UNIVERSITY OF CALIFORNIA, SAN DIEGO

From Microbial Communities to Human Cancer: Methods for Exploring Diversity Across Varying Levels of Biological Organization

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

Doruk Beyter

Committee in charge:

Professor Vineet Bafna, Chair
Professor Jonathan B. Shurin, Co-Chair
Professor Nuno Bandeira
Professor Steven Briggs
Professor Pavel Pevzner

2017

The dissertation of Doruk Beyter is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2017

DEDICATION

To my mother Fadime, my father Zafer,

and my brother Borikim.

EPIGRAPH

It's a dangerous business, Frodo, going out your door. You step onto the road, and if you don't keep your feet, there's no knowing where you might be swept off to.

—Bilbo Baggins (J.R.R. Tolkien)

Those who set off never to rest, shall never tire.

—*Mustafa Kemal Atatürk*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank first and foremost to my advisor, Vineet Bafna, for his repeated assistance, patience, kindness, and lectures, with which I embarked my career in Bioinformatics. I consider myself lucky that he let me attend his class during my exchange studies at UCSD even though I lacked certain pre-requisites, and especially that he kept me as an intern that summer, which became my first research experience. I would also like to thank my committee members Nuno Bandeira for numerous discussions and his extensive help and suggestions, Pavel Pevzner for solidifying my thoughts of pursuing a PhD in Computational Molecular Biology via his legendary undergraduate lecture and useful feedback and questions during my presentations, Jonathan Shurin for the several meetings we had and being an overall great collaborator on my first publication, and finally to Steve Briggs for being my very first collaborator several years ago, even before my PhD studies, and being here until the end, by serving in my committee.

I have to thank Tayfun Ozcelik, my Molecular Biology professor during my freshman year, who presented me the existence of Bioinformatics as a focus-area in Computer Science, and Can Alkan for being a superb example from the same department and alma mater as mine, looking for answers to questions in Molecular Biology. I should also thank Cigdem Gunduz Demir for her advices regarding doing exchange studies at UCSD, and taking classes from Pavel Pevzner and Vineet Bafna, which I apparently followed, and Nitin Udpa for being my very first TA in Bioinformatics, and all his help and guidance though my early years of PhD, and Natalie Castellana for being a great first mentor.

I would like to thank Chris Saunders, and Morten Kallberg, my supervisor and mentor during my internship at Illumina, for their guidance, patience and time for answering my questions, and the overall invaluable experience they provided

me.

Several of my lab mates, colleagues and collaborators also deserve my sincere thanks, for our numerous discussions, their feedback, and guidance. I would like to give my special thanks to Viraj Deshpande, Miin Lin, and Robert McBride for all the work we accomplished, and currently are preparing. I would also like to thank Ali Akbari, Arya Iranmehr, Seungjin Na, Seong Won Cha, Ben Pullman, and Eric Scott, for our several helpful discussions, and Jocelyne Bruand, Hosein Mohimani, Anand Patel, and Kyowon Jeong for their guidance in my first years of PhD, as senior students.

I want to thank my friends Baris Aksanli, Ozgur Yigit Balkan, Furkan Kavasoglu, Efecan Poyraz, Can Bal, Celal Ziftci, Tugcan Aktas, Pinar Sen-Aktas, Matt Sitek, Kirill Makarov, Brendan Duncan, Siddartha Nath, and Rodrigo Ehecatl Duran, Alican Nalci, Hasan Al-Rubaye, Ege Iseri, Can Cagdas Cengiz, and Oguz Ozan Kartal for rendering life simply better.

I would like to thank Victoria Morgan for her constant support and encouragement, for listening to my day, and sharing her experiences with me. I am grateful to have met her.

Most importantly, I would like to thank my family, for their ever unconditional love and support, for being there with me at all times, through the good and bad, regardless the ten thousand miles between us. I would like to thank my mother Fadime, my father Zafer, and my dear brother Bora for being my Borikim, to which I dedicate this dissertation. No word in any languages I am aware of can describe the love and gratitude I have for them – for all that they have done, and the people that they have been.

Chapter 2, in full, is a reprint of the material as it appears in: "Doruk Beyter, Pei-Zhong Tang, Scott Becker, Tony Hoang, Damla Bilgin, Yan Wei Lim, Todd C.

Peterson, Stephen Mayfield, Farzad Haerizadeh, Jonathan B. Shurin, Vineet Bafna, Robert McBride. Diversity, Productivity and Stability of an Industrial Microbial Ecosystem. Applied and Environmental Microbiology, 82(8), 2494-2505, 2016.". The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is currently being prepared for submission for publication of the material, by Doruk Beyter, Miin S. Lin, and Vineet Bafna. The dissertation author was the primary investigator and author of this material.

Chapter 4, in part, is a reformatted reprint of the material as it appears in: "Kristen M. Turner, Viraj Deshpande, Doruk Beyter, Tomoyuki Koga, Jessica Rusert, Catherine Lee, Bin Li, Karen Arden, Bing Ren, David A. Nathanson, Harley I. Kornblum, Michael D. Taylor, Sharmeela Kaushal, Webster K. Cavenee, Robert Wechsler-Reya, Frank Furnari, Scott R. Vandenberg, P. Nagesh Rao, Geoffrey M. Wahl, Vineet Bafna, Paul S. Mischel. Extrachromosomal oncogene amplification drives tumor evolution and the development of genetic heterogeneity in human cancer. Nature, 543(7643), 122-125, 2017.". The dissertation author was a joint primary investigator and author of this material.

VITA

2011	B. S. in Computer Science, Bilkent University, Ankara, Turkey
2011-2017	Graduate Student Researcher, University of California, San Diego, CA
2014	M. S. in Computer Science, University of California, San Diego, CA
2014	Teaching Assistant, University of California, San Diego, CA
2014	Visiting Scientist Intern, Illumina Inc., San Diego, CA
2017	Ph. D. in Computer Science, University of California, San Diego, CA
2017-	Research Scientist in deCODE Genetics/Amgen, Inc., Reykjavik, Iceland

PUBLICATIONS

Kristen M. Turner*, Viraj Deshpande*, Doruk Beyter*, Tomoyuki Koga, Jessica Rusert, Catherine Lee, Bin Li, Karen Arden, Bing Ren, David A. Nathanson, Harley I. Kornblum, Michael D. Taylor, Sharmeela Kaushal, Webster K. Cavenee, Robert Wechsler-Reya, Frank Furnari, Scott R. Vandenberg, P. Nagesh Rao, Geoffrey M. Wahl, Vineet Bafna, Paul S. Mischel. "Extrachromosomal oncogene amplification drives tumor evolution and the development of genetic heterogeneity in human cancer", *Nature*, 543(7643), 122-125, 2017.

Doruk Beyter, Pei-Zhong Tang, Scott Becker, Tony Hoang, Damla Bilgin, Yan Wei Lim, Todd C. Peterson, Stephen Mayfield, Farzad Haerizadeh, Jonathan B. Shurin, Vineet Bafna, Robert McBride. "Diversity, Productivity and Stability of an Industrial Microbial Ecosystem", *Applied and Environmental Microbiology*, 82(8), 2494-2505, 2016.

Sangtae Kim, Konrad Scheffler, Aaron L. Halpern, Mitchell A. Bekritsky, Eunho Noh, Morten Kllberg, Xiaoyu Chen, Doruk Beyter, Peter Krusche, Christopher T. Saunders. "Strelka2: Fast and accurate variant calling for clinical sequencing applications" (submitted)

Doruk Beyter*, Miin S. Lin*, Vineet Bafna. "ProteoStorm: An ultrafast metaproteomics database search framework enabled by multi-staged efficient and sensitive filtering of massive databases" (in preparation for submission) (*) These authors contributed equally.

ABSTRACT OF THE DISSERTATION

**From Microbial Communities to Human Cancer:
Methods for Exploring Diversity Across Varying Levels of Biological Organization**

by

Doruk Beyter

Doctor of Philosophy in Computer Science

University of California, San Diego, 2017

Professor Vineet Bafna, Chair
Professor Jonathan B. Shurin, Co-Chair

Biological diversity can be defined as the total variation of life across levels of biological organization from genes/cells to communities/ecosystems. Exploiting the observed diversity can be of vital interest for environmental, or clinical applications as it may translate into improved responses in community management or patient treatment. Advancements in biological data acquisition technologies such as next-generation sequencing, tandem mass spectrometry or cell imaging enabled scientists explore diversity in complex samples. The high volume of data, however, created

the need for efficient and sensitive computational techniques, to perform useful analyses. In this dissertation, I present three studies, where we explore the presence and the level of biological diversity together with the computational tools and analyses developed for three different data modalities.

First, I describe our computational analysis of the bacterial small subunit rRNA (16S) and the eukaryotic internal transcribed spacer 2 (ITS2) sequencing data of industrial scale open algae ponds, where we explored the associations of community composition and ecosystem variables, over a year. We found that periods of high eukaryotic diversity were associated with high and more stable biomass productivity.

Second, I present ProteoStorm, our computational workflow on performing efficient and sensitive peptide identifications of metaproteomics samples on massive microbial protein databases. Our approach focuses on efficiently reducing the set of candidate peptides for each spectrum, thus obtaining 100 to 1000-fold speedup at the expense of minimal sensitivity. Our re-analysis of urinary tract infection datasets using a comprehensive database, identified bacteria genera previously unknown to be associated with said samples.

Last, I present our study on the landscape of extrachromosomal DNA (ecDNA) in human cancer, where we employed whole-genome sequencing, structural modelling and cytogenetic analyses of 17 different cancer types, including metaphase of $2,572$ dividing cells. I focus on the exploration of the presence and diversity of ecDNA in tumor cells, which we conducted using ECdetect, an image anaysis software I developed. We discovered that ecDNA was found in nearly half of human cancers, and was almost never found in normal cells. Using ECdetect, we were also able to provide estimations on the ecDNA count diversity in tumor cell lines.

Chapter 1

Introduction

Biological diversity can be defined as the total variation of life across levels of biological organization from genes/cells to communities/ecosystems. Exploiting the observed diversity can be of vital interest for environmental, or clinical applications as it may translate into improved responses in community management, more extensive information regarding environmental concerns such as declining species, and the reasons behind them, monitoring for pre-emptive virulence patterns over the globe to prevent epidemics, or simply more targeted approaches during patient treatment. The ever-advancing data collection technologies are continuously producing more data, demanding for more efficient and sensitive computational techniques for comprehensive analyses. This phenomenon is especially pronounced in environments with high potential variation and diversity, as high volumes of data are of special interest in order to detect all variations, common and rare, in a deluge of possibilities.

In this dissertation, I present three studies, where we explore the presence and the level of biological diversity together with the computational tools and analyses developed for three different data modalities. I would like to present some

background and overview in advance, to help the reader.

1.1 Estimating diversity using marker gene sequencing

The staggering speed observed in genomic technologies in the last decade enabled access to high-throughput in a quick and low-cost fashion. Although particularly in multi-species environment sampling studies, traditional approaches including microscopy cell counting, or species culturing are still in use in various settings (e.g., clinical) their low-throughput, labor and time intensive nature, and potential low efficacy is paving the way for the more large scale use of genomic technologies, where information about highly complex environments can be deduced not only in highly practical time frames, but also more accurately.

Open algae ponds, for instance, can be a prime example for a managed ecosystem where constant interaction between three kingdoms (Viridiplantae, Bacteria, and Fungi) is observed – all regulated/affected by the environmental conditions (ecosystem variables) that the ponds are not concealed from. Marker gene (e.g. 16S, 18S, ITS2, ... etc) sequencing technologies are specifically geared towards exploiting the universal existence of "fingerprinting" genes that can be used for identification purposes. Using such technologies, the understanding of the content, behavior, and interaction of highly complex communities, can merely be reduced to assigning the fingerprints to the correct donors. Altough marker genes may not be able to provide a final answer on either diversity, or presence, it can safely provide valuable comparative information accross analyzed samples via seeking convergence patterns in incrementally subsampled data in larger and larger sizes.

In chapter 2, I describe our computational analysis of the bacterial small sub-

unit rRNA (16S) and the eukaryotic internal transcribed spacer 2 (ITS2) sequencing data in industrial scale open algae ponds, where we explored the associations of community composition and ecosystem variables, over a year. We found that periods of high eukaryotic diversity were associated with high and more stable biomass productivity, when controlled for temperature. In addition, bacteria and eukaryotic diversity were inversely correlated over time, possibly due to their opposite response to temperature. Although the addition of temperature as a potential confounding factor has not had a loss of significance in the relationship between diversity and productivity, in the relationship between bacteria and eukaryotic diversity, temperature alone was enough to explain this behavior. Our results have indicated that maintaining diverse communities may be essential to engineering stable and productive bioenergy ecosystems using micro-organisms.

1.2 Searching for peptide needles in particularly large protein haystacks

Genomic technologies such as marker gene sequencing or shotgun (meta) genomics can be highly effective in estimating the content or *potential* expression and available function by setting the universe of all available genes in the analyzed samples; however, it is via transcriptomics and proteomics that *context-specific* expression and function can be learned. Proteomics, specifically, since it can reveal the final product of mRNA, can achieve additional context and information that cannot be obtained in transcriptomics studies.

High throughput tandem mass spectrometry (MS/MS), similar to marker gene sequencing or shotgun genomics enabled comprehensive studies without the focus on single proteins, genes, or other entities of interest. In a multi-species

context, the high throughput nature of MS/MS renders functional information at scale, from regardless how complex the environment of interest can be. This benefit, however, also presents its own challenges in extracting useful information, particularly in higher complexity and data-intensive settings. In a metaproteomics setting, for instance, where spectra are obtained from a large list of different and most usually unknown set of organisms, a common technique will be searching the spectra against a protein database [ESCT11] consisted of the suspected species, called "database search". Whether extra genomic data is provided alongside the proteomic data or not; in highly complex samples, or sample cohorts, using a large search database will be necessary for comprehensive analyses. This necessity, further creates computational challenges by requiring either prohibiting memory requirements or impractical time frames using conventional search engines.

In chapter 3, I present ProteoStorm, our computational workflow on performing efficient and sensitive peptide identifications of metaproteomics samples on massive microbial protein databases. Our approach focuses on efficiently reducing the set of candidate peptides for each spectrum, thus obtaining 100 to 1000-fold speedup at the expense of minimal sensitivity on tested data. Most importantly, our re-analysis of urinary tract infection datasets using a comprehensive database, identified bacteria genera previously unknown to be associated with said samples.

1.3 The landscape of extrachromosomal DNA (ecDNA) in cancer

Circular extrachromosomal DNA as free standing DNA loops in cancer cells, although long known of [BSH63], were yet rarely understood, or have not been accurately characterized/cataloged due to their potential low overall prevalence,

therefore difficult detection. More recently, previous work [FML$^+$11, MJM16, SSG$^+$13] on their prevalence have indicated very low overall occurrence (1.4%), or elevated (31.7%) occurrence on samples analyzed specific to neuroblastoma. Although next generation sequencing (NGS) techniques can be an effective tool in analyzing the content of genes, *localizing* DNA material as extrachromosomal has been out of the capabilities of such technologies. Extensive microscopy imaging, however, at relatively high numbers for each sample analyzed, has proven to be a useful approach in deciphering/estimating the presence of ecDNA in cancer.

The high overlap of several extrachromosomally amplified genes with known oncogenes rendered the ecDNA a functioning mechanism rather than a mere structural oddity.

Most importantly, however, following the localization (thus the counting) of ecDNA, the heterogeneity (i.e. diversity in counts) in ecDNA counts has been of particular intrigue due to its providing of the raw material for a suspected evolution of subject cell lines.

In chapter 4, I present our study on the landscape of extrachromosomal DNA (ecDNA) in human cancer, where we employed whole-genome sequencing, structural modelling and cytogenetic analyses of 17 different cancer types, including metaphase of 2,572 dividing cells. I will focus on the exploration of the presence and diversity of ecDNA in tumor cells, which we conducted using ECdetect, an image analysis software I developed. We discovered that ecDNA was found in nearly half of human cancers, and was almost never found in normal cells. Using ECdetect, we were also able to provide estimations on the ecDNA count diversity in tumor cell lines.

Chapter 2

Diversity, Productivity, and Stability of an Industrial Microbial Ecosystem

2.1 Abstract

Managing ecosystems to maintain biodiversity may be one approach to insuring their dynamic stability, productivity, and delivery of vital services. The applicability of this approach to industrial ecosystems that harness the metabolic activities of microbes has been proposed but never been tested at relevant scales. We used a tag-sequencing approach of bacterial small subunit rRNA (16S) genes and eukaryotic ITS2 to measuring taxonomic composition and diversity of bacteria and eukaryotes in an open pond managed for bioenergy production by micro-algae over a year. Periods of high eukaryotic diversity were associated with high and more stable biomass productivity. In addition, bacteria and eukaryotic diversity were inversely correlated over time, possibly due to their opposite response to

temperature. The results indicate that maintaining diverse communities may be essential to engineering stable and productive bioenergy ecosystems using microorganisms.

2.2 Introduction

Microalgae are one of the most productive photosynthetic organisms on the planet, using sunlight to convert CO_2 and nutrients into biomass which can be used to generate products ranging from high value chemicals such as pigments or nutritional oils to commodities such as protein and biofuels. They can be cultivated on agricultural scales in open ponds using non-arable land and non-potable water, and as such are attractive candidates for the production of low cost biomass [SDBR98, Wal09]. A large limiting factor for reliable low cost biomass production in open ponds is contamination [Cha93, Ric04, Tre04, STHS+08, Shi04]. Managing biological contamination is costly and while it has been achieved in open ponds for the production of high value algae biomass [Lee01, BM13], managing algae stably in open ponds for the production of low cost algae biomass remains challenging [RCZB+09].

Agricultural pesticides or chemicals have been deployed to mitigate the challenges of contamination in algal production systems [MBB+13,ZR13,WR09,LHK83]. Approaches to managing contamination using precepts from ecology have been suggested as a viable low cost alternative [SAD+13, KAS12].This perspective is informed by the idea that traits that determine fitness are not independent and often experience tradeoffs [SAD+13]. For instance, Shurin et al. [SMA14] showed that species that are good N and P competitors generally grow poorly at low light levels. Tradeoffs between other ecologically important functions have also been shown

among algal taxa [LK08]. These tradeoffs can give rise to negative associations between fitness under different conditions or abilities to perform functions such as compete for resources or resist consumers [Chi92, LKSF07, EKL11]. Tradeoffs also imply that in heterogeneous environments open to invaders, maintaining a stable monoculture will be challenging or impossible. In contrast, poly-cultures or ecosystems may be more stable and productive than monocultures [CSD+06]. This assertion has been validated in natural and constructed algal assemblages, where increasing diversity was associated with higher productivity [SGHS12]. Other experiments have indicated that assemblages of algae are more efficient at taking up nutrients and resisting invasion than monocultures [SAD+13], however, more basic research is needed to determine if consortia are a viable option for algae biomass production at industrial scales. Open ponds are very distinct from natural environments experienced by most strains of algae, where they encounter nutrient limitation, consumers and pathogens, sinking, and fluctuating environmental conditions. Whether algae in the nutrient replete and highly productive environments of managed open ponds follow the same patterns observed in natural communities and lab experiments is still an open question.

In this study, we monitored the bacteria and eukaryotic composition of an algae pond managed to optimize biomass productivity over the course of a year. We used 16S and ITS2 (Internal Transcribed Spacer 2) Ion Torrent Personal Genome Machine (PGM) tag-sequencing to assess the bacteria and eukaryotic taxonomic composition and diversity of the pond. We simultaneously monitored a number of aspects of ecosystem structure and function (e.g. nitrate, phosphate, dry weight, fluorescence) to examine the intra- and inter- relationships of ecosystem structure with genomic composition, particularly between microbial diversity and biomass productivity. We asked whether the positive relationships among diversity,

stability and productivity observed in natural and experimental communities of algae were also seen in an engineered environment managed for bioenergy production. Our study seeks to establish the applicability of ecological principles to industrial ecosystems at scales relevant to production of biomass to generate energy or specialized products. Based on ecological theory [PSA+08, SGHS12, SAD+13, SMA14], we expect that periods of high taxonomic diversity should be associated with high and more stable biomass production.

2.3 Material and methods

2.3.1 Pond data collection

The algae were grown in a dirt-lined half acre pond on the Las Cruces, New Mexico, Test Site of Sapphire Energy Incorporated. The pond was filled with water on June 2011, became colonized by green algae and nutrients were added. The pond had a volume of 400 000 liters and was circulated via a pump at an average speed of 10cm/s. The maximum depth of the pond was 30cm. The pH of the pond was maintained at 9 via the addition of CO_2, and biomass was maintained between 0.4 and 0.8g/L, by harvesting (see Figure A.1 for harvest data and biomass). The media, i.e. initial concentrations of the pond, was made up of a salt component to simulate a possible commercial level total dissolved solids (TDS) and salt composition of water not suitable for most agricultural practices. The composition of the media on a liter basis are 3.675g $NaHCO_3$, 4.766g Na_2SO_4, 0.490g KCl, 1.090g NaCl, 0.518g $MgSO_4$-$7H_2O$, 0.146g NaF. The nutrient component of this media on a liter basis is comprised of urea 0.3g, 8.5% H_3PO_4 (v/v) 0.344mL, trace 0.06mL (1g sodium EDTA, 0.194g ferric chloride, 0.072g manganese chloride, 0.021g zinc chloride, 0.013g sodium molybdate, and 0.004g cobalt (II) chloride into 1 L DI H_2O,

sterilized using a Corning 0.22mM filter system and Fe 0.024mL (per liter - versene powder 336.3g and ferix-3 100g). Nutrient addition such as urea, NH_4, NO_3, and PO_4 was performed to maintain the initial state of the pond media, a N level of 100 ppm, and PO_4 level of 40ppm (see Figures A.2 and A.3 for N and PO_4 addition data, together with measured urea and PO_4 levels). The pond was treated on four separate occasions (days 152, 168, 177, and 190) with two commercial fungicides to address a decline in biomass that was suspected to be the result of fungal pathogen outbreak. The active ingredients in the fungicides applied were Fluazinam and Pyraclostrobin. 1 ppm of Fluazinam was applied on days 152, 177, and 190; and 1 ppm of Pyraclostrobin on day 168. McBride et al. [MBB+12] shows the effect of Fluazinam, and Pyraclostrobin on uncontaminated and contaminated algae for various dosage levels, including 1ppm, by observing the culture density (OD 750 nm). According to the study, Fluazinam has a microalgae toxicity for doses greater than 7.5ppm, and Pyroclostrobin for doses greater than 15ppm. Indeed, results in cited document demonstrate higher optical density values at applications of 1ppm doses of Fluazinam or Pyroclostrobin in contaminated algae, whereas these doses show no visible adverse effects to the optical density values on uncontaminated algae [MBB+12].

The pond was regularly monitored for a number of parameters, referred to as "pond ecosystem values" in this paper. Standard measurements such as temperature, pH, OD750, OD560, fluorescence 430/685nm, fluorescence 363/685nm, fluorescence 590/650nm, fluorescence 450/685nm, pond volume, Fv/Fm (PAM), dry weight g/L, alkalinity, NH_4, urea, NO_3, NO_2, PO_4, and harvest volume data were collected. OD and florescence were collected on a SpectraMax plate reader (Molecular devices, Sunnyvale, CA). PAM measures were collected using a PAM Fluorometer (Walz, Effeltrich, Germany). Alkalinity was measured on a TitroLine (Si-Analytics, Mainz,

Germany), NH_4 and urea were measured using colometric assays (Sapphire Energy assay, similar to Seal Analytical, Mequon, Wisconsin), NO_3, NO_2, and PO_4 were measured using an iron chromatography (IC). Dry weight was collected using standard techniques [ZL97].

Approximately every seven days a biological sample was collected from the pond in a 50 mL tube. Samples were taken at a depth of around 15cm from the same location of the pond, which was near its southwest corner. The sample was flash frozen in liquid nitrogen within 4 hours (maximum duration) of collection and stored at $-80°$C until processed for this evaluation. Most samples were collected within 1 hour. Although the maximum duration could have skewed some prokaryotic relative abundance data, Cuthbertson et al. [CRW$^+$14] present an acceptable window of up to 12 hours without significant divergence in bacterial community, though suggests within 1 hour collection as optimal window, as was performed in most of our samples.

The first tag-sequencing sample used in this project corresponds to November 2011.

2.3.2 Sample Sequencing

The PCR amplified products of 16S and ITS2 (see "DNA preparation" in Supplementary Methods) were applied for bi-directional sequencing using Ion Torrent PGM following a modified protocol of "Long Amplicon (400bp) Libraries using a modified long reads Ion XpressTM Plus Fragment Library Kit" (Life Technologies, Carlsbad, CA - Ion Torrent Community website). Briefly, PCR products that contained a phosphate at $5'$ end of each strand were directly ligated to a pair of Ion adaptors, P1 (universal) and A (barcoding) provided in the kit. The 34 samples derived from 16S gene PCR were ligated to Ion barcode-1 to -34

respectively, while the 34 samples derived from ITS2 gene PCR were ligated to Ion barcode-37 to -70 respectively in a 96-well plate. The ligation was performed in a 25μL reaction containing 50-100ng PCR sample, 2μL Ligation Buffer (5x), 1μL dNTP (10mM), 1μL DNA ligase (5 units/μL), 2μL Nick Repair Polymerase, 1μL Adaptor P1 and 1μL of barcoding adaptor A, incubated for 15 minutes at 15°C, and 5 minutes at 72°C. After clean up and size selection using "Magnetic Bead Cleanup Module" (Life Technologies), the ligated samples were pooled together and PCR amplified in 110μL of reaction containing 100μL HiFi Platinum®Taq Supermix (Life Technologies), 5μL P1 and A primer mix and 5μL of pooled samples, under the condition of initial 95°C for 5 minutes followed by 8 cycles of 95°C for 15 seconds, 58°C for 15 seconds and 70°C for 1 minute. After clean-up, the PCR product was quantified by qPCR. The multiple emulsion PCRs were performed to generate template-positive Ion Sphere Particles (ISPs) following the protocol of Ion OneTouchTM 2 System using Template OT2 400 kit (Life Technologies). About 25 to 30 million template-positive ISPs were loaded to each Ion PGM 318 chip. For 16S genes, three chips were sequenced on PGM; while for ITS2 genes, four chips were sequenced on PGM. The FASTQ files from Torrent Server were downloaded and used for downstream data processing.

2.3.3 Sequence analysis

Our ITS2 Ion Torrent PGM data contained an abundance of algal sequences. While OTU-based sequence analysis approaches are widely used and provide pipelines for 16S or fungal-only ITS data in determining sample composition, they are not readily available for ITS data from other eukaryotic taxa. e.g. green algae. Grattepanche et al. [GSMK14] suggest that the cutoffs applied in OTU-based analyses are taxon-dependent, and that tools developed for bacterial

studies (16S data analysis) are not directly applicable for all eukaryotic species (see Supplementary Methods for further discussion). Therefore, we mapped the 16S and ITS2 reads onto selected databases after applying certain quality controls. We also compared the results of our 16S mapping results to an OTU-based approach (see Supplementary Methods) for validation purposes. We obtained Mantel test r-statistics of 0.99, 0.98, 0.94, 0.94, 0.91 with $P = 0.001$ for 999 repetitions, for ranks phylum, class, order, family, and genus, respectively. Diversity results from both approaches had a Pearson $R = 0.96$ with $P = 2.60 \cdot 10^{-14}$. Therefore, we confirm that the taxonomic relative abundance and diversity results in both approaches are highly similar.

In the three chips used for 16S sequencing, we obtained 1.6, 3.8, and 4.4 million reads, while the four ITS2 chips resulted in 3.7, 4.7, 4.6, and 3.7 million reads, respectively (see Figures A.4, A.5, and A.6 for read length distributions). In order to estimate sample compositions and associated taxonomic information, we mapped our reads to the following databases: for 16S data, we used the GreenGenes 16S sequence database (version May 2013, 1.3 million sequences) [DHL+06], and for ITS2 data, we constructed a custom database from NCBI [NCB15a] using the keywords "ITS2" or "internal trancscribed spacer" for sequences with length smaller than $100,000$ under the "Nucleotide" database section, which resulted in 1.1 million sequences. For mapping purposes, we used the alignment software TMAP [NH15b], optimized to deal with variable read lengths, and Ion Torrent specific error profiles [NH15a]. See Supplementary Methods for detailed usage of TMAP. We filtered any read having length shorter than 50 nucleotides, and an error rate higher than 2.0 for 16S reads, and 4.0 for ITS reads, due to their longer average size compared to 16S (see Figure A.6). We accepted any mapping that breached a query coverage of 70% and percent identity of 95% per hit, as applied by

"16S Ribosomal RNA Reference Sequence Similarity Search" by NCBI [NCB15b] (see Supplementary Results and Figure A.7, and A.8 for mapping statistics). For practical purposes, among the $26,135$ and $9,631$ total reference sequences hits in all chips, we picked the top 2000 and 200, corresponding to 97.16% and 96.31% of all hit reads, for 16S and ITS2 data, and used their normalized hitting read counts to represent a sample composition.

We obtained the taxonomic composition of our 16S samples using the GreenGenes taxonomy, and for ITS2 samples, we used the taxonomy database of NCBI using Biopython [CAC+09]. We measured sample composition similarities across all 26 genomic samples, together with the 8 technical replicates, using Bray-Curtis dissimilarity on the top 2000 and 200 sequences' relative abundances, for 16S and ITS2 data, respectively. We provide intra-sample reproducibility assessment in Supplementary Results and Figure A.9.

Here we present our results from chip 3, for both 16S and ITS2 data, due to the higher percent of mapping reads (Figures A.7c and A.8b), while simultaneously confirming reproducibility among different chips using Mantel tests (see Supplementary Methods) achieving r-statistics in the range 0.98-0.99 with all other chips in both datasets.

2.3.4 Diversity analysis

We used Hill numbers to measure diversity with sensitivity parameter, $q = 1$, which is equal to exp(Shannon entropy H) [Hil73, Jos06, CCJ10, LC12]. We computed the Shannon entropy at genus level, after using a rarefaction of 5000 hit reads in all samples for Bacteria, Eukaryota, Viridiplantae, and algae diversities; and 500 for Fungi diversity due to the comparably small number of hits, using 100 iterations. As shown in Supplementary Figures A.11, all rarefaction

curves converged. We used the functions "rrarefy" and "diversity" in package "vegan" [OBK+13] in R. Diversity estimations using the top 200/2000 reference sequences, vs all hits resulted in a Pearson R > 0.99 due to the large portion (96-97%) the top sequences comprised in the community (rarefaction curves shown in Figure A.10).

2.3.5 Ecosystem variables

We collected fifteen pond ecosystem variables on a regular basis, ranging from every day for some variables to a few times a week for others, over the span of a year. We imputed the missing data points on dry weight (g/L) using 750 OD (optical density) as it had more frequent measurements, and it was the variable most strongly correlated with dry weight (Pearson R = 0.85). Since dry weight is the major input in the computation of productivity, and standard deviation in productivity was of interest, imputation was a necessary step in order to have approximately similar number of sample sizes (Figure A.19) in varying time windows. See Figure A.20 for variance patterns in original and imputed DW data. Other ecosystem variables than dry weight (g/L) were either sufficiently sampled or did not require such pre-processing as their standard deviations were not of interest. We used the function "mice.impute.norm.predict", in package "mice" [vGO11] in R.

We removed the outliers (see Supplementary Methods), applied linear interpolation on missing data, and a 7-day central moving average smoothing. To reduce the redundancy in ecosystem variables, we identified highly positively correlated (Pearson R) groups using CAST (Cluster Affinity Search Technique) [BDSY99], with a $\theta = 0.5$, where we measured the pairwise similarities using the Pearson correlation coefficient. After finding the highly collinear ecosystem variable clusters,

we standardized ($\mu = 0$, $\sigma = 1$) all variables, and represented each such cluster using the first principal component of variables inside the cluster – as a technique to represent/combine highly positively correlated variables [Fre11, GNLJ11], and capture the maximum variance in the subject cluster. For example, the six ecosystem variables (560 OD AVG, 750 OD AVG, DW g/L, Chloro1 450/685 nm AVG, Green1 430/685 nm AVG, and Cyano1 383/685 nm AVG) found in one of the ecosystem variable clusters, are variables related to optical density, dry weight, and fluorescence — all sharing biological relevance to each other (see Figure A.18). Since the standardized forms of all six variables in this cluster showed similar patterns in time, i.e. have high correlation to each other, we decided to represent this cluster using their standardized first principal component, explaining 86.18% of the variance of the six ecosystem variables the cluster included.

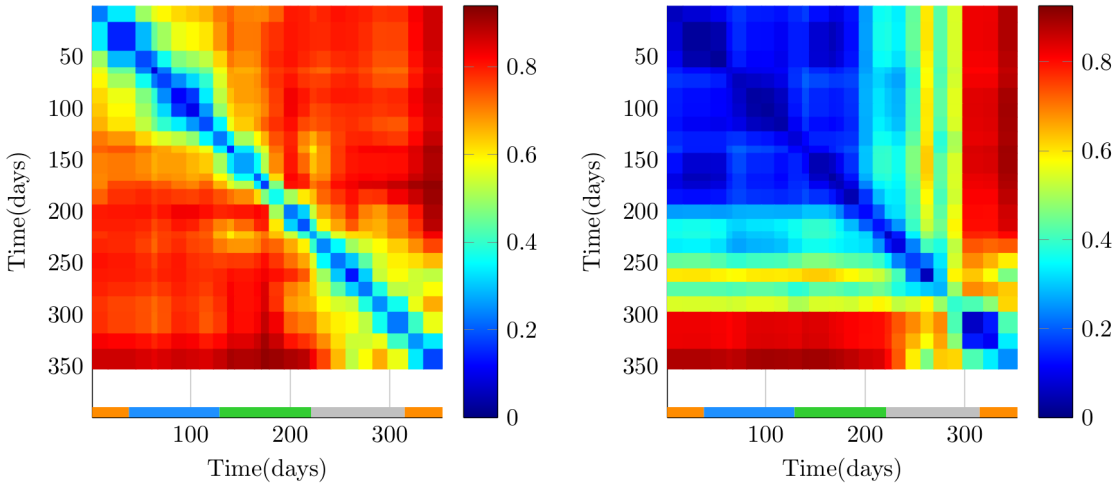
We computed dry weight in terms of kg, by multiplying dry weight (g/L) and pond volume (L) divided by 10^3, and applying a 7-day central moving average smoothing. Finally, we obtained productivity (kg d^{-1}) by subtracting the two consecutive dry weight (kg) measurements in time with no harvesting in between. We also applied the same smoothing approach to our productivity variable. We chose to measure stability in terms of variability, and used standard deviation as the metric, following previous studies [KLH$^+$14, Pim84]. Thus, high stability is associated with low standard deviation in productivity over a window of days.

Ecosystem variables present a time series of data points X_t. To reduce the variance in measurement, we computed statistics (mean and standard deviation) for ecosystem variables over a sliding window of length $2h + 1$ days ($X_{t-h}, \ldots, X_{t+h}$), centered at each time point of the sample. The choice of window-size is based on a trade-off between reduction of measurement noise versus retention of true signal, and we used a published empirical method to identify the appropriate

window size [CCP+11]. Specifically, we experimented with h values of 1 to 6 weeks. The noise in mean, and standard deviation patterns reduced around 3-4 weeks, and stabilized thereafter (see Figure A.22). Moreover, the difference between two distinct peaks (days 165 and 228) in the dry weight (kg) data (see Figure A.12) was a duration of approximately 8 weeks. Therefore, we chose windows with $h = 4$ weeks for our figures; however, we also reported final analysis results on varying window sizes (see discussion on "Relationship between algal diversity and productivity measures" under the Results section).

2.4 Results

2.4.1 Sample dissimilarity over time



(a) 16S sample divergence.

(b) ITS2 sample divergence.

Figure 2.1: Sample Dissimilarities: Panel a shows the Bray-Curtis dissimilarities among the samples between the bacterial (16S) samples, and panel b shows the dissimilarities between the eukaryotic (ITS2) samples. Seasons are denoted with a color bar atop the x axis as fall (orange), winter (blue), spring (green), and summer (silver).

The Bray-Curtis dissimilarities among all samples in 16S data in Figure 2.1a demonstrated two distinct time regions (days 1-100, and 200-350) in composition. Both time regions showed gradual dissimilarity increase over time, however, samples in one of the distinct time regions were at roughly similar distance to all samples in the other. ITS2 data dissimilarities across all samples in Figure 2.1b showed three main distinct regions in time (days 1-200, 200-300, and 300-350), with a fourth inner region (days 250-280). Sample compositions remained highly similar in the first 200 days (a dissimilarity of 0-0.2), and were highly different (dissimilarity of 0.8-0.9) around days 300-350, with an intermediary region of days 200-300. In both bacterial (16S) and eukaryotic (ITS2) samples, we observed that sample compositions have changed overall compositional state and showed high/increasing dissimilarity after around day 200. This roughly corresponded to the beginning of the recovery of algal dry weight (see Figure A.12) after its sharp fall, possibly as a response to the fungicide treatment (see Results section 3.2 for more detail).

2.4.2 Temporal bacteria and eukaryotic taxonomic profile

We observed the temporal taxonomic changes in the bacteria (16S) and eukaryotic (ITS2) composition of our samples using area plots for various taxonomic ranks (Figure 2.2). Area plots consist of stacked relative abundances over time of taxa at different levels of taxonomic resolution. Relative abundances less than 1% (1.3% for bacteria genera) are masked as "Other" for clarity. To examine patterns at an even finer resolution than genus level, we also included histomaps of the top 2000 and 200 reference sequences for 16S and ITS2 data in Figure A.13. We also placed a black shading on the plots between days 152 - 190 as the duration of the four dosages of fungicide, followed by the algal dry weight recovery (day 200). We referred the samples before day 200 as "pre" and the ones after as "post" recovery

samples.

Phylum level 16S composition (Figure 2.2a) revealed that Verrucomicrobia, Proteobacteria, and Cyanobacteria comprised the majority of the taxonomic profile. Proteobacteria decreased in the post recovery samples, while Verrucomicrobia and Cyanobacteria increased in abundance. Analyses at class (Figure 2.2c) and genus levels (Figure 2.2e) revealed that few taxa dominated the phyla present, such as the class Alphaproteobacteria in Proteobacteria, and the genus *Luteolibacter* in Verrucomicrobia. The abundance pattern shifts in these taxa also correspond to the algal dry weight recovery, rendering *Luteolibacter* the most abundant genus in the "post" samples, starting day 204 reaching a high 48%, and occupying 37% of the sample composition by day 350.

The eukaryotic (ITS2) taxonomy analysis at the kingdom level (Figure 2.2b) shows that Viridiplantae, which mainly consists of algal species in our samples, and Fungi constitute the dominant community members across all time points. Although class level composition (Figure 2.2d) was dominated by Chlorophyceae, the genus level analysis (Figure 2.2f) reveals striking changes in the abundance patterns of two genera: *Coalestrum* and *Scenedesmus*. The consistent dominance of *Coalestrum* changed in the "post" recovery samples, followed by a sharp decline by day 300 to be overtaken by *Scenedesmus*.

The decline in the algal dry weight (kg) measurements (see Figure A.12) triggering the fungicide application prior to day 152 (first dosage) coincided with an increased fungal relative abundance period, whereas the time interval between days 152 - 190, where all 4 dosages have been applied, correspond to low (lower than overall mean) fungal relative abundances. We observed that the top five most abundant fungal sequences had high percent identities to Cryptomycota, Chytridiomycota, and Amoeboaphelidium sp., which are reported as algae pathogens

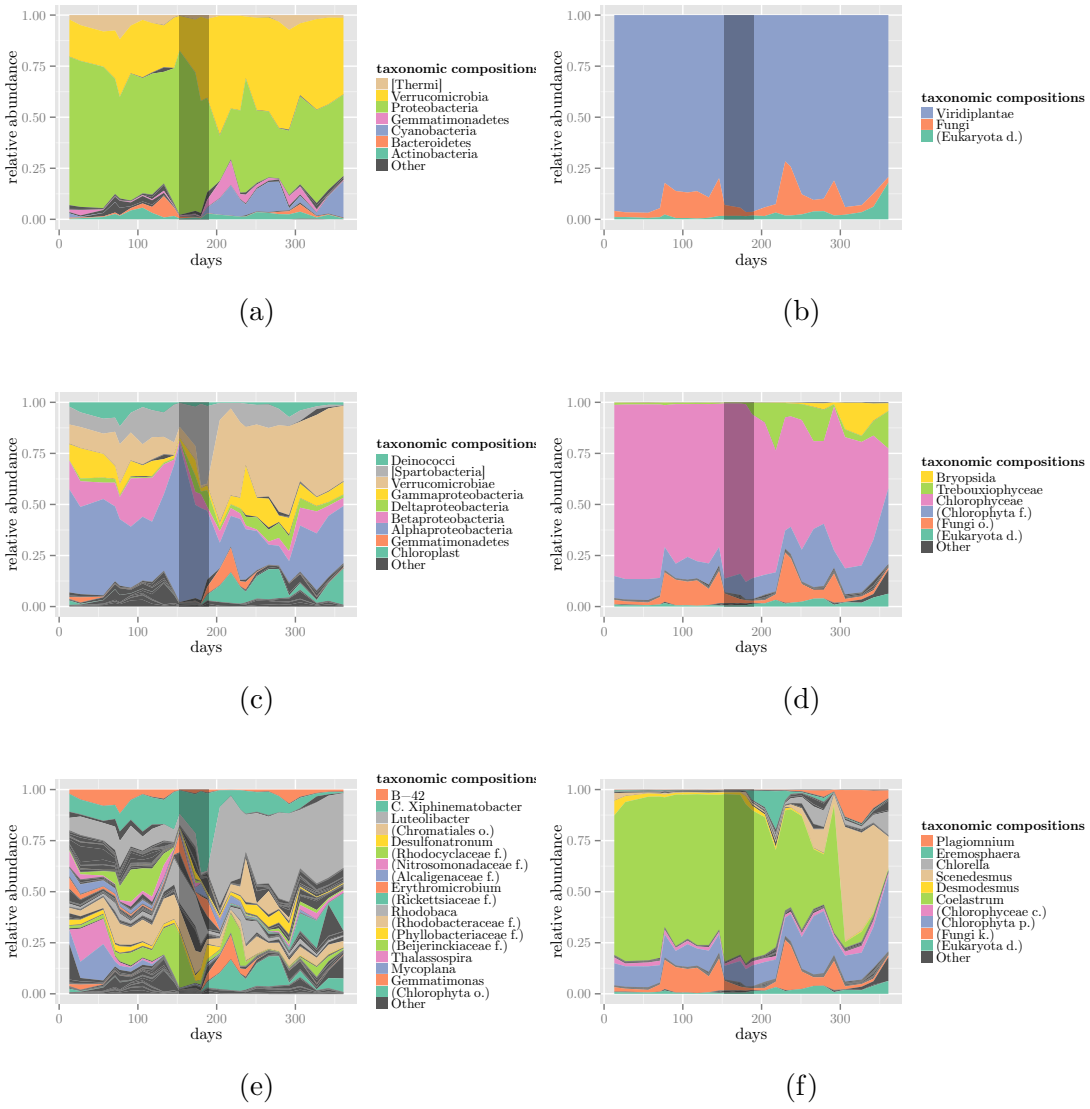


Figure 2.2: Area plots: The plots depict the relative abundances of various taxa and are organized with increasing level of rank in their corresponding taxonomy for 16S (left hand side) and ITS2 (right hand side) compositions. Plots **2.2a** and **2.2b** represent the relative abundances at phlyum and kingdom level, whereas **2.2c/2.2d** and **2.2e/2.2f** further analyzes the compositions at the class and genus levels, respectively. Taxa that had no information at their respective rank are shown in paranthesis using the lowest available taxonomic rank. The black shading between days 152-190 represents the time interval that includes the 4 time point of fungicide application.

in a previous study on a Sapphire Energy open algae pond [LLS+13]. Specifically, references gi|532165669, and gi|532165968 had percent identities (PID) of 87%, and 89% with a Cryptomycota sp. Reference gi|194354257 had 89% PID with a Chytridiomycota sp. (see Figure A.16 for a distance tree result), whereas references gi|532165358 and gi|532166006 had 85% with Amoeboaphelidium sp. PML-2014 isolate FD01, a sequence previously reported by Letcher et al. [LLS+13] on Sapphire Energy ponds. All hit subject sequences were the highest scoring BLAST hits, which contained at least a phylum level annotation, except for gi|532165358. See Figure A.15 for sequence mapping results.

2.4.3 Bacteria and eukaryotic diversity over time

We measured diversity at genus level using Hill numbers, with sensitivity parameter, $q = 1$ after rarefying to an equal number of subsampling on all time samples (see Methods).

We detected a structural break in the temporal diversity trends around the algal dry weight recovery (day 200) in both datasets as shown in Figures 2.3a and 2.3b. The bacteria diversity was high and decreasing in the "pre" period, and remained low in the "post" period, while the eukaryotic diversity showed the opposite trend.

A Chow Test revealed a significant improvement in fit was achieved by modeling the data on two subintervals rather than a regression across the entire time series available ($P < 0.01$ for both 16S and ITS2 data, respectively). In addition, a two-sided Wilcoxon rank sum test showed a significant difference between the median diversities in the two different periods for both bacteria and eukaryotes, where the signal was stronger ($P = 3.05 \cdot 10^{-3}$, and $2.07 \cdot 10^{-7}$, respectively), as shown in Figures 2.3c and 2.3d.

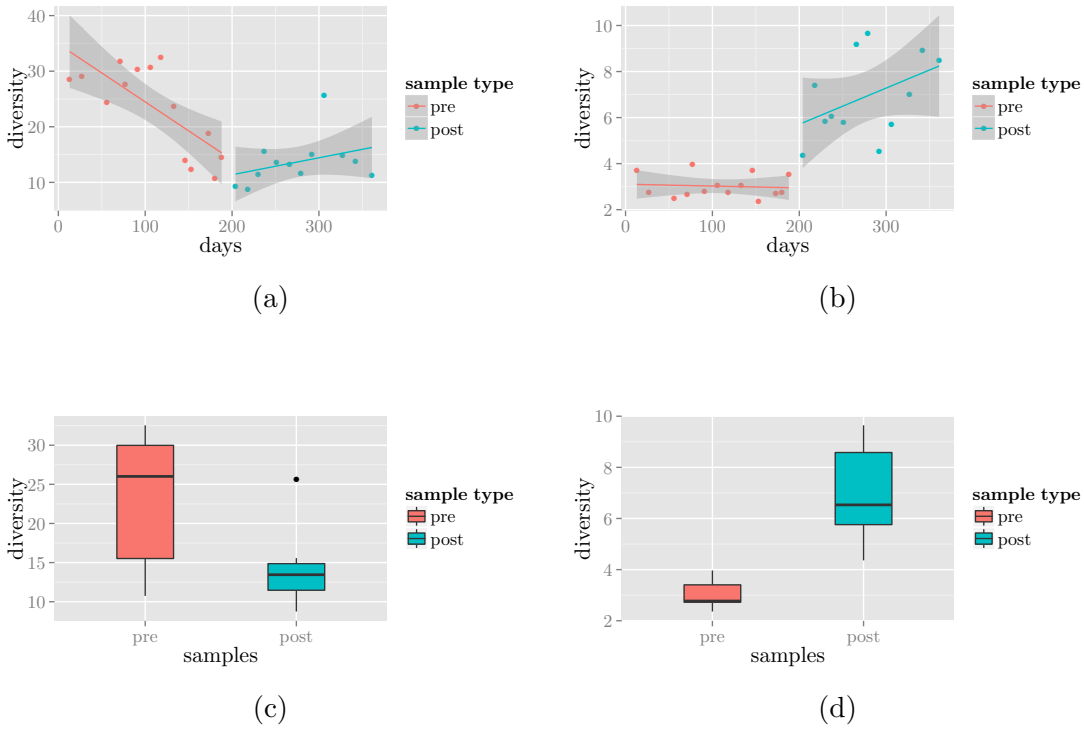


Figure 2.3: Diversity patterns: **2.3a** and **2.3b** show the diversity patterns of bacteria (16S) and eukaryotic (ITS2) data, respectively, in time. **2.3c** (16S) and **2.3d** (ITS2) show the distributions of the diversities at the two different time periods.

We initially observed a significant negative correlation (Pearson R = −0.56, P = < 0.01) between the bacteria and eukaryotic diversities. Controlling for temperature and fungal relative abundance (suspected algal pathogen levels and the effect of fungicide on it), however, revealed that bacterial and eukaryotic diversities had no significant explanatory value to each other (P = 0.62). We also confirmed that fungal relative abundance did not have a significant explanatory value on bacterial or eukaryotic diversity (P = 0.35, and 0.57), after controlling for temperature. We, therefore, think that the initial negative correlation between bacteria and eukaryota diversities could be due to their different responses to temperature. See Supplementary Methods, section 1.7 for controlling for confounding variables and associated model comparison.

2.4.4 Correlations between the pond ecosystem and taxonomic composition

Although the pond was managed to maintain a stable environment through biomass harvesting and nutrient additions, we observed seasonal shifts in the availability of energy and nutrients. Figure A.18 shows seasonal patterns in temperature (an indicator of day length and light availability), the concentration of urea, and Fv/Fm (photosynthetic health). Urea availability peaked in winter (around days 100 and 400), while temperature peaked between days 200 - 300 (summer). Fv/Fm fluctuated strongly, but showed apparent peaks in Spring and Fall (around days 150 and 350), with a decrease in summer, possibly due to the reduction in urea, similar to patterns in some natural phytoplankton communities [ELK13]. The sharp fall in Fv/Fm prior to day 200 could probably be associated with the dry weight fall (see Figure A.12).

The ecosystem variables in the pond showed patterns of collinearity as well

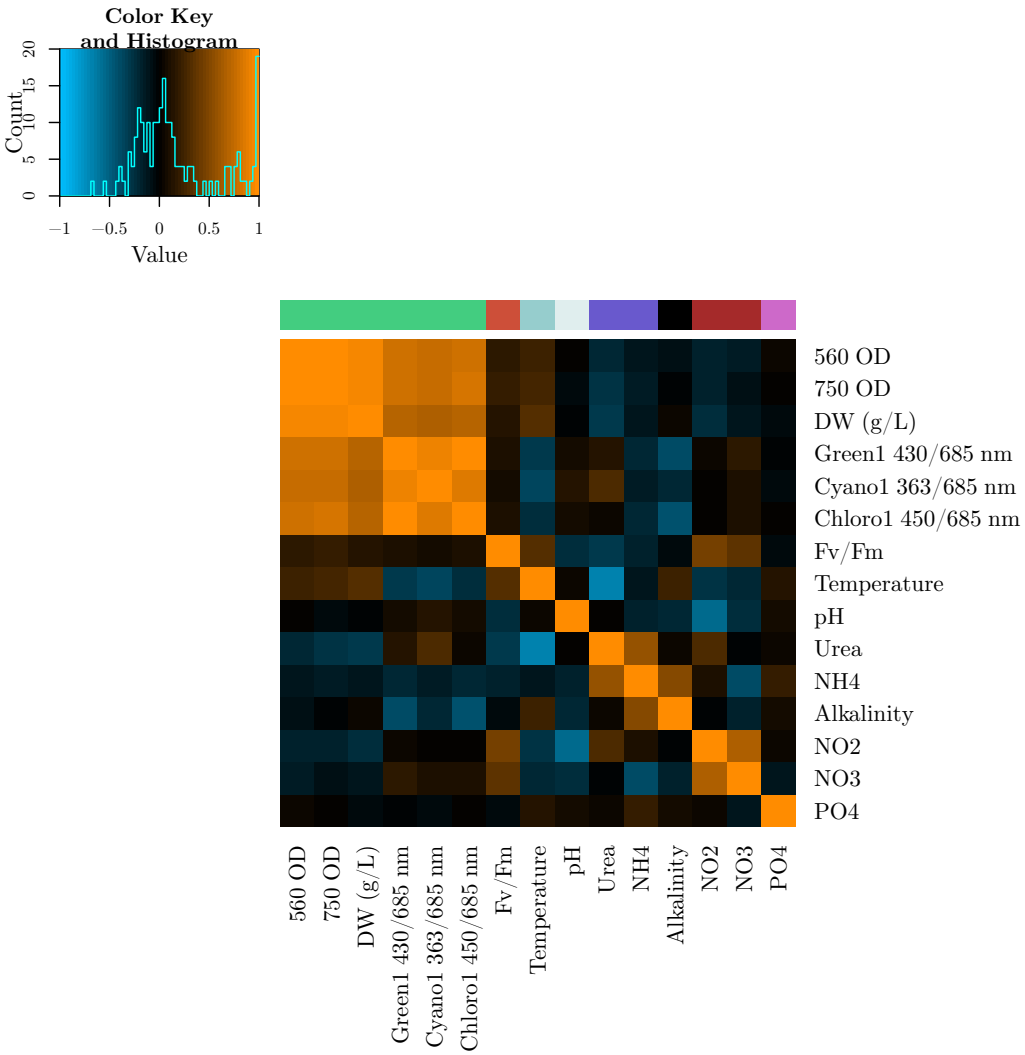


Figure 2.4: Correlation matrix of all phenotypic variables: Ecosystem variables forming a clique using the CAST algorithm are represented with a single color in the colorbar, as also suggested by the orange correlation blocks in the matrix. Cell colors are based on the Pearson correlation coefficients, according to the given colormap.

as associations with the genomic data. Figure 2.4 shows several variables that cluster in blocks of high correlation. We clustered the ecosystem variables using Cluster Affinity Search Technique (CAST) [BDSY99], where pairwise similarities were measured using Pearson correlation. The 15 variables could be described by 8 independent clusters, with a θ of 0.5, which all showed expected grouping (see Table 2.1), including for example the clustering of NO_2 and NO_3. Figure A.21 displays another example ecosystem cluster consisting optical density, fluroescence and dry weight measurements, alongside their standardized first principal component. Since the first principal components of all clusters explained over 75% of their variance as shown in Table 2.1, the final pond ecosystem versus taxonomic composition correlations are conducted using these first principal components.

Heatmaps in Figure 2.5 show the Pearson correlations for kingdom diversities, and bacterial phyla relative abundances versus ecosystem clusters. Kingdom level diversity - pond ecosystem correlation analysis (Figure 2.5a) showed that Bacteria and Viridiplantae had antagonistic correlations with temperature and urea-NH_4 group. Viridiplantae, in addition, showed positive correlation with the DW-fluorescence group, as well. Fungi diversity, on the other hand, was positively correlated with alkalinity, urea-NH_4, and negatively with DW-fluorescence.

Temperature, pH, urea-NH_4, and NO_2-NO_3 groups were the major ecosystem variables to show correlation with the relative abundances of bacteria phyla, as displayed in Figure 2.5b. The row dendrogram also showed that there were two major clusters of bacterial relative abundance patterns at phylum level, based on the correlations with ecosystem variables.

Table 2.1: Phenotype Table: Ecosystem clusters, associated individual ecosystem variables and percent variances explained by their first principle component.

Cluster Name	Ecosystem variables	% variance explained by first PC
DW - fluorescence group	Green1 430/685 nm AVG Chloro1 450/685 nm AVG Cyano1 363/685 nm AVG 560 OD AVG 750 OD AVG DW (g/L)	86.18
Fv/Fm	Fv/Fm AVG	100
pH	pH probe	100
Temperature	Temp probe	100
urea - NH_4 group	urea(ppm) NH_4 (ppm)	78.64
PO_4	PO_4 (ppm)	100
Alkalinity	Alkalinity (ppm)	100
NO_2 - NO_3 group	NO_2 (ppm) NO_3 (ppm)	82.83

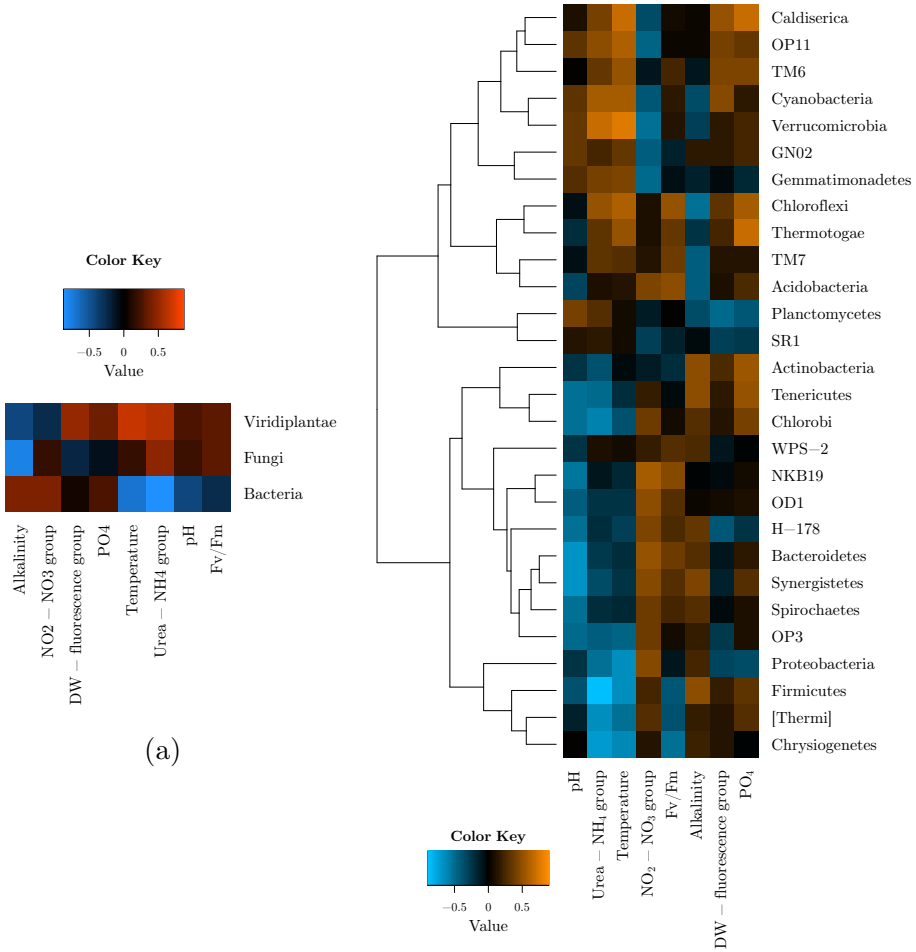


Figure 2.5: Pond ecosystem and taxonomic composition correlations: **2.5a** shows the correlations between ecosystem clusters and diversities at kingdom level, whereas **2.5b** show bacteria phyla relative abundance correlations.

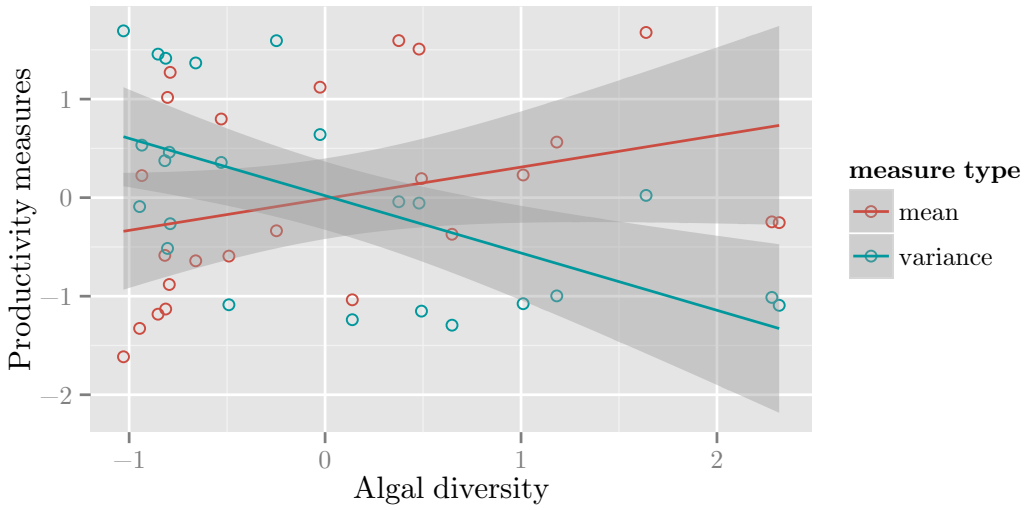


Figure 2.6: Correlations of productivity mean and standard deviation versus algal diversity: The scatter plot shows the correlations between algal diversity vs. mean and standard deviation of productivity measurements centered around genomic sampling days for $2h = 8$ weeks (see Methods) using the regression lines. Algal diversity is positively correlated with mean productivity (Pearson $R = 0.33$, $P = 1.1 \cdot 10^{-1}$) and negatively correlated with standard deviation in productivity, (Pearson $R = -0.6$, $P = 1.9 \cdot 10^{-3}$).

2.4.5 Relationship between algal diversity and productivity measures

We investigated the relationship between algal diversity and the mean and standard deviation of pond productivity measurements (kg d^{-1}), centered at genomic sampling dates (see Methods). We removed the only non-algal genus *Plagiomnium* (class *Bryopsida* (moss)) from the Viridiplantae composition for calculating algal diversity. Figure 2.6 shows the relationship between algal diversity and pond productivity statistics. Algal diversity was positively correlated with mean (Pearson $R = 0.33$, $P = 1.1 \cdot 10^{-1}$) and negatively correlated with standard deviation (sd) in productivity, (Pearson $R = -0.6$, $P = 1.9 \cdot 10^{-3}$), suggesting high stability in biomass production.

In order to control for temperature and fungal relative abundance (suspected algal pathogen levels and the effect of fungicide on it) as potential confounding variables, we used a model comparison using F-test to examine the explanatory value/power of algal diversity on productivity mean and standard deviation. We conducted our analysis on various window sizes (h from 16 to 36). Our results show that algal diversity has significant explanatory value on both productivity mean and sd ($P < 0.5$) for $h = 22$ through $h = 32$ (window sizes of 45 to 65 days), and on productivity sd for $h = 34$, and $h = 36$ as well, as Table 2.2 indicates. Temperature, however, did not have a significant explanatory value (when controlled for algal diversity and fungal relative abundance) on any of the window sizes experimented (see Table 2.3). Although the explanatory values of temperature for $h = 24$, through $h = 28$ had $P < 0.1$ for productivity mean; they had $P > 0.3$ for productivity sd in all window sizes. Since other ecosystem variables (such as urea, NH_4, or PO_4) were highly affected by the maintaining of nutrient supply, unlike temperature, we refrained from adding them into a predictive model.

2.5 Discussion

Open algae ponds as an agricultural platform have the potential to revolutionize the production of low cost biomass for food, fuel and specialty chemicals if their productivity can be optimized and their stability maintained. Research into this effort has generated progress in terms of the scale, productivity and stability of these ponds, however, substantive challenges remain. A novel and potentially transformative solution is to switch from the traditional agricultural paradigm of monocultures to one which deploys multiple strains (polycultures). The first step in this process is to understand if the benefits that have been ascribed to

Table 2.2: Algal diversity explanatory values: Explanatory values of algal diversity controlling for temperature and fungal relative abundance across various h values (half window size) shown as column names. Statistically significant explanatory values are shown in bold.

	16	18	20	22	24	26	28	30	32	34	36
mean	8.2e-02	8.4e-02	7.3e-02	**4.3e-02**	**2.3e-02**	**1.5e-02**	**1.6e-02**	**2.3e-02**	**3.9e-02**	6.9e-02	1e-01
sd	9.9e-02	7.6e-02	5.4e-02	**4.8e-02**	**4.6e-02**	**4.2e-02**	**1.5e-02**	**1.3e-02**	**1.2e-02**	**9.1e-03**	**5.4e-03**

Table 2.3: Temperature explanatory values: Explanatory values of temperature controlling for algal diversity and fungal relative abundance across various h values (half window size) shown as column names

	16	18	20	22	24	26	28	30	32	34	36
mean	1.1e-01	1.3e-01	1.4e-01	1.2e-01	9.3e-02	7.8e-02	8.5e-02	1.3e-01	2.1e-01	3.4e-01	4.6e-01
sd	5.8e-01	5.2e-01	4.2e-01	3.9e-01	4.5e-01	5.6e-01	8.4e-01	8.6e-01	8.6e-01	7.5e-01	6.2e-01

increased diversity in natural systems also occur in open ponds, which are very distinct from most natural systems as algae typically are maintained at a high density and not limited by any resource except for light. In this study, we observed the relationships between algal diversity and both algal productivity and standard deviation of productivity in an open algae pond managed to maintain productivity but open to colonization from aerial sources of microbes. We found a positive relationship between productivity and algal diversity, and a negative relationship between standard deviation in productivity and algal diversity, suggesting that research into how to construct and manage consortia for deployment in open ponds may be an effective tool for pond management, as indicated by studies of natural and experimental systems. [PSA$^+$08, SGHS12, SAD$^+$13, SMA14].

Our study reveals that managed open algae ponds for the production of biomass energy sustain a diversity of microbial life and a dynamic variability. The most common bacteria phyla observed in our study included the Proteobacteria, Verrucomicrobia, and Cyanobacteria, the same groups that dominate natural aquatic assemblages [NJE$^+$11]. Interestingly, the most abundant genus during the high and stable algal biomass yield period, *Luteolibacter*, under Verrucomicrobia, contains species that utilize algal metabolites as carbon and nutrient source, such as *Luteolibacter yonseiensis* and *Luteolibacter algae* [PBW$^+$13, YMA$^+$08]. Community composition also showed seasonal shifts comparable to natural assemblages [WKC$^+$15] even though the environment was managed to achieve relative homeostasis. Our results indicate that diversity and dynamic variability are unavoidable features of open algae ponds that should be incorporated as part of their design and management.

Kingdom level eukaryotic taxonomic composition analysis (Figure 2.2b) revealed three time intervals (days 77-146, days 230-251, and around day 292) with

continuous high (higher than overall mean) fungal relative abundance , with a decrease between the first and second. This decreased fungal relative abundance period (days 147-229) encompassed the four fungicide application time points (days 152, 168, 177, 190). Although we observed a dry weight fall soon after the first high fungal relative abundance time interval, we did not see a similar fall in biomass during/after the other two intervals. We would like to note, however, that the algae community composition was different in across the intervals. While the first time interval coincided with low algal diversity, a more diverse algal community was observed on the other two time intervals. Indeed, Smith. et al. [SC14], and Shurin et al. [SAD+13] discuss the possibility of crop protection against disease/predation through the use of mixed-species communities. Research also shows increased associational resistance against consumers in prey algae assemblages [HC04] due to various possible mechanisms [Duf02]. Although our observation supports the cited findings, control experiments would be required to deduce concrete conclusions.

Disentangling the causal association between diversity and productivity is complicated as diversity can be either a driving factor or a consequence of variation in productivity [CHH+09]. A positive association between pond biomass productivity and diversity of eukaryotes may reflect several underlying processes. First, a more diverse algal community may acquire abiotic resources such as different mineral nutrients [Til81, PSA+08] or wavelengths of light [SHdJ+04] more efficiently due to niche partitioning among species. Sampling effects of randomly selecting high productivity species may occur in assembled communities. Finally, the supply of resources may determine diversity, with a loss of species under pulses of high resource supply [IHH04]. However, nutrients were supplied to our community at a constant high level throughout the course of the study and biomass was maintained by harvesting. Alternatively diversity may not be the ultimate cause of high

productivity or stability but rather may be an associated variable, for unknown reasons. However, our results agree with studies of natural systems showing positive associations between ecosystem productivity and stability and the diversity of the phytoplankton community [PSA$^+$08, ZC14].

Our results showed that algal diversity had significant explanatory value on productivity mean and standard deviation, after controlling for temperature and fungal relative abundance (and the effect of fungicide on it). We acknowledge that the effect of algal diversity on productivity and stability could be confounded by temperature, and the usage of fungicide. Although controlling for temperature is simple, we believe that controlling for the possible confounding effect of fungicide is harder because it is a merely four time point application. Therefore, we chose to use fungal relative abundance as an extra covariate, given the microalgae toxicity values shown in the patent (7.5ppm, and 15ppm), which were higher than the used doses (1ppm) [MBB$^+$12].

Our observations indicate that fungal pathogens may place strong limitations on the productivity and composition of algal biofuel assemblages. These results agree well with data from other algal bioenergy studies [SAD$^+$13, CL14] and natural freshwater ecosystems [KdBIVD07]. Fungal pathogens have been shown to be important in terminating blooms of diatoms [IDBK$^+$04, GdSDNW$^+$13], however their role in maintaining productivity is not well known. Our results indicate that fungi may impose top-down control of productivity similar in magnitude to mesozooplankton grazers like crustaceans, and may therefore shape algal community composition.

Associations between diversity and ecosystem function varied among kingdoms. While we observed a negative correlation between temperature and bacteria diversity, eukaryotic (mostly green algae) diversity showed a positive correlation

with temperature. Indeed, Stomp et al. suggest a positive association between temperature and phytoplankton richness [SHM+11]. It has also been reported that many green algae genera we observed in our samples and Cyanobacteria have optima in higher temperatures, which correspond to the higher spring/summer temperatures at our research site [LdTPK+10]. The bacterial phylum Verrucomicrobia has been shown to be positively correlated with temperature [LKVAZ05], and to include genera (e.g. *Luteolibacter*) to have potential associations with Cyanobacteria [WKC+15]. Our data shows increased *Luteolibacter* relative abundance in periods of increased Cyanobacteria relative abundance and temperature (post day 200, see Figure 2.2e), which have led to the decrease in overall bacterial diversity in higher temperature periods particularly due to the dominance caused by the single genus *Luteolibacter*.

The negative correlation we observed between diversity of phytoplankton and bacteria over time provides some indications of the nature of the eukaryotic and bacteria components of the ecosystem. Producers and microbes engage in a range of pathogenic and mutualistic interactions that may drive positive or negative feedbacks in diversity between the two groups [BWA97]. Phytoplankton and bacterial communities show synchronous dynamics in nature, indicating that bacterial taxa are engaged in specific interactions with phytoplankton taxa [RVGS+05, KYR+07]. Our data indicate that conditions favoring high phytoplankton diversity and productivity are accompanied by low bacterial diversity. The causal basis for this association is unknown; however the correlation could be explained by an opposite response to temperature, since bacterial diversity had no explanatory power on eukaryotic diversity, after controling for temperature. As discussed previously, the relative abundance increases in Luteolibacter and Cyanobacteria during higher temperatures, patterns also observed by [LdTPK+10] and [WKC+15], could have been

the main reasons for diversity loss in bacteria in higher temperatures. Alongside the rising temperature, continuous invasion by airborne propagules of microalgae during a high light availability period could be another possible reason for increased eukaryotic diversity in the post algal dry weight recovery period [SSDB10]. The data therefore give no indication of a causal association between diversity of prokaryotes and eukaryotes.

Managing consortia using traditional tools such as pesticide application could be challenging for consortia stability. The data we collected showed a dramatic impact of pesticide (fungicide) application on the fungal relative abundance, and the recovery of algal dry weight. As mentioned earlier, our data do not allow us to discriminate among several possible causal relationships for this pattern. That said, the fungicide application may have reduced the fitness of the target algae and provided an opportunity for other competing green algae species to begin to enter, thus increasing diversity. Other traditional management tools for open algae ponds may similarly impact consortia in unintended ways. For example, some ponds are harvested using dissolved air flotation (DAF) technology which is commonly used in wastewater treatment. This technology relies on the deployment of a polymer which binds to and aggregates algae based on the surface charge of that algae. The aggregates are then floated to the surface of a DAF tank and skimmed off for further concentration. Without accounting for differential selectivity of this approach on a consortia of algae, harvesting using this strategy would undoubtedly also impact the makeup and stability of a deployed consortia.

Our results indicate that ecological principles relating ecosystem productivity to community diversity are applicable to industrial ecosystems for the cultivation of photosynthetic microbes. Intensifying biomass yield and fostering resilience against the vagaries of the environment or contaminating organisms are keys to

commercializing the industrial growth of microbial products [CRWC10, KAS12, SSDB10, SAD+13]. Most research efforts in this area involve understanding the genetic basis for phenotypic traits related to production of specific compounds [GM12a]. Ecological engineering for productivity and stability has been proposed and discussed [SGHS12], but never demonstrated beyond the laboratory scale. Many ecological processes are highly scale and context dependent [Car96], therefore principles demonstrated in tightly controlled laboratory studies must be validated at whole-system scale under natural regimes of environmental variation in order to ascertain their applicability. Our study indicates that managing microbial polycultures for productivity and stability may form the basis of a viable industrial practice to advance the commercial potential of phytoplankton for bioenergy or other more high value products.

2.6 Acknowledgements

We would like to thank Kalli Lambeth for collecting samples, Sapphire Energy Inc. Las Cruces production team for collecting data on ecosystem variables, and Shibu Yooseph for his insightful comments. DB, RM, JS, SM, TP, FH, YWL and VB were involved in designing the study, as well as providing methods and materials. DB, RM, TH, PT, SB, FH and DB performed the research. DB, RM, JS and VB wrote the manuscript.

Chapter 2, in full, is a reprint of the material as it appears in: "Doruk Beyter, Pei-Zhong Tang, Scott Becker, Tony Hoang, Damla Bilgin, Yan Wei Lim, Todd C. Peterson, Stephen Mayfield, Farzad Haerizadeh, Jonathan B. Shurin, Vineet Bafna, Robert McBride. Diversity, Productivity and Stability of an Industrial Microbial Ecosystem. Applied and Environmental Microbiology, 82(8), 2494-2505, 2016.".

The dissertation author was the primary investigator and author of this material.

Chapter 3

ProteoStorm: An ultrafast metaproteomics database search framework enabled by multi-staged efficient and sensitive filtering of massive databases

3.1 Abstract

Shotgun metaproteomics has been shown to be an effective approach in exploring the functional landscape of complex microbial communities. The necessary usage of large databases in complex samples with unknown bacterial species or strains creates the challenges of heavy computational workload and reduced

sensitivity. We present *ProteoStorm*, an ultrafast multi-staged database search framework, where each stage consists of i) a mass based partitioning of database and spectra, ii) an efficient and sensitive ion mass-indexing based database filtration, and finally iii) the computation of statistically calibrated peptide-spectrum match (PSM) scores via MSGF+. We achieve 100 to 1000-fold speedup compared to using the same search engine without the presented framework on a semi-tryptic search with no variable modifications, on particular large microbial datasets, at the expense of minimal sensitivity. Our re-analysis of urinary tract infection datasets using a comprehensive database, identified bacteria genera previously unknown to be associated with said samples. We further discuss the speed benefit of the usage of partitioned and filtered database searches for practically any search engine with a statistically calibrated and database independent PSM score or p-value.

3.2 Introduction

Metaproteomics, or whole community proteomics is a useful molecular technology in deciphering the functional realm of complex microbial environments employing high throughput tandem mass spectrometry (MS/MS). In a systems-biology perspective, while genomics sets the universe of all available genes in the analyzed samples for *potential* expression, and transcriptomics provides detailed data on the expressed metagenome, proteomics can reveal the final product of mRNA, and give additional context and information that cannot be obtained in RNA transcriptomics studies [HPCG13]. Existing studies [HP13,EMY94] also focus on the potential low correlation of mRNA and peptides, and suggest integrated analyses.

The interpretation of MS/MS data depends on accurately identifying exper-

imental spectra via assigning it to a peptide sequence. One common strategy is to search the acquired MS/MS spectra against a protein database using available database search tools such as SEQUEST [EMY94], Mascot [CBB99], Comet [EJH13], or MS-GF+ [KP14], among others. Following the database search, peptide-spectrum matches (PSMs) and peptides are reported after necessary false-discovery rate filtering [EG10].

Conventional database search algorithms are mainly assumed to operate on small size protein sequence databases (50-100M FASTA files). Although this presents no obstacle in the analysis of single or known and limited number of species proteomics samples, it can be a major computational challenge in complex samples with no prior information of the sample composition, for which large sample-independent databases will have to be used. Alternatively, additional coupled metagenomics or marker gene sequencing (e.g., 16S, 18S, ITS) data may be used to construct a sample-dependent, more focused database as a means to reduce the search database size. Indeed, in a study by Tanca et al. [TPD$^+$13] search results using different size and complexity databases, including reference proteomes and matched metagenomes, are evaluated, however, database specific peptide identifications are established. Similarly, Erickson et al. [ECL$^+$12] reports similar findings, and suggests matched metagenome and reference databases to be complimentary in peptide identification. A recent study [ZNM$^+$16], removes the need for using reference sequences by compiling a database ($>$ 1 million entries) for searching human and mouse microbiata spectra, using more than 1200 metagenomic samples. This approach is only applicable when the environment of interest is well studied via previous metagenomics studies. Also, since the compiled database is already greater than 10^6 entries, the computational challenge remains. Another approach to building a reference database is to use marker gene (e.g. 16S) sequencing

to identify relevant taxa. This, however, requires additional sample preparation and the resulting compiled database may still be too large for a practical time frame run [FPS+12].

Although the usage of large search databases may be prohibitive in achieving high identification rates due to the target-decoy (TD) based false-discovery rate (FDR) control, the absence protein sequences, that are expected to be identified, in the search database can result in matches to incorrect species. As reported in [CSP+16], the re-analysis [KC11] of a honey bee-derived protein sample [BHW+10] using an all species in the NCBI non-redundant database (80 million entries) identified several spectra, previously concluded to be viral and fungal, as honey-bee peptides due to their higher PSM scores. As a result, the usage of comprehensive databases can be crucial.

Existing metaproteomics database search strategies include two-stage searches [JGK+13], where a constrained second level database is formed by performing an initial search, and including the proteins that matched to at least one spectrum regardless of the match score. The conventional search strategy applied in both the intial and second steps, however, may yet pose a challenge in runtime in sufficiently large database and spectra sets. Another strategy [ZNM+16] makes extensive use of previously published environment-specific metagenomic datasets for the construction of a comprehensive yet restricted search database, before using the aforementioned two-stage searching approach. As mentioned above, since rich metagenome sequencing data may not be available for all desired environments, such an approach is limited to the specific environments such data exists. A recent search system published by Chatterjee et al. [CSP+16] addresses the challenges of searching large databases via the usage of MongoDB databases, and the distribution of the workload across several available machines, using an existing search

engine. Although high speedups are achieved in this study, it requires the usage of pre-loaded peptide data via database structures on servers with high RAM capacity (96G). Other common practices include the dividing of the spectra and database and performing several independent searches in parallel.

In all of the strategies mentioned above, where existing conventional search algorithms are used, each spectrum is compared/scored with all peptides within the parent-mass tolerance window of the respective spectrum. This exhaustive scoring scheme results in impractically long runtimes or high memory usage when large search databases are used due to the increased candidate peptides. To address all such challenges, we present ProteoStorm, an ultrafast metaproteomics database search framework enabled by multi-staged efficient and sensitive filtering of massive databases. ProteoStorm makes use of a peptide mass based data partitioning, an ion mass-indexing based database filtration, and an existing sensitive peptide-spectrum p-value generating function by MS-GF+ [KP14]. ProteoStorm achieves orders-of-magnitude speedup, particularly when employed with both large databases and spectra at the expense of minimal sensitivity.

3.3 Methods

3.3.1 Multi-stage ProteoStorm Pipeline

A main assumption in our framework is that in order for a protein to exist in a microbial sample, at least one fully-tryptic peptide belonging to it, needs to be assigned to a spectrum, with a sufficiently high scoring (or low p-value) PSM, with no variable modifications. This assumption motivates ProteoStorm to employ a multi-stage strategy (see Fig 3.1), where an initial fully-tryptic search is followed by a semi-tryptic search. For the semi-tryptic (second) stage, the original large search

database is reduced to merely the proteins containing the fully-tryptic peptides identified in the first stage.

Each stage in ProteoStorm is composed of three core units: i) database and spectra partitioning, ii) efficient and sensitive peptide filtering, and finally iii)peptide-spectrum match (PSM) p-value computing via MS-GF+ [KP14]. Since our focus is the identification of microbial peptides, we remove any spectra matching to Uniprot human proteome with 1% PSM-level FDR, before engaging in the multi-stage microbial database search, in order to be able to perform a conservative microbial peptide identification.

At the end of the second stage, we report the peptides below 1% peptide-level FDR peptides as identifications.

3.3.2 Spectra and database partitioning

When large search databases are used, a common practice is to divide the large FASTA file D into k arbitrary small sized files (chunks) d_i, where $\bigcup_{i=1}^{k} d_i = D$, and search all m spectra files $s_1, ..., s_m$ against all such d_i small databases (typically $\leq$ 200M), so that each of the mn search can be completed with a practical memory requirement, where the spectra file sizes are also limited. One major drawback of this practice is that each of the database d_i and spectra s_j files will be loaded m, and k times redundantly, with k redundant candidate peptide consideration for each spectra set s_j. Furthermore, duplicate peptides across n databases will be re-searched/scored against matching spectra, thus increasing the total runtime.

As shown in Fig 3.2, ProteoStorm addresses these drawbacks by first performing an in-silico digestion of the original large microbial database and retaining the unique set of mass sorted digested peptides. It then bins the set of unique peptides into *database partitions* with a pre-defined mass window in Daltons (Methods). Sim-

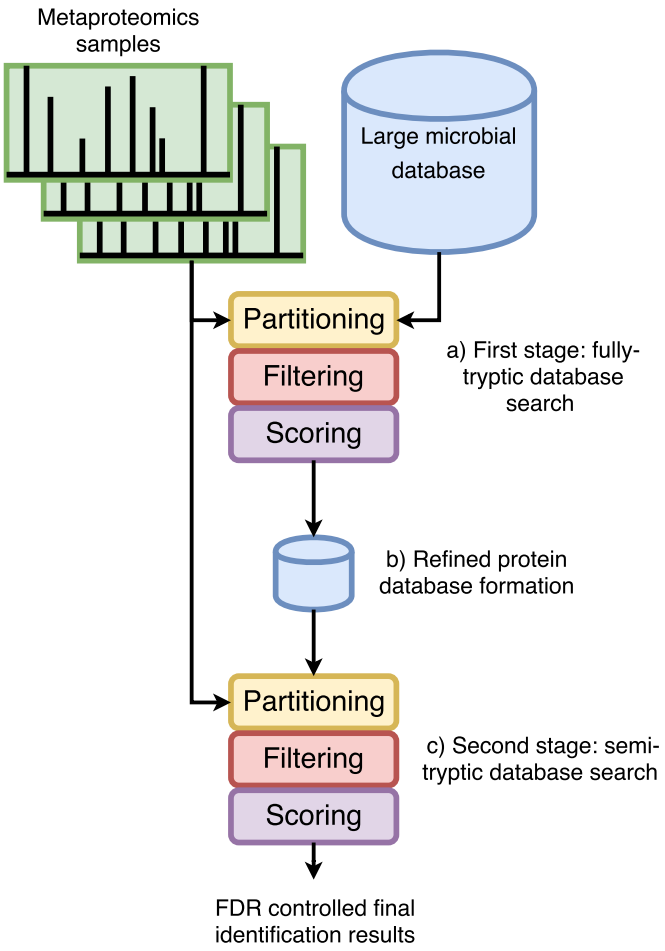


Figure 3.1: ProteoStorm pipeline: ProteoStorm employs two consecutive stages to identify the peptides, in which a fully-tryptic search is followed by a semi-tryptic search on a much smaller database as follows: **(a)** The first stage in silico digests the original large microbial database, partitions both the peptides and spectra by mass, filters the any peptide with insufficient matched peaks with spectra, and finally scores the remaining peptides against the spectra using MSGF+. **(b)** A refined protein database is constructed, with a much smaller size compared to the original microbial database based on the fully-tryptic spectra identifications. **(c)** Similar to **(a)**, the semi-tryptic stage only differs in the the smaller protein database used, and the digestion level. Semi-tryptic peptides are partitioned, filtered, and scored, after which final FDR control is made, and results reported.

ilarly, it places the spectra into the corresponding matching partitions, according to their parent masses to remove the unnecessary candidacy consideration of peptides with distant masses, a priori. As a result, spectra and unique peptide sequences are loaded into the memory only once, and no redundant peptide-spectrum scoring is performed, with the exception of partition boundary spectra for each spectra partition, which are loaded twice (Methods). Database partitioning can be done before the acquiring of spectra, without the knowledge of any parameters regarding the mass spectrometry experiment, for the initial fully-tryptic search stage.

3.3.3 ProteoStorm Filtering

After the partitioning of peptides and spectra by mass into n corresponding pairs, redundant I/O and spectrum-peptide scoring is eliminated. However, in order to be able to compute a match score between a spectrum and a peptide, each spectrum is scored against all peptides within their parent mass tolerance regardless of the overlap between the b- or y-ions of the peptide and the spectrum peaks, in conventional search engines. A potential low overlap may result in the computation of a low match score, which will be discarded in the presence of a higher scoring peptide. This extensive scoring of every peptide within the parent mass tolerance of a spectrum can be quite costly in runtime, especially in the presence of a large database.

Since several b- and y-ions of peptides in a partition may share the same mass (within the fragment tolerance) and can therefore be matched with spectrum peaks or discarded at once, low scoring peptides can indeed be quickly filtered from any extensive match score calculation. To address this, ProteoStorm performs an ultrafast, efficient and sensitive peptide filtering, which we refer as *ProteoStorm Filtering*, by matching the prominent spectra peaks and the theoretical ions of

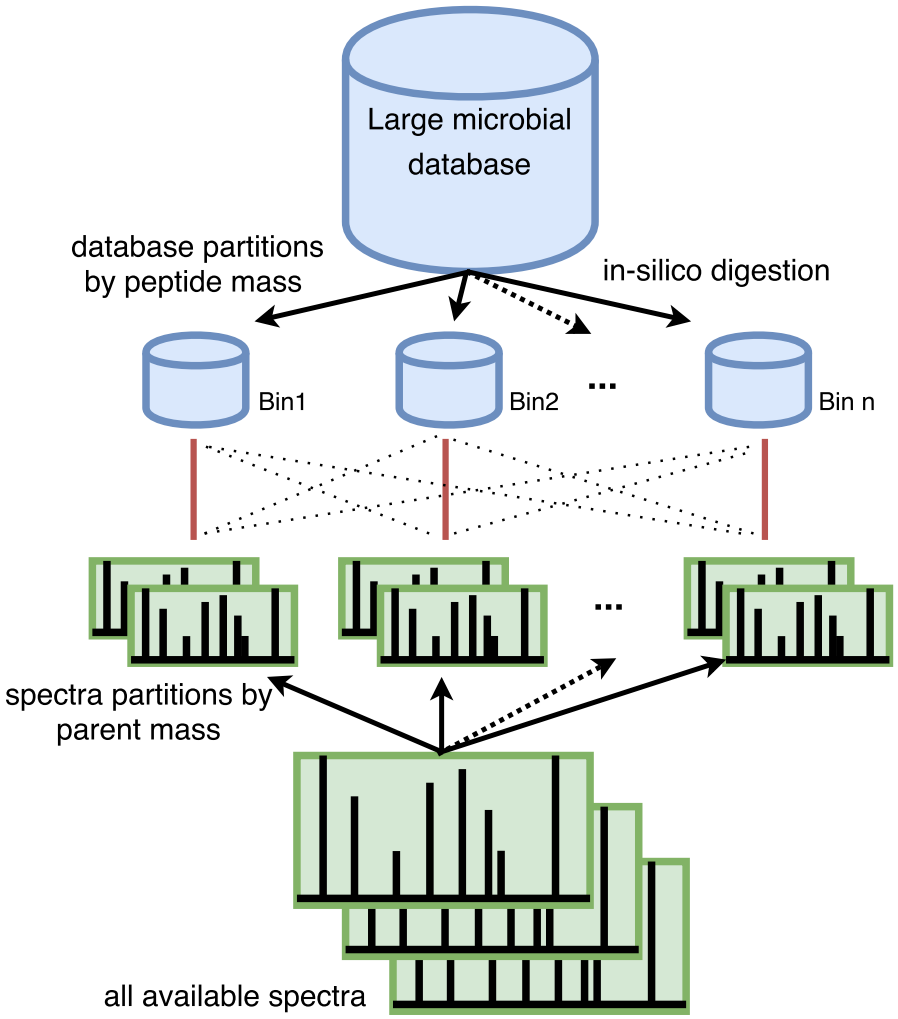


Figure 3.2: Partitioning of peptides: ProteoStorm in-silico digests and partitions the database peptides into n bins based on mass. It finds appropriate mass intervals for each bin, according to the database, and ensures each bin not to be larger than a specified size. Similarly, spectra are also distributed into respective bins that corresponds to the database partitions, given the parent mass tolerance. The red vertical bars between the database-spectra partition pairs shown the only necessary peptide-spectrum comparisons, achieved by the mass binning strategy. The black dotted lines across the partitions depict the alternative situation where all spectra and database partitions would have to be compared against each other, in the absence of any mass based binning, resulting in a much slower procedure.

peptides using an ion mass-indexing based data structure (Methods) similar to [RSW+08, KLA+17, BMT+17], also referred as "peaks-in-common screening" in [Ste95]. The ion-mass index based data structure is the aggregation of the ions of all peptides in a database partition, where the ion-masses are binned into indices, each holding a reference to a list of peptides sorted by parent mass, containing the indexed ion. As presented in Fig 3.3, every spectrum is peak filtered, and searched against an ion-mass index based peptide set data structure via shared spectra peak and theoretical ion indices only. This ion-mass based indexing enables optimal querying of a spectrum in a peptide database as it bypasses all peptides with no matched peaks, and matches all shared spectra peaks and theoretical ions simultaneously. The number of matched peaks between a spectrum and peptide is stored for all candidate peptides.

In order to be sensitive (retain true matches) and efficient (filter as many peptides as possible), for every spectrum, we filtered any candidate peptide with score less than $max(M_{min}, M_{max} - 1)$ ions, where

$$M_{max} = \max_{p_i \in P_s}(f(s, p_i)), \tag{3.1}$$

in which we empirically picked a low $M_{min} = 7$. $f(s, p_i)$ is the number of matched ions between spectrum s and peptide p_i, where P_s is the set of candidate peptides of spectrum s, i.e. peptides with mass within the parent mass tolerance range of spectrum s.

3.3.4 Peptide-spectrum match P-value computation

Finally, to report a peptide-spectrum match (PSM) and compute a P-value for the match, we modified the database search engine MS-GF+ [KP14] in order

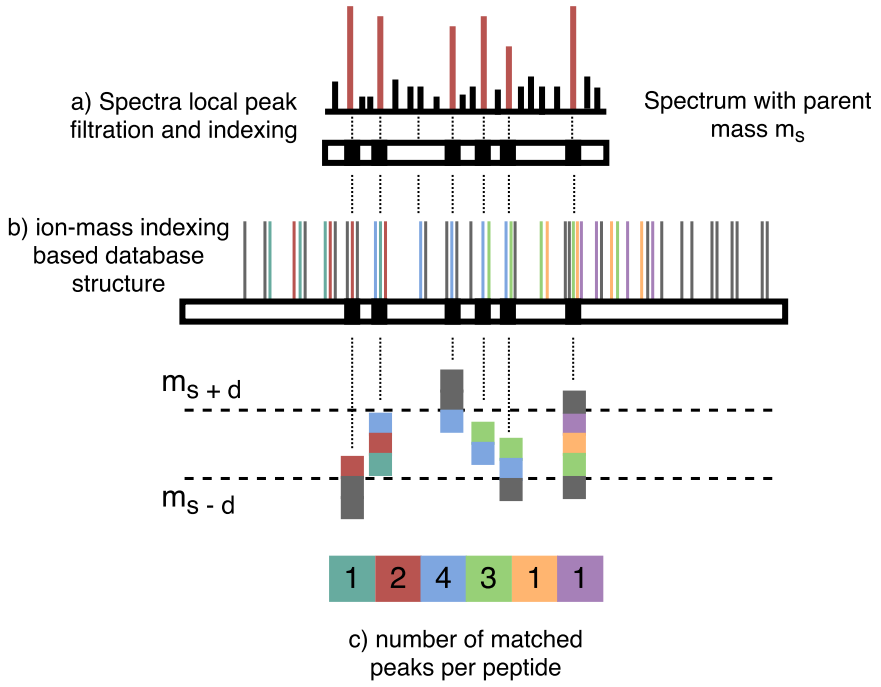


Figure 3.3: Fast filtering of peptides: During the peptide filtering phase, *ProteoStorm* **(a)** applies a window-based filtering on the experimental peaks, and indexes their masses, for every spectra in the currently analyzed spectra partition. **(b)** It then aggregates ions of all peptides the current database partition, and similarly records the ion-mass indices. Each ion-mass index holds a reference to a list of peptides containing an ion of the same ion-mass index (defined by fragment tolerance), sorted by their peptide mass. Each color here represents a unique peptide within the parent mass tolerance window. Peptides outside the window are depicted with grey color. **(c)** Number of matched peaks between the spectra and all candidate peptides (peptides within a parent mass tolerance of d) are computed.

to be able to use its P-value calculation method without performing a database search. Using the spectrum-peptide pairs ProteoStorm Filtering provides, modified MS-GF+ first finds the maximum scoring peptide for the spectrum, then calculates the P-value of the match.

3.3.5 Refined protein database formation

Following the first-stage where fully-tryptic peptides are searched, ProteoStorm performs a 5% peptide-level FDR detection to get a liberal set of tryptic peptide evidence. We chose 5% as a reasonable cutoff for ensuring efficiency and sensitivity. This concludes the first-stage, where we identify a slightly relaxed set of fully-tryptic peptides. ProteoStorm then constructs a refined protein database containing every protein with an exact match to any of the fully-tryptic peptides found in the step above, i.e. without protein inference. The motivation here is to provide the maximal set of sequence variation, given the original microbial database, for the identification of the semi-tryptic peptides – the second-stage.

3.3.6 Second-stage search

In the second-stage, ProteoStorm follows the same three core units as in the first-stage, as shown in Fig 3.1. It in-silico generates all possible semi-tryptic peptides using the refined protein database, sorts by mass and partitions both the peptides and spectra into respective mass bins. One major difference here is that the time it takes to finish this step is included in the total ProteoStorm runtime as the database partitions constructed here are *spectra-specific*, thus not usable for any other set of spectra. Peptides in this stage are filtered using $M_{min} = 6$ for enhanced sensitivity purposes.

3.4 Results

3.4.1 ProteoStorm efficiently searches massive databases with minimal sensitivity loss

We evaluated the performance of ProteoStorm on a urine metaproteomics dataset from urinary-tract infection (UTI) suspected individuals and healthy controls used in [YSBG+15, YSS+17], and compared its performance to a conventional usage of MS-GF+ [KP14]. We used 1.6M spectra from 25 individuals (13 suspected UTI cases, 12 healthy controls), and used the Uniprot KB bacterial database, a 6G fasta file, with 16M entries.

At 1% peptide-level FDR, ProteoStorm identified 13,213 peptides in 0.65 days, whereas conventional MS-GF+ identified 11,834 peptides in an estimated 20 weeks (Methods), achieving 215-fold speedup. Most importantly, 95% of the peptides found in the conventional MS-GF+ search have been also found by ProteoStorm indicating minimal sensitivity loss. This also confirms our initial assumption suggesting the searching of semi-tryptic peptides in a protein, only if there is a fully-tryptic evidence for the protein.

Using a larger spectra dataset (8M spectra), in which we analyzed 122 individuals (110 suspected UTI cases, 12 healthy controls), ProteoStorm completed the search in 2.41 days, whereas conventional MS-GF+ is estimated to complete in 100 weeks, achieving a 290-fold speedup.

Refined databases created in the above datasets were 212MB, and 375MB, respectively.

3.4.2 ProteoStorm reveals previously unknown genera associated with analyzed samples

ProteoStorm have been able to identify bacteria species that were previously unknown to be associated with the samples analyzed in a previous study [YSBG$^+$15]. Among the the top 10 genera based on the unpooled (per individual, per replicate) 1% PSM-level FDR PSMs from pooled 1% peptide-level FDR peptides that are genus-specific, the species *Propionimicrobium lymphophilum*, has not previously been associated with any sample because it was not a part of the search database used in the study. *Propionimicrobium lymphophilum* has also been found to be associated with urinary tract infections in two separate studies [Wil15, IHCD08], using 16S and metagenomic data, respectively.

3.5 Discussion

Thanks to the advancing sequencing efforts, reference protein databases are expected to grow larger in the near future, and will further increase the need for efficient computational tools for the analysis of complex multi-species environments. Although metaproteomics datasets are best suited for the usage of ProteoStorm our workflow can also be employed for practically any dataset with a database size $>$ 200Mb. We believe proteogenomic studies which make use of six-frame translation can also provide good candidate datasets for ProteoStorm due to the large database that may be required.

ProteoStorm Filtering can be an effective stand-alone tool that can be combined with practically any search engine capable of reporting statistically calibrated peptide-spectrum scores, independent of database size or composition. Since ProteoStorm Filtering is a highly sensitive procedure, we suggest its usage

as an efficient means to report a shortlist of spectrum-peptide pairs that can be re-scored with any statistically more rigorous scoring function.

3.6 Acknowledgements

Chapter 3, in full, is currently being prepared for submission for publication of the material, by Doruk Beyter, Miin S. Lin, and Vineet Bafna. The dissertation author was the primary investigator and author of this material.

Chapter 4

Extrachromosomal oncogene amplification drives tumor evolution and the development of genetic heterogeneity in human cancer

4.1 Abstract

Human cells have twenty-three pairs of chromosomes. In cancer, however, genes can be amplified in chromosomes or in circular extrachromosomal DNA (ecDNA), although the frequency and functional importance of ecDNA are not understood [VPV+13, SDGW89, Sch84, FML+11]. We performed whole-genome sequencing, structural modelling and cytogenetic analyses of 17 different cancer types, including analysis of the structure and function of chromosomes during

metaphase of 2,572 dividing cells, and developed a software package called ECdetect to conduct unbiased, integrated ecDNA detection and analysis. Here we show that ecDNA was found in nearly half of human cancers; its frequency varied by tumour type, but it was almost never found in normal cells. Driver oncogenes were amplified most commonly in ecDNA, thereby increasing transcript level. Mathematical modelling predicted that ecDNA amplification would increase oncogene copy number and intratumoural heterogeneity more effectively than chromosomal amplification. We validated these predictions by quantitative analyses of cancer samples. The results presented here suggest that ecDNA contributes to accelerated evolution in cancer.

4.2 Letter

Cancers evolve in rapidly changing environments from single cells into genetically heterogeneous masses. Darwinian evolution selects for survival of the fittest cells, that is, those that are best suited to their environment. Heterogeneity provides a pool of mutations upon which selection can act [VPV$^+$13, Now76, MS15, MAP12, YC12, GM12b]. Cells that acquire fitness-enhancing mutations are more likely to pass these mutations on to daughter cells, driving neoplastic progression and therapeutic resistance [AGJ$^+$16, GVG12]. One common type of cancer mutation, oncogene amplification, can be found either in chromosomes or in nuclear ecDNA elements, including double minutes [SDGW89, Sch84, FML$^+$11, VHNVY$^+$88, GMC$^+$14, CDG$^+$88]. Relative to chromosomal amplicons, ecDNA is less stable, segregating unequally to daughter cells [WDY$^+$91, KOW01]. Double minutes are reported to occur in 1.4% of cancers with a maximum of 31.7% in neuroblastoma, based on the Mitelman database [FML$^+$11, MJM16]. However,

the scope of ecDNA in cancer has not been accurately quantified, the oncogenes contained therein have not been systematically examined and the impact of ecDNA on tumour evolution has yet to be determined.

DNA sequencing permits unbiased analysis of cancer genomes, but it cannot spatially resolve amplicons to specific chromosomal or extrachromosomal regions. Bioinformatic analyses can potentially infer DNA circularity [SSG$^+$13], but the number of extrachromosomal amplicons may vary from cell to cell. Consequently, copies of oncogenes amplified on ecDNA may be greatly underestimated. Cytogenetic analysis of tumour cells during metaphase can localize amplicons, but this technique does not permit unbiased analysis. To quantify the spectrum of ecDNA in human cancer cells and systematically analyse the contents of the ecDNA, we integrated whole-genome sequencing of 117 cancer cell lines, patient-derived tumour cell cultures and tumour tissues from a range of cancer types (Fig. 4.1a) with bioinformatic and cytogenetic analysis of $2,049$ cells in metaphase from 72 cancer cell samples for which cells during metaphase could be obtained. Additionally, 290 cells in metaphase from 10 immortalized cell cultures, and 233 cells in metaphase from 8 normal tissue cultures were analysed, with a total of $2,572$ cells in metaphase analysed.

The fluorescent dye DAPI (4,6-diamidino-2-phenylindole) allows ecDNA detection (Fig.4.1b), which was confirmed using genomic DNA and centromeric FISH (fluorescence in situ hybridization) probes (Fig. 4.1b-d and Extended Data Fig. B.1). We developed an image analysis software package called ECdetect (Fig. 4.1e and Methods), providing a robust, reproducible and highly accurate method for quantifying ecDNA from DAPI-stained metaphases in an unbiased, semi-automated fashion. ECdetect accurately detected ecDNA and this detection rate was highly correlated with visual detection ($r = 0.98$, $P < 2.2 \cdot 10^{-16}$; Fig. 4.1f), allowing the

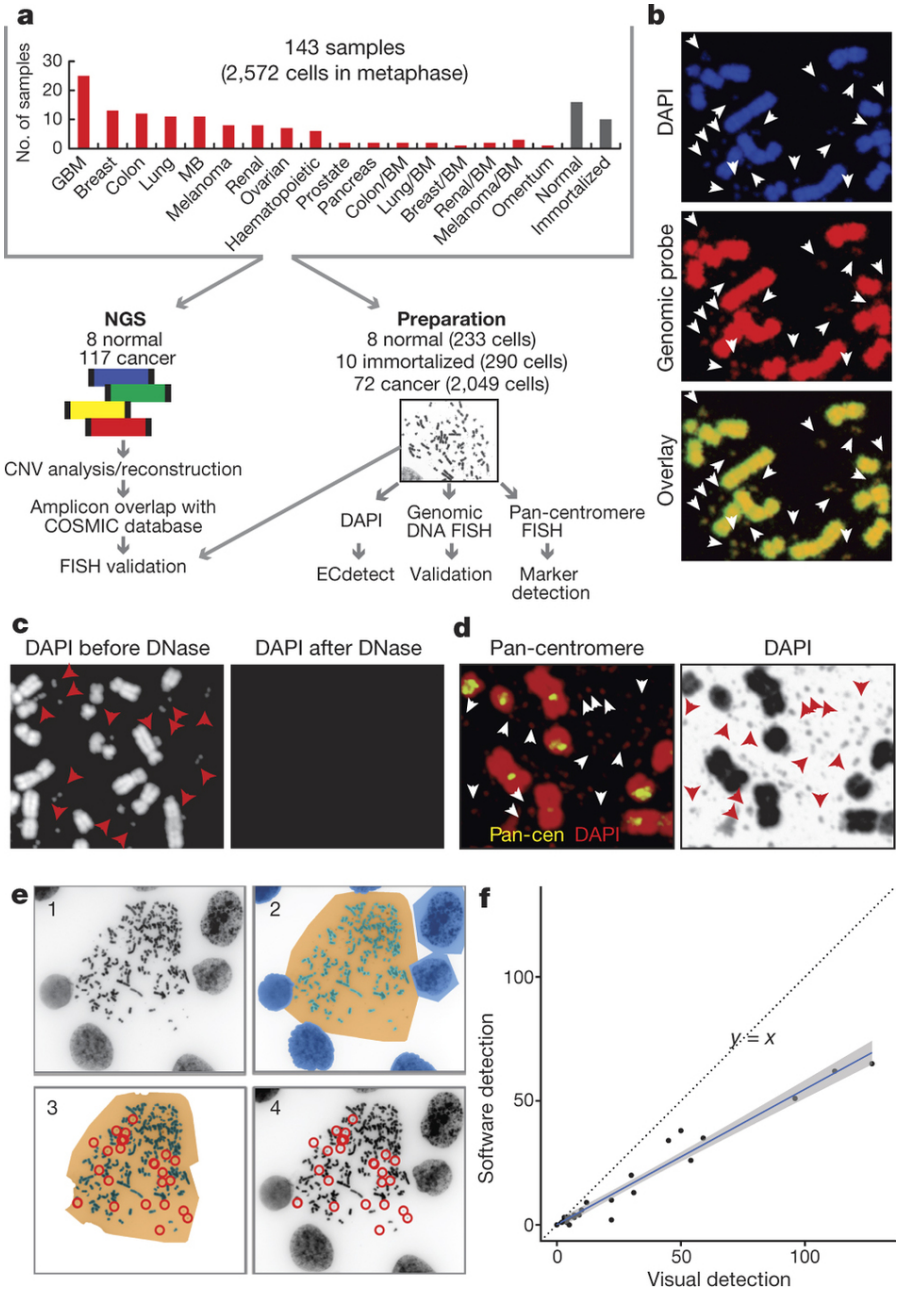


Figure 4.1: Integrated next-generation DNA sequencing and cytogenetic analysis of ecDNA: a, Schematic diagram of experimental flow. BM, brain metastasis; GBM, glioblastoma; MB, medulloblastoma. b, Representative cells during metaphase stained with DAPI and a genomic DNA FISH probe (ecDNA, arrows). c, DNase treatment abolishes DAPI staining of chromosomal and ecDNA (arrows). d, Pan-centromeric FISH shows that a centromere in the ecDNA is absent (arrows). e, Schematic illustration of ECdetect. (1) DAPI-stained metaphase as input, (2) semi-automated identification of ecDNA search region through segmentation, (3) conservative filtering, removing non-ecDNA components and (4) ecDNA detection and visualization. f, Pearson correlation between software-detected and manual calls of ecDNA ($r = 0.98$, $P < 2.2 \cdot 1016$)

quantification of 2,572 cells in metaphase, including at least 20 cells in metaphase from each sample.

ecDNA was abundant in the cancer samples (Fig. 4.2a), but was rarely found in normal cells. Approximately 30% of the ecDNA were paired double minutes . ecDNA levels varied among tumour types, with substantially higher levels in patient-derived cultures (Fig. 4.2b). Using the conservative metric of at least two ecDNA copies in $\geq$ 10% (2 out of 20) cells in metaphase, ecDNA was detected in nearly 40% of tumour cell lines and nearly 90% of patient-derived brain tumour models (Fig. 4.2c, d, Extended Data Fig. B.2 and Methods). No significant associations between ecDNA level and primary tumour or metastatic status; untreated or treated samples; or un-irradiated or post-irradiated tumours were detected . The diverse array of treatments relative to the sample size limited our ability to conclusively determine the effect of specific therapies on ecDNA levels. ecDNA number varied greatly from cell to cell within a tumour culture (Fig. 4.2e-g, Extended Data Fig. B.3 and), as quantified by the Shannon diversity index [ACR$^+$14]. These data demonstrate that ecDNA is common in cancer cells, varies greatly from cell to cell and is very rare in cells derived from normal tissue.

Whole-genome sequencing with a median coverage of $1.19X$ (Extended Data Fig. B.4) showed focal amplifications that were nearly identical to the amplifications found in The Cancer Genome Atlas (TCGA) analyses of the same cancer types (Fig. 4.3a), including amplified oncogenes found in a pan-cancer analysis of 13 different cancer types [ZSC$^+$13]. All of the amplified oncogenes tested were found solely in the ecDNA, or concurrently in ecDNA and chromosomal homogenous staining regions (HSRs) (Fig. 4.3b, c and Extended Data Figs B.5, B.6). Oncogenes amplified in ecDNA showed high expression levels of mRNA transcripts (Fig. 4.3d) and the copy-number diversity of commonly amplified oncogenes in ecDNA far

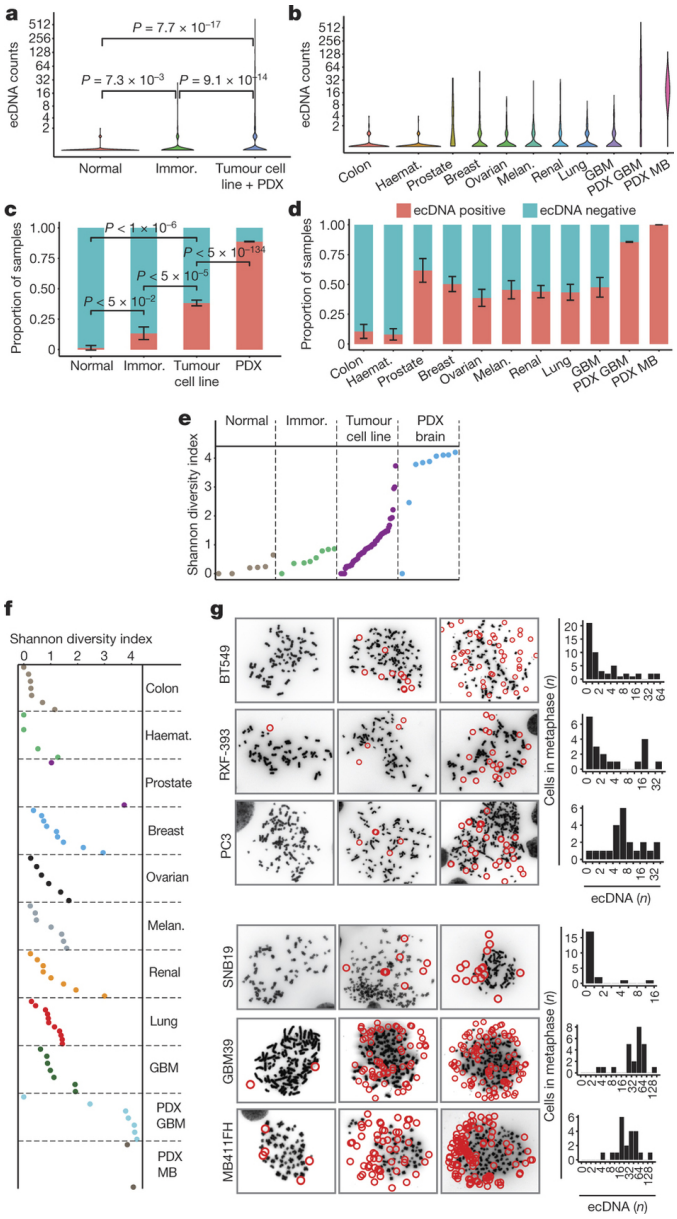


Figure 4.2: ecDNA is found in nearly half of cancers and contributes to intratumoural heterogeneity: a, Distribution of ecDNA elements per cell in metaphase from 72 cancer, 10 immortalized and 8 normal cell cultures, Wilcoxon rank-sum test. PDX, patient-derived xenograft. b, ecDNA distribution per cell in metaphase stratified by tumour type. c, Proportion of samples with two or more ecDNA elements in at least two out of 20 cells (positive for ecDNA) in metaphase. Data shown as mean$\pm s.e.m.$ (Methods). d, Proportion of tumour cultures positive for ecDNA by tumour type. e, Shannon diversity index. Each dot represents an individual cell line sampled with$\geq$ 20 cells in metaphase. f, Shannon diversity index by tumour type. g, DAPI-stained cells in metaphase of cell lines with histograms.

exceeded oncogene copy-number diversity if the oncogenes were located on other chromosomal loci (Extended Data Fig. B.7).

To determine whether extra- and intrachromosomal structures had a common origin, we developed AmpliconArchitect to elucidate the finer genomic structure using sequencing data (Methods). To better understand the relationship between subnuclear location and amplicon structure, we took advantage of a spontaneously occurring subclone of GBM39 cells in which a high copy EGFR mutant, EGFRvIII (an EGFR mutant with exons 27 deleted), shifted from the ecDNA exclusively to HSRs. Independent replicates of GBM39 containing an ecDNA amplicon, showed a consistent circular structure of 1.29Mb containing one copy of EGFRvIII (Extended Data Fig. B.8). Notably, the GBM39 subclone containing EGFRvIII exclusively on HSRs had an identical structure with tandem duplications containing multiple copies of EGFRvIII, indicating that the HSRs arose from reintegration of the EGFRvIII-containing ecDNA elements [CDG$^+$88] (Extended Data Fig. B.8). In GBM39 cells, resistance to EGFR tyrosine kinase inhibitors is caused by reversible loss of EGFRvIII from ecDNA [NGM$^+$14]. Structural analysis revealed a conservation of the fine structure of the EGFRvIII amplicon containing ecDNA in naive cells, during treatment and upon regrowth after discontinuation of therapy (Extended Data Fig. B.9), indicating that ecDNA can dynamically relocate to chromosomal HSRs while maintaining key structural features [CDG$^+$88, SLG$^+$10].

We next investigated whether ecDNA localization conferred a particular benefit, relative to chromosomal amplification. We hypothesized that ecDNA amplification may enable an oncogene to rapidly reach higher copy number because of the unequal segregation to daughter cells [WDY$^+$91] than would be possible by intrachromosomal amplification. We used a simplified GaltonWatson branching process to model the evolution of a tumour [BAO$^+$10], where each cell in the current

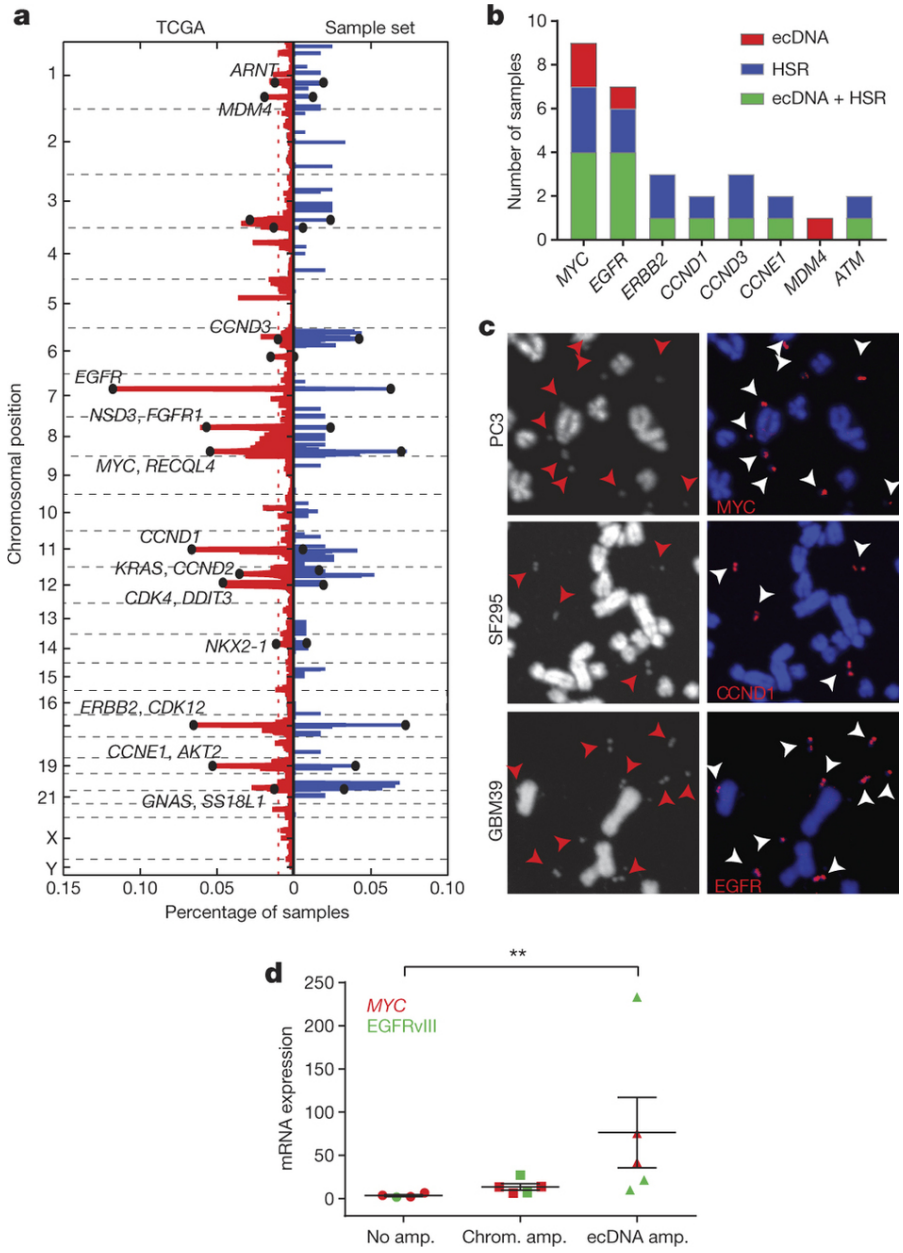


Figure 4.3: The most common focal amplifications in cancer are contained on ecDNA: a, Comparison of the frequency of focal amplifications detected by next generation sequencing of the 117 cancer samples studied here (blue) with those of matched tumour types in the TCGA (red) demonstrates significant overlap and representative sampling (P-value 10^{-6} based upon random permutations of TCGA amplicons; Methods). b, Localization of oncogenes by FISH. c, Representative FISH images of focal amplifications on ecDNA (arrows). d, EGFRvIII and MYC mRNA level, measured by qPCR ($P < 0.001$, MannWhitney U-test). Data are mean$\pm s.e.m.$; $n = 17$; each data point represents an average qPCR value of three technical replicates.

generation either replicates or dies to create the next generation. A cell with k copies of the amplicon is selected for replication with probability b_k as defined by $\frac{b_k}{(1-b_k)} = 1 + sf_m(k)$. We provided a positive selection bias towards cells with higher ecDNA counts by choosing s in the range of 0.5 to 1, and different selection regimes for f. Specifically, $f_m(k)$ increases to a maximum value $f_m(15) = 1$, then declines in a logistic manner with $f_m(m) = 0.5$ to reflect metabolic constraints (Methods). We allowed the amplicon copy number to grow to $1,000$ copies (Extended Data Fig. B.10), but set $b_k = 0$ for $k \geq 10^3$. During cell division, the $2k$ copies resulting from the replication of each of the k ecDNA copies segregate independently into the two daughter cells. We contrasted this with an intrachromosomal model of duplication with identical selection constraints, but with the change in copy number affected by mitotic recombination, and achieved by increasing or decreasing k by 1, with duplication probability P_d. A range of values for P_d, $(0.01 \leq P_d \leq 0.1)$ was used, where the upper boundary reflects a change in copy number once every five divisions. . Starting with an initial population of 10^5 cells, with $s = 0.5$, $m = 100$ and a selection function $f_{100}(k)$ (Fig. 4.4a), we find that an oncogene can reach a much higher copy number in a tumour if it is amplified on ecDNA, rather than on a chromosome (Fig. 4.4b). As predicted by the model, we detected a significantly higher copy number of the most frequently amplified oncogenes EGFR (including EGFRvIII) and MYC, when they were contained within ecDNA instead of within chromosomes (Fig. 4.4c). We also reasoned that if an oncogene is amplified intrachromosomally, the heterogeneity of the tumour (in terms of the distribution of copies of the oncogene) would stabilize at a much lower level. By contrast, unequal segregation of ecDNA would probably rapidly enhance heterogeneity and maintain it. Our model consistently confirmed this prediction (Fig. 4.4d) for a wide range of simulation parameters . The heterogeneity of copy-number change

stabilizes and even decreases over time [AGJ+16, LGP+14], much as predicted in Fig. 4.4d. We also tested the validity of the model by comparing the Shannon diversity index against the average number of amplicons per cell in our tumour samples. Heterogeneity of a tumour with respect to oncogene copy number would be more likely to rise relatively slowly if it is present on a chromosome, but would rise more rapidly and be maintained much longer, if that oncogene is present on ecDNA, as confirmed by a plot of Shannon diversity index versus copy number (Fig. 4.4e). Moreover, the predicted correlation in Fig. 4.4e is completely recapitulated by the experimental data (Fig. 4.4f), thereby validating the central tenets of the model.

There is growing evidence that genetically heterogeneous tumours are remarkably difficult to treat [AGJ+16]. The data presented here identifies a mechanism by which tumours maintain cell-to-cell variability in the copy number and transcriptional level of oncogenes that drive tumour progression and drug resistance. We suggest that extrachromosomal oncogene amplification may enable tumours to adapt more effectively to variable environmental conditions by increasing the likelihood that a subpopulation of cells will express that oncogene at a level that maximizes tumour proliferation and survival [VHNVY+88,NGM+14,MW16,SKAK78,NSG+14, BSH63], rendering tumours progressively more aggressive and difficult to treat over time. Even when using a selection function that only mildly depends on copy number, we detected a very large difference between intra- and extrachromosomal amplification mechanisms leading to a higher copy number of amplicons and greater heterogeneity in copy number. Thus, even small increases in selection advantage conferred by oncogenes amplified on ecDNA would be expected to yield a very high fitness advantage . The notably high frequency of ecDNA in cancer, as shown here, coupled to the benefits to tumours of extrachromosomal gene amplification

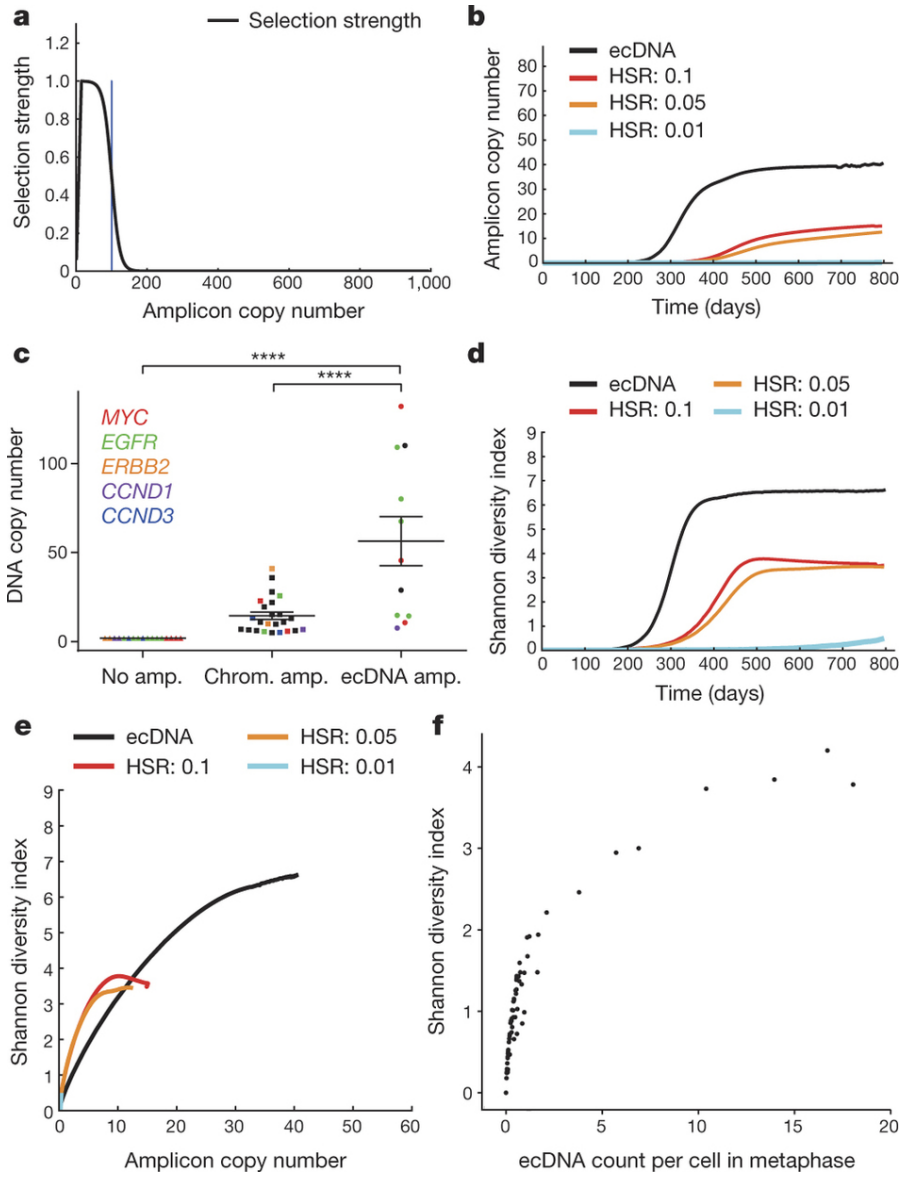


Figure 4.4: Theoretical model for focal amplification via extrachromosomal and intrachromosomal mechanisms: Simulated change in copy number via random segregation (ecDNA) or mitotic recombination (HSR), starting with 10^5 cells, 100 of which carry amplifications. a, The selection function $f_{100}(k)$ reaches a maximum for $k = 15$, then decays logistically. b, Growth in amplicon copy number over time. c, DNA copy number stratified by oncogene location. ($P < 0.001$, ANOVA/Tukeys multiple comparison). $n = 52$; data points include top five amplified oncogenes, mean$\pm s.e.m.d$, Change in heterogeneity (Shannon diversity index) over time. e, Correlation between copy number and heterogeneity. f, Experimental data showing correlation between ecDNA counts and heterogeneity matches the simulation in e.

relative to chromosomal inheritance, suggest that oncogene amplification on ecDNA may be a driving force in tumour evolution and the development of genetic heterogeneity in human cancer. Understanding the underlying molecular mechanisms of tumour evolution, including oncogene amplification in ecDNA, may help to identify more effective treatments that either prevent cancer progression or more effectively eradicate tumours.

4.3 Methods

4.3.1 Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

4.3.2 Cytogenetics

Metaphase cells were obtained by treating cells with Karyomax (Gibco) at a final concentration of 0.01μg ml^1 for 1-3 h. Cells were collected, washed in PBS, and resuspended in 0.075M KCl for 15-30 min. Carnoys fixative (3 : 1 methanol:glacial acetic acid) was added dropwise to stop the reaction. Cells were washed an additional three times with Carnoys fixative, before being dropped onto humidified glass sides for metaphase cell preparations. For ECdetect analyses, DAPI was added to the slides. Images in the main figures were captured with an Olympus FV1000 confocal microscope. All other images were captured at a magnification of $1,000X$ with an Olympus BX43 microscope equipped with a QiClick cooled camera. FISH was performed by adding the appropriate DNA FISH probe onto the fixed metaphase spreads. A coverslip was added and sealed with

rubber cement. DNA denaturation was carried out at 75°C for 3-5min and the slides were allowed to hybridize overnight at 37°C in a humidified chamber. Slides were subsequently washed in $0.4X$ SSC at 50°C for 2min, followed by a final wash in $2X$ SSC containing 0.05% Tween-20. Metaphase cells and interphase nuclei were counterstained with DAPI, a coverslip was applied and images were captured.

4.3.3 Cell culture

The NCI-60 cell line panel (gift from A. Shiau, obtained from NCI) was grown in RPMI-1640 with 10% FBS under standard culture conditions. Cell lines were not authenticated, as they were obtained from the NCI. The PDX cell lines were cultured in DMEM/F-12 medium supplemented with glutamax, B27, EGF, FGF and heparin. Lymphoblastoid cells (gift from B. Ren) were grown in RPMI-1640, supplemented with 2mM glutamine and 15% FBS. IMR90 and ALS6-Kin4 (gift from J. Ravits and D. Cleveland) cells were grown in DMEM/F-12 supplemented with 20% FBS. Normal human astrocytes (NHA) and normal human dermal fibroblasts (NHDF) were obtained from Lonza and cultured according to Lonza-specific recommendations. Cell lines were not tested for mycoplasma contamination.

4.3.4 Tissue samples

Tissues were obtained from the Moores Cancer Center Biorepository Tissue Shared Resource with IRB approval (#090401). All samples were de-identified and patient consent was obtained. Additional tissue samples that were obtained were approved by the UCSD IRB (#120920).

4.3.5 DNA library preparation

DNA was sonicated to produce 300-500bp fragments. DNA end repair was performed using End-it (Epicentre), DNA library adapters (Illumina) were ligated and the DNA libraries were amplified. Paired-end next-generation sequencing was performed and samples were run on the Illumina Hi-Seq using 100 cycles.

4.3.6 DNA extraction

Cells were collected and washed with $1X$cold PBS. Cell pellets were resuspended in buffer 1 (50mM Tris pH 7.5, 10mM EDTA, 50μg ml^1 RNase A), and incubated in buffer 2 (1.2% SDS) for 5min on ice. DNA was acidified by the addition of buffer 3 (3M CsCl, 1M potassium acetate, 0.67M acetic acid) and incubated for 15min on ice. Samples were centrifuged at $14,000$g for 15min at 4°C. The supernatant was added to a Qiagen column and briefly centrifuged. The column was washed (60% ethanol, 10mM Tris pH 7.5, 50μM EDTA, 80mM potassium acetate) and eluted in water.

4.3.7 DNase treatment

Metaphase cells were dropped onto slides and visualized with DAPI. Coverslips were removed and slides washed in $2X$SSC, and subsequently treated with 2.5% trypsin, and incubated at 25°C for 3min. Slides were then washed in $2X$SSC, DNase solution (1mg ml^1) was applied to the slide and cells were incubated at 37°C for 3h. Slides were washed in $2X$SSC and DAPI was again applied to the slide to visualize DNA.

4.3.8 ecDNA count statistics

In Fig. 4.2a, b the violin plots represent the distribution of ecDNA counts in different sample types. In order to compare the ecDNA counts between the different samples, we use a one-sided Wilcoxon rank-sum test, where the null hypothesis assumes that the mean ecDNA-count ranks of the compared sample types are equal.

4.3.9 Estimation of frequency of samples containing ecDNA

There is a wide variation in the number of ecDNA across different samples and within metaphases of the same sample. We want to estimate and compare the frequency of samples containing ecDNA for each sample type. We label a sample as being ecDNA positive by using the pathology standard: a sample is deemed to be ecDNA positive if we observe$\geq$ 2 ecDNA in$\geq$ 2 out of 20 metaphase images. Therefore, we ensure that every sample contains at least 20 metaphases.

We define indicator variable $X_{ij} = 1$ if metaphase image j in sample i has $\geq$ 2 ECDNA; $X_{ij} = 0$ otherwise. Let n_i be the number of metaphase images acquired from sample i. We assume that X_{ij} is the outcome of the j-th Bernoulli trial, where the probability of success p_i is drawn at random from a beta distribution with parameters determined by $\sum_j X_{ij}$. Formally,

$$p_i|\alpha_i, \beta_i \sim \text{Beta}(\alpha_i = \max\{\epsilon, \sum_j X_{ij}\}, \beta_i = \max\{\epsilon, n_i - \alpha_i\}). \tag{4.1}$$

We model the likelihood of observing k successes in $n = 20$ trials using the binomial density function as:

$$k|p_i \sim \text{Binom}(p_i, n = 20) \tag{4.2}$$

Finally, the *predictive* distribution $p(k)$, is computed using the product of the Binomial likelihood and Beta prior, modeled as a "beta-binomial distribution" [Lee12].

$$\begin{aligned} p(k) = \mathbb{E}_{p_i}[k|p_i] &= \int_0^1 k|p_i \cdot p_i|\alpha_i, \beta_i \, \mathrm{d}p_i \\ &= \int_0^1 \binom{n}{k} p_i^k (1-p_i)^{n-k} \cdot \frac{1}{\mathrm{B}(\alpha_i, \beta_i)} p_i^{\alpha_i - 1} (1-p_i)^{\beta_i - 1} \, \mathrm{d}p_i \\ &= \binom{n}{k} \frac{1}{\mathrm{B}(\alpha_i, \beta_i)} \int_0^1 p_i^{k+\alpha_i-1} (1-p_i)^{n-k+\beta_i-1} \, \mathrm{d}p_i \\ &= \binom{n}{k} \frac{\mathrm{B}(k+\alpha_i, n-k+\beta_i)}{\mathrm{B}(\alpha_i, \beta_i)} \end{aligned} \tag{4.3}$$

We model the probability for sample i being EC-positive with the random variable Y_i such that:

$$\begin{aligned} Y_i &= 1 - Pr(\text{sample } i \text{ is EC-negative}) \\ &= 1 - (k=1|p_i) - (k=0|p_i) \end{aligned} \tag{4.4}$$

The expected value of Y_i is:

$$\begin{aligned} \mathbb{E}_{p_i}(Y_i) &= 1 - p(k=1) - p(k=0) \\ &= 1 - \binom{20}{1} \frac{\mathrm{B}(1+\alpha_i, 19+\beta_i)}{\mathrm{B}(\alpha_i, \beta_i)} - \binom{20}{0} \frac{\mathrm{B}(\alpha_i, 20+\beta_i)}{\mathrm{B}(\alpha_i, \beta_i)} \end{aligned} \tag{4.5}$$

The variance of Y_i is:

$$\mathrm{Var}(Y_i) = \mathrm{Var}(k=1|p_i) + \mathrm{Var}(k=0|p_i) + 2\mathrm{Cov}(k=1|p_i, k=0|p_i), \tag{4.6}$$

where,

$$\begin{aligned}
\mathrm{Var}(k|p_i) &= \mathbb{E}_{p_i}[(k|p_i)^2] - \mathbb{E}_{p_i}[k|p_i]^2 && (4.7)\\
&= \int_0^1 (k|p_i)^2 \cdot p_i|\alpha_i, \beta_i \,\mathrm{d}p_i - (\int_0^1 k|p_i \cdot p_i|\alpha_i, \beta_i)^2 \,\mathrm{d}p_i \\
&= \binom{n}{k}\binom{n}{k}\frac{1}{\mathrm{B}(\alpha_i, \beta_i)}\int_0^1 p_i^{2k+\alpha_i-1}(1-p_i)^{2n-2k+\beta_i-1}\,\mathrm{d}p_i && (4.8)\\
&\qquad - \binom{n}{k}\binom{n}{k}\frac{\mathrm{B}(k+\alpha_i, n-k+\beta_i)^2}{\mathrm{B}(\alpha_i, \beta_i)^2} \\
&= \binom{n}{k}\binom{n}{k}\frac{1}{\mathrm{B}(\alpha_i, \beta_i)}[\mathrm{B}(2k+\alpha_i, 2n-2k+\beta_i) - \frac{\mathrm{B}(k+\alpha_i, n-k+\beta_i)^2}{\mathrm{B}(\alpha_i, \beta_i)}],
\end{aligned}$$

and

$$\mathrm{Cov}(k=1|p_i, k=0|p_i) = \mathbb{E}_{p_i}[k=1|p_i \cdot k=0|p_i] - \mathbb{E}_{p_i}[k=0|p_i]\,\mathbb{E}_{p_i}[k=1|p_i] \tag{4.9}$$

$$\begin{aligned}
&= \binom{n}{0}\binom{n}{1}\frac{1}{\mathrm{B}(\alpha_i, \beta_i)}[\int_0^1 p_i^{1+\alpha_i-1}(1-p_i)^{2n-1+\beta_i-1}\,\mathrm{d}p_i && (4.10)\\
&\qquad - \frac{\mathrm{B}(\alpha_i, n+\beta_i)\mathrm{B}(1+\alpha_i, n-1+\beta_i)}{\mathrm{B}(\alpha_i, \beta_i)} \\
&= \binom{n}{0}\binom{n}{1}\frac{1}{\mathrm{B}(\alpha_i, \beta_i)}[\mathrm{B}(1+\alpha_i, 2n-1+\beta_i) && (4.11)\\
&\qquad - \frac{\mathrm{B}(\alpha_i, n+\beta_i)\mathrm{B}(1+\alpha_i, n-1+\beta_i)}{\mathrm{B}(\alpha_i, \beta_i)}].
\end{aligned}$$

Let T be the set of samples belonging to a certain sample type t, e.g. immortalized samples. We define

$$Y_T = \frac{\sum_{i \in T} Y_i}{|T|} \tag{4.12}$$

We estimate the frequency of samples under sample t containing ECDNA (bar

heights on Figures 2C and 2D) as

$$\mathbb{E}[Y_T] = \frac{\sum_{i \in T} \mathbb{E}[Y_i]}{|T|} \tag{4.13}$$

and error bar heights (Figure 2C and 2D) as:

$$\mathrm{sd}(Y_T) = \frac{(\sum_{i \in T} \mathrm{Var}[Y_i])^{\frac{1}{2}}}{|T|} \tag{4.14}$$

assuming independence among samples $i \in T$. For any α_i or $\beta_i = 0$, we assign them a sufficiently small ϵ.

4.3.10 Comparison of ecDNA presence between different sample types

We construct binary ecDNA-presence distributions, based on the ecDNA counts, such that an image with≥ 2 ecDNA is represented as a 1, and 0 otherwise. In order to compare the ecDNA presence between the different samples, we use a one-sided Wilcoxon rank-sum test using the binary ecDNA-presence distributions, where the null hypothesis assumes the mean ranks of the compared sample types are equal.

4.3.11 ECdetect: software for detection of extrachromosomal DNA from DAPI staining metaphase images

The software applies an initial coarse adaptive thresholding [Mot15, BR07] on the DAPI images to detect the major components in the image with a window size of $150X150$ pixels, and $T = 10\%$. Components over $3,000$ pixels and 80% of solidity are masked, and small components discarded. Weakly connected components of the

remaining binary image are computed to find the separate chromosomal regions. Connected components over a cumulative pixel count of $5,000$ are considered as candidate search regions, and their convex hull with a dilation of 100 pixels are added into the ecDNA search region. Following the manual masking and verification of the ecDNA search region, a second finer adaptive thresholding with a window size of $20X20$ pixels and $T = 7\%$ is performed. Components that are greater than 75 pixels are designated as non-ecDNA structures and their 15-pixel neighbourhood is removed from the ecDNA search region. Any component detected with a size less than or equal to 75 pixels and greater than or equal to 3 pixels inside the search region is detected as ecDNA. For more detail, please see Appendix C.

4.3.12 Bioinformatic datasets

We sequenced 117 tumour samples including 63 cell lines, 19 neurospheres and 35 cancer tissues with coverage ranging from $0.6X$ to $3.89X$ and an additional 8 normal tissues as controls. See Extended Data Fig. B.4 for the coverage distribution across samples. We mapped the sequencing reads from each sample to the hg19 (GRCh37) human reference genome [LLB$^+$01] from the UCSC genome browser [KSF$^+$02] using BWA software version 0.7.9a (ref. [LD09]). We inferred an initial set of copy-number variants (CNVs) from these mapped sequence samples using the ReadDepth CNV software [MHCM11] version 0.9.8.4 with parameters FDR=0.05 and overDispersion=1.

We downloaded CNV calls for $11,079$ paired tumournormal samples covering 33 different tumour types from TCGA. We applied similar filtering criteria to ReadDepth output and TCGA calls to eliminate false copy number amplification calls from repetitive genomic regions and hotspots for mapping artefacts.

We used the filtered set of CNV calls from ReadDepth as input probes

for AmpliconArchitect which revealed the final set of amplified intervals and the architectures of the amplicons.

4.3.13 Reconstruction using AmpliconArchitect

We developed a novel tool AmpliconArchitect, to automatically identify connected amplified genomic regions and reconstruct plausible amplicon architectures. For each sample, AmpliconArchitect takes as input an initial list of amplified intervals and whole-genome sequencing paired-end reads aligned to the human reference. It implements the following steps to reconstruct the one or more architectures for each amplicon present in the sample: (1) use discordant read-pair alignments and coverage information to iteratively visit and extend connected genomic regions with high copy numbers; (2) for each set of connected amplified regions, segment the regions based on depth of coverage using a mean-shift segmentation to detect copy-number changes and discordant read-pair clusters to identify genomic breaks; (3) construct a breakpoint graph connecting segments using discordant read-pair clusters; (4) compute a maximum-likelihood network to estimate copy counts of genomic segments; and (5) report paths and cycles in the graph that identify the dominant linear and circular structures of the amplicon .

4.3.14 Comparison of CNV gains between the sequencing sample set and TCGA

We compared our sample set against TCGA samples to test the assumption that the genomic intervals amplified in our sample set are broadly representative of a pan-cancer dataset, by comparing against TCGA samples. Here, we deal with an abstract notation to represent different datasets and describe a generic procedure

to compare amplified regions. Consider a set of K samples. For any $k \in [1, ..., K]$, let S_k denote the set of amplified intervals in sample k.

Let c be the cancer subtype for sample k. We compare S_k against TCGA samples with subtype c. Let T denote the set of all genomic regions which are amplified in at least 1% of TCGA samples of subtype c. For each interval $t \in T$, let f_t denote its frequency in TCGA samples of subtype c. We define a match score

$$d_k = \sum_{t \in S_k, T} f_t = \{t \in T, s.t. toverlapsanintervalinS_k\} \tag{4.15}$$

The cumulative match score for all samples is defined as:

$$D = \sum_{t \leq k \leq K} d_k \tag{4.16}$$

To compute the significance of statistic D, we do a permutation test. We generate N random permutations of the TCGA intervals for subtype c and estimate the distribution of match scores of our sample set against the random permutations. We choose a random assignment of locations of all intervals in T, while retaining their frequencies. For the jth permuted set T_j, we computed the cumulative match score D_j relative to our sample set. Thus the significance of overlap between amplified intervals in our sample set and the TCGA set is estimated by the fraction of random permutations with D_j/gtD. Computing 1 million random permutations generated exactly one permutation breaching the TCGA score D, implying a $P \leq 10^6$.

4.3.15 Oncogene enrichment

We compared the rank correlation of the most frequent oncogenes in our sample set with the top oncogenes as reported by TCGA pan-cancer analysis in

ref. [ZSC+13]. We identified 14 oncogenes occurring in 2 or more samples of our sample set and compared these to the top 10 oncogenes from the TCGA pan-cancer analysis. We found that 7 out of the top 10 oncogenes were represented in our list of 14 oncogenes. Considering 490 oncogenes in the COSMIC database, the significance of observing 7 or more oncogenes in common in the two datasets is given by the hypergeometric probability

$$P = \sum_{i=7}^{10} \frac{\binom{480}{14-i}\binom{10}{i}}{\binom{490}{14}} = 3.07 \cdot 10^{-10} \tag{4.17}$$

4.3.16 Amplicon structure similarity

We found high similarity between amplicon structures of biological replicates (for example, Extended Data Fig. B.8). We estimate the probability of common origin between two samples by measuring the pairwise similarity between amplicon structures. In reconstructing the structures, we identify a set of locations representing change in copy number and we use the locations of change in copy number to estimate the similarity in amplicon structures.

Let L be the total length of amplified intervals. These intervals are binned into windows of size r, resulting in $N_b = \frac{L}{r}$ bins. We use a segmentation algorithm that determines if there is a change in copy number in any bin, within a resolution of $r = 10,000$bp. Note that this is an overestimate, because with split-reads and high-density sequencing data, we can often get the resolution down to a few base pairs. Let S_1 and S_2 represent the set of bins with copy-number changes in the two samples, respectively. S_1 and S_2 are selected from a candidate set of locations N_b. Under the null hypothesis that S_2 is random with respect to S_1, we expect $I = S_1 \cap S_2$ to be small. Let $m = min(|S_1|, |S_2|)$, and $M = max(|S1|, |S2|)$. A P-value is computed as follows:

$$P = \sum_{i=|I|}^{m} \frac{\binom{N_b - m}{M-i}\binom{m}{i}}{\binom{N_b}{M}} \tag{4.18}$$

4.3.17 A branching process model for oncogene amplification

Consider an initial population of N_0 cells, of which N_a cells contain a single extra copy of an oncogene. We model the population using a discrete generation Galton-Watson branching process [BAO+10]. In this simplified model, each cell in the current generation containing k amplicons (amplifying an oncogene) either dies with probability d_k, or replicates with probability b_k to create the next generation. We set the selective advantage

$$\frac{b_k}{d_k} = \begin{cases} 1 + sf_m(k), & 0 \leq k < M_a \\ 0 & \text{otherwise} \end{cases} \tag{4.19}$$

$$d_k = 1 - b_k \tag{4.20}$$

In other words, cells with k copies of the amplicon stop dividing after reaching a limit of M_a amplicons. Otherwise, they have a selective advantage for $0 < k \leq M_a$, where the strength of selection is described by $f_m(k)$, as follows:

$$f_m(k) = \begin{cases} \frac{k}{M_s} & (0 \leq k \leq M_s)\ , \\ \frac{1}{1+e^{-\alpha(k-m)}} & (M_s < k < M_a). \end{cases} \tag{4.21}$$

Here, s denotes the selection-coefficient, and parameters m and α are the 'mid-point', and 'steepness' parameters of the logistic function, respectively. Initially, $f_m(k)$ grows linearly, reaching a peak value of $f_m(k) = 1$ for $k = M_s$. As the viability of cells with large number of amplicons is limited by available nutrition [PT16],

$f_m(k)$ decreases logistically in value for $k > M_s$ reaching $f_m(k) \to 0$ for $k \geq M_a$. We model the decrease by a sigmoid function with a single mid-point parameter m s.t. $f_m(m) = \frac{1}{2}$. The 'steepness' parameter α is automatically adjusted to ensure that $\max\{1 - f_m(M_s), f_m(M_a)\} \to 0$.

The copy number change is effected by different mechanisms for extrachromosomal (EC) and intrachromosomal (HSR) models. In the EC model, the available k amplicons are on EC elements which replicate and segregate independently. We assume complete replication of EC elements so that there are $2k$ copies which are partitioned into the two daughter cells via independent segregation. Formally, the daughter cells end up with k_1 and k_2 amplicons respectively, where

$$k_1 \sim \mathcal{B}(2k, \frac{1}{2}) \tag{4.22}$$

$$k_2 = 2k - k_1 \tag{4.23}$$

In contrast, in the intrachromosomal model, the change in copy number happens via mitotic recombination, and the daughter cell of a cell with k amplicons will acquire either $k + 1$ amplicons or $k - 1$ amplicons, each with probability p_d. With probability $1 - 2p_d$, the daughter cell retains k amplicons.

4.3.18 Data availability

Whole-genome sequencing data are deposited in the NCBI Sequence Read Archive (SRA) under Bioproject (accession number: PRJNA338012). DAPI and FISH metaphase images are available for download on figshare at https://figshare.com/s/ab6a214738aa43833391.

4.3.19 Acknowledgements

We thank R. Kolodner, W. Mischel, D. Geschwind, members of the Mischel laboratory, A. Akbari, A. Iranmehr and A. Patel for helpful comments. This work was supported by the Ludwig Institute for Cancer Research (P.S.M., B.R., K.A., W.K.C., F.B.F.), Defeat GBM Program of the National Brain Tumor Society (P.S.M., F.B.F.), The Ben and Catherine Ivy Foundation (P.S.M.), generous donations from the Ziering Family Foundation in memory of Sigi Ziering (P.S.M.); The Susan G. Komen Foundation (SAC110036), The Leona M. and Harry B. Helmsley Charitable Trust (2012-PG-MED002) and The Breast Cancer Research Foundation (BCRF) to G.M.W.; CureSearch for Childrens Cancer and a Leadership Award from the California Institute for Regenerative Medicine to R.W.R. This work was also supported by the following NIH grants: NS73831 (P.S.M.), GM114362 (V.B., V.D., D.B.), NS80939 (F.B.F.), CA014195 and CA159859 (G.M.W.) and CA151819 (D.A.N.) and T32CA121938 (K.M.T.) and NSF grants: NSF-IIS-1318386 and NSF-DBI-1458557 (V.B., V.D., D.B.).

Chapter 4, in part, is a reformatted reprint of the material as it appears in: "Kristen M. Turner, Viraj Deshpande, Doruk Beyter, Tomoyuki Koga, Jessica Rusert, Catherine Lee, Bin Li, Karen Arden, Bing Ren, David A. Nathanson, Harley I. Kornblum, Michael D. Taylor, Sharmeela Kaushal, Webster K. Cavenee, Robert Wechsler-Reya, Frank Furnari, Scott R. Vandenberg, P. Nagesh Rao, Geoffrey M. Wahl, Vineet Bafna, Paul S. Mischel. Extrachromosomal oncogene amplification drives tumor evolution and the development of genetic heterogeneity in human cancer. Nature, 543(7643), 122-125, 2017.". The dissertation author was a joint primary investigator and author of this material.

Appendix A

Supplementary Material for Chapter 2

A.1 Supplementary Methods

A.1.1 DNA preparation

Each 50 ml biological sample was thawed, homogenized, and two 15ml subsamples withdrawn from the original sample and placed in 15ml tubes. These were centrifuged at 3500rpm for 20 minutes. The supernatant from each sample was combined and transferred to a 50mL tube. This was then concentrated using Amicon Ultra Centrifugal Filters (EMD Milipore, 2015). 15 mL of supernatant was added to Amicon Ultra Centrifugal Filters. These were centrifuged at max (3750rpm) for 1 hour. The liquid was disposed. The remaining supernatant was added to the filter which was again centrifuged at max (3750rpm) for 1 hour. 200μL from the top of the filter was transferred into a new centrifuge tube and stored. This liquid was then added to the pellet from the original centrifuge and DNA extracted using the PowerLyser PowerSoil DNA isolation Kit (Mo Bio Laboratories

Inc., 2015).

The DNA from the extraction was amplified using primers designed to target both the V4 region of 16S rRNA gene and the ITS2 region of prokaryotic and eukaryotic genomes. Primers were ordered with with 5' PHO modifications to ensure compatibility with labeling for the sequencing steps. The amplicon for the 16S should fall approximately between the 100-400bp range and the primers were designed to universally target Archea and Bacteria (Forward: S-D-Bact-0564-a-S-15 (41345) AYTGGGYDTAAAGNG, Reverse: S-D-Bact-0785-b-A-18 (41346) TACNVGGGTATCTAATCC). The amplicon for the ITS2 primer should fall approximately between 200-400bp and were selected because they universally target eukaryotes (Forward: (41343) GCATCGATGAAGAACGCAGC, Reverse: (41344) TCCTCCGCTTATTGATATGC).

The PCR was set up in a 96 well plate as follows: 20.0μL 5X HF buffer (Phusion kit), 4.0μL 10 mM dNTPs (NEB), 4.0μL DMSO (Phusion kit), 10.0μL 5M Betaine, 5.0μL 10μM of each primer, 0.8μL Phusion polymerase, 6.0μL DNA template. To cover the diversity represented gradient PCR was performed with the following PCR protocol: 98°C 0:30, 25X (98°C 0:10, 43°C-53°C 0:30, 72°C 0:30), 72°C 5:00, 4°C hold. Gels were run to ensure correct band sizes. The DNA was then pooled and cleaned using Invitrogen PureLink Pro 96 PCR purification Kit (Life Technologies, 2015). The resultant DNA was then quantified to ensure 2 micrograms and prepped for sequencing.

A.1.2 TMAP usage

We applied the "map2" algorithm (based off of the BWA long-read algorithm [LD10]), designed for reads longer than 150bps, due to the read sizes (a mean of 240bps for 16S and 420 for ITS2 sequences – see Figures SA.4, SA.5, and SA.6 for

read length distributions in all chips and samples; individually, and all combined) and other default parameters associated with it. For every read, TMAP returns the mapping with the best score. If multiple sequences had the same best score, a random mapping among them was returned.

A.1.3 OTU-based analysis for 16S data

Several OTU-based pipelines such as UPARSE [Edg13], QIIME [CKS+10], MOTHUR [SWR+09] have been developed for the analysis of Illumina or 454 pyrosequencing 16S and fungal only ITS2 marker-gene sequencing data. Very recently, a pipeline that includes 16S Ion Torrent PGM sequencing is developed [PRM+14], and used it in the Brazilian Microbiome Project (BMP) [Pyl15]. The BMP 16S profiling analysis pipeline makes use of the UPARSE OTU clustering, and QIIME taxonomy assignment, using Ribosomal Database Project (RDP) naive classifier [WGTC07].

In order to compare our 16S data analysis results with OTU-based pipelines, we used the pipeline suggested by BMP. We began by truncating the reads at length 200 as the read ends are assumed to have lowered quality, and discarded any read with a smaller length. We then removed any read having an expected error rate of 1.0, a suggested value in the UPARSE documentation [Edg15b]. We applied dereplication that removes the identical reads for faster querying, and removed any singleton reads. We clustered the OTUs, and applied a reference based chimera filtering using a gold database, which contains the ChimeraSlayer reference database from the Broad Microbiome Utilities version microbiomeutil-r20110519, as described in [Edg15a], using the plus strand, as specified. We finally assigned all quality filtered reads, including the singletons, to the constructed OTUs at 97% identity. All analysis until this point was performed using usearch v7.0.1090_i86linux32. We

gathered the taxonomy information using assign_taxonomy.py version 1.7.0 from QIIME, choosing RDP classifier as taxonomy assignment algorithm with the default bootstrap confidence threshold of 80%, and OTUs pre-constructed from GreenGenes (version May 2013) at 97% identity, as training sequences.

A.1.4 Comparison of sequence mapping and OTU-based approaches and reproducibility assessment among chips

We performed a Mantel test between the sample taxonomy composition results of our approach and the BMP pipeline for 16S data analysis as follows: at ranks phylum, class, order, family and genus, respectively we obtained the taxonomies of both analysis results. We took the union of the taxonomies observed in the two analyses, and assigned abundance values of 0 to any taxonomy in the union set not observed in individual results, for all 26 time point samples. Thus, for each approach, we had pairs of relative abundance values for all taxonomies in the union set at all time points as a matrix, which we called *a taxonomy abundance matrix*, for each of the aforementioned rank. We compared these pairs of taxonomy abundance matrices using the package "ade4" [DD07] in R with the function "mantel.rtest" using 999 replicates. We achieved Mantel r statistics of 0.99, 0.98, 0.94, 0.94, 0.91 for ranks phylum, class, order, family, and genus, respectively, all with p-value 0.001, suggesting high result similarity. Since the RDP classifier is not capable in classification beyond the genus level, we have no comparison available with the BMP pipeline at species/sequence level of resolution. BMP pipeline area plots at ranks phylum, class, and genus are shown in Figure SA.14, for visual comparison purposes.

We also note that a 16S genus level diversity comparison between the two approaches yield a nearly identical pattern: the linear regression describing the relationship between the two was: $r^2 = 0.96$, $P = 2.60 \cdot 10^{-14}$.

The reproducibility assessment among chips for 16S and ITS2 data also follows the same Mantel test approach, with the single difference of containing the top 2000 and 200 sequence relative abundances (instead of taxa relative abundances) in the compared pairs of abundance matrices coming from different chips.

A.1.5 Challenges in OTU-based approaches and taxonomy assignment on ITS2 data

Given the high variance in the ITS2 region length, ranging from 100bps to 700bps [YSL+10]; length trimming, a critically important step in an OTU-based approach [Edg15b], is not practical. Moreover, the taxon dependent OTU clustering identity percentages on microbial eukaryotes [GSMK14], may render the OTU clustering step erroneous. The taxon dependency of OTU clustering identity percentages also makes the RDP naive Bayesian classifier taxonomy assignment (used in OTU-based approach) challenging, as its reference taxonomy database is expected to be clustered at a certain identity percentage. Another challenge in contructing a clustered ITS2 database from NCBI would lie in determining the correct boundaries of the ITS2 region, previous to clustering, due to the flanking 18S, ITS1, 5.8S, and 28S regions in the NCBI nucleotide entries. Previous research [PALKX14] reports that taxonomy classification results using BLASTN, a mapping based approach, and RDP naive Bayesian classifier are very similar on ITS2 data. Considering these challenges and findings, we preferred to determine the taxa relative abundances using a mapping approach.

A.1.6 Outlier removal on time series ecosystem data

We initially subtracted the 7-day local central mean from each data point. We performed this step in order to reduce the dependency between successive points in our time series ecosystem data and to satisfy the idenpendent, identicaly distribution requirement for a normal distribution. We, then, tested for normality using "shapiro.test" in R, using the package "stats" [R C14]. Upon confirming for normality, we removed any data point that exceeded 3σ of distance from mean. We did not perform outlier detection for NH_4, urea, NO_3, NO_2, and PO_4, due to the expected high fluctuations stemming from pond nutrient management.

A.1.7 Model comparison using F-test

In order to explore the explanatory values of certain factors on a target, controlling for other factor(s), we compared two models: a reduced and a full model. The reduced model contains the factor we would like to control for, whereas the full model contains additional factor(s), which we are interested to explore the effect on our target.

$$\begin{aligned} \text{Reduced Model} \quad y =& \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon_r \\ \text{Full Model} \quad y =& \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} x_{k+1} + \cdots + \beta_p x_p + \varepsilon_f \end{aligned} \tag{A.1}$$

where in one our tests, for instance, y was chosen as the eukaryotic diversity we were targeting, $x_1, \ldots, x_k$ as the factors we controlled for such as temperature and bacteria diversity, and $x_{k+1}, \ldots, x_p$ as any factor(s) we explored the effect it had on the target, such as pre- and post-pesticide sampling. We tested if we could reject the null hypothesis:

$$H_0 : \beta_{k+1} = \cdots = \beta_p = 0$$

to see if our full model added a significant explanatory value over the reduced model, using an F statistic:

$$F = \frac{(RSS_{reduced} - RSS_{full})/(p-k)}{RSS_{full}/(n-p-1)} \tag{A.2}$$

where RSS_i is the residual sum of squares of model i.

A.2 Supplementary Results

A.2.1 Mapping statistics

We initially discarded any read having length shorter than 50 nucleotides, and an error rate higher than 2.0 for 16S reads, and 4.0 for ITS reads, due to their longer average size compared to 16S. After mapping the remaining 16S and ITS2 reads to respective databases, we calculated percent identity, and *query-coverage*, defined as the fraction of the query sequence matching to the target, for assessing mapping quality. For these measures, the quality was uniformly high with a mean percent identity of 97% and 96%, and mean coverage over 94% and 82% across all 16S and ITS2 reads that mapped their respective database. (Figures SA.7 and SA.8). Following the cutoffs applied by "16S Ribosomal RNA Reference Sequence Similarity Search" by NCBI [NCB15b], we used a 95% percent identity and 70% of query-coverage cutoff. On average among all chips, 75% of the 16S and 77% of the ITS2 reads exceeded our chosen cut-offs, and were used in subsequent analyses.

A.2.2 Intra-sample reproducibility assessment

In order to assess robustness in the sample composition analyses, two redundant samples were used as technical replicates for each of samples 4, 11, 19 and 24, in the design (samples 27 and 31 were replicates of sample 4, 28 and 32 for 11, 29 and 33 for 19, and 30 and 34 for 24). Figure SA.9 demonstrates that the technical replicates consistently show low dissimilarity values (mean Bray Curtis dissimilarity values of 0.06, 0.03, 0.04, 0.02 and 0.04, 0.07, 0.50, 0.06, for the two replicates of samples 4, 11, 19 and 24 for 16S and ITS2, chip 3.) suggesting good reproducibility, except sample 19 for ITS2 data only. We note the replicates for sample 19 (samples 29 and 33, ITS2 data) had a skewed read length distribution, compared to sample 19 itself, (see Figure SA.5b), which might be a possible reason for the observed noise.

A.2.3 Pre- and post-fungicide relationship of productivity variability and temperature

We investigated whether temperature, based on its pre-fungicide era relationship with productivity variability (standard deviation), could predict the post-fungicide productivity standard deviation (sd) trends. Figure SA.17 shows linear relationship between temperature and productivity sd in different periods. During the pre-fungicide period, temperature showed a positive correlation with productivity sd, whereas it had a negative correlation during the post-fungicide period, therefore temperature alone cannot explain the change in the productivity variability observed after the fungicide application.

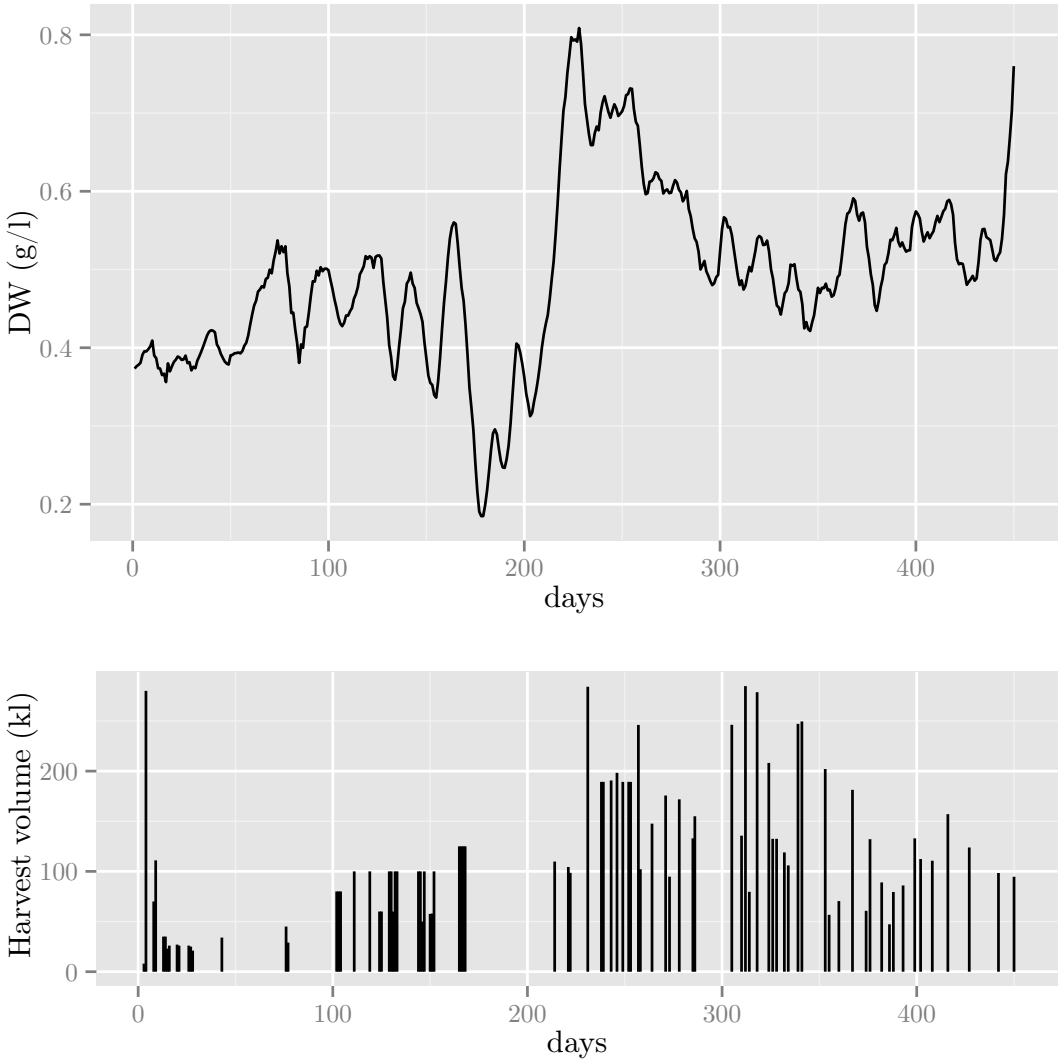


Figure A.1: DW (g/l) and harvest volume (kl) in time.

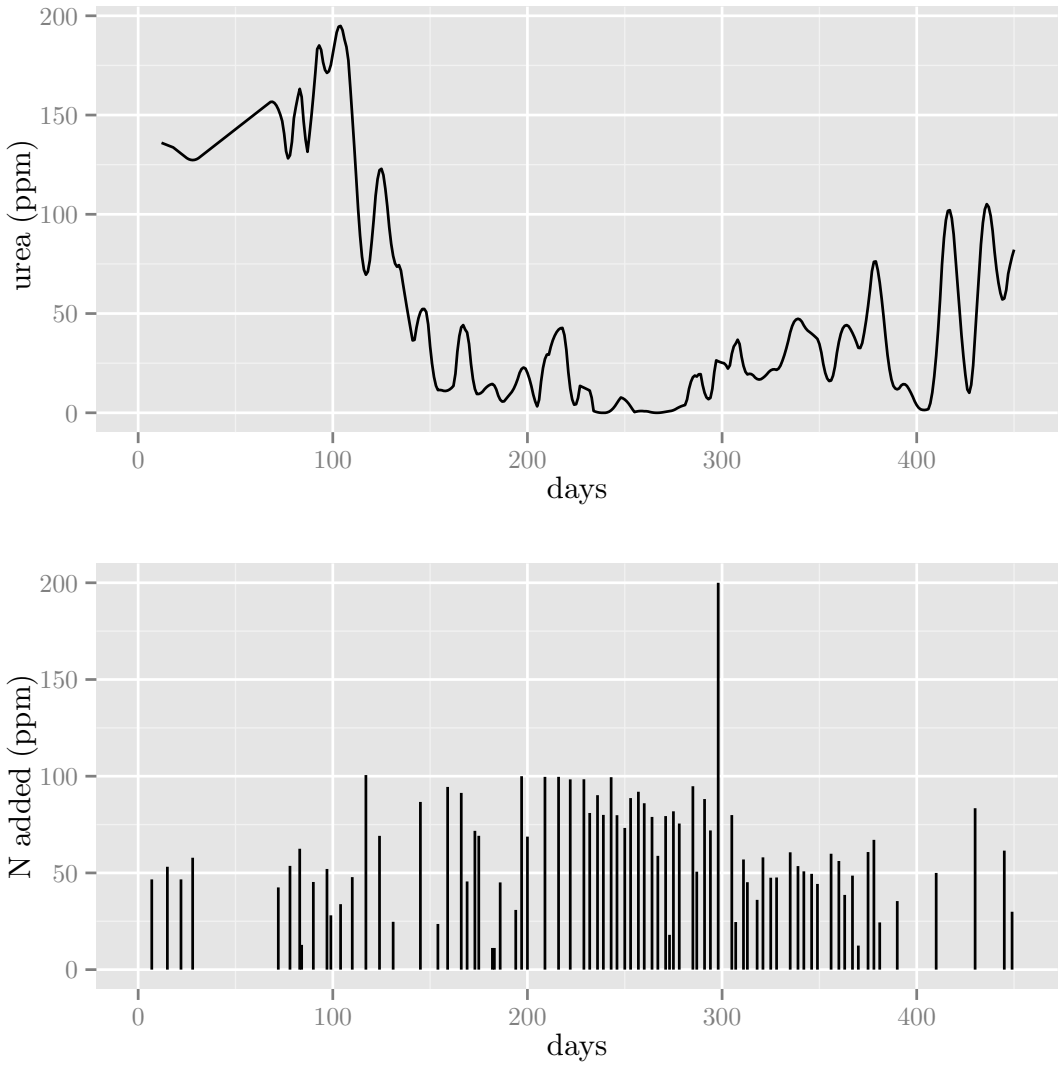


Figure A.2: Measured urea levels and N addition (mostly through urea addition) data.

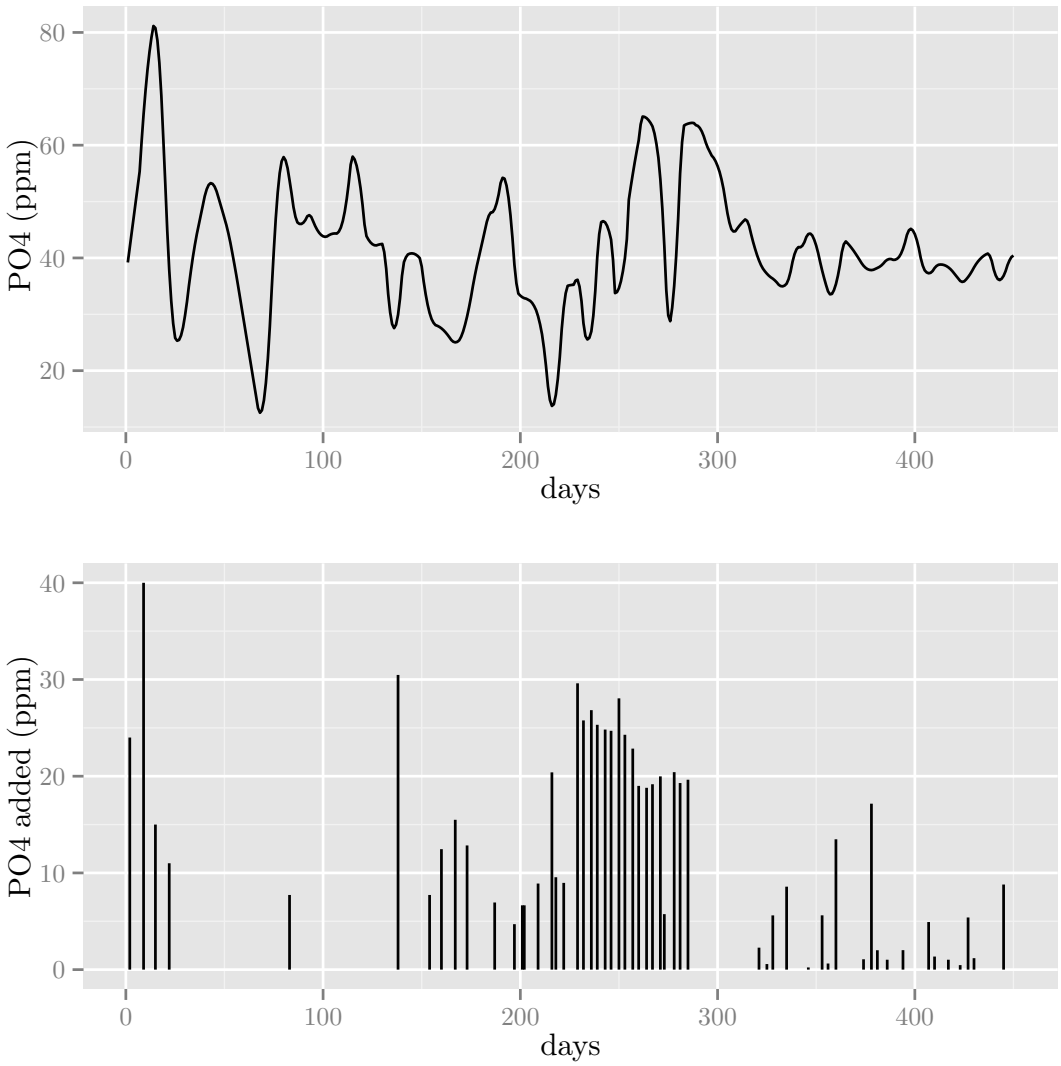


Figure A.3: Measured PO4 levels and PO4 addition data.

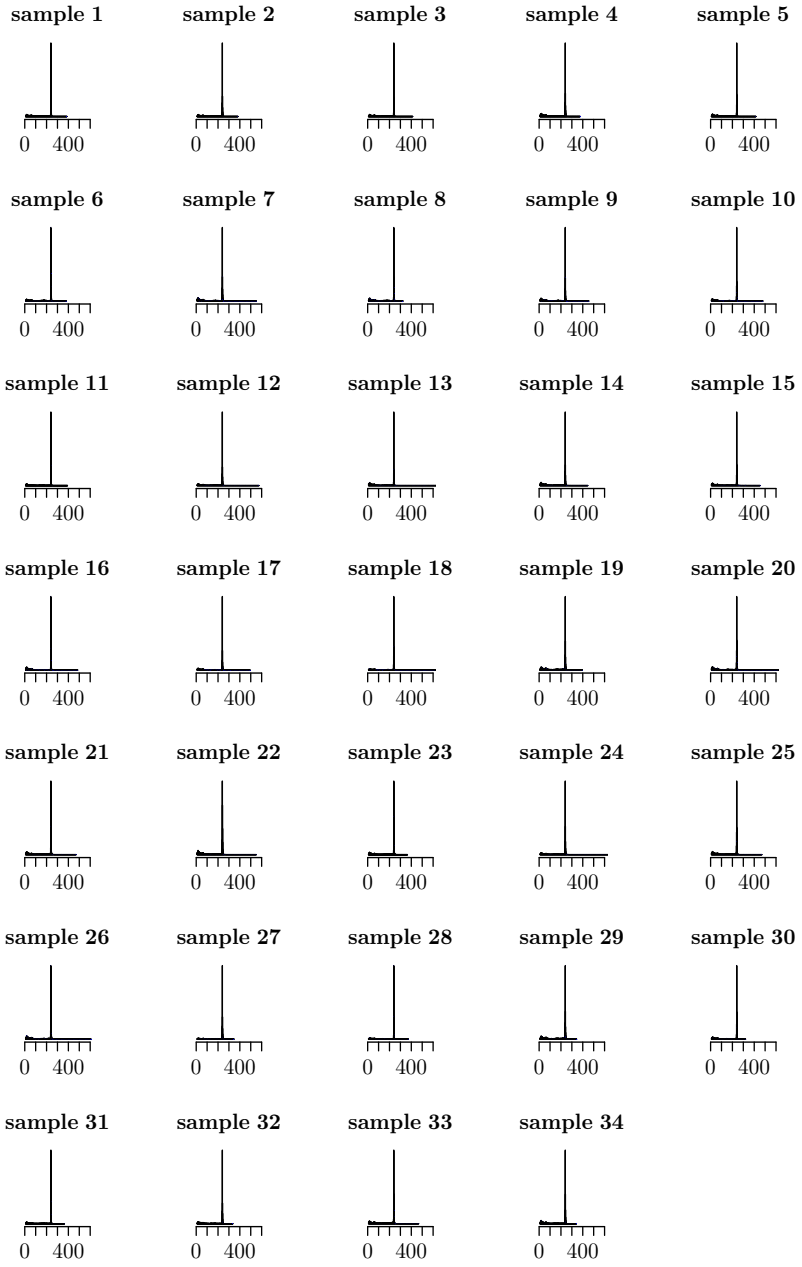


Figure A.4: Read length distribution for 16S data, chips 1, 2 and, 3.

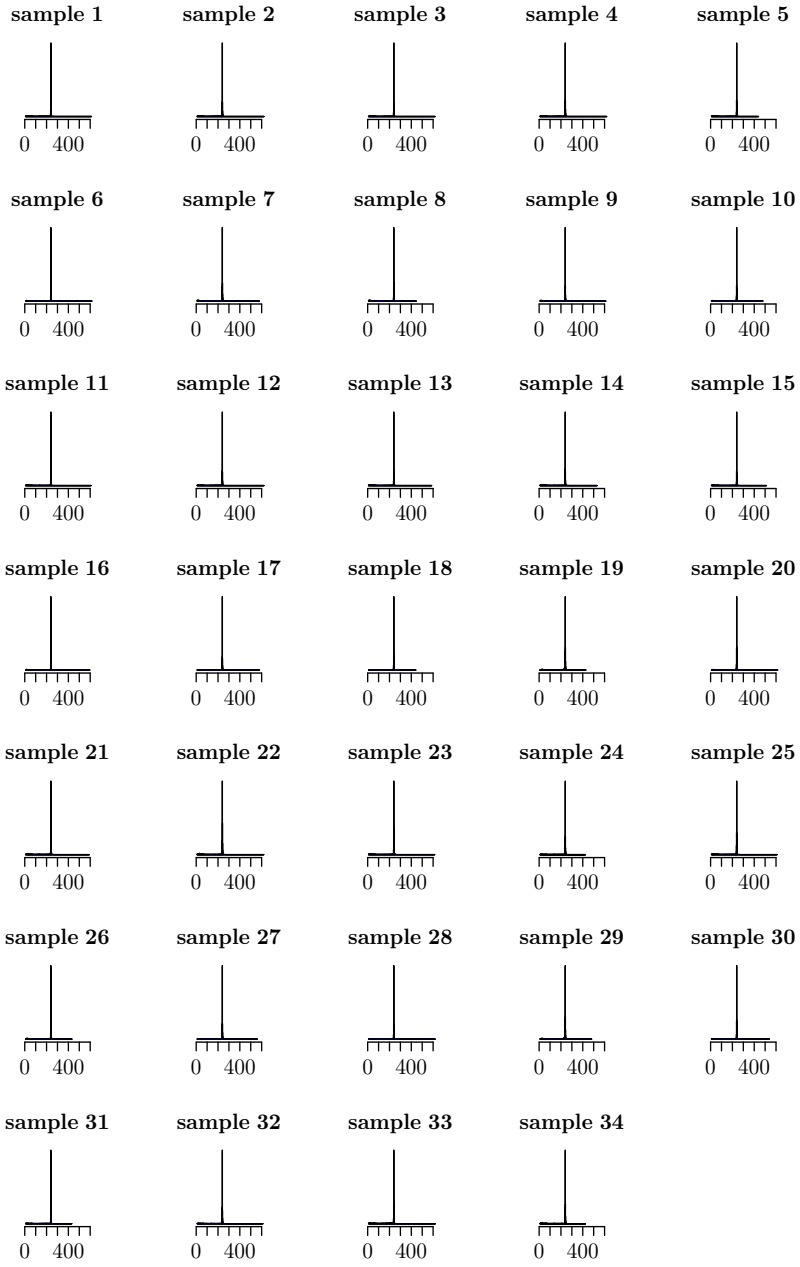


Figure A.4: Read length distribution for 16S data, chips 1, 2 and, 3, continued.

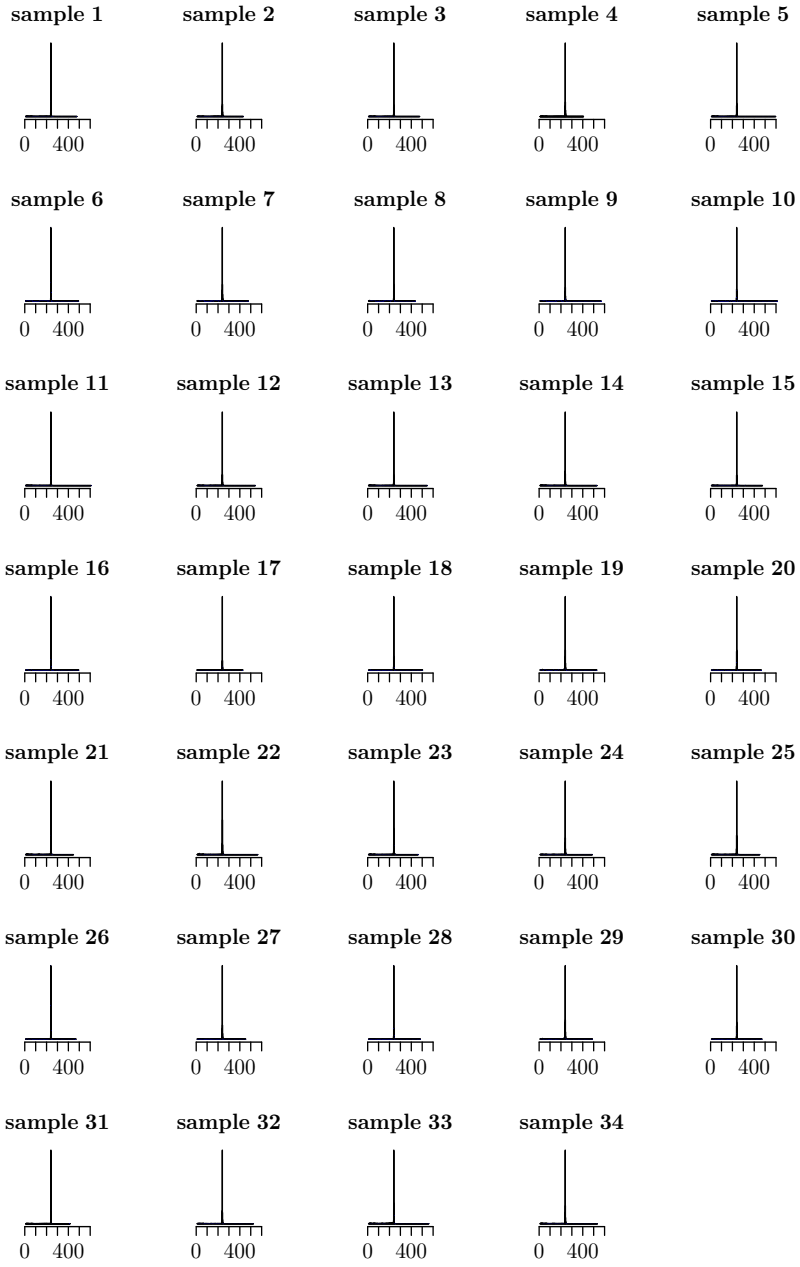


Figure A.4: Read length distribution for 16S data, chips 1, 2 and, 3, continued.

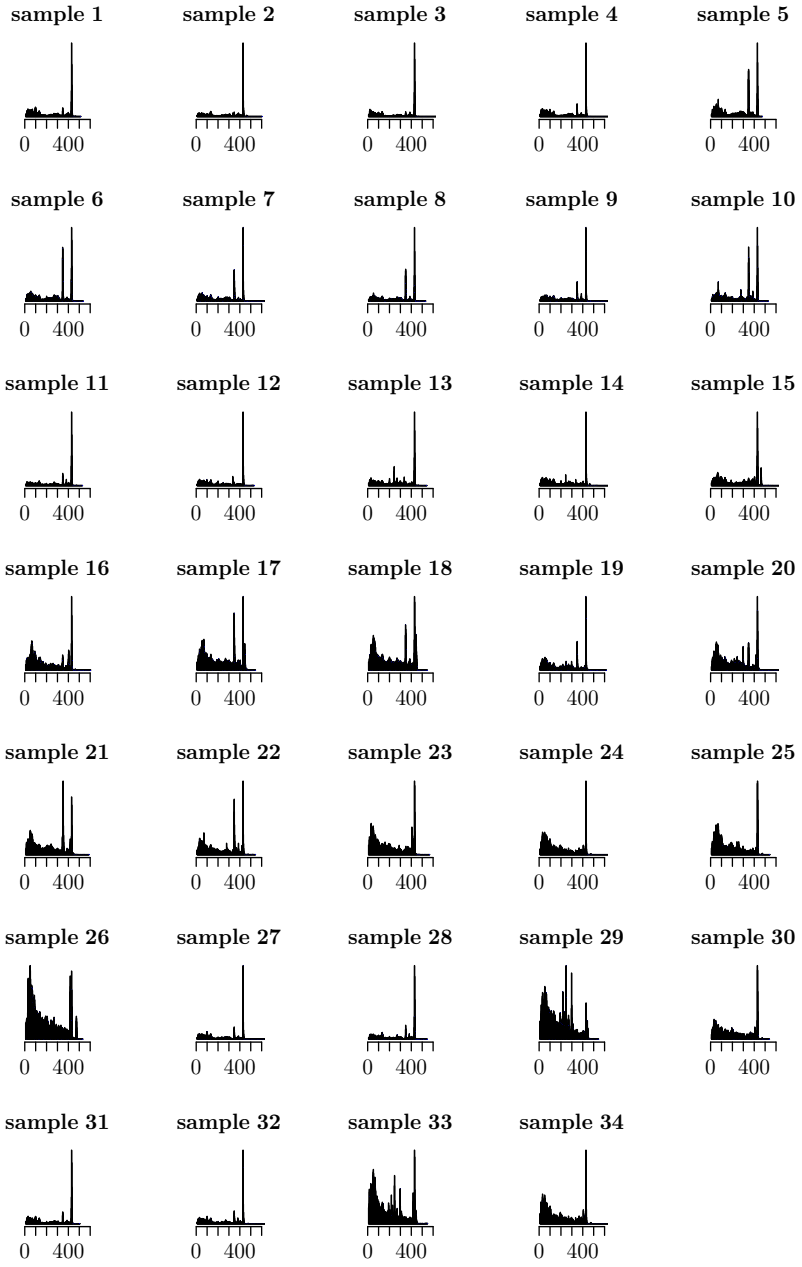


Figure A.5: Read length distribution for ITS2 data, chips 2, 3, 4, and 5.

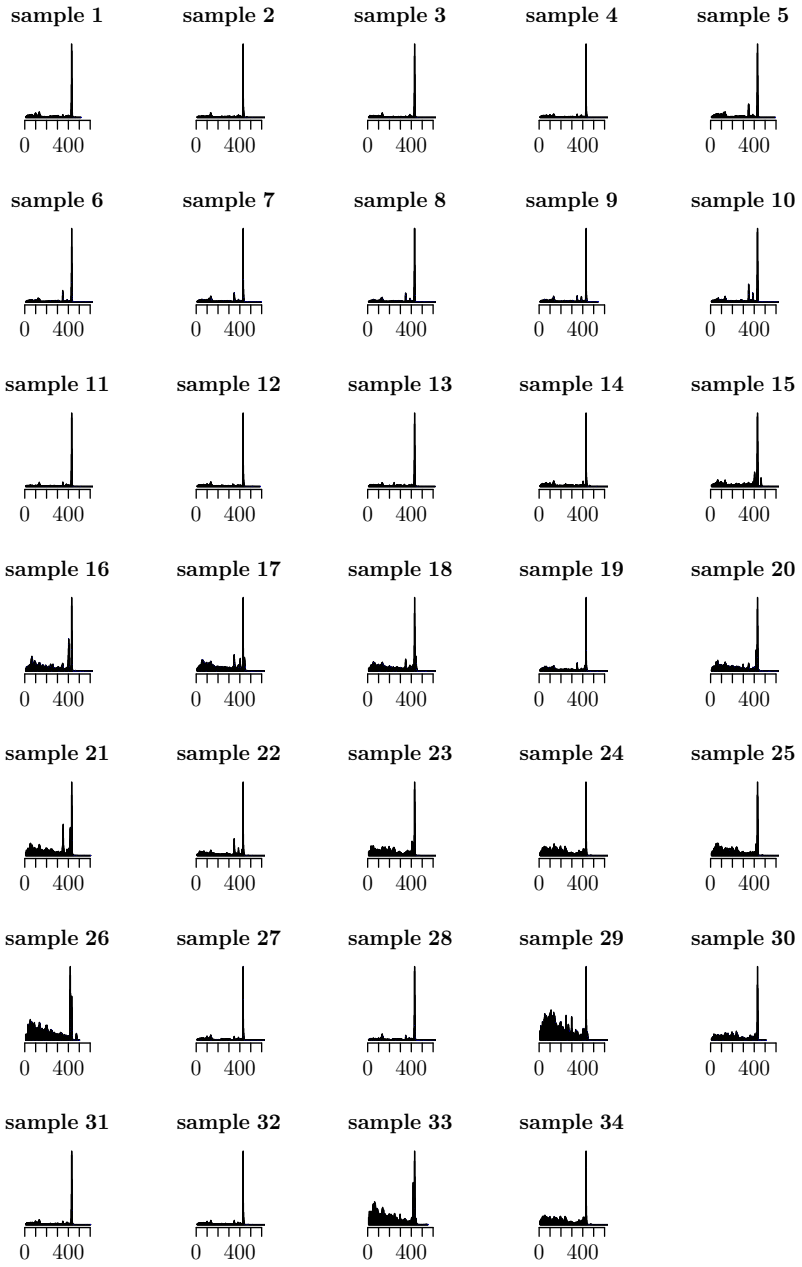


Figure A.5: Read length distribution for ITS2 data, chips 2, 3, 4, and 5, continued.

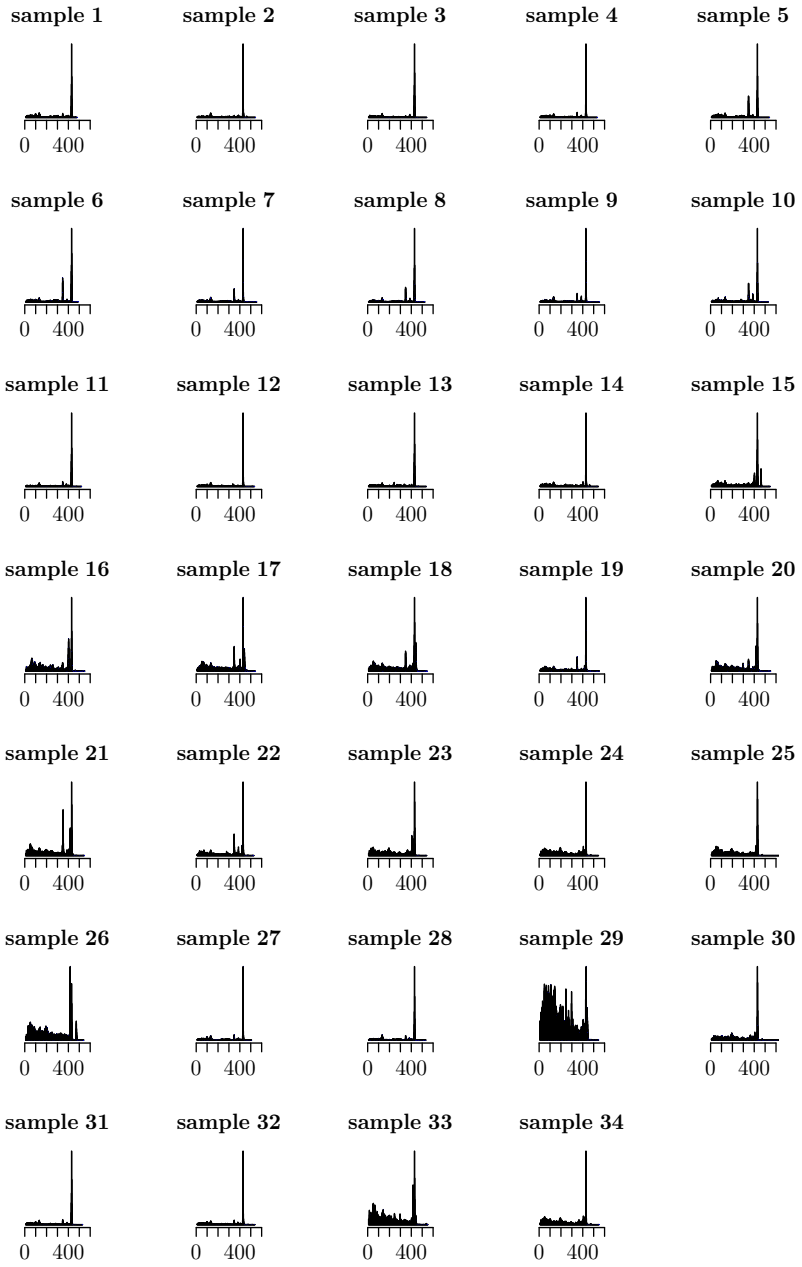


Figure A.5: Read length distribution for ITS2 data, chips 2, 3, 4, and 5, continued.

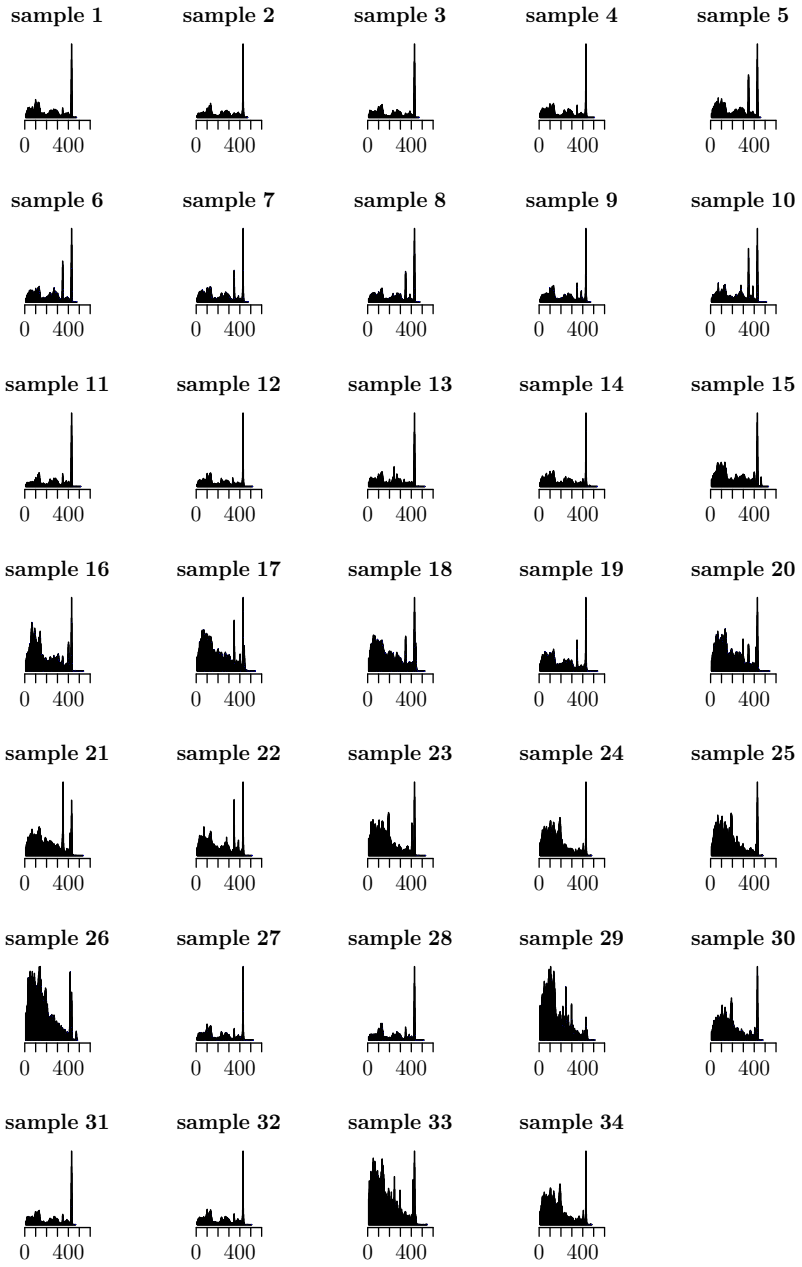


Figure A.5: Read length distribution for ITS2 data, chips 2, 3, 4, and 5, continued.

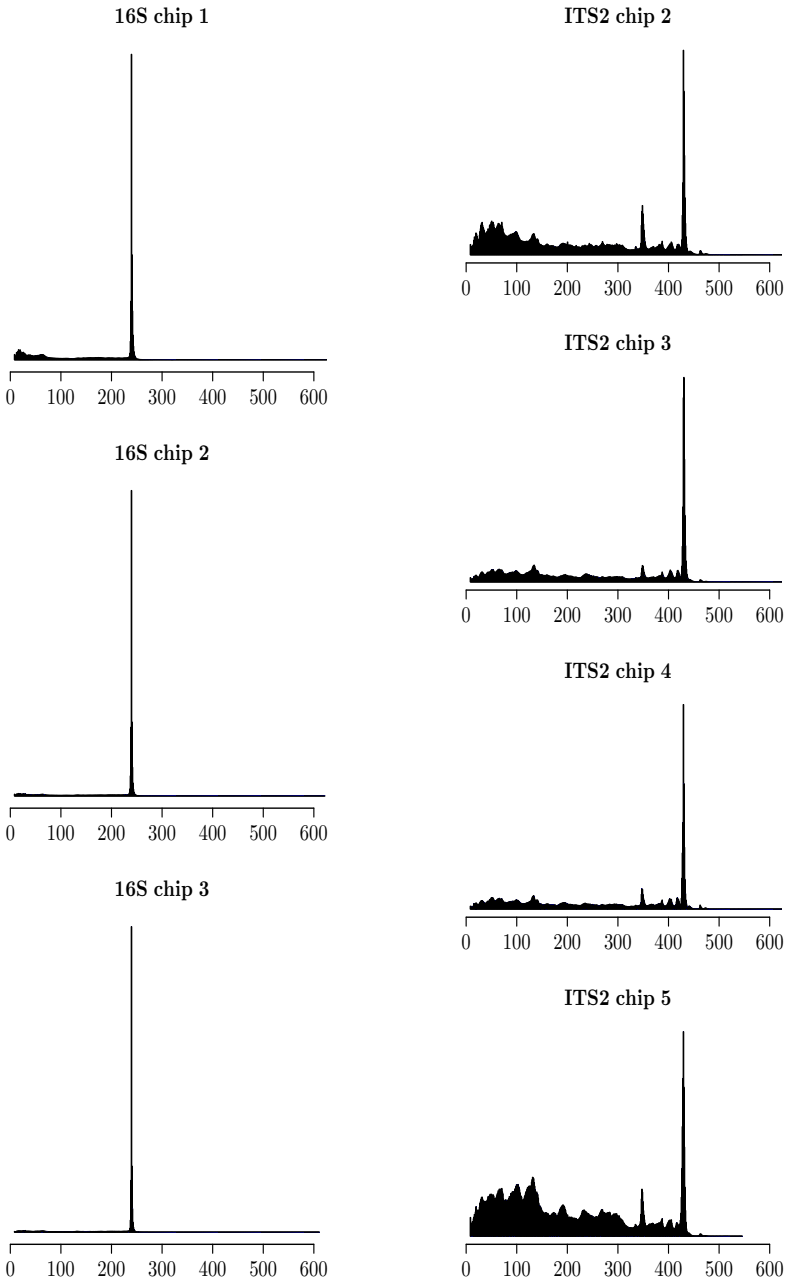
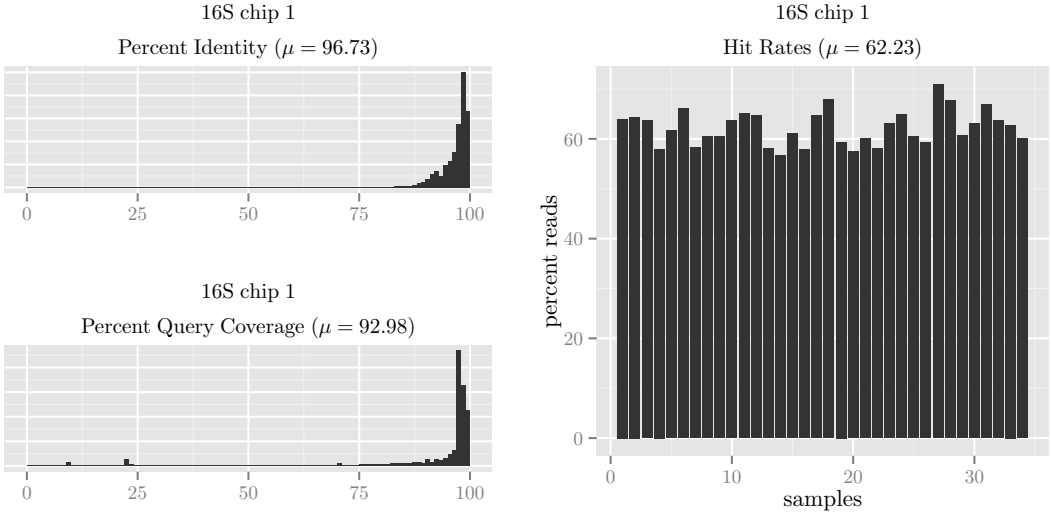
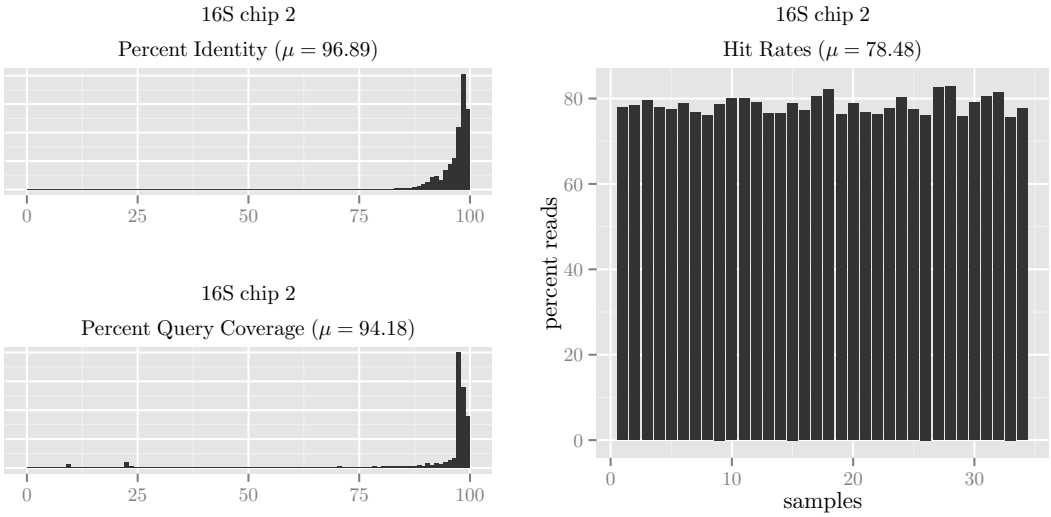


Figure A.6: Read length distributions for all 16S (A.6a) and ITS2 (A.6b) data.

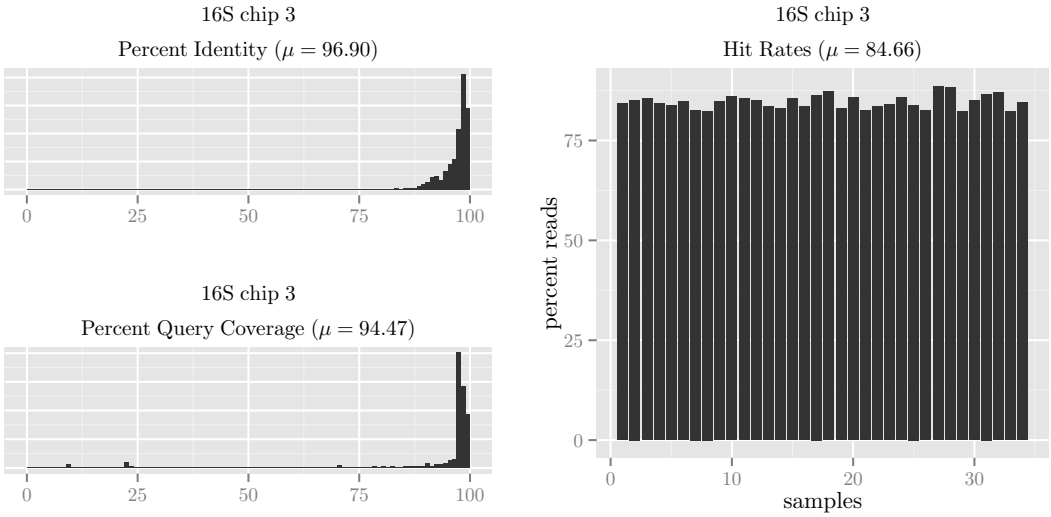


(a)



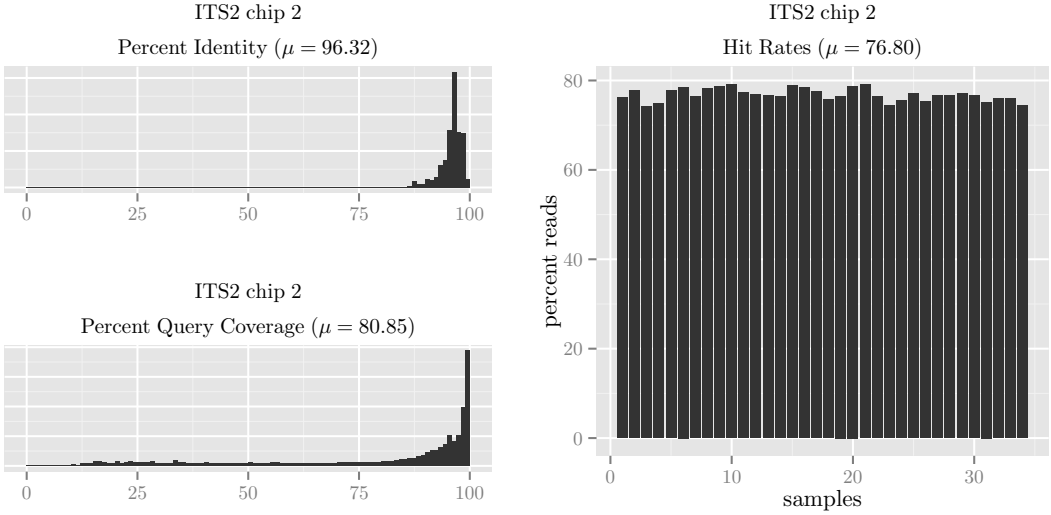
(b)

Figure A.7: Percent identities (%ID) and query coverages (%COV) of mapping sequences for all 16S chips: Figures A.7a, A.7b, A.7c shows the percent identities (%ID) and query coverages (%COV) of mapping sequences for chips 1, 2, 3; together with the percentages of sequences that are accepted as hit, after applying the 80% and 90% %COV and %ID cutoffs for all 34 samples.

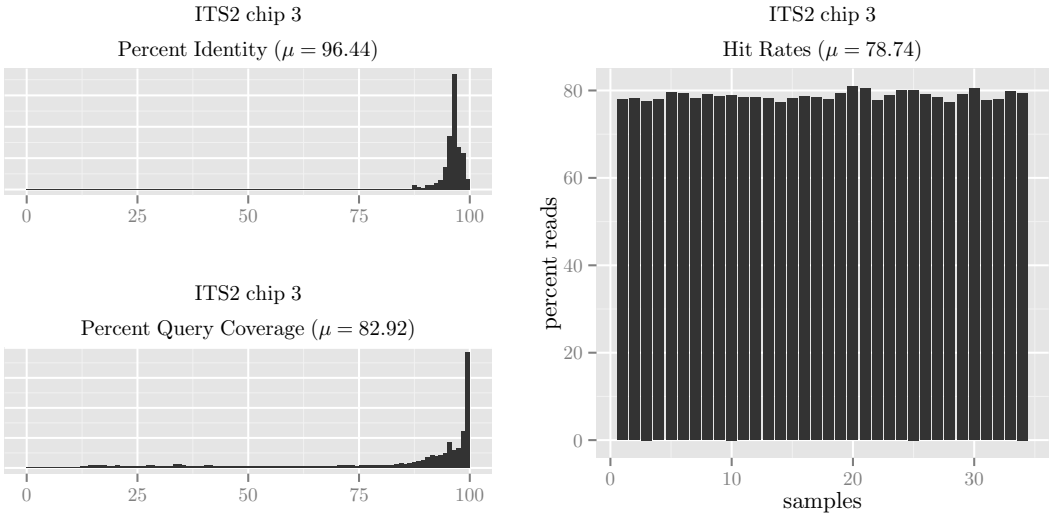


(c)

Figure A.7: Percent identities (%ID) and query coverages (%COV) of mapping sequences for all 16S chips: Figures A.7a, A.7b, A.7c shows the percent identities (%ID) and query coverages (%COV) of mapping sequences for chips 1, 2, 3; together with the percentages of sequences that are accepted as hit, after applying the 80% and 90% %COV and %ID cutoffs for all 34 samples, continued.

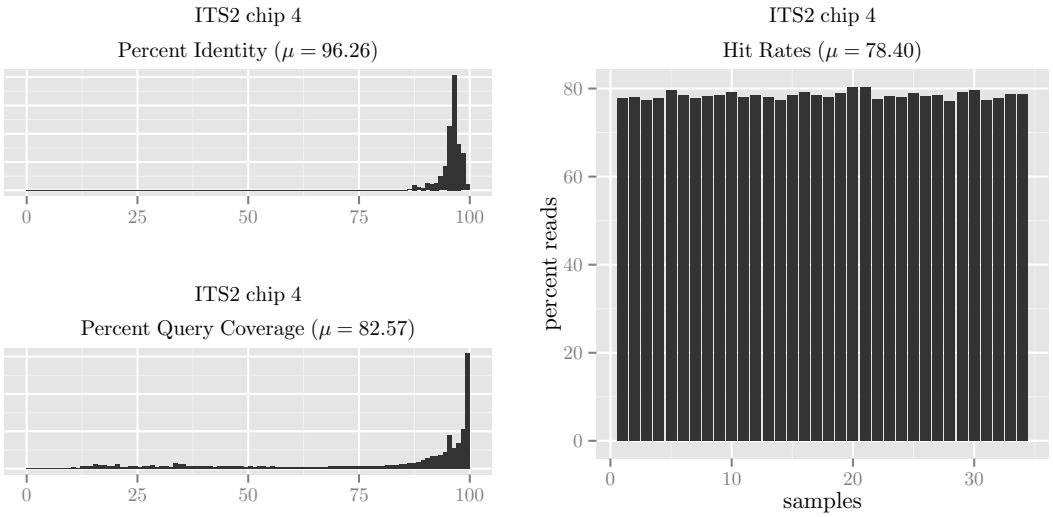


(a)

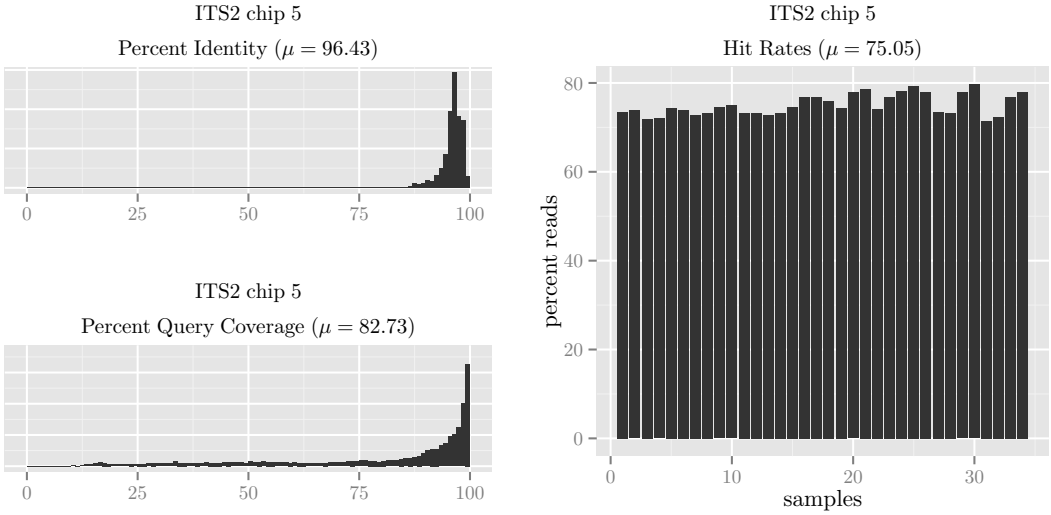


(b)

Figure A.8: Percent identities (%ID) and query coverages (%COV) of mapping sequences for all ITS2 chips: Figures A.8a, A.8b, A.8c, A.8d shows the percent identities (%ID) and query coverages (%COV) of mapping sequences for chips 2, 3, 4, 5; together with the percentages of sequences that are accepted as hit, after applying the 80% and 90% %COV and %ID cutoffs for all 34 samples.



(c)



(d)

Figure A.8: Percent identities (%ID) and query coverages (%COV) of mapping sequences for all ITS2 chips: Figures A.8a, A.8b, A.8c, A.8d shows the percent identities (%ID) and query coverages (%COV) of mapping sequences for chips 2, 3, 4, 5; together with the percentages of sequences that are accepted as hit, after applying the 80% and 90% %COV and %ID cutoffs for all 34 samples, continued.

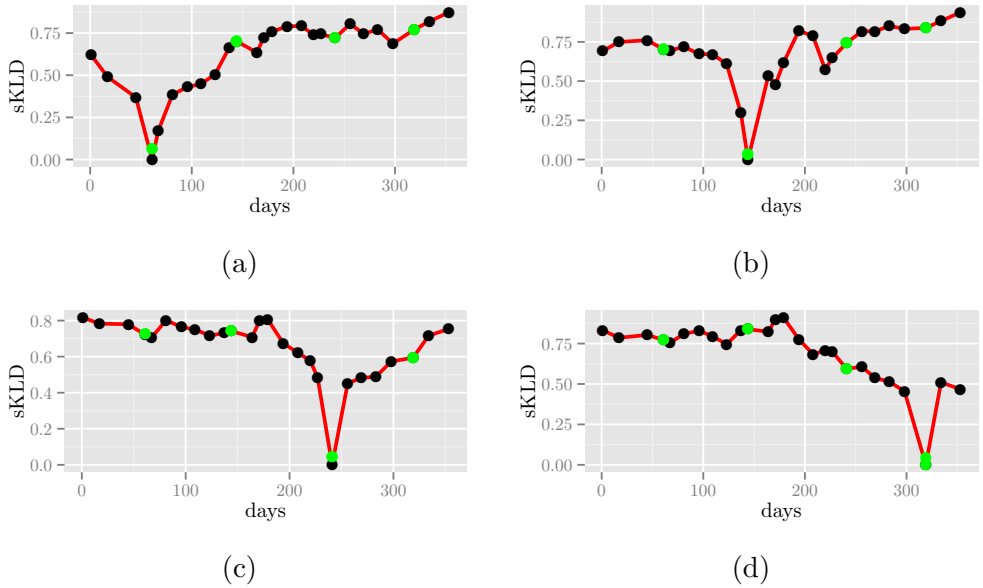


Figure A.9: Divergences across selected samples: A.9a, A.9b, A.9c, and A.9d shows the distances between sample 4, 11, 19, 24, and all other samples, respectively for 16S data, whereas A.9e, A.9f, A.9g, and A.9h shows it for ITS2 data. Grey points correspond to original samples, while green points represent the technical replicates of the samples sharing their x-axis value. The zero KL distance (y-axis) on each plot indicates which sample all other samples are compared against. Good reproducibility is achieved when the green points superimposed over the fixed samples (4, 11, 19, 24) also have zero KLD values.

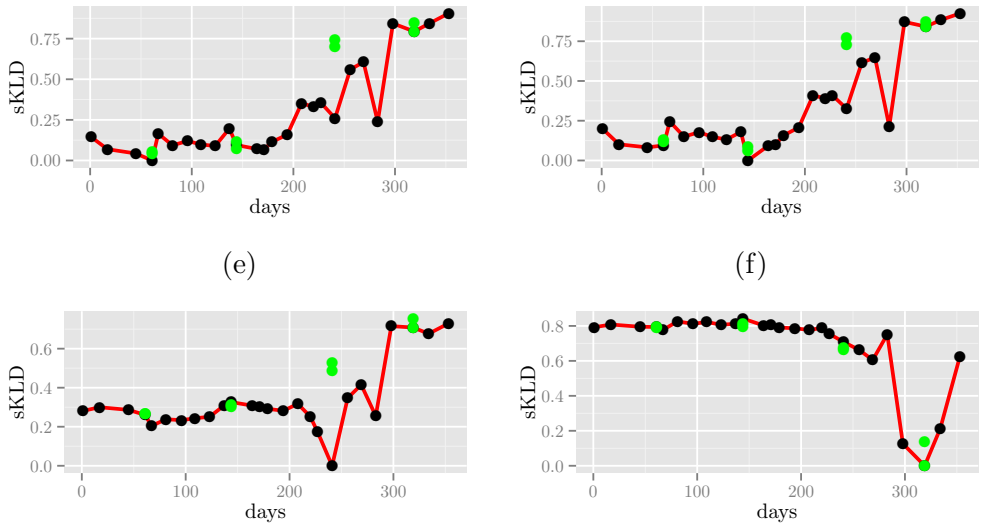
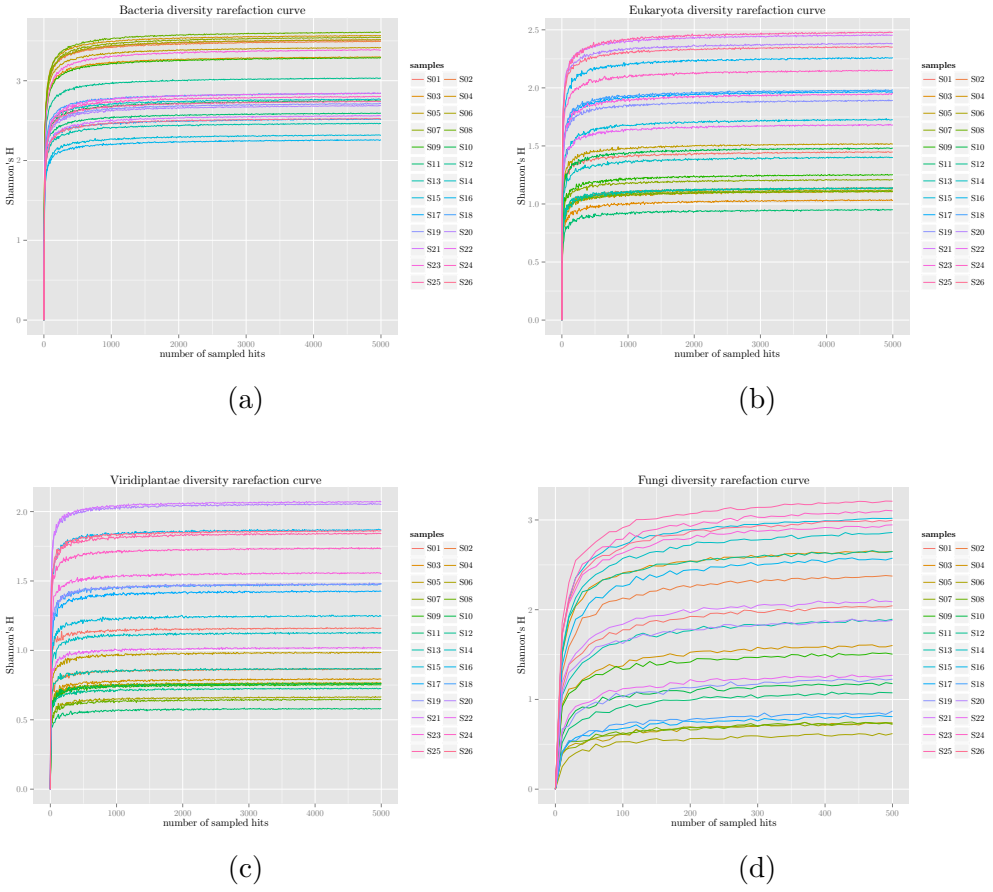


Figure A.9: Divergences across selected samples: A.9a, A.9b, A.9c, and A.9d shows the distances between sample 4, 11, 19, 24, and all other samples, respectively for 16S data, whereas A.9e, A.9f, A.9g, and A.9h shows it for ITS2 data. Grey points correspond to original samples, while green points represent the technical replicates of the samples sharing their x-axis value. The zero KL distance (y-axis) on each plot indicates which sample all other samples are compared against. Good reproducibility is achieved when the green points superimposed over the fixed samples (4, 11, 19, 24) also have zero KLD values, continued.



(a) (b)

(c) (d)

Figure A.10: Rarefaction Curves: Depicts the converging diversity (Shannon H) rarefaction curves for Bacteria, Eukaryota, Viridiplantae, algae, and Fungi, over all 16S and ITS2 reference sequences, averaged over 100 interations.

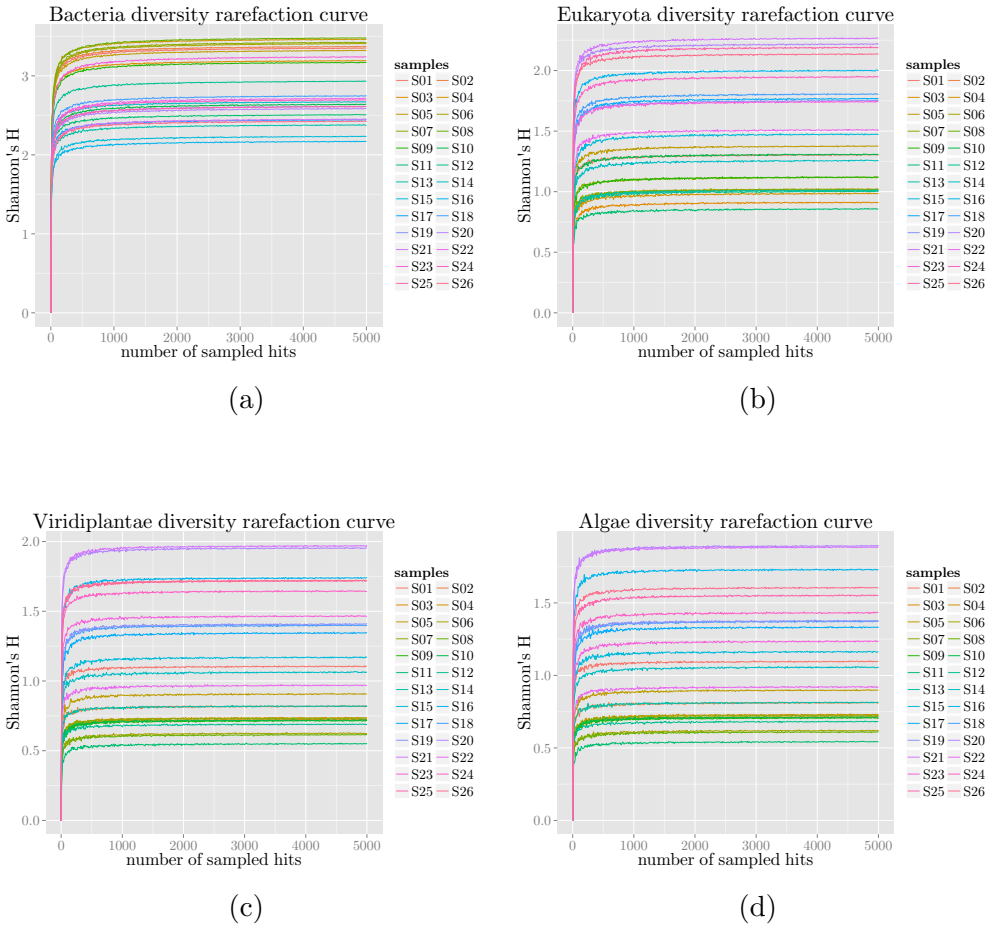
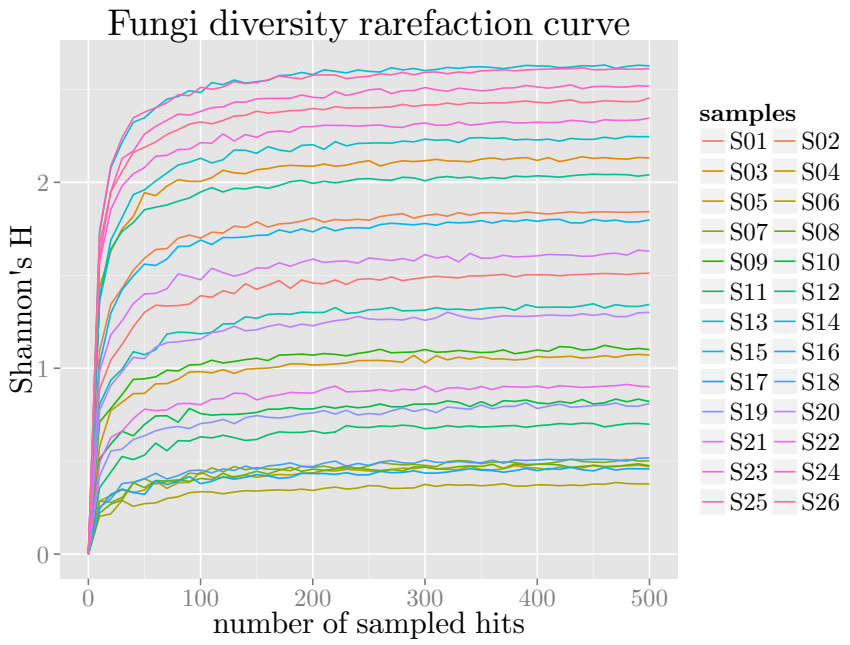


Figure A.11: Rarefaction Curves (top species): Depicts the converging diversity (Shannon H) rarefaction curves for Bacteria, Eukaryota, Viridiplantae, algae, and Fungi, over the top 2000 and 200 16S and ITS2 reference sequences, averaged over 100 interations.



(e)

Figure A.11: Rarefaction Curves (top species): Depicts the converging diversity (Shannon H) rarefaction curves for Bacteria, Eukaryota, Viridiplantae, algae, and Fungi, over the top 2000 and 200 16S and ITS2 reference sequences, averaged over 100 interations, continued.

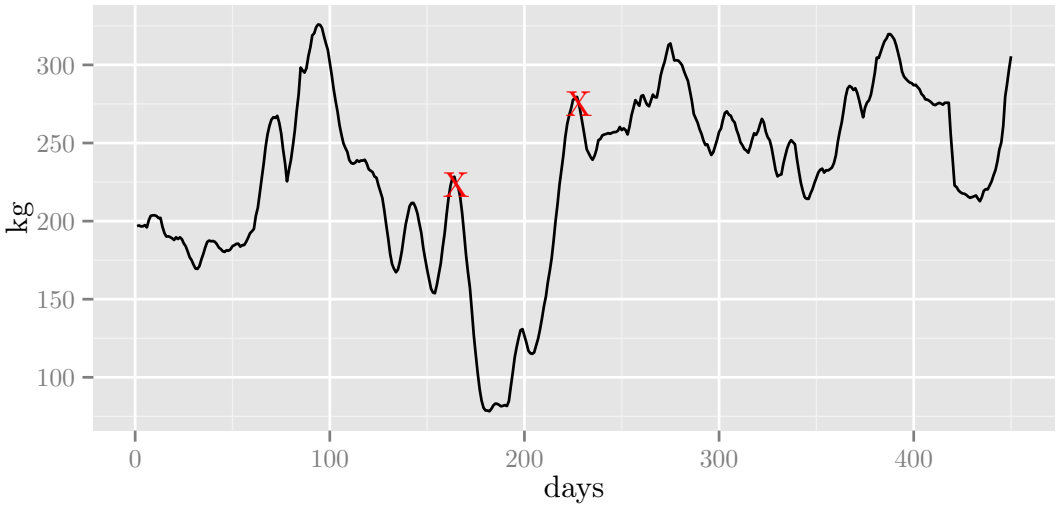
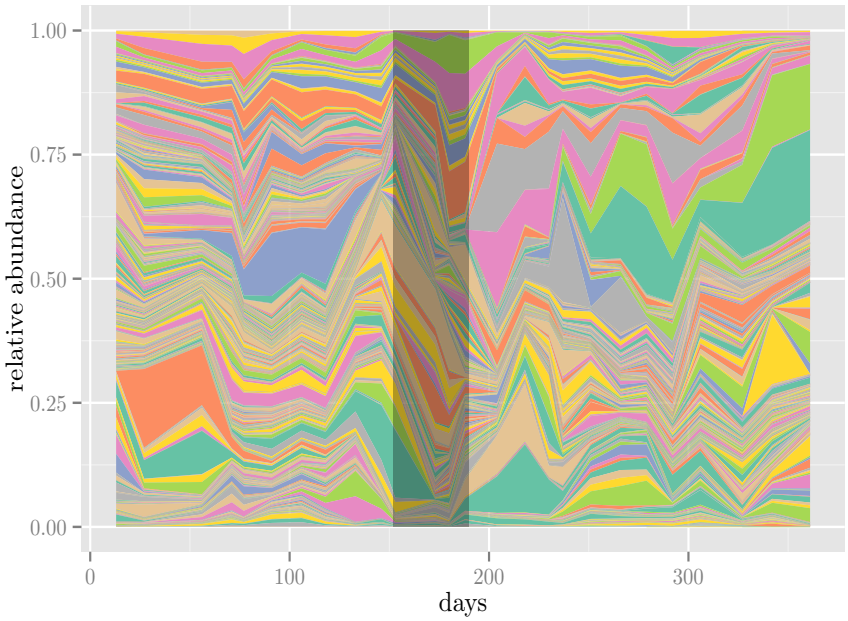
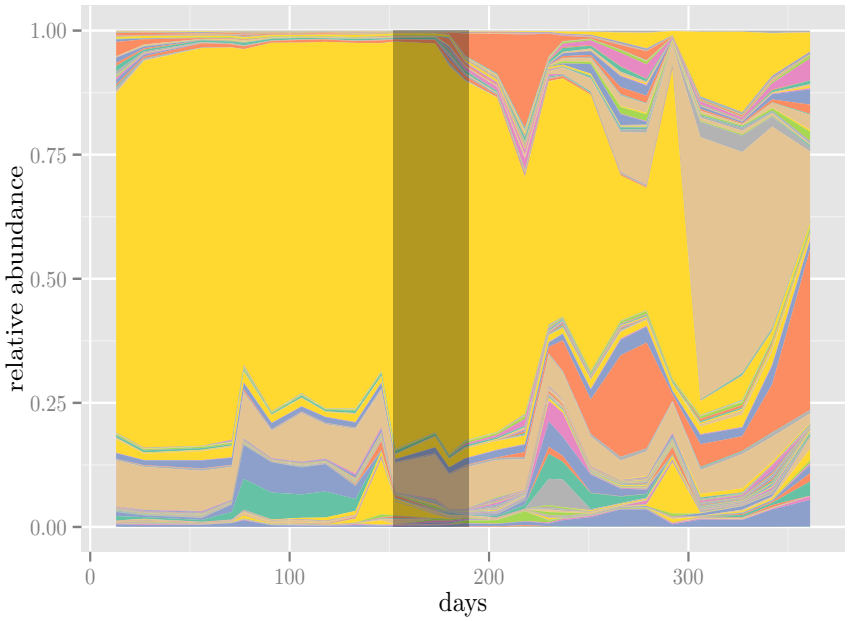


Figure A.12: Dry weight (kg): Algal dry weight in kg, with peaks on days 165, and 228 marked.

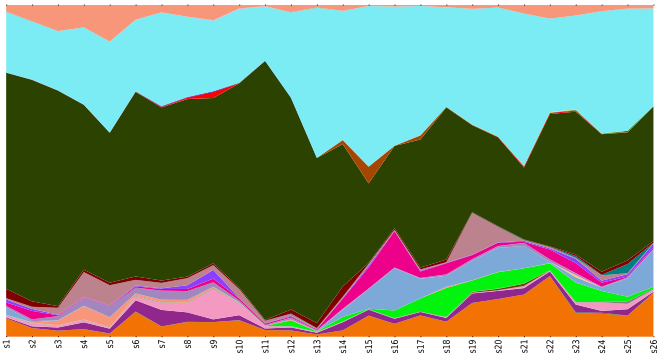


(a) Top 1000 sequences hit in GreenGenes.

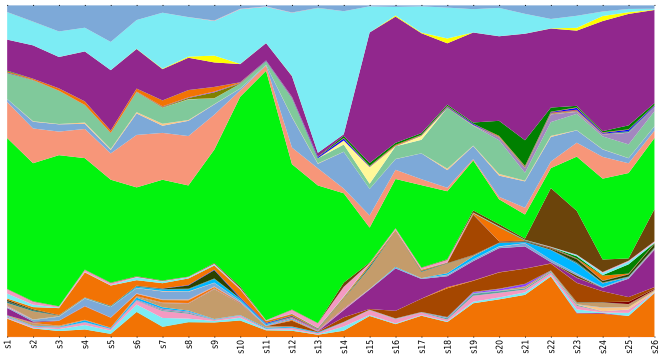


(b) Top 200 sequences hit in constructed ITS2 database from NCBI.

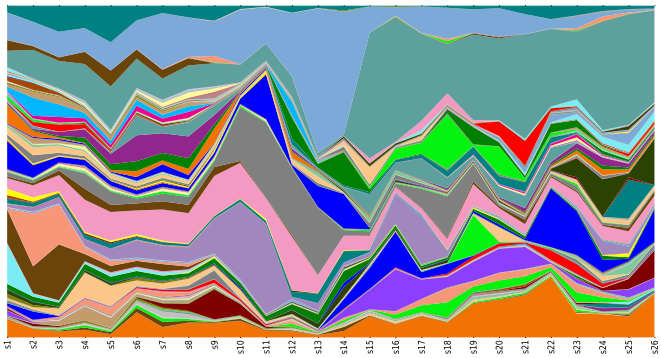
Figure A.13: Finest granularity (sequence level) area plots: Top hit reference sequences in 16S, using two different databases, and ITS2 data, respectively.



(a)

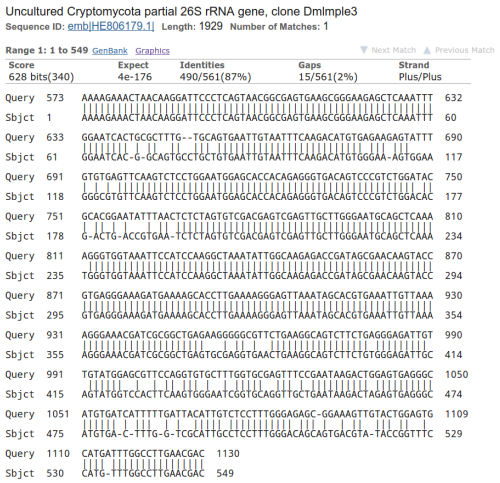


(b)

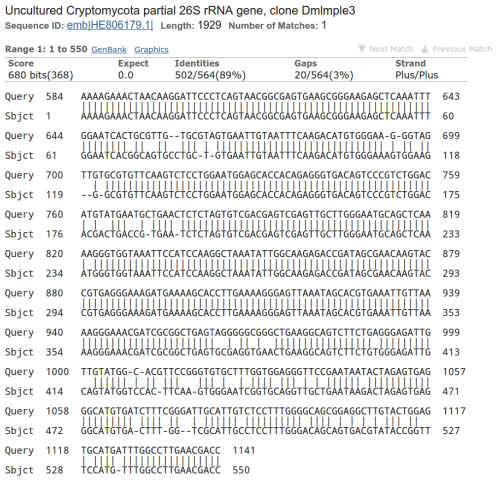


(c)

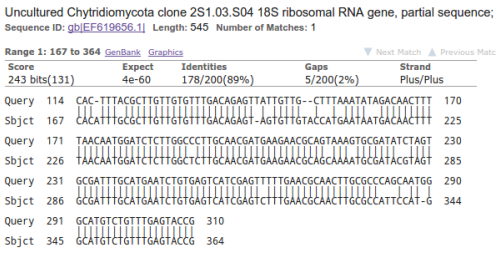
Figure A.14: Bralizian Microbiome Pipeline area plots at phylum (A.14a), class (A.14b), and genus (A.14c) levels for 16S data. Taxa not shown.



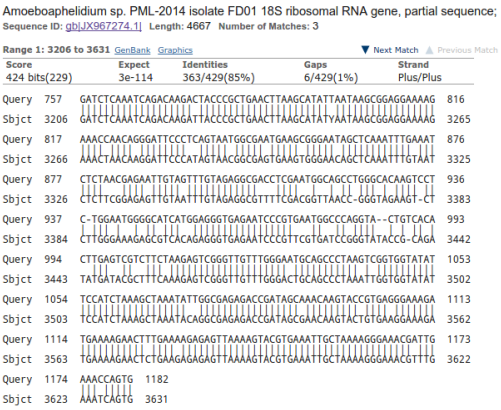
(a) Alignment of GI: 532165669



(b) Alignment of GI: 532165968



(c) Alignment of GI: 194354257



(d) Alignment of GI: 532165358

Figure A.15: Alignment results of the five most abundant fungal sequences to their highest scoring BLAST hits of known phylum level taxonomy.

Amoeboaphelidium sp. PML-2014 isolate FD01 18S ribosomal RNA gene, partial sequence;
Sequence ID: gb|JX967274.1| Length: 4667 Number of Matches: 3

Range 1: 3205 to 3622 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
431 bits(233)	2e-116	359/421(85%)	6/421(1%)	Plus/Plus

```
Query  639   CGATCTCAAATCAGACAAGACTACCCGCTGAACTTAAGCATATTAATAAGCGGAGGAAAA  698
             ||||||||||||||||||| ||||||||||||||||||||| |||||||||||||||||
Sbjct  3205  CGATCTCAAATCAGACAAGATTACCCGCTGAACTTAAGCATATYAATAAGCGGAGGAAAA  3264

Query  699   GAAACCAACAGGGATTCCCCCAGTAATGGCGAATGAAGCGGGAATAGCTCAAATTTTTAA  758
             ||||| |||| |||||||| |||||| ||||| |||||| |||||||||||| |||||
Sbjct  3265  GAAACTAACAAGGATTCCCATAGTAACGGCGAGTGAAGTGGGAACAGCTCAAATTTGTAA  3324

Query  759   TCTCTTCGGAGAGTTGTAATTTGAAGAGGTGACATCGTCGTCTTTGCCTGGTCAAAGTCT  818
             |||||||||||||||||||| |||||| |  ||| ||  ||  || |||  ||||||
Sbjct  3325  TCTCTTCGGAGAGTTGTAATTTGTAGAGGCGTTTTCGACG-GTTAACCGGGTAGAAGTCT  3383

Query  819   CCTGGAAAGGAGCAACATGGAGGGTGAAATTCCCGTATC-CGA-CCAGGTTGAAGGC-GC  875
             | |||| |||| |||  || |||||||| |||||| ||  || || ||  |||  | ||
Sbjct  3384  CTTGGGAAAGAGCGTCACAGAGGGTGAGAATCCCGT-TCGTGATCCGGGTATACCGCAGA  3442

Query  876   TCTTGATTCATTCTCAAAGAGTCGGGTTGCTTGAGACTGCAGCCCAAAGTGGGTGGTATA  935
             | |||| | |||||||||||||||||||||||||||||||||||| |||||||||||||
Sbjct  3443  T-ATGATACGCTTTCAAAGAGTCGGGTTGTTTGGGACTGCAGCCCTAAATTGGTGGTATA  3501

Query  936   TTCCATCTAAAGCTAAATATTGGCGAGAGACCGATAGCAAACAAGTACCGTGAGGGAAAG  995
             |||||||||||||||||| ||||||||||||||||||||| |||||||||| |||||||
Sbjct  3502  TTCCATCTAAAGCTAAATACAGGCGAGAGACCGATAGCGAACAAGTACTGTGAAGGAAAG  3561

Query  996   ATGAAAAGAACTTTGAAAAGAGAGTTAAAAGTACGTGAAATTGCTAAAAGGGAAACGTTT  1055
             |||||||||| |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3562  ATGAAAAGAACTCTGAAGAGAGAGTTAAAAGTACGTGAAATTGCTAAAAGGGAAACGTTT  3621

Query  1056  G  1056
             |
Sbjct  3622  G  3622
```

(e) Alignment of GI: 532166006

Figure A.15: Alignment results of the five most abundant fungal sequences to their highest scoring BLAST hits of known phylum level taxonomy, continued.

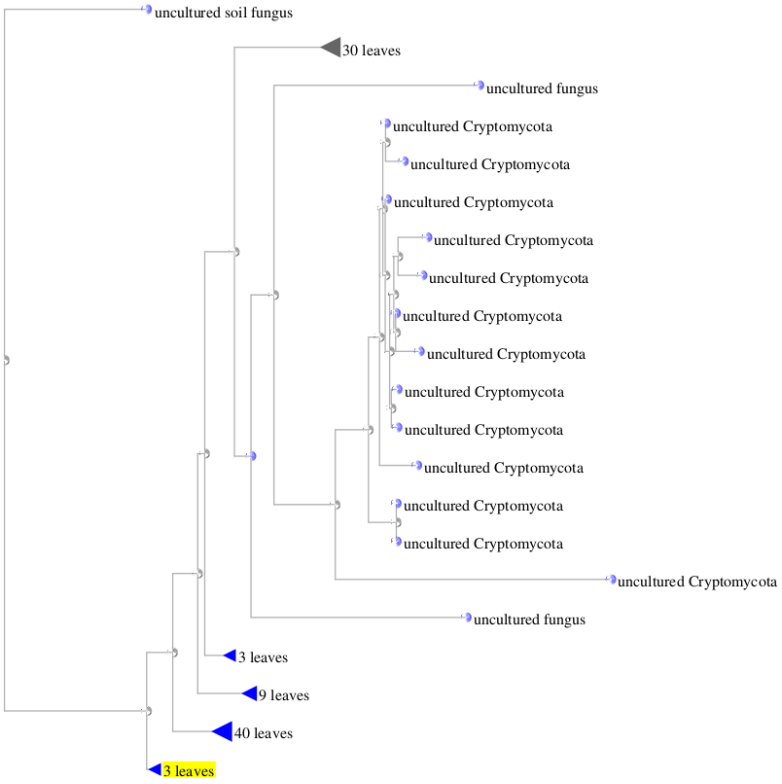


Figure A.16: Distance tree for sequence of interest: Distance tree for GI: 532165669, and GI: 532165968, collapsed on the branch highlighted in yellow.

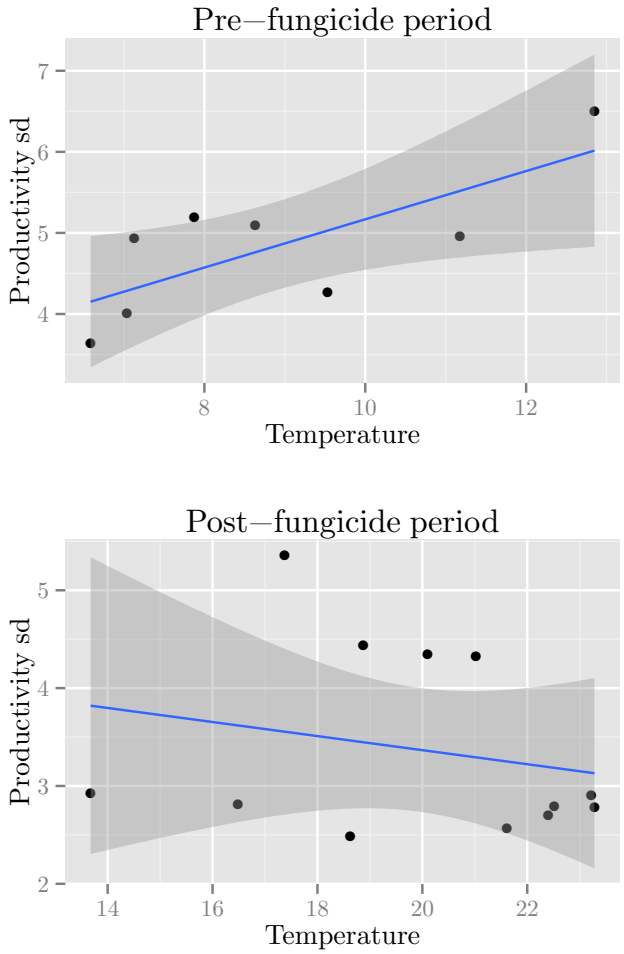


Figure A.17: Pre- and post-fungicide temperature and productivity variability relationship.

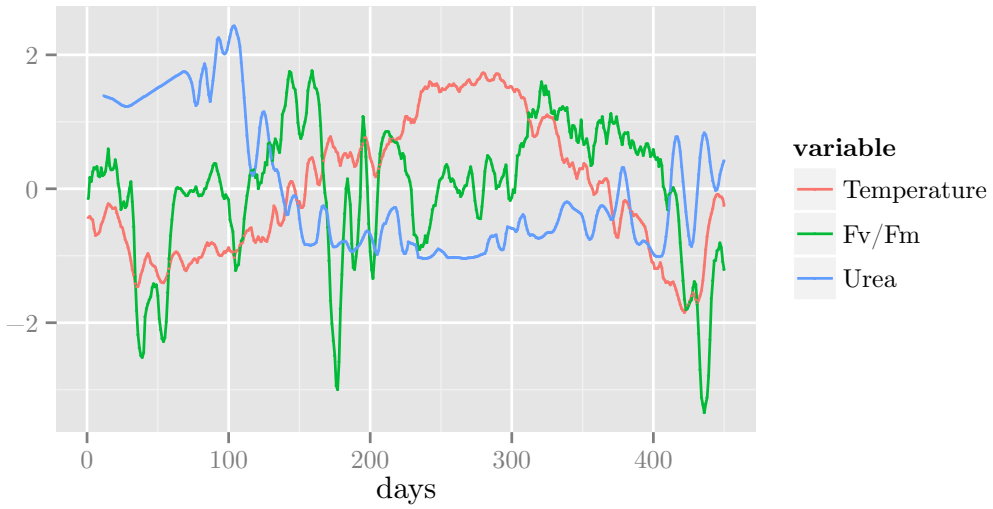


Figure A.18: Select Phenotypes: Relationship of temperature, urea, and photosynthetic health (F_v/F_m) over time, standardised by centering around their mean and division by their standard deviation.

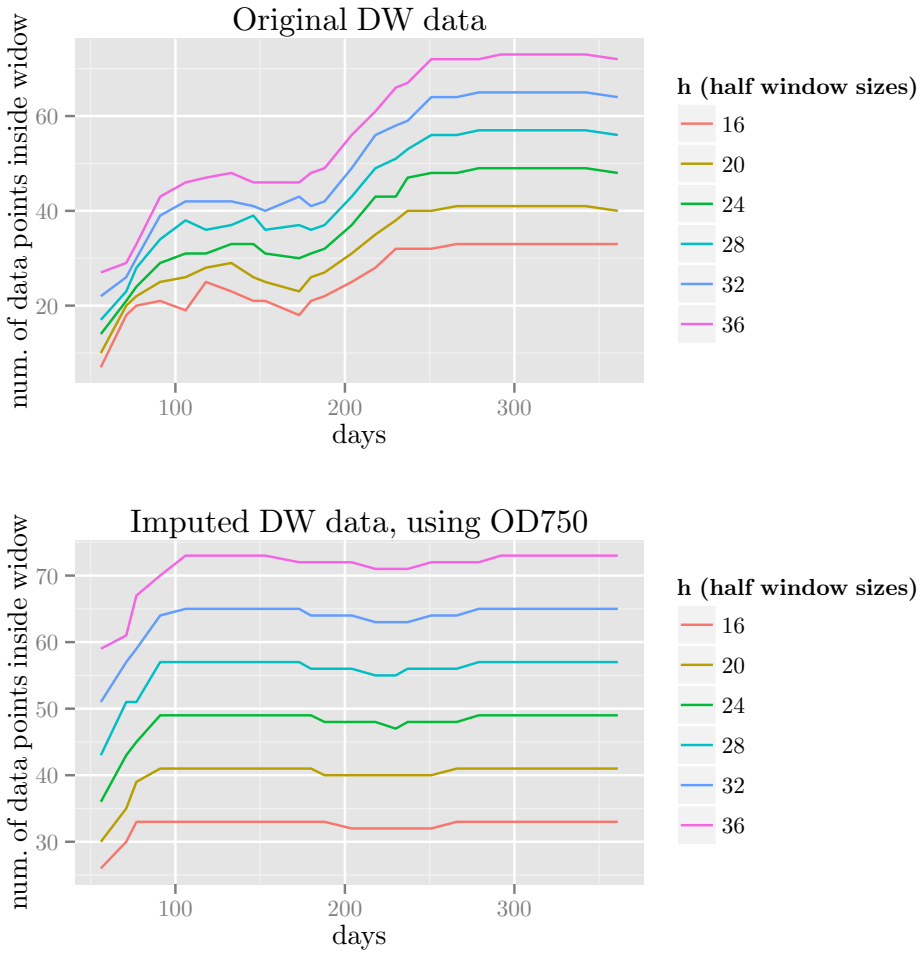


Figure A.19: Number of available data points inside given half window (h) in original and imputed (using OD 750) DW (g/l) data.

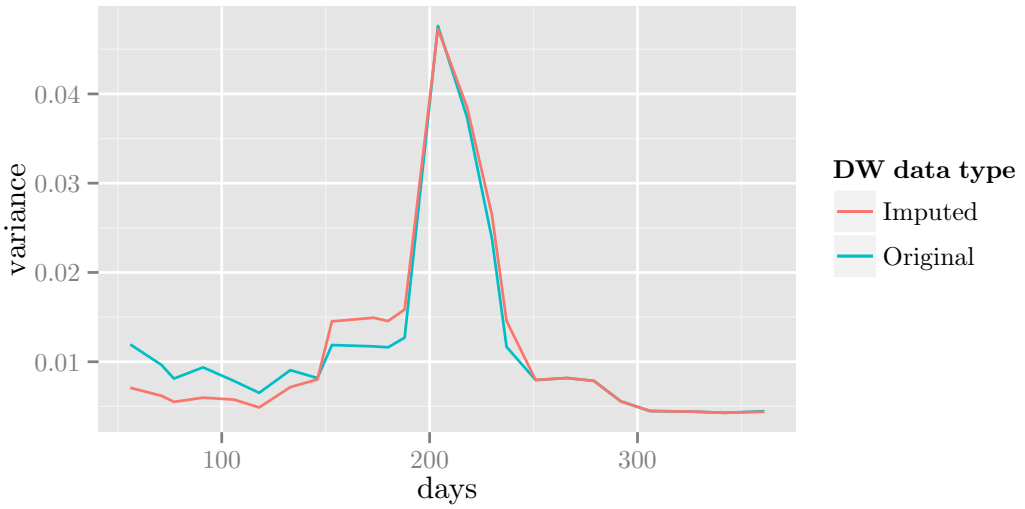


Figure A.20: Variance patterns of original and imputed (using OD 750) DW (g/l) data using half window size of $h = 28$ days.

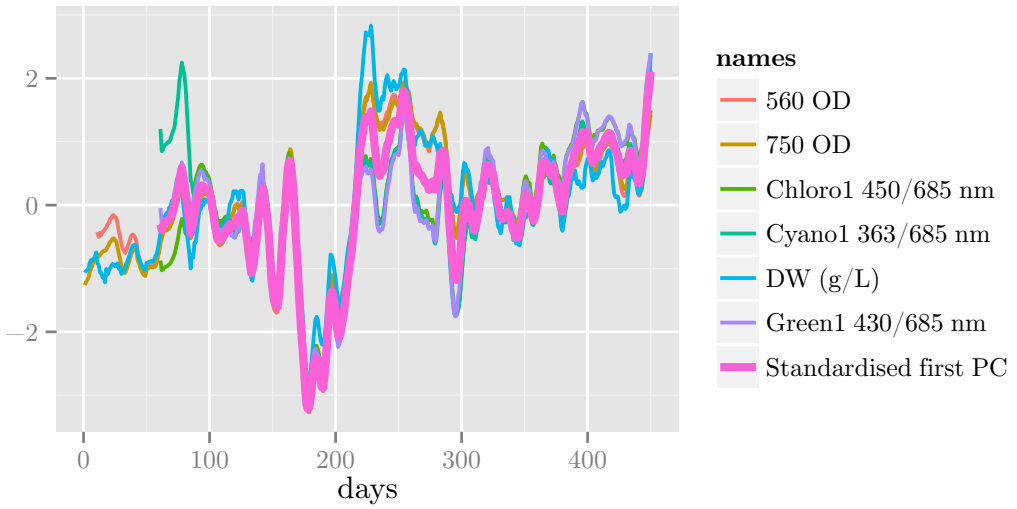
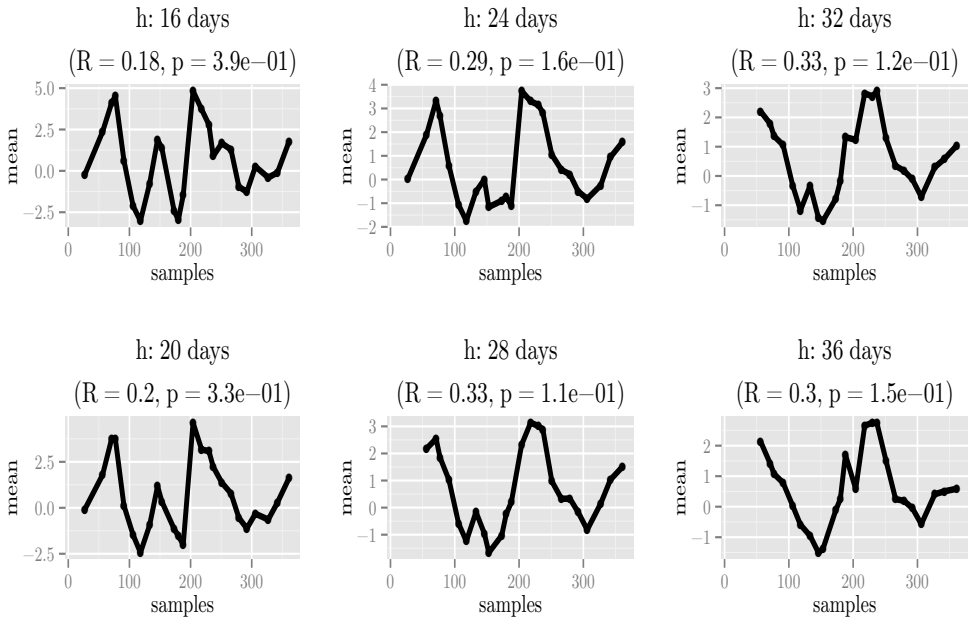
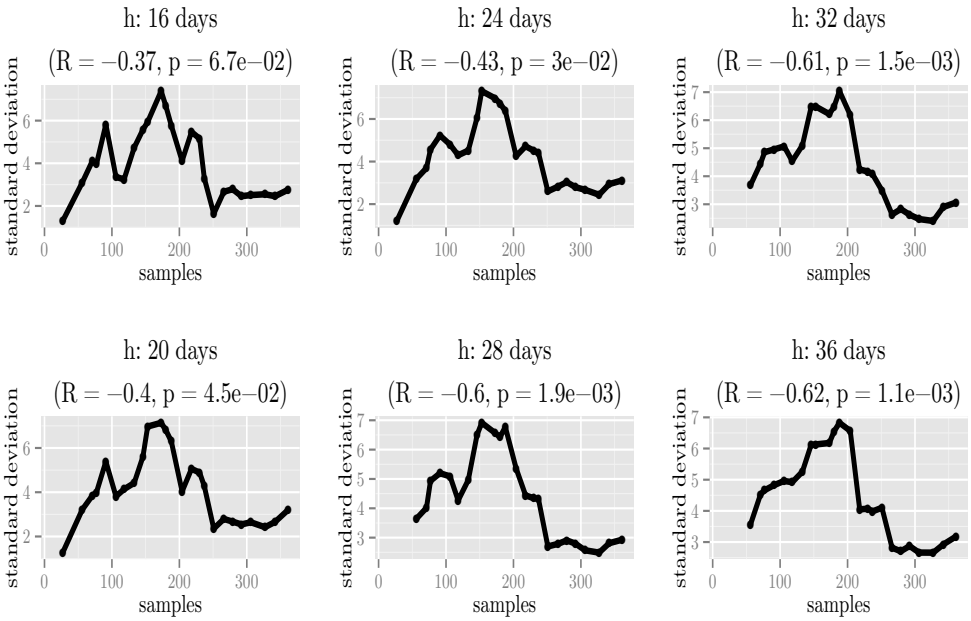


Figure A.21: Example highly correlated phenotypic variable cluster: 7 phenotype variables (560 OD AVG, 750 OD AVG, DW g/L, Chloro1 450/685 nm AVG, Green1 430/685 nm AVG, KG, Cyano1 383/685 nm AVG) that mainly consist of various fluorescence levels and dry weight measures. Normalized variables, together with their first normalized principle component (dashed red), explaining 87.3% of the variance of the cluster.



(a) Productivity mean for h:16-36 days



(b) Productivity standard deviation for h:16-36 days

Figure A.22: Productivity statistics trends for various h (half window) sizes changing from 16 to 36 days.

Appendix B

Extended Figures for Chapter 3

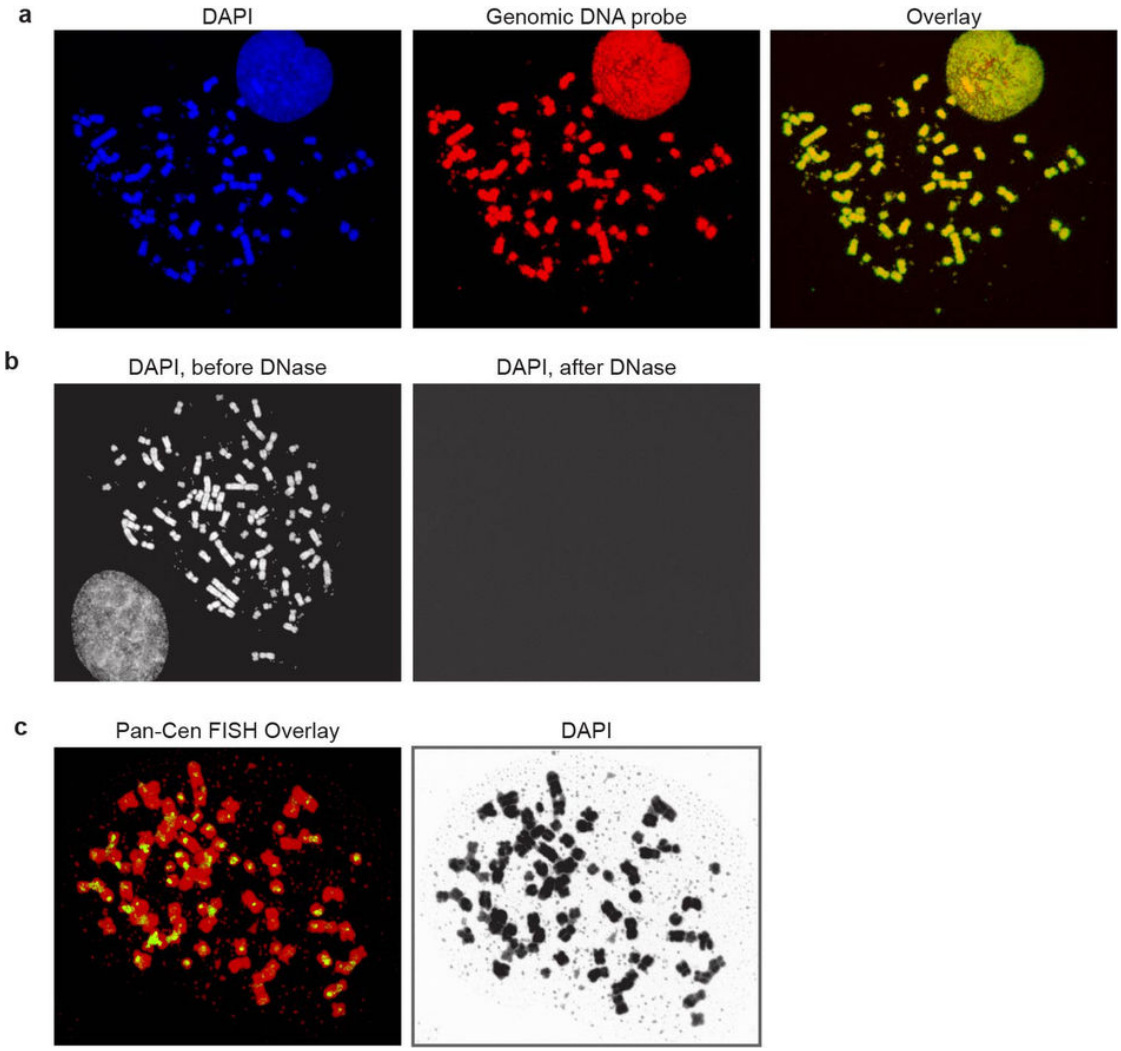


Figure B.1: Full select metaphase spreads: Full metaphase spreads corresponding to the partial metaphase spreads shown in Fig. 4.1a, Images corresponding to Fig. 4.1b. b, Images corresponding to Fig. 4.1c. c, Images corresponding to Fig. 4.1d.

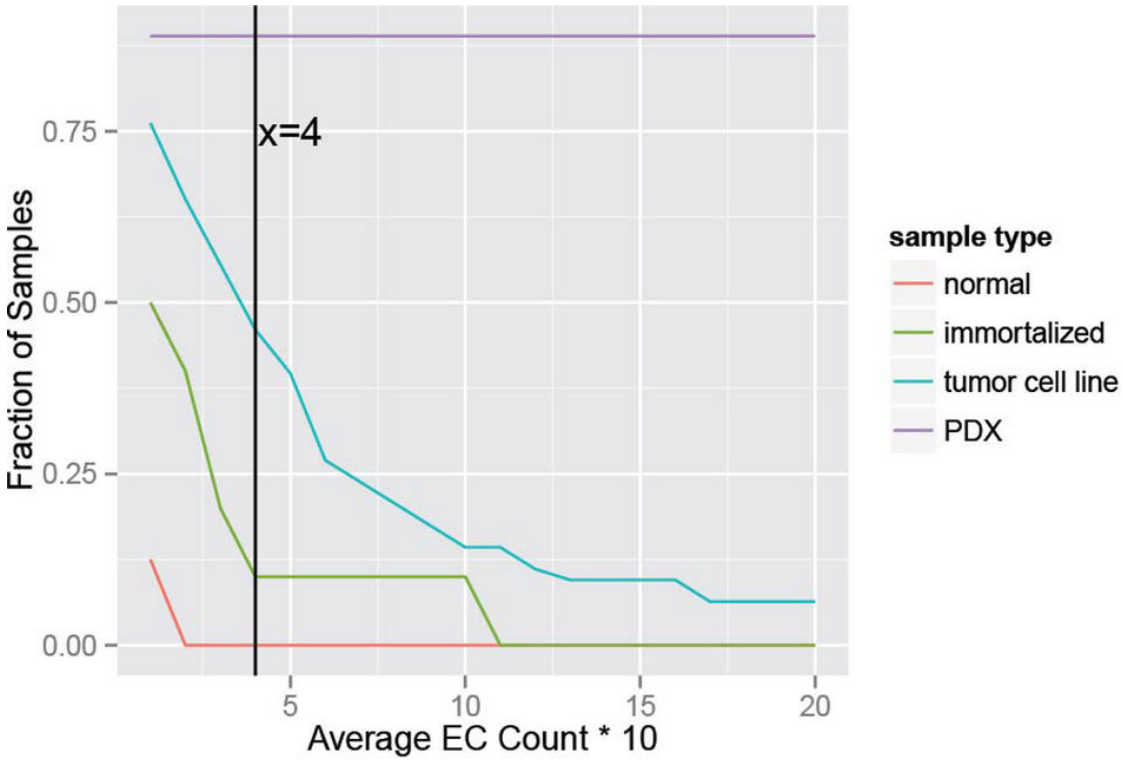


Figure B.2: Alternative analysis of ecDNA presence according to varying criteria, stratified by sample type: Samples with a minimum number of ecDNA elements per 10 cells in metaphase in average shown in x axis are classified ecDNA positive, and their fraction is displayed on the y axis. The vertical line at $x = 4$ shows that for a minimum of 4 ecDNA elements per 10 cells in metaphase on average, 0% of normal, 10% of immortalized, 46% of tumour cell line and 89% of PDX samples are classified as ecDNA positive.

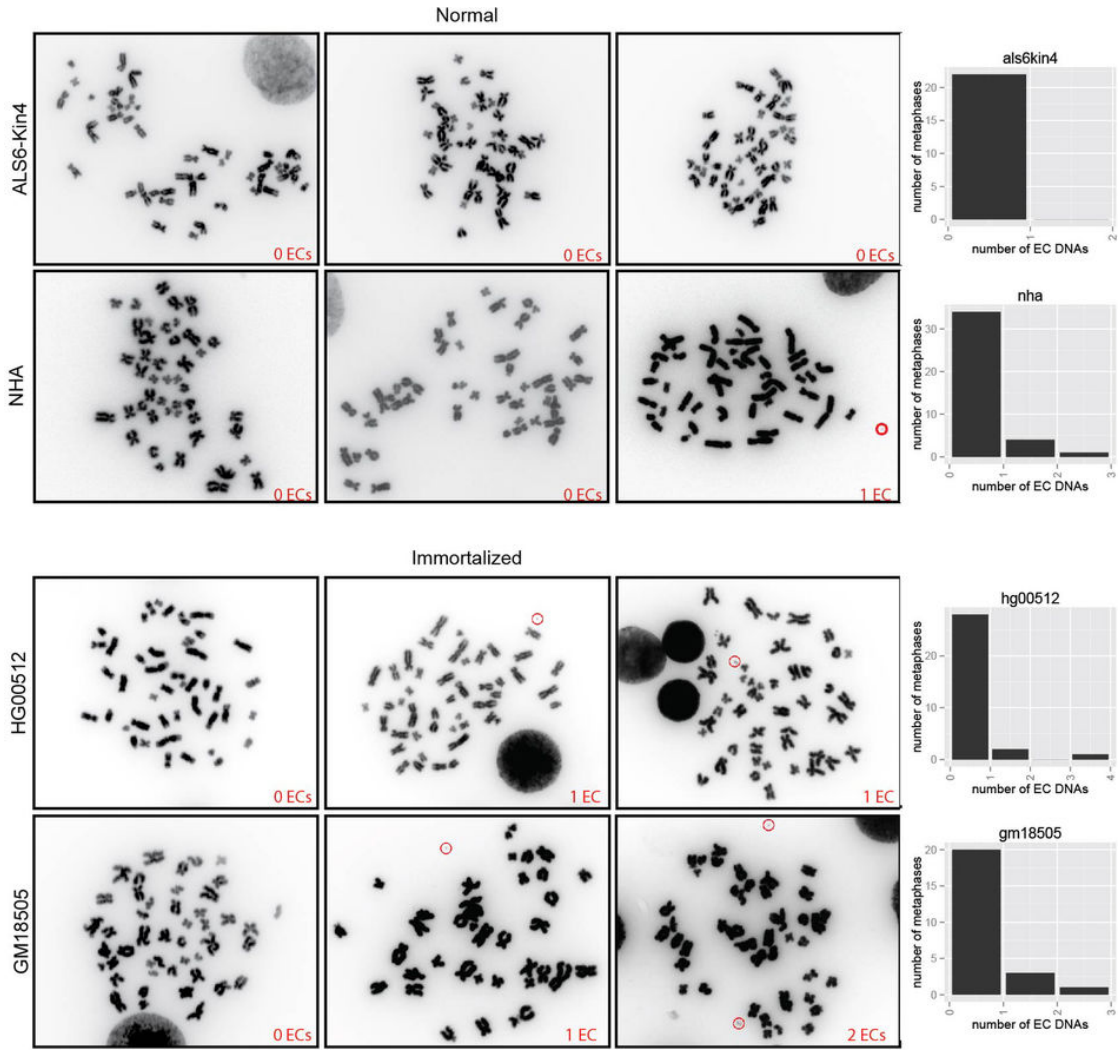


Figure B.3: ecDNA counts in normal and immortalized cells

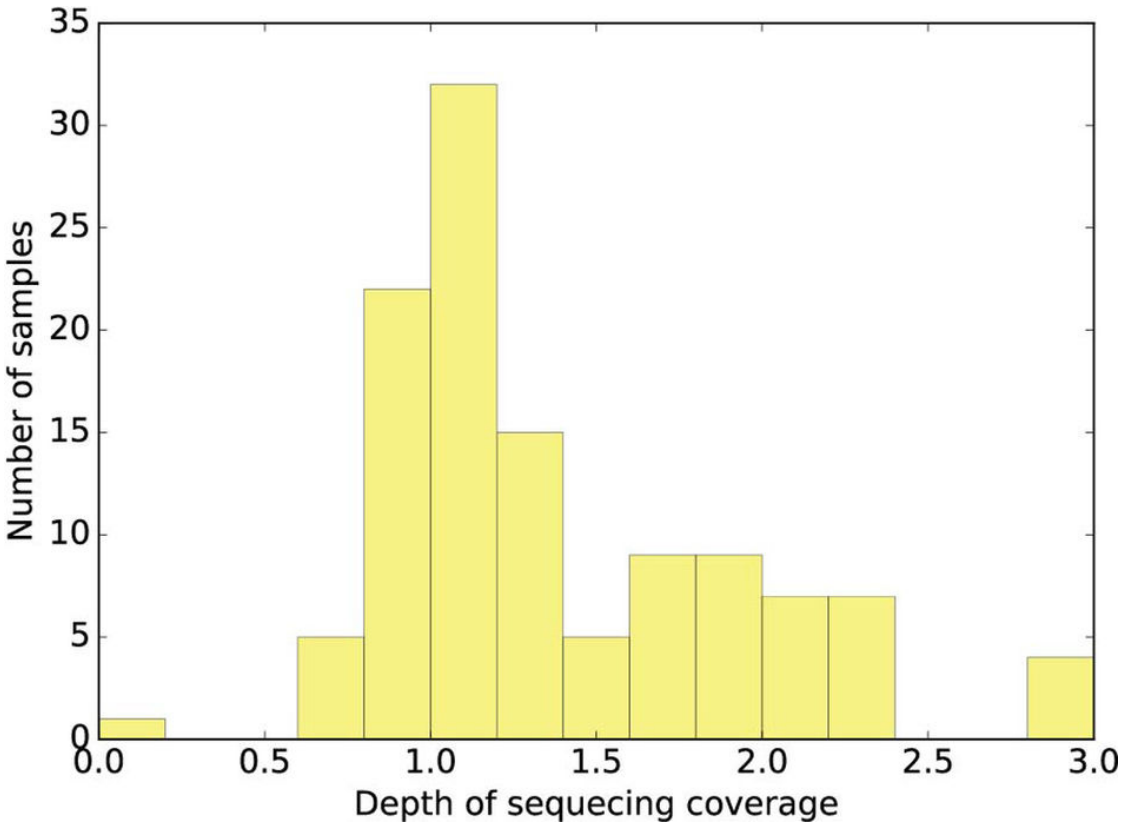


Figure B.4: Histogram of depth of coverage for next-generation sequencing of tumour samples: We sequenced 117 tumour samples including 63 cell lines, 19 neurospheres (PDX) and 35 cancer tissues with coverage ranging from $0.6X$to $3.89X$(excluding one sample with $0.06X$ coverage) with median coverage of $1.19X$.

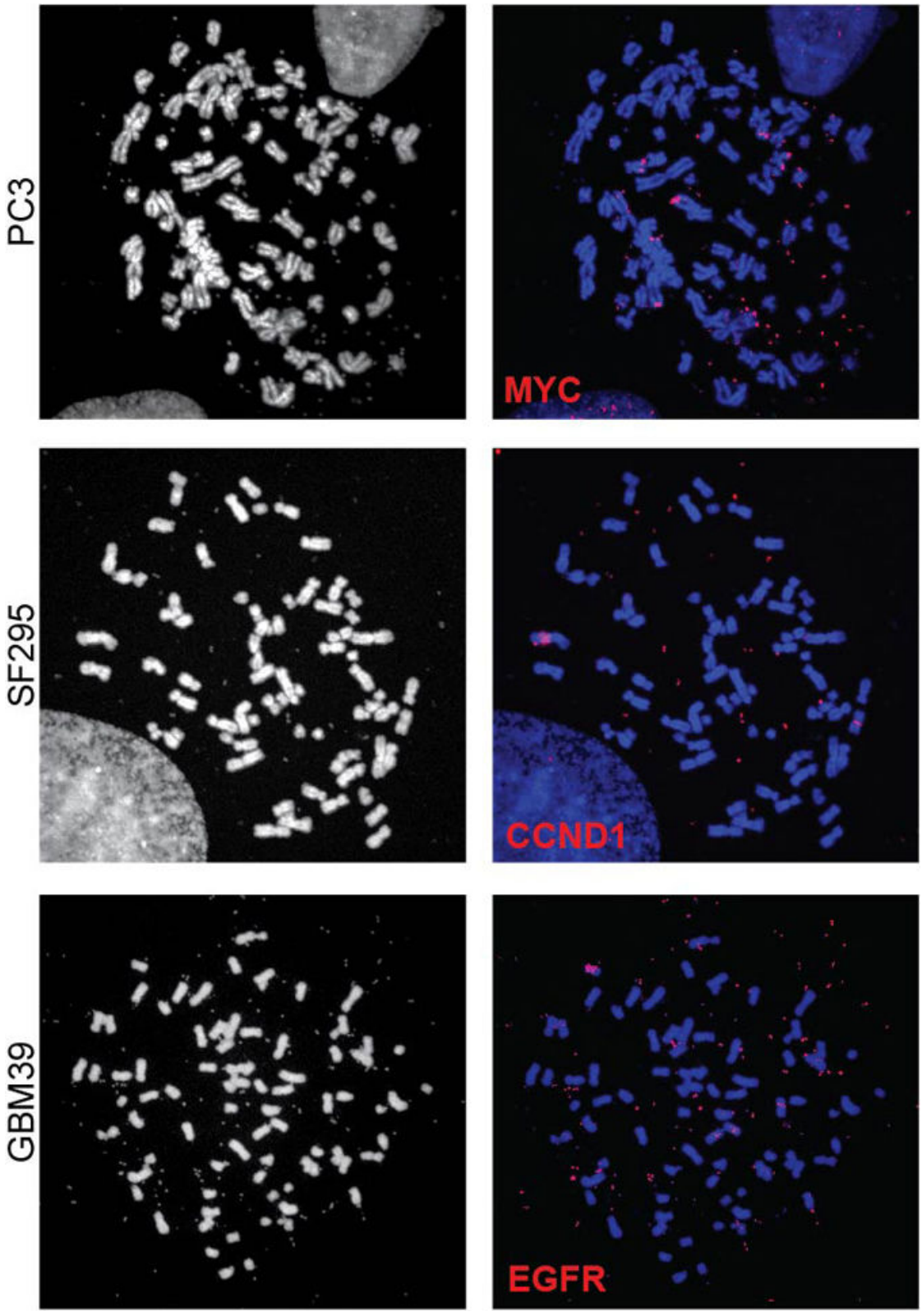


Figure B.5: Full select metaphase spreads: Full metaphase spreads corresponding to the partial metaphase spreads shown in Fig. 4.3c

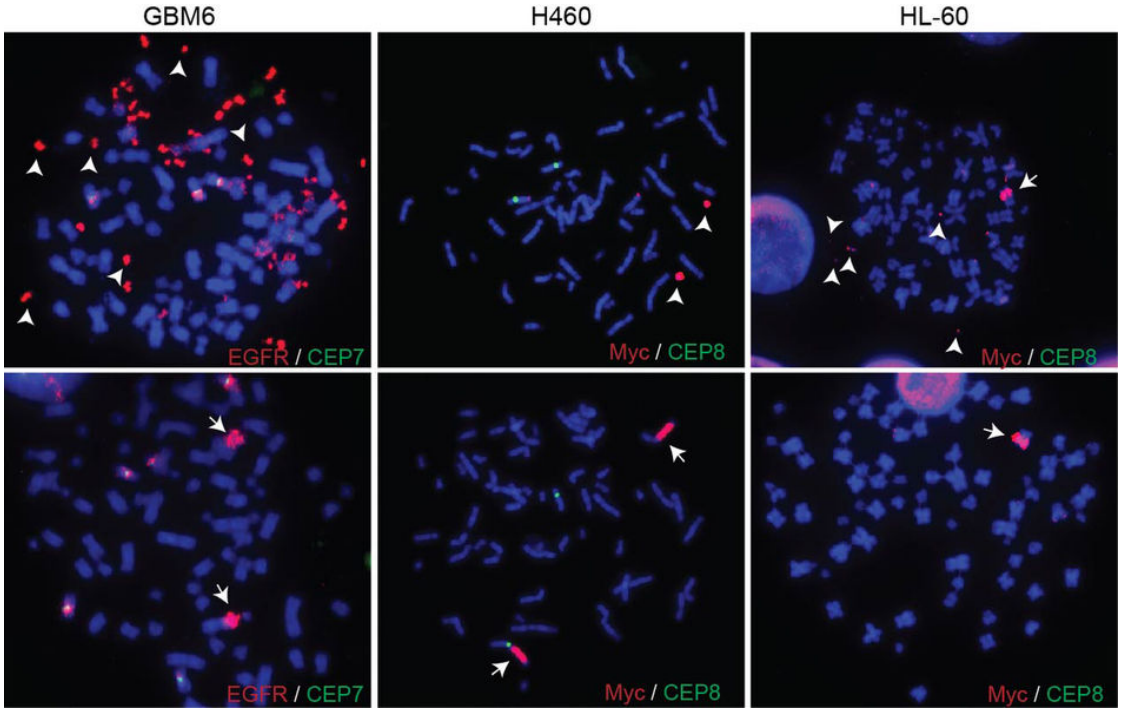


Figure B.6: FISH images displaying both ecDNA elements and HSRs in cells from the same sample

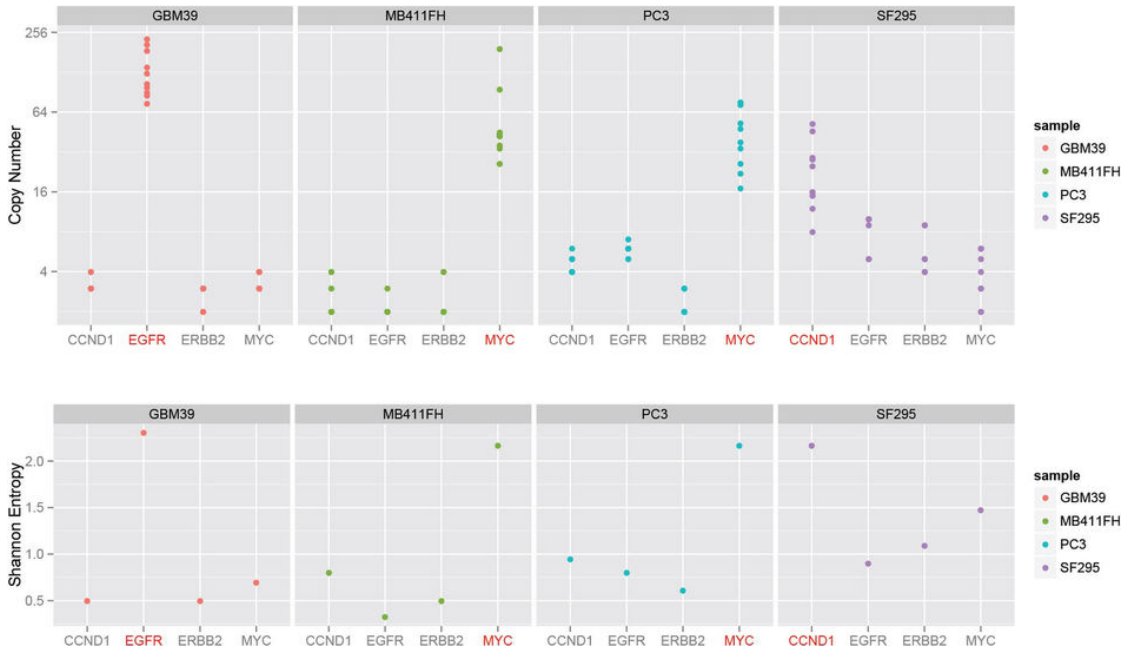


Figure B.7: Copy-number amplification and diversity due to ecDNA: To test how much of the copy-number amplification and diversity could be attributed to ecDNA, we chose FISH probes that bind to four of the most commonly amplified oncogenes in our sample set, EGFR, MYC, CCND1 or ERBB2, and quantified the cell-to-cell variability in their DNA copy number in metaphase spreads, from four tumour cell lines: GBM39, MB411FH, SF295 and PC3 cancer cells. For each cell line, only the target oncogene marked in red is known to be amplified on ecDNA (EGFR in GBM39; MYC in MB411FH and PC3, and CCND1 in SF295). The other 3 genes reside on chromosomal loci. The target oncogene shows consistently higher copy numbers (top) and diversity (bottom)

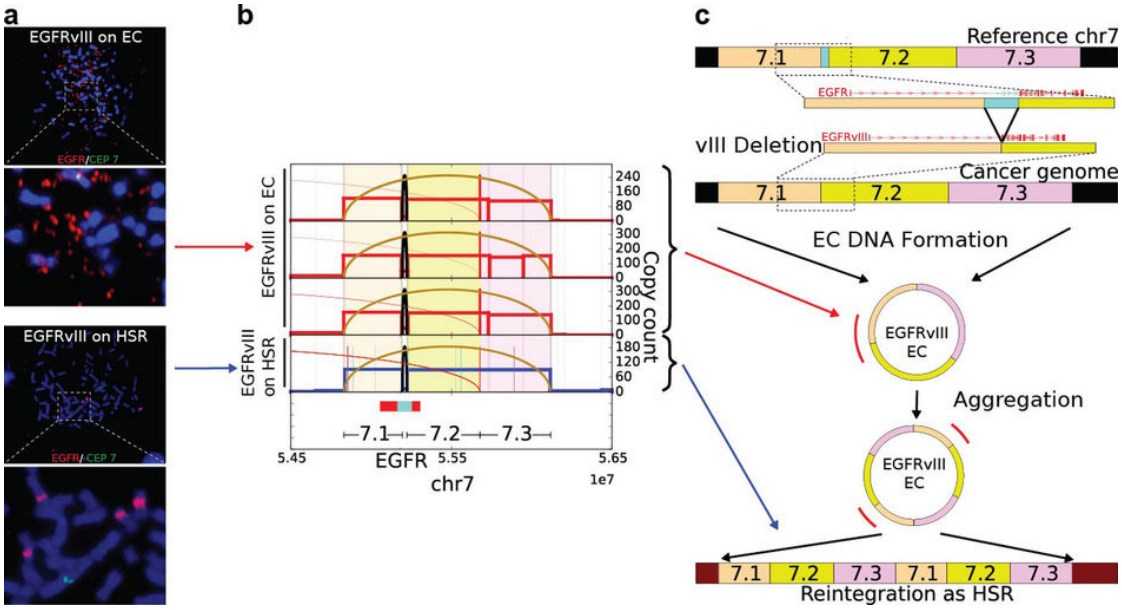


Figure B.8: Fine structure analysis of EGFRvIII amplification in extrachromosomal or chromosomal DNA in GBM39 cells

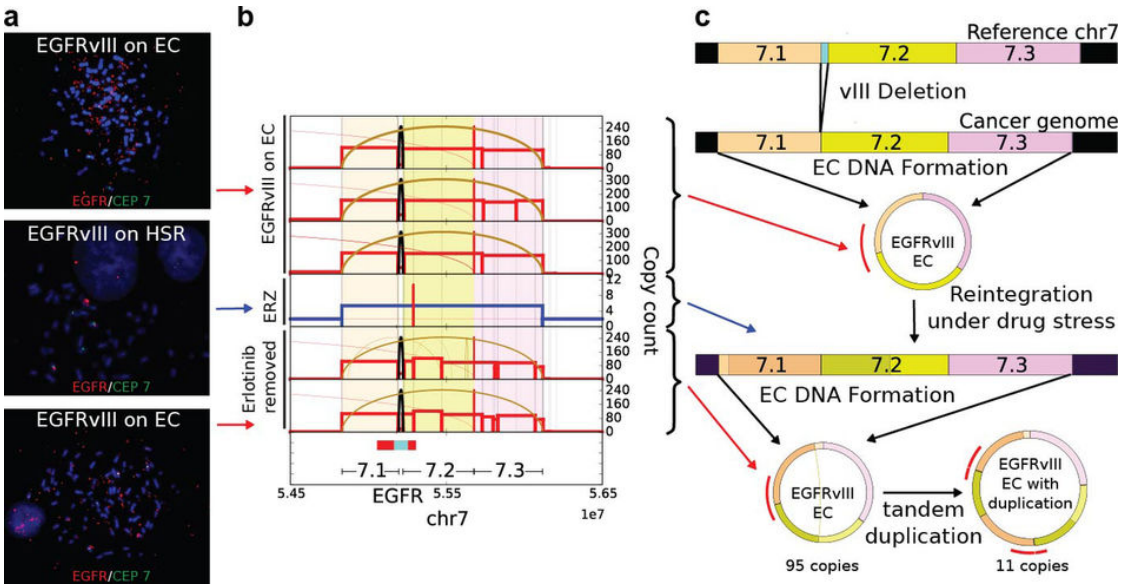


Figure B.9: Fine structure analysis of EGFRvIII amplification in extrachromosomal or chromosomal DNA in naive GBM39 cells and in response to erlotinib treatment and drug withdrawal

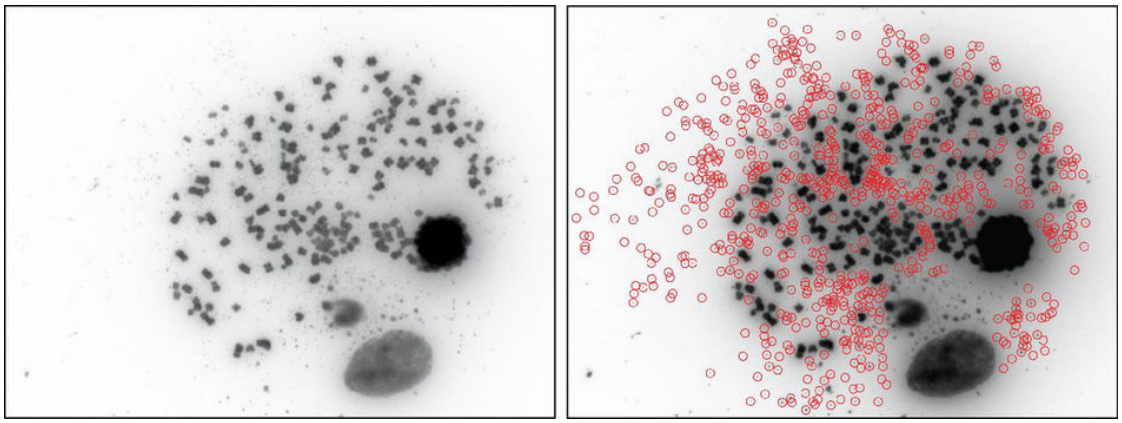


Figure B.10: A GBM cell in metaphase with large ecDNA counts (¿600), as determined by manual counting and ECdetect

Appendix C

ECdetect: Software for detection of extrachromosomal DNA from DAPI staining metaphase images

C.1 Introduction

The DAPI staining metaphase image extrachromosomal DNA (ECDNA) detection software provides a conservative estimation to the number of ECDNA in DAPI staining metaphase images. The software performs a pre-segmentation of the image in order to distinguish chromosomal and non-chromosomal structures, and computes an ECDNA search region of interest (ROI). The designated ROI is displayed on a user interface for the investigator to modify via masking and unmasking desired regions on the image, to correct for potential inaccurate segmentation and/or exclude debris from the ROI. The modifications made on the ROI are saved once verified, and are available for future usage. The output of the software includes the original images with ECDNA detections overlayed, the

count of ECDNA found, and their coordinates in the image. ECdetect does not require a pan-centromeric probe, and works on DAPI staining metaphase images only, therefore any detected ECDNA is assumed to not contain a centromere.

C.2 Software

Input

The ECDNA detection software uses Tagged Image File Format (.tiff) DAPI staining metaphase images. In this project we used 2572 images, after checking for duplicates, each at resolution 1392x1040. The investigator needs to provide the parent folder containing all imaging data as input and no other parameter will be required. The software will recursively process every tiff image under the parent folder.

Image pre-segmentation

The software applies an initial coarse adaptive thresholding [Mot15, BR07] to detect the major components in the image, with a window size of $150x150$ pixels, and $T = 10\%$. After filling the closed structures, components breaching 3000 pixels and 80% of solidity (the ratio of the area of the component to the area of its convex hull) are masked as non-chromosomal regions in order to remove the intact nuclei regions from subsequent analysis. Small components are also discarded, and the remaining image is accepted as the binary chromosomal image (BCI). The weakly connected components of the BCI are computed to find the separate chromosomal regions. The weakly connected components breaching a cumulative pixel count of 5000 are considered as candidate search regions, and their convex hull with a dilation of 100 pixels are added into the ECDNA search region of interest (ROI).

ROI verification

The software provides a user interface as shown in Figure C.1, where the original DAPI image is displayed next to its segmentation result, alongside an overview image.

We manually masked any non-chromosomal region that the software failed to discard during the pre-segmentation as shown in Figure C.2. Similarly, we also unmasked any region that the software mistakenly discarded as non-chromosomal region. The segmentation results are displayed in three colors: teal (chromosomal region qualified to be inside of the search region), dark blue (non-chromosomal/masked region), and green (chromosomal or small components not qualified to be inside of the search region). The color orange shows the current ECDNA search ROI. At the end of every masking/un-masking, the ECDNA search ROI is recomputed based on the newly generated BCI and displayed.

ECDNA detection

Figure C.3 shows the steps of ECDNA detection. After the verification of the ECDNA search ROI (Figure C.3a), the software applies a 2-D Gaussian smoothing to the image with standard deviation of 0.5, performs a second finer adaptive thresholding, with a window size of $20x20$ pixels and $T = 7\%$, and fills any closed structures. Components that are greater than 75 pixels are designated as non-ECDNA structures and their 15-pixel neighborhood is removed from the ECDNA search ROI, in order not to mistakenly call chromosomal extensions or other near intact nuclei structures as ECDNA (Figure C.3b). Any component detected with a size less than or equal to 75 and greater than or equal to 3 pixels inside the final search ROI is returned as ECDNA (Figure C.3c).

Output

The detected ECDNA elements are shown in the original image with overlayed red circles, as well as their coordinates in a separate file for every image. The total ECDNA count per image is also recorded.

Manual ECDNA marking

For ECDNA detection evaluation purposes, we allowed the investigator to manually select the ECDNA structures while being able to have access to the verified ECDNA search region (including the chromosome region neighborhood) and segmentation results, alongside zooming, if desired. Figure C.4 shows an example set of marked ECDNA at a specified zooming level.

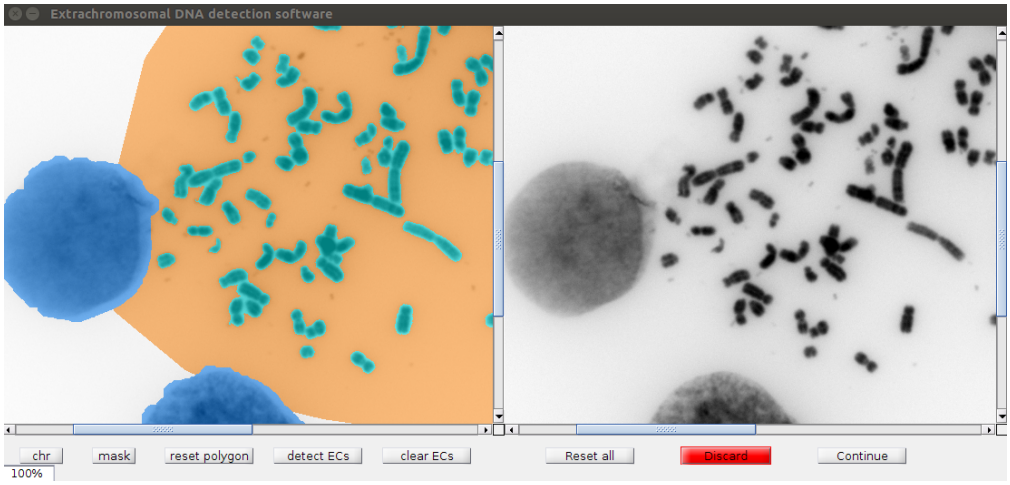
Comparison of software vs. visual inspection

The ECDNA coordinates detected by the software and selected by manual marking are compared and they are accepted to match if the distance between them is no more than 7 pixels. A sample comparison result is shown in Figure C.5. The green circles show the software detected ECDNA coordinates that agree with manually marked ECDNA, blue circles show manually marked ECDNA that the software missed, and red circles show software detected ECDNA that were not manually marked. Notice that a majority of blue circles appear in the immediate neighborhood of chromosomal structures, which we deliberately removed from the ECDNA search ROI. The red circles appear to have faint pixel intensities, which the visual inspection may have missed or discarded.

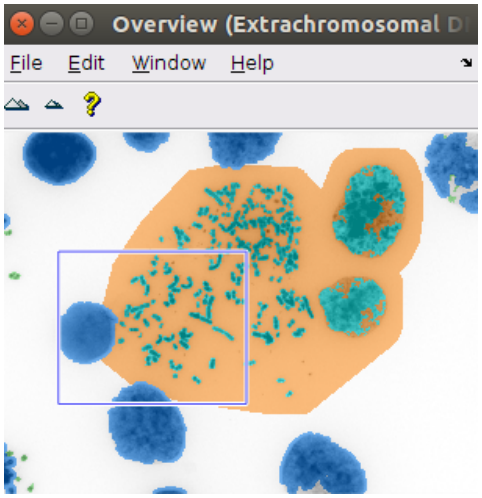
C.3 Results

We arbitrarily chose 28 images, in which we could confidently mark the ECDNA, while also aiming for a large range of ECDNA count across images, from various different tumor cell lines for purposes of robustness. We evaluated the performance of the ECDNA detection software by comparing it with manual ECDNA marking on the aforementioned 28 DAPI metaphase images from various tumor cell lines with varying count of ECDNAs.

Out of 406 detected ECDNA, 392 of them (97%) agreed with manually marked ECDNAs, however among the 737 total manually marked ECDNAs, the software missed 345 of them, resulting in a under-estimation by 53%. We would like to emphasize, however, that it was by design to discard the regions at the immediate neighborhood of non-ECDNA structures, e.g. chromosomal regions, from the ECDNA search ROI and undercall ECDNAs in order not to accept any questionable structure as extrachromosomal DNA. Indeed, 88% of the ECDNAs missed by the software compared to manual marking resides in the aforementioned discarded region. The software provides a conservative estimate of the total ECDNA signal; it achieves high precision at the expense of sensitivity compared to visual inspection, which may also have imperfections. Figure 1F shows the high correlation (Pearson; $r = 0.98$, $P < 2.2 \times 10^{-16}$) achieved between the ECDNA counts detected by the software and manual marking, suggesting a balanced undercalling of ECDNAs accross images, and a reliable estimation for correlative studies.

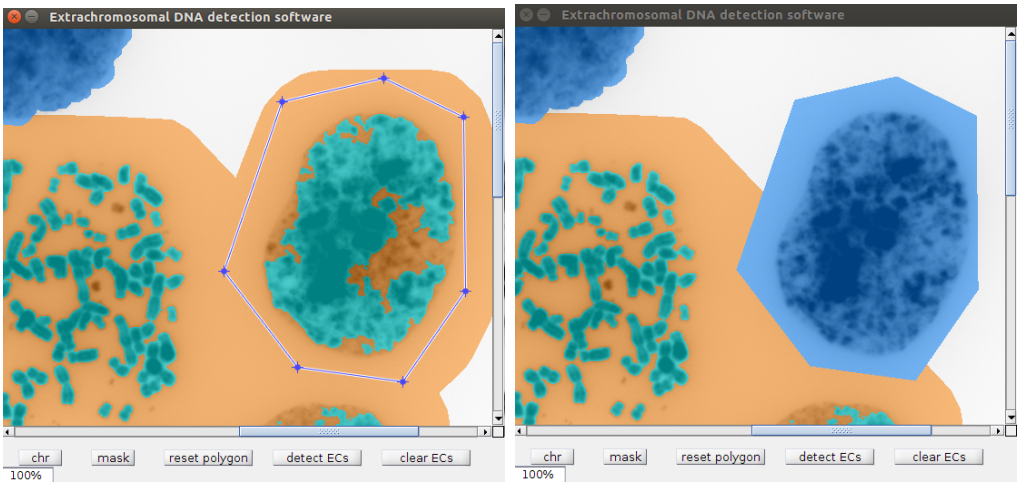


(a) Pre-segmented and original DAPI images



(b) Overview of pre-segmentation

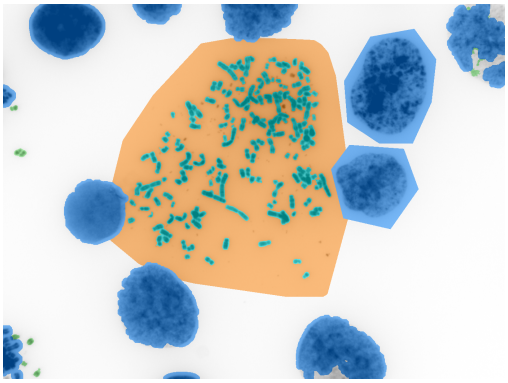
Figure C.1: User interface for EC DNA search ROI verification



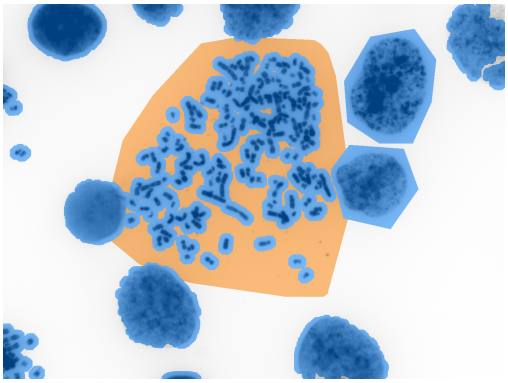
(a) Selection the undesired region

(b) Masking and removing from the EC search ROI

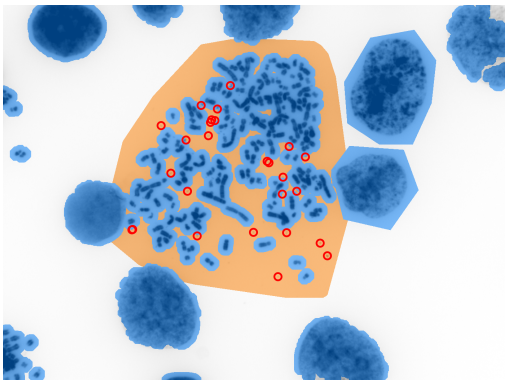
Figure C.2: Non-chromosomal region masking



(a) Step 1: Verified EC DNA search ROI.



(b) Step 2: 15-pixel neighborhood of any larger than EC DNA structure is removed.



(c) Step 3: EC DNA detection on final search ROI.

Figure C.3: EC DNA detection steps.

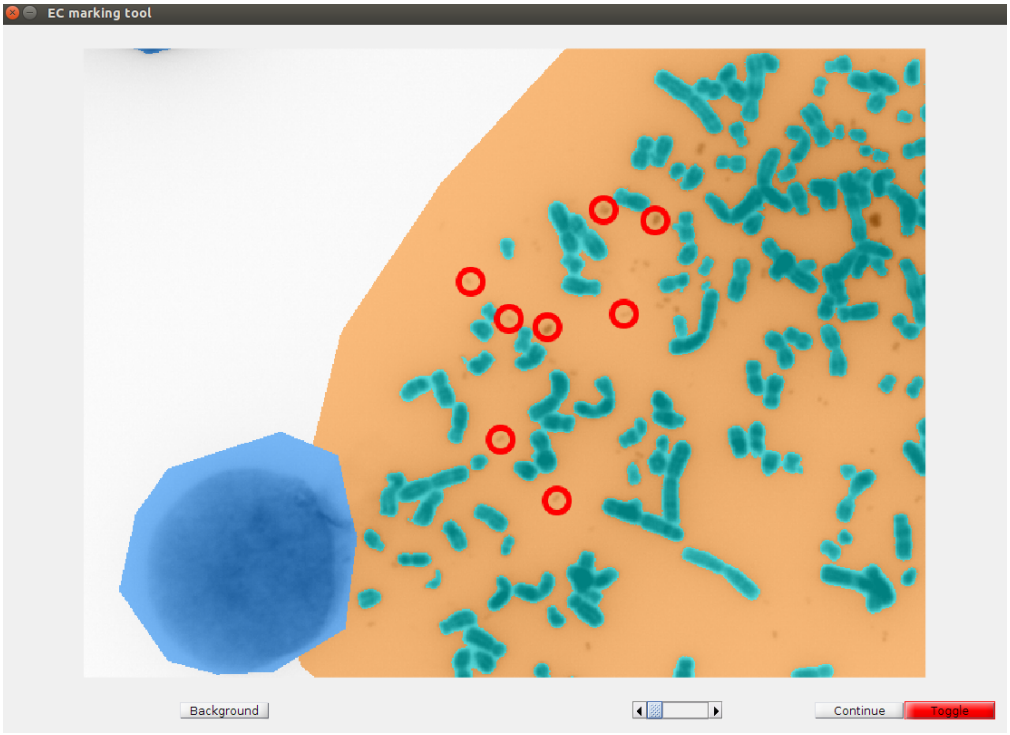


Figure C.4: Manual marking of EC DNA

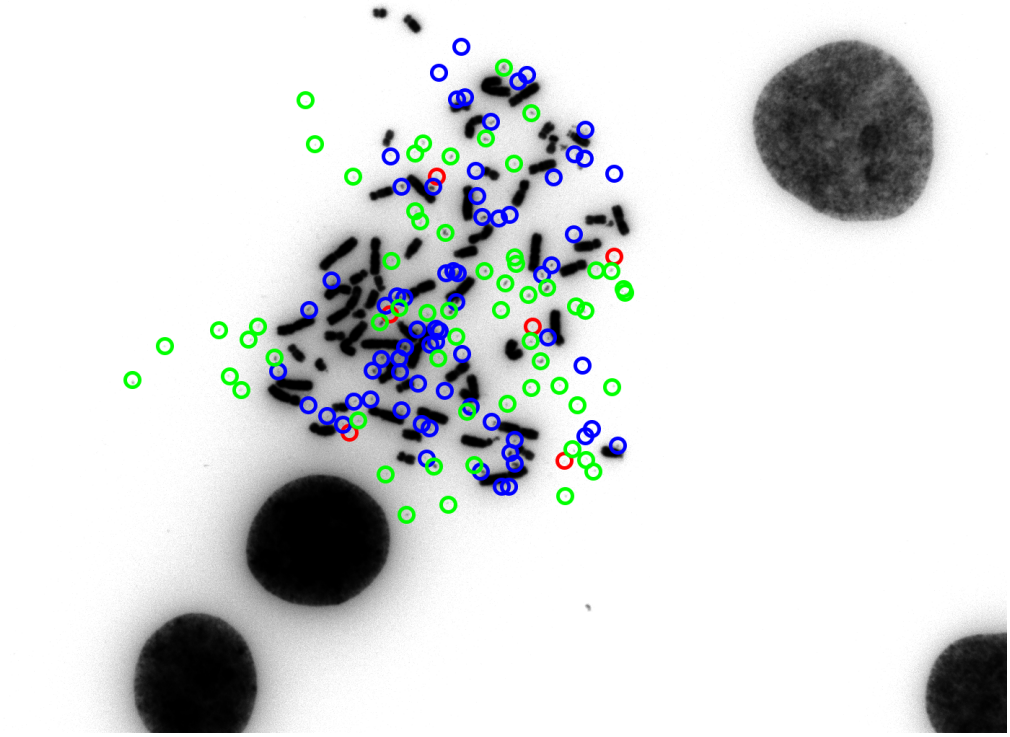


Figure C.5: ECdetect evaluation via manual marking

Bibliography

[ACR+14] V. Almendro, Y. K. Cheng, A. Randles, S. Itzkovitz, A. Marusyk, E. Ametller, X. Gonzalez-Farre, M. Munoz, H. G. Russnes, A. Helland, I. H. Rye, A. L. Borresen-Dale, R. Maruyama, A. van Oudenaarden, M. Dowsett, R. L. Jones, J. Reis-Filho, P. Gascon, M. Gonen, F. Michor, and K. Polyak. Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell Rep*, 6(3):514–527, Feb 2014.

[AGJ+16] Noemi Andor, Trevor A Graham, Marnix Jansen, Li C Xia, C Athena Aktipis, Claudia Petritsch, Hanlee P Ji, and Carlo C Maley. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature medicine*, 22(1):105–113, 2016.

[BAO+10] Ivana Bozic, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W Kinzler, Bert Vogelstein, and Martin A Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43):18545–18550, 2010.

[BDSY99] Amir Ben-Dor, Ron Shamir, and Zohar Yakhini. Clustering gene expression patterns. *Journal of computational biology*, 6(3-4):281–297, 1999.

[BHW+10] J. J. Bromenshenk, C. B. Henderson, C. H. Wick, M. F. Stanford, A. W. Zulich, R. E. Jabbour, S. V. Deshpande, P. E. McCubbin, R. A. Seccomb, P. M. Welch, T. Williams, D. R. Firth, E. Skowronski, M. M. Lehmann, S. L. Bilimoria, J. Gress, K. W. Wanner, and R. A. Cramer. Iridovirus and microsporidian linked to honey bee colony decline. *PLoS ONE*, 5(10):e13181, Oct 2010.

[BM13] Michael A Borowitzka and Navid Reza Moheimani. Open pond culture systems. In *Algae for Biofuels and Energy*, pages 133–152. Springer, 2013.

[BMT+17] Meghan C Burke, Yuri A Mirokhin, Dmitrii V Tchekhovskoi, Sanford P Markey, Jenny Heidbrink Thompson, Christopher Larkin, and Stephen E Stein. The hybrid search: A mass spectral library search method for discovery of modifications in proteomics. *Journal of Proteome Research*, 16(5):1924–1935, 2017.

[BR07] Derek Bradley and Gerhard Roth. Adaptive thresholding using the integral image. *Journal of Graphics Tools*, 12(2):13–21, 2007.

[BSH63] June L Biedler, Anthony W Schrecker, and Dorris J Hutchison. Selection of chromosomal variant in amethopterin-resistant sublines of leukemia l1210 with increased levels of dihydrofolate reductase 2. *Journal of the National Cancer Institute*, 31(3):575–601, 1963.

[BWA97] James D Bever, Kristi M Westover, and Janis Antonovics. Incorporating the soil community into plant population dynamics: the utility of the feedback approach. *Journal of Ecology*, pages 561–573, 1997.

[CAC+09] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, Jun 2009.

[Car96] Stephen R Carpenter. Microcosm experiments have limited relevance for community and ecosystem ecology. *Ecology*, 77(3):677–680, 1996.

[CBB99] Karl R Clauser, Peter Baker, and Alma L Burlingame. Role of accurate mass measurement (±10 ppm) in protein identification strategies employing ms or ms/ms and database searching. *Analytical chemistry*, 71(14):2871–2882, 1999.

[CCJ10] Anne Chao, Chun-Huo Chiu, and Lou Jost. Phylogenetic diversity measures based on hill numbers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1558):3599–3609, 2010.

[CCP+11] S. R. Carpenter, J. J. Cole, M. L. Pace, R. Batt, W. A. Brock, T. Cline, J. Coloso, J. R. Hodgson, J. F. Kitchell, D. A. Seekell, L. Smith, and B. Weidel. Early warnings of regime shifts: a whole-ecosystem experiment. *Science*, 332(6033):1079–1082, May 2011.

[CDG+88] SM Carroll, ML DeRose, P Gaudray, CM Moore, DR Needham-Vandevanter, DD Von Hoff, and GM Wahl. Double minute chromosomes can be produced from precursors derived from a chromosomal deletion. *Molecular and cellular biology*, 8(4):1525–1533, 1988.

[Cha93] Daniel Chaumont. Biotechnology of algal biomass production: a review of systems for outdoor mass culture. *Journal of Applied Phycology*, 5(6):593–604, 1993.

[CHH+09] Bradley J Cardinale, Helmut Hillebrand, WS Harpole, Kevin Gross, and Robert Ptacnik. Separating the influence of resource availabilityfrom resource imbalanceon productivity–diversity relationships. *Ecology Letters*, 12(6):475–487, 2009.

[Chi92] Sallie W Chisholm. Phytoplankton size. In *Primary productivity and biogeochemical cycles in the sea*, pages 213–237. Springer, 1992.

[CKS+10] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, 7(5):335–336, May 2010.

[CL14] Laura T Carney and Todd W Lane. Parasites in algae mass culture. *Frontiers in microbiology*, 5, 2014.

[CRW+14] Leah Cuthbertson, Geraint B Rogers, Alan W Walker, Anna Oliver, Tarana Hafiz, Lucas R Hoffman, Mary P Carroll, Julian Parkhill, Kenneth D Bruce, and Christopher J van der Gast. Time between collection and storage significantly influences bacterial sequence composition in sputum samples from cystic fibrosis respiratory infections. *Journal of clinical microbiology*, 52(8):3011–3016, 2014.

[CRWC10] Andres F Clarens, Eleazer P Resurreccion, Mark A White, and Lisa M Colosi. Environmental life cycle comparison of algae to other bioenergy feedstocks. *Environmental science & technology*, 44(5):1813–1819, 2010.

[CSD+06] Bradley J Cardinale, Diane S Srivastava, J Emmett Duffy, Justin P Wright, Amy L Downing, Mahesh Sankaran, and Claire Jouseau. Effects of biodiversity on the functioning of trophic groups and ecosystems. *Nature*, 443(7114):989–992, 2006.

[CSP+16] Sandip Chatterjee, Gregory S Stupp, Sung Kyu Robin Park, Jean-Christophe Ducom, John R Yates, Andrew I Su, and Dennis W Wolan. A comprehensive and scalable database search system for metaproteomics. *BMC genomics*, 17(1):642, 2016.

[DD07] S. Dray and A.B. Dufour. The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4):1–20, 2007.

[DHL+06] Todd Z DeSantis, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen. Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, 72(7):5069–5072, 2006.

[Duf02] J Emmett Duffy. Biodiversity and ecosystem function: the consumer connection. *Oikos*, 99(2):201–219, 2002.

[ECL+12] A. R. Erickson, B. L. Cantarel, R. Lamendella, Y. Darzi, E. F. Mongodin, C. Pan, M. Shah, J. Halfvarson, C. Tysk, B. Henrissat, J. Raes, N. C. Verberkmoes, C. M. Fraser, R. L. Hettich, and J. K. Jansson. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS ONE*, 7(11):e49138, 2012.

[Edg13] Robert C Edgar. Uparse: highly accurate otu sequences from microbial amplicon reads. *Nature methods*, 10(10):996–998, 2013.

[Edg15a] Robert C Edgar. UCHIME. Available at: http://drive5.com/uchime/uchime_download.html, 2015. Last Accessed: 01 April 2015.

[Edg15b] Robert C Edgar. UPARSE Pipeline. Available at: http://drive5.com/usearch/manual/uparse_pipeline.html, 2015. Last Accessed: 01 April 2015.

[EG10] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for mass spectrometry-based proteomics. *Proteome bioinformatics*, pages 55–71, 2010.

[EJH13] Jimmy K Eng, Tahmina A Jahan, and Michael R Hoopmann. Comet: An open-source ms/ms sequence database search tool. *Proteomics*, 13(1):22–24, 2013.

[EKL11] Kyle F Edwards, Christopher A Klausmeier, and Elena Litchman. Evidence for a three-way trade-off between nitrogen and phosphorus competitive abilities and cell size in phytoplankton. *Ecology*, 92(11):2085–2095, 2011.

[ELK13] Kyle F Edwards, Elena Litchman, and Christopher A Klausmeier. Functional traits explain phytoplankton community structure and seasonal dynamics in a marine ecosystem. *Ecology letters*, 16(1):56–63, 2013.

[EMD15] EMD Milipore. UFC903008 — Amicon Ultra-15 Centrifugal Filter Unit with Ultracel-30 membrane. Available at: http://www.millipore.com/catalogue/item/ufc903008, 2015. Last Accessed: 01 April 2015.

[EMY94] Jimmy K Eng, Ashley L McCormack, and John R Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.

[ESCT11] Jimmy K Eng, Brian C Searle, Karl R Clauser, and David L Tabb. A face in the crowd: recognizing peptides through database search. *Molecular & Cellular Proteomics*, 10(11):R111–009522, 2011.

[FML+11] Y. Fan, R. Mao, H. Lv, J. Xu, L. Yan, Y. Liu, M. Shi, G. Ji, Y. Yu, J. Bai, Y. Jin, and S. Fu. Frequency of double minute chromosomes and combined cytogenetic abnormalities and their characteristics. *J. Appl. Genet.*, 52(1):53–59, Feb 2011.

[FPS+12] D. E. Fouts, R. Pieper, S. Szpakowski, H. Pohl, S. Knoblach, M. J. Suh, S. T. Huang, I. Ljungberg, B. M. Sprague, S. K. Lucas, M. Torralba, K. E. Nelson, and S. L. Groah. Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. *J Transl Med*, 10:174, Aug 2012.

[Fre11] Robert P Freckleton. Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. *Behavioral Ecology and Sociobiology*, 65(1):91–101, 2011.

[GdSDNW+13] Alena S Gsell, Lisette N de Senerpont Domis, Suzanne MH Naus-Wiezer, Nico R Helmsing, Ellen Van Donk, and Bas W Ibelings. Spatiotemporal variation in the distribution of chytrid parasites in diatom host populations. *Freshwater Biology*, 58(3):523–537, 2013.

[GM12a] D Ryan Georgianna and Stephen P Mayfield. Exploiting diversity and synthetic biology for the production of algal biofuels. *Nature*, 488(7411):329–335, 2012.

[GM12b] Mel Greaves and Carlo C Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012.

[GMC+14] D. W. Garsed, O. J. Marshall, V. D. Corbin, A. Hsu, L. Di Stefano, J. Schroder, J. Li, Z. P. Feng, B. W. Kim, M. Kowarsky, B. Lansdell, R. Brookwell, O. Myklebost, L. Meza-Zepeda, A. J. Holloway, F. Pedeutour, K. H. Choo, M. A. Damore, A. J. Deans, A. T. Papenfuss, and D. M. Thomas. The architecture and evolution of cancer neochromosomes. *Cancer Cell*, 26(5):653–667, Nov 2014.

[GNLJ11] CE Grueber, S Nakagawa, RJ Laws, and IG Jamieson. Multimodel inference in ecology and evolution: challenges and solutions. *Journal of evolutionary biology*, 24(4):699–711, 2011.

[GSMK14] Jean-David Grattepanche, Luciana F Santoferrara, George B McManus, and Laura A Katz. Diversity of diversity: conceptual and methodological differences in biodiversity estimates of eukaryotic microbes as compared to bacteria. *Trends in microbiology*, 22(8):432–437, 2014.

[GVG12] Robert J Gillies, Daniel Verduzco, and Robert A Gatenby. Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. *Nature Reviews Cancer*, 12(7):487–493, 2012.

[HC04] Helmut Hillebrand and Bradley J Cardinale. Consumer effects decline with prey diversity. *Ecology Letters*, 7(3):192–201, 2004.

[Hil73] Mark O Hill. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973.

[HP13] Saad Haider and Ranadip Pal. Integrated analysis of transcriptomic and proteomic data. *Current genomics*, 14(2):91–110, 2013.

[HPCG13] Robert L Hettich, Chongle Pan, Karuna Chourey, and Richard J Giannone. Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities, 2013.

[IDBK+04] Bas W Ibelings, Arnout De Bruin, Maiko Kagami, Machteld Rijkeboer, Michaela Brehm, and Ellen Van Donk. Host parasite interactions between freshwater phytoplankton and chytrid fungi (chytridiomycota) 1. *Journal of Phycology*, 40(3):437–453, 2004.

[IHCD08] C Imirzalioglu, T Hain, T Chakraborty, and E Domann. Hidden pathogens uncovered: metagenomic analysis of urinary tract infections. *Andrologia*, 40(2):66–71, 2008.

[IHH04] Xabier Irigoien, Jef Huisman, and Roger P Harris. Global biodiversity patterns of marine phytoplankton and zooplankton. *Nature*, 429(6994):863–867, 2004.

[JGK+13] Pratik Jagtap, Jill Goslinga, Joel A Kooren, Thomas McGowan, Matthew S Wroblewski, Sean L Seymour, and Timothy J Griffin. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics*, 13(8):1352–1357, 2013.

[Jos06] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[KAS12] Elena Kazamia, David C Aldridge, and Alison G Smith. Synthetic ecology–a way forward for sustainable algal biofuel production? *Journal of Biotechnology*, 162(1):163–169, 2012.

[KC11] Giselle M Knudsen and Robert J Chalkley. The effect of using an inappropriate protein database for proteomic data analysis. *PloS one*, 6(6):e20873, 2011.

[KdBIVD07] Maiko Kagami, Arnout de Bruin, Bas W Ibelings, and Ellen Van Donk. Parasitic chytrids: their effects on phytoplankton communities and food-web dynamics. *Hydrobiologia*, 578(1):113–129, 2007.

[KLA+17] Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, and Alexey I Nesvizhskii. Msfragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, 14(5):513–520, 2017.

[KLH+14] Wanda Keersmaecker, Stef Lhermitte, Olivier Honnay, Jamshid Farifteh, Ben Somers, and Pol Coppin. How to measure ecosystem stability? an evaluation of the reliability of stability metrics based on remote sensing time series across the major global ecosystems. *Global change biology*, 20(7):2149–2161, 2014.

[KOW01] Teru Kanda, Michele Otter, and Geoffrey M Wahl. Mitotic segregation of viral and cellular acentric extrachromosomal molecules by chromosome tethering. *Journal of cell science*, 114(1):49–58, 2001.

[KP14] Sangtae Kim and Pavel A Pevzner. Universal database search tool for proteomics. *Nature communications*, 5:5277, 2014.

[KSF+02] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

[KYR+07] Angela D Kent, Anthony C Yannarell, James A Rusak, Eric W Triplett, and Katherine D McMahon. Synchrony in aquatic microbial community dynamics. *The ISME journal*, 1(1):38–47, 2007.

[LC12] Tom Leinster and Christina A Cobbold. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489, 2012.

[LD09] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[LD10] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[LdTPK+10] Elena Litchman, Paula de Tezanos Pinto, Christopher A Klausmeier, Mridul K Thomas, and Kohei Yoshiyama. Linking traits to species diversity and community structure in phytoplankton. *Hydrobiologia*, 653(1):15–28, 2010.

[Lee01] Yuan-Kun Lee. Microalgal mass culture systems and methods: their limitation and potential. *Journal of Applied Phycology*, 13(4):307–315, 2001.

[Lee12] Peter M Lee. *Bayesian statistics: an introduction*. John Wiley & Sons, 2012.

[LGP+14] X. Li, P. C. Galipeau, T. G. Paulson, C. A. Sanchez, J. Arnaudo, K. Liu, C. L. Sather, R. L. Kostadinov, R. D. Odze, M. K. Kuhner, C. C. Maley, S. G. Self, T. L. Vaughan, P. L. Blount, and B. J. Reid. Temporal and spatial evolution of somatic chromosomal alterations: a case-cohort study of Barrett's esophagus. *Cancer Prev Res (Phila)*, 7(1):114–127, Jan 2014.

[LHK83] EP Lincoln, TW Hall, and Ben Koopman. Zooplankton control in mass algal cultures. *Aquaculture*, 32(3):331–337, 1983.

[Lif15] Life Technologies. Purelink Pro 96 PCR Purification Kit. Available at: http://products.invitrogen.com/ivgn/product/K310096A, 2015. Last Accessed: 01 April 2015.

[LK08] Elena Litchman and Christopher A Klausmeier. Trait-based community ecology of phytoplankton. *Annual Review of Ecology, Evolution, and Systematics*, pages 615–639, 2008.

[LKSF07] Elena Litchman, Christopher A Klausmeier, Oscar M Schofield, and Paul G Falkowski. The role of functional traits and trade-offs in structuring phytoplankton communities: scaling from cellular to ecosystem level. *Ecology letters*, 10(12):1170–1181, 2007.

[LKVAZ05] Eva S Lindström, Miranda P Kamst-Van Agterveld, and Gabriel Zwart. Distribution of typical freshwater bacterial groups is associated with ph, temperature, and lake water retention time. *Applied and environmental microbiology*, 71(12):8201–8206, 2005.

[LLB+01] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W.

Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, and J. Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[LLS+13] Peter M Letcher, Salvador Lopez, Robert Schmieder, Philip A Lee, Craig Behnke, Martha J Powell, and Robert C McBride. Characterization of amoeboaphelidium protococcarum, an algal parasite new to the cryptomycota isolated from an outdoor algal pond used for the production of biofuel. *PloS one*, 8(2):e56232, 2013.

[MAP12] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334, 2012.

[MBB+12] Robert C McBride, Craig A Behnke, Kyle M Botsch, Nicole A Heaps, and Christopher Del Meenach. Use of fungicides in liquid systems, October 12 2012. US Patent App. 14/351,540.

[MBB+13] Robert C Mcbride, Craig A Behnke, Kyle M Botsch, Nicole A Heaps, and Christopher Del Meenach. Use of fungicides in liquid systems, 2013. US Patent 2,013,056,166.

[MHCM11] Christopher A Miller, Oliver Hampton, Cristian Coarfa, and Aleksandar Milosavljevic. Readdepth: a parallel r package for detecting

copy number alterations from short sequencing reads. *PloS one*, 6(1):e16327, 2011.

[MJM16] F. Mitelman, B. Johansson, and F. Mertens. Mitelman database of chromosome aberrations and gene fusions in cancer. Available at: http://cgap.nci.nih.gov/Chromosomes/Mitelman, 2016.

[Mo 15] Mo Bio Laboratories Inc. Powerlyzer Powersoil DNA Isolation Kit. Available at: http://www.mobio.com/soil-dna-isolation/powerlyzer-powersoil-dna-isolation-kit.html, 2015. Last Accessed: 01 April 2015.

[Mot15] J. Bradley Motl. local image thresholding. Available at: https://www.mathworks.com/matlabcentral/fileexchange/40854, 2015.

[MS15] Nicholas McGranahan and Charles Swanton. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell*, 27(1):15–26, 2015.

[MW16] Sweta Mishra and Johnathan R Whetstine. Different facets of copy number changes: permanent, transient, and adaptive. *Molecular and cellular biology*, 36(7):1050–1063, 2016.

[NCB15a] NCBI. Available at: http://www.ncbi.nlm.nih.gov/, 2015. Last Accessed: 01 April 2015.

[NCB15b] NCBI. Available at: http://www.ncbi.nlm.nih.gov/genomes/16S/help.html#query, 2015. Last Accessed: 12 July 2015.

[NGM+14] D. A. Nathanson, B. Gini, J. Mottahedeh, K. Visnyei, T. Koga, G. Gomez, A. Eskin, K. Hwang, J. Wang, K. Masui, A. Paucar, H. Yang, M. Ohashi, S. Zhu, J. Wykosky, R. Reed, S. F. Nelson, T. F. Cloughesy, C. D. James, P. N. Rao, H. I. Kornblum, J. R. Heath, W. K. Cavenee, F. B. Furnari, and P. S. Mischel. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. *Science*, 343(6166):72–76, Jan 2014.

[NH15a] Minita Shah Nils Homer, Michael Lyons. TMAP Technical Note. Available at: http://mendel.iontorrent.com/ion-docs/Technical-Note---TMAP-Alignment_9012907.html, 2015. Last Accessed: 01 April 2015.

[NH15b] Minita Shah Nils Homer, Michael Lyons. Torrent Mapping Alignment Program. Available at: https://github.com/iontorrent/TMAP, 2015. Last Accessed: 01 April 2015.

[NJE+11] Ryan J Newton, Stuart E Jones, Alexander Eiler, Katherine D McMahon, and Stefan Bertilsson. A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews*, 75(1):14–49, 2011.

[Now76] Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[NSG+14] S. Nikolaev, F. Santoni, M. Garieri, P. Makrythanasis, E. Falconnet, M. Guipponi, A. Vannier, I. Radovanovic, F. Bena, F. Forestier, K. Schaller, V. Dutoit, V. Clement-Schatlo, P. Y. Dietrich, and S. E. Antonarakis. Extrachromosomal driver mutations in glioblastoma and low-grade glioma. *Nat Commun*, 5:5690, Dec 2014.

[OBK+13] Jari Oksanen, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, and Helene Wagner. *vegan: Community Ecology Package*, 2013. R package version 2.0-10.

[PALKX14] Andrea Porras-Alfaro, Kuan-Liang Liu, Cheryl R Kuske, and Gary Xie. From genus to phylum: large-subunit and internal transcribed spacer rrna operon regions show similar classification accuracies influenced by database composition. *Applied and environmental microbiology*, 80(3):829–840, 2014.

[PBW+13] Joonhong Park, Gyu Seok Baek, Sung-Geun Woo, Jangho Lee, Jihoon Yang, and Juyoun Lee. Luteolibacter yonseiensis sp. nov., isolated from activated sludge using algal metabolites. *International journal of systematic and evolutionary microbiology*, 63(Pt 5):1891–1895, 2013.

[Pim84] Stuart L Pimm. The complexity and stability of ecosystems. *Nature*, 307(5949):321–326, 1984.

[PRM+14] Victor S Pylro, Luiz Fernando W Roesch, Daniel K Morais, Ian M Clark, Penny R Hirsch, and Marcos R Tótola. Data analysis for 16s microbial profiling from different benchtop sequencing platforms. *Journal of microbiological methods*, 107:30–37, 2014.

[PSA+08] Robert Ptacnik, Angelo G Solimini, Tom Andersen, Timo Tamminen, Pål Brettum, Liisa Lepistö, Eva Willén, and Seppo Rekolainen. Diversity predicts stability and resource use efficiency in natural phytoplankton communities. *Proceedings of the national academy of Sciences*, 105(13):5134–5138, 2008.

[PT16] Natalya N Pavlova and Craig B Thompson. The emerging hallmarks of cancer metabolism. *Cell metabolism*, 23(1):27–47, 2016.

[Pyl15] Victor Satler Pylro. Brazilian Microbiome Project. Available at: http://www.brmicrobiome.org, 2015. Last Accessed: 01 April 2015.

[R C14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[RCZB+09] Liliana Rodolfi, Graziella Chini Zittelli, Niccolò Bassi, Giulia Padovani, Natascia Biondi, Gimena Bonini, and Mario R Tredici. Microalgae for oil: Strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. *Biotechnology and bioengineering*, 102(1):100–112, 2009.

[Ric04] Amos Richmond. Biological principles of mass cultivation. *Handbook of microalgal culture: Biotechnology and applied phycology*, pages 125–177, 2004.

[RSW+08] Oliver Rinner, Jan Seebacher, Thomas Walzthoeni, Lukas Mueller, Martin Beck, Alexander Schmidt, Markus Mueller, and Ruedi Aebersold. Identification of cross-linked peptides from large sequence databases. *Nature methods*, 5(4):315–318, 2008.

[RVGS+05] Juliette N Rooney-Varga, Michael W Giewat, Mary C Savin, S Sood, M LeGresley, and JL Martin. Links between phytoplankton and bacterial community dynamics in a coastal marine environment. *Microbial Ecology*, 49(1):163–175, 2005.

[SAD+13] Jonathan B Shurin, Rachel L Abbott, Michael S Deal, Garfield T Kwan, Elena Litchman, Robert C McBride, Shovon Mandal, and Val H Smith. Industrial-strength ecology: trade-offs and opportunities in algal biofuel production. *Ecology letters*, 16(11):1393–1404, 2013.

[SC14] Val H Smith and Timothy Crews. Applying ecological principles of crop cultivation in large-scale algal biomass production. *Algal Research*, 4:23–34, 2014.

[Sch84] Robert T Schimke. Gene amplification in cultured animal cells. *Cell*, 37(3):705–713, 1984.

[SDBR98] John Sheehan, Terri Dunahay, John Benemann, and Paul Roessler. *A look back at the US Department of Energy's aquatic species program: biodiesel from algae*, volume 328. National Renewable Energy Laboratory Golden, 1998.

[SDGW89] George R Stark, Michelle Debatisse, Elena Giulotto, and Geoffrey M Wahl. Recent progress in understanding mechanisms of mammalian dna amplification. *Cell*, 57(6):901–908, 1989.

[SGHS12] Maria Stockenreiter, Anne-Kathrin Graber, Florian Haupt, and Herwig Stibor. The effect of species diversity on lipid production by micro-algal communities. *Journal of Applied Phycology*, 24(1):45–54, 2012.

[SHdJ+04] Maayke Stomp, Jef Huisman, Floris de Jongh, Annelies J Veraart, Daan Gerla, Machteld Rijkeboer, Bas W Ibelings, Ute IA Wollenzien, and Lucas J Stal. Adaptive divergence in pigment composition promotes phytoplankton biodiversity. *Nature*, 432(7013):104–107, 2004.

[Shi04] Hidenori Shimamatsu. Mass production of spirulina, an edible microalga. In *Asian Pacific Phycology in the 21st Century: Prospects and Challenges*, pages 39–44. Springer, 2004.

[SHM+11] Maayke Stomp, Jef Huisman, Gary G Mittelbach, Elena Litchman, and Christopher A Klausmeier. Large-scale biodiversity patterns in freshwater phytoplankton. *Ecology*, 92(11):2096–2107, 2011.

[SKAK78] Robert T Schimke, Randal J Kaufman, Fred W Alt, and Rodney F Kellems. Gene amplification and drug resistance in cultured murine cells. *Science*, 202(4372):1051–1055, 1978.

[SLG+10] C. T. Storlazzi, A. Lonoce, M. C. Guastadisegni, D. Trombetta, P. D'Addabbo, G. Daniele, A. L'Abbate, G. Macchia, C. Surace, K. Kok, R. Ullmann, S. Purgato, O. Palumbo, M. Carella, P. F. Ambros, and M. Rocchi. Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. *Genome Res.*, 20(9):1198–1206, Sep 2010.

[SMA14] Jonathan B Shurin, Shovon Mandal, and Rachel L Abbott. Trait diversity enhances yield in algal biofuel assemblages. *Journal of applied ecology*, 51(3):603–611, 2014.

[SSDB10] V. H. Smith, B. S. Sturm, F. J. Denoyelles, and S. A. Billings. The ecology of algal biodiesel production. *Trends Ecol. Evol. (Amst.)*, 25(5):301–309, May 2010.

[SSG+13] J Zachary Sanborn, Sofie R Salama, Mia Grifford, Cameron W Brennan, Tom Mikkelsen, Suresh Jhanwar, Sol Katzman, Lynda Chin, and David Haussler. Double minute chromosomes in glioblastoma multiforme are revealed by precise reconstruction of oncogenic amplicons. *Cancer research*, 73(19):6036–6045, 2013.

[Ste95] Stephen E Stein. Chemical substructure identification by mass spectral library searching. *Journal of the American Society for Mass Spectrometry*, 6(8):644–655, 1995.

[STHS+08] Peer M Schenk, Skye R Thomas-Hall, Evan Stephens, Ute C Marx, Jan H Mussgnug, Clemens Posten, Olaf Kruse, and Ben Hankamer. Second generation biofuels: high-efficiency microalgae for biodiesel production. *Bioenergy Research*, 1(1):20–43, 2008.

[SWR+09] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, 75(23):7537–7541, Dec 2009.

[Til81] David Tilman. Tests of resource competition theory using four species of lake michigan algae. *Ecology*, pages 802–815, 1981.

[TPD+13] Alessandro Tanca, Antonio Palomba, Massimo Deligios, Tiziana Cubeddu, Cristina Fraumene, Grazia Biosa, Daniela Pagnozzi, Maria Filippa Addis, and Sergio Uzzau. Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PloS one*, 8(12):e82981, 2013.

[Tre04] Mario R Tredici. Mass production of microalgae: photobioreactors. *Handbook of microalgal culture: Biotechnology and applied phycology*, 1:178–214, 2004.

[vGO11] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

[VHNVY+88] Daniel D Von Hoff, Donald R Needham-VanDevanter, Jennifer Yucel, Bradford E Windle, and Geoffrey M Wahl. Amplified human myc oncogenes localized to replicating submicroscopic circular dna molecules. *Proceedings of the National Academy of Sciences*, 85(13):4804–4808, 1988.

[VPV+13] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[Wal09] Emily Waltz. Biotech's green gold? *Nature Biotechnology*, 27(1):15–18, 2009.

[WDY+91] Brad Windle, Bruce W Draper, YX Yin, Stephen O'Gorman, and Geoffrey M Wahl. A central role for chromosome breakage in gene amplification, deletion formation, and amplicon integration. *Genes & development*, 5(2):160–174, 1991.

[WGTC07] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–5267, 2007.

[Wil15] Glynne D Williams. Two cases of urinary tract infection caused by propionimicrobium lymphophilum. *Journal of clinical microbiology*, 53(9):3077–3080, 2015.

[WKC+15] Jason Nicholas Woodhouse, Andrew Stephen Kinsela, Richard Nicholas Collins, Lee Chester Bowling, Gordon L Honeyman, Jon K Holliday, and Brett Anthony Neilan. Microbial communities reflect temporal changes in cyanobacterial composition in a shallow ephemeral freshwater lake. *The ISME journal*, 2015.

[WR09] Joseph Weissman and Guido Radaelli. Systems and methods for maintaining the dominance of nannochloropsis in an algae cultivation system, January 22 2009. US Patent App. 12/321,767.

[YC12] Lucy R Yates and Peter J Campbell. Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11):795–806, 2012.

[YMA+08] Jaewoo Yoon, Yoshihide Matsuo, Kyoko Adachi, Midori Nozawa, Satoru Matsuda, Hiroaki Kasai, and Akira Yokota. Description of persicirhabdus sediminis gen. nov., sp. nov., roseibacillus ishigakijimensis gen. nov., sp. nov., roseibacillus ponti sp. nov., roseibacillus persicicus sp. nov., luteolibacter pohnpeiensis gen. nov., sp. nov. and luteolibacter algae sp. nov., six marine members of the phylum verrucomicrobia, and emended descriptions of the class verrucomicrobiae, the order verrucomicrobiales and the family verrucomicrobiaceae. *International journal of systematic and evolutionary microbiology*, 58(4):998–1007, 2008.

[YSBG+15] Yanbao Yu, Patricia Sikorski, Cynthia Bowman-Gholston, Nicolas Cacciabeve, Karen E Nelson, and Rembert Pieper. Diagnosing inflammation and infection in the urinary system via proteomics. *Journal of translational medicine*, 13(1):111, 2015.

[YSL+10] H. Yao, J. Song, C. Liu, K. Luo, J. Han, Y. Li, X. Pang, H. Xu, Y. Zhu, P. Xiao, and S. Chen. Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS ONE*, 5(10), Oct 2010.

[YSS+17] Yanbao Yu, Patricia Sikorski, Madeline Smith, Cynthia Bowman-Gholston, Nicolas Cacciabeve, Karen E Nelson, and Rembert Pieper. Comprehensive metaproteomic analyses of urine in the presence and absence of neutrophil-associated inflammation in the urinary tract. *Theranostics*, 7(2):238, 2017.

[ZC14] Emily K Zimmerman and Bradley J Cardinale. Is the relationship between algal diversity and biomass in north american lakes consistent with biodiversity experiments? *Oikos*, 123(3):267–278, 2014.

[ZL97] CJ Zhu and YK Lee. Determination of biomass dry weight of marine microalgae. *Journal of Applied Phycology*, 9(2):189–194, 1997.

[ZNM+16] X. Zhang, Z. Ning, J. Mayne, J. I. Moore, J. Li, J. Butcher, S. A. Deeke, R. Chen, C. K. Chiang, M. Wen, D. Mack, A. Stintzi, and D. Figeys. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome*, 4(1):31, Jun 2016.

[ZR13] Oded Zmora and Amos Richmond. Microalgae for aquaculture: Microalgae production for aquaculture. *A Richmond, Q Hu, eds. Handbook of Microalgal Culture: Biotechnology and Applied Phycology, 2nd Edition. Oxford*, pages 365–379, 2013.

[ZSC+13] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C. Z. Zhsng, J. Wala, C. H. Mermel, C. Sougnez, S. B. Gabriel, B. Hernandez, H. Shen, P. W. Laird, G. Getz, M. Meyerson, and R. Beroukhim. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, 45(10):1134–1140, Oct 2013.