Characterizing the hologenome of *Lasallia pustulata* and tracing genomic footprints of lichenization

Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften

vorgelegt beim Fachbereich Biowissenschaften
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Bastian Greshake Tzovaras
aus Münster

Frankfurt (2017)
(D 30)

vom Fachbereich Biowissenschaften der

Johann Wolfgang Goethe-Universität als Dissertation angekommen.

Dekanin: Prof. Dr. Meike Piepenbring
(Mykologie)
Institut für Ökologie, Evolution und Diversität
Johann Wolfgang Goethe-Universität
D-60438 Frankfurt am Main

Gutachter: Prof. Dr. Ingo Ebersberger
(Angewandte Bioinformatik)
Institut für Zellbiologie und Neurowissenschaften
Johann Wolfgang Goethe-Universität
D-60438 Frankfurt am Main

PD Dr. Markus Pfenninger
(Molekulare Ökologie)
Institut für Ökologie, Evolution und Diversität
Johann Wolfgang Goethe-Universität
D-60438 Frankfurt am Main

Datum der Disputation: _____

The view of evolution as a chronic bloody competition among individuals and species, a popular distortion of Darwin's notion of "survival of the fittest," dissolves before a new view of continual cooperation, strong interaction, and mutual dependence among life forms. Life did not take over the globe by combat, but by networking. Life forms multiplied and complexified by co-opting others, not just by killing them.

LYNN MARGULIS, *MICROCOSMOS: FOUR BILLION YEARS OF EVOLUTION FROM OUR MICROBIAL ANCESTORS* (1986)

Abstract

The lichen symbiosis – consisting of fungal mycobionts and photoautotroph photobionts (green algae or cyanobacteria) – is globally successful. It covers an estimated 6% of the global surface with habitats ranging from deserts to the arctic. This success is reflected in the diversity of the mycobionts, with around 21% of all fungal species participating in lichen symbioses that can be facultative or obligate. Lichenization is furthermore evolutionary old, with fossil evidence for lichens reaching back 415 million years. For an individual fungal lineage, the Lecanoromycetes, the lichenization happened around 300 million years ago. This longstanding symbiotic relationship and the diversity of observed symbiotic dependency make them promising models to study the genomic consequences that follow the establishment of symbioses. Despite this, only little is known about the genomic effects of lichenization and extreme symbiotic dependency. To fill this gap we sequenced the hologenome of the lichen *Lasallia pustulata*, where the mycobiont could so far not been cultivated, suggesting that it might be more dependent on its symbionts.

As the poor culturability of lichen symbionts renders their genomes inaccessible to standard sequencing practices, we evaluated the extent to which different metagenome sequencing- and *de novo* assembly-strategies can be used to sequence and reconstruct the genomes of the individual symbionts. We find that the abundances of individual genomes present in the *L. pustulata* hologenome vary substantially, with the mycobiont being most abundant. Using *in silico* generated data sets and real *Illumina* sequencing data for *L. pustulata* we observe that the skewed abundances prevent a contiguous assembly of the underrepresented genomes when using only short-read sequencing. We conclude that short-read sequencing can offer first insights into lichen hologenomes. The fragmentation of the reconstructions hinders downstream analyses into the genomic consequences of lichenization though, as these are focused on identifying the gain and loss of genes.

We thus demonstrate a hybrid genome assembly strategy that is based on both short- and long-read sequencing. We show that this strategy is capable of creating highly contiguous genome reconstructions, not only for the *L. pustulata* mycobiont but also its photobiont *Trebouxia sp.*, along with substantial amounts of the bacterial microbiome. A subsequent analysis of the microbiome of *L. pustulata* – performed over nine different samples collected in Germany and Italy – showed a stable taxonomic composition across the geographic range. We find that Acidobacteriaceae, which are known to thrive in nutrient poor habitats, are the dominant taxa. These would make them well adapted for the co-habitation with *L. pustulata*, which largely grows on rocks. Whether the Acidobacteriaceae are functionally involved in the lichen symbiosis is unclear so far.

As further comparative genomic studies rely on comprehensive genome annotations, we evaluate the completeness and fidelity of the gene annotations for the mycobiont *L. pustulata* as well as four further Lecanoromycetes. This reveals that un- and mis-annotated genes impact all evaluated genomes, with artificially joined genes and unannotated genes having the largest impact. In addition to these factors we find that the sequence composition – especially G/C-rich inverted repeats – lead

to sequencing errors that interfere with the gene prediction. We minimize the effects of these artifacts through a rigorous curation.

Given the extremely sparse taxon sampling of available green alga genomes, we focus our search for the genomic footprints of lichenization on the mycobionts. We compare the genomes of the Lecanoromycetes to their closest relatives, the Eurotiomycetes and Dothideomycetes. This reveals that the last common ancestor of the Lecanoromycetes has lost around 10% of its genes after they split from the non-lichenized ancestor they share with the Eurotiomycetes. These losses are furthermore enriched, showing an excessive loss of genes involved with the degradation of polysaccharides. The loss of these genes fits a change from an ancestral saprotrophic lifestyle that depends on degrading complex plant matter, to the symbiotic lifestyle that relies on simpler nutrients provided by the photobionts. While the last common ancestor of the Lecanoromycetes additionally gained around 400 genes these could so far not be further characterized due to a lack of functionally annotated reference data.

As the mycobiont *L. pustulata* could so far not been grown in axenic culture, we initially expected to find an extensive genomic remodeling compared to the other mycobionts that easily grow in culture. We do not find evidence for this. Analyzing both the contraction of gene families and the loss of genes, we observe that *L. pustulata* and *Umbilicaria muehlenbergii* – its close relative that is easily grown in culture – share most of these. Furthermore, *L. pustulata* does not show an excessive loss of evolutionary old and well-conserved genes. These effects are mirrored on the functional level, as neither gene family contractions nor gene losses show a functional enrichment. This is partially due to the lack of functional reference data, analogous to the genes gained in the Lecanoromycetes, rendering their characterization hard. Thus, further studies on the genomic consequences of lichenization and differences in symbiotic dependence will have to be conducted, including larger taxon sets. This will be even more important for the photobionts, as the Chlorophyta are even more sparsely sampled today, hindering an effective functional and evolutionary study.

Zusammenfassung

Flechten bilden eine global erfolgreiche Gemeinschaft. Sie sind symbiotische Organismen die aus einem Pilz (Mykobionten) und einem oder mehreren photoautotrophen Photobionten bestehen. Die Photobionten können dabei Grünalgen oder Cyanobakterien sein. Schätzungen zufolge bedecken Flechten etwa 6% der Erdoberfläche. Dabei besiedeln Flechten diverse Habitate, die auch nährstoffarme arktische Regionen und Wüsten umfassen. Der Erfolg der Flechtengemeinschaft zeigt sich darüber hinaus auch an der Diversität der beteiligten Symbionten: Geschätzte 21% aller Pilzarten sind als Mykobionten an Flechten beteiligt. Darüber hinaus ist Lichenisierung ein altes Phänomen. Fossile Evidenzen für Flechten lassen sich auf über 415 Millionen Jahre vor heute datieren. Des weiteren gibt es molekulare Studien die zeigen, dass die Lecanoromyceten, eine Klasse von Pilzen die fast ausschließlich aus Mykobionten besteht, vor in etwa 300 Millionen Jahren entstanden ist und bereits lichenisiert war.

Die symbiotische Abhängigkeit einzelner Mykobionten variiert. Einige lichenisierte Pilze, wie *Cladonia grayi* oder *Umbilicaria muehlenbergii*, sind physiologisch fakultative Symbionten und können vergleichsweise einfach in axenischen Kulturen gehalten werden. Dies war bislang nicht möglich für Mykobionten wie *Lasallia pustulata* und Vertreter der Peltigerales. Sowohl das evolutionäre Alter der Flechten, als auch die Variabilität ihrer symbiotischen Abhängigkeit macht sie dabei zu interessanten Modellorganismen um die genomischen Auswirkungen von Symbiosen zu untersuchen. Die relativ kompakten Genome von sowohl den Mykobionten als auch den Photobionten machen Flechten darüber auch von einem technischen Standpunkt aus interessant. Dies gilt insbesondere im Vergleich zu anderen langanhaltenden Symbiosen – wie denen zwischen Mykorrhizapilzen und vaskulären Pflanzen. Dank der geringen Größe ist es daher auch möglich das Hologenom der Flechte zu analysieren. Neben dem Photobionten und Mykobionten kann daher auch das bakterielle Mikrobiome betrachtet werden. Dem zu trotz gibt es bislang nur wenige Studien zu den genomischen Auswirkungen der Lichenisierung. Die meisten Studien haben sich bislang auf vereinzelte Mykobionten beschränkt und gezielt einzelne Genen oder Genfamilien untersucht.

Unser Ziel ist es daher das Feld der Flechtengenomik zu erweitern und zu analysieren welche genomischen Konsequenzen die Lichenisierung hat. Darüber hinaus wollen wir untersuchen in wie weit die genomischen Änderungen sich zwischen individuellen Symbionten unterschiedet. Wir sequenzieren daher das Hologenom der Flechte *L. pustulata*, für die der Mykobiont bislang nicht in axenischer Kultur gehalten werden konnte. Dies deutet auf weitergehende genomische Konsequenzen hin. Da das separate Sequenzieren der einzelnen Symbionten ob der schlechten Kultivierbarkeit nicht möglich ist, ist es hier nötig metagenomische Sequenzierungsverfahren zu nutzen welche alle Genome in einer Sequenzierlibrary vereinen. Die Rekonstruktion einzelner Genome aus solchen Datensätzen hat sich bereits in bakteriellen Metagenomen als komplex erwiesen. Daher entwickeln wir einen simulationsbasierten Ansatz um zu evaluieren ob metagenomische Daten aus einer einzelnen *Illumina* Sequenzierlibrary geeignet sind um die Genome von Photobiont und Mykobiont zu rekonstruieren. Dazu entwickeln wir die Idee von *Twin Data Sets* – *in silico* generierte Sequenzreads die auf den Parametern einer echten Sequenzierlibrary basieren. Mit diesen *Twin Data Sets*

testen wir in wie weit die Genomrekonstruktionsqualität verschiedener *de novo* Genomassembler von dem Readcoverageverhältnis zwischen Mykobiont und Photobiont abhängt. Die *Twin Data Sets* zeigen, dass ein Ungleichgewicht zwischen den beiden Genomen einen erheblichen Einfluss auf die Qualität der Genomrekonstruktionen hat. Bei starken Ungleichgewichten sind sowohl die Kontiguität, als auch die assemblierte Genomgröße, beeinträchtigt. Dieser Effekt ist ein Resultat der geringen Readcoverage für das unterrepräsentierte Genome, welche dazu führt das genomische Regionen nicht ausreichend tief sequenziert wurden um sie eindeutig zu rekonstruieren oder sogar gar nicht in den Sequenzreads repräsentiert sind. Nicht alle Genomassembler sind dabei in gleichem Maße sensibel gegenüber solchen Ungleichgewichten. Sowohl der Overlap-basierte Assembler *MIRA* als auch der multi-*k*-mer basierte Assembler *SPAdes* sind dabei besonders unempfindlich für solche geringen Readcoverages. Darüber hinaus zeigt die *Twin Data Set*-Studie, dass die oft zur Assemblyparameteroptimierung verwendete Methode der N50-Maximierung problematisch ist wenn sie auf metagenomische Datensätze angewandt wird. Bei einer extremen Ungleichverteilung der Readcoverages zwischen den beiden Genomen finden wir, dass die N50-Maximierung dazu führt dass das unterrepräsentierte Genom überhaupt nicht rekonstruiert wird. Die Parameteroptimierung für metagenomische Assemblies ist daher nicht trivial.

Wir wenden die verschiedenen Genomassemblierungsmethoden nachfolgend auf ein einen echten *Illumina* Datensatz von *Lasallia pustulata* an. Die Assemblies sind fragmentierter als erwartet gegeben den *Twin Data Sets* und zeigen ein komplexes bakterielles Mikrobiome sowie ein deutliches Ungleichgewicht in der Readcoverage zugunsten des Mykobionten. Beide Eigenschaften führen dazu dass die Readcoverages insbesondere für *Trebouxia sp.* – den Photobionten von *L. pustulata* – so weit reduziert werden, dass die Genomrekonstruktion extrem fragmentiert bleibt. Frühere Studien, basierend auf einzelnen Genomen haben gezeigt, dass längere Sequenzreads bei der Genomrekonstruktion unterstützend wirken können. Aus diesem Grund entwickeln wir eine erweiterte, hybride Assemblierungsstrategie welche neben kurzen *Illumina* Sequenzreads auch lange *PacBio* Sequenzreads benutzt. Dazu verwenden wir verschiedene Assemblierungsmethoden, welche die längeren Sequenzreads auf mehrere Arten verwenden. Dies erlaubt uns die verschiedenen Genome mit ihren unterschiedlichen Readcoverages ideal zu assemblieren. Die resultierenden Assemblies werden nachträglich miteinander kombiniert und führen zu einem hologenomweiten Assembly welches für alle Organismen vollständiger und weniger fragmentiert ist als die individuellen Assemblies. Diese Methodik ist daher vielversprechend um auch weitere Hologenome und komplexe Mikrobiome mit starken Differenzen in den Readcoverages zwischen einzelnen Taxa zu assemblieren.

Die Verwendung verschiedener Sequenzierlibraries erlaubt uns darüber hinaus das bakterielle Mikrobiom von *L. pustulata* besser zu charakterisieren. Insgesamt nutzen wir dafür Sequenzdaten aus neun verschiedenen Libraries welche von Flechtenthalli aus Deutschland und Italien generiert wurden. Der Vergleich dieser Daten zeigt, dass es ein stabiles Mikrobiome gibt, welches über alle Datensätze und damit auch die gesamte geographische Breite zu finden ist. Acidobacteriaceae dominieren das Mikrobiom in alle Libraries. Diese Familie von

Bakterien ist tolerant gegenüber starken Änderungen in der Wasserverfügbarkeit. Darüber hinaus können sie ihre metabolische Rate anpassen wenn die Nährstoffverfügbarkeit gering ist. Durch diese Eigenschaften sind sie potentiell adaptiert auf das Zusammenleben mit *L. pustulata,* welche kahlen Fels in Südrichtung als Habitat bevorzugt. Obwohl das konsistente Vorkommen der Acidobacteriaceae für eine funktionelle Involvierung an der Flechtensymbiose spricht, so kann eine geteilte Habitatpräferenz als einziger Grund für das gemeinsame Auffinden nicht ausgeschlossen werden. Weitere Studien zur Lokalisation auf bzw. im Flechtenthallus sowie zu den funktionellen Möglichkeiten der Acidobacteriaceae werden nötig sein um dies abschließend zu klären. Unsere Genomrekonstruktionen der Acidobacteriaceae bieten dafür ideale Ausgangsbedingungen.

Tiefergehende evolutionäre Analysen erfordern komplette und korrekte Genomannotationen. Dies trifft insbesondere auf die korrekte Vorhersage von hinzugewonnen und verloren Genen zu. Aus diesem Grund annotieren wir nicht nur die Gene für die Symbionten von *L. pustulata*, sondern evaluieren darüber hinaus auch die Performance einzelner Genvorhersagemethoden und potentielle Annotationsfehler. Ein Vergleich der Allzweck-Genvorhersagesoftware *MAKER2* mit der pilzspezifischen Annotationspipeline *funannotate* zeigt deutliche Unterschiede. Vor allem *MAKER2* sagt zusätzliche Gene vorher, welche weder durch RNAseq-Daten noch Orthologe in anderen Taxa unterstützt werden, und verpasst Gene die durch diese Evidenzen unterstützt werden. Als weitere Vorbereitung auf die evolutionären Analysen des genomischen Fußabdrucks der Lichenisierung untersuchen wir darüber hinaus die Genauigkeit der Genvorhersage in *L. pustulata* sowie in vier weiteren, bereits annotierten Lecanoromyceten. Für den Vergleich nutzen wir das evolutionär interessante Set an Genen die bereits im letzten gemeinsamen Vorfahren der Lecanoromyceten vorhanden waren und nur in einem der Genome verloren gingen. Der seltene Verlust alter Gene weist dabei auf eine wichtige Rolle für die Lecanoromyceten hin, ein Verlust ist daher ein Hinweis für weitreichende genomische Konsequenzen. Ihre zentrale Rolle bedeutet allerdings auch, dass ein beobachteter Verlust eine hohe Wahrscheinlichkeit hat eine falsch-positive Vorhersage zu sein. Wir fokussieren uns daher auf Genen die in einer ersten Suche als ein privater Genverlust erkannt wurden. Solch privat verlorene Gene suchen wir dann mit weiteren Methoden während wir schrittweise die Suchsensitivität erhöhen. Dabei können für die verschiedenen Genome nur zwischen 9% und 25% der initial beobachteten Genverluste nicht als falsche Vorhersagen bestätigt werden. Unsere extensive Analyse der Falsch-positiven zeigt das verschiedene Artefakte dazu beitragen. Vor allem artifizielle Genfusionen beeinträchtigen dabei das Finden von Genverlusten. Dabei führen kurze Intergenbereiche und überlappende, untranslatierte 5’- und 3’-Regionen dazu das Genannotationen von benachbarten Genen von den Genannotationsmethoden als ein durchgehendes Gen vorhergesagt werden. Weitere Faktoren sind unannotierte Gene, welche in den assemblierten Genomen vorhanden sind aber nicht annotiert wurden, sowie nicht-assemblierte Genomregionen. Darüber hinaus finden wir das die bislang häufig vernachlässigte Sequenzkomposition einen starken Einfluss auf die korrekte Genvorhersage und damit die korrekte Identifizierung von Genverlusten hat. Insbesondere stark G/C-haltige Genomregionen, in besonders pathogenen Fällen mit invertierten Repeats

gepaart, führen dabei zu *in silico* nicht leicht zu identifizierenden Sequenzierfehlern, welche artifizielle Insertionen und Deletionen erzeugen. Diese erzeugen künstliche Stop-Codons, welche die Genvorhersage beeinträchtigen. In extremen Fällen können G/C-reiche invertierte Repeats sogar dazu führen, dass die entsprechenden Regionen nicht sequenziert werden können, entsprechende Gene in diesen Regionen sind dann im Assembly nicht repräsentiert. Eine extensive Kuratierung von annotierten Genen ist daher nötig um sicherzustellen das Genverluste korrekt vorhergesagt werden können.

Die funktionelle Annotation der Gene der Lecanoromyceten und des photobionten *Trebouxia sp.* zeigt deutliche Unterschiede. Während etwa 50% aller Gene der Lecanoromyceten mit Gene Ontology-Terms annotiert werden können, so ist dies nur möglich für 35% der *Trebouxia sp.*-Gene. Dies deutet auf einen markanten Unterschied in der Verfügbarkeit von Referenzdaten hin, welche für die für die funktionelle Annotation benötigt werden. Dies spiegelt sich auch in der funktionellen Kapazität wieder die wir in den Lecanoromyceten und den Grünalgen finden. Die betrachteten Lecanoromyceten zeigen ein stark überlappendes Set an Genfunktionen. In etwa 80% der in den Lecanoromyceten annotierten Funktionen werden in allen Mykobionten gefunden. Die Chlorophyta hingegen zeigen eine deutlich größere Diversität zwischen verschiedenen Taxa, was auf eine größere Divergenz zwischen den repräsentierten Grünalgen hindeutet. Eine Phylogenierekonstruktion die sowohl die Lecanoromyceten als auch die Chlorophyta umfasst bestätigt dies. Die Lecanoromyceten, welche intern relativ kurze Astlängen aufzeigen, werden in ein dichtes Taxonsampling zu den nächst verwandten Gruppen, den Dothideomyceten und Eurotiomyceten, platziert. Die Chlorophyta hingegen zeigen lange Astlängen auf, trotz der Verwendung aller verfügbaren Genome von einzelligen Grünalgen. Ob dieses Mangels an Referenzdaten fokussieren wir uns daher auf die Lecanoromyceten bei der Suche nach genomischen Änderungen im Zusammenhang mit der Lichenisierung.

Der Wechsel von einem solitären Organismus zu einem symbiotischen führt zu einer genomischen Adaptation. Vorhandene Gene werden obsolet, während neue Gene für die Interaktion mit den Symbionten benötigt werden. Solche Anpassungen im Zuge einer Symbiose und die damit erweiterte bzw. eingeschränkte funktionale Kapazität wurden für Bakterien und Mykorrhiza bereits gezeigt. Wir erwarten daher, dass die Lichenisierung einen vergleichbaren Effekt auf die Genome der Lecanoromyceten hat. Aus diesem Grund nutzen wir komparative Genomanalysen zwischen den Lecanoromyceten, Eurotiomyceten und Dothideomyceten um entsprechende Effekte zu identifizieren. Darüber hinaus untersuchen wir die Diversität der genomischen Anpassungen an die Flechtensymbiose. Wir erwarten das der Mykobiont *L. pustulata*, welcher bislang nicht kultiviert werden konnte, weitreichendere genomische Effekte zeigt als leicht kultivierbare Mykobionten. Wir vergleichen daher das Genom von *L. pustulata* mit anderen Lecanoromyceten, welche bereits erfolgreich in axenischen Kulturen gehalten wurden.

Der Vergleich der Lecanoromyceten mit nicht-lichenisierten Verwandten zeigt keine Anzeichen für eine rapide Kontraktion von Genfamilien in der frühen Evolution der Lecanoromyceten. Die Anzahl kontrahierter Genfamilien ist vergleichbar für alle drei Klassen. Auf Ebene individueller Gene hingegen finden wir, dass der letzte gemeinsame Vorfahre der Lecanoromyceten in etwa 800 der Gene

verloren hat welche im Vorfahren der Eurotiomyceten und Lecanoromyceten vorhanden waren (10%). Diese Genverluste zeigen eine funktionelle Überrepräsentierung und enthalten Gene welche in der Degradation von Polysacchariden involviert. Der Verlust dieser Gene kann durch die Adaptation von einem saprotrophen zu einem symbiotischen Lebenswandel erklärt werden. Während der gemeinsame Vorfahre der Lecanoromyceten und Eurotiomyceten auf die Zersetzung von komplexen Pflanzenstoffen angewiesen war, so können die lichenisierten Lecanoromyceten auf simplere Nährstoffe von ihren Photobionten zurückgreifen. Darüber hinaus finden wir in etwa 400 Gene welche der letzte gemeinsame Vorfahre der Lecanoromyceten hinzugewonnen hat. Mangels Referenzdaten können wir die meisten dieser Gene nicht funktionell charakterisieren. Dementsprechend gibt es keine Anzeichen für eine Überrepräsentierung einzelner Funktionen.

Die komparative Analyse der Genome der Lecanoromyceten zeigt deutliche Unterschiede in der Anzahl der sekretierten Proteine zwischen den individuellen Taxa. Während die Sekretomgrößen von *L. pustulata* und *U. muehlenbergii* vergleichbar sind mit denen der überwiegend parasitisch lebenden Eurotiomyceten, so sind die Sekretome der restlichen Lecanoromyceten deutlich größer und vergleichbar mit den saprotroph lebenden Dothideomyceten. Diese Unterschiede korrelieren mit dem präferierten Substrat auf dem die einzelnen Lecanoromyceten wachsen. Während *L. pustulata* und *U. muehlenbergii* auf Felsen zu finden sind, so leben die restlichen Lecanoromyceten überwiegend auf pflanzlichem Substrat, wie lebenden oder toten Bäumen. Dies deutet darauf hin dass das organische Substrat von den entsprechenden Lecanoromyceten genutzt wird.

Für das Genom von *L. pustulata* finden wir keine deutlichen Hinweise auf eine extensive Remodellierung des Genoms welche die schlechte Kultivierbarkeit erklären würde. Die Genfamilienevolution von *L. pustulata* ist vergleichbar mit der des nahen verwandten Mykobionten *U. muehlenbergii*. Die meisten Expansionen und Kontraktionen von Genfamilien findet man entsprechend im gemeinsamen Vorfahren der beiden. Ein ähnliches Bild ergibt sich für den Verlust einzelner Gene: *U. muehlenbergii* und *L. pustulata* zeigen ähnlich viele Genverluste, während deutlich mehr Genverluste im gemeinsamen Vorfahren zu beobachten sind. Darüberhinaus hat *L. pustulata* nicht deutlich mehr evolutionär alte Gene verloren als die anderen Lecanoromyceten. Die Suche nach einer funktionellen Anreicherung unter den Genverlusten und Veränderungen der Genfamilien zeigt darüber hinaus keine signifikante Veränderung in der funktionalen Kapazität von *L. pustulata*. Der Mangel an funktionell annotierten Referenzdaten ist potentiell daran beteiligt, da ein Großteil der verlorenen Gene nicht eindeutig klassifiziert werden kann. Daher lässt sich kein abschließendes Urteil fällen wieso *L. pustulata* bislang nicht in axenischer Kultur gehalten werden konnte. Weitere Studien zu den genomischen Konsequenzen der Lichenisierung und der unterschieden in der symbiotischen Abhängigkeit sind daher nötig, insbesondere unter Berücksichtigung das bessere Referenzdatensätze benötigt werden um eine bessere funktionale Auflösung zu erreichen. Dies wird vor allem für *Trebouxia sp.,* und die Chlorophyta allgemein, zentral. Die Abwesenheit von brauchbaren Referenzen macht eine Analyse der genomischen Auswirkungen einer entstehenden Symbiose für die Chlorophyta bislang unmöglich.

Table of Contents

Publications and Data Availability

in silico Evaluation of the assembly problem in metagenomic contexts

This chapter has been published as Greshake B, Zehr S, Dal Grande F, Meiser A, Schmitt I, Ebersberger I. (2016) *Potential and pitfalls of eukaryotic metagenome skimming: a test case for lichens*. Molecular Ecology Resources, 16: 511–523. doi:10.1111/1755-0998.12463

The data associated with this chapter have been archived at *DRYAD* at http://dx.doi.org/10.5061/dryad.8h95q.

Assembly & characterization of the *L. pustulata* metagenome; Genome annotation, artifacts and solutions; Evolutionary consequences of lichenization

A manuscript based on these chapters is in preparation with the draft title *Evolutionary Grounds and Consequence of lichen symbiosis on the example Lasallia pustulata*.

The data associated with these chapters have been archived at *zenodo* at http://dx.doi.org/10.5281/zenodo.894741.

Index of Figures

Index of Tables

1 Introduction

Fungi are a diverse kingdom with an estimated 2.2 to 3.8 million species (Hawksworth and Lücking 2017), found in almost all terrestrial habitats (Mueller et al. 2007; Treseder and Lennon 2015). Their diversity and global distribution is reflected in their wide range of lifestyles, allowing them to utilize highly different energy and nutrient sources (Lewis 1973). While it is estimated that around half of fungi survive by degrading dead organic material (saprophytes), there is a considerable diversity of fungi that are living in symbioses with other eukaryotic organisms (biotrophes) (Hawksworth 1988). A sizeable fraction of these symbiotic fungi live as parasites, using other fungi (Lawrey and Diederich 2003), animals (Fisher et al. 2012; Spatafora et al. 2007) or plants (Newton et al. 2010) as their hosts. Partially, these fungi are living as hemibiotrophes that parasitize and kill their respective hosts, and subsequently live saprotrophically (Horbach et al. 2011). The parasitic fungi of vascular plants have been shown to form specialized hypha that invade cells and tissues to tap into their hosts' metabolism (Strange and Scott 2005; Dean et al. 2012).

In contrast to these parasitic interactions, a substantial amount of fungi live in commensalistic or mutualistic symbioses (Lewis 1973). Mycorrhiza-forming fungi, and particularly the arbuscle-forming representatives found in the Mucoromycotina (K. J. Field et al. 2015) and the Glomeromycotina (Schüβler, Schwarzott, and Walker 2001), are closely studied cases of mutualistic symbioses. In these groups, fungi and their vascular plant hosts form a symbiosis that is well adapted to living at the interface of soil and atmosphere, with the plant partner utilizing light and CO_2 to produce carbohydrates, and the fungal partner exploiting resources like nitrogen and phosphates from the rhizosphere (Bonfante and Genre 2010). This type of

interaction has been shown to be so successful that around 80% of vascular land plants engage in mycorrhizal symbioses (Smith and Read 2008).

1.1 The lichen symbiosis

Mutualistic interactions between fungi and photoautotrophic organisms extend beyond the vascular land plants. Lichens are communities that largely consist of fungi (mycobionts) and photosynthesizing photobionts, which belong to either the green algae or cyanobacteria (Lewis 1973) . The exact nature of the symbiosis between the mycobiont and photobiont is unclear, and the proposed symbiotic interactions range from parasitic or commensalistic to mutualistic (Lewis 1973; Ahmadjian and Jacobs 1981; Ahmadjian 1993; Lücking et al. 2009). The interplay between mycobionts and photobionts is so close that they were initially classified as plants (Honegger 2000); and it was only in the late 19th century – thanks to early lichenologists like Simon Schwendener and Beatrix Potter – that their symbiotic nature was recognized (Margulis 2003).

Traditionally, lichens have been seen as the symbiosis of a single mycobiont and one or two photobionts, but recent studies have shown that both additional fungi (Spribille et al. 2016) as well as a bacterial microbiome (Hodkinson et al. 2012) can be involved in the lichen symbiosis. It appears that this symbiosis makes lichens globally successful, with lichens playing a key role in making fungi an integral part of terrestrial biodiversity in nearly all ecosystems (Ahmadjian 1993). Lichens cover an estimated 6% of the global surface (Gadd 2010) and their habitats range from arctic (Muller 1952) to desert environments (Kidron and Temina 2010). Furthermore, lichens are often amongst the first to colonize new terrestrial habitats (Caruso and Rudolphi 2009).

The success of the lichenized lifestyle is reflected in the diversity of the mycobionts, with an estimated 21% of all fungi being lichenized (Lewis 1973).

Their spread over 39 orders suggests between 20-30 independent lichenization events (Lucking, Hodkinson, and Leavitt 2016). Lichenization itself appears to be an evolutionarily old phenomenon, with fossil evidence reaching back 415 million years (Honegger, Edwards, and Axe 2013). Individual lineages, like the fungal class of the Lecanoromycetes, are especially indicative of the long-standing lichen symbiosis. Nearly all members of the Lecanoromycetes, which started diverging around 300 million years ago (Amo de Paz et al. 2011), are living as mycobionts, suggesting that the last common ancestor of them was already living in a lichen symbiosis.

1.2 The genomic consequences of symbiosis

The establishment of biotrophic interactions has been shown to often lead to a drastic genomic remodeling, for both parasitic as well as mutualistic interactions (Bonfante and Genre 2010; Newton et al. 2010; S. M. Schmidt and Panstruga 2011). Adapting from a solitary to a symbiotic lifestyle frequently requires the gain of new functions, and thus new genes, to interact with respective symbionts (F. Martin et al. 2008; Kohler et al. 2015; Bonfante and Genre 2010). At the same time, the symbiosis renders some functions obsolete, as the symbionts can provide these, leading to a loss of genes and thus molecular functionality (Kohler et al. 2015; McCutcheon and Moran 2011; Ochman and Moran 2001; F. Martin et al. 2008). These effects have been the subject of study for parasitic fungi (Chaudhari et al. 2014; S. M. Schmidt and Panstruga 2011; Lowe and Howlett 2012) and mycorrhizal fungi (Garcia et al. 2015; F. Martin et al. 2008; Kohler et al. 2015; Tisserant et al. 2013). For the mutualistic mycorrhizal fungi, studies have found a gain of secreted proteins linked to the interaction with their plant symbionts, as well as a contraction of gene families involved in plant cell wall degradation (F. Martin et al. 2008). This contraction of gene families has also been related to the inability to grow solitarily for some mycorrhizal fungi (Tisserant et al. 2013).

It is increasingly recognized that a thorough study of symbiotic interactions needs to take the hologenome, which includes all participants involved in a symbiotic ecosystem, into account (Theis et al. 2016). A comparative analysis then often depends on the completeness and fidelity of the genome reconstructions (Denton et al. 2014). Performing such a comprehensive, hologenome-wide study of the genomic consequences of symbiosis is problematic in the case of mycorrhiza. This has two reasons: 1. Their symbiotic plants have sizeable genomes, ranging from 500 Mbp (Tuskan et al. 2006) to over 20 Gbp (Neale et al. 2014), rendering a complete reconstruction of their photoautotroph symbionts challenging. 2. Their soil-based ecosystem often does not allow for a clear demarcation of the hologenome as accidentally sampled species lead to confounding. The symbiotic dependence of lichens is similar to that of mycorrhizal fungi – with some lichenized fungi only growing poorly in culture (McDonald, Gaya, and Lutzoni 2013). But in contrast to mycorrhiza, the individual genomes found in lichen hologenomes are rather small; with even the photobiont genomes having sizes of around 50 Mbp. Additionally, many lichens grow on virtually uncolonized surfaces like rocks that are poor in nutrients, allowing for an easy demarcation of the lichen hologenome. Thus, lichens can make a promising model system to study the genomic effects of symbioses.

1.3 Studying the genomic effects of lichenization

Despite these interesting characteristics, the hologenomes of lichens have so far not been studied in depth. DNA barcoding studies have investigated the diversity of mycobionts and photobionts (Magain and Sérusiaux 2015; P. Moya et al. 2017). The genomes of physiologically facultative symbionts – that can be grown in axenic cultures (Martínez-Alberola 2015) – have been investigated for various aspects: The genome of the mycobiont *Cladonia*

uncialis has been mined for gene clusters for the synthesis of polyketides that are of biotechnological interest (Abdel-Hameed et al. 2015). A genome reconstruction of the mycobiont *Endocarpon pusillum* revealed some genetic mechanisms of drought-tolerance along with first indications that secreted proteins and transporters facilitate the communication between mycobiont and photobiont (Wang et al. 2014). Additionally, metagenomic sequencing is frequently utilized for the study of mycobiont genomes. Metagenomic data were used to identify novel microsatellite markers (Lutsak et al. 2016); analyze the evolution of ammonium transporters and ammonia permeases amongst different mycobionts (McDonald et al. 2013); and to investigate adaptations to different habitats and climates (Dal Grande et al. 2017; Junttila and Rudd 2012). Metagenomic sequencing has furthermore allowed a first characterization of the bacterial microbiome found in lichens (Grube and Berg 2009; Grube et al. 2015; Hodkinson et al. 2012).

Yet, so far there have been only limited efforts to gain a comprehensive picture of how lichenization shaped the genomes of the participating organisms and how the degree of symbiotic dependence influences the genome evolution. The mycobionts belonging to the Lecanoromycetes offer an interesting model to study these genomic consequences. It is assumed that their change to a lichenized lifestyle already took place in their last common ancestor, after their split from the lineage to the Eurotiomycetes (Gueidan et al. 2008). Thus, the species in the Lecanoromycetes have experienced long-term adaptations to the lichenized lifestyle over a period of around 300 million years (Amo de Paz et al. 2011), resulting in ample of time for genomic effects to manifest in their genomes. Additionally, inside the Lecanoromycetes the degree of symbiotic dependence varies between species. A comparative study of physiologically facultative and obligate symbionts (Lewis 1973; McDonald, Gaya, and Lutzoni 2013) furthermore allows the investigation of

more extreme symbiotic dependences and the accompanying genomic remodeling.

We thus sequenced the hologenome of the rock-dwelling lichen *Lasallia pustulata* (Hestmark et al. 1997) to facilitate evolutionary comparative genomics studies into the genomic consequences of lichenization. To the best of our knowledge it was so far not possible to grow the mycobiont *L. pustulata* in culture. Given this we expect that this will furthermore allow the analysis of different degrees of symbiotic dependence and genomic remodeling. Earlier studies based on metagenomic data have often been hampered by the fragmentation of the resulting draft genomes (McDonald et al. 2013). In Chapter 2, "*in silico* evaluation of the assembly problem in metagenomic contexts", we therefore investigated to what extent genome skimming can be applied to characterize the eukaryotic genomes of a lichen community and the quality of the genome reconstructions one can expect given such data. We do this by applying a simulation-based framework, which models *in silico*-generated sequencing data mirroring the results of a real sequencing experiment. Subsequently we evaluate *de novo* assembler performance and its implications for the reconstruction of the hologenomes of lichens.
In Chapter 3, "Assembly & characterization of the *L. pustulata* metagenome", we describe our approaches to sequence, assemble and characterize the taxonomic composition of the *L. pustulata* hologenome. We demonstrate the utility of a hybrid *de novo* assembly strategy, which uses both long and short read sequencing data, to comprehensively assemble the hologenome of *L. pustulata*. This approach allows us to cope with the different genome coverages observed in our sequencing data. After the initial reconstruction of the hologenome we characterize its taxonomic diversity. We compare the taxonomic composition found in the bacterial microbiome observed in different *L. pustulata* thalli sampled from different sampling sites.

Chapter 4, "Genome annotation, artifacts and solutions", describes the annotation of the hologenome of *L. pustulata*, predicting genes and functionally annotating them. As both assembly and gene annotation artifacts can bias further evolutionary inferences, we describe a detailed, stepwise procedure to estimate and minimize the artifacts we observe in *L. pustulata* as well as in four other, public Lecanoromycetes genomes.

Subsequently, we perform detailed evolutionary genomics analyses as described in Chapter 5, "Evolutionary consequences of lichenization", to search for both the evolutionary footprint of lichenization in general, as well as the genomic origins of the poor culturability observed in *L. pustulata*. To that end, we investigate the evolution of secreted proteins and gene families, as well as the gain and loss of individual genes between the Lecanoromycetes and their closest relatives, the Eurotiomycetes and the Dothideomycetes.

2 *in silico* evaluation of the assembly problem in metagenomic contexts

2.1 Introduction

The successful commercialization of second-generation sequencing techniques, generating millions of short sequencing reads, has enabled genetic and genomic studies in a large number of fields (Mardis 2008; Schatz, Delcher, and Salzberg 2010). Second generation sequencing facilitates the *de novo* assembly of a large number of new genomes (R. Li et al. 2010; Read et al. 2013; Gnerre et al. 2011; Quail et al. 2012), the comparative study of transcriptomes to investigate niche adaptations (H. Schmidt et al. 2013; Feldmeyer et al. 2015; Elmer et al. 2010), as well as population-scale resequencing projects (Dal Grande et al. 2017; Auton et al. 2015; Lek et al. 2016). Second generation sequencing methods are not only used for sequencing individual taxa, but are also being applied to survey the genetic complexity of metagenomic communities (Tully et al. 2016; Sangwan et al. 2016). Metagenomic sequencing has been used to study the taxonomic and functional diversity as well as the evolutionary history of microbial communities (Olm et al. 2017; Olson et al. 2017).

2.1.1 (Meta)genome skimming

A popular method for assessing the genomes of so-far unsequenced organisms is genome skimming, which requires only a single sequencing library and limited sequencing coverage (Elgar et al. 1999). This enables rapid and low-cost genomic surveys that can be applied to studies ranging from phylogenomics or phylogeography to population genetics (Bock et al. 2014; Malé et al. 2014; Weitemier et al. 2015). A further application of the genome skimming approach is its use on metagenomic data sets instead of data

derived from a single organism. The sequencing of metagenomic samples facilitates the study of complex ecological niches (e.g. soil or gut) to characterize their functional and taxonomic diversity (e.g. Franzosa et al. 2014; Forsberg et al. 2014; Rondon et al. 2000). While most metagenome skimming has focused on microbial communities, nascent metagenomic sequencing efforts of lichen communities (Grube and Berg 2009; Cardinale, Puglia, and Grube 2006; McDonald et al. 2013; Erlacher et al. 2015; Kampa et al. 2013) essentially resemble such metagenome skimming experiments. In contrast to many microbial communities though, lichens contain at least one eukaryotic member, often even more (Millanes, Diederich, and Wedin 2016; Spribille et al. 2016).

2.1.2 Genome assembly strategies

Genome skimming, like many other sequencing applications, relies on shotgun sequencing, in which the template nucleotide sequences are randomly fragmented, thus requiring an extensive post-sequencing reconstruction of the target sequences (Miller, Koren, and Sutton 2010). Depending on data availability, this can either be achieved by a reference-based genome/transcriptome assembly (Chevreux, Wetter, and Suhai 1999; Trapnell et al. 2010) or through a *de novo* assembly of the sequenced data (Kumar and Blaxter 2010; Baker 2012). While the use of reference-based assemblies is typically preferred for its lower computational demands, this is often not possible due to a lack of reference data. For this reason a sizeable number of *de novo* genome assembly algorithms have been developed (Schatz et al. 2012; Magoc et al. 2013; Kajitani et al. 2014). Most of these tools fall conceptually into one of two categories, applying either overlap graphs (OLG) or de Bruijn graphs (DBG) (Miller, Koren, and Sutton 2010).

2.1.2.1 Evaluating genome assembly quality

The first step after performing any *de novo* assembly is to address the question of how good the resulting genome reconstruction is. To address this, post-assembly methods have been developed to assess the assembly quality, based on either comparisons to a known reference genome (E. Bao, Song, and Lan 2017), or making use of inherent assembly statistics (Gurevich et al. 2013; Hunt et al. 2013). Frequently used metrics in the evaluation of *de novo* genome assemblies include:

1. ***Total assembly length***: Evaluates how closely the achieved assembly size resembles the expected assembly size.
2. ***N50***: The length of the contig that, along with all longer contigs, makes up 50% of the concatenated assembly length.
3. ***NG50***: Analogous to the N50, but taking the expected genome length instead of the assembly as the total length.
4. ***NGA50***: Analogous to the NG50, but only counting contig blocks that can be consistently aligned to the reference genome(s).
5. ***Number of misassemblies:*** The number of splits that need to be performed to correctly map the assembly to the reference.

In the absence of a reference, a focus is often put on both the total assembly length, as well as on the N50 as a proxy for the assembly contiguity. Furthermore, methods to evaluate the completeness of a genome have been developed. These are based on the recall of well conserved genes, which are either ubiquitously present in all genomes (Parra, Bradnam, and Korf 2007), or are clade-specific for a defined group of taxa (Simão et al. 2015).

2.1.2.2 Overlap graph based methods

Methods that rely on overlap graphs (OLGs) use each sequencing-read as a continuous stretch of a given template DNA that was sequenced. By finding reads that have pairwise overlaps, and subsequently arranging them in a

consistent layout, they reconstruct "contigs" (contiguous consensus sequences, *i.e.* free of gaps). Examples of OLG assemblers include the overlap layout consensus-based tools like *CAP3* (X. Huang and Madan 1999) and *Celera* (Myers et al. 2000); and other implementations like *MIRA* (Chevreux, Wetter, and Suhai 1999). These assemblers often assume a uniform read coverage over the whole genome, which is not fulfilled in metagenomic data with varying abundances between taxa. As this can lead to individual genomes being misidentified as repetitive regions or sequencing errors, dedicated metagenome assemblers based on OLGs, like *Omega* (Haider et al. 2014), have been developed.

As OLGs allow for overlaps of varying size, these methods can in principle work with coverages close to 1, as long as the found minimal overlaps create unique joins between sequencing reads. At the same time, OLG-based methods often cannot be easily applied to eukaryotic whole genome sequencing projects that rely on large second generation sequencing data sets with high coverages (Shendure and Ji 2008). This is a result of the run time of OLG-based methods that increases quadratically with the amount of the input data, as all pairwise overlaps need to be calculated (Miller, Koren, and Sutton 2010). Due to these limitations, alternative approaches have been developed, which either use a more efficient implementation of the overlap-criterion, like string graphs (Simpson and Durbin 2012; Simpson and Durbin 2010), or alternatively use de Bruijn graphs (DBG).

2.1.2.3 De Bruijn graph-based methods

Unlike OLG-based methods, DBG assemblers don't overlap individual reads. Instead, they extract words of length k (k-mers) from the sequencing reads, representing the words that appear in the genomes, as well as words generated by sequencing errors. The word frequencies can then be used to differentiate between genomic and erroneous words. The k-mers are then

used to build a de Bruijn graph, with *k-1*-mers as nodes. Two nodes are connected if the *k*-mer formed by two overlapping *k-1*-mers is found in the list of genomic *k*-mers that was extracted from the sequencing data. Paths in the graph that don't have any branches represent contiguous sequences stretches in the template DNA. There is a sizeable number of general-purpose DBG assemblers, such as *Velvet* (Zerbino and Birney 2008), *SOAPdenovo2* (Luo et al. 2012) or *ABySS* (Simpson et al. 2009). As most DBG assemblers – analogous to most OLG methods – assume a uniform read coverage that is not found in many metagenomic data, some DBG assemblers are specifically designed for metagenomic data. These, like *MetaVelvet* (Namiki et al. 2012) or *IDBA-UD* (Peng et al. 2012), use additional information to find sub-graphs representing the different genomes in the overall DBG, for example coverage differences and graph connectivity. As DBG-based assemblers don't use a minimum overlap criterion but are mostly bound to a given *k*, the choice of the right value for *k* is central when using these methods. Finding an appropriate *k* is usually done by different rounds of trial-and-error, optimizing for a given assembly statistic. In absence of a known reference genome, a frequently used statistic is the contiguity as measured by the N50 size (Bradnam et al. 2013). Alternatively, an adequate *k* can be estimated through the overall number of observed *k*-mers that appear frequent enough to be likely of genomic origin rather than sequencing errors (Chikhi and Medvedev 2014; Q. Zhang et al. 2014). Assemblers that use multi-sized DBGs, such as *SPAdes* (Bankevich et al. 2012; Nurk et al. 2016), try to avoid the problem of having to choose a single "best" *k*, by iteratively increasing the size of *k*. This procedure allows the use of small values of *k* to assemble low coverage regions, while the subsequent use of larger *k* can enable the resolution of repetitive regions given that the coverage is high enough.

2.1.3 Choosing the right tools

The different assembly approaches come with their own benefits and drawbacks. While OLG methods can be useful for metagenome skimming data, where coverages for individual genomes can be low, their run time and memory requirements can limit their use for larger data sets. In contrast, while DBG-based methods are more time and memory efficient, their frequent need to optimize parameters that are dependent on read coverage often limit their applicability for complex metagenomes. This, along with the large and growing number of different *de novo* genome and metagenome assemblers, makes it challenging for users to decide which tool and parameters should be used. This is especially problematic, as different assemblers can yield different assemblies, even when using identical input data (Earl et al. 2011).

This problem has sparked numerous proposed solutions. There are various benchmarking approaches, such as the Assemblathons (Earl et al. 2011; Bradnam et al. 2013), GAGE (Schatz et al. 2012; Magoc et al. 2013), and the Critical Assessment of Metagenomic Interpretation (Sczyrba et al. 2017), which have evaluated assembler performance on simulated and real data sets. Furthermore, similar studies have been conducted for individual sequencing projects (Vollmers, Wiegand, and Kaster 2017; Z. Li et al. 2012; Kumar and Blaxter 2010). However, all these benchmarks differ in the sequencing method and chemistry employed; in sequencing library layouts and resulting read lengths/insert sizes; and in the complexity of the genomes under investigation. All of these factors influence key parameters of sequencing data, affecting different assemblers in different ways (Nagarajan and Pop 2013). As a result, different assemblers perform optimal on different data sets, making it hard to generalize assembler performance based on these benchmarks (Bradnam et al. 2013).

While tools have been developed to automate the large-scale creation of assemblies across a set of available genome assemblers (Mapleson, Drou, and

Swarbreck 2015), the creation of all these assemblies is still a time-consuming step, especially for large data sets. Additionally, these approaches rely on extensive sequencing data already being present, limiting how much the joint parameter space of different genomes, sequencing technologies and assembly methods can be explored.

2.1.4 Simulating metagenome skimming data from lichens

So far it has not been evaluated if and how metagenome-skimming data from lichens can be used for genome assemblies. Prior studies on lichen metagenomes were largely performed with data from a single *Illumina* or *454* sequencing library that was subsequently assembled using general purpose assemblers (Kampa et al. 2013; Lutsak et al. 2016; McDonald et al. 2013). As lichens feature at least one eukaryotic genome they frequently contain more repetitive non-coding regions than bacterial metagenomes in addition to the differences in species abundances that impact all metagenomic assemblies. Thus it is unclear how the selection of different assembly methods and parameters influences the extent to which the individual genomes can be reconstructed from the sequencing data. For this reason we introduce the idea of twin data sets, which are *in silico*-generated data sets of genomes that are expected to be of a similar composition and complexity as the targeted genomes. Such twin sets furthermore specifically mirror the sequencing parameters of a particular real data set. Simulated data has already been used for benchmarking assemblers (Mende et al. 2012; Earl et al. 2011). Additionally, a large number of programs are able to simulate whole genome shotgun sequencing reads, according to empirical sequencing error models for different sequencing techniques (Döring et al. 2008; McElroy, Luciani, and Thomas 2012; Richter et al. 2008; W. Huang et al. 2012; H. Li et al. 2009). By simulating from related organisms, our twin data sets have the benefit of not only being similar to the real data set in terms of library layout and

sequencing method, but should also mirror the expected genomic complexity to at least some degree. At the same time, the assemblies done on these twin data sets can be directly compared to the seed genomes from which the data was simulated. This enables assessing the influence of both the sequencing strategy and the genome assembly method on the quality of the genome reconstruction, as measured by contiguity, completeness and correctness (Gurevich et al. 2013). This allows one to rank the different methods according to a gold standard, thereby identifying which assembler is most capable to cope with the complexities of the twin sets. Assuming that our references from which we simulated are not substantially different to the target genomes, this ranking can then be used to inform the assembly of the real data. It furthermore allows teasing out the effects that different parameter choices have on the assembly outcome. Lastly, this approach can help in estimating the quality that can be expected from an assembly done on real data under a given sequencing strategy, thus enabling more targeted sequencing approaches. Given a real *Illumina* sequencing experiment, done on the lichen *Lasallia pustulata*, we thus simulate 11 twin data sets - using the lichenized fungi *Cladonia grayi* and its photobiont *Asterochloris sp.* as templates for the simulations. Subsequently, we demonstrate how such twin sets can guide the planning and execution of metagenome skimming projects in lichens and evaluate the impact of the metagenomic complexity on the assembler performance.

2.2 Methods

2.2.1 Metagenomic shotgun sequencing of *Lasallia pustulata*

We collected a *L. pustulata* thallus with a diameter of 10 cm in Olbia, Sardinia, Italy, in May 2013. From this, genomic DNA was extracted with the CTAB method (Cubero and Crespo 2002) and subsequently purified using the *PowerClean DNA Clean-Up Kit* (MO BIO, Carlsbad, CA, USA). A metagenomic shotgun library was constructed by the *Illumina TruSeq DNA Sample Prep v2* (Illumina, San Diego, CA, USA), selecting for fragments of a mean length of 450 bp with the *SPRIselect reagent kit* (Beckman Coulter, Krefeld, Germany). Members of the Schmitt group, at the Senckenberg Biodiversity and Climate Research Centre, performed the sampling and library preparation. The subsequent sequencing of 2 x 251 bp paired end reads was performed on an *Illumina MiSeq* machine by StarSEQ (Mainz, Germany). *FASTQC* (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for the quality control of the generated reads. Sequencing adapters were removed by *cutadapt* (M. Martin 2011) and overlapping read pairs were merged into longer super reads with *FLASH* (Magoč and Salzberg 2011) if reads had a minimum overlap of 10 bp. From the observed super read lengths we calculated the insert size distribution of the library by fitting a Weibull distribution in *R*.

2.2.2 Simulating lichen metagenome sequencing twin data sets

We concatenated the draft genome scaffolds of the lichenized fungus *Cladonia grayi* (http://genome.jgi.doe.gov/Clagr2/Clagr2.home.html) and its photobiont *Asterochloris sp.* (http://genome.jgi-psf.org/Astpho1/Astpho1.home.html), removing all ambiguous bases, thus creating two pseudogenomes to simulate sequencing data sets from. To characterize these pseudogenomes we assessed the self-similarity of them through *Gepard* (Krumsiek, Arnold, and Rattei 2007). Additionally, we annotated repetitive regions by generating libraries of

repetitive elements through *RepeatScout* v. 1.0.5 (Price, Jones, and Pevzner 2005) and gave these as input to *RepeatMasker* v. 4.0.5 (Smit, Hubley, and Green 2015).

We used *ART* v. 2.1.8 (W. Huang et al. 2012) to simulate whole-genome shotgun data sets from the pseudogenomes. Applying the *Illumina MiSeq-250* error profiles to simulate sequencing errors, we generated 2 x 250 bp paired-end reads. To model the insert size distribution observed in our real data, we did separate simulations with insert sizes ranging from 250 bp to 664 bp in step sizes of 1. Each individual simulation was performed with *ART*'s standard deviation for the insert size set to 1. We simulated a total of 13.9 million read pairs for each pseudogenome, following the average number of reads obtained through a single *Illumina MiSeq* run. Given the lengths of the pseudogenomes this yields mean coverages of 182x for the fungal and 125x for the algal genome. Furthermore we simulated 9 metagenomic *MiSeq* data sets, mixing reads from the pseudogenomes of *Asterochloris sp.* and *Cladonia grayi.* To estimate the influence of the coverage ratio between the genomes on downstream analyses we varied the fungal-to-algal coverage ratio from 1:9 (algal sequences are highly overrepresented) to 9:1 (fungal sequences are highly overrepresented) in steps of 1, while keeping the total number of read pairs for the joint data set close to 13.9 million.

2.2.3 Assembling the simulated data sets

We selected 6 *de novo* genome assemblers to assemble the 11 simulated data sets. These assemblers represent both Overlap Layout graph (OLG) as well as de Bruijn graph (DBG) based methods and additionally cover metagenomic and general assemblers (see Table 2-1). We preprocessed the reads in accordance with the individual program's instructions: For *MIRA* (Chevreux et al. 2004) and *Omega* (Haider et al. 2014), *FLASH*-merged super reads as well as unmerged read pairs were used as input, and all unmerged paired reads

were given for the other assemblers. While *MIRA* and *Velvet* (Zerbino and Birney 2008) come with internal error correction routines, we applied *bbmap* (http://sourceforge.net/projects/bbmap/) for correcting the reads prior to the assembly with *Omega*, and used *BayesHammer* (Nikolenko, Korobeynikov, and Alekseyev 2013) prior to the assembly with *SPAdes* (Bankevich et al. 2012). For *sga* (Simpson and Durbin 2010) we applied its internal error correction pipeline with a *k*-mer size of 31.

Table 2-1: The assemblers we evaluated and the respective methods they utilize.

	Assembler	Method	Description	Reference
OLG[1]	MIRA	Overlap Graph / Alignment	General purpose assembler	(Chevreux, Wetter & Suhai 1999)
	Omega	Overlap Graph	Metagenome assembler, uses coverage to discern between organisms	(Haider et al. 2014)
	sga	String Graph	Memory efficient implementation of overlap graphs	(Simpson & Durbin 2012)
DBG[2]	Velvet	DBG	General purpose assembler	(Zerbino & Birney 2008)
	MetaVelvet	DBG	Extension of Velvet for assembling metagenomics data, uses k-mer coverage to discern between organisms	(Namiki et al. 2012)
	SPAdes	Multisized DBG	Merges DBGs of different k mer sizes to handle differences in coverage	(Bankevich et al. 2012)

[1]OLG: Overlap layout graph; [2]DBG: de Bruijn Graph

Both *Omega* and *sga* require the specification of the minimum overlap length as an input parameter. Following the authors' recommendations for *Omega* we tested minimum overlap lengths between 100 and 250 in steps of 50. For *sga* we followed the example applications provided by the developers and tested minimum overlap sizes of 71, 75, 81, 91 and 101 bp. Similarly, *Velvet* and *MetaVelvet* (Namiki et al. 2012) require the user to provide the size of *k*. Here we used *VelvetOptimiser* v. 2.2.5 (https://github.com/Victorian-

Bioinformatics-Consortium/VelvetOptimiser/) to test *k*-mer sizes between 51 and 221 in step sizes of 10. In case of *Velvet*, we set *VelvetOptimiser* to optimize the coverage cutoff parameter to maximize the total number of base pairs in contigs larger 1 Kbp. For all 4 assemblers we observed the change of the N50 size for the different input parameters and used these parameters that maximized the N50 for each input data (Earl et al. 2011).

Following the documentation, the quick-flags *genome, denovo, accurate* were set for *MIRA*, while *SPAdes* was started with *–k 21,33,55,77,99,127 --careful*.

The assemblies were performed on a single machine with 4x Intel Xeon CPU E5-4607 @ 2.2 Ghz (24 cores, 48 threads) and 512 GB of RAM. Due to memory limits only a single size for *k* could be tested at a given time, instead of the standard parameter of four. As *sga* lacks a default setting, the number of threads was set to eight. For the other assemblers the default settings were used.

2.2.4 Evaluating the assemblies

We analyzed the resulting assemblies for the 6 different assemblers and 11 different data sets with *QUAST* (Gurevich et al. 2013), giving the pseudogenomes of *Asterochloris sp.* and *Cladonia grayi* as references. Additionally we checked for contigs that contain uniquely mapping sequences stemming from both genomes. For this we mapped the simulated reads back to the assemblies, using *bowtie2* (Langmead and Salzberg 2012), removing non-uniquely mapping reads using *samtools* (H. Li et al. 2009) and *grep*. The filtered mapping was then analyzed for the origin of the mapping reads using *pysam* (https://github.com/pysam-developers/pysam). *AUGUSTUS* (Stanke and Waack 2003) was trained with the 9588 *Cladonia grayi* reference genes and subsequently used to predict genes in the assembled genome sequences. We then matched the genes predicted in the assemblies to the reference genes with *blastp* (Altschul et al. 1997).

2.2.5 Assembling the metagenome of *L. pustulata*

We *de novo* assembled the *L. pustulata* metagenome from the *MiSeq* data set following the same protocol as described for the simulated twin data sets. We assigned the taxonomy of the resulting contigs by first comparing them to a custom database of the proteomes with *DIAMOND* (Buchfink, Xie, and Huson 2014). To differentiate between the eukaryotic symbionts and identify potential eukaryotic contaminations, our database consisted of 121 fungi, 16 plants and 8 animals (see Table A-1 on page 187 for the taxa included). We further enriched the database with 1,471 bacteria and 560 viruses downloaded from the *NCBI* genome database, randomly selecting one representative for each bacterial species if multiple strains were available. The results of the *DIAMOND* search were then used for the taxonomic assignment with *MEGAN4* (Huson et al. 2011), with *min-support = 1, min-score = 50, top-hit = 10%, no low complexity filtering* as parameters. For the assemblies of *MIRA* and *SPAdes*, we performed a *bowtie2* mapping of the MiSeq data to the contigs that could be taxonomically classified. The average coverages over the fungal, algal and bacterial contigs were used to estimate the mean coverages for the respective genomes, ignoring short contigs with a length smaller 2 Kbp.

2.3 Results

2.3.1 Simulating twin data sets mirroring a lichen metagenome

2.3.1.1 Sequencing Lasallia pustulata

Performing a single *Illumina MiSeq* sequencing run on a metagenomic shotgun library, we generated 14,013,249 read pairs with a read length of 251 bp from a single thallus of *Lasallia pustulata*. By merging 12,107,565 overlapping read pairs with *FLASH* (Magoč and Salzberg 2011) and subsequently analyzing the length distribution of the resulting super reads, we estimated the insert size distribution of our metagenomic library. Fitting a Weibull distribution, we found the mean insert size to be 336 bp with a standard deviation of 55 bp (Figure 2-1).

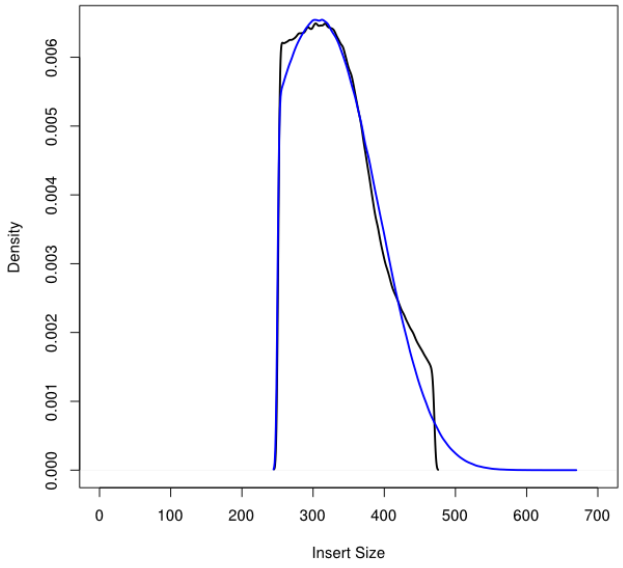


Figure 2-1: Observed distribution (black) for the insert sizes of the metagenomic *L. pustulata* shotgun library and the fitted Weibull distribution (blue curve). The fit extrapolates the insert size distribution for ranges where paired end reads no longer overlap.

2.3.1.2 Pseudogenome templates

We then simulated twin sets using our *L. pustulata* library, using the draft genomes of the lichenized fungus *Cladonia grayi* and its photobiont *Asterochloris sp.* as templates. Concatenating the 1,506 fungal and 153 algal contigs, we generated pseudogenomes for both, yielding a 38 Mbp fungal template (44% G/C) and a 55 Mbp algal template (58% G/C). The dot plot showed a rather large self-similarity in the pseudogenome of *Cladonia grayi*, which is pronounced towards the end, where the shortest contigs of the original assembly were concatenated (Figure 2-2 A). This finding was supported by our *RepeatMasker* (Smit, Hubley, and Green 2015) analysis, which found 5% of the overall sequence to be repetitive, with 4.2% of the genome falling in interspersed repeats. The visual inspection of the pseudogenome of Asterochloris sp. revealed less self-similarity (Figure 2-2 B), in line with the smaller fraction classified as repetitive (2.8%).

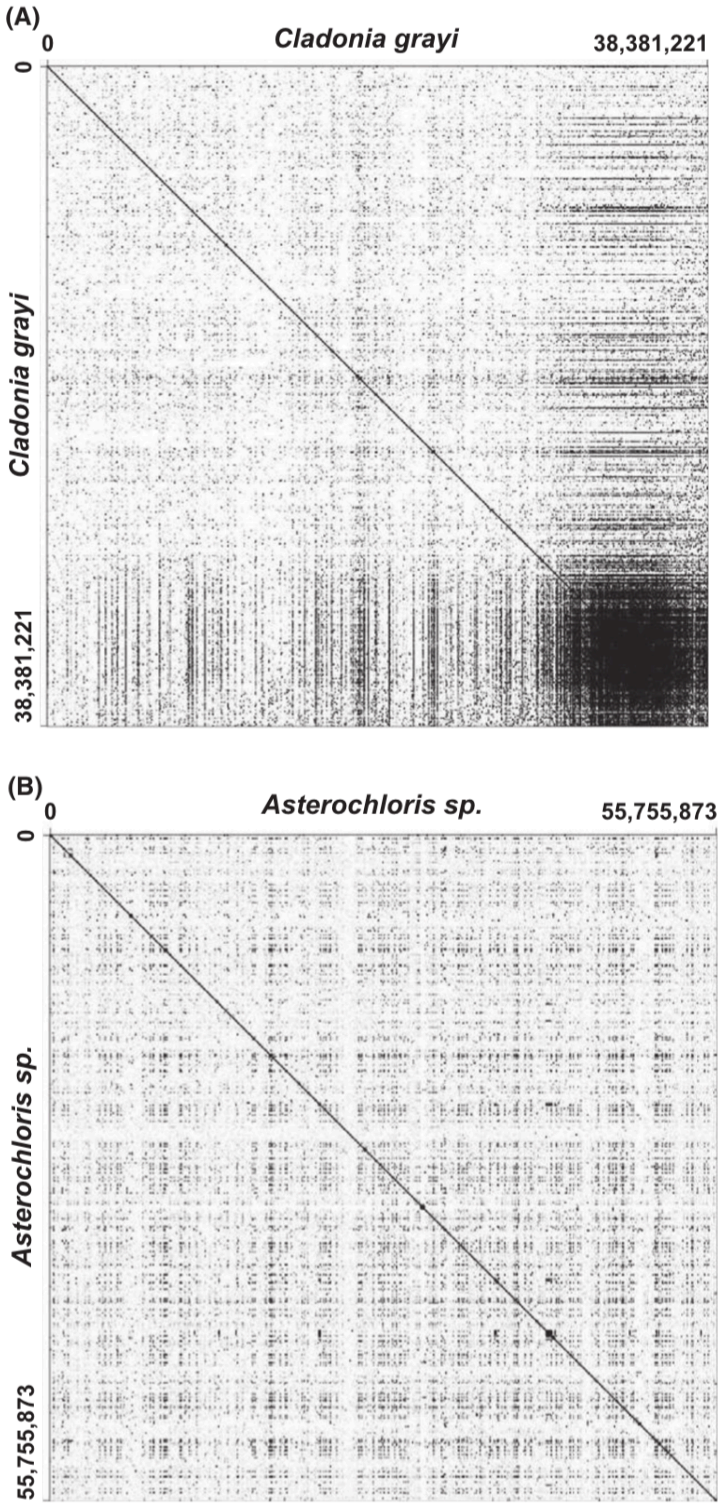


Figure 2-2: Dot plot for the pseudogenomes of *Cladonia grayi* (A) and *Asterochloris sp.* (B).

2.3.1.3 Twin data set creation

Based on the pseudogenomes of *Asterochloris sp.* and *C. grayi* we simulated read sets for both organisms. Furthermore, we also simulated nine data sets, mixing reads from both genomes in different coverage ratios (see Table 2-2). We assembled all twin data sets using six different *de novo* genome assemblers to benchmark their respective performance and sensitivity to different coverages in a given dataset. We selected three assemblers that are based on overlap layout algorithms (*Omega*, *MIRA* and *sga*) and three that are based on de Bruijn graphs (*Velvet*, *MetaVelvet* and *SPAdes*). With *MetaVelvet* and *Omega* the set includes two assemblers that are designed for the *de novo* assembly of metagenomic data, the others are general-purpose *de novo* assemblers.

Table 2-2: The eleven data sets simulated for the twin sets, with the number of reads used from *Asterochloris sp.* and *Cladonia grayi*. The differences in absolute coverages per organism are due to the different sizes of the two genomes.

Coverage Ratio *C. grayi* : *Asterochloris sp.*	Read Pairs *C. grayi*	Read Pairs *Asterochloris sp.*	Coverage *C. grayi*	Coverage *Asterochloris sp.*
0:10	-	13,976,839	0x	125x
1:9	993,072	12,983,735	13x	116x
2:8	2,052,139	11,924,657	26x	107x
3:7	3,184,074	10,792,727	40x	97x
4:6	4,396,524	9,580,266	56x	86x
5:5	5,698,537	8,278,247	74x	74x
6:4	7,100,343	6,876,442	92x	61x
7:3	8,613,884	5,362,851	112x	48x
8:2	10,253,117	3,723,658	134x	33x
9:1	12,034,317	1,942,453	157x	17x
10:0	13,976,783	-	182x	0x

2.3.1.4 *Assembler parameter optimization*

We explored the assembler parameter space of minimum overlap lengths for OLG-based methods (*sga* and *Omega*), and of *k*-mer sizes for DBG-based methods (*Velvet* and *MetaVelvet*) for the eleven twin data sets. We found that both *Omega* and *sga* consistently favor a single minimum overlap length for all twin sets. There is more variability in case of *Velvet* and *MetaVelvet*, with especially the latter showing larger differences between different metagenomic twin sets (Table 2-3). A drop in *k*-mer size is especially noticeable for the 7:3 and 8:2 data sets, when compared to the 6:4 and 9:1 sets.

Table 2-3: The selected parameter values yielding the highest N50 sizes for the individual assembler/data set combinations (All values given in base pairs).

Data Set	*Velvet* *k*-mer-size	*MetaVelvet* *k*-mer-size	*sga* overlap length	*Omega* overlap length
0:10	171	141	101	200
1:9	171	131	101	200
2:8	171	131	101	200
3:7	161	141	101	200
4:6	161	141	101	200
5:5	161	91	101	200
6:4	151	131	101	200
7:3	151	51	101	200
8:2	191	51	101	200
9:1	191	151	91	200
10:0	191	131	91	200

2.3.2 Baseline assembler performance on single-species data sets

In a first step we assembled the two single-species data sets, to establish a baseline performance of what the individual assemblers can achieve on a given genome skimming sequencing library, without the confounding factor of a species mixture. While all methods were successful in assembling the pseudogenomes over the full length, with no assembly differing more than 2% in length from the reference (Figure 2-3, columns 10:0 and 0:10), there

were marked differences in the contiguity and number of misassemblies between the two pseudogenomes and the different assemblers. In case of the algal genome both the NG50, as well as the assembly-error corrected NGA50, ranged from a high of >4 Mbp for *MIRA* to a low of only 0.28 Mbp for *sga* (Figure 2-3, column 0:10). We observed a similar spread for the fungal genome (Figure 2-3, column 10:0), though here the NG50 and NGA50 were generally smaller by an order of magnitude.

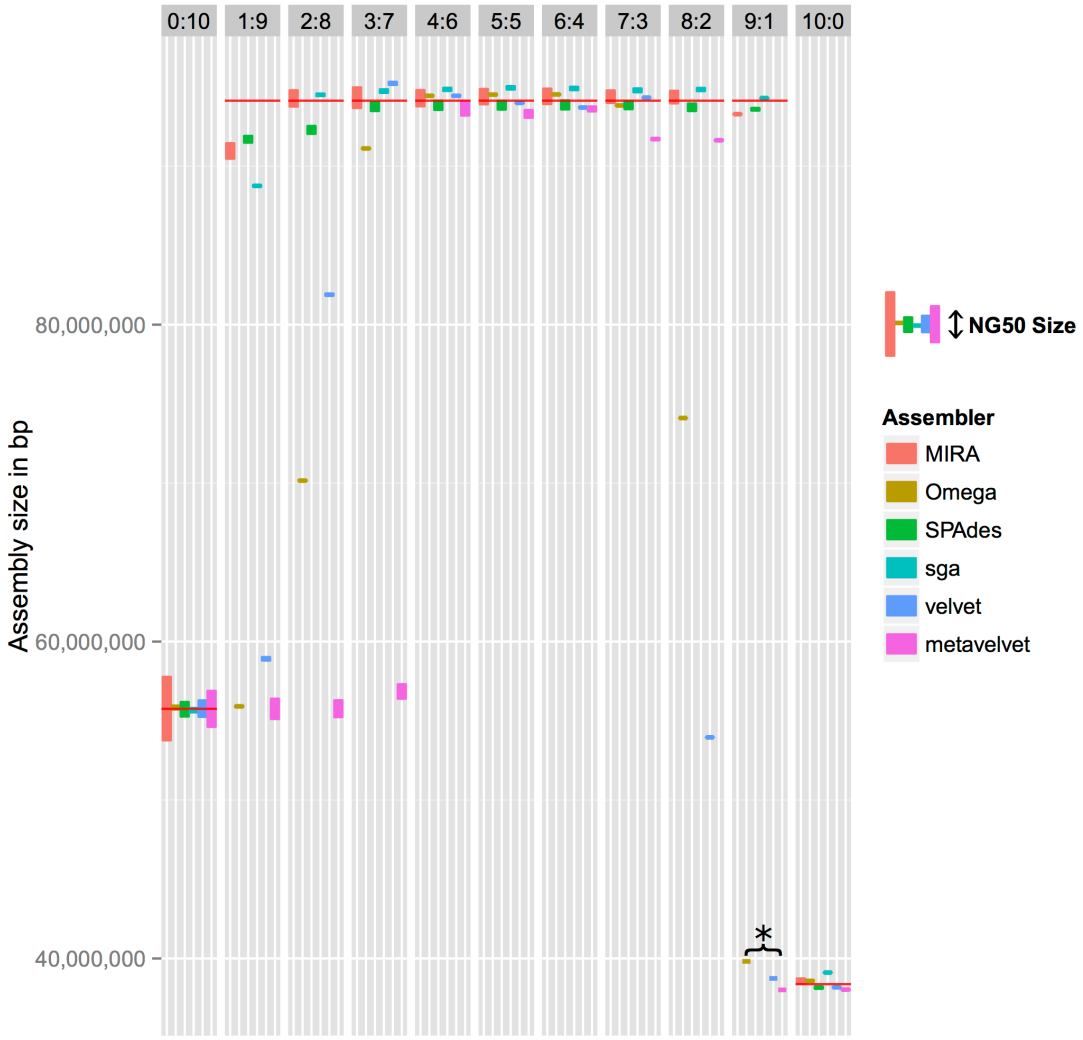


Figure 2-3: Assembler performance on the different data sets. Individual bars are centered at the total assembly lengths; the height of each bar represents the NG50 size. The orange lines give the reference genome lengths, 55.8 Mbp for the alga (0:10), 38.4 Mbp for the fungus (10:0) and 94.2 Mbp for the metagenomes (1:9 – 9:1), the NG50 is calculated with those numbers. The star denotes assemblies where the total assembly length is less than 50% of the target genome size, thus no NG50 could be calculated. The bar height was thus set to a default value.

We explored the reason for this difference by mapping the contigs from the six fungal genome assemblies as well as the original contigs to the fungal pseudogenome. We noticed that the borders of our assembled *C. grayi* contigs largely coincide with contig borders of the original assembly that was used to generate the pseudogenome (Figure 2-4A), leading to a highly similar contig length distribution. Doing the same for the *Asterochloris sp.* assemblies, we did not observe the same correlation of contig boundaries and contig length distributions (Figure 2-4B). We further investigated the potential reason for this correlation in the *C. grayi* assemblies by analyzing the repeat content in the respective pseudogenome. We observed that there is a considerable enrichment of repetitive elements within 100 bp of the contig borders. While the sum of contig borders makes up only 1 % of the total genome length, we found that 14.5% of all identified repeats fall in these regions (see Appendix, Figure A-1 on page 215).

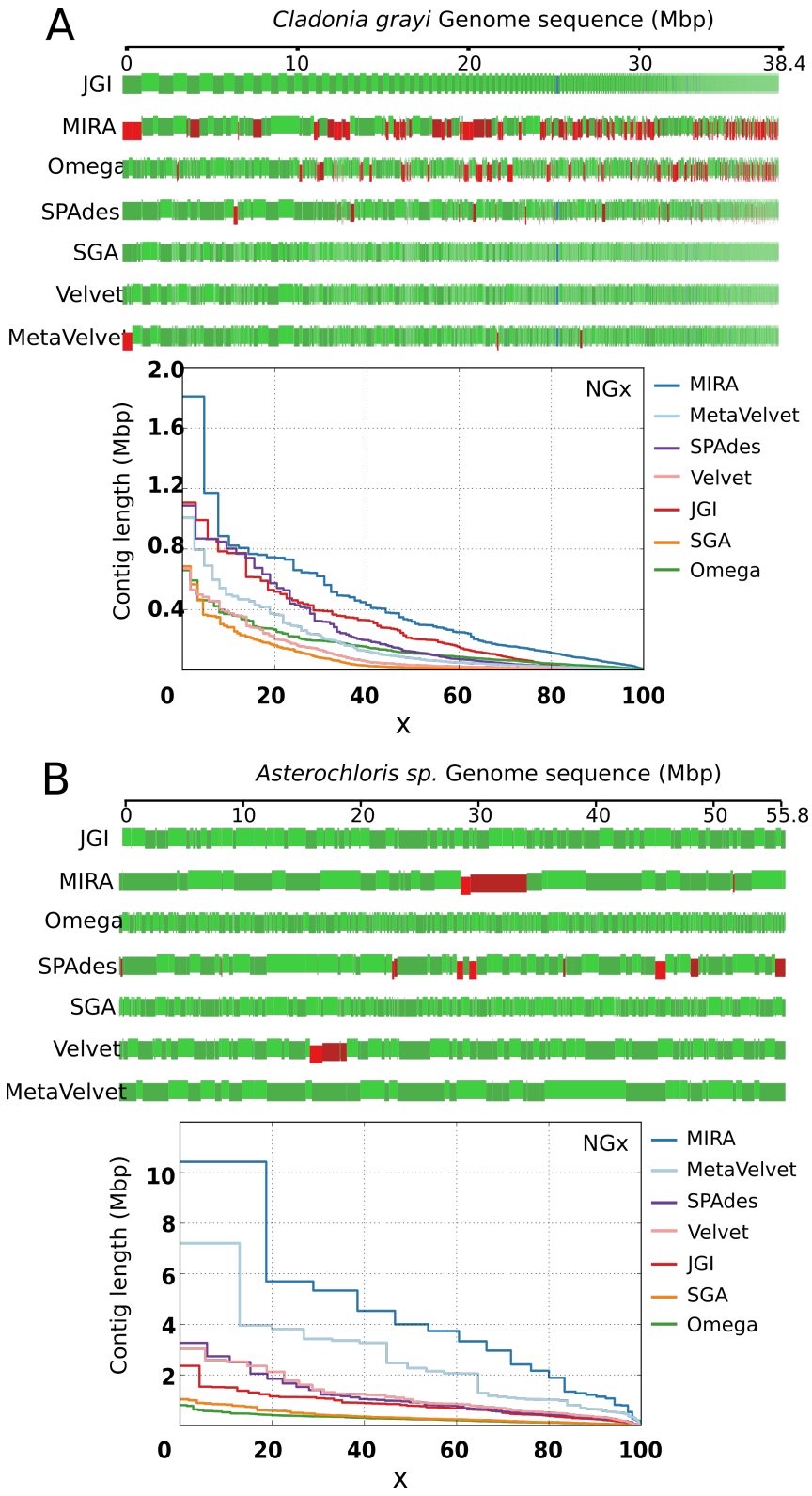


Figure 2-4: Mapping of the original contigs (JGI) and the contigs assembled from the *C. grayi* (A) and *Asterochloris sp.* (B) single species data sets. Red boxes indicate contigs with at least one misassembly. Erroneously fused contigs are split to map to the corresponding regions, boxes will be highlighted in red. The NGx plots give the length-ordered contig length distribution relative to the length of the target genome covered. All assemblies of *C. grayi* largely resemble the original input assembly, while the *Asterochloris sp.* assemblies in many instances extend over the contig borders of the JGI assembly.

In addition to the large differences in contiguity we also found substantial differences in the number of misassemblies between the assemblies of *C. grayi* and *Asterochloris sp.* data. While we did not find more than 10 errors in any of the algal assemblies, there were up to 258 misassemblies in the fungal assemblies (Table 2-4).

Table 2-4: The number of misassemblies per data set and assembler.

Data set	MIRA	Omega	SPAdes	sga	Velvet	MetaVelvet
0:10	1	1	5	0	1	0
1:9	839	3	753	27	45	2
2:8	414	178	639	1	231	3
3:7	339	246	172	0	535	57
4:6	318	200	96	0	377	24
5:5	130	222	96	0	40	30
6:4	126	216	86	0	2	13
7:3	94	311	94	0	4	124
8:2	99	182	116	0	9	135
9:1	151	117	295	0	8	5
10:0	112	258	77	0	3	2

We furthermore examined how method-dependent trade-offs between the contiguity and accuracy of an assembly influences the subsequent gene prediction. For this we predicted genes in all *C. grayi* assemblies with *AUGUSTUS* (Stanke and Waack 2003) and compared them to the gene predictions done on the pseudogenome (Table 2-5). *Omega* and *MIRA*, the two overlap-based assemblers, recovered the largest number of the reference genes, missing only 21 out of 10,740 genes, while all other assemblers were missing at least twice the amount. Furthermore, *MIRA* showed the smallest number of spurious gene predictions.

Table 2-5: Number of genes predicted in the different assemblies of *C. grayi*. The reference genes describe the genes predicted in the *C. grayi* pseudogenome.

Assembly	NGA50	# Gene predictions	# Reference genes recovered	# of additional gene predictions
Reference	n.a.	10,740	10,740	n.a.
MIRA	263 kb	10,835	10,719	23
Omega	101 kb	10,932	10,720	27
sga	16 kb	12,333	10,679	117
SPAdes	120 kb	10,903	10,694	33
Velvet	32 kb	11,068	10,678	30
MetaVelvet	77 kb	10,876	10,678	39

2.3.3 Assembling the metagenomic twin data sets

After estimating the baseline performance of the different assemblers on data sets of the individual genomes of *C. grayi* and *Asterochloris sp.* we continued to investigate the influence of mixing data from two eukaryotic species. To that end, we simulated data sets with different ratios of fungal to algal genomes (c.f. Table 2-2, page 26), as potentially encountered in metagenomic DNA isolates of lichens. We noticed that the performance of individual assemblers is highly affected by the species mixture in the metagenomic data sets, negatively affecting the total assembly length and the NG50 size. In all instances both assembly statistics became progressively worse when the coverage became skewed towards one of the two genomes (Figure 2-3, page 28).

MetaVelvet, and to a lesser extent *Velvet* and *Omega*, were strong examples for this trend. The *MetaVelvet* assemblies for the coverage ratios of 1:9 up to 4:6 only reached a length that is close to that of the algal genome. This indicates

that the fungal genome was not or only partially assembled. Similarly, *MetaVelvet* did not reconstruct the algal genome over the full length once the fungal-to-algal coverage skew exceeded 6:4. *MIRA*, *SPAdes*, and *sga* on the other hand appeared to be less sensitive to highly skewed coverage ratios in the input data, as their assemblies reached the expected sizes even if one genome is strongly underrepresented. We additionally found large differences between the individual assemblers when looking at the other metrics to assess assembly quality. In terms of NG50 size, especially *MIRA* performed well, regardless of the coverage ratio. For the ratios of 3:7 to 8:2 it outperformed all other tested assemblers (Figure 3). We observed the same when correcting for misassemblies by taking the NGA50, which splits contigs at misassemblies (see Table 2-4 for misassemblies found in each assembler/data set combination).

Only in three data sets, the algal-dominated 1:9/2:8 sets, and the fungal-dominated 9:1 set, *MIRA* did not generate the assemblies with the largest NG(A)50. In case of the algal-dominated data *MetaVelvet* achieved higher NG(A)50 values. Comparing the total assembly lengths (Figure 2-3, page 28), we found that these higher NG(A)50 values for *MetaVelvet* are achieved through the exclusion of the fungal genome, which remained unassembled. For the fungal-dominated 9:1 data set *SPAdes* outperformed *MIRA* for both the total assembly length as well as the NG50 ($MIRA_{\text{NG50}}$: 33 Kbp, $SPAdes_{\text{NG50}}$: 154 Kbp.; $MIRA_{\text{Assembly length}}$: 93,285,544, $SPAdes_{\text{Assembly length}}$: 93,594,650).

Additionally, we searched for chimeric contigs in all assemblies that consist of both fungal and algal data. For this, we mapped our simulated sequencing reads back to the individual assemblies. We then searched for contigs that have uniquely mapping reads of both the fungal and the algal fractions. In the 229,256 contigs, spread over 54 assemblies, we identified only a single chimeric one. This misassembly is found in a 3,645,792 bp long contig

generated by the *MIRA* assembly of the 6:4 data set. A closer inspection of this contig revealed a 32 bp long dinucleotide repeat, found in both a fungal and an algal read, to be the source of this wrong join of the two contigs from different species.

Table 2-6: The number of *C. grayi* reference genes recovered for the individual assembler/coverage ratio combinations.

Assembly	1:9	2:8	3:7
MIRA	10,348	10,715	10,718
Omega	72	5,825	10,302
sga	10,100	10,674	10,675
SPAdes	10,656	10,666	10,683
Velvet	2,845	8,817	10,530
MetaVelvet	4	66	1,657

In a last step, we analyzed how the downstream gene prediction is affected by the different genome coverage ratios in the sequencing data, and the choice of the assembler (Table 2-6). We limit our investigation to the fungal genome for the 1:9 to 3:7 data sets, as these show the highest differences in assembler performance, with *Omega*, *Velvet* and *MetaVelvet* having total assembly sizes much smaller than expected. As a consequence, we find that the assemblies of *MetaVelvet*, and to a lesser extent *Velvet*, and *Omega*, contained only a miniscule fraction of the reference genes of *C. grayi*. On the other hand, large parts of the reference genes were recovered from the assemblies done with *MIRA*, *SPAdes* and *sga*.

2.3.4 Assembling the metagenome of *Lasallia pustulata*

We evaluated the performance of the individual assemblers on our metagenome skimming data set from *Lasallia pustulata*. In the *L. pustulata* data *Omega*, *Velvet* and *MetaVelvet* generated assemblies with total lengths between 34 and 43 Mbp, while *Mira* and *SPAdes* yielded assemblies of 2x – 3x the size (Table 2-7). We noticed that the results largely resemble the findings of the 9:1 twin data set analysis.

Table 2-7: Results of the six assemblers on the *L. pustulata* data set. The N50-optimized parameters are given where applicable. Parameter optimization for *sga* and *Omega* was done for the minimum overlap length. For *Velvet* and *MetaVelvet* the *k*-mer size was optimized.

	MIRA	***Omega***	***sga***	***SPAdes***	***Velvet***	***MetaVelvet***
Parameter (bp)	-	150	101	-	211	191
Number of contigs	11,758	4,438	14,164	15,530	3,532	2,434
Total length (Mbp)	75.7	43.2	68.9	140.1	36.0	33.6
Largest contig (bp)	520,743	180,979	256,303	1,303,928	115,444	144,213
N50 (bp)	11,323	15,584	5,529	18,221	15,703	19,617

After looking at the overall assemblies, we performed a taxonomic assignment on the different contig sets to estimate the assembly sizes and assembly contiguities of the different taxonomic fractions (Table 2-8). We found that there are substantial size differences for the fungal genome; on the lower end *sga* reconstructed only 27.7 Mbp in total (*sga*), compared to 39.6 Mbp for *SPAdes* on the upper end. The assembly size spread is even more extreme for the algal genome, where the total assembly length spanned from 0.17 Mbp generated by *Velvet* to 40.7 Mbp found in the *SPAdes* assembly. In addition to this we uncovered a sizeable bacterial fraction in the *Lasallia pustulata* metagenome, which ranged from 0.11 Mbp in the *MetaVelvet* assembly to 43.5 in the *SPAdes* contigs. Overall, we found that *MIRA* and *SPAdes* are best at recovering large fractions of the metagenome in their respective assemblies, with *SPAdes* leading in terms of both the assembly

lengths across the different taxonomic fractions as well as the N50 size, followed by *MIRA* as the runner-up. A subsequent mapping of our sequencing reads against the fungal and algal assembly fractions of *SPAdes* and *MIRA* revealed that the coverage distribution has its peak at 110x for the fungal and ~10x for the algal genome (see Appendix, Figure A-2 on page 215).

Table 2-8: Assembly results for the different taxonomic fractions present in the *L. pustulata* metagenome. The largest contig and N50 are given in bp.

		MIRA	***Omega***	***sga***	***SPAdes***	***Velvet***	***MetaVelvet***
Fungi	**# contigs**	2,561	2,476	3,794	4,069	2,572	2,300
	Total length	34.1 Mb	30.2 Mb	27.7 Mb	39.6 Mb	32.3 Mb	32.63 Mb
	Largest contig	161,762	134,352	61,191	197,284	115,444	144,213
	N50	21,046	16,495	9,662	22,506	17,160	19,704
Algae	**# contigs**	3,008	208	3,579	2,606	17	15
	Total length	10.3 Mb	0.76 Mb	12.1 Mb	40.7 Mb	0.17 Mb	0.19 Mb
	Largest contig	26,589	101,190	21,759	129,862	63,433	63,393
	N50	3,382	3,229	3,372	23,085	30,739	20,075
Bacteria	**# contigs**	2,667	1,355	3,011	6,707	753	10
	Total length	19.4 Mb	10.5 Mb	17.2 Mb	43,5 Mb	2.5 Mb	0.11 Mb
	Largest contig	520,743	180,979	256,303	1,303,928	23,430	25,282
	N50	13,081	13,303	7,093	9,280	3,308	20,051

2.4 Discussion

Genome skimming has previously been shown to be an effective way to quickly assess the genomes of individual species at a low cost (Elgar et al. 1999; Malé et al. 2014; Bock et al. 2014). Additionally, metagenome skimming is now being applied to get insights into symbiotic communities, as found in lichens (McDonald et al. 2013; Sigurbjörnsdóttir et al. 2015; Grube et al. 2015). Our analysis provides insight into whether the data generated by such metagenome skimming approaches can be used for fine-grained comparative genomics studies and to what extent the answer to that depends on the choice of methods in processing this data. We additionally explore to what extent the composition of the data itself influences the potential assembly outcomes, as previously found in other genome assembly contexts (Earl et al. 2011; Bradnam et al. 2013; Deng et al. 2015). As the choice of optimal assembler is dependent on the underlying data that is being reconstructed, benchmarks cannot be generalized for the assembly of all data sets, limiting the applicability of general benchmarking efforts (Earl et al. 2011; Bradnam et al. 2013).

To minimize the biases from taxon-specific idiosyncrasies like G/C content or repeat content in the respective genomes we tried to generate twin data sets based on the already sequenced genomes of lichen symbionts. Given that the lineages of *L. pustulata* and *C. grayi* split about 250 million years ago (Amo de Paz et al. 2011), the twin data for the mycobiont can nevertheless only approximate the *L. pustulata* genome. These twin sets are furthermore generated to model the observed parameters from a given data set in terms of sequencing method and library layout. While the use of real data over simulated data for this kind benchmarking is generally preferable, the twin data come with their own benefits: As they can be generated quickly *in silico*, the twin sets facilitate rapidly testing various library layouts and combinations of different sequencing techniques. Additionally the use of

known reference genomes, which are ideally closely related to the genome of interest, allows evaluating the assembly results to a known ground truth of similar genomic complexity. We apply the idea of the twin data sets to a benchmark of assembler performance on lichen metagenomes. Using our metagenome skimming data from the lichen *Lasallia pustulata* we estimated the parameters of the *Illumina MiSeq* library. With these parameters we generated eleven twin data sets that mix simulated sequencing reads of the lichenized fungus *Cladonia grayi* and its algal photobiont *Asterochloris sp.* in different ratios.

2.4.1 Assembling single-species twin sets

Starting with the two data sets that contain only data of a single species, we established the baseline assembler performance for the individual genomes. This revealed that there are marked differences in the quality of the genome reconstructions between different assemblers, even when done on the same input data. While all assemblers are capable of assembling the full lengths of the genomes, the contiguity of these assemblies varies widely between methods. *MIRA* (Chevreux, Wetter, and Suhai 1999), the best-performing assembler for both the algal as well as the fungal genome, achieves N50 and NG50 values that are twice as large as the second-best methods, *MetaVelvet* (Namiki et al. 2012) and *SPAdes* (Bankevich et al. 2012) respectively.

An earlier assembler benchmark on bacterial genomes had found that *SPAdes* outperforms *MIRA* (Magoc et al. 2013), highlighting the influence that the underlying data has on the performance of individual tools (Bradnam et al. 2013). This point is also driven home by the differences in assembly quality between the fungal and algal genome. The N(G)50 sizes for the algal genome are nearly an order of magnitude larger than that for the repeat-rich fungal genome (Figure 2-3, page 28), despite the latter having a larger mean coverage (c.f. Table 2-2, page 26). It appears that the repetitive regions (Figure 2-2, page

25) cannot be resolved by any of the assemblers due to the lack of long-range information like mate pair libraries. This shows the importance of choosing sequencing strategies that are appropriate for the complexity of a given genome. Despite this, the relative assembler ranking remains largely unchanged, hinting that the choice of appropriate sequencing methods has a larger impact than the intrinsic characteristics of a given genome.

The comparison of the fungal assemblies additionally allowed us to evaluate how the different kinds of assemblers treat repetitive regions. Classical DBG assemblers such as *Velvet* and *MetaVelvet* use a single value for *k*, thus adopting a conservative strategy. The ideal choice of *k* needs to strike a balance between resolving repetitive regions with a large value of *k* and avoiding disruptions in the graph caused by low coverage regions, achieved by small values for *k*. Following the graph simplification, only unambiguous parts will be returned as contigs (Zerbino and Birney 2008). This strategy helps to avoid repeat-induced misassemblies, at the cost of low NG50 values in repeat-rich genomes (Table 2-4 on page 31, Figure 2-3 on page 28). Overlap-based assemblers, such as *MIRA* and *Omega*, and assemblers that use multiple values of *k* (*SPAdes*) on the other hand are more capable at resolving repeats, leading to higher NG50 values at the cost of introducing additional misassemblies (c.f. Table 2-4 on page 31, Figure 2-4 on page 30). This increase in observed misassemblies does not necessarily reflect a larger misassembly rate. Even given a constant misassembly rate the total number of misassemblies would linearly increase with the number of sequences that the assemblers join. Indeed, after splitting those misassemblies we find that the resulting NGA50 values remain larger for those assemblers that generate larger NG50 values, when compared to the more conservative DBG assemblers (see Appendix, Figure A-3 on page 216). Thus, we do not find evidence for a marked increase in the misassembly rate for *MIRA* and *SPades*.

In line with the previous observation, when comparing the *AUGUSTUS* gene predictions done on the different assemblies we find that an increase in NG50 leads to an increase in the number of reference genes found despite the higher number of misassemblies. At the same time the higher NG50 decreases the number of genes being split on different contigs. This is despite the small average gene size of 1.8 Kbp for *C. grayi*, which will make it unlikely that a contig break will fall into the middle of gene. Rather most contig break points will fall into intergenetic regions, thus not affecting the gene prediction. Given the low occurrence of misassemblies (112 events over a length of 38 Mbp for *MIRA*), we hypothesize the benefit of longer NG50 values on gene predictions to increase with average gene length. We thus find evidence that, while these misassemblies might be generally unwanted, these are an acceptable trade-off in the case of metagenomic skimming.

2.4.2 Assembling metagenomic twin sets

The assembly of metagenomic data comes with additional complexities not present in single-species data. Most importantly, we find that the common practice of N50 maximization for selecting the optimal parameter choice (Cha and Bird 2016; Earl et al. 2011) fails when the coverages between the individual species are highly uneven. Given the lack of a known reference, the NG50 most often cannot be used, especially when working with metagenomic data, where the true total genome length is highly dependent on the species present in the data. We show that the practice of maximizing the N50 can lead to the partial or even complete exclusion of the underrepresented genome from the assembly (Figure 2-3, page 28). This not only happens for DBG assemblers like *Velvet* or *MetaVelvet*, but also for overlap-based assemblers like *Omega* that require the user to set a minimum overlap length. While the intention of the N50-maximizing parameter choice is to choose values that allow the bridging of repeats, it can alternatively choose values that favor

values that leave those regions unassembled which would lead to short contigs. This prevents the formation of short contigs, at the cost of missing out on parts of the genome.

The N50-maximizing parameter choice for the 11 *MetaVelvet* assemblies (Table 2-3 on page 27), compared to the overall assembly lengths (Figure 2-3) gives an idea of this interdependence between the choice of *k*, read coverage, *k*-mer coverage and the resulting assembly (Chikhi and Medvedev 2014; Compeau, Pevzner, and Tesler 2011). In the data sets with read coverages ranging from 13x – 40x for the fungal genome (twin sets 1:9–3:7) basically the same value of *k* is chosen as for the data set that only consists of algal data (0:10). For these large values of *k* the *k*-mers of the fungal genome are found only so rarely, that their *k*-mer coverage is more or less the same as for the *k*-mers that are generated by sequencing errors. Consequently they are excluded as errors during the assembly procedure. We observe a similar effect in the data sets where the less repetitive algal genome is underrepresented. In the 7:3 and 8:2 data sets, a *k* of 51 is still sufficient to assemble the algal genome, despite the low read coverage of 48x and 33x respectively. Only in the 9:1 data set the algal read coverage drops to 17x and the parameter optimization leads to a jump to 151 for the the value of *k*, thus preventing the formation of short algal contigs (Figure 2-5).

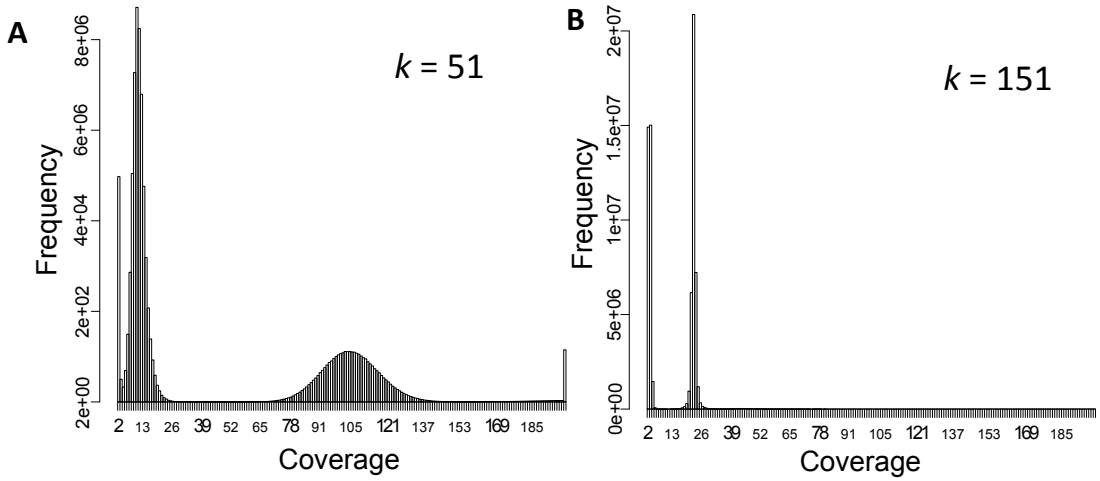


Figure 2-5: *k-mer* coverages for the 9:1 twin set. A *k* of 51 results in a clear bimodal distribution, with the two peaks representing the fungal and algal *k-mers* and a peak of 1 representing the sequencing errors (A). Increasing the *k*-mer size to 151 (B) shifts the distribution to the left, making the lower-coverage algal *k*-mers overlap with the sequencing error *k*-mers.

We further evaluated this effect by doing a *MetaVelvet* assembly of only the fungal data from the 9:1 twin set, again optimizing the value of *k*. Under these circumstances a *k* of only 131 is found as optimal, generating an NG50 size of 70 kb, which is about 18 kb longer than what is achieved with the mixed species data and a *k* of 151. Thus a selection against the formation of short contigs comes at the cost of a suboptimal assembly even for the overrepresented genome.

Given these results the parameter selection for metagenomic data should ideally not only take the N50, but also expected assembly size into account. In case of the 9:1 data set and *MetaVelvet*, a *k* of 51 does lead to an assembly that covers 99% of both genomes, with an N50 of 20 kb. Unfortunately, this approach is infeasible for most metagenomic data sets, as the joint target genome size is unknown, even with lichens, whose metagenome are supposedly simple, as they host a large number of different bacteria (Cardinale, Puglia, and Grube 2006; Erlacher et al. 2015; Grube et al. 2015). Possible solutions to this problem are the inspection of the *k*-mer coverage histograms as well as taking the number of unassembled/unmapped reads

into account. Lastly, the selection of an assembler less sensitive to uneven coverages can be helpful.

2.4.3 Assembling the *Lasallia pustulata* metagenome skimming data

Based on our study of assembler performance amongst the twin data sets we expect the best genome reconstructions for our *Lasallia pustulata* metagenome to be generated by *MIRA* – if the coverage ratio skew is not too extreme – or *SPAdes*, if the coverages in the metagenome are very uneven. Evaluating all six assemblers on this data set we find that *SPAdes* outperforms all other assemblers, with *MIRA* in second place (Table 2-7, Table 2-8). As the assembly-based coverage estimation for the fungal and algal genome shows an average read coverage of 110x for the former and of only 10x for the latter, we find that these results are fully in line with our predictions based on the simulated twin data sets.

Given the twin data sets we did expect the *SPAdes* assembly to facilitate a comprehensive characterization of the genomes involved in the lichen *L. pustulata*. As the assembly length for the fungus *Lasallia pustulata* is in line with what we expected given other lichenized fungi (Wang et al. 2014; McDonald et al. 2013), we assume that most (if not all) of the genome could be assembled. In contrast, the N50 size of 23 Kbp that we could achieve for the fungal genome is smaller than the expected value of ~120 Kbp (Figure 2-3, page 28). This might either be a result of a higher repeat content or in *L. pustulata* when compared to the pseudogenome of *C. grayi*, due to the higher complexity of the assembly problem introduced by the presence of bacteria, or a combination of both. The presence of bacteria might also be the reason why the algal genome could only be assembled partially. With a total assembly length of 41 Mbp it is around 13 Mbp smaller than we would expect given the length of the *Asterochloris sp.* genome. Additionally, a preliminary gene annotation of the draft genome sequence of *Trebouxia sp.* yielded only

around 5,100 genes of an average length of 460 bp, in stark contrast to the 10,025 genes (average length 1,379 bp) found in the draft genome of *Asterochloris sp.* (http://genome.jgi-psf.org/Astpho2/Astpho2.info.html). As about 1/3rd of all contigs are of bacterial origin, these do not only potentially increase the assembly complexity, but they also decrease the overall coverages for the fungus and alga. The low coverage could either result in parts of the genome not having being sequenced deeply enough or in contigs that are too short to reliably being identified as algal by *MEGAN* (Huson et al. 2011). Additional sequencing to increase the overall coverage for the algal genome will need to be performed to recover the full length of the *Trebouxia sp.* genome.

2.5 Conclusions

We have shown the suitability of simulated twin data sets to benchmark the performance of six different methods when assembling metagenome skimming sequencing libraries. This approach allows estimating the effectiveness of given combinations of assembly method, sequencing library layout and target genomes before undertaking extensive and expensive sequencing. We conclude that genome skimming can be suitable to facilitate a preliminary assembly and analysis of a metagenome. In fact, as long as the genome coverage ratios are close to being even, the choice of the assembler does not have a marked influence on the results of the genome assembly. However, this changes once the coverage ratios become skewed to a single organism. We found that some assemblers are highly sensitive to such uneven genome coverages in the metagenomic data, especially when a naïve N50-maximization criterion is used to optimize the input parameters. This already applies in simple metagenomes where only two genomes are present.

The assemblies done on the lichen metagenome of *Lasallia pustulata* revealed that the genome ratios inside the lichen are indeed highly skewed and contain a sizeable fraction of bacterial sequences in addition to the fungus and alga, adding to the complexity of the assembly problem.

3 Assembly & characterization of the *L. pustulata* metagenome

3.1 Introduction

Our initial exploration of the metagenome of the lichen *Lasallia pustulata* revealed that the species abundances of *L. pustulata* and *Trebouxia sp.* appear to be highly skewed towards the mycobiont. Our twin data set-based exploration showed that such extreme differences in abundances are hard to assemble, leading to fragmented and incomplete genome reconstructions. While we found that these consequences are more marked for the underrepresented genome, it even affects the highly covered genome, albeit to a lesser extent. This was reflected in the preliminary assembly done on the *L. pustulata* metagenome skimming data, which yielded only a heavily fragmented and probably incomplete genome for the photobiont *Asterochloris sp.*. The assembly of the mycobiont *L. pustulata* on the other hand reached the expected genome size, but remained largely fragmented as well. Given the results of the twin data sets we do not expect that further assembly methods based only on the metagenome skimming data would substantially improve the genome contiguity and completeness.

Furthermore, we found that a sizeable fraction of sequencing reads belonged to bacterial taxa. This is not unexpected, as studies have shown associations of bacteria with lichens (Aschenbrenner et al. 2016). Bacterial communities have been described for a variety of lichens (Cardinale, Puglia, and Grube 2006; Cernava et al. 2015; Hodkinson et al. 2012; C. H. Park et al. 2016), with growing evidence that these plays a functional role in the lichen symbiosis (Erlacher et al. 2015; Grube et al. 2015; Kampa et al. 2013; Sigurbjörnsdóttir et al. 2015; Cernava et al. 2017). Thus a hologenome-wide perspective should be included when studying the lichen symbiosis.

3.1.1 Lichen microbiomes

The composition of lichen microbiomes has been found to be influenced by their habitat and the type of photobiont present in them, with the orders of Acidobacteriales, Rhodospirillales, Rhizobiales and Sphingomonadales being dominant in different taxa (Hodkinson et al. 2012; C. H. Park et al. 2016). Investigations of the microbiome of *Lobaria pulmonaria* found that the Rhizobiales potentially play a functional role in auxin & vitamin production, nitrogen fixation, and stress protection (Erlacher et al. 2015). Furthermore, an antagonistic potential of the *L. pulmonaria* microbiome against pathogens was found (Cernava et al. 2015), while Chthoniobacterales are hypothesized to be additionally involved in the protection against oxidative stress (Cernava et al. 2017). The Rhizobiales, which were found as key contributors in most lichens from temperate climates (Hodkinson et al. 2012), were rarely found in Antarctic lichens, which instead feature members of the Acidobacteria (C. H. Park et al. 2016). The Acidobacteria are a widespread and phylogenetically diverse phylum, with the first representative, *Acidobacterium capsulatum*, only recently discovered in 1991 (Kishimoto, Kosako, and Tano 1991) and the phylum classification following in 1997 (Kuske, Barns, and Busch 1997). So far, the ecology of the Acidobacteria is not well characterized, partially due to the difficulty of growing them in culture (Kielak et al. 2016). Work on arctic tundra soil hypothesized that they utilize and synthesize diverse polysaccharides and are resilient to the fluctuating temperatures and low-nutrient conditions in those habitats (Rawat et al. 2012). While genomic studies have found hints for the use of nitrite as nitrogen source and responses to soil nutrients and soil acidity, physiological evidence for this is lacking so far, due to the poor culturability of Acidobacteria (Kielak et al. 2016).

3.1.2 Improving *de novo* assemblies of metagenomes

The complexity of the lichen hologenome, with eukaryotic symbionts as well as a diverse microbiome, has so far hindered the comprehensive sequencing of lichen metagenomes, leading to fragmented genome reconstructions (McDonald et al. 2013). For this reason, research has largely focused on the genomes of those lichenized fungi that can be grown in culture (S. Y. Park et al. 2014; S. Y. Park, Choi, Kim, Jeong, et al. 2013; S. Y. Park, Choi, Kim, Yu, et al. 2013; McDonald, Gaya, and Lutzoni 2013). While this facilitates the sequencing of the respective lichenized fungi, it only represents a fraction of the lichen diversity and does not allow an analysis of the hologenome. The use of new, third-generation sequencing methods can help to overcome these limitations. Sequences generated by these methods, while usually more error-prone than short-reads, can reach lengths of some 10-100 Kbp in the case of SMRT sequencing with *PacBio,* or even up to 882 Kbp with *Nanopore* sequencing (H. Lee et al. 2016; Jain et al. 2017). These methods have largely been used to generate assemblies of large genomes as wheat (Clavijo et al. 2017), or for the assembly of phased human genomes (Chin et al. 2016). Nevertheless, these techniques are also successfully being used for metagenomic data (Tsai et al. 2016; Edwards et al. 2016; Frank et al. 2016; Driscoll et al. 2017), where the long read length improves assembly contiguity, facilitating the downstream analysis.

Due to the high error rates of third-generation sequencing, the data generated by these require processing for error removal (Bleidorn 2015). While there are different methods, their applicability depends highly on read coverage of the sequencing data. If the long-read coverage is high enough, the data can be intrinsically error corrected by generating consensus sequences prior to the assembly (Chin et al. 2016; Koren et al. 2017). In the absence of such high coverages less error-prone short read data can be used to extrinsically correct these errors (Berlin et al. 2015). In the case of metagenomic data this can be

challenging, as especially lichens contain species in markedly different abundances (Hodkinson et al. 2012), rendering assemblies done purely on long reads infeasible. At the same time, relying only on hybrid approaches does not make use of most of the benefits of long-reads. It has been proposed that meta-assemblies, which merge the results of different assemblies, can be a solution to facilitate the reconstruction of genomes (Wences and Schatz 2015; Chakraborty et al. 2016).

Here we demonstrate that a joint use of second and third generation sequencing data to assemble a lichen metagenome is feasible. Through a hybrid assembly strategy, using different assembly methods and subsequently merging them, we can generate highly contiguous, full-length genome reconstructions for the both the mycobiont and photobiont of *L. pustulata*. Additionally, we can also assemble large fractions of the lichen microbiome into contiguous sequences, facilitating a further microbiome characterization. Furthermore, we compare the bacterial microbiomes between 9 geographically different samples of *L. pustulata*, finding a stable bacterial microbiome that potentially supports the lichen.

3.2 Methods

3.2.1 Estimating fungal-to-algal genome ratios

Following up on the preliminary estimates for the ratio of fungal to algal genomes in thalli of *L. pustulata*, we further analyzed these ratios through quantitative polymerase chain reactions (qPCR), performed on single copy genes of the fungus (a MCM subunit) and the alga (the COP-II coat subunit). All steps of the fungal-to-algal ratio estimation were performed by the group of Thomas Hankeln, at the Institute of Molecular Genetics of the Johannes Gutenberg University in Mainz, Germany. We measured the DNA concentrations of four thalli with the *Qubit dsDNA High Sensitivity Kit* (Life Technologies) according to the manufacturer's protocol. For the qPCRs, we used the *GoTag qPCR Master Mix* (Promega) with a total volume of 10 µl and performed a three-step PCR protocol with an annealing temperature of 55°C on an *ABI 7500 Fast Real Time PCR system cycler (*Applied Biosystems).

Each sample was measured in three technical replicates, with each assay being measured in triplicate. Total copy numbers were estimated by a standard curve approach (Nolan, Hands, and Bustin 2006) with serial ten-fold dilutions of plasmids that were engineered to contain only single copy PCR templates. From this we calculated mean quantities for each run, as well as the mean quantities and standard deviations across the three technical replicates.

3.2.2 Sample collection and sequencing strategies

In addition to the *Illumina MiSeq* library generated from a single thallus of *Lasallia pustulata*, as discussed in 2.2.1 (page 17), we collected further thalli near Olbia (Sardinia, Italy) and Orscholz (Saarland, Germany) between May 2013 and December 2014. We constructed additional libraries from these thalli, to perform sequencing on *Illumina HiSeq*, *Illumina MiSeq* and *PacBio RS II* machines. The sample collection, DNA isolation and the library preparation

for *Illumina MiSeq* sequencing was done by members of the Schmitt group, at the Senckenberg Biodiversity and Climate Research Centre in Frankfurt am Main, Germany. GenXPro GmbH (Frankfurt am Main, Germany) performed the library preparation and subsequent sequencing for the *Illumina HiSeq.* The group of Yahya Anvar at the Leiden Genome Technology Center of Leiden University Medical Center (Netherlands) did the library preparation and subsequent sequencing with the *PacBio RS II*.

For the whole genome samples we extracted the DNA using the CTAB method (Cubero and Crespo 2002) and subsequently purified it with the *PowerClean DNA Clean-Up Kit* (MO BIO, Carlsbad, CA, USA) according to the manufacturer's instructions. Additionally, we isolated RNA from one thallus with the method by Rubio-Piña & Zapata-Pérez (Rubio-Piña and Zapata-Pérez 2011) and purified it with the *RNeasy MinElute Clean-up Kit* (Qiagen).

We generated a whole-genome mate pair library with the *Nextera Mate Pair Sample Prep Kit* (Illumina, San Diego, CA, USA), aiming for an insert size of 5,000 bp. The RNAseq library was created with the *TrueSeq RNA Kit* (Illumina, San Diego, CA, USA). Long-read sequencing was performed on the *PacBio RS II* system (Pacific Biosystems of California, Menlo Park, CA, USA), using 16 SMRT cells in total. Furthermore, we constructed PoolSeq libraries with a target insert size of 200-300 bp. PoolSeqs for 6 populations of *L. pustulata*, consisting of 100 thalli each, collected in Sardinia, Italy (Dal Grande et al. 2017) were sequenced on an *Illumina HiSeq2000* machine, generating read pairs of 100 bp length.

3.2.3 Preprocessing of the sequencing data

3.2.3.1 Illumina data sets

All *Illumina* read pairs were processed with *Trimmomatic* v0.32 (Bolger, Lohse, and Usadel 2014) to remove low quality 3'-ends as well as remaining adapter sequences. We used a library of *Illumina* sequencing adapters as well as

ILLUMINACLIP:2:30:10 as the parameters for the trimming. The mate pairs were processed with *nextclip* v0.8 (Leggett et al. 2014), to remove adapters and bin them according to their read orientation. We used *--min_length 25 --number_of_reads 18978822 --trim_ends 0 --remove_duplicates* as the parameters.

3.2.3.2 PacBio data sets

The *PacBio* data from 16 SMRT cells were pooled. Best practices for *PacBio* pre-processing depend on the read coverages available. As sequencing depths of individual species in the metagenomic *L. pustulata* data vary, we pre-processed them by two different strategies to correct for the *PacBio*-specific sequencing errors. The data was intrinsically error corrected by generating consensus sequences by the pipeline of *canu* v1.20 (Berlin et al. 2015), employing standard parameters and estimating a total metagenome size of 150 Mbp. To also correct reads stemming from less abundant genomes, we additionally used *Illumina* data as extrinsic information for the error correction. We first merged the *Illumina* read pairs with *FLASH* (Magoč and Salzberg 2011) using standard parameters, and subsequently *de novo* assembled the processed *Illumina* read- and mate-pairs with *MIRA* v.4.0 (Chevreux, Wetter, and Suhai 1999) with the *genome,denovo,accurate* flags. With the help of *ECTools* (https://github.com/jgurtowski/ectools), the *PacBio* sequence reads were then error corrected according to those *Illumina* contigs, employing a minimum alignment length of 200 bp, a *WIGGLE_PCT* of 0.05 (allowing 5% from the end of a contig to not match while calculating overlaps) and a *CONTAINED_PCT_ID* of 0.8 for the contig mappings. We kept only error-corrected *PacBio* reads of a minimum length of 1,000 bp and trimmed all regions where the mapping identity was below 96%.

3.2.4 A stepwise, targeted assembly of the *L. pustulata* metagenome

As the uneven coverages in the metagenomic libraries require different assembly approaches (c.f. 2.3.3), we utilize different assembly methods and subsequently merge them to represent the different genomes of our metagenome (Chakraborty et al. 2016). A workflow for the data processing and subsequent assembly and merging is given in Figure 3-1.

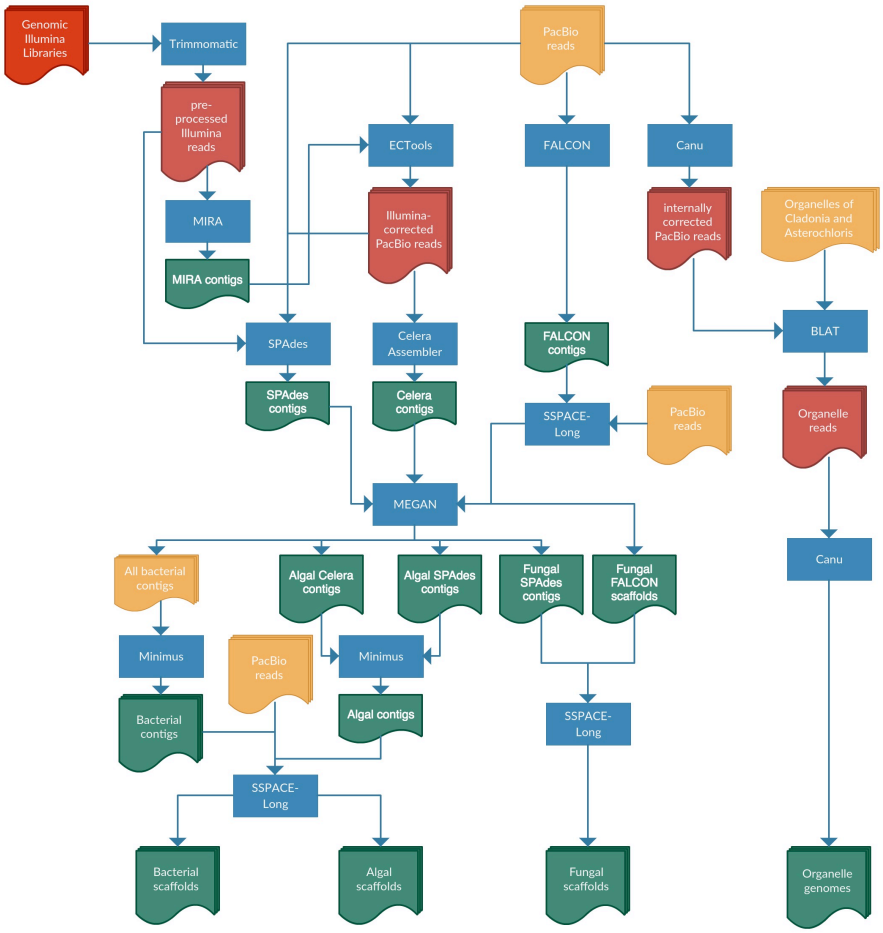


Figure 3-1: The complete preprocessing and assembly workflow, describing how the different data sets are used to target different genomes and are subsequently merged.

3.2.4.1 Performing the individual assemblies

The initial assembly of the *L. pustulata* metagenome was done with *FALCON* v0.2.1 (Chin et al. 2016). For this we used the unprocessed *PacBio* reads with a *length_cutoff* of 3,500 for the initial mapping and the pre-assembly steps. The pre-assembly was furthermore done with the *--min_idt 0.70 --min_cov 3 --local_match_count_threshold 2 --max_n_read 200* flags. The subsequent overlap

filtering was performed with *--max_diff 100 --max_cov 100 --min_cov 2 --bestn 10* as parameters.

To assemble also the lower coverage fractions, we employed two additional assembly strategies. For the first assembly we used the error-corrected *PacBio* reads as generated by *ECTools* and assembled them with the *Celera assembler wgs* v8.3rc2 (Berlin et al. 2015), using the default parameters. The second, hybrid-assembly was generated with *SPAdes* v3.5.0 (Bankevich et al. 2012). All processed whole genome *Illumina* libraries were given as input alongside both uncorrected and *ECTools*-preprocessed *PacBio* data. The standard parameters of *SPAdes* were applied.

As none of the three assemblers reconstructed full-length organelles we subsequently made use of a baiting strategy for their assemblies: The *canu*-corrected *PacBio* reads were aligned against the organelle genomes of *Cladonia grayi* and *Asterochloris sp.* by *BLAT* v35 (Kent 2002), using no cut-offs. All reads that were aligned were then extracted and the resulting bins were subsequently individually assembled with *canu* v1.20. The *canu* standard parameters were used; additionally setting the *estimated target genome size* parameter to 60 Kbp (*L. pustulata* mitochondrial genome), 120 Kbp (*Trebouxia sp.* mitochondrial genome) and 250 Kbp (*Trebouxia sp.* chloroplast genome) respectively.

3.2.4.2 Merging & finishing the assemblies

We first scaffolded the *FALCON* assembly with the raw *PacBio* reads and *SSPACE-Long* v1.1 (Boetzer and Pirovano 2014), run with standard parameters. The resulting scaffolds were then taxonomically classified. For this we mapped them against our custom database (see 2.2.5 on page 21 for a detailed description and Table A-1 on page 187 in the Appendix for taxonomic composition) using *DIAMOND* v.0.6.12.47 (Buchfink, Xie, and Huson 2014). *MEGAN* (Huson et al. 2011) subsequently assigned the scaffolds

to a taxonomic unit based on these mapping results. Alignments with a minimum bit-score of 50 were used for the LCA assignment, while no low-complexity filtering was performed. In the same way, the sequences assembled by *SPAdes* and *Celera* were taxonomically classified. For the genome of the fungus *L. pustulata*, the fungal *SPAdes* contigs of a length greater 3 Kbp were used to further scaffold the *FALCON* assembly with *SSPACE-Long*. For the algal genome we took the contigs that were assigned to the Viridiplantae from three assemblies and merged them into a single assembly using *minimus2* (Treangen et al. 2011). The same procedure was applied to merge bacterial sequences generated across the different assemblers. For the organellar genomes we aligned the *canu*-corrected *PacBio* reads back to them and took those mappings to circularize them with *circlator* v.1.2.0 (Hunt et al. 2015).

Additionally, we removed so far unidentified insertion and deletion (indel) errors, which are frequently found in *PacBio*-generated sequences (Chakraborty et al. 2016). We mapped the preprocessed metagenomic *Illumina* reads to the fungal, bacterial, algal and organellar assemblies and gave these as input to *Pilon* v1.15 (Walker et al. 2014). For the genomes of the fungus and the organelles, which have a high *Illumina* read coverage, we only mapped the mate pair library for the correction, as these are generated from a sample that is geographically close to the *PacBio* samples (c.f. Table 3-2, page 60). For the bacterial and algal assemblies we took both the read pairs as well as the mate pairs into account, as the *Illumina* coverage would not suffice otherwise. We performed a preliminary gene annotation for the fungal and bacterial assemblies with *MAKER2* (Holt and Yandell 2011), see 4.2.1 (page 85) for details on the gene prediction methods. The predicted genes were subsequently taxonomically classified through *MEGAN* (Huson et al. 2011) to identify chimeric scaffolds, in which contigs of two different genomes were joined. We furthermore mapped the reads of all sequencing libraries back to

the finished assemblies. For this we used *bowtie2* (Langmead and Salzberg 2012) for the genomic Illumina libraries and *bwa-mem* (H. Li 2013) for the *PacBio* data. The read coverages across the different data sets were then visualized by *anvi´o* (Eren et al. 2015).

3.2.5 Analyzing the microbiome composition

To estimate the microbial diversity in *L. pustulata* we analyzed the taxonomic composition for each individual data set. We performed a standard *DIAMOND* search of all pre-processed genomic *Illumina* libraries, as well as the uncorrected *PacBio* reads against the NCBI nr database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/nr, downloaded 2015-09-02). The results of that search were then used for the taxonomic assignment with *MEGAN5*. A minimum assignment score of 50 and no low complexity filtering were applied. All eukaryotic assignments were removed for subsequent analyses. The absolute read counts for each library where subsequently normalized in *MEGAN5* using the *sub-sampled counts* function to enable comparisons between them. These counts were then used to visualize the joint metagenome across the samples with *KronaTools* v2.6 (Ondov, Bergman, and Phillippy 2011). A visual comparison of the taxonomic diversity across the different samples was done with the *streamgraph* package (https://github.com/hrbrmstr/streamgraph) for *R*. To estimate how an assembly changes the observed metagenomic diversity, the same *MEGAN5* and *KronaTools* workflow was applied to the finished bacterial scaffolds. We first calculated the abundances of the different taxa based on the number of contigs assigned to the respective taxonomic groups. In a second step we corrected these numbers by accounting for the read coverages. For this we calculate the sum of reads that map to given contig (see *anvi´o* procedure above) using *samtools* (H. Li et al. 2009) and visualize these numbers using *KronaTools*. This was done for the *Illumina* read pairs.

3.3 Results

3.3.1 The fungus-to-alga ratio in *L. pustulata*

Following the preliminary coverage ratio estimation for the fungal to algal genome, based on the read-coverage found in an *Illumina MiSeq* read pair library (see 2.3.4, page 35), we performed quantitative PCRs to gain further insight into the expected over- and underrepresentation of the two genomes in our data sets. Based on the measurements done on four thalli of *L. pustulata*, we found that 1 ng of DNA extracted from a thallus contained between 26,962 and 39,063 copies of the fungal genome, in stark contrast to the 1,568 and 2,730 algal genome copies we observed. We found that the DNA isolated from a given thallus contained on average 16 copies of the genome of the fungus *L. pustulata* for each genome of *Trebouxia sp.*, the photobiont (Table 3-1).

Table 3-1: qPCR results from four thalli of *L. pustulata*. For each thallus the mean copy number of the triplicates was calculated and the mean fungal over algal copy number was calculated.

		Copy Number per Replicate				
Genome	**Thallus**	**#1**	**#2**	**#3**	**Mean**	**Mean Fungal:Algal Ratio**
Fungal	**#1**	24,618	28,317	27,952	26,962.58	13,98288708
	#2	27,375	28,196	30,826	28,799.43	12,66489031
	#3	36,805	39,654	40,730	39,063.67	14,30774526
	#4	36,769	37,257	38,935	37,654.29	24,00854662
Algal	**#1**	1,868	1,972	1,943	1,928.26	
	#2	1,994	2,368	2,458	2,273.96	
	#3	2,581	2,863	2,745	2,730.25	
	#4	1,478	1,531	1,694	1,568.37	

3.3.2 Hybrid-sequencing of the complex *L. pustulata* metagenome

Given the complexity of the *L. pustulata* metagenome, which includes a sizeable bacterial fraction (c.f. 2.3.4, page 35), strong coverage skews (see above) and a potentially repeat rich fungal genome (c.f. 2.4.3, page 43), we chose a more complex, hybrid sequencing setup that complements further

Illumina data with *PacBio* long-read data. We sequenced a mate pair library on the *Illumina MiSeq* to obtain long-range information, in addition to long-read sequencing using the *PacBio RSII* system. Furthermore we sequenced an RNAseq library with *Illumina MiSeq*, to aid in the post-assembly gene prediction. Lastly we sequenced 6 PoolSeq samples, consisting of 100 thalli each, on an *Illumina HiSeq 2000*, to get an estimate of the diversity found in *L. pustulata*. The sequencing statistics for all libraries are given in Table 3-2.

Table 3-2: Sequencing results for the individual libraries used.

Sequencing Technology	Library	Sampling Site	Read Length	Number of Read Pairs
Illumina MiSeq	MiSeq Metagenome	Sardinia, Italy	250 bp	14,013,249
	MiSeq Mate Pair	Saarland, Germany		18,978,822
	MiSeq RNASeq	Saarland, Germany		15,604,975
Illumina HiSeq2000	PoolSeq #1	Sardinia, Italy	100 bp	32,432,033
	PoolSeq #2			20,089,612
	PoolSeq #3			29,645,501
	PoolSeq #4			34,775,676
	PoolSeq #5			34,977,457
	PoolSeq #6			35,427,219
PacBio RS II	16 SMRT Cells	Saarland, Germany	6452 ± 3642 bp	1,851,141*

The PoolSeq data was only used for taxonomic assignment and not trimmed prior to that. * PacBio data are unpaired, total read counts are given.

The read pairs generated by *Illumina MiSeq* data were trimmed for low quality ends and sequencing adapters by *Trimmomatic* (Bolger, Lohse, and Usadel 2014), removing between 0.5 and 6.3% of all nucleotides. The *Illumina* mate

pairs were trimmed and sorted according to the read orientation by *nextclip* (Leggett et al. 2014), which classified 13,461,793 (70.9 %) as reliable mate pairs. The 16 SMRT cells sequenced through the *PacBio RS II* yielded 1,851,141 reads. As PacBio sequencing is performed on circular templates, individual fragments can be sequenced more than once. This leads to a primary read in the first pass of the sequencing and possibly to one or more sub-reads in subsequent passes. We found that 62.7% of our reads are being primary/main reads, with the remaining reads being sub-reads of those. As the *PacBio RS II* does not deliver fixed read lengths, we observed a marked difference in read lengths (Figure 3-2). The median primary read length is 7,194 bp (6,536 bp for sub-reads).

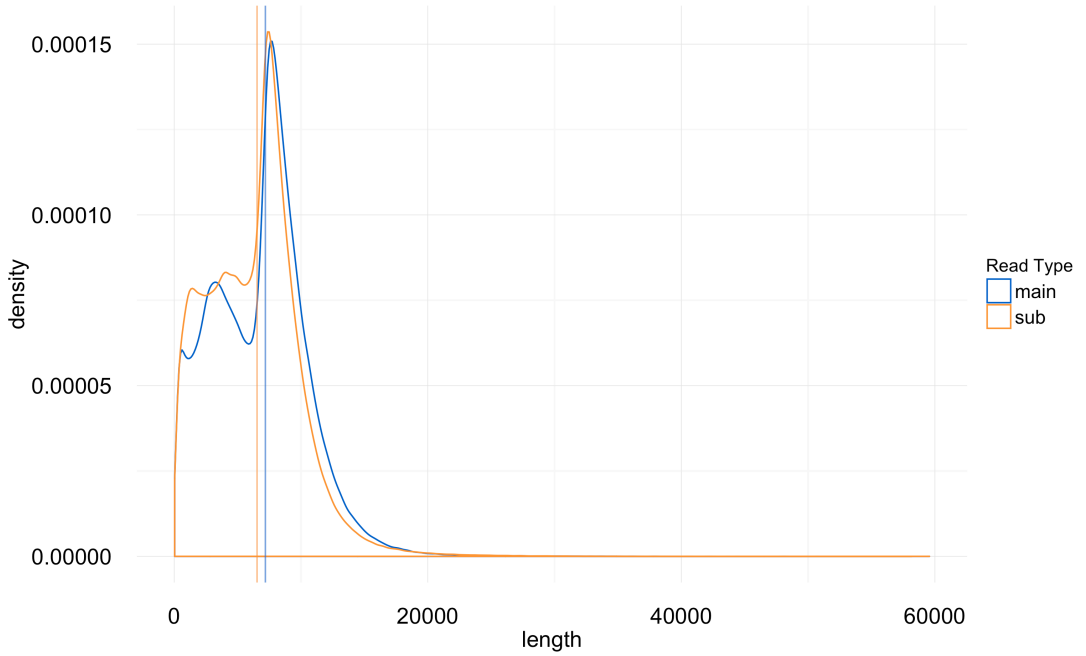


Figure 3-2: Read lengths of the main reads (blue) and sub reads (yellow) generated by the 16 SMRT cells that were sequenced on the PacBio RS II. Vertical lines give median read lengths.

We pre-processed the *PacBio* reads by error correcting them to reduce the methods inherently higher sequencing error rate. Depending on their coverage this can be achieved intrinsically, or extrinsically in hybrid-assemblies that also utilize short-read data (Berlin et al. 2015; H. Lee et al. 2014; H. Lee et al. 2016). An intrinsic error correction of our *PacBio* reads, performed with *canu* (Koren et al. 2017), yielded 594,604 error-corrected reads

with a mean read length of 7,726.9 bp (standard deviation 2,522.66 bp). An independent error correction, performed with a preliminary Illumina assembly, generated 1,328,383 reads with a mean length of 5,338.4 bp (standard deviation 3103.7 bp).

3.3.3 A step-wise assembly of the *L. pustulata* metagenome

To get the most out of the different types of sequencing data, it has been proposed to subsequently merge the results of different assembly strategies (Wences and Schatz 2015; Chakraborty et al. 2016). As we expected substantial coverage differences between the different genomes present in the *L. pustulata* metagenome, based on the qPCR and our preliminary assembly of a metagenomic shotgun data set (c.f. 2.3.4, page 35), we pursued both a pure *PacBio* sequencing approach as well as two hybrid-assembly approaches, subsequently merging those assemblies (c.f. Figure 3-1 on page 54 for the complete workflow).

Initial assemblies were performed with *FALCON* (Chin et al. 2016), using only *PacBio* data; with the *Celera* assembler (Berlin et al. 2015), which takes extrinsically error corrected *PacBio* reads as input; and *SPAdes* (Bankevich et al. 2012), which works with *Illumina* reads as well as corrected and uncorrected *PacBio* reads. We then continued to perform a taxonomic assignment on the individual assemblies, using *MEGAN* (Huson et al. 2011). We observed marked differences between the initial assemblies, both in total assembly length as well as in the contiguity as measured by the N50 value (Table 3-3). Overall, *FALCON* produced the shortest, though most contiguous assembly, while *Celera* yielded the longest though least contiguous assembly.

The taxonomic assignment of the individual contigs revealed marked differences to which degree the fungal, algal and bacterial fractions of the *L. pustulata* metagenome were assembled by the different methods. We found that *FALCON* preferentially assembled the highly abundant fungal genome,

with over 50% of the total assembly length coming from highly contiguous fungal sequences. *Celera* on the other hand assembled a much larger total sequence length for all three fractions. For the genome of the mycobiont *L. pustulata* it assembled 113 Mbp, while we expected a fungal genome size of around 35 Mbp. *Celera* furthermore yielded an assembly length 5.7x the size for the algal *Trebouxia sp.* genome compared to *FALCON*. The overall contiguity of the individual genomes is notably smaller for all fractions. We found that *SPAdes* performed especially well on the bacterial fraction of the *L. pustulata* metagenome, generating not only the longest total assembly length, but also the most contiguous bacterial fraction. Furthermore *SPAdes* generated the most contiguous algal fraction as well.

Table 3-3: Assembly results of the three different assemblers. Number of scaffolds, length and N50 are given for the total assembly plus the assigned fractions.

Assembly	Fraction	# of Scaffolds	Length	N50 (bp)
FALCON	All	2,343	62 Mbp	322,812
	Fungal	120	32 Mbp	550,723
	Algal	709	9 Mbp	17,208
	Bacterial	790	15 Mbp	56,167
Celera	All	22,216	216 Mbp	11,162
	Fungal	12,230	113 Mbp	10,395
	Algal	3,557	52 Mbp	16,617
	Bacterial	2,804	17 Mbp	7,798
SPAdes	All	21,900	123 Mbp	224,806
	Fungal	5,736	35 Mbp	159,267
	Algal	257	47 Mbp	461,016
	Bacterial	1,193	26 Mbp	91,450

We individually scaffolded and merged the fungal sequences generated by the different assemblers, using a combination of *SSPACE-Long* (Boetzer and Pirovano 2014) and *minimus2* (Treangen et al. 2011), to generate maximally

complete and contiguous assemblies. We subsequently also performed scaffolding and merging individually for the algal and bacterial assembly fractions respectively. Even after an initial read-based error correction, insertion/deletion (indel) errors introduced through the PacBio sequencing remain in assemblies (Chakraborty et al. 2016). We used *Pilon* (Walker et al. 2014), an *Illumina* mapping-based approach to correct for such indel errors. We found varying numbers of indel errors in the three assemblies. The fungal genome assembly included a total of 78,481 indel errors that were corrected by *Pilon*, while only 9,327 and 7,331 indel errors were observed in the algal and bacterial assembly respectively.

To further correct for the potential creation of chimeric scaffolds, which join sequences originating from two different genomes, we performed a preliminary gene annotation for the fungal and algal genome, which was subsequently taxonomically assigned through *MEGAN*. We found three instances where the exclusive presence of bacterial genes indicate single bacterial contigs of length 40 Kbp, 66 Kbp and 80 Kbp, which had been joined to the ends of longer, fungal scaffolds (1.9 Mbp, 1.7 Mbp and 180 Kbp respectively). These instances were manually corrected by splitting the scaffolds at those joins.

Table 3-4: Final assembly statistics after merging and correcting chimeric contigs.

Fraction	# of Scaffolds	Length	N50 (bp)
Fungal	43	33 Mbp	1,808,250
Algal	225	53 Mbp	848,255
Bacterial	499	35 Mbp	250,871

After this post-processing, we could reduce the total number of scaffolds to less than 800 (Table 3-4). The merging additionally resulted in a large increase in the assembly contiguity for all three taxonomic groups, with the fungal fraction having an N50 that is more than 3x larger than the individual

assemblies. Similar increases were found for the algal fraction, with a 1.8x increase; and the bacterial fraction (2.7x increase). The bacterial fraction furthermore includes two nearly complete genomes of species of the genus *Acidobacterium* (scaffolds 3.6 Mbp and 3.4 Mbp). As none of the assembled scaffolds included complete organelle genomes for *L. pustulata* and *Trebouxia sp.*, we performed a targeted assembly approach. We mapped the error-corrected *PacBio* reads to the organelle genomes of *Asterochloris sp.* and *Cladonia grayi.* This yielded 12,341 reads for the chloroplast of *Trebouxia sp.* and 6,017 and 18,315 reads for the mitochondrial genomes of *Trebouxia sp.* and *L. pustulata* respectively. Each of the three read sets was subsequently assembled with *Canu*, circularized with *circlator* (Hunt et al. 2015) and annotated with *mfannot* (https://github.com/BFL-lab/MFannot_data). This procedure led to gap-free, circularized genomes in all three cases, with lengths between 95 Kbp and 271 Kbp (Figure 3-3).

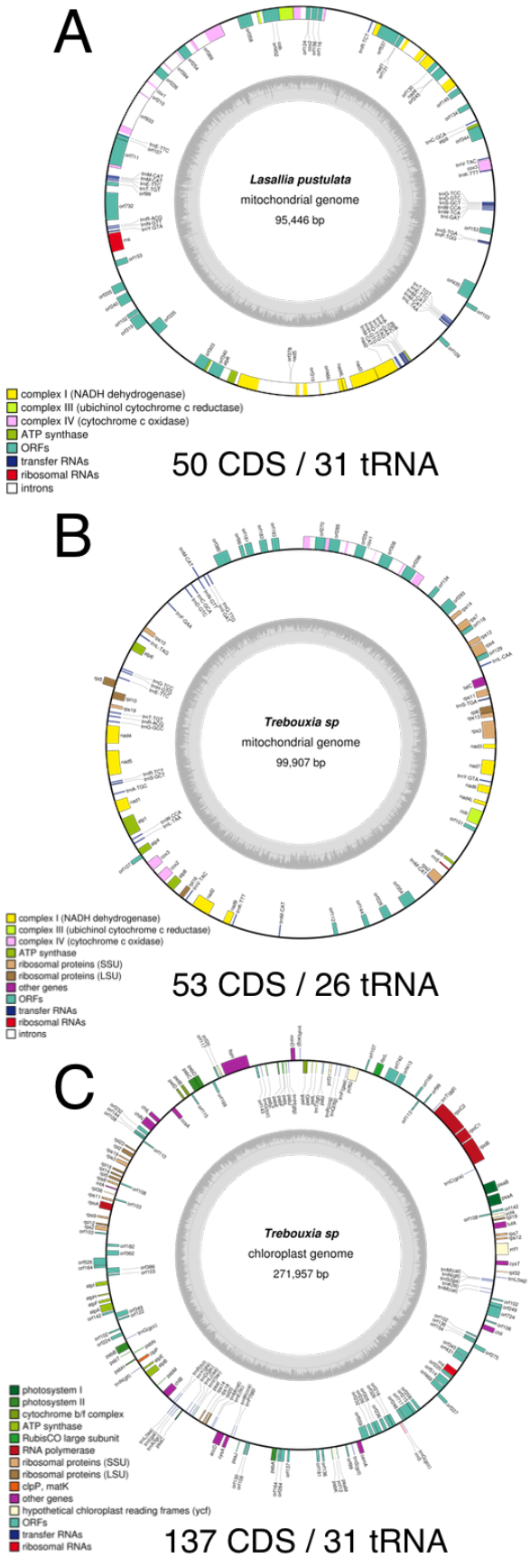


Figure 3-3: The organellar genomes of *L. pustulata* and *Trebouxia sp.* along with the respective gene predictions. Genes are color coded according to their function.

To investigate the compositional complexity of the lichen metagenome of *L. pustulata*, we mapped the reads from our different genomic sequencing libraries back to the finished genome assemblies and subsequently visualized those mappings (Eren et al. 2015). This revealed pronounced coverage differences between the genomes, which are rather consistent across the different sequencing libraries (Figure 3-4). Normalizing the coverages to the lowly abundant nuclear genome of *Trebouxia sp.*, we observed that for each copy of it, there are 16 algal mitochondrial and chloroplast genomes on average. Furthermore, we calculated that nearly 20 nuclear and 290 mitochondrial fungal genomes are found for each nuclear *Trebouxia sp.* genome (see Appendix, Table A-2 on page 191).

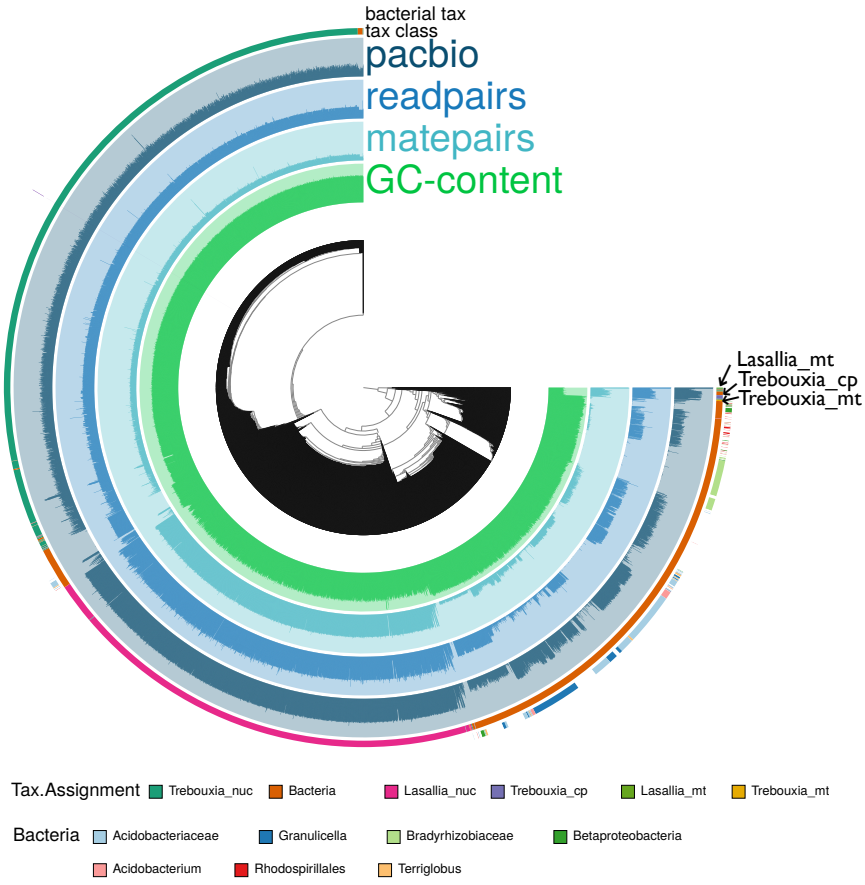


Figure 3-4: The read coverages and G/C content across the different genomes and different sequencing libraries. Sequences were split into chunks of 20 Kbp, the bars in each ring give the mean G/C and coverage for each chunk. The two outer rings give the taxonomic classification on a higher level as well as on a finer level for the bacteria.

3.3.4 The microbiome of *L. pustulata*

To further analyze the taxonomic composition of the bacterial fraction of the *L. pustulata* metagenome, we investigated the composition on the read level and across each of our sequencing libraries, including 6 PoolSeq experiments (see Table 3-2, page 60). By removing the eukaryotic fraction and normalizing the assignments – to account for the different number of sequences in the individual libraries – we generated a view of the joint microbiome of *L. pustulata*. This revealed that the Acidobacteriaceae, with 25% of all bacterial sequences being assigned to them, make up a substantial fraction of the *L. pustulata* microbiome. On the other hand, the Rhizobiales, which have been found as key contributors in other lichen microbiomes (Sigurbjörnsdóttir et al. 2015; Erlacher et al. 2015; Hodkinson et al. 2012; Grube et al. 2015), made up only 4% of the *L. pustulata* microbiome (Figure 3-5), see Table A-3 (page 191) in the Appendix for the 20 most abundant taxa on the genus level.

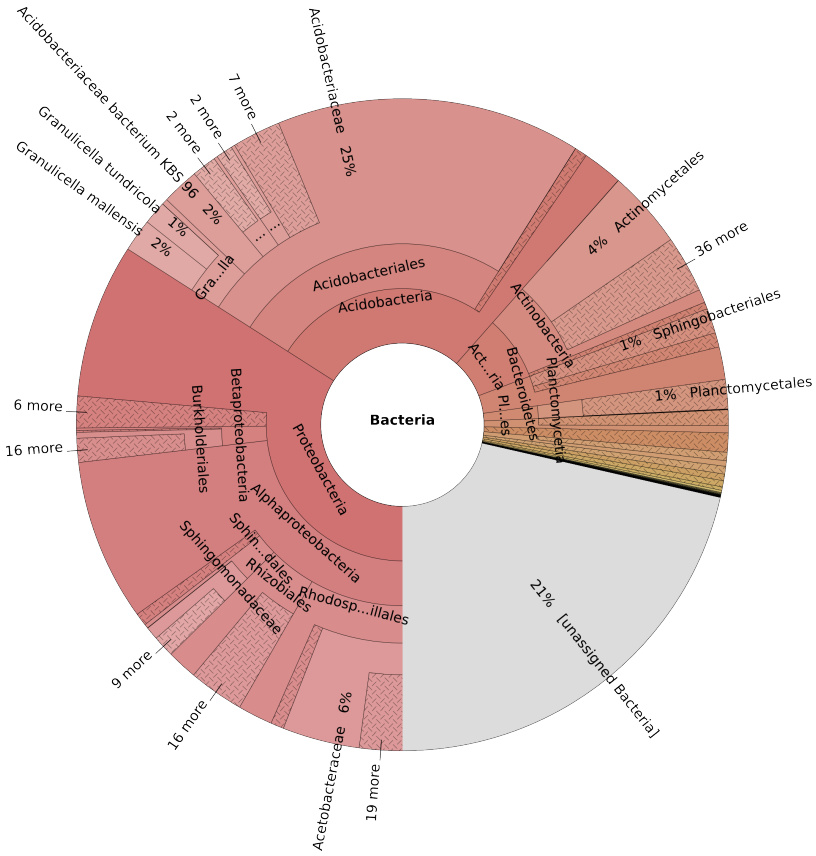


Figure 3-5: Taxonomic composition of bacterial pan-microbiome of *L. pustulata* based on nine sequencing libraries. Read-counts were normalized to account for differences in sequenced reads.

To estimate the compositional variation between different thalli and sequencing technologies, we additionally took the taxonomic assignments for the reads of each sequencing library and visually compared the abundances. We find that the taxonomic composition between the different samples is largely stable on the family level (Figure 3-6), even between samples from Germany (*PacBio* & *Illumina* read pair libraries) and Italy (*Illumina* PoolSeq and read pair libraries).

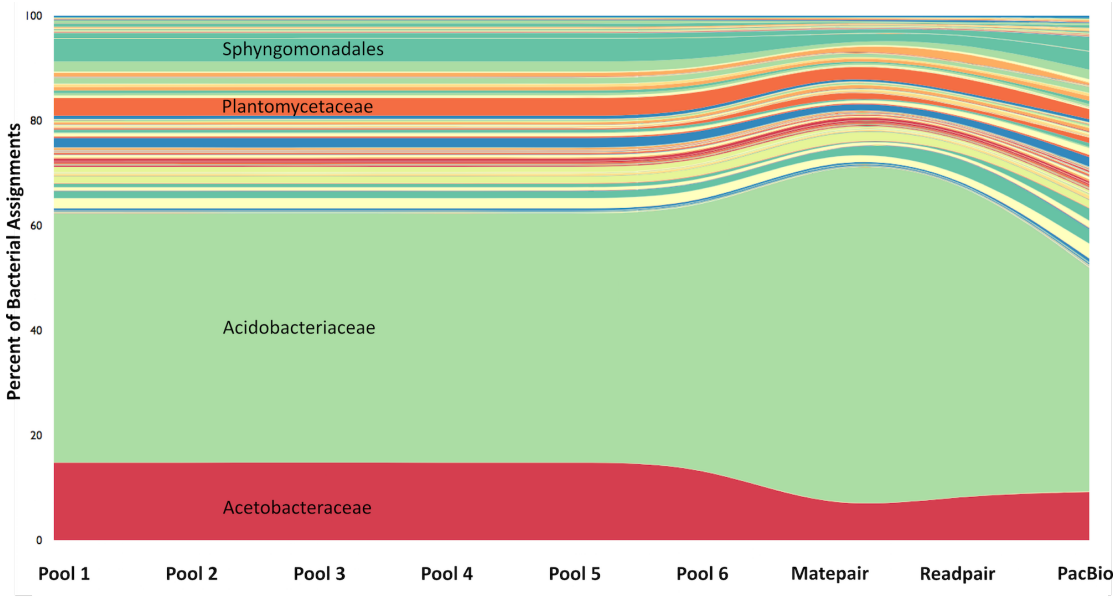


Figure 3-6: Taxonomic composition of the bacterial microbiome between the different samples.

We investigated the influence of methodological differences by additionally performing a taxonomic assignment in the same way on the 499 bacterial scaffolds. We found that the fraction of sequences assigned to the Acidobacteriaceae in this case dropped to 18%, while the proportion of the Rhizobiales increased to 11% (Figure 3-7 A). A total of 84 scaffolds were assigned to the Acidobacteriaceae, accounting for a total length of 10.8 Mbp, while the 48 scaffolds of the Rhizobiales sum up to 3 Mbp.

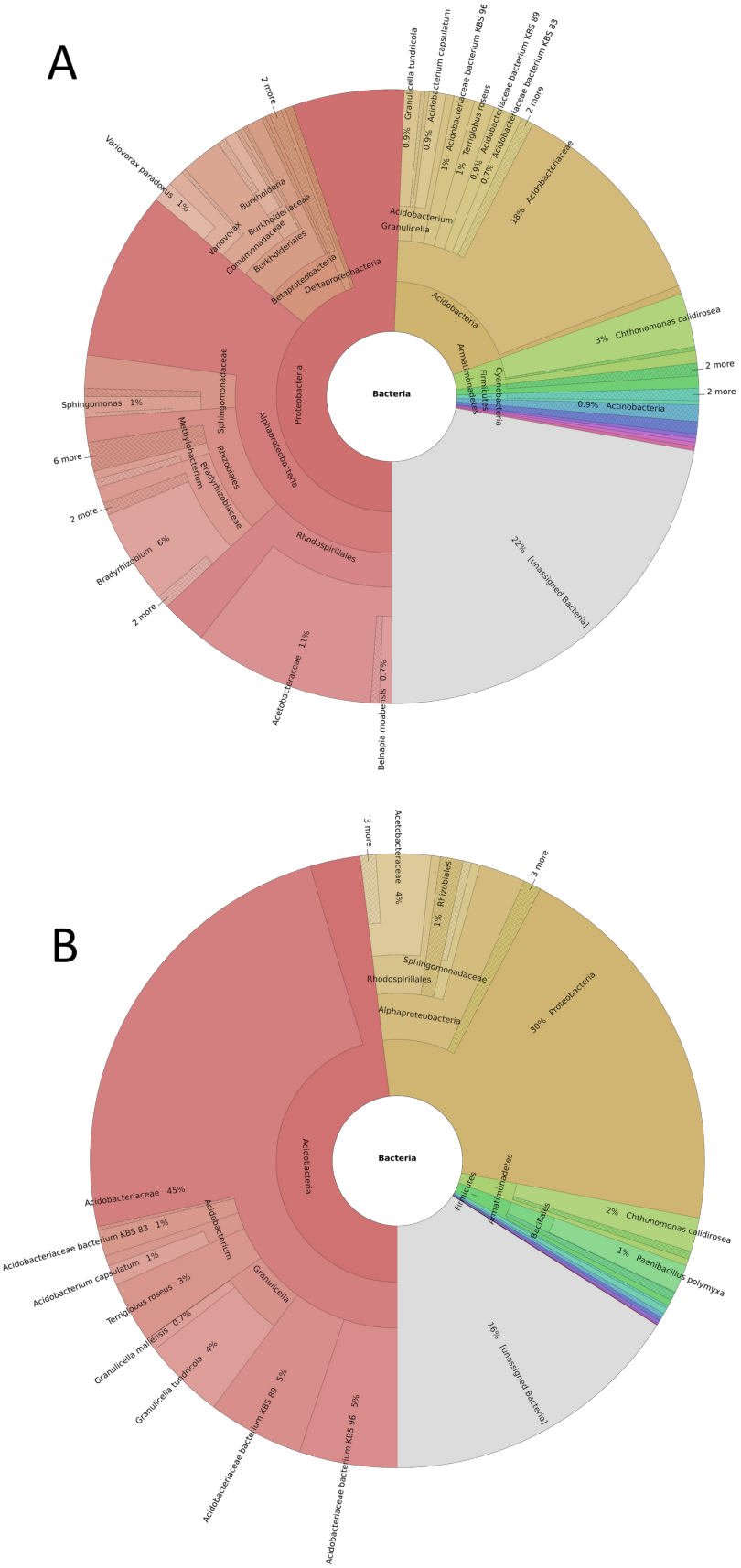


Figure 3-7: Taxonomic composition based on the assembled microbiome, without (A) and with (B) taking the read coverage into account.

To account for the differences in read coverages for the individual contigs, we mapped the read pair library back to the bacterial scaffolds (see Figure 3-8 on page 77 for a schematic of the different approaches of assigning the taxonomy) and corrected the taxonomic counts with the number of mapped reads for each sequence (Figure 3-7 B). This shifted the taxonomic composition markedly, with 45% of all sequences being assigned to the Acidobacteriaceae and only 1% to the Rhizobiales. Analyzing the coverages for the two groups we observed that the Rhizobiales only have a median coverage of 3x, while the Acidobacteriaceae have one of 19x.

3.4 Discussion

The assembly of metagenomes is a computational challenging task, which has been evaluated on different microbial communities (Awad, Irber, and Brown 2017; Mavromatis et al. 2007; Sangwan et al. 2016; Ghurye, Cepeda-Espinoza, and Pop 2016; Kyrpides et al. 2014; Marbouty et al. 2017). Based on our own research into the assembly of lichen metagenomes from genome skimming (c.f. 2.3.4, page 35), we further investigated the genome ratios found in thalli of *L. pustulata*, based on qPCR. We observe that the ratio between fungal and algal genome is even more skewed than estimated by the initial read mapping, hindering the generation of long, contiguous sequences for the underrepresented species in the assembly procedure. For this reason we approached the problem of lichen metagenome assembly with a hybrid-strategy that makes extensive use of second- and third-generation sequencing methods.

3.4.1 Assembling the metagenome of *L. pustulata*

Third-generation sequencing methods are capable of generating long but error-prone reads, compared to second-generation methods (H. Lee et al. 2016; Bleidorn 2015). Multiple ways of dealing with these errors have been proposed: Due to the largely uniform distribution of sequencing errors along these reads (Ross et al. 2013), these errors can be corrected by generating consensus sequences, given a high enough sequencing coverage (Chin et al. 2016; Koren et al. 2017; H. Lee et al. 2014). In the absence of such high coverages, these errors can be corrected externally, by mapping short sequencing reads, which are less error prone, to the long reads to generate the consensus (H. Lee et al. 2016; Berlin et al. 2015).

In case of the lichen metagenome, with the varying coverages for the individual genomes, we expect that no single assembly solution can utilize both *PacBio* long reads as well as *Illumina* short-read data to the full extent.

This is confirmed when looking at the assembly results generated by different methods (Table 3-3, page 63). An assembly performed with *FALCON* (Chin et al. 2016), using only *PacBio* data, generated a highly contiguous assembly for the fungal genome, while the algal and bacterial fractions are highly underrepresented. Similarly, a *Celera* assembly (Berlin et al. 2015) performed with *PacBio* data, which was error-corrected based on an *Illumina* assembly, could improve on both the algal as well as the bacterial fraction, but at the cost for the fungal genome, which is less contiguous. This might be a result of the pre-assembly error correction procedure, which shortens the *PacBio* reads, based on the mappings. An assembly with *SPAdes*, which uses the *PacBio* data largely for contig scaffolding, can thus further improve the contiguity of the bacterial/algal fraction, though at the cost of generating a shorter algal genome.

By merging the results of these three assemblies we improved on the overall assembly outcome, in line with prior evidence for the utility of assembly merging (Aganezov and Alekseyev 2016; Chakraborty et al. 2016). Ultimately, this strategy yielded an algal assembly with an N50 of 848 Kbp and a fungal assembly with an N50 of 1.8 Mbp, in addition to two largely reassembled bacterial genomes of the genus *Acidobacterium* (Table 3-4, page 64). Despite these improvements, this approach failed to generate full-length organelle genomes for *Trebouxia sp.* and *L. pustulata*. For the mitochondrial genome of *L. pustulata*, this is potentially a result of the high coverage that is around 14.5 times higher than the nuclear genome. For the organellar genomes of *Trebouxia sp.* it is so far unclear why the general assembly procedures did not yield full-length organelles, as their coverage is similar to that of the nuclear *L. pustulata* genome. Fully circularized organelle genomes could only be generated by using a baiting-based approach that simplified the assembly problem (Figure 3-3, page 66). This procedure further highlights the need for bespoke approaches when assembling complex metagenomes.

Following the assembly, we found that indel sequencing errors, most likely introduced by the *PacBio* sequences, remain despite a ~160x coverage for the fungal genome. Around 78,000 such indel errors (238 per 100 Kbp) were found by correcting the genome of *L. pustulata* through the *Illumina* data. This is in line with the error rate observed in an unpolished assembly of *Drosophila melanogaster* that was performed using uncorrected *PacBio* data and a dedicated *PacBio* assembler (*HGAP*), where 180 indel errors per 100 Kbp were observed (Chakraborty et al. 2016). In the algal and bacterial assemblies the number of detected indel errors is around a magnitude smaller (around 9,000 and 7,000 respectively). This discrepancy can likely be explained by the different assembly methods that generated the bulk of the final assemblies for the individual taxa. As can be seen in Table 3-3 (page 63), *FALCON* managed to reconstruct virtually all of the final fungal assembly, using only the indel-prone *PacBio* reads. The situation is different for the bacterial and algal assemblies; here *FALCON* generated only a fraction of the final assemblies. Instead most of the final assemblies of the bacteria and alga were generated in the two hybrid-assemblies that heavily rely on the less error-prone *Illumina* data. Thus, *PacBio* data is less used in these assemblies; consequently the *PacBio* indel error plays less of a role. We also need to consider a second explanation for the difference in corrected indel errors: Overall, the *Illumina* sequencing coverage is lower for the bacterial and algal genomes. As *Pilon* requires a minimum *Illumina* coverage for the error correction (Walker et al. 2014), we might in individual cases miss indel errors as we do not have enough evidence to reliably identify and correct them.

3.4.2 The diversity of the *L. pustulata* microbiome

A first insight into the microbiome comes from mapping back the sequencing reads of all nine libraries back to our scaffolds. In line with our initial observations, based on a single *Illumina* MiSeq library and the qPCR results,

we find marked differences in the scaffold coverages between the algal and fungal genomes. We furthermore find large coverage differences between individual genomes present in the bacterial fraction. Interestingly, the abundances of individual taxa in the bacterial fraction appear to be rather consistent across samples, as seen in Figure 3-4 (page 67). Our further taxonomic classification on the read level, done across all genomic libraries, shows that there is a consistent microbiome found in *L. pustulata*, which stable on the family level across the nine libraries (Figure 3-5 on page 68, Figure 3-6 on page 69).

This stable taxonomic composition hints that the *L. pustulata* microbiome might have a functional role in the lichen symbiosis, as hypothesized for other lichen microbiomes (Erlacher et al. 2015; Grube et al. 2015). The microbiome of *L. pustulata* appears to be dominated by Acidobacteriaceae (25% of all bacterial assignments). In soils the presence of Acidobacteria is an indicator for oligotrophic conditions (Castro et al. 2010). Additionally, Acidobacteriaceae are capable of slow metabolic rates when faced with nutrient-deprived environments and are known to be tolerant to changes in hydration (Ward et al. 2009). These traits make them potentially well adapted to co-habitate with *L. pustulata*, which preferentially grows on nutrient-poor rocks (Hestmark et al. 1997; Dal Grande et al. 2017). While these characteristics would allow them to potentially play a functional role in the lichen symbiosis, we cannot rule out that their association with *L. pustulata* is only due to a shared habitat preference. Such co-occurrences, based on nice requirements, have been stipulated in earlier studies, which found that geography could explain differences in microbiome composition (Hodkinson et al. 2012).

Strikingly, the Rhizobiales, key contributors in other lichen microbiomes, are largely absent, only making up 4% of all bacterial reads. In contrast, in *Lobaria pulmonata* (Grube et al. 2015; Erlacher et al. 2015) Rhizobiales make up 32.2%

of all identified bacteria. Assuming the methodology used for the assignment might explain this stark difference, we further analyzed the bacterial composition on the scaffold level, as the taxonomic assignment by Erlacher et al. (2015) was performed on the assembly level. It seems intuitive to perform an assembly prior to the taxonomic classification, as it increases the sequence lengths, thus facilitating an easier classification. At the same time, this procedure introduces biases. Reads for highly abundant taxa will be collapsed into few, long contigs which are highly covered, but will only be counted once in the typical *MEGAN* (Huson et al. 2011) workflow. Low-abundance taxa on the other hand will not be assembled at all, and are thus missing for the subsequent taxonomic assignment (Vollmers, Wiegand, and Kaster 2017). Figure 3-8 gives a hypothetical example of the effects of the different strategies.

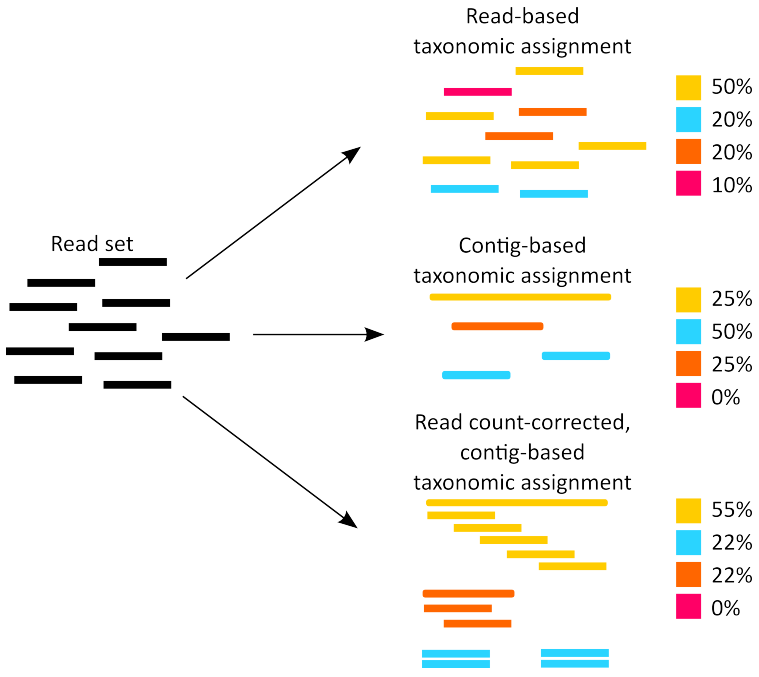


Figure 3-8: An overview over the potential biases introduced by the different taxonomic assignment strategies. Performing the assignment on the read level (top-right) ideally quantifies all reads present in the read set. A taxonomic assignment on the contig-level (middle-right) counts long contigs, consisting of many reads, only once, thus the abundance of the yellow fraction artificially decreases, while the orange and blue fractions increase in abundance. The red fraction was unassembled, and thus is completely missing. Correcting for the number of reads that could be mapped against the individual contigs (bottom-right) ameliorates this bias, the abundance of the yellow fraction increases again. Though, the red fraction remains absent as it was unassembled.

By estimating the taxon abundances on the scaffold level we observed exactly this effect. When looking at the number of scaffolds, regardless of length or coverage, the percentage for the Rhizobiales nearly triples to 11%, while it dorps to 18% for the Acidobacteriaceae drop (see Figure 3-7 A on page 70). This gives evidence for the effect of collapsing reads into contigs/scaffolds, as the counting the scaffolds ignores length and coverage differences, with the latter being more than 6 times larger for the Acidobacteriaceae. To counter this effect, we corrected these counts by using the number of mapped reads to each scaffold for the basis of the classification (Figure 3-7 B). In this case, the abundance of the Rhizobiales drops to 1%, while the Acidobacteriaceae grow to 45% of all bacterial sequences. This demonstrates the effect of unassembled sequences. While the highly abundant Acidobacteriaceae could be largely assembled (with the two largest contigs being species of the Acidobacteriaceae), low-abundance species could not be assembled to the same degree and thus no reads can be mapped against their contigs/scaffolds. This further highlights the potential sources of bias when analyzing the taxonomic composition of microbiomes (Nayfach and Pollard 2016).

3.5 Conclusion

Our analyses show that the metagenome of *L. pustulata* is a diverse ecosystem, which not only consists of the mycobiont and photobiont, but also includes a rich bacterial community. The composition of this community is largely stable across a wide geographic range. This complexity of the *L. pustulata* metagenome renders the *de novo* genome assembly of all included genomes algorithmically challenging, as there are strong abundance differences between the individual taxa, leading to a highly skewed sequencing coverage. Despite this, we find that an assembly of such complex metagenomes is possible, by utilizing the strengths of different assembly algorithms and sequencing methods, thus facilitating comparative studies of lichen metagenomes.

4 Genome annotation, artifacts and solutions

4.1 Introduction

The annotation of newly sequenced genomes is a routine task performed subsequent to their assembly (Yandell and Ence 2012). The annotation links genomic subsequences to a variety of biological information. It includes, but is not limited to, identifying genes (Mathe et al. 2002), repetitive elements (Jiang 2013), regulatory motifs (D'haeseleer 2006), regulatory RNA elements like microRNAs (Carthew and Sontheimer 2009) or long non-coding RNAs (Signal, Gloss, and Dinger 2016). This annotation allows for insights into the biochemical and functional potential encoded in those genomes (Stein 2001). For this reason, a variety of *in silico* methods have been developed to facilitate the prediction of these sequence elements, making use of sequence modeling, reference databases or *ab initio* predictions based on inherent sequence information (Stein 2001; Klasberg, Bitard-Feildel, and Mallet 2016). For example, transposable elements like long-terminal repeats (LTR) and other interspersed elements are annotated by searching for their characteristic sequence motifs (Ellinghaus, Kurtz, and Willhoeft 2008); through reference databases (Hubley et al. 2016; W. Bao, Kojima, and Kohany 2015); by compiling own *ab initio* predictions for the repetitive elements (Price, Jones, and Pevzner 2005; Edgar and Myers 2005); or using a combination of these methods (Smit, Hubley, and Green 2015).

Similarly, there is a broad set of methods for the *de novo* annotation and prediction of genes along genomic sequences (Hoff and Stanke 2015). Given the structural differences between eukaryotic and prokaryotic genomes, the individual methods do not only differ between their underlying strategies, but also in the organisms they are designed to annotate (Hyatt et al. 2010;

Seemann 2014). For eukaryotic genomes, gene annotation tools can be broadly classified into three categories (Hoff and Stanke 2015):

1. *Ab initio* predictors, which require a set of already curated and known gene models to subsequently train statistical models for the prediction of additional genes (Stanke and Waack 2003; Korf 2004; Besemer and Borodovsky 2005).
2. Transcript-based predictors, which map RNAseq sequencing data against the genome to identify the presence of genes (Trapnell et al. 2010; Trapnell, Pachter, and Salzberg 2009).
3. Homology-based predictors, which use the mapping of homologous genes against the unannotated genome to find genes (Birney, Clamp, and Durbin 2004; Slater and Birney 2005; Dunne and Kelly 2017; Wyman, Jansen, and Boore 2004).

Benchmarks of the different methods – and the tools using them – have shown widely varying accuracy and sensitivity (Guigó et al. 2006; Steijger et al. 2013). Recent efforts have focused on developing ensemble-based methods, like *MAKER2* (Holt and Yandell 2011), *BRAKER1* (Hoff et al. 2016) or *EVM* (Haas et al. 2008), which combine the different strategies for gene predictions. Further benchmarking efforts have shown that these ensemble methods generally achieve higher accuracy than single methods (Coghlan et al. 2008). Given gene structure differences between taxonomic groups, there has been a recent trend to develop taxon-specific ensemble methods that try to further optimize the quality of gene predictions. For example, fungal genes are known to be rather short, but densely packed with overlapping untranslated regions (UTRs) (David et al. 2006). As this routinely leads to wrongly annotated genes, this has lead to fungal-specific gene prediction tools such as *CodingQuarry* (Testa et al. 2015), *SnowyOwl* (Min, Grigoriev, and Choi 2017), *FunGAP* (Min, Grigoriev, and Choi 2017), or *funannotate* (https://github.com/nextgenusfs/funannotate). This is of importance, as the

accurate and complete prediction of genes has been shown to be a necessary prerequisite for comparative genomic studies (Veeckman, Ruttink, and Vandepoele 2016; Denton et al. 2014; Dunne and Kelly 2017). To further increase the annotation accuracy, methods that help with the manual curation of predicted genes have been proposed (E. Lee et al. 2013; Drăgan et al. 2016).

Subsequent to the prediction of genes, their functional annotation plays a relevant role in generating biological knowledge from newly sequenced genomes. Similar to the prediction of genes itself, annotation methods rely on various approaches and vary in their performance (Loewenstein et al. 2009; Radivojac et al. 2013). Hidden Markov Models (HMM) and neural networks have been trained to identify protein characteristics like subcellular localization, signal peptides, cleavage sites and transmembrane helices amongst genes (Emanuelsson et al. 2007). Furthermore, HMMs are widely used to quickly identify conserved domains across unannotated gene sets (Letunic et al. 2006; Finn et al. 2016; Hubbard et al. 1997).

Sequence similarity-based methods are frequently employed to transfer functional annotations between already functionally annotated and so-far unannotated genes. Such functional annotation transfers are frequently performed with the Gene Ontology (Ashburner, Ball, and Blake 2000; Conesa et al. 2005) and KEGG (Kanehisa et al. 2006; Kanehisa, Sato, and Morishima 2016). The Gene Ontology (GO) offers three ontologies that contain a standardized vocabulary of terms that describe molecular functions, biological processes and cellular compartments. Sequence similarity can be used to link unannotated genes to those that are already associated with GO terms (Conesa et al. 2005). KEGG (the *Kyoto Encyclopedia of Genes and Genomes*) contains information about biological pathways and the genes that are involved in them. Unannotated genes can be associated to functions and pathways by similarity to genes that are already available in KEGG (Kanehisa, Sato, and Morishima 2016). As homology-based methods are

limited by the availability of annotated homologous sequences, there are increasing efforts to enable *ab initio* functional prediction using Random Forests trained on sequence inherent features (Weber et al. 2015; Peled et al. 2016).

Predicting the functional capabilities found in hologenomes is central to understand their interactions and metabolic capacities (Frank et al. 2016; Bose et al. 2015). Here, we present an annotation of the lichen *Lasallia pustulata* by annotating the genomes of the mycobiont and the photobiont. Performing a repeat and gene annotation, we find considerable differences in both genome organization and gene structure for between the genomes of *Trebouxia sp.* and *L. pustulata*. Comparing the gene predictions done by *MAKER2* and *funannotate* for the genome of *L. pustulata* we find marked differences in the prediction performance with respect to both the sensitivity and accuracy. A subsequent comparative analysis of the gene predictions of five Lecanoromycetes reveals how incomplete and inaccurate annotations impact downstream evolutionary analyses.

4.2 Methods

4.2.1 Gene annotation

The genes for the mycobiont of *Lasallia pustulata* were annotated with both *MAKER2* v2.31.8 (Holt and Yandell 2011) and *funannotate* v0.5.7 (https://github.com/nextgenusfs/funannotate). While *MAKER2* is a general gene annotation pipeline, the *funannotate* pipeline was designed for fungal genomes, which makes extensive use of RNAseq data to avoid fungal-specific annotation errors (Hoff et al. 2016). For *MAKER2* we followed an iterative procedure, which contains two rounds of annotation (Kumar 2013), as additionally outlined in the instructions available online at *Github*: https://github.com/sujaikumar/assemblage/blob/master/README-annotation.md. For the first pass of *MAKER2* we identified well-conserved genes with *CEGMA* v2.5 (Parra, Bradnam, and Korf 2007) and converted these predictions into Hidden Markov Models (HMMs) for *SNAP* v2006-07-28 (Korf 2004). Additionally we ran *GeneMark* v4.21 (Besemer and Borodovsky 2005) on the fungal genome. Furthermore, we assembled RNAseq data from *L. pustulata* with *Trinity* release 2013-11-10 (Grabherr et al. 2011), using the *–jaccard-clip –normalize_reads* parameters. The assembled transcripts, as well as the proteomes of *Cladonia grayi* (http://genome.jgi.doe.gov/Clagr3/) and *Xanthoria parietina* (http://genome.jgi.doe.gov/Xanpa2) were used as further evidence for *MAKER2*. The gene predictions of this first pass were subsequently converted to HMMs for *SNAP* and *AUGUSTUS* v3.1 (Stanke and Waack 2003; Stanke et al. 2006). For the second pass of *MAKER2* the new HMMs of *SNAP*, *GeneMark* and *AUGUSTUS* were used, along with the RNAseq and proteome evidence.

For *funannotate* the gene prediction was performed in accordance with the documentation (https://github.com/nextgenusfs/funannotate/wiki). In addition to the already *de novo* assembled RNAseq data, these were also

assembled using *Trinity*'s *–reference-guided* mode, using the assembly of *L. pustulata* as further input. The output of both *Trinity* runs was then the basis to identify transcript assemblies on the *L. pustulata* genome assembly, using *TransDecoder* in the *PASA* pipeline (Haas et al. 2008). Additionally *HISAT2* (Kim, Langmead, and Salzberg 2015) mapped the RNAseq data to the fungal genome, giving *–max-intronlen 3000*. Ultimately, the results of *PASA*, *HISAT2*, and *Trinity* were given as input to *funannotate*, along with the proteomes of *X. parietina* and *C. grayi*. The same workflow for *funannotate* was applied to an unpublished draft genome of *Lasallia hispanica* and the unannotated draft genome of *Umbilicaria muehlenbergii* (S. Y. Park et al. 2014).

The gene prediction for *Trebouxia sp.* was done with *MAKER2*, analogous to the annotation described for *L. pustulata* above. Instead of *Cladonia grayi* and *Xanthoria parietina*, the protein set of the photobiont *Asterochloris sp.* (http://genome.jgi.doe.gov/Astpho2/) was used as external protein evidence.

4.2.2 Comparing the fungal gene predictions

The prediction results of *MAKER2* and *funannotate* for *L. pustulata* were evaluated by comparing them to each other. In a first step, we used *bedtools* v2.17.0 (http://bedtools.readthedocs.io/) to find to which extent the gene predictions occur in the same/similar positions, by finding partially and completely overlapping regions in the predictions. We furthermore analyzed the RNAseq coverage for all predicted genes, based on the *HISAT2* mapping generated for *funannotate*. In a last step, we analyzed the fidelity of the predicted genes by searching for orthologs in other species. We performed 3 ortholog searches with *OMA* v1.0.3 (Altenhoff et al. 2015), searching for orthologs in *X. parietina* and *C. grayi*, additionally adding non-overlapping sets of 3 Dothideomycetes and 3 Eurotiomycetes each (Table 4-1). For the individual runs we counted how many of the genes exclusively predicted by *MAKER2* and *funannotate* were assigned to orthologous groups (OG) and

what the group size of the OG is. Furthermore, we predicted *in silico* gene fusions using the Rosetta Stone approach (Marcotte and Marcotte 2002), which analyzes two genomes to find instances where a single gene present in one organism is found as two separate genes in another one. We applied the Rosetta Stone method to the gene predictions of *L. pustulata* and *C. grayi* to estimate their fidelity.

Table 4-1: Taxon sets used for the orthology prediction with *OMA* to compare the gene predictions of *MAKER2* and *funannotate*.

	Set I	**Set II**	**Set III**
Dothideomycetes	*Lepidopterella palustris* *Sporormia fimetaria* *Zasmidium cellare*	*Lentithecium fluviatile* *Myriangium duriaei* *Trichodelitschia bisporula*	*Lindgomyces ingoldianus* *Massarina eburnea* *Westerdykella ornata*
Eurotiomycetes	*Eurotium rubrum* *Microsporum canis* *Trichophyton verrucosum*	*Gymnascella aurantiaca* *Histoplasma capsulatum* *Trichophyton rubrum*	*Coccidioides immitis* *Gymnascella citrina* *Arthroderma benhamiae*
Lecanoromycetes	*Lasallia pustulata* *Xanthoria parietina* *Cladonia grayi*		

4.2.3 Repeat annotation

Libraries of repetitive elements in *L. pustulata*, *L. hispanica*, *U. muehlenbergii* and *Trebouxia sp.* were generated using *RepeatModeler* and subsequently used with *Repeatmasker* v4.0.5 (Smit, Hubley, and Green 2015) to identify repeats in

the respective genomes. We additionally applied the *Inverted Repeat Finder* (*IRF*) that identifies candidate inverted repeats (IR) through an exact matching of reverse-complemented *k*-mer matches which are clustered by positions, subsequently verifying these through a Smith-Waterman alignment (Warburton et al. 2004). The *IRF* was run with default parameters to predict inverted repeats in the genomes of the Lecanoromycetes *L. pustulata*, *U. muehlenbergii*, *C. grayi*, *Usnea florida* and *X. parietina*.

4.2.4 Functional annotation

The predicted genes of the mycobionts and *Trebouxia sp.* were assigned to *KEGG* Orthologs using *BlastKOALA* v2.1 (Kanehisa, Sato, and Morishima 2016). Gene Ontology terms (Ashburner, Ball, and Blake 2000) were assigned to the predicted protein sequences of all Lecanoromycetes and *Trebouxia sp.* by *BLAST2GO* v4.1.7 (Götz et al. 2008), following a *BLAST* against the NCBI-nr database. Pfam domains were annotated using *PfamScan* (Finn et al. 2016). To compare the functions represented in our fungal gene predictions with those of the publicly available Lecanoromycetes *Xanthoria parietina, Usnea florida* and *Cladonia grayi*, we annotated KO groups, GO terms and Pfam domains their gene sets in the same way.

4.2.5 Investigating annotation and assembly errors

We searched for orthologs between *L. pustulata, U. muehlenbergii*, *X. parietina*, *C. grayi* and *U. florida* with *OMA* v1.0.3. Hierarchical Orthologous Groups (HOGs) that contained sequences of 4 of the 5 taxa were analyzed in detail. All candidates for private gene losses where then searched with a more inclusive approach, using *HaMStR* v13.2.6 (Ebersberger, Strauss, and von Haeseler 2009, https://www.sourceforge.net/projects/hamstr). For this, each candidate HOG was aligned with *mafft* v.7.305b (Katoh and Toh 2008) and subsequently converted into an HMM with *hmmer* v.3.1b2 (Eddy 2011). The resulting HMMs were then used as core-ortholog sets for *HaMStR*, which was

run using the closest relative for each taxon as the reference species. The lengths of the orthologs found by *HaMStR* were then compared to that of the genes already present in the HOGs. HOGs for which no ortholog was identified via *HaMStR* were additionally searched using *exonerate* (Slater and Birney 2005) with the *protein2genome* model to bypass the gene annotation, directly searching in the genomic sequences instead. Each sequence present in a given HOG was aligned against all five Lecanoromycetes genomes. For each sequence we thus generated three kinds of alignment scores:

1. The *self*-alignments, in which the protein sequence was aligned to the genome in which the protein-sequence was predicted.
2. The *found*-alignments, in which the protein sequence was aligned to a genome in which an ortholog to the protein-sequence was found in the set of annotated genes.
3. The *missing*-alignments, in which the protein sequence was aligned to the genome in which no ortholog to the sequence could be found in the set of annotated genes.

We subsequently individually compared the alignment score distribution of the *missing*-alignments to that of the *found*-alignments for each HOG. A one-tailed Wilcoxon-Mann-Whitney test was performed to evaluate whether the *missing*-alignment scores were significantly lower than the *found*-alignment scores. For HOGs where we did not observe a significantly lower alignment score we additionally evaluated the positions of the *missing*-alignments using *bedtools* v2.25.0, to see whether those alignments overlap with already predicted genes or are in so-far unannotated genomic regions.

We additionally compared the lengths of the genes that overlapped with the *exonerate* alignments to the sequence lengths of proteins found in the HOGs. For *L. pustulata*, *U. florida*, *C. grayi*, and *X. parietina* we then also searched in the assembled transcript data for those HOGs that showed significantly lower

missing-alignment scores. This was done using *HaMStR* in the *–est* mode and the already compiled HMMs.

4.3 Results

4.3.1 Annotating genes

We predicted genes for *L. pustulata* using both *MAKER2* (Holt and Yandell 2011) and *funannotate* (https://github.com/nextgenusfs/funannotate). *MAKER2* predicted 10,420 protein-coding genes with a mean number of 3.26 exons per gene, and a mean gene length of 1,688 bp. *Funannotate* predicted 9,825 coding genes with 3.29 exons per gene on average, and a mean gene length 1,594 bp. Comparing the positions of the genes predicted by both methods we found that *MAKER2* predicted 1,604 genes that do not overlap with the predictions done by *funannotate*. Vice versa, we observed that *funannotate* predicted 839 genes that showed no overlap with the predictions of *MAKER2*. We subsequently used a read mapping of RNAseq data to evaluate the extent to which those genes are covered with transcriptomic data (Figure 4-1).

Figure 4-1: Fraction of the predicted genes length that is covered by RNAseq. Data is split into categories: Genes that were predicted by both methods (top), predicted only in the *funannotate* set (middle), and only in the *MAKER2* set (bottom).

For the genes that were predicted by both methods we discovered that the majority of these genes are covered by RNAseq data over the full length of the predicted transcripts, with only 2.8% of them having no position covered.

Similarly, for the genes exclusively predicted by *funannotate* we observed that only 14.1% are having no positions covered by the RNAseq data. On the other hand, for the genes predicted exclusively by *MAKER2* we found a marked increase in genes that have limited RNAseq coverage, with 69.5% of them having no position covered by a single RNAseq read.

To further investigate the genes exclusively predicted by the different methods, we used *OMA* to search for orthologs to these genes in 2 further Lecanoromycetes (*C. grayi* and *X. parietina*), 3 Dothideomycetes and 3 Eurotiomycetes each. To minimize the influence of taxon sampling, we performed the orthology prediction in three replicates, using non-overlapping sets of Dothideomycetes and Eurotiomycetes (c.f. Table 4-1, page 87).

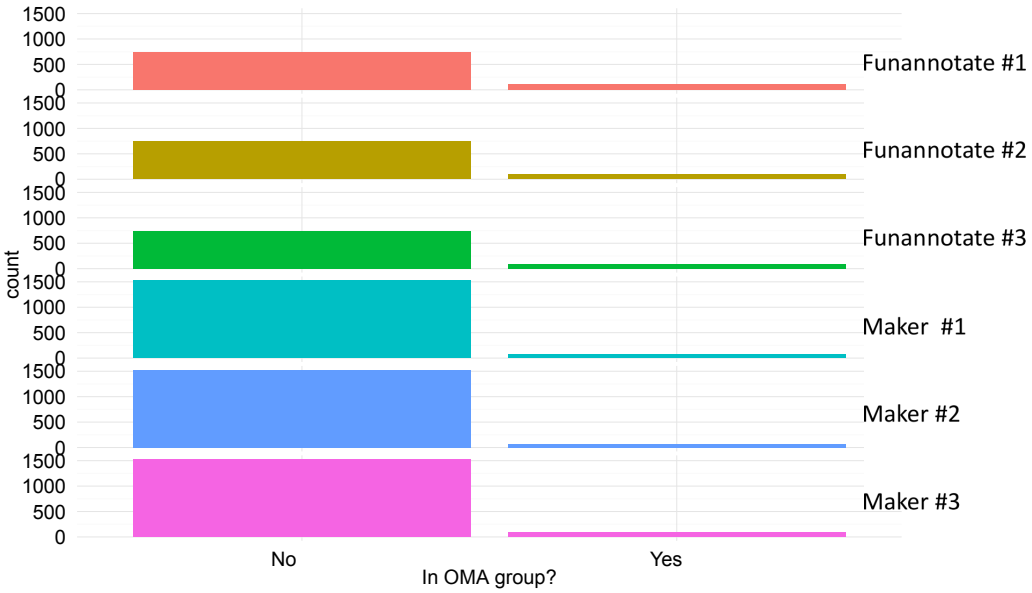


Figure 4-2: Number of exclusively predicted genes for *funannotate* and *MAKER2* that are assigned to orthologous groups for the three replicates used for the orthology prediction.

The majority of genes exclusively predicted by *funannotate* or *MAKER2* could not be assigned to any *OMA* orthologous groups (Figure 4-2). While on average 78 (4.9%) of the *MAKER2*-exclusive genes could be assigned to any orthologous group, on average 97 (11.5%) of the *funannotate*-exclusive genes were placed into an orthologous group.

Additionally, we investigated the size of the orthologous groups into which these genes were placed. This revealed that over half of the *MAKER2* genes were placed into orthologous groups that contain only a single other species (Figure 4-3). For the genes only predicted by *funannotate* this distribution markedly shifts towards larger orthologous groups, with over 50% of them being in groups that include orthologs in at least 3 additional species and only 29% of them being placed in *OMA* groups with a single other species.

Lastly, we did a preliminary search for *in silico* gene fusions/fissions with the *Rosetta Stone* method (Marcotte and Marcotte 2002). Using the protein set of *C. grayi* as a reference, we searched for *L. pustulata* genes in which distinct regions can be aligned to different *C. grayi* proteins without conflicts. Here, *MAKER2* showed larger numbers of fused genes for *L. pustulata*, with 218 predictions having significant evidence for fusions, involving 656 *Cladonia grayi* genes (see Table A-4 on page 192). In contrast, for the *funannotate* predictions we found 136 gene fusion predictions that involve 451 *C. grayi* genes (see Table A-5 on 202).

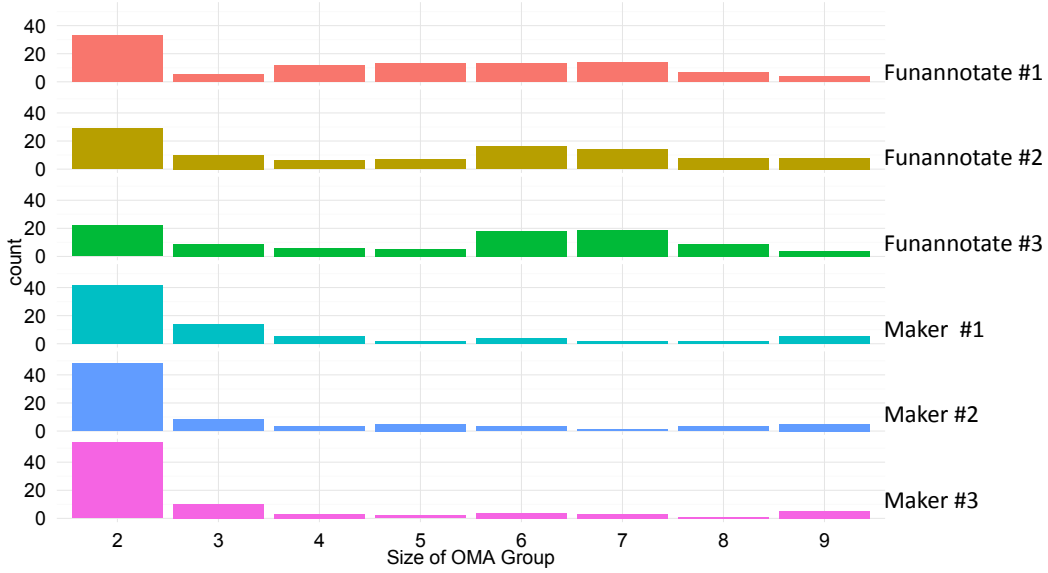


Figure 4-3: Sizes of the *OMA* orthologous groups in which the exclusively predicted genes were placed. The *OMA* orthologous groups, containing only a single sequence per species, were calculated for three taxon sets with 9 species each.

Given that genes predicted by *funannotate* appear to be of higher quality, as they are more likely to be supported by RNA sequencing data and the presence of orthologs in other fungi, we applied *funannotate* to predict genes in the unannotated draft genomes of *Umbilicaria muehlenbergii* (S. Y. Park et al. 2014) and *Lasallia hispanica* to facilitate downstream evolutionary analyses. For *U. muehlenbergii* we predicted 8,822 protein-coding genes with an average gene length of 1,739 bp and a mean number of 3.23 exons. In *L. hispanica* we predicted 8,488 coding genes (average length 1,602 bp, average number of exons 3.17). As there are no algal-specific gene annotation pipelines at this time, we predicted genes for the photobiont *Trebouxia sp.* with *MAKER2*. This yielded 14,134 protein-coding genes, which on average contain 6.8 exons for a mean gene length of 3,733 bp.

4.3.2 Repeat annotation

We annotated repetitive elements in the genomes of *L. pustulata*, *L. hispanica*, *U. muehlenbergii* and *Trebouxia sp.* with the *RepeatModeler/RepeatMasker* pipeline (Smit, Hubley, and Green 2015). For all four species only less than two percent of the genome are identified as simple repeats. Instead, larger parts of the genomes are identified as interspersed repeats (see Table 4-2). With 32.27% of the genome being classified as interspersed repeats, *L. hispanica* showed the largest proportion of them. In contrast, only 21.39% of the *L. pustulata* genome and 22.70% of the *U. muehlenbergii* genome were classified as interspersed repeats. With respect to the repeat content, the genome of the photobiont *Trebouxia sp.* deviates markedly from the genomes of the Lecanoromycetes, with only 4.87% of the genome being identified as interspersed repeats.

4.3.3 Functional annotation

We proceeded to functionally annotate the genes we predicted in the genomes of the mycobionts and the photobiont *Trebouxia sp.*. For the fungal genomes

about 1/3rd of genes were linked to KEGG orthologous groups (see Table 4-2). In addition, we observed a substantial overlap among the KO groups present in the 5 Lecanoromycetes genomes (see Figure 4-4).

Table 4-2: Annotation results for the four genomes, including the repeat and functional classifications.

Taxon	Number of Genes	Genome in Interspersed Repeats	Genes with GO Term assignments	Genes linked to KEGG Orthologous Groups	Genes with Pfam domains
L. pustulata	9,825	21.39%	4,718	3,226	6,241
L. hispanica	8,488	32.27%	4,096	2,812	5,380
U. muehlenbergii	8,822	22.70%	4,452	3,118	6,099
Trebouxia sp.	14,134	4.87%	4,970	3,474	7,067

Using the Gene Ontology (GO) for a further functional classification, we could annotate around 50% of the genes predicted in the fungi with at least one GO term (see Appendix, Figure A-4, as an example of the GO terms found in *L. pustulata*). Analogous to the KO groups, we observed that nearly all GO terms were found in all five lecanoromycete genomes (the overlap in GO terms found amongst the lecanoromycete genomes is shown in Appendix, Figure A-5). Furthermore, about 2/3rd of the fungal genes could be annotated with Pfam domains. In line with the GO terms and the KO assignments, these are largely shared between the five Lecanoromycetes (see Appendix, Figure A-6, for the overlaps in Pfam annotations between the Lecanoromycetes).

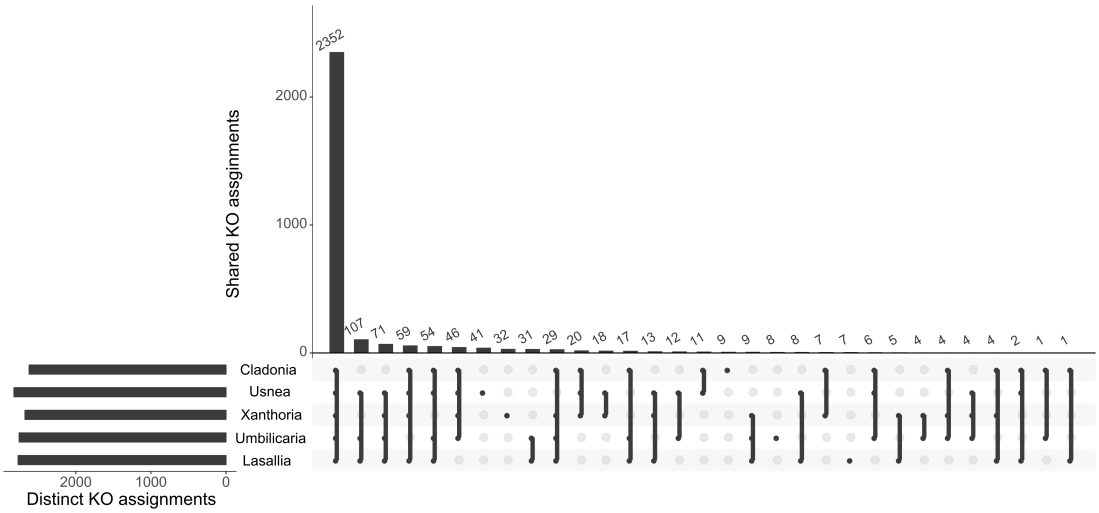


Figure 4-4: The bottom left bars give the total number of distinct KOs assigned to the five genomes of the Lecanoromycetes. The top bar charts show the number of KO assignments shared between taxa. Black dots denote the inclusion of an organism in the respective intersection.

For the genome of the green alga *Trebouxia sp.* 24.6% of the genes were placed into KEGG orthologous groups. Furthermore, 35.2% of the genes could be assigned to GO terms (see Figure A-7 in the Appendix for a graphical representation of the GO terms found in *Trebouxia sp.*), and 50% contained Pfam domains. Given the overall low number of functionally annotated genes, we did not compare the KO and GO annotations of *Trebouxia sp.* with further Chlorophyta in detail, but see Appendix Figure A-8 (page 218) for the overlap in Pfam annotations between *Trebouxia sp.* and 5 further Chlorophyta that were annotated by Nelson et al. (2017).

4.3.4 Surveying the effects of assembly and annotation errors

Having identified repeats as well as genes and additionally having annotated the functions of the genes, we performed the typical steps of a genome annotation. Missing genes – and functions – thus would be interpreted as evolutionary loss events (c.f. 1.2, page 3). However, this would take the annotation results of face value, ignoring the influence of genome assembly and genome annotations artifacts, which can lead to wrong evolutionary inferences (Denton et al. 2014). We thus evaluated the impact of assembly and annotation artifacts on downstream analyses by searching for genes that are

well conserved in the Lecanoromycetes. The previously analyzed *OMA* orthologous groups (c.f. 4.2.2, page 86) require that all genes in a group are orthologous to each other, making it very stringent and leading to a low recall (Altenhoff et al. 2016). We thus changed our focus to Hierarchical Orthologous Groups (HOGs) as predicted by *OMA*. HOGs, which can include in-paralogs, have been shown to have a good trade-off between recall and specificity (Altenhoff et al. 2016). Searching for orthologs between *L. pustulata*, *U. muehlenbergii*, *X. parietina*, *C. grayi* and *Usnea florida* we found a total of 9,081 HOGs. Of these, 4,607 HOGs contained sequences of all five taxa. Furthermore, we found 1,402 HOGs that contained sequences of 4 of the 5 taxa. Given our taxon sampling, these genes are present in two clades that coalesce in the last common ancestor of the Lecanoromycetes (LCA_{Lec}) and we thus assume that these were already present in the LCA_{Lec}. Furthermore, given their presence in all but one taxon, they are rarely lost. We find substantial numbers of genes that appear to be privately lost genes for all five taxa, ranging from 193 in *U. florida* to 371 in *X. parietina* (Table 4-3).

Both the quality of gene predictions in individual genomes, as well as the stringency of the orthology prediction can influence these numbers. In a first survey, we used the *in silico* gene fusions and fissions between *L. pustulata* and *C. grayi* that were predicted by the *Rosetta Stone* method to estimate how many genes are being missed based on those. We observed that 15 of the 136 artificial gene fusions found in *L. pustulata* involve genes of *C. grayi* that are placed in HOGs where *L. pustulata* is exclusively missing (see Appendix, Table A-5 on page 202). To further estimate the influence of gene fusions as well as otherwise missed genes, we performed a two-step, refined search for the LCA_{Lec} genes. In the first step we searched for the absent genes with the more inclusive, targeted orthology prediction method *HaMStR*, which does not apply a sequence-length cut off during the orthology search (Ebersberger,

Strauss, and von Haeseler 2009). This identified further orthologs for around half of the genes that so far were predicted to be privately lost (Table 4-3).

Table 4-3: Predicted gene losses after searching for orthologs using *OMA*, *HaMStR* and *exonerate*.

Taxon	Missing LCA_{Lec} orthologs		
	after *OMA* search	after *HaMStR* search	after *exonerate* search
L. pustulata	325	142	76
U. muehlenbergii	228	126	36
C. grayi	285	157	58
U. florida	193	90	55
X. parietina	371	203	159

Analyzing the differences in lengths between the additional genes found by *HaMStR* and the sequences already present in the corresponding HOGs, we observed marked length differences for the majority of the genes, regardless of the taxon. Figure 4-5 shows that a significant number of the newly found orthologs differ by a length larger than the length-cutoff employed by *OMA*. Additionally, we noted that 5 of the orthologs found for *L. pustulata* match gene fusions predicted by the *Rosetta Stone* method (see Appendix, Table A-5 on page 202).

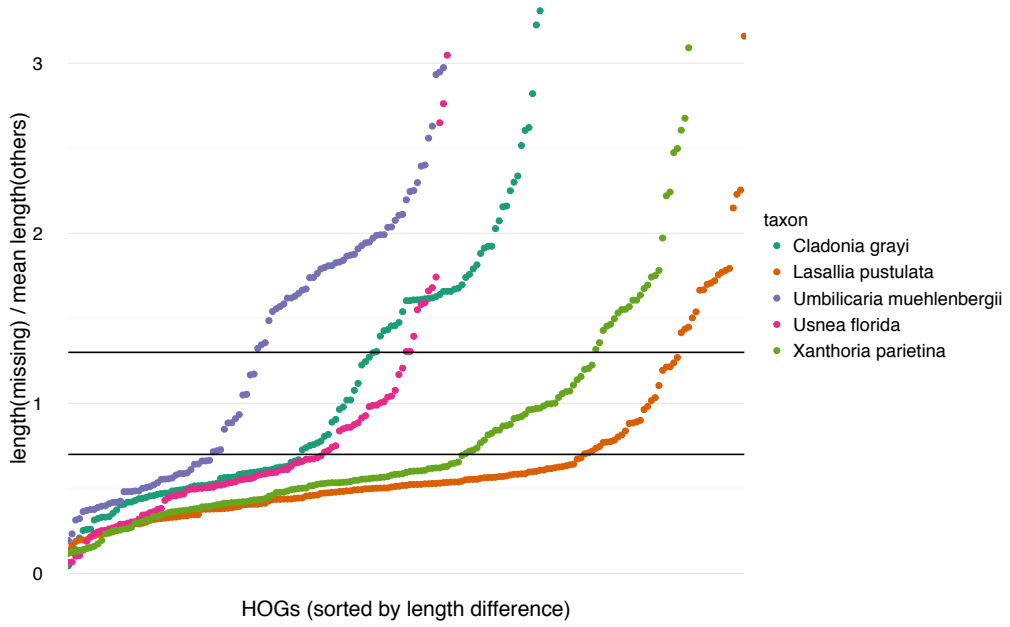


Figure 4-5: The relative sequence lengths of the additional orthologs found by *HaMStR*, compared to the mean sequence length of the orthologs already in a given *OMA* HOG. The horizontal bars give the upper and lower bound of the length cut-off filter used in the ortholog identification of *OMA*.

In the second step, we additionally screen the genomes for genes that have been overlooked by the gene prediction methods. To this end we focus on the genomes themselves rather than on the set of predicted genes. We searched for the still privately absent LCA_{Lec} genes with *exonerate*; aligning all sequences in a given HOG to the genome sequences of all five Lecanoromycetes. The protein alignment scores for the genomes in which an ortholog was found were then compared to the protein alignment scores for the genome in which so far no ortholog was found (see online supplementary material at doi:10.5281/zenodo.894741 for the *exonerate* alignment scores), using a one-tailed Wilcoxon-Mann-Whitney test. Significantly worse ($p \leq 0.05$) alignment scores in the genome in which no ortholog was found were taken as further evidence for a true absence of an ortholog. Genes that did not show a significantly worse alignment were classified as either a gene missed in the gene prediction – if they do not overlap an already annotated gene – or otherwise as a potentially missed ortholog. This approach uncovered further genes that were initially predicted to be absent in the respective genomes. We observed that this approach yielded not only evidence for additional orthologs, which were not uncovered so far, but also genes that were not in the set of predicted proteins (Table 4-4). After this filtering we notice that only between 15% (*U. muehlenbergii*) and 42% (*X. parietina*) of the original predicted losses could be confirmed.

Table 4-4: *Exonerate* evidence for orthologs that were found by neither *OMA* nor *HaMStR*. If the exonerate evidence overlaps with already annotated genes, this was classified as a potentially missed ortholog. If the exonerate evidence hints to a so far unannotated genomic region, it is classified as a missed gene annotation.

Taxon	Missed Ortholog	Missed Gene Annotation
L. pustulata	39	27
U. muehlenbergii	65	25
U. florida	15	17
C. grayi	43	59
X. parietina	28	16

Analogous to the additional orthologs found by *HaMStR*, we compared the lengths of the orthologs found by *exonerate* to the lengths of the sequences in a given HOG. We limited this to the not significantly worse *exonerate* hits that span a region in which genes were already predicted. We again note substantial length differences, with the sequences found by *exonerate* on average being longer than the sequences found in the HOGs. This especially affected the gene predictions of *U. muehlenbergii*, where all but one additional ortholog are longer than the average gene in the corresponding HOG. We additionally observed that 9 of the *L. pustulata* genes found were predicted as gene fusions by the *Rosetta Stone* method (see Appendix, Table A-5 on page 202).

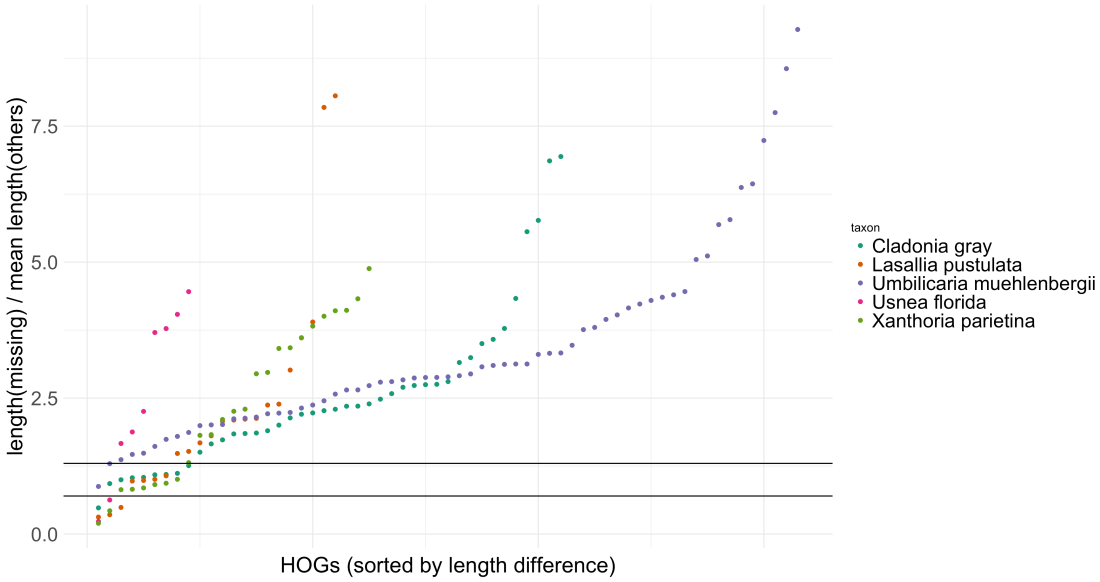


Figure 4-6: The relative sequence lengths of additional sequences found by *exonerate*, compared to the mean sequence length of the orthologs already in a given *OMA* HOG. The horizontal bars give the upper and lower bound of the length cut-off filter used in the ortholog identification of *OMA*.

After estimating the impact of lacking orthology and gene predictions, we additionally analyzed the impact of incomplete genome assemblies. By searching for orthologs to the so far not identified LCA_{Lec} genes in the assembled transcriptomes of all Lecanoromycetes, except *U. muehlenbergii*, where no transcript data was available, we further reduced the number of the

privately lost genes. This yielded a loss of 33 LCA_{Lec} genes for *L. pustulata*, 45 for *C. grayi*, 52 for *U. florida*, and 90 for *X. parietina*.

4.3.5 Tracing the sources of annotation artifacts

To gain further insights on why the gene prediction or orthology search failed, we manually curated individual examples of genes initially predicted to be absent in *L. pustulata* but subsequently found in our search for the privately lost LCA_{Lec} genes. Genes being missed during the gene annotation, despite being in a sequenced and assembled region, is one potential source for false-positive gene loss predictions. An illustrative example is shown in Figure 4-7. Despite the presence of RNAseq data, *funannotate* did not predict a gene at this position. The *exonerate* analysis suggests that this happened for 27 LCA_{Lec} genes in *L. pustulata*.

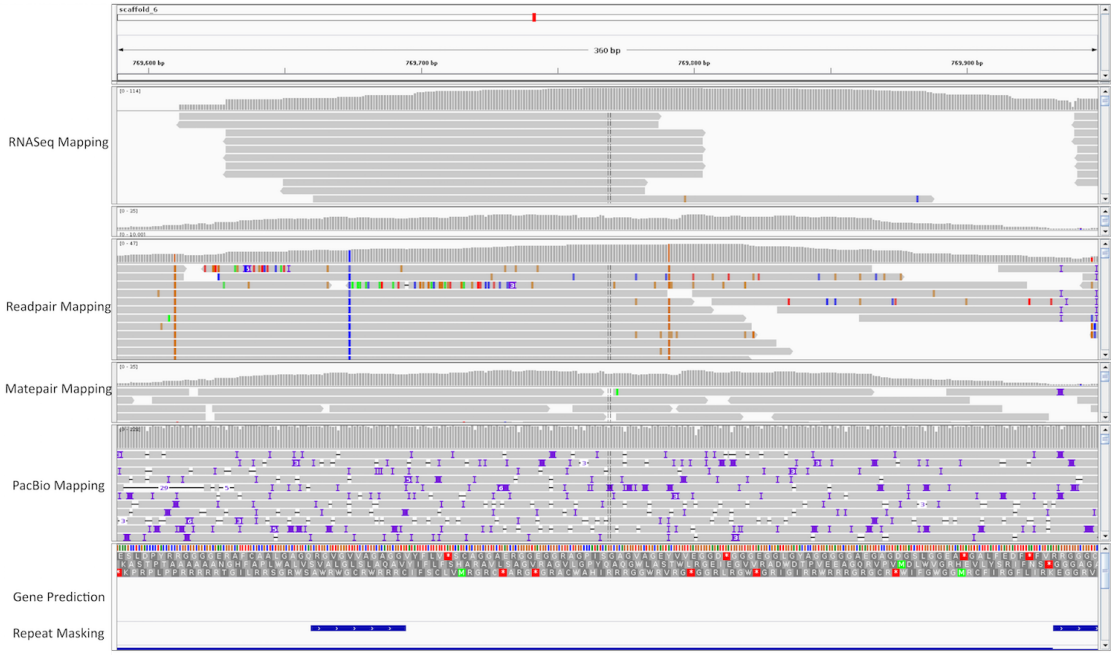


Figure 4-7: A 360 bp-long region in which *exonerate* successfully identified a so far overlooked gene. The top track shows the presence of mapped RNAseq data with a read coverage of up to 114x, which was used to guide the gene prediction.

Given the results of the *Rosetta Stone* method (see 4.3.1 on page 91) and the length discrepancies found in the orthologs discovered by *HaMStR* and *exonerate*, we investigated potential *in silico* gene fissions/fusions as well. In

such cases the gene prediction either falsely joins two neighboring genes or cuts a gene into parts. We observe that the former often happened when the intergenic space between these genes is small. Figure 4-8 shows an example of this, where two genes were fused over the intergenic region despite mapped RNAseq reads not supporting this. Here *funannotate* predicted the two terminal exons of the left and right gene to be internal exons of a joint gene.

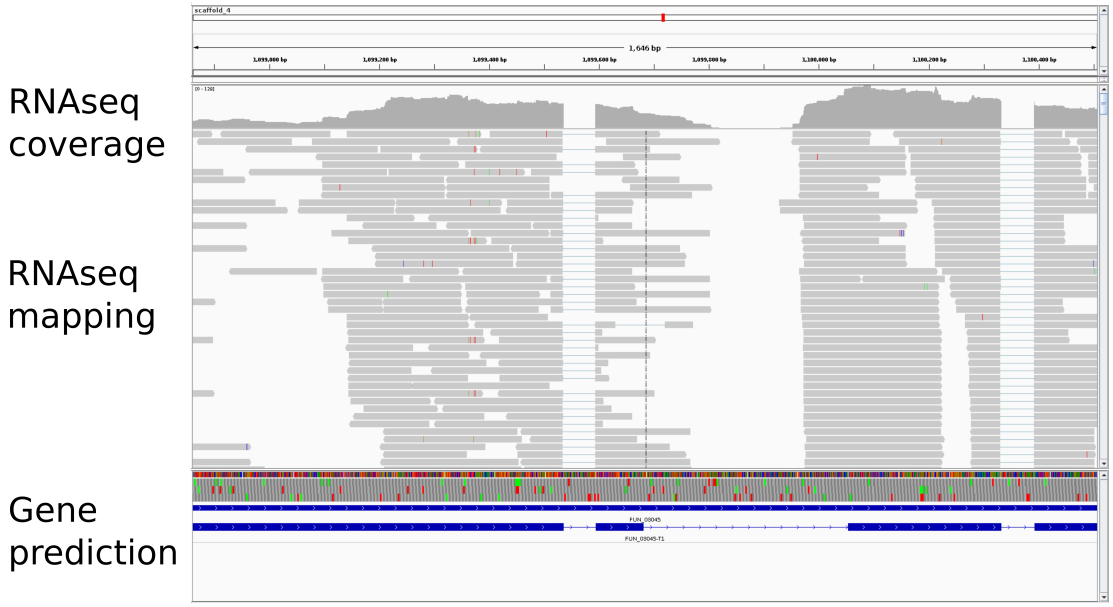


Figure 4-8: A 1,646 bp window in which two genes were fused. The gene prediction track (bottom track) shows four exons that are joined into a single gene. The RNAseq coverage (top track), as well as the RNAseq mapping shows that there are no read mappings which support joining the two central exons into a single gene. Despite this, *funannotate* predicted a 109 bp long intergenic region to be an intron.

In other cases we found that the coverage for the RNAseq data on first glance appeared to support the joining of genes. For example, Figure 4-9 shows a case where *funannotate* predicted a single gene with a long terminal exon. A closer look at the RNAseq coverage revealed that the left and right ends of this exon differ substantially with respect to the coverage. Furthermore, the central region of the exon showed a substantially lower coverage than both flanking regions and there is no RNAseq read that bridges this 180 bp long low-coverage region, despite an RNAseq read length of 250 bp.

Figure 4-9: A 2,465 bp region in which *funannotate* (bottom track) predicted a single, terminal exon. The RNAseq coverage (top track) decreases to 10x in the central region, but does not reach zero. The region to the left of this low-coverage area shows a lower mean coverage (63x) than the right flanking region (150x). The RNAseq read alignment shows no reads fully span over the 180 bp long, lowly covered region. Thus, both the coverage pattern as well as the mapping providing strong evidence for an overlapping UTR and thus an artificial gene fusion.

For *L. pustulata*, 33 LCA_{Lec} genes appeared absent after the additional analyses using *HaMStR*, *exonerate* and additional RNAseq data. As these are evolutionary old, the absence of these genes could have considerable functional implications. We thus manually curated each of these genes.

This revealed another source of error, which was not detected by the previous methods, impacting a total of 5 LCA_{Lec} genes. The best example of these is the dihydrofolate reductase (DHFR), a gene involved in the basal metabolism of nucleotides. As such its absence would imply wide-ranging changes in the metabolism of *L. pustulata*. By searching for the neighboring genes to the DHFR in *L. hispanica* we identified the corresponding genomic region in *L. pustulata* (Figure 4-10). In that genomic region we observed a marked lack of read coverage for our *Illumina* sequencing data, with only individual reads mapping to it, while the coverage for the *PacBio* reads remained high.

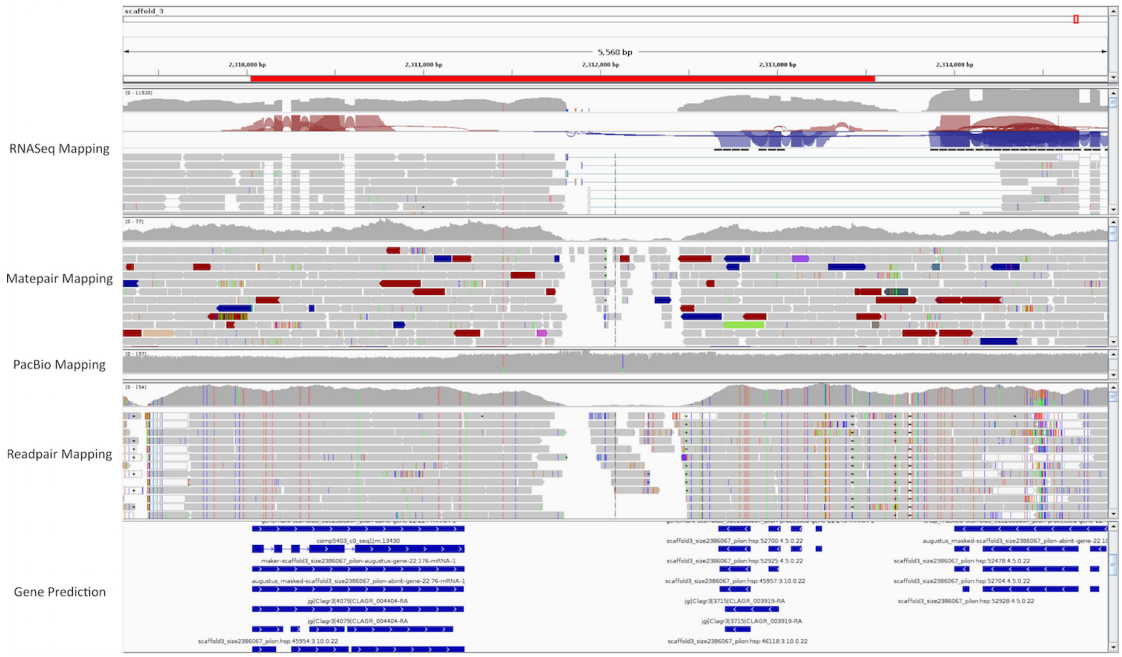


Figure 4-10: The 5,568 bp window where the DHFR should be located; centered on the position of the DHFR. The RNAseq track shows the coverage (top), spliced alignments (red/blue joins) as well as the read mapping. For the *Illumina* read pairs and mate pairs the read coverages as well as the read mappings are shown. All *Illumina* libraries show a lack of coverage for the central region, where the DHFR would be located. The *PacBio* sequences, for which only the read coverage is shown, show no decrease in coverage. The *funannotate* gene predictions show the genes flanking the DHFR to the left and right as correctly predicted.

A *genewise* (Birney, Clamp, and Durbin 2004) alignment of DHFR sequences against this genomic region revealed a high-scoring alignment (230.82 bits) in the region, showing four indels. The resulting frame shifts in the DHFR coding sequence would render it a pseudogene. However, the alignment of the sparsely available *Illumina* reads indicates that these indels are artifacts caused by sequencing errors which are found in around 1/3rd of the *PacBio* reads. Correcting the genomic sequence guided by the *Illumina* reads showed the gene to be intact, with no evidence for a pseudogenization. The region of the DHFR exhibited a high G/C content of over 70%, with *RepeatMasker* predicting the region as repetitive. We furthermore observed that the *Inverted Repeat Finder* (Warburton et al. 2004) predicted an inverted repeat (IR) in the genomic area surrounding the DHFR gene.

To assess the extent to which G/C-rich inverted repeats could impact the correct identification of genes, we investigated the inverted repeats found in

L. pustulata, analyzing their G/C-content and length. We repeated this analysis for the other four Lecanoromycetes to compare the IRs found in *L. pustulata* to the other genomes. For *L. pustulata* we found 1,464 IR, with a median length of 819.5 bp. We observed marked differences in the number of IR between the Lecanoromycetes, with a lower bound of 670 in *X. parietina* and an upper bound of 29,396 in *C. grayi* (see Table A-6 on page 207 in the Appendix). Furthermore, there was a strong bias in G/C content for the observed IR. *Cladonia grayi* showed the lowest median G/C content, with 11.13%. In contrast, the G/C content of the IR found in *L. pustulata* and *U. muehlenbergii* was bimodally distributed. While we observed peaks at 51% and ~75% G/C for *L. pustulata*, the peak with the highest G/C content for *U. muehlenbergii* was at ~60% (Figure 4-11). Generally, only the genome of *L. pustulata* showed a substantial amount of IR with a G/C content of over 70%. Additionally, we observe that 467 IR – with a mean G/C content of 67.8% – overlap regions in which the read coverage for the *Illumina* libraries drops to <10x while the *PacBio* coverage remains uniform, analogous to the case of the DHFR (see Appendix, Figure A-10 on page 219, for a length and G/C content distribution of these regions).

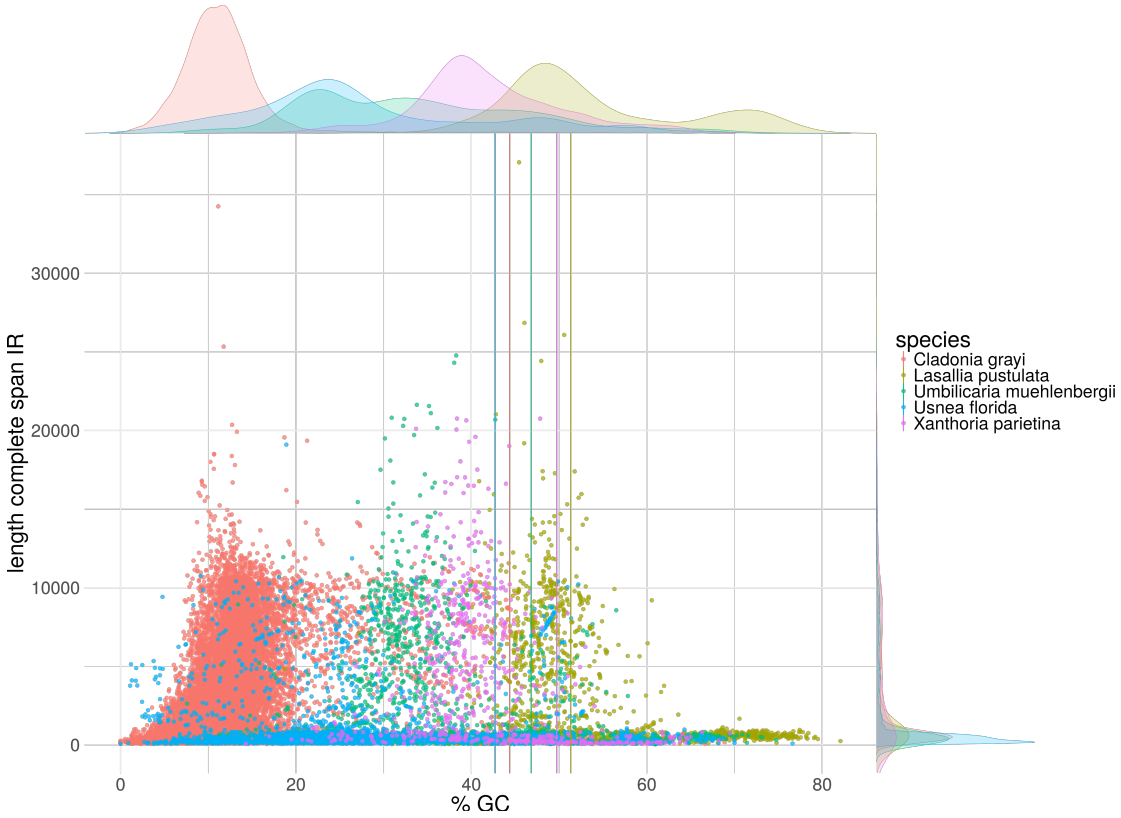


Figure 4-11: G/C content and length of the inverted repeats for the five Lecanoromycetes. The marginal plots give the kernel density estimates for each taxon. Vertical lines show the mean G/C content for the individual genomes.

4.4 Discussion

The annotation of genomes includes a variety of tasks, from annotating repetitive elements (Price, Jones, and Pevzner 2005) and motifs (D'haeseleer 2006) to the prediction of gene structures (Klasberg, Bitard-Feildel, and Mallet 2016). It is the latter that often remains a central element in making sense of newly sequenced genomes, as it allows a subsequent functional annotation of these genes (Finn et al. 2016; Kanehisa, Sato, and Morishima 2016; Ashburner, Ball, and Blake 2000). Furthermore, the search for potentially orthologous sequences to the newly predicted genes is frequently used to transfer functional annotations between those sequences (Conesa et al. 2005; Kanehisa, Sato, and Morishima 2016), motivated by the ortholog conjecture (Altenhoff et al. 2012; Studer and Robinson-Rechavi 2009).

4.4.1 Annotating genomes

As genome organization and gene structures differ between taxa (Yandell and Ence 2012), we utilized a number of different methods to annotate the hologenome of the lichen *Lasallia pustulata*. These differences are reflected in the repeat content for the genomes.

4.4.1.1 Annotating repeats

For the genome of the alga *Trebouxia sp.* we identified a low interspersed repeat content, with less than 5% of the genome being identified as interspersed repeats. This is contrasted by the fungal genomes where between 21.39% in *L. pustulata* to 32.27% in *Lasallia hispanica* of the genome were classified as interspersed repeats. This lower repeat content for *Trebouxia sp.* is in line with the observation made for the genome of the green alga *Chlorella variabilis* NC64A, where a total of 12% of the genome was found to be repetitive, including not only interspersed repeats but also gene duplications (Blanc et al. 2010). The surprising variability between the two closely related

members of *Lasallia* might reflect a real growth of repetitive elements in *L. hispanica*, as seen for other fungi (Muszewska et al. 2011), though it might also be due to misassemblies in the genome of *L. hispanica*. Misassemblies frequently lead to an artificial repeat expansion that not only fragments the overall assembly but also inflates the genome assembly size (Phillippy, Schatz, and Pop 2008). The genome of *L. hispanica* is more fragmented (N50: 137 Kbp) than the genome of *L. pustulata* (N50: 1.8 Mbp) and exhibits a larger assembly size (*L. hispanica*: 41 Mbp, *L. pustulata*: 33 Mbp). Furthermore, the observed repeat content for *Umbilicaria muehlenbergii* reflects that of *L. pustulata*. Thus, misassemblies in *L. hispanica* are the more likely reason for the observed repeat expansion.

4.4.1.2 Annotating genes

While the structure of fungal genes is conducive to annotating them (Galagan et al. 2005), their dense spacing in the genome leads to short intergene regions, which have been shown to make correct predictions challenging. This has lead to the development of dedicated methods for the annotation of fungal genomes (Testa et al. 2015; Ter-Hovhannisyan et al. 2008). For the prediction of genes in the fungal genomes we thus compared the performance of *MAKER2* (Holt and Yandell 2011), a general purpose genome annotation pipeline, with that of *funannotate*, a pipeline specifically designed for annotating fungal genomes. Internally, *funannotate* utilizes *BRAKER1*, which has been shown to outperform *MAKER2* as well as dedicated fungal gene predictors in the annotation of *Schizosaccharomyces pombe* (Hoff et al. 2016). While *MAKER2* predicted 1,604 genes not found in the annotation of *funannotate*, we observed that the vast majority of these were not supported by our existing RNAseq data (Figure 4-1) and furthermore no orthologs could be found for them (Figure 4-2). On the other hand, *funannotate* had a smaller number of privately predicted genes, but a larger fraction of those could be

supported by RNAseq coverage and additionally be placed into Hierarchical Orthologous Groups (HOGs). Furthermore, the preliminary search for gene fusions with the *Rosetta Stone* method (Marcotte and Marcotte 2002) showed that the *MAKER2* more often joined separate genes than *funannotate* (see Table A-4 on page 192 and Table A-5 on page 202). For this reason we subsequently annotated the genes of *Lasallia hispanica* and *Umbilicaria muehlenbergii* with the *funannotate* pipeline. Given the lack of dedicated gene prediction pipelines for green algae, we annotated the genome of Trebouxia sp. with *MAKER2.* We observed that gene models of *Trebouxia sp.* on average have twice the size of the models for *L. pustulata* and contain twice the number of exons. These differences in gene structure resemble what has been found for the genes of fungi and algae (Merchant et al. 2010; Galagan et al. 2005).

4.4.1.3 Functionally annotating genes

Our functional annotation recovered Gene Ontology (GO) terms for around half of the predicted fungal genes, as well as KEGG orthology (KO) assignments to around 1/3rd of them (Table 4-2). The methods for the assignment of both, KO and GO identifiers, rely heavily on sequence similarity to already annotated sequences (Kanehisa, Sato, and Morishima 2016; Conesa et al. 2005). A lack of reference sequences thus potentially hinders the functional annotation. This is demonstrated by the functional annotation of the genes for the green alga *Trebouxia sp.*, for which less reference data exists (Bhattacharya et al. 2015). Here, only 24.6% of the *Trebouxia sp.* genes could be assigned to a KO group and only 35.2% could be annotated with GO terms. Evidence for the lack of reference data for the functional annotation of the fungal genes comes from the orthology predictions for the Lecanoromycetes. For the 9,825 predicted *L. pustulata* genes we found that 2,381 do not have orthologs to any other

Lecanoromycetes and further 906 genes only have an ortholog in a single other lecanoromycete species.

4.4.2 Gene annotation artifacts and downstream impacts

Generally, the results of a typical genome and gene annotation (c.f. 4.3.1 on page 91 and 4.3.3 on page 94) would be used for comparative genomic studies (Sharpton et al. 2009). The absence of genes and their respective function are then interpreted as evolutionary losses (c.f. 1.2, page 3). Given that our subsequent analyses (c.f. 5.3 on page 125) are focused on loss events, which are known to be strongly influenced by annotation artifacts (Dunne and Kelly 2017), we further estimated the impact of gene annotation and assembly artifacts on our downstream analyses. For this reason we searched for privately absent orthologs, which are otherwise conserved among the Lecanoromycetes. Based on the presence in 4 out of 5 Lecanoromycetes we assumed these to have been present in the last common ancestor of the Lecanoromycetes (LCA_{Lec}). The accurate prediction of such loss events is of evolutionary interest, as such conserved genes potentially play a key biological role for the studied organisms, usually not allowing for a loss (Zhao et al. 2015; Ptitsyn and Moroz 2012). At the same time the accurate prediction of these loss events is problematic, as the detrimental nature of such losses makes them rare events, increasing the likelihood of observing false positives, analogous to the high false-positive rates observed when predicting loss-of-function mutations (MacArthur et al. 2012). This is indeed what we observed by using increasingly sensitive methods to search for those "privately lost" LCA_{Lec} genes. Over the five genomes of the Lecanoromycetes we find that in total only 15% of the initial predictions could not be rejected as false positives.

4.4.2.1 Not identified orthologs

Our search framework allowed us to investigate the reasons for the false positive predictions. In total, 80% of gene loss predictions were due to genes

that were in the predicted protein sets but not found during the orthology prediction (Table 4-3 on page 98 and Table 4-4 on page 99). The use of a more sensitive orthology prediction method, *HaMStR*, could partially ameliorate this problem. We found that the orthologs missed by *OMA*, but detected by *HaMStR*, differ in part substantially in length from the other sequences in the orthologous group (Figure 4-5). Therefore, the length cutoff implemented in *OMA* (http://omabrowser.org/standalone/#parameters) prevented the addition of these sequences into the orthologous group. In contrast, *HaMStR*, which does not include an explicit length cutoff, accepted these sequences as orthologous to the other group members.

A further investigation of these cases showed that many of these discrepancies are a consequence of gene prediction artifacts, which either split genes or falsely fuse neighboring genes, with fused genes being the result of small intergene spaces (Figure 4-8 on page 102). For some of the fused genes we observed apparently consistent RNAseq coverage (Figure 4-9 on page 103). This is a result of overlapping untranslated regions (UTRs). These are a common feature in fungal genomes (Donaldson et al. 2017; Xu et al. 2009) and have previously been seen to interfere with gene predictions (Testa et al. 2015).

For the missing LCA_{Lec} genes of *U. muehlenbergii* in particular we find that both *HaMStR* and *exonerate* uncover sequences that vary markedly in length when compared to the other members of the corresponding orthologous group. This is most likely a result of the gene annotation, as no RNAseq data for *U. muehlenbergii* was available and only data from *L. pustulata* could be used. It appears that due to this the reliable identification of gene starts and ends was hindered, leading to a large number of *in silico* gene fusion events.

4.4.2.2 Incomplete genomes

Furthermore, all the genomes we investigated were draft genomes, with an unknown fraction of the genome not represented in the draft sequence. Approaches to estimate the completeness of draft genomes have been developed. These use the number of found core eukaryotic genes (Parra, Bradnam, and Korf 2007) or lineage specific universal single copy genes (Simão et al. 2015) as a proxy for the draft genome completeness. While these methods can give a first insight into the quality of a draft genome, their use is problematic when expecting the loss of evolutionary old and well-conserved genes as such methods would interpret such losses as evidence of an incompletely sequenced genome. Using transcript data we could identify false positive loss predictions that were due to the absence of the corresponding regions in the different genome assemblies. This underlines the usefulness of transcriptome data to identify gene losses (c.f. Guzman and Conaco 2016).

4.4.2.3 Not predicted genes

We additionally observed that missed gene annotations were another source of false LCA_{Lec} gene loss predictions. Substantial numbers of genes were found when bypassing the gene annotations and searching on the genomic sequences. Similar incomplete annotations were observed in earlier studies (Veeckman, Ruttink, and Vandepoele 2016; Denton et al. 2014). To minimize such artifacts, prediction methods that either make more extensive use of aligned homologous sequences (van der Burgt et al. 2014; Dunne and Kelly 2017) or of community curation efforts (E. Lee et al. 2013) have been developed. Such methods can also be helpful to find annotations missed due to assembly errors, which can prevent the prediction of genes.

4.4.2.4 Compositional sequencing biases

Investigating the LCA_{Lec} genes predicted to be lost after all of our quality checks, we identified an extreme example of such an inaccurately reconstructed genomic region. In case of *L. pustulata* the set of predicted lost genes contained the dihydrofolate reductase (DHFR), which is essential for purine and thymidylate synthesis (Schnell, Dyson, and Wright 2004). Due to its central role in basic cell growth a loss of the DHFR is highly unlikely. Doing an extensive search for it, we found the DHFR, which was predicted as lost due to a missing annotation. This was, in turn, a result of a pathogenic assembly error that introduced insertion/deletions (indels), which led to premature stop codons (Figure 4-10). The sequence-content of the DHFR and its genomic neighborhood offer an explanation for these errors. The complete region consists of over 70% guanines and cytosines. Additionally, the high G/C content results in runs of homopolymeric stretches of Gs and Cs, which are arranged in a way that they form short inverted repeats. A high G/C content as well as homopolymers have been found to lead to increases in sequencing errors in both *Illumina* and *PacBio* sequencing data (Schirmer et al. 2016; Ross et al. 2013). Additionally, high-G/C regions lead to PCR-induced biases in the library preparation for *Illumina* sequencing, which results in corresponding regions not being sequenced (Benjamini and Speed 2012). Inverted repeats in itself have furthermore been shown to lead to an increase in *Illumina* sequencing errors (Nakamura et al. 2011; Allhoff et al. 2013; Star et al. 2014). Regions that are not only G/C-rich but also contain inverted repeats, which largely made up of homopolymers, correspondingly amplify the sequencing problem. It is likely that due to these biases virtually no *Illumina* reads for such regions could be sequenced, making the DHFR appear to be absent even in the transcriptome data. Due to this overall lack of *Illumina* data, the short-read based error correction of *Pilon* (Walker et al. 2014) could not identify indels at this position. As such, even the sequencing of genomes

to a high depth and using multiple sequencing methods is not always a solution to reliably identify the loss of predicted genes.

The genome of *L. pustulata* was furthermore the only lecanoromycete genome to show a large number of IRs with a G/C content of over 70%, while even its close relative *U. muehlenbergii* only included IRs with a G/C of about 60% (see Figure 4-11 on page 106). The used sequencing techniques offer an explanation for this. To our knowledge only genome of *L. pustulata* was sequenced using *PacBio* long reads. The genomes of all other Lecanoromycetes were sequenced using only short reads. Due to this it is likely that regions that include G/C-rich IRs were not sequenced for these genomes and are thus absent from the final assembly. Similar effects of G/C-rich regions were recently reported for avian genomes, with around 15% of the genomes not being contained in the final assembly, largely due to G/C-biases during the sequencing and assembly (Botero-Castro et al. 2017).

4.5 Conclusion

The annotation of genomes is a central part of every genomic study. It forms the basis of comparative evolutionary analyses, such as the investigation of individual gene families, reconstructing phylogenies and inferring the functional impact of gene gain and losses, relating to changes in habitats and lifestyles. We annotated the hologenome of *L. pustulata*, largely focusing on the mycobiont due to the availability of reference data. Despite the considerably dense fungal taxon sampling, the *in silico* annotation of genes and their function remains challenging. In particular, the composition of genomic sequences can bias the initial sequencing, leading to errors that interfere with the prediction of genes. These errors often remain unnoticed, as the corresponding regions, like G/C-rich inverted repeats in case of *L. pustulata*, will not be represented in the draft genomes at all. Our in-depth comparative analysis of the gene predictions of 5 Lecanoromycetes showed that all draft genomes were only incompletely annotated, strongly impacting the reliable prediction of gene losses. We thus find that Martin Rees' idiom – *"Absence of evidence is not evidence of absence"* (Oliver and Billingham 1971) – applies to the prediction of gene losses as well, especially when searching for unlikely or rare events.

5 Evolutionary consequences of lichenization

5.1 Introduction

The establishment of symbioses frequently leads to interdependences between the involved organisms (F. Martin, Uroz, and Barker 2017). Such interdependence leaves footprints in the genomes of the participating symbionts, with their genomes being remodeled to adapt to the requirements of the symbiosis (A. Moya et al. 2008). Such genomic footprints of symbiosis frequently include adaptations of the secretome (F. Martin et al. 2008), changes in gene family sizes (Dahan et al. 2015; Duncan et al. 2016; Zuccaro, Lahrmann, and Langen 2014), and large-scale reductions in gene set sizes (Bennett et al. 2014; Sabree, Degnan, and Moran 2010; Ochman and Moran 2001).

5.1.1 Footprints of symbioses in mycorrhizal fungi

Genomic adaptations to symbiosis have been studied in mycorrhizal fungi. It has been shown that they need their secretomes to establish and maintain a mutualistic symbiosis with their hosts (Garcia et al. 2015; F. Martin et al. 2008), using a set of genes that is partially homologous to pathogenic fungi for the invasion of the plant partner (Tollot et al. 2009; Heupel et al. 2010). Similar to secretome sizes, the evolution of gene families of fungi shows evidence for host-specific expansions and contractions of individual families.

Lifestyle-specific changes in gene family sizes of mycorrhizal fungi have been observed, leading to the proposal of a symbiosis-toolkit (Kohler et al. 2015): For the genomes of *Laccaria bicolor* and *Rhizophagus irregularis* gene families which are involved in protein-protein interactions and signal transduction have been found to be expanded (Tisserant et al. 2013; F. Martin et al. 2008).

Furthermore, both genomes display significant contractions of gene families that are associated with the degradation of plant cell walls, which is supposed to support their mycorrhizal lifestyle. Despite the absence of genes involved in plant cell wall degradation, the genome of *L. bicolor* encodes enzymes for the degradation of other polysaccharides, potentially because of its lifestyle that is both saprotroph and mycorrhizal (F. Martin et al. 2008). The obligate mycorrhizal fungus *R. irregularis,* on the other hand, additionally lacks further degrading enzymes, secondary metabolic enzymes and secreted invertases as well as sucrose transporters. It has been hypothesized that these losses are linked to its inability to grow *in vitro* (Tisserant et al. 2013). Given the similarities of lichens to the mycorrhizal symbioses (Ahmadjian and Jacobs 1981; Ahmadjian 1993; Lücking et al. 2009), it appears likely that similar effects should be visible in the genomes of lichen symbionts.

5.1.2 Prerequisites for identifying genomic footprints of symbiosis

The comprehensive investigation of the genomic effects of symbiosis relies on the availability of well-annotated reference genomes that can be used in comparative studies (C. W. Dunn and Munro 2016). A dense taxon sampling is needed to allow for a temporal resolution that allows to resolve when genes were gained and lost (Martín-Durán et al. 2017; Havird and Miyamoto 2010). For lichen symbionts the availability of reference data is heavily skewed, with notoriously few algal genomes being sequenced to date (Bhattacharya et al. 2015). While significantly more fungal genomes have been sequenced (Grigoriev et al. 2014), only a small subset of these are involved in lichen symbioses (McDonald et al. 2013). As most functional gene annotations are inferred *in silico* through reference data (Kanehisa, Sato, and Morishima 2016; Finn et al. 2016; Conesa et al. 2005), the sparse availability of reference data is

reflected in the ability to annotate functions to the genes of the green alga *Trebouxia sp.* and the lichenized fungi (see section 4.4.1.3 on page 109).

The quality of genome reconstructions and their annotations also play a crucial role for their use in comparative studies. Incomplete and fragmented draft genomes have been shown to lead to misestimated gene family sizes, as genes are absent or only present in fragments which are spread across scaffolds (Denton et al. 2014). Furthermore, even the genomes of model organisms like *Saccharomyces cerevisiae* have been shown to be incompletely annotated (Dunne and Kelly 2017), negatively impacting the evolutionary inferences drawn from gene presences & absences. In section 4.4.2 (page 110) we have demonstrated that the available draft genomes of the mycobionts inside the Lecanoromycetes are affected by such errors, hindering especially the analysis of gene losses.

5.1.3 The state of evolutionary lichen genomics

The genomic and functional consequences of lichenization are so far poorly understood. To our knowledge, there were so far only limited comparative studies that tried to analyze the impact that entering lichen symbiosis has on the symbionts' genomes. The genome of mycobiont *Endocarpon pusillum* of the fungal class of Verrucariales was compared to 14 other, non-lichenized fungi (Wang et al. 2014). This comparison showed gene family contractions and expansions specific to the lineage of *E. pusillum*, revealing a marked gain of signal transduction proteins and nitrogen transporters as well as a substantial loss of sugar transporters. As no further mycobionts were included in the analysis, it is unclear whether these changes are directly related to the establishment of the lichen symbiosis or specific to the genome of *E. pusillum.*

The only comparative study of mycobionts focused on the loss of genes of the ammonium transporter/ammonia permease (AMTP) gene family (McDonald

et al. 2013). Here, it was observed that the 8 studied mycobionts retain members of the AMTPs that are lost in most other fungi.

The situation for the photobionts of lichens is similar. So far, the genomes of only a few photobionts have been studied in detail (Martínez-Alberola 2015). Furthermore, these studies were focused on the diversity of the photobionts that are associated with given mycobionts (P. Moya et al. 2017). Due to this, the effects of lichenization on the symbiotic partners' genomes are poorly understood when compared to the effects observed in mycorrhizal fungi.

To fill this gap, we utilize a comparative genomics framework, investigating the genomic consequences of lichenization and potential genomic reasons that explain why *Lasallia pustulata* could so far not be grown in axenic culture. We compare the genomes of *L. pustulata* and four further, culturable Lecanoromycetes to representatives of non-lichenized Eurotiomycetes and Dothideomycetes. This reveals that the Lecanoromycetes lost their capabilities to catabolize polysaccharides early in their evolution. Comparing *L. pustulata* to those Lecanoromycetes that are easily capable of growing in axenic cultures, we find no evidence for a strongly elevated rate of loss for otherwise well-conserved Lecanoromycetes genes. Similarly, we do not find support for a marked loss of functions among the gene family contractions and gene losses in *L. pustulata*.

5.2 Methods

5.2.1 Phylogeny reconstruction

We calculated a maximum likelihood tree, containing 48 fungi, 11 microsporidia, 12 algae, 7 other Viridiplantae and 9 further eukaryotes using a set of 80 well-conserved proteins that are ubiquitously found in the eukaryotes. The sequences were aligned with *CLUSTALW* v.2.1 (Thompson, Gibson, and Higgins 2002) and standard parameters. These alignments were concatenated and columns with more than 50% missing data were filtered out. The resulting alignment was used as input for *RAxML* (Stamatakis 2006) to infer the maximum likelihood tree, running 100 bootstrap replicates with the *PROTGAMMAILGF* model. Vinh Tran performed all steps of the phylogeny reconstruction.

5.2.2 The secretome of Lecanoromycetes

We predicted secreted proteins for the protein sets of nine Dothideomycetes, nine Eurotiomycetes (see Table 4-1 on page 87 for the taxa), and the Lecanoromycetes. We used *TargetP* v1.1 with standard parameters to predict the cellular localization of the individual genes. For those genes that were predicted to be secreted and not localized in the mitochondrion, we subsequently ran *TMHMM* v.2 (Emanuelsson et al. 2007) to further exclude those genes which were predicted to be transmembrane proteins.

5.2.3 Gene family expansions & contractions

We analyzed the gene family size evolution among the same 9 Eurotiomycetes and 9 Dothideomycetes further including *L. pustulata*, *C. grayi* and *X. parietina*. An all-versus-all search was performed on all proteome sets with *Diamond* (Buchfink, Xie, and Huson 2014), and the results subsequently clustered by the Markov Clustering implemented in *mcl* v12-135 (Enright, Van

Dongen, and Ouzounis 2002), using an inflation parameter of 1.4. We created a time-calibrated tree with *BEAST* v2.4.2 (Bouckaert et al. 2014), with the 80 genes that were used for the maximum likelihood tree and the divergence estimate for the included Lecanoromycetes (Amo de Paz et al. 2011) as an calibration point.

The calibrated tree and the clustering results were used as input for *CAFÉ* (Bie et al. 2006) to find significant gene family expansions and contractions on the branches. The results were analyzed and visualized using *ETE 3* (Huerta-Cepas, Serra, and Bork 2016). For a further analysis of the gene family expansions and contractions inside the Lecanoromycetes, we analyzed the expansions between *Umbilicaria muehlenbergii*, *Lasallia pustulata*, *Cladonia grayi*, *Usnea florida*, and *Xanthoria parietina*, using the dothideomycete *Zasmidium cellare* and the eurotiomycete *Eurotium rubrum* as an outgroup.

We functionally annotated gene families that showed significant expansions or contractions by assigning them to KEGG Orthologous Groups with *BlastKOALA* (Kanehisa, Sato, and Morishima 2016) and to Gene Ontology terms with *Blast2GO* (Conesa et al. 2005). We then tested for overrepresented functions amongst the significant gene family expansions and contractions using Fisher's Exact Test.

5.2.4 Gene gains and losses in the Lecanoromycetes

We predicted orthologs between Lecanoromycetes, Eurotiomycetes and Dothideomycetes with *OMA* v1.0.3 to investigate lineage specific gains and losses of genes. To rule out potential biases due to taxon sampling, we did this for three, non-overlapping sets of Eurotiomycetes and Dothideomycetes (the groupings are given in Table 4-1 on page 87) and added the five Lecanoromycetes *U. muehlenbergii*, *L. pustulata*, *C. grayi*, *U. florida*, and *X. parietina* to each of those three. The resulting Hierarchical Orthologous

Groups (HOGs) describe a hierarchy of orthologous groups in a given taxonomic range, based on pairwise orthology relationships. The HOGs were used to determine the sets of genes that descended from a common ancestor at each internal node in the species tree. These gains were subsequently searched for absent taxa, indicating lineage-specific losses of genes. We applied Dollo parsimony to minimize the number of individual losses, attributing losses only to higher clades if they are observed in all taxa of a given subtree. In this way we reconstructed ancestral gene sets for each node in the species tree. The scripts to calculate the gains and losses along the tree were kindly provided by Bardya Djahanschiri. A Fisher's Exact Test with False-Discovery Rate correction (Benjamini and Hochberg 1995) was performed to test for overrepresented functions based on Gene Ontology terms in the gains and losses in the last common ancestor of the Lecanoromycetes and in *Lasallia pustulata*.

5.2.5 Private losses of evolutionary conserved genes in the Lecanoromycetes

We followed the step-wise procedure described in 4.3.4 (see page 96) to yield a set of high-confidence candidates for the private loss of genes in the genomes of the Lecanoromycetes *U. muehlenbergii*, *L. pustulata*, *C. grayi*, *U. florida*, and *X. parietina.* We then additionally searched for the genes that were predicted to be privately lost in *L. pustulata* in its sister species, *Lasallia hispanica*. A *HaMStR* (Ebersberger, Strauss, and von Haeseler 2009) orthology prediction was performed with the gene predictions of *L. hispanica* and the Hidden Markov Models generated from the HOGs predicted to be absent in *L. pustulata*. We additionally did an *exonerate* (Slater and Birney 2005) search against the unannotated genome of *L. hispanica*, using the *protein2genome* model, analogous to the stepwise procedure in 4.2.5 (page 88).

To analyze their phyletic distribution, we used *HaMStR* to search for orthologs to the high-confidence gene losses that were private to *L. pustulata* (see Appendix for the list of taxa in which the search was performed, Table A-7 on page 208). Vinh Tran kindly provided the pipeline for the orthology search across these taxa. The presence and absence of those orthologs across these species was subsequently visualized using *PhyloProfile* (https://github.com/trvinh/phyloprofile, commit *a21dff5314*).

5.2.6 Searching for horizontally acquired genes

We performed a taxonomic assignment on the gene predictions for *L. pustulata* to find genes that were potentially a result of a horizontal gene transfer (HGT). We performed a *DIAMOND* (Buchfink, Xie, and Huson 2014) search against our custom database (see 2.2.5 on page 21 for the database composition). We subsequently used these search results for the taxonomic assignment with *MEGAN5* (Huson et al. 2011), applying a minimum score of 50 and no low complexity filter for the taxonomic assignment. We took all genes as candidates for HGT that were assigned to nodes in the taxonomy tree that are not on the path leading to the Fungi. To rule out assembly and assignment artifacts, we took only such HGT candidates into account that were flanked by fungal genes on both sides. The identification of these genes was done with *bedtools* v2.17.0 (http://bedtools.readthedocs.io/). The list of candidates was then further investigated by screening the results of the *DIAMOND* search.

5.3 Results

5.3.1 Phylogenetic placement of *L. pustulata*

We investigated the phylogenetic relationships of the mycobiont and photobiont of the lichen *Lasallia pustulata* by reconstructing a Maximum Likelihood tree of 87 species, covering large parts of eukaryotic diversity and focusing on the Fungi and Chlorophyta. The tree, based on 80 well-conserved proteins that are broadly found in all eukaryotes, was well resolved with most nodes having a bootstrap support of 100 (Figure 5-1). The Umbilicariaceae, represented by *Lasallia pustulata* and *Umbilicaria muehlenbergii*, were placed as the sister group to the rest of the Lecanoromycetes. The Lecanoromycetes themselves group with the Eurotiomycetes, excluding the earlier branching Dothideomycetes. Inside the Chlorophyta we find that *Trebouxia sp.* is placed as a sister taxon to *Asterochloris sp.*, inside a clade of four photobionts (Thüs et al. 2011). We noted that the Chlorophyta are only sparsely represented, despite using the genomes of all single celled algae that were publically available at the time of the analyses. Inside the clade of photobionts, we observed a substantial divergence between the different taxa, with *Trebouxia sp.* and its closest relative *Asterochloris sp.* being separated by a patristic distance of 0.7. The Lecanoromycetes on the other hand show markedly smaller branch lengths, with even longest branches inside the clade – between *Cladonia grayi* and *Xanthoria parietina* – having a patristic distance of 0.4. Furthermore, the Lecanoromycetes are embedded in a dense sampling of taxa, facilitating subsequent comparative analyses with a high resolution. For this reason we focused on the evolution of the Lecanoromycetes in general and *L. pustulata* in particular.

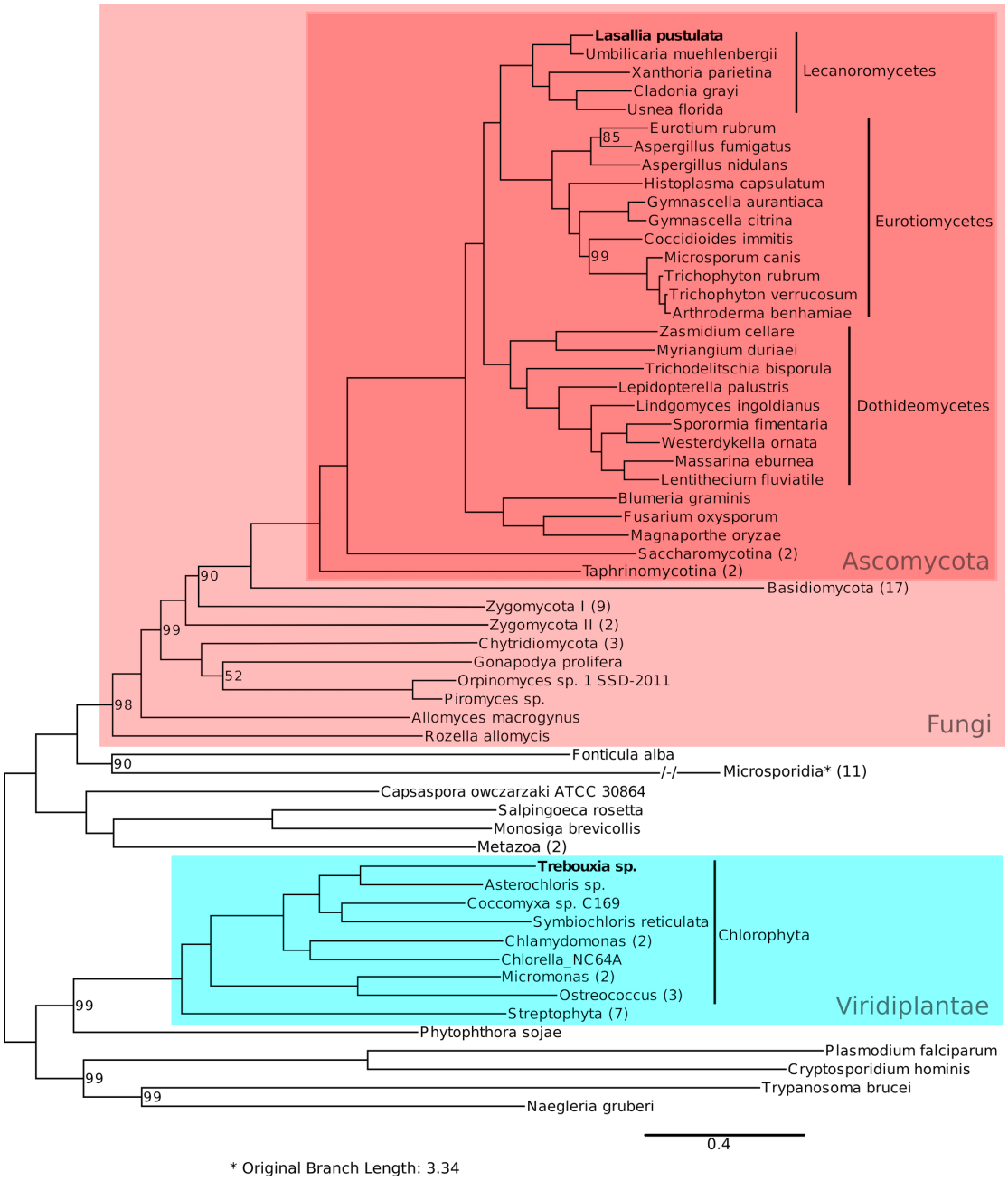


Figure 5-1: Maximum likelihood tree over 87 species. The Ascomycota (dark red), Fungi (light red) and Viridiplantae (blue) are highlighted. The lichen symbiont-containing groups, the Chlorophyta and the Lecanoromycetes, are delineated by the vertical bars, as are the closest relatives to the Lecanoromycetes, the Dothideomycetes and Eurotiomycetes. Node labels denote percent bootstrap support, and only bootstrap values <100 are shown. The eukaryotic symbionts of the lichen *L. pustulata* are highlighted in bold face. The full tree without collapsed taxonomic groups is depicted in Figure A-11 on page 220 in the Appendix.

5.3.2 The secretome of *L. pustulata*

Fungi use an array of secreted proteins to mine their environment for resources like carbon, phosphate and nitrogen (Bouws, Wattenberg, and Zorn 2008) and differences in lifestyles and habitats have been linked to the overall number of secreted proteins (Pellegrin et al. 2015). We found 721 secreted proteins in the genome of *Lasallia pustulata*. We subsequently compared the secretome size of *L. pustulata* with that of four other Lecanoromycetes, 9 Eurotiomycetes and 9 Dothideomycetes. We observed that both *L. pustulata* and *U. muehlenbergii* have notably smaller secretomes (Figure 5-2B), compared to the other Lecanoromycetes, which are more similar to the secretome sizes of the largely parasitic Eurotiomycetes (Figure 5-2A). For the other Lecanoromycetes we saw secretome sizes that are more comparable to those of the mainly saprophytic Dothideomycetes. This divide suggests that different habitats may have a stronger influence on secretome sizes of the Lecanoromycetes than differences in lifestyle.

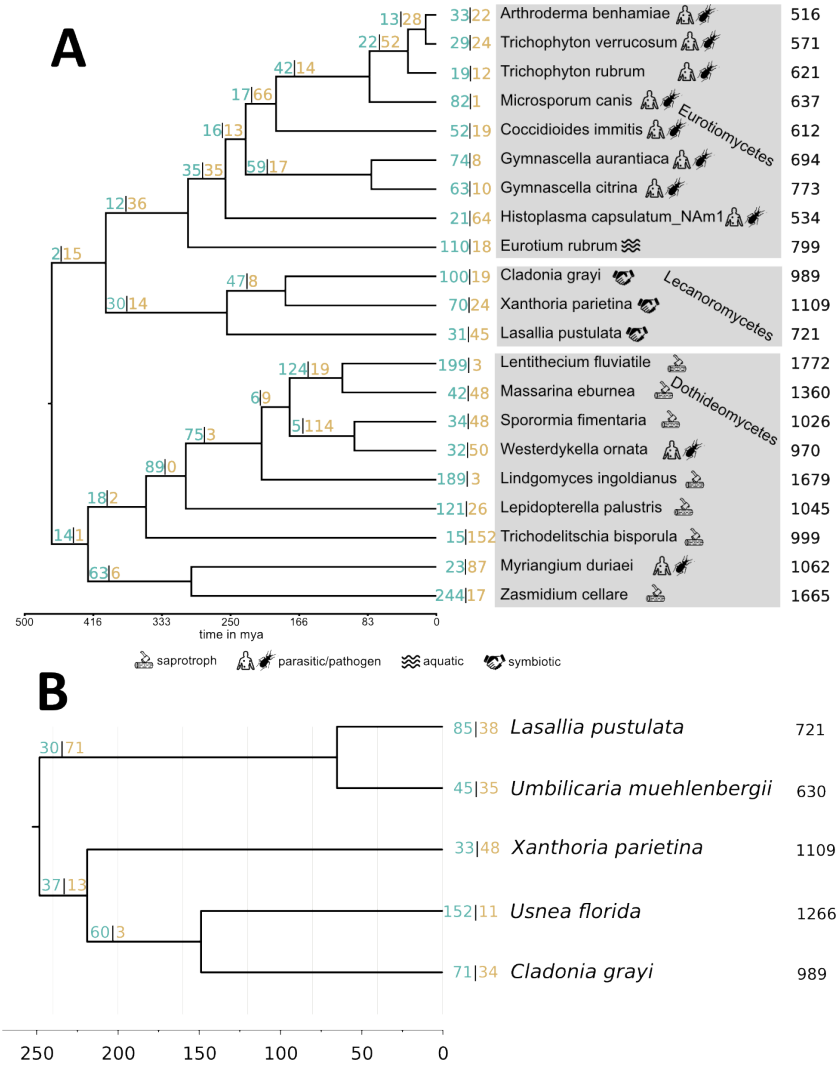


Figure 5-2: The large-scale gene set evolution amongst the Lecanoromycetes, Eurotiomycetes, and Dothideomycetes (A) and focusing on the Lecanoromycetes (B). The pictograms in (A) denote the nutritional lifestyles of the individual taxa. The branch lengths were time-calibrated with a published divergence estimate for the included Lecanoromycetes (Amo de Paz et al. 2011)**, the axes give the divergence time in million years ago (mya). The secretome sizes are given in black behind the taxon names. Gene family expansions (blue) and contractions (yellow) are given on the branches.**

5.3.3 The evolution of gene families in *Lasallia pustulata*

We investigated how living in longstanding symbiotic communities can shape genomes by analyzing the gene set evolution in the Lecanoromycetes. We did this on the level of gene families and individual genes. As a first layer of evidence for how lichenization leads to a genomic dependence on the symbiotic partners, we inferred expansions and contractions of gene families in the individual Lecanoromycetes relative to the closest non-lichenized

relatives. A clustering of the gene sets of 21 Eurotiomycetes, Dothideomycetes, and Lecanoromycetes with the Markov Clustering Algorithm (Enright, Van Dongen, and Ouzounis 2002) revealed 6,054 gene families. We then identified significant expansions/contractions of those families on individual lineages using *CAFÉ* (Bie et al. 2006). We observed that only a surprisingly small number of gene families have contracted on the lineage leading to the Lecanoromycetes, when compared to their sister clade, the largely parasitic Eurotiomycetes (Figure 5-2A). Instead those numbers resemble the contractions found in the saprophytic Dothideomycetes. This is in contrast to the gene family expansions, where Lecanoromycetes show numbers similar to those of the Eurotiomycetes and not the Dothideomycetes. However, we also found that most contractions and expansions did not happen on ancestral branches, but rather towards the terminal branches. For the Lecanoromycetes we observed the smallest number of expansions and the biggest number of contractions in *Lasallia pustulata*, when comparing it to *Cladonia grayi* and *Xanthoria parietina* (Figure 5-2A). We subsequently focused on the Lecanoromycetes, extending the taxon selection. This revealed that the majority of contractions are not private to *L. pustulata*, but rather happened on the branch leading to the Umbilicariaceae (Figure 5-2B).

We subsequently investigated whether the contractions and expansions in *L. pustulata* show evidence for significant alternations in the molecular functions that result from these changes in the underlying gene set. Using the Gene Ontology annotation, we did not find a significant functional enrichment. The functional classification of some of the gene families points into an interesting direction though. Amongst the significantly contracted gene families in *L. pustulata* we observed a major facilitator superfamily (MFS) transporter of the ACS family (KO identifier: K08192). Members of the ACS family of MFS

transporters are transmembrane proteins, which facilitate small solute transport of amino acids, sugars and further metabolites (Pao, Paulsen, and Saier 1998). Additionally the family of putative multidrug resistance ABC transporters (KO identifier: K05658), which facilitate the movement of xenobiotics (Barabote et al. 2011), was contracted as well. Amongst the expansion of gene families in *L. pustulata* we found the GTPase-activating protein family SAC7 (KO-identifier K19845), which was found to be necessary for a normal growth at low temperatures in yeast (T. M. Dunn and Shortle 1990). Furthermore an overexpression of SAC7 can inhibit vegetative growth in yeast (A. Schmidt et al. 1997).

5.3.4 The evolution of Lecanoromycetes gene sets

To further increase the resolution of our investigation into the genomic dependency on symbiotic partners, we analyzed the fate of individual genes inside the Lecanoromycetes. Utilizing *OMA* (Altenhoff et al. 2015), we identified 11,789 hierarchical orthologous groups (HOGs) from the gene sets of 5 Lecanoromycetes, 3 Dothideomycetes and 3 Eurotiomycetes. We subsequently inferred the gain and loss of individual genes along the phylogeny from those HOGs, using parsimony. In line with our observation for the gene family size evolution, we found that only a comparatively small number of genes are gained early in the evolution of the Lecanoromycetes (Figure 5-3). Instead, most genes are newly on the terminal branches. This effect becomes even more pronounced when taking the estimated ages of the splits into account (compare with Figure 5-2). For example, the LCA of *L. pustulata* and *U. muehlenbergii* acquired only 504 genes in the time between its split from the other Lecanoromycetes 250 mya and its subsequent diversification 70 mya. On the other hand, the lineage leading to *L. pustulata*

acquired an additional 2,314 genes since its separation from the *U. muehlenbergii* lineage at 70 mya.

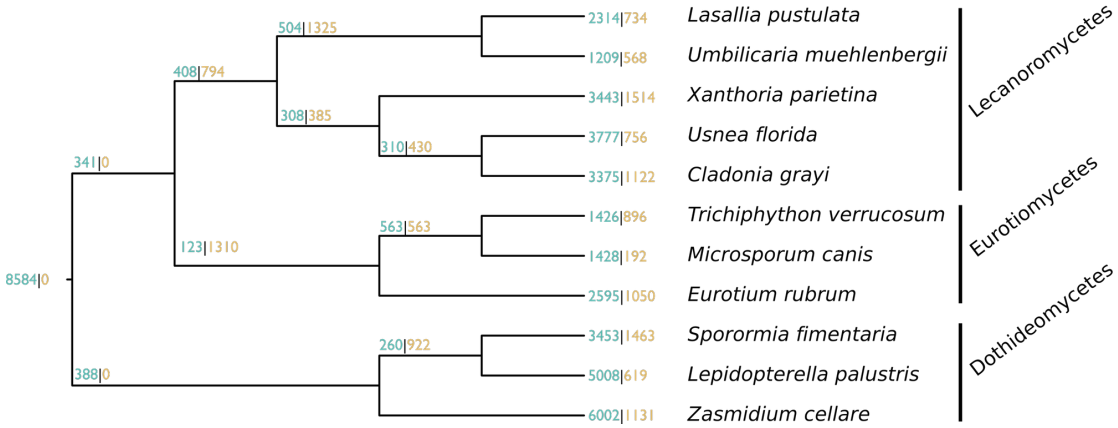


Figure 5-3: Cladogram for the Lecanoromycetes, Eurotiomycetes and Dothideomycetes included in the analysis of the HOGs. Gains (blue) and losses (yellow) for the individual branches were inferred from the HOGs using Dollo parsimony.

The picture is different for the loss of genes. We identified 794 genes that were lost after the last common ancestor (LCA) of the Lecanoromycetes split from the Eurotiomycetes, but prior to the diversification of the Lecanoromycetes. Alternating the taxon sampling, with two different, non-overlapping taxon sets for the Eurotiomycetes and Dothideomycetes had no substantial influence on the number of predicted gene gains and losses inside the Lecanoromycetes (see Appendix, Figure A-12 on page 219 for the results with the different taxon sets).

We subsequently investigated the observed gene sets changes during the early stage of lichenization, using GO terms. We first compared the gene losses for the LCA of the Lecanoromycetes to its reconstructed ancestral gene set. This enrichment analysis yielded 36 GO terms for biological processes (see Appendix, Table A-8, page 211), which are significantly overrepresented amongst the 794 genes that are lost on the lineage leading to the LCA of the Lecanoromycetes. Amongst the functions overrepresented in the lost genes, we found broad terms for polysaccharide catabolic processes as well as

specific ones for metabolic processes of xylan and other hemicelluloses. Furthermore, transmembrane transporters were enriched in the losses. Comparing the genes gained in the LCA of the Lecanoromycetes, we did not find evidence for a functional enrichment.

We then focused on the genes acquired and lost in *L. pustulata* to evaluate whether the functional consequences of the genomic remodeling can explain its poor culturability. Our enrichment analysis showed no evidence for functional enrichments amongst the genes acquired or lost in *L. pustulata*.

5.3.5 Loss of evolutionary conserved genes

We hypothesized that the long-established lichen symbiosis allows the loss of otherwise essential genes in individual Lecanoromycetes, potentially explaining why *L. pustulata* could so far not be grown in culture. We thus focused our analysis on the well-conserved genes that we traced back to the last common ancestor of the Lecanoromycetes (LCA_{Lec}) and which were privately lost in only a single taxon. Our sensitive search for those genes – making use of *OMA*, *HaMStR* (Ebersberger, Strauss, and von Haeseler 2009), exonerate (Slater and Birney 2005), RNAseq data and manual curation (described in detail in in 4.3.4 on page 96), yielded 28 LCA_{Lec} genes that were lost in *Lasallia pustulata*, 45 for *Cladonia grayi*, 52 for *Usnea florida*, and 90 for *Xanthoria parietina*. A reliable estimate of lost LCA_{Lec} genes was not possible for *Umbilicaria muehlenbergii*, due to the absence of RNAseq data, leading to a potential underestimation of retained LCA_{Lec} genes. We subsequently used the branching times of the individual taxa, to calculate the rate of gene loss amongst the Lecanoromycetes (see Table A-9 on page 212). While *L. pustulata* showed the highest LCA_{Lec} loss rate, the number of lost genes is small for all taxa and more data would be required to decide whether this indicates an accelerated loss.

We further narrowed down when the genes in *L. pustulata* have been lost, by searching for the absent 28 LCA_{Lec} genes in its sister species *Lasallia hispanica*. As the genome of *L. hispanica* was only recently sequenced, and is still in an early draft stage that is most likely less complete than the other genomes, it was not included from the start. Performing both a *HaMStR* (Ebersberger, Strauss, and von Haeseler 2009) orthology search in the annotated gene set, as well as an *exonerate* (Slater and Birney 2005) search in the unannotated genome, we identified orthologs for 4 of the 28 LCA_{Lec} genes. One of the privately lost LCA_{Lec} genes has GO terms assigned. According to the GO terms, this gene is a non-specific serine/threonine protein kinase.

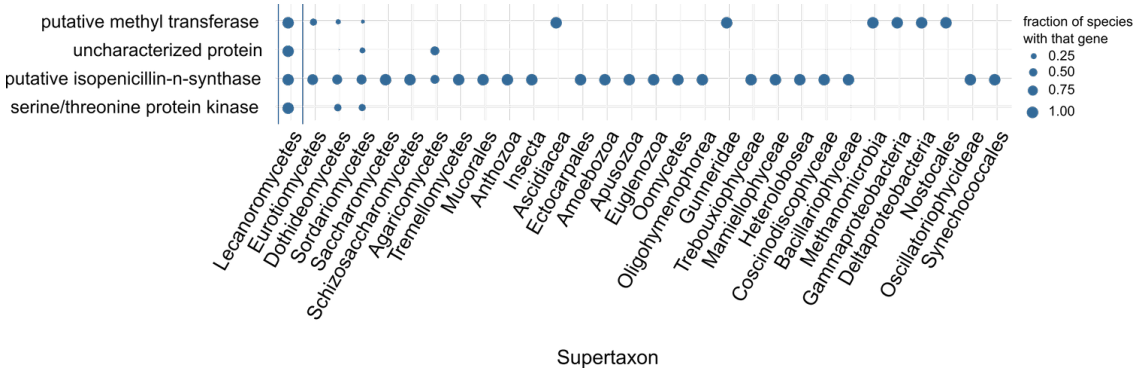


Figure 5-4: The phylogenetic profile of the four LCA_{Lec} genes privately lost in *L. pustulata* that were present in the draft genome of *L. hispanica*. Taxonomic groups, supertaxons, on the x-axis are sorted by increasing taxonomic distance to the Lecanoromycetes. On the y-axis the genes identified to be privately lost and their tentative annotation are given. The sizes of the circles give the fraction of the taxa in which an ortholog to the respective LCA_{Lec} gene was found.

For the three other absent LCA_{Lec} genes, which are annotated by neither KEGG nor GO, we searched for Pfam (Finn et al. 2010) domains to approximate their function. Additionally, we determined the presence of these four genes across the tree of life and visualized the results in a phylogenetic profile (Figure 5-4). This revealed a putative isopenicillin-n-synthase, which is present in a variety of fungal classes, in addition to a putative Methyltransferase that is only sparsely found outside the

Lecanoromycetes. Furthermore, we identified a so far uncharacterized gene that does not include known domains and was only rarely found outside the Lecanoromycetes. A similar picture emerged for the 24 absent LCA_{Lec} genes for which no ortholog in *L. hispanica* was identified (see Appendix Table A-10 on page 212 for the full functional annotation of these and Figure A-13 on page 222 for their phylogenetic profile). Only two of these could be annotated through KEGG; a carbonyl reductase that is found only sparsely in other fungal classes but found frequently in more distantly related taxa, as well as an ubiquitin-conjugating enzyme that was not found outside the Lecanoromycetes. In all, only four of these LCA_{Lec} genes are frequently observed in other taxonomic groups.

5.3.6 No evidence for horizontal gene transfer in *Lasallia pustulata*

We hypothesized that the long-standing symbiosis of *L. pustulata* might provide ideal conditions for a mutual exchange of genetic material between the symbionts. We thus searched for evidence of horizontally acquired genes of algal or bacterial origin in the fungal genome. We found 10 genes that *MEGAN* (Huson et al. 2011) identified to be of algal origin, in addition to 12 genes of bacterial origin. Genes of predicted fungal origin flank all of these candidate genes. The subsequent manual curation of the 22 candidates showed that the taxonomic assignment for each of them was not well supported: Comparing the best bit scores of our HGT candidates with the vertically inherited ones (see supplementary materials provided with this thesis for the *DIAMOND* results), we found that the candidates have a much lower sequence similarity (mean, length-normalized bit score 0.34) than the vertically inherited ones (mean, length-normalized bit score 1.04). Furthermore, the sequence similarities of the best fungal hits were only barely below the inclusion threshold for our candidates. This lack of support

suggests that these candidate genes are false taxonomic assignments rather than evidence for a horizontal gene transfer.

5.4 Discussion

Adaptive evolution is highly driven by a reshaping of the genetic repertoire (Albalat and Cañestro 2016). Rapid changes in gene families sizes (Baroncelli et al. 2016; Gan et al. 2016) as well as wide-spread gene losses (Wolf and Koonin 2013; Williams and Wernegreen 2015) playing key roles in this remodeling. These effects have also been observed during the establishment of symbioses (Bennett et al. 2014), leading to an interdependence between the involved organisms (F. Martin, Uroz, and Barker 2017). This seems to also be the case in lichens. Despite their global success in extreme habitats (Ahmadjian 1993; Ilse Kranner et al. 2009), some mycobionts have been found to hardly grow solitarily in culture (McDonald, Gaya, and Lutzoni 2013). This points to a strong reliance of the mycobionts on their symbiotic partners, to an extent that affects solitary survival. Given this, we expected a stronger genomic reshaping in the so far uncultured *Lasallia pustulata*, compared to other, culturable Lecanoromycetes and non-lichenized fungi. We thus investigated the genomic remodeling in a comparative framework, searching for changes in gene family sizes and losses of old, well-conserved genes.

5.4.1 The genomic footprint of lichenization

Comparing the evolution of gene families between the Lecanoromycetes, Dothideomycetes and Eurotiomycetes we found that different gene families show significant expansions or contractions on the different lineages. This observation fits that of earlier studies, which have shown that changes in gene family sizes are often connected to the adaptation to ecological niches or specific hosts (Baroncelli et al. 2016; Gan et al. 2016; Gazis et al. 2016; Morales-Cruz et al. 2015; Sharpton et al. 2009; P. D. Spanu et al. 2010). However, it appears that the nutrient exchange between the Lecanoromycetes and their

photobionts did not lead to substantial changes of particular gene families. On the level of individual gene losses, as found through the HOGs, we observed a reductive genome evolution for the Lecanoromycetes as well as the Eurotiomycetes. For both, we observed that their respective last common ancestors lose around 10% of the genes that were present in their shared ancestor. Genome reductions have been found to happen subsequently to fungi becoming biotrophes (P. Spanu 2012; McDowell 2011). As these losses are non-overlapping these can be indicative of the different modes of biotrophy that the two lineages started adapting to.

Amongst the Lecanoromycetes we noted the largest number of contracted gene families in the ancestor of *Umbilicaria muehlenbergii* and *Lasallia pustulata*. Furthermore, both species show low numbers of contracted gene families on their respective terminal branch. Thus, the number of contracted gene families does not explain the poor culturability of *L. pustulata*, as *U. muehlenbergii* was successfully grown in axenic cultures (S. Y. Park et al. 2014). We observed a similar picture, when focusing on the losses of individual genes as predicted through the HOGs. We found no signs of an increased gene loss in *L. pustulata;* instead more gene losses were shared between *U. muehlenbergii* and *L. pustulata*. As such, a widespread loss of genes cannot explain why *L. pustulata* could so far not been grown in axenic culture.

5.4.2 Loss of evolutionary conserved genes

Instead of being the outcome of a widespread loss of genes, the poor culturability of *Lasallia pustulata* could be a result of losing central, otherwise well-conserved genes. For this reason, we investigated the loss of genes that we could trace back to the last common ancestor of the Lecanoromycetes and that are conserved in all but one taxon. We found no substantial difference in the LCA_{Lec} loss rate between the Lecanoromycetes. Furthermore, we could

only verify the private loss of 4 of such genes in *L. pustulata*, when comparing it to *Lasallia hispanica*. Thus, the observed loss of conserved genes cannot explain *L. pustulata*'s lack of culturability either. This should be considered a conservative estimate for two reasons though. Foremost, the genome of *Lasallia hispanica* was in an early draft stage, with no RNAseq data being available. Thus, genes potentially present in *L. hispanica* might have been overlooked. Additionally, incomplete gene annotations for the other four Lecanoromycetes might lead us to underestimate the number of exclusively lost genes in *L. pustulata*, as genes will appear to be lost more than one time. Nevertheless, the observed low rate of lost LCA_{Lec} genes might be due to the nature of these old and conserved genes. On the sequence level it has been seen that multifunctional and pleiotropic genes, which have effects on multiple phenotypic traits, evolve slower due to their effects on multiple pathways (Salathe, Ackermann, and Bonhoeffer 2006; He and Zhang 2006; Dudley et al. 2005). Similar effects have been observed for genes that are highly co-expressed with many interaction partners (Jordan et al. 2004). Analogous to these observations, one can hypothesize that such central genes cannot even be lost in organisms that appear to be more strongly dependent on their symbiont, as the loss would affect a multitude of metabolic pathways, including those that the symbionts cannot supplement.

5.4.3 Functional consequences of genomic remodeling following lichenization

In addition to a decrease of gene content, the reshaping of genomes after the establishment of symbioses can also lead to a reduction in the functional repertoire, when functionalities shared between symbiotic partners render individual genes redundant (Williams and Wernegreen 2015). Such losses in functional capacity can occur either spread out over various genetic pathways

or be concentrated in a small set of related functionalities (Kohler et al. 2015). The analysis of the functional changes in the Lecanoromycetes allows for potential insights into the nature of the lichen symbiosis.

Our analysis of the secretomes of the Lecanoromycetes revealed large differences between individual species. The sizes of secretomes in fungi have previously been linked to different habitats and lifestyles (Pellegrin et al. 2015; Lowe and Howlett 2012). As expected, we observe that the secretomes of the rock-dwelling *Umbilicaria muehlenbergii* and *Lasallia pustulata* are smaller than those of *Xanthoria parietina*, *Cladonia grayi* and *Usnea florida*, all of which prefer organic substrates like living or dead trees. In line with this, the draft genomes of *Caloplaca flavorubescens* and *Cladonia macilenta*, both living on tree bark, exhibit even bigger secretome sizes, with 1,800 and 1,300 secreted proteins respectively (S. Y. Park, Choi, Kim, Yu, et al. 2013; S. Y. Park, Choi, Kim, Jeong, et al. 2013). This indicates that the reduced secretomes of *L. pustulata* and *U. muehlenbergii* might be an adaptation to their nutrient poor substrate, which potentially lacks additional interaction partners. Such adaptations have already been described for plant-pathogens and mycorrhizal fungi (McCotter, Horianopoulos, and Kronstad 2016).

Using the HOGs to analyze the genes that were lost in the last common ancestor of the Lecanoromycetes, we found evidence for a functional enrichment of these losses. Interestingly, we find that genes that have functions relating to the general catabolism of polysaccharides, and for xylan in particular, are preferentially lost. This is especially striking, as they are otherwise highly abundant in most fungi, including the closest relatives of the Lecanoromycetes, the Dothideomycetes and Eurotiomycetes (Berlemont 2017). The excessive loss of genes involved in the polysaccharide catabolism can be related to the adaptation from a saprotrophic lifestyle, which relies

heavily on polysaccharide degradation, to the less complex nutrient sources presented to the Lecanoromycetes by their photobionts. Similar losses of genes involved in the polysaccharide catabolism have earlier been observed for the mycorrhizal fungi that also evolved from being saprophytic to biotrophic (F. Martin et al. 2008; Kohler et al. 2015). The functional enrichment of the gene losses is contrasted by a lack thereof for the genes that were gained in the LCA of the Lecanoromycetes. However, instead of real absence of enrichment, this is potentially a consequence of the lack of annotations that are available for the genes gained in the Lecanoromycetes LCA. As a result of the sparsely available reference data, only 91 out of the 408 genes gained could be functionally annotated with GO terms.

We evaluated the functional effects of the genomic remodeling in *Lasallia pustulata*, searching for explanations for its poor culturability and stronger dependence on its photobiont. On the level of expanded and contracted gene families we did not find evidence for a systematic functional remodeling based on the functional annotation with Gene Ontology and KEGG pathways. Despite this, we observed that gene families for the major facilitator superfamilies (MFS) and ABC transporters are contracted on the lineage of *L. pustulata*, and to a lesser extent in the LCA of the Lecanoromycetes. Both gene families have been shown to have a central role in fungi that are plant-pathogens and furthermore for the efflux of toxic compounds (Coleman and Mylonakis 2009). This potentially relates to the move towards a commensal/mutualistic lifestyle that happened during the lichenization, rendering these functions less important. Analyzing the genes that the HOGs predicted to be gained or lost in the lineage of *L. pustulata*, we did not find a functional enrichment.

Additionally, the functional characterization of the LCA_{Lec} genes that were lost privately in *L. pustulata* did not show substantial decreases of the functional potential either. Our conservative estimate, including only genes absent in *L. pustulata* that were confirmed to be present in *L. hispanica*, revealed only four LCA_{Lec} genes to be absent, all of which were poorly annotated. The putative isopenicillin-n-synthase, which was found to have orthologs across the tree of life, appears to be the most interesting candidate on first look. Its function cannot be conclusively annotated though. All mononuclear non-heme Fe(II)- and 2-oxoglutarate (2OG)-dependent oxygenases share the Pfam domains found in the isopenicillin-n-synthase. Furthermore, Fe(II)-2OG-dependent oxygenases have been shown to be involved in numerous biological functions, ranging from fatty acid metabolism over DNA/RNA repair to biosynthesis of secondary metabolites and antibiotics (Martinez and Hausinger 2015). Similarly, serine/threonine protein kinases (Dickman and Yarden 1999) as well as methyl transferases (Katz, Dlakić, and Clarke 2003) are so numerous, that the identification of the corresponding Pfam domains provides only limited insight into the biological processes in which they are involved. The lack of a detailed annotation analogously hinders a detailed evaluation of the functional consequences of the additional 24 lost LCA_{Lec} genes in *L. pustulata* that we could not confirm in the genome of *L. hispanica*. Additional reference data will be needed to fully evaluate the impact of these gene losses.

The absence of a marked decrease in functional potential in *L. pustulata* fits recent observations of an endosymbiotic organism that also did not display a marked decrease in functional capability (Hehenberger et al. 2016). Further support for the absence of decreasing functional capacity comes from the large overlap in functional annotations we observed between the 5

Lecanoromycetes (see 4.3.3 on page 94, and Figure A-5, Figure A-6 on page 217), as at least amongst the functionally annotated genes it appears that the Lecanoromycetes have a rather consistent and conserved functional capacity.

5.4.4 No evidence for recent horizontal gene transfer into the Lecanoromycetes

Horizontal gene transfer (HGT) has been described as a source for genomic innovations, including HGTs between eukaryotes and bacteria or even between two eukaryotes. For fungi, including the Lecanoromycetes, there is some evidence that supports evolutionary old HGT events in which genes were acquired from prokaryotes (Schmitt and Lumbsch 2009; McDonald et al. 2013; McDonald, Dietrich, and Lutzoni 2012; Lawrence et al. 2011) or even from plants (Richards et al. 2009). Reliably inferring HGT is not simple, with the proposed HGT into a tardigrade genome being a prominent example (Boothby et al. 2015). A subsequent analysis showed that the initial finding – of 17% of all tardigrade genes being acquired through HGT – was a methodological artifact (Delmont and Eren 2016; Koutsovoulos et al. 2016).

Searching for recent horizontal gene transfers from plants or bacteria into the genome of *Lasallia pustulata*, we focused our analysis on HGT events that happened after the split of the Lecanoromycetes. Our assignments rely on sequence similarities to a database of sequences with known taxonomic classification. *MEGAN* (Huson et al. 2011) then uses all top-ranking hits found during that search, if they pass a liberal bit score threshold and have a score that is within a given range to the best hit found for a given sequence. The closer investigation of the 22 candidates we identified revealed that the taxonomic assignment for all of them was spurious. Already minor changes to the parameters used for the taxonomic classification in *MEGAN* lead to the weak HGT signal disappear. Based on these findings, we conclude that these

assignments are methodological artifacts and that a more parsimonious phenomenon of vertical transfer is likely to be correct.

5.5 Conclusion

Searching for the genomic footprint of lichenization we find that the early evolution of the Lecanoromycetes reveals a loss of around 10% of the genes present in the last common ancestor of the Lecanoromycetes and Eurotiomycetes. Both groups exhibit no substantial amount of shared gene losses of genes. This hints that both groups adapted in different ways to their biotrophic lifestyles. For the Lecanoromycetes we observed that these gene losses were functionally enriched, including a marked number of genes involved with the degradation of polysaccharides. These losses are most likely a consequence of an adaptation to a symbiotic, lichenized lifestyle that does not rely on the degradation of plant material as nutrition source.

Our analogous search for a genomic remodeling in *Lasallia pustulata*, which could explain its poor culturability, did not show a strong signal for losses in functional capabilities. Ignoring the draft genome of *Lasallia hispanica*, we found only 28 genes that have been privately lost in *L. pustulata*, with the majority of them being poorly functionally annotated. This is also reflected in the large set of genes that were privately gained in *L. pustulata* and could also not be characterized functionally.

6 Discussion & Outlook

Large parts of the fungal diversity are found in lichens, with an estimated 21% of fungal species living in lichen symbioses (Hawksworth 1988). Their global distribution, covering around 6% of the Earth's surface (Gadd 2010), makes them a key part of most terrestrial ecosystems, including habitats that are not open to many other macroscopic organisms (Muller 1952; Kidron and Temina 2010). The dependence of lichenized fungi on their symbionts varies, from facultative to obligate mycobionts (Lewis 1973), which is reflected in their varying abilities to grow in axenic cultures (McDonald, Gaya, and Lutzoni 2013). Lichens thus make an interesting group to study evolutionary adaptations between symbiotic partners (Grube and Spribille 2012). So far, most of these studies have focused either on the bacterial microbiome of the lichen symbiosis (Grube and Berg 2009; Grube et al. 2015) or on culturable mycobionts (Wang et al. 2014; S. Y. Park, Choi, Kim, Jeong, et al. 2013; S. Y. Park, Choi, Kim, Yu, et al. 2013). Our goal was to extend the field of genomic studies to include lichens that show a potently stronger symbiotic dependency by using *Lasallia pustulata* as a model. Given that it was so far not possible to grow the mycobiont *L. pustulata* in axenic cultures, we suspected that the long-standing symbiosis should have left substantial footprints on its genome and potentially the lichen hologenome in general.

6.1 The feasibility of metagenomic hologenome reconstructions

The genome reconstruction of symbionts that are not culturable has to rely on metagenomic data. For this reason we evaluated to what extent it is possible to assemble the genomes of eukaryotic lichen symbionts from a simple metagenome skimming data set. Using a simulation-based approach we evaluated the performance of different assembly strategies and algorithms in Chapter 2. We observed that skewed abundances for the organisms – leading

to a low coverage for the underrepresented genome (Desai et al. 2013) – negatively impact all assemblers. Only two of the tested methods, the overlap-based assembler *MIRA* (Chevreux, Wetter, and Suhai 1999) and the multi-*k*-mer assembler *SPAdes* (Bankevich et al. 2012), were less affected by this. We furthermore found that standard procedures for finding an optimal *k*-mer size (Namiki et al. 2012; Chikhi and Medvedev 2014) are detrimental when applied to highly coverage-ratio skewed metagenomic data, as it will effectively lead to an exclusion of the lower represented genome from the assembly to optimize the overall contiguity. While some assemblers, like *metaSPAdes* (Nurk et al. 2016) and *MEGAHIT* (D. Li et al. 2015) try to solve the parameter choice internally, we find that effective automated ways for finding ideal assembly parameters are still lacking (cf. Awad, Irber, and Brown 2017). As metagenome assemblies are becoming increasingly important, it will be necessary to further analyze how to optimize the assembly parameters in an unbiased way.

Applying these different methods on the real genome skimming data of *Lasallia pustulata*, we observed a pronounced skew towards the fungal genome along with a sizeable fraction of the data belonging to the bacterial microbiome. Both factors led to genome assemblies that remained more fragmented and incomplete than expected given our simulations. We thus concluded that a single metagenome skimming experiment is capable of recovering substantial amounts of the mycobiont genome and allows for a first hologenome characterization, but it is insufficient to facilitate the contiguous assembly of the photobiont.

6.2 Assembling & characterizing the *Lasallia pustulata* hologenome

Incomplete and fragmented genomes hinder high-resolution comparative phylogenomic analyses. This becomes even more important when the goal is to identify changes in the genetic and functional repertoire encoded in the

analyzed genomes rather than identifying conservation. Long-read sequencing can improve the assembly contiguity (Chakraborty et al. 2016) and aids in the reconstruction of complex genomes (Loman, Quick, and Simpson 2015; Jain et al. 2017; Frank et al. 2016; Tuskan et al. 2006). For this reason, we complemented our short-reads with additional long-read sequencing data. This data was then used to assemble and subsequently characterize the *L. pustulata* hologenome as described in Chapter 3.

We devised a custom-tailored hybrid assembly strategy, combining different data sources and assembly methods that were subsequently merged. As expected given prior studies on non-metagenomic data (Wences and Schatz 2015; Chakraborty et al. 2016), this substantially increased the completeness and contiguity of the *Lasallia pustulata* hologenome, including the photobiont and bacterial microbiome. Given these results it seems promising to further evaluate and benchmark the use of hybrid methods for the assembly of complex metagenomes. While there are first tries to further automate hybrid assemblies (Ye et al. 2016; Kajitani et al. 2014), these methods are currently limited by the impact that different long- and short-read coverages have on the resulting assembly quality, making it non-trivial to find the best combination of assembler and read-coverages (Chakraborty et al. 2016). In addition to these limitations, these methods have so far only been sparsely evaluated on metagenomic data (Frank et al. 2016), where differential coverages will potentially have an even bigger impact on the assembly quality. Further, more systematic, evaluations will be needed to estimate whether merging metagenomic hybrid-assemblies can help to overcome the limitations we observed for short-read based metagenomic assemblies.

We found evidence that the hybrid assembly approach does not guarantee error-free genome assemblies. We uncovered that the sequence composition itself can lead to pathogenic sequencing errors that subsequently interfere with the correct prediction of genes. G/C-rich inverted repeats largely

prevented the sequencing of these regions with *Illumina*. Genes located in these regions are thus potentially absent from the short-read sequencing data itself and consequently lead to an artificially increased prediction of gene losses. While our long-read sequencing could overcome such pathogenic sequence regions, they are more prone to contain insertion/deletion-sequencing errors, as short-read based error corrections cannot be performed. To our knowledge these effects have being scarcely reported so far (c.f. Botero-Castro et al. 2017), thus a more focused study of these effects will be necessary to evaluate their impact on genome assembly and evolutionary conclusions drawn from them.

Analyzing the hologenomic composition we observed that *L. pustulata* appears to be supported by a stable bacterial microbiome, which remains largely unchanged across sequencing libraries generated from samples collected in different locations in Germany and Italy. The *L. pustulata* microbiome is dominated by members of the Acidobacteriaceae, which have been found to be highly prevalent in the microbiomes of some other lichens (Hodkinson et al. 2012). Acidobacteriaceae survive marked changes in hydration (Ward et al. 2009) and can live in oligotrophic conditions (Castro et al. 2010) – slowing their metabolic rates when being nutrient-deprived. These characteristics make them well adapted for co-habitating with *L. pustulata*, which faces similar conditions as it grows on heavily sun-exposed, vertically inclined rocks and cliffs (Hestmark et al. 1997). At this point we cannot differentiate on whether the consistent presence of the Acidobacteriaceae hints at a functional involvement in the lichen symbiosis or whether their co-localization is due to a shared habitat preference. Further analyses into the localization of the Acidobacteriaceae in the lichen thalli, as well as the functional capabilities of the Acidobacteriaceae, are needed to clarify this. Our assembly of the bacterial microbiome resulted in two nearly completely

reconstructed Acidobacteriaceae genomes. These genomes can offer excellent starting points for such further analyses.

Additionally, we performed an initial investigation into how estimated taxonomic abundances in a microbiome can be biased by the strategy used to calculate those abundances. We found that contig-based methods tend to over- or underestimate taxa, depending on their abundance on the read level. It correspondingly appears ideal to perform the taxonomic assignment directly on the read level whenever possible. This strategy becomes problematic though if either the sequence read lengths are too short or the gene density in the sequenced bacteria is too low. Only reads that are (partially) encoding a protein can be taxonomically assigned, as the taxonomic assignment is routinely performed by searching with the translated DNA sequencing reads against a protein database. Given these trade-offs, a further, systematic evaluation should be performed to evaluate these effects, guiding the choice of how to analyze further microbiomes in an unbiased way.

6.3 Annotating the *Lasallia pustulata* hologenome

Subsequent evolutionary analyses into the genomic consequences of lichenization and the interactions between the symbiotic partners are heavily dependent on complete and accurate genome annotations (Denton et al. 2014). We thus not only annotated genes for the individual genomes found in *L. pustulata,* but also extensively evaluated potential annotation errors in Chapter 4. We compared the gene prediction fidelity in the mycobiont *L. pustulata* and four additional – already annotated – Lecanoromycetes genomes. We used an evolutionarily interesting subset of 1,402 genes that we could identify to have been present in the last common ancestor of these Lecanoromycetes (LCA_{Lec}) and that appear missing in only one of the genomes. These genes are supposedly central, as they are widely conserved.

Thus, losses of them are indicative of an extensive genomic and functional remodeling, potentially relating to the lichenization. At the same time they are potentially enriched for overlooked genes, as their importance makes these losses unlikely (c.f. MacArthur et al. 2012). Our focus on these genes is conceptually similar to the gene set completeness approaches employed by *CEGMA* (Parra, Bradnam, and Korf 2007) and *BUSCO* (Simão et al. 2015). Both methods try to estimate the quality of genome assemblies through the presence of evolutionary old and well-conserved genes. While *CEGMA* and *BUSCO* interpret the absence of these genes as signs of an incomplete genome assembly – as a loss of these genes is considerably unlikely – we are searching for genuine losses among such genes. Our search for those potentially lost genes was able to recover around 85% of those LCA_{Lec} genes initially believed to be lost. Artificial gene fusions, non-predicted genes and – to a minor extent – non-assembled regions all contributed to these false positive loss predictions. These artifacts, found across all 5 genomes, affected the gene predictions, even when using a gene predictor dedicated to fungal genomes. Thus extensive curation of gene predictions is still key to drawing sound, evolutionary conclusions.

6.4 Finding footprints of lichenization

We searched for footprints that the lichenization has left on the genomes of the individual symbionts. To this end we set out to investigate gains and losses of genes and associated functions. The genomes of mycobionts and photobionts can be screened for these footprints. Unfortunately, only few and furthermore only evolutionary distantly related genomes are available for the Chlorophyta. This limits the temporal resolution for evolutionary analyses, as a comparative study relies on closely related taxa to pinpoint when events happened. It furthermore hinders the functional annotation of the chlorophyte genomes as there is only little reference data for the annotation

transfer, as demonstrated by the low functional gene annotation rate for *Trebouxia sp.* In contrast, the Lecanoromycetes are embedded in a considerably denser taxon sampling including non-lichenized sister clades. For this reason we confined our comparative analyses on the mycobionts. We investigated the general footprint of lichenization that is shared by the Lecanoromycetes. Subsequently we searched for genomic events that could explain why *L. pustulata* – unlike the other studied Lecanoromycetes – could so far not been grown in axenic culture. For our comparative analyses we thus used a set of Lecanoromycetes and their closest relatives, the Eurotiomycetes and Dothideomycetes (Chapter 5).

We found first evidence for how the establishment of symbioses shaped the genomes of the early Lecanoromycetes. After the split from the Eurotiomycetes, the LCA of the Lecanoromycetes has lost about 10% of its genes, coinciding with the onset of lichenization for this class. Genes involved in the polysaccharide degradation were particularly affected by these losses. The loss of these genes is probably a consequence of the lichenization, as the photobionts provide a set of different sugars to the mycobionts. Similar effects were observed for mycorrhizal fungi that do not rely on degrading plant material (F. Martin et al. 2008; Kohler et al. 2015). Whether these losses are an active adaptation to the symbiosis itself, as hypothesized for the mycorrhizal fungi (Tisserant et al. 2013), or the beneficial loss of obsolete genes (Koskiniemi et al. 2012; Morris, Lenski, and Zinser 2012) will have to be studied further. While the LCA of the Lecanoromycetes also gained around 400 genes after the lichenization, there is currently no model-lecanoromycete. Furthermore, so far there were no molecular biological studies that investigated the function of lecanoromycete genes. Thus, we cannot transfer functions to lecanoromycete-specific genes. In these instances even the denser taxon sampling inside the fungi seems to be insufficient to facilitate high-resolution functional analyses. More data for mycobionts, photobionts, and

their relatives will be needed to enable more comprehensive functional insights. Given that *L. pustulata* could so far not be grown it culture, we expected to find signs of a genomic remodeling that affects particular pathways or includes the loss of individual, functionally central genes. However, our analyses revealed no such remodeling in *L. pustulata*. Its gene family evolution is similar to that of its closest relative, *Umbilicaria muehlenbergii*; and most gene family extractions/expansions indeed happened in the last common ancestor of the two species. A similar picture emerged for the losses of individual genes, with *L. pustulata* and *U. muehlenbergii* sharing a sizeable amount of gene losses in their LCA. Furthermore, neither the gene family contractions nor the gene losses found in *L. pustulata* showed any signs for a significant decrease in function amongst them. Searching for the private loss of LCA_{Lec} amongst the Lecanoromycetes we find no increased loss of well-conserved genes in *L. pustulata*. Thus, we did not find clear signs that can explain the poor culturability of *L. pustulata*, when compared to the other Lecanoromycetes.

It might be that the relevant genes have been lost more than once, with only *L. pustulata* exhibiting a combination of lost genes that interrupts a given function. These genes would be overlooked in our approach as it only allows for private gene losses. Furthermore, there is the possibility that genes are still present in *L. pustulata*, but have lost parts of their functions. It has been hypothesized that especially multi-functional genes are well conserved over long time (He and Zhang 2006; Dudley et al. 2005; Jordan et al. 2004; Salathe, Ackermann, and Bonhoeffer 2006). This might explain why we did not observe marked losses in *L. pustulata*, as the LCA_{Lec} genes might play a role in numerous essential pathways. Due to this, the genes of *L. pustulata* might only have lost partial functionality with some interaction partners, instead of being outright lost. A gene-wise comparison of annotations like Pfam domains and GO terms might help to reveal such functional differentiations between the

genes of *L. pustulata* and their corresponding orthologs in the other Lecanoromycetes.

6.5 Summary

Lichens are interesting models for comparative genomic and evolutionary research: They are evolutionary old symbioses that are frequently found as pioneers in new habitats. Lichens furthermore offer a usually well-defined symbiotic system – with only few interaction partners – that can be easily delineated. Due to this, they can provide insights into how symbiotic organisms form a holobiont. However, high-resolution comparative genomic studies are required for this. In turn, such studies need to be based on high-quality genome reconstructions. The completeness of genome reconstructions is of even more importance when analyses try to focus on evolutionary changes instead of conservation. We see that metagenome skimming alone – while allowing for a first hologenome characterization and draft genome reconstructions – does not suffice to generate adequate assemblies. The additional use of third-generation sequencing methods facilitates a hologenome-reconstruction that is substantially improved in completeness and contiguity. These genomes then enable a detailed characterization of the organisms that are found in lichen symbiosis.

Our analysis of the genomes of the Lecanoromycetes reveals that comparative studies of gene gains and losses are hindered by both the completeness of the gene annotations and the functional annotations. Most available draft genomes of the Lecanoromycetes are incompletely sequenced and annotated. Thus, an additional annotation of so far overlooked genes and the curation of existing gene annotations are needed for unbiased analyses.

So far there is no model organism among the Lecanoromycetes. For this reason we depend on a functional annotation transfer from more distantly related organisms. We thus have only limited or no information about the

function of many genes, particularly for those that are specific to the Lecanoromycetes. Despite these limitations, we find that the Lecanoromycetes have lost a significant number of genes that are involved in the polysaccharide catabolism. In contrast, on the level of genes lost and functional capacities encoded in the genome o*f L. pustulata*, we do not find changes that would explain its observed poor culturability.

To fully understand the full range of interactions of lichen symbionts and the symbiotic nature of the lichen symbiosis (Ahmadjian 1993; Honegger 1998), further analyses will have to rely on comparisons that do not only include the genomes of the different mycobionts, but also their photobionts. Such studies are currently hindered by sparse availability of sequenced genomes, especially of the photobionts (Graham, Wilcox, and Knack 2015; Bhattacharya et al. 2015). We are only at the beginning of exploiting lichens as model organisms for studying the evolution of symbioses and the formation of holobionts. Further metagenomic sequencing of lichen hologenomes will thus be central for gaining a deeper understanding of how symbioses can shape genomes.

References

Abdel-Hameed, Mona, Robert L. Bertrand, Michele D. Piercey-Normore, and John L. Sorensen. 2015. "Putative Identification of the Usnic Acid Biosynthetic Gene Cluster by de Novo Whole-Genome Sequencing of a Lichen-Forming Fungus." *Fungal Biology*. Elsevier Ltd, 1–11. doi:10.1016/j.funbio.2015.10.009.

Aganezov, Sergey S, and Max A Alekseyev. 2016. "CAMSA: A Tool for Comparative Analysis and Merging of Scaffold Assemblies." *bioRxiv*. doi:10.1101/069153.

Ahmadjian, Vernon. 1993. *The Lichen Symbiosis*. New York: John Willey and Sons, Inc.

Ahmadjian, Vernon, and Jerome B. Jacobs. 1981. "Relationship between Fungus and Alga in the Lichen Cladonia Cristatella Tuck." *Nature* 289: 169–72. doi:10.1038/289169a0.

Albalat, Ricard, and Cristian Cañestro. 2016. "Evolution by Gene Loss." *Nat Rev Genet* 17 (7): 379–91. doi:10.1038/nrg.2016.39.

Allhoff, Manuel, Alexander Schönhuth, Marcel Martin, Ivan G Costa, Sven Rahmann, and Tobias Marschall. 2013. "Discovering Motifs That Induce Sequencing Errors." *BMC Bioinformatics* 14 Suppl 5 (Suppl 5). BioMed Central: S1. doi:10.1186/1471-2105-14-S5-S1.

Altenhoff, A M, B Boeckmann, S Capella-Gutierrez, D A Dalquen, T DeLuca, K Forslund, J Huerta-Cepas, et al. 2016. "Standardized Benchmarking in the Quest for Orthologs." *Nat Methods* 13 (5): 425–30. doi:10.1038/nmeth.3830.

Altenhoff, A M, N Škunca, N Glover, C M Train, A Sueki, I Piližota, K Gori, et al. 2015. "The OMA Orthology Database in 2015: Function Predictions, Better Plant Support, Synteny View and Other Improvements." *Nucleic Acids Res* 43 (Database issue): D240-9. doi:10.1093/nar/gku1158.

Altenhoff, A M, RA Studer, M Robinson-Rechavi, and C Dessimoz. 2012. "Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs." Edited by Jonathan A. Eisen. *PLOS Computational Biology* 8 (5). Public Library of Science: e1002514. http://dx.plos.org/10.1371/journal.pcbi.1002514.

Altschul, Stephen F, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402. http://nar.oxfordjournals.org/cgi/content/abstract/25/17/3389.

Amo de Paz, Guillermo, Paloma Cubas, Pradeep K. Divakar, H. Thorsten Lumbsch, and Ana Crespo. 2011. "Origin and Diversification of Major Clades in Parmelioid Lichens (Parmeliaceae, Ascomycota) during the Paleogene Inferred by Bayesian Analysis." Edited by Robert DeSalle. *PLoS ONE* 6 (12). Museo

Regionale di Scienze Naturali: e28161. doi:10.1371/journal.pone.0028161.

Aschenbrenner, Ines Aline, Tomislav Cernava, Gabriele Berg, and Martin Grube. 2016. "Understanding Microbial Multi-Species Symbioses." *Frontiers in Microbiology*. Frontiers Media SA. doi:10.3389/fmicb.2016.00180.

Ashburner, M, CA Ball, and JA Blake. 2000. "Gene Ontology : Tool for the Unification of Biology." *Nature Genetics* 25 (may): 25–29. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc3037419/.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. doi:10.1038/nature15393.

Awad, Sherine, Luiz Irber, and C. Titus Brown. 2017. "Evaluating Metagenome Assembly on a Simple Defined Community with Many Strain Variants." *bioRxiv*. http://www.biorxiv.org/content/early/2017/06/25/155358.

Baker, M. 2012. "De Novo Genome Assembly: What Every Biologist Should Know." *Nature Methods* 9 (4). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 333–37. doi:10.1038/nmeth.1935.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey a Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, et al. 2012. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *J Comput Biol* 19 (5): 455–77. doi:10.1089/cmb.2012.0021.

Bao, Ergude, Changjin Song, and Lingxiao Lan. 2017. "ReMILO: Reference Assisted Misassembly Detection Algorithm Using Short and Long Reads." *Bioinformatics*, August. doi:10.1093/bioinformatics/btx524.

Bao, Weidong, Kenji K. Kojima, and Oleksiy Kohany. 2015. "Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes." *Mobile DNA*. doi:10.1186/s13100-015-0041-9.

Barabote, Ravi D, Jose Thekkiniath, Richard E Strauss, Govindsamy Vediyappan, Joe A Fralick, and Michael J San Francisco. 2011. "Xenobiotic Efflux in Bacteria and Fungi: A Genomics Update." *Advances in Enzymology and Related Areas of Molecular Biology* 77. NIH Public Access: 237–306. http://www.ncbi.nlm.nih.gov/pubmed/21692371.

Baroncelli, R, D B Amby, A Zapparata, S Sarrocco, G Vannacci, G Le Floch, R J Harrison, et al. 2016. "Gene Family Expansions and Contractions Are Associated with Host Range in Plant Pathogens of the Genus Colletotrichum." *BMC Genomics* 17: 555. doi:10.1186/s12864-016-2917-6.

Benjamini, Y, and Y Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300. http://www.jstor.org/stable/10.2307/2346101.

Benjamini, Y, and T P Speed. 2012. "Summarizing and Correcting the GC Content

Bias in High-Throughput Sequencing." *Nucleic Acids Res* 40 (10): e72. doi:10.1093/nar/gks001.

Bennett, G M, J P McCutcheon, B R MacDonald, D Romanovicz, and N A Moran. 2014. "Differential Genome Evolution Between Companion Symbionts in an Insect-Bacterial Symbiosis." *Mbio* 5 (5). doi:10.1128/mBio.01697-14.

Berlemont, Renaud. 2017. "Distribution and Diversity of Enzymes for Polysaccharide Degradation in Fungi." *Scientific Reports*. doi:10.1038/s41598-017-00258-w.

Berlin, Konstantin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. 2015. "Assembling Large Genomes with Single-Molecule Sequencing and Locality-Sensitive Hashing." *Nature Biotechnology*, no. August 2014. Nature Publishing Group: 1–11. doi:10.1038/nbt.3238.

Besemer, J, and M Borodovsky. 2005. "GeneMark: Web Software for Gene Finding in Prokaryotes, Eukaryotes and Viruses." *Nucleic Acids Res* 33 (Web Server issue): W451-4. doi:10.1093/nar/gki487.

Bhattacharya, D, H Qiu, D C Price, and H S Yoon. 2015. "Why We Need More Algal Genomes." *J Phycol* 51 (1): 1–5. doi:10.1111/jpy.12267.

Bie, Tijl De, Nello Cristianini, Jeffery P Demuth, and Matthew W Hahn. 2006. "CAFE: A Computational Tool for the Study of Gene Family Evolution." doi:10.1093/bioinformatics/btl097.

Birney, E, M Clamp, and R Durbin. 2004. "GeneWise and Genomewise." *Genome Research* 14 (5): 988–95. doi:10.1101/gr.1865504.

Blanc, Guillaume, Garry Duncan, Irina Agarkova, Mark Borodovsky, James Gurnon, Alan Kuo, Erika Lindquist, et al. 2010. "The Chlorella Variabilis NC64A Genome Reveals Adaptation to Photosymbiosis, Coevolution with Viruses, and Cryptic Sex." *The Plant Cell* 22 (9). American Society of Plant Biologists: 2943–55. doi:10.1105/tpc.110.076406.

Bleidorn, Christoph. 2015. "Third Generation Sequencing: Technology and Its Potential Impact on Evolutionary Biodiversity Research." *Systematics and Biodiversity* 2000 (January). Taylor & Francis: 1–8. doi:10.1080/14772000.2015.1099575.

Bock, Dan G., Nolan C. Kane, Daniel P. Ebert, and Loren H. Rieseberg. 2014. "Genome Skimming Reveals the Origin of the Jerusalem Artichoke Tuber Crop Species: Neither from Jerusalem nor an Artichoke." *New Phytologist* 201 (3): 1021–30. doi:10.1111/nph.12560.

Boetzer, M, and W Pirovano. 2014. "SSPACE-LongRead: Scaffolding Bacterial Draft Genomes Using Long Read Sequence Information." *BMC Bioinformatics* 15: 211. doi:10.1186/1471-2105-15-211.

Bolger, AM M, Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20. doi:10.1093/bioinformatics/btu170.

Bonfante, Paola, and Andrea Genre. 2010. "Mechanisms Underlying Beneficial Plant-

Fungus Interactions in Mycorrhizal Symbiosis." *Nature Communications* 1 (4). Nature Publishing Group: 48. doi:10.1038/ncomms1046.

Boothby, Thomas C, Jennifer R Tenlen, Frank W Smith, Jeremy R Wang, Kiera A Patanella, Erin Osborne, Sophia C Tintori, et al. 2015. "Evidence for Extensive Horizontal Gene Transfer from the Draft Genome of a Tardigrade," 1–6. doi:10.1073/pnas.1510461112.

Bose, Tungadri, Mohammed Monzoorul Haque, CVSK Reddy, Sharmila S. Mande, RJ Ram, and PM Richardson. 2015. "COGNIZER: A Framework for Functional Annotation of Metagenomic Datasets." Edited by Gajendra P. S. Raghava. *PLOS ONE* 10 (11). Public Library of Science: e0142102. doi:10.1371/journal.pone.0142102.

Botero-Castro, Fidel, Emeric Figuet, Marie-ka Tilak, Benoit Nabholz, and Nicolas Galtier. 2017. "Avian Genomes Revisited: Hidden Genes Uncovered and the Rates vs. Traits Paradox in Birds." *Molecular Biology and Evolution*, September. doi:10.1093/molbev/msx236.

Bouckaert, R, J Heled, D Kühnert, T Vaughan, C H Wu, D Xie, M A Suchard, A Rambaut, and A J Drummond. 2014. "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis." *PLoS Comput Biol* 10 (4): e1003537. doi:10.1371/journal.pcbi.1003537.

Bouws, H, A Wattenberg, and H Zorn. 2008. "Fungal Secretomes - Nature's Toolbox for White Biotechnology." *Applied Microbiology and Biotechnology* 80 (3): 381–88. doi:10.1007/s00253-008-1572-5.

Bradnam, Keith R, Joseph N Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, et al. 2013. "Assemblathon 2: Evaluating de Novo Methods of Genome Assembly in Three Vertebrate Species." *Gigascience* 2 (1): 10. doi:10.1186/2047-217X-2-10.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2014. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nat Methods* 12 (1): 59–60. doi:10.1038/nmeth.3176.

Cardinale, Massimiliano, Anna Maria Puglia, and Martin Grube. 2006. "Molecular Analysis of Lichen-Associated Bacterial Communities." *FEMS Microbiology Ecology* 57 (3): 484–95. doi:10.1111/j.1574-6941.2006.00133.x.

Carthew, Richard W, and Erik J Sontheimer. 2009. "Origins and Mechanisms of miRNAs and siRNAs." *Cell* 136 (4). NIH Public Access: 642–55. doi:10.1016/j.cell.2009.01.035.

Caruso, Alexandro, and Jörgen Rudolphi. 2009. "Influence of Substrate Age and Quality on Species Diversity of Lichens and Bryophytes on Stumps." *The Bryologist* 112 (3). The American Bryological and Lichenological Society, Inc New York Botanical Garden Bronx, NY 10458-5126 : 520–31. doi:10.1639/0007-2745-112.3.520.

Castro, H F, A T Classen, E E Austin, R J Norby, and C W Schadt. 2010. "Soil

Microbial Community Responses to Multiple Experimental Climate Change Drivers." *Appl Environ Microbiol* 76 (4): 999–1007. doi:10.1128/AEM.02874-09.

Cernava, Tomislav, Armin Erlacher, Ines Aline Aschenbrenner, Lisa Krug, Christian Lassek, Katharina Riedel, Martin Grube, and Gabriele Berg. 2017. "Deciphering Functional Diversification within the Lichen Microbiota by Meta-Omics." *Microbiome* 5 (1): 82. doi:10.1186/s40168-017-0303-5.

Cernava, Tomislav, Henry Müller, Ines A Aschenbrenner, Martin Grube, and Gabriele Berg. 2015. "Analyzing the Antagonistic Potential of the Lichen Microbiome against Pathogens by Bridging Metagenomic with Culture Studies." *Frontiers in Microbiology* 6 (June): 620. doi:10.3389/fmicb.2015.00620.

Cha, Soyeon, and David McK Bird. 2016. "Optimizing K-Mer Size Using a Variant Grid Search to Enhance de Novo Genome Assembly." *Bioinformation* 12 (2). Biomedical Informatics Publishing Group: 36–40. doi:10.6026/97320630012036.

Chakraborty, M, J G Baldwin-Brown, A D Long, and J J Emerson. 2016. "Contiguous and Accurate de Novo Assembly of Metazoan Genomes with Modest Long Read Coverage." *Nucleic Acids Res*. doi:10.1093/nar/gkw654.

Chaudhari, P, B Ahmed, D L Joly, and H Germain. 2014. "Effector Biology during Biotrophic Invasion of Plant Cells." *Virulence* 5 (7): 703–9. doi:10.4161/viru.29652.

Chevreux, B, T Pfisterer, B Drescher, AJ Driesel, WEG Müller, T Wetter, and S Suhai. 2004. "Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs." *Genome Research* 14 (6): 1147–59. doi:10.1101/gr.1917404.

Chevreux, B, T Wetter, and S Suhai. 1999. "Genome Sequence Assembly Using Trace Signals and Additional Sequence Information." *German Conference on Bioinformatics*, no. 1995. http://www.bioinfo.de/isb/gcb99/talks/chevreux/main.html.

Chikhi, R., and P. Medvedev. 2014. "Informed and Automated K-Mer Size Selection for Genome Assembly." *Bioinformatics* 30 (1): 31–37. doi:10.1093/bioinformatics/btt310.

Chin, Chen-Shan, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, Alicia Clum, Christopher Dunn, et al. 2016. "Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing." *Nature Methods*. doi:10.1038/nmeth.4035.

Clavijo, Bernardo J, Luca Venturini, Christian Schudoma, Gonzalo Garcia Accinelli, Gemy Kaithakottil, Jonathan Wright, Philippa Borrill, et al. 2017. "An Improved Assembly and Annotation of the Allohexaploid Wheat Genome Identifies Complete Families of Agronomic Genes and Provides Genomic Evidence for Chromosomal Translocations." *Genome Research* 27 (5). Cold Spring Harbor Laboratory Press: 885–96. doi:10.1101/gr.217117.116.

Coghlan, Avril, Tristan J Fiedler, Sheldon J McKay, Paul Flicek, Todd W Harris, Darin Blasiar, and Lincoln D Stein. 2008. "nGASP--the Nematode Genome

Annotation Assessment Project." *BMC Bioinformatics*. doi:10.1186/1471-2105-9-549.

Coleman, J J, and E Mylonakis. 2009. "Efflux in Fungi: La Pièce de Résistance." *PLoS Pathog* 5 (6): e1000486. doi:10.1371/journal.ppat.1000486.

Compeau, Phillip E C, Pavel A Pevzner, and Glenn Tesler. 2011. "How to Apply de Bruijn Graphs to Genome Assembly." *Nature Biotechnology* 29 (11): 987–91. doi:10.1038/nbt.2023.

Conesa, A, S Götz, JM García-Gómez, J Terol, M Talón, and M Robles. 2005. "Blast2GO: A Universal Tool for Annotation, Visualization and Analysis in Functional Genomics Research." *Bioinformatics* 21 (18): 3674–76. doi:10.1093/bioinformatics/bti610.

Cubero, Oscar F., and Ana Crespo. 2002. "Isolation of Nucleic Acids from Lichens." In *Protocols in Lichenology*, edited by I Kranner, R Beckett, and A Varma, 381–391. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-56359-1_23.

D'haeseleer, Patrik. 2006. "What Are DNA Sequence Motifs?" *Nature Biotechnology* 24 (4). Nature Publishing Group: 423–25. doi:10.1038/nbt0406-423.

Dahan, Romain A, Rebecca P Duncan, Alex C C Wilson, and Liliana M Dávalos. 2015. "Amino Acid Transporter Expansions Associated with the Evolution of Obligate Endosymbiosis in Sap-Feeding Insects (Hemiptera: Sternorrhyncha)." *BMC Evolutionary Biology* 15 (March). BioMed Central: 52. doi:10.1186/s12862-015-0315-3.

Dal Grande, Francesco, Rahul Sharma, Anjuli Meiser, Gregor Rolshausen, Burkhard Büdel, Bagdevi Mishra, Marco Thines, Jürgen Otte, Markus Pfenninger, and Imke Schmitt. 2017. "Adaptive Differentiation Coincides with Local Bioclimatic Conditions along an Elevational Cline in Populations of a Lichen-Forming Fungus." *BMC Evolutionary Biology* 17 (1): 93. doi:10.1186/s12862-017-0929-8.

David, Lior, Wolfgang Huber, Marina Granovskaia, Joern Toedling, Curtis J Palm, Lee Bofkin, Ted Jones, Ronald W Davis, and Lars M Steinmetz. 2006. "A High-Resolution Map of Transcription in the Yeast Genome." *Proceedings of the National Academy of Sciences of the United States of America* 103 (14). National Academy of Sciences: 5320–25. doi:10.1073/pnas.0601091103.

Dean, Ralph, Jan A L Van Kan, Zacharias A. Pretorius, Kim E. Hammond-Kosack, Antonio Di Pietro, Pietro D. Spanu, Jason J. Rudd, et al. 2012. "The Top 10 Fungal Pathogens in Molecular Plant Pathology." *Molecular Plant Pathology*. doi:10.1111/j.1364-3703.2011.00783.x.

Delmont, T O, and A M Eren. 2016. "Identifying Contamination with Advanced Visualization and Analysis Practices: Metagenomic Approaches for Eukaryotic Genome Assemblies." *PeerJ* 4: e1839. doi:10.7717/peerj.1839.

Deng, X., S. N. Naccache, T. Ng, S. Federman, L. Li, C. Y. Chiu, and E. L. Delwart. 2015. "An Ensemble Strategy That Significantly Improves de Novo Assembly of

Microbial Genomes from Metagenomic next-Generation Sequencing Data." *Nucleic Acids Research* 5771 (18): 1–11. doi:10.1093/nar/gkv002.

Denton, James F., Jose Lugo-Martinez, Abraham E. Tucker, Daniel R. DR Schrider, Wesley C. Warren, and MW Matthew W. Hahn. 2014. "Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies." Edited by Roderic Guigo. *PLoS Comput Biol* 10 (12). Public Library of Science: e1003998. doi:10.1371/journal.pcbi.1003998.

Desai, Aarti, Veer Singh Marwah, Akshay Yadav, Vineet Jha, Kishor Dhaygude, Ujwala Bangar, Vivek Kulkarni, and Abhay Jere. 2013. "Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data." Edited by Shu-Dong Zhang. *PLoS ONE* 8 (4). Public Library of Science: e60204. doi:10.1371/journal.pone.0060204.

Dickman, M B, and O Yarden. 1999. "Serine/threonine Protein Kinases and Phosphatases in Filamentious Fungi." *Fungal Genetics and Biology : FG & B* 26 (2). Academic Press: 99–117. doi:10.1006/fgbi.1999.1118.

Donaldson, Michael E, Lauren A Ostrowski, Kristi M Goulet, and Barry J Saville. 2017. "Transcriptome Analysis of Smut Fungi Reveals Widespread Intergenic Transcription and Conserved Antisense Transcript Expression." *BMC Genomics* 18 (1). BioMed Central: 340. doi:10.1186/s12864-017-3720-8.

Döring, Andreas, David Weese, Tobias Rausch, and Knut Reinert. 2008. "SeqAn An Efficient, Generic C++ Library for Sequence Analysis." *BMC Bioinformatics* 9 (1): 11. doi:10.1186/1471-2105-9-11.

Drăgan, Monica-Andreea, Ismail Moghul, Anurag Priyam, Claudio Bustos, and Yannick Wurm. 2016. "GeneValidator: Identify Problems with Protein-Coding Gene Predictions." *Bioinformatics* 32 (10). Oxford University Press: 1559–61. doi:10.1093/bioinformatics/btw015.

Driscoll, Connor B, Timothy G Otten, Nathan M Brown, and Theo W Dreher. 2017. "Towards Long-Read Metagenomics: Complete Assembly of Three Novel Genomes from Bacteria Dependent on a Diazotrophic Cyanobacterium in a Freshwater Lake Co-Culture." *Standards in Genomic Sciences*. doi:10.1186/s40793-017-0224-8.

Dudley, Aimée Marie, Daniel Maarten Janse, Amos Tanay, Ron Shamir, and George McDonald Church. 2005. "A Global View of Pleiotropy and Phenotypically Derived Gene Function in Yeast." *Molecular Systems Biology* 1. European Molecular Biology Organization: 2005.0001. doi:10.1038/msb4100004.

Duncan, Rebecca P., Honglin Feng, Douglas M. Nguyen, Alex C. C. Wilson, and Abbot P. 2016. "Gene Family Expansions in Aphids Maintained by Endosymbiotic and Nonsymbiotic Traits." *Genome Biology and Evolution* 8 (3). ACM Press, New York: 753–64. doi:10.1093/gbe/evw020.

Dunn, Casey W., and Catriona Munro. 2016. "Comparative Genomics and the

Diversity of Life." *Zoologica Scripta* 45 (S1): 5–13. doi:10.1111/zsc.12211.

Dunn, T M, and D Shortle. 1990. "Null Alleles of SAC7 Suppress Temperature-Sensitive Actin Mutations in Saccharomyces Cerevisiae." *Molecular and Cellular Biology* 10 (5): 2308–14. http://www.ncbi.nlm.nih.gov/pubmed/2183030.

Dunne, M P, and S Kelly. 2017. "OrthoFiller: Utilising Data from Multiple Species to Improve the Completeness of Genome Annotations." *BMC Genomics* 18 (1): 390. doi:10.1186/s12864-017-3771-x.

Earl, Dent, Keith Bradnam, John St. John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, et al. 2011. "Assemblathon 1: A Competitive Assessment of de Novo Short Read Assembly Methods." *Genome Research* 21 (12): 2224–41. doi:10.1101/gr.126599.111.

Ebersberger, I, S Strauss, and A von Haeseler. 2009. "HaMStR: Profile Hidden Markov Model Based Search for Orthologs in ESTs." *BMC Evolutionary Biology* 9 (January): 157. doi:10.1186/1471-2148-9-157.

Eddy, S R. 2011. "Accelerated Profile HMM Searches." *Plos Computational Biology* 7 (10). doi:10.1371/journal.pcbi.1002195.

Edgar, R. C., and E. W. Myers. 2005. "PILER: Identification and Classification of Genomic Repeats." *Bioinformatics* 21 (Suppl 1): i152–58. doi:10.1093/bioinformatics/bti1003.

Edwards, Arwyn, Aliyah R Debbonaire, Birgit Sattler, Luis AJ Mur, and Andrew J Hodson. 2016. "Extreme Metagenomics Using Nanopore DNA Sequencing: A Field Report from Svalbard, 78 N." *bioRxiv*. http://www.biorxiv.org/node/20341.full.

Elgar, G, M S Clark, S Meek, S Smith, S Warner, Y J Edwards, N Bouchireb, et al. 1999. "Generation and Analysis of 25 Mb of Genomic DNA from the Pufferfish Fugu Rubripes by Sequence Scanning." *Genome Research* 9 (10). Cold Spring Harbor Laboratory Press: 960–71. doi:10.1101/GR.9.10.960.

Ellinghaus, David, Stefan Kurtz, and Ute Willhoeft. 2008. "LTRharvest, an Efficient and Flexible Software for de Novo Detection of LTR Retrotransposons." *BMC Bioinformatics*. doi:10.1186/1471-2105-9-18.

Elmer, KR, S Fan, HM Gunter, JC Jones, S Boekhoff, S Kuraku, and A Meyer. 2010. "Rapid Evolution and Selection Inferred from the Transcriptomes of Sympatric Crater Lake Cichlid Fishes." *Molecular Ecology* 19 Suppl 1 (March): 197–211. http://www.ncbi.nlm.nih.gov/pubmed/20331780.

Emanuelsson, O, S Brunak, G von Heijne, and H Nielsen. 2007. "Locating Proteins in the Cell Using TargetP, SignalP and Related Tools." *Nat Protoc* 2 (4): 953–71. doi:10.1038/nprot.2007.131.

Enright, A J, S Van Dongen, and C A Ouzounis. 2002. "An Efficient Algorithm for Large-Scale Detection of Protein Families." *Nucleic Acids Res* 30 (7): 1575–84. https://www.ncbi.nlm.nih.gov/pubmed/11917018.

Eren, A M, Ö C Esen, C Quince, J H Vineis, H G Morrison, M L Sogin, and T O

Delmont. 2015. “Anvi’o: An Advanced Analysis and Visualization Platform for ’Omics Data.” *PeerJ* 3: e1319. doi:10.7717/peerj.1319.

Erlacher, Armin, Tomislav Cernava, Massimiliano Cardinale, Jung Soh, Christoph W. Sensen, Martin Grube, and Gabriele Berg. 2015. “Rhizobiales as Functional and Endosymbiontic Members in the Lichen Symbiosis of Lobaria Pulmonaria L.” *Frontiers in Microbiology* 6 (February): 1–9. doi:10.3389/fmicb.2015.00053.

Feldmeyer, Barbara, Bastian Greshake, Elisabeth Funke, Ingo Ebersberger, and Markus Pfenninger. 2015. “Positive Selection in Development and Growth Rate Regulation Genes Involved in Species Divergence of the Genus Radix.” *BMC Evolutionary Biology* 15 (1). BMC Evolutionary Biology: 164. doi:10.1186/s12862-015-0434-x.

Field, Katie J, William R Rimington, Martin I Bidartondo, Kate E Allinson, David J Beerling, Duncan D Cameron, Jeffrey G Duckett, Jonathan R Leake, and Silvia Pressel. 2015. “Functional Analysis of Liverworts in Dual Symbiosis with Glomeromycota and Mucoromycotina Fungi under a Simulated Palaeozoic CO2 Decline.” *The ISME Journal* 10 (6). Nature Publishing Group: 1–13. doi:10.1038/ismej.2015.204.

Finn, Robert D, P Coggill, R Y Eberhardt, S R Eddy, J Mistry, A L Mitchell, S C Potter, et al. 2016. “The Pfam Protein Families Database: Towards a More Sustainable Future.” *Nucleic Acids Res* 44 (D1): D279-85. doi:10.1093/nar/gkv1344.

Finn, Robert D, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E Pollington, O Luke Gavin, et al. 2010. “The Pfam Protein Families Database.” *Nucleic Acids Research* 38 (Database issue): D211-22. doi:10.1093/nar/gkp985.

Fisher, Matthew C, Daniel A Henk, Cheryl J Briggs, John S Brownstein, Lawrence C Madoff, Sarah L McCraw, and Sarah J Gurr. 2012. “Emerging Fungal Threats to Animal, Plant and Ecosystem Health.” *Nature* 484 (7393). NIH Public Access: 186–94. doi:10.1038/nature10947.

Forsberg, Kevin J, Sanket Patel, Molly K Gibson, Christian L Lauber, Rob Knight, Noah Fierer, and Gautam Dantas. 2014. “Bacterial Phylogeny Structures Soil Resistomes across Habitats.” *Nature* 509 (7502): 612–16. http://www.ncbi.nlm.nih.gov/pubmed/24847883.

Frank, J A, Y Pan, A Tooming-Klunderud, V G H Eijsink, A C McHardy, A J Nederbragt, and P B Pope. 2016. “Improved Metagenome Assemblies and Taxonomic Binning Using Long-Read Circular Consensus Sequence Data.” *Scientific Reports*. doi:10.1038/srep25373.

Franzosa, E. A., X. C. Morgan, N. Segata, L. Waldron, J. Reyes, A. M. Earl, G. Giannoukos, et al. 2014. “Relating the Metatranscriptome and Metagenome of the Human Gut.” *Proceedings of the National Academy of Sciences* 111 (22): E2329–38. doi:10.1073/pnas.1319284111.

Gadd, G. M. 2010. “Metals, Minerals and Microbes: Geomicrobiology and Bioremediation.” *Microbiology* 156 (3). Microbiology Society: 609–43.

doi:10.1099/mic.0.037143-0.

Galagan, James E, Matthew R Henn, Li-Jun Ma, Christina A Cuomo, and Bruce Birren. 2005. "Genomics of the Fungal Kingdom: Insights into Eukaryotic Biology." *Genome Research* 15 (12). Cold Spring Harbor Laboratory Press: 1620–31. doi:10.1101/gr.3767105.

Gan, P, M Narusaka, N Kumakura, A Tsushima, Y Takano, Y Narusaka, and K Shirasu. 2016. "Genus-Wide Comparative Genome Analyses of Colletotrichum Species Reveal Specific Gene Family Losses and Gains during Adaptation to Specific Infection Lifestyles." *Genome Biol Evol* 8 (5): 1467–81. doi:10.1093/gbe/evw089.

Garcia, K, P M Delaux, K R Cope, and J M Ané. 2015. "Molecular Signals Required for the Establishment and Maintenance of Ectomycorrhizal Symbioses." *New Phytol* 208 (1): 79–87. doi:10.1111/nph.13423.

Gazis, R, A Kuo, R Riley, K LaButti, A Lipzen, J Lin, M Amirebrahimi, et al. 2016. "The Genome of Xylona Heveae Provides a Window into Fungal Endophytism." *Fungal Biol* 120 (1): 26–42. doi:10.1016/j.funbio.2015.10.002.

Ghurye, Jay S, Victoria Cepeda-Espinoza, and Mihai Pop. 2016. "Metagenomic Assembly: Overview, Challenges and Applications." *The Yale Journal of Biology and Medicine* 89 (3). Yale Journal of Biology and Medicine: 353–62. http://www.ncbi.nlm.nih.gov/pubmed/27698619.

Gnerre, S, I Maccallum, D Przybylski, FJ Ribeiro, JN Burton, BJ Walker, T Sharpe, et al. 2011. "High-Quality Draft Assemblies of Mammalian Genomes from Massively Parallel Sequence Data." *Proceedings of the National Academy of Sciences* 108 (4): 1513–18. doi:10.1073/pnas.1017351108.

Götz, Stefan, Juan Miguel García-Gómez, Javier Terol, Tim D Williams, Shivashankar H Nagaraj, María José Nueda, Montserrat Robles, et al. 2008. "High-Throughput Functional Annotation and Data Mining with the Blast2GO Suite." *Nucleic Acids Research* 36 (10): 3420–35. doi:10.1093/nar/gkn176.

Grabherr, MG, BJ Haas, M Yassour, JZ Levin, DA Thompson, I Amit, X Adiconis, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology* 29 (7). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 644–52. doi:10.1038/nbt.1883.

Graham, L E, L W Wilcox, and J J Knack. 2015. "Why We Need More Algal metagenomes(1)." *J Phycol* 51 (6): 1029–36. doi:10.1111/jpy.12344.

Grigoriev, Igor V., Roman Nikitin, Sajeet Haridas, Alan Kuo, Robin Ohm, Robert Otillar, Robert Riley, et al. 2014. "MycoCosm Portal: Gearing up for 1000 Fungal Genomes." *Nucleic Acids Research* 42 (D1). Elsevier, Amsterdam: D699–704. doi:10.1093/nar/gkt1183.

Grube, Martin, and Gabriele Berg. 2009. "Microbial Consortia of Bacteria and Fungi with Focus on the Lichen Symbiosis." *Fungal Biology Reviews* 23 (3). Elsevier Ltd:

72–85. doi:10.1016/j.fbr.2009.10.001.

Grube, Martin, Tomislav Cernava, Jung Soh, Stephan Fuchs, Ines Aschenbrenner, Christian Lassek, Uwe Wegner, et al. 2015. "Exploring Functional Contexts of Symbiotic Sustain within Lichen-Associated Bacteria by Comparative Omics." *The ISME Journal* 9 (2): 1–13. doi:10.1038/ismej.2014.138.

Grube, Martin, and Toby Spribille. 2012. "Exploring Symbiont Management in Lichens." *Molecular Ecology* 21 (13): 3098–99. doi:10.1111/j.1365-294X.2012.05647.x.

Gueidan, C, C R Villaseñor, G S de Hoog, A A Gorbushina, W A Untereiner, and F Lutzoni. 2008. "A Rock-Inhabiting Ancestor for Mutualistic and Pathogen-Rich Fungal Lineages." *Studies in Mycology* 61. CBS Fungal Biodiversity Centre: 111–19. doi:10.3114/sim.2008.61.11.

Guigó, Roderic, Paul Flicek, Josep F. Abril, Alexandre Reymond, Julien Lagarde, France Denoeud, Stylianos Antonarakis, et al. 2006. "EGASP: The Human ENCODE Genome Annotation Assessment Project." *Genome Biology*. doi:10.1186/gb-2006-7-s1-s2.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. "QUAST: Quality Assessment Tool for Genome Assemblies." *Bioinformatics* 29 (8). Oxford University Press: 1072–75. doi:10.1093/bioinformatics/btt086.

Guzman, Christine, and Cecilia Conaco. 2016. "Comparative Transcriptome Analysis Reveals Insights into the Streamlined Genomes of Haplosclerid Demosponges." *Scientific Reports*. doi:10.1038/srep18774.

Haas, Brian J, Steven L Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E Allen, Joshua Orvis, Owen White, C Robin Buell, and Jennifer R Wortman. 2008. "Automated Eukaryotic Gene Structure Annotation Using EVidenceModeler and the Program to Assemble Spliced Alignments." *Genome Biology* 9 (1). BioMed Central: R7. doi:10.1186/gb-2008-9-1-r7.

Haider, Bahlul, TH Ahn, Brian Bushnell, and Juanjuan Chai. 2014. "Omega: An Overlap-Graph de Novo Assembler for Metagenomics." *Bioinformatics*, 1–6. http://bioinformatics.oxfordjournals.org/content/early/2014/07/06/bioinformatics.btu395.short.

Havird, Justin C., and Michael M. Miyamoto. 2010. "The Importance of Taxon Sampling in Genomic Studies: An Example from the Cyclooxygenases of Teleost Fishes." *Molecular Phylogenetics and Evolution* 56 (1): 451–55. doi:10.1016/j.ympev.2010.04.003.

Hawksworth, D. L. 1988. "The Variety of Fungal-algal Symbioses, Their Evolutionary Significance, and the Nature of Lichens." *Botanical Journal of the Linnean Society* 96 (1). Academic Press, New York: 3–20. doi:10.1111/j.1095-8339.1988.tb00623.x.

Hawksworth, D. L., and Robert Lücking. 2017. "Fungal Diversity Revisited: 2.2 to 3.8 Million Species." *Microbiology Spectrum* 5 (4). asm Pub2Web. doi:10.1128/microbiolspec.FUNK-0052-2016.

He, Xionglei, and Jianzhi Zhang. 2006. “Toward a Molecular Understanding of Pleiotropy.” *Genetics* 173 (4). Genetics Society of America: 1885–91. doi:10.1534/genetics.106.060269.

Hehenberger, E, F Burki, M Kolisko, and P J Keeling. 2016. “Functional Relationship between a Dinoflagellate Host and Its Diatom Endosymbiont.” *Mol Biol Evol*. doi:10.1093/molbev/msw109.

Hestmark, G., G. Hestmark, B. Schroeter, B. Schroeter, L. Kappen, and L. Kappen. 1997. “Intrathalline and Size-Dependent Patterns of Activity in Lasallia Pustulata and Their Possible Consequences for Competitive Interactions.” *Functional Ecology* 11 (3). Blackwell Science Ltd: 318–22. doi:10.1046/j.1365-2435.1997.00086.x.

Heupel, Stephanie, Birgit Roser, Hannah Kuhn, Marc-Henri Lebrun, Francois Villalba, and Natalia Requena. 2010. “Erl1, a Novel Era-Like GTPase from *Magnaporthe Oryzae* , Is Required for Full Root Virulence and Is Conserved in the Mutualistic Symbiont *Glomus Intraradices*.” *Molecular Plant-Microbe Interactions* 23 (1): 67–81. doi:10.1094/MPMI-23-1-0067.

Hodkinson, Brendan P., Neil R. Gottel, Christopher W. Schadt, and Fran??ois Lutzoni. 2012. “Photoautotrophic Symbiont and Geography Are Major Factors Affecting Highly Structured and Diverse Bacterial Communities in the Lichen Microbiome.” *Environmental Microbiology* 14 (1): 147–61. doi:10.1111/j.1462-2920.2011.02560.x.

Hoff, Katharina J., Simone Lange, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke. 2016. “BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1.” *Bioinformatics* 32 (5): 767–69. doi:10.1093/bioinformatics/btv661.

Hoff, Katharina J., and M Stanke. 2015. “Current Methods for Automated Annotation of Protein-Coding Genes.” *Current Opinion in Insect Science* 7 (February): 8–14. doi:10.1016/j.cois.2015.02.008.

Holt, C, and M Yandell. 2011. “MAKER2: An Annotation Pipeline and Genome-Database Management Tool for Second-Generation Genome Projects.” *BMC Bioinformatics* 12: 491. doi:10.1186/1471-2105-12-491.

Honegger, Rosmarie. 1998. “The Lichen Symbiosis - What Is so Spectacular about It?” *The Lichenologist* 30 (3). Cambridge University Press: 193. doi:10.1017/S002428299200015X.

Honegger, Rosemarie. 2000. “Simon Schwendener (1829–1919) and the Dual Hypothesis of Lichens.” *The Bryologist* 103 (4): 710–19. doi:10.1639/0007-2745(2000)103.

Honegger, Rosmarie, D Edwards, and L Axe. 2013. “The Earliest Records of Internally Stratified Cyanobacterial and Algal Lichens from the Lower Devonian of the Welsh Borderland.” *New Phytol* 197 (1): 264–75. doi:10.1111/nph.12009.

Horbach, Ralf, Aura Rocio Navarro-Quesada, Wolfgang Knogge, and Holger B. Deising. 2011. "When and How to Kill a Plant Cell: Infection Strategies of Plant Pathogenic Fungi." *Journal of Plant Physiology* 168 (1): 51–62. doi:10.1016/j.jplph.2010.06.014.

Huang, Weichun, Leping Li, Jason R Myers, and Gabor T Marth. 2012. "ART: A next-Generation Sequencing Read Simulator." *Bioinformatics (Oxford, England)* 28 (4): 593–94. doi:10.1093/bioinformatics/btr708.

Huang, X, and a Madan. 1999. "CAP3: A DNA Sequence Assembly Program." *Genome Research* 9 (9): 868–77. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=310812&tool=pmcentrez&rendertype=abstract.

Hubbard, T J, A G Murzin, S E Brenner, and C Chothia. 1997. "SCOP: A Structural Classification of Proteins Database." *Nucleic Acids Research* 25 (1). Oxford University Press: 236–39. http://www.ncbi.nlm.nih.gov/pubmed/9016544.

Hubley, Robert, Robert D Finn, Jody Clements, Sean R Eddy, Thomas A Jones, Weidong Bao, Arian F A Smit, and Travis J Wheeler. 2016. "The Dfam Database of Repetitive DNA Families." *Nucleic Acids Research* 44 (D1). Oxford University Press: D81-9. doi:10.1093/nar/gkv1272.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data." doi:10.1093/molbev/msw046.

Hunt, Martin, Taisei Kikuchi, Mandy Sanders, Chris Newbold, Matthew Berriman, and Thomas D Otto. 2013. "REAPR: A Universal Tool for Genome Assembly Evaluation." *Genome Biology* 14 (5). BioMed Central Ltd: R47. doi:10.1186/gb-2013-14-5-r47.

Hunt, Martin, N D Silva, T D Otto, J Parkhill, J A Keane, and S R Harris. 2015. "Circlator: Automated Circularization of Genome Assemblies Using Long Sequencing Reads." *Genome Biol* 16: 294. doi:10.1186/s13059-015-0849-0.

Huson, Daniel H, Suparna Mitra, Hans-Joachim J Ruscheweyh, Nico Weber, and Stephan C Schuster. 2011. "Integrative Analysis of Environmental Sequences Using MEGAN4." *Genome Research* 21 (9). Cold Spring Harbor Laboratory Press: 1552–60. doi:10.1101/gr.120618.111.

Hyatt, Doug, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11 (1): 119. doi:10.1186/1471-2105-11-119.

Jain, Miten, Sergey Koren, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, et al. 2017. "Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads." *bioRxiv*. http://www.biorxiv.org/content/early/2017/04/20/128835.

Jiang, Ning. 2013. "Overview of Repeat Annotation and De Novo Repeat

Identification." In *Methods in Molecular Biology (Clifton, N.J.)*, 1057:275–87. doi:10.1007/978-1-62703-568-2_20.

Jordan, I. King, Leonardo Mariño-Ramírez, Yuri I. Wolf, and Eugene V. Koonin. 2004. "Conservation and Coevolution in the Scale-Free Human Gene Coexpression Network." *Molecular Biology and Evolution* 21 (11). Oxford University Press: 2058–70. doi:10.1093/molbev/msh222.

Junttila, Sini M, and Stephen Rudd. 2012. "Characterization of a Transcriptome from a Non-Model Organism, Cladonia Rangiferina, the Grey Reindeer Lichen, Using High-Throughput next Generation Sequencing and EST Sequence Data." *BMC Genomics* 13: 575. doi:10.1186/1471-2164-13-575.

Kajitani, Rei, Kouta Toshimoto, Hideki Noguchi, Atsushi Toyoda, Yoshitoshi Ogura, Miki Okuno, Mitsuru Yabana, et al. 2014. "Efficient de Novo Assembly of Highly Heterozygous Genomes from Whole-Genome Shotgun Short Reads." *Genome Research* 24: 1384–95. doi:10.1101/gr.170720.113.

Kampa, Annette, Andrey N Gagunashvili, Tobias A M Gulder, Brandon I Morinaka, Cristina Daolio, Markus Godejohann, Vivian P W Miao, Jörn Piel, and Ólafur S Andrésson. 2013. "Metagenomic Natural Product Discovery in Lichen Provides Evidence for a Family of Biosynthetic Pathways in Diverse Symbioses." *Proceedings of the National Academy of Sciences of the United States of America* 110 (33): E3129-37. doi:10.1073/pnas.1305867110.

Kanehisa, Minoru, Susumu Goto, Masahiro Hattori, Kiyoko F Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. 2006. "From Genomics to Chemical Genomics: New Developments in KEGG." *Nucleic Acids Research* 34 (Database issue): D354-7. doi:10.1093/nar/gkj102.

Kanehisa, Minoru, Y Sato, and K Morishima. 2016. "BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences." *J Mol Biol* 428 (4): 726–31. doi:10.1016/j.jmb.2015.11.006.

Katoh, Kazutaka, and Hiroyuki Toh. 2008. "Recent Developments in the MAFFT Multiple Sequence Alignment Program." *Briefings in Bioinformatics*. doi:10.1093/bib/bbn013.

Katz, Jonathan E, Mensur Dlakić, and Steven Clarke. 2003. "Automated Identification of Putative Methyltransferases from Genomic Open Reading Frames." *Molecular & Cellular Proteomics : MCP* 2 (8). American Society for Biochemistry and Molecular Biology: 525–40. doi:10.1074/mcp.M300037-MCP200.

Kent, W. J. 2002. "BLAT--the BLAST-like Alignment Tool." *Genome Res* 12 (4): 656–64. doi:10.1101/gr.229202. Article published online before March 2002.

Kidron, Giora J., and Marina Temina. 2010. "Lichen Colonization on Cobbles in the Negev Desert Following 15 Years in the Field." *Geomicrobiology Journal* 27 (5). Taylor & Francis Group : 455–63. doi:10.1080/01490450903490805.

Kielak, Anna M., Cristine C. Barreto, George A. Kowalchuk, Johannes A. van Veen,

and Eiko E. Kuramae. 2016. "The Ecology of Acidobacteria: Moving beyond Genes and Genomes." *Frontiers in Microbiology*. Frontiers. doi:10.3389/fmicb.2016.00744.

Kim, D, B Langmead, and S L Salzberg. 2015. "HISAT: A Fast Spliced Aligner with Low Memory Requirements." *Nat Methods* 12 (4): 357–60. doi:10.1038/nmeth.3317.

Kishimoto, Noriaki, Yoshimasa Kosako, and Tatsuo Tano. 1991. "Acidobacterium Capsulatum Gen. Nov., Sp. Nov.: An Acidophilic Chemoorganotrophic Bacterium Containing Menaquinone from Acidic Mineral Environment." *Current Microbiology* 22 (1). Springer-Verlag: 1–7. doi:10.1007/BF02106205.

Klasberg, Steffen, Tristan Bitard-Feildel, and Ludovic Mallet. 2016. "Computational Identification of Novel Genes: Current and Future Perspectives." *Bioinformatics and Biology Insights* 10. SAGE Publications: 121–31. doi:10.4137/BBI.S39950.

Kohler, Annegret, Alan Kuo, Laszlo G Nagy, Emmanuelle Morin, Kerrie W Barry, Francois Buscot, Björn Canbäck, et al. 2015. "Convergent Losses of Decay Mechanisms and Rapid Turnover of Symbiosis Genes in Mycorrhizal Mutualists." *Nature Genetics* 47 (4): 410–15. doi:10.1038/ng.3223.

Koren, Sergey, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. 2017. "Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation." *Genome Research* 27 (5). Cold Spring Harbor Laboratory Press: 722–36. doi:10.1101/gr.215087.116.

Korf, Ian. 2004. "Gene Finding in Novel Genomes." *BMC Bioinformatics* 5 (1): 59. doi:10.1186/1471-2105-5-59.

Koskiniemi, Sanna, Song Sun, Otto G. Berg, Dan I. Andersson, and DH Boxer. 2012. "Selection-Driven Gene Loss in Bacteria." Edited by Josep Casadesús. *PLoS Genetics* 8 (6). Public Library of Science: e1002787. doi:10.1371/journal.pgen.1002787.

Koutsovoulos, G, S Kumar, D R Laetsch, L Stevens, J Daub, C Conlon, H Maroon, F Thomas, A A Aboobaker, and M Blaxter. 2016. "No Evidence for Extensive Horizontal Gene Transfer in the Genome of the Tardigrade Hypsibius Dujardini." *Proc Natl Acad Sci U S A* 113 (18): 5053–58. doi:10.1073/pnas.1600338113.

Kranner, Ilse, Richard Beckett, Ayala Hochman, and Thomas H Nash III. 2009. "Desiccation-Tolerance in Lichens: A Review." *The Bryologist*. doi:10.1639/0007-2745-111.4.576.

Krumsiek, J., R. Arnold, and T. Rattei. 2007. "Gepard: A Rapid and Sensitive Tool for Creating Dotplots on Genome Scale." *Bioinformatics* 23 (8). Springer,: 1026–28. doi:10.1093/bioinformatics/btm039.

Kumar, Sujai. 2013. "Next-Generation Nematode Genomes." The University of Edinburgh. https://www.era.lib.ed.ac.uk/handle/1842/7609.

Kumar, Sujai, and Mark L Blaxter. 2010. "Comparing de Novo Assemblers for 454 Transcriptome Data." *BMC Genomics* 11 (January): 571. doi:10.1186/1471-2164-11-571.

Kuske, C R, S M Barns, and J D Busch. 1997. "Diverse Uncultivated Bacterial Groups from Soils of the Arid Southwestern United States That Are Present in Many Geographic Regions." *Applied and Environmental Microbiology* 63 (9). American Society for Microbiology (ASM): 3614–21. http://www.ncbi.nlm.nih.gov/pubmed/9293013.

Kyrpides, Nikos C., Philip Hugenholtz, Jonathan A. Eisen, Tanja Woyke, Markus Göker, Charles T. Parker, Rudolf Amann, et al. 2014. "Genomic Encyclopedia of Bacteria and Archaea: Sequencing a Myriad of Type Strains." *PLoS Biology* 12 (8). Public Library of Science: e1001920. doi:10.1371/journal.pbio.1001920.

Langmead, B, and SL Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 357–59. doi:10.1038/nmeth.1923.

Lawrence, D P, S Kroken, B M Pryor, and A E Arnold. 2011. "Interkingdom Gene Transfer of a Hybrid NPS/PKS from Bacteria to Filamentous Ascomycota." *PLoS One* 6 (11): e28231. doi:10.1371/journal.pone.0028231.

Lawrey, James D., and Paul Diederich. 2003. "Lichenicolous Fungi: Interactions, Evolution, and Biodiversity." *The Bryologist* 106 (1): 80–120. doi:10.1639/0007-2745(2003)106[0080:LFIEAB]2.0.CO;2.

Lee, Eduardo, Gregg a Helt, Justin T Reese, Monica C Munoz-Torres, Chris P Childers, Robert M Buels, Lincoln Stein, Ian H Holmes, Christine G Elsik, and Suzanna E Lewis. 2013. "Web Apollo: A Web-Based Genomic Annotation Editing Platform." *Genome Biology*. doi:10.1186/gb-2013-14-8-r93.

Lee, Hayan, James Gurtowski, Shinjae Yoo, and Shoshana Marcus. 2014. "Error Correction and Assembly Complexity of Single Molecule Sequencing Reads." *bioRxiv*, 1–17. doi:10.1101/006395.

Lee, Hayan, James Gurtowski, Shinjae Yoo, Maria Nattestad, Shoshana Marcus, Sara Goodwin, W. Richard McCombie, and Michael Schatz. 2016. "Third-Generation Sequencing and the Future of Genomics." *bioRxiv*. http://www.biorxiv.org/content/early/2016/04/13/048603.

Leggett, R M, B J Clavijo, L Clissold, M D Clark, and M Caccamo. 2014. "NextClip: An Analysis and Read Preparation Tool for Nextera Long Mate Pair Libraries." *Bioinformatics* 30 (4): 566–68. doi:10.1093/bioinformatics/btt702.

Lek, Monkol, Konrad J Karczewski, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O, James S Ware, et al. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." *bioRxiv* 536 (7616): 30338. doi:10.1101/030338.

Letunic, I., Richard R. Copley, Birgit Pils, Stefan Pinkert, Jörg Schultz, and Peer Bork. 2006. "SMART 5: Domains in the Context of Genomes and Networks." *Nucleic Acids Research* 34 (90001). Oxford University Press: D257–60.

doi:10.1093/nar/gkj079.

Lewis, D H. 1973. "Concepts in Fungal Nutrition and the Origin of Biotrophy." *Biological Reviews of the Cambridge Philosophical Society* 48 (2): 261–78. doi:10.1111/j.1469-185X.1973.tb00982.x.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. "MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph." *Bioinformatics* 31 (10): 1674–76. doi:10.1093/bioinformatics/btv033.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *arXiv*. https://arxiv.org/abs/1303.3997.

Li, Heng, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, and R Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Li, Ruiqiang, Wei Fan, Geng Tian, Hongmei Zhu, Lin He, Jing Cai, Quanfei Huang, et al. 2010. "The Sequence and de Novo Assembly of the Giant Panda Genome." *Nature* 463 (7279): 311–17. doi:10.1038/nature08696.

Li, Zhenyu, Yanxiang Chen, Desheng Mu, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, et al. 2012. "Comparison of the Two Major Classes of Assembly Algorithms: Overlap-Layout-Consensus and de-Bruijn-Graph." *Briefings in Functional Genomics* 11 (1): 25–37. doi:10.1093/bfgp/elr035.

Loewenstein, Yaniv, Domenico Raimondo, Oliver C Redfern, James Watson, Dmitrij Frishman, Michal Linial, Christine Orengo, Janet Thornton, and Anna Tramontano. 2009. "Protein Function Annotation by Homology-Based Inference." *Genome Biology*. doi:10.1186/gb-2009-10-2-207.

Loman, Nicholas J, Joshua Quick, and Jared T Simpson. 2015. "A Complete Bacterial Genome Assembled de Novo Using Only Nanopore Sequencing Data." *bioRxiv*, 1–21.

Lowe, R G, and B J Howlett. 2012. "Indifferent, Affectionate, or Deceitful: Lifestyles and Secretomes of Fungi." *PLoS Pathog* 8 (3): e1002515. doi:10.1371/journal.ppat.1002515.

Lucking, R, B P Hodkinson, and S D Leavitt. 2016. "The 2016 Classification of Lichenized Fungi in the Ascomycota and Basidiomycota - Approaching One Thousand Genera." *Bryologist* 119 (4): 361–416. doi:10.1639/0007-2745-119.4.361.

Lücking, Robert, James D Lawrey, Masoumeh Sikaroodi, Patrick M Gillevet, José Luis Chaves, Harrie J M Sipman, and Frank Bungartz. 2009. "Do Lichens Domesticate Photobionts like Farmers Domesticate Crops? Evidence from a Previously Unrecognized Lineage of Filamentous Cyanobacteria." *American Journal of Botany* 96 (8). Botanical Society of America: 1409–18. doi:10.3732/ajb.0800258.

Luo, Ruibang, B Liu, Yinlong Xie, Z Li, Weihua Huang, and Jianying Yuan. 2012. "SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read de

Novo Assembler." … 1 (1): 18. doi:10.1186/2047-217X-1-18.

Lutsak, Tetiana, Fernando Fernández-Mendoza, Bastian Greshake, Francesco Dal Grande, Ingo Ebersberger, Sieglinde Ott, and Christian Printzen. 2016. "Characterization of Microsatellite Loci in the Lichen-Forming Fungus *Cetraria Aculeata* (Parmeliaceae, Ascomycota)." *Applications in Plant Sciences* 4 (9): 1600047. doi:10.3732/apps.1600047.

MacArthur, D. G., S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, et al. 2012. "A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes." *Science*. doi:10.1126/science.1215040.

Magain, Nicolas, and Emmanuël Sérusiaux. 2015. "Dismantling the Treasured Flagship Lichen Sticta Fuliginosa (Peltigerales) into Four Species in Western Europe." *Mycological Progress* 14 (10): 97. doi:10.1007/s11557-015-1109-0.

Magoc, Tanja, Stephan Pabinger, Stefan Canzar, Xinyue Liu, Qi Su, Daniela Puiu, Luke J. Tallon, and Steven L. Salzberg. 2013. "GAGE-B: An Evaluation of Genome Assemblers for Bacterial Organisms." *Bioinformatics* 29 (14). Oxford University Press: 1718–25. doi:10.1093/bioinformatics/btt273.

Magoč, Tanja, and Steven L Salzberg. 2011. "FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies." *Bioinformatics* 27 (21): 2957–63. doi:10.1093/bioinformatics/btr507.

Malé, Pierre-Jean Jean G., Léa Bardon, Guillaume Besnard, Eric Coissac, Frédéric Delsuc, Julien Engel, Emeline Lhuillier, Caroline Scotti-Saintagne, Alexandra Tinaut, and Jérôme Chave. 2014. "Genome Skimming by Shotgun Sequencing Helps Resolve the Phylogeny of a Pantropical Tree Family." *Molecular Ecology Resources* 14 (5): 966–75. doi:10.1111/1755-0998.12246.

Mapleson, Daniel, Nizar Drou, and David Swarbreck. 2015. "RAMPART: A Workflow Management System for de Novo Genome Assembly." *Bioinformatics* 31 (11). Oxford University Press: 1–2. doi:10.1093/bioinformatics/btv056.

Marbouty, Martial, Lyam Baudry, Axel Cournac, Romain Koszul, Oleg Reva, Don Arthur Cowan, KE Nelson, W Li, KH Williams, and JF Banfield. 2017. "Scaffolding Bacterial Genomes and Probing Host-Virus Interactions in Gut Microbiome by Proximity Ligation (Chromosome Capture) Assay." *Science Advances* 3 (2). BioMed Central: e1602105. doi:10.1126/sciadv.1602105.

Marcotte, C J, and E M Marcotte. 2002. "Predicting Functional Linkages from Gene Fusions with Confidence." *Appl Bioinformatics* 1 (2): 93–100. https://www.ncbi.nlm.nih.gov/pubmed/15130848.

Mardis, Elaine R. 2008. "The Impact of next-Generation Sequencing Technology on Genetics." *Trends in Genetics* 24 (3): 133–41. doi:10.1016/j.tig.2007.12.007.

Margulis, Lynn. 2003. "Microbial Actors in the Evolutionary Drama." *BioScience* 53 (2): 179–80. doi:10.1641/0006-3568(2003)053[0179:MAITED]2.0.CO;2.

Martín-Durán, José M, Joseph F Ryan, Bruno C Vellutini, Kevin Pang, and Andreas Hejnol. 2017. "Increased Taxon Sampling Reveals Thousands of Hidden

Orthologs in Flatworms." *Genome Research* 27 (7). Cold Spring Harbor Laboratory Press: 1263–72. doi:10.1101/gr.216226.116.

Martin, F, A Aerts, D Ahrén, A Brun, E Danchin, F Duchaussoy, J Gibon, et al. 2008. "The Genome of Laccaria Bicolor Provides Insights into Mycorrhizal Symbiosis." *Nature* 452 (7183): 88–92. doi:10.1038/nature06556.

Martin, F, Stéphane Uroz, and David G. Barker. 2017. "Ancestral Alliances: Plant Mutualistic Symbioses with Fungi and Bacteria." *Science* 356 (6340). http://science.sciencemag.org/content/356/6340/eaad4501.

Martin, M. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.journal* 17: 10. doi:10.14806/ej.17.1.200.

Martínez-Alberola, Fernando. 2015. "Genome Characterization of the Symbiotic Microalga Trebouxia Sp. TR9 Isolated from the Lichen Ramalina Farinacea (L.) Ach. by Means of NGS Techniques." University of Valencia.

Martinez, Salette, and Robert P Hausinger. 2015. "Catalytic Mechanisms of Fe(II)- and 2-Oxoglutarate-Dependent Oxygenases." *The Journal of Biological Chemistry* 290 (34). American Society for Biochemistry and Molecular Biology: 20702–11. doi:10.1074/jbc.R115.648691.

Mathe, C., Marie-France Sagot, Thomas Schiex, and Pierre Rouzé. 2002. "Current Methods of Gene Prediction, Their Strengths and Weaknesses." *Nucleic Acids Research* 30 (19). Oxford University Press: 4103–17. doi:10.1093/nar/gkf543.

Mavromatis, Konstantinos, Natalia Ivanova, Kerrie Barry, Harris Shapiro, Eugene Goltsman, Alice C McHardy, Isidore Rigoutsos, et al. 2007. "Use of Simulated Data Sets to Evaluate the Fidelity of Metagenomic Processing Methods." *Nature Methods* 4 (6): 495–500. doi:10.1038/nmeth1043.

McCotter, S W, L C Horianopoulos, and J W Kronstad. 2016. "Regulation of the Fungal Secretome." *Curr Genet* 62 (3): 533–45. doi:10.1007/s00294-016-0578-2.

McCutcheon, John P., and Nancy A. Moran. 2011. "Extreme Genome Reduction in Symbiotic Bacteria." *Nature Reviews Microbiology* 10 (1). Nature Publishing Group: 13. doi:10.1038/nrmicro2670.

McDonald, Tami R., Fred S. Dietrich, and Francxois Lutzoni. 2012. "Multiple Horizontal Gene Transfers of Ammonium Transporters/ammonia Permeases from Prokaryotes to Eukaryotes: Toward a New Functional and Evolutionary Classification." *Molecular Biology and Evolution* 29 (1): 51–60. doi:10.1093/molbev/msr123.

McDonald, Tami R., Ester Gaya, and François Lutzoni. 2013. "Twenty-Five Cultures of Lichenizing Fungi Available for Experimental Studies on Symbiotic Systems." *Symbiosis* 59 (3): 165–71. doi:10.1007/s13199-013-0228-0.

McDonald, Tami R, Olaf Mueller, Fred S Dietrich, and François Lutzoni. 2013. "High-Throughput Genome Sequencing of Lichenizing Fungi to Assess Gene Loss in the Ammonium Transporter/ammonia Permease Gene Family." *BMC Genomics* 14: 225. doi:10.1186/1471-2164-14-225.

McDowell, John M. 2011. "Genomes of Obligate Plant Pathogens Reveal Adaptations for Obligate Parasitism." *Proceedings of the National Academy of Sciences of the United States of America* 108 (22). National Academy of Sciences: 8921–22. doi:10.1073/pnas.1105802108.

McElroy, Kerensa E, Fabio Luciani, and Torsten Thomas. 2012. "GemSIM: General, Error-Model Based Simulator of next-Generation Sequencing Data." *BMC Genomics* 13 (1). BioMed Central Ltd: 74. doi:10.1186/1471-2164-13-74.

Mende, Daniel R., Alison S. Waller, Shinichi Sunagawa, Aino I. Järvelin, Michelle M. Chan, Manimozhiyan Arumugam, Jeroen Raes, and Peer Bork. 2012. "Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data." Edited by John Parkinson. *PLoS ONE* 7 (2). Public Library of Science: e31386. doi:10.1371/journal.pone.0031386.

Merchant, Sabeeha S, Simon E Prochnik, Olivier Vallon, Elizabeth H Harris, J Karpowicz, George B Witman, Astrid Terry, et al. 2010. "The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions." *Science*. doi:10.1126/science.1143609.The.

Millanes, Ana M., Paul Diederich, and Mats Wedin. 2016. "Cyphobasidium Gen. Nov., a New Lichen-Inhabiting Lineage in the Cystobasidiomycetes (Pucciniomycotina, Basidiomycota, Fungi)." *Fungal Biology* 120 (11): 1468–77. doi:10.1016/j.funbio.2015.12.003.

Miller, JR, S Koren, and G Sutton. 2010. "Assembly Algorithms for next-Generation Sequencing Data." *Genomics* 95 (6): 315–27. doi:10.1016/j.ygeno.2010.03.001.

Min, Byoungnam, Igor V. Grigoriev, and In-Geol Choi. 2017. "FunGAP: Fungal Genome Annotation Pipeline Using Evidence-Based Gene Model Evaluation." *Bioinformatics* 34 (June): W435–39. doi:10.1093/bioinformatics/btx353.

Morales-Cruz, A, K C Amrine, B Blanco-Ulate, D P Lawrence, R Travadon, P E Rolshausen, K Baumgartner, and D Cantu. 2015. "Distinctive Expansion of Gene Families Associated with Plant Cell Wall Degradation, Secondary Metabolism, and Nutrient Uptake in the Genomes of Grapevine Trunk Pathogens." *BMC Genomics* 16: 469. doi:10.1186/s12864-015-1624-z.

Morris, J Jeffrey, Richard E Lenski, and Erik R Zinser. 2012. "The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss." *mBio* 3 (2). American Society for Microbiology: e00036-12. doi:10.1128/mBio.00036-12.

Moya, Andrés, Juli Peretó, Rosario Gil, and Amparo Latorre. 2008. "Learning How to Live Together: Genomic Insights into Prokaryote–animal Symbioses." *Nature Reviews Genetics* 9 (3): 218–29. doi:10.1038/nrg2319.

Moya, Patricia, Arántzazu Molins, Fernando Martínez-Alberola, Lucia Muggia, Eva Barreno, M Podar, and YJ Yao. 2017. "Unexpected Associated Microalgal Diversity in the Lichen Ramalina Farinacea Is Uncovered by Pyrosequencing Analyses." Edited by Gabriel Moreno-Hagelsieb. *PLOS ONE* 12 (4). Pearson Educación S.A: e0175091. doi:10.1371/journal.pone.0175091.

Mueller, G M, J P Schmit, P R Leacock, B Buyck, J Cifuentes, D E Desjardin, R E Halling, et al. 2007. "Global Diversity and Distribution of Macrofungi." *Biodiversity and Conservation* 16 (1): 37–48. doi:10.1007/s10531-006-9108-8.

Muller, Cornelius H. 1952. "Plant Succession in Arctic Heath and Tundra in Northern Scandinavia." *Bulletin of the Torrey Botanical Club* 79 (4). Torrey Botanical Society: 296. doi:10.2307/2482004.

Muszewska, Anna, Marta Hoffman-Sommer, Marcin Grynberg, P Nanjappa, and C Phillips. 2011. "LTR Retrotransposons in Fungi." Edited by Rosemary J. Redfield. *PLoS ONE* 6 (12). ASM Press: e29425. doi:10.1371/journal.pone.0029425.

Myers, Eugene W., Granger G. Sutton, Art L. Delcher, Ian M. Dew, Dan P. Fasulo, Michael J. Flanigan, Saul A. Kravitz, et al. 2000. "A Whole-Genome Assembly of Drosophila." *Science* 287 (5461). http://science.sciencemag.org/content/287/5461/2196.full.

Nagarajan, Niranjan, and Mihai Pop. 2013. "Sequence Assembly Demystified." *Nature Reviews Genetics* 14 (3). Nature Publishing Group: 157–67. doi:10.1038/nrg3367.

Nakamura, Kensuke, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, et al. 2011. "Sequence-Specific Error Profile of Illumina Sequencers." *Nucleic Acids Research* 39 (13). Oxford University Press: e90. doi:10.1093/nar/gkr344.

Namiki, Toshiaki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. 2012. "MetaVelvet: An Extension of Velvet Assembler to de Novo Metagenome Assembly from Short Sequence Reads." *Nucleic Acids Research* 40 (20): e155. doi:10.1093/nar/gks678.

Nayfach, Stephen, and Katherine S Pollard. 2016. "Toward Accurate and Quantitative Comparative Metagenomics." *Cell* 166 (5). NIH Public Access: 1103–16. doi:10.1016/j.cell.2016.08.007.

Neale, D B, J L Wegrzyn, K A Stevens, A V Zimin, D Puiu, M W Crepeau, C Cardeno, et al. 2014. "Decoding the Massive Genome of Loblolly Pine Using Haploid DNA and Novel Assembly Strategies." *Genome Biol* 15 (3): R59. doi:10.1186/gb-2014-15-3-r59.

Nelson, David R, Basel Khraiwesh, Weiqi Fu, Saleh Alseekh, Ashish Jaiswal, Amphun Chaiboonchoe, Khaled M Hazzouri, et al. 2017. "The Genome and Phenome of the Green Alga *Chloroidium Sp.* UTEX 3007 Reveal Adaptive Traits for Desert Acclimatization." *eLife* 6 (June). doi:10.7554/eLife.25783.

Newton, Adrian C., Bruce D L Fitt, Simon D. Atkins, Dale R. Walters, and Tim J. Daniell. 2010. "Pathogenesis, Parasitism and Mutualism in the Trophic Space of Microbe-Plant Interactions." *Trends in Microbiology* 18 (8). Elsevier Ltd: 365–73. doi:10.1016/j.tim.2010.06.002.

Nikolenko, Sergey I, Anton I Korobeynikov, and Max a Alekseyev. 2013.

"BayesHammer: Bayesian Clustering for Error Correction in Single-Cell Sequencing." *BMC Genomics* 14 Suppl 1 (Suppl 1). BioMed Central Ltd: S7. doi:10.1186/1471-2164-14-S1-S7.

Nolan, T, Re Hands, and Sa Bustin. 2006. "Nature Protocols: Quantification of mRNA Using Real-Time RT-PCR." *Nature Protocols*.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel Pevzner. 2016. "metaSPAdes: A New Versatile de Novo Metagenomics Assembler," April. http://arxiv.org/abs/1604.03071.

Ochman, H., and N A Moran. 2001. "Genes Lost and Genes Found: Evolution of Bacterial Pathogenesis and Symbiosis." *Science (New York, N.Y.)* 292 (5519): 1096–99. http://www.ncbi.nlm.nih.gov/pubmed/11352062.

Oliver, Bernard M, and John Billingham. 1971. "Project Cyclops: A Design Study of a System for Detecting Extraterrestrial Intelligent Life." *The 1971 NASA/ASEE Summer Facultu Fellowship Program*. NASA, 243. https://www.researchgate.net/publication/234347473_Project_Cyclops_A_Design_Study_of_a_System_for_Detecting_Extraterrestrial_Intelligent_Life.

Olm, Matthew R, Cristina N Butterfield, Alex Copeland, T Christian Boles, Brian C Thomas, and Jillian F Banfield. 2017. "The Source and Evolutionary History of a Microbial Contaminant Identified Through Soil Metagenomic Analysis." *mBio* 8 (1). American Society for Microbiology: e01969-16. doi:10.1128/mBio.01969-16.

Olson, Nathan D., Todd J. Treangen, Christopher M. Hill, Victoria Cepeda-Espinoza, Jay Ghurye, Sergey Koren, and Mihai Pop. 2017. "Metagenomic Assembly through the Lens of Validation: Recent Advances in Assessing and Improving the Quality of Genomes Assembled from Metagenomes." *Briefings in Bioinformatics* 13 (August): 751–54. doi:10.1093/bib/bbx098.

Ondov, B D, N H Bergman, and A M Phillippy. 2011. "Interactive Metagenomic Visualization in a Web Browser." *BMC Bioinformatics* 12: 385. doi:10.1186/1471-2105-12-385.

Pao, S S, I T Paulsen, and M H Saier. 1998. "Major Facilitator Superfamily." *Microbiology and Molecular Biology Reviews : MMBR* 62 (1). American Society for Microbiology: 1–34. http://www.ncbi.nlm.nih.gov/pubmed/9529885.

Park, Chae Haeng, Kyung Mo Kim, Ok-Sun Kim, Gajin Jeong, and Soon Gyu Hong. 2016. "Bacterial Communities in Antarctic Lichens." *Antarctic Science* 28 (6). Cambridge University Press: 455–61. doi:10.1017/S0954102016000286.

Park, S Y, J Choi, J A Kim, M H Jeong, S Kim, Y H Lee, and J S Hur. 2013. "Draft Genome Sequence of Cladonia Macilenta KoLRI003786, a Lichen-Forming Fungus Producing Biruloquinone." *Genome Announc* 1 (5). doi:10.1128/genomeA.00695-13.

Park, S Y, J Choi, J A Kim, N H Yu, S Kim, S Y Kondratyuk, Y H Lee, and J S Hur. 2013. "Draft Genome Sequence of Lichen-Forming Fungus Caloplaca Flavorubescens Strain KoLRI002931." *Genome Announc* 1 (4).

doi:10.1128/genomeA.00678-13.

Park, S Y, J Choi, G W Lee, M H Jeong, J A Kim, S O Oh, Y H Lee, and J S Hur. 2014. "Draft Genome Sequence of Umbilicaria Muehlenbergii KoLRILF000956, a Lichen-Forming Fungus Amenable to Genetic Manipulation." *Genome Announc* 2 (2). doi:10.1128/genomeA.00357-14.

Parra, Genis, Keith Bradnam, and Ian Korf. 2007. "CEGMA: A Pipeline to Accurately Annotate Core Genes in Eukaryotic Genomes." *Bioinformatics* 23 (9): 1061–67. doi:10.1093/bioinformatics/btm071.

Peled, Sapir, Olga Leiderman, Rotem Charar, Gilat Efroni, Yaron Shav-Tal, and Yanay Ofran. 2016. "De-Novo Protein Function Prediction Using DNA Binding and RNA Binding Proteins as a Test Case." *Nature Communications*. doi:10.1038/ncomms13424.

Pellegrin, C, E Morin, F M Martin, and C Veneault-Fourrey. 2015. "Comparative Analysis of Secretomes from Ectomycorrhizal Fungi with an Emphasis on Small-Secreted Proteins." *Frontiers in Microbiology* 6. doi:10.3389/fmicb.2015.01278.

Peng, Y., H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin. 2012. "IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth." *Bioinformatics* 28 (11): 1420–28. doi:10.1093/bioinformatics/bts174.

Phillippy, Adam M, Michael C Schatz, and Mihai Pop. 2008. "Genome Assembly Forensics: Finding the Elusive Mis-Assembly." *Genome Biology* 9 (3). BioMed Central: R55. doi:10.1186/gb-2008-9-3-r55.

Price, A. L., N. C. Jones, and P. A. Pevzner. 2005. "De Novo Identification of Repeat Families in Large Genomes." *Bioinformatics* 21 Suppl 1 (Suppl 1). Oxford University Press: i351-8. doi:10.1093/bioinformatics/bti1018.

Ptitsyn, Andrey, and Leonid L Moroz. 2012. "Computational Workflow for Analysis of Gain and Loss of Genes in Distantly Related Genomes." *BMC Bioinformatics*. doi:10.1186/1471-2105-13-S15-S5.

Quail, MA, M Smith, P Coupland, TD Otto, SR Harris, TR Connor, A Bertoni, HP Swerdlow, and Y Gu. 2012. "A Tale of Three next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers." *BMC Genomics* 13 (1): 341. doi:10.1186/1471-2164-13-341.

Radivojac, Predrag, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, et al. 2013. "A Large-Scale Evaluation of Computational Protein Function Prediction." *Nature Methods*. doi:10.1038/nmeth.2340.

Rawat, Suman R., Minna K. Männistö, Yana Bromberg, and Max M. Häggblom. 2012. "Comparative Genomic and Physiological Analysis Provides Insights into the Role of Acidobacteria in Organic Carbon Utilization in Arctic Tundra Soils." *FEMS Microbiology Ecology* 82 (2): 341–55. doi:10.1111/j.1574-6941.2012.01381.x.

Read, Timothy D., Robert A. Petit, Sandeep J. Joseph, Md. Tauqeer Alam, M. Ryan Weil, Maida Ahmad, Ravila Bhimani, et al. 2013. "Draft Sequencing and Assembly of the Genome of the World's Largest Fish, the Whale Shark: Rhincodon Typus Smith 1828 Timothy," 1–5. doi:10.7287/peerj.preprints.141v2.

Richards, Thomas A, Darren M Soanes, Peter G Foster, Guy Leonard, Christopher R Thornton, and Nicholas J Talbot. 2009. "Phylogenomic Analysis Demonstrates a Pattern of Rare and Ancient Horizontal Gene Transfer between Plants and Fungi." *The Plant Cell* 21 (7): 1897–1911. doi:10.1105/tpc.109.065805.

Richter, Daniel C., Felix Ott, Alexander F. Auch, Ramona Schmid, and Daniel H. Huson. 2008. "MetaSim—A Sequencing Simulator for Genomics and Metagenomics." Edited by Dawn Field. *PLoS ONE* 3 (10). Academic Press: e3373. doi:10.1371/journal.pone.0003373.

Rondon, M R, P R August, A D Bettermann, S F Brady, T H Grossman, M R Liles, K A Loiacono, et al. 2000. "Cloning the Soil Metagenome: A Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms." *Applied and Environmental Microbiology* 66 (6): 2541–47. http://www.ncbi.nlm.nih.gov/pubmed/10831436.

Ross, Michael G, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum, et al. 2013. "Characterizing and Measuring Bias in Sequence Data." *Genome Biol* 14 (5). BioMed Central: R51. doi:10.1186/gb-2013-14-5-r51.

Rubio-Piña, Jorge A, and Omar Zapata-Pérez. 2011. "Isolation of Total RNA from Tissues Rich in Polyphenols and Polysaccharides of Mangrove Plants." *Electronic Journal of Biotechnology* Vol 14 (5).

Sabree, Z L, P H Degnan, and N A Moran. 2010. "Chromosome Stability and Gene Loss in Cockroach Endosymbionts." *Applied and Environmental Microbiology* 76 (12): 4076–79. doi:10.1128/AEM.00291-10.

Salathe, M., Martin Ackermann, and Sebastian Bonhoeffer. 2006. "The Effect of Multifunctionality on the Rate of Evolution in Yeast." *Molecular Biology and Evolution* 23 (4). Oxford University Press: 721–22. doi:10.1093/molbev/msj086.

Sangwan, Naseer, Fangfang Xia, Jack A. Gilbert, A Ramette, JM Tiedje, GW Tyson, J Chapman, et al. 2016. "Recovering Complete and Draft Population Genomes from Metagenome Datasets." *Microbiome* 4 (1): 8. doi:10.1186/s40168-016-0154-5.

Schatz, Michael C., Arthur L. Delcher, and Steven L. Salzberg. 2010. "Assembly of Large Genomes Using Second-Generation Sequencing." *Genome Research* 20 (9). Cold Spring Harbor Laboratory Press: 1165–73. doi:10.1101/gr.101360.109.

Schatz, Michael C, a L Delcher, M Roberts, G Marcçais, M Pop, and J a Yorke. 2012. "GAGE: A Critical Evaluation of Genome Assemblies and Assembly Algorithms Steven L. Salzberg, Adam M. Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen." *Genome Research* 22: 557–67. doi:10.1101/gr.131383.111.

Schirmer, M, R D'Amore, U Z Ijaz, N Hall, and C Quince. 2016. "Illumina Error Profiles: Resolving Fine-Scale Variation in Metagenomic Sequencing Data." *BMC Bioinformatics* 17: 125. doi:10.1186/s12859-016-0976-y.

Schmidt, A, M Bickle, T Beck, and M N Hall. 1997. "The Yeast Phosphatidylinositol Kinase Homolog TOR2 Activates RHO1 and RHO2 via the Exchange Factor ROM2." *Cell* 88 (4): 531–42. doi:10.1016/S0092-8674(00)81893-0.

Schmidt, Hanno, Bastian Greshake, Barbara Feldmeyer, Thomas Hankeln, and Markus Pfenninger. 2013. "Genomic Basis of Ecological Niche Divergence among Cryptic Sister Species of Non-Biting Midges." *BMC Genomics* 14 (1). BMC Genomics: 384. doi:10.1186/1471-2164-14-384.

Schmidt, S M, and R Panstruga. 2011. "Pathogenomics of Fungal Plant Parasites: What Have We Learnt about Pathogenesis?" *Curr Opin Plant Biol* 14 (4): 392–99. doi:10.1016/j.pbi.2011.03.006.

Schmitt, Imke, and H. Thorsten Lumbsch. 2009. "Ancient Horizontal Gene Transfer from Bacteria Enhances Biosynthetic Capabilities of Fungi." *PLoS ONE* 4 (2): 1–8. doi:10.1371/journal.pone.0004437.

Schnell, Jason R., H. Jane Dyson, and Peter E. Wright. 2004. "Structure, Dynamics, and Catalytic Function of Dihydrofolate Reductase." *Annual Review of Biophysics and Biomolecular Structure* 33 (1). Annual Reviews: 119–40. doi:10.1146/annurev.biophys.33.110502.133613.

Schüβler, Arthur, Daniel Schwarzott, and Christopher Walker. 2001. "A New Fungal Phylum, the Glomeromycota: Phylogeny and Evolution." *Mycological Research* 105 (12): 1413–21. doi:10.1017/S0953756201005196.

Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Droege, Ivan Gregor, et al. 2017. "Critical Assessment of Metagenome Interpretation − a Benchmark of Computational Metagenomics Software." *bioRxiv*. http://www.biorxiv.org/content/early/2017/06/12/099127.full.pdf+html.

Seemann, T. 2014. "Prokka: Rapid Prokaryotic Genome Annotation." *Bioinformatics* 30 (14): 2068–69. doi:10.1093/bioinformatics/btu153.

Sharpton, T J, J E Stajich, S D Rounsley, M J Gardner, J R Wortman, V S Jordar, R Maiti, et al. 2009. "Comparative Genomic Analyses of the Human Fungal Pathogens Coccidioides and Their Relatives." *Genome Res* 19 (10): 1722–31. doi:10.1101/gr.087551.108.

Shendure, Jay, and Hanlee Ji. 2008. "Next-Generation DNA Sequencing." *Nature Biotechnology* 26 (10). Nature Publishing Group: 1135–45. doi:10.1038/nbt1486.

Signal, Bethany, Brian S. Gloss, and Marcel E. Dinger. 2016. "Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs." *Trends in Genetics*. doi:10.1016/j.tig.2016.08.004.

Sigurbjörnsdóttir, Margrét Auður, Ólafur S Andrésson, Oddur Vilhelmsson, M. a. Sigurbjornsdottir, O. S. Andresson, and Oddur Vilhelmsson. 2015. "Analysis of the Peltigera Membranacea Metagenome Indicates That Lichen-Associated

Bacteria Are Involved in Phosphate Solubilization." *Microbiology* 161 (Pt 5): 989–96. doi:10.1099/mic.0.000069.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31 (19). Oxford University Press: 3210–12. doi:10.1093/bioinformatics/btv351.

Simpson, Jared T., K. Wong, S. D. Jackman, J. E. Schein, S. J.M. Jones, and I. Birol. 2009. "ABySS: A Parallel Assembler for Short Read Sequence Data." *Genome Research* 19 (6): 1117–23. doi:10.1101/gr.089532.108.

Simpson, Jared T, and Richard Durbin. 2010. "Efficient Construction of an Assembly String Graph Using the FM-Index." *Bioinformatics (Oxford, England)* 26 (12): i367-73. doi:10.1093/bioinformatics/btq217.

Simpson, Jared T, and Richard Durbin. 2012. "Efficient de Novo Assembly of Large Genomes Using Compressed Data Structures." *Genome Research*, 1–8. doi:10.1101/gr.126953.111.Freely.

Slater, G S, and E Birney. 2005. "Automated Generation of Heuristics for Biological Sequence Comparison." *Bmc Bioinformatics* 6. doi:10.1186/1471-2105-6-31.

Smit, AFA, R Hubley, and P Green. 2015. "RepeatMasker Open-4.0." http://www.repeatmasker.org.

Smith, Sally E., and David Read. 2008. *Mycorrhizal Symbiosis*. *Mycorrhizal Symbiosis*. Elsevier. doi:10.1016/B978-012370526-6.50019-2.

Spanu, PD. 2012. "The Genomics of Obligate (and Nonobligate) Biotrophs." *Annual Review of Phytopathology* 50 (1). Annual Reviews: 91–109. doi:10.1146/annurev-phyto-081211-173024.

Spanu, Pietro D, James C Abbott, Joelle Amselem, Timothy A Burgis, Darren M Soanes, Kurt Stüber, E van Themaat, et al. 2010. "Genome Expansion and Gene Loss in Powdery Mildew Fungi Reveal Tradeoffs in Extreme Parasitism." *Science* 330 (6010): 1543–46. doi:10.1126/science.1194573.

Spatafora, J. W., G.-H. Sung, J.-M. Sung, N. L. Hywel-Jones, and J. F. White. 2007. "Phylogenetic Evidence for an Animal Pathogen Origin of Ergot and the Grass Endophytes." *Molecular Ecology* 16 (8). Blackwell Publishing Ltd: 1701–11. doi:10.1111/j.1365-294X.2007.03225.x.

Spribille, Toby, Veera Tuovinen, Philipp Resl, Dan Vanderpool, Heimo Wolinski, M. Catherine Aime, Kevin Schneider, et al. 2016. "Basidiomycete Yeasts in the Cortex of Ascomycete Macrolichens." *Science* 353 (6298). http://science.sciencemag.org/content/353/6298/488.

Stamatakis, A. 2006. "RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models." *Bioinformatics* 22 (21): 2688–90. doi:10.1093/bioinformatics/btl446.

Stanke, Mario, O Schöffmann, B Morgenstern, and S Waack. 2006. "Gene Prediction

in Eukaryotes with a Generalized Hidden Markov Model That Uses Hints from External Sources." *BMC Bioinformatics* 7: 62. doi:10.1186/1471-2105-7-62.

Stanke, Mario, and Stephan Waack. 2003. "Gene Prediction with a Hidden Markov Model and a New Intron Submodel." *Bioinformatics* 19: 215–25. doi:10.1093/bioinformatics/btg1080.

Star, Bastiaan, Alexander J. Nederbragt, Marianne H. S. Hansen, Morten Skage, Gregor D. Gilfillan, Ian R. Bradbury, Christophe Pampoulie, Nils Chr Stenseth, Kjetill S. Jakobsen, and Sissel Jentoft. 2014. "Palindromic Sequence Artifacts Generated during Next Generation Sequencing Library Preparation from Historic and Ancient DNA." Edited by Ludovic Orlando. *PLoS ONE* 9 (3). Public Library of Science: e89676. doi:10.1371/journal.pone.0089676.

Steijger, Tamara, Josep F Abril, Pär G Engström, Felix Kokocinski, Martin Akerman, Tyler Alioto, Giovanna Ambrosini, et al. 2013. "Assessment of Transcript Reconstruction Methods for RNA-Seq." *Nature Methods*. doi:10.1038/nmeth.2714.

Stein, Lincoln. 2001. "Genome Annotation: From Sequence to Biology." *Nature Reviews. Genetics* 2 (7): 493–503. doi:10.1038/35080529.

Strange, Richard N., and Peter R. Scott. 2005. "Plant Disease: A Threat to Global Food Security." *Annual Review of Phytopathology* 43 (1). Annual Reviews : 83–116. doi:10.1146/annurev.phyto.43.113004.133839.

Studer, RA, and M Robinson-Rechavi. 2009. "How Confident Can We Be That Orthologs Are Similar, but Paralogs Differ?" *Trends in Genetics* 25 (5): 210–16. http://dx.doi.org/10.1016/j.tig.2009.03.004.

Ter-Hovhannisyan, V., A. Lomsadze, Y. O. Chernoff, and M. Borodovsky. 2008. "Gene Prediction in Novel Fungal Genomes Using an Ab Initio Algorithm with Unsupervised Training." *Genome Research* 18 (12): 1979–90. doi:10.1101/gr.081612.108.

Testa, A C, J K Hane, S R Ellwood, and R P Oliver. 2015. "CodingQuarry: Highly Accurate Hidden Markov Model Gene Prediction in Fungal Genomes Using RNA-Seq Transcripts." *BMC Genomics* 16: 170. doi:10.1186/s12864-015-1344-4.

Theis, K R, N M Dheilly, J L Klassen, R M Brucker, J F Baines, T C Bosch, J F Cryan, et al. 2016. "Getting the Hologenome Concept Right: An Eco-Evolutionary Framework for Hosts and Their Microbiomes." *mSystems* 1 (2). doi:10.1128/mSystems.00028-16.

Thompson, J D, T J Gibson, and D G Higgins. 2002. "Multiple Sequence Alignment Using ClustalW and ClustalX." *Curr Protoc Bioinformatics* Chapter 2: Unit 2.3. doi:10.1002/0471250953.bi0203s00.

Thüs, Holger, Lucia Muggia, Sergio Pérez-Ortega, Sergio E. Favero-Longo, Suzanne Joneson, Heath O'Brien, Matthew P. Nelsen, et al. 2011. "Revisiting Photobiont Diversity in the Lichen Family Verrucariaceae (Ascomycota)." *European Journal of Phycology* 46 (4). Taylor & Francis Group : 399–415. doi:10.1080/09670262.2011.629788.

Tisserant, Emilie, Mathilde Malbreil, Alan Kuo, Annegret Kohler, Aikaterini Symeonidi, Raffaella Balestrini, Philippe Charron, et al. 2013. "Genome of an Arbuscular Mycorrhizal Fungus Provides Insight into the Oldest Plant Symbiosis." *Proceedings of the National Academy of Sciences of the United States of America* 110 (50). National Academy of Sciences: 20117–22. doi:10.1073/pnas.1313452110.

Tollot, Marie, Joanne Wong Sak Hoi, Diederik Van Tuinen, Christine Arnould, Odile Chatagnier, Bernard Dumas, Vivienne Gianinazzi-Pearson, and Pascale M. A. Seddas. 2009. "An STE12 Gene Identified in the Mycorrhizal Fungus Glomus Intraradices Restores Infectivity of a Hemibiotrophic Plant Pathogen." *New Phytologist* 181 (3): 693–707. doi:10.1111/j.1469-8137.2008.02696.x.

Trapnell, Cole, L Pachter, and S L Salzberg. 2009. "TopHat: Discovering Splice Junctions with RNA-Seq." *Bioinformatics* 25 (9): 1105–11. doi:10.1093/bioinformatics/btp120.

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation." *Nature Biotechnology* 28 (5): 511–15. doi:10.1038/nbt.1621.

Treangen, T J, D D Sommer, F E Angly, S Koren, and M Pop. 2011. "Next Generation Sequence Assembly with AMOS." *Curr Protoc Bioinformatics* Chapter 11: Unit 11.8. doi:10.1002/0471250953.bi1108s33.

Treseder, K K, and J T Lennon. 2015. "Fungal Traits That Drive Ecosystem Dynamics on Land." *Microbiol Mol Biol Rev* 79 (2): 243–62. doi:10.1128/MMBR.00001-15.

Tsai, Yu-Chih, Sean Conlan, Clayton Deming, NISC Comparative Sequencing NISC Comparative Sequencing Program, Julia A Segre, Heidi H Kong, Jonas Korlach, and Julia Oh. 2016. "Resolving the Complexity of Human Skin Metagenomes Using Single-Molecule Sequencing." *mBio* 7 (1). American Society for Microbiology: e01948-15. doi:10.1128/mBio.01948-15.

Tully, Benjamin J, Rohan Sachdeva, Elaina D Graham, and John F Heidelberg. 2016. "290 Metagenome-Assembled Genomes from the Mediterranean Sea: Ongoing Effort to Generate Genomes from the Tara Oceans Dataset." *bioRxiv*. doi:10.1101/069484.

Tuskan, G A, S Difazio, S Jansson, J Bohlmann, I Grigoriev, U Hellsten, N Putnam, et al. 2006. "The Genome of Black Cottonwood, Populus Trichocarpa (Torr. & Gray)." *Science* 313 (5793): 1596–1604. doi:10.1126/science.1128691.

van der Burgt, Ate, Edouard Severing, Jérôme Collemare, and Pierre Jgm de Wit. 2014. "Automated Alignment-Based Curation of Gene Models in Filamentous Fungi." *BMC Bioinformatics*. doi:10.1186/1471-2105-15-19.

Veeckman, Elisabeth, Tom Ruttink, and Klaas Vandepoele. 2016. "Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences." *The Plant*

Cell. doi:10.1105/tpc.16.00349.

Vollmers, John, Sandra Wiegand, and Anne Kristin Kaster. 2017. "Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters!" *PLoS ONE*. doi:10.1371/journal.pone.0169662.

Walker, B J, T Abeel, T Shea, M Priest, A Abouelliel, S Sakthikumar, C A Cuomo, et al. 2014. "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement." *PLoS One* 9 (11): e112963. doi:10.1371/journal.pone.0112963.

Wang, YY Y, Bin Liu, XY Y L Zhang, QM M Zhou, T Zhang, H Li, Y F Yu, et al. 2014. "Genome Characteristics Reveal the Impact of Lichenization on Lichen-Forming Fungus Endocarpon Pusillum Hedwig (Verrucariales, Ascomycota)." *BMC Genomics* 15: 1–18. doi:10.1186/1471-2164-15-34.

Warburton, P. E., Joti Giordano, Fanny Cheung, Yefgeniy Gelfand, and Gary Benson. 2004. "Inverted Repeat Structure of the Human Genome: The X-Chromosome Contains a Preponderance of Large, Highly Homologous Inverted Repeats That Contain Testes Genes." *Genome Research* 14 (10a): 1861–69. doi:10.1101/gr.2542904.

Ward, Naomi L., Jean F. Challacombe, Peter H. Janssen, Bernard Henrissat, Pedro M. Coutinho, Martin Wu, Gary Xie, et al. 2009. "Three Genomes from the Phylum Acidobacteria Provide Insight into the Lifestyles of These Microorganisms in Soils." *Appl Environ Microbiol* 75 (7): 2046–56. doi:10.1128/AEM.02294-08.

Weber, Tilmann, Kai Blin, Srikanth Duddela, Daniel Krug, Hyun Uk Kim, Robert Bruccoleri, Sang Yup Lee, et al. 2015. "antiSMASH 3.0-a Comprehensive Resource for the Genome Mining of Biosynthetic Gene Clusters." *Nucleic Acids Research* 43 (W1). Oxford University Press: W237-43. doi:10.1093/nar/gkv437.

Weitemier, Kevin, Shannon C.K. Straub, Mark Fishbein, and Aaron Liston. 2015. "Intragenomic Polymorphisms among High-Copy Loci: A Genus-Wide Study of Nuclear Ribosomal DNA in *Asclepias* (Apocynaceae)." *PeerJ* 3 (January). PeerJ Inc.: e718. doi:10.7717/peerj.718.

Wences, Alejandro Hernandez, and Michael C Schatz. 2015. "Metassembler: Merging and Optimizing de Novo Genome Assemblies." *bioRxiv*.

Williams, L E, and J J Wernegreen. 2015. "Genome Evolution in an Ancient Bacteria-Ant Symbiosis: Parallel Gene Loss among Blochmannia Spanning the Origin of the Ant Tribe Camponotini." *PeerJ* 3: e881. doi:10.7717/peerj.881.

Wolf, Y I, and E V Koonin. 2013. "Genome Reduction as the Dominant Mode of Evolution." *Bioessays* 35 (9): 829–37. doi:10.1002/bies.201300037.

Wyman, S. K., R. K. Jansen, and J. L. Boore. 2004. "Automatic Annotation of Organellar Genomes with DOGMA." *Bioinformatics* 20 (17): 3252–55. doi:10.1093/bioinformatics/bth352.

Xu, Zhenyu, Wu Wei, Julien Gagneur, Fabiana Perocchi, Sandra Clauder-Münster, Jurgi Camblong, Elisa Guffanti, Françoise Stutz, Wolfgang Huber, and Lars M

Steinmetz. 2009. “Bidirectional Promoters Generate Pervasive Transcription in Yeast.” *Nature* 457 (7232). NIH Public Access: 1033–37. doi:10.1038/nature07728.

Yandell, Mark, and Daniel Ence. 2012. “A Beginner’s Guide to Eukaryotic Genome Annotation.” *Nature Reviews Genetics* 13 (5). Nature Publishing Group: 329–42. doi:10.1038/nrg3174.

Ye, Chengxi, Christopher M Hill, Shigang Wu, Jue Ruan, and Zhanshan Sam Ma. 2016. “DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies.” *Scientific Reports* 6 (August). Nature Publishing Group: 31900. doi:10.1038/srep31900.

Zerbino, Daniel R, and Ewan Birney. 2008. “Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs.” *Genome Research* 18 (5): 821–29. doi:10.1101/gr.074492.107.

Zhang, Qingpeng, Jason Pell, Rosangela Canino-Koning, Adina Chuang Howe, and C. Titus Brown. 2014. “These Are Not the K-Mers You Are Looking For: Efficient Online K-Mer Counting Using a Probabilistic Data Structure.” Edited by Dongxiao Zhu. *PLoS ONE* 9 (7). Public Library of Science: e101271. doi:10.1371/journal.pone.0101271.

Zhao, Yi, Liang Tang, Zhe Li, Jinpu Jin, Jingchu Luo, and Ge Gao. 2015. “Identification and Analysis of Unitary Loss of Long-Established Protein-Coding Genes in Poaceae Shows Evidences for Biased Gene Loss and Putatively Functional Transcription of Relics.” *BMC Evolutionary Biology*. doi:10.1186/s12862-015-0345-x.

Zuccaro, Alga, Urs Lahrmann, and Gregor Langen. 2014. “Broad Compatibility in Fungal Root Symbioses.” *Current Opinion in Plant Biology* 20 (August): 135–45. doi:10.1016/j.pbi.2014.05.013.

A. Appendix

Tables

Table A-1: Eukaryotic species represented in the *DIAMOND* database we used for our subsequent taxonomic assignments with *MEGAN*.

Kingdom	Phylum	Class	Species
-	-	Bangiophyceae	*Cyanidioschyzon merolae strain 10D*
-	-	Glaucocystophyceae	*Cyanophora paradoxa*
Fungi	Ascomycota	Arthoniomycetes	*Arthonia rubrocincta*
Fungi	Ascomycaota	Dothideomycetes	*Acidomyces richmondensis*
Fungi	Ascomycota	Dothideomycetes	*Alternaria brassicicola*
Fungi	Ascomycota	Dothideomycetes	*Aplosporella prunicola CBS 121167*
Fungi	Ascomycota	Dothideomycetes	*Aulographum hederae*
Fungi	Ascomycota	Dothideomycetes	*Aureobasidium melanogenum CBS 110374*
Fungi	Ascomycota	Dothideomycetes	*Aureobasidium namibiae CBS 147.97*
Fungi	Ascomycota	Dothideomycetes	*Aureobasidium pullulans EXF-150*
Fungi	Ascomycota	Dothideomycetes	*Aureobasidium subglaciale EXF-2481*
Fungi	Ascomycota	Dothideomycetes	*Baudoinia panamericana UAMH 10762*
Fungi	Ascomycota	Dothideomycetes	*Bipolaris maydis ATCC 48331*
Fungi	Ascomycota	Dothideomycetes	*Bipolaris maydis C5*
Fungi	Ascomycota	Dothideomycetes	*Bipolaris oryzae ATCC 44560*
Fungi	Ascomycota	Dothideomycetes	*Bipolaris sorokiniana ND90Pr*
Fungi	Ascomycota	Dothideomycetes	*Bipolaris victoriae FI3*
Fungi	Ascomycota	Dothideomycetes	*Bipolaris zeicola 26-R-13*
Fungi	Ascomycota	Dothideomycetes	*Botryosphaeria dothidea*
Fungi	Ascomycota	Dothideomycetes	*Cenococcum geophilum 1.58*
Fungi	Ascomycota	Dothideomycetes	*Cercospora zeae-maydis*
Fungi	Ascomycota	Dothideomycetes	*Cucurbitaria berberidis CBS 394.84*
Fungi	Ascomycota	Dothideomycetes	*Curvularia lunata m118*
Fungi	Ascomycota	Dothideomycetes	*Didymella exigua CBS 183.55*
Fungi	Ascomycota	Dothideomycetes	*Dissoconium aciculare*
Fungi	Ascomycota	Dothideomycetes	*Dothidotthia symphoricarpi*
Fungi	Ascomycota	Dothideomycetes	*Dothistroma septosporum NZE10*
Fungi	Ascomycota	Dothideomycetes	*Hysterium pulicare*
Fungi	Ascomycota	Dothideomycetes	*Lentithecium fluviatile*
Fungi	Ascomycota	Dothideomycetes	*Lepidopterella palustris*

Fungi	Ascomycota	Dothideomycetes	*Leptosphaeria maculans*
Fungi	Ascomycota	Dothideomycetes	*Lophiostoma macrostomum*
Fungi	Ascomycota	Dothideomycetes	*Macrophomina phaseolina MS6*
Fungi	Ascomycota	Dothideomycetes	*Melanomma pulvis-pyrius*
Fungi	Ascomycota	Dothideomycetes	*Myriangium duriaei CBS 260.36*
Fungi	Ascomycota	Dothideomycetes	*Neofusicoccum parvum UCRNP2*
Fungi	Ascomycota	Dothideomycetes	*Parastagonospora nodorum SN15*
Fungi	Ascomycota	Dothideomycetes	*Passalora fulva*
Fungi	Ascomycota	Dothideomycetes	*Patellaria atrata*
Fungi	Ascomycota	Dothideomycetes	*Piedraia hortae*
Fungi	Ascomycota	Dothideomycetes	*Pleomassaria siparia*
Fungi	Ascomycota	Dothideomycetes	*Polychaeton citri*
Fungi	Ascomycota	Dothideomycetes	*Pseudocercospora fijiensis*
Fungi	Ascomycota	Dothideomycetes	*Pyrenophora teres f. teres*
Fungi	Ascomycota	Dothideomycetes	*Pyrenophora tritici-repentis*
Fungi	Ascomycota	Dothideomycetes	*Rhytidhysteron rufulum*
Fungi	Ascomycota	Dothideomycetes	*Setosphaeria turcica Et28A*
Fungi	Ascomycota	Dothideomycetes	*Sphaerulina musiva SO2202*
Fungi	Ascomycota	Dothideomycetes	*Sphaerulina populicola*
Fungi	Ascomycota	Dothideomycetes	*Sporormia fimetaria*
Fungi	Ascomycota	Dothideomycetes	*Trypethelium eluteriae*
Fungi	Ascomycota	Dothideomycetes	*Zasmidium cellare ATCC 36951*
Fungi	Ascomycota	Dothideomycetes	*Zopfia rhizophila*
Fungi	Ascomycota	Dothideomycetes	*Zymoseptoria tritici*
Fungi	Ascomycota	Eurotiomycetes	*Arthroderma otae CBS 113480*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus acidus*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus aculeatus ATCC 16872*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus brasiliensis*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus carbonarius ITEM 5010*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus clavatus NRRL 1*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus fischeri NRRL 181*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus flavus NRRL3357*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus fumigatus var. fumigatus*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus glaucus*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus kawachii*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus nidulans*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus niger ATCC 1015*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus oryzae RIB40*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus ruber*

Fungi	Ascomycota	Eurotiomycetes	*Aspergillus sydowii*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus terreus NIH2624*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus tubingensis*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus versicolor*
Fungi	Ascomycota	Eurotiomycetes	*Aspergillus wentii*
Fungi	Ascomycota	Eurotiomycetes	*Coccidioides immitis RS*
Fungi	Ascomycota	Eurotiomycetes	*Coccidioides posadasii C735 delta SOWgp*
Fungi	Ascomycota	Eurotiomycetes	*Endocarpon pallidulum*
Fungi	Ascomycota	Eurotiomycetes	*Endocarpon pusillum Z07020*
Fungi	Ascomycota	Eurotiomycetes	*Gymnascella aurantiaca*
Fungi	Ascomycota	Eurotiomycetes	*Gymnascella citrina*
Fungi	Ascomycota	Eurotiomycetes	*Histoplasma capsulatum NAm1*
Fungi	Ascomycota	Eurotiomycetes	*Monascus purpureus*
Fungi	Ascomycota	Eurotiomycetes	*Paracoccidioides brasiliensis Pb03*
Fungi	Ascomycota	Eurotiomycetes	*Penicilliopsis zonata*
Fungi	Ascomycota	Eurotiomycetes	*Penicillium bilaiae ATCC 20851*
Fungi	Ascomycota	Eurotiomycetes	*Penicillium brevicompactum AgRF18*
Fungi	Ascomycota	Eurotiomycetes	*Penicillium canescens ATCC 10419*
Fungi	Ascomycota	Eurotiomycetes	*Penicillium digitatum PHI26*
Fungi	Ascomycota	Eurotiomycetes	*Penicillium expansum ATCC 24692*
Fungi	Ascomycota	Eurotiomycetes	*Penicillium fellutanum ATCC 48694*
Fungi	Ascomycota	Eurotiomycetes	*Penicillium glabrum DAOM 239074*
Fungi	Ascomycota	Eurotiomycetes	*Penicillium janthinellum ATCC 10455*
Fungi	Ascomycota	Eurotiomycetes	*Penicillium lanosocoeruleum ATCC 48919*
Fungi	Ascomycota	Eurotiomycetes	*Penicillium oxalicum 114-2*
Fungi	Ascomycota	Eurotiomycetes	*Penicillium raistrickii ATCC 10490*
Fungi	Ascomycota	Eurotiomycetes	*Penicillium rubens Wisconsin 54-1255*
Fungi	Ascomycota	Eurotiomycetes	*Talaromyces aculeatus ATCC 10409*
Fungi	Ascomycota	Eurotiomycetes	*Talaromyces marneffei ATCC 18224*
Fungi	Ascomycota	Eurotiomycetes	*Talaromyces stipitatus ATCC 10500*
Fungi	Ascomycota	Eurotiomycetes	*Thermoascus aurantiacus*
Fungi	Ascomycota	Eurotiomycetes	*Trichophyton benhamiae CBS 112371*
Fungi	Ascomycota	Eurotiomycetes	*Trichophyton rubrum CBS 118892*
Fungi	Ascomycota	Eurotiomycetes	*Trichophyton verrucosum HKI 0517*
Fungi	Ascomycota	Eurotiomycetes	*Uncinocarpus reesii 1704*
Fungi	Ascomycota	Lecanoromycetes	*Acarospora sp.*
Fungi	Ascomycota	Lecanoromycetes	*Cladonia grayi*
Fungi	Ascomycota	Lecanoromycetes	*Cladonia macilenta KoLRI003786*
Fungi	Ascomycota	Lecanoromycetes	*Cladonia metacorallifera KoLRI002260*

Fungi	Ascomycota	Lecanoromycetes	*Graphis scripta*
Fungi	Ascomycota	Lecanoromycetes	*Gyalolechia flavorubescens KoLRI002931*
Fungi	Ascomycota	Lecanoromycetes	*Xanthoria parietina 46-1-SA22*
Fungi	Ascomycota	Leotiomycetes	*Botrytis cinerea*
Fungi	Ascomycota	Pezizomycetes	*Tuber melanosporum*
Fungi	Ascomycota	Saccharomycetes	*Candida glabrata CBS 138*
Fungi	Ascomycota	Saccharomycetes	*Saccharomyces cerevisiae*
Fungi	Ascomycota	Schizosaccharomycetes	*Schizosaccharomyces pombe*
Fungi	Ascomycota	Sordariomycetes	*Fusarium oxysporum CL57*
Fungi	Ascomycota	Sordariomycetes	*Neurospora crassa*
Fungi	Basidiomycota	Agaricomycetes	*Agaricus bisporus var. bisporus H97*
Fungi	Basidiomycota	Pucciniomycetes	*Puccinia graminis*
Fungi	Basidiomycota	Ustilaginomycetes	*Ustilago maydis*
Fungi	Microsporidia	-	*Nosema ceranae*
Fungi	Mucoromycota	-	*Rhizopus oryzae*
Metazoa	Arthropoda	Insecta	*Drosophila melanogaster*
Metazoa	Chordata	Actinopteri	*Danio rerio*
Metazoa	Chordata	Aves	*Gallus gallus*
Metazoa	Chordata	Mammalia	*Homo sapiens*
Metazoa	Chordata	Mammalia	*Mus musculus*
Metazoa	Cnidaria	Anthozoa	*Nematostella vectensis*
Metazoa	Mollusca	Gastropoda	*Lottia gigantea*
Metazoa	Nematoda	Chromadorea	*Caenorhabditis elegans*
Viridiplantae	Chlorophyta	Chlorophyceae	*Chlamydomonas reinhardtii CC3269*
Viridiplantae	Chlorophyta	Chlorophyceae	*Volvox carteri f. nagariensis*
Viridiplantae	Chlorophyta	Mamiellophyceae	*Bathycoccus prasinos RCC1105*
Viridiplantae	Chlorophyta	Mamiellophyceae	*Micromonas pusilla*
Viridiplantae	Chlorophyta	Mamiellophyceae	*Ostreococcus 'lucimarinus'*
Viridiplantae	Chlorophyta	Trebouxiophyceae	*Auxenochlorella protothecoides sp 0710*
Viridiplantae	Chlorophyta	Trebouxiophyceae	*Chlorella sp.*
Viridiplantae	Chlorophyta	Trebouxiophyceae	*Coccomyxa sp.*
Viridiplantae	Streptophyta	-	*Amborella trichopoda*
Viridiplantae	Streptophyta	-	*Arabidopsis thaliana*
Viridiplantae	Streptophyta	-	*Malus domestica*
Viridiplantae	Streptophyta	-	*Medicago truncatula*
Viridiplantae	Streptophyta	-	*Populus trichocarpa*
Viridiplantae	Streptophyta	Bryopsida	*Physcomitrella patens*
Viridiplantae	Streptophyta	Liliopsida	*Oryza sativa*
Viridiplantae	Streptophyta	Liliopsida	*Sorghum bicolor*

Table A-2: Coverage ratios for the eukaryotic nuclear and organellar genomes, as well as for the 5 largest bacterial scaffolds. All values normalized to the nuclear genome of *Trebouxia sp*.

Library	pacbio	readpairs	matepairs	pool1	pool3	pool2	pool4	pool5	pool6	average
Trebouxia nuclear	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1,00	1.00	1.00
Trebouxia cp	23.79	8.00	8.68	21.64	14.28	18.83	19.85	16.72	13.63	16.47
Trebouxia mt	21.32	9.05	4.21	25.63	16.74	21.46	19.16	14.25	17.36	16.48
Lasallia nuclear	13.82	11.74	16.99	19.66	19.48	16.01	24.11	27.96	29.69	18.72
Lasallia mt	329.86	229.01	239.63	261.37	195.99	228.06	435.73	379.37	322.75	287.38
Bacteria 70	0.68	2.22	0.58	0.82	8.37	3.09	4.03	14.74	15.54	4.32
Bacteria 233	2.01	0.39	0.91	0.89	0.73	0.86	0.75	0.89	1.43	0.93
Bacteria 16	2.46	0.43	1.00	0.50	0.72	0.56	0.57	0.78	0.68	0.88
Bacteria 111	0.09	3.55	0.42	1.94	5.12	1.31	2.78	2.06	1.34	2.16
Bacteria 35	3.53	0.34	1.35	1.46	0.94	1.30	1.11	1.38	2.05	1.43
Bacteria 81	2.12	0.53	0.82	0.88	0.87	0.92	0.85	0.85	1.78	0.98

Table A-3: Top 20 genera that are found in the nine sequencing libraries. Normalized read counts were summed up over all libraries.

Rank	Genus	Normalized Count
1	unclassified Acidobacteriaceae	81168
2	*Granulicella*	57066
3	*Terriglobus*	25908
4	*Acidobacterium*	18842
5	*Singulisphaera*	15773
6	*Sphingomonas*	15333
7	*Methylobacterium*	12918
8	*Chthonomonas*	12185
9	*Bradyrhizobium*	10250
10	*Burkholderia*	9787
11	*Roseomonas*	9196
12	*Edaphobacter*	7155
13	*Candidatus Solibacter*	7131
14	*Acidiphilium*	6957
15	*Streptomyces*	6473
16	*Belnapia*	5671
17	*Granulibacter*	5018
18	*Mycobacterium*	4612
19	unclassified Gemmatimonadetes	4451
20	*Fimbriimonas*	4058

Table A-4: *L. pustulata* genes predicted by *MAKER2*, which the *Rosetta Stone* method identified to be a gene fusion. Comparisons for the *Rosetta Stone* method were made against the genes of *Cladonia grayi*. Each row represents one possible gene fusion that involves two *C. grayi* genes. *L. pustulata* genes can appear more than once, if the fusion involves more than two *C. grayi* genes or if multiple *C. grayi* genes would match the same fusion. The start and end positions of the *C. grayi* genes within the fused *L. pustulata* gene are given.

Fused gene	Fusion partner #1	Start position in fused gene	End position in fused gene	Fusion partner #2	Start position in fused gene	End position in fused gene
scaffold7-gene-9.22-mRNA-1	CLAGR_008330-RA	149	311	CLAGR_002929-RA	417	518
scaffold3-gene-14.114-mRNA-1	CLAGR_010481-RA	28	432	CLAGR_010480-RA	434	874
scaffold5-gene-15.58-mRNA-1	CLAGR_001430-RA	1	754	CLAGR_001442-RA	784	997
scaffold6-gene-2.33-mRNA-1	CLAGR_000328-RA	41	124	CLAGR_000327-RA	171	293
scaffold9-gene-3.93-mRNA-1	CLAGR_004409-RA	103	196	CLAGR_002647-RA	238	331
scaffold9-gene-3.93-mRNA-1	CLAGR_004409-RA	103	196	CLAGR_004410-RA	242	332
scaffold2-gene-4.93-mRNA-1	CLAGR_008510-RA	1	391	CLAGR_002656-RA	562	786
scaffold13-gene-0.73-mRNA-1	CLAGR_005526-RA	1	65	CLAGR_005525-RA	101	889
scaffold12-gene-5.42-mRNA-1	CLAGR_003178-RA	7	1735	CLAGR_007860-RA	2485	2647
scaffold12-gene-5.42-mRNA-1	CLAGR_009968-RA	1039	2462	CLAGR_007860-RA	2485	2647
scaffold2-gene-21.76-mRNA-1	CLAGR_001206-RA	3	143	CLAGR_007281-RA	203	352
scaffold3-gene-4.74-mRNA-1	CLAGR_004073-RA	9	519	CLAGR_004079-RA	539	801
scaffold3-gene-4.74-mRNA-1	CLAGR_004073-RA	9	519	CLAGR_008547-RA	546	801
scaffold8-gene-8.3-mRNA-1	CLAGR_005513-RA	7	389	CLAGR_005514-RA	413	701
scaffold16-gene-2.63-mRNA-1	CLAGR_003105-RA	33	819	CLAGR_003103-RA	901	1136
scaffold9-gene-10.88-mRNA-1	CLAGR_009372-RA	1	122	CLAGR_009371-RA	180	414
scaffold11-gene-6.140-mRNA-1	CLAGR_001276-RA	5	1978	CLAGR_001275-RA	2010	2424
scaffold11-gene-6.140-mRNA-1	CLAGR_001275-RA	2010	2424	CLAGR_001291-RA	2464	3272
scaffold11-gene-6.140-mRNA-1	CLAGR_001276-RA	5	1978	CLAGR_001291-RA	2464	3272
scaffold1-gene-7.64-mRNA-1	CLAGR_004634-RA	10	444	CLAGR_004633-RA	837	923
scaffold1-gene-21.31-mRNA-1	CLAGR_007667-RA	19	217	CLAGR_007668-RA	252	384
scaffold5-gene-8.169-mRNA-1	CLAGR_003606-RA	1	269	CLAGR_009144-RA	308	411
scaffold5-gene-8.17-mRNA-1	CLAGR_003627-RA	152	233	CLAGR_003626-RA	336	736
scaffold5-gene-8.17-mRNA-1	CLAGR_003628-RA	38	143	CLAGR_003626-RA	336	736
scaffold5-gene-8.17-mRNA-1	CLAGR_003628-RA	38	143	CLAGR_003627-RA	152	233
scaffold3-gene-2.31-mRNA-1	CLAGR_000044-RB	594	995	CLAGR_000044-RA	1078	1409
scaffold3-gene-2.31-mRNA-1	CLAGR_000226-RA	68	593	CLAGR_000044-RA	1078	1409
scaffold3-gene-2.31-mRNA-1	CLAGR_010021-RA	207	303	CLAGR_000044-RA	1078	1409
scaffold3-gene-2.31-mRNA-1	CLAGR_000226-RA	68	593	CLAGR_000044-RB	594	995

scaffold3-gene-2.31-mRNA-1	CLAGR_000226-RA	68	593	CLAGR_004534-RA	1087	1401
scaffold3-gene-2.31-mRNA-1	CLAGR_010021-RA	207	303	CLAGR_000044-RB	594	995
scaffold3-gene-2.31-mRNA-1	CLAGR_010021-RA	207	303	CLAGR_004534-RA	1087	1401
scaffold2-gene-16.49-mRNA-1	CLAGR_005101-RA	1	306	CLAGR_005096-RA	340	662
scaffold2-gene-19.33-mRNA-1	CLAGR_000840-RA	16	664	CLAGR_005145-RA	1117	1423
scaffold2-gene-19.33-mRNA-1	CLAGR_004536-RA	16	1044	CLAGR_005145-RA	1117	1423
scaffold2-gene-19.33-mRNA-1	CLAGR_007824-RA	1	1040	CLAGR_005145-RA	1117	1423
scaffold2-gene-19.33-mRNA-1	CLAGR_010299-RA	13	1044	CLAGR_005145-RA	1117	1423
scaffold8-gene-11.89-mRNA-1	CLAGR_007693-RA	39	143	CLAGR_007719-RA	183	413
scaffold6-gene-3.105-mRNA-1	CLAGR_000109-RA	3	122	CLAGR_000110-RA	168	415
scaffold16-gene-0.69-mRNA-1	CLAGR_000225-RA	19	627	CLAGR_004220-RA	652	889
scaffold16-gene-0.69-mRNA-1	CLAGR_004218-RA	49	348	CLAGR_004220-RA	652	889
scaffold14-gene-6.12-mRNA-1	CLAGR_001703-RA	14	345	CLAGR_004625-RA	643	706
scaffold14-gene-6.12-mRNA-1	CLAGR_003813-RA	41	346	CLAGR_004625-RA	643	706
scaffold14-gene-6.12-mRNA-1	CLAGR_008733-RA	14	344	CLAGR_004625-RA	643	706
scaffold10-gene-16.114-mRNA-1	CLAGR_006517-RA	1	275	CLAGR_006516-RA	951	1365
scaffold4-gene-16.80-mRNA-1	CLAGR_010656-RA	5	526	CLAGR_010655-RA	566	867
scaffold4-gene-13.20-mRNA-1	CLAGR_000958-RA	1	979	CLAGR_000899-RA	980	1123
scaffold4-gene-13.20-mRNA-1	CLAGR_000958-RA	1	979	CLAGR_000957-RA	984	1121
scaffold4-gene-13.20-mRNA-1	CLAGR_000958-RA	1	979	CLAGR_000998-RA	985	1121
scaffold18-gene-0.4-mRNA-1	CLAGR_005033-RA	253	335	CLAGR_005032-RA	994	1250
scaffold18-gene-0.41-mRNA-1	CLAGR_006334-RA	32	382	CLAGR_006333-RA	387	577
scaffold3-gene-13.59-mRNA-1	CLAGR_005946-RA	14	434	CLAGR_005945-RA	519	590
scaffold3-gene-13.63-mRNA-1	CLAGR_010502-RA	162	388	CLAGR_010503-RA	431	661
scaffold5-gene-0.52-mRNA-1	CLAGR_001642-RA	10	86	CLAGR_004993-RA	224	923
scaffold1-gene-11.17-mRNA-1	CLAGR_010285-RA	32	299	CLAGR_010975-RA	365	590
scaffold10-gene-10.11-mRNA-1	CLAGR_003917-RA	11	150	CLAGR_009297-RA	212	311
scaffold10-gene-10.11-mRNA-1	CLAGR_011204-RA	18	151	CLAGR_009297-RA	212	311
scaffold10-gene-11.9-mRNA-1	CLAGR_005271-RA	21	407	CLAGR_003198-RA	410	1024
scaffold10-gene-6.156-mRNA-1	CLAGR_008422-RA	59	268	CLAGR_009701-RA	356	409
scaffold2-gene-1.23-mRNA-1	CLAGR_006486-RA	5	201	CLAGR_006489-RA	202	562
scaffold1-gene-26.114-mRNA-1	CLAGR_006320-RA	3	224	CLAGR_009548-RA	254	780
scaffold12-gene-17.53-mRNA-1	CLAGR_002271-RA	373	778	CLAGR_004362-RA	998	1757
scaffold12-gene-17.53-mRNA-1	CLAGR_004807-RA	17	308	CLAGR_004362-RA	998	1757
scaffold12-gene-	CLAGR_010105-	373	994	CLAGR_004362-	998	1757

17.53-mRNA-1	RA			RA		
scaffold12-gene-17.53-mRNA-1	CLAGR_002271-RA	373	778	CLAGR_007633-RA	1117	1645
scaffold12-gene-17.53-mRNA-1	CLAGR_004807-RA	17	308	CLAGR_007633-RA	1117	1645
scaffold12-gene-17.53-mRNA-1	CLAGR_010105-RA	373	994	CLAGR_007633-RA	1117	1645
scaffold12-gene-17.53-mRNA-1	CLAGR_011250-RA	56	1020	CLAGR_007633-RA	1117	1645
scaffold12-gene-17.53-mRNA-1	CLAGR_011250-RA	56	1020	CLAGR_011257-RA	1292	1401
scaffold3-gene-1.132-mRNA-1	CLAGR_008977-RA	158	223	CLAGR_008978-RA	373	776
scaffold14-gene-5.82-mRNA-1	CLAGR_008982-RA	5	357	CLAGR_008139-RA	414	663
scaffold3-gene-17.72-mRNA-1	CLAGR_010226-RA	1	279	CLAGR_010225-RA	315	720
scaffold6-gene-15.54-mRNA-1	CLAGR_003841-RA	188	480	CLAGR_003845-RA	633	802
scaffold2-gene-10.66-mRNA-1	CLAGR_009178-RA	1	1250	CLAGR_009181-RA	1443	1559
scaffold6-gene-15.30-mRNA-1	CLAGR_010775-RA	9	248	CLAGR_010774-RA	267	461
scaffold17-gene-4.81-mRNA-1	CLAGR_003284-RA	37	170	CLAGR_003285-RA	298	346
scaffold20-gene-1.76-mRNA-1	CLAGR_008154-RA	5	148	CLAGR_000883-RA	206	1498
scaffold20-gene-1.76-mRNA-1	CLAGR_008154-RA	5	148	CLAGR_002878-RA	430	1081
scaffold20-gene-1.76-mRNA-1	CLAGR_008154-RA	5	148	CLAGR_002705-RA	189	987
scaffold20-gene-1.76-mRNA-1	CLAGR_008154-RA	5	148	CLAGR_006942-RA	195	1252
scaffold20-gene-1.76-mRNA-1	CLAGR_008154-RA	5	148	CLAGR_009814-RA	206	1355
scaffold20-gene-1.76-mRNA-1	CLAGR_008154-RA	5	148	CLAGR_010985-RA	823	1158
scaffold20-gene-1.89-mRNA-1	CLAGR_001750-RA	7	245	CLAGR_001749-RA	282	625
scaffold2-gene-12.34-mRNA-1	CLAGR_007012-RA	1	570	CLAGR_007015-RA	630	843
scaffold8-gene-1.109-mRNA-1	CLAGR_010668-RA	1	461	CLAGR_010670-RA	495	537
scaffold8-gene-1.109-mRNA-1	CLAGR_010668-RA	1	461	CLAGR_010671-RA	542	1138
scaffold8-gene-1.109-mRNA-1	CLAGR_010670-RA	495	537	CLAGR_010671-RA	542	1138
scaffold1-gene-32.129-mRNA-1	CLAGR_009556-RA	103	334	CLAGR_006221-RA	361	697
scaffold3-gene-19.13-mRNA-1	CLAGR_011007-RA	59	402	CLAGR_004698-RA	564	707
scaffold3-gene-19.13-mRNA-1	CLAGR_011007-RA	59	402	CLAGR_009527-RA	569	714
scaffold7-gene-5.68-mRNA-1	CLAGR_000524-RA	60	300	CLAGR_000525-RA	479	1938
scaffold13-gene-8.60-mRNA-1	CLAGR_011062-RA	1	40	CLAGR_010079-RA	225	299
scaffold4-gene-11.45-mRNA-1	CLAGR_000936-RA	9	216	CLAGR_004022-RA	233	383
scaffold9-gene-6.39-mRNA-1	CLAGR_009858-RB	1	75	CLAGR_009858-RA	80	308
scaffold12-gene-1.56-mRNA-1	CLAGR_009110-RA	1	483	CLAGR_009109-RA	556	826
scaffold17-gene-3.83-mRNA-1	CLAGR_005224-RA	15	273	CLAGR_005223-RA	400	794
scaffold8-gene-16.149-mRNA-1	CLAGR_000768-RA	225	319	CLAGR_002995-RA	462	719
scaffold8-gene-16.149-mRNA-1	CLAGR_000768-RA	225	319	CLAGR_003003-RA	508	719

scaffold8-gene-16.149-mRNA-1	CLAGR_000768-RA	225	319	CLAGR_003006-RA	588	719
scaffold1-gene-27.59-mRNA-1	CLAGR_006315-RA	1	1107	CLAGR_006316-RA	1124	1340
scaffold1-gene-17.83-mRNA-1	CLAGR_001612-RA	23	68	CLAGR_005759-RA	263	437
scaffold10-gene-11.69-mRNA-1	CLAGR_005276-RA	10	742	CLAGR_005277-RA	872	1594
scaffold15-gene-7.115-mRNA-1	CLAGR_002785-RA	74	175	CLAGR_001044-RB	176	530
scaffold15-gene-7.115-mRNA-1	CLAGR_002785-RA	74	175	CLAGR_002786-RA	203	594
scaffold3-gene-9.87-mRNA-1	CLAGR_010952-RA	286	460	CLAGR_010951-RA	809	904
scaffold10-gene-10.100-mRNA-1	CLAGR_008533-RA	1	255	CLAGR_000043-RA	324	590
scaffold10-gene-10.100-mRNA-1	CLAGR_008533-RA	1	255	CLAGR_005826-RA	351	458
scaffold7-gene-11.45-mRNA-1	CLAGR_005781-RA	1	166	CLAGR_005780-RA	218	381
scaffold2-gene-25.51-mRNA-1	CLAGR_010320-RA	50	460	CLAGR_006351-RA	474	779
scaffold7-gene-3.75-mRNA-1	CLAGR_001879-RA	1	364	CLAGR_001928-RA	493	895
scaffold5-gene-14.14-mRNA-1	CLAGR_002298-RA	1	389	CLAGR_002299-RA	391	1373
scaffold15-gene-2.123-mRNA-1	CLAGR_005645-RA	1	93	CLAGR_005644-RA	127	281
scaffold12-gene-15.30-mRNA-1	CLAGR_002195-RA	14	330	CLAGR_008947-RA	387	1056
scaffold12-gene-15.30-mRNA-1	CLAGR_003518-RA	156	251	CLAGR_008947-RA	387	1056
scaffold12-gene-15.30-mRNA-1	CLAGR_008364-RA	17	358	CLAGR_008947-RA	387	1056
scaffold12-gene-15.30-mRNA-1	CLAGR_008808-RA	15	184	CLAGR_008947-RA	387	1056
scaffold12-gene-15.30-mRNA-1	CLAGR_009311-RA	9	346	CLAGR_008947-RA	387	1056
scaffold12-gene-15.30-mRNA-1	CLAGR_010141-RA	27	290	CLAGR_008947-RA	387	1056
scaffold12-gene-15.30-mRNA-1	CLAGR_010939-RA	9	309	CLAGR_008947-RA	387	1056
scaffold2-gene-8.100-mRNA-1	CLAGR_006132-RA	38	726	CLAGR_006133-RA	828	1141
scaffold2-gene-8.100-mRNA-1	CLAGR_008655-RA	45	642	CLAGR_006133-RA	828	1141
scaffold12-gene-17.4-mRNA-1	CLAGR_003533-RA	1	189	CLAGR_003532-RA	278	531
scaffold1-gene-32.65-mRNA-1	CLAGR_009285-RA	8	358	CLAGR_010789-RA	366	599
scaffold5-gene-7.41-mRNA-1	CLAGR_003678-RA	17	74	CLAGR_003677-RA	106	828
scaffold3-gene-9.82-mRNA-1	CLAGR_010330-RA	283	385	CLAGR_010337-RA	620	705
scaffold3-gene-9.82-mRNA-1	CLAGR_010333-RA	288	393	CLAGR_010337-RA	620	705
scaffold4-gene-1.71-mRNA-1	CLAGR_008001-RA	118	491	CLAGR_008004-RA	526	690
scaffold13-gene-2.7-mRNA-1	CLAGR_002538-RA	532	592	CLAGR_003704-RA	664	978
scaffold13-gene-2.7-mRNA-1	CLAGR_003703-RA	40	613	CLAGR_003704-RA	664	978
scaffold13-gene-7.149-mRNA-1	CLAGR_001522-RA	38	348	CLAGR_001521-RA	478	1481
scaffold14-gene-3.80-mRNA-1	CLAGR_001794-RA	47	111	CLAGR_001793-RA	161	486
scaffold13-gene-3.3-mRNA-1	CLAGR_003759-RA	32	330	CLAGR_003718-RA	370	627
scaffold20-gene-	CLAGR_005620-	3	81	CLAGR_005619-	84	890

3.69-mRNA-1	RA			RA		
scaffold20-gene-3.69-mRNA-1	CLAGR_005620-RA	3	81	CLAGR_009692-RA	304	552
scaffold4-gene-4.13-mRNA-1	CLAGR_002673-RA	665	779	CLAGR_000239-RA	800	1099
scaffold4-gene-4.13-mRNA-1	CLAGR_002673-RA	665	779	CLAGR_005266-RA	825	1077
scaffold4-gene-4.13-mRNA-1	CLAGR_002673-RA	665	779	CLAGR_011236-RA	860	1087
scaffold15-gene-1.67-mRNA-1	CLAGR_000786-RA	377	1151	CLAGR_000787-RA	1181	1366
scaffold15-gene-1.67-mRNA-1	CLAGR_003336-RA	633	729	CLAGR_000787-RA	1181	1366
scaffold3-gene-15.18-mRNA-1	CLAGR_002948-RA	1	419	CLAGR_002950-RA	633	1019
scaffold10-gene-15.34-mRNA-1	CLAGR_007238-RA	58	416	CLAGR_007228-RA	481	865
scaffold4-gene-1.80-mRNA-1	CLAGR_008010-RA	31	548	CLAGR_008011-RA	581	688
scaffold7-gene-5.31-mRNA-1	CLAGR_000568-RA	605	722	CLAGR_000567-RA	726	1627
scaffold7-gene-5.31-mRNA-1	CLAGR_000569-RA	1	498	CLAGR_000567-RA	726	1627
scaffold7-gene-5.31-mRNA-1	CLAGR_000569-RA	1	498	CLAGR_000568-RA	605	722
scaffold5-gene-5.104-mRNA-1	CLAGR_006121-RA	48	340	CLAGR_009461-RA	357	601
scaffold5-gene-5.104-mRNA-1	CLAGR_007886-RA	50	340	CLAGR_009461-RA	357	601
scaffold3-gene-18.92-mRNA-1	CLAGR_001969-RA	31	302	CLAGR_001968-RA	321	513
scaffold1-gene-11.23-mRNA-1	CLAGR_004513-RA	101	267	CLAGR_004514-RA	288	869
scaffold1-gene-11.23-mRNA-1	CLAGR_007230-RA	138	261	CLAGR_004514-RA	288	869
scaffold7-gene-12.59-mRNA-1	CLAGR_006772-RA	15	238	CLAGR_000377-RA	320	548
scaffold3-gene-15.42-mRNA-1	CLAGR_002941-RA	47	183	CLAGR_002953-RA	226	491
scaffold9-gene-10.127-mRNA-1	CLAGR_000137-RA	53	585	CLAGR_000135-RA	822	944
scaffold6-gene-7.81-mRNA-1	CLAGR_008621-RA	7	503	CLAGR_008622-RA	589	2375
scaffold25-gene-0.113-mRNA-1	CLAGR_001502-RA	37	452	CLAGR_001503-RA	551	765
scaffold1-gene-16.91-mRNA-1	CLAGR_003450-RA	1	121	CLAGR_003449-RA	316	509
scaffold16-gene-6.107-mRNA-1	CLAGR_007329-RA	115	414	CLAGR_007330-RA	441	524
scaffold2-gene-7.19-mRNA-1	CLAGR_001897-RA	123	283	CLAGR_001896-RA	390	575
scaffold8-gene-16.136-mRNA-1	CLAGR_004703-RA	41	124	CLAGR_004704-RA	184	599
scaffold2-gene-24.73-mRNA-1	CLAGR_009289-RA	1	249	CLAGR_009640-RA	317	448
scaffold3-gene-3.29-mRNA-1	CLAGR_008614-RA	53	91	CLAGR_008613-RA	134	256
scaffold5-gene-6.64-mRNA-1	CLAGR_006069-RA	5	64	CLAGR_006068-RA	73	346
scaffold4-gene-8.21-mRNA-1	CLAGR_006371-RA	17	450	CLAGR_006372-RA	529	621
scaffold5-gene-17.123-mRNA-1	CLAGR_001587-RA	1	83	CLAGR_001586-RA	107	828
scaffold8-gene-14.22-mRNA-1	CLAGR_009636-RA	33	132	CLAGR_009635-RA	173	413
scaffold8-gene-14.25-mRNA-1	CLAGR_004738-RA	7	861	CLAGR_004735-RA	938	1250
scaffold8-gene-14.25-mRNA-1	CLAGR_005609-RA	33	645	CLAGR_004735-RA	938	1250

scaffold5-gene-4.71-mRNA-1	CLAGR_006858-RA	1	546	CLAGR_005699-RA	632	884
scaffold2-gene-11.123-mRNA-1	CLAGR_006167-RA	251	623	CLAGR_006166-RA	700	1448
scaffold1-gene-6.17-mRNA-1	CLAGR_004122-RA	456	1003	CLAGR_004121-RA	1073	1405
scaffold1-gene-6.17-mRNA-1	CLAGR_004123-RA	1	429	CLAGR_004121-RA	1073	1405
scaffold1-gene-6.17-mRNA-1	CLAGR_004123-RA	1	429	CLAGR_004122-RA	456	1003
scaffold2-gene-26.15-mRNA-1	CLAGR_000220-RA	347	512	CLAGR_004146-RA	552	774
scaffold2-gene-26.15-mRNA-1	CLAGR_002185-RA	10	528	CLAGR_004146-RA	552	774
scaffold2-gene-26.15-mRNA-1	CLAGR_004776-RA	3	546	CLAGR_004146-RA	552	774
scaffold1-gene-25.79-mRNA-1	CLAGR_010780-RA	586	795	CLAGR_006911-RA	1752	2274
scaffold1-gene-25.79-mRNA-1	CLAGR_010243-RA	50	1024	CLAGR_006911-RA	1752	2274
scaffold14-gene-7.2-mRNA-1	CLAGR_001637-RA	23	198	CLAGR_001638-RA	222	783
scaffold8-gene-12.2-mRNA-1	CLAGR_009997-RA	26	873	CLAGR_010012-RA	889	1310
scaffold11-gene-5.54-mRNA-1	CLAGR_006244-RA	25	61	CLAGR_006245-RA	66	524
scaffold16-gene-3.70-mRNA-1	CLAGR_010497-RA	1	670	CLAGR_010496-RA	1023	1179
scaffold8-gene-10.89-mRNA-1	CLAGR_007895-RA	5	60	CLAGR_007250-RA	66	403
scaffold3-gene-17.13-mRNA-1	CLAGR_005954-RA	1	744	CLAGR_005953-RA	822	1743
scaffold8-gene-3.13-mRNA-1	CLAGR_002264-RA	7	1709	CLAGR_001747-RA	1787	2121
scaffold10-gene-5.61-mRNA-1	CLAGR_002684-RA	1	81	CLAGR_002683-RA	97	205
scaffold8-gene-6.106-mRNA-1	CLAGR_003237-RA	5	172	CLAGR_002845-RA	405	614
scaffold8-gene-6.106-mRNA-1	CLAGR_003237-RA	5	172	CLAGR_003238-RA	200	618
scaffold8-gene-6.106-mRNA-1	CLAGR_003237-RA	5	172	CLAGR_009712-RA	403	608
scaffold10-gene-3.106-mRNA-1	CLAGR_007807-RA	6	260	CLAGR_000337-RA	385	613
scaffold10-gene-3.106-mRNA-1	CLAGR_010693-RA	18	307	CLAGR_000337-RA	385	613
scaffold10-gene-3.106-mRNA-1	CLAGR_007807-RA	6	260	CLAGR_004095-RA	336	565
scaffold10-gene-3.106-mRNA-1	CLAGR_007807-RA	6	260	CLAGR_010574-RA	385	562
scaffold10-gene-3.106-mRNA-1	CLAGR_010693-RA	18	307	CLAGR_004095-RA	336	565
scaffold10-gene-3.106-mRNA-1	CLAGR_010693-RA	18	307	CLAGR_010574-RA	385	562
scaffold27-gene-0.41-mRNA-1	CLAGR_010999-RA	89	174	CLAGR_011000-RA	237	326
scaffold1-gene-3.129-mRNA-1	CLAGR_008081-RA	1	310	CLAGR_009727-RA	360	606
scaffold8-gene-1.151-mRNA-1	CLAGR_003359-RA	1	574	CLAGR_003360-RA	664	815
scaffold20-gene-0.65-mRNA-1	CLAGR_001825-RA	92	346	CLAGR_002317-RA	384	610
scaffold1-gene-18.88-mRNA-1	CLAGR_000697-RA	79	497	CLAGR_000698-RA	545	1104
scaffold1-gene-28.146-mRNA-1	CLAGR_005800-RA	18	335	CLAGR_002449-RA	337	900
scaffold17-gene-1.13-mRNA-1	CLAGR_001194-RA	51	689	CLAGR_005503-RA	1289	1434
scaffold17-gene-	CLAGR_001194-	51	689	CLAGR_010900-	752	1002

1.13-mRNA-1	RA			RA		
scaffold17-gene-1.13-mRNA-1	CLAGR_010900-RA	752	1002	CLAGR_005503-RA	1289	1434
scaffold19-gene-3.2-mRNA-1	CLAGR_006210-RA	9	277	CLAGR_006211-RA	297	480
scaffold19-gene-3.2-mRNA-1	CLAGR_006210-RA	9	277	CLAGR_011195-RA	694	828
scaffold19-gene-3.2-mRNA-1	CLAGR_006211-RA	297	480	CLAGR_011195-RA	694	828
scaffold4-gene-3.43-mRNA-1	CLAGR_010274-RA	1	1139	CLAGR_007423-RA	1164	1360
scaffold10-gene-7.21-mRNA-1	CLAGR_004944-RA	1	508	CLAGR_010529-RA	999	1203
scaffold10-gene-7.21-mRNA-1	CLAGR_004944-RA	1	508	CLAGR_010530-RA	1045	1202
scaffold25-gene-0.48-mRNA-1	CLAGR_000624-RA	1	439	CLAGR_009331-RA	592	1036
scaffold25-gene-0.48-mRNA-1	CLAGR_009332-RA	1	434	CLAGR_009331-RA	592	1036
scaffold3-gene-16.45-mRNA-1	CLAGR_010224-RA	1	348	CLAGR_002920-RA	375	464
scaffold3-gene-19.74-mRNA-1	CLAGR_009023-RA	62	302	CLAGR_009022-RA	376	520
scaffold3-gene-19.75-mRNA-1	CLAGR_010933-RA	1	133	CLAGR_008989-RA	196	853
scaffold14-gene-2.82-mRNA-1	CLAGR_001803-RA	12	659	CLAGR_001802-RA	663	1402
scaffold10-gene-16.139-mRNA-1	CLAGR_000498-RA	23	445	CLAGR_000497-RA	456	916
scaffold10-gene-16.139-mRNA-1	CLAGR_000498-RA	23	445	CLAGR_007161-RA	588	740
scaffold10-gene-16.139-mRNA-1	CLAGR_000498-RA	23	445	CLAGR_009456-RA	820	899
scaffold10-gene-16.139-mRNA-1	CLAGR_000498-RA	23	445	CLAGR_009035-RA	763	1099
scaffold10-gene-16.139-mRNA-1	CLAGR_000498-RA	23	445	CLAGR_011129-RA	742	868
scaffold10-gene-16.139-mRNA-1	CLAGR_007161-RA	588	740	CLAGR_009035-RA	763	1099
scaffold10-gene-16.139-mRNA-1	CLAGR_007161-RA	588	740	CLAGR_011129-RA	742	868
scaffold1-gene-18.125-mRNA-1	CLAGR_000765-RA	1	835	CLAGR_000764-RA	868	1188
scaffold1-gene-18.66-mRNA-1	CLAGR_000844-RA	1	318	CLAGR_000845-RA	363	602
scaffold5-gene-12.19-mRNA-1	CLAGR_002238-RA	12	553	CLAGR_002231-RA	679	1058
scaffold14-gene-5.59-mRNA-1	CLAGR_008128-RA	1	1727	CLAGR_008127-RA	1740	2610
scaffold4-gene-14.31-mRNA-1	CLAGR_007978-RA	74	777	CLAGR_009783-RA	4346	5384
scaffold1-gene-11.22-mRNA-1	CLAGR_001111-RA	20	263	CLAGR_004654-RA	1048	1200
scaffold1-gene-11.22-mRNA-1	CLAGR_001111-RA	20	263	CLAGR_009335-RA	298	1422
scaffold1-gene-11.22-mRNA-1	CLAGR_001111-RA	20	263	CLAGR_009526-RA	330	1583
scaffold1-gene-11.22-mRNA-1	CLAGR_004367-RA	7	272	CLAGR_004654-RA	1048	1200
scaffold1-gene-11.22-mRNA-1	CLAGR_004367-RA	7	272	CLAGR_009335-RA	298	1422
scaffold1-gene-11.22-mRNA-1	CLAGR_004367-RA	7	272	CLAGR_009526-RA	330	1583
scaffold1-gene-11.22-mRNA-1	CLAGR_005383-RA	6	907	CLAGR_004654-RA	1048	1200
scaffold1-gene-11.22-mRNA-1	CLAGR_009083-RA	23	265	CLAGR_009335-RA	298	1422
scaffold1-gene-11.22-mRNA-1	CLAGR_009953-RA	7	945	CLAGR_004654-RA	1048	1200

scaffold22-gene-2.6-mRNA-1	CLAGR_004254-RA	1	89	CLAGR_004255-RA	141	348
scaffold8-gene-15.71-mRNA-1	CLAGR_004725-RA	1	111	CLAGR_004726-RA	132	284
scaffold2-gene-25.78-mRNA-1	CLAGR_009280-RA	1	56	CLAGR_009281-RA	73	413
scaffold5-gene-7.8-mRNA-1	CLAGR_003656-RA	1	127	CLAGR_003655-RA	223	744
scaffold6-gene-17.69-mRNA-1	CLAGR_003820-RA	23	369	CLAGR_006251-RA	408	520
scaffold15-gene-3.121-mRNA-1	CLAGR_005673-RA	1	81	CLAGR_004502-RA	175	303
scaffold14-gene-6.86-mRNA-1	CLAGR_002374-RA	99	351	CLAGR_001196-RA	1145	1364
scaffold14-gene-6.86-mRNA-1	CLAGR_003856-RA	67	521	CLAGR_001196-RA	1145	1364
scaffold14-gene-6.86-mRNA-1	CLAGR_002374-RA	99	351	CLAGR_003009-RA	394	1505
scaffold14-gene-6.86-mRNA-1	CLAGR_002374-RA	99	351	CLAGR_006981-RA	960	2268
scaffold14-gene-6.86-mRNA-1	CLAGR_003856-RA	67	521	CLAGR_006981-RA	960	2268
scaffold14-gene-6.86-mRNA-1	CLAGR_002374-RA	99	351	CLAGR_008627-RA	354	1248
scaffold5-gene-6.5-mRNA-1	CLAGR_006092-RA	2	324	CLAGR_006093-RA	357	785
scaffold1-gene-31.112-mRNA-1	CLAGR_006238-RA	1	276	CLAGR_006239-RA	293	542
scaffold9-gene-9.116-mRNA-1	CLAGR_006794-RA	55	125	CLAGR_006793-RA	341	545
scaffold21-gene-2.35-mRNA-1	CLAGR_007053-RA	12	108	CLAGR_007054-RA	117	348
scaffold3-gene-9.76-mRNA-1	CLAGR_009020-RA	8	438	CLAGR_002834-RA	569	706
scaffold17-gene-4.94-mRNA-1	CLAGR_008061-RA	1	115	CLAGR_008060-RA	131	662
scaffold6-gene-3.122-mRNA-1	CLAGR_007354-RA	24	96	CLAGR_007353-RA	111	432
scaffold12-gene-15.20-mRNA-1	CLAGR_008957-RA	27	107	CLAGR_008956-RA	297	389
scaffold15-gene-0.62-mRNA-1	CLAGR_001732-RA	1	48	CLAGR_001731-RA	83	733
scaffold13-gene-3.34-mRNA-1	CLAGR_003776-RA	265	888	CLAGR_003775-RA	1017	1548
scaffold4-gene-12.112-mRNA-1	CLAGR_007936-RA	1	2099	CLAGR_002886-RA	2133	2694
scaffold10-gene-8.71-mRNA-1	CLAGR_004952-RA	10	209	CLAGR_004953-RA	266	593
scaffold16-gene-3.76-mRNA-1	CLAGR_006013-RA	1	1110	CLAGR_008454-RA	1331	1598
scaffold6-gene-2.126-mRNA-1	CLAGR_000315-RA	14	835	CLAGR_000207-RA	883	1031
scaffold6-gene-15.67-mRNA-1	CLAGR_004281-RA	184	455	CLAGR_010986-RA	933	1268
scaffold27-gene-0.140-mRNA-1	CLAGR_002823-RA	42	100	CLAGR_010152-RA	256	437
scaffold9-gene-11.49-mRNA-1	CLAGR_000123-RA	1	1823	CLAGR_000124-RA	1850	1896
scaffold14-gene-5.74-mRNA-1	CLAGR_005292-RA	1	51	CLAGR_005291-RA	57	700
scaffold14-gene-5.32-mRNA-1	CLAGR_008144-RA	27	307	CLAGR_008145-RA	317	1344
scaffold8-gene-15.38-mRNA-1	CLAGR_004742-RA	8	407	CLAGR_004739-RA	428	997
scaffold7-gene-9.32-mRNA-1	CLAGR_006824-RA	1	100	CLAGR_006823-RA	109	529
scaffold3-gene-4.24-mRNA-1	CLAGR_001574-RA	16	137	CLAGR_004544-RA	1551	2619
scaffold3-gene-	CLAGR_004546-	17	109	CLAGR_004544-	1551	2619

4.24-mRNA-1	RA			RA		
scaffold3-gene-4.24-mRNA-1	CLAGR_007687-RA	19	441	CLAGR_004544-RA	1551	2619
scaffold3-gene-4.24-mRNA-1	CLAGR_001574-RA	16	137	CLAGR_007684-RA	2348	2617
scaffold3-gene-4.24-mRNA-1	CLAGR_004546-RA	17	109	CLAGR_007684-RA	2348	2617
scaffold3-gene-4.24-mRNA-1	CLAGR_007685-RA	2041	2316	CLAGR_007684-RA	2348	2617
scaffold3-gene-4.24-mRNA-1	CLAGR_007687-RA	19	441	CLAGR_007684-RA	2348	2617
scaffold3-gene-4.24-mRNA-1	CLAGR_007687-RA	19	441	CLAGR_007685-RA	2041	2316
scaffold7-gene-9.2-mRNA-1	CLAGR_006790-RA	33	134	CLAGR_006789-RA	139	626
scaffold1-gene-27.136-mRNA-1	CLAGR_002437-RA	1	537	CLAGR_002440-RA	584	860
scaffold11-gene-0.29-mRNA-1	CLAGR_010874-RA	1	151	CLAGR_007071-RA	248	695
scaffold8-gene-16.126-mRNA-1	CLAGR_004682-RA	13	555	CLAGR_004683-RA	590	953
scaffold28-gene-0.46-mRNA-1	CLAGR_005941-RA	1	2014	CLAGR_005949-RA	2109	2980
scaffold24-gene-1.51-mRNA-1	CLAGR_004588-RA	19	458	CLAGR_011136-RA	493	725
scaffold24-gene-1.23-mRNA-1	CLAGR_004584-RA	9	352	CLAGR_004585-RA	449	600
scaffold8-gene-9.66-mRNA-1	CLAGR_006718-RA	334	1355	CLAGR_005357-RA	1847	2225
scaffold5-gene-0.38-mRNA-1	CLAGR_003104-RB	1	124	CLAGR_004311-RA	279	438
scaffold5-gene-0.38-mRNA-1	CLAGR_003425-RA	1	220	CLAGR_004311-RA	279	438
scaffold5-gene-0.38-mRNA-1	CLAGR_005321-RA	1	216	CLAGR_004311-RA	279	438
scaffold5-gene-0.38-mRNA-1	CLAGR_009475-RA	1	240	CLAGR_004311-RA	279	438
scaffold23-gene-1.62-mRNA-1	CLAGR_006035-RA	1	111	CLAGR_006034-RA	114	440
scaffold1-gene-20.39-mRNA-1	CLAGR_003497-RA	1	805	CLAGR_003526-RA	822	1177
scaffold8-gene-11.3-mRNA-1	CLAGR_001573-RA	403	506	CLAGR_001572-RA	1976	2464
scaffold4-gene-9.7-mRNA-1	CLAGR_009118-RA	71	468	CLAGR_006403-RA	552	728
scaffold1-gene-14.66-mRNA-1	CLAGR_000807-RA	25	65	CLAGR_000806-RA	101	710
scaffold13-gene-2.8-mRNA-1	CLAGR_003692-RA	9	263	CLAGR_003693-RA	270	500
scaffold12-gene-11.47-mRNA-1	CLAGR_004051-RA	2	63	CLAGR_004052-RA	106	855
scaffold9-gene-0.38-mRNA-1	CLAGR_010917-RA	601	638	CLAGR_010087-RA	1261	1325
scaffold4-gene-16.69-mRNA-1	CLAGR_006653-RA	1	340	CLAGR_008875-RA	405	554
scaffold9-gene-11.143-mRNA-1	CLAGR_000142-RA	37	113	CLAGR_000141-RA	204	539
scaffold5-gene-14.67-mRNA-1	CLAGR_002383-RA	3	274	CLAGR_001407-RA	380	587
scaffold3-gene-22.223-mRNA-1	CLAGR_004454-RA	41	148	CLAGR_004403-RA	322	650
scaffold3-gene-22.223-mRNA-1	CLAGR_009704-RA	13	319	CLAGR_004403-RA	322	650
scaffold3-gene-14.117-mRNA-1	CLAGR_002952-RA	1	122	CLAGR_002951-RA	241	562
scaffold10-gene-5.129-mRNA-1	CLAGR_008308-RA	1	557	CLAGR_008309-RA	649	850
scaffold19-gene-0.118-mRNA-1	CLAGR_001349-RA	1	104	CLAGR_001348-RA	106	381

scaffold19-gene-0.123-mRNA-1	CLAGR_002085-RA	6	198	CLAGR_001152-RA	261	501
scaffold19-gene-0.123-mRNA-1	CLAGR_002085-RA	6	198	CLAGR_002086-RA	252	621
scaffold8-gene-14.3-mRNA-1	CLAGR_010805-RA	1	363	CLAGR_010815-RA	413	910
scaffold1-gene-18.86-mRNA-1	CLAGR_000398-RA	84	296	CLAGR_010514-RA	369	733
scaffold1-gene-31.61-mRNA-1	CLAGR_002734-RA	268	456	CLAGR_009288-RA	563	796
scaffold1-gene-31.61-mRNA-1	CLAGR_008453-RA	61	443	CLAGR_009288-RA	563	796
scaffold6-gene-0.52-mRNA-1	CLAGR_000061-RA	32	128	CLAGR_000062-RA	272	411
scaffold15-gene-0.117-mRNA-1	CLAGR_001678-RA	8	1367	CLAGR_001729-RA	1369	2000
scaffold17-gene-4.118-mRNA-1	CLAGR_008051-RA	1	131	CLAGR_008052-RA	517	654
scaffold4-gene-8.35-mRNA-1	CLAGR_006370-RA	9	572	CLAGR_006373-RA	576	768
scaffold9-gene-14.211-mRNA-1	CLAGR_001382-RA	17	333	CLAGR_001383-RA	338	864
scaffold17-gene-4.151-mRNA-1	CLAGR_003272-RA	1	830	CLAGR_003269-RA	880	1351
scaffold23-gene-2.166-mRNA-1	CLAGR_005490-RA	4	66	CLAGR_010467-RA	124	355
scaffold1-gene-3.18-mRNA-1	CLAGR_008064-RA	74	526	CLAGR_008075-RA	601	726
scaffold2-gene-26.197-mRNA-1	CLAGR_007536-RA	1	313	CLAGR_000249-RA	320	563
scaffold2-gene-1.49-mRNA-1	CLAGR_006478-RA	4	452	CLAGR_006477-RA	473	593
scaffold2-gene-1.76-mRNA-1	CLAGR_006864-RA	1	27	CLAGR_006863-RA	54	104
scaffold4-gene-19.38-mRNA-1	CLAGR_001494-RA	76	202	CLAGR_001493-RA	249	410
scaffold13-gene-8.75-mRNA-1	CLAGR_008673-RA	1	192	CLAGR_008672-RA	285	574
scaffold4-gene-9.19-mRNA-1	CLAGR_007938-RA	77	415	CLAGR_000980-RA	445	754
scaffold9-gene-9.100-mRNA-1	CLAGR_001822-RA	16	528	CLAGR_007873-RA	584	878
scaffold9-gene-9.100-mRNA-1	CLAGR_007874-RA	1	566	CLAGR_007873-RA	584	878
scaffold9-gene-6.8-mRNA-2	CLAGR_006189-RA	9	58	CLAGR_005149-RA	297	602

Table A-5: *L. pustulata* genes predicted by *funannotate*, which the *Rosetta Stone* method identified to be a gene fusion. Comparisons for the *Rosetta Stone* method were made against the genes of *Cladonia grayi*. Each row represents one possible gene fusion that involves two *C. grayi* genes. *L. pustulata* genes can appear more than once, if the fusion involves more than two *C. grayi* genes or if multiple *C. grayi* genes match the same fusion. The start and end positions of the *C. grayi* genes within the fused *L. pustulata* gene are given. *C. grayi* genes that are part of an LCA_{Lec} group in which *L. pustulata* is absent are highlighted. *C. grayi* genes are highlighted in green if the *L. pustulata* ortholog was found via *HaMStR* and in blue if the *L. pustulata* ortholog was found by *Exonerate*.

Fused gene	Fusion partner #1	Start position in fused gene	End position in fused gene	Fusion partner #2	Start position in fused gene	End position in fused gene
FUN_01172	CLAGR_006478-RA	4	452	CLAGR_006477-RA	473	593
FUN_09282	CLAGR_006187-RA	272	437	CLAGR_008444-RA	702	750
FUN_04770	CLAGR_003678-RA	17	74	CLAGR_003677-RA	106	828
FUN_03670	CLAGR_004917-RA	1	150	CLAGR_005022-RA	226	418
FUN_05857	CLAGR_000104-RA	4	1448	CLAGR_000325-RA	1454	1618
FUN_05857	CLAGR_000104-RA	4	1448	CLAGR_000484-RA	1451	1634
FUN_05857	CLAGR_000104-RA	4	1448	CLAGR_005753-RA	1476	1601
FUN_06909	CLAGR_000007-RA	90	180	CLAGR_001421-RA	192	563
FUN_06909	CLAGR_000007-RA	90	180	CLAGR_006427-RA	307	502
FUN_02492	CLAGR_010226-RA	1	279	CLAGR_010225-RA	315	720
FUN_09239	CLAGR_001825-RA	92	346	CLAGR_002317-RA	384	610
FUN_05192	CLAGR_010670-RA	34	74	CLAGR_010671-RA	79	711
FUN_06709	CLAGR_010917-RA	601	638	CLAGR_010087-RA	1261	1325
FUN_09697	CLAGR_001502-RA	29	444	CLAGR_001503-RA	543	757
FUN_05709	CLAGR_000061-RA	32	128	CLAGR_000062-RA	272	411
FUN_08949	CLAGR_005490-RA	4	66	CLAGR_010467-RA	124	355
FUN_08024	CLAGR_005283-RA	1	96	CLAGR_005285-RA	120	205
FUN_00192	CLAGR_004107-RA	464	1966	CLAGR_004098-RA	2018	2692
FUN_08020	CLAGR_008144-RA	291	571	CLAGR_008145-RA	581	1549
FUN_02036	CLAGR_008977-RA	158	223	CLAGR_008978-RA	373	776
FUN_03170	CLAGR_007978-RA	1	1554	CLAGR_009783-RA	5192	6230
FUN_02406	CLAGR_010481-RA	28	432	CLAGR_010480-RA	434	874
FUN_00530	CLAGR_003450-RA	1	121	CLAGR_003449-RA	316	509
FUN_02264	CLAGR_010952-RA	286	460	CLAGR_010951-RA	809	904
FUN_05480	CLAGR_001573-RA	488	591	CLAGR_001572-RA	2061	2549
FUN_05480	CLAGR_004546-RA	40	134	CLAGR_001572-RA	2061	2549
FUN_05480	CLAGR_004546-RA	40	134	CLAGR_001573-RA	488	591
FUN_05480	CLAGR_004546-RA	40	134	CLAGR_009720-RA	433	2546
FUN_08153	CLAGR_001732-RA	1	48	CLAGR_001731-RA	83	733
FUN_06830	CLAGR_009858-RB	1	75	CLAGR_009858-RA	80	308
FUN_02295	CLAGR_010330-RA	283	385	CLAGR_010337-RA	620	705

FUN_02295	CLAGR_010333-RA	288	393	CLAGR_010337-RA	620	705
FUN_01389	CLAGR_001848-RA	1	260	CLAGR_001849-RA	265	649
FUN_04797	CLAGR_003628-RA	38	144	CLAGR_003627-RA	152	295
FUN_07838	CLAGR_001778-RA	17	351	CLAGR_002376-RA	414	932
FUN_07838	CLAGR_001778-RA	17	351	CLAGR_010946-RA	765	895
FUN_03283	CLAGR_001494-RA	76	202	CLAGR_001493-RA	249	410
FUN_05347	CLAGR_003237-RA	5	172	CLAGR_002845-RA	405	614
FUN_05347	CLAGR_003237-RA	5	172	CLAGR_003238-RA	200	618
FUN_08074	CLAGR_002271-RA	374	779	CLAGR_001196-RA	1052	1271
FUN_08074	CLAGR_005383-RA	45	941	CLAGR_001196-RA	1052	1271
FUN_08074	CLAGR_001196-RA	1052	1271	CLAGR_007633-RA	1325	1846
FUN_08074	CLAGR_005031-RA	79	1298	CLAGR_007633-RA	1325	1846
FUN_08074	CLAGR_002271-RA	374	779	CLAGR_007633-RA	1325	1846
FUN_08074	CLAGR_005383-RA	45	941	CLAGR_007633-RA	1325	1846
FUN_08074	CLAGR_008627-RA	323	1204	CLAGR_007633-RA	1325	1846
FUN_01817	CLAGR_001206-RA	3	143	CLAGR_007281-RA	203	352
FUN_06385	CLAGR_000569-RA	70	567	CLAGR_000568-RA	674	791
FUN_06385	CLAGR_008728-RA	517	648	CLAGR_000568-RA	674	791
FUN_07015	CLAGR_000123-RA	1	1823	CLAGR_000124-RA	1850	1896
FUN_00334	CLAGR_010285-RA	32	299	CLAGR_010975-RA	365	590
FUN_01718	CLAGR_002570-RA	90	202	CLAGR_001085-RA	243	634
FUN_01718	CLAGR_002570-RA	90	202	CLAGR_002240-RA	279	762
FUN_01718	CLAGR_002570-RA	90	202	CLAGR_003013-RA	334	642
FUN_01718	CLAGR_002570-RA	90	202	CLAGR_003560-RA	353	692
FUN_01718	CLAGR_002570-RA	90	202	CLAGR_007982-RA	349	527
FUN_01718	CLAGR_002570-RA	90	202	CLAGR_008410-RA	338	759
FUN_01718	CLAGR_002570-RA	90	202	CLAGR_009904-RA	332	762
FUN_08204	CLAGR_000786-RA	1	1136	CLAGR_000787-RA	1166	1351
FUN_08204	CLAGR_003336-RA	618	714	CLAGR_000787-RA	1166	1351
FUN_04768	CLAGR_003656-RA	1	127	CLAGR_003655-RA	223	744
FUN_08804	CLAGR_008061-RA	1	115	CLAGR_008060-RA	131	679
FUN_08611	CLAGR_007329-RA	62	399	CLAGR_007330-RA	426	486
FUN_00442	CLAGR_000807-RA	25	65	CLAGR_000806-RA	101	710
FUN_09180	CLAGR_007053-RA	12	108	CLAGR_007054-RA	117	348
FUN_06973	CLAGR_009372-RA	1	122	CLAGR_009371-RA	180	414
FUN_00215	CLAGR_004634-RA	204	638	CLAGR_004633-RA	1031	1117
FUN_07805	CLAGR_004247-RA	2	34	CLAGR_002744-RA	43	270
FUN_07805	CLAGR_004247-RA	2	34	CLAGR_004248-RA	36	823
FUN_02540	CLAGR_001969-RA	31	302	CLAGR_001968-RA	321	460

FUN_08893	CLAGR_006035-RA	1	111	CLAGR_006034-RA	114	440
FUN_05930	CLAGR_008621-RA	1	397	CLAGR_008622-RA	483	2269
FUN_01665	CLAGR_005123-RA	3	320	CLAGR_005122-RA	437	986
FUN_07404	CLAGR_002032-RA	108	261	CLAGR_002033-RA	313	440
FUN_06802	CLAGR_007432-RA	2	131	CLAGR_008268-RA	155	1437
FUN_01165	CLAGR_006864-RA	1	27	CLAGR_006863-RA	54	104
FUN_03249	CLAGR_001146-RA	370	537	CLAGR_007866-RA	613	822
FUN_03249	CLAGR_004015-RA	12	537	CLAGR_007866-RA	613	822
FUN_09318	CLAGR_005620-RA	1	68	CLAGR_005870-RA	293	707
FUN_09318	CLAGR_005620-RA	1	68	CLAGR_005619-RA	71	884
FUN_09318	CLAGR_005620-RA	1	68	CLAGR_007331-RA	372	471
FUN_09252	CLAGR_008154-RA	5	148	CLAGR_000883-RA	206	1498
FUN_09252	CLAGR_008154-RA	5	148	CLAGR_002878-RA	430	1081
FUN_09252	CLAGR_004841-RA	234	987	CLAGR_005266-RA	1032	1280
FUN_09252	CLAGR_008154-RA	5	148	CLAGR_004841-RA	234	987
FUN_09252	CLAGR_008154-RA	5	148	CLAGR_002667-RA	965	1528
FUN_09252	CLAGR_008154-RA	5	148	CLAGR_005266-RA	1032	1280
FUN_09252	CLAGR_008154-RA	5	148	CLAGR_009814-RA	206	1355
FUN_09252	CLAGR_008154-RA	5	148	CLAGR_010856-RA	903	1147
FUN_09252	CLAGR_008154-RA	5	148	CLAGR_010985-RA	823	1158
FUN_08037	CLAGR_005292-RA	1	51	CLAGR_005291-RA	57	700
FUN_06956	CLAGR_006794-RA	55	125	CLAGR_006793-RA	341	545
FUN_06508	CLAGR_005781-RA	1	166	CLAGR_005780-RA	218	381
FUN_02124	CLAGR_007685-RA	805	1080	CLAGR_007684-RA	1112	1381
FUN_08560	CLAGR_003157-RA	8	60	CLAGR_003158-RA	181	768
FUN_06377	CLAGR_000524-RA	60	300	CLAGR_000525-RA	479	1938
FUN_01689	CLAGR_005044-RA	970	1356	CLAGR_003320-RA	1598	1656
FUN_01689	CLAGR_005044-RA	970	1356	CLAGR_005004-RA	1561	1668
FUN_05276	CLAGR_007644-RA	1	351	CLAGR_004897-RA	361	530
FUN_04306	CLAGR_000054-RA	1	308	CLAGR_007582-RA	439	674
FUN_02277	CLAGR_009020-RA	8	438	CLAGR_002834-RA	569	706
FUN_02277	CLAGR_011233-RA	42	380	CLAGR_002834-RA	569	706
FUN_04496	CLAGR_003533-RA	1	189	CLAGR_003532-RA	278	531
FUN_02371	CLAGR_010502-RA	217	443	CLAGR_010503-RA	486	716
FUN_04669	CLAGR_006858-RA	1	546	CLAGR_005699-RA	632	884
FUN_06840	CLAGR_006189-RA	4	35	CLAGR_005149-RA	254	538
FUN_06840	CLAGR_006189-RA	4	35	CLAGR_006190-RA	47	277
FUN_06840	CLAGR_006189-RA	4	35	CLAGR_006720-RA	44	543
FUN_04579	CLAGR_001642-RA	10	86	CLAGR_004993-RA	224	923

FUN_04576	CLAGR_003104-RB	1	124	CLAGR_004311-RA	279	438
FUN_04576	CLAGR_003425-RA	1	220	CLAGR_004311-RA	279	438
FUN_04576	CLAGR_005321-RA	1	216	CLAGR_004311-RA	279	438
FUN_04576	CLAGR_009475-RA	1	240	CLAGR_004311-RA	279	438
FUN_00256	CLAGR_004552-RA	3	73	CLAGR_004551-RA	122	326
FUN_00702	CLAGR_007667-RA	19	217	CLAGR_007668-RA	252	384
FUN_07099	CLAGR_001382-RA	17	345	CLAGR_001383-RA	350	876
FUN_07650	CLAGR_003772-RA	1	734	CLAGR_007456-RA	769	1014
FUN_09665	CLAGR_002823-RA	42	100	CLAGR_010152-RA	256	437
FUN_06183	CLAGR_010775-RA	9	248	CLAGR_010774-RA	267	461
FUN_09663	CLAGR_010999-RA	36	121	CLAGR_011000-RA	184	273
FUN_00871	CLAGR_009784-RA	39	1333	CLAGR_006911-RA	1752	2274
FUN_00871	CLAGR_009784-RA	39	1333	CLAGR_010779-RA	1850	1945
FUN_00871	CLAGR_009784-RA	39	1333	CLAGR_010242-RA	1685	2587
FUN_00871	CLAGR_010780-RA	586	795	CLAGR_006911-RA	1752	2274
FUN_00871	CLAGR_010780-RA	586	795	CLAGR_010779-RA	1850	1945
FUN_00871	CLAGR_010780-RA	586	795	CLAGR_010242-RA	1685	2587
FUN_00871	CLAGR_010243-RA	50	1024	CLAGR_006911-RA	1752	2274
FUN_00871	CLAGR_010243-RA	50	1024	CLAGR_010779-RA	1850	1945
FUN_00871	CLAGR_010243-RA	50	1024	CLAGR_010242-RA	1685	2587
FUN_08214	CLAGR_001661-RA	54	458	CLAGR_001662-RA	498	587
FUN_00583	CLAGR_001612-RA	23	68	CLAGR_005759-RA	263	437
FUN_08818	CLAGR_008051-RA	1	143	CLAGR_008052-RA	532	669
FUN_08748	CLAGR_005224-RA	15	273	CLAGR_005223-RA	400	751
FUN_01875	CLAGR_009289-RA	1	249	CLAGR_009640-RA	317	448
FUN_06440	CLAGR_006790-RA	33	134	CLAGR_006789-RA	139	507
FUN_06443	CLAGR_006792-RA	5	398	CLAGR_006795-RA	470	545
FUN_07506	CLAGR_005526-RA	1	65	CLAGR_005525-RA	101	889
FUN_06042	CLAGR_003035-RA	12	452	CLAGR_003914-RA	509	629
FUN_09721	CLAGR_000624-RA	1	439	CLAGR_009331-RA	592	1036
FUN_09721	CLAGR_009332-RA	1	434	CLAGR_009331-RA	592	1036
FUN_03230	CLAGR_006653-RA	1	318	CLAGR_006652-RA	396	1193
FUN_03230	CLAGR_006653-RA	1	318	CLAGR_008875-RA	521	671
FUN_07182	CLAGR_006244-RA	25	61	CLAGR_006245-RA	66	524
FUN_05792	CLAGR_000109-RA	3	117	CLAGR_000110-RA	161	408
FUN_07041	CLAGR_000162-RA	1	56	CLAGR_007333-RA	73	851
FUN_01426	CLAGR_009169-RA	12	95	CLAGR_009170-RA	138	490
FUN_04144	CLAGR_003178-RA	7	1735	CLAGR_007860-RA	2485	2647
FUN_04144	CLAGR_009968-RA	1039	2462	CLAGR_007860-RA	2485	2647

FUN_02745	CLAGR_008010-RA	1	510	CLAGR_008011-RA	543	650
FUN_01437	CLAGR_006180-RA	9	211	CLAGR_006182-RA	291	489
FUN_03046	CLAGR_000936-RA	9	216	CLAGR_004022-RA	233	383
FUN_02248	CLAGR_010339-RA	129	272	CLAGR_001960-RA	1039	1435
FUN_02248	CLAGR_010339-RA	129	272	CLAGR_008758-RA	1085	1445
FUN_03752	CLAGR_005276-RA	10	742	CLAGR_005277-RA	872	1594
FUN_02486	CLAGR_005954-RA	44	795	CLAGR_005953-RA	873	1794
FUN_05417	CLAGR_006718-RA	334	1355	CLAGR_005357-RA	1847	2225
FUN_02566	CLAGR_010933-RA	1	133	CLAGR_008989-RA	196	853
FUN_04000	CLAGR_009110-RA	1	483	CLAGR_009109-RA	556	826
FUN_09476	CLAGR_001950-RA	1	230	CLAGR_000982-RA	707	783
FUN_08680	CLAGR_009335-RA	734	1880	CLAGR_005503-RA	2006	2155
FUN_08680	CLAGR_011257-RA	1815	1924	CLAGR_005503-RA	2006	2155
FUN_07723	CLAGR_008673-RA	1	192	CLAGR_008672-RA	285	574
FUN_07724	CLAGR_011062-RA	1	40	CLAGR_010079-RA	225	299
FUN_05735	CLAGR_000328-RA	41	124	CLAGR_000327-RA	171	293
FUN_02410	CLAGR_002952-RA	1	122	CLAGR_002951-RA	241	562
FUN_00169	CLAGR_004123-RA	1	441	CLAGR_004122-RA	468	1094
FUN_06037	CLAGR_010984-RA	769	947	CLAGR_011236-RA	1629	1876
FUN_09582	CLAGR_004584-RA	9	352	CLAGR_004585-RA	449	600
FUN_08415	CLAGR_002785-RA	74	175	CLAGR_001044-RB	176	530
FUN_08415	CLAGR_002785-RA	74	175	CLAGR_002786-RA	203	594
FUN_08224	CLAGR_005645-RA	1	93	CLAGR_005644-RA	127	281
FUN_00531	CLAGR_007523-RA	193	413	CLAGR_003448-RA	484	1097
FUN_06453	CLAGR_006824-RA	1	100	CLAGR_006823-RA	109	529
FUN_03523	CLAGR_002684-RA	135	247	CLAGR_002683-RA	263	371
FUN_09358	CLAGR_006334-RA	32	382	CLAGR_006333-RA	387	577
FUN_01637	CLAGR_005101-RA	1	306	CLAGR_005096-RA	340	662
FUN_05257	CLAGR_002264-RA	5	1707	CLAGR_001747-RA	1785	2119
FUN_04736	CLAGR_006069-RA	5	64	CLAGR_006068-RA	73	317
FUN_00342	CLAGR_001111-RA	20	263	CLAGR_002142-RA	366	1284
FUN_00342	CLAGR_001111-RA	20	263	CLAGR_004654-RA	1030	1182
FUN_00342	CLAGR_001111-RA	20	263	CLAGR_009526-RA	325	1565
FUN_00342	CLAGR_001111-RA	20	263	CLAGR_010900-RA	1234	1512
FUN_00342	CLAGR_004367-RA	7	272	CLAGR_002142-RA	366	1284
FUN_00342	CLAGR_004367-RA	7	272	CLAGR_004654-RA	1030	1182
FUN_00342	CLAGR_004367-RA	7	272	CLAGR_009526-RA	325	1565
FUN_00342	CLAGR_004367-RA	7	272	CLAGR_010900-RA	1234	1512
FUN_01456	CLAGR_006167-RA	251	623	CLAGR_006166-RA	700	1448

FUN_08990	CLAGR_002085-RA	6	192	CLAGR_002086-RA	246	615
FUN_01900	CLAGR_009280-RA	1	56	CLAGR_009281-RA	73	413
FUN_05087	CLAGR_001587-RA	1	74	CLAGR_001586-RA	98	819
FUN_05675	CLAGR_000768-RA	251	345	CLAGR_002995-RA	488	745
FUN_05675	CLAGR_000768-RA	251	345	CLAGR_003006-RA	614	745
FUN_03918	CLAGR_007161-RA	356	508	CLAGR_011129-RA	510	636
FUN_07800	CLAGR_003500-RA	1	155	CLAGR_003501-RA	303	654
FUN_07693	CLAGR_001522-RA	38	348	CLAGR_001521-RA	478	1481
FUN_02106	CLAGR_008614-RA	22	52	CLAGR_008613-RA	63	186
FUN_03591	CLAGR_008422-RA	59	268	CLAGR_009701-RA	356	409
FUN_03451	CLAGR_007807-RA	28	274	CLAGR_000337-RA	387	615
FUN_03451	CLAGR_010693-RA	39	309	CLAGR_000337-RA	387	615
FUN_03451	CLAGR_007807-RA	28	274	CLAGR_004095-RA	338	567
FUN_03451	CLAGR_007807-RA	28	274	CLAGR_010574-RA	387	564
FUN_03451	CLAGR_010693-RA	39	309	CLAGR_004095-RA	338	567
FUN_03451	CLAGR_010693-RA	39	309	CLAGR_010574-RA	387	564
FUN_01311	CLAGR_001897-RA	123	283	CLAGR_001896-RA	390	575
FUN_09083	CLAGR_006210-RA	1	269	CLAGR_006211-RA	289	472
FUN_09083	CLAGR_006210-RA	1	269	CLAGR_011195-RA	664	798
FUN_09083	CLAGR_006211-RA	289	472	CLAGR_011195-RA	664	798

Table A-6: The inverted repeats (IR) found in the 5 Lecanoromycetes, along with median length and G/C content.

Taxon	Number of IR	Median IR length	Median % G/C in IR	Genomewide % G/C
Lasallia pustulata	1,464	819.5	50.54	51.34
Umbilicaria muehlenbergii	1,992	541	32.25	46.82
Cladonia grayi	29,396	807	11.13	44.37
Usnea florida	2643	450	24.70	42.71
Xanthoria parietina	670	2050.5	40.47	49.73

Table A-7: The 101 taxa that were used to find orthologs to the LCA_{Lec} genes that were privately lost in *L. pustulata*.

Kingdom	Phylum	Family	Species
Fungi	Ascomycota	Argynnaceae	*Lepidopterella palustris*
Fungi	Ascomycota	Arthrodermataceae	*Microsporum canis*
Fungi	Ascomycota	Arthrodermataceae	*Nannizzia gypsea CBS 118893*
Fungi	Ascomycota	Arthrodermataceae	*Trichophyton equinum*
Fungi	Ascomycota	Arthrodermataceae	*Trichophyton verrucosum*
Fungi	Ascomycota	Aspergillaceae	*Aspergillus clavatus*
Fungi	Ascomycota	Aspergillaceae	*Aspergillus nidulans*
Fungi	Ascomycota	Aspergillaceae	*Aspergillus oryzae*
Fungi	Ascomycota	Aspergillaceae	*Aspergillus ruber*
Fungi	Ascomycota	Aspergillaceae	*Aspergillus terreus*
Fungi	Ascomycota	Aspergillaceae	*Penicillium chrysogenum*
Fungi	Ascomycota	Aulographaceae	*Aulographum hederae*
Fungi	Ascomycota	Botryosphaeriaceae	*Botryosphaeria dothidea*
Fungi	Ascomycota	Botryosphaeriaceae	*Macrophomina phaseolina MS6*
Fungi	Ascomycota	Chaetomiaceae	*Chaetomium globosum*
Fungi	Ascomycota	Chaetomiaceae	*Thielavia terrestris*
Fungi	Ascomycota	Cladoniaceae	*Cladonia grayi*
Fungi	Ascomycota	Debaryomycetaceae	*Candida albicans*
Fungi	Ascomycota	Debaryomycetaceae	*Candida dubliniensis*
Fungi	Ascomycota	Debaryomycetaceae	*Debaryomyces hansenii*
Fungi	Ascomycota	Gloniaceae	*Cenococcum geophilum 1.58*
Fungi	Ascomycota	Hypocreaceae	*Trichoderma atroviride*
Fungi	Ascomycota	Hypocreaceae	*Trichoderma reesei*
Fungi	Ascomycota	Hysteriaceae	*Hysterium pulicare*
Fungi	Ascomycota	Hysteriaceae	*Rhytidhysteron rufulum*
Fungi	Ascomycota	Mycosphaerellaceae	*Passalora fulva*
Fungi	Ascomycota	Mycosphaerellaceae	*Sphaerulina musiva SO2202*
Fungi	Ascomycota	Mycosphaerellaceae	*Sphaerulina populicola*
Fungi	Ascomycota	Mycosphaerellaceae	*Zasmidium cellare*
Fungi	Ascomycota	Mycosphaerellaceae	*Zymoseptoria tritici*
Fungi	Ascomycota	Myriangiaceae	*Myriangium duriaei CBS 260.36*
Fungi	Ascomycota	Nectriaceae	*Fusarium graminearum*
Fungi	Ascomycota	Parmeliaceae	*Usnea florida*
Fungi	Ascomycota	Patellariaceae	*Patellaria atrata*
Fungi	Ascomycota	Piedraiaceae	*Piedraia hortae*

Fungi	Ascomycota	Plectosphaerellaceae	*Verticillium dahliae*
Fungi	Ascomycota	Pleomassariaceae	*Pleomassaria siparia*
Fungi	Ascomycota	Pleosporaceae	*Bipolaris maydis*
Fungi	Ascomycota	Pleosporaceae	*Bipolaris maydis ATCC 48331*
Fungi	Ascomycota	Pleosporaceae	*Bipolaris sorokiniana ND90Pr*
Fungi	Ascomycota	Schizosaccharomycetaceae	*Schizosaccharomyces cryophilus OY26*
Fungi	Ascomycota	Sordariaceae	*Neurospora crassa*
Fungi	Ascomycota	Sordariaceae	*Neurospora discreta*
Fungi	Ascomycota	Sporormiaceae	*Sporormia fimetaria*
Fungi	Ascomycota	Teloschistaceae	*Xanthoria parietina*
Fungi	Ascomycota	Trichocomaceae	*Thermomyces*
Fungi	Ascomycota	Trypetheliaceae	*Trypethelium eluteriae*
Fungi	Ascomycota	Umbilicariaceae	*Umbilicaria muehlenbergii*
Fungi	Ascomycota	Zopfiaceae	*Zopfia rhizophila*
Fungi	Basidiomycota	Cryptococcaceae	*Cryptococcus neoformans var. neoformans*
Fungi	Basidiomycota	Dacryobolaceae	*Postia placenta*
Fungi	Basidiomycota	Psathyrellaceae	*Coprinopsis cinerea*
Fungi	Basidiomycota	Schizophyllaceae	*Schizophyllum commune*
Fungi	Basidiomycota	Serpulaceae	*Serpula lacrymans*
Fungi	Basidiomycota	Tricholomataceae	*Laccaria bicolor*
Fungi	Mucoromycota	Phycomycetaceae	*Phycomyces blakesleeanus*
Metazoa	Arthropoda	Apidae	*Apis mellifera*
Metazoa	Arthropoda	Bombycidae	*Bombyx mori*
Metazoa	Arthropoda	Culicidae	*Aedes aegypti*
Metazoa	Arthropoda	Culicidae	*Anopheles gambiae*
Metazoa	Arthropoda	Drosophilidae	*Drosophila ananassae*
Metazoa	Arthropoda	Drosophilidae	*Drosophila erecta*
Metazoa	Arthropoda	Drosophilidae	*Drosophila grimshawi*
Metazoa	Arthropoda	Drosophilidae	*Drosophila melanogaster*
Metazoa	Arthropoda	Drosophilidae	*Drosophila mojavensis*
Metazoa	Arthropoda	Drosophilidae	*Drosophila persimilis*
Metazoa	Arthropoda	Drosophilidae	*Drosophila pseudoobscura*
Metazoa	Arthropoda	Drosophilidae	*Drosophila sechellia*
Metazoa	Arthropoda	Drosophilidae	*Drosophila simulans*
Metazoa	Arthropoda	Drosophilidae	*Drosophila virilis*
Metazoa	Arthropoda	Drosophilidae	*Drosophila willistoni*
Metazoa	Arthropoda	Drosophilidae	*Drosophila yakuba*
Metazoa	Chordata	Cionidae	*Ciona intestinalis*
Metazoa	Cnidaria	Edwardsiidae	*Nematostella vectensis*

Ochrophyta	Phaeophyceae	Ectocarpaceae	*Ectocarpus siliculosus*
Viridiplantae	Chlorophyta	Coccomyxaceae	*Coccomyxa subellipsoidea C-169*
Viridiplantae	Chlorophyta	Mamiellaceae	*Micromonas commoda*
Viridiplantae	Streptophyta	Brassicaceae	*Brassica rapa*
Viridiplantae	Streptophyta	Rutaceae	*Citrus sinensis*
-	-	-	*Phytophthora sojae*
-	-	-	*Polysphondylium pallidum*
-	Apusozoa	Apusomonadidae	*Thecamonas trahens*
-	Bacillariophyta	Bacillariaceae	*Fragilariopsis*
-	Bacillariophyta	Phaeodactylaceae	*Phaeodactylum tricornutum*
-	Bacillariophyta	Thalassiosiraceae	*Thalassiosira pseudonana*
-	Ciliophora	Parameciidae	*Paramecium tetraurelia*
-	Euglenozoa	Trypanosomatidae	*Leishmania infantum*
-	Euglenozoa	Trypanosomatidae	*Leishmania major strain Friedlin*
-	Percolozoa	Vahlkampfiidae	*Naegleria gruberi*
Archaea	Euryarchaeota	Methanocorpusculaceae	*Methanocorpusculum labreanum Z*
Eubacteria	Cyanobacteria	Acaryochloridaceae	*Acaryochloris marina*
Eubacteria	Cyanobacteria	Acaryochloridaceae	*Acaryochloris marina MBIC11017*
Eubacteria	Cyanobacteria	Cyanothecaceae	*Cyanothece sp. PCC 7425*
Eubacteria	Cyanobacteria	Nostocaceae	*Anabaena variabilis ATCC 29413*
Eubacteria	Cyanobacteria	Prochloraceae	*Prochlorococcus marinus str. MIT 9303*
Eubacteria	Cyanobacteria	Prochloraceae	*Prochlorococcus marinus str. MIT 9313*
Eubacteria	Cyanobacteria	Rivulariaceae	*Calothrix sp. PCC 7507*
Eubacteria	Cyanobacteria	Synechococcaceae	*Cyanobium gracile PCC 6307*
Eubacteria	Cyanobacteria	Synechococcaceae	*Synechococcus sp. JA-2-3B'a(2-13)*
Eubacteria	Proteobacteria	Cardiobacteriaceae	*Dichelobacter nodosus VCS1703A*
Eubacteria	Proteobacteria	Polyangiaceae	*Sorangium cellulosum So ce56*

Table A-8: GO-terms that are significantly enriched among genes lost in the last common ancestor of the Lecanoromycetes, along with their description and false-discovery rate corrected p-values.

GO term	GO description	FDR-corrected p-value
GO:0006351	transcription, DNA-templated	0
GO:0006357	regulation of transcription from RNA polymerase II promoter	0
GO:0097659	nucleic acid-templated transcription	0
GO:0032774	RNA biosynthetic process	0
GO:0019219	regulation of nucleobase-containing compound metabolic process	0
GO:0006355	regulation of transcription, DNA-templated	0
GO:0051252	regulation of RNA metabolic process	0
GO:0055085	transmembrane transport	0
GO:1903506	regulation of nucleic acid-templated transcription	0
GO:2001141	regulation of RNA biosynthetic process	0
GO:0055114	oxidation-reduction process	0
GO:0008150	biological_process	0
GO:0008152	metabolic process	0
GO:0010468	regulation of gene expression	0
GO:2000112	regulation of cellular macromolecule biosynthetic process	0
GO:0010556	regulation of macromolecule biosynthetic process	0
GO:0000272	polysaccharide catabolic process	0
GO:0051171	regulation of nitrogen compound metabolic process	0
GO:0031326	regulation of cellular biosynthetic process	0
GO:0009889	regulation of biosynthetic process	0
GO:0034654	nucleobase-containing compound biosynthetic process	0
GO:0044710	single-organism metabolic process	0
GO:0016052	carbohydrate catabolic process	0
GO:0005975	carbohydrate metabolic process	0
GO:0005976	polysaccharide metabolic process	0
GO:0060255	regulation of macromolecule metabolic process	0
GO:0019438	aromatic compound biosynthetic process	0.002
GO:0031323	regulation of cellular metabolic process	0.004
GO:0010410	hemicellulose metabolic process	0.004
GO:0008643	carbohydrate transport	0.004
GO:0080090	regulation of primary metabolic process	0.004
GO:1901362	organic cyclic compound biosynthetic process	0.004
GO:0019222	regulation of metabolic process	0.004
GO:0018130	heterocycle biosynthetic process	0.01
GO:0045491	xylan metabolic process	0.02
GO:0045493	xylan catabolic process	0.02

Table A-9: Rate of LCA_{Lec} gene loss for the Lecanoromycetes. Given the lack of RNAseq data for *U. muehlenbergii*, no rate could be calculated.

	Lasallia pustulata	*Cladonia grayi*	*Usnea florida*	*Xanthoria parietina*	*Umbilicaria muehlenbergii*
Lost LCA_{Lec} per mya	0.47	0.30	0.26	0.40	N/A

Table A-10: The putative functional annotation for the 28 private, high-confidence LCA_{Lec} loss candidates for *L. pustulata*. The four genes in which an ortholog could be identified in *L. hispanica* are highlighted in yellow. All genes found in these LCA_{Lec} HOGs were assigned to KEGG orthologous groups (KO) and annotated with Gene Ontology (GO) terms and Pfam domains.

HOG	KO identifier	GO terms	Pfam domains
HOG962	-	EC:3.2.1.18, hydrolase activity (GO:0016787), hydrolase activity, acting on glycosyl bonds (GO:0016798), exo-alpha-sialidase activity (GO:0004308)	DUF346 (PF03984.11)
HOG3669	-	hydrolase activity (GO:0016787), catalytic activity (GO:0003824)	Abhydrolase_1 (PF00561.18)
HOG6657	-	-	-
HOG2339	-	oxidation-reduction process (GO:0055114), metabolic process (GO:0044710), oxidoreductase activity (GO:0016491), regulation of biological process (GO:0050789), metal ion binding (GO:0046872)	2OG-FeII_Oxy (PF03171.18), DIOX_N (PF14226.4)
HOG985	-	-	Peptidase_S8 (PF00082.20)
HOG5535	-	-	HET (PF06985.9)
HOG6976	-	zinc ion binding (GO:0008270), EC:1.1.1.1, oxidation-reduction process (GO:0055114), alcohol dehydrogenase (NAD) activity (GO:0004022), oxidoreductase activity (GO:0016491)	ADH_zinc_N (PF00107.24), ADH_N (PF08240.10)
HOG3285	-	-	CorA (PF01544.16)
HOG4477	-	membrane (GO:0016020)	MAPEG (PF01124.16)
HOG3066	-	-	-
HOG826	-	-	Methyltransf_25 (PF13649.4), Methyltransf_11 (PF08241.10)
HOG22	-	-	Pkinase (PF00069.23)
HOG5972	-	nuclease activity (GO:0004518), nucleic acid metabolic process (GO:0090304)	Exo_endo_phos (PF03372.21)
HOG5783	-	-	-
HOG3759	K06688	nucleus (GO:0005634), cytoplasm (GO:0005737), ubiquitin protein ligase activity (GO:0061630), regulation of mitotic metaphase/anaphase transition (GO:0030071), ubiquitin protein ligase binding	UQ_con (PF00179.24)

		(GO:0031625), protein polyubiquitination (GO:0000209), anaphase-promoting complex-dependent catabolic process (GO:0031145), ATP binding (GO:0005524)	
HOG4273	-	-	Methyltransf_23 (PF13489.4)
HOG6659	-	-	-
HOG1675	-	metabolic process (GO:0008152), S-adenosylmethionine-dependent methyltransferase activity (GO:0008757), cytoplasm (GO:0005737), methylation (GO:0032259), methyltransferase activity (GO:0008168), integral component of membrane (GO:0016021), membrane (GO:0016020)	Methyltransf_11 (PF08241.10)
HOG911	-	-	-
HOG1265	-	-	-
HOG493	-	-	-
HOG2532	-	transferase activity (GO:0016740)	Fructosamin_kin (PF03881.12)
HOG1290	-	catalytic activity (GO:0003824)	FAD_binding_4 (PF01565.21), BBE (PF08031.10)
HOG4614	K00079	oxidoreductase activity (GO:0016491), oxidation-reduction process (GO:0055114)	adh_short (PF00106.23)
HOG3350	-	-	-
HOG5287	-	metabolic process (GO:0008152), catalytic activity (GO:0003824), transferase activity (GO:0016740)	Ketoacyl-synt_C (PF02801.20), KR (PF08659.8), PS-DH (PF14765.4), KAsynt_C_assoc (PF16197.3), Acyl_transf_1 (PF00698.19), ADH_zinc_N_2 (PF13602.4), ADH_N (PF08240.10), ketoacyl-synt (PF00109.24), ADH_zinc_N (PF00107.24)
HOG588	-	nucleus (GO:0005634), mitochondrion (GO:0005739), intracellular part (GO:0044424), cytosol (GO:0005829)	-
HOG4755	-	-	Ank_2 (PF12796.5)

Figures

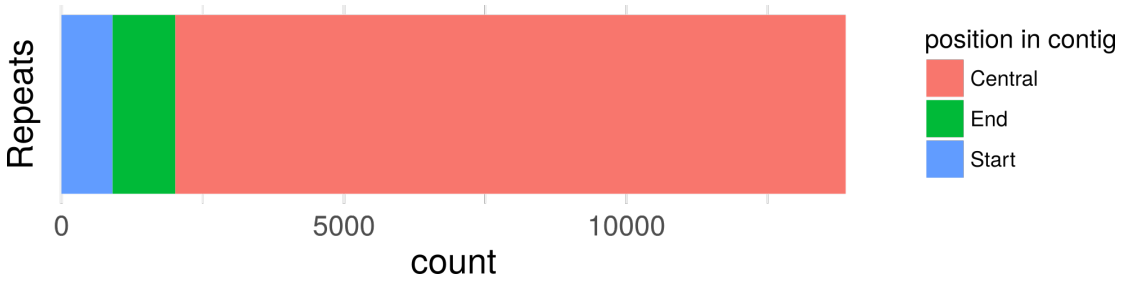


Figure A-1: Positions of the repetitive elements that were predicted in the *C. grayi* pseudogenome. If repeats overlap within the first or last 100bp of the original contig borders they were classified accordingly. Otherwise they were classified to be located in the central part of the contigs.

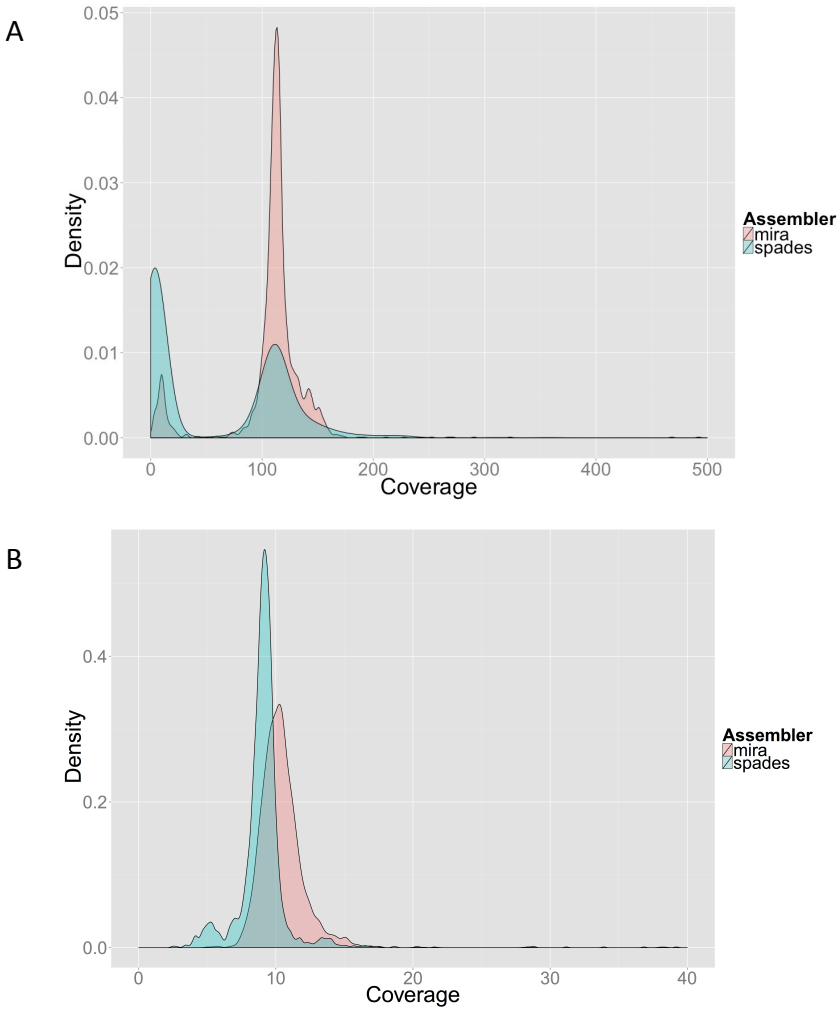


Figure A-2: Readcoverages for the *MIRA* and *SPAdes* assemblies of the genomes of *L. pustulata* (A) and *Trebouxia sp.* (B).

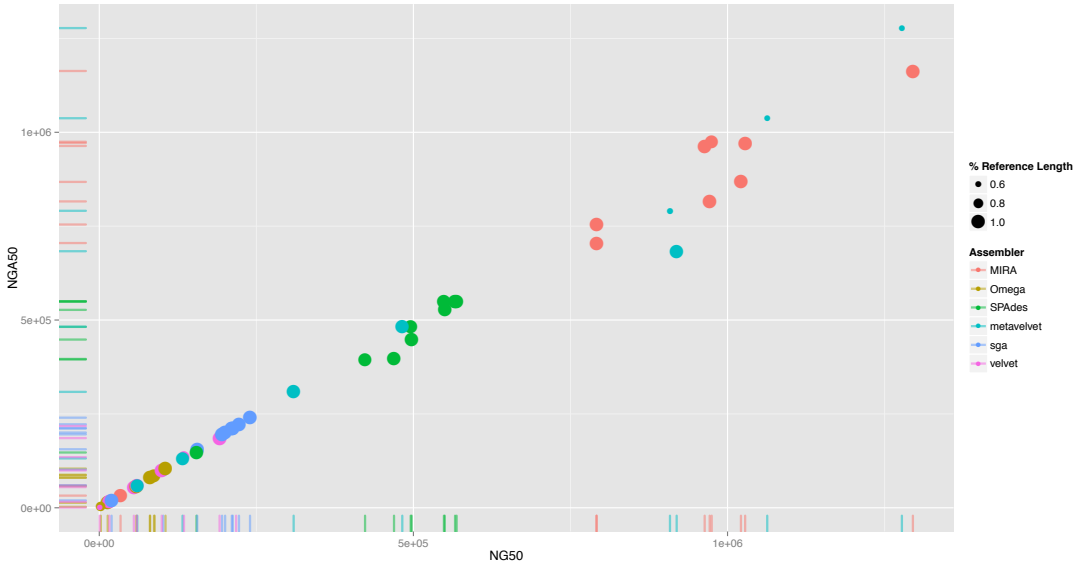


Figure A-3: NG50 vs NGA50 for the assemblies across all 11 data sets and the six assemblers. Misassemblies generated by wrongly joining non-adjacent sections of the reference genome will decrease the NGA50 value compared to the NG50 value. Different assemblers are represented by color. The size of the points reflects the percentage of the reference genome covered by the assembly. The rug plot along the x- and the y-axis identifies the exact values for the individual assemblers.

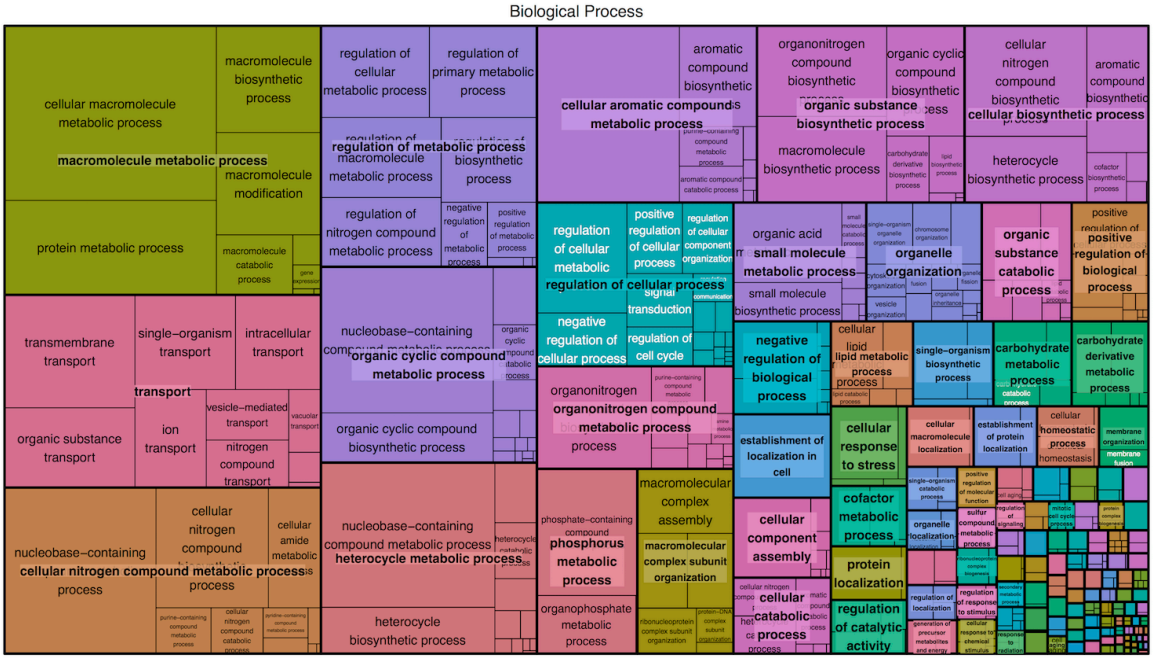


Figure A-4: Treemap of Biological Process Gene Ontology terms annotated in *Lasallia pustulata*.

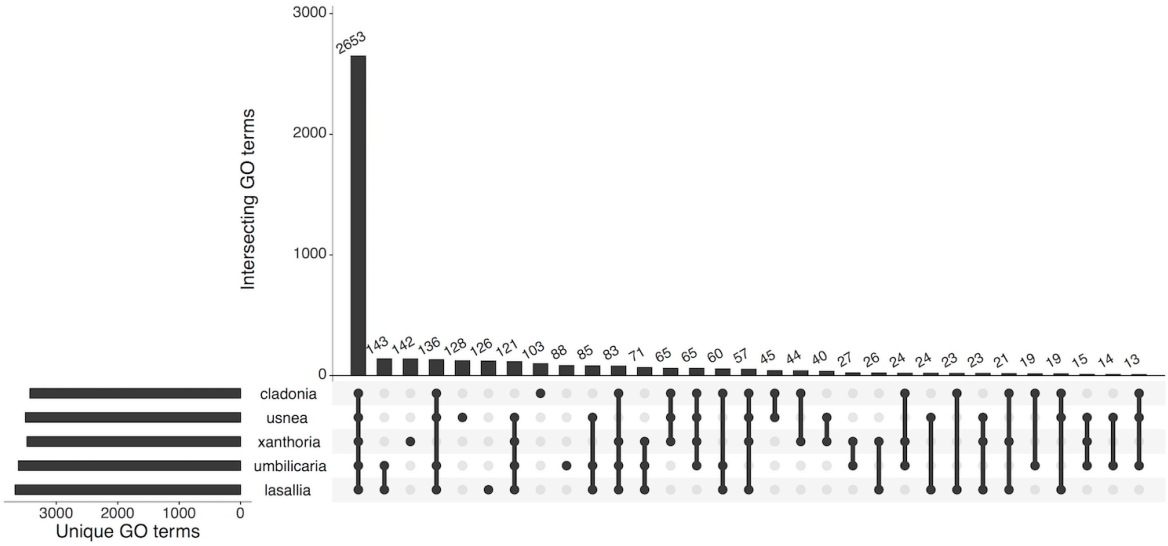


Figure A-5: Intersections amongst the unique GO terms that were assigned to the five Lecanoromycetes species.

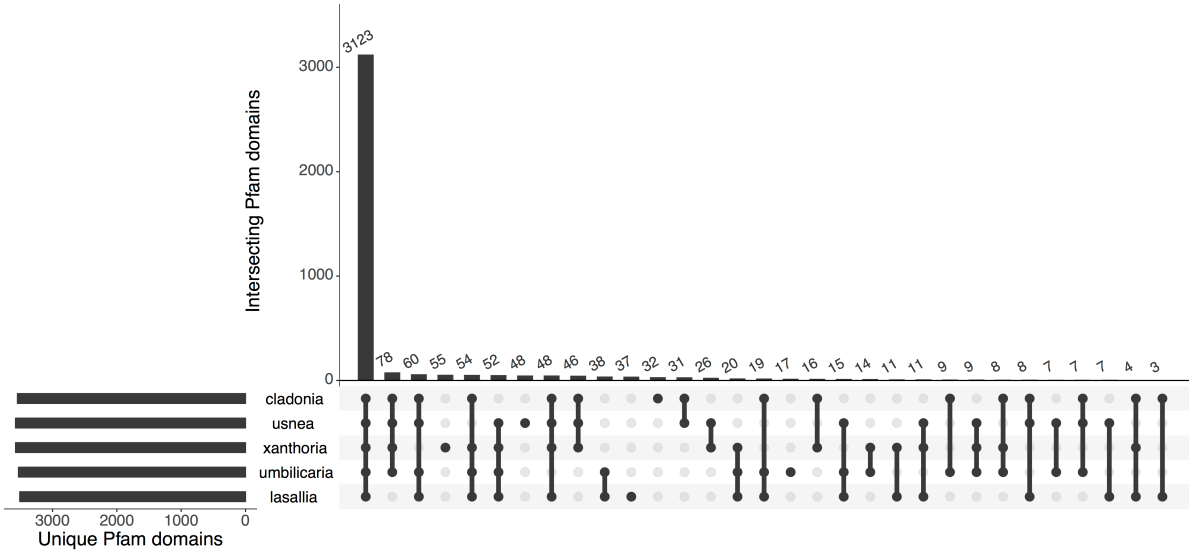


Figure A-6: Intersections amongst the unique Pfam domains annotated to the five Lecanoromycetes species.

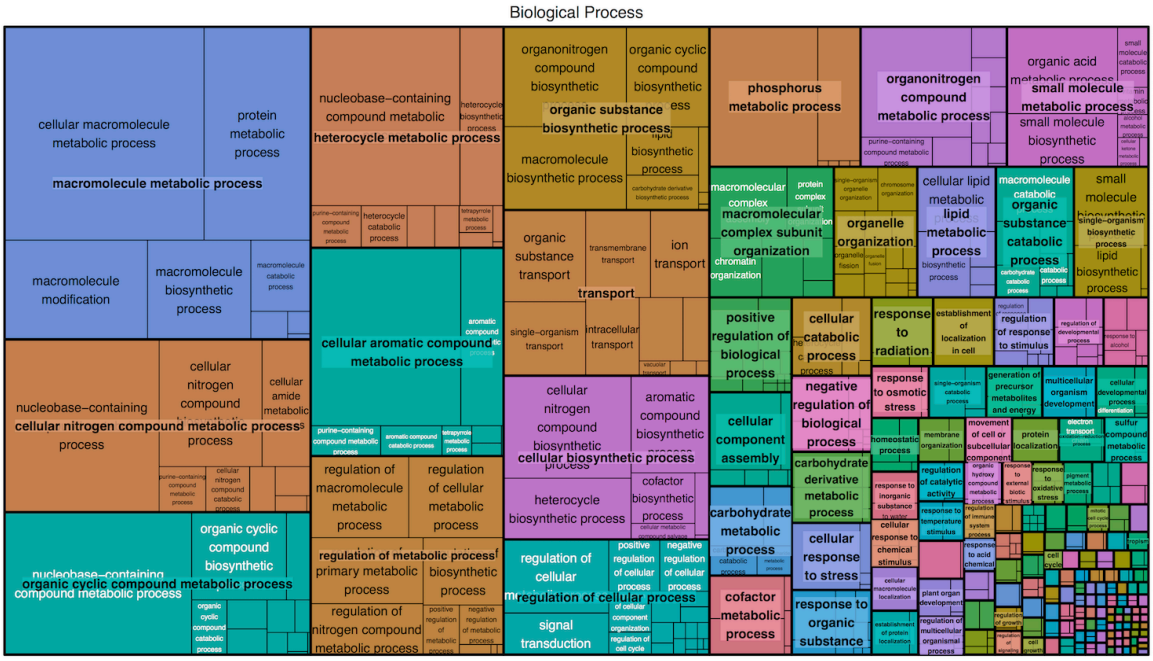


Figure A-7: Treemap of the Biological Process Gene Ontology terms assigned to the annotated genes of *Trebouxia sp.*

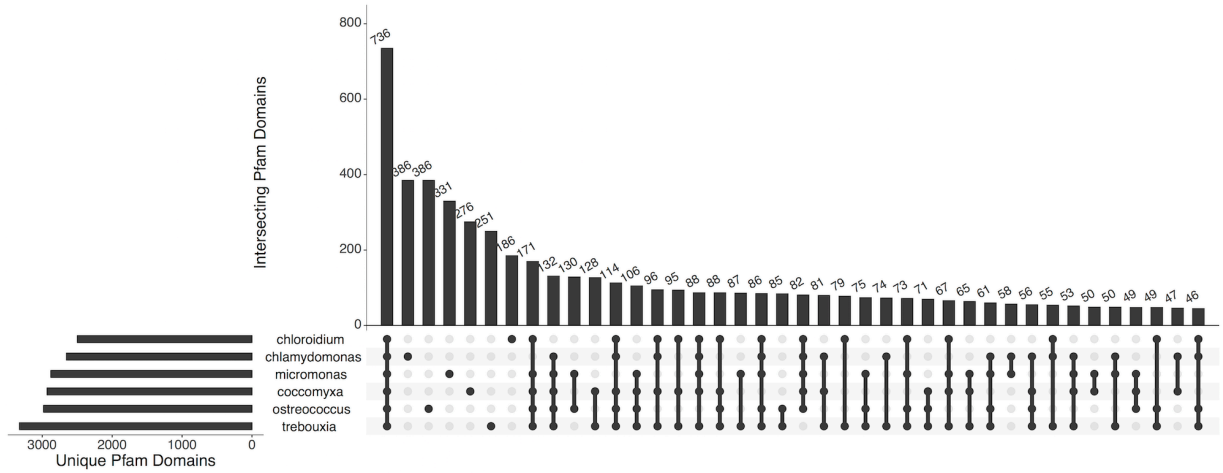


Figure A-8: Intersections among the unique Pfam domains found between *Trebouxia sp.* and five other Chlorophyta.

Figure A-9: Insertion/Deletion errors (purple bars) that are consistently observed in mapping of the read pairs of both the Illumina read- and mate-pair libraries compared to the assembly.

Figure A-10: Regions in which the Illumina coverage drops below 10x while the PacBio coverage remains constant. X-axis shows the percent G/C for the region, the Y-Axis shows the length. The color code shows whether the region overlaps a predicted inverted repeat (blue) or not (red). The marginal plots show the density of length and G/C for the two categories.

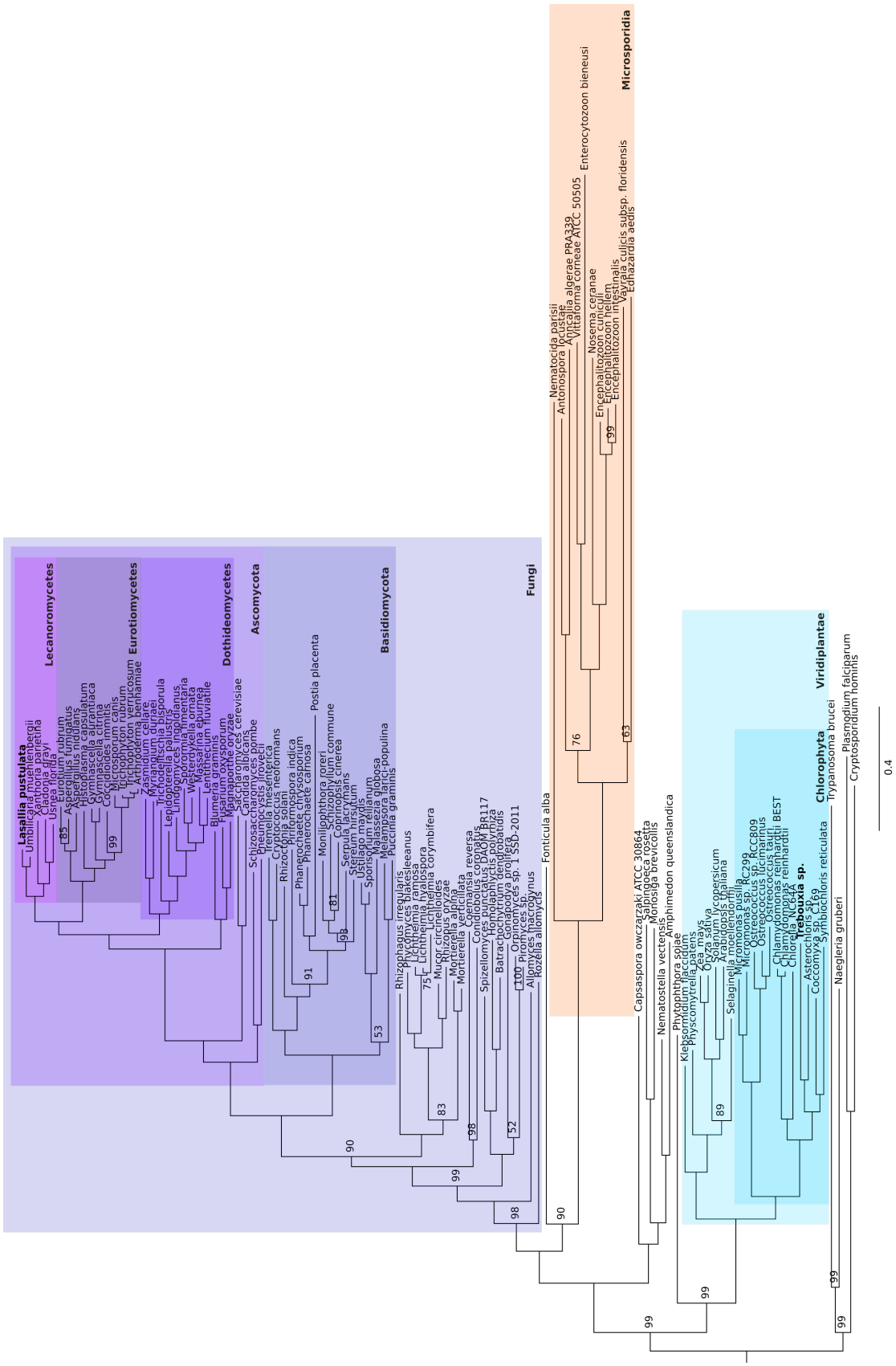
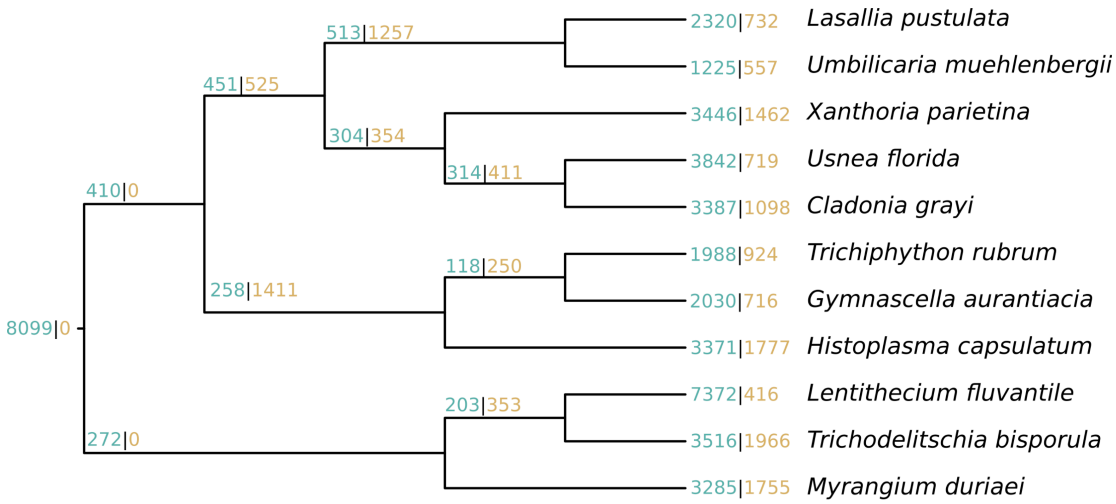


Figure A-11: Complete maximum likelihood tree that was used to place the *L. pustulata* and *Trebouxia sp.*. Node labels give the bootstrap support. Only bootstrap values <100 are shown.



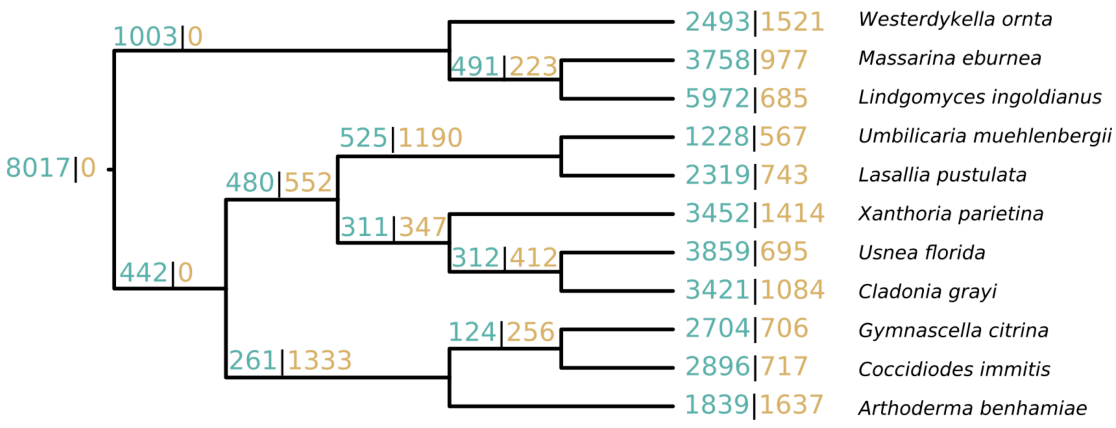


Figure A-12: Gene gains and losses as predicted by two further, non-overlapping sets of Eurotiomycetes and Dothideomycetes

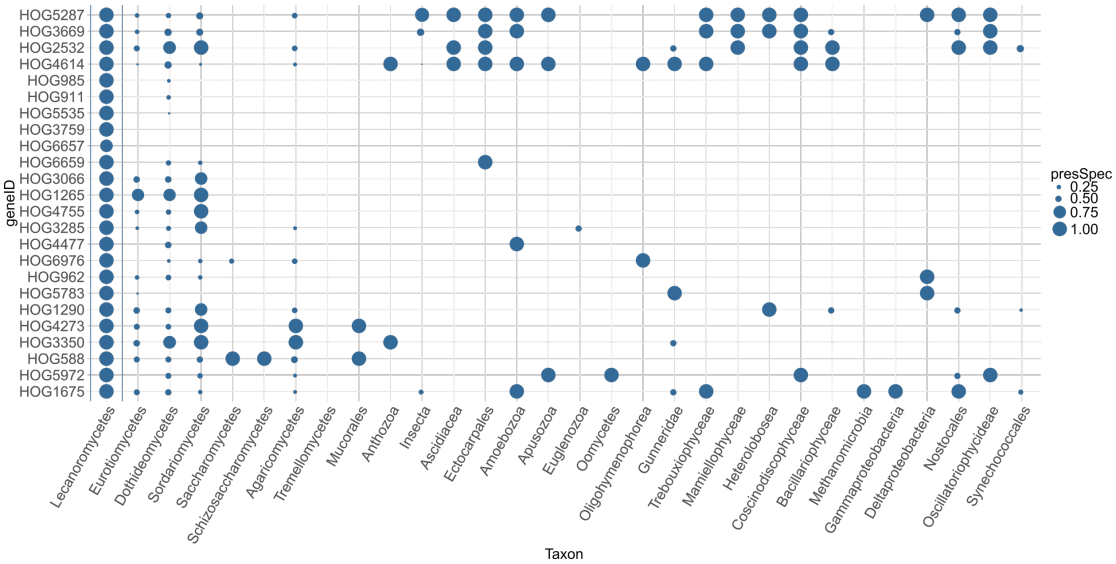


Figure A-13: The phylogenetic profile for the further 24 LCA_{Lec} genes that are privately lost in L. pustulata, but could not be verified by a found ortholog in the draft genome of *L. hispanica*.

Acknowledgements

Cognition is […] not an individual process of any theoretical "particular consciousness." Rather it is the result of a social activity, since the existing stock of knowledge exceeds the range available to any one individual.

> LUDWIK FLECK, *ENTSTEHUNG UND ENTWICKLUNG EINER WISSENSCHAFTLICHEN TATSACHE. EINFÜHRUNG IN DIE LEHRE VOM DENKSTIL UND DENKKOLLEKTIV* (1935)

Finishing a PhD thesis is more than the literal and proverbial "*blood, toil, tears and sweat*". Foremost it is a lot of teamwork and would not have been possible without the contributions and support of many individuals and communities. I am indebted to each of them.

I am grateful for each member (current and former) of the Applied Bioinformatics group: Special thanks go to Ingo Ebersberger, my mentor and supervisor who showed me the ropes of bioinformatics, engaged in many useful discussions about evolutionary biology at large, provided ample of useful feedback over the years, and gave me the freedom to teach, learn and grow. Furthermore, I have to thank Anne Hänel, who was always patient enough to help me with my phobia of all things bureaucracy and filling forms and always had an open ear for problems, academic and non-academic alike. Stefan Biermann provided the best tech support one could wish for – at the most unlikely hours even – and trusted me to not accidentally screw up the system with *sudo* privileges (the password is *ChangeMe1234*). Vinh Tran and Bardya Djahanschiri provided their assistance and code, and helped with numerous analyses over the years. Holger Bergmann, Arpit Jain and Sachli Zafari provided additional support with fruitful discussions, new insights and ideas for further scientific explorations. Jan Koch, Laura Kiesewetter and Simonida Zehr not only helped to keep me caffeinated, but also provided the best moral support I could wish for; I will dearly miss our coffee breaks.

Without external collaborations this work would not have been possible: I thank all the members of group of Imke Schmitt at the *Senckenberg Biodiversity and Climate Research Centre*. They collected the samples and performed the central wet lab work. I am especially grateful to Francesco Dal Grande, who always took time to answer any open questions I had. The Molecular Ecology group around Markus Pfenninger – who generously agreed to review this work – additionally was a great source of collaborations and scientific discussions. Thomas Hankeln and his group at the Institute of Molecular Genetics at the Johannes Gutenberg University in Mainz also kindly provided their expertise and time in the wet lab.

There are numerous additional communities that were instrumental in my learning and success: Thanks go to the whole bioinformatics open source community that is a central part of the wider field. Without their culture of sharing code, expertise and knowledge I would probably still be pipetting in a wet lab. Special

thanks go to Peter Cock (and all other Biopython contributors), Michael Crusoe, Titus Brown and Kai Blin. Thanks also to the whole bioinformatics crowd on Twitter that was instrumental in my education (there is no better way to learn about the field than discussing with Nick Loman, Mick Watson et al.), and all the authors who are putting their preprints up on *bioRxiv, peerj* etc. and were in a 100% of the cases more than willing to answer questions about their work.

Further thanks go to Philipp Bayer and Helge Rausch, who always had my back when it comes to software development and deployment and engaged in the occasional discussion about arcane details of even more arcane software problems. Philipp Bayer and Madeleine Price Ball did not only offer their precious time for cheering me up, but also for reading a first draft of this thesis. Alexandra Elbakyan provided unlimited help in making sure that this thesis would not be heavily undercited.

My parents were and are instrumental in all my successes and achievements. Without their ongoing support, strong belief in me, and all the freedom they ever gave me and made possible for me I would not be where I am today.

Last but not least there are the people who – over the years – were key in shaping who I am today and strove to help me become a better person. Thanks Silke, Julia and Athina.

Curriculum Vitae

BASTIAN GRESHAKE TZOVARAS
(BORN 1985-01-10 IN MÜNSTER, GERMANY)

ADDRESS	Hafeninsel 19, 63067 Offenbach am Main, Germany
EMAIL	bgreshake@googlemail.com
WEB	ruleofthirds.de
PHONE	+49 176 213 044 66

EDUCATION

since 2013 — **PhD in Bioinformatics, Goethe University, Frankfurt am Main, Germany.**
Title: *Characterizing the hologenome of Lasallia pustulata and tracing genomic footprints of lichenization*
Supervision: Prof. Dr. Ingo Ebersberger

2011-2013 — **MSc in Ecology & Evolution, Goethe University, Frankfurt am Main, Germany.**
Title: *Comparative transcriptome analysis in the genus* Radix
Supervision: PD Dr. Markus Pfenninger

2006-2010 — **BSc in Life Sciences, Westfälische Wilhelms-Universität, Münster, Germany**
Title: *Transcriptome analysis based on Next Generation Sequencing Data.*
Supervision: Prof. Dr. Erich Bornberg-Bauer & Dr. Philine Feulner

COMMUNITY WORK

since 2011 — **- Co-Founder of openSNP.org**
Creating an open data repository for personal genomics data sets as generated by Direct-To-Consumer genetic testing companies, hosts ~3,700+ data sets in Sept. 2017.

since 2016 — **- FORCE11**
Member of the *Scholarly Commons Workgroup* steering committee and organized a workshop in San Diego.

- Mozilla Science Lab
Mentored 4 people in the *Open Leadership Training Series* and taught in the *Working Open Workshops*.

- BIOINFORMATICS OPEN SOURCE CONFERENCE
Member of the organizing committee, co-organized the 200+ people COSI meeting in Prague at ISMB in 2017.

2016 **- GOOGLE SUMMER OF CODE**
Mentored 3 students realizing projects on openSNP

SINCE 2015 **- OPENCON**
Alumni of the *Open Data/Open Education/Open Access* group

- PART OF THE OPEN BIOINFORMATICS FOUNDATION

GRANTS, SCHOLARSHIPS & AWARDS

2017 **- YOUNG INVESTIGATORS AWARD OF THE *SAGE BIONETWORKS ASSEMBLY***

- SCHOLARSHIP *CREATIVE COMMONS GLOBAL SUMMIT*

2016 **- FELLOW OF *FORCE11***

2015 **- WINNER OF *THE WINNOWER* WRITING COMPETITION "*THE REWARDS OF OPEN SCIENCE*"**

2013 **- WINNER *GRANTS4APPS* BY BAYER PHARMACEUTICALS**
Grant for further the development and sustaining of *openSNP*. (5,000 €)

2012 **- BEST ORAL PRESENTATION AT 3RD *ANNUAL PHD SYMPOSIUM ON COMPUTATIONAL BIOLOGY AND INNOVATION*, DUBLIN**

2011 **- WINNER OF THE *PLOS/MENDELEY BINARY BATTLE***
First place for openSNP, awarded for the best/most creative use of the respective APIs (USD 10,000)

- WINNER OF THE WIKIMEDIA *WISSENSWERT* CONTEST
Granted to openSNP to get underrepresented minorities involved in open science and personal genomics (5,000 €)

SELECTED PUBLICATIONS

2017 - Bosman J, Bruo I, Chapman C, **Greshake Tzovaras B**, Jacobs N, Kramer B, Martone M, Murphy F, O'Donnell DP, Bar-Sinai M, Hagstrom S, Utley J, Veksler L. *The Scholarly Commons – principles and practices to guide research communication* OSF Preprints, doi:10.17605/OSF.IO/6C2XT

- Schell T, Feldmeyer B, Schmidt H, **Greshake B**, Tills O, Truebano M, Rundle SD, Paule J, Ebersberger I, Pfenninger M. *An annotated draft genome for Radix auricularia* Genome Biology and Evolution, (9)3, doi:10.1093/gbe/evx032

- Himmelstein, DS, Romero AR, McLaughlin SR, **Greshake Tzovaras B**, Greene CS. *Sci-Hub provides access to nearly all scholarly literature* PeerJ Preprints, 5:e3100v1 doi:10.7287/peerj.preprints.3100v1

- Haeusermann T*, **Greshake B***, Blasimme A, Irdam D, Richards M, Vayena E. *Open sharing of genomic data: who does it and why?* PLOS ONE 12(5): e0177158 doi:10.1371/journal.pone.0177158

- **Greshake B**. *Looking Into Pandora's Box: The Content Of Sci-Hub And Its Usage* F1000Research 6:541 doi: 10.12688/f1000research.11366.1

2016

- Grimm DB, Roqueiro D, Salome P, Kleeberger S, **Greshake B**, Zhu W, Liu C, Lippert C, Stegle O, Schölkopf B, Weigel D, Borgwardt K. *easyGWAS: A Cloud-Based Platform for Comparing the Results of Genome-Wide Association Studies The Plant Cell* vol. 29 no. 1 5-19

- Greshake B. *Correlating the Sci-Hub data with World Bank Indicators and Identifying Academic Use* The Winnower 4:e146485.57797 (2016). DOI: 10.15200/winn.146485.57797

- Greshake B, Zehr S, Dal Grande F, Meiser A, Schmitt I, Ebersberger I. *Potential and pitfalls of eukaryotic metagenome skimming: a test case for lichens* Molecular Ecology Resources (online ahead of print)

2015

- Vayena E, Brownsword R, Edwards SJ, **Greshake B**, Kahn JP, Ladher N, Montgomery J, O'Connor D, O'Neill O, Richards MP, Rid A, Sheehan M, Wicks P, Tasioulas J. *Research led by participants: a new social contract for a new kind of research* Journal of Medical Ethics, medethics-2015-102663

- Corpas M, Valdivia-Granda W*, Torres N*, **Greshake B***, Coletta A, Knaus A, Harrison AP, Cariaso M, Moran F, Nielsen F, Swan D, Weiss Sols DY, Krawitz P,

Schacherer P, Schols P, Yang H, Borry P, Glusman G, Robinson PN. *Crowdsourced direct-to-consumer genomic analysis of a family quartet* BMC Genomics 16 (1), 910

2014 - **Greshake B***, Bayer P*, Rausch H, Zimmer F, Reda J. *openSNP - a crowdsourced web resource for personal genomics*, PLOS ONE 9 (3), e89204

2013 - Schmidt PA, Balint M, **Greshake B**, Bandow C, Römbke J, Schmitt I. *Illumina metabarcoding of a soil fungal community* Soil Biology and Biochemistry 65, 128- 132

The full list can be found at http://orcid.org/0000-0002-9925-9623, * denotes shared authorship.

SELECTED TALKS/PANELS

2017 **- SAGE ASSEMBLY**
Seattle, invited closing remarks
- LGBTSTEMINAR
Sheffield, invited keynote.

2016 **- PERSONALIZED HEALTH IN THE DIGITAL AGE**
Geneva, *The Personal Data is Political,* invited keynote.
- BIOINFORMATICS OPEN SOURCE CONFERENCE
Orlando, *Growing and sustaining open source communities*, invited panel with Abigail Cabunoc Mayes, Jamie Whitacre, John Chilton and Natasha Wood
- SOCIETY FOR MOLECULAR BIOLOGY & EVOLUTION
Brisbane, *The genomic footprint of lichenization: comparative genomics of Lecanoromycetes*, accepted talk

2015 **- LIFT**
Basel, *Opening Up a Million Genomes (for Starters)*, invited keynote.
- INTERNATIONAL CONFERENCE ON GENOMICS
Shenzhen, *Transforming Direct-To-Consumer to Direct-To-Crowd Genetic Testing*, invited talk.
- CITIZEN SCIENCE
Zurich, Workshop at the ETH, invited talk.