OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

The transcriptomic networks controlling the sporulation in *Physarum polycephalum*

Dissertation
zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

genehmigt durch die Fakultät für Naturwissenschaften
der Otto-von-Guericke-Universität Magdeburg

von Walter Israel Barrantes Bustinza (Licenciado in Biology)
geb. am 14. August 1974 in Lima, Peru.

Gutachter: Prof. Dr. Wolfgang Marwan
Prof. Dr. Ludwig Eichinger

eingereicht am: 26. Juni 2014
verteidigt am: 17. März 2015

Vorgelegt von Dipl. Biol. Barrantes Bustinza, Walter Israel

"The transcriptomic networks controlling the sporulation in *Physarum polycephalum*"

Abstract

Physarum polycephalum is a unicellular eukaryote that belongs to the Amoebozoa group of organisms. Its complex life cycle involves various cell types that differ in morphology and biochemical composition. Sporulation, one step in the life cycle, is a simple form of differentiation that can be experimentally induced by far-red light. Well-established Genetics and the occurrence of macroscopic cells with naturally synchronous dividing nuclei make *Physarum* a model organism for studying the process of cell differentiation. In this thesis, next generation sequencing technologies were employed, specifically RNA sequencing (RNA-seq), together with multiple computational approaches, to study the transcriptomic changes during the commitment to sporulation in plasmodial cells. This work involved: (*i*) The generation of a transcriptome from cell pools; (*ii*) the identification of the transcriptome in single plasmodial cells; and (*iii*) combining the transcriptomes with the novel genome sequence data release to characterize the reference transcriptome. First, differentially expressed genes were identified in cell populations, and their products integrated into interaction networks using information from orthologs and the literature. Differential expression analyses showed that after light induction of a plasmodium the expression of transcripts linked to cell division and DNA repair is downregulated. In contrast, light-induction stimulated the expression of genes associated with the protein turnover, the cell cycle progression, and the maintenance of cell integrity and cytokinesis. Additionally, different groups of calcium-binding proteins are either down- or upregulated after light exposure. These differentially expressed genes are associated to a network of actin-binding proteins, whose products might accomplish different tasks in each stage. Later, high- coverage RNA-seq was performed with samples of individual plasmodial cells from *Physarum*, to characterize the the differentiation-dependent gene expression at the single-cell level. In this case, the observed regulation patterns correlate well with the results on cell populations, particularly regarding genes linked to signaling and actin-binding activities. Finally, a reference transcriptome for *Physarum* was generated from its first public draft genome. Novel RNA-seq analyses together with other available cDNA databanks supported the identification of 25,649 encoded transcripts. Genetic networks linked to cell differentiation were annotated, and molecular complexes involved in signal transduction and development were found within these large interactions. In addition, other major RNA families were mapped. This work contributes to necessary basic knowledge to understand the mechanisms of cell differentiation in this organism, through the characterization of networks and complexes specific to these molecular functions. Aside from the genome and transcriptomic sequences and their analyses, this thesis also offers a working pipeline and protocols that can be taken as a blueprint for the analysis of future transcriptomic sequences.

Vorgelegt von Dipl. Biol. Barrantes Bustinza, Walter Israel

"The transcriptomic networks controlling the sporulation in *Physarum polycephalum*"

Zusammenfassung

Physarum polycephalum ist ein einzelliger Eukaryot, welcher der Gruppe der Amoebozoen angehört. Sein komplexer Lebenszyklus umfasst verschiedene Zelltypen, die sich in der Morphologie und biochemischen Zusammensetzung unterscheiden. Die Sporenbildung, ein Abschnitt des Lebenszyklus, ist eine einfache Form der Differenzierung, die experimentell mit dunkelrotem Licht induziert werden kann. Die gut etablierte klassische Genetik und das Vorkommen von makroskopischen Zellen mit sich natürlicherweise synchron verhaltenden Zellkernen machen *Physarum* zu einem Modellorganismus für das Studium der Zelldifferenzierung. In dieser Doktorarbeit wurden bioinformatische Methoden zur Analyse von Daten aus Sequenzierungen der nächsten Generation angewandt, insbesondere der RNA Sequenzierung (RNA-Seq), um die Transkriptom-Änderungen während der Determination zur Sporulation plasmodialer Zellen zu untersuchen. Diese Doktorarbeit beinhaltet: (*i*) Die Analyse des Transkriptoms aus Zellpools; (*ii*) die Identifizierung des Transkriptoms einzelner Plasmodien-Zellen; und (*iii*) die Kombination von Transkriptomdaten mit der noch unveröffentlichten Genomsequenz, um ein Referenz-Transkriptom zu erstellen. Zunächst wurden differentiell exprimierte Gene in Zellpopulationen identifiziert und ihre Produkte in Interaktionsnetzwerken angeordnet, die mithilfe publizierter Informationen über Orthologe erstellt wurden. Differentielle Expressionsanalysen zeigten, dass nach Lichtinduktion eines Plasmodiums, das Expressionsniveau von Transkripten, welche im Zusammenhang mit der Zellteilung und der DNA-Reparatur stehen, herunterreguliert ist. Im Gegensatz dazu stimulierte die Licht Induktion die Expression von Genen, die für den Protein-Turnover, die Zellzyklus-Progression, die Aufrechterhaltung der Zellintegrität und die Zellteilung verantwortlich sind. Desweiteren werden verschiedene Gruppen von Calcium-bindenden Proteinen nach der Belichtung entweder nach unten oder nach oben reguliert. Diese differentiell exprimierten Gene sind Teil eines Netzwerkes von Aktin-bindenden Proteinen, dessen Produkte verschiedene Funktionen bei den genannten Prozessen vermitteln können. In einem weiter gehenden Ansatz wurden RNA-Seq Daten von *Physarum* Einzelzellen analysiert, um das Transkriptom in Abhängigkeit vom Differenzierungszustand auch auf Einzelzellebene zu charakterisieren. Die beobachteten Regulationsmuster korrelieren gut mit ersten Ergebnissen dieser Doktorarbeit hinsichtlich der Zellpopulationen, besonders im Zusammenhang mit Proteinen, die an der Aktin-Bindung und Signalverarbeitung beteiligt sind. Schließlich wurde ein Referenz Transkriptom für *Physarum* von der noch unveröffentlichen Genomsequenz erzeugt. Neue RNA-Seq-Analysen zusammen mit anderen verfügbaren cDNA-Datenbanken erlaubten die Identifikation von 25.649 kodierenden Transkripten. Genetische Netzwerke, die an Zelldifferenzierung gekoppelt sind wurden annotiert und Molekülkomplexe, die an Signaltransduktion und Entwicklung beteiligt sind, wurden anhand ihrer putativen Wechselwirkungen identifiziert. Darüber hinaus wurden Mitglieder andere wichtiger RNA Familien identifiziert.

Die vorliegende Doktorarbeit liefert essentielle Grundlagen, um die Mechanismen der Zelldifferenzierung in diesem Organismus zu verstehen, durch die Charakterisierung von Netzwerken und Komplexen, welche spezifisch für die entsprechenden molekularen Funktionen sind. Abgesehen von den Genom- und Transkriptom-Sequenzen und ihrer Analyse, wurde im Verlauf der Doktorarbeit auch eine bioinformatische Pipeline nebst Protokollen etabliert, die für zukünftige Analysen von Transkriptom-Daten verwendet werden kann.

Thesis Summary

Physarum polycephalum (“slime mold”), is a unicellular eukaryote that belongs to the Amoebozoa group of organisms. Its complex life cycle involves various cell types that differ in morphology, function, and biochemical composition. Sporulation, one step in the life cycle, is a simple form of cell differentiation that can be artificially induced by red light. Well-established genetics and the occurrence of macroscopic cells with a naturally synchronous population of nuclei as source of homogeneous cell material make *Physarum* a model organism for studying the process of cell differentiation. *Physarum* gene expression has been shown to be cell type-specific, but existing studies have been focused only on individual genes. In addition, cDNA libraries from macroplasmodia and other cell types have been reported (Martel et al. 1988; Watkins and Gray 2008; Glöckner et al. 2008).

In this work, the next generation sequencing technologies were employed, especifically RNA-sequencing (RNA-seq), together with multiple computational approaches, to study the transcriptomic changes during the commitment to sporulation in plasmodial cells. These analyses were carried out at three different levels: (*i*) The generation of a expressed transcriptome from cell pools; (*ii*) The identification of the expressed transcriptome in single plasmodial cells; and (*iii*) Combining the expressed transcriptomes with the novel genome release to characterize the reference transcriptome.

First, the global changes in expression that occur during light-induced sporulation of *Physarum* were analyzed, via low coverage RNA-seq (454 sequencing). In this manner, differentially expressed genes were identified, and their products integrated into interaction networks using information from orthologs and the literature. It was found that after light induction of a plasmodium the expression of transcripts linked to cell division and DNA repair is downregulated. In contrast, light-induction stimulated the expression of genes associated with protein turnover, genes related to cell cycle progression, and genes involved in the maintenance of cell integrity and cytokinesis. Additionally, different groups of calcium-binding proteins are either down- or upregulated after light exposure. These changes were associated with a network of actin-binding proteins, whose products might accomplish different tasks in each stage: the reorganization of the subcellular compartments in order to inhibit migration during starvation on one hand, and cell polarization and cytoskeletal redistribution after photoinduction mediated by a group of actin-binding proteins on the other.

Later, the availability of the high- coverage RNA-seq through the Illumina platform was combined with the simplicity for obtaining single cells from *Physarum*, to characterize the expressed transcriptome through the differentiation of this lower eukaryote, at the single-cell level. The observed regulation patterns correlate well with previous results on the differential gene expression during the commitment to sporulation in the slime mold, particularly with respect to proteins involved in signaling and actin-binding.

Finally, a reference transcriptome for the slime mold was generated and annotated, over its first public draft genome. Novel RNA-seq analyses together with other available cDNA databanks, supported the identification of 25,649 encoded transcripts. Genetic networks linked to cell differentiation were annotated, and molecular complexes involved in signal transduction and development were found within these large interactions. In addition, other major RNA families were mapped.

This work contributes the necessary basic knowledge to understand the mechanisms of cell differentiation in this organism, especially through the characterization of networks and complexes specific to these molecular functions. Furthermore, it provides a starting point for further exploration of the biology of *Physarum*, and its utility as a model organism. It is expected that the precise representation of the differentiation networks may become available as gene knockout experiments, proteomic data, and other high- throughput approaches are integrated in future studies of this organism. Aside from the genome and transcriptomic sequences and their analyses, this work also offers a working pipeline and annotation protocols, which can be taken as a blueprint for the analysis of future genomic and transcriptomic studies.

Contributions

Cell cultures and other biological samples were contributed by Prof. Dr. Wolfgang Marwan. RNA-seq sequencing procedures were carried out with the support of Vertis Biotechnologie AG (Fresing, Germany). The initial analysis of the 454 sequencing data was done with the participation of Sonja Meyer (Max Planck Institute for Dynamics of Complex Systems) and Gernot Glöckner (Leibniz Institute for Freshwater Ecology and Inland Fisheries, Berlin). Dr. Glöckner also contributed with valuable discussions during the 454 RNA-seq study and the design of the genomic annotation pipeline. The differential expression analysis of the Illumina data was carried out in collaboration with Jeremy Leipzig (Center for Biomedical Informatics, Children Hospital of Philadelphia, PA). Data to calculate the average gene size was provided by Daniel Ence (Eccles Institute of Human Genetics, University of Utah). The *Physarum* genome assembly analyzed in Chapter 5 was provided by Patrick Minx (The Genome Institute, Washington University School of Medicine, St.Louis, MO). The CEGMA analysis of the genome sequences was performed by Keith Bradnam (Genome Center, University of California, Davis). This thesis work was developed both in the Max Planck Institute for Dynamics of Complex Systems (2007 – 2010), and later in the Institut für Biologie of the Otto-von-Guericke University (2011-2013), Magdeburg.

Published Materials

Parts of this thesis work have been published in the following journal articles:

Barrantes I., Glockner G., Meyer S., Marwan W.
Transcriptomic changes arising during light-induced sporulation in *Physarum polycephalum*.
BMC Genomics. 2010 Feb 17; 11: 115.

Barrantes I., Leipzig J., Marwan W.
A next-generation sequencing approach to study the transcriptomic changes during the differentiation of *Physarum* at the single-cell level.
Gene Regul Syst Bio. 2012; 6: 127 – 137.

Supervision

Thesis Director:	Prof. Dr. Wolfgang Marwan Magdeburg Centre for Systems Biology, and Institute of Biology Otto-von-Guericke University

Acknowledgements

First and foremost to all my labmates ("the molnets"), for their support. Especially to Prof. Wolfgang Marwan and Dr. Markus Haas, and also to Jan-Thierry Wegener, for his help with computational issues.

To the people at the IMPRS and the Graduate School of the Otto von Guericke University, especially to Dr. Barbara Witter.

To those who helped during the application process: Esteban Ballestar (Centro Nacional de Investigaciones Oncológicas, Madrid), Alvaro Marcelo (Universidad Nacional Mayor de San Marcos, Lima), and Celso Romero (Universidade de Sao Paulo).

To those who made easier my transition to Germany: Mercedes Ames and Jeremy Leipzig;

And to those who made easier the stay: Arturo Padilla, Juan Pablo Gonzalez Gómez, Nini Phan, Jérémy Huard, Alexandra Grusche, and Jenny Wondrak.

To the *Physarum* Genome Consortium, especially to Patrick Minx and Richard K. Wilson and The Genome Institute, Washington University School of Medicine (St.Louis, MO), for providing the early sequence assembly that is featured on Chapter 5. This sequencing and characterization of the *Physarum polycephalum* genome was in turn funded by The National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH; Bethesda, MD).

This thesis work was supported by an International Max Planck Research School - Otto von Guericke University fellowship, and by grants of the German Federal Ministry of Education and Research (FORSYS - MaCS, Magdeburg Centre for Systems Biology).

Contents

1. Introduction 1
Physarum polycephalum........ 1
The life cycle of *Physarum* 4
Genomic Organization and Strains of *Physarum*........ 5
Gene Regulation during the Life Cycle of *Physarum*........ 6
    Microtubules........ 7
    Actin Cytoskeleton........ 8
    Actin-Binding Proteins........ 8
    Calcium-Binding Proteins........ 9
    Signal Transduction........ 10
    Epigenetic Modifications........ 11
    Other Stage- Specific Proteins and Metabolites........ 12
Physarum as a model organism........ 14
    Cell Motility........ 14
    Cell Differentiation........ 15
    DNA replication........ 16
    RNA editing........ 17
    Epigenetics........ 17
    Gravitational Biology........ 18
    Behavioral and Computer Sciences........ 19
Transcriptomes........ 19
Experimental methods for studying the transcriptome........ 20
    Microarrays........ 22
    EST Libraries........ 22
    Tag-Based Approaches........ 23
Next- Generation Sequencing........ 24
    Template preparation........ 25
    Sequencing........ 25
    Applications........ 28
    RNA sequencing (RNA-seq)........ 30
Computational methods for studying the transcriptome........ 32
    Assembly........ 33
    Expression........ 36
    Annotation........ 37
The Transcriptome of *Physarum*........ 39
    Comparative Transcriptomics........ 40
Objectives and Approach........ 42
Thesis Organization........ 42
2. Methods 44
I. Materials........ 44
II. Methods........ 52
Analysis of the expressed transcriptome during the differentiation of *Physarum* cell pools........ 52
    Culture and light-induction of plasmodial cells........ 52
    cDNA Library Construction and Sequencing........ 52
    Sequence Annotation and Network Inference........ 53

Analysis of the expressed transcriptome during the differentiation of *Physarum* single cells 55
Culture and sequencing 55
Data Analysis 55
Identification and Annotation of the Reference Transcriptome of *Physarum polycephalum* 58
RNA Sequencing of the White Strain (First Batch) 58
RNA Sequencing of the White Strain (Second Batch) 59
Genome sequencing and annotation 60
3. The expressed transcriptome during the differentiation of *Physarum* cell pools 63
Background 63
Results 63
Sequencing and Profiling of cDNAs expressed in competent and light-induced plasmodia 63
Gene Ontology Annotation of the Transcriptome 68
Pathway classification of transcripts 70
Inference of Interaction Networks 71
Discussion 79
A network of actin-binding proteins is associated to changes during light-induced sporulation in Physarum 79
Transcripts linked to cell division and DNA repair are downregulated in the light-induced plasmodium 80
Calcium- binding proteins exhibit diverse regulation patterns in the light-induced plasmodium 82
Actin-binding proteins associated to development are overexpressed in the light-induced plasmodium 83
GTP signaling genes involved in different processes are upregulated in the light-induced plasmodium 85
Transcripts annotated for cell death are overrepresented in the light-induced plasmodium 85
Conclusions 87
Summary 88
4. The expressed transcriptome during the differentiation of *Physarum* single cells 89
Background 89
Results and Discussion 90
Conclusions 103
Summary 104
5. The reference transcriptome of *Physarum polycephalum* 106
Background 106
Results 107
The Physarum Genome 107
Repeat Annotation 109
Non-coding RNA (ncRNA) Annotation 112
Mapping RNA-seq short reads to the *Physarum* genome 120
EST mapping against the *Physarum* genome 123
Inference of the Protein-coding gene models 123
Annotation of the Gene Models 129

Network Analysis ... 135
Validation and Completeness of the Genome and the Gene Models ... 152
Quality of the Assembly ... 153
Sequencing Contaminants ... 154
Mapping ESTs as an estimate of completeness ... 155
Mapping *Physarum* GenBank sequences to the novel assembly ... 156
Mapping RNA-seq short reads to assess the gene coverage ... 157
Mapping long RNA-seq reads against the genome ... 158
Comparison of Evidence- based Models versus Predictions ... 160
Mapping CEGMA datasets as an estimate of completeness of the genome assembly and the protein models ... 164
Protein coding potential of the transcript models ... 165
Comparative Analyses ... 168
Genome Assemblies ... 168
Repetitive Sequences ... 170
Encoded Genes and Proteins ... 171
Discussion ... 189
Genome Annotation ... 189
Validation and Completeness ... 192
Comparative Genomics ... 194
Conclusions ... 196
Summary ... 197
6. Bibliography ... 199
7. Appendix ... 225
Appendix 1. Calculation of RNA-seq Reads per mRNA molecule. ... 225
Appendix 2. Estimating the number of protein coding genes in *Physarum* ... 227

List of Figures

1 The life cycle of *Physarum polycephalum* 2
2 Heterothallic and Apogamic Cycles 3
3 Next generation sequencing flowchart 24
4 The Illumina method 26
5 The 454 method 27
6 RNA-seq flowchart 31
7 Experimental design for the analysis in cell pools 54
8 Experimental design for the analysis in single cells 57
9 Genome annotation pipeline 61
10 Hits Distribution of Transcript Species 65
11 Relative frequencies of transcripts in libraries prepared from competent and photoinduced plasmodia 67
12 Metabolic Atlas of *Physarum polycephalum* 71
13 Interactions with the Actin Cytoskeleton of Transcripts with Higher Relative Frequencies 72
14 Interaction of the Most Upregulated and Downregulated Transcripts with the Actin Cytoskeleton 78
15 Quality assessment of the Illumina RNA-seq outputs 95
16 Assessment of the reproducibility of the approach 96
17 Fold change and significance 97
18 Gene Ontology classification of differentially expressed transcripts 99
19 Distribution of scaffolds and contigs in the *Physarum* genome 108
20 Distribution of repeat families at the *Physarum*- specific custom repeat library 112
21 Relationship between the gene and genome sizes from a representative set of species 127
22 Example of evidences forming a gene model 130
23 Top 10 most frequent gene annotation descriptions 132
24 Top 10 most represented species in the orthologs 133
25 Top 10 most frequent gene ontology associations 133
26 The reference metabolic map of *Physarum* 134
27 Summary of genes associated to cell differentiation 137
28 Complexes linked to the embryo development ontology 139
29 Complexes linked to the signal transduction ontology 140
30 Complexes linked to the cell differentiation ontology 141
31 Comparison of gene ontologies between complexes and subnetworks annotated for cell differentiation 150
32 Comparison of gene ontologies between complexes and subnetworks annotated for signal transduction 151
33 Comparison of gene ontologies between complexes and subnetworks annotated for embryo development 152
34 Coverage identity plot of contaminants 155
35 Gene-body coverage of short RNA-seq reads from the strain WT31 159
36 Phylogenetic tree of slime molds 168
37 tRNA genes and codon counts in Mycetozoans 177
38 Comparison of ontologies between *Physarum* and dictyostelids. 181
39 The reference metabolic maps of *Physarum* and dictyostelids 182

40 Phylogenetic pattern of *Dictyostelium purpureum* and *Physarum* 187
41 SimiTri profile of the protein models from *Physarum* 188

List of Tables

1 Experimental methods for studying whole transcriptomes 21
2 Commercially available next generation sequencing platforms 29
3 Strains used in this work 44
4 Hardware 45
5 Nucleotide Databases and Datasets 46
6 Protein Databases and Datasets 47
7 RNA-seq datasets 48
8 Programs used for assembly and mapping RNA-seq data 48
9 Programs used for identification of repetitive sequences and non-coding RNA annotation 49
10 Programs used in the annotation and the comparative analysis of the coding transcriptome. 50
11 Programs used for the identification of gene models 51
12 Annotated transcripts with relative frequencies higher than 0.005 67
13 Overrepresented Gene Ontology terms in Upregulated Transcripts 69
14 Summary of the transcriptome sequencing and annotation 69
15 Top 20 Transcripts Downregulated in Light-induced Plasmodia 73
16 Top 20 Transcripts Upregulated in Light-induced Plasmodia 74
17 Top 20 Annotated Transcripts Downregulated in Light-induced Plasmodia 76
18 Top 20 Annotated Transcripts Upregulated in Light-induced Plasmodia 77
19 Summary of the single-cell transcriptome sequencing 92
20 Top 20 Annotated Transcripts Upregulated after Photoinduction 100
21 Top 20 Annotated Transcripts Downregulated after Photoinduction. 101
22 Sequencing summary for the genome assembly 108
23 Distribution of repetitive elements on the *Physarum* genome 111
24 Small nucleolar RNAs in the *Physarum* genome 114
25 MicroRNAs in the *Physarum* genome 115
26 Group II intron splicing non-coding RNAs in the *Physarum* genome 115
27 Bacterial non-coding RNAs found in the *Physarum* genome 116
28 Other non-coding RNAs found in the *Physarum* genome 116
29 tRNA Genes in the *Physarum* Genome 117
30 tRNA Genes including introns in the *Physarum* Genome 118
31 rRNA genes identified using RNAMMER 119
32 Summary of noncoding RNA predictions 119
33 RNA samples and sequencings from the white strain batches 120
34 Summary of Illumina RNA sequencing mappings 122
35 Transcripts identified by mapping of RNA-seq short reads 122
36 Gene and genome sizes from a representative set of species 126
37 Identification of protein gene models 128
38 Features of the predicted reference gene models 129
39 Top 10 most frequent KEGG orthologs 132
40 Summary of the conceptual protein complexes linked to the Cell Differentiation ontology 136
41 A protein complex annotated for Embryo Development 142
42 Protein orthologs annotated with the Embryo Development ontology that do not form protein complexes. 142

43 Proteins from the Signal Transduction Complex ST1 ... 143
44 Kinases from the Signal Transduction Complex ST1, and members of the protein complexes ST2, ST3 and ST4 ... 144
45 Signal transduction proteins that do not form complexes ... 145
46 Proteins from the Cell Differentiation complex CDN1 ... 146
47 Kinases from the Cell Differentiation complex CDN1 ... 147
48 Proteins from the Cell Differentiation complexes CDN2 and CDN3 ... 147
49 Proteins from the Cell Differentiation subnetwork CDO1 ... 148
50 Enzymes and proteins associated to the subnetwork CDO1 ... 149
51 Scaffolds and gaps from the most recent genome releases ... 154
52 Mapping clustered ESTs to the *Physarum* genome ... 157
53 Mapping ESTs to the novel protein models ... 157
54 Adaptors used for sequencing of the amoeba RNA samples ... 159
55 Runs from the 454 sequencing of amoeba transcript library ... 160
56 Evidence- and prediction- based exons and introns ... 162
57 Correspondences between the gene models and gene predictions ... 163
58 Comparison of genes and transcripts identified though RNA-seq, *ab initio* gene prediction and consensus gene modeling ... 163
59 CEGMA datasets ... 165
60 Mapping of CEGMA datasets to different genome versions ... 166
61 Statistics of the completeness of the genome using CEGMA ... 167
62 Predicted coding potential of transcripts and ESTs ... 167
63 Sequencing summaries of *Physarum* and other Mycetozoa ... 169
64 Distribution of repetitive elements on two dictyostelids ... 172
65 Protein-coding genes in protists and lower eukaryotes ... 173
66 Codon usage pattern in *Physarum* ... 175
67 Codon usage pattern in *Dictyostelium discoideum* ... 176
68 Biological process ontologies in *Physarum* ... 178
69 Biological process ontologies in *Dictyostelium discoideum* ... 179
70 Biological process ontologies in *Dictyostelium purpureum* ... 180
71 Mapping CEGMA datasets to dictyostelid genomes ... 185
72 Mapping CEGMA datasets to *Physarum* and dictyostelid proteomes ... 185
73 OrthoMCL analysis of the *Physarum* and dictyostelid proteomes ... 185
74 SimiTri yeast matches ... 186

Abbreviations

AED	Annotation edit distance
AMP	Adenosine monophosphate
BAC	Bacterial artificial chromosome
BWA	Burrows - Wheeler aligner
CCD	Charge-coupled device camera
cDNA	Complementary DNA
CEGMA	Core eukaryotic genes mapping approach
CEGs	Core eukaryotic genes
CNV	Copy number variation
CRAC	Cytosolic regulator of adenylyl cyclases
CRT	Cyclic reversible termination
dNTP	Deoxynucleotide triphosphate
emPCR	Emulsion PCR
ENA	European Nucleotide Archive
EST	Expressed sequence tag
FDR	False discovery rate
GFF	Generic Feature Format (also GFF3)
GMP	Guanosine monophosphate
GO	Gene Ontology
GOLD	Genomes Online Database
GLM	Generalized Linear Model
GTP	Guanosine triphosphate
KAAS	KEGG Automatic Annotation Server
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG ortholog
LGT	Lateral gene transfer
LINE	Long interspersed nuclear elements
LTR	Long terminal repeat
m5C	5-methyl-cytosine
m6A	N6-methyl- adenine
miRNA	Micro RNA
MLE	Maximum Likelihood Estimation
MMLV	Moloney murine leukemia virus
MTOC	Microtubule organising centre
ncRNA	Non-coding RNA
NGS	Next generation sequencing
ORF	Open reading frame
PCR	Polymerase chain reaction
PKA	cAMP- dependent protein kinase A
PMLA	Beta poly L-malate
PNK	Polynucleotide kinase
PPI	Protein – protein interaction
PPi	Inorganic phosphate
PTP	Picotiter plate
rDNA	Ribosomal DNA
RNA-seq	Deep RNA sequencing
RPKM	Reads per kilobase of transcript per million of mapped reads

rRNA	Ribosomal RNA
RSCU	Relative synonymous codon usage
RT-PCR	Reverse transcriptase PCR
SAGE	Serial analysis of gene expression
SBL	Sequencing by ligation
SINE	Short interspersed nuclear elements
snoRNA	Small nucleolar RNA
SNV	Single nucleotide variant
SRA	NCBI Sequence Read Archive
STAT	Signal transducer and activator of transcription
TAP	Tobacco acid pyrophosphatase
TIRF	Total internal reflection fluorescence
tRNA	Transfer RNA
UTR	Untranslated region

Chapter 1. Introduction

***Physarum polycephalum*.**

The slime mold *Physarum polycephalum* is a protist belonging to the clade of mycetozoans, a group whose members live either as individual amoebae (Class Dictyostelia or cellular slime molds; *e.g.*, *Dictyostelium*), or are able to fuse into large syncitia called plasmodia (Class Myxomycetes or plasmodial slime molds; *e.g., Physarum*). Other groups such as the acrasid slime molds have also been classified as mycetozoans, although there is no consensus about this inclusion (Blanton 2001; Adl et al. 2012).

Physarum was first grouped together under the lower fungi, but in recent years it has been accepted the following classification under the Protozoa (Baldauf and Doolittle 1997; Blanton 2001; Adl et al. 2012; The Marine Biological Laboratory 2013):

Division	Protozoa
Subdivision	Mycetozoa
Class	Myxomycetes
Order	Physarales
Family	Physaraceae
Genus	*Physarum*
Species	*Physarum polycephalum* Schweinitz 1822

Slime molds are cosmopolitan, with most species described in temperate forests. They are free- living heterotrophs, *i.e.,* they cannot fix carbon and therefore they rely on other organisms as sources of organic molecules, typically engulfing bacteria and other decaying matter found in soil of their natural habitats -and as such, they are secondary decomposers (Burland et al. 1993; Blanton 2001).

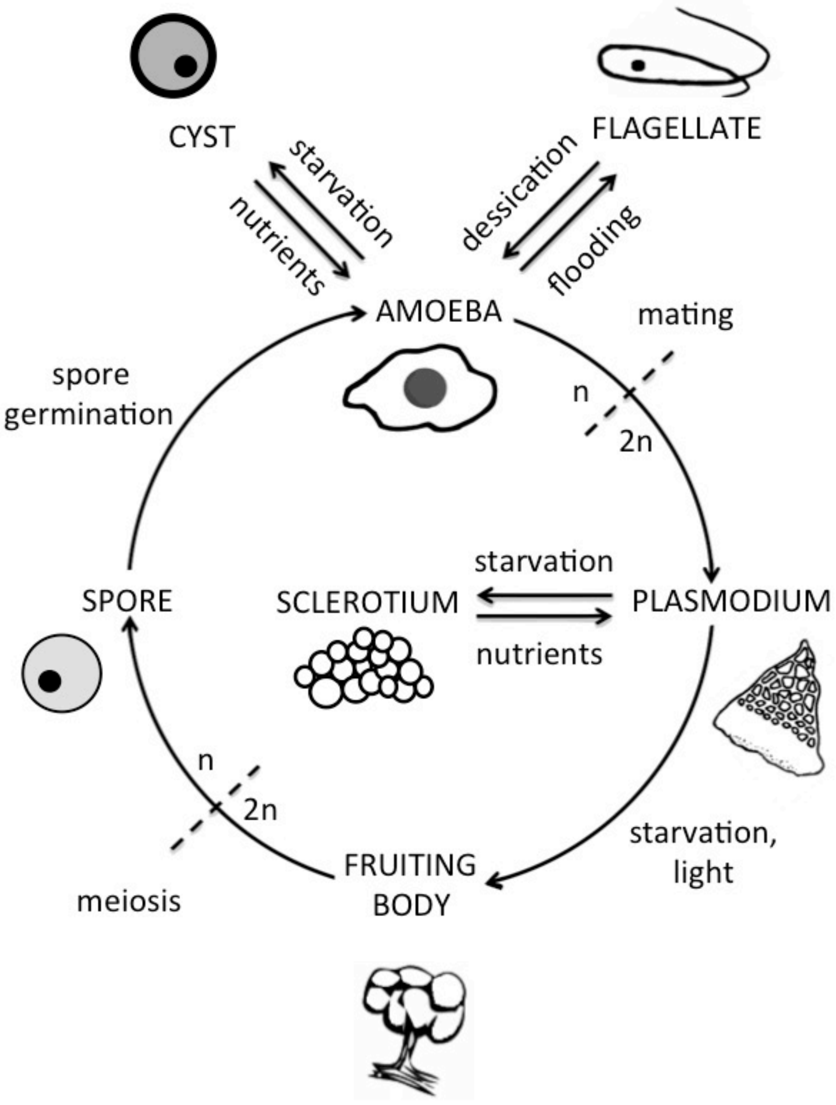


Figure 1. The life cycle of *Physarum polycephalum*. Spores, released from mature fruiting bodies, germinate into mononucleate amoebae (*n*), which propagate by mitosis. At high population density, amoebae of different mating type are able to mate, to form a zygote (*2n*). This diploid cell later develops into a multinuclear plasmodium (*2n*), through multiple nuclear divisions. Following starvation, the plasmodium can be induced to sporulation by visible light. Later, the plasmodial mass develops into individual fruiting bodies, which will subsequently yield haploid spores (*n*). Adapted from several sources (Burland et al. 1993; Marwan 2003).

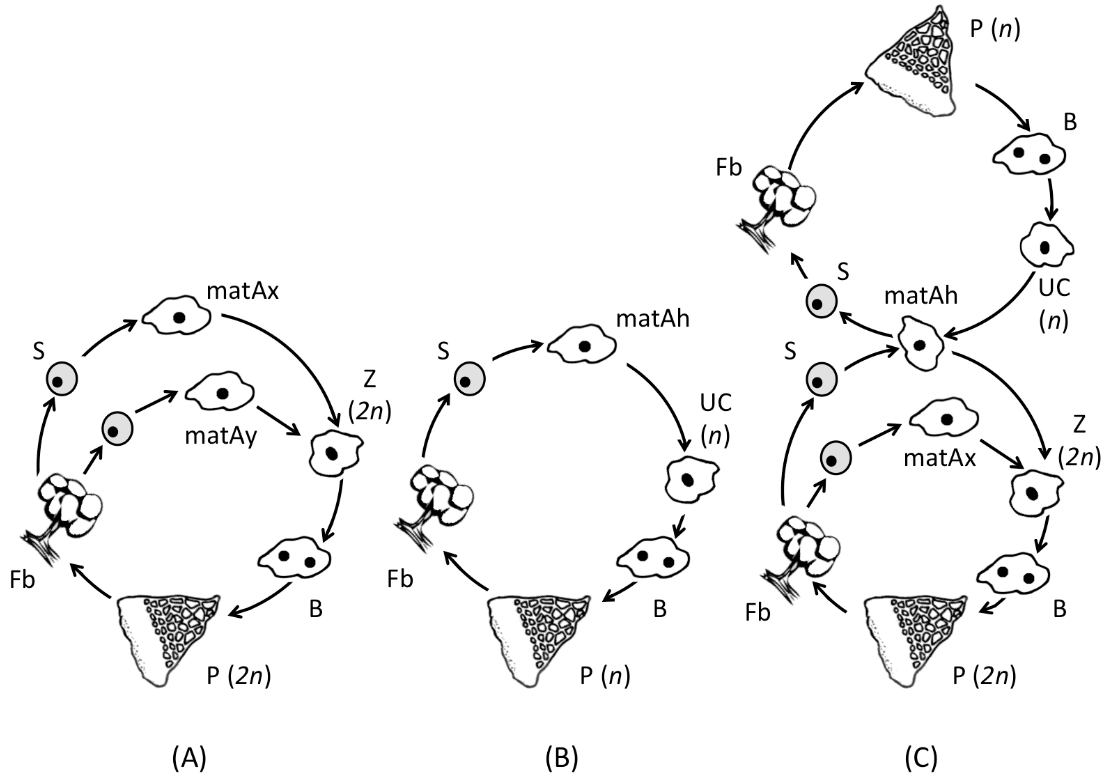


Figure 2. Heterothallic and Apogamic Cycles. During the heterothallic cycle (*A*), a the plasmodium (*P*) develops into a fruiting body (*Fb*), which will produce spores (*S*). From these, amoebae with wild-type alleles of the mating type locus (*matAx, matAy*) can re-enter the cycle by fusing into a diploid zygote (*Z*). Sucessive divisions of nuclei occur without cytokinesis, generating a binucleate cell (*B*), that develops into a multinucleate plasmodium. In the apogamic cycle (*B*), an uninucleate haploid cell committed to plasmodium formation (*UC*) will develop directly from amoebae carrying the mutant allele of matAh. In *C*, the cross of apogamic amoebae (genotype *matAh*) with heterothallic amoebae (genotype *matAx*) gives progenies of both types. Redrawn from several sources (Dee 1987; Anderson and Dee 1990).

The life cycle of *Physarum*.

The life cycle of the slime mold entails the alternation between uni- and multinucleate stages, from which only the amoeba and the plasmodium are able to proliferate (Figure 1; Burland et al. 1993). The cell cycles of these two stages, under similar growth conditions, are the same length. The uninucleate stage is the amoeba, an haploid cell of 10-20 μm that feeds by phagocytosis of fungal spores and bacteria. Amoebae divide by an open mitosis, which is followed by cytokinesis, and further divisions produce colonies of genetically identical amoebae. Upon transfer to water, amoebae transform into biflagellated cells which change their movements from amoeboid crawling to swimming, and then they will in turn swim to dryer regions (Burland et al. 1993; Bailey 1997). Flagellates are not able to feed or divide; therefore, when flagellates settle on a surface, their flagella are resorbed, and the cell reverts to its amoeboid state. Under stress conditions (such as starvation or low temperatures), the amoebae synthesize a resistant wall, and develop into cysts. These cysts hatch to release the contained amoebae, when favourable conditions return. Later, at high population density, the mating of two amoebae of compatible mating types produces a diploid zygote, which by multiple nuclear divisions develops into a diploid plasmodium (Bailey 1997; Marwan 2003). This change, unlike the two others to flagellate or cyst forms, is irreversible (Burland et al. 1993).

The plasmodium is the multinucleate stage. This cell feeds by phagocytosis of bacteria and other microbes, but they are also capable of pinocytosis, through secretion of extracellular enzymes, in order to break down the materials. They can be grown in liquid shaking cultures in the form of *microplasmodia*, which in turn will fuse into a *macroplasmodium* when transferred to a surface, regularly forming a large, yellow macroscopic syncitium of 1 x 10^7 - 1 x 10^{10} nuclei, or even more, depending on its size. However, this union will occur only between plasmodia sharing the same alleles from the three fusion type loci (*fusA*, *fusB* and *fusC*). Upon plasmodial fusion, nuclei and cytoplasm mix, but the nuclei do not merge, and therefore a macroplasmodium may be a heterokaryon if genetically different plasmodia (of the same fusion type) have fused (Burland et al. 1993; Bailey 1997). Plasmodial cells move with the help of a network of veins which generate a cytoplasmic streaming. Nuclei are also transported across the cell during this streaming, and the direction of the movement changes almost every minute or less. The plasmodial mass and nuclei double with the completion of a full cell cycle. During division, DNA synthesis and mitosis occur synchronously within all the nuclei, but the lack of cytokinesis ensures the continuation of the syncitial form. This synchronization means that all nuclei within the same macroplasmodial cell are in the same cell cycle and developmental stage (Guttes and Guttes 1964). Despite their unlimited ability to grow and divide, plasmodia are unable to transform into flagellate cells, and thus can follow other alternate differentiation pathways, depending on the environment and cell size. Under adverse conditions, plasmodia can enclose themselves into dormant resistant sclerotia. Furthermore, starved plasmodia have three developmental options: (*i*) they will go through sporulation if they are illuminated or exposed to heat shock while grown in a humid chamber; (*ii*) they will spherulate if they are submersed in water; or else (*iii*) they will re-enter the regular growth program if they find a nutrient source. During sporulation, the cell develops into fruiting bodies, in which haploid mononucleate spores are formed by meiosis. In turn, these spores will produce haploid amoebae, closing the life cycle (Bailey 1997; Kohama and Nakamura 2001; Marwan 2003).

Genomic Organization and Strains of *Physarum*

The size of the nuclear genome is yet unclear, although it is believed to consist of approximately 270 – 300 Mb (Mohberg and Rusch 1971; Mohberg 1977; Burland et al. 1993; Glöckner et al. 2008), with diploid stages entailing 40 chromosomes (Mohberg 1977; Burland et al. 1993). The GC content of the genome is approximately 40% (Gordon and Hardman 1988), and around 7% of the cytosine residues are methylated (Whittaker and Hardman 1980). Two thirds of the genome are single copy, and the repetitive regions comprise both inverted and direct repeats (Burland et al. 1993). The gene number is not known, although several preliminary approaches to characterize the transcriptome point to 20,000 protein- coding genes (see below; Watkins and Gray 2008; Glöckner et al. 2008). In addition, the 63-Kb circular mitochondrial genome has an A+T content of 74.1%, and possess 20 genes: eleven proteins related to the electron transport chain, one ribosomal protein, two rRNA genes, and five tRNA genes (Takano et al. 2001). RNAs transcribed from the mitochondrial genome suffer considerable editing, most notably the insertion of single Cs, with Us and dinucleotides, although the function of this editing is not defined (Bundschuh et al. 2011a).

Strains. Mutants of several classes have been isolated in different laboratories, some of them displaying natural polymorphisms or carry different allelles at several loci; nevertheless this heterogeneity does not interfere with genetic analyses. According to their ability to form plasmodia, two groups of amoebae strains can be identified (heterothallic and apogamic; Figure 2). As mentioned above, heterothallic strains are those that proliferate as amoebae, and produce plasmodia solely through crosses –and only amoebae with compatible genotypes can mate. Crosses are under the control of three multiallelic loci: The *matB* and *matC* loci affect the efficiency of crossing, and different alleles for the *matA* locus are required so a diploid amoeba can develop into a diploid plasmodium. Conversely, the apogamic strains are those that generate haploid plasmodia in clonal cultures; therefore in early studies they were widely used in gene expression studies, as amoebae and plasmodia were of the same genotype. These apogamic strains can be difficult to cultivate as amoebae, because of their readiness to form plasmodia, however these problems can be avoided by changing the culture conditions, or by altering the genotype of the strain (Figure 2; Anderson and Dee 1990).

Gene Regulation during the Life Cycle of *Physarum*

Amoebae and plasmodia display cell type-specific gene expression, with specificity of microtubular and actin cytoskeleton structures in both cell types. Early studies of the molecular biology of *Physarum* revealed that up to a quarter of the abundant proteins show different expression levels in both stages (Larue et al. 1982; Turnock et al. 1981). Cell- specific cDNA libraries revealed a 5 – 10% cell-type specific expression for both amoebae and plasmodia, and the change of expression patterns is initiated when unicellular forms become committed to the formation of a plasmodium (Sweeney et al. 1987; Bailey 1997). The gene families that show cell stage- specific patterns of expression are those encoding microtubular, cytoskeletal, actin- and calcium- binding proteins, as well as others associated to the GTP signaling and some with unknown functions; these genes are controlled either by genetic or epigenetic regulation. In the following paragraphs these differences will be summarized across the cell stages during the life cycle of *Physarum*.

Microtubules. The microtubules play a fundamental role during nuclear division, and in the maintenance of the cell shape and polarity. Four alpha-tubulin (named *altA*, *altB*, *altC* and *altD*), and three beta-tubulin genes (*betA*, *betB* and *betC*) have been described. All these genes are unlinked, and some of these exhibit cell-type specific gene expression (Schedl et al. 1984; Burland et al. 1993; Bailey 1997). The synthesis of tubulins T1 and T2 were found to be induced during fruiting body formation (Putzer et al. 1984). In the amoebae, the microtubules radiate from the so-called nucleus-associated microtubule organizing centre (MTOC). These microtubules pertain to three tubulin isotypes (alpha-1, alpha-3, and beta-1). The alpha-3 isotype results from the post-translational modification of alpha-1 tubulins, and these alpha-3 subtypes can be found in flagellate and amoebae but cannot be detected in plasmodia. In plasmodial cells, the microtubules radiate from the cytoplasmic foci, and have no specific orientation. Plasmodia express the alpha-1, alpha-2, beta-1 and beta-2 tubulin isotypes. The beta-2 tubulin can be found in plasmodial cells but not in amoebae, although developing uninucleates can form beta-2(+) flagella. The expression pattern of this plasmodium-specific isotype is similar to the actin-binding protein profilin P (*proP*; Bailey et al. 1999; see below). The beta-2 isotype is first detected in plasmodial mitosis (Bailey et al. 1999), and displayed after the commitment, and in turn, the alpha-3 tubulin levels decrease as the beta-2 isotype increases (Bailey et al. 1999; Bailey 1997). However, the lack of the alpha-3 isotype, and the accumulation of the beta-2 tubulin alone are not sufficient to provoke the reorganization of microtubules during development (Bailey 1997).

Actin Cytoskeleton. The actin cytoskeleton is a key component during processes such as locomotion and cell division. A family of actin genes has been described, designated *ardA* to *ardE*. The actin genes *ardA*, *ardB* and *ardC* are expressed at high levels during all stages of the cell cycle; the specific expression changes of *ardE* are unknown. All these genes generate identical proteins; and therefore changes in actin gene expression are not responsible of changes in actin organization. In amoebae, the actin layer is located underneath the cell membrane, with a higher concentration of actin in the pseudopodia, and in the cytokinetic furrow of cells during mitosis. In flagellates, an actin-rich support layer runs along the dorsal axis of the cell, from the anterior to posterior regions. In turn, plasmodia, like amoebae, contain an actin layer just beneath the cell membrane, although they are arranged in a much more complex microfilament network than in the amoebae. Microfilaments in plasmodia form a three-dimensional network in areas that lack veins. The contraction of actin networks gives the propulsive force for cytoplasmic streaming and plasmodial locomotion. The only actin specifically associated to plasmodia is the product of the *ardD* gene, which is expressed during spherulation (Bailey 1997). In contrast, the amount of actin mRNA decreased during sporulation. Actin transcripts were found to be abundant in amoebae, growing plasmodia, and light- induced plasmodium, but remained in low levels at 4 hours after the light pulse and throughout sporulation (Martel et al. 1988).

Actin-Binding Proteins. The actin-binding proteins are all cell-type specific, except for the myosin light chain, and the myosin-like *mlpA* protein, which are ubiquitously distributed. In spite of their relevance in several cell processes, little is known about their differences in function. The main cell-type specific gene families are the profilin, the myosin heavy chain, the 18 *KDa*-myosin light chain, the fragmin genes, and coronin (Binette et al. 1990; Bailey 1997; T'jampens et al. 1999; Bailey et al. 1999; Minami et al. 2009). These actin-binding proteins possess at least one gene member expressed in amoebae, and another present in plasmodia, except for the coronin, that has been observed only in diploid plasmodia so far (Bailey 1997; Minami et al. 2009). For example, there are amoebal- (*frgA*) and plasmodial- specific (*frgP* and *frg60*) fragmins, as well as amoebal and plasmodial profilins (*proA* and *proP*, respectively). Antibody studies also suggest the presence of amoebal- and plasmodial- specific myosin genes (Bailey et al. 1999). Both fragmins and profilins are developmentally regulated (T'jampens et al. 1999; Bailey et al. 1999; Binette et al. 1990). The fragmins are calcium- dependent regulators of the microfilament system, that enhance the phosphorylation of the actin-formin heterodimer, through an actin-fragmin specific kinase (*afk*; T'jampens et al. 1999). As for the profilins, the plasmodial-specific *proP* is not found in sexually developing cells but in apogamically developing cells, reaching its maximum levels in the plasmodial stage (Bailey et al. 1999). In turn, the developmentally regulated myosin D (*mynD*), is similar to the tail of the myosin II heavy chain, and colocalizes with actins in the microfilament network (Bailey et al. 1999). Coronin, on the other hand, is a protein found in various eukaryotes involved in several cytoskeletal- based processes, such as cell migration, cell division and membrane trafficking (Minami et al. 2009). The coronin from *Physarum* is a 449 amino acid protein encoded by a single copy gene, and it possess 60% identity with its *Dictyostelium* ortholog, a protein that has been linked to the G- protein mediated signal transduction (Minami et al. 2009). Taking together, these observations support that changes in expression of genes coding actin-binding proteins are coincidental with alterations in the cell organization and behavior, *e.g.*, the transformation of an amoeba into a flagellate form involves the reorganization of the actin cytoskeleton. However, it remains to be studied if the differential gene expression is the cause of the alterations of the actin organization (Bailey 1997; Bailey et al. 1999).

Calcium-Binding Proteins. Although its precise function is not yet clear, calcium surely plays a key role during the differentiation of the slime mold, as it is released from plasmodia right after the exposure to light, concentrations of calcium and malate are necessary for autocrine signaling in the absence of light during sporulation (Renzel et al. 2000), and high concentrations of calcium inhibit the actin – myosin interaction (Nakamura and Kohama 1999). Genes encoding calcium binding proteins so far described for *Physarum* include several types of spherulins (Savard et al. 1989; Pinchai et al. 2006), the regulated in development *redB* gene product (Bailey et al. 1999), and LAV1-2 (Laroche et al. 1989). The spherulins entail a heterogeneous group of mRNAs first detected during plasmodial encystment (spherulation), whose stability is calcium-dependent (Savard et al. 1989; Pinchai et al. 2006). From the spherulin cDNAs cloned, two have been more carefully studied, the spherulins 3a and 3b, both sharing sequence similarities between them, and containing Greek key calcium-binding domains from the βγ-crystallins (Savard et al. 1989; Pinchai et al. 2006), a group of proteins found in vertebrate eye lenses (Slingsby et al. 2013). The regulated in development transcripts, *redA* and *redB*, were isolated from a cDNA library obtained from apogamically developing cells, sharing the same expression patterns: higher levels during apogamic development, low expression in macroplasmodia, and no detection in amoebae (Bailey et al. 1999). Only one of these, *redB*, contained two calcium- binding domains, and shared significant identity with sarcoplasmic calcium- binding proteins from invertebrates (Bailey et al. 1999). Finally, LAV1-2 is a plasmodial- specific RNA of unknown function, whose gene product acts as a substrate of transaminases. The sequence of LAV1-2 contains an EF-hand type domain with a calcium- binding loop, and its calcium-binding activities have been observed in vitro, although its function is unknown (Laroche et al. 1989; Mottahedeh and Marsh 1998; Iwasaki et al. 1999).

Signal Transduction. Three groups of GTP- mediated signaling genes linked to proliferation and differentiation have been extensively studied in *Physarum*: *lig1* (Kroneder et al., 1999), the nitric oxide synthases A and B (Golderer et al., 2001), and the GTP cyclohydrolase I, a key enzyme that is part of the folate and biopterin biosynthesis pathways (Werner-Felmayer et al., 1994). The light-induced gene *lig1* is a homolog of the yeast gene *hus1,* a component of an evolutionarily conserved, genotoxin-activated checkpoint complex that is involved in the cell cycle arrest in response to DNA damage (Kroneder et al. 1999; Weiss et al. 2000). *lig1* is expressed in the starved plasmodia, and induced up to 60-fold upon the photoinduction (Kroneder et al., 1999). On the other hand, the nitric oxide synthases *nosA* and *nosB* are inducible isoenzymes whose sequences lack of the calcium- dependent region observed in the structures of orthologs in other species, and their mRNA levels are strongly induced during sporulation, specifically at the end of the starvation (Golderer et al. 2001).

The activity of these three genes is closely connected: The inhibition of either the nitric oxide synthase activity, or the formation of cyclic GMP, impairs the *lig1* expression and prevents sporulation (Kroneder et al. 1999; Golderer et al. 2001). In addition, during starvation, the addition of glucose to the growing media, suppresses the nitric oxide synthase activity, while at the same time induces the expression of the GTP cyclohydrolase gene (Golderer et al. 2001). Furthermore, the nitric oxide synthases use biopterin as cofactor, and it has been observed that the GTP cyclohydrolase controls the supply of biopterin (Golderer et al. 2001). Thus it is expected that they belong to the same gene regulatory network controlling the differentiation of the slime mold (Kroneder et al. 1999; Golderer et al. 2001; Marwan 2003).

Epigenetic Modifications. The *Physarum* genome is significantly methylated, with patterns that remind those of vertebrates –approximately 7% of cytosines are modified, and these are mostly clustered in HpaII- repeated regions (Gordon and Hardman 1988). About a third of the *Physarum* genome is composed of repetitive elements, which are mostly clusters of retrotransposon- like sequences (Rothnie et al. 1991), and many of these sequences might be controlled by epigenetic modifications. The slime mold genome also contains HTF islands (*HpaII* tiny fragments), similar to those found in vertebrates, but here in contrast, almost a half of these segments are derived from rDNA minichromosomal regions, and are mostly unmethylated (Gordon and Hardman 1988). It is very likely that these methylation levels are transient in many cases, changing throughout the developmental cycle (Fronk and Magiera 1994), and in fact, it has been observed that DNA methylation inhibitors (azacytidine, aza-deoxycytidine, L-ethionine and S- adenosyl homocysteine) prevent sporulation (Hildebrandt, 1986). Other less common DNA modifications, like N6-methyl-adenine (m6A), have also been reported in *Physarum* to be present in cyst but not in growing cell DNA (Ratel et al. 2006). To date, DNA modifications have been reported for only one gene involved in the differentiation of the slime mold, spherulin-4. This gene displays specific 5-methyl-cytosine (m5C) patterns correlated to different sporulation stages, and these levels might change throughout the developmental cycle (Fronk and Magiera 1994). On the other hand, it is well known that DNA methylation is typically accompanied by other types of epigenetic marks, such as histone modifications (Strahl and Allis 2000) and small RNAs (Grewal and Elgin 2007), and together form complex regulatory networks. In this regard, some chromatin marks like histone H1 methylation (Jerzmanowski and Moraczewska 1988), histone H4 acetylation (Waterborg et al. 1983; Pesis and Matthews 1986; Loidl and Gröbner 1986), histone H4 methylation (Waterborg et al. 1983), simultaneous changes in acetylation patterns in H3 and H4 histones (Waterborg and Matthews 1984), and have been observed during cellular differentiation in *Physarum*. Increased levels of the histone H1 during early spherulation (Heads and Carpenter 1990), and changes in the histone acetyl transferase activities (Lusser et al. 1997) have been also reported. Furthermore, the RNA interference mechanisms has been also observed in the slime mold (Haindl and Holler 2005). However, the current knowledge of chromatin regulation in this organism rather insufficient, and thus many molecular regulation phenomena that could be better explained at the epigenetic level, such as developmental pathways, have not been described yet. Furthermore, the process of sporulation is a good candidate for the control via epigenetic regulation, because it is driven by environmental stimuli and requires rapid changes in expression before reproduction, typical for this type of expression control systems (Jaenisch and Bird 2003).

Other Stage- Specific Proteins and Metabolites. In addition to the above mentioned genes, other molecules with unknown function have been identified as cell type- specific in the slime mold: The “hydrophobic abundant proteins,” *hapP* and *hapS* (Martel et al. 1988), as well as malate and beta- poly L-malate (Renzel et al. 2000; Pinchai et al. 2006). In the former case, the hydrophobic abundant proteins were first detected on plasmodial cDNA libraries, in a similar manner to the discovery of *LAV1-2*, and were exclusively distinguished in two cell stages, the plasmodium- specific *hapP*, and another observed only during sporulation, the *hapS*. Although *hapP* is abundant both in growing and starving plasmodia, photoinduction triggers the degradation of *hapP* and leads to sporulation. Conversely, the sporulation- specific hapS is absent in growing or starved plasmodia, and appears after 12 hours of light induction (Martel et al. 1988). HapP can be found in at least two allelic forms, assigned *hapP1* (carried by plasmodia from the M3C-O strain) and *hapP2* (present in LU648 x LU688 plasmodia). Both alleles code for similar proteins of 187 amino acids which are 90.4% similar to each other, and have an identity of 50% to HapS. However, neither of these encoded products share similarity to other known proteins (Lépine et al. 1995). The function of the *hapS* and *hapP* genes and their products remains to be discovered, although it has been hypothesized that hapS might encode a cell wall protein (Martel et al. 1988).

Moreover, Renzel et al. (2000) established a new manner to achieve sporulation in the light or in the dark; they accompanied their methodology with the use of a solid matrix in order to test for the secretion of sporulation- promoting factors. In this way, they found that three substances were involved in sporulation: calcium (described above in the "calcium binding proteins" paragraph), malate, and beta- poly L-malate (PMLA). PMLA is a water-soluble molecule synthesized from malate, that accumulates in the nuclei of plasmodia in amounts similar to those of the DNA and histones, and whose abundance in plasmodia is associated to the NKA48 transcript. The NKA48 sequence resembled the spherule- specific transcript spherulin 3a, and therefore it was named spherulin 3b (Pinchai et al. 2006). Renzel et al. (2000) observed that calcium and malate promoted sporulation in absence of light, while the polymalate acted as a sporulation control factor, who also might work as a source for calcium ions and malate. Nevertheless, the precise functions of PMLA in the sporulation process remains to be studied in detail.

Physarum as a model organism.

Since the second half of the last century, the slime mold has been not only an important model in several areas of Biology, but also in Physics and Computer Science. As happens with many microorganisms, *Physarum* is easily grown either on agar plates (as plasmodium), or in a culture broth (as amoebae or microplasmodia), and in this way large amounts of cells can be obtained with ordinary laboratory equipment from the many available well- characterized strains (Kohama and Nakamura 2001; Anderson and Dee 1990). Novel strains and mutants can also be generated by several means: cell fusion, complementation, transfection of plasmid vectors, etc. (Anderson and Dee 1990; Marwan 2003). As mentioned before, plasmodial nuclei are synchronous, *i.e.*, all the nuclei from a big plasmodial mass stay in the same physiological state. Furthermore, cell differentiation (sporulation) can be easily induced artificially by exposure to red light, and it is also highly synchronous (Martel et al. 1988). All these features are complemented by the disposal of standard molecular biology tools and methods, such as RNA interference, microinjection, transformation and cDNA libraries (Anderson and Dee 1990; Marwan 2003; Haindl and Holler 2005; Glöckner et al. 2008). These advantages have allowed to employ *Physarum* in areas as diverse as cell motility, cell differentiation, RNA editing, DNA replication, artificial intelligence and other topics, detailed below.

Cell Motility. Motility has been defined as "*the ability of living systems to exhibit motion and to perform mechanical work at the expense of metabolic energy*" (Allen 1981), and includes a wide range of biological processes, including cytoplasmic streaming, organellar and flagellar movement, cytokinesis, contractility, etc. Studies of the movement of *Physarum* date back 1937, with the classical works of Seifriz on shuttle streaming, and later continued by his student Kamiya, with measurements of the motive forces, as well as the analysis of the effects of diverse factors and substances on the motility of the slime mold (Seifriz 1937; Kamiya 1940; Allen 1981). Actomyosin-like solutions induced by ATP were then described in *Physarum* extracts, when Loewy employed the slime mold as the first nonmuscle motile system to study the muscle biochemistry (Loewy 1952;

Allen 1981). Afterwards, Huxley and collaborators (1970) showed the evolutionary conservation of the interaction between skeletal proteins, when they matched rabbit myosins with actins extracted from the slime mold (Nachmias and Huxley 1970). Later, mutants defective in cell movement were developed (Jacobson and Dove 1975), and the properties of the streaming in this organism have been also studied at the single-cell level (Wohlfarth-Bottermann 1979). Over many years, the cytoskeleton of *Physarum* has stood as an important focus of research about the roles of actin, myosin, tubulin and other cytoskeletal proteins in motility (Burland et al. 1993).

Cell Differentiation. Sporulation in the slime mold displays typical features present in the process of differentiation: competence, irreversible commitment, morphogenesis, and metabolites and gene expression unique for the differentiated state (Sauer et al. 1969). The different cell types and developmental pathways of the slime mold provide a natural resource for studying the differentiation in a simple manner (Burland et al. 1993). In addition, understanding the development of individual cells of multicellular organisms (which for many reasons cannot be easily studied in isolation) in a simpler system, such as the sporulation of the slime mold, may help to clarify the precise mechanisms employed by higher eukaryotes (Bailey 1997). Since sporulation can be easily induced by starving a plasmodium and then exposing it to light, and the conversion of plasmodia into the differentiated state allow biochemical approaches, these characteristics have established *Physarum* as a model for studying the differentiation in eukaryotic cells in the form of events that refer to a defining startting point (Sauer et al. 1969; Burland et al. 1993; Bailey 1997). In this respect, many cell- type specific and differentially expressed genes have been associated to the process of sporulation, with many of these genes coding for cytoskeletal proteins such as tubulins, profilins and actin-binding proteins, making the slime mold also a suitable model to study these proteins (Bailey 1995; Bailey 1997; Glöckner et al. 2008). The identification of developmentally regulated genes, and the nuclei synchronization inside a plasmodium, later allowed to study the relationship between the differentiation status and the DNA replication (see below; Pierron et al. 1989; Maric et al. 2003)

DNA replication. Eukaryotic genomes replicate through steps without changes in temporal order, a fact that was first demonstrated in the slime mold (Braun et al. 1965). Since then, the plasmodium of *Physarum* has been recognized as a model for studies of the control of cell division, because of the simplicity of the cell fusion methods, even between cells at different cell cycle stages, and the fact that the nuclei enter S-phase immediately after mitosis, which occurs synchronously (Braun et al. 1965; Sachsenmaier et al. 1972). In addition, the genome contains many repetitive regions, and therefore the proteins required for replication may be available in high amounts, which can be easily prepared from synchronous extracts at specific stages; this allowed the identification of the initiation from several close origins of replication for the first time in ribosomal DNA from *Physarum* (Vogt and Braun 1977; Daniel and Johnson 1989). To date, research focused mostly in the relationship of the histone modification status and the replication firing, and the developmental usage of replication origins (Loidl 1988; Borde and Duguet 1998; Thiriet and Hayes 2005; Thiriet and Hayes 2009; Pierron et al. 1989; Cunningham and Dove 1993; Maric et al. 2002; Maric et al. 2003; Bénard et al. 2007). In this respect, the natural synchrony of the cell cycle in the plasmodium allowed the mapping of replication origins associated to highly expressed genes (Bénard and Pierron 1992). Nuclear synchrony also enabled the discovery of cell cycle- dependent telomerase activation (Shimada et al. 1997) and topoisomerase II sites (Borde and Duguet 1998). Later, the comparison of replication patterns in amoebae and plasmodia showed a reprogramming of the cell cycle S-phase associated to the reprogramming of transcription during the differentiation, *i.e.* genes that display cell- type specific gene expression, are actively replicated from promoter-proximal origins in cell stages where they are highly expressed (Maric et al. 2003; Bénard et al. 2007). Other studies involved the analysis of structural features, such as the frequency of formation of the post-replicative X-shaped DNA molecules (*e.g.*, Maric et al. 2010).

RNA editing. The phenomenon now known as RNA editing was first described for the mitochondrial cytochrome oxidase *cox2* mRNA in *Trypanosoma* (Benne et al. 1986). A few years later, it was also observed in the mitochondrial ATPase subunit 1 *atp1* mRNA of *Physarum* (Mahendran et al. 1991). RNA editing involves modifications of mRNA molecules (insertions, deletions and substitutions), which produce final RNAs that differ from the original genomic template sequences. RNA editing has been described in many species, and in slime molds it occurs exclusively in the mitochondrion, where up to 25 nucleotides are edited in almost every gene (Bundschuh et al. 2011b). There are at least four types of RNA editing in *Physarum*: The most common form is the insertion of individual cytosines, and other possible modifications are the insertion of individual Us or dinucleotide pairs, substitutions of Cs by Us, and deletions. Interestingly, RNA editing in *Physarum* is highly accurate (Visomirski-Robic and Gott 1995), and occurs co-transcriptionally (Visomirski-Robic and Gott 1997), and therefore it must be associated to the RNA polymerase machinery, making this system the only non- viral co- transcriptional RNA editing process known so far (Knoop 2011). However, the mechanism of site- recognition and the editing machinery itself are not known so far, making this an active field of research (Knoop 2011; Chen et al. 2012).

Epigenetics. *Physarum* stands out as a promising model organism for epigenetic studies as well, because its genome is significantly methylated, with patterns that remind those of vertebrates –approximately 7% of cytosines are modified, and these are mostly clustered in HpaII- repeated regions. This hypermethylated regions together comprise ca. 20% of the genome (Gordon and Hardman 1988). For a brief review of most methylation and histone mark studies, see page 6. In addition, recent developments will enable future assessment of the histone marks in the course of the developmental stages in this organism. For instance, the study of chromatin regulation in this species would be almost impossible because of the current lack of antibodies directed against specific histone modifications, but it was demonstrated that *Physarum* can not only take up foreign histones (both native and recombinant molecules from *Xenopus* expressed in *E.coli*; Prior et al. 1980; Thiriet and Hayes 1999) but also it can incorporate them into its chromatin, which will allow to monitor the differential binding of these proteins to the studied DNA regions, and to study the influence of histone modifications on the regulation of gene expression (Thiriet 2004; Thiriet and Hayes 2005). Furthermore, other recent methodological advances, such as the development of RNA interference in *Physarum* (Haindl and Holler 2005), or epigenetic tools created in related organisms (such as *Dictyostelium*), will also be helpful to address associated biological questions (Kaller et al. 2006).

Gravitational Biology (Astrobiology). Multi- and unicellular organisms typically display different levels of gravisensitivities, and use the direction of the gravity vector for spatial orientation (*gravitaxis*). *Physarum* is no exception, as it reacts to many environmental stimuli, such as light, chemicals, but also to gravity (Block et al. 1995). On an early experiment, slime molds were sent to orbit during the Kosmos 1129 (Bion 5) unmanned space mission, which was part of the Soviet biosatellite program. This experience proved that the slime mold reduced its growth but maintained its migration ability after exposure to microgravity (Tairbekov et al. 1981). Later, demonstrations of gravisensitivity (Block et al. 1986) and gravitaxis in this organism (Wolke et al. 1987), paved the way for its use as a model in this field. *Physarum* was then chosen as the subject for studying the effects of microgravity on single cells for four missions during the Space Shuttle program: Spacelab D1 (STS-61A Challenger, 1984), IML-1 (STS-42 Discovery, 1992), IML-2 (STS-65 Columbia, 1994) and BRIC-06 (STS-69 Endeavour, 1995). In these experiments, it was observed the gravitaxis and the rhythmic contraction activity in weightlessness conditions (Block et al. 1986; Block et al. 1994), its low acceleration gravisensitivity (Block et al. 1995), and the involvement of cAMP in the signal transduction associated to the perception of gravity in the slime mold (Block et al. 1998). Although *Physarum* has not been employed as a research model in recent studies, it will presumably stay as a future choice for experimentation on astrobiology because of its long lasting stimulus response, which is revealed in multiple manners (oscillating contractions, changes in second messenger levels, differential gene expression, etc.; Block et al. 1995; Block et al. 1998; Putzer et al. 1984; Bernier et al. 1986; Sweeney et al. 1987; Martel et al. 1988).

Behavioral and Computer Sciences. When the slime mold looks for food supplies, it develops tubular structures that link the provisions it finds through a cost- efficient, robust network (Navlakha and Bar-Joseph 2011). *Physarum* requires these networks not only to transport the resources, but also to store related information, learn and recall associated events (Reid and Beekman 2013; De la Fuente et al. 2013), and even to solve complex nutritional problems and develop balanced diets (Dussutour et al. 2010). Furthermore, slime molds are able to find the minimum- length solution in different mazes (Nakagaki et al. 2000; Reid and Beekman 2013), construct robust networks to maximize nutrient uptakes (Nakagaki et al. 2004), and recall environmental stimuli and adaptation to changes (Saigusa et al. 2008), which suggest the existence of simple forms of biological devices for intelligence, and memory processing and storage (De la Fuente et al. 2013). These problem solving strategies in *Physarum* have been implemented into mathematical models, and applied to develop novel algorithms for network design (Tero et al. 2010), and its oscillatory behavior of adaptation to stimuli used to control the locomotion of a robot (Tsuda et al. 2007). Therefore, using biological processes such as the foraging behavior of the slime mold, modeled as natural algorithms, has the potential of solving complex real-world problems under a myriad of different conditions (Navlakha and Bar-Joseph 2011).

Transcriptomes

The transcriptome is defined as the population of all RNAs in the cell, or its RNA complement. Transcriptomes are the first phenotypic manifestation of the genome, and as such, are the basis of cellular specificity and higher-order phenotypes, through the mediation of all phenotypic changes encoded in the DNA sequence. This unfolding of instructions is started by the transcription of DNA into RNA, followed by the processing of RNA transcripts into functional mature RNAs. Recently, a vast number of novel RNA species have been described. Some of these species belong to novel splice forms of known protein-coding genes, but others do not seem to encode proteins, and correspond to novel families of small or multi-exonic noncoding RNAs. The specific roles of most of these species are still unknown, although many appear to be involved in the regulation of gene expression. Therefore, RNAs not only function as carriers of information from DNA to proteins, but also they play complex roles in cellular homeostasis and biological regulation (Guigó 2013).

The aims of studying the transcriptome include cataloguing all transcript species, to establish the gene structure (exons, introns and untranslated features), as well as quantifying the changes in transcript expression under different conditions. For the sake of simplicity, in the following paragraphs I will use the definitions by Guigó (2013) of protein-coding transcriptome –the set of genes that encode proteins- and the non-coding transcriptome –the set of transcripts that are not translated into proteins. I will also include the concepts of *reference transcriptome* –the set of all genes and transcripts potentially encoded in a genome- and *expressed transcriptome* –the set of genes and transcripts that are expressed in a given condition, and which are then responsible for cellular specificity (Guigó 2013).

Experimental methods for studying the transcriptome

Currently, the most commonly used experimental approaches to study the transcriptomes, include hybridization- and sequence- based technologies. The hybridization- based methods, include the genomic tiling microarrays (Schena et al., 1995), which employ a set of overlapping oligonucleotide probes that represent a cDNA subset or the whole transcriptome at a very high resolution, and the sequence- based approaches comprise the expressed sequence tag (EST) library construction (Adams et al. 1991), the tag- based methods, and more recently the next- generation sequencing technologies, RNA-seq in particular (Table 1; Nagalakshmi et al. 2008; Guigó 2013).

Table 1. Experimental methods for studying whole transcriptomes (Modified from Wang et al. 2009).

Technology	**Microarrays**	**EST library**	**RNA-Seq**
Principle	Oligonucleotide Hybridization	Sanger sequencing	Next generation sequencing
Resolution	Several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	No
Background noise	High	Low	Low
Simultaneous map and expression	Yes	Limited	Yes
Dynamic range for expression	Up to a few-hundredfold	Not practical	Over several thousandfold
Isoforms detection	Limited	Yes	Yes
Allelic expression detection	Limited	Yes	Yes
Required amount of RNA	High	High	Low
Study cost	High	High	Relatively low

Early approaches to study transcriptomes involved the analysis of total RNA, often comparing different organisms, growing conditions, tissues, cell types and disease states, in order to identify and quantify the expression of a given gene of interest (Morozova et al. 2009; Guigó 2013). The first of these studies (commonly known as *candidate gene approaches*), used a method named Northern blot, which consisted of a low throughput approach to identify RNAs by hybridization to radioactive probes (Alwine et al. 1977). The complexity of the method, and the requirement of large amounts of the analyzed nucleic acids, limited the Northern blot to the analysis of few known transcripts per experiment (Morozova et al. 2009). Later, the development of the polymerase chain reaction (Saiki et al. 1988), and particularly of the reverse transcriptase methods (RT-PCR), reduced the dependence on large amounts of starting materials, while at the same time increased the throughput. These methods however, are still limited to a maximum of hundreds of transcripts analyzed at the same time (Morozova et al. 2009).

Microarrays

Also known as DNA chips, microarrays are collections of microscopic spots attached to a solid surface, where each spot contains thousands of copies of the same DNA molecule ("*probe*"), representing each spot a given gene. Each microarray slide is employed to hybridize cDNAs present in a target sample, whose annealing is then captured and quantified by light- detection methods (Guigó 2013). In the last two decades, the microarrays have been the most commonly used method to monitor the amounts of transcripts at a whole-transcriptome level, effectively replacing the single- gene approaches by enabling the simultaneous characterization of thousands of RNAs (Morozova et al. 2009; Guigó 2013). However, the microarrays are not exempted of problems, and their major limitations can be summarized in three categories: (*i*) they are unable to detect novel transcripts or those that are not previously captured during the fabrication of the array; (*ii*) it is difficult to distinguish alternative forms of transcripts, as the probes usually cover small regions (typically the 3' ends); and (*iii*) the quantitative data obtained is noisy, because the transcript amount is inferred from the intensity of hybridization, which is sensitive to inhomogeneities of the chip surface properties (Morozova et al. 2009; Guigó 2013).

EST Libraries

Traditionally, the manner of studying RNAs involves first the synthesis of complementary DNAs (cDNAs) by reverse transcription, using the RNA molecule as a template. These cDNAs can be cloned into appropriate vectors, from which these molecules can be sequenced. Using oligonucleotides that are complementary to the poly-A tail present in eukaryotic mRNAs, cDNA libraries can be created, entailing copies of transcripts expressed in a given cell type or condition, and these libraries can be subsequently sequenced. However, sequencing of large numbers of full-length cDNAs is costly and labor intensive. Instead, a common strategy to analyze cDNA libraries is the single-pass sequencing of random cDNA clones, which produces a collection of partial sequences from specific transcripts, termed "expressed sequence tags" (ESTs;

Adams et al. 1991). In this way, it was possible to overcome the cost limitation of sequencing full-length cDNAs, although this method was still too expensive and complex to be performed routinely on a whole transcriptomic scale. Besides, when a very large sequencing capacity is not available, the wide range of mRNA abundances makes random sequencing of cDNA clones inefficient for discovering rare transcripts, because the most abundant cDNAs will be predominantly sequenced. Nevertheless, and in spite of being low throughput, mostly non-quantitative, and relatively expensive, as compared with the current sequencing technologies, usually EST libraries are still the first approach to study the transcriptome (Wang et al. 2009; Morozova et al. 2009; Guigó 2013).

Tag-Based Approaches

These methods are high- throughput and provide information on the gene expression level. Tag- based approaches include SAGE ("serial analysis of gene expression"), which uses 14-20 bp sequence tags from the 3' ends of transcripts, to measure expression levels (Velculescu et al. 1995); CAGE, which uses the 5' end instead; and MPSS, which determines 15-20 bp signatures from cDNA ends using multiple cycles of cleavage and ligation. The development of SAGE was an important advance in transcriptomics as it enabled the use of Sanger sequencing for expression profiling. Unlike microarrays, tab-based approaches were able to detect novel transcripts and splice variants, as well as allowing their direct quantitation; however as tag-based methods rely on Sanger sequencing, their implementation could be expensive, and also they require complex cloning procedures. Furthermore, in many cases the short tags cannot be uniquely mapped in genomes, and it is difficult to distinguish between transcripts with similar sequences (Wang et al. 2009; Morozova et al. 2009).

Microarrays, EST library sequencing and tag- based technologies have been complemented in recent years by the development of next generation sequencing (NGS) methods, and especially of the deep RNA sequencing, or RNA-seq (Nagalakshmi et al. 2008). RNA profiling by RNA-seq through multiple conditions of expressed transcriptomes (cell types, cell cycle stages, cellular compartments, *etc*.) is revealing an unexpected complexity of the eukaryotic transcriptome. This, combined with other molecules involved in the RNA synthesis and processing (epigenetic modifications, transcription factors, enzymes, regulators, *etc*.), gives now a more complete view of the transcriptional activity inside the cell, allowing the application of systems biology approaches to the modeling the pathways involved in RNA metabolism (Guigó 2013).

Next- Generation Sequencing

Since its arrival, the dideoxynucleotide method (Sanger et al. 1977) has dominated the sequencing approaches, and it has led to monumental accomplishments, such as the first reported genomes (Fleischmann et al. 1995; Adams et al. 2000; Lander et al. 2001). However, limitations mainly in producing large volumes of data cheaply, motivated the development of several strategies in the recent years, collectively known as *next- generation sequencing* (NGS) technologies (Metzker 2010). These strategies rely on a mixture of template preparation, sequencing and imaging, and genome alignment and assembly methods, in several commercially available platforms (Table 2 and Figure 3).

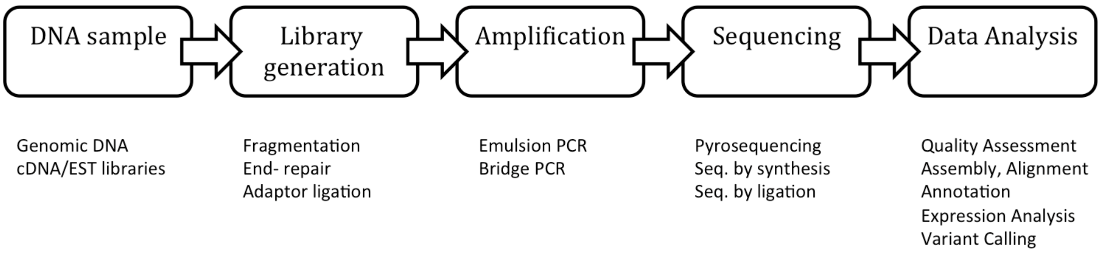


Figure 3. A typical next generation sequencing (NGS) flowchart. Adapted from several sources (Metzker 2010; Rehm et al. 2013).

Template preparation. Current methods of template preparation comprise breaking the nucleic acids into smaller pieces from which either fragment templates (by random shearing) or mate-pair templates (from circularized DNA) are made. The template is then attached or immobilized to a solid support. This immobilization of millions of separate template fragments allows thousands to millions of sequencing reactions to be carried out simultaneously. As most imaging devices do not possess the ability to detect single molecule signals, reactions from amplified templates are required. Amplification of immobilized templates then occurs either on as single strands captured on beads ("emulsion PCR", or emPCR), or directly on templates covalently attached to high-density glass slides ("solid-phase amplification"). This last strategy is employed on the popular Illumina platform (Bentley et al. 2008). The emulsion PCR beads can then be immobilized on a glass surface through chemical crosslinking (as in the Life Technologies APG SOLiD method), or deposited into picotiter plate (PTP) wells (Roche-454 method; Margulies et al. 2005). Other early methods performed the sequencing without relying on previously amplified material, to avoid the bias introduced by PCR methods. For example, in the Helicos platform this was achieved through immobilization of single molecule nucleic acids, either primers or sequencing templates; and in the Pacific Biosciences system, where the immobilized molecules are single DNA polymerases (Metzker 2010).

Sequencing. Clonal amplification results in a population of identical template molecules, each of which has gone through the sequencing reaction. Upon imaging, the observed signal is a consensus of the bases or probes added to the same templates for a given cycle; this puts a greater demand on the efficiency of the addition process, and incomplete extension of the template might result in a lagging-strand dephasing, *i.e.*, signal decay over time. Addition of multiple bases or probes can also occur in a given cycle, producing dephasing at the leading-strand. Furthermore, signal dephasing increases noise during fluorescence imaging, causing base-calling errors and shorter reads (Metzker 2010).

Three approaches are currently employed for high throughput sequencing, to overcome dephasing and base-calling errors: Cyclic reversible termination, sequencing by ligation, and pyrosequencing. Cyclic reversible termination (CRT), as the name indicates, uses reversible terminators in cyclic method that requires nucleotide addition, fluorescence imaging and cleavage. CRT involves three steps: (*i*) a DNA polymerase incorporates just one fluorescent nucleotide; (*ii*) the remaining unincorporated nucleotides are washed away, and imaging is then carried out to determine the identity of the incorporated nucleotide; and (*iii*), cleavage of the terminating or inhibiting group and the fluorescent dye. Additional washing is performed before the next CRT cycle. One instance of this approach is the CRT cycle used by the Illumina platform (Figure 4; Bentley et al. 2008) during sequencing and imaging, which detects four colors utilizing two lasers by total internal reflection fluorescence (TIRF). Fragments produced on the Illumina systems are typically of a hundred bases, with total outputs in the range of 3 – 20 Gb. The most common errors in this system are substitutions and underrepresentation of AT-rich and GC-rich regions (Metzker 2010).

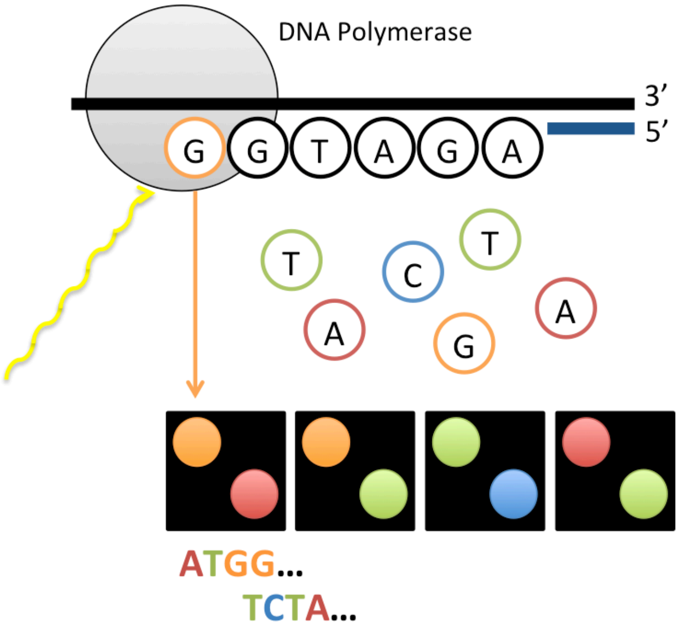


Figure 4. The Illumina method. This approach uses sequencing by synthesis and fluorescently labeled nucleotide analogues, that are incorporated in reversible reactions. These reactions occur in millions of spots (dark panels in the graphic), allowing the sequencing of many fragments simultaneously. Modified from Morozova et al. (2009).

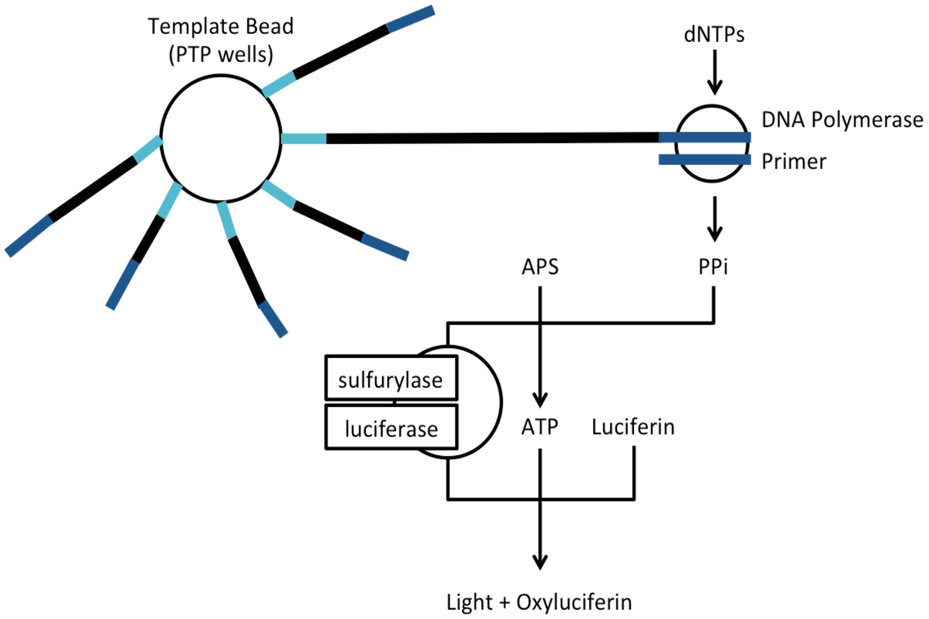


Figure 5. The Roche – 454 method. Beads with template DNA, amplified by emulsion PCR, are incorporated into individual picotiter plate (PTP) wells, together with additional beads, coupled with sulphurylases and luciferases. Then dNTPs are added across the PTP wells, and the inorganic phosphate (PPi) released starts an enzymatic cascade, that ends with the generation of light, that is detected by a charge-coupled device (CCD) camera from each PTP well. Adapted from several sources (Metzker 2010; Mutz et al. 2013).

On the other hand, the sequencing by ligation (SBL) approach differs from CRT in its use of a DNA ligase. SBL employs either one- or two-base encoded probes, and involves the hybridization of the fluorescent probe to the complementary sequences adjacent to the primed template, followed by the addition of the DNA ligase, which will join the dye-labeled probe to the primer. Probes that were not ligated are washed away, and the identity of the ligated probe is determined by fluorescence imaging. This cycle can be repeated either through cleavable probes (to remove the fluorescent dye and regenerate a 5' phosphate group for subsequent ligation steps), or by removing and hybridizing a new primer to the template. The SOLiD platform (Life Technologies) uses the SBL approach, with templates amplified by emulsion PCR (Metzker 2010).

Finally, the pyrosequencing, or single-nucleotide addition method (Figure 5), is a bioluminescence, non-electrophoretic approach that measures the release of inorganic pyrophosphate, and proportionally converting it into visible light using a series of enzymatic reactions (Margulies et al. 2005). The addition of single deoxynucleotide triphosphates (dNTPs) in limiting amounts allows controlling the DNA polymerase extension, and the order and intensity of the light peaks are recorded as flowgrams, which reveal the original DNA sequence. Margulies et al. (2005) described the first NGS platform, integrating pyrosequencing over picotiter plate (PTP) wells, which will be later commercially available as the Roche-454 technology. In this platform, DNA templates are fixed to beads and amplified by emPCR inside the PTP wells. Then smaller beads are loaded into the wells containing the amplified templates, that carry both sulphurylase and luciferase enzymes attached to them to facilitate light production. This is followed by a stream of individual dNTPs across the wells, which are dispensed in a predetermined sequential order, and the generated bioluminescence is captured with a charge-coupled device (CCD) camera (Margulies et al. 2005; Metzker 2010). Output fragments are in the range of several hundreds of nucleotide bases, with a total output of 0.6 Gb (Table 2); the platform has difficulties with homopolymeric regions, and the most common errors are insertions, followed by deletions (Metzker 2010).

Applications. The potential and use of the NGS technologies is akin to the early days of the polymerase chain reaction. Some example applications include: the discovery of sequence variants, through resequencing of targeted regions of interest, or whole genomes; the sequencing and de novo assemblies of bacterial and non-model organisms; the genome-wide profiling of epigenetic marks and chromatin structure (through *ChIP-seq*, *methyl-seq*, *DNase-seq* and others); species classification and/or gene discovery by metagenomics studies; analyzing mutations and variants in species populations; examination of personal genomes; studying the evolutionary relationships of ancient genomes; assessing the role of non-coding RNAs (ncRNAs); the qualitative and quantitative cataloguing of transcriptomes of cells, tissues and organisms (commonly known as *RNA-seq*; Table 1 and following paragraphs); between others (Nagalakshmi et al. 2008; Wang et al. 2009; Metzker 2010).

Table 2. Commercially available next generation sequencing platforms. Adapted from several sources (Morozova et al. 2009; Nowrousian 2010; Metzker 2010; Mardis 2011).

Platform	**Roche 454**	**Illumina**	**Life Tech. SOLiD**
Sequencing Principle	Pyrosequencing	Sequencing by synthesis with cyclic reversible terminators	Sequencing by ligation
Sequencing reaction	Polymerase-mediated	Ligase-mediated	Polymerase-mediated
Template amplification method	Emulsion PCR	Bridge PCR	Emulsion PCR
Incorporated chemicals	Unlabelled nucleotides	Fluorescent oligonucleotides	End-blocked fluorescent nucleotides
Post incorporation method	Not applicable	Chemical cleavage to remove fluorescent dye and 3′ end of oligonucleotide	Chemical cleavage to remove fluorescent dye and 3′ blocking group
Detection method	Light emission	Fluorescent emission	Fluorescent emission
Error model	Substitution errors rare, insertion or deletion errors at homopolymers	End of read substitution errors	End of read substitution errors
Raw read accuracy (%)	≥99	≥98–99	≥99.94
Read length (fragment/paired end)	400 bp/variable length mate pairs	75 bp/ 50+25 bp	150 bp/ 100+100 bp
Total output (Gb)	0.6	3–20	50–100
Pros	Longer reads improve mapping in repetitive regions; fast run times	Currently the most widely used platform in the field	Two-base encoding provides inherent error correction
Cons	High reagent cost; high error rates in homo-polymer repeats	Low multiplexing capability of samples	Long run times

RNA sequencing (*RNA-seq*). This approach was developed to overcome the difficulties with the classic approaches to study the transcriptomes. RNA-seq, in simple terms, involves the sequencing and quantitative characterization of cDNA copies of RNA molecules, or in some cases to sequence raw unamplified RNA molecules directly, through next- generation sequencing (NGS) methods (Nagalakshmi et al. 2008; Mortazavi et al. 2008; Ozsolak and Milos 2011). More specifically, in RNA-seq a population of total or poly-A fractionated RNA is converted to a cDNA library, with adapters attached to one or both fragment ends. Then each molecule, with or without a previous amplification step, is sequenced in a high- throughput manner, producing short sequences from one or both ends (single- or paired- end sequencing, respectively). These short sequences, called reads, are typically 30 – 400 bp long, depending on the sequencing technology used. The obtained sequences are either aligned to a reference genome, or assembled de novo, and this produces a whole genome transcriptional map that entails both the transcriptional structures and the expression level for each mRNA (Figure 6; Nagalakshmi et al. 2008; Wang et al. 2009).

RNA-seq presents many advantages over the hybridization- and Sanger sequencing- based methods. For example, unlike the hybridization- based approaches, RNA-seq is not limited to detect transcripts from known genomic sequences. This makes RNA-seq practical for organisms whose genome sequences are yet to be determined, such as *Physarum,* or for non-model organisms. RNA-seq can reveal the location of transcriptional boundaries at single- base resolution -30 bp. RNA-seq reads are long enough to depict how two exons are joined, while longer and paired- end reads show connections between multiple exons. The single base resolution also allows RNA-seq to reveal sequence variations, i.e. single nucleotide polymorphisms (SNPs), in transcribed regions. Another advantage over hybridization methods is that RNA-seq has a very low or almost none background signal. This is due to the fact that DNA sequences can be mapped almost unambiguously to unique regions in the genome. On the other hand, RNA-seq, unlike microarrays, does not have an upper limit for quantification, as the expression correlates with the number of sequences obtained (Wang et al. 2009). One particularly powerful advantage of

RNA-seq over microarrays, is that it can capture the transcriptome dynamics across different cell types or conditions with simple normalization, and in a time-resolved manner (Mortazavi et al. 2008; Wilhelm et al. 2008; Cloonan et al. 2008). Moreover, microarrays lack sensitivity for genes expressed at low or very high levels. Consequently, RNA-seq has a larger and more dynamic range of detection of transcript expression levels than hybridization- based methods. Besides, RNA-seq studies have also shown to be highly accurate and reproducible for quantification of expression levels, for both technical and biological replicates. Finally, and because RNA-seq requires no previous cloning steps (and some RNA-seq technologies like those from Helicos need no preceding amplification step), it requires less starting RNA sample than Sanger sequencing and hybridization methods. Considering all these advantages together, RNA-seq is the first sequencing- based method that allows a whole transcriptome survey in a high- throughput and quantitative manner (Wang et al. 2009).

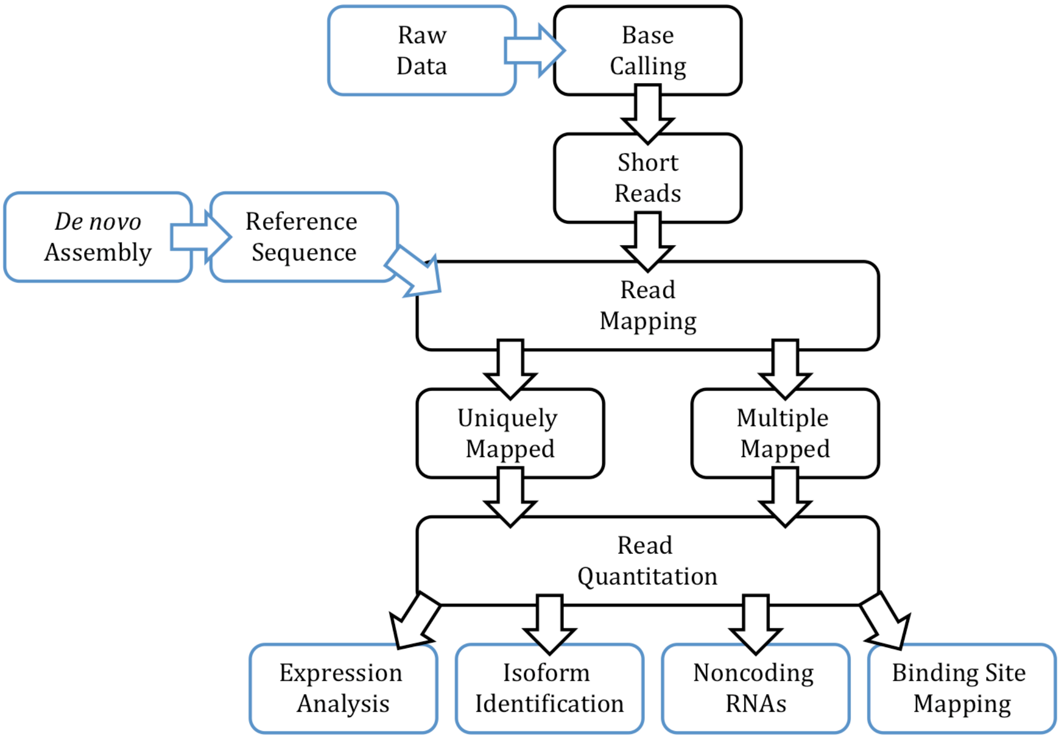


Figure 6. Analysis pipeline for RNA-seq data. Black boxes represent computational steps exclusive to the RNA-seq workflows. Modified from Mutz et al. (2013).

RNA-seq studies have generated an unprecedented view of the transcriptome and its organization, for several species and cell types. For example, before the development of RNA-seq, it was known that a larger than expected fraction of the genomes is transcribed, and in some particular cases such in yeast and human cells, RNA-seq have enabled the discovery of novel, distinct gene isoforms. Nevertheless, the transcriptional boundaries (start and ends of exons) of most genes have not yet been fully resolved, and the extent of heterogeneity due to splicing remains poorly understood. On the other hand, the single base resolution of RNA-seq have helped to revise existing gene annotations, including gene and exon – intron limits for known protein- coding genes, as well as the identification of novel transcribed regions, and the discovery of several novel features in the eukaryotic gene organization, e.g. many yeast genes overlap with others at their 3' ends. RNA-seq has also opened the possibility to unravel the extensive transcriptomic complexity, e.g. through transcription start site mapping, strand-specific measurements, gene fusion detection, small RNA characterization, quantitative examination of the splicing diversity by searching reads that map to splice junctions, and also by finding novel transcription regions that were not identified before using other methods such as transposon tagging and microarrays (Mortazavi et al. 2008; Ozsolak and Milos 2011).

Computational methods for studying the transcriptome

Computational methods are essential to investigate the transcriptional set of a given genome, and this is mostly because the data produced by transcriptome analyses cannot be processed without sophisticated computational resources, and experimental approaches are limited to analyze small fractions of reference genome. Currently, these methods employ many different heterogeneous sources of data, which are later processed and integrated as information, through a series of complex computational and statistical models. These sources generally entail three main types: (*i*) comparisons across genomes, at the sequence level; (*ii*) intrinsic features of sequences, such as specific signals or statistical biases on genomic regions; and (*iii*) transcribed sequences, such as those derived from cDNA sequencing (ESTs, RNA-seq) or proteins from related species (Guigó 2013). Regardless of the lab methodology used to obtain the transcriptomic data, the processing typically involves first the reconstruction of the transcriptome (*assembly*), to then proceed to assign functional features to each transcript (*annotation*), and analyze their expression patterns (Garber et al. 2011).

Assembly. The transcriptome reconstruction is a process in which a map of all transcripts, including their isoforms, is defined for a particular cell type or sample, and generally requires the assembly of Sanger sequencing fragments or RNA-seq reads into transcriptional units (Garber et al. 2011). The assembly is therefore a hierarchical data structure in which the sequence data is mapped or built into a putative transcriptome, by grouping reads and fragments into contigs, which entail multiple alignments of reads and consensus sequences (Miller et al. 2010).

Transcriptome reconstruction is a highly demanding computational task: It is affected by the several orders of magnitude that span the abundance of individual transcripts; generally, in the samples there is a mixture of mature and processed transcripts, increasing the difficulty in identifying the mature mRNAs; and genes can have isoforms, so it is challenging to establish which isoform produced a given sequencing read (Garber et al. 2011). Assembly algorithms (and their implementations) are therefore typically complex, and their operation can require high- performance computational platforms. For example, current high- throughput sequencing methods produce relatively short reads, and the assembly of these small fragments then requires a high coverage, in order to satisfy a minimum detectable overlap to allow the formation of contigs; high coverage in turn increases the complexity and the computational issues associated to large datasets. Nevertheless, the use of heuristics can help to overcome these and other common problems in real data, as well as the physical limitations of the computational equipments currently available (Miller et al. 2010).

The assembly of the short fragments (commonly known as *reads*) produced by the NGS methods is achieved by aligning the reads to a reference sequence (typically a pre-existing genome, "genome- guided methods"), or assembled *de novo* (without using a reference sequence, *"de novo* methods"; Metzker 2010). The decision of using either strategy is usually based on the intended biological application, the existence and completedness of a reference genome, the availability of sequencing and computing resources, the type of data generated by the sequencing approach, and the goal of the research project, as well as other technical considerations (cost, effort, time, *etc.*; Metzker 2010; Martin and Wang 2011). Regardless of the strategy used, the quality of both reference- based or *de novo* assemblies can be improved by increasing the read coverage and/or applying different platforms on the same target sequence (Aury et al. 2008; Reinhardt et al. 2009; Metzker 2010).

Genome- guided methods (also known as 'reference-based' or '*ab initio*' assembly), rely on a previously existing reference genome, to be used as a target where all the reads are mapped, and this coordinate system of spliced reads is then employed to build all the transcripts (Garber et al. 2011; Martin and Wang 2011). This strategy is substantially cheaper and faster than carrying out Sanger sequencing, and single nucleotide variants (SNVs) can be readily identified, although validation of findings through other methods, and repeated NGS experiments is mostly required (Metzker 2010). Genome- guided methods are then preferable for instances where a high quality reference genome is available. An interesting advantage of this approach is that, because it reduces a large assembly problem into many smaller by-locus assemblies, these can be easily solved through parallel computing in machines with only few gigabytes of RAM. Besides, contamination or artifacts are not a major concern, as these will not be mapped into the genome, and therefore they will not appear into the output assembly. Furthermore, as this method can incorporate low abundance transcripts, and gaps caused by lack of coverage can be filled by the reference genome, thus genome- guided approaches tend to generate longer untranslated regions (UTRs), allowing the discovery of novel transcripts that might not be included in the current annotation. However, genome- guided methods are not flawless, and their success depends on the quality of the reference genome used (Garber et al. 2011; Martin and Wang 2011). For example, genome- guided methods are capable of placing reads within repetitive regions (which can be solved using mate-pair sequencing), or placing reads in regions that may not exist in the reference sequence (which might result in sequence gaps caused by structural variants; Metzker 2010). In addition, these approaches are obviously not possible for species lacking a reference genome, and also they usually miss spliced reads spanning large introns, and trans- spliced genes (Garber et al. 2011; Martin and Wang 2011).

Conversely, the *de novo* methods are genome- independent, performing the assembly directly from the overlapping reads. For organisms lacking a finished or high- quality genome, *de novo* assemblies can provide an initial set of transcripts, thus allowing differential expression studies; and even when a reference genome is available, *de novo* assemblies can recover transcripts from segments missing in the genome assembly. De novo methods do not depend on alignments to known or predicted splice sites, and thus long introns or trans-spliced genes are not a concern. However, the resources needed to assemble large transcriptomes by this method can be overwhelming (they require complex computational facilities, and a much larger sequencing depth than reference- based strategies). Besides, de novo assemblies are very sensitive to sequencing errors, contaminants, chimeric molecules, and other sequencing artifacts. In spite of all these problems, de novo assembly of bacterial and lower eukaryotic transcriptomes is straightforward, and has led to important discoveries in recent years (Garber et al. 2011; Martin and Wang 2011). For example, *de novo* assemblies have been reported at the level of bacterial genomes, mammalian bacterial artificial chromosomes (BACs), and lower eukaryotic transcriptomes, although considerable challenges exist for their application to large plant and animal transcriptomes (Metzker 2010).

In addition, there is the possibility of combining both *ab initio* and *de novo* strategies, and in this way, one can take the advantage of the high sensitivity of the reference- based approaches, together with the ability of capturing novel transcripts and trans- spliced genes brought by the de novo assemblers. This combined strategy can be carried out either way: first by aligning the reads to the genome, or by de novo assembly of the reads –the choice of one or another depending on several factors. In case a high quality reference genome is available, the combined approach should start by aligning the reads against the genome, followed by the de novo assembly of reads that could not be mapped to the genome. In this way it is possible to filter quickly all unwanted sequences, contaminants and artifacts before the assembly. On the other hand, when the quality of the reference genome is called into question, an assemble-then-align should be carried out, first by de novo assembly of reads, and then to extend the contigs by alignment against the genome. The two obvious advantages here are that errors in the genome are not passed into the assembled transcripts, and also that gaps between fragments can be easily filled by the reference sequence. In this case it is also possible to use protein sequences for scaffolding, if the similarity at the RNA sequence is not enough for sequence extension (Martin and Wang 2011).

The accuracy of an assembly is difficult to measure, and therefore before carrying out functional assignments in transcriptomes, an assessment of the readiness of the assembly for annotation and differential expression analyses is needed (Miller et al. 2010; Yandell and Ence 2012). To this end, there are commonly used statistics that help to describe the completeness and contiguity of an assembly, as well as to evaluate its accuracy, *e.g.* the assembly size, the sequencing coverage, the contig N50, gaps percentage in the assembly, *etc.* (Yandell and Ence 2012). Nevertheless, these metrics require a previously existing transcriptome for comparison, and as standard criteria for assessing the quality of assemblies have not been established yet, the use and interpretation of these rules varies among analysts (Martin and Wang 2011).

Expression. During the assembly of fragments obtained from next generation sequencing experiments ("reads"), one of the processes involved the alignment or mapping of these reads against a reference genome or transcriptome, in order to reconstruct the full transcript sequence. In RNA-seq, the count of mapped sequencing reads for each gene in a given condition is used to measure the gene expression (Guigó 2013). This estimate requires normalization in order to obtain significant results (Garber et al. 2011). A common metric to this end is the number of reads per kilobase of transcript per million of mapped reads (RPKM), which normalizes the read count of a transcript against its length and the total of mapped reads in the RNA-seq sample (Mortazavi et al. 2008). Upon normalization, the difference of expression levels across conditions is analyzed. In this respect, as most RNA-seq studies deal with little or no sample replicates, the current methods to model biological variation and to provide significance in differential expression use various different parametric approaches, for example negative binomial distributions. However, these results must be interpreted carefully, because as with any biological measurement, replicates provide the only way to observe intrinsic variability. The assessment of differential gene expression is usually assisted by annotation results, e.g. by clustering genes with related functions and similar normalized read count patterns (Garber et al. 2011).

Annotation. This is a term that entails two types of processes, one being the identification of all genes and their exon-intron structures (the *structural* annotation), and the other is the assignment of metadata to structural annotations, such as gene ontology terms (*functional* annotation). Annotation can be done manually, but it is so laborious that, although it results in high-quality annotation sets, for reasons of budget generally projects are increasingly relying on automated procedures. Annotation of gene structures at the whole-genome or transcriptome level is generally divided in two different phases: (*i*) A computation phase, where cDNAs, ESTs, proteins and other coding sequences, are aligned to the genome or transcriptome, and *ab initio* or evidence based predictions are produced; and (*ii*) an annotation phase, where all the computed data are merged into gene descriptions. Because these processes are very complex and involve the use of many different tools, the programs that utilize the computed data to create annotations are typically referred to as annotation pipelines. Although there is no standard way to annotate genomic data, the used pipelines share some common features, such as the use of experimental sequence evidences in order to improve the accuracy of the predicted models (Yandell and Ence 2012).

The first step in the computation phase is the identification of repeats, which comprise two types of sequences: (*i*) low complexity sequences, such as homopolymeric regions; and (*ii*) mobile elements, including viruses and transposons. Eukaryotic genomes can be very rich in repeats, and these complicate annotation because their borders are not clearly defined and they are poorly conserved. Therefore, repeat identification requires the creation of species- specific repeat libraries, which are then used as probes for similarity search tools. Upon detection, the stretches containing repeats need to be 'masked,' i.e. marked as repeats, in order to avoid producing false evidences for annotation. After this, annotation pipelines align coding evidences (protein, EST and RNA-seq data) to the assembly, in two steps: First, approximate regions of similarity are defined, and these alignments are usually filtered for marginal matches; and second, the remaining data is clustered, to reveal overlapping segments, grouping different results into a single cluster, and removing redundant evidences. Then, *ab initio* gene prediction tools are employed. These tools use mathematical models rather than evidences (EST and protein data) to identify genes and gene structures. The advantage of this process is therefore that it does not need external evidences; however, in practice these tools are not sensitive or specific enough, and they necessitate previous training. To improve the accuracy of *ab initio* predictions, many tools use external evidences, in processes referred as evidence- based predictions. These tools are in practice difficult to use, being at the present time one of the main bottlenecks in annotation (Yandell and Ence 2012).

For the second step, the annotation phase, the simplest form to proceed is to combine the results from different gene predictions, choosing a single prediction that represents best the consensus of models among overlapping predictions. Another common approach is to supply alignment evidences to the predictors during the identification of the coding sequence, in order to improve accuracy, and then the predictions can be combined as mentioned before. In any case, the decision of using either approach will depend mainly in the amount and type of evidences available, the phylogenetic status of the organism studied (*i.e.*, if there are related organisms with well annotated genomes), and difficulties inherent to projects such as the required effort versus the desired accuracy of the goals, *etc*. (Yandell and Ence 2012).

The Transcriptome of *Physarum*

A first approach to study the transcriptome of *Physarum* was carried out by Glöckner et al. (2008). To this end, they used the plasmodia competent for sporulation, as cells in this state still have the potential to follow different routes of growth and differentiation, and therefore they could find more transcripts that are relevant to commitment than using either the vegetative of committed states. In this way, 15,680 complementary DNAs (cDNAs) were sequenced, and assembled into 5,856 contiguous fragments (contigs). These contigs represented single genes as their cDNA library was generated from the 5' end, and it was enriched for long fragments. They estimated that this sequence databank was roughly 30% of a tentative complete transcriptome of twenty thousand protein-coding genes, and the remaining transcripts were expected to be cell stage-specific, or expressed in very low amounts. From the 5,856 transcripts, 3,282 had orthologs on the TrEMBL database (Boeckmann et al. 2003), and 490 had no similarity to any entry from sequence databases, although they contained InterPro domains (Hunter et al. 2009). The main metabolic and housekeeping genes were also found in this cell stage. However, as they used a state where the cell is waiting for an external input, in this study they paid special attention to signal transduction pathways, and specifically to receptors. In this regard, they encountered 27 cDNAs with receptor domains from photolyases, a family of light detection proteins involved in light- induced signaling processes in other species (e.g., UV damage repair pathways); they postulated that these sensing proteins could participate in the activation during the early stages of sporulation. Glöckner et al. also found 529 potential alternatively spliced transcripts, after the observation of alignment gaps during clustering, and a close examination of these sites showed that the splice site consensus did not differ from the canonical splicing motif GT-AG. This phenomenon has been also observed in *Dictyostelium*, but not in sequences from animals or plants (Glöckner et al. 2008).

Comparative Transcriptomics of Physarum and Other Amoebozoa

Over 40% of the transcripts identified by Glöckner et al. (2008) were similar to proteins from *Dictyostelium discoideum*. However, 895 of their cDNAs (15.28%) entailed orthologs to sequences other than those from *Dictyostelium*. They attributed this difference to genes lost during the divergence of *Physarum* and *Dictyostelium*, and thus evolutionary speciation might account for these differences. The transcripts without counterparts in *Dictyostelium* were not involved in primary metabolism, so either these genes might have been lost in *Dictyostelium*, or these transcripts might have evolved faster in *Dictyostelium* than in other species, although they did not discard the possibility of false positive matches to other datasets (Glöckner et al. 2008).

In a similar manner, Watkins and Gray (2008), performed a comparative study of EST libraries of two free-living amoebae (*Acanthamoeba castellanii*, *Hartmannella vermiformis*) and three slime molds (*Physarum polycephalum*, *Hyperamoeba dachnaya* and *Hyperamoeba* sp.). They included in their analysis the genome sequences of *Dictyostelium* and *Entamoeba histolytica*, and the partial genome available for *A. castellanii*, and compare them to the EST data to identify genes that are unique exclusive to the Amoebozoa. In this way, they found a single gene cluster, called *cudA*, as the only strongly evidence of an amoebozoa- specific gene. This gene is key for the slug culmination in *Dictyostelium*, a terminal phase of the differentiation whose outcome is the fruiting body, needed for sporulation. CudA is essential for asexual reproduction, and it is also associated to the stalk cell differentiation, as *cudA* mutants are unable to go through the early phases of this process. It is likely that CudA acts as transcriptional regulator of the differentiation through interaction with DNA binding proteins, as it localizes to the nucleus. In turn, CudA seems to be controlled by a network of proteins involved in the regulation of morphogenesis that detect environmental and endogenous signals. All these findings correspond to the *Dictyostelium* ortholog, as the *cudA* functions in other amoebozoa remain to be studied (Watkins and Gray, 2008).

Regarding the genes that are specific to the slime molds, Watkins and Gray (2008) found a 23 common genes between the *Dictyostelium* and *Physarum* transcript libraries. Five of these were annotated, with three associated to development, and one involved in signal transduction, acting as a G-protein receptor. One of these genes exclusive to the slime molds is the cytosolic regulator of adenylyl cyclases (CRAC), a specific mediator of the response to extracellular cyclic AMP, and whose sequence cluster show no similarity hits to any other eukaryotic taxa. Interestingly, the mycetozoan- specific genes appear to be related to the differentiation, specifically to the sporulation, and this is probably due to the fact that the mechanisms of multicellularity might have evolved independently in slime molds and other eukaryotes (Watkins and Gray, 2008).

The *Physarum* EST library contained three core meiotic genes, one of them (Rad51) also observed in the *Acanthamoeba* library. This gene is crucial for crossing over during meiosis, and also possesses a key role in the double- strand break repair. The other two meiosis- related genes found in the *Physarum* library were the less studied Rad50 and Dmc1 (Watkins and Gray, 2008). In addition, Watkins and Gray (2008) identified orthologs encoding enzymes from the biosynthetic pathways of trehalose and mannitol. These pathways have been associated to the stress tolerance and adaptation in plants and fungi, and here they observed a considerable range of enzyme diversity within the Amoebozoa.

The plasmodial cDNA libraries studied by Watkins and Gray (2008) also exhibited considerable lateral gene transfer (LGT), comprising 25 EST clusters in *Physarum* that were not found in *Dictyostelium* candidates, although they did not discard the possibility of finding them in other amoebas. One of these clusters was depicted before by Benard et al. (1992) as a late replicating gene without further functional characterization, while the remaining ones are related to the secreted subtilisin-like serine proteases from beta- proteobacteria. These subtilisin-like protease clusters seemed to be fixed and also experienced further duplications and diversification. In addition, some LGT genes annotated as enzymes from the alternate trehalose synthesis pathway, appear to be shared between *Hartmannella vermiformis, H.dachnaya* and *Physarum* (Watkins and Gray, 2008).

Objectives and Approach

The thesis work presented was motivated by the availability of the first transcriptomes (Glöckner et al. 2008; Watkins and Gray 2008) and the first draft of the genome of the slime mold (The Genome Institute, Washington University School of Medicine), as well as the development of the high- throughput, next sequencing technologies, RNA-seq (Nagalakshmi et al. 2008; Wang et al. 2009) in particular, which would allow us to reveal all genes linked to the differentiation of the slime mold. Thus the general aim of this thesis was to identify all transcripts associated to the sporulation, a simple form of cell differentiation present in *Physarum*. More specifically I aimed to: (*i*) compare the transcriptomic changes during the sporulation of plasmodial cell pools, using RNA-seq; (*ii*) develop a single-cell approach to study these whole transcriptome differences under the same experimental conditions; and (*iii*) reveal all theoretically encoded transcripts in the genome, that could be linked to the cell differentiation of the slime mold.

Thesis Organization

In Chapter 2, I will describe the materials and methods employed during the course of this thesis work. Then, in Chapter 3, I will introduce the analysis of the whole transcriptome of the slime mold during sporulation. Here it was found that the most up- and downregulated transcripts could be associated through a protein interaction network involving actin-binding activities. This was achieved via RNA-seq, using the 454 sequencing platform, and comparing cell pool samples before and after the induction for sporulation (exposure to red light). In Chapter 4, I will detail the development of a novel approach, in order to study the transcriptomic changes during the sporulation at the single-cell level. To this end, similar growth and induction conditions were applied to those employed in the previous chapter, and performed the RNA-seq experiments using Illumina sequencing, to obtain the largest coverage possible. In this case, a similar network of actin-binding proteins was observed, thus confirming our results in cell pools, but also as supporting the feasibility of using *Physarum* as a model for transcriptomic studies at the single-cell level. Finally, in Chapter 5, I analyze the latest unpublished version of the genome of the slime mold, in order to identify all possible coding sequences, using a combination of experimental evidences (ESTs, RNA-seq, proteins), and computational predictions. In this manner, I was able to extend the sporulation network, displaying in detail the interaction groups associated to the differentiation, development, and signal transduction in the slime mold. This study not only provides sets of putative candidates that could be used in future experimental studies in the genetic nature of the sporulation in the slime mold, but also a pipeline to annotate the genome of *Physarum*, including both the coding and noncoding transcriptomes.

Chapter 2. Methods.

I. Materials.

Three different strains were employed in this study: WT31 (Chapters 4 and 5; Glöckner et al. 2008), LU352 (Chapter 5; Dee et al. 1989), and the cross LU897 x LU898 (Chapters 3 and 5; Sujatha et al. 2005). These are described in detail the Table 3. Computer equipment, data sources, and data generated during this thesis work, are listed in tables 4 to 7. Finally, the programs used are included in tables 8 to 11.

Table 3. Strains used in this work. The LU strain prefix stands for Leicester University (Anderson and Dee 1990). The strain LU352 was provided by Gerard Pierron (Villejuif, France) to the Washington University Sequencing Center (St.Louis, Missouri) for the genome project. The LU897 x LU898 cross derivative and the WT31 strain were developed and are available locally.

Strain Name	**Origin**	**Genotype**	**Phenotype**	**Apogamic Growth**	**Sporulation**	**Reference**
LU352	Cld-AXE x LU213 cross	*matA2, gadAh, npfC, matB3, fusA1, axe, whiA*$^+$	yellow	yes	yes	Dee et al. 1989
LU897 x LU898	LU897 x LU898 cross	*matA1, matA2, fusA1, fusA2, fusB1, fusC1, whiA1*	white	no	yes	Glöckner et al. 2008
WT31	LU352 x LU897 cross	*matA2, fusA1, gadAh, npfC*$^+$*, whiA*$^+$	yellow	yes	yes	Sujatha et al. 2005

Table 4. Hardware.

Platform	**IBM x3755 M3**	**Mac Pro 4,1**	**MacBook Pro 6,2**	**PC**
Processor(s)	Four 12-Core AMD Opteron 6172 (2.1 GHz)	Two Quad-Core Intel Xeon	One Intel Core i7 (2.66 GHz)	One Intel Core-2 Duo E6850 (3 GHz)
RAM Memory	256 Gb (*see note below*)	17 Gb, 1066 MHz DDR3	4 Gb, 1067 MHz DDR3	3.2 Gb
Storage	Eight 2-Tb SAS NearLine	639.79 Gb SATA Western Digital	One 499,76 Gb SATA Seagate	Two 250 Gb SATA Seagate
Operating System	SUSE Linux Enterprise Server 11 (x86_64)	Mac OS X version 10.6.8	Mac OS X version 10.6.8	Ubuntu 10.04; Windows XP SP3
Purpose	Data Analysis; Data Storage; Scripting	Data Analysis; Scripting; Statistics	Data Analysis; Scripting; Statistics	Statistics; Scripting

Note: Some bioinformatics applications, *e.g. de novo* assembly of large genomes and transcriptomes, require large amounts of RAM memory. In general, an approximation formula was derived by Simon Gladman (CSIRO, Australia) to calculate the RAM needed for *de novo* assembly (Gladman 2009):

$$RAM = -109635 + 18977 * ReadSize + 86326 * GenomeSize + 233353 * NumReads - 51092 * Kmer$$

Where the read size is given in base pairs, the genome size in megabases, and the number of reads is in millions; the result can be divided by 1,048,576 to convert it to gigabytes. For example, over 512 Gb of RAM are needed to work with the human genome. Moreover, to carry out the assembly of the *Physarum* genome (~300 Mb), with shorts reads (36 bp), on a mid- coverage resolution (24 million reads) using the velvet assembler (*k*-mers: 31, 41, 51; Zerbino and Birney 2008), the amount of RAM required would be in the range between 28.10 and 29.08 Gb.

Table 5. Nucleotide Databases and Datasets

Database	**Purpose**	**Version**	**Reference**
RepBase	Repeat annotation and masking	20120418	Jurka et al. 2005
Rfam	Noncoding RNA annotation	11.0	Griffiths-Jones et al. 2005
Physarum rRNA	rRNA annotation. Obtained from GenBank via Entrez.	Accessed on 29/01/2013	Benson et al. 2011
Physarum nucleotide sequences	Annotation. Obtained from GenBank via Entrez.	Accessed on 30/01/2013	Benson et al. 2011
Physarum transcriptomic ESTs	EST clustering; Mapping ESTs to Protein Models	N.A.	Glöckner et al. 2008
Physarum transcriptomic ESTs	EST clustering and mapping. Obtained from GenBank	N.A.	Watkins and Gray 2008
Physarum amoeba transcriptome long reads	Long read mapping. Obtained from the European Nucleotide Archive (ENA).	Accession SRP000013	Unpublished
D. discoideum EST sequences	Annotation; Mapping ESTs to Protein Models	Version date 12/19/2008	Chisholm et al. 2006
D. discoideum coding sequences	Annotation; comparative genomics of coding potential	Version date 18/02/2014	Chisholm et al. 2006
D. discoideum genomic scaffolds	Mapping CEGMA against *Dictyostelium* genomes	Accessed on 26/02/2013	Chisholm et al. 2006
D. purpureum genomic scaffolds	Mapping CEGMA against *Dictyostelium* genomes	Accessed on 26/02/2013	Sucgang et al. 2011

Table 6. Protein Databases and Datasets

Database	**Purpose**	**Version**	**Reference**
Physarum amino acid sequences	Annotation. Obtained from GenBank via Entrez.	Accessed on 30/01/2013	Benson et al. 2011
UniProt	Protein- coding gene annotation	Version date 11/2012	The UniProt Consortium 2010
CEGMA (core eukaryotic genes)	Assessment of Completeness; SimiTri Analysis; Mapping CEGMA against Dicty and Physarum genomes; comparison against protein models	N.A.	Parra et al. 2007; Parra et al. 2009
OrthoMCL-DB	Mapping Protein Models to Ortholog Clusters	Version 5 (31/03/2011)	Li et al. 2003
KEGG GENES	Mapping Protein Models to KEGG Orthologs; Mapping Dicty Proteins to KEGG Orthologs	Release 66, 04/2013	Kanehisa et al. 2010
Monosiga brevicollis Protein sequences	Comparative genomics via SimiTri Analysis. Obtained from the DOE Joint Genome Institute.	Accessed on 05/08/2009	King et al. 2008
Saccharomyces cerevisiae Protein sequences	Comparative genomics via SimiTri Analysis. Downloaded from Saccharomyces Genome Database (SGD).	Accessed on 05/08/2009	Issel-Tarver et al. 2002
D. purpureum Protein sequences	Comparative genomics via SimiTri Analysis. Obtained from the DOE Joint Genome Institute.	Version date 02/04/2010	Sucgang et al. 2011
D. discoideum Protein sequences	SimiTri Analysis; Mapping CEGMA proteins against *Dictyostelium* genomes. Obtained from DictyBase.	Version date 09/12/2011	Chisholm et al. 2006

Table 7. RNA-seq datasets. Each accession corresponds to a different sequencing run or experiment. All sets were generated during the course of this study, except for the data from the LU352 amoebae, which was sequenced at the Washington University Sequencing Center (St.Louis, Missouri), from an RNA sample preparated by Gerard Pierron (Patrick Minx, personal communication).

Strain	Source	Method	Reads	Database	Accession	Reference
LU352	amoebae	Roche 454	8,064,625	SRA	SRP000013	Unpublished
LU897 x LU898	plasmodia	Roche 454	405,363	SRA	SRP001397	This work
WT31	plasmodia	Illumina	77,023,388	ENA	ERP001220	This work
LU897 x LU898	plasmodia	Illumina	15,844,226	SRA	SRP009381	This work
LU897 x LU898	plasmodia	Illumina	98,803,609	N.A.	Unsubmitted	This work

Table 8. Programs used for assembly and mapping RNA-seq data

Program	Purpose	Version	Reference
cap3	Sequence Assembly	12/21/07	Huang and Madan 1999
BLAST+	Sequence clustering	2.2.27+	Camacho et al. 2009
Bowtie	Short reads mapping	0.12.7	Langmead et al. 2009
TopHat	Mapping short sequences from spliced transcripts to the genome	1.4.0	Trapnell et al. 2009
Cufflinks	Transcriptome assembly	1.3.0	Trapnell et al. 2012
samtools	Short reads manipulation	0.1.18 (r982:295)	Li et al. 2009
USEARCH	Sequence clustering	5.2.32	Edgar 2010

Table 9. Programs used for identification of repetitive sequences and non-coding RNA annotation. All programs are Linux versions, executed under SuSE, except for CPC, whose source was modified to run on OSX 10.6.8

Program	**Purpose**	**Version**	**Reference**
RepeatMasker	Repeat annotation; repeat masking	open-4.0.0	Smit et al. 2010
RepeatModeler	Repeat annotation, masking, and modeling of repetitive sequences	open-1-0-7	Smit and Hubley 2010
TRF (Tandem Repeats Finder)	Repeat annotation, masking, and modeling of repetitive sequences	4.07b	Benson 1999
RECON	Modeling of repetitive sequences	1.07	Bao and Eddy 2002
RepeatScout	Modeling of repetitive sequences	1.0.5	Price et al. 2005
RMBLASTN	Nucleotide-nucleotide BLAST with RepeatMasker extensions	2.2.27+	Smit et al. (2010)
BLAST+	rRNA annotation	2.2.27+	Camacho et al. 2009
tRNAscan-SE	tRNA annotation	1.23	Lowe and Eddy 1997
RNAmmer	rRNA mapping	1.2	Lagesen et al. 2007
Infernal	Noncoding RNA annotation	1.0.2	Nawrocki et al. 2009
bedtools	Noncoding RNA annotation (ncRNA GFF statistics)	2.17.0	Quinlan and Hall 2010
CPC (coding potential calculator)	Noncoding RNA annotation (Protein Coding Potential)	0.9-r2	Kong et al. 2007

Table 10. Programs used in the annotation and the comparative analysis of the coding transcriptome.

Program	**Purpose**	**Version**	**Reference**
InterProScan	Protein signature (domains) recognition	4.8	Zdobnov and Apweiler 2001
Blast2GO	Annotation	2.5	Conesa et al. 2005
OrthoMCL	Mapping Protein Models to Ortholog Clusters. Webserver accessed 17/05/2013	2.0;	Chen et al. 2007
KAAS	Mapping Protein Models to KEGG Orthologs. Webserver accessed 17/05/2013	1.67x	Moriya et al. 2007
GOSlimViewer	summary of GO annotation	Webserver v.2.00	McCarthy et al. 2006
iPath	Comparative Metabolic Pathways of Physarum and Dictys	Webserver version 2.0	Yamada et al. 2011
Blast	Simitri analysis	2.2.26 (OSX)	Altschul et al. 1997
SimiTri	Simitri analysis		Peregrín-Alvarez and Parkinson 2009
SOBAcl	command line tool for analyzing GFF3 annotations		Eilbeck et al. 2005
CodonW	Codon usage	1.4.4 compiled under Ubuntu 10.04.4	Peden, John; codonw.sourceforge.net

Table 11. Programs used for the identification of gene models. All programs are Linux versions, executed under SuSE, except for CPC, whose source was modified to run on OSX 10.6.8

Program	**Purpose**	**Version**	**Reference**
blat	Sequence alignment	35x1	Kent 2002
Genemark-ES	Gene prediction	2.3e	Borodovsky and Lomsadze 2011
MAKER2	Protein and Transcript Model Identification	2.1 (OSX)	Holt and Yandell 2011
BLAST+	Gene prediction (under MAKER2)	2.2.27+	Camacho et al. 2009
Exonerate	Gene prediction (under MAKER2)	2.2.0 (OSX)	Slater and Birney 2005
Augustus	Gene prediction (under MAKER2)	2.5.5 (source modified for OSX)	Stanke et al. 2008
SNAP	Gene prediction (under MAKER2)	2006-07-28 (OSX)	Korf 2004
Eval	Comparison of predicted gene models	v.2.2.8	Keibler and Brent 2003

Analysis of the expressed transcriptome during the differentiation of *Physarum* cell pools

Culture and light-induction of plasmodial cells. *Physarum* plasmodia of the white strain (LU897 × LU898 cross) were hatched from spherules, and grown as microplasmodial suspensions for four days. The plasmodial mass was then applied to starvation agar plates. Microplasmodia spontaneously fused to give a single plasmodium on each plate. Plasmodia were then starved for six days in the dark at 22°C to obtain maximal competence for sporulation. To verify the sporulation-competent state, plasmodia were cut into two halves. One half was immediately frozen in liquid nitrogen for RNA extraction, and the other half was returned to the dark and incubated until the next day to verify that the plasmodium had not been induced to sporulation. To obtain light-induced plasmodia, competent plasmodia were irradiated for 30 min with far red light and then returned to the dark. Six hours after the start of irradiation, plasmodia were cut into two halves. One half was frozen in liquid nitrogen for RNA extraction. The other half was returned to the dark and incubated until the next day to verify the sporulation status (Roland Kroneder et al. 1999; Golderer, Werner, Leitner, Grobner, et al. 2001).

cDNA Library Construction and Sequencing. Transcript poly(A)+ RNAs were isolated by oligo-dT chromatography. cDNAs were prepared from these RNAs by the full-length enriched synthesis method (vertis Biotechnologie, Freising-Weihenstephan, Germany). First strand cDNA was synthesized using oligo(dT) adapter primers and MMLV H-reverse transcriptase. Following RNA hydrolysis, an adapter primer was annealed to the 3' end, and the produced fragments were PCR-amplified for 22 cycles with a proofreading enzyme. The cDNA libraries were then directly sequenced using the 454 GS FLX system (Roche Diagnostics, Mannheim, Germany; Margulies et al. 2005). Chromatograms were scored for quality, and the produced sequences were trimmed of adapter sequences, and coassembled into contigs using previously available transcriptomic data

(Glöckner et al. 2008). For expression comparisons I obtained for each contig: (i) the number of reads (defined here as "hit counts") in both libraries; and (ii) their relative frequencies (reads of a given contig divided by the total number of reads). Statistical significance between the two hit counts for each contig species was then assessed (Audic and Claverie 1997).

Sequence Annotation and Network Inference. Similarity searches against protein databases were performed using BLASTX (Altschul et al. 1990; Altschul et al. 1997). I employed nine protein databases in this comparison: Swiss-Prot and TrEMBL, versions 56.3 and 39.3 (Boeckmann et al. 2003), dictyBase (Chisholm et al. 2006) and RefSeq database subsets: mammalian, other vertebrates, invertebrate, protozoa, plant and microbial, release 31 (Pruitt et al. 2007). Functional annotation was carried out using BLAST2GO, version 2.2.3 (Götz et al. 2008). This procedure consisted of a similarity search against the non-redundant GenBank database (Benson et al. 2008a), using BLASTX (e-value 1E-3), followed by Gene Ontology (GO) (Gene Ontology Consortium 2008) mappings extracted from similarity results and InterPro domain matches (InterPro release 18.0) (Hunter, Apweiler, Attwood, Bairoch, Bateman, Binns, Bork, Das, Daugherty, Duquenne, Finn, Gough, Haft, Hulo, Kahn, Kelly, Laugraud, Letunic, Lonsdale, Lopez, Madera, Maslen, McAnulla, McDowall, Mistry, Mitchell, Mulder, Natale, Orengo, Quinn, Selengut, C. J. a Sigrist, et al. 2009). Annotation of sequences (cutoff value 1E-6) was followed by their validation, and these annotations were extended using ANNEX (Myhre et al. 2006). Statistical analysis of GO annotations between differentially expressed cDNAs was carried out using the Fisher exact test, as implemented in the GOSSIP module (Bluthgen et al. 2005) of BLAST2GO. Sequences were also categorized in metabolic and signaling pathways, via similarity search against orthologs present in the KEGG database using the KAAS server (Kanehisa et al. 2008; Moriya et al. 2007). In this case, I employed default parameters for ESTs. KEGG orthologs (KOs) were then plotted into the whole metabolic atlas, utilizing the KEGG mapping tool (Okuda et al. 2008a). Putative networks of correlated genetic interactions were generated from annotation information, using the MLE algorithm (Deng et al. 2002), as implemented in the Cytoprophet plugin of the program Cytoscape (Morcos et al. 2008; Shannon et al. 2003). For a summary of these procedures, see Figure 7.

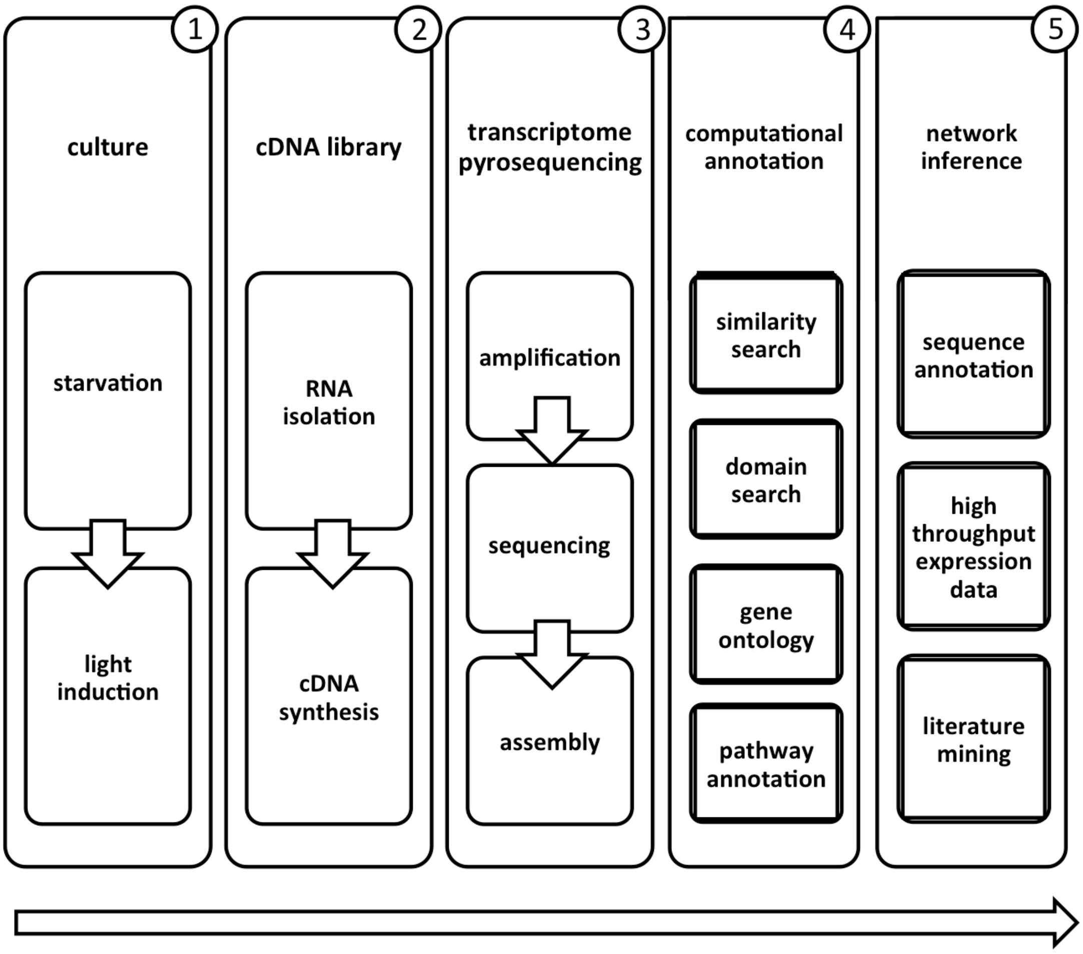


Figure 7. Overview of the experimental design for the analysis of the transcriptome in cell pools. A summary of experiments and computational analyses is depicted for the analysis of the transcriptome during the differentiation of cell pools. RNA samples were taken from competent plasmodia after six days of starvation in the dark, and from competent plasmodia at six hours after exposition to a 30 minutes pulse of red light (≥ 700 nm) (1). cDNAs were synthesized from extracted RNAs (2), and sequenced and quantitated using the 454 Life Sciences platform (3). Contigs generated were then annotated at every bioinformatic level (4), and network interactions (5) were obtained both by a combination of manual curation of literature, expression data, and predictions from annotations

Analysis of the expressed transcriptome during the differentiation of *Physarum* single cells

Culture and sequencing. *Physarum* macroplasmodia, apogamic strain WT31 (Starostzik and Marwan 1998), were cultured as previously described in the previous chapter. Cells were grown and collected under two different conditions: (*i*) a plasmodium starved for 6 days (competent D1 and D2 individual cell samples); and (*ii*) a plasmodium starved for 6 days, exposed to far red light for 30 minutes, and returned to the dark for 6.5 hours (photoinduced L1 and L2 cells; Table 1). During this time period the cell becomes irreversibly committed to sporulation (Starostzik and Marwan 1995). Samples were frozen and PolyA+ RNA was isolated from the total RNA samples (by two rounds of oligo-dT affinity chromatography), and fragmented with ultrasound (4 pulses of 30 sec at 4°C). Subsequently, the RNA fragments were poly(A)-tailed using poly(A) polymerase, followed by treatment with tobacco acid pyrophosphatase (TAP). Then a RNA adapter was ligated to the 5'-monophosphate of the RNA. First-strand cDNA synthesis was performed using an oligo(dT)-adapter primer and the M-MLV reverse transcriptase. The resulting cDNAs were PCR-amplified for 14-15 cycles to about 20-30 ng/µl, including distinctive 4-bp 5'-barcodes for each sample (Table 1), and using a high fidelity DNA polymerase. The PCR products were purified with the Agencourt AMPure XP kit (Beckman Coulter Genomics), and pooled in equivalent amounts. cDNAs in the range of 200 to 400 bp were fractionated from agarose gels and sequenced using the Illumina HiSeq 2000 system.

Data Analysis. The 100-bp sequencing outputs were then trimmed for quality (Phred score > 33), and later assembled de novo, using velvet (Zerbino and Birney 2008) and oases (Schulz et al. 2012). *k*-mers of 31, 41, and 51 nucleotides long were used for these assemblies. Later, CAP3 (Huang and Madan 1999) was employed to reduce redundancy in the assembly. The annotation of this assembly was carried out first through BLAST (Altschul, Madden, A. A. Schäffer, et al. 1997) searches (*e*-value 1E-3) against the SwissProt (Boeckmann et al. 2003) protein database. A search for *Physarum* noncoding RNAs was not included due to the lack of complete gene models and a finished reference genome in this species. Afterwards, domains and protein signature patterns were associated from matches to the InterPro database, and Gene Ontology (GO) based annotations were assigned using Blast2GO (Götz et al. 2008), from annotations pertaining to orthologs (annotation *e*-value cutoff < 1E-6). Gene names and descriptions were filtered using the Blast Description Annotation tool from Blast2GO (Götz et al. 2008). Significant differences in GO annotations between sets of up- and downregulated genes from each cDNA library were evaluated using Fisher exact tests, as implemented in Blast2GO.

To assess the differential expression between the several single cells, the sequencing output was splitted using the barcode information for each sample. Then the decoded outputs were mapped to the novel assembly with Bowtie (Langmead, Trapnell, Pop & S. Salzberg 2009). Samtools (Li et al. 2009) and Tablet (Milne et al. 2010) were later used to obtain mapped read counts. For expression comparisons it was obtained for each transcript: (*i*) the number of mapped reads; and (*ii*) the normalized expression value, as measured in reads per kilobase per million mapped reads, RPKM (Mortazavi et al. 2008). To identify differentially expressed transcripts between starved and photoinduced cells, the non-normalized mapped read count data was analyzed using the R-based package DESeq (Anders and Huber 2010). Transcript abundances for each gene were estimated as a weighted mean of mapped read counts from each replicate, normalised to the library size. *P*-values (adjusted for false discovery rate) were generated for each gene in pairwise comparisons between different conditions (competent and induced cells). I used the per-condition method and fit-only sharing mode. A summary of experiments and bioinformatic analyses is depicted in Figure 8.

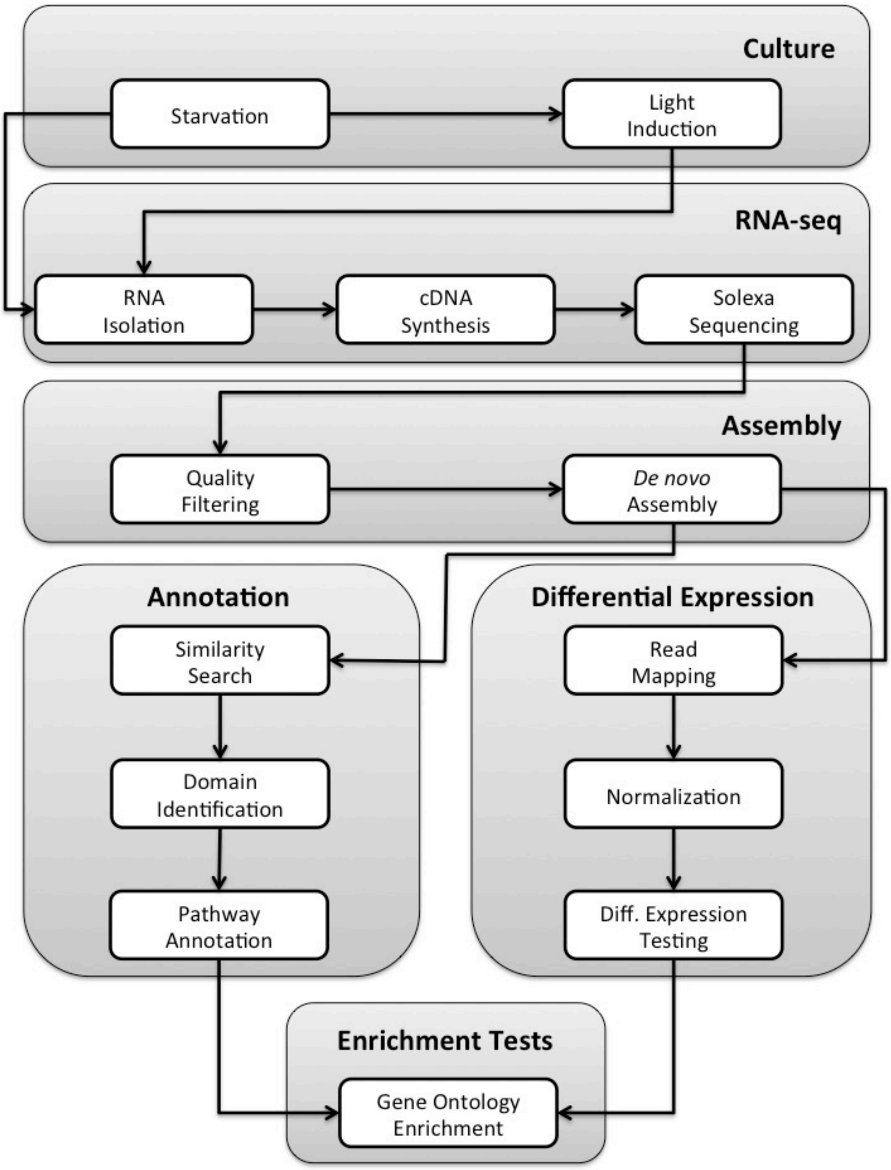


Figure 8. Overview of the experimental design for the analysis of the transcriptome in single cells. A summary of experiments and computational analyses is shown. RNA samples were taken from competent and light-induced plasmodia (*culture*). These RNAs were employed as templates for cDNA synthesis, which were later sequenced using the Illumina HiSeq 2000 platform (*RNA-seq*). Reads were assembled (*assembly*), and the obtained contigs were annotated at every bioinformatic level (*annotation*). Then, to evaluate differentially expressed transcripts, reads were mapped to the assembly and normalized (*differential expression*). Finally, gene ontology annotations were tested for enrichment between up- and downregulated transcripts (*enrichment tests*).

Identification and Annotation of the Reference Transcriptome of *Physarum polycephalum*

RNA Sequencing of the White Strain (First Batch). In order to obtain the maximum number of expressed transcripts, RNA-seq was carried out in two batches, from samples of *Physarum* plasmodia of the white strain (LU897 x LU898 cross; Table 3). For the first white strain sequencing batch (here named "LULU1"), macroplasmodial single-cells were grown and collected under three different conditions: (*i*) a plasmodium starved for 6 days (competent D cell sample); (*ii*) a plasmodium starved for 6 days, exposed to far red light for 30 minutes, and returned to the dark for 2 hours (L2 photoinduced cell); and (*iii*) a plasmodium starved for 6 days, exposed to far red light for 30 minutes, and returned to the dark for 6 hours (L6 photoinduced cell). Upon collection, the three samples were ground under liquid nitrogen. Total RNA was then isolated from the frozen samples using the mirVana miRNA isolation kit (Ambion). The total RNAs were tested for their integrity by capillary electrophoresis. Afterwards, to enrich for mRNA in the total RNA preparation, the RNA samples were incubated with Terminator exonuclease (New England Biolabs), which specifically degrades RNA species which carry a 5' phosphate. The obtained full-length mRNAs were then treated with a tobacco acid pyrophosphatase, to release the 5' CAP structure. This was followed by ligation of a RNA adapter to the 5´-phosphate of the decapped mRNAs. First-strand cDNA synthesis was carried out with a N6 randomized adapter primer and M-MLV-RNase H- reverse transcriptase. The resulting cDNAs were amplified with cycles of LA (long and accurate)-PCR. For Illumina sequencing (Bentley et al. 2008), the cDNAs were pooled in equal amounts and from this pool, the cDNAs in the size range of 200 – 450 bp were eluted from a preparative agarose gel. An aliquot of the size fractionated cDNA was analyzed by capillary electrophoresis. The output was encoded in Illumina Phred-64 FASTQ format (Cock et al., 2010), and deposited in the NCBI Sequence Read Archive (accession SRP009381; Table 7).

The primers used for PCR amplification were designed for amplicon sequencing according to the instructions of Illumina/Solexa. The following adapter sequences flanked the cDNA inserts (Illumina adapter sequences are underlined):

5’- end (53 bases):
5´- AAT GAT ACG GCG ACC ACC GAC AGG TTC AGA GTT CTA CAG TCC GAC GAT C-NNNN-3’
3’-end (39 bases):
5’- CAA GCA GAA GAC GGC ATA CGA-TCA GGC AGA GGA CGA GAA-3’

RNA Sequencing of the White Strain (Second Batch). For the second RNA-seq batch (“LULU2”), I included single-cell white strain samples from a competent plasmodium, and photoinduced cells collected after 3.5, 8 and 10 hours after far red light exposure. Poly(A)+ RNA was then isolated from total RNA, and fragmented with ultrasound (4 cycles at 4°C for 30 seconds). The RNA samples were then dephosphorylated using antarctic phosphatase and re-phosphorylated with polynucleotide kinase (PNK). Afterwards, the RNA fragments were poly(A)-tailed using a poly(A) polymerase. Then an adapter was ligated to the 5´-phosphate end. First- strand cDNA synthesis was carried out using an oligo(dT)-adapter primer and a Moloney murine leukemia virus reverse transcriptase (M-MLV). The obtained cDNAs were PCR-amplified to about 20-30 ng/µl using a high fidelity DNA polymerase, with primers including the barcoded TruSeq sequencing adapters (Illumina; cycle programs are indicated in the Table). Subsequently, the cDNA samples were pooled in 3 different pools, and then eluted from agarose gels in the size range of 200- 500 bp. Aliquots of the fractionated cDNA were analyzed by capillary electrophoresis. The cDNA pools were then sequenced (single reads) on a Illumina HiSeq 2000 system. The output was encoded in Phred-33 FASTQ (Sanger) format (Cock et al., 2010).

The following adapter sequences flank the DNA inserts (combined length of the flanking sequences is 146 bases):

TrueSeq Sense primer:
5´- AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T-3'
TrueSeq Antisense primer (N6- Barcode):
5'-CAA GCA GAA GAC GGC ATA CGA GAT-N(6)-GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC (dT25)-3'.

For both RNA sequencing batches, the RNA material was provided by Wolfgang Marwan (Otto von Guericke University), and RNA preparation and sequencing as described here were carried out by vertis Biotechnologie (Freising-Weihenstephan, Germany).

Genome sequencing and annotation. The *Physarum* genome assembly, version 7.3.1, was obtained from The Genome Institute, Washington University School of Medicine (St Louis, MO). First, a search for rRNAs, tRNAs and other noncoding RNAs, via a combination of similarity (BLAST+), *ab initio* (RNAmmer, tRNAscan-SE, CPC) and motif finding strategies (Infernal; Table 9) was performed (Camacho et al. 2009; Lagesen et al. 2007; Lowe and Eddy 1997; Kong et al. 2007; Nawrocki et al. 2009). Most of the following analyses were carried out over repeat masked sequences. To this end, a species- specific repeat library was created, using the RepeatModeler package, and then the repeats were identified with a combination of programs: TRF for tandem repeats, RECON and RepeatScout for *ab initio* repeat detection, and RMBLASTN and RepeatMasker for known repeats and transposons present in the RepBase database (see Table 9; Benson 1999; Bao and Eddy 2002; Price et al. 2005). The masked genome obtained was then (*i*) mapped for *Physarum* and *Dictyostelium* ESTs (Table 5; Chisholm et al. 2006; Glöckner et al. 2008; Watkins and Gray 2008); (*ii*) searched for ortholog proteins from the UniProt database (Table 6; The UniProt Consortium 2010); (*iii*) aligned against long (454) and short (Illumina) RNA-seq outputs, obtained in previous experiments (Table 7); and (*iv*) used for gene prediction with the GeneMark ES program (Table 11; Borodovsky and Lomsadze 2011). Then these four types of evidences (ESTs, proteins, RNA-seq reads and predicted genes) were combined over the masked genome into the final protein-coding gene models (Holt and Yandell 2011). The accuracy of these predicted gene models, in terms of sensitivity and specificity, was measured with Eval (Keibler and Brent 2003).

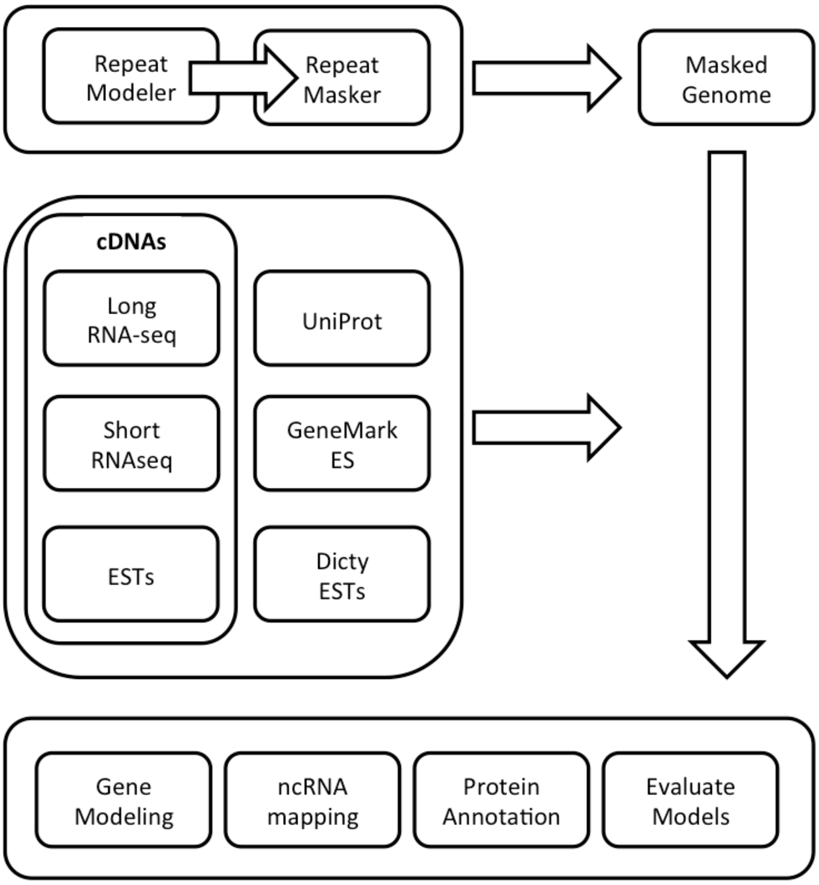


Figure 9. Gene identification and genome annotation pipeline. The genome was first searched for noncoding RNAs and repetitive sequences, and masked for the repeats found. Then known proteins, cDNAs (ESTs), as well as predicted exons were mapped and combined into gene models. Finally, the proteins encoded in these gene models were annotated and compared against closely related proteomes. The coding and noncoding RNA annotations were integrated into a uniform database.

The encoded proteins were then annotated (Altschul et al. 1997; Conesa et al. 2005; Hunter et al. 2009; Zdobnov and Apweiler 2001; Okuda et al. 2008), and their sequences and annotations were used for comparative genomics against the proteomes from *D.discoideum*, *D.purpureum*, *Monosiga* and yeast. Proteins annotated for Gene Ontologies associated to cell differentiation, signal transduction, and embryo development, were selected and joined into interaction networks, with the Cytoscape program. These procedures are shown in Figure 9.

Chapter 3. The expressed transcriptome during the differentiation of *Physarum* cell pools

Background

In order to identify the differentially expressed genes associated with the commitment to sporulation, a characterization and comparison of two cDNA libraries prepared from competent and light-induced plasmodia using massive parallel sequencing of RNAs, or RNA-seq (Margulies et al. 2005; Nagalakshmi et al. 2008), was carried out. This method was employed because it does not rely on reference transcripts for quantitation, previous cloning steps are not required, it does not have an upper limit for quantitation, and it is a relatively unbiased approach (Wang et al. 2009). The comparison of annotations and transcript quantitations show that most differentially expressed genes encode proteins associated to a network of actin-binding proteins. Components of this putative interaction network are associated to development, DNA repair, cell division, calcium release, cell death, and maintenance of cell integrity.

Results

Sequencing and Profiling of cDNAs expressed in competent and light-induced plasmodia

Separate cDNA libraries were constructed from polyA+ RNA isolated from two sources: (*i*) competent plasmodia; and (*ii*) sporulation- induced plasmodia (competent plasmodia harvested six hours after exposure to far-red light). The cDNAs libraries were then analyzed using massive parallel sequencing (Margulies et al. 2005; Wang et al. 2009). Transcripts were annotated at every bioinformatic level, and the annotation data was used to infer hypothetical interaction networks from differentially regulated genes. The whole approach is summarized in the Figure 7.

From the pyrosequencing, a total output of 61.9 Mb from two runs was obtained, corresponding to the starved (26.1 Mb) and light-induced (35.8 Mb) plasmodia libraries. As *Physarum* possess a 300 Mb genome (Glöckner et al. 2008), and assuming that 10% is encoding genes, therefore a 2.06X coverage of protein coding sequences was estimated. The assembled sequencing output consisted of 26,037 sequences, and large cDNAs from this assembly (>500 bp) were then joined to a previously available sequence dataset (Glöckner et al. 2008), to form a comprehensive set of representative transcripts. This analysis produced 16,669 sequences, 13,169 of these containing transcript abundance data: 125,456 reads from competent and 99,632 reads from light-induced plasmodia, respectively. For practical reasons, this entire transcript abundance dataset is not included here, but can be accessed at:
http://www.biomedcentral.com/content/supplementary/1471-2164-11-115-s2.xls

This abundance data (number of reads for each assembled transcript) was then used as a measure of expression, which defined here as "*hit counts*". The remaining contigs without hit counts consisted of previously sequenced clones from a normalized cDNA library prepared from competent plasmodia (Glöckner et al. 2008), indicating that the normalization produced a broader coverage of transcripts. From 11,399 cDNA contigs detected in the competent plasmodia library (10,689 in light-induced plasmodia), over 4,227 were represented with at least five hits (3,553 in light-induced plasmodia; Figure 10). Conversely, 8,711 transcripts (52,3%) were found with 5 or less sequence hits in both samples. For statistical reasons, no statement on the differential expression from this fraction could be made. Between contigs with lowest hit counts, 2,437 cDNA species were represented by just one hit (competent plasmodia), and 2,621 from light-induced sample (Figure 10).

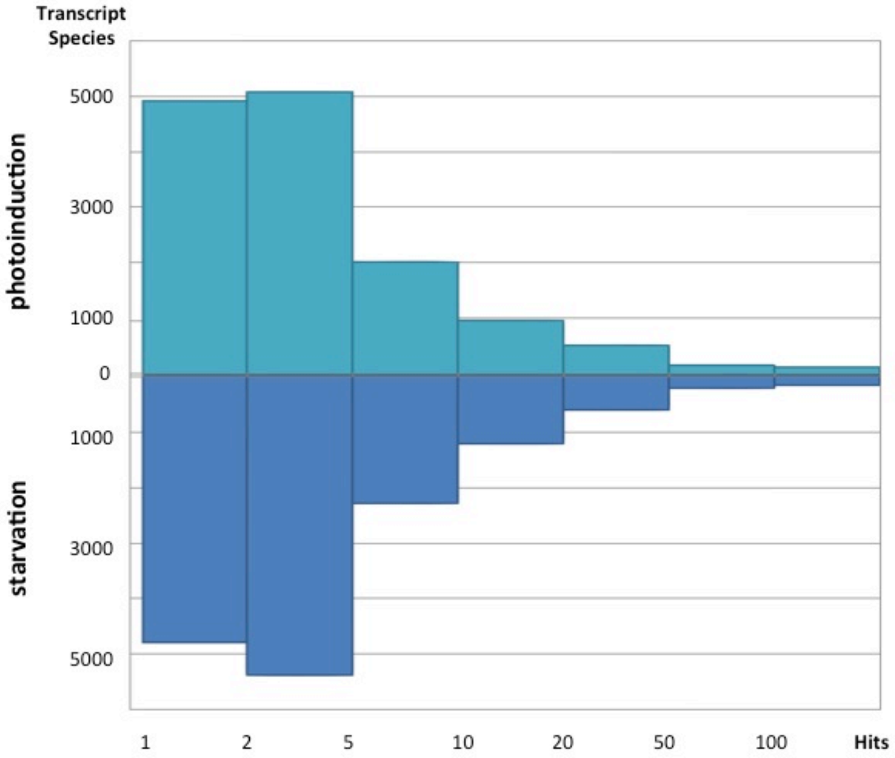


Figure 10. Hits Distribution of Transcript Species. The distribution of pyrosequencing hit counts respect to the number of transcript species on each library (starvation and light-induced) is depicted on a semi-logarithmic scale. Hit counts are included in the adjacent upper ranges to the right; for example, transcripts with 2 hits are present in the 2-5 range. Similar distributions of contig species were found on both libraries, and most transcripts were represented by 1 to 5 hits only.

Then a comparison of the transcript hit counts between different libraries as a measure of differential gene expression was necessary. As most contig species were represented by low hit counts, the number of hits was normalized. To this end, first the relative frequency (number of hits divided by the total hits on a given condition) was obtained, and later the relative frequencies were calculated for each contig in the two cDNA samples compared to each other. Given that each EST belongs to a single gene, the significance of its differential expression depends only on the number of hits, respect to the total number of hits on each library (Audic and Claverie 1997). Following these assumptions, 2,772 cDNAs were found that displayed significant differential expression (*P-value* < 0.05). All contig species, regardless of whether differentially expressed or not were submitted to the Sequence Read Archive subset of GenBank (Benson et al. 2008).

The sequencings were deposited under the accession numbers SRX012830 and SRX012831.

The newly assembled contigs were compared against sequence databases using BLASTX (Altschul et al. 1990; Altschul et al. 1997). This analysis revealed that 3,310 sequences have significant similarity (≤ 1E-15) to existing sequences in SwissProt (Boeckmann et al. 2003), 3,651 to the protozoa subset from RefSeq (Pruitt et al. 2007), and 3,345 to proteins of the related model organism *Dictyostelium discoideum*, present in dictyBase (Chisholm et al. 2006). From the 13,169 sequences with hit counts data, orthologs were identified for 5,544 transcripts (1,287 of these with significant differential expression). The similarity data for the entire transcript set is available at:
http://www.biomedcentral.com/content/supplementary/1471-2164-11-115-s4.xls

Later, in order to identify differentially regulated genes, the contig species were clustered into expression groups according to their relative frequencies in both conditions. As a result, contigs encoding orthologs related to cell division (meiosis-related protein *MEI2*; DNA polymerase beta; actin) and protein synthesis and degradation (elongation factor 1- alpha; cathepsin-L cysteine protease) were found, with higher relative frequencies in the competent plasmodial library. Similarly, orthologs related to the cytoskeleton (spire; actophorin; cell wall integrity and stress response component, *WSC1*) and cell differentiation genes (*CudA*) were found with greater relative frequencies in the light- induction library (Figure 11 and Table 12).

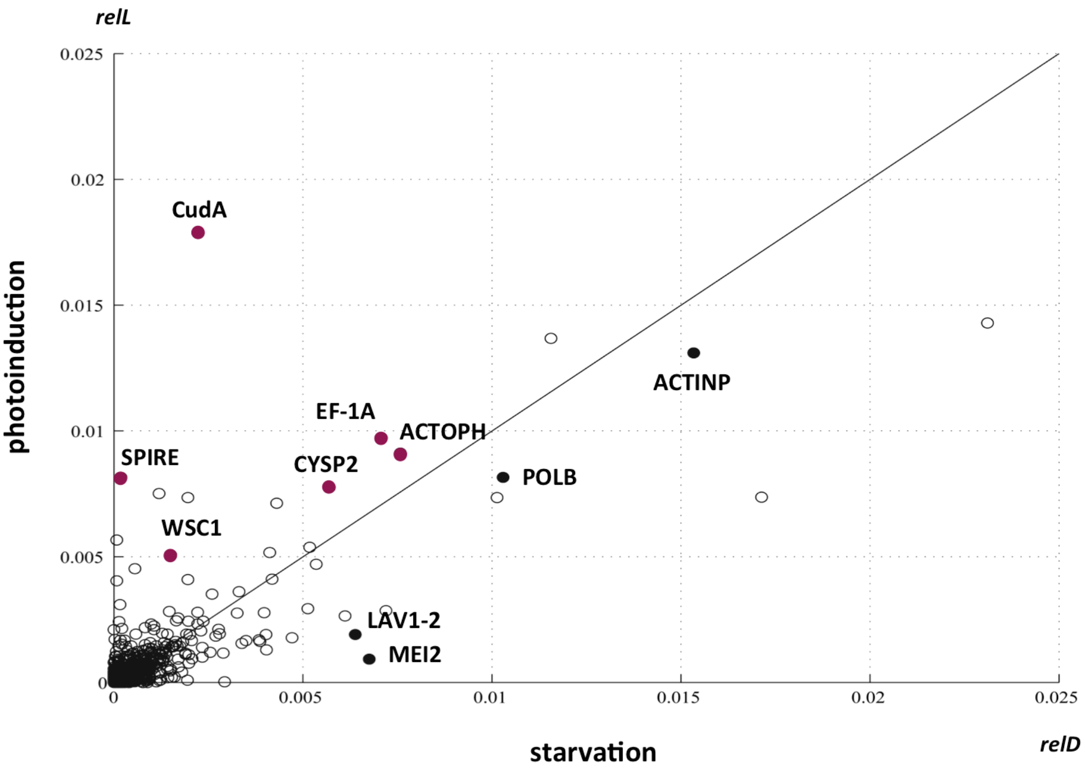


Figure 11. Relative frequencies of transcripts in libraries prepared from competent and photoinduced plasmodia. Each circle represents a single cDNA, plotted according to its relative frequencies (number of hits per transcript divided by the total number of hits) on each cDNA library. relL and relD represent the relative frequencies in the libraries prepared from light-induced and competent plasmodia, respectively. Transcripts more abundant in light-induced (red dots, above the diagonal) or in competent, not light-induced plasmodia (black dots, below the diagonal) are shown, and SwissProt orthologs are indicated for 10 contigs with relative frequencies greater than 0.005.

Table 12. Annotated transcripts with relative frequencies higher than 0.005. A list of transcripts obtained from the scatterplot of relative frequencies (Figure 11) is depicted.

Contig ID	*Annotation*	*hits(D)*	*hits(L)*	*P-value*
ppN1d50g09	Transcriptional Regulator CudA	280	1779	0.00
ppN1d38e09	Elongation Factor 1-alpha, EF1A	887	969	3.31E-12
contig04302	Actophorin	950	908	3.33E-05
contig12806	Cysteine Proteinase 2, CYSP2	714	773	1.17E-09
contig04331	Cell wall integrity and stress, WSC1	189	506	1.03E-52
ppN1d106h10	Spire	23	813	5.38E-250
contig12440	DNA Polymerase beta, POLB	1292	812	7.17E-08
ppN0a10e04	Plasmodial-specific protein LAV1-2	801	190	1.74E-62
ppN1d32d11	Meiosis protein MEI2	848	93	1.26E-118
ppN0a11e12	Actin P, plasmodial isoform	1924	1306	5.56E-06

Gene Ontology Annotation of the Transcriptome

The Gene Ontology (GO) project (Gene Ontology Consortium 2000) is an annotation framework that provides a standardized vocabulary that is used to assign function to uncharacterized sequences, based on three main categories: biological processes (BP), molecular functions (MF) and cellular components (CC). I employed BLAST2GO (Götz, Juan M García-Gómez, et al. 2008), a tool that associates GO terms to sequences based in several annotation evidences, to classify gene function in our dataset. Using the BLASTX hits (annotation e-value cutoff < 1E-6), together with GO terms previously extracted from InterPro domain searches (Hunter, Apweiler, Attwood, Bairoch, Bateman, Binns, Bork, Das, Daugherty, Duquenne, Finn, Gough, Haft, Hulo, Kahn, Kelly, Laugraud, Letunic, Lonsdale, Lopez, Madera, Maslen, McAnulla, McDowall, Mistry, Mitchell, Mulder, Natale, Orengo, Quinn, Selengut, C. J. a Sigrist, et al. 2009), 13,068 GO annotations for 3,304 (20%) cDNAs were inferred, with 11,446 annotations belonging to 2,459 sequences with hit counts data. Annotations of all sequences, including those with or without hit counts data, can be consulted at the following URL:
http://www.biomedcentral.com/content/supplementary/1471-2164-11-115-s7.xls

Transcripts were associated to biological processes (n = 2,437; 15%), molecular functions (n = 2,801; 17%), and cellular components (n = 2,023; 12%). As many as 2,136 (13%), 1,663 (10%) and 1,645 (10%) sequences were annotated with a combination of MF and BP terms, MF and CC, and BP and CC terms respectively, and 1,487 cDNAs were annotated with MF, BP and CC terms altogether. Later, in order to analyze the differences between the two condition groups with respect to the GO annotations, Fisher exact tests were conducted using the Gossip module (Bluthgen et al. 2005) from BLAST2GO. The GO terms '*cell development*' (GO:0048468), '*cell death*' (GO:0008219) and '*death*' (GO:0016265) were found to be overrepresented in cDNAs with higher relative frequencies in light-induced plasmodia (false discovery rate < 0.01), as compared to competent plasmodia (Table 13).

Table 13. Overrepresented Gene Ontology terms in Upregulated Transcripts. Full lists of GO terms from up- and downregulated contigs were compared against each other using the Fisher's exact test from the GOSSIP program (Bluthgen et al., 2005), as implemented in BLAST2GO (Götz et al., 2008). A two-tailed test with the false discovery rate (*FDR*) filter was employed. The number of GO-annotated transcripts used for comparison between up- (*Test*) and downregulated (*Ref*) groups of cDNAs is shown. All overrepresented GO terms belong to the biological process (*BP*) category.

GO term	**GO description**	**FDR**	***P*-value**	**Test**	**Ref**
GO:0048468	cell development	0.009272	0.000314	35	8
GO:0008219	cell death	0.009272	0.000314	35	8
GO:0016265	death	0.009272	0.000314	35	8

Table 14. Summary of the transcriptome sequencing and annotation. [a]Contigs with relative frequencies higher in the competent plasmodial library (*relD*/*relL* > 1), are classified as downregulated, and conversely [b]upregulated transcripts are those with relative frequencies higher in the light-induced plasmodial library (*relL*/*relD* > 1). The significance of differential expression was determined according to the model by Audic and Claverie (1997).

sequencing	
Total 454 reads	405,363
Total sequencing output (Mb)	61.9
Reads from the competent plasmodia library	125,456
Reads from the light-induced plasmodia library	99,632
contigs	
Total contigs	16,669
Contigs with hit counts	13,169
Contigs with at least 5 hits in both libraries	2,103
More abundant in the library from competent plasmodia	3,947[a]
More abundant in the library from light-induced plasmodia	4,972[b]
Downregulated, significant differential expression	1,149
Upregulated, significant differential expression	1,623
similarity search	
Total contigs with blastx results (e-value < 1E-3)	7,778
Contigs with blastx results and hit counts	5,544
Contigs with blastx results and significant differential expression	1,287
annotations	
Total contigs with GO annotations	3,304
Total contigs with KEGG orthologs	2,716
Total contigs with InterPro results	6,813
Contigs with GO annotations and hit counts	2,459
Contigs with KEGG orthologs and hit counts	1,904
Contigs with InterPro results and hit counts	5,180

Pathway classification of transcripts

Functional annotation can also be classified using the pathway-based definition of ortholog genes from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al. 2008). In order to categorize the transcripts in KEGG pathways, the KAAS server (Moriya et al. 2007), a tool that uses similarity information to assign a sequence to a KEGG ortholog (KO) identifier, was employed with default parameters for ESTs. 2,716 (16%) transcripts were mapped to 114 reference metabolic pathways, 1,904 including hit counts data, from which 770 correspond to cDNAs with higher relative frequencies in the library prepared from competent plasmodia, and 743 cDNAs in the library prepared from light-induced plasmodia respectively. In addition, 496 sequences in total were assigned to the KEGG BRITE hierarchies. Transcripts associated to the nucleotide metabolism (n = 110) and citrate cycle (n = 40) had the highest representation for the reference metabolic pathways, and the Wnt, TGF-beta and Jak- STAT signaling pathways were also depicted for the whole dataset (n = 49, 42 and 32 matches respectively). In the whole dataset 420 cDNAs were identified with potential roles in cell differentiation, with molecular entities associated to kinases (n = 140) and GTP binding (n = 110) having the highest representation in the BRITE hierarchies. In addition, 1,159 total enzyme commission (EC) numbers (418 unique) were mapped with 380 unique enzyme names in 851 transcripts, using the EC-module of BLAST2GO (Götz et al. 2008). Later, in order to assess the global metabolic changes that occur after light induction, transcripts with KO identifiers were mapped using the KEGG Atlas tool (Okuda et al. 2008a). For transcripts with higher relative frequencies in the competent plasmodia library, enzymes were mapped for the lipid biosynthesis (map00061) and oxidative phosphorylation (map00190) pathways. Conversely, enzymes for the N-glycan biosynthesis (map00510), urea cycle (map00220) and fatty acid metabolism (map00071) pathways were identified in transcripts with higher relative frequencies in the light-induced plasmodial library (Figure 12). In the end, 2,567 contigs annotated for GO terms, KEGG orthologs, and InterPro hits together were obtained. A summary of sequencing annotations and statistics is listed on the Table 14.

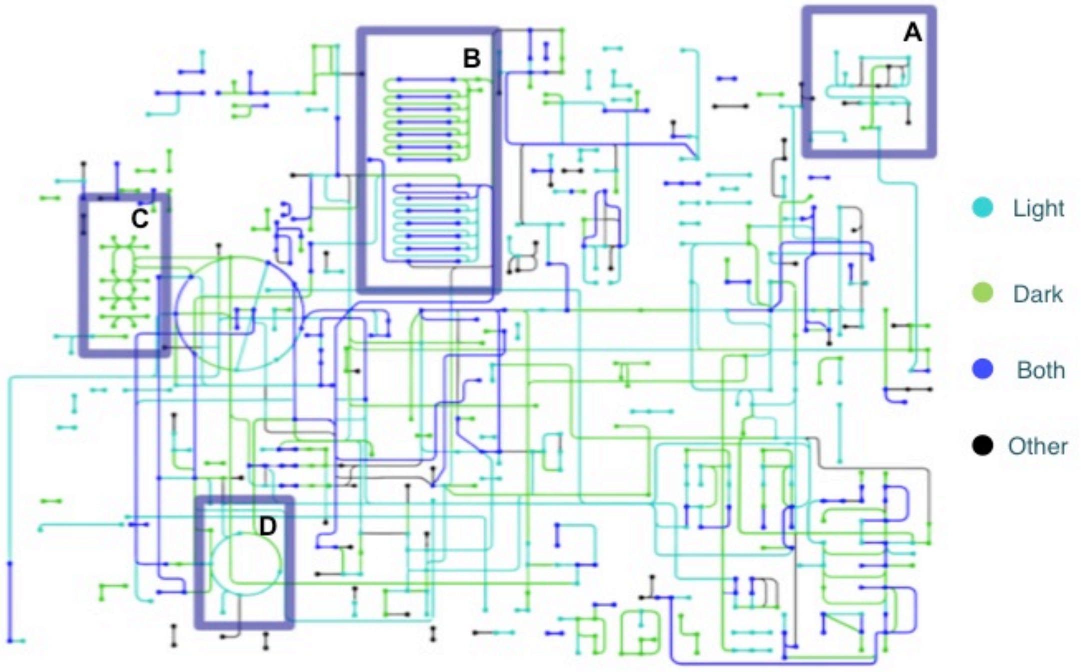


Figure 12. Metabolic Atlas of *Physarum polycephalum*. All *P. polycephalum* cDNAs (Watkins and Gray 2008; Glöckner et al. 2008; and our results) were sent for KEGG Ortholog (KO; Kanehisa et al. 2008) prediction using the KAAS server (Moriya et al. 2007). The output list of orthologs was used to plot this atlas with the KEGG mapping tool (Okuda et al. 2008a). Nodes represent metabolites and edges (lines) correspond to enzymatic reactions. Colors are assigned to either down- (green) or upregulated (light blue) transcripts, and the significance of up- or downregulation was calculated via the model of Audic and Claverie (1997). Transcripts with equivalent relative frequencies in both novel cDNA libraries ($relL/relD$ = 1) are also depicted (blue lines and nodes); black represent those cDNAs with no expression data. After photoinduction, most enzymes from the N-glycan biosynthesis (*A*) and the urea cycle (*D*) pathways are upregulated. In contrast, cDNAs mapped to the oxidative phosphorylation (*C*) had higher relative frequencies in competent plasmodia, whereas a change from fatty acid synthesis to degradation is seen after photoinduction (*B*).

Inference of Interaction Networks

In order to identify the functional relationships between the annotated cDNAs, known interactions in the literature were searched. First, the cDNAs that were previously clustered according to their relative frequencies (Figure 11; Table 12) were used, and included additional proteins whose interactions have been observed in the literature for *Physarum*. Using the "*guilt by association*" heuristic to link coexpressed transcripts into functional groups (Ge et al. 2001; Fraser et al. 2004), an interaction network between those transcripts was inferred. This network is based primarily on actin- binding activities (Figure 13).

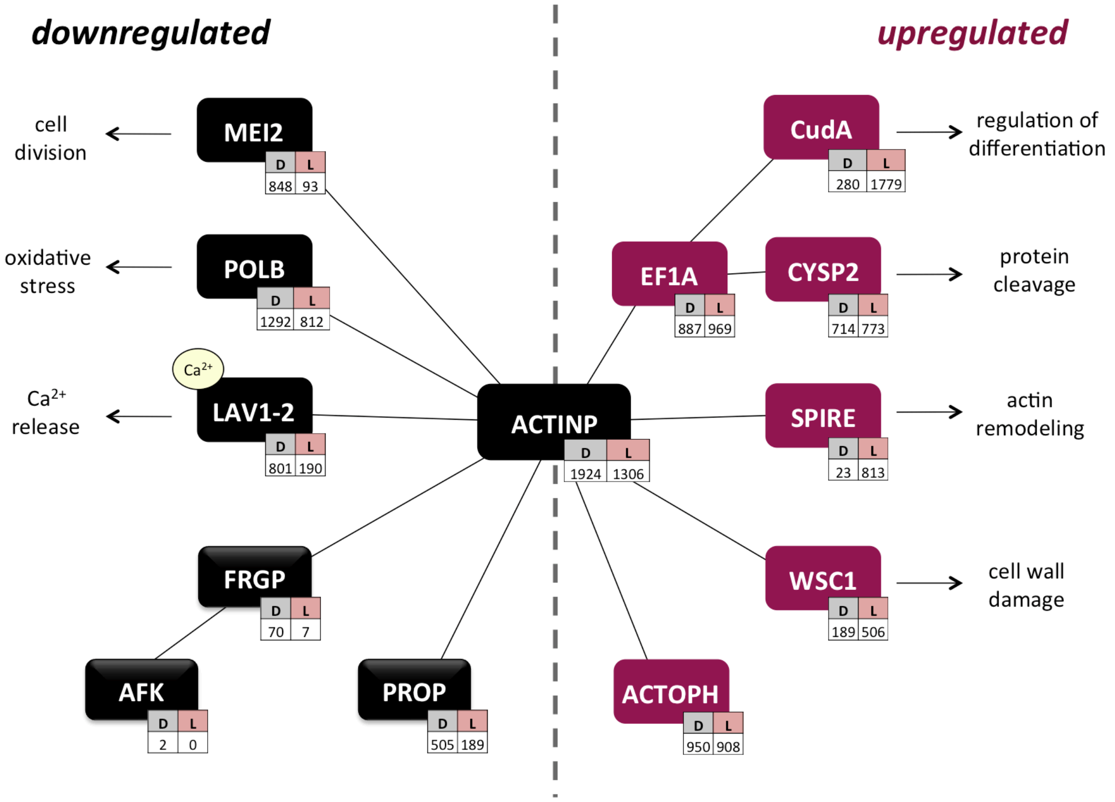


Figure 13. Interactions with the Actin Cytoskeleton of Transcripts with Higher Relative Frequencies. The network was hypothesized from interaction data reported in the literature, using transcripts previously clustered according to their relative frequencies (Figure 11 and Table 12). The transcripts shown are a subset of those from Figure 1, except for certain gene products (*FRGP*, *AFK*, and *PROP*) which were also included as their interactions have been previously observed in Physarum. cDNAs are displayed in colors corresponding to their expression status: down- (black) or up-regulated (red) upon photoinduction, as separated by the dotted vertical gray reference line. Each contig is shown with its hit number counts in both libraries (*D*: competent plasmodia, *L*: light-induced plasmodia).

Later, to identify genes with similar regulation, those transcripts with highest rates of relative frequencies, counted in both cDNA libraries (Tables 15 and 16) were listed. As most of these highly differentially regulated transcripts did not show any sequence similarity to previously annotated genes, the subset of cDNAs with similarity to annotated genes were clustered according to two parameters: (*i*) their rate of relative frequencies; and (*ii*) their statistical significance of differential expression (Audic and Claverie 1997). In this way those 20 transcripts with annotations that were most up- or most downregulated in light-induced plasmodia were listed, based on the statistical significance of their differential expression ($P < 0.05$; Tables 17 and 18). Despite the apparent diversity in biochemical functions, a search for known interactions between these two groups of transcripts was performed.

Table 15. Top 20 Transcripts Downregulated in Light-induced Plasmodia. Transcripts with the highest rates of downregulation (relD/relL > 1.0), are listed. BLAST2GO (Götz et al., 2008) automatic annotations were used, and manual corrections were included in some cases. Transcripts with unknown orthologs are described with "---NA---."

Contig ID	**SwissProt**	**Annotation**	***D***	***L***	**Rate**	***P*-value**
contig12399	---NA---	---NA---	368	3	97.31	1.58E-88
contig12495	---NA---	---NA---	141	2	55.92	1.84E-33
contig00052	---NA---	---NA---	45	1	35.69	4.49E-11
contig10338	P36618	Cell division control protein 16	40	1	31.73	7.48E-10
contig10470	P20072	Annexin A7	68	2	26.97	1.54E-15
contig00397	Q5BMR2	Phospholipase D	62	2	24.59	4.31E-14
contig01934	---NA---	---NA---	27	1	21.42	1.04E-06
ppN0a05b03	---NA---	---NA---	50	2	19.83	3.20E-11
contig00525	Q7EYV7	Poly [ADP-ribose] polymerase 1	244	10	19.36	5.19E-49
contig11321	P38750	Transporter YHL008C	24	1	19.04	5.43E-06
contig03338	---NA---	---NA---	23	1	18.24	9.39E-06
contig02945	---NA---	---NA---	22	1	17.45	1.62E-05
contig02169	---NA---	---NA---	22	1	17.45	1.62E-05
contig00994	---NA---	---NA---	22	1	17.45	1.62E-05
ppNOa14b03	---NA---	---NA---	151	7	17.11	4.08E-30
contig00901	P16064	Subtilisin inhibitor 1	21	1	16.66	2.79E-05
ppN1a03a12	Q07346	Glutamate decarboxylase	20	1	15.87	4.80E-05
contig00391	---NA---	---NA---	20	1	15.87	4.80E-05
ppN1a02c07	P34121	Coactosin A	56	3	14.81	1.07E-11
contig00477	---NA---	---NA---	110	6	14.54	1.67E-21

Table 16. Top 20 Transcripts Upregulated in Light-induced Plasmodia. Transcripts with the highest rates of upregulation (relL/relD > 1.0), are listed. BLAST2GO (Götz et al., 2008) automatic annotations were used, and manual corrections were included in some cases. Columns follow the same convention as in Table 15.

Contig ID	**SwissProt**	**Annotation**	***D***	***L***	**Rate**	***P*-value**
contig10367	---NA---	---NA---	1	79	99.94	2.20E-27
ppN0a10a04	---NA---	---NA---	9	565	79.11	4.30E-184
ppN1d39e07	O08623	Sequestosome 1	3	171	71.69	2.03E-56
contig00236	---NA---	---NA---	1	54	68.31	1.08E-18
contig12905	---NA---	---NA---	2	82	51.87	4.73E-27
contig01485	---NA---	---NA---	1	41	51.86	3.32E-14
contig12498	---NA---	---NA---	10	402	50.68	2.39E-126
contig02685	Q54IV3	ATP-dependent RNA helicase DDX42	1	37	46.81	7.87E-13
ppN1d106h10	Q9U1K1	Spire	23	813	44.55	5.38E-250
contig11969	---NA---	---NA---	1	30	37.95	1.95E-10
contig03550	---NA---	---NA---	3	81	33.96	1.66E-25
contig07470	---NA---	Hypothetical protein EHI83570	1	26	32.89	4.45E-09
contig12244	---NA---	---NA---	1	25	31.62	9.71E-09
ppN1a08g07	O08849	Regulator of G-protein signaling 2	1	22	27.83	9.99E-08
contig12659	---NA---	---NA---	1	22	27.83	9.99E-08
contig05590	Q8H100	GTPase-activating protein 8, AGD8	1	21	26.57	2.17E-07
contig12288	---NA---	---NA---	3	63	26.42	1.91E-19
contig12864	---NA---	---NA---	5	104	26.22	4.67E-31
ppN1a14d12	Q07283	Trichohyalin, TCHH	1	20	25.29	4.69E-07
contig07949	---NA---	---NA---	1	20	25.29	4.69E-07

From annotations of the top up- and down-regulated transcripts (Tables 17 and 18), and including the transcripts from the above mentioned analysis (Figure 13), the initial putative network was extended using Cytoprophet (Morcos et al. 2008). This Cytoscape (Shannon et al. 2003) plugin predicts networks based on information from interaction databases, associated to SwissProt matches of newly annotated genes (Deng et al. 2002). Accordingly, I found that most of these genes encoded proteins predicted to interact in a network of actin-binding proteins (coaA, ABP120, actobindin, *FRGP*, *AFK*, *PROP*; Figure 14). These genes encoding proteins orthologs of which are associated to cell division (*MEI2*, *PUM2*, *CDC16*), DNA repair (*POLB*, *FEN1*), signal transduction (*PP2C*, *CDC16*), and calcium-binding (*LAV1-2*, *KCNIP2*, *GAD*) are downregulated in light-induced plasmodia (Tables 12 and 17). In turn, a different group of developmentally regulated genes is preferentially expressed after photoinduction, including genes the products of which are involved in signaling (*DCR2*, *RGS2*, *YPTC6*, *pakA*), protein processing (*FKBP70*, sequestosome-1, *PSMA7*, *RR7*), cell integrity (*WSC1*, *CDC31*), calcium-binding (*MLR1*, *TRHY*, *PAT1*), and developmentally regulated actin-binding, such as the elongation factor 1 alpha (*EF1A*), spire, and actophorin (Tables 12 and 18; Figures 13 and 14). Interestingly, the previously featured network (Figure 13) connects the two groups of up- and downregulated transcripts in this figure. However, as Cytoprophet gathers experimental interaction data from specialized databases, some interactions depicted in Figure 13 are not shown (e.g., between *POLB* and *ACTINP*), because this data is not present on those source databases used by Cytoprophet for prediction.

Table 17. Top 20 Annotated Transcripts Downregulated in Light-induced Plasmodia. Transcripts with unambiguous annotations, significant differential expression ($P < 0.05$), and that possess the highest levels of downregulation (relD/relL > 1.0), are listed. BLAST2GO (Götz et al., 2008) automatic annotations were used, and manual corrections of annotations were included in some cases.

Contig ID	**SwissProt**	**Annotation**	***D***	***L***	**Rate**	***P*-value**
contig10338	P36618	Cell division control protein 16, CDC16	40	1	31.73	7.48E-10
contig10470	P20072	Annexin A7, ANXA7	68	2	26.97	1.54E-15
contig00397	Q5BMR2	Phospholipase D, PLD1	62	2	24.59	4.31E-14
contig00525	Q7EYV7	Poly ADP-ribose polymerase 1, PARP1	244	10	19.36	5.19E-49
contig11321	P38750	Transporter YHL008C, YHA8	24	1	19.04	5.43E-06
contig00901	P16064	Subtilisin inhibitor 1, ICI1	21	1	16.66	2.79E-05
ppN1a03a12	Q07346	Glutamate decarboxylase, GAD	20	1	15.87	4.80E-05
ppN1a02c07	P34121	Coactosin, COAA	56	3	14.81	1.07E-11
contig11574	P39749	Flap endonuclease 1, FEN1	18	1	14.28	0.000141
contig10414	Q5UNX2	Putative ankyrin repeat protein, YL715	90	5	14.28	8.82E-18
contig03548	O49286	F-box/LRR-repeat protein 5, FBL5	17	1	13.49	0.000242
contig00369	Q80U58	Pumilio homolog 2, PUM2	17	1	13.49	0.000242
contig10457	Q8WN03	Kv channel-interacting protein 2, KCNIP2	16	1	12.69	0.000412
contig00264	P13466	Actin-binding protein 120, ABP120	32	2	12.69	5.26E-07
contig02333	Q8RWN7	Poltergeist phosphatase 2C 32, PP2C	15	1	11.89	0.000701
contig01650	O24496	Glyoxalase II, GLO2C	15	1	11.89	0.000701
contig01322	Q94B74	NADH pyro phosphatase, NUDT2	15	1	11.89	0.000701
contig08310	Q10MW3	Pyruvate decarboxylase isozyme 2, PDC2	73	5	11.58	6.84E-14
contig00558	P18281	Actobindin, ACTO	29	2	11.50	2.55E-06
contig11873	O10296	Apoptosis inhibitor 1, IAP1	28	2	11.11	4.30E-06

Table 18. Top 20 Annotated Transcripts Upregulated in Light-induced Plasmodia. A list of transcripts with unambiguous annotations, significant differential expression (P < 0.05) with the highest levels of upregulation (relL/relD > 1.0), is shown. Annotations, SwissProt accessions, hit counts, and probability values follow the same convention as in Table 17.

Contig ID	**SwissProt**	**Annotation**	***D***	***L***	**Rate**	***P*-value**
ppN1d39e07	O08623	Sequestosome-1, SQSTM1	3	171	71.69	2.03E-56
contig02685	Q54IV3	ATP-dependent RNA helicase, DDX42	1	37	46.81	7.87E-13
ppN1d106h10	Q9U1K1	Spire, SPIR	23	813	44.55	5.40E-250
ppN1a08g07	O08849	Regulator of G-protein signaling 2, RGS2	1	22	27.83	9.99E-08
contig05590	Q8H100	GTPase-activating, AGD8	1	21	26.57	2.17E-07
ppN1a14d12	Q07283	Trichohyalin, TRHY	1	20	25.29	4.69E-07
contig11781	Q55D99	Serine/threonine-protein kinase, PAKA	2	34	21.51	8.84E-11
contig06420	Q9UUG5	Myosin regulatory light chain 1, MLR1	1	17	21.50	4.69E-06
contig08470	Q54MI7	Uncharacterized DDB_G0285917, Y6747	1	16	20.24	1.01E-05
contig12553	Q5R826	Transmembrane protein 63A, TM63A	20	308	19.41	1.89E-83
ppN1d18d06	Q05924	Dosage-dependent cycle regulator 2, DCR2	1	15	18.98	2.15E-05
contig08799	Q43207	Rotamase, FKBP70	1	14	17.71	4.59E-05
contig12445	Q7S045	Non-histone chromosomal 6, NHP6	1	13	16.45	9.76E-05
contig11110	P54678	Calcium-transporting ATPase, PAT1	1	13	16.45	9.76E-05
contig08929	Q39572	Ras-related Protein, YPTC6	1	13	16.45	9.76E-05
contig08360	Q6TQE1	Zinc finger- containing protein 18, NHN1	1	12	15.17	2.06E-04
contig04102	Q9D0C1	Rab RING finger 7, RR7	4	47	14.80	2.86E-13
contig03233	P06704	Cell division control protein 31, CDC31	2	23	14.55	3.44E-07
contig02500	Q5UPW6	FNIP repeat-containing protein, L281	2	23	14.55	3.44E-07
contig08917	Q9PTW9	Proteasome subunit alpha type-7, PSMA7	1	11	13.91	4.35E-04

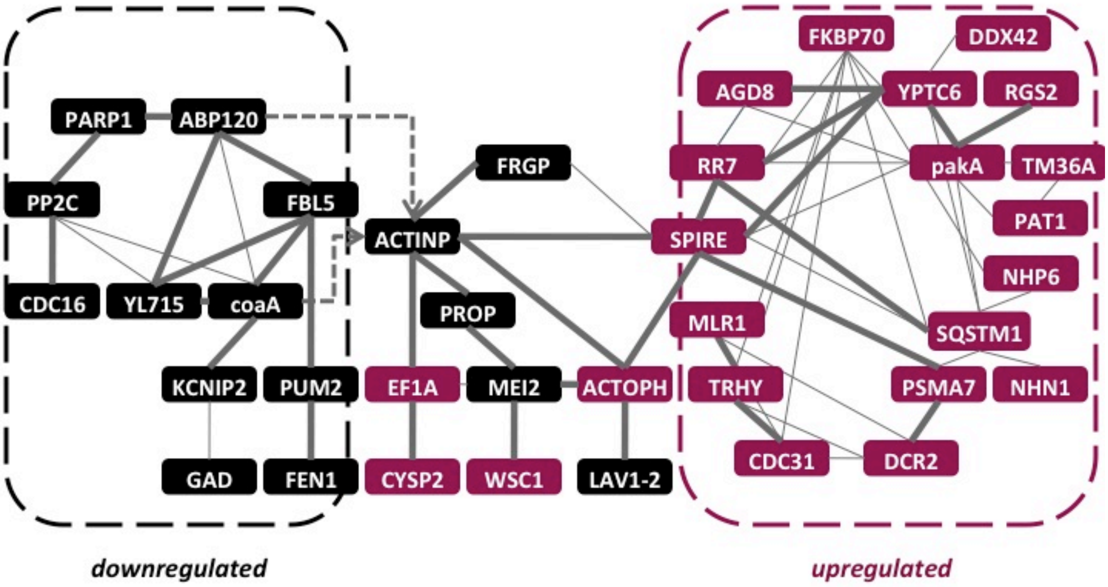


Figure 14. Interaction of the Most Upregulated and Downregulated Transcripts with the Actin Cytoskeleton. The conceptual network was predicted using the Cytoprophet module of Cytoscape, and therefore is solely based on information included on specialized interaction databases. Input transcripts included those from the top up- and down- regulated transcripts (Tables 17 and 18), and cDNAs taken from the previous interaction network (Figure 13). A significant probability of interaction (*P*-value > 0.9) is indicated as a thick edge. Node colors follow the same convention as in Figure 13. This network includes 64 interactions (33 with P > 0.9) between 38 gene products. Genes without Cytoprophet-predicted interactions are not included, except for two interactions with Actin-P that were not predicted by Cytoprophet but that can be found in the literature (indicated with arrows).

Discussion.

The development of plasmodia competent for sporulation includes growth arrest, condensation of cellular constituents, and mitosis (Bailey 1997). Sporulation of competent plasmodia can be triggered by a light pulse. Some proteins associated with the light-induced pathways that lead to sporulation have been described (Martel et al. 1988; Kroneder et al. 1999; Golderer et al. 2001), suggesting that several signaling mechanisms are involved, but there are no studies that describe changes at the level of the whole transcriptome. In the present study the most up- and downregulated transcripts, which are associated to a network of putative interactions, were identified (Figure 14). The network is hypothetical as interactions used for inference are based on data obtained from different organisms. For the sake of simplicity, the discussion will be focused on genes with predicted significant interactions ($P > 0.9$).

A network of actin-binding proteins is associated to changes during light-induced sporulation in Physarum

The actin cytoskeleton of *Physarum* is essential for locomotion, division, and other biological processes (Bailey 1997). Assembly and disassembly of actin filaments is controlled by a group of actin-binding proteins, whose activities in turn are regulated by specific signaling pathways. *Physarum* cell types differ in actin organization but express the same actin genes, suggesting that changes in actin-binding proteins are responsible for the differences in actin organization (Bailey et al. 1999). *Physarum* possesses several classes of actin- binding proteins, and most of these proteins display cell type-specific patterns of expression, but their precise roles are not known (Shirai et al. 2006; Binette et al. 1990). Nevertheless, expression changes in genes coding for actin-binding proteins correlate with modifications in cell organization and behavior (Bailey 1997). In the present study, some actin-binding genes were linked specifically to stages before and after photoinduction in the starved *Physarum* plasmodium.

Specifically, protist orthologs for actin-binding proteins were identified, including Dictyostelium *coaA* (Coactosin A) and *ABP-120* (actin-binding protein 120), and actobindin from *Acanthamoeba*, which binds actin monomers (Table 17; De Hostos et al. 1993; Vandekerckhove et al. 1990). Coactosin A interferes with the capping of F-actin filaments (Röhrig et al. 1995), and is differentially expressed after metal exposure in worms (Brulle et al. 2008). *ABP-120* organizes filamentous actin into networks of fibers, and *Dictyostelium* cells lacking *ABP-120* have a severe phototaxis defect at the multicellular slug stage (Khaire et al. 2007). In addition, transcripts coding for *Physarum* plasmodia-specific actin-binding proteins, such as profilin P (*PROP*; Binette et al. 1990) and fragmin P (*FRGP*; T'Jampens et al. 1999), are downregulated after photoinduction (Figure 13). *FRGP* enables actin phosphorylation by the actin-fragmin kinase (*AFK*), and binds phosphorylated actin (Shirai et al. 2006; T'Jampens et al. 1999). Therefore it is possible that during sporulation these proteins are involved in the reorganization of the subcellular compartments via interactions with the actin cytoskeleton.

Transcripts linked to cell division and DNA repair are downregulated in the light-induced plasmodium

After several days of starvation, cell processes must be limited in order to save energy. Coordination of several biological processes is then required, and thus regulation of these phenomena needs a pleiotropic transducer like the cAMP, which targets several signaling pathways, including those that limit cell proliferation (Howe 2004). Cell differentiation pathways regulated by cAMP levels have been described in *Dictyostelium*, a closely related protist (Aubry and Firtel 1999). In *Physarum*, the *MEI2* gene, which is controlled via cAMP levels, is downregulated in the light-induced plasmodium (Table 12 and Figure 13). *MEI2* is an RNA-binding protein that encodes a cAMP-regulated positive regulator of meiosis in the yeast *S.pombe* (Stettler et al. 1996; Fujioka and Shimoda 1989). This gene product is functionally related to the actin cytoskeleton via the cAMP-dependent protein kinase A (*PKA*; Howe 2004; Aubry and Firtel 1999). Other transcripts downregulated in light-induced plasmodia associated to cell division and DNA repair comprised *FEN1*, *CDC16* and *PUM2*. First, the Flap endonuclease 1 (*FEN1*) appears in several processes linked to the maintenance of the genome integrity, such as the UV-induced DNA repair, as well as in DNA replication and DNA recombination (Christmann et al. 2005; Larsen et al. 2008). Second, the yeast cell division control protein 16 (*CDC16*), constitutes the catalytic subunit of the spg1p GTPase-activating protein, that is involved in the signal transduction controlling septum formation. *CDC16* is involved in cytokinesis and is essential for proliferation, as spores lacking a functional *CDC16* gene complete mitosis without undergoing cell cleavage (Cerutti and Simanis 1999; Fankhauser et al. 1993). Finally, *PUM2* (Pumilio 2) encodes a RNA-binding protein associated to the control of meiosis during development (Lin and Spradling 1997). Consequently, starvation seems to be the signal that regulates cell division while protecting the cells from oxidative stress, through cAMP-regulated pathways (Figure 13).

Other downregulated transcripts in the light-induced plasmodium comprised orthologs of transducers, such as *FBL5*, a leucine-repeat protein linked to phosphorylation-dependent ubiquitination (Jin et al. 2004), *PARP1*, an *Oryza* poly ADP-ribose polymerase, a phospholipase D from *Phytophtora* (*PLD1*), and the *Arabidopsis* phosphatase 2C (*PP2C*, also known as Poltergeist). In plants, G-proteins are involved in phospholipase D activation, and this also seems to be the case for *Phytophtora* (Meijer et al. 2005); on the other hand, *PP2C* operates in several signaling pathways that regulate stem cell differentiation (Yu et al. 2003). It is then reasonable to consider that the differential expression of these transducers is also associated with the control of signaling mechanisms for differentiation, but more profound studies are needed to establish precise causal relationships.

Calcium-binding proteins exhibit diverse regulation patterns in the light-induced plasmodium

Transcripts identified as calcium-binding proteins displayed different patterns of expression regulation. These were either down- (*LAV1-2, KCNIP2* and *GAD*) or upregulated (*MLR1, TRHY*, and *PAT1*) after light induction. *LAV1-2* is a plasmodium-specific RNA of unknown function that encodes a protein containing an EF-hand type domain whose calcium-binding activity has been observed *in vitro* in *Physarum* (Iwasaki et al. 1999). *LAV1-2* seems to act as a sensor of cell damage, releasing Ca^{2+} that leads to the activation of a plasmodium-specific transglutaminase, which separates damaged areas of a plasmodium (Mottahedeh and Marsh 1998). Other transcripts encoding orthologs of calcium-binding proteins, such as *KCNIP2* and *GAD*, were also downregulated in the photoinduced plasmodium and have not been previously described for in *Physarum*. *KCNIP2* encodes a potassium channel-interacting protein that probably modulates channels density in a Ca^{2+}- dependent manner. In turn, the activation of glutamate decarboxylase (*GAD*) by calcium-bound calmodulin (CaM) is required for normal growth in plants (Yap et al. 2003). Previous studies have shown that the intracellular increase of calcium levels is correlated with increased concentrations of cAMP and with sporulation and differentiation in both *Physarum* and *Dictyostelium* (Schlatterer et al. 1994; Renzel et al. 2000). Moreover, actin filament crosslinking is affected by changes in intracellular calcium levels, which ultimately influences the cell contractility (Furukawa et al. 2003). Therefore it seems possible that these calcium-binding proteins coordinate the Ca^{2+} release as a means to influence the cell contractility through the interaction with the actin cytoskeleton (Figure 14).

Furthermore, the upregulated subset of calcium-binding proteins included *MLR1*, which inhibits cytokinesis in yeasts; trychohyalin (*TRHY*), which is involved in its own calcium-dependent processing during differentiation; and the *Dictyostelium PAT1* ATPase. *PAT1* is localized in the membrane of contractile vacuoles, and is a component of a calcium sequestration and excretion pathway, which functions to help maintain homeostasis, especially under conditions of Ca^{2+} stress (Moniakis et al. 1999). Thus these are candidates to control the intracellular calcium levels after light induction of starved plasmodia.

Actin-binding proteins associated to development are overexpressed in the light-induced plasmodium

After photoinduction, a group of actin-binding proteins is upregulated including the elongation factor 1 alpha (*EF1A*), Spire, and actophorin (Figures 13 and 14; Tables 12 and 18). Spire is a *Drosophila* gene involved in development through actin assembly. This gene is also widely distributed across the metazoan genomes. Spire mammalian isoforms are MAP kinase substrates, and data suggest that Spire evolved as an alternative independent mechanism of actin polymerization, necessary for cell polarization in multicellular organisms (Quinlan et al. 2005). Actophorin, in turn, binds actin monomers and separates actin filaments in a dose-dependent manner. Phosphorylation of actophorin blocks actin binding (Blanchoin et al. 2000). In turn, *EF1A*, aside from its role in the protein synthesis, has a separate conserved actin-binding activity in eukaryota, initially observed in *Dictyostelium*, where it is predominantly found in actin-bound form (Yang et al. 1990; Edmonds et al. 1998). *EF1A* regulates the stoichiometry of cytoskeletal components, and the conservation of the *EF1A*-actin interaction across eukaryotes suggests its importance for cytoskeletal maintenance (Gross and Kinzy 2007). Overexpression of *EF1A* in yeast results in effects on cell growth, and influences the actin distribution, morphology and budding in a dosage-dependent manner, although this increase of *EF1A* has no effect over the protein synthesis (Munshi et al. 2001). In addition, changes in cytoskeletal redistribution of *EF1A* seem to be linked to the differentiation status, where the association between *EF1A* and microtubules gradually increases in differentiating cultures (Bluem et al. 2007). Furthermore, *EF1A* stimulates actin remodeling and induces the formation of filopodia, and possibly connects these processes with signaling pathways (Jeganathan et al. 2008; Li et al. 2007).

Remarkably, two coexpressed transcripts (the cysteine proteinase *CYSP2* and the developmentally regulated gene *CudA*) are related to *EF1A*. First, cysteine proteinases are believed to participate in protein cleavage during the differentiation of *Dictyostelium* as a response to starvation (Datta and Firtel 1987), and these peptidases were copurified with *EF1A* in yeasts (Pope and Lee 2005). *CudA*, on the other hand, is associated to the transition from slug migration to culmination in *Dictyostelium*, and *CudA* expression levels depend on local cAMP concentration (Fukuzawa and Williams 2000). Recent evidences show that *CudA* contains a novel DNA-binding site that is distantly related to the metazoan STAT domains, which participate in the regulation of developmentally controlled genes (Yamada et al. 2008), and whose orthologs coexpress with *EF1A* (Li et al. 2007). Yamada et al. (2008) also proved a relationship between *Dictyostelium CudA* and a cDNA from *Physarum*, which corresponds to the contig reported here as a *CudA* ortholog. For these reasons, *EF1A* could work as a link between regulation of the protein synthesis, cytoskeletal maintenance, and signal transduction in slime molds (Figure 13).

Other developmentally regulated transcripts associated to the actin cytoskeleton included the cell wall integrity and stress response component (*WSC1*), which is a yeast membrane protein that acts as a sensor of cell wall damage (Gualtieri et al. 2004), and *CDC31*, a constituent of the nuclear pore complex that is also involved in the maintenance of cell morphology (Table 18 and Figure 14). *WSC1* is essential to keep the cell integrity, behaving like a stress-specific signal transducer that is involved in the reorganization of the actin cytoskeleton in response to osmotic shock (Serrano et al. 2006; Delley and Hall 1999). *WSC1* is involved in the depolarization of the actin cytoskeleton (Delley and Hall 1999), and, like *CDC16* (downregulated in light-induced plasmodia), is entailed in cytokinesis (Cerutti and Simanis 1999).

GTP signaling genes involved in different processes are upregulated in the light-induced plasmodium

Orthologs of certain genes highly upregulated in light-induced plasmodia are involved in signal transduction. These include transcripts linked to the GTP signaling (*AGD8*, *YPTC6*, *RGS2*), kinases (*pakA*) and phosphatases (*DCR2*). The serine/threonine-kinase *pakA* is a regulator of the myosin component of the cytoskeleton, required for cytokinesis and the regulation of the cytoskeleton during chemotaxis in *Dictyostelium* (Chung and Firtel 1999). In turn, the yeast dosage-dependent cell cycle regulator 2 (*DCR2*), is a phosphatase whose increased dosage alters cell cycle progression, while its loss delays the progression in the G1 phase (Pathak et al. 2004). In addition, upregulated GTP signaling transducers included a putative GTPase- activating protein from *Arabidopsis* (*AGD8*); a *Chlamydomonas* GTP-binding protein (*YPTC6*); and *RGS2*, which acts as a negative regulator of G-protein signaling, a function that is evolutionarily conserved in yeast, *C. elegans* and mammals. Increased *RGS2* expression is primarily mediated by the cAMP/PKA pathway (Miles et al. 2000), therefore it is possible that *RGS2* is carrying out similar tasks in slime molds, where it could work in coordination with the other transducers, as hypothesized in Figure 14.

Transcripts annotated for cell death are overrepresented in the light-induced plasmodium

Comparison of GO terms between up- and downregulated groups showed that transcripts annotated for '*cell development*' (GO:0048468), '*cell death*' (GO:0008219) and '*death*' (GO:0016265) were overrepresented in the upregulated group (Table 13). However, all these ontologies belong to the same hierarchy, meaning that '*cell death*' can be the product of either development or organismal death, and hence '*cell death*' is the only difference between both expression groups. One of these cDNAs annotated for '*cell death*' is Sequestosome 1, which is also included on the list of upregulated transcripts (Table 18). Sequestosome 1, also known as *p62*, is a multifunctional protein that targets polyubiquitinated proteins to degradation by proteasomes and autophagy (Seibenhener et al. 2007). *p62* knockouts significantly increased cell death (Bjorkoy et al. 2005), and this is probably linked to the interaction with atypical protein kinase C isoforms that are involved in pathways that control differentiation and apoptosis (Puls et al. 1997). Therefore it is likely that this gene product regulates cell death pathways linked to the commitment for sporulation.

Furthermore, other highly upregulated genes are also functionally linked to the protein turnover. These include the *FKBP70* rotamase, which accelerates the folding of proteins during synthesis; the *PSMA7* proteasome subunit, which together with the other subunits, suffer changes during the meiotic cell cycle (Tokumoto et al. 2000); and the endosome-lysosome vesicle traffic-related *RR7* (Mizuno et al. 2003). It is likely then that these gene products, together with Sequestosome 1, are linked to the control of differentiation through post-transcriptional regulation.

Conclusions

The gain of sporulation-competence of *Physarum* plasmodia involves growth arrest, condensation of constituents, and mitosis and is a prerequisite before sporulation can be induced by light (Bailey 1997). *Physarum* gene expression has been shown to be cell type-specific, but existing studies have been focused only on individual genes (Martel et al. 1988; Kroneder et al. 1999; Golderer et al. 2001). Previously, a library of 5,856 sequences obtained from plasmodia competent for the induction of sporulation was reported by our group (Glöckner et al. 2008). In this chapter the use of the massive parallel sequencing technology at the level of the whole transcriptome (Margulies et al. 2005; Wang et al. 2009) was described in order to identify global changes in expression that occur during light-induced sporulation of *Physarum*. The differentially expressed cDNAs were integrated into networks using interaction information from orthologs and the literature. The results show that after light induction of a plasmodium the expression of transcripts linked to cell division and DNA repair is downregulated. In contrast, light-induction stimulated the expression of genes associated with protein turnover (proteases and proteasome transcripts), genes related to cell cycle progression, and genes involved in the maintenance of cell integrity and cytokinesis. These latter gene products might protect the cell against osmotic shock. Additionally, different groups of calcium-binding proteins are either down- or upregulated after light exposure. These gene products are candidates to control the intracellular calcium levels during sporulation. Here it is postulated that these changes are associated with a network of actin-binding proteins (Figures 13 and 14), the components of which are differentially regulated upon plasmodial photoinduction. These gene products might accomplish different tasks in each stage: the reorganization of the subcellular compartments in order to inhibit migration during starvation on one hand, and cell polarization and cytoskeletal redistribution after photoinduction mediated by a group of actin-binding proteins on the other. The precise representation of the proposed interaction networks is therefore expected to become available as gene knockout experiments, proteomic data, and comparative interactomics are integrated in future studies of this organism.

Summary

Physarum polycephalum displays a complex life cycle, including alternation between single- and multinucleate stages through sporulation. This process of sporulation is a simple form of cell differentiation can be experimentally induced by several external factors, such as red light. In order to identify the genes associated to the light-induced sporulation in *Physarum*, especially those related to signal transduction, RNA was isolated before and after photoinduction from sporulation- competent cells, and used these RNAs to synthesize cDNAs, which were then analyzed using the 454 sequencing technology. 16,669 cDNAs were obtained, which were then annotated at every computational level. 13,169 transcripts included hit count data, from which 2,772 displayed significant differential expression (upregulated: 1,623; downregulated: 1,149). Transcripts with valid annotations and significant differential expression were later integrated into putative networks using interaction information from orthologs. After the integration of annotations, the gene ontology analysis suggested that most significantly downregulated genes are linked to DNA repair, cell division, inhibition of cell migration, and calcium release, while highly upregulated genes were involved in cell death, cell polarization, maintenance of integrity, and differentiation. In addition, transcripts related to cell death were overrepresented between the upregulated transcripts. These changes are associated to a network of actin-binding proteins encoded by genes that are differentially regulated before and after light induction.

Chapter 4. The expressed transcriptome during the differentiation of *Physarum* single cells

Differentiation follows spatial and temporal changes in transcript abundance in a cell type specific manner. Stochastic variations in gene expression presumably do impact cell-fate decisions (Wang and Bodovitz 2010), and therefore the time-resolved analysis of gene expression patterns in individual cells would provide valuable insight as compared to averaged data from measurements obtained on cell populations (Wang and Bodovitz 2010; Tang et al. 2011). Expression patterns of single-cells have been analysed using deep RNA sequencing, or RNA-seq (Nagalakshmi et al. 2008), to characterize the transcriptomes of individual embryonic mouse cells separated by technically complex procedures, and relying on the mouse genomic information for transcript assembly and mapping (Tang et al. 2010; Islam et al. 2011).

At the time of this study, the *Physarum* genome was deposited into the GenBank database (Benson et al. 2011) in the form of 454 sequencing reads (Margulies et al. 2005), but the data was still not assembled into a complete genome sequence (The Genome Institute, Washington University School of Medicine). Therefore, here I evaluated the possibility of studying the global transcriptional changes during the differentiation of *Physarum* single cells through RNA-seq and without relying on genomic information. In this manner an approach was developed to analyze the differential expression at several time points during the commitment of a plasmodial cell to sporulation. The results show that the detected differential expression patterns correlate well with those obtained in cell pools, especially regarding the annotations of the most up- and downregulated transcripts, which are also associated to actin-binding activities, as reported in the previous chapter.

Results and Discussion

Four datasets consisting a total of 77.07 million 100-base reads from single cell *Physarum* plasmodia were obtained from the Illumina sequencing (77.02 million reads with Phred score > 33; 7.12 Gb). This RNA-seq output was deposited in the European Nucleotide Archive (Leinonen, Akhtar, et al. 2011), as the study accession ERP001220 (see Methods, Table 7). The number of reads obtained for each Illumina run (18.28 – 19.94 million reads) is close to the reported optimal range for the creation of a representative *de novo* assembly (20 – 30 millions; Francis et al. 2013). Replicate data distributions were 1.85 and 1.82 Gb corresponding to the starved plasmodium (cDNA library replicates D1 and D2), and 1.67 and 1.78 Gb for the cells collected 6 hours after photoinduction (libraries L1 and L2; Table 19). Therefore, assuming a 10% of protein-encoding genes (see preceding chapter), a 237.32x coverage was obtained for the 300 Mb genome of *Physarum*. The data was then trimmed and filtered for quality (Figure 15), and assembled *de novo* using a combination of the velvet and oases programs (Zerbino and Birney 2008; Schulz et al. 2012). A basic statistic for describing the contiguity of an assembly is the N50 number, which is the length of the shortest sequence contig such that the sum of contigs of equal length or longer is at least 50% of the total length of all assembled contigs (Yandell and Ence 2012). In this case, the assembly consisted of 909,505 sequences, with a N50 contig size of 371 bp.

Large cDNAs from this assembly (>500 bp) were then clustered into 16,822 contigs (N50 length: 778 bp) with CAP3 (Huang and Madan 1999), to create a comprehensive set of representative transcripts. The number of sequence reads that align to an assembled transcript is commonly called a mapped read (or tag count), and this is conventionally used as a measure of gene expression. In this novel transcript assembly, 10,278 of these contigs included transcript abundance data in the form of mapped reads, in at least one cell sample. Transcriptionally active genes were then defined as contigs with at least one mapped read present in all four samples or differentiation stages; in this regard, 8,149 transcripts encompassed mapped reads in all libraries.

To make the mapped read counts for each transcriptionally active gene comparable among samples, a normalization must be performed, which is commonly done as reads per kilobase per million mapped reads (RPKM; Mortazavi et al. 2008). This method is a standard widely used in RNA-seq studies, and consists of two calculations: (*i*) A normalization to library size, which consists in dividing the mapped reads by the total reads in the library; and (*ii*) A normalization to transcript length, that is to divide the mapped read counts by the length of the assembled transcript in kilobases. In the first case, the normalization to library size is done because different replicates with different library sizes would produce different mapped reads for the same gene, and the second is to avoid a fragmentation bias, caused by the fragmentation step during RNA-seq library preparation followed by size selection, where longer transcripts would produce more fragments than shorter ones. Therefore a normalization according to the following equation was performed (Mortazavi et al. 2008):

$$RPKM = \frac{(1 \times 10^{6}\,\mathrm{reads})(\mathrm{transcript\ reads})}{(\mathrm{total\ reads})(\mathrm{transcript\ length\ bp}/10^{3}\mathrm{bp})}$$

$$or:$$

$$RPKM = \frac{10^{9} \cdot \mathrm{transcript\ reads}}{\mathrm{total\ reads} \cdot \mathrm{transcript\ length\ (bp)}}$$

Then a comparison of the RPKM- normalized data for competent plasmodia (D1 and D2 cells) and photoinduced cells (L1 and L2) that were separately processed was carried out (Figure 16). Accordingly, cells from related developmental stages exhibit very similar transcriptomic profiles (competent cells: r = 0.99; light- induced plasmodia: r = 0.98, where r is the symbol of the correlation coefficient). On the other hand, lower correlations were found between competent and photoinduced cells (r = 0.96 between D2 and L1; r = 0.97 in all other cases; Figure 16). However, further studies, involving comparisons between several cell types and cell stages, are required to establish if these lower correlations account for the variations between individual cells.

Table 19. Summary of the single-cell transcriptome sequencing. A summary of the single-cell RNA libraries is shown, corresponding to the competent (libraries D1 and D2), and photoinduced individual cell replicates (L1 and L2).

Culture	***Competent Plasmodium 1***	***Competent Plasmodium 2***	***Induced Plasmodium 1***	***Induced Plasmodium 2***
Replicate ID	D1	D2	L1	L2
Starvation time (days)	6.0	6.0	6.0	6.0
Light exposure (hrs)	----	----	0.5	0.5
Collection time after exposure (hrs)	----	----	6.5	6.5
RNA Samples				
Concentration (ng/ul)	274.0	250.2	472.8	666.4
Amount (ug)	120.6	81.3	93.6	129.3
cDNA amplification cycles	14	15	14	15
Sequencing				
Barcode	CAGATC	ACTTGA	GATCAG	TAGCTT
Total Bases (bp)	1,994,171,100	1,949,941,700	1,827,921,300	1,934,609,000
Total Reads	19,941,711	19,499,417	18,279,213	19,346,090
Reads used for assembly	19,930,198	19,489,244	18,269,297	19,334,649
%GC of reads used for assembly	34%	34%	31%	32%
Total mapped reads	2,128,193	2,076,582	1,567,249	1,937,222
Assembled transcripts with mapped reads	9,365	9,222	9,087	9,226

Afterwards, an analysis of which transcripts were being expressed at different levels in the two studied conditions was needed. Therefore, an estimation of the differentially expressed transcripts was performed from the raw mapped reads, with the R package DESeq (Anders and Huber 2010). The aim of this package is to assess the statistical significance of the differences in gene expression measured in RNA-seq experiments. Mapped read count data follows a Poisson distribution, but because in RNA-seq genes with larger mean counts have greater variances, the DESeq package uses an approximation of the count data with a negative binomial distribution. Briefly, the read count data is first normalized against the geometric mean of the counts, for each gene and across all samples. This step however, is not used to transform the data, but to generate normalization factors that will be employed during the statistical testing. Then, the dispersion within DESeq is estimated from the library coverage, gene expression (mean counts for each individual gene and for each condition), and the variance between genes, which under the model introduced by DEseq, is assumed to be a function of the mean. Finally, the differential expression is tested through the calculation of a probability of null hypothesis, *i.e.* that the gene is expressed at the same level in all conditions. This *P*-value is obtained through a generalized linear model (GLM) test, which is analogous to a Fisher exact test, but using a negative binomial distribution instead.

In the present differential expression analyses, only contigs with a combined count of 300 mapped reads among all the samples were considered, *i.e.*, 3,164 contigs were then selected that fitted these criteria. This mapped read count threshold was selected to reduce the noise caused by spurious contigs and alignments. Upon normalization, the distribution of mapped reads reflected the presence of differentially expressed transcripts and genes with other kinds of regulation, with a slightly greater set of genes with higher expression in light-induced cells (Figure 17). Specifically, 556 upregulated transcripts were identified (*P*-value < 0.05), 504 of these with false discovery rate (FDR) less than 0.1, and 531 downregulated (475 with FDR < 0.1), between the photoinduced and competent cell libraries for transcriptionally active contigs with mapped reads (Figure 17).

Subsequently, to assign functions to the novel sequences, annotations were associated to the transcriptome assembly. In this way 92,641 Gene Ontology (GO) terms were obtained (Gene Ontology Consortium 2008), corresponding to 5,722 SwissProt orthologs (The UniProt Consortium 2010), where 64,730 GO terms belong to 4,222 sequences with mapped reads. cDNAs were linked biological processes (*n* = 1,135), molecular functions (*n* = 1,558), and cellular components (*n* = 576). From the transcriptionally active genes with mapped reads, 231 annotated transcripts were upregulated, and 264 downregulated. These expression data results are fully and publicly available at:
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3469328/bin/Supplementary.xls

A comparison of GO annotations between sets of up- and downregulated transcripts revealed two terms exclusively featured in upregulated genes ('symplast,' GO:0055044; and 'auxiliary transport protein,' GO:0015457). Both annotations are related to the extracellular transport via pores. Conversely, six GO terms were identified only in downregulated transcripts ('*synapse*,' GO:0045202; '*synapse part*,' GO:0044456; '*antioxidant activity*,' GO:0016209; '*translation regulator activity*,' GO:0045182; '*immune system process*,' GO:0002376; and '*viral reproduction*,' GO: 0016032; Figure 18). These groups of GO annotations are associated to the regulation of translation. Next, the enrichment of GO terms in up- and downregulated contigs was tested, against the full list of annotated transcripts, using the Fisher's exact test as implemented in Blast2GO (Götz, Juan Miguel García-Gómez, et al. 2008). In this manner, significant overrepresentation was found only in upregulated transcripts (FDR = 0.037; *P*-value = 0.046), with all GO terms belonging to the molecular function category of ontologies: metal ion binding (GO:0046872), calcium ion binding (GO:0005509), ion binding (GO:0043167), and cation binding (GO:0043169). All these functions belong to the same hierarchy of ontologies, so all these can be summarized with the lower and more specific category, *i.e.*, the '*calcium ion binding*' GO term. Both analyses of GO annotations correlate well with the results shown in the previous chapter, that point to the upregulation of genes associated to the ion transport in the light-induced plasmodium.

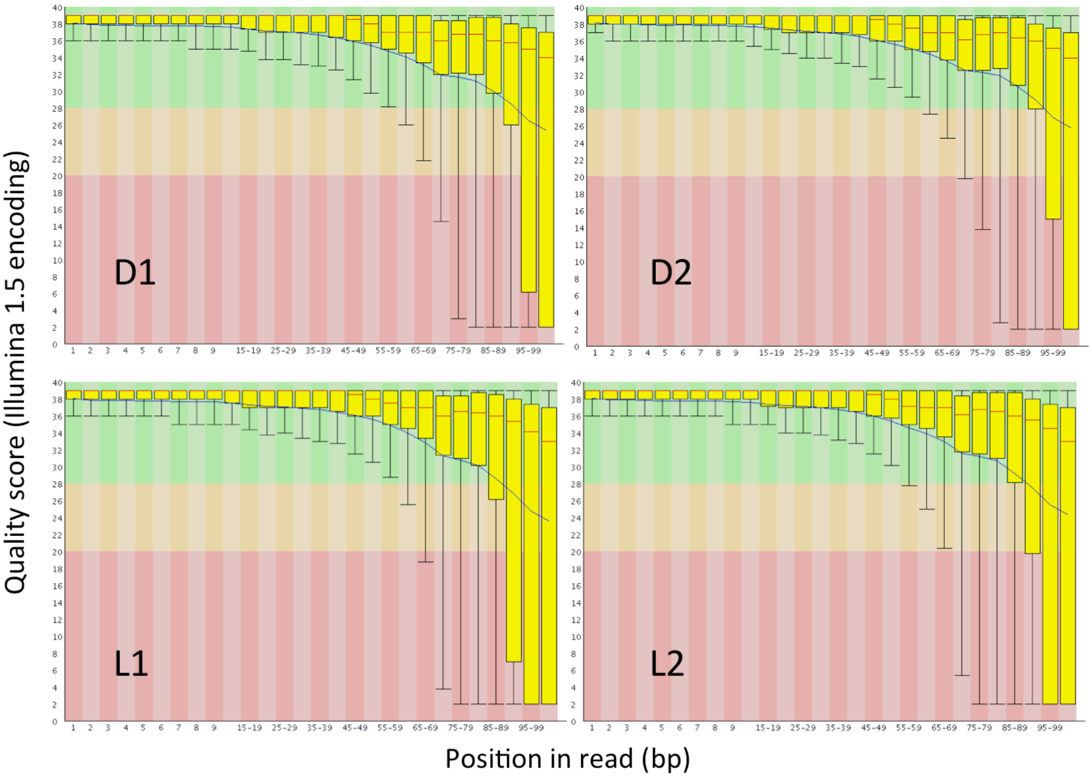


Figure 15. Quality assessment of the Illumina RNA-seq outputs. Each sequencing position in the read, given in base pairs (bp), is plotted against their corresponding sequencing quality values, measured in Phred-64 scores (vertical axis; Cock et al. 2010). These quality values belong to all sequencing reads from each RNA library (L1, L2, D1 and D2; see Table 19), and were obtained from the direct sequencing output (FASTQ format including sequence and quality). Bases with Phred scores over 28 are of very good quality (green area), bases in the brown area are of acceptable quality (Phred score 20-28), while those with score below 20 are of poor quality. The vertical yellow bars indicate the interquartile range, *i.e.* the distance between the upper and lower quartiles, which contains 50% of the plotted values around to the median (indicated with a red line inside the yellow bar). After this quality check, all bases with scores lower than 33 (base call accuracy > 99.95%) were trimmed.

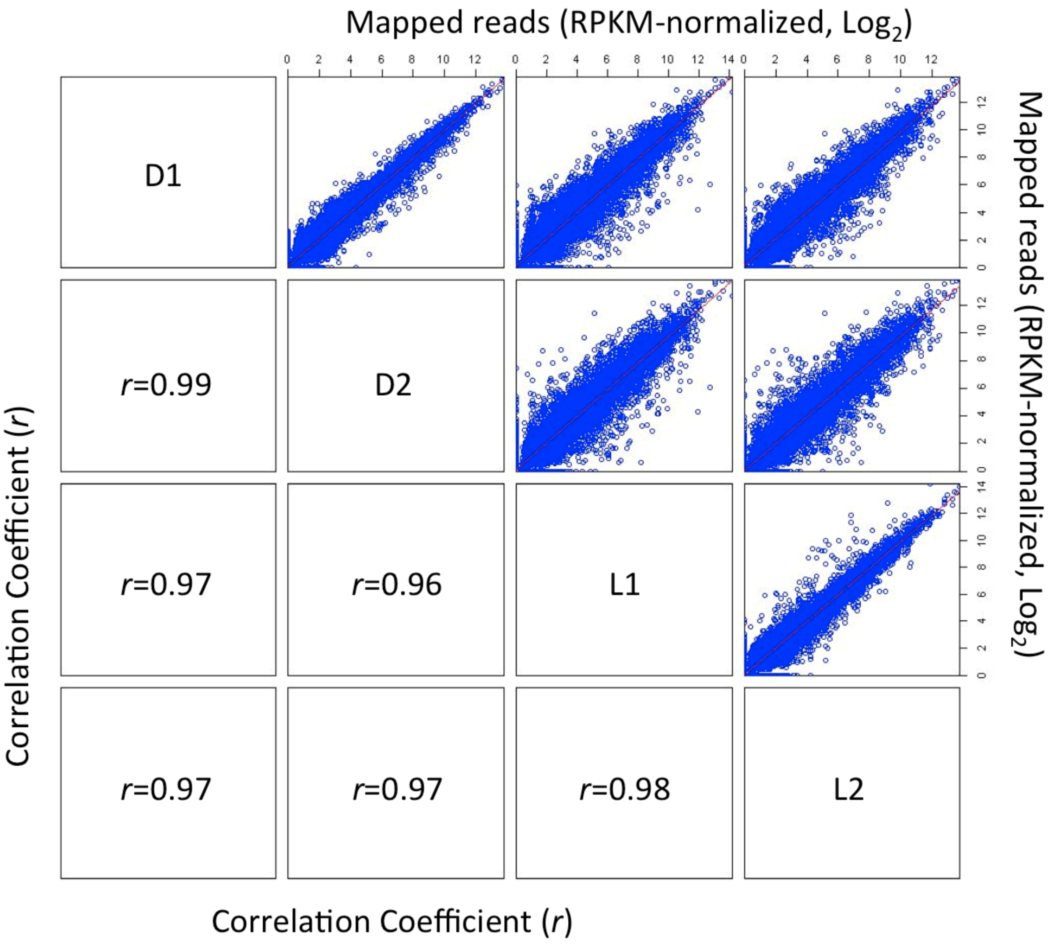


Figure 16. Assessment of the reproducibility of the approach. Correlation plots of the RPKM- normalized reads (Mortazavi et al. 2008) for competent (D1 and D2) and light- induced (L1 and L2) plasmodia were employed to assess the reproducibility of the single-cell RNA-seq in *Physarum*. Reads mapped to the novel transcriptomic assembly were used for plotting. Values of correlation coefficients (*r*) are shown in the corresponding boxes and the red lines indicate no fold changes in expression. Labels of both *x*- and *y*- axis are the Log$_2$ of the RPKM- normalized reads.

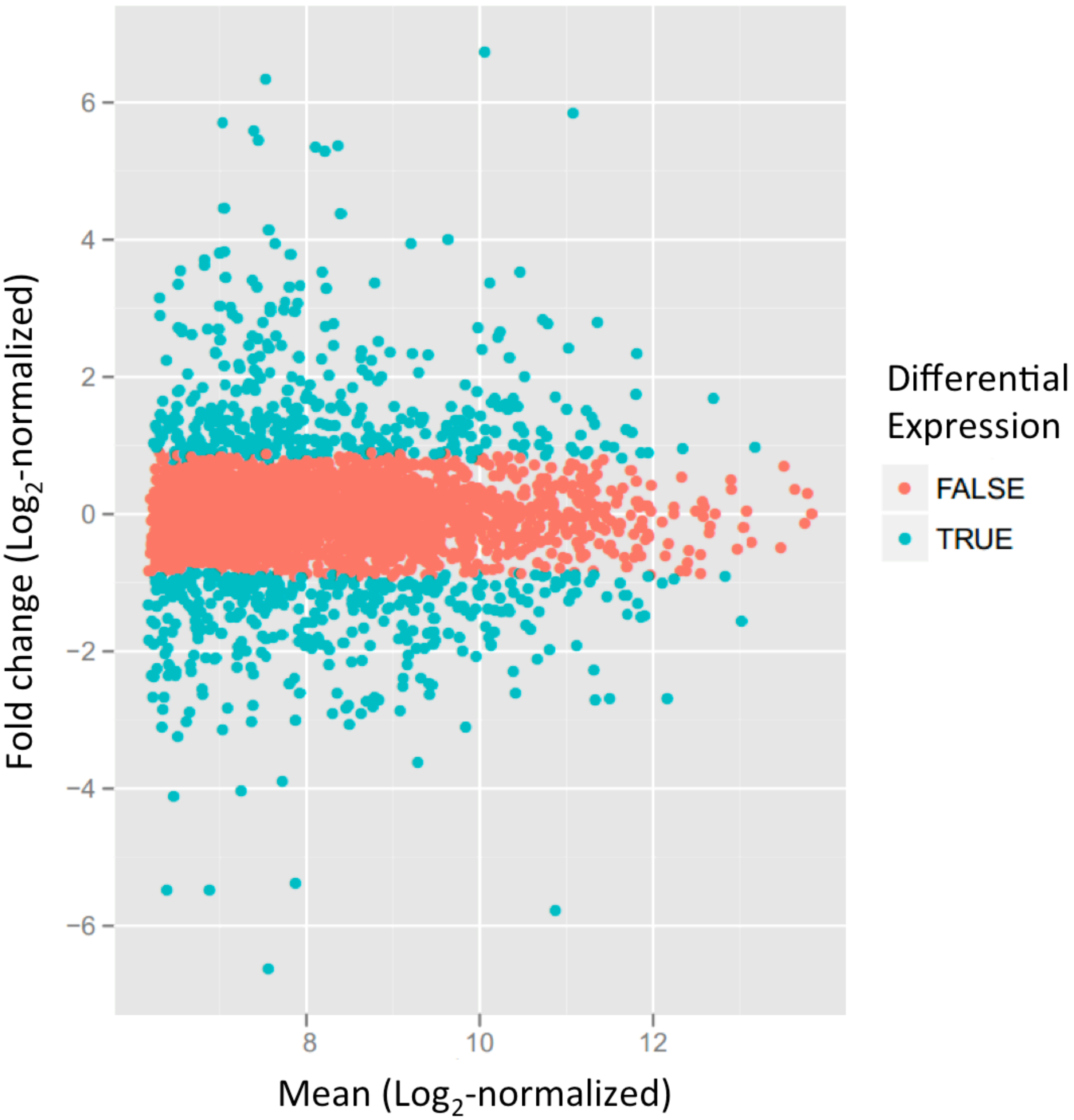


Figure 17. Fold change and significance. Log_2 fold-changes of normalized mapped reads are plotted on the *y*-axis, and log_2-normalized means are plotted on the *x*-axis. Differentially expressed transcripts (turquoise points) were identified between photoinduced and starved single cells of *Physarum,* through the use of a generalized linear model test, as implemented in the R package DEseq (version 1.6.1, false discovery rate < 0.05; Anders and Huber 2010). Transcripts with high fold change may not be significantly diferentially expressed simply due to high variance.

Later, to evaluate the genes with similar regulation, the fully annotated transcripts were clustered for the highest statistically significant up- and downregulation levels, as compared to the starved plasmodium cell libraries (Tables 20 and 21). In spite of the apparent diversity on the annotations, potential functions were inferred based on ortholog identities and gene ontology assignments. In this way, upregulated transcripts were identified encoding endopeptidases (*PHYSA*), phospholipases (*PLDG*) and stress response proteins (*BPM1*, *NAH1*), as well as genes related to biosynthetic processes (*COAD*, *IOD1*, *PYR1*), development (*STX3*), chromatin remodeling (*YA27*), and signaling (*ARF1*, *CYH4*, *SAR1*, *LTBP2*), that are highly expressed 6.5 hours after photoinduction (Table 20).

On the other hand, a different group of genes is downregulated upon light exposure. In this case, transcripts associated to actin- binding (*MYS2*, *COMA*), FMN- binding (*NOS*, *NCPR*), signaling (*VWKA*), sugar- (*TCT1*) and cation- binding proteins (*XANP*, *BOT2*), transporter (*PEP3*) and transferases (*SET5*, *HMNT*), were found as downregulated 6.5 hours after light induction (Table 21). These measurements of transcriptional regulation at different time points correlate well with previous results in cell pools, where actin-binding and signaling proteins were identified as core members of the regulatory network during sporulation (see previous chapter).

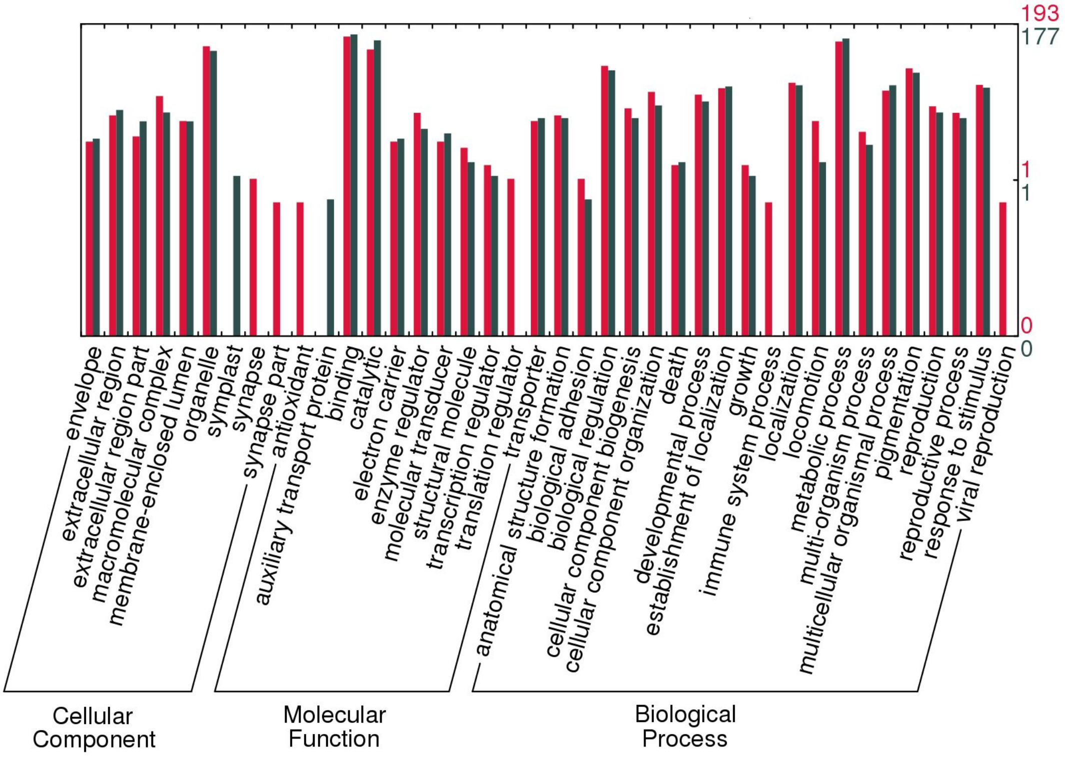


Figure 18. Gene Ontology (GO) classification of differentially expressed transcripts. A comparison of the three main GO categories (Biological Process, Molecular Function, and Cellular Component) between the different expression groups using WEGO (Ye et al. 2006), is shown. Up- and downregulated transcripts are indicated with dark green and red colors, respectively. The y-axis represents the number of transcripts for each GO category, plotted in logarithmic scale.

Table 20. Top 20 Annotated Transcripts Upregulated after Photoinduction. A list of transcripts with unambiguous annotations, significant differential expression ($P < 0.05$), and with the highest levels of upregulation between the competent and light-induced libraries, is shown. Sums of mapped reads (D: starved, L: photoinduced) and fold changes are indicated for each transcript on a given condition under the column "Fold". Blast2GO (Götz, Juan Miguel García-Gómez, et al. 2008) automatic annotations were used, and manual corrections of annotations were included in some cases.

Contig ID	**UniProt**	**Annotation**	**D**	**L**	**Fold**	***P*-value**
s432k3t27235	O00909	ADP-ribosylation factor, ARF1	43	3,475	106.36	3.75E-50
s431k3t6841	Q8MZS4	Physarolisin, PHYSA	10	655	81.16	6.79E-39
s431k4t817	P0CR31	Small COPII coat GTPase SAR1, SAR1	161	7,079	57.20	6.92E-47
s431k4t520	Q86AV9	Phospholipase D, PLDG	15	607	48.14	1.39E-34
s431k4t26745	Q8MZS4	Physarolisin, PHYSA	32	1,043	39.06	5.44E-37
s422k4t53788	Q8L765	BTB/POZ and MATH domain-1, BPM1	67	1,057	20.81	6.62E-28
s422k4t53789	Q8L765	BTB/POZ and MATH domain-1, BPM1	44	589	17.67	8.25E-24
s431k3t89	P49894	Iodothyronine deiodinase, IOD1	201	2,474	16.14	6.09E-24
s431k3t4056	Q20797	Syntaxin-3, STX3	157	1,885	15.39	2.41E-23
s431k3t8494	Q8L765	BTB/POZ and MATH domain-1, BPM1	53	619	15.39	1.31E-22
s432k3t59570	Q28019	Latent-transforming growth factor beta-binding 2, LTBP2	38	459	14.12	2.17E-22
s432k4t17	P20054	PYR1-3 CAD homolog, PYR1	37	441	13.92	3.32E-22
s422k4t53790	Q8L765	BTB/POZ and MATH domain-1, BPM1	66	692	13.83	4.95E-22
s432k3t79109	Q5UPW1	F-box and FNIP repeat-protein, YR286	37	382	12.29	1.83E-20
s422k4t53791	Q8L765	BTB/POZ and MATH domain-1, BPM1	101	884	11.49	2.26E-20
s432k4t1366	Q09698	Uncharacterized C2F7.07c, YA27	493	4,849	11.47	2.67E-22
s431k3t50779	Q99271	Na+/H+ antiporter, NAH1	48	410	10.96	4.53E-18
s431k4t11834	Q80YW0	Cytohesin-4, CYH4	425	3,365	10.32	8.72E-19
ctg9928	P08955	CAD protein, PYR1	90	737	9.90	4.73E-19
s431k4t20190	C1DIB2	Phosphopantetheine adenylyltransferase, COAD	68	519	9.89	4.5E-17

Table 21. Top 20 Annotated Transcripts Downregulated after Photoinduction. Transcripts with unambiguous annotations, significant differential expression (*P*-value < 0.05), with mapped reads in all libraries, and that possess the highest levels of downregulation between the competent and light-induced samples, are listed. Annotations, UniProt accessions, mapped reads, and probability values follow the same convention as in Table 20.

Contig ID	**UniProt**	**Annotation**	**D**	**L**	**Fold**	***P*-value**
s424k4t102	Q53G44	Interferon-induced 44-like, IF44L	792	7	0.01	6.4E-36
s422k3t11297	Q6WP50	Botrydial synthesis protein 2, BOT2	7,869	124	0.02	1.8E-39
s424k4t12560	O74467	SET domain-containing 5, SET5	616	30	0.06	1.5E-23
s424k3t14992	O74467	SET domain-containing 5, SET5	852	46	0.07	2.0E-21
s424k3t34437	P29473	Nitric oxide synthase, NOS	504	48	0.11	8.8E-17
s424k3t13022	Q60106	Xanthomonalisin, XANP	1,507	174	0.13	5.3E-13
s424k4t29841	Q27597	NADPH-cytochrome P450 reductase, NCPR	380	45	0.14	2.5E-14
ctg353	O61063	Tectonin-1, TCT1	2,055	243	0.14	3.6E-13
ctg4361	P08799	Myosin-2, MYS2	311	39	0.14	7.0E-13
s424k3t6526	Q9EST2	Histamine methyl transferase, HNMT	1,322	165	0.14	2.5E-12
ctg5105	O01840	Peptide transporter 3, PEPT3	1,646	199	0.14	1.3E-12
s424k4t30424	Q27597	NADPH-cytochrome P450 reductase, NCPR	1,344	166	0.14	2.1E-12
s424k4t31287	Q60106	Xanthomonalisin, XANP	1,584	208	0.15	7.9E-12
s424k4t4957	Q8T8C0	Nitric oxide synthase, NOS	1,535	195	0.15	4.4E-12
s424k4t2011	Q6B9X6	α-protein kinase, VWKA	9,604	1,216	0.15	1.6E-13
s424k4t319	Q03380	Comitin, COMA	17,238	2,208	0.16	9.7E-13
s422k4t3112	Q03380	Comitin, COMA	1,664	217	0.16	6.2E-12
s422k4t11955	P08799	Myosin-2, MYS2	311	44	0.16	8.4E-12
s422k3t4555	Q03380	Comitin, COMA	1,223	163	0.16	8.6E-12
s422k3t4907	Q6B9X6	α-protein kinase, VWKA	5,034	683	0.16	8.6E-13

Interestingly, the expression of multiple transcript isoforms in the same cell at the same time point, both in up- (*PYR1*, *BPM1*, *PHYSA*; Table 20), as in downregulated transcripts (*COMA*, *VWKA*, *NOS*, *NCPR*, *XANP*, *MYS2*, *SET5*; Table 21) was observed. This phenomenon has been also observed in previous single-cell studies, and has been attributed to the complexity of transcript variants (Tang et al. 2009). Whether these genes encode isoforms controlling stage-specific signalling pathways, remains to be studied in detail.

Before this work, two studies have reported the RNA-seq analysis of transcriptomes in eukaryotic organisms, using single embryonic cells as models (Tang et al. 2010; Islam et al. 2011). In these works, both the assembly and mapping procedures were achieved using the mouse genome as a reference. Here, using the power of RNA-seq to obtain whole transcriptomes without relying on previous genomic information, a characterization of a large set of expressed genes in different samples during the sporulation of *Physarum*, an organism without a known genomic sequence, was performed. Furthermore, in order to obtain single cells, all former studies on single-cell multiplex gene expression analysis required complex separation methods, such as pipetting cells manually, or using laser microdissection or fluorescence-activated cell sorting (Tang et al. 2011). In this study, the plasmodium was used, a natural macroscopic multinucleate single-cell stage from *Physarum*, whose culture and handling is straightforward, and for which there are several well established methods for genetic manipulation (Burland et al. 1993; Bailey 1997; Wolfgang Marwan 2003).

Conclusions

By combining the power of the next generation sequencing technologies, and the simplicity for obtaining single cells from *Physarum*, an approach to characterize the whole transcriptome through the differentiation of this lower eukaryote was developed, at the single-cell level. The observed regulation patterns correlate well with previous studies on the differential gene expression during the commitment to sporulation in the slime mold, particularly with respect to proteins involved in signaling and actin-binding. It is expected that improvements in single-cell transcriptomics, such as the discrimination in sense and antisense transcripts, the ability to sequence a more diverse range of nucleic acid species, and other future developments, will help to display a more precise picture of the regulatory network controlling the differentiation in this organism.

Summary

Cell fate decisions are influenced by stochastic variations in gene expression observed between cells in a population. In recent years, several studies attemped to cope with these variations through the analysis of single cells, which provides a better picture of the expression behavior, as compared to averaged data obtained on cell populations. These studies generally involved complex procedures to separate individual cells, high throughput methods to assess the expression (such as RNA-seq), and the use of the mouse genomic information for transcript assembly and mapping.

Here, an approach for studying the transcriptomic changes during the differentiation of the slime mold in individual cells was developed. This approach combines the use of the *Physarum* plasmodium, a natural macroscopic single-cell, with the power of RNA-seq to obtain whole transcriptomes without relying on previous genomic information.

To test the validity of this approach, first its reproducibility was evaluated through the correlation of expression patterns. Here it was observed that cells from related developmental stages exhibited very similar transcriptomic profiles (competent cells: r = 0.99; light- induced plasmodia: r = 0.98), while lower correlations were found between competent and photoinduced cells (r = 0.96 between D2 and L1; r = 0.97 in all other cases).

Then the gene regulation patterns and transcriptionally active genes were analyzed. In this manner 556 upregulated and 531 downregulated transcripts were identified when comparing the photoinduced against the competent cell RNA-seq libraries. Some of these transcriptionally active genes were associated to annotations (231 and 264 from the up- and downregulated transcripts, respectively), and the combination of expression and annotation data correlate well with previous results in cell pools, where actin-binding and signaling proteins were indicated as core members of the regulatory network during sporulation.

The expression of multiple transcript isoforms was also noticed in the same cell at the same time point. This phenomenon has been also observed in previous single-cell studies, and has been attributed to the complexity of transcript variants. Finally, analyses of gene ontology classifications and enrichment also

correlate well with the results shown in the previous chapter, that point to the upregulation of genes associated to the ion transport in the light-induced plasmodium. It is expected that this single-cell transcriptomics approach will enable in the future to display a more precise picture of the regulatory network controlling the differentiation in this organism.

Chapter 5. The reference transcriptome of *Physarum polycephalum*

Background

Many aspects of the biology of an organism are encoded in its genome. Genomes display phenotypically in a given condition through their expressed transcriptomes, while the whole set of transcripts comprised in a given genome is its reference transcriptome. Recent technological advances, and particularly the development of the next generation sequencing methods, make possible to survey the transcriptional complement of the genomes at the single base level. When the full sequence of the reference transcriptome is known, the effort then shifts in finding the biological function of the encoded genes (U.S. Department of Energy 1992; Guigó 2013).

In the case of *Physarum*, although genetic manipulation is possible, and comprehensive genomic and transcriptomic information are available for several closely related organisms, such as *Dictyostelium discoideum* and *D.purpureum*, the study of biological functions in the slime mold at the molecular level is still restricted to small groups of genes. In this respect, next generation sequencing technologies have nowadays allowed the study of several model organisms, even of those that are not amenable to classic genetic methods (National Institutes of Health 2004). Given the potential of *Physarum* as a model in many research areas, a genome consortium was formed (Physarum Genome Sequencing Consortium 2013), which sequenced and delivered a draft of the genome assembly.

Here, in order to identify all genes possibly associated with the sporulation in the slime mold, the *Physarum* genome was characterized and all its protein coding genes annotated, which were later organized into putative regulatory networks linked to biological processes such as signal transduction and differentiation. The process specifically involved searching and masking repetitive regions, and then the masked genome was used to map cDNAs (derived from *Physarum* ESTs and RNA-seq, and *D.discoideum* ESTs), and proteins from the UniProt database. In parallel, novel noncoding RNAs (ncRNAs) were also identified and mapped.

Outputs from these computational experiments were integrated for annotation, and the protein coding gene models evaluated for certainty and completeness (Figure 2). The analyses of the genome and the putative reference transcriptome presented here not only provide the first steps to a better understanding of the slime mold biology at the transcriptomic level, but also serve as a pilot pipeline that can be used for the annotation of the final genome release.

Results

The Physarum Genome

The genomic DNA sample was obtained from haploid amoeba (strain LU352; Table 3), by Gerard Gernot and Marianne Bernard (Integrated Research Cancer Institute, Villejuif, France), and sequenced using a whole genome shotgun strategy, at The Genome Institute of the Washington University School of Medicine (St Louis, MO), under the supervision of Patrick Minx. The platforms used were Roche 454 instruments, and the combined sequence reads were assembled via the Newbler package, version 2.6 (454 Life Sciences, Roche). The contaminating contigs, as well as redundant contigs resulting from high levels heterozygosity, have been removed from the final assembly. The retained contigs were then scaffolded, *i.e.* reunited into scaffolds, by introducing artificial gaps (represented by *N*s), whose lengths were calculated from the clone or sequence libraries of origin. Afterwards, the scaffolds of at least 200 bases were submitted to GenBank, where they were stored under the accession number 709848. This draft assembly is referred as version 7.3.1, with coverage of 54.6X (Patrick Minx, personal communication). According to the present analyses, this *Physarum* genome release contains 126,782 scaffolds, with a total length of 239,752,614 base pairs (189,684,779 bp excluding undefined bases), and a GC-level of 41.16%. The results here shown however may differ from the GenBank version, as the NCBI staff performs further filtering of contaminants and sequences prior to the public release. A summary of statistics of the genomic contigs and scaffolds is listed at the Table 22, and the distribution of these fragments is shown in the Figure 19.

Table 22. Sequencing summary for the genome assembly, release 7.3.1

	Scaffolds	***Contigs***
Total (Mb)	239.75	189.68
Undefined (Mb)	50.08	0.015
Real bases (Mb)	189.67	189.67
Sequences	126,782	189,840
Mean Size	1,891.1	999.2
Smallest	17	17
Largest	821,622	74,487
Fragment N50	97,377	2,096
N50 length	119,912,848	94,842,743

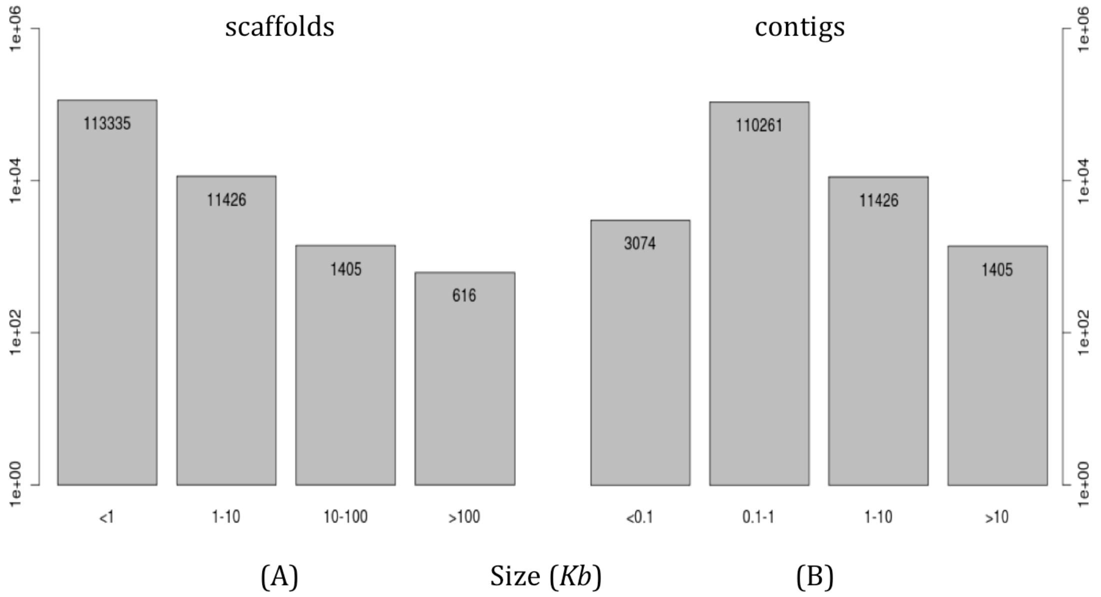


Figure 19. Distribution of scaffolds and contigs in the *Physarum* genome. The fragment size range (in Kb, *x*-axis), is plotted against the number of sequences (*y*-axis), for all the scaffolds (*A*) and contigs (*B*) in the genome, version 7.3.1 (GenBank Accession 709848; Unpublished). Number of fragments for a given size range are indicated inside the bars (*e.g.*, there are 616 genomic scaffolds with sizes over 100 Kb).

Repeat Annotation

Identification and masking of repetitive sequences is widely regarded as the first step towards genome annotation. Two kinds of sequences are considered repeats: homopolymeric tracts ("low- complexity regions"), and transposable elements, such as short and long interspersed nuclear elements (SINEs and LINEs, respectively). Repeats are extensively distributed in eukaryotic genomes, and their borders usually overlap or occur inside other repetitive elements, with most repeats rarely found complete (Lerat 2010). After repeat searches, genome sequences are typically masked, *i.e.*, each nucleotide regarded as part of a repetitive element is changed for an "*N*." Given that most programs used for gene annotation are sensitive to low- complexity sequences, they complicate genomic characterizations; and genomes without masking can produce millions of spurious similarity alignments, repetitive elements must be identified before genes are mapped and modeled (Yandell and Ence 2012). Here, a repeat search was carried out with the RepeatMasker software (version open-3.3.0; Smit et al. 2010). This Perl program uses a search engine like BLAST with a library of transposable elements, satellites, and typical low complexity sequences, to detect these in novel genomes (Tempel 2012). The default mode was employed, supported by the Tandem Repeats Finder (version 4.07b; Benson 1999), RMBLASTN (version 2.2.27+) and the RepBase database (update 20120418; Jurka et al. 2005). In this manner, a total of 34,875,330 bp (14.55%) were masked from the scaffolds. Following the classification of eukaryotic transposable elements (Wicker et al. 2007), most elements were found entailing LINEs (337,725 bp; 0.14%), simple repeats (13,11 Mbp; 5.47%), and low complexity regions (21,35 Mbp; 8.9%). The output of the RepeatMasker analysis was a masked genome of 154,830,967 bp excluding undefined bases, that was used later as the target for EST and cDNA mapping.

However, the RepBase library release contains only three *Physarum*- specific sequences, and 2 ancestral from the Mycetozoan lineage, from a total of 179 sequences. This, together with the fact that the *Physarum* genome is larger and more fragmented than those from related social amoebae (Hardman et al. 1980; Sucgang et al. 2011; Eichinger et al. 2005), encouraged the building of a species-specific transposable elements database. For this purpose, the repeat modeling package RepeatModeler was employed (version open-1.0.7; Smit and Hubley 2010). RepeatModeler internally calls two *de novo* repeat finding programs (RECON and RepeatScout; Bao and Eddy 2002; Price et al. 2005), and uses their outputs to build a library of putative interspersed repeats. Following the recommendations of Hu (2011), only sequences longer than 100 Kb were employed to build the custom library (616 scaffolds; 118,143,527 total bp; 86,429,144 bp excluding undefined bases). The obtained library consists of 338 nucleotide sequences distributed among 23 repeat families, with a total of 289,971 bp (N50: 1640 bp; N50 length: 144,972; see Figure 20). This novel repeat library was then used for a new repeat search using RepeatMasker. Here a significant increase in total number of elements was observed, which went from 0.17% with the default RepBase library, to covering up to 15.29% of the genome with the custom library, and also increasing the proportion of the genome masked (from 14.55 to 27.59%; see Table 23). However, these results must be taken carefully, as these are algorithm predictions, and more wet-lab research is needed to verify the nature of these candidate regions. Nevertheless, the custom masked genome output still entails an important resource, and thus it was included later during the modeling of protein coding genes (see page 119).

Table 23. Distribution of repetitive elements on the *Physarum* genome. These regions were identified with RepeatMasker, using either RepBase (Update 20120418, "*default*" library), or a RepeatModeler- custom built library ("*custom*" library). The elements listed in the first column follow the classification of Wicker et al. (2007). Column parameters represent the number of elements found, the length covered in bp., and the proportion in percentages (*Perc %*).

Library	**Default**			**Custom**		
Parameter	*Elements*	*Length*	*Perc (%)*	*Elements*	*Length*	*Perc (%)*
SINEs	198	27,026	0.01	0	0	0.00
ALUs	145	22,094	0.01	0	0	0.00
MIRs	35	3,999	0.00	0	0	0.00
LINEs	1,494	337,725	0.14	6,713	2,893,999	1.21
LINE1	1,340	327,197	0.14	2,404	26,625	0.20
LINE2	53	3,724	0.00	0	0	0.00
L3/CR1	83	5,504	0.00	0	0	0.00
LTR elements	168	28,573	0.01	17,579	6,657,620	2.78
ERVL	21	4,008	0.00	0	0	0.00
ERVL-MaLRs	29	5,470	0.00	0	0	0.00
ERV_classI	100	14,868	0.01	909	176,670	0.07
ERV_classII	12	3,285	0.00	0	0	0.00
DNA elements	106	17,495	0.01	12,836	3,128,705	1.30
hAT-Charlie	18	2,155	0.00	0	0	0.00
TcMar-Tigger	37	3,154	0.00	0	0	0.00
Unclassified	2	210	0.00	130,568	23,979,251	10.00
Total	---	411,029	0.17	---	36,659,575	15.29
Small RNA	397	36,756	0.02	0	0	0.00
Satellites	101	23,133	0.01	1	241	0.00
Simple repeats	180,381	13,113,566	5.47	306,223	23,422,801	9.77
Low complexity	262,728	21,348,986	8.90	78,942	6,662,193	2.78
Bases masked	---	34,875,330	14.55	---	66,150,424	27.59

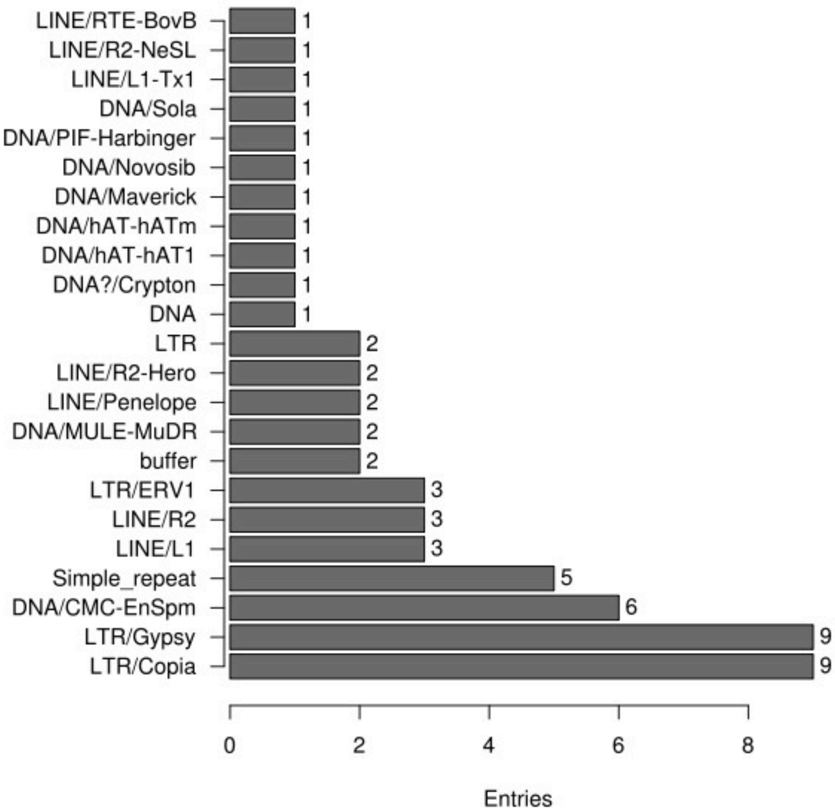


Figure 20. Distribution of repeat families at the *Physarum*- specific custom repeat library. The *y*-axis specifies the family name, and the number of entries for each family is plotted on the *x*-axis.

Non-coding RNA (ncRNA) Annotation

In eukaryotes, most of the genomic DNA comprises non-protein-coding transcripts. These RNAs consist of many heterogeneous groups, and the best-characterized ncRNA classes are known to form secondary structures that are relevant for their function. These classes include ribosomal RNAs (rRNA), small nuclear RNAs (snRNAs), and transfer RNAs (tRNAs) that are involved in messenger RNA (mRNA) splicing and translation. Also in this group are catalytic RNAs such as snRNAs, RNase P RNA, and other ribozymes, and regulatory RNAs such as microRNAs and spliceosomal RNAs, which direct protein complexes to RNA targets. In addition, ncRNAs are also known to be involved in the regulation of gene expression, chromosome replication, RNA processing and modification, mRNA stability, protein degradation and translocation (Dhanasekaran et al. 2013).

In this work, a combination of similarity, pattern and *ab initio* approaches were used to find all noncoding RNA classes in the *Physarum* genome. To this end, first all possible noncoding RNAs were identified using the Infernal package (version 1.0.2; Nawrocki et al. 2009). This program combines the use of probabilistic models of known consensus RNAs, built from collections of RNA families present in the Rfam database (version 11.0, August 2012; Griffiths-Jones et al. 2005; Burge et al. 2013), with similarity searches against the sequences of these consensus models. For this analysis and in order to reduce the computation time, a prefiltering was first performed through a BLASTN search (Altschul et al. 1990; Altschul et al. 1997) against the noncoding RNA sequences present in the Rfam database, with an E-value of 0.01. This was achieved using a perl script obtained from Rfam (*rfam_scan.pl*, version 1.0.4), modified for multi-thread execution. The filtering and the Infernal search allowed the identification of 1,436 ncRNAs, comprising 144 small nucleolar RNAs (snoRNAs; Table 24), 777 micro-RNAs (as indexed in miRBase, see Table 25; Kozomara and Griffiths-Jones 2011), 16 RNAs involved in Group II intron splicing (Table 26), 29 bacterial small RNAs (Table 27; possible vector contaminants), and 183 members of other diverse non-coding RNA families (Table 28).

Subsequently, tRNA gene structures were predicted with the tRNAscan-SE program, with default parameters (version 1.23; Lowe and Eddy 1997). tRNAscan-SE uses probabilistic structure profiles built from known RNAs, to find novel tRNAs in uncharacterized sequences. tRNAscan-SE predicted 325 tRNA genes in this release of the *Physarum* genome, 281 of these encoding the twenty standard amino acids, and including eleven Selenocysteine (Sec) tRNAs (Table 29). Furthermore, 29 were predicted as tRNA pseudogenes, 30 containing introns (see Table 30), and 4 encoded undetermined tRNA isotypes. No suppressor tRNAs (CTA and TTA anticodons) were found.

Table 24. Small nucleolar RNAs (snoRNAs) in the *Physarum* genome. A total of 144 snoRNA homologs were found, corresponding to three main taxonomic groups: 115 from *Plasmodium falciparum*, two from a human homolog, and several types belonging to Trypanosomatid ncRNA sequences (27 snoRNAs). Accession numbers are listed as Rfam entries (Griffiths-Jones et al. 2005).

Type	**Origin**	**ID(s)**	**Accession(s)**	**Number**
snoR11	*P.falciparum*	Single member	RF01589	115
SNORA17	Human	Single member	RF00560	2
snoTBR	*Trypanosoma*	snoTBR17 snoTBR5 snoTBR7	RF00294 RF00292 RF00295	3
TB10	*Trypanosoma*	TB10Cs1H1 TB10Cs1H2 TB10Cs1H3 TB10Cs2H1 TB10Cs3H1 TB10Cs3H2 TB10Cs4H2	RF01522 RF01523 RF01524 RF01525 RF01531 RF01532 RF01862	7
TB11	*Trypanosoma*	TB11Cs2H1 TB11Cs4H1 TB11Cs4H2	RF01537 RF01539 RF01540	4
TB6	*Trypanosoma*	TB6Cs1H1 TB6Cs1H3 TB6Cs1H4	RF01546 RF01547 RF01548	3
TB8	*Trypanosoma*	TB8Cs2H1 TB8Cs3H1	RF01549 RF01550	2
TB9	*Trypanosoma*	TB9Cs1H1 TB9Cs1H2 TB9Cs1H3 TB9Cs2H1 TB9Cs3H1 TB9Cs3H2 TB9Cs4H2	RF01861 RF01552 RF01553 RF01554 RF01555 RF01556 RF01558	8

Table 25. Micro RNAs (miRNAs) in the *Physarum* genome. miRNAs found are all involved in post-transcriptional regulation and belong to diverse species, and were obtained from miRBase (Kozomara and Griffiths-Jones 2011). Accession, gene ID, and the number of genes found (*Number*), follow the same convention as in the Table 24.

Accession	**ID**	**Number**
RF00639	mir-515	2
RF00665	mir-290	1
RF00690	MIR408	193
RF00692	MIR171_	1
RF00708	mir-450	42
RF00729	mir-278	285
RF00736	mir-320	1
RF00758	mir-346	1
RF00788	mir-287	105
RF00834	mir-268	2
RF00871	mir-689	1
RF00876	mir-684	19
RF00885	MIR821	2
RF00886	MIR807	2
RF00911	mir-672	45
RF00929	mir-574	14
RF00994	mir-1255	1
RF01005	MIR530	1
RF01021	mir-558	1
RF01059	mir-598	50
RF01063	mir-324	8

Table 26. Group II intron splicing non-coding RNAs found in the *Physarum* genome. These molecules are a class of self- catalytic ribozymes and mobile elements. Accession, number and ID columns follow the same convention as in the Table 24.

Accession	**ID**	**Number**
RF00029	Intron gpII	8
RF01998	group-II-D1D4-1	1
RF01999	group-II-D1D4-2	1
RF02001	group-II-D1D4-3	1
RF02003	group-II-D1D4-4	3
RF02012	group-II-D1D4-7	2

Table 27. Bacterial non-coding RNAs found in the *Physarum* genome. Accession and numbers follow the same convention as in the Table 25.

Accession	**Functional Category**	**Number**
RF02076	Gammaproteobacterial sRNA STnc100	1
RF02221	*Xanthomonas* small RNA, sRNA-Xcc1	1
RF02278	Betaproteobacteria toxic small RNA	2
RF00624	P9, small RNA from *P.aeruginosa*	1
RF00106	regulation of DNA replication RNAI	6
RF00391	bacterial cis-regulatory element RtT	2
RF00442	Detoxification in *B.subtilis,* ykkC-yxkD	1
RF01699	RNA motif from Clostridial bacteria, Clostridiales-1	8
RF01725	SAM riboswitch, Gram-positive bacteria, SAM-I-IV-variant	1
RF01757	DNA repair in Burkholderiaceae, sbcD	1
RF01766	cold shock response in Enterobacteriales, cspA	3
RF01497	frameshifting in bacteria, ALIL	1
RF00240	Inhibition of IS10 transposase expression, RNA-OUT	1

Table 28. Other non-coding RNAs found in the *Physarum* genome.

Accession	**Functional Category**	**Number**
RF00009	tRNA processing, RNAse P	1
RF00019	Ro ribonucleoprotein particle (Ro RNP), Y-RNA	33
RF00032	Histone mRNA 3'-end processing	3
RF00039	mRNA binding, DicF	1
RF00174	cobalamin binding	1
RF00198	nuclear mRNA trans splicing SL1, via spliceosome	2
RF01656	small RNAs ceN72-3 ceN74-2, function unknown	6
RF01666	Cis-regulatory element rox2	51
RF00003	Splicing U1	20
RF00004	Splicing U2	31
RF00007	Splicing U12	1
RF00015	Splicing U4	6
RF00020	Splicing U5	10
RF00026	Splicing U6	13
RF00619	Splicing U6atac	4

Table 29. tRNA Genes in the *Physarum* Genome. The corresponding codons are listed from the universal genetic code, in IUPAC notation. The number of tRNA genes for each codon is indicated between parentheses. The presence of the 20 standard proteinogenic amino acids for eukaryotes confirms the completion of the draft genome.

Amino acid	**Codons**	**tRNA genes**
Ala/A	GCT (9), GCC, GCA (8), GCG (4)	21
Arg/R	CGT (6), CGC, CGA (3), CGG (2), AGA (4), AGG (4)	19
Asn/N	AAT, AAC (6)	6
Asp/D	GAT, GAC (6)	6
Cys/C	TGT (1), TGC (7)	8
Gln/Q	CAA (6), CAG (4)	10
Glu/E	GAA (7), GAG (6)	13
Gly/G	GGT (1), GGC (13), GGA (10), GGG (3)	27
His/H	CAT, CAC (10)	10
Ile/I	ATT (7), ATC (1), ATA (3)	11
Leu/L	TTA (4), TTG (5), CTT (7), CTC, CTA (3), CTG (4)	23
Lys/K	AAA (8), AAG (12)	20
Met/M	ATG (22)	22
Phe/F	TTT, TTC (7)	7
Pro/P	CCT (9), CCC, CCA (6), CCG (2)	17
Sec/U	TGA (11)	11
Ser/S	TCT (4), TCC (6), TCA (2), TCG (1), AGT, AGC (8)	21
Thr/T	ACT (7), ACC (1), ACA (3), ACG (2)	13
Trp/W	TGG (5)	5
Tyr/Y	TAT (1), TAC (7)	8
Val/V	GTT (5), GTC (1), GTA (3), GTG (5)	14

Finally, the focus was directed to the mapping of ribosomal RNAs using methods different than the Infernal search of RNAs present in the Rfam database, that was previously carried out. Thus, additional 19 rRNAs were predicted *ab initio* by RNAmmer (version 1.2; Lagesen et al. 2007), with default parameters. This program uses probabilistic models built over known ribosomal RNAs present in the European ribosomal database project. rRNAs found by this method include seventeen 8S, one 18S and one 28S rRNA (Table 31).

Table 30. tRNA Genes including introns in the *Physarum* Genome. This list includes 5 pseudogenes, corresponding to the TGC (Ala; 3tRNA pseudogenes), TCG (Arg, 1 tRNA) and TCT (Arg, 1 tRNA) anticodons.

Amino acid	**Codon**	**Anticodon**	**Intron genes**
Ala	GCA	TGC	4
Gly	GGC	GCC	1
Gly	GGA	TCC	3
Arg	CGA	TCG	1
Arg	AGA	TCT	2
Leu	CTG	CAG	1
Lys	AAG	CTT	1
Lys	AAA	TTT	1
Gln	CAA	TTG	2
Ile	ATA	TAT	3
Tyr	TAC	GTA	7
Sec	TGA	TCA	4

Furthermore, 19 previously characterized *Physarum* ribosomal RNA sequences present in GenBank were also used for similarity searches. In this manner, these sequences were mapped to 893 positions, via BLASTN alignment to the unmasked genome sequence, with an e-value of 1E-5 (version 2.2.27+; Camacho et al. 2009). To minimize the redundancy between the noncoding RNA genes predicted with different methods and programs, overlaps between the positions in the genome (also known as "annotated genomic intervals") were identified with the intersect tool of the bedtools program (version 2.17.0; Quinlan and Hall 2010). Upon filtering of overlapping ribosomal and transfer RNAs intervals, the final set consisted of 928 rRNA (873 from BLASTN, 19 from RNAmmer and 36 from Infernal) and 347 tRNA (96 from tRNAscan and 251 from Infernal) annotations. A summary of these noncoding annotations is displayed in the Table 32.

Table 31. rRNA genes identified using RNAMMER (Lagesen et al., 2007). The positions in the genomic scaffolds (start and end, in base pairs) are indicated for each predicted rRNA molecule.

rRNA type	Scaffold ID	Start	End
28S rRNA	Scaffold2079	93	6,644
8S rRNA	Scaffold8822	8	122
8S rRNA	Scaffold34229	583	697
8S rRNA	Scaffold143	110,478	110,586
8S rRNA	Scaffold38711	242	356
8S rRNA	Scaffold91028	7	121
8S rRNA	Scaffold108587	13	127
8S rRNA	Scaffold58903	184	298
8S rRNA	Scaffold19812	149	263
8S rRNA	Scaffold54413	102	216
8S rRNA	Scaffold42285	354	468
8S rRNA	Scaffold958	9,791	9,905
8S rRNA	Scaffold14262	1,754	1,868
8S rRNA	Scaffold164	201,223	201,337
8S rRNA	Scaffold93590	36	150
8S rRNA	Scaffold389	81,909	82,023
8S rRNA	Scaffold8558	1,125	1,241
8S rRNA	Scaffold13630	73	185
18S rRNA	Scaffold2079	6,902	9,103

Table 32. Summary of noncoding RNA predictions.

Program	Molecule	Total	Unique
BLASTN	rRNA	893	873
RNAmmer	rRNA	19	19
Rfam	rRNA	36	36
tRNAscan-SE	tRNA	325	96
Rfam	tRNA	251	251

Mapping RNA-seq short reads to the Physarum genome

As covered in previous chapters, RNA-seq offers the capacity of finding new genes, due to its ability to reconstruct transcripts from short cDNA fragments, even being able to reveal alternative splice isoforms and context-specific transcripts. For these reasons, it was decided to include the identification of unannotated transcripts using mapped reads from RNA-seq experiments. To this end, *Physarum* macroplasmodia from the white strain (LU897 × LU898 cross; Table 3) were cultured and collected through several points of the sporulation cycle, and their RNAs were then sequenced in two batches (LULU1 and LULU2; Table 33), using the Illumina platform (see Methods; Bentley et al. 2008; Nagalakshmi et al. 2008). RNA isolation, cDNA synthesis and cDNA library preparations were carried out by vertis Biotechnologie (Freising-Weihenstephan, Germany). A summary of these RNA sequencing outputs can be seen in the Table 33.

Table 33. RNA samples and sequencings from the white strain batches.

Sample Group	**RNA sample**	**Time Point**	**PCR cycles**	**Sample Barcode**	**Conc. (ng/ul)**	**Amount (ug)**	**Output Reads**
LULU1	dS72	0	23	AGAC	1097	88	5,915,413
LULU1	dS10	2	24	TCCA	476	38	4,273,727
LULU1	dS87	6	23	GAGT	133	11	5,678,394
LULU2	dS54	0	13	CGATGT	638	44,0	15,242,846
LULU2	dS16	3.5	12	ATCACG	359	29,0	20,000,414
LULU2	dS37	8	14	TTAGGC	256	19,0	40,986,624
LULU2	dS101	10.5	13	TGACCA	212	14,0	23,138,471

In addition, the RNA-seq output from the single-cell experiment (strain WT31; European Nucleotide Archive, accession ERP001220; Chapter 4) was also included in the following analyses. All sequencing outputs were then decoded if necessary, and trimmed for quality using the FASTX Toolkit, using an arbitrary minimum Phred score of 33, which is equivalent to 99.94988% of base call accuracy (version 0.0.13; Gordon 2008). The RNA-seq datasets were then processed separately, according to their respective sequencing experiment, and following a standard procedure for short-read mapping (Trapnell et al. 2012).

First, the unmasked *Physarum* genome scaffolds were prepared as a target database, and then each RNA-seq output was mapped to the genome using the Bowtie aligner (version 0.12.7; Langmead et al. 2009). In this way candidate exons were obtained, with their potential splice junctions identified with TopHat (version 1.4.0; Trapnell et al. 2009). A summary of the mapping statistics can be found in the Table 34.

Afterwards, the reconstruction of candidate transcript models was carried out with Cufflinks (version 1.3.0; Roberts et al. 2011), with default settings, from mapped reads and splice sites predicted by Bowtie and TopHat. Statistics about mapped reads and exon/intron structures were estimated with samtools (version 0.1.7; Li et al. 2009) and eval (Keibler and Brent 2003), respectively. In this manner, a range of 25 to 82 thousand genes was obtained, corresponding to over 26 – 92 thousand transcripts. Later, the cufflinks2gff3 tool from MAKER2 (Holt and Yandell 2011) was employed to filter these mappings, reducing the transcript range to 8 to 51 thousand protein-coding genes (Table 35). Finally, with the help of the bedtools package (version 2.17.0; Quinlan and Hall 2010), the number of predicted transcripts shared between the different RNA-seq outputs was assessed. The bedtools program achieves this by comparing all the genomic intervals where the transcript are located, eliminating redundancies between overlapping genomic positions (Quinlan and Hall 2010). In this manner 30,283 transcript intervals were found shared between the two white strain RNA-seq outputs, and 39,539 intervals shared by all three Illumina sequencing groups (Eilbeck et al. 2005).

Table 34. Summary of Illumina RNA sequencing mappings. Sample groups correspond to WT31 (European Nucleotide Archive, accession ERP001220; see Chapter 4); LULU1 (Sequence Read Archive, accession SRP009381); and LULU2 (not submitted to databases).

Sample group	**RNA sample**	**Database accession**	**Total reads**	**Reads used for mapping**	**Mapped reads**	**Percentage mapped**
WT31	422	ERS09485	19,941,711	19,930,198	8,412,392	42.21
WT31	424	ERS09485	19,499,417	19,489,244	8,421,530	43.21
WT31	431	ERS09485	18,279,213	18,269,297	6,610,504	36.18
WT31	432	ERS09485	19,346,090	19,334,649	7,777,779	40.23
LULU1	dS10	SRX10602	4,273,727	4,268,022	2,808,910	65.81
LULU1	dS72	SRX10602	5,915,413	5,907,188	3,994,729	67.62
LULU1	dS87	SRX10602	5,678,394	5,669,016	3,639,024	64.19
LULU2	dS101	---NA---	23,138,471	22,995,174	10,289,027	44.74
LULU2	dS16	---NA---	20,000,414	19,896,573	8,923,108	44.85
LULU2	dS37	---NA---	40,986,624	40,763,966	18,976,937	46.55
LULU2	dS54	---NA---	15,242,846	15,147,896	6,630,770	43.77

Table 35. Transcripts identified by mapping of RNA-seq short reads. This search was done with the Cufflinks and TopHat programs (Langmead, Trapnell, Pop & S. L. Salzberg 2009; Trapnell et al. 2009; Roberts et al. 2011). The protein-coding genes and transcript statistics were obtained using eval (Keibler and Brent, 2003), except for those that were passed to MAKER2, which were analyzed with SOBAcl (Eilbeck et al., 2005). Differences in the number of transcripts and their total lengths are likely linked to differences on the RNA-seq dataset sizes (Table 34). Default settings were employed in all cases.

	Sample group		
Sequencing Batch	WT31	LULU1	LULU2
Cufflinks Gene Count	68,872	25,737	82,584
Total Transcripts	73,836	26,554	92,109
Transcript Average Length	952.87	322.48	1,378.54
Transcript Total Length	70,356,304	8,563,115	126,975,784
Transcripts passed to MAKER2	32,298	8,939	51,763

Clustering cDNAs for EST mapping against the Physarum genome

In order to create summaries of large datasets, clustering approaches are typically applied to enlist commonly occurring sequence signatures (Hawkins et al. 2010). Here, to avoid redundancies in the cDNA reference dataset, all *Physarum* EST sequences (Glöckner et al. 2008; Watkins and Gray 2008), were clustered together with the obtained 454 sequencing output, via the UCLUST algorithm from USEARCH (version 5.2.32; Edgar 2010). An identity threshold of 100% was used for this clustering. This produced 22,632 clusters that were later combined with the CAP3 assembler (version date: 12/21/07; Huang and Madan 1999). The final non-redundant cDNA set consisted of 17,931 sequences, with 1,797 contigs and 16,134 EST singlets. This cDNA dataset was included in the next step (gene modeling) and during the estimation for completeness.

Inference of the Protein-coding gene models

The protein- coding gene models were predicted with the annotation program MAKER2 (Holt and Yandell 2011). This is an automated pipeline that aligns EST and protein data using several tools (BLAST, exonerate; Altschul et al. 1997; Slater and Birney 2005), and it is also capable to include other types of annotations, such as RNA-seq outputs and *ab initio* gene predictions, to create consensus gene models located in the genome. By default, MAKER2 requires two types of information ("biological evidences"): *ab initio* gene predictions, and alignments of transcripts and proteins to the genome. For each locus with existing gene predictions, MAKER2 evaluates if there are evidences of gene expression (RNA-seq and protein alignments), and if there are other overlapping evidences such as ESTs, the program chooses which prediction better matches the evidences, raising the prediction to annotation, *i.e.* a novel gene model. Predictions without overlapping evidences are not incorporated into the annotations, but they are still saved for future references (Holt and Yandell 2011).

In the case of the *Physarum* genome, three types of evidences were used for the modeling: (*i*) The entire protein dataset containing all non- redundant sequences from all organisms, included in UniProt (Release 2012/08; The UniProt Consortium 2010); (*ii*) two EST sets: A *Physarum* EST databank formed by clustering all existing cDNAs with the 454 data (8.94 Mb, 17,931 sequences; Chapter 3), and a collection of ESTs from *Dictyostelium discoideum*, from dictyBase (86.44 Mb, 163,182 sequences; Gaudet et al. 2011); and (*iii*) the three groups of transcript models obtained from the mapping of short RNA-seq reads to the genome (LULU1, LULU2, and WT31 datasets; Table 35).

As recommended before (Vonk et al. 2013; Gioti et al. 2013), a total of three consecutive iterative runs of MAKER2 were carried out to produce the final gene set, all of them using UniProt proteins, the *Physarum* EST evidences, and in the absence of a trained gene predictor. A different Illumina RNA-seq evidence was included for each run (LULU1, LULU2, and WT31), the *Dictyostelium* ESTs only in the second run, and data from the masking of repeats using a *Physarum* specific-custom library solely during the first run (see *Repeat Annotation*). No protein mappings to the genome were included for the modeling, but rather only those who matched a cDNA evidence were kept, although these evidences were still analyzed and saved for future reference.

In addition, to calculate the minimum size of a genomic scaffold to be analyzed, I used as a rule of thumb an estimate of the average length of a protein- coding gene in *Physarum*. For this purpose, the regression curve from the relationship between genome and gene sizes in average genomes was employed (Yandell and Ence 2012). There are, of course, exceptions to gene size – genome correlation, but for the sake of simplicity it was assumed this rule applies to *Physarum.* Here, the gene and genome data were plotted in a logarithmic scale (Table 36 and Figure 21), obtaining the following regression curve:

$$y = 0.4138\,x + 2.5482$$

Where:

$$y = \log(gene.size); x = \log(genome.size)$$

Given that the *Physarum* genome is approximately 300 Mb (Mohberg and Rusch 1971; Glöckner et al. 2008), then:

$$\log(gene.size) = 0.4138 \log 300 + 2.5482$$

$$gene.size = 10^{3.5732} = 3{,}743.11$$

Therefore the average *Physarum* gene should be 3,743 bp long, or around 4 Kb, and this number was used as the minimum contig size that should be analyzed by MAKER2 (Table 37). Finally, the output of each iteration from MAKER2 was converted into a GFF3 gene model formatted file (Eilbeck et al. 2005), to be provided as input in the following run (Gioti et al. 2013). During the first MAKER2 run no coding sequences were predicted because the parameter *est2genome*, which enables the mapping of EST data to the genome via the exonerate spliced aligner (Slater and Birney 2005), was disabled (Table 37). In this manner the ESTs were mapped directly to the genome with blastn and their coordinates recorded in the first MAKER2 output. Later this data was passed to the second run, in which 31,429 transcripts were obtained (N50 1,102 bp; average length 827.2 bp). Finally, after the third iteration, a set of 25,649 protein- coding transcripts was established (AED score < 0.49; 5,197 with AED < 0.2), encoded in 5,422 unique scaffolds (*i.e.*, 4.73 transcripts per scaffold on average). Four scaffolds contained more than a hundred transcripts, eight hundred encoded at least ten proteins, and 3,659 scaffolds comprised at least one transcript. Moreover, 2,906 transcripts used ESTs as evidences (identified with exonerate), 22,315 come from RNA-seq sequences alone (candidate gene models from Cufflinks), and 428 possessed both EST and RNA-seq evidences (combined exonerate and Cufflinks predictions). These transcripts have an average length of 601.5 base pairs and N50 of 746 bp, and an average number of 3.36 exons per gene. The highest number of exons on a gene is 27, and 483 genes are single exonic. A summary of the gene model statistics is listed in the Table 38, and an example of the predicted gene models can be seen on the Figure 22.

Table 36. Gene and genome sizes from a representative set of species (D.Ence, personal communication; Yandell and Ence 2012). The genome sizes are given in megabase pairs (Mbp), and the average gene sizes in base pairs (bp).

Species	**Genome Size**	**Average Gene Size**
Escherichia coli	4.74	806
Saccharomyces cerevisiae	12	1,070
Schizosaccharomyces pombe	14	1,866
Caenorhabditis elegans	100	1,967
Arabidopsis thaliana	120	1,847
Volvox carteri	138	3,851
Drosophila melanogaster	169	1,841
Citrus clementina	296	2,931
Takifugu rubripes	392	5,307
Oryza sativa	430	2,705
Populus trichocarpa	500	2,278
Eucalyptus grandis	641	2,473
Gallus gallus	1,080	9,693
Danio rerio	1,400	12,138
Ornithorhynchus anatinus	1,900	9,628
Zea mays	2,070	2,746
Ailuropoda melanoleuca	2,400	13,024
Mus musculus	2,720	15,819
Homo sapiens	2,870	20,590

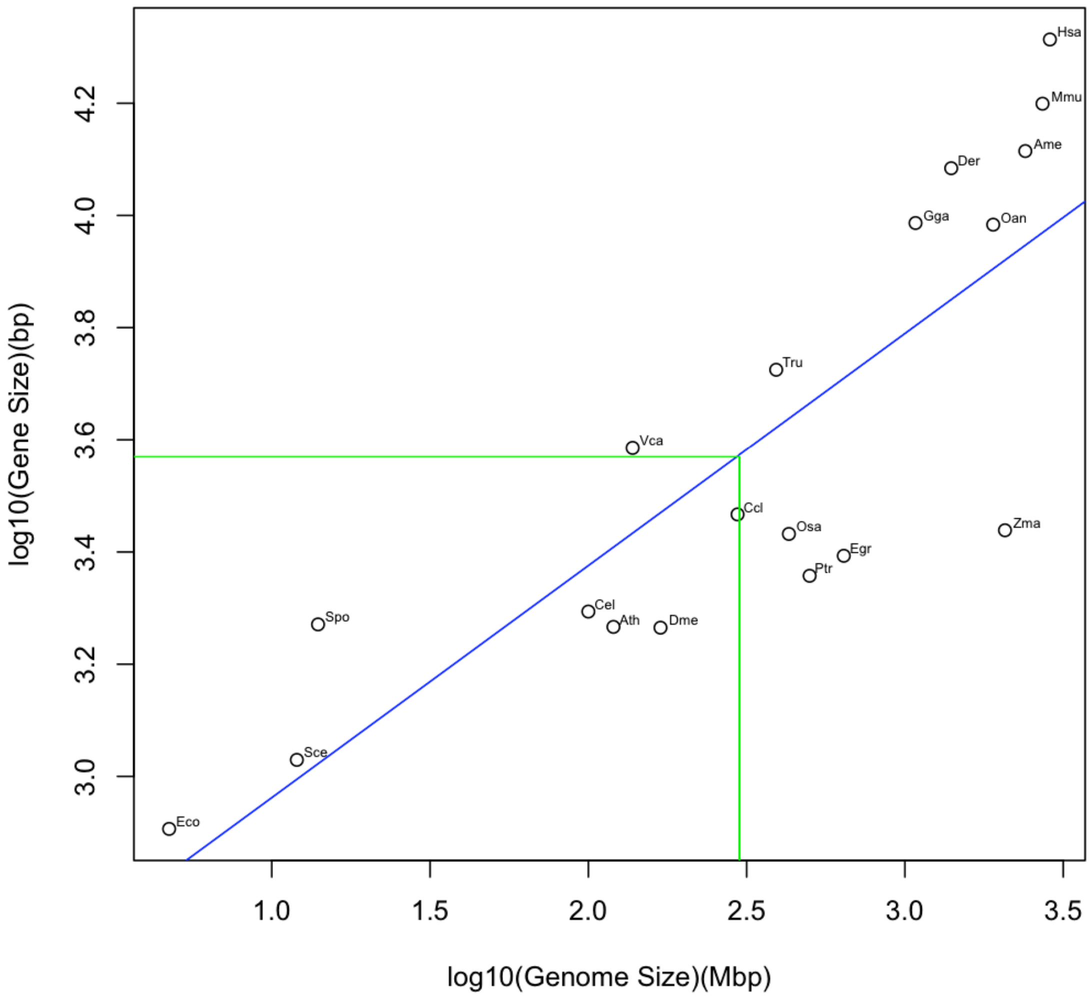


Figure 21. Relationship between the gene and genome sizes from a representative set of species. The common logarithms of the genome sizes (*x*-axis, in Mbp) were plotted against the logarithms of their corresponding average gene lengths (*y*-axis, in bp) for each species. A regression curve was obtained (blue line), and the average size for a *Physarum* gene was projected from the *x*- to the *y*- axis as a reference (green line), using the approximate size of the *Physarum* genome (Mohberg and Rusch 1971). Data was obtained from Daniel Ence (University of Utah, personal communication; Table 36), and the figure redrawn from Yandell and Ence (2012)

Table 37. Identification of protein gene models. The employed biological evidences, further parameters and outputs are listed for each MAKER2 iteration (Holt and Yandell 2011).

Evidences and input data			
Iteration	**First**	**Second**	**Third**
Input models	---	GFF3 output from the first iteration	GFF3 output from the second iteration
Primary EST evidence	*Physarum* non redundant clusters	*Physarum* non redundant clusters	*Physarum* non redundant clusters
Secondary EST evidence	---	*Dictyostelium* ESTs from dictyBase	---
RNA-seq evidence	Transcripts from mapped LULU1 reads	Transcripts from mapped LULU2 reads	Transcripts from mapped WT31 reads
Protein evidences	Uniprot	Uniprot	Uniprot
Repeat evidences	*Physarum* custom library	RepBase	RepBase
Modeling parameters			
Date complete	Nov 18 2012	Jan 24 2013	Feb 27 2013
Running time	~15 days	~34 days	~26 days
Parallel CPUs	6	6	8
Minimum contig size (bp)	1	9,999	4,000
Mapping EST data to genome	no	yes	yes
Identify single exonic genes	no	no	yes

Table 38. Features of the predicted reference gene models. These statistics correspond to the three GFF3- formatted outputs from the MAKER2 runs, obtained with the SOBAcl program (Eilbeck et al. 2005; Holt and Yandell 2011; Moore et al. 2010).

Gene models			
Iteration	**First**	**Second**	**Third**
mRNAs	none	31,429	25,649
Genes	0	28,379	24,615
Exons	0	131,097	84,152
Coding sequences	0	125,363	75,448
Matching evidences			
Expressed	56,822	93,583	92,537
Protein	446,897	265,209	512,450
Translated	0	5,854	0
Transcript statistics			
Total bases	---	25,999,231	15,426,914
Minimum size	---	22	26
Maximum size	---	9,717	7,016
Average size	---	827.2	601.5
N50 length	---	12,999,712	7,713,703
N50 value	---	1,102	746

Annotation of the Gene Models

In order to characterize the gene functions in the *Physarum* genome, first the set of transcripts and proteins (corresponding to the gene models identified with MAKER2 in the previous step), was extracted using the fasta_merge utility included in this pipeline (Holt and Yandell 2011). Then a *blastp* similarity search (version 2.2.27; parameters: e-value 1E-3; maximum target sequences = 20) of the encoded proteins was performed against the UniProt database (Altschul et al. 1997; The UniProt Consortium 2010). Simultaneously, motifs and domains present in the InterPro database were obtained in these protein sequences, with the InterProScan program, including the gene ontology (GO) annotations for each domain found (Hunter et al. 2009; Quevillon et al. 2005; Ashburner et al. 2000).

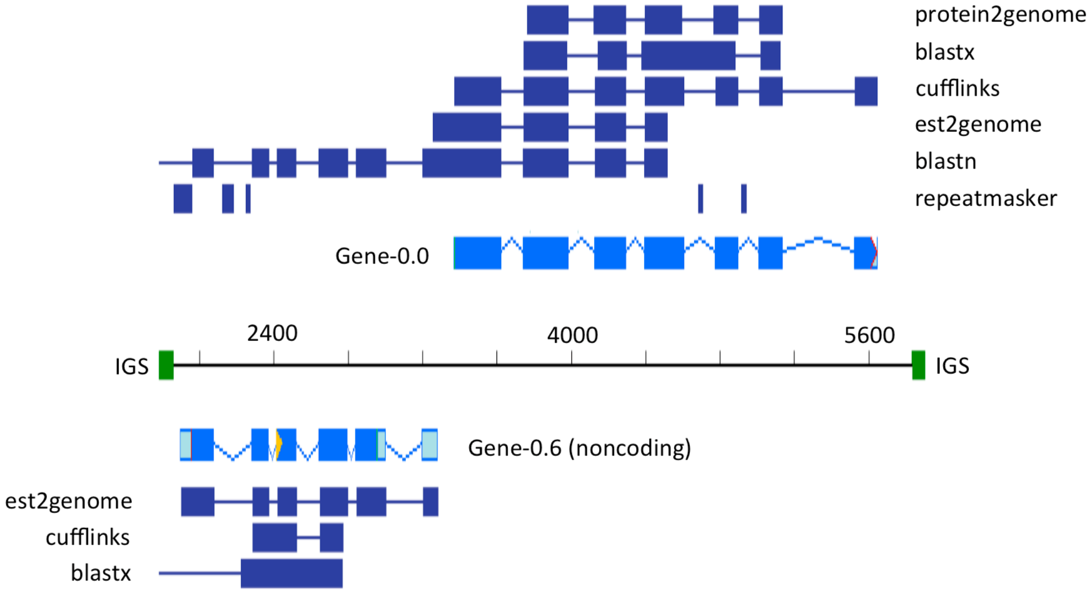


Figure 22. Example of evidences forming a gene model. A plot of the several mapped evidences against a region of the genomic Scaffold1 is presented. This region corresponds to an interval between approximately 2 – 6 Kb, and separated by two predicted intergenic spacers (IGS). A gene was identified in the forward strand (Gene-0.0; above), encoding an homolog of the U4/U6 small nuclear ribonucleoprotein PRP31. In this case, the EST and RNA-seq evidences (found with blastn, est2genome and cufflinks), are in agreement with the mapped protein, identified via the blastx and protein2genome programs inside the MAKER2 pipeline. Conversely, in the opposite strand, a gene model was predicted as noncoding (Gene-0.6), given that it does not possess complete overlapping physical evidences (RNA-seq, EST or protein alignments). In all cases, the default gene naming convention of the MAKER2 pipeline (such as Gene-0.0 and Gene-0.6 in this example) was employed.

Then the annotations from the outputs from both the blastp and InterProScan searches were integrated, by obtaining the gene ontology information from the UniProt entries, and adding those from the Interpro database, with the Blast2GO pipeline (version 2.5; Götz et al. 2008; Conesa et al. 2005). The Blast2GO annotation database employed was the version *b2g_aug12*, accessed online at http://publicdb.blast2go.com. All these processes were executed through command line- batch protocols. In this manner, 4,915 sequences were linked to UniProt homologs, 5,752 were associated to gene ontology (GO) annotations, and 15,914 contained InterPro domains, including 7,080 sequences (27.60%) that possessed PFAM domains (Finn et al. 2008). The UniProt protein homologs pertained to 3,549 unique annotation descriptions and 483 species, with the most common gene description found was the DNA ligase (38 orthologs; Figure 23), and *Dictyostelium discoideum* is the most represented species, together with a considerable number of other animal and fungal species (Figure 24). Most encoded proteins were associated to binding, kinase and other gene ontologies related to the interaction with nucleic acids (Figure 25). Moreover, the novel proteins were also searched for similarity against the KEGG orthologs, in order to study the representation of enzymes and metabolic pathways in the *Physarum* genome (Kanehisa et al. 2008).

To this end, the bidirectional best hit mode and the GENES dataset from the KAAS annotation server were employed (version 1.67; Moriya et al. 2007). Here, 2,066 transcripts with KEGG orthologs were found (1,779 unique); the most common of these entailed transferases (*AKR1*, *NatA*, *mhkB*, *omt5*, *ppkA*) and dehydrogenases (*CHDH*, *ptpB*, *PP2C*; Table 39). Finally, these KEGG orthologs were mapped to the KEGG Atlas of metabolic pathways, using the version 2 of ipath tool (Okuda et al. 2008b; Yamada et al. 2011; Letunic et al. 2008). In this manner, 741 KEGG orthologs (KOs) were linked to the metabolic primary map, 439 to the regulatory and 202 to the biosynthetic pathway, and, as shown in Figure 26, most KOs belong to the main macromolecular pathways (carbohydrate, lipid, amino acids, nucleotide and energy metabolism).

Table 39. Top 10 most frequent KEGG orthologs.

Accession	**Annotation**	**Transcripts**
K06867	Palmitoyl transferase, *AKR1*	8
K00108	Choline dehydrogenase, *CHDH*	6
K00670	N(alpha)-acetyltransferase, *NatA*	5
K00924	Phosphotransferase *mhkB*	5
K01104	Protein-tyrosine phosphatase *ptpB*	5
K01802	Peptidylprolyl isomerase *impA*	5
K07126	sel-1 suppressor of lin-12-like 2, *SEL1L2*	5
K08884	Serine/threonine protein kinase *ppkA*	5
K00599	O-methyltransferase family 3 protein *omt5*	4
K01090	Protein phosphatase 2C-related protein *PP2C*	4

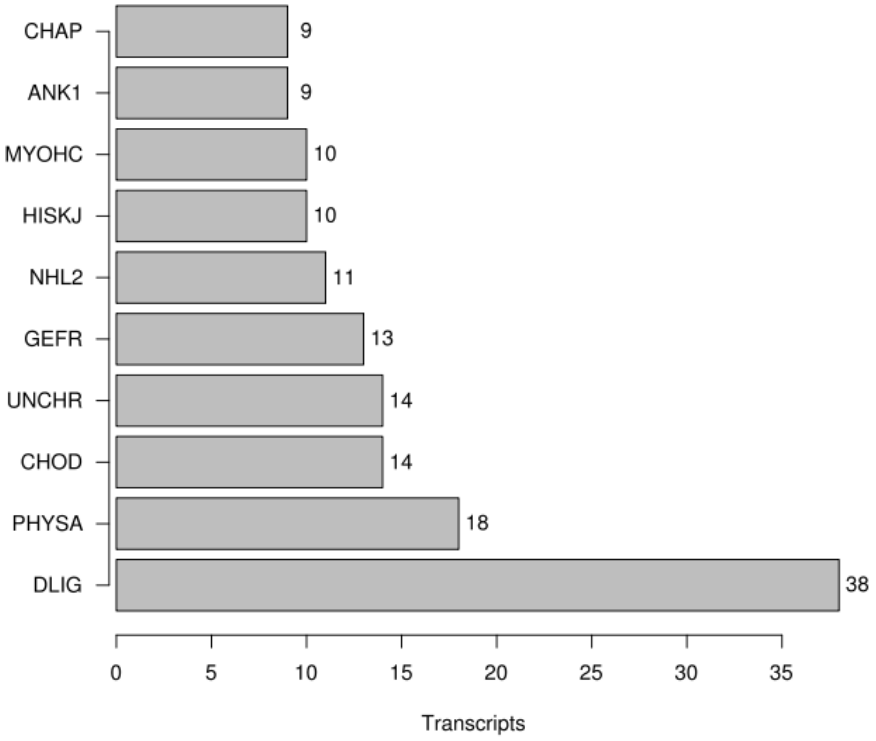


Figure 23. Top 10 most frequent gene annotation descriptions. The number of most common unique annotation descriptions was plotted against their frequency in transcripts. Displayed annotations correspond to the following genes: DNA ligase (*DLIG*), physarolisin (*PHYSA*), choline dehydrogenase (*CHOD*), an uncharacterized protein (*UNCHR*), the guanine exchange factor for RAC 30 (*GEFR*), the NHL repeat-containing protein 2 (*NHL2*), the hybrid signal transduction histidine kinase J (*HISKJ*), myosin-I heavy chain (*MYOHC*), ankyrin-1 (*ANK1*), and a chaperone protein (*CHAP*). Annotation data was analyzed and plotted using the R statistical environment (R Core Team 2013).

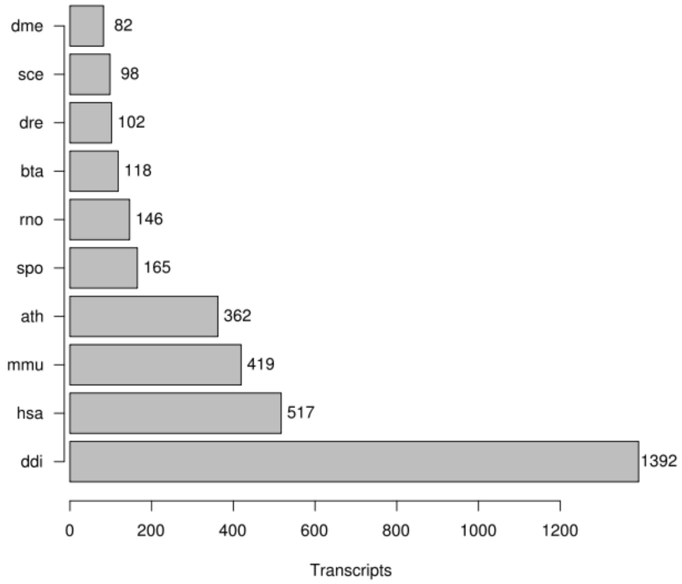


Figure 24. Top 10 most represented species in the orthologs. Plotted species correspond to the cellular slime mold *Dictyostelium discoideum* (*ddi*), human (*hsa*), mouse (*mmu*), the thale cress *Arabidopsis thaliana* (*ath*), the fission yeast *Schizosaccharomyces pombe* (*spo*), rat (*rno*), cattle (*bta*), zebrafish (*dre*), the budding yeast (*sce*), and the fruitfly (*dme*). Statistics were obtained in a similar manner as in the Figure 23.

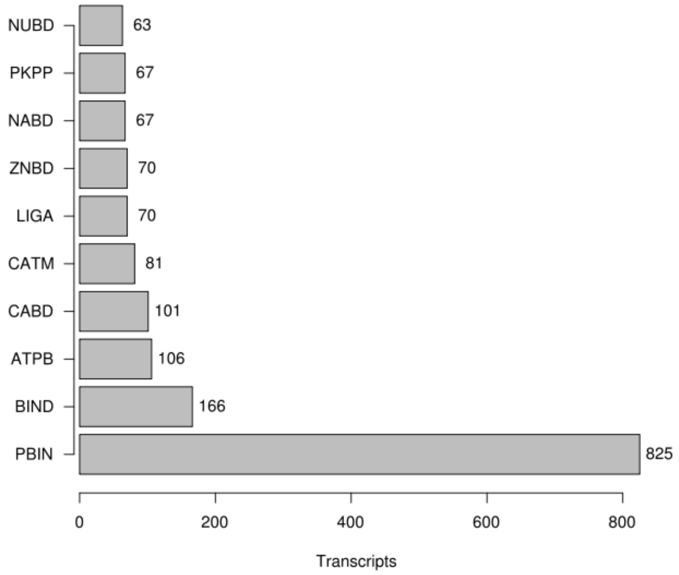


Figure 25. Top 10 most frequent gene ontology associations. Ontologies plotted belong to the following descriptions: protein binding (*PBIN*); binding (*BIND*); ATP binding (*ATPB*); calcium ion binding (*CABD*); catalytic activity, metabolic process (*CATM*); DNA ligase activity (*LIGA*); zinc ion binding (*ZNBD*); nucleotide binding (*NABD*); protein kinase activity (*PKPP*); and nucleotide binding (*NUBD*). The plot follows the same conventions as in Figures 23 and 24.

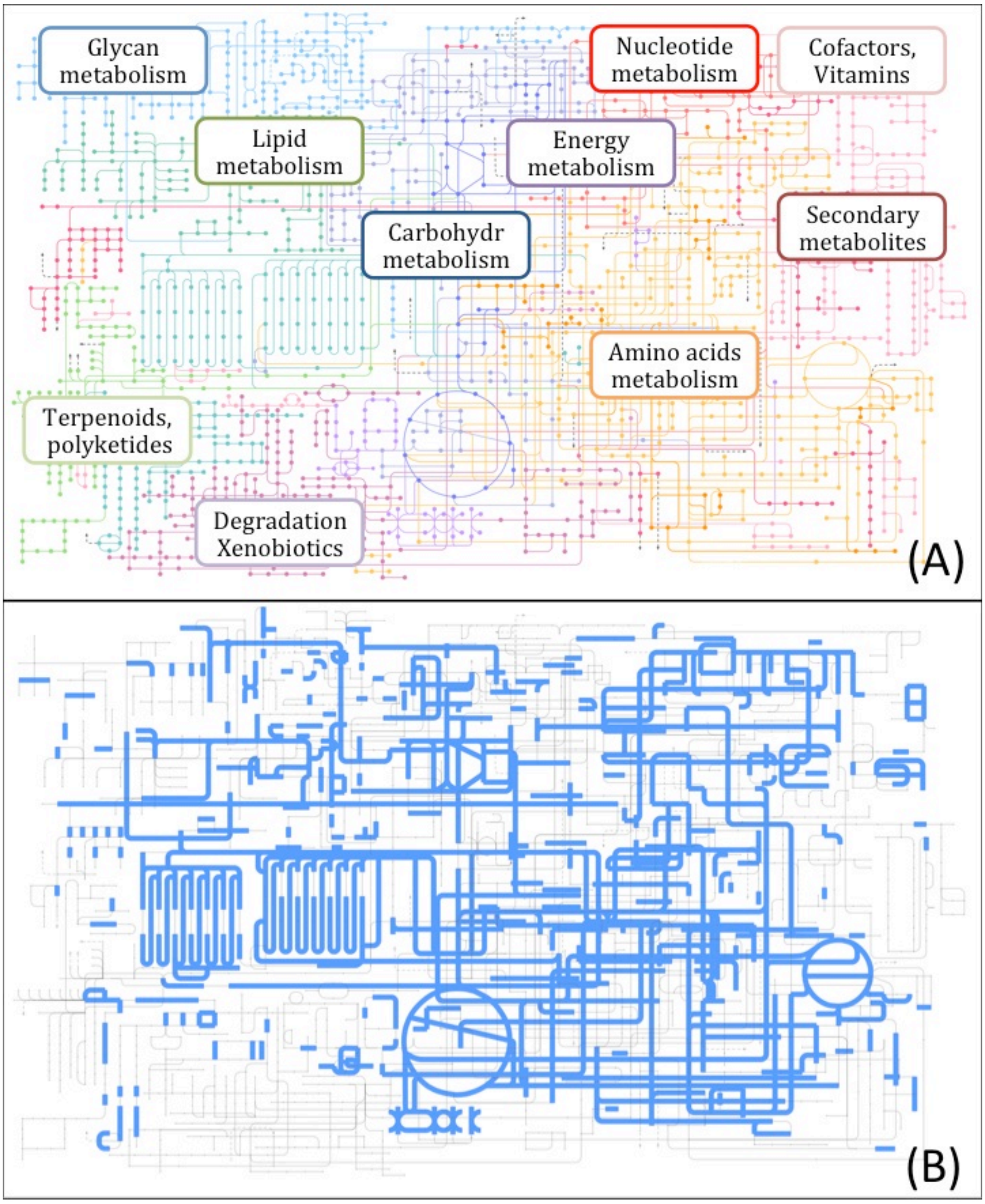


Figure 26. The reference metabolic map of *Physarum*. The transcripts were assigned to KEGG orthologs through similarity search, and these were mapped to pathways with the ipath tool. Above (*A*), the generic map of the whole metabolism is displayed, with reference colors to each pathway. Below (*B*), all predicted metabolic reactions in the slime mold, according to the reference transcriptome obtained in this thesis work.

Network Analysis

First, all UniProt IDs were extracted from the obtained annotations under the R environment, and these were mapped to their extended annotations, stored in the UniProt database. Then the gene ontologies (GO) were analyzed, to select those that are annotated for "*cell differentiation*" (GO: 0030154), using the GO Retriever and GO Slim Viewer from AgBase version 2 (McCarthy et al. 2006a). This produced a dataset of 432 proteins (277 unique entries). A treemap of the ontology terms was then plotted for these 432 proteins, in order to summarize the annotations, with a modified R script from REVIGO (Figure 27; Supek et al. 2011). From these results, two subsets from this differentiation dataset were chose: one annotated with the GO:0009790 ("*embryo development*"; 40 unique entries), and another with the GO:0007165 ("*signal transduction*"; 111 entries) ontologies. This was done to simplify the network reconstruction, and because these annotations indicate that a given protein is more likely to be actively involved in the differentiation process. The remaining entries that were not annotated from any of these two ontologies were kept for later analyses (150 unique entries). Then these three subsets were loaded into Cytoscape, a biological network visualization and analysis software (Shannon et al. 2003; Smoot et al. 2011), and the conceptual interactions between proteins of each subset predicted with the Cytoprophet plugin (Morcos et al. 2008). This tool draws potential networks based on the domain composition and experimental assays from the input proteins, gathered from databases of protein interactions through their UniProt accessions. Here the default mode of Cytoprophet was used, *i.e.*, the maximum likelihood estimation (MLE) algorithm, and protein – protein interactions (PPI).

At this point, the predicted networks were composed of large numbers of edges: There were 2,047 interactions predicted for the proteins annotated for signal transduction, 171 for those with the embryo development ontology, and 1,948 between those annotated for cell differentiation, but not included in the two former ontologies. Therefore the most closely connected regions in these large Cytoprophet- predicted networks were searched, using the MCODE clustering algorithm (Bader and Hogue 2003). MCODE is an automated method to encounter all the highly interconnected subgraphs as protein complexes in large

PPI networks. This procedure is generally recommended in standard network analysis protocols in order to simplify even further these interactions (Cline et al. 2007). MCODE gives each predicted complex a score, equivalent to the network density multiplied by the number of nodes; where the density of a graph is the number of edges divided by the maximum theoretical number of edges. In this manner, the existence of one protein complex within the network of proteins annotated for the embryo development ontology (EDC, Table 41), four complexes for those with the signal transduction GO (Tables 43 and 44), and three for those annotated with cell differentiation alone were inferred (Tables 46 – 48). These protein complex predictions are summarized on Table 40, and displayed in Figures 28 – 30.

Table 40. Summary of the conceptual protein complexes linked to the Cell Differentiation ontology. A list of complexes predicted by the MCODE tool, inside the networks previously created with Cytoprophet, is displayed (Bader and Hogue 2003; Morcos et al. 2008; Cline et al. 2007). These complexes were classified according to a given ontology (embryo development or signal transduction), and those who did not belong to those two ontologies (Cell Differentiation ontology alone). Scores are standard MCODE scores. Nodes and edges represent proteins and interactions, respectively, and the Node IDs are the UniProt entries associated to a given complex.

Embryo Development (GO:0009790)			
Complex	*Score*	*Nodes*	*Edges*
ED1	5,665	15	85
Signal Transduction (GO:0007165)			
Complex	*Score*	*Nodes*	*Edges*
ST1	24,868	53	1,318
ST2	2,875	8	23
ST3	2	5	10
ST4	1	3	3
Cell Differentiation (GO: 0030154)			
Complex	*Score*	*Nodes*	*Edges*
CDN1	20,878	49	1,023
CDN2	2,583	12	31
CDN3	1	3	3

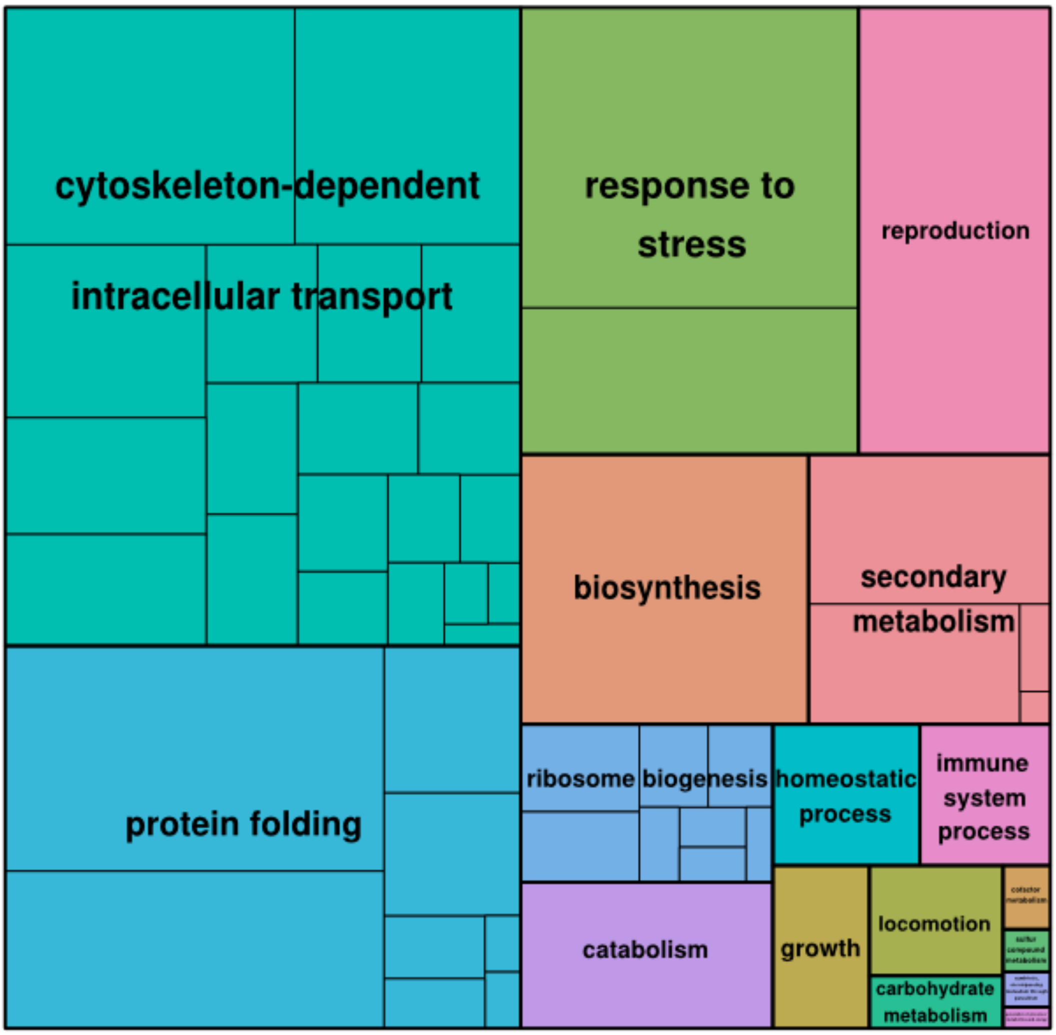


Figure 27. Summary of ontologies for the genes associated to cell differentiation. In these genes, 17 ontologies were identified as the most frequent: cytoskeleton-dependent intracellular transport; protein folding; response to stress; reproduction; biosynthesis; secondary metabolism; ribosome biogenesis; catabolism; homeostatic process; immune system process; growth; locomotion; carbohydrate metabolism; cofactor metabolism; sulfur compound metabolism; symbiosis; generation of precursor metabolites and energy. Each color represents a group of ontologies that share a parent (e.g. "response to stress" contains two ontologies: "signal transduction" and "response to stress"). Each lower level of ontology is indicated as a separate box, and the size of each box is proportional to the frequency of these ontologies in the analyzed gene dataset. The analysis was performed with REVIGO (Supek et al. 2011).

Those subnetworks predicted by Cytoprophet but without MCODE subcomplexes, were saved and analyzed separately: One subnetwork for signal transduction (Table 45), four for embryo development (Table 42) and three for cell differentiation (Tables 49 and 50) respectively. No analyses of the expression statuses (*i.e.*, differential expression) of the transcripts included in these networks were carried out in this thesis work. Later, in order to distinguish the processes and functions for each complex and network, the ontology annotations from these associations were then compared, through the WEGO online tool (Ye et al. 2006). Here, the results show that the largest complex (CDN1) is annotated for the following ontologies: membrane enclosed lumen, transcriptional regulator, adhesion, locomotion, and multiorganism process; while the second complex (CDN2), is associated to the auxiliary transport and enzyme regulation (Figure 31A). The complexes CDN1 and CDN3 shared most gene ontologies. On the other hand, the network CDO1 is linked to multiple ontologies: envelope, complex, and lumen (cellular component GO); electron carrier, structural molecule, transcriptional regulator, and transporter (molecular function GO); and anatomical structure formation, adhesion, death, and immune system process (biological process GO; Figure 31B). As in the case before, the largest network (CDO1) and the smallest (CDO3) shared most ontology associations. Then, regarding the analysis of the signal transduction entries, the signal transduction complex ST1 was found to be related to the transcriptional regulator, adhesion, death, rhythmic process, and viral reproduction ontologies, while the ST3 complex is exclusively annotated for the extracellular region category (Figure 32A). At the same time, the transduction network STNC1 alone entailed the auxiliary transport, transducer, transcription regulator, transporter, and immune system process ontologies; and those proteins not forming networks or complexes, that are annotated for signal transduction are exclusively linked to the electron carrier ontology (Figure 32B). Finally, it was observed that the proteins from the embryo development complex (EDC) are annotated for enzyme regulation and cell adhesion, while all other subnetworks are associated to electron carrier, molecular transducer, transcriptional regulator, growth, and rhythmic processes (Figure 33).

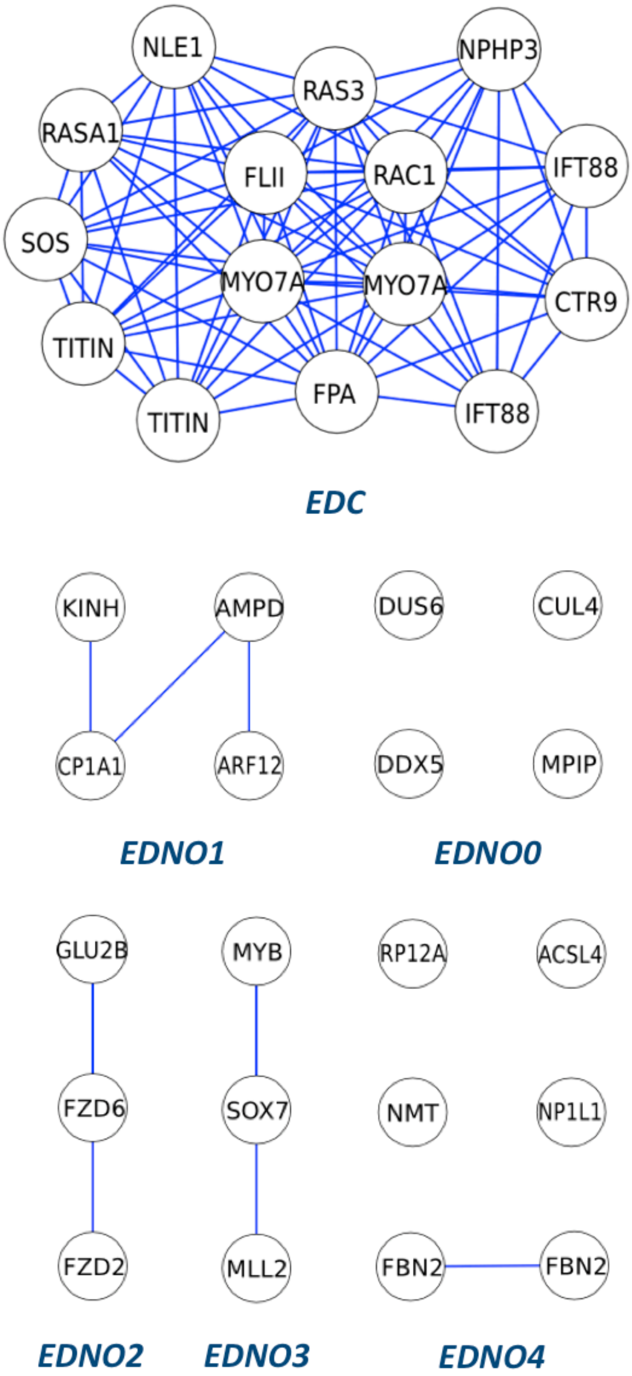


Figure 28. Complexes and subnetworks linked to the embryo development ontology. These modules were identified first by extracting proteins annotated for the cell differentiation and embryo development gene ontologies, then by predicting their interactions obtained from the bibliography, and then by locating protein complex with the MCODE tool. Annotations from each protein entry are listed in Tables 41 and 42.

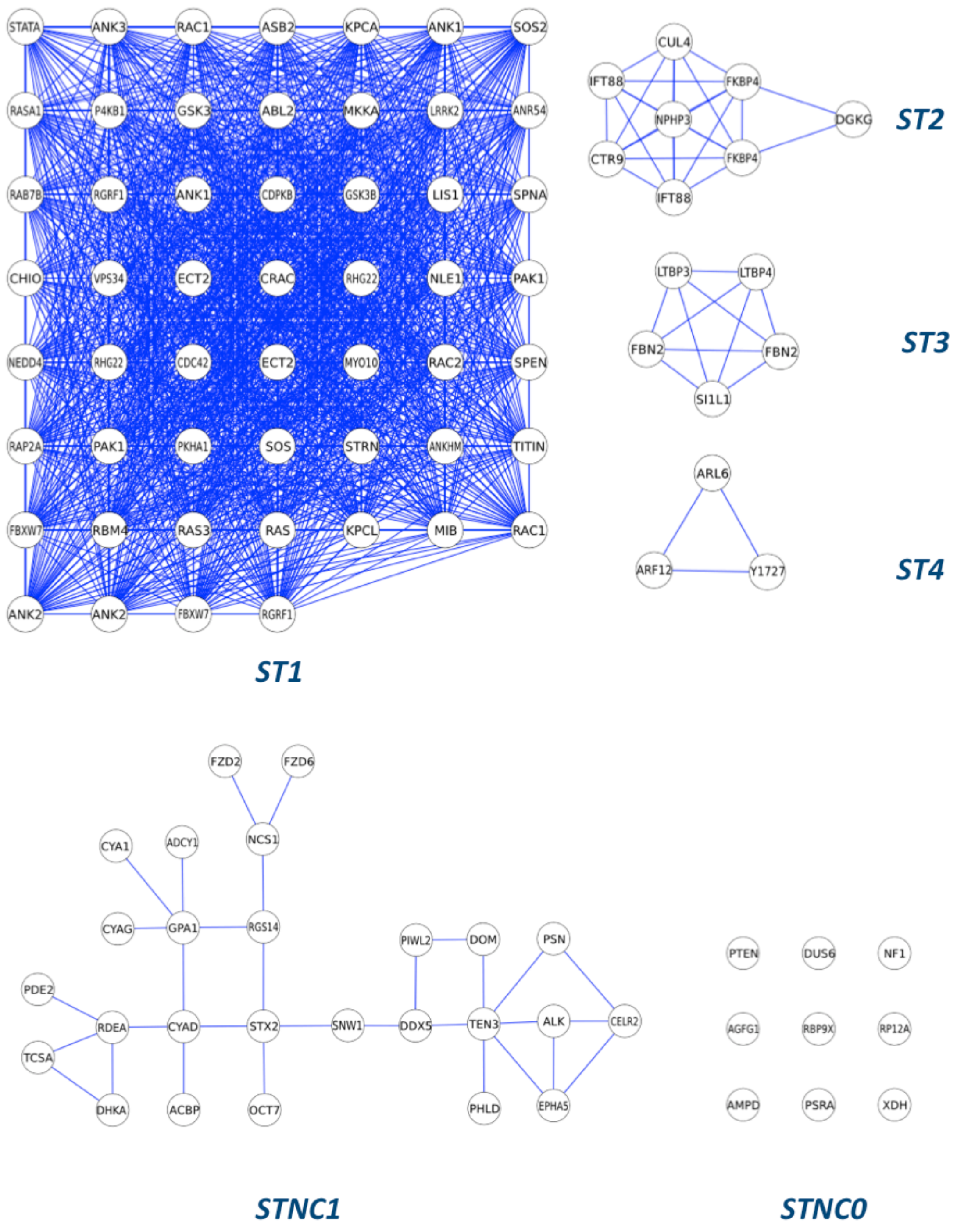
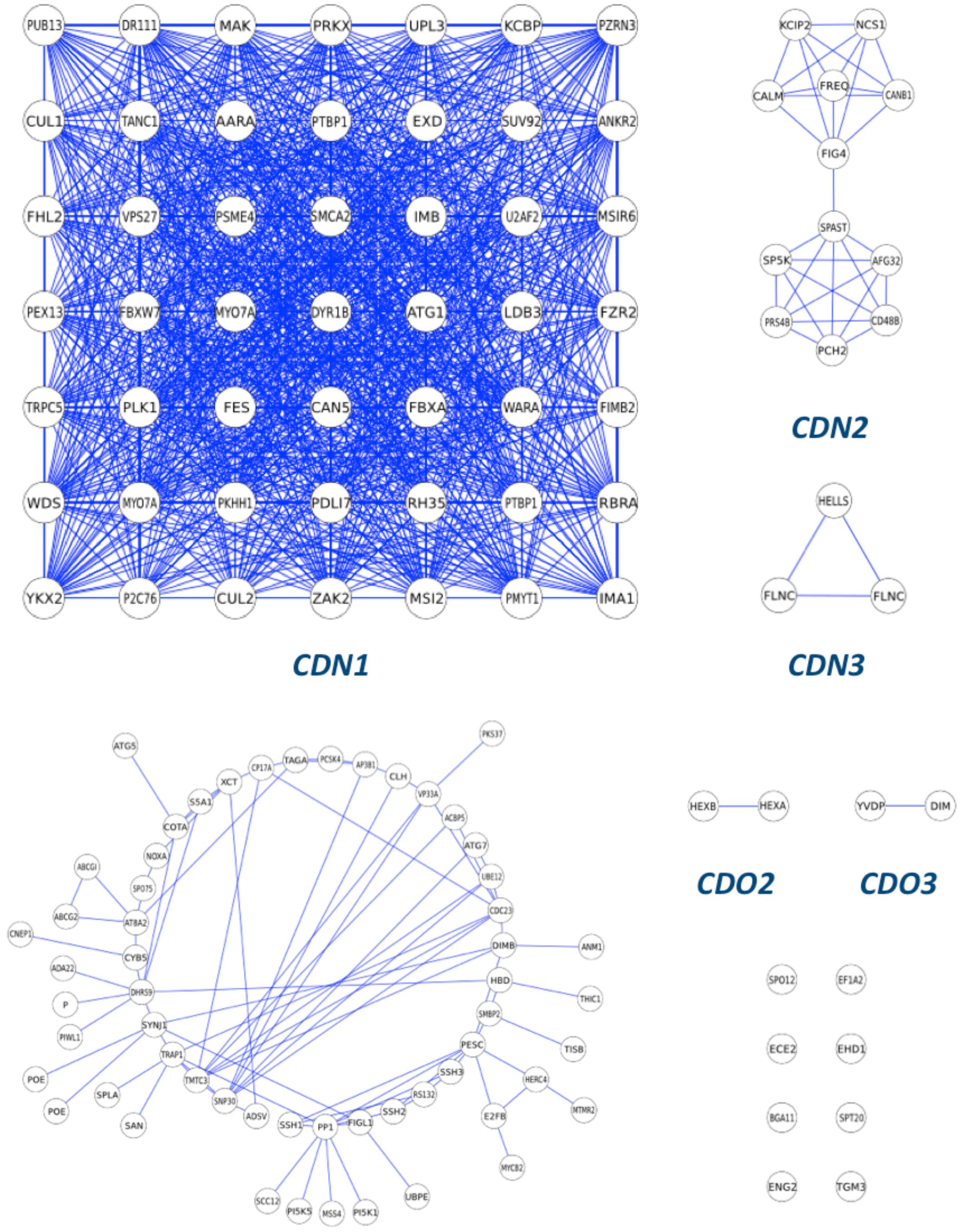


Figure 29. Complexes and subnetworks linked to the signal transduction ontology. The procedure to obtain these modules, and the annotations for each entry, follows the same convention as in the Figure 28.



CDN1 **CDN2** **CDN3**

CDO1 **CDO2** **CDO3** **CDO0**

Figure 30. Complexes and subnetworks linked to the cell differentiation ontology. The procedure to obtain these modules, and the annotations for each entry, follows the same convention as in the Figure 28.

Table 41. A protein complex annotated with the Embryo Development ontology (GO: 0009790). Entries are specified as UniProt accession numbers.

Protein	Entries	Annotation
CTR9	Q62018	RNA polymerase-associated protein CTR9 homolog
FLII	Q24020	Flightless-I
FPA	Q8LPQ9	Flowering time control protein FPA
IFT88	Q13099, Q61371	Intraflagellar transport protein 88 homolog
MYO7A	Q13402, P97479	Unconventional myosin-VIIa
NLE1	Q58D20	Notchless protein homolog 1
NPHP3	Q7TNH6	Nephrocystin-3
RAC1	Q6RUV5	Ras-related C3 botulinum toxin substrate 1
RAS3	P08645	Ras-like protein 3
RASA1	P50904	Ras GTPase-activating protein 1
SOS	P26675	Protein son of sevenless
TITIN	Q8WZ42, A2ASS6	Titin

Table 42. Protein orthologs annotated with the Embryo Development ontology (GO:0009790) that do not form protein complexes. Subnetwork names are indicated for each protein entry, and other fields follow the same convention as in Table 41. The inclusion of proteins such as kinesin and the cytochrome P450 obeys only to their presence in their annotations at UniProt.

Network	Protein	Entries	Annotation
EDNO0	ACSL4	Q9QUJ7	Long-chain-fatty-acid--CoA ligase 4
EDNO0	CUL4	Q8LGH4	Cullin-4
EDNO0	DDX5	Q61656	ATP-dependent RNA helicase DDX5
EDNO0	DUS6	Q9DBB1	Dual specificity protein phosphatase 6
EDNO0	MPIP	P20483	M-phase inducer phosphatase
EDNO0	NMT	O61613	Glycylpeptide N-tetradecanoyltransferase
EDNO0	NP1L1	Q28EB4	Nucleosome assembly protein 1-like 1
EDNO0	RP12A	Q9SGW3	26S proteasome non-ATPase regulatory subunit
EDNO1	ARF12	Q10943	ADP-ribosylation factor 1-like 2
EDNO1	AMPD	O80452	AMP deaminase
EDNO1	CP1A1	P00185	Cytochrome P450 1A1
EDNO1	KINH	P17210	Kinesin heavy chain
EDNO2	FZD2	Q08464	Frizzled-2
EDNO2	FZD6	Q8WMU5	Frizzled-6
EDNO2	GLU2B	O08795	Glucosidase 2 subunit beta
EDNO3	MLL2	Q6PDK2	Histone-lysine N-methyltransferase MLL2
EDNO3	MYB	P10242	Transcriptional activator Myb
EDNO3	SOX7	Q28GD5	Transcription factor Sox-7
EDNO4	FBN2	P35556, Q61555	Fibrillin-2

Table 43. Proteins from the Signal Transduction Complex ST1. Listed are those entries whose annotations are other than Kinases, and containing the signal transduction ontology (GO: 0007165). Fields follow the same convention as in Table 42.

Protein	**Entries**	**Annotation**
ANK1	P16157, Q02357	Ankyrin-1
ANK2	Q01484, Q8C8R3	Ankyrin-2
ANK3	Q12955	Ankyrin-3
ANKHM	Q9VCA8	Ankyrin repeat and KH domain-containing mask
ANR54	Q91WK7	Ankyrin repeat domain-containing protein 54
ASB2	Q8K0L0	Ankyrin repeat and SOCS box protein 2
CDC42	Q4R4R6	Cell division control protein 42 homolog
CHIO	Q03070	Beta-chimaerin
CRAC	P35401	Protein CRAC
ECT2	Q9H8V3, Q07139	Protein ECT2
FBXW7	Q969H0, Q8VBV4	F-box/WD repeat-containing protein 7
LIS1	Q8I0F4	Lissencephaly-1 homolog
MIB	Q9VUX2	E3 ubiquitin-protein ligase mind-bomb
MYO10	Q9HD67	Unconventional myosin-X
NEDD4	P46935	E3 ubiquitin-protein ligase NEDD4
NLE1	Q58D20	Notchless protein homolog 1
PKHA1	Q8BUL6	Pleckstrin homology domain-containing A member 1
RAB7B	Q96AH8	Ras-related protein Rab-7b
RAC1	Q03206	Ras-related protein ced-10
RAC1	Q6RUV5	Ras-related C3 botulinum toxin substrate 1
RAC2	P15153	Ras-related C3 botulinum toxin substrate 2
RAP2A	Q5R988	Ras-related protein Rap-2a
RAS3	P08645	Ras-like protein 3
RASA1	P50904	Ras GTPase-activating protein 1
RAS	P08647	Ras-like protein 1
RBM4	Q4R979	RNA-binding protein 4
RGRF1	Q13972, P28818	Ras-specific guanine nucleotide-releasing factor 1
RHG22	Q7Z5H3, Q8BL80	Rho GTPase-activating protein 22

Table 44. Kinases from the Signal Transduction Complex ST1, and members of the protein complexes ST2, ST3 and ST4. Complexes are indicated for each entry, and other fields follow the same convention as in Table 43.

Complex	**Protein**	**Entries**	**Annotation**
ST1	ABL2	P42684	Abelson tyrosine-protein kinase 2
ST1	CDPKB	Q39016	Calcium-dependent protein kinase 11
ST1	GSK3B	Q91757	Glycogen synthase kinase-3 beta
ST1	GSK3	P51136	Glycogen synthase kinase-3
ST1	KPCA	P10102	Protein kinase C alpha
ST1	KPCL	P24723	Protein kinase C eta
ST1	LRRK2	Q5S006	Leucine-rich repeat Ser/Thr kinase 2
ST1	MKKA	Q54R82	MAPK/ERK kinase 1
ST1	P4KB1	Q9FMJ0	Phosphatidylinositol 4-kinase beta 1
ST1	PAK1	Q13153, O88643	Serine/threonine-protein kinase PAK 1
ST1	SOS2	Q07890	Son of sevenless homolog 2
ST1	SOS	P26675	Protein son of sevenless
ST1	SPEN	Q8SX83	Protein split ends
ST1	SPNA	O15743	Ser/Thr phosphatase spalten
ST1	STATA	O00910	Signal transducer, activator of transcription A
ST1	STRN	O43815	Striatin
ST1	TITIN	A2ASS6	Titin
ST1	VPS34	P50520	Phosphatidylinositol 3-kinase vps34
ST2	CTR9	Q62018	RNA polymerase-associated CTR9 homolog
ST2	CUL4	Q8LGH4	Cullin-4
ST2	DGKG	P49620	Diacylglycerol kinase gamma
ST2	FKBP4	Q02790, P30416	Peptidyl-prolyl cis-trans isomerase FKBP4
ST2	IFT88	Q13099, Q61371	Intraflagellar transport protein 88 homolog
ST2	NPHP3	Q7TNH6	Nephrocystin-3
ST3	FBN2	P35556, Q61555	Fibrillin-2
ST3	LTBP3	Q9NS15	Latent-transforming growth factor β-binding 3
ST3	LTBP4	Q8N2S1	Latent-transforming growth factor β-binding 4
ST3	SI1L1	O43166	Signal-induced proliferation-associated 1-like 1
ST4	ARF12	Q10943	ADP-ribosylation factor 1-like 2
ST4	ARL6	O88848	ADP-ribosylation factor-like protein 6
ST4	Y1727	Q9VYY9	TBC1 domain family member CG11727

Table 45. Proteins annotated for the signal transduction ontology that do not form complexes. These entries do not form any complexes. Subnetworks are indicated for each entry, and other fields follow the same convention as in Table 43.

Network	**Protein**	**Entries**	**Annotation**
STNC1	ACBP	Q5FXM5	Acyl-CoA-binding protein
STNC1	ADCY1	O88444	Adenylate cyclase type 1
STNC1	ALK	P97793	ALK tyrosine kinase receptor
STNC1	CELR2	Q9HCU4	Cadherin EGF LAG seven-pass G-type receptor 2
STNC1	CYA1	P32870	Ca(2+)/calmodulin-responsive adenylate cyclase
STNC1	CYAD	Q55F68	Adenylate cyclase, terminal-differentiation specific
STNC1	CYAG	Q03101	Adenylate cyclase, germination specific
STNC1	DDX5	Q61656	Probable ATP-dependent RNA helicase DDX5
STNC1	DHKA	Q54U87	Hybrid signal transduction histidine kinase A
STNC1	DOM	Q9NDJ2	Helicase domino
STNC1	EPHA5	Q60629	Ephrin type-A receptor 5
STNC1	FZD2	Q08464	Frizzled-2
STNC1	FZD6	Q8WMU5	Frizzled-6
STNC1	GPA1	P16894	Guanine nucleotide-binding protein alpha-1 subunit
STNC1	NCS1	P62168	Neuronal calcium sensor 1
STNC1	OCT7	Q940M4	Organic cation/carnitine transporter 7
STNC1	PDE2	Q23917	3',5'-cyclic-nucleotide phosphodiesterase regA
STNC1	PHLD	Q8R2H5	Phosphatidylinositol-glycan-specific phospholipase D
STNC1	PIWL2	A2CEI6	Piwi-like protein 2
STNC1	PSN	P52166	Presenilin sel-12
STNC1	RDEA	Q54RR8	Phosphorelay intermediate protein rdeA
STNC1	RGS14	O43566	Regulator of G-protein signaling 14
STNC1	SNW1	Q5R7R9	SNW domain-containing protein 1
STNC1	STX2	P32856	Syntaxin-2
STNC1	TCSA	Q9P896	Two-component system protein A
STNC1	TEN3	Q9W7R4	Teneurin-3
STNC0	RP12A	Q9SGW3	26S proteasome non-ATPase regulatory subunit
STNC0	XDH	P47989	Xanthine dehydrogenase/oxidase
STNC0	PTEN	O08586	Dual-specificity protein phosphatase
STNC0	DUS6	Q9DBB1	Dual specificity protein phosphatase 6
STNC0	RBP9X	Q4Z8K6	Ran-binding proteins 9/10 homolog
STNC0	AGFG1	Q4KLH5	Arf-GAP domain and FG repeat-containing protein 1
STNC0	AMPD	O80452	AMP deaminase
STNC0	NF1	P97526	Neurofibromin
STNC0	PSRA	Q54VB6	Ser/Thr phosphatase 2A regulatory subunit

Table 46. Proteins from the Cell Differentiation complex CDN1. Featured here are those whose annotations are for proteins other than kinases. Fields follow the same convention as in Table 43.

Protein	Entries	Annotation
AARA	Q54I71	Protein aardvark
ANKR2	Q9WV06	Ankyrin repeat domain-containing 2
CAN5	Q22036	Calpain-5
CUL1	O60999	Cullin-1
CUL2	Q9XZJ3	Cullin-2
DR111	P42698	DNA-damage-repair/toleration protein DRT111
EXD	P40427	Homeobox protein extradenticle
FBXA	Q9Y0T2	F-box/WD repeat-containing protein A
FBXW7	Q9VZF4	F-box/WD repeat-containing protein 7
FHL2	O70433	Four and a half LIM domains protein 2
FIMB2	Q9FKI0	Fimbrin-like protein 2
FZR2	Q8L3Z8	Protein FIZZY-RELATED 2
IMA1	Q96321	Importin subunit alpha-1
IMB	O18388	Importin subunit beta
KCBP	Q9FHN8	Kinesin-like calmodulin-binding protein
LDB3	Q9JKS4	LIM domain-binding protein 3
MSI2	O22468	WD-40 repeat-containing protein MSI2
MSIR6	Q9VVE5	RNA-binding protein Musashi homolog Rbp6
MYO7A	Q17LW0, Q9V3Z6	Myosin-VIIa
P2C76	Q94AT1	Probable protein phosphatase 2C 76
PDLI7	Q679P3	PDZ and LIM domain protein 7
PEX13	Q92968	Peroxisomal membrane protein PEX13
PKHH1	Q00IB7	Pleckstrin domain-containing H1
PSME4	Q5SSW2	Proteasome activator complex subunit 4
PTBP1	P26599, Q00438	Polypyrimidine tract-binding protein 1
PUB13	Q9SNC6	U-box domain-containing protein 13
PZRN3	Q9UPQ7	E3 ubiquitin-protein ligase PDZRN3
RBRA	Q6T486	Probable E3 ubiquitin-protein ligase rbrA
RH35	Q9LU46	DEAD-box ATP-dependent RNA helicase 35
SMCA2	Q6DIC0	Probable global transcription activator SNF2L2
SUV92	Q5F3W5	Histone-lysine N-methyltransferase SUV39H2
TANC1	Q0VGY8	Protein TANC1
TRPC5	Q9UL62	Short transient receptor potential channel 5
U2AF2	Q24562	Splicing factor U2AF 50 kDa subunit
UPL3	Q6WWW4	E3 ubiquitin-protein ligase UPL3
VPS27	O13821	Vacuolar protein sorting-associated 27
WARA	Q54F46	Homeobox protein Wariai
WDS	Q9V3J8	Protein will die slowly
YKX2	Q9P3U4	Uncharacterized RING finger C328.02

Table 47. Kinases from the cell differentiation complex CDN1. Fields follow the same convention as in Table 43.

Protein	**Entries**	**Annotation**
ATG1	Q86CS2	Serine/threonine-protein kinase atg1
DYR1B	Q9Z188	Dual specificity Tyr-phosphorylation-regulated kinase 1B
FES	P14238	Tyrosine-protein kinase Fes/Fps
MAK	P20794	Serine/threonine-protein kinase MAK
PLK1	P70032	Serine/threonine-protein kinase PLK1
PMYT1	Q9NI63	Membrane Tyr/Threonine-specific cdc2-inhibitory kinase
PRKX	P51817	cAMP-dependent protein kinase catalytic subunit PRKX
ZAK2	Q552C6	Dual specificity protein kinase zak2

Table 48. Proteins from the Cell Differentiation complexes CDN2 and CDN3. Fields follow the same convention as in Table 44.

Complex	**Protein**	**Entries**	**Annotation**
CDN2	AFG32	Q8JZQ2	AFG3-like protein 2
CDN2	CALM	P05933	Calmodulin
CDN2	CANB1	Q55G87	Calcineurin subunit B type 1
CDN2	CD48B	Q9ZPR1	Cell division control protein 48 homolog B
CDN2	FIG4	Q92562	Polyphosphoinositide phosphatase
CDN2	FREQ	P37236	Frequenin-1
CDN2	KCIP2	Q9JM59	Kv channel-interacting protein 2
CDN2	NCS1	Q5RC90	Neuronal calcium sensor 1
CDN2	PCH2	Q5XHZ9	Pachytene checkpoint protein 2 homolog
CDN2	PRS4B	Q9SL67	26S proteasome regulatory subunit 4B
CDN2	SP5K	P27643	Stage V sporulation protein K
CDN2	SPAST	Q719N1	Spastin
CDN3	HELLS	Q60848	Lymphocyte-specific helicase
CDN3	FLNC	Q14315, Q8VHX6	Filamin-C

Table 49. Proteins from the Cell Differentiation subnetwork CDO1. Featured here are those whose annotations are other than enzymes. Fields follow the same convention as in Table 43.

Protein	**Entries**	**Annotation**
ABCG2	Q9NGP5	ABC transporter G family member 2
ABCGI	Q8ST66	ABC transporter G family member 18
ACBP5	Q8RWD9	Acyl-CoA-binding domain-containing protein 5
ADSV	Q28046	Adseverin
AP3B1	Q32PG1	AP-3 complex subunit beta-1
ATG5	Q3MQ24	Autophagy protein 5
CDC23	Q9STS3	Anaphase-promoting complex subunit 8
CLH	P25870	Clathrin heavy chain
COTA	P07788	Spore coat protein A
CYB5	Q9V4N3	Cytochrome b5
DIMB	Q54ER9	Basic-leucine zipper transcription factor B
E2FB	Q9FV71	Transcription factor E2FB
FIGL1	Q8BPY9	Fidgetin-like protein 1
MTMR2	Q9Z2D1	Myotubularin-related protein 2
PCSK4	P29121	Proprotein convertase subtilisin/kexin type 4
PESC	P79741	Pescadillo
PIWL1	Q96J94	Piwi-like protein 1
POE	Q9VLT5, Q29L39	Protein purity of essence
P	Q04671	P protein
RS132	P59224	40S ribosomal protein S13-2
SCC12	Q9FQ20	Sister chromatid cohesion 1 protein 2
SMBP2	P40694	DNA-binding protein SMUBP-2
SNP30	Q9LMG8	Putative SNAP25 homologous protein SNAP30
SPO75	Q07798	Sporulation-specific protein 75
SYNJ1	Q62910	Synaptojanin-1
TISB	P23950	Zinc finger protein 36, C3H1 type-like 1
TMTC3	Q6ZXV5	Transmembrane and TPR repeat-containing protein 3
TRAP1	Q86L04	TNF receptor-associated 1, mitochondrial
VP33A	Q9D2N9	Vacuolar protein sorting-associated protein 33A
XCT	Q9UPY5	Cystine/glutamate transporter

Table 50. Enzymes and proteins associated to cell differentiation from the subnetwork CDO1. Featured here are those belonging to the subnetworks CDO2 and CDO3, as well as those that do not form interactions (CDO0). Fields follow the same convention as in Table 45.

Network	Protein	Entries	Annotation
CDO1	ADA22	Q9P0K1	Disintegrin and metalloproteinase domain 22
CDO1	ANM1	Q54EF2	Protein arginine N-methyltransferase 1
CDO1	AT8A2	Q9NTI2	Probable phospholipid-transporting ATPase IB
CDO1	ATG7	Q86CR9	Ubiquitin-like modifier-activating enzyme atg7
CDO1	CNEP1	Q8JIL9	CTD nuclear envelope phosphatase 1
CDO1	CP17A	P11715	Steroid 17-alpha-hydroxylase/17,20 lyase
CDO1	DHRS9	Q9BPW9	Dehydrogenase/reductase SDR family member 9
CDO1	HBD	P45856	Probable 3-hydroxybutyryl-CoA dehydrogenase
CDO1	HERC4	Q5GLZ8	Probable E3 ubiquitin-protein ligase HERC4
CDO1	MYCB2	Q7TPH6	Probable E3 ubiquitin-protein ligase MYCBP2
CDO1	NOXA	Q9XYS3	Superoxide-generating NADPH oxidase heavy chain A
CDO1	PI5K1	Q6EX42	Phosphatidylinositol 4-phosphate 5-kinase 1
CDO1	PI5K5	Q9SLG9	Phosphatidylinositol 4-phosphate 5-kinase 5
CDO1	PKS37	Q54FI3	Probable polyketide synthase 37
CDO1	PP1	Q9UW86	Serine/threonine-protein phosphatase PP1
CDO1	S5A1	Q28891	3-oxo-5-alpha-steroid 4-dehydrogenase 1
CDO1	SAN	Q9NHD5	Probable N-acetyltransferase san
CDO1	SPLA	P18160	Dual specificity protein kinase splA
CDO1	SSH1	Q8WYL5	Protein phosphatase Slingshot homolog 1
CDO1	SSH2	Q76I76	Protein phosphatase Slingshot homolog 2
CDO1	SSH3	Q5XIS1	Protein phosphatase Slingshot homolog 3
CDO1	TAGA	Q9GTN7	Serine protease/ABC transporter B family protein tagA
CDO1	THIC1	Q8S4Y1	Acetyl-CoA acetyltransferase, cytosolic 1
CDO1	UBE12	P92974	Ubiquitin-activating enzyme E1 2
CDO1	UBPE	Q24574	Ubiquitin carboxyl-terminal hydrolase 64E
CDO2	HEXA	Q0V8R6	Beta-hexosaminidase subunit alpha
CDO2	HEXB	P49614	Beta-hexosaminidase subunit beta
CDO3	YVDP	O06997	FAD-linked oxidoreductase YvdP
CDO3	DIM	Q39085	Delta(24)-sterol reductase
CDO0	ENG2	Q09850	Putative endo-1,3(4)-beta-glucanase 2
CDO0	TGM3	Q08189	Protein-glutamine gamma-glutamyltransferase E
CDO0	SPT20	Q8TB22	Spermatogenesis-associated protein 20
CDO0	EHD1	Q641Z6	EH domain-containing protein 1
CDO0	SPO12	Q9M4A1	Meiotic recombination protein SPO11-2
CDO0	EF1A2	P05303	Elongation factor 1-alpha 2
CDO0	ECE2	O60344	Endothelin-converting enzyme 2
CDO0	BGA11	Q9SCV1	Beta-galactosidase 11

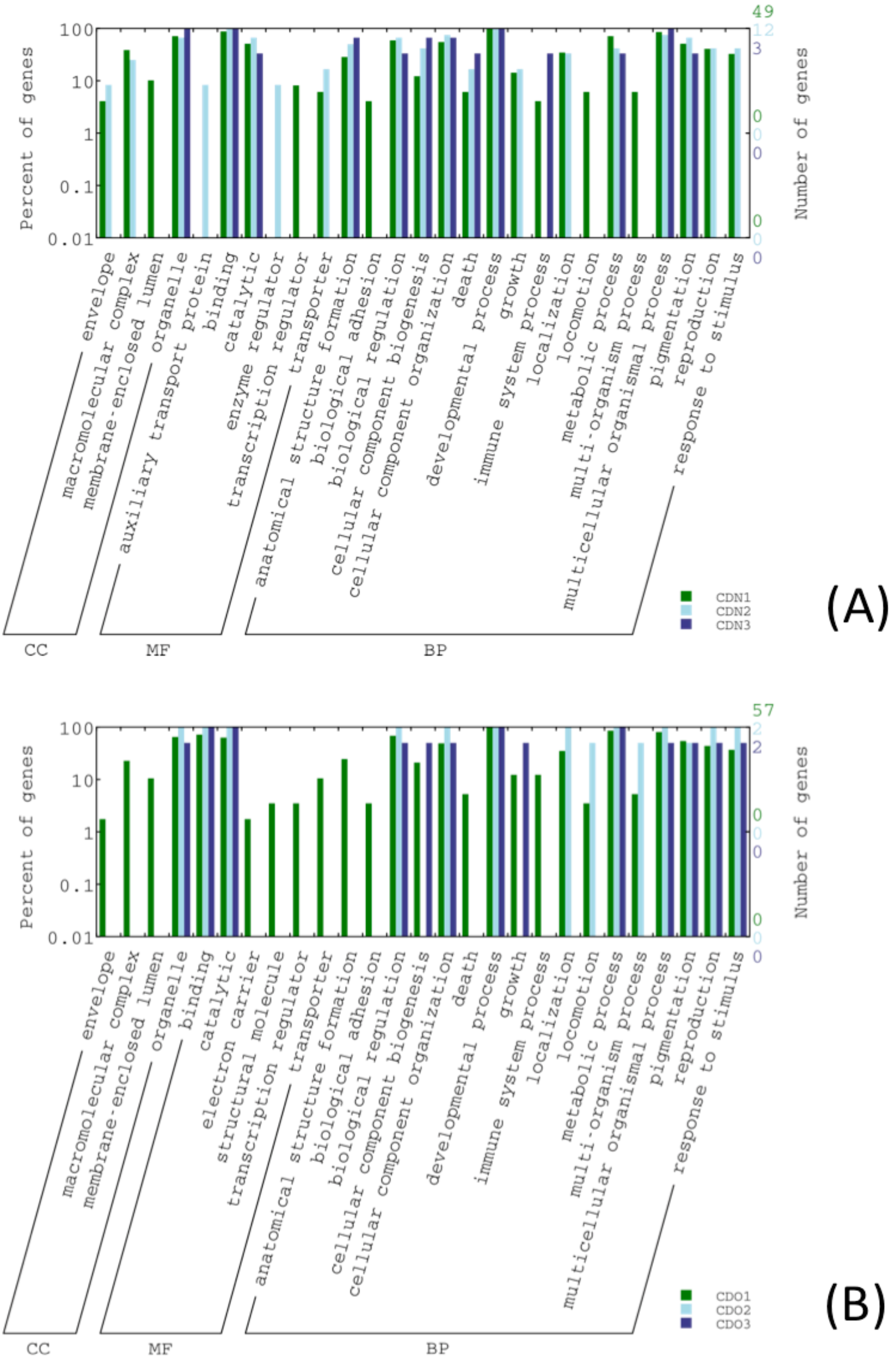


Figure 31. Comparison of gene ontologies between complexes and subnetworks annotated for cell differentiation. Above (*A*), the predicted protein complexes, and below (*B*), interaction networks without complexes. CC, MF and BP correspond to the ontology categories (cellular component, molecular function, and biological process, respectively). The comparison was plotted using the WEGO tool (Ye et al. 2006).

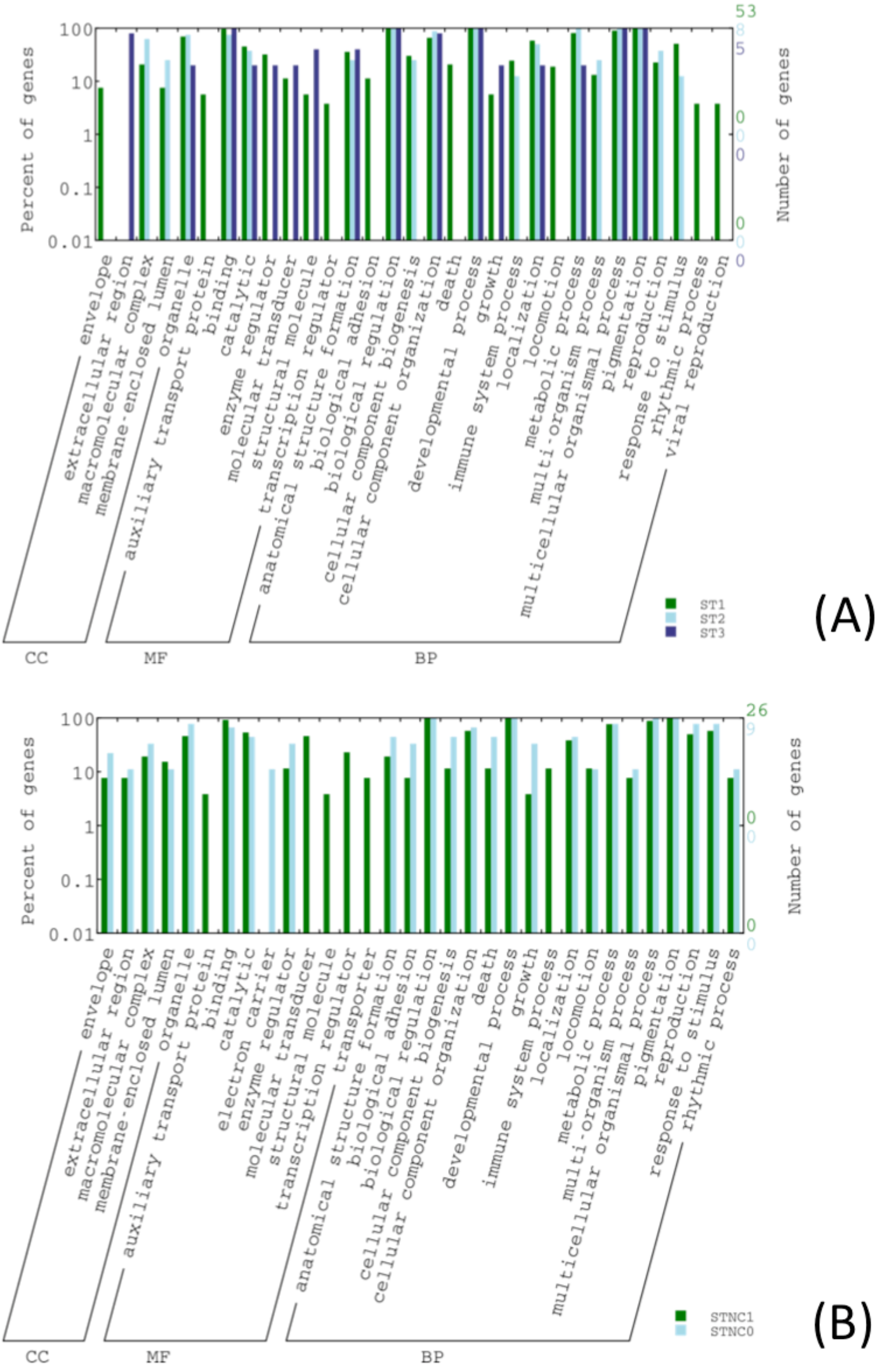


Figure 32. Comparison of gene ontologies between complexes and subnetworks annotated for signal transduction. Above (*A*), the predicted protein complexes, and below (*B*), interaction networks without complexes. The procedure to obtain these plots, and the ontology category conventions, are the same as in the Figure 31.

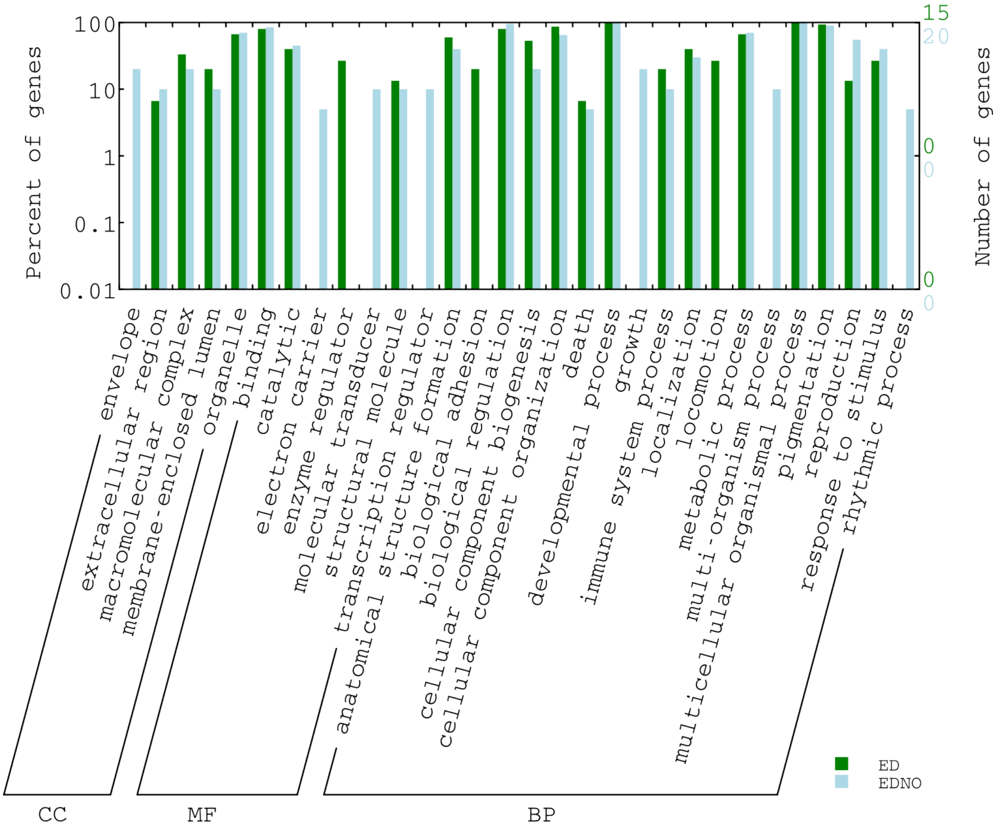


Figure 33. Comparison of gene ontologies between complexes and subnetworks annotated for embryo development. *ED*, stands for the single complex found under this ontology (Table 41), and *EDNO*, for those entries that do not form complexes (Table 42). The procedure to obtain these plots, and the ontology category conventions, are the same as in the Figure 31.

Validation and Completeness of the Genome and the Gene Models

The desired goal for a genome project is to achieve a high- quality draft assembly. In this work, the genome reported by the sequencing facility at Washington University (St.Louis MO) was employed, whose assembly combined short and long genomic reads, to achieve a maximum of completeness. Besides, assessing the accuracy of the annotation is important, given that even the best gene prediction programs and annotation pipelines hardly exceed the 80% accuracy at the exon level. Here, a set of metrics (N50 scaffold, coding potential) and evidences (contaminants, ESTs, RNA-seq, gene prediction, CEGMA) were used to determine whether this genome assembly, and its derived gene models, meet the minimum standards required for submission to databases (Yandell and Ence 2012). These analyses are detailed in the following paragraphs.

Quality of the Assembly. Although there are no general rules for establishing what is a ‘good’ or ‘high-quality’ draft assembly, there are several summary statistics that can be used to describe its completedness and contiguity, and the most commonly used are the N50 scaffold, the gap distributions, and the percent coverages. The scaffold N50 is calculated by ordering each scaffold from longer to shortest, and then the scaffold lengths are summed starting from the longest fragment, until the sum equals half of the total length of all scaffolds. Therefore, the longer the scaffold N50, the better the genome assembly is, and this is useful to compare between different assembly releases from a given species or biological sample. A derived rule of thumb is that an acceptable assembly should have a gene sized N50 scaffold length; *i.e.*, if the N50 scaffold equals the average gene length, then approximately 50% of the genes will be contained in a single scaffold (Yandell and Ence 2012). In this case, the assembly has a N50 scaffold of 97,377 bp (Table 22), a theoretical gene size of 3,743 bp (see *Inference of the Protein-coding gene models*, page 119), and the average of the obtained transcript model is 601.5 bp (Table 38). Thus the average gene and transcript sizes are well below the N50, and therefore this genome release can be counted as a reliable source for gene model annotations. Furthermore, a comparison between the present and former genome releases, using several common assembly descriptors (scaffold N50, gaps and percent coverage), shows that the version 7.3.1 contains less gaps (50.08 versus 50.11 and 77.14 Mb of its immediate predecessors), while having a higher N50 scaffold that the previous version (97.38 vs. 88.91 Kb). These analyses are summarized in the Table 51.

Table 51. Scaffolds and gaps from the most recent genome releases. Percent gaps are the rates between the total gap size and the genome size, and the percent coverage (genomic) was calculated by dividing the obtained genome size against the expected genome value (approx. 300 Mb; Mohberg and Rusch 1971).

Genome

Release	4.0	5.0	7.0	7.3	7.3.1
Date	07/2009	06/2010	05/2011	07/2011	12/2011
Total Size (Mb)	137.54	125.56	272.23	254.79	239.75
Percent coverage (%)	45.85	41.85	90.74	84.93	79.92
Scaffolds					
Total	12,974	5,049	181,840	129,575	126,782
Largest (bp)	97,839	122,657	761,234	821,622	821,622
Smallest (bp)	1,986	11,204	74	17	17
Mean Size (bp)	10,601.2	24,867.8	1,497.1	1,966.4	1,891.1
N50 (bp)	15,456	27,536	114,306	88,913	97,377
Gaps					
Total Gap Size (Mb)	36.48	31.08	77.14	50.11	50.08
Percent gaps (%)	26.52	24.75	28.34	19.67	20.89
Smallest Size (bp)	1	1	1	1	1
Smallest gaps	2,754	4,470	14,349	14,103	14,059
Largest Size (bp)	3,493	3,090	7,021	1,005	1,005
Largest gaps	1	1	1	1	1
Most frequent gap size (bp)	1	1	1	1,000	1,000
Most frequent gaps	2,754	4,470	14,349	39,162	39,137

Sequencing Contaminants. Newly obtained sequences, when acquired from impure DNA preparations, might contain contaminants (sequences from sources different than the intended sample). These contaminants limit the quality of the data, and the conclusions than can be reached in downstream analyses. Consequently, it was decided to check the genome sequence for potential contaminants with the online version of the DeconSeq tool (Schmieder and Edwards 2011). Here the program was ran against the unmasked genome sequences, and results from this analysis can be seen in the Figure 34. By setting the coverage to equal or over 90%, and an identity threshold of 94%, 771 sequences (0.61%) were found matching the contaminant "Remove" databases (archaeal and bacterial genomes). Both the contaminant and clean sequences were kept for further analyses.

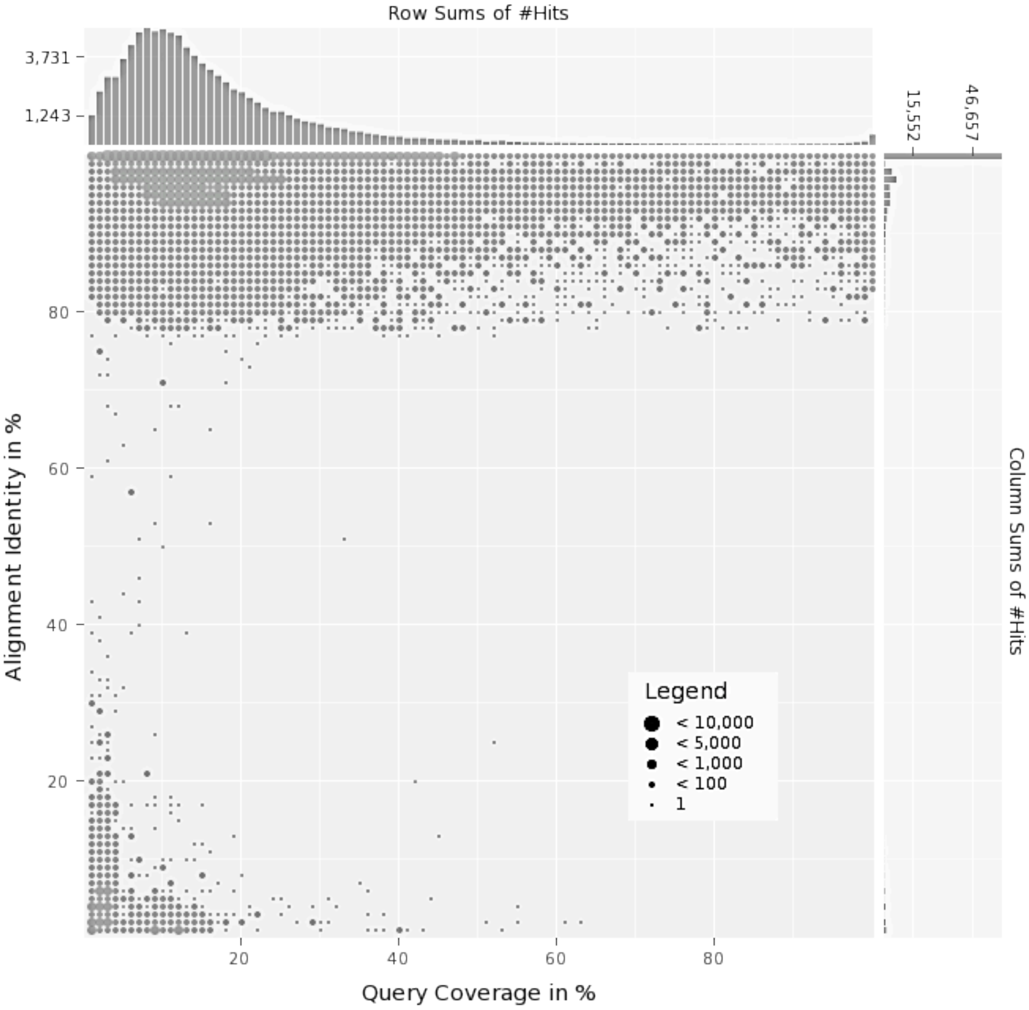


Figure 34. Coverage identity plot of contaminants. Hits against the DeconSeq Remove database are displayed. Multiple hits for one query with different covegare and identity values may be plotted (e.g., two hits with 90% coverage / 90% identity and 89% coverage / 95% identity). This plot was obtained with the DeconSeq program (Schmieder and Edwards 2011).

Mapping ESTs as an estimate of completeness. To estimate the completeness of the sequenced genome and the gene models, a mapping of the clustered ESTs (Glöckner et al. 2008; Watkins and Gray 2008) was performed. To this end, the genome was first masked for repeats with RepeatMasker (Smit et al. 2010; Tempel 2012), using default values. Then a BLASTN search was carried out (version 2.2.27+; Altschul et al. 1990; Camacho et al. 2009) against the clustered cDNAs of *Physarum* (page 115), with an e-value of 1 x 10-8. This value which has been indicated as appropriate for EST mapping (Korf et al. 2003). In this way, 17,577 contigs were successfully matched against the unmasked genome (17,500 for the masked version; see Table 52), which represent the 98.03% of the total cDNAs clustered (97.60% considering the mapping against the masked genome). Later, different datasets of *Physarum* cDNAs were mapped against the protein models obtained in this thesis work. This was done to assess the representation of these transcript sequences in the final reference proteome. The included cDNAs entailed published sequences (Glöckner et al. 2008; Watkins and Gray 2008), the transcriptomic assembly of the 454 RNA-seq (Chapter 2), the clustered cDNAs of *Physarum* (page 115), as well as the Illumina short read mappings completed earlier in this chapter using tophat (page 115; samples WT31, LULU1 and LULU2). The procedure first involved conversion of the tophat outputs into FASTA assemblies, and then all cDNAs were used on BLASTX searches (e-value of 1 x 10-6). The results can be seen in the Table 53. The lowest representation, both at the number of ESTs and proteins matched, corresponded to the tophat assembly of the LULU1 strain: ~4 thousand ESTs and proteins, or roughly 15% of both datasets. Proportionally, from the previosly reported EST banks, the most represented was the dataset reported by Watkins and Gray (2008), with 78.58% of their cDNAs found in the protein models. In addition, the tophat assembly of the WT31 strain has the largest representation in the final models: 22,934 protein matches (89.41% of the proteome).

Mapping *Physarum* GenBank sequences to the novel assembly. In order to annotate the genome with previously characterized *Physarum* genes, all the GenBank (Benson et al. 2008b; Benson et al. 2011) nucleotide and protein sequences were obtained, excluding those of mitochondrial origin, for this organism (NCBI taxonomic id: 5791). This resulted in 253 nucleic acid and 297 amino acid downloaded sequences (Date obtained: January 30th, 2013). Then BLAT was used (version 35x1; Kent 2002) to map these sequences, with default parameters in both searches. In this manner, 261 protein (87.87%) and 231 GenBank nucleotide (91.30%) sequences were located in the masked genome. The fact that approximately one out of ten sequences were not mapped to the genome, might be due to the fragmentation of the genome version analyzed. The BLAT outputs were later converted to the GFF3 standard format, as described in the Sequence Ontology project (Eilbeck et al., 2005), to be incorporated in a future MAKER2 annotation.

Table 52. Mapping clustered ESTs to the *Physarum* genome.

Genome	**Unmasked**	**Masked**
Total mapped	119,859	119,177
ESTs with hits	17,577	17,500
without hits	354	431
Total ESTs	17,931	17,931
Percentage mapped	98.03	97.60

Table 53. Mapping ESTs to the novel protein models.

Dataset	Total ESTs in dataset	Total EST matches	EST matches over 70% identity	Uniquely mapped ESTs	Represented protein models
Watkins and Gray 2008	9,713	33,446	9,415	7,632	4,126
Glöckner et al. 2008	15,680	56,016	13,113	11,684	6,635
454 cDNAs	16,669	56,469	10,879	9,752	11,119
Clustered cDNAs	17,931	64,227	12,320	10,842	12,042
LULU1	26,554	12,093	4,296	4,104	3,930
LULU2	92,109	294,806	42,028	42,601	22,465
WT31	73,836	186,018	39,551	35,990	22,934

Mapping RNA-seq short reads to assess the gene coverage. As a manner of quality control of the accuracy of the Illumina RNA-seq experiments to represent the full transcriptomic set, the gene body coverage of the mapped short reads was investigated. To this end, first the GFF3 annotation file output from MAKER (Holt and Yandell 2011), was converted to the BED format using the gff2bed and sortBed tools from bedops version 2.0.0b (Neph et al., 2012). Then the Illumina sequencing RNA-seq runs from the strain WT31 (Chapter 4) were mapped against the masked genome, version 7.3.1, using bowtie version 0.12.7 (Langmead, Trapnell, Pop & S. L. Salzberg 2009). Previous masking was achieved by employing the RepeatMasker program version 3.0 (Smit et al., 2010; Tempel, 2012), against the *Physarum*- specific custom library, built with RepeatModeler 1.0.7 (Smit et al., 2010). Finally, the mapped reads coverage over the annotated gene bodies were obtained with ever-seq version 1.0.7 (Wang 2012), and plotted using the R statistical environment, version 3.0.0 (R Core Team 2013). The plots below (Figure 35) show uneven distributions of mapped reads, probably due to degradation of the initial RNA sample.

Mapping long RNA-seq reads against the genome. The availability of datasets of long reads allowed to check the presence of these transcript fragments in this genome release. For this, RNAs were obtained from *Physarum polycephalum* amoebae, strain LU352 (Dee et al. 1989), by Marianne Bénard and Gerard Pierron (Institut Gustave-Roussy, Paris XI University, France). cDNAs were synthesized from these RNAs, and then used for sequencing on the 454 FLX platform; the adaptors used are listed on Table 54 (Pat Minx, personal communication). This was carried out at The Genome Institute, Washington University School of Medicine (St Louis, MO). 598,725 spots were obtained (two reads per spot), spanning 155 Mb and distributed in 5 datasets, which were deposited in the Sequence Read Archive (Leinonen, Sugawara, et al. 2011), under the accession SRX000019 (Table 55).

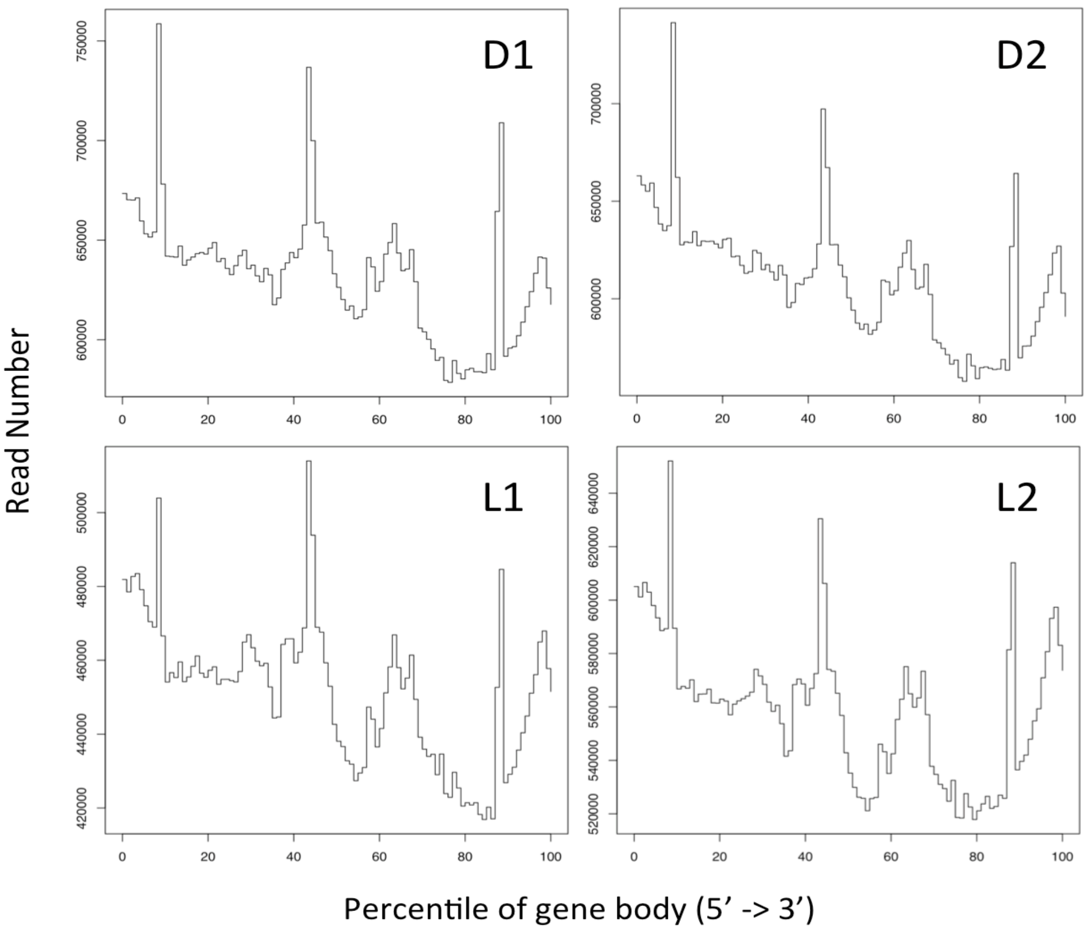


Figure 35. Gene-body coverage of short RNA-seq reads from *Physarum,* strain WT31. The plot shows distributions of mapped short reads for gene bodies of different Illumina RNA-seq outputs (starved D1 and D2, and photoinduced L1 and L2 samples). Gene bodies were defined as the sequence between the transcriptional start and termination sites annotated in the genome version 7.3.1.

Table 54. Adaptors used for cDNA synthesis and sequencing of the amoeba RNA samples.

Name	Step	Sequence (5' → 3')
5' Smart	cDNA	AAG CAG TGG TAA CAA CGC ATC CGA CGC rGrGrG
3' SmartIIA	cDNA	AAG CAG TGG TAA CAA CGC ATC CGA
N-SmartIIA	cDNA	AAG CAG TGG TAA CAA CGC ATC CGA C
Adaptor A	454	CCA TCT CAT CCC TGC GTG TCC CAT CTG TTC CCT CCC TGT CTC
Adaptor B	454	CCT ATC CCC TGT GTG CCT TGC CTA TCC CCT GTT GCG TGT CTC

Table 55. Runs from the 454 sequencing of amoeba transcript library (SRA accession SRX000019). This alignment data was obtained with the flagstat algorithm of samtools, version 0.1.7 (Li et al. 2009), and analyzed using R (R Core Team 2013).

Dataset	Total Reads	Mapped Reads	Rate (%)
SRR000117	6,759	6,695	99.05
SRR000118	496,176	491,109	98.98
SRR000119	446,732	442,103	98.96
SRR000120	3,677	3,659	99.51
SRR000121	6,414	6,350	99.00
SRR000122	3,410	3,382	99.18

The RNA-seq long reads outputs from these experiments were then downloaded from the Sequence Read Archive, and their reverse transcription and sequencing adaptors (Table 54) removed with the TagCleaner tool (Schmieder et al. 2010). For mapping, the datasets were filtered (quality threshold: 25) with the FASTX toolkit (Gordon 2008), and the genome was masked for repeats with the default RepeatMasker library. Then the long reads were compared against the genome with the Burrows - Wheeler aligner (BWA, version 0.6.1-r104; Li and Durbin 2009). This program was chosen because: (*i*) it manages better the alignment to large genomes, through the *bwtsw* option; (*ii*) it entails algorithms optimized for long reads (*bwa-sw* and *bwa-mem*); and (*iii*) the *bwa-sw* algorithm is more sensitive to frequent alignment gaps, which are present in this genome release. Therefore, through the bwa-sw algorithm of the BWA aligner, over 98% of the sequencing reads were mapped (Table 55). These high rates of mapping reflect the fact that these RNA-seq outputs were employed during the finishing of the genome assembly (Pat Minx, personal communication).

Comparison of Evidence- based Gene Models versus Gene Prediction. To assess the reliability of the evidence- based gene modeling, *i.e.* the prediction of protein coding genes using the consensus of experimental sources (cDNA, EST, RNA-seq and protein data; see page 116, *Inference of the Protein-coding gene models*), the obtained gene model structures were compared against models predicted *ab initio* (gene prediction using gene content statistics and signals). To this end, the slime mold genome was searched for novel genes with the

GeneMark program (Borodovsky and Lomsadze 2011). Specifically, the GeneMark-ES algorithm was employed, which unlike most gene prediction software, does not require a previous training, *i.e.*, the obtention of the set of rules and genome- specific parameters that allow the gene identification (gene content, UTR and splice signals, *etc.*). Instead, GeneMark-ES uses the input genomic sequence to derive these rules and parameters, and thus this self-training is an attractive feature for organisms lacking reliable sources of full gene models, as it is the case of *Physarum* (Ter-Hovhannisyan et al. 2008; Borodovsky and Lomsadze 2011). As previously recommended (Shulaev et al. 2011), GeneMark-ES was ran against the repeat- masked genome, switching off the branch point submodel of the program (Lomsadze et al. 2005). The results of these predictions were then compared first against the second and third iteration of MAKER2 (see *Inference of the Protein-coding gene models*; Tables 56 and 57), and later also against the transcript models derived from mapping short RNA-seq reads (see *Mapping RNA-seq short reads to the Physarum genome*; Table 58). These comparisons were carried out with the EVAL program (Keibler and Brent 2003), and using a standard set of metrics for the evaluation of gene prediction programs (Burset and Guigo 1996). These results show that the evidence- based identification produced only a third of exons predicted *ab initio*, but these exons were more than double of the size on average (133.78 bp for the third MAKER iteration versus 59.47 of GeneMark; Table 56). The number of introns also decreased in the final gene models (87 versus 48 thousand), but these introns were more than double the size longer, and the total intronic regions in the genome increased from 14.09 with GeneMark, to 21.43 for the final MAKER2 iteration. Taken these results together, it can be affirmed that the evidence- based gene modeling brought less false positive exons, while being more sensitive in mapping introns. These results are also supported by the fact that these numbers follow a progression of improvement between the second and the third and final MAKER run (in both cases decreasing the exon number while increasing the intron total length; Tables 56 and 57). Moreover, the sensitivity (proportion of coding nucleotides correctly predicted as coding; Burset and Guigo 1996) to detect genes and transcripts, was increased over 30-fold from the GeneMark predictions to the MAKER2 second iteration, while the specificity (proportion of predicted coding bases that are actually coding), was over a hundred times larger in the MAKER2 second iteration than in the *ab initio* predictions (Table 57). The annotation edit distance (AED), which indicates the correspondence between annotations and supporting evidences (Eilbeck et al. 2009), decreased from 99.64 – 99.66% in genes and transcripts in the GeneMark prediction, to ~83% in the MAKER2 second pass distance, meaning that the support of the annotations increased from less than 1% in *ab initio* predictions to close to 17% in evidence- based gene modeling (*i.e.*, the lower the AED value, the better the agreement between the annotation set and its evidence; Table 57). Furthermore, a comparison between the mappings of RNA-seq short reads from several strains, the GeneMark predictions, and the evidence- based MAKER2 gene models, showed larger transcript per gene rates in the RNA-seq and MAKER2 analyses than in the *ab initio* predictions, which might account for a better identification of alternative spliced transcripts using these two methods (Table 58). Finally, the average number of exons was also larger in the RNA-seq mappings and MAKER2 gene models, except in the case of the RNA-seq sample with the lower coverage (LULU1; Table 58).

Table 56. Evidence- and prediction- based exons and introns. A comparison of gene models predicted by MAKER2 (second and third iterations; *M2* and *M3* respectively), and GeneMark (*GMES*), is listed. These results include all types of exons: single, initial, internal and terminal.

Model Source	M3	M2	GMES
	Exons		
Count	73,670	114,992	231,981
Average Length	133.78	142.64	59.47
Total Length (Mb)	9.86	16.40	13.80
	Introns		
Count	48,606	85,636	87,444
Average Length	440.86	451.38	161.09
Total Length (Mb)	21.43	38.65	14.09

Table 57. Correspondences between gene models and *ab initio* gene predictions. These statistics entail comparisons of annotations from the second iteration and the gene prediction, against those from the final gene model set (third MAKER2 iteration; M2 and GMES versus M3). Results were obtained with eval (Keibler and Brent 2003), and the parameters used, *i.e.* specificity and sensitivity, are based on the recommendations by Burset and Guigó (1996), except for the accuracy and annotation edit distance (*AED*), which were manually calculated from the sensitivity and specificity values (Yandell and Ence 2012).

Model Source	M2	GMES	M2	GMES
	Gene		*Transcript*	
Sensitivity	18.13%	0.58%	17.81%	0.55%
Specificity	15.71%	0.14%	14.59%	0.14%
Accuracy	16.92%	0.36%	16.2%	0.34%
AED	83.08%	99.64%	83.8%	99.66%
	Exon		*Nucleotide*	
Sensitivity	48.49%	1.89%	96.95%	76.08%
Specificity	31.06%	0.64%	86.01%	83.90%
Accuracy	39.78%	1.27%	91.48%	79.99%
AED	60.23%	98.74%	8.52%	20.01%

Table 58. Comparison of genes and transcripts identified though RNA-seq, *ab initio* gene prediction and consensus gene modeling. WT31, LULU1 and LULU2 represent the RNA-seq reads from said strains, mapped against the genome using TopHat (see page 116); GMES is the gene identification with GeneMark-ES; and M2 and M3 are the second and third MAKER2 runs. Statistics obtained with eval (Keibler and Brent 2003), except for the exons count of the RNA-seq mapping outputs, that were calculated using a bash/perl script one-liner.

Model Source	WT31	LULU1	LULU2	GMES	M2	M3
			Genes			
Count	68,872	25,737	82,584	190,995	28,379	24,615
Total Transcripts	73,836	26,554	92,109	190,995	31,429	25,649
Transcripts Per Gene	1.07	1.03	1.12	1	1.11	1.04
			Transcripts			
Count	73,836	26,554	92,109	190,995	31,429	25,649
Average Length	952.87	322.48	1,378.54	172.7	1,915.42	1,253.5
Total Length (Mb)	70.36	8.56	126.98	32.98	60.20	32.15
Ave Exons Per	1.99	1.38	2.76	1.53	3.99	2.94
Total Exons	146,644	36,597	254,347	292,177	125,363	75,448

Mapping CEGMA datasets as an estimate of completeness of the genome assembly and the protein models. CEGMA (*core eukaryotic genes mapping approach*), is a method to assess the reliability of a set of annotations, which includes a computational pipeline and sets of conserved, single-copy protein families present in a extensive range of eukaryotes (Parra et al. 2007). Two CEGMA sets of protein families are available (Table 59, Parra et al. 2007; Parra et al. 2009), and the comparison of novel genome scaffolds and their annotations against these protein sets can be also used as an estimate of completeness and contiguity of a reported assembly (Yandell and Ence 2012). Here, to evaluate the completeness of the current genome release and the reported protein- coding gene models, the coverage of the two CEGMA protein sets in these two sequence datasets was analyzed.

First, for the evaluation of the genome sequence, I used the GenBlastA software (She et al. 2009; She et al. 2011), following a previously reported protocol that employed the default settings of the program (Wang et al. 2011). GenBlastA is a program that filters the high scoring sequence outputs from a BLAST similarity search, in order to identify candidate homologous genes (She et al. 2009). In this case, 451 proteins from the core set (98.47%) and 245 from the second set (98.79%) were matched by GenBlastA to the current genome release; 206 (83.06%) and 386 (84.27%) possessed over 50% identity coverage. In addition, more than a half of the core eukaryotic genes (CEGs) were found, regardless of using masked or unmasked sequences in the analysis (Table 60), and similar results were obtained with both CEGMA datasets (54.03% with the CEGMA-248 sequences, and 58.52% for the core set; Table 60). Later, a comparison against the CEGMA-248 set was carried out, this time using the CEGMA pipeline itself (Keith Bradnam, personal communication). This procedure involved a combination of similarity and motif searches, in order to find represented CEGs on a given genome. For this analysis, the CEGMA proteins were separated into four groups, according to their sequence conservation, where the Group 1 contains the most divergent, and the Group 4 the most conserved set (Parra et al. 2009). This separate assessment is recommended for highly divergent genomes, in order to avoid the bias caused by evolutionary distance. According to these results (Table 61), the novel genome assembly of *Physarum* can be classified as a highly divergent sequence, as the partial matches range from 50% completeness in the Group 1, to 83.08% for the Group 4, and this pattern is expected in this type of genomes; however, considering the complete sequences, the *Physarum* genome could be considered either incomplete or divergent, given that the conservation ranges from 33.33% to 55.74 (Parra et al. 2009). Finally, the predicted reference proteome and the CEGs were compared using blastp (e-value 1E-6; Altschul et al. 1997). 84 matches (18.34%) were found against the dataset with 458 proteins, and 43 against the smaller CEGMA dataset (17.34%) with over 70% of identity coverage. These results are on disagreement with previous results with GenBlastA from this thesis work, while at the same time supporting the idea that this genome release could be incomplete, at least in its protein- coding regions. Nevertheless, it is also possible that the results from the CEGMA pipeline in this case are not conclusive due to the fact that this estimate of completeness may not be accurate in highly divergent genomes (Parra et al. 2009).

Protein coding potential of the transcript models. In terms of mass, most cellular RNA is noncoding, and while there are several experimental methods that allow the identification of RNA molecules as such, these procedures are limited by cost and number of samples, and therefore the incorporation of computational predictions might help in the search for putative noncoding RNA genes. Here, to find coding and noncoding sequences among the transcript models, the Coding Potential Calculator (CPC) program (Kong et al. 2007) was used as previously described (Young et al. 2012).

Table 59. CEGMA datasets. The core set are the original CEGs (core eukaryotic genes), and the 248 dataset is a group of genes with are generally in low copy number, and therefore are better for studying the completeness of genomes (Parra et al. 2009).

Dataset	Core	248
Total proteins	2,748	1,488
Unique CEGs	458	248
Reference	Parra et al. 2007	Parra et al. 2009

Table 60. Mapping of CEGMA datasets to different genome versions. For this search, GenBlastA with default settings of the different genome releases was used, including the current unmasked (*U*) and masked for repeats (*M*) versions. The protein matches listed below correspond to those over >= 70% identity coverage, and the percentage stands for the proportion of proteins found on a given genome release, as compared to the total number of proteins in the analyzed CEGMA dataset.

CEGMA Dataset	248		Core	
Genome Release	*Matches*	*Percentage*	*Matches*	*Percentage*
4.0	70	28,23	158	34,50
7.0	127	51,21	249	54,37
7.3	173	69,76	324	70,74
7.3.1 (U)	136	54,84	268	58,52
7.3.1 (M)	134	54,03	268	58,52

CPC is a *de novo* noncoding RNA predictor that classifies novel sequences as coding or not, based on several sequence features, *e.g.* a coding transcript will have more similarity search hits (and with lower e-values) with known proteins than a noncoding one, and these hits usually reside within one frame (Kong et al. 2007). Several tests with noncoding RNA databases, reported CPC as the most sensitive of its type (Wang et al. 2013). In this manner, 19,254 transcripts (75.07%) were predicted as noncoding, from which 17,823 lacked UniProt annotations, and 9,306 could not be associated with InterPro domains (Table 62). Conversely, 3,214 transcripts (12.53%) were predicted as coding by CPC while having an ortholog in the UniProt database. It was also noticed that the ESTs libraries obtained by cDNA cloning and Sanger sequencing contained proportionally more coding sequences than the 454 and MAKER gene models (Table 62). This must be due to a large percentage of RNA-seq reads that are expressed from noncoding regions. In summary, further screenings from noncoding RNAs should be performed in these candidate genes, *e.g.* RFAM and tRNA searches, to identify true long noncoding sequences in this genome.

Table 61. Statistics of the completeness of the genome using CEGMA. This analysis is based on the 248 CEGs dataset (Parra et al. 2009), and was carried out at the Genome Center of the University of California, Davis (Keith Bradnam, personal communication). CEGs were divided into divergence groups (1 to 4), and the *complete* and *partial* matches are included; complete proteins will also be included with the partial matches. *Proteins*, are the number of CEGs present in the genome; *Completeness*, is the percentage of CEGs present; *Total*, number of CEGs including putative orthologs; *Average*, is the average number of orthologs per CEG; and *Orthologs*, are percentages of CEGs that have more than one ortholog.

Complete

	Proteins	*Completeness*	*Total*	*Average*	*Orthologs*
Total	115	46.37	148	1.29	20.87
Group 1	22	33.33	28	1.27	18.18
Group 2	23	41.07	27	1.17	17.39
Group 3	34	55.74	46	1.35	26.47
Group 4	36	55.38	47	1.31	19.44

Partial

	Proteins	*Completeness*	*Total*	*Average*	*Orthologs*
Total	183	73.79	290	1.58	38.80
Group 1	33	50.00	44	1.33	24.24
Group 2	45	80.36	66	1.47	33.33
Group 3	51	83.61	83	1.63	47.06
Group 4	54	83.08	97	1.80	44.44

Table 62. Predicted coding potential of transcripts and ESTs. These predictions were obtained with CPC (Kong et al. 2007) using default values, against two EST libraries (Glöckner et al. 2008; Watkins and Gray 2008), the 454 assembly (Chapter 3), the clustering of all cDNAs, and the second and third MAKER2 runs (M2 and M3).

EST dataset	Total cDNAs	Coding Transcripts	ncRNAs	Percentage Coding
Glöckner 2008	15,680	7,159	8,521	45.66%
Watkins and Gray 2008	9,713	6,639	3,074	68.35%
454 assembled cDNAs	16,669	4,173	12,496	25.03%
Clustered cDNAs	17,931	4,881	13,050	27.22%
M2 transcript models	31,429	9,702	21,727	30.87%
M3 transcript models	25,649	6,125	19,524	23.88%

Comparative Analyses.

Genome Assemblies. The genome of *Physarum* is considerably larger than the closest Mycetozoans, being close to five times the size of the genome of *D.discoideum*, and eight times the assembly from *D.purpureum* (Table 63). The number of undefined bases also exceeds those from these *Dictyostelium* species (50.08 Mb versus 0.03 and 0.11 Mb), and the GC-content is almost the double from these taxa (41.16% versus 21.99 and 24.47%; Table 63). In both *Physarum* and *D.purpureum*, the Scaffold N50 is over 50 Kb, with their average gene sizes of 1,689 bp (dictybase website; Gaudet et al. 2011) and 1,253.5 bp (see Table 58) respectively, and this means that in theory more than 50% of the genes will be contained on a single scaffold (Yandell and Ence 2012). Similarly, the scaffold N50 of *D.discoideum* (3,809 bp) still fits an average gene (1,756 bp; data from dictybase, accessed September 9, 2010), although to a lesser extent. The larger scaffold sizes and N50 in *Physarum* and *D.purpureum*, might account for the differences in sequencing technologies used in these projects (next generation sequencing versus Sanger in *D.discoideum*; Sucgang et al. 2011).

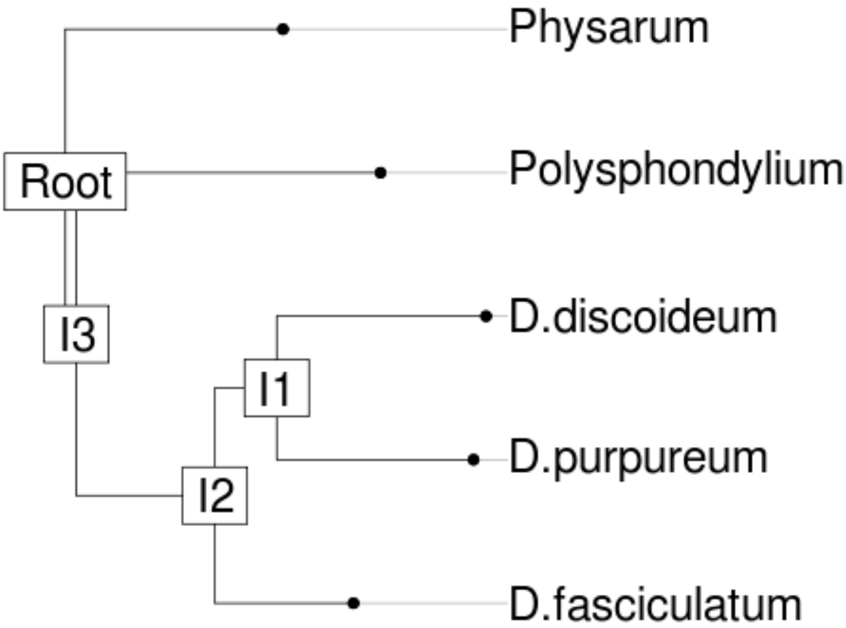


Figure 36. Phylogenetic tree of Mycetozoans. This plot is based on the multiple alignments of conserved coding sequence blocks, calculated with mauve (Darling et al. 2004).

Table 63. Sequencing summary of the genomes of *Physarum* and other Mycetozoa. All specified values are in base pairs (bp). These statistics were obtained using the faSize program, from the Jim Kent source tree, except for the GC level percentage, which was obtained from the RepeatMasker output, and the N50- related values, which were calculated using in-house Perl scripts.

Species	***D.discoideum***	***D.purpureum***	***P.polycephalum***
Data obtained	26 Feb 2013	26 Feb 2013	05 Dec 2011
Data source	Dictybase	JGI	WUSTL
Site	dictybase.org	genome.jgi-psf.org	genome.wustl.edu
Reference	Eichinger et al. 2005	Sucgang et al. 2011	Unpublished
Total bases	50,649,189	32,967,507	239,752,614
Undefined bases	36,046	115,529	50,083,098
Real bases	50,613,143	32,851,978	189,669,516
GC-level (%)	21.99	24.47	41.16
Sequences	13,475	799	126,782
Mean Size	3,758.8	41,261.0	1,891.1
Smallest size	1084	3,010	17
Smallest scaffold	DDB_G0294661	scaffold_821	Scaffold244352
Largest size	35,422	285,244	821,622
Largest scaffold	DDB_G0292696	scaffold_1	Scaffold1
Scaffold N50	3,809	66,881	97,377
N50 length	25,325,737	16,520,785	119,912,848

The general features of the *Physarum* genome show that this novel assembly might form a separate clade within the Mycetozoans. To analyze this further, a phylogenetic analysis at the whole- transcriptome level, involving multiple alignment of all conserved coding blocks, was built with the mauve program, release 2.3.1 (Darling et al. 2004). Here, sequences from *P.polycephalum, Dictyostelium discoideum* (Eichinger et al. 2005)*, D.purpureum* (Sucgang et al. 2011)*, D.fasciculatum,* and *Polysphondylium pallidum* (Heidel et al. 2011) were compared. After the alignments, a rooted tree was plotted from the Newick output from mauve, using the ADE4 library from the R statistical environment (Dray and Dufour 2007; R Core Team 2013). The tree shows the expected separation of Physarum from the *Dictyostelium* clade (Figure 36).

Repetitive Sequences. Similar repeat searches were performed over the genomes of the social amoebae *Dictyostelium discoideum* (Eichinger et al. 2005) and *D.purpureum* (Sucgang et al. 2011), as I did before with the *Physarum* assembly. To this end, genomic sequences were downloaded from dictybase (Fey et al., 2009) and the DoE Joint Genome Institute websites, respectively, and these datasets were then analyzed using RepeatMasker with default settings, and compared against the default repeat library. Results from these analyses are displayed on the Table 64. First, it was observed that the total length of SINE elements is small, compared to the *Physarum* genome searched with the default repeat library: 308 and 139 bp in dictyostelids, versus 27,026 bp in *Physarum* (Tables 23 and 64). The extent of LINEs, LTR and DNA elements is 12.67x, 52.62x and 107.33 times larger in *Physarum* than in *D.discoideum*; similar results were obtained with *D.purpureum*, although RepeatMasker found no LTR elements in this species (Table 64). The predicted total length of small RNAs in *Physarum* is between the values of both dictyostelids, although *D.discoideum* presents a larger proportion, accounting for its smaller genome size (0.11% of small RNAs). Furthermore, the total extent of simple repeats and low complexity regions in *Physarum* (13.11 Mb of simple repeats and 21.35 Mb of low complexity sequences), is also larger in *Physarum* than in dictyostelids, although proportionally the *Dictyostelium* genus have larger simple repeat content (17.75% in *D.discoideum* and 9.87% in *D.purpureum*). Finally, no satellites were detected in dictyostelids, while 23,133 bp of this type of elements were found in the *Physarum* genome. Similar results were obtained when using the custom library for *Physarum*, except for detecting no small RNAs in this case (Table 23).

Moreover, it was also noticed that the employed version of the transposable elements database (RepBase v.20120418) contained 179 sequences matching *Physarum polycephalum* (56,836 bp): 176 ancestral and ubiquitous sequences (two of these belonging to the Mycetozoa clade), with a total length of 50,916 bp, and three lineage- specific sequences (5,920 bp), corresponding to the retrotransposon- related Tp1 (274 bp; Rothnie et al., 1991) and Tp2 elements (1,679 bp; McCurrach et al., 1990), and a HERO non-LTR retrotransposon (3,967 bp; Kapitonov and Jurka, 2009). In comparison, this database possesses 18 lineage specific sequences (74,309 bp in total) for *Dictyostelium discoideum*, and none for *D.purpureum*. These differences in the default library might account for some of the disparities on the results.

Encoded Genes and Proteins. The number of protein models predicted in this work for *Physarum* far exceeds those for the dictyostelids: there are 25,649 in the slime mold, as compared to over 12 thousand genes in both *D.discoideum* and *D.purpureum* (Table 65). In fact, this large number of proteins, within protists, is only akin to those from free-living Ciliophora. The reasons why these organisms feature larger numbers of protein coding genes is unknown, although they also possess large genome sizes (ciliophora genomes range from 72 Mb in *Paramecium*, to 103 Mb for the *Tetrahymena* macronuclear sequence; Liolios et al. 2010). In addition, to compare the number of noncoding transcripts between *Physarum* and other mycetozoans, I calculated the coding potential in the *D.discoideum* transcripts, using CPC (Kong et al. 2007). 11,128 transcripts (90.33%) were classified as coding in *Dictyostelium*, a number proportionally higher than the number found in *Physarum* (6,125 coding transcripts or 23.88%; Table 62). These results might be related to the higher number of repetitive elements found in the slime mold genome (see paragraph above).

Afterwards, the tRNA gene subsets from *Physarum* and *Dictyostelium* were compared, using the predicted data obtained with tRNAscan-SE for the first (Table 29), and data from dictybase (Gaudet et al. 2011) for the latter. Specifically, it was studied whether the number of tRNA genes is correlated to the codon usage. To this end, first the total number of codons, and then the relative frequencies of occurrence of synonymous codons for a specific amino acid (also called “relative synonymous codon usage,” or RSCU; Oresic and Shalloway 1998) were calculated. This was carried out with the codonw program, in transcripts of both species, as previously described (Peden 2005; Behura and Severson 2011). Results from these analyses can be seen on the Tables 66 and 67.

Table 64. Distribution of repetitive elements identified with RepeatMasker on two species of the *Dictyostelium* genus, *D.discoideum* and *D.purpureum*. Column parameters follow the same convention as in Table 23.

Species	***Dictyostelium discoideum***			***Dictyostelium purpureum***		
Parameter	*Elements*	*Length*	*Perc (%)*	*Elements*	*Length*	*Perc (%)*
SINEs	6	308	0	3	139	0
ALUs	1	253	0.00	0.00	0.00	0.00
MIRs	0.00	0.00	0.00	0.00	0.00	0.00
LINEs	141	26,660	0.05	76	15,659	0.05
LINE1	140	26,625	0.05	48	13,844	0.04
LINE2	0.00	0.00	0.00	6	351	0.00
L3/CR1	1	35	0.00	22	1,464	0.00
LTR elements	5	543	0.00	0.00	0.00	0.00
ERVL	0.00	0.00	0.00	0.00	0.00	0.00
ERVL-MaLRs	0.00	0.00	0.00	0.00	0.00	0.00
ERV_classI	4	490	0.00	0	0	0.00
ERV_classII	1	53	0.00	0	0	0.00
DNA elements	3	163	0.00	5	269	0.00
hAT-Charlie	1	66	0.00	1	41	0.00
TcMar-Tigger	1	51	0.00	1	59	0.00
Unclassified	0	0	0.00	0	0	0.00
Total	---	27,674	0.05	---	16,067	0.05
Small RNA	640	56,454	0.11	230	25,304	0.08
Satellites	0	0	0	0	0	0
Simple repeats	158,246	8,990,383	17.75	58,402	3,252,880	9.87
Low complexity	34,931	2,668,725	5.27	15,060	959,594	2.91
Bases masked	---	11,453,026	22.61	---	4,158,906	12.62

Table 65. Protein-coding gene numbers between protists and other lower eukaryotes. *Abbreviations.* Where: Autotrophic (A); free-living (F); mixotrophic (M); parasite (P); and saprophytic (S). Sources: dictybase (*D;* Chisholm et al. 2006); GOLD (*G;* Liolios et al. 2010); Heidel et al. 2011 (*H*); and the DoE Joint Genome Institute (*J;* Grigoriev et al. 2012).

Organism	**Genes**	**Group**	**Life**	**Source**
Phaeodactylum	9,479	Bacillariophyta	A	G
Thalassiosira	11,242	Bacillariophyta	A	G
Paulinella	922	Cercozoa	A	G
Guillardia	553	Cryptophyta	A	G
Cyanidioschyzon	5,331	Rhodophyta	A	G
Tetrahymena	27,000	Ciliophora	F	G
Paramecium	40,000	Ciliophora	F	G
D.purpureum	12,410	Mycetozoa	F	J
D.discoideum	12,646	Mycetozoa	F	D
D.lacteum	11,477	Mycetozoa	F	H
D.fasciculatum	12,173	Mycetozoa	F	H
Polysphondylium	12,373	Mycetozoa	F	H
Naeglaeria	15,753	Heterolobosea	F	J
Bigelowiella natans	21,708	Rhizaria	M	J
T.gondii	8,155	Apicomplexa	P	G
Babesia	3,773	Apicomplexa	P	G
Theileria	4,159	Apicomplexa	P	G
Cryptosporidium	3,956	Apicomplexa	P	G
C.parvum	3,886	Apicomplexa	P	G
P.falciparum	5,298	Apicomplexa	P	G
P.yoelii	7,910	Apicomplexa	P	G
Giardia	6,598	Diplomonadida	P	G
E.histolytica	10,202	Entamoebidae	P	G
T.brucei	10,253	Euglenozoa	P	G
L.infantum	7,993	Euglenozoa	P	G
T.cruzi	22,570	Euglenozoa	P	G
L.major	1,579	Euglenozoa	P	G
Phytophthora	17,797	Stramenopiles	S	G
Monosiga	9,174	Choanoflagellates	F	G
Hydra	18,950	Cnidaria	F	G
Trichoplax	11,520	Placozoa	F	G

These results were then compared to the number of tRNA genes on each of these two genomes, and a significant association ($P < 0.05$) between the number of tRNA genes and the codon usage in both *Physarum* and *Dictyostelium* was found (Figure 37). These results are analogous to those obtained in bacteria, yeast, *C.elegans*, *Drosophila*, and the mosquitoes *Aedes* and *Anopheles* (Behura and Severson 2011), which also showed a significant correlation between the codon usage and the number of tRNA genes.

Subsequently, the differences in gene ontology (GO) and KEGG orthologs (KOs) annotations of *Physarum* and the two species of *Dictyostelium* studied above were investigated. To this end, first a GO slim analysis was carried out, *i.e.*, obtaining a summary of gene ontologies for a large annotation set. The GOslimViewer server (McCarthy et al. 2006a) was used, with the obtained gene ontology annotations from *D.discoideum* and *D.purpureum* from dictybase (Gaudet et al. 2011), together with the annotations from the reference transcriptome obtained for *Physarum* (Tables 68 – 70). Here it was observed that the number of genes in *Physarum* associated to the signal transduction ontology (GO:0007165, 513 genes or 20.28% of biological process annotations), is proportionally larger than those in the dictyostelids: 1,078 (8.21%) and 287 (10.63%) of the biological process ontologies for *D.discoideum* and *D.purpureum* respectively. This might reflect the fact that most well studied genes and proteins in *Physarum* are related to this type of signaling processes. Conversely, the proportion of transport ontologies (GO:0006810), is larger in the dictyostelids than in *Physarum* (over 20% in the former group, versus 13.17% in the slime mold). Other categories of genes, particularly those related to metabolic processes (GO:0009058, biosynthesis process; and GO:0009056, catabolic process), showed similar proportions in the three studied taxa. Later, these results were extended, using the original annotations from *Physarum* and those from dictyostelids (downloaded from dictybase), to plot the differences of all gene ontologies with the WEGO online tool (Ye et al. 2006).

Table 66. Codon usage pattern in *Physarum*. Total codon counts (*N*) and *RSCU* (relative synonymous codon usage) values are displayed for each codon. Calculations were obtained with the codonw program (Peden 2005).

AA	**Codon**	**N**	**RSCU**	**AA**	**Codon**	**N**	**RSCU**
Phe	UUU	207,307	1.35	*Ser*	UCU	82,131	1.18
	UUC	100,429	0.65		UCC	79,893	1.14
Leu	UUA	83,520	1.04		UCA	78,331	1.12
	UUG	106,377	1.33		UCG	47,285	0.68
	CUU	89,686	1.12	*Pro*	CCU	74,987	1.09
	CUC	80,874	1.01		CCC	78,470	1.14
	CUA	58,741	0.73		CCA	87,537	1.27
	CUG	61,685	0.77		CCG	35,202	0.51
Ile	AUU	126,451	1.33	*Thr*	ACU	73,503	0.97
	AUC	73,847	0.78		ACC	68,568	0.91
	AUA	85,299	0.9		ACA	107,789	1.42
Met	AUG	86,848	1		ACG	53,004	0.7
Val	GUU	73,968	1.2	*Ala*	GCU	63,651	0.99
	GUC	37,969	0.62		GCC	51,510	0.8
	GUA	60,835	0.99		GCA	86,886	1.36
	GUG	73,657	1.2		GCG	54,245	0.85
Tyr	UAU	88,176	1.1	*Cys*	UGU	76,939	1.02
	UAC	71,971	0.9		UGC	74,072	0.98
Stop	UAA	71,120	1.15	Stop	UGA	70,007	1.13
Stop	UAG	44,993	0.73	*Trp*	UGG	81,894	1
His	CAU	71,602	0.92	*Arg*	CGU	40,185	0.69
	CAC	83,494	1.08		CGC	54,678	0.94
Gln	CAA	137,930	1.34		CGA	54,459	0.94
	CAG	67,204	0.66		CGG	30,429	0.52
Asn	AAU	127,576	1.15	*Ser*	AGU	64,750	0.93
	AAC	94,993	0.85		AGC	66,344	0.95
Lys	AAA	246,097	1.35	*Arg*	AGA	99,508	1.71
	AAG	117,932	0.65		AGG	68,877	1.19
Asp	GAU	83,079	1.18	*Gly*	GGU	50,266	0.83
	GAC	57,768	0.82		GGC	48,859	0.8
Glu	GAA	128,061	1.2		GGA	90,326	1.49
	GAG	85,711	0.8		GGG	53,391	0.88

Table 67. Codon usage pattern in *Dictyostelium discoideum*. Total codon counts (*N*) and *RSCU* (relative synonymous codon usage) values are shown for each codon, and follow the same conventions as in Table 66.

AA	**Codon**	**N**	**RSCU**	**AA**	**Codon**	**N**	**RSCU**
Phe	TTT	227,897	1.42	*Ser*	TCT	106,729	0.96
	TTC	93,939	0.58		TCC	27,186	0.24
Leu	TTA	391,116	3.97		TCA	347,541	3.11
	TTG	73,016	0.74		TCG	15,982	0.14
	CTT	65,402	0.66	*Pro*	CCT	40,628	0.59
	CTC	22,612	0.23		CCC	7,951	0.12
	CTA	35,635	0.36		CCA	221,685	3.24
	CTG	2,693	0.03		CCG	3,631	0.05
Ile	ATT	354,934	1.83	*Thr*	ACT	143,962	1.39
	ATC	76,747	0.4		ACC	53,928	0.52
	ATA	148,697	0.77		ACA	209,559	2.02
Met	ATG	108,057	1		ACG	7,066	0.07
Val	GTT	162,975	2.22	*Ala*	GCT	69,223	1.31
	GTC	22,662	0.31		GCC	22,814	0.43
	GTA	91,612	1.25		GCA	114,611	2.17
	GTG	15,886	0.22		GCG	4,222	0.08
Tyr	TAT	207,489	1.7	*Cys*	TGT	87,110	1.79
	TAC	35,969	0.3		TGC	10,378	0.21
Stop	TAA	12,012	2.56	Stop	TGA	1,029	0.22
Stop	TAG	1,013	0.22	*Trp*	TGG	50,404	1
His	CAT	103,772	1.7	*Arg*	CGT	40,125	1.24
	CAC	18,299	0.3		CGC	644	0.02
Gln	CAA	338,481	1.92		CGA	3,729	0.12
	CAG	13,466	0.08		CGG	414	0.01
Asn	AAT	700,498	1.79	*Ser*	AGT	155,436	1.39
	AAC	81,800	0.21		AGC	17,384	0.16
Lys	AAA	448,344	1.69	*Arg*	AGA	138,697	4.3
	AAG	80,922	0.31		AGG	9,865	0.31
Asp	GAT	327,835	1.83	*Gly*	GGT	223,372	2.91
	GAC	31,099	0.17		GGC	14,691	0.19
Glu	GAA	340,872	1.69		GGA	62,249	0.81
	GAG	61,898	0.31		GGG	6,482	0.08

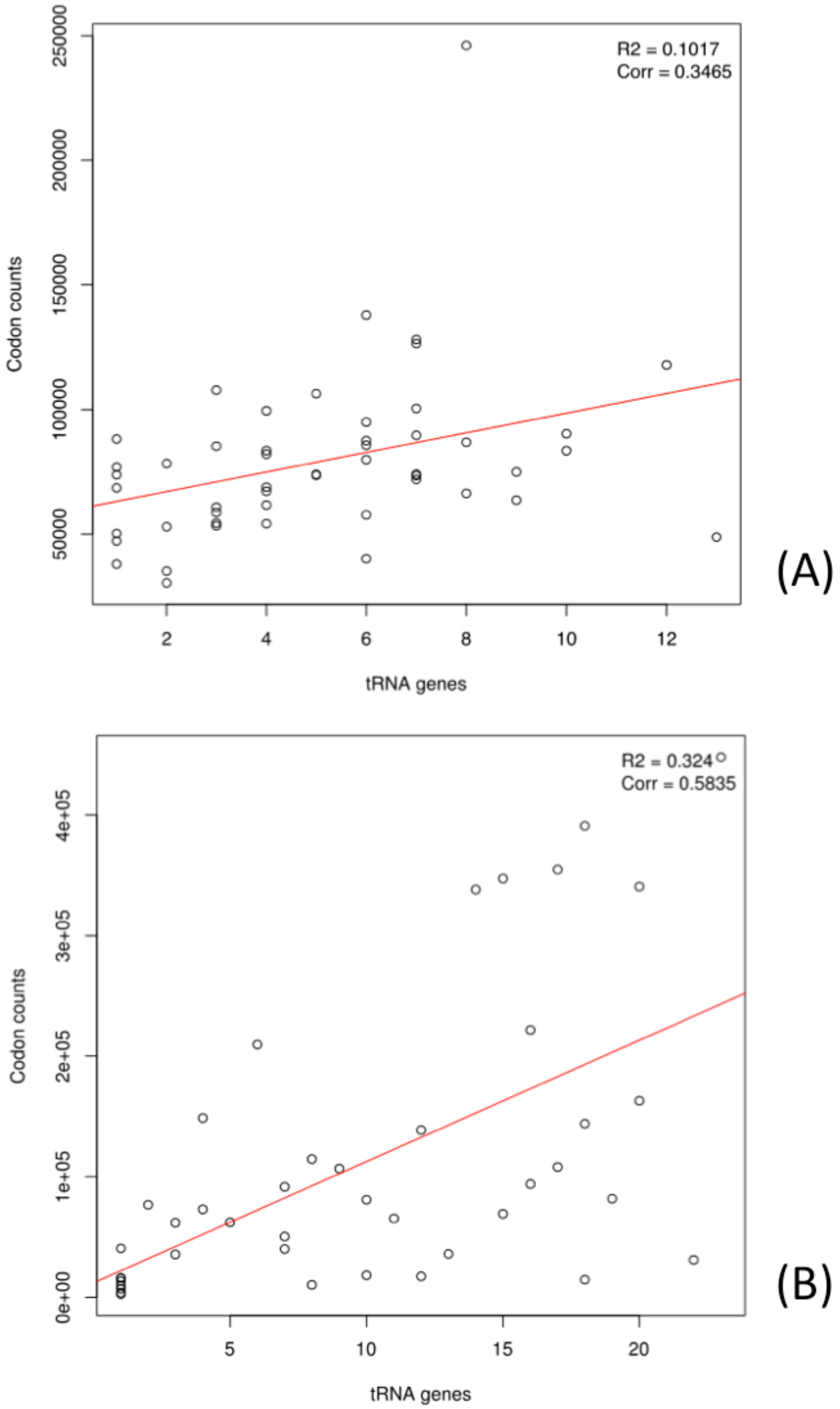


Figure 37. Regression analysis of tRNA genes and codon counts in Mycetozoans. Displayed here are calculations based on the codons and tRNAs from the *Physarum* (A) and *D.discoideum* (B) genomes. A positive correlation was observed in both cases (R^2 = 0.10 for *Physarum* and R^2 = 0.32 for *Dictyostelium,* respectively).

Table 68. Summary of biological process ontologies in *Physarum*. The top 20 gene ontologies for the biological process category are listed, which were obtained from the analysis of the UniProt annotations with the GOSlimViewer server (McCarthy et al. 2006b). Gene counts for each id and descriptions are displayed.

GO id	*GO description*	*Count*
GO:0008150	Biological process	2529
GO:0007165	signal transduction	513
GO:0006464	cellular protein modification process	415
GO:0034641	cellular nitrogen compound metabolic process	356
GO:0009058	biosynthetic process	356
GO:0006810	transport	333
GO:0044281	small molecule metabolic process	317
GO:0005975	carbohydrate metabolic process	260
GO:0006259	DNA metabolic process	258
GO:0009056	catabolic process	219
GO:0055085	transmembrane transport	174
GO:0006950	response to stress	145
GO:0006520	cellular amino acid metabolic process	140
GO:0006412	translation	124
GO:0006629	lipid metabolic process	112
GO:0034655	nucleobase-containing compound catabolic process	107
GO:0016192	vesicle-mediated transport	73
GO:0006457	protein folding	55
GO:0006399	tRNA metabolic process	52
GO:0065003	macromolecular complex assembly	39

Figure 38 shows that that more *D.discoideum* genes are linked to cell differentiation- related ontologies (such as death, developmental process, growth, locomotion, and reproduction). The proteasome regulator and cell killing ontologies were located only in *D.purpureum*, but no GOs were counted solely for *Physarum*. The remaining ontologies show similar distributions in all species. Later, to study the alterations in metabolic maps, the proteins from *Physarum* and dictyostelids were mapped to the KEGG database with the KAAS server (Kanehisa et al. 2008; Moriya et al. 2007). The version 1.67x, the method "bidirectional best hit," and the GENES subset from KEGG were employed to this end.

Table 69. Summary of biological process ontologies in *Dictyostelium discoideum*. The method to obtain these results, and the meaning of the columns follow the same convention as the table 68.

GO id	*GO description*	*Count*
GO:0008150	Biological process	13124
GO:0006810	transport	3869
GO:0016192	vesicle-mediated transport	2612
GO:0034641	cellular nitrogen compound metabolic process	2399
GO:0009058	biosynthetic process	2165
GO:0044281	small molecule metabolic process	1416
GO:0007165	signal transduction	1078
GO:0006950	response to stress	1038
GO:0006464	cellular protein modification process	949
GO:0009056	catabolic process	923
GO:0006259	DNA metabolic process	798
GO:0006629	lipid metabolic process	625
GO:0006412	translation	511
GO:0006520	cellular amino acid metabolic process	481
GO:0000003	reproduction	474
GO:0034655	nucleobase-containing compound catabolic process	449
GO:0048856	anatomical structure development	431
GO:0005975	carbohydrate metabolic process	416
GO:0007010	cytoskeleton organization	328
GO:0055085	transmembrane transport	259

Outputs from the KEGG mappings were then plotted to the reference metabolic map, on the ipath server version 2 (Yamada et al. 2011), for each *Dictyostelium* species against the *Physarum* proteins (Figure 39). By comparison with the generic metabolic map (Figure 26A), it was noticed that in both *Physarum* and dictyostelids, the terpenoid, polyketide and secondary metabolites pathways, as well as the glycan metabolism, are poorly represented. Besides, the fatty acid biosynthesis reactions are present only in dictyostelids (Figure 39; black box at the center left).

Table 70. Summary of biological process ontologies in *Dictyostelium purpureum*. The method to obtain these results, and the meaning of the columns follow the same convention as the table 68.

GO id	*GO description*	*Count*
GO:0008150	Biological process	2700
GO:0006464	cellular protein modification process	669
GO:0009058	biosynthetic process	641
GO:0006810	transport	638
GO:0034641	cellular nitrogen compound metabolic process	483
GO:0044281	small molecule metabolic process	315
GO:0007165	signal transduction	287
GO:0006412	translation	280
GO:0006259	DNA metabolic process	210
GO:0006520	cellular amino acid metabolic process	207
GO:0006629	lipid metabolic process	187
GO:0005975	carbohydrate metabolic process	160
GO:0009056	catabolic process	147
GO:0006399	tRNA metabolic process	115
GO:0006950	response to stress	92
GO:0006457	protein folding	68
GO:0051276	chromosome organization	55
GO:0051186	cofactor metabolic process	42
GO:0016192	vesicle-mediated transport	39
GO:0007155	cell adhesion	37

Conversely, the urea cycle, which is associated to the arginine synthesis and is used by mammals and fish to remove excess nitrogen, is displayed only in *Physarum* (Figure 39; black box at the bottom right). The lack of this and other amino acid synthetic pathways in *Dictyostelium* has been confirmed both by computational and experimental approaches (Payne and Loomis 2006), and shows the evolutionary divergence at the metabolic level of the dictyostelids from the *Physarum* genus. Hence, a more detailed study of the metabolism of *Physarum* is needed to firmly establish the differences in the metabolism between the slime mold and the dictyostelids.

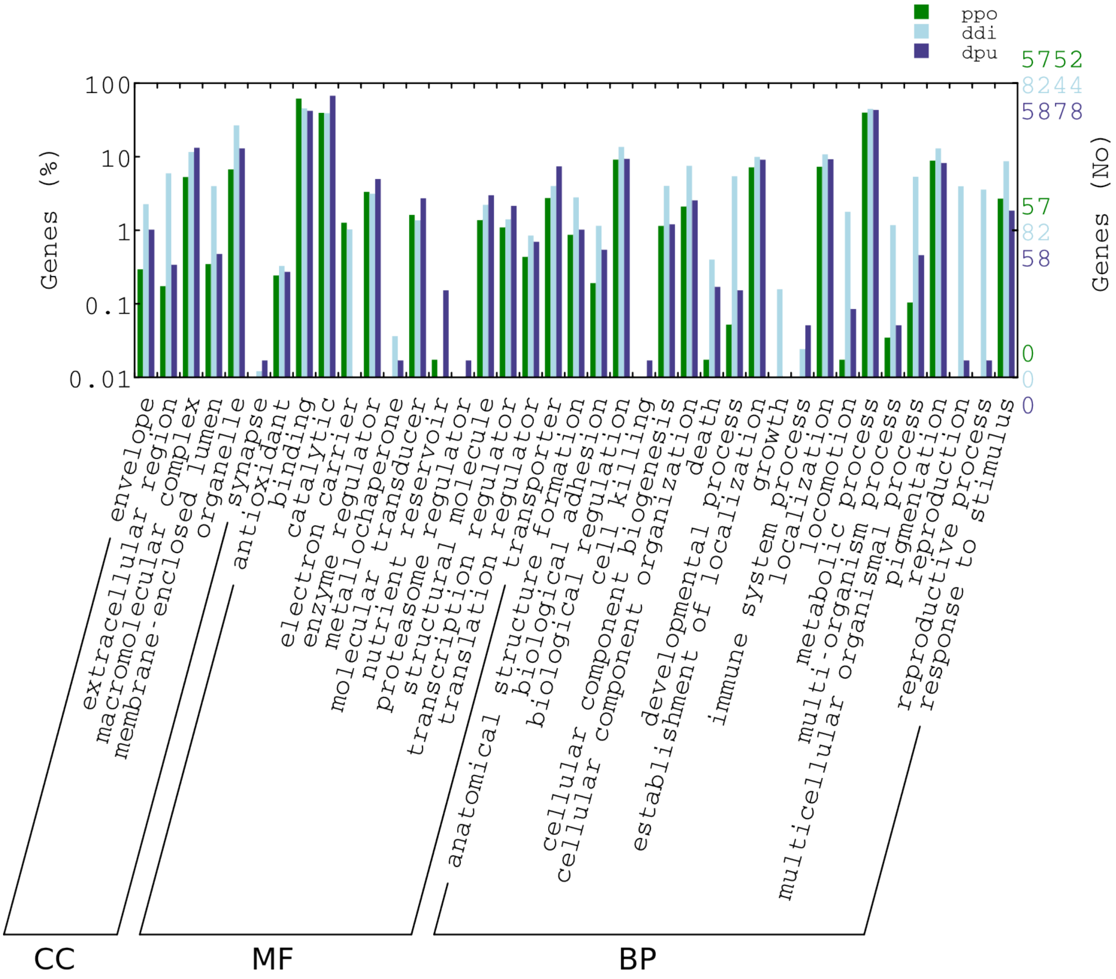


Figure 38. Comparison of gene ontologies between *Physarum* and dictyostelids. Level 3 ontologies were plotted, where the green bars belong to the *Physarum* genes, light blue for *D.discoideum*, and violet for *D.purpureum*. CC, MF and BP correspond to the ontology categories (cellular component, molecular function, and biological process, respectively). The vertical values on the left indicate the percentage of genes, while at the right they denote the gene number. The graphic was plotted using the WEGO tool (Ye et al. 2006).

Then the ESTs from *D.discoideum* were compared against the reference proteome of *Physarum*, using the blastx algorithm from blast (Altschul et al. 1990). These EST sequences comprised 163,182 sequences, that were obtained from dictybase (Chisholm et al. 2006). In this case 259,791 matches were found (68,205 unique), with 45,635 cDNA matches over 70% identity. These dictyostelid matches represent 4,574 *Physarum* protein models (17.83%).

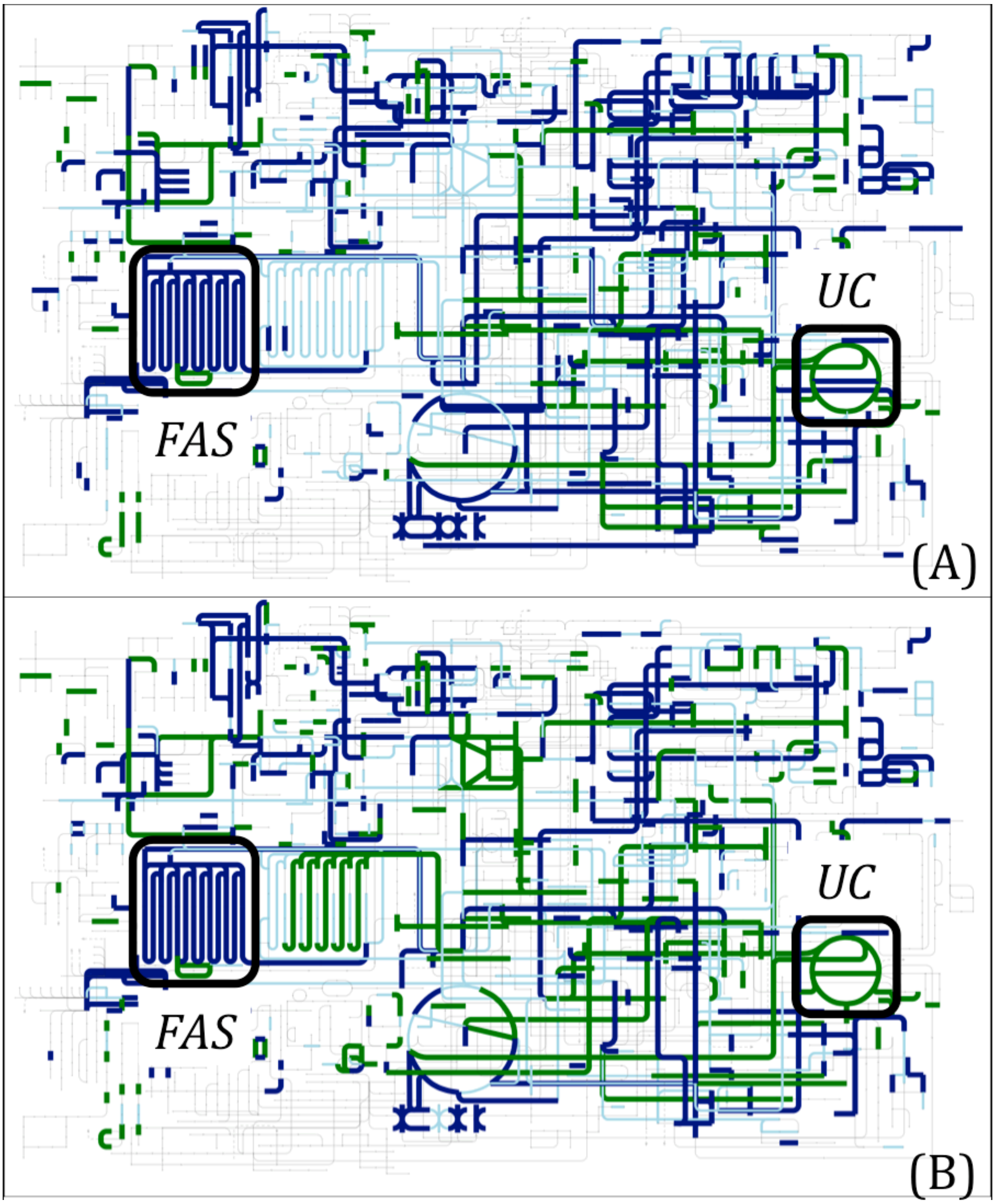


Figure 39. The reference metabolic maps of *Physarum* and dictyostelids. Transcripts were assigned to KEGG orthologs through similarity search, and these were mapped to the primary metabolic pathways with the ipath tool. Above (*A*), comparison of the whole metabolism of *D.discoideum* is displayed; below (*B*), an analogue comparison against *D.purpureum*. The black boxes indicate the fatty acid biosynthesis (*FAS*, left) and the urea cycle (*UC*, right). In both cases, the green lines stand for metabolic reactions in *Physarum*, navy blue for the dictyostelids, and light blue for those reactions that occur in both cases. Data from dictybase and this thesis work.

Furthermore, the dictyostelid genomes and proteomes were also compared with the core eukaryotic gene (CEGMA) datasets (Parra et al. 2007; Parra et al. 2009). For this, first the genomic scaffolds from *D.discoideum* and *D.purpureum* were obtained, and mapped the two CEGMA sets to these dictyostelid genomes (Table 71) with the genblasta program (She et al. 2009), as previously described (Wang et al. 2011). Here it was observed that the dictyostelid genomes have a very large coverage of the CEGMA sets (over 92%; Table), as opposed to the *Physarum* genome, whose matches to the core eukaryotic genes range from 54.03 to 58.52% (Table 60). Afterwards, a similarity search of the dictyostelid proteomes was performed versus the most recent CEGMA dataset (Parra et al. 2009), with the blastp algorithm of blast (e-value 1E-6, over >= 70% identity coverage). Here, the reference proteome of *Physarum* was found to cover more core eukaryotic proteins than the dictyostelid proteomes, with up to 17.34% orthologs of the CEGMA proteins in *Physarum*, and less than 9% for both species of the genus *Dictyostelium* (Table 72). These apparent opposite results between the genome and proteome mappings might be due to curation and annotation of the dictyostelid genes.

Later, the OrthoMCL server (Li et al. 2003) was employed to find groups of unique and conserved ortholog genes in the *Physarum* and dictyostelid proteomes, with default parameters. This program uses the similarity search (blastp, e-value 1E-5 and 50% identity match; Altschul et al. 1990) against a database of conserved proteins (OrthoMCL-DB). Queried proteins that are reported above the cutoff, are assigned to the respective ortholog group, otherwise they are classified as “no group.” Here, almost all proteins from *D.discoideum* (12,272 matches, 99.64%) were assigned to ortholog groups, while *D.purpureum* and *Physarum* had proportionally less assignments (71.35 and 36.33% respectively; Table 73). This might be because the proteins from dictybase were a primary source for building the OrthoMCL ortholog groups. Later, the phylogenetic patterns of species in ortholog groups from dictyostelids and *Physarum* was obtained. As for *D.discoideum*, 12,261 of the mapped orthologs belong to the species itself, and its taxonomic representation included other nine species, with 2 or 1 orthologs. Conversely, the phylogenetic patterns of *D.purpureum* and *Physarum* entailed many species, and with similar frequency distributions (Figure 40). Together, these results reflect the fact that both proteomes, of *D.purpureum* and *Physarum*, used common sources for gene modeling and annotation, particularly the inclusion of the *D.discoideum* data.

Finally, an evaluation of how related is the *Physarum* proteome to other eukaryotic proteomes was performed. For these analyses, the SimiTri program (Parkinson and Blaxter 2003) was utilized as previously described (Peregrín-Alvarez and Parkinson 2009; Wang et al. 2007). This Java application allows the simultaneous comparison of similarities of a given query, to three different sequence databases, and the visualization of the evolutionary relationships between these sequence sets. The outputs of similarity searches were employed as inputs, between the predicted protein models of *Physarum*, and the proteome sets from *D.discoideum* and *D.purpureum* (Chisholm et al. 2006; Parikh et al. 2010), the best filtered protein models from the choanoflagellate *Monosiga brevicollis* (King et al. 2008; Grigoriev et al. 2012), the translations of curated ORFs from the yeast *Saccharomyces cerevisiae* (Cherry et al. 2012), and the most recent core eukaryotic gene (CEGMA) dataset (Parra et al. 2009). The similarity searches were carried out with the blastp algorithm of blast (Altschul et al. 1990; Altschul et al. 1997), using an e-value of 1E-6. The outputs were adapted to SimiTri with a combination of several in-house bash and perl scripts, and plotted in groups of three proteomes: (*i*) *D.discoideum*, *D.purpureum*, and yeast; (*ii*) *D.discoideum*, *Monosiga*, and yeast; and (*iii*) *D.discoideum*, CEGMA proteins, and yeast (Figure 41). First, when comparing the *Physarum* proteome with those of the dictyostelids and yeast, the slime mold sequences group either in the center, or closer to those from *D.discoideum* or yeast (Figure 41A). Replacement of the *D.purpureum* with the choanoflagellate proteome, group the matches either in the center or closer to *D.discoideum* (Figure 41B). Finally, the substitution of the *Monosiga* proteome with the core eukaryotic gene (CEGMA) dataset, resulted in the *Physarum* sequences clearly aligning with at the *D.discoideum* and yeast side, with most proteins mapping at the dictyostelid corner (Figure 41C). Also it is interesting to noticed that in two of the SimiTri plots, a clearly detached group of dots are closer to the yeast proteome vertex (Figure 41A and 41C). Separate examination of *Physarum* genes corresponding to these dots showed that most of these are fungal homologs linked to the primary metabolism: The acetolactate synthase ILV2, the R export factor ELF1, the fatty acid synthase subunit alpha FAS2, and the sulfite reductase subunit beta SIR1 (Table 74). Annotations of these proteins map to the amino acid (ILV2, gltB, and SIR1) and fatty acid biosynthesis (FAS2); and three of them possess oxidoreductase activities (FAS2, gltB, and SIR1). These results are in agreement with previous observations in this thesis work, regarding differences in the fatty acid synthesis and amino acid metabolism between the *Physarum* and dictyostelid genomes (Figure 39).

Table 71. Mapping CEGMA datasets to dictyostelid genomes. For this search GenBlastA was used with default settings. The protein matches listed and the percentages follow the same convention as in the Table 60.

CEGMA Dataset	248		Core	
Species	Matches	Percentage	Matches	Percentage
D.discoideum	237	95.56	431	94.10
D.purpureum	234	94.35	424	92.58

Table 72. Mapping CEGMA datasets to *Physarum* and dictyostelid proteomes. For this search I used the most recent core eukaryotic genes dataset (CEGMA; 248 entries). The protein matches listed below (over 70% of identity coverage), and the percentages follow the same convention as in the Table 60.

Proteome	Matches	Percentage
Physarum	43	17.34
D.discoideum	21	8.47
D.purpureum	19	7.66

Table 73. OrthoMCL analysis of the *Physarum* and dictyostelid proteomes.

Proteome	Sequences in Proteome	Proteins assigned	No Group	Unique matches	Represented species
D.discoideum	12,316	12,272	2,403	11,754	10
D.purpureum	12,410	8,854	1,188	7,027	108
Physarum	25,649	9,318	345	7,568	146

Table 74. SimiTri yeast matches.

Physarum Gene id	UniProt id	Annotation
maker-Scaffold361-est_gff_Cufflinks-gene-0.3-mRNA-1	ILVB_CRYNV	Acetolactate synthase, ILV2
maker-Scaffold370-est_gff_Cufflinks-gene-1.6-mRNA-1	ELF1_SCHPO	R export factor ELF1
maker-Scaffold152-est_gff_Cufflinks-gene-0.10-mRNA-2	FAS2_CANAX	Fatty acid synthase subunit alpha, FAS2
maker-Scaffold1925-est_gff_Cufflinks-gene-0.0-mRNA-1	GLTB_BACSU	Glutamate synthase small chain, gltB
maker-Scaffold251-est_gff_Cufflinks-exonerate_est2genome-gene-1.0-mRNA-1	MET5_SCHPO	Sulfite reductase subunit beta, SIR1

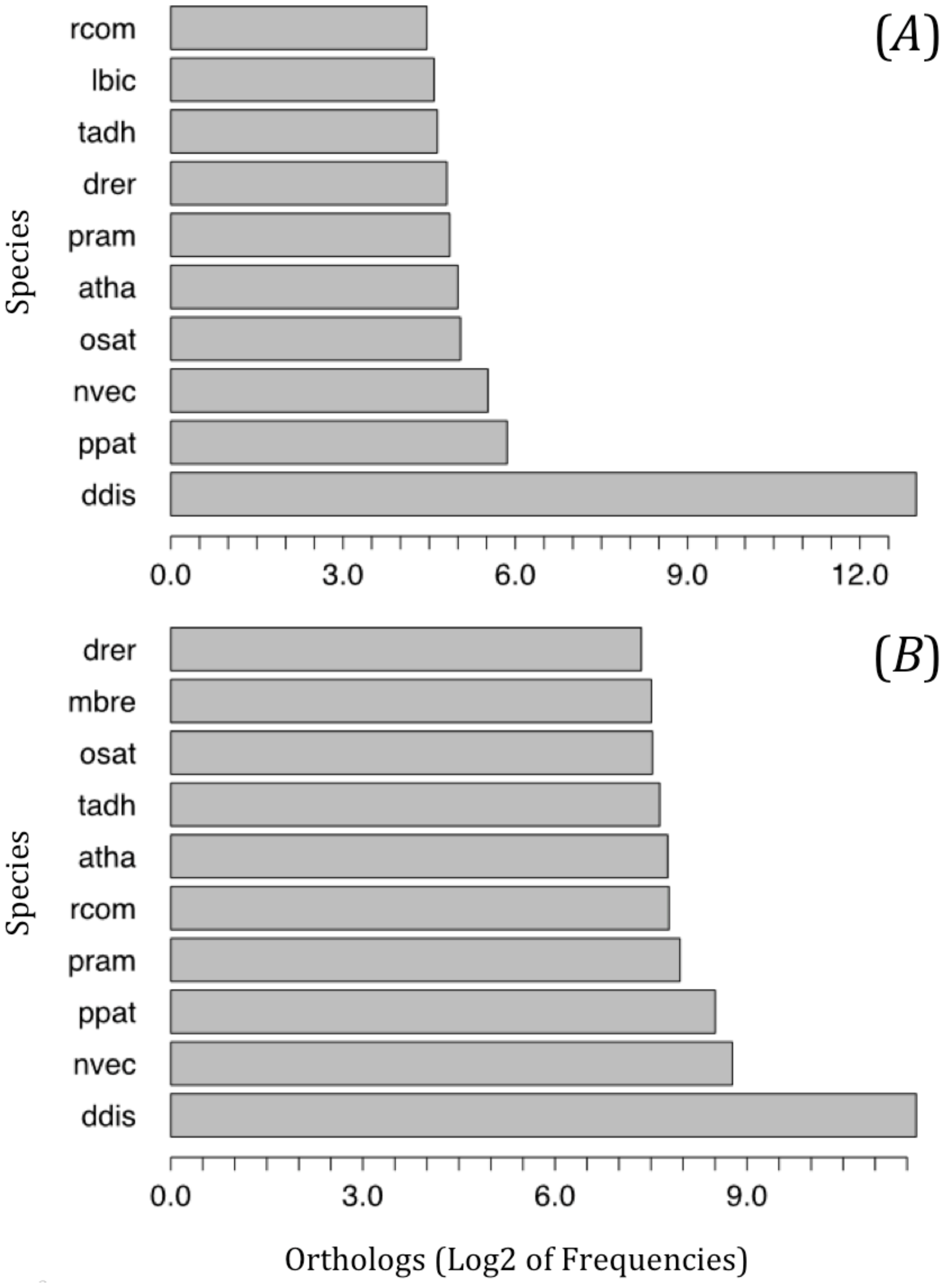


Figure 40. Phylogenetic pattern of *Dictyostelium purpureum* and *Physarum*. The proteomes of *D.purpureum* (*A*) and *Physarum* (*B*) were searched for conserved ortholog genes with the OrthoMCL server, with default values. The plot shows the top ten species with the most orthologs, whose frequencies are in logarithmic (Log2) scale. Species listed include *Dictyostelium discoideum* (ddis), the fungus *Phytophthora ramorum* (pram) and *Laccaria bicolor* (lbic), the choanoflagellate *Monosiga brevicollis* (mbre), the placozoon *Trichoplax adhaerens* (tadh), the sea anemone *Nematostella vectensis* (nvec), the zebrafish *Danio rerio* (drer), and the plants *Physcomitrella patens* (ppat), *Ricinus communis* (rcom), *Arabidopsis thaliana* (atha), and *Oryza sativa* (osat).

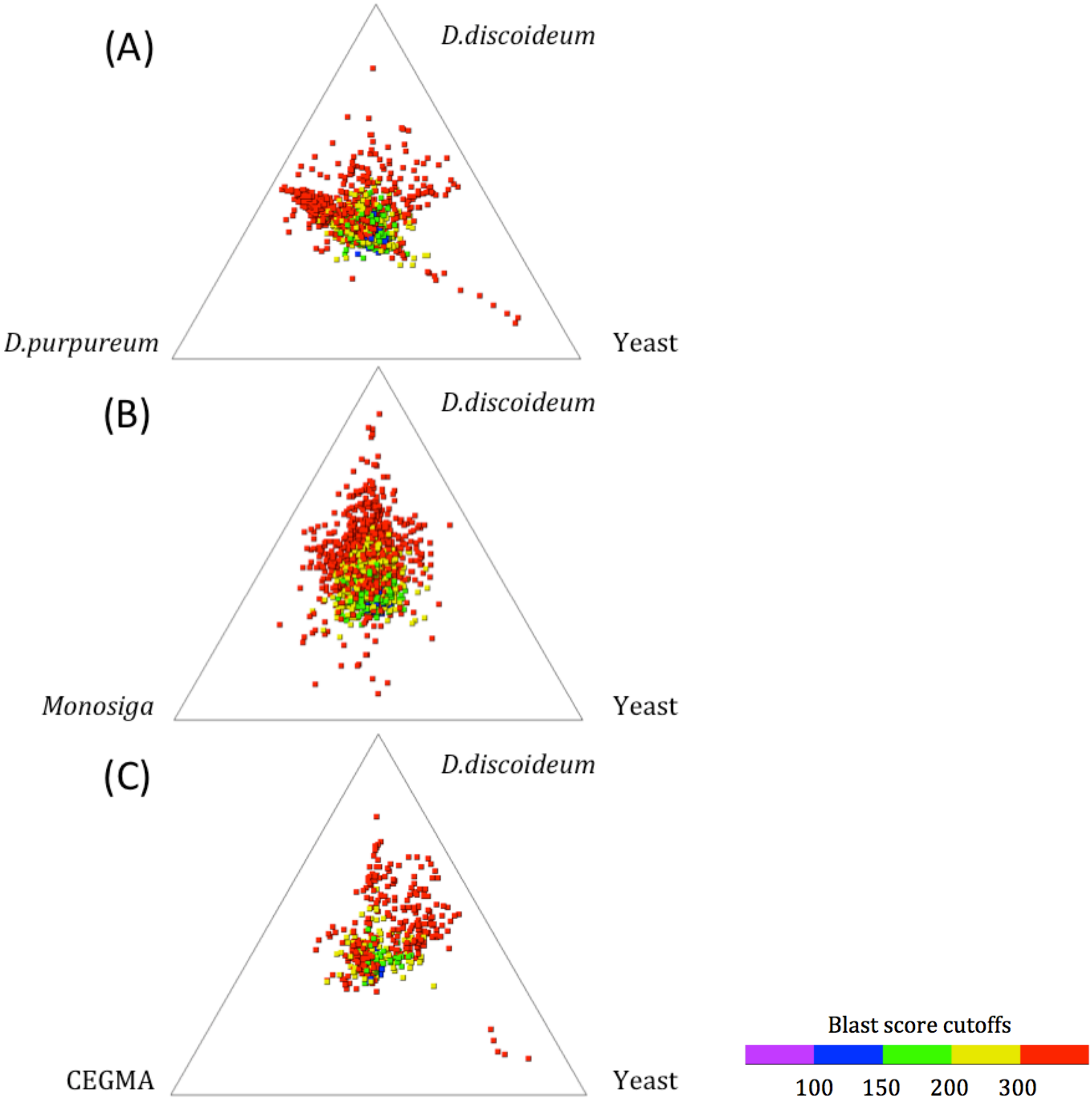


Figure 41. SimiTri profiles of the *Physarum* protein models. The reference proteome of *Physarum* was searched for similarity against the protein datasets from *Dictyostelium discoideum*, *D.purpureum, Monosiga brevicollis,* the yeast *Saccharomyces cerevisiae,* and the set of core eukaryotic genes (CEGMA), with blastp (e-value 1E-6). Outputs from the blastp alignments were then loaded and plotted in sets of three proteomes (*A* – *C*), with the SimiTri application (Parkinson and Blaxter 2003). The position of each dot represents its similarity to a given protein set, specified in blast scores (Altschul et al. 1990), and their color is coded according to the highest of these blast scores.

Discussion

Genome Annotation. The draft genome assembly of this *Physarum* genome release comprises 239.75 Mb (Table 22). This value is below the latest expected size of 300 Mb (Glöckner et al. 2008). Early experiments showed that the amount of DNA per nucleus (*C*-values) is between 0.25 and 0.3 pg. (Mohberg and Rusch 1971; Mohberg 1977). Using the equivalence of *C*-values or masses to base pairs (Dolezel et al. 2003), in this work it was estimated that the genome ranges between 244.5 and 293.4 Mb, with the first value fitting closely the obtained size for the working draft.

Respect to the noncoding fraction of the genome, this assembly has a high GC-content (41.16%), with a repeat content of 14.55% (27.59% using a *Physarum*-specific library). Most repetitive elements found are simple repeats (5.47% of the total assembly) and low complexity regions (8.9%). Furthermore, 1,436 noncoding RNA genes were also identified, most of them pertaining to the ribosomal RNA, microRNA, small nucleolar RNA and transfer RNA families (928, 777, 144 and 347 genes, respectively; Tables 24, 25 and 32). Selenocysteine tRNAs were found within the tRNA gene set, and more importantly, all twenty standard amino acids (Table 29). However, some predicted noncoding RNAs might be sample contaminants, particularly those annotated as bacterial small RNAs (Table 27).

As for the coding regions of the genome, three main sources were used to search for transcripts: ESTs, RNA-seq, and well-annotated proteins from the UniProt database. First, the clustering of all available cDNA data (Watkins and Gray 2008; Glöckner et al. 2008) together with the transcriptome obtained by 454 sequencing (Chapter 3), produced a nonredundant dataset of 17,931 coding sequences. Later, to use the RNA-seq outputs in gene modeling, a previous mapping of the short reads against the genome was required. This resulted in 36.18 to 67.62% mapped reads (Table 34). Why an average of 49.03% of reads do not map to the assembly might be due to several reasons, *e.g.* (*i*) the generation of chimeric sequences during the PCR amplification previous to the Illumina sequencing; (*ii*) the quality and coverage depth of the RNA-seq output (which is varies between samples and strains); (*iii*) the alignment method use, and the criteria for these alignments; (*iv*) many reads align to the mitochondrial genome, which is in a large copy number excess, relative to the genome; (*v*) a percentage of reads map to microRNA precursors and several types of 5' or 3' end untranslated regions (UTR), such as promoters, spliced exons, *etc*.; and (*vi*) the number of mutations, paralogs or CNVs (copy number variations) between different strains (Hansey et al. 2012; Mortazavi et al. 2008). In future RNA-seq experiments (with increased depths), a better correspondence between the read mapping and the predicted transcripts is expected.

Afterwards, the combination of the TopHat and Cufflinks programs generated approximately from 25 to 82 thousand genes from these mappings (Table 35). The obtained wide range might be due to the difficulty in obtaining full-length cDNAs from short-read high-throughput sequencing experiments, although other factors, such as post-transcriptional modifications that occur in this species might be contributing to the fragmentation. In spite of this, the number of reads per Illumina run in the RNA-seq outputs from the WT31 and LULU2 samples was within the considered optimal range to generate a representative *de novo* assembly (20 – 30 millions; Francis et al. 2013), and therefore these datasets represent a reliable source for protein- coding gene modeling. The average transcript length ranged between 322 - 1,379 bp, which is well below the expected value of 3,743 bp (see *Inference of Protein Models*). For the protein-coding gene modeling, the RNA-seq outputs required a previous conversion, utilizing the cufflinks2gff3 tool from MAKER2, leaving a transcript range of 8 – 51 thousand protein-coding genes (Table 35). The reason of this lower number of transcripts is that by default this tool will ignore features that correspond to single exon models, because these could pertain to repetitive elements and pseudogenes (Holt and Yandell 2011). In the end, 39,539 transcripts intervals were shared by all Illumina sequencings. These results suggest that the number of protein-coding genes in *Physarum* predicted by RNA-seq mapping might be overestimated, while the shorter than expected transcript lengths would eventually fuse into larger cDNAs, thus diminishing the final protein-coding gene number.

The EST and RNA-seq data, together with proteins from UniProt and cDNAs from *Dictyostelium discoideum*, were then used as evidences to predict the protein coding gene models, which resulted in 25,649 transcripts identified in 5,422 unique genomic scaffolds. The total transcript extension is 15.43 Mb, *i.e.* 6.44% of the genome is coding (Table 38). Most of these transcripts came from RNA-seq evidences alone (22,315 sequences or 87%), while 428 had both EST and RNA-seq previous data. In this respect, and as a manner of experimental control, an equation based on the relationship between gene numbers and genome sizes was used, to estimate the expected number of protein coding genes in *Physarum* (Hou and Lin 2009). This calculation yielded 38,188 genes (Appendix 2), which is larger than the number of obtained models (25,649 transcripts), almost the double of the last estimation (20 thousand genes; Glöckner et al., 2008), but close to the number obtained by the RNA-seq mappings (39,539 transcripts). The discrepance indicates not only that the gene number versus genome size equation does not apply to *Physarum*, but also that most genes assembled from RNA-seq mappings must have fused into longer transcripts in the final gene set. In addition, an estimate the transcriptome size using the distribution of *k*-mers in the sequencing outputs (Marçais and Kingsford 2011) was also attempted, but the *k*-mer distributions showed no peak in any of the samples separately, or combining all of them into a single source (data not shown), and therefore no prediction of the transcriptome size was obtained in this manner.

Then, the predicted protein- coding genes were annotated using several sources of biological information, which resulted in 4,915 sequences associated to UniProt homologs, 5,752 with gene ontology annotations, 15,914 containing InterPro domains, and 2,066 linked to KEGG orthologs; 1,629 transcripts were annotated at all these levels. The most common species in the UniProt orthologs was *Dictyostelium discoideum*, reflecting the high degree of annotation of the genome of this species. Then, in order to study those genes involved in cell differentiation, genes linked to this gene ontology were selected (GO:0030154; 432 genes). The encoded proteins were then separated into three groups, one with those associated to the "embryo development" ontology (GO:0009790; 40 unique proteins), another with those with the "signal transduction" annotation

(GO:0007165; 111 proteins), and a third group with those lacking these two gene ontologies. Within these three groups, potential protein – protein interactions were searched, and certain groups inside these large interactions networks were classified as macromolecular complexes (Table 40). From the associated ontologies, it was observed that these complexes feature distinctive biological functions, and therefore they constitute valuable candidates to study different aspects of the regulation of the cell differentiation in this organism in future studies.

Validation and Completeness. To evaluate whether this assembly is a reliable source for gene annotation, several measures of completeness were used (Yandell and Ence 2012). First, the current assembly was compared against older versions, and found that this release contains fewer gaps than its predecessors (Table 51). The N50 scaffold is also larger than in former releases (97.38 Kb), a value also greater than our estimations for the average gene size (3,743 bp; see *Inference of Protein Models*). This result secures that more than 50% of the genes will be contained in a single scaffold; otherwise additional sequencing would be required to extend the N50 scaffold length (Yandell and Ence 2012). Then, the genome was checked for contaminants, and found 771 sequences matching bacterial and archaeal genome. These entries should be removed to avoid false annotations in future releases.

Afterwards, several coding sources were mapped against the genome assembly: ESTs, RNA-seq short and long reads, GenBank sequences and CEGMA datasets. Most ESTs (17,577 cDNAs, or 98.03%), GenBank sequences (231 nucleotide entries, or 91.3%) and RNA-seq long reads (over 98%; Table 55) were matched in the genome. Unmatched sequences might have been lost due to the fragmentation of the genome. In a similar manner, most CEGMA proteins were found in the genome: 98.47% of the core and 98.79% of the most recent dataset. In parallel, a CEGMA analysis was performed with its original pipeline, a procedure recommended for highly divergent genomes. This resulted in protein conservation ranging between 33.33% to 55.74% of the core eukaryotic genes, and therefore this genome release can be considered either incomplete or divergent (Parra et al. 2007; Parra et al. 2009; Yandell and Ence 2012).

Then, the annotated gene models were compared against *ab initio* gene predictions. In this case, the results show that the evidence- based gene identification produced less false positive exons, while increasing the length of these coding fragments (Table 56); similarly, the number of introns decreased, while at the same time being longer. The sensitivity increased up to 30-fold, and the specificity reached over 100-fold, when comparing the second run of MAKER2 with the GeneMark *ab initio* predictions (Table 57). The support of annotations is 17 times larger in the second MAKER2 iteration than in the GeneMark predictions. Therefore, the method used for gene identification is validated for this genome release, as being more sensitive, more specific, and more supportive of annotations that standard *ab initio* gene finding procedures.

Furthermore, two more measures of completeness were employed with the annotated genes: the number of encoded tRNA genes, and the percentage of detected domains. First, a complete genome is expected to encode for all standard amino acids for protein translation; therefore, I searched for the tRNA genes, and found all those coding for the twenty standard amino acids (Table 29). Later, a measure of annotation quality can be obtained by calculating the percentage of annotated proteins with known domains from the InterPro (Hunter et al. 2009) or PFAM (Finn et al. 2008) databases. For example, the domain content in well annotated model organisms such as human, mouse and *Drosophila* range from 57 to 75% (Yandell and Ence 2012). Here it was observed that 25,649 predicted protein- coding gene models included 7,080 sequences (27.60%) that possessed PFAM domains, which is just over the lower threshold for poor gene predictions (5 – 25% PFAM content; Yandell and Ence 2012). However, as these gene models were obtained using default annotation distance (AED) values, it is expected that the proportion of genes with PFAM domains will increase with higher AED thresholds, and these parameters should be incorporated in future annotation releases.

Comparative Genomics. The *Physarum* genome is 4.73 and 7.27 times larger than its *D.discoideum* and *D.purpureum* counterparts, and almost doubling their GC-content – 41.16% in *Physarum* against 21.99% and 24.47% in *D.discoideum* and *D.purpureum*, respectively. However, the number of undefined bases is over a thousand fold larger in *Physarum* than in these dictyostelids (Table 63). On the other hand, the N50 value is in all cases larger than the average gene size, with larger N50 scaffold values in *D.purpureum* and *Physarum*, reflecting the more recent sequencing technologies used (see page 164).

Afterwards, the repetitive sequences of *Physarum* and dictyostelids were compared. First, it was noticed that, although the extent of masked bases is larger in *Physarum*, the percentage of bases masked is greater in *D.discoideum* than in *Physarum* (Tables 23 and 64). This might be due to the use of the default RepBase database (Jurka et al. 2005), which contains more repetitive elements from dictyostelids than from those discovered in *Physarum*. It was also found that the length of repetitive elements is larger in *Physarum* than in *D.discoideum* and *D.purpureum*, ranging from 52- (LTR transposons) to more than a hundred fold (DNA elements). Furthermore, the *Physarum* genome is the only of the three Mycetozoans analyzed that contains satellites. These results must be taken cautiously, however, as these sequences were not found when using a custom library of repetitive sequences (Table 23). All these results so far suggested that *Physarum* might form a separate clade within the Mycetozoans; this hypothesis was tested and verified through the multiple alignments of conserved coding blocks (Figure 36).

Later, it was decided to further contrast the *Physarum* and the dictyostelid genomes in terms of their coding regions. First, the number of protein coding genes in *Physarum* doubled those in *D.discoideum* and *D.purpureum* (25,649 versus ~12 thousand genes; Table 65). However, less than 24% of the *Physarum* transcripts were predicted as coding (23.88%, Table 62), while most *D.discoideum* genes were predicted as such (90.33% coding). These results could be linked to the fact that the method used (CPC; Kong et al. 2007) employs similarity to annotated proteins, and because *D.discoideum* possess more annotated proteins in UniProt, therefore more sequences will be predicted as coding. In a similar manner, both dictyostelids have larger proteome coverages in the CEGMA proteins (larger than 92%, Table 71), and most of the *D.discoideum* proteins were also assigned to orthologs groups with OrthoMCL (12,272 matches or 99.64%; Table 73). The results of the noncoding prediction with CPC, the mapping versus CEGMA, and the OrthoMCL classification showed that the *D.discoideum* genome was employed as a primary source of annotations in these databases. Furthermore, the Gene Ontology analyses proved that a larger proportion of the protein- coding genes is associated to signal transduction in *Physarum* than in other dictyostelids, while *D.discoideum* and *D.purpureum* displayed more genes related to transport ontologies. Other ontologies linked to cell differentiation were found mostly or exclusively in dictyostelids, reflecting the larger degree of experimental annotation of these two genomes (Figure 38). The proteins encoding metabolic enzymes and their reactions in *Physarum* also differ with those from the dictyostelids: In this work, the urea cycle reactions were only observed in the slime mold, and the fatty acid biosynthesis reactions exclusively in dictyostelids (Figure 39). These results are in agreement with the SimiTri analyses early in this chapter, which showed that most proteins cluster closely to dictyostelid proteins (Figure 41), except for a small group of metabolic proteins that are highly similar to their yeast counterparts (Table 74). However, these predicted differences would require experimental studies for confirmation.

Conclusions

This study provides the first genomic survey of the slime mold *Physarum polycephalum*. These novel deep RNA sequencings, together with formerly obtained cDNAs, support a reference transcriptome of 25,649 encoded nucleotide sequences. In addition, other major RNA families were mapped. These analyses contribute the necessary basic knowledge to understand the mechanisms of cell differentiation in this organism, especially through the characterization of networks and complexes specific to these molecular functions. Furthermore, it provides a starting point for further exploration of the biology of *Physarum*, and its utility as a model organism. Aside from the genome and transcriptomic sequences and their analyses, this study also offers a working pipeline and annotation protocols, which can be taken as a blueprint for the analysis of future genome releases.

Summary

In this chapter, the analysis of the first draft of the genome of the slime mold *Physarum polycephalum* (NCBI accession 709848) was presented. This genomic assembly entails 239.75 Mb (scaffold N50: 97.38 Kb). The genome is high on GC-content (41.16%) and repetitive sequences (14.55 to 27.59%, depending on the repeat library used).

Novel RNA sequencings (RNA-seq) of several strains, sample types (cell pools, single cells), growth conditions (starved, sporulation- induced, *etc.*), and different time points of the sporulation cycle were also carried out. These data, combined with previous RNA-seq studies from this thesis work (see Chapters 3 and 4), and formerly published EST sequencings from different cell stages (plasmodium, amoeba) support a total of 25,649 transcripts. 4,915 of these sequences were associated to UniProt homologs, 5,752 to gene ontologies, 15,914 to InterPro domains, and 2,066 linked to KEGG pathway orthologs. No automatic annotations or predictions were used as evidences for finding protein-coding genes. Genes annotated for the cell differentiation (GO:0030154) were joined into interaction networks, including subsets involved in signal transduction and development. Protein complexes within these networks were also identified. In addition, complete sets of 347 transfer, 928 ribosomal and other 161 noncoding RNAs, were also mapped in the genome.

The genome annotation is validated through mapping of *Physarum*- specific coding evidences (EST and RNA-seq data) and sets of core eukaryotic genes. Furthermore, tRNA genes for all twenty standard amino acids were found, and the protein domain content (27.6%) is within the range of reliable gene identifications.

Compared to the dictyostelid genomes, the *Physarum* genome is larger and richer in GC-content and repetitive sequences. The number of protein- coding genes is twice as large in *Physarum* than in *D.discoideum* and *D.purpureum*, with more genes annotated for the signal transduction ontology in *Physarum*, while more genes linked to transport and cell differentiation ontologies were found in dictyostelids. Annotations pertaining to metabolic pathways also support

considerable differences between *Physarum* and the dictyostelids, although these predictions require experimental confirmation.

As far as the literature shows, this is the first global analysis of the genome of *Physarum* and its encoded genes, offering valuable information that adds to the current knowledge of the slime mold biology. Furthermore, this work also offers novel databanks of transcriptomic sequences and a working annotation pipeline, which can be taken as a blueprint for the analysis of future genome releases.

6. Bibliography

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., … Venter, J. C. (2000). The Genome Sequence of *Drosophila melanogaster*. *Science*, *287*(5461), 2185–2195.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., … Moreno, R. F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, *252*(5013), 1651–1656.

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., … Spiegel, F. W. (2012). The revised classification of eukaryotes. *J Euk Microbiol*, *59*(5), 429–93.

Allen, R. D. (1981). Motility. *J Cell Biol*, *91*(3 Pt 2), 148s–155s.

Altschul, S. F., Madden, T. L., Schäffer, a a, Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, *25*(17), 3389–402.

Altschul, S., Gish, W., & Miller, W. (1990). Basic local alignment search tool. *J Mol Biol, 215*, 403–410.

Alwine, J. C., Kemp, D. J., & Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci USA*, *74*(12), 5350–4.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, *11*(10), R106.

Anderson, R., & Dee, J. (1990). Culture, Development and Genetics of Physarum - A Practical Manual. In *9th European Physarum Conference*. Leicester.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: A tool for the unification of biology. *Nature Genet*, *25*, 25–29.

Aubry, L., & Firtel, R. (1999). Integration of Signaling Networks that regulate *Dictyostelium* differentiation. *Ann Rev Cell Dev Biol*, *15*(1), 469–517.

Audic, S., & Claverie, J. M. (1997). The significance of digital gene expression profiles. *Genome Res*, *7*(10), 986–995.

Aury, J. M., Cruaud, C., Barbe, V., Rogier, O., Mangenot, S., Samson, G., … Wincker, P. (2008). High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics*, *9*(1), 603.

Bader, G., & Hogue, C. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, *4*(1), 2.

Bailey, J. (1995). Plasmodium development in the myxomycete *Physarum polycephalum*: genetic control and cellular events. *Microbiol*, *141 (Pt 1*, 2355–65.

Bailey, J. (1997). Building a plasmodium: Development in the acellular slime mould *Physarum polycephalum*. *BioEssays*, *19*(11), 985–992.

Bailey, J., Cook, L. J., Kilmer-Barber, R., Swanston, E., Solnica-Krezel, L., Lohman, K., … Anderson, R. W. (1999). Identification of three genes expressed primarily during development in *Physarum polycephalum*. *Archives Microbiol*, *172*(6), 364–376.

Baldauf, S. L., & Doolittle, W. F. (1997). Origin and evolution of the slime molds (Mycetozoa). *Proc Natl Acad Sci USA*, *94*(22), 12007–12.

Bao, Z., & Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*, *12*(8), 1269–76.

Behura, S., & Severson, D. (2011). Coadaptation of isoacceptor tRNA genes and codon usage bias for translation efficiency in Aedes aegypti and Anopheles gambiae. *Insect Mol Biol*, *20*(2), 177–187.

Bénard, M., Maric, C., & Pierron, G. (2007). Low rate of replication fork progression lengthens the replication timing of a locus containing an early firing origin. *Nucleic Acids Res*, *35*(17), 5763–74.

Benard, M., Pallotta, D., & Pierron, G. (1992). Structure and identity of a late-replicating and transcriptionally active gene. *Exp Cell Res*, *201*(2), 506–13.

Bénard, M., & Pierron, G. (1992). Mapping of a *Physarum* chromosomal origin of replication tightly linked to a developmentally-regulated profilin gene. *Nucleic Acids Res*, *20*(13), 3309–15.

Benne, R., Van den Burg, J., Brakenhoff, J. P., Sloof, P., Van Boom, J. H., & Tromp, M. C. (1986). Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell*, *46*(6), 819–26.

Benson, D. a, Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2011). GenBank. *Nucleic Acids Res*, *39*(Database issue), D32–7.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2008). GenBank. *Nucleic Acids Res*, *36*(Database issue), D25–D30.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, *27*(2), 573–80.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–9.

Bernier, F., Pallotta, D., & Lemieux, G. (1986). Molecular cloning of mRNAs expressed specifically during spherulation of *Physarum polycephalum*. *Biochim Biophys Acta*, *867*(4), 234–243.

Binette, F., Bénard, M., Laroche, A., Pierron, G., Lemieux, G., & Pallotta, D. (1990). Cell-Specific Expression of a Profilin Gene Family. *DNA Cell Biol*, *9*(5), 323–334.

Bjorkoy, G., Lamark, T., Brech, A., Outzen, H., Perander, M., Overvatn, A., … Johansen, T. (2005). p62/SQSTM1 forms protein aggregates degraded by autophagy and has a protective effect on huntingtin-induced cell death. *J Cell Biol*, *171*(4), 603–614.

Blanchoin, L., Robinson, R. C., Choe, S., & Pollard, T. D. (2000). Phosphorylation of *Acanthamoeba* actophorin (ADF/cofilin) blocks interaction with actin without a change in atomic structure1. *J Mol Biol*, *295*(2), 203–211.

Blanton, R. L. (2001). Slime Moulds. In *Encyclopedia of Life Sciences*. Chichester, UK: John Wiley & Sons, Ltd.

Block, I., Briegleb, W., & Wohlfarth-Bottermann, K. E. (1986). Gravisensitivity of the acellular slime mold *Physarum polycephalum* demonstrated on the fast-rotating clinostat. *Eur J Cell Biol*, *41*, 44–50.

Block, I., Rabien, H., & Ivanova, K. (1998). Involvement of the second messenger camp in gravity-signal transduction in *Physarum*. *Adv Space Res*, *21*(8), 1311–1314.

Block, I., Wolke, A., Briegleb, W., & Ivanova, K. (1995). Gravity perception and signal transduction in single cells. *Acta Astronautica*, *36*(8-12), 479–86.

Block, I., Wolke, A., Briegleb, W., Wohlfarth-Bottermann, K.-E., Merbold, U., Brinckmann, E., & Brillouet, C. (1994). Gravity-related behaviour of the acellular slime mold *Physarum polycephalum*. In H. Oser & T. ~D. Guyenne (Eds.), *Life Sciences Research in Space* (Vol. 366, p. 43).

Bluem, R., Schmidt, E., Corvey, C., Karas, M., Schlicksupp, A., Kirsch, J., & Kuhse, J. (2007). Components of the Translational Machinery Are Associated with Juvenile Glycine Receptors and Are Redistributed to the Cytoskeleton upon Aging and Synaptic Activity. *J Biol Chem*, *282*(52), 37783–37793.

Bluthgen, N., Brand, K., Cajavec, B., Swat, M., Herzel, H., & Beule, D. (2005). Biological profiling of gene groups utilizing gene ontology. *Genome Informat*, *16*, 106–115.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., … Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, *31*(1), 365–370.

Borde, V., & Duguet, M. (1998). DNA topoisomerase II sites in the histone H4 gene during the highly synchronous cell cycle of *Physarum polycephalum*. *Nucleic Acids Res*, *26*(9), 2042–49.

Borodovsky, M., & Lomsadze, A. (2011). Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Current Protocols in Bioinformatics,* Andreas D. Baxevanis *et al.* (Eds.), Chapter 4, Unit 4.6.1–10.

Braun, R., Mittermayer, C., & Rusch, H. P. (1965). Sequential temporal replication of DNA in *Physarum polycephalum*. *Proc Natl Acad Sci USA*, *53*(5), 924–31.

Brulle, F., Cocquerelle, C., Mitta, G., Castric, V., Douay, F., Leprêtre, A., & Vandenbulcke, F. (2008). Identification and expression profile of gene transcripts differentially expressed during metallic exposure in *Eisenia fetida coelomocytes*. *Dev Comp Immunol*, *32*(12), 1441–1453.

Bundschuh, R., Altmüller, J., Becker, C., Nürnberg, P., & Gott, J. M. (2011). Complete characterization of the edited transcriptome of the mitochondrion of *Physarum polycephalum* using deep sequencing of RNA. *Nucleic Acids Res*, *39*(14), 6044–55.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., … Bateman, A. (2013). Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res*, *41*(Database issue), D226–32.

Burland, T. G., Solnica-Krezel, L., Bailey, J., Cunningham, D. B., & Dove, W. F. (1993). Patterns of inheritance, development and the mitotic cycle in the protist *Physarum polycephalum*. *Adv Microbial Physiol*, *35*, 1–69.

Burset, M., & Guigo, R. (1996). Evaluation of gene structure prediction programs. *Genomics*, *34*(3), 353–367.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421.

Cerutti, L., & Simanis, V. (1999). Asymmetry of the spindle pole bodies and spg1p GAP segregation during mitosis in fission yeast. *J Cell Sci*, *112*(14), 2313–2321.

Chen, C., Frankhouser, D., & Bundschuh, R. (2012). Comparison of insertional RNA editing in Myxomycetes. *PLoS Comput Biology*, *8*(2), e1002400.

Chen, F., Mackey, A. J., Vermunt, J. K., & Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS One*, *2*(4), e383.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., … Wong, E. D. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res*, *40*(Database issue), D700–5.

Chisholm, R. L., Gaudet, P., Just, E. M., Pilcher, K. E., Fey, P., Merchant, S. N., & Kibbe, W. A. (2006). dictyBase, the model organism database for *Dictyostelium discoideum*. *Nucl. Acids Res.*, *34*(suppl_1), D423–427.

Christmann, M., Tomicic, M. T., Origer, J., & Kaina, B. (2005). Fen1 is induced p53 dependently and involved in the recovery from UV-light-induced replication inhibition. *Oncogene*, *24*(56), 8304–8313.

Chung, C. Y., & Firtel, R. A. (1999). PAKa, a putative PAK family member, is required for cytokinesis and the regulation of the cytoskeleton in Dictyostelium discoideum cells during chemotaxis. *J Cell Biol*, *147*(3), 559–576.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., … Bader, G. D. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, *2*(10), 2366–2382.

Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., … Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, *5*(7), 613–619.

Cock, P. J. a, Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*, *38*(6), 1767–71.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–6.

The Gene Ontology Consortium (2008). The Gene Ontology project in 2008. *Nucleic Acids Res*, *36*(suppl 1), D440–D444.

The UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*, *38*(Database issue), D142–8.

Cunningham, D. B., & Dove, W. F. (1993). Two alleles of a developmentally regulated alpha-tubulin locus in *Physarum polycephalum* replicate on different schedules. *Mol Cell Biol*, *13*(1), 449–61.

Daniel, D. C., & Johnson, E. M. (1989). Selective initiation of replication at origin sequences of the rDNA molecule of Physarum polycephalum using synchronous plasmodial extracts. *Nucleic Acids Res*, *17*(20), 8343–62.

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

Datta, S., & Firtel, R. A. (1987). Identification of the sequences controlling cyclic AMP regulation and cell-type-specific expression of a prestalk-specific gene in Dictyostelium discoideum. *Mol Cell Biol*, *7*(1), 149–159.

De Hostos, E. L., Bradtke, B., Lottspeich, F., & Gerisch, G. (1993). Coactosin, a 17 kDa F-actin binding protein from *Dictyostelium discoideum*. *Cell Motil Cytoskeleton*, *26*(3), 181–191.

De la Fuente, I. M., Cortes, J. M., Pelta, D. A., & Veguillas, J. (2013). Attractor metabolic networks. *PloS One*, *8*(3), e58284.

Dee, J. (1987). Genes and development in *Physarum*. *Trends Genet*, *3*(8), 208–213.

Dee, J., Foxon, J. L., & Anderson, R. W. (1989). Growth, Development and Genetic Characteristics of *Physarum polycephalum* Amoebae Able to Grow in Liquid, Axenic Medium. *Microbiology*, *135*(6), 1567–1588.

Delley, P. A., & Hall, M. N. (1999). Cell wall stress depolarizes cell growth via hyperactivation of RHO1. *J Cell Biol*, *147*(1), 163–174.

Deng, M., Mehta, S., Sun, F., & Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, *12*(10), 1540–1548.

Dhanasekaran, K., Kumari, S., & Kanduri, C. (2013). Noncoding RNAs in Chromatin Organization and Transcription Regulation: An Epigenetic View. In T. K. Kundu (Ed.), *Epigenetics: Development and Disease* (Vol. 61, pp. 343–372). Dordrecht: Springer Netherlands.

Dolezel, J., Bartos, J., Voglmayr, H., & Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry*, *51*(2), 127–8

Dray, S., & Dufour, A.-B. (2007). The ade4 package: implementing the duality diagram for ecologists. *J Stat Software*, *22*(4), 1–20.

Dussutour, A., Latty, T., Beekman, M., & Simpson, S. J. (2010). Amoeboid organism solves complex nutritional challenges. *Proc Natl Acad Sci USA*, *107*(10), 4607–11.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–1.

Edmonds, B. T., Bell, A., Wyckoff, J., Condeelis, J., & Leyh, T. S. (1998). The Effect of F-actin on the Binding and Hydrolysis of Guanine Nucleotide by *Dictyostelium* Elongation Factor 1A. *J Bio. Chem*, *273*(17), 10288–10295.

Eichinger, L., Pachebat, J. A., Glöckner, G., Rajandream, M.-A. A., Sucgang, R., Berriman, M., … Kuspa, A. (2005). The genome of the social amoeba *Dictyostelium discoideum*. *Nature*, *435*(7038), 43–57.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol*, *6*(5), R44.

Eilbeck, K., Moore, B., Holt, C., & Yandell, M. (2009). Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics*, *10*(1), 67.

Fankhauser, C., Marks, J., Reymond, A., & Simanis, V. (1993). The *S. pombe* cdc16 gene is required both for maintenance of p34cdc2 kinase activity and regulation of septum formation: a link between mitosis and cytokinesis? *EMBO J*, *12*(7), 2697–2704.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J. J., Hotz, H.-R. R., … Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Res*, *36*(Database issue), D281–D288.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., … Merrick, J. M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, *269*(5223), 496–512.

Francis, W. R., Christianson, L. M., Kiko, R., Powers, M. L., Shaner, N. C., & Haddock, S. H. D. (2013). A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC Genomics*, *14*(1), 167.

Fraser, H. B., Hirsh, A. E., Wall, D. P., & Eisen, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci USA*, *101*(24), 9033–9038.

Fronk, J., & Magiera, R. (1994). DNA methylation during differentiation of a lower eukaryote, *Physarum polycephalum*. *Biochem J*, *304 (Pt 1)*, 101–104.

Fujioka, H., & Shimoda, C. (1989). A mating-type-specific sterility gene map1 is required for transcription of a mating-type gene mat1-Pi in the fission yeast *Schizosaccharomyces pombe*. *FEMS Microbiol Lett*, *51*(1), 45–48.

Fukuzawa, M., & Williams, J. G. (2000). Analysis of the promoter of the cudA gene reveals novel mechanisms of *Dictyostelium* cell type differentiation. *Development*, *127*(12), 2705–2713.

Furukawa, R., Maselli, A., Thomson, S. A. M., Lim, R. W. L., Stokes, J. V, & Fechheimer, M. (2003). Calcium regulation of actin crosslinking is important for function of the actin cytoskeleton in *Dictyostelium*. *J Cell Sci*, *116*(1), 187–196.

Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, *8*(6), 469–477.

Gaudet, P., Fey, P., Basu, S., Bushmanova, Y. a, Dodson, R., Sheppard, K. a, … Chisholm, R. L. (2011). dictyBase update 2011: web 2.0 functionality and the initial steps towards a genome portal for the Amoebozoa. *Nucleic Acids Res*, *39*(Database issue), D620–4.

Ge, H., Liu, Z., Church, G. M., & Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat Genet*, *29*(4), 482–486.

Gioti, A., Nystedt, B., Li, W., Xu, J., Andersson, A., Averette, A. F., … Scheynius, A. (2013). Genomic insights into the atopic eczema-associated skin commensal yeast *Malassezia sympodialis*. *mBio*, *4*(1), e00572–12.

Gladman, S. (2009). Velvetg running time. Retrieved from http://listserver.ebi.ac.uk/pipermail/velvet-users/2009-July/000474.html

Glöckner, G., Golderer, G., Werner-Felmayer, G., Meyer, S., & Marwan, W. (2008). A first glimpse at the transcriptome of *Physarum polycephalum*. *BMC Genomics*, *9*, 6.

Golderer, G., Werner, E. R., Leitner, S., Grobner, P., & Werner-Felmayer, G. (2001). Nitric oxide synthase is induced in sporulation of *Physarum polycephalum*. *Genes Dev.*, *15*(10), 1299–1309.

Gordon, A. (2008). FASTX toolkit, Version 0.013. Retrieved from http://hannonlab.cshl.edu/fastx_toolkit

Gordon, M., & Hardman, N. (1988). Detection and analysis of CpG-rich islands in the genome of *Physarum polycephalum*. *Curr Genet*, *14*(1), 23–28.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., … Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*, *36*(10), 3420–3435.

Grewal, S. I. S., & Elgin, S. C. R. (2007). Transcription and RNA interference in the formation of heterochromatin. *Nature*, *447*(7143), 399–406.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., & Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, *33*(Database issue), D121–4.

Grigoriev, I. V, Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., … Dubchak, I. (2012). The genome portal of the Department of Energy - Joint Genome Institute. *Nucleic Acids Res*, *40*(Database issue), D26–32.

Gross, S. R., & Kinzy, T. G. (2007). Improper Organization of the Actin Cytoskeleton Affects Protein Synthesis at Initiation. *Mol. Cell. Biol.*, *27*(5), 1974–1989.

Gualtieri, T., Ragni, E., Mizzi, L., Fascio, U., & Popolo, L. (2004). The cell wall sensor Wsc1p is involved in reorganization of actin cytoskeleton in response to hypo-osmotic shock in *Saccharomyces cerevisiae*. *Yeast*, *21*(13), 1107–1120.

Guigó, R. (2013). The Coding and the Non-coding Transcriptome. In *Handbook of Systems Biology* (pp. 27–41). Elsevier.

Guttes, E., & Guttes, S. (1964). Chapter 3: Mitotic Synchrony in the Plasmodia of *Physarum polycephalum* and Mitotic Synchronization by Coalescence of Microplasmodia (Vol. 1, pp. 43–54). Elsevier.

Haindl, M., & Holler, E. (2005). Use of the giant multinucleate plasmodium of *Physarum polycephalum* to study RNA interference in the myxomycete. *Anal Bioch*, *342*(2), 194–9.

Hansey, C. N., Vaillancourt, B., Sekhon, R. S., de Leon, N., Kaeppler, S. M., & Buell, C. R. (2012). Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing. *PloS One*, *7*(3), e33071.

Hardman, N., Jack, P. L., Fergie, R. C., & Gerrie, L. M. (1980). Sequence organisation in nuclear DNA from *Physarum polycephalum*. Interspersion of repetitive and single-copy sequences. *Eur J Biochem / FEBS*, *103*(2), 247–57.

Hawkins, R. D., Hon, G. C., & Ren, B. (2010). Next-generation genomics: an integrative approach. *Nature Rev Genet*, *11*(7), 476–86.

Heads, R., & Carpenter, B. (1990). Differential synthesis of histone H1 during early spherulation in *Physarum polycephalum*. *Biochim Biophys Acta*, *1053*(1), 56–62.

Heidel, A. J., Lawal, H. M., Felder, M., Schilde, C., Helps, N. R., Tunggal, B., … Glöckner, G. (2011). Phylogeny-wide analysis of social amoeba genomes highlights ancient origins for complex intercellular communication. *Genome Res*, *21*(11), 1882–91.

Hildebrandt, A. (1986). Methylation is an early and necessary step in the sporulation programme of the slime mold *Physarum polycephalum*. *Exp Cell Res*, *167*(1), 271–275.

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*, 491.

Hou, Y., & Lin, S. (2009). Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes. *PLoS ONE*, *4*(9), e6978.

Howe, A. (2004). Regulation of actin-based cell migration by cAMP/PKA. *Biochim Biophys Acta,* 2–3, 159–174.

Hu, H., Bandyopadhyay, P. K., Olivera, B. M., & Yandell, M. (2011). Characterization of the *Conus bullatus* genome and its venom-duct transcriptome. *BMC Genomics*, *12*(1), 60.

Huang, X., & Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. *Genome Res*, *9*(9), 868–877.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., … Yeats, C. (2009a). InterPro: the integrative protein signature database. *Nucleic Acids Res*, *37*(Database issue), D211–5.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., … Yeats, C. (2009b). InterPro: the integrative protein signature database. *Nucleic Acids Res*, *37*(Database issue), D211–D215.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B. B., Lönnerberg, P., & Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res*, *21*(7), 1160–1167.

Issel-Tarver, L., Christie, K. R., Dolinski, K., Andrada, R., Balakrishnan, R., Ball, C. A., … Cherry, J. M. (2002). *Saccharomyces* Genome Database. *Methods in Enzymology*, *350*, 329–46.

Iwasaki, W., Sasaki, H., Nakamura, A., Kohama, K., & Tanokura, M. (1999). Crystallization and preliminary X-ray diffraction studies of a 40 kDa calcium binding protein specifically expressed in plasmodia of *Physarum polycephalum. J Biochem*, *126*(1), 7–9.

Jacobson, D. N., & Dove, W. F. (1975). The amoebal cell of Physarum polycephalum: colony formation and growth. *Dev Biol*, *47*(1), 97–105.

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genet*, *33 Suppl*, 245–54.

Jeganathan, S., Morrow, A., Amiri, A., & Lee, J. M. (2008). Eukaryotic elongation factor 1A2 cooperates with phosphatidylinositol-4 kinase III beta to stimulate production of filopodia through increased phosphatidylinositol-4,5 bisphosphate generation. *Mol Cell Biol*, *28*(14), 4549–4561.

Jerzmanowski, A., & Moraczewska, J. (1988). Distribution of postsynthetic methylation sites in *Physarum* histone H1. *Mol Biol Rep*, *13*(2), 97–101.

Jin, J., Cardozo, T., Lovering, R. C., Elledge, S. J., Pagano, M., & Harper, J. W. (2004). Systematic analysis and nomenclature of mammalian F-box proteins. *Genes Dev*, *18*(21), 2573–2580.

Jurka, J., Kapitonov, V. V, Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, *110*(1-4), 462–7.

Kaller, M., Nellen, W., & Chubb, J. R. (2006). Epigenetics in *Dictyostelium discoideum*. In L. Eichinger & F. Rivero (Eds.), *Methods Mol Biol* (Vol. 346, pp. 491–505). New Jersey: Humana Press.

Kamiya, N. (1940). The Control Of Protoplasmic Streaming. *Science*, *92*(2394), 462–3.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., … Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, *36*(Database issue), D480–484.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, *38*(Database issue), D355–60.

Keibler, E., & Brent, M. R. (2003). Eval: a software package for analysis of genome annotations. *BMC Bioinformatics*, *4*, 50.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res*, *12*(4), 656–64.

Khaire, N., Müller, R., Blau-Wasser, R., Eichinger, L., Schleicher, M., Rief, M., … Noegel, A. A. (2007). Filamin-regulated F-actin assembly is essential for morphogenesis and controls phototaxis in *Dictyostelium*. *J Biol Chem*, *282*(3), 1948–1955.

King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., … Rokhsar, D. (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature*, *451*(7180), 783–8.

Knoop, V. (2011). When you can't trust the DNA: RNA editing changes transcript sequences. *Cell Mol Life Sci*, *68*(4), 567–86.

Kohama, K., & Nakamura, A. (2001). *Physarum* Cell Culture. Retrieved from http://mrw.interscience.wiley.com/emrw/9780470015902/els/article/a0002580/current/abstract

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., & Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*, *35*(Web Server issue), W345–9.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*(1), 59. doi:10.1186/1471-2105-5-59

Korf, I., Yandell, M., & Bedell, J. (2003). *BLAST*. Sebastopol, CA, USA: O'Reilly & Associates, Inc.

Kozomara, A., & Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, *39*(Database issue), D152–7.

Kroneder, R., Cashmore, A. R., & Marwan, W. (1999). Phytochrome-induced expression of lig1, a homologue of the fission yeast cell-cycle checkpoint gene hus1, is associated with the developmental switch in P*hysarum polycephalum* plasmodia. *Curr Genet*, *36*(1), 86–93.

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*, *35*(9), 3100–8.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Pan. (2001). Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. *Nature*, *409*(6814), 860–921.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, *10*(3), R25–10.

Laroche, A., Lemieux, G., & Pallotta, D. (1989). The nucleotide sequence of a developmentally regulated cDNA from *Physarum polycephalum*. *Nucleic Acids Res*, *17*(24), 10502. R

Larsen, E., Kleppa, L., Meza, T. J., Meza-Zepeda, L. A., Rada, C., Castellanos, C. G., … Klungland, A. (2008). Early-onset lymphoma and extensive embryonic apoptosis in two domain-specific Fen1 mice mutants. *Cancer Research*, *68*(12), 4571–4579.

Larue, H., Masson, S., Lafontaine, J. G., Nadeau, P., & Pallotta, D. (1982). Changes in protein and RNA during asexual differentiation of *Physarum polycephalum*. *Canadian J Microbiol*, *28*(4), 438–447.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., … Cochrane, G. (2011). The European Nucleotide Archive. *Nucleic Acids Res*, *39*(Database issue), D28–31.

Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Res*, *39*(Database issue), D19–D21.

Lépine, G., Laroche, A., Lemieux, G., & Pallotta, D. (1995). The two alleles of the hapP gene in *Physarum polycephalum* code for different proteins. *Biochim Biophys Acta*, *1264*(3), 271–4.

Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity*, *104*(6), 520–33.

Letunic, I., Yamada, T., Kanehisa, M., & Bork, P. (2008). iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci*, *33*(3), 101–103.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–9.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, *13*(9), 2178–89.

Li, P., Maines-Bandiera, S., Kuo, W. L., Guan, Y., Sun, Y., Hills, M., … Auersperg, N. (2007). Multiple roles of the candidate oncogene ZNF217 in ovarian epithelial neoplastic progression. *Int J Cancer*, *120*, 1863–1873.

Lin, H., & Spradling, A. C. (1997). A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. *Development*, *124*(12), 2463–2476.

Liolios, K., Chen, I.-M. A., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V. M., & Kyrpides, N. C. (2010). The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*, *38*(Database issue), D346–54.

Loewy, A. (1952). An actomyosin-like substance from the plasmodium of a myxomycete. *J Cell Physiol*, *40*(1), 127–56.

Loidl, P. (1988). Towards an understanding of the biological function of histone acetylation. *FEBS Letters*, *227*(2), 91–5.

Loidl, P., & Gröbner, P. (1986). Biosynthesis and posttranslational acetylation of histones during spherulation of *Physarum polycephalum. Nucleic Acids Res*, *14*(9), 3745–62.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., & Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*, *33*(20), 6494–506.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, *25*(5), 955–64.

Lusser, A., Brosch, G., López-Rodas, G., & Loidl, P. (1997). Histone acetyltransferases during the cell cycle and differentiation of *Physarum polycephalum*. *Eur J Cell Biol*, *74*(1), 102–10.

Mahendran, R., Spottswood, M. R., & Miller, D. L. (1991). RNA editing by cytidine insertion in mitochondria of *Physarum polycephalum*. *Nature*, *349*(6308), 434–8. doi:10.1038/349434a0

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–70.

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, *470*(7333), 198–203.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380.

Maric, C., Bénard, M., & Pierron, G. (2003). Developmentally regulated usage of *Physarum* DNA replication origins. *EMBO Reports*, *4*(5), 474–8.

Maric, C., Bénard, M., & Pierron, G. (2010). RNase-dependent discontinuities associated with the crossovers of spontaneously formed joint DNA molecules in *Physarum polycephalum*. *Chromosoma*, *119*(6), 601–11.

Maric, C., Swanston, E., Bailey, J., & Pierron, G. (2002). Replicational organization of three weakly expressed loci in *Physarum polycephalum*. *Nucleic Acids Res*, *30*(11), 2261–9.

Martel, R., Tessier, A., Pallotta, D., & Lemieux, G. (1988). Selective gene expression during sporulation of Physarum polycephalum. *J. Bacteriol.*, *170*(10), 4784–4790.

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews. Genetics*, *12*(10), 671–682.

Marwan, W. (2003). Detecting functional interactions in a gene and signaling network by time-resolved somatic complementation analysis. *BioEssays*, *25*(10), 950–960.

Marwan, W. (2003). Theory of time-resolved somatic complementation and its use to explore the sporulation control network in *Physarum polycephalum*. *Genetics*, *164*(1), 105–15.

McCarthy, F. M., Wang, N., Magee, G. B., Nanduri, B., Lawrence, M. L., Camon, E. B., … Burgess, S. C. (2006a). AgBase: a functional genomics resource for agriculture. *BMC Genomics*, *7*, 229.

McCarthy, F. M., Wang, N., Magee, G. B., Nanduri, B., Lawrence, M. L., Camon, E. B., … Burgess, S. C. (2006b). AgBase: a functional genomics resource for agriculture. *BMC Genomics*, *7*(1), 229.

Meijer, H. J., Latijnhouwers, M., Ligterink, W., & Govers, F. (2005). A transmembrane phospholipase D in Phytophthora; a novel PLD subfamily. *Gene*, *350*(2), 173–182.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, *11*(1), 31–46.

Miles, R. R., Sluka, J. P., Santerre, R. F., Hale, L. V, Bloem, L., Boguslawski, G., … Onyia, J. E. (2000). Dynamic Regulation of RGS2 in Bone: Potential New Insights into Parathyroid Hormone Signaling Mechanisms. *Endocrinology*, *141*(1), 28–36.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315–327.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., & Marshall, D. (2010). Tablet—next generation sequence assembly visualization. *Bioinformatics*, *26*(3), 401–402.

Minami, Y., Ishihara, M., Hayase, M., Sakaguchi, T., & Yubisui, T. (2009). cDNA Cloning and Life-Cycle Stage-Specific Expression of Coronin from *Physarum polycephalum*. *Biosci Biotechnol Biochem*, *73*(3), 747–9.

Mizuno, K., Kitamura, A., & Sasaki, T. (2003). Rabring7, a novel Rab7 target protein with a RING finger motif. *Mol Biol Cell*, *14*(9), 3741–3752.

Mohberg, J. (1977). Nuclear DNA content and chromosome numbers throughout the life cycle of the Colonia strain of the myxomycete, *Physarum polycephalum*. *J Cell Sci*, *24*, 95–108.

Mohberg, J., & Rusch, H. P. (1971). Isolation and DNA content of nuclei of *Physarum polycephalum*. *Exp Cell Res*, *66*(2), 305–16.

Moniakis, J., Coukell, M. B., & Janiec, A. (1999). Involvement of the Ca2+-ATPase PAT1 and the contractile vacuole in calcium regulation in *Dictyostelium discoideum*. *J Cell Sci*, *112*(3), 405–414.

Moore, B., Fan, G., & Eilbeck, K. (2010). SOBA: sequence ontology bioinformatics analysis. *Nucleic Acids Res*, *38*(Web Server issue), W161–4.

Morcos, F., Lamanna, C., Sikora, M., & Izaguirre, J. (2008). Cytoprophet: a Cytoscape plug-in for protein and domain interaction networks inference. *Bioinformatics*, *24*(19), 2265–6.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*, *35*(Web Server issue), W182–5.

Morozova, O., Hirst, M., & Marra, M. A. (2009). Applications of new sequencing technologies for transcriptome analysis. *Ann Rev Genomics Hum Genet*, *10*, 135–51.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628.

Mottahedeh, J., & Marsh, R. (1998). Characterization of 101-kDa Transglutaminase from *Physarum polycephalum* and Identification of LAV1-2 as Substrate. *J Biol Chem*, *273*(45), 29888–29895.

Munshi, R., Kandl, K. A., Carr-Schmid, A., Whitacre, J. L., Adams, A. E., & Kinzy, T. G. (2001). Overexpression of Translation Elongation Factor 1A Affects the Organization and Function of the Actin Cytoskeleton in Yeast. *Genetics*, *157*(4), 1425–1436.

Mutz, K.-O., Heilkenbrinker, A., Lönne, M., Walter, J.-G., & Stahl, F. (2013). Transcriptome analysis using next-generation sequencing. *Curr Opinion Biotechnol*, *24*(1), 22–30.

Myhre, S., Tveit, H., Mollestad, T., & Lægreid, A. (2006). Additional Gene Ontology structure for improved biological reasoning. *Bioinformatics*, *22*(16), 2020–2027.

Nachmias, V. T., & Huxley, H. E. (1970). Electron microscope observations on actomyosin and actin preparations from *Physarum polycephalum*, and on their interaction with heavy meromyosin subfragment I from muscle myosin. *J Mol Biol*, *50*(1), 83–90.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, *320*(5881), 1344–1349.

Nakagaki, T., Yamada, H., & Hara, M. (2004). Smart network solutions in an amoeboid organism. *Biophys Chem*, *107*(1), 1–5.

Nakagaki, T., Yamada, H., & Tóth, A. (2000). Maze-solving by an amoeboid organism. *Nature*, *407*(6803), 470.

Nakamura, A., & Kohama, K. (1999). Calcium regulation of the actin-myosin interaction of Physarum polycephalum. *Intl Rev Cytol*, *191*, 53–98.

National Institutes of Health. (2004). Tools for genetic and genomic studies in emerging model organisms. Retrieved from http://grants.nih.gov/grants/guide/pa-files/PA-04-135.html

Navlakha, S., & Bar-Joseph, Z. (2011). Algorithms in nature: the convergence of systems biology and computational thinking. *Mol Syst Biol*, *7*, 546.

Nawrocki, E. P., Kolbe, D. L., & Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics*, *25*(10), 1335–7.

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., … Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics*, *28*(14), 1919–20.

Nowrousian, M. (2010). Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Euk Cell*, *9*(9), 1300–1310.

Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., … Kanehisa, M. (2008). KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*, *36*(Web Server issue), W423–6.

Oresic, M., & Shalloway, D. (1998). Specific correlations between relative synonymous codon usage and protein secondary structure. *J Mol Biol*, *281*(1), 31–48.

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*, *12*(2), 87–98.

Parikh, A., Miranda, E. R., Katoh-Kurasawa, M., Fuller, D., Rot, G., Zagar, L., … Shaulsky, G. (2010). Conserved developmental transcriptomes in evolutionarily divergent species. *Genome Biol*, *11*(3), R35.

Parkinson, J., & Blaxter, M. (2003). SimiTri--visualizing similarity relationships for groups of sequences. *Bioinformatics*, *19*(3), 390–5.

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*(9), 1061–7.

Parra, G., Bradnam, K., Ning, Z., Keane, T., & Korf, I. (2009). Assessing the gene space in draft genomes. *Nucleic Acids Res*, *37*(1), 289–97.

Pathak, R., Bogomolnaya, L. M., Guo, J., & Polymenis, M. (2004). Gid8p (Dcr1p) and Dcr2p Function in a Common Pathway To Promote START Completion in Saccharomyces cerevisiae. *Euk Cell*, *3*(6), 1627–1638.

Payne, S., & Loomis, W. (2006). Retention and loss of amino acid biosynthetic pathways based on analysis of whole-genome sequences. *Euk Cell*, *5*(2), 272–276.

Peden, J. (2005). Correspondence Analysis of Codon Usage. Retrieved June 11, 2013, from http://codonw.sourceforge.net/

Peregrín-Alvarez, J. M., & Parkinson, J. (2009). Phylogenomic analysis of EST datasets. *Methods in Molecular Biology (Clifton, N.J.)*, *533*, 257–76.

Pesis, K. H., & Matthews, H. R. (1986). Histone acetylation in replication and transcription: turnover at specific acetylation sites in histone H4 from *Physarum polycephalum*. *Archiv Biochem Biophys*, *251*(2), 665–73.

Physarum Genome Sequencing Consortium. (2013). Physarum Genome Resources. Retrieved from http://www.physarum-blast.ovgu.de/

Pierron, G., Benard, M., Puvion, E., Flanagan, R., Sauer, H. W., & Pallotta, D. (1989). Replication timing of 10 developmentally regulated genes in *Physarum polycephalum*. *Nucleic Acids Res*, *17*(2), 553–66.

Pinchai, N., Lee, B. S., & Holler, E. (2006). Stage specific expression of poly(malic acid)-affiliated genes in the life cycle of *Physarum polycephalum*. Spherulin 3b and polymalatase. *FEBS J*, *273*(5), 1046–1055.

Pope, S. N., & Lee, I. R. (2005). Yeast two-hybrid identification of prostatic proteins interacting with human sex hormone-binding globulin. *J Steroid Biochem Mol Biol*, *94*(1-3), 203–208.

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, *21 Suppl 1*, i351–8.

Prior, C. P., Cantor, C. R., Johnson, E. M., & Allfrey, V. G. (1980). Incorporation of exogenous pyrene-labeled histone into *Physarum* chromatin: a system for studying changes in nucleosomes assembled in vivo. *Cell*, *20*(3), 597–608.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, *35*(Database issue):D61-5.

Puls, A., Schmidt, S., Grawe, F., & Stabel, S. (1997). Interaction of protein kinase C zeta with ZIP, a novel protein kinase C-binding protein. *Proc Natl Acad Sci USA*, *94*(12), 6191–6196.

Putzer, H., Verfuerth, C., Claviez, M., & Schreckenbach, T. (1984). Photomorphogenesis in Physarum: induction of tubulins and sporulation-specific proteins and of their mRNAs. *Proc Natl Acad Sci USA*, *81*(22), 7117–7121.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic Acids Res*, *33*(Web Server issue), W116–20.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–2.

Quinlan, M. E., Heuser, J. E., Kerkhoff, E., & Mullins, R. D. (2005). Drosophila Spire is an actin nucleation factor. *Nature*, *433*(7024), 382–388.

R Core Team. (2013). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.r-project.org

Ratel, D., Ravanat, J.-L., Berger, F., & Wion, D. (2006). N6-methyladenine: the other methylated base of DNA. *BioEssays*, *28*(3), 309–315.

Rehm, H. L., Bale, S. J., Bayrak-Toydemir, P., Berg, J. S., Brown, K. K., Deignan, J. L., … Lyon, E. (2013). ACMG clinical laboratory standards for next-generation sequencing. *Genetics in Medicine*, *15*(9), 733–47.

Reid, C. R., & Beekman, M. (2013). Solving the Towers of Hanoi - how an amoeboid organism efficiently constructs transport networks. *J Exp Biol*, *216*(Pt 9), 1546–51.

Reinhardt, J. A., Baltrus, D. A., Nishimura, M. T., Jeck, W. R., Jones, C. D., & Dangl, J. L. (2009). De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. oryzae. *Genome Res*, *19*(2), 294–305.

Renzel, S., Esselborn, S., Sauer, H. W., & Hildebrandt, A. (2000). Calcium and Malate Are Sporulation-Promoting Factors of Physarum polycephalum. *J. Bacteriol.*, *182*(24), 6900–6905.

Roberts, A., Pimentel, H., Trapnell, C., & Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, *27*(17), 2325–9.

Röhrig, U., Gerisch, G., Morozova, L., Schleicher, M., & Wegner, A. (1995). Coactosin interferes with the capping of actin filaments. *FEBS Letters*, *374*(2), 284–286.

Rothnie, H. M., McCurrach, K. J., Glover, L. A., & Hardman, N. (1991). Retrotransposon-like nature of Tp1 elements: implications for the organisation of highly repetitive, hypermethylated DNA in the genome of *Physarum polycephalum*. *Nucleic Acids Res*, *19*(2), 279–286.

Sachsenmaier, W., Remy, U., & Plattner-Schobel, R. (1972). Initiation of synchronous mitosis in *Physarum polycephalum*. A model of the control of cell division in eukariots. *Exp Cell Res*, *73*(1), 41–8.

Saigusa, T., Tero, A., Nakagaki, T., & Kuramoto, Y. (2008). Amoebae anticipate periodic events. *Phys Rev Letters*, *100*(1).

Saiki, R., Gelfand, D., Stoffel, S., Scharf, S., Higuchi, R., Horn, G., … Erlich, H. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, *239*(4839), 487–491.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, *74*(12), 5463–5467. doi:10.1073/pnas.74.12.5463

Sauer, H. W., Babcock, K. L., & Rusch, H. P. (1969). Sporulation in *Physarum polycephalum*: a model system for studies on differentiation. *Exp Cell Res*, *57*(2), 319–327.

Savard, L., Laroche, A., Lemieux, G., & Pallota, D. (1989). Developmentally regulated late mRNAs in the encystment of *Physarum polycephalum* plasmodia. *Biochim Biophys Acta*, *1007*(3), 264–269.

Schedl, T., Owens, J., Dove, W. F., & Burland, T. G. (1984). Genetics of the tubulin gene families of *Physarum*. *Genetics*, *108*(1), 143–164.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*(5235), 467–470.

Schlatterer, C., Gollnick, F., Schmidt, E., Meyer, R., & Knoll, G. (1994). Challenge with high concentrations of cyclic AMP induces transient changes in the cytosolic free calcium concentration in *Dictyostelium discoideum*. *J Cell Sci*, *107*(8), 2107–2115.

Schmieder, R., & Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS One*, *6*(3), e17288.

Schmieder, R., Lim, Y. W., Rohwer, F., & Edwards, R. (2010). TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics*, *11*(1), 341.

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*(8), 1086–92.

Seibenhener, M. L., Geetha, T., & Wooten, M. W. (2007). Sequestosome 1/p62 - More than just a scaffold. *FEBS Lett*, *581*(2), 175–179.

Seifriz, W. (1937). A theory of protoplasmic streaming. *Science*, *86*(2235), 397–8.

Serrano, R., Martin, H., Casamayor, A., & Arino, J. (2006). Signaling Alkaline pH Stress in the Yeast Saccharomyces cerevisiae through the Wsc1 Cell Surface Sensor and the Slt2 MAPK Pathway. *J Biol Chem*, *281*(52), 39785–39795.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, *13*(11), 2498–2504.

She, R., Chu, J. S.-C., Uyar, B., Wang, J., Wang, K., & Chen, N. (2011). genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics (Oxford, England)*, *27*(15), 2141–3.

She, R., Chu, J., Wang, K., Pei, J., & Chen, N. (2009). genBlastA: Enabling BLAST to identify homologous gene sequences. *Genome Res*, (19), 143–149.

Shimada, Y., Nakano, M., Kanda, N., Murakami-Murofushi, K., Kim, J. K., Ide, T., & Murofushi, H. (1997). Cell cycle-dependent activation of telomerase in naturally synchronized culture of a true slime mold, *Physarum polycephalum*. *Biochem Biophys Res Communications*, *232*(2), 492–6.

Shirai, Y., Sasaki, N., Kishi, Y., Izumi, A., Itoh, K., Sameshima, M., … Murakami-Murofushi, K. (2006). Regulation of levels of actin threonine phosphorylation during life cycle of *Physarum polycephalum*. *Cell Motil Cytoskeleton*, *63*(2), 77–87.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., … Folta, K. M. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nature Genet*, *43*(2), 109–16.

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, 31.

Slingsby, C., Wistow, G. J., & Clark, A. R. (2013). Evolution of crystallins for a role in the vertebrate eye lens. *Protein Science*, *22*(4), 367–80.

Smit, A., & Hubley, R. (2010). RepeatModeler. Retrieved from http://www.repeatmasker.org

Smit, A., Hubley, R., & Green, P. (2010). RepeatMasker. Retrieved from http://www.repeatmasker.org

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., & Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, *27*(3), 431–2.

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*(5), 637–44.

Starostzik, C., & Marwan, W. (1995). Functional mapping of the branched signal transduction pathway that controls sporulation in *Physarum polycephalum*. *Photochemistry and Photobiology*, *62*(5), 930–933.

Starostzik, C., & Marwan, W. (1998). Kinetic analysis of a signal-transduction pathway by time-resolved somatic complementation of mutants. *J Exp Biol*, *201*(Pt 13), 1991–1999.

Stettler, S., Warbrick, E., Prochnik, S., Mackie, S., & Fantes, P. (1996). The wis1 signal transduction pathway is required for expression of cAMP-repressed genes in fission yeast. *J Cell Sci*, *109 (Pt 7)*, 1927–1935.

Strahl, B. D., & Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, *403*(6765), 41–45.

Sucgang, R., Kuo, A., Tian, X., Salerno, W., Parikh, A., Feasley, C. L., … Grigoriev, I. V. (2011). Comparative genomics of the social amoebae Dictyostelium discoideum and *Dictyostelium purpureum*. *Genome Biol*, *12*(2), R20.

Sujatha, A., Balaji, S., Devi, R., & Marwan, W. (2005). Isolation of *Physarum polycephalum* plasmodial mutants altered in sporulation by chemical mutagenesis of flagellates. *Eur J Protistol*, *41*(1), 19–27.

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One*, *6*(7), e21800.

Sweeney, G. E., Watts, D. I., & Turnock, G. (1987). Differential gene expression during the amoebal-plasmodial transition in *Physarum*. *Nucleic Acids Res*, *15*(3), 933–945.

T'jampens, D., Bailey, J., Cook, L. J., Constantin, B., Vandekerckhove, J., & Gettemans, J. (1999). *Physarum* amoebae express a distinct fragmin-like actin-binding protein that controls *in vitro* phosphorylation of actin by the actin-fragmin kinase. *Eur J Biochem*, *265*(1), 240–250.

Tairbekov, M. G., Parfyonov, G. P., Platonova, R. W., Abramova, V. M., Golov, V. K., Rostopshina, A. V., … Chuchkin, V. G. (1981). Biological investigations aboard the biosatellite Cosmos-1129. *Adv Space Res*, *1*(14), 89–94.

Takano, H., Abe, T., Sakurai, R., Moriyama, Y., Miyazawa, Y., Nozaki, H., … Kuroiwa, T. (2001). The complete DNA sequence of the mitochondrial genome of Physarum polycephalum. *Mol Gen Genet*, *264*(5), 539–45.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., … Surani, M. A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell*, *6*(5), 468–478.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., … Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382.

Tang, F., Lao, K., & Surani, M. A. (2011). Development and applications of single-cell transcriptome analysis. *Nature Methods*, *8*(4 Suppl).

Tempel, S. (2012). Using and understanding RepeatMasker. *Methods in Molecular Biology (Clifton, N.J.)*, *859*, 29–51.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., & Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Res*, *18*(12), 1979–90.

Tero, A., Takagi, S., Saigusa, T., Ito, K., Bebber, D. P., Fricker, M. D., … Nakagaki, T. (2010). Rules for biologically inspired adaptive network design. *Science*, *327*(5964), 439–442.

The Marine Biological Laboratory. (2013). uBio. Retrieved November 19, 2013, from http://uio.mbl.edu/

Thiriet, C. (2004). Analysis of chromatin assembled in vivo using exogenous histones in *Physarum polycephalum*. *Methods*, *33*(1), 86–92.

Thiriet, C., & Hayes, J. J. (1999). Histone proteins in vivo: cell-cycle-dependent physiological effects of exogenous linker histones incorporated into *Physarum polycephalum*. *Methods*, *17*(2), 140–150.

Thiriet, C., & Hayes, J. J. (2005). Replication-independent core histone dynamics at transcriptionally active loci in vivo. *Genes Dev*, *19*(6), 677–82.

Thiriet, C., & Hayes, J. J. (2009). Linker histone phosphorylation regulates global timing of replication origin firing. *J Biol Chem*, *284*(5), 2823–9.

Tokumoto, M., Horiguchi, R., Nagahama, Y., Ishikawa, K., & Tokumoto, T. (2000). Two proteins, a goldfish 20S proteasome subunit and the protein interacting with 26S proteasome, change in the meiotic cell cycle. *Eur J Biochem*, *267*(1), 97–103.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–11.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., … Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–78.

Tsuda, S., Zauner, K.-P., & Gunji, Y.-P. (2007). Robot control with biological cells. *BioSystems*, *87*(2-3), 215–23.

Turnock, G., Morris, S. R., & Dee, J. (1981). A comparison of the proteins of the amoebal and plasmodial phases of the slime mould, Physarum polycephalum. *Eur J Biochem*, *115*(3), 533–538.

U.S. Department of Energy. (1992). DOE Human Genome Program: Primer on Molecular Genetics, 1–44.

Vandekerckhove, J., Van Damme, J., Vancompernolle, K., Bubb, M. R., Lambooy, P. K., & Korn, E. D. (1990). The covalent structure of Acanthamoeba actobindin. *J. Biol. Chem.*, *265*(22), 12801–12805.

Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, *270*(5235), 484–487.

Visomirski-Robic, L. M., & Gott, J. M. (1995). Accurate and efficient insertional RNA editing in isolated *Physarum* mitochondria. *RNA (New York, N.Y.)*, *1*(7), 681–91.

Visomirski-Robic, L. M., & Gott, J. M. (1997). Insertional editing of nascent mitochondrial RNAs in *Physarum*. *Proc Natl Acad Sci USA*, *94*(9), 4324–9.

Vogt, V. M., & Braun, R. (1977). The replication of ribosomal DNA in Physarum polycephalum. *Eur J Biochem*, *80*(2), 557–66.

Vonk, F. J., Casewell, N. R., Henkel, C. V, Heimberg, A. M., Jansen, H. J., McCleary, R. J. R., … Richardson, M. K. (2013). The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc Natl Acad Sci USA*, *110*(51), 20651–6.

Wang, C., Wei, L., Guo, M., & Zou, Q. (2013). Computational approaches in detecting non- coding RNA. *Curr Genomics*, *14*(6), 371–7.

Wang, D., & Bodovitz, S. (2010). Single cell analysis: the new frontier in "omics". *Trends in Biotechnology*, *28*(6), 281–290.

Wang, L. (2012). EVER-seq (Evaluate Experiment of RNA-seq). Retrieved from http://code.google.com/p/ever-seq/

Wang, M., Guerrero, F. D., Pertea, G., & Nene, V. M. (2007). Global comparative analysis of ESTs from the southern cattle tick, Rhipicephalus (Boophilus) microplus. *BMC Genomics*, *8*(1), 368.

Wang, X., Chen, W., Huang, Y., Sun, J., Men, J., Liu, H., … Yu, X. (2011). The draft genome of the carcinogenic human liver fluke *Clonorchis sinensis*. *Genome Biol*, *12*(10), R107.

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

Waterborg, J. H., Fried, S. R., & Matthews, H. R. (1983). Acetylation and methylation sites in histone H4 from *Physarum polycephalum*. *Eur J Biochem*, *136*(2), 245–52.

Waterborg, J. H., & Matthews, H. R. (1984). Patterns of histone acetylation in *Physarum polycephalum*. H2A and H2B acetylation is functionally distinct from H3 and H4 acetylation. *Eur J Biochem*, *142*(2), 329–35.

Watkins, R. F., & Gray, M. W. (2008). Sampling gene diversity across the supergroup Amoebozoa: large EST data sets from *Acanthamoeba castellanii, Hartmannella vermiformis, Physarum polycephalum, Hyperamoeba dachnaya* and *Hyperamoeba* sp. *Protist*, *159*(2), 269–281.

Weiss, R. S., Leder, P., & Enoch, T. (2000). A conserved role for the Hus1 checkpoint protein in eukaryotic genome maintenance. *Cold Spring Harbor Symp Quant Biol*, *65*, 457–66.

Werner-Felmayer, G., Golderer, G., Werner, E. R., Gröbner, P., & Wachter, H. (1994). Pteridine biosynthesis and nitric oxide synthase in *Physarum polycephalum*. *Biochem J*, *304 (Pt 1)*, 105–11.

Whittaker, P. A., & Hardman, N. (1980). Methylation of nuclear DNA in Physarum polycephalum. *Biochem J*, *191*(3), 859–62.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., … Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Rev. Genet*, *8*(12), 973–82.

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., … Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, *453*(7199), 1239–1243.

Wohlfarth-Bottermann, K. E. (1979). Oscillatory contraction activity in *Physarum*. *J Exp Biol*, *81*, 15–32.

Wolke, A., Niemeyer, F., & Achenbach, F. (1987). Geotactic behavior of the acellular myxomycete *Physarum polycephalum*. *Cell Biol Intl Rep*, *11*(7), 525–528.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., & Bork, P. (2011). iPath2.0: interactive pathway explorer. *Nucleic Acids Res*, *39*(Web Server issue), W412–5.

Yamada, Y., Wang, H. Y., Fukuzawa, M., Barton, G. J., & Williams, J. G. (2008). A new family of transcription factors. *Development*, *135*(18), 3093–3101.

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Rev Genet*, *13*(5), 329–42.

Yang, F., Demma, M., Warren, V., Dharmawardhane, S., & Condeelis, J. (1990). Identification of an actin-binding protein from *Dictyostelium* as elongation factor 1a. *Nature*, *347*(6292), 494–496.

Yap, K. L., Yuan, T., Mal, T. K., Vogel, H. J., & Ikura, M. (2003). Structural basis for simultaneous binding of two carboxy-terminal peptides of plant glutamate decarboxylase to calmodulin. *J Mol Biol*, *328*(1), 193–204.

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., … Wang, J. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res*, *34*(Web Server issue), W293–7.

Young, R. S., Marques, A. C., Tibbit, C., Haerty, W., Bassett, A. R., Liu, J.-L., & Ponting, C. P. (2012). Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. *Genome Biol Evol*, *4*(4), 427–42.

Yu, L. P., Miller, A. K., & Clark, S. E. (2003). POLTERGEIST encodes a protein phosphatase 2C that regulates CLAVATA pathways controlling stem cell identity at *Arabidopsis* shoot and flower meristems. *Curr Biol*, *13*(3), 179–188.

Zdobnov, E. M., & Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics, 17*(9), 847–8.

Zerbino, D. R., & Birney, E. (2008a). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, *18*(5), 821–829.

7. Appendix

Appendix 1. Calculation of RNA-seq Reads per mRNA molecule.

The analysis of differential gene expression usually requires normalization, to adjust the samples and sequencing runs, into a single and common scale. The most common normalization method for RNA-seq sequencing outputs is the number of reads per kilobase of transcript per million of mapped reads (RPKM), which normalizes the read counts of a given transcript to its length and the total number of mapped reads (Mortazavi et al. 2008). However, when working with single cells and to have a practical cutoff value of expression, it is also possible to calculate the number of RNA-seq reads per each nuclei, and therefore on each single cell (Parikh et al. 2010).

The procedure involves using the extracted mRNA mass, the molar mass of a ribonucleotide, and the Avogadro number, in the mass and number concentration equations. For example, from each sample of a wild-type single-cell *Physarum* plasmodium, 100 ug of total RNA were extracted on average. The average assembled contig length is 847 bp (Chapter 4) and the average molecular weight (or molar mass) of a ribonucleotide monophosphate is 339.5 gr/mol. Assuming that total RNA contains 4% mRNA (4 ug), therefore I estimated the number of transcripts per cell represented by each RNA-seq read as follows:

$$\frac{4\times10^{-6}\,gr\ \mathrm{mRNA}\times 6.022\times10^{23}}{847bp\times 339.5gr/mol}=8.4\times10^{12}\ transcripts\ per\ cell$$

Since the plasmodium consists of 10^8 nuclei (Burland et al. 1993), the number of transcripts per nucleus is:

$$\frac{8.4 \times 10^{12}\, transcripts/cell}{10^{8}\, nuclei/cell} = 84{,}000\ transcripts\ per\ nucleus$$

Considering an average of 2×10^7 mRNA reads per RNA-seq lane (Chapter 4), then the number of transcripts represented by a sequencing read is:

$$\frac{84{,}000\, transcripts/nucleus}{2 \times 10^{7}\, reads/run} = 0.0042\, transcripts/read$$

Therefore, each RNA-seq mapped read represents approximately 0.004 transcripts per nucleus, so 240 reads represent approximately 1 mRNA molecule per nucleus in our analyses of the WT31 strain (Chapter 4). In this same study, the differential expression analysis with the deseq library from R (Anders and Huber 2010; R Core Team 2013), was performed over contigs with combined count of 300 mapped reads (1,26 mRNA molecule per nucleus), to reduce noisecaused by spurious contigs and alignments.

Appendix 2. Estimating the number of protein coding genes in *Physarum*

Hou Y, Lin S (2009) Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes. *PLoS ONE* 4(9): e6978.

Hou and Lin (2009) found a relation between the genome size (S, in Kb), and the protein-coding gene number (P):

$$y' = \ln\ (-46.2 + 22.217\, x')$$

Where x' and y' represent:

$$x' = \log\ (S)\, ; y' = \log\ (P)$$

Therefore, for the *Physarum* genome (S = 3 x 10^6 Kb):

$$P = 10^{\ln\ (-46.2+22.217\ \log 3\ \times 10^6)} = 38{,}187.75\ \approx 38{,}188$$

there must be over 38 thousand protein-coding genes.

Selbstständigkeitserklärung

Walter Israel Barrantes Bustinza, Kurt-Schumacher-Strasse 12, 38102
Braunschweig

Hiermit erkläre ich, dass ich die von mir eingereichte Dissertation zum Thema

„**The transcriptomic networks controlling the sporulation in *Physarum polycephalum***"

selbstständig verfasst, nicht bereits als Dissertation verwendet und die benutzten Hilfsmittel und Quellen vollständig angegeben habe.

Weiterhin erkläre ich, dass ich weder diese noch eine andere Arbeit zur Erlangung des akademischen Grades doctor rerum naturalium (Dr. rer. nat.) an anderen Einrichtungen eingereicht habe.

Braunschweig, 01 September 2015

Walter Israel Barrantes Bustinza