Bioinformatic Analysis of Genomic and Transcriptomic Variation in Fungi

Gehrmann, Thies

DOI
10.4233/uuid:b7844e47-91c4-49a4-9178-d64ba4b45713

Publication date
2018

Document Version
Final published version

Citation (APA)
Gehrmann, T. (2018). *Bioinformatic Analysis of Genomic and Transcriptomic Variation in Fungi*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:b7844e47-91c4-49a4-9178-d64ba4b45713

Bioinformatic Analysis of Genomic and Transcriptomic Variation in Fungi

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Friday 6 April 2018 at 15:00 o'clock

by

Thies Gehrmann

Master of Science in Computer Science,
Leiden University, The Netherlands,
born in Hannover, Germany.

This dissertation has been approved by the
promotor: Prof. dr. ir. M.J.T. Reinders and
copromotor: Dr. T. Abeel

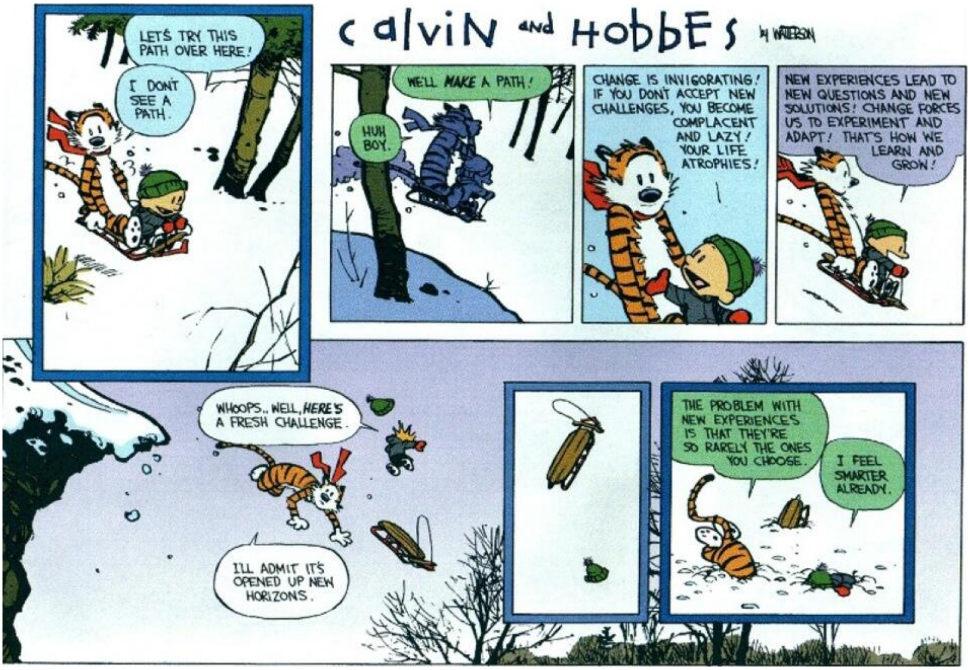
Composition of the doctoral committee:

Rector Magnificus,	Chairperson
Prof. dr. ir. M.J.T. Reinders,	Delft University of Technology, promotor
Dr. T. Abeel,	Delft University of Technology, copromotor
Independent members:	
Prof. dr. J. Pronk,	Delft University of Technology
Prof. dr. B. Snel,	Utrecht University
Prof. dr. P.A.C. ‘t Hoen	Radboud University Medical Center Nijmegen
Prof. dr. K. Vandepoele,	Ghent University
Prof. dr. H.A.B. Wösten,	Utrecht University, other member
Prof. dr. R.C.H.J. van Ham,	Delft University of Technology, reserve member

Parts of this thesis have received funding from The Netherlands Technology Foundation (STW), as part of STW Project 11615 (“Push the White Button”).

ISBN 978-94-6186-915-9
DOI 10.4233/uuid:b7844e47-91c4-49a4-9178-d64ba4b45713
Published by Uitgeverij BOXPress || Proefschriftmaken.nl

An electronic version of this dissertation is available at: `http://repository.tudelft.nl/`


Calvin and Hobbes
by WATTERSON
LET'S TRY THIS PATH OVER HERE!
I DON'T SEE A PATH.
WE'LL MAKE A PATH!
HUH BOY.
CHANGE IS INVIGORATING! IF YOU DON'T ACCEPT NEW CHALLENGES, YOU BECOME COMPLACENT AND LAZY! YOUR LIFE ATROPHIES!
NEW EXPERIENCES LEAD TO NEW QUESTIONS AND NEW SOLUTIONS! CHANGE FORCES US TO EXPERIMENT AND ADAPT! THAT'S HOW WE LEARN AND GROW!
WHOOPS.. WELL, HERE'S A FRESH CHALLENGE.
I'LL ADMIT IT'S OPENED UP NEW HORIZONS.
THE PROBLEM WITH NEW EXPERIENCES IS THAT THEY'RE SO RARELY THE ONES YOU CHOOSE.
I FEEL SMARTER ALREADY.

Bill Watterson, Andrews McMeel Syndication. All Rights Reserved.

Contents

CHAPTER 1
INTRODUCTION

To most people, fungi are the “black stains in their shower”. However, fungi are an incredibly varied group of organisms that are genetically separated from bacteria, plants and animals, and do far more than make showers disgusting. In fact, fungi thrive in a slew of environments, and form vital parts of every conceivable ecosystem. Fungi recycle energy, degrading energy in complex stores such as wood fibres [1–3] into more easily consumable energy, which are later used by other organisms, or as input to biofuel generators. Most plants form symbiotic relationships with a specific group of fungi, called mycorrhizal fungi, and exchange carbon for water and minerals[4]. The fungal networks in soil even transmit signals between plants when they are attacked by pathogens[5]. Fungi live all over the human body, on the skin, in the mouth and in the gut[6], in both healthy and sick people. Different fungi are present in different body parts, and form complex interactions with other fungi, bacteria and viruses to maintain human health. Even in artificial ecosystems, fungi are exploited to produce specific compounds. In the case of beer, it is usually a yeast fungus[7] that consumes sugars and produces alcohol, carbon dioxide and various other flavour enhancing molecules. In the case of medicines, fungi produce antibiotics that save the lives of millions of people each year[8]. Fungi also feed us as mushrooms, turning what are essentially waste products (manure and biomass) into a nutritious and delicious dish.

Despite all the good that fungi can do for us, they also pose a great threat. Fungi are responsible for disastrous loss of biodiversity, including the decimation of elm trees in Europe and the Americas by the Dutch elm disease[9], and the devastation of banana plantations by the Panama disease[10]. To this day, fungi threaten our food security[11] and global biodiversity, causing the extinction of more plants and animals than any other infectious disease[12]. Not to mention their impact on our own health[13]. Fungi have been found in brain tissues of Alzheimer’s patients[14], and the same fungi that peacefully exist in our mouth and gut can become pathogenic when our immune systems are weakened. Fungi are increasingly prevalent in hospital environments[15], and their impact on health is an ongoing concern.

If we are to exploit fungi to produce better beer, higher crop yields and cleaner energy, or even to combat them in hospitals and fields, we need a better understanding of the regulation of an incredible variety of biological processes in these fungi. This is a tremendous task, and it is made much more complicated by several biological traits that impede analysis. Genetic variation between fungi is very high[16], making comparative analysis difficult, and consequently hard to transfer knowledge from one fungus to another. Additionally, the genes in fungi are often located very close to one another in their genomes, making it difficult to separate them properly. Further, the genome organi-

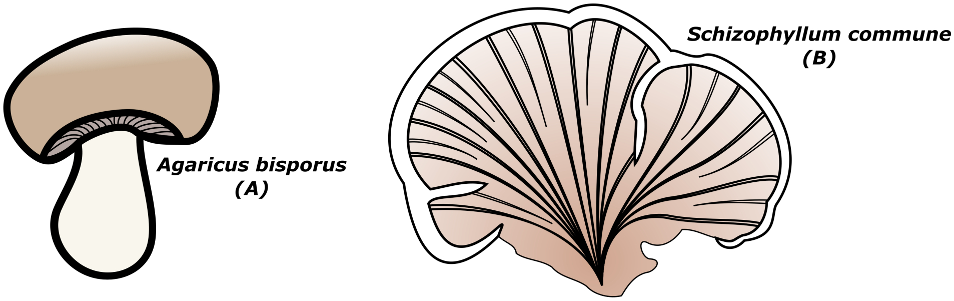


Figure 1.1: *Mushrooms take many different forms. (A) Agaricus bisporus produces its spores underneath an umbrella-like cap, whereas Schizophyllum commune (B), produces its spores in gills which split inside each other as it grows.*

zation is highly sophisticated and varies considerably between different fungal species. Many have multiple nuclei, each containing different genomes. This makes it difficult to decipher the origin of phenotypic variation. Beyond this, even in the same species of fungi, where variation should be minimal, mutations that accumulate through growth in the laboratory confound experimental results.

AN INTRODUCTION TO BIOLOGICAL VARIATION IN FUNGI

In order to understand variation in any organism, we must first have a basic understanding of the way in which information is represented and utilized in a biological system. Fungal organisms are composed of one or more cells. Within each cell, we find at least one nucleus, a structure that separates the genome from the rest of the cell. The genome is a set of extremely long molecules, called chromosomes. On these chromosomes, genes are encoded which represent instructions for the operation of the organism. Given a certain environmental factor, specific genes are copied from the genome into free-floating molecules called RNAs in a process called transcription. RNAs on their own are not usually functional, and must first be converted to a protein through a process called translation. It is the proteins that eventually act as the functional components of the cell, reacting to changes in the environment, performing the necessary tasks for survival, and even changing the appearance of an organism. Each step of this process can be measured, and a change in the genome can trickle into a change in the regulation of transcription, translation, and growth.

The most obvious kind of variation is in the physical appearance of a fungus – the morphological variation. For example, *Agaricus bisporus* (Figure 1.1A), and *Schizophyllum commune* (Figure 1.1B), are obviously different in their outer appearance. *Agaricus bisporus* is probably the most familiar form of mushroom. It forms gills under an umbrella-like cap, in it produces its spores. *Schizophyllum commune* on the other hand produces a fan-shaped mushroom, with gills reaching from the base to the tip, splitting as they dry, leading to its name – the split gill mushroom. Other fungi are individual cells, while others produce big networks of cells, and others produce special tissues.

Fungi can also vary in their genome organisation. Broadly, we can describe a genome organization in two dimensions, their karyotype, indicating how many nuclei are in each cell, and their ploidy, describing how many copies of each chromosome exist in each nucleus. Cells with one copy of the genome are haploid, those with two are diploid. We are most familiar with diploid genomes, such as humans, cats, and dogs, in which each cell contains one nucleus with two copies of each chromosome. If the number of copies are different for each chromosome, then it is called aneuploid. Cells with one nucleus are called monokaryotic, and cells with two nuclei are called dikaryotic. If a cell contains one or more nuclei, each which contains the same genomic information, it is called a homokaryon, whereas if it contains more than one nucleus each contains different genetic material, it is called a heterokaryon. Fungi can have a varying number of nuclei and nuclei may contain also a varying number of chromosomes[17].

Even within a single species, the genome organization can differ, depending upon the sexual stage of the organism. *Schizophyllum commune* starts from a spore, a specialized cell designed to survive harsh conditions, that germinates to form a haploid homokaryon. When a homokaryon encounters another compatible homokaryon, they fuse and form fertile, haploid dikaryons, eventually forming a mushroom. In the mushroom, special cells called basidia fuse nuclei, producing diploid homokaryons, and eventually haploid homokaryon spores. This generality does not extend to all fungi, and the number of nuclei and chromosomes can also vary dramatically.

In each developmental stage, the organism may also behave differently. This means that it reacts differently to its environment at one stage in development than in another, producing different levels of RNA and proteins. This variation in gene regulation is compounded by the slightly different ways in which same genes are encoded in the different nuclear types. A slight genetic variation can produce a different protein and this may have a large impact on function. Often, only one of the nuclear types is known. In this case, the genetic variation in the other nuclear type hides a functional variant that may do something entirely different.

In many complex organisms, a single gene can encode multiple proteins. In a process known as alternative splicing, pieces of RNA are removed (spliced) in alternative ways before translation, producing different proteins. It is very difficult to identify these differences, especially in fungi, where genes lie very close to each other, making it difficult to identify the start and stop of RNA products, and thereby their alternate forms. These alternative products are part of the instruction set of the organism, and the way these different products are used varies depending upon the environment and the organism's state. Again, the differences in genes between the different nuclear types can also interfere with identifying the variation in this stage.

Genomic variation is introduced through the regular process of growth. Biological growth can take essentially two different forms. Some fungi work as individual cells and simply become bigger, while others form large multicellular communities of many clearly separated components. In both cases, it very often becomes necessary for an organism to duplicate its genome, and form two smaller cells. Fungi that work as large communities become larger by dividing the cells. Each time a cell divides, it forms a copy of its genome, and passes one copy to the offspring cell. Each time a genome is duplicated, errors may occur. These errors can accumulate over time, and this gained genomic variation can cause variation in phenotype. In Glomeromycota, it is hypothesized that having multiple nuclei helps to overcome deleterious mutations introduced in this manner[18], but it is not clear how frequent these spontaneous mutations are, and how they impact the functioning of the cell.

MUSHROOM FORMATION IN *Agaricus bisporus* AND *Schizophyllum commune*

We are particularly interested in the process of mushroom formation (also known as fructification) in *Agaricus bisporus* (Figure 1.1A), which is the most cultivated mushroom forming fungus in the world. It is known as the champignon, the white and brown button mushroom, and the Portobello. The increased knowledge on the process of mushroom formation aims to optimally produce mushrooms as a food source, i.e. optimizing the growth of *A. bisporus* to produce more mushrooms, faster, cheaper, and more efficiently than currently is possible. In nature, *Agaricus bisporus* grows on humus, the carbon rich layer of soil that is the result of degraded biological material such as leaves and animal droppings. There, it colonizes the soil (Figure 1.2B) until it has built up enough energy stores to begin forming mushrooms, when hyphae in the soil clump together to form tiny structures known as initials. These initials, in a process called pinning (Figure 1.2B), develop into primordia which have a pileal section that will develop into the cap of the mushroom, and a stipeal section that develops into the stipe of the mushroom. Later, these sections develop into different tissues (Figure 1.2C), including the gills, where spores are produced. *A. bisporus* produces fertile dikaryotic spores, which undergo very little crossover[19]. After a while, the mushroom will age and release the spores.

Although *Agaricus bisporus* is industrially cultivated, it is very difficult to study due to its inability to grow on a simple medium in the lab. In the lab it is grown under industrial cultivation conditions, i.e. it is grown on compost produced from horse manure, straw and gypsum, topped with a so-called casing layer of peat. The mycelium is permitted to colonize the compost, and grows through the peat to form mushrooms as in the wild. However differently from in the wild, the industrial setting convinces the colony to produce multiple harvests (Figure 1.2C-E). By harvesting mushrooms, and controlling oxy-

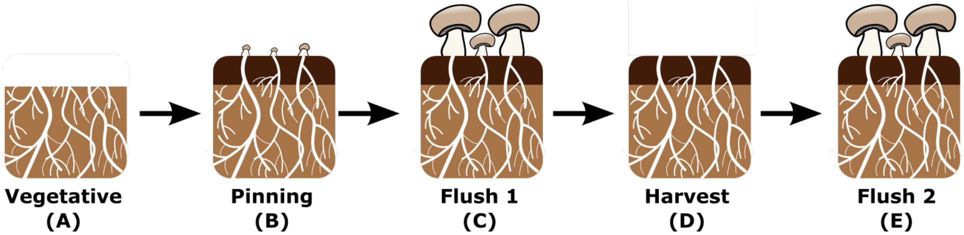


Figure 1.2: *Industrial cultivation of Agaricus bisporus. (A) The mycelium of A. bisporus colonizes compost, gathering energy to produce mushrooms. Once the compost has been fully colonized, a nutritionally poor casing layer is placed over the compost (B, top segment). Mushroom formation is initiated, and the pins of the mushroom begin to form (B), which eventually form the entire mushroom (C). Industrially, these mushrooms are harvested (D), and induced to grow again (E).*

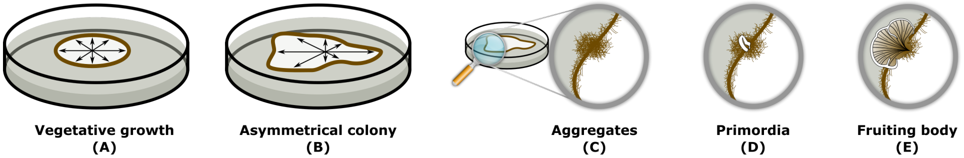


Figure 1.3: *The growth stages of S. commune. The fungus grows vegetatively (A) but at some point grows asymmetrically (B). Along the edge of the colonies, aggregates form (C), which develop into primordia (D), the initial stages of mushrooms (E).*

gen, temperature and light levels, the colony is cajoled into producing multiple flushes of mushrooms. Industrially, this flushing will be continued until after the third flush, where the compost is discarded and the process started again. Besides the relative difficulty in growing *A. bisporus* in the lab, it also lacks efficient genetic modification tools, which limits the ability to study mushroom formation in *A. bisporus* comprehensively.

Consequently, fructification is generally studied in the mushroom model organism *Schizophyllum commune* (Figure 1.1B) as it has the advantage of being culturable in a petridish, and it is possible to genetically modify it [20]. *Schizophyllum commune* is a mushroom forming fungus that exists on all continents with the exception of Antarctica [21, 22]. Although it has culinary value, it is not particularly enjoyed in Europe.

The life cycle of *S. commune* starts as a spore which was produced from a previous mushroom. Under favorable environmental conditions, the spore germinates into a homokaryotic mycelium. This mycelium grows vegetatively and, under normal circumstances, it cannot produce a mushroom on its own. It will continue to grow vegetatively until it comes across another *S. commune* mycelium which has a compatible mating type. At this point two cells from the two strains of *S. commune* fuse, in a process known as hyphal anastomosis, to form a fertile dikaryon by nuclear exchange. Mushroom formation is initiated when the environmental conditions are met. As *S. commune* is a wood rot fungus, it grows in the wood fibres where there is no light, low oxygen and high carbon dioxide levels. The fungus grows through the wood until it experiences a drop in carbon dioxide levels, and blue light, signalling that it has reached the outer surface of the wood. At this point the hyphae of the fungus grow out, forming aerial hyphae that protrude from the medium it has grown in. These hyphae aggregate into bunches called aggregates (Figure 1.3C), and eventually into primordia (Figure 1.3D) which turn into the fruiting bodies (Figure 1.3E). Inside the fruiting bodies, new spores are created. In *S. commune*, spores are the result of the fusion of the compatible nuclei in the basidia and underdoing crossover and chromosomal trading. The spores are released to begin the cycle anew.

Although *S. commune* has been a subject of research for a long time, relatively little is known about its genetics. The mating type compatibility loci have been established for a few strains, but the precise differences that are required between the two strains are not exactly known. A few of the regulatory genes involved in mushroom formation have been identified[21, 22]. These regulators, when removed (knocked out) have been shown to halt the formation of mushrooms in different developmental stages, or to produce different phenotypes[22]. These regulatory genes were identified by growing *S. commune*

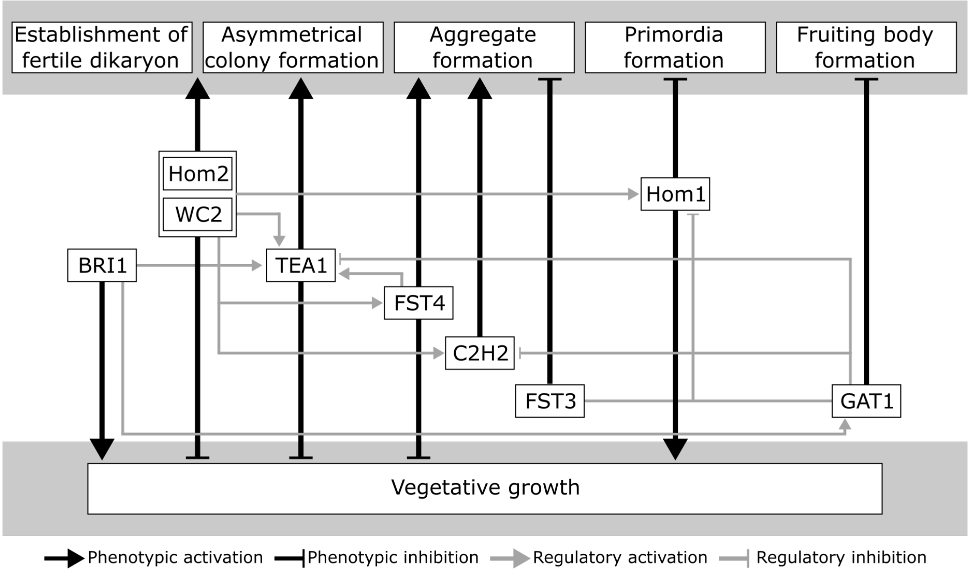


Figure 1.4: *Impact of transcription factors on phenotype and gene regulation in S. commune. The boxes in the gray area represent the different developmental stages, and the boxes in the middle represent transcription factors. The black arrows indicate a phenotypic effect on mushroom development, while the gray arrows indicate a regulatory influence of the different transcription factors.*

in the lab (Figure 1.4). There, it grows very differently from in the wild. Rather than growing on wood, it grows on a sugar-agarose medium in a petri dish, which is a much simpler medium than wood. In the lab, the environmental conditions must be regulated externally, as they are not provided by nature. Thus, the plates are kept in the dark, and in high carbon dioxide levels.

Despite the work in *S. commune* to understand the genetic regulation behind mushroom formation, there is very little known about this process in *A. bisporus*. Many of the genes found to be involved in fructification in *S. commune* exist in *A. bisporus* [23], however, the lack of genetic modification tools and the relatively difficult culturing conditions, have limited the ability to confirm their functional similarity. This is further complicated due to the high diversity of fungal genomes making a comparative analysis between *S. commune* and *A. bisporus* not straightforward.

ORTHOLOGOUS GENES BETWEEN SPECIES

When we claim that many of the fructification genes in *S. commune* also exist in *A. bisporus*, what we are saying is that we have found genes in *A. bisporus* that look a lot like genes found in *S. commune*. The relationship we believe to have established is one of orthology - the gene in *A. bisporus* and the gene in *S. commune* are derived from the same gene in a common ancestral genome. We hypothesize that orthologous genes have the same function in both organisms. However, the detection of orthology relationships between species can be marred with difficulties.

Orthologous genes do not need to have the same gene sequence. Often, not all parts of a gene are necessary to fulfill the function of the gene. Thus, some pieces may be lost in some organisms, and some pieces may be added in others. The function might still be the same, but the gene sequence is now different. This makes it harder to identify the ortholog among all the genes in the species. Even further, the gene nucleotide sequence can be written in an entirely different code in both species. The nucleotide sequence may appear completely different, but the protein sequence they produce will be very similar. Altogether, the sequences between two orthologous genes may differ substantially and retain the same function.

An additional complication lies in that there may be several copies of a single gene that can obscure the true orthology relationship. Through evolution, a gene may be duplicated. Many genes exist in multiple copies. This can be beneficial for an organism as it can survive the loss or damage of one of the copies. If one copy is relieved from evolutionary pressures to maintain the important parts of its sequence, it may change its function.

Thus, if we find a gene in another organism that matches our gene of interest, there may be two scenarios. It might be an ortholog - the same gene in the two species derived from an original gene in a common ancestral genome, or it may be a paralog - a copy of the same gene in a common ancestor that was duplicated. The consequences of this distinction is not fully clear either. Like orthologs, paralogs may also share the same function. Often, we cannot distinguish between them. In both of these scenarios, we can refer to them as homolgous.

To elevate a homologous relationship to an orthologous relationship, we can use additional information. Do we find similar relationships in other species? Are both homologs expressed in a similar way in both species, across different conditions? In the two species, are the genes surrounding the homologs also homologous? Accumulating evidence allows us to be more condifent in an orthologous relationship.

Knowing that two genes are orthologous is the second step in the transfer of knowledge from one organism to the other. The first step is to gather knowledge in one organism.

TRANSCRIPTOMICS REVEAL GENES INVOLVED IN MUSHROOM FRUCTIFICATION

Although the genome can tell us a lot about the functional capabilities of an organism, in the end it is the proteins that perform biological tasks. Hence, measuring only the genome tells us nothing about the functional activity of an organism. Yet, measuring protein abundance is too difficult to do at a large scale. With the recent upswing of Next Generation Sequencing[24] (Figure 1.5), it has become possible to efficiently sequence RNA from cells. This gives us a view on the functional activity of an organism as RNA is a proxy for protein abundances, under the assumption that the levels of RNA and protein are correlated[25, 26]. RNA sequencing starts by first isolating mature RNAs, mRNAs, from a sample. The long pieces of RNA are broken into small pieces that are sequenced individually, producing short fragments of sequences known as reads. These small pieces can be found on the reference genome in a process called alignment.

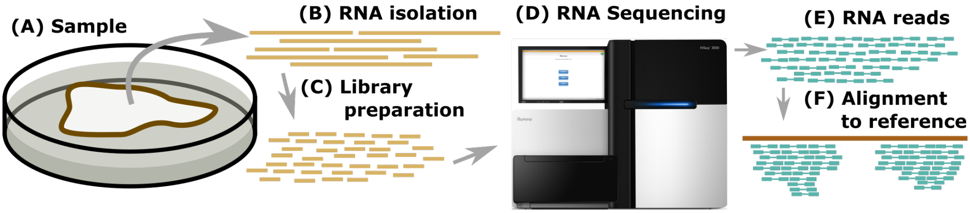


Figure 1.5: *RNA-Sequencing. After the isolation of RNA from a biological sample (A-B), the pieces of RNA are fragmented into small pieces that can be sequenced. These small pieces are sequenced using a short-read sequencer, such as the illumine platform (D), which produces paired-end reads (E). These paired-end reads are aligned to the reference genome.*

ALIGNMENT

Given a reference genome and a read, sequence alignment refers to the task of identifying the locations in which the same sequence can be found. Because of the size of a genome, this can be a time consuming problem, exacerbated by repeat regions, sequencing errors and biological variation. In reality, an RNA molecule originates from one single region on the genome. In practice however, a region may be repeated on the genome, and those regions look so similar that the alignment algorithm does not know which one of them it actually originates from, resulting in an ambiguous alignment. Additionally, the reads we obtain can contain sequencing errors, meaning that a base is replaced by another. Although the error rate is quite low, it is sufficiently large that we need to allow our alignment to permit sub-optimal alignments. On top of this, it is often the case that our reference genome was created from a different individual from the one we are currently sequencing. Then, by the biological variation between those two individuals, there can also be differences in the reads we obtain. Thereby, within a single read, the biological variation is indistinguishable from a sequencing error. To alleviate these problems, many sequencing projects are performed in a paired end format. This means that the RNA molecule is sequenced at its two ends, obtaining two reads where the distance between them on the genome is approximately known. Therefore it is known that the two ends of the read must align nearby each other on the genome. This helps to resolve ambiguous alignments, provides evidence of expression for the bases in between the two ends, and gives some more structural information on the RNA molecule.

DIFFERENTIAL EXPRESSION

When a gene is highly expressed in one sample, but lowly expressed in another, we say that this gene is differentially expressed between the two samples. Typically, one performs RNA-Seq studies with the intention of identifying genes that are differentially expressed with respect to some condition such as growth on different substrates, or different developmental stages. By counting the number of reads that are aligned within a gene region, we attain a measurement of the expression of that gene. This number is normalized with respect to the number of reads that aligned to the genome in that sample. The resulting measure represents the expression of that gene relative to the other genes in the organism in that sample. Each experiment is performed in replicates. This means that for each condition, you perform the measurement at least twice (although ideally, more often than that), and align the reads obtained separately. These replicates help to build up a distribution of the expression of a gene under a given condition. A statistical test is then performed to ascertain whether the distributions of expression of a gene under the two conditions are significantly different, or not.

However, in any given experiment, hundreds of genes will be differentially expressed, and it is difficult to identify which are the most important, or interesting. Often, they will be prioritized based on the observed difference in expression, but may also be prioritized on functional attributes, such as specific protein domains that indicate regulatory roles like transcription factors. Once these have been identified, the next goal is to understand the role of these genes. For this, knockout studies help us understand their impact on phenotype and on gene regulation.

KNOCKOUT STUDIES

In a gene knockout study, a gene is deactivated through genetic modification, in an attempt to ascertain the function of that gene. Phenotypic changes, such as the loss of an ability to produce a specific compound, or a disruption in development will give an indication of the role of the gene in phenotype. Additionally, if the RNA is measured in a strain with a knocked out gene, we can observe the regulatory effects of the loss of this gene. However, many gene knockouts will not produce a visible phenotypic effect[27], as they may be redundant, or involved in an inactive component of the cell.

Previously[21, 22], several transcription factors, genes known to function as regulators, have been identified in *S. commune*. In Figure 1.4, we show the current genetic model of mushroom formation in *S. commune*. The phenotypic role of the different genes was determined by observing the effect of the gene knockout on the development of the mushroom. For example c2h2, when knocked out, produced fewer aggregate structures, and therefore, fewer mushrooms. This indicates that c2h2 provides a supportive role in the regulation of aggregate formation. These kinds of observations permit a reconstruction of the genes involved in different stages of mushroom development.

Furthermore, we can identify interactions between genes. In a knockout study, if we observe a differential expression between a gene between a wildtype (not genetically modified) organism, and a knockout organism, we can use this as evidence of a regulatory interaction between genes themselves. Typically we represent the interactions between genes as either activating, or inhibiting expression of another gene. For example, when we knocked out the Fst3 gene, we observed an up-regulation of Hom1. Thus, Fst3 (or some downstream gene currently unknown to us) is responsible for inhibiting the expression of Hom1.

While a knockout can give us information about what happens when the gene is gone, the opposite effect - overexpression - can tell us what happens if this gene is more activated. If we observe a different phenotype as a result of the hyper-expression of a specific gene, it can hint to us that the gene is involved in the regulation of that phenotype. Similarly, in the case of gene expression, it can provide us the opposite information of a knockout. Based on our observations of the Fst3 knockout, we would expect to see lower levels of Hom1 in an overexpressor of Fst3.

ALTERNATIVE SPLICING INFLUENCES REGULATION

After DNA is transcribed into RNA, it still contains many segments, called introns, that are not part of the final protein product. These segments are removed from the RNA sequence in a process called splicing. However, sometimes, splicing can have various outcomes, each producing a slightly different RNA sequence that can change the protein sequence, and thereby the function of the protein. This is known as alternative splicing, producing multiple different RNA transcripts from one gene. Alternative splicing is more often studied in humans and many other mammals, and has been shown to have important roles in development and health. In fungi, however, it is largely ignored as a simplifying assumption. Additionally, it has been difficult to identify alternatively spliced genes due to complicating factors.

The genomes of fungi are exceptionally dense. While in humans there are approximately 4,000 nucleotides between genes[28], in *S. commune*, this distance is only 539 nucleotides. This is a problem because transcripts contain slightly more sequence than only the gene itself. Pieces at the start and end of the molecules, known as UnTranslated Regions (UTRs) are present on the transcript, but do not form part of the final protein. These regions can be very long, exceeding the distance between genes[29], and frequently flow into neighbouring genes. This means that, when we measure the RNA of a fungal sample and break it into millions of small pieces, we lose track of which gene that piece of RNA actually originated from. We know that it must originate from one of the two, but our algorithms cannot determine which is correct, and fails.

In this thesis, we developed a simplified method to avoid the problem of overlapping UTRs. By doing this we are able to identify different alternative splicing events and reconstruct the structures of alternatively spliced gene products. With this, we could examine the expression of the different transcripts throughout mushroom development, but also the differences in functionality of the different transcripts. We find widespread alternative splicing, more than has ever been observed in any fungus, and associate the differential regulation of different transcripts to development of the mushroom. Furthermore, we find that many of the different transcripts have alternative functional properties. Taken together, this work revealed a large amount of hidden functionality that was previously obscured by the overlapping structure of the densely packed genes, and the interference of UTRs.

NUCLEAR SPECIFIC EXPRESSION OF GENES.

S. commune and *A. bisporus* exist as heterokaryons, meaning each cell has at least two nuclei, each with their own DNA. The differences between the two nuclei have shown to impact growth characteristics in homokaryons constructed from the wildtype heterokaryon[1]. The phenotypic variation is caused by differences in protein abundances in each of the nuclei, which should also reflect in differences in RNA expression levels. Currently, all transcriptomic studies in all fungi align the sequenced reads to one of the genomes inside the cell, or even a combined genome of the different constituent nuclear types. This is usually a fundamental limitation caused by the lack of genomic information for the individual nuclear types. Despite this difficulty, it is still possible to examine the nuclear type specific expression. By identifying sequence variation in the expression data, one can delineate reads that originate from different nuclear types. In this scenario, however, it is not possible to know which nuclear type it came from. This means that it is impossible to make any claims about the functional or regulatory contribution of one genome or the other.

In this thesis, we uniquely take advantage of the fact that the genomes *A. bisporus* homokaryons have both been independently sequenced and assembled. With this knowledge, we are able to discern the nuclear origin of many RNA-Seq reads. We developed a method to identify distinguishing markers (based on genomic variation) between the genes in the two genomes, and determine transcriptomic contribution of each nucleus. We find that one nuclear type produces more RNA than the other, although there appears to be some variation throughout development. Further, we find that one nuclear type produces a higher expression of many genes involved in nutrient degradation than the other, hinting at a possible superiority in vegetative growth. These results show that the two nuclear types, although required for sexual reproduction, still maintain some autonomy and respond differently to the same environmental stimuli. Again, we uncover a new world of variation and functionality, that has previously been hidden

EVOLUTION IN THE LABORATORY

It is often thought that evolution occurs at the level of a species, but actually, it occurs at the level of individual reproducing organisms, but also at the level of individual cells. A cell which gains an advantage over its neighbours in some respect can quickly become dominant. A common and devastating example of this outside the fungal domain is cancer. The primary mechanism of such evolution is when DNA is copied, and small errors occur, as mentioned above. It is such mutations that can give rise to bacterial antibiotic immunity, the development of tumors, and many other human diseases. In a laboratory setting, it is generally undesirable to have such mutations, as the validity of scientific conclusions rests on the assumption that the organism studied today, is the same as the one studied the week before. For this, research institutes make use of strain preservation systems to reduce the number of cell duplications that occur in the time between measurements. Nevertheless, these mutations occur, but are generally ignored.

Not all mutations will give a phenotypic difference large enough to be observed. Further, they can interfere with experiments. Suppose a gene knockout is produced, but in the process of culturing the cells to measure the effect of this knockout, a new mutation with damaging effects is gained in another gene. From this measurement alone, it is not possible to know whether it is the the knockout that is responsible for the observed measurement differences, or the mutation. Such interference is inevitable, but it is not often known to what extent it occurs in the lab. In this thesis, we developed a method to identify these mutations from RNA-Seq measurement data, and to identify where in the experimental design those mutations originated. We found that the mutation rate in the laboratory is the same as it is in the wild, despite the attempts of the strain preservation system to curb the influence of the mutation rate. We find thousands of mutations across our samples, even in important regulatory genes. Additionally, we find that the presence of SNPs can have an impact on the expression of nearby genes. These findings show that the mutations accumulating in the standard laboratory experiments should not be discounted, and that such variation could potentially have large effects on scientific conclusions.

OUTLINE OF THIS THESIS

In this thesis, we investigate the consequences of the various simplifying assumptions that are made in the study of fungi. We do this by assessing the variation that exists within and between *S. commune* and *A. bisporus*.

Therefore, as a starting point, we propose an improved algorithm to study synteny (Figure 1.6) in this thesis. We realized that most synteny tools work on the nucleotide level, and, consequently, suffered dramatically from the high diversity between the two genomes. We build a synteny on exon level that is robust to higher variation and introduce a clever strategy to merge homologous exons into syntenic blocks. We use this tool in Chapter 2, to investigate the genetic similarity between *S. commune* and *A. bisporus*. As mentioned earlier, the identification of gene orthology is a difficult problem to solve. Often, it helps to examine the genomes from a much higher level, and to examine similarities across large groups of genes, rather than individual genes. Similarity in the order that genes occur is called synteny, and we developed a tool to investigate synteny between pairs of highly divergent fungal species. We find that although at the small scale there is very little similarity, at the large scale there are a lot of similarities between *S. commune* and *A. bisporus*.

Knowledge transfer between the two organisms has been very difficult due to the large differences that the genomes exhibit, but in Chapter 3, we performed the first transfer of knowledge between *S. commune* and *A. bisporus*. When a gene which we have linked to mushroom formation in *S. commune* is overexpressed in *A. bisporus*, we observe the opposite effect of what we saw in *S. commune* when it was knocked out. This indicates that this particular gene has the same function in *A. bisporus*, as in *S. commune*, and that although there is a substantial amount of variation between the two genomes, a significant component of mushroom fructification remains intact.

Chapter 4 discusses the variation induced through the process of alternative splicing in *S. commune*. Using the RNA measurements that we obtained from several stages of mushroom development, we were able to identify that *S. commune* regulates the different products of splicing differently throughout development. Further, we show that the different products can even have different functions. This shows that the simplification of ignoring alternative splicing makes us oblivious to the existence of thousands of additional protein products which could be involved in interesting biological processes in *S. commune*, and indeed, many other fungi.

The variation that is included in the differences between the constituent homokaryons of *A. bisporus* have been hidden from view due to the collapsed reference genome structure. However, the recent publication of the genomes of these individual homokaryons[19] gave us the unique ability to investigate the transcriptomic variation as a result of two different nuclei in the same cell in chapter 5. Using the RNA measurements we took from *A. bisporus*, we find that many genes are differently regulated between the two nuclear types, with one nuclear type being more active than another. Further, different functional groups are more active than others in one nuclear type than the other. The simplification that the two genomes are collapsed into one

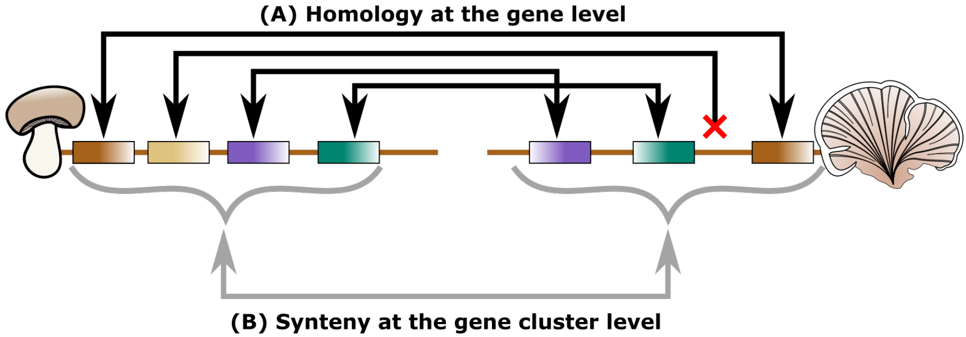

(A) Homology at the gene level
(B) Synteny at the gene cluster level

Figure 1.6: *Synteny is constructed from homology relationships. Sequence similarities between gene sequences tell us about homology relationships between genes in different organisms. (A) We depict a group of genes on the A. bisporus and S. commune genomes. The colored boxes indicate gene regions along a genome, and the black arrows depict homology relationships based on sequence similarities. Here, the light brown gene is present in the A. bisporus genome, but not in the S. commune genome. (B) The genes are homologous to each other, but beyond that, the grouping of genes is also conserved between these two genomes. It is such a conserved grouping of genes that we refer to as synteny.*

genome hides a complex interaction between the two nuclei throughout growth, and the different characteristics of the two.

Spontaneous mutations that accumulate in an experimental design are possible confounders in every experiment. In chapter 6 we show that, using the RNA measurements from *S. commune*, we are able to identify SNPs that originated in different stages of the experimental design. These SNPs elucidate many unanswered questions about the recombination rate and the efficiency of the strain preservation system. It also revealed a strong warning about the likelihood of a confounding mutation. Of the 40 measurements we took, we found potentially function altering mutations in 40 genes, including regulators and basic cell function genes. These results inform us that the experimental design needs to be done carefully, and that the strain preservation system does not appear to slow down the mutation rate. Scientists need to be aware of the likelihood of these mutations in their research, and that somatic mutations cannot be ignored. In Chapter 7, we discuss the conclusions of this work and provide some general remarks about the field and what work still needs to be done.

Following this, there is a summary of the thesis in English and Dutch.

BIBLIOGRAPHY

[1] E. Morin, et al. Genome sequence of the button mushroom agaricus bisporus reveals mechanisms governing adaptation to a humic-rich ecological niche. *Proceedings of the National Academy of Sciences* 109(43):17501–17506 (2012).

[2] A. Patyshakuliyeva, et al. Uncovering the abilities of agaricus bisporus to degrade plant biomass throughout its life cycle. *Environmental Microbiology* 17(8):3098–3109 (2015).

[3] R. a. Ohm, et al. Genome sequence of the model mushroom schizophyllum commune. *Nature biotechnology* 28(9):957–63 (2010).

[4] K. Lin, et al. Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. *PLoS Genetics* 10(1) (2014).

[5] Z. Babikova, et al. Underground signals carried through common mycelial networks warn neighbouring plants of aphid attack. *Ecology Letters* n/a–n/a (2013).

[6] L. Cui, A. Morris, E. Ghedin. The human mycobiome in health and disease. *Genome Medicine* 5(63):1–12 (2013).

[7] M. van den Broek, et al. Chromosomal copy number variation in saccharomyces pastorianus is evidence for extensive genome dynamics in industrial lager brewing strains. *Applied and Environmental Microbiology* 81(18):6253–6267 (2015).

[8] R. Bigelis, H. He, H. Y. Yang, L.-P. Chang, M. Greenstein. Production of fungal antibiotics using polymeric solid supports in solid-state and liquid fermentation. *Journal of Industrial Microbiology & Biotechnology* 33(10):815–826 (2006).

[9] S. Khoshraftar, et al. Sequencing and annotation of the ophiostoma ulmi genome. *BMC genomics* 14(1):162 (2013).

[10] L. Guo, et al. Genome and transcriptome analysis of the fungal pathogen fusarium oxysporum f. sp. cubense causing banana vascular wilt disease. *PLoS ONE* 9(4) (2014).

[11] R. Dean, et al. The top 10 fungal pathogens in molecular plant pathology. *Molecular Plant Pathology* 13(4):414–430 (2012).

[12] M. C. Fisher, et al. Emerging fungal threats to animal, plant and ecosystem health. *Nature* 484(7393):186–94 (2012).

[13] G. D. Brown, et al. Hidden killers: Human fungal infections. *Science Translational Medicine* 4(165rv13):1–9 (2012).

[14] D. Pisa, R. Alonso, A. Rábano, I. Rodal, L. Carrasco. Different brain regions are infected with fungi in alzheimer's disease. *Scientific reports* 5:15015 (2015).

[15] A. Warris, A. Voss, T. G. Abrahamsen, P. E. Verweij. Contamination of hospital water with aspergillus fumigatus and other molds. *Clinical Infectious Diseases* 34(8):1159–1160 (2002).

[16] J. E. Galagan, M. R. Henn, L. J. Ma, C. A. Cuomo, B. Birren. Genomics of the fungal kingdom: Insights into eukaryotic biology. *Genome Research* 15(12):1620–1631 (2005).

[17] L. Bertier, L. Leus, L. D'hondt, A. W. A. M. De Cock, M. Höfte. Host adaptation and speciation through hybridization and polyploidy in phytophthora. *PLoS ONE* 8(12):1–15 (2013).

[18] J.-L. Jany, T. E. Pawlowska. Multinucleate spores contribute to evolutionary longevity of asexual glomeromycota. *The American naturalist* 175(4):424–435 (2010).

[19] A. S. M. Sonnenberg, et al. A detailed analysis of the recombination landscape of the button mushroom agaricus bisporus var. bisporus. *Fungal Genetics and Biology* 93:35–45 (2016).

[20] R. a. Ohm, et al. An efficient gene deletion procedure for the mushroom-forming basidiomycete schizophyllum commune. *World journal of microbiology & biotechnology* 26(10):1919–1923 (2010).

[21] R. A. Ohm. *Regulation of mushroom formation in Schizophyllum commune.* Ph.D. thesis, Utrecht University (2010).

[22] J. Pelkmans. *Environmetal signalling and regulation of mushroom formation.* Ph.D. thesis, Utrecht University (2016).

[23] T. Gehrmann, M. J. Reinders. Proteny: discovering and visualizing statistically significant syntenic clusters at the proteome level. *Bioinformatics* 31(21):3437–3444 (2015).

[24] S. Goodwin, J. D. Mcpherson, W. R. Mccombie. Coming of age : ten years of next- generation sequencing technologies. *Nature Publishing Group* 17(6):333–351 (2016).

[25] D. Greenbaum, C. Colangelo, K. Williams, M. Gernstein. Comparing protein abundance and mrna expression levels on a genomic scale. *Genome Biol* 4:117 (2003).

[26] T. Maier, M. Güell, L. Serrano. Correlation of mrna and protein in complex biological samples. *FEBS Letters* 583(24):3966–3973 (2009).

[27] I. Barbaric, G. Miller, T. N. Dear. Appearances can be deceiving: Phenotypes of knockout mice. *Briefings in Functional Genomics and Proteomics* 6(2):91–103 (2007).

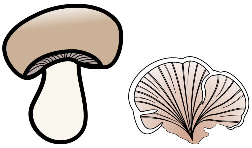
[28] S. Djebali, et al. Landscape of transcription in human cells. *Nature* 489(7414):101–8 (2012).

[29] T. Gehrmann, et al. Schizophyllum commune has an extensive and functional alternative splicing repertoire. *Scientific Reports* 6(1):33640 (2016).

CHAPTER 2

PROTENY: DISCOVERING AND VISUALIZING STATISTICALLY SIGNIFICANT SYNTENIC CLUSTERS AT THE PROTEOME LEVEL

Thies Gehrmann
Marcel J. T. Reinders



This chapter is published as: *Bioinformatics*, Volume 31, Issue 21, 1 November 2015, Pages 3437–3444 ()doi: 10.1093/bioinformatics/btv389)
Supplementary material is available online at: `http://bit.ly/2FTPYdN` (oup.com)

ABSTRACT

With more and more genomes being sequenced, detecting synteny between genomes becomes more and more important. However, for microorganisms the genomic divergence quickly becomes large, resulting in different codon usage and shuffling of gene order and gene elements such as exons.

We present Proteny, a methodology to detect synteny between diverged genomes. It operates on the amino acid sequence level to be insensitive to codon usage adaptations, and clusters groups of exons disregarding order to handle diversity in genomic ordering between genomes. Furthermore, Proteny assigns significance levels to the syntenic clusters such that they can be selected on statistical grounds. Finally, Proteny provides novel ways to visualize results at different scales, facilitating the exploration and interpretation of syntenic regions.

We test the performance of Proteny on a standard ground truth dataset, and we illustrate the use of Proteny on two closely related genomes (two different strains of Aspergillus niger) and on two distant genomes (two species of Basidiomycota). In comparison to other tools, we find that Proteny finds clusters with more true homologies in fewer clusters that contain more genes, i.e. Proteny is able to identify a more consistent synteny. Further, we show how genome rearrangements, assembly errors, gene duplications and the conservation of specific genes can be easily studied with Proteny.

Proteny is freely available at the Delft Bioinformatics Lab website http://bioinformatics.tudelft.nl/dbl/software.

2.1. INTRODUCTION

A synteny analysis is a useful way to compare organisms, that allows us to study the evolution between genomes, make claims about functional conservation [1, 2], identify genome rearrangements [3], aide genome annotation [4], and even predict genome assembly errors.

Numerous tools are already available to detect synteny. Tools like Mugsy [5], Mauve [6], Multiz [7] and Sibelia [8], focus only on highly related genomes. OrthoCluster [9] and SyMAP [10] operate at the DNA level and discover groups of genes with their gene order being conserved. These assumptions are too strict when considering more distant genomes (see Suppementary Material S1).

i-ADHoRe [11–13] works at the protein level, and builds a homologous gene matrix based on protein-protein alignments, detecting clusters of genes by identifying diagonal groups of genes, allowing for a maximum gap size. However, for more distant genomes, exons may be inserted and removed from genes, while splice variants remain conserved [14]. Consequently, it seems more reasonable to detect synteny between more distant genomes by considering the protein level at the resolution of exons, rather than the genes.

We introduce a method, called Proteny, which can discover statistically significant syntenic clusters between diverged genomes that may have different codon usages. Proteny analyses synteny at the exon level, so that more distant homologies can be revealed. As Proteny assigns a significance level to the detected syntenic clusters, it only requires setting a p-value cutoff and an intuitive parameter balancing the conservation ratio of the detected clusters.

Traditionally, synteny is visualized using dot-matrix plots such as those in R2Cat [15] and SyMAP [10] which is useful to visualize the synteny between entire genomes, but not when closely inspecting specific regions. Novel techniques such as [16] can visualize synteny between many genomes at a lower level, but quickly produces complicated figures when looking at very large regions or sufficiently different organisms. Easyfig [17] can look at different levels and can be used to annotate interesting regions, however, it must be done manually. Cinteny [3] provides multi-level visualizations to display synteny between multiple organisms, but it can not visualize exons. With Proteny we also introduce a user-friendly visualization for examining the discovered syntenic regions, which are especially useful when genomes are more distant.

Proteny is quantitatively benchmarked against a dataset from the Yeast Gene Order Browser that includes a gold standard of orthology relationships[18], and it is compared to i-ADHoRe. We demonstrate the utility of Proteny on two fungal datasets: *1*) two *A. niger* genomes which are known to be highly related, illustrating how Proteny can be used to explore the similarities and differences between two genomes, and *2*) two mushroom forming fungi (of the phylum basidiomycota) *Schizophyllum commune* and *Agaricus bisporus*, demonstrating the power of Proteny to detect syntenic regions between more distant genomes which also differ in their codon usage (see Suppementary Material S1). As there is no gold standard for these datasets, we qualitatively analyse the discovered clusters.

2.2. METHODS

General overview: Proteny detects syntenic clusters by translating all exon regions into protein sequences, and producing a set of BLASTp hits (figure 2.1a). Proteny then calculates a distance between all hits based on genomic distance, resulting in a distance matrix. From this distance matrix, Proteny builds a dendrogram where each node represents a cluster of hits (figure 2.1b). The dendrogram is traversed in a depth first procedure, searching for clusters with significant scores based on a statistical test. Each cluster is scored depending on the hits which are found within the cluster, and the number of unaccounted exons (exons without hits) that lie within the genomic regions that the cluster covers. When a significant cluster is found (and its child is not *more* significant), the branch is cut (i.e. no smaller clusters are evaluated in that branch). Proteny terminates when no more significant clusters can be found, culminating in a set of significant clusters of hits (figure 2.1c). These clusters can then be visualized by looking at the individual hits (figure 2.1d), or at a higher level (figure 2.1e).

Obtaining a mapping: A mapping from organism β to organism γ is a set of pairs, whereby a locus in organism β is linked to a locus in organism γ. Proteny links loci on their translated sequence similarity. For that, all exons in each organism are translated to construct two BLAST databases [19] and two sequence sets for each genome. A bi-directional BLASTp (using default parameters) then produces a mapping, i.e. a set of bi-directional hits $h_i \in H$, between sequences from the two organisms describing a similarity between two sequences. Consequently, a hit represents two regions, $h_i = (r_i^\beta, r_i^\gamma)$,

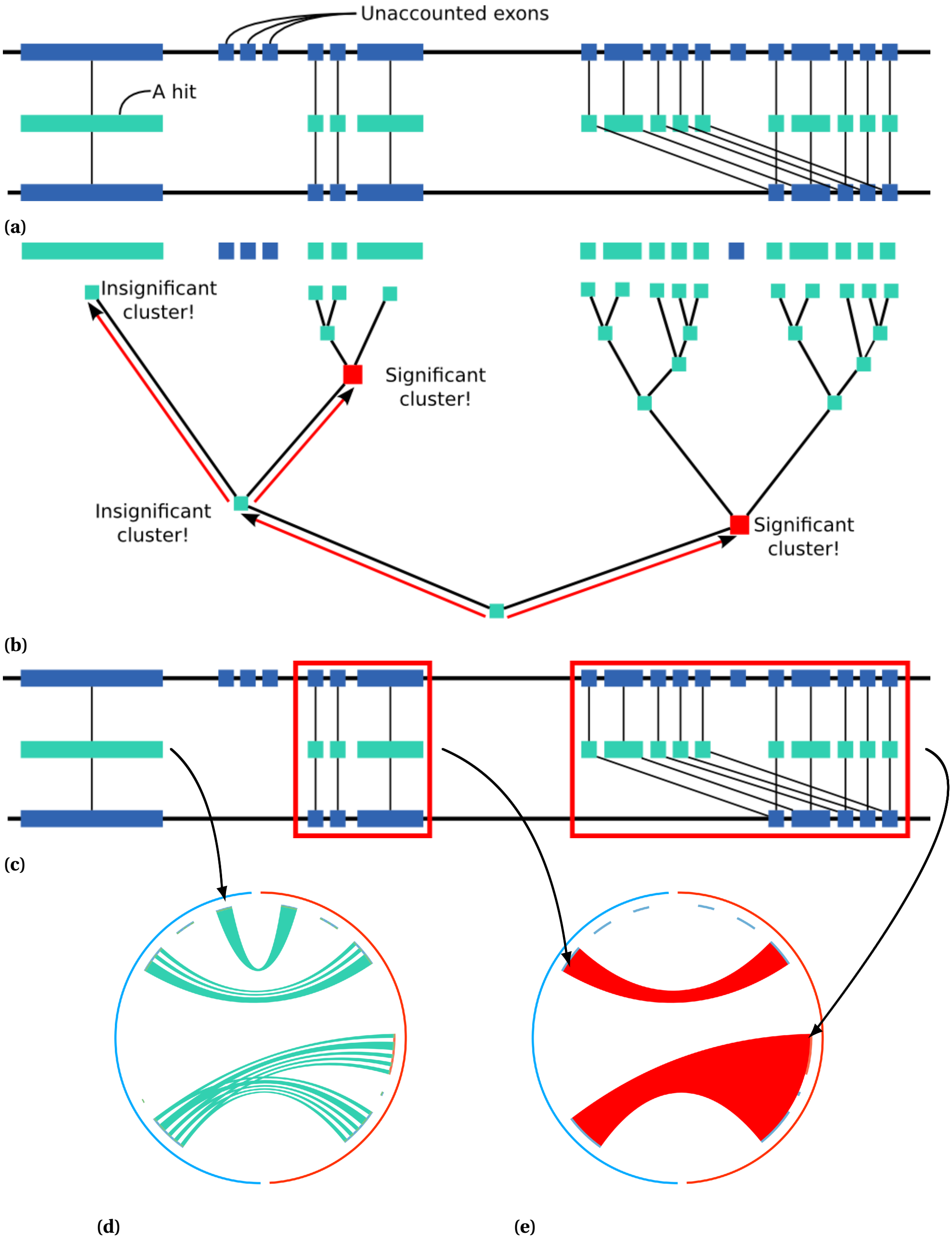


Figure 2.1: *An illustration of how Proteny works.* ***(a)*** *First, BLASTp is used to produce a set of hits, which are used to build* ***(b)*** *a dendrogram which is traversed to find* ***(c)*** *significant clusters (red boxes).* ***(d)*** *Individual hits are displayed (here in turquoise) in a region visualization, while* ***(e)*** *significant clusters are displayed (here in red) in a chromosome visualization.*

which correspond to the genomic location of subsequences of exons in the genomes of organisms β and γ respectively. A region has a start and an end, i.e. $r_i^\chi = (s_i^\chi, e_i^\chi)$, where χ corresponds to an organism. All these variables are clarified in Supplementary Figure S3a.

Distances between hits: In order to cluster hits we need a definition of similarity between them, which we base on the distance between their associated regions. The distance between two regions on the same genome is given by equation (2.1).

$$dist(r_i^\chi, r_j^\chi) = \max\left\{0, \max(s_i^\chi, s_j^\chi) - \min(e_i^\chi, e_j^\chi)\right\} \tag{2.1}$$

The distance between two hits is then defined as the sum of the distance between the two regions on one genome, and the distance between the two regions on the other genome.

$$d(h_i, h_j) = dist(r_i^\beta, r_j^\beta) + dist(r_i^\gamma, r_j^\gamma) \tag{2.2}$$

Note that when two regions overlap (i.e. their distance is zero), they do not contribute to the distance between two hits. Supplementary Figure S3b illustrates the distance between two hits as the sum of the distances between the regions they involve. In Supplementary Figure S3c, we see an example of an exon duplication, and two hits referring to the same exon.

Constructing a dendrogram: Using this distance measure between hits, we construct a dendrogram through a single linkage hierarchical clustering. An example is shown in Figure 2.1b. We first group all hits by the chromosomes on which the hits fall. For any pair of chromosomes (each on a different genome), Proteny builds a dendrogram, in which each node represents a cluster of hits. It is important to note that we cluster not exons, but hits. This strategy results in an exon being present in multiple clusters (and multiple dendrograms), allowing us to handle duplications.

However, the height of the tree reflects only the density of hits, not the quality of conservation within. Cutting the tree to produce clusters based on height alone will therefore not be sufficient. Instead, we define a cluster score which reflects our expecations of syntenic clusters.

A cluster score: As in [20], we consider a syntenic region good if it maximizes the similarity within the cluster, and minimizes the similarity between them. We characterize each cluster with a cluster score, which describes the similarity within the cluster, but is punished by the similarity to other regions. The similarity within the cluster is described by the quality of the hits which lie within the region, and the similarity to other regions is described by the quality of hits which fall within the genomic region defined by the cluster, but have no hits within the cluster (unaccounted exons). The quality of a hit should reflect the coverage of the hit over the exons it covers, and the significance of this hit. We therefore define a quality score $K(h_i)$ for a hit $h_i \in H$ between two exon sequences:

$$K(h_i) = \{1 - \min(1, E(h_i))\} \cdot \frac{||r_i^\beta|| + ||r_i^\gamma||}{||x_i^\beta|| + ||x_i^\gamma||}, \tag{2.3}$$

where x_i^{χ} is the exon the hit h_i refers to on genome χ, $||\cdot||$ is the length of a given sequence or region, and $E(h_i)$ is the e-value of the hit. The ratio represents the fraction of the size of the exons which are covered by the hits, favoring hits which cover the whole exon. This ratio is multiplied by $1-E(h_i)$ to factor in the significance of the hit, so that insignificant hits will deteriorate the score. Note that $K(\cdot) \in [0,1]$ where 1 is the perfect score.

Then, the cluster score, $s(C)$, accumulates the hit scores for the hits within the cluster, but is penalized by exons within the cluster which do not have a hit in the cluster C:

$$s(C) = 2 \cdot \sum_{h_i \in C} K(h_i) - \sum_{e \in U_{C}^{\beta}, \cup U_{C}^{\gamma}} \max_{h_j \in H_e} K(h_j)\,, \tag{2.4}$$

where U_C^{χ} is the set of exons on genome χ which are located within cluster C, but are unaccounted for within the cluster, and H_e are all the hits to exon e (for e from organism β or γ). If H_e is empty (i.e., the unaccounted exon has no hit to the other genome), then the cluster is not penalized (see Supplementary Figure S3d).

Note that the penalization for unaccounted exons is based on the maximum hit score. The main motivation for this is that if an unaccounted exon has a better hit somewhere else then it should not be in the current cluster. However, if the unaccounted exon does not have a hit anywhere on the other genome (H_e being empty), then, without knowing anything more about it, it should not affect the cluster score.

A dynamic cutting algorithm: Proteny cuts the dendrogram at a given node depending upon the significance of the cluster score assigned to that node (see next section). However, some clusters contain so many good hits that they may contain many large gaps (unaccounted exons), while still being significant. To counter that, we restrict ourselves to clusters which satisfy a minimum 'conservation ratio', given by the user-specified parameter τ. The conservation ratio τ_C of a cluster C, is defined as $\frac{n_C+1}{n_C^{\beta}+n_C^{\gamma}+1}$, where $n_C = |C|$, the number of hits in the cluster, $n_C^{\beta} = |U_C^{\beta} \cap H|$, the number of unaccounted exons on genome β which have a hit elsewere, and $n_C^{\gamma} = |U_C^{\gamma} \cap H|$, the number of unaccounted exons on genome γ, that have a hit elsewhere.

The dendrogram will therefore not be cut at a single height, but at different heights depending on the significance and the conservation ratio. For such an approach, a "dynamic tree cut", other methods exist [21, 22], but those do not rely upon a statistical significance to cut. We use a greedy cutting algorithm, given in Supplementary Equation SE-8. Starting at the root node, check if the current node satisfies the conservation ratio and has a lower p-value than its child nodes. If both are true, and the node is significant, then we cut at this node and we do not descend further into the tree. Alternatively, if the current node is not significant, or either of the child nodes have a lower p-value and satisfy the conservation ratio, we descend instead to the children.

Testing the significance of a cluster: In order to calculate the significance of a cluster, we must build a null distribution of cluster scores. Other methods which calculate the statistical signifigance of a cluster such as [23] do not take into account the similarity between clusters that our cluster score does. Therefore, we must build our own null distribution of cluster scores for each particular size of cluster (i.e. combination of n_C, n_C^{β}, and n_C^{γ}). Although a null distribution constructed from hits with random scores that are randomly distributed along the genome would be ideal, it is computationally infeasible as we would need to re-cluster at every iteration. Instead, we permute hit scores after the clustering, thereby assuming no fixed structure in successive hits, as would be the case if the hits were randomly distributed. Hence, the cluster score for one permutation becomes:

$$s_p(C) = 2 \cdot \sum_{k=1}^{n_C} P_k - \left(\sum_{k=1}^{n_C^{\beta}} P_k^{\beta *} + \sum_{k=1}^{n_C^{\gamma}} P_k^{\gamma *} \right) \tag{2.5}$$

where P_k is the k^{th} element of a randomly shuffled list of hit scores H (created by random reordering), and $P_k^{\chi *}$ is the k^{th} element of a randomly permuted list of best hits for each exon in organism χ (by taking only the best hit for each exon).

P-values can now easily be obtained by comparing the actual cluster score to the permuted scores. However, since many nodes in the dendrograms are tested, we need to correct for multiple testing. For a pair of organisms with $|H|$ hits between them, we would in the worst case perform $2|H|$ tests, calling for a bonferroni correction of $2|H|$. With such a correction, and a p-value threshold of α, at least $\left\lceil \frac{2|H|}{\alpha} \right\rceil$ permutations would be required just to achieve the resolution required to detect a cluster. This correction factor, and thereby the number of permutations can become very high, and we therefore wish to limit the number of permutations when possible.

Unfortunately, the inheritance procedure of [24], which controls the family-wise error rate for hierarchical tests does not apply, since our problem does not fulfill the condition that significant tests must have significant parent tests. Similarly, the same condition for the false-discovery-rate correction for trees of [25] are not met. We integrate four approaches which help us making the number of permutations more tractable.

Not considering all clusters: As we are interested in synteny (beyond homology statements between genes), we are not interested in clusters which are smaller than 2 genes, nodes in the dendrogram which contain fewer than 2 genes are not tested.

Early stopping: We can apply the method of [26], to limit further permutations when the number of exceedences is already sufficient. The cluster which is then not significant, will never be significant with more permutations. As we only wish to detect significant clusters, we can apply this strategy.

Analytical solution: If a cluster is large enough (see Supplementary Material S4.1), we can make use of an analytical description of the null distribution, based on the central limit theorem (CLT) described in equation Supplementary Equation SE-6. The cluster score is a sum of three different distributions, each component being a summation over random variables. Consequently, where possible, we use the CLT approximation for the null distribution. Only if the cluster size is too small do we revert to the permutation method.

Dynamic correction: Rather than using a worst-case multiple testing scenario to determine the number of tests to correct for, we determine the number of tests dynamically. That is, we start out by setting the initial number of tests to 1, and, following the dynamic cutting algorithm Supplementary Equation SE-8, we increase the correction factor only when we descend to a child node in the tree. Alternatively, if a node is called significant, we do not need to increase the correction factor.

For insignificant nodes, this is always allowed since it will only be *more* insignificant at higher correction factors. However, significant nodes will have to be revisited (since, with the larger correction factor they may become insignificant). The advantage here, is that when we need to revisit a node, we only need to do the additional permutations; i.e. we can still make use of the earlier permutations. This procedure is performed iteratively until no further tests are performed.

Visualization: Proteny provides two different types of visualization: i) a *chromosome*-level visualization, and ii) a *region*-level visualization. Chromosome-level visualizations allow us to have an overview of the relationships between two genomes. In this visualization (e.g. figure 2.3).The outer ring is the genome, the inner ring represents the genes (blue and orange representing genes on the forward and reverse strand respectively), and the ribbons between two loci represent a conserved cluster. The query chromosome is shown first, in a clockwise-fashion from 12 o'clock onward.

The region-level visualizations show only a few loci from both genomes (e.g. figure 2.4e). Again, the outer ring represents the regions on the genome and the inner ring represents genes. Now, additional green boxes within the genes represent exons. The intensity of each link represents the quality score $K(\cdot)$ of the hit. The ribbons no longer represent clusters, rather, they are the original BLASTp hits between exons.

Implementation details: For data handling we use the Ibidas [27] data query and manipulation suite, and the Circos[28] utility is used to visualize the discovered clusters. For more information, see Supplementary Material S6.

2.3. RESULTS

YEAST GENE ORDER BROWSER DATASET

The YGOB [18] provides a ground truth through a large set of ortholog relationships between 20 yeast genomes. We use the data and scores described in [20] to compare Proteny to i-ADHoRe. Since Proteny performs a pairwise synteny discovery analysis, the two scores are equivalent. We use the same parameters for i-ADHoRe (version 3.0.01) and fasta36 [29] as given in [20]. For Proteny, we used protein blast and a p-value threshold of 0.05 and a conservation threshold of 1. We score the clusters that i-adhore and Proteny find using the relaxed score in [20].

Figure 2.2 shows the means of the relaxed scores for all clusters in each of the pairwise tests for both Proteny and i-ADHoRe. Proteny had a higher average cluster score in 16 out of 28 experiments. In 15 of these, the relaxed cluster score distributions were significantly different (by a bonferroni corrected Kolmogorov-Smirnov test), see also Supplementary Figure S12c. Although i-ADHoRe had higher average relaxed scores in 12 experiments, in 10 of these cases, the distributions of relaxed scores are not significantly different. Based on this, Proteny either performs comparably to, or outperforms i-ADHoRe on this dataset.

Aspergillus niger

We study two strains of *Aspergillus niger*, which have been separated by 50 years of evolution, n402 and CBS513.88. *A. niger* CBS513.88 is an industrial strain which is used as

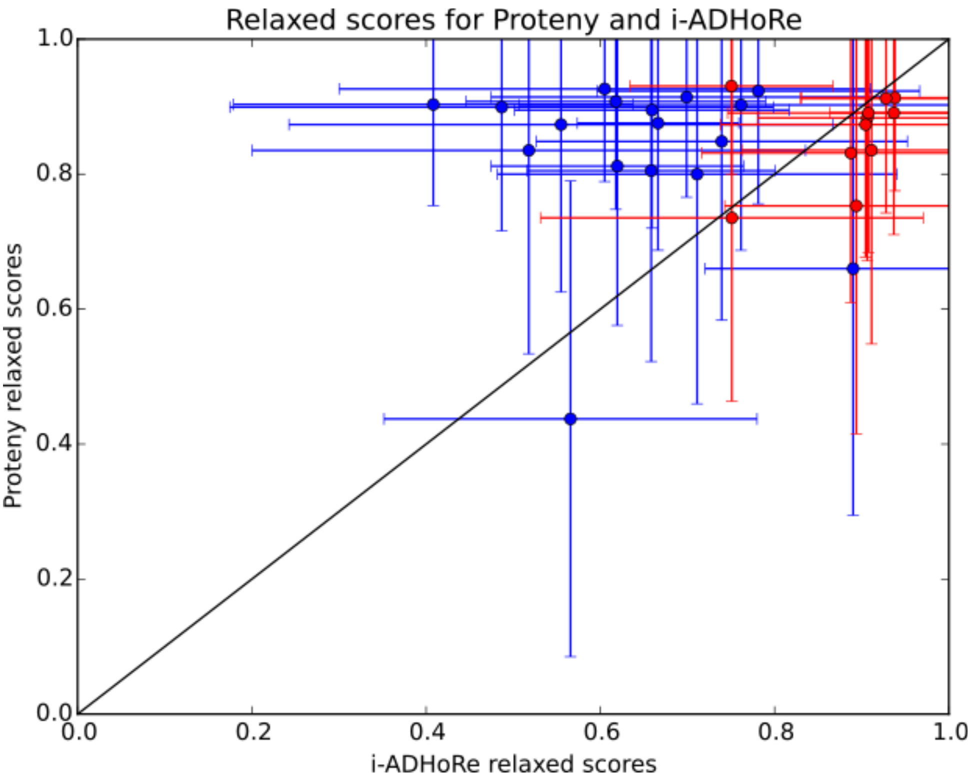
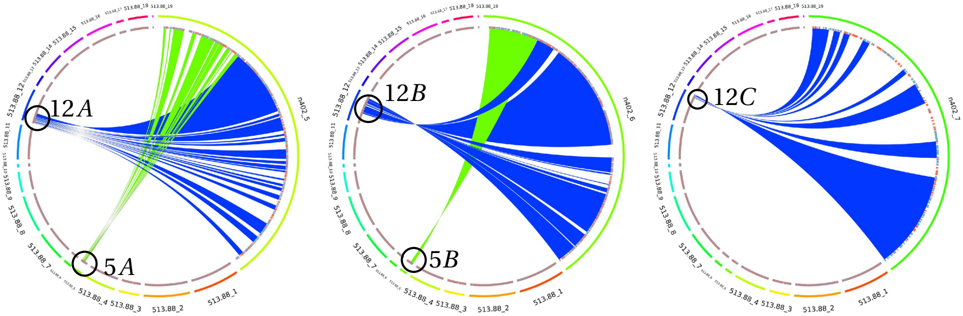


Figure 2.2: *The means and their standard deviations of relaxed scores across all syntenic clusters for Proteny (y-axis) and i-ADHoRe (x-axis) clusters, for each pairwise test. Red points are cases where the relaxed score distributions are not significantly different between Proteny and i-ADHoRe (*$q < 0.05$ *Kolmogorov-Smirnov test with bonferroni multiple testing correction).*

a cell factory for enzyme and metabolite production, while n402 is a lab strain used in research. We use this to demonstrate the performance of the method. Since we know that the two strains must be highly related, we expect to find large similarities between the two genomes. The *Aspergillus niger* CBS513.88 genome [30] and annotation was retrieved from the Aspergillus genome database [31]. The de novo genome sequence of lab strain *A. niger* n402 is unpublished at time of writing (see acknowledgement). The n402 strain has 13,612 genes, while the industrial strain CBS513.88 has 14,067 genes. Due to incomplete genome assemblies, we deal with scaffolds rather than chromosomes. The n402 and CBS513.88 strains have 24 and 19 scaffolds, respectively. For this dataset, we set $\tau = 2$, because we assume that the two strains are quite similar.

General synteny: Proteny finds high conservation between the two strains. In total, Proteny finds 119 syntenic clusters, covering 10.880 n402 genes, and 10.956 genes in CBS513.88 (see Supplementary Table S2). We compare Proteny's results to those of i-ADHoRe, as it is the only tool that also works on the protein level and is not specifically designed for similar genomes. i-ADHoRe finds 189 syntenic clusters, covering 9.667



(a) *n402 scaffold 5 (Proteny)* **(b)** *n402 scaffold 6 (Proteny)* **(c)** *n402 scaffold 7 (Proteny)*

Figure 2.3: *Syntenic clusters found for Aspergillus niger on* ***(a)*** *scaffold 5,* ***(b)*** *scaffold 6, and* ***(c)*** *scaffold 7 in n402. For scaffolds 5 and 6, one can see that n402 has undergone a rearrangement.*

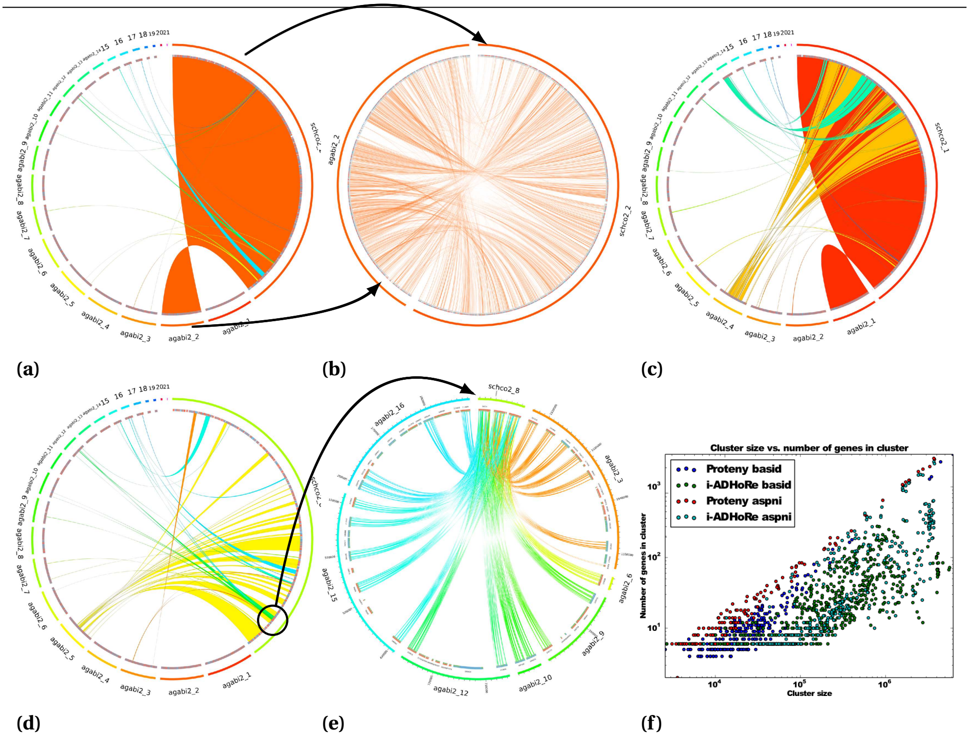
(9.310 in common with Proteny), and 9728 genes (9343 in common) from the n402 and CBS513.88 strains, respectively. We find that 66.5% of the area covered by the clusters discovered by Proteny and i-ADHoRe is found by both algorithms.

By calculating the score for each i-ADHoRe cluster using our scoring function, we find that only 93 (49.2%) of the clusters that i-ADHoRe finds are significant (see Supplementary Table S6), and most have a very small conservation ratio (see Supplementary Figure S10b). Furthermore, we see in Figure 2.4f, that Proteny generally has more genes in i-ADHoRe clusters of the same size. From this we conclude that Proteny finds more genes in fewer clusters. Apparently Proteny discovers informative clusters that tightly describe the syntenic genes.

Identifying a genome rearrangement: Figures 2.3a and 2.3b show the syntenic clusters Proteny discovers for n402 scaffolds 5 and 6, respectively. These figures show that *A. niger* n402 was formed by a rearrangement: parts 5A and 12A from CBS513.88 (see figure 2.3a) have fused together to form scaffold 5 in the n402 strain. Likewise, scaffold 6 from n402 was formed by the fusion of parts 5B and 12B (see figure 2.3b). From these detected syntenic regions, one can conclude that scaffolds 5 and 12 of CBS513.88 have split in two and fused together over time to form two scaffolds in n402. When comparing to i-ADHoRe, (Supplementary Figures S5a-c), we see that Proteny gives a clearer synteny (i.e. i-ADHoRe is cluttered with other supposed syntenies), and at the same time Proteny gives more detail on the fused or separated syntenic regions.

Assisting genome assembly: Proteny can also assist in genome assembly. By studying the visualizations, we can quickly inform ourselves about the results of an assembly. Figure 2.3c shows that scaffold 7 of n402 maps to part 5C from CBS513.88, whereas, according to figure 2.3b scaffold 6 of n402 maps to the connecting part 5B in CBS513.88. This indicates that scaffolds 6 and 7 in the n402 assembly could be joined together.

The effect is even more pronounced in Supplementary Figures S7a-c. We see that three chromosomes in n402 map to a single chromosome in the CBS513.88 genome. Proteny can guide an assembly and suggest that they be joined together in the n402 genome, as in the CBS513.88 genome.



(a) (b) (c)

(d) (e) (f)

Figure 2.4: ***(a)*** *The synteny between scaffold 2 in S. commune and the scaffolds in A. bisporus.* ***(b)*** *The hits between scaffold 2 in S. commune and scaffold 2 in A. bisporus show a scattered synteny.* ***(c)*** *The syntenic relationships between scaffold 1 in S. commune and the scaffolds in A. bisporus. The three scaffolds 1, 4, and 14 share a lot in common with scaffold 1 in S. commune, and it follows that, at some point the two species diverged when the scaffold split in the branch of A. bisporus, but not in the branch of S. commune.* ***(d)****-* ***(e)*** *A repeated gene elucidates a divergent trait between A. bisporus and S. commune.* ***(f)*** *A Proteny cluster of a given size generally has more genes than an i-ADHoRe cluster of the same size. In this figure, 'basid' refers to the basidiomycete analysis, and 'aspni' refers to the A. niger analysis. It can be seen in this figure that Proteny cluster gene densities are higher than those of i-ADHoRe and a t-test with unequal variance assumptions states that the distributions of cluster scores are separated with a p-value of* 9.1×10^{-75}.

BASIDIOMYCOTA

Next, we applied Proteny to *Schizophyllum commune* [32], a model organism for mushroom formation, and *Agaricus bisporus* [33], which is a commercially valuable mushroom but has a relatively large evolutionary distance to *S. commune*. We retrieved the genomes and genome annotation files for *Schizophyllum commune v*2.0 and *Agaricus bisporus v*2.0 from the JGI genome portal [34]. *S. commune* has 14.652 genes, and *A. bisporus* has 10.438 genes. As before, these organisms have incomplete genome assemblies, with 36 and 31 scaffolds respectively. For this experiment, we set $\tau = 1$, because while we assume a lot of divergence, we are interested in conserved clusters.

General synteny: Proteny finds 345 significant clusters, covering 5.828 *S. commune* genes lying within conserved regions, and 4.572 *A. bisporus* genes (see table Supplementary Table S2). Many exons do not have a bi-directional BLASTp hit, resulting in many smaller clusters. i-ADHoRe discovers 377 clusters which cover 2.588 *S. commune* genes (2.889 in common with Proteny) and 4.090 *A. bisporus* genes (2.662 in common). 41.4% of the area covered by the clusters discovered by Proteny and i-ADHoRe is found by both algorithms. From these clusters (see Supplementary Table S7), we find that, using our scoring function, only 327 (57.1%) i-ADHoRe clusters are significant. The reason we find so many more genes than i-ADHoRe stems from the orderless detection of the clusters. The results from Proteny show that although both mushrooms are evolutionarily distant, a large portion of the genes remain conserved.

Figure 2.4f shows the genomic size of a Proteny cluster is smaller than that of an i-ADHoRe cluster containing the same number of genes. The figure also shows that a Proteny cluster of a given size generally has more genes than an i-ADHoRe cluster of the same size. i-ADHoRe clusters contain more unaccounted exons (gaps), confirmed by Supplementary Figures S10b and e. Again, we conclude that Proteny finds fewer clusters which harbor more genes.

Large similarities: We even observe large similarities between the diverged genomes, as shown in figure 2.4a between scaffold 2 in *S. commune* and scaffold 2 in *A. bisporus* When we look at this cluster more closely in Figure 2.4b we see that the hits are very dense. Figure 2.4a also shows that Proteny results in a much clearer synteny between the scaffolds than i-ADHoRe (Supplementary Figure S6b), which includes many other clusters which occlude the results. This can be attributed to the result of the orderless synteny detection of Proteny. i-ADHoRe discovers more clusters which contain large stretches of gaps between genes. For example, the additional cluster between chromosome 2 in *S. commune* and chromosome 7 in *A. bisporus* seen in Supplementary Figure S6b, which is not found by Proteny in figure 2.4a, contains only a few spurious hits between a few genes with many unaccounted exons (Supplementary Table S6, no. 141).

Scattered synteny: Figure 2.4c shows that scaffold 1 of the *S. commune* assembly consists primarily of three scaffolds in the *A. bisporus* assembly, over a number of syntenic blocks. Clearly, the three *A. bisporus* scaffolds 1, 4, and 14 have a lot in common with scaffold 1 in *S. commune*, and it follows that at some point the two species diverged when the scaffold split in the branch of *A. bisporus*, but not in the branch of *S. commune*. Alternatively, it is possibe that this 'scattered' effect arises from that scaffolds 1, 4 and 14 in *A. bisporus* are not correctly assembled. Despite the syntenic regions between the two fungi is highly scattered, (either through evolution by an incomplete assembly) Proteny is able to detect this and the visualisations allow us to explore them intuitively. For i-ADHoRe the results are harder to examine, see Supplementary Figure S6a.

Gene duplication: Figure 2.4d shows the syntenic clusters detected in scaffold 8 of *S. commune*. We see an interesting phenomenon here: there is a region in *S. commune* which is repeated several times in *A. bisporus*. Figure 2.4e zooms in on this region. Here it becomes clear that there are three genes which are duplicated many times in *A. bisporus*. These genes are cytochrome P450 monooxygenases which are involved in metabolism detoxification [35], and are expected to be involved in the detoxification of byproducts from lignin degradation. The fact that *S. commune* has fewer copies of the

P450 compared to *A. bisporus* highlights the fact that *S. commune* does not have the ability to degrade lignin, while *A. bisporus* does. The speciation event which separated *S. commune* and *A. bisporus* came before *A. bisporus* was able to degrade lignin, and can be derived from the number of P450 copies in *A. bisporus*. Figure 2.4e reveals that many of these duplications are not entirely conserved, often exons are missing, or new ones are there instead, exemplifying the benefit of the exon-level analysis. This again shows the capabilities of Proteny (i-ADHoRe does not detect this region, Supplementary Figure S6c).

Developmental proteins are conserved: We are particularly interested in eight transcription factors and a light sensing protein which have been linked to mushroom formation in *S. commune* [36]. To increase the confidence that these transcription factors are functionally similar in both *S. commune* and *A. bisporus*, we wish to find that these genes lie in syntenic regions. Proteny reveals that six of these nine developmental proteins lie within conserved clusters. Supplementary Figure S8c shows the region-level plot for the cluster which contains the transcription factor *gat1*. The figure clearly shows that the transcription factor lies in a well conserved region, i.e. neighboring genes in *A. bisporus* match to neighboring genes in *S. commune*. Supplementary Figure S8 shows the region-level plots for the clusters of the other developmental proteins found in syntenic clusters, and Supplementary Figure S9 shows the developmental proteins which were *not* found.

2.4. DISCUSSION

We presented Proteny, a methodology which identifies significant conserved syntenic clusters of exons between two genomes through a novel method for cutting dendrograms, and a new dynamic multiple testing correction algorithm. Knowledge of the discovered clusters allow us to uncover genome rearrangement events (as shown for both the *A. niger* strains and the Basidiomycota), make more motivated statements about functional conservation (as for *S. commune*), identify possible errors in the assembly of related genomes (like in *A. niger*), and study the evolution between species (as in looking at the cytochrome P450 monooxygenases in *A. bisporus* and *S. commune*).

When comparing to i-ADHoRe, the most competitive tool, on a ground truth dataset, we find Proteny outperforms i-ADHoRe. Qualitatively, we observe that i-ADHoRe finds more clusters, covering fewer genes than Proteny clusters. Proteny finds gene-dense clusters of high quality, as verified by the cluster scores achieved by Proteny on the YGOB dataset. This can be attributed to the statistical testing procedure, and the conservation ratio we enforce in Proteny.

One practical advantage of Proteny over other synteny tools is that, besides the BLASTp settings (for which we used default values in our experiments), it only requires specifying a significance threshold (which can be set by statistical reasoning), and a conservation ratio parameter τ. It should be noted that Proteny could work with other aligners also, and that BLASTp could be replaced by other protein sequence aligners.

On the other hand, i-ADHoRe, and many other tools are able to perform an analysis on more than two genomes at a time. Proteny could be generalized towards any number of species by a progressive heuristic similar to the star multiple alignment heuristic [37], which uses a central sequence with pairwise sequence alignments to guide the multiple alignment.

It is important to note that the cluster score of Proteny does not account for the conservation of the order of the exons within the cluster. This can most prominently be seen in the synteny between scaffold 2 of *S. commune* and scaffold 2 of *A. bisporus*, (Figures 2.4a-b), where the order of the hits is scrambled. Although the ordering of the exons does play a role when constructing the dendrogram (nearby hits are merged first), we chose that the ordering should not play a role when scoring the clusters. This was a deliberate choice since Proteny was designed to find synteny between relatively divergent organisms in a microbiology context where evolution is fast; insertions, inversions, strand changes and gene shuffling occur frequently. Clearly, in other problem settings the order may be important, in which case the cluster score in Proteny should be adjusted. However, one should be careful when designing a corresponding permutation scheme, e.g. a circular permutation of hit scores to preserve the order of hits in that set, as this might result in computational difficulties, as (for example) the CLT approximation will not hold anymore.

By searching for synteny at the exon level, we exclude the influence of noncoding regions of the genome, which are typically not well conserved between divergent genomes. While an analysis at the gene level is interesting, we reasoned that it makes more sense to look at the conservation of individual exons within the gene. The region-level visualizations indeed show that conservation is higher at the exon level than at the gene level, i.e. some exons may be missing, making the gene less conserved, while individual exons are conserved.

The ability to give each cluster a p-value is an important contribution. However, the null distribution assumes that there is a completely random relationship between the organisms, which is not true. Currently, the τ parameter, representing a lower bound on their conservation ratio (in terms of the ratio of conserved and non-conserved exons) is used to regulate the clusters. Future contributions could develop a null model which takes into account evolution between two species. For example, a better permutation may shuffle groups of exons (genes), rather than individual exons. Yet as indicated earlier, this will give rise to computational difficulties.

Another important consideration is that exons which do not have a hit with the other organism increase the distance between hits when constructing the dendrogram, but they do not penalize the cluster score. Again, this was a deliberate choice in order to be less sensitive to evolutionary insertions and deletions, but could be changed by using a different distance measure.

Altogether, Proteny is a powerful tool which can detect synteny between relatively divergent genomes at the amino acid sequence level. It detects clusters of exons based on a significance test and provides a rich visualization which supports the interpretation of the detected syntenic regions.

BIBLIOGRAPHY

[1] R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, N. Maltsev. Use of contiguity on the chromosome to predict functional coupling. *In silico biology* 1(2):93–108 (1999).

[2] P. E. McClean, et al. Synteny mapping between common bean and soybean reveals extensive blocks of shared loci. *BMC genomics* 11:184 (2010).

[3] A. U. Sinha, J. Meller. Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC bioinformatics* 8:82 (2007).

[4] D. Vallenet, et al. MaGe: a microbial genome annotation system supported by synteny results. *Nucleic acids research* 34(1):53–65 (2006).

[5] S. V. Angiuoli, S. L. Salzberg. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics (Oxford, England)* 27(3):334–42 (2011).

[6] A. C. E. Darling, B. Mau, F. R. Blattner, N. T. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research* 14(7):1394–403 (2004).

[7] M. Blanchette, et al. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome research* 708–715 (2004).

[8] I. Minkin, A. Patel, M. Kolmogorov. Sibelia: A scalable and comprehensive synteny block generation tool for closely related microbial genomes. *Algorithms in Bioinformatics* 8126(1) (2013).

[9] X. Zeng, M. Nesbitt, J. Pei, K. Wang. OrthoCluster: a new tool for mining synteny blocks and applications in comparative genomics. *Proceedings of the 11th international conference on Extending database technology: Advances in database technology* 656–667 (2008).

[10] C. Soderlund, M. Bomhoff, W. M. Nelson. SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic acids research* 39(10):e68 (2011).

[11] K. Vandepoele, Y. Saeys, C. Simillion, J. Raes, Y. Van De Peer. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome research* 12(11):1792–801 (2002).

[12] C. Simillion, K. Janssens, L. Sterck, Y. Van de Peer. i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics (Oxford, England)* 24(1):127–8 (2008).

[13] S. Proost, et al. i-ADHoRe 3.0–fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic acids research* 40(2):e11 (2012).

[14] M. Long, E. Betrán, K. Thornton, W. Wang. The origin of new genes: glimpses from the young and old. *Nature reviews Genetics* 4(11):865–75 (2003).

[15] P. Husemann, J. Stoye. R2Cat: Synteny Plots and Comparative Assembly. *Bioinformatics (Oxford, England)* 26(4):570–1 (2010).

[16] C. D. Shaw. Genomic Spring-Synteny Visualization with IMAS. *2008 Fifth International Conference BioMedical Visualization: Information Visualization in Medical and Biomedical Informatics* 3–8 (2008).

[17] M. J. Sullivan, N. K. Petty, S. a. Beatson. Easyfig: a genome comparison visualizer. *Bioinformatics (Oxford, England)* 27(7):1009–10 (2011).

[18] K. P. Byrne, K. H. Wolfe. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome research* 15(10):1456–61 (2005).

[19] S. F. Altschul, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25(17):3389–402 (1997).

[20] C. G. Ghiurcuta, B. M. E. Moret. Evaluating synteny for improved comparative studies. *Bioinformatics (Oxford, England)* 30(12):i9–18 (2014).

[21] P. Langfelder, B. Zhang, S. Horvath. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics (Oxford, England)* 24(5):719–20 (2008).

[22] M. J. Mason, G. Fan, K. Plath, Q. Zhou, S. Horvath. Signed weighted gene coexpression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC genomics* 10:327 (2009).

[23] K. Jahn, S. Winter, J. Stoye, S. Böcker. Statistics for approximate gene clusters. *BMC bioinformatics* 14 Suppl 15(Suppl 15):S14 (2013).

[24] J. J. Goeman, L. Finos. The inheritance procedure: multiple testing of tree-structured hypotheses. *Statistical applications in genetics and molecular biology* 11(1):Article 11 (2012).

[25] D. Yekutieli. Hierarchical False Discovery Rate–Controlling Methodology. *Journal of the American Statistical Association* 103(481):309–316 (2008).

[26] T. a. Knijnenburg, L. F. a. Wessels, M. J. T. Reinders, I. Shmulevich. Fewer permutations, more accurate P-values. *Bioinformatics (Oxford, England)* 25(12):i161–8 (2009).

[27] M. Hulsman, J. J. Bot, A. P. de Vries, M. J. T. Reinders. Ibidas: Querying Flexible Data Structures to Explore Heterogeneous Bioinformatics Data. *Data Integration in the Life Sciences* 23–37 (2013).

[28] M. Krzywinski, et al. Circos: an information aesthetic for comparative genomics. *Genome research* 19(9):1639–45 (2009).

[29] W. R. Pearson. Empirical statistical estimates for sequence similarity searches. *Journal of molecular biology* 276(1):71–84 (1998).

[30] H. J. Pel, et al. Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. *Nature biotechnology* 25(2):221–31 (2007).

[31] M. B. Arnaud, et al. The Aspergillus Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the Aspergillus research community. *Nucleic acids research* 38(Database issue):D420–7 (2010).

[32] R. a. Ohm, et al. Genome sequence of the model mushroom Schizophyllum commune. *Nature biotechnology* 28(9):957–63 (2010).

[33] A. R. Baker, et al. Genome sequence of the button mushroom Agaricus bisporus reveals mechanisms governing adaptation to a humic-rich ecological niche. *Proceedings of the National Academy of Sciences of the United States of America* 110(10):4146 (2013).

[34] I. V. Grigoriev, et al. The genome portal of the Department of Energy Joint Genome Institute. *Nucleic acids research* 40(Database issue):D26–32 (2012).

[35] B. Crešnar, S. Petrič. Cytochrome P450 enzymes in the fungal kingdom. *Biochimica et biophysica acta* 1814(1):29–35 (2011).

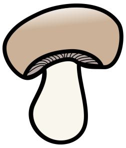
[36] R. a. Ohm, D. Aerts, H. a. B. Wösten, L. G. Lugones. The blue light receptor complex WC-1/2 of Schizophyllum commune is involved in mushroom formation and protection against phototoxicity. *Environmental microbiology* (2012).

[37] S. F. Altschul, D. J. Lipman. Trees, Stars, and Multiple Biological Sequence Alignment. *SIAM Journal on Applied Mathematics* 49(1):197–209 (1989).

Chapter 3

The transcriptional regulator *c2h2* accelerates mushroom formation in *Agaricus bisporus*

Jordi F. Pelkmans
Aurin M. Vos
Karin Scholtmeijer
Ed Hendrix
Johan J.P. Baars
Thies Gehrmann
Marcel J. T. Reinders
Luis G. Lugones
Han A. B. Wösten



This chapter is published as: *Applied Microbiology and Biotechnology*, August 2016, Volume 100, Issue 16, pp 7151–7159 (doi: 10.1007/s00253-016-7574-9)
Supplementary material is available online at: `http://bit.ly/2mNy9o5` (springer.com)

ABSTRACT

The Cys2His2 zinc finger protein gene c2h2 of Schizophyllum commune is involved in mushroom formation. Its inactivation results in a strain that is arrested at the stage of aggregate formation. In this study, the c2h2 orthologue of Agaricus bisporus was over-expressed in this white button mushroom forming basidiomycete using Agrobacterium¬-mediated transformation. Morphology, cap expansion rate, and total number and biomass of mushrooms was not affected by over-expression of c2h2. However, yield-per-day of the c2h2 over-expression strains peaked one day earlier. These data and expression analysis indicate that C2H2 impacts timing of mushroom formation at an early stage of development, making its encoding gene a target for breeding of commercial mushroom strains.

INTRODUCTION

The basidiomycete *Agaricus bisporus* is cultivated globally for the production of white button mushrooms. These fruiting bodies have a relative high protein content and contain fibers, vitamins, minerals, and bioactive compounds. *A. bisporus* is grown on compost formed from wheat straw, horse or chicken manure, and gypsum. During colonization, the compost is topped with a casing layer needed for moisture and microbial flora [1–3]. Induction of mushroom formation depends on environmental signals. The volatile 1-octen-3-ol represses early development, while high temperature (i.e. 25 °C instead of 18 °C) inhibits development from smooth to elongated primordia. On the other hand, CO2 impacts the number of fruiting bodies that are formed [4, 5]. Development of *A. bisporus* is a complex process [6]. It starts with aggregation of hyphae into hyphal knots [7]. These structures develop into 1-2 mm initials, also called primordia, that differentiate by forming cap and stem tissues [7]. Up to 10% of differentiated primordia develop into mushrooms [8]. Breaking of the veil of these fruiting bodies enables airborne dispersal of basidiospores that had been formed in the gill tissue within the cap.

Production conditions of white button mushrooms have been optimized with respect to yield and quality of fruiting bodies [9]. However, molecular mechanisms underlying mushroom formation are poorly understood. For instance, transcription factors (TFs) involved in white button development have not been identified so far. Such regulatory proteins have been identified in the model organism *Schizophyllum commune* [10–12]. Formation of its fruiting bodies is induced by blue light and is repressed by high CO2 [12–15]. The blue light receptor complex consists of the sensor WC-1 and TF WC-2. Inactivation of wc-1 and / or wc-2 results in a blind strain not able to produce aggregates, primordia, and fruiting bodies [12]. Strains in which the homeodomain gene hom2 or the zinc finger TF gene fst4 have been inactivated are also not able to produce aggregates [10, 11] In contrast, inactivation of the gene encoding the Cys2His2 zinc finger protein C2H2 results in a strain that does form aggregates but primordia and fruiting bodies are not formed [11]. Strains in which genes are inactivated that encode the zinc finger protein Fst3, the GATA type zinc finger protein Gat1, or the homeodomain protein Hom1 form smaller fruiting bodies but in higher numbers [11]. These proteins were proposed to play a role in repression of outgrowth of primordia into fruiting bodies or to play a role in expansion of the fruiting body.

Homologues of the *S. commune* TFs involved in fruiting body development have been identified in other mushroom forming fungi. Expression analysis in *A. bisporus*, *Lacaria bicolor*, and *Coprinopsis cinerea* suggest that mushroom development in the Basidiomycota follows a core regulatory program with species specific variations that explain differences in morphology and sensitivity to environmental signals [10, 16–18]. In this study, the *A. bisporus c2h2* homologue was over-expressed in the commercial A15 strain of this mushroom forming fungus. This resulted in an accelerated rate of mushroom production. Experimental data indicate that C2H2 functions both early and late in mushroom development and that it is an interesting target for breeding of commercial strains.

MATERIALS AND METHODS

CULTURE CONDITIONS AND STRAINS

The heterokaryotic *A. bisporus* strain A15 (obtained from the fungal collection of Plant Breeding Wageningen UR, the Netherlands) and its derivatives AT273-1 and AT273-5 that over-express *c2h2* were routinely grown at 25 °C on malt extract agar medium (MEA; 20 gr l-1 malt extract agar [BD biosciences, Franklin Lakes, USA], 2.1 gr l-1 MOPS, pH 7.0). Spawn substrate was produced by heating 75 gr of Sorghum seeds in water at 100 °C for 20 min, after which 24 g kg-1 CaSO4 and 6.87 g kg-1 CaCO3 was added. Spawn was colonized for 3 weeks at 25 °C using 2 1-week-old *A. bisporus* colonies as inoculum. Mushrooms were produced by inoculating boxes (40 cm width x 60 cm length x 22 cm height) containing 16 kg phase 2 compost (CNC, Milsbeek, The Netherlands) with 75 gr of spawn. Compost temperature was maintained at 25 °C with an air temperature of 22 °C. Relative humidity in growth cells was kept at 95%, while CO2 levels fluctuated between 1500 ppm and 2000 ppm. 10 boxes were inoculated per strain and were randomly distributed in the growth cell. After 16 days, the compost in each box was topped with 7 kg casing layer (CNC, Milsbeek, The Netherlands). Growth was prolonged for 14 days before venting. The casing was manually broken 4 days prior to venting and mixed to create fast regenerative growth and a more equal distribution of *A. bisporus* in the casing layer. Venting resulted in a gradual decrease of compost and air temperature to 19 and 18 °C, respectively. Relative humidity and CO2 levels decreased gradually to 85% and 1200 ppm, respectively. The first buttons were removed from the bed 9 days after venting.

ANALYSIS OF MUSHROOM FORMATION

Photos of casing layer surfaces were taken in a fixed rig at 24 h intervals from venting until the start of the first flush. Emergence of mushrooms and growth rate of the caps was monitored using ImageJ[19]. Harvesting of mushrooms was done by a professional picker as performed in commercial production. Prior to the flushes some buttons were removed to open up the space between developing buttons. Fruiting bodies with a diameter between 40 and 60 mm were always harvested, while fruiting bodies with a diameter of ≤ 40 mm were picked from densely populated areas to provide more space, water, and nutrients to the remaining mushrooms, thereby ensuring optimal yield. Mushrooms were classified as size 40 (mushrooms with a cap ≤ 40 mm) and size 60 (mushrooms with a cap between 40-60 mm). Mushrooms were harvested in two flushes. All mushrooms had reached a size ≥ 40 mm during the second flush at day 22 and all fruiting bodies were therefore harvested, thus completing the experiment. Yield per box was expressed as the biomass and the number of harvested mushrooms. Height and width of cap and stem were determined of 10 randomly selected mushrooms per box during the peak day of the first flush. Dry weight of the mushrooms was assessed by drying 200 gr wet weight fruiting bodies at 100 °C. Relative dry weight is defined as the dry weight compared to the original wet weight.

OVER-EXPRESSION OF *c2h2*

Primer pair McSpBH_F/McSpBH_R (Table 3.1) was used to introduce PacI and AscI sites into pBHg [20], creating pBHgPA. Gene *c2h2* of *A. bisporus* (ProteinID 230069, http://genome.jgi.doe.gov/Agabi_varbisH97_2) encompassing its coding region with 750 bp up- and downstream sequences was amplified by PCR using genomic DNA of *A. bisporus* A15, primer pair C2h2Abo7wnF/ C2h2ABownR (Table 3.1) and Phusion Hot Start II High-Fidelity DNA polymerase (Thermo Fisher Scientific, Waltham, USA). The amplicon contained PacI and AscI linkers at its 5' and 3' ends, respectively, enabling its introduction in pBHgPA that had been cut with PacI and AscI. The resulting plasmid pKS273 was introduced in Agrobacterium tumefaciens AGL-1[20]. Transformation of *A. bisporus* A15 gills was performed as described (Romaine and Chen 2005). Transformants were screened on MEA plates containing 25 μg ml-1 hygromycin, 200 μM cefotaxime, and 100 µg ml-1 moxalactum. Transformants were transferred to a second selection plate containing 40 μg ml-1 hygromycin, 200 μM cefotaxime, and 17 μg ml-1 tetracycline.

Table 3.1: *Primers used in this study*

Primer name	Sequence
C2h2ABownF	CGCTTAATTAACCTGGCAAAAAAGTGAAC
C2h2ABownR	ATATGGCGCGCCACTACGTCGATGATCATG
McSpBH_F	GATCGTTAATTAAGAATTCAGATCTCAATTGGGCGCGCC
McSpBH_R	GGCGCGCCCAATTGAGATCTGAATTCTTAATTAAC

WHOLE GENOME EXPRESSION ANALYSIS

Mycelium in the casing layer, initials, stage I and stage II buttons, and young fruiting bodies of *A. bisporus* strain A15 were harvested 9 days after venting from two distinct places of the casing bed (thereby creating biological duplo's). Due to the method of cultivation at the commercial hand-picking grower Maatschap van den Heuvel, de Rips, The Netherlands, all developmental stages were present on the casing bed at this time point. Casing mycelium was harvested with casing soil. The initials were pooled to obtain sufficient material for RNA isolation. A single stage I button was divided in cap and stipe using a scalpel. A stage II button and a young fruiting body were divided into components of the stipe (skin, underlying tissue and center) and cap (skin, underlying tissue, gill tissue and veil). Samples were immediately frozen in liquid nitrogen. The casing mycelium sample was broken in pieces and kept frozen with liquid nitrogen while harvesting mycelium using cooled tweezers. Samples were homogenized using the TissueLyser II (Qiagen, Düsseldorf, Germany) and RNA was purified using the NucleoSpin RNA kit (Macherey-Nagel, Düren, Germany). Quality was assessed by gel electrophoresis and sent to ServiceXS (Leiden, the Netherlands) for Illumina Next Generation Sequencing. RNA sequencing data have been deposited at NCBI under accession PRJNA309475.

Sequencing revealed between 20,002,387 and 38,840,092 reads. The RNA-Seq pipeline used the TRIMMOMATIC read trimmer version 0.32[21] to remove low quality regions and the ILLUMINA adapters from the 125 bp paired end reads. These filtered reads that made up 79-86% of the initial reads were aligned to the *A. bisporus* v3.0 genome (Sonnenberg, unpublished data) using STAR aligner version 2.4.0f1 [22]. The size of the introns was limited to 1500 bp based on the largest intron sizes in the genome annotation provided by the Joint Genome Institute of the Department of Energy (JGI DOE). This resulted in an alignment of 80-93% of the filtered reads. Abundance estimation was calculated with Cufflinks version 2.1.1 [23], and differential expression tests were performed by Cuffdiff using a Benjamini Hochberg false discovery rate of 0.05 [24]. Proteins annotated to contain a DNA-binding or regulatory protein domain in the InterPro annotation predictions provided by JGI DOE were considered TFs.

STATISTICAL ANALYSES

For the statistical tests analysing the differences in mushroom measurements, we performed permutation tests to circumvent the assumption of normality of the data[1]. Within each test 1,000,000 permutations were performed. P-values were corrected with a Benjamini Hochberg procedure using a false discovery rate of 0.05.

RESULTS

WHOLE GENOME EXPRESSION ANALYSIS

RNA composition of casing mycelium, initials, stage I and II buttons, and young fruiting bodies of the commercial *A. bisporus* strain A15 were determined. Initials consisted of 1-2 mm hyphal knots, while stage I buttons were 4-5 mm in diameter. Stage II buttons showed differentiation within the cap and stipe tissue. For instance, gills had developed. Young fruiting bodies were between 15 and 20 mm in diameter and still had their gills covered with veil. Stage I and II buttons and young fruiting bodies were dissected into stipe and cap. Stipes of stage II buttons and young fruiting bodies were subdivided into skin, underlying tissue, and central tissue, while caps were subdivided in skin, underlying tissue, gill tissue, and veil tissue. 875 and 707 genes were up- and downregulated, respectively, in initials when compared to casing mycelium (Figure 3.1A). In stipes and caps of stage I buttons 1194 and 1496 genes were upregulated when compared to initials, while 1788 and 2225 genes were downregulated. The number of genes that were upregulated in tissues of stage II buttons ranged from 39 to 147, while 105 to 503 genes were downregulated when compared to the caps and stipes of stage I buttons. The number

[1] In this chapter, we present the original text as in the publication. However, it has come to our attention that the permutation test with 1,000,000 permutations is not an appropriate procedure when there are only 10 measurements per group (permutations are guaranteed to be performed more than once). Therefore, we also performed wilcoxon rank sum tests to test for population differences. We achieved the same significant results as with the permutation test.

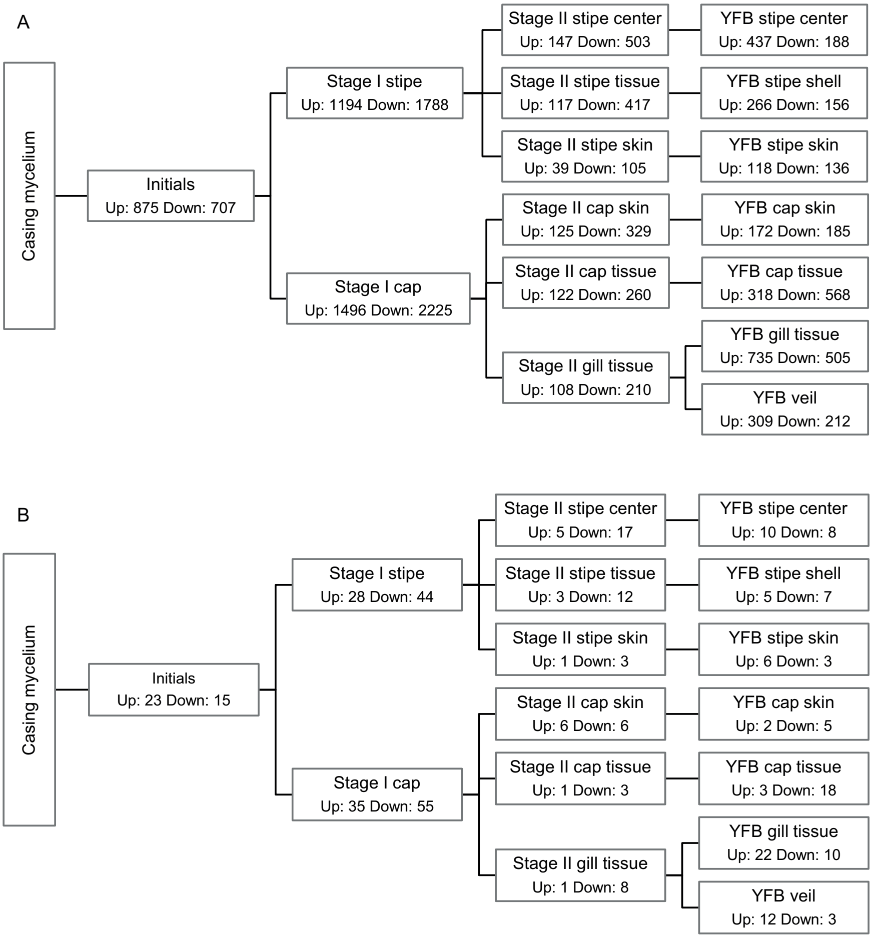


Figure 3.1: *Total up- and downregulated genes (A) and up- and downregulated TF genes (B) comparing initials, stage I buttons, stage II buttons, and young fruiting bodies (YFB) with the preceding developmental stage.*

of upregulated genes in young fruiting body tissues ranged from 118 to 735 compared to the stage II button tissues, while 136 to 568 genes were downregulated.

The overall number of TF genes that were upregulated ranged from 1 to 35 when consecutive stages were compared, while the number of downregulated regulatory genes ranged from 3 to 55 (Figure 3.1B, Table S1 in the Supplementary Material). The most prominent changes were observed when initials and caps of stage I buttons were compared (90 differentially expressed TF genes). Only 4 TF genes were differentially expressed in the transition of stipes of stage I buttons into stipe skin of stage II buttons and from stage I caps to stage II cap tissues (Figure 3.1B).

Expression of the *A. bisporus* orthologues of the blue light sensor gene wc-1 and the TF genes wc-2, hom2, fst4, *c2h2*, fst3, gat1 and hom1 of *S. commune* [16] was analyzed. To this end, expression levels at the different stages of development were compared with mycelium in the casing layer. Transcript levels of wc-2 and *c2h2* increased > 2-fold in initials compared to casing mycelium, while hom1 levels decreased > 2-fold (Table 3.2; Table S2 in the Supplementary Material). Expression of wc-1 and wc-2 was in general higher in aerial structures when compared to the casing mycelium, in stipes when compared to caps, and in outer tissues when compared to inner tissues of the aerial structures. Genes hom2 and fst4 were ≥ 2-fold upregulated when initials had developed in stage I buttons. Like wc-1 and wc-2, they were more highly expressed in stipes when compared to caps but in this case there was no difference between outer and inner tissues of the stage II buttons and young fruiting bodies. Gene *c2h2* showed high expression at different stages of fruiting body development. Expression levels ≥ 4-fold were observed in initials, caps of stage I buttons, gill tissue of stage II buttons, and veil tissue of young fruiting bodies. Expression of *c2h2* was reduced ≥ 2-fold when compared to casing mycelium in stipe and cap skin and in inner cap tissue of young fruiting bodies. Increased (≥ 2-fold) levels of fst3 were only observed in stipes of stage I buttons and in stipe skin and tissue of stage II buttons. Gene hom1, and in particular gat1 was in general downregulated when compared to casing mycelium.

Table 3.2: *Fold changes in expression during A. bisporus development of TF orthologues involved in mushroom formation in S. commune.*

	wc-1	*wc-2*	*hom2*	*fst4*	*c2h2*	*fst3*	*gat1*	*hom1*
Casing mycelium	1	1	1	1	1	1	1	1
Initials	1.4	6	1.6	1.9	4.4	1.1	0.9	0.4
Stage I buttons								
Stipe	2.3	3.4	5.6	2.3	3.1	2	0.4	0.6
Cap	1.1	2.8	1.4	2.1	8.5	0.9	0.4	0.3
Stage II buttons								
Stipe center	1.6	5.5	6	2.1	3.2	1.3	0.2	0.7
Stipe tissue	3.4	11.2	6	2.1	1.2	2.2	0.2	0.9
Stipe skin	2.7	8.2	4.7	2	1.6	2	0.4	0.9
Cap skin	2.1	3.2	0.6	1.3	2.1	0.9	0.4	0.6
Cap tissue	1.3	1.5	1.1	1.6	3.8	0.9	0.3	0.5
Gill tissue	0.8	0.8	1.7	1.9	7.6	0.8	0.3	0.4
Young fruiting bodies								
Stipe center	3.1	7.6	8.3	2.1	1.2	1.5	0.5	1.4
Stipe shell	4.6	18.8	8	2	1.6	1.9	0.3	1.4
Stipe skin	4.1	24	4.6	1.7	0.3	1.9	0..3	1.3
Cap skin	3.7	8.2	0.5	0.9	0.4	1	0.4	1.1
Cap tissue	2.8	2.3	1	0.8	0.1	1	0.3	0.8
Gill tissue	3.8	9.1	1.4	1.3	3	1.4	0.4	1.1
Veil	2.3	3.3	1.7	1.8	6.7	1.4	0.4	0.7

Table 3.3: *Total weight and number of mushrooms per box of A. bisporus strains A15, AT273-1 and AT273-5 after 2 flushes (n=10).*

	Weight (g)		Number of mushrooms	
Strain	Average	Standard deviation	Average	Standard deviation
A15	4197.6	200.1	600.1	114.6
AT273-1	4209.8	439.2	573.1	149.1
AT273-5	4324.4	235.5	634.7	90.9

OVER EXPRESSION OF *c2h2* OF *A. bisporus* RESULTS IN FASTER PRODUCTION OF MUSHROOMS

Gene *c2h2* (protein ID 230069) of *A. bisporus* shares 79% identity with its homologue of *S. commune* (protein ID 1194000; http://genome.jgi.doe.gov/Schco3). Expression construct pKS273 (see Material and methods) encompassing *A. bisporus* gene *c2h2* was introduced into *A. bisporus* A15 using A. tumefaciens mediated transformation. This resulted in 10 transformants, 2 of which were picked for further analysis. qPCR showed a 30- and a 2.5-fold increase in *c2h2* expression in *A. bisporus* AT273-1 and AT273-5, respectively, when grown on MEA. Growth of these strains on malt extract medium was similar to the parental strain.

Mushroom production of *A. bisporus* AT273-5 and AT273-1 was assessed in a semi-commercial setting (see Material and methods). The first flush started 9 days after venting and progressed until day 14. The second flush took place between day 19 and day 22 (Figure **??**). Biomass of mushrooms harvested at day 9-11 and at day 19-20 was higher for *A. bisporus* AT273-1 when compared to A15 (p = <0.01, 0.01, 0.02, 0.04 and 0.00, respectively) (Figure **??**). *A. bisporus* AT273-5 showed higher harvested mushroom biomass at day 11, 19, and 20 when compared to A15 (p = 0.04, <0.01 and <0.01, respectively). The latter strain produced more biomass at day 13 and 14 compared to *A. bisporus* AT273-1 (p = 0.01 and <0.01, respectively) and *A. bisporus* AT273-5 (p = 0.01 and <0.01, respectively) and more biomass compared to *A. bisporus* AT273-1 on day 22 (p = 0.02). Total production of mushrooms was similar for the 3 strains (Table 3.3). A higher number of A15 mushrooms was harvested at day 13 when compared to *A. bisporus* AT273-5 (p = 0.01), day 14 compared to both transformants (p = <0.01 for both), and day 22 compared to *A. bisporus* AT273-1 (p-values 0.01). A higher number of *A. bisporus* AT273-1 mushrooms were harvested at day 9 and 20 (p = 0.04 and 0.01, respectively) and of *A. bisporus* AT273-5 mushrooms at day 19 and 20 (p = 0.01 for both) when compared to A15 (Figure **??**). Together, these data show that over-expression of *c2h2* accelerates development of mushrooms.

Harvested mushrooms were classified based on size 40 (cap ≤ 40 mm) and size 60 (cap between 40-60 mm) (Figure 3.3). During the first flush, A15 produced more size 40 mushrooms (57%) (p = 0.01), while the *c2h2* over-expressing strains produced more size 60 mushrooms (56% and 55%, respectively) (p = 0.01 for both). The ratio between cap and stem dimensions were similar for all strains. A relative dry weight of 8% was found for the mushrooms of the 3 strains at day 12 and 13 (Figure 3.4). All strains produced more size 40 mushrooms in the second flush (64% for A15 versus 73% and 71%

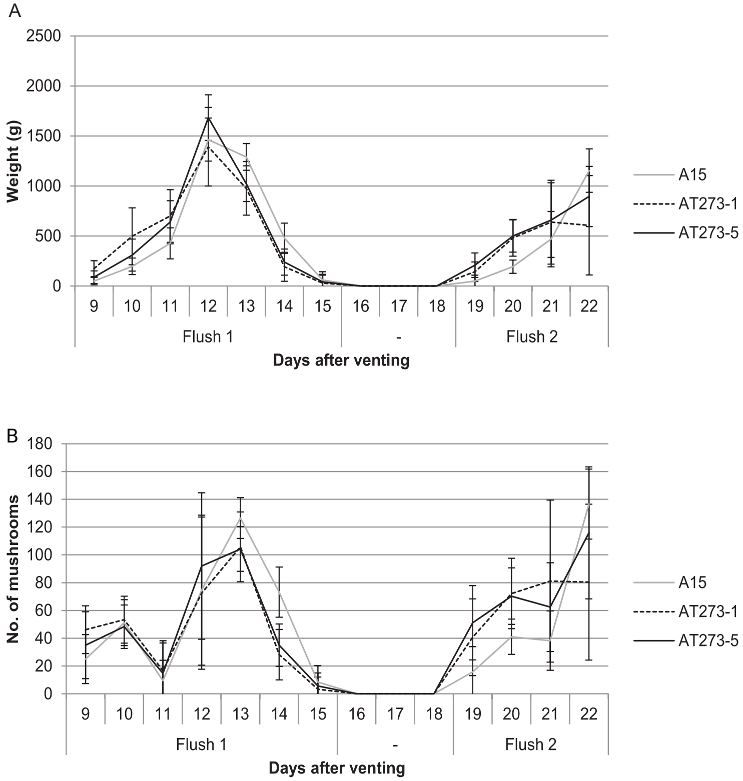


Figure 3.2: *Average biomass (a) and number of mushrooms (b) of A15, AT273-1 and AT273-5 mushrooms per box (n=10) picked during a 22-day period after venting (t = 0). Bars represent standard deviation.*

for AT273-1 and AT273-5) (p = <0.01 for all). Relative dry weight at day 21 amounted between 6.2% and 6.7% for the three strains (Figure 3.4). Together, these data show that over-expression of *c2h2* promotes size in the first flush.

Mushroom formation was monitored by analyzing photos taken in 24 h intervals. Cap expansion was similar for the 3 strains (Figure 3.5B). The number of buttons emerging from the casing was not significantly different between the three strains but there was a trend that the *c2h2* over-expression strains showed accelerated button emergence (Figure 3.5A).

DISCUSSION

Formation of mushrooms is a highly complex developmental process [25]. After a submerged mycelium has been formed, hyphae escape the substrate to grow into the air. These hyphae form aggregates with a diameter < 1 mm. They result from a single hypha that branches intensely or arise from branches of neighboring aerial hyphae that grow towards and alongside each other. The dark-grown aggregates of *C. cinerea* can develop

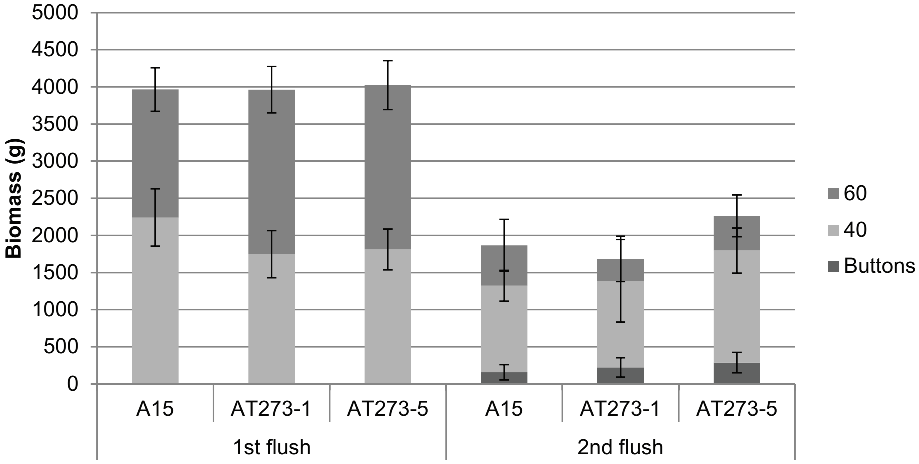


Figure 3.3: *Size composition of A15, AT273-1 and AT273-5 mushrooms harvested during two flushes (n=100). Bars represent standard deviation.*

into different structures depending on light conditions. Continuation of growth in the dark results in formation of sclerotia, while a 12 h day-night cycle induces development of initials. Initials, or primordia, are the first fruiting body specific structures and can be selectively stained with Janus green [26].

The switch from aggregate to primordia can be considered key in development since it determines the structure to develop into a sexual reproductive structure. The fact that *c2h2* of *S. commune* is involved in the switch from aggregates to primordia [11] makes it a gene of high interest for fruiting body development. Here the *c2h2* orthologue of *A. bisporus* was over-expressed resulting in accelerated mushroom formation.

Expression of *c2h2* in different developmental stages was compared to that in casing mycelium. Gene *c2h2* was ≥ 4-fold upregulated in initials of *A. bisporus*. Upregulation was also found in stage I and stage II buttons, in particular in cap tissue. Expression of *c2h2* was reduced or ≤ 2-fold upregulated in the tissues of young fruiting bodies with the exception of veil and gill tissue. These expression data indicate that *c2h2* functions early in fruiting body development, while it also seems to have a role in selective tissues of young mushrooms.

Gene *c2h2* of *A. bisporus* was over-expressed in the commercial *A. bisporus* strain A15. Two transformants were selected for phenotypic analysis. These strains, called AT273-1 and AT273-5, displayed 30-fold and 2.5-fold increased *c2h2* expression, respectively. Phenotypes of these strains were similar, indicating that a few fold increased expression of *c2h2* is sufficient to obtain a full effect of over-expression. Morphology, cap expansion rate, and total number and biomass of harvested mushrooms were not affected by over-expression of *c2h2*. However, formation of mushrooms with a cap size ≥ 4 cm was accelerated. Biomass of harvested mushrooms was increased on day 9 to11 (first flush) and 19 to 20 (second flush) in *A. bisporus* AT273-1 and day 11, 19 and 20 in *A. bisporus* AT273-5 when compared to A15. On the other hand, A15 produced more biomass

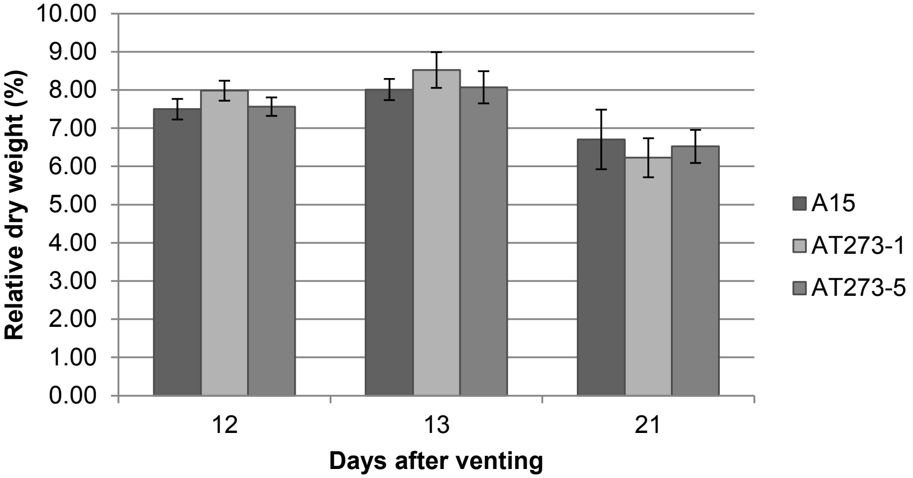


Figure 3.4: *Relative dry weight of mushrooms at days 12, 13 and 21 after vent-off for A. bisporus strains A15, AT273-1 and AT273-5 (n=10). Bars represent standard deviation.*

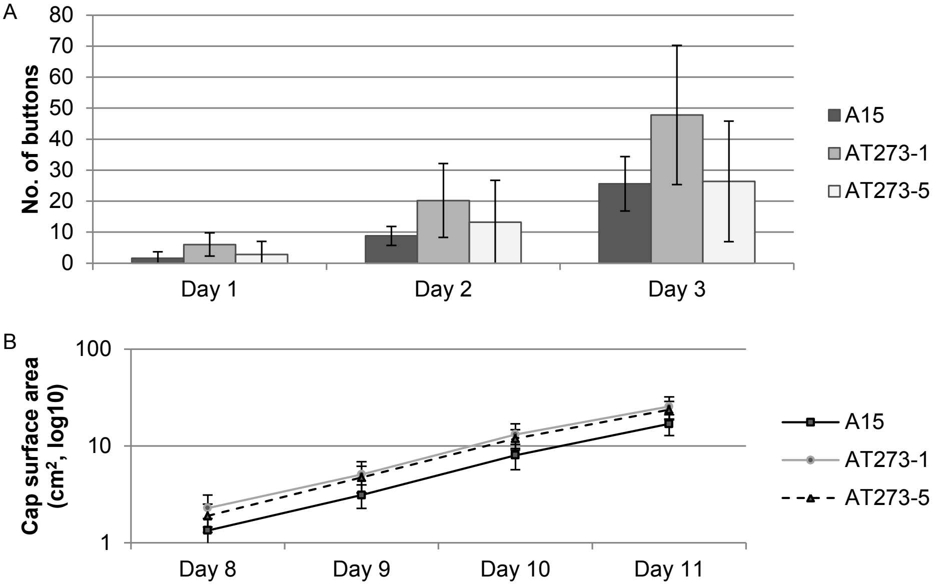


Figure 3.5: *Number of A15, AT273-1 and AT273-5 buttons that had emerged from the casing soil 1, 2 and 3 days after venting (A) and expansion of mushroom caps of the three strains in time (B) (n=10). Bars represent standard deviation.*

on day 13, 14 and 22. The number of harvested mushrooms also indicated accelerated growth of the *c2h2* over-expressors. The fact that expansion rate of mushrooms was similar between the transformants and A15 implies that accelerated mushroom formation is caused at the level of outgrowth of initials. This is supported by a trend that the transformants had formed more initials when compared to A15 1, 2, and 3 days after venting.

It is difficult to compare our results with other whole genome expression studies of mushroom development [10–12, 16–18]. In this work RNA from a fertile casing mycelium was used as a reference for differential expression, while Plaza et. al. [17] used fertile vegetative mycelium from complete medium. [10–12] compared whole cultures of a sterile monokaryotic vegetative mycelium with whole cultures of the fertile dikaryon at different developmental stages. In contrast, we here used pure developmental structures and tissues. Therefore, we have only focused on expression of genes known to play a role in fruiting body development in *S. commune*. Expression of *c2h2* in *S. commune* is highest in primordia and mature fruiting bodies [11], which is in agreement with the findings in *A. bisporus*. The genes encoding the blue light sensing complex components Wc-1 and Wc-2 are also most highly expressed in primordia and fruiting bodies of *S. commune*. This agrees with the finding that the *A. bisporus* homologues were more highly expressed in aerial structures when compared to casing mycelium, in stipe tissue when compared to cap tissue, and in outer tissues of the aerial structures when compared to inner tissues. *A. bisporus* does not require blue light to produce mushrooms. Yet, blue light sensing is also required to induce UV light-related DNA damage repair (e.g. photolyase) and in conversion of toxic porphyrin intermediates in heme (ferrochelatase)[12]. Expression of hom2 of *S. commune* does not change during development until the stage of mature fruiting bodies when expression drops. Expression of fst3 and fst4 is considered constitutive in *S. commune*. Gene hom2 was ≥ 2-fold over-expressed when initials had developed into stage I buttons. Like wc-1 and wc-2, expression of hom2, fst4, and fst3 were more highly expressed in stipes when compared to caps but only small differences were observed between outer and inner tissues of the stage II buttons and young fruiting bodies. Genes gat1 and hom1 of *S. commune* are most highly expressed in late stages of mushroom development, although their upregulation is modest. These genes have a different expression profile in *A. bisporus*. They were generally downregulated in the developmental structures when compared to casing mycelium. This effect was most prominent for gat1. It may thus be that their role in *S. commune* and *A. bisporus* is different. Recently, expression profiles of wc-2, hom2, fst4, *c2h2*, fst3, hom1 and gat1 were determined in stipe and cap during fruiting body development in *C. cinerea* [18]. Expression of wc-2 increased during initial development and was higher in the stipe when compared to cap. Genes fst4 and fst3 were also more highly expressed in the stipe. This is similar to the *A. bisporus* expression profiles presented in this study. Expression of hom2 remained constant during the early development, but in contrast to *A. bisporus*, was higher in the cap during later stages. Transcript levels of gat1 slightly increased during development in *C. cinerea*, while they diminished in *A. bisporus* at this stage. Expression of *c2h2* was higher in the cap compared to the stipe tissues early in development while this was reversed later in development, a situation similar to *A. bisporus*. Together, these data support the view that mushroom development in the Basidiomycota follows a core regulatory program with species specific variations that may explain differences in morphology and sensitivity to environmental signals.

BIBLIOGRAPHY

[1] H. Bels-Konig. Experiments with casing soils, water supply and climate. *Mushroom Science* 1:78–84 (1950).

[2] P. B. Flegg. The casing layer in the cultivation of the mushroom (psalliota hortensis). *Journal of Soil Science* 7(1):168–176 (1956).

[3] P. P. Kalberer. Water potentials of casing and substrate and osmotic potentials of fruit bodies of agaricus bisporus. *Scientia Horticulturae* 32(3-4):175–182 (1987).

[4] R. Noble, A. Dobrovin-Pennington, P. J. Hobbs, J. Pederby, A. Rodger. Volatile c8 compounds and pseudomonads influence primordium formation of agaricus bisporus. *Mycologia* 101(5):583–591 (2009).

[5] D. C. Eastwood, et al. Environmental regulation of reproductive phase change in agaricus bisporus by 1-octen-3-ol, temperature and co2. *Fungal Genetics and Biology* 55:54–66 (2013).

[6] U. Kües, M. Navarro-González. How do agaricomycetes shape their fruiting bodies? 1. morphological aspects of development. *Fungal Biology Reviews* 29(2):63–97 (2015).

[7] M. H. Umar, L. J. Van Griensven. Morphological studies on the life span, developmental stages, senescence and death of fruit bodies of agaricus bisporus. *Mycological Research* 101(12):1409–1422 (1997).

[8] R. Noble, et al. Primordia initiation of mushroom (agaricus bisporus) strains on axenic casing materials. *Mycologia* 95(4):620 (2003).

[9] G. Straatsma, A. S. Sonnenberg, L. J. van Griensven. Development and growth of fruit bodies and crops of the button mushroom, agaricus bisporus. *Fungal Biology* 117(10):697–707 (2013).

[10] R. a. Ohm, et al. Genome sequence of the model mushroom schizophyllum commune. *Nature biotechnology* 28(9):957–63 (2010).

[11] R. a. Ohm, J. F. de Jong, C. de Bekker, H. a. B. Wösten, L. G. Lugones. Transcription factor genes of schizophyllum commune involved in regulation of mushroom formation. *Molecular Microbiology* 81(6):1433–1445 (2011).

[12] R. a. Ohm, D. Aerts, H. a. B. Wösten, L. G. Lugones. The blue light receptor complex wc-1/2 of schizophyllum commune is involved in mushroom formation and protection against phototoxicity. *Environmental Microbiology* 15(3):943–955 (2013).

[13] J. H. Perkins, S. A. Gordon. Morphogenesis in schizophyllum commune. ii. effects of monochromatic light. *PLANT PHYSIOLOGY* 44(12):1712–1716 (1969).

[14] D. J. Niederpruem. Role of carbon dioxide in the control of fruiting of schizophyllum commune. *Journal of bacteriology* 85:1300–8 (1963).

[15] M. Raudaskoski, H. Viitanen. Effect of aeration and light on fruit body induction in schizophyllum commune. *Transactions of the British Mycological Society* 78(1):89–96 (1982).

[16] E. Morin, et al. Genome sequence of the button mushroom agaricus bisporus reveals mechanisms governing adaptation to a humic-rich ecological niche. *Proceedings of the National Academy of Sciences* 109(43):17501–17506 (2012).

[17] D. Plaza, C.-W. Lin, N. S. van der Velden, M. Aebi, M. Künzler. Comparative transcriptomics of the model mushroom coprinopsis cinerea reveals tissue-specific armories and a conserved circuitry for sexual development. *BMC Genomics* 15(1):492 (2014).

[18] H. Muraguchi, et al. Strand-specific rna-seq analyses of fruiting body development in coprinopsis cinerea. *PLOS ONE* 10(10):e0141586 (2015).

[19] C. A. Schneider, W. S. Rasband, K. W. Eliceiri. Nih image to imagej: 25 years of image analysis. *Nature Methods* 9(7):671–675 (2012).

[20] X. Chen, M. Stone, C. Schlagnhaufer, C. P. Romaine. A fruiting body tissue method for efficient agrobacterium-mediated transformation of agaricus bisporus. *Applied and Environmental Microbiology* 66(10):4510–4513 (2000).

[21] A. M. Bolger, M. Lohse, B. Usadel. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* 30(15):2114–2120 (2014).

[22] A. Dobin, et al. Star: ultrafast universal rna-seq aligner. *Bioinformatics (Oxford, England)* 29(1):15–21 (2013).

[23] C. Trapnell, et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology* 31(1):46–53 (2012).

[24] C. Trapnell, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* 7(3):562–578 (2012).

[25] U. Kues. Life history and developmental processes in the basidiomycete coprinus cinereus. *Microbiology and Molecular Biology Reviews* 64(2):316–353 (2000).

[26] C. Sánchez, D. Moore. Conventional histological stains selectively stain fruit body initials of basidiomycetes. *Mycological Research* 103(3):315–318 (1999).

Chapter 4

Schizophyllum commune HAS AN EXTENSIVE AND FUNCTIONAL ALTERNATIVE SPLICING REPERTOIRE

Thies Gehrmann
Jordi F. Pelkmans
Luis G. Lugones
Han A. B. Wösten
Thomas Abeel
Marcel J. T. Reinders

This chapter is published as: *Scientific Reports* September 2016, Volume 6, Article number: 33640 (doi: 10.1038/srep33640)
Supplementary material is available online at: `http://go.nature.com/2mU6TFi`

ABSTRACT

Recent genome-wide studies have demonstrated that fungi possess the machinery to alternatively splice pre-mRNA. However, there has not been a systematic categorization of the functional impact of alternative splicing in a fungus. We investigate alternative splicing and its functional consequences in the model mushroom forming fungus Schizophyllum commune.

Alternative splicing was demonstrated for 2,285 out of 12,988 expressed genes, resulting in 20% additional transcripts. Intron retentions were the most common alternative splicing events, accounting for 33% of all splicing events, and 43% of the events in coding regions. On the other hand, exon skipping events were rare in coding regions (1%) but enriched in UTRs where they accounted for 57% of the events.

Specific functional groups, including transcription factors, contained alternatively spliced genes. Alternatively spliced transcripts were regulated differently throughout development in 19% of the 2,285 alternatively spliced genes. Notably, 69% of alternatively spliced genes have predicted alternative functionality by loss or gain of functional domains, or by acquiring alternative subcellular locations. S. commune exhibits more alternative splicing than any other studied fungus. Taken together, alternative splicing increases the complexity of the S. commune proteome considerably and provides it with a rich repertoire of alternative functionality that is exploited dynamically.

INTRODUCTION

Alternative splicing is an important regulatory mechanism, which in many eukaryotes provides additional complexity to the proteome required to sustain cell functions [1]. In humans, the overwhelming majority of genes are alternatively spliced, and mutations in alternative transcripts can give rise to diseases such as cancer [2]. Alternative splicing events have also been shown to contribute to the evolution of new species [3].

Splicing results in removal of introns in pre-mRNAs, and alternative splicing gives rise to different mRNA variants of a gene. All resulting mRNAs consist of a coding region flanked by UnTranslated Regions (UTRs). We distinguish primitive events, i.e. basic splice operations which only affect a single transcript, from composite events that relate to observed primitive events over multiple transcripts from the same gene (Figure 4.1a-b). Primite events include intron retentions (IR, when an intron is not spliced out) exon skipping (ES, when exons are spliced out), alternative 5' splicing sites (A5SS, when an exon starts at alternative locations), and 3' splicing sites (A3SS, when an exon ends at alternative locations). We examine three composite events, mutually exclusive exons (MUT, when two or more transcripts skip different exons such that the two exons never occur together in a single transcript), and multiple alternative 5- and 3' splicing sites (MA5SS and MA3SS, when alternative transcripts have more than two alternative 5' or 3' splicing sites on the same exon). Composite events represent regions of especially high sequence heterogeneity between transcripts. Alternative splicing events occur not only in coding regions but also in UTRs, resulting in transcripts with alternative coding regions. Consequently, alternative splicing increases the complexity of the proteome which may also result in functional changes of the proteome when, for example, alternative splicing results in gain or loss of protein domains and functional elements.

Alternative splicing events may be triggered by changes in environmental conditions [4] or are part of developmental pathways [5]. On the other hand, alternative splicing may also be an error of the splicing machinery and not necessarily lead to alternative functionality. For instance, *RPL30* of *Saccharomyces cerevisiae* produces a transcript that fails to undergo splicing and later decays [6].

Alternative splicing has been shown to occur in fungi. The first reported example of alternative splicing was in the *glaA* gene of *Aspergillus niger*, resulting in two isoforms of the glucoamylase enzyme [7]. Since then, other alternative splicing events have been described. For instance, the primary transcript of *ZAS1+* of *Schizosaccharomyces pombe* contains two zinc finger domains, while its secondary transcript contains three of these domains [8]. Genome-wide comparative studies have shown that alternative splicing is common in most fungi, but that many yeasts exhibit few alternative splicing events [9, 10]. Developmentally more complex fungi appear to have more alternative splicing (see Figure 4.1e, where we compared the phylogeny of fungal species with reported richness of their alternative splicing). Genome-wide studies also showed that development impacts alternative splicing in *Fusarium graminearum* [11], while growth medium composition impacts splicing in *Trichoderma longibrachiatum* [12]. In all studies, intron retentions were the most common primitive splicing events.

Alternative splicing at a genome-wide scale can be detected from expressed sequence tags [9, 10]. With the advent of Next Generation Sequencing, detecting alternative splicing in fungi is mostly done by splice junction analysis. By comparing splice junctions that appear as gaps in the alignment of short reads to a reference genome, primary splicing events can be detected [12]. To reconstruct the structure of alternative transcripts, probabilistic modeling methods are used, which are based on primary event co-occurrence in single reads [13, 14]. Along with splice junction analysis, this approach has also been separately used to study alternative splicing in fungi [11]. However, the high gene density and polyscystronic transcripts in many fungal genomes[15], make this approach inapplicable. Gene UTRs overlapping with transcripts of neighboring genes cause the transcript reconstruction methods to erroneously fuse transcripts from different genes together (for more details see supplementary Note S2). This shortcoming has prevented a systematic characterization of the alternative splicing repertoire for fungi with gene-dense genomes.

Despite the existing work to identify alternative splicing in fungi, a systematic analysis of the functional impact of alternative splicing is missing. In this work, we show the potential functional impact of alternative splicing in the mushroom forming fungus *S. commune*. We designed a region-restricted probabilistic modeling method to detect alternative splicing for the gene-dense genome of *S. commune*. We detected alternative splice variants from RNA-seq data at different developmental stages of *S. commune*. We further analyzed the functional impact of alternative splicing and show its relationships with transcriptional regulation, subcellular localization, and alternative use of functional domains.

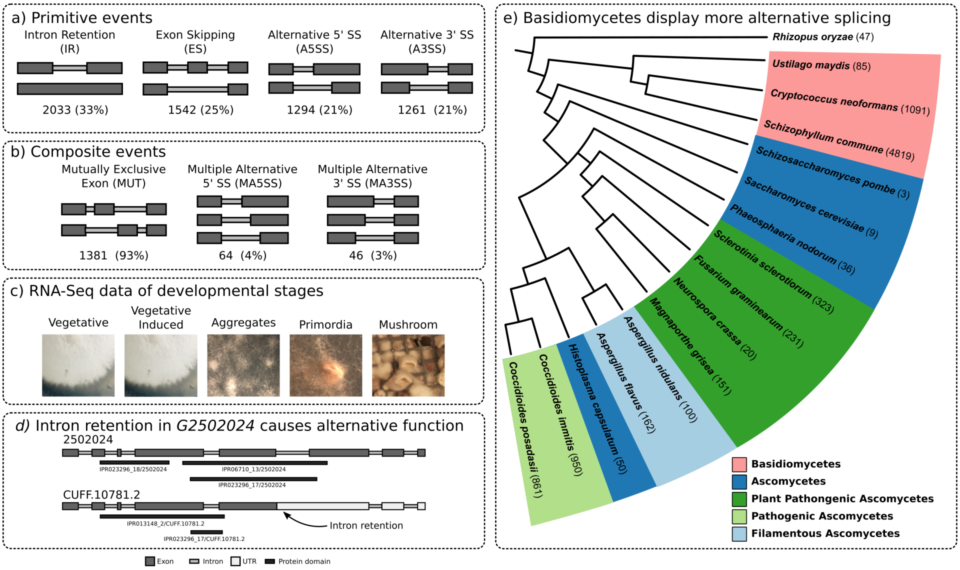

a) Primitive events
Intron Retention (IR)
Exon Skipping (ES)
Alternative 5' SS (A5SS)
Alternative 3' SS (A3SS)
2033 (33%)
1542 (25%)
1294 (21%)
1261 (21%)
b) Composite events
Mutually Exclusive Exon (MUT)
Multiple Alternative 5' SS (MA5SS)
Multiple Alternative 3' SS (MA3SS)
1381 (93%)
64 (4%)
46 (3%)
c) RNA-Seq data of developmental stages
Vegetative
Vegetative Induced
Aggregates
Primordia
Mushroom
d) Intron retention in G2502024 causes alternative function
2502024
IPR023296_18/2502024
IPR06710_13/2502024
IPR023296_17/2502024
CUFF.10781.2
IPR013148_2/CUFF.10781.2
IPR023296_17/CUFF.10781.2
Intron retention
Exon
Intron
UTR
Protein domain
e) Basidiomycetes display more alternative splicing
Rhizopus oryzae (47)
Ustilago maydis (85)
Cryptococcus neoformans (1091)
Schizophyllum commune (4819)
Schizosaccharomyces pombe (3)
Saccharomyces cerevisiae (9)
Phaeosphaeria nodorum (36)
Sclerotinia sclerotiorum (323)
Fusarium graminearum (231)
Neurospora crassa (20)
Magnaporthe grisea (151)
Aspergillus nidulans (100)
Aspergillus flavus (162)
Histoplasma capsulatum (50)
Coccidioides immitis (950)
Coccidioides posadasii (861)
Basidiomycetes
Ascomycetes
Plant Pathongenic Ascomycetes
Pathogenic Ascomycetes
Filamentous Ascomycetes

Figure 4.1: *Alternative splicing in S. commune.* ***a)*** *Primitive alternative splicing events that affect a single transcript. Two alternative transcripts are shown above each other. Thick dark gray bars represent exons, and skinny light gray bars indicate introns. The frequencies of each event in S. commune are provided beneath the figure.* ***b)*** *Composite splicing events emerge from splicing events across multiple alternative transcripts.* ***c)*** *Alternative splicing was analysed using RNA-seq data from five developmental stages of the mushroom-forming S. commune fungus.* ***d)*** *An example of alternative splicing in gene G2502024. An intron retention (indicated by an arrow) introduces a premature stop codon. Due to this, the alternative transcript CUFF.10781.2 loses two protein domain annotations.* ***e)*** *The number of alternatively spliced transcripts identified in various species (numbers from [10]). Basidiomycetes are indicated in red, while different groups ascomycetes are indicated in blue and green. More phenotypically complex fungi have more alternatively spliced genes (see Supplementary Note S1).*

RESULTS

ALTERNATIVE SPLICING IS A COMMON MECHANISM IN *S. commune*

Sexual development of *S. commune* proceeds via different stages. Dikaryotic colonies resulting from mating of two compatible partners grow symmetrical in the dark and at high CO2 conditions. Asymmetrical growth and the formation of aggregates is induced when the colony is exposed to light and ambient CO2. The aggregates further develop into primordia and mature mushrooms. mRNA of five different stages of development of *S. commune* (Figure 4.1c) was sequenced to study the incidence and functional impact of alternative splicing. In addition, RNA was used from deletion strains that are affected at different stages of mushroom formation [16–18] (see Materials and Methods).

Based on RNA sequencing of all samples, transcripts were predicted using our region-restricted probabilistic modeling method (RRPM, see Materials and Methods). Figure 4.2 gives an overview of features of the found transcripts. 3,331 genes that were not expressed or whose exon boundaries were not supported by the RNA-seq data were excluded from the 16,319 genes in the *S. commune* v3.0 reference genome. 15,522 transcripts were predicted for the remaining 12,988 genes. 10,703 genes (82%) had only

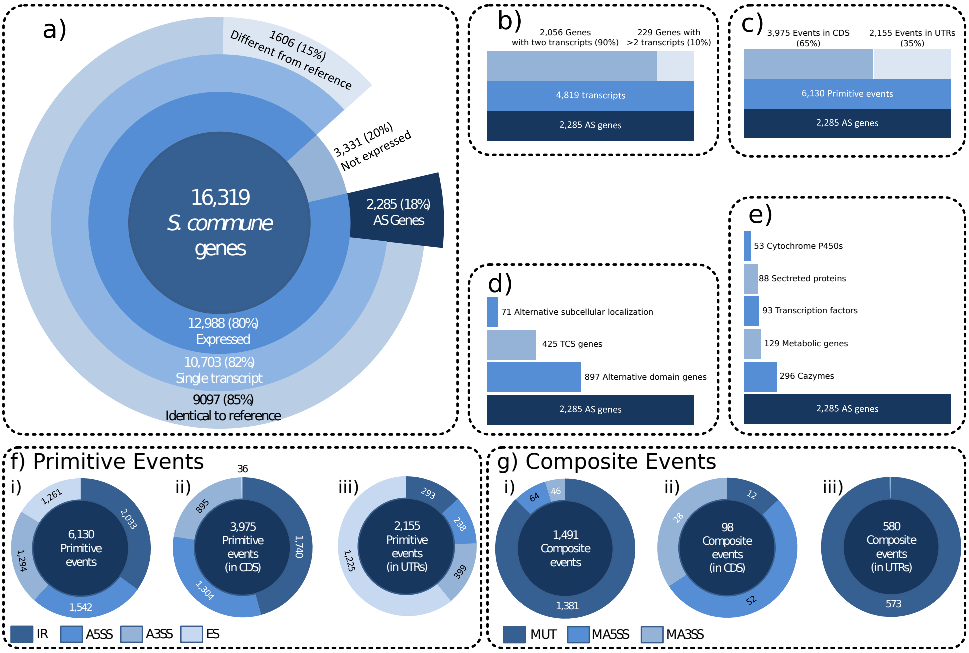


Figure 4.2: *Statistics of alternative splicing.* ***a)*** *The extend of alternative splicing revealed by analysing the RNA-seq data over different developmental stages of S. commune.* ***b)*** *The distribution of transcripts per gene.* ***c)*** *The number of alternative splicing events within and outside of coding regions.* ***d)*** *The number of genes that are functionally impacted by the alternative splicing on domains, subcellular localization and development.* ***e)*** *The amount of alternative spliced genes for a number of functional groups.* ***f)*** *The proportion of primitive events i) genome wide, ii) within coding regions, and iii) in UTRs (supplemental Note S10).* ***g)*** *The proportion of composite events i) genome wide, ii) within coding regions, and iii) in UTRs (supplemental Note S10).*

one transcript, 85% of them being identical to the original annotation (Figure 4.2a). A total number of 4,819 transcripts were found for the other 2,285 genes (see Figure 4.2b). Most of these genes (90%) had two splice variants (Figure 4.2b and supplementary Note S3) and 72% of the alternative transcripts had only one alternative splicing event (supplementary Note S4). The alternatively spliced genes are located over the entire genome (supplementary Note S5). Together, 6,130 primitive splicing events resulted in the 4,819 transcripts of the 2,285 genes of which 65% (3,075) occur in the coding region of the gene, and 35% (2,155 genes) in the UTR (Figure 4.2c).

Intron retention (33%) and alternative 5' splicing site (25%) were the most common primitive events (Figure 4.2f.i). Alternative 3' splicing sites and exon skipping events occurred at similar levels (21%). Mutually exclusive exons accounted for 93% of the 1,491 composite events, while multiple alternative 5' and 3' splicing sites were relatively rare at 4% and 3%, respectively (Figure 4.2g.i). These mutually exclusive exon events occurred in 171 genes representing 8% of the alternative spliced genes. Most alternative splicing events are not reading frame neutral (Supplementary Note S6).

UTRs of genes might stretch so far that they overlap with a neighboring gene. As a consequence, paired-end reads may lie within the boundary of another gene. To investigate whether such reads have a strong effect on event counts, we selected only those genes that are flanked on both sides by intergenic regions with zero coverage (i.e. no overlap). A total of 306 alternatively spliced genes were found in the selected set of 2,516 genes. We observed the same distribution of event counts in this set as compared to all genes (Supplementary Note S7). From this, we conclude that our analysis is not confounded by reads from neighboring genes.

ALTERNATIVE SPLICING IS NOT LIMITED TO CODING REGIONS

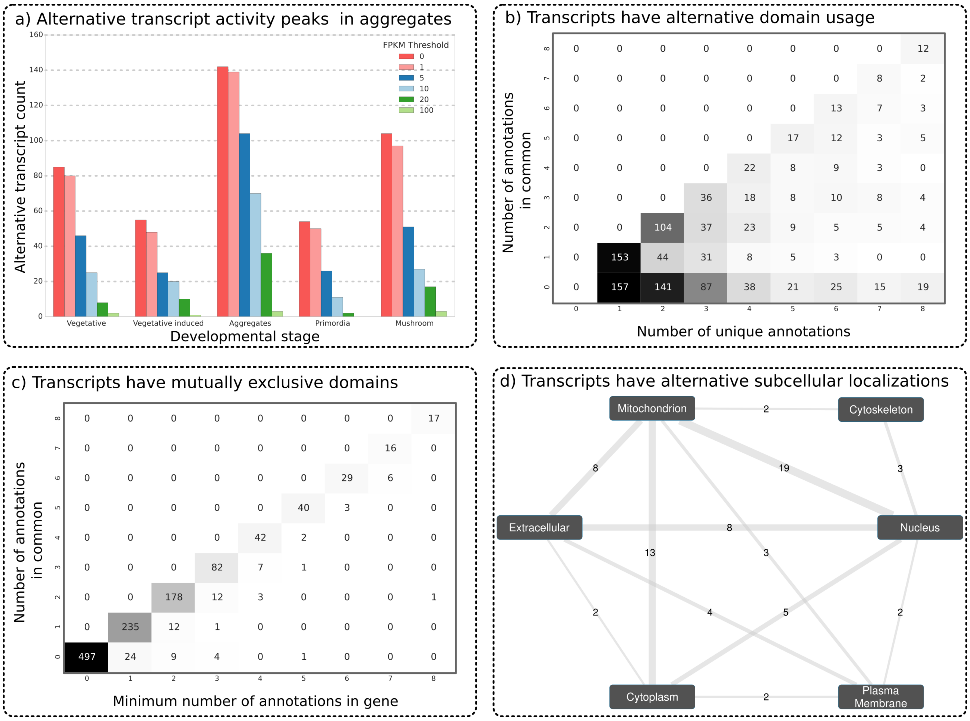
Alternative splicing occurs not only in the translated regions of transcripts, but also in the UTRs of transcripts. RRPM is able to detect alternative splicing events that lie within the original gene boundaries of known genes. RRPM can detect alternative splicing events after premature stop codons introduced by alternative splicing events that still lie within the original gene boundaries (Supplementary Note S8). RRPM cannot investigate alternative splicing events in the UTRs extending outside of these original gene boundaries, and which are generally unknown. In the unknown UTR regions, we can use splice junction analysis (supplementary Note S9).

A total of 3,975 (65%) alternative splicing events were found within coding transcript regions (Figure 4.2c). Of these, 44% are intron retentions, and only 1% were exon skipping events (Figure 4.2f.ii). Due to splicing, some regions at the 5' and 3' ends of the original coding regions may become non-coding UTRs, and this is where the remaining 2,155 (35%) events are located. Of these, 1,225 (57%) were classified as exon skipping events and 293 (14%) as intron retentions (Figure 4.2f.iii and supplementary Note S10). Mutually exclusive exons occur primarily in the untranslated regions of transcripts (Figure 4.2g.ii and Figure 4.2g.iii). They account for 99% of all composite events in the UTR regions, but for only 15% of those in the coding regions. Multiple alternative 5' and 3' splicing sites, which are rare, occur mostly in coding regions.

Splice junction analysis enables a semi-quantitative analysis of alternative splicing in coding and non-coding regions (see Methods). We observed 186,221 splice junctions across all our RNA-seq samples. 29% of these occur only once, and are likely errors of the splicing machinery or sequencing. In the remaining 71%, we find 15,054 primary alternative splicing events. 62% lie within the coding regions of defined genes, and 38% are in the UTRs of these genes. We found 7,851 alternative (5' or 3') splicing sites and 1,527 exon skipping events inside coding regions, while we discovered 4,574 alternative splicing sites and 1,102 exon skipping events in UTRs. Alternative splicing is therefore clearly not limited to translated regions.

EXPRESSION OVER DEVELOPMENTAL STAGES IMPLICATES ALTERNATIVE SPLICING IN DEVELOPMENT

Time Course Switching (TCS) genes differentially express splice variants during development [5]. The primary transcript (defined as the transcript that has the highest average expression over development), is replaced in expression by one of its alternative transcripts at some point in time. *S. commune* was found to have 425 of such genes, representing 865 transcripts. The aggregate stage contained more expressed alternative transcripts than any other stage (Figure 4.3a). At this stage, the alternative transcripts of 142 genes were more abundant than their primary transcript. For example, the alter-



	0	1	2	3	4	5	6	7	8
8	0	0	0	0	0	0	0	0	12
7	0	0	0	0	0	0	0	8	2
6	0	0	0	0	0	0	13	7	3
5	0	0	0	0	0	17	12	3	5
4	0	0	0	0	22	8	9	3	0
3	0	0	0	36	18	8	10	8	4
2	0	0	104	37	23	9	5	5	4
1	0	153	44	31	8	5	3	0	0
0	0	157	141	87	38	21	25	15	19

	0	1	2	3	4	5	6	7	8
8	0	0	0	0	0	0	0	0	17
7	0	0	0	0	0	0	0	16	0
6	0	0	0	0	0	0	29	6	0
5	0	0	0	0	0	40	3	0	0
4	0	0	0	0	42	2	0	0	0
3	0	0	0	82	7	1	0	0	0
2	0	0	178	12	3	0	0	0	1
1	0	235	12	1	0	0	0	0	0
0	497	24	9	4	0	1	0	0	0

Figure 4.3: ***a)*** *Alternative splice variants of time course switch genes being more the abundant when compared to the annotated variant at different stages of activity of time course switch genes through development. We count the number of alternative transcripts which are active at each developmental stage, exceeding a particular expression threshold, and find that there are more active at the aggregates stage of development than at any other, regardless of expression threshold.* ***b)*** *Implications of alternative splicing on domain predictions. The number of unique annotations are given on the x-axis, and the number of annotations in common between all transcripts is given on the y-axis. If the numbers are not equal, that means that not all transcripts in a gene are annotated with the same domains. In the corresponding cells are given the number of genes with this combination of annotations. Most genes do not exhibit alternative functionality (see diagonal), but quite a number do (see below the diagonal).* ***c)*** *The smallest number of annotations of a transcript in a gene is given on the x-axis (e.g. a gene with two transcripts having 2 and 3 annotations would have a value of 2) and the number of annotations in common between all transcripts is given on the y-axis. On the diagonal are genes in which the number of annotations in common is limited by a transcript for which annotations are lost due to alternative splicing. Off the diagonal are genes that have enhanced their functional abilities through alternative splicing (see supplemental Note S11).* ***d)*** *Alternative splicing changes subcellular location of genes. Nodes represent a subcellular location. Edges represent genes that have two alternative splice variants, each encoding for a different subcellular location (connected nodes). Counts associated with an edge indicate the amount of genes having alternative transcripts that encode for the associated subcellular locations. For example, there are 8 genes for which the protein encoded by one transcript is secreted (extracellular) and the protein encoded by the other transcript is transported to the mitochondrion. See supplemental Note S12 for genes whose subcellular location do not change.*

native transcript (ID CUFF.11357.2) of the carbohydrate active enzyme (cazyme) gene *G2634198* (Figure 4.4a) entirely replaces the primary transcript (ID 2634198) during the aggregate stage, after which its abundance decreased to zero at the primordia stage, and in turn, replaced by the primary transcript. Gene *G2629174* (Figure 4.4b) which encodes a secreted protein also exhibits this behavior. The polypeptide encoded by the alternative transcript lacks a large part of its N-terminus due to intron retention.

GENES WITH ALTERNATIVE TRANSCRIPTS EXHIBIT ALTERNATIVE FUNCTIONALITY THROUGH PROTEIN DOMAINS

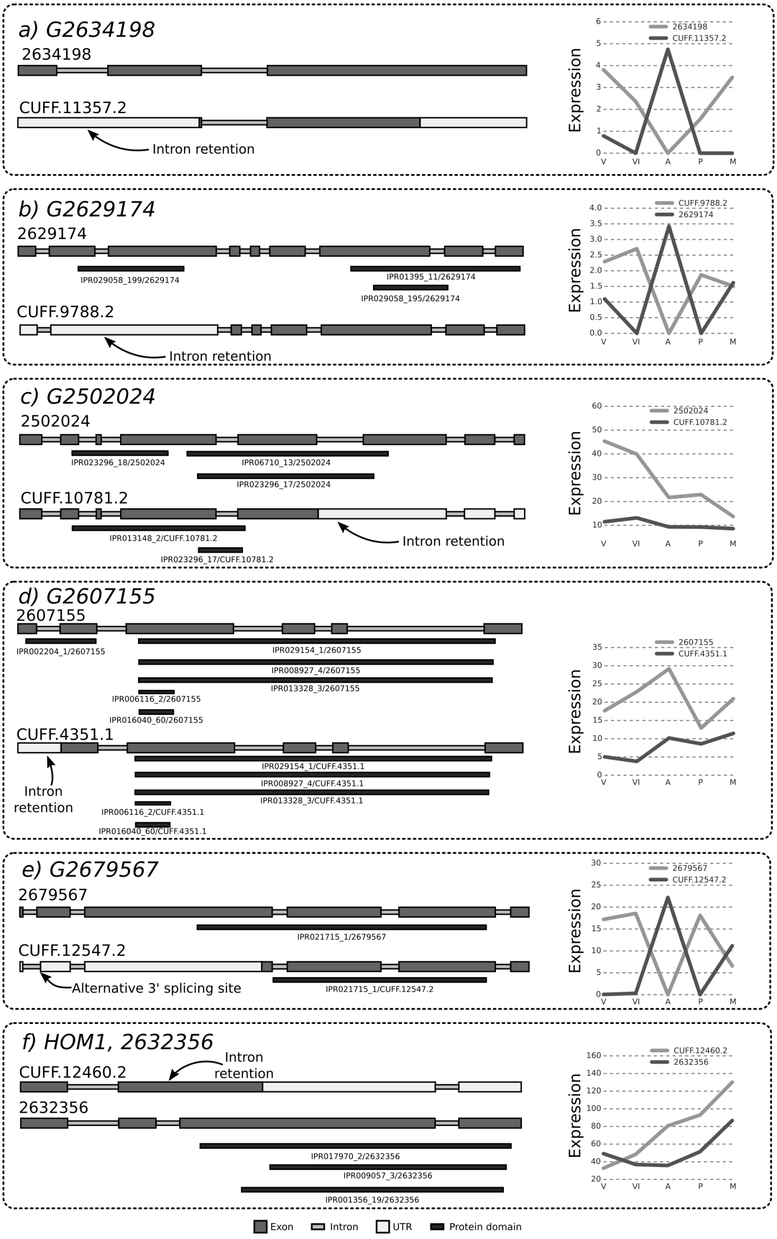
Alternative splicing may result in the gain or loss of functional domains. Out of the 2,285 genes with alternative transcripts, 1,257 (55%) have predicted functional domains. A functional domain is lost/gained in an alternative transcript in 897 (70%) genes (Figure 4.3b-c and supplementary Note S11). For 274 (12%) genes a domain is exchanged for another (Figure 4.4c). For example, gene *G2502024*, encoding a predicted secreted cazyme, has two splice variants (Figure 4.4c) both annotated with a glycosyl hydrolase domain (IPR domain IPR023296) involved in the hydrolysis of polysaccharide carbohydrates. The primary transcript (ID 2502024) is annotated with a glycoside hydrolase family 43 domain (IPR domain IPR006710), while the alternative transcript (ID CUFF.10781.2) has a different glycosyl hydrolase domain (IPR013148). The presence of alternative domain families in the two transcripts indicates alternative carbohydrate degradation abilities. 102 (24%) of the time course switching genes are predicted to have alternative functionality through alternative domain usage, implicating a change of function of those genes throughput development.

ALTERNATIVE SPLICING CAN AFFECT THE SUBCELLULAR LOCALIZATION OF ENCODED PROTEINS

Alternative splicing may result in the gain or loss of a localization signal leading to a function in another compartment of the cell. For 936 alternatively spliced genes (41%) we were unable to predict a subcellular location for any transcripts (see Supplementary Note S12), and for 1278 genes (56%) we were unable to predict a subcellular location for all transcripts, but in 71 genes (3%), transcripts have different predicted localization signals (Figure 4.3d).

Gene *G2607155* is an example of a predicted metabolic gene with two transcripts encoding proteins with alternative subcellular localizations (see Figure 4.4d). The primary and alternative transcript, 2607155 and CUFF.4351.1, respectively, are both expressed. The protein encoded by the primary transcript is predicted to be transported to the mitochondrion, while the polypeptide encoded by the alternative transcript has an intron retention near the 5' end of the transcript. As a result, it loses its first exon, destroying a valine catabolism-related3-hydroxyisobutyrate dehydrogenase site (IPR002204) specific to mitochondria, and the signaling peptides. The alternative transcript still retains the four deoxyhydrogenase (IPR029154, IPR008927, IPR013328, IPR006115) and the NADP binding domain (IPR016040) of the protein encoded by the primary transcript. Based on its remaining protein sequence, it is predicted that this secondary transcript is located in the cytoplasm, which indicates that alternative splicing may regulate the location of proteins in *S. commune*.

Figure 4.4: *Structures of predicted alternative transcripts for selected genes. Genes are oriented from 5' to 3'. Gray boxes are exons, white boxes are UTRs, thin light gray boxes are introns and dark gray boxes are protein domains. The primary (on average most expressed) transcript is always above the alternative transcript. The expression of the corresponding transcripts are shown on the left hand side of the figure, throughout the five developmental stages: Vegetative (V), Vegetative Induced (VI), Aggregates (A), Primordia (P) and, Mushroom (M). Refer to supplemental File 1 for all transcript structures, and supplemental Table 1 for expression and functional information for each transcript* ***a)*** *Gene G2634198, encoding a carbohydrate active protein, with primary transcript 2634198 and secondary transcript CUFF.11357.2. An intron retention occurs in CUFF.11357.2 fusing the first two exons. (scaffold_9:981266-981925).* ***b)*** *Gene G2629174, encoding a secreted protein. The second intron is retained. (scaffold_7:1,406,122-1,408,401).* ***c)*** *Gene G2502024, encoding a carbohydrate active protein. The 5th intron is retained. (scaffold_6:428624-430098).* ***d)*** *Gene G2607155, a metabolic gene. The first intron is retained. (scaffold_1:2733233-2735448).* ***e)*** *Gene G2679567, a transcription factor. The secondary transcript has an A5SS on its second exon. (scaffold_8:1832047-1833917).* ***f)*** *Gene HOM1, (G2632356), a transcription factor involved in mushroom formation in S. commune. The primary transcript has a secondary transcript in which the second intron is retained. (scaffold_8:1757475-1758654)*



ALTERNATIVE SPLICING IS PRESENT IN KEY FUNCTIONAL GROUPS

Of particular interest are five gene functional groups: i) transcription factors, ii) cazymes, iii) secreted proteins, iv) cytochrome P450s, and v) metabolic genes. Alternative splicing occurs in all these groups (see Figure 4.2e and Supplementary Note S13), and the set of cytochrome P450s are enriched in the set of alternatively spliced genes. The peak of alternative transcript expression in the aggregate stage persists across all these functional groups regardless of expression threshold, demonstrating that alternative splicing is actively regulated in these functional groups (see Supplementary Note S14).

Alternative splicing in a transcription factor gene may impact a DNA binding domain or a protein interaction domain. As an example, the alternative transcript of gene *G2679567* has an alternative 5' splicing site in its second exon (see Figure 4.4e) resulting in a shortened DNA binding domain (IPR021715) due to the loss of most of the third exon. This alternative transcript is differentially expressed across developmental stages, suggesting that the regulation program also changes with alternative splicing. Genes encoding secreted proteins that are alternatively spliced may impact the signal sequence for secretion located in the N-terminal of the protein. The alternative transcript of the TCS gene *G2629174* is the result of an intron retention (see Figure 4.4b). As a result, the N-terminal signal sequence is affected, which is predicted to prevent the protein from being be secreted. If this is the case then, based on the expression of these transcripts, this protein is secreted at different rates throughout development. Many secreted proteins may take on alternative locations in the cell (Figure 4.3d). In 22 cases, secreted proteins (in the extracellular compartment) are predicted to end up in other subcellular locations (8 in the mitochondrion, 8 in the nucleus, 4 in the plasma membrane and 2 in the cytoplasm), and in 62 cases, we were unable to predict the subcellular location (Supplementary Note S12).

IMPLICATIONS ON PREVIOUSLY STUDIED GENES OF *S. commune*

Only 84 *S. commune* genes have been assigned names, 13 of them have alternative splicing variants (see supplementary Note S15. Among these genes are the mating type genes *AAY4*, *ABU6*, and *BBP2-1*, transcription factor genes crea and hom1 that encode the carbon catabolite repression regulator, and the fruiting transcription factor *HOM1*. The alternative transcripts in the mating type genes have no effect on function. In contrast, alternative splicing impacts *HOM1*. Its primary transcript retains the second intron (see Figure 4.4f). This produces a frame shift that prematurely ends the coding region. Due to this, it loses its homeobox domain and thus will not function as a transcription factor. Notably, this *HOM1* transcript is the most abundant throughout development.

DISCUSSION

Previous studies characterizing alternative splicing in fungi unequivocally found intron retentions to be the most common events [10–12]. This study confirmed this finding with intron retentions accounting for 33% of the events. Notably, they are enriched in coding regions, where they account for 43% of the events, while exon skipping occurred primarily in UTRs. With 1% of the events, exon skipping was rare in coding regions, but they represented 57% of the events in UTRs. The prevalence of intron retention contrasts the dominance of exon skipping in mammals [19, 20]. Our results strongly suggest that fungi use alternative splicing events differently when compared to mammalian genomes. Based on event prevalence they are more similar to plants, where intron retentions are also the most common event [21].

Most alternative splicing events (64%) were not reading frame neutral, indicating that there is no strong selective pressure for reading frame neutral alternative splicing events. This is also in contrast with alternative splicing in mammals, where there appears to be a selection for reading-frame neutral events [19, 20].

The incidence of alternative splicing was much higher compared to previous studies. Previous studies examined primarily ascomycetes and few basidiomycetes, only two of which form mushrooms [9, 10]. This may imply that more developmentally complex fungal species have more alternative splicing (see Figure 4.1e and supplementary Note S1) with basidiomycetes having the most alternative splicing, followed by pathogenic ascomycetes, then plant pathogenic ascomycetes and finally, saprobic ascomycetes and yeasts.

The genome of *S. commune* encodes 16,319 genes. The median intergenic distance is 539 bp, making the gene density very high when compared to plants and humans [22, 23]. This hampers studying alternative splicing in *S. commune*, and therefore, we focused only on annotated genes. This prevents the discovery of splicing events at novel loci but allowed us to identify alternatively spliced transcripts of the annotated genes and to study the functional impact of these splicing events. Recent studies from the ENCODE project [23] have shown that significant portions of non-coding regions of the human genome are expressed as mRNAs, and are spliced. By analyzing splice junctions, we observed 5,676 alternative splicing events in non-coding regions, demonstrating that *S. commune* also makes use of alternative splicing in non-coding regions. Due to the density of the genome, events that lie in UTRs cannot be associated with a gene using short read data. Use of long read technologies will alleviate technological problems in assembling the splice variants as those reads will cover entire transcripts.

In predicting the protein sequences of our transcripts, we assumed that the longest Open Reading Frame (ORF) is the one that is most likely to be translated. This is a common assumption in gene prediction software, yet, as a consequence, different transcripts within a gene may have different start codons. The Kozak consensus sequence is a good indicator of translation initiation [24]. The coding sequences predicted in our alternative transcripts indeed are supported by Kozak consensus sequences (see supplement note S17). But, translation initiation is still poorly understood and predicting exactly where translation should initiate (i.e. which ORF is actually translated) is still difficult [25]. It has been shown that there are alternative AUG start codons which produce different proteins from the same transcript [26, 27]. Even beyond that, there are ORFs in the upstream UTRs that are also translated [28]. Future studies of alternative splicing should extend their perspective from the current "one gene, many transcripts", to the "one transcript many proteins" viewpoint.

Finally, we should note that we have been conservative in our prediction of alternative transcripts and alternative splicing events. Additionally, although we sampled across the whole development of *S. commune*, all samples were grown under the same environmental conditions (see Methods) in an isogenic strain (see supplementary Note

S16). Hence, alternative splicing may be much more abundant and provide much more functionality than we observe in this study.

CONCLUSIONS

Alternative splicing was found to increase proteome complexity in *S. commune* by the gain or loss of functional domains or localization signals. Alternative splicing was found in 18% of expressed genes, resulting in 20% more transcripts. Yet, the real incidence of alternative splicing is probably higher. More than two thirds of the alternatively spliced genes exhibited alternative functionality by coding for different domains or by acquiring alternative subcellular localizations. Selected functional groups were shown to contain alternatively spliced genes, with the cytochrome P450 genes showing the highest proportion. Time Course Switching genes that differentially express gene variants throughout development represented 19% of the alternatively spliced genes. Of these genes, 24% were predicted to have alternative functionality. We thus show that alternative splicing dynamically varies the complexity of the proteome encoded in the dense genome of *S. commune*.

It has long been believed that alternative splicing does not contribute functionality to the cell in plants, or fungi, despite the presence of alternative splicing events [29]. Alternative splicing events were generally thought to result in non-functional proteins or to be similar to other transcripts of the same gene, resulting in retained functionality. Similarly, in the early stages of the ENCODE project it was also commonly believed that the vast majority of alternative mammalian transcripts do not have functional roles [30]. Now, the role of alternative splicing is commonly accepted in humans [31] and its tremendous potential influence on function is being accepted in plants [32–34]. With this work, we extend this notion to fungi.

METHODS

S. commune GENOME AND ANNOTATIONS

The genome sequence of *S. commune* v3.0 was retrieved from the Joint Genome Initiative database [35], together with gene definitions (GFF), Interpro and GO annotations.

MRNA ISOLATION

mRNA was isolated from *S. commune* strain H4-8 [36] grown at 25 °C on minimal medium containing 1% glucose and 1.5% agar [37]. RNA was sampled during (i) vegetative growth, (ii) induced vegetative growth, (iii) aggregate formation ,(iv) primordia formation, and (v) when mushroom had developed. In the induced vegetative growth stage, the colony has been exposed to light for some time but has not yet started forming aggregates. RNA was also isolated from 9 dikaryotic deletion strains at two time points [16–18]: Δ*WC-1*Δ*WC-1*; Δ*WC-2*Δ*WC-2*; Δ*HOM1*Δ*HOM1*; Δ*HOM2*Δ*HOM2*; Δ*FST3*Δ*FST3*; Δ*FST4*Δ*FST4*; Δ*BRI1*Δ*BRI1*; Δ*GAT1*Δ*GAT1*; and Δ*C2H2*Δ*C2H2*. The sample at the first time point was taken at the time where a simultaneously growing wildtype sample reached the aggregates stage of development. The sample at the second time point was taken when the wildtype had formed mature mushrooms (see supplementary Note S18).

RNA SEQUENCING

Samples were sequenced with 100bp paired-end sequencing on the HiSeq 2000 Illumina platform with an average of 1.7Gb of reads per sample (average 43x genome coverage). Exceptions are the wildtype stages i) vegetative growth, ii) induced vegetative growth and v) primordia, which were sequenced at a later date with 125bp paired end sequences on the HiSeq 2500 platform with an average of 5.0Gb of reads per sample. All samples were produced in duplo for use in transcript abundance estimation. Raw RNA-seq reads are filtered for quality using trimmomatic [38]. These RNA-seq samples have been made available under BioProject PRJNA323434.

REGION-RESTRICTED PROBABILISTIC MODELING (RRPM) FOR ALTERNATIVE TRANSCRIPT DISCOVERY

Due to the high gene-density of *S. commune*, very few neighboring genes have intergenic regions without expression. In order to reduce the interference of neighboring UTRs, we restricted the discovery of alternative transcripts to gene coding regions only. To this end, the genome was split at gene boundaries into as many fragments as there are genes, each fragment representing a different annotated gene region on the *S. commune* genome.

Reads from all samples were aligned to the collection of gene-based fragments using STAR [39], producing a splice junction database used to optimize STAR alignments in a second alignment round (we use all samples to have a comprehensive database). The second round of STAR produced BAM files sorted by coordinate. Reads not aligning entirely within a gene-based fragment were discarded. Using these BAM files, Cufflinks [14] was run in Reference Annotation Based Transcript Assembly (RABT) mode, to produce a set of predicted transcripts. Although the transcripts were predicted on the basis of all samples, it does not mean that these transcripts are also active in every sample. These transcripts were then projected back onto the original genome to restore their context. See Supplementary Note S19 for a schematic of RRPM, and Supplementary Note S20 for parameter settings used for the tools mentioned above.

TRANSCRIPT FILTERING

The output from Cufflinks was filtered to remove transcripts resulting from noise: 1) predicted transcripts which do not lie on the same strand as the original gene description were removed (they cannot be accurately called without strand-specific reads); 2) transcripts which had no expression or did not have reads supporting its splice junctions were removed; 3) transcripts which did not contain Open Reading Frames (ORFs) of at least length 40nt were removed. See Supplementary Note S21 for an overview of how many genes and transcripts were filtered at each step. The final output is a list of genes with their corresponding transcripts in GFF format.

CALLING ALTERNATIVE SPLICING EVENTS

To identify alternative splicing events, we must compare the found transcripts against a reference. Comparing transcripts to the original reference annotation may result in peculiarities due to a lagging quality of annotation. For example, the original annotation may describe a large exon, whereas all predicted transcripts indicate only two smaller exons. To find a suitable reference, we established a consensus annotation based on all expressed transcripts for that gene. The consensus annotation contains all distinctly observed exons in a gene, without intron retentions (see supplementary Note S22). Consequently, the consensus annotation is not necessarily one of the transcripts of the gene. In fact, it might not even contain a valid ORF. It just serves as a reference to describe observed splice variations.

We called primary splice events by aligning the exons of an observed transcript to the consensus annotation. From this alignment we determine the absence of an exon (ES), the fusion of exons (IR), or the extension or contraction of exon boundaries (A5SS, A3SS) (see Supplementary Note S23). Composite events "multiple alternative 5'" and "multiple alternative 3'" splicing sites are classified by grouping exons of transcripts within one gene and counting the number of alternative exon boundaries for each exon. Mutually exclusive exons were detected by counting how often pairs of exons occur across all transcripts. If a pair was never observed across all transcripts in a gene but individually the exons are observed, then we called that pair a mutually exclusive exon event.

We counted the number of alternative splicing events at three levels. At the event level, we counted each alternative splicing event. If, for example, three transcripts exist in a gene, and in two of these transcripts the third exon is skipped, we count this as two events. At the transcript level, we counted each transcript with a particular event. Thus, when a transcript skips two exons, we count it only once. At the gene level, we count the number of genes with a specific alternative splicing event. For example, when a gene has two transcripts that each skip different exons we would counted that as one event.

DETECTION OF UNTRANSLATED SPLICING EVENTS

We quantify splicing events outside of the translated regions similar to Xie et. al.[12] through the study of clusters of splice junctions [12]. The slight differences are described in supplementary Note S9.

ESTIMATING TRANSCRIPT EXPRESSION PROFILE OF DEVELOPMENT

Abundance estimates for each transcript were computed by the Cufflinks suite. For each sample separately, we aligned both replicates to the entire genome with STAR in the same two-round procedure as used for transcript prediction. We provided the resulting BAM files of both replicates to Cufflinks, which estimates the abundances of the transcripts in both replicated samples together. Doing this for each developmental stage results in an expression profile for each predicted transcript.

TIME COURSE SWITCHES

Based on the expression profiles of each transcript, we detected time course switching genes for each gene using the method described by Lees et al. [5]. Briefly, we compare the expression of the primary transcript (which has the highest average expression across time) as well as the expression of an alternative transcript. Intuitively speaking, this comparison scores the relative expression of both transcripts in such a way that the score is larger when the expression profiles of the two transcripts are more dissimilar in shape. As in Lees et. al.[5], genes which have a score greater than 0.5 are considered to be time course switches. This results in a list of genes which have alternative transcripts that swap expression with another.

FUNCTIONAL ANNOTATIONS AND GROUP DEFINITIONS

Functional domain annotations: InterPro domain annotations were predicted for the translated protein sequence of the longest ORF in each predicted transcript using InterProScan [40]. We re-index these annotations based on location. If a gene has the same annotation twice, but in two different locations, we assign each of them new identifiers. This allows us to more specifically identify alternative functionality by domain annotations.

Transcription factor definitions: Transcription factors were predicted based on 83 InterProDNA-binding or regulatory domains (see supplementary Note S24) suggested by the Fungal Transcription Factor Database [41, 42]. Transcription factors are predicted for the original gene definitions. If one of these InterPro domains was present in the protein sequence of a gene, then the gene was called as a transcription factor.

Carbohydrate-active enzymes prediction: Using the Cazymes Analysis Toolkit (CAT) [43], we predicted carbohydrate-active enzymes based on the original gene definitions. If a gene's protein sequence was predicted to be a cazyme by either the sequence-based annotation method or the Pfam-based annotation method then we considered it a cazyme.

Secreted Proteins prediction: We used the same procedure as [44] to predict secreted proteins. Briefly, genes with SignalP [45] signal peptides, or a TargetP [46] Loc=S were kept. The remaining genes were further filtered with TMHMM [47], keeping only genes with zero or one transmembrane domains. Finally, genes were filtered using Wolf PSort [48] to select genes with a Wolf PSort extracellular score greater than 17.

Metabolic and Cytochrome P450 gene groups: Genes with the GO annotation "metabolic process" (annotation ID: GO:0008152) were called as metabolism genes. Genes with the IPR annotation IPR001128 were used as Cytochrome P450 genes.

Subcellular location prediction: We used Wolf PSort [48] to predict the subcellular localization of alternatively spliced transcripts. Protein sequences for each transcript were constructed, and provided to Wolf PSort. All annotations with a score below 17 were removed [44].

BIBLIOGRAPHY

[1] M. W. Medina, R. M. Krauss. Alternative splicing in the regulation of cholesterol homeostasis. *Current Opinion in Lipidology* 24(2):147–152 (2013).

[2] S. Oltean, D. O. Bates. Hallmarks of alternative splicing in cancer. *Oncogene* 33(November):1–8 (2013).

[3] S. Gueroussov, et al. An alternative splicing event amplifies evolutionary differences between vertebrates. *Science* 349(6250):868–873 (2015).

[4] Y.-J. Kwon, M.-J. Park, S.-G. Kim, I. T. Baldwin, C.-M. Park. Alternative splicing and nonsense-mediated decay of circadian clock genes under environmental stress conditions in Arabidopsis. *BMC Plant Biology* 14(1):136 (2014).

[5] J. G. Lees, J. A. Ranea, C. A. Orengo. Identifying and characterising key alternative splicing events in Drosophila development. *BMC Genomics* 16(1):608 (2015).

[6] J. Vilardell, P. Chartrand, R. H. Singer, J. R. Warner. The odyssey of a regulated transcript. *RNA (New York, NY)* 6(12):1773–1780 (2000).

[7] E. Boel, I. Hjort, B. Svensson, F. Norris, N. P. Fiil. Glucoamylases G1 and G2 from Aspergillus niger are synthesized from two different but closely related mRNAs. *The EMBO journal* 3(5):1097–1102 (1984).

[8] K. Okazaki, O. Niwa. mRNAs encoding zinc finger protein isoforms are expressed by alternative splicing of an in-frame intron in fission yeast. *DNA research : an international journal for rapid publication of reports on genes and genomes* 7(1):27–30 (2000).

[9] K. Grützmann, et al. Fungal alternative splicing is associated with multicellular complexity and virulence: A genome-wide multi-species study. *DNA Research* 21(October 2013):27–39 (2014).

[10] A. M. McGuire, M. D. Pearson, D. E. Neafsey, J. E. Galagan. Cross-kingdom patterns of alternative splicing and splice recognition. *Genome biology* 9(3):R50 (2008).

[11] C. Zhao, C. Waalwijk, P. J. G. M. de Wit, D. Tang, T. van der Lee. RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen Fusarium graminearum. *BMC genomics* 14(1):21 (2013).

[12] B.-B. Xie, et al. Deep RNA sequencing reveals a high frequency of alternative splicing events in the fungus Trichoderma longibrachiatum. *BMC genomics* 16(1):54 (2015).

[13] M. G. Grabherr, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29(7):644–52 (2011).

[14] C. Trapnell, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* 28(5):511–5 (2010).

[15] S. P. Gordon, et al. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS ONE* 10(7):1–15 (2015).

[16] R. a. Ohm, D. Aerts, H. a. B. Wösten, L. G. Lugones. The blue light receptor complex WC-1/2 of Schizophyllum commune is involved in mushroom formation and protection against phototoxicity. *Environmental Microbiology* 15(3):943–955 (2013).

[17] R. a. Ohm, et al. An efficient gene deletion procedure for the mushroom-forming basidiomycete Schizophyllum commune. *World journal of microbiology & biotechnology* 26(10):1919–1923 (2010).

[18] R. a. Ohm, J. F. de Jong, C. de Bekker, H. a. B. Wösten, L. G. Lugones. Transcription factor genes of Schizophyllum commune involved in regulation of mushroom formation. *Molecular Microbiology* 81(6):1433–1445 (2011).

[19] M. Sammeth, S. Foissac, R. Guigó. A general definition and nomenclature for alternative splicing events. *PLoS Computational Biology* 4(8) (2008).

[20] Y. Xing, C. Lee. Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. *Nature reviews Genetics* 7(7):499–509 (2006).

[21] S. Chamala, G. Feng, C. Chavarro, W. B. Barbazuk. Genome-Wide Identification of Evolutionarily Conserved Alternative Splicing Events in Flowering Plants. *Frontiers in Bioengineering and Biotechnology* 3(March) (2015).

[22] N. N. Alexandrov, et al. Features of Arabidopsis genes and genome discovered using full-length cDNAs. *Plant Molecular Biology* 60(1):69–85 (2006).

[23] S. Djebali, et al. Landscape of transcription in human cells. *Nature* 489(7414):101–8 (2012).

[24] M. Kozak. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic acids research* 15(20):8125–48 (1987).

[25] H. Zur, T. Tuller. New universal rules of eukaryotic translation initiation fidelity. *PLoS computational biology* 9(7):e1003136 (2013).

[26] N. T. Ingolia. Ribosome profiling: new views of translation, from single codons to genome scale. *Nature reviews Genetics* 15(3):205–13 (2014).

[27] A. V. Kochetov. Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *BioEssays* 30(7):683–691 (2008).

[28] S. R. Starck, et al. Translation from the 5' untranslated region shapes the integrated stress response. *Science* 351(6272):aad3867–aad3867 (2016).

[29] N. H. Syed, M. Kalyna, Y. Marquez, A. Barta, J. W. S. Brown. Alternative splicing in plants - coming of age. *Trends in Plant Science* 17(10):616–623 (2012).

[30] M. L. Tress, et al. The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences of the United States of America* 104(13):5495–5500 (2007).

[31] H. Keren, G. Lev-Maor, G. Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nature reviews Genetics* 11(5):345–355 (2010).

[32] W. B. Barbazuk, Y. Fu, K. M. McGinnis. Genome-wide analyses of alternative splicing in plants: Opportunities and challenges. *Genome Research* 18(9):1381–1392 (2008).

[33] A. S. N. Reddy, Y. Marquez, M. Kalyna, A. Barta. Complexity of the Alternative Splicing Landscape in Plants. *The Plant Cell* 25(10):3657–3683 (2013).

[34] S. Filichkin, H. D. Priest, M. Megraw, T. C. Mockler. Alternative splicing in plants: directing traffic at the crossroads of adaptation and environmental stress. *Current Opinion in Plant Biology* 24:125–135 (2015).

[35] I. V. Grigoriev, et al. The genome portal of the Department of Energy Joint Genome Institute. *Nucleic acids research* 40(Database issue):D26–32 (2012).

[36] R. a. Ohm, et al. Genome sequence of the model mushroom Schizophyllum commune. *Nature biotechnology* 28(9):957–63 (2010).

[37] A. F. Van Peer, C. De Bekker, A. Vinck, H. a. B. Wösten, L. G. Lugones. Phleomycin increases transformation efficiency and promotes single integrations in schizophyllum commune. *Applied and Environmental Microbiology* 75(5):1243–1247 (2009).

[38] A. M. Bolger, M. Lohse, B. Usadel. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120 (2014).

[39] A. Dobin, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* 29(1):15–21 (2013).

[40] E. M. Zdobnov, R. Apweiler. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics (Oxford, England)* 17(9):847–848 (2001).

[41] J. Park, et al. CFGP: A web-based, comparative fungal genomics platform. *Nucleic Acids Research* 36(SUPPL. 1):D562–71 (2008).

[42] J. Park, et al. FTFD: an informatics pipeline supporting phylogenomic analysis of fungal transcription factors. *Bioinformatics (Oxford, England)* 24(7):1024–5 (2008).

[43] V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, B. Henrissat. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research* 42(D1):D490–D495 (2014).

[44] A. Morais do Amaral, J. Antoniw, J. J. Rudd, K. E. Hammond-Kosack. Defining the Predicted Protein Secretome of the Fungal Wheat Leaf Pathogen Mycosphaerella graminicola. *PLoS ONE* 7(12):1–19 (2012).

[45] T. N. Petersen, S. r. Brunak, G. von Heijne, H. Nielsen. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* 8(10):785–786 (2011).

[46] O. Emanuelsson, H. Nielsen. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of molecular ...* 300(4):1005–16 (2000).

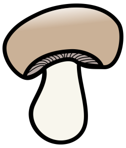
[47] A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology* 305(3):567–580 (2001).

[48] P. Horton, et al. WoLF PSORT: protein localization predictor. *Nucleic Acids Research* 35(Web Server):W585–W587 (2007).

Chapter 5

Nucleus specific expression in the multinucleated mushroom-forming fungus *Agaricus bisporus* reveals different nuclear regulatory programs

Thies Gehrmann
Jordi F. Pelkmans
Robin A. Ohm
Aurin M. Vos
Anton S.M Sonnenberg
Johan J.P. Baars
Han A. B. Wösten
Marcel J. T. Reinders
Thomas Abeel



A preprint of this chapter has been published as: *bioR$_\chi$ iv*, July 2017 (doi: 10.1101/141689)
Supplementary material is available online at: `http://bit.ly/2rpGgwX` (biorxiv.org)

ABSTRACT

Motivation: *Fungi are essential in nutrient recycling in nature. They also form symbiotic, commensal, parasitic and pathogenic interactions with other organisms including plants, animals and humans. Many fungi are polykaryotic, containing multiple nuclei per cell. In the case of heterokaryons, there are even different nuclear types within a cell. It is unknown what the different nuclear types contribute in terms of mRNA expression levels in fungal heterokaryons. Each cell of the cultivated, mushroom forming basidiomycete Agaricus bisporus contains 2 to 25 nuclei of two nuclear types that originate from two parental strains. Using RNA-Seq data, we wish to assess the differential mRNA contribution of individual nuclear types in heterokaryotic cells and its functional impact.*

Results: *We studied differential expression between genes of the two nuclear types throughout mushroom development of A. bisporus in various tissue types. The two nuclear types, P1 and P2, produced specific mRNA profiles which changed through development of the mushroom. The differential regulation occurred at the gene level, rather than at locus, chromosomal or nuclear level. Although the P1 nuclear type dominates the mRNA production throughout development, the P2 type showed more differentially upregulated genes in important functional groups including genes involved in metabolism and genes encoding secreted proteins. Out of 5,090 karyolelle pairs, i.e. genes with different alleles in the two nuclear types, 411 were differentially expressed, of which 246 were up-regulated by the P2 type. In the vegetative mycelium, the P2 nucleus up-regulated almost three-fold more metabolic genes and cazymes than P1, suggesting phenotypic differences in growth. A total of 10% of the differential karyollele expression is associated with differential methylation states, indicating that epigenetic mechanisms may be partly responsible for nuclear specific expression.*

Conclusion: *We have identified widespread transcriptomic variation between the two nuclear types of A. bisporus. Our novel method enables studying karyollelle specific expression which likely influences the phenotype of a fungus in a polykaryotic stage. This is thus relevant for the performance of these fungi as a crop and for improving this species for breeding. Our findings could have a wider impact to better understand fungi as pathogens. This work provides the first insight into the transcriptomic variation introduced by genomic nuclear separation. Introduction*

INTRODUCTION

Fungi are vital to many ecosystems, contributing to soil health, plant growth, and nutrient recycling[1]. They are key players in the degradation of plant waste[2, 3], form mutually beneficial relationships with plants by sharing minerals in exchange for carbon sources[4, 5] and by inhibiting the growth of root pathogens[6, 7]. They even form networks between plants, which can signal each other when attacked by parasites[8]. Yet, some are plant pathogens responsible for huge economic losses in crops[9–11].

The genome organization of fungi is incredibly diverse and can change during the life cycle. For instance, sexual spores can be haploid with one or more nuclei or can be diploid. Sexual spores of mushroom forming fungi are mostly haploid and they form monokaryotic (one haploid nucleus per cell) or homokaryotic (two or more copies of genetically identical haploid nuclei) mycelia upon germination. Mating between two

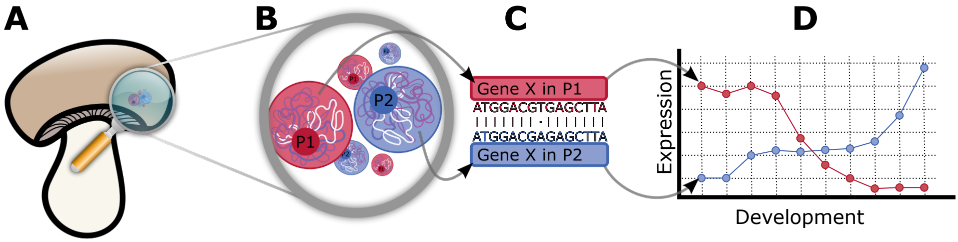


Figure 5.1: *Nuclear type specific expression in A. bisporus.* ***A)*** *The A. bisporus mushroom is composed of different tissues that consist of hyphae comprised of cellular compartments.* ***B)*** *Each cellular compartment is a heterokaryon containing between 2 and 25 nuclei. In our strain, each nucleus is either of type P1 (red) or P2 (blue). Both nuclear types are haploid, and contain exactly one copy of each gene. However, because there are multiple nuclei, there may be multiple copies of each gene in the cell.* ***C)*** *Furthermore, the gene in the two types, which we call karyolleles, may differ in their genetic sequences.* ***D)*** *These differences in transcript sequence allow us to quantify expression of each karyollele in each tissue and to investigate nucleus specific expression.*

such mycelia results in a fertile dikaryon (one copy of the parental nuclei per cell) or heterokaryon (two or more copies of each parental nuclei) when they have different mating loci[12]. In contrast to eukaryotes of other kingdoms, the nuclei do not fuse into di- or polyploid nuclei but remain side by side during the main part of the life cycle. Only just before spores are formed in mushrooms, do these nuclei fuse, starting the cycle anew.

Agaricus bisporus is the most widely produced and consumed edible mushroom in the world[2]. Heterokaryotic mycelia of the button mushroom *Agaricus bisporus* var. bisporus (Sylvan A15 strains) have between 2 and 25 nuclei per cell[13, 14] (Figure 5.1). The genomes of both ancestral homokaryons have been sequenced[1, 15] showing that DNA sequence variation is associated with different vegetative growth capabilities[1]. Due to the two nuclear types, each gene exists at two alleles separated by nuclear membranes, which we call karyolleles. Although there have been a few studies investigating the expression of genetic variety in the transcriptome[16, 17], the differential transcriptomic activity of two (or more) nuclear types has never been systematically investigated in a heterokaryon at the genome wide scale. Based on SNPs identified in mRNA sequencing, it has been suggested that allele specific expression is tightly linked to the ratio of the nuclear types in a basidiomycete[18].

Allele specific expression in mononuclear cells has been studied in fungi[19], plants[20], animals[21], and humans[22]. Such studies have shown that allele heterogeneity is linked to differential allele expression and cis-regulatory effects[21–23], and even sub-genome dominance[24]. *A. bisporus* is in many ways an excellent model organism to investigate differential karyollele expression. It only has two nuclear types in the heterokaryon contrasting to the mycorrhizae that can have more nuclear types[25, 26], making computational deconvolution of mRNA sequence data intractable with currently available tools. Additionally, the recently published genomes of the two nuclear types of Sylvan A15[15] exhibit a SNP density of 1 in 98 bp allowing differentiation of transcripts in high throughput sequencing data. Finally, bulk RNA-Seq datasets of different stages of development and of different tissues of the fruiting bodies are available[2, 27]. Here, we show that differential karyollele expression exists in *Agaricus bisporus* Sylvan A15 strain, which changes across tissue type and development and affects different functional groups. Further, we show that differential karyollele expression associates with differential methylation states, suggesting that epigenetic factors may be a cause for the differential regulation of karyolleles.

RESULTS

KARYOLLELE SPECIFIC EXPRESSION THROUGH SEQUENCE DIFFERENCES

To assign expression levels to individual karyolleles, we exploit sequence differences between karyollele pairs in the P1 and P2 homokaryon genomes of *A. bisporus* A15 strain (Materials). Briefly, the sequence differences define marker sequences for which the RNA-Seq reads uniquely match to either the P1 or the P2 variant, effectively deconvolving the mRNA expression from the two nuclear types (see Methods). There are a total of 5,090 distinguishable karyollele pairs between the P1 and P2 genomes, corresponding to 46% of all genes. The remaining genes could not be unambiguously matched, or the karyollele pairs had too few sequences differences. Most (80%) distinguishable karyollele pairs had the same number of markers in each homokaryon. For the remaining pairs (20%), the number of markers per karyollele was different (see Supplementary Material Note A). This variation can be explained by the non-symmetric number of markers produced by the different kinds of variation. While a SNP will result in one marker in each karyollele, an indel (if longer than 21bp) will result in one marker in one karyollele, and at least two in the other. Karyollele specific expression is expressed as a read count ratio that reflects the relative abundance of mRNAs originating from the P1 or P2 nuclear types (Equation 5.3, Methods).

We studied *A. bisporus'* karyollele specific expression for different tissues and development in two RNA-Seq datasets, one studying the mycelium in compost throughout mushroom harvest, and one studying different mushroom tissues throughout mushroom formation (Figure 5.2, Supplementary Material Note C, and Materials). Measured difference in expression between nuclear types is not correlated with the number of markers ($p > 0.05$) for any of the samples, nor is it correlated with CG content (see Supplementary Material Note B).

P1 AND P2 MRNA PRODUCTION DIFFERS PER TISSUE AND ACROSS DEVELOPMENT

First, we assess the total mRNA production of the P1 and P2 nuclear types and their relative contributions during development. To do this, we considered the total number of reads uniquely matching to P1 with respect to P2. Figure 5.2 shows that this nuclear type read count ratio (NRR, see Equation 5.5, Methods) changes throughout development and across tissue types. For example, during the 'Differentiated' stage, the P2 nuclei are dominant in the skin, but in the 'Young Fruiting Body', the P1 nuclei dominate the skin (two right most panels in Figure 5.2). In contrast, the 'Stipe Center' is dominated by P1 nuclei in the differentiated stage, while later the expression of P2 nuclei dominates.

The transcription patterns throughout the mushroom development differ between the karyolleles. Based on a principal component analysis of the expression profiles of each nuclear type, we observe that the expression profiles of P1 and P2 group together in

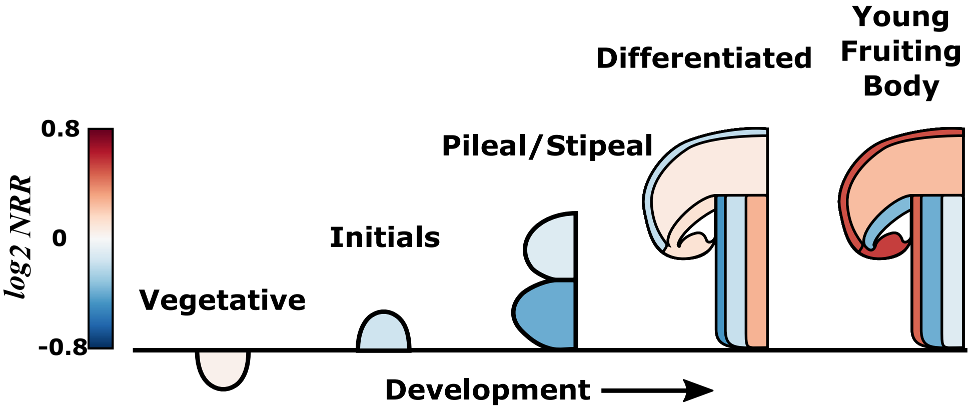


Figure 5.2: *Read count ratios at the nuclear type level (Equation 5.5) of Agaricus bisporus throughout its development. Red colour indicates higher P1 activity, blue colour higher P2 activity. The scale bar indicates the log2 fold change in activity between the P1 and P2 nuclear types. We observe a differential mRNA activity in different mushroom tissues.*

different clusters, based on the first and second principal components (Supplementary Material Note D). This clustering is indicative of distinct regulatory programs. It appears as though the first principal component represents the tissue type, and the second represents the nuclear type. Interestingly, measurements of the same tissue from P1 and P2 do not have exactly the same value for the first principal component, indicating that the difference in nuclear type does not entirely explain the variation between P1 and P2.

WITHIN A SAMPLE, MRNA PRODUCTION OF P1 AND P2 VARY BETWEEN CHROMOSOMES

Figure 5.3 shows the Chromosome Read count Ratios (CRR, Equation 5.4), demonstrating that some chromosomes are more active in P1 (e.g. chromosome 8) throughout development, while others are more active in P2 (e.g. chromosome 9). Expression of other chromosomes depend on the developmental state, changing in time (e.g. chromosome 2). The chromosome log2 fold changes lie between [-0.60, 0.79]. In the vegetative mycelium we see less drastic differences in mRNA production throughout development than in the mushroom tissues, with expression log2 fold changes between [-0.28, 0.36] (see Supplementary Material Note C).

A FEW HIGHLY EXPRESSED GENES REPRESENT A LARGE COMPONENT OF CHROMOSOME MRNA ACTIVITY

In Figure 5.3, we showed that more mRNA originates from P2 in the case of, for example, chromosome 9 than from P1. This was in part due to a few genes which were very highly expressed. These highly expressed genes skew the read count ratios (see Supplementary Material Note E). The differences can be quite extreme; In one case, a P1 karyollele accounted for <1% of all reads originating from chromosome 9, while its P2 karyollele accounted for 21% of all the chromosome 9 reads. Hence, most of the observed differences

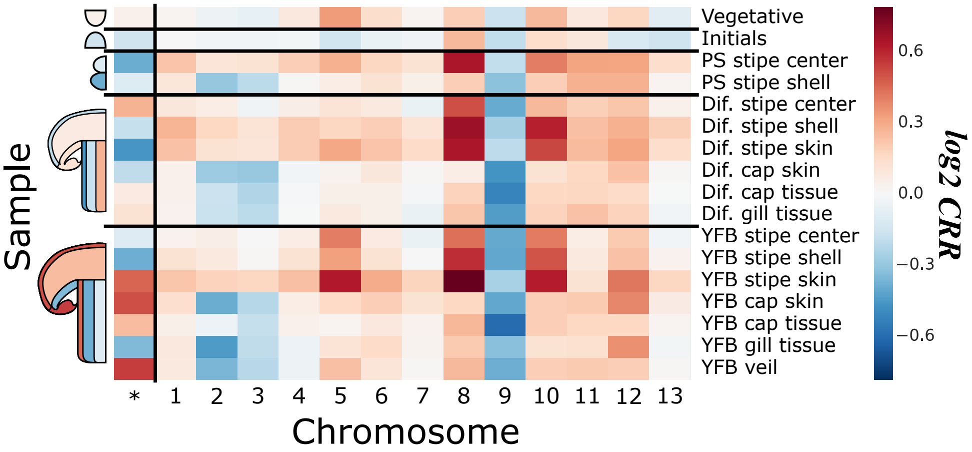


Figure 5.3: *The read count ratios at the chromosome level (Equation 5.4) throughout development of the mushroom. The different tissues are shown along the y-axis, and the different scaffolds are given along the x-axis. Red colour indicates higher P1 activity, blue colour higher P2 activity. The first column provides the read count ratios at the nuclear type level (NRR) from Figure 5.1. Supplementary Material Note G provides the read count ratios at the chromosome level in the vegetative mycelium dataset.*

for chromosome 9 (Figure 5.3) is explained by such highly expressed genes (Supplementay Note F).

In total, we identified 22 genes whose contribution exceeds 10% of the total expression of the chromosome it is located on. Most of these genes are differentially expressed between the two nuclear types, with 16 showing fold changes larger than 2 (Supplementary Material Note F). These genes are primarily metabolic.

GENE READ RATIOS REVEAL A DOMINANT P1 TYPE IN MUSHROOM TISSUE, BUT NOT IN MYCELIUM

To investigate whether either nuclear type is truly dominant we correct for extremely highly expressed genes by limiting their impact on the chromosome and tissue level ratios by using per-gene activity ratios per chromosome (CGR, Equation 5.6), instead of read ratios. This revealed that, in addition to P1 producing more mRNA than P2, P1 karyolleles were also more frequently higher expressed than their P2 counterpart (Figure 5.4). Looking across all tissues and chromosomes, P1 is significantly dominant over P2, i.e. teragehe average of the log-transformed CGR is significantly larger than zero, following a t-test in mushroom tissue, with $p < 0.01$, (see Supplementary Material Note G). Using the Chromosome Gene Ratio has a notable impact on chromosome 9. Although P2 produces most chromosome 9 mRNA (Figure 5.3), it is not the case that more P2 karyolleles are more highly expressed than P1 karyolleles.

We do not observe such a dominance of P1 in the mycelium ($p > 0.05$, with t-test as in mushroom dataset), where neither P1 nor P2 show a dominant mRNA activity (see Supplementary Material Note H).

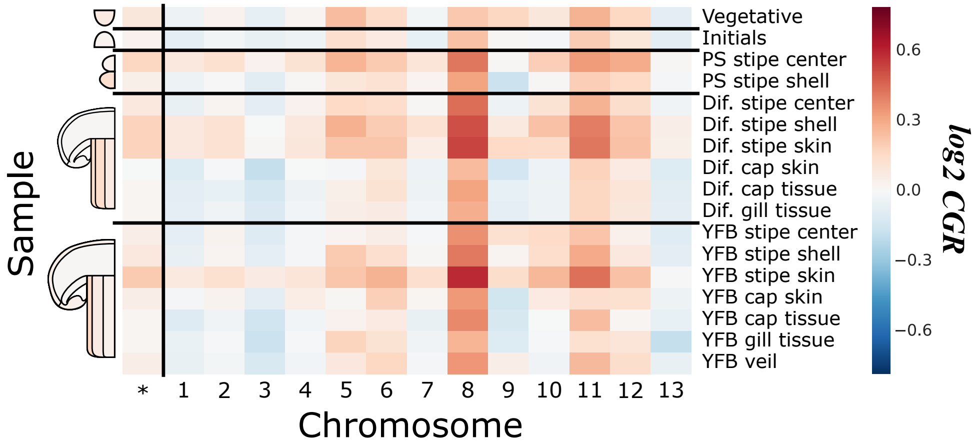


Figure 5.4: *The chromosome gene ratios (Equation 5.6) per nuclear type at the chromosome level (Equation 5.4) (rows) and per chromosome (columns). Red colour indicates higher P1 activity, blue colour higher P2 activity. The first column (indicated with a star) represents gene ratio measurements for each tissue (see Supplementary Material Note G). P1 produces more mRNA per gene on average than P2. This is particularly striking in scaffold 8. See Supplementary Material Note H for the gene ratio measures in the vegetative mycelium dataset.*

A SUBSTANTIAL PORTION OF KARYOLLELES ARE DIFFERENTIALLY EXPRESSED

In each tissue, we determined the set of karyolleles which are statistically significantly differentially expressed between the two nuclear types. Although the dominance of the P1 nuclear type indicates a general trend of higher activity across many genes, some karyollele pairs have a much larger difference pointing towards a functional role. In total, we find 411 genes that are differentially expressed (see Methods) in a mushroom tissue or in vegetative mycelium throughout development (Table 1); 368 genes are differentially expressed in mushroom tissues, and 82 in the vegetative mycelium. Interestingly, when a karyollele pair is differentially expressed, with only a few exceptions (see Supplementary Material Note I), it will always be observed to be more highly expressed in the same nuclear type, i.e. if a gene is observed to be more highly expressed in P1 than in P2, than it will never be observed to be more highly expressed in P2 than in P1 in other tissues, and vice versa. The only exceptions to this rule lie in the set of genes that are differentially expressed in both the mushroom dataset and the mycelium dataset.

The set of differentially higher expressed genes between the nuclear types in mushroom and mycelium sets overlap with only 39 genes. In this intersection set, more genes are higher expressed in P2 than in P1. Ten genes had a higher expression in P1, and 24 had a higher expression in P2. Five were more highly expressed in P2 in the mycelium, but switched their origin of primary expression to P1 in the mushroom (see Supplementary Material Note I). The lack of a substantial overlap of differentially expressed genes between the two nuclear types is indicative of different regulatory processes during the vegetative stage and a mushroom stage.

Although P2 upregulates more differentially expressed genes than P1 does, more genes show a consistently higher expression in P1 than in P2. We identify consistently

Table 5.1: *Karyolleles differentially expressed between P1 and P2 in mushroom tissue and vegetative mycelium across development. In the first row we indicate the number of differentially expressed genes that are higher expressed in the different nuclear types for the two datasets (columns). The second row gives the total number of differentially expressed genes in the two different datasets. Row three shows the number of differentially expressed genes in a dataset that are not differentially expressed in the other dataset. In the last row, we show the number of differentially expressed genes that overlap between the two datasets.*

	Mushroom tissue dataset		**Mycelium tissue dataset**	
	P1 up	**P2 up**	**P1 up**	**P2 up**
Diff. ex.	176	193	30	52
Total/dataset	368		82	
Unique/dataset	329		43	
Overlap	39 (411 total)			

higher expressed genes that show a higher expression in one nuclear type over the other across all samples (Methods). In the mushroom tissue dataset, we find 1,115 genes that are consistently higher expressed in P1, and 785 genes that are consistently higher expressed in P2. Similarly, in the vegetative mycelium, we find 832 genes that are consistently higher expressed in P1 and 645 that are consistently higher expressed in P2. The two datasets overlap with 470 and 256 genes for P1 and P2, respectively.

CO-LOCALIZED GENE CLUSTERS ARE CO-REGULATED

To investigate the level at which genes are regulated, we investigated whether there are regions where the majority of genes were consistently higher expressed in one homokaryon than in the other. We detected many of such regions, given in Table 2 and Figure 5.5 (Methods, Supplementary Material Note J), hinting towards a sub-chromosomal level of regulation. This is supported by observations in Figures 3 and 4, where we see that within one tissue chromosomes are differently regulated, excluding a regulation at the nuclear level. Because we observe that co-regulated gene are co-localized in regions, regulation can also not occur at the chromosome level, because then we would have expected regions of co-regulation of the size of whole chromosomes.

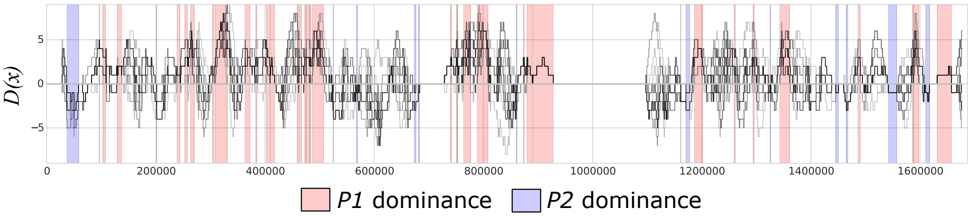


Figure 5.5: *Co-localized genes are often co-regulated. Pictured here are the co-localized and co- regulated gene clusters along chromosome 10 in the mushroom tissue dataset. Along the x-axis is the genomic co-ordinate. For each sample (gray lines), we plot the difference between the number of genes more highly expressed by P1 and the number of genes more highly expressed by P2 (a value of 0 indicates an equal distribution). We also highlight the regions that are consistently upregulated in P1 (red regions) and the number of genes that are consistently upregulated in P2 (blue regions). See Supplementary Material Note J for other chromosomes.*

Table 5.2: *The number of regions in which the majority of the genes are coregulated (Methods), across the mushroom and mycelium datasets and with the number of genes in these regions. P1 and P2 columns indicate whether the region is consistently higher in for the P1 kayollele or the P2 karyollele, respectiverly. Row Both indicates overlapping regions between the mushroom and vegetative mycelium datasets. Supplementary Material Note J offers detailed expression profiles of these regions.*

	P1		P2	
Dataset	**#Regions**	**#Genes**	**#Regions**	**#Genes**
Mushroom	207	741	73	233
Vegetative Mycelium	414	1955	43	140
Both	151	484	7	17

Co-regulated regions are more frequently upregulated for the P1 karyollele than for the P2 karyolleles. This observation is in agreement with the observed P1 nuclear type dominance. We observe relatively little overlap between the Mushroom and Vegetative Mycelium datasets (Table 2), indicative of different regulatory programs between the vegetative mycelium and mushroom tissue cells.

MANGANESE PEROXIDASE IS UP-REGULATED IN VEGETATIVE MYCELIUM BY P2, BUT IN MUSHROOM BY P1

Of 90 genes with named annotations in *A. bisporus* (see Methods), 42 are identified as differentiable karyollele pairs, and one, manganese peroxidase (*mnp1*) was differentially expressed between the two nuclear types in any stage of development. *mnp1* is known to be highly expressed in early stages of development, and drops to much lower levels (log fold change of -2.8) after mushroom formation[2, 28]. In our datasets, the individual contributions of P1 and P2 to *mnp1* expression are largely different. In the vegetative mycelium, we find that P2 produces four-fold more *mnp1* immediately before mushroom formation than P1 (see Supplementary Material Note K). In the mushroom tissue, however, *mnp1* is expressed on average 4.2-fold higher by P1 in the stem of the fruiting body throughout development (see Supplementary Material Note K). Whether this switching behavior is functionally relevant remains unclear, as two karyolleles of *mnp1* have the same protein domain annotations in the P1 and P2 homokaryon genomes.

BROAD RANGE OF FUNCTIONALITY AFFECTED BY KARYOLLELE SPECIFIC EXPRESSION THROUGHOUT DEVELOPMENT

Next, we set out to examine the functional annotations of the differentially expressed karyollele pairs, considering the following categories: (i) transcription factors, (ii) metabolic genes, (iii) secondary metabolism genes, (iv) cytochrome P450 genes, (v) carbohydrate active enzymes (cazymes) and (vi) secreted proteins. These categories, with the exception of secondary metabolite genes, are all enriched in the set of differential genes ($p < 0.05$ by a chi-squared approximation to the fisher's exact test with FDR correction).

Figures 5.6a and 5.6b show the division of the 411 differentially expressed genes across the functional categories in all the different samples. None of the differentially expressed genes were transcription factors. For the other functional categories, we saw

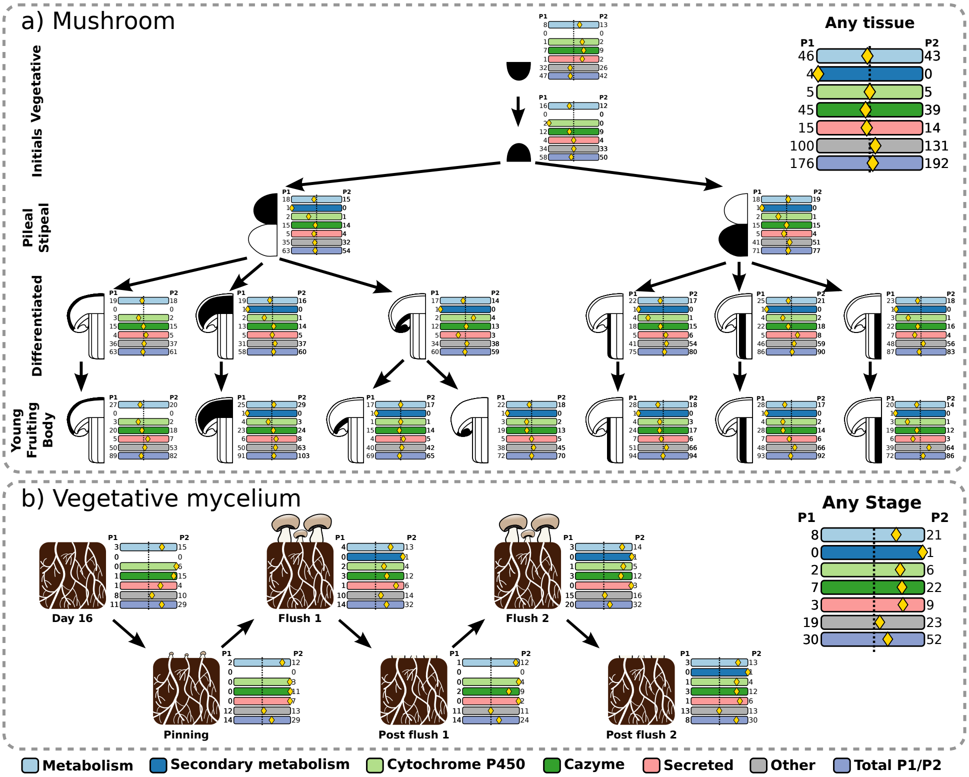


Figure 5.6: *Differential regulation of functional groups through mushroom development. We investigate metabolic genes (light blue), secondary metabolic genes (dark blue), cytochrome P450 genes (light green), carbohydrate active enzymes (dark green), secreted protein genes (light red), and all others not fitting into any of the previous groups (grey). At each developmental stage, we observe how many genes of each group are differentially upregulated in P1 (left) and in P2 (right). The yellow diamond indicates the ratio of these counts. (a) For the mushroom tissue dataset, the development of different mushroom tissues is illustrated as a tree. (b) Differential regulation of functional groups in compost mycelium.*

a more or less equal amount of up-regulated karyolleles in P1 and P2 (Figure 5.6a) in the mushroom tissues (except the vegetative stage), and a more skewed distribution of activity in the mycelium dataset (and the vegetative stage of the mushroom dataset). In these cases, P2 had more differentially expressed genes in these functional categories (Figure 5.6b).

The P2 type had a higher expression of significantly more karyolleles than P1 in mycelium (see Supplementary Material Note L). In the mycelium, P2 had an enriched expression of cytochrome P450 genes, secondary metabolite genes, and cazymes ($p < 0.05$, with an FDR corrected chi-squared approximation to the fisher's exact test). Furthermore, cazymes and metabolic genes in mycelium were more likely to be more highly expressed in P2 ($p < 0.05$, with an FDR corrected binomial test).

Nineteen of the 39 previously identified differentially expressed genes that are shared between the mycelium and mushroom datasets had the following functional annotations: 14 were annotated as metabolic genes, 14 as cazymes, five as secreted proteins, and two as cytochrome P450s (some genes have multiple annotations). Additionally, five of these 39 overlapping genes have different domain annotations, indicating different functional properties between the P1 and P2 karyolleles.

To further elucidate the functional impact of the 411 differentially expressed genes, we mapped them onto the KEGG pathway database. Sixteen of the genes that are differentially expressed in mushroom tissue or vegetative mycelium samples are found in 20 pathways. Interestingly, three differentially expressed genes are found in the Aminoacyl-tRNA biosynthesis (M00359) pathway (Supplementary Material Note M). Two genes belong to valine and methionine tRNAs pathways and were upregulated in P1. One gene in the pathway producing aspartamine tRNAs pathway was upregulated in P2. Together, this suggests that P1 is able to produce more valine and methionine tRNAs than P2.

Next we studied whether differential expression of a karyollele also resulted in the production of a functionally different protein due to sequence differences between the karyolleles. 216 of the 5,090 distinguishable karyolleles had sequence differences that led to an alternative protein domain annotation, and 36 of these 216 have alternative domain annotations. 36 of these 216 karyollele pairs are differentially expressed between P1 and P2 (see Supplementary Material Note N).

METHYLATION IMPLICATED IN KARYOLLELE SPECIFIC EXPRESSION

To investigate the biological mechanism causing differential expression, we measured methylation on the A15 genome. Assuming that the relative Cystosine/Thymine coverage at each base relates to a differential methylation state between the two nuclear types, we conclude that 277 genes are differentially methylated (Methods). 42 of these 277 genes were also found to be differentially expressed between the two nuclear types at some point in development. Although this is a significant proportion ($p < 0.05$, χ^2 test, Supplementary Material Note O), methylation only explains at most 10% of the differential expression we observe. Noteworthy is that 40 of the 42 differentially expressed and differentially methylated genes are differentially expressed in mushroom tissues (Supplementary Material Note O), whereas only 9 are differentially expressed in the vegetative mycelium. This indicates that the largest impact of differential methylation is much later in mushroom development, suggesting that methylation has a delayed effect on expression.

DISCUSSION

Differently from most eukaryotes, nuclei remain side by side during most of the life cycle of basidiomycete fungi. Whether each nucleus is contributing equally to the phenotype and, if not, how this is regulated is largely unknown. In an attempt to understand this, we studied the expression of alleles in both constituent nuclei (P1 and P2) of the button mushroom cultivar Sylvan 15. From the observed average gene expression, we conclude that the expression of nuclear type P1 of the *Agaricus bisporus* sylvan A15 strain is dominant over nuclear type P2. Remarkably, this dominance is present across all developmental stages in the heterokaryon. We can link this phenomenon to the human case, where in fibroblasts[29], it has been shown that individual cells preferentially express one allele over the other, which is not evident over a collection of many cells. Whereas in a diploid genome the cell must rely on heterochromatin DNA packing and RNAi regulatory pathways[30], heterokaryotic cells could instead control the energy usage of a specific nuclear type.

In the mushroom tissue dataset, the number of up-regulated karyolleles in P1 is approximately equal to those in P2, but in the vegetative mycelium dataset, P2 has more up-regulated karyolleles relative to P1. The contrast between a dominant P1, yet more differentially over-expressed genes in P2 in mushroom tissue is paradoxical. However, there are many genes that show a consistently higher expression in either P1 or P2, with more genes showing a consistently higher expression for P1. Is it possible that the P1 homokaryon is responsible for the basal mRNA production, while P2 plays a more reactive regulatory role? Mechanisms for this kind of regulation are not known. In plants, sub-genome dominance may be linked to methylation of transposable elements[24]. Might it be possible that something similar happens in *A. bisporus*?

Although an imbalance in the number of nuclei could very well explain the dominance of P1, we have shown that genes that are consistently higher expressed in one of the karyolleles do co-localize in sub-chromosomal regions. If there were more P1 nuclei than P2 nuclei, we would have expected a general higher expression of genes of one nuclear type across all chromosomes, which we do not observe.

For many differentially expressed genes, the protein sequence differences between the two karyolleles in the two nuclear types encode for different protein domains. This suggests a functional impact of karyollele specific expression. We also observe a broad range of functionality being differentially expressed between the P1 and the P2 nuclear types. For example, the P2 upregulation of cazymes and metabolic genes in P2 in compost highlight the importance of the P2 homokaryon in development. H97, one of the homokaryons in the cultivar Horst U1, from which Sylvan A15 is derived, displays stronger vegetative growth characteristics than its counterpart H39[1]. This metabolic strength may be passed down from the H97 homokaryon to the Sylvan A15 P2 homokaryon, and the differentially expressed karyolleles may in part be responsible for this. *mnp1*, for example, is an important gene for growth on compost and P2 has indeed inherited the relevant chromosome 2 from H97 [15]. Such characteristics are relevant for breeding strategies.

Surprisingly, *mnp1* is expressed and even up-regulated in the mushroom tissues. *mnp1* is known to be involved in lignin degradation, which occurs in the vegetative mycelium[2, 28]. In compost, the abundance decreases dramatically throughout development (Supplementary Material Note K). Therefore, the abundance of *mnp1* in the stipe of the fruiting body is unexpected, although it has been shown that proteins produced in the mycelium can find their way into the mushroom[31]. However, it does not explain the fact that the P1 karyollele exists in higher abundance in the mushroom tissues, while the P2 karyollele is higher expressed in the vegetative mycelium. Transport of the P2 karyollele from the vegetative mycelium into the mushroom conflicts with the abundances of the P1 karyollele observed in the mushroom tissues.

A significant proportion of differentially methylated karyolleles were also differentially expressed, most differentially expressed genes are not observed to be methylated. The overlap we observe between methylated genes and differentially expressed genes in different developmental stages explain an effect in the mushroom tissue. However, we cannot link the methylation to a preference of nuclear type. For example, the five differentially expressed genes between compost and mushroom that change their nuclear dominance are not methylated. Although, methylation seems to play a role in the differential use of nuclear type for mRNA production, it only explains 10% of the observed differential expression. This may be due to a limitation of our methylation dataset, (which only comprises vegetative growth), but it may also hint towards other regulatory mechanisms.

In addition to methylation, we also observe co-localization of co-expressed genes. This may be indicative of a difference in genome organization, whereby the DNA is less accessible in certain regions in P1 than in P2 through different levels of chromatin compaction. It has been shown that gene expression is strongly linked to DNA availability, and further, that such chromatin organization is heritable[32].

The sequences of a pair of karyolleles need to be sufficiently different for our algorithm to be able to uniquely assign reads to each karyollele. These sequence differences between nuclear types may have an effect on various regulatory mechanisms of transcription, such as transcription factor binding efficiencies, transcription efficiency, differences in mRNA stability, or differences in epigenetic factors. Future research might shed light on whether these differences are related to observed differential karyollele expression.

Causative mechanisms of karyollele specific expression can further be elucidated by population studies across multiple spore isolates. Sylvan A15 is derived as a heterokaryotic single spore isolate from Horst U1. In such heterokaryons, non-sister nuclei are paired in one spore. Combined with the restriction of recombination to chromosome ends, such heterokaryons are genetically very similar to the parent and differ only in the distribution of parental type chromosomes over both nuclei. Karyollele expression could thus be studied in different heterokaryotic single spore isolates having different distributions of otherwise very similar chromosomes over both nuclei. If the expression patterns are consistent with nuclear chromosome organization across different single spore isolates, it will suggest that expression of specific karyolleles can be controlled by selecting isolates where karyolleles lie in the desired nuclei.

CONCLUSION

We show that karyolleles, the different copies of a gene separated by nuclear membranes in a heterokaryon, are differentially expressed between the two different nuclear types in the *Agaricus bisporus* Sylvan A15 strain. Each nuclear type contributes varying amounts of mRNA to the cell, and differential expression occurs at the gene level. Despite a dominant P1 type, we see no evidence that would suggest an imbalance in the number of copies P1 and P2 nuclei in any cell type, though it may vary from cell to cell.

Genes with various vital functions are differentially expressed. The P2 homokaryon significantly up-regulates cazymes and metabolic genes, which may indicate a difference in vegetative growth strengths. This corroborates what was observed in the constituent homokaryons of the Horst U1 cultivar from which P1 and P2 are essentially derived.. Manganese peroxidase is one of the differentially expressed genes, and exhibits interesting, previously unknown behavior. The cause of these differential regulations is still not known, but it is possible that epigenetic mechanisms, like methylation, play a role.

The biological gene regulation mechanisms between heterokaryons need to be investigated. Unfortunately, such research is hindered by current mRNA isolation procedures. As mRNA transcripts are secreted from the nuclei and mixed in the cytoplasm of the cell, traditional sequencing methods will be unable to generate a full resolution of both homokaryon expression from full cell isolates. Single nucleus sequencing[33, 34] would circumnavigate this issue by isolating mRNA from individual nuclei. As we have shown that the two nuclear types exhibit distinguishable regulatory programs, it will be possible to distinguish them based on their expression profiles.

The impact of differential expression between nuclei of heterokaryotic organisms is underappreciated. Heterokaryotic fungi have major impact in clinical and biotechnological applications, and impact our economy and society as animal pathogens such as Cryptococcus neoformans[35], plant pathogens such as Ustilago maydis[36], plant and soil symbionts such as mycorrizal fungi[26], bioreactors such as Schizophyllum commune[37], and of course the subject of this study, the cultivated, edible mushroom *Agaricus bisporus* [15]. It is known that different homokaryons in these species will produce different phenotypes[2] which no doubt need to be treated, nourished or utilized differently.

We have demonstrated differential nuclear regulation of a fungal organism and we showed that variation between homokaryons results in functional differences that were previously unknown. With this work, we hope to draw attention to the impact of sequence and regulatory variation in different nuclei on the function and behavior of the cell in order to further our understanding of the role of fungi in our environment.

MATERIALS AND METHODS

RNA-Seq data: We used two RNA-seq datasets from the *Agaricus bisporus* (A15) strain: (a) tissue samples through mushroom development (BioProject: PRJNA309475)[27], and (b) vegetative mycelium samples taken from compost through mushroom development (BioProject PRJNA275107)[2]. Throughout the text, when we refer to the mushroom tissue, we also refer to all samples in dataset (a), including the first sample, which technically is a sample of the vegetative mycelium. The compost dataset exhibited high amounts of PCR duplicates (Supplementary Material Note P). This can be attributed to the difficulty in isolating RNA from soil. To remedy the biases involved with this, we removed all PCR duplicates using FastUniq[38].

Methylation data: A sample of vegetative stage mycelium of A15 was treated with the EpiTect Bisulphite conversion and cleanup kit and sequenced with the Illumina HiSeq 2000. Raw reads were trimmed using TRIMMOMATIC[39] and aligned to the A15 P1 genome using Bismark[40] and bowtie2[41]. Methylated bases were analyzed with Methylkit[42]. Only bases which had a minimum coverage of 10 were retained. For samples with mixed methylation states, we will observe what appear to be incomplete conversions of unmethylated cytosines but in reality represents the mixed methylation states of those bases. Therefore, to include only differentially methylated bases between the two nuclei (i.e. methylated in one homokaryon, but not in the other), we considered only those bases which were measured to be methylated between 40 and 60% of all reads (Supplementary Material Note O). While 164,290 bases had an indication of methylation signal, 10,325 bases had methylation signals of about 50%, suggestive of differential methylation states. Methylated bases were mapped to genes when between the start and stop codons, or 1000bp up/downstream (Supplementary Material Note Q).

Homokaryon genome and annotations: The P1 and P2 genomes[15] were annotated with BRAKER1[21] using the pooled RNA-seq data described above. In order to prevent chimeric genes (neighboring genes that are erroneously fused into one predicted gene) the following procedure was used. After the first round of gene prediction, predicted introns were identified that were at least 150 bp in size and not supported by RNA-seq reads. The midpoint of these introns were labeled as intergenic regions in the next round of gene prediction using AUGUSTUS 3.0.2[43] and the parameter set produced in the first round of gene prediction. The SNP density between the genomes was estimated using MUMMER's[44] show-snps tool.

Karyollele pair discovery: The genome annotations were used to produce predicted mRNA sequences for each gene. The genes in the two parental genomes were matched using a reciprocal best BLAST [45] hit. Hits which had E-values greater than 10-100 were removed. This resulted in a conservative orthology prediction between the two homokaryons that are our set of karyolleles. Karyollele pairs which have a 100% sequence identity were removed, as it would be impossible to identify distinguishing markers for these identical pairs.

Marker Discovery: For each discovered karyollele pair, we identify markers that uniquely identify each element of the pair. This is done by constructing all possible kmers for each sequence, resulting in two sets per pair. The kmers overlapping in these sets are removed, resulting in distinguishing pairs of markers. Once distinguishing markers have been discovered for all pairs, we remove all non-unique markers. Finally, the set of markers is made non redundant by scanning the position-sorted list of markers from left to right and removing any marker that overlaps with the previous marker. Finally, we ensure that the markers are unique throughout the whole genome by removing markers that are present anywhere else in either genome. In order to guarantee sufficient evidence across the whole gene, we remove karyollele pairs which do not have at least five markers each.

Marker quantification: We scan all RNA-Seq reads for the detected markers using the Aho-Corasick algorithm[46]. We insert all markers and their reverse complements into an Aho-Corasick tree and count each marker only once for each fragment (a marker may be present twice, if the read mates overlap). We calculate a gene expression score as the average of each marker count for a gene. This results in an expression score E_h for each gene g in each sample s for each replicate r, per homokaryon h (Equation 5.1):

$$E_h(r,s,g) = \frac{1}{|M_h(g)|} \sum_{m \in M_h(g)} C_h(r,s,m) \tag{5.1}$$

where $M_h(g)$ is the set of markers in a gene g, and $C_h(r,s,m)$ is the count for marker m in replicate r, sample s, for homokaryon h.

Differential expression: Using DE-Seq[47], we perform a differential expression test for each karyollele pair in a tissue, i.e. we test if a gene has a differential expression in P1 or P2. DESeq requires a size factor to be calculated, which normalizes for the library sizes of each sample. Since however, the counts from P1 and from P2 originate from the same sample, these must have the same size factor. Size factors are therefore calculated manually, by counting the total number of reads for each sample, and dividing it by the largest value for any sample (Equation 5.2).

$$sf(s,r) = \frac{\sum_{h} \sum_{m\in M_h(g)} C_h(r,s,m)}{\max_{(s',r')} \left(\sum_{h} \sum_{m\in M_h(g)} C_h(r',s',m) \right)} \tag{5.2}$$

The P1 and P2 counts originating from the same sample will then be assigned the same size factor. The expression counts for each gene in each replicate in each tissue (Equation 5.1) are provided to DE-Seq with the provided size factor (Equation 5.2). The normalized read counts per gene$D_h(s,g)$ are returned by DE-Seq, together with significance values for each test. We select only differentially expressed genes that have a q-value < 0.05, and a fold change of at least three.

Read ratio calculation: Using the normalized read counts from DE-Seq [47], we calculate the ratio of the number of reads originating from the two homokaryons at the gene (GRR), chromosome (CRR) and nuclear type level (NRR).

$$GRR(s,g) = \frac{D_{P1}(s,g)}{D_{P2}(s,g)} \tag{5.3}$$

$$CRR(s,c) = \frac{\sum_{g\in c} D_{P1}(s,g)}{\sum_{g\in c} D_{P2}(s,g)} \tag{5.4}$$

$$NRR(s) = \frac{\sum_{c\in C} \sum_{g\in c} D_{P1}(s,g)}{\sum_{c\in C} \sum_{g\in c} D_{P2}(s,g)} \tag{5.5}$$

Gene ratio calculation: Using the normalized read counts from DE-Seq [47], we calculate the ratio of the number of reads originating from the two homokarons at the gene level, and use those ratios to calculate the geometric mean of the relative expression activities at the chromosome (CGR, Equation 5.6) and nuclear type level (NGR, Equation 5.7). The geometric mean is more suitable than the arithmetic mean for averaging ratios.

$$CGR(s,c) = \sqrt[|c|]{\prod_{g\in c} GRR(s,g)} \tag{5.6}$$

$$NGR(s) = \sqrt[|C|]{\prod_{c\in C} CGR(s,c)}, \tag{5.7}$$

where c represents a set of genes on a specific chromosome, and C represents the set of all chromosomes.

Identifying consistent genes: For each gene, we observe the relative expression in each sample (Equation 5.3). We refer to a gene as being consistently expressed if it is more highly expressed in the same nuclear type in each sample. I.e. the GRR is always greater than one, or always less than 1.

Identifying co-regulated clusters: We slide a window of size 20,001bp (10,000- up and down-stream) across each chromosome. In this window, we count the number of genes that are more highly expressed by P1 and by P2, and calculate the difference per sample. I.e.

$$D(x,s) = \sum_{g \in W(x-10000, x+10000)} \begin{cases} 1 & \text{if } GRR(g,s) > 1 \\ -1 & \text{if } GRR(g,s) < 1 \end{cases} \tag{5.8}$$

, where W(x,y) is the set of genes between genomic location x and y, and s is a sample. This difference is shown in Figure 5.5. Next, we identify regions where each sample in the dataset shows consistent regulation. That is to say, in these regions, $D(x,s) > 0 \forall s \in S$, or $D(x,s) < 0 \forall s \in S$, where S is the set of all samples. These regions contain co-localized genes that are co-regulated across all samples.

Functional predictions:

PFAM: Conserved protein domains were predicted using PFAM version 27[48, 49]. Transcription factor definitions: Predicted proteins with a known transcription factor-related (DNA-binding) domain (based on the PFAM annotations) were considered to be transcription factors.

Carbohydrate-active enzymes prediction: Using the Cazymes Analysis Toolkit (CAT) [50], we predicted carbohydrate-active enzymes based on the original gene definitions. If a gene's protein sequence was predicted to be a cazyme by either the sequence-based annotation method or the PFAM-based annotation method then we considered it a cazyme.

Secreted Proteins prediction: We used the same procedure as [51] to predict secreted proteins. Briefly, genes with SignalP [52] signal peptides, or a TargetP [53] Loc=S were kept. The remaining genes were further filtered with TMHMM [54], keeping only genes with zero or one transmembrane domains. Finally, genes were filtered using Wolf PSort [55] to select genes with a Wolf PSort extracellular score greater than 17.

Metabolic and Cytochrome P450 gene groups: Genes with the GO annotation "metabolic process" (annotation ID: GO:0008152) were called as metabolism genes. Genes with the PFAM annotation PF00067 were used as Cytochrome P450 genes.

KEGG: KEGG annotations were made with the KAAS KEGG [56] annotation pipeline, using genes from all available fungi, with the exception of leotiomycetes, Dothideomycetes, and Microsporidians, due to the limitation of the number of species (Selected organisms by ID: cne, cgi, ppl, mpr, scm, uma, mgl, sce, ago, kla, vpo, zro, cgr, ncs, tpf, ppa, dha, pic, pgu, lel, cal, yli, clu, ncr, mgr, fgr, nhe, maw, ani, afm, aor, ang, nfi, pcs, cim, cpw, pbl, ure, spo, tml). The GHOSTX and BBH options were selected. Predictions were made individually for both the P1 and P2 genomes, using the translated protein sequences.

Named genes: Named genes for *Agaricus bisporus* version 2 were downloaded from the JGI DOE Genome Portal (http://genome.jgi.doe.gov/pages/search-for-genes.jsf?organism=Agabi_varbisH97_2) by searching for genes with 'Name' in the 'user annotations' attribute. Gene names were transferred from *A. bisporus* v. 2 using reciprocal best blast hit to P1 and P2, and then selecting the best match (in the single case of an ambiguity). See Supplementary Material Note R.

Software and code availability: Marker discovery and abundance calculations was done in Scala, while downstream analysis was performed in python using the ibidas data query and manipulation suite [57]. All source code, together with a small artificial example dataset is available at: https://github.com/thiesgehrmann/Homokaryon-Expression

Data Availability: The RNA-Seq data was previously generated and can be found at bioprojects PRJNA309475 and PRJNA275107. The bisulphite sequencing data can be accessed at SAMN06284058.

Supplementary information: Together with this manuscript, we provide a file of Supplementary Notes, and Supplementary Tables 1-4 to support our findings.

BIBLIOGRAPHY

[1] E. Morin, et al. Genome sequence of the button mushroom agaricus bisporus reveals mechanisms governing adaptation to a humic-rich ecological niche. *Proceedings of the National Academy of Sciences* 109(43):17501–17506 (2012).

[2] A. Patyshakuliyeva, et al. Uncovering the abilities of agaricus bisporus to degrade plant biomass throughout its life cycle. *Environmental Microbiology* 17(8):3098–3109 (2015).

[3] R. a. Ohm, et al. Genome sequence of the model mushroom schizophyllum commune. *Nature biotechnology* 28(9):957–63 (2010).

[4] T. E. Pawlowska. Genetic processes in arbuscular mycorrhizal fungi. *FEMS Microbiology Letters* 251(2):185–192 (2005).

[5] M. ud din Khanday, et al. Arbuscular mycorrhizal fungi boon for plant nutrition and soil health. In *Soil Science: Agricultural and Environmental Prospectives*, 317–332. Springer International Publishing, Cham (2016).

[6] C. Sun, et al. The beneficial fungus piriformospora indica protects arabidopsis from verticillium dahliae infection by downregulation plant defense responses. *BMC Plant Biology* 14:268 (2014).

[7] B. D. Harrach, H. Baltruschat, B. Barna, J. Fodor, K.-H. Kogel. The mutualistic fungus piriformospora indica protects barley roots from a loss of antioxidant capacity caused by the necrotrophic pathogen fusarium culmorum. *Molecular plant-microbe interactions : MPMI* 26(5):599–605 (2013).

[8] Z. Babikova, et al. Underground signals carried through common mycelial networks warn neighbouring plants of aphid attack. *Ecology Letters* n/a–n/a (2013).

[9] C. Collins, et al. Genomic and proteomic dissection of the ubiquitous plant pathogen, armillaria mellea: Toward a new infection model system. *Journal of Proteome Research* 12(6):2552–2570 (2013).

[10] S. Khoshraftar, et al. Sequencing and annotation of the ophiostoma ulmi genome. *BMC genomics* 14(1):162 (2013).

[11] L. Guo, et al. Genome and transcriptome analysis of the fungal pathogen fusarium oxysporum f. sp. cubense causing banana vascular wilt disease. *PLoS ONE* 9(4) (2014).

[12] C. A. Specht. Isolation of the ba and bb mating-type loci of schizophyllum commune. *Current Genetics* 28(4):374–379 (1995).

[13] K. N. Saksena, R. Marino, M. N. Haller, P. a. Lemke. Study on development of Agaricus bisporus by fluorescent microscopy and scanning electron microscopy. *Journal of bacteriology* 126(1):417–28 (1976).

[14] G. D. Craig, R. J. Newsam, K. Gull, D. A. Wood. An ultrastructural and autoradiographic study of stipe elongation inagaricus bisporus. *Protoplasma* 98(1-2):15–29 (1979).

[15] A. S. M. Sonnenberg, et al. A detailed analysis of the recombination landscape of the button mushroom agaricus bisporus var. bisporus. *Fungal Genetics and Biology* 93:35–45 (2016).

[16] R. B. Todd, M. a. Davis, M. J. Hynes. Genetic manipulation of aspergillus nidulans: heterokaryons and diploids for dominance, complementation and haploidization analyses. *Nature protocols* 2(4):822–830 (2007).

[17] E. Boon, E. Zimmerman, B. F. Lang, M. Hijri. Intra-isolate genome variation in arbuscular mycorrhizal fungi persists in the transcriptome. *Journal of Evolutionary Biology* 23(7):1519–1527 (2010).

[18] T. Y. James, J. Stenlid, K. Olson, H. Johannesson. Evolutionary significance of imbalanced nuclear ratios within heterokaryons of the basidiomycete fungus heterobasidion parviporum. *Evolution* 62(9):2279–2296 (2008).

[19] D. Muzzey, G. Sherlock, J. S. Weissman. Extensive and coordinated control of allele-specific expression by both transcription and translation in candida albicans. *Genome Research* 24(6):963–973 (2014).

[20] X. Wei, X. Wang. A computational workflow to identify allele-specific expression and epigenetic modification in maize. *Genomics, proteomics & bioinformatics* 11(4):247–52 (2013).

[21] J. J. Crowley, et al. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature Genetics* 47(4):353–360 (2015).

[22] P. R. Buckland. Allele-specific gene expression differences in humans. *Human Molecular Genetics* 13(REV. ISS. 2):255–260 (2004).

[23] P. V. K. Pant, et al. Analysis of allelic differential expression in human white blood cells. *Genome Research* 16(3):331–339 (2006).

[24] P. P. Edger, R. Smith, M. R. Mckain, A. M. Cooley, M. Vallejo-marin. Subgenome dominance in an interspecific hybrid , synthetic allopolyploid , and a 140 year old naturally established neo-allopolyploid monkeyflower. *bioRxiv* 1–27 (2016).

[25] T. R. Horton. The number of nuclei in basidiospores of 63 species of ectomycorrhizal homobasidiomycetes. *Mycologia* 98(2):233–238 (2006).

[26] K. Lin, et al. Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. *PLoS Genetics* 10(1) (2014).

[27] J. F. Pelkmans, et al. The transcriptional regulator c2h2 accelerates mushroom formation in agaricus bisporus. *Applied Microbiology and Biotechnology* 2 (2016).

[28] A. M. Bonnen, L. H. Anton, A. B. Orth. Lignin-degrading enzymes of the commercial button mushroom, agaricus bisporus. *Applied and environmental microbiology* 60(3):960–5 (1994).

[29] C. Borel, et al. Biased allelic expression in human primary fibroblast single cells. *American Journal of Human Genetics* 96(1):70–80 (2015).

[30] T. A. Volpe, et al. Regulation of heterochromatic silencing and histone h3 lysine-9 methylation by rnai. *Science (New York, NY)* 297(5588):1833–7 (2002).

[31] B. M. Woolston, et al. Long-distance translocation of protein during morphogenesis of the fruiting body in the filamentous fungus, agaricus bisporus. *PLoS ONE* 6(12) (2011).

[32] R. McDaniell, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* 328(5975):235–239 (2010).

[33] B. B. Lake, et al. Neuronal subtypes and diversity revealed by single-nucleus rna sequencing of the human brain. *Science* 352(6293):1586–1590 (2016).

[34] S. R. Krishnaswami, et al. Using single nuclei for rna-seq to capture the transcriptome of postmortem neurons. *Nature protocols* 11(3):499–524 (2016).

[35] B. J. Loftus, et al. The genome of the basidiomycetous yeast and human pathogen cryptococcus neoformans tl - 307. *Science* 307 VN -(5713):1321–1324 (2005).

[36] J. Kämper, et al. Insights from the genome of the biotrophic fungal plant pathogen ustilago maydis. *Nature* 444(7115):97–101 (2006).

[37] C. H. Shu, H. J. Hsu. Production of schizophyllan glucan by schizophyllum commune atcc 38548 from detoxificated hydrolysate of rice hull. *Journal of the Taiwan Institute of Chemical Engineers* 42(3):387–393 (2011).

[38] H. Xu, et al. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLoS ONE* 7(12):1–6 (2012).

[39] A. M. Bolger, M. Lohse, B. Usadel. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* 30(15):2114–2120 (2014).

[40] F. Krueger, S. R. Andrews. Bismark: A flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* 27(11):1571–1572 (2011).

[41] B. Langmead, S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods* 9(4):357–359 (2012).

[42] A. Akalin, et al. methylkit: a comprehensive r package for the analysis of genome-wide dna methylation profiles. *Genome biology* 13(10):R87 (2012).

[43] M. Stanke, M. Diekhans, R. Baertsch, D. Haussler. Using native and syntenically mapped cdna alignments to improve de novo gene finding. *Bioinformatics* 24(5):637–644 (2008).

[44] S. Kurtz, et al. Versatile and open software for comparing large genomes. *Genome biology* 5(2):R12 (2004).

[45] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology* 215(3):403–10 (1990).

[46] A. V. Aho, M. J. Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM* 18(6):333–340 (1975).

[47] S. Anders, et al. Differential expression analysis for sequence count data. *Genome Biology* 11(10):R106 (2010).

[48] R. D. Finn, et al. The pfam protein families database. *Nucleic acids research* 36(Database issue):D281–D288 (2008).

[49] R. D. Finn, et al. Pfam: The protein families database. *Nucleic Acids Research* 42(D1):222–230 (2014).

[50] V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, B. Henrissat. The carbohydrate-active enzymes database (cazy) in 2013. *Nucleic Acids Research* 42(D1):D490–D495 (2014).

[51] A. Morais do Amaral, J. Antoniw, J. J. Rudd, K. E. Hammond-Kosack. Defining the predicted protein secretome of the fungal wheat leaf pathogen mycosphaerella graminicola. *PLoS ONE* 7(12):1–19 (2012).

[52] T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen. Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* 8(10):785–786 (2011).

[53] O. Emanuelsson, H. Nielsen, S. Brunak, G. von Heijne. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *Journal of Molecular Biology* 300(4):1005–1016 (2000).

[54] A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology* 305(3):567–580 (2001).

[55] P. Horton, et al. Wolf psort: protein localization predictor. *Nucleic Acids Research* 35(Web Server):W585–W587 (2007).

[56] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, M. Kanehisa. Kaas: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* 35(SUPPL.2):182–185 (2007).

[57] M. Hulsman, J. J. Bot, A. P. de Vries, M. J. T. Reinders. Ibidas: Querying flexible data structures to explore heterogeneous bioinformatics data. *Data Integration in the Life Sciences* 23–37 (2013).

Chapter 6

Variants in RNA-Seq data show a continued mutation rate during strain preservation of *Schizophyllum commune*

Thies Gehrmann
Jordi F. Pelkmans
Luis G. Lugones
Han A. B. Wösten
Thomas Abeel
Marcel J. T. Reinders

A preprint of this chapter has been published as: *bioR$_\chi$ iv*, October 2017 (doi: 10.1101/201012)
Supplementary material is available online at: `http://bit.ly/2FTWyBl` (biorxiv.org)

ABSTRACT

Background: *Typical microorganism studies link genetic markers to physiological observations, like growth and survival. Experiments are carefully designed, comparing, for example, wildtype strains with knockout strains, and replications are conducted to capture biological variation. To maintain monoclonal strains, strain preservation systems are used to keep the number of generations between the primary stock and the experimental measurement low, to decrease the influence of spontaneous mutations on the experimental outcome. The impact of spontaneous mutations during the minimal number of growth cycles for the experimental design is, however, poorly understood.*

Results: *We set out to characterize the mutation landscape for a transcriptome dataset of an experiment with Schizophyllum commune, a model for mushroom formation. We designed a methodology to detect SNPs from the RNA-seq data, and found a mutation rate of* 1.923×10^{-8} *per haploid genome per base per generation, highly similar to the previously described mutation rate of S. commune in the wild. Knock-outs did not influence the mutation rate considerably and chromosomal recombination occurring at mating type loci was frequent. We found that missense and nonsense SNPs were selected against throughout the experiment. Also, most mutations show a low variant allele frequency and appear only in a small part of the population. Yet, we found 40 genes that gained a nonsense mutation affecting one of its annotated protein domains, and more than 400 genes having a missense mutation inside an annotated protein domain. Further, we found transcription factors, metabolic genes and cazymes having gained a mutation. Hence, the mutation landscape is wide-spread and has many functional annotations.*

Conclusions: *We have shown that spontaneous mutations accumulate in typical microorganism experiments, where one usually assumes that these do not to happen. As these mutations possibly confound experiments they should be minimized as much as possible, or, at least, be trackable. Therefore, we recommend labs to implement a sample tracking system, and to ensure that biological replicates originate from different parental plates, as much as possible.*

INTRODUCTION

Experiments with microorganisms rely on the assumption that the organisms used in independent experiments are identical to ensure that differences in phenotypic characteristics are not the result of an underlying genetic heterogeneity. In reality, spontaneous mutations are regularly acquired during cell division[1], invalidating this assumption. These spontaneous mutations represent confounding factors in the original experiments [2], as well as for independent replication experiments. These confounders usually go unnoticed, or are disregarded. However, a change of a specific phenotypic trait can reveal the occurrence of spontaneous mutations. For example, Saccharomyces cerevisiae is often plagued by the petite phenotype, caused by deletion of mitochondrial DNA[3, 4]. As another example, the mutation rate of 2.0 × 10⁻8 per base per haploid genome per generation[5] of *Schizophyllum commune* (model wood rot mushroom) regularly interferes with biological experiments[6], and consequently several mutations frequently occur in laboratory settings: the thn mutation prevents aerial hyphae formation[7], the streak mutant results in a blue color[8], and the fbf mutant prevents

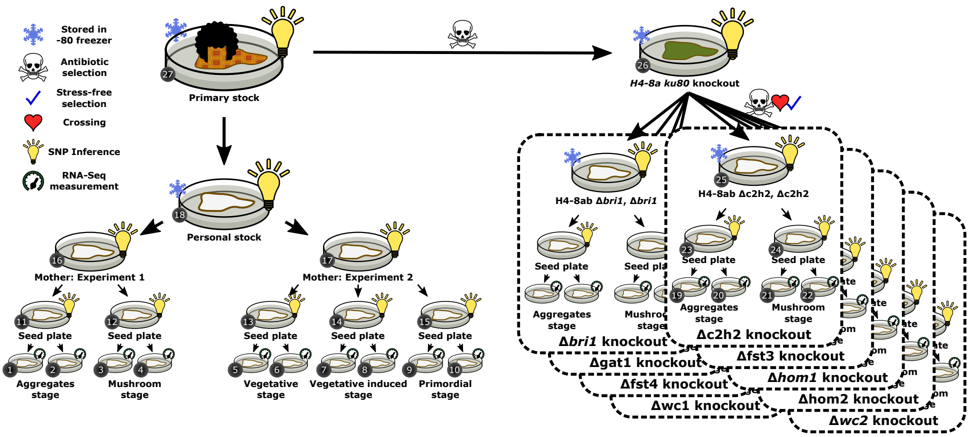


Figure 6.1: *The experimental design of our S. commune experiment. The experimental design is heavily influenced by the strain preservation system, and takes the form of a tree. The labelled nodes (1-27) are the same as those shown in 6.2. RNA-Seq data is derived from the leaves of the tree (indicated by the measurement icon), while SNPs are inferred for the internal nodes of the tree. All nodes are samples that are originally derived from the -80 freezer (node 27), including the ku80 knockout (node 26), from which all knockout samples are derived. Initially, a researcher took a sample from the original stock, and created his own personal stock (node 18). From this, two experimental runs were conducted (nodes 16 and 18, Materials).*

mushroom formation[9].

To prevent these commonly occurring mutations, and mutations in general, from seeping into other experiments, labs utilize a strain preservation system (Supplementary Note 1). In general, a strain preservation system attempts to minimize the number of generations between the primary stock and the strains from which measurements or materials are eventually sampled. This involves ensuring the long-term preservation of the primary stock of a given strain. Each lab worker creates their own personal stock, subcultured from the primary stock, and samples exclusively from this stock to perform experiments. For each experiment, a ‘mother plate’ is subcultured from the personal stock, from which all subsequent measurements are made. If a personal stock is depleted, it is recreated from the primary stock. This procedure is followed for each strain, including mutant strains derived from the primary stock. To ensure statistical rigor, experiments are replicated, often seeded simultaneously from the same mother plate to reduce human error. Although quite some mitosis steps take place in such experiments, it is presumed that there are no genetic alterations with respect to the original strain except for the intentionally introduced mutations. But, it is not clear how spontaneous mutations impact phenotypic or transcriptomic differences.

We set out to capture and characterize the mutations acquired during a microorganism experiment. We chose the *Schizophyllum commune* mushroom because this species suffers frequently from spontaneous mutations that change the phenotype of the strain. We use a near-isogenic dikaryonic strain of *S. commune* (H4-8), meaning that each hyphal compartment contains two nuclei. These nuclei have been backcrossed together in such a way that their genomic material differs only in their mating type loci (Materials, Supplementary Note 2). As *S. commune* grows linearly outwards, with cell division occurring only at the hyphal tips, a mutation gained at any stage of this growth will naturally be passed along to its decendants.

Our dataset is composed of 46 RNA-Seq measurements (see Materials and Methods) from the dikaryonic wildtype and knockout strains of several genes involved in mushroom fructification (BRI1,FST3,FST4,HOM1,HOM2,GAT1,WC-1,WC-2, and C2H2). The heredity of these 46 samples is defined in the sample tree shown in 6.1. The wild type has been sampled at five different developmental stages, across two different sequencing runs (aggregates and mushroom in the first, vegetative, vegetative induced and primordia in the second). The knockout strains were sampled at two different developmental stages (aggregates and mushroom, see Materials and Methods, Supplementary Note 3). All knockout strains are derived from an additional knockout strain, in which the KU80 gene is knocked out (node 26, 6.1, Supplementary Note 4) to repress chromosomal repair for which KU80 is responsible[10]. We studied the accumulation of spontaneous genomic mutations throughout an experimental design encompassing wildtype and knockout strains in several developmental stages through the transcriptome (6.1, Materials) [11, 12].

RESULTS

SNPS CAN BE IDENTIFIED FROM RNA-SEQ DATA.

To characterize the mutation landscape in various steps of an experimental design, we developed a method with which single nucleotide polymorphisms (SNPs) can be identified from the transcriptome (Methods). As the genome of *S. commune* is very dense, and neighboring genes have overlapping UTRs, the transcriptome spans 89% of the genome[11]. This not only permits us to identify mutations that accumulated during culturing of the strain samples across a large portion of the genome but also gives us the ability to study their phenotypic effect across various growth conditions. Using the 'infinite sites assumption'[13], which assumes that mutations are only gained once at novel loci and never lost, we are able to associate individual SNPs to intermediate steps in our experimental setup (6.1). We sampled the H4-8 *S. commune* strain and nine derived knock-out strains across a total of five developmental stages (Materials). RNA was sequenced with an average coverage of 100X per sample (Materials). Using a method which leverages the lineage information in our experimental design, we identified SNPs in our RNA-Seq data (Methods).

We identified 13,249 SNPs across all our samples (6.1). Depending on the read depth, we detected 1413 SNPs on average per sample (Supplementary Note 5). 94% of SNPs are heterozygous. 43 SNPs mapped to mating type regions (Materials, Supplementary Note 2). The majority (71%) of SNPs lie in gene coding regions, 27% lie in intergenic regions, and 2% lie within intron regions. We are able to capture intergenic and intronic SNPs due to the high gene density of *S. commune*, and the overlapping UTRs of transcripts. Most (92%) are present at lower abundances than the reference base, resulting in low Variant Allele Frequencies (VAF) (Supplementary Note 6). However, care should be taken to interpret the VAF, as in this case it represents a non-deconvolvable expression of subcolonal populations and nuclear specific expression. SNPs present in a large number of samples also mapped to genes that have a high RNA expression level (Supplementary

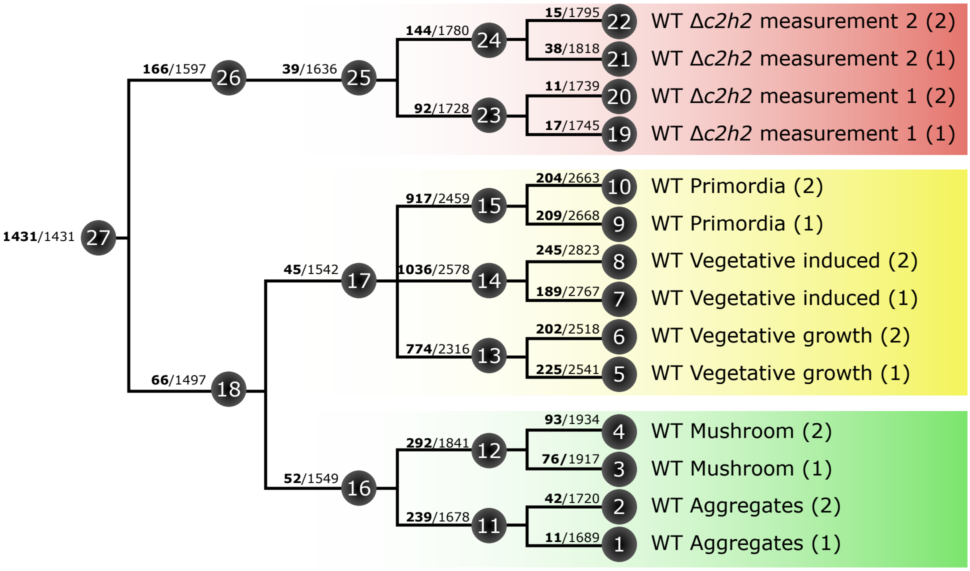


Figure 6.2: *SNPs identified at each plate in the sample tree. The numbered nodes in the tree are the same as those numbered in Figure 6.1. The bold number refers to the SNPs which are gained on that plate. The non-bold number indicates the total number of SNPs observed on that sample (cumulative from root). The full tree can be seen in supplementary note P.*

Note 7). 9% of the SNPs (1252) lie in predicted alternatively spliced genes, based on the transcripts in[11].

SNP ORIGINS REVEAL COMPOUNDING SPONTANEOUS MUTATIONS.

To investigate the accumulation of mutations throughout the experimental design, we identified the origin of SNPs in the sample tree. At each stage in the sample tree, mutations are gained with respect to the previous stage (6.2, see Supplementary Note 9 for the full tree). Most SNPs (85%) whose origin is identified at an internal node in the tree are supported by all its child samples (Methods), and only a small proportion (15%) are supported by a subset of child samples. As most variants match with the sample tree, it supports that the SNPs are DNA mutations rather than RNA editing substitutions. 11% (1431) of the detected SNPs are predicted to be present in the primary stock of H4-8. The remaining 11,818 SNPs are gained at some point in the experimental procedure. Of the 783 homozygous SNPs, 85% (663) are associated to the primary stock. 7% (52) homozygous SNPs have been introduced in the ku80 knockout. The remaining 68 heterozygous SNPs are scattered throughout the experimental design, which, although unexpected due to the infinite sites assumption, may be the result of allele specific expression or silencing.

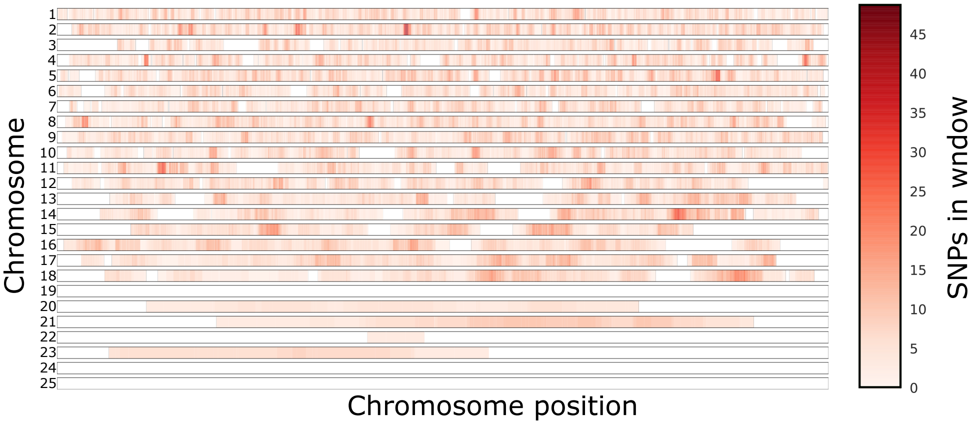


Figure 6.3: *SNP hotspots across the genome. We move a sliding window across the genome (Methods), counting the SNPs in that window. There are certain peaks, several of which coincide with mating type loci (chromosomes 2 and 11).*

ESTIMATED MUTATION RATE IS SIMILAR TO THE NATURAL MUTATION RATE.

To determine if the mutation rate in the laboratory, where evolutionary stresses are removed, is different from the natural environment, we calculated a mutation rate from the observed mutations. Each plate represents approximately 200 cell divisions, and thereby 200 dikaryotic DNA duplications (Methods, Supplementary Note 10). From this, and given the 13,249 SNPs we identified, we estimated the mutation rate to be 1.9233×10^{-8} (95% CI: 1.3899×10^{-8}, 2.4568×10^{-8}) per haploid genome per base per generation (Methods). This estimated mutation rate, in a strain preservation system, is almost identical to the mutation rate known for wild *S. commune* strains[5]. The mutation rate varies across the genome (6.3), with 27 non-overlapping loci of 20,000bp containing more than 20 SNPs. In these regions, we observe a mutation rate of 1.652×10^{-8} (95% CI: 6.349×10^{-9}, 2.670×10^{-8}) per haploid genome per base per generation. The mating type loci exhibit a slightly (but not significantly) lower mutation rate of 1.652×10^{-8} (95% CI: $6.349 \times 10 - 9, 2.670 \times 10^{-8}$), and consequently are not enriched for or depleted of mutations (Supplementary Note 11).

S. commune SELECTS AGAINST NONSENSE AND MISSENSE MUTATIONS.

In order to characterize the functional impact of these SNPs, we examine the impact of these SNPs on the coding sequence. Excluding the SNPs present in the original sample, 71% (8820) of the SNPs that are gained at some point in the experimental procedure lie in coding regions (between start and stop-codons) of predicted genes. 48% (4234) are synonymous mutations, 50% (4769) are missense mutations, and 2% (205) are nonsense mutations. This is significantly more synonymous, less missense and less nonsense mutations than expected by random chance (p-value < 0.05, χ^2-test) when taking into account the codon usage of *S. commune* (Supplementary Note 12). This indicates that deleterious mutations are still under negative selection in the population.

Table 6.1: *Number of genes with SNPs in any part of a gene in the different functional groups. The SNPs are split into Synonymous, Missense and Nonsense mutations. The total row indicates any functionally annotated gene.*

	Synonymous	**Missense**	**Nonsense**
Transcription Factors	128	132	13
Cytochrome P450s	23	21	3
Metabolic proteins	177	183	13
Cazymes	525	542	30
Total	2860	2934	**202**

Table 6.2: *Number of genes with SNPs before the end of the last protein domain. The total row indicates all genes with a functional annotation.*

	Synonymous	**Missense**	**Nonsense**
Transcription Factors	31	33	1
Cytochrome P450s	8	5	1
Metabolic proteins	55	77	2
Cazymes	176	166	9
Total	**575**	**574**	**40**

Table 6.1 shows that the SNPs in coding regions occur without specific enrichments across the functional groups "transcription factors", "cytochrome P450s", "metabolic proteins" and "carbohydrate active enzymes (cazymes)", even for the nonsense mutations. Examining the functional groups in which these mutations lie can help us understand the impact of these SNPs on the functionality of the genes, we investigate the location of the mutation in the gene relative to the annotated protein domains. 11% (967) of the SNPs in coding regions are located within predicted domain regions of 737 genes. 47% (453) are synonymous, 50% (486) are missense, and 3% (28) are nonsense mutations. No transcription factors have mutations in their binding domains (see Table 6.3). 16% (1,369) of the SNPs in coding regions are located before the end of the last annotated protein domain, also potentially altering the functional components of the protein. They lie within 1,023 genes, are annotated across various function groups (Table 6.2), and are similarly distributed across synonymous 48% (661), missense 49% (668), and nonsense 3% (40) mutations as all detected SNPs. As an example, gene G2683529 is a predicted transcription factor with a nonsense mutation upstream of its DNA binding domain. The SNP is gained in the seed plate of the first-time measurement of the Hom2 knockout. The variant has a low VAF (< 4%), and is unlikely to severely impact colony behavior due to its low expression and prevalence in the population.

SPONTANEOUS SNPS MAY INFLUENCE GENE EXPRESSION.

Next, we set out to study the impact of detected mutations on the RNA expression level of neighboring genes. There are no SNPs within gene coding regions that do influence gene expression (FDR corrected p-value > 0.05, two sample t-test with independent variance assumption between expression of samples with and without SNP). On the other hand,

Table 6.3: *Number of genes with SNPs in protein domains. The total row indicates all genes with a functional annotation.*

	Synonymous	Missense	Nonsense
Transcription Factors	11	12	0
Cytochrome P450s	7	5	1
Metabolic proteins	47	71	2
Cazymes	128	130	7
Total	**394**	**411**	**28**

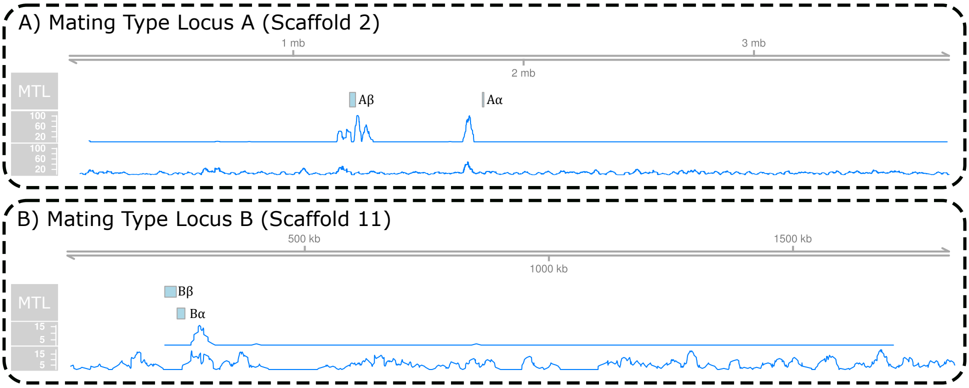


Figure 6.4: *SNP hotspots around the mating type loci MatA on chromosome 2 (A) and MatB on chromosome 11 (B). The top line indicates a sliding window of 10,000bp of SNPs whose origin was at the root node of the sample tree, while the second line indicates the SNPs whose origin was anywhere except the root node of the sample tree.*

this is different for SNPs in (predicted) promoter regions of genes (Methods). 2% (36) of the 1,995 SNPs found in promoter regions of genes showed a significant difference in the RNA expression level of the associated genes (FDR corrected p-value < 0.01). 29 of these genes have been observed to be differentially expressed between different developmental groups (Supplementary Note 13), so that the difference in expression level might not be caused by the SNP but is a result of development. For the other seven genes there was no alternative explanation. Four of them have a lower expression when the SNP is present, and three a higher. One encodes for protein ID 2539542, a (predicted) transcription factor. The SNP appeared for the first time in the aggregate stage seed plate for the gat1 knockout and has a variant allele frequency of 11%. The observed difference between the means of the RNA expression levels is 13% with the gene having a higher RNA expression level when the SNP is present.

DURING KARYOGAMY, *S. commune* PERFORMS FREQUENT CHROMOSOMAL CROSSOVER.

SNPs identified at the root node in the strain tree allow us to investigate recombination in *S. commune* (see Materials). The 1,431 SNPs predicted to be associated with the primary stock are distributed more or less evenly over the genome. Scaffold 2 and 11 show hotspots (Supplementary Note 14) that map to the A and B mating type loci (6.4). Their exceedingly high number of mutations suggest that these regions are at the breakpoints of a chromosomal recombination. This is supported by the observation that neighboring regions are depleted of mutations, indicative of the isogenic nature of the H4-8 strain. To explain the observed recombination events at these regions, a minimum number of six chromosomal recombinations are required, amongst which one should have happened upstream, and one downstream of each mating type locus (four recombinations for the MatA loci, and two recombinations for the MatB loci). Having at least six chromosomal recombinations at these specific loci within seven backcrosses is indicative of a high chromosomal recombination frequency. We do find a few similar hotspots for the knock-out samples (Supplementary Note 15).

THE GENE-KNOCKOUT IN *S. commune* PROCEDURE DOES NOT INDUCE SNP MUTATIONS.

We were wondering whether the stress conditions during the knockout procedure (see Materials) would introduce an increased load of mutations. Between the wildtype and the ku80 knockout, we found 166 SNPs. This is not significantly more than at any other reproductive step (node) in the tree, even after correcting for read depth (p-value > 0.05, 1 sample t-test). On the other hand, between the ku80 knockout and the derived dikaryotic knockouts, we find significantly fewer mutations SNPs when compared to the other reproductive steps (p-value < 0.05, 2 sample t-test with unequal variance). This indicates that the stresses of a gene knockout do not induce further spontaneous mutations.

DISCUSSION

The mutation rate we observe is similar to the previously reported mutation rate identified in in wild *S. commune* strains. From this observation, it is obvious that a strain preservation system does not protect against (or lower the number of) spontaneous mutations. Consequently, even in such a laboratory setting, spontaneous mutations will confound experiments, underpinning the necessity for replication experiments. However, we do need to be careful in designing the setup of these replication experiments. If we would replicate from the same parental seed plate, similar spontaneous mutations can confound results. Hence, replications should preferably be done from different seed plates.

We observed a similar mutation rate in the mating type loci as in the rest of the genome. James [14] argued that mushrooms have evolved a high outcrossing efficiency to ensure and encourage diversity in the population. *S. commune* has an extremely high outcrossing efficiency[15]. Our observation, which implies that the mating type loci are not protected from mutations, could point towards a biological mechanism that drives the high outcrossing efficiency in mushrooms. That is, the mating type loci mutations may create additional mating types. In theory, this could make it possible for hyphal anastomosis within a monokaryotic colony to result in a compatible, fertile dikaryon.

We identified a large number of SNPs originating in the primary stock (root node of the sample tree) near the mating type loci. These are indicative of recombination sites. Not having linkage information complicated the calculation of a recombination rate for

S. commune. Based on the mutation pattern near the mating loci, we expect a high chromosomal recombination frequency for *S. commune*. Previously it has been shown that *S. commune* performs crossover at regions of high homology[16], and we see that *S. commune* also recombined very closely to mating type regions. A. bisporus[17] (for which *S. commune* serves as a model for mushroom formation) performs crossover only near the telomeric regions[18]. Given the seemingly alternative crossover mechanisms, we suggest caution when comparing evolutionary mechanisms between *S. commune* and A. bisporus.

The ku80 knock-out strain derived from the primary stock showed a similar number of mutations as seen for the other derived strains. Initially we were expecting a higher number of mutations due to stress (e.g. passenger mutations by selection with antibiotic resistance) induced by creating the knock-out. This might not have occurred because ku80 is involved in the DNA repair mechanism for double strand breaks. Hence, the absence of ku80 might not have an impact on single nucleotide polymorphisms, but rather induce structural variations such as indels or inversions. Our method to detect variations of DNA from RNA sequence data was not designed to detect these larger variations. For the knock-out strains that are derived from the ku80 strain, we observed a lower mutation rate. This might be the result of crossover during the backcrossing with the primary stock wildtype to restore the ku80 gene. This would remove some mutations that were gained on one allele but not the other. The backcrossing with the primary stock wildtype might also explain the relatively high number of homozygous SNPs in the ku80 knockout strain.

As we derived mutations from RNA sequencing data, we were able to relate detected spontaneous mutations to changes in RNA expression levels (in the same samples). We found no SNPs in coding regions that influence RNA expression of the corresponding gene, and only a handful of SNPs in predicted promotor regions were associated to changes in expression levels of the corresponding genes. This might suggest that regulatory regions are vulnerable to evolutionary drift, especially in intermediate plates which never complete their life cycle and where a large part of the functional repertoire of the organism is not utilized. This effect might even be larger, since we only used the simplistic rule of associating a regulatory SNP to a gene via its upstream promotor region. Enhancers and promotors, however, lie scattered across the genome, forming complex interactions[19], which can be activated and deactivated by the 3D conformation of the genome[20]. Detected SNPs in these regulatory regions might influence expression of a gene much further away than we now know account for. To estimate this effect, we do, however, need a more accurate picture of the complex genomic and regulatory interactions in higher fungi.

Although we did not find spontaneous mutations in coding regions to change RNA expression, we did find mutations in coding regions that led to functionally different proteins. That is, we found 411 missense SNPs in protein domains, and 40 nonsense SNPs that change the protein domain configuration of a protein. These missense and nonsense mutations are underrepresented in our observations, again pointing towards an evolutionary conservation of these regions. Nevertheless, they do involve regulatory genes and important metabolic genes. While it is known that non-essential genes evolve faster than essential genes[21], we found no functional group enriched for SNPs. As *S.*

commune does not need its full functional repertoire in our experiment, we expected some groups to be more mutated than others. For example, our samples are grown exclusively on minimal medium[11], implying that the need for carbohydrate active enzymes (cazymes) is reduced. Yet, we do not observe more mutations in this group of genes due to a lesser evolutionary pressure. We should, however, realize that the effect of selective pressure might be limited due to the relatively small number of generations. Based on the detected mutation rate, we can expect approximately 300 SNPs to accumulate through the growth of a single plate in a dikaryotic *S. commune* strain. Given the incidence of nonsense mutations in our dataset, we can anticipate that approximately 5 will induce a stop codon.

It is possible that a number of the SNPs we discover are the result of post-transcriptional modification, such as RNA editing. RNA-editing has been shown to occur in fungi. However, the SNPs we identify are confidently associated to nodes in an evolutionary lineage, which is not what is to be expected from RNA editing events. Additionally, the mutation rate we estimate corresponds with the known mutation rate of *S. commune* in the wild, and the SNPs around the mating type loci correctly coincide with expected recombination sites. Together, these observations indicate that the substitutions we observe are actually genetic variants, rather than post-translational modifications. To resolve these conflicting observations would require an additional study in which DNA and RNA are sampled simultaneously, such as simul-Seq[22].

It has been shown that errors in the repair of damaged DNA (and possibly cDNA) are linked to faulty variant identification[23]. Such errors could explain the majority of variants with low variant allele frequencies. And, in our case, the majority of SNPs do have low VAFs. However, most of our SNPs are identified across a large number of samples. Hence, it is unlikely that the DNA is damaged and incorrectly repaired at identical locations over multiple samples. In this work, we developed our own method to detect SNPs in RNA-Seq data, which makes sense for *S. commune* since the transcriptome covers 89% of the genome. There have been previous attempts to call SNPs from RNA-Seq reads[24–28]. With the exception of JACUSA[28], which was designed for the identification of RNA editing events, other approaches generally rely on GATK[29], which was primarily not designed for the study of variants in RNA-Seq data. Most importantly, GATK assumes an approximately uniform distribution of reads across the genome, which is certainly not the case for RNA-seq data. Furthermore, the allelic imbalance due to allele specific expression (or, in our case, karyollele specific expression) severely hampers the performance. The best practices as described by Broad Institute indicate that results are only acceptable when strict filters are used[30]. When we applied the GATK pipeline to our data, we found only 351 SNPs that associated to our experimental design tree. Therefore, we chose to develop our own method. Our initial SNP calling step is permissive and will call many spurious SNPs. Our method, therefore, strongly relies on a second step to filter spurious SNPs, e.g. it only permits SNPs with low RNA-seq coverage if they are present across several samples. Without knowing the relationship between the samples, this becomes very difficult.

As we derived mutations from RNA-Seq data, we do have to make a note of caution on our findings as they depend on the expression level of a gene. That is, when a gene is not expressed, no mutation can be detected. We remedy this by exploiting the (full) experimental setup. Throughout the complete experiment, only 1,612 (9.8%) of all predicted genes were considered to be not expressed (FPKM < 1) in any sample. Thus, although we do not capture the entire genome, we capture a considerable portion of it, and the reported mutation rate takes this into account.

CONCLUSION

In the laboratory, the selection pressures that shaped the genotype and phenotype of wildtype organisms are replaced, relaxed, or even lost. We have shown that *S. commune*, a model organism for mushroom formation, has the same high mutation rate in the lab as in the wild. With this work, we like to remind researchers that spontaneous mutations will accumulate in their experiments. Even the best strain preservation system cannot escape this. We showed that SNPs are introduced in a variety of important functional groups, and that they can have an effect on the function and regulation of genes. It is not clear that there is a better way to prevent the accumulation of spontaneous mutations, other than reducing the number of generations between the primary stock and the experimental strains derived thereof. We recommend that labs implement a sample tracking system in the lab, whereby each sample that enters a freezer is registered with its ancestor sample. This will enable the isolation of mutations should they later be discovered. Additionally, the experimental design should take into account the additional mutations that could accumulate, and replicates should originate from different parental plates. Although this may result in higher biological variation between the replicates, it will eliminate differences that result from confounding mutations that accumulated in the tree.

MATERIALS AND METHODS

H4-8 *S. commune* strain: The H4-8 *S. commune* strain[31] is a co-isogenic dikaryon, meaning that it is a heterokaryon whose constituent homokaryons are supposedly identical with the exception of the mating type loci. It is the result of an integration of the H4-8a and H4-8b homokaryons. The H4-8b strain was achieved through 9 backcrosses between H4-8a and 4.40, selecting in each stage for a crossing that had a compatible mating type to H4-8a. During meiosis, the chromosomes are exchanged and (often) undergo crossover at locations of genetic similarity[16]. The exact efficiency of this backcrossing procedure in terms of homozygosity, especially in the chromosomes containing the mating type loci is unknown.

Mating type loci: The two homokaryons of H4-8 differ in their A and B mating type loci[31]. These loci were identified in version 3 of the H4-8 genome by mapping the genes annotated in the $matA\alpha$, $matA\beta$, $matB\alpha$ and $matB\beta$ of version 2 to the version 3 genome using the BLAST functionality of the JGI DOE website. See Supplementary Note 2.

Knock-out strains: The knockout strains all originate from a ku80 knockout[32] (Supplementary Note 4), which was used to generate a series of regulatory gene knockouts[12], all stored in the -80 freezer. The ku80 knockout is the result of several stressful interventions (Supplementary Note 4), over an unknown number of generations. Beyond the phenotypic and transcriptomic differences induced by the knockouts[12, 33], it is not known what additional sequence variation is induced by the knockout of the ku80 gene and the final regulatory genes. After the knockout of the second gene, the ku80 gene is crossed back into the genome.

RNA-Seq data: RNA-Seq samples were retrieved from BioProject PRJNA323434[11]. To produce these samples, mRNA was isolated from *S. commune* strain H4-8 grown at 25°C on minimal medium containing 1% glucose and 1.5% agar[34]. The wildtype strain was initially sampled twice, once in the aggregates stage of development, and once in fruiting body of the mushroom and sequenced on an Illumina Hi-Seq 2000. Knockout strains were sampled when they reached the aggregates state and mushroom stage. If a knockout was halted in an earlier developmental stage (Supplementary Note 3), then they were sampled when the wildtype reached the aggregates or mushroom stage. A later second sequencing run sequenced wildtype samples at three additional developmental stages, vegetative growth, induced vegetative growth (after exposure to O_2 and light), and primordia. Details on the sequencing runs can be found in Supplementary Note 16.

Read Alignment: Raw reads were trimmed using TRIMMOMATIC[35] and the resulting reads were aligned to the reference genome using two-pass STAR[36], where the second pass used the splice junctions detected in all samples during the first alignment pass (Supplementary Note 17). Reads that ambiguously mapped were discarded. STAR was used to sort the resulting BAM files based on read alignment co-ordinate. Duplicate reads were flagged with PICARD (http://picard.sourceforge.net/).

Detecting SNPs from RNA-Seq data: We process the aligned BAM files, ignoring duplicate reads, counting the number of observed nucleotides at each position in the genome. If the quality of a base is less than 30 in the PHRED scoring system, it is not counted. Based on the CIGAR strings in the BAM file, we also maintain a record of insertions and deletions observed in the alignments. For each position on the genome, we test a base for SNPs only if the base is not within 4 bases of a possible insertion/deletion site/splice junction, and that base is not an 'N' in the reference genome. For each nucleotide, we calculate the probability that it is observed erroneously. To do this, we assume that each erroneous observation of a nucleotide at a specific locus follows a Bernouilli trial with a small probability of success. With multiple observations, we build a binomial distribution around the probability of observing a specific nucleotide by error. Thus, when a locus has a depth of x, and x_n counts of nucleotide n, then $P(X_n > x_n)$ expresses the probability of observing more than x_n counts erroneously, where $X_n \sim B(x, 0.01)$. Clearly, with increasing observations of the nucleotide (x_n), the probability of seeing that nucleotide at that locus as the result of an error becomes smaller. If this probability becomes smaller than 0.05 we conclude that the nucleotide is not observed erroneously, and thus is truly observed. We do so for all four nucleotides and when one of them passes this test and it is not equal to the reference nucleotide of that locus, we call a potential SNP at that base. Any SNP in a gene knockout region that originates from that knockout sample is removed. All positive and negative base calls are output in VCF format.

Assigning SNP origin in the generation tree: We use the lineage information in the sample tree to enhance our confidence in a SNP, and to remove spurious SNP calls. The VCF files of all the samples are merged and sorted on base coordinates. SNP calls from different samples are grouped together at each base. A generation tree is constructed, such as the one shown in 6.1. For each SNP, we determine which nodes in the tree are possible candidates for the origin of the SNP. To do this, we calculate three metrics for each node of the tree: Sn, the number of descendent leaf nodes that have this SNP; En, the number of reads supporting this SNP across all the child nodes, and Pn, the possibility of this node harboring the SNP, being either 'yes', 'no' or 'maybe'. For leaf-nodes a SNP is 'yes' when the SNP is present, 'no' when the SNP is not present, and 'maybe' when there is not enough depth to make a SNP call (i.e. depth < 3). For non-leaf nodes, the possibility of a SNP is 'yes' when all its descendent leaves are 'yes', 'no' when at least one of its descendent leaves is 'no', and 'maybe' when all its descendent leaves are 'yes' or 'maybe'. For each SNP, we select the nodes highest (towards the root) in the tree where Pn is 'yes' or 'maybe', and either $E_n > 3$ or $S_n > 1$, as the node of origin for that SNP. This results in SNPs that are either supported by sufficient depth within at least one sample, or supported by multiple samples. SNPs with multiple alternative nucleotides, or SNPs whose origin can't be resolved (i.e. no origin found, or multiple origins found) are discarded.

Estimating transcript abundance: To calculate transcript abundance, we pre-processed the reads with TRIMMOMATIC[35] and aligned the reads to the genome using a two-pass STAR, as in the read alignment above, only in this case we permitted ambiguous alignments. Expression of each transcript for each sample was quantified and normalized with the Cufflinks[37] toolkit.

Associating SNPs to genes and assessing deleteriousness: If a SNP lies within the coding region of a gene, then we can assess the deleteriousness of the SNP. If the transcript with the SNP produces the same amino acid sequence as without, then the SNP is considered synonymous. If, on the other hand, the amino acid sequence is changed, then it is a missense mutation, and if the sequence is shortened, then it is described as nonsense. If a SNP lies within 500bp upstream of the start codon of a gene, then we say that the SNP falls within the promotor region of that gene.

Calculating a mutation rate To calculate a mutation rate, we consider the number of mutations associated to each node in the tree. As we do not precisely know how many steps were involved in creating the original double knockouts plates, the ku80 knockout plate, and the primary stock plate, we excluded those samples from the calculation. The number of SNPs in each sample are divided by the number of bases considered, and multiplied by the number of generations that each plate represents (200, see Supplementary Note 10 and [38]). The mutation rates for each plate are averaged to arrive at a cross-sample mutation rate. A confidence interval is calculated assuming a normal distribution. For the genome-wide mutation rate, we used the number of bases covered with at least 5 reads (excluding mating type loci) multiplied by two, representing the callable part of the diploid genome. Because the number of detected mutations is dependent upon read depth, we need to correct for the library sizes in each sample. However, since the ability to call a SNP depends on the coverage of each base, we should, more specifically, correct for the number of confidently callable bases per sample. For each sample i, $k(i)$ denotes the number of bases with at least 5x coverage (Supplementary Note 16). For non-leaf nodes, we infer $k(i)$ to be the maximum across its leaves. The number of SNPs detected in each sample is multiplied by $\max_j(k(j)/k(i)$, where $max(k(i))$ is the maximum number of confidently callable bases across all the samples. This scales the number of mutations up for those samples with less read depth.

Detecting mutation hotspots: A sliding window of 10,000bp up- and down-stream of a detected SNP, which contains at least 20 SNPs is considered a mutation hotspot. Functional annotations. Interpro domain annotations were taken from the JGI DOE website, filtered with a score threshold < 0.05. Transcription factors were predicted based on a curated list of fungal DNA binding domains, as in [11]. Cytochrome P450 genes were predicted based on the Interpro domain IPR001128, and metabolic genes based on the GO annotation GO:0008152. Carbohydrate active proteins were predicted using the CAT[39, 40] tool, selecting only those proteins which are predicted both with the PFAM and sequence predictors. Alternatively spliced genes were taken from [11].

Conflict of interest: The authors have no conflict of interest to declare.

Acknowledgements: The sequence and annotation data of *S. commune* H4-8 version 3 were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community. This research is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs. Additionally, we thank Aurin M. Vos and Robin A Ohm for fruitful discussions.

Author contributions: TG, HABW, TA, and MJTR wrote the manuscript. JFP and LGL performed the biological experiments. TG, TA, and MJTR designed the analyses. TG performed the analyses. All authors aided in biological interpretation of the results. All authors reviewed the manuscript.

BIBLIOGRAPHY

[1] C. F. Baer, M. M. Miyamoto, D. R. Denver. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews Genetics* 8(8):619–631 (2007).

[2] J. E. Barrick, R. E. Lenski. Genome dynamics during experimental evolution. *Nature Reviews Genetics* 14(12):827–839 (2013).

[3] C. Zeyl, J. a. DeVisser. Estimates of the rate and distribution of fitness effects of spontaneous mutation in saccharomyces cerevisiae. *Genetics* 157(1):53–61 (2001).

[4] S. B. Joseph, D. W. Hall. Spontaneous mutations in diploid saccharomyces cerevisiae: More beneficial than expected. *Genetics* 168(4):1817–1825 (2004).

[5] M. a. Baranova, et al. Extraordinary genetic diversity in a wood decay mushroom. *Molecular Biology and Evolution* 32(10):2775–2783 (2015).

[6] J. Raper, P. Miles. The genetics of schizophyllum commune. *Genetics* (1958).

[7] J. G. Wessels, O. M. de Vries, S. a. Asgeirsdottir, J. Springer. The thn mutation of schizophyllum commune, which suppresses formation of aerial hyphae, affects expression of the sc3 hydrophobin gene. *J GenMicrobiol* 137(10):2439–2445 (1991).

[8] P. G. Miles, H. Lund, J. R. Raper. The identification of indigo as a pigment produced by a mutant culture of schizophyllum commune. *Archives of Biochemistry and Biophysics* 62(1):1–5 (1956).

[9] J. Springer, J. G. H. Wessels. A frequently occurring mutation that blocks the expression of fruiting genes in schizophyllum commune. *MGG Molecular & General Genetics* 219(3):486–488 (1989).

[10] J. F. De Jong, R. A. Ohm, C. De Bekker, H. A. Wösten, L. G. Lugones. Inactivation of ku80 in the mushroom-forming fungus schizophyllum commune increases the relative incidence of homologous recombination. *FEMS Microbiology Letters* 310(1):91–95 (2010).

[11] T. Gehrmann, et al. Schizophyllum commune has an extensive and functional alternative splicing repertoire. *Scientific Reports* 6(1):33640 (2016).

[12] J. Pelkmans. *Environmetal signalling and regulation of mushroom formation.* Ph.D. thesis, Utrecht University (2015).

[13] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61(4):893–903 (1969).

[14] T. Y. James. Why mushrooms have evolved to be so promiscuous: Insights from evolutionary and ecological patterns. *Fungal Biology Reviews* 29(3-4):167–178 (2015).

[15] J. R. Raper, G. S. Krongelb, M. G. Baxter. The number and distribution of incompatibility factors in schizophyllum. *The American Naturalist* 92(865):221–232 (1958).

[16] V. B. Seplyarskiy, et al. Crossing-over in a hypervariable species preferentially occurs in regions of high local similarity. *Molecular Biology and Evolution* 31(11):3016–3025 (2014).

[17] J. F. Pelkmans, et al. The transcriptional regulator c2h2 accelerates mushroom formation in agaricus bisporus. *Applied Microbiology and Biotechnology* 2 (2016).

[18] A. S. M. Sonnenberg, et al. A detailed analysis of the recombination landscape of the button mushroom agaricus bisporus var. bisporus. *Fungal Genetics and Biology* 93:35–45 (2016).

[19] M. W. Vermunt, et al. Large-scale identification of coregulated enhancer networks in the adult human brain. *Cell Reports* 9(2):767–779 (2014).

[20] S. Babaei, W. Akhtar, J. de Jong, M. Reinders, J. de Ridder. 3d hotspots of recurrent retroviral insertions reveal long-range interactions with cancer genes. *Nature communications* 6:6381 (2015).

[21] I. K. Jordan, I. B. Rogozin, Y. I. Wolf, E. V. Koonin. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Research* 12(6):962–968 (2002).

[22] J. A. Reuter, D. V. Spacek, R. K. Pai, M. P. Snyder. Simul-seq: combined dna and rna sequencing for whole-genome and transcriptome profiling. *Nature Methods* (2016).

[23] L. Chen, P. Liu, T. C. Evans, L. M. Ettwiller. Dna damage is a major cause of sequencing errors, directly confounding variant identification. *bioRxiv* 756(February):070334 (2016).

[24] R. H. Ramirez-Gonzalez, et al. Rna-seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat. *Plant Biotechnology Journal* 13(5):613–624 (2015).

[25] P. Deelen, et al. Calling genotypes from public rna-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med* 7(1):30 (2015).

[26] R. Piskol, G. Ramaswami, J. B. Li. Reliable identification of genomic variants from rna-seq data. *American Journal of Human Genetics* 93(4):641–651 (2013).

[27] E. M. Quinn, et al. Development of strategies for snp detection in rna-seq data: Application to lymphoblastoid cell lines and evaluation using 1000 genomes data. *PLoS ONE* 8(3) (2013).

[28] M. Piechotta, E. Wyler, U. Ohler, M. Landthaler, C. Dieterich. Jacusa: site-specific identification of rna editing events from replicate sequencing data. *BMC Bioinformatics* 18(1):7 (2017).

[29] A. McKenna, et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research* 20(9):1297–1303 (2010).

[30] G. V. der Auwera. Calling variants in rna-seq (2017).

[31] R. a. Ohm, et al. Genome sequence of the model mushroom schizophyllum commune. *Nature biotechnology* 28(9):957–63 (2010).

[32] R. a. Ohm, et al. An efficient gene deletion procedure for the mushroom-forming basidiomycete Schizophyllum commune. *World Journal of Microbiology and Biotechnology* 26(10):1919–1923 (2010).

[33] R. a. Ohm, J. F. de Jong, C. de Bekker, H. a. B. Wösten, L. G. Lugones. Transcription factor genes of schizophyllum commune involved in regulation of mushroom formation. *Molecular Microbiology* 81(6):1433–1445 (2011).

[34] A. F. Van Peer, C. De Bekker, A. Vinck, H. a. B. Wösten, L. G. Lugones. Phleomycin increases transformation efficiency and promotes single integrations in schizophyllum commune. *Applied and Environmental Microbiology* 75(5):1243–1247 (2009).

[35] A. M. Bolger, M. Lohse, B. Usadel. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* 30(15):2114–2120 (2014).

[36] A. Dobin, et al. Star: ultrafast universal rna-seq aligner. *Bioinformatics (Oxford, England)* 29(1):15–21 (2013).

[37] C. Trapnell, et al. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols* 7(3):562–78 (2012).

[38] M. Raudaskoski, M. Salonen. Interrelationships between vegetative development and basidiocarp initiation. In D. H. Jennings, A. D. M. Rayner (eds.), *The ecology and physiology of the fungal mycelium*, chap. 13, 291–322. Cambridge University Press, Cambridge (1984).

[39] B. H. Park, T. V. Karpinets, M. H. Syed, M. R. Leuze, E. C. Uberbacher. Cazymes analysis toolkit (cat): Web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using cazy database. *Glycobiology* 20(12):1574–1584 (2010).

[40] V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, B. Henrissat. The carbohydrate-active enzymes database (cazy) in 2013. *Nucleic Acids Research* 42(D1):D490–D495 (2014).

CHAPTER 7
DISCUSSION

In this thesis, we have explored the variation at the genetic level between fungi (Chapter 2) and within a single fungus (Chapters 5 and 6). We transferred some existing knowledge from one fungus to another (Chapter 3) and we have also explored transcriptomic variation between them (Chapters 4 and 5). In Chapter 2, we developed an algorithm that was able to find non-collinear syntenic regions between diverged genomes at the protein level. This revealed that the two fungi, *Schizophyllum commune* and *Agaricus bisporus*, have large regions that are in synteny, albeit separated by large genomic rearrangements, and that many regulatory genes previously identified in *S. commune* lie in syntenic clusters between *S. commune* and *A. bisporus*. In Chapter 3, we showed that one of the genes known to be involved in mushroom formation in *S. commune* also is involved in mushroom formation in *A. bisporus*. In *S. commune*, the overexpression of this gene produces more mushrooms at a faster rate, and we observe the same effect in *A. bisporus*. This validates the use of *S. commune* as a model for mushroom formation in *A. bisporus*. Chapter 4 investigated alternative splicing and its effect on functionality in *S. commune*. We identified a large number of alternatively spliced genes that produced multiple transcripts, and linked their expression to mushroom development. Further, we showed that the function of isoforms can be different and can have quite a large impact. In Chapter 5, we explored the transcriptomic activity of the different nuclear types in *A. bisporus*. We found that many karyolleles are produced at different rates in the different nuclei. Specifically, we found that many cazymes were more highly expressed in the P2 nuclear type. This has been the first report ever on genome-wide nuclear specific expression in a fungus that contains different nuclear types. In Chapter 6, we investigated the variation that accumulated across RNA-Seq experiments in *S. commune*. We detected SNP mutations from RNA-Seq data, and were able to estimate a mutation rate that was highly similar to the one found in the wild. Further, we linked some of these variations to differences in expression.

Taken together, this thesis represents an investigation of the complex variation between and within two fungi. In this concluding Chapter, we make some final notes on our research in the context of fungal research, and on bioinformatics and biological research in general.

BIOLOGICAL COMPLEXITY COMPLICATES REASONING

Biological systems are teeming with variation, and one kind of variation can easily be mistaken for another kind. The slight sequence variations between the same gene in different homokaryons, could be mistaken for alternative splicing, or a spontaneous mutation. An expression difference between two developmental states could actually be an expression change in only one homokaryon. All of these sources of variation, between genomes, through alternative splicing, by expression, and by mutation are tied tightly together. Not to mention other sources of variation that are still being revealed, such as polycystronic transcription[1], and epigenetic[2] and epitranscriptomic [3] mechanisms. In our work, we addressed each of these sources independently, ignoring that these sources work in concert and thus confound analyses. But, how would that impact our results?

In Chapter 2, we identified syntenic regions between *S. commune* and *A. bisporus*. Although the genome of *A. bisporus* is heterokaryotic, the reference is based on H97, a heterokaryotic strain of *A. bisporus*. As a result, we construct an incomplete view of the syntenic relationships between *S. commune* and *A. bisporus*, especially considering that *A. bisporus* does not exist as a homokaryon in the wild. Ideally, we should compare *S. commune* to both homokaryotic strains in a heterokaryon.

In Chapter 3, we verified that the c2h2 gene, known to be involved in mushroom formation in *S. commune*, also plays a role in mushroom formation in *A. bisporus*. We examined the expression of genes throughout mushroom growth, and in different tissues. In this Chapter, we did not consider the different nuclear types of *A. bisporus* that we investigated in Chapter 5. Although the transcription factors fst4 and hom2 exist in different forms in the P1 and P2 homokaryons, c2h2 does not. None of these transcription factors are differentially expressed between the two homokaryons at any time throughout development (Chapter 5). Therefore, we can make the assumption that the functional difference between the two forms is not significant, and that the different homokaryons do not influence the conclusion that c2h2 is involved in mushroom formation.

In Chapter 4, we identified alternative splicing events in the *S. commune* heterokaryon. Although it is a heterokaryotic species, the strain we used was co-isogenic. This, for the most part removes the interference of different karyolleles in the two nuclear types. Although we observed SNPs between the two homokaryons (i.e. the two nuclear types are not entirely co-isogenic, Chapter 6), relatively few of these SNPs are present in alternatively spliced genes, so interference is also minimal.

In Chapter 5, we studied the differential expression between karyolleles in the different nuclear types in the heterokaryon of *A. bisporus*. There, we averaged all marker counts in a gene, effectively collapsing all possible alternative splicing products into one transcript and ignoring the possibility of alternative splicing. Further, we did not take into account the markers that would identify different isoforms. Consequently, the observed differential expression could be caused by an alternative use of alternatively spliced transcripts, rather than alternative expression of all transcripts.

In Chapter 6, we identified SNPs from RNA-seq data, and inferred the moment they occurred in the experimental design. We ignored SNPs that appear to have been introduced at multiple moments in time. These observations conflict with our 'infinite sites assumption', which assumes that a mutation is only gained once and never lost. In reality, such mutations could be the result of a mutation that persists in a mutant for some time, before eventually dying and being removed from the population. Alternatively, such observations could be the result of SNPs that lie on alternatively spliced transcripts that are only expressed in specific stages of development. Thus, by doing so, we do not use all information. Taken together, these examples highlight that to be able to understand the biological complexity, the field really needs to shift towards measuring more sources of variation and, more importantly, a comprehensive analysis of the data.

BIOLOGICAL VALIDATIONS ARE REQUIRED

In almost every Chapter of this thesis, we put forward a hypothesis based on our observations under some initial assumptions, but we refrained from biological validations (mostly because of the significant effort required to perform them). Here, we propose possible validation experiments.

In Chapter 4, we concluded that alternative splicing produces functionally different proteins, based on the observation that alternatively spliced transcripts have different functional annotations. An alternative explanation would be that the different transcripts are errors of the splicing machinery, or merely a mechanism to halt translation[4]. We think this is unlikely as the expression of these different transcripts is linked to development, and that the types of alternative splicing events are similar to those in other fungi. To support this claim, we need to show: 1) the presence of the alternative transcript, 2) the differential expression of the transcripts, and 3) the differential function of the two transcripts. To validate the existence of the alternative transcript, we would need to isolate RNA from developmental stages where the alternate forms are present, design primers that distinguish the different isoforms, and perform (q)PCR to amplify that specific form. To confirm that the different transcripts have a functionally different role, we would need to perform a knockout or knockdown experiment, and observe a phenotypic difference. Alternatively, one could perform long read RNA-sequencing at the different stages, which provides a genome-wide validation of the presence of alternatively spliced transcripts.

Likewise, we provided evidence of alternative nuclear type activity in *Agaricus bisporus* (Chapter 5). We detected sequence differences between the two nuclear types, observed a differential expression of these different sequences, and concluded that this difference is regulated by the organism. To support this claim, we would need to perform essentially the same validation experiment as for the alternatively spliced transcripts: validation of different sequences, of different expression, and differential function. The additional complexity is that genetic modification tools do not currently exist in *A. bisporus*.

When we identified SNPs in our RNA-Seq data (Chapter 6), we assumed that mutations accumulate in our plates, while they could just as well be RNA-editing events. We postulated that this is unlikely as they show relatively low variant allele frequencies, and originate at nodes in the sample tree that do not represent phenotypically related samples. To validate these outcomes one needs to repeat the experiment and sequence both the DNA and RNA content for the same samples. Also, detected SNPs should be re-validated using deep sequencingor or by Sanger sequencing.

With this section, we stress the importance of biological validations. Clearly, smart reasoning helps in rejecting alternative explanations. But finally, a firm validation should underpin conclusions drawn from the data. Unfortunately, these are always time consuming. The primary challenge is to draw the right conclusions from the data to test. Future technologies will ease analysis

Long read technology will solve many sequence analysis problems in fungi. In Chapter 4, we had a problem with overlapping transcripts, which made it difficult to deconvolve the reads from neighbouring genes. This problem forced us to discard any reads up or downstream of the open reading frames. As a result, we lose a lot of information about the possible structures of transcripts. With long read sequencing, we will be able to identify transcription start sites, the spliced transcript sequences and the UTRs of the transcripts. This will provide a much better insight into alternative splicing tools in dense genomes.

In Chapter 6, we faced a practical limitation because we relied upon sequence differences between the two nuclear types. When orthologous genes in the two nuclear types are identical, there are no sequence differences that we can exploit to assign the nuclear origin to a gene. With existing technologies, the analysis of nuclear specific expression is fundamentally limited by the mRNA capture technology. Even in a single cell, the mRNA produced by the two homokaryons are mixed in the cytoplasm of the cell, and for sequences with no genomic variation it becomes impossible to link them to homokaryon differences. New technologies in sequence isolation can help us do better. Single nuclei sequencing[5], isolating and sequencing mRNAs from individual nuclei, will make it possible to study genes in a homokaryon specific nuclear context, giving a genome wide view of karyollele specific expression.

Another novel technology is Simul-Seq[6], which makes it possible to isolate DNA and RNA from the same sample. This technology will be useful to disentangle SNPs caused by evolutionary events from those that result from RNA editing events, clarifying our findings on experimentally acquired genomic variations as studied in Chapter 6.

The future of fungal research is intertwined with the assembly of many genomes. This is of course not trivial as colonies need to be grown, DNA needs to be isolated, and library preparation needs to be done. Many fungi grow on very non-uniform media and may not even grow on their own (if they are symbionts). Even ignoring that, DNA isolation and library preparation takes a lot of time. Therefore, new technologies, such as the Voltrax[7] that automate this process are extremely interesting to increase the throughput of raw sample material to DNA sequence information, as well as to reduce human error during protocol execution.

These examples show that technology in this field advances at an incredible speed and overcomes a lot of the current limitations, better and even faster than bioinformatics solutions. As bioinformaticians, we are thrilled by these developments as they surpass bioinformatic efforts to resolve these limitations. These advances allow us to reserve more resources to the so much needed downstream data analysis, enabling more in depth understanding of the biological processes we study.

FURTHER APPLICATIONS OF THIS WORK BEYOND FUNGI

The methodologies presented in this thesis can be generalized to other situations as well. For example, the clustering algorithm of Chapter 2 cuts a hierarchical clustering tree of objects depending on the significance of a statistical test at each node. Described in this way, the algorithm can be applied to any clustering problem. For example, when clustering genes based on their co-expression, and cutting the corresponding tree based on a functional enrichment at each node. Since we propose a rigorous statistical test to select cuts, we can avoid making arbitrary choices as is mostly custom practice[8].

The methodology to identify the nuclear origin of RNA-Seq reads as presented Chapter 6, can be very useful for metagenomics studies. We differentiate reads to their nuclear origin based on unique markers for their genome of origin. When applied to metagenomic reads, reads can be scanned for presence of different organisms in the sample, instead of aligning reads to genomes. Interestingly, the recently published KRAKEN[9] procedure works in a similar fashion. As another example, the same idea can be used to diagnose a patient with a mixed bacterial infection. Having a database of thousands of existing bacterial strains, we can identify unique markers that specify each one, and identify mixed infections by scanning reads of a patient for the unique markers representing each individual strain, instead of aligning reads first to the thousands of strains, or limiting ourselves to one specific reference strain.

An interesting alternative application of the SNP detection from RNA-Seq data, as proposed in Chapter 6, is to determine the lineage of cells and cell types in, for example, cancer development[10]. Knowing such lineages gives valuable information about the development and physiology of the tissue of consideration. Basically, one is interested in reconstructing the phylogeny of individual cells, often using mRNA expression. With SNPs from RNA-Seq samples it becomes possible to enrich this analysis with genetic information.

GENERAL BIOINFORMATICS CHALLENGES

As mentioned above, we considered each source of variation more or less independently from one another. The ultimate goal would be to integrate all the different sources of variation into a general model. In our case, a model explaining the expression of haplotype specific alternatively spliced transcripts across samples and influenced by the acquired SNPs. It is hard to imagine what form such a model would take, or whether it is even possible to derive such a model using only RNA-seq data. Novel tools that try to infer haplotype specific alternative splicing transcripts[11] indeed use additional data to perform their analysis. The integration of different sources of data, together with the integration of the different concepts will continue to be the greatest challenge in bioinformatics, and imperative to progress our understanding in biology.

Identifying true orthology, or even further, true gene identity between different species is difficult. Due to duplication events, the orthology relationships can become ambiguous. As a result, it becomes easy to make mistakes. Synteny can help resolve orthology inferences, but even this is imperfect. Take as an example the c2h2 gene in *S. commune* and *A. bisporus*. This gene did not lie in a syntenic cluster between the two genomes, but the gene had a similar function in the two fungi. This may be enough to infer orthology, but in this case synteny did not help us in transferring the knowledge. Moreover, the homology and syntenic relationships we derived between the two fungi was based on protein sequence only. Between fungi, detecting orthology at the DNA level is nearly impossible. Consequently, identifying functional similarity between genomes is still not fully resolved.

Orthology prediction is at the basis of transferring gene annotations from one species to the other. Currently, we still make use of models based on highly distant organisms to make predictions on the structure and function of various functional elements. For common cellular functions this works quite well, but for highly specialized features this is difficult. Given the high number of highly specialized functions across the tree of life, it remains a challenge to build models that predict specialized functionality from relatively distant genomes.

In Chapter 6, we identified SNPs in a variety of samples and associated them to nodes in the sample tree. This was possible because we knew the tree beforehand. In case a sample tree is unknown, we need to reconstruct the sample tree, which is not an easy task. A similar problem occurs when one wants to reconstruct the clonal evolution or expansion of cells. In that case, SNPs and their Variant Allele Frequencies (VAFs) are used to infer a tree that best explains the observed SNPs and their VAFs[10]. This would not work for SNPs derived from RNA-Seq data since in that case VAFs are highly influenced by the abundance of expression. For heterokaryotic fungi, this is further complicated due to the nuclear composition of cells. Perhaps structural variants, such as indels and inversions, may help. However, to identify structural variants from RNA-Seq data we need new detection tools.

Science relies on the ability of others to reproduce and replicate work. In bioinformatics, tools are often delivered stand alone, and work together with other tools provided by others. There has been a significant effort to make the sharing of tools easier, and efforts such as the R bioconductor package repository encourage developers to produce high quality software with documentation. This facilitates the use of packages together with the correct version of their dependencies. However, for many tools, these efforts do not help. Often, tools rely on specific server configurations to be run efficiently, and these are certainly not easy to describe or replicate. Initiatives such as workflow systems like Galaxy and Taverna[12], virtual machines, environment managers like Conda (`https://conda.io`), together with large package repositories like Bioconda[13], and software containers like Docker[14] are very promising. Solutions are essential before reproducibility, reusability and dependency of software become too big a problem.

FUTURE WORK IN THE FUNGAL DOMAIN

Fungi are largely underappreciated in scientific research. As a result, the roles they play in various ecosystems are still unknown. In some cases, such as metagenomics, fungi are largely ignored due to fungi specific difficulties such as the low abundance of fungi in metagenomic environments and difficulty in isolating DNA. Also, the assembly of fungal genomes has been limited due to the difficulty in isolating DNA, and the complexity in their genome structure when compared to bacteria. With more sequenced fungal genomes, our understanding of niche functions and the roles these fungi play in our environment will increase. Hence, this should be a priority in fungi research.

Even if the fungal research community agrees to sequence everything, the question still remains: Which genomes do we sequence first? As fungi play important roles in many important ecosystems, it might be reasonable to prioritize based on their presence in microbiomes. With barcoding information of fungal species, one can easily determine abundances of species in microbiomes. The order in which genomes should be sequenced could then be related to the frequency at which the different species are observed in different microbiomes, instead of random sampling.

GWAS (Genome Wide Association Study) has been a very successful tool to associate mutations to phenotypes. In fungi, this has hardly been done, specifically due to a lack of population studies where enough species variety has been captured. A complicating factor might be that fungal species have much higher interspecies variation than mammals, for example *S. commune* can still recombine with 14% gene sequence difference [15]. Although this may seem like a complicating factor, recall that the size of the *S. commune* genome is only a fraction of the size of the human genome, and a 0.02% sequence difference between humans corresponds to the same number of SNPs as a 15% difference between *S. commune* strains. Therefore, given enough samples, there should not be a power problem. Consequently, we advocate GWAS studies in fungi to increase our understanding of genotype/phenotype relationships.

In our research, we only had access to one epigenetic dataset, i.e. we used methylation profile of a single vegetatively growing *A. bisporus* sample to associate differential methylation to differential karyollele expression. One sample is hardly sufficient to make any strong claims about the cause of differential karyollele expression. In order to make a stronger claim, more methylation data is needed. Hence, more epigenetic processes should be investigated in fungi as their impact can be just as great as sequence differences, as we currently see in human data [16].

In humans and human model organisms, there are many data sources available, i.e. genomic, transcriptomic, epigenetic, translational, post-translational and proteomic data are available for a variety of populations. Currently, the state-of-the-art in most fungal species is transcriptomic data. The marked exception is yeast, for which a wide variety of epigenomic studies have been done. Yeast, however, is arguably a much more simple fungus than fungi with more complicated phenotypes [17]. Clearly, when a field such as mycology is spread over hundreds of species, there will be less variety of data available for each individual species. We must remember that phenotypic differences between species is the intertwined result of various processes going on in a cell. In an ideal scenario, mainstream fungal research would expand from yeast to another, more complicated, fungus. One which exhibits a more complicated life-cycle, forms multiple tissue types, grows on more complicated media, whose genetics have already been partially studied, and which can be manipulated in the lab. *Schizophyllum commune* meets all these criteria, and could thus provide an excellent model for higher fungi to study cell type differentiation, biomass degradation, genome organization, hyphal growth, epigenetic interactions, and more; as it currently does for mushroom formation.

BIBLIOGRAPHY

[1] S. P. Gordon, et al. Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. *PLOS ONE* 10(7):e0132628 (2015).

[2] Y. Saletore, et al. The birth of the epitranscriptome: deciphering the function of rna modifications. *Genome biology* 13(10):175 (2012).

[3] Y. Fu, D. Dominissini, G. Rechavi, C. He. Gene expression regulation mediated through reversible m6a rna methylation. *Nature Reviews Genetics* 15(5):293–306 (2014).

[4] Y.-J. Kwon, M.-J. Park, S.-G. Kim, I. T. Baldwin, C.-M. Park. Alternative splicing

and nonsense-mediated decay of circadian clock genes under environmental stress conditions in arabidopsis. *BMC Plant Biology* 14(1):136 (2014).

[5] S. R. Krishnaswami, et al. Using single nuclei for rna-seq to capture the transcriptome of postmortem neurons. *Nature protocols* 11(3):499–524 (2016).

[6] J. A. Reuter, D. V. Spacek, R. K. Pai, M. P. Snyder. Simul-seq: combined dna and rna sequencing for whole-genome and transcriptome profiling. *Nature Methods* (2016).

[7] ONT. About voltrax (2016).

[8] A. Mahfouz, M. N. Ziats, O. M. Rennert, B. P. F. Lelieveldt, M. J. T. Reinders. Shared pathways among autism candidate genes determined by co-expression network analysis of the developing human brain transcriptome. *Journal of Molecular Neuroscience* 57(4):580–594 (2015).

[9] D. E. Wood, S. L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* 15(3):R46 (2014).

[10] M. El-Kebir, L. Oesper, H. Acheson-Field, B. J. Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* 31(12):i62–i70 (2015).

[11] S. Mangul, et al. *HapIso: An Accurate Method for the Haplotype-Specific Isoforms Reconstruction from Long Single-Molecule Reads*, 80–92. Springer International Publishing, Cham (2016).

[12] J. Leipzig. A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics* 18(3):530–536 (2017).

[13] B. Grüning, et al. Practical computational reproducibility in the life sciences. *bioRxiv* (2017).

[14] C. Boettiger. An introduction to docker for reproducible research. *ACM SIGOPS Operating Systems Review* 49(1):71–79 (2015).

[15] V. B. Seplyarskiy, et al. Crossing-over in a hypervariable species preferentially occurs in regions of high local similarity. *Molecular Biology and Evolution* 31(11):3016–3025 (2014).

[16] ENCODE Project, et al. An integrated encyclopedia of dna elements in the human genome. *Nature* 489(7414):57–74 (2012).

[17] J. W. Taylor, C. E. Ellison. Mushrooms: morphological complexity in the fungi. *Proceedings of the National Academy of Sciences of the United States of America* 107(26):11655–11656 (2010).

SUMMARY

Fungi are microorganisms whose astounding variety can be found in every conceivable ecosystem on the planet. Fungi are nutrient recyclers, playing an irreplaceable role in the carbon cycle. They grow on land and in the sea, on plants and animals and in the soil. They feed us as mushrooms, and drive our economy as bioreactors. They leaven our bread and brew our beer, nourish our crops and spoil our food. They even directly play a role in human health. Fungi are, however, far more complex organisms than their simple phenotypes lead us to believe. In order to harness the potential of fungi, and to address the threats they pose, we must gain a better understanding of fungi. However, their substantial genomic and regulatory diversity impede our reasoning. Thus, to understand fungi, we need to understand their genetic and regulatory diversity.

In this thesis, I developed and utilized bioinformatics methods to understand variation within and between fungi. We focussed on two fungi: *Agaricus bisporus* (the champignon, or white button mushroom) because of commercial interest, and *Schizophyllum commune* (the split-gill mushroom) because it is used as a model organism for mushroom formation (for, amongst others, *A. bisporus*).

Capturing variation between fungi helps us understand evolutionary differences and whether functional implication within one organism can be transferred to the other. I found that genomic differences between *S. commune* and *A. bisporus* are smaller than expected. Not only are genes conserved between the two genomes, but I showed that gene neighbourhoods are also conserved for a large portion of the genome.

Variation in gene regulation occurs throughout growth and development and can be captured by looking at expression differences across such stages. Part of this variation occurs because an individual gene can produce multiple proteins through so-called alternative splicing. In fungi, it is difficult to study this phenomenon, and as a result the predominant belief was that alternative splicing does not happen. Using a method that I developed specifically for gene-dense fungi, I show that thousands of cases of alternative splicing occur in *S. commune*. Moreover, I show that different alternative splice forms are being expressed at different stages of mushroom development.

Another source of variation that occurs in fungi relates to the fact that many fungi, such as *A. bisporus*, have cells with two (or more) nuclei, each carrying a different (parental) copy of the same gene. I have shown that variations occur in activity of the two nuclei during development, resulting in varying expression of the two parental genes. The impact of this finding is that we need to consider which nucleus is active when one identifies genes relevant for a specific phenotype.

Finally, I looked at genomic variation driven by evolution that occurs across generations. I have developed a method with which genomic variation can be derived from expression (gene regulation) data. By re-evaluating expression data, I have shown that genomic variation as a result of evolution occurs even in standardized experiments in which one assumed that this does not happen. These insights are helpful when design-

ing new experiments to control the confounding effects of this evolutionary variation when detecting phenotypic differences.

With this thesis, I have introduced the first investigation into several novel sources of variation in fungi, with the hope that it will trigger further research into the complexity of fungal genomes and genome regulation.

SAMENVATTING

Schimmels zijn micro-organismen met een verbazingwekkende variëteit en komen in elk denkbaar ecosysteem op de planeet voor. Schimmels spelen een belangrijke rol in de natuur omdat ze voedingsstoffen hergebruiken. Zo spelen ze bijvoorbeeld een onvervangbare rol in de koolstofcyclus. Ze groeien op het land en in de zee, maar ook op planten en dieren. Het zijn lekkernijen in ons eten, ze zuren ons brood, brouwen ons bier, en voeden onze gewassen. Ze spelen zelfs een directe rol in de gezondheid van de mens. Kortom, ze spelen een belangrijke rol in ons leven. Om hun potentieel beter te kunnen benutten moeten we een beter begrip van hen krijgen. Paddenstoelvormende schimmels zijn echter veel complexer dan hun eenvoudige vorm ons doet geloven. Dit wordt verder belemmert doordat hun DNA enorm varieert. Om paddenstoelvormende schimmels beter te berijpen moeten we dus een beter begrip krijgen van hun genomische diversiteit.

Dit proefschrift beschrijft bioinformatica methoden die speciaal ontwikkeld zijn om de variatie binnen en tussen schimmels te meten. Om variaties in kaart te brengen en ze beter te begrijpen worden de methoden toegepast op data van twee paddenstoelvormende schimmels: *Agaricus bisporus*, de eetbare paddenstoel, en *Schizophyllum commune*, een paddenstoel die gebruikt wordt als model voor paddenstoelvormige schimmels.

Het vastleggen van variatie tussen paddenstoelvormende schimmels helpt om evolutionaire verschillen te begrijpen. Daarnaast helpt dit om bekende functionele implicaties van genen in de ene schimmel over te zetten naar genen van de andere schimmel. Dit proefschrift laat zien dat de genomische verschillen tussen *S. commune* en *A. bisporus* veel kleiner zijn dan op voorhand verwacht werd. Niet alleen worden individuele genen geconserveerd, maar er zijn zelfs hele gebieden van genen die geconserveerd zijn.

Naast verschillen in het DNA varieert de hoeveelheid eiwit dat een gen produceert gedurende de ontwikkeling van een organisme. Bovendien kan een individueel gen verschillende eiwitten aanmaken door middel van een proces dat 'alternative splicing' heet. Dit proces controleert welke stukjes DNA van een gen worden gebruikt om het eiwit samen te stellen. De overheersende gedachte was echter dat bij schimmels geen alternative splicing bestond. Dit kwam omdat het bij schimmels erg moeilijk is om dit fenomeen te meten. Dit proefschrift beschrijft echter een methode om alternative splicing varianten van een gen te detecteren specifiek voor schimmels. Op basis van deze methode blijkt dat er duizenden gevallen van alternative splicing voorkomen in *S. commune*. Bovendien blijkt dat de verschillende alternatieve splicing varianten in verschillende stadia van paddenstoelontwikkeling actief zijn, en dat de bijbehorende verschillende eiwitten dus een verschillende functie voor de paddenstoelvormende schimmel hebben.

Een andere bron van variatie die optreedt bij schimmels houdt verband met het feit dat veel schimmels, zoals *A. bisporus*, cellen hebben met twee (of meer) kernen, die elk een andere (ouderlijke) kopie van hetzelfde gen dragen. Dit proefschrift toont aan dat er variaties optreden in de activiteit van de twee kernen tijdens de ontwikkeling, en dat deze variatie ook gen-specifiek is.

Tenslotte bestudeerd dit proefschrift ook variaties die in het DNA optreden gedurende verschillende opvolgende generaties. Allereerst introdcuceert dit proefschrift een methode die variaties in het DNA kan meten aan de hand van geproduceerde mRNA moleculen van genen in plaats van het DNA zelf. Hiermee kan de aanwezigheid van variaties ook gerelateerd worden aan de activiteit van een gen. Met behulp van deze methode wordt in dit proefschrift bekeken of er ook variaties optreden in 'strain-preservation' systemen. Dit zijn systemen waarin een specificieke stam bewaard wordt om daarmee nieuwe experimenten te kunnen uitvoeren zodat experimenten onderling vergeleken kunnen worden. Dit proefschrift laat zien dat in deze systemen normale evolutie plaatsvindt en dus evenveel variatie geïntroduceerd wordt als in de natuur. Dit is belangrijk aangezien men er tot nog toe vanuit ging dat er geen variatie optrad tussen experimenten gebaseerd op deze systemen. De resultaten in dit proefschrift laten echter zien dat er wel degelijk variaties zijn en dat deze variatie een systematisch verstorend effect kunnen hebben op de statistische analyses van data uit deze experimenten.

Samenvattend beschrijft dit proefschrift genomische variatie in paddenstoelvormende schimmels waarmee begrip van hun complexiteit een stapje dichterbij gekomen is.

CURRICULUM VITÆ

Thies Gehrmann was born on April 8^{th} 1989 in Hannover, Germany. In 2010, he obtained his Bachelor's degree in Computer Science from Heriot Watt University, Scotland. Afterwards, Thies moved to the Netherlands and started his Master's studies at Leiden University, in Leiden, The Netherlands, in collaboration with Delft University of Technology, in Delft, The Netherlands. There, he worked on applying Conditional Random Fields to the problem of protein function prediction. He obtained his Master's degree in 2012.

In November 2012, Thies started his PhD studies in the Delft Bioinformatics Lab, part of the Pattern Recognition and Bioinformatics group of the Faculty of Electrical Engineering Mathematics and Computer Science at TU Delft in Delft. His research project involved examining gene expression and regulation in two mushroom forming fungi. This project was in close collaboration with the Fungal Microbiology group at Utrecht University. The research project was funded by the Dutch Open Technology Program STW.

Following this, Thies worked for one year as a postdoctoral researcher at the Westerdijk Institute of Fungal Biodiversity at the Royal Netherlands Academy of Arts and Sciences (KNAW) in Utrecht. Since January 2018, Thies is working Molecular Epidemiology group at the Leiden University Medical Center, together with the Max Plank institute institute for Biological Ageing in Köln, Germany, as a postdoctoral researcher on healthy human ageing.

List of Publications

- **T. Gehrmann**, J. F. Pelkmans, L. G. Lugones, H. A. B. Wösten, T. Abeel, M. J. T. Reinders. Variants in RNA-Seq data show a continued mutation rate during strain preservation of *Schizophyllum commune*. *bioRxiv*, page 201012, 2017.
- E. T. Diepeveen, V. Pourquie, **T. Gehrmann**, T. Abeel, and L. Laan. Patterns of conservation and diversification in the fungal polarization network. *bioRxiv*, page 154641, 2017.
- **T. Gehrmann**, J. F. Pelkmans, R. A. Ohm, A. M. Vos, A. S. Sonnenberg, J. J. Baars, H. A. Wösten, M. J. T. Reinders, and T. Abeel. Nucleus specific expression in the multinucleated mushroom-forming fungus *Agaricus bisporus* reveals different nuclear regulatory programs. *bioRxiv*, page 141689, 2017.
- A. L. Manson, T. Abeel, J. E. Galagan, J. C. Sundaramurthi, A. Salazar, **T. Gehrmann**, S. K. Shanmugam, K. Palaniyandi, S. Narayanan, S. Swaminathan, et al. *Mycobacterium tuberculosis* whole genome sequences from southern india suggest novel resistance mechanisms and the need for region-specific diagnostics. *Clinical Infectious Diseases*, 64(11):1494–1501, 2017.
- J. F. Pelkmans, M. B. Patil, **T. Gehrmann**, M. J. T. Reinders, H. A. Wösten, and L. G. Lugones. Transcription factors of *Schizophyllum commune* involved in mushroom formation and modulation of vegetative growth. *Scientific Reports*, 7(1):310, 2017.
- J. F. Pelkmans, A. M. Vos, K. Scholtmeijer, E. Hendrix, J. J. Baars, **T. Gehrmann**, M. J. T. Reinders, L. G. Lugones, and H. A. Wösten. The transcriptional regulator c2h2. *Applied microbiology and biotechnology*, 100(16):7151–7159, 2016.
- **T. Gehrmann**, J. F. Pelkmans, L. G. Lugones, H. A. Wösten, T. Abeel, and M. J. T. Reinders. *Schizophyllum commune* has an extensive and functional alternative splicing repertoire. *Scientific reports*, 6, 2016.
- **T. Gehrmann** and M. J. T. Reinders. Proteny: discovering and visualizing statistically significant syntenic clusters at the proteome level. *Bioinformatics*, 31(21):3437–3444, 2015.
- **T. Gehrmann**, M. Loog, M. J. T. Reinders, and D. de Ridder. Conditional random fields for protein function prediction. In *IAPR International Conference on Pattern Recognition in Bioinformatics*, pages 184–195. Springer, Berlin, Heidelberg, 2013.