MASTERARBEIT | MASTER'S THESIS

Titel | Title

Investigating the evolutionary origin of animal cell differentiation by establishing sn-multiomics in unicellular holozoans

verfasst von | submitted by

Antonia Grausgruber BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 831
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Zoologie
Betreut von | Supervisor:	Assoz. Prof. Dr. Oleg Simakov
Mitbetreut von | Co-Supervisor:	Roger Revilla Domingo MSc PhD

Acknowledgements

Working on this master project opened up fantastic opportunities and unique experiences. I could not have undertaken this really exiting as well as challenging journey without many people supporting me along the way.

Firstly, I am amazingly grateful to my group leader Roger Revilla-i-Domingo for being a never-ending source of new ideas and for supporting me in a multitude of different contexts. He showed me incredible trust by handing me over both, exceptionally great responsibility as well as freedom in the implementation of diverse projects. I am equally thankful to Carolina Atria, the bioinformatical superpower of our research group and best companion in any imaginable working situation. Her endless euphoria as well as her way of thinking and developing smart problem-solving strategies never failed to amaze me. In this sense I would also like to acknowledge two former members of our group, Elena Paveggio and Sina Kirchem, who guided me in the initial phase of taking over my project. Especially Elena became the most supportive and patient helping hand in and outside the lab that I could imagine.
I would also like to extend my special thanks to Oleg Simakov for taking over the official supervision of my project.

The whole study would not have been possible without our amazing collaborators of the Multicellgenome Lab in Barcelona. Therefore, I would like to express my deepest gratitude to group leader Iñaki Ruiz-Trillo and co-group leader Elena Casacuberta. Not only for providing cells to set up my own *Capsaspora*, *Abeoforma* and *Corallochytrium* culture in Vienna, but also for giving me the possibility to visit their research institute in Barcelona and learn all techniques directly from the real experts in this field. I am particularly grateful to Victoria Shabardina who mainly supervised me during my stay in Barcelona and kept being a fantastic and exceptional inspiring support once I continued my work back in Vienna. Thanks also to Marta Álvarez who as well was part of many discussions about latest observations and possible solutions for some of the many arising problems while working with those non-model research organisms.

The sn-Multiomics sequencing and ATAC tagmentation tests were performed by Thomas Grentzinger of the Next Generation Sequencing Facility at Vienna BioCenter Core Facilities (VBCF), whom I would also like to give special thanks to.

Also, realising my whole study would probably have been way more difficult without Paul Knabl and Sophie Frampton, who shared their experience and knowledge with me and never ran out of new ideas how to improve experimental designs. Especially Paul became a major support and important constant over time. I would also like to thank my Aqualab colleague Georg Brenneis for encouraging me all along the way, not only by enriching my view by an external scientific opinion but also by keeping up my positive mood throughout many fail attempts.

Lastly, I would like to thank my parents, my two sisters Franziska and Johanna, my close friends Patrick-Amir, Augustina, Judith, David and Constantin and all other colleagues and friends for their constant support and understanding. Thank you for helping me to put things into relation and keep also a non-scientific view.

Abstract

Two main features characterizing metazoans are multicellularity and the ability to form specified cell types by gene expression-driven cell differentiation. In order to understand how these crucial traits happened to emerge at the evolutionary origin of animals, it is inevitable to look beyond the borders of the metazoan phylum and dive into their closest relatives, the group of unicellular holozoans. In fact, many of those single-celled organisms show complex life cycles covering a multitude of cell morphologies and functions which indicates that already unicellular sisters of animals are capable of some kind of cell differentiation. Also, respective underlying gene regulatory modules appear to be surprisingly elaborated, involving a quantity of gene orthologs and gene expression controlling mechanisms that previously have been suspected to be metazoan-specific. These discoveries support the idea that animal cell differentiation is based on gene regulatory modules that have already been present in their unicellular ancestor and that got recycled, expanded and refined in the context of animal evolution. Thus, our long-term aim is to compare the gene regulatory changes taking place along the transition between unicellular cell states and those happening along cell differentiation trajectory in basally branching animals. One tool which is able to resolve changes in transcriptomic and epigenetic constitutions on the level of individual cell nuclei is single-nucleus Multiomics. The main goal of my study is to optimize nuclei extraction protocols for the two unicellular species *Abeoforma whisleri* and *Capsaspora owczarzaki* to be able to subsequently perform sn-Multiomics sequencing on those species. For the very first time, sn-Multiomics data is obtained on unicellular holozoans which may be a major step to successively complete our picture on the evolution of animal cell differentiation.

Zusammenfassung

Zwei wesentliche Merkmale, die Metazoa charakterisieren, sind Vielzelligkeit und die Fähigkeit, durch Geneexpressions-gesteuerte Zelldifferenzierung spezialisierte Zelltypen auszubilden. Um zu verstehen, wie diese fundamentalen Eigenschaften am evolutionären Ursprung der Tiere entstanden sind, ist es unerlässlich über die Grenzen des Phylums der Metazoa hinauszublicken und in die nächstverwandte Organismengruppen einzutauchen, die einzelligen Holozoa. Tatsächlich weisen einige dieser unizellulären Organismen einen komplexen Lebenszyklus auf, der eine Vielzahl an verschiedenen Zell-Morphologien und Funktionen umfasst. Das weist darauf hin, dass auch schon die einzelligen Schwesterngruppen der Tiere in bestimmter Form zur Zelldifferenzierung fähig sind. Auch entsprechende zugrundeliegende genregulatorische Module scheinen überraschend elaboriert zu sein und einige Gen-Orthologe sowie Genexpression-kontrollierende Mechanismen zu inkludieren, die zuvor als Metazoa-spezifisch betrachtet wurden. Diese Beobachtungen untermauern die Idee, dass tierische Zelldifferenzierung auf genregulatorischen Modulen beruht, die bereits in einzelligen Vorfahren vorhanden waren und im Kontext der Evolution der Tiere wiederverwendet, erweitert und verfeinert wurden. Unser langfristiges Ziel ist es dementsprechend, die genregulatorischen Veränderungen entlang der Transition zwischen verschiedenen Einzeller-Zellstadien mit jenen zu vergleichen, die entlang der Differenzierungsbahn in basal gestellten Tieren stattfinden. Ein Werkzeug, das erlaubt Änderungen im transkriptionellen und epigenetischen Zustand auf der Ebene von einzelnen Zellkernen aufzulösen, ist „single-nucleus Multiomics“. Dementsprechend ist eine wesentliche Aufgabe meiner Arbeit die Protokolloptimierung für Nuklei-Extraktion aus den zwei Einzeller-Arten *Abeoforma whisleri* und *Capsaspora owczarzaki*, um anschließend sn-Multiomics-Sequenzierung durchführen zu können. Für das aller erste Mal werden sn-Multiomics Daten für einzellige Holozoa-Vertreter generiert, die ein wesentliche Schritt darin sein können, sukzessive unser Bild zur Evolution der tierischen Zelldifferenzierung zu vervollständigen.

Table of contents

1. INTRODUCTION – a conceptional framework

More than 600 million years ago multicellular animals (metazoa) evolved from a single-celled ancestor, setting one of the most important milestones in the history of life (King, 2004; Ligrone, 2019b). Growing a multicellular body is accompanied by an incredible gain of developmental possibilities, but at the same time it encounters certain challenges of coordinating such a system. While in a unicellular organism a single cell has to fulfil all biological functions necessary for life und reproduction, cells within a multicellular construct can develop functional specialisations allowing division of labour among different spatial units (Brunet & King, 2017; Ligrone, 2019a).

This is what modern developmental biology describes as “spatiotemporal cell differentiation”. During embryogenesis a complex, multicellular animal body develops out of a hollow sphere of undifferentiated cells (Sebé-Pedrós et al., 2017; Ruiz-Trillo & de Mendoza, 2020). These pluripotent cells divide, form new cells and give rise to all kinds of specialised unipotent cells. Approaching cell differentiation in animals is therefore the result of gradual reduction of potency whereby the potential that remains at the end of this process defines a certain cell identity, called “cell type” (Marioni & Arendt, 2017). While all cell types still rely on the same genomic information content, huge parts of the genome are inactivated in fully differentiated cells. Their identity is fundamentally determined by the distinct set of active transcription factors and effector genes, establishing a cell type-specific gene expression profile (Arendt et al., 2016; Paps, 2018). This profile is directly associated with a particular cell morphology and defines functional and biological capacities (Sebé-Pedrós et al., 2017; Ligrone, 2019a). To ensure that the formation of a certain animal body plan is a highly reproducible event, the process of cell differentiation needs to be strictly controlled and orchestrated on a temporal as well as spatial axis. This is done by the cooperation of hierarchical gene regulatory networks and epigenetic mechanisms determining which distinct gene expression profile a certain cell develops in the course of its differentiation process (Davidson, 2006).

Accordingly, animals’ multicellularity seems to be tightly connected to their ability to form an unparalleled diversity of specialised cell types (Rokas, 2008). Together these two traits may have formed the most important and powerful prerequisites for the evolutionary success of metazoans. They allowed for the emergence of seemingly endless variety of body plans and developmental modes that we know from fossil records and recent taxa (Paps, 2018). The great rise of the animal kingdom was not only influenced by past environmental conditions, but it also shaped the given environment in return as animals were able to construct and inhabit other ecological niches than their unicellular ancestors. This highlights even more the unique role that the emergence of animals played for the history of our whole planet (Ligrone, 2019a & 2019b).

On the other hand, building a multicellular animal also brings up new challenges. While a single-celled organism is comparatively flexible in reacting to environmental cues, cells within a multicellular construct need to be able to communicate and interact with each other. They not only have to reacting to external stimuli but also to intra-structural signals (Ligrone, 2019a; Ros-Rocher et al., 2021). Thus, they need modules enabling mechanisms like cell adhesion, signal transduction and reception, resource transfer or mediation of positional information (Grosberg & Strathmann, 2007; Knoll, 2011).

Given this introduction on what characterises animals, it might appear as if there is a multitude of aspects clearly distinguishing animals from their unicellular relatives. However, in order to truly capture what makes up an animal we have to understand where they came from and how the transition from a single-celled ancestor to the first multicellular metazoan happened (Mikhailov et al., 2009). Despite the fundamental importance of this question, centuries of research have not yet been enough to give a sufficient and satisfactory answer. Especially findings of recent decades showed that what might seemed obvious at first, is surprisingly complex and unexpected once looking at it in more detail (Sebé-Pedrós et al., 2011; Sebé-Pedrós et al., 2016). But what indeed became clear is that investigating unicellular relatives of animals will be an essential key to our in-depth understanding of metazoan evolution. Preadaptation happening in single-celled ancestors gradually set the stage to allow the emergence of animals. Recent unicellular relatives of animals can help us to reconstruct the hypothetical last common ancestor of all metazoans as well as the mechanisms underlying the transition towards animals (Ocãna-Pallarès et al., 2022; Ros-Rocher et al., 2021; Ruiz-Trillo et al., 2023).

1.1. The phylogenetic relations of Holozoa

Having a robust and reliable representation of the phylogenetic relationships embedding the origin of animals is one of the most essential requirements to be able to interpret new findings in an evolutionary context (Sebé-Pedrós et al., 2017; Ros-Rocher et al., 2021).

One of the longest known and studied groups within unicellular opisthokonta are choanoflagellates, comprising about 400 described species. They include comparatively well-investigated representatives like *Monosiga brevicollis* and *Salpingoeca rosetta* (Hoffmeyer & Burkhardt, 2016; Figure 1). Already in the 19th century, morphological studies described them as the direct sister group of metazoans (Cavalier-Smith, 2017). This could later on be confirmed by several molecular analyses, supporting it as notably robust branching point (Medina et al., 2003; King, 2004; Carr et al., 2008; Ruiz-Trillo et al., 2008; Shalchian-Tabrizi et al., 2008). However, other parts of the phylogenetic tree surrounding metazoa were widely unclear for a very long time. Only intensive research of recent decades in describing new

unicellular organisms and performing large-scale genome comparisons, could reconstruct a much more comprehensive picture (Ruiz-Trillo et al., 2004 & 2007; Torruella et al., 2011 & 2015; Glocking et al., 2013; Hehenberger et al., 2017). Latest findings distinguish three additional classes of single-celled species, apart from the choanoflagellates, which form the closest relatives of animals: namely these are filasterea, ichthyosporea and corallochytrea (Grau-Bové et al., 2017; Ros-Rocher et al., 2021). The group of “holozoa” covers metazoans as well as these four groups of closest related unicellular organisms (see Figure 1) (Lang et al., 2002; Ros-Rocher et al.; 2021).

Filasterea are suspected to represent the second closely related unicellular group to animals (Paps & Holland, 2018; Schultz et al., 2023). Initially they only included the genus *Capsaspora* with the single species *Capsaspora owczarzaki* (Hertel et al., 2002; Ruiz-Trillo et al., 2004). Until now four other species of filastereans have been described (Tikhonenkov et al., 2020; Kożyczkowska et al., 2021), supporting it as a distinct and robust sister group to animals and choanoflagellates.

The exact phylogenetic resolution of ichthyosporea and corallochytrea, however, is less clear. Current perceptions are grouping them together as “teretosporea” (Torruella et al., 2014; Tikhonenkov et al., 2020), forming a joint sister group to filasterea, choanoflagellates and metazoa. Today, 25 different species are described within ichthyosporea, while only two corallochytrea species are known so far (Hehenberger et al., 2017; Ruiz-Trillo et al., 2023).

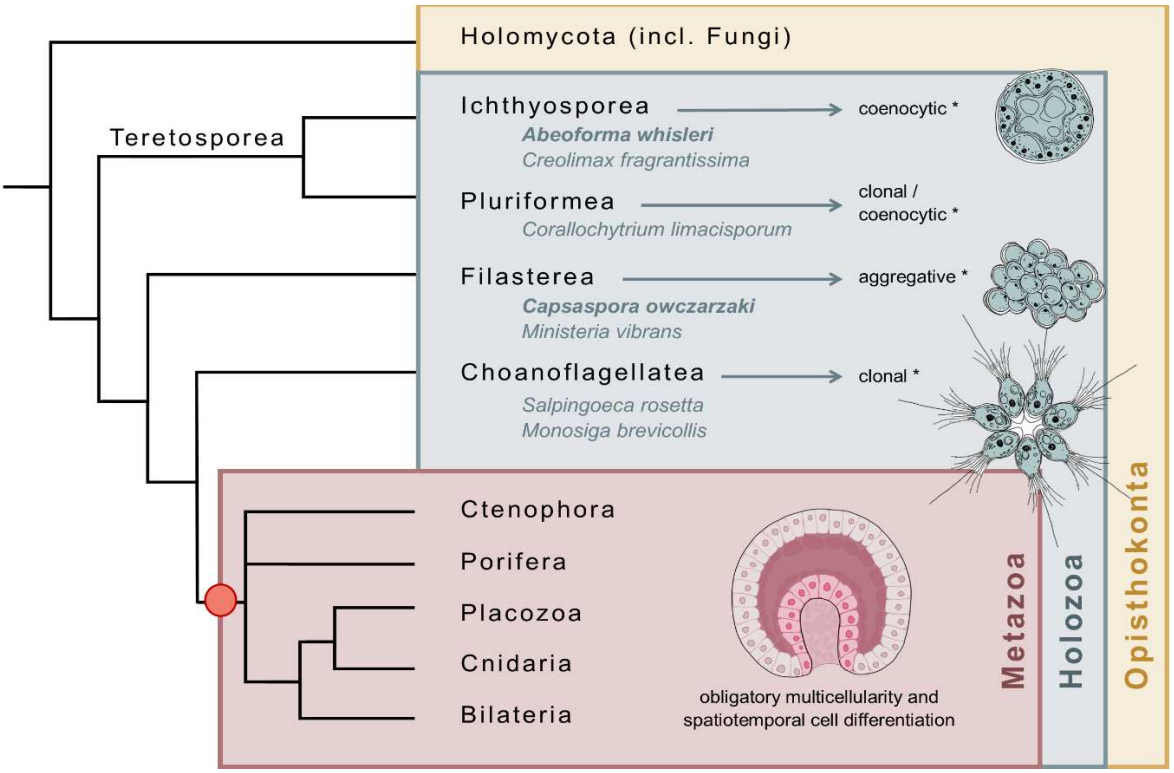

Figure 1) Overview of the opisthokonta phylogeny: examples of described representatives are given for each of the four unicellular holozoan lineages. (*) mark forms of temporal multicellular or multinucleated states which could be observed in the respective groups. The red circle marks the origin of animal multicellularity and cell differentiation (adapted from Ruiz-Trillo et al., 2023).

Even though research efforts of the last 15 years could considerably improve the picture that we can draw on the phylogenetic framework embedding animals, there is strong evidence that a great part of the diversity of unicellular holozoans is still hidden. Studies analysing environmental DNA from freshwater and marine habitats discovered that a big percentage of collected DNA could be rooted within the opisthokonta or even unicellular holozoans but could not be assigned to any known species (Arroyo et al., 2018; Arroyo et al., 2020). They suggest a great diversity of, potentially also higher-ranked, unicellular holozoan groups which is not yet described (Ruiz-Trillo et al. 2023).
We therefore must be aware that the current phylogenetic tree of holozoa is the best possible reconstruction we can get based on available data, but it might be put to the test by the discovery of new species at any time.

1.2. The surprising potential of unicellular relatives of animals

For the longest time, research investigating the origin of animals was strongly focussed on basally branching animals like porifera or ctenophores and choanoflagellates as a direct sister group (Brooke & Holland, 2003; King et al., 2008). Based on these comparisons many features that are known to be involved in animal development were traditionally classified as “metazoan-specific” (Sebé-Pedrós et al., 2016a). However, once scientists started to look more closely into other groups of unicellular holozoans they were facing a surprisingly high potential for complex life that is present already in these single-celled animal relatives. This applies to a biological and ecological level as well as to the level of underlying genomic and gene expression regulatory capabilities (Ruiz-Trillo et al., 2006; Mikhailov et al., 2009; Sebé-Pedrós et al., 2016a & 2017). These findings strongly reshaped our recent thinking about the unicellular ancestry of animals and conditions required to allow the transition from unicellular to multicellular life.

1.2.1. Biological potential

Probably the very first thing that stands out when looking at unicellular holozoans, is their remarkable variability in terms of realised cell morphologies and life forms. Many groups show complex life cycles, covering a multitude of different cell states that can considerably vary in visual appearance (size and shape), proliferation activity or mobility. Dependent on given environmental cues a single cell can undergo multiple transitions between distinct cell states within its lifetime (Ros-Rocher et al., 2021; Ruiz-Trillo et al., 2023). Even though this is an apparently dynamical process, these transitions appear to happen directional and controlled rather than randomly (Sebé-Pedrós et al., 2017). What is especially interesting is that many species show some form of transient multicellularity in their life cycle. But even these

multicellular states differ in how they are organised and originating (see Figure 1; Brunet & King, 2017).

Choanoflagellates show a rather simple life cycle. Free-swimming, flagellated cells can form multicellular colonies, which can either be floating or attached to the substrate (King, 2005; Hoffmeyer & Burkhardt, 2016). They result from clonal division of one individual cell, meaning that all cells involved in such a colony are of the same genetic origin (Fairclough et al., 2010; Sebé-Pedrós et al., 2017).

In comparison, life cycles found among the known filasterean species are a bit more complex and diverse. Single cells can have an amoeboid to spherical shape, potentially they attach to the substrate using filopodia or flagella or form cystic states (Kożyczkowska et al., 2021; Ros-Rocher et al., 2021). However, what most species have in common is the formation of multicellular aggregates throughout their lifetime (Cavalier-Smith & Chao, 2003; Hehenberger et al., 2017; Tikhonenkov et al., 2020). Such aggregative multicellularity significantly differs from clonal structures found in choanoflagellates since it is not the result of cell division but of genetically distinct cells clustering together and building up a common construct (Brunet & King, 2017; Sebé-Pedrós et al., 2017).

Finally, the group of teretosporea is seemingly exploding in possibilities regarding cell morphologies and life forms. The specific example of *Abeoforma whisleri* (ichthyosporea) is described in more detail in paragraph 2.1. Overall, individual cells can greatly vary in terms of size and shape, the presence of protrusions or flagella and their mobility (Mendoza et al., 2002; Marshall et al., 2008; Marshall & Berbee, 2011; Kożyczkowska et al., 2021). Most importantly, many teretosporea species form large polynucleated cell states: these so-called coenocytes are generated by nuclei division without following cytokinesis. Mature coenocytes can undergo cellularisation and release endospore-like or amoeboid-shaped cells, setting up a new generation of mononucleated cells (Marshall & Berbee, 2011; Ros-Rocher et al., 2021; Ruiz-Trillo et al., 2023).

The attempt of a conceptional interpretation

To summarize, this inherently suggests that already single-celled relatives appear to hold the potential for two of the most important features allowing animal life: multicellularity and the ability to generate differently specialised cells (Ros-Rocher et al., 2021). Still the exact details of these characteristics significantly differ between animals and non-animal holozoans.

Firstly, cell state transitions in single-celled organisms are typically described as “temporal cell differentiation” (Mikhailov et al., 2009). This means that the distinct cell states an individual cell passes through during its lifetime are separated from each other in time – in comparison, spatiotemporal cell differentiation enables multicellular animals to combine multiple, functionally specialised and spatially separated, units within one organism at the same time

(Sebé-Pedrós et al., 2017). In most literature, the distinction between pure temporal and spatiotemporal cell differentiation is referenced as a main trait discriminating animals and their unicellular relatives. Such a scenario would incorporate that spatial cell differentiation evolved after the phylogenetic split of choanoflagellates and metazoans (Mikhailov et al., 2009; Brunet & King, 2017). However, latest research indicates that it might not be possible to draw such a strict black-and-white picture there. Studies on the choanoflagellate *Salpingoeca rosetta* showed that multicellular colonies can include cells that notably differ in appearance and might represent distinguishable cell types (Laundon et al., 2019; Naumann & Burkhardt, 2019). Similar evidence is also given for the filasterean *Capsaspora owczarzaki* (Ros-Rocher et a., 2021). This indicates that some preliminary form of spatial cell differentiation might be already present in unicellular holozoans.

In any respect it is important to highlight that the principle of spatiotemporal cell differentiation itself can only be realised under multicellular conditions. Therefore, it might not be meaningful to talk about different manifestations of cell differentiation if this is not interpreted in the context of the presence or absence of multicellularity. In return, it is known that multicellularity itself emerged multiple times in the scope of evolution and transient cluster formation is a common phenomenon in diverse single-celled opisthokonta. Multicellular constructs have evolved from independent origins, developing different particularities and potentials (Knoll, 2011; Grosberg & Strathmann, 2007; Sebé-Pedrós et al. 2017; Ligrone, 2019a). This further complicates the attempt to interpret the occurrence of multicellular constructs and related properties (as the potential to show spatial cell differentiation) in unicellular relatives of animals.

Secondly, it appears that we are comparing processes with fairly different consequences when we speak of cell differentiation in animals and their unicellular relatives. As described at the very beginning, cell differentiation in animals is tightly connected to the reduction of cellular potency (Sebé-Pedrós et al., 2017). In most known cases this process is unidirectional and finalizes in terminal cell differentiation. The resulting unipotent cell is seemingly “trapped” in its highly specialised function (Arendt et al., 2016; Marioni & Arendt, 2017). However, cell differentiation in unicellular organisms seems to be the entry into a very dynamic process where cells do not seem to terminate in a single function. Instead, they are able to enter certain cell states and again “exit” them whenever environmental stimuli trigger them to do so (Ros-Rocher et al., 2016). To put it very metaphorically, cell specialisation in unicellular organisms seems less lapidary and final than in animals. Therefore, the concept of cell differentiation that researchers developed based on studying multicellular animals might not be completely sufficient to describe cell state transitions in unicellular holozoans.

Accordingly, comparing cell differentiation and multicellularity of animals with features of their single-celled relatives on a purely morphological and conceptional level is clearly not sufficient

to provide satisfactory answers. We thus have to look way more in-depth and investigate underlying genomic groundworks of these processes (Sebé-Pedrós et al., 2016a).

1.2.2 Genomic and gene regulatory potential

1.2.2.1. Looking at genomes

To address the challenge raised above, research of the last years made great efforts to explore the genomic capacities of premetazoan organisms. Within the last six years, 11 new genomes of different unicellular species have been sequenced; making up a total of 15 recently available whole-genome datasets covering representatives of all four non-animal holozoan clades (Sebé-Pedrós et al., 2017; Ruiz-Trillo et al., 2023). This opened up unprecedented resources on a broad-taxon level to gain new insights into the potential genomic equipment of the hypothetical single-celled ancestor of animals.

In fact, analyses showed that unicellular holozoan genomes carry an unexpectedly elaborated repertoire of genes which are known to be involved in animal cell differentiation and multicellularity (Sebé-Pedrós et al., 2011; Suga et al., 2013). The incorporation of more and more species in these comparisons indeed led to a striking reduction in the number of genes considered to be animal-specific (Ruiz-Trillo et al., 2023). As already mentioned, early research strongly focused on choanoflagellates. However, they show substantial secondary losses in genomic content, leading to many genes being miscategorized as metazoan-specific in the past (Richter & King, 2013; O'Malley et al., 2016). Based on these misplaced expectations, results obtained on a broader range of unicellular species were even more surprising.

Firstly, investigated genomes covered many orthologs of genes that are required for, or involved in cell adhesion and cell interaction (Ros-Rocher et al., 2021). These are crucial functions to enable animal multicellularity and development. Well-investigated examples are cadherins, integrins (secondarily lost in choanoflagellates) or genes encoding for proteins that are associated with extracellular matrix composition. A more detailed overview of found genes / gene families, including respective references, is given in Table 1.

Secondly, they found quite a large number of well-known animal transcription factors (TFs) to be encoded in these premetazoan genomes (Sebé-Pedrós et al., 2016a). Through binding to characteristic DNA-motifs TFs can regulate (activate or repress) the transcription of defined sets of target genes in a spatiotemporal and cell type-specific manner (Spitz & Furlong, 2012). As they are main conductors of downstream gene regulatory networks but also involved in cell signalling pathways, they represent a very important interconnecting point between different processes involved in cell differentiation (Weidemüller et al., 2021). Besides *Brachyury,* which is crucial for animal gastrulation, also several TFs related to stem cell differentiation and cell proliferation, like *p53* or the proto-oncogene *Myc,* were found to be present in unicellular

holozoans (see Table 1). Interestingly, more in-depth experiments using Chip-qPCR provided evidence that also the *Brachyury* target networks seem to be partly shared with metazoans. Indications for conserved targets are also given for *Myc* (Sebé-Pedrós et al., 2016a).

Table 1) Summary of genomic features found in premetazoan genomes. This list should give an overview (without guarantee of completeness) of the most important examples. Not all genes / gene families listed are found in all investigated unicellular holozoan lineages, however their distribution strongly suggests a premetazoan origin.

		Function in metazoa	References
genes involved in cell adhesion / interaction	cadherins	cell-cell adhesion	King et al., 2003; King et al., 2008; Abedin & King, 2008; Nichols et al., 2012; Levin et al., 2014 Richter et al., 2018; Tikhonenkov et al., 2020
	c-type lectins	cell-cell adhesion	
	integrins	cell-cell adhesion, cell-ECM adhesion	Sebé-Pedrós et al., 2010; Sebé-Pedrós et al., 2013c; Hehenberger et al., 2017; Richter et al., 2018; Parra-Acero et al., 2020; Tikhonenkov et al., 2020
	laminin, collagen, fibronectin	extracellular matrix or basal lamina composition	Sebé-Pedrós et al., 2013c; Suga et al., 2013; Shabardina et al., 2020; Tikhonenkov et al., 2020
transcription factors	*Brachyury*	animal gastrulation, mesoderm differentiation	Sebé-Pedrós et al., 2011; Sebé-Pedrós et al., 2013a; Ruiz-Trillo et al., 2023
	+ downstream targets	transcription regulation, cell polarity, metabolism, phagocytosis	
	p53	stem cell differentiation, cancer-related	Sebé-Pedrós et al. 2011; Young et al., 2011; Sebé-Pedrós et al., 2013a; Sebé-Pedrós et al., 2016a
	Myc	stem cell differentiation, proliferation, proto-oncogene	
	+ downstream targets	cell motility, proliferation	
	NF- κB	immunity response, proliferation, apoptosis	
	RUNX	hematopoietic stem cell differentiation	
key developmental signalling pathways	receptor tyrosine kinase RTKs	cell growth, proliferation, cancer-related	King et al., 2003; Manning et al., 2008; Schultheiss et al., 2014; Suga et al., 2014
	Hippo pathway	cell proliferation, cell differentiation organ-size control, regeneration, apoptosis	Sebé-Pedrós et al., 2017; Paps & Holland, 2018; Phillips et al., 2022
	G-protein	interconnects receptor and second messenger systems	de Mendoza et al., 2014; Paps, 2018
	Delta / *Notch*	embryonic development	Ros-Rocher et al., 2021

Lastly, studies revealed that also other components of important cellular signalling pathways are present in premetazoan genomes. These mechanisms allow for signal transduction between cells as well as the reception of, and reaction to external stimuli in a multicellular animal (Ros-Rocher et al., 2021). Probably the best studied example is the receptor tyrosine kinase pathway, which is represented by a large repertoire of components in all four clades of unicellular holozoans (see Table 1). In the case of the Hippo signalling pathway, most intracellular components of the pathway are already present in single-celled holozoans, whereas certain upstream receptor ligands known from animals could not be found. This, however, seems to be a common pattern also among other examples of present signalling pathways (see Table 1; Sebé-Pedrós et al., 2017; Paps & Holland, 2018).

1.2.2.2. Looking at transcriptomes

An arising subsequent question regarding the genetic foundation of cell differentiation is whether these genes or pathways are employed in a cell state-dependent manner. In order to allow for the establishment of cell identities based on differential gene expression it is necessary that these modules operate in a highly particular pattern rather than generically. Accordingly, the next step is to explore cell state specific transcriptomes and how these change throughout the life cycle (Sebé-Pedrós et al., 2017).

Respective experiments carried out on diverse unicellular species indeed showed that distinct cell states significantly differ in their transcriptional profile. For instance, Sebé-Pedrós et al. (2013c) found that cell states in *Capsaspora owczarzaki* are characterised by hundreds of differentially expressed genes. This involves integrin adhesomes and genes encoding for ECM proteins which are specifically upregulated in the aggregative state, while the filopodial state shows increased expression of genes related to the cytoskeleton (Sebé-Pedrós et al., 2017). In fact, subsequent experiments could show that integrins are substantial for aggregate formation as inhibition of integrins represses aggregation (Parra-Acero et al., 2020). Also in the choanoflagellate *Salpingoeca rosetta* the gene expression profile of free-swimming individual cells significantly differs from the one of colonial cells (Fairclough et al., 2013). A well-studied example is the C-type lectin *Rosetteless* which is upregulated in colonial cells and essential to enable colony formation (Levin et al., 2014).
Furthermore, Sebé-Pedrós et al. (2013c) showed that not only certain sets of genes show differential expression across cell states, but also signalling pathways are employed in a particular pattern. For example, the filopodial state of *Capsaspora owczarzaki* is characterised by increased activity of the tyrosine kinase signalling pathway, while the hippo pathway seems to be upregulated in the aggregative cell state (Sebé-Pedrós et al., 2017).

1.2.2.3 Looking at mechanisms

Up to this point, studies already showed that unicellular holozoans do have a broad genomic potential involving many differentially activated transcription factors and signalling pathways important for gene regulation. However, this does not yet give a comprehensive picture on the mechanistic repertoire of unicellular holozoans to control gene regulation.

The probably most inclusive study on gene regulatory capacities of unicellular holozoans was conducted by Sebé-Pedrós et al. (2016a). Results showed that *Capsaspora* owns a surprisingly broad set of modules for gene regulation involved in cell state transition. These include dynamic alternative splicing events and cis-regulatory sites, differential lincRNA (long intergenic non-coding RNAs) expression as well as dynamic changes of chromatin states regulated by distinct histone modifications (Sebé-Pedrós et al., 2013b; Sebé-Pedrós et al., 2016a). Certain cell states are associated with a specific phosphorylation pattern leading to downstream signalling pathways being state-specifically activated (Sebé-Pedrós et al., 2016b). Overlapping results could also be obtained for *Creolimax fragrantissima,* showing that differential employment of alternative splicing and lincRNAs are important mechanisms happening during cell state transitions (de Mendoza et al., 2015). Finally, recent studies on ichthyosporea even discovered a whole set of animal-type microRNA genes and genes homologous to the animal microRNAs biogenesis machinery (including *Drosha* and *Pasha;* Grimson et al., 2008; Bråte et al., 2018).

1.3. Rewriting our idea of the unicellular ancestor of animals

Altogether studies of the last decades revealed unexpected complexity of premetazoan species and thus truly rewrote the common perception on the evolutionary emergence of animals.

Without giving any definite answer to this question, one very important conclusion can be made that focuses on remodelling events in the manifestation of cell differentiation and multicellularity. We can assume that the unicellular ancestor of all animals already had the potential for cell differentiation based on complex mechanisms controlling differential gene expression. This applies for sure to a temporal context, put potentially also to a spatial one (Ruiz-Trillo et al., 2023). Additionally, the presence of certain cell signalling pathways enabled this organism to transit between life stages triggered by distinct environmental stimuli (Ros-Rocher et al., 2021). Our unicellular ancestor presumably had a complex life cycle, including at least one multicellular stage as well as a trophic state among several plastic morphologies. Accordingly, the potential for cell differentiation and for the formation of multicellular constructs did evolve before the transition to obligatory multicellularity in animals happened. Underlying genomic and gene regulatory features were reused, co-opted, extended, and refined in the

context of animal evolution (Sebé-Pedrós et al., 2017; Ros-Rocher et al., 2021; Ruiz-Trillo et al., 2023). These changes in the manifestation of already existing tools and additional innovations enabled the first animal to develop spatiotemporal cell differentiation in an obligatorily multicellular context. Also, rewiring of existing networks and the evolution of new signalling pathway allowed the transition from a mainly environment-dependent organism to an autonomous construct which is able to transmit and receive body-intern signals and mediate positional information leading to body axis formation (King et al., 2008; Suga et al., 2013; Sebé-Pedrós & de Mendoza, 2015; Ros-Rocher et al., 2021).

This modern view on the origin of animals completely inverts the presumed order of evolutionary events, compared to the long believed traditional view (Levit et al., 2021). It was strongly shaped by Ernst Haeckel's "Gastraea theory" and suggested that the first animals originated from a simple multicellular construct consisting of uniform choanoflagellate-like cells. When cells started invaginating at one side of this construct, forming an outer and inner cell layer, the first sights of cellular differentiation emerged (Mikhailov et al., 2009; Levit et al., 2021). This inherently states that multicellularity predated the emergence of cell differentiation at the root of the animal phylum (Paps, 2018). However, our current perception on an already highly developed unicellular ancestor turns this traditional view upside down.

1.4. Describing core mechanisms involved in metazoan emergence

Based on this updated reconstruction of the last common ancestor of animals, one can formulate four main core mechanisms that most likely played an important role in the transition to multicellular animals.

1.4.1 Innovation of new features

As far as our recent state of knowledge can tell, several new features evolved at the evolutionary origin of animals (Paps, 2018; Paps & Holland, 2018). This applies to gene families as well as whole signalling pathways or regulatory mechanisms.
Good candidates for being metazoan-specific are for example the transcription factor families of ETSs, SMADs, IRFs or nuclear receptor families (Sebé-Pedrós et al., 2011; Sebé-Pedrós et al., 2013a; Sebé-Pedrós & de Mendoza, 2015). Signalling pathways which are crucial in animal development but have not been reported outside of metazoans are Wnt / ß-catenin, TGF-ß or the JAK-STAT pathway (King et al., 2008; Suga et al., 2013). The acquisition of these signalling pathways could have been crucial for allowing body axis formation and mediation of positional information in a multicellular animal (Ros-Rocher et al., 2021). Most likely also new regulatory mechanisms arose at the origin of animals, which got interconnected in already existing networks. This might be the case for distal enhancers which implement an additional layer of regulatory possibilities (Sebé-Pedrós et al. 2017).

Also, the presence of repressive histone modification marks might be an innovation of animals. Having additional repressive controllers could have been one aspect allowing animals to develop their very profound form of terminal cell differentiation, based on progressive restriction of genomic capacities and long-term silencing of genome regions (Sebé-Pedrós et al., 2016a; Sebé-Pedrós et al., 2017).

1.4.2. Co-option of ancestral genes

"Co-option" describes the phenomenon of ancestral features taking over a new function during their evolution (True & Carroll, 2002). In many cases gene co-option is associated with gene duplication events: two emerging daughter cells can diverge by accumulation of sequence changes, leading to the acquisition of different functions in the paralogous genes (True & Carroll, 2002; Holland et al., 2017). However, co-option can also happen on a gene regulatory or mechanistic level (Reike et al., 2013; Sebé-Pedrós et al., 2017).
One example are cadherins which may have developed their function in cell-cell adhesion only within the animal lineage. Experiments carried out on choanoflagellates indicated that the premetazoan role of cadherins was presumably not associated with cell adhesion but rather with prey capture in bacterivorous feeding (Abedin & King, 2008; Sebé-Pedrós et al., 2017).

1.4.3. Loss of features

Just as new traits emerge in the course of evolution, others also get lost at the same time. In fact, recent studies highlighted the great importance of gene losses for evolutionary developments (O'Malley et al., 2016). It happens frequently on diverse levels and in several stem lineages. Certain clades can develop higher tendencies for gene loss, for example due to drastic changes in their lifestyle (as it might be the case for choanoflagellates). Taking into account that secondary loss of features can substantially hamper our ability to reconstruct the evolution of certain traits, broad-taxon sampling in comparative studies appears even more important (Richter & King, 2013; Sebé-Pedrós et al., 2010; O'Malley et al., 2016).

1.4.4. Reuse and expansion of existing features

Another commonly observable principle is the reuse and expansion of already existing systems. On the one hand, this can happen within one gene family, due to gene duplication events (Sebé-Pedrós et al., 2017). On the other hand, also regulatory modules can get extended in their function by the addition of new upstream regulatory units or elaboration of downstream target networks. One example is the Hippo pathway which potentially got expanded in metazoans by the addition of new animal-specific upstream receptors (Sebé-Pedrós et al., 2012; Ros-Rocher et al., 2021).

1.5. The importance & opportunities of single-nucleus Multiomics sequencing

If we want to find out what remodelling events exactly took place at the transition from unicellular to multicellular life in animals, we need to compare the genomic and regulatory changes happening during cell differentiation in recent species. For this purpose, it becomes necessary to look at transcriptomes and epigenomes of single cells.
Up to this point, all investigations carried out on unicellular holozoans were done on a whole population level of cells (bulk). Such approaches automatically set fundamental limitations restricting the possibilities for designing and setting up experiments. Investigating distinct cell states requires that the life cycle of used species can be highly synchronized in cell culture and thus allows harvesting of only one particular cell state at a time (Sebé-Pedrós et al., 2017). Also, cell states can only be differentiated up to a level where clear morphological discriminations can be made (Kharchenko, 2021). The ability to do so is considerable hampered for unicellular organisms with complex, hardly synchronizable life cycles involving a lot of intermediate states that cannot be clearly distinguished. All these issues encountered by cell population-based approaches can be overcome by using single-cell- or single-nucleus-based technology.

The development of single-cell RNA-sequencing truly revolutionized the field of evolutionary and developmental biology as well as clinical research in the last decades (Marioni & Arendt, 2017; Huo et al., 2019; Kharchenko, 2021). By combining cell-specific barcoding of all transcripts with Next-Generation Sequencing methods it is possible to trace each transcript in a sample back to its cellular origin and reconstruct unique gene expression profiles of single cells (Kharchenko, 2021). This not only enables inference about cell differentiation trajectories between cell types, but it can also resolve intermediate cell states and possible sub-populations of cells which cannot be distinguished by any traditional method (Tirosh et al., 2016; Villani et al., 2017).

Constant development of the single-cell research field brought up several extensions of this approach in the last years (Belhocine et al., 2021). One of them is single-nucleus Multiomics sequencing (10x Genomics: Next GEM Single Cell Multiome ATAC + Gene Expression). This method combines single-nucleus ATAC-sequencing (assessing the epigenetic state by reflecting chromatin accessibility) with single-nucleus RNA-sequencing (assessing the transcriptional state by profiling gene expression). It finds a way to apply these two tools simultaneously on individual nuclei. Therefore, it does not rely on the bioinformatical integration of two separate datasets for sc-RNA- and sc-ATAC-seq. This makes sn-Multiomics an extraordinary powerful instrument to study gene expression and the chromatin landscape of individual cells at the same time (Belhocine et al., 2021; Guo et al., 2023). Detailed steps of the sn-Multiomics workflow are depicted in Figure 2.

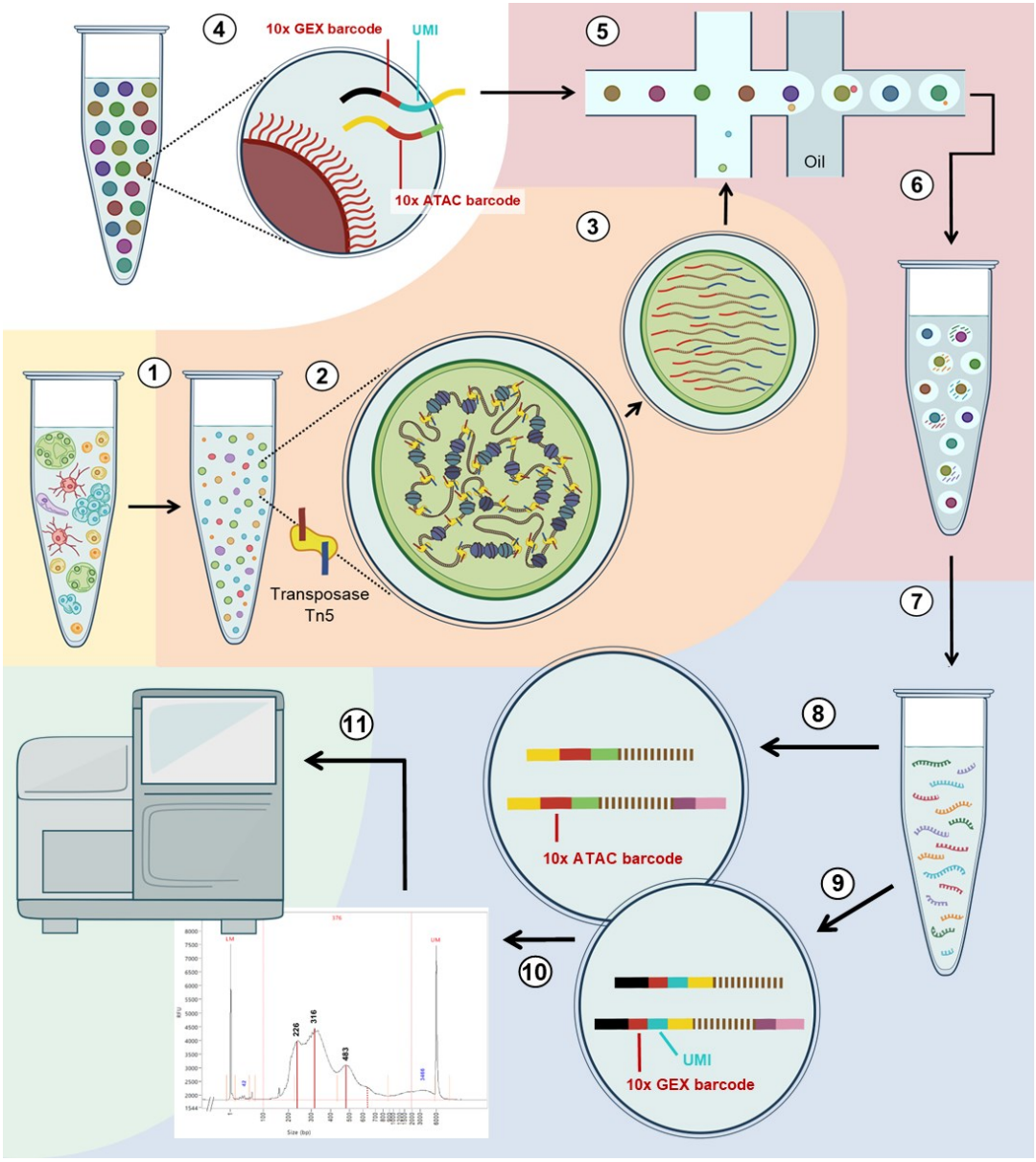

Figure 2) Schematic representation of single-nucleus Multiomics workflow. **1)** Cell lysis: single cells are broken up to extract intact nuclei. **2)** Transposition (for ATAC): Transposase Tn5 enters nuclei and binds to open chromatin regions which are not bound by nucleosomes. It cuts accessible DNA and tags it simultaneously with respective adaptors. This happens without breaking up nuclei. **3)** Single nuclei contain mRNA transcripts as well as tagged DNA fragments of different sizes (dependent on the number of nucleosomes these sequences have been bound to). **4)** Single Cell Multiome ATAC + GEX Gel Beads carry distinct poly(dT) sequences on their surface: these cover either 10x ATAC barcodes (specific for bead) or 10x GEX barcodes (specific for bead) and UMI sequences (specific for transcript). **5)** GEM (Gel Beads-in-emulsion) generation inside of the 10x Genomics Chromium controller: microscopical small droplets are formed which contain a single barcoded Gel Bead and a single transposed nucleus. To make sure that each GEM encloses a maximum of one nucleus, nuclei are delivered in high dilution, resulting in the majority of created GEMs containing no nucleus. **6)** Nuclei within intact GEMs are broken up: chromatin fragments can bind to the ATAC-barcode-labelled poly(dT) sequences on the surface of the Gel Bead, and mRNA transcripts can bind to the GEX-barcode-labelled sequences. This allows GEM-specific tagging of each fragment. **7)** GEMs are broken up and all fragments are pooled. A pre-amplification step happens. **8)** ATAC library construction: ATAC libraries include GEM-specific ATAC barcodes as well as all necessary sequencing adaptors. **9)** Gene expression library construction: cDNA synthesis happens to further create gene expression libraries. These include GEM-specific GEX barcodes, transcript-specific UMIs and sequencing adaptors. **10)** Quality check of constructed libraries. **11)** Illumina sequencing: libraries of sufficient quality are sent for sequencing.
Different background colours depict sections of the workflow that happen on different molecular levels: **yellow** = whole cells; **orange** (2 + 3) = whole nuclei in bulk**; red** (5 + 6) = single nuclei / generated GEMs; **blue** (7 – 10) = tagged fragments in bulk; **green** (11) = single fragments.

In comparison to other methods like sc-RNA-seq, sn-Multiomics relies on a single-nucleus instead of single-cell basis. Therefore, it is particularly advantageous when working with polynucleated cells, where the single-cell level does not automatically correspond to the level of single nuclei. In these cases, it is necessary to look at separate nuclei in order to exploit the highest possible information potential (studies comparing the ability of single-cell and single-nucleus RNA-seq data to predict cell types: Lake et al., 2017; Wu et al., 2019). Since this will become important in the given study, I will consciously speak of “single-nucleus” rather than “single-cell”, even though most literature still refers to “single cell”.

Obtaining sn-Multiomics data on unicellular holozoans will be useful to move forward in a broad spectrum of research interests. Our group in particular is interested in exploring the gene regulatory groundwork responsible for cell state transitions in unicellular holozoans and comparing it with cell differentiation trajectories in basally branching metazoans. Sn-Multiomics data will enable us to investigate changes in the transcriptional and epigenetic profiles of cells along their differentiation pathway from one cell state to another. Finding similarities as well as differences between involved transcription factors and target genes, downstream regulatory networks and profiles of chromatin accessibility will help us to reconstruct the genomic and regulatory equipment of a hypothetical unicellular ancestor of all animals (Ros-Rocher et al., 2021).

Apart from these concrete research interests, having sn-Multiomics data on unicellular organisms bears a lot of additional information potential. For example, it could help resolving complex life cycles and intermediate cell states of less-investigated species. Also, it will facilitate the characterization of transient multicellular / multinucleated states found in diverse unicellular holozoans (Marshall & Berbee, 2011; Ruiz-Trillo et al., 2023).

Therefore, the information content carried by single-nucleus Multiomics data on unicellular holozoans can not only enhance our understanding of certain organisms. Its impact could potentially reach up to a conceptual level of how we think about fundamental questions surrounding the evolutionary origin of animals (Marioni & Arendt, 2017; Paps, 2018): how does unidirectional cell differentiation in animals can be compared to, or distinguished from cell state transitions in unicellular organisms? How do cells reach their specification? How can a cell type itself be defined?

1.6. The goal of my thesis

The goal of my thesis was to generate sn-Multiomics datasets for representatives of two different unicellular holozoan lineages: *Abeoforma whisleri* (ichthyosporea) and *Capsaspora owczarzaki* (filasterea). In order to benefit from newly developed tools like sn-Multiomics also outside of clinical sectors, it is unavoidable to adapt them for the use on non-model organisms. Available protocols have been developed on well-investigated animal model organisms or cell lines. However, since each type of organism and cell comes with certain particularities, they can strongly differ in how they react to a defined experimental treatment. Accordingly, applying standardized protocols on poorly understood non-model organisms requires systematic preliminary experiments, testing and specifying potentially needed protocol adjustments.

Putting reasonable effort in organism-specific protocol optimization can be the key to obtain meaningful results. In the case of sn-Multiomics sequencing the very first step of nuclei isolation is particularly important, as it sets the stage for all following steps. Accordingly, the main focus of my Master's project was to optimize cell lysis and nuclei isolation protocols species-specifically for the two unicellular holozoans *Abeoforma whisleri* and *Capsaspora owczarzaki*. Being able to extract good-quality nuclei in sufficient quantity finally allowed the performance of sn-Multiomics on these species. Therefore, I was able to generate the first single-nuclei (or single-cell) dataset available for unicellular holozoans. These new datasets could thus deliver unprecedented insights into premetazoan cell differentiation that cannot be provided by already existing bulk ATAC- or RNA-seq data.

2. RESEARCH ORGANISMS

2.1. *Abeoforma whisleri*

Among known unicellular holozoans *Abeoforma whisleri* is one of the less well-investigated representatives. Despite major lacks in knowledge and understanding, it appears especially interesting for our research question due to its phylogenetic positioning and complex life cycle.

2.1.1. Discovery and ecology

Abeoforma whisleri was discovered in samples isolated from the digestive tract of the mussel *Mytilus* sp. and can be classified within ichthyosporea (Marshall & Berbee, 2011; Torruella et al., 2015). Most described ichthyosporea species are living as highly host-specific parasites in diverse invertebrates or vertebrates what makes many of them not culturable in the lab. However, *A. whisleri* is one of the few known representatives that can be cultured axenically independently of any host cells which is a major advantage for the given experimental set-up (Whisler, 1960; Marshall & Berbee, 2011; Glocking et al., 2011).

2.1.2 Morphological appearance & life cycle description

As the Latin name (“abeo” – changing; “forma” - shape) already indicates, the taxon *Abeoforma* shows a great variety of different cell states covering diverse morphological and functional characteristics. This made it comparatively difficult to capture all definitive cell states of this species and how they are connected to each other within the life cycle (Marshall & Berbee, 2011; Kożyczkowska et al., 2021; see Figure 3).

One particular cell state that can be observed with high abundance in cell culture are mononucleated spherical cells of about 10-15 µm diameter, with large central vacuoles and a single- or double-layered cell wall (see (1) in Figure 3). Starting from there, *Abeoforma* seems to undergo different ways of reproduction, passing diverse intermediate cell states to finally come back to such a mononucleated cell again. Characteristic of ichthyosporea, a mononucleated cell can grow up to a size of about 50 µm by undergoing multiple rounds of nuclei division (without cell division) and forming a multinucleated coenocyte (see (4) and (5) in Figure 3; Marshall & Berbee, 2011). Mature coenocytes themselves can develop various shapes deviating from the spherical one (described as plasmodial form in Marshall & Berbee, 2011) and can eventually also give rise to an actively moving cell state (see (6) and (7) in Figure 3). What is especially interesting about this mobile state is that it potentially combines two differentially specialised cell forms in one construct and therefore it might represent a form of spatial cell differentiation (unpublished data, Victoria Shabardina – Multicellgenome Lab; see Oxford Nanopore Technologies, 2022 for further information). A multinucleated unit can finally undergo cellularization and release endospore-like structures, producing a multitude of mononucleated spherical cells again (Marshall & Berbee, 2011). Also frequently observed are

amoeboid-like cells which again occur in multiple different forms and complicate the life cycle of *Abeoforma* even further (Marshall & Berbee, 2011; Kożyczkowska et al., 2021).

Altogether it seems that *Abeoforma whisleri* fundamentally follows what can be commonly found among many ichthyosporea, only with many alternative intermediate cell states and variations on top (Kożyczkowska et al., 2021). How those different cell states are all connected to each other within the life cycle and in what extent they represent stringently obligatory or facultative cell states, is not described in detail so far. This is part of Victoria Shabardina's work in the Multicellgenome Lab, which is yet to be published.

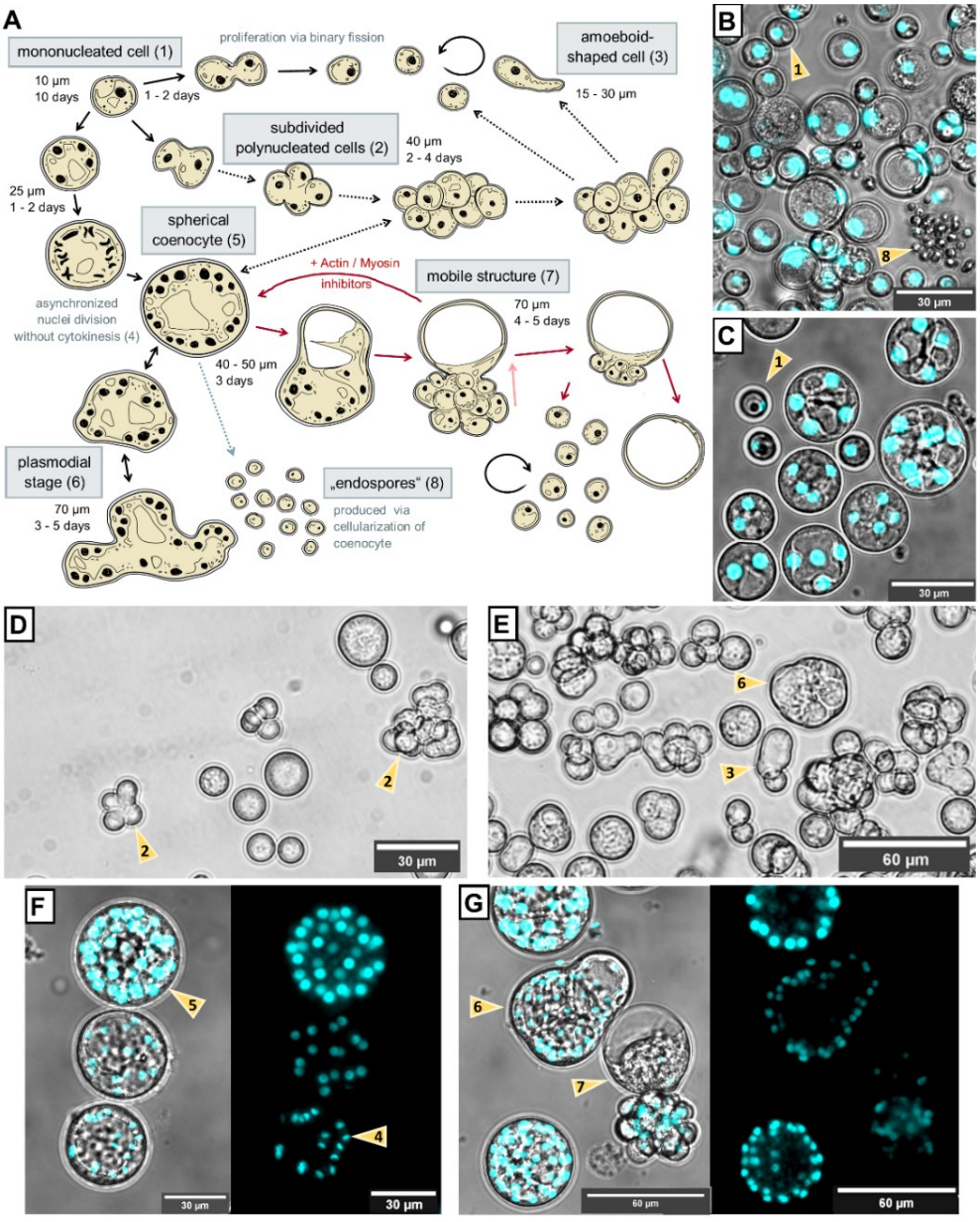

Figure 3) Life stages of Abeoforma whisleri. **A)** Illustration of cell states observable in cell culture and potential transition pathways; dotted arrows mark hypothetical transitions which have not been resolved yet. The pink arrow indicates the direction of movement of the mobile state. In pictures B-G different cell states are indicated with yellow arrows and respective numbers referenced in A. B, C, F, G additionally show Hoechst staining of nuclei in alive cells. **B)** Mother culture before setting up fresh subcultures shows mostly mononucleated cells. **C+D)** Subculture after 27hrs includes spherical (C) and subdivided (D) polynucleated cells. **E)** Subculture after 3 days includes spherical to plasmodial, but also subdivided polynucleated and amoeboid-shaped cells**. F+G)** Subculture after 4 days includes polynucleated coenocytes and mobile cells (G).

2.2. *Capsaspora owczarzaki*

The filasterean species *Capsaspora owczarzaki* is probably the best investigated unicellular holozoan outside of the comparatively well-studied choanoflagellates (Hoffmeyer & Burkhardt, 2016) and comes along with a broad set of already established tools (Kożyczkowska et al., 2021; Sebé-Pedrós et al., 2016a). Research carried out on *Capsaspora* gave fundamental insights into premetazoan genomic constitutions, making it an indispensable research organism in this field.

2.2.1. Discovery and ecology

Capsaspora owczarzaki was first isolated as an amoeboid endosymbiont in the haemolymph of the freshwater gastropod *Biomphalaria glabrata* (Stibbs et al., 1978). Genomic and molecular analyses place it within filasterea, the second closest related sister-group of metazoans (Hertel et al., 2002; Torruella et al., 2011).

Biomphalaria glabrata is an intermediate host of the trematode *Schistosoma mansoni* which is the main cause for intestinal schistosomiasis in humans being the definitive hosts. *Capsaspora owczarzaki* was discovered to be involved in the snail's resistance towards *S. mansoni* infection by killing and degrading sporocysts of the trematode (Stibbs et al., 1978; Owczarzaki et al., 1980).

2.2.2. Morphological appearance & life cycle description

Capsaspora cells are about 3 to 7 µm in diameter, housing a single nucleus which is about one-third to one-half of the cells' size (Stibbs et al., 1978; Hertel et al., 2002). As it is known for most unicellular holozoans, they transit between distinct cell states triggered by environmental stimuli (see Figure 4). However, by now *Capsaspora owczarzaki* is the only filasterean whose life cycle has been described in detail (Sebé-Pedrós et al., 2017).

Three morphologically distinguishable stages can be observed, which also differ in their cell function and proliferation activity (see Figure 4.A). First, amoeboid cells show a single-layered cell membrane and form several filopodial: by extending and retracting those extensions the cells can move while they are attached to the surface (Stibbs et al., 1978). Amoeboid cells are phagocytic trophic and proliferate exponentially by cytokinesis (Stibbs et al., 1978; Sebé-Pedrós et al., 2016; Sebé-Pedrós et al., 2017). In cell culture, this state is the first one that can be observed in a fresh subculture where it builds a monolayer of bottom-attached cells (Sebé-Pedrós et al., 2013b). Increasing culture density, which leads to successive depletion of available nutrient resources, results in *Capsaspora* cells starting to transit into a cystic state. Cells lose their filopodia, round up and form a floating cyst. These cells build a double cell wall and do not longer proliferate (Stibbs et al., 1978; Hertel et al., 2002).

Finally, a third state can be described where multiple cells without filopodia cluster together. They form floating cell aggregates which are held together by a common extracellular matrix (Sebé-Pedrós et al., 2013b). Those multicellular constructs can occur randomly or can be specifically induced by osmotic stress, agitation or release of starvation stress; still it is not yet understood what this aggregate formation is a reaction to. Sebé-Pedrós et al. (2013c) found experimental proof that these clusters are the result of real aggregation and not clonal division as proliferation activity of cells within these clusters is minimal. This form of temporal multicellularity based on multiple, genetically distinct cells clustering together was reported for the first time within unicellular holozoans (Sebé-Pedrós et al., 2017).

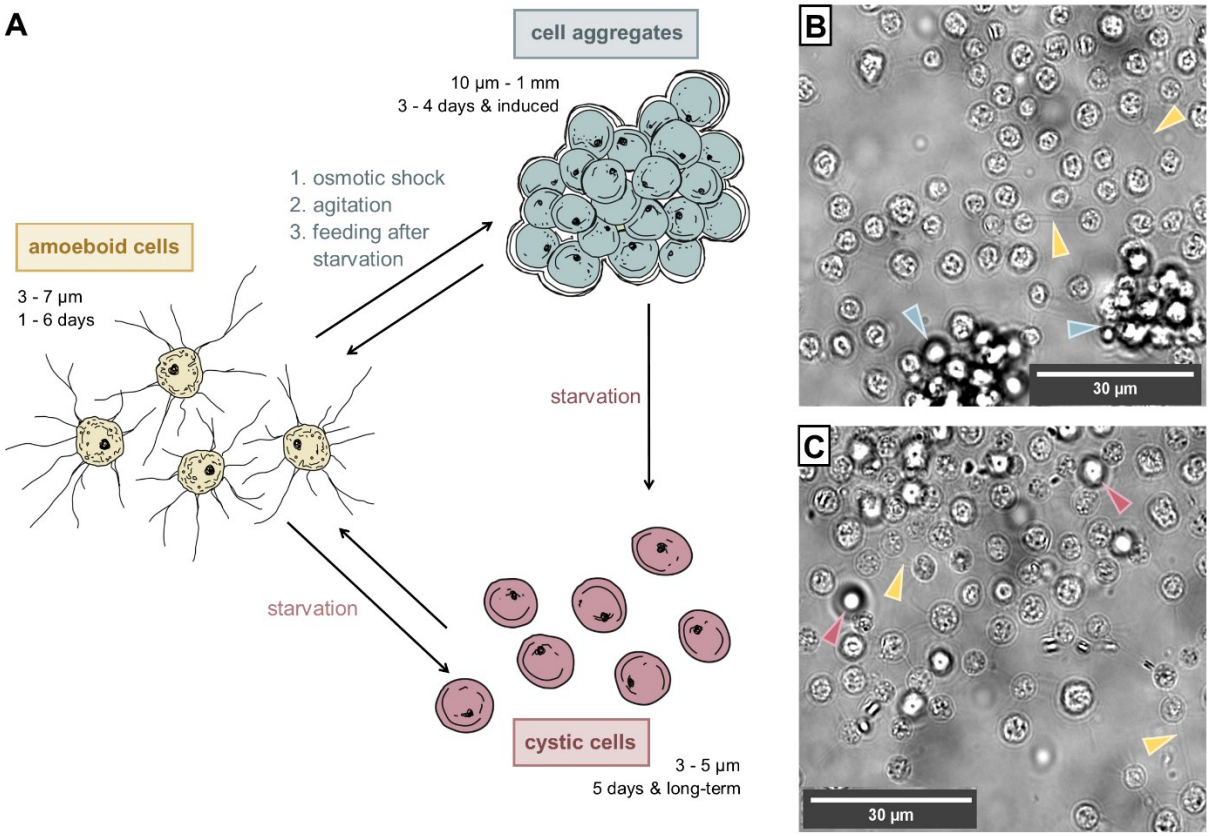

Figure 4) Life stages of *Capsaspora owczarzaki*. A) Illustration of cell states observable in cell culture and transitions between those within the life cycle. **B)** 3 days old cell culture contains amoeboid cells attached to the ground (in focus, filopodia of these cells are indicated by yellow arrows) as well as cell aggregates in the floating phase (out of focus, marked by blue arrows). **C)** 6 days old culture contains amoeboid cells (in focus, filopodia are indicated by yellow arrows) as well as cystic cells in the floating phase (out of focus, marked by red arrows).

3. MATERIAL & METHODS

3.1. Cell culture and growth conditions

3.1.1. Maintenance of cultures for optimization experiments

The cells used to initially set up the cell cultures of *Capsaspora* and *Abeoforma* were provided by the Multicellgenome Lab of Iñaki Ruiz-Trillo and Elena Casacuberta in February 2023. For observing cells and estimating the state of cultures Hoechst 33342 staining of alive cells was used.

Abeoforma whisleri:

Standard culturing conditions: *Abeoforma* cells were grown axenically in 25 cm^2 cell culture flasks using Marine Broth medium 2216 (Ref. 279110, BD-Difco). For culture maintenance, cells were kept in an incubator at 17°C and subcultures were initiated every 4-5 days by pipetting 150 µl old culture into 5 ml fresh Marine Broth medium.

For experiments done during the optimization process a mixture of cultures of different ages was used. For most experiments a 2-4 days old culture was combined with a more mature culture, 6-9 days of age. No special treatments were applied to enrich certain cell states; culturing conditions were selected to achieve asynchronized cultures in terms of present cell states.

Capsaspora owczarzaki:

Standard culturing conditions: *Capsaspora* cells (strain ATCC30864) were cultured axenically in 25 cm^2 cell culture flasks (Ref. 156340, Nunc™ EasYFlask™ Cell culture flasks) using ATCC medium 1034 (modified PYNFH medium; see supplementary material). For culture maintenance, cells were kept in an incubator at 21°C. Subcultures were initiated every 5-7 days by thoroughly scraping off cells and using 100 µl initial culture in 5 ml fresh medium.

During the optimization process, experiments were carried out on mixed samples (containing all different cell states) as well as on samples having either the filopodial or cystic cell state particularly enriched. For mixed samples, a 2-4 days old culture and a 5-8 days old culture were combined and attached cells were scraped off thoroughly before harvesting. To obtain samples containing mostly filopodial cells, 6-9 days old cultures were used which have been prepared respectively: by replacing the cell culture medium every 2 days without scraping off attached cells, the majority of cells in the floating state (cystic or aggregates) was regularly removed, while attached filopodial cells could form a dense monolayer on the bottom of the culture flask. The last medium exchange was carried out about 20-24 hours before the respective experiment was carried out and cells were carefully scraped off when finally

harvested. For cystic stage samples, 7-14 days old cultures were used. In this case only cells in the floating phase of the culture were collected, without scrapping off attached cells.

3.1.2. Cell culture set-up for final sn-Multiomics experiment

In the final sn-Multiomics experiment as many different cell states and transient conditions as possible should be covered. In order to do so, rather asynchronized cultures of different ages and stages were combined to one final sample. For both species seven different subcultures were used, covering a culture age range of one day (D1) to seven days (D7): starting with the D7 subsample seven days before carrying out the final experiment, all following samples have been set up with a time difference of about 24hrs to each other respectively. Subculturing was done by using well-developed 5-8 days old cultures; the used volumes of initial cultures is given in the following:

Abeoforma: D1 = 200 µl initial culture in 5 ml medium; D2 = 175 µl in 5 ml medium; D3 / D4 / D5 = 150 µl in 5 ml medium; D6 / D7 = 100 µl in 5 ml medium.

Capsaspora: D1 = 250 µl initial culture in 5 ml medium; D2 = 150 µl in 5 ml medium; D3 = 125 µl in 5 ml medium; D4 = 100 µl in 5 ml medium; D5 = 100 µl in 5 ml medium; D6 = 80 µl in 5 ml medium; D7 = 50 µl in 5 ml medium.

All used cultures have been grown under standard culturing conditions (see 3.1.1) without receiving any cell state-specific enrichment treatment. The proportional occurrence of different cell states in used cultures is therefore supposed to be representative for their abundance in untreated cultures of the respective age.

3.2. Cell lysis and nuclei extraction

3.2.1 Optimization experiments

The initial protocol that was used as starting point to go on with *Capsaspora*- or *Abeoforma*-specific adaptations was provided by Thomas Grentzinger of the NGS facility at Vienna BioCenter Core Facilities. This, on the other hand, is strongly based on the official 10x Genomics protocol CG000169 “Nuclei Isolation for Single Cell ATAC Sequencing” (10x Genomics Inc, 2022a) (identically, except of the usage of DTT in the lysis and wash buffer).

3.2.1.1. Tested parameters

In order to come up with improved, *Capsaspora*- and *Abeoforma*-specific lysis conditions, systematic changes in the following parameters were carried out: concentration of lysis buffer, volume of lysis buffer (equate to cell density), lysis time and time point of filtering. For *Capsaspora* additional experiments were done testing if there is any detectable difference in lysis results for cells of different states. Accordingly, samples were assigned to the categories “filopodial” (see 3.1.1), “cystic” and “mixed” for samples containing cells of all three different states.

For each experiment the age and overall state of used cultures (see 3.1.1) as well as the exact number of input cells was documented. The concentration of cells in used cultures was measured and calculated by using a Neubauer improved haemocytometer (0,10 mm depth; Superior Marienfeld, Stuttgart, Germany) on a standard brightfield microscope.

3.2.1.2. Evaluation of lysis results

The lysis success of each experiment was subsequently judged in terms of quantity of single nuclei isolated, occurrence of nuclei aggregates, viability of cells after lysis (percentage of non-lysed cells) as well as the visual appearance of the sample under the microscope (taking into account e.g. cleanliness of the sample and the amount of remaining cell parts which could potentially hamper following processing steps).

Since the aim of this study was to generate sn-Multiomics data covering all observable cell states as comprehensive as possible, one important aspect is monitoring if the lysis reaction works comparatively well for all used cell states. Due to the very distinct characteristics of *Capsaspora* and *Abeoforma* cells*,* testing the influence of “cell state belonging” on lysis efficiency was not possible in the same extend for both species. The life cycle of *Capsaspora* is significantly simpler and fully resolved. It is easily possible to enrich a specific state in cell culture that can then be harvested separately from the others. Thus, for *Capsaspora* it was possible to carry out additional experiments testing the same lysis conditions on different cell states. They were compared among each other as well as in relation to mixed samples including all cell states. *Abeoforma*, however, shows a very complex life cycle including many

unresolved intermediate states which does not allow harvest of separated cell states. In this case, homogeneity of lysis efficiency could only be evaluated superficially by comparing samples of different culture ages which contained different cell states in varying proportions.

All parameters used for the evaluation of lysis success were analysed by using propidium iodide staining of isolated nuclei and subsequent observation using the Nikon Eclipse 80i Upright Microscope (Nikon Instruments, Melville, USA). For quantification of nuclei a haemocytometer and in addition a Cellometer X2 Fluorescent Viability Counter (Nexcelom Bioscience, Lawrence, USA) were used.

Statistical analysis of lysis success

Among all parameters used to evaluate lysis success, “quantity of isolated nuclei” and “percentage of non-lysed cells” were chosen to be the two most important and most objective ones. Therefore, those two parameters were used as dependent variables to carry out further statistical analysis on the significances of applied changes (as independent predictor variables) in lysis buffer concentration, lysis time as well as used cell states (only in the case of *Capsaspora*). All calculations were performed in SAS 9.4 M2 (SAS Institute, North Carolina, USA), while for some preparative steps the software PAST 4.15 (Natural History Museum Oslo, Norway) was used in addition. Graphics were created in SAS 9.4 M2 and subsequently processed using the illustration software Inkscape 1.3.1 (Boston, USA).

1) For testing the statistical effect of used cell states on lysis success in *Capsaspora* an ANOVA-based test approach was used (“MIXED procedure” in SAS) on previously square root transformed data. The test was carried out separately for “quantity of isolated nuclei” and “post-lysis viability”. For both, a reduced dataset of n = 33 was used, only containing samples that were assigned to the categories “filopodial” or “cystic” within the predictor variable “cell state”, since “mixed” samples would contain cells of both types in varying proportions. Preliminary an ANCOVA-test was carried out considering “lysis buffer concentration” and “lysis time” as possible covariables besides the main predictor “cell state”. This test provided no relevant significant results and showed that applying a one-way ANOVA describes the given context sufficiently, without including “lysis buffer concentration” and “lysis time” as additional covariables.
2) The effects of lysis buffer concentration and lysis time were analysed in parallel to each other by using a multiple regression approach (“RSREG procedure” in SAS): a response surface regression model was calculated which accounts for linear, quadratic as well as cross-product effects of the tested predictor variables. Again, this test was carried out separately for “quantity of isolated nuclei” and “percentage of non-lysed cells”. In the case of *Capsaspora* the regression model was based on a dataset of n = 62 which had been previously square root transformed; for *Abeoforma* a dataset of

n = 40 was used that had been square root (in case of “quantity of isolated nuclei”) or inverse square root transformed (in case of “percentage of non-lysed cells”).
The dependency of „quantity of isolated nuclei” and “percentage of non-lysed cells” on the predictor variables “lysis buffer concentration” and “lysis time” was visually depicted by creating a scatter plot including a curvilinear regression plot as well as a response contour plot in SAS.

Effects of different cell densities (lysis buffer volumes) and time points of filtering were not evaluated statistically due to the comparatively small sample size targeting those parameters, which did not allow for meaningful statistics.

Therefore, the described statistical analyses do not cover all parameters that have been investigated in the scope of lysis success evaluation and can therefore not exclusively explain all decisions and adjustments made during the optimization process. Other parameters bearing a rather non-quantifiable nature, like the visual appearance of extracted nuclei under the microscope were documented separately. Combined with the results of the first tagmentation test (see 3.2.1.3.), those three ways of evaluation could altogether provide the necessary information basis to specify the final optimized lysis protocol.

3.2.1.3. Quality assessment through tagmentation test

Close to the end of the optimization process a bulk ATAC tagmentation test was carried out, in order to assess to chromatin integrity of extracted nuclei. For this test three different samples were handed in: one *Abeoforma* sample and two differently processed *Capsaspora s*amples. The overall lysis procedure and sample handling (such as quantification of cells, centrifugation steps etc.) for these samples was similar to what is described below in 3.2.2. In the following, only those parameters will be outlined which specifically differ between the final optimized protocol and the protocol applied to extract nuclei for the tagmentation test.

For the *Abeoforma* sample a 3-days old culture was combined with a 4-days old culture. A sample volume corresponding to 900.000 cells was used for the extraction. The cells were lysed for 5 min, using 200 µl 1x lysis buffer. For the first *Capsaspora* sample a 4-days old culture was used: cells were scraped off from the cultivating flask to specifically harvest a majority of filopodial cells. A sample volume corresponding to 2.500.000 cells was used and cells were lysed for 1 min in 100 µl 14x lysis buffer. For the second *Capsaspora* sample a 7-days old culture was used; mostly cells in the floating phase were harvested to specifically obtain mainly cystic cells. Again, a sample volume corresponding to 2.500.000 cells was used and lysis happened for 1 min, using 100 µl of 16x lysis buffer. For all three samples extracted nuclei were resuspended in 100 µl nuclei buffer solution.

After lysis, the extracted nuclei were directly passed on (kept on ice in the meantime) to the NGS Facility at Vienna BioCenter Core Facilities for further processing.

3.2.2. Nuclei extraction set-up for final sn-Multiomics experiment

3.2.2.1. Optimized protocol for nuclei isolation for *Abeoforma whisleri*

Abeoforma cells were harvested, combining aliquots of the following volumes of the seven different subcultures in a 2 ml Eppendorf tube: 400 µl of D1 subculture, 400 µl of D2, 400 µl of D3, 300 µl of D4, 200 µl of D5, 150 µl of D6, 150 µl of D7. A haemocytometer and a 10 µl propidium iodide-stained aliquot were used to count alive (not stained by PI, only visible under brightfield) as well as broken cells (stained by PI), in order to calculate concentration of cells and viability. Viability of used cultures should be > 85%; in healthy cultures of this age, it is usually close to 100%. A volume of the well-mixed, final sample corresponding to about 700.000 cells was transferred into a new 2 ml Eppendorf tube.

During all following steps of nuclei isolation, the sample and all used reagents were kept on ice and waiting times were minimizes. This is important to ensure the best possible quality of isolated nuclei.
1.800 µl final wash buffer (WB, see supplementary material) was made by adding DTT (final concentration: 0,001 M) and RNase inhibitor (final concentration: 1 U/µl) to already prepared wash buffer solution and kept on ice. Cells were pelleted, by centrifuging the sample for 5 min, at 4°C and 500 g. The supernatant medium was removed, and cells were resuspended in 200 µl 1x lysis buffer (LB: 194 µl WB incl. DTT and RNase inhibitor + 2 µl NP40 10% (final concentration: 0,1 M) + 4 µl Digitonin 0,5% (final concentration: 0,01)), by pipetting carefully but thoroughly for 15 seconds. Cell lysis was stopped after 2 min by adding 1.600 µl WB. The total sample volume of 1.800 µl was filtered through a 30 µm MACS SmartStrainer (Miltenyi Biotec, Bergisch Gladbach, Germany). The sample was centrifuged for 10 min, at 4 °C and 500 g. The supernatant was removed, and nuclei were resuspended in 100 µl nuclei buffer solution (NBS), pipetting carefully for 15 seconds. Acridine orange / propidium iodide staining of a 10 µl aliquot and a Cellometer X2 Fluorescent Viability Counter (Nexcelom Bioscience, Lawrence, USA) were used to determine the concentration of isolated nuclei in the final sample. For the Cellometer the following settings were used: 5 msec exposure for brightfield image, 4 sec exposure for AO/PI image, 7-50 µm accepted cell diameter, 3-10 µm accepted nuclei diameter. Following the given protocol, a sample with a concentration of 8.540 nuclei per 1 µl could be obtained for the final sn-Multiomics experiment.
For more detailed information on the optimized protocols and additional notes see the supplementary material.

3.2.2.2. Optimized protocol for nuclei isolation for *Capsaspora owczarzaki*

Capsaspora cells were harvested, combining aliquots of the following volumes of the seven different subcultures in a 2 ml Eppendorf tube: 300 µl of D1 culture, 270 µl of D2, 250 µl of D3, 220 µl of D4, 180 µl of D5, 150 µl of D6, 130 µl of D7. The culture state and cell concentration of the mixed, final samples were estimated in the same way as already described in 3.2.2.1 for *Abeoforma.* A sample volume containing about 2.500.000 cells was transferred into a new 1,5 ml Eppendorf tube.

During all following steps of nuclei isolation, the sample and all used reagents were kept on ice and waiting times were minimizes.
900 µl final wash buffer was made by adding DTT and RNase inhibitor to already prepared wash buffer solution and kept on ice. Cells were pelleted, by centrifuging the sample for 5 min, at 4°C and 500 g. The supernatant medium was removed, and cells were resuspended in 100 µl 16x LB (88,8 µl WB incl. DTT and RNase inhibitor + 8 µl NP40 20% (final concentration: 1,6 M) + 3,2 µl Digitonin <5% (final concentration: 0,16)), by pipetting carefully but thoroughly for 15 seconds. Cell lysis was stopped after 1 min lysis time by adding 800 µl WB. The total sample volume of 900 µl was filtered through a 30 µm MACS SmartStrainer. The sample was centrifuged for 10 min, at 4 °C and 500 g. The supernatant buffer solution was removed, and nuclei were resuspended in 150 µl NBS, pipetting carefully for 15 seconds. As for *Abeoforma* a Cellometer was used to determine the concentration of isolated nuclei in the final sample. Since *Capsaspora* nuclei are way smaller than those of *Abeoforma*, the settings for the Cellometer had to be strongly adapted: 5 msec exposure for brightfield image, 12 sec exposure for AO/PI image, 3-7 µm accepted cell diameter, 2-25 µm accepted nuclei diameter. Using these settings, the machine might not be able to distinguish between single nuclei and nuclei aggregates, but this was judged to be a minor disadvantage in favour of the very beneficial time aspect given by using the Cellometer. Following the described protocol, a sample with a concentration of about 7.420 nuclei / 1 µl could be obtained for the final sn-Multiomics experiment.

3.3. sn-Multiomics sequencing

Once nuclei were extracted and quantified, the samples were kept on ice while they were passed on to Thomas Grentzinger of the NGS Facility at VBC Core Facilities. He performed the tagmentation and preparation of sn-ATAC and sn-RNA 10x barcoded libraries according to the official 10x Genomics protocol CG000338 “Chromium Next GEM Single Cell Multiome ATAC + Gene Expression” (10x Genomics Inc., 2022b). Subsequently, Ilumina sequencing of the 10x libraries was performed (Illumina NovaSeq S4 PE150 XP).

3.3.1. Bioinformatical analysis of sn-Multiomics data

All bioinformatical analysis on the obtained sn-Multiomics data were performed by Carolina Atria. For the *Abeoforma* sample the raw sn-Mutliomics sequencing data, together with the annotated reference genome was processed in 10x Genomics Cell Ranger ARC. To perform quality control of the gene expression data, the raw Cell Ranger ARC output was processed with the R-package DropletUtils. By doing so, a Barcode rank plot was created which gives information about the distinction between real cells and empty droplets or doublets. For ATAC data quality control the raw data was processed using the tools chromap and bwa and the output was subsequently used in the R-package ATACseqQC. An insert size distribution plot was created that assesses tagmentation success and chromatin integrity in used nuclei.
The *Capsaspora* sample was initially treated similar but since these first analyses did not show interpretable results a second alternative approach was used in addition. For doing so, the raw sn-ATAC and sn-RNA sequencing results were handled as separate datasets. Firstly, the raw sn-RNA-seq data was processed in 10x Genomics Cell Ranger, using the annotate genome as a reference. The Cell Ranger output was subsequently processed with the R-package DropletUtils, in order to create a barcode rank plot as RNA-seq quality control. Secondly, the raw sn-ATAC sequencing results were processed using 10x Genomics Cell Ranger ATAC. Then the raw Cell Ranger ATAC output was used in the R-package ATACseqQC to generate a quality control insert size distribution plot.
Annotated genomes used as a reference were provided by the Multicellgenome Lab of Ruiz-Trillo. The *Abeoforma* genome was generated in collaboration with the Lab of Alex de Mendoza, the used *Capsaspora* genome is still unpublished data of Hiroshi Suga and Arnau Sebé-Pedrós. The quality and completeness of all used genomes and transcriptomes was tested by using BUSCO 5.4.5 against eukaryotes.

4. RESULTS

4.1. Establishment of a sn-Multiomics protocol for unicellular holozoans

Single nucleus-Multiomics has been established in a number of model organisms and well-investigated cell lines. Main steps of the sn-Multiomics workflow include nuclei extraction, nuclei transposition, GEM generation, library construction and sequencing (see Figure 2 for detailed information). While some steps like library construction are likely to be applicable to samples of all different kinds of organisms, this is not the case for cell lysis and nuclei extraction. Cells of different types and species can heavily differ in how they behave during the lysis reaction. Thus, certain adaptations of established protocols are most probably necessary to account for respective particularities of each cell type. Therefore, my main task in establishing sn-Multiomics on the unicellular holozoans *Abeoforma whisleri* and *Capsaspora owczarzaki* was to develop optimized protocols for nuclei extraction by adjusting lysis parameters species-specifically. Also, as nuclei are expected to be rather sensitive once they are isolated, another important purpose of protocol optimization was to ensure that all processing steps work as quickly and smoothly as possible to not impact nuclei quality in the final experiment.

4.1.1 Quantifying initial cell concentration and determining input cell number

Before performing the actual step of nuclei extraction, concentration and viability of harvested cells need to be estimated in order to calculate the required culture volume (corresponding to a particular input number of cells). The original protocol suggested this step to be done by using a haemocytometer and trypan blue staining as it is a very classical method in cell biology to distinguish between alive and broken cells. However, preliminary tests showed that application of this dye did not work for either of the two tested unicellular species. Trypan blue either immediately precipitated in the used culturing medium (Marine Broth medium, in the case of *Abeoforma*) or didn't show any meaningful staining of dead cells (in the case of *Capsaspora*) (see Figure 5). Therefore, it became necessary to test alternative methods for measuring viability of used cell cultures. I found that staining with the fluorescent dye propidium iodide (PI) shows appropriate results: it specifically enters and stains dead *Abeoforma* and *Capsaspora* cells without harming living ones, similar to what is known from other organisms (see Figure 5). Accordingly, PI has been used to assess cell viability throughout the following study.

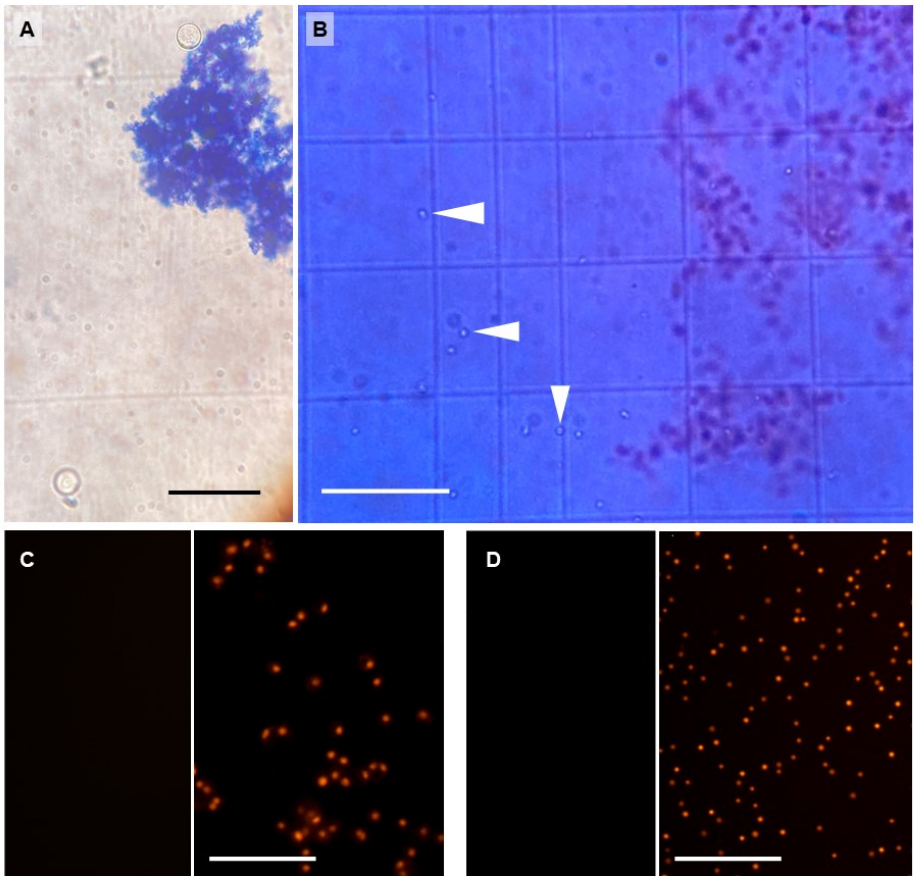

Figure 5) Comparison between trypan blue (TB) and propidium iodide (PI) staining to distinguish dead from alive cells. **A)** Alive *Abeoforma whisleri* cells stained with TB: trypan blue precipitated immediately. **B)** EtOH killed *Capsaspora owczarzaki* cells stained with TB: white arrows indicate dead but not stained cells. **C + D)** *Abeoforma* (C) or *Capsaspora* (D) stained with PI before (left) and after (right) killing cells with EtOH: PI only enters and stains dead cells. Given scale bars measure 50 µm.

Being able to quantify the viability and number of cells at the beginning of the process, allowed me to establish the optimal amount of cells used as an input for the lysis reaction. Existing protocols suggest starting off with 250.000 cells to obtain a desired nuclei concentration of 16-22.000 nuclei per 5 µl (~ 3.000 - 4.000 cells per µl) after cell lysis. However, preliminary experiments indicated that using a significantly higher input cell number seems to be beneficial in the given case. This concerns time efficiency (additional centrifugation and concentration steps are not anymore needed) as well as cell loss which appears to proportionally decrease with increasing cell number (data not shown). Using 700.000 *Abeoforma* cells or 2.500.000 *Capsaspora* cells as an input turned out to show satisfactory and reproducible results, achieving a nuclei concentration of about 6.000 - 8.000 nuclei per µl (diluted to a final concentration of 3.000 - 4.000 nuclei per µl).

4.1.2 Optimizing lysis efficiency and chromatin integrity

During the lysis process, single cells are resuspended in lysis buffer. This buffer contains the detergents Tween-20, digitonin and NP40 which serve to break up the outer cell membrane and degrade remaining cytoplasmatic parts of the cells. Ideally only single, intact nuclei should remain. Once this has happened, lysis has to be stopped before the lysing reagents can damage the nuclear membrane and, even more critically, the contained chromatin. This step is critical as the quality of obtained sn-Multiomics data is heavily dependent on the number of nuclei used in the experiment and the integrity of contained chromatin and mRNA.
The efficiency of lysis can be estimated by the number of single nuclei that can be obtained out of a certain amount of input cells (quantity of isolated nuclei) and the number of cells that stay unbroken after lysis (post-lysis viability = percentage of non-lysed cells). To maximize lysis efficiency for *Abeoforma* and *Capsaspora* cells, changes in several parameters throughout the standard lysis protocol were tested.

4.1.2.1. Optimization process for *Abeoforma* cells

In order to do so, I started off by applying the standard 10x Genomics nuclei extraction protocol which has been established on PBMCs cell lines (10x Genomics Inc., 2022a). By using PI staining combined with fluorescence and brightfield microscopy I monitored the quantity and morphological appearance of isolated nuclei as well as remaining non-lysed cells. Already standard lysis conditions (100 µl 1x lysis buffer, 5 min lysis time) allowed successful lysis of most *Abeoforma* cells (about 5% non-lysed cells remaining; see Table 2, yellow column). However, a rather large proportion of the nuclei extracted formed aggregates (see Table 2). Given the workflow of 10x Genomics sn-Multiomics (Figure 2), excessive occurrence of such nuclei aggregates can seriously hamper the ability to extract information on a single nuclei resolution level. Also, microscopic observation of samples revealed the presence of large cell structures of up to 70 µm which remained unbroken after lysis (see Figure 8.C). These seemingly empty, bubble-like structures might result from cells that took up water during lysis due to the hypotonic conditions of the buffer but failed to lyse. In any case, the morphological appearance could not be referred to any specific cell state that is known from *Abeoforma.* Independent of their origin, it was clear that those large structures remaining unbroken after lysis can bear a serious risk in further processing, since such big particles could clog the microfluidics system of the 10x Genomics controller (see step 5 in Figure 2).
Therefore, two main challenges needed to be overcome: drastic reduction of nuclei aggregates and complete elimination of non-lysed cells to prevent occurrence of such large cell structures.

Since first experiments showed that already lysis buffer of standard concentration (1x) successfully lyses a considerably large percentage of *Abeoforma* cells, I focused following experiments on increasing lysis efficiency (and decreasing occurrence of aggregates) without

increasing lysis buffer concentration. Systematic changes of lysis conditions showed that lowering the cell density to half (by using the double amount of lysis buffer) results in nearly complete elimination of non-lysed cell structures and significant reduction of nuclei aggregates by nearly 20 % (see Table 2, blue column). Accordingly, lowering cell density turned out to be a key step to solve the two challenges formulated above.

Table 2) Summary of lysis efficiency using different cell densities and lysis buffer concentrations. The standard protocol suggests the usage of 100 µl 1x lysis buffer (LB). Experiments on *Abeoforma* cells showed that halving cell density by using 200 µl LB allows to clearly lower the percentage of remaining non-lysed cells and isolated nuclei forming aggregates, while achieving a comparable quantity of isolated nuclei.

lysis buffer volume (cell density)	**100 µl LB** (3.500 cells per µl)			**200 µl LB** (1.750 cells per µl)		
lysis buffer concentration	**0,2x**	**0,5x**	**1x**	**0,2x**	**0,5x**	**1x**
quantity of isolated nuclei (per 5 µl)	20.750	6.750	12.250	34.750	9.250	12.250
percentage of non-lysed cells	23,8 %	22,2 %	5,4 %	3,1 %	3,7 %	1,6 %
percentage of nuclei forming aggregates	78,5 %	45,2 %	22,2 %	31,9 %	15,0 %	4,0 %

As Table 2 depicts, additional experiments have been carried out using lower lysis buffer concentrations (0,5x and 0,2x) to explore the possibilities of even milder lysis conditions. However, since treatment with lower concentrated lysis buffer could not eliminate non-lysed cell structures and nuclei aggregates as efficiently (Table 2), I considered working with 1x lysis buffer to be optimal. These correlations could also be confirmed by statistical analyses carried out on obtained experimental data (see Table 3 and Figure 7). They showed significant dependence of the percentage of non-lysed cells and the quantity of isolated nuclei on the concentration of used lysis buffer. Also, the hypothesized optimum of lysis efficiency could be reconstructed at a lysis buffer concentration between 1x and 1,5x.

In order to effectively progress in protocol optimization, the necessary next step was to test the integrity of chromatin in nuclei that have been isolated by using the conditions specified above (lowered cell density in 1x concentrated lysis buffer). This can be done through a tagmentation assay. For this purpose, the transposition step of the sn-Multiomics workflow is carried out on extracted nuclei (see step 2 in Figure 2). During tagmentation the transposase enzyme Tn5 enters single nuclei and cuts the double-stranded DNA specifically at accessible chromatin regions which are not bound by nucleosomes. The distribution of generated fragments can subsequently be measured by using a Fragment Analyzer. Results are displayed in a fragment size distribution plot which will give confirmation about how well tagmentation itself worked and about chromatin integrity on a bulk level. Figure 6 shows the respective fragment size distribution plot that could be obtained for *Abeoforma* nuclei isolated by using 200 µl 1x lysis buffer. It displays the count of chromatin fragments (y-axis) of a certain length (x-axis) that are present in the sample after tagmentation has been performed.

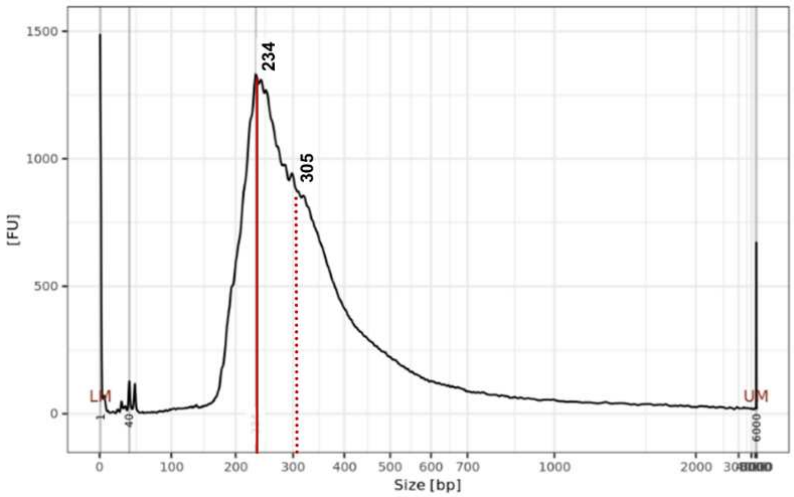

Figure 6) Fragment size distribution plot for the *Abeoforma* sample (treated with 200 µl 1x LB, 5 min lysis time): number of fragments (FU, y-axis) is plotted against fragment size (in bp, x-axis). The red line marks the peak representing mono-nucleosome-bound DNA regions; the dotted red line indicates di-nucleosome-bound regions.

In an ideal case, the distribution plot shows several, clearly distinguishable peaks representing chromatin regions that have been bound by one (peak at ~200 bp), two (350 - 400 bp) or three (500-600 bp) nucleosomes (Yan et al., 2020; 10x Genomics Inc., 2022b). Fragments shorter than 200 bp are suspected to originate from nucleosome-free regions.

The tagmentation test results obtained for isolated *Abeoforma* nuclei indicates that the tagmentation process itself worked (no high content of fragments bigger than 500 bp), but chromatin integrity itself was not optimal. The distribution of present fragments is clearly shifted to a rather short size of about 200 to 350 bp, indicating that chromatin got partly fragmented already during the lysis process which leads to a reduced number of long DNA fragments. Still, the results showed a clear peak for mono-nucleosome-bound fragments (solid red line, Figure 6) and a slight shoulder of the distribution curve between 300-350 bp (dotted red line) suggests a certain quantity of di-nucleosome-bound fragments.

Improving chromatin integrity

Results shown in Figure 6 strongly suggested that figuring out milder lysis conditions is most likely necessary to improve the quality of isolated *Abeoforma* nuclei and ensure high quality of resulting sn-Multiomics data. Using lower lysis buffer concentration to alleviate lysis conditions has already been tested in previous experiments but turned out to result in increased presence of non-lysed cells and nuclei aggregates (see Table 2). Therefore, I decided to test changing one condition which has not been examined yet: the duration of lysis which is suggested to be 5 min in the standard protocol (10x Genomics Inc., 2022a).

Accordingly, shorter lysis durations (4, 3, 2 and 1 min) were tested in order to find the minimum lysis time that still shows high lysis efficiency. Statistical tests comparing the lysis outcome obtained for different lysis times did not detect any significant difference in terms of quantity of isolated nuclei or percentage of non-lysed cells (Table 3 and Figure 7).

This indicates that the main process of breaking up cells and extracting nuclei already happens within the first one to two minutes of exposure to the lysis buffer. Even using a standard lysis buffer concentration (1x), longer lysis times might lead to successive damage of the nuclei membrane. Even though this potential partial damage is not reflected in decreased nuclei quantity (Table 3: “lysis time” is statistically not significant for quantity of isolated nuclei) it could explain the sub-optimal chromatin integrity shown in Figure 6.

Table 3) Summary of the results for statistical analysis on lysis success for *Abeoforma*; using multiple regression to model dependency of “quantity of isolated nuclei” and “percentage of non-lysed cells” on the two predictor variables “concentration of lysis buffer” (conc) and “lysis time” (time). Significant correlations, showing a *p*-value < 0.05, are marked in light red. Concentration of lysis buffer has a significant linear as well as quadratic effect, while lysis time appears to have no significant influence on lysis efficiency. More detailed results are given in Supplementary Table 1 & 2.

p-values for response surface regression model		
	nuclei quantity	% of non-lysed cells
intercept	0,2713	0,6373
time	0,0992	0,8970
conc	0,001	0,0003
time*time	0,1026	0,8497
conc*time	.	.
conc*conc	0,0026	0,0051

At the same time, microscopic observation of the same samples indicated that very short lysis duration of 1 min also leads to a relatively high proportion of debris (probably remaining cell parts) in respective samples (data not shown). Among all tested conditions, a lysis time of 2 min appeared to be the lowest possible lysis duration that still results in high lysis efficiency as well as acceptable amounts of debris. I therefore concluded that optimal lysis results can be obtained by using lowered cell density in lysis buffer of standard concentration (1x) and a lysis time of 2 min. For details on the optimized isolation protocol see 3.2.2.1 or supplementary material.

Due to constrains in time and given capacities no additional tagmentation test was carried out to validate the improvement in chromatin integrity after shortening the lysis duration from 5 min to 2 min. In this context it should be stated that being confronted with polynucleated states in the case of *Abeoforma*, strongly limits to possibilities for alternative solutions to fall back on. Using a less inclusive method like sc-RNA-sequencing which does not require preceding nuclei isolation but works on a single cell-based level would imply considerable loss of information. This made nuclei isolation become an unavoidable step.
Finally, fragment analysis of the real sn-Multiomics experiment performed on *Abeoforma whisleri* under optimized conditions (2 min lysis duration) indeed confirms an improved tagmentation (see Fig. 9 in section 4.2.1.1 of this thesis).

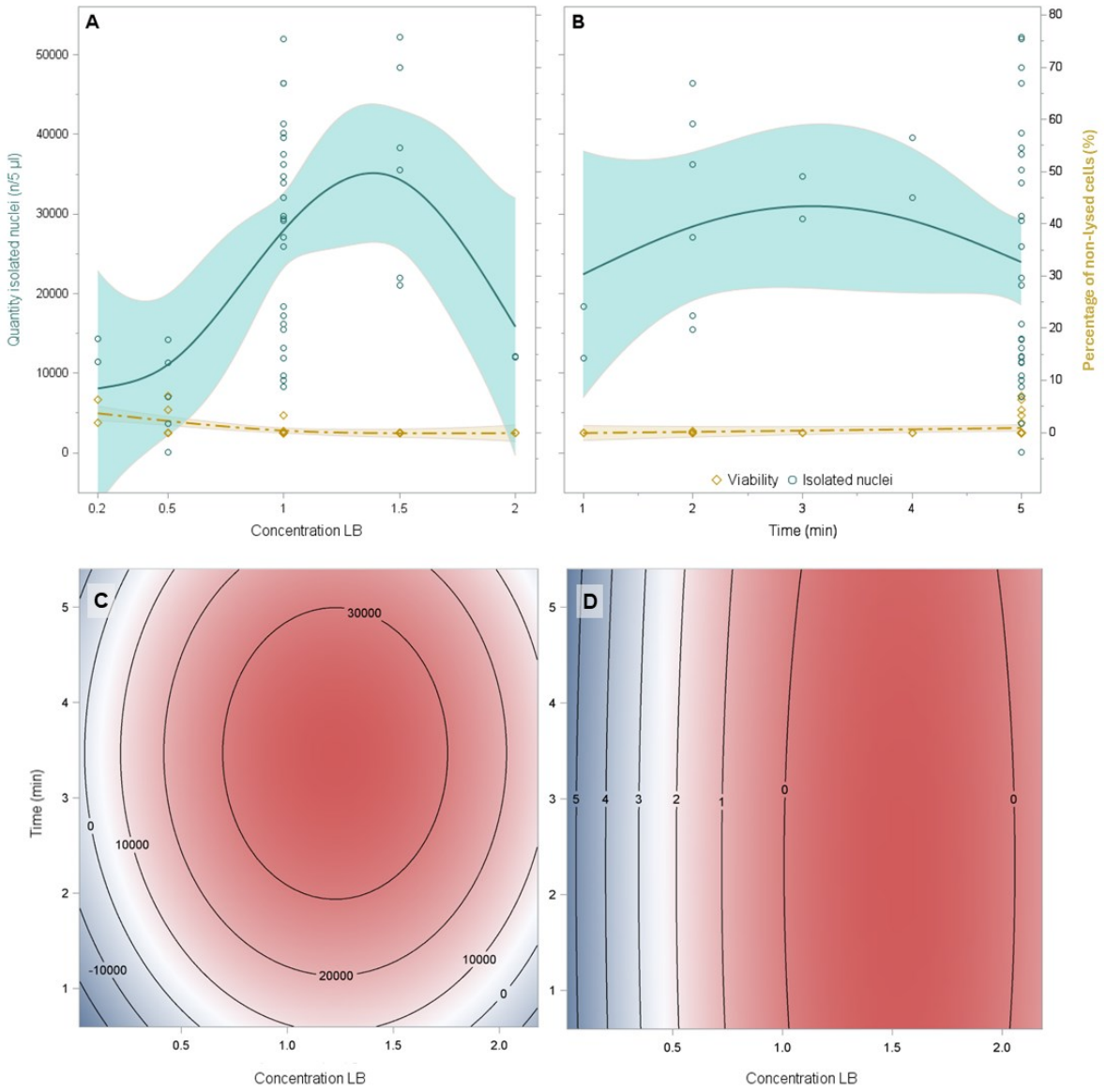

Figure 7) Regression model and response contour plot reconstructing optimal lysis conditions for *Abeoforma* (based on obtained lysis results). **A & B)** Scatter plot including a curvilinear regression plot: data points displayed as yellow diamonds in the scatter plot represent the percentage of non-lysed cells measured for 40 differently treated samples (in %; y-axis labelling on the right-hand side of B); light blue circles display the quantity of isolated nuclei for the same 40 samples (absolute number in 5 µl final samples; y-axis labelling on the left-hand side of A). The dotted yellow line represents the curvilinear regression model fitted into the given dataset for "% of non-lysed cells"; the regression model fitted into the corresponding dataset for "quantity of isolated nuclei" is displayed by the solid blue line (both incl. 95% confidence interval). **A** shows distribution of data points as a function of the used lysis buffer concentration (x-axis, labelled ticks show all tested LB concentrations). Minimal percentage of non-lysed cells and a maximum of isolated nuclei can be achieved by using LB concentrations between 1x and 1,5x. **B** shows distribution of data points as a function of applied lysis time (x-axis, labelled ticks show all tested lysis durations). Highest quantity of isolated nuclei and lowest percentage of non-lysed cells is given at a lysis time of 3 min (BUT: lysis time is NO significant effector – see Supplementary Table 1 & 2). **C & D)** Response contour plot modelling the dependency of "quantity of isolated nuclei" (C) and "percentage of non-lysed cells" (D) on the concentration of used LB (x-axis) and applied lysis time (y-axis). The gradient from red to grey-blue displays a gradient from highest to lowest lysis success. **C)** Response of the variable "quantity of isolated nuclei" to changes in applied lysis time and used LB concentration: red areas indicate lysis conditions associated with high numbers of isolated nuclei, with the black lines describing estimated amounts of nuclei that can be achieved by the respective lysis conditions defining this region. The highest quantity of isolated nuclei can be reconstructed in an area defined by 1-1,5x concentrated LB and 2-5 min lysis time. **D)** Response of the variable "percentage of non-lysed cells" to changes in applied lysis time and used LB concentration: red areas indicate low percentage of non-lysed cells. 0% post-lysis viability (black line labelled "0") can be achieved by applying 1-2x concentrated LB, without any significant effect of lysis time.

4.1.2.2 Optimization process for *Capsaspora* cells

In comparison to what could be observed for *Abeoforma*, fairly different adjustments of the lysis protocol were necessary to obtain satisfactory lysis results for *Capsaspora*. Nuclei isolation using standard lysis conditions on *Capsaspora* cells resulted in 75-90% of non-lysed cells (Figure 10.A), showing that applying default conditions is clearly not sufficient to isolate nuclei efficiently. These results suggested that much harsher lysis conditions are necessary to lyse *Capsaspora* cells.

Accordingly, the first steps of lysis optimization were focused on testing a considerably wide range of lysis buffer concentrations reaching from the standard concentration (1x) to 25-times that concentration (25x). Results of these experiments revealed that increasing the concentration of lysis buffer up to a certain point results in a significant increase in the number of recovered nuclei as well as a strong decrease in the percentage of non-lysed cells (reaching almost 0%, see Figure 8.A). On the other hand, an extreme intensification of lysis conditions using 25x concentrated lysis buffer, resulted in decreasing numbers of recovered nuclei (Figure 8.A). This indicates that raising the lysis buffer concentration over an optimal point leads to the nuclei being strongly overlysed and even degraded. Based on these results, the optimal lysis buffer concentration was determined to be in a range between 14-times (14x) and 18-times (18x) the standard lysis concentration (1x).
In parallel, experiments were performed to test optimal cell density, since this was one substantial adjustment made for *Abeoforma*. Using a cell density of about 250.000 cells per µl lysis buffer showed satisfactory and reproducible results, whereas experiments using lower cell densities indicated a tendency for more inconsistent outcomes and higher cell / nuclei loss (data no shown).

As the results described above suggest, treating cells with a high concentrated lysis buffer bears a great risk of seriously damaging nuclei and decrease chromatin quality. It therefore appears important to remove isolated nuclei from the influence of the lysing reagents as soon as possible, in order to maintain the highest possible integrity of chromatin. Thus, I tested shortened lysis duration to detect the minimal lysis time needed to achieve high lysis success (see Figure 8.B). Results of these experiments revealed that, by using a lysis buffer concentration of 14-18 x, lysing cells for 1 min is apparently sufficient to reach almost 0% of non-lysed cells (Figure 8.D). *Capsaspora* cells seem to get broken up already in the first minute of lysis, showing the major drop in cell viability within this time frame. Also, extending lysis duration over 1 min led to a decrease in the number of recovered nuclei (Figure 8.C), indicating that nuclei get progressively damaged and degrade at longer lysis times.

All these insights suggested that *Capsaspora* cells need comparatively (in relation to *Abeoforma* or other organisms) harsh conditions to be lysed, but breaking up of cells happens

very fast during lysis. These conclusions are supported by statistical analyses that model the dependency of “quantity of isolated nuclei” and “% of non-lysed cells” on “lysis buffer concentration” and “lysis time” (see Table 4 and Figure 8). Both, concentration as well as time, turned out to be statistically significant predictor variables which appear to act independent from each other (interaction not significant). Interestingly, the number of nuclei that could be isolated seems to underly especially strong time influences and weaker (but still significant) effects of the lysis buffer concentration (see Table 4, “nuclei quantity”). While exactly the opposite was found for the number of non-lysed cells, which is mainly influenced by the lysis buffer concentration and less by lysis time (see Table 4, “% of non-lysed cells”). Further analyses showed that among all tested conditions and combinations, the optimal lysis conditions can be specified at a lysis time of 1 min and a lysis buffer concentration around 16x (see Figure 8).

Table 4) Summary of the results for statistical analysis on lysis success for *Capsaspora*; using multiple regression to model dependency of “quantity of isolated nuclei” and “percentage of non-lysed cells” on the two predictor variables “concentration of lysis buffer” (conc) and “lysis time” (time). Significant correlations, showing a *p*-value < 0.05, are marked in light red. Concentration of lysis buffer as well as lysis buffer appear to have a significant influence on nuclei quantity and % of non-lysed cells; both, in a linear and quadratic respect. Interaction between the predictor variable, however, has no significant effect. More detailed results are given in Supplementary Table 3 & 4.

p-values for response surface regression model		
	nuclei quantity	% of non-lysed cells
intercept	< 0,0001	0,0028
time	< 0,0001	0,0379
conc	0,0434	0,0002
time*time	< 0,0001	0,0782
conc*time	0,7658	0,8389
conc*conc	0,0204	0,0005

Another important aspect to investigate was whether distinct cell states of *Capsaspora* show any detectable differences in lysis susceptibility. For example, as the cystic cells are thought to play a role in overcoming harsh conditions, this cell state was suspected to be more robust, potentially also during lysis. Accordingly, the same experimental conditions as described above were applied on separate samples containing either mostly filopodial, mostly cystic or all cell states in varying proportions. For none of the applied conditions a clear difference in lysis efficiency could be discovered between samples of different cell states. This was also confirmed by subsequent statistical analyses carried out on this data. It showed that neither “quantity of isolated nuclei” (*p*-value = 0.7129) nor “% of non-lysed cells” (*p*-value = 0.1522) are significantly dependent on the cell state of used cells (for details see Supplementary Table 9). Therefore, using a lysis duration of 1 min and a lysis buffer concentration between 14x and 18x were determined to be optimal for lysing *Capsaspora* cells independent of used cell states.

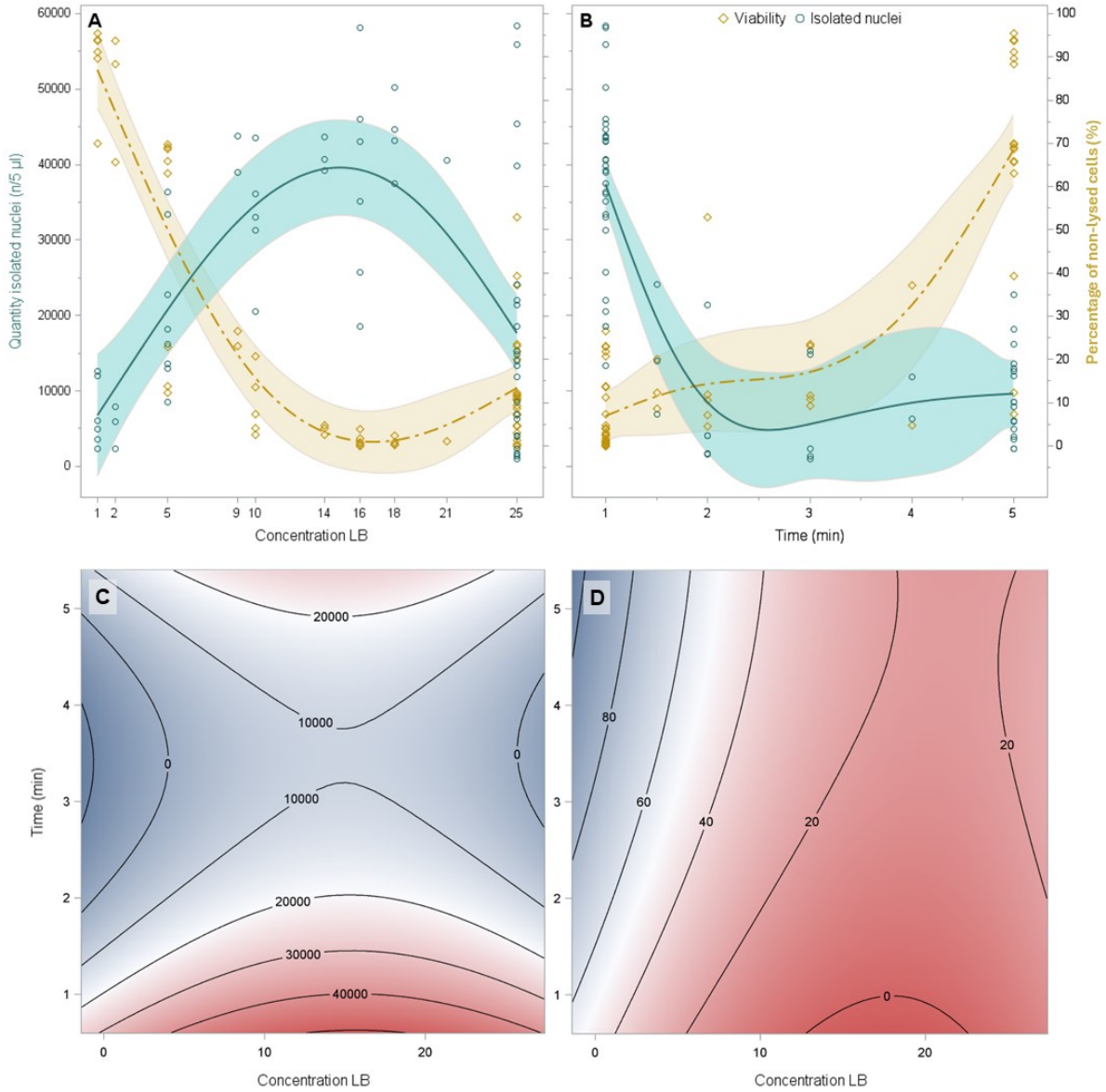

Figure 8) Regression model and response contour plot reconstructing optimal lysis conditions for *Capsaspora* (based on obtained lysis results). **A & B)** Scatter plot including a curvilinear regression plot: data points displayed as yellow diamonds in the scatter plot represent the percentage of non-lysed cells measured for 62 differently treated samples (in %; y-axis labelling on the right-hand side of B); light blue circles display the quantity of isolated nuclei for the same 62 samples (absolute number in 5 µl final samples; y-axis labelling on the left-hand side of A). The dotted yellow line represents the curvilinear regression model fitted into the given dataset for "percentage of non-lysed cells"; the regression model fitted into the corresponding dataset for "quantity of isolated nuclei" is displayed by the solid blue line (both incl. 95% confidence interval). **A** shows distribution of data points as a function of the used lysis buffer concentration (x-axis, labelled ticks show all tested LB concentrations). Minimal percentage of non-lysed cells and a maximum of isolated nuclei could be achieved by using LB concentrations between 14x and 18x. **B** shows distribution of data points as a function of applied lysis time (x-axis, labelled ticks show all tested lysis durations). Highest quantity of isolated nuclei and lowest percentage of non-lysed cells are given at a lysis time of 1 min. **C & D)** Response contour plot modelling the dependency of "quantity of isolated nuclei" (C) and "percentage of non-lysed cells" (D) on the concentration of used LB (x-axis) and applied lysis time (y-axis). The gradient from red to grey blue displays a gradient from highest to lowest lysis success. **C)** Response of the variable "quantity of isolated nuclei" to changes in the applied lysis time and used LB concentration: red areas indicate lysis conditions associated with high numbers of isolated nuclei, with the black lines describing estimated amounts of nuclei that can be achieved by the respective lysis conditions defining this region. The highest quantity of isolated nuclei can be reconstructed in an area defined by <1 min lysis time and 16x LB. **D**) response of the variable "percentage of non-lysed cells" to changes in applied lysis time and used LB concentration: red areas indicate low percentage of non-lysed cells. The black line labelled "0" marks the area where 0% viability can be achieved which corresponds to a lysis time of <1 min and LB concentration between 14x and 22x.

The next important step to specify optimal lysis conditions for *Capsaspora* cells was to carry out a tagmentation test in order to assess chromatin integrity of isolated nuclei. As the size of *Capsaspora* nuclei is very small (< 1 µm) which makes it hardly possible to judge quality of nuclei by their microscopical appearance, this test was especially valuable.

The tagmentation test was carried out on two different samples that varied in the proportions of contained filopodial and cystic cells (mostly filopodial cells (Figure 9.A) or mostly cystic cells (Figure 9.B)) and the concentration of lysis buffer that was used for their treatment (14x or 16x). As it can be seen in Figure 9, fragment size distribution profiles could be obtained for these two samples that show a very similar pattern. This suggests no visible difference in chromatin quality between those samples. Both fragment size distribution plots show clear peaks representing fragments that have been bound by one, two, three or even four nucleosomes (solid red lines, Figure 9). While further notable peaks are visible at about 800 bp, 950 bp and 1200 bp (dotted red lines, Figure 9), the number of fragments having a size shorter than 150 bp seems to be vanishingly small. These results indicated an optimal chromatin integrity of both samples. Importantly, these results also showed that although an unusually high concentration of lysis buffer was needed to lyse *Capsaspora* cells, chromatin damage could be prevented by applying a shortened lysis duration. In conclusion, the above optimization steps allowed me to establish a protocol for the isolation of nuclei of very good quality and satisfactory quantity.

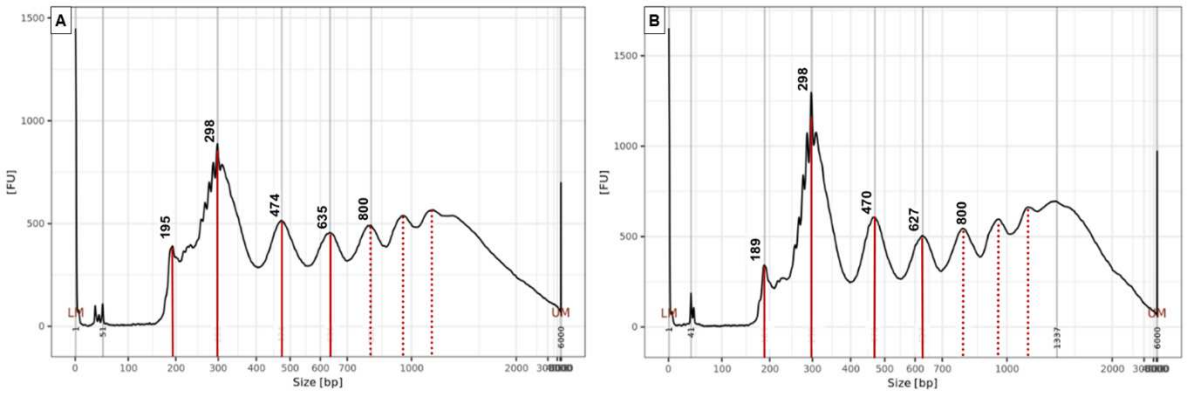

Figure 9) Fragment size distribution plot for two differently treated *Capsaspora* samples: number of fragments (FU, y-axis) is plotted against fragment size (in bp, x-axis). Solid red lines mark peaks corresponding to DNA regions bound by one, two, three or four nucleosomes; the dotted red line indicates peaks at even higher fragment sizes. **A)** Result for sample that contained mostly filopodial cells and was treated with 14x lysis buffer. **B)** Result for sample which contained mostly cystic cells and was treated with 16x lysis buffer.

Since results of the tagmentation test did not show clear differences between the 14x- and the 16x-lysis buffer treated sample and since statistical analysis detected the highest lysis efficiency at a lysis buffer concentration of 16x, this was finally specified as the preferred choice for the optimized lysis protocol.

4.1.3 Quantifying concentration of isolated nuclei

The quality of sn-Multiomics experiments is highly dependent on the number of nuclei that is loaded onto the 10x Genomics Chromium controller (see step 5 in Figure 2). On the one hand, the larger the number of processed and ultimately sequenced nuclei, the more informative a sn-Multiomics dataset can be. On the other hand, loading too many nuclei will result in an increasing percentage of GEMs that capture more than one nucleus, leading to partial loss of the desired single-nucleus character (10x Genomics Inc., 2022b). According to standard protocols, ideally 16.000 – 22.000 nuclei (in 5 µl sample volume) should be used as an input to achieve a targeted sn-Multiomics dataset covering 10.000 nuclei.
Therefore, the last important step towards the final lysis protocol was finding a fast but still accurate way to quantify isolated nuclei. Minimizing the time that it takes to quantify nuclei is essential to maintain them in good quality, since nuclei are very sensitive once they have been isolated (10x Genomics Inc., 2022b).

During the optimization process nuclei quantification was done by manually counting propidium iodide-stained nuclei using a fluorescence microscope. This method was very beneficial for the optimization process, since it allowed detailed observation and documentation of the visual appearance of samples after lysis. Nevertheless, it is really time-consuming and is therefore not suitable for quantifying nuclei that will be used in the real sn-Multiomics experiment. Thus, it became necessary to substitute manual counting by a more time-efficient instrument.
For this purpose, I established a quantification method based on qPCR using the Kapa SYBR Fast Kit (Merck KGaA, Darmstadt, Germany). This technique can deliver highly accurate quantification results within 25-35 minutes which would be an acceptable time span to not impair nuclei quality. However, testing this method on nuclei isolated from *Abeoforma* or *Capsaspora* cells indicated that reproducibility and reliability of obtained results was not high enough for the given purpose.
As an alternative approach, I tested using a Cellometer for nuclei quantification. It has the major advantage of providing results within only a few minutes. Experiments comparing results of manual counting and measurements of the Cellometer showed that the machine is able to deliver sufficiently accurate and reproducible results for isolated *Abeoforma* nuclei, already by using default setting conditions. For *Capsaspora,* settings of the Cellometer needed to be strongly adapted to account for the very small size of *Capsaspora* nuclei (for details see 3.2.2.2). By applying these modified setting conditions, it was finally possible to achieve rather precise quantification of *Capsaspora* nuclei as well. Given those results, using the Cellometer by applying specifically adapted settings was chosen to be the preferred nuclei quantification tool for the real sn-Multiomics experiment in both unicellular organisms.

To summarize, by testing systematic changes in concrete lysis parameters, I was able to develop species-specifically adapted nuclei extraction protocols for both unicellular species, *Abeoforma whisleri* and *Capsaspora owczarzaki*. How optimizing these protocols could improve lysis efficiency is also depicted in Figure 10, comparing results for standard and optimized lysis conditions. Therefore, I could set the stage to finally perform single-nucleus Multiomics-sequencing for the very first time in unicellular holozoans.

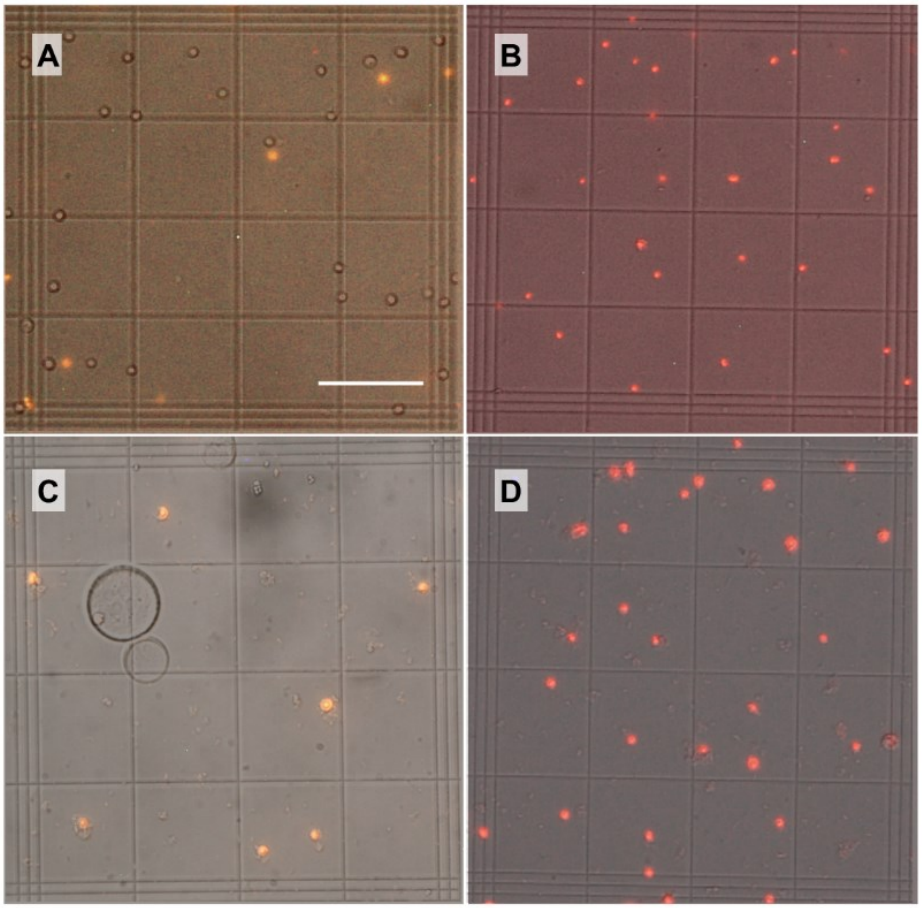

Figure 10) Comparison of lysis results for *Abeoforma* and *Capsaspora* cells, using standard (A, C) vs. optimized protocols (B, D). Nuclei have been stained with PI and are displayed in orange (A, C) or red (B, D). The given scale bar measures 50 µm; 40x magnification. **A)** Lysis of *Capsaspora* cells using 100µl 1x LB and 5 min lysis time. Lots of unbroken cells (not stained) remain. **B)** Lysis of *Capsaspora* cells using 100 µl 16x LB and 1 min lysis time. No unbroken cells are remaining after lysis. **C)** Lysis of *Abeoforma* cell using 100 µl 1x LB and 5 min lysis time. Big blown-up cell structures remain unbroken, while some nuclei could successfully be extracted. **D)** Lysis of *Abeoforma* cells using 200 µl 1x LB and 2 min lysis time. No unbroken cells are remaining after lysis.

4.2. Quality assessment of generated sn-Multiomics data

Finding conditions for optimal cell lysis is one major requirement to be able to realise sn-Multiomics sequencing. However, the informative value of an obtained dataset is strongly determined by its comprehensiveness. The setting for a sn-Multiomics experiment must therefore be designed in a way that the cells used as an input for lysis include as many different cell states and transient constitutions as possible.

For both species, *Abeoforma whisleri* as well as *Capsaspora owczarzaki,* transitioning into all different cell states described above (section 2 of this thesis) happens autonomously in cell cultures and does not require any additional inducing factors. Changes in cell state are naturally regulated by the increase of cell density and the decrease of nutrient availability in aging cell cultures. By thorough observation of developing cultures, I could come up with an experimental set-up for the real sn-Multiomics experiment that enabled me to include not only clearly distinguishable cell states but also all transient states as comprehensively as possible. I decided to use a combination of cells from seven different asynchronized cultures in certain proportions (see 3.1.2 “Cell culture set-up for final sn-Multiomics experiment”). Given this, I was able to gather samples representative for the whole life cycle of the respective species that I could process using the optimised lysis protocol described above. This provides the two most important prerequisites to obtain informative sn-Multiomics data of good quality. Details on the exact experimental design of the final sn-Multiomics experiment can be found in Material and Methods 3.1.2 and 3.2.2 as well as in the supplementary materials.

4.2.1. *Abeoforma whisleri*

4.2.1.1. Quality control of constructed libraries

For the final sn-Multiomics experiment nuclei were extracted out of a heterogenous cell sample by using 200 µl 1 x lysis buffer and a lysis duration of 2 min. As it is depicted in Figure 2, nuclei isolation is just the very first step in the workflow of sn-Multiomics-seq. After tagmentation has happened, GEM generation is the crucial step that enables resolution of information on a single-nucleus level: in microscopically small droplets individual transposed nuclei are combined with single beads that carry certain oligonucleotide sequences. After breaking up the nuclear membrane, the single DNA fragments and mRNA transcripts can bind to these oligonucleotides. This allows tagging each DNA or mRNA sequence with nuclei- or transcript-specific labels (ATAC- or GEX-10x Barcodes + UMIs). Individually marked sequences are pre-amplified and used for ATAC as well as gene expression library construction. Generated ATAC library constructs will include nuclei-specific ATAC-barcodes and all necessary sequencing adaptors; constructs of the gene expression library will cover transcript-specific unique molecular identifiers (UMIs) besides nuclei-specific gene expression barcodes and sequencing

adaptors. At this point the quality of prepared libraries can be checked by analysing the distribution of fragment sizes before sending them for sequencing.

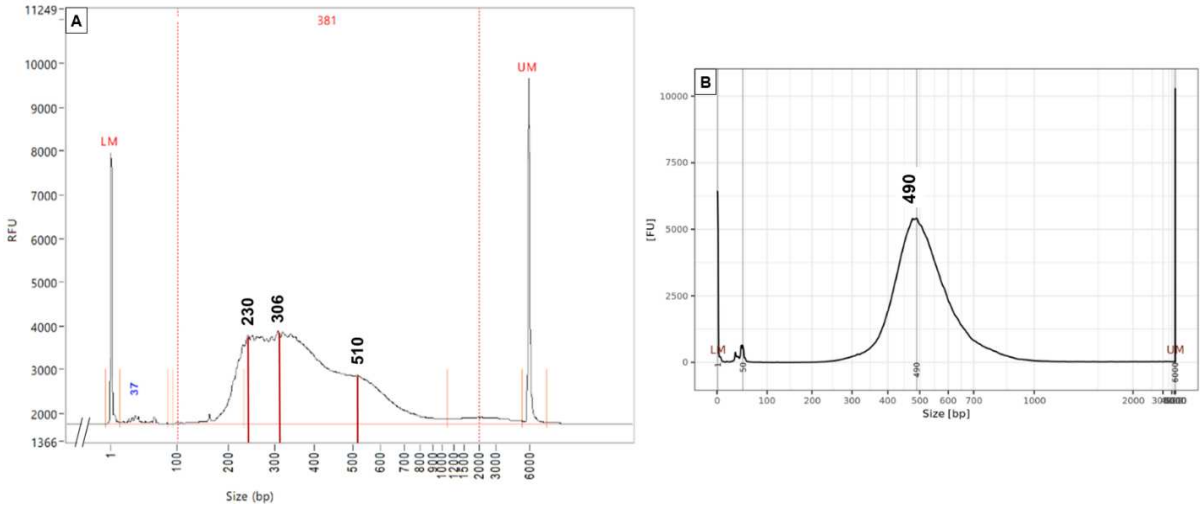

Figure 11) Fragment size distribution plots for *Abeoforma* - after library construction: number of fragments (RFU / FU, y-axis) is plotted against fragment size (in bp, x-axis). **A)** Results for ATAC library: solid red lines mark peaks corresponding to nucleosome-free and mono-nucleosome bound chromatin regions. **B)** Result for gene expression library: a single main peak is visible at 490 bp.

Similar to the fragment size distribution plot obtained for the tagmentation test that was described in section 4.1.2.1, the distribution plot for the ATAC library (Figure 11.A) mirrors the distribution of fragments that have been bound by a certain number of nucleosomes. The given profile for the constructed *Abeoforma* ATAC library shows two compressed, but clearly visible peaks at 230 bp and 306 bp. These peaks represent library constructs of nucleosome-free and mono-nucleosome-bound fragments (solid red lines, Figure 11.A). A rather shallow shoulder can be seen at 510 bp, corresponding to DNA regions that have been bound by two nucleosomes. Even if this profile might not show perfectly explicit and prominent peaks it indicates a clear improvement of chromatin integrity in comparison to the fragment distribution profile obtained after the first tagmentation test (Figure 6). It therefore suggests that the protocol adjustments which were implemented after the tagmentation test (shortening lysis time) indeed helped to obtain increased chromatin integrity.

The ideal distribution profile of the RNA-seq library should show only a single peak, since gene expression library construction includes an additional fragmentation step that aims to achieve more uniform fragment lengths of < 500 µm. This is exactly what we can see for the *Abeoforma* sample (Figure 11.B), indicating that also the gene expression library is of good quality.

Accordingly, the obtained distribution profiles imply that quality of both, the ATAC and gene expression library, seems to be sufficient to deliver good sn-Multiomics sequencing results. Therefore, both libraries were submitted for sequencing.

4.2.1.2. Quality control of sn-Multiomics sequences

The following analyses of obtained sn-Multiomics sequencing data are based on the mapping of reads against a reference genome (sn-ATAC-seq) or reference transcriptome (sn-RNA-seq; the used transcriptome is based on the annotated genome). Therefore, also a good-quality annotated genome is essential to be able to generate meaningful results. A tool to estimate the quality of a genome assembly and annotation completeness is BUSCO (Benchmarking Universal Single-Copy Orthologs; Simão et al., 2015). It compares against a set of single-copy gene orthologs that are conserved between most eukaryotic genomes (here n=255) and distinguishes the query genes according to the four categories “complete and single-copy” (preferably high), “complete and duplicated” (preferably low), “fragmented” (preferably low) and “missing” (preferably low).

Results of the BUSCO analysis for the *Abeoforma* genome used (Table 5) show a very low number of complete genes (53,7%) and a rather high number of missing genes (27,9%). Even though these BUSCO scores indicate considerably low quality of the assembly, it is already a notable improvement compared to previous genomes (~20% complete genes; see Ruiz-Trillo et al., 2023; personal communication, Victoria Shabardina, 2024). For the transcriptome constructed out of the annotated genome, resulting BUSCO scores indicate a significantly better overall quality (Table 5).

Table 5) Results of the BUSCO assessment for the used *Abeoforma* reference genome and transcriptome. Query reads are sorted into four categories: complete single-copy, complete duplicated, fragmented and missing.

Abeoforma	**complete**	**single-copy**	**duplicated**	**fragmented**	**missing**	#Busco markers
Genome	53,7%	53,3%	0,4%	18,4%	27,9%	255
Transcriptome	91,4%	72,2%	19,2%	4,7%	3,9%	255

As the given genome assembly is the best resource available for *Abeoforma whisleri*, it was still used as a reference for sn-Multiomics sequencing data. However, when it comes to interpreting results, one should keep in mind that the preconditions for mapping to the genome are not ideal in this case.

The generated sn-Multiomics sequencing reads were mapped to the reference genome using default settings of the software Cell Ranger ARC provided by the 10x Genomics manufacturers (10x Genomics Inc., 2020b). The next important step towards evaluating the quality of given data is to investigate which of the mapped reads are likely to originate from “real” single nuclei and which come from other sources. As described above (see Figure 2 and 4.2.1.1) the generation of GEMs serves to capture single nuclei in microscopic droplets and tag all contained ATAC fragments and mRNA transcripts with GEM-specific barcodes (ATAC or GEX barcodes). Each mRNA molecule is marked by an additionally transcript-specific UMI sequence. These identifier sequences enable us to trace back every molecule to its nuclear

origin and in return they enable us to estimate the read content of one specific GEM (represented by a certain ATAC as well as GEX barcode). This very important measurement allows Cell Ranger ARC to distinguish between barcodes associated with “real” single nuclei (= nuclei-associated barcodes) and barcodes representing GEMs which most likely captured more than one nucleus (unrealistically high number of reads associated with this barcode) or only background noise (very low number of reads). Cell Ranger ARC identifies only those nuclei as meaningful or “real” that are represented by a gene expression and corresponding ATAC barcode which are both recognized as valid measurements: GEX barcodes are judged as valid if the number of associated UMIs (mRNA molecules) is within a desired range; ATAC barcodes are judges as valid if the number of assigned fragments that overlap with ATAC peaks is high enough (if ATAC reads are not overlapping with peaks they might result from random genome regions rather than real open chromatin regions). This ensures that only nuclei represented by good quality mRNA as well as ATAC signal are selected for further analyses.

For the obtained *Abeoforma* sn-Multiomics dataset, Cell Ranger ARC could predict 11.232 “real” nuclei which is well within the desired target range. Using the set of barcodes associated with these selected nuclei, several controls have been carried out to further investigate the quality of the sn-ATAC-seq and sn-RNA-seq signal.

4.2.1.2.1. sn-ATAC-seq

One important quality control to check the sn-ATAC-seq content of given Multiomics data is generating an insert size distribution plot. Similar to the fragment size distribution plots already discussed before, this plot depicts the count of reads (normalized read density, y-axis) of a certain length (x-axis). In an ideal profile one main peak is visible at <100 bp, representing reads associated with the nucleosome-free chromatin regions. In a periodicity of ~ 150 bp two additional peaks of decreasing size are following which represent reads that originate from mono- and di-nucleosome-bound fragments (10x Genomics Inc., 2020b).
The insert distribution plot obtained for the *Abeoforma* sn-ATAC-seq dataset is displayed in Figure 12. It shows one main peak at < 100 bp representing reads of the nucleosome-free region (marked by the solid red line, Figure 12). Although further peaks cannot be seen as distinctly, two shallow shoulders are visible at expected positions (~ 200 bp and ~400 bp; dotted red lines, Figure 12). Therefore, although this profile clearly differs from an ideal one, it suggests that the level of chromatin integrity might be sufficient to obtain meaningful ATAC signal on open chromatin regions.

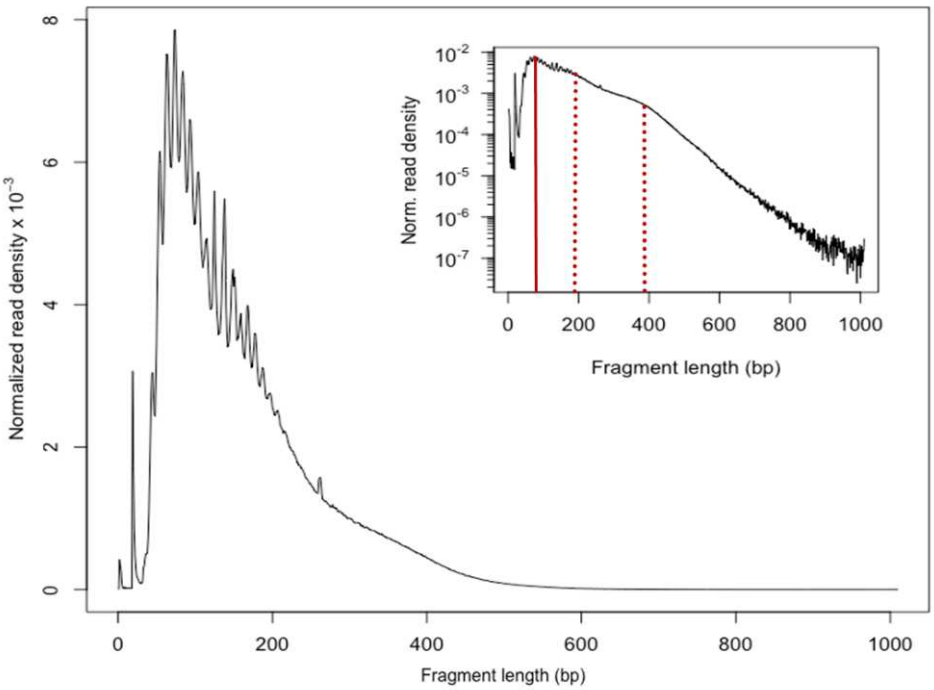

Figure 12) Insert size distribution plot for *Abeoforma* sn-ATAC-seq data: normalized read density (y-axis) is plotted against fragment size (in bp, x-axis). The solid red line marks a peak corresponding to fragments originating from the nucleosome-free region; dotted red lines mark indications of shallow shoulder, representing mono- and di-nucleosome-bound regions.

To get more profound insights into the quality of obtained ATAC data, I additionally used IGV to visually investigate the ATAC signal mapped to the reference genome. Representative examples are shown in Figure 13. As explained above, to generate ATAC data the chromatin is tagmented by an enzyme which specifically cuts at open chromatin regions and simultaneously tags resulting fragments with sequencing adaptors. These fragments are amplified and mapped to the genome. Specific algorithms are used to call ATAC peaks based on the distribution of those reads. The more accessible a particular region is, the more reads will result from this locality and the more prominent the associated peak will be.

For a gene to be actively transcribed, it is necessary that its transcriptional start site (TSS) is accessible to allow binding of the transcriptional machinery. Therefore, we expect a sharp peak in the ATAC profile at TSSs of active genes. This is what can be seen in both examples shown in Figure 13. Distinct peaks are located right around the transcriptional start site of the genes while the ATAC profile completely flattens down along the gene bodies. Such a pattern can also be found in the rest of the genome. This indicates that obtained ATAC data is sufficient to call meaningful peaks allowing reliable interpretations of open chromatin regions and epigenetic constitution.

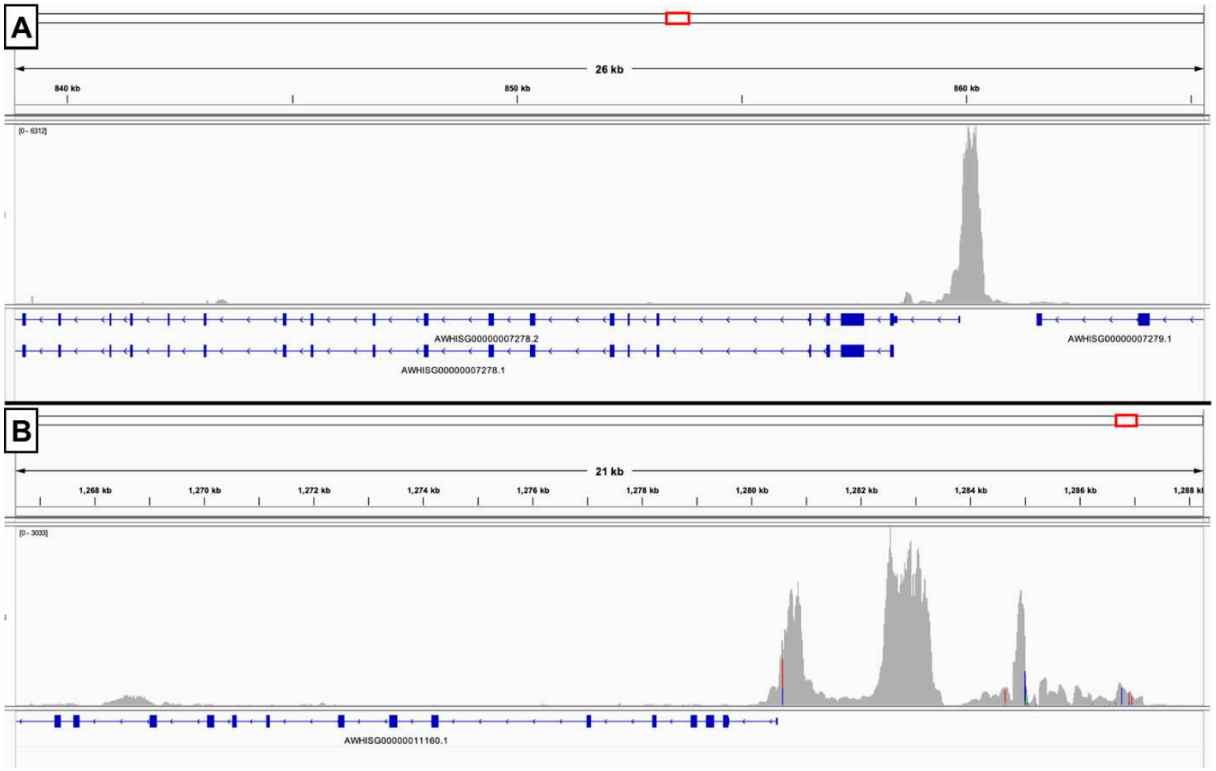

Figure 13) Examples for ATAC peaks of *Abeoforma* sn-Multiomics data. In both cases sharp peaks can be seen at the transcriptional start site of the genes while no higher signal is detectable along the gene body itself (marked in blue underneath the ATAC profile). **A)** Not annotated gene. **B)** ABC transporter gene.

4.2.1.2.2. sn-RNA-seq

In order to evaluate the quality of obtained gene expression data, a barcode rank plot was created. In this plot all barcodes (irrespective of nucleus or non-nucleus assignment) are ranked according to the number of UMIs which are associated with a particular barcode. As already described in 4.2.1.2. the UMI count is an important value to assess whether a barcode can be considered as valid representation of a real nucleus or not. Barcodes associated with GEMs that contained an intact nucleus are likely to be associated with a significantly higher number of UMIs than barcodes of GEMs containing just free-floating mRNA. Ideally, this difference in UMI content can be seen as sharp drop and stereotypical knee-like shape of the curve in the barcode rank plot. This sudden decrease in UMI content enables clear distinction between nuclei-associated barcodes and background noise.

This can be seen also for the barcode rank plot generated for *Abeoforma* sn-RNA-seq data (Figure 14). The area covering nucleus-associated barcodes is highlighted in light blue and includes about 11.000 barcodes which is consistent with the results for the amount of valid nuclei obtained by using the Cell Ranger ARC software.

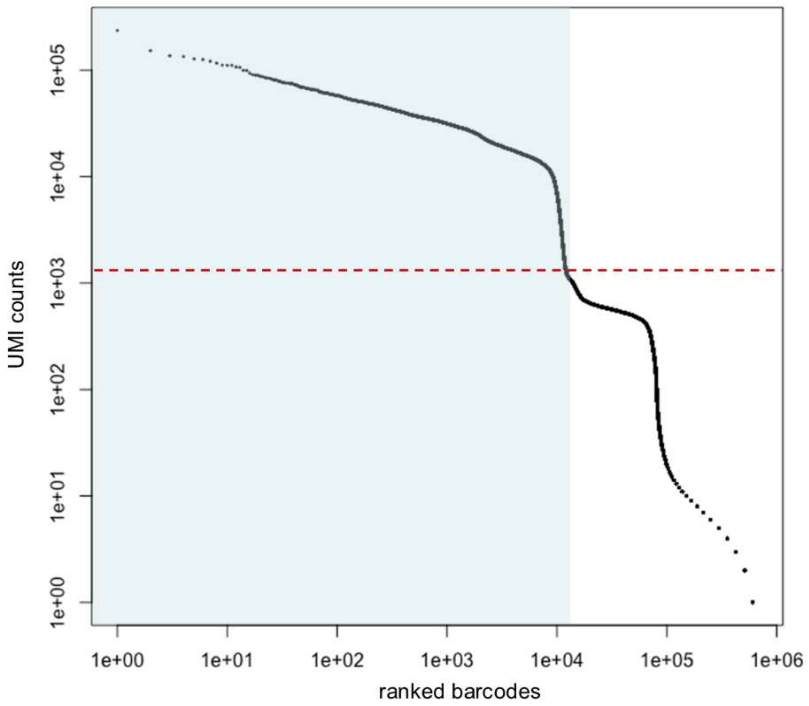

Figure 14) Barcode rank plot for *Abeoforma* sn-RNA-seq data: barcodes (x-axis) are ranked according to the amount of associated UMIs (y-axis). The area highlighted in light blue covers nucleus-associated barcodes; all barcodes with rank > 11.232 (with area) are non-nucleus associated. The dashed red line indicates the UMI count marking the distinction between nucleus- and non-nucleus assignment.

Further analysis using Cell Ranger ARC shows that the great majority of all mRNA sequences can be mapped to the annotated genome (92,3 %) and that the median number of genes per nucleus is 730. Both values are well within expected ranges. This further supports that *Abeoforma* sn-RNA-seq data is of good quality.

4.2.1.2.3. Clustering of obtained sn-Multiomics data

The selected set of nucleus-associated barcodes can further be used to generate single cell clustering where each dot represents one nucleus (Figure 15). This clustering can either be based on the sn-ATAC data or the corresponding sn-RNA data. When performing sn-ATAC-seq clustering, the ATAC profiles of individual nuclei are compared among each other, and nuclei are grouped according to their similarity. Data points sharing very similar ATAC profiles (defined by amount and height of ATAC peaks) are displayed in close proximity to each other, while nuclei with clearly distinct profiles are located further apart. Sn-RNA-seq clustering uses the same approach but assess the similarity between nuclei based on their gene expression profiles (defined by the number of transcripts for each gene expressed). Since we expect all nuclei belonging to the same cell state to show similar epigenetic and transcriptional constitution, we expect them to group together and form clusters representative for specific cell states. The ideal clustering of a heterogeneous sample should show several clearly distinct clusters, corresponding to differentiated cell states, as well as potential interconnecting trajectories formed by intermediate states.

Cluster visualisation inherently requires a drastic reduction of dimensions. Coming from a nucleus-specific ATAC or gene expression profile which is defined by several thousand dimensions (corresponding to the amount of ATAC peaks / amount of genes), this depiction happens in only two. How this dimensional reduction exactly takes place depends on the applied algorithm underlying clustering; in the given case a 2D-t-SNE reduction was used. Independent of the used clustering algorithm, such a reduction unavoidably comes with a certain loss of information which should be kept in mind when interpreting these plots.

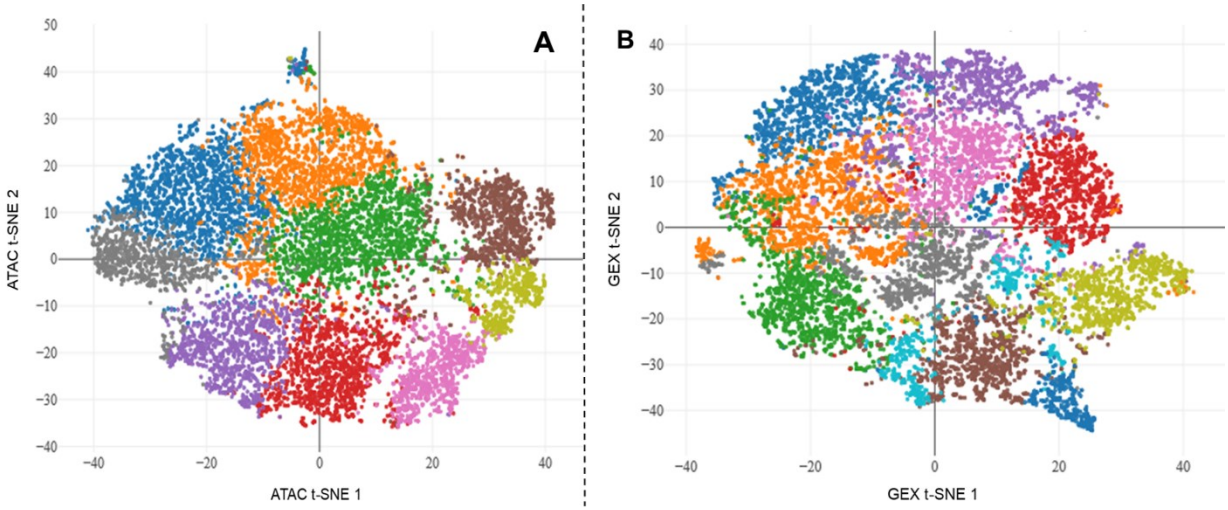

Figure 15) 2D-t-SNE plot for *Abeoforma* sn-Multiomics-seq data. Nuclei-associated barcodes used as data basis were selected by default settings of Cell Ranger ARC, no further processing in terms of cluster improvement was done. Different colours indicate different clusters. **A)** Clustering based on ATAC data, shown in 9 clusters. **B)** Clustering based on gene expression data, shown in 12 clusters.

The given results for clustering of *Abeoforma* sn-Multiomics data show the preliminary and unprocessed output of Cell Ranger ARC (Figure 15). For both, the ATAC- as well as gene expression-based clustering, several distinguishable groups could be detected (shown in different colours). Spatial distance between clusters is relatively small in most cases, creating a rather compressed overall picture. Overlap between clusters seems to be less in the case of ATAC-based clustering. However, without further analysis, the actual quality of clustering can hardly be judged. What can be said is that quality of ATAC as well as gene expression data is sufficient to allow cluster generation, even by applying Cell Ranger ARC default settings.

4.2.2. *Capsaspora owczarzaki*

In accordance with the detailed explanation given in 4.2. one important requirement to generate informative sn-Multiomics data also for *Capsaspora owczarzaki* was again to come up with an experimental design that allows capture of all distinct as well as intermediate cell states. For the final sn-Multiomics experiment nuclei were extracted out of a heterogenous cell sample (combining 7 different cultures) by using 100 µl 16x lysis buffer and a lysis duration of 1 min. After tagmentation, GEM generation and library construction happened, a quality control of created libraries was performed before sending them for sequencing.

4.2.2.1. Quality control of constructed libraries

As it was done for *Abeoforma*, the quality of constructed *Capsaspora* libraries was checked by investigating distribution of fragment sizes. The resulting profile of the ATAC library shows three clear peaks at 226, 316 and 483 bp, representing library constructs that correspond to nucleosome-free, mono- and di-nucleosome-bound chromatin regions (Figure 16.A). In addition, a quite shallow shoulder can be seen at a fragment length of about 630 bp indicating tri-nucleosome-bound regions. In contrast to the distribution profile obtained for the first tagmentation test, there are no other peaks visible at even higher fragment sizes. Still, this profile suggests high quality of the ATAC library and sufficient chromatin integrity to identify accessible chromatin regions.

The distribution profile for the corresponding gene expression library shows the targeted single peak at 485 bp and indicates adequate library quality as well (Figure 16.B). Given that both libraries showed good quality, they were submitted for sequencing.

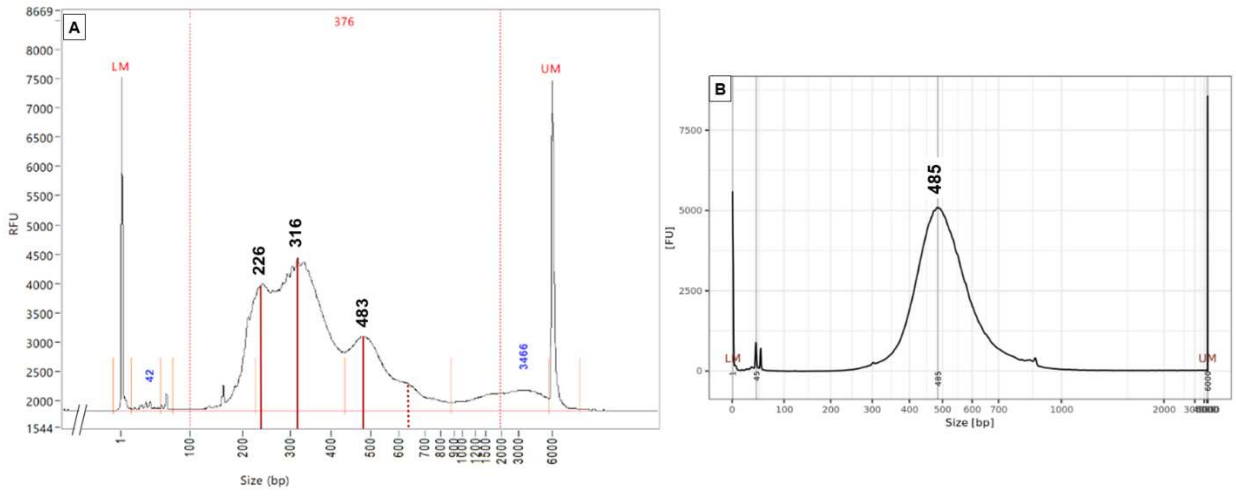

Figure 16) Fragment size distribution plot for *Capsaspora* - after library construction: number of fragments (RFU / FU, y-axis) is plotted against fragment size (in bp, x-axis). **A)** Results for ATAC library: solid red lines mark peaks corresponding to nucleosome-free, mono- and di-nucleosome-bound chromatin regions; the dotted red line indicates tri-nucleosome-bound regions. **B)** Result for gene expression library: a single peak is visible at 485 bp.

4.2.2.2. Quality control of sn-Multiomics sequences

Before further analysis was carried out, the quality of the used reference genome and transcriptome was estimated by applying BUSCO analysis. The results showed a desirably high completeness score for the genome (93,0%) as well as the transcriptome (97,7%) and a notable low percentage of fragmented or missing gene queries (see Table 5). It can therefore be assumed that the annotated genome used as a reference is of high quality.

Table 6) Results of the BUSCO assessment for the used *Capsaspora* reference genome and transcriptome. Query reads are sorted into four categories: complete single-copy, complete duplicated, fragmented and missing.

Capsaspora	**complete**	**single-copy**	**duplicated**	**fragmented**	**missing**	#Busco markers
Genome	93,0%	92,2%	0,8%	4,3%	2,7%	255
Transcriptome	97,7%	96,6%	0,8%	1,2%	1,1%	255

In order to perform quality control of obtained sn-Multiomics data, first the same default settings of the Cell Ranger ARC software were used as for analysis of the *Abeoforma* dataset. However, by using this software on the *Capsaspora* dataset only 190 “real” nuclei were detected. It could thus not be used for meaningful evaluation of the quality of given sn-Multiomics data. Although this could be the result of low-quality sequencing data, it could also be due to standard settings of the Cell Ranger ARC software being incompatible with specific characteristics of the *Capsaspora* dataset.

To test this, an alternative approach was implemented which was processing the sn-RNA- and sn-ATAC-seq data separately from each other by using appropriate alternative software (see Material and Methods, 3.3.1). On the one hand, by applying Cell Ranger ATAC (10x Genomics Inc., 2020a) on the sn-ATAC-seq dataset 11.634 barcodes associated to “real” nuclei could be identified. In comparison to Cell Ranger ARC, Cell Ranger ATAC detects “real” nuclei based on the validity of the assigned ATAC barcode without taking the quality of corresponding sn-RNA-seq signal into account. On the other hand, using the software Cell Ranger (10x Genomics Inc., 2020c) to process the sn-RNA-seq dataset resulted in the identification of 8.337 barcodes associated with “real” nuclei based on their RNA-seq signal (without taking quality of corresponding ATAC signal into account).

Very importantly, subsequent comparison between these two sets of barcodes selected by the separate criteria showed an overlap of 5.485 barcodes. In fact, this indicates that the generated dataset contains 5.485 nuclei showing acceptable quality in terms of their ATAC-seq as well as RNA-seq content.

Using the selected 11.634 barcodes for the ATAC-seq dataset and 8.337 barcodes for the RNA-seq dataset, subsequent analyses were carried out to further assess the quality of obtained sn-ATAC-seq and sn-RNA-seq data.

4.2.2.2.1. sn-ATAC-seq

Analysis of the mapping of sn-ATAC-seq reads onto the *Capsaspora* reference genome, revealed 55 % of non-mapped ATAC read pairs. Given the demonstrated high quality of the used reference genome (see BUSCO results, Table 6, in 4.2.2.2.), this was rather surprising. Superficial investigations of these non-mapped reads could not reveal obvious contamination from other organisms. Instead, they gave indication that a high proportion of these non-mapped reads might be associated with telomeric DNA. Although no in-depth experiments investigating the source of the non-mapped reads were carried out yet, quality control of the sn-ATAC-seq data based on mapped reads (shown below), indicates that the high number of non-mapped reads does not jeopardise the quality of remaining data.

As described above for *Abeoforma*, the first step in ATAC-seq quality control was to generate an insert size distribution plot. The profile obtained for *Capsaspora* shows three clear peaks corresponding to reads originating from nucleosome-free (< 100 bp), mono- and di-nucleosome-bound regions (~180 and 370 bp) (see Figure 17). Three more shallow "shoulders" follow the profile in a periodicity of around 150 to 200 bp where the curve flattens locally before dropping again. These can be referred to reads that originate from chromatin regions which were bound by even higher numbers of nucleosomes. Similar to the results of the tagmentation test and the quality control of the constructed libraries, the obtained distribution curve mirrors a nearly ideal profile and indicates high quality of ATAC fragments.

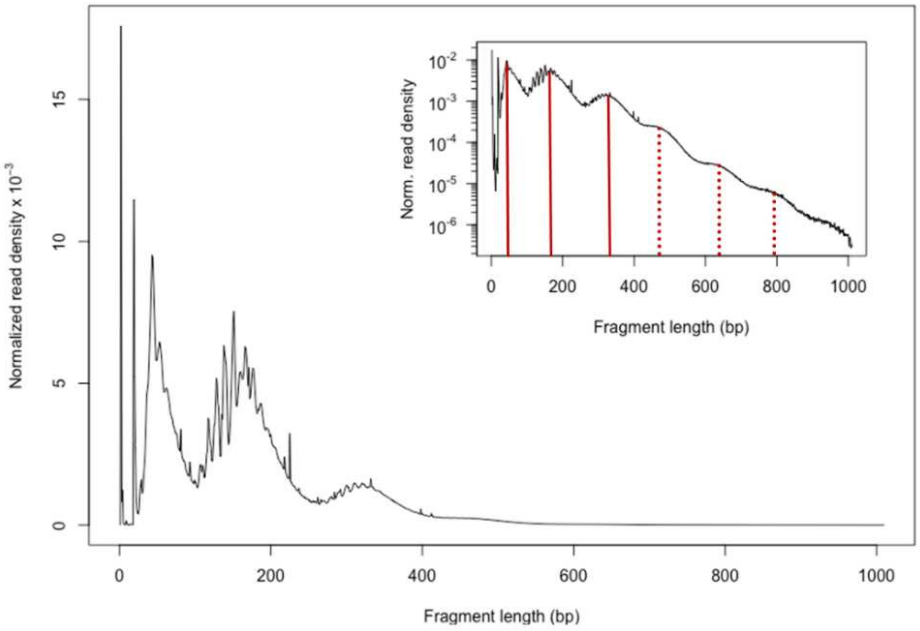

Figure 17) Insert size distribution plot for *Capsaspora* sn-ATAC-seq data: normalized read density (y-axis) is plotted against fragment size (in bp, x-axis). The solid red lines mark peaks corresponding to fragments originating from the nucleosome-free, mono- and di-nucleosome-bound region; dotted red lines mark indications of shallow shoulders, representing regions bound by three, four and five nucleosomes.

Again, also the visual appearance of the ATAC profile was investigated by mapping to the genome and peak calling using IGV. The examples shown in Figure 18 display genes homologous to the ones that were shown for *Abeoforma* before. For both genes clear peaks are visible around their transcriptional start side. In comparison to what we saw for *Abeoforma*, the peaks in Figure 18 appear broader and less sharp. Still, they are clearly distinct from the low ATAC signal along the gene bodies. These widened peaks could be observed all across the ATAC profile. Given the results above, confirming high quality of the sn-ATAC-seq dataset, this diverging appearance of peaks might be due to biological particularities of *Capsaspora*, rather than to insufficient data quality.

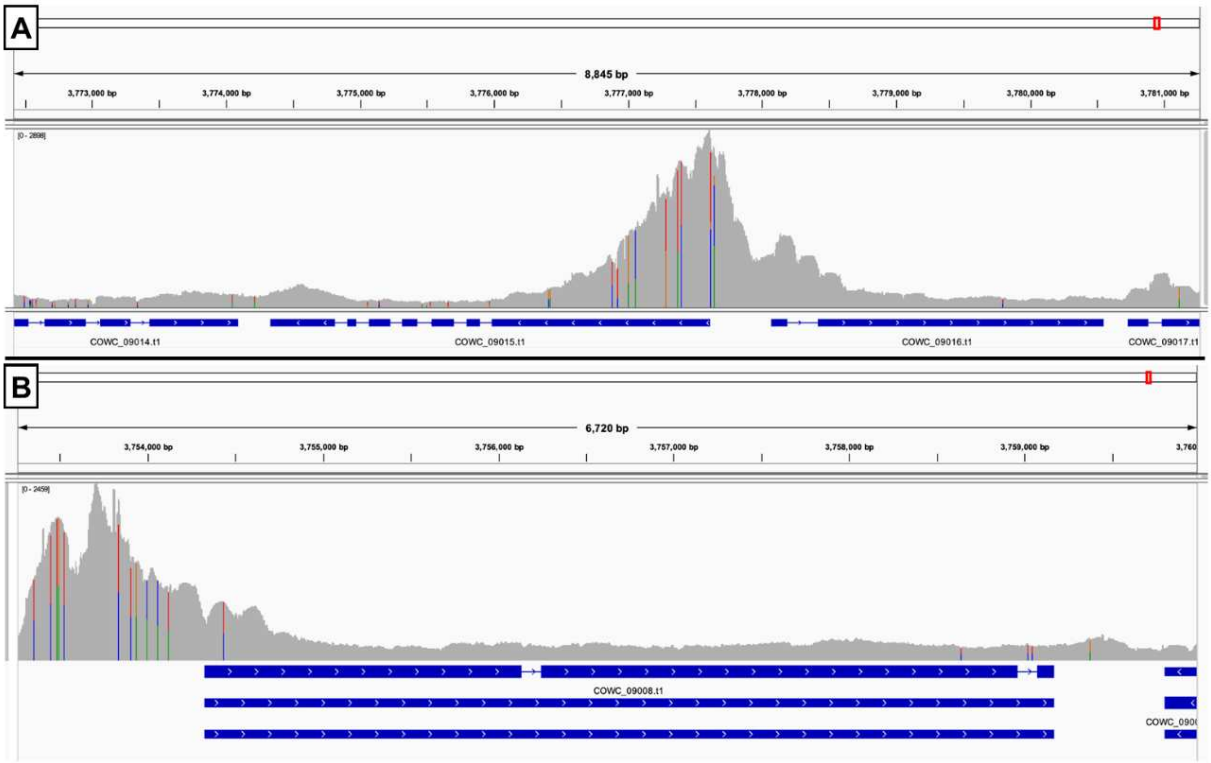

Figure 18) Examples for ATAC peaks of *Capsaspora* sn-Multiomics data. In both cases broadened peaks can be seen at the transcriptional start site of the genes while no high signal is detectable along the gene body itself (marked in blue underneath the ATAC profile). **A)** Not annotated gene. **B)** ABC transporter gene.

4.2.2.2.2 sn-RNA-seq

A barcode rank plot was generated in order to check the quality of obtained sn-RNS-seq data (Figure 19). The resulting curve shows a knee-like shape, even though the drop in UMI counts distinguishing between nucleus- and non-nucleus-assigned barcodes appears less steep compared to the *Abeoforma* sample. Although the discrimination between background noise (non-nucleus associated barcodes) and real signal is less clear in the case of *Capsaspora*, it is still sufficient to distinguish nuclei from noise. Figure 19 confirms the identification of 8.337 gene expression barcodes associated to valid nuclei.

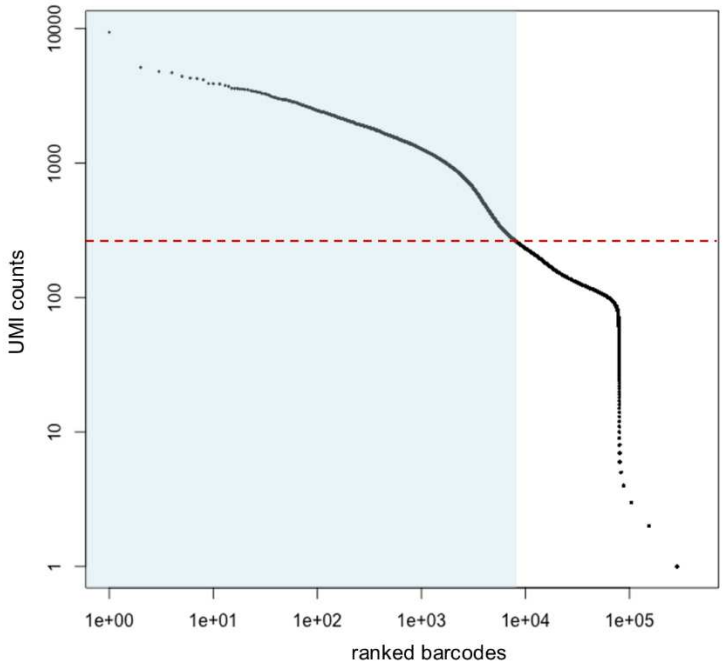

Figure 19) Barcode rank plot for *Capsaspora* sn-RNA-seq data: barcodes (x-axis) are ranked according to the amount of associated UMIs (y-axis). The area highlighted in light blue covers nucleus-associated barcodes; all barcodes with rank > 8.337 (with area) are non-nucleus associated. The dashed red line indicates the UMI count marking the distinction between nucleus- and non-nucleus assignment.

Statistics of the raw Cell Ranger output show that 94.2% of all reads could be mapped to the genome, while the median for genes expressed per nuclei is 315. Both values are within an adequate range and further support good quality of obtained sn-RNA-seq data.

4.2.2.2.3. Clustering of sn-ATAC-seq and sn-RNA-seq datasets

The sets of valid barcodes which were selected for the ATAC and the gene expression dataset were used respectively to generate single nuclei clustering. As the results depicted in Figure 20 show, clustering using the default settings of Cell Ranger (ATAC) was possible for both datasets. Clustering based on gene expression data slightly indicates separation between two main groups and one potentially smaller (pink cluster) group of cells (Figure 20.B). This shape cannot be seen as clearly in the clustering based on ATAC-seq data (Figure 20.A). Similar to the clustering results for A*beoforma*, the overall picture appears rather compressed. However,

these results show that both, the sn-ATAC-seq and sn-RNA-seq data, are of sufficient quality to allow clustering and therefore identification of nuclei showing distinct transcriptomes and epigenetic profiles.

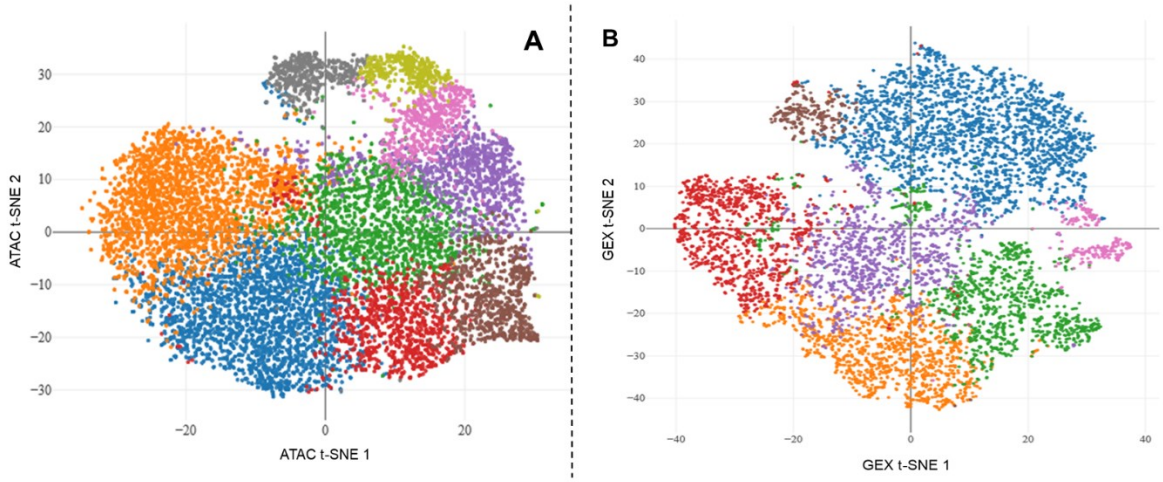

Figure 20) 2D-t-SNE plot for separated *Capsaspora* datasets. Nuclei-associated barcodes used as data basis were selected by default settings of Cell Ranger ATAC (A) or Cell Ranger (B); no further processing in terms of cluster improvement was done. Different colours indicate different clusters. **A)** Clustering based on ATAC data, shown in 9 clusters. **B)** Clustering based on gene expression data, here depiction in 7 clusters was chosen.

5. DISCUSSION

Optimization process

Reviewing all findings gathered during the optimization process, it becomes clear that the two studied unicellular species required fairly different protocol adjustments to be lysed with high efficiency. Thus, premetazoan holozoans are not only very diverse in terms of their morphological potential, life cycle, and occurrence along ecological niches; also, how cells react to particular experimental treatments and how they have to be handled in the lab differs significantly. While *Abeoforma*-specific changes of the standard protocol happened in seemingly small but delicate steps (decreased cell density could improve lysis efficiency sufficiently, without the need to increase lysis buffer concentration), *Capsaspora* cells could only be lysed under unexpectedly harsh conditions (using 16-times higher concentrated lysis buffer). However, in both cases breaking the cellular membrane happened within the very first minute of lysis. This made drastic minimization of lysis time one of the most important aspects to ensure quality of isolated nuclei for both species.

These observations highlight the fundamental importance of doing such optimization work in order to make standardized protocols, established on model cell lines, applicable to a broader range of organisms and research questions. Especially work done on *Capsaspora* showed that intensive protocol adjustment allowed me to implement a technique on a non-model organism which would not have been feasible under standard conditions. Also, knowledge gained during this process might be valuable information to facilitate future work on obtaining sn-Multiomics-seq data on other unicellular holozoan species.

Obtaining first sn-Multiomics datasets for unicellular holozoans

To summarize, I was able to generate sn-Multiomics data for both species, *Abeoforma whisleri* and *Capsaspora owczarzaki*. While data obtained for *Abeoforma* could be analysed using a standard 10x Genomics software that processes sn-ATAC-seq and sn-RNA-seq data in parallel, analyses of the *Capsaspora* dataset required separate bioinformatics processing of the two data modalities. By doing so, sets of over 11.000 (*Abeoforma*) and nearly 5.500 (*Capsaspora*) nuclei could be identified which show high-quality sn-Multiomics signal. To our knowledge, these represent the first datasets delivering information on a single-nucleus resolution level for any unicellular holozoan.

Some insights gathered on the quality of generated data should be mentioned in particular. Firstly, while fragment size distribution analyses of *Abeoforma* ATAC data showed suboptimal profiles, several quality controls indicated very high quality of obtained sn-ATAC-seq data: with over 11.000 identified “real” nuclei, the number of nuclei carrying meaningful ATAC-seq signal is even higher than the targeted nuclei recovery of 10.000 (10x Genomics Inc., 2022b). Also, quality of the data was sufficient to allow clustering and visual inspections of the obtained

ATAC-seq profile showed desirably sharp peaks with low background noise around transcriptional start sites of genes.
A second aspect that should be noted is the high percentage of *Capsaspora* ATAC read pairs that could not be mapped to the genome (55%). Given the demonstrated high completeness of the used reference genome (BUSCO completeness score 93%) and an ideal fragment size distribution profile of the generated sn-ATAC-seq data, this was rather surprising. Very first glimpses into these unmapped reads revealed frequent occurrence of telomeric repetitive sequences. This could explain such extraordinarily high rates of non-mapped reads. Also, it would suggest that *Capsaspora* nuclei showed a high content of accessible telomeric chromatin when the tagmentation reaction happened. Even though, no detailed tests have been carried out so far to explain or confirm this idea, we could not find any evidence that the high rate of non-mapped genes has any negative impact on the quality of existing data.

Besides of that, one particularly important result of this thesis is that generated sn-Multiomics data of both unicellular organisms could be clustered. Even if resulting clusters have not yet been annotated or validated, they suggest the presence of distinguishable transcriptional profiles and chromatin states among analysed nuclei. While previous studies already showed that distinct cell states of *Capsaspora* are associated with differentially employed transcriptional and epigenetic constitutions (Sebé-Pedrós et al., 2013b; 2013c; 2016a; 2016b), this is the first report of its kind for *Abeoforma*.
As described in 4.2., the 2D-t-SNE representation of generated clusters appears very compact and lacks visible trajectories between the barely separated clusters. This stands in contrast to the ideal t-SNE clustering one would aim for when working with single cell data of heterogeneous animal samples: this should show clearly spatially separated clusters that are connected by trajectories formed by intermediate differentiation states. Further analyses (also regarding finding appropriate species-specific adaptations of used clustering parameters) will definitely be necessary to investigate whether this observation bears any biological relevance. It could point to a fundamental difference in the manifestation of cell types in animals and cell states in single-celled holozoans. For example, it is possible that unicellular organisms never reach such distinct and profound gene expression patterns or epigenetic constitutions as we can see in animals (see Sebé-Pedrós et al., 2017). Also, in the special case of *Abeoforma,* we expect many unresolved intermediate states to be covered in the used sample which might form dense interconnections between more distinct cell states. This could additionally contribute to the clustering appear more cohesive overall. The data generated in this study, opens the door to further investigate many of these fundamental questions.

In order to validate and annotate the clustering results it will be necessary to find cell state-specific marker genes that can be tested experimentally. For *Capsaspora*, already published data may facilitate these analyses. For example, results of Sebé-Pedrós at al. (2016) provide

evidence that the *Capsaspora Brachyury* ortholog is potentially associated with the filopodial and aggregative states. In contrast, cluster annotation might be significantly more challenging for *Abeoforma*, since there is neither a detailed description of *Abeoforma* life stages or how they interconnect with each other, nor any report on cell state-specific marker genes. But even more important and illuminating will this data become on a long-term perspective. Sn-Multiomics-seq is able to overcome issues that arise for cell population-based approaches due to *Abeoforma's* complex life cycle and the limited possibilities to generate synchronized culture and separated cell states (Sebé-Pedrós et al., 2017). In addition, bulk-RNA-sequencing experiments have been carried out on samples of different culture states by the Multicellgenome Lab in Barcelona (personal communication Victoria Shabardina, 2024). Combining results of these experiments could not only allow the discovery of potential state-specific marker genes, but also it will help us unravelling the hidden details of the *Abeoforma* life cycle.

Upon cluster annotation, another important step will be to look for distinct cell state transition trajectories. There are particular tools available like the algorithm CytoTRACE (Gulati et al., 2020) which are used to estimate the differentiation state of cells based on the number of expressed genes. Cells expressing the highest number of unique genes are assumed to bear the highest differentiation potential, representing stem cell-like characteristics. Thus, CytoTRACE strongly builds on the biological concept of unidirectional cell differentiation that developed from studies on multicellular animals. To meaningfully apply it on the sn-Multiomics datasets of *Abeoforma* and *Capsaspora* would require that this concept of cell differentiation also holds true for unicellular holozoans, at least partly. Even if we do not expect to find unidirectional cell differentiation processes in the studied unicellular species, CytoTRACE could help us finding cell states with higher differentiation potential und distinguish them from other, more specified states. To what extant such bioinformatical tools can be applied to obtained datasets of unicellular holozoans is yet to be discovered. Nevertheless, they will provide important instruments for initial analyses and even finding out that they are not sufficient to comprehensively describe cell differentiation in unicellular organisms could be a very important discovery to make.

Being able to identify specific cell state transition trajectories will be the basis to trace how transcriptional and epigenetic profiles of cells specifically change along those trajectories. This can be used to identify sets of genes that seem to be correlated in their changes of expression during differentiation pathways. Using further tools like motif enrichment analysis will help to investigate real interactions between genes and provide the first step towards reconstruction gene regulatory modules underlying unicellular cell state transition. Insights potentially gained from such investigations could help answering diverse questions that have arisen in the field. For example, work of Sebé-Pedrós et al. (2017) and Sogabe et al. (2019) suggested that the

function of the animal proto-oncogene *Myc* might be very conserved among holozoans. Finding potential downstream targets of *Myc* as part of reconstructing gene regulatory networks involved in cell state transitions of unicellular holozoans, could help to resolve the details about its ancestral function.

Apart from the described purpose to explore gene regulatory modules for cell differentiation in unicellular holozoans, the herein described dataset can be utilized in many other perspectives to help answering a multitude of different research questions. For example, it could open up a window to investigate potential heterogeneity within clusters, meaning heterogeneity within cell states that cannot be resolved on a morphological level. In the case of *Capsaspora* this will enable us to study the similarity of cells involved in aggregates and therefore find important proof or disproof of the assumption that all cells contributing to these aggregate formations are of the same type (Ruiz-Trillo, et al., 2023). For *Abeoforma* exploring heterogeneity within cell states would be especially interesting in terms of the polynucleated cell state. It is known that coenocytes having multiple nuclei can cellularize and give rise to multiple mononucleated cells of different shapes and behaviours (Marshall & Berbee, 2011; personal communication, Victoria Shabardina). These observations led to the idea that there may be a certain heterogeneity given among the different nuclei of the same coenocytic cell. The presented sn-Mutliomics datasets could help us getting one step closer to uncover these and many similar questions.

6. CONCLUSION - and future perspectives

In the scope of my study, I was able to generate the first sn-Multiomics datasets for unicellular holozoan representatives that resolves epigenetic and transcriptional profiling on the information level of single nuclei. As just discussed, these data will enable us to tackle a multitude of different research questions to stepwise deepen our understanding of cell differentiation in unicellular holozoans and how premetazoan genomic potential could have led to the evolutionary origin of multicellular animals.

As this work is located in the heart of a research field that has considerably grown in size and importance over the last two decades, we suspect this to be basis for many different directions of future research. A multitude of questions is unanswered, and every new insight is accompanied by many new arising questions which are worth investigating.
To do so, it will be necessary to perform sn-Multiomics-sequencing on other unicellular species (like *Corallochytrium limacisporum*) to allow for even more comprehensive and robust comparisons across the phylogenetic tree of holozoans. Any differences as well as similarities that can be discovered in comparison to basally branching metazoans will help us to further specify our idea of the last common unicellular ancestor of all animals. In parallel, research efforts to further explore the hidden diversity of unicellular holozoans will be essential to consolidate phylogenetic relationships and enable reliable data interpretation in an evolutionary context (Arroyo et al., 2018; Arroyo et al., 2020). Also, it will become substantial to experimentally validate results gained from using sn-Multiomics data. For example, further development of stable transfected cell lines for unicellular organisms (see Kożyczkowska et al., 2021 for more detail on already available techniques) will build an important experimental baseline. In this context, also broad-range functional analyses involving tools like genome editing need to be developed (Levin et al., 2014; Kożyczkowska et al., 2021; Ruiz Trillo et al., 2023). This could allow deep insights into ancestral functions of gene families and transcription factors known from present animals (so far only integrin function in *C.owczarzaki* and *Rosetteless* function in *S.rosetta* could be experimentally validated; Levin et al., 2014; Parra-Acero et al., 2020). But the need for functional analyses applies not only to a genetic context but also to a biological one. For example, experiments testing the ability of multicellular constructs for intra-structural signal transduction and reception are necessary to understand in what extent those transient multicellular structures already bear the potential for cell-cell communication similar to principles that we know from multicellular animals.

At the same time also other research disciplines will progress which will successively improve our understanding of the environmental conditions that were present during the evolutionary origin of animals. This could for example help to reconstruct and understand possible benefits of the emergence of an obligatory multicellular lifestyle in the ancient world.

To summarize, future developments in a lot of different research disciplines should be combined to push this fundamental topic of evolutionary biology further forward. The establishment of many different techniques on representatives of unicellular holozoans could lead to an unimaginable gain of knowledge regarding the unicellular ancestry of animals and the transition to multicellular metazoans. However, if we do not want these new findings to overextend our interpretational capacities, it might be inevitable to adopt our own way of thinking: our concepts on the definition of animals, the characteristics of cell differentiation and stem cell potential, but also our idea of cell types themselves, could be truly rewritten and shifted by potential future knowledge – as previous studies of the last two centuries already proofed. All together these findings will add on to a deepened understanding of the unicellular to multicellular transition, but most importantly we must not be afraid of adopting our mindsets to these new insights.

References

Abedin M., King N. (2008) The Premetazoan Ancestry of Cadherins. Science, 319 (5865): 946-948. DOI: 10.1126/science.1151084

Arendt D., Musser JM., Baker CVH. et al. (2016): The origin and evolution of cell types. Nature Reviews Genetics, 17(12): 744-757. DOI: 10.1038/nrg.2016.127

Arroyo AS., Lannes R., Bapteste E. et al. (2020): Gene Similarity Networks Unveil a Potential Novel Unicellular Group Closely Related to Animals from the Tara Oceans Expedition. Genome Biology and Evolution, 12 (9): 1664-1678. DOI: 10.1093/gbe/evaa117

Arroyo AS., López-Escardó D., Kim E. et al. (2018): Novel Diversity of Deeply Branching Holomycota and Unicellular Holozoans Revealed by Metabarcoding in Middle Paraná River, Argentina. Frontiers in Ecology and Evolution, 6: 99. DOI:10.3389/fevo.2018.00099

Belhocine K., DeMare L., Habern O. (2021): Single-Cell Multiomics: Simultaneous Epigenetic and Transcriptional Profiling. Genetic Engineering & Biotechnology News, 41 (1): 66-68. DOI: 10.1089/gen.41.01.17

Bråte J., Neumann RS., Fromm B. et al. (2018) Unicellular Origin of the Animal MicroRNA Machinery. Current Biology, 28 (20): 3288-3295. DOI: 10.1016/j.cub.2018.08.018

Brooke NM., Holland PWH. (2003): The evolution of multicellularity and early animal genomes. Current Opinion in Genetics & Development, 13: 599-603. DOI: 10.1016/j.gde.2003.09.002

Brunet T., King N. (2017): The origin of animal multicellularity and cell differentiation. Developmental Cell Review, 43: 124-140. DOI: 10.1016/j.devcel.2017.09.016

Carr M., Leadbeater BSC., Hassan R. et al. (2008): Molecular phylogeny of choanoflagellates, the sister group to Metazoa. PNAS, 105 (43): 16641-16646. DOI: 10.1073/pnas.080166710

Cavalier-Smith T. (2017): Origin of animal multicellularity: precursors, causes, consequences - the choanoflagellate/sponge transition, neurogenesis and the Cambrian explosion. Philosophical Transactions, The Royal Society, 372: 20150476. DOI: 10.1098/rstb.2015.0476

Cavalier-Smith T., Chao EE-Y. (2003): Phylogeny of Choanozoa, Apusozoa, and Other Protozoa and Early Eukaryote Megaevolution. Journal of Molecular Evolution, 56: 540-563. DOI: 10.1007/s00239-002-2424-z

Davidson EH. (2006): The Regulatory Genome. Gene Regulatory Networks in Development and Evolution. Academic Press. DOI: 10.1016/B978-0-12-088563-3.X5018-4

de Mendoza A., Sebé-Pedrós A. (2019) Origin and evolution of eukaryotic transcription factors. Current Opinion in Genetics & Development, 58-59: 25-32. DOI: 10.1016/j.gde.2019.07.010

de Mendoza A., Sebé-Pedrós A., Ruiz-Trillo I. (2014): The evolution of the GPCR signaling system in eukaryotes: modularity, conservation, and the transition to metazoan multicellularity. Genome Biol Evol., 6 (3): 606-619. DOI: 10.1093/gbe/evu038

de Mendoza A., Sebé-Pedrós A., Šestak MS. et al. (2013): Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. PNAS, 110 (50): E4858-E4866. DOI: 10.1073/pnas.1311818110

de Mendoza A., Suga H., Permanyer J. et al. (2015): Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. eLife, 4: e08904. DOI: 10.7554/eLife.08904

Fairclough SR., Chen Z., Kramer E. et al. (2013): Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. Genome Biology, 14: R15. DOI: 10.1186/gb-2013-14-2-r15

Fairclough SR., Dayel MJ., King N. (2010): Multicellular development in a choanoflagellate. Current Biology, 20 (20): R875-R876. DOI: 10.1016/j.cub.2010.09.014

Glocking SL., Marshall WL., Gleason FH. (2013): Phylogenetic interpretations and ecological potentials of the Mesomycetozoea (Ichthyosporea). Fungal Ecology, 6: 237-247. DOI: 10.1016/j.funeco.2013.03.005

Grau-Bové X., Torruella G., Donachie S. et al. (2017): Dynamics of genomic innovation in the unicellular ancestry of animals. eLife, 6: e26036. DOI: 10.7554/eLife.26036

Grimson A., Srivastava M., Fahey B. et al. (2008): Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature, 455: 1193-1197. DOI: 10.1038/nature07415

Grosberg RK., Strathmann RR. (2007): The Evolution of Multicellularity: A Minor Major Transition? Annual Review in Ecology, Evolution, and Systematics, 38: 621-654. DOI: 10.1146/annurev.ecolsys.36.102403.114735

Gulati GS., Sikandar SS., Wesche DJ. et al. (2020): Single-cell transcriptional diversity is a hallmark of developmental potential. Science, 367 (6476): 405-411. DOI: 10.1126/science.aax0249

Guo J., Tang H., Huang P. et al. (2023): Integrative single-cell RNA and ATAC sequencing reveals that the FOXO1-PRDX2-TNF axis regulates tendinopathy. Frontiers in Immunology, 14: 1092778. Doi: 10.3389/fimmu.2023.1092778

Hehenberger E., Tikhonenkov DV., Kolisko M. et al. (2017): Novel Predators Reshape Holozoan Phylogeny and Reveal the Presence of a Two-Component Signaling System in the Ancestor of Animals. Current Biology, 27 (13): 2043-2050. DOI: 10.1016/j.cub.2017.06.006

Hertel LA., Bayne CJ., Loker ES. (2002): The symbiont *Capsaspora owczarzaki*, nov. gen. nov. sp., isolated from three strains of the pulmonate snail *Biomphalaria glabrata* is related to members of the Mesomycetozoea. International Journal of Parasitology, 32: 1183-1192. DOI: 10.1016/s0020-7519(02)00066-8

Hoffmeyer TT., Burkhardt P. (2016): Choanoflagellate models — *Monosiga brevicollis* and *Salpingoeca rosetta*. Current Opinion in Genetics & Development, 39: 42-47. DOI: 10.1016/j.gde.2016.05.016

Holland PWH., Marlétaz F., Maeso I. et al. (2017): New genes from old: asymmetric divergence of gene duplicates and the evolution of development. Philosophical Transactions, The Royal Society, 372 (1713): 20150480. DOI: 10.1098/rstb.2015.0480

Huo L., Jiao Li J., Chen L. et al. (2019): Single-cell multi-omics sequencing: application trends, COVID-19, data analysis issues and prospects. Briefings in Bioinformatics, 22 (6): 1-19. DOI: 10.1093/bib/bbab229

Kharchenko PV. (2021): The triumphs and limitations of computational methods for scRNA-seq. Nature Methods, 18: 723-732. DOI: 0.1038/s41592-021-01171-x

King N. (2004): The Unicellular Ancestry of Animal Development. Developmental Cell, 7 (3): 313-325. DOI: 10.1016/j.devcel.2004.08.010

King N. (2005): Choanoflagellates. Current Biology, 15 (4): R113-R114. DOI: 10.1016/j.cub.2005.02.004

King N., Hittinger CT., Carroll SB. (2003): Evolution of Key Cell Signaling and Adhesion Protein Families Predates Animal Origins. Science, 301 (5631): 361-363. DOI: 10.1126/science.1083853

King N., Westbrook MJ., Young SL. et al. (2008): The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature, 451: 783-788. DOI: 10.1038/nature06617

Knoll AH. (2011): The Multiple Origins of Complex Multicellularity. Annual Review of Earth and Planetary Sciences, 39: 217-239. DOI: 10.1146/annurev.earth.031208.100209

Kożyczkowska A., Ruiz-Trillo I., Casacuberta E. (2021). Unicellular Relatives of Animals. In: *Handbook of Marine Model Organisms in Experimental Biology* (1st). CRC Press, p.49-66. DOI: 10.1201/9781003217503-3

Lake BB., Codeluppi S., Yung YC. et al. (2017): A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. Scientific Reports, 7: 6031. DOI: 10.1038/s41598-017-04426-w

Lang BF., O'Kelly C., Nerad T. et al. (2002): The closest unicellular relatives of animals. Current Biology, 12: 1773-1778. DOI: 10.1016/S0960-9822(02)01187-9

Laundon D., Larson BT., McDonald K. et al. (2019): The architecture of cell differentiation in choanoflagellates and sponge choanocytes. PLoS Biology, 17 (4): e3000226. DOI: 10.1371/journal.pbio.3000226

Levin TC., Greaney AJ., Wetzel L. et al. (2014): The *rosetteless* gene controls development in the choanoflagellate *S. rosetta*. eLife: e04070. DOI: 10.7554/eLife.04070

Levit GS., Hoßfeld U., Naumann B. et al. (2021): The biogenetic law and the Gastraea theory: from Ernst Haeckel's discoveries to contemporary views. Molecular and Developmental Evolution, 338: 13-27. DOI: 10.1002/jez.b.23039

Ligrone R. (2019a): Chapter 8 – Multicellularity. In: *Biological Innovations that Built the World*. Springer International Publishing. DOI: 10.1007/978-3-030-16057-9

Ligrone R. (2019b): Chapter 10 – The Animals. In: *Biological Innovations that Built the World*. Springer International Publishing. DOI: 10.1007/978-3-030-16057-9

Manning G., Young SL., Miller WT. et al. (2008): The protist, *Monosiga brevicollis*, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. PNAS, 105 (28): 9674-9679. DOI: 10.1073/pnas.0801314105

Marioni JC., Arendt D. (2017): How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology. Annual Review of Cell and Developmental Biology. DOI: 10.1146/annurev-cellbio-100616-060818

Marshall WL., Berbee ML. (2011): Facing Unknowns: Living Cultures (*Pirum gemmata* gen. nov., sp. nov., and *Abeoforma whisleri*, gen. nov., sp. nov.) from Invertebrate Digestive Tracts Represent an Undescribed Clade within the Unicellular Opisthokont Lineage Ichthyosporea (Mesomycetozoea). Protist, 162: 33-57. DOI: 10.1016/j.protis.2010.06.002

Marshall WL., Celio G., McLaughlin DJ. et al. (2008): Multiple isolations of a culturable, motile Ichthyosporean (Mesomycetozoa, Opisthokonta), *Creolimax fragrantissima* n. gen., n. sp., from marine invertebrate digestive tracts. Protist, 159: 415-433. DOI: 10.1016/j.protis.2008.03.003

Medina M., Collins AG., Taylor JW. et al. (2003): Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa. International Journal of Astrobiology, 2 (3): 203-211. DOI: 10.1017/S1473550403001551

Mendoza L., Taylor JW., Ajello L. (2002): The Class Mesomycetozoea: A Heterogeneous Group of Microorganisms at the Animal-Fungal Boundary. Annual Review of Microbiology, 56: 315-344. DOI: 10.1146/annurev.micro.56.012302.160950

Mikhailov KV., Konstantinova AV., Nikitin MA. (2009): The origin of Metazoa: a transition from temporal to spatial cell differentiation. BioEssays, 31: 758-768. DOI: 10.1002/bies.200800214

Naumann B., Burkhardt P. (2019): Spatial Cell Disparity in the Colonial Choanoflagellate *Salpingoeca rosetta*. Frontiers in Cell and Developmental Biology, 7: 231. DOI: 10.3389/fcell.2019.00231

Nichols SA., Roberts BW., Richter DJ. et al. (2012): Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/β-catenin complex. PNAS, 109 (32): 13046-13051. DOI: 10.1073/pnas.1120685109

O'Malley MA., Wideman JG., Ruiz Trillo I. (2016): Losing Complexity: The Role of Simplification in Macroevolution. Trends in Ecology & Evolution, 31 (8): 608-621. DOI: 10.1016/j.tree.2016.04.004

Ocăna-Pallarès E., Williams TA., López-Escardó D. et al. (2022): Divergent genomic trajectories predate the origin of animals and fungi. Nature, 609: 747-753. DOI: 10.1038/s41586-022-05110-4

Owczarzaki A., Stibbs HH., Bayne CJ. (1980): The destruction of *Schistosoma mansoni* mother sporocysts in vitro by amoebae isolated from *Biomphalaria glabrata*: an ultrastructural study. Journal of Invertebrate Pathology, 35 (1): 26-33. DOI: 10.1016/0022-2011(80)90079-8

Oxford Nanopore Technologies (2022, June 8): *Victoria Shabardina: Genome of Abeoforma whisleri helps our understanding of the origin of animals*. [Video]. YouTube. https://www.youtube.com/watch?v=X7VmHqfq-Us

Paps J. (2018): What Makes an Animal? The Molecular Quest for the Origin of the Animal Kingdom. Integrative and Comparative Biology, 58 (4): 654-665. DOI: 10.1093/icb/icy036

Paps J., Holland PWH. (2018): Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nature Communications, 9: 1730. DOI: 10.1038/s41467-018-04136-5

Parra-Acero H., Harcet M., Sánchez-Pons N. et al. (2020): Integrin-Mediated Adhesion in the Unicellular Holozoan *Capsaspora owczarzaki*. Current Biology, 30 (21): 4270-4275. DOI: 10.1016/j.cub.2020.08.015

Phillips JE., Santos M., Konchwala M. et al. (2022): Genome editing in the unicellular holozoan *Capsaspora owczarzaki* suggests a premetazoan role for the Hippo pathway in multicellular morphogenesis. eLIfe, 11: e77598. DOI: 10.7554/eLife.77598

Reinke AW., Baek J., Ashenberg O. et al. (2013): Networks of bZIP Protein-Protein Interactions Diversified Over a Billion Years of Evolution. Science, 340 (6133): 730-734. DOI: 10.1126/science.1233465

Richter D., King N. (2013): The genomic and cellular foundations of animal origins. Annual Reviews of Genetics, 47: 509-537. DOI: 10.1146/annurev-genet-111212-133456

Richter DJ., Fozouni P., Eisen MB. et al. (2018): Gene family innovation, conservation and loss on the animal stem lineage. eLife, 7: e34226. DOI: 10.7554/eLife.34226

Rokas A. (2008): The molecular origins of multicellular transitions. Current Opinion in Genetics & Development, 18 (6): 472-478. DOI: 10.1016/j.gde.2008.09.004

Ros-Rocher N., Pérez-Posada A., Leger MM. et al. (2021): The origin of animals: an ancestral reconstruction of the unicellular-to-multicellular transition. Open Biology, the Royal Society, 11: 200359. DOI: 10.1098/rsob.200359

Ruiz-Trillo I., Burger G., Holland PWH. et al. (2007): The origins of multicellularity: a multi-taxon genome initiative. Trends in Genetics, 23 (3): 113-118. DOI: 10.1016/j.tig.2007.01.005

Ruiz-Trillo I., de Mendoza A. (2020): Towards understanding the origin of animal development. Development, 147 (23): dev192575. DOI: 10.1242/dev.192575

Ruiz-Trillo I., Inagaki Y., Davis LA. et al. (2004): *Capsaspora owczarzaki* is an independent opisthokont lineage. Current Biology, 14 (22): R946-R947. DOI: 10.1016/j.cub.2004.10.037

Ruiz-Trillo I., Kin K., Casacuberta E. (2023): The origin of metazoan multicellularity: a potential microbial black swan event. Annual Review of Microbiology, 77: 499-516. DOI: 10.1146/annurev-micro-03242-120023

Ruiz-Trillo I., Roger AJ., Burger G. et al. (2008): A phylogenomic investigation into the origin of metazoa. Molecular Biology and Evolution, 25 (4): 664-672. DOI: 10.1093/molbev/msn006

Schultheiss KP., Craddock BP., Suga H. et al. (2013): Regulation of Src and Csk Nonreceptor Tyrosine Kinases in the Filasterean *Ministeria vibrans*. Biochemistry, 53 (8): 1320-1329. DOI: 10.1021/bi4016499

Schultz D., Haddock SHD., Bredeson JV. et al. (2023): Ancient gene linkages support ctenophores as sister to other animals. Nature, 618: 110-117. DOI: 10.1038/s41586-023-05936-6

Sebé-Pedrós A., Ariza-Cosano A., Weirach MT. et al. (2013a): Early evolution of the T-box transcription factor family. PNAS, 110 (40): 16050-16055. DOI: 10.1073/pnas.1309748110

Sebé-Pedrós A., Ballaré C., Parra-Acero H. et al. (2016a): The Dynamic Regulatory Genome of *Capsaspora* and the Origin of Animal Multicellularity. Cell, 165 (5): 1224-1237. DOI: 10.1016/j.cell.2016.03.034

Sebé-Pedrós A., Burkhardt P., Sánchez-Pons N. et al. (2013b): Insights into the Origin of Metazoan Filopodia and Microvilli. Molecular Biology and Evolution, 30 (9): 2013-2023. DOI: 10.1093/molbev/mst110

Sebé-Pedrós A., de Mendoza A., Lang BF. et al. (2011): Unexpected repertoire of metazoan transcription factors in the unicellular holozoan Capsaspora owczarzaki. Molecular biology and evolution, 28(3): 1241-1254. DOI: 10.1093/molbev/msq309

Sebé-Pedrós A., Degnan BM., Ruiz-Trillo I (2017): The origin of Metazoa: a unicellular perspective. Nature Reviews Genetics, 18: 498-512. DOI: 10.1038/nrg.2017.21

Sebé-Pedrós A., Grau-Bové X., Richards TA. et al. (2014): Evolution and classification of myosins, a paneukaryotic whole-genome approach. Genome Biol Evolution, 6 (2): 290-305. DOI: 10.1093/gbe/evu013

Sebé-Pedrós A., Irimia M., del Campo J. et al. (2013c): Regulated aggregative multicellularity in a close unicellular relative of metazoa. eLife, 2: e01287. DOI: 10.7554/eLife.01287

Sebé-Pedrós A., Roger AJ., Lang FB. et al. (2010): Ancient origin of the integrin-mediated adhesion and signaling machinery. PNAS, 107 (22): 10142-10147. DOI: 10.1073/pnas.1002257107

Sebé-Pedrós A., Zheng Y., Ruiz-Trillo I. et al. (2012): Premetazoan origin of the hippo signaling pathway. Cell Report, 1 (1): 13-20. DOI: 10.1016/j.celrep.2011.11.004

Shabardina V., Kashima Y., Suzuki Y. et al. (2020): Emergence and Evolution of ERM Proteins and Merlin in Metazoans. Genome Biology and Evolution, 12 (1): 3710-3724. DOI: 10.1093/gbe/evz265

Shalchian-Tabrizi K., Minge MA., Espelund M. et al. (2008): Multigene Phylogeny of Choanozoa and the Origin of Animals. PLoS ONE, 3 (5): e2098. DOI: 10.1371/journal.pone.0002098

Simão FA., Waterhouse RM., Ioannidis P. et al. (2015): BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31 (19): 3210-3212. DOI: 10.1093/bioinformatics/btv351

Sogabe S., Hatleberg WL., Kocot KM. et al. (2019): Pluripotency and the origin of animal multicellularity. Nature, 570: 519-522. DOI: 10.1038/s41586-019-1290-4

Spitz F., Furlong EEM. (2012): Transcription factors: from enhancer binding to developmental control. Nature Reviews Genetics, 13: 613-626. DOI: 10.1038/nrg3207

Stibbs HH., Owczarzak A., Bayne CJ. et al. (1979): Schistosome sporocyst-killing Amoebae isolated from *Biomphalaria glabrata*. Journal of Invertebrate Pathology, 33: 159-170. DOI: 10.1016/0022-2011(79)90149-6

Suga H., Ruiz-Trillo I. (2013): Development of ichthyosporeans sheds light on the origin of metazoan multicellularity. Developmental biology, 377(1): 284-292. DOI: 10.1016/j.ydbio.2013.01.009

Suga H., Torruella G., Burger G. et al. (2014): Earliest Holozoan expansion of phosphotyrosine signaling. Molecular Biology and Evolution, 31 (3): 517-528. DOI: 10.1093/molbev/mst241

Tikhonenkov DV., Hehenberger E., Esaulov AS. et al. (2020): Insights into the origin of metazoan multicellularity from predatory unicellular relatives of animals. BMC Biology, 18: 39. DOI: 10.1186/s12915-020-0762-1

Tirosh I., Venteicher AS., Hebert C. et al. (2016): Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature, 539 (7628): 309-313. DOI: 10.1038/nature20123

Torruella G., de Mendoza A., Grau-Bové X. et al. (2015): Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. Current Biology, 25 (18): 2404-2410. DOI: 10.1016/j.cub.2015.07.053

Torruella G., Derelle R., Paps J. et al. (2011): Phylogenetic Relationships within the Opisthokonta Based on Phylogenomic Analyses of Conserved Single-Copy Protein Domains. Molecular Biology and Evolution, 29 (2): 531-544. DOI: 10.1093/molbev/msr185

True JR., Carroll SB. (2002): Gene co-option in physiological and morphological evolution. Annual Review of Cell and Developmental Biology, 18: 53-80. DOI: 10.1146/annurev.cellbio.18.020402.140619

Villani A-C., Satija R., Reynolds G. et al. (2017): Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science, 356 (6335): eaah4573. DOI: 10.1126/science.aah4573

Weidemüller P., Kholmatov M., Petsalaki E. et al. (2021): Transcription factors: Bridge between cell signaling and gene regulation. Proteomics, 21: 23-24. DOI: 10.1002/pmic.202000034

Whisleri HC. (1960): Pure Culture of the Trichomycete, *Amoebidium parasiticum*. Nature, 186: 732-733.

Wu H., Kirita Y., Donnelly E. et al. (2019): Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. Journal of the American Society of Nephrology, 30 (1): 23-32. DOI: 10.1681/ASN.2018090912

Yan F., Powell DR., Curtis DJ. et al. (2020): From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biology, 21: 22. DOI: 10.1186/s13059-020-1929-3

Young SL., Diolaiti D., Conacci-Sorrell M. et al. (2011): Premetazoan ancestry of the Myc-Max network. Molecular Biology and Evolution, 28 (10): 2961-2971. DOI: 10.1093/molbev/msr132

Used protocols and manuals of 10x Genomics:

10x Genomics, Inc. (2020a): Technical Note CG000329 – Interpreting Cell Ranger Web Summary Files for Single Cell Gene Expression Assay.

10x Genomics, Inc. (2020b): Technical Note CG000368 – Interpreting Cell Ranger ARC Web Summary Files for Single Cell Multiome ATAC + Gene Expression Assay.

10x Genomics, Inc. (2020c): Technical Note CG000202 – Interpreting Cell Ranger ATAC Web Summary Files for Single Cell ATAC Assay.

10x Genomics, Inc. (2022a): Demonstrated Protocol CG000169 – Nuclei Isolation for Single Cell ATAC Sequencing.

10x Genomics, Inc. (2022b): User Guide CG000338 Rev F – Chromium Next GEM Single Cell Multiome ATAC + Gene Expression.

Supplementary Material

S1. Growth media for cell culture

***Abeoforma whisleri*:**

Marine Broth medium 2216 (Ref. 279110, BD-Difco)

Prepared medium according to the manufacturer's instructions (suspend 37,4 g in 1 L of Milli-Q water; boil shortly to dissolve powder completely), filter it using a 0.22 µm filter and autoclave for 15 min at 121°C. Store the medium at 4°C.

***Capsaspora owczarzaki*:**

ATCC medium 1034 (modified PYNFH medium)

Suspend 10 g Peptone, 10 g Yeast extract, 1 g Yeast nucleic acid (Ribonucleic Acid, type VI from Torula yeast), 15 mg Folic Acid in 880 Milli-Q water. Autoclave the mixture for 15 min at 121°C. After letting it cool down, add 400 µl of Hemin stock solution[1], 20 ml Buffer solution[2] and 100 ml of heat-inactivated Fetal Bovine Serum. Filter-sterilise it using a 0,22 µm filter and store the final medium at 4°C.

[1] **Hemin stock solution (for 50 ml)**: dissolve 100 mg NaOH in 50 ml Milli-Q water and add 125 mg Hemin. Autoclave it for 20 min at 121°C. Store it in 400 µl aliquots at 4°C.

[2] **Buffer solution (for 1 l)**: dissolve 18.1 g KH_2PO_4 and 25 g Na_2HPO_4 in 1 l Milli-Q water. Adjust the pH value to 6.5 using HCl 37%. After filtering it using a 0,22 µm filter, it can be stored at 4°C.

Reagent	Vendor	CAS-NO (product number):
Peptone	Carl Roth	8013-01-2 (2904.3)
Yeast extract	Carl Roth	91079-46-8 (2365.4)
Ribonucleic acid from torula yeast – type VI	Sigma-Aldrich	63231-63-0 (R6625-100G)
Folic Acid	Sigma-Aldrich	59-30-3 (F8758)
Fetal Bovine Serum	Sigma-Aldrich	(F9665-100ML)
Hemin	Sigma-Aldrich	16009-13-5 (H9039-1G)
KH_2PO_4	Merck	7778-77-0 (1.04873.1000)
Na_2HPO_4	Carl Roth	13472-35-0 (T879.2)

S2. Reagents for nuclei extraction

Wash Buffer *	stock conc.	final conc.	needed vol. (for 25 ml WB)
Tris-HCl (pH 7.4)	1 M	0,01 M	250 µl
NaCl	5 M	0,01 M	50 µl
$MgCl_2$	1 M	0,003 M	75 µl
BSA	10 %	1 %	2.500 µl
Tween-20	10 %	1 %	250 µl
DTT **	1 M	0,001 M	25 µl
RNase inhibitor **	40 U/µl	1 U/µl	625 µl
Nuclease-free water			21.225 µl

* Wash Buffer without DTT and RNase inhibitors can be stored at 4°C for > 1 month.
** DTT and RNase inhibitors should be added to the needed amount of WB right before usage: for one Capsaspora + one Abeoforma sample, 2.700 µl final WB are needed. 2.760 µl ml final WB = 2.688 µl WB stock + 69 µl RNase inhibitor + 2,76 µl DTT.

16x Lysis Buffer *** **used for *C. owczarzaki***	stock conc.	final conc.	needed vol. (for 100 µl LB-C)
final Wash buffer			88,8 µl
NP40-Alternative	20 %	1,6 %	8 µl
Digitonin	5 %	0,16 %	3,2 µl

1x Lysis Buffer *** **used for *A. whisleri***	stock conc.	final conc.	needed vol. (for 200 µl LB-A)
final Wash buffer			194 µl
NP40-Alternative	10 %	0,1 %	2 µl
Digitonin	0,5 %	0,01 %	4 µl

Nuclei Buffer solution ***	stock conc.	final conc.	needed vol. (for 300 µl NBS)
20X nuclei buffer	20X	1X	15 µl
BSA	10 %	0,1 %	30 µl
DTT	1 M	0,001 M	0,3 µl
RNase inhibitors	40 U/µl	1 U/µl	7,5 µl
Nuclease-free water			247,2 µl

*** Lysis Buffers and Nuclei buffer solution should be prepared freshly.

Reagent	Vendor	CAS-NO (product number):
Tris-HCl	Sigma-Aldrich	1185-53-1 (T3253-1KG)
NaCl	Sigma-Aldrich	7647-14-5 (31434-1KG-M)
$MgCl_2$	Carl Roth	7786-30-3 (KK36.2)
Bovine Serum Albumin	Sigma-Aldrich	9048-46-8 (A7906-500G)
Tween-20	Sigma-Aldrich	9005-64-5 (P9416-50ML)
DTT	Sigma-Aldrich	3483-12-3 (43816-10ML)
NP40-Alternative	Calbiochem	9016-45-9 (492016-100ML)
Digitonin 5%	Invitrogen, Thermo Fisher Scientific	11024-24-1 (BN2006-1ML)
20X nuclei buffer	10x Genomics (supplied in Chromium Next GEM sc-Multiome ATAC Kit)	2000153/2000207
RNase inhibitors 40 U/µl	TaKaRa	2313A

S3. Statistical evaluation of lysis efficiency

S3.1. *Abeoforma whisleri*

Supplementary Table 1] Summary of results for multiple regression (using Response Surface method), modelling the dependency of number of isolated nuclei on the two ordinal categorical predictor variables "concentration of lysis buffer" and "lysis time" in *Abeoforma whisleri*. For this analysis a square root transformed dataset of n = 40 was used, including replicates out of 8 different experiments: each of these replicates was assigned to one out of five different categories within the variable "concentration" (0.2x, 0.5x, 1x, 1.5x, 2x) and to one out of five categories within the variable "time" (5min, 4min, 3min, 2min, 1min).
All significant *p*-values are highlighted in red (α = 0.05): R^2 being 0.3466 shows that the fit to the regression model is not very high, and testing the lack of fit gives a result that is only just insignificant. For the regression model only the variable "concentration" (conc) and its square value clearly have significant effects, while the variable "time" (time, time²) and the interaction between the two predictor variables seem to not have any significant effect.

Dependency of quantity of isolated nuclei on LB on concentration and lysis time

Response Surface regression model

fit to regression model	R^2 = 0.3466		source	estimate	*p* (t-test)
lack of fit	*p* = 0.062		intercept	- 80.63	0.2713
type of regression	R^2	*p* (F-test)	time	76.35	0.0992
linear	0.0726	0.1583	conc	233.30	0.001
quadratic	0.274	0.0022	time*time	- 11.22	0.1026
crossproduct	.	.	conc*time	.	.
total model	0.3466	0.0041	conc*conc	- 92.32	0.0026

Supplementary Table 2] Summary of results for multiple regression (using Response Surface method), modelling the dependency of viability on the two ordinal categorical predictor variables "concentration of lysis buffer" and "lysis time" for *Abeoforma whisleri*. For this analysis an inverse square root transformed dataset of n = 40 was used, including replicates out of 8 different experiments: each of these replicates was assigned to one out of five different classes within the variable "concentration" (0.2x, 0.5x, 1x, 1.5x, 2x) and to one out of five classes within the variable "time" (5min, 4min, 3min, 2min, 1min). It is important to take into account that the viability values obtained from this dataset did not follow a normal distribution: this is most likely due to a very high number of "0"-observations present in the data, leading to a distortion of the data distribution. Results of this analysis therefore must be interpreted critically.
All significant *p*-values are highlighted in red (α = 0.05). With R^2 = 0.3907 and a highly unsignificant "lack of fit" the fitting of the dataset to the regression model is slightly better, compared to the results display in Summary Table 1. In the regression model the variable "concentration" shows a highly significant effect as both, linear as well as quadratic component. "Time" (time, time²) and the interaction between "concentration" and "time", in contrast, seem to not have any significant effect.

Dependency of % of non-lysed cells on concentration and lysis time

Response Surface regression model

fit to regression model	R^2 = 0.3907		source	estimate	*p* (t-test)
lack of fit	*p* = 0.6989		intercept	- 23.47	0.6373
type of regression	R^2	*p* (F-test)	time	- 4.02	0.8970
linear	0.2352	0.0033	conc	176.50	0.0003
quadratic	0.1555	0.0187	time*time	0.87	0.8497
crossproduct	.	.	conc*time	.	.
total model	0.3907	0.0013	conc*conc	- 58.25	0.0051

S3.2. *Capsaspora owczarzaki*

Supplementary Table 3] Results for analysis of variances and means in *Capsaspora owczarzaki*, testing the dependency of the number of isolated nuclei and viability on the cell state of used cells (handled as nominal categorical variable). For this analysis a reduced dataset of n = 33 was used, that has been square root transformed to follow normal distribution required for ANOVA. The dataset includes replicates of 5 different experiments, only covering samples for which either only filopodial or cystic cells were used (assigned to either the category "filopodial" or "cystic" within the variable "cell state") – since mixed samples contained cells of both states in varying proportions. The analysis of variances (ANOVA) shows that neither for quantity of isolated nuclei, nor for viability of cells after lysis a significant difference could be detected between "filopodial" and "cystic" sample (with α = 0.05). The same insignificance also results for the following post hoc t-test using Tukey-Kramer adjustment, comparing the means.

Dependency of quantity of isolated nuclei on cell state			
ANOVA		post-hoc t-test	
χ^2-test	p = 0.7103	t-test	p = 0.7129
Fisher's	p = 0.7129		

Dependency of % of non-lysed cells on cell state			
ANOVA		post-hoc t-test	
χ^2-test	p = 0.1418	t-test	p = 0.1522
Fisher's	p = 0.1522		

Supplementary Table 4] Summary of results for multiple regression (using Response Surface method), modelling the dependency of number of isolated nuclei on the two ordinal categorical predictor variables "concentration of lysis buffer" and "lysis time" in *Capsaspora owczarzaki*. For this analysis a square root transformed dataset of n = 62 was used, including replicates out of 14 different experiments: each of these replicates was assigned to one out of nine different categories within the variable "concentration" (1x, 2x, 5x, 10x, 14x, 16x, 18x, 21x, 25x) and to one out of five categories within the variable "time" (5min, 4min, 3min, 2min, 1min).
All *p*-values showing significance are highlighted in red (α = 0.05). With R^2 = 0.7201 the fit to the calculated regression model is clearly better than what was obtained for *Abeoforma*, which might be due to a bigger underlying dataset. Apart from the interaction between the two predictor variables all sources seem to have a significant effect on the number of isolated nuclei. "Time" as linear component shows the highest significance with a *p*-value smaller than 0.0001, while "concentration" as linear component seems to have the lowest, still significant effect (p = 0.0434).

Dependency of quantity of isolated nuclei on LB on concentration and lysis time

Response Surface regression model

fit to regression model	R^2 = 0.7201
lack of fit	p = 0.2966

type of regression	R^2	p (F-test)
linear	0.5616	< 0.0001
quadratic	0.1581	< 0.0001
crossproduct	0.0004	0.7658
total model	0.7201	< 0.0001

source	estimate	p (t-test)
intercept	239.42	< 0.0001
time	- 130.67	< 0.0001
conc	9.76	0.0434
time*time	19.45	< 0.0001
conc*time	- 0.12	0.7658
conc*conc	- 0.32	0.0204

Supplementary Table 5] Summary of results for multiple regression (using Response Surface method)**,** modelling the dependency of viability after lysis on the two ordinal categorical predictor variables “concentration of lysis buffer” and “lysis time” in *Capsaspora owczarzaki*. For this analysis a box-cox transformed dataset of n = 62 was used, including replicates out of 14 different experiments: each of these replicates was assigned to one out of nine different categories within the variable “concentration” (1x, 2x, 5x, 10x, 14x, 16x, 18x, 21x, 25x) and to one out of five categories within the variable “time” (5min, 4min, 3min, 2min, 1min).
All *p*-values showing significance are highlighted in red (α = 0.05). With R^2 = 0.8245 the dataset shows a very good fit to the calculated regression model. The factor “concentration” seems to have a highly significant effect on viability, the as a linear as well as quadratic component. “Time” appears the have a less important significant effect, and only as linear component while interaction between both predictors does not lead to any significant effect.

Dependency of % of non-lysed cells on concentration and lysis time

Response Surface regression model

fit to regression model	R^2 = 0.8245
lack of fit	*p* = 0.5689

type of regression	R^2	*p* (F-test)
linear	0.7472	< 0.0001
quadratic	0.0771	< 0.0001
crossproduct	0.0001	0.8389
total model	0.8245	< 0.0001

source	estimate	*p* (t-test)
intercept	6.18	0.0028
time	2.45	0.0379
conc	- 0.75	0.0002
time*time	- 0.32	0.0782
conc*time	0.01	0.8389
conc*conc	0.02	0.0005

S4. Optimized protocols including supplementary notes

S4.1. Nuclei isolation for sn-Multiomics in *Abeoforma whisleri*

Harvesting cells and estimating initial cell concentration & viability

1. Depending on the concentration of used cultures harvest between 500 and 2 000 µl of cells. In case multiple subsamples of different states / ages are combined, make sure to mix the final sample carefully but thoroughly.
2. Take 10 µl aliquot and stain with propidium iodide (1:1000).
3. Use a hemocytometer to count alive as well as broken / dead cells:
   - alive cells: only visible in brightfield, but not stained with PI
   - broken cells: potentially deformed in brightfield, stained with PI

   1st.Note: in other protocols it is suggested to estimate cell viability by using trypan blue staining of dead cells. However, this didn't work here, since trypan blue seems to be incompatible with the medium used for culturing Abeoforma (form heavy precipitation without staining dead cells).
4. Calculate concentration of cells per ml and viability in the initial sample, using the counts obtained for alive and dead cells.
   cell viability [%] = (1 – (#dead cells / #alive cells)) * 100
5. Calculate the needed sample volume corresponding to ~700 000 cells

Nuclei Extraction

Every step apart from step 2 should happen on ice. Keep all buffers that are used also on ice while working!

1. Prepare 1.800 µl Wash Buffer per sample. Add DTT and RNase inhibitors in respective amounts (see section S2.).
2. Take the sample volume corresponding to ~700 000 cells into a 2 ml Epi.
3. Centrifuge for 5 min, at 4°C and 0,5 rcf.
4. In the meantime prepare 200 µl 1x Lysis Buffer per sample:
   194 µl WB (incl. DTT) + 2 µl NP40 10% (final conc.: 0,1 M) + 4 µl Digitonin 0,5% (final conc.: 0,01 M)
5. Remove supernatant
6. Add 200 µl 1x LB and lyse for 2 min

   2nd.Note: 15 sec of carful pipetting to thoroughly resuspend cell pellet in LB.
7. Stop lysis after 2 min by adding 1 600 µl WB (incl. DTT) into the Epi.

   3rd.Note: just adding WB is enough to stop lysis sufficiently, no further pipetting is needed at this point.
8. Filter sample using a 30 µm MACS SmartStrainer filter – use the whole sample volume which should be 1.800 µl at this point and try to gather as much sample as possible.

9. Centrifuge for 10 min, at 4°C and 0,5 rcf.
10. In the meantime prepare 100 µl Nuclei Buffer Solution per sample
11. Remove supernatant
12. Resuspend nuclei in 100 µl NBS

    4th.Note: If necessary additional washes with NBS can be done (centrifuging 5-10 min, at 4°C and 0,5 rcf). Those can be used to clean up sample or to achieve a higher concentrated final sample using a smaller volume of NBS for resuspending nuclei in the final step.

Quantify nuclei concentration and cell viability after lysis

1. use AO/PI staining and a Cellometer to count remaining alive cells and isolated nuclei.

    5th.Note: after nuclei have been isolated, they should be processed as fast as possible to ensure good nuclei quality. Therefore, using a Cellometer for nuclei quantification is recommended in terms of saving time. For preliminary experiments on testing nuclei isolation without aiming to process isolated nuclei further, it might be very useful to use manual counting (PI-staining and a hemocytometer) in parallel. This might help to estimate sample quality since the visual appearance of nuclei can be judged in way more detail.

2. Following settings on the cell counter turned out to show good results:
   - exposure brightfield image: 5 msec
   - exposure PI image: 4 000 msec
   - cell diameter, brightfield image: 7 µm – 50 µm
   - nuclei diameter, PI image: 3 µm – 10 µm

    6th.Note: dependent on the abilities of the nuclei counter to differentiate between single nuclei and aggregates it might be useful to count a second time, changing the nuclei diameter to a wider range up to ~ 40µm. This second count shouldn’t be too apart from the first one, if nuclei are mostly present in a singular form.

3. By following the given protocol one should be able to achieve a post-lysis viability of ~0% and a nuclei concentration of 5 000 – 8 000 nuclei / 1µl.

S4.2 Nuclei isolation for sn-Mutliomics in *Capsaspora owczarzaki*

Harvesting cells and estimating initial cell concentration & viability

1. Depending on the concentration of used cultures harvest between 200 and 1 000 µl of cells. In case multiple subsamples of different states / ages are combined, make sure to mix the final sample carefully but thoroughly.
2. Take 10 µl aliquot and stain with propidium iodide (1:1000).
3. Use a hemocytometer to count alive as well as broken / dead cells:
   - alive cells: only visible in brightfield, but not stained with PI
   - broken cells: potentially deformed in brightfield, stained with PI

   1st.Note: in other protocols it is suggested to estimate cell viability by using trypan blue staining of dead cells. However, this didn't work in the given case, since I wasn't able to obtain any meaningful trypan blue staining of killed cells.
4. Calculate concentration of cells per ml and viability in the initial sample, using the counts obtained for alive and dead cells.
   cell viability [%] = (1 – (#dead cells / #alive cells)) * 100
5. Calculate the needed sample volume corresponding to ~2 500 000 cells

Nuclei Extraction

Every step apart from step 2 should happen on ice. Keep all buffers that are used also on ice while working!

1. Prepare 900 µl Wash Buffer per sample. Add DTT and RNase inhibitors in respective amounts (see section S2).
2. Take the sample volume corresponding to ~2 500 000 cells into a 1,5 ml Epi.
3. Centrifuge for 5 min, at 4°C and 0,5 rcf.
4. In the meantime prepare 100 µl 16x Lysis Buffer per sample:
   88,8 µl WB (incl. DTT) + 8 µl NP40 20% (final conc.: 1,6 M) + 3,2 µl Digitonin 5% (final conc.: 0,16 M)
5. Remove supernatant
6. Add 100 µl 1x LB and lyse for 1 min

   2nd.Note: 15 sec of carful pipetting to thoroughly resuspend cell pellet in LB.
7. Stop lysis after 1 min by adding 800 µl WB (incl. DTT) into the Epi.

   3rd.Note: just adding WB is enough to stop lysis sufficiently, no further pipetting is needed at this point.
8. Filter sample using a 30 µm MACS SmartStrainer – use the whole sample volume which should be 900 µl at this point and try to gather as much sample as possible.
9. Centrifuge for 10 min, at 4°C and 0,5 rcf.
10. In the meantime prepare 100 µl Nuclei Buffer Solution per sample

11. Remove supernatant
12. Resuspend nuclei in 150 µl NBS

    4th.Note: If necessary additional washes with NBS can be done (centrifuging 5-10 min, at 4°C and 0,5 rcf). Those can be used to clean up sample or to achieve a higher concentrated final sample using a smaller volume of NBS for resuspending nuclei in the final step.

Quantify nuclei concentration and cell viability after lysis

1. use AO/PI staining and a Cellometer to count remaining alive cells and isolated nuclei.

    5th.Note: after nuclei have been isolated, they should be processed as fast as possible to ensure good nuclei quality. Therefore, using a Cellometer for nuclei quantification is recommended in terms of saving time. For preliminary experiments on testing nuclei isolation without aiming to process isolated nuclei further, it might be very useful to use manual counting (PI-staining and a hemocytometer) in parallel. This might help to estimate sample quality since the visual appearance of nuclei can be judged in way more detail.

2. Following settings on the cell counter turned out to show good results:
   - exposure brightfield image: 5 msec
   - exposure PI image: 12 000 msec
   - cell diameter, brightfield image: 3 µm - 7 µm
   - nuclei diameter, PI image: 2 µm – 25 µm

    6th.Note: since extracted *Capsaspora* nuclei are really small and can take up just a relatively small amount of PI dye, a very long exposure time is needed to achieve a detectable signal. Most probably the nuclei counter will not be able to differentiate between single nuclei and aggregates under these conditions – in order to capture the single nuclei as well as bigger aggregates the range for the accepted nuclei diameter is defined very broad.

3. By following the given protocol one should be able to achieve a viability of 0-5% and a nuclei concentration of about 6 000 – 8 000 nuclei / µl