




**ADVERTIMENT.** L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  <https://creativecommons.org/licenses/?lang=ca>

**ADVERTENCIA.** El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <https://creativecommons.org/licenses/?lang=es>

**WARNING.** The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

# **Molecular complexity of the differential growth of freshwater diatom species along pH gradients**

Dissertation presented by  
**Xènia Rodríguez Miret**  
to obtain the degree of Doctor

Under the supervision of

Dr. Jordi Catalan Aguilà  
CSIC-CREAF

Dr. Marisol Felip Benach  
UB-CREAF

Doctoral program of Terrestrial Ecology  
Universitat Autònoma de Barcelona (UAB)  
Centre de Recerca Ecològica i Aplicacions Forestals (CREAF)

Cerdanyola del Vallès  
June 2024

---





# Abstract

This thesis provides a comprehensive overview of the molecular mechanisms responding to a wide range of pH conditions in phylogenetically distinct freshwater diatom species. Diatoms are a remarkably diverse group of eukaryotic algae belonging to the Stramenopiles (Heterokonts). They constitute one of the most dominant and ubiquitous groups in aquatic environments, in which they play significant ecological and biogeochemical roles. Diatoms originated in marine waters around 190 million years ago, and many clades have subsequently invaded freshwater habitats. Given the buffered pH conditions in the ocean, continental pH gradients are likely among the primary drivers of evolutionary change. Freshwater diatoms are highly sensitive to pH changes in their environment. However, there is limited information on the genetic mechanisms underlying their sensitivity to pH and the resulting species segregation along the pH gradient.

The main goal of this thesis was to investigate the molecular responses of phylogenetically distant freshwater diatoms to a wide range of pH conditions to elucidate potential adaptive mechanisms determining diatom sensitivity to pH, particularly to acidic environments. To this end, twelve diatom strains were grown under acidic, neutral, and alkaline conditions in a common garden experiment. The twelve strains encompassed a broad phylogenetic range within the raphid pennate clade of diatoms, including species from the genera *Nitzschia*, *Tryblionella*, *Eunotia*, *Navicula*, *Achnanthyidium*, *Gomphonema*, and *Encyonopsis*. The twelve strains exhibited contrasting growth patterns along the pH gradient, with an apparent phylogenetic pH niche conservatism restricted to lower taxonomic ranks and with variable strength among clades.

Environmental pH changes among acidic, neutral, and alkaline pH conditions caused the regulation of a myriad of molecular functions and biological processes across diatoms. Many affected proteins were involved in expression tuning in response to stimuli, presumably to meet physiological requirements dictated by environmental factors. This study demonstrates that the responses of known functions to pH changes exhibited substantial strain-specificity despite strains sharing a significant proportion of these functions. This may be a consequence of the substantial genetic differentiation observed across strains. The study also identifies pH-responsive molecules potentially related to proton extrusion, diatom carbon-concentrating mechanisms (CCM) and

---

---

silica biomineralization.

Growth and molecular responses showed that acidic pH is a more ecologically distinct environment than neutral and alkaline pH conditions for diatoms. The distinctiveness of acidic pH as an eco-evolutionary challenge is probably related to the marine ancestry of diatoms and the widespread regional distribution of pH-circumneutral fresh waters across the planet. Survival at acidic pH likely requires the emergence of new adaptations specifically targeted to low pH, which is consistent with mutational randomness playing a relevant role in adaptive evolution.

This research is the first essential step for uncovering the mechanisms underlying the diatom sensitivity to pH, which has been overlooked until now despite pH being a main ecological factor of diatom species segregation in inland waters. This study shows remarkable inter-specific variability in responses to pH variation and provides evidence that acidic environments represent a challenging environment for diatoms from an eco-evolutionary perspective. The anthropogenic increase in acidity in freshwater environments further underscores the significance of elucidating the genetic mechanisms underlying pH tolerance, particularly at low pH.

---

# Resum

Aquesta tesi proporciona una visió general dels mecanismes moleculars que responen a una àmplia gamma de condicions de pH en espècies de diatomees d'aigua dolça filogenèticament diferents. Les diatomees són un grup divers d'algues eucariotes dins dels estramenòpils (heteroconts). Constitueixen un dels grups aquàtics més dominants i ubics, duent a terme funcions ecològiques i biogeoquímiques importants. Les diatomees es van originar en aigües marines fa uns 190 milions d'anys i, posteriorment, diferents clades han envaït hàbitats d'aigua dolça. Degut a l'amortiment del pH de l'oceà, és probable que els gradients continentals de pH siguin un dels principals impulsors del canvi evolutiu. Les diatomees d'aigua dolça són molt sensibles als canvis ambientals de pH. No obstant això, existeix poca informació sobre els mecanismes genètics que expliquen aquesta sensibilitat i la resultant segregació d'espècies al llarg del gradient de pH.

L'objectiu principal d'aquesta tesi era investigar les respostes moleculars de diatomees d'aigua dolça filogenèticament distants a una àmplia gamma de condicions de pH, en particular als ambients àcids. Per a això, es van conrear dotze soques de diatomees en condicions àcides, neutres i alcalines en un experiment de transplantació. Les dotze soques englobaven un ampli ventall filogenètic dins del clade de diatomees pennades rafídiques, incloent-hi espècies dels gèneres *Nitzschia*, *Tryblionella*, *Eunotia*, *Navicula*, *Achnanthidium*, *Gomphonema* i *Encyonopsis*. Les dotze soques van mostrar patrons de creixement contrastats al llarg del gradient de pH, amb una aparent conservació filogenètica del nínxol de pH restringida a rangs taxonòmics inferiors i amb força variable entre clades.

Els canvis ambientals de pH van provocar la regulació d'una gran quantitat de funcions moleculars i processos biològics en les diatomees. Moltes proteïnes afectades estaven implicades en l'ajust de l'expressió en resposta a estímuls, presumiblement per a satisfer requisits fisiològics dictats per factors ambientals. Aquest estudi demostra que les respostes als canvis de pH de funcions conegudes van mostrar una especificitat de soca substancial malgrat que les soques compartien una proporció significativa d'aquestes funcions. Això pot ser conseqüència de la considerable diferenciació genètica observada entre soques. L'estudi també identifica molècules sensibles al pH potencialment relacionades amb l'extrusió de protons, els mecanismes de concentració de carboni (CCM) de les diatomees i la biomineralització

---

---

del silici.

El creixement i les respostes moleculars van mostrar que el pH àcid és un entorn ecològicament més diferent que les condicions de pH neutre i alcalí per a les diatomees. El caràcter distintiu del pH àcid com a desafiament eco-evolutiu està probablement relacionat amb l'origen marí de les diatomees i l'àmplia distribució regional d'aigües dolces amb pH neutre en tot el planeta. És probable que la supervivència a pH àcid requereixi l'aparició de noves adaptacions específicament dirigides a aquestes condicions, la qual cosa concorda amb el fet que l'aleatorietat mutacional exerceixi un paper rellevant en l'evolució adaptativa.

Aquesta recerca és el primer pas essencial per a descobrir els mecanismes que expliquen la sensibilitat de les diatomees al pH, que fins ara s'havia passat per alt malgrat que el pH és un factor ecològic principal de la segregació d'espècies de diatomees en aigües continentals. Aquest estudi mostra una notable variabilitat interespecífica en les respostes a la variació del pH i aporta proves que els ambients àcids representen un entorn desafiador per a les diatomees des d'una perspectiva eco-evolutiva. L'augment antropogènic de l'acidesa en ambients d'aigua dolça accentua encara més la importància de dilucidar els mecanismes genètics que expliquen la tolerància al pH, particularment als ambients àcids.

---

# Acknowledgements

This endeavor would not have been possible without the supervision of Jordi Catalan and Marisol Felip. I have learned a lot from them, which is something I deeply appreciate because they have a profound knowledge of ecology and evolution that I have been able to discuss with them. I am also profoundly grateful for their trust in me and this project. Jordi has taught me, among many other valuable advice, to focus on the main discoveries and avoid getting lost in the details. I am getting better at it, I hope. As for Marisol, I appreciate her patience and support whenever I needed it. She is an enthusiastic problem solver, as well as a remarkable scientist.

I want to thank all the members of the thesis examining board, Daniel Richter, Josep Piñol, Lucia Campese, as well as Hannah Benisty and Eveline Pinseel, for accepting to evaluate this thesis and for their enthusiasm for this work. It is a pleasure for me to be able to benefit from their expertise.

Special appreciation is due to Dr. Eric Pelletier and his research group at Genoscope-CEA (France) for giving me the opportunity to intern with their team for a few months and granting me access to powerful bioinformatic resources. Eric is always ready to help. I fondly remember my stay at Genoscope in 2021, although it rained a lot there.

I would also like to express my sincere gratitude to Elena Fagín for all the work and personal support she has given me. From her, I have learned some of the lab procedures used in this thesis, and she has always been very willing to help. Thanks should also go to Lluís Camarero for his collaboration in the fieldwork, always with an energetic attitude, and to Saúl Blanco, who taxonomically identified the diatom strains. I am also grateful for the fantastic work environment and healthy community at CREAF.

Last but certainly not least, heartfelt thanks to my family: my mother, my brother, and my father, as well as my friends for their constant spiritual and emotional support throughout the thesis process and in all aspects of my life. They were always there to listen to my hopes and disappointments, reminding me of the importance of believing in myself and my work, and of pushing for what it is worth.

---

# Summary

<b>List of abbreviations</b>	<b>11</b>
<b>Chapter 1 Introduction</b>	<b>13</b>
1.1 Background to the study . . . . .	13
1.1.1 Generalities of diatoms and their ecology . . . . .	13
1.1.2 Diatoms evolution and freshwater colonization . . . . .	14
1.1.3 pH: a key environmental factor affecting diatoms . . . . .	15
1.2 The research gap . . . . .	16
1.3 Research aim, objectives, and questions . . . . .	17
1.4 Significance . . . . .	18
1.5 The structural outline . . . . .	18
<b>Chapter 2 Methods: Experimental and data analysis procedures</b>	<b>21</b>
2.1 Generation of the diatom monoclonal cultures . . . . .	21
2.1.1 Sample collection . . . . .	21
2.1.2 Single-cell isolation . . . . .	22
2.2 Common garden experiment . . . . .	24
2.2.1 Experimental design . . . . .	24
2.2.2 Growth monitoring by fluorescence . . . . .	25
2.2.3 Statistical analyses for the growth rate . . . . .	25
2.2.4 RNA purification and sequencing . . . . .	26
2.3 Bioinformatics workflow . . . . .	27
2.3.1 <i>De novo</i> transcriptomic assembly . . . . .	28
2.3.2 Protein prediction and annotation . . . . .	29
2.3.3 Differential expression analysis . . . . .	30
2.3.4 Functional enrichment analyses . . . . .	31
2.3.5 Response clustering using IndVal . . . . .	34
2.3.6 Responses to acidic pH . . . . .	35
2.3.7 Acid-specific adaptations . . . . .	36
2.3.8 Growth-related genes and proteins . . . . .	38
<b>Chapter 3 Diatom growth along the pH gradient</b>	<b>39</b>
3.1 Results . . . . .	39
3.1.1 Growth patterns along the pH gradient . . . . .	39

---

3.1.2	Growth rates at pH 4.7 . . . . .	43
3.1.3	Observed maximum growth capacity . . . . .	45
3.2	Discussion . . . . .	45
3.2.1	pH niche was conserved within most diatom genera . . . . .	45
3.2.2	Acidic pH was more restrictive for diatom growth . . . . .	47
3.2.3	Speed of declining growth at pH 4.7 as a consequence of stress level in acid-intolerant strains . . . . .	48
3.2.4	Acidophilic strains may have a lower growth capacity . . . . .	49
<b>Chapter 4</b>	<b>Strain transcriptomes and molecular responses to pH</b>	<b>51</b>
4.1	Results . . . . .	51
4.1.1	Transcriptome characteristics . . . . .	51
4.1.2	Orthogroups and functional sets . . . . .	55
4.1.3	Differential expression analysis . . . . .	57
4.1.4	Enrichment analysis . . . . .	59
4.1.5	Response clustering . . . . .	64
4.2	Discussion . . . . .	69
4.2.1	Phylogenetic relationship and niche tolerance as key factors for protein-coding genome sizes . . . . .	69
4.2.2	Cellular resources are reallocated when environmental pH changes, with biosynthesis, transport, and repair playing a significant role . . . . .	70
4.2.3	Gene families involved in plastic responses to a pH condition vary greatly among diatom species . . . . .	71
<b>Chapter 5</b>	<b>Transcriptomic responses to acidic environmental pH</b>	<b>73</b>
5.1	Results . . . . .	73
5.1.1	Functional classification and distribution of enrichments and depletions in acid-tolerant strains . . . . .	73
5.1.2	Gene sets enriched or depleted at pH 4.7 in acid-tolerant strains	78
5.1.3	Gene sets enriched or depleted at pH 4.7 in generalists or acidophiles. . . . .	82
5.1.4	Growth rate as a confounding factor . . . . .	84
5.1.5	Enrichment patterns at pH 4.7 based on the significant acid-tolerant group and pH comparison . . . . .	86
5.2	Discussion . . . . .	87
5.2.1	Adaptations to acidic environments can follow many molecular pathways . . . . .	87
5.2.2	Signal transduction and adaptive proteins adjustment as potential key molecular mechanisms for adaptation to acidic pH .	88
5.2.3	Phosphate may play a key role in acid adaptation across strains	89
5.2.4	FCP but not fucoxanthin biosynthesis is enriched under acidic pH conditions . . . . .	90

---



5.2.5	Acid-enriched CHMP5/Vps60 and alternative ESCRT-III filaments	90
<b>Chapter 6</b>	<b>Potential acid-specific adaptations in diatoms</b>	<b>93</b>
6.1	Results	94
6.1.1	Orthogroups and functions exclusively detected among acid-tolerant strains	94
6.1.2	Adding shared enrichment patterns to gene set exclusivity	94
6.1.3	Gene sets primarily enriched or depleted at pH 4.7 in acid-tolerant strains	97
6.1.4	Exclusive gene sets constitutively expressed or strain-specifically enriched or depleted at pH 4.7	102
6.1.5	Gene sets primarily enriched or depleted at pH 4.7 in generalists or acidophiles	105
6.1.6	Growth rate as a confounding factor	105
6.2	Discussion	108
6.2.1	Most gene sets specifically enriched at pH 4.7 are widely distributed in diatoms	108
6.2.2	Several acid-specific adaptations could be constitutive or strain-specific inducible	109
6.2.3	Most potential acid-specific adaptations could be group- or strain-specific	110
<b>Chapter 7</b>	<b>General discussion and conclusions</b>	<b>113</b>
7.1	Molecular mechanisms for adaptation to varying pH	113
7.1.1	The complexity in plastic responses across strains	113
7.1.2	Diatoms and the challenge of acidity	115
7.1.3	Acidophiles and generalists: distinct strategies for thriving at low pH	117
7.1.4	Adaptive plasticity and canalization in acidophiles	118
7.2	Research limitations and future research	120
7.3	Conclusions	121
<b>Bibliography</b>		<b>123</b>
<b>Appendices</b>		<b>142</b>
<b>Appendix A</b>	<b>Supplementary methods</b>	<b>143</b>
<b>Appendix B</b>	<b>Summary plots for the enrichment analyses</b>	<b>147</b>

---

# List of abbreviations

$\mu$  growth rate. 25, 48

**bp** base pair. 52, 53

**CA** carbonic anhydrase. 63, 64, 78, 114

**CCM** carbon-concentrating mechanism. 16, 114

**CGE** common garden experiment. 17, 21, 24, 27, 117, 120

**Chl** chlorophyll. 14, 90

**CHMP5** charged multivesicular body protein 5. , 73, 78, 80, 89–91, 116, 149

**CLD** compact letter display. 26, 42, 44, 46

**cORA** over-representation analysis using chi-square. 32, 61–63, 77, 80, 99, 121, 148

**CPM** count per million. 30, 51

**DE** differential expression. , 21, 30–32, 81, 97, 98, 100–102, 148

**DEG** differentially expressed gene. 17, 27, 31–33, 57–59, 70, 78–81, 110

**DEI** differentially expressed isoform. 27, 79, 81, 110

**DET** differentially expressed transcript. 17, 31

**ESCRT** endosomal sorting complex required for transport. , 90, 91

**F<sub>0</sub>** dark-adapted minimal fluorescence. 25

**FC** fold change. 30, 32, 33

**FCP** fucoxanthin-chlorophyll protein. , 73, 78, 81, 89, 90, 116, 150

**FCS** functional class scoring. 31–33, 60–63, 77, 80, 99, 121, 148

**FDR** false discovery rate. 12, 26, 31–33, 38, 40–45, 56, 59–63, 81, 84, 97, 98, 100–102

**FUNDC** FUN14 domain-containing protein. 109, 116, 117

**GDPD** glycerophosphodiester phosphodiesterase. 73, 78–80, 89, 90, 101, 109, 116, 117, 148, 155

**GSA** gene set analysis. 31

**IndVal** indicator value. , 21, 34–37, 65, 66, 93, 94, 97–101, 106, 108–111

**IQR** interquartile range. 40, 43, 45

**LCIB** Limiting CO<sub>2</sub>-inducible B protein. 78

**MCP** methyl-accepting chemotaxis protein. 76, 88, 93, 97, 102, 109, 151

- ORA** over-representation analysis. 31, 33, 60
- ORF** open reading frame. 29
- pK<sub>a</sub>** acid dissociation constant. 90
- PAM** partitioning around medoids. 34, 65, 66
- pH<sub>μmax</sub>** maximum-growth pH. 26, 45–47, 49, 50
- Pi** inorganic phosphate. 89, 90
- RFU** relative fluorescence units. 24, 25
- Rubisco** ribulose-1,5-biphosphate carboxylase/oxygenase. 16
- SLFDR** signed ln(FDR). 31–33, 85, 107
- Spearman's  $\rho$**  Spearman's rank correlation coefficient. 38, 84, 85, 107
- UGS** unidirectional gene set. 31–33, 60–63, 77, 80, 99, 121, 148
- VDE** violaxanthin de-epoxidase. 81
- VDL1** violaxanthin de-epoxidase-like 1. 81
- VDL2** violaxanthin de-epoxidase-like 2. 81
- ZEP** zeaxanthin epoxidase. 81
- ZEP1** zeaxanthin epoxidase 1. 81

# Chapter 1

## Introduction

Diatoms are a widely distributed group of microalgae that have a remarkable role in evaluating anthropogenic climate change and habitat degradation because they are excellent indicators of environmental conditions, rapidly responding to their shifts. Diatoms are particularly sensitive to environmental pH, which significantly influences the species composition of diatom freshwater communities. However, the molecular basis of the diatom tolerance to pH fluctuations remains unknown. This thesis aims to provide the first comprehensive view of transcriptomic responses to a wide range of pH conditions in phylogenetically distinct freshwater diatom species, discussing the degree of influence of historical constraints (phylogeny), randomness (because undirected selection), and necessity (physically unavoidable issues) in the adaptation to the selection pressures imposed by continental wide-pH gradients compared to the buffered alkaline environment of the sea, from where the main diatom clades originated. This thesis constitutes the first step to identifying fundamental adaptive molecular mechanisms to pH tolerance and, ultimately, the mechanisms underlying diatom species segregation along the pH gradient. This chapter introduces the study by first discussing the background and context, followed by defining the research problem, aims, objectives and questions, the significance, and finally, the limitations of the investigation.

### 1.1 Background to the study

#### 1.1.1 Generalities of diatoms and their ecology

Diatoms (Bacillariophyta) are a remarkably diverse group of eukaryotic algae belonging to the Stramenopiles, also known as Heterokonts (Lee, 2008). They constitute one of the most dominant and ubiquitous groups in aquatic environments, with a widespread distribution in freshwater and marine ecosystems in which they play significant ecological and biogeochemical roles. Diatoms are unicellular organisms, and colonial forms are present in some taxa. Most species are photosynthetic autotrophs, while others lack pigments and function as colorless heterotrophs or form symbiotic relationships with other organisms. Their lifestyle can be either planktonic

or benthic. The diatom chloroplasts include chlorophylls (Chls)  $a$ ,  $c_1$ , and  $c_2$ , and the most abundant carotenoid is the golden-brown fucoxanthin, responsible for the color of diatom cells.

Diatom cells are contained within a cell wall of silica called the frustule (Babenko et al., 2022; Lee, 2008). This external wall comprises two similar halves called theca that fit together like a Petri dish, the outer theca being the epitheca and the inner, the hypotheca. The flattened plate of the theca is the valve. Diatoms have been traditionally classified and taxonomically identified based on the morphology of valves and the pattern of ornamentation in the valve surface (Blanco, 2020; Williams, 2020). Based on valve symmetry, diatoms are classified as centric if they have a radiating pore arrangement, or as pennate if the pores are arranged with elongated, bilateral symmetry. Pennate diatoms can be araphid or raphid. The raphe is a groove in the frustule that runs longitudinally along the valve and is involved in the gliding motility of the cells. This structure varies among different groups of raphid diatoms (Blanco, 2020). Phylogenetic trees depict four major classes within diatoms: Coscinodiscophyceae (radial centric), Mediophyceae (polar centric and radial Thalassiosirales), Fragillariophyceae (araphid pennate) and Bacillariophyceae (raphid pennate), from the oldest to more recently evolutionarily emerged (Medlin, 2016; Nakov et al., 2018).

Diatoms produce and secrete extracellular polysaccharide mucilages in the form of stalks, apical pads, tubes, fibrils, adhering films, and cell coats (Hoagland et al., 1993; Kumar et al., 2015; Lee, 2008). They are key to cell movement, attachment, habitat stabilization, and anti-desiccation. Motility is typically restricted to raphid pennate diatoms, even though active motility has been reported in a few araphid pennate and pennate-like centric diatoms using other mechanisms (Poulsen et al., 2022).

### **1.1.2 Diatoms evolution and freshwater colonization**

Diatoms likely arose through two endosymbiotic events (Benoiston et al., 2017; Gould et al., 2008; Keeling, 2010; Yoon et al., 2004). Ancestral eukaryotes first engulfed a cyanobacterium around 1.5 billion years ago, leading to photosynthetic eukaryotes like red and green algae. Later, 1,200–700 million years ago, a red alga (perhaps preceded by a green alga) was engulfed by a non-photosynthetic eukaryote, with the endosymbiont undergoing significant cellular reduction. This second endosymbiosis gave rise over time to different groups including heterokonts, within which diatoms emerged. As a result, diatom chloroplasts have a singular structure consisting of four membranes (Scarsini et al., 2019): one derived from the phagocytic membrane of the host eukaryote; one, from the plasma membrane of the secondary endosymbiont; and the two innermost membranes, from the plastid of the secondary endosymbiont, in turn corresponding to two limiting membranes of the cyanobacterial primary endosymbiont.

Diatoms emerged within heterokonts probably from Chrysophyceae or Bolidophyceae

(Lee, 2008). According to molecular clocks, diatoms originated in the ocean around 190 Mya (Nakov et al., 2018). The oldest fossil diatom record generally supports the time of emergence of diatoms, yet it has been challenged (Brylka et al., 2023). They originated after the Permian-Triassic great mass extinction caused by extensive Siberian volcanism that led to anoxic and more acidic oceans. The exceptional conditions favored the radiation of phytoplanktonic groups with red algal plastids, particularly dinoflagellates, haptophytes, and diatoms (Falkowski et al., 2004; Medlin, 2011, 2016). Ancestral diatoms lived in a marine planktonic environment, although the marine benthos was rapidly colonized (Nakov et al., 2019). Diatoms colonized freshwater environments 120–130 Mya ago for the first time, 60–70 Mya after the group origin, and multiple colonization events in distinct diatom main clades have occurred since (Nakov et al., 2019). Oceans show elevated salinity ( $\sim 35 \text{ g L}^{-1}$ ) and buffered slightly alkaline pH (7.8–8.2), while inland waters are much more dilute ( $< 1 \text{ g L}^{-1}$ ) and comprise a broad pH range ( $< 4.5$ – $> 9$ ) (Wetzel, 2001). Diatoms have notably diversified within freshwater plankton and benthos, at a faster rate than their marine counterparts (Nakov et al., 2019). Provided the buffered pH conditions in the ocean, continental pH gradients are likely among the primary drivers of evolutionary change.

### 1.1.3 pH: a key environmental factor affecting diatoms

Freshwater diatoms are highly sensitive to changes in pH in their environment, with their communities responding significantly within a few days (Hirst et al., 2004; Patrick et al., 1968). This trait, together with other characteristics such as their high abundance and diversity in natural habitats, has led to the widespread use of freshwater diatoms as indicators of environmental pH to assess the ecological status of ecosystems (Battarbee et al., 2010; Julius & Theriot, 2010; Stevenson et al., 2010). The pH tolerance and optima of numerous diatom species have been determined (e.g., ter Braak and van Dame (1989), Van Dam et al. (1994), and Duda et al. (2023)). The relationship between the distribution of diatoms and pH gradients is more complex than had been initially observed. Changes in pH in inland aquatic ecosystems are generally linked to variations in the acid-buffering capacity of the water, known as the alkalinity, although pH and alkalinity are not strictly equivalent (Catalan, Curtis, & Kernan, 2009). In this context, there are diatom species that are better indicators of pH and others that are better indicators of alkalinity (Catalan, Pla, et al., 2009).

Previous studies have linked pH tolerance in living organisms with the ability to maintain a nearly neutral cytosolic pH, even when cells are surrounded by an acidic or basic environment (Gross, 2000; Madshus, 1988), although it may go slightly further at extreme external pH (Gimmler & Degenhard, 2001; Messerli et al., 2005). The strict regulation of cytosolic pH is critical for keeping many fundamental activities and cellular pathways functional (Putnam, 2012). Mechanisms related to the cytosolic pH homeostasis could consist of limiting unfavorable proton fluxes through plasma

membrane impermeabilization or an altered membrane potential, buffering the acid or base loading or extrusion by chemical pH buffering systems or pH-dependent metabolic conversions (biochemical pH-stat), and transporting acids or bases across membranes (biophysical pH-stat) (Baker-Austin & Dopson, 2007; Gross, 2000; Mirete et al., 2017; Sakano, 2001). The biochemical pH-stat mechanisms may generally only mitigate the impact of large fluctuations in the short term (Reid & Smith, 2002).

Some mechanisms respond indirectly to pH changes due to constraints of other pH-related gradients. For instance, carbon dioxide concentration decreases with alkalization because it transforms into bicarbonate. The low concentration of carbon dioxide in alkaline seawater activates diatom carbon-concentrating mechanisms (CCMs) in some species (Hopkinson et al., 2016; Tsuji et al., 2017). Diatom CCMs typically consist of bicarbonate pumps and carbonic anhydrases to concentrate carbon dioxide near the ribulose-1,5-biphosphate carboxylase/oxygenase (Rubisco) enzyme to increase its efficiency in carbon fixation. Maintaining an adequate photosynthetic rate to offset the energy-consuming process of cellular respiration is crucial for survival.

## 1.2 The research gap

Despite the widespread use of freshwater diatoms for ecological assessment, mainly as indicators of trophic state and acidic conditions in inland waters, there is limited information on the molecular mechanisms underlying their sensitivity to pH and the resulting species segregation along the pH gradient. Most studies on physiological diatom responses to pH changes have usually focused on predicting the effects of anthropogenic ocean acidification on marine diatoms (e.g., Petrou et al. (2019)). These studies typically encompass a limited range of taxonomic diversity and pH variation, often neglecting the acidic end of the pH gradient. Investigating diatom adaptations to acidic freshwater environments represents an opportunity to fill a significant knowledge gap. Acidic environments have a high biological and evolutionary interest because they show a remarkable ecological distance to the marine environment in which ancestral diatoms evolved and adapted, potentially requiring unknown novel adaptations for surviving in those habitats. These environments offer an opportunity to expand our understanding of evolutionary trajectories and their components of history, chance, and necessity (Lenski et al., 1991).

The birth of next-generation sequencing at the beginning of the 21<sup>st</sup> century has allowed researchers to generate large amounts of nucleotide sequence data without the cost being too high (Mardanov et al., 2018), which has made these technologies very useful to completely sequence the genome and transcriptome of several model and non-model organisms (Ekblom & Galindo, 2010). In comparison with other techniques such as PCR or FISH, which focus on particular target genetic sequences,

scanning and comparing whole transcriptomes is a relatively new technique that can provide more comprehensive results in determining the genetic basis of ecological patterns in trait variation (Ekblom & Galindo, 2010). A representative number of sequenced genomes within the phylogenetical clade of interest is desired to resolve eco-evolutionary hypotheses. However, only a few diatom genomes have been sequenced to date, most of them representing marine species (Osuna-Cruz et al., 2020). Consequently, sequences specific to freshwater diatoms are probably underrepresented in functional databases.

## **1.3 Research aim, objectives, and questions**

The main goal of this thesis was to investigate the molecular responses of phylogenetically distant freshwater diatoms to a wide range of pH conditions to elucidate potential adaptive mechanisms determining diatom sensitivity to pH.

Research objectives:

1. To evaluate growth rates within a range of pH conditions from acidic to alkaline in phylogenetically distinct diatom strains. This will be approached by performing a common garden experiment (CGE), growing diatom monocultures under controlled conditions and different pH levels.
2. To explore the similarities and divergences in the protein-coding transcriptome of the studied diatom strains.
3. To assess the quality of predicted proteomes by determining the completeness of protein-coding transcriptomes and the fragmentation of predicted proteins.
4. To identify differentially expressed genes (DEGs) and differentially expressed transcripts (DETs) related to responses to pH for each diatom strain and pairwise pH comparison.
5. To compare functional characterization and homology relationships among the predicted proteomes.
6. To determine functional categories and orthogroups enriched and depleted in each pH comparison and strain.
7. To identify and discuss shared plastic responses to pH across strains.
8. To identify and discuss diverse response strategies specific and unspecific to acidic pH within diatoms.
9. To discuss the influence of freshwater diatom evolution on observed patterns.
10. To compare our findings with the patterns observed in other organisms.



Research questions:

1. Are there differences in pH optima and tolerances across diatom strains isolated from distinct lakes along the natural pH gradients?
2. How similar are the studied strains in terms of their genetic background?
3. How similar are the studied strains in terms of known functions?
4. Which functions and orthogroups are environmentally induced or repressed?
5. What is the extent of inter-specific variability in molecular responses to pH?
6. Which functions and orthogroups were regulated at low pH?
7. Are there specific adaptations to low pH?
8. Are there shared plastic responses to pH across strains?
9. Do observed plastic responses have an adaptive component?
10. What eco-evolutionary factors could have driven the observed growth and molecular patterns among strains?
11. Do diatoms share adaptations to pH with other organisms?

## **1.4 Significance**

This thesis addresses a critical gap in knowledge on pH sensitivity in diatoms by contributing to the first comprehensive investigation of transcriptomic responses to a wide range of pH conditions in phylogenetically distinct freshwater diatom species. This study demonstrates that the responses of known functions to pH changes exhibited substantial strain-specificity despite strains sharing a significant proportion of these functions. This pattern may result from the substantial genetic differentiation observed across strains, suggesting a large component of historical contingency in the evolutionary changes and species radiation across pH continental gradients. In addition, this thesis provides evidence that acidic environments represent a challenging environment for diatoms from an eco-evolutionary perspective. The documented anthropogenic increase in acidity in freshwater environments further underscores the significance of elucidating the genetic mechanisms underlying pH tolerance, particularly at low pH.

## **1.5 The structural outline**

Chapter 1 introduces the ecological importance of diatoms and the challenges they encounter with pH changes. It outlines the research gap on how diatoms respond to different pH conditions and defines the specific aims and objectives of the thesis.

Chapter 2 describes in detail the methodology used for this study. The first part of the chapter describes the field and lab work. The second part describes the bioinformatic pipeline conducted to detect enriched and depleted gene sets and identify shared responses among strains.

Chapter 3 explores the diatom growth under acidic, neutral, and alkaline pH conditions and discusses the potential underlying factors determining the observed growth patterns.

Chapter 4 describes the protein-coding transcriptomes of the twelve diatom strains and provides an overview of the transcriptomic responses to the three pH conditions examined, discussing the potential eco-evolutionary processes beneath the observed responses.

Chapter 5 provides an overview of molecular responses to acidic pH, describing core molecular responses among acid-tolerant strains in detail.

Chapter 6 identifies functions specifically regulated at low pH and/or exclusively present in acid-tolerant strains as potential acid-specific adaptations.

Chapter 7 provides a comprehensive discussion of the findings obtained throughout the thesis and the potential impact and value of the research, acknowledging limitations that might have affected our findings. Finally, the conclusions of the thesis are presented.



## Chapter 2

# Methods: Experimental and data analysis procedures

This chapter describes the methods to explore the growth and gene expression responses to distinct pH conditions in twelve diatom strains isolated from Pyrenean lakes. The chapter is divided into two main sections: culture experimentation and transcriptomic analyses. Twelve diatom strains were isolated from epilithic samples collected from four Pyrenean lakes with distinct pH conditions. Each strain was grown under acidic, neutral, and alkaline conditions in a common garden experiment. Growth was monitored by fluorescence, and growth rates were estimated. At the end of the experiment, RNA was extracted and sequenced. For each strain, *de novo* transcriptome assembly was performed, and the predicted proteins from transcripts were functionally annotated using InterPro, Gene Ontology, and KEGG databases. Differential expression analysis was conducted to identify genes and proteins whose expression was affected by pH. Based on the differential expression results, distinct enrichment analyses were performed to explore functional sets and orthogroups enriched at each pH condition. Responses across strains and pH comparisons were clustered based on similarity and the IndVal index. Then, gene and protein expression responses of diatoms to acidic pH were investigated using distinct approaches, including the IndVal index. Finally, correlation analysis was used to identify genes, proteins, and gene sets whose expression was related to growth rate. The outcome of these methods was described in Chapters 4 to 5.

## 2.1 Generation of the diatom monoclonal cultures

### 2.1.1 Sample collection

To obtain diatom species adapted to different environmental conditions, we sampled four Pyrenean lakes with distinct pH and alkalinity conditions (Table 2.1): the acidic Lake Aixeus, the pH-neutral Lake Redon and Lake Redó, and the alkaline Lake Estanya. The four lakes are located in the eastern part of the Pyrenean mountain

range. Water samples were collected in each of the four lakes for pH measurement. Then, the epilithic biofilms of 15–20 submerged littoral rocks were scrapped using a clean brush, focusing on the top surfaces, and the content was placed into sterile tubes (Figure 2.1).

**Table 2.1. Water pH and sampling date of sampled lakes.**

Lake	Water pH	Sampling date
Aixeus	4.89	31/10/2020
Redon (Conangles)	6.97	01/11/2020
Redó (Aigüestortes)	7.04	29/10/2020
Estanya	8.11	22/04/2021



**Figure 2.1. Sampling in the Pyrenean lakes.** Submerged rocks were retrieved, and their biofilm was collected.

## 2.1.2 Single-cell isolation

Twelve diatom cells from different species were isolated from the collected samples by single-cell pipetting and suction using capillary pipettes and slides with cavities under an inverted microscope. Isolated cells were placed individually in sterile culture flasks with their sterile best-fitted growth medium. The resulting monocultures were reinoculated periodically under sterile conditions. The monocultures were maintained at 15 °C with a 12:12 light-dark cycle and an irradiance of  $\sim 90 \mu\text{mol photons m}^{-2} \text{ s}^{-1}$ .

Three distinct growth media were used based on the nutrient concentrations and adjusted pH (Table 2.2): WC medium (Guillard & Lorenzen, 1972) adjusted to pH 7.0 (WC7) was applied for diatoms isolated from Redó and Redon lake samples, WC medium adjusted to pH 8.2 (WC8) was used for diatoms isolated from Estanya and PM medium (Urbánková et al., n.d.) adjusted at 4.7 (PM4) was used for diatoms isolated from Aixeus. For medium preparation, macronutrient and micronutrient stock solutions were mixed following the concentrations detailed in Table 2.2, and the resulting medium solution was sterilized in the autoclave for 20 minutes at 121 °C. Then, the required quantity of vitamin stock solution was filter-sterilized using a syringe with a 0.2  $\mu\text{m}$  filter and added to the medium solution. Media and stock solutions were kept in the dark at 4 °C.

Slide preparation for taxonomic identification of the strains was performed following Battarbee et al. (2001). First, an aliquot of each monoculture was placed in a heatproof tube and treated with 20 mL  $\text{H}_2\text{O}_2$  30% on a hot plate to eliminate organic matter. After rinsing, cleaned specimens were mounted in Naphrax and analyzed with ZEISS

## 2.1. Generation of the diatom monoclonal cultures

**Table 2.2. Nutrient and vitamin concentrations in WC and PM media.** All concentrations are shown in mg L<sup>-1</sup>.

	WC medium	PM medium	PM to WC
<b>Macronutrients</b>			
CaCl <sub>2</sub> ·2 H <sub>2</sub> O	36.76	0.37	1/100
K <sub>2</sub> HPO <sub>4</sub>	8.71	2.90	1/3
MgSO <sub>4</sub> ·7 H <sub>2</sub> O	36.97	3.70	1/10
Na <sub>2</sub> SiO <sub>3</sub> ·9 H <sub>2</sub> O	28.42	14.21	1/2
NaHCO <sub>3</sub>	12.60	3.15	1/4
NaNO <sub>3</sub>	85.01	56.70	2/3
<b>Micronutrients</b>			
Na <sub>2</sub> ·EDTA	4.36	2.18	1/2
FeCl <sub>3</sub> ·6 H <sub>2</sub> O	3.15	1.58	1/2
CuSO <sub>4</sub> ·5 H <sub>2</sub> O	0.01	0.005	1/2
MnCl <sub>2</sub> ·4 H <sub>2</sub> O	0.18	0.09	1/2
ZnSO <sub>4</sub> ·7 H <sub>2</sub> O	0.022	0.011	1/2
CoCl <sub>2</sub> ·6 H <sub>2</sub> O	0.01	0.005	1/2
Na <sub>2</sub> MoO <sub>4</sub> ·2 H <sub>2</sub> O	0.006	0.003	1/2
H <sub>3</sub> BO <sub>3</sub>	1.00	0.50	1/2
<b>Vitamins</b>			
B <sub>1</sub>	1.00 × 10 <sup>-1</sup>	1.00 × 10 <sup>-1</sup>	1
H	5.00 × 10 <sup>-4</sup>	5.00 × 10 <sup>-4</sup>	1
B <sub>12</sub>	5.00 × 10 <sup>-4</sup>	5.00 × 10 <sup>-4</sup>	1

AXIO IMAGER A1 light microscope equipped with differential interference contrast (DIC) optics at 1000× magnification. Strain pictures were taken using the same microscope (Figure A.1). All obtained monocultures were pennate diatoms from the Bacillariophyceae class but belonging to five different orders: Bacillariales, Eunotiales,

**Table 2.3. Taxonomic position and origin lake of the twelve isolated strains.** The strain name is based on a common code for species abbreviation and the strain lab identification number.

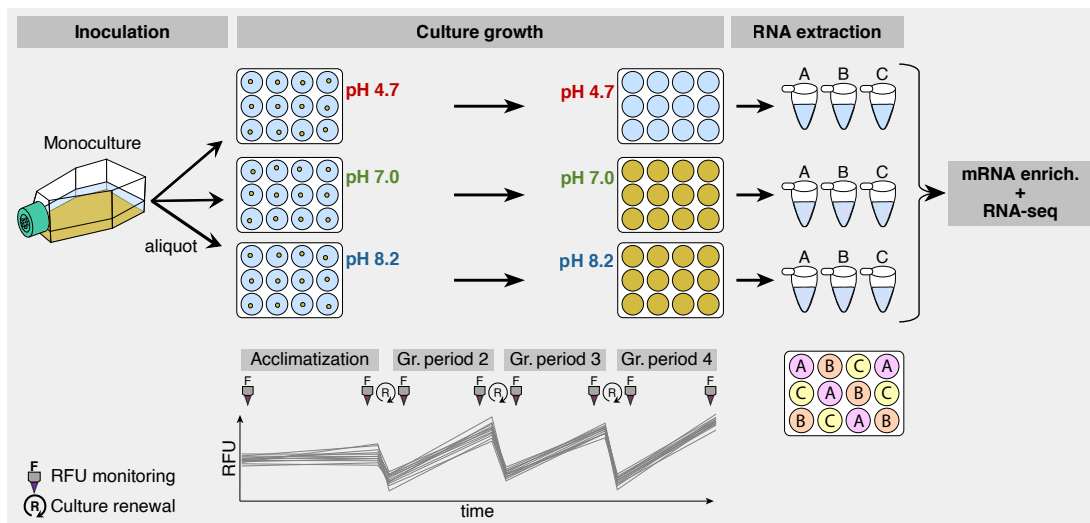
Strain name	Origin lake	Taxonomy
		Bacillariophyceae
		Bacillariales
		Bacillariaceae
NIPM-01	Redó	<i>Nitzschia perminuta</i> Grunow
NPAL-12	Estanya	<i>Nitzschia palea</i> (Kützinger) W.Smith
TAPI-17	Estanya	<i>Tryblionella apiculata</i> W.Gregory
		Eunotiales
		Eunotiaceae
EUNS-26	Aixeus	<i>Eunotia</i> sp.
EUPA-20	Aixeus	<i>Eunotia paludosa</i> Grunow
		Naviculales
		Naviculaceae
NVEN-18	Estanya	<i>Navicula veneta</i> Kützinger
NRAD-29	Estanya	<i>Navicula radiosa</i> Kützinger
		Achnanthales
		Achnanthidiaceae
ACSC-11	Estanya	<i>Achnanthidium</i> aff. <i>sehuencoense</i> E.Morales
ADMI-08	Estanya	<i>Achnanthidium minutissimum</i> (Kützinger) Czarnecki s.s.
ACAF-21	Redon	<i>Achnanthidium affine</i> (Grunow) Czarnecki
		Cymbellales
		Gomphonemataceae
GGDI-23	Redon	<i>Gomphonema graciledictum</i> E.Reichardt
		Cymbellaceae
ECES-28	Estanya	<i>Encyonopsis cesatii</i> (Rabenhorst) Krammer

Naviculales, Achnanthes and Cymbellales (Table 2.3). Only one genus represented Eunotiales, Naviculales, and Achnanthes, whereas Bacillariales and Cymbellales included species from two different genera. The twelve strains were named with a commonly used code for the species name and the strain number from our lab culture collection.

## 2.2 Common garden experiment

### 2.2.1 Experimental design

Common garden experiments consist of growing genetically distinct taxa under the same environmental conditions. By comparing their responses, one can separate the influence of genes from environmental factors. This study used twelve diatom strains and three distinct environmental pH conditions: pH 4.7, 7.0, and 8.2. The experimental design is represented in Figure 2.2. For each strain, three sterile 12-well plates were filled with growth medium: one plate was filled with PM4 medium, one with WC7 medium, and one with WC8 medium. An aliquot of the strain culture was inoculated into each well of the three sterile 12-well plates. The plates were kept at 15 °C with a 12:12 light-dark cycle and an irradiance of  $\sim 90 \mu\text{mol photons m}^{-2} \text{s}^{-1}$  throughout the experiment.



**Figure 2.2. Common garden experiment design.** This process was performed for each of the twelve studied diatom strains. Yellow-brownish surfaces represent diatom biofilms. The growth and RFU pattern shown here is an example used to illustrate the process more comprehensively. Culture renewal can be growth medium renewal or full reinoculation, depending on the culture growth phase. The plate annotated with letters A, B, and C indicates how culture samples from wells were pooled at the end of the experiment into the replicates used for RNA extraction and sequencing.

Experimental cultures were renewed every three days (except four for the acclimatization period). The renewal process was crucial for keeping the population in exponential growth as well as maintaining the original experimental environmental parameters. Experimental culture renewal generally consisted of growth medium

renewal; however, if the population size approached the carrying capacity, the population was reduced by scrapping and partly removing the biofilm in the well to return to an earlier stage of exponential growth. To minimize the effect of daily cellular cycles in fluorescence biomonitoring (see subsection 2.2.2), this process was always conducted at a similar time of day for all plates and strains.

### 2.2.2 Growth monitoring by fluorescence

The growth rate of our experimental populations was assessed by fluorescence monitoring during distinct consecutive growth periods. The dark-adapted minimal fluorescence ( $F_0$ ) parameter has been identified as a reliable fluorescence-based biomass proxy for benthic diatoms (Stock et al., 2019). To measure  $F_0$ , culture plates were kept in the dark for 15 minutes before the fluorescence measurement. A single fluorescence measurement was obtained from the central top of each well with the Varioskan LUX multimode microplate reader. An excitation wavelength of 440 nm and an emission wavelength of 680 nm were used based on previous studies (Consalvey et al., 2005; Herbstová et al., 2015; Nagai et al., 2013). Fluorescence measurements were conducted after the initial inoculation, before and after each experimental culture renewal, and before the total RNA extraction. Therefore, the growth periods used generally comprised three days, starting after either the initial inoculation or experimental culture renewal and ending before the next renewal or the RNA extraction (Figure 2.2).

To calculate the growth rate, the two growth periods after growth rate stabilization (acclimatization period) were selected for subsequent analysis except for rapidly decaying samples, for which only one period could be selected. For each well and growth period, the growth rate ( $\mu$ ) was calculated as:

$$\mu = \frac{\ln(N_1) - \ln(N_0)}{t_1 - t_0}$$

where  $N_0$  represents relative fluorescence units (RFU) at the beginning of the growth period ( $t_0$ ), and  $N_1$  represents RFU at the end of the growth period ( $t_1$ ). As a result, for one strain and pH condition, each growth period included 12 growth rate estimations, one per each well of the 12-well plate.

### 2.2.3 Statistical analyses for the growth rate

Three distinct Kruskal-Wallis and post-hoc Dunn's tests for multiple comparisons were performed. First, Kruskal-Wallis and post-hoc Dunn's tests were conducted for each strain to determine whether and which experimental pH conditions showed significantly different growth rate medians. From the resulting output, strains were classified into distinct response groups based on their response curve. For the second analysis, the goal was to focus on growth at pH 4.7. Only the growth data at pH



4.7 was kept and grouped into the distinct response groups retrieved in the first analysis. Kruskal-Wallis and post-hoc Dunn's tests were conducted to determine the existence of and identify significant differences among the distinct response groups in their growth rate at pH 4.7. The third analysis compared growth rates under the maximum-growth pH ( $\text{pH}_{\mu\text{max}}$ ), which is the pH condition/s where the growth rate was maximum for each strain. The growth data at  $\text{pH}_{\mu\text{max}}$  was kept for each strain. Then, growth data for the same  $\text{pH}_{\mu\text{max}}$  was grouped, and Kruskal-Wallis and post-hoc Dunn's tests were conducted to detect significant differences among the growth rates depending on the  $\text{pH}_{\mu\text{max}}$ .

The analyses were conducted using R v 4.3.1 (R Core Team, 2023). The Kruskal-Wallis tests were performed using the `kruskal.test` function (with default settings). The Dunn's tests were performed using the `DunnTest` function (with default settings and `method = "fdr"`) from the `DescTools` R package (Signorell, 2024). For all the Kruskal-Wallis and Dunn's tests performed for this subsection, the significance threshold used was false discovery rate (FDR) = 0.05. The `multcompLetters` function (with default settings) from the `multcompView` R package (Graves et al., 2024) was used for the compact letter display (CLD) assignment (Piepho, 2004).

## 2.2.4 RNA purification and sequencing

After at least ten days of experimental culture growth since the first inoculation, the biofilm from wells was completely scrapped and total RNA was extracted and purified using the Direct-zol RNA Miniprep Kit (Zymo Research). Three replicates were obtained from each plate (i.e., for each strain and pH condition) by pooling the samples from four wells, as described in Figure 2.2. Total RNA was quantified using the Qubit fluorometer with the Qubit RNA High Sensitivity (HS) Assay kit. In total, 80 RNA samples were obtained (Table 2.4). Insufficient RNA concentration was detected for replicates at pH 4.7 of strains with populations dying at pH 4.7, at pH 8.2 in strain EUNS-26, and for replicate B at pH 4.7 in strain EUPA-20.

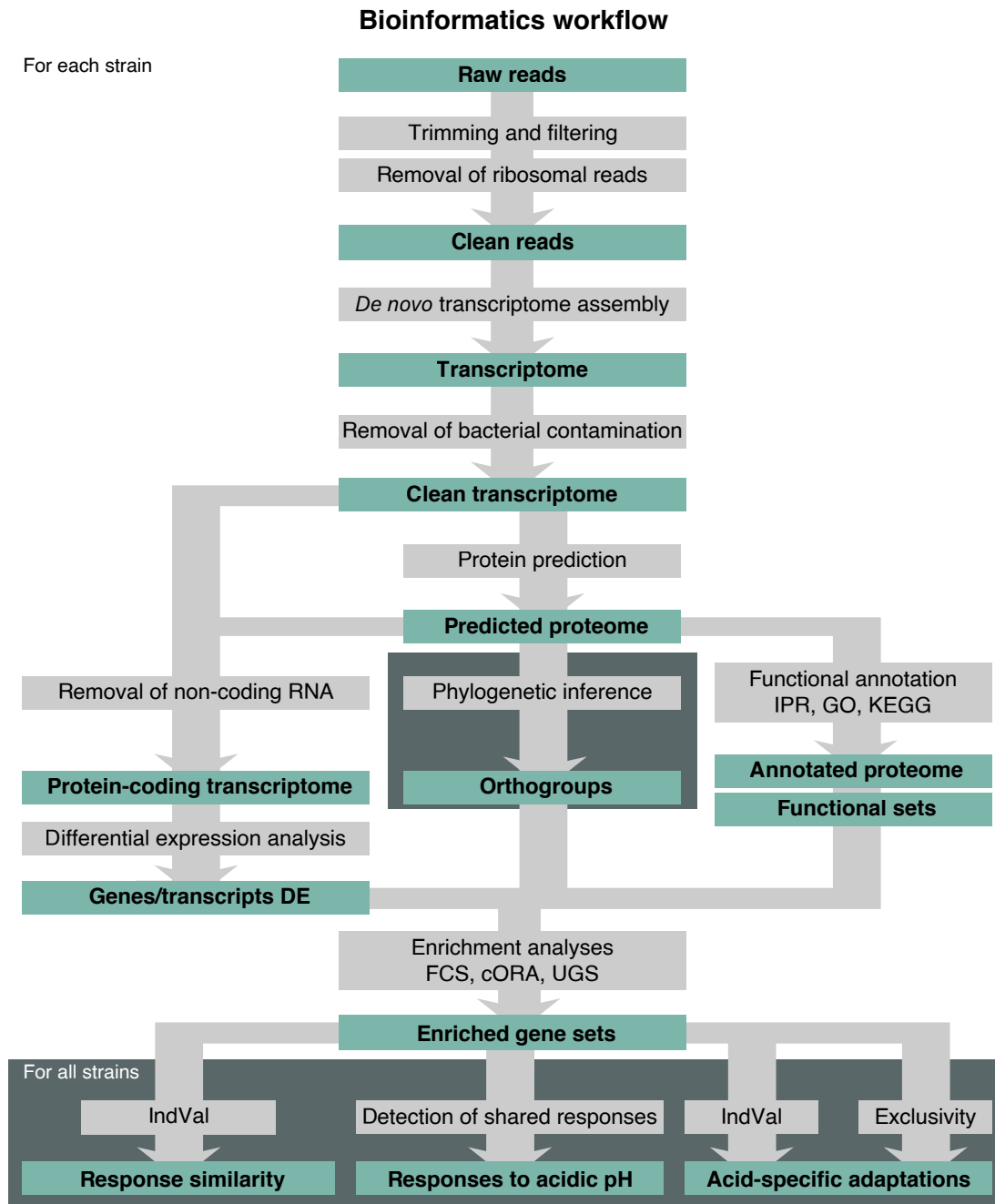
mRNA enrichment, library preparation, and sequencing were performed by the Novogene branch in the UK following their standard protocols (Figure A.2). To selectively capture mRNA from the total RNA sample, magnetic beads coated with poly-T oligos were used. Next, the library was constructed by mRNA fragmentation, reverse transcription using random hexamer primers, end repair, A-tailing, adapter ligation, size selection, amplification, and purification. The quality of the library was assessed using Qubit and real-time PCR for quantification and bioanalyzer for size distribution detection. Finally, libraries were pooled and sequenced on Illumina NovaSeq PE150. Reads that passed the Illumina quality filters are named raw reads.

**Table 2.4. RNA concentration in sample replicates of strains and pH conditions used in the common garden experiment.** For each replicate, 20–50  $\mu\text{L}$  of RNA extract were sent for sequencing.

Strain	Replicate	Concentration (in $\text{ng } \mu\text{L}^{-1}$ )		
		pH 4.7	pH 7.0	pH 8.2
NIPM-01	A	-	37.2	62.0
	B	-	38.4	54.4
	C	-	29.8	64.6
NPAL-12	A	-	41.0	52.0
	B	-	36.2	93.6
	C	-	48.2	45.8
TAPI-17	A	-	33.0	30.4
	B	-	45.0	44.2
	C	-	61.2	48.2
EUNS-26	A	11.3	6.0	-
	B	17.7	4.2	-
	C	11.2	5.2	-
EUPA-20	A	6.4	8.0	12.5
	B	-	5.7	12.2
	C	16.4	5.0	7.8
NVEN-18	A	-	20.8	70.2
	B	-	24.0	41.2
	C	-	28.4	51.2
NRAD-29	A	-	27.2	77.6
	B	-	30.0	75.2
	C	-	28.8	88.0
ACSC-11	A	-	16.8	18.8
	B	-	11.2	36.4
	C	-	20.0	29.8
ADMI-08	A	-	23.2	38.0
	B	-	22.2	39.8
	C	-	18.4	35.4
ACAF-21	A	26.0	25.0	33.4
	B	19.4	22.4	17.7
	C	30.8	22.0	24.2
GGDI-23	A	13.7	10.3	26.4
	B	10.3	11.1	21.2
	C	16.4	10.3	23.4
ECES-28	A	-	5.2	13.8
	B	-	5.9	15.6
	C	-	5.1	14.1

## 2.3 Bioinformatics workflow

The complete bioinformatics workflow used is schematically represented in Figure 2.3. It entailed the transformation of the raw reads obtained from Illumina sequencing to DEGs and differentially expressed isoforms (DEIs) and enriched functions, as well as their use in response clustering among strains and responses to acidic pH.



**Figure 2.3. Bioinformatics workflow.** Processes on a white background were run for each strain independently, whereas those on a dark blue background involved all strains.

### 2.3.1 *De novo* transcriptomic assembly

The raw reads were processed into clean no ribosomal reads by Genoscope (France) using its custom filtering and quality control methods, detailed in Alberti et al. (2017). Filtering steps included: 1) read trimming and cleaning using the Genoscope's internal software fastx\_clean, 2) removing reads and their mates that mapped onto Enterobacteria phage PhiX174 sequences using SOAP aligner (R. Li et al., 2008), and 3) removing ribosomal reads and their mates using SortMeRNA v 1.0 (Kopylova & N, n.d.). Data quality control entailed taxonomic assignation by alignment with Mega BLAST (Blast 2.2.15 suite) (Morgulis et al., 2008) and MEGAN v 3.9 (Huson et al.,

2007), the percentage of merged paired-end reads using the Genoscope's internal software `fastx_mergepairs`, and duplicated sequences rates.

For each strain, clean reads from all samples were used as input for the *de novo* transcriptome assembly performed with Trinity v 2.11.0 (Grabherr et al., 2011). Transcripts were examined using BLAST+ v 2.12.0 (Camacho et al., 2009) to identify potential bacterial contamination. Transcripts with bacterial hits (with default settings and E-value  $\leq 10^{-20}$ , percent identity  $\geq 90\%$ , percent query coverage per hsp  $\geq 40\%$ ) and no Ochrophyta hit based on either `blastn` or `blastx` results were removed and hence not included in subsequent analyses.

The summary of general transcriptome assembly metrics was obtained using the `TrinityStats.pl` Trinity script. The metrics included the number of genes, the number of transcripts, the GC content, Nx metrics, average and median transcript lengths, and total assembled bases.

### 2.3.2 Protein prediction and annotation

Scripts from TransDecoder v 5.5.0 () were employed to predict proteins from the transcripts. First, open reading frames (ORFs) on the top strand containing at least 100 amino acids were retrieved using the `TransDecoder.LongOrfs` script (with default settings and `-S`). Then, two searches on the predicted ORFs for homology to known proteins were performed: a `blastp` search against the Swissprot protein database using BLAST+ v 2.12.0 (Camacho et al., 2009) (with default settings and `-max_target_seqs 1`, `-evalue 1e-5`) and a protein domain search with Pfam v 35.0 (Bateman, 2002) and HMMER v 3.3.2 (Finn et al., 2011) (with default settings and `-E 1e-10`). These two homology searches were included as ORF retention criteria in the second TransDecoder script used, `TransDecoder.Predict` (with default settings). In addition, the longest ORF from each transcript was also kept as a predicted protein. The longest ORF per transcript was retrieved using the `get_longest_ORF_per_transcript.pl` script from TransDecoder. The protein prediction pipeline used resulted in some transcripts encoding multiple proteins. Although eukaryotic gene expression has typically been considered monocistronic, polycistronic gene expression has been reported in diatoms and other protists (Gallaher et al., 2021; Michaeli, 2014; Rogato et al., 2014).

Predicted proteins were characterized using distinct software and functional databases. Protein sequences were annotated with InterPro (Paysan-Lafosse et al., 2022) and Gene Ontology (Ashburner et al., 2000; Consortium et al., 2023) terms using InterProScan v 5.61.93.0 (Jones et al., 2014). Parent terms of the Gene Ontology terms assigned by InterProScan were also assigned to each sequence using the hierarchy among terms from the R package `GO.db` v 3.17.0 (Carlson, 2023). Parent terms are broader GO terms and are closer to the root of the GO hierarchy than their children. Proteins were also annotated with the KEGG Orthology (KO), the KEGG

MODULE, and the KEGG BRITE protein family from the KEGG database (Kanehisa, 2000; Kanehisa et al., 2022). KO identifiers (KOs) were assigned to the proteins using KofamScan v 1.3.0 (Aramaki et al., 2019). KEGG BRITE and KEGG MODULE terms were associated to predicted proteins from their links with KO identifiers according to the KEGG database. The prediction of protein location was obtained from TargetP v 2.0 (Almagro Armenteros et al., 2019) with the "non-plant" organism group selected (default settings and -org non-pl) and ASAFind v 2 (Gruber et al., 2015) (with default settings and -ppc), since this combination is suitable for diatoms (Gruber et al., 2020). The number of transmembrane helices within each protein sequence was predicted using TMHMM v 2.0 (Krogh et al., 2001).

Phylogenetic relationships inference among the proteins from the twelve strains was performed with OrthoFinder v 2.5.4 (D. M. Emms & Kelly, 2015, 2019). This process was performed in two steps. First, different possible species trees were obtained by using proteins on the longest isoform for each gene as the input for OrthoFinder (with default settings). OrthoFinder uses STAG (D. Emms & Kelly, 2018) for species tree inference and STRIDE (D. M. Emms & Kelly, 2017) for rooting the unrooted species tree. The resulting species tree consistent with the diatom phylogeny in Nakov et al. (2018) was selected and used in subsequent analyses. Then, OrthoFinder was rerun, now using the whole set of proteins and the obtained species tree as inputs (with default settings). As a result, all proteins in the twelve strains were assigned to an orthogroup. In addition, transcripts were examined using BLAST+ v 2.12.0 (Camacho et al., 2009) to identify hits with the *Phaeodactylum tricornutum* Bohlin nonredundant protein database (with default settings and E-value  $\leq 10^{-5}$ ).

### 2.3.3 Differential expression analysis

Differential expression (DE) analyses were conducted independently for each strain using Trinity v 2.11.0 utilities (Grabherr et al., 2011). First, the align\_and\_estimate\_abundance.pl Trinity script (with default settings and est\_method RSEM, aln\_method bowtie2, SS\_lib\_type FR, prep\_reference) was used for the alignment-based quantification of transcripts and genes. This utility required RSEM v 1.3.3 (B. Li & Dewey, 2011). Transcript and gene expression matrices were created with the abundance\_estimates\_to\_matrix.pl Trinity script (with default settings and est\_method RSEM). Two independent DE analyses, one for transcripts and one for genes, were performed for each pairwise pH comparison using the run\_DE\_analysis.pl Trinity script (with default settings and method edgeR), which required edgeR software package (Robinson et al., 2009). Only transcripts and genes with  $\geq 1$  counts per million (CPMs) in at least two replicates (default setting for min\_reps\_min\_cpm) were included in the analyses. In all comparisons, the lowest pH was compared to the highest pH, and the fold change (FC) of gene or protein expression was calculated as the ratio of lowest pH to highest pH average expression. For example, when comparing pH 4.7 versus 7.0, a FC = 0.50

indicates that the average gene expression at pH 7.0 is twice that of pH 4.7. Then, the `analyze_diff_expr.pl` Trinity script (with default settings and `-P 0.01`, `-C 0`) was employed to extract DEGs and DETs for each pairwise pH comparison, with a FDR cut-off = 0.01 for significance. This script also provided the clustering of DEGs and DETs, as well as the sample correlation matrix heatmap, according to their expression patterns among samples. Finally, `replicates_to_sample_averages_matrix.pl`, `DE_results_to_pairwise_summary.pl` and `pairwise_DE_summary_to_DE_classification.pl` Trinity scripts (all with default settings, and `avg_log_val` in the first script) were used to merge the results of pairwise pH comparisons into a single output entailing all pH comparisons. Expression data and DE results of each transcript and gene were assigned to the proteins they encoded, generating the transcript-based and gene-based protein datasets, respectively.

### 2.3.4 Functional enrichment analyses

Enrichment analyses were performed to identify sets of genes or proteins enriched in a particular pH. Six set types were included, namely Gene Ontology, InterPro, KEGG KO, KEGG BRITE, and KEGG MODULE terms, and orthogroups. Three different methods for enrichment analysis were used, which are described in Table 2.5. Two of them belong to widely used gene set analysis (GSA) categories, functional class scoring (FCS) and over-representation analysis (ORA) (Khatri et al., 2012; Maleki et al., 2020; Mora, 2019). The third enrichment analysis is a method we named the unidirectional gene set (UGS) analysis. All three analyses were performed for each strain, pairwise pH comparison, set type, and dataset. Two distinct datasets were used: the genes of the gene-based protein DE dataset and the proteins of the transcript-based protein DE dataset. For simplicity, the former will be referred to as the gene dataset and the latter as the protein dataset. For subsequent analyses, a gene set was considered to be enriched when the enrichment was significant based on at least one of the three methods. The analyses were conducted using R v 4.3.1 (R Core Team, 2023). In the following paragraphs, the functional enrichment analyses used will be described for the gene dataset, but the process was also similarly performed for the protein dataset.

FCS methods consider all the genes and rank them by an expression metric and then, for each gene set, combine all gene metrics into a gene set statistic and assess its significance (Khatri et al., 2012; Maleki et al., 2020; Mora, 2019). We performed FCS analysis in three steps. First, similarity among gene sets was computed by the Cohen's Kappa statistic (Cohen, 1960). Gene sets containing exactly the same genes ( $Kappa = 1$ ) were clustered together, and a unique gene set representative of each cluster was included in the FCS analysis to avoid redundancy. After the FCS analysis, enrichment FDR values of representative gene sets were assigned to their whole cluster. FCS analysis was performed with the `runGSA` function of the R package `piano` (Väremo et al., 2013). The gene signed  $\ln(\text{FDR})$  (FDR) was adopted

**Table 2.5. Description of the three enrichment analysis methods.**

<b>Functional class scoring (FCS) analysis</b>	
<b>Significant sets</b>	Sets with a general tendency of their genes to be upregulated at that pH condition.
<b>Statistical test</b>	Wilcoxon rank-sum (non-parametric), with significance threshold: FDR = .01
<b>Gene metric</b>	Gene SLFDR
<b>Pros</b>	<ul style="list-style-type: none"> <li>• Focuses on the general expression pattern of the set.</li> <li>• Uses the whole DEs significance continuum instead of an arbitrary threshold for classifying genes.</li> </ul>
<b>Considerations</b>	<ul style="list-style-type: none"> <li>• Enriched sets might have only few DEGs, or many genes upregulated at each pH condition.</li> <li>• Does not focus on sets with many genes upregulated at the two compared pH conditions.</li> </ul>
<b>Over-representation analysis using chi-square (cORA)</b>	
<b>Significant sets</b>	Sets with a higher proportion of upregulations at that pH condition than the proportion outside the set.
<b>Statistical test</b>	Chi-square (non-parametric), with significance threshold: FDR = .01
<b>Gene metric</b>	Gene DE, with significance threshold: FDR = .01
<b>Pros</b>	<ul style="list-style-type: none"> <li>• Focuses on sets with the highest number of affected genes in each pH condition.</li> <li>• Enrichments can be detected at both compared pH conditions in each set.</li> </ul>
<b>Considerations</b>	<ul style="list-style-type: none"> <li>• Uses an arbitrary threshold for determining DEGs.</li> <li>• Affected by the proportion of up- and downregulations in the whole transcriptome.</li> <li>• Assumes independence between genes, which is not biologically realistic</li> </ul>
<b>Unidirectional gene set (UGS) analysis</b>	
<b>Significant sets</b>	Sets that have all the DEGs upregulated at the same pH condition, regardless of the proportion they represent in the set.
<b>Statistical test</b>	None
<b>Gene metric</b>	Gene DE, with significance threshold: FDR = .01
<b>Pros</b>	<ul style="list-style-type: none"> <li>• Focuses on unidirectional responses, even if the affected proteins represent a small proportion within the set.</li> <li>• Works appropriately with small sets, which may not be detected in FCS nor cORA.</li> <li>• Significance independent of background distribution of DEGs.</li> </ul>
<b>Considerations</b>	<ul style="list-style-type: none"> <li>• Uses an arbitrary threshold for determining DEGs.</li> <li>• Sets can be filtered by DEGs proportion using an arbitrary threshold.</li> </ul>

as the gene level statistic, the global significance test used was the non-parametric Wilcoxon rank-sum test (Barry et al., 2005), and gene sampling was the method for significance assessment of gene sets. Gene sets with FCS FDR  $\leq 0.01$  were considered statistically significant for enrichment.

The SLFDR is defined as:

$$SLFDR = \ln(FDR) \times \frac{\log_2(FC)}{|\log_2(FC)|}$$

For a gene or protein, the more significant the difference in expression between two

pH conditions, the highest the absolute value of SLFDR, and the SLFDR value has the same sign as the  $\log_2(FC)$ . The FC is the ratio between the expression at the lowest pH compared to that at the highest pH. The FC will be higher than one if expression at the lowest pH is greater than expression at the highest pH, and then the logarithm of FC will be positive. Conversely, the FC will be lower than one if the expression at the lowest pH is smaller than the expression at the highest pH, and the logarithm of FC will be negative. For example, a gene with a higher expression at pH 7.0 compared to pH 4.7 with FDR = 0.01 has a SLFDR = -4.61, since the FC is smaller than 1 and the logarithm is negative, and  $\ln(0.01) = 4.61$ .

The second enrichment analysis was the ORA. This analysis statistically evaluates whether there is an association between differential expression and membership in a gene set (Khatri et al., 2012; Maleki et al., 2020; Mora, 2019). This is achieved by comparing the proportion of significant up- and downregulated genes within to the proportions outside each gene set, which can be illustrated as a contingency table. The chi-square test was computed using the `chisq.test` function from the R package `stats` to assess the global statistical significance of over- or under-representation for each gene set. The standardized residual of each gene expression pattern was used to compute its p-value by calculating the upper tail probability of the standardized residual absolute value under the standard normal distribution and multiplying by two (García-Pérez et al., 2014; Haberman, 1973; Kateri, 2014). The p-values were adjusted by FDR. Gene sets with ORA FDR  $\leq 0.01$  were considered statistically significant for enrichment. As in FCS, the ORA analysis used a single gene set representative of each cluster of gene sets containing exactly the same genes ( $Kappa = 1$ ) to avoid redundancy, and afterward, enrichment FDR values of representative gene sets were assigned to their whole cluster.

The third enrichment analysis performed was named the UGS. This method detects gene sets that have all DEGs upregulated towards the same pH condition, regardless of the proportion that DEGs represents within the gene set. These gene sets are considered significantly enriched. Unlike the other two methods for enrichment analysis, this method does not rely on a significance test. Instead, it focuses on finding expression differences that, although they may represent only some portion of the whole gene set, could be biologically relevant. For example, suppose an enzyme is encoded by four different genes; three of them are not differentially expressed, but one of them is upregulated at pH 4.7. The enzyme will probably not appear as significant in the FCS and ORA analyses because the single DEG represents a small proportion of total genes encoding the enzyme. However, changes in the expression of the enzyme isoform of this DEG are sufficient to influence the whole activity of the enzyme at pH 4.7. A more stringent approach can be used by setting a minimum threshold for the proportion of DEGs within the gene set required to be considered enriched.



### 2.3.5 Response clustering using IndVal

With enrichment analyses completed, we searched for shared features and patterns in the transcriptomic responses among strains and pH comparisons. The objective was to cluster contrasts with similar responses. The whole process described in this section was performed independently for orthogroups and functional gene sets. The first step was to generate a binary matrix for the enrichment of gene sets in each contrast, where columns represented contrasts and rows represented enriched gene sets. A value of “1” indicated that the corresponding gene set was enriched in that specific contrast. The gene sets that were enriched uniquely in one contrast were removed from the matrix to focus only on shared gene sets. From this binary matrix, the dissimilarity between contrast pairs was computed by using the Jaccard distance (Jaccard, 1901):

$$d_J = 1 - s \qquad s = \frac{a}{a + b + c}$$

where  $d_J$  is the Jaccard distance,  $s$  is the Jaccard index (or similarity coefficient),  $a$  are 1 to 1 matches,  $b$  are 1 to 0 differences and  $c$  are 0 to 1 differences. The square root was used to approximate the Jaccard distance to an Euclidean-like measure (Legendre & Legendre, 2012). The Euclidean-like Jaccard distance was calculated with the `dist.binary` function from the `ade4` R package (Dray & Dufour, 2007). A dendrogram was computed using the `hclust` R function (with default settings and `method = “complete”`) to sort contrasts by response similarity.

The transformed Jaccard distances were used to determine the optimal number of clusters comprising contrasts with the most similar responses. The following procedure (involving three steps) was performed iteratively, once for each possible number of clusters ( $k$ ) from  $k = 2$  to  $k = 35$ , since there were 36 contrasts in total. First, the transformed Jaccard distances were employed for grouping the contrasts by similarity into  $k$  clusters with the partitioning around medoids (PAM) algorithm, described in Kaufman and Rousseeuw (1987, 1990). Then, for each cluster, the indicator value (IndVal) (Dufrêne & Legendre, 1997) of each significant indicator gene sets ( $p\text{-value} \leq 0.01$ ) was calculated using the `multipatt` function (with default settings and `control = how(nperm = 100000)`) from the `indicspecies` R package (De Cáceres & Legendre, 2009). Finally, all IndVal values from all clusters were added to get the total sum of IndVal. The optimal number of clusters ( $k_{\text{optimal}}$ ) was the  $k$  with the highest total sum of IndVal.

The IndVal index was originally used to assess the indicator value of species for site groups (Dufrêne & Legendre, 1997). In the present work, the IndVal index was computed using contrasts as “sites” and gene sets as “species”, considering a gene set as “present” if it was enriched in that contrast. The IndVal index for a particular gene set and a cluster of contrasts is computed as:

$$IndVal_{ij} = A_{ij} \times B_{ij} \times 100$$

$$A_{ij} = \frac{Ngsets_{ij}}{Ngsets_i}$$

$$B_{ij} = \frac{Ncontrasts_{ij}}{Ncontrasts_j}$$

where  $A_{ij}$  and  $B_{ij}$  are the specificity and fidelity terms of gene set  $i$  to the cluster of contrasts  $j$ , respectively. The specificity (0–1) is calculated by comparing  $Ngsets_{ij}$ , which is the mean presence of gene set  $i$  across contrasts of cluster  $j$ , to  $Ngsets_i$ , which is the sum of the mean presence of gene set  $i$  over all clusters. The specificity is maximum when gene set  $i$  was enriched only in contrasts from cluster  $j$ . The fidelity (0–1) is computed by comparing  $Ncontrasts_{ij}$ , which is the number of contrasts in cluster  $j$  in which gene set  $i$  was enriched, to  $Ncontrasts_j$ , which is the total number of contrasts in cluster  $j$ . The fidelity is maximum when gene set  $i$  is enriched in all contrasts of cluster  $j$ .

To determine the hierarchical relationship among the  $k_{optimal}$  clusters, a symmetric matrix was generated with the number of times each pair of contrasts was grouped in the same cluster from  $k = 2$  to  $k = k_{optimal}$  as a measure of the degree of similarity between contrasts. This matrix was used as the input for the pheatmap function of the pheatmap R package (Kolde, 2019), which creates a clustered heatmap. In addition, a sankey plot with the number of significant indicator gene sets per cluster from  $k = 2$  to  $k = k_{optimal}$  was created with ggplot2 (Wickham, 2016) and ggsankey (Sjoberg, 2024) R packages.

### 2.3.6 Responses to acidic pH

The output of strain-level enrichment analysis was used to identify global and group-shared responses to acidic pH. Two factors were used to classify the shared responses to acidic pH. The first factor was the pH comparison at three levels: enriched at pH 4.7 compared to both pH 7.0 and 8.2, exclusively to 7.0, and exclusively to 8.2. The second factor was the strain group analyzed, which used the response group assignment to strains from subsection 2.2.3. Three strain groups were used: acid-tolerant strains, which included both generalists and *Eunotia*, and generalist and *Eunotia* independently. This analysis identified gene sets enriched at pH 4.7 across all acid-tolerant strains, specifically in generalists and specifically in *Eunotia* strains; and the same for gene sets significant exclusively for the pH 4.7 versus 7.0 comparison and exclusively for the pH 4.7 versus 8.2 comparison. For the purpose of this analysis, the response pattern of gene sets at pH 7.0 compared to pH 8.2 was not considered for gene set classification.

### 2.3.7 Acid-specific adaptations

To investigate further acid adaptations, we explored adaptations specific to acidic pH, meaning they were mostly not enriched in the pH 7.0 versus 8.2 comparison. These specific adaptations were identified from two distinct approaches. Both approaches utilized the response group assignment to strains from subsection 2.2.3 results to focus on acid-tolerant strains, classified into acidophiles and generalists. The first approach used IndVal to determine gene sets primarily enriched or primarily depleted at pH 4.7 across contrasts, meaning it was generally enriched or depleted consistently at pH 4.7 but not enriched in most contrasts for the pH 7.0 versus 8.2 comparison. The IndVal analysis was performed with the `multipatt` function (with default settings and `control = how(nperm = 100000)`) from the `indicspecies` R package (De Cáceres & Legendre, 2009). The p-value threshold used for significance was 0.001. The same binary matrices generated in subsection 2.3.5 were used as input for the IndVal computation.

	Analysis					
	pH4gl	pH4gr	pH4vs7gl	pH4vs7gr	pH4vs8gl	pH4vs8gr
EUNS-26_4UP7	2	2	2	2	4	4
EUPA-20_4UP7	2	2	2	2	4	4
EUPA-20_4UP8	2	2	4	4	2	2
ACAF-21_4UP7	2	3	2	3	4	4
GGDI-23_4UP7	2	3	2	3	4	4
ACAF-21_4UP8	2	3	4	4	2	3
GGDI-23_4UP8	2	3	4	4	2	3
EUNS-26_4DN7	3	4	3	5	5	7
EUPA-20_4DN7	3	4	3	5	5	7
EUPA-20_4DN8	3	4	5	7	3	5
ACAF-21_4DN7	3	5	3	6	5	7
GGDI-23_4DN7	3	5	3	6	5	7
ACAF-21_4DN8	3	5	5	7	3	6
GGDI-23_4DN8	3	5	5	7	3	6
Others	1	1	1	1	1	1

**Figure 2.4. Clustering of contrasts used for IndVal analyses for acid-specific responses.**

Colored cells indicate clusters for which indicator gene sets were searched, each color indicating a distinct cluster. Cluster number and colors are independent for each analysis. The analysis labels (*x* axis) combine the pH comparison with the strain grouping analyzed, with no separator. “pH4” analyses identified shared responses that were significant when comparing pH 4.7 to both 7.0 and 8.2, “pH4vs7” specifically to pH 7.0, and “pH4vs8” specifically to pH 8.2. “gl” analyses were aimed at detecting gene sets specific to acid-tolerant strains, and “gr” analyses acidophiles- or generalists-specific gene sets. The contrast labels (*y* axis) combine the strain name with the simplified label of the response pattern, separated by an underscore (.).

Six different IndVal analyses were performed, one for each pairwise combination of levels from two factors: pH comparison at three levels, “pH4”, “pH4vs7”, “pH4vs8”;

and strain grouping at two levels, “gl” (global) and “gr” (group). The clusters defined for each analysis are shown in Figure 2.4. Cluster 1 included all contrasts comparing pH 7.0 to 8.2. Acidophiles and generalists contrasts involving the same expression pattern were placed in the same cluster in -gl analyses and in two different clusters in -gr analyses. When looking for gene sets enriched or depleted at pH 4.7 compared to 7.0, contrasts entailing pH 4.7 versus 8.2 comparisons were placed in clusters apart from cluster 1 even though they were not searched for indicator gene sets. Likewise, contrasts involving pH 4.7 versus 7.0 comparisons were placed in clusters apart from cluster 1 when aiming at obtaining gene sets enriched or depleted at pH 4.7 compared to 8.2. This strategy allows us to target specifically each pH comparison involving pH 4.7. `restcomb` argument of the `multipatt` function was used to indicate which clusters were to be searched for indicator gene sets. Obtained indicator gene sets were subjected to a cluster specificity filter. Only indicator gene sets with at least 80% of contrasts belonging to the designated cluster were retained as significant.

The second approach for identifying specific adaptations to pH 4.7 consisted of detecting gene sets exclusively present in acid-tolerant strains, grouping them according to which strains contained the gene set. We specifically focused on gene sets exclusively present in the four acid-tolerant strains, exclusively present in acidophiles, and exclusively present in generalists.

**Table 2.6. Classification of shared acid-specific adaptations based on the results from IndVal and exclusivity analyses.** This classification was applied to acid-tolerant, generalist, and acidophile adaptations.

	Exclusive+IndVal	Exclusive+Const subtype Partial	Exclusive+Const subtype Total	Nonexclusive+IndVal
Exclusivity	✓	✓	✓	✗
Significant IndVal	✓	✗	✗	✓
Acid response	✓ <sup>a</sup>	Partial	✗	✓ <sup>a</sup>
No response	✗ <sup>b</sup>	Partial	✓	✗ <sup>b</sup>
Other responses	✗ <sup>b</sup>	✗	✗	✗ <sup>b</sup>

<sup>a</sup>This response type can be absent in certain contrasts, but absences are very limited.

<sup>b</sup>This response type can be present in certain contrasts, but it is very limited.

The IndVal and the exclusivity analyses were combined to classify significant gene sets into the four categories described in Table 2.6. Exclusivity and response specificity to acid pH were adapted accordingly for acid-tolerant-, generalist- and *Eunotia*-specific adaptations particular to pH 4.7. Gene sets that were primarily enriched or primarily depleted at pH 4.7 (significant IndVal) in the group under consideration were classified according to their exclusivity to the group: they were assigned to the “Exclusive+IndVal” category if the gene set was detected exclusively in all strains from the group and to the “Nonexclusive+IndVal” category if the gene set was detected outside the strain group under consideration. Another two categories were dedicated to gene sets exclusively detected in all strains from the group under consideration: exclusive gene sets were assigned to the “Exclusive+Const subtype Total” category when they were not enriched in any contrast across strains and pH

comparisons; and exclusive gene sets were assigned to the “Exclusive+Const subtype Total” category if they were enriched or depleted consistently at pH 4.7 in some contrasts but not enriched in the remaining, including contrasts from the pH 7.0 versus 8.2 comparison.

### 2.3.8 Growth-related genes and proteins

Correlation analyses were used to identify genes and proteins positively or negatively related to growth. The procedure will be described for genes, but a similar approach was applied to the protein dataset. As detailed in section 2.2, expression data was obtained for each of the three replicates retrieved from each culture plate, each replicate entailing four wells, whereas the growth rate was estimated for each well of each culture plate. The mean for the growth rates entailing the four wells included in each expression replicate was calculated so that each replicate had a unique growth rate value and a unique expression value for each gene. Average growth rate and gene expression were standardized by using the z-score:

$$Z_{ij} = \frac{x_{ij} - \bar{x}_{.j}}{\sigma_{.j}}$$

where  $Z_{ij}$  is the variable z-score for the replicate  $i$  and strain  $j$ ,  $x_{ij}$  is the variable value for the replicate  $i$  and strain  $j$ , and the  $\bar{x}_{.j}$  and  $\sigma_{.j}$  are the variable mean and the standard deviation among replicates of the same strain, respectively. This formula was applied to each gene independently for the gene expression data.

Three distinct scales were used for the correlation analyses performed between growth rate and gene expression z-scores: 1) a correlation analysis for each gene in each strain, 2) a correlation analysis for each orthogroup in each strain, and 3) a correlation analysis for each orthogroup combining all strains. When working at the orthogroup scale, each orthogroup comprised the expression data of all its constituent genes. The correlation tests were performed with the `cor.test` R function (with default settings and `method = “spearman”`) using the Spearman correlation. The Spearman’s rank correlation coefficient (Spearman’s  $\rho$ ), p-value, and FDR were retrieved for each correlation.

## Chapter 3

# Diatom growth along the pH gradient

This chapter investigates the growth response of the twelve diatom strains to varying pH conditions. Based on their pH preference, the strains were classified into three response groups: acidophiles (two strains) preferred acidic pH but survived at neutral and alkaline pH, generalists (two strains) grew considerably across all pH conditions, and acid-intolerant strains (eight strains) died under acidic pH. Strains from the same genus generally showed similar growth patterns, suggesting a link between pH tolerance and recent evolutionary history (niche conservatism). Acidic environments were more restrictive for growth, which may reflect the marine origin of diatoms and the regionally limited availability of acidic freshwater habitats. Adaptation to acidic pH likely required the emergence of costly innovations, which will be explored in the following chapters. Acid-intolerant strains exhibited varying declining rates at pH 4.7. This result suggests a difference in the stress intensity and resource allocation between these groups at acidic pH, which may be related to ecological distance to optimal pH and pH tolerance. Optimal pH could also be linked with maximum growth rate, which tended to be faster for strains with an optimal pH closer to the ocean and slower in acidophiles. However, other factors like cell size are likely to affect the maximum growth rate, and a more comprehensive study would be required.

## 3.1 Results

### 3.1.1 Growth patterns along the pH gradient

**Acidophiles** The two *Eunotia* strains, EUPA-20 and EUNS-26, had a positive mean growth rate under all three experimental pH conditions (Figure 3.1a). Because their highest average growth occurred at pH 4.7, these two strains were assigned to the acidophiles guild. Growth was more variable at pH 7.0 and 8.2 than at pH 4.7, particularly regarding the total range: minimum growth rate values were more negative, and maximum growth rate values were more positive at pH 7.0 and 8.2. The

highest interquartile range (IQR) occurred at pH 8.2 in both strains. The growth rates were negative in some samples under pH 7.0 and 8.2.

For EUPA-20, the median growth rate under acidic conditions was  $0.298 \text{ day}^{-1}$ , significantly higher than the value at pH 7.0 (Dunn FDR = 0.001) and at pH 8.2 (Dunn FDR = 0.001). The highest average, minimum, Q25, and Q75 values were observed at pH 4.7. On the other hand, growth at pH 7.0 was comparable to growth at pH 8.2 (Dunn FDR = 0.923), with median values of  $0.167$  and  $0.156 \text{ day}^{-1}$ , respectively. Some outliers with positive and particularly negative growth rates were present at pH 7.0 and 8.2. As for strain EUNS-26, the growth rate at pH 4.7 was higher than that at pH 7.0 (Dunn FDR = 0.044), with median values of  $0.222$  and  $0.125 \text{ day}^{-1}$ , respectively. Contrary to what was observed for strain EUPA-20, growth at pH 8.2 had a median value of  $0.160 \text{ day}^{-1}$ , being in between growth at pH 4.7 and growth at pH 7.0 and not significantly different from them (Dunn FDR = 0.264 for both comparisons). The highest average, minimum, and Q25 values occurred at pH 4.7, whereas the highest maximum and Q75 values occurred at pH 8.2.

**Generalists** The *Achnantheidium* ACAF-21 and the *Gomphonema* GGDI-23 had substantial positive growth under all three experimental pH conditions, with median growth rates above  $0.270 \text{ day}^{-1}$  in all cases (Figure 3.1a). For this feature, they were denoted in this study as generalists. Both strains had their lowest median growth rate value at pH 4.7, but for strain ACAF-21, it was equiparable to the median growth rate at pH 7.0.

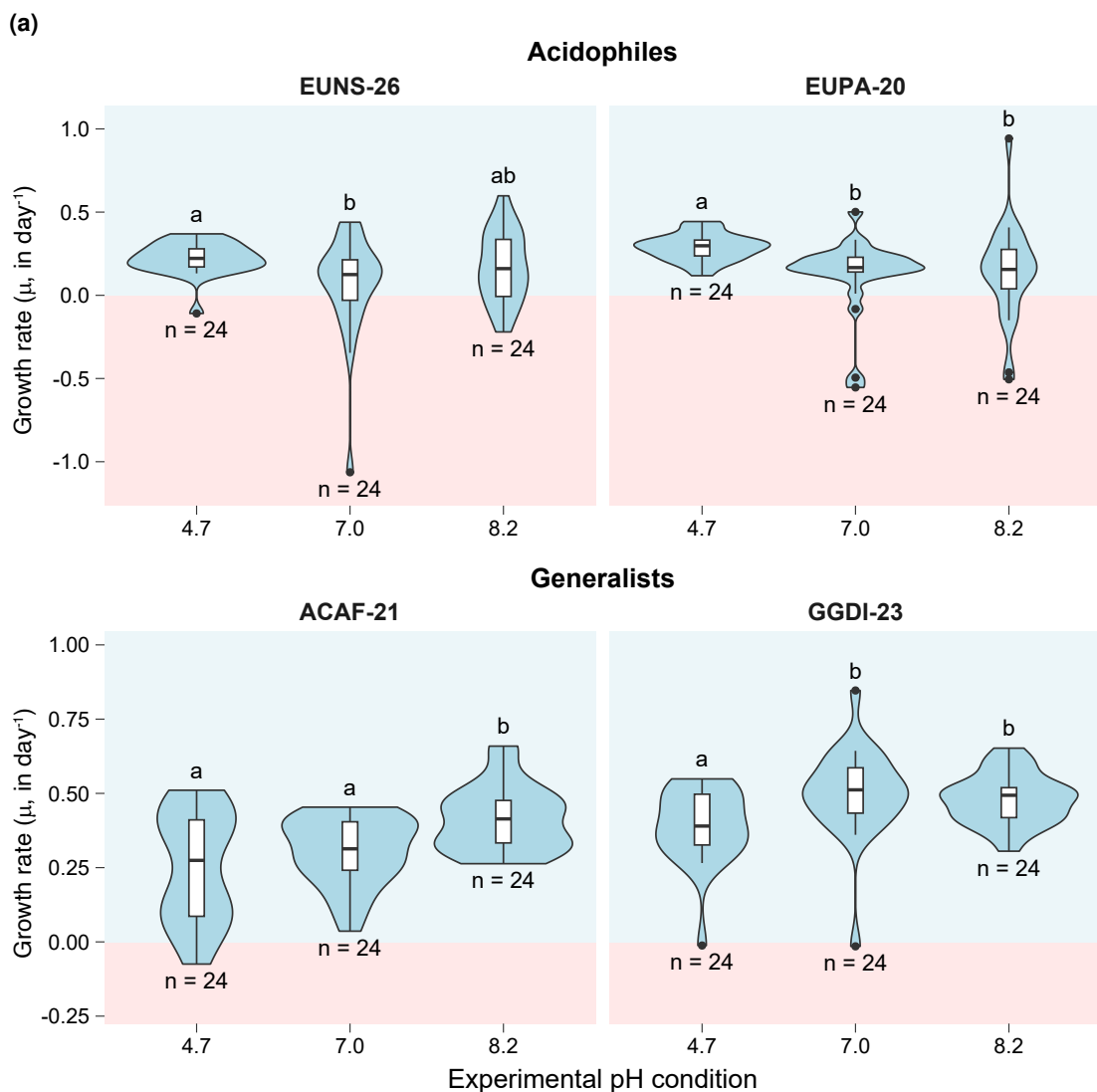
For strain ACAF-21, the maximum growth rate was observed at pH 8.2 (Dunn FDR = 0.017 and 0.011 compared to pH 7.0 and 4.7, respectively) and had a median value of  $0.414 \text{ day}^{-1}$ . Under these conditions, this strain had its greatest values for all descriptive parameters: mean, median, minimum, maximum, Q25, and Q75 values. The growth rate was slower at pH 7.0 and 4.7, with median values of  $0.313$  and  $0.274 \text{ day}^{-1}$ , respectively. This difference between pH 7.0 and 4.7 was not significant (Dunn FDR = 0.705). The growth rate was more variable at pH 4.7 than at pH 7.0: maximum and Q75 values were higher, and minimum and Q25 values were lower at pH 4.7 than at pH 7.0, resulting in a wider range and a wider IQR. Slightly negative growth (up to  $-0.075 \text{ day}^{-1}$ ) was observed in two samples under acidic conditions.

For strain GGDI-23, the growth rates were maximum at pH 7.0 and 8.2, with a median value of  $0.512$  and  $0.494 \text{ day}^{-1}$ , respectively. The difference in growth between both pH conditions was not significant (Dunn FDR = 0.452). The IQR and especially the range for the growth rate were wider at pH 7.0, and Q25 and particularly the Q75 and maximum growth rate values were higher at pH 7.0 than at pH 8.2. At pH 4.7, the growth rate was slower compared to pH 7.0 (Dunn FDR = 0.006) and 8.2 (Dunn FDR = 0.027), with a median value of  $0.390 \text{ day}^{-1}$ . Besides the median, the mean, the maximum, the Q25, and the Q75 values were much lower than those found at pH 7.0 and 8.2. The growth rate was above  $0.250 \text{ day}^{-1}$  in practically all samples under

the three experimental pH conditions.

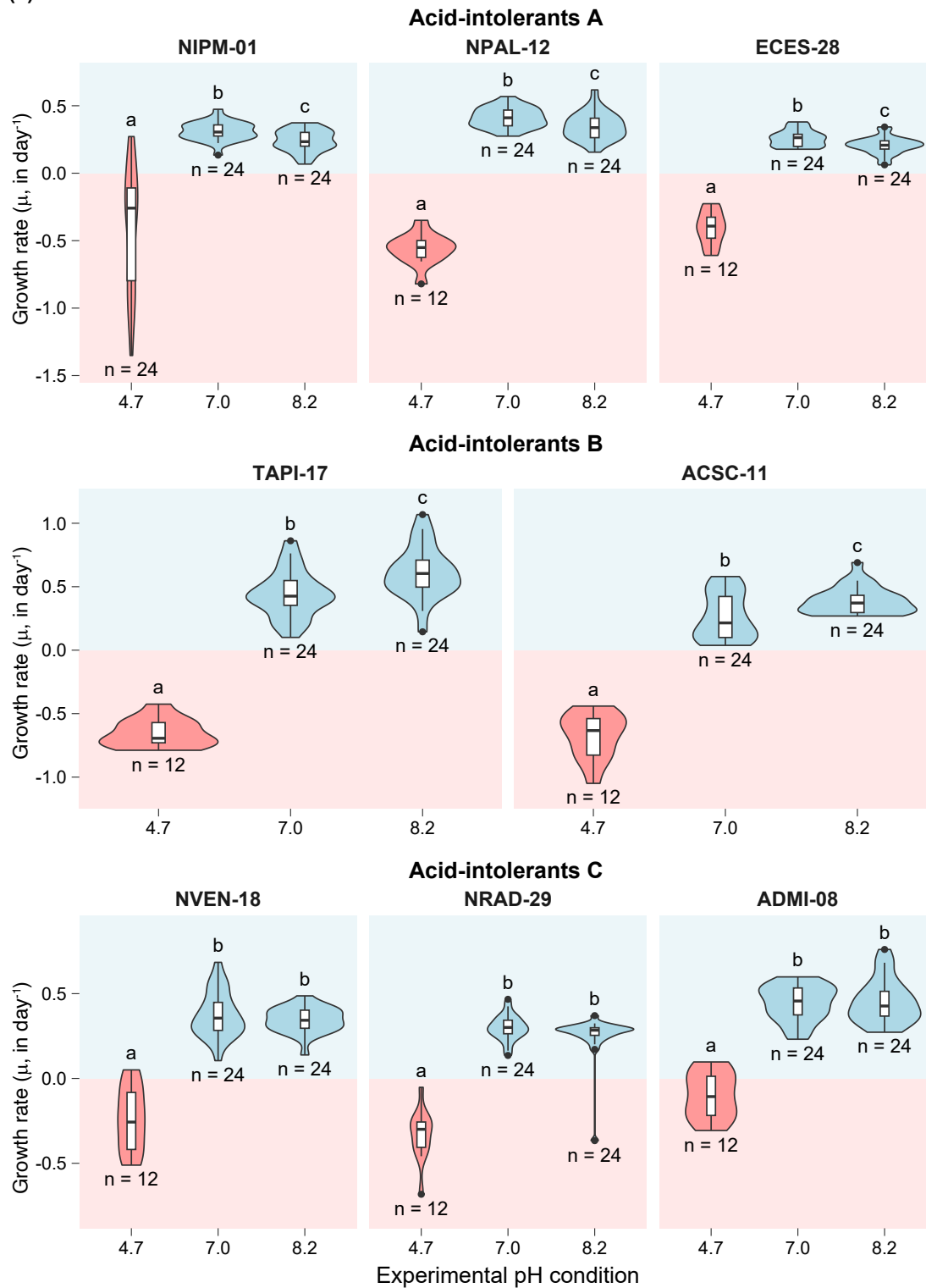
**Acid-intolerant strains** The *Tryblionella* TAPI-17, the *Nitzschia* NIPM-01 and NPAL-12, the *Navicula* NVEN-18 and NRAD-29, the *Achnanthisidium* ACSC-11 and ADMI-08 and the *Encyonopsis* ECES-28 had positive growth at pH 7.0 and 8.2, but their populations died under acidic conditions (Figure 3.1b). For that reason, they were labeled as acid-intolerant strains in this study. Dunn FDR < 0.001 for all contrasts comparing pH 4.7 to pH 7.0 and 8.2 in the eight strains. Depending on their most favorable pH condition, acid-intolerant strains were classified into three groups: strains with a higher positive growth under neutral conditions were assigned to group A; strains that grew better at pH 8.2, to group B; and strains for which the median growth was similar between pH 7.0 and 8.2, to group C.

Acid-intolerant strains that preferred neutral conditions (group A) included the two *Nitzschia* strains, NIPM-01 and NPAL-12, and strain ECES-28. NPAL-12 had a median growth rate of  $0.411 \text{ day}^{-1}$  at pH 7.0, higher than the median of  $0.338 \text{ day}^{-1}$  at pH





(b)



**Figure 3.1. Growth rate at experimental pH 4.7, 7.0, and 8.2 for the twelve diatom strains.** Blue and red violins indicate a positive and a negative median value for the growth rate, respectively. CLD on top of the violins identify statistically indistinguishable groups.

8.2 (Dunn FDR = 0.047). The mean, minimum, Q25, and Q75 values were higher at pH 7.0 than at pH 8.2, but the range was wider at pH 8.2. The other *Nitzschia* strain,

NIPM-01, had a median growth rate of 0.305 at pH 7.0 and it was slower at pH 8.2, with 0.234 day<sup>-1</sup> (Dunn FDR = 0.049). The mean, minimum, maximum, Q25, and Q75 values were also higher at pH 7.0 than at 8.2 for this strain. Lastly, ECES-28 median growth rates values were 0.263 day<sup>-1</sup> at pH 7.0 and 0.207 at pH 8.2, this variation being significant (Dunn FDR = 0.044). At pH 7.0, the mean, minimum, maximum, Q25, and Q75 values for growth rates were also higher than at 8.2.

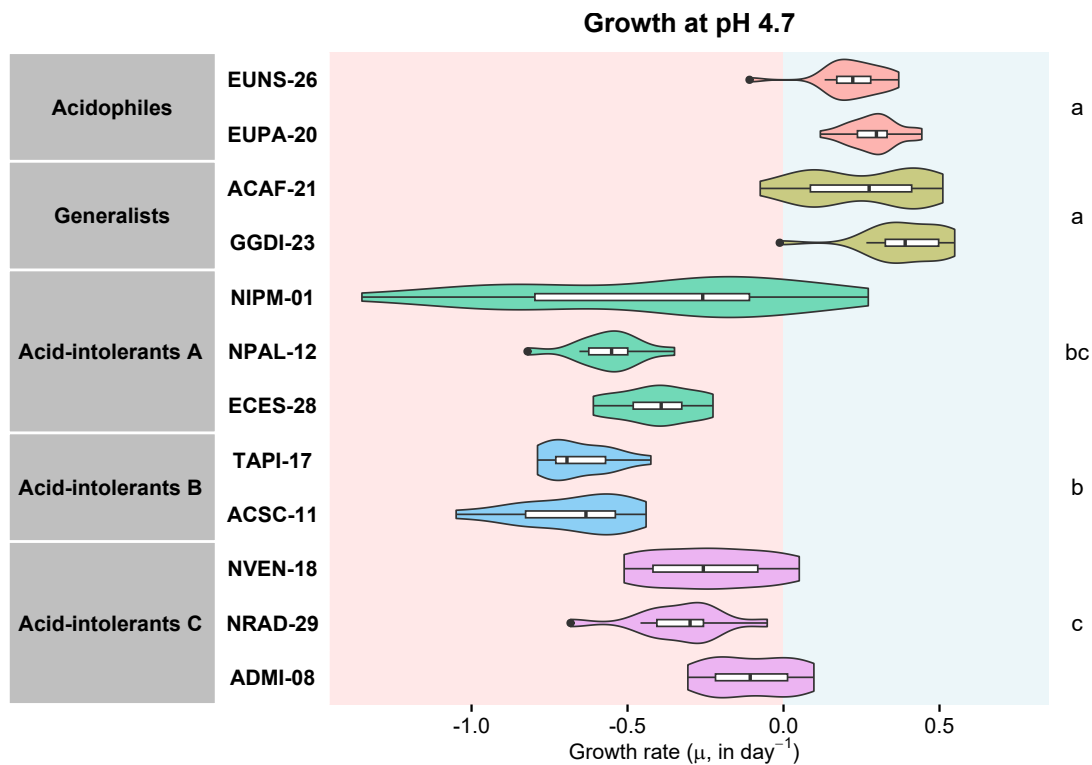
Acid-intolerant strains that preferred alkaline conditions (group B) included strains TAPI-17 and ACSC-11. In the case of TAPI-17, the median growth rate at pH 8.2 was 0.604 day<sup>-1</sup>. At pH 7.0, the median was lower, of 0.426 day<sup>-1</sup> (Dunn FDR = 0.022). Besides the median, also the mean, minimum, maximum, Q25, and Q75 values were higher at pH 8.2. ACSC-11 median growth rate was 0.371 day<sup>-1</sup> at pH 8.2 and 0.215 day<sup>-1</sup> at pH 7.0, which was lower (Dunn FDR = 0.047). In this strain, growth was much more variable under neutral conditions than at pH 8.2, particularly regarding the IQR. This difference in the IQR value resulted from the fact that, although Q75 values were similar under both pH conditions, the Q25 value was clearly smaller at pH 7.0.

Lastly, indifferent acid-intolerant strains (group C), with a similar growth between pH 7.0 and pH 8.2, included both *Navicula* strains, NVEN-18 and NRAD-29, and ADMI-08. NRAD-29 had a median growth rate of 0.301 day<sup>-1</sup> at pH 7.0 and 0.285 day<sup>-1</sup> at pH 8.2. These values were not significantly distinct (Dunn FDR = 0.258). At pH 7.0, this strain had the highest minimum, maximum and Q75 growth rate values. NVEN-18 had a similar median growth rate at pH 7.0 and 8.2 of 0.356 and 0.343 day<sup>-1</sup>, respectively (Dunn FDR = 0.779). Similar to NRAD-29, NVEN-18 had higher maximum and Q75 values for growth rate at pH 7.0 than at pH 8.2. In addition, growth was more variable at pH 7.0 than at pH 8.2 in this strain. Finally, ADMI-08 median growth rate was 0.457 day<sup>-1</sup> at pH 7.0 and 0.427 day<sup>-1</sup> at pH 8.2, these not being statistically different (Dunn FDR = 0.843). For this strain, the growth rate had higher minimum and maximum values at pH 8.2 compared to pH 7.0.

### 3.1.2 Growth rates at pH 4.7

While pH 7.0 and 8.2 supported growth for all twelve strains, only certain strains could maintain positive growth at pH 4.7, namely the acidophiles EUNS-26 and EUPA-20 and the generalists ACAF-21 and GGDI-23 (Figure 3.1 and Figure 3.2). Even though acidophiles had their best growth and generalists had their worst growth at pH 4.7, the median growth rate was not significantly different between the two groups under this pH condition (Dunn FDR = 0.168). The median growth rate under acidic conditions was 0.222 day<sup>-1</sup> for EUNS-26, 0.274 day<sup>-1</sup> for ACAF-21, 0.298 day<sup>-1</sup> for EUPA-20 and 0.390 day<sup>-1</sup> for GGDI-23.

On the other hand, acid-intolerant populations decreased at pH 4.7. The median growth rates between each acid-intolerant group and either the generalists or the acidophiles were noticeably different (Dunn FDR < 0.001 for the six comparisons).



**Figure 3.2. Growth rate at experimental pH 4.7 for the twelve diatom strains.** Each group is represented with a different color. CLD on the right identifies statistically indistinguishable groups.

The intensity of the population decline varied among groups (Figure 3.2). A fast decline at pH 4.7 was observed in group B, with a group B median growth rate of  $-0.664 \text{ day}^{-1}$ . The two strains from this group, TAPI-17 and ACSC-11, showed the fastest median declines among all strains, with rates of  $-0.694 \text{ day}^{-1}$  and  $-0.633 \text{ day}^{-1}$ , respectively. The acid-intolerant group C had a median growth rate of  $-0.246 \text{ day}^{-1}$  at pH 4.7, representing a significantly slower population decline at pH 4.7 than that found for group B (Dunn FDR = 0.001). The median growth rate at pH 4.7 was  $-0.300 \text{ day}^{-1}$  for NRAD-29,  $-0.257 \text{ day}^{-1}$  for NVEN-18 and  $-0.107 \text{ day}^{-1}$  for ADMI-08, the latter being the slowest decline at pH 4.7 among acid-intolerant strains.

Acid-intolerant strains from group A showed a growth rate at pH 4.7 that was intermediate between those of groups B and C, yet the difference was almost significant for both comparisons (Dunn FDR = 0.070 with group B and 0.072 with group C). The median growth rate at pH 4.7 for group A was  $-0.454 \text{ day}^{-1}$ . At the strain level, NPAL-12 populations decreased at pH 4.7 at a median rate of  $-0.551 \text{ day}^{-1}$  and ECES-28 populations of  $-0.392 \text{ day}^{-1}$ . NIPM-01 from group A had a negative median growth rate of  $-0.259 \text{ day}^{-1}$  at pH 4.7, but the strain showed a highly variable growth rate at this pH, with two growth periods with clearly differentiated declining rates. When the analysis was repeated, removing NIPM-01 and keeping NPAL-12 and ECES-28 as the sole constituents of group A, this group still had a growth rate not significantly distinct from either group B or C (Dunn FDR > 0.05 in both cases).

### 3.1.3 Observed maximum growth capacity

In Figure 3.3a, the growth rates under the most favorable pH condition, referred to hereafter as the observed  $\text{pH}_{\mu\text{max}}$ , were compared among the twelve strains. Therefore, for acidophilic strains, samples at pH 4.7 were considered; for strains in the acid-intolerant group A, samples at pH 7.0; for the generalist ACAF-21 and strains in the acid-intolerant group B, samples at pH 8.2; and for the generalist GGDI-23 and strains in the acid-intolerant group C, samples at pH 7.0 and 8.2.

The strain with the fastest growth rate was the acid-intolerant TAPI-17 from group B, whose median value was  $0.604 \text{ day}^{-1}$  at  $\text{pH}_{\mu\text{max}}$ . Besides its median value, its mean, maximum, Q25, and Q75 values for growth rate at  $\text{pH}_{\mu\text{max}}$  were the highest across all studied strains. In addition, the growth rate range and IQR at  $\text{pH}_{\mu\text{max}}$  of this strain were the widest. The two generalist strains, namely GGDI-23 and ACAF-21, showed some of the highest growth rates at  $\text{pH}_{\mu\text{max}}$ , with median values of  $0.497 \text{ day}^{-1}$  and  $0.414 \text{ day}^{-1}$ , respectively. Like ACAF-21, the other two *Achnantheidium* strains also grew faster than most strains at  $\text{pH}_{\mu\text{max}}$ . This was particularly the case for ADMI-08, which, with a value of  $0.443 \text{ day}^{-1}$ , had the highest median growth rates at  $\text{pH}_{\mu\text{max}}$  among the three acid-intolerant strains from group C.

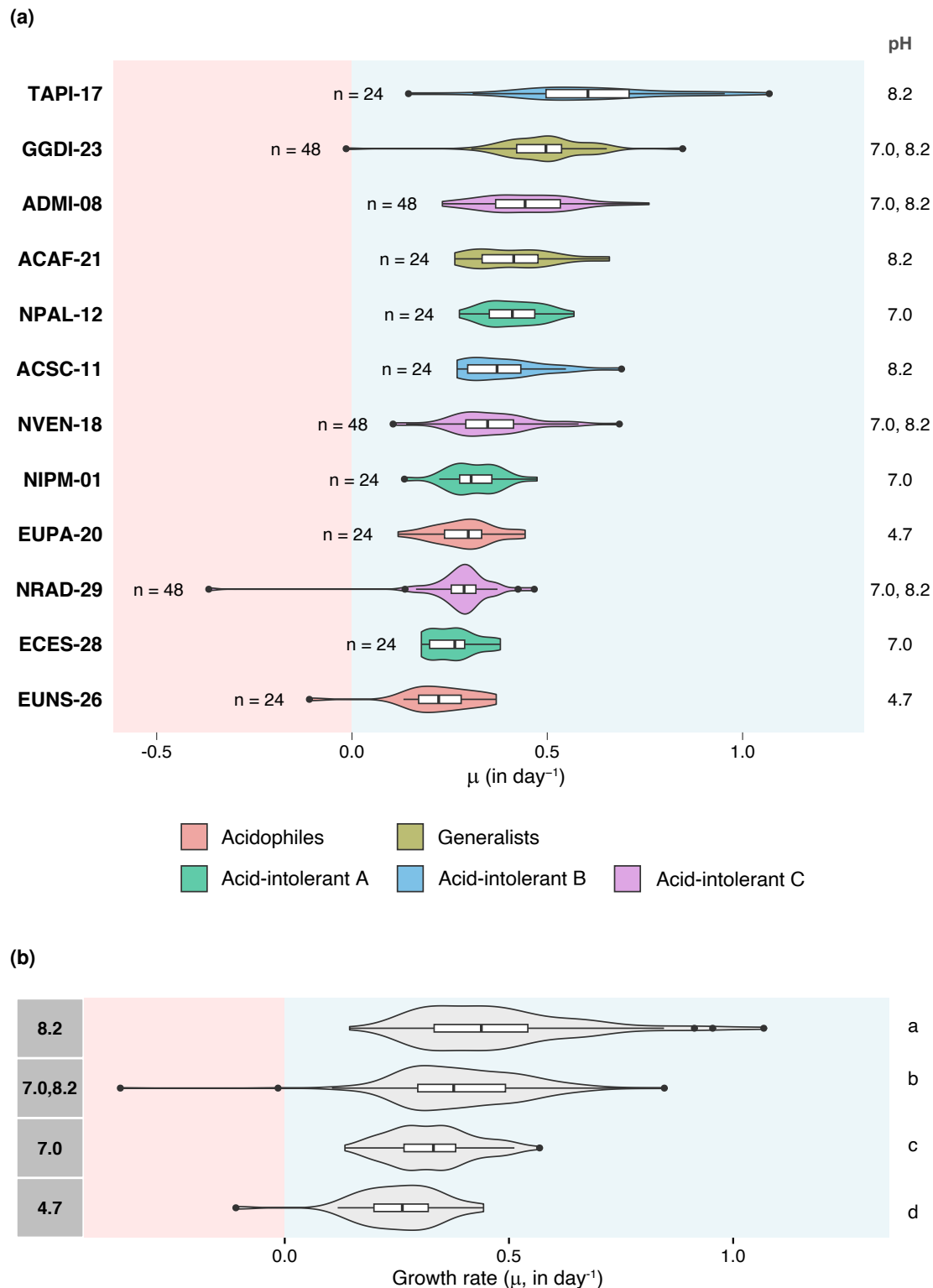
Conversely, the two acidophiles, *Eunotia* EUNS-26 and EUPA-20, were among the species with the slowest growth rates at  $\text{pH}_{\mu\text{max}}$ : their median growth rates were  $0.222$  and  $0.298 \text{ day}^{-1}$ , respectively. *Encyonopsis* ECES-28 from group A of acid-intolerant strains and *Navicula* NAD-29 from group C of acid-intolerant strains also had two of the slowest growth rates at  $\text{pH}_{\mu\text{max}}$ , growing at a median rate of  $0.263$  and  $0.287 \text{ day}^{-1}$ , respectively.

Growth data at  $\text{pH}_{\mu\text{max}}$  of strains that showed the same  $\text{pH}_{\mu\text{max}}$  was pooled. All resulting groups displayed significant differences (Dunn FDR < 0.01 for all comparisons) in their median growth rate at  $\text{pH}_{\mu\text{max}}$  (Figure 3.3b). The group of strains with pH 8.2 as their  $\text{pH}_{\mu\text{max}}$  showed the highest median growth rate at  $\text{pH}_{\mu\text{max}}$ , followed by those with comparable growth at pH 7.0 and 8.2. The lowest median growth rates at  $\text{pH}_{\mu\text{max}}$  were observed in the group with the  $\text{pH}_{\mu\text{max}}$  at 7.0 and, especially, in the group with  $\text{pH}_{\mu\text{max}}$  at 4.7.

## 3.2 Discussion

### 3.2.1 pH niche was conserved within most diatom genera

The twelve diatom strains studied in this investigation were classified into three ecological guilds according to their growth pattern along the pH gradient: the *Eunotia* EUNS-26 and EUPA-20 were described as acidophiles because their fastest growth was observed at pH 4.7, *Achnantheidium* ACAF-21 and *Gomphonema* GGDI-23 were assigned to the generalist guild because they showed considerable growth under the three experimental pH conditions, and the remaining eight strains were described



**Figure 3.3. Growth rate distributions for the twelve diatom strains under their most favorable pH condition (a) and by pH<sub>μmax</sub> (b).** In (a), each response group is represented with a different color, and pH<sub>μmax</sub> is displayed on the right. In (b), CLD on the right identifies statistically distinguishable groups.

as acid-intolerant strains because their populations died at pH 4.7 a few days after inoculation. Acid-intolerant strains were further classified into three groups depending on their  $\text{pH}_{\mu\text{max}}$ . Acidophiles were not considered generalists due to their slow-paced growth rates at neutral and alkaline pH conditions. This classification into guilds was used in the following sections and chapters.

Species from the same genus showed similar growth patterns, although it was more variable among the three *Achnantheidium* strains. This persistence of ancestral ecological characteristics within a lineage over evolutionary timescales suggests the presence of phylogenetic pH niche conservatism in diatoms (Webb et al., 2002; Wiens & Graham, 2005), although it seems restricted to lower taxonomic ranks and seems to have a variable strength among groups. In line with our results, Borrego-Ramos et al. (2023) found a significant phylogenetic signal for pH in the genera *Nitzschia* and *Gomphonema*, but not in *Achnantheidium*. Other studies showed potentially contrasting results (Keck et al., 2016), indicating that this question will require further investigation.

### 3.2.2 Acidic pH was more restrictive for diatom growth

All twelve strains analyzed in our study could survive and grow at neutral and mildly alkaline pH, even though they were isolated from lakes with distinct pH conditions and showed varying growth patterns along pH. Differences between growth at pH 7.0 and 8.2 were uniquely significant for some strains. On the other hand, acidic environmental pH was uniquely tolerated for growth by the two strains collected from acidic lake Aixeus, the *Eunotia* EUNS-26 and EUPA-20, and the two strains collected from lake Redon, the *Achnantheidium* ACAF-21 and the *Gomphonema* GGDI-23. Without being acidic, Lake Redon shows very low ionic strength waters. Therefore, the observed pattern agrees that acidic environments are generally more restrictive for diatom growth and diversification (Hirst et al., 2004; Patrick et al., 1968; Schneider et al., 2013), although regionally, it may depend on the lake pH commonness (Telford et al., 2006)

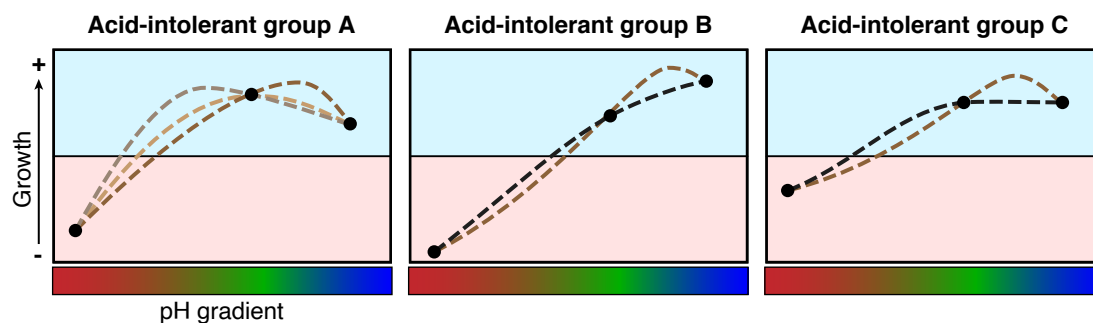
According to molecular clocks, diatoms originated in the ocean around 190 Mya (Nakov et al., 2018, 2019). The seawater has been mildly alkaline during the Phanerozoic time (Halevy & Bachan, 2017). When the first marine diatoms appeared, the mean ocean surface pH was around 7.6 and began to become more alkaline 100 Mya. Marine diatoms with optimal pH close to ocean pH (mildly alkaline) would be expected, as these conditions are where they emerged and evolved. Natural selection would have favored adaptations for optimal growth at this mildly alkaline pH, leading to diatoms best adapted to these pH conditions over time. Diatoms colonizing fresh waters might have been best adapted to environments with pH conditions similar to the ocean's. Acidic pH is more distant ecologically from the ocean pH where diatoms originally evolved, whereas pH 7.0 and 8.2 are closer between them and so is ocean pH. As a result, adaptation to pH 4.7 likely required the emergence of resource-demanding molecular and genetic innovations (Gostinčar et al., 2022). In

Chapter 4, we will provide an overview of the molecular adaptations of the twelve analyzed diatoms to acidic, neutral, and alkaline pH conditions, and Chapters 5 and 6 will delve deeper into the adaptations to acidic environments.

Acidic pH conditions are globally far less abundant than circumneutral and mildly alkaline pH conditions in fresh waters, including the Pyrenees (Catalan, Curtis, & Kernan, 2009; Pinheiro et al., 2021), and thus less probable to reach by the passive dispersal of diatoms (Kristiansen, 1996). Marine ancestry and lake pH commonness are two factors that have probably led to a limited proportion of diatom species developing the capacity to grow at pH 4.7. In our study, acidophile strains tolerated circumneutral or mildly alkaline pH environments, which might imply that they have not lost all adaptive genes to mildly alkaline pH present originally in their genome due to their marine ancestry.

### 3.2.3 Speed of declining growth at pH 4.7 as a consequence of stress level in acid-intolerant strains

Organisms often exhibit varying growth rates across environmental gradients, determining their ecological response curves. The most common response curves are the symmetrical unimodal curves (Austin, 2007; Duda et al., 2023; Rydgren et al., 2003). This pattern also seems to apply for diatoms and the pH gradient, followed by monotonic and skew responses (Birks et al., 1990; Duda et al., 2023). In these models, the growth rate is maximum at the organism's optimal pH range. As the environment deviates further from this optimum, the growth rate typically declines due to an increasing cost of stress tolerance on cellular functions in a context of finite resources (Bruggeman et al., 2023; Burnap, 2015; Nyström, 2004).



**Figure 3.4. Schematic growth response curves according to the acid-intolerant group.** The  $x$  axis represents the pH gradient and the  $y$  axis, the growth rate ( $\mu$ ) in  $\text{day}^{-1}$ . Acidic, neutral, and alkaline pH conditions are represented in red, green, and blue, respectively, in the  $y$  axis. Each response curve model example is represented with a different color: monotonic models are depicted in black; and unimodal models are depicted in dark brown, light brown, or gray when the optimal pH lies above, at, or below pH 7.0, respectively.

In our study, the declining growth speed of acid-intolerant populations varied from group B to C, with group A showing an intermediate decline rate. The growth rate of all three acid-intolerant groups greatly decreased from pH 7.0 to 4.7, moving from positive growth to collapse. However, these groups had different growth patterns between pH

7.0 and 8.2. In group B from acid-intolerant strains, the growth rate decreased from pH 8.2 to 7.0 to 4.7, which suggests that the response curve between these two pH conditions could be either monotonic or unimodal with the optimal pH just below pH 8.2 (Figure 3.4). For a monotonic response, the real optimal pH is equal to or higher than 8.2 for these strains. On the other hand, strains from group C had comparable growth at pH 7.0 and 8.2, suggesting either an unimodal response with an optimal pH between pH 7.0 and 8.2 or a monotonic response with a plateau including these two pH conditions (Figure 3.4).

An optimal pH further from pH 4.7 and/or a lower pH tolerance could be the cause for pH 4.7 representing a more stressing environment for group B than for group C and thus likely requiring a higher amount of resources to be used for stress tolerance. Since cellular resources are finite, directing more resources to stress tolerance would necessitate a corresponding decrease in resources dedicated to growth (Bruggeman et al., 2023; Burnap, 2015; Nyström, 2004). When the stress is too strong, the cell cannot provide enough energy to sustain minimal cellular functions, leading to cell death (Sokolova, 2013). Higher stress under pH 4.7 in group B strains would explain the faster population decline observed at this pH in this group. Strains from group A apparently have an unimodal response curve with a circumneutral optimal pH (Figure 3.4), which is likely closer to pH 4.7 than the group B optimal pH but this did not lead to a significant difference in declining rates between both groups. The differences were neither significant between groups A and C, thus representing an intermediate stress level to pH 4.7 between groups B and C.

### 3.2.4 Acidophilic strains may have a lower growth capacity

The  $\text{pH}_{\mu\text{max}}$  observed for each strain generally matched the pH of the origin lake, but there were some exceptions. The generalist *Achnanthes* ACAF-21 was obtained from Redon (Conangles), which is a circumneutral lake, but this strain grew faster at pH 8.2 than at 7.0. The *Nitzschia* NPAL-12 and the *Encyonopsis* ECES-28 were isolated from the alkaline lake Estanya but had a higher growth at pH 7.0. These findings indicate that diatom species can also be present at suboptimal pH conditions in natural environments. This is common for other microorganisms (Bodor et al., 2020), but may be less generalized in diatoms as suggested by the high specificity to pH for many diatoms in natural environments (Gottschalk & Kahlert, 2012; Van Dam et al., 1994). In our experiment, strains belonging to the same genus generally showed similar  $\text{pH}_{\mu\text{max}}$ . As discussed in subsection 3.2.1 for the growth patterns, the  $\text{pH}_{\mu\text{max}}$  may have a significant phylogenetic signal within genera.

In our study, the mean growth rate at  $\text{pH}_{\mu\text{max}}$  decreased as the  $\text{pH}_{\mu\text{max}}$  diverged from pH 8.2, with acidophiles *Eunotia* strains showing slow growth rates at  $\text{pH}_{\mu\text{max}}$ . This pattern could reflect the varying selective pressures encountered during their evolutionary diversification from a common marine ancestor. This is in agreement with subsection 3.2.2 and other studies (e.g., Patrick et al. (1968) and Hirst et al.



(2004)), which found acidic environments to be generally more limiting to the growth of diatoms. Microbial growth in harsh environments is typically characterized by slow growth rates, in part due to the significant energy allocation into costly cellular mechanisms essential for survival under these conditions (Gostinčar et al., 2022; López-Maury et al., 2008). In addition, the pressure for faster growth may be smaller in these environments because diversity and competition for resources could be lower (Gostinčar et al., 2022). It should be noted that optimal pH may not correspond with the  $\text{pH}_{\mu\text{max}}$  identified in our analysis, since we only explored three pH values of the response curve. Therefore, *Eunotia* optimal pH might be lower than 4.7, and the growth rate at optimal pH might be higher than the one observed in this study for these strains.

Another trait that has been previously shown to determine growth rates at optimal pH and that is worth consideration is cell size. Bigger diatom cell sizes have been associated with slower growth rates (Inomura et al., 2023; Lynch et al., 2022). This pattern applies when comparing the two big-sized and slow-growing *Encyonopsis* ECES-28 and *Navicula* NRAD-29 with the three small-sized and fast-growing *Achnanthes* strains, for instance. However, many other strains did not follow this pattern and, as a result, we did not detect any global relation between cell size and the growth rate at  $\text{pH}_{\mu\text{max}}$ . Precisely determining the relationship between diatom cell sizes and growth rates at optimal pH would require further examination. In summary, the maximum growth rate may be affected by distinct factors including cell size and optimal pH, so that the interactions among these factors may be crucial for a comprehensive study of growth capacity.

## Chapter 4

# Strain transcriptomes and molecular responses to pH

### 4.1 Results

This chapter explores the genetic background of diatom strains and their transcriptomic responses to pH changes. The number of protein-coding genes in studied transcriptomes ranged from 15,556 in ACSC-11 to 31,827 in NVEN-18. The GC content in protein-coding transcripts was higher in more ancient clades. The twelve protein-coding transcriptomes exhibited considerable completeness, with at least 77.0% of stramenopile BUSCOs found. There was apparently no relation between the number of protein-coding genes and the pH niche width of the strain, which might be in line with the presence of a large set of cellular mechanisms responding to most stress conditions. Many enriched functions were associated with biosynthetic, location, transport, and DNA-related processes across strains and pH comparisons. However, each strain activated different genes and functions within these processes, which could have resulted from differences in their adaptive landscape topography and associated evolutionary trajectories. These differences could have derived from the large number of young genes within each strain, which tend to be less evolutionarily constrained. The great proportion of functionally unannotated young genes indicates that many key niche-specialized functions in diatoms have probably not been described.

#### 4.1.1 Transcriptome characteristics

The total number of detected genes in the protein-coding transcriptomes ranged from 18,699 in ACSC-11 to 43,521 in NVEN-18. A relatively large proportion of genes showed low expression in each strain. Henceforth, when referring to genes and their proteins and functional annotation, we will consider the subset of genes or isoforms with an expression above the established minimum threshold (i.e., with CPM  $\geq 1$  in at least two replicates). Considering this, the number of protein-coding genes ranged from 15,556 in ACSC-11 to 31,827 in NVEN-18 (Table 4.1).

Two *Achnanthyidum*, ACSC-11 and ADMI-08, were the strains with the lowest number of protein-coding genes and transcripts detected. On the other hand, the two *Navicula*, NVEN-18 and NRAD-29, and TAPI-17 showed the largest number of protein-coding genes and transcripts. In general, the mean number of transcripts per gene for each strain spanned from 1.8 to 2.2, with at least 42.7% of genes per strain producing a single transcript. Exceptionally, in EUPA-20, the mean number of transcripts per gene was 1.4, with 75.4% of genes encoding a single transcript, whereas in NIPM-01, the number of transcripts per gene was 2.6 on average, and 34.3% of genes produced a single transcript. Excluding EUPA-20, the median longest transcript length ranged from 1,503 bp in ACAF-21 to 2,089 bp in NIPM-01. EUPA-20 had substantially smaller transcript lengths than the other eleven strains, with a median of 894 bp.

**Table 4.1. Summary of the *de novo* protein-coding transcriptome assembly for the twelve studied strains.** N10, N50, and median transcript lengths were measured in bps. Transcripts with their expression below the established minimum threshold were excluded.

	NIPM-01	NPAL-12	TAPI-17	EUNS-26
<b>Number of genes</b>				
Protein-coding genes	18,917	18,130	27,439	19,821
Protein-coding transcripts	48,816	40,320	60,029	37,148
Mean transcripts per gene	2.6	2.2	2.2	1.9
Genes with 1 isoform (%)	34.3	48.0	49.1	57.1
GC content (%)	47.1	46.7	46.2	47.9
<b>Longest transcript length</b>				
Transcript N10	6,305	6,458	5,369	6,659
Transcript N50	3,107	3,132	2,545	3,092
Median transcript length	2,089	2,062	1,658	1,827
	EUPA-20	NVEN-18	NRAD-29	ACSC-11
<b>Number of genes</b>				
Protein-coding genes	24,509	31,827	29,104	15,556
Protein-coding transcripts	35,086	71,226	62,178	29,551
Mean transcripts per gene	1.4	2.2	2.1	1.9
Genes with 1 isoform (%)	75.4	42.7	50.4	57.6
GC content (%)	48.7	46.8	50.0	39.9
<b>Longest transcript length</b>				
Transcript N10	3,000	5,698	5,786	5,929
Transcript N50	1,441	2,789	2,747	3,096
Median transcript length	894	1,844	1,848	1,690
	ACAF-21	ADMI-08	GGDI-23	ECES-28
<b>Number of genes</b>				
Protein-coding genes	19,808	16,070	23,157	22,588
Protein-coding transcripts	44,315	28,531	40,976	46,151
Mean transcripts per gene	2.2	1.8	1.8	2.0
Genes with 1 isoform (%)	48.3	61.9	61.5	51.6
GC content (%)	41.7	41.2	40.0	40.1
<b>Longest transcript length</b>				
Transcript N10	6,082	6,322	6,640	6,889
Transcript N50	3,060	3,364	3,070	3,100
Median transcript length	1,503	1,898	1,692	1,583

The GC content in protein-coding transcripts varied among phylogenetic clades (Table 4.1). The three *Achnanthyidum* and the two Cymbellales strains had a mean mRNA GC content of 39.9% to 41.7%. The GC content in the other seven strains was higher than in the *Achnanthyidum* and Cymbellales strains, ranging from 46.2% to 50.0%. These seven strains included the two *Navicula*, the two *Eunotia*, and the three

Bacillariales strains.

In the studied strains, 76.0% to 88.5% of isoforms contained a single homology-based predicted protein, except in EUPA-20, for which the percentage was 94.2% (Table 4.2). The average number of proteins per transcript was close to 1 for all twelve strains, ranging from 1.1 to 1.3 proteins per transcript. Hence, the number of potentially expressed proteins per strain notably depended on the total number of transcripts. Based on the protein dataset, the twelve protein-coding transcriptomes exhibited considerable completeness, with at least 77.0% of stramenopile BUSCOs found (considering both complete and fragmented).

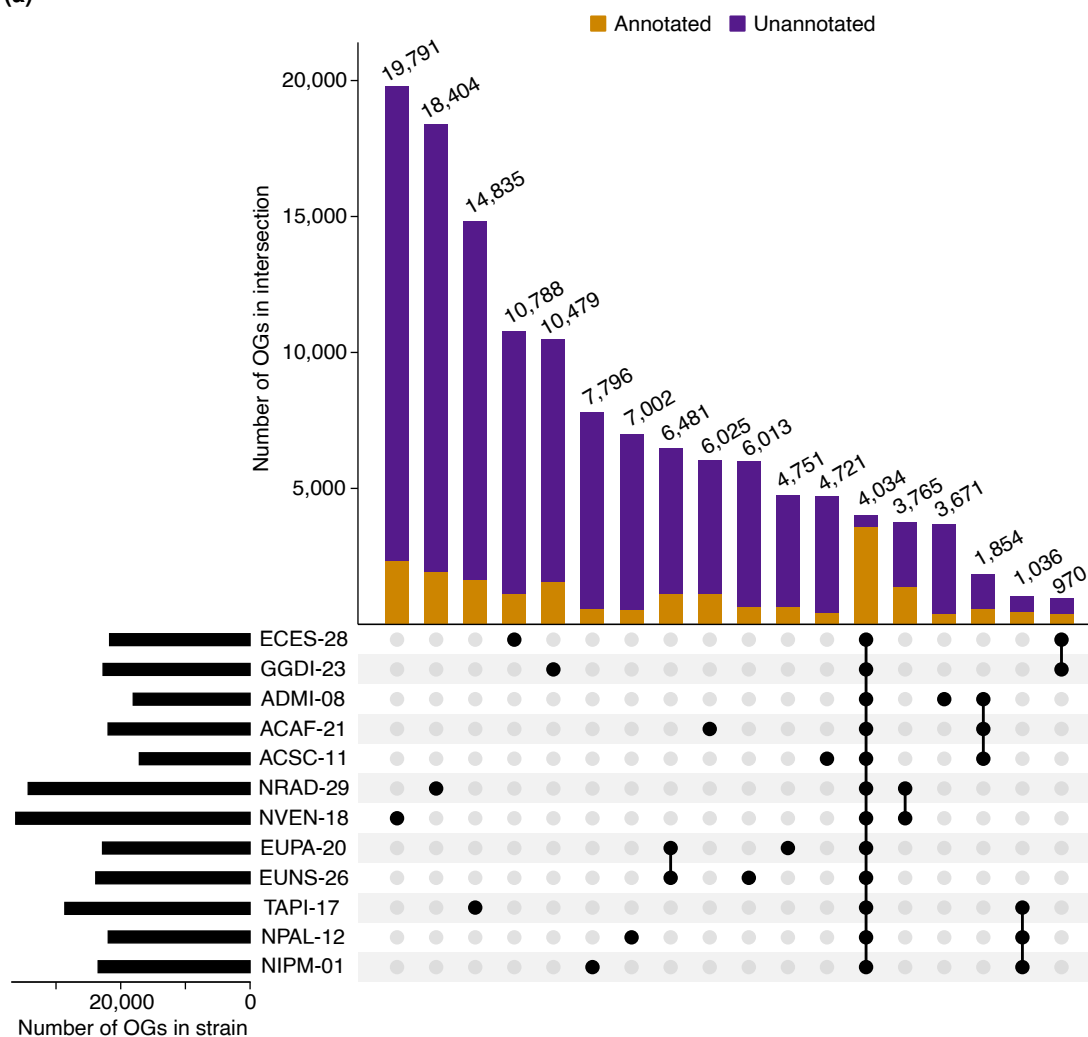
**Table 4.2. Summary of the protein dataset for the twelve studied strains.** BUSCOs completeness was assessed using the stramenopiles odb10 database. Median protein length was measured in bps. Proteins from transcripts with their expression below the established minimum threshold were excluded.

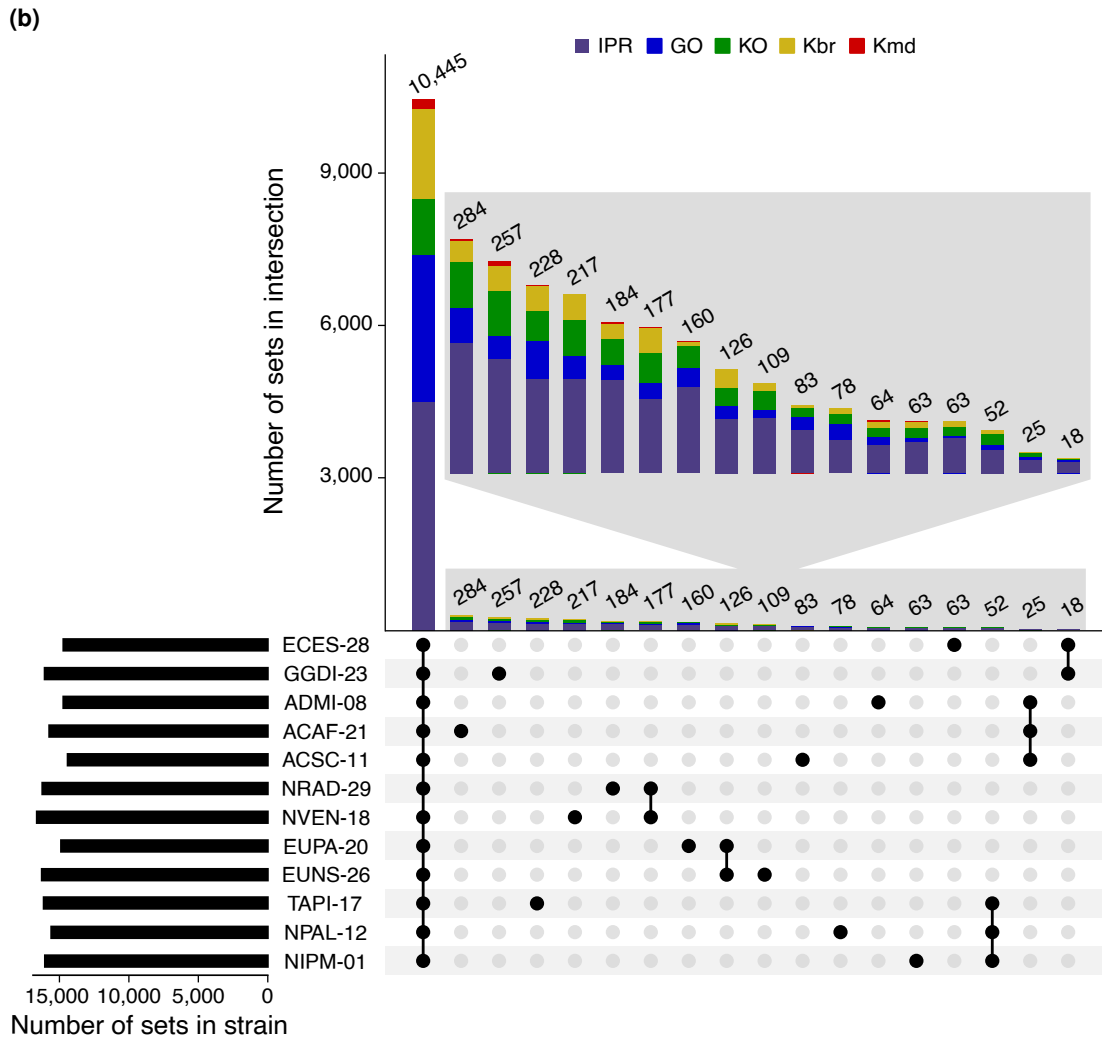
	NIPM-01	NPAL-12	TAPI-17	EUNS-26
<b>Number of proteins</b>				
Proteins	63,659	52,900	71,100	48,084
Orthogroups	23,486	21,926	28,634	23,853
Mean proteins per isoform	1.3	1.3	1.2	1.3
Isoforms with 1 protein (%)	76.5	76.0	84.8	78.3
<b>Dataset quality</b>				
Complete BUSCOs (%)	96.0	88.0	98.0	97.0
Fragmented BUSCOs (%)	1.0	3.0	1.0	1.0
Complete proteins	47,621	38,514	47,705	31,792
Complete proteins (%)	74.8	72.8	67.1	66.1
<b>Protein length</b>				
Median protein length	227	225	202	214
	EUPA-20	NVEN-18	NRAD-29	ACSC-11
<b>Number of proteins</b>				
Proteins	37,361	87,159	80,138	36,387
Orthogroups	22,807	36,225	34,303	17,118
Mean proteins per isoform	1.1	1.2	1.3	1.2
Isoforms with 1 protein (%)	94.2	82.0	77.4	81.7
<b>Dataset quality</b>				
Complete BUSCOs (%)	73.0	96.0	92.0	76.0
Fragmented BUSCOs (%)	10.0	2.0	0.0	6.0
Complete proteins	13,441	62,564	62,133	21,398
Complete proteins (%)	36.0	71.8	77.5	58.8
<b>Protein length</b>				
Median protein length	175	207	207	243
	ACAF-21	ADMI-08	GGDI-23	ECES-28
<b>Number of proteins</b>				
Proteins	54,736	36,825	46,637	53,724
Orthogroups	21,952	18,064	22,725	21,713
Mean proteins per isoform	1.2	1.3	1.1	1.2
Isoforms with 1 protein (%)	81.7	78.1	88.5	86.6
<b>Dataset quality</b>				
Complete BUSCOs (%)	77.0	83.0	89.0	73.0
Fragmented BUSCOs (%)	11.0	4.0	4.0	4.0
Complete proteins	32,029	23,417	28,693	30,499
Complete proteins (%)	58.5	63.6	61.5	56.8
<b>Protein length</b>				
Median protein length	229	248	235	224

A notably large proportion of the detected BUSCOs were complete and not fragmented

(Table 4.2). According to BUSCO, the protein-coding transcriptomes with the best quality were those from TAPI-17, EUNS-26, NVEN-18, and NIPM-01, with more than 96% of stramenopile BUSCOs being complete. On the other end, EUPA-20, ECES-28, ACSC-11, and ACAF-21 had the lowest number of complete BUSCOs, with percentages ranging from 73% to 77%. The percentage of complete proteins decreased for all strains when considering the whole protein dataset, not only BUSCOs. The strains with the highest protein sequence integrity were the two *Navicula*, NRAD-29 and NVEN-28, and the two *Nitzschia*, NIPM-01 and NPAL-12, for which 71.8% to 77.5% of proteins were complete. EUPA-20 showed a markedly high protein fragmentation level compared to the other strains, with complete proteins representing 36.0% of total proteins. The median protein length was slightly smaller in this strain compared to the other strains.

(a)





**Figure 4.1. Number of gene sets in the twelve studied strains and their age classes.** The total number of orthogroups (a) and functional sets (b) in each strain is plotted on the left side. The bar plot of the upset plot shows the number of orthogroups (a) and functional sets (b) found exclusively in all strains of the intersection indicated in the matrix below. For orthogroups (a), the bars are colored according to the proportion of orthogroups that are functionally annotated. An orthogroup was considered functionally annotated if it was annotated by at least one of the databases used. For functional sets (b), the bars are colored according to the database to which the functional set belongs. Note that the  $y$  axes of each plot have independent scales.

#### 4.1.2 Orthogroups and functional sets

Proteins were assigned to orthogroups based on sequence homology. There was consistency between the number of orthogroups and the protein count per strain. In total, 156,682 orthogroups were found. The lowest number of orthogroups across strains was 17,118, detected in ACSC-11, whereas NVEN-18 exhibited the highest count with 36,225 orthogroups (Figure 4.1a). 4,034 orthogroups were found in all twelve strains (pennate-specific), representing between 11.1 and 23.6% of total orthogroups per strain. The percentage of order-specific orthogroups in each strain was generally lower, ranging from 3.6% to 11.0% except in Eunotiales, which each strain contained 27.2 and 28.4% of order-specific orthogroups. It should be noted

that three out of the five diatom orders included in this study were represented by a single genus: Eunotiales, Naviculales, and Achnanthes. Lastly, at the strain level, the percentage of strain-specific orthogroups was mainly higher than the pennate and order-specific orthogroups in each strain, representing at least 20.3% and up to 54.6% of total orthogroups per strain. The differences in percentage in strain transcriptome with order- and pennate-specific were both significant (Dunn  $FDR \leq 0.001$  and  $= 0.005$ , respectively). The two *Navicula* strains and TAPI-17 contained the highest number and percentage of strain-specific orthogroups, whereas EUPA-20 and ADMI-08 had the lowest.

The number of functional sets identified in all twelve strains was greatly higher than those order- or species-specific (Figure 4.1b). The percentage of annotation for each functional annotation database was similar across strains (Table 4.3). In each strain, 40.0% to 50.5% of genes encoded at least one protein with annotations from the InterPro, Gene Ontology, and/or KEGG databases. Almost all of the annotated genes had annotations from the InterPro database. A considerable proportion of those genes also had Gene Ontology annotation. As for KEGG annotations, the number of annotated genes was smaller: 11.3% to 17.2% of genes in each strain were annotated with an entry from the KEGG Orthology database. Most of the KO-annotated genes were also annotated with terms from the KEGG BRITE database, whereas a smaller proportion was annotated using the KEGG MODULE database.

The proportion of functionally annotated orthogroups significantly diminished as the orthogroup age class was younger (Dunn  $FDR \leq 0.01$  for the three comparisons among pennate-, order-, and strain-specific percentages in strains; Figure 4.1a), considering an orthogroup as functionally annotated if it contained at least one annotated protein in any of the functional databases used. The highest percentage of annotation was found for orthogroups present in the twelve studied strains, with 88.7% of these orthogroups having at least one functional annotated protein. The proportion of functional annotation was notably lower for order-specific orthogroups, with 26.8%. Lastly, among strain-specific orthogroups, only 11.7% were functionally annotated.

The phylogenetic tree for the twelve strains inferred from the gene trees is shown in Figure 4.2. The Bacillariales lineage diverged from the common ancestor of the other strains relatively early in evolutionary history, followed by the two Eunotiales strains and then, the two Naviculales strains. Achnanthes and Cymbellales were the most closely related among the five lineages. Within the Bacillariales, the two *Nitzschia* strains were more closely related to each other than to the *Tryblionella* TAPI-17. The *Achnanthes* ADMI-08 and ACAF-21 were more closely related to each other than to the *Achnanthes* ACSC-11.

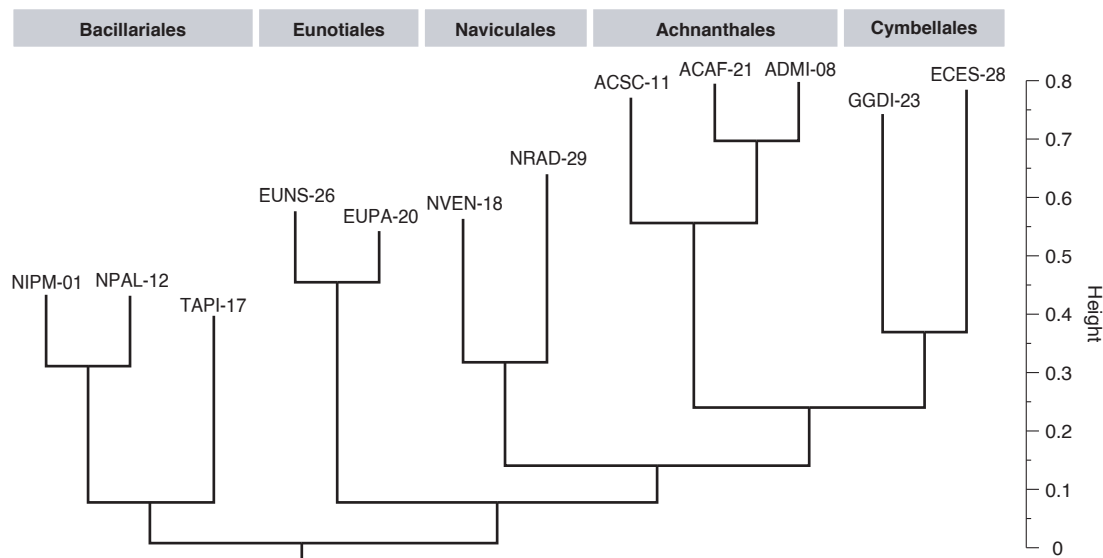
**Table 4.3. Summary of the protein dataset functional annotation for the twelve studied strains.**

	NIPM-01	NPAL-12	TAPI-17	EUNS-26
<b>Number of genes</b>				
Protein-coding genes	18,917	18,130	27,439	19,821
Annotated genes	9,407	9,051	11,541	9,386
Annotated genes (%)	49.7	49.9	42.1	47.4
<b>InterPro and GO</b>				
IPR-annotated	9,388	9,031	11,520	9,363
IPR-annotated (%)	49.6	49.8	42.0	47.2
GO-annotated	6,425	6,065	7,474	6,399
GO-annotated (%)	34.0	33.5	27.2	32.3
<b>KEGG</b>				
KO-annotated	3,262	3,033	3,322	3,311
KO-annotated (%)	17.2	16.7	12.1	16.7
Kbr-annotated	3,152	2,921	3,214	3,205
Kbr-annotated (%)	16.7	16.1	11.7	16.2
Kmd-annotated	569	545	555	567
Kmd-annotated (%)	3.0	3.0	2.0	2.9
	EUPA-20	NVEN-18	NRAD-29	ACSC-11
<b>Number of genes</b>				
Protein-coding genes	24,509	31,827	29,104	15,556
Annotated genes	10,291	12,727	11,797	7,643
Annotated genes (%)	42.0	40.0	40.5	49.1
<b>InterPro and GO</b>				
IPR-annotated	10,258	12,687	11,772	7,623
IPR-annotated (%)	41.9	39.9	40.4	49.0
GO-annotated	6,813	8,298	7,610	5,143
GO-annotated (%)	27.8	26.1	26.1	33.1
<b>KEGG</b>				
KO-annotated	2,775	3,631	3,421	2,444
KO-annotated (%)	11.3	11.4	11.8	15.7
Kbr-annotated	2,671	3,486	3,263	2,362
Kbr-annotated (%)	10.9	11.0	11.2	15.2
Kmd-annotated	527	623	590	470
Kmd-annotated (%)	2.2	2.0	2.0	3.0
	ACAF-21	ADMI-08	GGDI-23	ECES-28
<b>Number of genes</b>				
Protein-coding genes	19,808	16,070	23,157	22,588
Annotated genes	9,763	8,116	10,795	9,232
Annotated genes (%)	49.3	50.5	46.6	40.9
<b>InterPro and GO</b>				
IPR-annotated	9,741	8,100	10,770	9,210
IPR-annotated (%)	49.2	50.4	46.5	40.8
GO-annotated	6,603	5,516	7,127	5,955
GO-annotated (%)	33.3	34.3	30.8	26.4
<b>KEGG</b>				
KO-annotated	2,952	2,629	3,182	2,705
KO-annotated (%)	14.9	16.4	13.7	12.0
Kbr-annotated	2,848	2,537	3,070	2,612
Kbr-annotated (%)	14.4	15.8	13.3	11.6
Kmd-annotated	565	483	575	498
Kmd-annotated (%)	2.9	3.0	2.5	2.2

### 4.1.3 Differential expression analysis

When comparing expression levels of DEGs, replicates from the same pH condition clustered together in all strains, indicating that the expression of DEGs between replicates of the same pH condition was more similar than compared to samples from



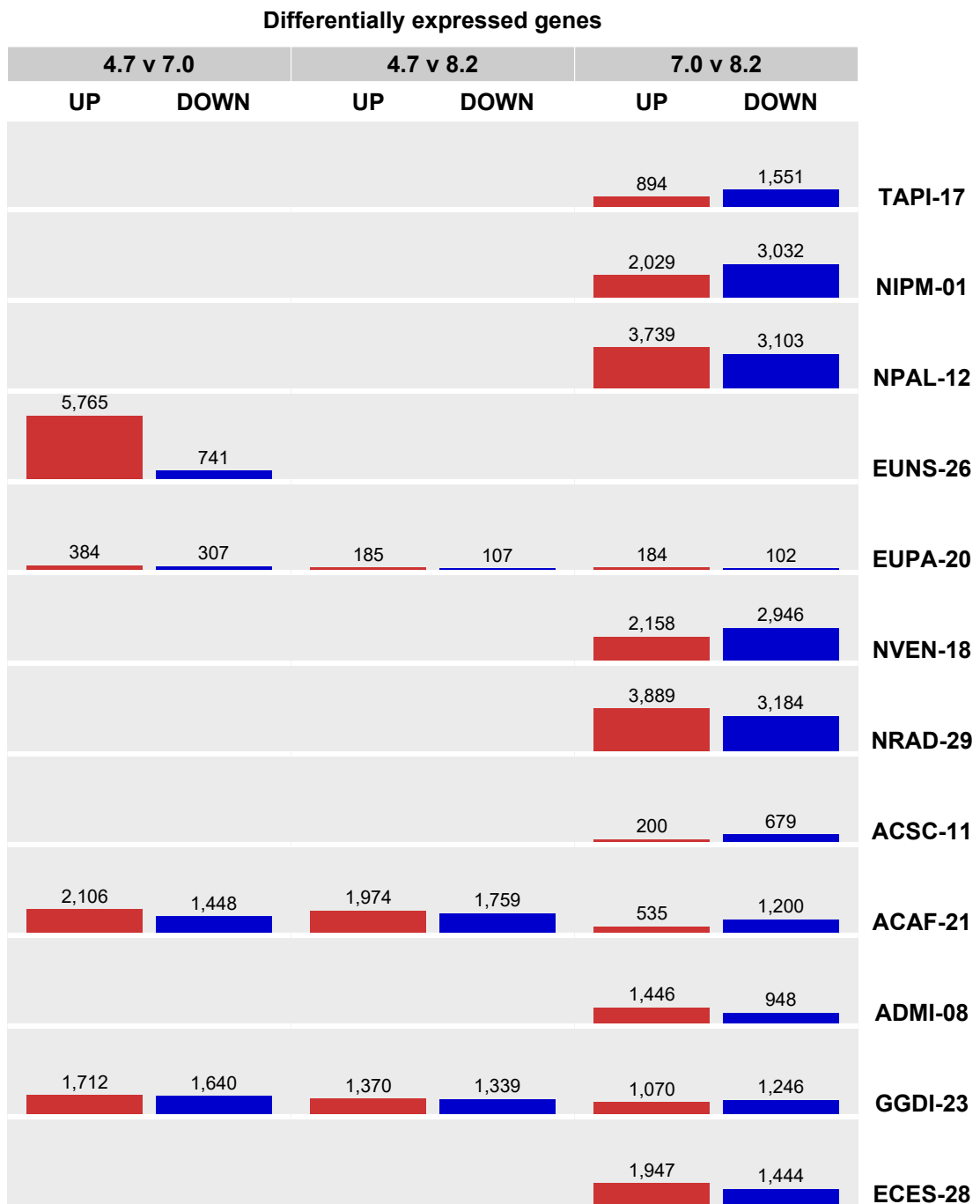


**Figure 4.2. Phylogenetic tree of the twelve studied strains.** The tree was inferred from gene trees of shared orthogroups using OrthoFinder and was manually rooted to be consistent with the Nakov et al. (2018) tree.

another pH condition. In the three strains for which we have transcriptomic data at the three pH conditions, namely EUPA-20, ACAF-21, and GGDI-23, the expression of DEGs was more similar between pH 7.0 and 8.2 than any of them compared to pH 4.7.

Both gene upregulations and downregulations were detected in each pH comparison and strain (Figure 4.3). Generally, there was no substantial difference between the numbers of upregulated and downregulated genes. Only in three cases the difference between the number of upregulated and downregulated genes was more than double: in EUNS-26, the number of upregulated genes at pH 4.7 compared to pH 7.0 was more than seven times higher than the downregulation, whereas in ACSC-11 and ACAF-21, the number of downregulations at pH 7.0 compared to pH 8.2 was 3.4 and 2.2 times higher than upregulations, respectively.

In the comparison between pH 7.0 and 8.2, the two *Nitzschia* and the two *Navicula* strains had the highest number of DEGs, with more than 5,000 genes in each strain, while in EUPA-20 and ACSC-11, the number of DEGs was the lowest, with less than 900 genes (Figure 4.3). A low amount of DEGs was also detected for EUPA-20 in both comparisons involving pH 4.7, yet the number was higher in the comparison between pH 4.7 and 7.0. ACAF-21 and GGDI-23 had a much higher number of affected genes in the comparisons involving pH 4.7 than EUPA-20. In fact, in both ACAF-21 and GGDI-23, more genes were differentially expressed at pH 4.7 than between pH 7.0 and 8.2, particularly in ACAF-21.



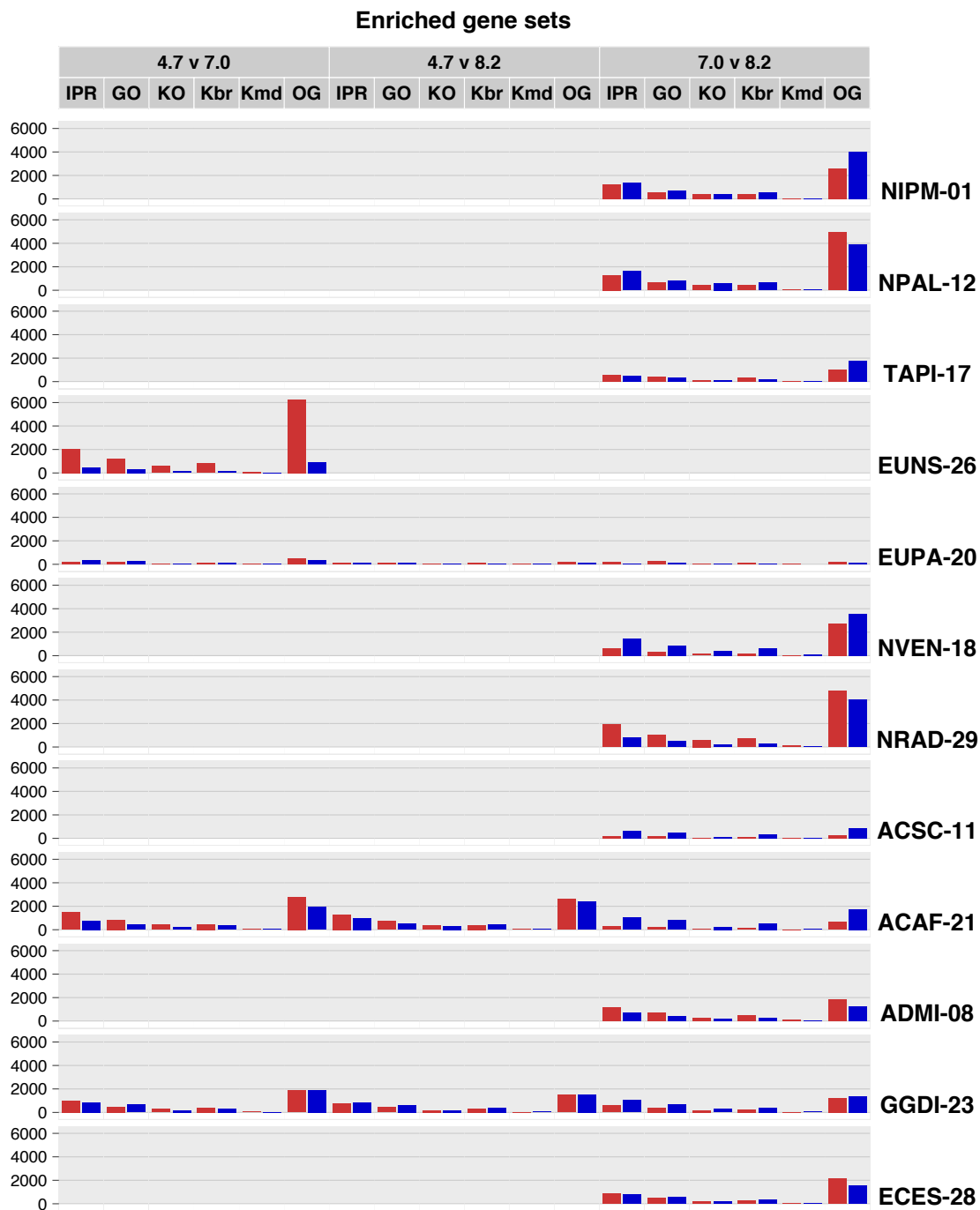
**Figure 4.3. Number of differentially expressed genes per comparison between pH conditions in the twelve studied strains.** For each contrast, genes were considered differentially expressed if  $FDR \leq 0.01$ .  $n = 3$  replicates per condition except in EUPA-20 at pH 4.7, with  $n = 2$ .

#### 4.1.4 Enrichment analysis

In line with the outcome from differential expression analysis, enriched gene sets were detected in each pH condition in all pH comparisons and strains (Figure 4.4). For each contrast, the proportion between sets enriched at each pH resembled the proportion found for DEGs, with contrasts showing a relatively similar amount of enriched terms in each pH condition. Exceptionally, ACAF-21 and ACSC-11 had more sets enriched

at pH 8.2 than at pH 7.0, and EUNS-26 had more sets enriched at pH 4.7 than at pH 7.0. The gene set type with the highest number of enriched terms for all contrasts was the orthogroups, which also had the greatest number of sets.

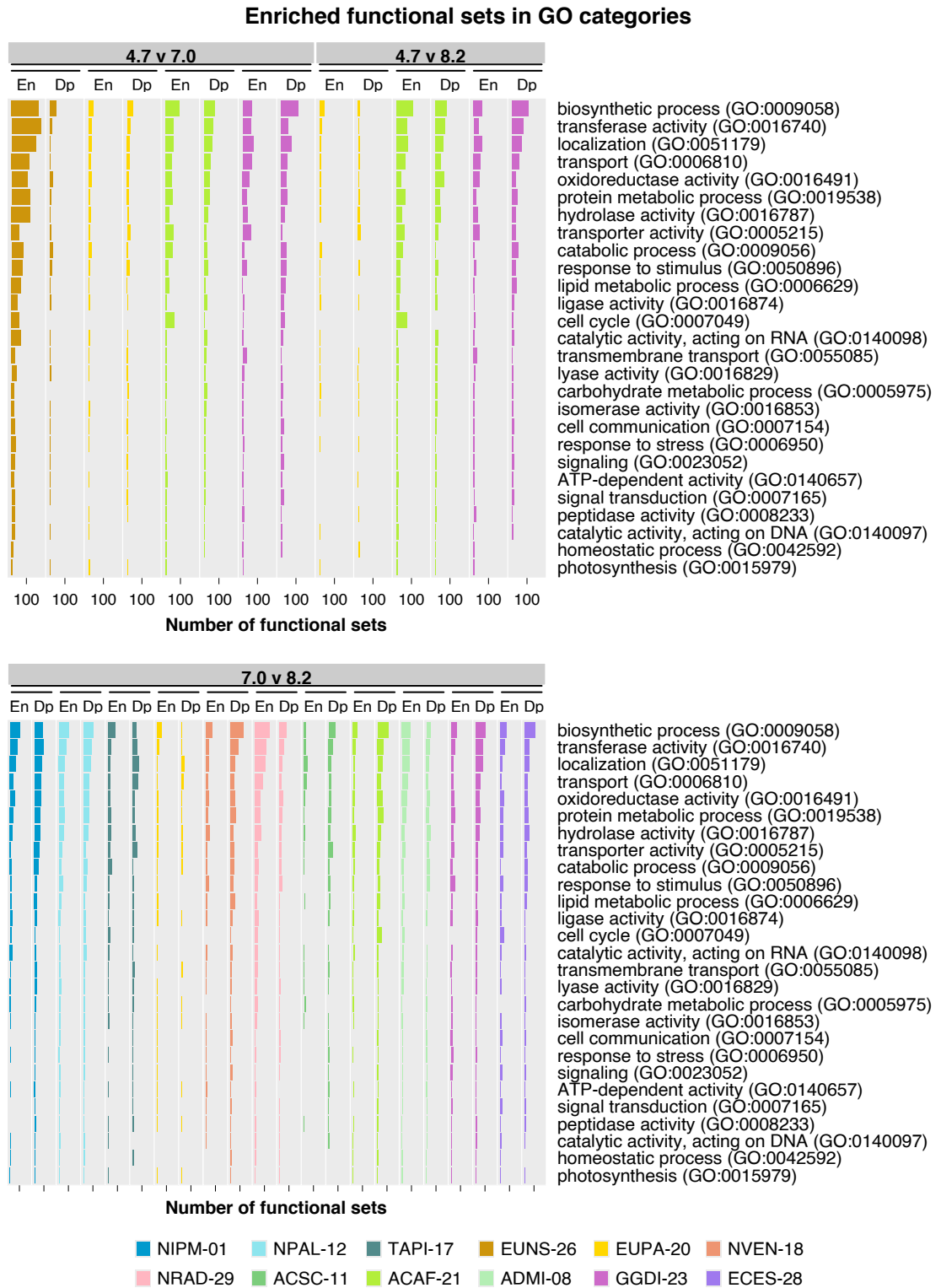
Gene Ontology and KEGG BRITE are two databases with a hierarchical structure among their entries (Ashburner et al., 2000; Consortium et al., 2023; Kanehisa, 2000; Kanehisa et al., 2022). Therefore, one can see which general terms encompass the



**Figure 4.4. Number of enriched and depleted gene sets per set type and pH comparison in the twelve studied strains.** For each contrast, sets were considered enriched or depleted if  $FDR \leq 0.01$ . A gene set was considered enriched when the enrichment was significant in at least one of the three methods (FCS, ORA, and UGS), and the same for depletions.

## 4.1. Results

largest number of affected entries. In this study, many GO sets involved in biosynthetic processes were enriched in each pH for all three pH comparisons (Figure 4.5). Catabolism was also notably affected, yet the number of associated enriched GO



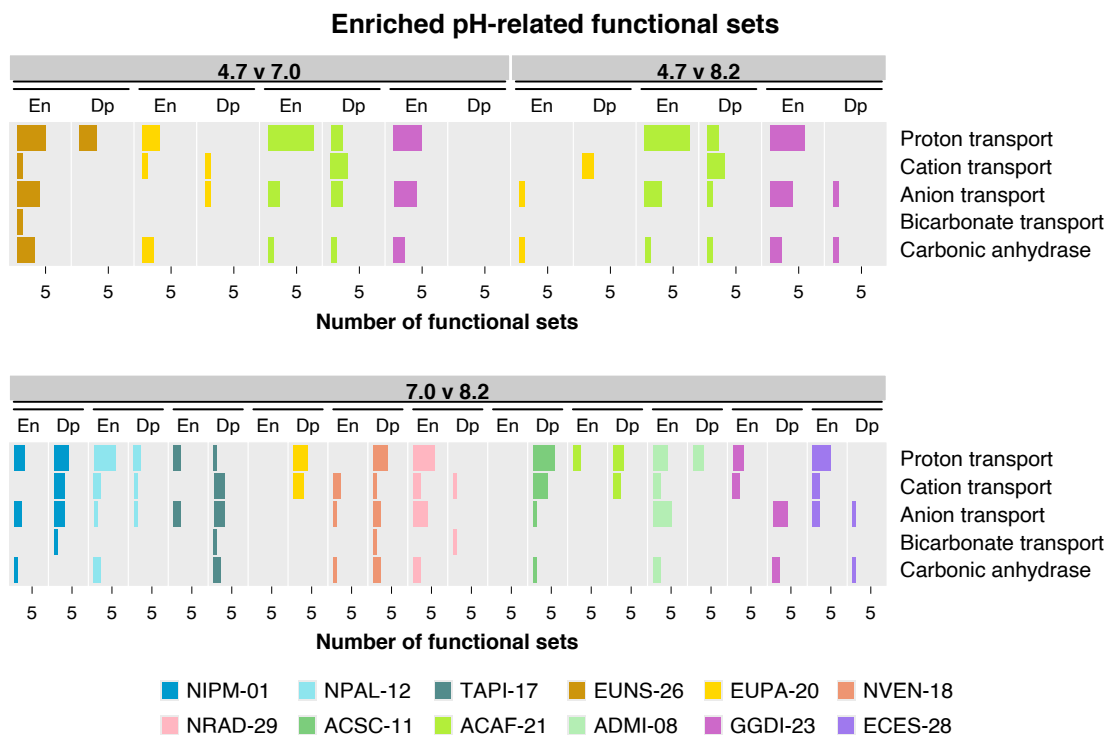
**Figure 4.5. General Gene Ontology categories with their number of enriched terms.** GO categories are sorted by the decreasing number of enriched terms across all strains. A GO term was considered enriched if it was significant in either of the three analyses: FCS, cORA, and UGS. For FCS and cORA, the significance threshold was  $FDR \leq 0.01$ .



many enriched GO terms participating in responses to stress or signal transduction. Other GO categories with a considerable number of enriched GO sets include lipid metabolic processes, cell cycle, and carbohydrate metabolic processes.

Consistent with enriched GO categories, membrane trafficking, transporters, and signal transduction were among the most affected KEGG BRITE categories in all contrasts (Figure 4.6). Proteins associated with the chromosome and proteins participating in DNA repair and recombination or in mRNA biogenesis showed a large number of enriched KEGG BRITE sets. Mitochondrial and ribosome biogenesis also contained a great amount of enriched KEGG BRITE sets, as well as peptidases and inhibitors and the ubiquitin system.

Some functional sets that could be directly related to pH homeostasis were examined. These gene sets were classified into five functional categories, including proton bicarbonate, and other cations and anions transport, and carbonic anhydrases (CAs). The list of selected functional sets within these categories is shown in Figure 5.3. The enrichment pattern of sets within each functional category generally varied across strains (Figure 4.7 and Figure 5.3 in Chapter 5). The cation and anion transport, including proton transport, were affected in most strains and pH comparisons. Proton-transporting V-type ATPase was enriched at pH 7.0 compared to pH 8.2 in many strains and depleted in the *Eunotia* EUPA-20. In contrast, bicarbonate transport



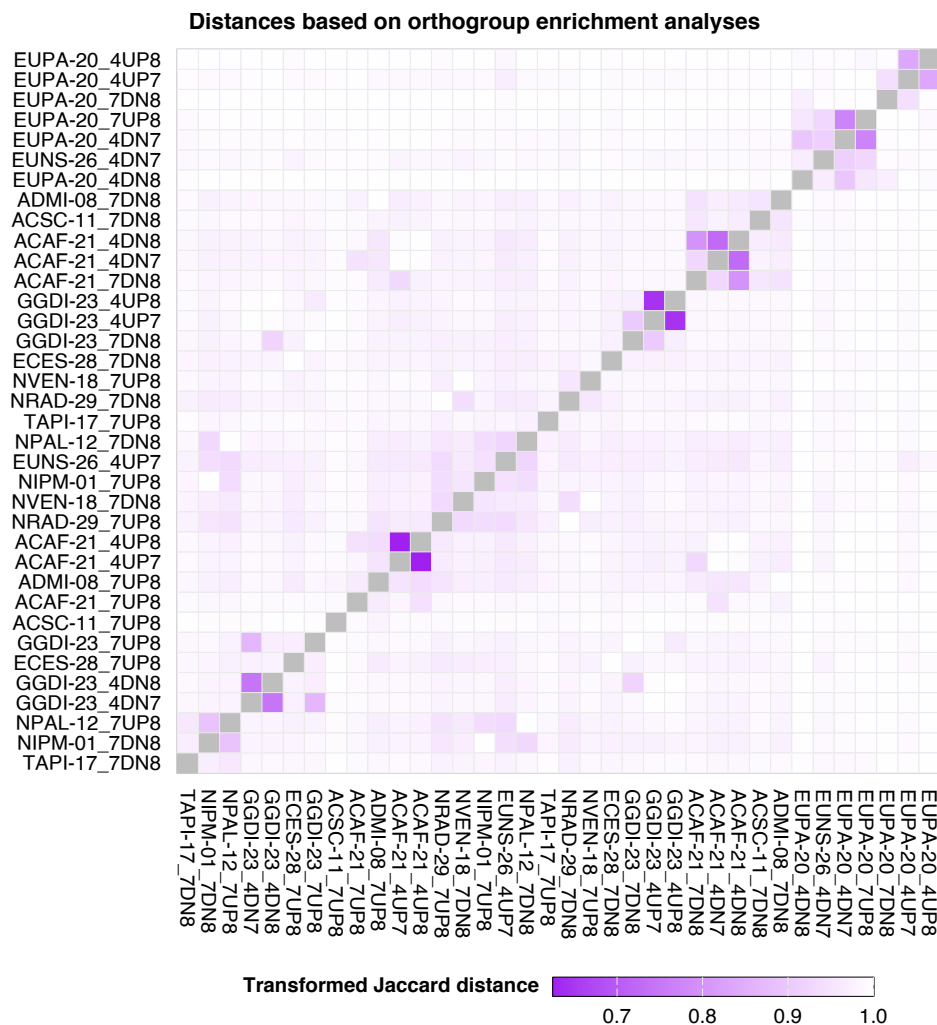
**Figure 4.7. Number of affected gene sets in the selected pH-related functional categories.** A gene set was considered enriched (“En”) or depleted (“Dp”) if it was significant in either of the three analyses: FCS, cORA, and UGS. For FCS and cORA, the significance threshold was  $\text{FDR} \leq 0.01$ . Note that only the transport of some anions and cations was selected.

was affected only in four strains at pH 7.0 versus 8.2, but in all cases, it was depleted at pH 7.0. Bicarbonate transport was also enriched at pH 4.7 in the *Eunotia* EUNS-26. CAs were affected in all strains and pH comparisons, with contrasting enrichment patterns and distinct subcellular localizations.

### 4.1.5 Response clustering

#### Orthogroups

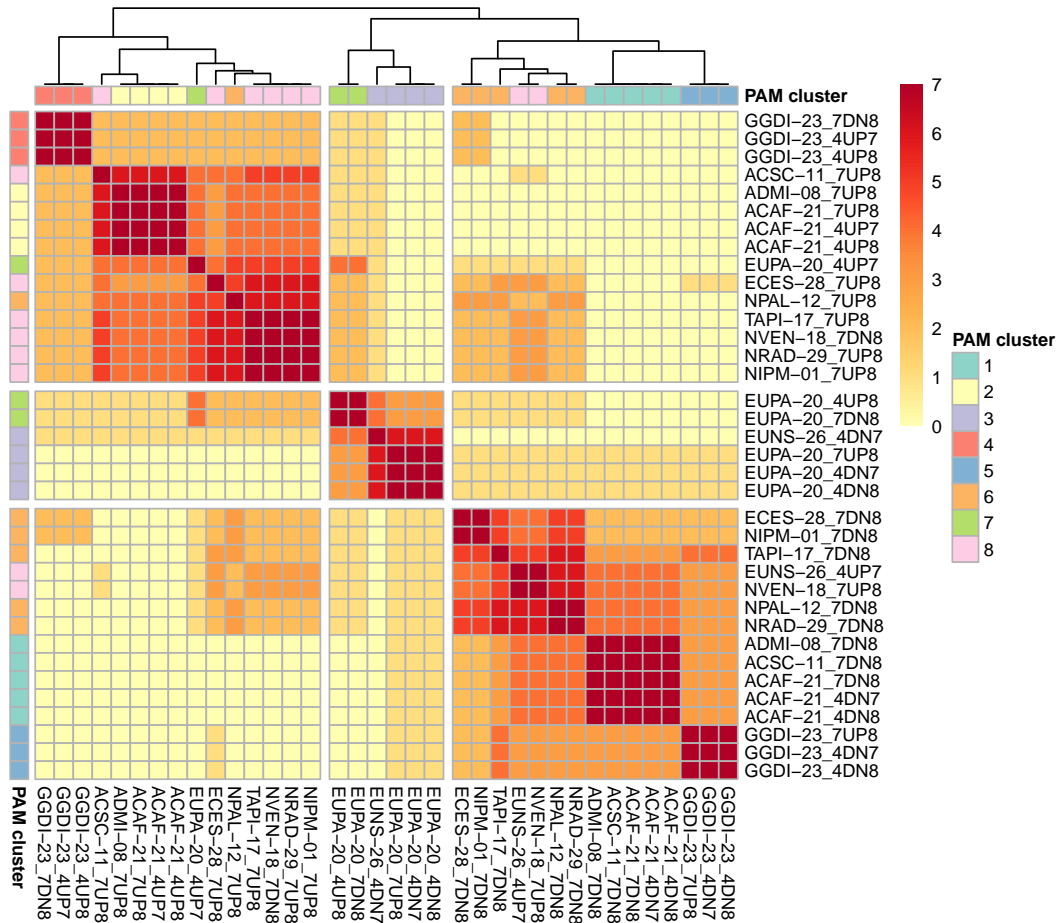
Generally, the affected orthogroups were remarkably different in each strain-specific pH contrast (Figure 4.8). Responses in orthogroups were more similar for comparisons involving pH 4.7 for the same strain and regulation direction, particularly upregulations. The most similar responses in orthogroups were the upregulations at pH 4.7 versus 7.0 and 8.2 in ACAF-21 and, independently, in GGDI-23, followed by downregulations in the same pH comparisons and for each of the same two



**Figure 4.8. Heatmap based on similarity of affected orthogroups among strain-specific pH contrasts.** Contrasts were sorted along the axes following the position in the dendrogram based on similarity (not shown). Contrasts were identified by the strain name and pH comparison. Note that, for visualization purposes, the fill gradient does not start at zero but at the minimum estimated distance.

strains. EUPA-20 upregulations at pH 4.7 versus 7.0 and 8.2 also showed a higher resemblance between them than the average. In EUPA-20, orthogroup downregulations at pH 4.7 compared to 7.0 were more similar to upregulations at pH 7.0 compared to 8.2 than to downregulations at pH 4.7 versus 8.2. Despite the low resemblance for the rest of the pairwise comparisons, phylogenetically closely related strains tended to cluster together.

The PAM algorithm and the IndVal metric were used to determine the optimal number of clusters for grouping the strain-specific pH contrasts by similarity in orthogroup regulation. Clustering our strain-specific pH contrasts in eight clusters resulted in the highest IndVal value with indicator orthogroups and at least two contrasts in all clusters. The eight clusters showed a notable phylogenetic structure, with most containing a single taxonomic group. Among the eight clusters, clusters 1 to 5 were more solid and showed more indicator orthogroups than the other four clusters. These five clusters were related to regulations at pH 4.7 in EUPA-20, ACAF-21, and GGDI-23. However, the response similarity between the contrasts involving pH 4.7



**Figure 4.9. Resemblance and robustness in clusters of strain-specific pH contrasts based on orthogroups.** The heatmap shows the number of co-occurrences in the same cluster from  $k = 2$  to  $k = 8$  between two strain-specific pH contrasts. The hierarchical clustering at the top of the heatmap groups strain-specific pH contrasts based on the number of co-occurrences.



and the pH 7.0 versus 8.2 contrast was different among the three strains: in ACAF-21, orthogroup upregulations at pH 4.7 shared more similarities to downregulations at pH 8.2 compared to 7.0 and vice versa, whereas in GGDI-23 and EUPA-20, orthogroup upregulations at pH 4.7 were more similar to upregulations at pH 8.2 in relation to 7.0 and vice versa. In *Achnanthes*, acidification-downregulated and -upregulated orthogroups are grouped in clusters 1 and 2, respectively. On the other hand, orthogroups downregulated under alkaline and particularly under acidic pH conditions are grouped in cluster 3 in *Eunotia* and in cluster 5 in GGDI-23, whereas the opposite pattern was found for EUPA-20 in cluster 7 and for GGDI-23 in cluster 4.

The number of coincidences in the same cluster from  $k = 2$  to  $k = 8$  was used to determine the similarity relationships among the eight clusters (Figure 4.9). Clusters with the same expression pattern tended to be more similar among them: acidification-stimulated orthogroups from Cymbellales, *Achnanthes*, Bacillariales or Naviculales were grouped together, as were orthogroups that were downregulated by acidification in the same strains. There were a few exceptions: in GGDI-23 and NVEN-18, orthogroups upregulated at pH 7.0 compared to 8.2 were grouped with acidification-downregulated orthogroups, and vice versa. On the other hand, *Eunotia* orthogroup upregulations and downregulations were more similar to each other than to any of the other clusters except upregulations at pH 4.7 in EUNS-26, which were most similar with NVEN-18 upregulations at pH 7.0 compared to 8.2. Upregulations at pH 7.0 compared to 8.2 in ACSC-11 were placed in cluster 8, but they were grouped with the other *Achnanthes* upregulations in cluster 2 from  $k = 2$  to  $k = 7$ .

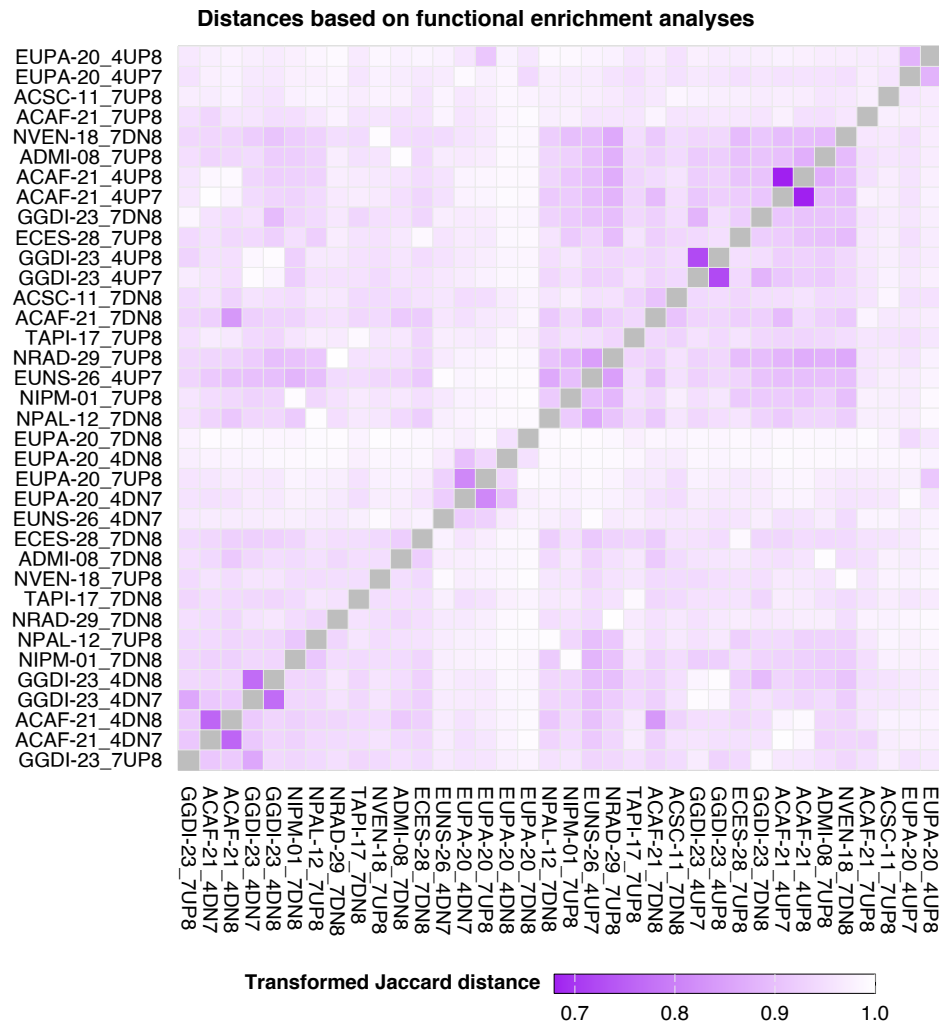
### Functional sets

Similar to the results obtained for orthogroups, there was generally a low resemblance among the affected functional sets in each strain-specific pH contrast Figure 4.10. Responses in functional sets were also more similar for comparisons involving pH 4.7 for the same strain and regulation direction, particularly for upregulations in ACAF-21 and, independently, in GGDI-23. Another resemblance with the results from orthogroup similarity is the higher similarity between orthogroup downregulations at pH 4.7 compared to 7.0 and upregulations at pH 7.0 compared to 8.2 than to downregulations at pH 4.7 versus 8.2 in EUPA-20.

The highest IndVal value was obtained when the strain-specific pH contrasts were clustered by the PAM algorithm in six clusters, considering only  $k$  with indicator orthogroups and more than one contrast in all clusters. The phylogenetic component of the six clusters was not as strong as that found for orthogroup similarity. The most solid clusters were clusters 2 to 4, whereas components of cluster 6 were more mobile across clusters depending on the  $k$ . However, clusters 1 and 6 had more indicator sets than clusters 2 and 3. Clusters 1 to 5 were related to regulations at pH 4.7 in EUPA-20, ACAF-21, and GGDI-23. In line with results from orthogroups, in ACAF-21, upregulations at pH 4.7 were more similar to upregulations at pH

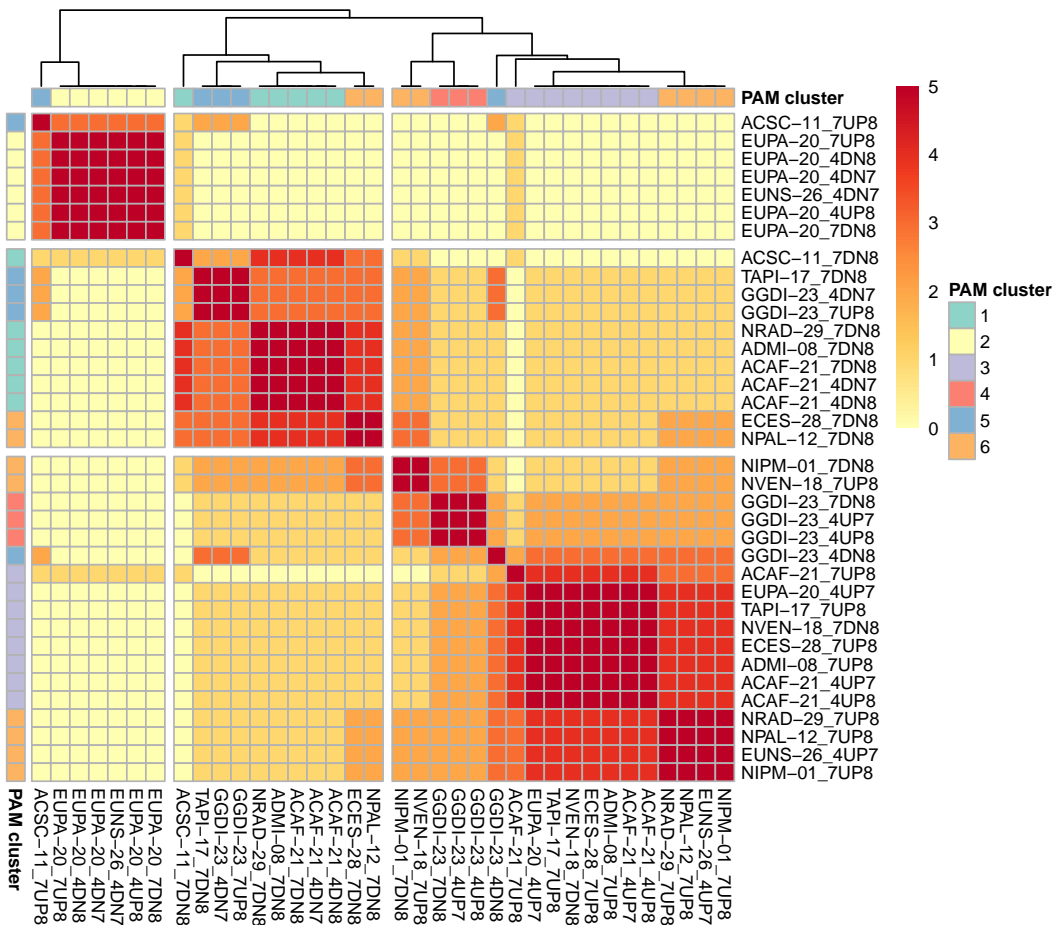
7.0 versus 8.2 and the same occurred for downregulations, whereas the opposite occurred for GGDI-23 and EUPA-20. Cluster 1 englobed acidification-downregulated orthogroups from *Achnanthidium* and NRAD-29; whereas acidification-stimulated orthogroups from *Achnanthidium*, ECES-28 and TAPI-17, together with upregulations at 4.7 versus 7.0 in *Eunotia* EUPA-20 and acidification-downregulated orthogroups from NVEN-18, formed cluster 3. Clusters 2 represented orthogroups regulated in *Eunotia*. Orthogroups upregulated at pH 8.2 and particularly at pH 4.7 in GGDI-23 are grouped in cluster 4. The opposite pattern for the same strain was found in cluster 5, together with upregulations and downregulations at pH 7.0 versus 8.2 in ACSC-11 and TAPI-17, respectively.

The hierarchical relationships among clusters based on functional sets similarity resembled those found using orthogroup similarity, where clusters with the same expression pattern tended to be more similar among them (Figure 4.11).



**Figure 4.10. Heatmap based on similarity of affected functional sets among strain-specific pH contrasts.** Contrasts were sorted along the axes following the position in the dendrogram based on similarity (not shown). Contrasts were identified by the strain name and pH comparison. Note that, for visualization purposes, the fill gradient does not start at zero but at the minimum estimated distance.

Acidification-stimulated orthogroups from all clades except *Eunotia* were generally in the same branch, whereas orthogroups downregulated by acidification in the same strains were also grouped together but in another branch. There were a few exceptions. In GGDI-23, NVEN-18 and NIPM-01 functional sets upregulated at 8.2 were clustered more frequently with acidification-upregulated sets, whereas upregulations at pH 7.0 versus 8.2 in GGDI-23 were associated more times with acidification-downregulated sets. Also, ACSC-11 upregulations at pH 7.0 in relation to 8.2 were more frequently clustered together with *Eunotia* responses. Similarly to the pattern for orthogroups, *Eunotia* upregulations and downregulations of functional sets were generally more similar to each other than to any of the other clusters. However, EUPA-20 and EUNS-26 upregulations at pH 4.7 compared to 7.0 were more similar to acidification-stimulated orthogroups from the rest of the clades.



**Figure 4.11. Resemblance and robustness in clusters of strain-specific pH contrasts based on functional sets.** The heatmap shows the number of co-occurrences in the same cluster from  $k = 2$  to  $k = 6$  between two strain-specific pH contrasts. The hierarchical clustering at the top of the heatmap groups strain-specific pH contrasts based on the number of co-occurrences.

## 4.2 Discussion

### 4.2.1 Phylogenetic relationship and niche tolerance as key factors for protein-coding genome sizes

The number and length of protein-coding genes found for our strains are similar to those found for the genome of other sequenced diatoms (Basu et al., 2017; Osuna-Cruz et al., 2020), and the phylogenetic relationships among the twelve strains were consistent with the diatom species tree inferred in Nakov et al. (2018). In this study, the number of protein-coding genes was similar among closely related strains despite the high proportion of strain-specific orthogroups. Interestingly, the *Navicula* NVEN-18 may be the sequenced diatom with the highest number of protein-coding genes to date, with a total of 43,521 protein-coding genes detected (considering genes either above or below the minimum expression threshold used). There could be even more genes in each strain, considering that the predicted proteomes were not entirely complete, according to BUSCO. In addition, protein identification in mRNA sequences was based on BLAST homology and Pfam domain content, so some identified proteins may not be actually translated or, on the contrary, some proteins with unique characteristics may not have been identified, potentially altering the number of protein-coding genes.

Looking at the twelve protein-coding transcriptomes, there was apparently no relation between the number of protein-coding genes and the pH niche width of the strain. The four strains that could grow under acidic, neutral, and alkaline pH conditions possess similar or even fewer genes and proteins than most acid-intolerant strains. This might be in line with other studies that observed a large set of cellular mechanisms responding to most stress conditions in certain organisms (Gasch et al., 2000; López-Maury et al., 2008). Applied to the present study, the capacity to thrive at acidic pH might require the activity of a small number of genes specifically targeted at low pH, while many regulated proteins at pH 4.7 could be involved in an unspecific core general response to stress. In diatoms, this core response may not be a one-size-fits-all response to stress but rather involves genes having contrasting expression patterns among different sets of stresses (Z. Li et al., 2023). Alternatively, the relation between the number of genes and the pH niche width in our strains could be masked by the influence of the phylogeny on the total number of protein-coding genes. When focusing on *Achnanthes* strains, one can see that the generalist ACAF-21 has more protein-coding genes than the two acid-intolerant ACSC-11 and ADMI-08, which is in agreement with a broader pH niche associated with a higher number of protein-coding genes.

The transcript-to-gene relationships described for EUPA-20 are apparently less complex than the other eleven strains, with fewer transcripts and proteins per gene. However, the significant fragmentation in EUPA-20 proteins and the smaller protein lengths indicate that portions of many protein sequences were missing. Hence,

transcripts and proteins from this strain are probably longer than what the data shows. Fortunately, the median protein length is not substantially smaller than those found in the other strains, meaning that the proportion of lost protein sequence is probably small. On the other hand, considering the difference in transcript length compared to other strains, even whole protein sequences could be missing if all the sequence was contained in the lost transcript fragment.

#### **4.2.2 Cellular resources are reallocated when environmental pH changes, with biosynthesis, transport, and repair playing a significant role**

In most stain-specific pH contrasts from our study, the number of upregulated and downregulated genes was not remarkably different, regardless of the change in the growth rate. Transcriptional change is one of the mechanisms used to regulate the concentration of adaptive proteins. An upregulated transcript typically leads to higher protein concentration, increasing the capacity for the cellular tasks this protein facilitates (Bruggeman et al., 2023). However, biosynthetic and physical resources are finite in cells. Hence, adaptive proteins are upregulated at the expense of proteins with a lesser essential role (Bruggeman et al., 2023; Burnap, 2015), although some protein downregulations can be direct adaptive responses to environmental change (de Nadal et al., 2011).

This distribution of resources typically results in trade-offs between the resources allocated to reproduction and survival tasks, including instantaneous growth, stress (or niche) tolerance, and capacity to acclimatize to environmental changes readily (Bruggeman et al., 2023; Burnap, 2015; Nyström, 2004). Generally, a higher growth rate is interpreted as a higher fitness and should be maximized (Bruggeman et al., 2023). In the contrasts in which the growth rate was affected, a proportion of the DEGs could be directly related to changes in the growth rate rather than responding directly to stress (López-Maury et al., 2008). The remaining proportion of DEGs would be associated with stress tolerance strategies. In the contrasts where the growth rate was not significantly affected, DEGs may reflect the reallocation among stress tolerance resources, maintaining the growth rate between both pH conditions but with different internal cellular activities based on the niche requirements. The gene expression is often more noisy for stress-related genes than for genes associated with growth (López-Maury et al., 2008), which could make detecting stress-related DEGs more difficult.

According to the functional annotation of DEGs, many enriched functions were associated with biosynthetic, location, transport and DNA-related processes across strains and pH comparisons, although each strain activated different genes and functions within these processes (see subsection 4.2.3). These results show the potential relevance of the pathways involving the biosynthesis of adaptive molecules and the transport to their target location in the cell for pH acclimatization. Also relevant

was the movement of ions across cellular membranes, the protein homeostasis, and the maintenance of DNA molecules. Most of these processes participate in the core environmental stress responses of *Thalassiosira pseudonana* Hasle & Heimdal (Z. Li et al., 2023)

#### **4.2.3 Gene families involved in plastic responses to a pH condition vary greatly among diatom species**

The twelve analyzed diatom strains possessed many strain-specific orthogroups, even when comparing phylogenetically close strains with a similar pH niche. Both the number and the proportion of strain-specific orthogroups increased with the number of protein-coding genes. This pattern was also found when comparing the genome of *P. tricornutum* with the larger genomes of *Thalassiosira oceanica* Hasle, *Seminais robusta* D.B.Danielidis & D.G.Mann, and *Synedra acus* Kützing (Osuna-Cruz et al., 2020). Young genes could have been recently acquired by gene duplication with functional divergence, functionalization of previously nonfunctional genomic sequences, gene fusions, foreign DNA acquisition, or horizontal gene transfer (Diner et al., 2017; Kaessmann, 2010). Although horizontally transferred genes have been detected in marine and freshwater diatoms in other studies, they represented a low proportion of total genes, and most belonged to older age classes (Vancaester et al., 2020). On the other hand, extensive tandem gene duplication events have been identified in the diatom *S. robusta*, leading to many species-specific gene family expansions (Osuna-Cruz et al., 2020).

The specificity of physiological mechanisms for adaptation to pH across species mentioned in subsection 4.2.2 could result from differences in their adaptive landscape topography and associated evolutionary trajectories. Together with the environment, the genetic background of each species defines its fitness landscape topography and the populations' location within it, which in turn determine the most favorable direction and the rate of adaptive evolution (Ogbunugafor & Eppstein, 2016, 2019; Payne & Wagner, 2018). The high proportion of young genes creates considerably divergent genetic backgrounds among the twelve studied strains. Young genes typically have non-essential functions (i.e., not related to growth) and low expression. These characteristics could make them more susceptible to mutation accumulation and faster evolution and, consequently, to become specialized, niche-adaptive genes (Burnap, 2015; Capra et al., 2010; Doughty et al., 2020; Osuna-Cruz et al., 2020). As a result, each strain seems to have followed a unique evolutionary trajectory toward one of their fitness peaks, resulting in the substantial dissimilarity observed among strains in their response to the same pH change. The low resemblance was observed from both homology and functional perspectives. The great proportion of functionally unannotated young genes indicates that many key niche-specialized functions in diatoms have probably not been described.

It should be noted that functional annotation was more extensive for orthogroups

present in the twelve strains and, as a result, the functional dissimilarity might be biased. A problem arising from using databases for functional annotations is that annotation is shaped by the proteins, genes, and species recorded in these databases. Gene families present only in certain species may not be annotated using these databases, in contrast with more universal gene families (Capra et al., 2010). This pattern was found in our data and, as a result, a large proportion of young orthogroups in our strains have unknown functions. Unannotated young orthogroups could be potentially relevant for pH adaptation, and some could share the same expression pattern and function in different strains, thus reducing the functional dissimilarity in the response to pH.

In the following chapters, we will focus on the physiological adaptations of the acidophiles EUNS-26 and EUPA-20 and the generalists ACAF-21 and GGDI-23 to acidic pH.

## Chapter 5

# Transcriptomic responses to acidic environmental pH

This chapter examines the enrichment patterns of four acid-tolerant diatom strains to identify the affected biological processes at low pH and determine their strain distribution. A limited number of gene sets showed the same enrichment pattern at low pH across all four strains. These gene sets included 5'-nucleotidases, GDPD domain-containing proteins, CHMP5/Vps60 and fucoxanthin-chlorophyll proteins as enrichments, and an Ankyrin repeat-containing protein as the unique shared depletion. In contrast, many gene sets displayed group- and, especially, strain-specific responses. Adaptations to acidic environmental pH entailed a wide variety of cellular functions and processes within each strain. Among them, pathways involving signal transduction, gene expression, and protein metabolism were broadly affected, with both enrichments and depletions. Adaptation to acidic pH likely requires a shift in resource allocation. This redistribution did not follow a common pattern across strains, as evidenced by the downregulation of generally distinct cellular processes among strains. These findings indicate that adaptation to acidic environments could be notably shaped by narrowly distributed mechanisms and highlight the involvement of a wide range of cellular functions in this process.

## 5.1 Results

### 5.1.1 Functional classification and distribution of enrichments and depletions in acid-tolerant strains

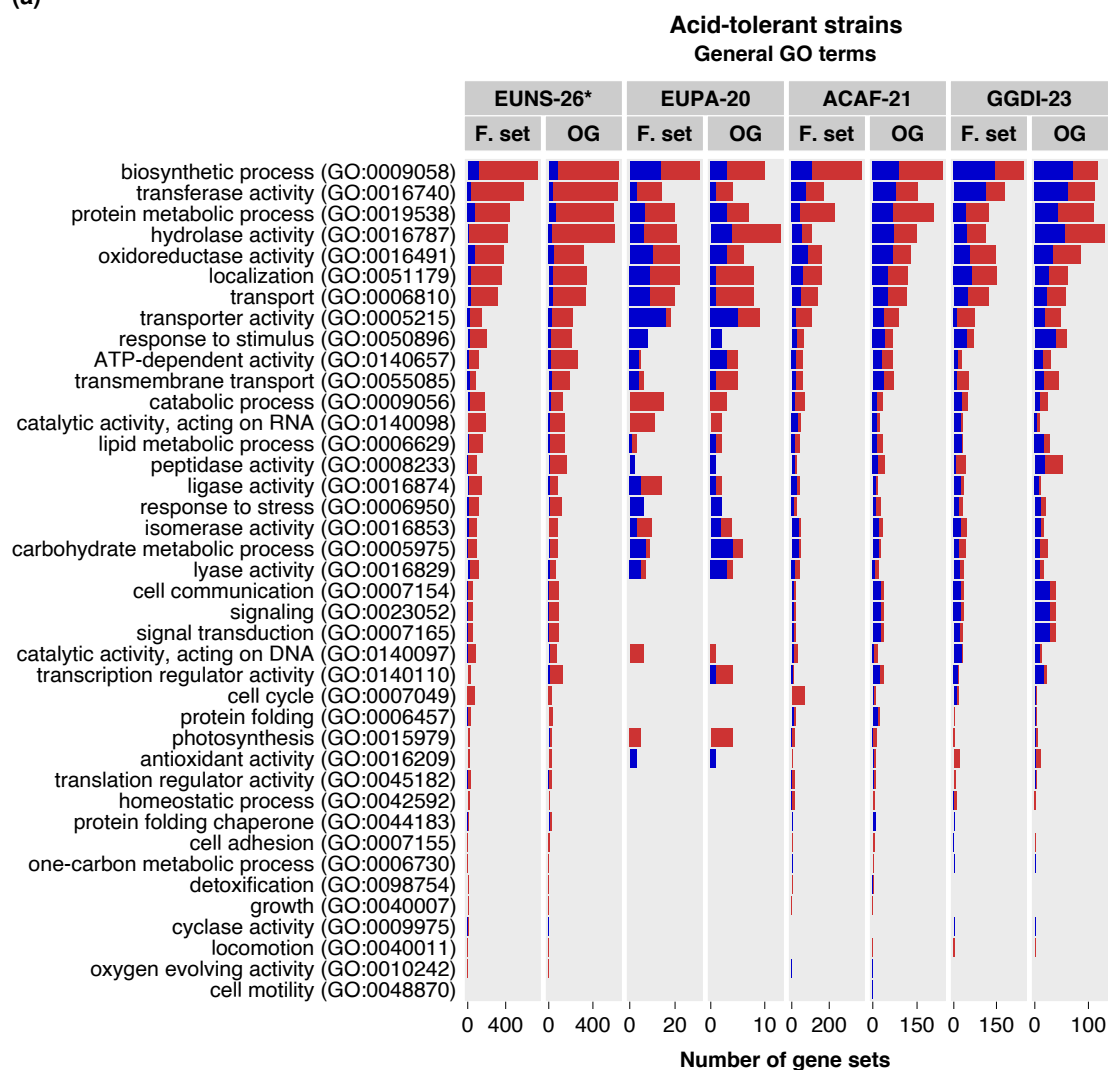
In this chapter, we will explore gene sets enriched or depleted at pH 4.7 compared to both pH 7.0 and 8.2, which will be referred to as gene sets enriched or depleted at pH 4.7 for simplicity. For a detailed description of the three enrichment methods used, see subsection 2.3.4 in Chapter 2. Four diatom strains sustained a positive growth at pH 4.7 and were therefore considered to be acid-tolerant strains (see Chapter 3), namely the two acidophilic *Eunotia* EUNS-26 and EUPA-20 and the two generalists



*Achnanthidium* ACAF-21 and *Gomphonema* GGDI-23. As there was no expression data for the pH 4.7 versus 8.2 comparison for EUNS-26, this strain was included in this chapter but with necessary considerations for the missing data.

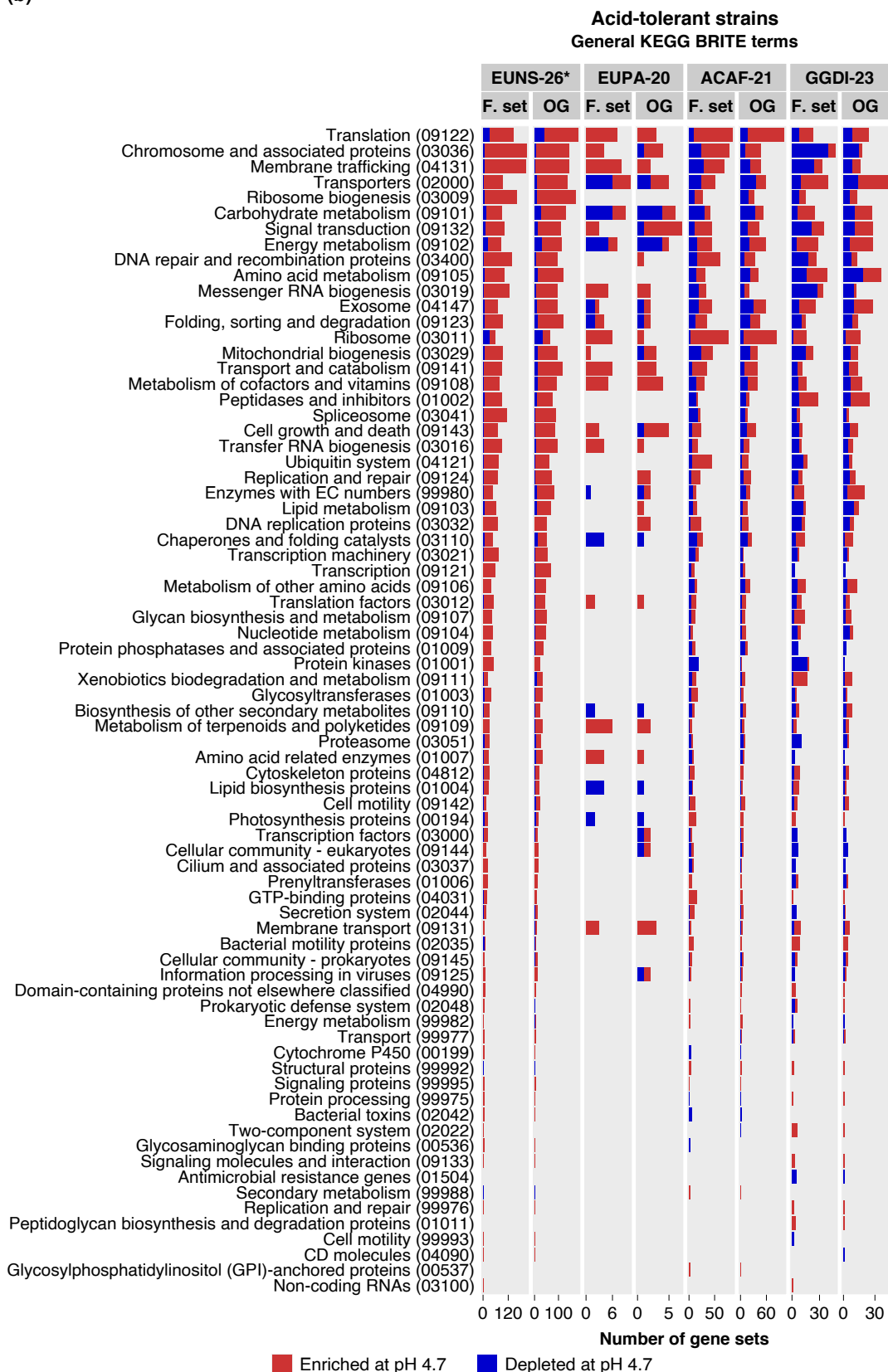
In total, 5,617 orthogroups and 6,351 functional sets were affected (either enriched or depleted) at pH 4.7 in at least one of EUPA-20, ACAF-21, and GGDI-23, representing 11.5% and 34.9% of all orthogroups and functional sets included in these contrasts and strains. Around half of the affected orthogroups had at least one InterPro, Gene Ontology, or KEGG annotation. General functional categories were selected based on high-level GO and KEGG BRITE terms to overview the functionality of enriched and depleted gene sets (Figure 5.1). 1,557 orthogroups and 3,490 functional sets affected in EUPA-20, ACAF-21 and/or GGDI-23 were included in the selected functional categories. Enrichments and depletions of EUNS-26 at pH 4.7 compared to 7.0 were also included in the plot. The most affected functional categories across acid-tolerant strains were related to the biosynthetic process, protein metabolic process, transport

(a)



## 5.1. Results

(b)

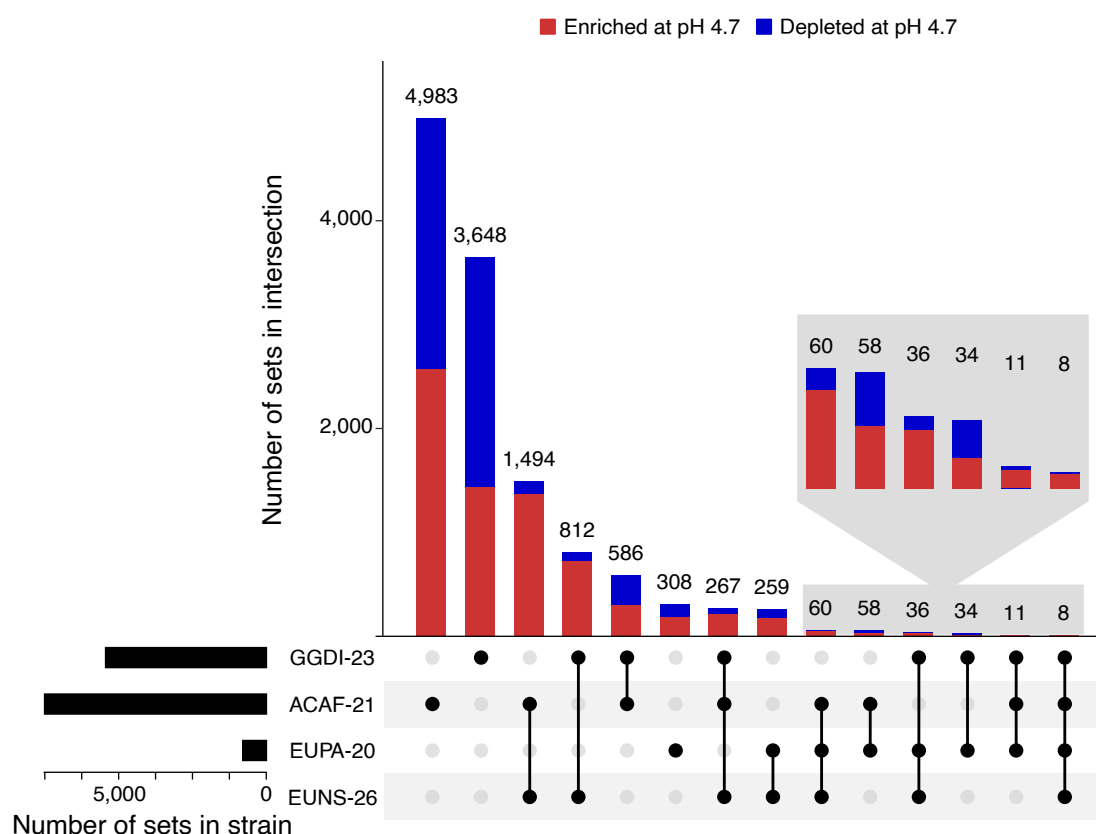


**Figure 5.1. Number of gene sets enriched or depleted at pH 4.7 per functional category in each acid-tolerant strain.** General GO and KEGG BRITE terms were selected to represent functional categories. The same gene set can be assigned to multiple functional categories. Note that the  $x$  axes of each facet have independent scales. \*For strain EUNS-26, uniquely the pH 4.7 versus 7.0 comparison was considered.

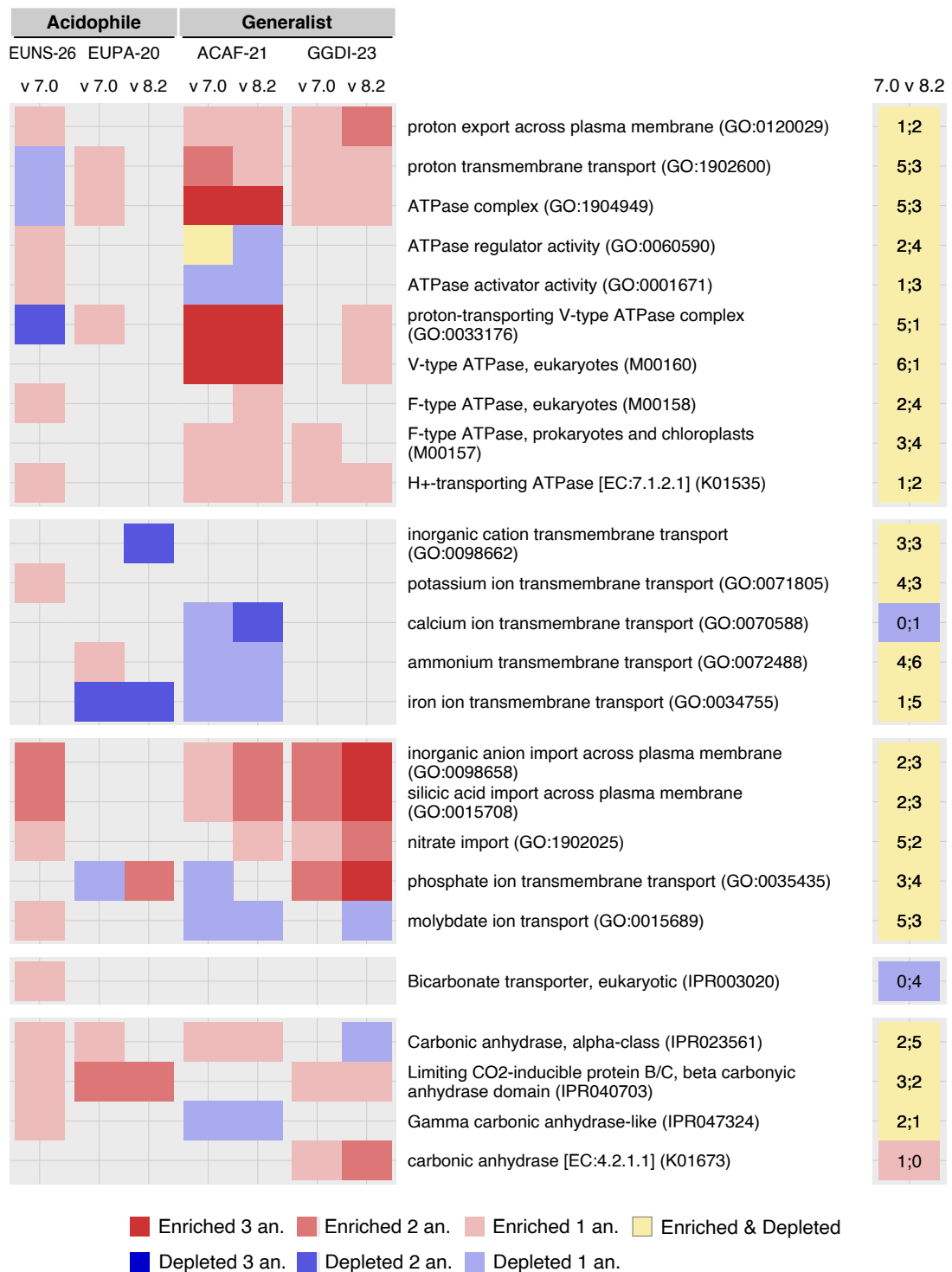
and localization, and response to stimulus. These categories had both enrichments and depletions across strains. Within transport, protein transport showed the highest number of enriched or depleted strains across strains, followed by monoatomic cation transport.

Among selected functional categories, few contained exclusively enrichments or exclusively depletions across strains. Locomotion had exclusively enriched gene sets in each acid-tolerant strain except in *Eunotia* EUPA-20, which had no affected gene set within this category. Enriched gene sets within this category were related to methyl-accepting chemotaxis proteins (MCPs). Other functional categories also exclusively contained enriched gene sets, but these categories were only affected in one or two acid-tolerant strains. No functional category contained exclusively depletions and was detected in more than one acid-tolerant strain.

The number of gene sets affected at pH 4.7 in the *Eunotia* EUPA-20 was notably smaller than in the two generalists. Most gene sets affected at pH 4.7 had a strain-specific response, particularly for both generalists (Figure 5.2). On the



**Figure 5.2. Number of gene sets enriched or depleted at pH 4.7 in the four acid-tolerant strains.** The total number of enriched and depleted gene sets at pH 4.7 in each strain is plotted on the left side except for EUNS-26, for which only intersections with other strains have been considered due to the availability uniquely of pH 4.7 versus 7.0 comparison. The bar plot of the upset plot shows the number of gene sets enriched or depleted exclusively consistently across all strains of the intersection indicated in the matrix below. The bars are colored based on the enrichment pattern.



**Figure 5.3. Enrichment pattern at pH 4.7 of selected functional sets related to pH homeostasis across acid-tolerant strains.** Gene sets belonging to similar functional categories were grouped in the same facet. The intensity of the color is proportional to the number of significant enrichment analyses (up to three, FCS, cORA, and UGS). The tile plot on the right indicates whether the gene set was enriched or depleted at pH 7.0 compared to 8.2 across strains, including acid-intolerant strains. Yellow tiles represent gene sets enriched and depleted in the same or distinct strains. The number combination in the tiles indicates the number of strains in which the gene set was enriched and depleted, respectively.

other hand, gene sets showing the same enrichment pattern exclusively within all acid-tolerant, generalist, or acidophile strains represented a small fraction of total gene sets affected at pH 4.7 in each strain. Gene sets with a shared response among all acid-tolerant strains represented at most 1.01% of total gene sets affected at pH 4.7 in each strain. Gene sets with a shared response exclusively within generalists represented 10.9% and 7.86% of total affected gene sets in *Gomphonema* GGD1-23 and *Achnanthyidium* ACAF-21, respectively; and exclusively within acidophiles, 33.6% of total affected gene sets in *Eunotia* EUPA-20.

Enrichment patterns within broad functional sets typically associated with pH homeostasis were explored, including proton transport, bicarbonate transport, other cations and anions transport, and CAs Figure 5.3. Although many gene sets were enriched or depleted at pH 4.7 in some strains, no selected functional sets shared the same enrichment pattern across strains or showed contrasting enrichment patterns consistently between acidophiles and generalists. Proton transmembrane transport, including export across the plasma membrane, proton-transporting ATPase (KEGG KO K01535), and silicic acid import across the plasma membrane were enriched at pH 4.7 in generalists, whereas the two acidophiles showed distinct enrichment patterns between them for these sets. Limiting CO<sub>2</sub>-inducible B protein (LCIB)-like proteins (InterPro IPR040703) were enriched at pH 4.7 in the two acidophiles and in the generalist GGD1-23, but they were not enriched nor depleted in ACAF-21.

In sections 5.1.2 to 5.1.5, we will explore the gene sets that showed the same enrichment pattern exclusively within all acid-tolerant, generalist, or acidophile strains. In contrast to Figure 5.2, gene sets enriched at pH 4.7 in one guild and enriched at pH 4.7 uniquely compared to either pH 7.0 or 8.2 in a strain from the other guild will not be considered in these sections, and the same will apply to depletions.

### **5.1.2 Gene sets enriched or depleted at pH 4.7 in acid-tolerant strains**

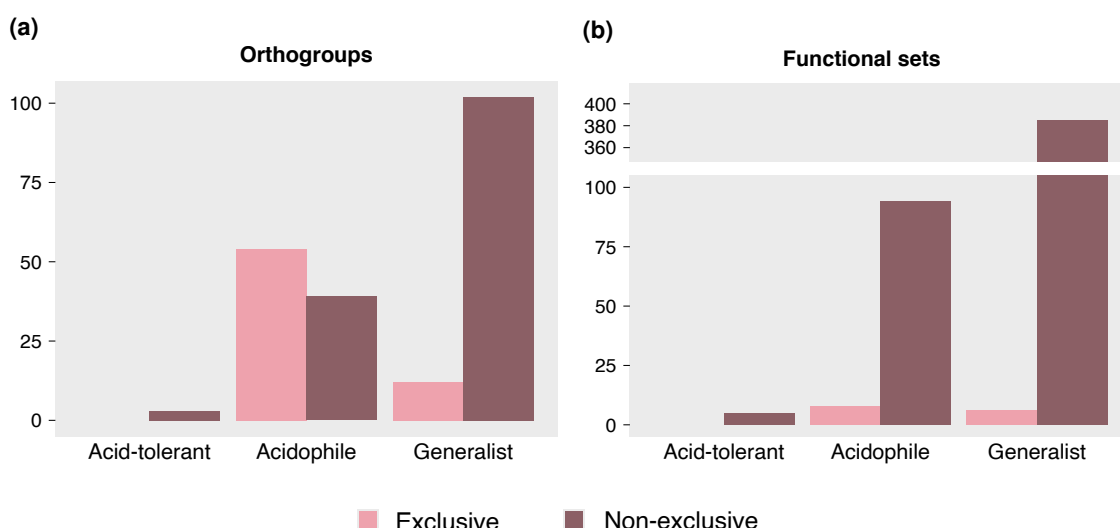
Two orthogroups and five functional sets were enriched and one orthogroup was depleted at pH 4.7 consistently in all four acid-tolerant strains (Figures 5.4 and 5.5). All of them were detected in at least one acid-intolerant strain. These gene sets were grouped into five distinct functions: 5'-nucleotidase, glycerophosphodiester phosphodiesterase (GDPD), charged multivesicular body protein 5 (CHMP5), fucoxanthin-chlorophyll protein (FCP), and an Ankyrin repeat-containing protein.

**5'-nucleotidase.** Three functional sets related to 5'-nucleotidases (InterPro IPR008334, IPR006179, and IPR036907) were enriched at pH 4.7 in the four acid-tolerant strains (Figures 5.5 and B.1). All DEGs in acid-tolerant strains annotated with these terms were upregulated at pH 4.7. In these strains, four orthogroups encoded 5'-nucleotidases, with each acid-tolerant strain containing one to three 5'-nucleotidase-encoding orthogroups. The genes upregulated at pH 4.7 belonged to distinct orthogroups among strains. In the *Eunotia* EUPA-20 and the *Gomphonema*

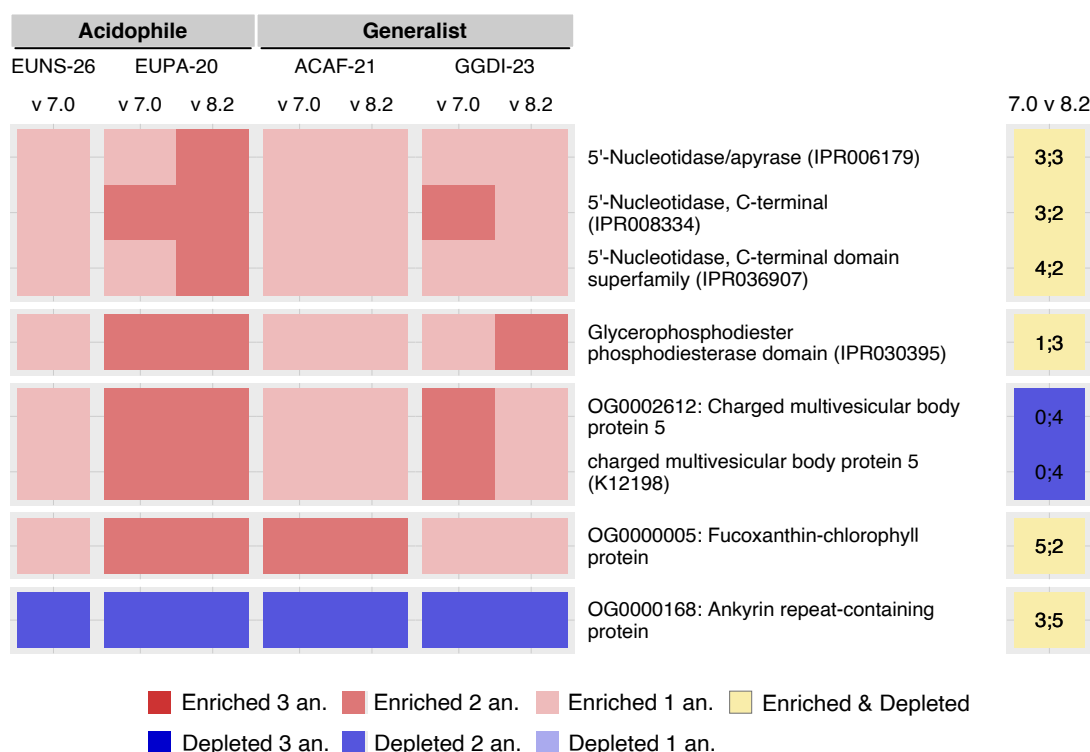
GGDI-23, affected genes belonged to OG0001672, and the proteins were predicted to contain a signal peptide and no transmembrane domains. This orthogroup was identified in the *Eunotia* EUNS-26 but not in *Achnanthidium* ACAF-21, among others. Affected genes in the *Eunotia* EUNS-26 and the *Achnanthidium* ACAF-21 belonged to OG0008409 and OG0019638, respectively, and no transit peptide nor transmembrane domain was detected for them. OG0008409 was detected in most strains with the exception of GGDI-23 and some acid-intolerant strains. OG0019638 was exclusively found in all three *Achnanthidium* strains. DEGs were detected in the pH 7.0 versus 8.2 comparison exclusively in some acid-intolerant strains, all belonging to orthogroups different from those upregulated at pH 4.7 in acid-tolerant strains and with no transit peptide identified.

### **Glycerophosphodiester phosphodiesterase domain-containing proteins.**

Glycerophosphodiester phosphodiesterase domain-containing proteins (InterPro IPR030395) were enriched at pH 4.7 in the four acid-tolerant strains (Figures 5.5 and B.2). All DEGs and DEIs in acid-tolerant strains annotated with this term were upregulated at pH 4.7. Multiple orthogroups were annotated with this term, with each acid-tolerant strain containing more than one of these orthogroups. The gene from OG0005209 was upregulated at pH 4.7 in the acid-tolerants ACAF-21, GGDI-23, and EUPA-20, but not in EUNS-26. Affected proteins from this orthogroup had a signal peptide and uniquely a transmembrane domain in ACAF-21. This orthogroup was detected in most of the acid-intolerant strains and predicted to be a GDPD (KEGG KO K01126) in ACSC-11. In GGDI-23, the gene from OG0154027 was upregulated at pH 4.7. OG0154027 was strain-specific and not further annotated. In EUNS-26, one



**Figure 5.4. Number of exclusive and non-exclusive gene sets enriched or depleted consistently at pH 4.7 in generalists, acidophiles, and both.** Acid-tolerant strains include both acidophiles and generalists. Exclusive gene sets are those exclusively detected within the group. Non-exclusive gene sets are those detected in at least one acid-intolerant strain. Note that the *y* axes of each plot have independent scales. For strain EUNS-26, uniquely the pH 4.7 versus 7.0 comparison was considered.



**Figure 5.5. Orthogroups and functional sets enriched or depleted at pH 4.7 consistently across acid-tolerant strains.** Gene sets belonging to the same function were grouped in the same facet. The intensity of the color is proportional to the number of significant enrichment analyses (up to three, FCS, cORA, and UGS). The tile plot on the right indicates whether the gene set was enriched or depleted at pH 7.0 compared to 8.2 across strains, including acid-intolerant strains. Yellow tiles represent gene sets enriched and depleted in the same or distinct strains. The number combination in the tiles indicates the number of strains in which the gene set was enriched and depleted, respectively.

isoform (but not the gene) from OG0002615 was upregulated at pH 4.7 compared to 7.0, and it was predicted to be a GDPD (KEGG KO K01126) with a signal peptide and no transmembrane domain. This orthogroup was identified in most of the twelve strains, including the other acid-tolerant strains. On the other hand, DEGs were detected in the pH 7.0 versus 8.2 comparison in GGDI-23 and some acid-intolerant strains, all belonging to orthogroups different from those upregulated at pH 4.7 in acid-tolerant strains except in GGDI-23, which was the same gene.

**CHMP5/Vps60.** The charged multivesicular body protein 5 (CHMP5 in mammals, Vps60 in yeast) (KEGG KO K12198) was exclusively encoded by orthogroup OG0002612, with all proteins from OG0002612 annotated with this term. Both KEGG KO K12198 and OG0002612 were enriched at pH 4.7 in the four acid-tolerant strains (Figures 5.5 and B.3). CHMP5/Vps60 was detected in all twelve analyzed strains, and had a single gene except in GGDI-23. The genes were upregulated at pH 4.7 in acid-tolerant strains except for one of the two genes in GGDI-23, which was upregulated at pH 4.7 only compared to 7.0. The gene was downregulated at pH 7.0 compared to 8.2 in EUPA-20 and some acid-intolerant strains. In GGDI-23, the gene upregulated at pH 4.7 compared to 7.0 was downregulated at pH 7.0 compared to 8.2.

**OG0000005: Fucoxanthin-chlorophyll protein.** Multiple orthogroups annotated as chlorophyll-binding proteins (InterPro IPR022796 and some also IPR001344) were affected at pH 4.7 in this study. The whole set (InterPro IPR022796) was enriched at pH 4.7 in strains EUPA-20, ACAF-21, and GGDI-23, although the latter was uniquely significant for isoforms (Figure B.4). In diatoms, chlorophyll-binding proteins are FCPs (Büchel et al., 2022; Grossman et al., 1990). Many of the affected orthogroups were regulated in a single strain. Only the orthogroup OG0000005 was enriched at pH 4.7 in both the acidophiles and the generalists (Figures 5.5 and B.4). OG0000005 had multiple gene copies in all strains except in EUPA-20, in which a single-copy gene was identified. Multiple genes from this orthogroup were upregulated at pH 4.7 in ACAF-21, and one gene was upregulated at pH 4.7 in GGDI-23. In EUPA-20, the gene was upregulated at pH 4.7 compared to 7.0, but some isoforms were also upregulated compared to 8.2. In EUNS-26, one of the genes was almost significantly upregulated at pH 4.7 versus 7.0 (DE FDR = 0.011), and two isoforms from that gene showed a significant upregulation. OG0000005 homologs were also identified in the nine acid-intolerant strains analyzed in this study, and this orthogroup was affected in the pH 7.0 versus 8.2 comparison for many of them.

Violaxanthin de-epoxidase-like 1 (VDL1), violaxanthin de-epoxidase-like 2 (VDL2) and zeaxanthin epoxidase 1 (ZEP1) play a key role in fucoxanthin biosynthesis in diatoms (Bai et al., 2022; C. Li et al., 2024). Based on the BLAST search against the *P. tricornutum* database, orthogroup OG0001490 is related to VDL1, OG0001859 is related to VDL2 and OG0007077 is related to ZEP1. Both OG0001490 and OG0001859 were annotated as violaxanthin de-epoxidase (VDE) (InterPro IPR044682 and KEGG KO K09839), and OG0007077, as zeaxanthin epoxidase (KEGG KO K09838). In addition, OG0010461 was related to VDE from *P. tricornutum* and annotated with the same InterPro term and KEGG KO as VDL1 and VDL2. OG0005979 was associated to a zeaxanthin epoxidase (ZEP)-like protein (ZEP2) from *P. tricornutum*, and OG0007233, to ZEP3. Although some genes and proteins from these orthogroups were differentially expressed at pH 4.7 compared to pH 7.0 or 8.2, none of these orthogroups was consistently enriched or depleted at pH 4.7 across acid-tolerant strains. Uniquely OG0010461 in strain ACAF-21 was depleted at pH 4.7.

**OG0000168: Ankyrin repeat-containing protein.** OG0000168 was depleted at pH 4.7 in all four acid-tolerant strains (Figures 5.5 and B.5). This orthogroup was annotated as an Ankyrin repeat (InterPro IPR002110)-containing protein and was detected in the twelve strains. The two generalists ACAF-21 and GGDI-23 contained a single-copy gene that was downregulated at pH 4.7. One gene in the *Eunotia* EUPA-20 was downregulated at pH 4.7. In the *Eunotia* EUNS-26, all three genes were downregulated at pH 4.7. Downregulated proteins with a predicted signal peptide were detected in the four acid-tolerant strains, and with no transit peptide in ACAF-21 and the two *Eunotia* strains. All affected proteins had no transmembrane domain. There were DEGs and DEIs for the pH 7.0 versus 8.2 comparison in most strains.

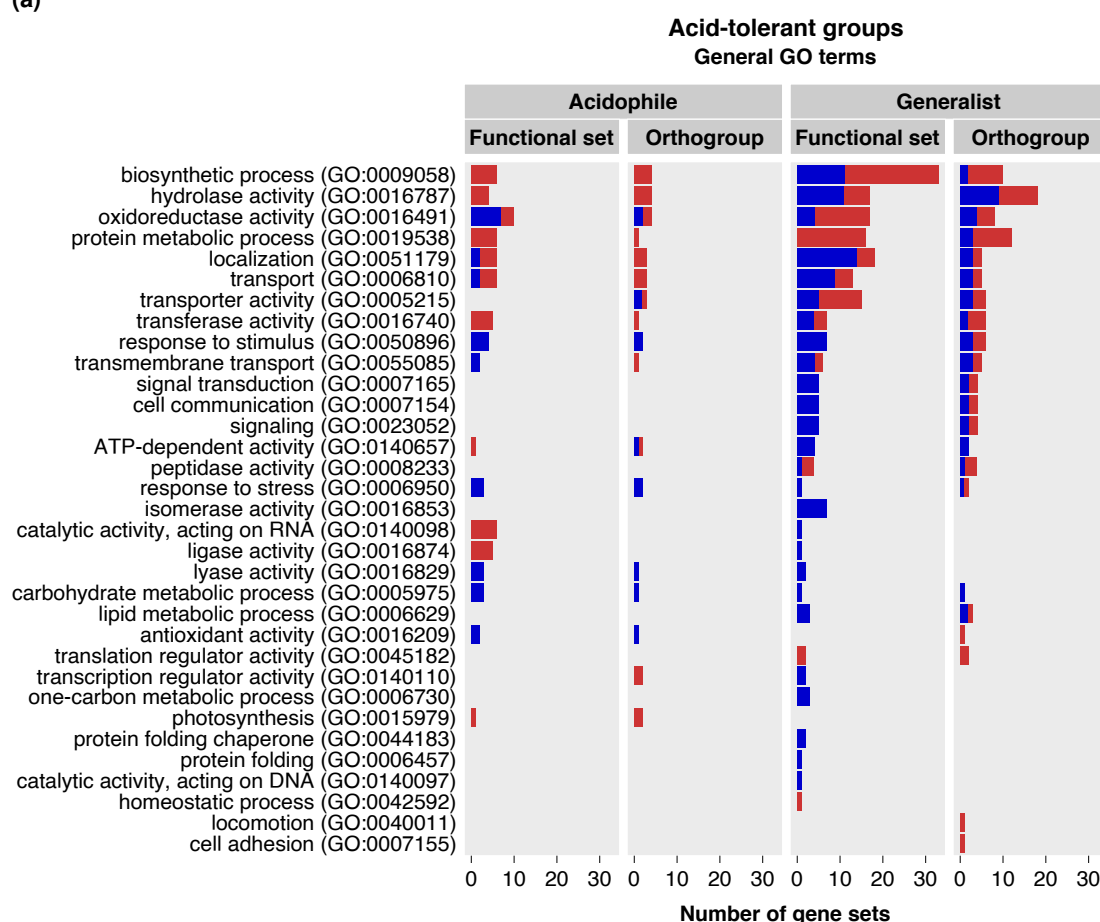


### 5.1.3 Gene sets enriched or depleted at pH 4.7 in generalists or acidophiles.

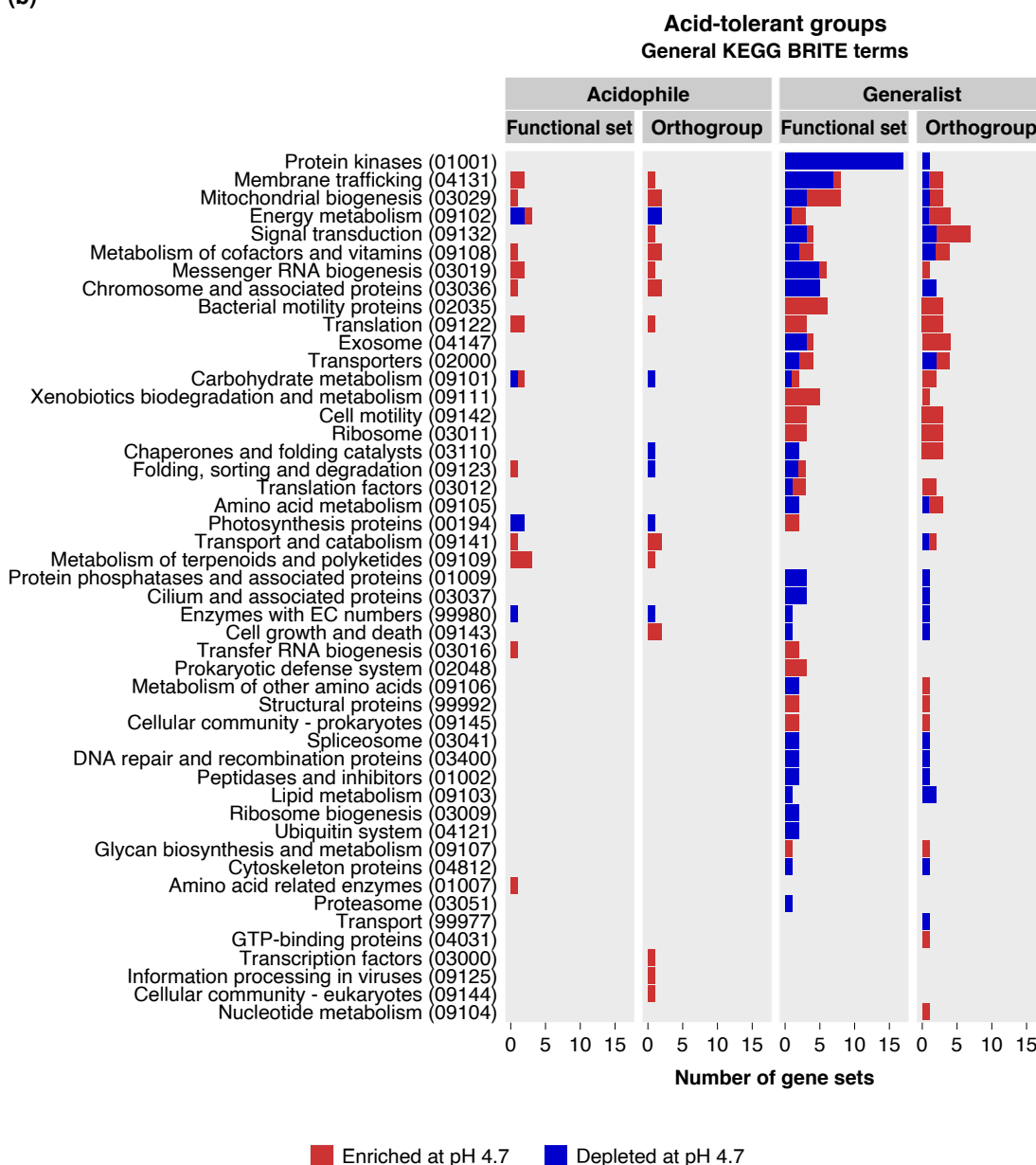
There were 78 orthogroups and 179 functional sets enriched, and 40 orthogroups and 214 functional sets depleted at pH 4.7 consistently in the two generalists ACAF-21 and GGDI-23 and not sharing the same pattern with either of the two acidophiles. Most of them were detected in at least one acid-intolerant strain (Figure 5.4), and those exclusively detected in the two generalists were all enriched at pH 4.7. 69.5% of affected orthogroups (82) had at least one functional annotation. Considering both orthogroups and functional sets, the most affected selected functional categories were the biosynthetic process, hydrolase activity, protein metabolic process, oxidoreductase activity, localization, and transporter activity (Figure 5.6).

Some selected functional categories affected in generalists contained exclusively enriched gene sets, others depleted gene sets, and the remaining both enriched and depleted gene sets. Functional categories containing exclusively enriched gene sets included cell motility, translation, ribosome, xenobiotics biodegradation and metabolism, and translation regulator activity, among others. Functional categories containing exclusively depleted gene sets included protein kinases, chromosome and

(a)



(b)



**Figure 5.6. Number of gene sets consistently enriched or depleted at pH 4.7 exclusively for generalists or acidophiles per functional category.** General GO and KEGG BRITE terms were selected to represent functional categories. The same gene set can be assigned to multiple functional categories. For strain EUNS-26, uniquely the pH 4.7 versus 7.0 comparison was considered.

associated proteins, isomerase activity, ATP-dependent activity, cilium and associated proteins, and protein phosphatases and associated proteins, among others.

As for the acidophiles, there were 65 orthogroups and 74 functional sets enriched and 28 orthogroups and 28 functional sets depleted at pH 4.7 consistently in the two acidophiles EUPA-20 and EUNS-26 and with a pattern not shared with either of the two generalists. Most of these functional sets were detected in at least one acid-intolerant strain, whereas many of the affected orthogroups were exclusively

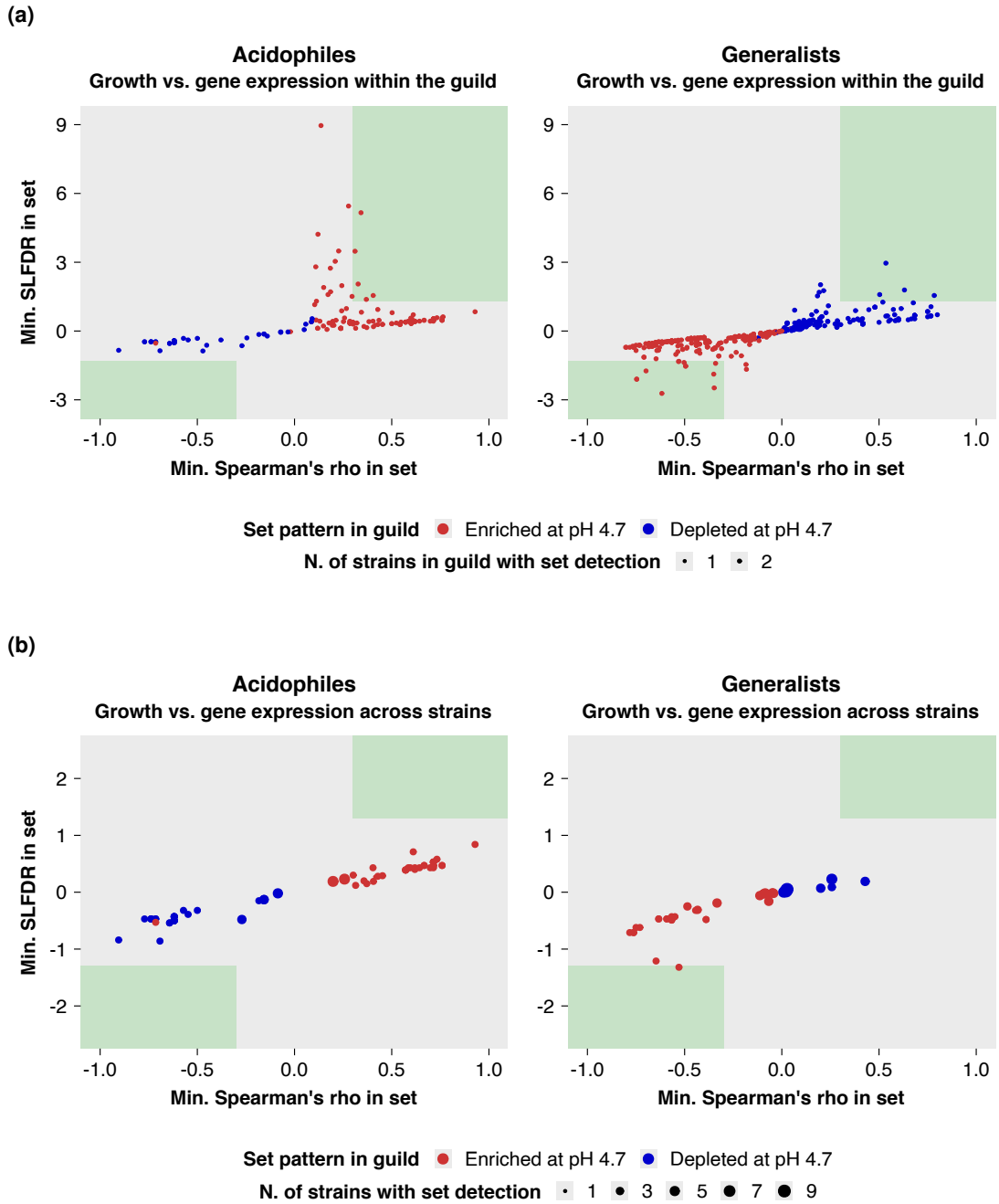
detected in acidophiles (Figure 5.4). Phosphoesterase (InterPro IPR007312) was exclusive to acidophiles and enriched at low pH. Only 35 of the affected orthogroups (37.6%) had at least one functional annotation. The selected functional categories with the highest number of affected gene sets were the oxidoreductase activity, biosynthetic process, localization, transport, hydrolase activity, and protein metabolic process (Figure 5.6).

There were functional categories exclusively containing enriched, depleted, or both enriched and depleted gene sets. Functional categories containing exclusively enriched gene sets included biosynthetic process (GO:0009058), hydrolase activity (GO:0016787), protein metabolic process (GO:0019538), catalytic activity acting on RNA (GO:0140098) and transferase activity (GO:0016740), among others. Functional categories containing exclusively depleted gene sets included response to stimulus (GO:0050896), response to stress (GO:0006950), carbohydrate metabolic process (GO:0005975), lyase activity (GO:0016829), photosynthesis proteins (KEGG BRITE 00194) and antioxidant activity (GO:0016209), among others.

#### **5.1.4 Growth rate as a confounding factor**

Gene sets enriched or depleted consistently at pH 4.7 exclusively in generalists or acidophiles could be potentially influenced by growth and not by pH because both guilds had contrasting growth patterns at pH 4.7 compared to both pH 7.0 and 8.2. A correlation analysis for each gene set and strain between the expression of its constituent genes and the strain growth rate across pH conditions was used to identify those gene sets whose expression was significantly related to growth within each guild and also across all strains in which the gene set was identified (including acid-intolerant strains with differential growth between pH 7.0 and 8.2). Gene sets were considered to be significantly correlated to growth when the strain correlations had in common the following three characteristics: the same direction, an  $FDR \leq 0.05$  and an absolute Spearman's  $\rho$  value  $\geq 0.30$ . In other words, the weaker strain correlations for each gene set had to be significant. For methodological details on the correlation analysis, see subsection 2.3.8 in Chapter 2.

Five functional sets enriched at pH 4.7 in acidophiles were positively related to growth in both strains (Figure 5.7a). These sets were calcineurin-like phosphoesterase domain, ApaH type (InterPro IPR004843); WW domain (InterPro IPR001202); clathrin/coatomer adaptor, adaptin-like, N-terminal (InterPro IPR002553); and nuclear envelope (GO:0005635) and nuclear pore complex (KEGG BRITE Nuclear pore complex). No gene set depleted at pH 4.7 in acidophiles was identified to be related to growth. Generalists contained twelve enriched and four depleted gene sets significantly related to growth in both strains (Figure 5.7a). Gene sets depleted at pH 4.7 and positively related to growth in both generalists included Tudor domain (InterPro IPR002999); proline-rich protein PRCC (K13105); aspartate decarboxylase-like domain superfamily (InterPro IPR009010); and OG0000155, which



**Figure 5.7. Significance of the correlations between gene sets expression and strain growth rate along the environmental pH gradient.** The  $x$  and  $y$  axes represent the minimum Spearman's  $\rho$  and the minimum SLFDR across considered strains for each gene set, respectively. The gene sets plotted were those enriched or depleted consistently at pH 4.7 exclusively in both acidophiles (left panels) or in both generalists (right panels) and showing the same correlation directionality between their expression and growth rate within the guild (a) or across all strains possessing the gene set (including acid-intolerant strains with differential growth between pH 7.0 and 8.2) (b). Green areas represent significant SLFDR (SLFDR  $\leq 0.05$ ) and Spearman's  $\rho$  (absolute Spearman's  $\rho \geq 0.30$ ).

was annotated as a cation efflux protein (InterPro IPR002524). Gene sets enriched at pH 4.7 and negatively related to growth in both generalists included activator of Hsp90 ATPase homologue 1-like (InterPro IPR013538); secretion system C-terminal sorting

domain (InterPro IPR026444); L-lactate/malate dehydrogenase (InterPro IPR001557); OmpA-like domain (InterPro IPR006665) and its superfamily (InterPro IPR036737); BON domain (InterPro IPR007055); protein of unknown function DUF347 (InterPro IPR007136); structural proteins (KEGG BRITE 99992); malate dehydrogenase and related metabolism (GO:0006108, KEGG BRITE 1.1.1.37 malate dehydrogenase and KEGG MODULE M00168) and mitochondrial translation factors (KEGG BRITE Mitochondrial translation factors).

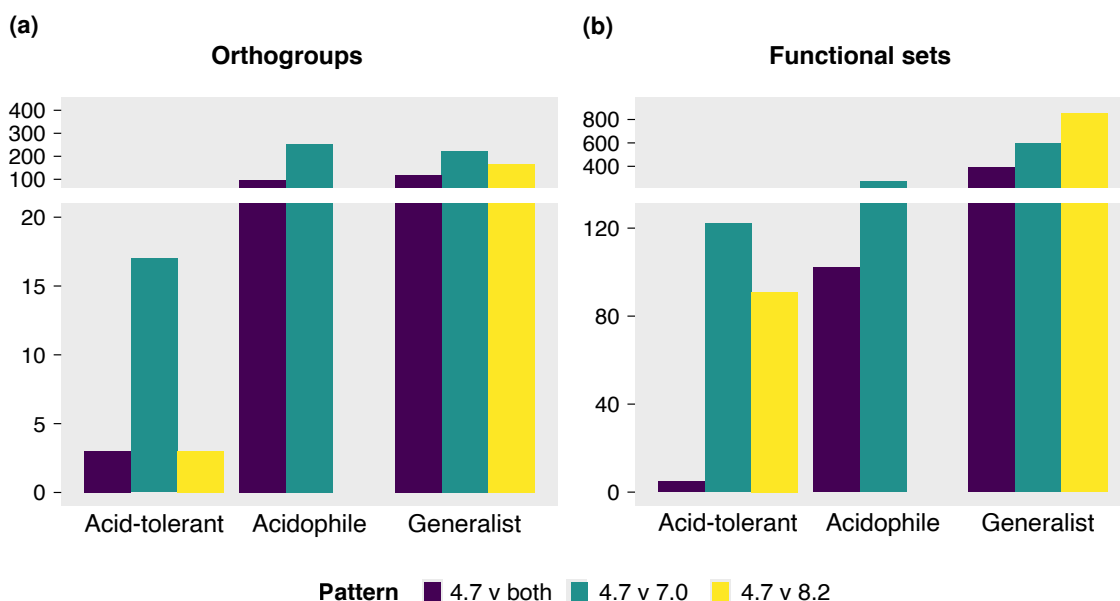
Figure 5.7b displays enriched and depleted gene sets in acidophiles or generalists that showed the same correlation direction (either statistically significant or not) between its expression and growth across all strains in which the gene set was detected, including acid-intolerant strains with differential growth at pH 7.0 and 8.2. For affected gene sets in both acidophile strains, the weaker strain correlations with growth were not significant for any of them. In addition, none of these enriched or depleted gene sets was detected in more than six strains, most detected exclusively in both acidophiles. For generalists, no affected gene set showed significant correlations between its expression and growth in all containing strains except for the InterPro term for the secretion system C-terminal sorting domain (InterPro IPR026444), which was negatively related to growth but was uniquely detected in both generalists. Many of the affected gene sets in generalists showing the same correlation direction of its expression with growth across strains were detected exclusively in the two generalists.

### **5.1.5 Enrichment patterns at pH 4.7 based on the significant acid-tolerant group and pH comparison**

Enrichment patterns at pH 4.7 were examined across strains, and the two pH comparisons were used to identify shared patterns among acid-tolerant groups. Three possibilities were considered based on the combination of significant pH comparisons: gene sets enriched or depleted at pH 4.7 (compared to both pH 7.0 and 8.2), exclusively significant compared to 7.0, or to 8.2.

23 orthogroups and 218 functional sets showed the same enrichment pattern at pH 4.7 in all four acid-tolerant strains (Figure 5.8). Three orthogroups and five functional sets were enriched or depleted consistently at pH 4.7 compared to both pH 7.0 and 8.2, which were described in detail in subsection 5.1.2. There was a higher number of gene sets enriched or depleted at pH 4.7 exclusively versus pH 8.2 than to both pH 7.0 and 8.2. The highest number was identified for enrichments or depletions at pH 4.7 exclusively versus pH 7.0.

As shown in subsection 5.1.1, the generalists and the acidophiles individually showed more shared gene sets with the same enrichment pattern at pH 4.7 exclusively within the group than shared among all acid-tolerant strains (Figure 5.8). The acidophiles had exclusively in common 704 orthogroups and 822 functional sets with the same enrichment pattern. In this guild, gene sets enriched or depleted at pH 4.7 compared



**Figure 5.8. Number of gene sets enriched or depleted at pH 4.7 consistently compared to at least one of pH 7.0 and 8.2 in the four acid-tolerant strains and per guild.** Gene sets have been counted as enriched or depleted in a particular group when they showed the same response pattern in all strains within the group and different from the responses in all strains outside the group. Acid-tolerant strains include both acidophiles and generalists. Note that the *y* axes of each plot have independent scales.

to both pH 7.0 and 8.2 were less abundant than those enriched or depleted at pH 4.7 exclusively versus 7.0. The pH 4.7 versus 8.2 comparison was not depicted for the acidophiles since transcriptomic data at pH 8.2 was only available for EUPA-20. As for the generalists, 502 orthogroups and 1,843 functional sets showed a common enrichment pattern at pH 4.7 exclusively within this group. Gene sets enriched or depleted at pH 4.7 compared to both pH 7.0 and 8.2 in this guild were less abundant than those enriched or depleted at pH 4.7 compared exclusively to either pH 7.0 or 8.2.

## 5.2 Discussion

### 5.2.1 Adaptations to acidic environments can follow many molecular pathways

The number of gene sets enriched or depleted at pH 4.7 consistently among all acid-tolerant strains was very limited, especially the depletions. The number of group- and especially strain-specific enrichments and depletions at pH 4.7 was notably higher, and also more evenly distributed between enrichments and depletions. These results suggest that adaptive proteins to pH 4.7 could be mainly narrowly distributed, particularly at the species level, although a more extensive proteome annotation is required to confirm it. Increasing the expression of adaptive proteins requires the redistribution of the finite resources that the cell possesses (Bruggeman et al., 2023; Burnap, 2015; Nyström, 2004). The scarcity of depleted gene sets in extended shared

responses but its commonality in specific responses seems to indicate that distinct pathways and cellular functions were downregulated in each strain to compensate for higher investment in adaptive proteins required to grow at acidic pH. Based on the number of enriched and depleted gene sets, this resource allocation redistribution is more prominent in the two generalists than in the two acidophiles.

It should be noted that enriched and depleted gene sets at pH 4.7 in GGDI-23, ACAF-21, and EUPA-20 could only be compared to enriched and depleted gene sets at pH 4.7 versus 7.0 in *Eunotia* EUNS-26 throughout the whole study. As a consequence, gene sets classified as shared responses at pH 4.7 among a strain group comprising EUNS-26 may not be significant for the EUNS-26 pH 4.7 versus 8.2 comparison and, in that case, they would not be retrieved as a shared response at pH 4.7 using our classification.

### **5.2.2 Signal transduction and adaptive proteins adjustment as potential key molecular mechanisms for adaptation to acidic pH**

Adaptation to acidic pH requires regulating many genes participating in signal transduction, gene transcription, and protein biosynthesis, metabolism, and transport. These processes are key mechanisms for algal responses to abiotic conditions, with signaling pathways activation and deactivation in response to stimuli leading to adjustments in the expression of proteins to meet the physiological needs shaped by environmental conditions (Kaur et al., 2022; López-Maury et al., 2008). Some specific functions were enriched within these processes, and some were depleted. Guild-specific adaptive proteins to acidic pH were mainly involved in the response pathway from signal transduction to protein localization as well. Besides this pathway, a myriad of other functions were also affected across strains, such as some related to lipid or carbohydrate metabolism, catalytic activity on DNA molecules, cell cycle, photosynthesis and energy metabolism, antioxidant activity, and locomotion, among others. This result implies that many aspects of cell functionality must be rearranged to grow under acidic conditions, which is likely widely extended across environmental adaptations (Brooks et al., 2011).

Despite sharing a considerable proportion of affected functional categories, categories containing exclusively enrichments or depletions were generally distinct between acidophiles and generalists. Growth was not identified as a global confounding factor in diatoms for the gene sets included within these functional categories. A small proportion of these gene sets were related to growth exclusively within each guild, and further research is needed to determine precisely how growth affects their expression. Therefore, most of these differences between guilds in their categories containing exclusively enrichments or depletions are likely related to variations in adaptation strategies to acidic pH. MCPs, which participate in chemotaxis, were the unique cellular function exclusively enriched at pH 4.7 in all acid-tolerant strains but the *Eunotia* EUPA-20, for which the enrichment was not significant. MCPs will be

discussed in more detail in chapter 6.

Common acid adaptations across all acid-tolerant strains were scarce but functionally diverse. Their putative activities will be discussed in sections 5.2.3 to 5.2.5, but precise information about their specific roles in cells is, in most cases, missing. Common enriched proteins included the 5'-nucleotidases, the GDPD domain-containing proteins, the CHMP5/Vps60, and an FCP, and the unique depleted protein was an Ankyrin-repeat-containing protein.

### 5.2.3 Phosphate may play a key role in acid adaptation across strains

5'-nucleotidases dephosphorylate 5'-ribonucleotides and 5'-deoxyribonucleotides by hydrolysis. In *P. tricornutum*, no extracellular but a membrane-bound 5'-nucleotidase was detected, presumably in the plasma membrane (Flynn et al., 1986). In the *Eunotia* EUPA-20 and *Gomphonema* GGDI-23, the affected proteins were predicted to contain a signal peptide and no transmembrane domain, which can be indicative of a secreted free protein (Erdene-Ochir et al., 2019; Fattorini & Maier, 2021), whereas the affected proteins from *Eunotia* EUNS-26 and *Achnanthyidium* ACAF-21 belong to different orthogroups, had no transit peptide nor transmembrane domain identified and could be free cytosolic proteins. These distinct types of 5'-nucleotidases could perform different cellular functions, including signal transduction, DNA repair, or cell-to-cell communication, among others (Zakataeva, 2021). Secreted 5'-nucleotidases could be involved in the regeneration of inorganic phosphate (Pi) and extracellular nucleotides, while intracellular 5' nucleotidases may be involved in regulating DNA and RNA pools. Some 5'-nucleotidases have been shown to be overexpressed under phosphorous deficiency (Alipanah et al., 2018; Chen et al., 2018).

The enzyme GDPD hydrolyzes glycerophosphodiester into sn-glycerol-3-phosphate and the corresponding alcohols. Affected proteins were generally predicted to have a signal peptide and no transmembrane domain and thus presumably to be secreted soluble proteins (Erdene-Ochir et al., 2019; Fattorini & Maier, 2021). GDPDs are highly conserved enzymes with variations in their catalytic activity and cellular roles across and within species (Corda et al., 2014), which makes it difficult to assess their specific role in adaptation to acidic pH. The overexpression of some GDPD isoforms has been linked to different stresses, particularly phosphorous deficiency (Dyhrman et al., 2012; Helliwell et al., 2021; Mehra et al., 2018). The glycerol-3-phosphate generated by GDPD can be used as a substrate for phosphatases to obtain glycerol and Pi, for phospholipid biosynthesis and remodeling, or for glycolysis or gluconeogenesis, both pathways for energy and carbohydrate metabolism (Hejazian et al., 2024).

5'-nucleotidases and GDPDs could both be related to Pi metabolism. Previous studies have linked Pi metabolism and acquisition to pH homeostasis (Eskes et al., 2017), although this association is still mainly uncharacterized. In marine diatoms, Pi uptake mostly depends on Na<sup>+</sup> (Matsui et al., 2023), but the cotransport of Pi with protons is



also extended in other organisms (Dick et al., 2014; Mimura & Reid, 2024). Altered proton gradients across the plasma membrane under acidic environmental conditions may influence phosphorous uptake, especially if this uptake is associated with proton cotransport. On the other hand, the  $\text{H}_2\text{PO}_4^-/\text{HPO}_4^{2-}$  is a common eukaryotic cellular buffer pair (Boron, 2004; Casey et al., 2009). This buffer system has a  $\text{pK}_a = 7.21$ , which is close to the typical cytosolic pH. As a result, the  $\text{H}_2\text{PO}_4^-/\text{HPO}_4^{2-}$  likely has a buffering power close to its maximal in the cytosolic pH (Boron, 2004). Uptake of Pi molecules could be useful for intracellular pH homeostasis. Besides its role in pH homeostasis, Pi is required for the biosynthesis of DNA, RNA, and cell membranes, energy metabolism, and signal transduction (Paytan & McLaughlin, 2007; Wagner, 2023). A more comprehensive study on the specific role of affected 5'-nucleotidases and GDPDs is required for a better understanding of these responses.

#### **5.2.4 FCP but not fucoxanthin biosynthesis is enriched under acidic pH conditions**

The diatom light-harvesting complexes consist of membrane intrinsic proteins that bind fucoxanthin, Chl *a*, Chl *c*<sub>1</sub>, and *c*<sub>2</sub> and, frequently, diadinoxanthin and diatoxanthin. For this reason, the diatom complexes are referred to as FCPs (Büchel et al., 2022; Grossman et al., 1990). Our results suggest that even though FCPs were generally enriched at pH 4.7, the response was mainly species-specific, at least based on homology, except for one orthogroup that was consistently enriched at pH 4.7 in the four acid-tolerant strains. The higher expression of FCPs at pH 4.7 seems not associated with a corresponding increase in fucoxanthin biosynthesis because key enzymes for this pathway were mostly unaffected.

Some diatom FCPs are upregulated with thermal (Hwang et al., 2008) and light (Büchel, 2014; Park et al., 2009; Truong et al., 2022) stresses. FCPs in brown algae were upregulated with salt and oxidative stresses (Dittami et al., 2009). In plants, chlorophyll-binding proteins also had a role in drought and salt stresses (Xu et al., 2011; Xue et al., 2024), with different genes responding to different stresses (X.-W. Li et al., 2020). Plant chlorophyll-binding proteins belong to the same family as diatom FCPs, but in a different subfamily (Engelken et al., 2011). These studies indicate that FCPs expression changes are a general response to multiple stresses, but the mechanism for this in the context of acid stress remains unknown. In diatom FCPs under acidic pH conditions, energy transfer pathways are rearranged, which results in a functional transition from light harvesting to energy quenching (Nagao et al., 2020). This transformation may be related to the response of these proteins to environmental pH 4.7.

#### **5.2.5 Acid-enriched CHMP5/Vps60 and alternative ESCRT-III filaments**

CHMP5/Vps60 belongs to the endosomal sorting complex required for transport (ESCRT) III, which is a component of the ESCRT machinery dedicated to membrane

remodeling and fission presumably on all cellular membranes, participating in functions such as endosomal intraluminal vesicle biogenesis, cytokinetic abscission, nuclear envelope remodeling or wound repair (McCullough et al., 2018; Pfitzner et al., 2023). Within the ESCRT-III complex, CHMP5/Vps60 acts as a regulatory subunit, apparently initiating CHMP5/Vps60-based filaments that recruit other ESCRT-III subunits (Pfitzner et al., 2023). These filaments do not share the same spatial and biochemical properties as typical Snf7-based ESCRT-III filaments, which could lead to a functional differentiation between both filament types. Based on our results, functions specific to CHMP5/Vps60-based filaments could be relevant in adaptation to acidic pH. Some studies have found CHMP5/Vps60 to have a role in responses to multiple stressors (Alqurashi et al., 2018; X. Ma et al., 2020; Zhao et al., 2020). In the case of low pH, membrane remodeling could help repair membrane damage produced by acid pH (Ammendolia et al., 2021; Lund et al., 2014; Schafer & Buettner, 2000; Wong-ekkabut et al., 2007).

Lastly, an Ankyrin repeat-containing protein was the unique gene set depleted consistently at pH 4.7 across acid-tolerant strains. This protein motif mediates protein-protein interactions and is widely extended in living organisms, particularly in eukaryotes, where it represents one of the most abundant repeat domains (Jernigan & Bordenstein, 2014; Kane & Spratt, 2021; J. Li et al., 2006). Due to the variety of functions different ankyrin repeat-containing proteins perform, a more detailed characterization of this orthogroup is required to understand its role in acid stress response.



## Chapter 6

# Potential acid-specific adaptations in diatoms

This chapter investigated potential adaptations specific to acidic environments in diatoms comparing gene expression data from four acid-tolerant and eight acid-intolerant strains. Four distinct collections of potential acid-specific adaptive gene sets shared among acid-tolerant strains were considered based on their exclusivity and enrichment patterns across strains and pH conditions. The “Nonexclusive+IndVal” collection included widespread gene sets in diatoms that might represent acid-specific stress response pathways or, alternatively, adaptive responses that evolved in acid-tolerant strains from a less plastic ancestral phenotype. Most of these gene sets were enriched, rather than depleted, at pH 4.7. Some gene sets from this collection involved cell division, DNA repair, or protein translation. The “Exclusive+IndVal” collection comprised gene sets that may represent inducible adaptations that are specifically essential for growth at low pH. A single function was identified in this collection, the MCPs, which was enriched at pH 4.7 and may have a role in pH-sensing. The “Exclusive+Const.T” collection might include constitutive adaptations allowing for faster acclimation to acidic environments, a reduction in regulatory costs, or the maintenance of the optimal phenotype (canalization). Lastly, the “Exclusive+Const.P” collection comprised gene sets constitutively expressed in some strains and enriched or depleted at low pH in others. The two collections including putative constitutive adaptations entailed some gene sets participating in biosynthesis, signal transduction (including chemotaxis), transport and localization, energy metabolism, phosphatases, and chaperones, among others. Most putative acid-specific adaptations could show a narrow distribution across strains, and the functions of many of these sets remain unknown.

## 6.1 Results

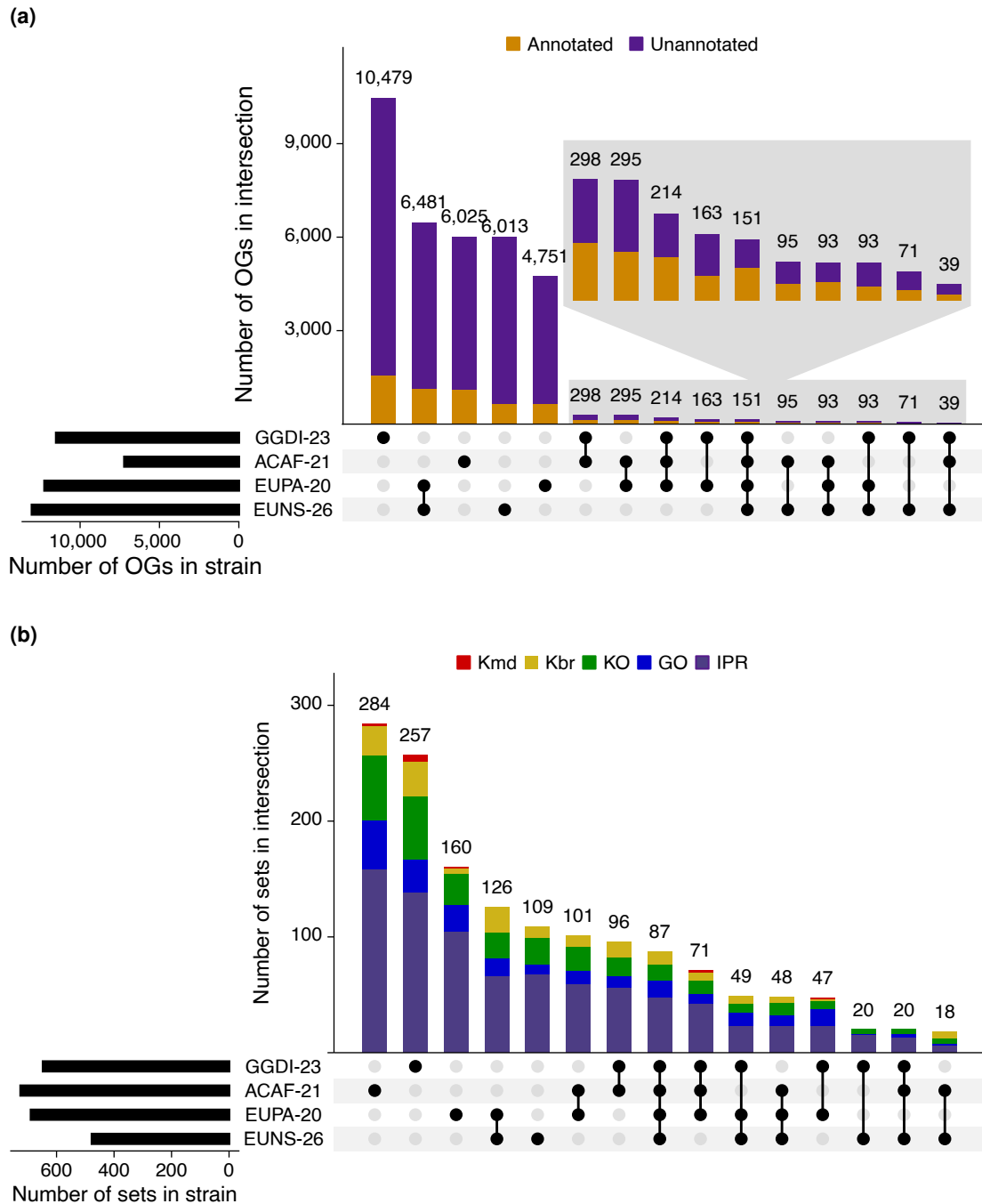
### 6.1.1 Orthogroups and functions exclusively detected among acid-tolerant strains

Gene sets exclusively detected in acid-tolerant strains were identified. A total of 35,261 orthogroups were present in at least one of the four acid-tolerant strains while being absent in acid-intolerant strains (Figure 6.1). Most of these orthogroups (77.3%) were strain-specific. Orthogroups shared exclusively between the two *Eunotia* EUNS-26 and EUPA-20 also represented a remarkable proportion (18.4%). The remaining possible combinations of strains had notably fewer shared exclusive orthogroups, including that comprising all four acid-tolerant strains. ACAF-21 shared a similar number of orthogroups exclusively with GGDI-23, with EUPA-20, and with the two strains. Compared to GGDI-23, ACAF-21 shared more orthogroups with other acid-tolerant strains but had fewer strain-specific orthogroups, leading to fewer total acid-tolerant-exclusive orthogroups in this strain. At the opposite end, the combinations of acid-tolerant strains with the least amount of shared exclusive orthogroups were those involving EUNS-26 and either one or the two generalists.

There were 1,493 functional sets present in at least one of the four acid-tolerant strains and absent in acid-intolerant strains. A considerable proportion of these exclusive functional sets were InterPro terms (56.3%). The highest number of exclusive functional sets was found for ACAF-21, GGDI-23, and EUPA-20 strain-specific sets, followed by functional sets shared exclusively by the two *Eunotia* strains and exclusive functional sets of EUNS-26. Similar to the pattern found for orthogroups, ACAF-21 shared more exclusive functional sets with EUPA-20 than GGDI-23, and EUNS-26 shared the least number of exclusive functional sets with either one or the two generalists.

### 6.1.2 Adding shared enrichment patterns to gene set exclusivity

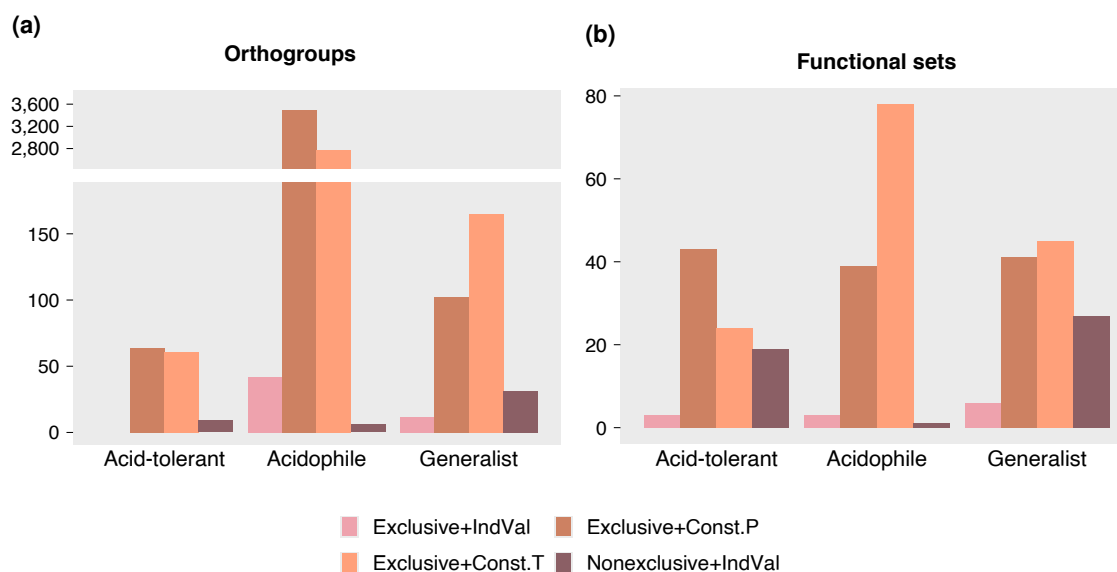
Based on their exclusivity to acid-tolerant strains and enrichment pattern at pH 4.7, gene sets were classified into four collections: 1) gene sets that were primarily enriched or primarily depleted at pH 4.7 consistently (i.e., significant based on the IndVal analysis) and exclusively detected in all four acid-tolerant strains (“Exclusive+IndVal”); 2) gene sets that were primarily enriched or primarily depleted at pH 4.7 consistently, but detected in at least one acid-intolerant strain (“Nonexclusive+IndVal”); and gene sets that were exclusively detected in all four acid-tolerant strains and were either 3) not enriched in any contrast (“Exclusive+Const subtype Total”) or 4) enriched or depleted consistently at pH 4.7 in some strains and not enriched in the others (“Exclusive+Const subtype Partial”). These four collections were non-overlapping. The same four collections were also adapted and applied to the two strain guilds within acid-tolerant strains, namely the acidophiles and the generalists, to identify guild-specific gene sets. For a more detailed description of the



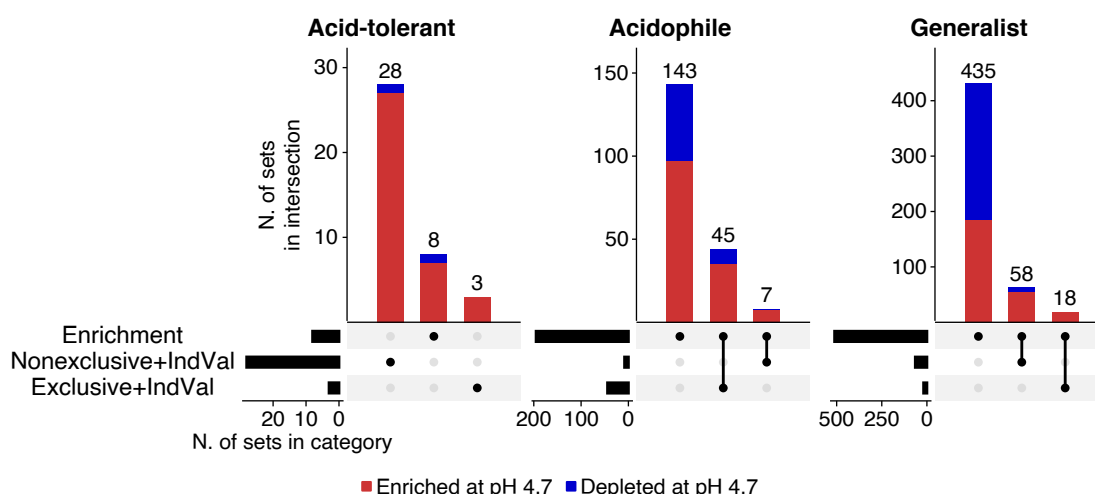
**Figure 6.1. Number of gene sets specific to acid-tolerant strains.** The total number of gene sets in each strain only present in acid-tolerant strains is plotted on the left side. The bar plot of the upset plot shows the number of gene sets found exclusively in all strains of the intersection indicated in the matrix below. For (a), the bars are colored according to the proportion of orthogroups that are functionally annotated. An orthogroup was considered functionally annotated if it was annotated by at least one of the databases used. For (b), the bars are colored according to the proportion of sets from each functional annotation database. Note that the  $y$  axes of each plot have independent scales.

methodology and classification used, see subsection 2.3.7 in Chapter 2.

The majority of identified acid-specific gene sets were classified as “Exclusive+Const”



**Figure 6.2. Number of gene sets specific to acid-tolerant strains according to IndVal analysis and their exclusivity.** Acid-tolerant strains include both acidophiles and generalists. Note that the  $y$  axes of each plot have independent scales. For strain EUNS-26, uniquely the pH 4.7 versus 7.0 comparison was considered.



**Figure 6.3. Number of gene sets in the response collections for pH 4.7.** The total number of gene sets in each response collection is plotted on the left side. The bar plot of the upset plot shows the number of gene sets found exclusively in all the response collections of the intersection indicated in the matrix below. The bars are colored according to the proportion of gene sets that are enriched or depleted at pH 4.7. Note that the  $y$  axes of each plot have independent scales. For strain EUNS-26, uniquely the pH 4.7 versus 7.0 comparison was considered.

regardless of the strain group analyzed and the type of gene set (orthogroup or functional set) (Figure 6.2). The predominance of the subtype “Total” or “Partial” varied across groups and gene set types. Most gene sets primarily enriched or depleted at pH 4.7 were detected in at least one acid-intolerant strain except for acidophiles, for which “Exclusive+IndVal” surpassed “Nonexclusive+IndVal”. Independent evaluation of the two acidophiles and the two generalist strains yielded a larger number of gene sets in each collection.

Figure 6.3 shows the gene set overlap between gene sets enriched or depleted consistently at pH 4.7 across all strains in each group (explored in detail in Chapter 5) and exclusive and non-exclusive gene sets retrieved as significant according to the IndVal for being primarily enriched or depleted consistently at pH 4.7. No gene set was consistently enriched or depleted exclusively in all seven contrasts involving pH 4.7 (i.e., two contrasts for EUPA-20, ACAF-21, and GGDI-23 and one for EUNS-26). As a result, there was no overlap between gene sets enriched or depleted consistently at pH 4.7 across all acid-tolerant strains and exclusive and non-exclusive gene sets retrieved as significant based on the IndVal analysis, with the “Nonexclusive+IndVal” collection being the most abundant with 28 gene sets. Conversely, all gene sets retrieved as significant for either the acidophiles or the generalists based on the IndVal analysis were enriched or depleted consistently at pH 4.7 across all strains in the group. For both generalists and acidophiles, “Enrichment” was the most abundant collection. For the acidophiles, exclusive and non-exclusive IndVal significant gene sets represented 26.7% of total gene sets enriched or depleted at pH 4.7 consistently in the two strains. For the generalists, the proportion was lower, at 14.9%. Within the three groups, significant exclusive and non-exclusive gene sets based on the IndVal analysis were mostly enrichments at pH 4.7 rather than depletions.

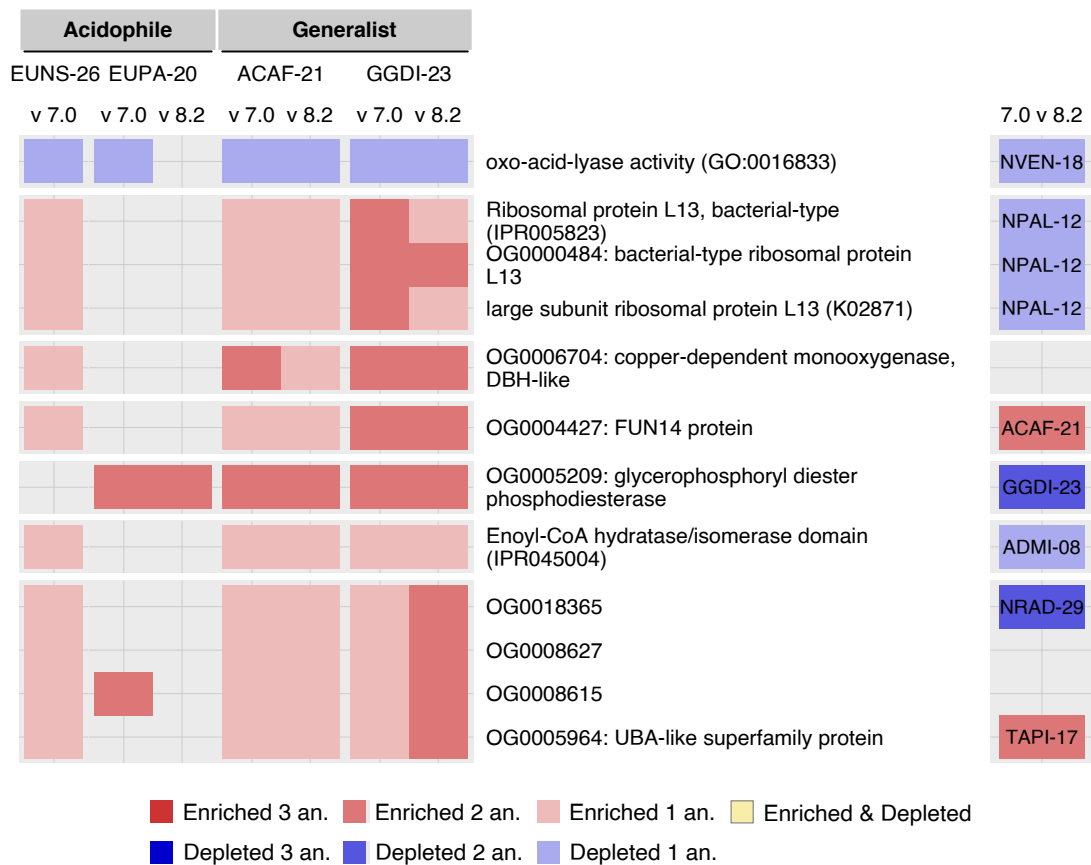
### 6.1.3 Gene sets primarily enriched or depleted at pH 4.7 in acid-tolerant strains

Nine orthogroups and 21 functional sets were primarily enriched at pH 4.7, and one functional set was primarily depleted at pH 4.7 across acid-tolerant strains (Figure 6.4).

**Methyl-accepting chemotaxis proteins.** Three functional sets were only detected in all four acid-tolerant strains and were also identified by the IndVal analysis as primarily enriched at pH 4.7: the KEGG KO K03406 and two associated KEGG BRITE terms, representing MCPs (Figures 6.4 and B.6). The InterPro IPR004089, which represents MCPs as well, was also primarily enriched at pH 4.7 but not exclusive of the four acid-tolerant strains. Some MCPs genes were upregulated at pH 4.7 in ACAF-21, GGDI-23, EUNS-26, and potentially in EUPA-20. MCPs were encoded by multiple genes and orthogroups in these four acid-tolerant strains. The only orthogroup shared among the four strains was OG0000438, which had upregulated genes at pH 4.7 in the two generalists ACAF-21 and GGDI-23. One gene from this orthogroup was almost upregulated at pH 4.7 compared to 7.0 in *Eunotia* EUPA-20 (DE FDR = 0.030). The upregulated gene in EUNS-26 belonged to OG0025945, which was only identified in the four acid-tolerant strains and had one poorly annotated upregulated gene at pH 4.7 versus 7.0 in ACAF-21. MCPs could also be present in acid-intolerant *Achnanthidium* ACSC-11 and ADMI-08: one unregulated gene in ACSC-11 was annotated as MCPs according to InterPro (InterPro IPR004089), whereas ADMI-08 had two genes annotated with KEGG KO K03406 but were below the established minimum expression threshold.







**Figure 6.4. Orthogroups and functional sets primarily enriched or depleted consistently at pH 4.7 in acid-tolerant strains.** Gene sets belonging to the same function were grouped in the same facet. The intensity of the color is proportional to the number of significant enrichment analyses (up to three, FCS, cORA, and UGS). For strain EUNS-26, uniquely the pH 4.7 versus 7.0 comparison was considered. The tile plot on the right indicates whether the gene set was enriched or depleted at pH 7.0 compared to 8.2 across strains, including acid-intolerant strains, and in which strain. Yellow tiles depict gene sets enriched and depleted in the same or distinct strains.

(GO:0051303), metaphase plate congression (GO:0051310), and attachment of mitotic spindle microtubules to kinetochore (GO:0051315).

**Type IA DNA topoisomerase.** Type IA DNA topoisomerases (InterPro IPR000380) were detected in the twelve strains, but were exclusively enriched at pH 4.7 according to the IndVal analysis (Figures 6.4 and B.8). Genes upregulated at pH 4.7 were identified in EUNS-26, EUPA-20, and ACAF-21, whereas contrasting expression patterns among distinct orthogroups were found in GGDI-23. Orthogroups OG0000902 and OG0003171 were the only type IA DNA topoisomerase orthogroups detected in all twelve strains and also the only ones annotated as DNA topoisomerase III (KEGG KO K03165). OG0000902 was upregulated at pH 4.7 in EUPA-20, and one isoform was upregulated compared to pH 7.0 in EUNS-26 and another, compared to pH 8.2 in GGDI-23. OG0003171 single-copy gene from EUNS-26 was upregulated at pH 4.7 versus 7.0, while the single-copy gene from GGDI-23 was downregulated at pH 4.7 compared to 8.2. Two orthogroups, OG0084147 and OG0090627, were not

further annotated as either type I or type III DNA topoisomerase, but the single-copy genes from these orthogroups were upregulated at pH 4.7 in ACAF-21. These two orthogroups could be exclusive of this strain. Other related terms that were also retrieved as primarily enriched at pH 4.7 by the IndVal analysis were DNA topoisomerase, type IA, central (InterPro IPR013497); DNA topoisomerase, type IA, core domain (InterPro IPR023405); DNA topoisomerase, type IA, central region, subdomain 3 (InterPro IPR013826); and DNA topoisomerase, type IA, active site (InterPro IPR023406)

**Oxo-acid-lyases.** Oxo-acid-lyases catalyze the cleavage of a carbon-carbon bond of a 3-hydroxy acid through a mechanism distinct from hydrolysis or oxidation. Oxo-acid-lyase activity (GO:0016833) was primarily depleted at pH 4.7 based on the IndVal analysis (Figures 6.4 and B.9). This activity included five different enzymes, all identified in the twelve analyzed strains: isocitrate lyase, anthranilate synthase, imidazole glycerol-phosphate synthase, hydroxymethylglutaryl-CoA lyase, and naphthoate synthase. None of them was individually retrieved as primarily depleted at pH 4.7 according to the IndVal. Isocitrate lyase (InterPro IPR006254 and KEGG KO K01637) was generally depleted at pH 4.7, but this enzyme was also enriched or depleted in some acid-intolerant strains when comparing pH 7.0 and 8.2.

**Bacterial-type ribosomal protein L13.** Mitochondrial/plastidial large subunit ribosomal protein L13 (InterPro IPR005823 and KEGG KO K02871) was mainly enriched at pH 4.7 according to the IndVal analysis (Figures 6.4 and B.10). This protein was detected in all twelve strains. Most genes with this annotation belonged to orthogroup OG0000484, which was also significant for enrichment primarily at pH 4.7. This orthogroup contained one gene in GGDI-23, two in ACAF-21, and three in the two acidophiles. One gene was upregulated at pH 4.7 in generalists ACAF-21 and GGDI-23. In EUNS-26, some isoforms of the same gene were upregulated compared to pH 7.0, whereas another gene was almost significantly downregulated (DE FDR = 0.013). No gene or isoform was differentially expressed in EUPA-20. In the acid-intolerant NPAL-12, one homologous gene was downregulated at pH 7.0 compared to 8.2.

**OG0006704, copper-dependent monooxygenase.** Orthogroup OG0006704 was retrieved as a set specifically enriched at pH 4.7 by the IndVal analysis (Figures 6.4 and B.11). This orthogroup was not found in EUPA-20, in the two *Nitzschia* strains, nor potentially in the other strain from Bacillariales, TAPI-17. The annotation assigned to this orthogroup described it as a copper-dependent monooxygenase (InterPro IPR045266). In *Gomphonema* GGDI-23, it was further annotated as a dopamine beta-hydroxylase-like protein (InterPro IPR000945). One gene from this orthogroup was upregulated at pH 4.7 compared to 7.0 in ACAF-21, GGDI-23, and EUNS-26. The same gene in ACAF-21 and one of the gene isoforms in GGDI-23 were also upregulated at pH 4.7 versus 8.2. Among acid-intolerant strains, one gene from the

two *Navicula* strains and ADMI-08 was almost significantly differentially expressed when comparing pH 7.0 and 8.2 (DE FDR < 0.05 in the three contrasts).

**OG0004427, FUN14 domain-containing protein.** Different orthogroups were annotated as FUN14 domain-containing proteins (InterPro IPR007014). One of them, OG0004427, was present in all analyzed strains but the three Bacillariales, and was specifically enriched at pH 4.7 according to the IndVal analysis (Figures 6.4 and B.12). The single-copy gene was upregulated at pH 4.7 versus 7.0 in the two generalists. For the pH 4.7 versus 8.2 comparison, the single-copy gene was only upregulated at pH 4.7 in the generalist GGDI-23, but in ACAF-21, one isoform was upregulated, and another was almost significantly upregulated (DE FDR = 0.018) at pH 4.7. The ACAF-21 isoform upregulated at pH 4.7 versus 8.2 was also upregulated at pH 7.0 compared to pH 8.2. In EUNS-26, one isoform was upregulated, but the other was almost significantly downregulated (DE FDR = 0.025) at pH 4.7 versus 7.0 in EUNS-26.

**OG0005209, glycerophosphodiester phosphodiesterase.** OG0005209 is one of the orthogroups encoding a GDPD (InterPro IPR030395 and KEGG KO K01126). This orthogroup was detected in most of the twelve studied strains, including the four acid-tolerant strains, and was identified by the IndVal analysis as preferentially enriched at pH 4.7 (Figures 6.4 and B.13). The single-copy gene from this orthogroup was upregulated at pH 4.7 in all acid-tolerant strains except in the *Eunotia* EUNS-26. The gene was downregulated at pH 7.0 compared to 8.2 in GGDI-23.

**Enoyl-CoA hydratase/isomerase domain-containing proteins.** Enoyl-CoA hydratase/isomerase domain (InterPro IPR045004)-containing proteins were found in the twelve studied strains. This InterPro term was primarily enriched at pH 4.7 based on the IndVal analysis (Figures 6.4 and B.14). One gene from OG0005092 was upregulated at pH 4.7 in ACAF-21. Another gene from another orthogroup, OG0003043, was upregulated at pH 4.7 compared to 7.0 in GGDI-23, and one isoform from the same gene was upregulated compared to 8.2. Two isoforms from EUNS-26, one belonging to OG0003043 and another to OG0005092, were upregulated at pH 4.7 versus 7.0. One gene from OG0005092 was downregulated at pH 7.0 versus 8.2 in the acid-intolerant ADMI-08.

**Poorly annotated and unannotated orthogroups.** Some poorly annotated and unannotated orthogroups were primarily affected at pH 4.7 based on the IndVal analysis (Figures 6.4 and B.15). OG0005964 was identified in most strains, including the four acid-tolerant, but was only annotated in ECES-28, as a UBA-like superfamily protein. This orthogroup had genes or isoforms upregulated at pH 4.7 in EUNS-26, ACAF-21, and GGDI-23, and at pH 7.0 compared to 8.2 in the acid-intolerant TAPI-17. One gene from EUPA-20 was almost significantly upregulated at pH 4.7 versus 7.0 (DE FDR = 0.015), and one from the acid-intolerant ECES-28, at pH 7.0 compared

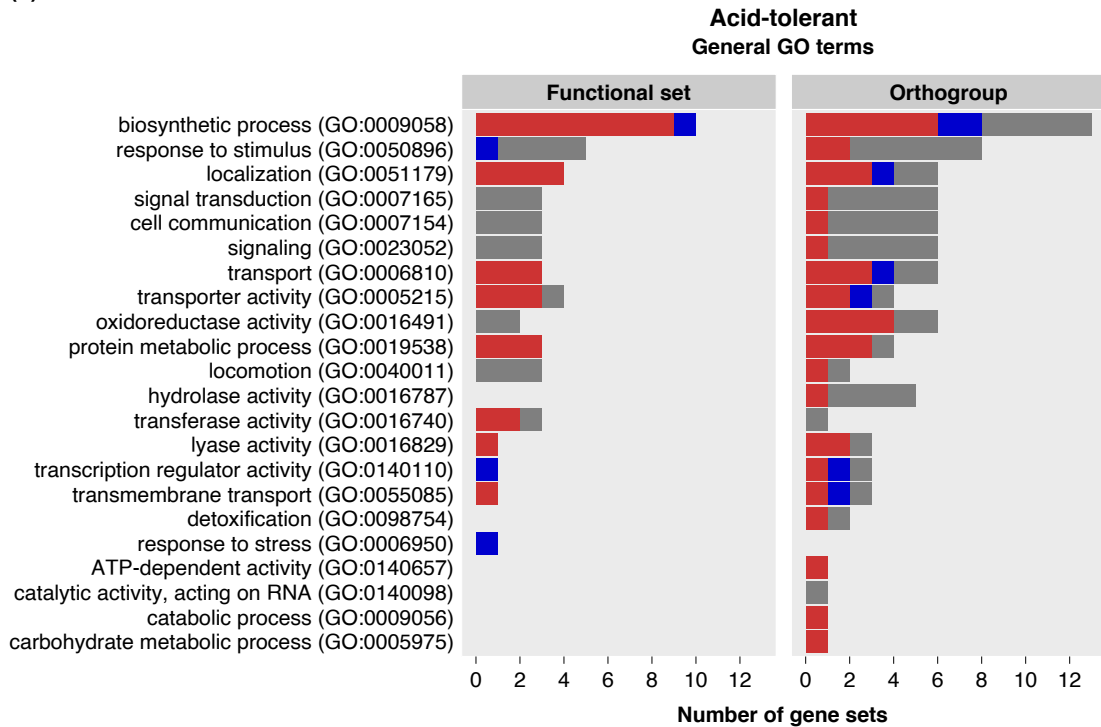
to pH 8.2 (DE FDR = 0.017). OG0008615 was detected in the four acid-tolerant and the acid-intolerant *Achnantheidium* ADMI-08. The single-copy gene was upregulated at pH 4.7 in generalists, and one isoform of the single-copy gene was upregulated at pH 4.7 versus 7.0 in the two acidophiles. Another isoform was almost upregulated at pH 4.7 compared to 8.2 in EUPA-20 (DE FDR = 0.031). OG0008627 was found in the four acid-tolerant and in the acid-intolerant *Achnantheidium* ADMI-08. The single-copy gene from this orthogroup was upregulated at pH 4.7 in generalists, and one isoform of the single-copy gene was upregulated at pH 4.7 versus 7.0 in *Eunotia* EUNS-26. Finally, OG0018365 was found in the four acid-tolerant and the acid-intolerant the *Navicula* NRAD-29. The single-copy gene was upregulated at pH 4.7 in the two generalists and at pH 4.7 versus 7.0 in EUNS-26. However, in the generalist GGD1-23, different orthogroups were identified in different transcripts from the same gene, and the OG0018365 isoforms were not differentially expressed. The gene from NRAD-29 was downregulated at pH 7.0 compared to 8.2.

#### **6.1.4 Exclusive gene sets constitutively expressed or strain-specifically enriched or depleted at pH 4.7**

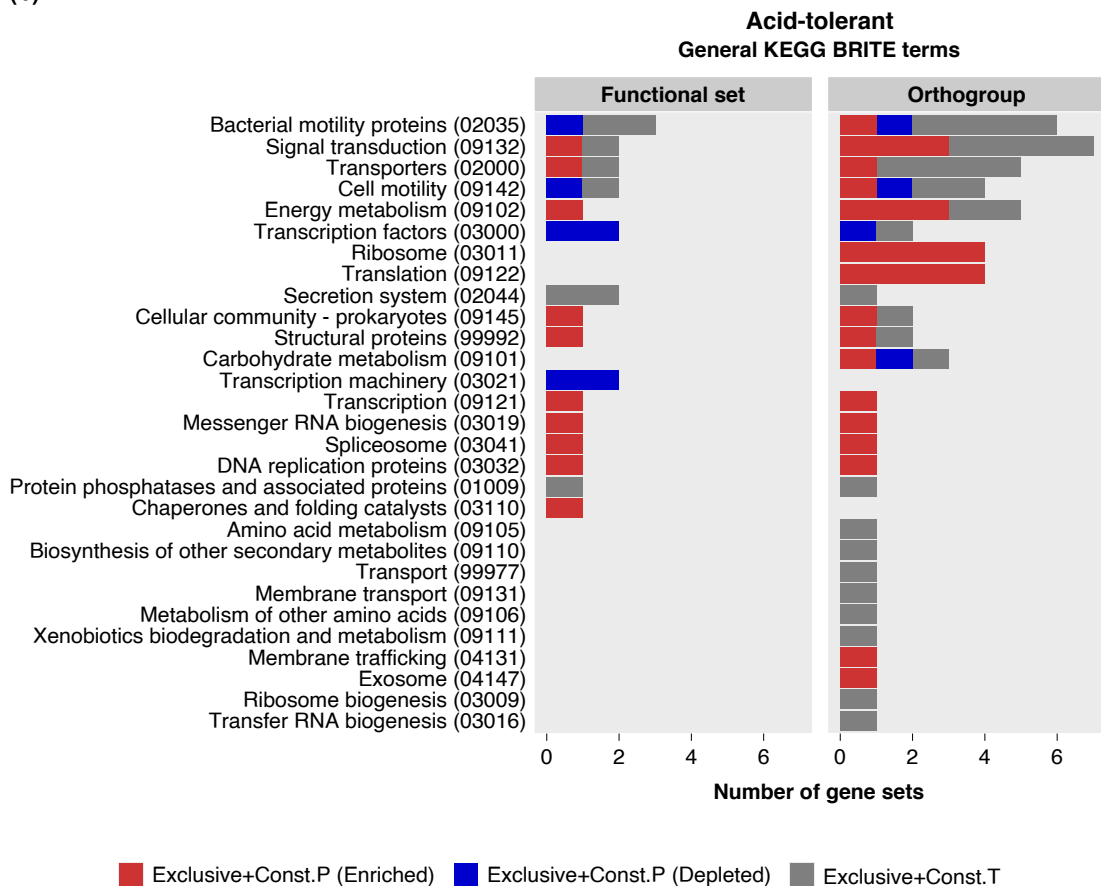
This section focuses on gene sets classified as “Exclusive+Const” for acid-tolerant strains. Among these gene sets, 60 orthogroups and 24 functional sets were not enriched nor depleted in any contrast, including those for the acid-intolerant strains, and were thus considered as the subtype “Total”. On the other hand, 38 orthogroups and 30 functional sets from the “Exclusive+Const” collection were enriched at pH 4.7 in some acid-tolerant strains and not enriched nor depleted in the remaining contrasts, including those for the acid-intolerant strains. The analogous pattern for depletions at pH 4.7 entailed 24 orthogroups and 13 functional sets. These strain-specifically enriched or strain-specifically depleted gene sets constitute the subtype “Partial”. Approximately half of all the “Exclusive+Const” orthogroups (62) were annotated in at least one functional database, either InterPro, Gene Ontology, or KEGG. 34 orthogroups had at least one GO annotation, and 31 orthogroups had at least one KEGG BRITE annotation. The number of gene sets assigned to “Exclusive+Const” in acid-tolerant strains for each selected functional category is displayed in Figure 6.5.

**Biosynthesis.** The biosynthetic process was one of the functional categories with a greater number of “Exclusive+Const” gene sets for acid-tolerant strains. Many of these sets were involved in gene expression, some participating in transcription and some in translation. Many of the gene sets involved in transcription were described as transcriptional regulators, and there were constitutively expressed, strain-specifically enriched, and strain-specifically depleted gene sets. Among them, one non-enriched orthogroup was annotated as a MCP (KEGG KO K03406). Gene sets involved in translation were bacterial-type ribosomal proteins, and all were strain-specifically enriched at pH 4.7. Also within the biosynthetic process, there were three orthogroups functionally associated with carboxylic acid biosynthesis. Two of them were related to

(a)



(b)



**Figure 6.5. Number of gene sets classified as “Exclusive+Cost” subtypes per functional category.** General GO and KEGG BRITE terms were selected to represent functional categories. The same gene set can be assigned to multiple functional categories. For strain EUNS-26, uniquely the pH 4.7 versus 7.0 comparison was considered.

amino acid biosynthesis, one being totally constitutive and the other strain-specifically enriched at pH 4.7. The third was annotated as the coenzyme PQQ synthesis protein D (InterPro IPR022479), and it was strain-specifically enriched at pH 4.7. This protein participates in the coenzyme pyrroloquinoline quinone biosynthesis (GO:0018189), a tricarboxylic acid biosynthetic process (GO:0072351). These two processes were also classified as “Exclusive+Const” enriched in some strains. Polyphosphate kinase (InterPro IPR003414) and polyphosphate biosynthetic process (GO:0006799) were also assigned to the “Exclusive+Const” collection and they were strain-specifically enriched.

**Response to stimulus.** Several gene sets involved in responding to stimuli were identified as “Exclusive+Const” for acid-tolerant strains. Most of these gene sets were implicated in signal transduction, with eleven gene sets constitutively expressed and four strain-specifically enriched. Some gene sets were related to the response to chemical stimuli, with two orthogroups involved in the response to toxic substances. The exclusive sets related to chemotaxis generally overlap between signal transduction and response to chemical stimuli, entailing four functional sets and three orthogroups. Within all the categories mentioned in this paragraph, some gene sets were constitutively expressed, and some strain-specifically enriched. Chemotaxis-related sets were also associated with locomotion and cell motility. The only gene set strain-specifically depleted at pH 4.7 was related to DNA repair and response to stress.

**Transport and localization.** “Exclusive+Const” sets for acid-tolerant strains included several transporters mediating the movement of distinct types of molecules. Three orthogroups were involved in the export or efflux from the cell, one constitutively expressed, one strain-specifically enriched, and one strain-specifically depleted. In addition, three functional sets and one orthogroup were related to polysaccharide export proteins, all strain-specifically enriched. The transport-associated OB type 2 domain (InterPro IPR013611) of the plasma membrane ABC transporter complex was also strain-specifically enriched at pH 4.7. Two orthogroups were associated with protein transport, one constitutively expressed and one strain-specifically enriched. One functional set and one orthogroup involved in glutamate/aspartate transport were constitutively expressed.

**Energy metabolism and respiration.** Five orthogroups and two functional sets in “Exclusive+Const” were involved in energy metabolism and cellular respiration. Two of these orthogroups participated in the glyoxylate metabolism, one being constitutively expressed and the other strain-specifically depleted. One strain-specifically enriched orthogroup was involved in glycolysis/gluconeogenesis or the Calvin-Benson cycle. One orthogroup and one functional set were related to S-(hydroxymethyl)glutathione synthase (KEGG KO K03396) from methane metabolism, and they were strain-specifically enriched. There was one functional set and one orthogroup

related to oxidative phosphorylation, and one exclusive orthogroup related to nitrogen metabolism. No strain-specifically depleted gene set was identified within the categories of energy metabolism and cellular respiration.

### **6.1.5 Gene sets primarily enriched or depleted at pH 4.7 in generalists or acidophiles**

Thirty-seven orthogroups and 31 functional sets were primarily enriched at pH 4.7 in generalists. Twelve and six of these orthogroups and functional sets, respectively, were exclusively detected in the two generalists, and two orthogroups and four functional sets in all four acid-tolerant strains. Six orthogroups and two functional sets were primarily depleted at pH 4.7 in generalists, all non-exclusive to generalists or to acid-tolerant strains. Twenty-four of the 43 enriched or depleted orthogroups were annotated. The most affected categories in generalists were hydrolase and oxidoreductase activities, protein metabolic process, transporter activity, peptidase activity, cell motility, biosynthetic process, and response to stimulus (Figure 6.6). Except for hydrolase and transporter activity, these categories exclusively contained gene sets primarily enriched at pH 4.7 in generalists.

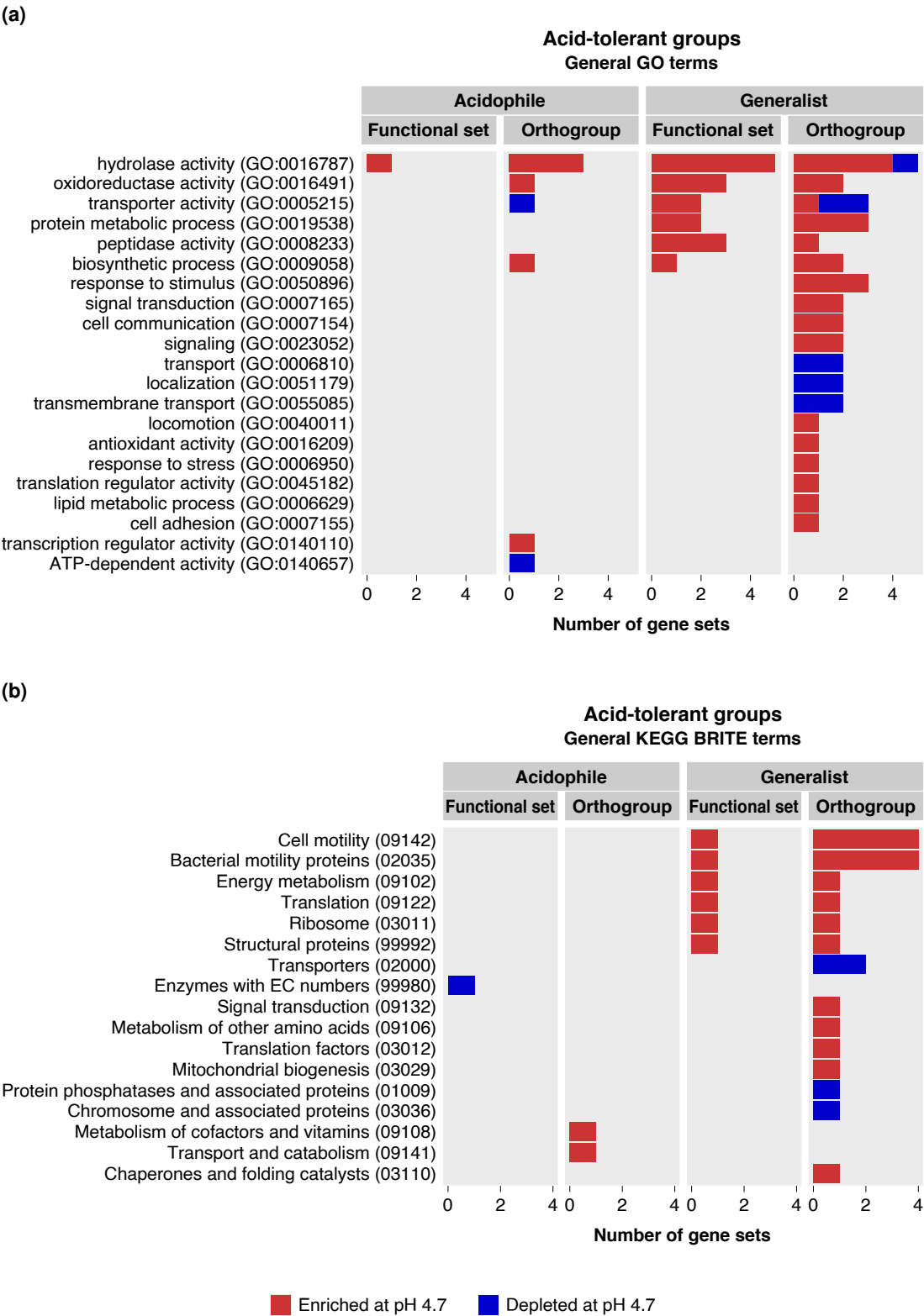
Forty orthogroups and two functional sets were primarily enriched at pH 4.7 in acidophiles. Thirty-five and one of these orthogroups and functional sets, respectively, were exclusively detected in the two acidophiles. Eight orthogroups and two functional sets were primarily depleted at pH 4.7 in generalists, all but one exclusively detected in the two acidophiles. Only nine of the 48 enriched or depleted orthogroups were functionally annotated. Compared to the generalists, there were few identified affected functional categories (Figure 6.6). The most affected category in acidophiles was hydrolase activity (GO:0016787).

### **6.1.6 Growth rate as a confounding factor**

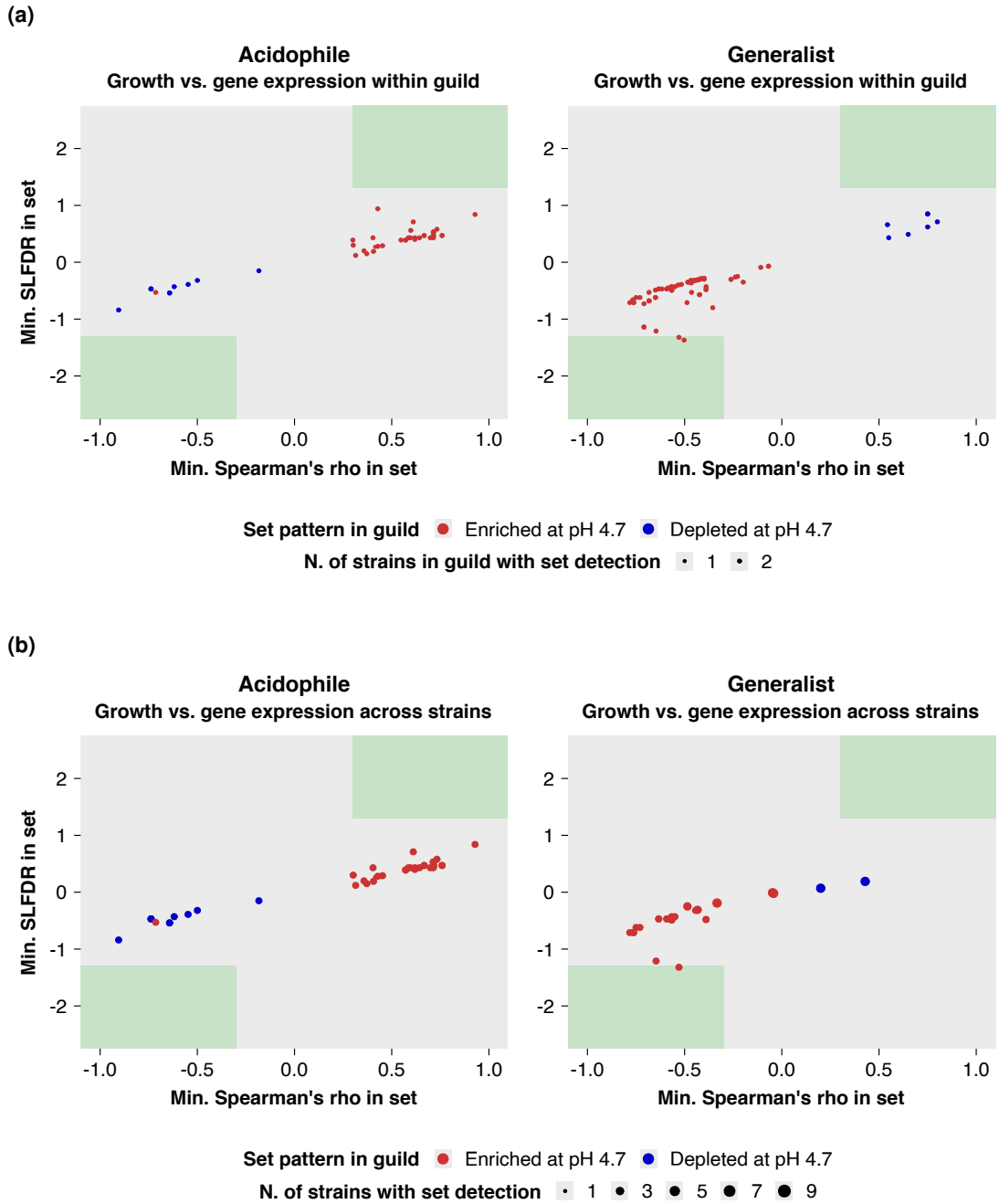
Gene sets whose expression was significantly related to growth were identified using correlation analyses. A correlation analysis was performed for each gene set and strain between the expression of its constituent genes and the strain growth rate across pH conditions. This analysis was performed both within each guild exclusively and across all strains in which the gene set was identified (including acid-intolerant strains with differential growth between pH 7.0 and 8.2). The same criteria as in subsection 5.1.4 in Chapter 5 were applied to establish significant correlations with growth across strains.

No gene set primarily enriched or depleted at pH 4.7 in acidophiles was related to growth in the two strains (Figure 6.7a) or across all strains possessing the gene set (Figure 6.7b). As for the generalists, they contained two gene sets primarily enriched at pH 4.7 in this guild that were negatively related to growth in the two strains, namely the activator of Hsp90 ATPase homolog 1-like (InterPro IPR013538) and the





**Figure 6.6. Number of gene sets significant in the IndVal analysis for generalists or acidophiles per functional category.** General GO and KEGG BRITE terms were selected to represent functional categories. The same gene set can be assigned to multiple functional categories. For strain EUNS-26, uniquely the pH 4.7 versus 7.0 comparison was considered.



**Figure 6.7. Significance of the correlations between gene sets expression and strain growth rate along the environmental pH gradient.** The  $x$  and  $y$  axes represent the minimum Spearman's  $\rho$  and the minimum SLFDR across considered strains for each gene set, respectively. The gene sets plotted were those primarily enriched or depleted at pH 4.7 exclusively in the two acidophiles (left panels) or in the two generalists (right panels) and showing the same correlation directionality between their expression and growth rate within the guild (a) or across all strains possessing the gene set (including acid-intolerant strains with differential growth between pH 7.0 and 8.2) (b). Green areas represent significant SLFDR (SLFDR  $\leq 0.05$ ) and Spearman's  $\rho$  (absolute Spearman's  $\rho \geq 0.30$ ).

secretion system C-terminal sorting domain (InterPro IPR026444). The activator of Hsp90 ATPase homolog 1-like was also identified in the acid-intolerant ACSC-11, but its correlation with growth was not significant for this strain, whereas the secretion

system C-terminal sorting domain was exclusively detected in the two generalists. No gene set primarily depleted at pH 4.7 in generalists was identified to be related to growth within generalists or across all strains possessing the gene set.

## 6.2 Discussion

### 6.2.1 Most gene sets specifically enriched at pH 4.7 are widely distributed in diatoms

This study identified potential acid-specific adaptations through four distinct collections based on the exclusivity to acid-tolerant strains and the enrichment pattern at pH 4.7 of these adaptations. Two of these collections were derived from the IndVal analysis. Functions and orthogroups enriched at pH 4.7 in most acid-tolerant strains and mostly non-enriched in the pH 7.0 versus 8.2 comparison were identified by the IndVal analysis as being primarily enriched at pH 4.7, and the same applies for depletions. Except for one gene set, all were primarily enriched, rather than depleted, at pH 4.7. Functions and orthogroups specifically enriched at pH 4.7 likely indicate a higher cellular demand for these biological functions or structures primarily under acidic pH (Bruggeman et al., 2023), suggesting that these sets are useful particularly for surviving at pH 4.7. Considering this evidence, common acid-specific responses at pH 4.7 among acid-tolerant strains might be generally associated with the induction of adaptive mechanisms rather than the repression of maladaptive proteins or compensatory downregulations.

The presence or absence of these enriched functions and orthogroups in acid-intolerant strains distinguishes the two IndVal-related collections, shedding further light on the specific mechanisms underlying acid tolerance. Enriched functions and orthogroups non-exclusive to acid-tolerant strains were typically identified in a great proportion of the acid-intolerant strains. These sets might represent widespread acid-specific stress response pathways in diatoms, activated for niche tolerance to mitigate the detrimental effects of acidic environments. Alternatively, the adaptive plastic response might have evolved uniquely in acid-tolerant strains from an ancestral phenotype constitutively expressed or less plastic (Crispo, 2007; Ghalambor et al., 2007; Yampolsky et al., 2014). These hypotheses could not be tested, since we did not collect expression data for dying diatom populations and hence, the enrichment pattern of those gene sets in acid-intolerant strains remains unknown. However, their detection in acid-intolerant strains indicates that the presence of these sets alone is insufficient for thriving at pH 4.7, pointing to a need for additional adaptive mechanisms or processes in acid-tolerant strains.

A limited number of functions were identified within the responsive non-exclusive collection. None shared the same enrichment pattern across all four acid-tolerant strains, suggesting there may not be a universal acid-specific adaptation to low pH in diatoms. However, determining the function of unannotated proteins could reveal

more similitudes in adaptive mechanisms among strains. Enriched non-exclusive functions and orthogroups entailed proteins participating in core functions, including cell division, genome stability, translation, bioenergetics, ROS production and motility. The kinetochore protein Ndc80 is localized to the kinetochore outer plate and is crucial for its maintenance and for stable kinetochore microtubule attachment (DeLuca et al., 2005). Type IA DNA topoisomerases typically control DNA topology by catalyzing the creation and re-ligation of single-stranded DNA breaks in a magnesium-dependent mode. Eukaryotic DNA topoisomerases III are universal and essential for genome stability and DNA repair (Forterre et al., 2007; McKie et al., 2021). 50S ribosomal protein L13 is essential for mitochondrial and plastidial translation and the functioning of these organelles in various organisms, and its expression changes under different stresses (Ke et al., 2018; Longworth et al., 2016; Song et al., 2013). FUN14 domain-containing proteins (FUNDGs) enhance mitochondrial autophagy (mitophagy), mitochondrial oxidative metabolism, cell proliferation, and chemotaxis, and decrease mitochondrial ROS and oxidized glutathione and mitochondrial-directed cell motility in mammals (J. Li et al., 2020; Liu et al., 2012). GDPDs exhibit a high degree of evolutionary conservation across species while showing variations in their catalytic activity and cellular roles between and within different species (Corda et al., 2014), which makes it difficult to assess their specific role in adaptation to acidic pH. The overexpression of some GDPD isoforms has been linked to different stresses, particularly phosphorous deficiency (Dyhrman et al., 2012; Helliwell et al., 2021; Mehra et al., 2018)]. Other detected orthogroups and functions putatively providing acid tolerance were not sufficiently described.

The second collection of IndVal-related gene sets is constituted of acid-specific enriched functions and orthogroups exclusively detected in acid-tolerant strains. These gene sets may encode inducible adaptations that are specifically essential for growth at low pH. The only detected function for this adaptation collection was the MCPs, and its exclusivity is questioned by its probable presence in acid-intolerant *Achnanthes* strains. MCPs are transmembrane chemoreceptors with methyl-dependent adaptation that participate in bacterial chemotaxis (Wadhams & Armitage, 2004), although they may participate in photoresponses in diatoms (Dibrova et al., 1985). In bacteria, regulation of chemotaxis-related genes has been described as a common acid stress adaptation that allows bacteria to move towards microenvironments with more suitable pH conditions (Huang et al., 2017; Schumacher et al., 2023). The precise role of MCP or MCP-like proteins in diatoms is yet to be determined.

### **6.2.2 Several acid-specific adaptations could be constitutive or strain-specific inducible**

The remaining two collections of potential acid-specific adaptations shared among acid-tolerant strains emerge from these functions and orthogroups exclusively present

in the four acid-tolerant strains and either not enriched in any contrast at all or only enriched at pH 4.7 and in some contrasts. The first collection includes exclusive non-enriched sets and may represent constitutive adaptations specific to acidic environments. This set of adaptations may act as a first line of defense, allowing the organism to acclimate more readily to environmental changes (in our case, to acidic pH), especially when the response strategy is too slow (Bruggeman et al., 2023; Geisel, 2011). Constitutive expression is also advantageous when maintaining the complex machinery for sensing and regulation requires a significant energy investment (Geisel, 2011). Alternatively, these adaptations could represent the buffer against pH changes (canalization) of fundamental physiological processes required to function correctly (Ghalambor et al., 2007). The collection of constitutive adaptations may comprise a mixture of these adaptive mechanisms. Constitutive adaptations might have arisen from modifications by directional selection in the regulatory pathway of originally plastic responses (Ghalambor et al., 2007). On the other hand, exclusive gene sets that are generally not enriched but uniquely enriched at pH 4.7 and in some strains could be adaptations specific to acidic pH exhibiting distinct regulatory strategies across strains: being constitutively expressed in some strains and inducible in others. The two collections of constitutive and strain-specific inducible exclusive gene sets individually contained notably more gene sets than IndVal-related collections. This raises the possibility that constitutive adaptations could be more prevalent across acid-tolerant strains than adaptive plasticity. However, future investigations are required to confirm the involvement in acid adaptation of gene sets from these two collections.

The functional categories of the gene sets in these two collections of potential adaptations generally partially overlap with the two previously described collections. Functional categories in this collection include biosynthesis, signal transduction (including chemotaxis), transport and localization, energy metabolism, phosphatases, and chaperones, among others. Both non-enriched and strain-specifically enriched or depleted gene sets were identified in most of these functional categories. All these are functions widely recognized as constituents of responses to environmental changes (Borowitzka, 2018; Kaur et al., 2022; López-Maury et al., 2008; Mikami et al., 2021; Sahoo et al., 2020; Storey & Storey, 2023; Xiong & Zhu, 2001). It should be noted that, based on the enrichment methods used, gene sets containing a smaller-than-average proportion of both upregulated and downregulated genes or isoforms were probably retrieved as not enriched. Further examination is required to determine whether these DEGs and DEIs could have a significant impact on the gene set activity.

### **6.2.3 Most potential acid-specific adaptations could be group- or strain-specific**

In previous sections, acid-specific adaptations potentially widely distributed among acid-tolerant diatom strains were discussed. However, more putative adaptations

specific to acidic environments was identified when considering gene sets enriched or depleted specifically within one of the two acid-tolerant guilds analyzed. Generalist- and acidophile-specific responsive adaptations specifically to acidic environments entailed functions and orthogroups participating in similar processes to those for constitutive or strain-specific inducible adaptations shared among acid-tolerant strains. Most of these gene sets were enriched at pH 4.7, especially for functional categories such as biosynthesis, oxidoreductase, and hydrolase activity in both acidophiles and generalists; protein metabolism, signaling, and antioxidant activity in generalists; and metabolism of cofactors and vitamins in acidophiles. As in the case of acid-tolerant-specific adaptations, the predominance of enrichments indicates that in our study, guild-specific responses at pH 4.7 were generally associated with the activation of adaptive mechanisms rather than the repression of maladaptive proteins. However, the function of many of the identified potential group-specific adaptive gene sets is still unknown. No significant correlation with growth was detected for any guild-specifically enriched or depleted gene sets. Two functional sets, the activator of Hsp90 ATPase homolog 1-like (InterPro IPR013538) and the secretion system C-terminal sorting domain (InterPro IPR026444), were negatively related with growth uniquely within generalists. The overall absence of these functional sets from other non-generalist strains hints at a narrowly distributed gene set specifically related to stress in some strains.

All gene sets retrieved as significant based on the IndVal analysis were enriched or depleted consistently at pH 4.7 in the two generalists or the two acidophiles. However, this collection of genes represented a small portion of all guild-specific enriched or depleted gene sets at pH 4.7, particularly in the generalists. In other words, most gene sets enriched or depleted at pH 4.7 in the two generalists or the two acidophiles were also enriched at a certain pH when comparing pH 7.0 to 8.2 in some strains (including the acid-intolerant strains). This finding is consistent with many guild-specific responsive adaptations to acidic pH also showing some advantage at neutral or alkaline pH. This point will be further discussed in Chapter 7. It would be valuable to determine whether this pattern extends beyond guild-specific adaptations to acidic pH. The “Enrichment” and the IndVal collections are not directly comparable for acid-tolerant-specific adaptations because none of the significant gene sets based on the IndVal analysis was enriched or depleted consistently at pH 4.7 in the four acid-tolerant strains.

Most gene sets uniquely detected in acid-tolerant strains are species-specific. A portion of these gene sets could be strain-specific adaptations specific to acidic environments. Also, this chapter focused the analysis on global and guild-based strain groups, but there were other strain combinations that provided many exclusive gene sets that could potentially be adaptations to acidic environments. As in the case of the other potential acid-specific adaptations, further research on these group- and strain-specific adaptations is necessary to unveil their putative role in acid adaptation.



## Chapter 7

# General discussion and conclusions

### 7.1 Molecular mechanisms for adaptation to varying pH

#### 7.1.1 The complexity in plastic responses across strains

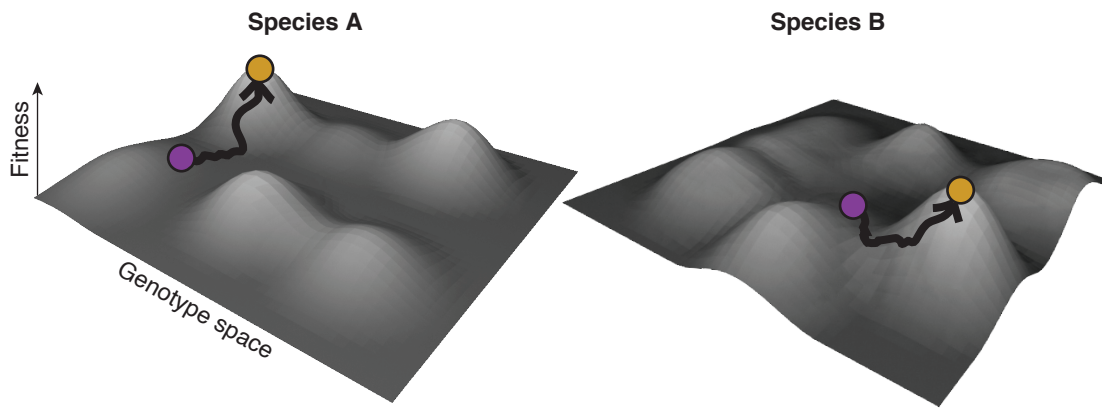
This study investigated the growth and molecular responses comparing acidic (4.7), neutral (7.0), and alkaline (8.2) environmental pH conditions of twelve freshwater diatom strains. These strains represented distinct species isolated from four Pyrenean lakes with pH conditions from acidic to alkaline. The twelve strains encompassed a broad phylogenetic range within the raphid pennate clade of diatoms (according to Nakov et al. (2018)), including species from the genera *Nitzschia*, *Tryblionella*, *Eunotia*, *Navicula*, *Achnantheidium*, *Gomphonema*, and *Encyonopsis*. Interestingly, *Navicula* NVEN-18 could be the sequenced diatom with the largest protein-coding transcriptome to date, with potentially more than 43,000 identified protein-coding genes. The twelve strains showed contrasting growth patterns along the pH gradient, with an apparent phylogenetic pH niche conservatism restricted to lower taxonomic ranks and with variable strength among clades. Strains were grouped into three distinct growth pattern types. The two *Eunotia* strains isolated from acidic Lake Aixeus, EUNS-26 and EUPA-20, were classified as acidophiles because their growth was faster at low pH. The two strains isolated from pH-neutral Lake Redon (Conangles), the *Achnantheidium* ACAF-21 and *Gomphonema* GGDI-23, were classified as generalists due to their capacity for considerable growth across all examined pH conditions. However, contrary to acidophiles, they showed their slowest growth rates at low pH. The remaining eight strains were isolated from either Lake Redó (Aigüestortes) or Lake Estanya and were classified as acid-intolerant since their populations collapsed at pH 4.7.

Acidic, neutral, and alkaline pH conditions caused the regulation of myriad molecular functions and biological processes among the strains studied. Despite sharing most of them, different species employed distinct molecular mechanisms to optimize



fitness within the same environmental pH condition. This specificity could have derived from the interaction of the niche characteristics in which the species evolved and the presence of numerous recently emerged, strain-specific genes in diatom protein-coding transcriptomes. Young genes tend to perform non-essential functions and hence are less constrained by selection than old genes, making them more likely to evolve to become specialized, niche-adaptive genes (Doughty et al., 2020; Osuna-Cruz et al., 2020). It seems probable that differences in past evolutionary niches and genetic backgrounds among lineages and species translated into differences in their adaptive landscape topologies, including the location and accessibility to the fitness peaks, even under the same environmental conditions. This resulted in strains following unique evolutionary trajectories and relying on strain-specific adaptive mechanisms for niche tolerance (Figure 7.1; Ogbunugafor and Epstein (2019); Anderson et al. (2021)).

Many affected proteins across strains were involved in the activation and deactivation of signaling pathways in response to stimuli and the resulting modifications in gene and protein expression, presumably to meet physiological requirements dictated by environmental factors. These pathways are key mechanisms for general algal responses to abiotic conditions (Kaur et al., 2022; López-Maury et al., 2008). As for pH homeostasis, proton transmembrane transport was regulated in most strains and pH changes, with contrasting enrichment patterns except for proton-transporting V-type ATPases, which seemed less relevant at alkaline pH conditions. This type of proton-transporting ATPases may participate in proton extrusion from the cytosol into the vacuole as a response to cytosolic acidification (Bethmann & Schönknecht, 2009) which is expected to be stronger at lower pH conditions. On the other hand, bicarbonate transport was identified as potentially relevant for adaptation to alkaline pH conditions but only in some Naviculales and Bacillariales strains. Increasing cytosolic concentrations of bicarbonate alkalinizes the cytoplasmic pH, which seems counterintuitive to be favored under environmental alkaline conditions. However, the enrichment of bicarbonate transporters under alkaline environments in some diatom strains may respond to a higher relevance of bicarbonate transport within CCMs to cope with the lower (almost absent) concentration of carbon dioxide in the external alkaline media. The induction of some members of the SLC4 family of bicarbonate transporters at low environmental carbon dioxide concentrations has been previously shown for marine diatoms (Nakajima et al., 2013). The diatom CCMs also include CAs, which catalyze the interconversion of carbon dioxide and water to bicarbonate and protons. CAs were affected in many strains but with contrasting expression patterns in most cases and different predicted distributions in subcellular compartments. This finding agrees with the previously described large diversity in CAs types, subcellular locations, and presumably functions across diatoms (Shen et al., 2017).



**Figure 7.1. Differences in adaptive landscapes can lead to distinct evolutionary trajectories.** This figure depicts the hypothetical adaptive landscape of two species, A and B, which have evolved in different environmental and/or genetic backgrounds. The  $x$ - $y$  plane represents the genotype space and the  $z$ -axis represents the fitness, so each location in the plane corresponds to a genotype and the height at that location corresponds to the genotype fitness. Black arrows represent the mutational path followed from the initial point (purple) towards one of the adaptive peaks (orange), favored by natural selection. Under the same environmental conditions (e.g., low pH), each species can follow a unique evolutionary trajectory towards an adaptive peak based on its genetic background. Evolving populations can be fixed at local, suboptimal peaks, hindering evolvability. Based on Fig. 3 in Payne and Wagner (2018).

### 7.1.2 Diatoms and the challenge of acidity

According to both growth and molecular responses, acidic pH represented a more ecologically distinct environment from neutral and alkaline pH conditions. Considering slower (including negative) growth rates to be related to higher cellular stress (Bruggeman et al., 2023; Sokolova, 2013), low pH represented a more stressing environment for all analyzed strains except the two *Eunotia* retrieved from the acidic Lake Aixens. The populations of strains isolated from pH-neutral or alkaline lakes collapsed in acidic environments except for the two strains from the pH-neutral Lake Redon (Conangles), *Achnanthisdium* ACAF-21 and *Gomphonema* GGD1-23, which grew at a slower pace. The two strains isolated from the acidic Lake Aixens could maintain a slightly positive growth at neutral and even alkaline conditions. The distinctiveness of acidic pH as an eco-evolutionary challenge is probably related to the marine ancestry of diatoms and the widespread regional distribution of pH-circumneutral fresh waters (described in Nakov et al. (2019) and Catalan, Curtis, and Kernan (2009)). The differential tolerance to low pH and the divergence in acid adaptations may have derived from the independent freshwater colonization events of diatom lineages, with *Eunotia*, *Achnanthisdium*, and *Gomphonema* representing early freshwater colonizers from independent colonization events (Nakov et al., 2019). Marine diatom colonizing fresh waters may require first invading fresh waters with a pH close to the sea and after allowing sufficient time for new, costly adaptations that facilitate survival at low pH to randomly emerge in some lineages within the clade (Gostinčar et al., 2022). The strain specificity of these acid adaptations could be derived from the presence of numerous recently emerged, strain-specific genes

detected in the studied diatom protein-coding transcriptomes. Young genes tend to perform non-essential functions and hence are less constrained by selection than old genes, making them more likely to evolve to become specialized, niche-adaptive genes (Doughty et al., 2020; Osuna-Cruz et al., 2020). Overall, our results are consistent with mutational randomness playing a relevant role in adaptive evolution (Lenski & Travisano, 1994).

The limited enrichments at low pH shared among acid-tolerant strains included the 5'-nucleotidases, the GDPD domain-containing proteins, the CHMP5/Vps60, and an FCP. These proteins were detected in acid-intolerant strains, indicating that they could be general diatom niche-response proteins. Previous studies identified them as responsive proteins to certain environmental gradients, such as light, drought or temperature (Alipanah et al., 2018; Alqurashi et al., 2018; Büchel, 2014; Chen et al., 2018; Dittami et al., 2009; Dyhrman et al., 2012; Helliwell et al., 2021; Hwang et al., 2008; X. Ma et al., 2020; Mehra et al., 2018; Park et al., 2009; Truong et al., 2022; Zhao et al., 2020). The CHMP5/Vps60 could be related to membrane remodeling, which is involved in numerous cellular functions such as vesicle biogenesis or membrane wound repair (McCullough et al., 2018; Pfitzner et al., 2023). FCPs participate in light harvesting and energy dissipation (Büchel et al., 2022; Grossman et al., 1990). Lastly, the role of 5'-nucleotidases and GDPD domain-containing proteins is not clear, but it may be related to phosphate metabolism and associated pH homeostasis, biosynthesis, or signal transduction (Eskes et al., 2017; Paytan & McLaughlin, 2007; Wagner, 2023). Although these common plastic responses were most distinctively expressed at low pH, shifting from neutral to alkaline pH also affected these functions in some strains (including acid-intolerants). This suggests that these functions could also be beneficial at neutral or alkaline pH in the case of adaptive plasticity. This observation contrasts with the specificity to acid pH observed for the entire set of acid responses in strains discussed in the previous paragraph. The enrichment patterns of these functions for the neutral versus alkaline pH comparison indicate that they are likely not general diatom responses to increasing acidity or diverging pH from neutrality. The only exception was the CHMP5, which showed an increasing enrichment as the pH diverged from neutrality in some strains.

As for widely shared acid-specific responses, they involved the kinetochore protein Ndc80, type IA DNA topoisomerases (probably type III), mitochondrial or plastidial 50S ribosomal protein L13, a FUNDCs protein, a GDPDs, and some poorly annotated orthogroups. These proteins were functionally non-overlapping with those involved in shared responses unspecific to low pH, except perhaps for the orthogroup encoding a GDPDs. Ndc80 is essential for chromosome segregation stability (DeLuca et al., 2005), and eukaryotic DNA topoisomerases III are crucial for genomic maintenance and stability (Forterre et al., 2007; McKie et al., 2021). 50S ribosomal protein L13 is required for mitochondrial and plastidial translation and the functioning of these organelles in various organisms (Ke et al., 2018; Longworth et al., 2016; Song et al.,

2013). FUNDCs could be related to mitochondrial metabolism and control, including ROS and motility (J. Li et al., 2020; Liu et al., 2012). The role of GDPDs has been discussed in the previous paragraph. Surprisingly, shared responses (either specific or unspecific) to acidic pH might not entail functions traditionally associated with pH homeostasis. Proton export across the plasma membrane was enriched at low pH in the two generalists and potentially in the acidophile EUNS-26, but not in EUPA-20. Also, as mentioned above, some enzymes potentially related to phosphate buffer in pH homeostasis were enriched at low pH, but more studies are required to test its involvement. The absence of more core plastic responses regulating pH homeostasis does not imply that this regulation is not relevant for acid adaptation in diatoms. Instead, based on the remarkable uniqueness of plastic responses to pH, different mechanisms for pH homeostasis might be narrowly distributed across strains, as discussed in subsection 7.1.1.

### **7.1.3 Acidophiles and generalists: distinct strategies for thriving at low pH**

The two acidophilic *Eunotia* EUNS-26 and EUPA-20 exhibited the most distinct set of plastic responses. These two strains showed a much larger number of orthogroups exclusively shared between them than any other pair of analyzed strains, even higher than their strain-exclusive orthogroups. In the CGE, the two *Eunotia* strains were identified as specialists on acidic pH, and *Achnanthes* ACAF-21 and *Gomphonema* GGDI-23 strains as generalists, which maintained considerable but slower growth at low pH. The pH niche position and breadth of strains probably reflect the average and the variation in environmental pH in which they evolved. Acidophiles probably evolved in constant acidic pH environments (Kassen, 2002; Woodcock et al., 2017). The loss of the capacity to maintain similar growth at neutral and alkaline pH conditions in acidophiles may have derived from the costs of adaptation, represented by two factors acting at different evolutionary scales (Kassen, 2002). In the short term, strongly selected adaptive genes at low pH may be deleterious at neutral and alkaline pH (antagonistic pleiotropy). The pH conditions of acidic environments greatly differ from the ancestral mildly alkaline pH where diatoms originally evolved, and thus, higher antagonistic pleiotropy would be expected for adaptations to acidic pH in acidophiles. Moreover, the two examined acidophiles could be best adapted to even lower pH conditions than pH 4.7, further increasing the expectation for a stronger antagonistic pleiotropy. In the long term, mutations showing neutral effects in acidic pH but deleterious at higher pH conditions may accumulate in the genome of strains evolving in constantly low pH conditions. When the cost of switching becomes too large, then the acidophile strategy should be selected (J. Ma & Levin, 2006). In contrast, the two generalists *Achnanthes* ACAF-21 and *Gomphonema* GGDI-23 contained more strain-exclusive functions and orthogroups than shared among them. Despite this high degree of individual uniqueness, the two generalists shared more acid responses among them than with *Eunotia* strains. Generalists ACAF-21 and GGDI-23 may

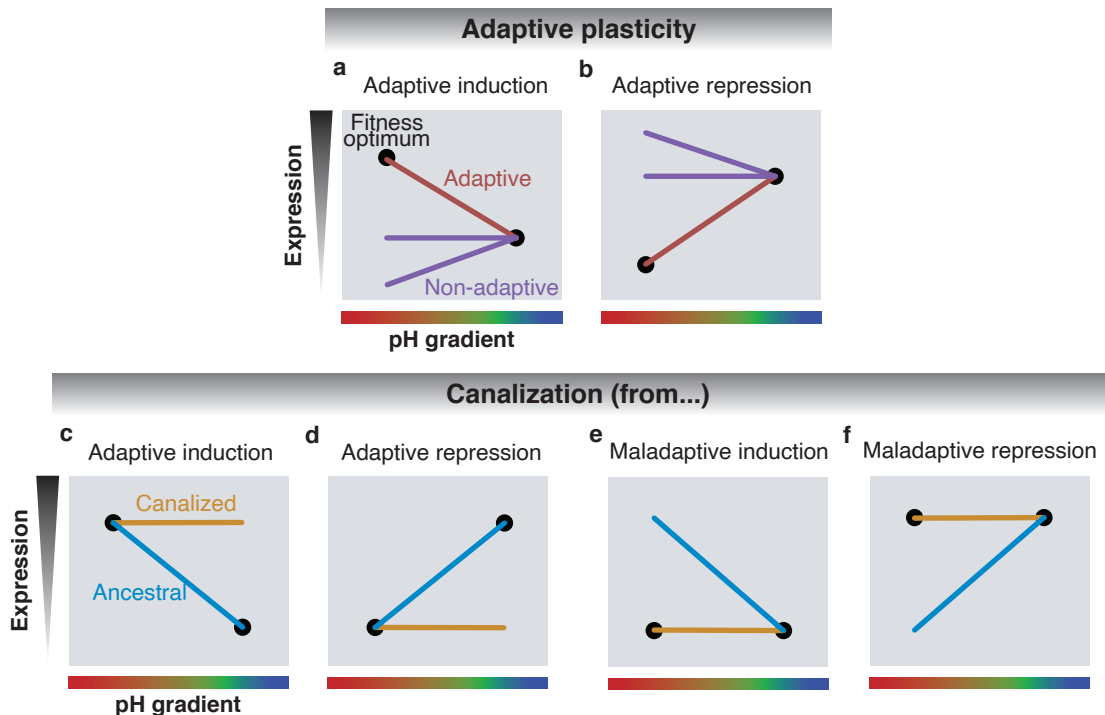
have experienced more similar selecting pressures than those experienced by the two acidophilic *Eunotia*, causing them to share more plastic acid responses. For example, the capacity of generalists to maintain homeostasis across a wide range of pH conditions may be associated with high resource costs in mechanisms for reducing environmental fitness variance, which may exclusively be beneficial in heterogeneous pH environments (Kassen, 2002; Van Tienderen, 1991; Woodcock et al., 2017).

Most functions responding specifically to low pH in the two acidophiles probably have evolved within this lineage, in contrast to the wider distribution across diatoms observed for acid-unspecific responses of acidophiles and acid (specific and unspecific) responses of generalists. In addition, many of these acid-unspecific responses of acidophiles entailed proteins encoded by an orthogroup exclusive to acidophiles, so these proteins may possess some unique characteristics that could be advantageous for developing their function at low pH. These results are in line with the relevant role of random mutation in generating adaptive innovations facilitating survival and growth at low pH suggested in subsection 7.1.2.

#### **7.1.4 Adaptive plasticity and canalization in acidophiles**

This study investigated molecular plasticity in response to varying pH conditions along the pH gradients. However, it can be challenging to determine whether this plasticity is adaptive, meaning it confers a selective advantage and increases fitness. The same observed plastic response along the pH gradient could be adaptive or maladaptive (see Figure 7.2). For observed plastic responses to low pH, one approach to determining their adaptive potential is to compare them to the direction of expression correlation with maximum acid tolerance among organisms (Campbell-Staton et al., 2021). When the expression of a particular trait is higher in organisms tolerating lower pH conditions (positive regulator), then higher expressions at low pH can be categorized as adaptive, and lower expressions at low pH as maladaptive plasticity (Figure 7.2a). Conversely, when the expression of a particular trait is lower in organisms tolerating lower pH conditions (negative regulator), then lower expressions at low pH can be categorized as adaptive, and higher expressions at low pH as maladaptive plasticity (Figure 7.2b).

Based on observed growth responses and native pH conditions, the two acidophiles might show greater tolerance to lower pH conditions than the two generalists. Under this assumption, traits with a higher expression in acidophiles than in generalists are probably positive regulators of acid tolerance, and hence, their induction at low pH would be categorized as adaptive plasticity. A set of potential positive regulators of acid tolerance identified in this study are those functions and orthogroups exclusively detected in acidophiles and enriched at low pH, including a family of eukaryotic acid phosphatases and a heat shock factor. The regulation of these types of proteins has been associated with acid tolerance in other organisms (Hirooka et al., 2017). Most orthogroups potentially acting as positive regulators were not functionally annotated,



**Figure 7.2. Hypothetical reaction norms in adaptive plasticity and canalization to low pH environments.** The same observed changes in the expression of a trait (e.g., a protein expression) along the environmental pH gradient can represent two distinct adaptive mechanisms depending on the position of fitness optima of that trait in each environment. Based on Fig. 1 in Ghalambor et al. (2007).

so further studies are required to characterize those genes and create a more comprehensive functional profile of identified potentially positive regulators.

Constant selective pressures for a particular adaptive plastic phenotype (e.g., in a constant environment) may result in its canalization, with that advantageous phenotype becoming constitutively expressed (Ghalambor et al. (2007); Figure 7.2c–d). Canalization of plastic phenotypes can also be selected for maladaptive plasticity (Figure 7.2e–f); e.g., to maintain homeostasis of fundamental processes across pH changes. In this study, many functions and, especially, orthogroups were exclusively identified in the two acidophiles and showed no response to varying pH conditions. Some of these constitutive functions might have arisen from the canalization of ancestrally plastic traits. However, this evolutionary mechanism is more likely to have emerged in those functions showing plastic responses to low pH in the two generalists while being constitutively expressed in acidophiles. Other sets identified in this study as potentially canalized traits as an adaptation to low pH are constitutive functions of the four acid-tolerant strains or only with generalists. Further studies are required to assess whether the observed expression patterns arose from canalization or increased plasticity in the response based on the ancestral state and discern between adaptive and maladaptive plasticity.

## 7.2 Research limitations and future research

While the research presented here contributes to understanding the molecular complexity of adaptation to varying pH conditions in diatoms, it is important to acknowledge some limitations. This study investigated the growth and molecular responses of twelve distinct diatom strains at three environmental pH conditions ranging from acidic to alkaline. Although the explored diatom strains were phylogenetically different and the pH conditions ranged from pH 4.7 to 8.2, the conclusions achieved by this study could not be general to diatoms and the pH gradient. For instance, the two acidophilic species were more closely related phylogenetically than the two generalist species. This raises the possibility that the observed pattern differences among acidophiles and generalists in their plastic responses to low pH might be influenced by phylogenetic distance. Future investigations could benefit from including a more diverse set of species and additional pH conditions (if possible, some representing more extreme environments), providing a more comprehensive picture of growth and molecular patterns along the pH gradient. In addition, the CGE performed was not aimed at obtaining transcriptomic data for dying populations. Comparing molecular data from dying to surviving populations will provide valuable insights into the adaptive component of observed growth and molecular responses (e.g., Yampolsky et al. (2014)). Another factor to consider is that the response to an environmental change not only varies among species but also could vary markedly among strains from the same diatom species (e.g., Pinseel et al. (2022)).

Transcriptomes for the twelve diatom strains were assembled *de novo*. This method allows for evolutionary analysis with no reference genome, which is helpful for non-model species like those included in this investigation. However, this approach comes at the cost of potential underestimation of diversity and heterozygosity, and biases in expression estimates (Freedman et al., 2020). The filtering steps of both lowly expressed and short transcripts and genes performed in the present study likely alleviated the biases in gene expression estimates but at the cost of excluding a potentially significant number of protein-coding genes (Freedman et al., 2020). In addition, it should be noted that associating mRNA responses to specific phenotypic responses may not be straightforward in many cases. The correlation between mRNA and protein abundance could be notably weak due to post-transcription, translation, and post-translation modifications, and protein interaction with other biomolecules and their catalytic activity also determine the resulting phenotype (Buccitelli & Selbach, 2020).

An analytical limitation encountered during this investigation was the relatively high percentage of unannotated proteins, even using multiple well-established and widely used functional databases. For example, many orthogroups identified as potential regulators of acid adaptation were unannotated, limiting understanding of their specific

contributions to acid tolerance in diatoms. From another perspective, knowing the regulation across pH conditions and the predicted location of proteins from unannotated orthogroups is a first step toward their functional annotation. Lastly, traditional gene set enrichment methods FCS and over-representation analysis using chi-square (cORA) and the additional UGS used in this investigation are based on some assumptions that are likely not realistic for biological systems, such as the independence among gene sets and assigning a similar weight to all genes within the same gene set (for a review, see Khatri et al. (2012), García-Campos et al. (2015), and Table 2.5).

In a world where ongoing climate change and anthropogenic activities have a substantial impact on aquatic ecosystems, the distribution of diatom species across pH conditions has been extensively studied for their use as bioindicators to infer the ecological status of freshwaters. However, the molecular mechanisms underlying the diatom species segregation along the pH gradients are unknown. This thesis contributes to the first comprehensive investigation of the molecular responses of freshwater diatoms along the pH gradient, with a special focus on acidic environments, representing a first step toward understanding the role of these genes in the acid and alkaline tolerance mechanisms of diatoms. This question is not only of fundamental biological interest but also holds the potential to improve the use of diatoms as bioindicators of pH conditions. Further studies should explore the functional role and the adaptive potential of identified plastic responses and exclusive molecular mechanisms; for example, by assessing species acid tolerance and establishing the precise correlation between gene expression and maximum acid tolerance (e.g., Campbell-Staton et al. (2021)). Although this investigation provided an overview of all plastic responses in each strain, species-specific plastic responses to pH likely played a relevant role in adaptation due to the high uniqueness of diatom transcriptomes, thus requiring a more in-depth examination.

## 7.3 Conclusions

The principal conclusions drawn from this thesis research are:

1. Environmental pH changes among acidic, neutral, and alkaline pH conditions determine the regulation of a myriad of molecular functions and biological processes across diatoms.
2. Many affected functions are involved in activating and deactivating signaling pathways in response to stimuli and the resulting transcriptional and translational tuning, presumably to meet physiological requirements dictated by environmental factors.
3. The responses of known functions to pH changes are markedly strain-specific despite strains sharing a substantial proportion of these functions, meaning that



diatom species employ distinct molecular mechanisms to improve fitness within the same environmental pH condition.

4. Survival at acidic pH likely requires the emergence of new adaptations targeted explicitly to low pH. The distinctiveness of acidic pH as an eco-evolutionary challenge is probably linked to the marine ancestry of diatoms and the widespread regional distribution of pH-circumneutral fresh waters. Colonizing acidic freshwaters may require sufficient time for new, costly adaptations that facilitate survival at low pH to emerge randomly in some lineages within the clade.
5. The high inter-specific variation in plastic responses to pH may result from the large genetic divergence among strains, which could favor the accumulation of historical contingencies in the evolutionary radiation across pH continental gradients. The specificity of the acid adaptations could be derived from numerous recently emerged, strain-specific genes, which are less constrained by selection and thus more likely to evolve to become specialized, niche-adaptive genes.
6. The limited shared enrichments at low pH in acid-tolerant strains included the 5'-nucleotidases, the GDPD domain-containing proteins, the CHMP5/Vps60, and an FCP, which could be general responsive proteins, also outside diatoms.
7. Acid-specific responses widely shared among acid-tolerant strains involved the kinetochore protein Ndc80, type IA DNA topoisomerases (probably type III), mitochondrial or plastidial 50S ribosomal protein L13, a FUNDCs protein, a GDPDs, and some poorly annotated orthogroups.
8. Detected differences between plastic responses to acidic pH in acidophiles and generalist strains might result from the distinct evolutionary response to the selective pressure of the pH gradient, with acidophiles potentially experiencing greater antagonistic pleiotropy for their acid adaptations.
9. This thesis identified plastic responses to varying pH from acidic to alkaline conditions, and exclusive functions and genes within acid-tolerant strains. Further studies should explore the functional role and the adaptive potential of these mechanisms. Identifying these genes is a significant finding, given the considerable proportion of unannotated proteins occurring in the examined transcriptomes.
10. The number of strains considered is far from comprehensively covering the diatom evolutionary tree. However, the variety of clades considered and the complexity of the molecular responses along the pH gradient are sufficient to conclude that pH tolerance has been resolved with much historical contingency.

# Bibliography

- Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., et al. (2017). Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Scientific Data*, 4(1). <https://doi.org/10.1038/sdata.2017.93>
- Alipanah, L., Winge, P., Rohloff, J., Najafi, J., Brembu, T., & Bones, A. M. (2018). Molecular adaptations to phosphorus deprivation and comparison with nitrogen deprivation responses in the diatom *Phaeodactylum tricornutum* (S. Lin, Ed.). *PLOS ONE*, 13(2), e0193335. <https://doi.org/10.1371/journal.pone.0193335>
- Almagro Armenteros, J. J., Salvatore, M., Emanuelsson, O., Winther, O., von Heijne, G., Elofsson, A., & Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance*, 2(5), e201900429. <https://doi.org/10.26508/lsa.201900429>
- Alqurashi, M., Chiapello, M., Bianchet, C., Paolocci, F., Lilley, K., & Gehring, C. (2018). Early Responses to Severe Drought Stress in the *Arabidopsis thaliana* Cell Suspension Culture Proteome. *Proteomes*, 6(4), 38. <https://doi.org/10.3390/proteomes6040038>
- Ammendolia, D. A., Bement, W. M., & Brumell, J. H. (2021). Plasma membrane integrity: implications for health and disease. *BMC Biology*, 19(1). <https://doi.org/10.1186/s12915-021-00972-y>
- Anderson, D. W., Baier, F., Yang, G., & Tokuriki, N. (2021). The adaptive landscape of a metallo-enzyme is shaped by environment-dependent epistasis. *Nature Communications*, 12(1). <https://doi.org/10.1038/s41467-021-23943-x>
- Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2019). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold (A. Valencia, Ed.). *Bioinformatics*, 36(7), 2251–2252. <https://doi.org/10.1093/bioinformatics/btz859>
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1), 25–29. <https://doi.org/10.1038/75556>

- Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, 200(1–2), 1–19. <https://doi.org/10.1016/j.ecolmodel.2006.07.005>
- Babenko, I., Friedrich, B. M., & Kröger, N. (2022). Structure and Morphogenesis of the Frustule. In A. Falciatore & T. Mock (Eds.), *The Molecular Life of Diatoms* (pp. 287–312). Springer International Publishing. [https://doi.org/10.1007/978-3-030-92499-7\\_11](https://doi.org/10.1007/978-3-030-92499-7_11)
- Bai, Y., Cao, T., Dautermann, O., Buschbeck, P., Cantrell, M. B., Chen, Y., Lein, C. D., Shi, X., Ware, M. A., Yang, F., Zhang, H., Zhang, L., Peers, G., Li, X., & Lohr, M. (2022). Green diatom mutants reveal an intricate biosynthetic pathway of fucoxanthin. *Proceedings of the National Academy of Sciences*, 119(38). <https://doi.org/10.1073/pnas.2203708119>
- Baker-Austin, C., & Dopson, M. (2007). Life in acid: pH homeostasis in acidophiles. *Trends in Microbiology*, 15(4), 165–171. <https://doi.org/10.1016/j.tim.2007.02.005>
- Barry, W. T., Nobel, A. B., & Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9), 1943–1949. <https://doi.org/10.1093/bioinformatics/bti260>
- Basu, S., Patil, S., Mapleson, D., Russo, M. T., Vitale, L., Fevola, C., Maumus, F., Casotti, R., Mock, T., Caccamo, M., Montresor, M., Sanges, R., & Ferrante, M. I. (2017). Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. *New Phytologist*, 215(1), 140–156. <https://doi.org/10.1111/nph.14557>
- Bateman, A. (2002). The Pfam Protein Families Database. *Nucleic Acids Research*, 30(1), 276–280. <https://doi.org/10.1093/nar/30.1.276>
- Battarbee, R. W., Charles, D. F., Bigler, C., Cumming, B. F., & Renberg, I. (2010, September). Diatoms as indicators of surface-water acidity. In J. P. Smol & E. F. Stoermer (Eds.), *The Diatoms: Applications for the Environmental and Earth Sciences* (pp. 98–121). Cambridge University Press. <https://doi.org/10.1017/cbo9780511763175.007>
- Battarbee, R. W., Jones, V. J., Flower, B., Bennion, H., Carvalho, L., & Juggins, S. (2001). Diatoms. In J. P. Smol, H. J. B. Birks, W. M. Last, R. S. Bradley, & K. Alverson (Eds.), *Tracking Environmental Change Using Lake Sediments: Terrestrial, Algal, and Siliceous Indicators* (pp. 155–202). Springer Netherlands. [https://doi.org/10.1007/0-306-47668-1\\_8](https://doi.org/10.1007/0-306-47668-1_8)
- Benoiston, A.-S., Ibarbalz, F. M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S., & Bowler, C. (2017). The evolution of diatoms and their biogeochemical functions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1728), 20160397. <https://doi.org/10.1098/rstb.2016.0397>

- Bethmann, B., & Schönknecht, G. (2009). pH regulation in an acidophilic green alga – a quantitative analysis. *New Phytologist*, 183(2), 327–339. <https://doi.org/10.1111/j.1469-8137.2009.02862.x>
- Birks, H. J. B., Line, J. M., Juggins, S., Stevenson, A. C., & Ter Braak, C. J. F. (1990). Diatoms and pH reconstruction. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 327(1240), 263–278. <https://doi.org/10.1098/rstb.1990.0062>
- Blanco, S. (2020). Diatom Taxonomy and Identification Keys. In G. Cristóbal, S. Blanco, & G. Bueno (Eds.), *Modern Trends in Diatom Identification* (pp. 25–38). Springer International Publishing. [https://doi.org/10.1007/978-3-030-39212-3\\_3](https://doi.org/10.1007/978-3-030-39212-3_3)
- Bodor, A., Bounedjoum, N., Vincze, G. E., Erdeiné Kis, Á., Laczi, K., Bende, G., Szilágyi, Á., Kovács, T., Perei, K., & Rákhely, G. (2020). Challenges of unculturable bacteria: environmental perspectives. *Reviews in Environmental Science and Bio/Technology*, 19(1), 1–22. <https://doi.org/10.1007/s11157-020-09522-4>
- Boron, W. F. (2004). Regulation of intracellular pH. *Advances in Physiology Education*, 28(4), 160–179. <https://doi.org/10.1152/advan.00045.2004>
- Borowitzka, M. A. (2018). The ‘stress’ concept in microalgal biology—homeostasis, acclimation and adaptation. *Journal of Applied Phycology*, 30(5), 2815–2825. <https://doi.org/10.1007/s10811-018-1399-0>
- Borrego-Ramos, M., Rimet, F., Bécares, E., & Blanco, S. (2023). Environmental drivers of genetic variability in common diatom genera: Implications for shallow lake biomonitoring. *Ecological Indicators*, 154, 110898. <https://doi.org/10.1016/j.ecolind.2023.110898>
- Brooks, A. N., Turkarslan, S., Beer, K. D., Yin Lo, F., & Baliga, N. S. (2011). Adaptation of cells to new environments. *WIREs Systems Biology and Medicine*, 3(5), 544–561. <https://doi.org/10.1002/wsbm.136>
- Bruggeman, F. J., Teusink, B., & Steuer, R. (2023). Trade-offs between the instantaneous growth rate and long-term fitness: Consequences for microbial physiology and predictive computational models. *BioEssays*, 45(10). <https://doi.org/10.1002/bies.202300015>
- Brylka, K., Alverson, A. J., Pickering, & Conley, D. J. (2023). Uncertainties surrounding the oldest fossil record of diatoms. *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-35078-8>
- Buccitelli, C., & Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10), 630–644. <https://doi.org/10.1038/s41576-020-0258-4>
- Büchel, C. (2014). Fucoxanthin-Chlorophyll-Proteins and Non-Photochemical Fluorescence Quenching of Diatoms. In *Non-Photochemical Quenching and Energy Dissipation in Plants, Algae and Cyanobacteria* (pp. 259–275). Springer Netherlands. [https://doi.org/10.1007/978-94-017-9032-1\\_11](https://doi.org/10.1007/978-94-017-9032-1_11)

- Büchel, C., Goss, R., Bailleul, B., Campbell, D. A., Lavaud, J., & Lepetit, B. (2022). Photosynthetic Light Reactions in Diatoms. I. The Lipids and Light-Harvesting Complexes of the Thylakoid Membrane. In *The Molecular Life of Diatoms* (pp. 397–422). Springer International Publishing. [https://doi.org/10.1007/978-3-030-92499-7\\_15](https://doi.org/10.1007/978-3-030-92499-7_15)
- Burnap, R. L. (2015). Systems and Photosystems: Cellular Limits of Autotrophic Productivity in Cyanobacteria. *Frontiers in Bioengineering and Biotechnology*, 3. <https://doi.org/10.3389/fbioe.2015.00001>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1). <https://doi.org/10.1186/1471-2105-10-421>
- Campbell-Staton, S. C., Velotta, J. P., & Winchell, K. M. (2021). Selection on adaptive and maladaptive gene expression plasticity during thermal adaptation to urban heat islands. *Nature Communications*, 12(1). <https://doi.org/10.1038/s41467-021-26334-4>
- Capra, J. A., Pollard, K. S., & Singh, M. (2010). Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biology*, 11(12), R127. <https://doi.org/10.1186/gb-2010-11-12-r127>
- Carlson, M. (2023). *GO.db: A set of annotation maps describing the entire Gene Ontology* [R package version 3.17.0].
- Casey, J. R., Grinstein, S., & Orlowski, J. (2009). Sensors and regulators of intracellular pH. *Nature Reviews Molecular Cell Biology*, 11(1), 50–61. <https://doi.org/10.1038/nrm2820>
- Catalan, J., Curtis, C. J., & Kernan, M. (2009). Remote European mountain lake ecosystems: regionalisation and ecological status. *Freshwater Biology*, 54(12), 2419–2432. <https://doi.org/10.1111/j.1365-2427.2009.02326.x>
- Catalan, J., Pla, S., García, J., & Camarero, L. (2009). Climate and CO<sub>2</sub> saturation in an alpine lake throughout the Holocene. *Limnology and Oceanography*, 54(6part2), 2542–2552. [https://doi.org/10.4319/lo.2009.54.6\\_part\\_2.2542](https://doi.org/10.4319/lo.2009.54.6_part_2.2542)
- Chen, X.-H., Li, Y.-Y., Zhang, H., Liu, J.-L., Xie, Z.-X., Lin, L., & Wang, D.-Z. (2018). Quantitative Proteomics Reveals Common and Specific Responses of a Marine Diatom *Thalassiosira pseudonana* to Different Macronutrient Deficiencies. *Frontiers in Microbiology*, 9. <https://doi.org/10.3389/fmicb.2018.02761>
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Consalvey, M., Perkins, R. G., Paterson, D. M., & Underwood, G. J. C. (2005). PAM FLUORESCENCE: A BEGINNERS GUIDE FOR BENTHIC DIATOMISTS. *Diatom Research*, 20(1), 1–22. <https://doi.org/10.1080/0269249x.2005.9705619>

- Consortium, T. G. O., Aleksander, S. A., Balhoff, Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, Van Auken, K., Ramsey, J., . . . Ruzicka. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1), iyad031. <https://doi.org/10.1093/genetics/iyad031>
- Corda, D., Mosca, M. G., Ohshima, N., Grauso, L., Yanaka, N., & Mariggiò, S. (2014). The emerging physiological roles of the glycerophosphodiesterase family. *The FEBS Journal*, 281(4), 998–1016. <https://doi.org/10.1111/febs.12699>
- Crispo, E. (2007). THE BALDWIN EFFECT AND GENETIC ASSIMILATION: REVISITING TWO MECHANISMS OF EVOLUTIONARY CHANGE MEDIATED BY PHENOTYPIC PLASTICITY. *Evolution*, 61(11), 2469–2479. <https://doi.org/10.1111/j.1558-5646.2007.00203.x>
- De Cáceres, M., & Legendre, P. (2009). Associations between species and groups of sites: indices and statistical inference. *Ecology*, 90, 3566–3574. <https://doi.org/10.1890/08-1823.1>
- DeLuca, J. G., Dong, Y., Hergert, P., Strauss, J., Hickey, J. M., Salmon, E. D., & McEwen, B. F. (2005). Hec1 and Nuf2 Are Core Components of the Kinetochore Outer Plate Essential for Organizing Microtubule Attachment Sites. *Molecular Biology of the Cell*, 16(2), 519–531. <https://doi.org/10.1091/mbc.e04-09-0852>
- de Nadal, E., Ammerer, G., & Posas, F. (2011). Controlling gene expression in response to stress. *Nature Reviews Genetics*, 12(12), 833–845. <https://doi.org/10.1038/nrg3055>
- Dibrova, E., Bibikov, S., Glagoleva, T., & Glagolev, A. (1985). The bacterial-type taxis and protein methylation in diatoms. *FEMS Microbiology Letters*, 26(3), 295–299. <https://doi.org/10.1111/j.1574-6968.1985.tb01614.x>
- Dick, C. F., Dos-Santos, A. L., & Meyer-Fernandes, J. R. (2014). Inorganic phosphate uptake in unicellular eukaryotes. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1840(7), 2123–2127. <https://doi.org/10.1016/j.bbagen.2014.03.014>
- Diner, R. E., Noddings, C. M., Lian, N. C., Kang, A. K., McQuaid, J. B., Jablanovic, J., Espinoza, J. L., Nguyen, N. A., Anzelmatti, M. A., Jansson, J., Bielinski, V. A., Karas, B. J., Dupont, C. L., Allen, A. E., & Weyman, P. D. (2017). Diatom centromeres suggest a mechanism for nuclear DNA acquisition. *Proceedings of the National Academy of Sciences*, 114(29). <https://doi.org/10.1073/pnas.1700764114>
- Dittami, S. M., Scornet, D., Petit, J.-L., Ségurens, B., Da Silva, C., Corre, E., Dondrup, M., Glattig, K.-H., König, R., Sterck, L., Rouzé, P., Van de Peer, Y., Cock, J. M., Boyen, C., & Tonon, T. (2009). Global expression analysis of the brown alga *Ectocarpus siliculosus* (Phaeophyceae) reveals large-scale reprogramming of the transcriptome in response to abiotic stress. *Genome Biology*, 10(6), R66. <https://doi.org/10.1186/gb-2009-10-6-r66>

- Doughty, T. W., Domenzain, I., Millan-Oropeza, A., Montini, N., de Groot, P. A., Pereira, R., Nielsen, J., Henry, C., Daran, J.-M. G., Siewers, V., & Morrissey, J. P. (2020). Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts. *Nature Communications*, 11(1). <https://doi.org/10.1038/s41467-020-16073-3>
- Dray, S., & Dufour, A.-B. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, 22(4). <https://doi.org/10.18637/jss.v022.i04>
- Duda, M. P., Sivarajah, B., Rühland, K. M., Paterson, A. M., Barrow, J. L., et al. (2023). Environmental optima for common diatoms from Ontario lakes along gradients of lakewater pH, total phosphorus concentration, and depth. *Journal of Paleolimnology*, 70(2), 131–158. <https://doi.org/10.1007/s10933-023-00288-7>
- Dufrêne, M., & Legendre, P. (1997). SPECIES ASSEMBLAGES AND INDICATOR SPECIES: THE NEED FOR A FLEXIBLE ASYMMETRICAL APPROACH. *Ecological Monographs*, 67(3), 345–366. [https://doi.org/10.1890/0012-9615\(1997\)067\[0345:saaist\]2.0.co;2](https://doi.org/10.1890/0012-9615(1997)067[0345:saaist]2.0.co;2)
- Dyhrman, S. T., Jenkins, B. D., Ryneerson, T. A., Saito, M. A., Mercier, M. L., Alexander, H., Whitney, L. P., Drzewianowski, A., Bulygin, V. V., Bertrand, E. M., Wu, Z., Benitez-Nelson, C., & Heithoff, A. (2012). The Transcriptome and Proteome of the Diatom *Thalassiosira pseudonana* Reveal a Diverse Phosphorus Stress Response (P. Santos, Ed.). *PLoS ONE*, 7(3), e33768. <https://doi.org/10.1371/journal.pone.0033768>
- Ekblom, R., & Galindo, J. (2010). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1), 1–15. <https://doi.org/10.1038/hdy.2010.152>
- Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1). <https://doi.org/10.1186/s13059-015-0721-2>
- Emms, D. M., & Kelly, S. (2017). STRIDE: Species Tree Root Inference from Gene Duplication Events. *Molecular Biology and Evolution*, 34(12), 3267–3278. <https://doi.org/10.1093/molbev/msx259>
- Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1). <https://doi.org/10.1186/s13059-019-1832-y>
- Emms, D., & Kelly, S. (2018). STAG: Species Tree Inference from All Genes, 267914. <https://doi.org/10.1101/267914>
- Engelken, J., Funk, C., & Adamska, I. (2011, August). The Extended Light-Harvesting Complex (LHC) Protein Superfamily: Classification and Evolutionary Dynamics. In *Advances in Photosynthesis and Respiration* (pp. 265–284). Springer Netherlands. [https://doi.org/10.1007/978-94-007-1533-2\\_11](https://doi.org/10.1007/978-94-007-1533-2_11)

- Erdene-Ochir, E., Shin, B.-K., Kwon, B., Jung, C., & Pan, C.-H. (2019). Identification and characterisation of the novel endogenous promoter HASP1 and its signal peptide from *Phaeodactylum tricornutum*. *Scientific Reports*, 9(1). <https://doi.org/10.1038/s41598-019-45786-9>
- Eskes, E., Deprez, M.-A., Wilms, T., & Winderickx, J. (2017). pH homeostasis in yeast; the phosphate perspective. *Current Genetics*, 64(1), 155–161. <https://doi.org/10.1007/s00294-017-0743-2>
- Falkowski, P. G., Katz, M. E., Knoll, A. H., Quigg, A., Raven, J. A., Schofield, O., & Taylor, F. J. R. (2004). The Evolution of Modern Eukaryotic Phytoplankton. *Science*, 305(5682), 354–360. <https://doi.org/10.1126/science.1095964>
- Fattorini, N., & Maier, U. G. (2021). Targeting of proteins to the cell wall of the diatom *Thalassiosira pseudonana*. *Discover Materials*, 1(1). <https://doi.org/10.1007/s43939-021-00005-z>
- Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl), W29–W37. <https://doi.org/10.1093/nar/gkr367>
- Flynn, K. J., Opik, H., & Syrett, P. J. (1986). Localization of the Alkaline Phosphatase and 5'-Nucleotidase Activities of the Diatom *Phaeodactylum tricornutum*. *Microbiology*, 132(2), 289–298. <https://doi.org/10.1099/00221287-132-2-289>
- Forterre, P., Gribaldo, S., Gadelle, D., & Serre, M.-C. (2007). Origin and evolution of DNA topoisomerases. *Biochimie*, 89(4), 427–446. <https://doi.org/10.1016/j.biochi.2006.12.009>
- Freedman, A. H., Clamp, M., & Sackton, T. B. (2020). Error, noise and bias in de novo transcriptome assemblies. *Molecular Ecology Resources*, 21(1), 18–29. <https://doi.org/10.1111/1755-0998.13156>
- Gallaher, S. D., Craig, R. J., Ganesan, I., Purvine, S. O., McCorkle, S. R., Grimwood, J., Strenkert, D., Davidi, L., Roth, M. S., Jeffers, T. L., Lipton, M. S., Niyogi, K. K., Schmutz, J., Theg, S. M., Blaby-Haas, C. E., & Merchant, S. S. (2021). Widespread polycistronic gene expression in green algae. *Proceedings of the National Academy of Sciences*, 118(7). <https://doi.org/10.1073/pnas.2017714118>
- García-Campos, M. A., Espinal-Enríquez, J., & Hernández-Lemus, E. (2015). Pathway Analysis: State of the Art. *Frontiers in Physiology*, 6. <https://doi.org/10.3389/fphys.2015.00383>
- García-Pérez, M. A., Núñez-Antón, V., & Alcalá-Quintana, R. (2014). Analysis of residuals in contingency tables: Another nail in the coffin of conditional approaches to significance testing. *Behavior Research Methods*, 47(1), 147–161. <https://doi.org/10.3758/s13428-014-0472-0>
- Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., & Brown, P. O. (2000). Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes (P. A. Silver, Ed.).



- Molecular Biology of the Cell*, 11(12), 4241–4257. <https://doi.org/10.1091/mbc.11.12.4241>
- Geisel, N. (2011). Constitutive versus Responsive Gene Expression Strategies for Growth in Changing Environments (V. Brezina, Ed.). *PLoS ONE*, 6(11), e27033. <https://doi.org/10.1371/journal.pone.0027033>
- Ghalambor, C. K., McKay, J. K., Carroll, S. P., & Reznick, D. N. (2007). Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. *Functional Ecology*, 21(3), 394–407. <https://doi.org/10.1111/j.1365-2435.2007.01283.x>
- Gimmler, H., & Degenhard, B. (2001). Alkaliphilic and Alkali-Tolerant Algae. In *Algal Adaptation to Environmental Stresses* (pp. 291–321). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-59491-5\\_10](https://doi.org/10.1007/978-3-642-59491-5_10)
- Gostinčar, C., Zalar, P., & Gunde-Cimerman, N. (2022). No need for speed: slow development of fungi in extreme environments. *Fungal Biology Reviews*, 39, 1–14. <https://doi.org/10.1016/j.fbr.2021.11.002>
- Gottschalk, S., & Kahlert, M. (2012). Shifts in taxonomical and guild composition of littoral diatom assemblages along environmental gradients. *Hydrobiologia*, 694(1), 41–56. <https://doi.org/10.1007/s10750-012-1128-7>
- Gould, S. B., Waller, R. F., & McFadden, G. I. (2008). Plastid Evolution. *Annual Review of Plant Biology*, 59(1), 491–517. <https://doi.org/10.1146/annurev.arplant.59.032607.092915>
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., . . . Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7), 644–652. <https://doi.org/10.1038/nbt.1883>
- Graves, S., Piepho, H.-P., & with help from Sundar Dorai-Raj, L. S. (2024). *multcompView: Visualizations of Paired Comparisons* [R package version 0.1-10]. <https://CRAN.R-project.org/package=multcompView>
- Gross, W. (2000). Ecophysiology of algae living in highly acidic environments. *Hydrobiologia*, 433(1-3), 31–37. <https://doi.org/10.1023/a:1004054317446>
- Grossman, A., Manodori, A., & Snyder, D. (1990). Light-harvesting proteins of diatoms: Their relationship to the chlorophyll a/b binding proteins of higher plants and their mode of transport into plastids. *Molecular and General Genetics MGG*, 224(1), 91–100. <https://doi.org/10.1007/bf00259455>
- Gruber, A., McKay, C., Rocap, G., & Oborník, M. (2020). Comparison of different versions of SignalP and TargetP for diatom plastid protein predictions with ASAFind. *Matters*.
- Gruber, A., Rocap, G., Kroth, P. G., Armbrust, E. V., & Mock, T. (2015). Plastid proteome prediction for diatoms and other algae with secondary plastids of

- the red lineage. *The Plant Journal*, 81(3), 519–528. <https://doi.org/10.1111/tpj.12734>
- Guillard, R. R. L., & Lorenzen, C. J. (1972). YELLOW-GREEN ALGAE WITH CHLOROPHYLLIDE C<sub>1,2</sub>. *Journal of Phycology*, 8(1), 10–14. <https://doi.org/10.1111/j.1529-8817.1972.tb03995.x>
- Haberman, S. J. (1973). The Analysis of Residuals in Cross-Classified Tables. *Biometrics*, 29(1), 205–220. <http://www.jstor.org/stable/2529686>
- Halevy, I., & Bachan, A. (2017). The geologic history of seawater pH. *Science*, 355(6329), 1069–1071. <https://doi.org/10.1126/science.aal4151>
- Hejazian, S. M., Pirmoradi, S., Zununi Vahed, S., Kumar Roy, R., & Hosseiniyan Khatibi, S. M. (2024). An update on Glycerophosphodiester Phosphodiesterases; From Bacteria to Human. *The Protein Journal*, 43(2), 187–199. <https://doi.org/10.1007/s10930-024-10190-4>
- Helliwell, K. E., Harrison, E. L., Christie-Oleza, J. A., Rees, A. P., Kleiner, F. H., Gaikwad, T., Downe, J., Aguilo-Ferretjans, M. M., Al-Moosawi, L., Brownlee, C., & Wheeler, G. L. (2021). A Novel Ca<sup>2+</sup> Signaling Pathway Coordinates Environmental Phosphorus Sensing and Nitrogen Metabolism in Marine Diatoms. *Current Biology*, 31(5), 978–989.e4. <https://doi.org/10.1016/j.cub.2020.11.073>
- Herbstová, M., Bína, D., Koník, P., Gardian, Z., Vácha, F., & Litvín, R. (2015). Molecular basis of chromatic adaptation in pennate diatom *Phaeodactylum tricornutum*. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1847(6–7), 534–543. <https://doi.org/10.1016/j.bbabi.2015.02.016>
- Hirooka, S., Hirose, Y., Kanesaki, Y., Higuchi, S., Fujiwara, T., Onuma, R., Era, A., Ohbayashi, R., Uzuka, A., Nozaki, H., Yoshikawa, H., & Miyagishima, S.-y. (2017). Acidophilic green algal genome provides insights into adaptation to an acidic environment. *Proceedings of the National Academy of Sciences*, 114(39). <https://doi.org/10.1073/pnas.1707072114>
- Hirst, H., Chaud, F., Delabie, C., Jüttner, I., & Ormerod, S. J. (2004). Assessing the short-term response of stream diatoms to acidity using inter-basin transplantations and chemical diffusing substrates. *Freshwater Biology*, 49(8), 1072–1088. <https://doi.org/10.1111/j.1365-2427.2004.01242.x>
- Hoagland, K. D., Rosowski, J. R., Gretz, M. R., & Roemer, S. C. (1993). DIATOM EXTRACELLULAR POLYMERIC SUBSTANCES: FUNCTION, FINE STRUCTURE, CHEMISTRY, AND PHYSIOLOGY. *Journal of Phycology*, 29(5), 537–566. <https://doi.org/10.1111/j.0022-3646.1993.00537.x>
- Hopkinson, B. M., Dupont, C. L., & Matsuda, Y. (2016). The physiology and genetics of CO<sub>2</sub> concentrating mechanisms in model diatoms. *Current Opinion in Plant Biology*, 31, 51–57. <https://doi.org/10.1016/j.pbi.2016.03.013>
- Huang, L., Wang, L., Lin, X., Su, Y., Qin, Y., Kong, W., Zhao, L., Xu, X., & Yan, Q. (2017). mcp, aer, cheB, and cheV contribute to the regulation of Vibrio

- alginolyticus (jscp¿NDj/scp¿-01) adhesion under gradients of environmental factors. *MicrobiologyOpen*, 6(6). <https://doi.org/10.1002/mbo3.517>
- Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17(3), 377–386. <https://doi.org/10.1101/gr.5969107>
- Hwang, Y.-s., Jung, G., & Jin, E. (2008). Transcriptome analysis of acclimatory responses to thermal stress in Antarctic algae. *Biochemical and Biophysical Research Communications*, 367(3), 635–641. <https://doi.org/10.1016/j.bbrc.2007.12.176>
- Inomura, K., Pierella Karlusich, J. J., Dutkiewicz, S., Deutsch, C., Harrison, P. J., & Bowler, C. (2023). High Growth Rate of Diatoms Explained by Reduced Carbon Requirement and Low Energy Cost of Silica Deposition (A. Veach, Ed.). *Microbiology Spectrum*, 11(3). <https://doi.org/10.1128/spectrum.03311-22>
- Jaccard, P. (1901). Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines. 37, 241–272. <https://doi.org/10.5169/seals-266440>
- Jernigan, K. K., & Bordenstein, S. R. (2014). Ankyrin domains across the Tree of Life. *PeerJ*, 2, e264. <https://doi.org/10.7717/peerj.264>
- Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9), 1236–1240. <https://doi.org/10.1093/bioinformatics/btu031>
- Julius, M. L., & Theriot, E. C. (2010, September). The diatoms: a primer. In J. P. Smol & E. F. Stoermer (Eds.), *The Diatoms: Applications for the Environmental and Earth Sciences* (pp. 8–22). Cambridge University Press. <https://doi.org/10.1017/cbo9780511763175.003>
- Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Research*, 20(10), 1313–1326. <https://doi.org/10.1101/gr.101386.109>
- Kane, E. I., & Spratt, D. E. (2021). Structural Insights into Ankyrin Repeat-Containing Proteins and Their Influence in Ubiquitylation. *International Journal of Molecular Sciences*, 22(2), 609. <https://doi.org/10.3390/ijms22020609>
- Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1), 27–30. <https://doi.org/10.1093/nar/28.1.27>
- Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., & Ishiguro-Watanabe, M. (2022). KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1), D587–D592. <https://doi.org/10.1093/nar/gkac963>
- Kassen, R. (2002). The experimental evolution of specialists, generalists, and the maintenance of diversity. *Journal of Evolutionary Biology*, 15(2), 173–190. <https://doi.org/10.1046/j.1420-9101.2002.00377.x>

- Kateri, M. (2014). Analysis of Two-way Tables. In *Contingency Table Analysis* (pp. 17–61). Springer New York. [https://doi.org/10.1007/978-0-8176-4811-4\\_2](https://doi.org/10.1007/978-0-8176-4811-4_2)
- Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. In Y. Dodge (Ed.), *Statistical Data Analysis based on the L1 Norm* (pp. 405–416). Elsevier, Amsterdam.
- Kaufman, L., & Rousseeuw, P. J. (1990, March). Partitioning Around Medoids (Program PAM). In L. Kaufman & P. J. Rousseeuw (Eds.), *Finding Groups in Data: An Introduction to Cluster Analysis* (pp. 68–125). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9780470316801.ch2>
- Kaur, M., Saini, K. C., Ojah, H., Sahoo, R., Gupta, K., Kumar, A., & Bast, F. (2022). Abiotic stress in algae: response, signaling and transgenic approaches. *Journal of Applied Phycology*, 34(4), 1843–1869. <https://doi.org/10.1007/s10811-022-02746-7>
- Ke, H., Dass, S., Morrissey, J. M., Mather, M. W., & Vaidya, A. B. (2018). The mitochondrial ribosomal protein L13 is critical for the structural and functional integrity of the mitochondrion in *Plasmodium falciparum*. *Journal of Biological Chemistry*, 293(21), 8128–8137. <https://doi.org/10.1074/jbc.ra118.002552>
- Keck, F., Rimet, F., Franc, A., & Bouchez, A. (2016). Phylogenetic signal in diatom ecology: perspectives for aquatic ecosystems biomonitoring. *Ecological Applications*, 26(3), 861–872. Retrieved May 7, 2024, from <http://www.jstor.org/stable/24701991>
- Keeling, P. J. (2010). The Endosymbiotic Origin, Diversification and Fate of Plastids. *Philosophical Transactions: Biological Sciences*, 365(1541), 729–748. Retrieved May 9, 2024, from <http://www.jstor.org/stable/40538240>
- Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges (C. A. Ouzounis, Ed.). *PLoS Computational Biology*, 8(2), e1002375. <https://doi.org/10.1371/journal.pcbi.1002375>
- Kolde, R. (2019). *pheatmap: Pretty Heatmaps* [R package version 1.0.12]. <https://CRAN.R-project.org/package=pheatmap>
- Kopylova, E., & N. (n.d.). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, 28(24). <https://doi.org/10.1093/bioinformatics/bts611>
- Kristiansen, J. (1996). 16. Dispersal of freshwater algae — a review. *Hydrobiologia*, 336(1–3), 151–157. <https://doi.org/10.1007/bf00010829>
- Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes<sup>1</sup> Edited by F. Cohen. *Journal of Molecular Biology*, 305(3), 567–580. <https://doi.org/10.1006/jmbi.2000.4315>
- Kumar, S., Baweja, P., & Sahoo, D. (2015). Diatoms: Yellow or Golden Brown Algae. In D. Sahoo & J. Seckbach (Eds.), *The Algae World. Cellular Origin, Life*

- in Extreme Habitats and Astrobiology* (pp. 235–258). Springer Netherlands. [https://doi.org/10.1007/978-94-017-7321-8\\_8](https://doi.org/10.1007/978-94-017-7321-8_8)
- Lee, R. E. (2008). Heterokontophyta, Bacillariophyceae. In *Phycology* (pp. 369–408). Cambridge University Press.
- Legendre, P., & Legendre, L. (2012). Ecological resemblance. In *Numerical Ecology* (pp. 265–335). Elsevier. <https://doi.org/10.1016/b978-0-444-53868-0.50007-1>
- Lenski, R. E., & Travisano, M. (1994). Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences*, 91(15), 6808–6814. <https://doi.org/10.1073/pnas.91.15.6808>
- Lenski, R. E., Rose, M. R., Simpson, S. C., & Tadler, S. C. (1991). Long-Term Experimental Evolution in *Escherichia coli*. I. Adaptation and Divergence During 2,000 Generations. *The American Naturalist*, 138(6), 1315–1341. <https://doi.org/10.1086/285289>
- Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1). <https://doi.org/10.1186/1471-2105-12-323>
- Li, C., Pan, Y., Yin, W., Liu, J., & Hu, H. (2024). A key gene, violaxanthin de-epoxidase-like 1, enhances fucoxanthin accumulation in *Phaeodactylum tricornutum*. *Biotechnology for Biofuels and Bioproducts*, 17(1). <https://doi.org/10.1186/s13068-024-02496-3>
- Li, J., Agarwal, E., Bertolini, I., Seo, J. H., Caino, M. C., Ghosh, J. C., Kossenkova, A. V., Liu, Q., Tang, H.-Y., Goldman, A. R., Languino, L. R., Speicher, D. W., & Altieri, D. C. (2020). The mitophagy effector FUNDC1 controls mitochondrial reprogramming and cellular plasticity in cancer cells. *Science Signaling*, 13(642). <https://doi.org/10.1126/scisignal.aaz8240>
- Li, J., Mahajan, A., & Tsai, M.-D. (2006). Ankyrin Repeat: A Unique Motif Mediating Protein-Protein Interactions. *Biochemistry*, 45(51), 15168–15178. <https://doi.org/10.1021/bi062188q>
- Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5), 713–714. <https://doi.org/10.1093/bioinformatics/btn025>
- Li, X.-W., Zhu, Y.-L., Chen, C.-Y., Geng, Z.-J., Li, X.-Y., Ye, T.-T., Mao, X.-N., & Du, F. (2020). Cloning and characterization of two chlorophyll A/B binding protein genes and analysis of their gene family in *Camellia sinensis*. *Scientific Reports*, 10(1). <https://doi.org/10.1038/s41598-020-61317-3>
- Li, Z., Zhang, Y., Li, W., Irwin, A. J., & Finkel, Z. V. (2023). Common environmental stress responses in a model marine diatom. *New Phytologist*, 240(1), 272–284. <https://doi.org/10.1111/nph.19147>
- Liu, L., Feng, D., Chen, G., Chen, M., Zheng, Ma, Q., Zhu, C., Wang, R., Qi, Xue, P., Li, B., Wang, X., Jin, H., Wang, J., Yang, F., Liu, P., Zhu, Y., Sui, S., & Chen, Q. (2012). Mitochondrial outer-membrane protein FUNDC1 mediates

- hypoxia-induced mitophagy in mammalian cells. *Nature Cell Biology*, 14(2), 177–185. <https://doi.org/10.1038/ncb2422>
- Longworth, J., Wu, D., Huete-Ortega, M., Wright, P. C., & Vaidyanathan, S. (2016). Proteome response of *Phaeodactylum tricornutum*, during lipid accumulation induced by nitrogen depletion. *Algal Research*, 18, 213–224. <https://doi.org/10.1016/j.algal.2016.06.015>
- López-Maury, L., Marguerat, S., & Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics*, 9(8), 583–593. <https://doi.org/10.1038/nrg2398>
- Lund, P., Tramonti, A., & De Biase, D. (2014). Coping with low pH: molecular strategies in neutrophilic bacteria. *FEMS Microbiology Reviews*, 38(6), 1091–1125. <https://doi.org/10.1111/1574-6976.12076>
- Lynch, M., Trickovic, B., & Kempes, C. P. (2022). Evolutionary scaling of maximum growth rate with organism size. *Scientific Reports*, 12(1). <https://doi.org/10.1038/s41598-022-23626-7>
- Ma, J., & Levin, S. A. (2006). The Evolution of Resource Adaptation: How Generalist and Specialist Consumers Evolve. *Bulletin of Mathematical Biology*, 68(5), 1111–1123. <https://doi.org/10.1007/s11538-006-9096-6>
- Ma, X., Zhang, B., Miao, R., Deng, X., Duan, Y., Cheng, Y., Zhang, W., Shi, M., Huang, K., & Xia, X.-Q. (2020). Transcriptomic and Physiological Responses to Oxidative Stress in a *Chlamydomonas reinhardtii* Glutathione Peroxidase Mutant. *Genes*, 11(4), 463. <https://doi.org/10.3390/genes11040463>
- Madshus, I. H. (1988). Regulation of intracellular pH in eukaryotic cells. *Biochemical Journal*, 250(1), 1–8. <https://doi.org/10.1042/bj2500001>
- Maleki, F., Ovens, K., Hogan, D. J., & Kuslik, A. J. (2020). Gene Set Analysis: Challenges, Opportunities, and Future Research. *Frontiers in Genetics*, 11. <https://doi.org/10.3389/fgene.2020.00654>
- Mardanov, A. V., Kadnikov, V. V., & Ravin, N. V. (2018). Metagenomics: A Paradigm Shift in Microbiology. In *Metagenomics* (pp. 1–13). Elsevier. <https://doi.org/10.1016/b978-0-08-102268-9.00001-x>
- Matsui, H., Harada, H., Maeda, K., Sugiyama, T., Fukuchi, Y., Kimura, N., Nawaly, H., Tsuji, Y., & Matsuda, Y. (2023). Coordinated phosphate uptake by extracellular alkaline phosphatase and solute carrier transporters in marine diatoms. *New Phytologist*, 241(3), 1210–1221. <https://doi.org/10.1111/nph.19410>
- McCullough, J., Frost, A., & Sundquist, W. I. (2018). Structures, Functions, and Dynamics of ESCRT-III/Vps4 Membrane Remodeling and Fission Complexes. *Annual Review of Cell and Developmental Biology*, 34(1), 85–109. <https://doi.org/10.1146/annurev-cellbio-100616-060600>
- McKie, S. J., Neuman, K. C., & Maxwell, A. (2021). DNA topoisomerases: Advances in understanding of cellular roles and multi-protein complexes via structure-function analysis. *BioEssays*, 43(4). <https://doi.org/10.1002/bies.202000286>

- Medlin, L. K. (2011). The Permian–Triassic mass extinction forces the radiation of the modern marine phytoplankton. *Phycologia*, 50(6), 684–693. <https://doi.org/10.2216/10-31.1>
- Medlin, L. K. (2016). Evolution of the diatoms: major steps in their evolution and a review of the supporting molecular and morphological evidence. *Phycologia*, 55(1), 79–103. <https://doi.org/10.2216/15-105.1>
- Mehra, P., Pandey, B. K., Verma, L., & Giri, J. (2018). A novel glycerophosphodiester phosphodiesterase improves phosphate deficiency tolerance in rice. *Plant, Cell & Environment*, 42(4), 1167–1179. <https://doi.org/10.1111/pce.13459>
- Messerli, M. A., Amaral-Zettler, L. A., Zettler, E., Jung, S.-K., Smith, P. J. S., & Sogin, M. L. (2005). Life at acidic pH imposes an increased energetic cost for a eukaryotic acidophile. *Journal of Experimental Biology*, 208(13), 2569–2579. <https://doi.org/10.1242/jeb.01660>
- Michaeli, S. (2014). Non-coding RNA and the complex regulation of the trypanosome life cycle. *Current Opinion in Microbiology*, 20, 146–152. <https://doi.org/10.1016/j.mib.2014.06.006>
- Mikami, K., Takio, S., Hiwatashi, Y., & Kumar, M. (2021). Editorial: Environmental Stress-Promoting Responses in Algae. *Frontiers in Marine Science*, 8. <https://doi.org/10.3389/fmars.2021.797613>
- Mimura, T., & Reid, R. (2024). Phosphate environment and phosphate uptake studies: past and future. *Journal of Plant Research*, 137(3), 307–314. <https://doi.org/10.1007/s10265-024-01520-9>
- Mirete, S., Morgante, V., & González-Pastor, J. E. (2017). Acidophiles: Diversity and Mechanisms of Adaptation to Acidic Environments. In *Adaption of Microbial Life to Environmental Extremes* (pp. 227–251). Springer International Publishing. [https://doi.org/10.1007/978-3-319-48327-6\\_9](https://doi.org/10.1007/978-3-319-48327-6_9)
- Mora, A. (2019). Gene set analysis methods for the functional interpretation of non-mRNA data—Genomic range and ncRNA data. *Briefings in Bioinformatics*, 21(5), 1495–1508. <https://doi.org/10.1093/bib/bbz090>
- Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schäffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics*, 24(16), 1757–1764. <https://doi.org/10.1093/bioinformatics/btn322>
- Nagai, T., Taya, K., Annoh, H., & Ishihara, S. (2013). Application of a fluorometric microplate algal toxicity assay for riverine periphytic algal species. *Ecotoxicology and Environmental Safety*, 94, 37–44. <https://doi.org/10.1016/j.ecoenv.2013.04.020>
- Nagao, R., Yokono, M., Ueno, Y., Shen, J.-R., & Akimoto, S. (2020). Acidic pH-Induced Modification of Energy Transfer in Diatom Fucoxanthin Chlorophyll a/c-Binding Proteins. *The Journal of Physical Chemistry B*, 124(24), 4919–4923. <https://doi.org/10.1021/acs.jpcc.0c04231>

- Nakajima, K., Tanaka, A., & Matsuda, Y. (2013). SLC4 family transporters in a marine diatom directly pump bicarbonate from seawater. *Proceedings of the National Academy of Sciences*, 110(5), 1767–1772. <https://doi.org/10.1073/pnas.1216234110>
- Nakov, T., Beaulieu, J. M., & Alverson, A. J. (2018). Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). *New Phytologist*, 219(1), 462–473. <https://doi.org/10.1111/nph.15137>
- Nakov, T., Beaulieu, J. M., & Alverson, A. J. (2019). Diatoms diversify and turn over faster in freshwater than marine environments\*. *Evolution*, 73(12), 2497–2511. <https://doi.org/10.1111/evo.13832>
- Nyström, T. (2004). MicroReview: Growth versus maintenance: a trade-off dictated by RNA polymerase availability and sigma factor competition? *Molecular Microbiology*, 54(4), 855–862. <https://doi.org/10.1111/j.1365-2958.2004.04342.x>
- Ogbunugafor, C. B., & Eppstein, M. J. (2016). Competition along trajectories governs adaptation rates towards antimicrobial resistance. *Nature Ecology & Evolution*, 1(1). <https://doi.org/10.1038/s41559-016-0007>
- Ogbunugafor, C. B., & Eppstein, M. J. (2019). Genetic Background Modifies the Topography of a Fitness Landscape, Influencing the Dynamics of Adaptive Evolution. *IEEE Access*, 7, 113675–113683. <https://doi.org/10.1109/ACCESS.2019.2935911>
- Osuna-Cruz, C. M., Bilcke, G., Vancaester, E., De Decker, S., Bones, A. M., et al. (2020). The *Seminais robusta* genome provides insights into the evolutionary adaptations of benthic diatoms. *Nature Communications*, 11(1). <https://doi.org/10.1038/s41467-020-17191-8>
- Park, S., Jung, G., Hwang, Y.-s., & Jin, E. (2009). Dynamic response of the transcriptome of a psychrophilic diatom, *Chaetoceros neogracile*, to high irradiance. *Planta*, 231(2), 349–360. <https://doi.org/10.1007/s00425-009-1044-x>
- Patrick, R., Roberts, N. A., & Davis, B. (1968). The effect of changes in pH on the structure of diatom communities.
- Payne, J. L., & Wagner, A. (2018). The causes of evolvability and their evolution. *Nature Reviews Genetics*, 20(1), 24–38. <https://doi.org/10.1038/s41576-018-0069-z>
- Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., et al. (2022). InterPro in 2022. *Nucleic Acids Research*, 51(D1), D418–D427. <https://doi.org/10.1093/nar/gkac993>
- Paytan, A., & McLaughlin, K. (2007). The Oceanic Phosphorus Cycle. *Chemical Reviews*, 107(2), 563–576. <https://doi.org/10.1021/cr0503613>
- Petrou, K., Baker, K. G., Nielsen, D. A., Hancock, A. M., Schulz, K. G., & Davidson, A. T. (2019). Acidification diminishes diatom silica production in the Southern



- Ocean. *Nature Climate Change*, 9(10), 781–786. <https://doi.org/10.1038/s41558-019-0557-y>
- Pfützner, A.-K., Zivkovic, H., Bernat-Silvestre, C., West, M., Peltier, T., Humbert, F., Odorizzi, G., & Roux, A. (2023). Vps60 initiates alternative ESCRT-III filaments. *Journal of Cell Biology*, 222(11). <https://doi.org/10.1083/jcb.202206028>
- Piepho, H.-P. (2004). An Algorithm for a Letter-Based Representation of All-Pairwise Comparisons. *Journal of Computational and Graphical Statistics*, 13(2), 456–466. <https://doi.org/10.1198/1061860043515>
- Pinheiro, J. P. S., Windsor, F. M., Wilson, R. W., & Tyler, C. R. (2021). Global variation in freshwater physico-chemistry and its influence on chemical toxicity in aquatic wildlife. *Biological Reviews*, 96(4), 1528–1546. <https://doi.org/10.1111/brv.12711>
- Pinseel, E., Nakov, T., Van den Berge, K., Downey, K. M., Judy, K. J., Kourtchenko, O., Kremp, A., Ruck, E. C., Sjöqvist, C., Töpel, M., Godhe, A., & Alverson, A. J. (2022). Strain-specific transcriptional responses overshadow salinity effects in a marine diatom sampled along the Baltic Sea salinity cline. *The ISME Journal*, 16(7), 1776–1787. <https://doi.org/10.1038/s41396-022-01230-x>
- Poulsen, N., Davutoglu, M. G., & Zackova Suchanova, J. (2022). Diatom Adhesion and Motility. In *The Molecular Life of Diatoms* (pp. 367–393). Springer International Publishing. [https://doi.org/10.1007/978-3-030-92499-7\\_14](https://doi.org/10.1007/978-3-030-92499-7_14)
- Putnam, R. W. (2012). Intracellular pH Regulation. In *Cell Physiology Source Book* (pp. 303–321). Elsevier. <https://doi.org/10.1016/b978-0-12-387738-3.00017-2>
- R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
- Reid, R. J., & Smith, F. A. (2002). The cytoplasmic pH stat. In Z. Rengel (Ed.), *Handbook of Plant Growth: pH as the Master Variable* (pp. 49–71). Marcel Dekker Inc.
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Rogato, A., Richard, H., Sarazin, A., Voss, B., Cheminant Navarro, S., Champeimont, R., Navarro, L., Carbone, A., Hess, W. R., & Falciatore, A. (2014). The diversity of small non-coding RNAs in the diatom *Phaeodactylum tricornutum*. *BMC Genomics*, 15(1). <https://doi.org/10.1186/1471-2164-15-698>
- Rydgren, K., Økland, R. H., & Økland, T. (2003). Species response curves along environmental gradients. A case study from SE Norwegian swamp forests. *Journal of Vegetation Science*, 14(6), 869–880. <https://doi.org/10.1111/j.1654-1103.2003.tb02220.x>
- Sahoo, S. A., Raghuvanshi, R., Srivastava, A. K., & Suprasanna, P. (2020). Phosphatases: The Critical Regulator of Abiotic Stress Tolerance in Plants. In *Protein Phosphatases and Stress Management in Plants* (pp. 163–201).

- Springer International Publishing. [https://doi.org/10.1007/978-3-030-48733-1\\_10](https://doi.org/10.1007/978-3-030-48733-1_10)
- Sakano, K. (2001). Metabolic regulation of pH in plant cells: Role of cytoplasmic pH in defense reaction and secondary metabolism. In *International Review of Cytology* (pp. 1–44). Elsevier. [https://doi.org/10.1016/s0074-7696\(01\)06018-1](https://doi.org/10.1016/s0074-7696(01)06018-1)
- Scarsini, M., Marchand, J., Manoylov, K. M., & Schoefs, B. (2019, June). Photosynthesis in Diatoms. In J. Seckbach & R. Gordon (Eds.), *Diatoms: Fundamentals and Applications* (pp. 191–211). Wiley. <https://doi.org/10.1002/9781119370741.ch8>
- Schafer, F. Q., & Buettner, G. R. (2000). Acidic pH amplifies iron-mediated lipid peroxidation in cells. *Free Radical Biology and Medicine*, 28(8), 1175–1181. [https://doi.org/10.1016/s0891-5849\(00\)00319-1](https://doi.org/10.1016/s0891-5849(00)00319-1)
- Schneider, S. C., Kahlert, M., & Kelly, M. G. (2013). Interactions between pH and nutrients on benthic algae in streams and consequences for ecological status assessment and species richness patterns. *Science of The Total Environment*, 444, 73–84. <https://doi.org/10.1016/j.scitotenv.2012.11.034>
- Schumacher, K., Brameyer, S., & Jung, K. (2023). Bacterial acid stress response: from cellular changes to antibiotic tolerance and phenotypic heterogeneity. *Current Opinion in Microbiology*, 75, 102367. <https://doi.org/10.1016/j.mib.2023.102367>
- Shen, C., Dupont, C. L., & Hopkinson, B. M. (2017). The diversity of CO<sub>2</sub>-concentrating mechanisms in marine diatoms as inferred from their genetic content. *Journal of Experimental Botany*, 68(14), 3937–3948. <https://doi.org/10.1093/jxb/erx163>
- Signorell, A. (2024). *DescTools: Tools for Descriptive Statistics* [R package version 0.99.54]. <https://CRAN.R-project.org/package=DescTools>
- Sjoberg, D. (2024). *ggsankey: Sankey, Alluvial and Sankey Bump Plots* [R package version 0.0.99999].
- Sokolova, I. M. (2013). Energy-Limited Tolerance to Stress as a Conceptual Framework to Integrate the Effects of Multiple Stressors. *Integrative and Comparative Biology*, 53(4), 597–608. <https://doi.org/10.1093/icb/ict028>
- Song, J., Wei, X., Shao, G., Sheng, Z., Chen, D., Liu, C., Jiao, G., Xie, L., Tang, S., & Hu, P. (2013). The rice nuclear gene WLP1 encoding a chloroplast ribosome L13 protein is needed for chloroplast development in rice grown under low temperature conditions. *Plant Molecular Biology*, 84(3), 301–314. <https://doi.org/10.1007/s11103-013-0134-0>
- Stevenson, R. J., Pan, Y., & van Dam, H. (2010, September). Assessing environmental conditions in rivers and streams with diatoms. In J. P. Smol & E. F. Stoermer (Eds.), *The Diatoms: Applications for the Environmental and Earth Sciences* (pp. 57–85). Cambridge University Press. <https://doi.org/10.1017/cbo9780511763175.005>

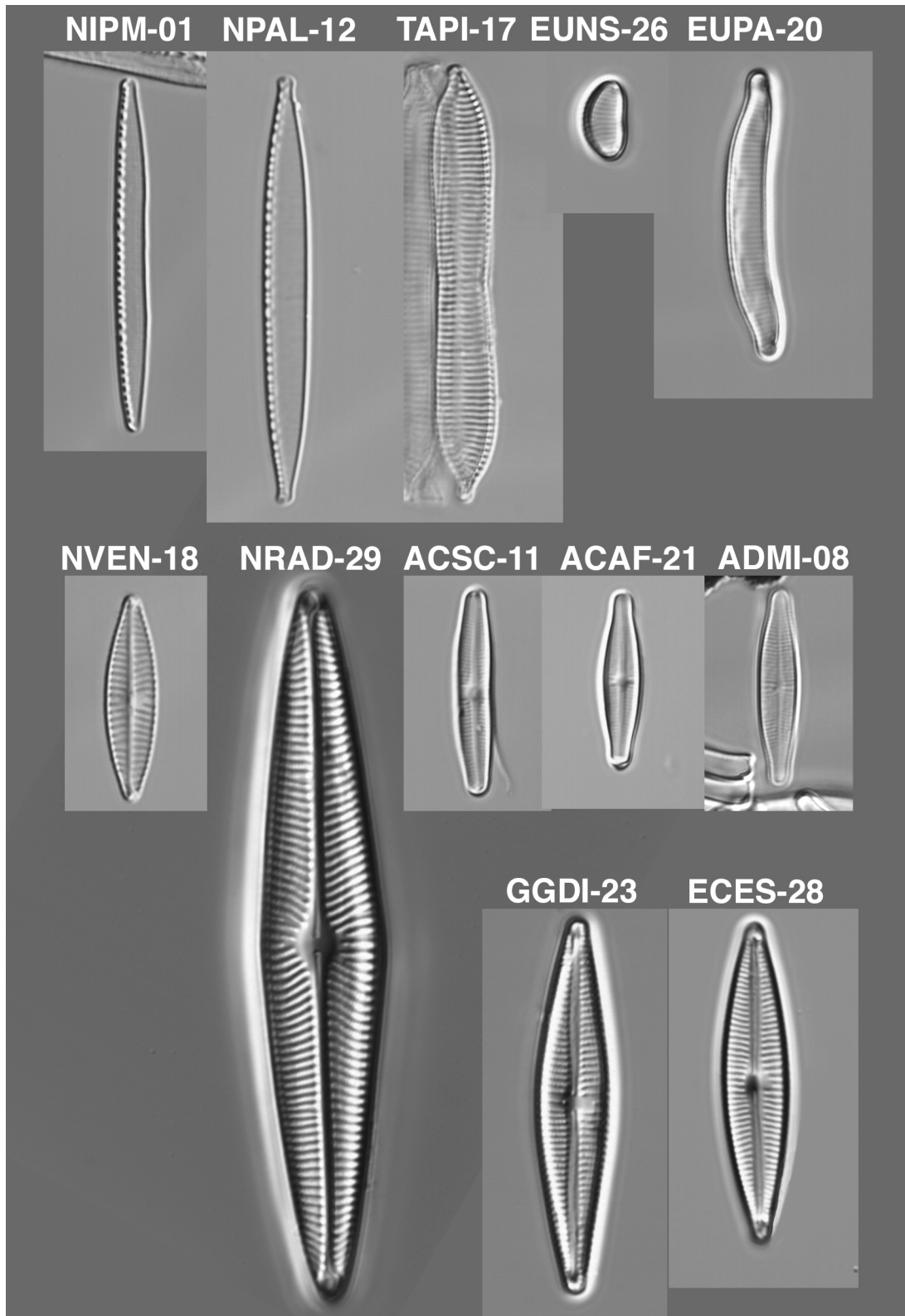
- Stock, W., Blommaert, L., Daveloose, I., Vyverman, W., & Sabbe, K. (2019). Assessing the suitability of Imaging-PAM fluorometry for monitoring growth of benthic diatoms. *Journal of Experimental Marine Biology and Ecology*, 513, 35–41. <https://doi.org/10.1016/j.jembe.2019.02.003>
- Storey, J. M., & Storey, K. B. (2023). Chaperone proteins: universal roles in surviving environmental stress. *Cell Stress and Chaperones*, 28(5), 455–466. <https://doi.org/10.1007/s12192-022-01312-x>
- Telford, R. J., Vandvik, V., & Birks, H. J. B. (2006). Dispersal Limitations Matter for Microbial Morphospecies. *Science*, 312(5776), 1015–1015. <https://doi.org/10.1126/science.1125669>
- ter Braak, C. J. F., & van Dame, H. (1989). Inferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia*, 178(3), 209–223. <https://doi.org/10.1007/bf00006028>
- Truong, T. Q., Park, Y. J., Koo, S. Y., Choi, J.-H., Enkhbayar, A., Song, D.-G., & Kim, S. M. (2022). Interdependence of fucoxanthin biosynthesis and fucoxanthin-chlorophyll a/c binding proteins in *Phaeodactylum tricornutum* under different light intensities. *Journal of Applied Phycology*, 35(1), 25–42. <https://doi.org/10.1007/s10811-022-02856-2>
- Tsuji, Y., Nakajima, K., & Matsuda, Y. (2017). Molecular aspects of the biophysical CO<sub>2</sub>-concentrating mechanism and its regulation in marine diatoms. *Journal of Experimental Botany*, 68(14), 3763–3772. <https://doi.org/10.1093/jxb/erx173>
- Urbánková, P., Vanormelingen, P., Sabbe, K., & Vyverman, W. (n.d.). *Experimentally determined pH preferences in a diatom species complex [Unpublished manuscript]*.
- Van Dam, H., Mertens, A., & Sinkeldam, J. (1994). A coded checklist and ecological indicator values of freshwater diatoms from The Netherlands. *Netherlands Journal of Aquatic Ecology*, 28(1), 117–133. <https://doi.org/10.1007/bf02334251>
- Van Tienderen, P. H. (1991). EVOLUTION OF GENERALISTS AND SPECIALISTS IN SPATIALLY HETEROGENEOUS ENVIRONMENTS. *Evolution*, 45(6), 1317–1331. <https://doi.org/10.1111/j.1558-5646.1991.tb02638.x>
- Vancaester, E., Depuydt, T., Osuna-Cruz, C. M., & Vandepoele, K. (2020). Comprehensive and Functional Analysis of Horizontal Gene Transfer Events in Diatoms (F. U. Battistuzzi, Ed.). *Molecular Biology and Evolution*, 37(11), 3243–3257. <https://doi.org/10.1093/molbev/msaa182>
- Väremo, L., Nielsen, J., & Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Research*, 41(8), 4378–4391. <https://doi.org/10.1093/nar/gkt111>
- Wadhams, G. H., & Armitage, J. P. (2004). Making sense of it all: bacterial chemotaxis. *Nature Reviews Molecular Cell Biology*, 5(12), 1024–1037. <https://doi.org/10.1038/nrm1524>
-

- Wagner, C. A. (2023). The basics of phosphate metabolism. *Nephrology Dialysis Transplantation*, 39(2), 190–201. <https://doi.org/10.1093/ndt/gfad188>
- Webb, C. O., Ackerly, D. D., McPeck, M. A., & Donoghue, M. J. (2002). Phylogenies and Community Ecology. *Annual Review of Ecology and Systematics*, 33(1), 475–505. <https://doi.org/10.1146/annurev.ecolsys.33.010802.150448>
- Wetzel, R. G. (2001). WATER AS A SUBSTANCE. In R. G. Wetzel (Ed.), *Limnology* (pp. 9–14). Elsevier. <https://doi.org/10.1016/b978-0-08-057439-4.50006-x>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wiens, J. J., & Graham, C. H. (2005). Niche Conservatism: Integrating Evolution, Ecology, and Conservation Biology. *Annual Review of Ecology, Evolution, and Systematics*, 36(1), 519–539. <https://doi.org/10.1146/annurev.ecolsys.36.102803.095431>
- Williams, D. M. (2020). Diatom Classifications: What Purpose Do They Serve? In G. Cristóbal, S. Blanco, & G. Bueno (Eds.), *Modern Trends in Diatom Identification* (pp. 11–24). Springer International Publishing. [https://doi.org/10.1007/978-3-030-39212-3\\_2](https://doi.org/10.1007/978-3-030-39212-3_2)
- Wong-ekkabut, J., Xu, Z., Triampo, W., Tang, I.-M., Peter Tieleman, D., & Monticelli, L. (2007). Effect of Lipid Peroxidation on the Properties of Lipid Bilayers: A Molecular Dynamics Study. *Biophysical Journal*, 93(12), 4225–4236. <https://doi.org/10.1529/biophysj.107.112565>
- Woodcock, D. J., Krusche, P., Strachan, N. J. C., Forbes, K. J., Cohan, F. M., Méric, G., & Sheppard, S. K. (2017). Genomic plasticity and rapid host switching can promote the evolution of generalism: a case study in the zoonotic pathogen *Campylobacter*. *Scientific Reports*, 7(1). <https://doi.org/10.1038/s41598-017-09483-9>
- Xiong, L., & Zhu, J.-K. (2001). Abiotic stress signal transduction in plants: Molecular and genetic perspectives. *Physiologia Plantarum*, 112(2), 152–166. <https://doi.org/10.1034/j.1399-3054.2001.1120202.x>
- Xu, Y.-H., Liu, R., Yan, L., Liu, Z.-Q., Jiang, S.-C., Shen, Y.-Y., Wang, X.-F., & Zhang, D.-P. (2011). Light-harvesting chlorophyll a/b-binding proteins are required for stomatal response to abscisic acid in *Arabidopsis*. *Journal of Experimental Botany*, 63(3), 1095–1106. <https://doi.org/10.1093/jxb/err315>
- Xue, T., Wan, H., Chen, J., He, S., Lujin, C., Xia, M., Wang, S., Dai, X., & Zeng, C. (2024). Genome-wide identification and expression analysis of the chlorophyll a/b binding protein gene family in oilseed (*Brassica napus* L.) under salt stress conditions. *Plant Stress*, 11, 100339. <https://doi.org/10.1016/j.stress.2023.100339>
- Yampolsky, L. Y., Zeng, E., Lopez, J., Williams, P. J., Dick, K. B., Colbourne, J. K., & Pfrender, M. E. (2014). Functional genomics of acclimation and adaptation in response to thermal stress in *Daphnia*. *BMC Genomics*, 15(1), 859. <https://doi.org/10.1186/1471-2164-15-859>

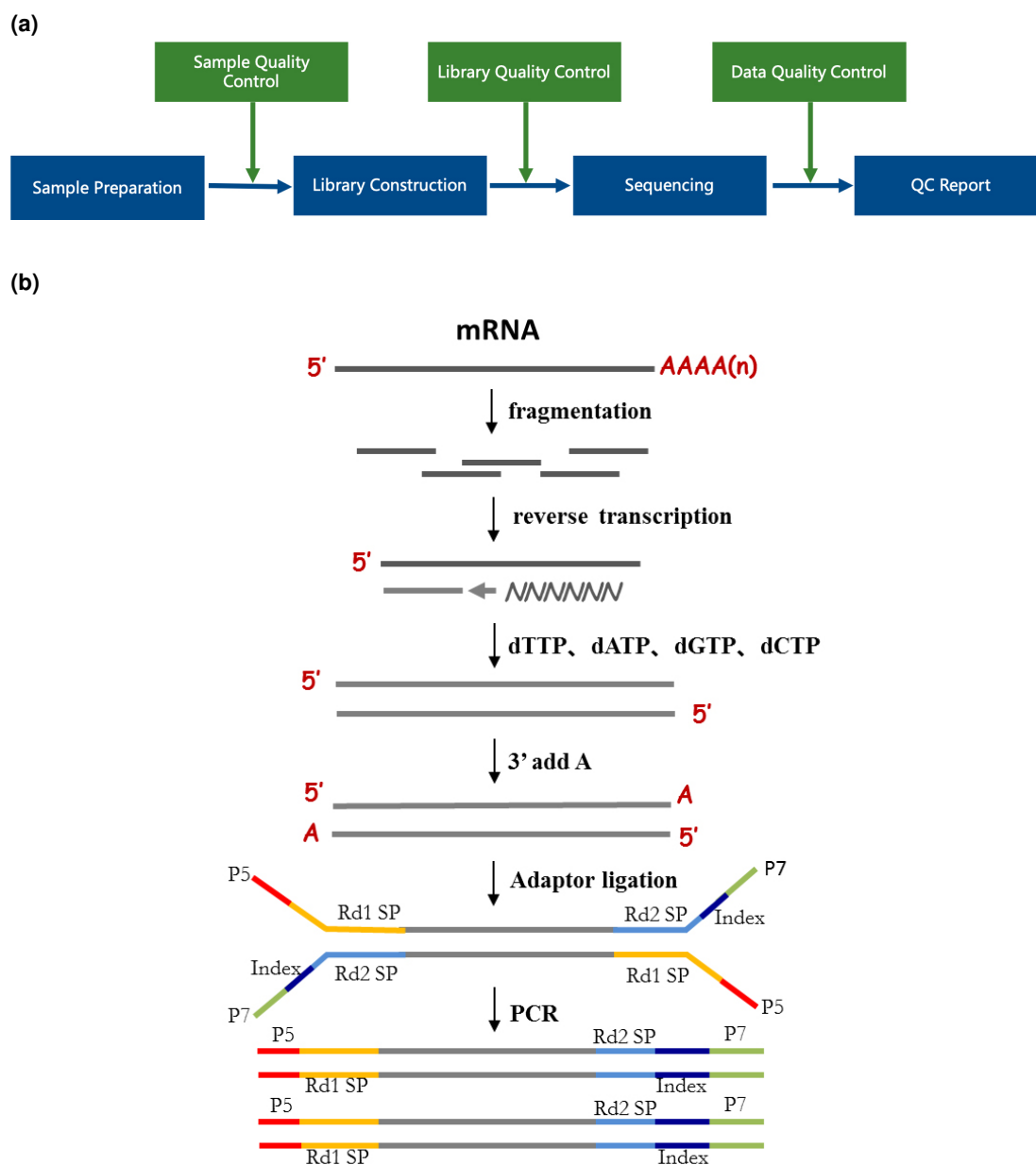
- Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G., & Bhattacharya, D. (2004). A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. *Molecular Biology and Evolution*, 21(5), 809–818. <https://doi.org/10.1093/molbev/msh075>
- Zakataeva, N. P. (2021). Microbial 5'-nucleotidases: their characteristics, roles in cellular metabolism, and possible practical applications. *Applied Microbiology and Biotechnology*, 105(20), 7661–7681. <https://doi.org/10.1007/s00253-021-11547-w>
- Zhao, Y.-y., Cao, C.-l., Liu, Y.-l., Wang, J., Li, S.-y., Li, J., & Deng, Y. (2020). Genetic analysis of oxidative and endoplasmic reticulum stress responses induced by cobalt toxicity in budding yeast. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1864(3), 129516. <https://doi.org/10.1016/j.bbagen.2020.129516>

## **Appendix A**

# **Supplementary methods**



**Figure A.1. Pictures of the twelve diatom strains.** Taxonomic information is described in Table 2.3. All strains are displayed on the same scale.



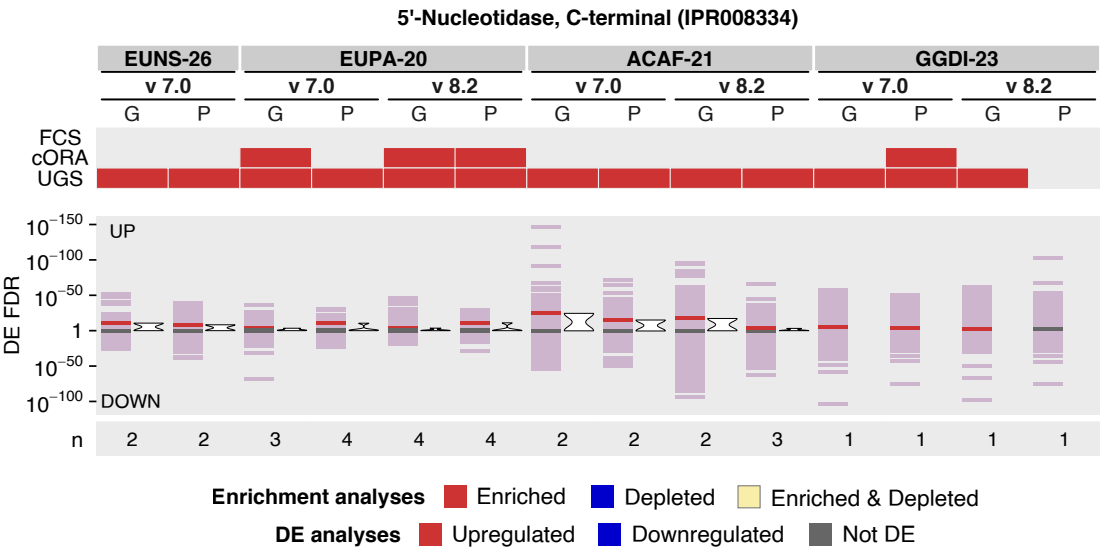
**Figure A.2. Protocol for library preparation and sequencing.** Images provided by Novogene Co., Ltd. (UK).



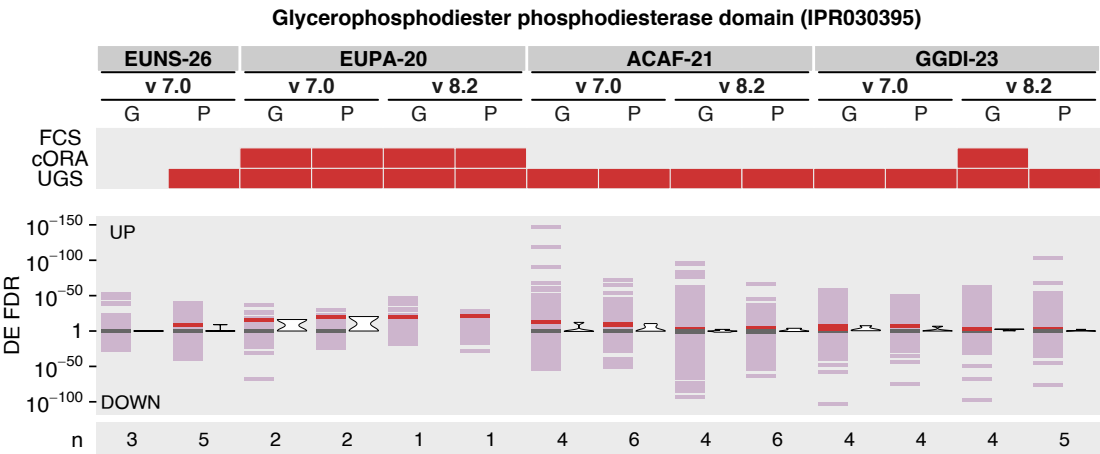


## **Appendix B**

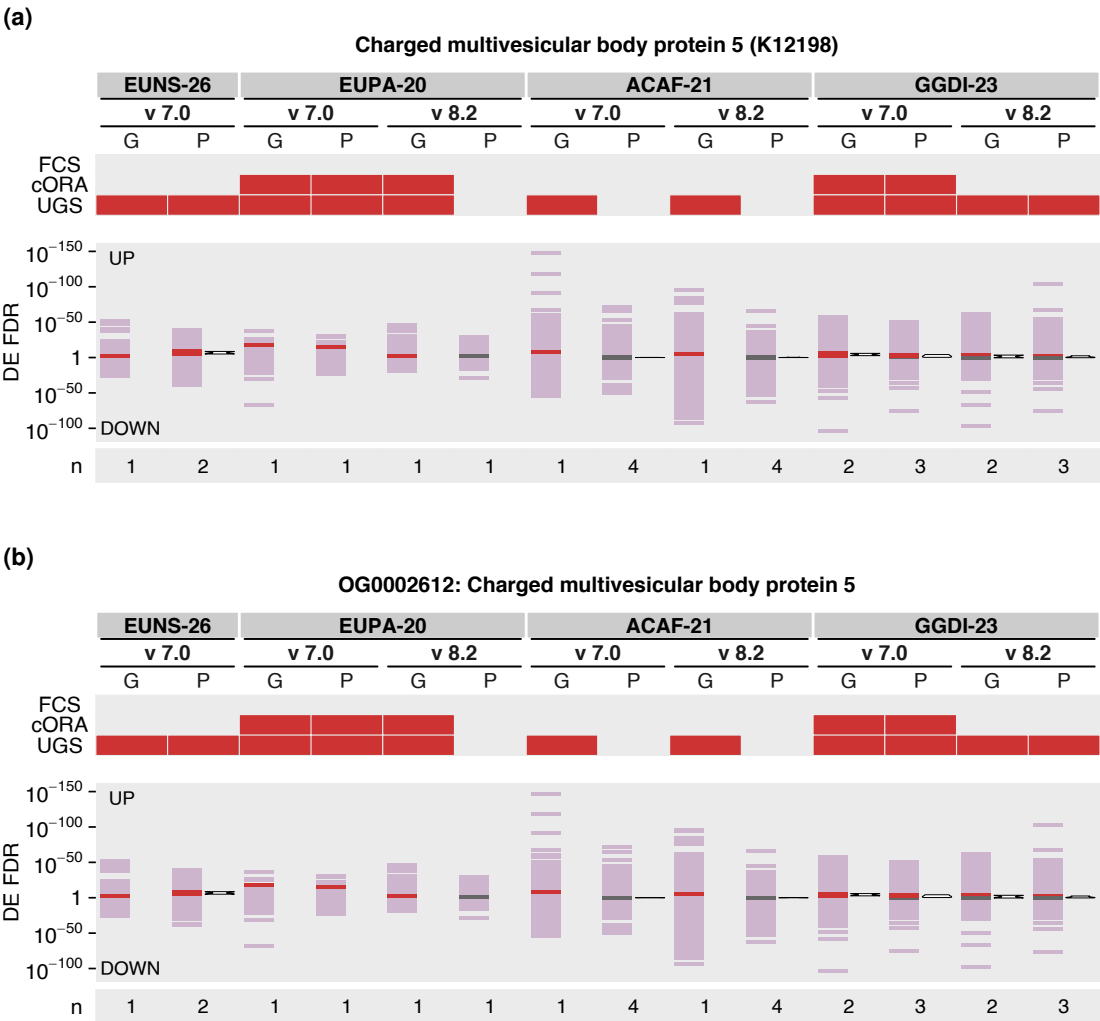
# **Summary plots for the enrichment analyses**



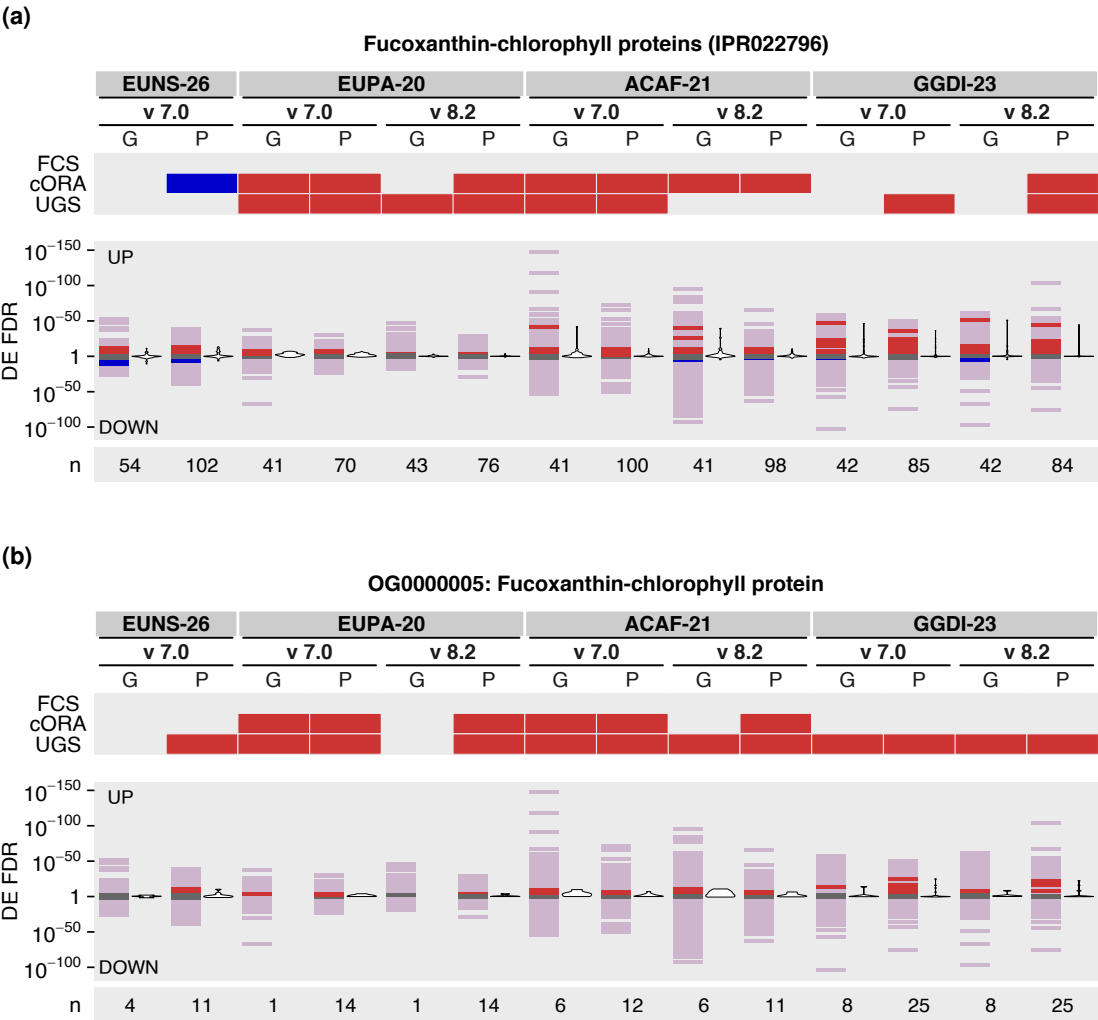
**Figure B.1. Summary of the enrichment analyses results at pH 4.7 for InterPro IPR008334, representing 5'-nucleotidases.** The top plot shows the significant enrichments and depletions at pH 4.7 across acid-tolerant strains according to each enrichment analysis (FCS, cORA, and UGS). Results from the gene ("G") and the protein ("P") datasets are shown. The bottom plot displays (following the same columns as the top plot) the position of the genes/proteins of the gene set along the significance gradient from the differential expression analyses, from most significantly upregulated at the top to the most significantly downregulated at the bottom. The position of these genes/proteins is represented with both tiles (colored according to DE result) and violins. Light purple tiles represent the position in the significance gradient of all genes/proteins of the strain included in the DE analyses (background). The number of genes/proteins from the gene set in each strain is displayed at the bottom of the plot.



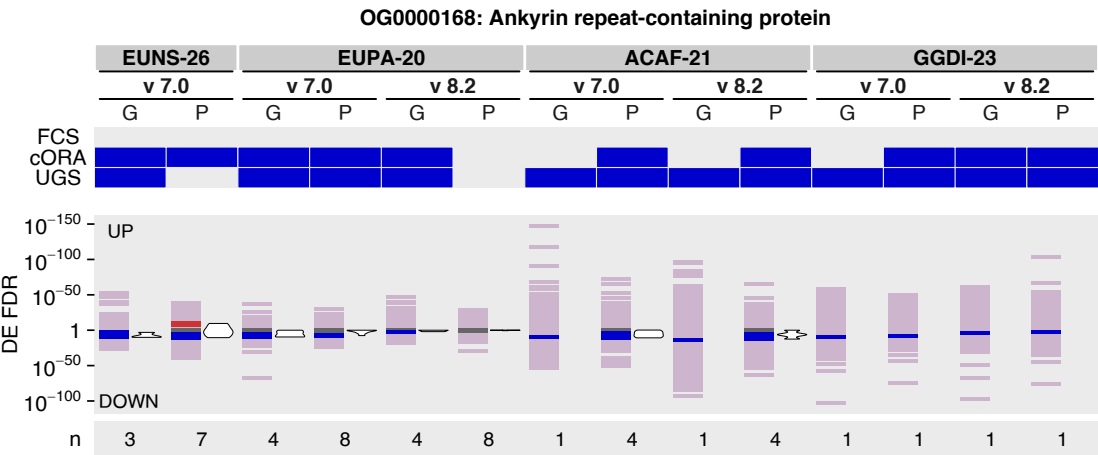
**Figure B.2. Summary of the enrichment analyses results at pH 4.7 for InterPro IPR030395, representing GDPD domain.** For a detailed description of the figure display, see Figure B.1.



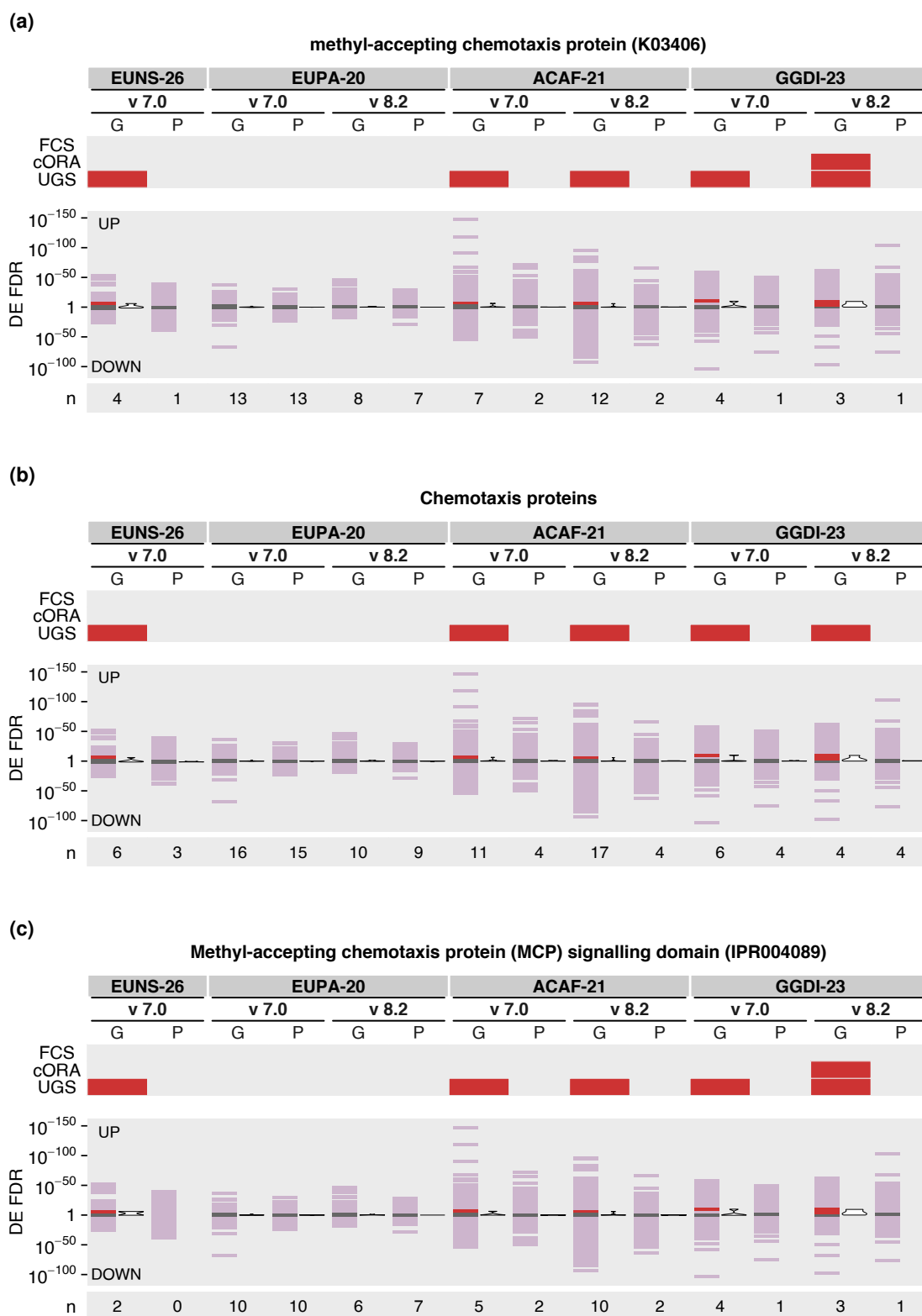
**Figure B.3. Summary of the enrichment analyses results at pH 4.7 for CHMP5** For a detailed description of the figure display, see Figure B.1.



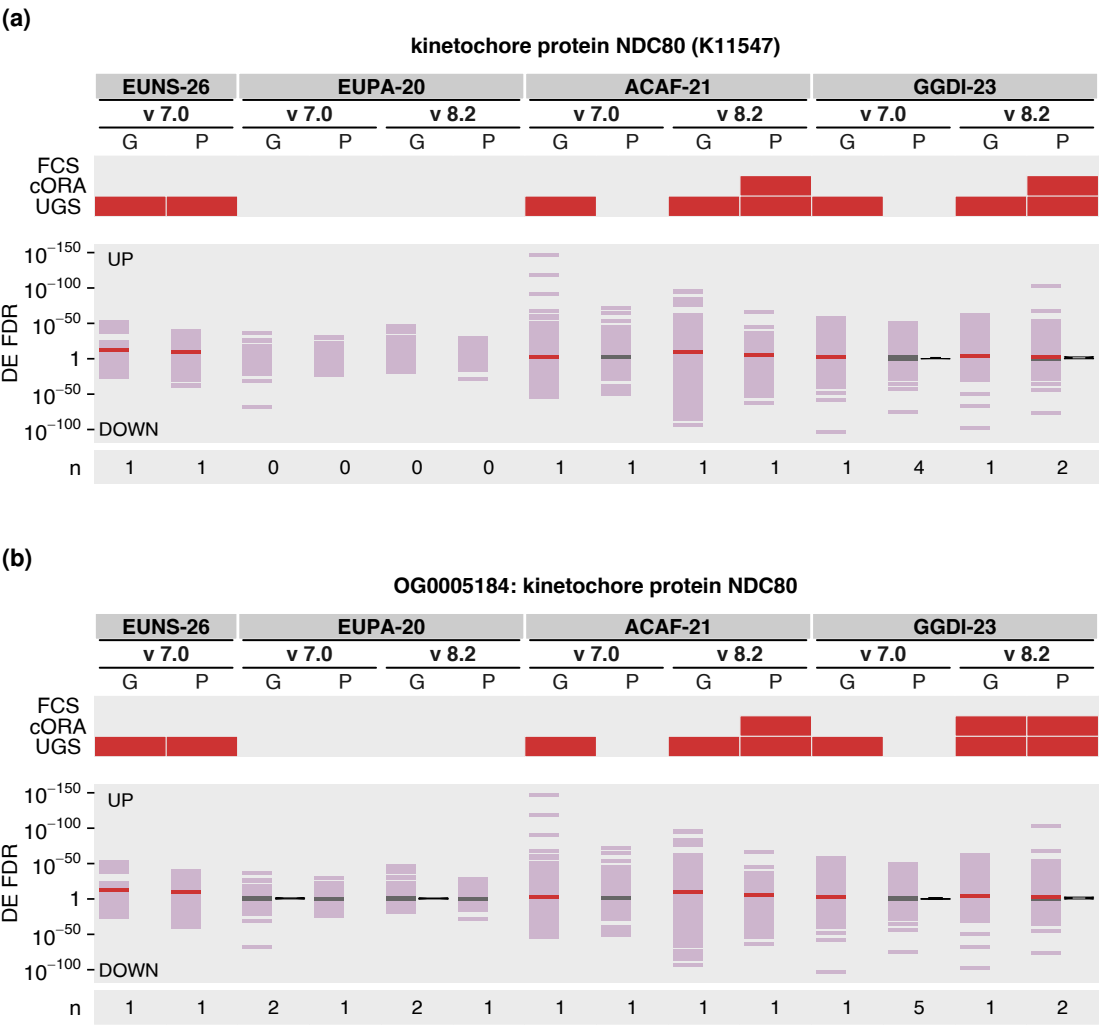
**Figure B.4. Summary of the enrichment analyses results at pH 4.7 for FCPs.** For a detailed description of the figure display, see Figure B.1.



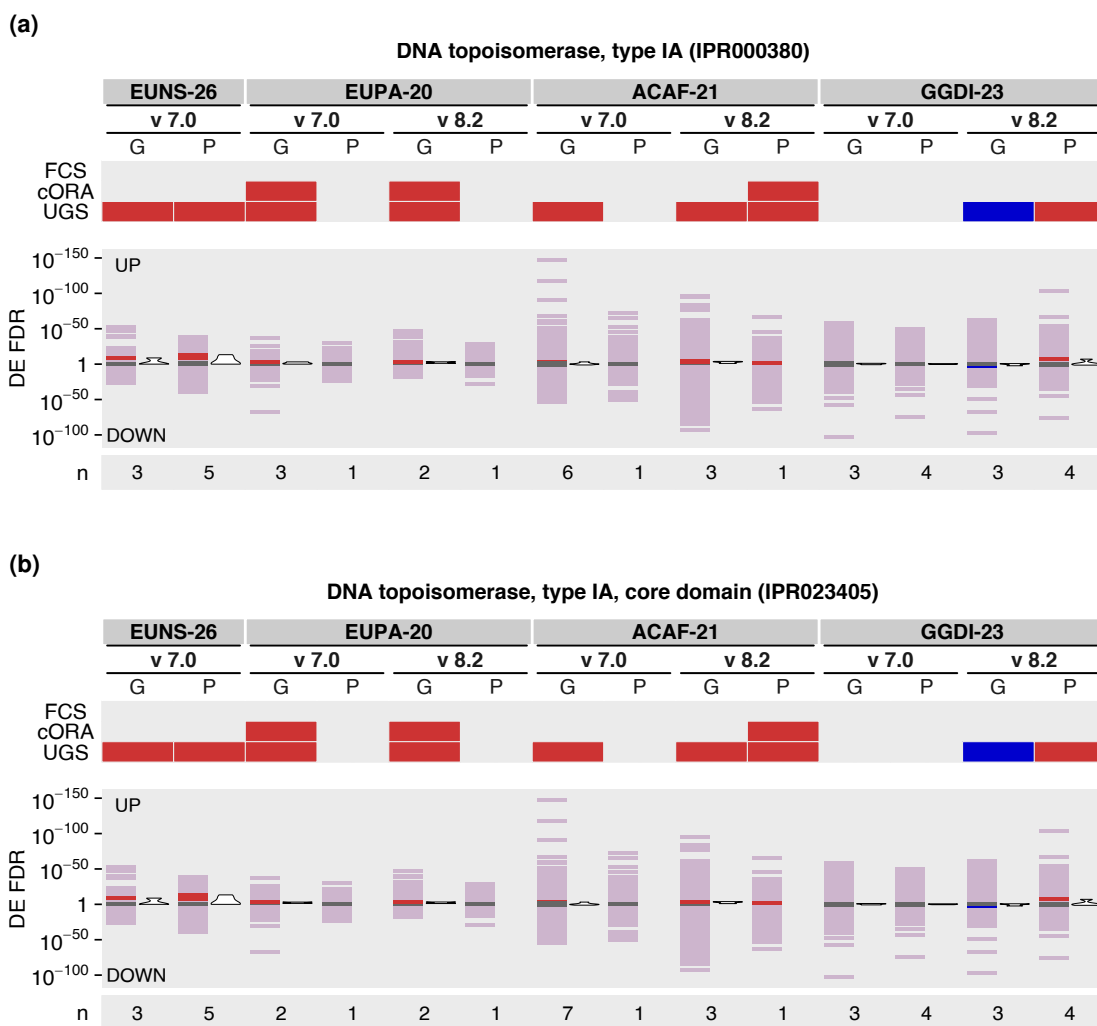
**Figure B.5. Summary of the enrichment analyses results at pH 4.7 for orthogroup OG0000168, representing an Ankyrin repeat-containing protein.** For a detailed description of the figure display, see Figure B.1.



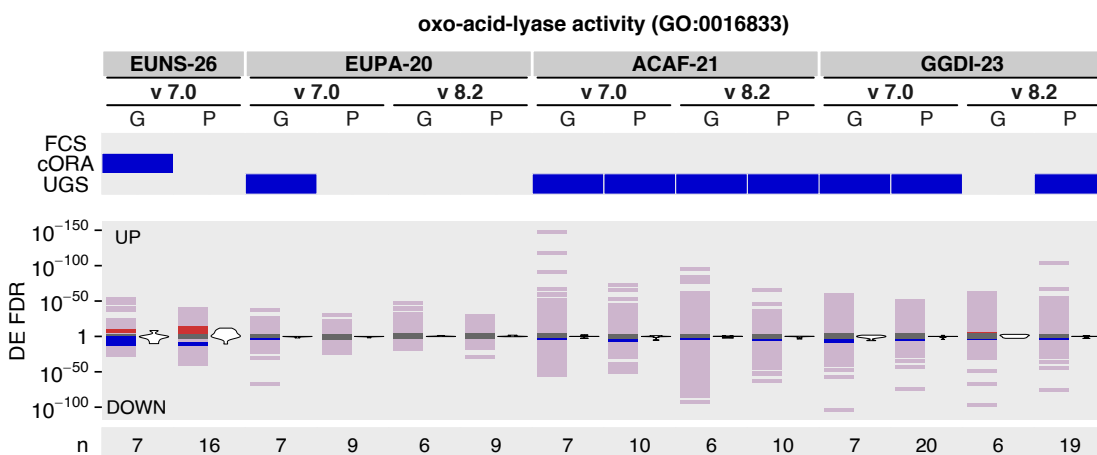
**Figure B.6. Summary of the enrichment analyses results at pH 4.7 for MCPs.** KEGG BRITE MCPs showed the same annotation as KEGG KO K03406 (a). For a detailed description of the figure display, see Figure B.1.



**Figure B.7. Summary of the enrichment analyses results at pH 4.7 for the kinetochore protein NDC80.** InterPro IPR005550 and GO terms GO:0007080, GO:0008608, GO:0050000, GO:0051303, GO:0051310, and GO:0051315 showed the same annotation as KEGG KO K11547 (a). InterPro IPR038273 was assigned to fewer proteins in GGDI-23 strain than the two sets displayed in this figure and is not shown. For a detailed description of the figure display, see Figure B.1.

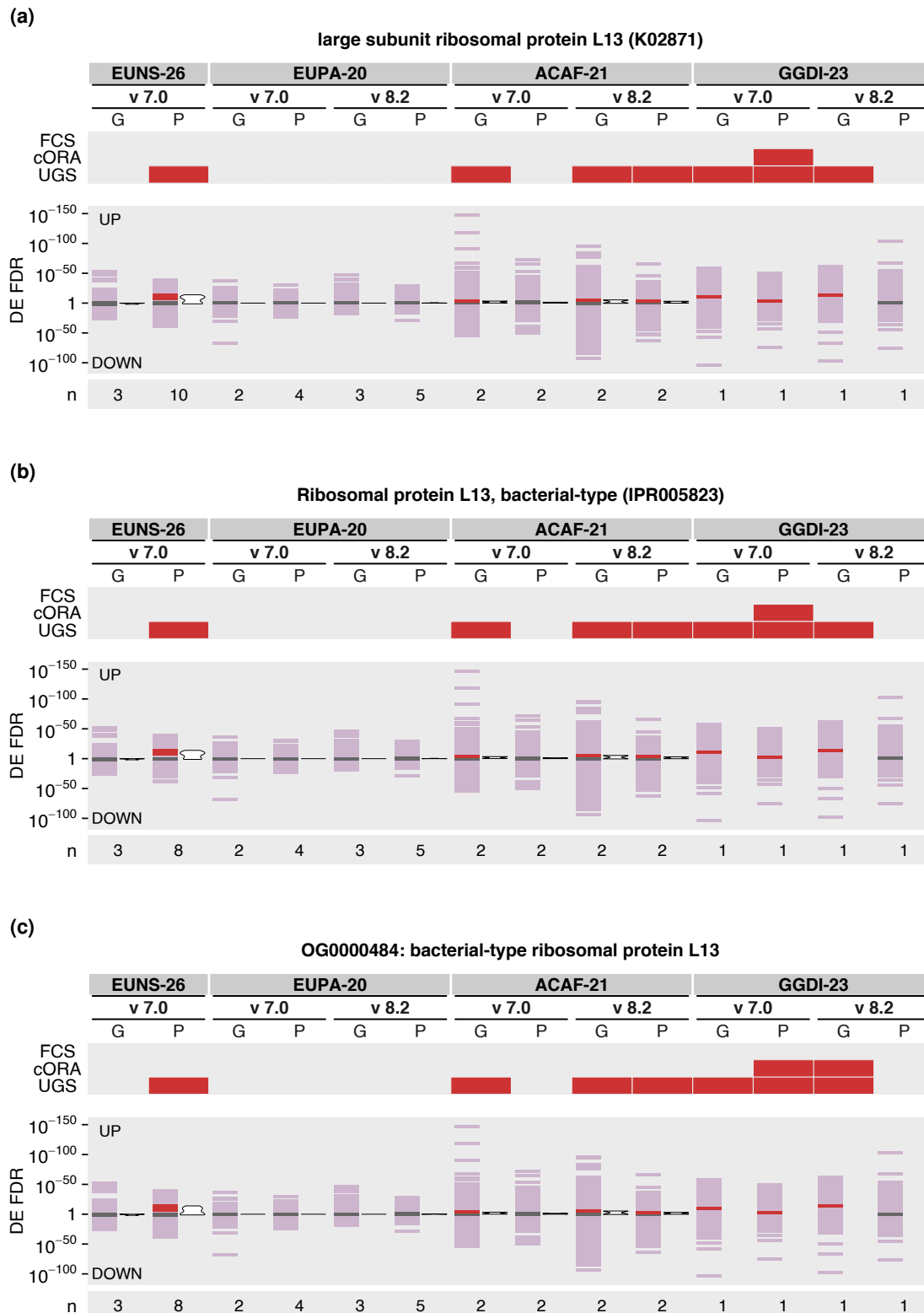


**Figure B.8. Summary of the enrichment analyses results at pH 4.7 for DNA topoisomerases type IA.** InterPro IPR013497, IPR013826, and IPR023406 were assigned to fewer proteins than the two InterPro sets displayed in this figure and are not shown. For a detailed description of the figure display, see Figure B.1.

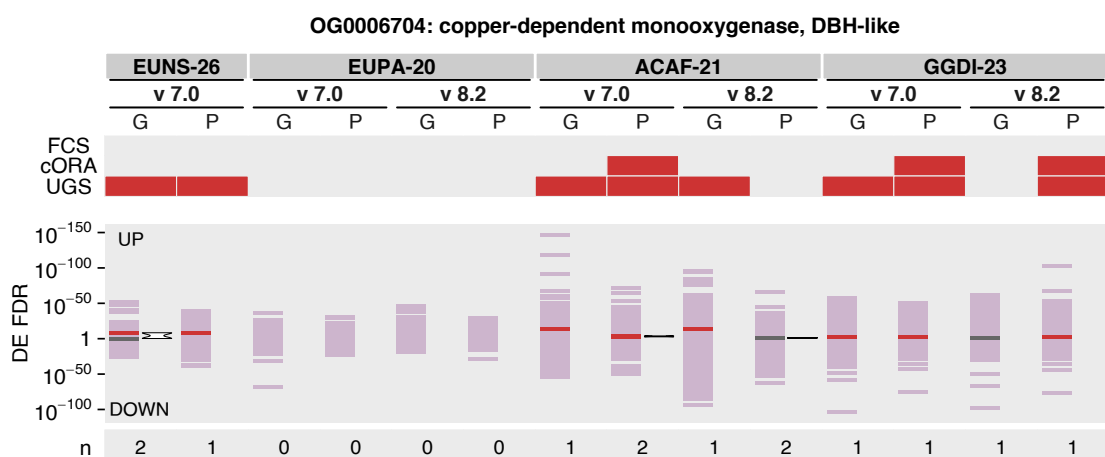


**Figure B.9. Summary of the enrichment analyses results at pH 4.7 for oxo-acid-lyases.** For a detailed description of the figure display, see Figure B.1.

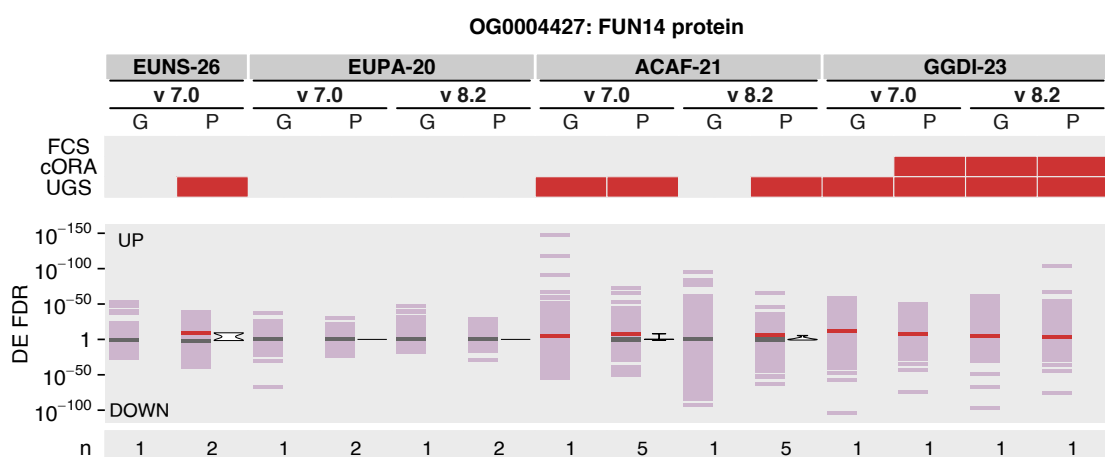




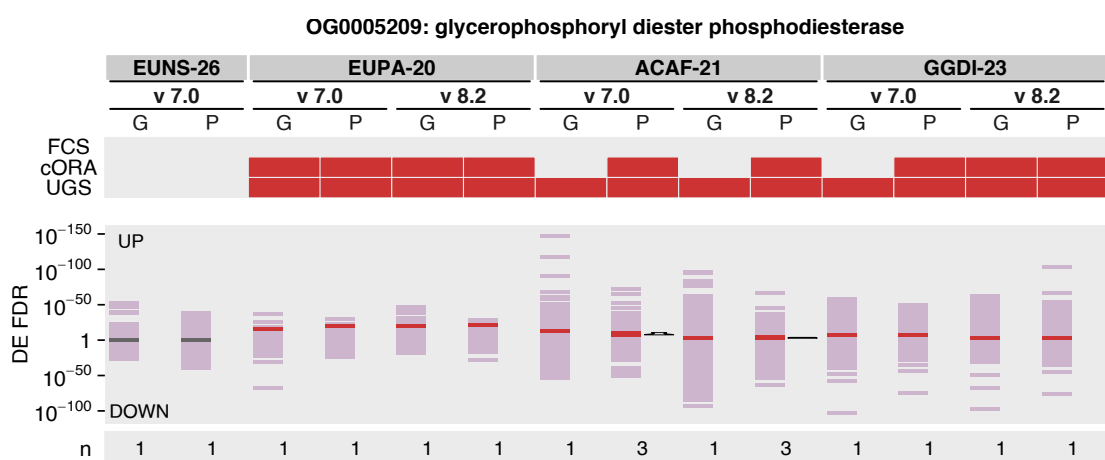
**Figure B.10. Summary of the enrichment analyses results at pH 4.7 for bacterial-type ribosomal protein L13.** For a detailed description of the figure display, see Figure B.1.



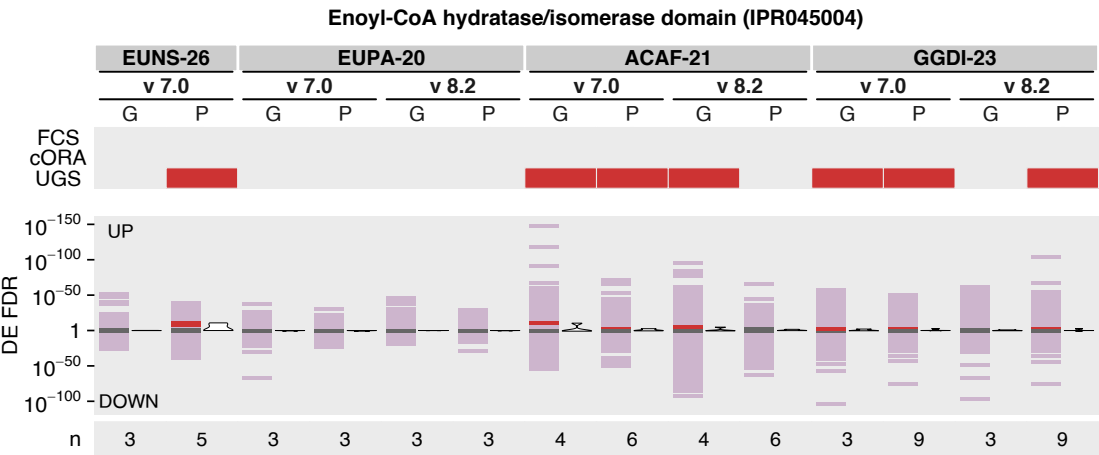
**Figure B.11.** Summary of the enrichment analyses results at pH 4.7 for orthogroup **OG0006704**, encoding a copper-dependent monooxygenase. For a detailed description of the figure display, see Figure B.1.



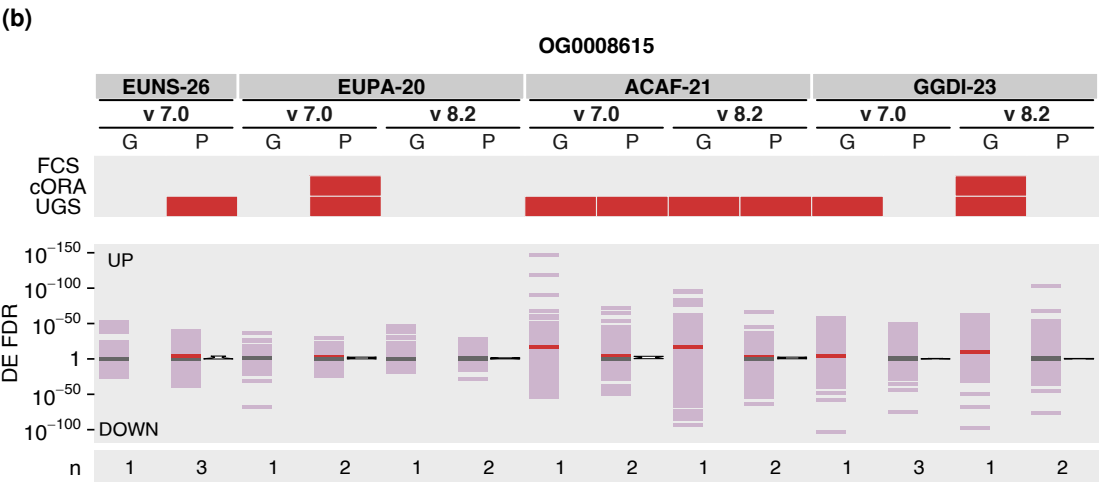
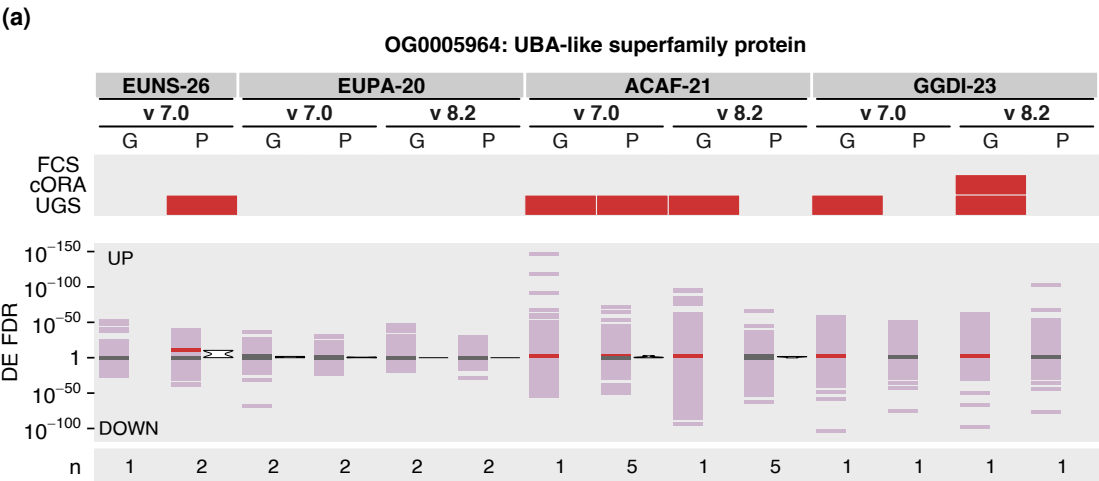
**Figure B.12.** Summary of the enrichment analyses results at pH 4.7 for orthogroup **OG0004427**, encoding a FUN14 domain-containing protein. For a detailed description of the figure display, see Figure B.1.

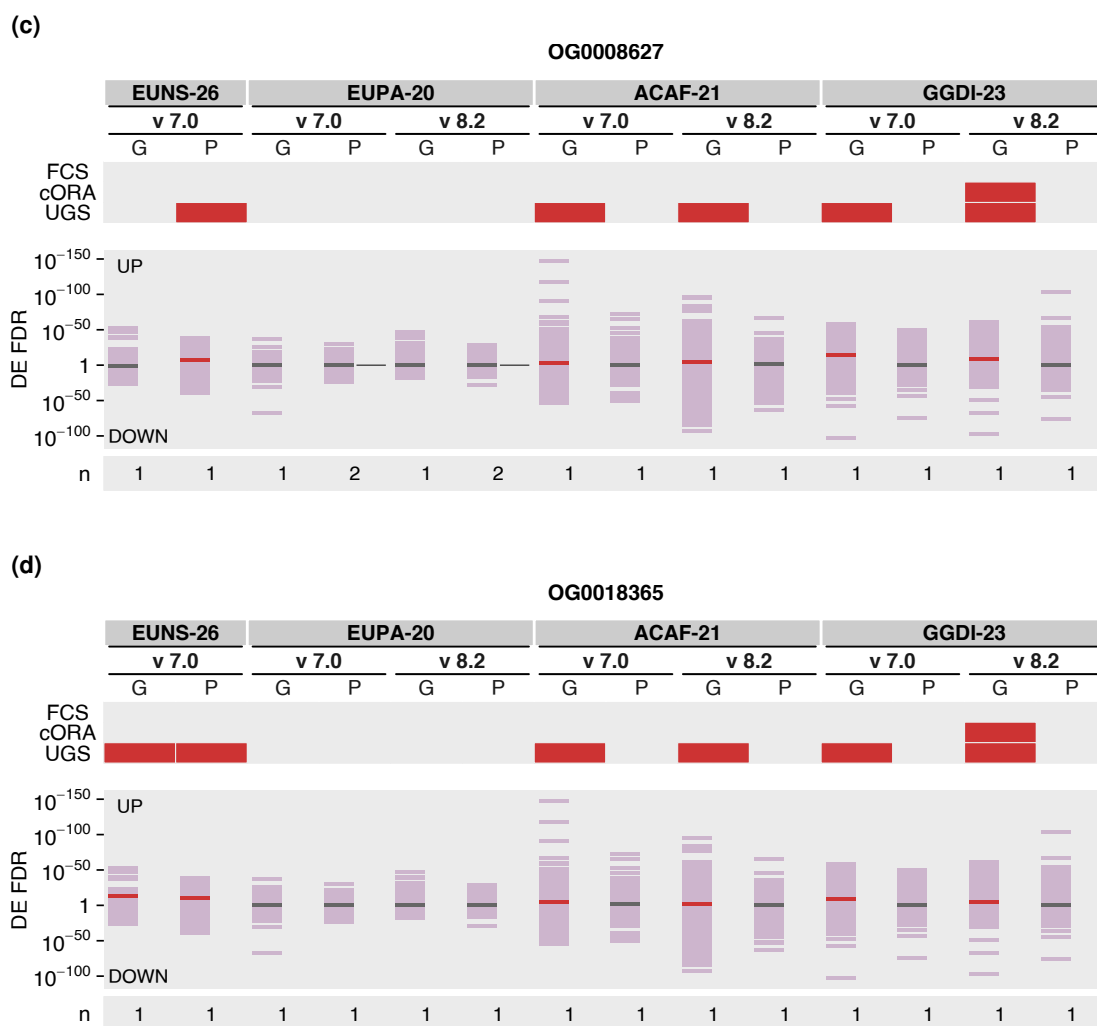


**Figure B.13.** Summary of the enrichment analyses results at pH 4.7 for orthogroup **OG0005209**, encoding a GDPD. For a detailed description of the figure display, see Figure B.1.



**Figure B.14. Summary of the enrichment analyses results at pH 4.7 for enoyl-CoA hydratase/isomerase domain-containing proteins.** For a detailed description of the figure display, see Figure B.1.





**Figure B.15. Summary of the enrichment analyses results at pH 4.7 for poorly annotated orthogroups OG0005964, OG0008615, OG0008627, OG0018365.** For a detailed description of the figure display, see Figure B.1.

