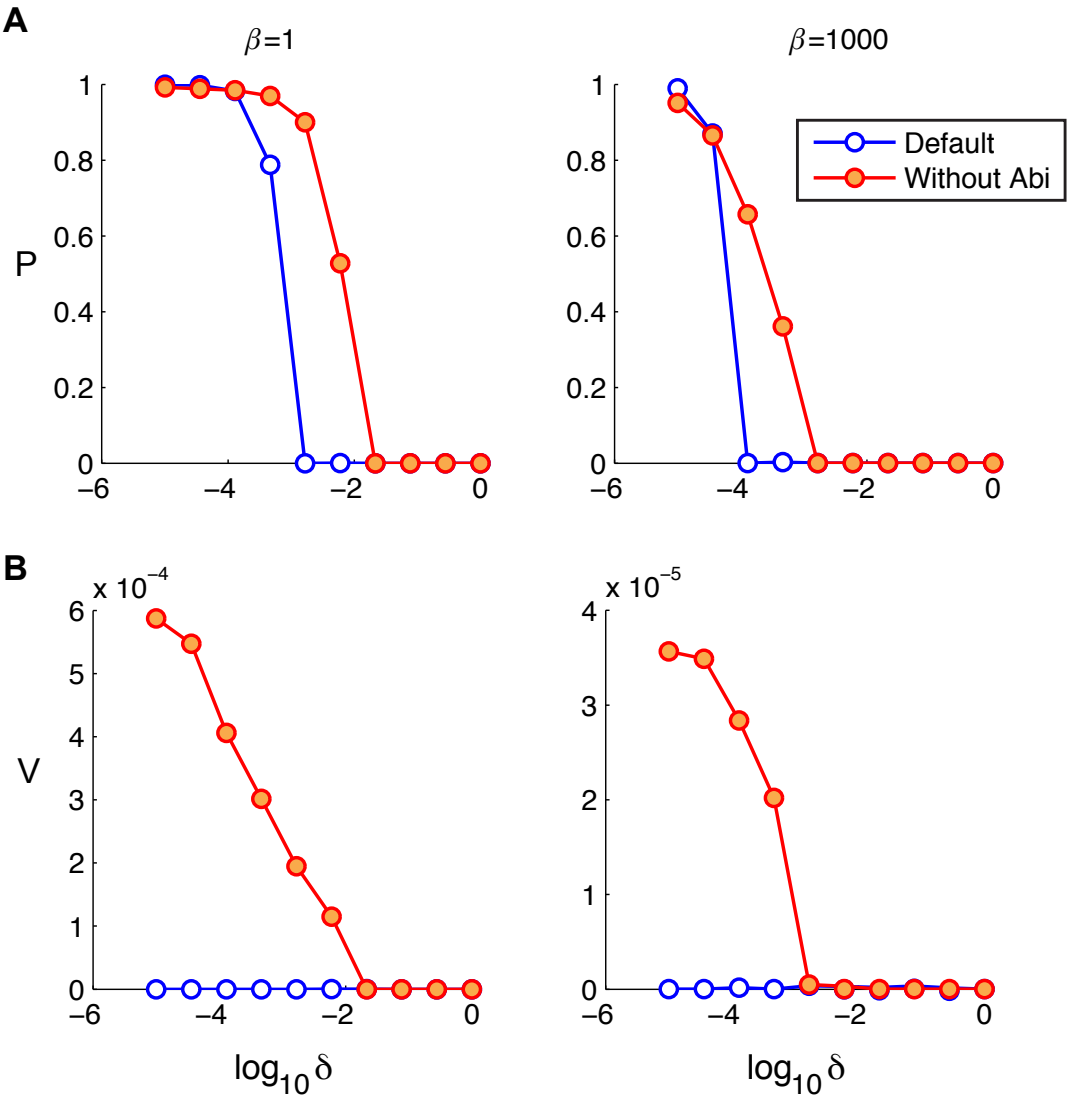

Figure 6.6: Elimination of ABI allows for improved phage densities. *Steady state densities of free cells (A) and phages (B) for various values of free-cell CRISPR activity, δ. Both coexistence and phage densities are improved without ABI. Above a critical value of δ, the system goes to extinction.*

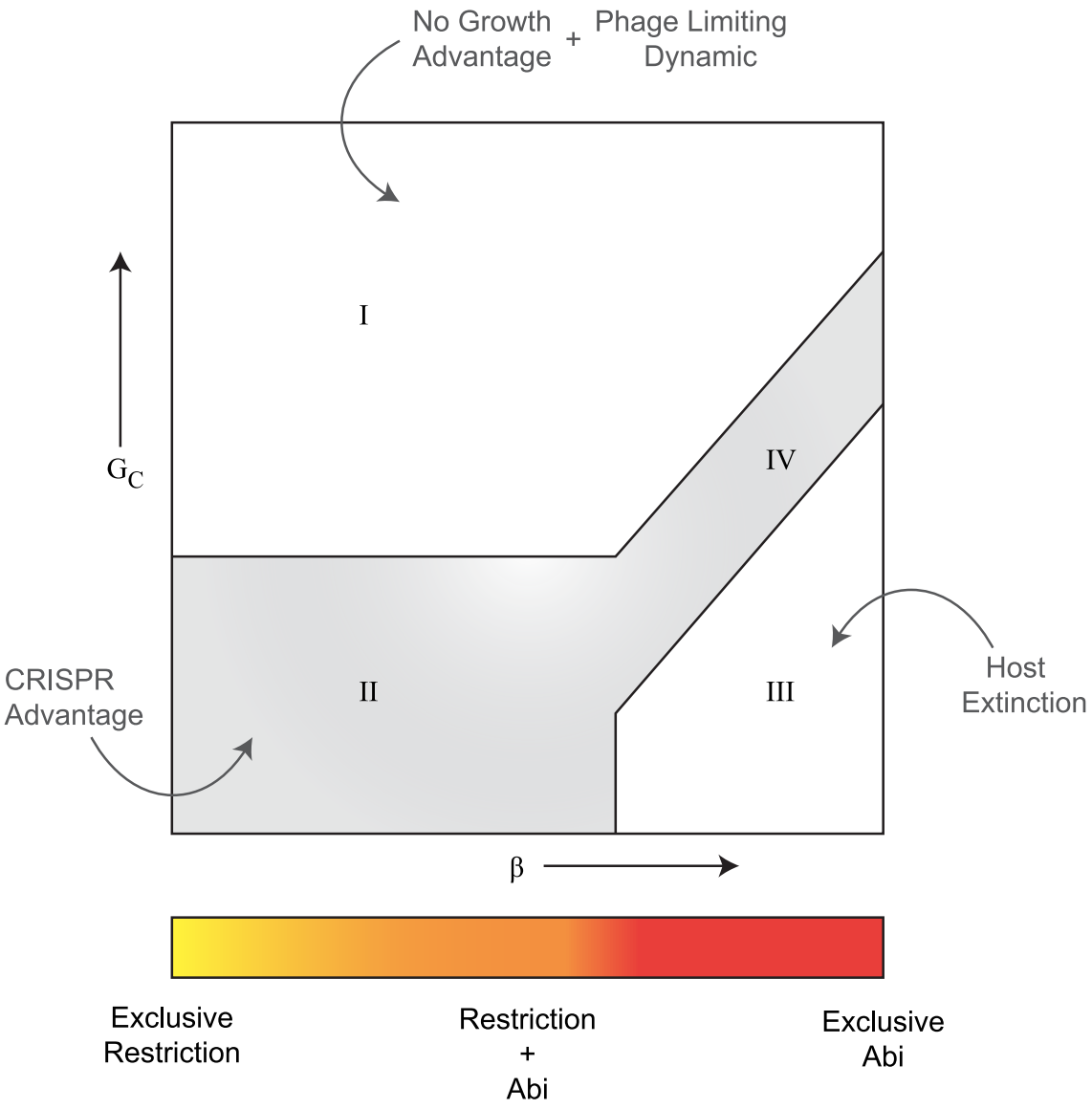

Figure 6.7: Qualitative behavior of regulated CRISPR modules. *Depending on the activation level of CRISPR activity in free cells (δ), the host to phage protospacer ratio (β), and the CRISPR specific spacer deletion to acquisition rate ratio (G_C), regulated CRISPR cassettes can fall in one of the four regimes: no advantage (regime I), advantageous to hosts by offering immune resistance and abortive infections (regime II and IV), or causing host extinction (regime III). Because per-spacer immune rates have been experimentally measured to be high, we do not study its influence specifically here. When CRISPR activity is completely repressed in free cells ($\delta = 0$), regime III vanishes, and regime IV expands into its place. Notice that a low β value corresponds to efficient SND during acquisition process.*

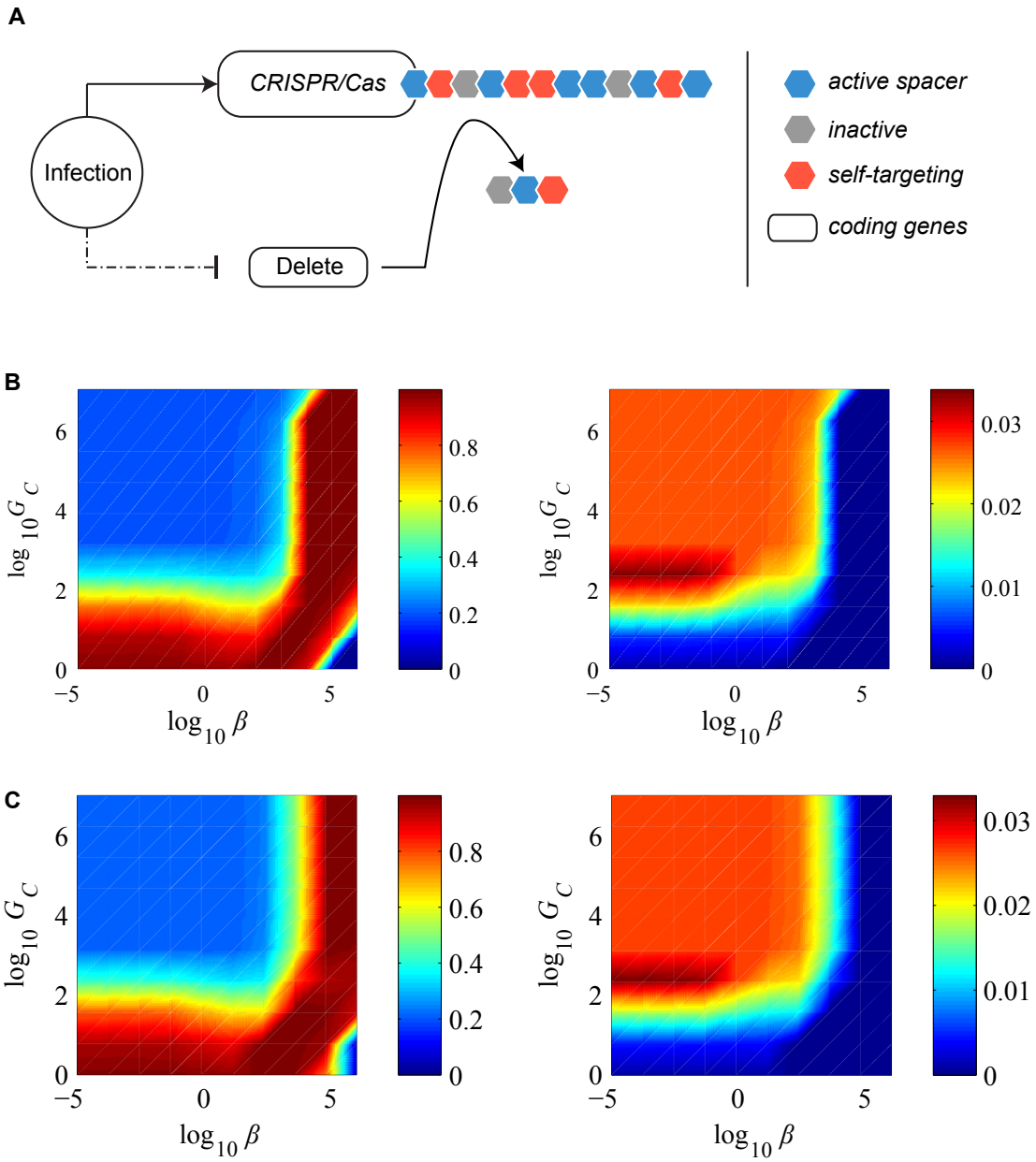

Figure 6.8: Decoupled behavior of a spacer deletion system. *(A) A schematic of the decoupled model of CRISPR regulation. Arrow indicates activation, and a blunt arrow indicates repression. The dashed arrow can be active (suppression when infected) or inactive (constitutive expression). We plot the steady state free cell densities (the first column) and the corresponding phage densities (the second column) for various values of* (β, G_C) *values at* $\delta = 10^{-4}$*. Comparison with Fig. 6.5B illustrates that decoupled spacer deletion systems as in (B) no regulation or (C) regulation in a direction opposite to that of the rest of the CRISPR system can tolerate higher host protospacer levels without requiring extremely high GC values. Note that* $\log_{10}\beta = 2$ *corresponds to* $100\times$ *the corresponding phage protospacer levels, a realistic condition in the case of E.coli vs. phage* λ*, where the expected number of host protospacers is a hundred fold.*

Part III

Appendix

Chapter 7

Ecological equivalence as a modeling strategy for metagenomic count data.

In chapter 4 of part I, we mentioned statistical inference of taxas (OTUs) observed in large-scale 16S metagenomic surveys is of considerable biological interest. The large number of taxa thus discovered (albeit with only a few dominating/abundant ones) and excess zeroes in the taxa count distributions, however, make it a challenge for performing statistical analyses.

In this appendix, we present a strategy that aims to mitigate these issues by aggregating counts of carefully chosen taxa that behave similarly to latent ecological factors and environmental processes. It is well known that the relative abundances of such ecologically equivalent/nearly-equivalent species are not necessarily influenced by changes in environmental conditions across local and regional scales, but their summed total abundance, however, is [232–235]. To this end, we aim to cluster the 16S metagenomic taxa into equivalence classes, and create a reduced dataset that presents these clusters of taxa as new units of analysis interest (termed "Equivalence Class Units") and their summed counts as new measurements across observations in an experiment. We suggest

a Bayesian nonparametric model of ecological equivalence, and establish posterior inference algorithms, for inferring equivalent classes of metagenomic 16s features, which are simply clusters of OTUs. Interesting prior probability distributions also appear which allow for both unknown number of ECUs in a dataset, and known relationships among the species to be clustered (example, a taxonomic tree).

Our approach is applicable to datasets with few thousands of species, and we demonstrate these ideas with metagenomic data arising from a few simple ecosystems. While several clusters of taxa showed significant enrichment of taxonomic identities there were also many clusters that did not demonstrate this behavior suggesting cross-taxonomic equivalence in these ecosystems. Examples illustrating the coherence of the clusters in terms of reflecting known biology are indicated. We present the models and derive the inference algorithms, before presenting these preliminary results with publicly available experimental datasets.

7.1 Model

We consider metagenomic features (OTUs) $i = 1 \dots p$, measured across different experimental conditions $g = 1 \dots G$, in samples $s = 1 \dots N_g$. Here N_g is the number of samples in group g. For now, we shall assume the number of equivalence classes (the OTU clusters we want to infer) to be fixed to K, and let $k = 1 \dots K$ index clusters. Z is a p length vector where each entry Z_i indicates which equivalence class $k \in \{1 \dots K\}$, feature i is a member of. Given a configuration of Z, X_{gskj} will denote the metagenomic count of the j^{th} feature in cluster k in sample s from experimental condition g. We let

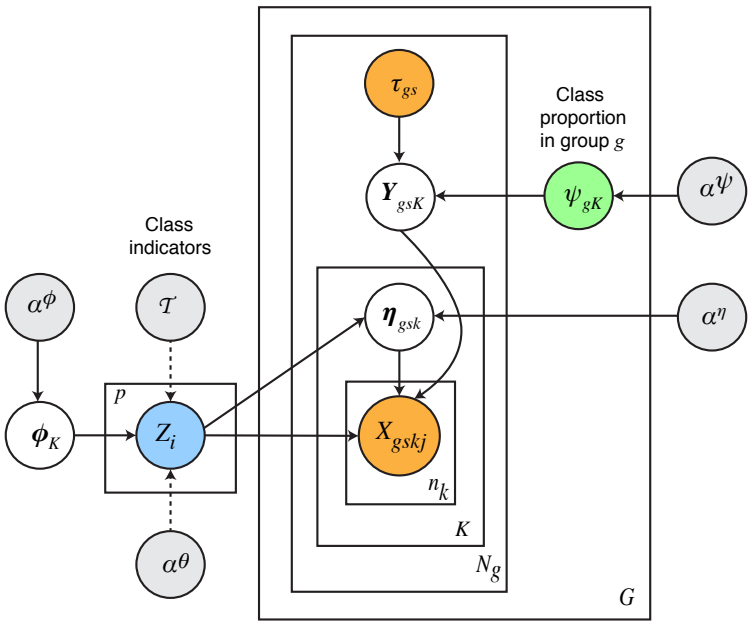

Figure 7.1: A plate model illustration of the proposed generative process underlying metagenomic counts. *The entire process is specific to every group g with N_g samples and p metagenomic features (OTUs). The total number of classes is fixed at K. The total number of OTUs in an equivalence class $k \in 1, \ldots K$ is given by n_k. Orange nodes indicate observed data, and the blue and green nodes indicate our target variables, the posterior of which is needed. The equivalence classes built are conditional on the available data, and with respect to the distinct groups a researcher has.*

$n_k = \sum_i I_{[Z_i==k]}$, where I is the indicator variable, be the total number of features assigned to equivalence class k. The abundance count corresponding to any given equivalence class k is simply defined as the summed total count of all features assigned to the equivalence class according to the configuration Z. Specifically, $Y_{gsk} = \sum_{i:Z_i=k} X_{gski}$ denotes the count of the k^{th} equivalence class in sample s from experimental condition g. We will use the $\cdot$ notation to denote vectorized quantities. So for instance, $Y_{gs\cdot}$ is a K length vector desribing the net count of K different equivalence classes in sample s from experimental condition g. We will also use $\odot$ represents a product / element-wise product (which will be clear from the context).

With this notation, we now present the baseline Bayesian hierarchical model below in 7.1, and illustrate it in Fig. 7.1. Our goal is to infer Z. We leave the prior on Z unspecified for now. We shall first derive the likelihood of the data, conditioned on the cluster assignments and the parameters. We then consider two distinct priors in the subsequent sections, and derive the resulting posteriors in each case.

$$
\begin{aligned}
Z_i|\alpha^\theta \ldots &\sim p(Z_i|\alpha^\theta, \ldots) \leftarrow \text{ prior on equivalence class assignments for each OTU} \\
\psi_{g\cdot}|K, \alpha^\psi &\sim Dirichlet(\alpha^\psi \odot 1_K) \leftarrow \text{prior on relative abundances of K equivalence classes} \\
Y_{gs\cdot}|\psi_{g\cdot} &\sim Multinomial(\tau_s, \psi_{gK}) \leftarrow \text{total summed abundance of equivalent species.} \\
\eta_{gsk}|Z, \alpha^\eta &\sim Dirichlet(\alpha^\eta \odot 1_{n_k}) \leftarrow \text{ models drift} \\
X_{gsk\cdot}|Z, \eta_{gsk}, Y_{gsk} &\sim Multinomial(Y_{gsk}, \psi_{gk} \times \eta_{gsk}) \leftarrow \text{observed data}
\end{aligned}
\tag{7.1}
$$

For convenience, we describe the dimensions of the variables in the model above. Z is a p length vector, each entry i holding the equivalence class of feature i. $\psi_{g\cdot}$ is K length equivalence class proportions vector, where the k^{th} entry describes the proportion of equivalence class k of $1 \ldots K$ classes. $Y_{gs\cdot}$ is a K length vector describing the count of each of the K equivalence classes in a sample s from condition g. η_{gsk} is an n_k length drift proportions vector where each j^{th} entry models assigns a proportion for the j^{th} member feature of cluster k in sample s from condition g. Finally, $X_{gsk\cdot}$ is an n_k length vector describing the count of all member features of cluster k in sample s from condition g.

7.2 Data likelihood

We now derive the likelihood $p(X|Z,\psi,\eta)$.

First, we shall restrict ourselves to describing the conditional likelihood of the feature count data in a single sample s from an experimental condition g. The key is to first observe from the last line in the generative process 7.1 that, the feature count data for any given equivalence class (cluster) k in sample s from condition g follows:

$$X_{gsk\cdot}|Z,\psi,\eta \sim Multinomial(\tau_{gs}, \psi_{gk} \times \eta_{gsk}) \tag{7.2}$$

So the entire feature count data for a given sample s in group g has the conditional distribution:

$$X_{gs\cdot\cdot}|Z,\psi,\eta \sim Multinomial(\tau_{gs}, \psi_{g\cdot} \otimes \eta_{gsk}) \tag{7.3}$$

Here $\otimes$ is used to denote a Hadamard product as follows: each entry of the K length equivalence class proportions vector $\psi_{g\cdot}$, ψ_{gk} multiplies the corresponding n_k length drift proportion vector η_{gsk}. The result is a p length vector of feature proportions that describe the average relative abundance of the feature count data in sample gs.

Writing eqn. 7.3 explicitly, we observe:

$$
\begin{aligned}
p(X_{gs\cdot\cdot}|Z,\psi,\eta) &= \frac{\tau_{gs}!}{\prod_{k=1}^{K}\prod_{j=1}^{n_k} X_{gskj}!}\prod_{j=1}^{p}\left(\psi_{gk}\eta_{gskj}\right)^{X_{gskj}} \\
&= \frac{\tau_{gs}!}{\prod_{k=1}^{K} Y_{gsk}!\prod_{j=1}^{n_k} X_{gskj}!}\prod_{k=1}^{K} Y_{gsk}!\prod_{j=1}^{n_k}\left(\psi_{gk}\eta_{gskj}\right)^{X_{gskj}} \\
&= \frac{\tau_{gs}!}{\prod_{k=1}^{K} Y_{gsk}!}\prod_{k=1}^{K}\left(\psi_{gk}\right)^{Y_{gsk}}\frac{Y_{gsk}!}{\prod_{j=1}^{n_k} X_{gskj}!}\prod_{j=1}^{n_k}\left(\eta_{gskj}\right)^{X_{gskj}} \\
&= \frac{\tau_{gs}!}{\prod_{k=1}^{K} Y_{gsk}!}\prod_{k=1}^{K}\left(\psi_{gk}\right)^{Y_{gsk}}\prod_{j=1}^{n_k} Multinomial(X_{gsk\cdot}|Y_{gsk},Z,\eta_{gsk})
\end{aligned}
$$

In the above derivation, $Y_{gsk} = \sum_{j=1}^{n_k} X_{gskj}$, where the entries in X_{gskj} are organized based on the equivalence class membership vector Z. Now, in a straightforward fashion, we can integrate out η_{gsk} yielding the likeilihood distribution per sample:

$$
\begin{aligned}
p(X_{gs\cdot\cdot}|Z,\psi,\alpha^{\eta}) &= \int_{\eta_{gsk}} \frac{\tau_{gs}!}{\prod_{k=1}^{K} Y_{gsk}!}\prod_{k=1}^{K}\left(\psi_{gk}\right)^{Y_{gsk}}\prod_{j=1}^{n_k} Multinomial(X_{gsk\cdot}|Y_{gsk},Z,\eta_{gsk})p(\eta_{gsk}|\alpha^{\eta})d\eta_{gsk} \\
&= \frac{\tau_{gs}!}{\prod_{k=1}^{K} Y_{gsk}!}\prod_{k=1}^{K}\left(\psi_{gk}\right)^{Y_{gsk}}\int_{\eta_{gsk}}\prod_{j=1}^{n_k} Multinomial(X_{gsk\cdot}|Y_{gsk},Z,\eta_{gsk})p(\eta_{gsk}|\alpha^{\eta})d\eta_{gsk} \\
&= \frac{\tau_{gs}!}{\prod_{k=1}^{K} Y_{gsk}!}\left[\prod_{k=1}^{K}\left(\psi_{gk}\right)^{Y_{gsk}} DM(X_{gsk}|Y_{gsk},\alpha^{\eta})\right]
\end{aligned}
$$

where $DM(X_{gsk\cdot}|Y_{gsk},\alpha^{\eta})$ represents a Dirichlet-Multinomial distribution of vector X_{gsk} with n_k features of total count Y_{gsk}, and concentration parameter $\alpha^{\eta}\odot 1_{n_k}$. Here 1_{n_k} represents an n_k length vector of 1s.

The condition data likelihood, for all independent samples from all experimental

conditions $g = 1 \ldots G$, is then given as:

$$
\begin{aligned}
p(X|Z,\psi,\alpha^{\eta}) &\propto \prod_{g=1}^{G}\prod_{s=1}^{N_g}\prod_{k=1}^{K}\left(\psi_{gk}\right)^{Y_{gsk}} DM(X_{gsk}|Y_{gsk},\alpha^{\eta}) \\
&= \prod_{g=1}^{G}\prod_{k=1}^{K}(\psi_k)^{\sum_{s=1}^{N_g} Y_{sk}} \prod_{s=1}^{N_g} DM(X_{sk}|Y_{sk},\alpha^{\eta}) \\
&= \prod_{g=1}^{G}\prod_{k=1}^{K}\left(\psi_{gk}\right)^{\sum_{s=1}^{N_g} Y_{gsk}} \prod_{s=1,\ Y_{gsk}>0}^{N_g} DM(X_{gsk\cdot}|Y_{gsk},\alpha^{\eta})
\end{aligned}
\tag{7.4}
$$

The last line arises because $Y_{gsk} = 0 \implies X_{gskj} = 0\ \forall j = 1 \ldots n_k$, and therefore, the corresponding Dirichlet-Multinomial evaluates to 1. This is seen easily using the Gamma representation of a Dirichlet Multinomial distribution.

7.3 Posteriors for ψ and Z

We shall now derive the posterior for the two classes of variables of interest. The first class of variables is the equivalence class proportions vector for each experimental group g, $\psi_{g\cdot}$, whose posterior is based on a standard Dirichlet-Multinomial hierarchy. The second variable of interest is the equivalence class membership vector Z, whose posterior is based on priors derived from relational data among the metagenomic features to be clustered. Additionally, we present the Z posteriors in both cases of fixed and unknown number of clusters (K).

7.3.1 Conditional posterior for ψ

While the cluster proportions vector ψ can in principle also be integrated out, our inferential interest centers also in the posterior estimate of ψ. We use the likelihood

prescribed by eqn. (7.4) to derive the conditional posterior of $\psi_{g\cdot}$ given the rest of the variables and parameters.

$$
\begin{aligned}
p(\psi_{g\cdot}|Z,X,...) &\propto p(X|Z,\psi,...)p(\psi|\alpha^{\psi}) \\
&\propto \left[\prod_{g=1}^{G}\prod_{k=1}^{K}\left(\psi_{gk}\right)^{\sum_{s=1}^{N_g} Y_{gsk}} \prod_{s=1,\ Y_{gsk}>0}^{N_g} DM(X_{gsk}|Y_{gsk},\alpha^{\eta})\right] p(\psi|\alpha^{\psi}) \\
&\propto \prod_{k=1}^{K}\left(\psi_{gk}\right)^{\sum_{s=1}^{N_g} Y_{gsk}}\ p(\psi|\alpha^{\psi}) \\
&\propto \prod_{k=1}^{K}\left(\psi_{gk}\right)^{\sum_{s=1}^{N_g} Y_{gsk}}\ \psi_{gk}^{\alpha_k^{\psi}-1} \\
&\propto \prod_{k=1}^{K}\left(\psi_{gk}\right)^{\sum_{s=1}^{N_g} Y_{gsk}+\alpha_k^{\psi}-1} \\
&\equiv Dirichlet(\psi_{g\cdot}|(\alpha^{\psi}\odot 1_{n_k})+Y_{g\cdot\cdot})
\end{aligned}
\tag{7.5}
$$

where the last line describes the parameter of the posterior Dirichlet distribution as an n_k length vector of α^{ψ} added to the total count of each of the K components across all samples s from group g.

Thus, we arrive at the posterior of one of the two target variables we are interested. We next derive the posterior for the equivalence class feature assignments vector Z.

7.3.2 Conditional posteriors for Z

To derive the posterior for Z, we will first need to specify its prior distribution. We take two routes below. The first is a standard move in Bayesian hierarchical clustering based on a Dirichlet distribution for component membership probabilities. The second is a more natural route for modeling metagenomic microbial features, a prior based on

taxonomic trees. It reflects the prior belief that metagenomic features belonging to similar taxonomic categories should are part of the same equivalence class.

In each of the above two cases, we first provide the posterior when the number of equivalence classes in the data K can be considered as known, fixed quantity. We then provide a non-parametric extension, which assumes a small, but unknown number of clusters in the data. This is based on the theory of infinite mixture models [236].

Case 1: Classic Dirichlet priors

Known K Suppose in our baseline model (7.1), we assume the following prior distribution for $p(Z|\dots)$.

$$\begin{aligned} \phi_K|\alpha^\phi, K &\sim Dirichlet(\alpha^\phi \odot 1_K) \\ Z_i|\phi_K &\sim Multinomial(\phi_K, n=1)\ \forall i = 1\dots p \end{aligned} \tag{7.6}$$

which leads to the conditional prior:

$$p(Z_i = k|Z_{\setminus i}, \alpha^\phi) = \frac{n_{k\setminus i} + \alpha_k^\phi}{\sum_k \alpha_k^\phi + p - 1} \tag{7.7}$$

The posterior for Z_i is then given using standard calculations as:

$$p(Z_i|Z_{\setminus i}, X, \psi, \dots) \sim p(X|Z, \dots)\left[n_{k\setminus i} + \alpha^\phi\right] \tag{7.8}$$

where the likelihood term is given by (7.4), and $Z_{\setminus i}$ denotes the membership vector of all features except the i^{th} feature whose membership gets sampled with the above

posterior. Similarly, $n_{k\backslash i}$ denotes the number of features assigned to clsuter k except the i^{th} feature. We have also abused the notation above by mentioning α_k^{ϕ} to indicate the k^{th} component of the Dirichlet prior parameter, which is derived in eqn. 7.6 as a K length vector of a single value α^{ϕ} repeated.

Unknown K. A non-parametric extension We now want to generalize to arbitrary K, with the idea being that we would like to generate a small number of clusters. Consider $\tilde{\alpha}^{\phi} = [|\alpha^{\phi}/K|]_{k=1}^{K}$. Notice that a Dirichlet prior with a paramer < 1 for all entries favors fewer categories. We consider the limit $K \to \infty$. Then the conditional prior in eqn. 7.7 becomes:

1. For sampling a represented equivalence classes with at least one OTU:

$$p(Z_i = k|\mathbf{Z}_{\backslash i}, \tilde{\alpha}^{\phi}) = \frac{n_{k\backslash i}}{\sum_k \tilde{\alpha}_k^{\phi} + p - 1}$$

2. For creating a new equivalence class: (7.9)

$$p\left(Z_i = k^{new} \cap \left[\, k^{new} \neq Z_j \;\forall\; j \neq i\right] |\mathbf{Z}_{\backslash i}, \tilde{\alpha}^{\phi}\right) = \left(1 - \sum_{k, k \neq k^{new}}^{K} p(Z_i = k|\mathbf{Z}_{\backslash i}, \tilde{\alpha}^{\phi})\right)$$

$$= \frac{\alpha^{\phi}}{\sum_k \tilde{\alpha}_k^{\phi} + p - 1}$$

When applying the above non-parametric prior to derive the posterior for Z, we do not need to worry about the divergence of the inner sum in the prefactor $\Gamma(\sum_{k=1}^{K} \alpha_k^{\phi})$ terms in the $Dirichlet - Multinomial$ likelihood, as the inner sum (from our definition of α^{ϕ} above) is always given by $\tilde{\alpha}^{\phi}$. This mathematical convenience for convergence is reflected as the "sparse number of custers" assumption above.

Case 2: Tree priors for equivalence class memberships Often times, a metagenomic data analyst has additional relational information about the metagenomic features that one may wish to account for in the above clustering procedure. For instance, genomically closely related OTUs share the same cluster component. These similarlities are reflected in the taxonomic relationships among the microbial features or the edit-distances among the 16S RNA sequences themselves. These measures render themselves conveniently for a tree representation. The nodes $\mathscr{T}$ and α_θ in Fig. 7.1 precisely correspond to a process that accounts for such a prior. If not immediately available, such a tree can be constructed using the *Cho-Liu/Edmond's* algorithm based on other relational data among the metagenomic features. We consider prior distribution generated by the model in Fig. 7.2 below.

For any given undirected(/directed) tree prior, we would like a given taxon at the leaf assume a cluster membership similar to other taxons closer to it in the tree. We consider the following generative model, a simplified caricature of an evolutionary process:

$$\begin{aligned} p(Z_i = k | w_i, \mathscr{P}_i, \phi) &\sim Discrete(\phi_{i\, w_i}) \\ w_i | \mathscr{P}_i &\sim Discrete\left([|\theta_{ih} \prod_{u=1}^{h-1} (1-\theta_{iu})|]_{h=1}^{|\mathscr{P}_i|} \right) \\ \theta_{ih} &\sim Beta(a,b) \; \forall \; h \in \mathscr{P}_i \\ \phi_{ih} &\sim Dirichlet(\alpha^\phi \odot 1_K) \; \forall \; h = 1 \ldots H \end{aligned} \tag{7.10}$$

For a given taxon i, and a prior tree, there exists a path $\mathscr{P}_i$ from it to the root of the tree. For all internal nodes that lie in the path (denoted as $h \in \mathscr{P}_i$), we associate a multinomial parameter ϕ_{ih}, and a Bernoulli switching parameter θ_{ih}. The taxon i sends a

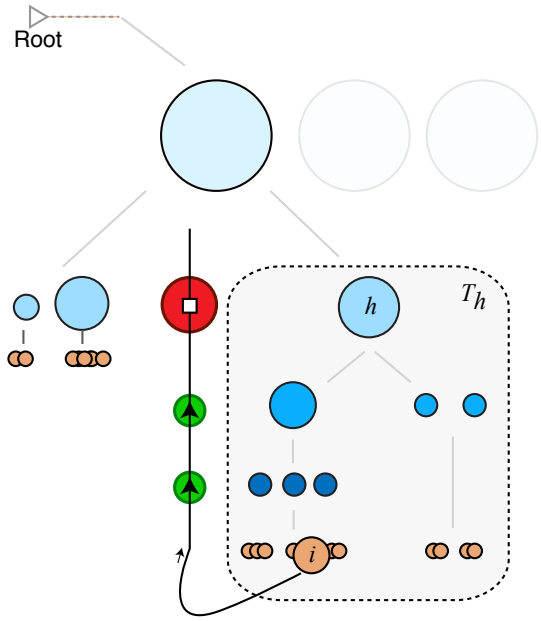

Figure 7.2: Prior distributions based on a tree of relationships among taxa. *For a given taxon i, a given tree of relationships implies a particular distance metric between the chosen node and all others. We consider a rooted tree of relationships among the taxa to be clustered. All the taxa are positioned in the leaves, and are colored brown. This is very similar to a phylogenetic tree. Blue nodes indicate distinct internal nodes in the tree, the deepness of color indicate the posititon/height level in the tree. Each internal node h is associated with a specific Dirichlet-Multinomial probability distribution ϕ_h of size K for components. Every taxon i is at the leaft the tree, and there exists a path $\mathscr{P}_i$ from it to the root that passes through a subset of internal nodes. A given taxon i chooses its cluster membership according to the ϕ_h prescribed by an $h \in \mathscr{P}_i$ node that it randomly chooses to stop at (red signal), as it visits each internal node $h \in \mathscr{P}_i$ from the bottom up (green signal). Such tree based priors closely reflects the commonly available tree of relationships in metagenomics, and to an extent, evolutionary divergence in real world systems.*

variable w_i up the tree which chooses to stop at $h \in \mathscr{P}_i$ with probability θ_{ih} (or jump one level up with probability $1-\theta_{ih}$). Stopping at a level h (if no internal $h \in \mathscr{P}_i$ was chosen, it stops at the root of the tree), a taxon chooses to derives its component membership (Z_i) according to ϕ_{ih} prescribed by node $h \in P_i$ with the Geometric probability $\theta_{ih}\prod_{u=1}^{h-1}(1-\theta_{iu})$. According to the above model then, on average, at a given level h, this product is close to $\left(\frac{a}{a+b}\right)\prod_{u=1}^{h-1}(1-\frac{a}{a+b})$. If $a > b$, this product $\longrightarrow 0$ as h grows, thus capturing our prior belief.

In summary, taxa decide on an internal node to sample their component assignments from, and conditioned on these choices, the entire vector of component assignments Z is generated independently across all these internal nodes, as prescribed by the Multinomial mixture at each node.

Posterior Equations for Z and W We can integrate out ϕ_{ih}, and θ_{ih} yielding:

$$p(Z|w,\mathscr{P}) \sim \prod_{h=1}^{H} DM(Z^{(h)}|\alpha^{\phi} \odot 1_{n_{kh}})$$

$$w_i|\mathscr{P}_i \sim Discrete\left(\left[|\left(\frac{a}{a+b}\right)\prod_{u=1}^{h-1}\left(1-\frac{a}{a+b}\right)|\right]_{h=1}^{|\mathscr{P}_i|}\right) \;\forall\, i = 1\ldots p$$

Here, *DM* is the Dirichlet-Multinomial distribution, $Z^{(h)}$ is the vector of taxons that derive their membership from internal node h, n_{kh} is the total number of taxon for which the component indicator is k, and derived with probability at level h. Notice hs in the set of internal nodes from which no taxon derives its membership from can be safely ingored as the DM simply evaluates to 1. Also notice, while in principle marginalizing over w is possible, the posterior for Z is easy to sample conditioned on w. We therefore consider

sampling both variables.

At a given level h in path, a single Dirichlet-Multinomial determines component assignments for all taxa that sample their component assignments from that level. Thus, we can straightforwardly write the conditional prior as:

$$p(Z_i = k|w_i = h, Z_{\setminus i}, w_{\setminus i}, \mathscr{P}, \alpha^{\phi}) = p(Z_i = k|w_i = h, Z_{\setminus i}^{(h)}, w_{\setminus i}^{(h)}, \mathscr{P}, \alpha^{\phi}) \propto n_{kh\setminus i} + \alpha^{\phi}$$

The posterior for w becomes:

$$\begin{aligned} p(w_i = h|Z,X) &\propto p(X|Z,w_i)p(Z|w_i)p(w_i) \\ &\propto p(Z|w_i)p(w_i) \\ &\propto \left(\frac{a}{a+b}\right)\prod_{u=1}^{h-1}\left(1-\frac{a}{a+b}\right)\left[\frac{n_{kh\setminus i}+\alpha^{\phi}}{n_{kh}\alpha^{\phi}+n_{kh}-1}\right] \\ &\propto \prod_{u=1}^{h-1}\left(1-\frac{a}{a+b}\right)\left[\frac{n_{kh\setminus i}+\alpha^{\phi}}{n_{kh}\alpha^{\phi}+n_{kh}-1}\right] \end{aligned} \tag{7.11}$$

The posterior for $Z|w$ becomes:

$$\begin{aligned} p(Z_i = k|X, Z_{\setminus i}, w_{\setminus i}, w_i = h) &\propto p(X|Z)p(Z_i|w_i, Z_{\setminus i}) \\ &\propto p(X|Z)\left[n_{kh\setminus i}+\alpha^{\phi}\right] \end{aligned} \tag{7.12}$$

Notice the trade off in these posterior equations. Suppose if a given taxon is truly from component k; while the posterior is proportional to the number of features that arise from the component, as we go higher up in the tree, the number of features that arise from the same component is likely to grow, giving rise to higher posterior probablities for $Z_i = k$. However, our prior on w_i can force the taxon to stay low in the tree.

Incorporating phylogenetic distance If one wants to take branch lengths in a (phylogenetic) tree into account, a simple and straightforward model would be to model the up-wise jumping probabilities $(1-(a/(a+b)))$ terms for each level h in the the w_i posterior equation above, as a decreasing function of the total distances spanned upto level $h+1$. This can be done by making them a solution of $\log\left(\frac{1-\theta_{ih}}{\theta_{ih}}\right) = -\delta_{h,h+1}$, where $\delta_{h,h+1}$ is the distance spanned from internal node h to $h+1$ both $\in \mathscr{P}_i$. With this expression, the posterior for w becomes:

$$\begin{aligned} p(w_i = h|Z,X) &\propto (\theta_{ih}) \prod_{u=1}^{h-1} (1-\theta_{ih}) \left[\frac{n_{kh\setminus i} + \alpha^{\phi}}{n_{kh}\alpha^{\phi} + n_{kh} - 1}\right] \\ &\propto \prod_{u=1}^{h-1} (1-\theta_{ih}) \left[\frac{n_{kh\setminus i} + \alpha^{\phi}}{n_{kh}\alpha^{\phi} + n_{kh} - 1}\right] \end{aligned} \tag{7.13}$$

– *Non-parametric extension* Solution is straightforward and looks similar to that derived in the Dirichlet case of previous section, with $n_{k\setminus i}$ replaced by $n_{kh\setminus i}$.

7.4 Applications

The non-parametric inference algorithm with the tree prior was applied to the mouse microbiome data [152] with roughly 1600 taxa. Relative to a standard Dirichlet-Multinomial model for the count data, the equivalence class model lead to a $> 300X$ increase in the data likelihood. Roughly 110 equivalent clusters were found, and some were found to be differentially abundant between cases (mice fed a "western" diet) and controls (mice fed a plant based "BK" diet). While the use of a taxonomy tree prior for the metagenomic features lead to more stable enrichments of taxa among the clusters identified (Fig. 7.3), this was not always the case, suggesting that ecologically equivalent clusters need not be

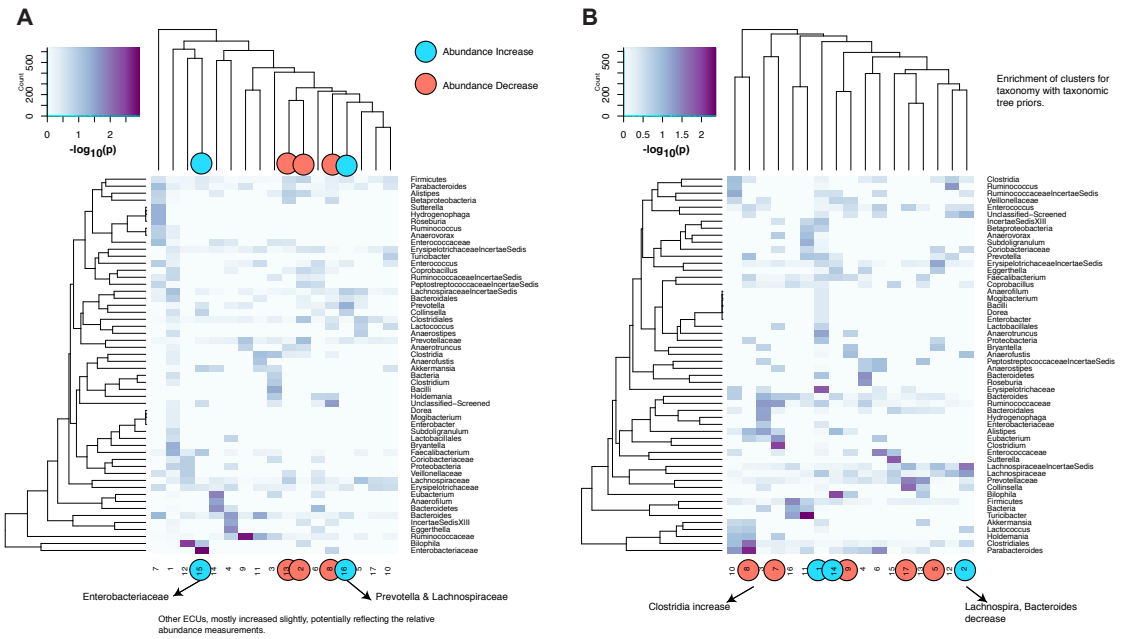

Figure 7.3: Tree priors improve taxonomy enrichments. *$-log_{10}(p-value)$ from Fisher exact tests for taxonomic categories in each EC without (A) and with tree priors (B).*

taxonomically closely related.

Applying the non-parametric inference algorithm with the tree prior to the Tara Oceans mirobiome data [8], about 450 equivalent clusters were found across the different *oceans* categories. Even though the clusters were built based on the different *oceans* categories alone, they recapitulated several OTU level properties. For instance, as shown in Fig. 7.4, given the general negative correlation between temperature and pressure in ocean layers, OTUs that were found to correlate positively with pressure, showed negative correlation with temperature. This behavior was retained when summarizing at the level of ECs as well.

We next sought to identify ECs, with interesting community level properties. The Tara oceans dataset [8] also has sample-specific relative abundance information on the en-

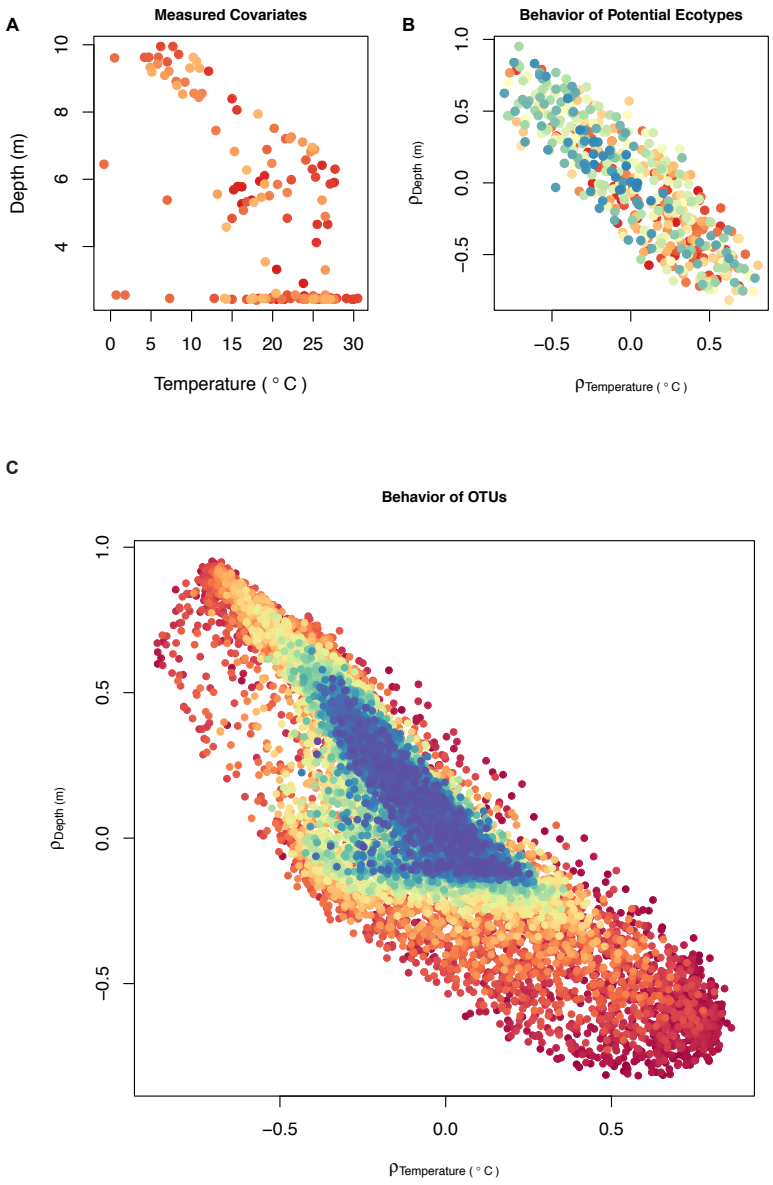

Figure 7.4: Equivalence classes capture environmental gradients. *(A) The depth and local temperature measurements for the various Tara Ocean's samples. (B) A plot of the correlations of each EC's abundance profile with depth and temperature. (C) Same as B, except the correlations are computed with OTU abundance profiles. We observe that ECs recapitulate the anti-correlations in temperature and pressure, similar to OTUs.*

coded functional/gene content (e.g., various ion transport channels, glycolysis pathways) by the sample's microbiome. So after ranking the ECs based on the p-values obtained from a differential abundance test [124] across oceans, we chose ECs that showed strong correlations to at least one of the functional categories. For each such chosen EC, we built a hierarchical clustering tree of the constituent OTUs, split the EC to finer clusters depending on the tree topology manually, and visualized the count profiles for each of these ECs across various categories samples. In several cases, we not only found a coherent set of OTUs that behave equivalently in their count profiles, but they also had sound potential for generating testable hypotheses. For instance, as illustrated in Fig. 7.5, cluster EC163 was depressed in its relative abundance in the samples from (deep ocean) mesopellagic layer, but had comparable relative abundances in the samples from surface (SRF) and deep chlorophyll maximum (DCM) oceanic layers. The EC was highly correlated with the relative abundances of the Iron (III) transport systems, which also exhibited higher abundances in DCM and SRF layers compared to the samples from MES layer across oceans. The EC was a multi-phyla cluster with member taxa from 9 different Phyla, all consisting of members with known roles in Iron metabolism and chelation with the aid of iron transport channels. In particular, several Proteobacteria is known to convert Iron(II) to Iron (III), which are further metabolized by the other members of phyla associated with EC163. These associations and correlations lead us to hypothesize that the decrease in Proteobacteria (for e.g., in MES) leads to reduced levels of circulating Iron(III); this is limiting for the growth of strains with a higher preference for Iron(III). The MES relative abundance of the EC163's Deferribacteres and Proteobacteria member taxa could explain the observed reduction in the iron (III) transport channels' relative abundances.

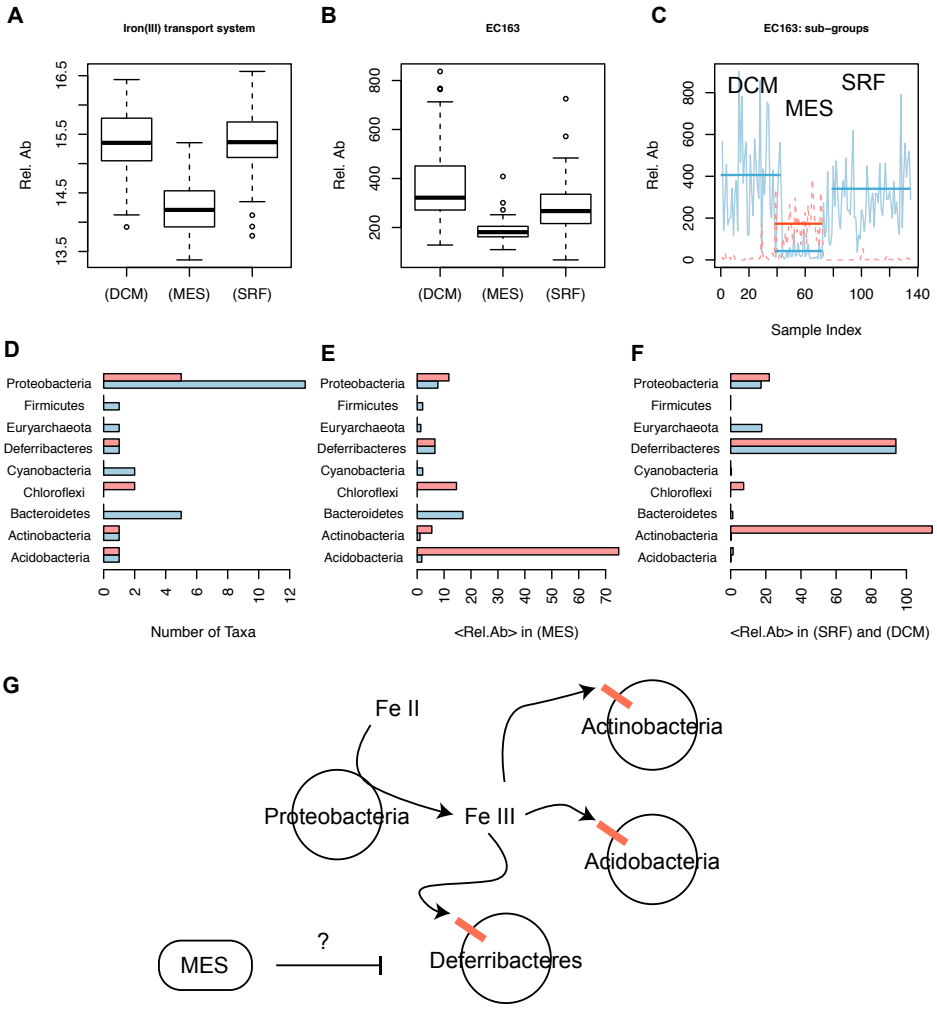

Figure 7.5: Equivalence classes of OTUs as better hypotheses generators. *(A,B) The count profile of EC163, a cluster of 16S OTUs, was found to highly correlate with Iron(III) transport system frequencies from the whole metagenome shotgun sequencing from the Tara Oceans project [8]. (C) Observing the count profiles of the two distinct subgroups within this cluster showed the coherent average abundance changes of distinct phyla in different depth layers of the oceans surveyed in the experiment. (D,E,F) indicate the distribution and relative abundances of the EC163 member taxa. (G) a potential hypothesis to explain these results.*

As future work, we aim to integrate appropriate statistics on convergence of the Gibbs samplers, introduce additional structure in the model to constrain clustering choices, and incorporate compositional correction factors to offer some prection for technical variation. While the inferences and hypotheses generated above appear meaningful, given the discussions in part I of this thesis, the fact that such inferences were based only on relative abundance information does not offer much confidence in them, and additional evidence must be sought. With rapidly growing dataset sizes, compute time will limit the application of the proposed approach, and faster algorithms based on other approximate inference methods (like variational inference and stochastic gradient descent) will be preferable.

Chapter 8

Evolutionary invasion analysis of altruistic post-infection suicidal genotypes in a well-mixed epidemiological model.

In this chapter, we sketch the conditions for the evolution of altruistic, post-infection suicidal mechanisms in a simplified well-mixed epidemiological model, using an adaptive dynamics approach. The derivation serves to illustrate the difficulty of explaining the surprising evolutionary origins, and the continued maintenance of an altruistic trait in well-mixed systems, in general.

As with chapter II, our inspiration for analysis arises from the following set of observations in microbiology. Prokaryotic Toxin/Anti-Toxin (TA) and Abortive Infection (Abi) defense systems work against parasite invasion by inducing dormancy/cellular suicide following infection [189, 205, 237]. To avoid excess population loss when uninfected, such systems are regulated so that their activation is restricted to infected populations [238]. Given the widespread occurrence and the high effectiveness of phage targeting machinery that clear a variety of phage infections without inducing host cell death (like that of restriction-modification (R-M) and CRISPR systems) [239,240] or the widely occurring envelope resistance mechanisms [241–243], the prevalance of suicidal defense

systems in prokaryotes is surprising [190, 237, 244]. Indeed, their complete absence in most endosymbionts is perhaps a testament to their associated fitness cost [244]. Thus the search for an explanation for their evolutionary origins and ecological maintenance in natural prokaryotic populations is quite interesting.

Several investigations have previously addressed the evolution of altruistic defense systems by incorporating epidemiological feedback in the context of evolutionary game theory or agent-based simulation models and extensively concluded that spatial structure is necessary to allow for the stable evolution of altruistic hosts [245–249]. Favorable spatial constraints can limit the dispersal of infectious agents, and post-infection suicide of altruistic hosts can reduce the local densities of infectious agents. Such effects can ultimately reduce the propensity of infections among locally dense altruists, providing the hosts with a fitness advantage through selective assortment [250]. Similar results for the evolution of altruism have been proposed in social game theory [251–253]. Kin selection and inclusive fitness theory offer another route [254, 255].

Epidemiological Model

We consider a simplified epidemiological model, with extensive similarities to the CRISPR model analyzed in Part II. Here, susceptible hosts (with density S) grow at a rate of b under a carrying capacity constraint of K, the intensity of which is measured by α. Susceptible hosts acquire infections from infecteds (with density I) at a rate of β. Infected hosts clear infections through background resistance mechanisms at a rate of ρ. The background host mortality rate is assumed to be γ. Infected hosts additionally undergo suicide

at a rate of ξ, and susceptible hosts regulate this autoimmunity through a suppression factor of $0 \leq \delta \leq 1$: a value of 1 implies no difference in autoimmunity rates between the susceptible and infected host states, while a value of 0 implies complete suppression in the susceptible host state. The excess mortality induced by the parasite (virulence) is indicated by λ. Denoting $\kappa = \frac{\alpha}{K}$, we can write the following non-dimensional system for the resident population (in which the parameters and variables are rescaled accordingly but we preserve the same notation):

$$\begin{aligned} \dot{S} &= b\left[1-\kappa(S+I)\right]S+\rho I-\beta SI-(\delta\xi+\gamma)S \\ \dot{I} &= \beta SI-(\rho+\lambda+\xi+\gamma)I \end{aligned} \tag{8.1}$$

8.1 Evolutionary Dynamics

As a defense mechanism, both regulation by suppression and abortive infection mediated resistance can be costly to a host [188, 213, 256–264].

Our goal is to perform a very simplified analysis and get a few insights on mechanisms that could lead to the evolution of altruist defense. In the above model, let us consider the evolution of altruist defense (ξ) at a fixed δ under the special case where there is no recovery $\rho = 0$ (but notice that the condition $\rho = 0$ can also reflect an instantaneous recovery/resistance where only a fraction of parasite-adsorbed susceptibles actually proceed to the infected stage). Assume that our resident host population is at a stable endemic equilibrium ($\overline{S} > 0, \overline{I} > 0$), encoding a no-suicide strategy ($\xi = 0$). We ask when a rare mutant encoding a strategy $\xi_m = 0 + \varepsilon, \varepsilon > 0$ small can invade this no-suicide

resident host population. Under the assumptions of the theory of adaptive dynamics, a mutant encoding a strategy ξ_m will invade a resident strategy ξ when its per-generational invasion fitness $\phi_m > 0$.

In our model, the invasion fitness for a mutant encoding a strategy ξ_m is given by:

$$\phi_m = b_m(1 - \kappa_m(\overline{S}+\overline{I})) - \beta_m\overline{I} - (\delta\xi_m + \gamma)$$

Here the $m-$subscripted parameters (other than ξ_m) indicate that they are as yet unspecified smooth functions of the mutant strategy ξ_m, and possibly also of the encoded resident strategy ξ. For instance, $b_m = b(\xi_m, \xi)$. Usually, transmission coefficient (β) is considered to be a parasite related phenotype but for now, we let that to be an as-yet unspecified smooth function of the encoded suicidal strategy as well. For evolution to lead away from the resident no-suicide resident defense phenotype, the fitness gradient at $\xi = 0$ must be positive. That is, $\frac{d\phi_m}{d\xi_m}|_{\xi_m=\xi=0} > 0$. This, by the chain rule, means:

$$\begin{aligned} &\frac{db_m}{d\xi_m} \cdot \left[1 - \kappa_m(\overline{S}+\overline{I})\right]|_{\xi_m=\xi=0} \\ &\quad - b_m \cdot \frac{d\kappa_m}{d\xi_m} \cdot (\overline{S}+\overline{I})|_{\xi_m=\xi=0} \\ &\qquad - \frac{d\beta_m}{d\xi_m} \cdot \overline{I}|_{\xi_m=\xi=0} \\ &> \delta \in [0,1] \end{aligned} \tag{8.2}$$

Now, the terms $(1 - \kappa_m(\overline{S}+\overline{I})) > 0, \overline{S}+\overline{I} > 0$, and $\overline{I} > 0$ under conditions needed for stable endemic equilibrium. So whether the above inequality holds ultimately rests on the interplay of other (derivative) terms. We will consider a series of simple cases; each of these will ultimately inform potential mechanisms that one can explore for the evolution

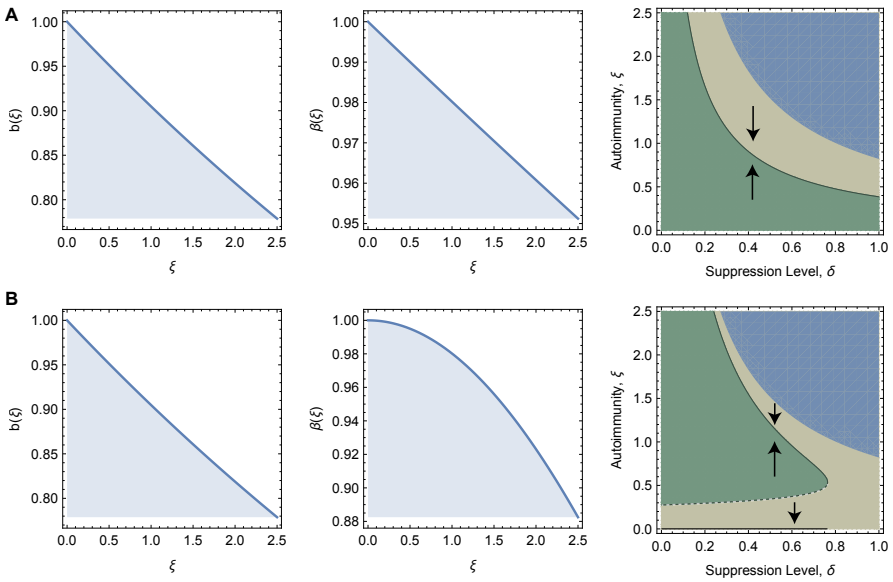

Figure 8.1: Evolution of host abortive infection potential. *Trade off curves and their shapes influence the evolutionary stability (A) and bistability (B) of the evolved suicidal defense strategy. Green and brown regions indicate regions of positive and negative selection respectively; black dark line indicates evolutionary stable states (ESS), dotted line indicates unstable states.*

of suicidal defense.

Case A If both the intraspecific and the transmission coefficients are fixed constants, and not a function of ξ_m ($\kappa_m \equiv \kappa$, $\beta_m \equiv \beta$), then $\frac{\partial b_m}{\partial \xi_m}|_{\xi_m=\xi=0} > \delta / \left[1-\kappa(\overline{S}+\overline{I})\right]|_{\xi_m=\xi=0} \geq 0$ is needed for suicidal defense to evolve. This means, that the hosts must experience a sufficient *reduction in intrinsic growth rates* (i.e., its gradient must be sufficiently positive) at $\xi = 0$. It is easy to see that opposite conclusions arise for the two other cases below.

Case B When $b_m \equiv b, \beta_m \equiv \beta$ but κ still a function of ξ_m and possibly also of ξ, $\frac{\partial \kappa_m}{\partial \xi_m}|_{\xi_m=\xi=0}$ needs to be sufficiently negative for the inequality to hold. That is, the resident hosts must suffer *increased intraspecific competition* at $\xi = 0$.

Case C Similarly, for $b_m \equiv b, \kappa_m \equiv \kappa$ but when β still left to vary with ξ_m and possibly also of ξ, $\frac{\partial \beta_m}{\partial \xi_m}|_{\xi_m=\xi=0}$ needs to be sufficiently negative for the inequality to hold.That is, the resident hosts must suffer *increased disease transmission rates* at $\xi = 0$.

This simple catalog then reveals that there are at least three distinct mechanisms by which adaptive evolutionary dynamics can lead to potential altruistic defense against parasites in a well-mixed SIS model. Indeed, each of these mechanisms can be viewed as a *punishment* mechanism that enforces cooperation among defecting individuals that do not support altruistic defense.

The above analysis addressed the question of whether a mutant encoding a suicidal strategy can invade a resident host population encoding no suicidal defense strategy, under the model assumptions. We can also ask when evolutionary singular points, in particular, attractors, occur in the interior of the ξ parameter space; in this case, the inequality in (8.2) sign above must be replaced with equality. It is then clear that under the classical const-benefit trade-off assumption of monotonically decreasing birth rates with higher resistance (in this case, suicidal defense ξ), unless one of the other gradients is sufficiently positive in (8.2), there is no way for altruism to evolve. This is illustrated in Fig. 8.1, where dependence of intrinsic growth rate and transmission rates on the encoded suicidal strategy allows stable evolution of altruist defense. Similar results are obtained numerically in the presence of background resistance and when the parasites are allowed to coevolve their virulence strategy.

Thus, for post-infection suicidal hosts to evolve in an evolutionarily stable manner, unless helped by other structural changes in the model, complete loss of suicidal

systems must be assumed to be costly. Such a helping mechanism for the evolution of altruistic traits [253, 265] operates by punishment [266–271], and is mimicked by natural prokaryotic toxin/anti-toxin systems that induce post-seggregational killing, or when they take part in essential cellular pathways, or when parasites specialize to infect less suicidal hosts.

Chapter 9

Multi-resolution analysis with bifurcation analysis of smoothing spline models.

Modern biological data are often measured along time and spatial coordinates. And relative to some baseline reference, they reveal trends at various scales. Arctic and antarctic sea ice cover vary by month, season and years. Relative to healthy controls, cancer associated epigenetic signals have been recorded over small chunks of DNA, yet one can also view them to be organized coherently over large blocks of the genome. The goal of this study is to develop some analytics to identify such trends at various scales in a systematic fashion. We exploit smoothing spline ANOVA models to model case-control longitudinal data, and propose carrying out bifurcation analysis of the fitted spline's roots (zero crossings) as a function of the regularization parameter λ (which determines the wiggliness of the fit) to identify qualitative changes in the spline topology. We illustrate the potential of the proposed approach in revealing sound inferences in the case of a few biological applications.

9.1 Smoothing Splines Models

We now give a brief introduction to the theory of smoothing splines, which is needed for our model construction below. For more mathematical details, we direct the readers to refs. [272–274].

Suppose we want to study the association of predictors x_i to outcome y_i, given the observations $i = 1 \ldots n$. The predictors x_i could arise from a discrete set of treatment groups $\in G = \{1, \ldots, K\}$ or a continuos covariate like time or age $\in \mathbb{R}$. More generally, the predictor variables could be multi-dimensional, arising from a product space of covariates.

As with other chapters in this thesis, we shall use the $\cdot$ notation to indicate vectorized quantities. Given the data $(x_\cdot, y_\cdot)$, the smoothing spline technology aims to fit general functions $f(x_i),\ x_i \in \chi, f \in \mathscr{H}$ to describe the mean outcome, by solving a penalized optimization problem:

$$\underset{f \in \mathscr{H}}{\arg\min}\, L(f; y_\cdot, x_\cdot) + \lambda J(f) \tag{9.1}$$

Here, χ is the input domain, and $\mathscr{H}$ is a *reproducing Kernel Hilbert space* (RKHS) of functions defined on χ. $L(\cdot)$ is a loss function, and arises usually from the likelihood model that one prescribes for the data. For instance, if the data generative process is a Gaussian process, one recovers the root mean square deviations of the predicted from the observed values, as $L(\cdot)$. $\lambda \in \mathbb{R}$ is a penalty parameter that acts on the "roughness" (wiggliness) measure of the function $J(f)$. Thus, roughly speaking, the goal of the above optimization problem is to find a reasonably smooth function whose predictions deviate minimally from the observations, while at the same time satisfying some constraint on its

roughness.

The key advantage of restricting ourselves to fitting functions from an RKHS $\mathscr{H}$[2] is that any function $f \in \mathscr{H}$ can be decomposed linearly in terms of their orthogonal projections based on the given data points $i = 1 \ldots n$ (like a standard finite $K-$ dimensional $\mathbb{R}^K$ vector space) as:

$$\begin{aligned} f(x) &= \sum_{i=1}^{n} \alpha_i R(x, x_i) + \nu(x), \text{with } \nu \in \mathscr{H} \ominus \{q : q = \sum_{i=1}^{n} \alpha_i R(\cdot, x_i)\} \\ &= \underbrace{\sum_{i=1}^{n} c_i R_A(x, x_i)}_{\in \mathscr{H}_A \subset \mathscr{H}} + \underbrace{\zeta(x)}_{\in \mathscr{H}_B \subset \mathscr{H}}, \text{ with } \mathscr{H} = \mathscr{H}_A \oplus \mathscr{H}_B \end{aligned} \tag{9.2}$$

Here $\alpha_\cdot, \xi_\cdot$ are scalars. $R(\cdot,\cdot)$ is a bivariate, symmetric non-negative definite function called the reproducing kernel (RK) and is unique for every RKHS. The first line in the above equation decomposes the function into two orthogonal pieces, the first term based on the space's RK, and a residual term ν. In the second line, the overall space $\mathscr{H}$ is decomposed in to two orthogonal closed subspaces $\mathscr{H}_A$ and $\mathscr{H}_B$, which are themselves RKHSs, and therefore have their own RKs; by $R_A(\cdot,\cdot)$, we mean the RK associated with the RKHS subspace $\mathscr{H}_A$. The term orthogonal is used in the standard sense: the space's associated inner product $(f,g)_{\mathscr{H}} = 0 \; \forall f \in \mathscr{H}_A, \; g \in \mathscr{H}_B$. So because all our decompositions above are orthogonal, it should be clear $(R(\cdot,x_i), \nu)_{\mathscr{H}} = 0$, and $(R_A(\cdot,x_i), \zeta)_{\mathscr{H}} = 0$. Notice we could write ζ in terms of the RK of $\mathscr{H}_B$ as well; for our purposes, that is immaterial.

In fact, one can go further and exploit this decompositional convenience to build

[2] a special type of *Hilbert space*, which is defined as a complete vector space endowed with an inner product.

a classical ANOVA like procedure as follows. Consider decomposing the subspace $\mathscr{H}_B$ further:

$$\begin{aligned} f(x) &= \sum_{i=1}^{n} \alpha_i R(x,x_i) \\ &= \underbrace{\sum_{i=1}^{n} c_i R_A(x,x_i)}_{\in \mathscr{H}_A \subset \mathscr{H}} + \underbrace{\zeta(x)}_{\in \mathscr{H}_B \subset \mathscr{H}}, \text{ with } \mathscr{H} = \mathscr{H}_A \oplus \mathscr{H}_B \\ &= \sum_{i=1}^{n} c_i R_A(x,x_i) + \underbrace{\sum_{j=1}^{m-1} d_j \phi_j(x)}_{\in \mathscr{H}_{B0} \subset \mathscr{H}} + \underbrace{\rho(x)}_{\in \mathscr{H}_{B1} \subset \mathscr{H}}, \text{ with } \rho \in \mathscr{H}_B \ominus \{q : q = \sum_{j=1}^{m-1} d_j \phi_j(\cdot)\} \end{aligned} \tag{9.3}$$

where $\phi_j, j = 1 \ldots m-1$ are some set of basis functions chosen such that they are orthogonal to the space spanned by $R_A(\cdot, x_i)\ \forall i = 1 \ldots n$. For instance, this basis can be chosen to include constant and linear order terms (like a general linear model). The higher order terms are then left intact in subspaces $\mathscr{H}_A$ and $\mathscr{H}_{B1}$. We shall come back to this point in the next subsection.

In summary, we have decompsed our function space of interest, the RKHS $\mathscr{H}$, into three distinct orthogonal subspaces: $\mathscr{H} = \mathscr{H}_A + \mathscr{H}_{B0} + \mathscr{H}_{B1}$.

9.2 Two specific instances of the problem

We now briefly illustrate two example smoothing spline models. These will serve to simplify our discussions for case-control longitudinal data next. Each such instance of the general problem considered in eqn. 9.1 involves (a) defining the loss function, (b) the space of functions we need to search over as an RKHS, and their corresponding RKs, and (c) the roughness penalty we would like to impose. Usually, the roughness penalty

measures the (inner product) induced squared norm of the function's projection onto the penalized $\mathscr{H}_1$ subspace, where higher order terms live.

9.2.1 Ridge regression

The standard one-way ANOVA can be cast as an instance of the penalized optimization problem mentioned in eqn. 9.1, by fitting discrete functions $f : \chi \to R$, where $\chi = \{1, \ldots K\}$ represents the experimental groups. It turns out the function space $\mathscr{H}$ defined this way is an RKHS with an RK $R(x,y) = I_{[x==y]}$. Let $\mathbf{1}$ be a K dimensional vector of 1s. We can decompose $\mathscr{H} \ni f$ into two orthogonal subspaces with individual RKs (that add to give the original space's RK as):

$$I_{[x==y]} = \underbrace{\frac{\mathbf{1}\mathbf{1}^{\mathbf{T}}}{K}}_{R_0} + \underbrace{\left(I - \frac{\mathbf{1}\mathbf{1}^{\mathbf{T}}}{K}\right)}_{R_1} \tag{9.4}$$

.

The RKHS subspaces $\mathscr{H}_0$ and $\mathscr{H}_1$ of $\mathscr{H}$ spanned by each of these RKs can be reasoned about by studying their respective linear combinations $\{\sum_i \alpha_i R_j(\cdot, x_i), \alpha_i \in \mathbb{R}, x_i \in \chi\}$ for $j \in \{0,1\}$ corresponding to each of the RKs R_0, and R_1. This corresponds to spaces $H_0 = \{f : f(1) = f(2) = \cdots = f(K)\}$, and $\mathscr{H}_1 = \{f : (f,g) = 0, f \in \mathscr{H}_0, g \in \mathscr{H}_1\}$, where the inner product $(f,g)_{\mathscr{H}} = f^T g$.

Then, when one considers the following instance of the optimization problem 9.1:

$$\underset{f \in \mathscr{H}}{\arg\min} \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(x_i))^2}_{=L(f;x_\cdot,y_\cdot)} + \underbrace{\lambda f^T f}_{=\lambda J(f)} \tag{9.5}$$

we are effectively solving the ridge regression problem.

9.2.2 Cubic smoothing splines

Consider fitting functions $f : \chi \to \mathbb{R}$, where $\chi = [0,1]$ to a continuos covariate $x \in \chi$. Let us restrict our attention to fitting functions whose second derivatives are square integrable: i.e., for $f \in \{g : g^{(2)} \in \mathscr{L}_2[0,1]\}$. It turns out one can decompose this RKHS for f as $\mathscr{H} = \mathscr{H}_0 \oplus \mathscr{H}_1$, where:

$$\begin{aligned} \mathscr{H}_0 &= \left\{ g : g^{(2)} = 0 \right\}, \text{and,} \\ \mathscr{H}_1 &= \left\{ g : g^{(0)} = g^{(1)} = 0, \int \left(g^{(2)} \right)^2 dx < \infty \right\} \end{aligned} \tag{9.6}$$

with corresponding RKs:

$$\begin{aligned} R_0(x,y) &= 1 + k_1(x)k_1(y) \\ R_1(x,y) &= k_2(x)k_2(y) - k_4(|x-y|), \end{aligned}$$

where :

$$\begin{aligned} k_1(x) &= x - 0.5 \\ k_2(x) &= \frac{1}{2}\left(k_1^2(x) - \frac{1}{12} \right) \\ k_4(x) &= \frac{1}{24}\left(k_1^4(x) - \frac{k_1^2(x)}{2} + \frac{7}{240} \right) \end{aligned} \tag{9.7}$$

are the Bernoulli polynomials.

The point here is to note that the zero and first order terms (in x) are attached to $\mathscr{H}_0$ and the non-linear higher order terms are restricted to $\mathscr{H}_1$. So one choice of basis

functions for $\mathscr{H}_0$ are given by $\phi_0(x) = 1, \phi_1(x) = x - .5$.

With such a setup, one can write $f(x)$ as:

$$f(x) = \underbrace{\sum_{j=0}^{1} d_j \phi_j}_{\in \mathscr{H}_0} + \underbrace{\sum_{i=1}^{n} R_1(x, x_i) + \rho(x)}_{\in \mathscr{H}_1} \tag{9.8}$$

where as before $\rho \in \mathscr{H}_1 \ominus g : \sum_{i=1}^{n} \alpha_i R_1(\cdot, x_i)$. Due to orthogonality $(R_1(\cdot, x_i), \rho) = (\phi_j, R_1(\cdot, x_i)) = (\phi_j, \rho) = 0$. One then considers the penalized problem for obtaining cubic smoothing splines f:

$$\underset{f \in \mathscr{H}}{\arg\min} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(x_i))^2}_{= L(f; x_\cdot, y_\cdot)} + \underbrace{\lambda \int_0^1 \left(f^{(2)} \right)^2 dx}_{= \lambda J(f)} \tag{9.9}$$

where the penalty mesures the roughness/curvature of the function.

9.2.3 Deriving the solution of the cubic smoothing spline problem

Substituting the decomposition of f derived above, and noting the orthogonalities of ρ, $R_1(\cdot, x_i)$, and ϕ_j, together with the identity that $\int_0^1 R_1^{(2)}(x, x_i) R_1^{(2)}(x_i, x) = R_1(x_i, x_j)$ we find that our specific cubic spline optimization problem in 9.9 is given in matrix terms as:

$$\underset{\mathbf{c}, \mathbf{d}}{\arg\min} (Y - S\mathbf{d} - Q\mathbf{c})^T (Y - S\mathbf{d} - Q\mathbf{c}) + n\lambda \mathbf{c}^T Q \mathbf{c} + n\lambda (\rho, \rho) \tag{9.10}$$

.

Here Y is the n length response vector, Q is an $n \times n$ matrix with $Q(i, j) = R_1(x_i, x_j)$,

S is the $n \times 2$ matrix with row-wise entries $(\phi_1(x_i), \phi_2(x_i))$. It should be clear that ρ appears only through its square norm term, which being independent of the parameters, merely serves to introduce a non-negative shift in the objective's location away from zero; clearly then, the objective is minimized when $\rho = 0$. One then exploits linear algebraic techniques to solve for **c** and **d** to estimate our cubic smoothing smoothing splines. As one would expect, these estimates are functions of the roughness penalty λ. We thus arrive at the celebrated Kimmeldorf-Wahba result [272] that polynomial smoothing splines reside in a closed, finite dimensional space $\mathscr{H}_0 \oplus \{g : g = \sum_{i=1}^{n} \alpha_i R_1(\cdot, x_i), \alpha_i \in \mathbb{R}\}$.

9.3 Proposed strategy for multi-resolution analysis of case-control longitudinal data

As we noted in the general formulation of the smoothing splines models in eqn. 9.1, by increasing the penalty parameter λ, one trades off data fit for model simplicity. For instance, in the case of cubic splines models, $\lambda \to \infty$ chooses a linear fit with no wiggliness, while $\lambda \to 0$ retrieves a cubic spline within the function space that interpolates all the data.

We make three key observations that lead us to the proposed algorithm. First, varying λ leads us to fitting models with varied complexities. This means, in terms of a difference (contrast) function that one might build for longitudinal data (e.g., time series measurements of a bacterium's abundance in cases and controls, or DNA coordinate-wise epigenetic measurements from cancer cells relative to healthy controls), lower values of λ would reveal finer blocks of changes, while higher values of λ will restrict our attention to

larger organizational blocks of changes in DNA. Second, given the continuous nature of functions we fit our data with, the only way by which a bigger block of change can arise at a higher value of λ, is by loss of one or more roots from the splines fitted at smaller λ values (that is via a reduction in the number of points where the difference function attains a value of zero). Thus by cataloging the number of roots obtained as we vary λ (and the function gradients at these points, discussed later), one obtains a quantitative picture of the special points in the λ space, where qualitative changes in the topology of the fitted splines occur. Finally, as assumptions behind the Implicit Function Theorem apply to spline solutions almost everywhere in the domain, one can go further, exploit numerical continuation theory to obtain the location of the roots smoothly as we vary λ.

9.4 Model construction for longitudinal case-control data

In the two simplified problem instances described above, we constructed functions for categorical and continuous data separately. Our goal now is to exploit smoothing spline technology to construct a contrast function that describes a continuous change in outcome along one continuous coordinate $t \in \tau$ (e.g., DNA coordinate or time) between two discrete experimental conditions $x \in \chi = 1,2$ (e.g., controls and cases).

It turns out one can construct RKHS model spaces for multi-variate functions as easily as constructing them for univariate functions. The main result we exploit here is that given two marginal RKHSs and their decompositions: $\mathscr{H}^{\chi} = \mathscr{H}_0^{\chi} + \mathscr{H}_1^{\chi}$ and

$\mathscr{H}^{\tau} = \mathscr{H}_0^{\tau} + \mathscr{H}_1^{\tau}$, a (tensor) product RKHS space can be constructed as :

$$\begin{aligned}\mathscr{H}^{\chi\otimes\tau} &= \{\mathscr{H}_0^{\chi} \oplus \mathscr{H}_1^{\chi}\} \otimes \{\mathscr{H}_0^{\tau} \oplus \mathscr{H}_1^{\tau}\} \\ &= \underbrace{\{\mathscr{H}_0^{\chi} \otimes \mathscr{H}_0^{\tau}\}}_{=\mathscr{H}_0^{\chi\otimes\tau}} \oplus \underbrace{\underbrace{\{\mathscr{H}_0^{\chi} \otimes \mathscr{H}_1^{\tau}\}}_{=\mathscr{H}_{10}^{\chi\otimes\tau}} \oplus \underbrace{\{\mathscr{H}_1^{\chi} \otimes \mathscr{H}_0^{\tau}\} \oplus \{\mathscr{H}_1^{\chi} \otimes \mathscr{H}_1^{\tau}\}}_{=\mathscr{H}_{11}^{\chi\otimes\tau}}}_{=\mathscr{H}_1^{\chi\otimes\tau}}\end{aligned} \tag{9.11}$$

where we have hierarchically decomposed our new product input space into two orthogonal RKHS [272–274]. The only remaining step needed to utilize the algorithms outlined in the previous section is to identify the RKs associated with these spaces. Conveniently, it turns out that RKs assigned to the marginal spaces also add and multiply accordingly!

$$\begin{aligned}R^{\chi\otimes\tau} &= \{R_0^{\chi} + R_1^{\chi}\} \times \{R_0^{\tau} + R_1^{\tau}\} \\ &= \underbrace{\{R_0^{\chi} \times R_0^{\tau}\}}_{=R_0^{\chi\otimes\tau}} + \underbrace{\underbrace{\{R_0^{\chi} \times R_1^{\tau}\}}_{=R_{10}^{\chi\otimes\tau}} + \underbrace{\{R_1^{\chi} \times R_0^{\tau}\} + \{R_1^{\chi} \times R_1^{\tau}\}}_{=R_{11}^{\chi\otimes\tau}}}_{R_1^{\chi\otimes\tau}}\end{aligned} \tag{9.12}$$

From the hierarchical decomposition of RKs and the marginal basis functions, it is clear that $H_0^{\chi\times\tau}$ is spanned by the basis $\{\phi_1(x,t) = 1, \phi_2(x,t) = t - 0.5\}$ and models the grand mean and linear main effect of τ; $H_{10}^{\chi\times\tau}$ describes the smooth main effect due to τ, and the third, interesting for our purposes, space $H_{11}^{\chi\times\tau}$ models the smooth-linear interaction and smooth-smooth interactions of x and τ. Thus, it is this subspace whose contributions to the fitted function completely specify overall treatment effects that we care for in this work. Specifically, we can write, for every function $f \in \mathscr{H}^{\chi\times\tau}$, and

denoting $z_i = (x_i, t_i)$

$$\begin{aligned} f(z) &= \sum_{j=0}^{1} d_j \phi_j(z) + \sum_{i=1}^{n} c_i R_1^{\chi\otimes\tau}(z, z_i) + \rho(z) \\ &= \sum_{j=0}^{1} \phi_j(z) + \sum_{i=1}^{n} c_i R_{10}^{\chi\otimes\tau}(z, z_i) + \underbrace{\sum_{i=1}^{n} c_i R_{11}^{\chi\otimes\tau}(z, z_i)}_{=\gamma(z)} + \rho(z) \end{aligned} \tag{9.13}$$

where $\rho \in \mathscr{H}_1^{\chi\otimes\tau} - \{\sum_{i=1}^{n} c_i R_1^{\chi\otimes\tau}(\cdot, z_i)\}$, and we have exploited the hierarchical decomposition of the RK corresponding to $\mathscr{H}_1^{\chi\otimes\tau}$. We emphasize that $\gamma(\cdot)$ is the overall effects function whose roots we are after, as a function of λ in the rest of this work.

9.4.1 Estimation and Notation

With this model space construction, for every λ, we can estimate $\mathbf{c}(\lambda)$ from the linear algebraic algorithms (Algorithm 3.4.2 [273]) available for smoothing spline models, and compute the contrast function $\gamma(z = (case, t))$ as:

$$\gamma(z = (case, t); \lambda) = \mathbf{R_{11}}^T(t)\mathbf{c}(\lambda) \tag{9.14}$$

Here $\mathbf{R_{11}}(t) = [|R_{11}(t_i, t)|]_{i=1}^{n}$. Henceforth, we drop the superscript $\chi \otimes \tau$ and denote by c the entire estimated vector of c_i s instead of a bold typeface. Furthermore, because we restrict our analysis to $x = case$, we will make the estimation functions' dependence on it implicit, and simply use $\gamma(t, \lambda)$ and $R_{11}(t)$.

9.5 Bifurcation analysis of $\gamma(t,\lambda)$ with λ as the control parameter

Noting that it is the roots of the contrast function $\gamma(t,\lambda)$ we are after, and that the function is continuously differentiable[1], Implicit Function Theorem guarantees the existence of a smooth solution to the equation $\gamma(t,\lambda)=0$ in the open neighborhood of a given root (t^*,λ^*) whenever $\frac{\partial\gamma}{\partial\lambda}(t^*,\lambda*)\neq 0$.

In fact, this is the central theory underlying numerical bifurcation analysis in dynamical systems theory, where qualitative behavior about steady states are mapped as a function of some control parameter. We had exploited such a technique in Part II of this thesis for the analysis of a CRISPR model.

For our purposes here, we developed the equivalent numerical continuation algorithms and implemented them in the R software language. Fold points were detected by simultaneously asking for $\gamma(t^*)=0$ and $\dot{\gamma}(t^*)=0$.

9.5.1 Confidence intervals for $\hat{t}$ given λ

For every λ, the confidence intervals in the fitted roots can be obtained with a linearization calculation as below. For a given value of λ, let the root along τ axis be given as $\hat{t}$. Expand around the true root t_0, when $\dot{\gamma}(\hat{t})\neq 0$:

$$\gamma(\hat{t})=\gamma(t_0)+\dot{\gamma}(\hat{t}-t_0)+O(|\hat{t}-t_0|^2)\implies \hat{t}\ \approx t_0+\frac{1}{\dot{\gamma}(\hat{t})}\gamma(\hat{t}), \tag{9.15}$$

[1] to be precise, the function is continuously differentiable almost everywhere, given the non-differentiable nature of the RK $R_{11}(z)$ at data points $x_i, i=1\ldots n$. But we do not worry about this complication in this work

which leaves us with the following approximate variance on the estimated root:

$$Var(\hat{t}) = \left[\frac{1}{\dot{\gamma}(\hat{t})}\right]^2 Var(\gamma(\hat{t})) \tag{9.16}$$

where $Var(\gamma(\hat{t}))$ is easily available as the posterior variance of the fitted function γ at $\hat{t}$ form the Bayesian calculations of Wahba [272] for polynomial smoothing splines. Based on that theory, a Gaussian process prior with a mean zero and a covariance function proportional to the RK $R_{11}(\cdot,\cdot)$ can be assumed for γ. When$\dot{\gamma}(\hat{t}) = 0$, which is the case at a fold point, a second order treatment is made in the above variance calculations. A $100(1-\frac{\alpha}{2})\%$ confidence interval can then be approximately obtained as: $\hat{t} \pm z_{\frac{\alpha}{2}} \sqrt{Var(\hat{t})}$, which is an interesting overlay to the classical bifurcation analysis methodology exploited for deterministic systems.

9.6 Applications

9.6.1 Metagenomic time series

To illustrate the potential of the proposal above in identifying multi-resolution changes in longitudinal data, we first considered the 16s metagenomic feature data from David et al., [275]. In this work, the authors measured microbial frequencies over time as an individual travelled abroad and returned back home. We performed a bifurcation analysis on the contrast function for a few dominant genera, where samples post-travel were considered as "cases" and those pre-travel were considered as "controls". The results are presented in Fig. 9.1. One of the main results in from David et al., is clearly recapitulated

in this plot: Bacteroides and Blautia undergo major long-term changes post- travel before they settle back to the pre-travel state. Bifurcation points along the time axis indicate at which points changes in the (relative) abundances started to occur. These changes are located at roughly similar time points for several features, indicating correlated factors underlying their observed changes. As to whether they are purely due to compositional effects discussed in Part I of the manuscript or truly owing to underlying biological reasons, we cannot conclude from this result alone.

9.6.2 Genome-Wide DNA Methylation Signals

We next applied the technique to characterize long and short-term changes in high resolution methylation signals throughout the genome in lung cancer tissue relative to healthy controls. In contrast to the metagenomic time series datasets above, which consist of a few hundred to a few thousand observations, we are now faced with the problem of analyzing millions of methylation intensity values averaged and recorded throughout the genome in 150 bp nucleosome sized windows [276]. This is a major computational challenge, which we currently address using the following modifications to the more accurate algorithm outlined in the previous section. First, we observed that changes in methylation often spanned over several thousands of base-pairs. Second, given a value of λ, finding its roots involve solving an one-dimensional root finding problem. Although non-linear, sound computational techniques exist that are very fast for this problem [277]. Finally, we exploited the faster estimation algorithm by Kim and Gu [278] (also see section 3.5.3 in [273]), which by only using a fewer $q < n$ set of observations to describe the $\mathscr{H}_1$ space (instead of all of the n observations to form $\sum_i c_{i=1}^{q} R_1(,x_i)$), improved the speed

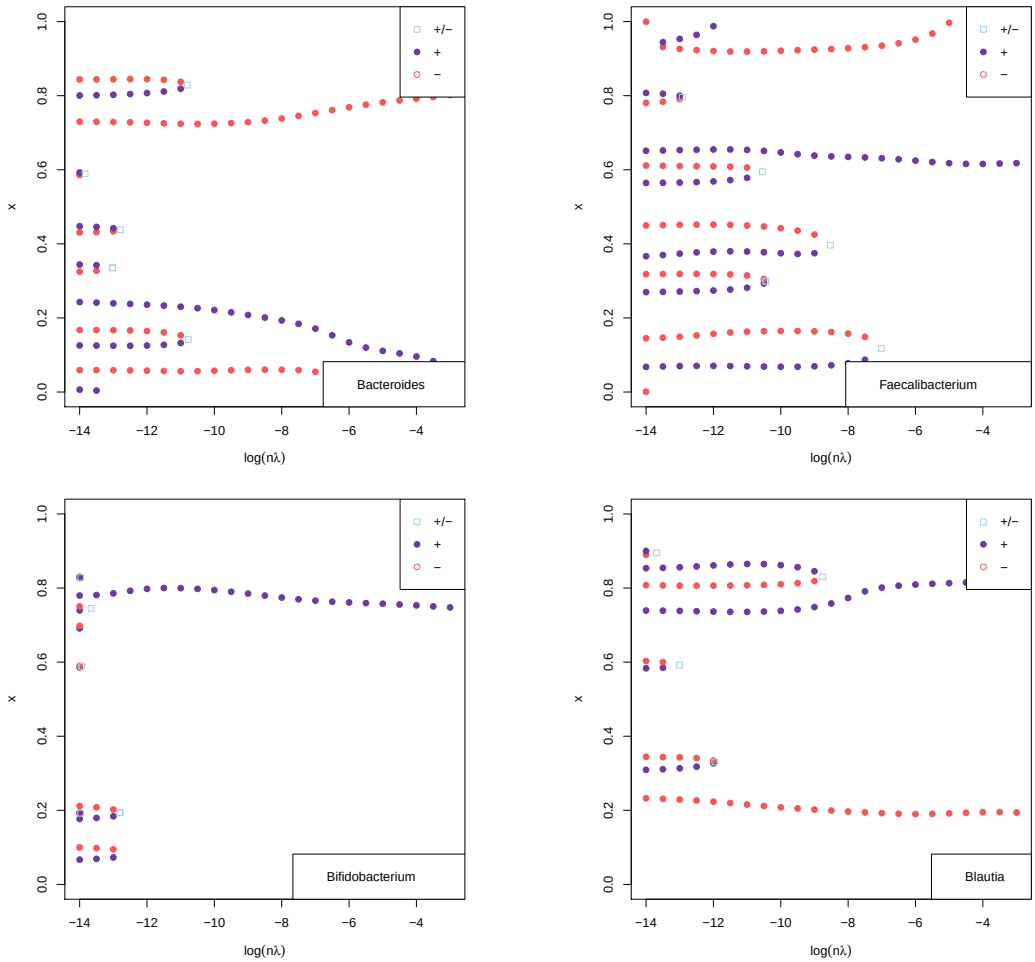

Figure 9.1: Long term and short-term differences in microbial time series pre- and post- travel. *Plotted are the roots of the contrast functions for various microbial genera $\gamma(x,\lambda)$ for various values of $n*\lambda$; n is the sample size, and x is rescaled time. Purple and red points indicate roots where the contrast function has positive and negative gradients with respect to the longitudinal coordinate (in this case, x) respectively. For any given value of λ then, the region between a consecutive (blue, red) pair is a region of positive difference, while a region between a consecutive (red, blue) pair of points indicate regions with a negative difference in cases relative to controls. Blue squares indicate fold points.*

of the algorithm several folds. In summary, if one does not care about fold points, these modifications to the original algorithm prescribed above, lead to deriving plots similar to Fig. 9.1 genome-wide in non-overlapping 1 megabase pair windows in less than 3 hours, with 3 parallel processes on a Macintosh laptop. With the more accurate algorithm, this

problem would have taken more than two weeks to solve, with 16 parallel processes.

After obtaining the contrast function for various values of the roughness penalty λ, we computed the lengths of differential regions/segments suggesting a negative difference (*hypo*-methylated) and positive change (*hyper*-methylated) relative to controls. In Fig. 9.2, bottom panel, we plot the growth in the median lengths of the hypo- and hyper-methylated regions vary (over 8x) as a function of λ. In general, for all resolutions(specified here by the value of $\log(n\lambda)$), we found higher median lengths for hypo- methylated regions than in hyper- methylated regions. The distributions of length values obtained at fine- ($\log(n\lambda) = -1$) and large-scale ($\log(n\lambda) = -14$) changes are shown in the top two panels.

Interestingly, we also found that across all resolutions, transcription factor binding sites, as measured with ChipSeq by the ENCODE consortia, were enriched in hypo-methylated regions genome-wide in lung cancer tissues. To illustrate this, we have plotted the binding site fraction in hypo-methylated regions in Fig. 9.3. These findings were confirmed by Fisher exact tests for >85% of the transcription factors as well.

Here is one possible explanation for the aforementioned results. The significantly longer hypo-methylation blocks in lung cancer and the enrichment of transcription factor binding sites in these regions could be caused by competitive binding of over-expressed transcription factors, or other DNA binding agents, preventing stable methylation establishment. A simple kinetic / stochastic process model of such binding events will indicate that this is a genuine possibility. If one has access to absolute concentration measurements of protein molecules/mRNA expression from genes, such a hypothesis can quickly be tested as well. As described in Part I of this thesis, relying on frequency based measure-

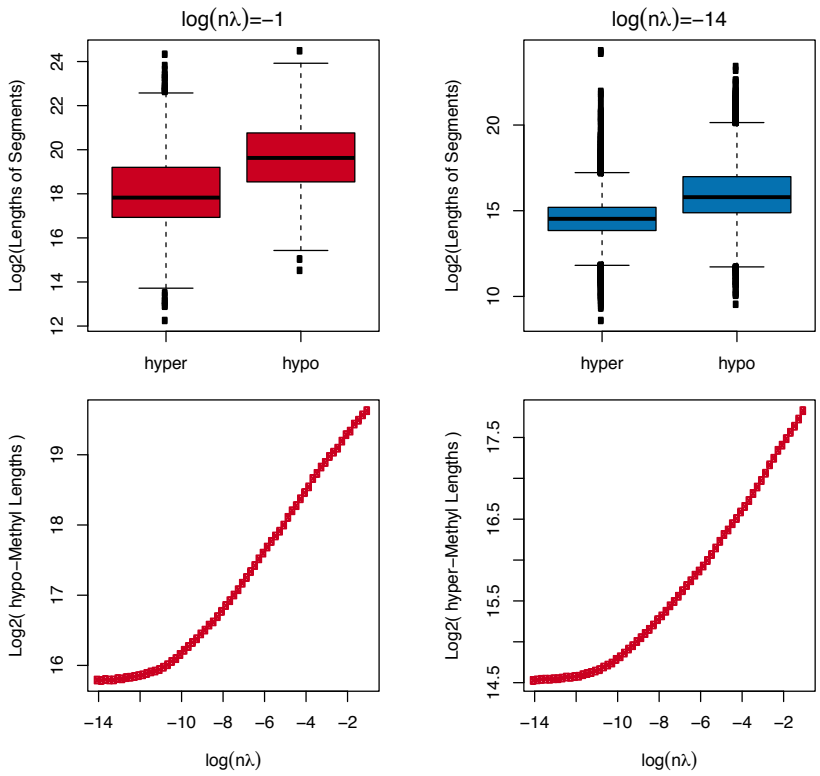

Figure 9.2: Scale specific genome-wide differences in DNA methylation in lung cancer tissue relative to controls. *Plotted are the median lengths of differentially methylated regions* $\gamma(x,\lambda)$ *estimated from lung cancer data relative to healthy controls for various resolutions (as measured by* $\log(n*\lambda)$*); n is the sample size.*

ments from RNAseq (unless resolved effectively with internal spike-in control features) need not always allow stable biologically relevant conclusions. Although scale normalization approaches for RNAseq can lead to effective inferences when most genes do not change in their expression values, cancer tissues exist where such an assumption is heavily violated [113].

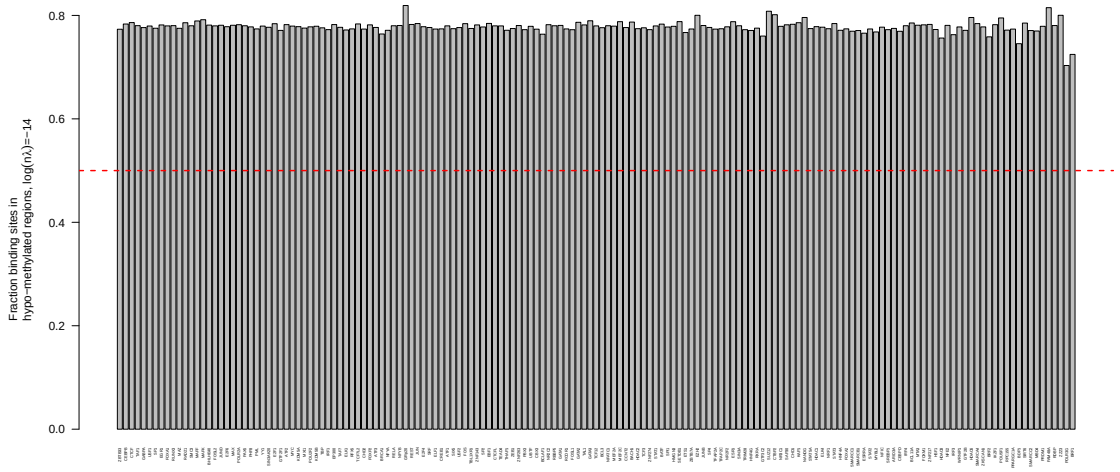

Figure 9.3: Enrichment of transcription factor binding sites in hypo-methylated regions. *For each transcription factor whose binding sites were characterized by the ENCODE consortia, we plot the fraction of binding sites found in lung cancer's hypo-methylated regions. Enrichment of transcription factor binding sites in hypo-methylated regions was generally the case for all resolutions as measured by $log(n\lambda)$. The red dashed line indicates a value of 0.5.*

Bibliography

[1] Bruce R Levin. Nasty viruses, costly plasmids, population dynamics, and the conditions for establishing and maintaining CRISPR-mediated adaptive immunity in bacteria. *PLoS genetics*, 6(10):e1001171, October 2010.

[2] Lauren M Childs, Nicole L Held, Mark J Young, Rachel J Whitaker, and Joshua S Weitz. Multiscale model of CRISPR-induced coevolutionary dynamics: diversification at the interface of Lamarck and Darwin. *Evolution; international journal of organic evolution*, 66(7):2015–2029, July 2012.

[3] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, July 2008.

[4] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[5] Joseph N. Paulson, O. Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 2013.

[6] Aaron T. L. Lun, Karsten Bach, and John C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17:75, 2016.

[7] Ying Yu, James C. Fuscoe, Chen Zhao, Chao Guo, Meiwen Jia, Tao Qing, Desmond I. Bannon, Lee Lancashire, Wenjun Bao, Tingting Du, Heng Luo, Zhenqiang Su, Wendell D. Jones, Carrie L. Moland, William S. Branham, Feng Qian, Baitang Ning, Yan Li, Huixiao Hong, Lei Guo, Nan Mei, Tieliu Shi, Kevin Y. Wang, Russell D. Wolfinger, Yuri Nikolsky, Stephen J. Walker, Penelope Duerksen-Hughes, Christopher E. Mason, Weida Tong, Jean Thierry-Mieg, Danielle Thierry-Mieg, Leming Shi, and Charles Wang. A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. *Nature Communications*, 5:3230, February 2014.

[8] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R. Mende, Adriana Alberti, Francisco M. Cornejo-Castillo, Paul I. Costea, Corinne Cruaud, Francesco d'Ovidio, Stefan Engelen, Isabel Ferrera, Josep M. Gasol, Lionel Guidi, Falk Hildebrand, Florian Kokoszka, Cyrille Lepoivre, Gipsi Lima-Mendez, Julie Poulain, Bonnie T. Poulos, Marta Royo-Llonch, Hugo Sarmento, Sara Vieira-Silva, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans Coordinators, Chris Bowler, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Fabrice Not, Hiroyuki Ogata, Stephane Pesant, Sabrina Speich, Lars Stemmann, Matthew B. Sullivan, Jean Weissenbach, Patrick Wincker, Eric Karsenti, Jeroen Raes, Silvia G. Acinas, and Peer Bork. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359, May 2015.

[9] Christos Argyropoulos, Alton Etheridge, Nikita Sakhanenko, and David Galas. Modeling bias and variation in the stochastic processes of small RNA sequencing. *Nucleic Acids Research*, 45(11):e104, June 2017.

[10] Gregor Mendel and Paul C. Mangelsdorf. *Experiments in Plant Hybridisation*. Harvard University Press, 1965. Google-Books-ID: pzSoD55L1W0C.

[11] Ronald Aylmer Fisher. *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press, 1930.

[12] John Burdon Haldane. *The causes of evolution*. Number 36. Princeton University Press, 1932.

[13] James D. Watson and Francis HC Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[14] Philip Hedrick. *Genetics of populations*. Jones & Bartlett Learning, 2011.

[15] Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *Journal of experimental medicine*, 79(2):137–158, 1944.

[16] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356, 1961.

[17] David Baltimore Lodish, et al Harvey. *Molecular Cell Biology, 4th Edition*. W H Freeman & Co, fourth edition edition edition, 2002.

[18] David L. Nelson, Albert L. Lehninger, and Michael M. Cox. *Lehninger principles of biochemistry*. Macmillan, 2008.

[19] Gregory J. Hannon. RNA interference. *nature*, 418(6894):244, 2002.

[20] Lin He and Gregory J. Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, 5(7):522, 2004.

[21] Hans V. Westerhoff and Bernhard O. Palsson. The evolution of molecular biology into systems biology. *Nature biotechnology*, 22(10):1249, 2004.

[22] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC, 2006.

[23] Peter A. Jones and Peter W. Laird. Cancer-epigenetics comes of age. *Nature genetics*, 21(2):163, 1999.

[24] Andrew P. Feinberg and Benjamin Tycko. The history of cancer epigenetics. *Nature Reviews Cancer*, 4(2):143, 2004.

[25] Robin Holliday. Epigenetics: a historical overview. *Epigenetics*, 1(2):76–80, 2006.

[26] Andrew P. Feinberg, Rolf Ohlsson, and Steven Henikoff. The epigenetic progenitor origin of human cancer. *Nature reviews genetics*, 7(1):21, 2006.

[27] Adrian Bird. Perceptions of epigenetics. *Nature*, 447(7143):396, 2007.

[28] Charles Darwin. *On the origin of species*. Routledge, 1859.

[29] Theodosius Dobzhansky and Theodosius Grigorievich Dobzhansky. *Genetics and the Origin of Species*, volume 11. Columbia university press, 1982.

[30] Dan Graur and Wen-Hsiung Li Li. *Fundamentals of Molecular Evolution*. Sinauer Associates is an imprint of Oxford University Press, Sunderland, Mass, 2 edition edition, January 2000.

[31] Andrew P. Feinberg and Rafael A. Irizarry. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences*, 107(suppl 1):1757–1764, 2010.

[32] David Botstein and Neil Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33(3s):228–237, March 2003.

[33] Pedro D'Orléans-Juste, Jean-Claude Honoré, Emilie Carrier, and Julie Labonté. Cardiovascular diseases: new insights from knockout mice. *Current Opinion in Pharmacology*, 3(2):181–185, April 2003.

[34] The Comprehensive Knockout Mouse Project Consortium, Christopher P. Austin, James F. Battey, Allan Bradley, Maja Bucan, Mario Capecchi, Francis S. Collins, William F. Dove, Geoffrey Duyk, Susan Dymecki, Janan T. Eppig, Franziska B. Grieder, Nathaniel Heintz, Geoff Hicks, Thomas R. Insel, Alexandra Joyner, Beverly H. Koller, K. C. Kent Lloyd, Terry Magnuson, Mark W. Moore, Andras Nagy, Jonathan D. Pollock, Allen D. Roses, Arthur T. Sands, Brian Seed, William C.

Skarnes, Jay Snoddy, Philippe Soriano, David J. Stewart, Francis Stewart, Bruce Stillman, Harold Varmus, Lyuba Varticovski, Inder M. Verma, Thomas F. Vogt, Harald von Melchner, Jan Witkowski, Richard P. Woychik, Wolfgang Wurst, George D. Yancopoulos, Stephen G. Young, and Brian Zambrowicz. The Knockout Mouse Project. *Nature Genetics*, 36:921–924, September 2004.

[35] Florence Vignols, Claire Bréhélin, Yolande Surdin-Kerjan, Dominique Thomas, and Yves Meyer. A yeast two-hybrid knockout strain to explore thioredoxin-interacting proteins in vivo. *Proceedings of the National Academy of Sciences*, 102(46):16729–16734, 2005.

[36] Tomoya Baba, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A. Datsenko, Masaru Tomita, Barry L. Wanner, and Hirotada Mori. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology*, 2(1), 2006.

[37] Juraj Gregan, Peter K. Rabitsch, Cornelia Rumpf, Maria Novatchkova, Alexander Schleiffer, and Kim Nasmyth. High-throughput knockout screen in fission yeast. *Nature protocols*, 1(5):2457, 2006.

[38] D. W. Threadgill, A. A. Dlugosz, L. A. Hansen, T. Tennenbaum, U. Lichti, D. Yee, C. LaMantia, T. Mourton, K. Herrup, and R. C. Harris. Targeted disruption of mouse EGF receptor: effect of genetic background on mutant phenotype. *Science (New York, N.Y.)*, 269(5221):230–234, July 1995.

[39] L. J. Kurihara, T. Kikuchi, K. Wada, and S. M. Tilghman. Loss of Uch-L1 and Uch-L3 leads to neurodegeneration, posterior paralysis and dysphagia. *Human Molecular Genetics*, 10(18):1963–1970, September 2001.

[40] Marek Drab, Paul Verkade, Marlies Elger, Michael Kasper, Matthias Lohn, Birgit Lauterbach, Jan Menne, Carsten Lindschau, Fanny Mende, Friedrich C. Luft, Andreas Schedl, Hermann Haller, and Teymuras V. Kurzchalia. Loss of Caveolae, Vascular Dysfunction, and Pulmonary Defects in Caveolin-1 Gene-Disrupted Mice. *Science*, 293(5539):2449–2452, September 2001.

[41] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–695, February 1977.

[42] F. Sanger. Sequences, sequences, and sequences. *Annual Review of Biochemistry*, 57:1–28, 1988.

[43] T. Hunkapiller, R. J. Kaiser, B. F. Koop, and L. Hood. Large-scale and automated DNA sequence determination. *Science (New York, N.Y.)*, 254(5028):59–67, October 1991.

[44] Clyde A. Hutchison. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research*, 35(18):6227–6237, 2007.

[45] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, October 2008.

[46] Timothy D. Harris, Phillip R. Buzby, Hazen Babcock, Eric Beer, Jayson Bowers, Ido Braslavsky, Marie Causey, Jennifer Colonell, James Dimeo, J. William Efcavitch, Eldar Giladi, Jaime Gill, John Healy, Mirna Jarosz, Dan Lapen, Keith Moulton, Stephen R. Quake, Kathleen Steinmann, Edward Thayer, Anastasia Tyurina, Rebecca Ward, Howard Weiss, and Zheng Xie. Single-molecule DNA sequencing of a viral genome. *Science (New York, N.Y.)*, 320(5872):106–109, April 2008.

[47] Eric S. Lander and Michael S. Waterman. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2(3):231–239, April 1988.

[48] M. F. Bonaldo, G. Lennon, and M. B. Soares. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Research*, 6(9):791–806, September 1996.

[49] Joshua S. Bloom, Zia Khan, Leonid Kruglyak, Mona Singh, and Amy A. Caudy. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, 10(1):221, May 2009.

[50] Alicia Oshlack and Matthew J. Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology direct*, 4(1):14, 2009.

[51] Matthew D. Young, Matthew J. Wakefield, Gordon K. Smyth, and Alicia Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome biology*, 11(2):R14, 2010.

[52] Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, March 2010.

[53] Alicia Oshlack, Mark D. Robinson, and Matthew D. Young. From RNA-seq reads to differential expression results. *Genome Biology*, 11(12):220, December 2010.

[54] Lior Pachter. Models for transcript quantification from RNA-Seq. *arXiv:1104.3889 [q-bio, stat]*, April 2011. arXiv: 1104.3889.

[55] Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics*, 12(1):480, December 2011.

[56] Simon T. Bennett, Colin Barnes, Anthony Cox, Lisa Davies, and Clive Brown. Toward the 1,000 dollars human genome. *Pharmacogenomics*, 6(4):373–382, June 2005.

[57] Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, Michael S. Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B. Dewell, Lei Du, Joseph M. Fierro, Xavier V. Gomes, Brian C. Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Chun He Ho, Gerard P. Irzyk, Szilveszter C. Jando, Maria L. I. Alenquer, Thomas P. Jarvie, Kshama B. Jirage, Jong-Bum Kim, James R. Knight, Janna R. Lanza, John H. Leamon, Steven M. Lefkowitz, Ming Lei, Jing Li, Kenton L. Lohman, Hong Lu, Vinod B. Makhijani, Keith E. McDade, Michael P. McKenna, Eugene W. Myers, Elizabeth Nickerson, John R. Nobile, Ramona Plant, Bernard P. Puc, Michael T. Ronan, George T. Roth, Gary J. Sarkis, Jan Fredrik Simons, John W. Simpson, Maithreyan Srinivasan, Karrie R. Tartaro, Alexander Tomasz, Kari A. Vogt, Greg A. Volkmer, Shally H. Wang, Yong Wang, Michael P. Weiner, Pengguang Yu, Richard F. Begley, and Jonathan M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, September 2005.

[58] Tarjei S. Mikkelsen, Manching Ku, David B. Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P. Koche, William Lee, Eric Mendenhall, Aisling O'Donovan, Aviva Presser, Carsten Russ, Xiaohui Xie, Alexander Meissner, Marius Wernig, Rudolf Jaenisch, Chad Nusbaum, Eric S. Lander, and Bradley E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, August 2007.

[59] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, 320(5881):1344–1349, June 2008.

[60] John C. Marioni, Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 2008.

[61] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, January 2009.

[62] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, February 2014.

[63] Peter V. Kharchenko, Michael Y. Tolstorukov, and Peter J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology*, 26(12):1351, 2008.

[64] Anton Valouev, David S. Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M. Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods*, 5(9):829, 2008.

[65] Peter J. Park. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, October 2009.

[66] Axel Visel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, and Feng Chen. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854, 2009.

[67] Terrence S. Furey. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, 13(12):840, 2012.

[68] Alexander Meissner, Andreas Gnirke, George W. Bell, Bernard Ramsahoye, Eric S. Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic acids research*, 33(18):5868–5877, 2005.

[69] Aaron L. Statham, Mark D. Robinson, Jenny Z. Song, Marcel W. Coolen, Clare Stirzaker, and Susan J. Clark. Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome research*, 2012.

[70] Sébastien A. Smallwood, Heather J. Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R. Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8):817, 2014.

[71] ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004.

[72] ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799, 2007.

[73] Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330, 2012.

[74] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61, 2012.

[75] Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519, 2012.

[76] Rehan Akbani, Patrick Kwok Shing Ng, Henrica MJ Werner, Maria Shahmoradgoli, Fan Zhang, Zhenlin Ju, Wenbin Liu, Ji-Yeon Yang, Kosuke Yoshihara, and Jun Li. A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nature communications*, 5:3887, 2014.

[77] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, and Nancy Young. The genotype-tissue expression (GTEx) project. *Nature genetics*, 45(6):580, 2013.

[78] Marta Melé, Pedro G. Ferreira, Ferran Reverter, David S. DeLuca, Jean Monlong, Michael Sammeth, Taylor R. Young, Jakob M. Goldmann, Dmitri D. Pervouchine, and Timothy J. Sullivan. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, 2015.

[79] Latarsha J. Carithers, Kristin Ardlie, Mary Barcus, Philip A. Branton, Angela Britton, Stephen A. Buia, Carolyn C. Compton, David S. DeLuca, Joanne Peter-Demchok, and Ellen T. Gelfand. A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreservation and biobanking*, 13(5):311–319, 2015.

[80] Greg Gibson. GTEx detects genetic effects. *Science*, 348(6235):640–641, 2015.

[81] GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[82] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

[83] Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. The human microbiome project. *Nature*, 449(7164):804, 2007.

[84] Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16s rDNA-based community profiling for human microbiome research. *PloS one*, 7(6):e39315, 2012.

[85] Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, Heather H. Creasy, Ashlee M. Earl, Michael G. FitzGerald, and Robert S. Fulton. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207, 2012.

[86] Barbara A. Methé, Karen E. Nelson, Mihai Pop, Heather H. Creasy, Michelle G. Giglio, Curtis Huttenhower, Dirk Gevers, Joseph F. Petrosino, Sahar Abubucker, and Jonathan H. Badger. A framework for human microbiome research. *Nature*, 486(7402):215, 2012.

[87] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R. Mende, Gabriel R. Fernandes, Julien Tap, Thomas Bruls, and Jean-Michel Batto. Enterotypes of the human gut microbiome. *nature*, 473(7346):174, 2011.

[88] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, and Takuji Yamada. A human gut microbial gene catalogue established by metagenomic sequencing. *nature*, 464(7285):59, 2010.

[89] S. Dusko Ehrlich and MetaHIT Consortium. MetaHIT: The European Union Project on metagenomics of the human intestinal tract. In *Metagenomics of the human body*, pages 307–316. Springer, 2011.

[90] Michael Balter. *Taking stock of the human microbiome and disease*. American Association for the Advancement of Science, 2012.

[91] Andrew B. Shreiner, John Y. Kao, and Vincent B. Young. The gut microbiome in health and in disease. *Current opinion in gastroenterology*, 31(1):69, 2015.

[92] Daniel L. Hartl, Andrew G. Clark, and Andrew G. Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997.

[93] Michael Lynch and Bruce Walsh. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.

[94] Doris Vandeputte, Gunter Kathagen, Kevin D'hoe, Sara Vieira-Silva, Mireia Valles-Colomer, João Sabino, Jun Wang, Raul Y. Tito, Lindsey De Commer, Youssef Darzi, Séverine Vermeire, Gwen Falony, and Jeroen Raes. Quantitative microbiome profiling links gut community variation to microbial load. *Nature*, 551(7681):507–511, 2017.

[95] Mukund Thattai and Alexander Van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614–8619, 2001.

[96] Michael B. Elowitz, Arnold J. Levine, Eric D. Siggia, and Peter S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[97] Peter S. Swain, Michael B. Elowitz, and Eric D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, 2002.

[98] Jonathan M. Raser and Erin K. O'shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–2013, 2005.

[99] Mukund Thattai. Universal Poisson Statistics of mRNAs with Complex Decay Pathways. *Biophysical Journal*, 110(2):301–305, January 2016.

[100] Joshua R. Stokell, Raad Z. Gharaibeh, Timothy J. Hamp, Malcolm J. Zapata, Anthony A. Fodor, and Todd R. Steck. Analysis of Changes in Diversity and Abundance of the Microbial Community in a Cystic Fibrosis Patient over a Multiyear Period. *Journal of Clinical Microbiology*, 53(1):237–247, January 2015.

[101] George A. O'Toole. Cystic Fibrosis Airway Microbiome: Overturning the Old, Opening the Way for the New. *Journal of Bacteriology*, 200(4):e00561–17, February 2018.

[102] SunHee Hong, John Bunge, Chesley Leslin, Sunok Jeon, and Slava S. Epstein. Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME Journal*, 3(12):1365, 2009.

[103] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733, 2010.

[104] Kasper D. Hansen, Steven E. Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38(12):e131–e131, 2010.

[105] Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research*, 40(10):e72–e72, 2012.

[106] Nicholas F. Lahens, Ibrahim Halil Kavakli, Ray Zhang, Katharina Hayer, Michael B. Black, Hannah Dueck, Angel Pizarro, Junhyong Kim, Rafael Irizarry, Russell S. Thomas, Gregory R. Grant, and John B. Hogenesch. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biology*, 15(6):R86, June 2014.

[107] J. Paul Brooks, David J. Edwards, Michael D. Harwich, Maria C. Rivera, Jennifer M. Fettweis, Myrna G. Serrano, Robert A. Reris, Nihar U. Sheth, Bernice Huang, Philippe Girerd, Vaginal Microbiome Consortium, Jerome F. Strauss, Kimberly K. Jefferson, and Gregory A. Buck. The truth about metagenomics: quantifying and counteracting bias in 16s rRNA studies. *BMC microbiology*, 15:66, March 2015.

[108] Paul I. Costea, Georg Zeller, Shinichi Sunagawa, Eric Pelletier, Adriana Alberti, Florence Levenez, Melanie Tramontano, Marja Driessen, Rajna Hercog, Ferris-Elias Jung, Jens Roat Kultima, Matthew R. Hayward, Luis Pedro Coelho, Emma Allen-Vercoe, Laurie Bertrand, Michael Blaut, Jillian R. M. Brown, Thomas Carton, Stéphanie Cools-Portier, Michelle Daigneault, Muriel Derrien, Anne Druesne, Willem M. de Vos, B. Brett Finlay, Harry J. Flint, Francisco Guarner, Masahira Hattori, Hans Heilig, Ruth Ann Luna, Johan van Hylckama Vlieg, Jana Junick, Ingeborg Klymiuk, Philippe Langella, Emmanuelle Le Chatelier, Volker Mai, Chaysavanh Manichanh, Jennifer C. Martin, Clémentine Mery, Hidetoshi Morita, Paul W. O'Toole, Céline Orvain, Kiran Raosaheb Patil, John Penders, Søren Persson, Nicolas Pons, Milena Popova, Anne Salonen, Delphine Saulnier, Karen P. Scott, Bhagirath Singh, Kathleen Slezak, Patrick Veiga, James Versalovic, Liping Zhao, Erwin G. Zoetendal, S. Dusko Ehrlich, Joel Dore, and Peer Bork. Towards standards for human fecal sample processing in metagenomic studies. *Nature Biotechnology*, 35(11):1069–1076, November 2017.

[109] Nathan D. Olson and Jayne B. Morrow. DNA extract characterization process for microbial detection methods development and validation. *BMC research notes*, 5:668, December 2012.

[110] Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, April 2009.

[111] Lichun Jiang, Felix Schlesinger, Carrie A. Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R. Gingeras, and Brian Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome research*, 21(9):1543–1551, 2011.

[112] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A. Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A. Teichmann, John C. Marioni, and Marcus G. Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095, November 2013.

[113] Jakob Lovén, David A. Orlando, Alla A. Sigova, Charles Y. Lin, Peter B. Rahl, Christopher B. Burge, David L. Levens, Tong Ihn Lee, and Richard A. Young. Revisiting global gene expression analysis. *Cell*, 151(3):476–482, October 2012.

[114] Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, March 2015.

[115] James H. Bullard, Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11:94, 2010.

[116] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[117] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177, 1982.

[118] Jonathan Friedman and Eric J. Alm. Inferring Correlation Networks from Genomic Survey Data. *PLoS Comput Biol*, 8(9):e1002687, September 2012.

[119] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, August 2012.

[120] Andrew D. Fernandes, Jennifer NS Reid, Jean M. Macklaim, Thomas A. McMurrough, David R. Edgell, and Gregory B. Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16s rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2:15, 2014.

[121] Huaying Fang, Chengcheng Huang, Hongyu Zhao, and Minghua Deng. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics*, page btv349, June 2015.

[122] David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. Proportionality: A Valid Alternative to Correlation for Relative Data. *PLOS Comput Biol*, 11(3):e1004075, March 2015.

[123] Kaifu Chen, Zheng Hu, Zheng Xia, Dongyu Zhao, Wei Li, and Jessica K. Tyler. The overlooked fact: fundamental need of spike-in controls for virtually all genome-wide analyses. *Molecular and Cellular Biology*, pages MCB.00970–14, December 2015.

[124] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010.

[125] Robert Schmieder and Robert Edwards. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLOS ONE*, 6(3):e17288, March 2011.

[126] Susannah J. Salter, Michael J. Cox, Elena M. Turek, Szymon T. Calus, William O. Cookson, Miriam F. Moffatt, Paul Turner, Julian Parkhill, Nicholas J. Loman, and Alan W. Walker. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, 12:87, 2014.

[127] Christopher L. Hemme, Qichao Tu, Zhou Shi, Yujia Qin, Weimin Gao, Ye Deng, Joy D. Van Nostrand, Liyou Wu, Zhili He, Patrick S. G. Chain, Susannah G. Tringe, Matthew W. Fields, Edward M. Rubin, James M. Tiedje, Terry C. Hazen, Adam P. Arkin, and Jizhong Zhou. Comparative metagenomics reveals impact of contaminants on groundwater microbiomes. *Frontiers in Microbiology*, 6, October 2015.

[128] Carl R. Woese, GEORGE E. Fox, Lawrence Zablen, Tsuneko Uchida, Linda Bonen, Kenneth Pechman, Bobby J. Lewis, and David Stahl. Conservation of primary structure in 16s ribosomal RNA. *Nature*, 254(5495):83, 1975.

[129] Carl R. Woese and George E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

[130] George E. Fox, Linda J. Magrum, William E. Balch, Ralph S. Wolfe, and Carl R. Woese. Classification of methanogenic bacteria by 16s ribosomal RNA characterization. *Proceedings of the National Academy of Sciences*, 74(10):4537–4541, 1977.

[131] George E. Fox, Kenneth R. Pechman, and Carl R. Woese. Comparative cataloging of 16s ribosomal ribonucleic acid: molecular approach to procaryotic systematics. *International Journal of Systematic and Evolutionary Microbiology*, 27(1):44–57, 1977.

[132] Carl R. Woese. Bacterial evolution. *Microbiological reviews*, 51(2):221, 1987.

[133] George E. Fox, Jeffrey D. Wisotzkey, and Peter Jurtshuk JR. How close is close: 16s rRNA sequence identity may not be sufficient to guarantee species identity. *International Journal of Systematic and Evolutionary Microbiology*, 42(1):166–170, 1992.

[134] Norman R. Pace, Jan Sapp, and Nigel Goldenfeld. Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1011–1018, January 2012.

[135] John C. Wooley, Adam Godzik, and Iddo Friedberg. A Primer on Metagenomics. *PLOS Comput Biol*, 6(2):e1000667, February 2010.

[136] Philip Hugenholtz and Gene W. Tyson. Microbiology: metagenomics. *Nature*, 455(7212):481, 2008.

[137] Morgan GI Langille, Jesse Zaneveld, J. Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A. Reyes, Jose C. Clemente, Deron E. Burkepile, Rebecca L. Vega Thurber, and Rob Knight. Predictive functional profiling of microbial communities using 16s rRNA marker gene sequences. *Nature biotechnology*, 31(9):814, 2013.

[138] Susannah Green Tringe and Edward M. Rubin. Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics*, 6(11):805–814, November 2005.

[139] Mihai Pop, Alan W Walker, Joseph Paulson, Brianna Lindsay, Martin Antonio, M Anowar Hossain, Joseph Oundo, Boubou Tamboura, Volker Mai, Irina Astrovskaya, Hector Corrada Bravo, Richard Rance, Mark Stares, Myron M Levine, Sandra Panchalingam, Karen Kotloff, Usman N Ikumapayi, Chinelo Ebruke, Mitchell Adeyemi, Dilruba Ahmed, Firoz Ahmed, Meer Taifur Alam, Ruhul Amin, Sabbir Siddiqui, John B Ochieng, Emmanuel Ouma, Jane Juma, Euince Mailu, Richard Omore, J Glenn Morris, Robert F Breiman, Debasish Saha, Julian Parkhill, James P Nataro, and O Colin Stine. Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome Biology*, 15(6):R76, 2014.

[140] Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Comput Biol*, 11(5):e1004226, May 2015.

[141] Matthew C. B. Tsilimigras and Anthony A. Fodor. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of Epidemiology*, 26(5):330–335, May 2016.

[142] The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

[143] Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Article3, 2004.

[144] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[145] M. Senthil Kumar, Eric V. Slud, Kwame Okrah, Stephanie C. Hicks, Sridhar Hannenhalli, and Héctor Corrada Bravo. Analysis and correction of compositional bias in sparse sequencing count data. *BMC genomics*, 19(1):799, November 2018.

[146] Stilianos Louca, Laura Wegener Parfrey, and Michael Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, September 2016.

[147] David M. Karl, Lucas Beversdorf, Karin M. Björkman, Matthew J. Church, Asuncion Martinez, and Edward F. Delong. Aerobic production of methane in the sea. *Nature Geoscience*, 1(7):473, July 2008.

[148] Sara Borin, Lorenzo Brusetti, Francesca Mapelli, Giuseppe D'Auria, Tullio Brusa, Massimo Marzorati, Aurora Rizzi, Michail Yakimov, Danielle Marty, Gert J. De Lange, Paul Van der Wielen, Henk Bolhuis, Terry J. McGenity, Paraskevi N. Polymenakou, Elisa Malinverno, Laura Giuliano, Cesare Corselli, and Daniele Daffonchio. Sulfur cycling and methanogenesis primarily drive microbial colonization of the highly sulfidic Urania deep hypersaline basin. *Proceedings of the National Academy of Sciences*, 106(23):9151–9156, June 2009.

[149] Beth N. Orcutt, Jason B. Sylvan, Nina J. Knab, and Katrina J. Edwards. Microbial Ecology of the Dark Ocean above, at, and below the Seafloor. *Microbiology and Molecular Biology Reviews*, 75(2):361–422, June 2011.

[150] Edward F. DeLong, Christina M. Preston, Tracy Mincer, Virginia Rich, Steven J. Hallam, Niels-Ulrik Frigaard, Asuncion Martinez, Matthew B. Sullivan, Robert Edwards, Beltran Rodriguez Brito, Sallie W. Chisholm, and David M. Karl. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science*, 311(5760):496–503, January 2006.

[151] Brandon K. Swan, Manuel Martinez-Garcia, Christina M. Preston, Alexander Sczyrba, Tanja Woyke, Dominique Lamy, Thomas Reinthaler, Nicole J. Poulton, E. Dashiell P. Masland, Monica Lluesma Gomez, Michael E. Sieracki, Edward F. DeLong, Gerhard J. Herndl, and Ramunas Stepanauskas. Potential for Chemolithoautotrophy Among Ubiquitous Bacteria Lineages in the Dark Ocean. *Science*, 333(6047):1296–1300, September 2011.

[152] Peter J. Turnbaugh, Vanessa K. Ridaura, Jeremiah J. Faith, Federico E. Rey, Rob Knight, and Jeffrey I. Gordon. The effect of diet on the human gut microbiome:

a metagenomic analysis in humanized gnotobiotic mice. *Science Translational Medicine*, 1(6):6ra14, November 2009.

[153] Stephanie C. Hicks, Kwame Okrah, Joseph N. Paulson, John Quackenbush, Rafael A. Irizarry, and Hector Corrada Bravo. Smooth Quantile Normalization. *bioRxiv*, page 085175, November 2016.

[154] Patricio S. La Rosa, J. Paul Brooks, Elena Deych, Edward L. Boone, David J. Edwards, Qin Wang, Erica Sodergren, George Weinstock, and William D. Shannon. Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. *PLOS ONE*, 7(12):e52078, December 2012.

[155] Editorial. Toll Bridges. *Nature Immunology*, 5(10):969, October 2004.

[156] Paul Rossiter, Elizabeth S. Williams, Linda Munson, and Seamus Kennedy. Morbilliviral diseases. *Infectious diseases of wild mammals*, pages 37–76, 2001.

[157] Yusuke Yanagi, Makoto Takeda, and Shinji Ohno. Measles virus: cellular receptors, tropism and pathogenesis. *Journal of General Virology*, 87(10):2767–2779, 2006.

[158] Stephen E. Straus, Jeffrey M. Ostrove, Genevieve InchauspÉ, James M. Felser, Alison Freifeld, Kenneth D. Croen, and Mark H. Sawyer. Varicella-zoster virus infections: biology, natural history, treatment, and prevention. *Annals of internal medicine*, 108(2):221–237, 1988.

[159] Kasper Hoebe, Edith Janssen, and Bruce Beutler. The interface between innate and adaptive immunity. *Nature Immunology*, 5:971–974, October 2004.

[160] Akiko Iwasaki and Ruslan Medzhitov. Regulation of Adaptive Immunity by the Innate Immune System. *Science*, 327(5963):291–295, January 2010.

[161] Ruud Jansen, Jan D. A. van Embden, Wim Gaastra, and Leo M. Schouls. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*, 43(6):1565–1575, March 2002.

[162] Luciano A. Marraffini and Erik J. Sontheimer. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature Reviews Genetics*, 11(3):181–190, March 2010.

[163] Eric S. Lander. The Heroes of CRISPR. *Cell*, 164(1):18–28, January 2016.

[164] Y. Ishino, H. Shinagawa, K. Makino, M. Amemura, and A. Nakata. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *Journal of Bacteriology*, 169(12):5429–5433, December 1987.

[165] A. Nakata, M. Amemura, and K. Makino. Unusual nucleotide arrangement with repeated sequences in the Escherichia coli K-12 chromosome. *Journal of Bacteriology*, 171(6):3553–3556, June 1989.

[166] P. W. Hermans, D. van Soolingen, E. M. Bik, P. E. de Haas, J. W. Dale, and J. D. van Embden. Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains. *Infection and Immunity*, 59(8):2695–2705, August 1991.

[167] F. J. Mojica, C. Ferrer, G. Juez, and F. Rodríguez-Valera. Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. *Molecular Microbiology*, 17(1):85–93, July 1995.

[168] B. Masepohl, K. Görlitz, and H. Böhme. Long tandemly repeated repetitive (LTRR) sequences in the filamentous cyanobacterium Anabaena sp. PCC 7120. *Biochimica Et Biophysica Acta*, 1307(1):26–30, June 1996.

[169] H. P. Klenk, R. A. Clayton, J. F. Tomb, O. White, K. E. Nelson, K. A. Ketchum, R. J. Dodson, M. Gwinn, E. K. Hickey, J. D. Peterson, D. L. Richardson, A. R. Kerlavage, D. E. Graham, N. C. Kyrpides, R. D. Fleischmann, J. Quackenbush, N. H. Lee, G. G. Sutton, S. Gill, E. F. Kirkness, B. A. Dougherty, K. McKenney, M. D. Adams, B. Loftus, S. Peterson, C. I. Reich, L. K. McNeil, J. H. Badger, A. Glodek, L. Zhou, R. Overbeek, J. D. Gocayne, J. F. Weidman, L. McDonald, T. Utterback, M. D. Cotton, T. Spriggs, P. Artiach, B. P. Kaine, S. M. Sykes, P. W. Sadow, K. P. D'Andrea, C. Bowman, C. Fujii, S. A. Garland, T. M. Mason, G. J. Olsen, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. *Nature*, 390(6658):364–370, November 1997.

[170] C. J. Bult, O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. Geoghagen, and J. C. Venter. Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science (New York, N.Y.)*, 273(5278):1058–1073, August 1996.

[171] Y. Kawarabayasi, Y. Hino, H. Horikawa, S. Yamazaki, Y. Haikawa, K. Jin-no, M. Takahashi, M. Sekine, S. Baba, A. Ankai, H. Kosugi, A. Hosoyama, S. Fukui, Y. Nagai, K. Nishijima, H. Nakazawa, M. Takamiya, S. Masuda, T. Funahashi, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, and H. Kikuchi. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, Aeropyrum pernix K1. *DNA research: an international journal for rapid publication of reports on genes and genomes*, 6(2):83–101, 145–152, April 1999.

[172] F. J. Mojica, C. Díez-Villaseñor, E. Soria, and G. Juez. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular Microbiology*, 36(1):244–246, April 2000.

[173] Alexander Bolotin, Benoit Quinquis, Alexei Sorokin, and S Dusko Ehrlich. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of ex-

trachromosomal origin. *Microbiology (Reading, England)*, 151(Pt 8):2551–2561, August 2005.

[174] Francisco J M Mojica, César Díez-Villaseñor, Jesús García-Martínez, and Elena Soria. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of molecular evolution*, 60(2):174–182, February 2005.

[175] C. Pourcel, G. Salvignol, and G. Vergnaud. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology (Reading, England)*, 151(Pt 3):653–663, March 2005.

[176] Reidun K Lillestøl, Peter Redder, Roger A Garrett, and Kim Brügger. A putative viral defence mechanism in archaeal cells. *Archaea (Vancouver, B.C.)*, 2(1):59–72, August 2006.

[177] Kira S Makarova, Nick V Grishin, Svetlana A Shabalina, Yuri I Wolf, and Eugene V Koonin. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology direct*, 1:7, 2006.

[178] Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A Romero, and Philippe Horvath. CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.)*, 315(5819):1709–1712, March 2007.

[179] Luciano A Marraffini and Erik J Sontheimer. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science (New York, N.Y.)*, 322(5909):1843–1845, December 2008.

[180] Hélène Deveau, Rodolphe Barrangou, Josiane E Garneau, Jessica Labonté, Christophe Fremaux, Patrick Boyaval, Dennis A Romero, Philippe Horvath, and Sylvain Moineau. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *Journal of bacteriology*, 190(4):1390–1400, February 2008.

[181] Philippe Horvath, Dennis A. Romero, Anne-Claire Coûté-Monvoisin, Melissa Richards, Hélène Deveau, Sylvain Moineau, Patrick Boyaval, Christophe Fremaux, and Rodolphe Barrangou. Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus thermophilus. *Journal of Bacteriology*, 190(4):1401–1412, February 2008.

[182] Gene W. Tyson and Jillian F. Banfield. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environmental Microbiology*, 10(1):200–207, 2008.

[183] F. J. M. Mojica, C. Díez-Villaseñor, J. García-Martínez, and C. Almendros. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, 155(3):733–740, March 2009.

[184] Rodolphe Barrangou and Philippe Horvath. CRISPR: new horizons in phage resistance and strain identification. *Annual review of food science and technology*, 3:143–162, 2012.

[185] Ido Yosef, Moran G Goren, and Udi Qimron. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic acids research*, 40(12):5569–5576, July 2012.

[186] Avital Brodt, Mor N. Lurie-Weinberger, and Uri Gophna. CRISPR loci reveal networks of gene exchange in archaea. *Biology Direct*, 6(1):65, December 2011.

[187] Adi Stern, Leeat Keren, Omri Wurtzel, Gil Amitai, and Rotem Sorek. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends in genetics: TIG*, 26(8):335–340, August 2010.

[188] Pedro F. Vale, Guillaume Lafforgue, Francois Gatchitch, Rozenn Gardan, Sylvain Moineau, and Sylvain Gandon. Costs of CRISPR-Cas-mediated resistance in Streptococcus thermophilus. *Proc. R. Soc. B*, 282(1812):20151270, August 2015.

[189] Adi Stern and Rotem Sorek. The phage-host arms race: shaping the evolution of microbes. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 33(1):43–51, January 2011.

[190] Kira S Makarova, Yuri I Wolf, and Eugene V Koonin. Comparative genomics of defense systems in archaea and bacteria. *Nucleic acids research*, 41(8):4360–4377, April 2013.

[191] Rotem Sorek, Victor Kunin, and Philip Hugenholtz. CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature Reviews Microbiology*, 6(3):181–186, March 2008.

[192] M. Senthil Kumar and Kevin C. Chen. Evolution of animal Piwi-interacting RNAs and prokaryotic CRISPRs. *Briefings in Functional Genomics*, 11(4):277–288, July 2012.

[193] David Bikard and Luciano A. Marraffini. Control of gene expression by CRISPR-Cas systems. *F1000Prime Reports*, 5, November 2013.

[194] Eugene V Koonin and Kira S Makarova. CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA biology*, 10(5):679–686, May 2013.

[195] Christine L Sun, Rodolphe Barrangou, Brian C Thomas, Philippe Horvath, Christophe Fremaux, and Jillian F Banfield. Phage mutations in response to CRISPR diversification in a bacterial population. *Environmental microbiology*, 15(2):463–470, February 2013.

[196] Timothy R. Sampson and David S. Weiss. Alternative Roles for CRISPR/Cas Systems in Bacterial Pathogenesis. *PLoS Pathog*, 9(10):e1003621, October 2013.

[197] Reuben B Vercoe, James T Chang, Ron L Dy, Corinda Taylor, Tamzin Gristwood, James S Clulow, Corinna Richter, Rita Przybilski, Andrew R Pitman, and Peter C Fineran. Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLoS genetics*, 9(4):e1003454, April 2013.

[198] Rotem Edgar and Udi Qimron. The Escherichia coli CRISPR system protects from λ lysogenization, lysogens, and prophage induction. *Journal of bacteriology*, 192(23):6291–6294, December 2010.

[199] David Paez-Espino, Wesley Morovic, Christine L Sun, Brian C Thomas, Ken-ichi Ueda, Buffy Stahl, Rodolphe Barrangou, and Jillian F Banfield. Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nature communications*, 4:1430, 2013.

[200] Wenyan Jiang, Inbal Maniv, Fawaz Arain, Yaying Wang, Bruce R Levin, and Luciano A Marraffini. Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids. *PLoS genetics*, 9(9):e1003844, 2013.

[201] Ron L Dy, Andrew R Pitman, and Peter C Fineran. Chromosomal targeting by CRISPR-Cas systems can contribute to genome plasticity in bacteria. *Mobile genetic elements*, 3(5):e26831, September 2013.

[202] Joseph Bondy-Denomy and Alan R Davidson. To acquire or resist: the complex biological effects of CRISPR-Cas systems. *Trends in microbiology*, February 2014.

[203] Iwona Mruk and Ichizo Kobayashi. To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. *Nucleic acids research*, 42(1):70–86, January 2014.

[204] Luciano A Marraffini and Erik J Sontheimer. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature*, 463(7280):568–571, January 2010.

[205] Kenn Gerdes, Susanne K Christensen, and Anders Løbner-Olesen. Prokaryotic toxin-antitoxin stress response loci. *Nature reviews. Microbiology*, 3(5):371–382, May 2005.

[206] Tessa E F Quax, Marleen Voet, Odile Sismeiro, Marie-Agnes Dillies, Bernd Jagla, Jean-Yves Coppée, Guennadi Sezonov, Patrick Forterre, John van der Oost, Rob Lavigne, and David Prangishvili. Massive activation of archaeal defense genes during viral infection. *Journal of virology*, 87(15):8419–8428, August 2013.

[207] Jacque C Young, Brian D Dill, Chongle Pan, Robert L Hettich, Jillian F Banfield, Manesh Shah, Christophe Fremaux, Philippe Horvath, Rodolphe Barrangou, and

Nathan C Verberkmoes. Phage-induced expression of CRISPR-associated proteins is revealed by shotgun proteomics in Streptococcus thermophilus. *PloS one*, 7(5):e38077, 2012.

[208] Corinna Richter, James T Chang, and Peter C Fineran. Function and regulation of clustered regularly interspaced short palindromic repeats (CRISPR) / CRISPR associated (Cas) systems. *Viruses*, 4(10):2291–2311, October 2012.

[209] Ksenia Pougach, Ekaterina Semenova, Ekaterina Bogdanova, Kirill A Datsenko, Marko Djordjevic, Barry L Wanner, and Konstantin Severinov. Transcription, processing and function of CRISPR cassettes in Escherichia coli. *Molecular microbiology*, 77(6):1367–1379, September 2010.

[210] Pu Han, Liang Ren Niestemski, Jeffrey E Barrick, and Michael W Deem. Physical model of the immune response of bacteria against bacteriophage through the adaptive CRISPR-Cas immune system. *Physical biology*, 10(2):025004, April 2013.

[211] Jaime Iranzo, Alexander E Lobkovsky, Yuri I Wolf, and Eugene V Koonin. Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context. *Journal of bacteriology*, 195(17):3834–3844, September 2013.

[212] Ariel D Weinberger, Yuri I Wolf, Alexander E Lobkovsky, Michael S Gilmore, and Eugene V Koonin. Viral diversity threshold for adaptive immunity in prokaryotes. *mBio*, 3(6):e00456–00412, 2012.

[213] S. Gandon and P. F. Vale. The evolution of resistance against good and bad infections. *Journal of Evolutionary Biology*, 27(2):303–312, February 2014.

[214] Leah Edelstein-Keshet. *Mathematical models in biology*, volume 46. Siam, 1988.

[215] Umit Pul, Reinhild Wurm, Zihni Arslan, René Geissen, Nina Hofmann, and Rolf Wagner. Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS. *Molecular microbiology*, 75(6):1495–1512, March 2010.

[216] Edze R Westra, Umit Pul, Nadja Heidrich, Matthijs M Jore, Magnus Lundgren, Thomas Stratmann, Reinhild Wurm, Amanda Raine, Melina Mescher, Luc Van Heereveld, Marieke Mastop, E Gerhart H Wagner, Karin Schnetz, John Van Der Oost, Rolf Wagner, and Stan J J Brouns. H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. *Molecular microbiology*, 77(6):1380–1393, September 2010.

[217] Ritsdeliz Perez-Rodriguez, Charles Haitjema, Qingqiu Huang, Ki Hyun Nam, Sarah Bernardis, Ailong Ke, and Matthew P DeLisa. Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in Escherichia coli. *Molecular microbiology*, 79(3):584–599, February 2011.

[218] Yoshihiro Agari, Keiko Sakamoto, Masatada Tamakoshi, Tairo Oshima, Seiki Kuramitsu, and Akeo Shinkai. Transcription profile of Thermus thermophilus CRISPR systems after phage infection. *Journal of molecular biology*, 395(2):270–281, January 2010.

[219] Bard Ermentrout. *Simulating, analyzing, and animating dynamical systems: a guide to XPPAUT for researchers and students*, volume 14. Siam, 2002.

[220] Bruce R Levin, Sylvain Moineau, Mary Bushman, and Rodolphe Barrangou. The population and evolutionary dynamics of phage and bacteria with CRISPR-mediated immunity. *PLoS genetics*, 9(3):e1003312, 2013.

[221] Dan I. Andersson and Diarmaid Hughes. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nature Reviews Microbiology*, 8(4):260–271, April 2010.

[222] Kelli L Palmer and Michael S Gilmore. Multidrug-resistant enterococci lack CRISPR-cas. *mBio*, 1(4), 2010.

[223] Timothy R. Sampson and David S. Weiss. Degeneration of a CRISPR/Cas system and its regulatory target during the evolution of a pathogen. *RNA biology*, 10(10):1618–1622, October 2013.

[224] Honghu Sun, Yinghui Li, Xiaolu Shi, Yiman Lin, Yaqun Qiu, Jinjin Zhang, Yao Liu, Min Jiang, Zhen Zhang, Qiongcheng Chen, Qun Sun, and Qinghua Hu. Association of CRISPR/Cas Evolution with Vibrio parahaemolyticus Virulence Factors and Genotypes. *Foodborne Pathogens and Disease*, 12(1):68–73, January 2015.

[225] Xiangjiao Guo, Yingfang Wang, Guangcai Duan, Zerun Xue, Linlin Wang, Pengfei Wang, Shaofu Qiu, Yuanlin Xi, and Haiyan Yang. Detection and Analysis of CRISPRs of Shigella. *Current Microbiology*, 70(1):85–90, January 2015.

[226] R. Louwen, D. Horst-Kreft, A. G. de Boer, L. van der Graaf, G. de Knegt, M. Hamersma, A. P. Heikema, A. R. Timms, B. C. Jacobs, J. A. Wagenaar, H. P. Endtz, J. van der Oost, J. M. Wells, E. E. S. Nieuwenhuis, A. H. M. van Vliet, P. T. J. Willemsen, P. van Baarlen, and A. van Belkum. A novel link between Campylobacter jejuni bacteriophage defence, virulence and Guillain-Barré syndrome. *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology*, 32(2):207–226, February 2013.

[227] Nigel F. Delaney, Susan Balenger, Camille Bonneaud, Christopher J. Marx, Geoffrey E. Hill, Naola Ferguson-Noel, Peter Tsai, Allen Rodrigo, and Scott V. Edwards. Ultrafast evolution and loss of CRISPRs following a host shift in a novel wildlife pathogen, Mycoplasma gallisepticum. *PLoS genetics*, 8(2):e1002511, February 2012.

[228] Tao Liu, Yingjun Li, Xiaodi Wang, Qing Ye, Huan Li, Yunxiang Liang, Qunxin She, and Nan Peng. Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition. *Nucleic Acids Research*, 43(2):1044–1055, January 2015.

[229] J. D. van Embden, T. van Gorkom, K. Kremer, R. Jansen, B. A. van Der Zeijst, and L. M. Schouls. Genetic variation and evolutionary origin of the direct repeat locus of Mycobacterium tuberculosis complex bacteria. *Journal of Bacteriology*, 182(9):2393–2401, May 2000.

[230] Anders F. Andersson and Jillian F. Banfield. Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. *Science*, 320(5879):1047–1050, May 2008.

[231] Asaf Levy, Moran G. Goren, Ido Yosef, Oren Auster, Miriam Manor, Gil Amitai, Rotem Edgar, Udi Qimron, and Rotem Sorek. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature*, 520(7548):505–510, April 2015.

[232] Stephen P. Hubbell. The unified neutral theory of species abundance and diversity. *Princeton University Press, Princeton, NJ. Hubbell, SP (2004) Quarterly Review of Biology*, 79:96–97, 2001.

[233] Jerôme Chave. Neutral theory and community ecology. *Ecology letters*, 7(3):241–253, 2004.

[234] Mathew A. Leibold and Mark A. McPeek. Coexistence of the niche and neutral perspectives in community ecology. *Ecology*, 87(6):1399–1410, 2006.

[235] Stephen P. Hubbell. Neutral theory in community ecology and the hypothesis of functional equivalence. *Functional ecology*, 19(1):166–172, 2005.

[236] Carl Edward Rasmussen. The infinite Gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000.

[237] Simon J. Labrie, Julie E. Samson, and Sylvain Moineau. Bacteriophage resistance mechanisms. *Nature Reviews. Microbiology*, 8(5):317–327, May 2010.

[238] Aurora M. Nedelcu, William W. Driscoll, Pierre M. Durand, Matthew D. Herron, and Armin Rashidi. On the Paradigm of Altruistic Suicide in the Unicellular World. *Evolution*, 65(1):3–20, January 2011.

[239] Kira S. Makarova, Vivek Anantharaman, L. Aravind, and Eugene V. Koonin. Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. *Biology Direct*, 7:40, 2012.

[240] Edze R. Westra, Angus Buckling, and Peter C. Fineran. CRISPR-Cas systems: beyond adaptive immunity. *Nature Reviews Microbiology*, 12(5):317–326, May 2014.

[241] Angus Buckling and Paul B. Rainey. Antagonistic coevolution between a bacterium and a bacteriophage. *Proceedings. Biological Sciences / The Royal Society*, 269(1494):931–936, May 2002.

[242] Virginie Poullain, Sylvain Gandon, Michael A. Brockhurst, Angus Buckling, and Michael E. Hochberg. The Evolution of Specificity in Evolving and Coevolving Antagonistic Interactions Between a Bacteria and Its Phage. *Evolution*, 62(1):1–11, January 2008.

[243] Alex R. Hall, Pauline D. Scanlan, Andrew D. Morgan, and Angus Buckling. Host-parasite coevolutionary arms races give way to fluctuating selection. *Ecology Letters*, 14(7):635–642, July 2011.

[244] Deo Prakash Pandey and Kenn Gerdes. Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Research*, 33(3):966–976, January 2005.

[245] F. Débarre, S. Lion, M. van Baalen, and S. Gandon. Evolution of host life-history traits in a spatially structured host-parasite system. *The American Naturalist*, 179(1):52–63, January 2012.

[246] Masaki Fukuyo, Akira Sasaki, and Ichizo Kobayashi. Success of a suicidal defense strategy against infection in a structured habitat. *Scientific Reports*, 2, January 2012.

[247] Thomas W. Berngruber, Sébastien Lion, and Sylvain Gandon. Evolution of suicide as a defence strategy against pathogens in a spatially structured environment. *Ecology Letters*, 16(4):446–453, April 2013.

[248] S. Lion and S. Gandon. Evolution of spatially structured host-parasite interactions. *Journal of Evolutionary Biology*, 28(1):10–28, January 2015.

[249] Jaime Iranzo, Alexander E. Lobkovsky, Yuri I. Wolf, and Eugene V. Koonin. Immunity, suicide or both? Ecological determinants for the combined evolution of anti-pathogen defense systems. *BMC Evolutionary Biology*, 15(1):43, March 2015.

[250] Michael Doebeli, Christoph Hauert, and Timothy Killingback. The Evolutionary Origin of Cooperators and Defectors. *Science*, 306(5697):859–862, October 2004.

[251] F. C. Santos and J. M. Pacheco. Scale-free networks provide a unifying framework for the emergence of cooperation. *Physical Review Letters*, 95(9):098104, August 2005.

[252] F. C. Santos, J. M. Pacheco, and Tom Lenaerts. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proceedings of the National Academy of Sciences of the United States of America*, 103(9):3490–3494, February 2006.

[253] Jeffrey A. Fletcher and Michael Doebeli. A simple and general explanation for the evolution of altruism. *Proceedings of the Royal Society of London B: Biological Sciences*, 276(1654):13–19, January 2009.

[254] Jürgen Heinze and Bartosz Walter. Moribund Ants Leave Their Nests to Die in Social Isolation. *Current Biology*, 20(3):249–252, February 2010.

[255] Dominik Refardt, Tobias Bergmiller, and Rolf Kümmerli. Altruism can evolve when relatedness is low: evidence from bacteria committing suicide upon phage infection. *Proceedings of the Royal Society B: Biological Sciences*, 280(1759), May 2013.

[256] Ellen L. Simms and Jim Triplett. Costs and benefits of plant responses to disease: resistance and tolerance. *Evolution*, pages 1973–1985, 1994.

[257] R G Bowers, M Boots, and M Begon. Life history trade-offs and the evolution of pathogen resistance: competition between host strains. *Proceedings. Biological sciences / The Royal Society*, 257(1350):247–253, September 1994.

[258] Wendy L. Fineblum, Mark D. Rausher, and others. Tradeoff between resistance and tolerance to herbivore damage in a morning glory. *Nature*, 377(6549):517–520, 1995.

[259] Michael Boots and Yoshihiro Haraguchi. The Evolution of Costly Resistance in Host-Parasite Systems. *The American Naturalist*, 153(4):359–370, April 1999.

[260] M. Boots and R. G. Bowers. Three mechanisms of host resistance to microparasites-avoidance, recovery and tolerance-show different evolutionary dynamics. *Journal of Theoretical Biology*, 201(1):13–23, November 1999.

[261] B A Roy and J W Kirchner. Evolutionary dynamics of pathogen resistance and tolerance. *Evolution; international journal of organic evolution*, 54(1):51–63, February 2000.

[262] Lars Råberg, Derek Sim, and Andrew F. Read. Disentangling genetic variation for resistance and tolerance to infectious diseases in animals. *Science (New York, N.Y.)*, 318(5851):812–814, November 2007.

[263] Michael Boots, Alex Best, Martin R. Miller, and Andrew White. The role of ecological feedbacks in the evolution of host defence: what does theory tell us? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364(1513):27–36, January 2009.

[264] Jessica A. Hill, Teresa R. O'Meara, and Leah E. Cowen. Fitness Trade-Offs Associated with the Evolution of Resistance to Antifungal Drug Combinations. *Cell Reports*, February 2015.

[265] Martin A. Nowak. Five Rules for the Evolution of Cooperation. *Science*, 314(5805):1560–1563, December 2006.

[266] Karl Sigmund, Christoph Hauert, and Martin A. Nowak. Reward and punishment. *Proceedings of the National Academy of Sciences*, 98(19):10757–10762, September 2001.

[267] James H. Fowler. Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19):7047–7049, May 2005.

[268] Stuart A. West, Ashleigh S. Griffin, Andy Gardner, and Stephen P. Diggle. Social evolution theory for microorganisms. *Nature Reviews Microbiology*, 4(8):597–607, August 2006.

[269] Mayuko Nakamaru and Yoh Iwasa. The coevolution of altruism and punishment: role of the selfish punisher. *Journal of Theoretical Biology*, 240(3):475–488, June 2006.

[270] Hannelore Brandt, Christoph Hauert, and Karl Sigmund. Punishing and abstaining for public goods. *Proceedings of the National Academy of Sciences*, 103(2):495–497, January 2006.

[271] Christoph Hauert, Arne Traulsen, Hannelore Brandt, Martin A. Nowak, and Karl Sigmund. Via Freedom to Coercion: The Emergence of Costly Punishment. *Science*, 316(5833):1905–1907, June 2007.

[272] Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

[273] Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media, 2013.

[274] Yuedong Wang. *Smoothing splines: methods and applications*. Chapman and Hall/CRC, 2011.

[275] Lawrence A. David, Arne C. Materna, Jonathan Friedman, Maria I. Campos-Baptista, Matthew C. Blackburn, Allison Perrotta, Susan E. Erdman, and Eric J. Alm. Host lifestyle affects human microbiota on daily timescales. *Genome Biology*, 15(7):R89, July 2014.

[276] Garrett Jenkinson, Elisabet Pujadas, John Goutsias, and Andrew P. Feinberg. Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nature Genetics*, 49(5):719–729, May 2017.

[277] Karline Soetaert and yale sparse matrix package authors. rootSolve: Nonlinear Root Finding, Equilibrium and Steady-State Analysis of Ordinary Differential Equations, December 2016.

[278] Young-Ju Kim and Chong Gu. Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):337–356, 2004.