3 Biotech (2021) 11:475
https://doi.org/10.1007/s13205-021-03032-y

REVIEW ARTICLE

Trends in biological data integration for the selection of enzymes and transcription factors related to cellulose and hemicellulose degradation in fungi

Jaire A. Ferreira Filho[1] · Rafaela R. Rosolen[1] · Deborah A. Almeida[1] · Paulo Henrique C. de Azevedo[1] · Maria Lorenza L. Motta[1] · Alexandre H. Aono[1] · Clelton A. dos Santos[1,2] · Maria Augusta C. Horta[1,3] · Anete P. de Souza[1,4]

Received: 11 May 2021 / Accepted: 15 October 2021 / Published online: 26 October 2021

Abstract

Fungi are key players in biotechnological applications. Although several studies focusing on fungal diversity and genetics have been performed, many details of fungal biology remain unknown, including how cellulolytic enzymes are modulated within these organisms to allow changes in main plant cell wall compounds, cellulose and hemicellulose, and subsequent biomass conversion. With the advent and consolidation of DNA/RNA sequencing technology, different types of information can be generated at the genomic, structural and functional levels, including the gene expression profiles and regulatory mechanisms of these organisms, during degradation-induced conditions. This increase in data generation made rapid computational development necessary to deal with the large amounts of data generated. In this context, the origination of bioinformatics, a hybrid science integrating biological data with various techniques for information storage, distribution and analysis, was a fundamental step toward the current state-of-the-art in the postgenomic era. The possibility of integrating biological big data has facilitated exciting discoveries, including identifying novel mechanisms and more efficient enzymes, increasing yields, reducing costs and expanding opportunities in the bioprocess field. In this review, we summarize the current status and trends of the integration of different types of biological data through bioinformatics approaches for biological data analysis and enzyme selection.

Keywords Fungi · Bioinformatics · Biotechnology · Data integration · Bioprocesses

Introduction

Fungi are heterotrophic eukaryotic organisms that are widely distributed in the environment (Maharachchikumbura et al. 2021; Liu et al. 2020). For centuries, fungi have been applied in processes of human interest, such as food production and agriculture. Fungi are an important source of enzymes, platform for the discovery of new enzymes, and/or host for the production of industrial enzyme products. Different classes of enzymes and bioactive compounds from fungi, such as hydrolases, peroxidases, lipases, and laccases, are used in second generation biofuel production, the pharmaceutical industry, food processing, pulp and paper industry, textiles and bioremediation (Singh et al. 2021; Huang et al. 2021; Anasonye et al. 2014; Saravanan et al. 2021; Tomer et al. 2021). With the advent of biotechnology, the application of fungi in bioprocesses has become commonplace (Singh and Gehlot 2020; Jouzani et al. 2020). The bioprospecting

✉ Anete P. de Souza
anete@unicamp.br

1 Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Campinas, SP, Brazil

2 Brazilian Biorenewables National Laboratory (LNBR), Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, SP, Brazil

3 Faculty of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, SP, Brazil

4 Department of Plant Biology, Institute of Biology, UNICAMP, Universidade Estadual de Campinas, Campinas, SP 13083-875, Brazil

of enzymes and compounds produced by fungi involves different scientific disciplines, such as genetics, microbiology, biotechnology, molecular biology, and bioinformatics (Wang et al. 2020; Saldarriaga-Hernández et al. 2020; Abrashev et al. 2021).

Bioinformatics is a scientific discipline that arose with the demand to analyze large-scale data produced by sequencing and other high-throughput techniques, encompassing knowledge of biotechnology, molecular biology, computing, and statistics (Pereira et al. 2020; Orlov and Baranova 2020; Jhalia and Swarnkar 2021). Bioinformatics can be defined as a science that links biological data storage, interpretation, and analyses with computational biology, enabling the integration of large-scale data produced through genomics, transcriptomics, and proteomics (Xu et al. 2020; Diniz and Canduri 2017; Saeys et al. 2007). Advances in bioinformatics have allowed the emergence of techniques for dealing with genome sequencing data, demanding a high computational processing capacity. Both fields evolved quickly in recent decades, and following the advances achieved by the Human Genome Project (Venter et al. 2001), many other industrially and clinically relevant fungal species lines have been sequenced (Martinez et al. 2008a; Machida et al. 2005; Van Den Berg et al. 2008; Batista et al. 2020).

The genomic sequencing of fungi is an essential technique for understanding the evolutionary principles of particular groups and searching for targets of biotechnological interest (Gurjar et al. 2019; Pramesh et al. 2020; Alberti et al. 2020). The process of genome assembly, prediction, and annotation involves diverse bioinformatics tools and analyses that are important for revealing the hypothetical function of a given gene or identifying clusters of nearby genes that are related to a particular process (Giani et al. 2020; Faksri et al. 2016; Pop and Salzberg 2008).

The complete genome of a fungus makes it possible to characterize the profile of a certain group of enzymes, such as carbohydrate-active enzymes (CAZymes) (Gujar et al. 2018; Zhao et al. 2013). CAZymes are enzymes whose activity is related to carbohydrates and are rationally grouped in the CAZy (http://www.cazy.org/) database (Huang et al. 2018). These enzymes are produced by all fungi (Glass et al. 2013) and are applied in bioprocesses related to water decontamination and the food, textile, paper, and cellulose industries. Important degradative CAZymes are also needed and are a target of prospecting by the energy sector for the production of biofuels from plant biomass (Bohra et al. 2018) through enzymatic hydrolysis. These enzymes are grouped into five catalytic modules: glycoside hydrolases (GHs), glycosyl transferases (GTs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), and auxiliary activities (AAs) (Cantarel et al. 2009; Levasseur et al. 2013). In addition, some enzymes are associated with binding modules known as carbohydrate-binding modules (CBMs) (Shoseyov et al. 2006). All of these modules are subdivided into families categorized based on level-specific information.

The study of gene expression levels under certain biological conditions is fundamental for understanding the complex expression systems of eukaryotes. RNA-sequencing (RNA-Seq) techniques associated with the refinement of bioinformatics tools reliably identify and quantify gene expression under different conditions and can even be used to study the exon composition in the alternative splicing process (Geniza and Jaiswal 2017; Costa-Silva et al. 2017; Ding et al. 2017; Luecken and Theis 2019). Different approaches are used to calculate the correlations between expression data, enabling researchers to predict the coregulation between genes. An understanding of the correlations between genes can allow researchers to predict the genes involved in a particular metabolic pathway or process at a systematic level (Hurley et al. 2012; Zhang et al. 2016).

Bioinformatics tools are also fundamental for predictions and in silico modeling for target selection during the biochemical and structural characterization of a protein (Martins-Santana et al. 2018; Guzmán-Chávez et al. 2018; Silva et al. 2020). Protein engineering is a unique process that requires in-depth knowledge of the organism being studied; thus, basic information, such as structural and functional information, on the genome of the organism must first be obtained to generate a set of enzymes that are modified and adapted for a particular bioprocess (Yang and Zhang 2018; Milić and Veprintsev 2015). The prospection of new proteins and the improvement of protein efficiency are strategies to achieve the highest efficiency in different industrial processes. In addition, the future development and improvement of bioinformatics and computational tools for predicting and assigning the functions of nonidentified and/or unannotated proteins show great potential.

The advancement of the application of computational knowledge to solve biological problems has generated a revolution in the way we understand complex systems involving the regulation of genes and proteins. The integration of different types of biotechnological data is only possible through the implementation of bioinformatics tools, the development of algorithms and the construction of databases (Fig. 1). In this review, we summarize the main approaches for the integration of different types of biological data from fungi to guide the selection and design of targets for application in bioprocesses.

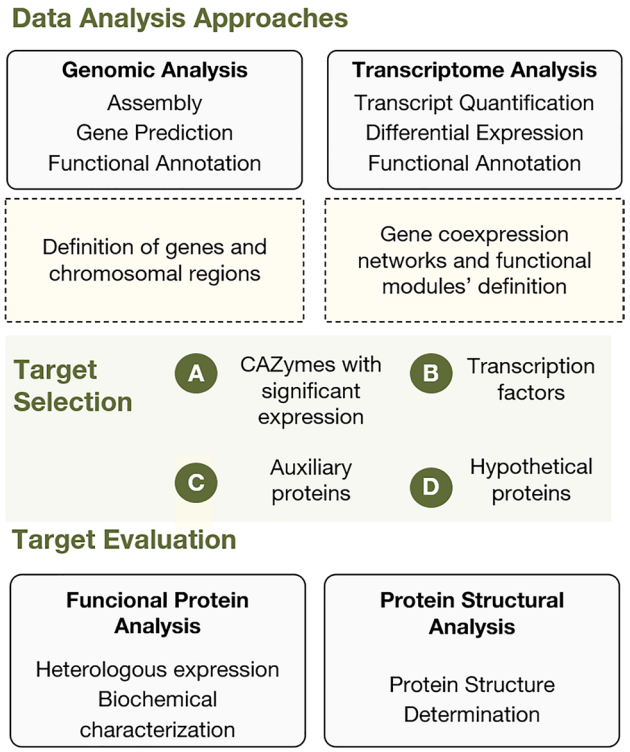

Fig. 1 Data integration approach for selecting enzymes for application in bioprocesses

Application of bioinformatics in the prospecting and engineering of enzymes

Genomic studies for mining enzymes of biotechnological interest in fungi

Genomics technology is a basic tool for searching for genes of interest that may have biotechnological applications. Fungi generally have small genomes relative to other organisms (such as plants) (Stajich 2017), which facilitates the execution of draft genome sequencing projects in fungal species. However, high-quality sequenced genomes are available for few species, which highlights the challenges in this type of project (Stajich 2017; Jauhal and Newcomb 2021). The main challenges in fungal genome assembly are related to the sequencing coverage necessary to fill gaps in the assembly with long and short reads and the assembly of repeated elements in the genome, especially in intron regions, which demands sufficient computational capacity for data processing (Amarasinghe et al. 2020; Magi et al. 2018).

Although many draft fungal genomes have been sequenced in recent years (Lange et al. 2021; Lopes et al. 2020), nearly complete genomes are available for only a few model fungi, such as *Saccharomyces cerevisiae* (Wei et al. 2007; Peter et al. 2018), *Trichoderma reesei* (Martinez et al. 2008a; Li et al. 2017), *Aspergillus fumigatus* (Nierman et al. 2005; Fedorova et al. 2008), and *Neurospora crassa* (Galagan et al. 2003; Wu et al. 2014). Even these genomes still include many genes with unknown functions (Ellison et al. 2014). The biological databases used for the automatic annotation of genes often include redundant information and show low curation, which hinders the process of gene annotation, and many genes/proteins are annotated as hypothetical in a fungal genomic project (Chen et al. 2017, 2020). FungiDB (https://fungidb.org/fungidb/app) is a platform that integrates different genomic and functional information from fungi to facilitate the process of curating genomic annotation (Basenko et al. 2018).

Sequenced fungal genomes can be used to search for genes with biotechnological applications through several different approaches, such as prospecting coregulated gene clusters and secondary metabolites or protein families, including CAZymes and phosphatases; prospecting for transcription factors (TFs) related to the regulation of genes involved in a particular bioprocess; SNP calling related to gene regions; and analyzing motifs in promoter and regulatory regions.

The genomic contents of CAZymes have been characterized in a large variety of fungi (Fig. 2). The identification of the CAZyme genome content is related to the fungal degradation capacity; however, the number of CAZyme families in the genome does not seem to directly correlate with the production levels and enzymatic efficiency exhibited by the fungus (Ferreira Filho et al. 2017; Druzhinina and Kubicek 2017; Martinez et al. 2008b). Nevertheless, the study of this group of enzymes provides insights into the evolution, mode of life, and biotechnological application of fungi. Barrett et al. (2020) used the predictions and annotations of the CAZyme secretomes of 465 Ascomycota and Basidiomycota genomes to examine the evolutionary relationships of the fungal CAZyme secretome. In another study, the CAZyme contents of the proteomes of *Aspergillus terreus*, *T. reesei*, *Myceliophthora thermophila*, *N. crassa,* and *Phanerochaete chrysosporium* in the presence of different lignocellulosic substrates were determined, indicating that the expression of these enzymes is directly related to substrate specificity (Arntzen et al. 2020). Additionally, the type of process (liquid-state (SmF) or solid-state fermentation (SSF)) influences fungal metabolism and subsequent enzyme production as a determinant of the enzymatic activity induced (Teigiserova et al. 2021). SSF culture, a fed batch culture, presents faster oxygenation but a slower sugar supply than SmF. The process is static without mechanical energy expenditures. In contrast, SmF cultures work as homogeneous systems requiring large energy expenditures to supply oxygen at sufficiently fast rates to address the large oxygen

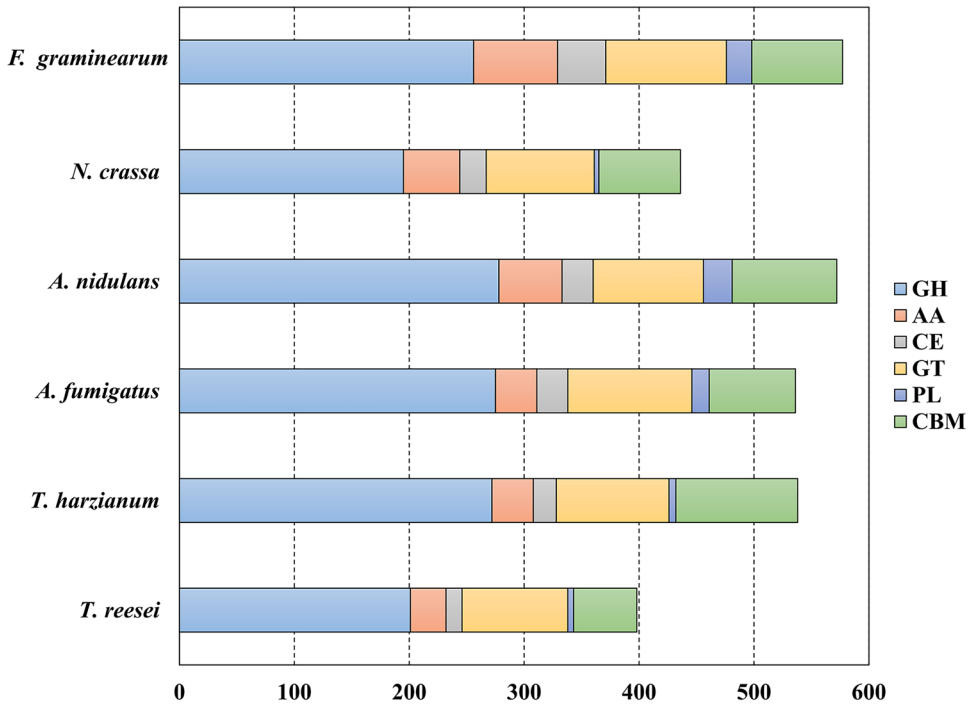

Fig. 2 Genomic annotation of CAZymes in different fungi of biotechnological interest

demand. Automated fed batch supply of substrates is necessary to avoid catabolite repression (Viniegra-González et al. 2003). SSF processes have been explored for protease production due to the higher yield achieved, and novel microorganisms are being tested (López-Gómez and Venus 2021; Benabda et al. 2019; Usman et al. 2021).

Fungi produce a wide variety of secondary metabolites that can be used in the pharmaceutical industry in particular, demonstrating the potential of genomic studies for the discovery of new secondary metabolites. Cheng et al. (2020) sequenced the genome of the fungus *Calcarisporium arbuscula* and found 65 clusters of secondary metabolite synthesis genes, including genes related to mycotoxin synthesis. Wei et al. (2020) identified a new molecule, pyranoviolin A, from the sequencing and analysis of clusters of genes from the *Aspergillus violaceofuscus* genome.

Software tools and scripts for data processing automatization are applied to improve genomic analyses, generating large amounts of processed data (Huber et al. 2015; Qu et al. 2016). In the stages of assembly, gene prediction, and automatic annotation, several developed software programs have been applied in each step of data processing (Table 1). For the processing of textual data in biological analysis, the most commonly used programming languages are Python and Perl, which have specific modules for bioinformatics analysis. For tabulation and statistical analysis, the main tools used are Excel and R tools.

Table 1 summarizes some bioinformatics tools for sequence assembly, gene prediction, and automatic annotation that can be used in a fungal genomic project. The choice of the tool will vary according to the objective of each project.

Transcriptomics applied in the selection of enzymes in fungi

Transcriptomic analysis is used to elucidate a specific cell response state. It is an important strategy for studying the expression of large gene sets under particular conditions. Fungi quickly adapt their metabolism to different environments, and the adaptation process is driven by coordinated gene expression responses. The set of expressed genes and their expression levels are directly related to the species and growth conditions, such as the carbon source, temperature, luminosity, and humidity. These characteristics make transcription analysis a precise tool for understanding the state of the cell as well as differences regarding transcriptional regulation in different states. It is essential to understand the transcriptome (the set of all transcribed RNAs) to interpret the functional elements of the genome and reveal the molecular constituents of cells and tissues in different stages of development (Bull et al. 2000; Wang et al. 2009).

Thus, transcriptomic analysis methods have evolved in recent decades from the analysis of expressed sequence tags (ESTs) to RNA-Seq, which is a high-throughput approach for deep transcriptome sequencing that can precisely determine all transcripts and their expression levels in a controlled state, while also providing insight into the regulatory mechanisms of gene expression control. From a transcription profile, the sets of fungal gene transcripts conferring

Table 1 Bioinformatics tools used for genome analysis

Software	Application	Operating system	Command line or graphic interface	References
SPAdes	De novo assembler	Linux	Command line	Nurk et al. (2013)
Celera assembler	De novo assembler	Unix system and Mac OS-X	Command line	Myers et al. (2000)
Pilon	Draft assembly improvement and variant detection	Unix System	Command line	Walker et al. (2014)
BWA	Error correction by low divergence alignment	Unix System	Command line	Li and Durbin (2009)
FGNESEH	Ab initio gene predictor	Windows and Unix system	Command line and Graphic interface	Solovyev et al. (2006)
AUGUSTUS	Ab initio gene predictor	Windows and Unix system	Command line and Graphic interface	Stanke and Morgenstern (2005)
MAKER	Genome annotator	Unix system and MAC OS-X	Command line	Cantarel et al. (2008)
LTR_FINDER	Repeat and transposable element annotator	Windows	Graphic interface	Xu and Wang (2007)
SignalP5.0	Signal peptide predictor	Windows	Graphic Interface	Armenteros et al. (2019)
BLAST*	alignment tool	Unix system, MAC OS-X and Windows	Command line and Graphic interface	Altschul et al. (1990)
HMMER3	Homology searches and alignment	Windows and Unix system	Command line and Graphic interface	Krogh et al. (1994)
InterProScan	Protein classification and functional analysis	Windows and Unix system	Command line and Graphic interface	Quevillon et al. (2005)
FunGAP	Annotator and evidence-based model predictor	Unix system	Command line	Min et al. (2017)
Artemis	Sequence or genome visualization, manual editing	Windows, Unix system and MAC OS-x	Command line and Graphic interface	Rutherford et al. (2000), Carver et al. (2005)

different enzymatic activities can be analyzed (Horta et al. 2014; Glass et al. 2013; dos Santos Castro et al. 2014), providing detailed information on the molecular mechanisms of enzyme production. This technology has also facilitated the genetic manipulation of strains to increase their production potential and the discovery of new enzymes or proteins (Borin et al. 2017; Zhao et al. 2018).

In fungi, ESTs provide initial information on the transcribed regions of genes. The first *Trichoderma* EST libraries were generated from *T. reesei* QM6a under biomass degradation conditions (Lorito et al. 2010). The successful application of ESTs in *T. harzianum* provided the first indication of gene expression in mycelium (Liu and Yang 2005) and later allowed the identification of genes with putative roles in mycoparasitism against *Fusarium solani* (Steindorff et al. 2012). Using ESTs, transcriptional processes have been explored in industrially relevant species; for example, glucose metabolism has been studied in the enzyme producer *T. reesei* (Chambergo et al. 2002), and genes with putative roles in crop contamination have been identified in *Aspergillus oryzae* (Akao et al. 2007) and *Aspergillus flavus* (Yu et al. 2004).

The arrival of next-generation sequencing (NGS) and RNA-Seq approaches has promoted many successful studies involving transcriptome analysis. These technologies produce millions of base pairs of sequencing reads in the form of short- to medium-length reads (30–400 bp reads, depending on the DNA sequencing technology) in a short time with low costs. Next-generation sequencing has accelerated RNA-Seq experiments, allowing studies of even more diverse conditions and species (Lorito et al. 2010; Wang et al. 2009). The transcriptomic field has developed very quickly, and RNA-Seq has become the preferred method for gene expression profiling (McGettigan 2013; Parkhomchuk et al. 2009; Corchete et al. 2020) by cDNA sequencing. The Illumina Genome Analyzer, HiSeq, MiSeq, and NextSeq platforms are some of the technologies for RNA-Seq that have become available since 2006. These NGS platforms are capable of paired-end sequencing, resulting in high coverage, large numbers of reads, and high-quality sequence data relative to single-end sequencing (Ambardar et al. 2016). Many studies have used RNA-Seq to elucidate enzymatic activities, protein structures, and synergistic reactions among enzymes (Horta et al. 2018; Santos et al. 2016; Mhuantong et al. 2021; Liu et al. 2021). On the basis of RNA-Seq data, genes encoding hydrolytic enzymes involved in plant cell wall degradation have been identified in different species of *Trichoderma*, providing important information for the selection of species and target genes for use in industrial enzyme production technologies (dos Santos Castro et al. 2014).

Transcriptomic studies on well-known filamentous fungi such as *T. reesei* and *Aspergillus niger* have been employed to improve enzymatic cocktail technologies, leading to the production of new enzymes and more efficient mutants to reduce costs involved in second-generation bioethanol production (de Gouvêa et al. 2018; de Souza et al. 2011; Borin et al. 2017; dos Santos Castro et al. 2014).

Many functionalities are currently available for transcriptomic analysis, ranging from specific programs to entire platforms dedicated solely to genomic analyses, including tools for RNA-Seq assembly, mapping, annotation, expression, and differential expression analyses, etc. Regarding mapping and quantification, various tools can be used, including reference genomes, spliced read aligners, unspliced read aligners, pseudoalignment and quasi-mapping, and various tools for the processing of all possible situations in RNA-Seq analysis can be found for (Costa-Silva et al. 2017). Table 2 lists the currently available tools that are commonly used for RNA-Seq data management.

Data integration: coregulation networks

Due to the development of molecular biology and high-throughput technologies, the use of systems biology and combinations of data sources from different omic levels to decipher the enormous network of cellular relationships has attracted increasing interest (Kitano 2002; Aderem 2005). Regarding the industrial enzymatic engineering of fungi, several studies encompassing this holistic perspective have been published (Kubicek 2013; Akcapinar and Sezerman 2016); however, the computational modeling of these interacting structures remains a challenge.

In recent decades, the use of complex networks has emerged as a powerful analytical tool for complex systems, with common applications in diverse areas of knowledge due to the common dynamics and architecture of such structures (Wilson 1999; Strogatz 2001). By using principles of graph theory, network science offers methodological resources for elucidating the dynamics and interactions of these complex systems (Barabási 2013), including molecular biology. When using omics data, there are different possible types of representations of cellular interactions through networks, with nodes (vertices) and links (edges) representing the components of the system and their relationships (e.g., proteins and their metabolic interactions, respectively) (Vazquez et al. 2003).

In fungal research, the first network approaches were mainly based on protein interactions for metabolic reconstruction and functional inferences using omic data and biological databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa 2002) and MetaCyc (Caspi et al. 2014). However, because of the high availability of data, other types of networks have now been employed (Koutrouli et al. 2020), such as protein–protein interaction networks, sequence similarity networks, gene regulatory networks, signal transduction networks, metabolic networks, and gene coexpression networks.

Several biological databases and tools assist with the construction of these structures (Mering et al. 2003; Stark et al. 2006; Ulrich and Zhulin 2007; Barrett et al. 2012; Caspi et al. 2014). The main challenge is the method to properly measure unknown relationships among genes, especially considering the potential presence of noise and nonlinear interactions (Koutrouli et al. 2020). A common and currently employed strategy to predict gene functions is to use

Table 2 Tools used for mapping and differential expression analysis in transcriptome studies

Software	Application	Operating system	Command line or graphic interface	References
HISAT2	Mapping reads	Linux	Command line	Kim et al. (2019)
STAR	Mapping reads	Linux and Mac OS-X	Command line	Dobin et al. (2013)
TopHat2	Mapping reads	Linux and Mac OS-X	Command line	Kim et al. (2013)
Salmon	Quantifying the expression of transcripts using RNA-Seq data	Linux	Command line	Patro et al. (2017)
Kallisto	Quantifying the abundance of transcripts	Linux and Mac OS-X	Command line	Bray et al. (2016)
BWA	Mapping reads	Unix system	Command line	Li and Durbin (2009)
BBMap	Short read aligner	Linux	Command line	Bushnell (2014)
Cufflinks	Transcriptome assembly and differential expression analysis for RNA-Seq	Unix system and MAC OS-X	Command line	Trapnell et al. (2012)
DESeq2	Bioconductor package to perform differential gene expression analysis	Windows, Linux and MAC OS-X	Graphic interface and command line	Love et al. (2014)
EdgeR	R package for differential expression analysis	Windows, Linux and MAC OS-X	Graphic interface and command line	Robinson et al. (2010)

coexpression and coregulation data, including coexpressed gene clusters (Lawler et al. 2013; Sieber et al. 2014), Bayesian inferences (Ferreira Filho et al. 2020; Sazal et al. 2020) and coexpression networks.

With the development of high-throughput technologies, including microarrays and RNA-Seq, large amounts of gene expression data are being continuously generated. Because of the large volume of transcriptome data, scalable methods are required to decipher novel relationships (Saha et al. 2017). Through gene coexpression networks, gene–gene expression dependencies can be measured and can supply insights into functional associations (Zhao et al. 2010). Genes with a similar expression pattern across multiple samples are grouped and often participate in the same biological process (Langfelder et al. 2013). By using this principle of guilt-by-association (Wolfe et al. 2005), a gene coexpression network analysis allows the prediction of the functions of unknown genes and, accordingly, the prediction of the pathways involving those genes.

Gene coexpression network construction is based on pairwise correlations between each possible pair of genes in a dataset. Some correlation measures that are widely used to compute pairwise correlations are Pearson's correlation and Spearman's correlation (Song et al. 2012; Usadel et al. 2009). Using the correlation matrix as a similarity measure, a network is constructed in which nodes represent genes that are connected to other genes by edges based on coexpression relationships. These correlation values can also be transformed to guarantee network properties, using, for example, weighted gene correlation network analysis (WGCNA) methodology (Langfelder et al. 2013), the highest reciprocal rank (HRR) method (Mutwill et al. 2011), correlation coefficient cutoffs (Burks and Azad 2016), and mutual ranks (Obayashi and Kinoshita 2009).

With this model structure, clustering approaches are generally employed to find groups of coexpressed genes using methods of hierarchical clustering coupled with a branch cutting method (Langfelder and Horvath 2008) for defining network modules and employing additional algorithms, such as MCODE (Xu and Hejzlar 2008), for defining more precise groups. Using the guilt-by-association principle, these modules can be interpreted through functional enrichment analysis to obtain biological insights about the coexpressed genes and their impact on molecular mechanisms. Generally, the Gene Ontology (GO) (Ashburner et al. 2000) and KEGG (Kanehisa and Goto 2000) databases are the most commonly used to characterize molecular functions and metabolic pathways, respectively.

In addition to the identification of genes acting as a module in a network, an important goal is to determine which gene(s) effectively represent the behavior of the module or may be important for module robustness (Koutrouli et al. 2020). For this purpose, network centrality measures are often employed (Barabási 2013), and the most common strategy is to calculate highly connected genes (hubs). These hub genes are the central nodes in a network structure (Koutrouli et al. 2020) and are considered the central players in this structure (Li et al. 2018), but they are not always directly associated with a trait of interest (Langfelder et al. 2013).

Following a scale-free topology criterion, the WGCNA methodology has been extensively used to infer novel functions. Instead of using binary information (connected = 1, unconnected = 0) with correlation cutoffs, WGCNA uses a 'soft' threshold to determine the weights of the edges connecting pairs of genes, which has been proven to yield more robust results than unweighted networks (Zhang and Horvath 2005).

Because of the improvement of 'omics' approaches used in biotechnological research, more studies are evaluating the molecular relationships among biological systems. For example, gene coexpression network analysis has been used in various biological contexts, such as human diseases (mostly cancer), plant biotechnology, plant-pathogen interactions, etc. This approach has also been applied in the study of hydrolytic microbial enzymes for industrial bioprocesses, including second-generation ethanol technology, as described below.

In a previous study, gene coexpression networks were inferred based on a transcriptome dataset of *A. niger* grown under two different experimental conditions (van den Berg et al. 2010). This approach allowed the prediction of biologically relevant modules and a search for enriched putative TF binding sites, which improved the understanding of higher-order regulatory structures in the studied fungi.

Other studies have reported the use of gene coexpression network analysis in this context. Using the transcriptomes of *T. reesei*, *T. harzianum*, and *T. atroviride* grown on cellulose and glucose, a coexpression network was inferred for each species (Horta et al. 2018). Based on the resulting data, a set of 80 genes shared among the three *Trichoderma* species was obtained, which could represent a common cellulose degradation system. Similarly, Almeida et al. (2021) constructed gene coexpression networks based on the transcriptome data of a novel strain of *T. harzianum* grown on cellulose and glucose. For this fungus, the results indicated that several genes may function in a coordinated manner during cellulose degradation, which revealed the capacity of *T. harzianum* as an enzyme producer (Almeida et al. 2021).

Additionally, for species of the filamentous genus *Trichoderma*, a gene coexpression network based on the transcriptome dataset of *T. harzianum* grown on cellulose and glucose was constructed to explore CLR2 regulator activity during biomass degradation (Ferreira Filho et al. 2020). In the same study, using gene expression data for secreted proteins, a Bayesian network of induced/repressed genes in *T.*

harzianum was inferred under cellulose-based growth conditions. The results revealed potential candidates for the validation of functions during the degradation of plant biomass in further studies.

Through WGCNA based on a transcriptome dataset of *T. reesei* grown on steam-exploded sugarcane bagasse, 28 highly connected gene modules were identified (Borin et al. 2018). One of these modules was enriched with the most representative core of cellulolytic enzymes, including their regulators and transporters. For several hub genes, DNA binding sites for the main activator of (hemi)cellulases, XYR1, were found in the promoter regions, which suggests a putative role of these hubs in bagasse cell wall breakdown (Borin et al. 2018).

To explore the biotechnological potential of *Trichoderma* spp., the WGCNA package was used to explore the genetic mechanisms related to the XYR1 and CRE1 TFs during cellulose degradation (Rosolen et al. 2020). For this purpose, the mycoparasite fungus *T. atroviride* and two mycoparasitic strains with the hydrolytic potential of *T. harzianum* were selected to model a network for each strain. Based on the transcriptome dataset obtained when the fungi were grown on cellulose and glucose, gene coexpression network analysis revealed that the strains developed different molecular mechanisms to control the regulation and the expression of genes encoding proteins related to cellulose degradation (Rosolen et al. 2020).

Recently, WGCNA was used to explore the molecular mechanisms used by *Penicillium oxalicum* during biomass degradation (Li et al. 2020a). Based on a transcriptome dataset for the fungus grown on two novel carbon sources, methylcellulose and 2-hydroxyethyl cellulose, 17 highly connected modules were generated. One of these modules included major cellulase- and xylanase-encoding genes of *P. oxalicum* as well as transcription factor- and transporter-encoding genes (Li et al. 2020a).

WGCNA is one of the most commonly used approaches for constructing and analyzing gene coexpression networks, it has recently been adapted for proteomics studies (Vella et al. 2017). Furthermore, a large-scale proteomics approach for investigating and comparing the enzymatic responses of the filamentous fungi *Aspergillus terreus*, *Trichoderma reesei*, *Myceliophthora thermophila*, *Neurospora crassa*, and *Phanerochaete chrysosporium* has been presented (Arntzen et al. 2020). In this study, these fungi were grown on five different substrates: grass (sugarcane bagasse), hardwood (birch), softwood (spruce), cellulose, and glucose. Gene coexpression networks were inferred for each fungus based on proteomics data and the use of the WGCNA package. The results allowed the identification of the adaptation profile of each fungus in response to the different substrates, and the results suggested the specificity of the CAZymes according to the substrate (Arntzen et al. 2020).

The gene coexpression networks of transcriptome and proteome data allow the application of computational simulations for predicting the behavior of the biological system over time or under different conditions. The knowledge acquired through these approaches paves the way for further experimental studies to validate gene function and to improve the use of filamentous fungi as enzyme producers in the biotechnology field.

In silico characterization of transcription factors related to bioprocesses

TFs control transcription by specifically binding to DNA sequences (Todeschini et al. 2014). Thus, studying these proteins can provide important resources for researchers who are interested in studying the regulation of gene expression. In this context, computational studies have been used to provide support for bioinformatics approaches for searching new targets across genomes. The obtained knowledge could be used as a basis for new biotechnological applications in the gene regulation field.

In filamentous fungi, the expression of the genes encoding plant biomass-degrading enzymes is tightly controlled by TFs, which function according to the carbon source available in the environment. In the presence of readily metabolizable sugars, such as glucose, fungal genes responsible for the expression of cellulolytic and hemicellulolytic enzymes are repressed, while in the presence of inductive carbon sources, such as cellulose, the expression of fungal genes encoding CAZymes is activated (de Paula et al. 2018).

Thus, the regulatory network responsible for cell wall deconstruction is regulated by several transcription factors, including the positive regulators XYR1 (Stricker et al. 2006), ACEII (Aro et al. 2001), LAE1 (Seiboth et al. 2012), BglR (Nitta et al. 2012), VEL1 (Karimi Aghcheh et al. 2014), and the HAP 2/3/5 complex (Zeilinger et al. 2001), and the negative regulators CRE1 (Portnoy et al. 2011), ACEI (Aro et al. 2003), and RCE1 (Cao et al. 2017). Additionally, holocellulose genes may also target other TFs, such as Xpp1 (Derntl et al. 2015), SxlR (Liu et al. 2017), and CRZ1 (Martins-Santana et al. 2020) (Fig. 3).

Due to the difficulty in the heterologous expression of some recombinant proteins in prokaryotic and eukaryotic hosts, the in silico characterization of TFs related to bioprocesses is useful to guide experiments and to provide new insights into the regulation of gene expression. These bioinformatics approaches have been reported to show great potential in exploring transcriptional regulation in important enzyme biofactories, particularly in industrial filamentous species, which is explored next.

The elucidation of the molecular mechanisms involved in the expression of genes encoding hydrolytic enzymes at the transcriptional level has attracted increasing interest to allow

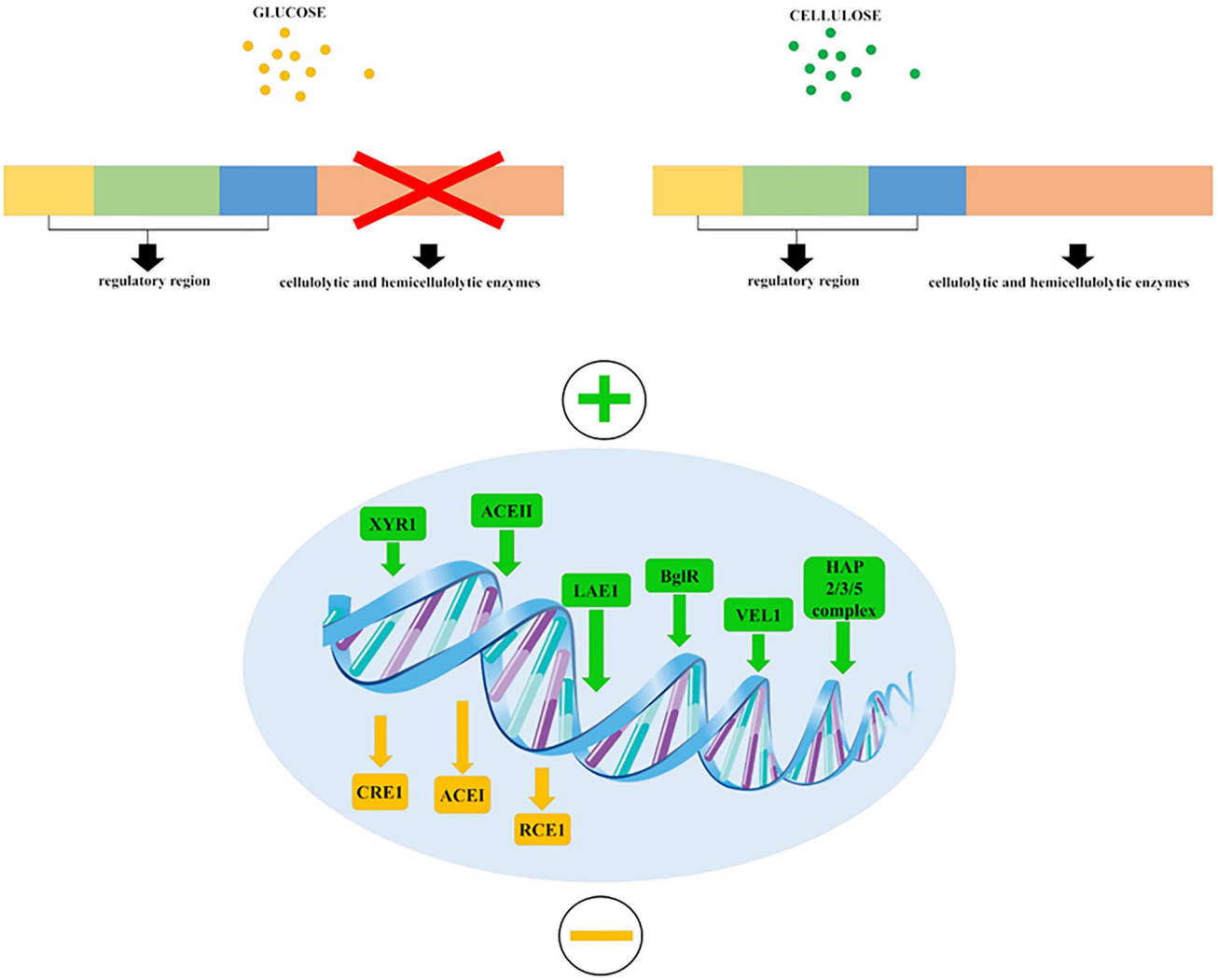

Fig. 3 TFs related to the expression of hydrolytic enzymes in fungi grown in the presence of glucose or cellulose

the rational engineering of fungi with increased enzyme production levels. These efforts might enable the identification of many regulatory proteins and signaling pathways responsible for the coordination of cellulase expression under different environmental conditions.

Regarding the main regulators of the expression of cellulase genes in *T. reesei*, the consensus binding sequences of XYR1 and CRE1 (5′-GGCWWW-3′ and 5′-SYGGRG-3′, respectively) were initially proposed based on comparisons with homologous regulators from other organisms (Kiesenhofer et al. 2018). However, these sequences cannot be used to distinguish between genes regulated by these regulators and those that are not (Furukawa et al. 2009).

The in silico identification of the cis-regulatory elements of XYR1 and CRE1 was subsequently reported. The authors identified potential binding sites for both regulatory proteins in 22 cellulase promoters and found that CRE1 affected *xyr1* expression (Silva-Rocha et al. 2014). Following a similar approach, a recent study investigated the role of the TF CRZ1 in *T. reesei* during biomass degradation. Using bioinformatics methods, CRZ1 binding motifs were identified in the promoter regions of genes encoding proteins related to plant cell wall depolymerization, such as cellulases, hemicellulases, sugar transporters, calcium transporters, and TFs. These findings are crucial for improving engineering attempts to construct new cellulase-responsive promoters and to understand the role of these regulators in *T. reesei* at the global scale.

Integrated transcriptomic data are also an option for identifying novel TFs related to the control of the gene expression of lignocellulosic hydrolases in filamentous fungi. In

a recent study, the authors used high-throughput data and a comparative genomics approach to identify novel potential TFs related to the control of hydrolytic enzyme-encoding genes in *T. reesei* and *A. nidulans* (Antonieto et al. 2019). The bioinformatics approach led to the characterization of an AZF1 homolog in *T. reesei* and the discovery of a novel Cys2His2-type zinc finger regulator involved in plant biomass deconstruction.

Identifying the genes regulated by a TF through in silico analysis is important for understanding not only the function of the regulatory protein but also the gene expression network of the cells. Although not all of these genes are considered direct target genes in the absence of experimental support, the insights provided by the bioinformatics approaches allow us to elucidate the biological reactions occurring at a molecular level, providing targets for further target-validation experiments.

Hypothetical proteins: approach for identification and characterization

Since the genomics era began in the 1990s, the identification of open reading frames (ORFs) whose products are annotated as hypothetical proteins has become increasingly common. In biochemistry, hypothetical proteins (also called "uncharacterized" or "unknown proteins") are by definition those for which amino acid residue comparisons according to sequence similarity with other gene/protein sequences deposited in a databank (GenBank, https://www.ncbi.nlm.nih.gov/genbank/, PDB, https://www.rcsb.org/) indicate that no accurate functions for such targets can be assigned and for which no evidence of in vivo existence is available (Sivashankari and Shanmughavel 2006). Interestingly, a quick look at genomic survey results reveals that ~ 50–70% of an organism's total genes generally have attributed biological roles, showing that a high percentage of unknown targets remain to be exploited (Bork 2000; Kawaji and Hayashizaki 2008; Armstrong et al. 2019).

Advances in the field of bioinformatics associated with studies using multiomic methodologies have contributed massively to the identification and functional assignment of hypothetical proteins (Subramanian et al. 2020). Because of these tools, it is possible to not only identify the expression profile of a target gene and the required metabolic conditions for its expression (transcriptome studies) but also to reveal whether the hypothetical protein is produced employing proteomic tools. The combination of such methodologies has enabled exciting discoveries in protein science (Pinu et al. 2019; Krassowski et al. 2020). However, the characterization of the function of a protein remains a bottleneck to be overcome in the postgenomic era.

Briefly, it is possible to enumerate up to five different strategies for the characterization of an unknown protein (Sivashankari and Shanmughavel 2006). The most accurate of these approaches is based on the resolution of the three-dimensional structure of a protein by X-ray diffraction studies or methodologies such as nuclear magnetic resonance spectroscopy (NMR) or single-particle cryoelectron microscopy (cryo-EM) (Ronda et al. 2015). There is a direct correlation between protein structure and function, so the high-resolution structure of a hypothetical protein can provide hints about its biochemical functions and reveal its biological role. Other approaches include methodologies based on protein–protein interactions (Marcotte et al. 1999). It is possible to infer the functions of proteins based on their interaction partners since proteins interact with one another to perform complementary functions. Thus, methods based on protein–protein interactions can be used to ensure the assignment of a protein role. At the genomic level, approaches based on comparative genomics and genome structure are also promising for protein function prediction (Pellegrini et al. 1999; Huynen et al. 2000). These methods focus on the relationship between genome structure and the function of a protein in a metabolic pathway, since it is expected that proteins that work together evolve together. The conservation of a gene neighborhood across different species (synteny) can also be exploited to identify protein function. Clustering approaches are useful as well, allowing the grouping of genes with a similar function in a specific cluster (Overbeek et al. 1999). Other modern protein function prediction methods can involve machine learning via high-throughput data analysis, which instead of focusing on specific protein domain prediction, addresses probability and employs accurate mathematical calculations to assign protein functions (Nakano et al. 2019; Jumper et al. 2021). Each of the approaches described here has advantages and disadvantages in the study of hypothetical proteins; however, regardless of the approach used, appropriate biochemical assays are always necessary to confirm a protein's function.

Interestingly, studies focusing on the identification and characterization of CAZymes routinely report the presence of several targets annotated as hypothetical conserved proteins (Horta et al. 2018). These targets, which are up- or downregulated according to the cultivation conditions of the studied organisms, indicate that the major puzzle of enzymatic biomass degradation is still incompletely understood. It is worth mentioning that the saccharification of biomass is a synergistic process; thus, not all key players in the process have been discovered and characterized, and corresponding biotechnological and industrial interventions have not yet reached their highest efficiency. In this context, there remains great potential for the future development and improvement of bioinformatics and computation tools for predicting and assigning the functions of nonidentified and/or unannotated proteins.

In addition to predicting the characteristics of unknown proteins, bioinformatics tools can be used for purposes such as in silico bioprospecting of genes/proteins. This method consists of discovering genes and metabolic pathways based solely on the large number of sequences found in databases, which may be used for biotechnological purposes (Ferrer et al. 2016). In silico bioprospecting can generally be divided into two stages: the search for databases and the use of bioinformatics tools to select and analyze potential candidates (Kamble et al. 2019).

Such techniques are possible due to the continuous decline in sequencing costs, which has led to the generation of massive amounts of information from highly diverse types of genomes. The combination of several tools and methods has facilitated the discovery of new enzymes with different functions and applications, in addition to assisting in their functional characterization (Gerlt 2017). Bioprospecting can also be performed with a combination of different approaches to obtain new bioproducts, such as newly designed molecular probes that can aid in prospecting enzymes or in building metagenomic libraries (Lee and Lee 2013). An example of the application of this technique is the discovery of new enzymes related to the degradation of lignocellulosic matter.

Several hydrolytic enzymes have been obtained via the application of metagenomics tools to noncultivable microorganisms (Berini et al. 2017; Popovic et al. 2015), as in the work of Salmeán et al. (2018), who identified CAZymes from metagenomics data from marine environments. In silico bioprospecting has been used to discover fungal esterases from the CAZyme database (Dilokpimol et al. 2018), and research groups have sought to develop computational tools for identifying secondary metabolite genes in microorganisms and plants (Medema 2018). This is possible due to the constant advances in bioinformatics, resulting in large amounts of available data on genomes, metabolic pathways, operons, omics information, gene regulatory networks, differential expression, protein characteristics, and structural data deposited in several databases (Fasim et al. 2021).

Application of bioinformatics in enzyme engineering

The integration of different data and bioinformatics approaches is a very useful tool for identifying new protein targets and for directing the study of protein engineering (Florindo et al. 2018; Lenfant et al. 2017; Zhou et al. 2019).

Genetic information from different levels must be integrated to implement a successful fungal engineering approach, ranging from the gene expression profile encoding an enzyme or natural product (NP) of interest to the structural properties of a certain protein. Thus, bioinformatics plays a central role in mining, integrating, and interpreting the data from these studies. Recently, significant advances in the identification, understanding, and engineering of fungal biosynthetic gene clusters (BGCs) have facilitated the biosynthesis of fungal NPs at the global, pathway, and enzyme levels using in vivo and in vitro approaches. It advances the progress in understanding how fungal BGCs are regulated, and the subsequent applications of these novel biosynthetic enzymes as biocatalysts. In 2019, Skellam (2019) described three principal categories for the methods used in fungal engineering of NPs as follows: (1) induction of transcriptional perturbations (e.g., through epigenetic modifications or the overexpression of global transcriptional regulators); (2) manipulating specific biosynthetic pathways either in the native host or in a heterologous host; and (3) the specific engineering of enzymes to synthesize novel NPs in vivo or in vitro. Each strategy varies according to the nature of the target compound and the specificities of the fungal species from which it was produced, leading to the discovery of new enzyme classes/functions, pathways, and heterologous expression systems, low-cost processes, and high-level production.

The biochemical characterization of proteins is essential for understanding the function and activity of a particular enzyme (Manavalan et al. 2017; Uechi et al. 2020; Dilokpimol et al. 2018). One approach that can be used in this type of study is to search RNA-Seq databases to identify key enzymes involved in a given process (Borin et al. 2017; Li et al. 2020b). In *T. harzianum* IOC-3844, studies on a β-glucosidase from the GH1 family (rThBgl) revealed high expression of the gene according to RNA-Seq data related to plant biomass degradation (Santos et al. 2016). This enzyme was cloned and expressed heterologously in *Escherichia coli,* and the product was subjected to structural crystallography. In another study, the accessory protein swollenin was identified in a cluster with CAZyme enzymes in BACs of *T. harzianum* IOC-3844 (Crucello et al. 2015). In this study, high synergism of swollenin with a xylanase from *T. viride* was found (Santos et al. 2017).

For a protein to be applied in a bioprocess many times, it needs to be adapted, which is known as protein engineering (DiTursi et al. 2006). This process involves genetically modified organisms and random or induced mutations in specific domains of the protein. In silico studies can direct this process according to the comparison of a protein with the desired properties and the target protein to be engineered; through multiple alignments, synthetic analysis and phylogeny, it is possible to design a tailor-made enzyme suitable for a given bioprocess (Pucci et al. 2017; Yang et al. 2019; Dilokpimol et al. 2018). Santos et al. (2019) performed the alignment of a highly glucose tolerant β-glucosidase from *Humicola insolens* (ideal for application in industrial cellulose degradation) with a low-tolerance enzyme from *T. harzianum* and found two amino acid residues responsible for this difference. Using the technique of site-directed mutation, it was then possible to obtain a β-glycosidase from *T. harzianum* with high glucose tolerance.

Conclusions

In this review, we present an overview of the major bioinformatics tools that could be applied for the bioprospection of hydrolytic enzymes and TFs related to cellulose and hemicellulose degradation. As fungi are producers of potential new enzymes, approaches including genomics, transcriptomics, proteomics, and systems biology are used to reveal the degradative mechanism employed by fungi, producing an unprecedented volume of biological data. We propose biological data integration as a methodological strategy that is useful for prospecting and producing enzymes appropriate for biotechnological process. The correlation of different types of data can contribute to a better understanding of how the expression of genes, enzymes, and regulators plays important roles in pathways or reactions of biotechnological interest that are suitable for improvement. The major challenge is to integrate data from different experiments or biological databases, which often have different formats and are not promptly correlated. Hence, we assume that data integration approaches will become increasingly sophisticated and accessible, thus facilitating the understanding and prediction of the actions of complex biological systems.

Acknowledgements Not applicable.

Author contributions APS and JAFF conceptualized the manuscript. JAFF, RRR, DAA, PHCA, MLLM, AHA, CAS, MACH and APS wrote the manuscript. JAFF, AHA and MLLM prepared the figures. JAFF and PHCA prepared the tables. CAS, MACH and APS revised the manuscript.

Funding This work was supported by grants from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP 2015/09202-0 and 2018/19660-4), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Biology Program 88882.160095/2013–01), and Conselho Nacional de Desenvolvimento Científico e Tecnológico for a Research Fellowship to APS (CNPq 312777/2018-3) and grants to CAS (CNPq Univeral 430350/2018-0). MACH received fellowship from FAPESP (2020/10536-9). AHA and RRR received PhD fellowships from FAPESP (2019/03232-6 and 2020/13420-1, respectively).

Data availability statement Not applicable as this is a review article that does not present any new data.

Declarations

Conflicts of interest The authors declare that they have no conflicts of interest.

Consent for publication All authors agree to the publication.

References

Abrashev R, Krumova E, Petrova P, Eneva R, Kostadinova N, Miteva-Staleva J, Engibarov S, Stoyancheva G, Gocheva Y, Kolyovska V (2021) Distribution of a novel enzyme of sialidase family among native filamentous fungi. Fungal Biol 125(5):412–425

Aderem A (2005) Systems biology: its practice and challenges. Cell 121(4):511–513

Akao T, Sano M, Yamada O, Akeno T, Fujii K, Goto K, Ohashi-Kunihiro S, Takase K, Yasukawa-Watanabe M, Yamaguchi K (2007) Analysis of expressed sequence tags from the fungus *Aspergillus oryzae* cultured under different conditions. DNA Res 14(2):47–57

Akcapinar GB, Sezerman OU (2016) Systems biological applications for fungal gene expression. In: Gene expression systems in fungi: advancements and applications. Springer, pp 385–393

Alberti F, Kaleem S, Weaver JA (2020) Recent developments of tools for genome and metabolome studies in basidiomycete fungi and their application to natural product research. Biol Open 9:12

Almeida DA, Horta MAC, Ferreira Filho JA, Murad NF, de Souza AP (2021) The synergistic actions of hydrolytic genes reveal the mechanism of *Trichoderma harzianum* for cellulose degradation. J Biotechnol 334:1–10

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q (2020) Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21(1):1–16

Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J (2016) High throughput sequencing: an overview of sequencing chemistry. Indian J Microbiol 56(4):394–404

Anasonye F, Winquist E, Kluczek-Turpeinen B, Räsänen M, Salonen K, Steffen KT, Tuomela M (2014) Fungal enzyme production and biodegradation of polychlorinated dibenzo-p-dioxins and dibenzofurans in contaminated sawmill soil. Chemosphere 110:85–90. https://doi.org/10.1016/j.chemosphere.2014.03.079

Antonieto ACC, Nogueira KMV, de Paula RG, Nora LC, Cassiano MHA, Guazzaroni M-E, Almeida F, da Silva TA, Ries LNA, de Assis LJ, Goldman GH, Silva RN, Silva-Rocha R (2019) A novel Cys2His2 zinc finger homolog of AZF1 modulates holocellulase expression in *Trichoderma reesei*. mSystems 4(4):e00161-00119. https://doi.org/10.1128/mSystems.00161-19

Armenteros JJA, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature Biotechnol 37(4):420–423

Armstrong J, Fiddes IT, Diekhans M, Paten B (2019) Whole-genome alignment and comparative annotation. Annu Rev Animal Biosci 7:41–64

Arntzen MØ, Bengtsson O, Várnai A, Delogu F, Mathiesen G, Eijsink VG (2020) Quantitative comparison of the biomass-degrading enzyme repertoires of five filamentous fungi. Sci Rep 10(1):1–17

Aro N, Saloheimo A, Ilmén M, Penttilä M (2001) ACEII, a novel transcriptional activator involved in regulation of cellulase and xylanase genes of *Trichoderma reesei*. J Biol Chem 276(26):24309–24314. https://doi.org/10.1074/jbc.M003624200

Aro N, Ilmén M, Saloheimo A, Penttilä M (2003) ACEI of *Trichoderma reesei* is a repressor of cellulase and xylanase expression. Appl Environ Microbiol 69(1):56. https://doi.org/10.1128/AEM.69.1.56-65.2003

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25–29

Barabási A-L (2013) Network science. Philos Trans R Soc Math Phys Eng Sci 371(1987):20120375

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M (2012) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41(D1):D991–D995

Barrett K, Jensen K, Meyer AS, Frisvad JC, Lange L (2020) Fungal secretome profile categorization of CAZymes by function and family corresponds to fungal phylogeny and taxonomy: example *Aspergillus* and *Penicillium*. Sci Rep 10(1):1–12

Basenko EY, Pulman JA, Shanmugasundram A, Harb OS, Crouch K, Starns D, Warrenfeltz S, Aurrecoechea C, Stoeckert CJ, Kissinger JC (2018) FungiDB: an integrated bioinformatic resource for fungi and oomycetes. J Fungi 4(1):39

Batista TM, Hilario HO, de Brito GAM, Moreira RG, Furtado C, de Menezes GCA, Rosa CA, Rosa LH, Franco GR (2020) Whole-genome sequencing of the endemic Antarctic fungus *Antarctomyces pellizariae* reveals an ice-binding protein, a scarce set of secondary metabolites gene clusters and provides insights on Thelebolales phylogeny. Genomics 112(5):2915–2921

Benabda O, M'hir S, Kasmi M, Mnif W, Hamdi M (2019) Optimization of protease and amylase production by *Rhizopus oryzae* cultivated on bread waste using solid-state fermentation. J Chem 2019:3738181. https://doi.org/10.1155/2019/3738181

Berini F, Casciello C, Marcone GL, Marinelli F (2017) Metagenomics: novel enzymes from non-culturable microbes. FEMS Microbiol Lett 364(21):fnx211

Bohra V, Dafale NA, Purohit HJ (2018) *Paenibacillus polymyxa* ND25: candidate genome for lignocellulosic biomass utilization. 3 Biotech 8(5):1–7

Borin GP, Sanchez CC, de Santana ES, Zanini GK, Dos Santos RAC, de Oliveira PA, de Souza AT, Dal RMMTS, Riaño-Pachón DM, Goldman GH (2017) Comparative transcriptome analysis reveals different strategies for degradation of steam-exploded sugarcane bagasse by *Aspergillus niger* and *Trichoderma reesei*. BMC Genomics 18(1):1–21

Borin GP, Carazzolle MF, Dos Santos RAC, Riaño-Pachón DM, Oliveira JVdC (2018) Gene co-expression network reveals potential new genes related to sugarcane bagasse degradation in *Trichoderma reesei* RUT-30. Front Bioeng Biotechnol 6:151

Bork P (2000) Powers and pitfalls in sequence analysis: the 70% hurdle. Genome Res 10(4):398–400

Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34(5):525–527

Bull AT, Ward AC, Goodfellow M (2000) Search and discovery strategies for biotechnology: the paradigm shift. Microbiol Mol Biol Rev 64(3):573–606

Burks DJ, Azad RK (2016) Identification and network-enabled characterization of auxin response factor genes in *Medicago truncatula*. Front Plant Sci 7:1857

Bushnell B (2014) BBMap: a fast, accurate, splice-aware aligner. In: Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States)

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18(1):188–196

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The carbohydrate-active enzymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res 37(Suppl 1):D233–D238

Cao Y, Zheng F, Wang L, Zhao G, Chen G, Zhang W, Liu W (2017) Rce1, a novel transcriptional repressor, regulates cellulase gene expression by antagonizing the transactivator Xyr1 in *Trichoderma reesei*. Mol Microbiol 105(1):65–83. https://doi.org/10.1111/mmi.13685

Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, Parkhill J (2005) ACT: the Artemis comparison tool. Bioinformatics 21(16):3422–3423

Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 42(D1):D459–D471

Chambergo FS, Bonaccorsi ED, Ferreira AJ, Ramos AS, Junior JRF, Farah JPS, El-Dorry H, Abrahão-Neto J (2002) Elucidation of the metabolic fate of glucose in the filamentous fungus *Trichoderma reesei* using expressed sequence tag (EST) analysis and cDNA microarrays. J Biol Chem 277(16):13983–13988

Chen Q, Zobel J, Verspoor K (2017) Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. Database 2017:baw163

Chen Q, Britto R, Erill I, Jeffery CJ, Liberzon A, Magrane M, Onami J-i, Robinson-Rechavi M, Sponarova J, Zobel J (2020) Quality matters: biocuration experts on the impact of duplication and other data quality issues in biological databases. Genomics Proteomics Bioinform 18(2):91

Cheng J-T, Cao F, Chen X-A, Li Y-Q, Mao X-M (2020) Genomic and transcriptomic survey of an endophytic fungus *Calcarisporium arbuscula* NRRL 3705 and potential overview of its secondary metabolites. BMC Genomics 21(1):1–13

Corchete LA, Rojas EA, Alonso-López D, De Las RJ, Gutiérrez NC, Burguillo FJ (2020) Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Sci Rep 10(1):19737. https://doi.org/10.1038/s41598-020-76881-x

Costa-Silva J, Domingues D, Lopes FM (2017) RNA-Seq differential expression analysis: an extended review and a software tool. PLoS ONE 12(12):e0190152

Crucello A, Sforça DA, Horta MAC, dos Santos CA, Viana AJC, Beloti LL, de Toledo MAS, Vincentz M, Kuroshu RM, de Souza AP (2015) Analysis of genomic regions of *Trichoderma harzianum* IOC-3844 related to biomass degradation. PLoS ONE 10(4):e0122122

de Gouvêa PF, Bernardi AV, Gerolamo LE, de Souza SE, Riaño-Pachón DM, Uyemura SA, Dinamarco TM (2018) Transcriptome and secretome analysis of *Aspergillus fumigatus* in the presence of sugarcane bagasse. BMC Genomics 19(1):1–18

de Paula RG, Antoniêto ACC, Ribeiro LFC, Carraro CB, Nogueira KMV, Lopes DCB, Silva AC, Zerbini MT, Pedersoli WR, Costa MdN, Silva RN (2018) New genomic approaches to enhance biomass degradation by the industrial fungus *Trichoderma reesei*. Int J Genomics 2018:1974151. https://doi.org/10.1155/2018/1974151

de Souza WR, de Gouvea PF, Savoldi M, Malavazi I, de Souza Bernardes LA, Goldman MHS, de Vries RP, de Castro Oliveira JV, Goldman GH (2011) Transcriptome analysis of *Aspergillus niger* grown on sugarcane bagasse. Biotechnol Biofuels 4(1):1–17

Derntl C, Rassinger A, Srebotnik E, Mach RL, Mach-Aigner AR (2015) Xpp1 regulates the expression of xylanases, but not of cellulases in *Trichoderma reesei*. Biotechnol Biofuels 8(1):112. https://doi.org/10.1186/s13068-015-0298-8

Dilokpimol A, Mäkelä MR, Cerullo G, Zhou M, Varriale S, Gidijala L, Brás JL, Jütten P, Piechot A, Verhaert R (2018) Fungal glucuronoyl esterases: genome mining based enzyme discovery and biochemical characterization. New Biotechnol 40:282–287

Ding L, Rath E, Bai Y (2017) Comparison of alternative splicing junction detection tools using RNASeq data. Curr Genomics 18(3):268–277

Diniz W, Canduri F (2017) Bioinformatics: an overview and its applications. Genet Mol Res 16(1):17

DiTursi MK, Kwon S-J, Reeder PJ, Dordick JS (2006) Bioinformatics-driven, rational engineering of protein thermostability. Protein Eng Des Sel 19(11):517–524

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21

dos Santos CL, Pedersoli WR, Antoniêto ACC, Steindorff AS, Silva-Rocha R, Martinez-Rossi NM, Rossi A, Brown NA, Goldman GH, Faça VM (2014) Comparative metabolism of cellulose, sophorose and glucose in *Trichoderma reesei* using high-throughput genomic and proteomic analyses. Biotechnol Biofuels 7(1):1–18

Druzhinina IS, Kubicek CP (2017) Genetic engineering of *Trichoderma reesei* cellulases and their production. Microb Biotechnol 10(6):1485–1499. https://doi.org/10.1111/1751-7915.12726

Ellison CE, Kowbel D, Glass NL, Taylor JW, Brem RB (2014) Discovering functions of unannotated genes from a transcriptome survey of wild fungal isolates. Mbio 5:2

Faksri K, Tan JH, Chaiprasert A, Teo Y-Y, Ong RT-H (2016) Bioinformatics tools and databases for whole genome sequence analysis of *Mycobacterium tuberculosis*. Infect Genet Evol 45:359–368

Fasim A, More VS, More SS (2021) Large-scale production of enzymes for biotechnology uses. Curr Opin Biotechnol 69:68–76

Fedorova ND, Khaldi N, Joardar VS, Maiti R, Amedeo P, Anderson MJ, Crabtree J, Silva JC, Badger JH, Albarraq A (2008) Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. PLoS Genet 4(4):e1000046

Ferreira Filho JA, Horta MAC, Beloti LL, Dos Santos CA, de Souza AP (2017) Carbohydrate-active enzymes in *Trichoderma harzianum*: a bioinformatic analysis bioprospecting for key enzymes for the biofuels industry. BMC Genomics 18(1):1–12

Ferreira Filho JA, Horta MAC, Dos Santos CA, Almeida DA, Murad NF, Mendes JS, Sforça DA, Silva CBC, Crucello A, de Souza AP (2020) Integrative genomic analysis of the bioprospection of regulators and accessory enzymes associated with cellulose degradation in a filamentous fungus (*Trichoderma harzianum*). BMC Genomics 21(1):1–14

Ferrer M, Martínez-Martínez M, Bargiela R, Streit WR, Golyshina OV, Golyshin PN (2016) Estimating the success of enzyme bioprospecting through metagenomics: current status and future trends. Microb Biotechnol 9(1):22–34

Florindo RN, Souza VP, Mutti HS, Camilo C, Manzine LR, Marana SR, Polikarpov I, Nascimento AS (2018) Structural insights into β-glucosidase transglycosylation based on biochemical, structural and computational analysis of two GH1 enzymes from *Trichoderma harzianum*. New Biotechnol 40:218–227

Furukawa T, Shida Y, Kitagami N, Mori K, Kato M, Kobayashi T, Okada H, Ogasawara W, Morikawa Y (2009) Identification of specific binding sites for XYR1, a transcriptional activator of cellulolytic and xylanolytic genes in *Trichoderma reesei*. Fungal Genet Biol 46(8):564–574. https://doi.org/10.1016/j.fgb.2009.04.001

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma L-J, Smirnov S, Purcell S (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422(6934):859–868

Geniza M, Jaiswal P (2017) Tools for building de novo transcriptome assembly. Curr Plant Biol 11:41–45

Gerlt JA (2017) Genomic enzymology: web tools for leveraging protein family sequence–function space and genome context to discover novel functions. Biochemistry 56(33):4293–4308

Giani AM, Gallo GR, Gianfranceschi L, Formenti G (2020) Long walk to genomics: history and current approaches to genome sequencing and assembly. Comput Struct Biotechnol J 18:9–19

Glass NL, Schmoll M, Cate JH, Coradetti S (2013) Plant cell wall deconstruction by ascomycete fungi. Annu Rev Microbiol 67:477–498

Gujar VV, Fuke P, Khardenavis AA, Purohit HJ (2018) Draft genome sequence of *Penicillium chrysogenum* strain HKF2, a fungus with potential for production of prebiotic synthesizing enzymes. 3 Biotech 8(2):1–5

Gurjar MS, Aggarwal R, Jogawat A, Kulshreshtha D, Sharma S, Solanke AU, Dubey H, Jain RK (2019) De novo genome sequencing and secretome analysis of *Tilletia indica* inciting Karnal bunt of wheat provides pathogenesis-related genes. 3 Biotech 9(6):1–11

Guzmán-Chávez F, Zwahlen RD, Bovenberg RA, Driessen AJ (2018) Engineering of the filamentous fungus *Penicillium chrysogenum* as cell factory for natural products. Front Microbiol 9:2768

Horta MAC, Vicentini R, da Silva DP, Laborda P, Crucello A, Freitas S, Kuroshu RM, Polikarpov I, da Cruz Pradella JG, Souza AP (2014) Transcriptome profile of *Trichoderma harzianum* IOC-3844 induced by sugarcane bagasse. PLoS ONE 9(2):e88689

Horta MAC, Ferreira Filho JA, Murad NF, de Oliveira SE, Dos Santos CA, Mendes JS, Brandão MM, Azzoni SF, de Souza AP (2018) Network of proteins, enzymes and genes linked to biomass degradation shared by *Trichoderma* species. Sci Rep 8(1):1–11

Huang L, Zhang H, Wu P, Entwistle S, Li X, Yohe T, Yi H, Yang Z, Yin Y (2018) dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation. Nucleic Acids Res 46(D1):D516–D521

Huang X, Men P, Tang S, Lu X (2021) *Aspergillus terreus* as an industrial filamentous fungus for pharmaceutical biotechnology. Curr Opin Biotechnol 69:273–280. https://doi.org/10.1016/j.copbio.2021.02.004

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T (2015) Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 12(2):115

Hurley D, Araki H, Tamada Y, Dunmore B, Sanders D, Humphreys S, Affara M, Imoto S, Yasuda K, Tomiyasu Y (2012) Gene network inference and visualization tools for biologists: application to new human transcriptome datasets. Nucleic Acids Res 40(6):2377–2398

Huynen M, Snel B, Lathe W, Bork P (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 10(8):1204–1210

Jauhal AA, Newcomb RD (2021) Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. Mol Ecol Resour 21(5):1416–1421. https://doi.org/10.1111/1755-0998.13364

Jhalia V, Swarnkar T (2021) A critical review on the application of artificial neural network in bioinformatics. Data Anal Bioinform A Mach Learn Perspect 2021:51–76

Jouzani GS, Tabatabaei M, Aghbashlo M (2020) Fungi in fuel biotechnology. Springer, Berlin

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2

Kamble A, Srinivasan S, Singh H (2019) In-silico bioprospecting: finding better enzymes. Mol Biotechnol 61(1):53–59

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30

Kanehisa M (2002) The KEGG database. In: Novartis Foundation Symposium, 2002. Wiley Online Library, pp 91–100
Karimi Aghcheh R, Németh Z, Atanasova L, Fekete E, Paholcsek M, Sándor E, Aquino B, Druzhinina IS, Karaffa L, Kubicek CP (2014) The VELVET A orthologue VEL1 of *Trichoderma reesei* regulates fungal development and is essential for cellulase gene expression. PLoS ONE 9(11):e112799. https://doi.org/10.1371/journal.pone.0112799
Kawaji H, Hayashizaki Y (2008) Genome annotation. Bioinformatics 2:125–139
Kiesenhofer DP, Mach RL, Mach-Aigner AR (2018) Influence of cis element arrangement on promoter strength in *Trichoderma reesei*. Appl Environ Microbiol 84(1):e01742-e11717. https://doi.org/10.1128/AEM.01742-17
Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):1–13
Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37(8):907–915
Kitano H (2002) Systems biology: a brief overview. Science 295(5560):1662–1664
Koutrouli M, Karatzas E, Paez-Espino D, Pavlopoulos GA (2020) A guide to conquer the biological network era using graph theory. Front Bioeng Biotechnol 8:34
Krassowski M, Das V, Sahu SK, Misra BB (2020) State of the Field in multi-omics research: from computational needs to data mining and sharing. Front Genet 2020:1
Krogh A, Brown M, Mian IS, Sjölander K, Haussler D (1994) Hidden Markov models in computational biology: applications to protein modeling. J Mol Biol 235(5):1501–1531
Kubicek CP (2013) Systems biological approaches towards understanding cellulase production by *Trichoderma reesei*. J Biotechnol 163(2):133–142
Lange L, Barrett K, Meyer AS (2021) New method for identifying fungal kingdom enzyme hotspots from genome sequences. J Fungi 7(3):207
Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinform 9(1):1–13
Langfelder P, Mischel PS, Horvath S (2013) When is hub gene selection better than standard meta-analysis? PLoS ONE 8(4):e61505
Lawler K, Hammond-Kosack K, Brazma A, Coulson RM (2013) Genomic clustering and co-regulation of transcriptional networks in the pathogenic fungus *Fusarium graminearum*. BMC Syst Biol 7(1):1–16
Lee MH, Lee S-W (2013) Bioprospecting potential of the soil metagenome: novel enzymes and bioactivities. Genomics Inform 11(3):114
Lenfant N, Hainaut M, Terrapon N, Drula E, Lombard V, Henrissat B (2017) A bioinformatics analysis of 3400 lytic polysaccharide oxidases from family AA9. Carbohyd Res 448:166–174
Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B (2013) Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol Biofuels 6(1):1–14
Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760
Li W-C, Huang C-H, Chen C-L, Chuang Y-C, Tung S-Y, Wang T-F (2017) *Trichoderma reesei* complete genome sequence, repeat-induced point mutation, and partitioning of CAZyme gene clusters. Biotechnol Biofuels 10(1):1–20
Li J, Zhou D, Qiu W, Shi Y, Yang J-J, Chen S, Wang Q, Pan H (2018) Application of weighted gene co-expression network analysis for data from paired design. Sci Rep 8(1):1–8
Li C-X, Zhao S, Luo X-M, Feng J-X (2020a) Weighted gene co-expression network analysis identifies critical genes for the production of cellulase and xylanase in *Penicillium oxalicum*. Front Microbiol 11:520
Li J-X, Zhang F, Jiang D-D, Li J, Wang F-L, Zhang Z, Wang W, Zhao X-Q (2020b) Diversity of cellulase-producing filamentous fungi from tibet and transcriptomic analysis of a superior cellulase producer *Trichoderma harzianum* LZ117. Front Microbiol 11:1617
Liu P-g, Yang Q (2005) Identification of genes with a biocontrol function in *Trichoderma harzianum* mycelium using the expressed sequence tag approach. Res Microbiol 156(3):416–423
Liu R, Chen L, Jiang Y, Zou G, Zhou Z (2017) A novel transcription factor specifically regulates GH11 xylanase genes in *Trichoderma reesei*. Biotechnol Biofuels 10(1):194. https://doi.org/10.1186/s13068-017-0878-x
Liu S, Wang H, Tian P, Yao X, Sun H, Wang Q, Delgado-Baquerizo M (2020) Decoupled diversity patterns in bacteria and fungi across continental forest ecosystems. Soil Biol Biochem 144:107763
Liu H, Wu H, Wang Y, Wang H, Chen S, Yin Z (2021) Comparative transcriptome profiling and co-expression network analysis uncover the key genes associated withearly-stage resistance to *Aspergillus flavus* in maize. BMC Plant Biol 21(1):216. https://doi.org/10.1186/s12870-021-02983-x
Lopes AMM, de Mélo AHF, Procopio DP, Teixeira GS, Carazzolle MF, de Carvalho LM, Adelantado N, Pereira GA, Ferrer P, Maugeri Filho F (2020) Genome sequence of *Acremonium strictum* AAJ6 strain isolated from the Cerrado biome in Brazil and CAZymes expression in thermotolerant industrial yeast for ethanol production. Process Biochem 98:139–150
López-Gómez JP, Venus J (2021) Potential role of sequential solid-state and submerged-liquid fermentations in a circular bioeconomy. Fermentation 7:2. https://doi.org/10.3390/fermentation7020076
Lorito M, Woo SL, Harman GE, Monte E (2010) Translational research on *Trichoderma*: from'omics to the field. Annu Rev Phytopathol 48:395–417
Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):1–21
Luecken MD, Theis FJ (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol 15(6):8746
Machida M, Asai K, Sano M, Tanaka T, Kumagai T, Terai G, Kusumoto K-I, Arima T, Akita O, Kashiwagi Y (2005) Genome sequencing and analysis of *Aspergillus oryzae*. Nature 438(7071):1157–1161
Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R (2018) Nanopore sequencing data analysis: state of the art, applications and challenges. Brief Bioinform 19(6):1256–1272
Maharachchikumbura SS, Wanasinghe DN, Cheewangkoon R, Al-Sadi AM (2021) Uncovering the hidden taxonomic diversity of fungi in Oman. Fungal Divers 2021:1–40
Manavalan T, Liu R, Zhou Z, Zou G (2017) Optimization of acetyl xylan esterase gene expression in *Trichoderma reesei* and its application to improve the saccharification efficiency on different biomasses. Process Biochem 58:160–166
Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285(5428):751–753
Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, Chapman J, Chertkov O, Coutinho PM, Cullen D (2008a) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. Hypocrea jecorina). Nature Biotechnol 26(5):553–560
Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, Chapman J, Chertkov O, Coutinho PM, Cullen D, Danchin EGJ, Grigoriev IV, Harris P, Jackson M, Kubicek CP, Han CS, Ho I, Larrondo LF, de Leon AL, Magnuson JK, Merino S, Misra M, Nelson B, Putnam N, Robbertse B, Salamov AA, Schmoll M,

Terry A, Thayer N, Westerholm-Parvinen A, Schoch CL, Yao J, Barabote R, Nelson MA, Detter C, Bruce D, Kuske CR, Xie G, Richardson P, Rokhsar DS, Lucas SM, Rubin EM, Dunn-Coleman N, Ward M, Brettin TS (2008b) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. Hypocrea jecorina). Nature Biotechnol 26(5):553–560. https://doi.org/10.1038/nbt1403

Martins-Santana L, Nora LC, Sanches-Medeiros A, Lovate GL, Cassiano MH, Silva-Rocha R (2018) Systems and synthetic biology approaches to engineer fungi for fine chemical production. Front Bioeng Biotechnol 6:117

Martins-Santana L, Paula RGd, Silva AG, Lopes DCB, Silva RdN, Silva-Rocha R (2020) CRZ1 regulator and calcium cooperatively modulate holocellulases gene expression in *Trichoderma reesei* QM6a. Genet Mol Biol 43(2):e20190244–e20190244. https://doi.org/10.1590/1678-4685-GMB-2019-0244

McGettigan PA (2013) Transcriptomics in the RNA-seq era. Curr Opin Chem Biol 17(1):4–11

Medema MH (2018) Computational genomics of specialized metabolism: from natural product discovery to microbiome ecology. Msystems 3:2

Mering CV, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 31(1):258–261

Mhuantong W, Charoensri S, Poonsrisawat A, Pootakham W, Tangphatsornruang S, Siamphan C, Suwannarangsee S, Eurwilaichitr L, Champreda V, Charoensawan V, Chantasingh D (2021) High quality aspergillus aculeatus genomes and transcriptomes: a platform for cellulase activity optimization toward industrial applications. Front Bioeng Biotechnol 1594:8. https://doi.org/10.3389/fbioe.2020.607176

Milić D, Veprintsev DB (2015) Large-scale production and protein engineering of G protein-coupled receptors for structural studies. Front Pharmacol 6:66

Min B, Grigoriev IV, Choi I-G (2017) FunGAP: fungal genome annotation pipeline using evidence-based gene model evaluation. Bioinformatics 33(18):2936–2937

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA (2000) A whole-genome assembly of *Drosophila*. Science 287(5461):2196–2204

Nakano FK, Lietaert M, Vens C (2019) Machine learning for discovering missing or wrong protein function annotations. BMC Bioinform 20(1):1–32

Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, Berriman M, Abe K, Archer DB, Bermejo C (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. Nature 438(7071):1151–1156

Nitta M, Furukawa T, Shida Y, Mori K, Kuhara S, Morikawa Y, Ogasawara W (2012) A new Zn(II)(2)Cys(6)-type transcription factor BglR regulates β-glucosidase expression in *Trichoderma reesei*. Fungal Genet Biol 49(5):388–397. https://doi.org/10.1016/j.fgb.2012.02.009

Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelsky A, Pyshkin A, Sirotkin A, Sirotkin Y (2013) Assembling genomes and mini-metagenomes from highly chimeric reads. In: Annual international conference on research in computational molecular biology, 2013. Springer, pp 158–170

Obayashi T, Kinoshita K (2009) Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. DNA Res 16(5):249–260

Orlov YL, Baranova AV (2020) bioinformatics of genome regulation and systems biology. Front Genet 2020:11

Overbeek R, Fonstein M, D'souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci 96(6):2896–2901

Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res 37(18):e123–e123

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14(4):417–419

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci 96(8):4285–4288

Pereira R, Oliveira J, Sousa M (2020) Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. J Clin Med 9(1):132

Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. Nature 556(7701):339–344

Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, Wishart D (2019) Systems biology and multi-omics integration: viewpoints from the metabolomics research community. Metabolites 9(4):76

Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. Trends Genet 24(3):142–149

Popovic A, Tchigvintsev A, Tran H, Chernikova TN, Golyshina OV, Yakimov MM, Golyshin PN, Yakunin AF (2015) Metagenomics as a tool for enzyme discovery: hydrolytic enzymes from marine-related metagenomes. Prokaryotic Syst Biol 2015:1–20

Portnoy T, Margeot A, Linke R, Atanasova L, Fekete E, Sándor E, Hartl L, Karaffa L, Druzhinina IS, Seiboth B, Le Crom S, Kubicek CP (2011) The CRE1 carbon catabolite repressor of the fungus *Trichoderma reesei*: a master regulator of carbon assimilation. BMC Genomics 12(1):269. https://doi.org/10.1186/1471-2164-12-269

Pramesh D, Prasannakumar MK, Muniraju KM, Mahesh H, Pushpa H, Manjunatha C, Saddamhusen A, Chidanandappa E, Yadav MK, Kumara MK (2020) Comparative genomics of rice false smut fungi *Ustilaginoidea virens* Uv-Gvt strain from India reveals genetic diversity and phylogenetic divergence. 3 Biotech 10(8):1–14

Pucci F, Kwasigroch JM, Rooman M (2017) SCooP: an accurate and fast predictor of protein stability curves as a function of temperature. Bioinformatics 33(21):3415–3422

Qu K, Garamszegi S, Wu F, Thorvaldsdottir H, Liefeld T, Ocana M, Borges-Rivera D, Pochet N, Robinson JT, Demchak B (2016) Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace. Nat Methods 13(3):245–247

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33(2):W116–W120

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140

Ronda L, Bruno S, Bettati S, Storici P, Mozzarelli A (2015) From protein structure to function via single crystal optical spectroscopy. Front Mol Biosci 2:12

Rosolen RR, Aono AH, Almeida DA, Ferreira Filho JA, Horta MAC, de Souza AP (2020) Comparative gene coexpression networks analysis reveals different strategies of *Trichoderma* spp. associated with XYR1 and CRE1 during cellulose degradation. bioRxiv preprint

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, Barrell B (2000) Artemis: sequence visualization and annotation. Bioinformatics 16(10):944–945

Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinforms 23(19):2507–2517

Saha A, Kim Y, Gewirtz AD, Jo B, Gao C, McDowell IC, Engelhardt BE, Battle A, Aguet F, Ardlie KG (2017) Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Res 27(11):1843–1858

Saldarriaga-Hernández S, Velasco-Ayala C, Flores PL-I, de Jesús Rostro-Alanis M, Parra-Saldivar R, Iqbal HM, Carrillo-Nieves D (2020) Biotransformation of lignocellulosic biomass into industrially relevant products with the aid of fungi-derived lignocellulolytic enzymes. Int J Biol Macromol 161:1099–1116

Salmeán AA, Guillouzo A, Duffieux D, Jam M, Matard-Mann M, Larocque R, Pedersen HL, Michel G, Czjzek M, Willats WG (2018) Double blind microarray-based polysaccharide profiling enables parallel identification of uncharacterized polysaccharides and carbohydrate-binding proteins with unknown specificities. Sci Rep 8(1):1–11

Santos CA, Zanphorlin LM, Crucello A, Tonoli CC, Ruller R, Horta MA, Murakami MT, de Souza AP (2016) Crystal structure and biochemical characterization of the recombinant ThBgl, a GH1 β-glucosidase overexpressed in *Trichoderma harzianum* under biomass degradation conditions. Biotechnol Biofuels 9(1):1–11

Santos CA, Ferreira-Filho JA, O'Donovan A, Gupta VK, Tuohy MG, Souza AP (2017) Production of a recombinant swollenin from *Trichoderma harzianum* in *Escherichia coli* and its potential synergistic role in biomass degradation. Microb Cell Fact 16(1):1–11

Santos CA, Morais MA, Terrett OM, Lyczakowski JJ, Zanphorlin LM, Ferreira-Filho JA, Tonoli CC, Murakami MT, Dupree P, Souza AP (2019) An engineered GH1 β-glucosidase displays enhanced glucose tolerance and increased sugar release from lignocellulosic materials. Sci Rep 9(1):1–10

Saravanan A, Kumar PS, Vo D-VN, Jeevanantham S, Karishma S, Yaashikaa PR (2021) A review on catalytic-enzyme degradation of toxic environmental pollutants: microbial enzymes. J Hazard Mater 419:126451. https://doi.org/10.1016/j.jhazmat.2021.126451

Sazal M, Mathee K, Ruiz-Perez D, Cickovski T, Narasimhan G (2020) Inferring directional relationships in microbial communities using signed Bayesian networks. BMC Genomics 21(6):663. https://doi.org/10.1186/s12864-020-07065-0

Seiboth B, Karimi RA, Phatale PA, Linke R, Hartl L, Sauer DG, Smith KM, Baker SE, Freitag M, Kubicek CP (2012) The putative protein methyltransferase LAE1 controls cellulase gene expression in *Trichoderma reesei*. Mol Microbiol 84(6):1150–1164. https://doi.org/10.1111/j.1365-2958.2012.08083.x

Shoseyov O, Shani Z, Levy I (2006) Carbohydrate binding modules: biochemical properties and novel applications. Microbiol Mol Biol Rev 70(2):283–295

Sieber CM, Lee W, Wong P, Münsterkötter M, Mewes H-W, Schmeitzl C, Varga E, Berthiller F, Adam G, Güldener U (2014) The *Fusarium graminearum* genome reveals more secondary metabolite gene clusters and hints of horizontal gene transfer. PLoS ONE 9(10):e110311

Silva F, Gonçalves D, Lopes D (2020) The use of bioinformatics tools to characterize a hypothetical protein from *Penicillium rubens*. Genet Mol Res 19(2):1–18

Silva-Rocha R, Castro LdS, Antoniêto ACC, Guazzaroni M-E, Persinoti GF, Silva RN (2014) Deciphering the Cis-regulatory elements for XYR1 and CRE1 regulators in *Trichoderma reesei*. PLoS ONE 9(6):e99366. https://doi.org/10.1371/journal.pone.0099366

Singh A, Bajar S, Devi A, Pant D (2021) An overview on the recent developments in fungal cellulase production and their industrial applications. Bioresourc Technol Rep 14:100652. https://doi.org/10.1016/j.biteb.2021.100652

Singh J, Gehlot P (2020) New and future developments in microbial biotechnology and bioengineering: recent advances in application of fungi and fungal metabolites. Curr Aspects (TechnicalReport)

Sivashankari S, Shanmughavel P (2006) Functional annotation of hypothetical proteins–a review. Bioinformation 1(8):335

Skellam E (2019) Strategies for engineering natural product biosynthesis in fungi. Trends Biotechnol 37(4):416–427. https://doi.org/10.1016/j.tibtech.2018.09.003

Solovyev V, Kosarev P, Seledsov I, Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol 7(1):1–12

Song L, Langfelder P, Horvath S (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinform 13(1):1–21

Stajich JE (2017) Fungal genomes and insights into the evolution of the kingdom. Fungal Kingd 2017:619–633

Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33(Suppl–2):W465–W467

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34(suppl_1):D535–D539

Steindorff AS, do Nascimento Silva R, Coelho ASG, Nagata T, Noronha EF, Ulhoa CJ (2012) *Trichoderma harzianum* expressed sequence tags for identification of genes with putative roles in mycoparasitism against *Fusarium solani*. Biol Control 61(2):134–140

Stricker AR, Grosstessner-Hain K, Würleitner E, Mach RL (2006) Xyr1 (xylanase regulator 1) regulates both the hydrolytic enzyme system and d-xylose metabolism in *Hypocrea jecorina*. Eukaryot Cell 5(12):2128. https://doi.org/10.1128/EC.00211-06

Strogatz SH (2001) Exploring complex networks. Nature 410(6825):268–276

Subramanian I, Verma S, Kumar S, Jere A, Anamika K (2020) Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights 14:1177932219899051

Teigiserova DA, Bourgine J, Thomsen M (2021) Closing the loop of cereal waste and residues with sustainable technologies: an overview of enzyme production via fungal solid-state fermentation. Sustain Prod Consumpt 27:845–857. https://doi.org/10.1016/j.spc.2021.02.010

Todeschini A-L, Georges A, Veitia RA (2014) Transcription factors: specific DNA binding and specific gene regulation. Trends Genet 30(6):211–219. https://doi.org/10.1016/j.tig.2014.04.002

Tomer A, Singh R, Singh SK, Dwivedi SA, Reddy CU, Keloth MRA, Rachel R (2021) Role of fungi in bioremediation and environmental sustainability. In: Prasad R, Nayak SC, Kharwar RN, Dubey NK (eds) Mycoremediation and environmental sustainability: Volume 3. Springer International Publishing, Cham, pp 187–200. https://doi.org/10.1007/978-3-030-54422-5_8

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3):562

Uechi K, Watanabe M, Fujii T, Kamachi S, Inoue H (2020) Identification and biochemical characterization of major β-mannanase in *Talaromyces cellulolyticus* mannanolytic system. Appl Biochem Biotechnol 2020:1–16

Ulrich LE, Zhulin IB (2007) MiST: a microbial signal transduction database. Nucleic Acids Res 35(suppl_1):D386–D390

Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ 32(12):1633–1651

Usman A, Mohammed S, Mamo J (2021) Production, optimization, and characterization of an acid protease from a filamentous fungus by solid-state fermentation. Int J Microbiol 2021:6685963–6685963. https://doi.org/10.1155/2021/6685963

Van Den Berg MA, Albang R, Albermann K, Badger JH, Daran J-M, Driessen AJ, Garcia-Estrada C, Fedorova ND, Harris DM, Heijne WH (2008) Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. Nat Biotechnol 26(10):1161–1168

van den Berg RA, Braaksma M, van der Veen D, van der Werf MJ, Punt PJ, van der Oost J, de Graaff LH (2010) Identification of modules in *Aspergillus niger* by gene co-expression network analysis. Fungal Genet Biol 47(6):539–550

Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Global protein function prediction from protein-protein interaction networks. Nat Biotechnol 21(6):697–700

Vella D, Zoppis I, Mauri G, Mauri P (2017) Di Silvestre D (2017) From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. EURASIP J Bioinf Syst Biol 1:1–16

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA (2001) The sequence of the human genome. Science 291(5507):1304–1351

Viniegra-González G, Favela-Torres E, Aguilar CN, Rómero-Gomez SdJ, Díaz-Godínez G, Augur C (2003) Advantages of fungal enzyme production in solid state over liquid fermentation systems. Biochem Eng J 13(2):157–167. https://doi.org/10.1016/S1369-703X(02)00128-6

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One 9 (11):e112963

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63

Wang B-T, Hu S, Yu X-Y, Jin L, Zhu Y-J, Jin F-J (2020) Studies of cellulose and starch utilization and the regulatory mechanisms of related enzymes in fungi. Polymers 12(3):530

Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, Gu Z, Bruno D, Miranda M, Nguyen M (2007) Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. Proc Natl Acad Sci 104(31):12825–12830

Wei X, Chen L, Tang J-W, Matsuda Y (2020) Discovery of pyranoviolin A and its biosynthetic gene cluster in *Aspergillus violaceofuscus*. Front Microbiol 11:2488

Wilson EO (1999) Consilience: The unity of knowledge, vol 31. Vintage, New York

Wolfe CJ, Kohane IS, Butte AJ (2005) Systematic survey reveals general applicability of" guilt-by-association" within gene coexpression networks. BMC Bioinform 6(1):1–10

Wu C, Yang F, Smith KM, Peterson M, Dekhang R, Zhang Y, Zucker J, Bredeweg EL, Mallappa C, Zhou X (2014) Genome-wide characterization of light-regulated genes in *Neurospora crassa*. G3: Genes, Genomes, Genetics 4(9):1731–1745

Xu T, Zheng X, Li B, Jin P, Qin Z, Wu H (2020) A comprehensive review of computational prediction of genome-wide features. Brief Bioinform 21(1):120–134

Xu Z, Hejzlar P (2008) MCODE, Version 2.2: an MCNP-ORIGEN depletion program. In: Massachusetts Institute of Technology. Center for Advanced Nuclear Energy

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35(suppl_2):W265–W268

Yang Z, Zhang Z (2018) Engineering strategies for enhanced production of protein and bio-products in Pichia pastoris: a review. Biotechnol Adv 36(1):182–195

Yang KK, Wu Z, Arnold FH (2019) Machine-learning-guided directed evolution for protein engineering. Nat Methods 16(8):687–694

Yu J, Chang P-K, Ehrlich KC, Cary JW, Bhatnagar D, Cleveland TE, Payne GA, Linz JE, Woloshuk CP, Bennett JW (2004) Clustered pathway genes in aflatoxin biosynthesis. Appl Environ Microbiol 70(3):1253–1262

Zeilinger S, Ebner A, Marosits T, Mach R, Kubicek C (2001) The Hypocrea jecorina HAP 2/3/5 protein complex binds to the inverted CCAAT-box (ATTGG) within the cbh2 (cellobiohydrolase II-gene) activating element. Mol Genet Genomics 266(1):56–63. https://doi.org/10.1007/s004380100518

Zhang X, Acencio ML, Lemke N (2016) Predicting essential genes and proteins based on machine learning and network topological features: a comprehensive review. Front Physiol 7:75

Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4:1

Zhao W, Langfelder P, Fuller T, Dong J, Li A, Hovarth S (2010) Weighted gene coexpression network analysis: state of the art. J Biopharm Stat 20(2):281–300

Zhao Z, Liu H, Wang C, Xu J-R (2013) Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics 14(1):1–15

Zhao X-Q, Zhang X-Y, Zhang F, Zhang R, Jiang B-J, Bai F-W (2018) Metabolic engineering of fungal strains for efficient production of cellulolytic enzymes. In: Fungal cellulolytic enzymes. Springer, pp 27–41

Zhou X, Qi X, Huang H, Zhu H (2019) Sequence and structural analysis of AA9 and AA10 LPMOs: an insight into the basis of substrate specificity and regioselectivity. Int J Mol Sci 20(18):4594