Nucleic Acids Research Advance Access published November 14, 2014

Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku1147

An integrated approach for genome annotation of the eukaryotic thermophile *Chaetomium thermophilum*

Thomas Bock[1], Wei-Hua Chen[1], Alessandro Ori[1], Nayab Malik[1], Noella Silva-Martin[1], Jaime Huerta-Cepas[1], Sean T. Powell[1], Panagiotis L. Kastritis[1], Georgy Smyshlyaev[1,2], Ivana Vonkova[1], Joanna Kirkpatrick[3], Tobias Doerks[1], Leo Nesme[1], Jochen Baßler[4], Martin Kos[4], Ed Hurt[4], Teresa Carlomagno[1], Anne-Claude Gavin[1], Orsolya Barabas[1], Christoph W. Müller[1], Vera van Noort[1], Martin Beck[1,*] and Peer Bork[1,*]

[1]European Molecular Biology Laboratory (EMBL), Structural and Computational Biology Unit, Meyerhofstrasse 1, D-69117 Heidelberg, Germany, [2]Institute of Cytology and Genetics, Laboratory of Molecular Genetic Systems, 630090 Novosibirsk, Russia, [3]European Molecular Biology Laboratory (EMBL), Proteomics Core Facility, Meyerhofstrasse 1, D-69117 Heidelberg, Germany and [4]Biochemie-Zentrum der Universität Heidelberg, INF328, D-69120 Heidelberg, Germany

Received August 11, 2014; Revised October 21, 2014; Accepted October 27, 2014

ABSTRACT

The thermophilic fungus *Chaetomium thermophilum* holds great promise for structural biology. To increase the efficiency of its biochemical and structural characterization and to explore its thermophilic properties beyond those of individual proteins, we obtained transcriptomics and proteomics data, and integrated them with computational annotation methods and a multitude of biochemical experiments conducted by the structural biology community. We considerably improved the genome annotation of *Chaetomium thermophilum* and characterized the transcripts and expression of thousands of genes. We furthermore show that the composition and structure of the expressed proteome of *Chaetomium thermophilum* is similar to its mesophilic relatives. Data were deposited in a publicly available repository and provide a rich source to the structural biology community.

INTRODUCTION

Thermophilic proteins are known to be more stable than their mesophilic counterparts (1,2) and have been successfully utilized by structural biologists for the three-dimensional characterization of many proteins and protein complexes. However, until recently, this approach had been restricted to bacteria and archaea, because thermophilic eukaryotes are rare and often difficult to culture in the laboratory. This limitation can be overcome with the recent genome sequencing of several thermophilic fungi. In particular *Chaetomium thermophilum* (*Ct*), which has an optimal growth temperature of 50–55°C and is cultivatable in several standard media, is developing into a powerful model organism in which several proof of principle studies on three-dimensional characterization have already been performed (3–8). In total, 19 Protein Data Bank (PDB) entries using *Ct* have already been deposited since its genomic sequence became available in 2011 (3), 12 of which occurred in 2013 (Figure 1), illustrating the importance of genomic information as a prerequisite for structural investigations.

The original analysis of the 28.3 Mb *Ct* genome (20 scaffolds) (3) was based on DNA sequencing, *ab initio* gene prediction and automatic annotation, which identified 7227 protein coding genes. The accuracy of gene models, solely based on genomic sequence data, relies entirely on gene prediction algorithms, which are known to be far from accurate (9). To further facilitate the use of *Ct* as a model organism for structural biology, we have refined its genome annotation using a strategy that integrates large-scale proteomics and next-generation RNA sequencing data sets with biochemical and bioinformatics analysis. We identified various novel and previously non-annotated genes, corrected intron-exon structures, improved confidence in expression of gene termini and annotated non-coding RNAs. We con-

*To whom correspondence should be addressed. Tel: +49 6221 387 8526; Fax: +49 6221 387 8517; Email: bork@embl.de
Correspondence may also be addressed to Martin Beck. Tel: +49 6221 387 8267; Fax: +49 6221 387 8519; Email: martin.beck@embl.de
Present Addresses:
Wei-Hua Chen, Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland.
Nayab Malik, University of Oxford, Trinity College, Oxford, OX1 3BH, UK.
Vera van Noort, KU Leuven, Center of Microbial and Plant Genetics, 3001 Leuven, Belgium.

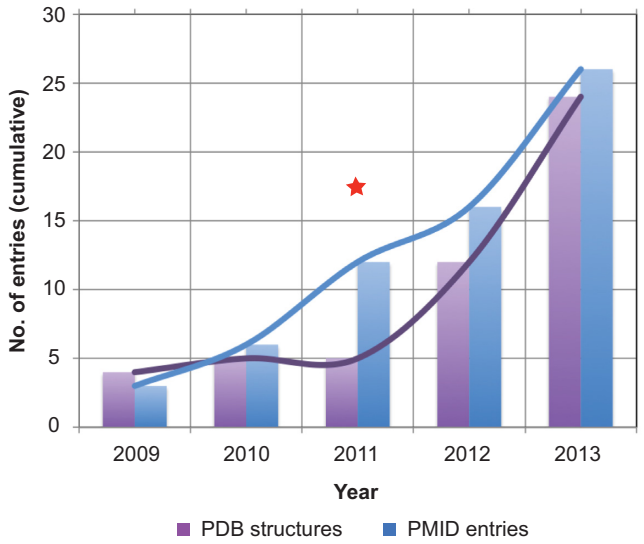

Figure 1. Literature overview of PDB-deposited structures and scientific publications derived from or referring to *Chaetomium thermophilum* proteins. Structures deposited before initial genome sequencing in 2011 were enabled by access to partial genome information.

sidered a multitude of individual biochemical experiments as a validation for the annotated gene models. We improved the functional characterization of 2853 genes and provided a detailed analysis of genomic repeats and potential transposable elements in *Ct*. The proteomic analysis alone is the first of its kind in a thermophilic eukaryote and quantifies the expression of 4297 genes at the protein level. By comparing the global protein expression pattern of *Ct* to a mesophilic relative, we demonstrate that there is little deviation of protein abundances within orthologous groups. These findings suggest that adaption to thermophily primarily occurs at the individual protein level and underline the suitability of Ct as a model organism for structural biology studies of eukaryotes. We have integrated the data into a publicly available database that should serve as a rich source for the scientific community.

MATERIALS AND METHODS

Cell culture

Chaetomium thermophilum was obtained from Deutsche Sammlung von Mikroorganismen und Zellkulturen (DMSZ No.: 1495). All cultures were grown under standard media conditions as defined by the DMSZ.

RNA sequencing

Whole RNA was extracted from *Ct* grown at 50°C using RNeasy kit (Quiagen) according to the procedure described by the manufacturer. The mRNA was derived by poly-A purification and subjected to next-generation sequencing (Illumina RNAseq). In total 74 million reads were obtained by paired-end reading (50 base), subjected to trimming using FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html), quality control (reads < = 30 bp in size after trimming were removed) and mapped to the published *Ct* genome (2011) using Bowtie 2 (10) with default parameters.

Preparation of proteomic samples

Proteins were extracted from *Ct* cells grown under standard conditions in 4 M urea buffer containing 0.2% (w/v) Rapigest® detergent. Carbamidomethylation, enzymatic digest and peptide purification was performed as previously described (11). Purified peptide mixtures were lyophilized in a vacuum concentrator and stored at –20°C until further use.

Peptide fractionation

To maximize proteome coverage, peptide digests from *Ct* lysate were subjected to different high pH reversed phase separation strategies as well as strong anion exchange (SAX) chromatography. The collected data were subsequently combined.

High pH reversed phase fractionation: peptides were fractionated on an Agilent 1200 Infinity HPLC system with a Gemini C_{18} column (3 μm, 110 Å, 100 × 1.0 mm, Phenomenex) using a linear 60 min gradient from 0% to 35% (v/v) acetonitrile in 20 mM ammonium formate (pH 10) at a flow rate of 0.1 ml/min. Alternatively, a stepwise gradient at the same flow rate was used (6 min steps of increasing acetonitrile (v/v) content: 11.1%, 14.5%, 17.4%, 20.8%, 45%). Elution of peptides was detected with a variable wavelength UV detector set to 254 nm. Thirty two fractions were collected along with the linear LC separation that were subsequently pooled into six fractions using a post-concatenation strategy as previously described (12). Five fractions were collected using the stepwise elution.

SAX *chromatography:* SAX fractionation of peptides was performed as previously described (13). In brief, six fractions were obtained by elution of peptides under varying pH conditions (flow through, pH 12, pH 8, pH 6, pH 4, pH 2). All fractions were dried in a vacuum concentrator and then stored at −80°C until LC-MS/MS analysis.

MS analysis

Peptides were separated with a BEH300 C18 (75 μm x 250 mm, 1.7 μm) nanoAcquity UPLC column (Waters) using a stepwise 145 min gradient from 3% to 85% (v/v) acetonitrile in 0.1% (v/v) formic acid at a flow rate of 300 nl/min. The LTQ-Orbitrap Velos Pro instrument was operated in data-dependent mode. Parameters for the CID-based method used one survey MS scan acquired in the orbitrap followed by up to 20 fragmentation scans (TOP20) of the most abundant ions analysed in the LTQ. Only charge states of two and higher were allowed for fragmentation. Essential MS settings were: full MS: AGC = 10^6, maximum ion time = 500 ms, m/z range = 375–1600, resolution = 30 000 FWHM; MS2: AGC = 30 000, maximum ion time = 50 ms, minimum signal threshold = 1500, dynamic exclusion time = 30 s, isolation width = 2 Da, normalized collision energy = 40, activation Q = 0.25.

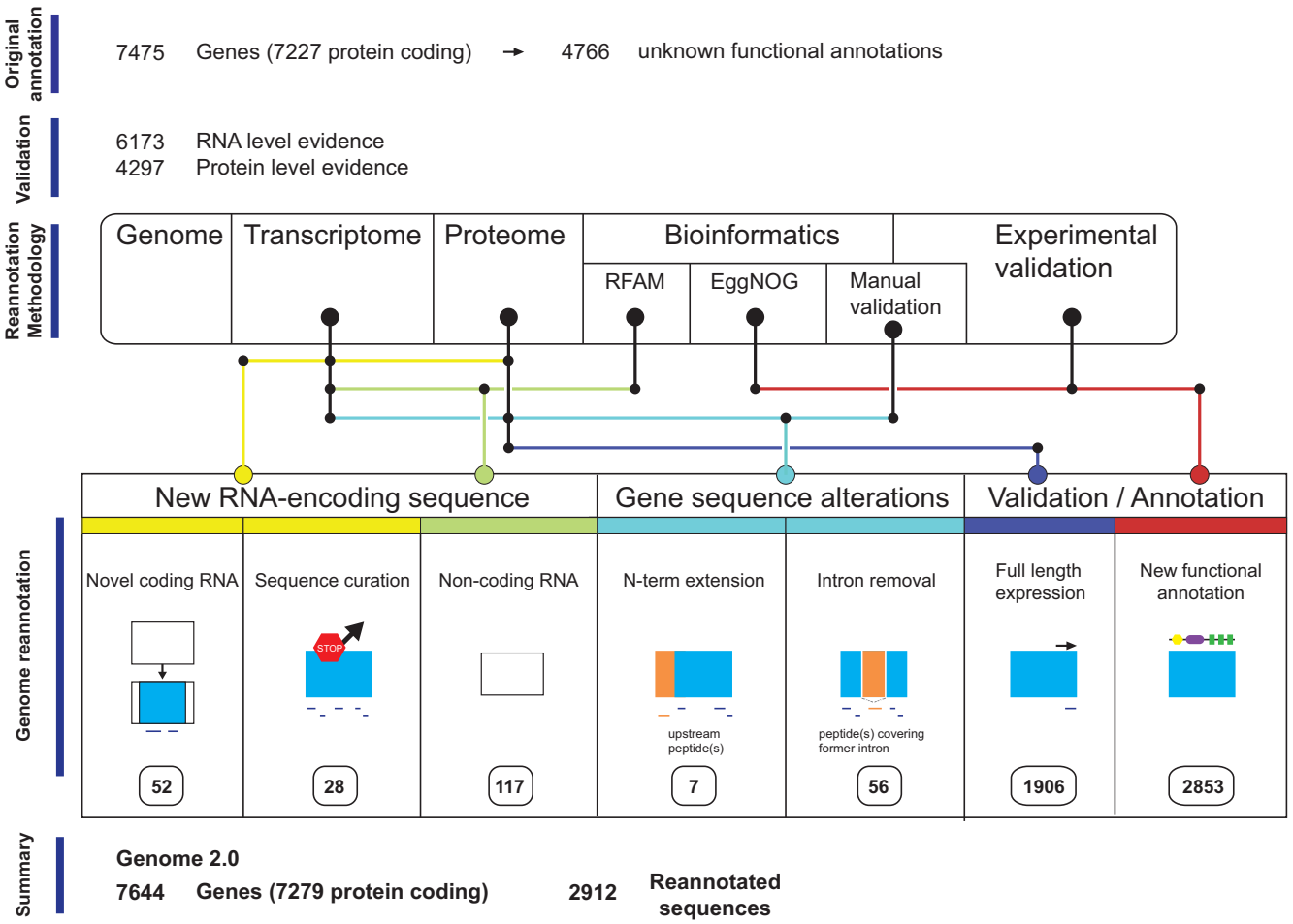

Figure 2. Summary and added value of the analyses performed on the experimental transcriptomics and proteomics data, which included: refinement of intron/exon structures, analysis of previously predicted ORFs that contain a stop codon, *de novo* ORF/peptide analysis and expression analysis of protein termini.

MS database search

MS peptide spectra were searched against the *Ct* database of nuclear genome-encoded proteins (published in 2011) (3) including common protein contaminants using the MaxQuant search engine (version 1.305) (14). A protein and peptide false discovery rate (FDR) of 1% was determined by target-decoy-based search (reverse database search). All peptides showing a score < 60 were excluded from the analysis.

Major protein ID was used for identified protein groups. For the proteomic validation of novel sequence annotations, dedicated databases were used: an additional 63 gene models were contained, which were originally excluded from the UniProt and EMBL genome databases because they contained a STOP codon close to the 5' region. For the validation of intron retention (de novo exons), a database was used in which 3098 protein-coding sequences were modified by including additional sequence stretches previously annotated as introns. For the validation of novel genes, a *6-frame* ORF database was created containing proteins concatenated to translations in all six reading frames of parts of the genome that do not contain annotated genes. Proteome data are available for download at the new genome browser website (see visualization paragraph in the results section).

Orthologous group mapping

Functional annotation was performed by mapping the unannotated *Ct* proteins to annotated orthologous groups in eggNOGv3 (15). This was done by a best-hit homology search using Basic Local Alignment Search Tool (BLAST). To prevent erroneous mappings, all groups within eggNOGv3 were analysed based on the group variance and outgroup scores. This helped to rule out paralogs and distant homologous genes and provide the closest possible functional annotation. The results were manually validated.

Artificial fusions

Potential fusion products were found by searching for orthologs in other genomes. If two separate genes in other fungal genomes were found to be fused in the Ct predicted protein coding set, it was marked as suspicious. Manual inspection revealed missed introns in many cases that then caused an artificial fusion of protein coding genes.

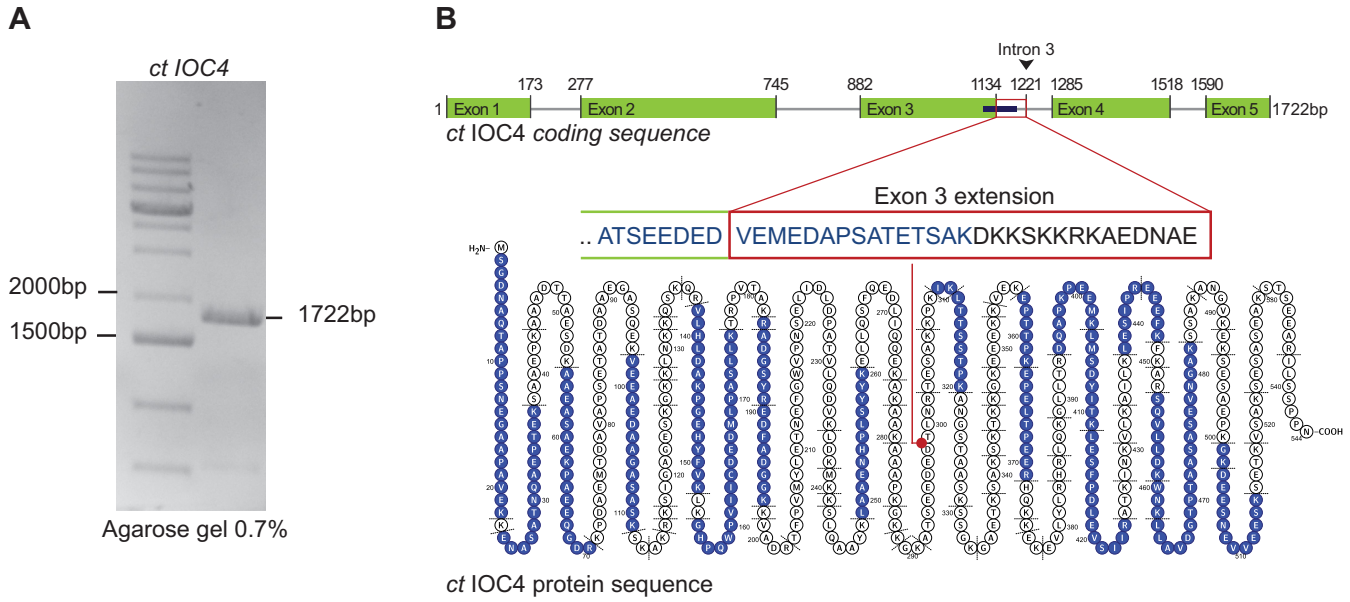

Figure 3. Example of experimental validation of gene sequence reannotation. (**A**) Reverse transcriptase polymerase chain reaction of the *ct* IOC4 (CTHT_0009460) gene. Size of gel band was high compared to annotated gene model (> 1722 nt instead of 1635 nt). Gene sequencing (RT-PCR) reveals 87 nt additional sequence extending exon 3. The extended sequence partially matches intron 3 of the originally annotated *ct* IOC4 sequence. (**B**) Original *Ct* IOC4 coding sequence (top) and MS-identified peptides overlaid to the Protter protein sequence view (bottom). The expression of assumed exon 3 extension initially found by RT-PCR (red box) was verified by MS-based identification of the peptide sequence ".ATSEEDEDVEMEDAPSATETSAK." (blue bar) which covers the MS-detectable part of the assumed exon 3 extension. The insert site of the exon 3 extended sequence is indicated (red dot) in the Protter (29) protein sequence image of *ct* IOC4, together with all other MS-identified peptides (highlighted in blue, N- and C-terminus and potential tryptic cleavage sites for MS indicated). An alternative splice variant for *ct* IOC4 containing the extended exon 3 sequence is included in the reannotated *Ct* genome.

RFAM predictions

To identify putative structural RNA elements, 'infernal' tool (16) (version 1.1) was used to search the *Ct* genome against the RFAM database (version 11) containing HMM models of 21 283 RNAs (17).

De novo gene prediction and validation

RNAseq transcripts that match the *Ct* genome, but do not match to any predicted gene model, were grouped in pseudo-contigs using Velvet (18). All six frames were translated into pseudo-proteins by the EMBOSS Transseq tool (19). Pseudo-proteins were concatenated to the originally published sequences. Search results were additionally validated by protein BLAST. Prediction of protein function was performed by Blast2GO (20) and manual BLAST search.

Genomic repeat analysis

Repetitive elements in the genome of *Ct* were mined and classified using RepeatModeller *de novo* repeat identification package (http://www.repeatmasker.org/RepeatModeler.html). The resulting data set of *de novo* identified repeats was combined with a collection of repetitive DNA deposited to the public repeat database, RepBase (http://www.girinst.org/server/RepBase/index.php), and integrated into a single library. This repeat library was then used to screen the *C. thermophilum* genome using RepeatMasker (http://www.repeatmasker.org/).

RESULTS

An integrated approach for annotation

To achieve a much more accurate and reliable genome annotation, the integration of experimental evidence obtained at the RNA and protein levels is invaluable as it validates the gene, pseudo-gene, intron, exon, START and STOP predictions inferred from sequence data. In order to employ such an integrative strategy for the thermophilic eukaryote *Ct,* we acquired two large-scale data sets using next-generation RNA sequencing and state-of-the-art shotgun proteomics technology. For the RNAseq data, approximately two thirds of the high-quality reads that mapped to the genome matched an existing gene model. These resulted in a high coverage of the *Ct* transcriptome, matching 6173 (85%) of the previously predicted open reading frames (ORFs; FPKM cut-off = 5). Another third of the genome mapped reads matched to the genome, but not to any previously predicted gene model. This finding implied various distinctly spliced, elongated, yet undetected genes or the existence of non-coding RNA that remained undetected during first genome annotation (Supplementary Figure S1A). For the proteome data, the combined data set allowed us to identify 4297 proteins from 44620 unique peptide sequences in *Ct*. Protein and peptide identifications are visualized in the new *Ct* genome browser (see visualization paragraph). Overall, the experimentally identified proteome covered 59.5% of the previously predicted *Ct* ORFs (Supplementary Figure S1B). This value is in line with large-scale data sets recently obtained of *Saccharomyces cerevisiae*

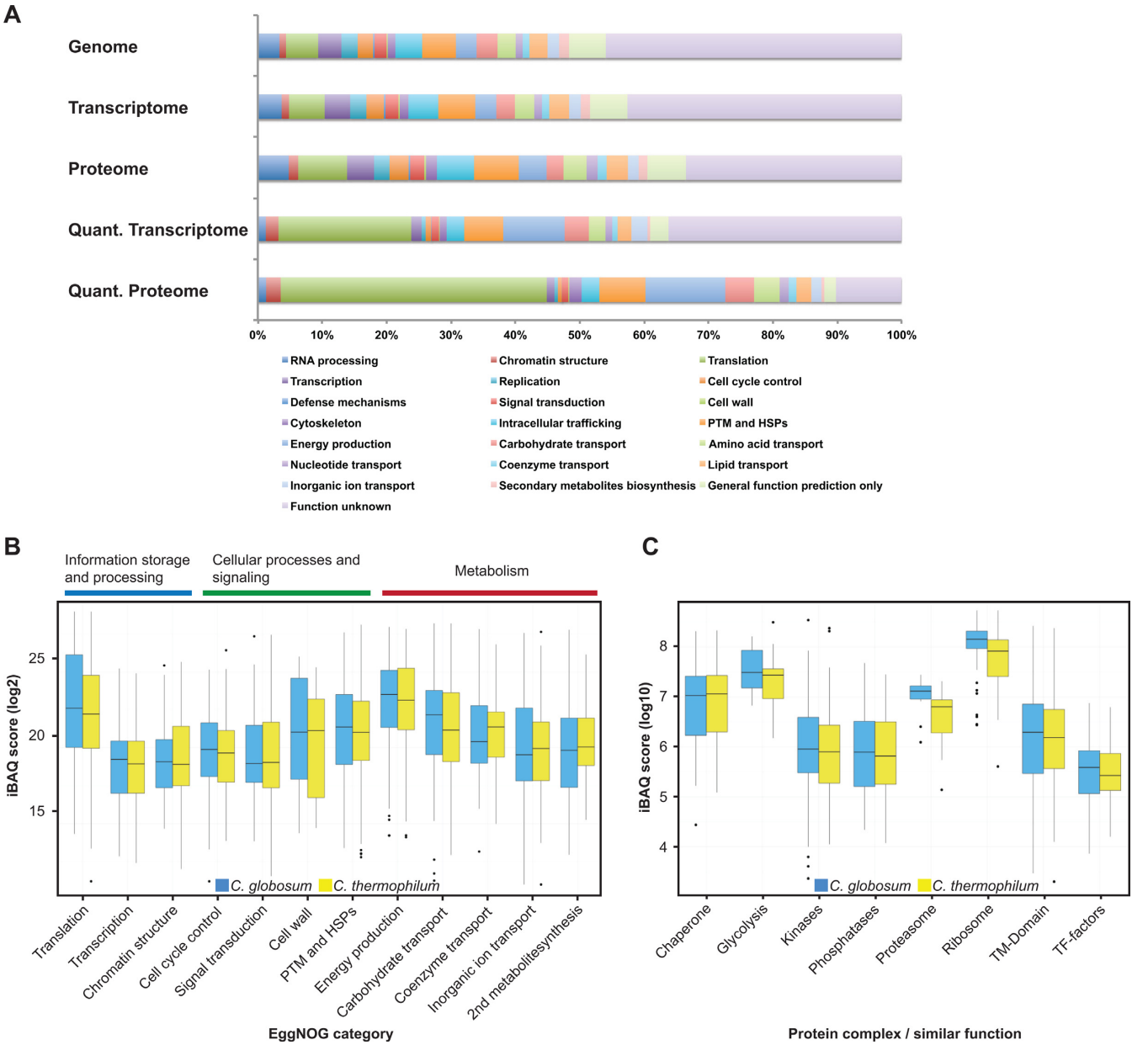

Figure 4. Abundance of functional protein category levels. (**A**) Qualitative and quantitative overview of the fraction of the genome, transcriptome and proteome dedicated to defined functional categories. 'Genome' refers to the number of genes, 'Transcriptome' to the number of identified mRNAs, 'Proteome' to the number of identified proteins, 'Quantitative transcriptome' to mRNA abundances and 'Quantitative proteome' to protein abundances within any given functional category. (**B**) Comparison of protein abundance changes in selected functional protein categories between the thermophile, *Chaetomium thermophilum,* and the mesophile, *Chaetomium globosum*. Functional categories were selected from the three present main categories available in eggNOG. (**C**) Comparison of protein abundance in selected major protein complexes and proteins of similar function between *Chaetomium thermophilum* and *Chaetomium globosum*. Protein abundance is based on "intensity-based absolute quantification" scores (iBAQ).

(67% (21), and human cells (50% (22,23)) and highlights the comprehensive nature of our analysis. To quantify the correlation of mRNA and protein abundances in a eukaryotic thermophile, we compared the log-transformed FPKM values of the transcriptome with log transformed protein abundance scores (intensity-based absolute quantification, iBAQ score (24)) from a comparable proteomic data subset. We found a moderate correlation ($R2 = 0.36$, P-value $< 2.2e^{-}16$) over ~4 orders of magnitude (Supplementary Figure S1C), which is very similar to values previously observed for other species (24). Next, we took advantage of the combined data set to refine the *Ct* genome annotation and employed several integrated strategies (Figure 2).

De novo gene search. To identify previously undetected protein coding genes, we generated pseudo-contigs out of the high-quality RNAseq reads that did not match to any annotated *Ct* ORF (one third of the high-quality genome mapped reads). We then compared these to our proteomic data set for expression evidence at the protein level. We

identified 72 peptides mapping to 51 of the pseudo-contigs (<0.01 FDR). The average length of the identified new genes was 273 nt (median: 174 nt). BLAST analysis identified homologous sequences for 18 out of the 51 pseudo-contigs. Taken together the multiple lines of evidence from RNAseq, MS-based proteomics and BLAST analysis we considered the 51 identified pseudo-contigs as new coding genes (Supplementary Table S1A). The suggested *de novo* genes were included into the reannotated genome of *Ct* and are highlighted in the new version of the genome browser (see visualization paragraph).

Intron retention search. A number of failed cloning attempts by collaborators (personal communication) hinted at incorrect annotation of gene structure and prompted us to validate splice sites and gene boundaries. We therefore searched our proteomic data for peptides that match into introns. We identified intron-retained sequences derived from 54 *Ct* proteins that originated, at least in part, from intron-labeled sequences and annotated 54 novel exons (Supplementary Table S1B). For this analysis, feedback from the co-authors was considered together with the transcriptome and proteome data (Figure 3A and B).

Pseudo-gene removal. We also used the peptide data to re-analyse 63 gene models, which were categorized as pseudogenes in the initial annotation (3), because of intrinsic STOP codons and were thus not listed in the public gene browser or protein repositories, although they have obvious homologues sequences in other species. In 28 out of the 63 cases, our proteomic data confirm the expression of the respective gene (Supplementary Table S1C). In total, 263 peptides were identified for these 28 proteins. For 13 of these 28 proteins, we identified peptides located at either side (N- and C-terminal) of the putative incorrectly annotated STOP codon. From the remaining 15 proteins, we identified peptides located at either the N-terminal side (four proteins) or the C-terminal side (eleven proteins) of the putative incorrectly annotated STOP codon. Our RNAseq data support the expression of the whole-length of the 28 protein-coding genes. All 28 sequences are now annotated as protein coding in the reannotated *Ct* genome and are shown in the new version of the genome browser (see visualization paragraph). In the other 35 cases, evidence at the protein level was lacking, the RNAseq data suggest extremely low abundance of the transcript and the data quality was insufficient to justify a reannotation.

N-terminal extension and full-length expression analysis. Encouraged by this result, we set out to confirm gene boundaries based on our proteomic data set. We identified 12 peptides located in the pre-N-terminal region of 7 *Ct* proteins. The peptides did not match to any other known protein coding ORF of *Ct*. We verified those N-terminal extensions using RNA transcript data and reannotated the genes with alternative splice variants (Supplementary Table S1D).

Inspired by the intron-retention detected in the cases of some *Ct* genes, we queried our proteomic data to verify the full-length expression of proteins (Table 1). We found peptides close to the predicted protein termini (distance of <25 AA) of 1738 N-termini and 1906 C-termini, corresponding to 40% and 44% of the MS-detectable proteome, respectively. For almost one quarter of all gene models for which evidence at the protein level was found (938 proteins), we detected peptide matches to both the N- and C-termini. Since the sequence coverage of proteins is stochastically dependent on peptide detectability in the mass spectrometer, this verification cannot be provided comprehensively. It is nevertheless very useful in the case where verified genes are targeted for low throughput biochemical and structural analyses.

Refined functional annotation by bioinformatics analysis. In relation to other model organisms, such as *Saccharomyces cerevisiae* (approximately 82% functionally annotated genes based on UniProt database entries), the originally published *Ct* genome contained a high number of proteins with unknown function (approximately 4669 of 7227 protein coding genes, 65%). To close that gap, we mapped unannotated *Ct* proteins to orthologous groups in eggNOGv3 (15). From that analysis, we added functional information to a large number of genes with previous unknown annotation. Thus, the number of *Ct* genes with a predicted function increased to 5684 (> 70% of all gene models) (Supplementary Table S2).

RFAM prediction. To annotate non-coding RNAs, we used *RFAM*-based predictions (17). In total, 219 putative structural RNA elements were identified. Of them, 102 overlap with known tRNAs and rRNAs. We thus also annotated 117 newly identified RNA elements. Most of the novel ncRNAs belong to the small nuclear RNA (snoRNA) or fungal signal recognition particle RNA families; they are often highly expressed, e.g. about 41% of them have expression abundance > = 10 RPKM (Reads Per Kilobase of transcript per Million mapped reads), as compared with 31% of the 102 known ones, implying functional importance under normal growth conditions.

Genomic repeat analysis. To gain insights into the part of the *Ct* genome that does not contain any gene models, we performed a genomic repeat analysis. We found a surprisingly small part (2.5%) of the *Ct* genome accounts for genomic repeats. These included tandem repeats, satellites, uncharacterized repeats and transposable elements (Table 2). We found that the most abundant group of transposable elements in the genome of Ct belongs to the Tad1 superfamily of non-long terminal repeat (LTR) retrotransposons, which is ubiquitous in fungi (25). These Tad1 transposons are present in highly degenerated and diverged copies, which are probably remnants of previously active elements. Together, they occupy about 1% of the genome. Notably, no intact copies of any other transposable element were detected in the *Ct* genome, suggesting that transposable elements are inactive in this thermophilic organism. In total, 3803 repeats were added to the new version of the genome. The lack of repeat sequences seems to be a general feature for thermophilic eukaryotes (26,27).

Ad hoc manual annotation and expression support. As *Ct* is becoming increasingly important for structural studies and the resource (version 1) is public (3), we received experimentally validated corrections or novel annotations of 28 genes,

Table 1. *Chaetomium thermophilum* proteins with direct proteomic evidence of N-terminal and C-terminal expression

	Termini match	1st Met omitted match[a]	Sum terminal matches	Match in 25 terminal AA	Fraction of total proteome
N-terminus	194	822	985	1738	24.0
C-terminus	665	-	665	1906	26.4

[a]Peptides matching N-terminus without the START codon (Met) are counted as termini match.

Table 2. Genomic repeat analysis in *Chaetomium thermophilum*

Repeat type	Length occupied (bp)	Percentage of sequence (%)
Interspersed repeats	721 491	2.55
Retroelements	342 067	1.21
LINEs	276 216	0.98
Tad1	236 085	0.83
LTR elements	64 479	0.23
Gypsy	37 960	0.13
Copia	10 775	0.04
DNA transposons	39 456	0.14
Unclassified	339 968	1.20
Satellites	3877	0.01
Simple repeats	9302	0.03

which we also included in the reannotated *Ct* genome. In 20 cases, genes were shortened based on RNAseq data or manual inspection. In two cases, an intron was removed from the gene. In one case, a new locus was identified. In another five cases, artificial gene fusions were corrected. Furthermore, we collected manually validated expression evidence of 66 *Ct* genes and 27 *Ct* protein domains (Supplementary Table S3; personal communication). The data from the manual feedback are highlighted in the new version of the genome browser (see visualization paragraph).

Visualization of the CTHT v2.0 genome/proteome. To promote the usage of *Ct* as a model organism, public deposition as well as the possibility to browse and conveniently visualize data is essential. Experimental data and updated reannotation data are publicly available at the EMBL Genome Browser database (based on Genome Maps (28)) and can be accessed under: http://ct.bork.embl.de/ctbrowser/. The updated reannotation data are further available at NCBI (Bioproject ID PRJNA47065). Parallel to the reannotated genome data, the browser contains separated tracks for non-coding RNAs and genomic repeats. As a new feature to the genome browser, the MS-based proteomic evidence for gene expression is overlaid. For this, *Ct* protein sequences are visualized by the Protter software (29) from N-terminus to C-terminus together with an overlay of the MS-identified peptides. Peptide sequences and their positions in the protein sequence are indicated with a blue background.

Protein expression patterns are highly similar to mesophiles

With an improved annotation of genes and proteins in hand, we explored whether the adaption of *Ct* to extreme conditions has altered its protein expression patterns and the overall proteomic structure. To estimate the general relevance of experimental data obtained from *Ct* as a model organism for other eukaryotes, we compared its protein expression levels to mesophilic relatives. Previous analyses of functional categories in eukaryotic mesophiles revealed, for example in *Saccharomyces cerevisiae*, that cellular core functions such as translation, energy production and metabolism, although being conducted by relatively few gene products, are typically overrepresented in the total proteome because of their high abundance (22,30). In contrast, proteins involved in so-called regulatory functions such as replication, cell cycle control, defense mechanisms and secondary metabolite biosynthesis are typically expressed at low copy numbers.

To compare the fraction of the total protein that *Ct* devotes to specific functional categories, we used orthologous group mapping by eggNOGv3 (15). In total, there are 6862 orthologous groups matched to the originally predicted 7227 *Ct* protein coding genes (95%). Of those, 4266 (62%) have proteomics support (Supplementary Figure S2A). We classified the orthologous groups into 13 broad functional categories provided by eggNOG (31) and analysed them for their abundance in the genome, transcriptome and proteome. This way, the number of genes dedicated to a certain function can be compared to the respective fraction of all mRNAs or even proteins. Previous studies found that, for example, few ribosomal genes encode for a large fraction of the quantitative proteome because of the high abundance of the ribosome, and, vice versa, many genes encoding, for example, transcription factors make up only a very minor fraction of all protein (22). We found that the overall *Ct* proteome is similarly structured and devotes a large amount of its total protein copies to cellular core functions (Figure 4A). One might nevertheless assume that thermophiles, because of their adaption to extreme environments, have to devote a larger fraction of their proteomes to the maintenance of proteostasis and thus might have a higher abundance of chaperones, ribosomes or proteases. To unambiguously address this question, we acquired technically consistent proteomic data sets of *Ct* and its closest mesophilic relative, *Chaetomium globosum,* both grown in identical standard medium, and estimated protein abundances. 2974 eggNOGv3 entries were matched between both data sets (Supplementary Figure S2B and S2C). Strikingly, no sig-

nificant difference between protein abundance in the selected major functional groups was detected between the mesophilic and the thermophilic fungus (Figure 4B). We also investigated whether protein complexes or specific protein groups would deviate in their abundance across both organisms (Figure 4C). In both species, the proteasome and ribosome belong to the most abundant protein complexes of the cell and share a narrow distribution of the protein complex components. Their abundance shows a trend toward lower copy numbers in *Ct*. Since a similar trend was observed for glycolytic enzymes, one might surmise that this behavior reflects slightly reduced growth rates at higher temperatures. In contrast, chaperones and heat shock proteins, as well as broader functional categories such as kinases, phosphatases, transmembrane domain containing proteins and transcription factors have similar abundances in both species. We conclude that the heat stable fungus *Ct* has a similarly composed proteome compared to mesophilic organisms.

DISCUSSION

Here, we describe for the first time the genome-wide proteome and transcriptome analysis of a eukaryotic thermophile, the filamentous fungus *Chaetomium thermophilum*. We have refined the annotation of thousands of genes and obtained experimental evidence for their expression under standard laboratory conditions. These publicly available data allow for a straightforward identification of orthologues, the expressed RNA and protein sequence as well as protein abundances in the cell and thus provide a rich source for the structural biology community. Based on the integration of biochemical data, and computational, as well as manual annotation, we have identified various new genes, corrected and verified gene structures of a considerable fraction of the genome, characterized and annotated long non-coding RNAs and genomic repeats, and also predicted functions of many uncharacterized proteins. The refinement of gene structure by our transcriptomics and proteomics approach in particular ensures individual gene expression studies and subsequent experimental characterizations. We believe that the annotation updates provided should reduce redundancy in efforts and should further advance structural biology.

In our study, we do not find major functional implications of high temperature on the cellular protein abundance level. Thus, our experimental findings fit well with recent bioinformatics and deep sequencing studies, which also conclude that changes in protein primary structure lead to thermo-stability of proteins and not the differential expression of thermo-inducible genes (32,33). Although we cannot exclude potential contributions to thermophily from special lipids for example found in archaeal thermophiles (34) or thermoprotection by particular cellular crowding, mostly found in archaea and bacterial thermophiles (35), our proteomics findings imply that the individual proteins are the basis for the adaptation of a thermophilic lifestyle. We believe this reconfirms our annotation effort, which should facilitate further exploitation of *Ct* molecules in diverse research disciplines such as structural biology, systems biology and biotechnology.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We would like to thank Dr Amparo Andres-Pons and Alessia Gagliardi for excellent technical support. We gratefully acknowledge support by EMBL's Proteomics Core facility and Genomics Core facility.

ACCESSION NUMBERS

CTHT_0075750, CTHT_0075780, CTHT_0075790,
CTHT_0075800, CTHT_0075810, CTHT_0075820,
CTHT_0075840, CTHT_0075850, CTHT_0075860,
CTHT_0075870, CTHT_0075880, CTHT_0075890,
CTHT_0075910, CTHT_0075920, CTHT_0075930,
CTHT_0075940, CTHT_0075950, CTHT_0075960,
CTHT_0075970, CTHT_0075990, CTHT_0076020,
CTHT_0076040, CTHT_0076060, CTHT_0076070,
CTHT_0076080, CTHT_0076090, CTHT_0076100,
CTHT_0076140, CTHT_0076150, CTHT_0076170,
CTHT_0076180, CTHT_0076190, CTHT_0076200,
CTHT_0076210, CTHT_0076220, CTHT_0076240,
CTHT_0076250, CTHT_0076260, CTHT_0076270,
CTHT_0076280, CTHT_0076290, CTHT_0076300,
CTHT_0076310, CTHT_0076320, CTHT_0076340,
CTHT_0076360, CTHT_0076380, CTHT_0076410,
CTHT_0076420, CTHT_0076430, CTHT_0076440,
CTHT_0076470, CTHT_0076480, CTHT_0076490,
CTHT_0076500, CTHT_0076510, CTHT_0076520,
CTHT_0076530, CTHT_0076540, CTHT_0076550,
CTHT_0076560, CTHT_0076570, CTHT_0076580,
CTHT_0076590, CTHT_0076600, CTHT_0076610,
CTHT_0076620, CTHT_0076630, CTHT_0076640,
CTHT_0076650, CTHT_0076660, CTHT_0076670,
CTHT_0076680, CTHT_0076690, CTHT_0076700,
CTHT_0076710, CTHT_0076720, CTHT_0076730,
CTHT_0076740, CTHT_0076750, CTHT_0076760,
CTHT_0076770, CTHT_0076780, CTHT_0076790,
CTHT_0076800, CTHT_0076810, CTHT_0076820,
CTHT_0076830, CTHT_0076840, CTHT_0076850,
CTHT_0076860, CTHT_0076870, CTHT_0076880,
CTHT_0076890, CTHT_0076900, CTHT_0076910,
CTHT_0076920, CTHT_0076930, CTHT_0076940,
CTHT_0076950, CTHT_0076960, CTHT_0076970,
CTHT_0076980, CTHT_0076990, CTHT_0077000,
CTHT_0077010, CTHT_0077020, CTHT_0077030,
CTHT_0077040, CTHT_0077050, CTHT_0077060,
CTHT_0077070, CTHT_0077080, CTHT_0077090,
CTHT_0077100, CTHT_0077110, CTHT_0077120,
CTHT_0077130, CTHT_0077140, CTHT_0077150,
CTHT_0077160, CTHT_0077170, CTHT_0077180,
CTHT_0077190, CTHT_0077200, CTHT_0077210,
CTHT_0077220, CTHT_0077230, CTHT_0077240,
CTHT_0077250, CTHT_0077260, CTHT_0077270,
CTHT_0077280, CTHT_0077290, CTHT_0077300,
CTHT_0077310, CTHT_0077320, CTHT_0077330,
CTHT_0077340, CTHT_0077350, CTHT_0077360,
CTHT_0077370, CTHT_0077380, CTHT_0077390,

CTHT_0077400, CTHT_0077410, CTHT_0077420, CTHT_0077430, CTHT_0077440, CTHT_0077450, CTHT_0077460, CTHT_0077470, CTHT_0077480, CTHT_0077490, CTHT_0077500, CTHT_0077510, CTHT_0077520, CTHT_0077530, CTHT_0077540, CTHT_0077550, CTHT_0077560, CTHT_0077570, CTHT_0077580, CTHT_0077590, CTHT_0077600, CTHT_0077610, CTHT_0077620 and CTHT_0077630.

FUNDING

European Molecular Biology Laboratory; European Union's Seventh Framework Programme project "SysteMTb" [eC-Bork-241587 to P.B.]; EMBL Interdisciplinary Postdoc Programme under Marie Curie Actions COFUND [to T.B., N.S.]; German Academic Exchange Service (DAAD) Scholarships [to G.S.]; CellNetworks (Excellence Initiative of the University of Heidelberg) [to M.B., P.B., C.W.M.]. Funding for open access charge: European Molecular Biology Laboratory.
Conflict of interest statement. None declared.

REFERENCES

1. Perutz,M.F. and Raidt,H. (1975) Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. *Nature*, **255**, 256–259.
2. Perutz,M.F. (1978) Electrostatic effects in proteins. *Science*, **201**, 1187–1191.
3. Amlacher,S., Sarges,P., Flemming,D., van Noort,V., Kunze,R., Devos,D.P., Arumugam,M., Bork,P. and Hurt,E. (2011) Insight into structure and assembly of the nuclear pore complex by utilizing the genome of a eukaryotic thermophile. *Cell*, **146**, 277–289.
4. Thierbach,K., von Appen,A., Thoms,M., Beck,M., Flemming,D. and Hurt,E. (2013) Protein interfaces of the conserved Nup84 complex from Chaetomium thermophilum shown by crosslinking mass spectrometry and electron microscopy. Structure, **21**, 1672–1682.
5. Monecke,T., Haselbach,D., Voß,B., Russek,A., Neumann,P., Thomson,E., Hurt,E., Zachariae,U., Stark,H., Grubmüller,H. *et al.* (2013) Structural basis for cooperativity of CRM1 export complex formation. *Proc. Natl Acad. Sci. U.S.A.*, **110**, 960–965.
6. Leidig,C., Bange,G., Kopp,J.U.R., Amlacher,S., Aravind,A., Wickles,S., Witte,G., Hurt,E., Beckmann,R. and Sinning,I. (2012) Structural characterization of a eukaryotic chaperone—the ribosome-associated complex. Nat. Struct. Mol. Biol., **20**, 23–28.
7. Hondele,M., Stuwe,T., Hassler,M., Halbach,F., Bowman,A., Zhang,E.T., Nijmeijer,B., Kotthoff,C., Rybin,V., Amlacher,S. *et al.* (2013) Structural basis of histone H2A-H2B recognition by the essential chaperone FACT. *Nature*, **499**, 111–114.
8. Lapinaite,A., Simon,B., Skjaerven,L., Rakwalska-Bange,M., Gabel,F. and Carlomagno,T. (2013) The structure of the box C/D enzyme reveals regulation of RNA methylation. *Nature*, **502**, 519–523.
9. Mathé,C., Sagot,M.-F., Schiex,T. and Rouzé,P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, **30**, 4103–4117.
10. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
11. Ori,A., Banterle,N., Iskar,M., Andrés-Pons,A., Escher,C., Khanh Bui,H., Sparks,L., Solis-Mezarino,V., Rinner,O., Bork,P. *et al.* (2013) Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. *Mol. Syst. Biol.*, **9**, 648.
12. Wang,Y., Yang,F., Gritsenko,M.A., Wang,Y., Clauss,T., Liu,T., Shen,Y., Monroe,M.E., Lopez-Ferrer,D., Reno,T. *et al.* (2011) Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics*, **11**, 2019–2026.
13. Wiśniewski,J.R., Zougman,A. and Mann,M. (2009) Combination of FASP and stagetip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *J. Proteome Res.*, **8**, 5674–5678.
14. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.
15. Powell,S., Szklarczyk,D., Trachana,K., Roth,A., Kuhn,M., Muller,J., Arnold,R., Rattei,T., Letunic,I., Doerks,T. *et al.* (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, **40**, D284–D289.
16. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
17. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
18. Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
19. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
20. Götz,S., García-Gómez,J.M., Terol,J., Williams,T.D., Nagaraj,S.H., Nueda,M.J., Robles,M., Talón,M., Dopazo,J. and Conesa,A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.*, **36**, 3420–3435.
21. de Godoy,L.M.F., Olsen,J.V., Cox,J., Nielsen,M.L., Hubner,N.C., Fröhlich,F., Walther,T.C. and Mann,M. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, **455**, 1251–1254.
22. Beck,M., Schmidt,A., Malmstroem,J., Claassen,M., Ori,A., Szymborska,A., Herzog,F., Rinner,O., Ellenberg,J. and Aebersold,R. (2011) The quantitative proteome of a human cell line. *Mol. Syst. Biol.*, **7**, 549.
23. Geiger,T., Wehner,A., Schaab,C., Cox,J. and Mann,M. (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics*, **11**, doi:10.1074/mcp.M111.014050.
24. Schwanhäusser,B., Busse,D., Li,N., Dittmar,G., Schuchhardt,J., Wolf,J., Chen,W. and Selbach,M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337–342.
25. Novikova,O., Fet,V. and Blinov,A. (2009) Non-LTR retrotransposons in fungi. *Funct. Integr. Genomics*, **9**, 27–42.
26. Schönknecht,G., Chen,W.-H., Ternes,C.M., Barbier,G.G., Shrestha,R.P., Stanke,M., Bräutigam,A., Baker,B.J., Banfield,J.F., Garavito,R.M. *et al.* (2013) Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, **339**, 1207–1210.
27. Berka,R.M., Grigoriev,I.V., Otillar,R., Salamov,A., Grimwood,J., Reid,I., Ishmael,N., John,T., Darmond,C., Moisan,M.-C. *et al.* (2011) Comparative genomic analysis of the thermophilic biomass-degrading fungi Myceliophthora thermophila and Thielavia terrestris. *Nat. Biotechnol.*, **29**, 922–927.
28. Medina,I., Salavert,F., Sanchez,R., de Maria,A., Alonso,R., Escobar,P., Bleda,M. and Dopazo,J. (2013) Genome Maps, a new generation genome browser. *Nucleic Acids Res.*, **41**, W41–W46.
29. Omasits,U., H Ahrens,C., Müller,S. and Wollscheid,B. (2013) Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinformatics , **30**, 884–886.
30. Ghaemmaghami,S., Huh,W.-K., Bower,K., Howson,R.W., Belle,A., Dephoure,N., O'Shea,E.K. and Weissman,J.S. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.
31. Jensen,L.J., Julien,P., Kuhn,M., von Mering,C., Muller,J., Doerks,T. and Bork,P. (2007) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.*, **36**, D250–D254.
32. van Noort,V., Bradatsch,B., Arumugam,M., Amlacher,S., Bange,G., Creevey,C., Falk,S., Mende,D.R., Sinning,I., Hurt,E. *et al.* (2013) Consistent mutational paths predict eukaryotic thermostability. *BMC Evol. Biol.*, **13**, 7.
33. Holder,T., Basquin,C., Ebert,J., Randel,N., Jollivet,D., Conti,E., Jékely,G. and Bono,F. (2013) Deep transcriptome-sequencing and proteome analysis of the hydrothermal vent annelid Alvinella pompejana identifies the CvP-bias as a robust measure of eukaryotic thermostability. *Biol. Direct*, **8**, 2.
34. Sprott,G.D. (1992) Structures of archaebacterial membrane lipids. *J. Bioenerg. Biomembr.*, **24**, 555–566.
35. Santos,H. and da Costa,M.S. (2002) Compatible solutes of organisms that live in hot saline environments. *Environ. Microbiol.*, **4**, 501–509.