Published in final edited form as:
Gene. 2010 November 1; 467(0): 41–51. doi:10.1016/j.gene.2010.07.009.

Detecting novel genes with sparse arrays

Mikko Arvas[a,1], **Niina Haiminen**[b,1], **Bart Smit**[c], **Jari Rautio**[d], **Marika Vitikainen**[a], **Marilyn Wiebe**[a], **Diego Martinez**[e], **Christine Chee**[e], **Joe Kunkel**[e], **Charles Sanchez**[e], **Mary Anne Nelson**[e], **Tiina Pakula**[a], **Markku Saloheimo**[a], **Merja Penttilä**[a], and **Teemu Kivioja**[f]

Mikko Arvas: mikko.arvas@vtt.fi; Niina Haiminen: niina.haiminen@cs.helsinki.fi; Bart Smit: bart.smit@campina.com; Jari Rautio: jari.rautio@plexpress.fi; Marika Vitikainen: marika.vitikainen@vtt.fi; Marilyn Wiebe: marilyn.wiebe@vtt.fi; Diego Martinez: admar@unm.edu; Christine Chee: cchee@unm.edu; Joe Kunkel: jkunkel@unm.edu; Charles Sanchez: csanche9@unm.edu; Mary Anne Nelson: manelson@unm.edu; Tiina Pakula: tiina.pakula@vtt.fi; Markku Saloheimo: markku.saloheimo@vtt.fi; Merja Penttilä: merja.penttila@vtt.fi; Teemu Kivioja: teemu.kivioja@helsinki.fi

[a]VTT Technical Research Centre of Finland, Tietotie 2, P.O. Box FI-1000, 02044 VTT, Espoo, Finland [b]HIIT, Department of Computer Science, P.O. Box 68, FI-00014 University of Helsinki, Finland [c]FrieslandCampina Innovation Europe, Nieuwe Kanaal 7C, 6709 PA Wageningen, The Netherlands [d]Plexpress, Helsinki, Viikinkaari 6, 00790 Helsinki, Finland [e]Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131, USA [f]Department of Computer Science, PO Box 63, FI-00014 University of Helsinki, Finland

Abstract

Species-specific genes play an important role in defining the phenotype of an organism. However, current gene prediction methods can only efficiently find genes that share features such as sequence similarity or general sequence characteristics with previously known genes. Novel sequencing methods and tiling arrays can be used to find genes without prior information and they have demonstrated that novel genes can still be found from extensively studied model organisms. Unfortunately, these methods are expensive and thus are not easily applicable, e.g., to finding genes that are expressed only in very specific conditions.

We demonstrate a method for finding novel genes with sparse arrays, applying it on the 33.9 Mb genome of the filamentous fungus *Trichoderma reesei*. Our computational method does not require normalisations between arrays and it takes into account the multiple-testing problem typical for analysis of microarray data. In contrast to tiling arrays, that use overlapping probes, only one 25mer microarray oligonucleotide probe was used for every 100 b. Thus, only relatively little space on a microarray slide was required to cover the intergenic regions of a genome. The analysis was done as a by-product of a conventional microarray experiment with no additional costs. We found at least 23 good candidates for novel transcripts that could code for proteins and all of which were expressed at high levels. Candidate genes were found to neighbour *ire1* and *cre1* and many other regulatory genes. Our simple, low-cost method can easily be applied to finding novel species-specific genes without prior knowledge of their sequence properties.

Correspondence to: Mikko Arvas, mikko.arvas@vtt.fi.

[1] The first two authors contributed equally to this work.

Supplementary materials related to this article can be found online at doi:10.1016/j.gene.2010.07.009.

Keywords

Microarray; Gene prediction; Moulds; Transcript

1. Introduction

Recent progress in sequencing technology is increasing rapidly the number of available genomes. Consequently, efficient discovery of basic functional elements, such as all protein or RNA encoding genes, from newly sequenced genomes is becoming a crucial bottleneck in the use of genomics to explore biological diversity. Automatic sequence-based prediction methods are very good at finding genes that have homology to known genes or have sequence features such as translation initiation and splice sites that are common to many genes of the organism (reviewed in (Brent, 2008)). Typically, a combination of approaches is used. Candidate genes are found by *de novo* gene prediction programs that are trained on the known genes of the organism. Alternatively, some programs can both learn typical characteristics of the genes and predict candidate genes in an iterative fashion given only the genome and an initial gene model (Ter-Hovhannisyan et al., 2008). In parallel, the genome is aligned to ESTs from the organism and related organisms and genomes of other organisms to find genes based on expression and/or conservation.

The pitfall of this otherwise successful strategy is that it is biased towards finding things that resemble what we already know. When sequencing new genomes, it is also important to find those genes that are truly different from what has been observed before. Recent study of prokaryotic genomes, with far simpler gene structure, suggests that current gene prediction methods may miss hundreds of conserved gene families (Warren et al., 2010). Thus, there is a need for complementary approaches that only utilize the genomic sequence and functional data from the organism in question to predict new genes. Dense tiling arrays (Selinger et al., 2000; Bertone et al., 2004; David et al., 2006) and direct transcript sequencing (Miura et al., 2006; Wilhelm et al., 2008; Nagalakshmi et al., 2008) are such methods, but their high cost limits expansion of their usage to the hundreds of less studied organism for which genome sequences are or will be available. In addition, finding an interesting novel transcript might require analysis of large amounts of samples, such as dense time series under different experimental conditions.

The fungal kingdom includes various industrially, medically and agriculturally important species and major model organisms such as *Saccharomyces cerevisiae*. Sequencing of fungal genomes allows us to tap into their diverse metabolism, such as lignocellulose or pectin degradation, and synthesis of antibiotics or other secondary metabolites. Many of the fungi with a large impact on society come from the phylum Ascomycota, such as the protein and citric acid producing *Aspergillus niger*, rice blast fungus *Magnaporthe grisea*, human pathogens *Aspergillus fumigatus* and *phCandida albicans* and baker's yeast *S. cerevisiae*. However, in particular genes related to the interesting metabolic functions mentioned earlier appear to be under fast evolution in Ascomycota (Arvas et al., 2007) and thus likely to give rise to lineage-specific or orphan genes. Lineage specificity has been proposed to arise in fungi, for example through duplication and divergence (Fedorova et al., 2008), accelerated evolutionary rates (Cai et al., 2006; Kawahara and Imanishi, 2007) and horizontal gene transfer (Khaldi et al., 2008; Lieckfeldt et al., 2000). Lineage-specific genes are also considered particularly important for virulence of human and plant pathogens.

In this work we studied the Ascomycota fungus *Trichoderma reesei (Hypocrea jecorina)*, an important model organism for lignocellulose degradation. *T. reesei* is used for commercial production of its native enzymes such as various cellulases and heterologous proteins. It can achieve protein yields above 100 g/l in industrial fermentations, a quantity not reported for any other organism (Cherry and Fidantsef, 2003). Degradation of lignocellulose from agricultural crop residues, grasses, wood and municipal solid waste by cellulases and other enzymes is a crucial step in transforming these biomasses to second generation biofuels. Hence, there is a dire need for understanding the protein-secretion process.

Typical oligonucleotide microarray slides can accommodate hundreds of thousands of probes. However, for example with 60mer probes, only 1 or 2 probes per transcript are routinely utilised. Fungi typically have from 5000 to 20,000 predicted genes, thus extra space is often available on a microarray to search for novel transcripts. To test this concept we covered the intergenic regions of the plus strand of the *T. reesei* genome with 187,641 25mer probes with approximately 100 b gap between two consecutive probes. In addition, our microarray contained 25mer probes also for the previously predicted genes, as in a conventional oligonucleotide microarray expression profiling experiment. In comparison, 6.5 million probes were used previously (David et al., 2006; Juneau et al., 2007) to study *S. cerevisiae*, whose genome size is roughly a third of the *T. reesei* genome size.

'Tiling array' refers to an array design where the probe positions overlap, we call our design a 'sparse array'. The low signal-to-noise ratio of the sparse microarray data makes it hard to distinguish true gene expression from the background, especially because the hybridization probes have different affinities to their targets. However, we demonstrate that it is still possible to assess the presence of a novel gene by comparing the expression levels of the group of probes within an open reading frame (ORF) to those of other ORFs. We did not want to predict new genes by comparing the expression levels to those of known genes, as that would require deciding which known genes are expressed in the experiment — a hard task in itself. Instead, we look for ORFs that contain many probes with high expression values. The significance of observing an ORF with a given number of highly expressed probes was determined by a comparison to the overall distribution of expression levels of probes in the (mostly) non-transcribed sequence. This was done *in silico* by permuting the locations of the probes. The randomization allows us to estimate the false-positive rate of our findings and to avoid problems due to multiple hypothesis testing, without making unrealistic assumptions about the data. A similar computational approach has been suggested earlier (Royce et al., 2005), but our study is the first one to consistently apply it to finding novel genes from sparse array data.

We show that it is possible to detect dozens of previously unknown transcripts from sparse array data that was collected without additional cost as a side-product of a conventional gene expression experiment. Furthermore, the novel transcripts show regulation and high expression in conditions relevant for protein production, making them key targets for further studies on fungal protein secretion.

2. Materials and methods

A work flow of the analysis is included in Supplementary file 6.

2.1. Data collection

The *Trichoderma reesei* strain Rut-C30 (Montenecourt and Eveleigh, 1979) was grown in chemostat cultivations as described in (Rautio et al., 2006). Strain Rut-C30 was used instead of the sequenced strain QM6a for its enhanced protein production capabilities. Cultivations were done in lactose-limited chemostats at three different conditions: specific constant growth rates of 0.03 h^{-1} (D03), 0.06 h^{-1} (D06) and 0.03 h^{-1} with high cell density (HD03). The high cell density was achieved by increasing the lactose concentration of the feed medium from 20 g/L to 80 g/L. Triplicate cultures were analysed for the three conditions.

Stable chemostat cultures were attained within two or three residence times, and three generations after steady states were attained, samples were withdrawn for microarray analysis. Mycelial samples were homogenised with FastPrep cell homogenizer (ThermoSavant, Dreirich, Germany) using 6 m/s for 45 s. RNA was extracted from the homogenate with Total RNA kit (A&A Biotechnologies, Gnydia, Poland). Quality and quantity of RNA was monitored by absorbance measurement at A260 (DNA Quant, Pharmacia, Uppsala, Sweden) and Agilent Bioanalyser and RNA 6000 Nano Assay kit (Agilent Technologies, Palo Alto, CA, USA).

Total RNA samples were submitted to microarray analysis by Roche Nimblegen (WI, USA). Probe design and synthesis, RNA labelling, hybridisation and signal quantification were carried out by Nimblegen. Design of the microarray and all subsequent analysis were carried out with the *T. reesei* genome (Martinez et al., 2008) version 1.2 (http://genome.jgi-psf.org/trire1/trire1.home.html). Plus strand of intergenic regions were covered with 187,641 25mer oligonucleotide probes with approximately 100 nt spacing.

2.2. Preprocessing and normalization

The *T. reesei* genome version 1.2 is composed of 1094 scaffolds. The GC % of the short scaffolds varies considerably (see Supplementary file 1). A cutoff for the lengths of the scaffolds to be included in the analysis was chosen so that the GC% for the chosen scaffolds is relatively constant. The cutoff was set to include 51 of the total 1094 scaffolds, covering 90.92% of the sequence data. The length of the shortest included scaffold was 131,123 bp.

The expression values of a microarray probe have been observed to be highly dependent on their GC percentage (Samanta et al., 2006; Royce et al., 2007). Therefore we first applied the GC-scaling scheme introduced in (Samanta et al., 2006) and discussed in (Royce et al., 2007) to the raw probe-level data. The expression values for all probes with a given GC% in a given experiment were divided by the median expression value of the intergenic probes in that experiment. This corrected the bias towards larger expression values of probes with a higher GC%.

Potential ORFs were found using the EMBOSS (Rice et al., 2000) program ‘getorf’ and those that overlap with a previously predicted gene in the genome version 1.2 were excluded.

2.3. Analysis of the novel genes

Similarity of probes to transcripts of the old genes was estimated by blastn (Altschul et al., 1990) with default settings (Supplementary file 2). To find homologues of the novel genes, their translations in six frames were used to search the EMBL protein and nucleotide sequence databases with tblastx, blastx (Altschul et al., 1990) and psi-blast (Altschul et al., 1997) with default settings. All available unpublished fungal genomes were collected from the Broad Institute, USA (http://www.broad.mit.edu/) and Joint Genome Institute, USA (http://genome.jgi-psf.org/) web sites and searched for matches.

In addition, after completion of experimental part of this study, *T. reesei* genome version 2.0 was released. Through additional sequencing the number of scaffolds was reduced from 1094 to 87. In addition gene modelling was improved based on extensive manual curation, reducing the number of genes from 9997 to 9129. In order to benefit from these improvements we also checked whether candidate genes overlapped with genes in v2.0.

Protein family databases were searched for matches by Inter-ProScan (Quevillon et al., 2005) and RNA families with Rfam (Griffiths-Jones et al., 2003) with default settings. Figs. 3 and 4 and Supplementary files 1–5 were produced by custom scripts with R (Development Core Team, 2008; Gentleman et al., 2004) and Fig. 1 with Bioperl (Stajich et al., 2002) and Gbrowse (Stein et al., 2002).

2.4. Reverse transcriptase-PCR

Specific primers were designed for the candidate genes with PerlPrimer version 1.1.16 (Marshall, 2004). Forward and reverse primers respectively for quantitative RT-PCR were: 8_1718: ACGTCTTGTT-CTTCTCTTCTC and AAAGGAGAGGTAAATGCACAG, 4_2310: CTCTCGCCTCACATAATCAC and GAATGAGAATTGGACCGCTG, 10_1091: TTCTGCTGCTTGATTGTTTCTC and GAACAGGTAGATTAAATGAGCGA, 14_172: GGGAAACGAAACAAAGAACAG and TATTATTTGGAGGT-GAGCGG, 10_1841: TTACTACTCTTCTT and ATTCGCAATGCTTGATACTC and for qualitative RT-PCR: 16_41: TGGCTTCTTGCTTTAAATGC and GAA-GAGTGGAGAAGAAAGGC, 19_1159: ATGACCAGCACCTTTGAATG and TACATGTACATACTCCAACCG, 19_455: ATGTTCTACGTCCATCACCC and TACATAGTGTGGAGAGGGAG, 21_266: ATGGGCTCCGATAGCAAA and CAAGAGATACTGATGGTGAAC, 28_34: ATGATATCCTCCGTCTTCTCC and TGTCATTGCAAACCGACG, 28_371: ATGCAGCTGCACTCGTAC and GTTCGCTCTGGCTGTTG, 5_2510: ATGCTATTTGCCATCTGCA and AGCAG-CAGGAGAGGAGAG, 7_1555: ATGGTCGGCATCACCTAT and CTGATGA-TATGATGCATGAGG, 8_1028: ATGACAGTTGAGCGACTGAC and AGGTACAGACAGAGCTGC, 9_273f: TGCATTGCATCAATACTACC and AGGGAATTGGGCTATCTAGA. Same total RNA samples as in the microarray analysis were used. Total RNA was treated with RNase-Free DNase Set (Qiagen N.V. Venlo, The Netherlands) prior to clean up with RNeasy Mini Kit

(Qiagen N.V. Venlo, The Netherlands). Purified total RNA samples were used as templates in the cDNA synthesis carried out with oligo(dT) primers and DyNAmo cDNA Synthesis Kit (Finnzymes, Espoo, Finland). Non-purified cDNA samples were then used as templates for RT-PCR.

Quantitative RT-PCR was carried out using LightCycler 480 SYBR Green I Master and Light Cycler Instrument II (Roche Diagnostics Ltd., Rotkreuz, Switzerland). Possible contamination of genomic DNA was monitored by carrying out cDNA synthesis reactions without the reverse transcriptase enzyme and then using the reactions as templates in RT-PCR with reference gene 60091 primers. Melting temperature analysis was performed after the amplification and the RT-PCR reactions were also analyzed in 1.2% agarose gel. Relative quantification of genes was calculated with LightCycler 480 SW1.5 basic relative quantification analysis with PCR efficiency correction. Endogenous reference genes used were 60091 and 106250 (v2.0 identifiers). They were selected for their strong and stable signal in a large non-public dataset. Their primers are 106250: CGGACGACAGA-CAATACCAG and GCTGGTGGATGGTCAAGATT and 60091: GCGACCTCGTCCTCTACAAG and GGTTATCGCCAACAATCCAG.

Three independent RT-PCRs were carried out for each of the five Top candidates in the four samples. Each RT-PCR contained reactions for both control genes, i.e. in total 120 RT-PCR reactions were carried out.

cDNA synthesis for qualitative RT-PCR was carried out as for quantitative RT-PCR except that RNA samples were pooled before synthesis. Qualitative RT-PCR was carried out using Dynazyme Ext PCR Kit (Finnzymes, Espoo, Finland). Possible contamination of genomic DNA was monitored by carrying out cDNA synthesis reactions without the reverse transcriptase enzyme and using these reactions also as templates for each individual primer pair. Genomic DNA was also used as template for each primer pair and all three reactions corresponding to single primer pair analyzed on 1.5% agarose gel.

2.5. Sequencing

Quantitative RT-PCR products were analyzed in 1.2% agarose gel and isolated with Qiaquick Gel Extraction Kit (Qiagen N.V. Venlo, The Netherlands) and sequenced in both forward and reverse direction using the same specific oligos as in the RT-PCR amplification. Sequencing was performed using Big Dye Terminator v3.1 Cycle Sequencing Kit (AB Applied Biosystems, Life Technologies Corporation, CA, USA) and analyzed with 3100 Genetic Analyer ((AB Applied Biosystems, Life Technologies Corporation, CA, USA).

3. Results

3.1. Obtaining sparse array data

In order to detect novel genes with sparse arrays from protein production conditions, the *T. reesei* strain Rut-C30 (Montenecourt and Eveleigh, 1979) was cultivated in lactose-limited chemostats. A chemostat is a bioreactor cultivation where some substrate component such as the main carbon source, e.g. lactose, limits biomass production and is fed at a constant rate which determines the specific growth rate of the organism. The highest specific rate of total secreted protein production is achieved at the specific growth rates close to 0.03 h^{-1} in *T. reesei* chemostat cultivations. At the relatively high specific growth rates of 0.05–0.07 h^{-1} more protein is synthesized in total, but less secreted protein is produced (Pakula et al., 2005).

The goal of our chemostat cultivations was to study the intracellular responses using genomic methods to growth rates 0.03 h^{-1} (D03), 0.06 h^{-1} (D06) and 0.03 h^{-1} with high cell density (HD03). The specific protein production rates, averaged for three replicates were found to be 6.5, 4.5, and 1.4 mg g^{-1} h^{-1} in D03, D06, and HD03 cultivations, respectively, confirming (Pakula et al., 2005) result.

Further results of this experiment will be published in detail elsewhere, along with the conventional transcriptional profiling data from genes previously predicted in *T. reesei*. In this article we describe a method for the analysis of sparse array data, using data from these chemostat cultivations. The data discussed in this publication have been deposited in NCBI’s Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series accession number GSE12960 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12960).

3.2. Identification of candidate genes

To remove the dependency between signal strength and the probe’s GC%, the signals were normalised with a GC% scaling as in (Samanta et al., 2006) (Fig. 1). All potential ORFs of length >150 b and flanked by a start and stop codon were extracted from intergenic regions of the *T. reesei* genome. There are three experimental conditions—D03, D06 and HD03—in the dataset, and three independent measurements for each condition. In the following analysis we averaged the results over the independent measurements, obtaining a single expression value for each probe under each experimental condition.

Consider an ORF spanning *N* probes, *h* of which have an expression level higher than the *k*th percentile of all intergenic probes. We computed the probability of observing such an ORF, denoted by a *p*-score (Eq. (1)). We limit our analysis to ORFs that have $N \geq 3$. The probability *p* (*p*-score) of observing at least *h* probes with expression values above the *k*th percentile (e.g., 75th) of all intergenic expression values for that experimental condition is computed according to Eq. (3) in (Royce et al., 2005)[2]:

$$p=\sum_{i=h}^{N} k^{N-i}(1-k)^{i}\begin{pmatrix} N \\ i \end{pmatrix}. \quad (1)$$

The choice of the percentile *k* is addressed in the next subsection. These *p*-scores are based on the assumption that all probes are statistically independent, which is unlikely to hold in our data. Therefore we adopt the scheme suggested in (Royce et al., 2005) and compare the *p*-scores in the original data to those in randomized versions of the data.

[2] Notice the error in the notation in Eq. (3) in (Royce et al., 2005): the powers should be *N–i* and *i* instead of *N–h* and *h*.

Randomization is performed by shuffling the locations of all intergenic probes, i.e., a random permutation of the probe locations is obtained, and the original probe values are assigned to those locations. The *p*-scores for each ORF are now computed in the same way as for the original data. The results can be used to assess the number of false positives that are included in the results for a given p-score cutoff.

3.3. Choice of the percentile and p-score cutoff

The chosen percentile *k* and the p-score cutoff affects the number of genes that are found to have significantly high expression, and the number of false positives among them. The ORFs in original and randomized data were listed in the order of increasing *p*-scores. The percentage of ORFs in the list up to some p-score cutoff c that results from randomized data, *false positives*, is the false discovery rate (FDR) with that p-score cutoff c. We approached the percentile selection problem by fixing the false discovery rate (FDR) to 20%.

We experimented with percentiles *k*={50,55,60,65,70,75,80,85} and in order to select the percentile that yielded the largest number of true positives when the p-score cutoff is such that including *p*-scores above the cutoff would increase the false discovery rate to above 20%. Here *true positives* are ORFs with *p*-scores below the cutoff in the original data, and false positives are ORFs with *p*-scores below the cutoff in the randomized data. For each alternative percentile, the randomization was repeated 100 times, and the average number of true positives was computed.

We found that the 75th percentile yields lists that contain the highest numbers of ORFs from the original data. The corresponding numbers of potential candidate ORFs were 75 for D03, 82 for D06, and 119 for HD03.

Fig. 2 shows the distribution of *p*-scores for the experimental condition D03 in the original data, and the average distribution in 100 randomizations. In this case the percentile *k*=75 was used in Eq. (1). The results for the other conditions D06 and HD03 are similar (data not shown).

We then obtained the intersection of the lists of the three conditions, resulting in 57 ORFs of interest. In cases where the ORF locations overlapped, those with the lowest *p*-scores were retained, resulting in 47 ORFs (65 for D03, 72 for D06, and 106 for HD03, see Supplementary file 8). Hereafter they are referred to as candidate genes, in contrast to the genes originally predicted for the genome, referred to as old genes. Detailed information on the candidate genes is shown in Table 1 and an example of a candidate gene is shown in Fig. 1.

3.4. Analysis of the candidate genes

To verify that signals detected from probes of candidate genes could not be explained by hybridisation to transcripts of old genes, i.e. cross hybridisation, we estimated the similarity of all probes to the old genes' transcripts by blastn (Altschul et al., 1990). The highest blastn bit scores were compared for each probe of candidate genes, all intergenic probes, and probes of old genes against transcripts other than for which the probe was designed. We found no significant differences between the bit score distributions (Supplementary file 2).

Repeat sequences of the *T. reesei* genome have been analysed in detail. Class I and II fungal transposons were found, none apparently active (Martinez et al., 2008). Regardless, we verified that none of the candidate genes overlapped with repeat sequences.

We then considered whether the candidate genes could be explained as 5′ or 3′ UnTranslated Regions (UTR) of neighboring old genes. In that case the UTR would at least span the candidate gene and the gap between the start or stop codon (whichever was nearer) of the closest old gene on the same strand. In order to compare this length to those of verified 5′ or 3′ UTRs, we gathered information about the expected length of UTRs in fungal genomes, many of which were from related *Trichoderma* spp, from the UTRdb (Mignone et al., 2005). With 6726 entries for 5′ UTRs and 8218 for 3′ UTRs and a good correspondence to size estimates presented for *S. cerevisiae* (David et al., 2006) and *Schizosaccharomyces pombe* (Wilhelm et al., 2008), two extremely distant Ascomycota fungi, the database set probably represents fungal UTRs well. The mean length of 5′ and 3′ UTRs was found to be 155 and 184 b, respectively, and less than 1% of UTRs were found to be longer than 1000 b, hence we assumed that candidate genes whose lengths as UTRs would be longer than 1000 b are unlikely to be actual UTRs.

There were 23 such candidates, hereafter referred to as Top candidates. 4 candidates (2_2919, 6_1818, 43_73, 45_18), did not fill this criteria and were not included in Top candidates. However, several separate ESTs for both these candidate genes and the nearest neighbouring old genes supported the hypothesis that these candidates are not UTRs.

tblastx, blastx (Altschul et al., 1990), psi-blast (Altschul et al., 1997), InterProScan (Quevillon et al., 2005) and Rfam (Griffiths-Jones et al., 2003) searches were carried out to find homologues of the candidate genes in DNA, RNA or protein sequences (Table 1). In six cases (Table 1: column '*T. reesei* v2.0 protein id.') the candidate gene derived from analysis of the genome version 1.2 was found to overlap with a gene in the genome version 2.0. In three of these cases the gene sequence in 1.2 contained gaps that had been removed in 2.0 by additional sequencing (Table 1). In the others the model has been changed to include the sequence covered by the candidate gene. As these genes are not novel they have been excluded from further analysis unless explicitly stated. For 33_421, a Top candidate, a predicted unknown homologous protein was found from *Fusarium verticillioides* and *Fusarium graminearum*. EST evidence exists also for the *Fusarium verticillioides* gene. In all other cases, based on detailed manual inspection of the search results, we concluded that database matches were nonsignificant, thus similarity searches yielded no insight into function of the candidate genes.

To study whether other gene prediction methods than the ones used for *T. reesei* genome (Martinez et al., 2008) could find the candidate genes, we predicted a completely new set of gene models with Augustus (Stanke et al., 2008). To train Augustus on the *T. reesei* 2.0 genome, we downloaded all *T. reesei* full length mRNAs (113) and ESTs (44,964) from GenBank release 174 and used this as input into the Augustus pipeline.

This effort resulted in 7197 gene models that cover 9,779,775 b over the entire T. reseei 2.0 genome, while 9129 genes that cover 16,175,832 b were originally predicted (Martinez et al., 2008). Two candidate genes were found to overlap with Augustus genes on the opposite strand (4_2591, 4_2591) and three on same strand (1_3454, 14_1144, 25_407, Table 1: column ‘Augustus overlap’). However, 1_3454 and 14_1144 already overlap with genes in genome version 2.0.

In order to find independent experimental evidence for the candidate genes, we analysed *T. reesei* ESTs from (Chambergo et al., 2002; Foreman et al., 2003; Arvas et al., 2006; Martinez et al., 2008). Overlapping ESTs were found for 33_421, mentioned above (Fig. 1), and for 3 cases which had been predicted to be genes in version 2.0, and for 14 other candidate genes, including 3 Top candidates, for which no database matches were found.

In order to further verify the existence and expression signal of candidate genes we carried out an quantitative and qualitative RT-PCR (Reverse Transcriptase-PCR) experiments. Of the 23 top candidates, for 15 primers could be designed.

Five Top candidates that covered the expression signal range were selected for quantitative RT-PCR. Their relative expression was measured against two control genes in four samples originally used for array experiments. Possible contamination of genomic DNA was monitored by carrying out cDNA synthesis reactions without the reverse transcriptase enzyme and then using these reactions as templates in RT-PCR. Control reactions showed that in one sample a minor amount of genomic DNA was left. In this control reaction there was amplification in very late cycles but since the amplification of the candidate and reference genes from cDNA started in much earlier cycles, the amount of genomic DNA in the sample did not influence the results. Expression of all the five candidate genes was detected with RT-PCR. Regardless of repeated efforts we were able to successfully sequence only two of the RT-PCR products and they were found to correctly match expected genomic sequence (EMBL accessions: 4_2310 FN651826, 10_1841 FN651825). Relative RT-PCR expression values for each sample averaged over technical repeats and control genes was compared to array expression signal (Supplementary file 3). Overall Pearson correlation between array and RT-PCR signal was 0.48. However, if 10_1841 data is removed the correlation is 0.94. 10_1841 has the lowest array expression signal.

The rest 10 Top candidates for which primers could be designed were analysed with qualitative RT-PCR. For all, but 28_34, a higher signal was detected from cDNA sample than from negative control (cDNA synthesis without reverse transcriptase enzyme) confirming their expression. Some genomic contamination is present in most of the reactions, in particular 28_371 is somewhat borderline. For 8_1028 and 19_435 a fragment of different size was derived from cDNA and genomic sample (Supplementary file 7).

To compare signal intensity and variation, i.e. gene regulation, of transcriptional profiling signals of candidate and old genes, we processed their probe-level data together into gene level normalised signals using Robust Multichip Average (RMA) (Irizarry et al., 2003). The mean and standard deviation of the three experimental conditions for each candidate gene can be found in Table 1. Furthermore, we compared the signals of candidate genes to the signals of old genes. The mean (11.7) and standard deviation (0.3) of all candidate genes’ signals are higher than those of old genes (8.4, 0.18) and only 16% of old genes have a signal above 10.0 (Fig. 4).

We also compared the regulation of candidate genes to its nearest neighbour on the same strand. For each candidate and old gene, we calculated the Pearson correlation of non-averaged signals and difference of averaged signals of the 9 samples to the signals of the gene's nearest neighbour on the same strand (Supplementary file 4, Table 1). The average difference between average signals of candidates genes and each nearest neighbour is 2.5 with a p-value of 7.7*e*–12 in paired t-test.

Candidate genes' mean signal intensity, standard deviation of signal, correlation or difference of signal to nearest neighbour were not found to be dependent on the EST or homology evidence for the candidate gene (Fig. 4, Supplementary file 4, Table 1).

Based on sequence database searches, the candidate genes appeared to be orphan, i.e. lineage-specific genes, only found in *T. reesei* or its close relatives. In fungi, orphan genes sometimes appear in genomic islands (Machida et al., 2005; Kamper et al., 2006) that can be found particularly near telomers (Rehmeyer et al., 2006; Fedorova et al., 2008). Consequently, we analysed the old genes surrounding candidate genes for their lineage specificity. The genome was divided into non-overlapping windows of six genes, genes in each window were mapped to protein clusters from (Arvas et al., 2007), and the median number of proteins in those clusters was calculated (Supplementary file 5). The data set in (Arvas et al., 2007) contains 33 fungal genomes, 27 of which belong to Ascomycota. In (Arvas et al., 2007) protein sequences from these genomes were clustered to form protein clusters that mostly contain one gene from each of a set of closely related species, i.e. they are groups of orthologous genes sometimes complemented by paralogues. Given the species distribution of the protein clusters, windows with median cluster size of 1–4 represent genomic islands roughly specific to subclass Hypocreomycetidae. In (Arvas et al., 2007) this subclass is represented by *T. reesei*, Nectria haematococca and Fusarium graminearum. Correspondingly, windows with median cluster size of 5–12 represent regions roughly specific to the class Sordariomycetes. Increasing numbers of paralogous genes confound the interpretation of clusters with more genes, thus no evident taxon corresponds to windows of higher median cluster sizes. Furthermore, clusters above that size often contain several different groups of orthologous genes. Although 5 candidate genes reside in windows of median size below 6, i.e. Hypocreomycetidae specific genomic islands (Table 1), candidate genes are not particularly enriched in these islands nor are they found particularly near the scaffold ends.

T. reesei's carbohydrate-active enzymes (CAZymes) have been shown to be located in loose chromosomal clusters (Martinez et al., 2008). As these are the main protein product genes studied in the cultivations we checked whether any candidate genes resided in CAZyme clusters. 11_1798 was found in CAZyme cluster 4 (scaffold-3:8500–238,500) with cellulases egl1 and cbh2 and 627 b downstream of the start of the unfolded protein response (UPR) regulatory factor ire1 transcript (Valkonen et al., 2004). In addition, 25_408 was found in CAZyme cluster 19 (scaffold-22:310,000–540,000). Other candidate genes were not found in CAZyme clusters.

In order to study whether other candidate genes than 11_1798 would have interesting neighbouring genes we collected 3 different gene sets for both the 47 candidate genes and the 23 Top candidates among them: two, six and ten nearest neighbouring genes. For each of these six different gene sets an enrichment analysis was carried out as in (Arvas et al., 2007). Briefly, for each *T. reesei* gene conserved Protfun function predictions (Jensen et al., 2003), Funcat functional categories (Ruepp et al., 2004) and InterPro identifiers (Mulder et al., 2005) were determined and the gene sets analysed for enrichment of any of the annotations using the hypergeometric distribution (Table 2). Only InterPro identifiers i.e. protein domains were found to be significantly enriched (p-value<0.05). Of the eight enriched domains, four were found to be related to signalling. This does not include the ire1, nor the carbon catabolite repressor *cre1* (Strauss et al., 1995). 2_1351 was found 1139 b downstream from the end of the *cre1*transcript with no intervening genes. In addition six candidate genes were found to be adjacent to a putative transcription factor of the ‘Fungal transcriptional regulatory protein’ IPR001138 family. However, as the IPR001138 is very abundant in Pezizomycota (Arvas et al., 2007) it is not significantly enriched in these gene sets.

4. Discussion

We selected ORFs defined by a start and stop codon and >150 b as the basic gene model among which to look for candidate genes. The *T. reesei* genome version 2.0 is 33.9 Mb long with 9129 genes. 40.4% of it is coding and the average size of protein coding exon is 486 b, while the gene length is 1793 b with 3.1 introns of 120 b, on average. With sparse spacing of array data and high coding percentage of the genome, we did not expect to find large genes with introns. Furthermore, genes missed by current gene prediction methods are likely to be small (Warren et al., 2010) i.e. intronless in *T. reesei*. Current *de novo* gene predictors use complex models based on for example characteristics of splice donor and acceptor sites and translation initiation and termination sites (reviewed in (Brent, 2008)). These characteristics are ultimately derived from known genes. As we wanted to find truly novel genes that current gene prediction methods would be unable to find we wanted our model to have minimal assumptions. The average length of the found candidate genes is 463 b. If the statistics of version 2.0 genome apply for the candidate genes, then we might miss some genes that have introns. However, a model accounting for introns would make more assumptions.

In order to apply our method to other species the characteristics of the genome would have to be considered to find a suitable minimal model. For example predictions of individual exons or a windowing schema without any use of sequence characteristics could be used. In our opinion our data was of too low quality for a segmentation approach.

We searched for candidate genes from sparse microarray data by calculating how likely it is to observe a signal from a candidate gene that is significantly higher than the background level. This was done by applying a GC% scaling and computing the probability p of observing at least h probes out of all N probes of a candidate gene that have a signal higher than 75% of all intergenic probes. The false discovery rate (FDR) of candidates was then defined by randomisation and an FDR of 20% set as a cut-off (Fig. 1) for the probabilities p.

This strategy is particularly appealing because it is not sensitive to outliers, it requires no arbitrary cut-offs for original expression values, there is no need to normalise across experiments, and it takes into account multiple statistical testing. It has been shown that the common normalisation strategy that forces signal distributions to be similar across samples hides true variation between samples (van de Peppel et al., 2003).

Similarity between the probes in the candidate genes and originally predicted transcripts was then estimated to confirm that cross-hybridisation could not explain the results more than in conventional transcription profiling experiments (Supplementary file 2).

The candidate transcribed regions we detected could be actual novel genes or UTRs of old genes. We used data from known UTRs of *T. reesei* and other fungal genes to estimate whether the candidate genes, on average, based on the length of the candidate gene and its distance to the nearest neighbour on the same strand could be explained as UTRs (Fig. 3). We found that 23 candidate genes, Top candidates, are unlikely to be UTRs.

The candidate genes were subsequently evaluated in the context of EST data, repeat sequences, microarray transcriptional profiling data from old genes, the *T. reesei* genome, and homology to other organisms. ESTs gave independent experimental evidence to support the hypothesis that the candidate genes are actually transcribed and comparison to microarray data from the same experiments showed that the expression of candidate genes is on average higher than that of old genes (Fig. 4). Expression and high signal was further verified for five candidate genes with quantitative RT-PCR (Supplementary file 3) and expression for at least eight candidate genes with qualitative RT-PCR (Supplementary file 7). As expected qualitative RT-PCR results are harder to interpret and give less evidence for actual expression or lack of it than quantitative RT-PCR experiments. In addition a gene prediction software not previously used for the *T. reesei* genome, Augustus (Stanke et al., 2008) was used to predict a completely new gene set for the whole genome. Not counting candidate genes that could be explained by improved quality of genome version 2.0, Augustus was able to find one of the candidate genes. Given the variety of methods integrated into the JGI (Joint Genome Institute) gene prediction pipeline and large number of ESTs available for *T. reesei*, it would be surprising if other contemporary gene prediction methods could fare better. Thus, current gene prediction methods cannot generally predict such genes as our candidate genes.

Microarrays allow efficient comparison of changes in the transcript amount of one gene between several samples. However, it has not been very clear whether differences in microarray signals between genes are correlated with actual differences between transcript amounts. Recent comparison between transcript profiling by tiling arrays and transcript sequencing shows that a good correlation between counts of sequenced ESTs and tiling array signal exists over all genes of an organism (Wilhelm et al., 2008). Furthermore our array and quantitative RT-PCR experiments correlated well. Thus, it is likely that the particularly strong signal detected for the candidate genes identified here reflects particularly high amounts of transcripts. High transcript amounts suggest that the transcripts are also physiologically relevant.

We also compared the regulation of candidate genes to that of their nearest neighbour on the same strand (Supplementary file 4). We found that the candidate genes' signals are on average significantly higher than those of their neighbours'. In addition, signals of most candidate genes do not show positive correlation with those of their neighbours'. Transcriptional regulation of neighbouring genes could correlate positively for a number of reasons, for example, they might share promoter or enhancer elements or be subjected to common regulation of the nucleosome structure. Common regulation of the candidate genes with their neighbours does not mean that the candidate genes would be UTRs of the neighbouring gene. In contrast, finding genes with such exceptionally long UTRs and such internal signal variation would be highly unlikely. EST evidence also supports the conclusion that four candidates, not included as Top candidates are independent transcribed regions instead of being UTRs.

Sparse array data appears to contain much noise and, in addition, most of the eukaryotic genome is randomly transcribed to some extent (Wilhelm et al., 2008). In order to find true novel genes of interest, regardless of these effects we designed a robust analysis. In practice the analysis found genes transcribed at particularly high level. In this study short 25mer probes were used, while longer probes have been shown to perform better (Hughes et al., 2001; Chou et al., 2004). Use of longer probes could improve transcript detection even with the same relatively sparse probe spacing.

Homologous genes for the candidate genes were found from the 2.0 version of the *T. reesei* genome and closely related species, even though they were not predicted as genes in genome version 1.2 used in the analysis. This further confirms that our method can detect protein coding transcripts. Gaps in the sequence had prevented gene detection in three cases in version 1.2, while our method detected the transcript. This demonstrates that our method would be useful in the annotation of even low quality genome sequences.

We could detect homology to a sequence (protein or DNA) or family (protein or RNA) from other organisms for only one candidate gene. As fungal lineage-specific genes sometimes appear in genomic islands, we checked whether our candidate genes were found in these islands particularly often, or were found particularly near scaffold ends. Their locations did not show such biases and thus they are not likely to be found in subtelomeric regions (Supplementary file 5). Subtelomeric regions appear to have potential for faster evolution than other chromosomal regions in fungi (Naumov et al., 1996; Rehmeyer et al., 2006; Fedorova et al., 2008). In our analysis of the *T. reesei* genome, the most prominent lineage-specific region is found at the end of scaffold 50 (Supplementary file 5). As subtelomeric positioning does not offer clues to the mechanism behind the rise of these lineage-specific candidate genes; other evolutionary forces must be driving their evolution.

As similarity searches and analysis of genome structure yielded no insight into the function of the candidate genes, we turned to their neighbouring genes (Table 2). These analyses suggest that candidate genes occur often near to genes related to regulation and signalling. Strikingly, two important regulatory factors related intimately to protein production, ire1 and cre1 were found adjacent to candidate genes. As full cDNA for these genes are known, unlike for many other genes adjacent to candidate genes, we can be sure that the candidate genes are not actually parts of ire1 or cre1 UTRs. These results invoke the possibility that the candidate genes would belong to a novel class of regulatory factors involved also in protein production. They would have been previously missed due to their small size.

The novel transcribed regions found in this work are each composed of an ORF. The polyadenylated tail of an RNA is used in the transcript labelling step of the transcriptome profiling technique used. Thus, the detected transcripts have the potential to code for proteins. As the majority of them have not been detected as protein coding genes in any genomes analysed, their sequence characteristics are not typical of other protein coding genes. However, the existence of longer, in contrast to short interfering RNAs, polyadenylated noncoding RNAs (ncRNA) has been shown in many metazoan species ((Zhao et al., 2008), reviewed in (Gingeras, 2007)) and in fungi (Miura et al., 2006; Samanta et al., 2006; Wilhelm et al., 2008; David et al., 2006; Nagalakshmi et al., 2008). The function of these is generally not known, but diverse examples of ncRNA roles related to chromatin architecture and epigenetic control exist (reviewed in (Amaral et al., 2008)). Particularly, polyadenylated mRNA-like (mlRNA) transcripts that are derived from independent exons, in contrast to anti-sense transcripts, are sometimes processed to microRNAs (Rodriguez et al., 2004). In addition, many genes might code for both, a functional RNA and a protein (reviewed in (Dinger et al., 2008))

It is possible that some candidate genes are in fact noncoding RNAs. Many ncRNAs are conserved over different mammals (Ponjavic et al., 2007). Nevertheless the evolution of their DNA sequences is likely to be faster and under different constraints than that of protein coding genes, which consequently makes homology detection harder. This could partly explain why we cannot find homologues for such highly expressed genes. The work presented here was geared towards finding novel transcripts of protein coding genes, however the computational principles could easily be adapted to finding other types of transcripts.

5. Conclusions

The ability to find all transcripts of an organism is essential to understanding its biology, particularly its species-specific attributes, such as a notably high protein or metabolite production capability or severe pathogenicity. We demonstrated a simple, low-cost alternative for finding novel genes that are left undetected by gene prediction software from sequenced genomes. Our method is complementary to existing gene prediction methods, geared towards finding exceptional genes from low quality data, thus we do not expect to find many novel genes. We apply sparse arrays in this task, discovering novel candidate genes with a fixed false discovery rate. We validate our findings with a quantitative and qualitative RT-PCR experiments and a comparison to *T. reesei* EST and other sequence data. We show that the discovered novel candidate genes are expressed at high level and thus likely to be of importance for the phenotype of *T. reesei*. In addition, the genes neighboring the candidate genes are enriched in genes related to regulation and signalling, proposing a regulatory role for them. It has been shown that more than 90% of a fungal genome can be transcribed (Wilhelm et al., 2008). Much of this expression is likely to be noise. In contrast, sparse array data analysed with our method reveals highly expressed, condition- and species-specific genes, that are prime targets for explaining species-specific attributes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Heikki Mannila for helpful comments on the manuscript and Aili Grundstrm for excellent technical assistance. This work was financially supported by the Academy of Finland grant 118653 (ALGODAN) and by the Academy of Finland postdoctoral researcher's fellowship 127715 (as part of and supported by the Finnish Centre of Excellence in White Biotechnology - Green Chemistry, Project No. 118573). Diego Martinez and Charles Sanchez were supported by the National Institutes of Health Grant GM060201. Christine Chee was supported by Award Number T34GM00851 from the National Institute of General Medical Sciences.

Abbreviations

RT	reverse transcriptase
RMA	Robust Multichip Average
UPR	Unfolded Protein Response
CAZyme	carbohydrate active enzyme
FDR	False Discovery Rate

References

Altschul S, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410. [PubMed: 2231712]

Amaral PP, Dinger ME, RMT, Mattick JS. The eukaryotic genome as an RNA machine. Science. 2008; 319:1787–1789. [PubMed: 18369136]

Arvas M, et al. Comparison of protein coding gene contents of the fungal phyla Pezizomycotina and Saccharomycotina. BMC Genomics. 2007:8. [PubMed: 17210083]

Arvas M, et al. Common features and interesting differences in transcriptional responses to secretion stress in the fungi *Trichoderma reesei* and *Saccharomyces cerevisiae*. BMC Genomics. 2006; 7:32. [PubMed: 16504068]

Bertone P, et al. Global identification of human transcribed sequences with genome tiling arrays. Science. 2004; 306:2242–2246. [PubMed: 15539566]

Brent M. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet. 2008; 9:62–74. [PubMed: 18087260]

Cai J, Woo P, Lau S, DKSKY. Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. J Mol Evol. 2006; 63:1–11. [PubMed: 16755356]

Chambergo F, et al. Elucidation of the metabolic fate of glucose in the filamentous fungus *Trichoderma reesei* using expressed sequence tag (EST) analysis and cDNA microarrays. J Biol Chem. 2002; 277:13983–13988. [PubMed: 11825887]

Cherry J, Fidantsef A. Directed evolution of industrial enzymes: an update. Curr Opin Biotechnol. 2003; 14:438–443. [PubMed: 12943855]

Chou C, Chen C, Lee T, Peck K. Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. Nucleic Acids Res. 2004; 32:e99. [PubMed: 15243142]

David L, et al. A high-resolution map of transcription in the yeast genome. Proc Nat Acad Sci. 2006; 103:5320–5325. [PubMed: 16569694]

Dinger M, Pang K, Mercer T, Mattick J. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol. 2008:4.

Edgar R, Domrachev M, Lash A. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30:207. [PubMed: 11752295]

Fedorova ND, et al. Genomic islands in the pathogenic filamentous fungus aspergillus fumigatus. PLoS Genet. 2008; 4:e1000046. [PubMed: 18404212]

Foreman P, et al. Transcriptional regulation of biomass-degrading enzymes in the filamentous fungus *Trichoderma reesei*. J Biol Chem. 2003; 278:31988–31997. [PubMed: 12788920]

Gentleman RC, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004; 5:R80. [PubMed: 15461798]

Gingeras TR. Origin of phenotypes: genes and transcripts. Genome Res. 2007; 17:682–690. http://www.genome.org/cgi/reprint/17/6/682.pdf. [PubMed: 17567989]

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy S. Rfam: an RNA family database. Nucleic Acids Res. 2003; 31:439. [PubMed: 12520045]

Hughes T, et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol. 2001; 19:342–347. [PubMed: 11283592]

Irizarry R, Bolstad B, Collin F, Cope L, Hobbs B, Speed T. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003; 31:e15. [PubMed: 12582260]

Jensen L, Ussery D, Brunak S. Functionality of system components: conservation of protein function in protein feature space. Genome Res. 2003; 13:2444. [PubMed: 14559779]

Juneau K, Palm C, Miranda M, Davis R. High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. Proc Nat Acad Sci. 2007; 104:1522. [PubMed: 17244705]

Kamper J, et al. Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis. Nature. 2006; 444:97–101. [PubMed: 17080091]

Kawahara Y, Imanishi T. A genome-wide survey of changes in protein evolutionary rates across four closely related species of Saccharomyces sensu stricto group. BMC Evol Biol. 2007; 7:13. [PubMed: 17284311]

Khaldi N, Collemare J, Lebrun M, Wolfe K. Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. Genome Biol. 2008; 9:R18. [PubMed: 18218086]

Lieckfeldt E, Kullnig C, Samuels G, Kubicek C. Sexually competent, sucrose-and nitrate-assimilating strains of Hypocrea jecorina (*Trichoderma reesei*) from South American soils. Mycologia. 2000; 92:374–380.

Machida M, et al. Genome sequencing and analysis of Aspergillus oryzae. Nature. 2005; 438:1157–1161. [PubMed: 16372010]

Marshall O. PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. Bioinformatics. 2004; 20:2471. [PubMed: 15073005]

Martinez D, et al. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. Hypocrea jecorina). Nat Biotechnol. 2008; 26:553–560. [PubMed: 18454138]

Mignone F, et al. UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res. 2005; 33:D141. [PubMed: 15608165]

Miura F, et al. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. Proc Nat Acad Sci. 2006; 103:17846. [PubMed: 17101987]

Montenecourt BS, Eveleigh DE. Selective screening methods for the isolation of high yielding cellulase mutants of *Trichoderma reesei*. Adv Chem Ser. 1979; 181:289–301.

Mulder N, et al. InterPro, progress and status in 2005. Nucleic Acids Res. 2005; 33:D201. [PubMed: 15608177]

Nagalakshmi U, et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science. 2008; 320:1344–1349. [PubMed: 18451266]

Naumov G, Naumova E, Sancho E, Korhbla M. Polymeric SUC genes in natural populations of Saccharomyces cerevisiae. FEMS Microbiol Lett. 1996; 135:31–35. [PubMed: 8598274]

Pakula T, Salonen K, Uusitalo J, Penttilä M. The effect of specific growth rate on protein synthesis and secretion in the filamentous fungus *Trichoderma reesei*. Microbiology. 2005; 151:135–143. [PubMed: 15632433]

van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege F. Monitoring global messenger RNA changes in externally controlled microarray experiments. EMBO Rep. 2003; 4:387. [PubMed: 12671682]

Ponjavic J, Ponting C, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. 2007; 17:556. [PubMed: 17387145]

Quevillon E, et al. Interproscan: protein domains identifier. Nucleic Acids Res. 2005; 33:W116–20. [PubMed: 15980438]

Development Core Team, R. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2008.

Rautio J, Smit B, Wiebe M, Penttilä M, Saloheimo M. Transcriptional monitoring of steady state and effects of anaerobic phases in chemostat cultures of the filamentous fungus *Trichoderma reesei*. BMC Genomics. 2006; 7:247. [PubMed: 17010217]

Rehmeyer C, et al. Organization of chromosome ends in the rice blast fungus, Magnaporthe oryzae. Nucleic Acids Res. 2006; 34:4685–4701. [PubMed: 16963777]

Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000; 16:276–277. [PubMed: 10827456]

Rodriguez A, Griffiths-Jones S, Ashurst J, Bradley A. Identification of mammalian microRNA host genes and transcription units. Genome Res. 2004; 14:1902. [PubMed: 15364901]

Royce T, et al. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. Trends Genet. 2005; 21:466–475. [PubMed: 15979196]

Royce T, Rozowsky J, Gerstein M. Assessing the need for sequence-based normalization in tiling microarray experiments. Bioinformatics. 2007; 23:988. [PubMed: 17387113]

Ruepp A, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. 2004; 32:5539. [PubMed: 15486203]

Samanta M, Tongprasit W, Sethi H, Chin CS, Stolc V. Global identification of noncoding RNAs in Saccharomyces cerevisiae by modulating an essential RNA processing pathway. Proc Nat Acad Sci. 2006; 103:4192–4197. [PubMed: 16537507]

Selinger D, et al. RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. Nat Biotechnol. 2000; 18:1262–1268. [PubMed: 11101804]

Stajich J, et al. The Bioperl toolkit: Perl Modules for the life sciences. Genome Res. 2002; 12:1611. [PubMed: 12368254]

Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008; 24:637. [PubMed: 18218656]

Stein L, et al. The generic genome browser: a building block for a model organism system database. Genome Res. 2002; 12:1599–1610. [PubMed: 12368253]

Strauss J, et al. Crel, the carbon catabolite repressor protein from *Trichoderma reesei*. FEBS Lett. 1995; 376:103–107. [PubMed: 8521952]

Ter-Hovhannisyan V, Lomsadze A, Chernoff Y, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 2008; 18:1979. [PubMed: 18757608]

Valkonen M, Penttilä M, Saloheimo M. The ire1 and ptc2 genes involved in the unfolded protein response pathway in the filamentous fungus *Trichoderma reesei*. Mol. Genet. Genomics. 2004; 272:443–451.

Warren A, Archuleta J, Feng W, Setubal J. Missing genes in the annotation of prokaryotic genomes. BMC Bioinform. 2010; 11:131.

Wilhelm B, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008; 453:1239–1243. [PubMed: 18488015]

Zhao Y, et al. MicroRNA regulation of messenger-like noncoding RNAs: a network of mutual microRNA control. Trends Genet. 2008; 24:323–327. [PubMed: 18514357]

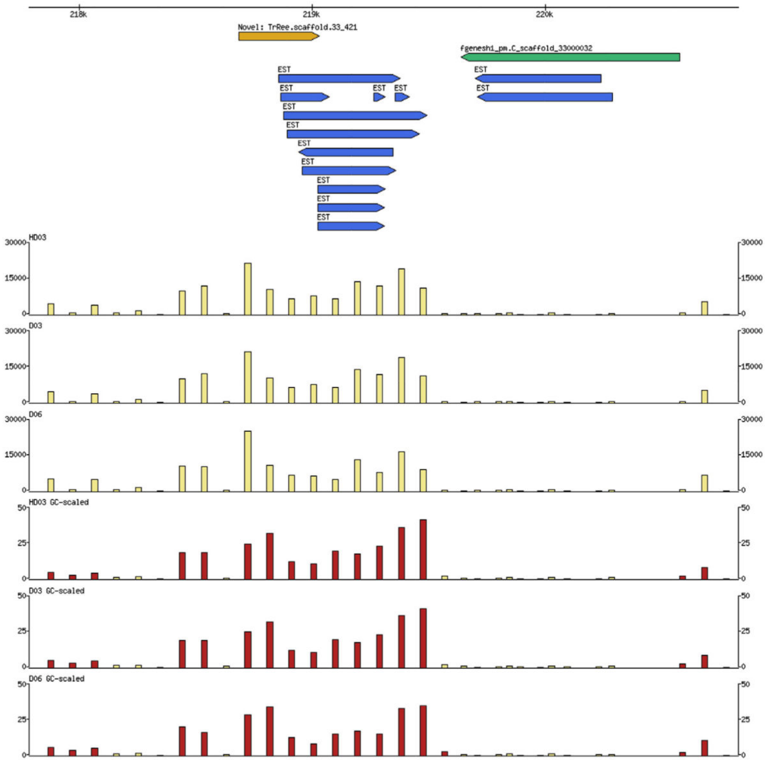

Fig. 1.

33_421 with probe signal and EST data. On top, a ruler with genomic coordinates in scaffold 33. Below a panel that shows positions of old genes (green), candidate gene (yellow) and ESTs (blue) as arrows. Arrows point in the direction of transcription. The bottom 6 panels show average signals of the three repeats of each condition from individual probes as vectical bars, before and after GC% scaling. Bars are positioned at the genomic location of the probes as specified by the top ruler. Bars for probes whose signal value is above the 75th percentile of signals of all intergenic probes are colored in red. The location of the candidate gene is based on a simple ORF prediction that does not take into account splicing nor UTRs.

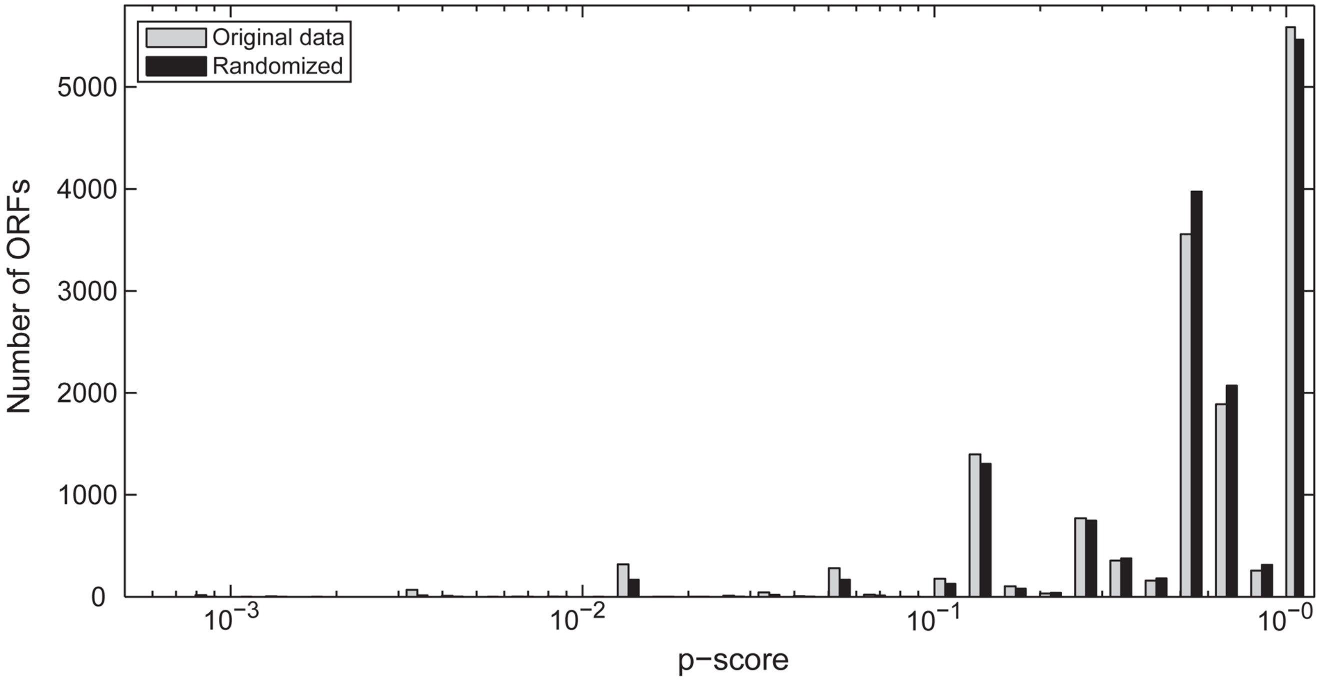

Fig. 2.

Histograms of p-score distributions. The distribution of the *p*-scores for each ORF is shown in the original data, as well as their averages in 100 randomizations. The results are for the experimental condition HD03, computed according to the 75th percentile.

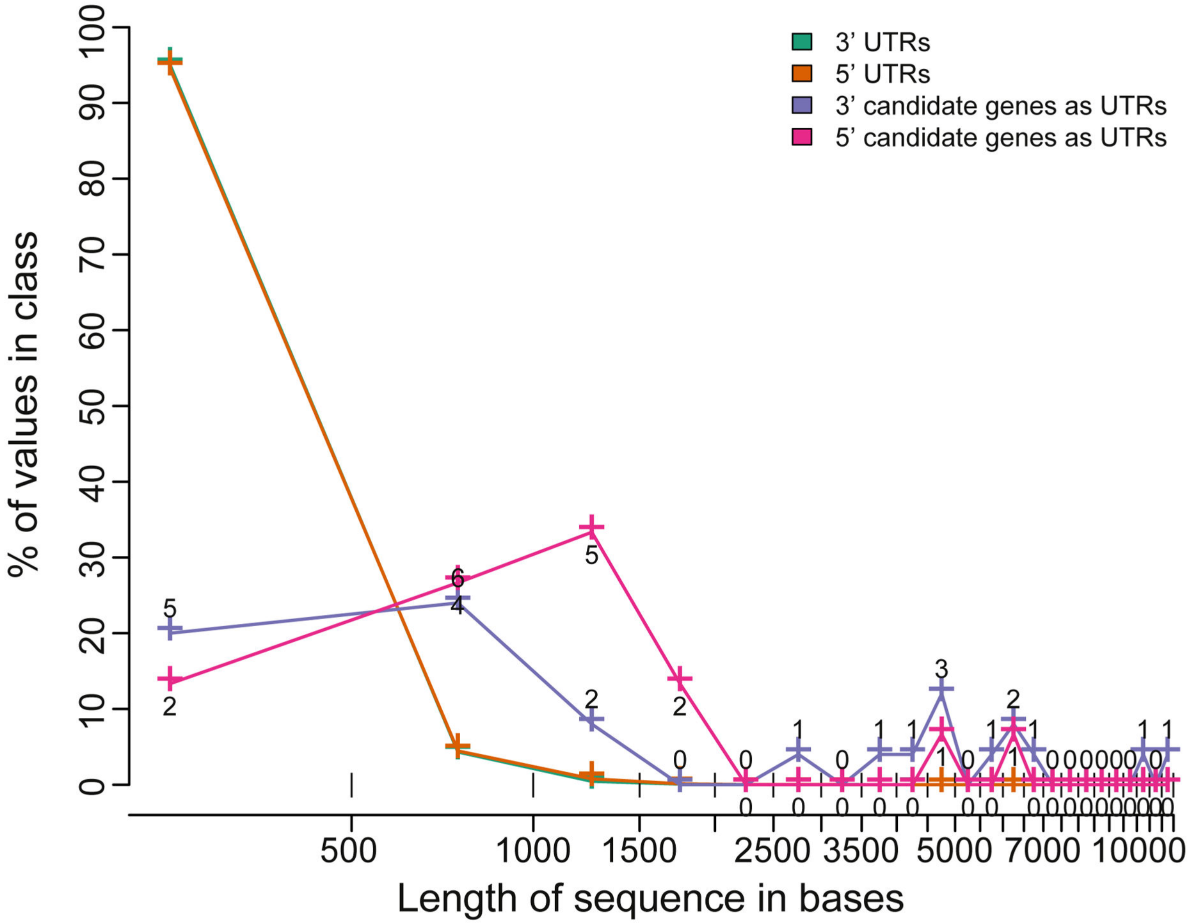

Fig. 3.

Length distribution of verified UTRs and candidate genes as UTRs. The distribution of the length of the candidate genes plus each distance to the nearest stop or start codon, which ever was closer, of an old gene, i.e. the length of candidate gene if it would be a UTR, and the distribution of experimentally verified UTRs in fungi. Plus signs indicate mid values of bins (500 b bins) and lines connect them. The counts of candidate genes are shown for each bin. Candidate genes that overlapped genes in version 2.0 have been excluded. The Y axis shows the percentage of values in a bin for the four categories.

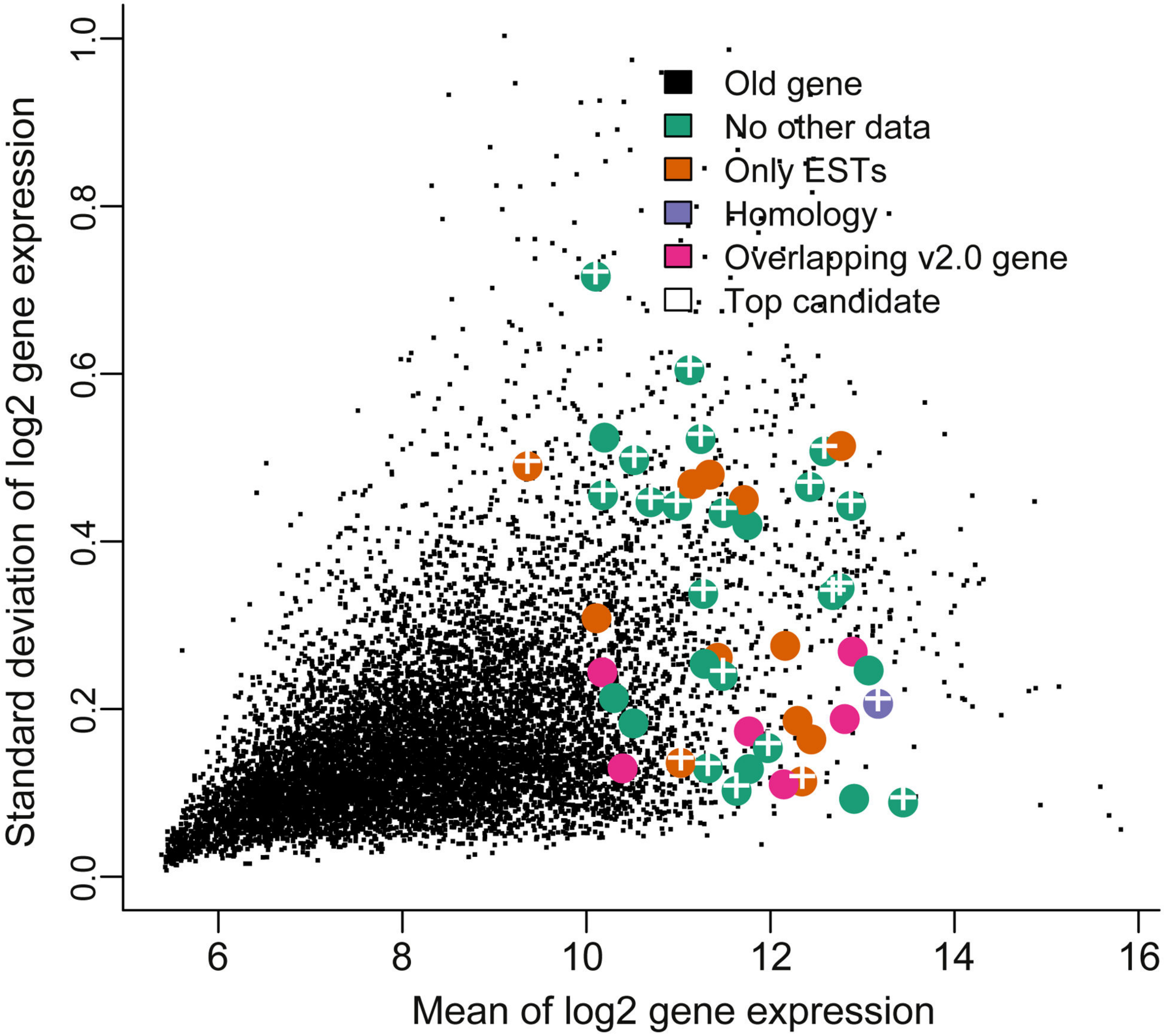

Fig. 4.

Scatterplot of gene expression signals. X axis shows the mean of log2 of gene expression signals for the three conditions and Y axis the respective standard deviation. Small black dots show the signals from old genes. Values for candidate genes are colored based on the evidence found for them: either no other evidence was found, ESTs were found, a homologue was found in another organism or the gene was successfully predicted in version 2.0 of the genome. White plus signs indicate the Top candidate genes which are particularly likely to be true novel genes based on analysis of UTR sequences.

Table 1

Candidate genes. Details of candidate genes: strand, scaffold and position in genome version 2.0, number of probes included in the gene, GC% of the gene, average p-score of the 3 conditions (HD, D03 and D06), *T. reesei* v2.0 protein identification number and the identification number in Fusarium verticillioides, counts of overlapping ESTs, microarray signal average and standard deviation of the 3 conditions after Robust Multichip Average (RMA) (Fig. 4), microarray signal correlation and average difference for the nine samples to the nearest neighbour gene on the same strand (Supplementary file 4), median size of the multispecies protein cluster in the genomic window (Supplementary file 5), distance to nearest old gene, Top candidate classification and if Augustus predicted an overlapping gene on the same strand.

Name	v2.0 scaffold	v2.0 strand	v2.0 start	v2.0 end	Number of probes	GC%	Average p-score	*T. reesei* v2.0 protein id.	EST count	Mean of expression	St. dev. of expression	Correlation to nearest	Mean difference to nearest	Median of window cluster size	Nearest start or stop	Top candidate	Augustus overlap
1_72	1	−	2,165,189	2,165,534	4	50	0.00391		0	10.1	0.7	−0.69	3.85	179.5	1135	Yes	
1_1624	1	−	1,422,866	1,423,277	4	55	0.00391		2	12.3	0.1	0.19	5.44	11	525		
1_3454	1	−	563,951	564,329	4	53	0.00391	73,654	2	11.8	0.2	0.58	1.54	16	0		Yes
2_1351	2	+	789,518	789,857	4	51	0.00391		0	12.8	0.3	−0.39	3.04	5.5	1140	Yes	
2_1404	2	+	814,943	815,342	4	58	0.00391		0	11.3	0.3	−0.71	1.13	32	65		
2_2919	2	+	1,548,526	1,548,883	4	43	0.00391		3	11.3	0.5	−0.39	2.88	18.5	187		
4_2310	5	−	469,257	470,004	9	60	0.00093		2	12.3	0.1	−0.50	2.12	20	376	Yes	
4_2591	5	−	330,157	330,439	4	60	0.00391	76,269	0	12.1	0.1	−0.30	2.41	151	0		
5_2195	7	−	327,787	329,404	18	54	0.0005	121,336	3	10.4	0.1	0.87	−0.52	7	0		
5_2510	7	−	180,823	181,177	4	62	0.00391		0	12.0	0.2	0.63	2.39	32	3669	Yes	
6_2389	27	+	252,264	252,558	4	46	0.00391		3	11.2	0.5	0.87	1.18	12.5	301		
7_1555	10	−	420,427	420,769	4	52	0.00391		0	11.5	0.2	−0.46	5.61	2	5772	Yes	
7_1746	10	−	335,489	335,840	4	53	0.00391		1	12.8	0.5	−0.28	2.73	34.5	334		
8_1028	4	−	1,206,242	1,206,716	5	69	0.00098		0	13.4	0.1	0.37	4.78	42.5	1342	Yes	
8_1718	4	−	894,365	894,938	7	58	0.00134		0	12.9	0.4	−0.44	4.12	14	740	Yes	
9_273	8	+	129,617	130,190	6	57	0.00024		0	11.2	0.5	−0.52	1.40	10.5	6016	Yes	
9_407	8	+	185,130	185,703	7	61	0.00134		0	11.8	0.1	−0.08	2.00	11	0		
9_1935	8	+	919,787	920,273	6	56	0.00024		0	11.5	0.4	0.81	2.07	38	1130	Yes	
10_196	9	+	88,874	89,228	4	57	0.00391		0	11.6	0.1	0.66	3.47	51.5	2478	Yes	
10_591	9	+	279,191	279,797	7	45	0.00134		3	10.1	0.3	−0.53	1.37	14	370		
10_1091	9	+	528,125	528,986	9	53	0.00134		0	11.3	0.3	−0.55	4.24	15.5	343	Yes	
10_1841	9	+	880,777	881,140	4	37	0.00391		2	9.4	0.5	−0.79	−0.67	20	896	Yes	
11_782	3	−	569,507	569,999	5	54	0.00098		0	12.6	0.5	0.91	2.17	20	4398	Yes	
11_1798	3	−	115,291	115,630	4	46	0.00391		2	12.2	0.3	−0.15	2.54	24.5	39		

Name	v2.0 scaffold	v2.0 strand	v2.0 start	v2.0 end	Number of probes	GC%	Average p-score	*T. reesei* v2.0 protein id.	EST count	Mean of expression	St. dev. of expression	Correlation to nearest	Mean difference to nearest	Median of window cluster size	Nearest start or stop	Top candidate	Augustus overlap
12_1290	15	−	288,959	289,244	4	39	0.00391		0	10.5	0.2	−0.35	1.89	18.5	13		
14_172	12	−	926,824	927,403	7	51	0.00519		0	10.2	0.5	−0.80	−0.55	49	5572	Yes	
14_1144	12	−	487,635	488,241	6	61	0.00024	63,395	0	12.8	0.2	0.24	4.38	49	0		Yes
14_1295	12	−	416,882	417,143	4	46	0.00391	78,988	2	10.2	0.2	0.00	1.29	20.5	0		
15_517	17	−	562,390	562,699	4	45	0.00391		0	10.7	0.4	0.19	4.60	3	4549	Yes	
16_41	1	−	3,734,232	3,734,556	4	57	0.00391		0	12.4	0.5	−0.60	4.01	179.5	5738	Yes	
17_578	18	−	402,235	402,751	6	58	0.00024		3	12.4	0.2	−0.22	1.90	7	10		
19_431	19	+	191,393	192,428	11	56	0.00331		0	10.3	0.2	−0.70	1.31	29	0		
19_455	19	+	203,731	204,390	4	48	0.00391		0	10.5	0.5	−0.88	1.53	33.5	9717	Yes	
19_1159	19	+	531,096	531,483	4	60	0.00391		0	11.3	0.1	−0.14	1.73	27	4563	Yes	
20_864	20	−	210,277	210,682	4	57	0.00391	110,403	0	12.9	0.3	0.44	6.81	25.5	0		
21_266	11	+	672,193	672,625	5	44	0.00098		0	11.1	0.6	0.67	0.84	13.5	3136	Yes	
21_435	11	+	752,591	752,972	4	60	0.00391		0	13.1	0.2	−0.50	3.32	19.5	7		
24_795	11	−	155,005	155,326	4	46	0.00391		3	11.4	0.3	0.36	2.21	26.5	246		
25_407	22	−	346,020	346,377	4	61	0.00391		0	12.9	0.1	−0.20	0.17	13.5	26		Yes
25_1011	22	−	33,336	33,750	5	48	0.00098		0	10.2	0.5	−0.33	2.19	14.5	16		
26_552	3	−	1,520,215	1,520,662	5	54	0.00098		0	11.8	0.4	−0.80	3.24	16	68		
28_34	24	−	485,172	485,547	4	57	0.00391		0	12.7	0.3	−0.78	5.85	24	5228	Yes	
28_371	24	−	329,677	329,983	4	56	0.00391		2	11.0	0.1	−0.09	2.35	6.5	1035	Yes	
33_421	8	+	1,271,898	1,272,243	4	59	0.00391	FVEG_02368.3	9	13.2	0.2	−0.25	5.79	4.5	10,884	Yes	
43_73	9	−	1,181,925	1,182,387	5	61	0.00098		7	12.3	0.2	0.67	0.54	28	108		
45_18	12	−	151,545	151,908	4	47	0.00391		3	11.7	0.4	−0.10	3.20	1	360		
50_81	36	−	96,949	97,273	4	48	0.00391		0	11.0	0.4	−0.83	−0.23	37.5	4303	Yes	

Table 2

Domain enrichment in genes surrounding candidate genes. For each enriched InterPro identifier: whether the identifier is enriched in Top candidate genes (tc), candidate genes (c) or both; count of neighbouring genes used for the enrichment test; description of the identifier; count of genes and p-value in the smallest gene set the identifier was found in and biological process the identifier is related to, as intepreted by authors.

Identifier	Gene set	Window	Description	Count of genes	p-value	Author intepretation
IPR000909	tc, c	2	Phospholipase C	2	0.003	Signalling
IPR000182	tc, c	6	GCN5-related N-acetyltransferase	5	0.006	Signalling
IPR001900	c	6	Ribonuclease II/R	2	0.007	RNA
IPR003661	c	6	Signal transduction histidine kinase	3	0.011	Signalling
IPR006634	c	6	TRAM/LAG1/CLN8 homology domain	2	0.022	Various
IPR003864	tc, c	6	Protein of unknown function DUF221	2	0.022	Unknown
IPR000219	c	6	Dbl homology (DH) domain	2	0.033	Signalling
IPR000086	tc	10	NUDIX hydrolase domain	3	0.042	Unknown