Horseradish peroxidase isoenzyme discovery and production in a new improved *Pichia pastoris* expression system

Doctoral thesis

Laura Hannele Näätsaari

This work was carried out under the supervision of Ao. Univ.-Prof. Mag.Dr.rer.nat. Anton Glieder at the Institute of Molecular Biotechnology, Graz University of Technology between June 2008 and June 2012.

Deutsche Fassung:

Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008

Genehmigung des Senates am 1.12.2008

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Graz, am

…………………………..…………………………………………..

(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

………………………………………………………………………………………………………..

date (signature)

TABLE OF CONTENTS

ABSTRACT

The popularity of the methylotrophic yeast *Pichia pastoris* has continuously increased in the last decades, making *P. pastoris* one of the major eukaryotic hosts for recombinant protein production. By knocking out the *KU70* gene coding for one of the proteins essential for the non-homologous end-joining pathway, a strain with high homologous and low random integration frequencies was made. In addition, a platform providing a broad spectrum of *P. pastoris* strains and *P. pastoris*/*Escherichia coli* shuttle vectors for heterologous protein production was generated. The simplicity of targeted integration in the *KU70* knock-out strain will facilitate further developments of this platform and aid the functional characterization of the genome and metabolic routes of *P. pastoris*. The *P. pastoris* platform generated was utilized for the successful production of 22 horseradish peroxidase (HRP) isoenzymes. Although HRPs have long been used for biomedical and biotechnical applications, the commercial enzymes are still isolated from the roots of *Armoracia rusticana*. Lack of sequence information and efficient heterologous production systems have been limiting the developments in the recombinant production of single isoenzymes. *A. rusticana* transcriptome sequencing, ORF analysis, automated search of conserved secretory peroxidase domains and manual verification of the resulting sequences yielded 28 HRP isoenzymes. Successful production of a set of HRP isoenzymes with a diversity of properties concerning substrate specificity, pI and activity provides a basis for the utilization of the enzymes in various applications. The diversity of the naturally occurring HRP isoenzymes was further increased by surface engineering isoenzyme C1A for oriented immobilization. Successful immobilization and application of the surface modified enzyme produced in *P. pastoris* for the degradation of endocrine disruptor compounds present in wastewaters showed the potential of the versatile set of HRP isoenzymes in novel applications.

Keywords: *Pichia pastoris*, *Komagataella pastoris*, *Armoracia rusticana*, horseradish peroxidase, expression platform, protein production, enzyme discovery, oriented enzyme immobilization

KURZFASSUNG

Die Bedeutung der methylotrophen Hefe *Pichia pastoris* hat in den letzten Jahrzehnten kontinuerlich zugenommen, wodurch *P. pastoris* zu einem der wichtigsten eukaryotischen Zellystemen für die rekombinante Proteinproduktion geworden ist. Durch die Entfernung des *KU70* Gens, das eines der Proteine kodiert, die für den nicht-homologen Reparaturmechanismus verantwortlich sind, wurde ein Stamm mit hoher homologer und niedriger zufälliger Integrationshäufigkeit generiert. Zusätzlich wurde eine Plattform bestehend aus einen breiten Spektrum an *P. pastoris* Stämmen und *P. pastoris/E. coli* Shuttlevektoren für die heterologe Proteinproduktion hergestellt. Die einfache gezielte Integration in den *KU70* knock-out Stamm wird die weitere Entwicklung dieser Plattform erleichtern und die funktionelle Charakterisierung des Genoms und metabolische Wege von *P. pastoris* unterstützen.

Die generierte *P. pastoris* Plattform wurde erfolgreich für die Produktion von 22 Peroxidase Isoenzymen aus Kren (HRP) verwendet. Obwohl HRPs schon seit längeren für biomedizinische und biotechnologische Anwendungen eingesetzt werden, werden die kommerziellen Enzyme noch immer aus den Wurzeln von Kren (*Armoracia rusticana*) isoliert. Das Fehlen von Sequenzinformation und effizienter Produktionssysteme hat bis jetzt die Entwicklungen in der rekombinanten Produktion der einzelnen reinen Isoenzyme limitiert. Die Transkriptom Sequenzierung von *A. rusticana* mit darauffolgender ORF Analyse, automatisierter Suche von konservierten Domänen von sekretorischen Peroxidasen und manueller Verifikation der resultierenden Sequenzen ergab 28 HRP Isoenzyme. Die erfolgreiche Produktion einer Palette an HRP Isoenzymen mit unterschiedlichen Eigenschaften bezüglich Substratspezifizität, pI und Aktivität stellt eine wichtige neue Basis für unterschiedliche Anwendungen dieser Enzyme dar. Weiters wurde die Diversität der natürlich vorkommenden HRP Isoenzyme durch das Engineering der Oberfläche des Isoenzyms C1A erhöht, um eine orientierte Immobilisierung zu ermöglichen. Die erfolgreiche Immobilisierung und Anwendung des oberflächenmodifizierten, in *P. pastoris* produzierten Enzyms für den Abbau schädlicher hormonähnlicher Substanzen, wie man sie im Abwasser findet, zeigt das Potential dieser vielseitigen Pallette and HRP Isoenzymen in neuen Anwendungen.

ACKNOWLEDGEMENTS

This study was carried out at the Institute of Molecular Biotechnology, Graz University of Technology during 2008-2012 under the supervision of Ao. Univ.-Prof. Mag.Dr.rer.nat. Anton Glieder, as a part of the PhD training program DK Molecular Enzymology. I wish to thank DK Molecular Enzymology and FWF for financial support (Fonds zur Förderung der wissenschaftlichen Forschung grant #W901 DK Molecular Enzymology), my supervisor Anton Glieder for giving me the opportunity to work on challenging and versatile projects, and the head of the institute Helmut Schwab (&other IMBT staff) for providing excellent research facilities.

My sincere thanks go to all collaborators & co-authors and past & present members of the Glieder group, especially to the 2nd floor gang: THANK YOU Amneris, Andrea, Andreas B., Andreas K., Claudia, Flo, Maria, Martina, Sasha, Tanja, Thomas, Viktorija and all others, this project would not have been possible without the exceptional group spirit, discussions (scientific and so-not-scientific) and contributions of each one of you.

I am deeply grateful to my parents for unconditional love and support during my thesis. *KIITOS, äiti ja isä!* Special thanks go also to my boyfriend Harald for all the love and patience, and to my bonus mom & dad Evy and Butch and sis Catalina for showing interest in my work and for believing in me.

Marjo, Kati, Riitta, Johku, Terja, Heli and all other dear friends, thank you for the reliability and understanding, the moments together, for giving me hope and a broader view of life.

INTRODUCTION

Background and aim of the thesis

Horseradish peroxidases isolated from the roots of the perennial plant *Armoracia rusticana* have been used in various applications for decades. Due to the increasing need of peroxidases (Figure 1), *A. rusticana* is cultivated around the world to guarantee consistent availability of the roots. Industrial units process tons of horseradish roots to isolate and chromatographically separate groups of peroxidases with variable properties. Although some of the traditional applications in molecular biology and medicine do not necessarily require preparations consisting of only one isoenzyme, new applications like ADEPT (antibody dependent enzyme prodrug therapy) (1), enzyme immobilization (2) and biocatalysis (3) are based on homogenous preparations with consistent quality. A set of characterized isoenzymes with versatile properties is needed for the expanding range of applications.

Attempts to produce HRP in heterologous hosts have been successful, but limited to only one isoenzyme HRP C1A and resulting in variable but generally low yields. The highest yields have been reached with the methylotrophic yeast *P. pastoris* (4). Heterologous expression of a variety of HRP isoenzymes has also been limited by lacking sequence information. The amino acid sequences of only eight isoenzymes have been published previously (Table 1).

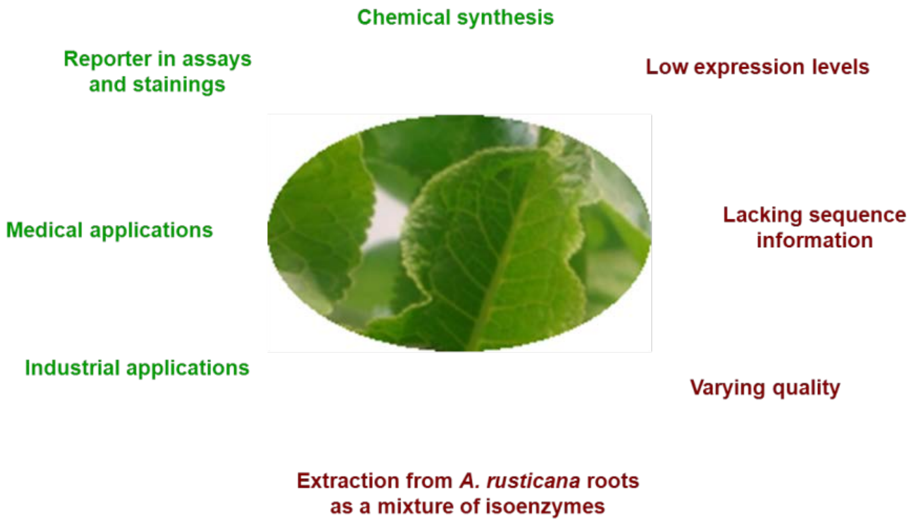

Chemical synthesis
Reporter in assays and stainings
Low expression levels
Medical applications
Lacking sequence information
Industrial applications
Varying quality
Extraction from *A. rusticana* roots as a mixture of isoenzymes

Figure 1: Horseradish peroxidases have traditionally been used as reporters in biological assays and stainings. Of late they have also gained interest in new fields like medical applications, chemical synthesis and other industrial applications. Despite the increasing importance of pure HRP preparations, due to the lacking sequence information and low expression levels reached in the heterologous hosts, the commercially available HRP is still extracted from the roots of *A. rusticana* as a mixture of isoenzymes, resulting in preparations of varying consistency and quality.

This study has three major aims:

1. Generation of a versatile *Pichia pastoris* (*Komagataella pastoris*) expression platform with research and production strains and expression vectors with multiple choices concerning, *e.g.*, markers, copy numbers, protein localization, promoter efficiency, carbon sources and co-expression of helper proteins (**Chapter 1**).
2. Exploration of the HRP isoenzyme diversity in *A. rusticana* to provide a group of enzymes with various properties concerning substrate specificity, specific activity, pH optimum and stability for the new and promising applications in medicine and industry, followed by sequence verification and heterologous production of the isoenzymes in *P. pastoris* (**Chapter 2**).
3. Improvement of the essential properties of a chosen isoenzyme for new applications by protein engineering (**Chapter 3**).
4. Collection and evaluation of a set of sensitive peroxidase assays (**Chapter 4**).

Figure 2: Horseradish peroxidase isoenzymes discovered from the transcriptome and genome of *A. rusticana* (A) were heterologously produced in the methylotrophic yeast *P. pastoris* (B) reaching very high cell densities (C) in bioreactor cultivations. The ABTS assay (D) is one of the most sensitive colorimetric assays described in **Chapter 4** for the detection of peroxidase activity.

Pichia pastoris (*Komagataella pastoris*)

The popularity of the methylotrophic yeast *Komagataella pastoris*, previously named and still commonly referred to as *Pichia pastoris*, has continuously increased in the last decades, making *P. pastoris* one of the major eukaryotic hosts for recombinant protein production. Hundreds of recombinant proteins have been produced in *P. pastoris* since the 1970's, when Philips Petroleum Company developed the first protocols for high cell density cultures on methanol –containing media. The advantages of *P. pastoris* in recombinant protein production have been extensively reviewed (5–8). In addition to the powerful, regulated promoters (9), ability to perform complex post-translational modifications and the ease and frugality of the manipulation and growth (5), *P. pastoris* is not known to contain viral or oncogenic nucleic acids like mammalian cells, or toxic cell wall components like *E. coli* (6, 10). Thus it has also gained acceptance as a host to produce antibody fragments (11), growth factors (12), serum components (13) and other pharmaceutical products passing through clinical trials. The characteristic high-mannose glycosylation pattern of proteins produced in *P. pastoris* has been reported to be of advantage with medical products needing fast clearance (14). Also the first strains with modified or humanized glycosylation have been made available, thus enabling also the production of pharmaceuticals where avoiding immunogenic responses is necessary (15).

The absence of a commonly accessible genome sequence of *P. pastoris* has been limiting the advancements and opportunities to modify the endogenous pathways of the species as well as strain engineering to improve the expression of difficult to express proteins such as the heme peroxidase HRP. Recently, the genome and mitochondrial genome sequences of the unmodified wild type strain CBS7435 (NRRL-Y11430, ATCC 76273) were published (16), offering a possibility to gain better understanding of, *e.g.*, the function of endogenous proteases, efficient secretion signals and glycosylation. However, the generation of specific changes like targeted knock-outs has been observed to function with variable success, limited by the non-homologous end-joining (NHEJ) mechanism leading to unspecific integration events. NHEJ has been reported to be common in filamentous fungi and higher eukaryotic organisms (17–20), and also seems to play a substantial role in the integration events of knock-out cassettes with flanking homologous sequences in *P. pastoris* (21, 22).

Chapter 1 describes the development and testing of a knock-out strain with reduced NHEJ and increased homologous integration frequencies. In this study, the genes *KU70* and *KU80* coding for the two subunits of the heterodimer Ku70p/Ku80p binding to double-strand DNA ends were identified in the genome sequence of the wild type strain CBS7435. The *KU70* gene was knocked out utilizing the FLP recombinase system from *Saccharomyces cerevisiae* 2µm plasmid (Figure 3) (23). The advantageous properties of this new platform strain were characterized targeting well known loci *HIS4* and *ADE1*, and further tested with genes *dihydroxyacetone synthase 1*, *dihydroxyacetone synthase 2* and *glycerol kinase 1*, from the methanol and glycerol assimilation pathways of *P. pastoris*. In addition, **Chapter 1** describes the development of a well-defined *P. pastoris* expression platform generated based on the wild-type strain CBS7435. The strains included in the platform were constructed with specific knock-outs limiting undesired mutations in the genome. The *E. coli*/*P. pastoris* shuttle vectors included in the expression platform enable the production of a wide variety of proteins and allows the choice of markers, promoter efficiency and carbon source, localization and co-expression of helper proteins to improve folding. Due to the properties of the *ku70* deletion strain, the platform can facilitate the functional characterization and further development of *P. pastoris* towards an even more versatile host for industrial protein production including the engineering of the host to improve HRP production.

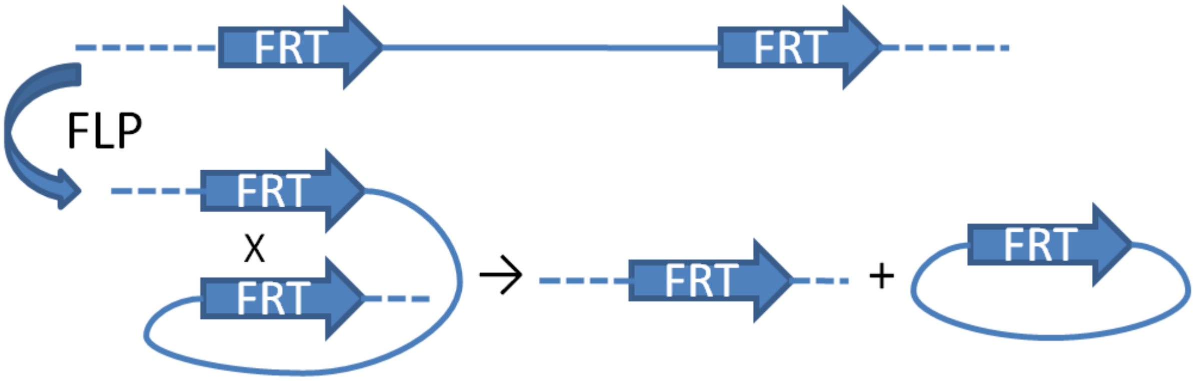


Figure 3: All platform strains generated during this study were constructed utilizing the FLP recombinase system from *Saccharomyces cerevisiae* 2µm plasmid (23). The site-specific recombinase FLP recognizes two 34bp target sequences (FRTs) both including an 8bp core sequence surrounded by two 13bp FLP contact sites. Placing two FRTs as direct repeats results in the excision of the sequence placed in between, leaving only one FRT in the genome. Thus the recombination system can be considered as a method to create knock-out strains without leaving any markers behind in the genome.

Horseradish peroxidases (HRPs)

Horseradish peroxidases (HRPs) produced by the perennial herb *Armoracia rusticana* (*Brassicaceae*) are heme-containing monomeric glycoproteins belonging to the class III plant peroxidase subfamily (24). The alpha-helices constituting a major part of the structure are highly conserved across plant peroxidases (2). The sizes of the known HRP isoenzymes prior to processing vary from 305 to 353 amino acids, corresponding to an approximate molecular weight of ~32 to 36 kDa for the mature protein. In addition, the heme prosthetic group and two calcium ions increase the molecular weight by ~0.7 kDa. A coordinate bond between the heme iron atom and the side chain of the proximal histidine (His170) is attaching the heme to the residue. The distal coordination site is available for hydrogen peroxide during catalysis. Each HRP molecule contains two calcium ions essential for the stability and activity of the peroxidases (24, 25), and one structural water molecule (Figure 5).

The mechanism of the HRP catalysis is presumed to proceed mainly through the classical cycle of peroxidase catalysis. The resting ferric state of heme is oxidized by H_2O_2 forming compound I two oxidizing equivalents above the resting state. Alternative hypotheses of the detailed reactions connected to compound I formation exist (26–28). Compound I contains a porphyrin-based π-cation radical and an oxoferryl center. The enzyme is returned to the resting state through two one-electron reduction steps connected to the substrate molecule oxidation. The π-cation radical is neutralized in the first step forming the second intermediate, compound II (oxidation state +4) and further reduced to the resting state (29). Especially residues Arg38, Phe41, His42 and Asn70 (numbering from mature isoenzyme C1A) are known to play a crucial role in the catalysis (24). In addition to the peroxidatic cycle described above, hydroxylic and oxidative cycles have been described to exist (30).

The glycosylation pattern of the known HRP isoenzymes is heterogenous but significant. HRP C1A has 8 occupied N-glycosylation sites (24) and the carbohydrate chains attached have been reported to have a molecular mass of ~9.4 kDa (31) in the enzyme isolated from *A. rusticana* roots. Although the first HRP isoenzymes have been characterized in the 1960’s (32) and the total number of isoenzymes has been estimated to be as high as 42 (33), the amino acid sequences of only eight isoenzymes (summarized in Table 1) are available in UniProtKB and the nucleotide sequences of six isoenzymes (C1A, C1B, C1C, C2, C3 and N) are accessible in the NCBI databases. The group C isoenzymes have been annotated to have an N-terminal signal peptide and reported to consist of four exons with identical splice site positions (24). A more exact analysis of the HRP sequences is presented in **Chapter 2**.

Table 1: Summary of the known HRP isoenzymes according to UniProtKB and calculations using Expasy Compute pI/Mw tool. Calculated properties in this table differ from the properties reported in Chapter 2 due to sequence divergence and signal sequence prediction. pI = isoelectric point, aa = amino acid, kDa = kiloDaltons

Isoenzyme name	Accession number	pI (mature protein)	Length aa (mature protein)	Mw kDa
C1A	P00433	6.35	308	33.9
C1B	P15232	5.84	323	35.6
C1C	P15233	6.12	323	35.6
C2	P17179	8.56	323	35.4
C3	P17180	7.71	320	35.3
A2	P80679	4.72	305	31.9
E5	P59121	9.13	306	33.7
N	Q42517	5.96	299	32.1

The large estimated number of existing HRP isoenzymes in older publications can be due to the limited possibilities of protein analysis mistaking, for example, glycovariants for separate isoenzymes. However, the roles of the HRP isoenzymes in nature are multiple and diverse, thus supporting the estimation of a large group of isoenzymes with versatile properties including optimal conditions and substrate specificities. Although the detailed functions of HRP isoenzymes in the plant are unknown, the roles of peroxidases in nature have been extensively reviewed (34), and are summarized in figure 3 (35).

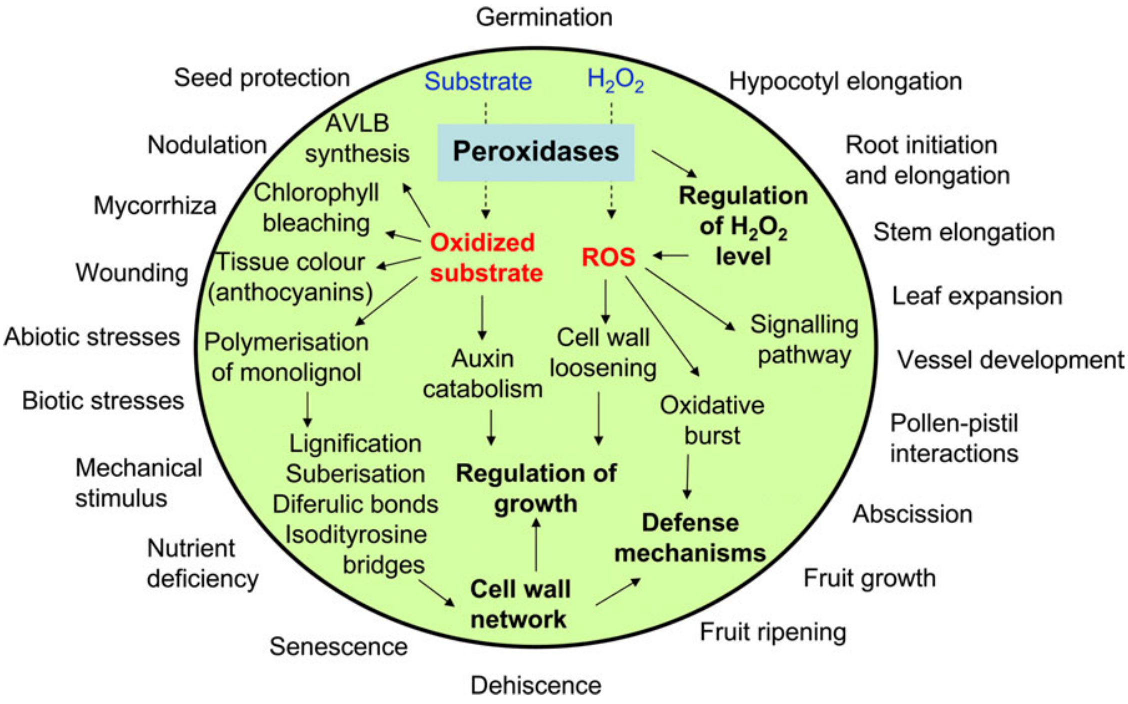


Figure 4: Roles of the peroxidases in nature as summarized by Cosio and Dunand (2009) (35).

Lacking sequence information and the low yields of HRP reached in heterologous hosts have impeded the production and commercial availability of individual HRP isoenzymes. All commercially available preparations are still extracted from the roots of *A. rusticana* as a mixture of isoenzymes. **Chapter 2** describes an approach to discover a large group of isoenzymes of non-model organism origin utilizing transcriptome sequencing, automatized ORF finding and characterization of the conserved domains, manual curation of the sequences and heterologous expression of the discovered, sequence verified isoenzymes in *P. pastoris*. A *de novo* assembly of the almost 600 000 reads resulting from the 454 pyrosequencing generated 14871 isotigs and further characterization and manual sequence verification led to the discovery of a total of 28 HRP isoenzymes. Twenty-two isoenzymes were successfully produced in *P. pastoris* and showed peroxidase activity. Thus, this study provided the basis for large scale production of single isoenzymes with consistent quality but versatile physical and chemical properties.

Essential applications for HRPs

Due to the enzyme's robustness and multiple existing chromogenic substrates, horseradish peroxidases have traditionally been used in histological staining and diagnostic assays. Diagnostic kits for medical analytics are being continuously further developed towards more simple, sensitive analytics in miniature size (36), covering a wide range of analytes such as glucose, lactose, cholesterol, toxins, pathogens and even cancer markers (37), extensively reviewed by (38). In the last decades, also the potential of horseradish peroxidases in biotechnological and industrial applications has been noted to be significant. HRP has been used in polymerization reactions of aromatic and phenolic amine compounds (39), in organic synthesis (3) and in degradation of recalcitrant organic compounds like substituted phenols (40) and industrial azo dyes (41).

HRPs have not only turned out to be beneficial in medical analytics, but also in new experimental therapies. Targeted cancer therapies with HRP in combination with indole-3-acetic acid (IAA) have proceeded to clinical trials (42). The non-toxic plant hormone IAA, also known as auxin, can be activated by HRP to produce cytotoxic compounds (peroxyl, skatolyl and indolyl radicals, ROS) activating caspase-3 and PARP (poly ADP-ribose polymerase) cleavage leading to death-receptor mediated apoptosis (43, 44) of the tumor cells with a distinct bystander effect. The skatolyl radical has also been reported to be reactive towards DNA in anoxic conditions (45) often prevailing in tumors. Targeting of HRP to the cancer cells has been suggested to occur with antibodies (ADEPT, antibody-directed enzyme-prodrug therapy), polymers (PDEPT) or genes (GDEPT) (1, 45, 46,). In ADEPT, the patient is first administered an antibody-enzyme conjugate or fusion with an antibody specific to a tumor-specific antigen. The antibody binds to the antigens presented on the surface of the tumor cells, and unbound antibody-enzyme complex is cleared from the circulation. Especially fusion proteins produced in *P. pastoris* have been reported to show enhanced clearance from healthy tissues, with very high tumor to normal tissue ratios (47). The fast clearance can be explained by the oligomannose structures occupying N-glycosylation sites of proteins produced in *P. pastoris*, leading to effective clearance through mannose receptors in hepatic cells (48). Following administration of IAA leads to the formation of cytotoxic compounds locally only around the tumor cells, thus reducing side effects of the therapy. Although the combination of HRP and IAA has proven efficient in preclinical trials, the results published are based on commercial HRP preparations or conjugates instead of fusion proteins. Unknown enzyme mixtures and the variable quality of conjugates limit the possibilities to perform studies with reproducible results. The successful expression of a large group of single isoenzymes with versatile substrate specificities (**Chapter 2**) offers a considerable opportunity to screen for an isoenzyme with high specific activity towards IAA.

Aromatic compounds like phenol derivatives originating from industrial production of, for example, textiles, abrasives, wood composites and coatings are one of the major pollutants in waste waters (38, 49, 50). Benzidine-based dyes are potentially carcinogenic and many bis-phenolic compounds, dioxin and pesticides can function as endocrine disruptors (51). Due to the toxicity of the compounds, their use and removal from waste waters is very tightly regulated in many countries, but require complicated, cost-intensive methods (52). Also hormones originating from medication and ending up in municipal waste waters can cause a threat to the natural ecosystems if let to the nature unprocessed (53). HRP has been reported to be useful in the bioremediation of polluted waters and soils (38, 51). However, processing large amounts of contaminated waters requires immobilization to enable enzyme recovery and retention. Immobilization of an enzyme to carrier materials can also improve stability, enable enzyme performance under optimal process conditions and make reuse and recycling of the enzyme possible permitting also continuous processes (54). Recently, other advantages such as modification of enantioselectivity, substrate selectivity and enhanced activity have been reported (54). Immobilization of HRP has been reported to improve resistance to proteolysis and to increase storage stability (51), and substantially high retention of catalytic activity has been achieved in orientated immobilization of surface mutated HRP produced in *E. coli* (2).

Enzyme immobilization technologies are constantly under development. The existing methods (Figure 4) have been extensively reviewed (54, 55) and can be divided in five types:

1. **Entrapment (typically a polymer network)**
    - Provides enzyme protection from mechanical sheer, gas bubbles and solvents
    - Drawbacks including low enzyme loading and mass transfer limitations
2. **Encapsulation (cross-linked polymers, alginate capsules with silicate shell)**
    - Provides enzyme protection from the environment
    - Has problems with mass transfer limitations especially with larger substrates
3. **Solid support immobilization via:**
    a. Adsorption
    - Simple, inexpensive method without any chemical modification of the enzyme
    - The method is unspecific and enzyme can be leaching out

    b. Ionic binding (functionalized polysaccharide biopolymers)
    - Simple, reversible method where re-use of the support is possible
    - Connected to problems of enzyme leaching out

    c. Covalent binding typically through ε-amino group of lysines (various supports activated with epoxide functional groups)

   - Provides enzyme protection from the environment, increased stability and rigidity through multi-point attachment. Fissured surfaces have high loading capacity and multifunctional supports exist for enhanced binding
   - Chemical modifications can lead to partial enzyme inactivation and steric hindrance cause decreased activity

4. **Self-immobilization (bi-functional cross-linkers, *e.g.* glutaraldehyde)**
   - The method has an especially high specific and volumetric activity combined with high mechanical stability and high toleration of organic solvents
   - Extensive protein purification and optimization are needed and the enzyme used has to be crystallizable
5. **Spontaneous self-immobilization**
   - Can be used only for enzymes spontaneously forming large active aggregates

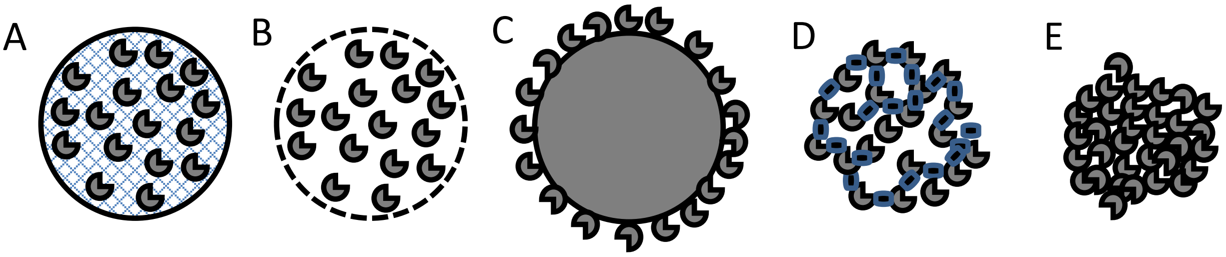


Figure 5: Enyzme immobilization methods. A) Entrapment B) Encapsulation C) Immobilization to a solid support D) Self-immobilization with cross-linkers E) Spontaneous self-immobilization to active aggregates

Covalent linkages have been reported to be the most robust techniques for immobilization (56, 57) and would thus be suitable for applications like waste-water treatment. Lysine residues provide good bond stability and above average reactivity. Therefore lysines located on the surface of proteins are typically used for covalent attachment to activated supports with epoxide functional groups covering the reactive surface (58). **Chapter 3** describes a surface engineering approach to simultaneously increase the stability of HRP isoenzyme C1A, and to produce a surface modified enzyme preparation for oriented enzyme immobilization using covalent binding. Three of the six lysine residues on the surface of HRP C1A have been reported to be accessible to chemicals (Figure 5), and thus can also react with the carrier materials (59). Substitutions to remove the lysines have been demonstrated to increase the stability of the enzyme towards both solvents and heat (60). In this study, single and double mutants to remove the lysines proximal to the substrate access channels are created and produced in *P. pastoris* to provide heterologously produced single isoenzyme with known structure and surface composition to determine the optimal orientation for immobilization and avoidance of steric hindrances of the reaction. Previous HRP immobilization studies have been using either nonglycosoylated HRP produced in *E. coli* (2) or commercial HRP preparations isolated from the horseradish plant (51, 61, 62). The glycosylation grade of HRP produced in *P. pastoris* is known to be higher and more heterogeneous than that of the plant enzyme (4), with a total carbohydrate content of up to 65%. Possible advantages such as stabilization and problems such as blocking of active residues caused by the high glycosylation grade are assessed in **Chapter 3**. In addition, the possibilities to utilize raw cultivation supernatants without purification and other immobilization methods are compared in terms of binding efficiency and retention of catalytic activity.

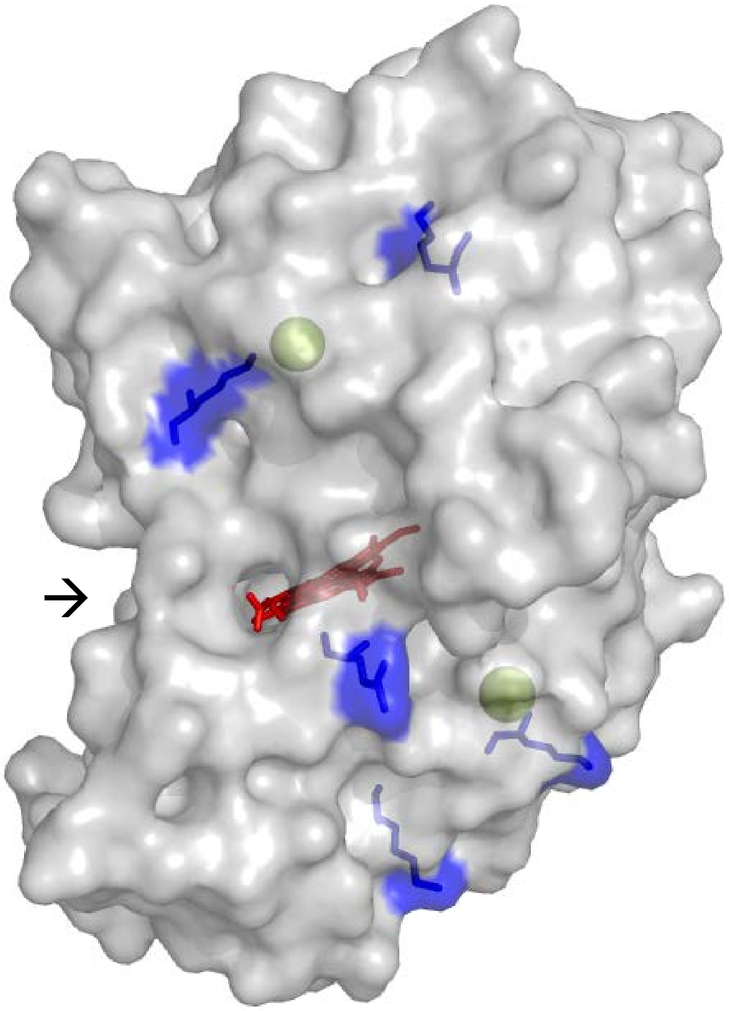


Figure 6: Structure of the horseradish peroxidase isoenzyme C1A used as a basis to construct surface variants for oriented enzyme immobilization. The heme molecule visible through one of the channels leading to the active site is marked red. The arrow on the left side indicates the location of the second, main substrate access channel leading to the active site. Two calcium ions in the structure are marked with green. Lysine residues on the surface are marked with blue. In addition to the five lysine residues visible, one is placed on the reverse side of the molecule. Only three of the lysines (K174, K232 and K241) are accessible to chemical modifications (59) and thus also for interactions with surfaces.

References

1. Greco,O., Rossiter,S., Kanthou,C., Folkes,L.K., Wardman,P., Tozer,G.M. and Dachs,G.U. (2001) Horseradish Peroxidase-mediated Gene Therapy : Choice of Prodrugs in Oxic and Anoxic Tumor Conditions. *Mol. Cancer Ther.*, **1**, 151-160.

2. Ryan,B.J. and O'Fágáin,C. (2007) Arginine-to-lysine substitutions influence recombinant horseradish peroxidase stability and immobilisation effectiveness. *BMC Biotechnol.*, **7**, 86.

3. van de Velde,F., van Rantwijk,F. and Sheldon,R.A. (2001) Improving the catalytic performance of peroxidases in organic synthesis. *Trends Biotechnol.*, **19**, 73-80.

4. Morawski,B., Lin,Z., Cirino,P., Joo,H., Bandara,G. and Arnold,F.H. (2000) Functional expression of horseradish peroxidase in *Saccharomyces cerevisiae* and *Pichia pastoris*. *Protein Eng.*, **13**, 377-84.

5. Lin Cereghino,G.P., Lin Cereghino,J., Ilgen,C. and Cregg,J.M. (2002) Production of recombinant proteins in fermenter cultures of the yeast *Pichia pastoris*. *Curr. Opin. Biotechnol.*, **13**, 329-33.

6. Cregg,J.M., Vedvick,T.S. and Raschke,W.C. (1993) Recent advances in the expression of foreign genes in *Pichia pastoris*. *Bio/technology*, **11**, 905-10.

7. Macauley-Patrick,S., Fazenda,M.L., McNeil,B. and Harvey,L.M. (2005) Heterologous protein production using the *Pichia pastoris* expression system. *Yeast*, **22**, 249-270.

8. Daly,R. and Hearn,M.T.W. (2005) Expression of heterologous proteins in *Pichia pastoris*: a useful experimental tool in protein engineering and production. *J. Mol. Recognit.*, **18**, 119-3.

9. Cregg,J.M., Madden,K.R., Barringer,K.J., Thill,G.P. and Stillman,C.A. (1989) Functional characterization of the two alcohol oxidase genes from the yeast *Pichia pastoris*. *Mol. Cell. Biol.*, **9**, 1316-23.

10. Romanos,M. (1995) Advances in the use of *Pichia pastoris* for high-level gene expression. *Curr. Opin. Biotechnol.*, **6**, 527-533.

11. Freyre,F.M., Vázquez,J.E., Ayala,M., Canaán-Haden,L., Bell,H., Rodríguez,I., González,A., Cintado,A. and Gavilondo,J.V. (2000) Very high expression of an anti-carcinoembryonic antigen single chain Fv antibody fragment in the yeast *Pichia pastoris*. *J. Biotechnol.*, **76**, 157-63.

12. Cereghino,J.L. and Cregg,J.M. (2000) Heterologous protein expression in the methylotrophic yeast *Pichia pastoris*. *FEMS Microbiol. Rev.*, **24**, 45-66.

13. Ohtani,W., Nawa,Y., Takeshima,K., Kamuro,H., Kobayashi,K. and Ohmura,T. (1998) Physicochemical and immunochemical properties of recombinant human serum albumin from *Pichia pastoris*. *Anal. Biochem.*, **256**, 56-62.

14. Sharma,S.K., Pedley,R.B., Bhatia,J., Boxer,G.M., El-Emir,E., Qureshi,U., Tolner,B., Lowe,H., Michael,N.P., Minton,N., *et al*. (2005) Sustained tumor regression of human colorectal cancer xenografts using a multifunctional mannosylated fusion protein in antibody-directed enzyme prodrug therapy. *Clin. Cancer Res.*, **11**, 814-25.

15. Li,H., Sethuraman,N., Stadheim,T.A., Zha,D., Prinz,B., Ballew,N., Bobrowicz,P., Choi,B.-K., Cook,W.J., Cukan,M., et al. (2006) Optimization of humanized IgGs in glycoengineered *Pichia pastoris. Nat. Biotechnol.*, **24**, 210-5.

16. Küberl,A., Schneider,J., Thallinger,G.G., Anderl,I., Wibberg,D., Hajek,T., Jaenicke,S., Brinkrolf,K., Goesmann,A., Szczepanowski,R., et al. (2011) High-quality genome sequence of *Pichia pastoris* CBS7435. *J. Biotechnol.*, **154**, 312-20.

17. Daley,J.M., Palmbos,P.L., Wu,D. and Wilson,T.E. (2005) Nonhomologous end joining in yeast. *Annu. Rev. Genet.*, **39**, 431-51.

18. Critchlow,S.E. and Jackson,S.P. (1998) DNA end-joining: from yeast to man. *Trends Biochem Sci.*, **23**, 394-398.

19. Dudásová,Z., Dudás,A. and Chovanec,M. (2004) Non-homologous end-joining factors of *Saccharomyces cerevisiae*. *FEMS Microbiol. Rev.*, **28**, 581-601.

20. Yamana,Y., Maeda,T., Ohba,H., Usui,T., Ogawa,H.I. and Kusano,K. (2005) Regulation of homologous integration in yeast by the DNA repair proteins Ku70 and RecQ. *Mol. Genet. Genomics*, **273**, 167-76.

21. Higgins,D.R. and Cregg,J.M. eds. (1998) *Pichia* Protocols Humana Press, Totowa, New Jersey.

22. Li,P., Anumanthan,A., Gao,X.-G., Ilangovan,K., Suzara,V.V., Düzgüneş,N. and Renugopalakrishnan,V. (2007) Expression of Recombinant Proteins in *Pichia Pastoris*. *Appl. Biochem. Biotechnol.*, **142**, 105-124.

23. Broach,J.R., Guarascio,V.R. and Jayaram,M. (1982) Recombination within the yeast plasmid 2mu circle is site-specific. *Cell*, **29**, 227-34.

24. Veitch,N.C. (2004) Horseradish peroxidase: a modern view of a classic enzyme. *Phytochemistry*, **65**, 249-259.

25. Friedhoff,J.M. and Haschke,R.H. (1978) CALCIUM-RELATED PROPERTIES OF HORSERADISH PEROXIDASE. *Biochem. Biophys. Res. Commun.*, **80**, 1039-1042.

26. Poulos,T.L. and Kraut,J. (1980) The stereochemistry of peroxidase catalysis. *J. Biol. Chem.*, **255**, 8199-205.

27. Vidossich,P., Fiorin,G., Alfonso-Prieto,M., Derat,E., Shaik,S. and Rovira,C. (2010) On the role of water in peroxidase catalysis: a theoretical investigation of HRP compound I formation. *J. Phys. Chem.*, **114**, 5161-5169.

28. Derat,E., Shaik,S., Rovira,C., Vidossich,P. and Alfonso-Prieto,M. (2007) The effect of a water molecule on the mechanism of formation of compound 0 in horseradish peroxidase. *J. Am. Chem. Soc.*, **129**, 6346-7.

29. Berglund,G.I., Carlsson,G.H., Smith,A.T., Szöke,H., Henriksen,A. and Hajdu,J. (2002) The catalytic pathway of horseradish peroxidase at high resolution. *Nature*, **417**, 463-8.

30. Chen,S.X. and Schopfer,P. (1999) Hydroxyl-radical production in physiological reactions. A novel function of peroxidase. *Eur. J. Biochem.*, **260**, 726-35.

31. Welinder,K.G. and Mazza,G. (1977) Amino-acid sequences of heme-linked, histidine-containing peptides of five peroxidases from horseradish and turnip. *Eur. J. Biochem.*, **73**, 353-8.

32. Shannon,L.M., Kay,E. and Lew,J.Y. (1966) from Horseradish Roots. *J. Biol. Chem.*, **241**, 2166-2172.

33. Hoyle,M.C. (1977) High resolution of peroxidase-indoleacetic acid oxidase isoenzymes from horseradish by isoelectric focusing. *Plant Physiol.*, **60**, 787-93.

34. Passardi,F., Cosio,C., Penel,C. and Dunand,C. (2005) Peroxidases have more functions than a Swiss army knife. *Plant Cell Rep.*, **24**, 255-265.

35. Cosio,C. and Dunand,C. (2009) Specific functions of individual class III peroxidase genes. *J. Exp. Bot.*, **60**, 391-408.

36. Cho,J.-H., Paek,E.-H., Cho,I.-H. and Paek,S.-H. (2005) An enzyme immunoanalytical system based on sequential cross-flow chromatography. *Anal. Chem.*, **77**, 4091-7.

37. Ruzgas,T., Csöregi,E., Katakis,I., Kenausis,G. and Gorton,L. (1996) Preliminary investigations of an amperometric oligosaccharide dehydrogenase-based electrode for the detection of glucose and some other low molecular weight saccharides. *J Mol Recognit.*, **9**, 480-484.

38. Regalado,C., Garcia-Almandarez,B.E. and Duarte-Vazquez,M.A. (2004) Biotechnological applications of peroxidases. *Phytochem. Rev.*, **3**, 243-256.

39. Oguchi,T., Tawaki,S.-ichiro, Uyama,H. and Kobayashi,S. (1999) Soluble polyphenol. *Macromol.Rapid Commun.*, **20**, 401-403.

40. Tatsumi,K. and Wada,S. (1996) Communication to the Editor Removal of Chlorophenols from Wastewater. *Biotechnology*, **51**, 126-130.

41. Bhunia,A., Durani,S. and Wangikar,P.P. (2001) Horseradish Peroxidase Catalyzed Important Dyes. *Biotechnology*, 1-6.

42. Folkes,L.K., Candeias,L.P. and Wardman,P. (1998) Toward targeted “oxidation therapy” of cancer: peroxidase-catalysed cytotoxicity of indole-3-acetic acids. *Int. J. Radiat. Oncol., Biol., Phys.*, **42**, 917-920.

43. Kim,D.-S., Jeon,S.-E., Jeong,Y.-M., Kim,S.-Y., Kwon,S.-B. and Park,K.-C. (2006) Hydrogen peroxide is a mediator of indole-3-acetic acid/horseradish peroxidase-induced apoptosis. *FEBS letters*, **580**, 1439-46.

44. Folkes,L.K., Dennis,M.F., Stratford,M.R., Candeias,L.P. and Wardman,P. (1999) Peroxidase-catalyzed effects of indole-3-acetic acid and analogues on lipid membranes, DNA, and mammalian cells in vitro. *Biochem. Pharmacol.*, **57**, 375-82.

45. Wardman,P. (2002) Indole-3-acetic acids and horseradish peroxidase: a new prodrug/enzyme combination for targeted cancer therapy. *Curr. Pharm. Des.*, **8**, 1363-74.

46. Folkes,L.K. and Wardman,P. (2001) Oxidative activation of indole-3-acetic acids to cytotoxic species- a potential new role for plant auxins in cancer therapy. *Biochem. Pharmacol.*, **61**, 129-36.

47. Medzihradszky,K.F., Spencer,D.I.R., Sharma,S.K., Bhatia,J., Pedley,R.B., Read,D. A, Begent,R.H.J. and Chester,K. A (2004) Glycoforms obtained by expression in *Pichia pastoris* improve cancer targeting potential of a recombinant antibody-enzyme fusion protein. **14**, 27-37.

48. Kogelberg,H., Tolner,B., Sharma,S.K., Lowdell,M.W., Qureshi,U., Robson,M., Hillyer,T., Pedley,R.B., Vervecken,W., Contreras,R., *et al*. (2007) Clearance mechanism of a mannosylated antibody-enzyme fusion protein used in experimental cancer therapy. *Glycobiology*, **17**, 36-45.

49. Wagner,M. and Nicell,J.A. (2002) Detoxification of phenolic solutions with horseradish peroxidase and hydrogen peroxide. *Water Res.*, **36**, 4041-52.

50. Dalal,S. and Gupta,M.N. (2007) Treatment of phenolic wastewater by horseradish peroxidase immobilized by bioaffinity layering. *Chemosphere*, **67**, 741-7.

51. Bayramoglu,G., Altintas,B. and Yakup Arica,M. (2012) Cross-linking of horseradish peroxidase adsorbed on polycationic films: utilization for direct dye degradation. *Bioprocess Biosyst. Eng.*, 10.1007/s00449-012-0724-2.

52. Karam,J. and Nicell,J.A. (1997) Review Potential Applications of Enzymes in Waste Treatment. *J. Chem. Tech. Biotechnol.*, **69**, 141-153.

53. Auriol,M., Filali-Meknassi,Y., Adams,C.D., Tyagi,R.D., Noguerol,T.-N. and Piña,B. (2008) Removal of estrogenic activity of natural and synthetic hormones from a municipal wastewater: efficiency of horseradish peroxidase and laccase from *Trametes versicolor*. *Chemosphere*, **70**, 445-52.

54. Brady,D. and Jordaan,J. (2009) Advances in enzyme immobilisation. *Biotechnol. Lett.*, **31**, 1639-50.

55. Tischer,W. and Wedekind,F. (1999) Immobilized Enzymes : Methods and Applications. *Top. Curr. Chem.*, **200**, 96-123.

56. Dyal,A., Loos,K., Noto,M., Chang,S.W., Spagnoli,C., Shafi,K.V.P.M., Ulman,A., Cowman,M. and Gross,R. A. (2003) Activity of *Candida rugosa* lipase immobilized on gamma-Fe2O3 magnetic nanoparticles. *Journal of the American Chemical Society*, **125**, 1684-5.

57. Wang,W., Xu,Y., Wang,D.I.C. and Li,Z. (2009) Recyclable nanobiocatalyst for enantioselective sulfoxidation: facile fabrication and high performance of chloroperoxidase-coated magnetic nanoparticles with iron oxide core and polymer shell. *Journal of the American Chemical Society*, **131**, 12892-3.

58. Krenková,J. and Foret,F. (2004) Immobilized microfluidic enzymatic reactors. *Electrophoresis,* **25**, 3550-63.

59. O’Brien,A.M., O’Fagain,C., Nielsen,P.F. and Welinder,K.G. (2001) Location of Crosslinks in Chemically Stabilized Horseradish Peroxidase : Implications for Design of Crosslinks. *Biotechnol. Bioeng.*, **76**, 277-284.

60. Ryan,B.J. and O’Fágáin,C. (2008) Effects of mutations in the helix G region of horseradish peroxidase. *Biochimie,* **90**, 1414-1421, 10.1016/j.biochi.2008.05.008.

61. Wang,Q., Kromka,A., Houdkova,J., Babchenko,O., Rezek,B., Li,M., Boukherroub,R. and Szunerits,S. (2012) Nanomolar hydrogen peroxide detection using horseradish peroxidase covalently linked to undoped nanocrystalline diamond surfaces. *Langmuir : the ACS journal of surfaces and colloids*, **28**, 587-92.

62. Zhang,Q., Yang,S., Zhang,J., Zhang,L., Kang,P., Li,J., Xu,J., Zhou,H. and Song,X.-M. (2011) Fabrication of an electrochemical platform based on the self-assembly of graphene oxide-multiwall carbon nanotube nanocomposite and horseradish peroxidase: direct electrochemistry and electrocatalysis. *Nanotechnology*, **22**, 494010.

CHAPTER 1

Deletion of the *Pichia pastoris KU70* Homologue Facilitates Platform Strain Generation for Protein Expression and Synthetic Biology

Näätsaari Laura [1], Mistlberger Beate [2], Ruth Claudia [1,2], Hajek Tanja [1,2], Hartner Franz [1,2], Glieder Anton [*,1,2]

[1] Institute of Molecular Biotechnology, Graz University of Technology, Petersgasse 14, A-8010 Graz, Austria

[2] Austrian Centre of Industrial Biotechnology (ACIB GmbH), Petersgasse 14, A-8010 Graz, Austria

Manuscript submitted to PLoS ONE and accepted with revision April 2012

Parts of the study described in this chapter were initiated by and performed in co-operation with Beate Mistlberger (Pscheidt). Thus parts of the manuscript might have similarities to her Doctoral Thesis.

Abstract

Targeted gene replacement to generate knock-outs and knock-ins is a commonly used method to study the function of unknown genes. In the methylotrophic yeast *Pichia pastoris*, the importance of specific gene targeting has increased since the genome sequencing projects of the most commonly used strains have been accomplished, but rapid progress in the field has been impeded by inefficient mechanisms for accurate integration. To improve gene targeting efficiency in *P. pastoris*, we identified and deleted the *P. pastoris KU70* homologue. We observed a substantial increase in the targeting efficiency using the two commonly known and used integration loci *HIS4* and *ADE1*, reaching over 90% targeting efficiencies with only 250-bp flanking homologous DNA. Although the *ku70* deletion strain was noted to be more sensitive to UV rays than the corresponding wild-type strain, no lethality, severe growth retardation or loss of gene copy numbers could be detected during repetitive rounds of cultivation and induction of heterologous protein production. Furthermore, we demonstrated the use of the *ku70* deletion strain for fast and simple screening of genes in the search of new auxotrophic markers by targeting dihydroxyacetone synthase and glycerol kinase genes. Precise knock-out strains for the well-known *P. pastoris AOX1*, *ARG4* and *HIS4* genes and a whole series of expression vectors were generated based on the wild-type platform strain, providing a broad spectrum of precise tools for both intracellular and secreted production of heterologous proteins utilizing various selection markers and integration strategies for targeted or random integration of single and multiple genes. The simplicity of targeted integration in the *ku70* deletion strain will further support protein production strain generation and synthetic biology using *P. pastoris* strains as platform hosts.

Introduction

The methylotrophic yeast *Komagataella pastoris*, commonly known as *Pichia pastoris* has become one of the major eukaryotic hosts for recombinant protein production, mainly because of its strong and tightly regulated *AOX1* promoter (1), ease of manipulation, growth to high cell-densities in inexpensive media and ability to perform complex post-translational modifications (2). The genome sequence of the *P. pastoris* histidine auxotrophic variant GS115 has been published with a curated annotation for its 5313 protein coding genes (3). This most commonly used and commercially available strain had been derived by mutagenesis of the *P. pastoris* WT strain NRRL-Y11430 (ATCC 76273), which was also deposited in the Netherlands as *P. pastoris* CBS7435. More recently, the nuclear and mitochondrial genome sequences of the unmodified WT strain CBS7435 were published (4). Compared to previous sequences, most gaps were closed in the new genome sequence and some sequence and annotation mistakes were corrected. This new information shed light on the formerly poorly known, but important pathways that are responsible for methanol utilization, secretion, glycosylation, proteolytic processing and protein folding. Therefore, this information has increased the potential of *P. pastoris* and enabled its further development towards a customized and highly efficient host for heterologous protein production.

Due to the lack of stable plasmid systems for *P. pastoris*, gene expression cassettes are usually integrated into the genome. Stable integration of entire expression cassettes into the *P. pastoris* genome is based on homologous recombination (HR) and non-homologous end joining (NHEJ). HR dominant in *Saccharomyces cerevisiae* is an accurate pathway that repairs double-strand breaks (DSBs) by using the information from homologous sequences (5, 6). The far less specific process of NHEJ, more dominant in filamentous fungi and higher eukaryotic organisms, is not necessarily accurate and deletions of a few nucleotides are often introduced at DSB-sites (6–11). Gene replacement events in *P. pastoris* have been estimated to occur with a frequency of <0.1% when the total length of targeting fragments is <500bp (12) and with a frequency of 10-20% (13) or up to 30% (12) when extensive ~1kb regions of homology are used. However, in previous projects we have observed that success also depends on the genomic locus. NHEJ seems to have a substantial role in the integration events of expression cassettes with flanking homologous sequences and thus limits the efficient generation of specific changes like targeted knock-outs.

In NHEJ, double-strand breaks are recognized by the highly conserved and arguably process defining Ku70p/Ku80p heterodimer, which specifically binds to DNA ends and forms a complex with the DNA-dependent protein kinase catalytic subunit (14–17). This protein kinase, in a complex with a co-factor-like protein Xrcc4 and other auxiliary factors is known to stimulate end processing, followed by ligation by DNA ligase IV (7, 18). Studies in both yeasts and filamentous fungi revealed that by deleting components of the NHEJ-pathway, the random integration of DNA fragments is strongly reduced (19, 20). Therefore, in those NHEJ-pathway defective mutants, DNA integrates mainly via the HR-pathway giving rise to high homologous recombination frequencies. *Mus musculus* and *S. cerevisiae* cells that are deficient in the Ku80p counterpart of the Ku70p/Ku80p heterodimer have been reported to display telomeric shortening, be radiosensitive and exhibit V(D)J recombination defects and high levels of chromosomal aberrations (21–25). However, opposite results also exist (26) suggesting that negative effects of deletions in the NHEJ pathway could be dependent on specific species and culture conditions used. In *S. cerevisiae*, mutations in the Ku70p counterpart of the Ku70p/Ku80p heterodimer have been shown to greatly impair NHEJ without perturbing telomeric functions (27). Publications in this research field existed regarding CHO-cells (28), fungi like *Aspergillus* sp. (29–32), *Magnaporthe grisea* (33), *Neurospora* sp. (9) and yeasts like *Kluyveromyces lactis* (34), *Candida glabrata* (26), and *Saccharomyces cerevisiae* (9), but there was no published information regarding *P. pastoris* to date.

The objective of this study was to show that elimination of the normal function of Ku70p, a conserved DNA end-binding protein essential for NHEJ, significantly increases the homologous recombination efficiency and thus strongly reduces the problematic random integration of DNA fragments in *P. pastoris*. In order to evaluate possible advantages of a *ku70* mutant strain for the quick identification of target genes to generate new auxotrophic strains and selection systems, we targeted three significant proteins of carbon source utilisation pathways. *Dihydroxyacetone synthase 1*, *dihydroxyacetone synthase 2* and *glycerol kinase 1* from the methanol and glycerol assimilation pathways (Figure 1) were the first targets for deletion. In addition, the well-known auxotrophy marker genes *HIS4* and *ADE1* were targeted to characterize the effects of the *KU70* deletion and to generate genetically well-defined auxotrophic strains by precise single gene deletions instead of the commonly applied random mutagenesis of the whole genome. Finally, a new and well defined *P. pastoris* expression platform was generated based on the wild-type strain CBS7435 including complementary *E. coli*/*P. pastoris* shuttle vectors for protein expression.

Materials and Methods

Strains and culture conditions

All *P. pastoris* strains used and constructed during this study are based on the wild-type strain CBS7435 (NRRL-Y11430, ATCC 76273) and described in more detail in Table 5. Recombinant DNA manipulations were performed in *E. coli* strains DH5alpha and TOP10F' (Invitrogen Corp., Carlsbad, CA) according to standard protocols (35). All *P. pastoris* and *E. coli* strains were cultured in standard media using chemicals and other components as previously described by (36). Amino acids were added to 40µg/ml (histidine) or 50µg/ml (arginine), Ampicillin to 100µg/ml (*E. coli*) and Zeocin™ (Invitrogen) to 100µg/ml (*P. pastoris)* or 25µg/ml (*E. coli*) as required. *P. pastoris* competent cells were prepared with the condensed protocol and transformations were performed by electroporation, essentially as described by (37). Before plating out aliquots on corresponding selective media, the cells (80µl) were allowed to regenerate for two hours at 28°C with 500µl 1M sorbitol and 500µl YPD. To study the effect of knock-outs in the methanol assimilation, glycerol assimilation, purine biosynthesis and amino acid synthesis pathways, target genes were disrupted with a Zeocin™ resistance cassette surrounded by two homologous ends for HR in the target locus. All *P. pastoris* platform strains constructed during this study were generated with specific full or partial knock-out of the target gene using a flipper cassette described below and similar to the construct previously described by (38).

Growth rate studies

Growth rates of the *P. pastoris* wild-type strain CBS7435 and knock-out strains *aox1, aox1 arg4, his4, aox1 his4, ku70, ku70 gut1, das1, das1 das2* and the corresponding complemented *arg4, his4* and *gut1* mutant strains were defined by measuring optical density (OD_{595}) of triplicate cultures during exponential growth phase. An overnight culture was inoculated with a single colony in corresponding media and diluted to a starting OD of ~0.15. The main cultures were grown in standard conditions using 50ml of buffered minimal media BMD2% (D-glucose), BMG1% (glycerol) or BMM0.5% (methanol) in 250ml baffled shake flasks (28°C, 120rpm with 50mm amplitude). Amino acids were added to 40µg/ml (histidine) or 50µg/ml (arginine) as required for non-complemented auxotrophic strains.

Sequence analysis and optimization

Nucleotide sequence data were primarily obtained from the public database NCBI (www.ncbi.nih.gov). *P. pastoris* genes *glycerol kinase 1* (*GUT1*, FR839631 region 302996-304861), *dihydroxyacetone synthase 1* (*DAS1*, *FR839630* region 634689-636812), *dihydroxyacetone synthase 2* (*DAS2*, *FR839630* region 630077-632200), the trifunctional *HIS4 (phosphoribosyl-ATP pyrophosphohydrolase, phosphoribosyl-AMP cyclohydrolase, and histidinol dehydrogenase,* U14126.1), *argininosuccinate lyase 4* (*ARG4*, AF321097.1) and *KU70* (FR839630, region 1598101-1599963, XM_002492501.1) were identified either by their annotation or by a blastx search of the genome sequence of *P. pastoris* CBS7435. Other sequences of the *DAS1* and *DAS2* genes in public databases contained mistakes due to the wrong assembly of raw data caused by the high sequence similarity of these genes (4). Excision cassette constructs and plasmids were designed using VectorNTI (Invitrogen, Carlsbad, CA, USA). Primers were designed manually and analyzed with EditSeq (DNASTAR, Madison, WI, USA). ClustalW (39) was used for the pairwise and multiple alignments of known sequences and the pI/Mw tool from ExPASy Proteomics Server was used for calculating the isoelectric points and molecular weights of proteins. To confirm the expected knock-outs and plasmid compositions, the sequences obtained from Sanger sequencing (LGC Genomics, Berlin, Germany) were assembled using SeqMan (DNASTAR). Synthetic genes were codon compromised according to a combined average codon usage of *Pichia pastoris*, *Yarrowia lipolytica* and *Schizosaccharomyces pombe* using GeneDesigner (DNA 2.0, Menlo Park, CA, USA, 40). This strategy avoided very rare codons for any of these three hosts to provide vector elements for a broad host spectrum.

Excision cassettes and gene disruption

The structure of all excision cassettes used to create the new platform strains is, in principle, as described by (38), wherein the FLP recombinase system (41) is utilized to enable excision cassette removal. A marker free knock-out strain is created with one 34bp FLP recombinase recombination target sequence (*FRT*, GAAGTTCCTATACTTTCTAGAGAATAGGAACTTC) left in the locus. All cassette components used were either amplified from wild-type strains (*P. pastoris* CBS7435, *S. cerevisiae* BY4741) or synthetic fragments. The *P. pastoris AOX1* promoter was chosen to drive the regulated expression of the *S. cerevisiae FLP* recombinase terminated by the *S. cerevisiae CYC1* terminator. A Zeocin™ resistance cassette was amplified from pPpT2 (36). The above-mentioned parts were surrounded on both sides by identical 34bp *FRT*s placed in direct orientation. The outermost parts of each excision cassette, namely the 5’ and 3’ integration sequences, were locus-specific to guarantee a knock-out only in the target region. All parts were amplified and joined by standard overlap-extension PCR using HPLC-purified primers (Eurogentec, Seraing, Belgium) listed in Table S1 and Phusion^{TM} High-Fidelity DNA-polymerase (Finnzymes Oy, Espoo, Finland). For the overlap-extension PCRs, equimolar amounts of each fragment were used. In the case of the *AOX1* knock-out cassette, the *AOX1* promoter was placed 5’ of the first *FRT* to simultaneously function as a promoter for the FLP recombinase and as a 5’ integration sequence. The length of the integration sequences and the extent of the knock-outs (described in Table 1) were dependent on the sequence information available at the time of cassette design. To ensure the inactivation of the *KU70* gene, the start codon ATG was also modified to ATAC. An example of the cassette structure is illustrated in Figure 2a. Gene disruption cassettes to study both homologous recombination in the *ADE1* and *HIS4* loci of the *ku70* deletion strain and the effect of knock-outs in glycerol and methanol assimilation pathways were constructed by overlap-extension PCR joining the Zeocin™ resistance cassette with 5' and 3' locus-specific sequences for targeted homologous recombination (Figure 2b). The lengths of the homologous sequences used are depicted in Table 2. Figure 2b illustrates the method used to achieve gene inactivation in *HIS4* locus despite the extent of homologous integration

The overlap-extension PCR products were purified using Wizard® SV Gel and PCR Clean-Up System (Promega, Madison, WI, USA) and cloned into the pJET vector using CloneJET kit (Fermentas). The correct cassette structure was verified by sequencing followed by cassette amplification, purification and transformation into appropriate *P. pastoris* strains. Only a low amount of DNA (~500ng/80µl cells) was used to guarantee the prevalence of single-copy integration. Positive transformants were grown in 96-well deep-well plates (42) for 2.5 days and pinned onto either minimal methanol (*AOX1* knock-out) or minimal dextrose (*ARG4* and *HIS4* knock-outs) plates to distinguish between the strains with normal and slow-growth/no-growth phenotypes. The frequency of correct integration into the *ADE1* locus was defined by observing the transformation plates and calculating red colonies resulting from the accumulation of biosynthetic intermediate phosphoribosyl aminoimidazole. Due to a significant difference in the growth rates of the *ade1* and wild-type strains, the plating out of transformants was performed only after significant dilution, with the aim of observing approximately 10 colonies on each plate. Methanol induction to initiate the production of the FLP recombinase and thus achieve cassette excision was performed by two consecutive rounds of cultivation on minimal methanol plates. After a total of 5 days, single colonies were picked from the methanol plates and streaked out on YPD- Zeocin™ plates to identify the clones that had excised the cassette and were Zeocin™-sensitive.

Plasmid constructions

The sequences of the plasmids in our *P. pastoris* expression platform were deposited in the GenBank database at NCBI. The plasmids pPpB1_S (JQ519685), pPpB1GAP (JQ519686), pPpB1GAP_S (JQ519687), pPpB1_Alpha_S (JQ519688), pPpT4 (JQ519689), pPpT4_S (JQ519690), pPpT4_Alpha_S (JQ519691), pPpT4GAP_S (JQ519692), pPpT4GAP_Alpha_S (JQ519693), pPpKan_S (JQ519694) and pPpKan_Alpha_S (JQ519695) were constructed as described previously (36, 43, 44). Primers and the components and origins of components used in the plasmid construction are described in Tables S1 and S2. The vector maps are elucidated in Figure S1.

The complementation plasmids pPpARG4 (JQ519696) and pPpHIS4 (JQ519697) were constructed as follows. The codon usages of the coding sequences (cds) of both synthetic marker genes were designed to be the combined average codon usage of *P. pastoris*, *Yarrowia lipolytica* and *Schizosaccharomyces pombe.* Codons with an appearance of less than 8% in any of the above organisms were excluded. The expression unit, including the *AOX1* promoter, multiple cloning site and *AOX1* terminator, was amplified from pPpT2 (36). The *BLA* cds and pUC origin of replication were amplified from pUC8. The synthetic prokaryotic *EM72* promoter was constructed using HPLC purified primers. For pPpARG4, the natural promoter and terminator sequences were amplified from the wt genomic DNA. For the plasmid pPpHIS4, *ADH1* promoter and *TIF51A* terminator were chosen to be used. Both parts were amplified from *S. cerevisiae* wt genomic DNA. All parts were joined by overlap-extension PCR before the linear cassette was digested by XbaI (Fermentas) and circularized using T4 ligase (Fermentas).

A glycerol kinase complementation plasmid pPpGUT1 (JQ519698) was constructed using both pPpT2 for *AOX1* promoter driven expression and a 3275bp fragment from the glycerol kinase locus that included the natural promoter, cds and terminator for selection in the *P. pastoris gut1* deletion strain. This vector was designed as a replacement vector and therefore contained 5´and 3´homologous regions of the *GUT1* locus for site specific integration and replacement of the antibiotic marker in the *gut1* deletion strain.

Methods used in the characterization of the strains and plasmids

Genomic DNA was isolated from each strain using DNAeasy kit (Invitrogen Corp., Carlsbad, CA) according to the manufacturer’s protocol for large scale yeast DNA isolation. All targeted loci in the *P. pastoris* genome were PCR amplified and sequenced to confirm expected knock-outs using primers described in Table S1 and Phusion^{TM} High-Fidelity DNA-polymerase according to manufacturer’s recommendations. Southern blotting and hybridization were carried out according to standard protocols (45) to verify the correct integration of the excision cassettes. PCR amplification of the probes specific to the Zeocin™ gene and knock-out regions in genes *AOX1*, *HIS4* and *KU70* was performed using DIG-labeled dNTPs (Roche, Basel, Switzerland) and AmpliTaqGOLD® (Roche, Basel, Switzerland) polymerase according to the manufacturers’ instructions. Primer details can be found in Table S1.

Correct plasmid sequences were confirmed by Sanger sequencing. For testing the functionality, an improved version of green fluorescent protein (46) was cloned into the multiple cloning site of each plasmid. Correct insertion was verified by sequencing. Each plasmid was amplified in *E. coli*, linearized with BglII or SmiI (Fermentas) and transformed into the corresponding *P. pastoris* strains. Single colonies were transferred to 96-well deep-well plates for standard cultivation as described previously (42). To compare the expression levels, GFP fluorescence was measured as described by (47).

Genome walking experiments

In addition to the expected locus, gene disruption cassettes might also integrate at other loci of the genome. Genome walking was used to define the integration sites in strains where targeting of the cassettes seemed to be random. Therefore 2µg of genomic DNA of each strain was singly digested with BamHI (a), EcoRI (b) and HindIII (c) in order to get fragments of 1 - 5 kb size. The digestion was stopped by heat inactivation, the fragmented DNA precipitated with ethanol and the pellet dissolved in 30µl of distilled water.

An adaptor fragment was created by annealing adaptor strand 1 (5'-GTAATACGACTCACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGT-3') either to adaptor strand 2.a (3'-TCCCCGACCACTAG-5') for BamHI digested DNA, 2.b (3'-TCCCCGACCATTAA-5') for EcoRI digested DNA or 2.c (3'-TCCCCGACCATCGA-5') for HindIII digested DNA.

In the annealing reaction, adaptor strand 1 (100µM, 13,7µl) was mixed in a 1:1 molar ratio with adaptor strand 2.a, 2.b or 2.c (100µM, 4µl) and denatured in 95 °C for 5 min. The mixture was allowed to slowly cool down to room temperature. The three differently annealed adaptors 1+2a (BamHI), 1+2b (EcoRI) and 1+2c (HindIII) were ligated for 3 h at room temperature with T4 DNA ligase to the digested DNA fragments, considering the specific 5' overhangs that had been created by the restriction enzymes used. The ligation reaction was stopped by incubation for 5 min at 70°C and diluted with 70µl TE buffer.

Two gene-specific primers and two adaptor primers were designed (Table S1). The gene-specific primers were designed to bind approximately 100bp away from the end of the known sequence, considering that no restriction site of the restriction enzymes used for gDNA digestion lay between the primer-binding site and the end of the known sequence. Adaptor primer 1 and gene specific primer 1 were used as a primer pair for a first PCR with 1µl of the gDNA-adaptor ligation product as template DNA. 1µl of the first PCR mix was used as the template for a second PCR with adaptor primer 2 and gene specific primer 2. This second primer pair was designed to bind within the first PCR product. Both PCR steps were performed with an annealing temperature of 58°C and an elongation time of 50 seconds. A gene-specific DNA fragment as a product from the second PCR was isolated from a preparative agarose gel, purified and sent to Sanger sequencing (LGC Genomics GmbH, Berlin, Germany) using adaptor primer 2 and the corresponding gene specific primer 2.

Genetic stability tests

To test the stability of multiple expression cassettes in the wild-type and *ku70* deletion strains during repetitive phases of growth on glucose and methanol, the wt strain CBS7435 and deletion strains CBS7435 mut^s and CBS7435 *ku70* were transformed with 3µg of SmiI- linearized plasmid pPpT4_S_GFP. The resulting colonies were screened for GFP activity as described by (47) and the best expressing clones were preserved in glycerol stocks. Duplicates of each strain were inoculated to a starting OD of 0.55 in 50ml BMD1% in 250ml baffled shaking flasks. The strains were cultivated for 48h to reach a comparable cell density (28°C, 120rpm with 25mm amplitude), followed by a methanol induction phase for 72h (42). The growth and induction cycles were repeated for four times, always using the induced culture as a starting material for the next round. DNA was extracted from the starting strains and the strains after four rounds of methanol induction as described previously (48). Copy numbers were detected with qRT-PCR as described by (44). Each qRT-PCR was performed at least twice by independent researchers. Strains displaying a standard deviation of over one copy number between independent duplicates were omitted from the analysis.

Spread titer tests

To measure the sensitivity of the *ku70* deletion strain to UV light, a spread titer test was performed as follows. Overnight cultures of four biological replicates of each wt and *ku70* deletion strain cells grown in YPD were harvested and washed. The OD was adjusted to 1 prior to plating out 50µl and 100µl aliquots of 10^{-4} and 10^{-5} dilutions respectively on YPD agar. The plates were exposed to 0, 50 and 100J/m^2 UV rays (Bio-Link 254 UV crosslinker, Vilber Lourmat, Marne-la-Vallée Cedex, France), and incubated in the dark for 60h before calculating the amount of colony forming units on exposed and unexposed plates.

Results

Identification of the target genes *KU70, KU80, GUT1, DAS1* and *DAS2* in the wild-type *P. pastoris* CBS7435 genome

Non-homologous end-joining is one of the main pathways repairing broken genomic DNA in eukaryotes. One of the aims of our study was to investigate if the NHEJ pathway plays an important role in the site targeted specific integration in the genome of *Pichia pastoris* (NRRL-Y 11430, ATCC 76273) and if the NHEJ mechanism can be impaired by knock-out mutagenesis. Homologues to the nucleotide sequences of the *Saccharomyces cerevisiae* genes *HDF1* (NM_001182791.1) and *HDF2* (NM_001182606.1), corresponding to *KU70* and *KU80* respectively, were identified in the *P. pastoris* wild-type genome by blastx. The *KU70* homologue (FR839630, region 1598101-1599963) in chromosome 3 shared 28% sequence identity to *HDF1* and the *KU80* homologue (FR839631.1 Region 361953-362288, 362551-364041) in chromosome 4 showed 19% sequence identity to *HDF2* (both defined by pairwise alignment using ClustalW). Both genes had been annotated as subunits of the telomeric Ku complex (yeast Ku70p/Ku80p) by the automated annotation of the assembled genome. Remarkably, the sizes of the Ku70 and Ku80 proteins, compared to the corresponding proteins of *S. cerevisiae, H. sapiens* and *A. thaliana*, would better cohere with contrary annotation. The *P. pastoris KU70* and *KU80* cDNAs encode proteins of 620 and 608 amino acids with predicted molecular masses of 71.3 and 69.5kD. The protein sizes are similar to those of *S. cerevisiae* (602aa and 629aa, 70.6 and 71.2kDa), *H. sapiens* (P12956 609aa and P13010 732aa, 69.8 and 82.7kDa) and *A. thaliana* (Q9FQ08 621aa and Q9FQ09 680aa, 70.3 and 76.7kDa). However, the knock-out of either subunit has been reported to result in reciprocal down-regulation of the other subunit (49), thus making a confirmation of correct annotation by function troublesome. The sequence identities to already known Ku proteins are generally low. The *P. pastoris* Ku70 and Ku80 proteins share only 28% & 19%, 18% & 16% %, and 15% & 16% sequence identity with the proteins of *S. cerevisiae*, *H. sapiens* and *A. thaliana,* respectively.

For a first feasibility test to generate well defined auxotrophic strains based on the *P. pastoris* wt strain CBS7435, the nucleotide sequences of the three genes g*lycerol kinase* 1 (*GUT1*, FR839631 region 302996-304861, 621aa), *dihydroxyacetone synthase 1* (*DAS1*, *FR839630* region 634689-636812, 707aa) and *dihydroxyacetone synthase 2* (*DAS2*, *FR839630* region 630077-632200, 707aa) from the glycerol and methanol assimilation pathways were elucidated.

Construction and molecular characterization of a *P. pastoris ku70* deletion strain

The *KU70* flipper knock-out cassette was constructed with 1299bp 5’ and 1170bp 3’ flanking sequences from the *KU70* locus. Both integration sequences were placed partially on top of the coding sequence, so that the deletion caused by the knock-out cassette integration would be only 215bp. This short deletion leading to inactivation of the Ku70 protein is optimal for complementation of the locus if needed. The 6747bp assembled linear PCR product was successfully used to transform the *P. pastoris* wild-type strain CBS7435. Control PCRs were performed to confirm correct cassette integration, excision and expected knock-out in the cds of *KU70*. In the first PCR reaction, both primers (Table S1) were located outside the target region. This reaction functioned as a positive control, PCR amplifying ~2.7kb, ~2.5kb and ~6.7kb fragments from the wt, the *ku70* deletion strain after flipper excision and the *ku70* deletion strain before flipper excision. The second primer pair served as a negative control PCR, since one of the primers bound in the region knocked out in *ku70 deletion* strains. A ~1.3kb product was expected from the wt strain and any other strain having the excision cassette integrated somewhere other than in the *KU70* locus of the genome. One strain giving expected PCR product sizes from all control PCRs was chosen for further tests. High quality genomic DNA was isolated from the wt starting strain, the original transformant strain before flipper excision and the final *ku70 deletion* strain after flipper excision. The purified genomic DNA was subjected to southern blot analysis after two separate digestions with restriction enzymes BamHI and BglII. When using a probe specific to the knock-out region in the *KU70* locus, hybridizing fragments of expected sizes (~4.4kb for BamHI and ~4.1kb for BglII) were observed in the wild-type strain. Fragments were missing in both *ku70 deletion* strains (before and after cassette excision) verifying correct cassette integration and successful knock-out. To rule out knock-outs elsewhere in the genome, single-copy integration of the excision cassette in the expected locus was tested by hybridizing a probe specific to the Zeocin™ resistance cassette to both BamHI and BglII digested genomic DNA of the before mentioned strains. Fragments of the expected size (8.4kb BamHI, 4.8kb BglII) could be observed in the strain still carrying the resistance cassette before induction, suggesting single-copy integration in the expected locus. No bands were observed either from the wt strain or *ku70* knock-out strains (Figure S2). Furthermore a 1.4kb fragment of the deletion site was amplified from the genomic DNA of the *ku70 knock-out* strain and sequenced. This sequence (Figure S3) verified a correct gene deletion and the expected FRT sequence left in the locus.

The behaviour of the *ku70* deletion strain in heterologous protein production was tested with SmiI linearized pPpT4_GFP plasmid in comparison to the corresponding wt strain. The fluorescence intensity landscapes resulting from the screening of two 96-well plates per strain were noted to be more uniform for the *ku70* deletion strain (data not shown). No increased amount of inactive clones could be detected.

Increased targeting efficiency employing the *ku70* deletion strain

To compare the gene targeting efficiency of the *ku70* knock-out and wt strains, the loci of two biosynthetic genes *ADE1* and *HIS4* were chosen due to the reliable and simple detection of corresponding auxotrophies. *P. pastoris* colonies lacking *ADE1* develop red color when grown without adenine supplementation and *his4* knock-out colonies can be detected by replica plating on minimal media without histidine supplementation.

The excision cassettes to disrupt *ADE1* and *HIS4* genes were successfully constructed by overlap-extension PCR as described previously. The cassettes were planned to disrupt the target gene despite the fact that the excision cassette might recombine using either 5' or 3' integration sequence only. Increasing the length of homologous sequences has been reported to have a positive effect on the homologous recombination frequency in *N. crassa* (19). 1 kbp has been suggested to be the minimum length of 5' and 3' homologous regions required for 100% and >85% site-specific recombination and replacement of target genes in *ku70* deletion mutants of *N. crassa* and *S. macrospora* respectively (19, 20). Also, in *P. Pastoris*, longer homologous sequences are important for efficient integration into the genome (12). Therefore, we chose homologous flanking sequence lengths varying from 1350bp down to only 100bp to examine the recombination frequency in the *HIS4* locus of *P. pastoris*. The same study was done for the *ADE1* locus. However, since the homologous recombination frequencies for the *ku70* deletion strain with flanking sequence lengths of 1350bp down to 650bp were still about 100% in the *HIS4* locus, a set of shorter flanking sequences varying from 650bp to 50bp was chosen to point out strain specific differences for the *ADE1* locus. The homologous recombination frequencies in the wt and *ku70 knock-out* strains are depicted in Table 2. In contrast to the wt strain, almost all antibiotic resistant transformants of the *ku70 deletion* strain showed the expected phenotype – even when short homologous sequences of 150-250 bases were used for integration. This observation indicated correct site-specific integration for the majority of the *ku70* knock-out strain transformants whereas most of the constructs integrated at different sites in the case of the wt strain. Homologous recombination efficiency in the *ku70* deletion strain dropped to the level of the wt with integration sequence lengths as short as 50-100bp. The potential artefacts caused by variations between transformations and radically different growth rates of wt and *ade1* strains were minimized by repeating every transformation for four times and plating small aliquots of diluted cells. However, the possibility of wt colonies overgrowing *ade1* colonies and thus affecting the calculation of homologous recombination frequencies cannot be ruled out, especially for the wt host strain with almost no homologous recombination in the *ADE1* locus.

A majority of the cassettes targeted to either *HIS4* or *ADE1* loci of the wild-type strain did not result in auxotrophy. Therefore the integration loci of altogether 18 wt and mut^s strains were analysed by genome walking in order to verify the assumption that integration happens randomly in these strains. Although very low amounts of DNA were used for transformation, and thus most of the colonies were expected to be single-copy transformants, multi-copy integration could not be ruled out. Therefore only 14 strains delivering a strong main product in the nested PCR of genome walking and high sequence quality were included in the final analysis depicted in Table 3. In addition, the integration loci of four strains were verified by digestion with two different restriction enzymes.

Cassette integration was detected in every chromosome and, for most of the strains, seemed to be random. No micro-homologies (~10bp) at the exact site of integration were detected by sequencing the 5' end of the disruption cassette. Interestingly, the integration loci of the strains #4 and #11 in chromosome 4 were separated by only ~3.5kbp. In addition, the integration locus of strains #2 and #8 in chromosome 4 was identical. No explanation of these facts could be detected by sequence analysis. However, the identical integration locus of strains #2 and #8 could also be explained by a possible clone identity due to duplication of the transformant during the regeneration phase after electroporation. After transformation, the cells were regenerated in sorbitol/YPD for approximately 2 hours, which theoretically allows one cell division.

In NHEJ, the possibility of inaccuracy and introduction of small deletions at the DSB-sites has been reported (6–11). Also, in two of the 14 strains analyzed in this study, a single nucleotide deletion in the 5' end of the disruption cassette was detected. However, considering the fact that only 5' sequences were determined by genome walking and the relatively low number of analyzed random integrants, no reliable determination of the total frequency of deletions at the DSB sites can be made.

Application of the *P. pastoris ku70* deletion strain as a tool to screen targets for the generation of auxotrophic strains

Low rates of site-specific integration in the expected gene locus make it difficult to study the effect of site-specific integration or knock-out variants in the wt strains. Many transformants have to be analyzed on a molecular level (e.g. by colony PCR, sequencing, southern blot) since many of the selective markers integrate at other sites. For hard to access loci most of the transformants show integration of the selection marker at different sites (as for *HIS4* and *ADE1* described above), which can lead to the misinterpretation that the targeted gene might be essential. The use of the *P. pastoris ku70* deletion strain offers a possibility to reduce the number of positive transformants showing integration of the selection marker without disrupting the target gene and therefore facilitates and speeds up knock-out strain identification. After quick identification of a suitable locus, this can still be repeated with the wt strain or alternatively the *KU70* deletion can be complemented again. Since all clones where the *HIS4* and *ADE1* loci were targeted displayed the expected auxotrophic phenotype, we took advantage of the high rate of site-specific integration in the *ku70* deletion strain for a quick and simple evaluation of the feasibility for a few selected genes to serve as targets to generate new auxotrophic strains.

For the search of possible new *P. pastoris* auxotrophy strains, knock-out cassettes to disrupt *GUT1*, *DAS1* and *DAS2* were constructed and used for successful transformation of the *P. pastoris ku70* knock-out strain. Since the knock-out cassettes contain long (500-1000bp) homologous integration sequences, all eight clones tested had the cassette integrated in the right locus, resulting in the expected gene disruptions. Specific inactivation (partial replacement) of the *glycerol kinase (GUT1)* gene resulted in a *P. pastoris* strain that was observed to have abolished growth on glycerol as the sole carbon source. Initial experiments proved the possibility of complementing the auxotrophy with plasmid pPpGUT1 (Table 5) and of selecting corresponding transformants on minimal glycerol medium. Even though a disruption of the genes *DAS1* and *DAS2* lead to phenotypes with reduced growth on methanol (Table 5), the ability to grow on minimal media agar plates with methanol as the sole carbon source was not totally abolished, leading to too high background growth for use as a selection marker. We speculate that another enzyme, like transketolase *TKL1*, with similarity to the known dihydroxyacetone synthases could, to some extent, take over their role in the peroxisomes.

Comparative genetic stability of transformants

According to (50), the genetic stability of *P. pastoris,* strains carrying multiple (≥12) copies of an expression cassette is dependent on conditions which induce target gene expression. However, low-copy transformants (1-6 copies) were shown to exhibit high stability regardless of whether induced or not. To rule out that changes in the normal repair mechanisms of the *ku70* deletion strain could cause significantly increased genetic instability of expression strains, a set of *P. pastoris* transformants carrying four to seven copies of P$_{AOX}$-GFP (plasmid pPpT4_S) were employed to investigate the genetic stability of the *ku70* deletion strains during methanol induction. No changes in the copy numbers could be detected even after four 72h rounds of methanol induction, corresponding to a total cultivation time of 480h and a total induction time of 288h (over 100 generations) in non-selective media (the results are described in Table S3).

Susceptibility of the *ku70* deletion strain to DNA damage induced by UV light

The Ku70p/Ku80p heterodimer has been suggested to be involved in DNA-repair processes (14). To measure the sensitivity of the *P. pastoris ku70* deletion strain to UV light, a spread titer test was performed with four biological replicates and 0, 50 and 100J/m^2 exposure. The survival rates determined by calculating the amount of colony forming units after 60h incubation were 58% (±6,9%) for the wt strain and 48% (±6,8%) for the *ku70* deletion strain with 50J/m^2 exposure. The corresponding values for wt and *ku70* deletion strains when using an exposure of 100J/m^2 were 37% (±5,6%) and 27% (±5,6%), respectively, confirming the expected reduced survival rate.

New *P. pastoris* MutS, *arg4* deletion and *his4* deletion platform strains and complementary *E. coli*/*P. pastoris* shuttle vectors

In order to complete the new well characterized toolbox for protein expression based on the *P. pastoris* wt strain, a set of new platform strains was generated based on precise knock-outs in the wt strain CBS7435, thereby reducing the risk of undesired additional mutations in the genome that are commonly observed during traditional random mutagenesis approaches. Since negative effects of the *KU70* deletion on strain stability and heterologous protein production cannot be excluded yet, the platform strains were made on the basis of the non-mutagenized wt strain rather than on the Δ*KU70* variant. All excision cassettes (flipper-cassettes) were constructed with overlap-extension PCR having 5' and 3' flanking regions from *AOX1* (940bp and 1143bp, complete cds knock-out), *ARG4* (963bp and 1502bp, partial cds knock-out) and *HIS4* (844 and 882bp, partial cds knock-out) loci for targeted integration. Wild-type *P. pastoris* was successfully transformed with the respective 5302bp, 6761bp and 5985bp amplification products. Primary selection with Zeocin^{TM} ensured the integration of at least one copy of the cassette in the genome. Transformants were cultivated in 250µl YPD until reaching stationary phase, followed by stamping on both YPD and MM or MD plates. Normal growth on YPD combined with slow growth on MM (*aox1* knock-out strain) or no growth on MD (*arg4*, *his4 knock-out strains*) indicated cassette integration in the target locus. In the *AOX1*, *ARG4* and *HIS4* loci, approximately 2%, 3% and 1% of the excision cassettes were integrated in the target locus resulting in mut^s, *arg4* auxotrophic and *his4* auxotrophic phenotypes respectively. Variation in the integration rates might, for example, be caused by differences in the lengths of the homologous sequences used for integration, excision cassette sizes and secondary structures present in the target loci. Since very low amounts of DNA were used in the transformations, most of the positive transformants were expected to have only a single copy of the selection cassette integrated in the genome.

Flipper cassettes should stay integrated in the genome as long as the tightly controllable promoter chosen for *FLP* is not activated, and thus no FLP recombinase is produced. In the case of the *AOX1* promoter, activation is attained by adding methanol to culture media that contain no repressing carbon source such as glucose or glycerol (51). Theoretically the number of flipper-cassettes in every transformant can vary and every cell has an individual chance of cassette excision. Therefore, we produced clean single-colony streak-outs from each positive transformant colony and initially induced the promoter only for a short time until the first strains had excised the cassette. This approach should increase the chance to obtain Zeocin™ sensitive cells where only one selection marker was integrated and excised. Original Zeocin™ resistant colonies were streaked out on minimal methanol plates. After two days, no Zeocin™ sensitive colonies were found, indicating insufficient amounts of FLP had been produced to locate the flipper cassette and perform cassette excision. Therefore, the population was further streaked out on fresh minimal methanol plates and incubated for 3 days. In the case of the *AOX1* promoter in the *AOX1* excision cassette, a five-day induction with methanol on MM plates resulted in the excision of the cassette in ~7% (16/228) of the single colonies tested, thus rendering the cells Zeocin™ sensitive.

To verify correct cassette integration and excision, high quality genomic DNA was extracted from the wild-type strain and from all knock-out strains before and after cassette excision. Target loci were sequenced using primers located both outside and inside the integration sequences (Table S1). The lengths and sequences of the PCR products with all primer combinations were as expected. The sequences of the targeted sites (Figure S3) confirmed the expected total (*aox1* deletion strain) or partial (*his4, aox1his4, ku70his4, aox1arg4* deletion strains) deletions. Only one *FRT* sequence was left in the genome locus.

DNA samples from the *aox1* deletion strain and the auxotrophic strains before and after knock-out cassette excision were subjected to southern blot analysis to verify the expected knock-out and to define the location and number of excision cassettes integrated in the genome. The results and expected fragment sizes are summarized in Table S4. Wild-type strain CBS7435 was used as a control. *Aox1*, *his4, ku70his4* and *aox1his4* deletion strains were digested with at least two separate restriction enzymes each: NdeI and SspI for the AOX1 locus, and DraI and BglII for the HIS4 locus. Every probe targeted specifically to either *AOX1* or *HIS4* wild-type locus showed a band of expected size with the wt control strain CBS7435. As expected, the corresponding fragments were not present in the before and after induction strains with the wild-type locus excised. When using a probe targeted to the Zeocin™ resistance cassette, one band of expected size was observed in all before induction strains still including the excision cassette in the target locus. As expected, no bands could be observed with either wild-type starting strain or any of the after induction strains with the excision cassette flipped out of the genome.

These results suggest a successful, targeted, single-copy integration of the cassettes in the expected locus in all loci tested. Both PCR analysis and sequencing of the modified loci and the results from the southern blots verify the expected excision of the flipper- cassette, leaving only one 34bp *FRT* in the genome.

Also, a series of new *E .coli*/*P. pastoris* shuttle vectors were constructed during this study. GFP was used as a model protein to test the functionality of all vectors. An overview of the vector features and properties are summarized in Table 4. All pPpB1-based vectors (including the *S. cerevisiae ADH1* promoter and terminator to regulate the expression of the Zeocin^{TM} marker gene) were noted to display distinct background growth upon primary selection of positive transformants. Therefore, only the largest colonies from each plate should be chosen for further screening. However these vectors favour multi copy gene integration and copy numbers of up to 60 gene copies per cell have been reached with this group of vectors (personal communication with Andrea Mellitzer). Contrary to the pPpB1 vectors, transformations with pPpT4 –based vectors with ILV5 promoter and AOD terminator to regulate the expression of the Zeocin^{TM} marker gene were noted to lead to no or very little background growth and usually low copy numbers (up to 7 copies/cell). Single copy integration was preferred if a low amount of DNA was used for the transformation.

Comparison of growth rates of *P. pastoris* CBS7435 derived platform strains

The growth rates of all strains used and created during this study are depicted in Table 5. For all shake flask cultivations, buffered minimal media were used to achieve reproducible results. The growth rates of the *P. pastoris ku70* deletion strain in baffled shake flasks was observed to be only 11% (±0,00%), 10% (±0,01%) and 30% (±0,01%) lower than that of the wild-type starting strain CBS7435 in the respective dextrose, glycerol or methanol minimal media. The *aox1*, *his4* and *arg4* deletion strains constructed during this work showed expected growth rates. Since the commonly used *S. cerevisiae ADH1* promoter was chosen to activate the transcription of the *HIS4* gene in the complementation plasmid pPpHIS4, species specific activation and de-repression of yeast promoters (52) could cause the observed lower growth rate of the complemented *his4* deletion strain on glycerol.

Discussion

To enable quick and efficient mutant construction for target gene knock-out and characterization in *P. pastoris*, a new strain with reduced non-homologous recombination and thus precise and exclusive integration at the targeted sites is required. In filamentous fungi and higher eukaryotic organisms, the integration of foreign DNA has been reported to occur preferentially via the NHEJ pathway joining DNA ends with little or no homology (53), while in *S. cerevisiae* HR has been reported to be dominant and non-homologous recombination only observed if HR was blocked or homologous chromosome unavailable to serve as a template (54). The repair process is initiated by a Ku70p/Ku80p heterodimer (7). This study describes the elucidation of *P. pastoris KU70* and *KU80* genes and for the first time shows experimental data of the involvement of the *P. pastoris KU70* homologue in NHEJ.

To date, most of the modified *P. pastoris* strains have been constructed using unspecific mutagenesis methods or excessively long integration sequences. Gene replacement events have previously been described to occur with a frequency of <0.1% when a total of <500bp homology, short enough for high throughput screening, is used (12). We have constructed and characterized a *ku70* deletion strain with a specific 215bp knock-out in the *KU70* locus. Through analyzing the homologous recombination frequencies in this strain using two well-known auxotrophic loci and homologous sequence lengths varying from 50bp to 1350bp, we have shown that when using the *ku70* deletion strain, 100% homologous recombination frequencies can be achieved with as little as 650bp homologous flanking regions on each side of the integration cassette. For the tested sites, a reasonable number of specific disruptants (>85%) was achieved using a minimum flanking sequence length of 250bp to 650bp. This is short enough even to be made synthetically using 1 to 3 oligonucleotides for high throughput screening of genomic loci in *ku70* mutant strains. Only low numbers (0-17%, correlating with the size of the integration sequences) of disruptants can be achieved in the wt strain with any flanking sequence length tested. In the well accessible *HIS4* locus of the *ku70* knock-out strain, over one third of the cassettes showed correct integration with a homologous sequence length as short as 100bp. Furthermore, the differences between the *HIS4* and the *ADE1* loci confirmed the frequently discussed observations that the efficiency of site-specific integration and absolute knock-out efficiency are locus dependent. This has often caused significant troubles in achieving specific knock-out strains of *P. pastoris* previously (unpublished results). The *KU70* locus seems to play an essential role in the NHEJ mechanism of *P. pastoris*. Knocking out *KU70* reduced the fraction of randomly integrated DNA fragments dramatically. To demonstrate the use of the *ku70* mutant strain in the search of new auxotrophies, we successfully disrupted the *P. pastoris GUT1* gene from the glycerol assimilation pathway creating a new *gut1* deletion strain and complemented the locus with the *E. coli*/*P. pastoris* shuttle vector pPpGUT1. In addition to the use as an auxotrophic strain, a *gut1* deletion strain with complemented *KU70* locus might also be useful as a whole cell biocatalysis platform for biotransformations using glycerol as a substrate due to its inability to efficiently utilize glycerol as a carbon source for biomass production.

For most of the transformants where the selection marker was integrated and no auxotrophy was observed, the *HIS4* or *ADE1* disruption cassettes were expected to be randomly integrated in the genome. By sequencing the integration loci of such “false” and undesired transformants we detected no notable homology of the integration site to the 5’ end of the integrated disruption cassette. Cassettes were found in both coding and non-coding regions distributed among all four chromosomes. Only two strains showed integration of the cassette in the same region, thus provoking discussion about an integration “hotspot” with chromatin structure more permissive for integration. However, future studies with large sets of strains are required to confirm these suggestions. In accordance to previous studies (6–11) in yeasts, in two of the 14 strains there was also a single nucleotide deletion at the DSB site.

In addition to the role of Ku70p/Ku80p heterodimer in DNA end joining, it has also been shown to play an important role in the regulation of normal DNA end structure in telomeres (55, 56). Ku deficient *S. cerevisiae* strains have been reported to have growth defects markedly specific to the culture conditions. Especially cultivation at elevated temperatures has been reported to lead to growth arrest (22, 57). Ku-deficient *M. musculus* and *S. cerevisiae* cells have been reported to have V(D)J recombination defects, display telomeric shortening and high levels of chromosomal aberrations (21, 23–25). However, our results are more consistent with previous studies in *C. glabrata* (26) and *S. macrospora* (20), where no severe defects in the development and growth could be detected in normal growth conditions. The *P. pastoris ku70* mutant strain was observed to grow only 11% (±0,00%) slower than the corresponding wt strain in minimal dextrose media liquid cultures at 28°C, indicating at least that differences exist for *P. pastoris* too. As suggested previously (55), it is possible that the normal telomerase activity suffices and the activity of the Ku complex in establishing a correct terminal DNA structure is not required when the cells are not dividing too rapidly. However, the effects of the KU70 knock-out on protein expression in large-scale cultivations with higher stress levels have not been evaluated so far. Thus, at this stage direct industrial applications without restoring the normal activity of the Ku70p/Ku80p complex by complementing the locus are not recommended. The slower growth rate could also be caused by DNA damage-induced cell cycle arrest and apoptosis (58). When exposed to UV rays, the *ku70* mutant strain was observed to be more sensitive than the corresponding wt strain. This fact suggests that the *P. pastoris* Ku70 protein could be involved in the DNA-repair processes. However, despite the possible role of the Ku70 protein in DNA-repair, our results suggest that at least in case of overexpression of non-toxic heterologous proteins such as GFP, deleting the normal function of the *P. pastoris* Ku70 protein neither leads to severe genetic instability nor to an increased loss of expression cassettes within reasonable cultivation times corresponding to 200-300 generations, suggesting that the strain is a suitable host for simple lab-scale experiments. Therefore, the *ku70* mutant strain offers a noteworthy choice, for example, for primary functional studies and screening for selectable markers. Due to the decelerated growth and possible DNA-repair defects exposing the strain to increased genomic instability, complementing the KU70 locus or verification of promising results in the wild-type strain background is recommended. This is of particular importance if the strain generated is destined for industrial scale production of heterologous proteins.

In conclusion, in this study we have shown that the *ku70* deletion strain enables the quick construction of precise site directed genomic integrations and thus contributes, for example, to gene function analyses and the identification of target genes to generate new selection systems. In addition, effects of the overexpression of individual genes can be separated from undesired effects caused by random integration in other important loci of the genome. Although no lethality, severe growth retardation or loss of expression cassettes during four rounds of methanol induction was detected, the *ku70* deletion strain was observed to express decelerated growth and have possible defects in DNA-repair processes. Inactivation of *KU70* could also cause unknown metabolic changes. Therefore, when analyzing the specific functions of genes, the normal activity of *KU70* can also be restored after successful gene targeting. The simplicity of locus complementation is aided by the extremely high homologous integration frequency and the minimal deletion in the *KU70* locus.

The observed simplicity of targeted integration in the *ku70* deletion strain together with the wealth of information provided by the *P. pastoris* full genome sequence opens a new era for the creation and testing of designer strains, for applications in synthetic biology and for supplementing the very limited amount of selectable markers available for *P. pastoris*. Especially regarding gene disruption studies, the efforts to find more biosynthetic markers for the production of pharmaceuticals and the co-expression of multiple proteins or multi-subunit proteins are slow and cumbersome due to the problems caused by NHEJ. Basic expression systems for recombinant protein production in *P. pastoris* have been commercially available in the recent past, but diversity of the available tools has still been low. With the help of the new *ku70* deletion strain of this study we identified a *gut1* deletion strain, which shows no growth on glycerol. This deletion can be simply complemented again, offering antibiotic free selection based on specific carbon sources in the growth media. The *ku70* deletion strain provides quick access to new selection markers and highly specific targeted gene disruptions and insertions for foreign pathway construction. In order to provide a new second generation *P. pastoris* platform without unknown undesired mutations in the genome, we also constructed a series of precise new deletion strains based on the fully sequenced *P. pastoris* wt strain CBS7435, obtained from the Dutch strain collection (also deposited as ATCC 76273) and developed a complementary new versatile platform of *E. coli*/*P. pastoris* shuttle vectors enabling the production of a wide variety of proteins with different requirements concerning promoter efficiency, choice of markers, secretion and, due to the special properties of the *ku70* deletion strain, also precise site specific gene integration and knock-out. Furthermore, the *P. pastoris ku70* deletion strain facilitates the functional characterization of the genome and the metabolic routes of this important protein production host, and it will accelerate further developments of this host platform and corresponding metabolic models.

Funding

This work was supported by the Austrian Science Fund 'Fonds zur Förderung der wissenschaftlichen Forschung'[W901 DK Molecular Enzymology] and VTU Technology GmbH. Funding for open access charge: Fonds zur Förderung der wissenschaftlichen Forschung. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We would like to thank Viktorija Vidimce, Anna-Maria Hatzl, Birgit Brunner and Christian Schmid for excellent technical assistance.

References

1. Cregg JM, Madden KR, Barringer KJ, Thill GP, Stillman CA (1989) Functional characterization of the two alcohol oxidase genes from the yeast *Pichia pastoris*. Mol Cell Biol 9: 1316-23.

2. Lin Cereghino GP, Lin Cereghino J, Ilgen C, Cregg JM (2002) Production of recombinant proteins in fermenter cultures of the yeast *Pichia pastoris*. Curr Opin Biotechnol 13: 329-332.

3. De Schutter K, Lin Y-C, Tiels P, Van Hecke A, Glinka S *et al*. (2009) Genome sequence of the recombinant protein production host *Pichia pastoris*. Nat Biotechnol 27: 561-566.

4. Küberl A, Schneider J, Thallinger GG, Anderl I, Wibberg D *et al*. (2011) High-quality genome sequence of *Pichia pastoris* CBS7435. J Biotechnol 154: 312-320.

5. Pastwa E, Blasiak J (2003) Non-homologous DNA end joining. Acta Biochim Pol (Engl Transl) 50: 891–908.

6. Daley JM, Palmbos PL, Wu D, Wilson TE (2005) Nonhomologous end joining in yeast. Annu Rev Genet 39: 431-451.

7. Critchlow SE, Jackson SP (1998) DNA end-joining: from yeast to man. Trends Biochem Sci 23: 394-398.

8. Aylon Y, Kupiec M (2004) DSB repair: the yeast paradigm. DNA repair 3:797-815.

9. Dudásová Z, Dudás A, Chovanec M (2004) Non-homologous end-joining factors of *Saccharomyces cerevisiae*. FEMS Microbiol Rev 28: 581-601.

10. Klinner U, Schäfer B (2004) Genetic aspects of targeted insertion mutagenesis in yeasts. FEMS Microbiol Rev 28: 201-223.

11. Yamana Y, Maeda T, Ohba H, Usui T, Ogawa HI *et al*. (2005) Regulation of homologous integration in yeast by the DNA repair proteins Ku70 and RecQ. Mol Genet Genomics 273: 167-716.

12. Higgins DR, Cregg JM eds. (1998) *Pichia* Protocols. Totowa, New Jersey: Humana Press.

13. Li P, Anumanthan A, Gao X-G, Ilangovan K, Suzara VV *et al.* (2007) Expression of Recombinant Proteins in *Pichia Pastoris*. Appl. Biochem Biotechnol 142: 105-124.

14. Mimori T, Hardin JA (1986) Mechanism of interaction between Ku protein and DNA. J Biol Chem 261:10375-10379.

15. Paillard S, Strauss F (1991) Analysis of the mechanism of interaction of simian Ku protein with DNA. Nucleic acids res 19: 5619-5624.

16. Milne GT, Jin S, Shannon KB, Weaver DT (1996) Mutations in two Ku homologs define a DNA end-joining repair pathway in *Saccharomyces cerevisiae*. Mol Cell Biol 16: 4189-4198.

17. Walker JR, Corpina RA, Goldberg J (2001) Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. Nature 412: 607-614.

18. Moshous D, Callebaut I, de Chasseval R, Corneo B, Cavazzana-Calvo M *et al.* (2001) Artemis, a novel DNA double-strand break repair/V(D)J recombination protein is mutated in human severe combined immune deficiency. Cell 105: 177-86.

19. Ninomiya Y, Suzuki K, Ishii C, Inoue H (2004) Highly efficient gene replacements in *Neurospora* strains deficient for nonhomologous end-joining. Proc Natl Acad Sci U S A 101: 12248-12253.

20. Pöggeler S, Kück U (2006) Highly efficient generation of signal transduction knockout mutants using a fungal strain deficient in the mammalian ku70 ortholog. Gene 378: 1-10.

21. Zhu C, Bogue MA, Lim DS, Hasty P, Roth DB (1996) Ku86-deficient mice exhibit severe combined immunodeficiency and defective processing of V(D)J recombination intermediates. Cell 86: 379-389.

22. Boulton SJ, Jackson SP (1996) Identification of a *Saccharomyces cerevisiae* Ku80 homologue: roles in DNA double strand break rejoining and in telomeric maintenance. Nucleic Acids Res 24: 4639-4648.

23. Nussenzweig A, Sokol K, Burgman P, Li L, Li GC (1997) Hypersensitivity of Ku80-deficient cell lines and mice to DNA damage: the effects of ionizing radiation on growth, survival, and development. Proc Natl Acad Sci *U S A* 94: 13588-93.

24. Difilippantonio MJ, Zhu J, Chen HT, Meffre E, Nussenzweig MC *et al*. (2000) DNA repair protein Ku80 suppresses chromosomal aberrations and malignant transformation. Nature 404: 510-4.

25. Ferguson DO, Sekiguchi JM, Chang S, Frank KM, Gao Y *et al*. (2000) The nonhomologous end-joining pathway of DNA repair is required for genomic stability and the suppression of translocations. Proc Natl Acad Sci U S A 97: 6630-6633.

26. Ueno K, Uno J, Nakayama H, Sasamoto K, Mikami Y *et al.* (2007) Development of a highly efficient gene targeting system induced by transient repression of *YKU80* expression in *Candida glabrata*. Eukaryotic cell 6: 1239-1247.

27. Ribes-Zamora A, Mihalek I, Lichtarge O, Bertuch AA (2007) Distinct faces of the Ku heterodimer mediate DNA repair and telomeric functions. Nat Struct Mol Biol 14: 301-307.

28. Feldmann E, Schmiemann V, Goedecke W, Reichenberger S, Pfeiffer P (2000) DNA double-strand break repair in cell-free extracts from Ku80-deficient cells: implications for Ku serving as an alignment factor in non-homologous DNA end joining. Nucleic Acids Res 28: 2585-2596.

29. Krappmann S, Sasse C, Braus GH (2006) Gene Targeting in *Aspergillus fumigatus* by Homologous Recombination Is Facilitated in a Nonhomologous End- Joining-Deficient Genetic Background. Eukaryotic Cell 5: 212-215.

30. Nayak T, Szewczyk E, Oakley CE, Osmani A, Ukil L et al. (2006) A versatile and efficient gene-targeting system for *Aspergillus nidulans.* Genetics 172:1557-1566.

31. Takahashi T, Masuda T, Koyama Y (2006) Identification and analysis of Ku70 and Ku80 homologs in the koji molds *Aspergillus sojae* and *Aspergillus oryzae*. Biosci Biotechnol Biochem 70: 135-143.

32. Meyer V, Arentshorst M, El-Ghezal A, Drews A-C, Kooistra R et al. (2007) Highly efficient gene targeting in the *Aspergillus niger* kusA mutant. J Biotechnol 128: 770-775.

33. Villalba F, Collemare J, Landraud P, Lambou K, Brozek V et al. (2008) Improved gene targeting in *Magnaporthe grisea* by inactivation of *MgKU80* required for non-homologous end joining. Fungal Genet Biol 45: 68-75.

34. Kooistra R, Hooykaas PJJ, Steensma HY (2004) Efficient gene targeting in *Kluyveromyces lactis*. Yeast 21: 781-792.

35. Ausubel FM, Brent R, Kingston RE, Moore DD Seidman JG *et al.* eds. (2007) Current Protocols in Molecular Biology. New York: John Wiley & Sons.

36. Ruth C, Zuellig T, Mellitzer A, Weis R, Looser V *et al*. (2010) Variable production windows for porcine trypsinogen employing synthetic inducible promoter variants in *Pichia pastoris*. Syst Synth Biol 4: 181-191.

37. Lin-Cereghino J, Wong WW, Xiong S, Giang W, Luong LT, et al. (2008) Condensed protocol for competent cell preparation and transformation of the methylotrophic yeast *Pichia pastoris.* Biotechniques 38: 44-48.

38. Reuss O, Vik A, Kolter R, Morschhäuser J (2004) The *SAT1* flipper, an optimized tool for gene disruption in *Candida albicans*. Gene 341: 119-127.

39. Higgins DG, Thompson JD, Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. Methods Enzymol 266: 383-402.

40. Villalobos A, Ness JE, Gustafsson C, Minshull J, Govindarajan S (2006) Gene Designer: a synthetic biology tool for constructing artificial DNA segments. BMC Bioinf 7: 285.

41. Broach JR, Guarascio VR, Jayaram M (1982) Recombination within the yeast plasmid 2µ circle is site-specific. Cell 29: 227-234.

42. Weis R, Luiten R, Skranc W, Schwab H, Wubbolts M, *et al*. (2004) Reliable high-throughput screening with *Pichia pastoris* by limiting yeast cell death phenomena. FEMS Yeast Res 5: 179-189.

43. Schroer K, Luef KP, Hartner FS, Glieder A, Pscheidt B (2010) Engineering the *Pichia pastoris* methanol oxidation pathway for improved NADH regeneration during whole-cell biotransformation. Metab Eng 12: 8-17.

44. Abad S, Kitz K, Hörmann A, Schreiner U, Hartner FS *et al*. (2010) Real-time PCR-based determination of gene copy numbers in *Pichia pastoris*. Biotechnol J 5: 413-420.

45. Southern EM (1974) Improved Method from Electrophoresis for Transferring Nucleotides Strips to Thin Layers Cellulose of Ion-Exchange. Anal Biochem 62: 317-318.

46. Crameri A, Whitehorn EA, Tate E, Stemmer WP (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. Nat Biotechnol 14: 315-319.

47. Hartner FS, Ruth C, Langenegger D, Johnson SN, Hyka P *et al*. (2008) Promoter library designed for fine-tuned gene expression in *Pichia pastoris*. Nucleic Acids Res 36: e76 10.1093/nar/gkn369

48. Hoffman CS, Winston F (1987) A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*. Gene 57: 267-272.

49. Nussenzweig A, Chen C, da Costa Soares V, Sanchez M, Sokol K *et al*. (1996) Requirement for Ku80 in growth and immunoglobulin V(D)J recombination. Nature 382: 551-555.

50. Zhu T, Guo M, Sun C, Qian J, Zhuang Y, Chu J, Zhang S (2009) A systematical investigation on the genetic stability of multi-copy *Pichia pastoris* strains. Biotechnol Lett 31: 679-684.

51. Inan M, Meagher MM (2001) Non-repressing carbon sources for *alcohol oxidase* (*AOX1*) promoter of *Pichia pastoris*. J Biosci Bioeng 92: 585-589.

52. Raschke WC, Neiditch BR, Hendricks M, Cregg JM (1996) Inducible expression of a heterologous protein in *Hansenula polymorpha* using the *alcohol oxidase 1* promoter of *Pichia pastoris*. Gene 177: 163-167.

53. Chu G (1997) Double Strand Break Repair. J Biol Chem 272: 24097-24100.

54. Clikeman JA, Khalsa GJ Barton SL, Nickoloff JA (2000) Homologous Recombinational Repair of Double-Strand Breaks in Yeast Is Enhanced by MAT Heterozygosity Through yKU-Dependent and –Independent Mechanisms. Genetics 157: 579-589.

55. Gravel S, Larrivée M, Labrecque P, Wellinger RJ (1998) Yeast Ku as a Regulator of Chromosomal DNA End Structure. Science 280: 741-744.

56. Gravel S, Wellinger RJ (2002) Maintenance of Double-Stranded Telomeric Repeats as the Critical Determinant for Cell Viability in Yeast Cells Lacking Ku. Mol Cell Biol. 22: 2182-2193.

57. Feldmann H, Winnacker EL (1993) A putative homologue of the human autoantigen Ku from *Saccharomyces cerevisiae*. J Biol Chem 268: 12895-12900.

58. Schärer OD (2003) Chemistry and biology of DNA repair. Angew Chem Int Ed Engl 42: 2946-2974.

59. Pahlman AK, Granath K, Ansell R, Hohmann S, Adler L (2001) The yeast glycerol 3-phosphatases Gpp1p and Gpp2p are required for glycerol biosynthesis and differentially involved in the cellular responses to osmotic anaerobic and oxidative stress. J Biol Chem 276: 3555-3563.

Supplementary References

60. Oka Y, Ishida H, Morioka M, Numasaki Y, Yamafuji T *et al*. (1981) Combimicins, new kanamycin derivatives bioconverted by some *Micromonosporas*. Journal of Antibiotics 34: 777-781.

61. Wach A, Brachat A, Pöhlmann R, Philippsen P (1994) New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. Yeast 10: 1793-808.

62. Umlauf SW, Cox MM (1988) The functional significance of DNA sequence structure in a site-specific genetic recombination reaction. The EMBO journal 7: 1845-1852.

Tables

Table 1: Excision cassette structures

Locus	**5’ integration**	**3’ integration**	**Knock-out size**
KU70	-200 to +1099	+1315 to +2484	215bp
AOX1	-939 to -1	+1993 to +3136	1992bp
HIS4	-894 to -51	+2533 to +3414	2582bp
ARG4/1	-1500 to -1	+1399 to +2899	1398bp
ARG4/2	+22 to +984	+1399 to 2900	414bp
GK1	-169 to +752	+1138 to +2050	385bp
DAS1	+62 to +580	+581 to +1133	0 (disruption)
DAS2	+62 to +580	+581 to +1133	0 (disruption)

Homologous integration sequence lengths used in the 5’ and 3’ ends of the excision cassettes were defined by the structure of the target locus and the sequence data available at the time point of cassette design.

Table 2: Homologous recombination frequency (HRFb) in wt CBS7435 and *ku70* deletion strains.

Recipient strain	**Locus**	**Length of homology on each side**	**Auxotrophic/total number of transformants counted**	**HRFb%**
wt	*HIS4*	~1350 bp	44/266	16.5
		~1000 bp	31/268	11.6
		~650 bp	32/268	11.9
		~250 bp	8/268	3.0
		~100 bp	1/268	0
	ADE1	~650 bp	0/>10^3	0
		~400 bp	0/>10^3	0
		~150 bp	11/>10^3	0
		~50 bp	0/>10^3	0
ku70	*HIS4*	~1350 bp	267/268	100
		~1000 bp	268/268	100
		~650 bp	267/268	100
		~250 bp	86/89	96.6
		~100 bp	20/57	35.1
	ADE1	~650 bp	465/533	87.9
		~400 bp	114/172	75.4
		~150 bp	12/99	17.5
		~50 bp	1/317	0

Over 95% homologous recombination frequencies could be reached in the *ku70* deletion strain with as little as 250bp of homologous sequence on each side of the integration cassette. In the corresponding wild-type strain, only 16.5% homologous recombination frequency was reached with the longest (1350bp) homologous sequence tested.

Table 3: Selection marker integration.

Strain #	Digestion	Strain	Integration sequence	Integration locus	Gene hit
1	HindIII	wt	HIS4 250bp	Chr. 2 (1088794)	Protein Mis14
1	EcoRI	wt	HIS4 250bp	Chr. 2 (1088794)	Protein Mis14
2	BglII	wt	HIS4 250bp	Chr. 4 (1549929)	Cell morphogenesis protein Pag1
3	BamHI	mut^s	HIS4 250bp	Chr. 1 (2815628)	Non-coding: 64 bp at 5' side: Protein Tos1, 292 bp at 3' side: Uncharacterized protein YPL066W
3	BglII	mut^s	HIS4 250bp	Chr. 1 (2815628)	Non-coding: 64 bp at 5' side: Protein Tos1, 292 bp at 3' side: Uncharacterized protein YPL066W
4	BglII	mut^s	HIS4 250bp	Chr. 4 (1356618)	Non-coding: upstream of hypothetical protein
6	HindIII	wt	HIS4 250bp	Chr. 1 (1935839)	Non-coding: 98 bp at 5' side: Protein Lst4, 310 bp at 3' side: protein midasin
8	BamHI	mut^s	HIS4 250bp	Chr. 4 (1549929)	Cell morphogenesis protein Pag1
10	BglII	wt	HIS4 250bp	Chr. 3 (1866738)	Non-coding: 66 bp at 5' side: Zinc finger protein 167, 288 bp at 3' side: 1,3-beta-glucanosyltransferase
10	BamHI	wt	HIS4 250bp	Chr. 3 (1866738)	Non-coding: 66 bp at 5' side: Zinc finger protein 167, 288 bp at 3' side: 1,3-beta-glucanosyltransferase
11	HindIII	mut^s	HIS4 250bp	Chr. 4 (1354273)	Protein Ecm3
12	BamHI	mut^s	HIS4 250bp	Chr. 1 (2137406)	Ammonium transporter protein Mep2
13	HindIII	wt	HIS4 250bp	Chr. 1 (2706933)	Inositol 2-dehydrogenase
14	BamHI	mut^s	HIS4 250bp	Chr. 4 (846891)	Non-coding: 188 bp at 5' side: Protein Letm1 and EF-hand domain-containing protein anon-60Da, 143 bp at 3' side: phosphatidylinositol 3-kinase
19	EcoRI	wt	ADE1 150bp	Chr. 3 (87007)	Non-coding: 420 bp at 3' side: mannose-6-phosphate isomerase
20	BamHI	wt	ADE1 150bp	Chr. 1 (1994439)	Likely SIR2 family histone deacetylase
21	EcoRI	wt	ADE1 150bp	Chr. 3 (2074903)	Eukaryotic translation initiation factor 2 subunit alpha
21	HindIII	wt	ADE1 150bp	Chr. 3 (2074903)	Eukaryotic translation initiation factor 2 subunit alpha

Integration sites of the gene disruption cassettes in *P. pastoris* wt strains which remained autotroph after selection marker integration.

Table 4: New shuttle vectors constructed during this study

Name	**Accession #**	**Promoter**[a]	**Localization**[b]	**Linearization**	**Selection**
pPpB1_S	JQ519685	*AOX1*	Intracellular	Blunt	Zeocin^{TM}
pPpB1GAP	JQ519686	*GAP1*	Intracellular	Sticky-end	Zeocin^{TM}
pPpB1GAP_S	JQ519687	*GAP1*	Intracellular	Blunt	Zeocin^{TM}
pPpB1_Alpha_S	JQ519688	*AOX1*	Secreted	Blunt	Zeocin^{TM}
pPpT4	JQ519689	*AOX1*	Intracellular	Sticky-end	Zeocin^{TM}
pPpT4_S	JQ519690	*AOX1*	Intracellular	Blunt	Zeocin^{TM}
pPpT4_Alpha_S	JQ519691	*AOX1*	Secreted	Blunt	Zeocin^{TM}
pPpT4GAP_S	JQ519692	*GAP1*	Intracellular	Blunt	Zeocin^{TM}
pPpT4GAP_Alpha_S	JQ519693	*GAP1*	Secreted	Blunt	Zeocin^{TM}
pPpKan_S	JQ519694	*AOX1*	Intracellular	Blunt	KanMX6
pPpKan_Alpha_S	JQ519695	*AOX1*	Secreted	Blunt	KanMX6
pPpARG4	JQ519696	*AOX1*	Intracellular	Sticky-end	::*ARG4*
pPpHIS4	JQ519697	*AOX1*	Intracellular	Sticky-end	::*HIS4*
pPpGUT1	JQ519698	*AOX1*	Intracellular	Blunt	::*GUT1*

[a] Promoter to regulate the expression of the gene of interest

[b] Localization of the recombinant protein. Vectors aimed for intracellular production can be used for the secretory production by adding a signal sequence

Table 5: Strains of *P. pastoris* used and constructed during this work.

Strain	**Genotype**	**Markers**	**Source**	**BMD2%**	**BMG1%**	**BMM0.5%**
CBS7435	wild type[a]	-	CBS[b]	0.30 ±0.01	0.28±0.02	0.17±0.00
mut^s	*aox1*	-	This study	0.29 ±0.02	0.30±0.01	0.05±0.00
mut^s arg-	*aox1arg4*	-	This study	0.26 ±0.01	0.27±0.01	0.06±0.00
mut^s his-	*aox1his4*	-	This study	0.31 ±0.01	0.29±0.01	0.05±0.00
his-	*his4*	-	This study	0.30 ±0.01	0.29±0.01	0.15±0.00
ku70-	*ku70*	-	This study	0.27 ±0.00	0.26±0.01	0.12±0.00
mut^s arg- c.	*aox1arg4*::ARG4	*bla*	This study	0.32 ±0.01	0.25±0.01	0.03±0.00
his- c.	*his4*::HIS4	*bla*	This study	0.32 ±0.02	0.24±0.01	0.14±0.00
mut^s his- c.	*aox1 his4*::HIS4	*bla*	This study	0.31 ±0.01	0.23±0.04	0.01±0.00
ku70-das1-	*ku70 das1*	*Sh ble*	This study	0.17 ±0.01	0.17±0.04	0.04±0.01
ku70-das12-	*ku70 das1das2*	*Sh ble*	This study	0.21 ±0.00	0.22±0.01	0
ku70-gut-	*ku70 gut1*	*Sh ble*	This study	0.26 ±0.01	0	0.13±0.01
ku70-gut- c.	*ku70 gut1*::*GUT1*	-	This study	0.27 ±0.02	0.23±0.01	0.14±0.00
ku70-his-	*ku70 his4*	-	This study	n/d	n/d	n/d
ku70-his-	*ku70 his4*	*Sh ble*	This study	n/d	n/d	n/d
ku70-ade1-	*ku70 ade1*	*Sh ble*	This study	n/d	n/d	n/d

The growth rates reported correspond to the maximal growth rates ($^{h-1}$) reached in minimal media during the exponential growth phase. The standard deviation reported is calculated according to the growth rates of three biological replicates. c.=complemented. BM = buffered minimal media with glucose (D), glycerol (G) or methanol (M).

[a]NRRL Y-11430, ATCC 76273

[b]Centraalbureau voor Schimmelcultures

Figures

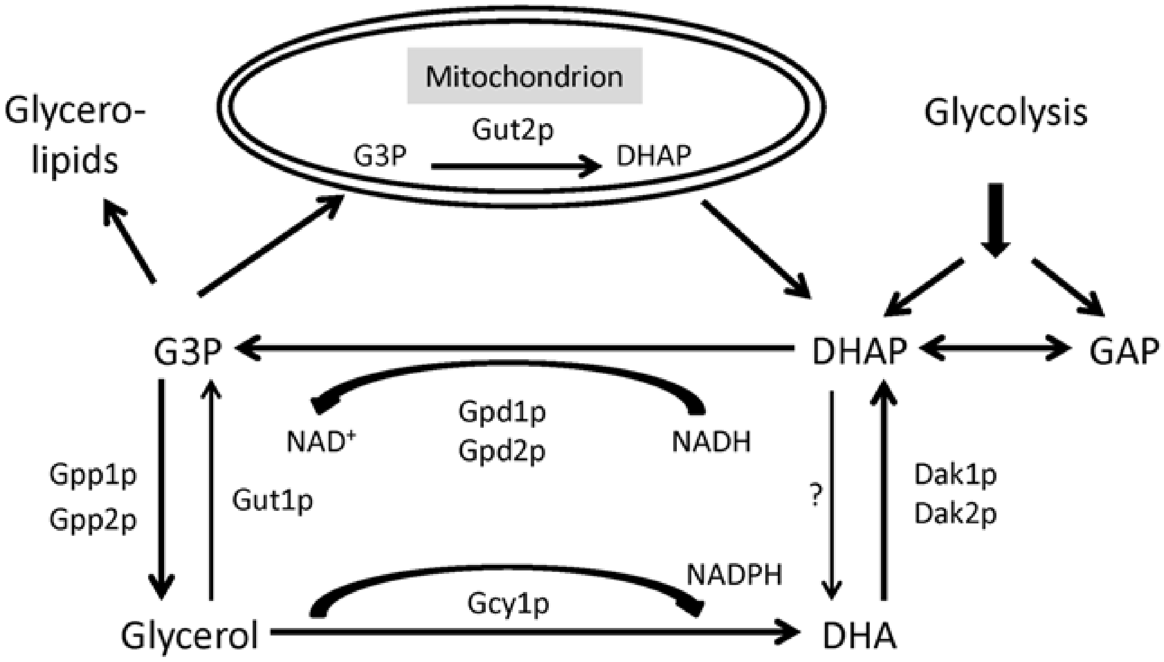


Figure 1: A simplified scheme of the glycerol assimilation pathway in yeast. Glycerol kinase 1 (Gut1p) knocked out in this study is involved in the first step of glycerol metabolism. Gpp1p and Gpp2p: glycerol-3-phosphatases 1 and 2, Gpd1p and Gpd2p: glycerol 3-phosphate dehydrogenases 1 and 2, Dak1 and Dak2p: dihydroxyacetone kinases 1 and 2, Gcy1p: putative NADP$^{(+)}$ coupled glycerol dehydrogenase, DHA: dihydroxyacetone, DHAP: dihydroxyacetone phosphate, G3P: Glycerol-3-phosphate, "?": the enzyme responsible for the conversion of DHAP to DHA is not verified. Under aerobic conditions, mitochondrial Gut2p (glycerol-3-phosphate dehydrogenase 2) can convert G3P to DHAP. Figure modified from (59).

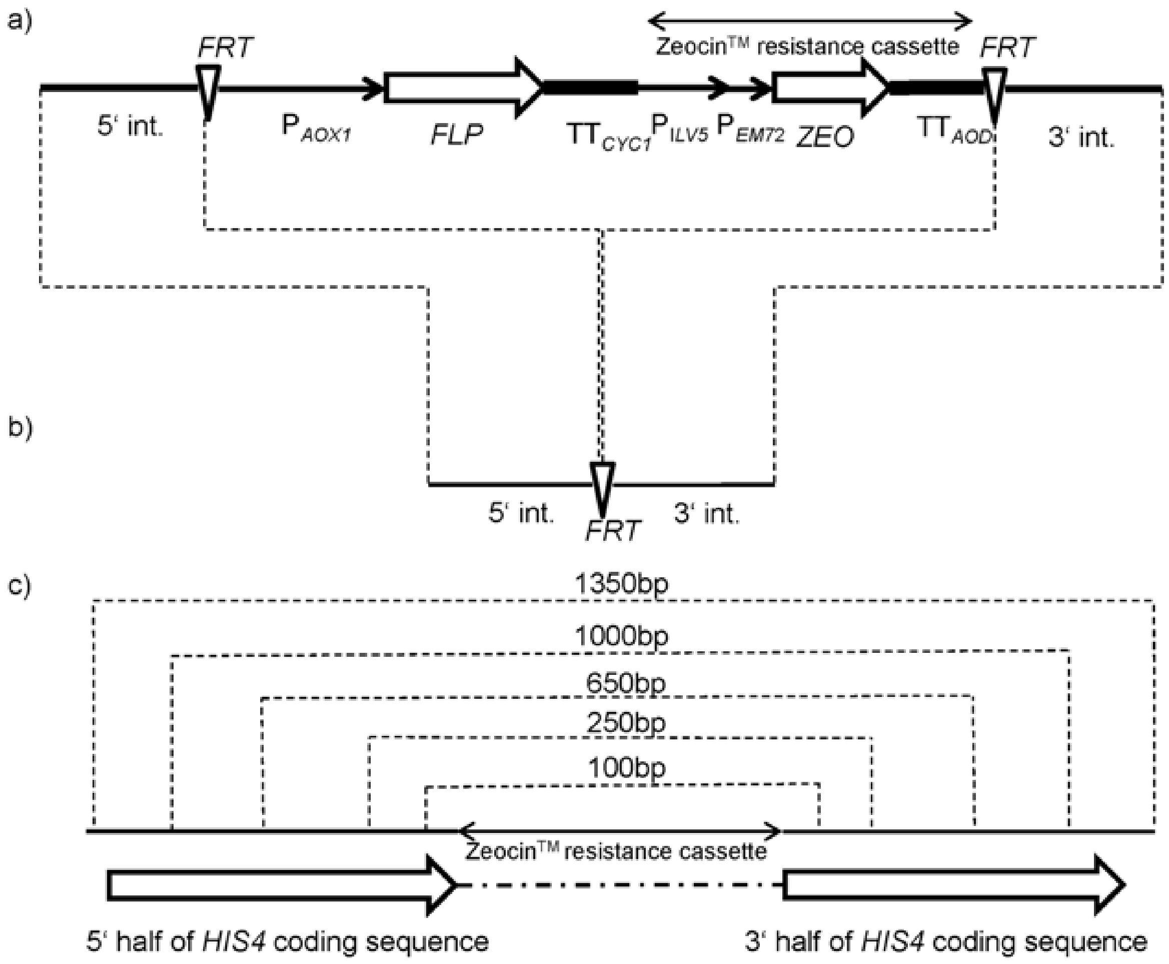

a)
FRT
Zeocin^{TM} resistance cassette
FRT
5' int.
P_{AOX1}
FLP
TT_{CYC1}
P_{ILV5}
P_{EM72}
ZEO
TT_{AOD}
3' int.
b)
5' int.
FRT
3' int.
c)
1350bp
1000bp
650bp
250bp
100bp
Zeocin^{TM} resistance cassette
5' half of HIS4 coding sequence
3' half of HIS4 coding sequence

Figure 2: Integration cassette composition and function. a) *KU70* disruption cassette based on the *S. cerevisiae* FLP recombinase system. On both sides the flipper cassette with *AOX1* promoter (P_{AOX1}), FLP recombinase (*FLP*), CYC1 terminator ($CYC1_{TT}$) and Zeocin^{TM} resistance cassette are surrounded by recombinase target sequences (*FRT*) and locus specific integration sequences (5'int and 3'int). Cassette components are not drawn to scale.

b) After methanol induced (P_{AOX1}) FLP production and subsequent *FRT* recognition leading to cassette excision only one *FRT* (34bp) is left in the locus in between the 3' and 5' integration sequences.

c) The lengths of the homologous sequences at 5' and 3' ends of the disruption cassettes used to compare the homologous recombination frequencies in wt and ku70 deletion strains varied from 100bp to 1350bp in the *HIS4* locus. Zeocin™ resistance cassette was placed in between the homologous sequences. Cassette components are not drawn to scale.

Supporting Information

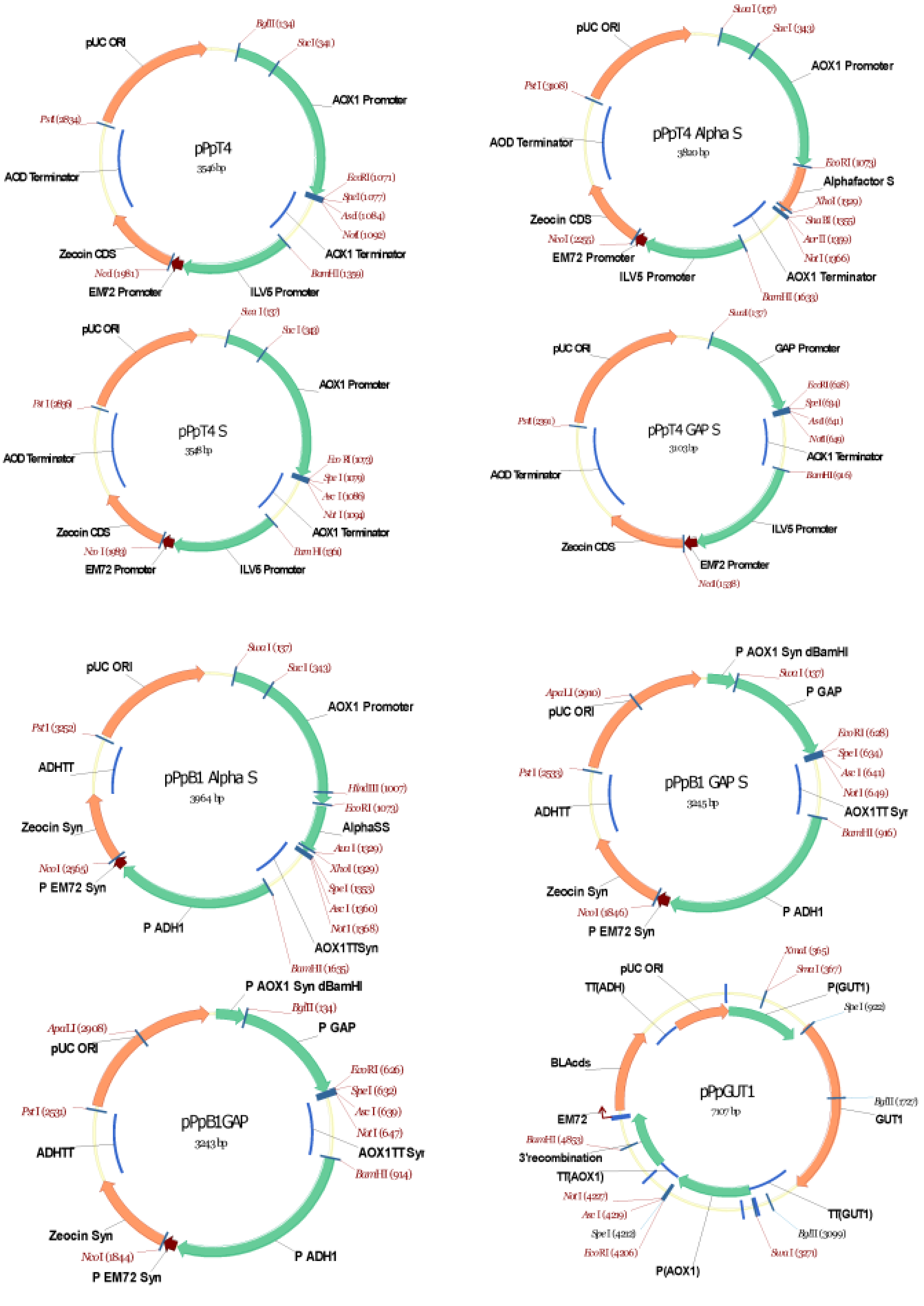


Figure S1. **Vector maps of the *E. coli/P. pastoris* shuttle vectors constructed during this study.** The origins and functions of plasmid components are depicted in Table S2.

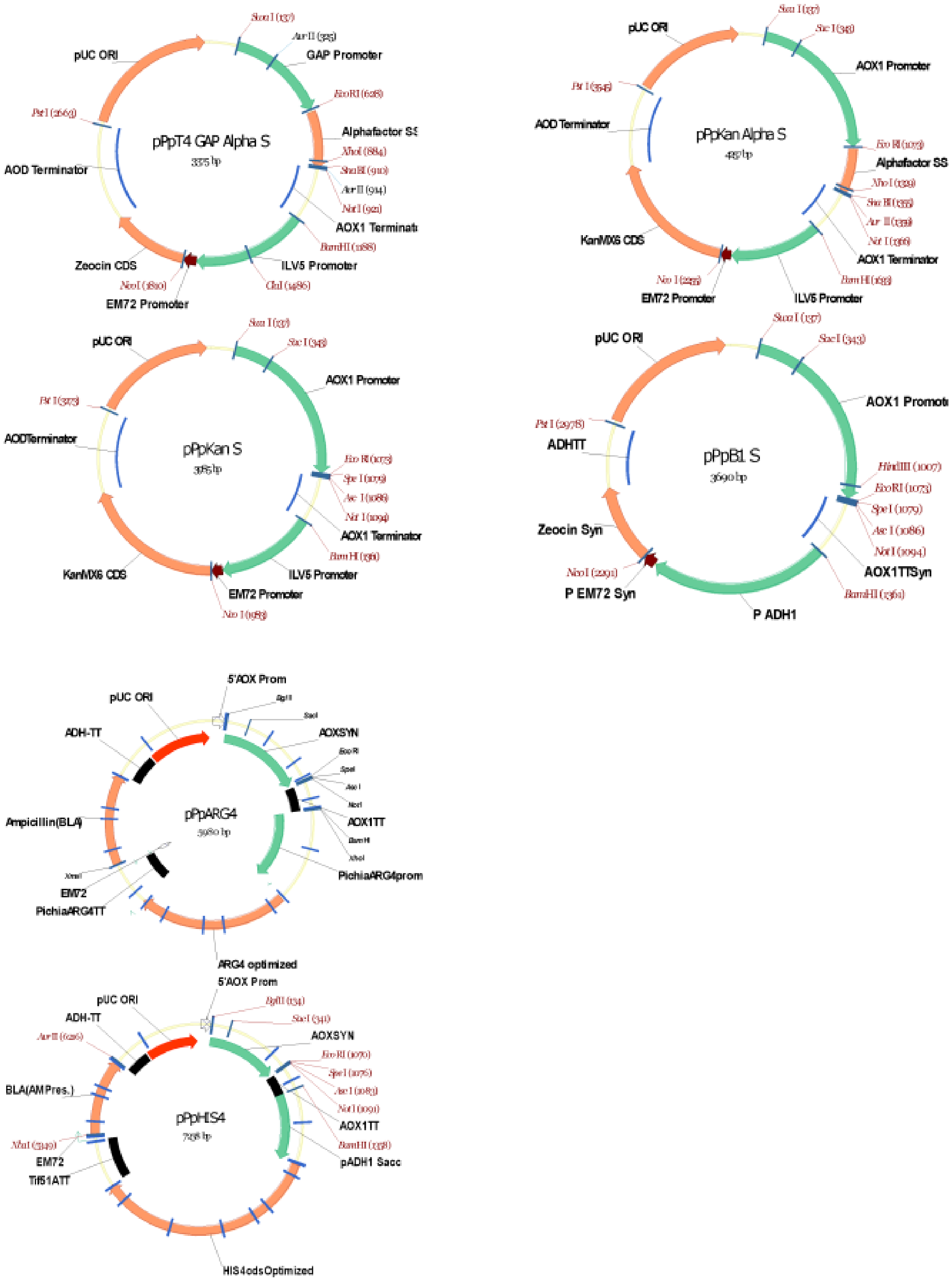


Figure S1 (continued). **Vector maps of the *E. coli*/*P. pastoris* shuttle vectors constructed during this study.** The origins and functions of plasmid components are depicted in Table S2.

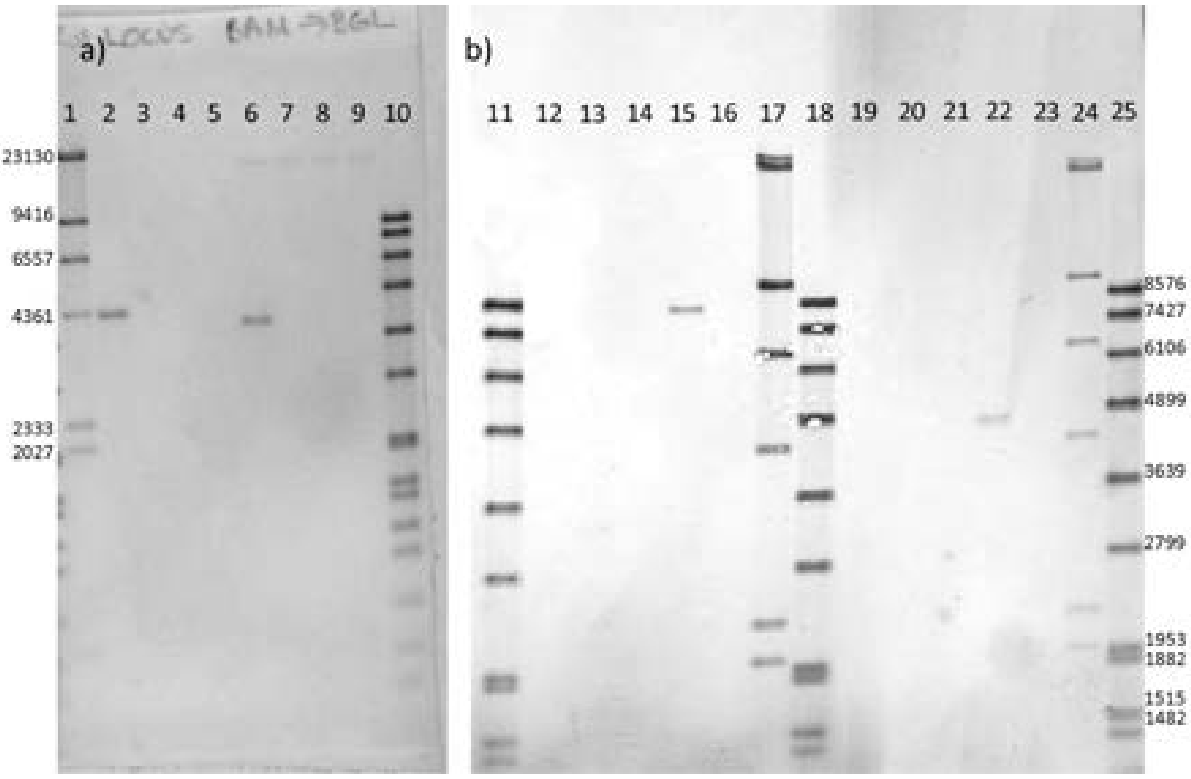
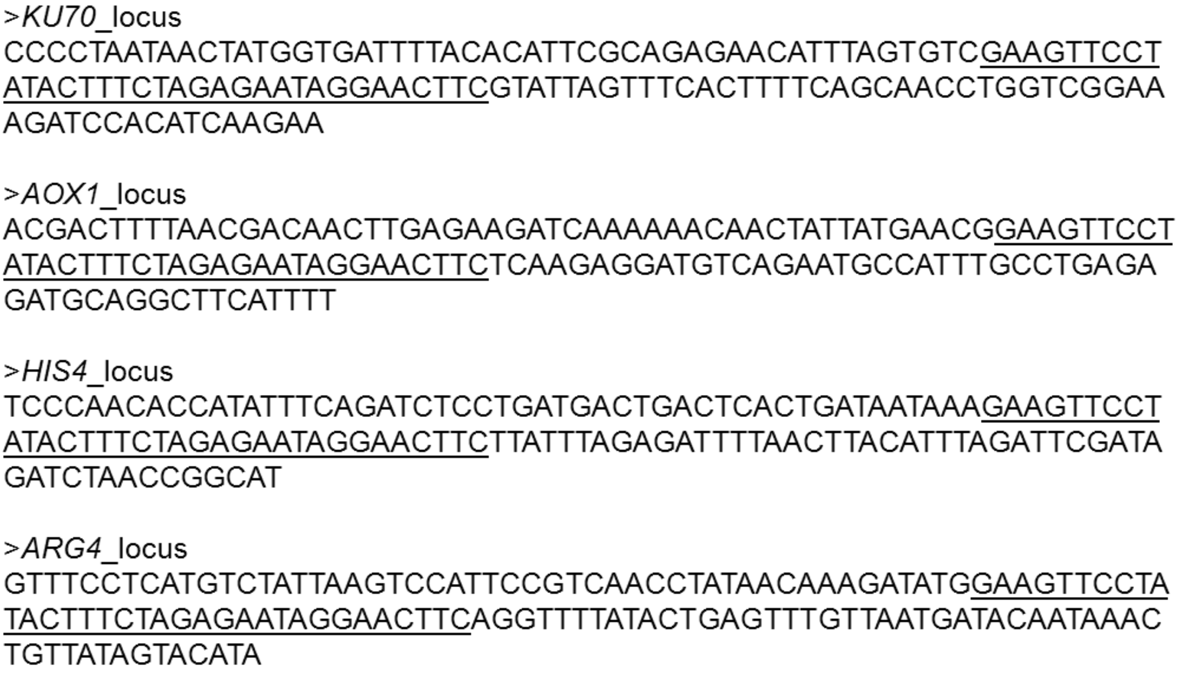


Figure S2. **Southern blots confirming the expected knock-out in the *KU70 locus***. a) BamHI (columns 2-5) and BglII (columns 6-9) digested DNA of the wt CBS7435 strain (columns 2 and 6), the *ku70* strain before induction (columns 3 and 7) and two strains after induction (columns 4-5 and 8-9) were detected with a probe specific to the knock-out region in the *KU70* locus. A band of the expected size can be observed in the wt strain only (4368bp for BamHI and 4127bp for BglII). b) BamHI (columns 12-15) and BglII (columns 19-22) digested DNA of the wt (columns 12 and 19), *ku70* after induction (13-14 and 20-21) and *ku70* before induction (columns 15 and 22) strains detected with Zeocin-specific probe. A band of the expected size can be observed in the before induction strain only (8412bp for BamHI and 4837bp for BglII). Lanes 16 and 23 are empty. Five µl of DIG-labelled ladder #II (Roche) was loaded to lanes 1, 17 and 24, and five µl of DIG-labelled ladder #VII to lanes 10, 11, 18 and 25.



>*KU70*_locus
CCCCTAATAACTATGGTGATTTTACACATTCGCAGAGAACATTTAGTGTCGAAGTTCCT
ATACTTTCTAGAGAATAGGAACTTCGTATTAGTTTCACTTTTCAGCAACCTGGTCGGAA
AGATCCACATCAAGAA

>*AOX1*_locus
ACGACTTTTAACGACAACTTGAGAAGATCAAAAAACAACTATTATGAACGGAAGTTCCT
ATACTTTCTAGAGAATAGGAACTTCTCAAGAGGATGTCAGAATGCCATTTGCCTGAGA
GATGCAGGCTTCATTTT

>*HIS4*_locus
TCCCAACACCATATTTCAGATCTCCTGATGACTGACTCACTGATAATAAAGAAGTTCCT
ATACTTTCTAGAGAATAGGAACTTCTTATTTAGAGATTTTAACTTACATTTAGATTCGATA
GATCTAACCGGCAT

>*ARG4*_locus
GTTTCCTCATGTCTATTAAGTCCATTCCGTCAACCTATAACAAAGATATGGAAGTTCCTA
TACTTTCTAGAGAATAGGAACTTCAGGTTTTATACTGAGTTTGTTAATGATACAATAAAC
TGTTATAGTACATA

Figure S3. **Knock-out locus sequences.** The sequences of the targeted sites confirmed the expected total (*aox1, aox1arg4*) or partial (*ku70, his4, aox1his4, ku70his4*) deletions. Only one *FRT* sequence (underlined) was left in the genome locus.

Table: S1. Primers used in this study. AOX1flipper: primers used in the construction and analysis of the *aox1* strain; probe: primers used for the amplification of the southern blot probes; Ku locus: primers used for the construction and analysis of the *ku70* deletion strain; ARG4flipper: primers used for the construction and analysis of the *arg4* strain; HIS4 flipper: primers used for the construction and analysis of the *his4* strain; Targeting: primers used for the construction and analysis of the disruption cassettes to evaluate targeting efficiencies; pPp: primers used for the construction and analysis of the complementation plasmids; GUT1disrupt: primers used for the construction and analysis of the *gut1* strain; seq: sequencing primer.

Oligo #	Oligo name	Sequence 5’-3’
P07614	AOX1flipper_1F	AGATCTAACATCCAAAGACGAAAGGTTGAATGAAACC
P07615	AOX1flipper_1R	GAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCCGTTTCAATAATTAGTTGTTTTTTG
P07616	AOX1flipper_2F	CTATACTTTCTAGAGAATAGGAACTTCATGCCACAATTTGATATATTATG
P07617	AOX1flipper_2R	CAAGACATTACTGAATAAGCTTACATTATGAAGAGCAGC
P07618	AOX1flipper_3F	GTAATATGCTGCTCTTCATAATGTAAGCTTATTCAGTAATGTCTTGTTTCTTTTG
P07619	AOX1flipper_3R	GAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCCTAAGGTAATCAGATCCAAG
P07620	AOX1flipper_4F	CTATACTTTCTAGAGAATAGGAACTTCTCAAGAGGATGTCAGAATG
P07621	AOX1flipper_4R	GATCTTGAGATAAATTTCACGTTTAAAATCAGCGTACCTTTTTCTCG
P07626	AOX1flipper_1F(short)	AGATCTAACATCCAAAGACG
P07627	AOX1flipper_4R(short)	GATCTTGAGATAAATTTCACG
P07628	AOX1flipper_1FRTFseq	GGTGCACCTGTGCCGAAACG
P07629	AOX1flipper_2FRTRseq	GTTCCGTTATGTGTAATCATCCAAC
P07630	AOX1flipper_3FRTFseq	CATATTGCCTACGCATGTATAGGTG
P07631	AOX1flipper_4FRTRseq	CGAGATAGGCTGATCAGGAG
P07636	AOX1flipper_2RnewCYC	CAAAGGAAAAGGGGCCTGTTTATATGCGTCTATTTATGTAG
P07637	AOX1flipper_CYC1F	ACATAAATAGACGCATATAAACAGGCCCCTTTTCCTTTGTCGATATC
P07638	AOX1flipper_CYC1R	GAAACAAGACATTACTGAAGTCGACAACTAAACTGGAATGTGAGG

P07639	AOX1flipper_3FnewCYC	CTCACATTCCAGTTTAGTTGTCGACTTCAGTAATGTCTTGTTTCTTTTG
P08-31	ZEOprobe_EM72F	GACACTTTATACTTCCGGCTCG
P08-32	ZEOprobe_ZEOR	CTGCTCTTCTGCGACGAAATGC
P08-33	AOX1cdsProbeF	TGCTTCTGATTACGATGACTTCC
P08-34	AOX1cdsProbeR	CAAAACCGGATCTTTGCAAAACCAAT
P08-35	AOX2cdsProbeF	GGTGACGCTAACATTCAAAAGAAG
P08-36	AOX2cdsProbeR	GGAAGTAACATGTCCTTCGTTTC
P0879	Kulocus_UpStrm_fwd	AACATGAAAGTAATATGGAACTCCG
P0880	Kulocus_UpStrm2_fwd	AAGCAAAGGGTTGTATAGGC
P0881	Kulocus_DnStrm_rev	TGAATCAGGCAGTCTGCATTCC
P0882	Kulocus_mutATG_rev	GCTTGCTGACAACACTGTATCTTGCAATGCTTTTTATTATTCTC
P0883	Kulocus_mutATG_fwd	GCAAGATACAGTGTTGTCAGCAAGC
P0884	Kulocus_orf_end_rv	CAACATAGGCAATGTGGTGG
P0885	Kulocus_mid_orf_fwd	CACATTCGCAGAGAACATTTACTTGTC
P0886	Kulocus_mid_orf_rev	GACAACTAAATGTTCTCTGCGAATGTG
P0887	Kulocus_kulink5_rev	CCTTTCGTCTTTGGATGTTAGATCTGACAACTAAATGTTCTCTGC
P0888	Kulocus_kulink3_fwd	CAGAGTACAGAAGATTAAGTGAGACCTTCGTTGTCTTTTACAATCCATGACC
P0889	Kulocus_flipend_rev	CGAAGGTCTCACTTAATCTTCTGTACTCTG
P0896	Kulocus_FRT2-KU_f	CTATACTTTCTAGAGAATAGGAACTTCGTATTAGTTTCACTTTTCAGCAAC
P08101	Kulocus_ko_miss_f	GTTCTTCCTGATAAAGCTCC
P08102	Kulocus_ko_mutATG_R	TGCTTGCTGACAACACTGT
P08103	Kulocus_plLV_f	ATCCTTCAGTAATGTCTTG
P08104	Kulocus_KU_end_f	CCCCCACGACAGTAGACC
P08105	Kulocus_KU_5-2_f	GTAGACATTACTTTGATTCCG
P08106	Kulocus_KU_5_1_r	CCAAATTCAGGTCTTGTAAC
P08107	ADE_seq1	TTTCATCGTGTTCACCCTG
P08108	ADE_seq2	CACAGACTGATACCTTTGG
P08109	ARG_seq1	CTCGATTTTGATAGCATCC
P08110	ARG_seq2	AGGGACAACTGATCAATACC
P08111	ARG_seq3	CAACCAGTGAGACCATC
P08112	URA_seq1	CGACATAATTGATGACTTCAC
P08113	URA_seq2	CATGGGCAATTGATCC
P08114	HIS_seq1	GAGTAATTAGAAGAGTCAGCC
P08115	HIS_seq2	TTCTGGATAGGACGACG
P08116	HIS_seq3	ATCTTGGCAGCAGTAACG
P08117	HIS_seq4	TGCTGGGTGTTCCTGC
P08119	Kulocus_KU70+plLV_f	CACATTCGCAGAGAACATTTAGTTGTCGATCCTTCAGTAATGTCTTG
P08134	ARG4flipper_P(AOX)F	GAAGTTCCTATACTTTCTAGAGAATAGGAACTTCAGATCTAACATCCAAAGACGAAAGG
P08135	ARG4flipper_P(AOX)R	CATAATATATCAAATTGTGGCATCGTTTCAATAATTAGTTGTTTTTTGATCTTCTCAAG
P08141	ARG4flipper_5'UTRseqF	GGTTGGATTCATCGTCTTCGTGC
P08142	ARG4flipper_3'UTRseqR	GACAGTTCTATCTACCCGAGGAAACC
P08158	Targeting_ADE_2_fw	GGCACCCTACATAAAGAATC
P08159	Targeting_ADE_3_fw	CCATGTGTCATCGCTTCC
P08160	Targeting_ADzeo1r	GAAGCTATGGTGTGTGGGCCAGTGATGTAACCTCTGACAATGGC
P08161	Targeting_ADzeo1f	GGCCCACACACCATAGCTTC
P08162	Targeting_zeoAD2r	TTGCTCACATGTTGGTCTCC

P08163	Targeting_zeoAD2f	GGAGACCAACATGTGAGCAAAAGGAAGTGCATGGAAAGAGTACAAGAAC
P08164	Targeting_ADE_1_rv	ATGATCATTGTTTACTAATTACC
P08165	Targeting_ADE_3_rv	GCGATTTACCCACTTGG
P08172	ARG4_5‘seqF	GAAAGATGACCGATACTATTGG
P08173	ARG4_3‘seqR	GCTTGTCTGACACATTCACC
P08174	P(AOX)seqR	GAGAAGAGGAGTGGAGGTCC
P08183	HIS4flipper_5H4intR	GAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCTTTATTATCAGTGAGTCAGTCATCAGG
P08184	HIS4flipper_3H4intF	GAAGTTCCTATACTTTCTAGAGAATAGGAACTTCTTATTTAGAGATTTTAACTTAC
P08193	HIS4flipper_5H4intF	CTCCACCAATCAATTCTGGGGATTTGGCTCC
P08194	HIS4flipper_3H4intR	CCTTGACTTTCAGCTGACGTTGGAGTTCG
P08195	HIS4flipper_H4intseqF	GAACAACTGGACTAACACCAGAACCTGC
P08196	HIS4flipper_H4intseqR	CCACATTTCCTACGAACTTGAGTATGGC
P08197	HIS4flipper_H4outseqF	CCAATGAAATTATTCAGCAATCGAGAGC
P08198	HIS4flipper_H4outseqR	CAAATCATCGATTTCACGCTGGTATCC
P08341	pPp_2AMP_ADHTT_R	CATAAGAAATTCGCCCTAGGTTACCAATGCTTAATCAGTGAGG
P08342	pPp_3AMP_ADHTT_F	CACTGATTAAGCATTGGTAACCTAGGGCGAATTTCTTATGATTTATG
P08343	pPp_4PADH_HIS_R	GAAGCAAGGGAAAGGTCATTGTATATGAGATAGTTGATTGTATGC
P08344	pPp_5PADH_HIS_F	CAACTATCTCATATACAATGACCTTTCCCTTGCTTC
P08345	pPp_6HIS_TIFTT_R	GATGTTAACCGGTGCGGCCTTAAATCAAACCAAGCTTCTCC
P08346	pPp_7HIS_TIFTT_F	GAGAAGCTTGGTTTGATTTAAGGCCGCACCGGTTAACATC
P08347	pPp_4PARG_AOXTT_R	GAGCAGGTAAAGCGGTCCTCGAGGGATCCGCACAAACGAAGGTC
P08348	pPp_5AOXTT_PARG_F	GTTTGTGCGGATCCCTCGAGGACCGCTTTACCTGCTCTTG
P08349	pPp_6CDSARG_PARG_R	CTCTCTTCCTGGTTAGACATAGATAGCTGGTAATAAGTTTAGAACAAAAG
P08350	pPp_7PARG_CDSARG_R	CTAAACTTATTACCAGCTATCTATGTCTAACCAGGAAGAGAG
P08351	pPp_8TTARG_CDSARG_R	CAAACTCAGTATAAAACCTATTAGGATTCAAGTTTCTCATTCAAG
P08352	pPp_9CDSARG_TTARG_F	GAATGAGAAACTTGAATCCTAATAGGTTTTATAATGAGTTTGTTAATGATAC
P08353	pPp_10EM72_ARGTT_R1	ATTATACGAGCCGGAAGTATAAAGTGTCAACACCTGTACCGGTTTACAGAAGG
P08354	pPp_P72_ARGTT_XmaR2	AAATTCCCGGGTTTAGTCCTCCTTACACCTTGTCGTATTATACGAGCCGGAAGTATAAAG
P08355	pPp_1AMP_Xmal_F	TAAACCCGGGATGAGTATTCAACATTTCCGTGTC
P08356	pPp_1AMP_Xbal_F	TAAATCTAGAATGAGTATTCAACATTTCCGTGTC
P08357	pPp_8TIFTT_EM72_R	GGAAGTATAAAGTGTCAACACCCTGCAGGACTCGAACCTG
P08358	pPp_9TIFTT_EM72_F	GGTTCGAGTCCTGCAGGGTGTTGACACTTTATACTTCC
P08359	pPp_10EM72_Xbal_R	ATATTCTAGATTTAGTCCTCCTTACACCTTG
P08478	pPp_HIScdsSeqR	GACTTGAAGCTCGGTGGACTGTG
P08479	pPp_HIScdsSeqF	CAGCTCTGGAACCAATCATAC
P08480	pPp_ARGcdsSeqR	CAGTGTAGACCTTTGTACCTTC
P08481	pPp_ARGcdsSeqF	CATCACATTTCTGGTGAATGTGTG
P08482	pPp_AMPcdsSeqR	CAAAAAAGCGGTTAGCTCCTTC
P08579	pPp_Seq1amp	CTGGATCTCAACAGCGGTAAG
P08580	pPp_Seq2amp	GTGACACCACGATGCCTGTAG
P08581	pPp_PAOX800F	CTGTTCTAACCCCTACTTG
P08582	pPp_AOXTTseqF	GTGGTAGGGGTTTGGGAAAATC
P08583	pPp_ARGcdsSeqF2	CTCTTGGTGCTGGAGCACTTG
P08584	pPp_ARGcdsSeqR2	CTCACCAGTGCTGTAGATG
P08585	pPp_ARGTTseqF	CTGACTGTCGTACGGCCTAG
P08586	pPp_PADHseqF	CATCATCATATCGAAGTTTCACTAC

P08587	pPp_HIScdsSeqF2	CTTACTCCTGAGGTCATCTATGTC
P08588	pPp_HIScdsSeqF3	GTTACTCGACGTAAAGGTGATG
P08589	pPp_HIScdsSeqF4	TTCTCTTACCACAGACCGTCCAG
P08697	pPp_ARG4cds1F	CTAACTAAAGACGAACTAAGTGAG
P08698	pPp_ARG4cds1R	CAATGTATTCACGATCAATTCCATAAG
P08699	pPp_ARG5intRs	GTGTGGAACCTCCTTCCACTTG
P08700	HISprobe1F	CACTAGAAGGAAAGGAGATGCCAAG
P08701	HISprobe1R	CTTTGGCAGGAACACCCAGCATC
P08721	pPp_pUCoriF	CTGCGCGTAATCTGCTGCTTGC
P08722	pPp_PAOXstartF	CTAACATCCAAAGACGAAAGGTTG
P08723	pPp_HIScdsSeqR2	CTCAATGCCAAGCAACTTCTGTG
P08763	Targeting_ADE_4_rv	GGACAGTTTTTGAGTTCTTG
P08764	Targeting_HI-1000-r	GCAGGATCAAGTGTTCAGG
P08765	Targeting_HI-250-r	CTATAGAGAGATCAATGGCTC
P08766	Targeting_HI-100-r	GGCTTTGTCACCATTTTG
P08767	Targeting_HI-1000-f	CCTAGATTTGGCAGAAAGAG
P08768	Targeting_HI-640-f	GGCTGACTCTTCTAATTACTCG
P08769	Targeting_HI-250-f	CTTGCAGAAGCTAAATCC
P08770	Targeting_HI-100-f	CCAAGCCAGGATACACC
P08779	Targeting_HI-zeo-1-r	GAAGCTATGGTGTGTGGGCCTTAGAAACGTCAATTTTGC
P08780	Targeting_HI-zeo-2-f	GGAGACCAACATGTGAGCAAAAGGCCTCCTCACAAGAAATTG
P08781	Targeting_UR-zeo-1-r	GAAGCTATGGTGTGTGGGCCTTGATATTGATGCTTGACAG
P08782	Targeting_UR-zeo-2-f	GGAGACCAACATGTGAGCAAAAGGAGGTGTCTACAAGATTGCAC
P08840	ARGprobe5intF	CTATTAGAAGGGTTTACGATGAGGAAG
P08841	ARGprobe5intR	CACCAATAGTATCGGTCATCTTTCTC
P09076	Targeting_UpStrm2_fwd	AAGCAAAGGGTTGTATAGGC
P09077	Targeting_DnStrm_rev	TGAATCAGGCAGTCTGCATTCC
P09078	Targeting_UpStrm2_BamHI_fwd	CGCAGGATCCAAGCAAAGGGTTGTATAGGC
P09079	Targeting_DnStrm_BamHI_rv	CTACCCGGGTGAATCAGGCAGTCTGCATTCC
P09147	Targeting_Upstrm_KU70_rev	GGATGTCGTATTGCTTGCTGAC
P09148	Targeting_Dnstrm_KU70_Mitte_fw	GGTGGATCAATTACGAAAATACG
P09149	Targeting_Dnstrm_KU70_Ende_fw	CAGATGATGCACAGAAACAACG
P09309	ARG4flipper_ARG4locusR3	CTCAGGAGATCCGCATCAGACGAAG
P09310	ARG4flipper_ARG4locusR2	GAGACTCTGTCGACAGTTCTATCTAC
P09311	ARG4flipper_ARG4locusR1	GTACAACGAAGTGCTCTTGTCATACC
P09312	ARG4flipper_ARGlocusF3	CCTGCTCTTGGAGACGTTTACTG
P09313	ARG4flipper_ARG4locusF2	GCTAATTTGGCTGCTGAGAAGGACG
P09314	ARG4flipper_ARG4locusF1	GAATAGTTGAACCCTTGAACGAAGAGG
P09522	pPp_GUTseq1	GATCTGGTGCGAAGCAACAG
P09523	pPp_GUTseq2	GGTACTTTGCCGACTCCTC
P09524	pPp_GUTseq3	GTCTTGCTGCTTGTTTAGTCAC
P09525	pPp_GUTseq4	CAGAACCAACTTCATGAACATTG
P09526	pPp_GUTseq5	CGTCGGACCATTGGCTTC
P09527	pPp_GUTseq6	CTTGGCTGCAGGGAACAC
P09528	pPp_GUTseq7	CTTTCTACTAGATATTCTGGAACTG
P09529	pPp_GUTseq8	CAACAAGCTCCGCATTACAC

P09530	pPp_GUTseq9	GATCAGCCTACTTCGCAG
P09531	pPp_GUTseq10	CCTTAGGATCCTTTTCTCTTCTAC
P09532	pPp_GUTseq11	CCAGAAACGCTGGTGAAAG
P09533	pPp_GUTseq12	CACGATGCCTGTAGCAATG
P09534	pPp_GUTseq13	CATGAGGTCGCTCTTATTGAC
P09535	pPp_GUTseq14	GGTTGGACTCAAGACGATAG
P09536	GUT1disrupt_GUTlocus_1F	CTTTTGCTGGCCTTTTGCTCACAATACCGAAAGGTTAAACAACTTCG
P09537	GUT1disrupt_GUTlocus_1R	CAACCTTT CGTCTTTGGATGTTATTTAAATTGCCAGAGCTGTCACATACTTG
P09540	pPp_3rec_3R	ATTAGTGAGACCTTCGTTTGTGCGCGTTGTATATTCGGTTGGTTTTCC
P09541	pPp_3recPml_3R	TATTAACACGTGCAAGTTGAACTAAAGAACGGAAC
P09542	pPp_AmpPml_4F	AATTACACGTGTTGACACTTTATACTTCCGGCTCG
P09543	pPp_pUCORl_4R	GAAGTTGTTTACCTTTCGGTATTGTGAGCAAAAGGCCAGCAAAAG
P10038	GUT1disrupt_GUTout3prR2	GGTTCTTGATGAAGCTTATATCG
P10039	P(AOX)_124R	GAGAAGAGGAGTGGAGGTC
P10040	GUT1disrupt_GUTout3prR1	ATGAAGTTAGTAAGGTTCTTGATGAAGC
P10041	GUT1disrupt_GUTout5prF1	CGCTCCTGACTGTTTCAAGTC
P10042	GUT1disrupt_GUTout5prF2	GCATTGTTCTTTGAAATCGAAATTGG
P10529	KuProbeF	CAATCCATGACCAAAAAATCCAAG
P10530	KuProbeR	CAATTTTGGGTGGCAGCTG
-	Adaptor primer 1	GTAATACGACTCACTATAGGGC
-	Adaptor primer 2	ACTATAGGGCACGCGTGGT
P11093	Gene specific primer 1	CCAAACCTTTAGTACGGGTAATTAACG
P11095	Gene specific primer 2	GTCCTCCACGAAGTCCCGG
P11075	ARG4flipper_ARG5intF1	ATCAAAATTGAAGATGACTTACTTGATAACATCC
P11076	ARG4flipper_ARG3intR1	TGAGTCATTACCGGAAGCTAGAAC
P11077	ARG4flipper_ARG5intF2	CTTATTACCAGCTATCTATACTCGAATCAAGAAGAAGGAC
P11078	ARG4flipper_ARG3intR2	ATCAAAATTGAAGATGACTTACTTGATAACATCC
P11079	ARG4flipper_ARG5seq1F	CGTACGACAGTCAGTTAGTAG
P11080	ARG4flipper_ARG5seq1R	TACCCTGATTAGATAATACAATAACCAAC
P11081	ARG4flipper_ARG5seq2F	CAAGTTGGCAGATGCTTATTCTAC
P11082	ARG4flipper_ARG3seq2R	CTAGGCCGTACGACAGTCAG
P11083	ARG4flipper_ARG5intR1	GAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCAGGTTTTATACTGAGTTTGTTAATGATAC
P11084	ARG4flipper_ARG3intF1	GAAGTTCCTATACTTTCTAGAGAATAGGAACTTCAGATAGCTGGTAATAAGTTTAGAACAAAAG
P11085	ARG4flipper_ARG5intR2	GAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCCATATCTTTGTTATAGGTTGAC
P11086	ARG4flipper_ARG3intF2	GAAGTTCCTATACTTTCTAGAGAATAGGAACTTCAGGTTTTATACTGAGTTTGTTAATGATACAATAAAC

Table S2. The origins and functions of the *E. coli*/*P. pastoris* shuttle vector components.

Element	**Origin**	**Function**
P_AOX1Syn_ dBamHI	Synthetic	Part of *Pichia pastoris AOX1* promoter
P_AOX1Syn	Synthetic	*AOX1* promoter for methanol induced expression of the target gene in *Pichia pastoris*
P_GAP	CBS7435, *Pichia pastoris* strain	*GAP1* promoter for constitutive expression of the target gene in *Pichia pastoris*
P_ADH1	BY4741, *Saccharomyces cerevisiae* strain	*ADH1* promoter for expression of respective antibiotic resistance gene in *Pichia pastoris*
P_ILV5	CBS7435, *Pichia pastoris* strain	*ILV5* promoter for expression of respective antibiotic resistance gene in *Pichia pastoris*
P_TEF1	BY4741, *Saccharomyces cerevisiae* strain	*TEF1* promoter for expression of respective antibiotic resistance gene in *Pichia pastoris*
P_EM72 Syn	Synthetic consensus sequence *E. coli* promoter	Constitutive promoter for expression of respective antibiotic resistance gene in *E. coli*
P_ARG4	CBS7435, *Pichia pastoris* strain	*ARG4* promoter for expression of *ARG4* gene
AOX1TTSyn	Synthetic	*AOX1* terminator for target gene transcription termination in *Pichia pastoris*.
TIF51ATT	BY4741, *Saccharomyces cerevisiae* strain	*TIF51A* terminator for target gene transcription termination in *Pichia pastoris*.
ADH1TT	BY4741, *Saccharomyces cerevisiae* strain	*ADH1* terminator for target gene transcription termination in *Pichia pastoris*.
AODTT	CBS7435, *Pichia pastoris* strain	*AOD* terminator for target gene transcription termination in *Pichia pastoris*.
CYC1TT	BY4741, *Saccharomyces cerevisiae* strain	*CYC1* terminator for target gene transcription termination in *Pichia pastoris*.
ARG4TT	CBS7435, *Pichia pastoris* strain	*ARG4* terminator for target gene transcription termination in *Pichia pastoris*.
Zeocin Syn	Synthetic, *Sh ble* codon optimized :Leto, Entelechon®, mixed codon usage (*E. coli*, *P. pastoris*)	Zeocin resistance, selection marker in *E.coli* and *Pichia pastoris*
KanMX6	KanMX6 (60, 61)	Kanamycin and Geneticin resistance in *E. coli* and *Pichia pastoris*, respectively; selection marker
BLAcds	*β-lactamase* gene from pUC8	Ampicillin resistance in *E. coli*, selection marker
pUC origin	pBR322	pUC replication origin for *E.Coli*
Alphafactor	Synthetic, codon optimized :Leto, Entelechon®, based on *Saccharomyces cerevisiae* prepro alpha-factor (MF alpha-2)	Secretion signal sequence
ARG4_optimized	Synthetic, codon optimized: Gene Designer, Leto (Entelechon®); *Pichia pastoris*, *Yarrowia lipolytica*, *Schizosaccharomyces pombe* average codon usage leaving out rarest codons from all 3 organisms	*Pichia* wild-type gene coding for *argininosuccinate lyase*, selection marker
HIS4_optimized	Synthetic, codon optimized: Gene Designer,	*Pichia* wild-type gene coding for

	Leto (Entelechon®); Synthetic *Pichia pastoris*, *Yarrowia lipolytica*, *Schizosaccharomyces pombe* average codon usage leaving out rarest codons from all 3 organisms	trifunctional *HIS4*, selection marker
KU70	CBS7435, *Pichia pastoris* strain	Homologue of *S. cerevisiae HDF1* in *Pichia pastoris*
FRT	Synthetic minimal FRT site (62)	FLP recombinase target sequence
FLP recombinase	BY4741, *Saccharomyces cerevisiae* strain	Site-specific recombinase, breakage and joining of four DNA strands between two target sequences
PDI704	CBS704, *Pichia pastoris* strain	Protein Disulfide Isomerase, Chaperone
Syn*PDI*	Synthetic, based on PDI from *Pichia pastoris* strain X-33, codon optimized with Leto 1.0, Entelechon®	Protein Disulfide Isomerase, Chaperone

Table S3. Estimation of the copy number of the GFP expression cassette. Copy numbers were calculated according to absolute (Abs. Q) and relative (Rel. Q) quantification. Q.1 = quantification before methanol induction, Q.2 = quantification after four rounds of cultivation and induction. No significant changes could be detected in the copy numbers of the wt, mut^s and Δ*KU70* strains. GFP production is shown as relative fluorescence units (RFU) normalized by OD.

Strain	Abs Q.1	Rel. Q.1	Copies	GFP(RFU)*OD-1	Abs Q.2	Rel. Q.2	Copies
CBS7435wt	4	4	4	2966	4,7	4,3	4-5
CBS7435mut^s	4,8	5	~5	3876	4,9	4,4	4-5
Δ*KU70*/1	4,4	4,6	4-5	1763	4,7	4,5	4-5
Δ*KU70*/2	7,5	7,5	7-8	3046	7,2	6,7	~7

Table S4. Southern blot analysis of the knock-out loci *AOX1* and *HIS4*. Southern blot analysis was performed to verify the expected knock-outs and to define the location and number of the excision cassettes integrated in the targeted genomes. Fragment = size of the hybridizing fragment in the analysis. Probes targeted to the coding sequence (cds) and Zeocin^{TM} resistance cassette (zeo) were used to detect the wild-type (wt) locus, the location and number of Zeocin^{TM} resistance cassettes in the before induction strain still carrying the excision cassette in the targeted locus (flipper in) and the expected knock-out and removal of the Zeocin^{TM} resistance cassette in the final strain after FLP recombinase induction (knock-out).

Locus	Enzyme	Probe	Fragment/wt	Fragment/flipper in	Fragment/knock-out
AOX1	NdeI	cds	9000bp	-	-
AOX1	NdeI	zeo	-	6786bp	-
AOX1	SspI	cds	5520bp	-	-
AOX1	SspI	zeo	-	963bp	-
HIS4	BglII	cds	2650bp	-	-
HIS4	DraI	zeo	-	2283bp	-

CHAPTER 2

Peroxidase gene discovery from the horseradish transcriptome

Laura Näätsaari[1], Florian W. Krainer[1], Michael Schubert[1], Gerhard G. Thallinger[2,3] and Anton Glieder[3]

[1] Institute of Molecular Biotechnology, Graz University of Technology, Petersgasse 14, A-8010 Graz, Austria

2 Institute for Genomics and Bioinformatics, Graz University of Technology, 8010 Graz, Austria

[3] Austrian Centre of Industrial Biotechnology (ACIB GmbH), Petersgasse 14, A-8010 Graz, Austria

[*] To whom correspondence should be addressed. Tel: +43 316 8739300; Fax: +43 316 873 9301;
Email: glieder@glieder.com

Manuscript prepared for submission in “Nucleic Acids Research”

Parts of the study described in this chapter were performed in co-operation with Florian Krainer. Thus parts of the manuscript, especially in the methodological section, might have similarities to his Master’s Thesis.

ABSTRACT

Horseradish peroxidases (HRPs) from *Armoracia rusticana* have long been utilized as reporters in various diagnostic assays and histochemical stainings. Regardless of their increasing importance in the field of life sciences and suggested uses in medical applications, chemical synthesis and other industrial applications, the HRP isoenzymes, their substrate specificities and enzymatic properties are poorly characterized. Due to lacking sequence information of natural isoenzymes and the low levels of HRP expression in heterologous hosts, commercially available HRP is still extracted as a mixture of isoenzymes from the roots of *A. rusticana*. In this study, a normalized, size-selected *A. rusticana* cDNA library was sequenced using 454 pyrosequencing. Sequence databases, ORF finding and ORF characterization were utilized to identify peroxidase genes from the 14871 isotigs generated by a *de novo* assembly, resulting in the discovery of 19 secretory peroxidases. The sequences were manually reviewed and verified with Sanger sequencing of PCR amplified genomic fragments. A total of 22 isoenzymes including allelic variants were successfully produced in *Pichia pastoris* and showed peroxidase activity, thus enabling their development into commercial pure isoenzymes. This study demonstrates that transcriptome sequencing combined with sequence motif search is a powerful concept for the discovery and quick supply of new enzymes and isoenzymes from any plant or other eukaryotic organism.

Keywords: peroxidase, transcriptome sequencing, *Pichia pastoris*, 454 sequencing

INTRODUCTION

Horseradish peroxidases (HRPs) originating from the perennial herb *Armoracia rusticana* (*Brassicaceae*) are heme-containing monomeric glycoproteins belonging to the class III plant peroxidase subfamily (1).These versatile enzymes have traditionally been utilized as reporters in various diagnostic assays and histochemical stainings but have also gained increasing interest in other life science and biotechnological applications ranging from cancer therapeutics (2), protein engineering (3) and transgenics (1) to bioremediation (4), biosensors (5) and biocatalysis (6). The *in vivo* functions of HRPs have not been fully elucidated owing to the estimated large number of isoenzymes (up to 42) (7), but are known to be very diverse (8) thus offering a wide range of substrate specificities and applications. Although HRP has been studied for decades and in spite of the large diversity of this enzyme family, protein engineering and heterologous expression have mainly been focused on one single isoenzyme C1A, thus neglecting the potential of all others. This was largely due to the lack of sequence information and the low efficiency of HRP expression in heterologous hosts. Commercial preparations are still extracted from the roots of *A. rusticana* and therefore consist of a mixture of various isoenzymes. The quality of these preparations varies greatly and depends on several biotic and abiotic factors, such as seasonal change or origin. Chromatographic purification is needed to isolate highly enriched isoenzyme fractions. Only very few purified isoenzymes were accessible so far and their substrate specificities and enzymatic properties are poorly characterized.

For only six isoenzymes (C1A, C1B, C1C, C2, C3, N), the nucleotide sequences have previously been published (9–11). In addition, the amino acid sequences of isoenzymes A2 (P80679) (12) and E5 (P59121) (13) have been determined. However, for decades, HRP has been known as a large group of enzymes with versatile physical properties. Already in 1966, Shannon *et al.* (14) described the physical properties of seven isoenzymes. Aibara *et al*. (15, 16) characterized five neutral isoenzymes (B1, B2, B3, C1 and C2) and six basic isoenzymes (E1-E6) in 1982 and 1981, respectively. A total number of 42 peroxidase isoenzymes have been identified by isoelectric focusing in commercially available HRP preparations (7), without knowing whether these isoenzymes differ in amino acid sequence or due to different posttranslational modifications.

This study has two major goals. First, we report the transcriptome sequence of *A. rusticana*, a perennial plant of industrial, medical and culinary importance. The availability of a large expressed sequence tag (EST) collection is crucial to support annotation in possible future *A. rusticana* genome sequencing projects. Secondly, we demonstrate the use of an efficient enzyme discovery pipeline including new generation cDNA sequencing technologies, *in silico* isoenzyme discovery and experimental sequence verification, gene synthesis and enzyme production and secretion by *Pichia pastoris* as a straightforward approach to discover and characterize new isoenzymes from plants or other eukaryotes. Transcriptomes deliver all sequences of expressed genes, at the same time avoiding sequencing introns and providing information about all expressed exons and alternative exon junctions. Studies in *Arabidopsis thaliana*, a model species of the *Brassicaceae* family, have demonstrated the power of massively parallel transcriptome sequencing in providing high-quality representation of transcripts needed for gene discovery (17). Similar transcriptome sequencing approaches have previously been applied, for example, in marker development, population genomics (18) and predictions of biosynthetic pathways (19–21). However, the concept of novel isoenzyme discovery from the large bulk of sequences generated by next generation sequencing (NGS) technologies needs a full pipeline of efficient tools. This is the first study using NGS transcriptome sequencing to discover, discriminate and characterize large numbers of sequence verified isoenzymes of non-model plant origin. Although a similar study was recently performed to identify fungal cellulases (22) the method described utilized combined secretome and transcriptome analyses and was only aiming to show cellulose activity of the cloned cDNA without the need of the full verified sequences to make all discovered isoenzymes available by recombinant expression. The method described in this study can be widely applied for the replenishment of the sequence data in any eukaryotic organism including fungi, plants and animal cell lines or tissues when detailed sequence and gene structure information of enzymes and isoenzymes is needed.

MATERIALS AND METHODS

Plant specimens, RNA extraction and quality analysis

Wild horseradish (*A. rusticana*) roots were purchased from local farmers and grown in the laboratory to obtain fresh roots, sprouts, stems and leaves. Tissues were collected in aliquots, frozen in liquid nitrogen and stored at -80°C. Total RNA from all available plant parts was isolated using RNaqueous kit (Applied Biosystems/Ambion, Austin, TX, USA) according to the manufacturer’s recommendations. Quality assessment to ensure RNA integrity was performed with Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

Normalized cDNA library construction and sequencing

Transcriptome sequence was obtained by a commercial service from LGC Genomics (Berlin, Germany). The methods used are roughly summarized as follows: mRNA was purified from total RNA using mRNA-ONLY™ Eukaryotic mRNA Isolation Kit (Epicentre, Madison, WI, USA). One µg of mRNA was used for first-strand cDNA synthesis and amplification according to the Mint-Universal cDNA Synthesis Kit user manual (Evrogen, Moscow, Russia), followed by a normalization reaction using the Trimmer Kit (Evrogen). Normalized material was re-amplified, digested (SfiI), size-selected (>800bp, LMP agarose), purified (Qiagen, Hilden, Germany) and ligated to pDNR-lib vector (Clontech, Saint-Germain-en-Laye, France) using the Fast Ligation Kit (New England Biolabs, Ipswich, MA, USA). The desalted ligation was used to transform NEB10b competent cells (New England Biolabs).

Roughly a million clones were plated on LB + chloramphenicol (Cm) plates, scraped off the plates and stored as glycerol stocks at -70°C. Plasmid DNA was prepared using standard methods

(Qiagen, Hilden, Germany), and digested with SfiI. cDNA inserts were gel-purified (LMP-Agarose/MinElute Gel Extraction Kit, Qiagen) and ligated to high-molecular-weight DNA using a proprietary SfiI-linker.

Library generation for the 454 FLX sequencing was carried out according to standard protocols (Roche/454 life sciences, Branford, CT 06405, USA). In short, the concatenated inserts were sheared randomly by nebulization to fragments ranging in size from 400bp to 900bp. These fragments were end polished and the 454 A and B adaptors that are required for the emulsion PCR and sequencing were ligated to the ends of the fragments. The resulting fragment library was sequenced on half a picotiterplate on the GS FLX using the Roche/454 Titanium chemistry.

Assembly of the sequence reads to transcripts

Raw reads produced by the pyrosequencing process were screened for the SfiI-linker that was used for concatenation and the linker sequences were clipped from the reads. Poly A/T sequences were mostly (~90%) removed with the linker. High-quality reads were selected using Newbler sequence filtering at default settings. The clipped, quality controlled reads were assembled into individual isotigs using the Roche/454 Newbler software at default settings (454 Life Sciences Corporation, Software Release: 2.5.3).

Discovery of peroxidases in the assembled contigs

The PSSM profile corresponding to all known horseradish peroxidases was obtained from NCBI's Conserved Domain Database (cd00693, CDD v3.01) (23). It was used in a tblastn search (e-value 1e-5) against a nucleotide BLAST database containing the assembled contigs to yield a preliminary set of HRP candidate sequences. This set was refined by discarding sequences whose translation mapped back to a domain other than the original profile in an rpsblast classification step or had a bit score lower than the NCBI specific hit threshold therein. The read composition of the refined set of contigs was manually reviewed using SeqMan (DNASTAR, Madison, Wisconsin, USA) and the ACE file produced by the Newbler assembler. Isotigs reflecting two different variants assembled into one contig by the assembler were split and discarded read ends were incorporated if they were concordant and part of the protein coding regions.

Genome walking and manual verification of horseradish peroxidase sequences

The sequences of the identified peroxidase genes were manually verified by Sanger sequencing of PCR amplified genomic fragments using primers listed in supplementary table 1. If no flanking regions were available for primer design, genome walking (24) was utilized to clarify and complete the sequences of the C- and N-termini. Therefore, 2µg aliquots of genomic DNA were singly digested with Bsp143I, HindIII, PsuI (Fermentas, St. Leon-Rot, Germany), BsaWI (New England Biolabs GmbH, Frankfurt am Main, Germany) or XhoII (Promega GmbH, Madison, WI, USA) in order to get fragments of 1 - 5kb size. The digestion was stopped by heat inactivation of the enzymes, the fragmented DNA was precipitated with ethanol and the pellet was dissolved in 30µL of distilled water. An adaptor was created by annealing adaptor strand 1 (5'-GTAATACGACTCACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGT-3') either to adaptor strand 2.a (3'-TCCCCGACCACTAG-5') for Bsp143I-/PsuI-/XhoII-digested DNA, 2.b (3'-TCCCCGACCATTAA-5') for BsaWI-digested DNA or 2.c (3'-TCCCCGACCATCGA-5') for HindIII-digested DNA.

In the annealing reaction, adaptor strand 1 was mixed in 1:1 molar ratio with adaptor strand 2.a/2.b/2.c (i.e. 13.7µL of 100µM adaptor strand 1 + 4.0µL of 100µM adaptor strand 2) and heated to 95 °C for 5 min. To anneal the two strands to a functional adaptor molecule, the mixture was allowed to slowly cool down to room temperature. The three differently annealed adaptors (1+2a, 1+2b, 1+2c) were ligated for 3 hours at room temperature with T4 DNA Ligase (Fermentas) to the digested DNA fragments, considering the specific 5' overhangs that have been created by the respective restriction enzymes. The ligation reaction was stopped by incubation for 5min at 70°C and 70µL of TE buffer were added. Two gene-specific primers and two adaptor primers were designed. The gene-specific primers were designed to bind approximately 100bp from the end of the known sequence, considering that no restriction site of the restriction enzymes used for DNA digestion laid between the primer-binding site and the end of the known sequence. AdaptorPrimer1 (5'-GTAATACGACTCACTATAGGGC-3') and GeneSpecificPrimer1 were used as a primer pair for a first PCR with 1µL of the DNA+adaptor ligation product as template DNA. One µL of the first PCR mix was used as template for a second PCR with AdaptorPrimer2 (5'-ACTATAGGGCACGCGTGGT-3') and GeneSpecificPrimer2. This second primer pair was designed to bind within the first PCR product. Both PCR steps were performed with an elongation time of 50s.

A gene-specific DNA fragment as product from the second PCR was isolated from a preparative agarose gel, purified (SV DNA extraction kit, Promega) and sent to Sanger sequencing (LGC Genomics GmbH, Berlin, Germany), using AdaptorPrimer2 and the corresponding GeneSpecificPrimer2. If any doubt of unspecific primer binding or polymorphisms existed, the PCR products were cloned into the pJET 1.2blunt vector (GeneJet cloning kit, Promega), transformed to *E. coli* Top10 F' and plasmids from single colonies were isolated for sequencing to ensure the read consisted of only one allele.

The resulting gene sequences were submitted to EMBL (table 1).

Codon usage, GC content, isoelectric point, signal sequence prediction, disulfide bridge prediction and phylogenetic analyses of the horseradish peroxidase isoenzymes

Codon usages and GC contents of the HRP isoenzymes were analyzed using CAIcal (25) (http://genomes.urv.es/CAIcal/) and Mega5 (26) (http://www.megasoftware.net/). Sequences were aligned with ClustalW2 (27) (http://www.ebi.ac.uk/Tools/msa/clustalw2/). The theoretical isoelectric points (pI) were calculated with ExPASy Compute pI/Mw tool (http://web.expasy.org/compute_pi/). Disulfide bridges were predicted with EDBCP tool (Ensemble-based Disulfide Bonding Connectivity Pattern prediction server, http://biomedical.ctust.edu.tw/edbcp/) (28, 29). Phylogenetic analyses were performed with "Phylogenetic Tree" online tool (http://www.cbrg.ethz.ch/services/PhylogeneticTree) and Mega5. Signal sequences were predicted using SignalP 3.0 (30) (http://www.cbs.dtu.dk/services/SignalP/).

Gene synthesis and heterologous expression in *Pichia pastoris*

The codon usages of 14 isoenzymes were optimized for the expression in *P. pastoris* using a novel algorithm (DNA2.0, Menlo Park, CA, USA, Mellitzer *et al.* manuscript in preparation). Further 12 isoenzymes including allelic variants were optimized using GeneDesigner 1.1.4.1 (DNA2.0) in accordance to the *P. pastoris* codon usage described by Abad *et al.* (31). Signal sequence variants were generated by PCR amplification and all HRP genes were cloned into the shuttle vector pPpT4_alpha_S of a newly generated open source expression platform (Näätsaari *et al*., submitted). The vector pPpT4_alpha_S is a basic low-copy (1-5 copies/genome), zeocin resistance based expression vector for efficient secretory expression of heterologous proteins. Sanger sequencing of the plasmids verified successful cloning into the right frame. The linearized expression cassettes were transformed into *P. pastoris* wild-type CBS7435 based mut^s strain using standard protocols (32), and selected on zeocin-containing plates. From each gene, 88 clones were picked to 96-well deep-well plates for cultivation and high-throughput screening of peroxidase activity. Two of the well expressing clones of each isoenzyme were streaked out to single colonies. Four single colonies of each clone were used for re-screening to estimate the reproducibility of the results. All media compositions and cultivation protocols used in this study were as previously described by (33). Minimal media BMD1% was supplemented with 5mM ferrous sulfate heptahydrate (Sigma-Aldrich Handels Gmbh, Vienna, Austria) to ensure sufficient iron supply for heme biosynthesis.

Peroxidase assays

ABTS (2,2'-azino-bis(3-ethylbenzthiazoline-6-sulfonic acid), TMB (3,3',5'5-tetramethyl benzidine), pyrogallol (1,2,3-Trihydroxybenzene) and guaiacol (2-methoxyphenol) assays were used to detect peroxidase activity essentially as described in (3, 34, 35). ABTS assays were performed in 50mM sodium acetate buffer pH4.5 with 1mM ABTS and 0.0026% (v/v) H_2O_2. For the TMB stock solution, TMB was dissolved in DMSO to a concentration of 4.16mM. For the assay solution, TMB stock solution and 30% (v/v) H_2O_2 were diluted with 20mM citrate buffer pH 5.5 to final concentrations of 0.416mM and 0.006% (v/v), respectively. Guaiacol assays were performed in 10mM sodium phosphate buffer pH7.0 with 5mM guaiacol and 0.0009% (v/v) H_2O_2. For the pyrogallol assay solution, pyrogallol (Sigma-Aldrich Handels Gmbh, Vienna, Austria) was dissolved in 10mM potassium phosphate buffer pH6.0 containing 0.027% (v/v) H_2O_2 to a concentration of 45mM. For all assays, 15µl cultivation supernatant was mixed with 140µl of the assay solution in a flat-bottom 96-well microtiterplate (Greiner Bio-One GmbH, Frickenhausen, Germany). The reaction kinetics were followed with Spectramax Plus^{384} spectrophotometer and SoftMax® Pro software (Molecular Devices, LLC) for 3-5min at wavelengths 405nm (ABTS), 650nm (TMB), 470nm (guaiacol) and 420nm (pyrogallol). Enzyme activity was calculated using only time points fitting to linear increase of the absorbance (ΔmAU min^{-1}).

RESULTS

Sample preparation and cDNA library generation

The high quality (RIN 9.4-9.8) of the RNA samples was confirmed with an Agilent 2100 bioanalyzer. In order to include the majority of all coded isoenzymes, a mixture of RNA from diverse plant parts, including leaves, roots, sprouts and stems, was chosen for mRNA isolation, cDNA synthesis, normalization, size selection and cloning (performed by LGC Genomics, Berlin, Germany). The normalized cDNA library was subjected to quality control experiments before using it for 454 pyrosequencing: a cDNA fragment size of over 800bp was ensured and the normalization efficiency was verified by sequencing 96 randomly selected clones (LGC genomics).

Sequencing and *de novo* assembly

The normalized cDNA library was sequenced on half a picotiterplate run on the GS FLX using Roche 454/Titanium chemistry. A total of 592507 sequence reads with an average read length of 353±122 nucleotides (range 719bp) were obtained. A total of 2.16% of clonal reads (exact, 3’ or 5’) were detected. Prior to assembly, the sequence reads were screened for the linker sequence used for concatenation, the linker sequences were clipped and the reads were quality checked (LGC Genomics). The resulting 556269 reads with an average length of 343bp (Figure 1A) were further filtered by Newbler sequence filtering to ensure consistent high quality of the reads used in the assembly. 490285 reads were aligned to individual transcripts using Newbler assembler at default settings. The *de novo* assembly generated 18511 contigs with an average length of 718bp (Figure 1B) and an average coverage of 11.4-fold (Figure 1C). The contigs were further processed to 14871 isotigs with an average length of 1133bp (Figure 1D). 35950 reads were left as singletons. A detailed summary of the alignment and assembly process is described in Table 2.

Plant peroxidase search and manual validation of assembled transcripts

The settings used for the initial tblastn search allowed for a very permissive filtering of putative peroxidases, thus including many false positives, but impeding the loss of valuable data for further analyses. This search yielded hits in 91 transcripts, which were classified in secretory peroxidases, ascorbate peroxidases, GSH peroxidases and peroxidase-like proteins by definition of the Conserved Domains Database (CDD, NCBI). All previously known horseradish peroxidases were classified as secretory peroxidases, so only the contigs comprising a secretory peroxidase conserved domain were kept. The horseradish transcriptome contigs of the 18 resulting secretory peroxidases were manually reviewed. In this process, the coding sequences of four contigs were extended with available assembly data, three contigs were split because of strongly conflicting reads, and two more contigs were discarded because of only a partial domain match that could not be resolved into a full-length sequence. In total, 18 peroxidase genes were identified in the transcriptome of *A. rusticana*, all featuring a full secretory peroxidase domain. In addition, isoenzymes C1A and C3 could be partially retrieved from the raw reads although they did not form a full-length contig. No read even partially corresponding to the previously published “neutral isoenzyme” N (Q42517) was found.

Sanger sequencing and genome walking

Sequences yielded by the transcriptome assemblies are not necessarily error free but can include incorrect information either caused by errors in the transcription machinery of the plant, introduced in the sequencing process or resulting from misassemblies. To estimate the reliability of the transcript sequences, the sequences of all peroxidase genes detected in the isotigs were verified on genome level by Sanger sequencing of amplified genomic DNA. In addition, the sequences of the isoenzymes C1A and C3 available in the databases and partially also found in the raw reads were revised. In case of five full-length contigs where no sufficient read data from untranslated regions was available to enable amplification and sequencing of the complete gene, a genome walking approach was successfully performed in order to verify the 5’ and/or 3’ regions of the respective gene. Divergent coding sequence information was observed for 10 isoenzymes in the form of possible allelic variants. PCR artifacts were ruled out by repeated experiments to increase the coverage of the positions. Thus, conflicting sequence information was postulated not to be due to sequencing errors, but rather due to the high sequence similarity of the HRP isoenzymes and the permissive settings used in the assembly. Supporting this assumption, putative allelic sequences could be found as separate raw reads. For altoghether 9 positions in contigs 22684, 06117, 17517, 23190, 04663 (Table 3) the nucleotide present in the transcriptome reads could not be found in the genomic DNA sequenced. The sequence of the previously published “neutral isoenzyme” N (Q42517) not present in the transcriptome sequences could not be amplified from gDNA either in the previously published form. Therefore, it was not included in the following analyses or experiments. The sequence of the transcript # 22489 (see Table 1) could not be verified.

GC content and codon usage

The average GC content of all 14871 isotigs was calculated to be 42.7% (Range 28%-62%) (Figure 4), almost identical to the average GC content, 42.5%, reported for *A. thaliana* (36). This is lower than the average GC content of 45.1% reported by the Codon Usage Database (http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=3704), suggesting that the small set of available *A. rusticana* genes (14 CDS) used in the calculation is not representative for the whole population. The GC content of the HRP isoenzymes was observed to vary between 42.9% (C2, contig #04627) and 51.0% (contig #22489), with an average GC content of 47.1%. The results of the analysis are closer described in table 1.

The codon usages of the HRP genes were compared in the form of a heatmap (Figure 3), depicting the fold change of the codon usage frequencies compared to the expected (1/64) frequency (ΔRSCU). The clustering of the isoenzymes according to their codon usage frequencies situated newly discovered isoenzymes with most divergent sequence and gene structure (03523, 05508, 22489, 17517, 23190) also furthers away from the previously known group C isoenzymes.

Gene structure and phylogenetic analyses

Phylogenetic relationships of the HRP isoenzymes are shown in figures 2 and 3. Interestingly, the previously known isoenzymes seem to be closely related to each other, while most of the new isoenzymes discovered in the transcriptome seem to share higher evolutionary distance to them. From the 20 peroxidase gene loci, 15 were confirmed to have three introns by comparing either transcript data or protein sequence data to the verified gDNA sequence. Further four genes (05508, 22489, 17517, 23190) were noted to have only two introns and one gene (03523) no introns (Table 1). The number of introns was correlating with the evolutionary distance so that genes having aberrant intron numbers were situated in separate branches close to each other in the phylogenetic topology. With the information obtained from the reads, no alternative splicing could be shown. According to SignalP, all of the isoenzymes have a signal sequence varying in length from 18 amino acids to 31 amino acids. The lengths of the signal sequences are described in Table 1, and an alignment of the amino acid sequences of the HRP isoenzymes is shown in Supplementary Figure 1.

Heterologous protein production in *Pichia pastoris*

A total of 26 horseradish peroxidase sequences including allelic variants were codon optimized for *P. pastoris* expression. From the 26 isoenzymes ordered as synthetic fragments, 22 were successfully expressed in *P. pastoris*. The peroxidase activities of the produced isoenzymes were detected with four substrates having variable assay pH optima. The substrate or pH-specific performance (Table 4) of the isoenzymes suggests a wide range of possible applications for this versatile group of peroxidases.

DISCUSSION

Although high throughput sequencing technologies and bioinformatics tools to handle the enormous amount of data generated have been rapidly developing in the recent past, the expressed sequence data of many organisms of wide importance are still not available. Our study demonstrates how NGS technologies can provide a rapid, low-cost basis for the discovery of isoenzymes required for specific industrial, medical or biological applications. Below we discuss the reliability of the approach in identifying and characterizing an important group of isoenzymes, challenges provided by the library generation, sequencing and assembly methods, and suitability of the data obtained from the pipeline for heterologous protein production without laborious manual verification of the sequences.

A normalized cDNA collection originating from multiple tissues was sequenced with 454 pyrosequencing. Studies in *A. thaliana* suggested that with the high coverage attainable by massively parallel sequencing, all transcripts can be well represented in the sequence data regardless of expression levels (17). However, the benefits of normalization in non-model species have not been well characterized. Since also the genome and transcriptome sizes of *A. rusticana* and the sequence data required to reach also isoenzymes with low expression levels are unknown, the cDNA library sequenced in this study was normalized. Normalization of the cDNA has been reported to be especially important in gene discovery when the cDNA used for sequencing is pooled from many tissues or individuals (37–39), and to considerably reduce the frequency of abundant transcripts thus increasing the possibility to reach also unique transcripts of isoenzymes with low expression levels. Half a plate run on a GS FLX platform resulted in over 500000 high-quality reads, corresponding to a relatively high average coverage (>10x) of the assembled 14871 isotigs with a high average length of 1133bp. Comparable to previous transcriptome sequencing studies (38, 39), 88% of the reads could be assembled into contigs. Many of the remaining singletons were of high quality and also represent an important source of information (18). Singletons could either result from 454 sequencing errors or contaminants from plant parasites (40), they could be caused by efficient normalization methods, or simply represent rare transcripts with thin coverage despite the normalized cDNA pool used for sequencing.

The 454 pyrosequencing technique generates relatively long reads including very few technical errors (mainly related to homopolymer runs), and is therefore well-suited for applications such as *de novo* transcriptome sequencing. Although the sequence length achieved by 454 pyrosequencing is still clearly shorter than by traditional Sanger sequencing, it has been reported to be adequate for reconstructing full length transcripts (17) and validated to be comparable in accuracy to Sanger sequencing (41, 42). The high average isotig length of 1133bp and the assembly of 90% (19/21) full length peroxidase isotigs reached in this study supports the statement of adequate read length and coverage to detect complete transcripts. Only isoenzymes C1A and C3 were not assembled to a contig due to low sequence coverage. Although the error rates associated with NGS methods have been reported to be low, they could still cause problems in reliable sequence polymorphism detection. The requirement of >90% match used in this study, combined with a minimal match length of 40bp was expected to provide a very high number of contigs without collapsing and joining similar isoenzymes to one contig (18). However, the isoenzyme sequences were known to be partially almost identical. To validate the assembly and ORF prediction correctness, and the existence of allelic variants, the isoenzyme sequences were amplified from genomic DNA. The combination of transcriptome sequencing and Sanger sequencing of amplified genomic DNA revealed 23 variable positions in the coding sequences of 9 genes. For 9 positions thereof (in 4 transcripts), the corresponding nucleotide could not be found in the genomic DNA. Manual confirmation of the transcriptome reads revealed that the coverages of all positions were high and the reads agreed. Therefore, sequencing mistakes could be estimated not to be the cause for the differences. The differences could be caused either by mistakes made by the reverse transcriptase enzyme during library generation, or reflect the natural variation between individual plants and plant parts, since the RNA used for the library generation did not originate from the same parts of the same individual as the genomic DNA used for manual sequence verification. The general error rate of the sequencing technique was noted to be very low, but isoenzymes coded by less common allelic variants would have been missed without manual sequence verification. Twenty-six of the resulting coding sequences, including allelic variants missed in the transcriptome sequencing and assembly processes, were codon optimized and transformed to *P. pastoris* for expression.

This study reveals that for the coverage of all isoenzymes including allelic variants represented by the cDNA library sequenced, manual work to verify the resulting transcript sequences cannot be avoided. However, the allelic variants represent only a minor part of the newly discovered enzymes. The quality of the sequences is very high and differences to genomic DNA minimal, confirming that the enzyme discovery method described in this study for high-throughput applications would not necessarily require manual verification of the sequencing by laborious Sanger sequencing of amplified genomic DNA. However, manual curation of the contigs of interest and splitting of the data in contigs with clear assembly conflicts can be done and could be worthwhile, since especially the additive effects of amino acid changes in collapsed contigs could cause problematic changes in the enzyme structure.

The primary enzyme discovery pipeline utilized in this study provides a functional approach to find proteins of interest for heterologous production, giving an example of an affordable standardized sequencing project. Despite the large amount of useful data produced by the NGS approach, our study showed that sequence confirmation and data validation should not be neglected. For the heterologous secretory production of single isoenzymes in *P. pastoris*, the codon usages of the coding sequences were optimized for efficient translation, and fragments corresponding to the predicted mature isoenzymes were produced synthetically. If the signal peptide prediction with SignalP led to two alternative signal peptide junctions, the mature peptides corresponding to the longer signal peptide variants were ordered as synthetic fragments, and the the signal peptide variants with shorter mature peptide were successfully amplified via PCR. Correct cloning of all genes into *P. pastoris* expression vectors was verified by Sanger sequencing. Since the used *P. pastoris* expression vector already contains the signal sequence of the *S. cerevisiae* mating factor α, the isoenzymes were produced without the predicted natural signal sequences. From the 26 isoenzymes produced, 22 showed peroxidase activity with at least one of the substrates used. All activities were measured with four different assays with a range of assay pH optima from 4.5 to seven. Interestingly, for various assays (Table 4) different isoenzymes showed the highest activities thus suggesting variable substrate specificities or pH optima. This observation further emphasizes the importance of the availability of a large group of individually produced pure isoenzymes to be able to comprehensively respond to the need of variable performance parameters including substrate specificity, activity, stability and operating pH optimum.

Interestingly, an isotig corresponding to isoenzyme C1A reported to be the most abundant isoenzyme in *A. rusticana* (1) and thus expected to be found in the transcriptome, was absent. Only two raw reads covering a minor part of the coding sequence could be detected. This might either suggest over-normalization of the cDNA library decreasing the total counts of the putatively most abundant transcripts to almost zero (43), or happen due to naturally occurring genetic variation with phenotypic correlation to adaptations to natural environments ranging from pathogens, light conditions or abiotic stress to a variety of other environmental perturbations (8, 36, 37). Although mRNA originating from all available plant parts was used to reach genes activated at diverse stages of *A. rusticana* growth, some developmental stages were not present and the absence of certain isoenzymes due to missing tissues cannot be ruled out. This finding might illustrate the high variance of HRP expression in *A. rusticana* plants and consequently the variance in the commercial HRP preparations, thus underlining the clear need for a reliable heterologous expression system that enables a consistent isoenzyme quality.

Peroxidase isoenzymes have been suggested to have multiple roles in the plant and thus also be variedly expressed depending on both biotic and abiotic factors (8). To roughly estimate the expression levels of the newly discovered peroxidase genes, their GC contents and codon usage biases were defined. Genes that are highly expressed have been suggested to possess a higher GC content and a more biased codon usage than genes with low expression levels (46). A majority of both eukaryotic and prokaryotic species with large population sizes have been reported to have non-random codon usage mainly due to Darwinian selection between synonyms (47–49). Highly expressed genes have been reported to use a restricted set of codons to ensure optimal translational efficiency (50, 51). In addition to gene expression levels, GC content has also been connected to gene regulation (52–55) and correlated with genomic features including methylation pattern (56), short intron length (57) and gene density (58) thus suggesting possible functional relevance. The codon usages and GC contents calculated using the verified coding sequences of the isoenzymes are described in Table 1. As expected, large variation between isoenzymes exist. These findings could suggest a spatial and temporal distribution of the isoenzymes in cellular processes (59).

Phylogenetic relationships of the HRP isoenzymes are shown in Figure 2. Most of the new isoenzymes discovered in the transcriptome seem to share higher evolutionary distance to the previously known HRPs. BLASTX analysis to the peroxidases of *A. thaliana* (Supplementary Table 2) revealed that the *A. rusticana* peroxidases share 81% to 95% sequence similarity to the most similar isoenzyme of *A. thaliana*. Evolutionary distance does not necessarily correlate with altered substrate specificity, specific activity or optimal reaction conditions, but the discovery of new evolutionary branches with higher structural diversity does offer optimal conditions for the generation of an enzyme assortment with diverse properties for a wide variety of biomedical and industrial applications.

A combination of cDNA sequencing and gDNA verification in this study also provided valuable information of the intron-exon boundaries of the HRP genes. The number of exons in the isoenzymes was noted to vary from one to four, corresponding to zero to three introns. A large majority (76% of the peroxidase loci) of the isoenzymes were found to have four exons and three introns. Intron numbers have been reported to be highly conserved, but total intron length (total sum of the sizes of all introns within a gene) rather correlated to the GC-content of the gene (60). Thus, intron number could be informative in terms of evolutionary origin and distance of the enzyme. In this study, intron numbers were found to correlate with the phylogenetic relationships of the amino acid sequences. Contigs with an unusual number of introns (none or two, Table 1) were situated in close proximity to each other furthest away from the previously known isoenzymes and clustered together when comparing the codon usage frequencies. With the information obtained from the reads, no alternative splicing could be shown.

The well-characterized isoenzyme HRP C1A has been reported to have a signal peptide consisting of 30 amino acids, and a carboxy-terminal extension suggested to target the protein to the vacuoles (61). Also other known isoenzymes of the group C (C1B, C1C, C2, and C3) have been reported to have signal peptides varying in length from 9 amino acids (C1C) to 29 amino acids (C3). By observing the alignment of all previously known and newly discovered isoenzymes (Supplementary Figure 1), existence of signal sequences also in other previously known and most of the new isoenzymes seemed very probable. According to the signal sequence prediction (SignalP) performed, all isoenzymes seem to have a signal sequence varying in length from 18 to 31 amino acids (Table 1, Supplementary Figure 1). Isoenzyme C1C, previously reported to have a signal sequence of nine amino acids, was confirmed to have - better corresponding to the sequences of the very closely related isoenzymes C1A and C1B - a signal sequence of 29 amino acids. In the case of unclear signal sequence prediction with more than one option for the length of the signal peptide, both forms were taken into consideration when planning the constructs for enzyme production in *P. pastoris*.

In summary, to facilitate the possibilities for heterologous expression and isoenzyme characterization, we have elucidated the nucleotide sequences of 28 horseradish peroxidase isoenzymes by using the data obtained from *A. rusticana* 454 transcriptome sequence analysis with manual verification of PCR amplified genomic DNA. Although studies including transcriptome analysis of non-model species have become increasingly popular since the emergence of the NGS technologies, methods for the utilization of the 454 technology for the purpose of isoenzyme discovery in non-model plant species have not been established. In this project, transcriptome sequencing reads are further processed with alternative assemblies and manual sequence verification to determine the nucleotide sequences of all HRP isoenzymes. This study does not only contribute a set of transcripts, which can be used for marker development and genomic studies to understand agriculturally important traits in *A. rusticana,* but also provides valuable information of the peroxidase gene structure. Twenty-two of the verified isoenzymes have been produced in an active form in *P. pastoris* utilizing a new *P. pastoris* expression platform (Näätsaari *et al*., submitted), validating the success of the approach and providing first insights into the versatility of this large group of isoenzymes discovered.

SUPPLEMENTARY INFORMATION

Supplementary data are available at NAR online: Supplementary Tables 1-2 and Supplementary Figure 1.

FUNDING

This work was supported by the Austrian Science Fund ‘Fonds zur Förderung der wissenschaftlichen Forschung‘ [W901 DK Molecular Enzymology] and in part by the Austrian Ministry of Science and Research GEN-AU project BIN [FFG Grant 820962]. Funding for open access charge: Fonds zur Förderung der wissenschaftlichen Forschung.

ACKNOWLEDGEMENTS

We would like to thank Christian Schmid and Christoph Fischer for technical assistance.

REFERENCES

1. Veitch,N.C. (2004) Horseradish peroxidase: a modern view of a classic enzyme. *Phytochemistry*, **65**, 249-259.

2. Greco,O., Rossiter,S., Kanthou,C., Folkes,L.K., Wardman,P., Tozer,G.M. and Dachs,G.U. (2001) Horseradish Peroxidase-mediated Gene Therapy : Choice of Prodrugs in Oxic and Anoxic Tumor Conditions. *Mol. Cancer Ther.*, **1**, 151-160.

3. Morawski,B., Quan,S. and Arnold,F.H. (2001) Functional expression and stabilization of horseradish peroxidase by directed evolution in *Saccharomyces cerevisiae. Biotechnol. Bioeng.*, **76**, 99-107.

4. Wagner,M. and Nicell,J.A. (2002) Detoxification of phenolic solutions with horseradish peroxidase and hydrogen peroxide. *Water Res.*, **36**, 4041-52.

5. Azevedo,A.M., Martins,V.C., Prazeres,D.M.F., Vojinović,V., Cabral,J.M.S. and Fonseca,L.P. (2003) Horseradish peroxidase: a valuable tool in biotechnology. *Biotechnol. Annu. Rev.*, **9**, 199-247.

6. van de Velde,F., van Rantwijk,F. and Sheldon, R.A. (2001) Improving the catalytic performance of peroxidases in organic synthesis. *Trends Biotechnol.*, **19**, 73-80.

7. Hoyle,M.C. (1977) High resolution of peroxidase-indoleacetic acid oxidase isoenzymes from horseradish by isoelectric focusing. *Plant Physiol.*, **60**, 787-93.

8. Passardi,F., Cosio,C., Penel,C. and Dunand,C. (2005) Peroxidases have more functions than a Swiss army knife. *Plant Cell Rep.*, **24**, 255-265.

9. Fujiyama,K., Takemura,H., Shibayama,S., Kobayashi,K., Choi,J.K., Shinmyo,A., Takano,M., Yamada,Y. and Okada,H. (1988) Structure of the horseradish peroxidase isozyme C genes. *Eur. J. Biochem.*, **173**, 681-7.

10. Fujiyama,K., Takemura,H., Shinmyo,A., Okada,H. and Takano,M. (1990) Genomic DNA structure of two new horseradish-peroxidase-encoding genes. *Gene*, **89**, 163-169.

11. Bartonek-Roxå,E., Eriksson,H. and Mattiasson,B. (1991) The cDNA sequence of a neutral horseradish peroxidase. *Biochim. Biophys. Acta*, **1088**, 245-50.

12. Nielsen,K.L., Indiani,C., Henriksen,A., Feis,A., Becucci,M., Gajhede,M., Smulevich,G. and Welinder,K.G. (2001) Differential Activity and Structure of Highly Similar Peroxidases. Spectroscopic , Crystallographic , and Enzymatic Analyses of Lignifying Arabidopsis thaliana Peroxidase A2 and Horseradish Peroxidase A2. *Biochemistry*, **40**, 11013-11021.

13. Morita Y., Mikami B., Yamashita H., Lee J.Y., Aibara S., Sato M., Katsube Y.,T.N. (1991) Primary and crystal structures of horseradish peroxidase isozyme E5. In Lobarzewski, J., Greppin, H., Penel, C. and Gaspar,T. (ed), *BIOCHEMICAL, MOLECULAR, AND PHYSIOLOGICAL ASPECTS OF PLANT PEROXIDASES*. University of M. Curie-Sklodowska and University of Geneva, Lublin and Geneva (1991), pp. 81-88.

14. Shannon,L.M., Kay,E. and Lew,J.Y. (1966) from Horseradish Roots. *J. Biol. Chem.*, **241**, 2166-2172.

15. Aibara,S., Yamashita,H., Mori,E., Kato,M. and Morita,Y. (1982) Isolation and characterization of five neutral isoenzymes of horseradish peroxidase. *J Biochem*, **92**, 531-539.

16. Aibara,S., Kobayashi,T. and Morita,Y. (1981) Isolation and properties of basic isoenzymes of horseradish peroxidase. *J Biochem*, **90**, 489-496.

17. Weber,A.P.M., Weber,K.L., Carr,K., Wilkerson,C. and Ohlrogge,J.B. (2007) Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiol.*, **144**, 32-42.

18. Parchman,T.L., Geist,K.S., Grahnen,J.A., Benkman,C.W. and Buerkle,C.A. (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC genomics*, **11**, 180, 10.1186/1471-2164-11-180.

19. Sun,C., Li,Y., Wu,Q., Luo,H., Sun,Y., Song,J., Lui,E.M.K. and Chen,S. (2010) *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. *BMC genomics*, **11**, 262.

20. Schmid,J., Müller-Hagen,D., Bekel,T., Funk,L., Stahl,U., Sieber,V. and Meyer,V. (2010) Transcriptome sequencing and comparative transcriptome analysis of the scleroglucan producer *Sclerotium rolfsii. BMC genomics*, **11**, 329.

21. Wong,M.M.L., Cannon,C.H. and Wickneswari,R. (2011) Identification of lignin genes and regulatory sequences involved in secondary cell wall formation in *Acacia auriculiformis* and *Acacia mangium* via *de novo* transcriptome sequencing. *BMC genomics*, **12**, 342, 10.1186/1471-2164-12-342.

22. Wang,T.-Y., Chen,H.-L., Lu,M.-Y.J., Chen,Y.-C., Sung,H.-M., Mao,C.-T., Cho,H.-Y., Ke,H.-M., Hwa,T.-Y., Ruan,S.-K., *et al*. (2011) Functional characterization of cellulases identified from the cow rumen fungus *Neocallimastix patriciarum* W5 by transcriptomic and secretomic analyses. *Biotechnol. Biofuels*, **4**, 24.

23. Marchler-Bauer,A., Lu,S., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R., *et al*. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.*, **39**, D225-9.

24. Shyamala,V. and Ferro-Luzzi Ames,G. (1989) Genome walking by single-specific-primer. *Gene*, **84**, 1-8.

25. Puigbò,P., Bravo,I.G. and Garcia-Vallve,S. (2008) CAIcal: a combined set of tools to assess codon usage adaptation. *Biol. Direct*, **3**, 38.

26. Tamura,K., Peterson,D., Peterson,N., Stecher,G., Nei,M. and Kumar,S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, **28**, 2731-9.

27. Higgins,D.G., Thompson,J.D. and Gibson,T.J. (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.*, **266**, 383-402.

28. Cheng,J., Saigo,H. and Baldi,P. (2006) Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. *Proteins*, **62**, 617-29.

29. Lippi,M., Passerini,A., Punta,M., Rost,B. and Frasconi,P. (2008) MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. *Bioinformatics*, **24**, 2094-5.

30. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783-95.

31. Abad,S., Kitz,K., Hörmann,A., Schreiner,U., Hartner,F.S. and Glieder,A. (2010) Real-time PCR-based determination of gene copy numbers in *Pichia pastoris. Biotechnol. J.*, **5**, 413-20.

32. Lin-Cereghino,J., Wong,W.W., Xiong,S., Giang,W., Luong,L.T., Vu,J., Johnson,S.D. and Lin-Cereghino,G.P. (2005) Condensed protocol for competent cell preparation and transformation of the methylotrophic yeast *Pichia pastoris*. *Biotechniques*, **38**, 44-48.

33. Weis,R., Luiten,R., Skranc,W., Schwab,H., Wubbolts,M. and Glieder,A. (2004) Reliable high-throughput screening with *Pichia pastoris* by limiting yeast cell death phenomena. *FEMS Yeast Res.*, **5**, 179-89.

34. Ryan,B.J. and O'Fágáin,C. (2007) Arginine-to-lysine substitutions influence recombinant horseradish peroxidase stability and immobilisation effectiveness. *BMC Biotechnol.*, **7**, 86.

35. Morawski,B., Lin,Z., Cirino,P., Joo,H., Bandara,G. and Arnold,F.H. (2000) Functional expression of horseradish peroxidase in *Saccharomyces cerevisiae* and *Pichia pastoris*. *Protein Eng.*, **13**, 377-84.

36. Garg,R., Patel,R.K., Tyagi,A.K. and Jain,M. (2011) *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res.*, **18**, 53-63.

37. Toth,A.L., Varala,K., Newman,T.C., Miguez,F.E., Hutchison,S.K., Willoughby,D.A., Simons,J.F., Egholm,M., Hunt,J.H., Hudson,M.E., *et al.* (2007) Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science*, **318**, 441-4.

38. Vera,J.C., Wheat,C.W., Fescemyer,H.W., Frilander,M.J., Crawford,D.L., Hanski,I. and Marden,J.H. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol.*, **17**, 1636-47.

39. Novaes,E., Drost,D.R., Farmerie,W.G., Pappas,G.J., Grattapaglia,D., Sederoff,R.R. and Kirst,M. (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC genomics*, **9**, 312.

40. Pop,M. and Salzberg,S.L. (2008) Bioinformatics challenges of new sequencing technology. *Trends Genet.*, **24**, 142-149.

41. Natarajan,P. and Parani,M. (2011) *De novo* assembly and transcriptome analysis of five major tissues of *Jatropha curcas L.* using GS FLX titanium platform of 454 pyrosequencing. *BMC genomics*, **12**, 191.

42. Huse,S.M., Huber,J.A., Morrison,H.G., Sogin,M.L. and Welch,D.M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.*, **8**, R143.

43. Hale,M.C., McCormick,C.R., Jackson,J.R. and Dewoody,J.A. (2009) Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. *BMC genomics*, **10**, 203.

44. Delker,C., Pöschl,Y., Raschke,A., Ullrich,K., Ettingshausen,S., Hauptmann,V., Grosse,I. and Quint,M. (2010) Natural variation of transcriptional auxin response networks in *Arabidopsis thaliana*. *The Plant cell*, **22**, 2184-200.

45. Alonso-Blanco,C., Aarts,M.G.M., Bentsink,L., Keurentjes,J.J.B., Reymond,M., Vreugdenhil,D. and Koornneef,M. (2009) What has natural variation taught us about plant development, physiology, and adaptation? *The Plant cell*, **21**, 1877-96.

46. Wright,F. (1990) The “effective number of codons” used in a gene. *Gene*, **87**, 23-29.

47. Gouy,M. and Gautier,C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*, **10**, 7055-7074.

48. Ikemura,T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13-34.

49. Sharp,P.M. and Matassi,G. (1994) Codon usage and genome evolution. *Curr. Opin. Genet. Dev.*, **4**, 851-860.

50. Bulmer,M. (1991) The Selection-Mutation-Drift Theory of Synonymous Codon Usage. *Genetics*, **129**, 897-907.

51. Li,W.H. (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.*, **24**, 337-345.

52. Carels,N. and Bernardi,G. (2000) The compositional organization and the expression of the *Arabidopsis* genome. *FEBS Lett.*, **472**, 302-6.

53. Carels,N. and Bernardi,G. (2000) Two classes of genes in plants. *Genetics*, **154**, 1819-25.

54. Vinogradov,A.E. (2003) DNA helix: the importance of being GC-rich. *Nucleic Acids Res.*, **31**, 1838-1844.

55. Zhang,L., Kasif,S., Cantor,C.R. and Broude,N.E. (2004) GC/AT-content spikes as genomic punctuation marks. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 16855-60.

56. Jabbari,K. and Bernardi,G. (1998) CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. *Gene*, **224**, 123-127.

57. Galtier,N., Piganeau,G., D.,M. and Duret,L. (2001) GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics*, **159**, 907-911.

58. Mouchiroud,D., D'Onofrio,G., Aïssani,B., Macaya,G., Gautier,C. and Bernardi,G. (1991) The distribution of genes in the human genome. *Gene*, **100**, 181-7.

59. Cosio,C. and Dunand,C. (2009) Specific functions of individual class III peroxidase genes. *J. Exp. Bot.*, **60**, 391-408.

60. Rayko,E., Jabbari,K. and Bernardi,G. (2006) The evolution of introns in human duplicated genes. *Gene*, **365**, 41-7, 10.1016/j.gene.2005.09.038.

61. Welinder,K.G. (1976) Covalent structure of the glycoprotein horseradish peroxidase (EC 1.11.1.7). *FEBS Lett.*, **72**, 19-23.

FIGURE LEGENDS

Figure 1. Overview of the sequencing and assembly of *the A. rusticana* transcriptome. (A) Size distribution of the quality-filtered reads. Total number of reads: 556269, average/median length 342.9/379.0 (B) Length distribution of the 18511 contigs. Average/median length of the contigs: 717.7/667.0 (C) Coverage distribution of the 18511 contigs. Average/median coverage of the contigs: 11.4/7.8 (D) Length distribution of the14871 isotigs. Average/median length of the isotigs: 1133.3/1113.0

Figure 2. An evolutionary distance tree (phylogram) of all isoenzymes known and discovered during this study.

Figure 3. A heat map of of the changes in the relative synonymous codon usage (ΔRSCU). Each column represents one codon indicated along the bottom, each row one isoenzyme marked to the right side of the row. Isoenzymes are clustered by their codon usage similarity. Green cells correspond to underrepresented codons, red cells to overrepresented codons. Missing codons are marked with a grey cell.

Figure 4. GC content of the *A. rusticana* transcripts.

(Figure 5). SDS-PAGE analysis of the horseradish peroxidase isoenzymes in *Pichia pastoris*.

TABLES

Table 1: Summary of the horseradish peroxidase isoenzymes and associated data produced during this study. The nucleotide sequences of 28 isoenzymes were submitted to EMBL. “*” indicates previously unknown isoenzymes. “CAI” = codon adaptation index. If signal sequence predictions gave more than one alternative result, both signal sequence lengths are shown, separated by “/”. Disulfide bridges were predicted using both alternatives of the mature protein (EDBCP = Ensemble-based Disulfide Bonding Connectivity Pattern prediction server)

Contig number	Name (* previously unknown)	Length nt	GC content %	Accession #	CAI calculated	Intron #	Signal sequence length	Disulfide bridges (EDBCP prediction)	pI mature protein
-	C1A	1062	43.69		0.812	3	30	[11-91] [44-49] [97-301] [177-209]	5.59
15901	C1B	1056	43.94		0.808	3	28	[11-91] [44-49] [97-301] [177-209]	5.84
25148	C1C	1059	45.14		0.809	3	29	[11-91] [44-49] [97-301] [177-209]	6.49
25148_2	C1D*	1059	45.04		0.810	3	29	[11-91] [44-49] [97-301] [177-209]	7.04
04627	C2	1044	42.91		0.800	3	24	[11-91] [44-49] [97-301] [177-209]	8.56
-	C3	1050	46.76		0.781	3	29	[11-91] [44-49] [97-300] [177-209]	7.71
Manual assembly	A2A*	1011	46.79		0.761	3	31	[11-91] [44-49] [97-299] [176-208]	4.93
Manual assembly	A2B*	1011	46.69		0.761	3	31	[11-91] [44-49] [97-299] [176-208]	4.93
04382	E5	1044	46.07		0.771	3	27	[11-91] [44-49] [97-300] [177-209]	8.84
01805	01805*	1065	44.41		0.797	3	31	[11-91] [44-49] [97-301] [177-209]	5.97
22684	22684.1*	1050	46.76		0.770	3	29	[11-91] [44-49] [97-300] [177-209]	6.98
22684_2	22684.2*	1050	46.67		0.772	3	29	[11-91] [44-49] [97-300] [177-209]	6.37
01350	01350*	975	50.97		0.707	3	28	[11-91] [44-49] [97-292] [176-201]	8.67
02021	02021*	996	46.08		0.788	3	29	[11-89] [44-49] [95-297]	9.46
23190	23190.1*	1080	49.26		0.724	2	31/42	[11-92] [44-49] [98-293] [178-205] or [22-103] [55-60] [109-304] [189-216]	8.40/6.58
23190_2	23190.2*	1080	49.17		0.722	2	31/42	[11-92] [44-49] [98-293] [178-205] or [22-103] [55-60] [109-304] [189-216]	8.60/7.09
04663	04663*	1077	47.82		0.748	3	31	[11-91] [44-49] [97-299] [176-208]	4.48
06351	06351*	945	43.39		0.786	3	18	[17-96] [50-55] [102-292] [180-206]	6.37
03523	03523*	960	44.58		0.761	0	22	[11-92] [44-49] [98-293] [177-203]	8.99
05508	05508.1*	966	49.28		0.735	2	24/30	[11-87] [44-49] [93-287] [171-198] or [17-93] [50-55] [99-293] [177-204]	8.49/8.47
05508_2	05508.2*	966	49.38		0.735	2	24/30	[11-87] [44-49] [93-287] [171-198] or [17-93] [50-55] [99-293] [177-204]	8.49/8.47

22489_1	22489.1*	978	51.02		0.726	2	23/34	[22-98] [55-60] [104-298] [182-209] or [11-87] [44-49] [93-287] [171-198]	8.93/8.51
22489_2	22489.2*	978	50.82		0.727	2	23/34	[22-98] [55-60] [104-298] [182-209] or [11-87] [44-49] [93-287] [171-198]	8.93/8.51
06117	06117*	1008	47.02		0.802	3	22/32	[11-91] [44-49] [97-298] [176-208] or [21-101] [54-59] [107-308] [186-218]	5.52/6.16
17517_1	17517.1*	972	48.46		0.737	2	23/24	[12-88] [45-50] [94-296] [171-203] or [11-87] [44-49] [93-295] [170-202]	9.49/9.39
17517_2	17517.2*	972	48.56		0.739	2	23/24	[12-88] [45-50] [94-296] [171-203] or [11-87] [44-49] [93-295] [170-202]	9.52/9.41
08562_1	08562.1*	996	47.09		0.779	3	22/28	[11-91] [44-49] [97-298] [176-208] or [17-97] [50-55] [103-304] [182-214]	9.01/9.03
08562_4	08562.2*	996	47.79		0.788	3	22/28	[11-91] [44-49] [97-298] [176-208] or [17-97] [50-55] [103-304] [182-214]	9.00/9.02

Table 2: Summary of the transcriptome sequencing and enzyme discovery processes

	Number	%
Transcriptome sequencing and assembly		
Total number of reads	592507	-
Clipped reads	556269	-
Reads after Newbler quality control	543439	100
Reads aligned	490285	90.2
Reads assembled	433179	88.4
Reads partially assembled	57064	10.5
Singletons	35950	7.3
Contigs	18511	-
Isogroups	10619	-
Isotigs	14871	-
Average isotig size	1133	-
Largest isotig size	3659	-
Average contig coverage	11.4	-
HRP isoenzyme discovery and verification		
# of isotigs with a secretory peroxidase domain	18	
# of full length peroxidase genes	18	
# of peroxidase genes after manual revision	20	
Total # of isoenzymes (including allelic variants)	28	100
Successfully verified from gDNA	26	89.7
Enzyme production in *Pichia pastoris*		
Synthetic genes for production	26	89.7
Successful production of an active isoenzyme	22	75.9

Table 3: Comparison of the HRP isoenzyme sequences between GenBank, UniProt, transcriptome and verified genome sequences. All nucleotide (nt) and amino acid (aa) positions were calculated from the start ATG. Variations between the nt positions of the transcriptome sequence compared to other sequence sources are either due to deletions or intronic sequences in the other sources. "-" indicates that no sequence information is available from the respective source. "ss" indicates a putative signal sequence. Deletions or missing sequences are marked with "*". "N/NX" indicates a variation at position X. "None" indicates that no polymorphisms or differences between transcriptome and genome sequence were found. The isoenzymes C1A and C3 were detected in the transcriptome raw reads with only partial/low coverage (0-2x), thus no consensus sequence was formed.

HRP	Sanger sequence		transcriptome sequence		GenBank sequence		UniProt sequence
	nt	aa (exon) /intron	nt	aa	nt	aa (exon) /intron	aa
C1A	TA 109-110	Y37	-	-	AT 109-110	I37	Y37
	C1159	intron	-	intron	G1159	intron	-
C1B	T/C253	intron	-	intron	T253	intron	-
	T/C859	intron	-	intron	C859	intron	-
C1C	ss nt1-60	ss aa1-20	ss nt1-60	ss aa1-20	*	*	*
	C178	R60	C178	R60	A118	S40	S40
	A/T1335	intron	-	intron	-	-	-
	A/G1888	T/A165	A/G493	T/A165	G433	A145	A145
	C1889	A165	C/T1889	A/V165			
	C/G1921	Q/E176	C/G526	Q/E176	G466	E156	E156
C2	CT 1250-1251	intron	-	intron	*	intron	-
	A1334	intron	-	intron	*	intron	-
C3	-	-	-	-	-	-	-
A2	ss nt1-93	ss aa1-31	ss 1-93	ss aa1-31	-	-	*
	AAT 231-234	N78	AAT 231-234	N47	-	-	D47
	GGA 996-998	G220	GGA 661-663	G220	-	-	N189
	AAT 999-1001	N221	AAT 664-666	N221	-	-	G190
	ACG 1185-1187	T284	ACG 850-852	T284	-	-	L253
	G/A1203	A/T290	G868	A290	-	-	A259
	AAT 1335-1337	N334	AAT 999-1002	N334	-	-	D303
E5	ss nt1-81	ss aa1-27	ss nt1-81	ss aa1-27	-	-	*
	T419	L82	C/T246	L82	-	-	L55
	C422	D83	T/C249	D83	-	-	D56
	C545	C124	T/C372	C124	-	-	C97
01805	none	none	none	none	-	-	-
22684	G1611	R337	A1010	K337	-	-	-
	TGA 1627-1629	D343	CGG 1026-1028	G343	-	-	-
01350	none	none	none	none	-	-	-

02021	none	none	none	none	-	-	-
03523	none	none	none	none	-	-	-
06117	T30	V10	C/T30	V10	-	-	-
	C1088	I269	T807	I269	-	-	-
17517	T190	Y64	C190	H64	-	-	-
	C1157	G282	T846	G282	-	-	-
	A1232	K307	G921	K307	-	-	-
08562.1	none	none	none	none	-	-	-
08562.4	none	none	none	none	-	-	-
23190	T1345	S109	G1345	S109	-	-	-
	C1423	G135	T1423	G135	-	-	-
	T1842	S222	T/C1842	S/P222			
	C1850	T224	A/C1850	T224	-	-	-
	A2221	E348	T/A2221	V/E348	-	-	-
04663	none	none	none	none	-	-	-
06351	none	none	none	none	-	-	-
05508	G/A346	A/T116	G/A346	A/T116	-	-	-
22489	-	-	G/A597	T199	-	-	-
	.	.	G/T715	A/S239	-	-	-

Table 4: Summary of isoenzyme expression and characterization. Isoenzymes showing obvious peroxidase activity with the assay used are marked with "+". Isoenzymes showing very low but detectable peroxidase activity with the assay used are marked with "(+)". Isoenzymes with no activity detected during an observation period of 2h are marked with "-". Allelic variants not produced heterologously are marked with n.d. (no data available). Isoenzymes discovered during this study are mareked with "*".

Contig number	Name (* previously unknown)	Activity/ABTS	Activity/TMB	Activity/Guaiacol	Activity/Pyrogallol
-	C1A	+	+	+	+
15901	C1B	-	+	-	-
25148	C1C	+	+	(+)	+
25148_2	C1D*	+	+	(+)	+
04627	C2	+	+	+	+
-	C3	+	+	+	+
Manual assembly	A2A*	+	+	+	+
Manual assembly	A2B*	+	+	(+)	(+)
04382 (Contig split manually)	E5	+	+	+	+
01805	01805* (B1?)	+	+	-	-
22684	22684.1* (B2A?)	+	+	-	-
22684_2	22684.2* (B2B?)	+	+	-	-
01350 (Contig split manually)	01350*	+	+	-	-
02021	02021*	-	-	-	-
23190	23190.1*	-	-	-	-
23190_2	23190.2*	n.d.	n.d.	n.d.	n.d.
04663	04663.1*	+	(+)	-	-
04663_2	04663.2*	n.d.	n.d.	n.d.	n.d.
06351	06351*	+	+	+	+
03523	03523*	-	-	-	-
05508	05508.1*	+	+	+	+
05508_2	05508.2*	+	+	n.d.	n.d.
22489_1	22489.1*	+	+	(+)	+
22489_2	22489.2*	+	+	(+)	+
06117	06117*	-	-	-	-
17517_1	17517.1*	+	-	-	-
17517_2	17517.2*	+	+	+	+
08562_1	08562.1*	+	+	-	-
08562_4 (Contig split manually)	08562.2*	+	+	-	+

FIGURES

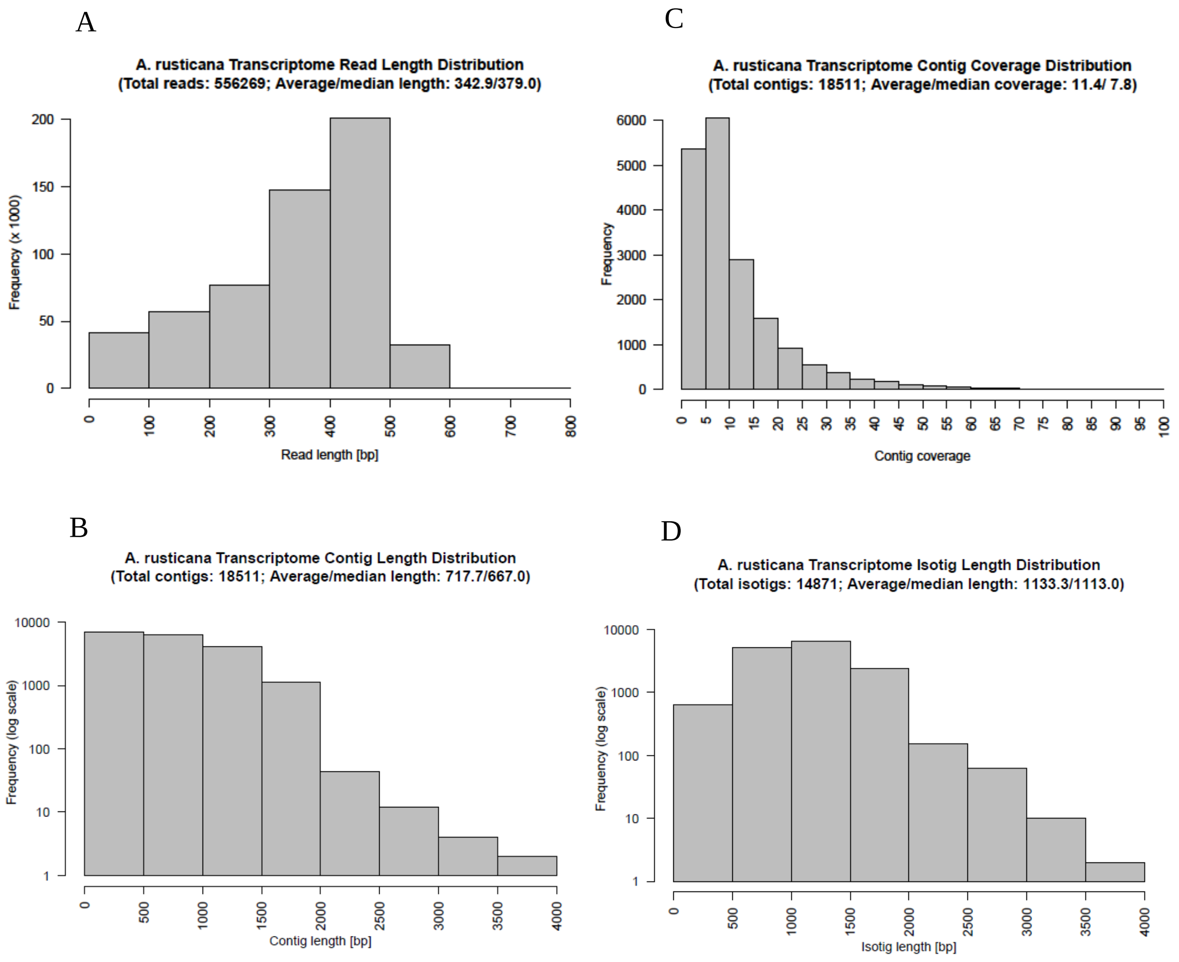


Figure 1: Overview of the sequencing and assembly of the *A. rusticana* transcriptome. (A) Size distribution of the quality-filtered reads. Total number of reads: 556269, average/median length 342.9/379.0 (B) Length distribution of the 18511 contigs. Average/median length of the contigs: 717.7/667.0 (C) Coverage distribution of the 18511 contigs. Average/median coverage of the contigs: 11.4/7.8 (D) Length distribution of the14871 isotigs. Average/median length of the isotigs: 1133.3/1113.0

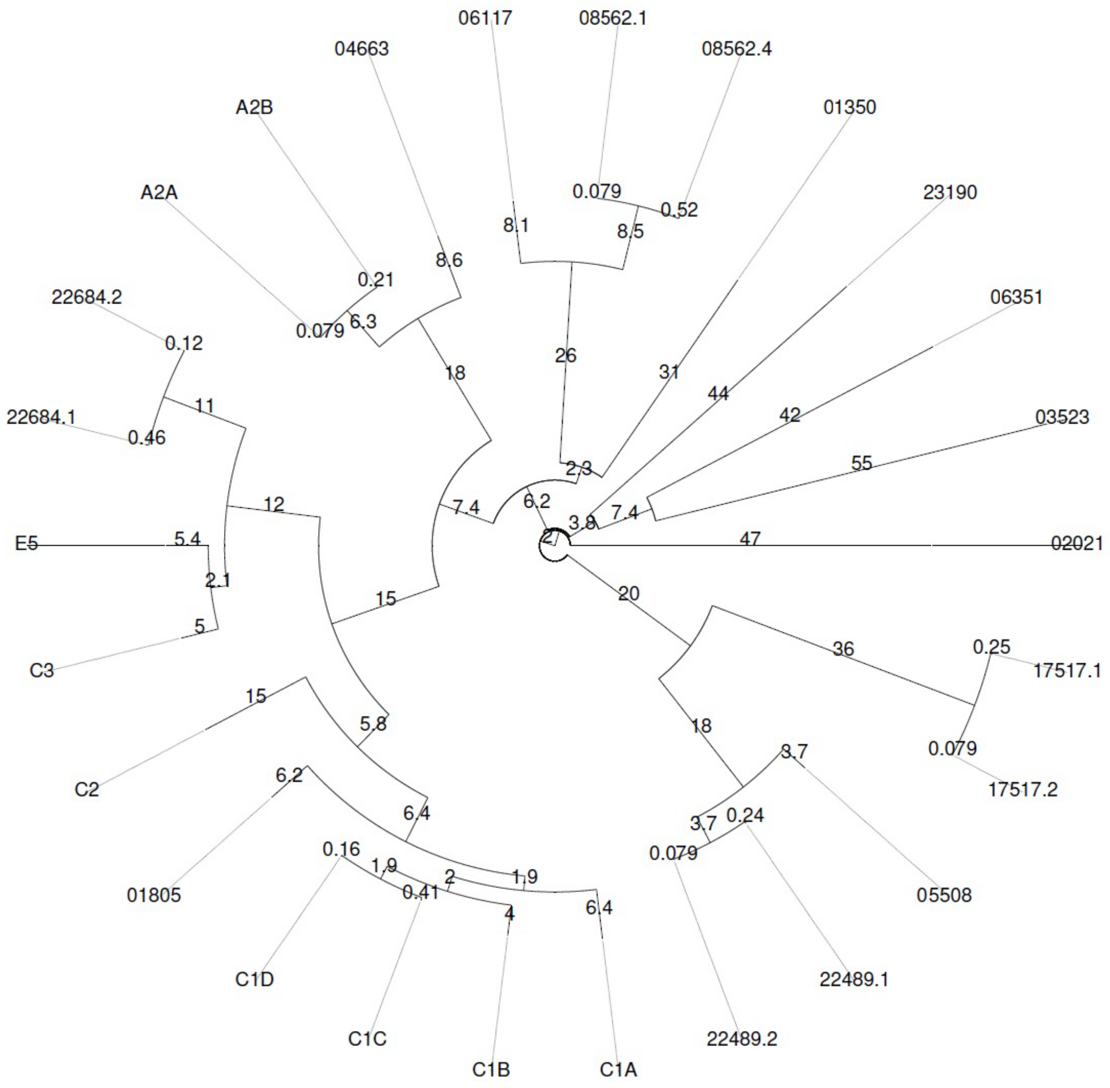


Figure 2: Phylogram of all isoenzymes known and discovered during this study.

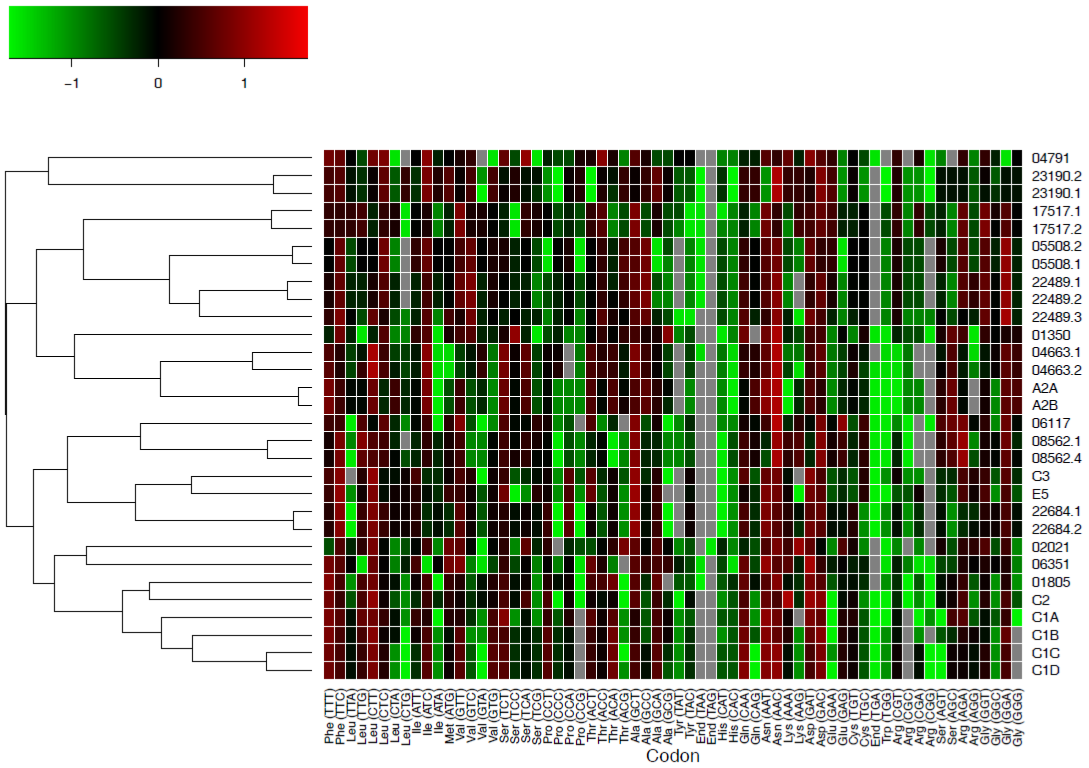


Figure 3: A heat map of of the changes in the relative synonymous codon usage (ΔRSCU) of all the HRP isoenzymes verified in this study. Each column represents one codon indicated along the bottom, each row one isoenzyme marked to the right side of the row. Isoenzymes are clustered by their codon usage similarity. Green cells correspond to underrepresented codons, red cells to overrepresented codons. Missing codons are marked with a grey cell.

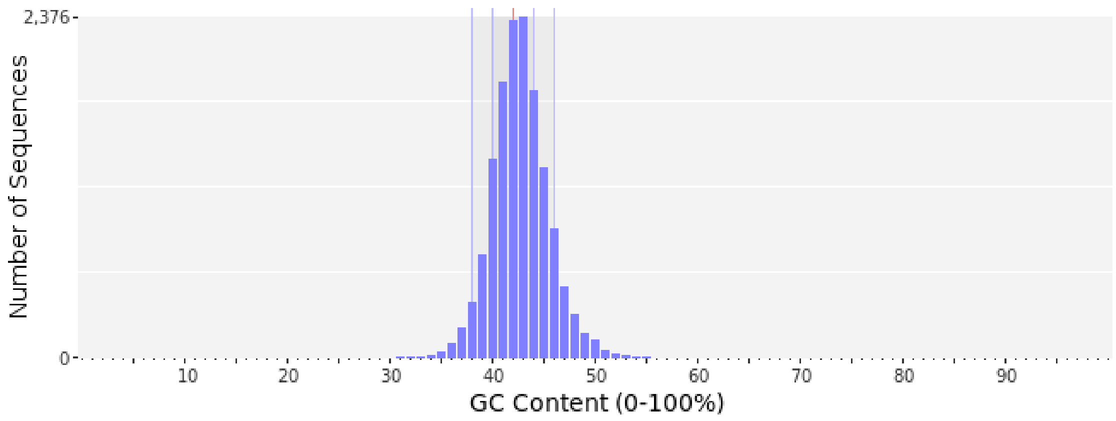


Figure 4: GC content of the *A. rusticana* transcripts varies from 28% (min) to 62% (max) with a range of 35%. The average GC content of all transcripts is 42.72%. Mode (x-axis value) 43, mode value (y-axis value) 2376.

SUPPLEMENTARY INFORMATION

Supplementary Table 1. Primers used in this study.

Number	Name	Sequence
P10-729	01350fw1	GGATGCGATGGTTCGATTTTAC
P10-727	01350fw2	GCGACACCGACAAAACAATTG
P10-728	01350rv1	GAGCGGAATTGCGGTTTG
P10-726	01350rv2	TGCGCGCCTGAAAAAAATGAAC
P10-725	06117fw1	GAGATTTAACAGCCAAGGTCTC
P10-724	06117rv1	CTCTTGAGAACTTAAGGAAGCG
P10-723	08562.1.4fw1	GAGCTGGGTGGTTTCATTAG
P10-722	08562.1.4rv1	GCTTTAGCTACGACTGATCTC
P10-721	08562.1.4rv2	CAAAAACCACACCGAGTTAGC
P10-720	17517fw1	CTAGTTACGTTTTTAGTATTGGTCG
P10-719	17517rv1	GACCAAATATGTAACTATTTGATTAATATACATTG
P10-573	A2_Nterm_Strepfw1	GTCCCACCCACAGTTCGAGAAGTCTGCACAATTGAATGCAACTTTCTATTCCG
P10-574	A2_Nterm_Strepfw2	ATATCTCGAGAAGAGAGAGGCCGAAGCTTGGTCCCACCCACAGTTCGAGAAG
P10-374	A2ex1rv	CGAAAAAGATCAAGAAGCCTTTATG
P10-381	A2ex4fw	CTTTGCTCAGTCCATGATCAAC
P10-382	A2ex4rv	GCCATTATGTGTTTTTTGAAAAACAAGAG
P10-375	A2in1ex2fw	CACGACTGCTTTGTTAATGTATAC
P10-376	A2in1ex2rv	GCTCTGGATGCTTCCAC
P10-377	A2in2fwa	CCGTGTCTTTGGTAATTAGTAATTAC
P10-379	A2in2fwb	GTTATATATATATAATTTGTCTTATAAATTATGTTTTAGTAATAATATAG
P10-378	A2in2rva	CTCCTGTCTGAAGATGATCAAAC
P10-380	A2in2rvb	GACTATCTCGTCTTCCTAATAATAC
P10-395	B1ex1fw	CTTAAACCAATAAAAGATAAGTTTCCTCTTAC
P10-396	B1ex1rv	GCGTATCCACAAAATCATCAATTAAAC
P10-397	B1ex2fw	GGTAGTAATTAAAGATTTGATTATGATTTCAAC
P10-398	B1ex2rv	GATAAAATTAAAACTAAAAAAAAGAGATAAATTAATGTTGC
P10-436	B2CtermSpecific1a	GTAAACCTATACAGCAAGAACACG
P10-438	B2CtermSpecific2	CTTCGGTGCATTCGTTGACG
P10-399	B2ex1fw	TAGTCTGTCTTCCTCTTGAAAAAAG
P10-400	B2ex1rv	CATGTTTTTTTCTTTTTCGGATAAAAAGATAAG
P10-403	B2ex2fw	GAGGCTAAAAGCAATTTTTAATTAATAAATACAATC
P10-404	B2ex2rv	ATAATATATTTTGATTAGTTTCACCCGATATAAG
P10-408	B2ex3rv	ATAATACATAGAGTAGGTTATATAGATTGAAAC
P10-411	B2ex4fw	TAAACATATATACAATGTATCTAAACTTTACTTTTTTTTG
P10-412	B2ex4rv	CGTCGTTCTCCATACCCT
P10-401	B2in1fw	CTCGTATTGCAGCTAGTCTC
P10-402	B2in1rv	GCATCTTTCTCGGACTGAAATG
P10-405	B2in2fw	CCTTGAGAAAGCTTGTCCTG
P10-406	B2in2rv	CTCTCCTCCCCAACAAAAC
P10-409	B2in3fw	GCGCCTCAGATCTAGTTG
P10-410	B2in3rv	GTCACAAGTAGACATTGTGCTC
P10-329	C1Aex3in3fw	GAGATAGCTTACAAGCATTTCTG
P10-330	C1Aex3in3rv	CCCTAACAAAAGAAAGAGAGATC
P10-331	C1Aex4fw	CACCATTTGATATAGTTGTATTTAGTGAG

P10-332	C1Aex4rv	GTAGCCACATATGGCGTC
P10-325	C1Ain1fw	GTAAATTACTACTTTTCATATTTCTATTTCGTTAC
P10-325	C1Ain1fw	GTAAATTACTACTTTTCATATTTCTATTTCGTTAC
P10-326	C1Ain1rv	GCGTCACAACCCTTTCAAAAATC
P10-326	C1Ain1rv	GCGTCACAACCCTTTCAAAAATC
P10-327	C1Ain2fw	CGCAGATTTGCTCACCATTG
P10-328	C1Ain2rv	CCGCCTATAATAATAACCATTTTGTTTTG
P10-333	C1Bin1fw	GTTTTATCCTTTAGATATTGATAAATCACCTC
P10-335	C1Bin2fw	GAAATGATGTTGTGTTGTCTAACAATATC
P10-336	C1Bin2rv	GTATGCTCCATTCATTACAAACATTG
P10-337	C1Bin3fw	GACCTCCTGCCTATAATATAATAG
P10-338	C1Bin3rv	GATGCCTTTGCAAAAGTTGGC
P10-339	C1Cex1fw	GCTTCATGCATCTTTTTCCAATG
P10-340	C1Cex1rv	CACAACGCAATAGAAATATGAAAAGTAATATC
P10-343	C1Cex2fw	GTTACCGGTGGTAAAGATTTAGC
P10-344	C1Cex2rv	GAGGAATTGATATATTAAAATTTAAAACAATGATTTTAAC
P10-347	C1Cex3fw	ACAACGATAGCCTAAGTTTGAAAAG
P10-348	C1Cex3rv	ATTAAACTATACCAAATGGTGTTTAGTTTTCT
P10-348	C1Cex3rv	ATTAAACTATACCAAATGGTGTTTAGTTTTCT
P10-341	C1Cin1fw	CATTATCAATGAGTTACGATCGG
P10-342	C1Cin1rv	GCCTTGATTCTGTCAACCACA
P10-345	C1Cin2fw	CATGCCCAAGAACTGTTTCATG
P10-345	C1Cin2fw	CATGCCCAAGAACTGTTTCATG
P10-350	C1Cin3ex4rv	GCAGATCGAAATCCACCAAG
P10-351	C2ex1fw	CAAACCTAACCAAAGAATTTTATCTTAGAG
P10-352	C2ex1rv	CAGGTTTTTTACTAGCGTTAATAAAAAGTC
P10-353	C2in1ex2fwa	CACTAGATTCAATATTTTCACGTATATATTAATTAAG
P10-355	C2in1ex2fwb	CAAAAGAGTTGAAATATACTAAAAATATAATCTTACTAG
P10-357	C2in1ex2fwc	CAAAAGAAAACACCGGAGTTAAATAAAAATAG
P10-354	C2in1ex2rva	ATCAATCATATCCGAAAACGACAAC
P10-356	C2in1ex2rvb	GCGATTGGTCCATTTATAAATACTG
P10-358	C2in1ex2rvc	CCATCCTATCCTGAGCTATC
P10-360	C2in1ex2rvd	TCCTAAATGATGTAGTGTTGTCTAG
P10-361	C2in2fw	CTTTTACAATGCTATGCATCTATACATC
P10-362	C2in2rv	CGCCTATTTTATTTTCATTTTCAAGTAAACAAT
P10-365	C2in3ex4fwb	CAACAAATACTATGTGAATCTCAAAGAG
P10-364	C2in3ex4rva	CTCGGACCAAAGGGATAG
P10-366	C2in3ex4rvb	GTAGTAAAGAAAAAGATAACATTGATCATTTATTATTAG
P10-549	contig01350fw	CAACTCCAACTCCAAGTCTATTC
P10-550	contig01350rv	GATCAAATCTTATTCTCACCGAATC
P10-551	contig02021fw	AAGAGTATTGAGAAACAAGATCGAG
P10-552	contig02021rv	AACAGAAAACATCACTTTCCGAC
P10-545	contig04791fw	TCTCACTTTCTCTCTTCCGC
P10-546	contig04791rv	CATATAGATTTATGTGTTTTAACATTAACCAAAAAC
P10-543	contig06117fw	CTTCTTCTTTCTTCTGGATCTAG
P10-544	contig06117rv	CCCAAAACATCTCTCATTTTTATTTCG
P10-555	contig08562fw	GTAATGGCAAGACTCACTAGC
P10-556	contig08562rv	CCCAATACTTTCCTCATTTCAAGA
P10-547	contig17517fw	CAACAACAACTTTTACAAAGCTCAAAG

P10-548	contig17517rv	CTATAGCAAGCTTATTCCATAAAATAAGTG
P10-439	E5CtermSpecific1	CAGCAGCAACACGTTATCG
P10-440	E5CtermSpecific2	CGTTCTTCGGAGCATTCGC
P10-383	E5ex1fw	CTCAAATCATAGTCTATCATCCTC
P10-384	E5ex1rv	GTCATGGTTTTTTTTATTAATAATAAAAACATAAGTTAAG
P10-387	E5ex2fw	CAAGTACAATCGTCATATAACGTATAATATC
P10-388	E5ex2rv	GACGTCAAAATTTCATAACATATTTTTATTAATTTCAC
P10-393	E5ex4fw	GAATTATAAGATAAGATGGTAAAACGACAAAAC
P10-394	E5ex4rv	CCCATCCTAATCATTGCATCAG
P10-385	E5in1fw	GCCCATCTGTTTTCAATATTATTAAGAATG
P10-386	E5in1rv	GGTTCGGAACGATTTGGAAG
P10-389	E5in2fw	TAGAACAGTGTCTTGTGCAGATA
P10-390	E5in2rv	CTCCCCAACGGAACTG
P10-391	E5in3fw	TTAAAAAAGCTTTTGCTGACGTTGGTT
P10-392	E5in3rv	GAGCTGTCACAAATAGGCATC
P10-432	gWalkingAdaptorStrand1	GTAATACGACTCACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGT
P10-433	gWalkingAdaptorStrand2a	GATCACCAGCCCCT
P10-434	gWalkingAdaptorStrand2b	CCGGACCAGCCCCT
P10-435	gWalkingAdaptorStrand2c	AGCTACCAGCCCCT
P10-218	HRPA2_3UTR_rv	CCAGAGCTTGCCATTATGTG
P10-219	HRPA2_Cterm_fw	CACAATAGCGGTTGTTACCTC
P10-073	HRPA2_Cterm_rev1	CAACTTCCATTAACCTTCTTACAGTC
P10-116	HRPB1_Cterm_rev	CCAAATATTCTTAAGTAATGTTTCGAGAAC
P10-103	HRPB1_ex1_rev	GATCTGATCTTAGCTCGTTCAC
P10-106	HRPB1_ex2_fw	CTGTGATGCATCGATTTTGTTAG
P10-107	HRPB1_ex2_rev	CTGCGCATGATACGGTTC
P10-110	HRPB1_ex3_fw	CAGGAGGTCCTTCTTGG
P10-111	HRPB1_ex3_rev	GTCCAACATTACTAAAGCTGGC
P10-115	HRPB1_ex4_fw2	CAACTCGCTGCTCCATG
P10-114	HRPB1_ex4fw1	GGTCACACATTTGGTAAAAACCAATG
P10-104	HRPB1_in1_fw	GTGAACGAGCTAAGATCAGATC
P10-105	HRPB1_in1_rev	GCTGCATCTTTCTCTGTTCG
P10-108	HRPB1_in2_fw	GAACCGTATCATGCGCAG
P10-109	HRPB1_in2_rev	CAAGATCAAAAAATGCTTGTACGC
P10-112	HRPB1_in3_fw	GCCAGCTTTAGTAATGTTGGAC
P10-113	HRPB1_in3_rev	GTTACTAAAGTTGTATAGTCTGTCC
P10-075	HRPB1_Nterm_fw	CAACTTAAACCAATAAAAGATAAGTTTCCTC
P10-001	HRPC1A_ex1F	CGTTTTGCCTATAAAAGGATTC
P10-002	HRPC1A_ex1R	AAATGATAAATGTAATTAGACAGTATG
P10-003	HRPC1A_ex2F	GACAAAAATGTTACATTGTTGC
P10-004	HRPC1A_ex2R	GTGTTGATATGTAAAGTGACTATTTG
P10-005	HRPC1A_ex3F	CACATATTTTTCTCTTAACACATTG
P10-006	HRPC1A_ex3R	CTAAATACAACTATATCAAATGGTG
P10-008	HRPC1A_ex4R	GATAGAGATCTTCTCATGCTC
P10-017	HRPC1B_ex1F	GGATTATATAAGATATGGACCTTAC
P10-018	HRPC1B_ex1R	CAAAACAAATTGACTCTTTGTATC
P10-019	HRPC1B_ex2F	GTTAGCTATGGATGTAATACATG
P10-020	HRPC1B_ex2R	GAACAGTTGTTGATACTAAATATAG
P10-021	HRPC1B_ex3F	CTTCCGAGTGACAAATTAATCTC

P10-022	HRPC1B_ex3R	CTAATTGAACTTATATCAAATGGTG
P10-023	HRPC1B_ex4F	CACCATTTGATATAAGTTCAATTAG
P10-024	HRPC1Bex4R	CTTCTCCTTCTCAAGTAACATC
P10-228	HRPC1C_ex1_rv	CACTTCCACGACTGCTTTG
P10-231	HRPC1C_ex2_fw	GGTTGTGACGCATCGATC
P10-232	HRPC1C_ex2_rev	CAGATTGTTGAGCTGCAATGG
P10-235	HRPC1C_ex3_fw	GCAGGAGGTCCTTCTTG
P10-236	HRPC1C_ex3_rv	GAGAGCAACGAGATCAGAAG
P10-239	HRPC1C_ex4_fw1	GGTCACACATTTGGTAAAAATCAATG
P10-240	HRPC1C_ex4_fw2	GAACTCAAGGAGAAATCAGGTTG
P10-188	HRPC1C_fw2	GCTTCATGCATCTTTTTCCAATG
P10-229	HRPC1C_in1_fw	CAAAGCAGTCGTGGAAGTG
P10-230	HRPC1C_in1_rv	GTTCGAAATGATGTTGTGTTGTC
P10-233	HRPC1C_in2_fw	CCATTGCAGCTCAACAATCTG
P10-186	HRPC1C_rv2	CACACTACACACCAATAAAGATATTC
P10-009	HRPC2_ex1F	CACTCAACTTCAAACCTAAC
P10-010	HRPC2_ex1R	GAAGTACTTAAACAGGTTTTTACTAG
P10-011	HRPC2_ex2F	GATTTAACTATGAATATGGTAGTTG
P10-012	HRPC2_ex2R	CCTAAAAATTAAAATCAATAAGATGATATG
P10-013	HRPC2_ex3F	CATATCATCTTATTGATTTTAATTTTTAGG
P10-014	HRPC2_ex3R	GAGTTTATTAATGACTACAACTATAG
P10-016	HRPC2_ex4R	CTATAGTTGTAGTCATTAATAAACTC
P10-195	HRPE5_rv1	GATATATATTCCCAACATAATCACATAGAAC
P10-449	newC1CNtermSpecific1a	GTAACTCATTGATAATGATGTCCC
P10-450	newC1CNtermSpecific2	TCATTGATAATGATGTCCCGTACTATG
P10-451	newNNtermSpecific1a	GACAATTTGTAAAAGATTCGGGCAC
P10-453	newNNtermSpecific2	GTGCCCTAACCGCTGAAC
P08-483	pJET1.2fw	CGACTCACTATAGGGAGAGCGGC
P08-484	pJET1.2rv	AAGAACATCGATTTTCCATGGCAG
P07-514	RT-synPDI-rev	ACTTGGACGATAACTGGCTCTTTAG
P09-338	Zeocin:fw	GACTCGGTTTCTCCCGTGACT
P09-337	Zeocin_rv	CTGCGGAGATGAACAGGGTAA
P12022	23190F1	GGAACAACAAGAAGCAGAGAAGAGAGAG
P12023	23190F2	CGACTGAAACAACAAAAAATGGCAATG
P12024	23190F3	ATGGCAATGAGTTATTCGATACGTGTC
P12025	23190R1	GAGATAGTCTTAGGCATTCCACAAACC
P12026	23190R2	GTTATTTTAGATCATGGAAAGAGCTTCC
P12027	23190R3	GGATATGAAACTTGCGGTGTTTCTGG
P12028	04663F1	CAAAGCTCTATCATTATTTGCAACAAAC
P12029	04663F2	CTGATTAATGGCTGCAACAAGCTCTTC
P12030	04663F3	ATGGCTGCAACAAGCTCTTCTACTAC
P12031	04663R1	GACTGAGCAAAAGCCTCAAAAAACAGG
P12032	04663R2	GTTTGGTTACTTGCAAAGGATTACAATC
P12033	04663R3	CTTGCAAAGGATTACAATCGCGATGG
P12034	06351F1	GATTACAAGATTTAAGATAGAAAATAATAAGATGG
P12035	06351F2	GATGGTTAGGGCAAATTTAGTGAGCG
P12036	06351F3	GCAAATTTAGTGAGCGTGATTCTGTTAATGC
P12037	06351R1	GGAGCATAAATACAAAAATCGGCCTAGGC
P12038	06351R2	GACATTATGACTCGAATTAATGACAGAGTAGG

P12039	06351R3	AAATCGGCCTAGGCTTAGTTAATAGTCC
P12040	05508F1	CTAAAAACACACTTGATCTTCTCTAAAACATCG
P12041	05508F2	CACTTGATCTTCTCTAAAACATCGAAACATAAATAC
P12042	05508F3	ATACAAGATGGGTTTGATTAGATCATTATGC
P12043	05508R1	ATCGGTTAATTAATTAATCGCAGAGCAAACC
P12044	05508R2	CACCACTGATTTTAATCGGTTAATTAATTAATCG
P12045	05508R3	CAAACCTTACGAATTTCCCCATTAGTCC
P12046	22489F1	CACTTGATCTTCTCTAAAACACTAAAATTATATATCC
P12047	22489F2	CACTAAAATTATATATCCAATATGGAGTTTGTTAG
P12048	22489F3	CCAATATGGAGTTTGTTAGATCATTATGC
P12049	22489R1	TCTGTTTATCACCACTGATTTTAATCGG
P12050	22489R2	GCAGAGCAAACCCTACGAATTTCC
P12051	22489R3	GATTTTAATCGGTTAATTAATTAACCGCAGAGC
P12195	23190seqF1	GAGAACAATCATCGATCCCGAAC
P12196	23190seqF2	GTCAACAAGCCTTTGTTGTCATCAATAACC
P12197	23190seqF3	CATCGGAATTGCGCATTGTCCGTC
P12198	23190seqF4	CAAGATCCAACCATGAACAAGTCTTTC
P12199	04663seqF1	CCATCGTACGCAGCACTATCCAGC
P12200	04663seqF2	GCAAGCTCTTCAATCCGACCCGAG
P12201	04663seqF3	TTGCCTCAGAGGCTTCCGTGTCTTTG
P12202	04663seqF4	CCTTCGAAGGCCTTAACAACATCAC
P12203	06351seqF1	GCTCTTCAAGCCGATCCCACTTTAGC
P12204	06351seqF2	CCACTTTAGCCGCAGGTCTTATACG
P12205	06351seqF3	CCTTTGGCAACCGTGGCTTCTCTCC
P12206	06351seqF4	GCAAGATGTTGTTGCTCTCTCTGG
P12207	05508seqF1	GCACCCGGAATATTGAGAATGC
P12208	05508seqF2	GCATTTCCACGACTGCTTCGTTCAAG
P12209	05508seqF3	CGTGACTCCGTCGCCGTTCAACAAC
P12210	05508seqF4	CAGTGCGCGTGGATCTCGATAC
P12211	22489seqF1	GGAATATTGAGAATGCATTTCCACG
P12212	22489seqF2	CGACTGCTTCGTTCTAGGTTGTGACG
P12213	22489seqF3	CGTGACTCCGTCGCCGTTCAACAAC
P12214	22489seqF4	CGACACCGGAAGTGGAACCAC
P12320	05508seqF5	ATGTGAGAAACGTTTTACGTGCGTG
P12321	05508seqR5	TTAACATGAAGAGTTTCTCAACG
P12322	05508seqR6	CATGAAGAGTTTCTCAACGTTAATTTTTCG
P11114	03523noSSfw1	ATATCTCGAGAAGAGAGAGGCCGAAGCTAGACTGACTACCAACTTCTAC
P11115	03523noSSrv1	TATAGCGGCCGCATTAGTTGATTG
P11116	04663noSSfw1	ATATCTCGAGAAGAGAGAGGCCGAAGCTCAGTTGAACGCAACCTTCTAC
P11117	04663noSSrv1	TATAGCGGCCGCATTACTTACACAAG
P11118	05508noSSfw1	ATATCTCGAGAAGAGAGAGGCCGAAGCTCAAGCTATCTCCATTTCCATTAC
P11119	05508noSSfw2	ATATCTCGAGAAGAGAGAGGCCGAAGCTACCATCAGAATTGGTTTCTACCTTAC
P11120	05508noSSrv1	TATAGCGGCCGCATTAGTTGATAGC
P11121	06351noSSfw1	ATATCTCGAGAAGAGAGAGGCCGAAGCTTTCCCATTCCACGCCAGAGGTTTG
P11122	06351noSSrv1	TATAGCGGCCGCATTAGTTGATGGTTC
P11123	23190noSSfw1	ATATCTCGAGAAGAGAGAGGCCGAAGCTAAGAAGCCACGTAGAGACGTTC
P11124	23190noSSfw2	ATATCTCGAGAAGAGAGAGGCCGAAGCTGGTTTGTCATGGAACTTCTAC
P11125	23190noSSrv1	TATAGCGGCCGCATTAGATCATAGAAAG
P11126	22489noSSfw1	ATATCTCGAGAAGAGAGAGGCCGAAGCTCAGGCTGCCGCTAGAAGACCAG

P11127	22489noSSfw2	ATATCTCGAGAAGAGAGAGGCCGAAGCTGGTACTAGAATTGGTTTCTACTTAAC
P11128	22489noSSrv1	TATAGCGGCCGCATTAGTTGACTGCTG
P11129	04791noSSfw1	ATATCTCGAGAAGAGAGAGGCCGAAGCTAGATTGACTACCAACTTCTACTCTAAG
P11130	04791noSSrv1	TATAGCGGCCGCATTAATTGATAG
P11131	06117noSSfw1	ATATCTCGAGAAGAGAGAGGCCGAAGCTGACGACGAGTCCAACTACGGTG
P11132	06117noSSfw2	ATATCTCGAGAAGAGAGAGGCCGAAGCTAAGTTGTTCCCTGGATTCTAC
P11133	06117noSSrv1	TATAGCGGCCGCATTAAGAGTTGATC
P11134	17517noSSfw1	ATATCTCGAGAAGAGAGAGGCCGAAGCTAGAAGACCTAGAGTTGGTTTC
P11135	17517noSSfw2	ATATCTCGAGAAGAGAGAGGCCGAAGCTAGACCTAGAGTTGGTTTCTAC
P11136	17517noSSrv1	TATAGCGGCCGCATTAGTTGATGGC
P11137	08562noSSfw1	ATATCTCGAGAAGAGAGAGGCCGAAGCTGACAAATCCTACGGTGGAAAG
P11138	08562noSSfw2	ATATCTCGAGAAGAGAGAGGCCGAAGCTAAGTTGTTCCCAGGTTTCTAC
P11139	08562noSSrv1	TATAGCGGCCGCATTAAGAGTTGATC
P12518	04663R6	ACACAAATGATTCAGCTTAGGAG
P12519	04663R7	GATTCAGCTTAGGAGTAATTATTCATGTATC
P12520	04663R8	CATGTATCATAAACAAACGGTG
P12521	04663R9	GTATCATAAACAAACGGTGAAACTGATCC
P12522	C3F1	GATTCACACCATCAGCCACAC
P12523	C3F2	CGCCTTAGCTAGATTCACACC
P12524	C3F3	GTTCCTTCTACGCCTTAGC
P12525	C3F4	CTCCAGCTACTCTATATAGTGTTCC
P12526	C3F5	CACATATAGAATAGAGGCCAAAAGG
P12527	C3F6	CATCGCCTCTCAAATATCAGTGC
P12528	C3F7	CTTCGATTTGGCTAATACAGCTCTTC
P12529	C3F8	CAACTGACAAACCATAAAAACTTAAACATGC
P12530	C3F9	GAATTAGGGGTATGGAGAACGATG
P12531	C3R1	CGACATGCATGTTCCCACACATTTTATG
P12532	C3R2	GTGAGAGCTTTTATTTAGTCGACATGC
P12533	C3R3	CACAAGTCATAACTCGTGAGAGC
P12534	C3R4	CAGTTGTAATCTCACAAGTCATAACTCG
P12535	C3R5	CTTTTCTTTCCTTTGGTTTTTTCAGTTG
P12536	C3R6	GAGTCGAGAAGAGCTCTTGG
P12537	C3R7	AAAGTCATGATTTTTTCGTTTTACTAATTCATG
P12538	C3R8	GCTTGAAGACATTTTGTTTGAGATGG

Supplementary Table 2. BLASTP of respective full length HRP amino acid sequence against non-redundant protein sequences (nr) database with *Arabidopsis thaliana* (taxid:3702) as organism.

No	HRP	*A. rusticana* peroxidase name (database)	*A. thaliana* gene accession #	*A. thaliana* peroxidase name (database)	identities	positives	max score	total score	query coverage %	E value	max ident %
1	C1A	C1A	NP_190481.1	peroxidase 34	320/353	338/353	649	649	100	0.0	91
2	C1B_15901	C1B	NP_190480.1	peroxidase 33	306/342	320/342	635	635	97	0.0	89
3	C1C_25148	C1C	NP_190480.1	peroxidase 33	303/332	317/332	631	631	100	0.0	91
4	C1CvarD_25148	C1D	NP_190480.1	peroxidase 33	303/332	316/332	630	630	100	0.0	91
5	C2_04627_optPp	C2	NP_192617.1	peroxidase 37	297/332	310/332	610	610	95	0.0	89
6	HRPC3_NCBI_D90116	C3	NP_181373.1	peroxidase 23	313/349	327/349	644	644	100	0.0	90
7	A2_manualassembly	A2A	NP_196290.1	peroxidase 53	318/336	327/336	635	635	100	0.0	95
8	A2variant_manualassembly	A2B	NP_196290.1	peroxidase 53	317/336	326/336	636	636	100	0.0	95
9	E5_manualassembly	E5	NP_181372.1	peroxidase 22	298/349	314/349	597	597	100	0.0	85
10	B1_01805	01805	NP_850652.1	peroxidase 32	326/349	337/349	677	677	98	0.0	93
11	B2gvariant_22684	22684.1	NP_181372.1	peroxidase 22	291/349	318/349	607	607	100	0.0	83
12	B2_22684	22684.2	NP_181372.1	peroxidase 22	289/349	317/349	603	603	100	0.0	83
13	01350	01350	NP_196153.1	peroxidase 52	304/324	312/324	600	600	100	0.0	94
14	02021	01350	NP_188814.1	peroxidase 30	300/331	312/331	613	613	100	0.0	91
15	23190	23190	NP_177313.1	peroxidase 12	327/359	340/359	613	613	99	0.0	91
16	04663	04663	NP_196291.1	peroxidase 54	267/292	278/292	481	481	98	6e-175	91
17	06351	06351	NP_567919.1	peroxidase 47	298/314	308/314	610	610	99	0.0	95
18	03523	03523	NP_189460.1	peroxidase 31	266/299	288/299	495	495	93	1e-180	89
19	05508	05508	NP_201217.1	peroxidase 71	267/329	295/329	538	538	99	0.0	81
20	22489.1	22489.1	NP_201217.1	peroxidase 71	279/329	298/329	563	563	99	0.0	85
21	22489.2	22489.2	NP_201217.1	peroxidase 71	280/329	298/329	565	565	99	0.0	85
22	04791	04791	NP_189460.1	peroxidase 31	266/299	288/299	495	495	93	1e-180	89
23	06117	06117	NP_179407.1	peroxidase 15	311/337	327/337	533	533	99	0.0	92
24	17517.1	17517.1	NP_201215.1	peroxidase 69	275/334	288/334	539	539	100	0.0	82
25	17517.2	17517.2	NP_201215.1	peroxidase 69	276/334	288/334	541	541	100	0.0	83
26	08562.1	08562.1	NP_195361.1	peroxidase 49	311/331	320/331	592	592	99	0.0	94
27	08562.4	08562.2	NP_195361.1	peroxidase 49	311/331	320/331	592	592	99	0.0	94

```
C1C_gDNA          ---MH--SPSSTSFTWATLITLGCLMLHASFS--------NAQLTPTFYDNSCPNVSNIV 47
C1D_gDNA          ---MH--SPSSTSFTWATLITLGCLMLHASFS--------NAQLTPTFYDNSCPNVSNIV 47
C1B_gDNA          ---MH--SPSSTSFTWI-LITLGCLAFYASLS--------DAQLTPTFYDTSCPNVSNIV 46
C1A_gDNA          ---MHF-SSSSTLFTCITLIPLVCLILHASLS--------DAQLTPTFYDNSCPNVSNIV 48
01805_gDNA        ---MHFSTSSSSLSTWTTLITLGCLMLHSFKS--------SAQLTPTFYDSTCPSVFSIV 49
C2_gDNA           ---MH---SSSSLIKLG----FLLLLLNVSLS--------HAQLSPSFYDKTCPQVFDIA 42
C3_gDNA           ---MG--FSPLISCSAMGALILSCLLLQASNS--------NAQLRPDFYFRTCPSVFNII 47
E5_gDNA           ---MV--VSPFFSCSAMGALILGCLLLQASN----------AQLRPDFYSRTCPSVFNII 45
22684.1_gDNA      ---MG--FSPSFSSSSIGVLILGCLLLQASNS--------NAKLRPDFYLKTCPSVFQII 47
22684.2_gDNA      ---MG--FSPSFSSSSIGVLILGCLLLQASNS--------NAKLRPDFYLKTCPSVFQII 47
A2A_gDNA          ---MA---VTNLSTTCDGLFIISLLVIVSSLFGT-----SSAQLNATFYSGTCPNASAIV 49
A2B_gDNA          ---MA---VTNLSTTCDGLFIISLLVIVSSLFGT-----SSAQLNATFYSGTCPNASAIV 49
04663_gDNA        ---MA---ATSSSTTCDGLFIISLLVIASSLFGT-----SSAQLNATFYSGTCPNASAIV 49
08562.4_gDNA      ---MAR--LTSILLLLSLLCFFPLCLCDKS--YG-----G--KLFPGFYAHSCPQAGEIV 46
08562.1_gDNA      ---MAR--LTSILLLLSLLCFFPLCLCDKS--YG-----G--KLFPGFYAHSCPQAGEIV 46
06117_gDNA        ---MAR--IGSFLVVISLACVLTLCICDDESNYG-----GQGKLFPGFYSSSCPKAEEIV 50
01350_gDNA        ---MAS--NQRISILVLVVTFLVQGNYNNV---------VEAQLTPNFYSTSCPNLLSTV 46
23190.1_gDNA      MAMSYSIRVLTFLMLISLMAVTLNLLSTAEAKKPRRDVPIVKGLSWNFYQRACPKVEKII 60
23190.2_gDNA      MAMSYSIRVLTFLMLISLMAVTLNLLSTAEAKKPRRDVPIVKGLSWNFYQRACPKVEKII 60
03523_gDNA        ---MAELKSLSLILLFTLLT------TTIESR-----------LTTNFYSKSCPRFFDIV 40
04791_gDNA        ---MAELKSLSLILLFTLLT------TTIESR-----------LTTNFYSKSCPRFFDIV 40
06351_gDNA        ---MVRANLVSVILLMHVIVG-----FPFHARG----------LSMTYYMMSCPMAEQIV 42
05508.1_gDNA      --------MGLIRSLCVFITFLSCIISSAHGQAISIS---IT-IRIGFYLTTCPTAEIIV 48
05508.2_gDNA      --------MGLIRSLCVFITFLSCIISSAHGQAISIS---IT-IRIGFYLTTCPTAEIIV 48
22489.1           --------MEFVRSLCVFITFLGCLISSAHGQAAARRPGPISGTRIGFYLTTCPTAEIIV 52
22489.2           --------MEFVRSLCVFITFLGCLISSAHGQAAARRPGPISGTRIGFYLTTCPTAEIIV 52
17517.1_gDNA      --------MGRGYNLLLILVTFLVLVAAVTARR----------PRVGFYGNRCRKVESIV 42
17517.2_gDNA      --------MGRGYNLLLILVTFLVLVAAVTARR----------PRVGFYGNRCRKVESIV 42
02021_gDNA        ---MRT--MKRLNVAVAVAVTATVLMGMLGSSE--------AQLQMNFYAKSCPNAEKII 47
                                                                 :*   *

C1C_gDNA          RDIIINELRSDPRIAASILRLHFHDCFVNGCDASILLDNTTSFRTEKDAFGNANSAR-GF 106
C1D_gDNA          RDIIINELRSDPRIAASILRLHFHDCFVNGCDASILLDNTTSFRTEKDAFGNANSAR-GF 106
C1B_gDNA          RDIIINELRSDPRITASILRLHFHDCFVNGCDASILLDNTTSFLTEKDALGNANSAR-GF 105
C1A_gDNA          RDTIVNELRSDPRIAASILRLHFHDCFVNGCDASILLDNTTSFRTEKDAFGNANSAR-GF 107
01805_gDNA        RDTIVNELRSDPRIAASILRLHFHDCFVNGCDASILLDNTTSFRTEKDAAPNANSAR-GF 108
C2_gDNA           TNTIKTALRSDPRIAASILRLHFHDCFVNGCDASILLDNTTSFRTEKDAFGNARSAR-GF 101
C3_gDNA           GDIIVDELRTDPRIAASLLRLHFHDCFVRGCDASILLDNSTSFRTEKDAAPNANSAR-GF 106
E5_gDNA           KNVIVDELQTDPRIAASILRLHFHDCFVRGCDASILLDTSKSFRTEKDAAPNVNSAR-GF 104
22684.1_gDNA      GNVIVDELQSDPRIAASLLRLHFHDCFVRGCDASVLLDNSTSFQSEKDAAPNANSAR-GF 106
22684.2_gDNA      GNVIVDELQSDPRIAASLLRLHFHDCFVRGCDASVLLDNSTSFQSEKDAAPNANSAR-GF 106
A2A_gDNA          RSTIQQAFQSDTRIGASLIRLHFHDCFVNGCDASILLDDSGSIQSEKNAGPNANSAR-GF 108
A2B_gDNA          RSTIQQAFQSDTRIGASLIRLHFHDCFVNGCDASILLDDSGSIQSEKNAGPNANSAR-GF 108
04663_gDNA        RSTIQQALQSDPRIGASLIRLHFHDCFVNGCDGSLLLDDTGSIQSEKNAPANANSAR-GF 108
08562.4_gDNA      RSVVAKAVARETRMAASLMRLHFHDCFVQGCDGSLLLDSSGRIVSEKGSNPNSRSAR-GF 105
08562.1_gDNA      RSVVAKAVARETRMAASLMRLHFHDCFVQGCDGSLLLDSSGKIVSEKGSNPNSRSAR-GF 105
06117_gDNA        RSVVAKAVARETRMAASLMRLHFHDCFVQGCDGSLLLDSSGSIVTEKNSNPNSRSAR-GF 109
01350_gDNA        QSAVKSAVNSEARMGASIVRLFFHDCFVNGCDGSILLDDTSSFTGEQNANPNRNSAR-GF 105
23190.1_gDNA      KKELKKVFKRDIGLAAAILRIHFHDCFVQGCEASVLLAGSASGPGEQSSIPNLTLRQQAF 120
23190.2_gDNA      KKELKKVFKRDIGLAAAILRIHFHDCFVQGCEASVLLAGSASGPGEQSSIPNLTLRQQAF 120
03523_gDNA        RDTISNKQITTPTTAAATIRLFFHDCFPNGCDASILISSTAFNTAERDSSINLSLPGDGF 100
04791_gDNA        RDTISNKQITTPTTAAATIRLFFHDCFPNGCDASILISSTAFNTAERDSSINLSLPGDGF 100
06351_gDNA        KNSVNNALQADPTLAAGLIRMLFHDCFIEGCDASILLDSTKDNTAEKDSPANLSLRG--Y 100
05508.1_gDNA      RNAVRAGFNSDPRIAPGILRMHFHDCFVQGCDGSVLISGS---NTERTAVPNLSLRG--F 103
05508.2_gDNA      RNAVRAGFNSDPRIAPGILRMHFHDCFVQGCDGSVLISGS---NTERTAVPNLSLRG--F 103
22489.1           RNAVRAGFNSDPRIAPGILRMHFHDCFVLGCDGSVLISGS---NTERTAVPNLNLRG--F 107
22489.2           RNAVRAGFNSDPRIAPGILRMHFHDCFVLGCDGSVLISGS---NTERTAVPNLNLRG--F 107
17517.1_gDNA      RSVVRSHFRCNPANAPGILRMYFHDCFVNGCDGSILLAGN---TSERTAGPNRSLRG--F 97
17517.2_gDNA      RSVVRSHFRCNPANAPGILRMHFHDCFVNGCDGSILLAGN---TSERTAGPNRSLRG--F 97
02021_gDNA        SDHIQKHIPSGPSLAAPLIRMHFHDCFVRGCDGSVLINSTSG-NAEKDSAPNLTLRG--F 104
                   . :           .  :*: *****  **:.*:*:  .     *: :  *       :

C1C_gDNA          PVVDRIKAAVERACPRTVSCADVLTIAAQQSVNLAGGPSWRVPLGRRDSRQAFLDLANAN 166
C1D_gDNA          PVVDRIKAAVERACPRTVSCADVLTIAAQQSVNLAGGPSWRVPLGRRDSRQAFLDLANTN 166
C1B_gDNA          PTVDRIKAAVERACPRTVSCADVLTIAAQQSVNLAGGPSWRVPLGRRDSLQAFLDLANAN 165
C1A_gDNA          PVIDRMKAAVESACPRTVSCADLLTIAAQQSVTLAGGPSWRVPLGRRDSLQAFLDLANAN 167
01805_gDNA        PVIDTMKAAVERACPRTVSCADLLTIAAQQSVNLAGGPSWRVPLGRRDSVQAFFDLANTN 168
C2_gDNA           DVIDTMKAAVEKACPKTVSCADLLAIAAQKSVVLAGGPSWKVPSGRRDSLRGFMDLANDN 161
C3_gDNA           GVIDRMKTSLERACPRTVSCADVLTIASQISVLLSGGPWWPVPLGRRDSVEAFFDLANTA 166
E5_gDNA           NVIDRMKTALERACPRTVSCADILTIASQISVLLSGGPSWAVPLGRRDSVEAFFDLANTA 164
22684.1_gDNA      DVVDRMKAALEKACPGTVSCADVLAISAQISVLLSGGPWWPVLLGRRDGVEAFFDLANTA 166
22684.2_gDNA      DVVDRMKAALEKACPGTVSCADVLAISAQISVLLSGGPWWPVLLGRRDGVEAFFDLANTA 166
A2A_gDNA          NVVDNIKTALENTCPGVVSCSDILALASEASVSLTGGPSWTVLLGRRDSLTANLAGANSA 168
A2B_gDNA          NVVDNIKTALENTCPGVVSCSDILALASEASVSLTGGPSWTVLLGRRDSLTANLAGANSA 168
04663_gDNA        NVVDDIKTALENACPGIVSCSDILALASEASVSLAGGPSWTVLVGRRDGLTANLSGANSS 168
08562.4_gDNA      DVVDQIKAELEKQCPGTVSCADALTLAARDSSVLTGGPSWVVSLGRRDSRSASLSGSNNN 165
08562.1_gDNA      DVVDQIKAELEKQCPGTVSCADALTLAARDSSVLTGGPSWVVSLGRRDSRSASLSGSNNN 165
06117_gDNA        EVVDEIKAALENECPNTVSCADALTLAARDSSVLTGGPSWMVPLGRRDSTSASLSGSNNN 169
```

```
01350_gDNA        NVIDNIKAAVEKACPGVVSCADILAIAARDSVVVLGGPNWTVKVGRRDARTASQAAANSN 165
23190.1_gDNA      VVINNLRALVQKQCGQVVSCSDILALAARDSIVLSGGPDYAVPLGRRDSLAFATPETTLA 180
23190.2_gDNA      VVINNLRALVQKQCGQVVSCSDILALAARDSIVLSGGPDYAVPLGRRDSLAFATPETTLA 180
03523_gDNA        DVIVRAKTAIELACPNTVSCSDIITVATRDLLVTVGGPYYDVYLGRRDSRISKSSLLTDL 160
04791_gDNA        DVIVRAKTAIELACPNTVSCSDIITVATRDLLVTVGGPYYDVYLGRRDSRISKSSLLTDL 160
06351_gDNA        EIIDDAKEKVENMCPGVVSCADIVAMAARDAVFWAGGPYYDIPKGRFDGKRSK-IEDTRN 159
05508.1_gDNA      EVIENAKTQLEATCPGVVSCADILALAARDTVVLTRGIGWQVPTGRRDGRVS-VASNANN 162
05508.2_gDNA      EVIENAKTQLEAACPGVVSCADILALAARDTVVLTRGIGWQVPTGRRDGRVS-VASNANN 162
22489.1           EVIDNAKTQLEATCPGVVSCADILALAARDTVVLTRGLGWQVPTGRRDGRVS-VASNANN 166
22489.2           EVIDNAKTQLEATCPGVVSCADILALAARDTVVLTRGLGWQVPTGRRDGRVS-VASNANN 166
17517.1_gDNA      EAIEEAKTRLENACPNTVSCADILTLAARDAVVWTGGKGWSVPLGRLDGRRS-EASDVN- 155
17517.2_gDNA      EAIEEAKTRLENACPNTVSCADILTLAARDAVVWTGGKGWSVPLGRLDGRRS-EASDVN- 155
02021_gDNA        GFVERIKTLLEAECPKTVSCADIIALTARDAVVATGGPSWKVPTGRRDGRISNTTEALNN 164
                    :   :  ::  *   ***:* :::::.       *  : :  ** *.

C1C_gDNA          -LPAPSFTLPELKAAFANVGLNRPSDLVALSGGHTFGKNQCRFIMDRLYNFSNTGLPDPT 225
C1D_gDNA          -LPAPSFTLPQLKAAFANVGLNRPSDLVALSGGHTFGKNQCRFIMDRLYNFSNTGLPDPT 225
C1B_gDNA          -LPAPFFTLPQLKDAFAKVGLDRPSDLVALSGGHTFGKNQCRFIMDRLYNFSNTGLPDPT 224
C1A_gDNA          -LPAPFFTLPQLKDSFRNVGLNRSSDLVALSGGHTFGKNQCRFIMDRLYNFSNTGLPDPT 226
01805_gDNA        -LPAPFFTLPQLKASFSNVGLDRPEDLVALSGGHTFGKNQCQFIMDRLYNFSNTGLPDPT 227
C2_gDNA           -LPGPSSTLQVLKDKFRNVGLDRPSDLVALSGGHTFGKNQCQFIMDRLYNFSNSGKPDPT 220
C3_gDNA           -LPSPFFTLAQLKKAFADVGLNRPSDLVALSGGHTFGRAQCQFVTPRLYNFNGTNRPDPT 225
E5_gDNA           -LPSPFFTLAQLKKAFADVGLNRPSDLVALSGGHTFGRARCLFVTARLYNFNGTNRPDPT 223
22684.1_gDNA      -LPNPFAPLTELKEKFADVGLKRASDLVALSGAHTFGRAQCLLVTPRLYNFSGTNKPDPT 225
22684.2_gDNA      -LPNPFAPLTELKEKFADVGLKRASDLVALSGAHTFGRAQCLLVTPRLYNFSGTNKPDPT 225
A2A_gDNA          -IPSPFEGLSNITSKFSAVGLN-TNDLVALSGAHTFGRARCGVFNNRLFNFSGTGNPDPT 226
A2B_gDNA          -IPSPFEGLSNITSKFSAVGLN-TNDLVALSGAHTFGRARCGVFNNRLFNFSGTGNPDPT 226
04663_gDNA        -LPSPFEGLNNITSKFLAVGLN-TTDVVVLSGAHTFGRGQCVTFNNRLFNFNGTGSPDPT 226
08562.4_gDNA      -IPAPNNTFQTILSKFNRQGLD-VTDLVALSGSHTIGFSRCTSFRQRLYNQSGNGRPDMT 223
08562.1_gDNA      -IPAPNNTFQTILSKFNRQGLD-VTDLVALSGSHTIGFSRCTSFRQRLYNQSGNGRPDMT 223
06117_gDNA        -IPAPNNTFNTILSRFNSQGLD-LTNVVALSGSHTIGFSRCTSFRQRLYNQSGNGSPDTT 227
01350_gDNA        -IPAPTSSLSQLISSFSAVGLS-TRDMVALSGAHTIGQSRCTSFRTRIYN-------ETN 216
23190.1_gDNA      NLPPPFANASQLISDFNDRNLN-ITDLVALSGGHTIGIAHCPSFTDRLYPNQ-----DPT 234
23190.2_gDNA      NLPPPFANASQLISDFNDRNLN-ITDLVALSGGHTIGIAHCPSFTDRLYPNQ-----DPT 234
03523_gDNA        -LPLPSSPISKTIRQFESKGFT-IQEMVALSGAHSIGFSHCKEFVNRVAGN------NTG 212
04791_gDNA        -LPLPSSPISKTIRQFESKGFT-IQEMVALSGAHSIGFSHCKEFVNRVAGN------NTG 212
06351_gDNA        -LPSPFLNASQLIQTFGNRGFS-PQDVVALSGAHTLGVARCSSFKARLTTP------DSS 211
05508.1_gDNA      -LPGPRDSVAVQQQKFSALGLN-TRDLVVLAGGHTLGTAGCGVFRDRLFNN-----TDPN 215
05508.2_gDNA      -LPGPRDSVAVQQQKFSALGLN-TRDLVVLAGGHTLGTAGCGVFRDRLFNN-----TDPN 215
22489.1           -LPGPRDSVAVQQQKFSAVGLN-TRDLVVLAGGHTIGTAGCGVFRDRLFNN-----TDPN 219
22489.2           -LPGPRDSVAVQQQKFSAVGLN-TRDLVVLAGGHTIGTAGCGVFRDRLFNN-----TDPN 219
17517.1_gDNA      -LPGPSDPVAKQKQDFAAKNLN-TLDLVTLVGGHTIGTAGCGLVRGRFFNFNGTGQPDPS 213
17517.2_gDNA      -LPGPSDPVAKQKQDFAAKNLN-TLDLVTLVGGHTIGTAGCGLVRGRFFNFNGTGQPDPS 213
02021_gDNA        -IPPPTSNFTTLQRLFANQGLN-LKDLVLLSGAHTIGVSHCSSMNTRLYNFSTTVKQDPS 222
                   :* *          *   .:    ::* * *.*::*   *  .  *.         :

C1C_gDNA          LNTTYLQTLRQ-QCPRNGN--QSVLVDFDLRTPTVFDNKYYVNLKEQKGLIQSDQELFSS 282
C1D_gDNA          LNTTYLQTLRQ-QCPRNGN--QSVLVDFDLRTPTVFDNKYYVNLKEQKGLIQSDQELFSS 282
C1B_gDNA          LNTTYLQTLRQ-QCPLNGN--QSVLVDFDLRTPTVFDNKYYVNLKEQKGLIQSDQELFSS 281
C1A_gDNA          LNTTYLQTLRG-LCPLNGN--LSALVDFDLRTPTIFDNKYYVNLEEQKGLIQSDQELFSS 283
01805_gDNA        LNTTYLQTLRV-QCPRNGN--QSVLVDFDLRTPTVFDNKYYVNLKEHKGLIQTDQELFSS 284
C2_gDNA           LDKSYLSTLRK-QCPRNGN--LSVLVDFDLRTPTIFDNKYYVNLKENKGLIQSDQELFSS 277
C3_gDNA           LDPTYLVQLRA-LCPQNGN--GTVLVNFDVVTPNTFDRQYYTNLRNGKGLIQSDQELFST 282
E5_gDNA           LNPSYLADLRR-LCPRNGN--GTVLVNFDVMTPNTFDNQFYTNLRNGKGLIQSDQELFST 280
22684.1_gDNA      LNPSYLVELRR-LCPQNGN--GTVLLNFDLVTPNAFDRQYYTNLRNGKGLIQSDQELFST 282
22684.2_gDNA      LNPSYLVELRR-LCPQNGN--GTVLLNFDLVTPNAFDRQYYTNLRNGKGLIQSDQELFST 282
A2A_gDNA          LNSTLLSSLQQ-LCPQNGS--ASTITNLDLSTPDAFDNNYFANLQSNNGLLQSDQELFST 283
A2B_gDNA          LNSTLLSSLQQ-LCPQNGS--ASTITNLDLSTPDAFDNNYFANLQSNNGLLQSDQELFST 283
04663_gDNA        LNSTLLSSLQQ-ICPQNGS--GSAITNLDLTTPDAFDSNYYTNLQSNNGLLQSDQELFSN 283
08562.4_gDNA      LEQSFAANLRQ-RCPRSGG--DQILSVLDIISAAKFDNSYFKNLIENKGLLNSDQVLFNS 280
08562.1_gDNA      LEQSFAANLRQ-RCPRSGG--DQILSVLDIISAAKFDNSYFKNLIENKGLLNSDQVLFSS 280
06117_gDNA        LEQSYAANLRH-RCPRSGG--DQNLSELDINSAGRFDNSYFKNLIENMGLLNSDQVLFSS 284
01350_gDNA        INAAFATTRQR-TCPRTSGSGDGNLAPLDVTTAASFDNNYFKNLMTQRGLLHSDQELFNG 275
23190.1_gDNA      MNKSFANSLKR-TCP--TAN-SSNTQVNDIRSPDVFDNKYYVDLMNRQGLFTSDQDLFVD 290
23190.2_gDNA      MNKSFANSLKR-TCP--TAN-SSNTQVNDIRSPDVFDNKYYVDLMNRQGLFTSDQDLFVD 290
03523_gDNA        YNPRFAQALKQ-ACSNYPKD-PTLSVFNDIMTPNRFDNMYYQNIPKGLGLLESDHGLYSD 270
04791_gDNA        YNPRFAQALKQ-ACSNYPKD-PTLSVFNDIMTPNRFDNMYYQNIPKGLGLLESDHGLYSD 270
06351_gDNA        LDSTFANTLTR-TCN--AGD-NAEQPFD--ATRNDFDNAYFNALQRKSGVLFSDQTLFNT 265
05508.1_gDNA      VDQPFLTQLQT-KCPRNGD--GSVRVDLDTGSGTTFDNSYFINLSRGRGVLESDHVLWTD 272
05508.2_gDNA      VDQPFLTQLQT-KCPRNGD--GSVRVDLDTGSGTTFDNSYFINLSRGRGVLESDHVLWTD 272
22489.1           VNQLFLTQLQT-QCPQNGD--GAVRVDLDTGSGTTFDNSYFINLSRGRGVLESDHVLWTD 276
22489.2           VNQLFLTQLQT-QCPQNGD--GSVRVDLDTGSGTTFDNSYFINLSRGRGVLESDHVLWTD 276
17517.1_gDNA      IDPSFVPLVQA-RCPQNGN--ATTRVDLDTGSAGDFDTSYLSNVRSSRVVLQSDLVLWKD 270
17517.2_gDNA      IDPSFVPLVQA-RCPQNGN--ATTRVDLDTGSAGDFDTSYLSNVRSSRVVLQSDLVLWKD 270
02021_gDNA        LDSEYAANLKANKCKSLND--NTTILEMDPGSSKTFDLSYYRLVLKRRGLFQSDSALTTN 280
                   :           *                 :   **  :   :     :: :*  *

C1C_gDNA          PNATDTIPLVRSYADGTQ---TFFNAFVEAMNRMGNITPLTG-TQGEIRLNCRVVNSNSL 338
```

```
C1D_gDNA          PNATDTIPLVRSYADGTQ---TFFNAFVEAMNRMGNITPLTG-TQGEIRLNCRVVNSNSL 338
C1B_gDNA          PNATDTIPLVRSFADGTQ---KFFNAFVEAMNRMGNITPLTG-TQGEIRLNCRVVNSNSL 337
C1A_gDNA          PNATDTIPLVRSFANSTQ---TFFNAFVEAMDRMGNITPLTG-TQGQIRLNCRVVNSNSL 339
01805_gDNA        PNAADTIPLVRSYADGTQ---KFFNAFMEAMNRMGNITPLTG-TQGQIRQNCRVINSNSL 340
C2_gDNA           PDASDTIPLVRAYADGQG---KFFDAFVEAMIRMGNLSPSTG-KQGEIRLNCRVVNSKPK 333
C3_gDNA           P-GADTIPLVNLYSSNTF---AFFGAFVDAMIRMGNLRPLTG-TQGEIRQNCRVVNSR-- 335
E5_gDNA           P-GADTIPLVNLYSSNTL---SFFGAFADAMIRMGNLRPLTG-TQGEIRQNCRVVNSR-- 333
22684.1_gDNA      P-GADTIPLVNLYSKNTF---AFFGAFVDAIIRMGNIQPLTG-TQGEIRQNCRVVNSR-- 335
22684.2_gDNA      P-GADTIPLVNLYSKNTF---AFFGAFVDAIIRMGNIQPLTG-TQGEIRQNCRVVNSR-- 335
A2A_gDNA          T-GSATIAVVTSFASNQT---LFFQAFAQSMINMGNISPLTG-SNGEIRLDCKKVNGS-- 336
A2B_gDNA          T-GSATITVVTSFASNQT---LFFQAFAQSMINMGNISPLTG-SNGEIRLDCKKVNGS-- 336
04663_gDNA        T-GSPTIAIVNSFASNQT---LFFEAFAQSMIKMGNISPLTG-TSGEIRQDCKAVNGQSS 338
08562.4_gDNA      N--EKSRELVKKYAEDQG---EFFEQFAESMIKMGNISPLTG-SSGEIRKNCRKINS--- 331
08562.1_gDNA      N--EKSRELVKKYAEDQG---EFFEQFAESMIKMGNISPLTG-SSGEIRKNCRKINS--- 331
06117_gDNA        N--DESRELVKKYAEDQE---EFFEQFAESMVKMGNISPLTG-SSGQIRKNCRKINS--- 335
01350_gDNA        G--S-TDSIVRGYSNNPS---SFSSDFAAAMIKMGDISPLTG-SSGEIRKVCGRTN---- 324
23190.1_gDNA      K---RTRGIVESFAIDQN---LFFDHFTVAMIKMGQMSVLTG-TQGEIRSNCSARNTASF 343
23190.2_gDNA      K---RTRGIVESFAIDQN---LFFDHFTVAMIKMGQMSVLTG-TQGEIRSNCSARNTASF 343
03523_gDNA        P---RTRPFVDLYARDQD---LFFKDFARAMQKLSLFGVKTG-RRGEIRRRCDAIN---- 319
04791_gDNA        P---RTRPFVDLYARDQD---LFFKDFARAMQKLSLFGVKTG-RRGEIRRRCDAIN---- 319
06351_gDNA        P---RTRNLVNGYALNQA---KFFFDFQQAMRKMSNLDVKLG-SQGEIRQNCRTIN---- 314
05508.1_gDNA      P---ATRPIVQQLMSSSG---NFNAEFARSMVKMSNIGVVTG-TNGEIRKVCSAIN---- 321
05508.2_gDNA      P---ATRPIVQQLMSSSG---NFNAEFARSMVKMSNIGVVTG-TNGEIRKVCSAIN---- 321
22489.1           P---ATRPIVQQLMSPRG---NFNAEFARSMVRMSNIGVVTG-ANGEIRRVCSAVN---- 325
22489.2           P---ATRPIVQQLMSPRG---NFNAEFARSMVRMSNIGVVTG-ANGEIRRVCSAVN---- 325
17517.1_gDNA      T---ETRAIIERLLGLRRPVLRFGSEFGKSMTKMSLIEVKTRLSDGEIRRVCSAIN---- 323
17517.2_gDNA      T---ETRAIIERLLGLRRPVLRFGSEFGKSMTKMSLIEVKTRLSDGEIRRVCSAIN---- 323
02021_gDNA        S---ATLKMINDLVNGPEK--KFLKAFAKSMEKMGRVKVKTG-SAGVIRTRCSVAGS--- 331
                       :  .:            *   *  :: .:. .        * **  *   .

C1C_gDNA          LHDIVEVVDFVSSM------ 352
C1D_gDNA          LHDIVEVVDFVSSM------ 352
C1B_gDNA          LHDIVEVVDFVSSM------ 351
C1A_gDNA          LHDMVEVVDFVSSM------ 353
01805_gDNA        LHDIVEIVDFVSSM------ 354
C2_gDNA           IMDVVDTNDFASSI------ 347
C3_gDNA           IRGMENDDGVVSSI------ 349
E5_gDNA           IRGMENDDGVVSSM------ 347
22684.1_gDNA      IKGMENDGGVVSSI------ 349
22684.2_gDNA      IRGMENDDGVVSSI------ 349
A2A_gDNA          --------------------
A2B_gDNA          --------------------
04663_gDNA        ATKAEDIQMQSDGPVSLADM 358
08562.4_gDNA      --------------------
08562.1_gDNA      --------------------
06117_gDNA        --------------------
01350_gDNA        --------------------
23190.1_gDNA      ISVLEEGIVEEALSMI---- 359
23190.2_gDNA      ISVLVEGIVEEALSMI---- 359
03523_gDNA        --------------------
04791_gDNA        --------------------
06351_gDNA        --------------------
05508.1_gDNA      --------------------
05508.2_gDNA      --------------------
22489.1           --------------------
22489.2           --------------------
17517.1_gDNA      --------------------
17517.2_gDNA      --------------------
02021_gDNA        --------------------
```

Supplementary figure 1. Alignment of the amino acid sequences of the HRP isoenzymes

CHAPTER 3

Expression of horseradish peroxidase surface variants in *Pichia pastoris*

Laura Näätsaari[1], Martin Kulterer[2], Philipp Nothdurft[2], Volker Ribitsch[2] and Anton Glieder[3]

[1] Institute of Molecular Biotechnology, Graz University of Technology, Petersgasse 14, A-8010 Graz, Austria
[2] Institute of Chemistry, University of Graz, Heinrichstraße 28, A-8010 Graz, Austria
[3] Austrian Centre of Industrial Biotechnology (ACIB GmbH), Petersgasse 14, A-8010 Graz, Austria
[*] To whom correspondence should be addressed. Tel: +43 316 8739300; Fax: +43 316 873 9301;
Email: glieder@glieder.com

Manuscript in preparation for submission in "Protein Expression and Purification"

Introduction

Horseradish peroxidases (HRPs) isolated from the roots *Armoracia rusticana* represent a diverse set of enzymes traditionally used in biomedical applications and assays due to the availability of substrates forming stable, coloured products upon HRP catalysed reaction with hydrogen peroxide. Lately, HRPs have also been gaining increasing interest in the field of biotechnology (1). Suggested applications in biocatalysis, biosensors and diagnostics require stable enzymes of consistent quality. HRP has also been reported to be useful in the bioremediation of polluted waters and soils. Aromatic compounds like phenol derivatives originating from industrial production of, for example, textiles, abrasives, wood composites and coatings are one of the major pollutants in wastewates (2–4). Benzidine-based dyes are potentially carcinogenic and many bis-phenolic compounds, dioxin and pesticides can function as endocrine disruptors (5). Due to the toxicity of the compounds, their use and removal from waste waters is very tightly regulated in many countries, but requires complicated, cost-intensive methods (6). Also hormones, e.g. 17α-ethynylestradiol originating from contraceptive medication and ending up in municipal waste waters can cause a threat to the natural ecosystems if let to the nature unprocessed (7). However, utilization of HRPs in large-scale applications such as wastewater treatment requires enzyme immobilization to enable enzyme recovery and retention. Immobilization of an enzyme to carrier materials can also improve stability and enable enzyme performance under optimal process conditions (8).

Immobilization of HRP has been reported to improve resistance to proteolysis and to increase storage stability (5). Substantially high retention of catalytic activity has been achieved in orientated immobilization of surface engineered HRP produced in *E. coli* (9). Traditional approaches employing adsorption as the immobilization technique suffer from several drawbacks including denaturation, non-specific binding and too weak binding capacity (10). Especially the immobilization of unmodified recombinant HRP C1A produced in *E.coli* to traditional adsorbent solid phases has been noted to be difficult due to the lack of carbohydrate residues on the surface (11). The glycosylation grade of HRP produced in *Pichia pastoris* is known to be higher and more heterogeneous than that of the plant enzyme (12), with a total carbohydrate content of up to 65%. The extremely high glycosylation grade of HRP C1A produced in this methylotrophic yeast could offer advantages for adsorption-based binding or use with novel multifunctional supports (8, 13). Furthermore, the hyper-glycosylation could increase the temperature, solvent and hydrogen peroxide stabilities of the enzyme.

In addition to the traditional, adsorption based immobilization methods, covalent linkages utilizing reactive amino acids have emerged and been reported to be the most robust techniques for immobilization (14, 15). Fissured surfaces have high loading capacity and multifunctional supports exist for enhanced binding (8, 13). Lysine residues provide good bond stability and above average reactivity. Therefore lysines located on the surface of proteins are typically used for covalent attachment to activated supports with epoxide functional groups covering the reactive surface (16). In several protein engineering strategies surface amino acids have been modified to lysine, simultaneously aiming to improve enzyme stability and the possibilities for oriented immobilization (17). In 2001, O'Brien *et al.* (18) used commercial plant HRP to identify the natural lysine residues accessible to crosslinking compound EGNHS used for protein stabilization. They concluded that from the 6 surface lysines (Lys65, Lys84, Lys149, Lys174, Lys232 and Lys241), only Lys174, Lys232 and Lys241 are well accessible to chemical modifications and thus also for immobilization. Ryan and O'Fagain analysed single- and double-mutants of solvent-exposed lysine and glutamic acid residues of HRP produced in *E. coli*. Two lysine single-mutants (K232N, K241F) and two double-mutants (K232N/K241N and K232F/K241N) exhibited increased hydrogen peroxide and heat tolerance (19, 20). Later in 2007, Ryan and O'Fagain (9) also replaced arginine residues opposite to the active site by lysines, providing a means of oriented multipoint covalent immobilization of *E. coli* produced, glycan-free HRP. However, although the oriented immobilization was successful, some loss of thermal, solvent and hydrogen peroxide stability were observed.

To the best of our knowledge, no immobilization studies utilizing engineered HRP surface variants produced by a eukaryotic host which allows typical eukaryotic posttranslational protein modifications exist until now. The multiple HRP isoenzymes with various glycosylation grades discovered in our previous studies (**Chapter 2**) provide a basis for experiments with immobilization techniques using both carbohydrate-mediated adsorption and covalent binding. In contrast to other studies where the number of lysine residues on the surface was increased, in this study single and double mutants to remove one or two lysines proximal to the substrate access channels were created. The proteins were produced in *P. pastoris* to provide heterologously produced single isoenzyme with known structure and surface composition to determine the optimal orientation for immobilization and avoidance of steric hindrances of the reaction.

Materials and methods

Chemicals, media and materials

All *P. pastoris* and *E. coli* strains were cultured in standard media using chemicals and other components as previously described by (21). For selection, Zeocin™ (Invitrogen Corp., Carlsbad, CA) was added to 100µg/ml (*P. pastoris)* or 25µg/ml (*E. coli*) as required. ABTS (2,2'-azino-bis(3-ethylbenzthiazoline-6-sulfonic acid) (30931-67-0), HRP type IV-A used as a positive control (9003-99-0), 17α-ethynylestradiol (57-63-6), Low Molecular Weight Chitosan (9012-76-4) and sodium carboxymethyl cellulose (9004-32-4) were purchased from Sigma-Aldrich Handels Gmbh (Vienna, Austria). Cellulose acetate was produced by Pentair/X-Flow (Netherlands). All other chemicals used in the immobilization were purchased from Roth (Karlsruhe, Germany).

Construction of the HRP expression vector

The HRP C1A surface variant genes were codon optimized to *Pichia pastoris* codon usage using GeneDesigner (22) according to the codon usage tables published previously (23). The synthetic fragments produced also contained a codon optimized *Saccharomyces cerevisiae* 267bp α-factor signal sequence known to be functional for secretion by *P. pastoris*. Cloning was performed with EcoRI and NotI into the *P. pastoris* shuttle vector pPpT4_S (**Chapter 1**) and the resulting expression vector was amplified using the *Escherichia coli* strain TOP10F'. Plasmids were isolated using GeneJET™ Plasmid Miniprep Kit (Fermentas, St. Leon-Rot, Germany) and the correct insertion was confirmed by Sanger sequencing the resulting plasmids.

Transformation of *P. pastoris* CBS7435 mut^s and primary screening of positive transformants

The pPpT4_S expression vectors containing HRP C1A variants were linearized using SmiI, and *P. pastoris* CBS7435 mut^s was transformed with 3.5µg of the purified linear fragments. *P. pastoris* competent cells were prepared with the condensed protocol and transformations were performed by electroporation, essentially as described by (24) using electroporation at 2.0kV, 200Ω and 25µF. After electroporation, 500µl of ice-cold 1M sorbitol and 500µl YPD were added. The cells were allowed to regenerate for 2h in 28°C, 130rpm prior to plating on YPD agar containing 100µg/ml Zeocin™ (Invitrogen). Screening for active clones was performed in 96-well deep-well plates as previously described by (25).

Rescreening and large-scale production of HRP C1A variants in *P. pastoris*

A set of four positive clones were streaked out to single colonies and rescreened to verify the clone homogeneity and the results from screening. Glycerol stocks were prepared from a single, rescreened colony according to standard protocols and stored in -80°C. Fermentations were performed in 5l working volume in BiostatC bioreactors (Sartorius AG, Göttingen, Germany). Chemicals, media compositions and cultivation protocols were applied as described previously (21) with the following modifications: batch medium contained glycerol as carbon source 40g l^{-1} and glycerol substrate feed contained 700g l^{-1} glycerol. Standard fermentation protocols were used with total glycerol and methanol feeds of 400g and 210g, 700g and 230g, 630g and 220g, 800g and 180g for HRP C1A and variants HRPC1AsynK232N/K241N, HRPC1AsynK174R/K241N and HRPC1AsynK174Q/K241F, respectively.

Downstream processing and SDS-PAGE analysis

The cells were separated from the protein-containing supernatant with repeated centrifugations. Essentially cell-free supernatant was filtered through 0.8µm and 0.45µm cellulose acetate filters (Sartorius, Göttingen, Germany). Protein concentration was performed using cross-flow ultrafiltration with 30 kDa cutoff and Vivaspin® centrifugal concentration units (both from Sartorius) with 10 kDa cutoff. The proteins were separated using precast NuPAGE® Bis-Tris Gels and SeeBlue Plus2 prestained protein ladder (both from Invitrogen) in 1xMOPS buffer in reducing conditions.

Immobilization of HRP to cellulose acetate membranes

To test the properties of the HRP C1A surface variants in enzyme immobilization to surfaces, silicon-dioxide wafers and hollow fibre membranes were coated with cellulose acetate (CA) and carboxymethyl cellulose (CMC) using methods described by Kulterer *et al.* (manuscript in preparation). Also the possibilities to use 1-(3-dimethylaminopropyl)-3-ethylcarbodiimine (EDC) coupling agent for covalent binding of HRP onto polysaccharide surfaces were mapped out as described by Kulterer *et al*. (manuscript in preparation). In short, for the covalent binding approach, the CMC surface was activated with EDC, washed twice with double-distilled water and dried prior to adding the raw fermentation supernatant including the enzyme.

HRP activity assays

Peroxidase activity of free and immobilized HRP was measured essentially as described previously (12) in 50mM sodium acetate buffer pH4.5 with 1mM ABTS and 0.0026% (v/v) $H_2O_{2.}$ For all assays of free enzyme, 15µl cultivation supernatant was mixed with 140µl of the assay solution in a flat-bottom 96-well microtiterplate (Greiner Bio-One GmbH, Frickenhausen, Germany). The reaction kinetics were followed with Spectramax Plus^{384} spectrophotometer and SoftMax® Pro software (Molecular Devices, LLC) for 3-5min at a wavelength of 405nm. The efficiency of HRP C1A surface variants in degrading one of the endocrine disrupting compounds, 17α-ethynylestradiol, was tested as described by Kulterer *et al*. (manuscript in preparation).

Results and discussion

Increased stability and the possibility for oriented immobilization of HRPs are desired features for the diverse uses of HRPs in biotechnological applications. Codon-optimized wild-type HRP C1A and 14 HRP C1A surface variants (Table 1) were cloned into the *E. coli/P. pastoris* shuttle vector pPpT4 with leader sequence for efficient secretion into the culture supernatant. Correct insertion was verified by sequencing. *P. pastoris* CBS7435 mut^s was successfully transformed with SmiI linearized expression vectors. The resulting single colonies growing on Zeocin™ containing media were transferred to 96-well deep-well plates for cultivation and induction of heterologous protein production. Peroxidase activity of the clones was assayed with ABTS assay after 72h of methanol induction. The expression levels of each enzyme variant after standard cultivation and methanol induction are depicted in Table 1. The peroxidase activities of all variants were detectable, but varied from very low activity to activities comparable to or even higher than the wild-type enzyme. To increase the chance to obtain active enzyme variants only amino acids occurring also in natural enzyme variants at the same mutated positions were used to replace the lysine residues. Due to the weak reproducibility of high-throughput cultivation methods in deep-well plates, the absolute activities reached were not compared between variants. However, they prove that at least (>100mABS/min using standard ABTS assay) 12 of the 14 surface variants can be produced in *P. pastoris*. The resulting landscapes were typical for the expression vector pPpT4_S preferring low-copy integration (**Chapter 1**). The results from rescreening of the best-producing clones of each variant were consistent with the primary screening results, suggesting expected clone homogeneity and stability of the expression strains.

The production of three surface variants HRPC1AsynK232N/K241N, HRPC1AsynK174R/K241N and HRPC1AsynK174Q/K241F was up-scaled to 5l reactor volume (BiostatC fermenter), yielding expression levels of 34,1U/ml; 30,2U/ml and 44,3U/ml, respectively, in the raw cultivation supernatant. The yield of codon optimized C1A without modifications was 5,3U/ml. The wild-type HRP C1A was not cultivated in the same batch as the surface variants. Thus, the lower enzyme activity of the wild-type HRP C1A can be connected to the more optimized feeds used for the cultivation of the surface variants rather than to the absolute productivity of the strain.

SDS-PAGE analysis of the concentrated cultivation supernatants shows enzymes appearing as heterogenous but identifiable main bands on the gel despite the lack of any purification steps (Figure 1). As described previously (12), the HRP C1A isoenzyme produced in *P. pastoris* is heterogenously glycosylated and thus appears as a smear around 60 kDa. Since purification of heavily glycosylated proteins can be challenging, the purification effect of protein secretion to the supernatant is of advantage. The first experiments to degrade 17α-ethynylestradiol with concentrated unpurified cultivation supernatants were successful. The wild-type isoenzyme C1A and all variants tested could degrade 17α-ethynylestradiol both in solution and when immobilized to CA surfaces. Further tests are needed to confirm the necessary activities and stabilities of the heterologously produced enzyme preparates for large-scale applications. The attachment of the enzymes to the carrier surfaces only through the surface lysines was not proven so far, since activated CMC surfaces can also react with carboxyl or sulfhydryl groups, but is very likely due to the preferred amide bonds with high stability (14, 15).

In conclusion, during this study two single and 12 double mutants of HRP C1A with solvent exposed lysine residues replaced with other residues naturally occurring at the same position in known HRP isoenzymes were created, codon optimized and expressed in *P. pastoris* to study the possibility of efficient heterologous production of surface-modified HRPs. None of these mutant combinations have been previously produced in *P. pastoris*. Exchange of lysine residue 174 located closest to the heme access channel has not been described previously. In this study, we also combine the removal of this active lysine to known stabilizing mutations of residues K232 and K241 (19). The upscaling of production is tested with codon-optimized wild-type HRP C1A and the surface double-mutants HRPC1AsynK232N/K241N, HRPC1AsynK174R/K241N and HRPC1AsynK174Q/K241F. In this work, the possibilities to immobilize recombinant, glycosylated HRP C1A on cellulose acetate hollow fibre membranes through a monolayer or multilayer of polyelectrolytes such as carboxymethyl cellulose (CMC) was evaluated. In addition to the wild-type enzyme, all 3 variants produced in larger scale and tested could be immobilized utilizing the stable bonds formed between the amino groups of the surface lysines, and the carboxy goups present on the surface of 1-(3-dimethylaminopropyl)-3-ethylcarbodiimine activated carboxymethyl cellulose. Finally, the degradation of one of the endocrine disruptors, 17α-ethynylestradiol, was investigated. All double-mutants HRPC1AsynK232N/K241N, HRPC1AsynK174R/K241N and HRPC1AsynK174Q/K241F produced in larger scale were found to be able catalyse steroid degradation.

Table 1. Maximum mABS/min reached screening one 96-well plate using standard ABTS assay. Activities are not comparable between plates, since no normalization was performed.

Variant #	AA variation	Activity mABS/min
0	wt	363
1	HRPC1AsynK232Q/K241N	236
2	HRPC1AsynK232Q/K241F	146
3	HRPC1AsynK232N	1428
4	HRPC1AsynK232N/K241N	293
5	HRPC1AsynK232N/K241F	77
6	HRPC1AsynK174R/K241N	135
7	HRPC1AsynK174R/K241F	45
8	HRPC1AsynK174R/K232Q	234
9	HRPC1AsynK174R/K232N	307
10	HRPC1AsynK174Q/K241N	157
11	HRPC1AsynK174Q/K241F	108
12	HRPC1AsynK174Q/K232Q	196
13	HRPC1AsynK174Q/K232N	133
14	HRPC1AsynT110V	56
15	HRPC1AsynK241F	134

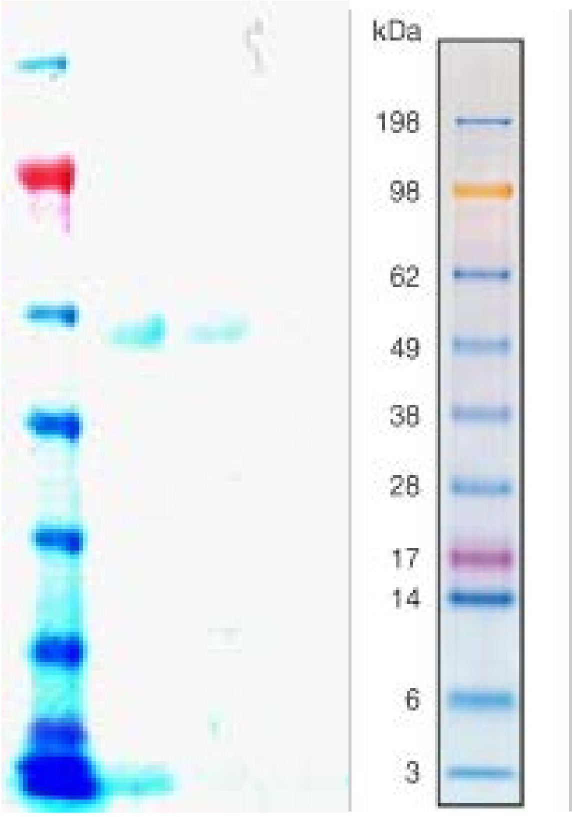


Figure 1. **10µg of the concentrated supernatant of variants HRPC1AsynK232N/K241N and HRPC1AsynK174R/K241N loaded on an Invitrogen NuPage SDS-PAGE Gel with SeeBlue Plus2 prestained protein ladder (Invitrogen).** The glycosylated HRP variants move as a smear around the 62kDa band

Author contributions

A.G., L.N., M.K. P.N. and V.R. contributed to the design of the study. L.N. performed the experiments in *P. pastoris* and drafted the manuscript. P.N. and M.K. performed the immobilization experiments.

Acknowledgements

This work was supported by the Austrian Science Fund ‘Fonds zur Förderung der wissenschaftlichen Forschung‘[W901 DK Molecular Enzymology]

References:

1. Ryan,B.J., Carolan,N. and O'Fágáin,C. (2006) Horseradish and soybean peroxidases: comparable tools for alternative niches? *Trends Biotechnol.*, **24**, 355-363.

2. Regalado,C., Garcia-Almandarez,B.E. and Duarte-Vazquez,M.A. (2004) Biotechnological applications of peroxidases. *Phytochem. Rev.*, **3**, 243-256.

3. Wagner,M. and Nicell,J.A. (2002) Detoxification of phenolic solutions with horseradish peroxidase and hydrogen peroxide. *Water Res.*, **36**, 4041-52.

4. Dalal,S. and Gupta,M.N. (2007) Treatment of phenolic wastewater by horseradish peroxidase immobilized by bioaffinity layering. *Chemosphere*, **67**, 741-7.

5. Bayramoglu,G., Altintas,B. and Yakup Arica,M. (2012) Cross-linking of horseradish peroxidase adsorbed on polycationic films: utilization for direct dye degradation. *Bioprocess Biosyst. Eng.*, 10.1007/s00449-012-0724-2.

6. Karam,J. and Nicell,J.A. (1997) Review Potential Applications of Enzymes in Waste Treatment. *J. Chem. Tech. Biotechnol.*, **69**, 141-153.

7. Auriol,M., Filali-Meknassi,Y., Adams,C.D., Tyagi,R.D., Noguerol,T.-N. and Piña,B. (2008) Removal of estrogenic activity of natural and synthetic hormones from a municipal wastewater: efficiency of horseradish peroxidase and laccase from *Trametes versicolor*. *Chemosphere*, **70**, 445-52.

8. Brady,D. and Jordaan,J. (2009) Advances in enzyme immobilisation. *Biotechnol. Lett.*, **31**, 1639-50.

9. Ryan,B.J. and O'Fágáin,C. (2007) Arginine-to-lysine substitutions influence recombinant horseradish peroxidase stability and immobilisation effectiveness. *BMC Biotechnol.*, **7**, 86.

10. Cretich,M., Damin,F., Pirri,G. and Chiari,M. (2006) Protein and peptide arrays: recent trends and new directions. *Biomol. Eng.*, **23**, 77-88.

11. Rojas-Melgarejo,F., Marin-Iniesta,F., Neptuno-Rodriquez-Lopez,J., Garcia-Canovas,F. and Garcia-Ruiz,P.A. (2006) Cinnamic carbohydrate esters show great versatility as supports for the immobilization of different enzymes. *Enzyme Microb. Technol.*, **38**, 748-755.

12. Morawski,B., Lin,Z., Cirino,P., Joo,H., Bandara,G. and Arnold,F.H. (2000) Functional expression of horseradish peroxidase in *Saccharomyces cerevisiae* and *Pichia pastoris*. *Protein Eng.*, **13**, 377-84.

13. Tischer,W. and Wedekind,F. (1999) Immobilized Enzymes : Methods and Applications. *Top. Curr. Chem.*, **200**, 96-123.

14. Dyal,A., Loos,K., Noto,M., Chang,S.W., Spagnoli,C., Shafi,K.V.P.M., Ulman,A., Cowman,M. and Gross,R. A (2003) Activity of *Candida rugosa* lipase immobilized on gamma-Fe2O3 magnetic nanoparticles. *Journal of the American Chemical Society*, **125**, 1684-5.

15. Wang,W., Xu,Y., Wang,D.I.C. and Li,Z. (2009) Recyclable nanobiocatalyst for enantioselective sulfoxidation: facile fabrication and high performance of chloroperoxidase-coated magnetic nanoparticles with iron oxide core and polymer shell. *Journal of the American Chemical Society*, **131**, 12892-3.

16. Krenková,J. and Foret,F. (2004) Immobilized microfluidic enzymatic reactors. *Electrophoresis*, **25**, 3550-63.

17. Abian,O., Grazu,V., Hermoso,J., Gonzalez,R., Garcia,J., Fernandez-Lafuente,R. and Guisan,J. (2004) Stabilization of penicillin G acylase from *Escherichia coli*: site-directed mutagenesis of the protein surface to increase multipoint covalent attachment. *Appl. Environ. Microbiol.*, **70**, 1249-1251.

18. O'Brien,A.M., O'Fagain,C., Nielsen,P.F. and Welinder,K.G. (2001) Location of Crosslinks in Chemically Stabilized Horseradish Peroxidase : Implications for Design of Crosslinks. *Biotechnol. Bioeng.*, **76**, 277-284.

19. Ryan,B.J. and O'Fágáin,C. (2008) Effects of mutations in the helix G region of horseradish peroxidase. *Biochimie*, **90**, 1414-1421.

20. Ryan,B.J. and O'Fagain,C. (2007) Effects of single mutations on the stability of horseradish peroxidase to hydrogen peroxide. *Biochimie*, **89**, 1029-1032.

21. Ruth,C., Zuellig,T., Mellitzer,A., Weis,R., Looser,V., Kovar,K. and Glieder,A. (2010) Variable production windows for porcine trypsinogen employing synthetic inducible promoter variants in *Pichia pastoris*. *Syst. Synth. Biol.*, **4**, 181-191.

22. Villalobos,A., Ness,J.E., Gustafsson,C., Minshull,J. and Govindarajan,S. (2006) Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinf.*, **7**, 285.

23. Abad,S., Kitz,K., Hörmann,A., Schreiner,U., Hartner,F.S. and Glieder,A. (2010) Real-time PCR-based determination of gene copy numbers in *Pichia pastoris*. *Biotechnol. J.*, **5**, 413-20.

24. Lin-Cereghino,J., Wong,W.W., Xiong,S., Giang,W., Luong,L.T., Vu,J., Johnson,S.D. and Lin-Cereghino,G.P. (2005) Condensed protocol for competent cell preparation and transformation of the methylotrophic yeast *Pichia pastoris*. *Biotechniques*, **38**, 44-48.

25. Weis,R., Luiten,R., Skranc,W., Schwab,H., Wubbolts,M. and Glieder,A. (2004) Reliable high-throughput screening with *Pichia pastoris* by limiting yeast cell death phenomena. *FEMS Yeast Res.*, **5**, 179-89.

Appendix

Sequences of *Armoracia rusticana* wild-type isoenzyme C1A, codon-optimized isoenzyme C1A and the 14 surface lysine variants made thereof, all produced in *Pichia pastoris.* All sequences contain the Kozak-sequence and alpha-factor signal sequence used for efficient secretory production

>HRPC1Asyn
AACGATGAGATTCCCATCTATTTTCACCGCTGTCTTGTTCGCTGCCTCCTCTGCATTGGCTGCCCCTGTTAACACTACCACTGA
AGACGAGACTGCTCAAATTCCAGCTGAAGCAGTTATCGGTTACTCTGACCTTGAGGGTGATTTCGACGTCGCTGTTTTGCCTTT
CTCTAACTCCACTAACAACGGTTTGTTGTTCATTAACACCACTATCGCTTCCATTGCTGCTAAGGAAGAGGGTGTCTCTCTCGA
GAAGAGAGAGGCCGAAGCTCAACTTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCA
TTGTCAATGAATTGAGATCAGATCCACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTG
ATGCTTCCATCTTGCTGGACAACACTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTT
CCAGTCATTGACAGAATGAAGGCTGCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGC
TCAGCAATCTGTTACCTTAGCTGGTGGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCT
TGCAAATGCTAACTTGCCTGCTCCATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATC
CGACTTGGTTGCCTTATCTGGAGGTCACACCTTTGGTAAGAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAA
CACCGGTTTGCCAGATCCTACTCTGAACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTG
CTCTGGTTGACTTCGATTTGCGTACTCCTACCATCTTCGACAACAAGTACTATGTCAACTTGGAGGAACAGAAGGGTCTTATC
CAATCTGACCAGGAGTTGTTCTCCTCTCCTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACC
TTCTTTAACGCTTTCGTCGAGGCAATGGACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAA
CTGCCGTGTTGTCAACTCTAACTCATAAT

>HRPC1AsynK232Q_K241N
aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTAAGAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCT
GAACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTAC
TCCTACCATCTTCGACAACcaaTACTATGTCAACTTGGAGGAACAGaacGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCTC
CTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATGG
ACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat

>HRPC1AsynK232Q_K241F
aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTAAGAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCT

```
GAACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTAC
TCCTACCATCTTCGACAACcaaTACTATGTCAACTTGGAGGAACAGttcGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCTC
CTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATGG
ACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat
```

>HRPC1AsynK232N

```
aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTAAGAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCT
GAACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTAC
TCCTACCATCTTCGACAACaacTACTATGTCAACTTGGAGGAACAGAAGGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCT
CCTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATG
GACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat
```

>HRPC1AsynK232N_K241N

```
aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTAAGAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCT
GAACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTAC
TCCTACCATCTTCGACAACaacTACTATGTCAACTTGGAGGAACAGaacGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCTC
CTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATGG
ACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat
```

>HRPC1AsynK232N_K241F

```
aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTAAGAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCT
GAACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTAC
TCCTACCATCTTCGACAACaacTACTATGTCAACTTGGAGGAACAGttcGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCTC
CTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATGG
ACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat
```

>HRPC1AsynK174R_K241N

aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTcgtAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCTG
AACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTACT
CCTACCATCTTCGACAACAAGTACTATGTCAACTTGGAGGAACAGaacGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCT
CCTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATG
GACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat

>HRPC1AsynK174R_K241F

aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTcgtAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCTG
AACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTACT
CCTACCATCTTCGACAACAAGTACTATGTCAACTTGGAGGAACAGttcGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCTC
CTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATGG
ACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat

>HRPC1AsynK174R_K232Q

aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTcgtAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCTG
AACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTACT
CCTACCATCTTCGACAACcaaTACTATGTCAACTTGGAGGAACAGAAGGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCT
CCTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATG
GACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat

>HRPC1AsynK174R_K232N

aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC

ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTcgtAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCTG
AACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTACT
CCTACCATCTTCGACAACaatTACTATGTCAACTTGGAGGAACAGAAGGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCTC
CTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATGG
ACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat

>HRPC1AsynK174Q_K241N
aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTcagAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCTG
AACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTACT
CCTACCATCTTCGACAACAAGTACTATGTCAACTTGGAGGAACAGaacGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCT
CCTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATG
GACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat

>HRPC1AsynK174Q_K241F
aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTcagAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCTG
AACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTACT
CCTACCATCTTCGACAACAAGTACTATGTCAACTTGGAGGAACAGttcGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCTC
CTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATGG
ACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat

>HRPC1AsynK174Q_K232Q
aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTcagAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCTG
AACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTACT
CCTACCATCTTCGACAACcagTACTATGTCAACTTGGAGGAACAGAAGGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCT
CCTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATG
GACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat

```
>HRPC1AsynK174Q_K232N
aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTcagAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCTG
AACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTACT
CCTACCATCTTCGACAACaatTACTATGTCAACTTGGAGGAACAGAAGGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCTC
CTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATGG
ACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat

>HRPC1Asyn_T110V
aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTgtgTTAGCTGGTG
GACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCCAT
TCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGGTC
ACACCTTTGGTAAGAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCTGA
ACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTACTC
CTACCATCTTCGACAACAAGTACTATGTCAACTTGGAGGAACAGAAGGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCT
CCTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATG
GACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat

>HRPC1Asyn_K241F
aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTAAGAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCT
GAACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTAC
TCCTACCATCTTCGACAACAAGTACTATGTCAACTTGGAGGAACAGttcGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCTCT
CCTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAATG
GACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAtaat

>HRPC1Asyn_original
Aacgatgagattcccatctattttcaccgctgtcttgttcgctgcctcctctgcattggctgcccctgttaacactaccactgaagacgagactgctcaaattccagctgaagcagttatcggttactctgaccttgag
ggtgatttcgacgtcgctgttttgcctttctctaactccactaacaacggtttgttgttcattaacaccactatcgcttccattgctgctaaggaagagggtgtctctctcgagaagagagaggccgaagctCAAC
TTACTCCAACCTTCTACGATAACTCTTGTCCTAATGTGTCCAACATCGTTAGAGACACCATTGTCAATGAATTGAGATCAGATC
CACGTATTGCTGCATCTATCTTGAGACTTCACTTTCATGACTGCTTCGTCAACGGTTGTGATGCTTCCATCTTGCTGGACAACA
CTACCTCTTTCAGAACTGAGAAGGACGCTTTCGGTAATGCCAACTCTGCTAGAGGATTTCCAGTCATTGACAGAATGAAGGCT
GCCGTTGAATCTGCATGTCCTAGAACTGTGTCATGTGCTGACCTTCTGACTATTGCCGCTCAGCAATCTGTTACCTTAGCTGGT
```

GGACCATCCTGGAGAGTTCCATTGGGTCGTAGAGACTCCCTTCAAGCCTTTCTGGACCTTGCAAATGCTAACTTGCCTGCTCC
ATTCTTTACCTTACCTCAATTGAAAGACTCTTTCAGAAACGTTGGTCTTAACAGATCATCCGACTTGGTTGCCTTATCTGGAGG
TCACACCTTTGGTAAGAACCAATGTAGATTCATCATGGATCGTCTGTACAACTTCTCTAACACCGGTTTGCCAGATCCTACTCT
GAACACCACTTACTTGCAAACCTTAAGAGGTTTGTGCCCACTTAACGGAAATCTGTCTGCTCTGGTTGACTTCGATTTGCGTAC
TCCTACCATCTTCGACAACAAGTACTATGTCAACTTGGAGGAACAGAAGGGTCTTATCCAATCTGACCAGGAGTTGTTCTCCT
CTCCTAACGCTACTGATACCATTCCATTGGTGAGATCCTTCGCAAACTCCACTCAAACCTTCTTTAACGCTTTCGTCGAGGCAA
TGGACAGAATGGGTAACATTACTCCTTTGACCGGTACTCAAGGACAGATTAGATTGAACTGCCGTGTTGTCAACTCTAACTCAt
aat

CHAPTER 4

Peroxidase assays

ACIB Protocol

Peroxidase assays

Author: **Laura Näätsaari**

Purpose and Field of Application

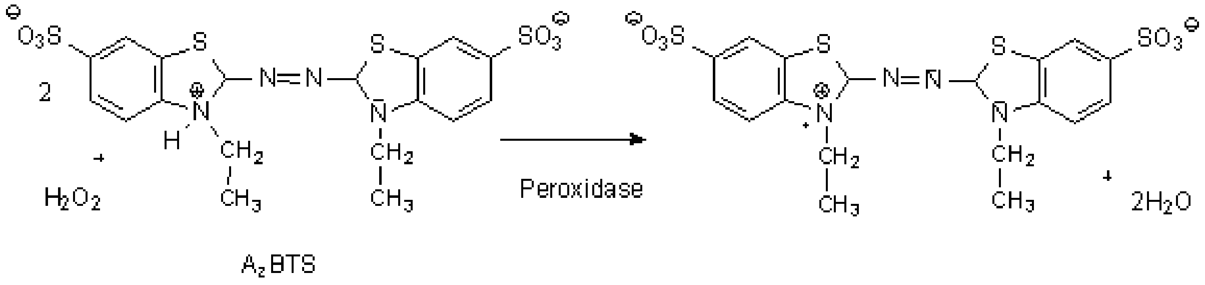
These assays can be used to determine the activity of peroxidases in culture supernatants and other aqueous solutions. In addition to different substrates, the set of assays described here comprises a broad range of operating pH values to be able to reliably estimate the activities of peroxidases with variable optimal pH values. The assays can be used with a concentration range dependant on the properties of the spectrophotometer used. Thus the concentration range has to be determined experimentally for each enzyme assayed.

Principle

ABTS assay

ABTS is a sensitive colorimetric substrate forming a relatively stable radical cation ($ABTS^{\cdot+}$) with green-blue colour, which can be measured at 405nm. H_2O_2 binds to heme in Fe^{III} state. H_2O_2 is cleaved, H_2O released followed by two-electron oxidation of heme to form compound I: oxoferryl species (FeIV=O) + ABTS radical cation ($ABTS.^{+}$). Successive one-electron reductions return the enzyme to its resting state via a second intermediate compound II: heme is retained in oxoferryl state (FeIV=O). The stability of this radical cation is dependent on the combination of several things including [ABTS], [H_2O_2], pH and antioxidants. It is known that it can undergo a disproportionation reaction giving ABTS and the azodication. The disproportionation reaction is not favoured in acidic solutions, therefore pH 4,5 is used and a decrease in the reaction rate can be observed when increasing the pH, regardless the enzyme's optimal pH. It seems that unreacted ABTS is stabilizing the radical cation.

Figure below from: Sigma-Aldrich (http://www.sigmaaldrich.com/life-science/metabolomics/enzyme-explorer/analytical-enzymes/peroxidase-enzymes.html)



TMB assay

TMB (3,3′,5,5′-Tetramethylbenzidine) is a sensitive non-toxic substrate producing an end product with pale blue colour. The reaction can be observed spectrophotometrically at 650nm. If no kinetics is needed, the reaction can be stopped with 2M H_2SO_4 and the yellow colour forming can be measured at 450nm. This should increase the sensitivity of the assay. Oxidation products are very stable at acidic pH. In the reaction with HRP and H_2O_2, three oxidation products have been characterized:

- Radical cation (one-electron oxidation)
- Diimine derivative (two-electron oxidation product)
- Diimine

A charge-transfer complex of TMB and its two-electron oxidation product (in equilibrium with the cation free radical) is responsible for the absorption at 650nm. Diimine is bright yellow and absorbs at 450nm.

Figure below from: Josephy *et al.* (1982) Cooxidation of the Clinical Reagent 3,5,3'5'-Tetramethylbenzidine by Prostaglandin Synthase.

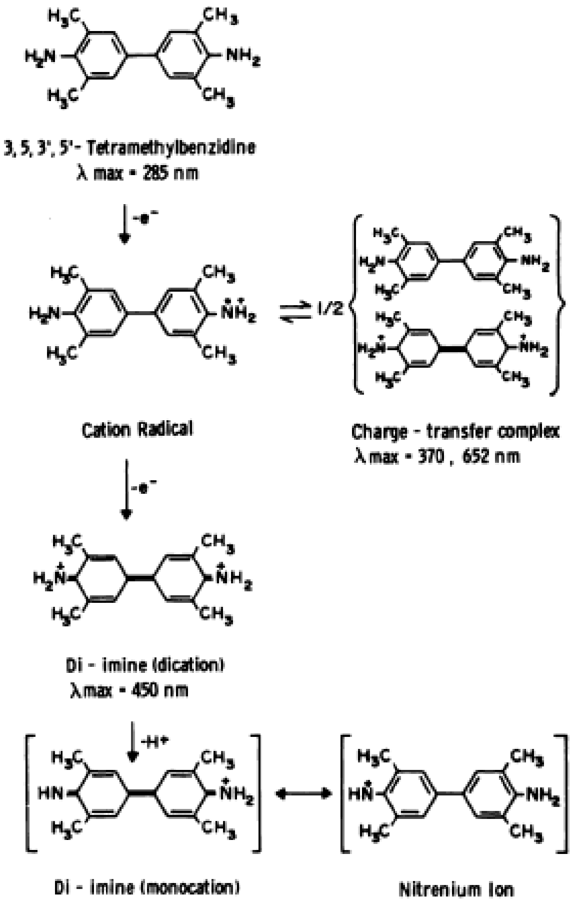
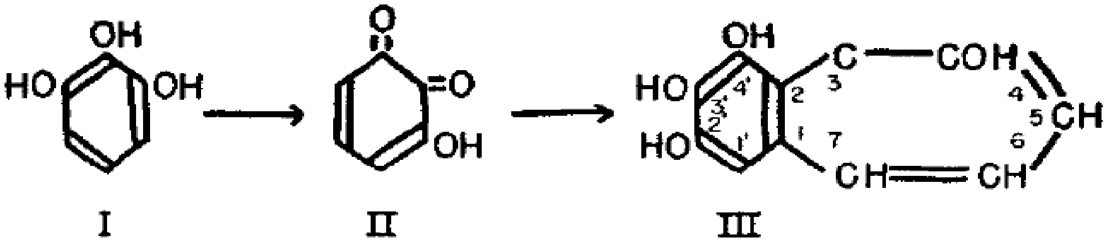


Chart 1. Oxidation of TMB: chemical structures.

Pyrogallol assay

Pyrogallol as a HRP substrate is not as sensitive as either ABTS or TMB, but has a higher assay pH optimum. The substrate is very toxic especially to aquatic organisms, and is therefore not recommended to be used as a primary HRP substrate. Formation of the colored end product upon oxidation can be measured at 420nm. Peroxidase oxidizes pyrogallol to orthoquinone (II). This orthoquinone (II), after further dehydrogenation and loss of a carbon is coupled with pyrogallol (I) to form purpurogallin (III). The usage of a slightly acidic pH for the assay is recommended, since at alkaline pH pyrogallol auto-oxidation has been suggested to be considerable. The pH range of the assay is rather narrow, since already at pH 5,3 purpurogallin formation is negligible.

The figure below from: Tauber (1953) Oxidation of pyrogallol to purpurogallin by crystalline catalase.



Guaiacol assay

Guaiacol is an aromatic, oily, yellow substrate used in the HRP assays. The chemical is toxic, has an extremely strong and unpleasant smell, and the assay is not as sensitive as the assays with ABTS or TMB. Thus, it should not be used as a primary assay for HRP activity. However, the assay can be used also in neutral or slightly alkaline buffers.

Due to the typical catechol impurities present in the commercial guaiacol preparations, the product of a reaction of guaiacol with H_2O_2 and HRP has been analysed to be a mixture of 3,3 '-dimethoxy-4,4'-biphenylquinone, 3,3′-dimethoxy-4,4′-dihydroxybiphenyl and 3-methoxy-2′,3′,4-trihydroxybiphenyl. The rate of the guaiacol dehydrogenation product formation can be measured spectrophotometrically at 470nm.

Key Words, Definitions & Abbreviations

Key words: HRP, horseradish peroxidase, enzyme activity, ABTS assay, TMB assay, pyrogallol assay, guaiacol assay.

ABTS: 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulphonic acid)
TMB: 3,3′,5,5′-Tetramethylbenzidine
Pyrogallol: 1,2,3-Trihydroxybenzene
Guaiacol: 2-Methoxyphenol (catechol monomethyl ether, pyrocatechol monomethyl ether)

Methodology

Reagents

List of the chemicals needed in addition to standard buffers.

Name	Formula	MW	Purity	Supplier	Order No.	Comments
TMB	$C_{16}H_{20}N_2$	240.34	≥99%	Sigma	860336, T2885	2–8 °C
Hydrogen peroxide	H_20_2	34.0147	30 %	Sigma	H1009	+4°C
Pyrogallol	$C_6H_3(OH)_3$	126.11, 1.129 g/ml at 25 °C(lit.)	≥99%	Sigma	16040	RT
Guaiacol	$(CH_3O)C_6H_4OH$	124.14 1,129g/ml at 25 °C(lit.)	n.d.	Sigma	G5502	RT (vapour density 4.27 vs air)
ABTS	$C_{18}H_{24}N_6O_6S_4$	548.68	≥98%	Sigma	A1888	+4°C

Solutions

Make sure all glassware or plastics used to prepare the solutions are clean and free of any cell material or other contaminants with oxidase/peroxidase activity. If possible, use autoclaved buffers and new/sterile plastics.

ABTS assay

1mM ABTS, 0,0026% (1,11mM) hydrogen peroxide in 50mM NaOAc buffer pH4,5:

50mM NaOAc buffer pH4.5 (can be autoclaved for longer storage at RT)

ABTS assay solution: 22mg ABTS to 20ml NaOAc buffer (very pale green colour)

20x ABTS stock solution can be prepared by dissolving 220mg ABTS in 20ml 50mM NaOAc buffer. Can be store in the dark at +4°C as long as no colour change is observed.
30% (v/v) hydrogen peroxide solution (1,75µl/20ml assay solution, use ready assay solution immediately)

TMB assay

0,416mM TMB, 0,006% (2,54mM) hydrogen peroxide in 20mM citrate buffer pH 5,5:

20mM citrate buffer pH 5,5 (autoclave for a longer storage at RT)

10x TMB stock solution: Dissolve TMB in DMSO to a concentration 1mg/ml. Stock solution can be stored in -20°C in aliquots, and should, according to the manufacturer, remain active for about 2 years. Discard any aliquots with a color change (oxidation). Dilute to 1x in 20mM citrate buffer pH5,5.

If you want to avoid the use of DMSO, order 3,3′,5,5′-Tetramethylbenzidine dihydrochloride hydrate instead of TMB (can be dissolved to aqueous solutions). Both forms should give identical results.

30% (v/v) hydrogen peroxide solution (4µl/20ml assay solution, use ready assay solution immediately)

Pyrogallol assay

42,8mM pyrogallol, 0,027% (11,4mM) hydrogen peroxide in 10mM phosphate buffer pH6:

10mM phosphate buffer pH6 (autoclave for a longer storage at RT)

Pyrogallol assay solution: 108mg pyrogallol to 20ml 10mM phosphate buffer pH6. Use immediately.

30% (v/v) hydrogen peroxide solution (29,9µl/20ml assay solution, use ready assay solution immediately).

Guaiacol assay

5mM guaiacol, 0,0009% (0,38mM) hydrogen peroxide in 50mM sodium phosphate buffer pH7:

50mM sodium phosphate buffer pH7 (autoclave for a longer storage at RT)

Guaiacol assay solution: add 11µl of guaiacol to 20ml phosphate buffer

30% (v/v) hydrogen peroxide solution (0,6µl/20ml assay solution, use ready assay solution immediately)

Materials

Name	Supplier	Order No.	Comments
50 ml polypropylene tubes	Greiner bio-one	227261	Any supplier ok
PS 96-well microplates, flat bottom, medium binding	Greiner bio-one	655 101	Use only low or medium binding plates
Standard lab glassware	Any		
Standard lab plasticware	Any		

Apparatus

Name	Supplier
Spectramax Plus 384	Molecular Devices, Ismaning/München, Germany
NanoDrop 2000c Spectrophotometer	peqlab Biotechnologie GmbH, Polling, Austria
Centrifuge 5810R	Eppendorf AG, Hamburg, Germany
Centrifuge 5415R	Eppendorf AG, Hamburg, Germany

Procedure

For secreted proteins, pellet the cells with 4000rpm for 5-10min. Transfer 15µl of the supernatant to a clean flat-bottom 96-well plate (longer storage seems to increase the amount of enzyme bound to the plastic surface, therefore longer storage on medium binding plates can't be recommended). Be careful not to disturb the pellet, the cells can exhibit very strong oxidase activity without any heterologous peroxidase produced. Dilute if necessary. All assay solutions should be prepared fresh right before measuring, thus prepare the plates with the enzyme solution first.

If the spectrophotometer shows "No fit" it can either mean no measurable activity (no colour change) or too strong sample/too much enzyme (substrate used up too fast, dilute 1/10, 1/100). If a well shows clear colour change, the corresponding clone has peroxidase activity. The optimal measuring range is usually, depending on the spectrophotometer, around 30-300 mAbs/min.

If the aim of the measurement is to calculate exact enzyme activity, the assays can be scaled up to be measured in a cuvette with known (1cm) path length, or the path length estimated using the PathCheck –function provided by the SoftMax Pro software (for 155µl total volume and Greiner Bio-one 655 101 plates the path length is 0.42cm)

ABTS assay

Dilute ABTS stock solution or prepare a fresh solution, add hydrogen peroxide. Prepare a spectrophotometer to observe the increase of absorption at 405nm, 25°C, for three to five minutes (at least 10 reads total with linear increase of absorption, every 10 to 20 seconds, mix between reads, use water or supernatant from wild-type cells to blank). Pour ready ABTS assay solution to a tray, add 140µl of the solution to 15µl of undiluted or diluted enzyme solution with an electronic multichannel pipette. Start measurement immediately.

The assay plate can be covered with a plastic foil, stored in the dark and observed after a few hours (end-point measurement at 405nm). If any color development can be observed while blanks stay colorless, peroxidase activity exists in the sample. As long as evaporation is prevented, the plate can be stored also for a longer time.

TMB assay

Melt TMB aliquots and dilute to 1x with citrate buffer, add 4µl fresh hydrogen peroxide per 20ml assay solution. Prepare a spectrophotometer to observe the increase of absorption at 650nm, 25°C, for three to five minutes (at least 10 reads total with linear increase of absorption, every 10 to 20 seconds, mix between reads, use water or supernatant from wild-type cells to blank). Pour ready TMB assay solution to a tray; add 140µl of the solution to 15µl of undiluted or diluted enzyme solution with an electronic multichannel pipette. Start measurement immediately.

The plate can be stored for a few hours if evaporation is prevented, however background increases and also final oxidation product diimine is formed.

Pyrogallol assay

Prepare a spectrophotometer to observe the increase of absorption at 420nm for three to five minutes (at least 10 reads total with linear increase of absorption, every 10 to 20 seconds, mix between reads, use water or supernatant from wild-type cells to blank). Dissolve 108mg pyrogallol to 20ml of phosphate buffer and add 29.9µl hydrogen peroxide. Pour ready pyrogallol assay solution to a tray; add 140µl of the solution to 15µl of undiluted or diluted enzyme solution with an electronic multichannel pipette. Start measurement immediately.
Pyrogallol waste should be collected separately (see Chapter 5 for safety precautions).

Guaiacol assay

Prepare the guaiacol assay solution well in time, so that 11µl guaiacol is properly mixed in 20ml of phosphate buffer. Work in the fume hood or take care of other proper ventilation. Prepare a spectrophotometer to observe the increase of absorption at 470nm for three to five minutes (at least 10 reads total with linear increase of absorption, every 10 to 20 seconds, mix between reads, use water or supernatant from wild-type cells to blank). Add 0.6µl of hydrogen peroxide (dilution 1/10 or 1/100 necessary for exact concentration) to the assay solution, mix properly and pour to a tray. Add 140µl of the solution to 15µl of undiluted or diluted enzyme solution with an electronic multichannel pipette. Start the measurement immediately. The plate can be sealed after addition of assay buffer to decrease the evaporation of guaiacol. Take care of proper ventilation.

Guaiacol waste should be collected separately (see Chapter 5 for safety precautions).

Calculations

Calculation of the molarity of 30% hydrogen peroxide:

Density (H_20_2): 1,45g/cm^3

Mw (H_20_2): 34.0147g/mol

0.3l of 100% hydrogen peroxide in one liter of 30% hydrogen peroxide has a mass of (0.3l)(1.45g/0.001l) = 432g

432g divided by the molar mass 34.01g/mol = 12.7M

Calculation of the enzyme activity

ε of oxidized ABTS is 34700 M^{-1} cm^{-1}

ε of oxidized guaiacol is 26000 M^{-1} cm^{-1}

ε of purpurogallin 26700 M^{-1}cm^{-1}

ε of oxidized TMB 39000 M^{-1}cm^{-1}

$$U = \frac{\Delta A \cdot V_{total} \cdot F_D \cdot 1000 \cdot d}{\varepsilon \cdot V_{sample}}$$

U = Units/ml of enzyme

ΔA = Absorbance change Abs/min

F_D = Dilution factor

d = Light path length (cm)

ε = Molar extinction coefficient of oxidized substrate at the wavelength used (M^{-1} cm^{-1})

V_{total} = Total assay volume in ml

V_{sample} = Amount of enzyme solution in ml

Unit definitions

Unit definitions vary by publication; some are connected to the amount of the substance, some to the weight of the substance oxidized in a certain amount of time (typically, but not necessarily 1 minute).

ABTS: The amount of enzyme oxidizing 1µmol of substrate in one minute at pH 4.5, 25°C.

TMB: The amount of enzyme oxidizing 1µmol of substrate in one minute at pH 5.5, 25°C.

Pyrogallol: The amount of enzyme which catalyses the production of one milligram of Purpurogallin in 20 seconds at 20°C and pH 6.0. This unit is, according to Sigma-Aldrich, equivalent to ~18 µM units per minute at 25 °C.

Guaiacol: The amount of enzyme oxidizing 1µmol of substrate in one minute at pH 7.0, 25°C.

Safety Precautions

For all the assays, if available please follow instructions described in "ACIB-Mitarbeiterleitfaden Gefahrstoff- und Laborordnung"

ABTS assay

ABTS has irritant properties, but is otherwise considered non-toxic and safe to use. No need to collect separately.
Hazard statements: H315-H319-H335
Precautionary statements: P261-P305 + P351 + P338
Risk statements: 36/37/38
Safety statements: 26-36

TMB assay (Xi)

TMB has irritant properties, but as ABTS, is otherwise considered safe to use.

Hazard statements: H315-H319-H335
Precautionary statements: P261-P305 + P351 + P338
Risk statements: 36/37/38
Safety statements: 26-36

Personal Protective Equipment: dust mask type N95 (US), Eyeshields, Faceshields, Gloves

Pyrogallol assay (Xn)

Waste and contaminated plastics have to be collected separately marked with "Organic solid, 1,2,3-Trihydroxybenzene". R68: possible risk of irreversible effects, handle carefully avoiding any dust formation.

Hazard statements: H302-H312-H332-H341-H412
Precautionary statements: P273-P280
Risk statements: 20/21/22-52/53-68
Safety statements: 36/37-61

Personal Protective Equipment: dust mask type N95 (US), Eyeshields, Faceshields, full-face particle respirator type N100 (US), Gloves, respirator cartridge type N100 (US), type P1 (EN143) respirator filter, type P3 (EN 143) respirator cartridges

Guaiacol assay (Xn)

Waste and contaminated plastics have to be collected separately marked with “Organic liquid, 2-methoxyphenol”. May affect genetic material and cause damage in the central nervous system. Take care of sufficient ventilation and avoid breathing vapours without suitable respiratory equipment.

Hazard statements: H302-H315-H319
Precautionary statements: P305 + P351 + P338
Risk statements: 22-36/38

Personal Protective Equipment: dust mask type N95 (US), Eyeshields, Faceshields, Gloves

Documentation

Practical examples of the use of the abovementioned assays can be found in the lab books of Laura Näätsaari (book # 2-7) and Florian Krainer (book # 1-3).

References

ABTS assay

The assay is modified from:

Morawski, B., Lin, Z., Cirino, P., Joo, H., Bandara, G., and Arnold, F. H. (2000) Functional expression of horseradish peroxidase in *Saccharomyces cerevisiae* and *Pichia pastoris*. *Protein engineering 13*, 377-84.

TMB assay

The assay is modified from:

O'Brien AM, O'Fágáin C, Nielsen PF, Welinder KG (2001). Location of crosslinks in chemically stabilized horseradish peroxidase: implications for design of crosslinks. *Biotechnol Bioeng.*, **76**, 277-84.

Other information:

Liem, H.H., *et al*., (1979) *Anal. Biochem*., **98**, 388-393.

Holland, V.R., *et al* (1974). *Tetrahedron* **30**, 3299.

Suzuki, K., *et al.* (1983*) Anal. Biochem*. 132, 345.

Marquez, L.A., and Dunford, H.B. (1997) *Biochemistry*, 36, 9349.

Pyrogallol assay

The assay is modified from:

Chance, B. and Maehly, A.C. (1955) *Methods in Enzymology*, II, 773-775 and http://www.sigmaaldrich.com/technical-documents/protocols/biology/enzymatic-assay-of-peroxidase.html

Guaiacol assay

Assay is performed as described by:

Morawski, B., Lin, Z., Cirino, P., Joo, H., Bandara, G., and Arnold, F. H. (2000) Functional expression of horseradish peroxidase in *Saccharomyces cerevisiae* and *Pichia pastoris*. *Protein engineering 13*, 377-84.

Other information:

Venisse, J.S., *et al.* (2001). Evidence for the involvement of an oxidative stress in the initiation of infection of pear by *Erwinia amylovora*. *Plant Physiology* 125: 2164-2172.

Desser, R.K., *et al.* (1972) *Arch. Biochem. Biophys*., 148, 452-465

CONCLUSIONS AND OUTLOOK

An expression platform for heterologous protein production in *P. pastoris* (*Komagataella pastoris*) was designed and constructed during the first part of this study. The platform is based on wild-type strain CBS7435 (ATCC 76273, NRRL -Y11430) and thus planned to give freedom to operate. Using the wild-type strain and specific knock-out methods, strains with deletion or disruption of the *AOX1*, *ARG4*, *HIS4*, *GUT1*, *DAS1* and *DAS2* loci were generated. In addition, the *P. pastoris KU70* gene coding for one of the subunits of the Ku70p/Ku80p heterodimer responsible for binding free double-stranded DNA ends in the non-homologous end-joining machinery of *P. pastoris* was identified and disrupted. The resulting *ku70* knock-out strain showed significantly increased homologous integration compared to random integration. This property supports the search and evaluation of specific gene knock outs, thereby providing for example quick access to new selection markers and facilitated pathway construction. The strain is not only facilitating the functional characterization of the genome and the metabolic routes of *P. pastoris*, but it will also accelerate further developments of this host platform and corresponding metabolic models. In addition to the platform strains, a whole series of expression vectors were generated providing a broad spectrum of precise tools for both intracellular and secreted production of heterologous proteins utilizing various selection markers and integration strategies for targeted or random integration of single and multiple genes.

During the second part of this study, a total of 28 horseradish peroxidase isoenzymes were discovered in the transcriptome and genome sequences of *A. rusticana*. Applying the new *P. pastoris* expression platform, twenty-two of the verified isoenzymes were successfully produced in an active form in *P. pastoris*, enabling their development into commercial pure isoenzymes. An analysis of the structures and phylogenetic relationships of the isoenzymes revealed that the previously known isoenzymes are more closely related to each other than most of the new isoenzymes discovered during this study. For the numerous applications for HRP, phylogenetic distance offers also the possibility of higher divergence in substrate specificity and optimal reaction conditions.

In the third part of this study, the surface composition of HRP isoenzyme C1A was modified for oriented enzyme immobilization on cellulose acetate hollow fibre membranes through monolayer or multilayer of polyelectrolytes. Active surface modified enzymes with only one lysine accessible to chemicals left to the surface were successfully produced in *P. pastoris*. First tests to immobilize recombinant HRP C1A on silicon dioxide wafers coated with carboxymethyl cellulose and 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC) linker molecule were very promising. All surface modified isoenzyme C1A produced in *P. pastoris* showed high binding capacity using covalent linkage and were active degrading endocrine disruptors present in municipal wastewaters.

The sequence data of the HRP isoenzymes acquired in this study, together with the new versatile *P. pastoris* expression platform, provide an advantageous starting point for producing a large group of single isoenzymes for biotechnological and medical applications. However, further work will be necessary to provide a platform of isoenzyme preparations for fast and effective screening towards commercial applications:

1. Large-scale production of all single isoenzymes.
2. Optimization of the purification protocols for the isoenzymes. The seemingly irregular, batch dependent glycosylation degree has made the trials to build a reproducible purification protocol very time consuming. Inspired by the results of the immobilization studies, purification trials using a carbodiimide linker molecule could be worth consideration.
3. Although the strong glycosylation and especially mannosylation of proteins produced in *P. pastoris* has been of advantage in applications like ADEPT, and could increase the stability of the protein and improve the immobilization properties, platform strains to reduce or standardize the glycosylation of the proteins might facilitate the production of purified single enzyme preparations of invariable consistency and quality.
4. The large group of isoenzymes described in chapter 2 can be screened to find the isoenzyme with highest specific activity towards IAA and also the highest stability in the physiological conditions of a tumor. According to previous studies, the production of antibody-enzyme fusion proteins is possible in *P. pastoris* (1, 2). However, no published results of studies aiming to produce HRP-antibody fusions could be found. The possibilities to produce active HRP with and functional antibody as a fusion protein could be first tested using an antibody with easily accessible antigen. The started cooperation with Peter Punt could be continued now with first trials to produce HRP-VHH fusions.
5. The first results of the oriented enzyme immobilization with HRP surface variants are very promising, since the enzymes from concentrated supernatants could be bound to the carrier membranes without any purification steps and showed high activity. Larger scale production of the surface variants would be required to test the application in industrial scale waste-water purification processes.

References:

1. Kogelberg,H., Tolner,B., Sharma,S.K., Lowdell,M.W., Qureshi,U., Robson,M., Hillyer,T., Pedley,R.B., Vervecken,W., Contreras,R., *et al.* (2007) Clearance mechanism of a mannosylated antibody-enzyme fusion protein used in experimental cancer therapy. *Glycobiology*, **17**, 36-45.

2. Medzihradszky,K.F., Spencer,D.I.R., Sharma,S.K., Bhatia,J., Pedley,R.B., Read,D. a, Begent,R.H.J. and Chester,K. A (2004) Glycoforms obtained by expression in *Pichia pastoris* improve cancer targeting potential of a recombinant antibody-enzyme fusion protein. **14**, 27-37.

APPENDIX

The appendix includes accession numbers and/or strain collection numbers of the *P. pastoris* strains and *E. coli*/*P. pastoris* shuttle vectors used and constructed during this study. In addition, since the accession numbers of the HRP isoenzymes submitted to EMBL were not available at the time point of thesis upload, the corresponding sequences of each isoenzyme, including allelic variants, can be found in the last part of the appendix.

Table 1. Platform strains. ARG4 knock-out strains have not been characterized, therefore more than one strain can be found in the collection.

CC #	Designation	Genotype	Mutation description
3444	CBS7435 WT	Wild-type	-
3445	3444/mut^s	*aox1*	*AOX1* knock-out
3519/3521	3445/arg-	**Unknown**	*Unknown*
3520/3522	3444/his-	*his4*	*HIS4* knock-out
3499	3444/ku-	*ku70*	*KU70* partial knock-out
3518	3499/his-	*ku70 his4*	*KU70* partial knock-out and *HIS4* knock-out
3583/3584	3445/his-	*aox1 his4*	*AOX1* knock-out and *HIS4* knock-out
6488	3444/arg-	*arg4*	*ARG4* knock-out
6492	3444/arg-	*arg4*	*ARG4* knock-out
6493	3444/arg-	*arg4*	*ARG4* knock-out
6494	3444/arg-	*arg4*	*ARG4* knock-out
6489	3499/arg-	*ku70 arg4*	*KU70* partial knock-out and *ARG4* knock-out
6499	3499/arg-	*ku70 arg4*	*KU70* partial knock-out and *ARG4* knock-out
6490	3445/his-arg-	*aox1 his4 arg4*	*AOX1*, HIS4, and *ARG4* knock-outs
6491	3445/his-arg-	*aox1 his4 arg4*	*AOX1*, HIS4, and *ARG4* knock-outs
6495	3445/arg-	*aox1 arg4*	*AOX1* knock-out and *ARG4* knock-out
6496	3445/arg-	*aox1 arg4*	*AOX1* knock-out and *ARG4* knock-out
6497	3445/arg-	*aox1 arg4*	*AOX1* knock-out and *ARG4* knock-out
6498	3445/arg-	*aox1 arg4*	*AOX1* knock-out and *ARG4* knock-out

Table 2. Shuttle vectors (empty and with HRP/GFP stuffer). If not stated otherwise in details, the vectors are in the *E. coli* host strain DH5α.

CC #	Name	Accession #	Details
6075	pPpB1_S	JQ519685	Amp-R; EcoRI/NotI/SpeI/AscI ; SmiI to linearize
5799	pPpB1GAP	JQ519686	Amp-R; EcoRI/NotI/SpeI/AscI ; BglII to linearize
6076	pPpB1GAP_S	JQ519687	Amp-R; EcoRI/NotI/SpeI/AscI ;SmiI to linearize
6337	pPpB1_Alpha_S_HRP	JQ519688	Amp-R; EcoRI/NotI/SpeI/AscI ;SmiI to linearize
5710	pPpT4 (pILV5_AODTT_T4)	JQ519689	Amp-R; EcoRI/NotI/SpeI/AscI ; BglII to linearize
6070	pPpT4_S	JQ519690	Amp-R; EcoRI/NotI/SpeI/AscI ;SmiI to linearize
6071	pPpT4_Alpha_S	JQ519691	Amp-R; EcoRI/NotI/SpeI/AscI ;SmiI to linearize
6072	pPpT4GAP_S	JQ519692	Amp-R; EcoRI/NotI/SpeI/AscI ;SmiI to linearize
6073	pPpT4GAP_Alpha_S	JQ519693	Amp-R; EcoRI/NotI/SpeI/AscI ;SmiI to linearize
6064	pPpKan_S	JQ519694	Amp-R; EcoRI/NotI/SpeI/AscI ;SmiI to linearize
6065	pPpKan_Alpha_S	JQ519695	Amp-R; EcoRI/NotI/SpeI/AscI ;SmiI to linearize
5823	pPpARG4	JQ519696	Amp-R; EcoRI/NotI/SpeI/AscI; BglII to linearize
5824	pPpHIS4	JQ519697	Amp-R; EcoRI/NotI/SpeI/AscI ;BglII to linearize
3088	pPpGUT1	JQ519698	Amp-R;EcoRI/NotI/SpeI/AscI; long3´rec
3089	pPpGUT1	-	pPpGUT1_short3´rec
6343	pPpB1_GFP	-	Cycle-3-GFP EcoRI/NotI cloned to B1_BglII, TOP10F’
6337	pPpB1_alpha1_HRP	-	HRPC1A (CC#3090) with SSalpha in pPpB1_BglII, TOP10F’
6333	pPpT4_SmiI_GFP	-	Cycle-3-GFP EcoRI/NotI cloned to T4_SmiI, TOP10F’

Table 3. HRP vectors and expression strains. All constructs marked with "alpha" have been planned for secreted production of HRP. Thus, they contain the *S. cerevisiae* 267bp synthetic alpha factor signal sequence (including recognition and cleavage sites). α = codon usage is optimized for *Pichia pastoris* using GeneDesigner (manually improved automatic optimization). β = codon usage is optimized for *Pichia pastoris* by DNA2.0 (only automatic optimization). γ = mature protein without predicted natural N-terminal signal sequence. ε = complete coding sequence including predicted signal sequences. λ = sequence of the isoenzyme in the genome of the recipient strain has been confirmed by sequencing the amplified genomic locus.

CC #	Designation	Mutation/genotype	Remarks	Strain
3090	pPpT4_alpha_SmiI_HRPC1ASyn#0	Wild-type aa sequence	α	K12 TOP10F'
3091	pPpT4_alpha_SmiI_HRPC1ASyn#1	HPRC1AsynK232Q_K241N	α	K12 TOP10F'
3092	pPpT4_alpha_SmiI_HRPC1ASyn#2	HPRC1AsynK232Q_K241F	α	K12 TOP10F'
3093	pPpT4_alpha_SmiI_HRPC1ASyn#3	HPRC1AsynK232N	α	K12 TOP10F'
3080	pPpT4_alpha_SmiI_HRPC1ASyn#4	HPRC1AsynK232N_K241N	α	K12 TOP10F'
3081	pPpT4_alpha_SmiI_HRPC1ASyn#5	HPRC1AsynK232N_K241F	α	K12 TOP10F'
3082	pPpT4_alpha_SmiI_HRPC1ASyn#6	HPRC1AsynK174R_K241N	α	K12 TOP10F'
3083	pPpT4_alpha_SmiI_HRPC1ASyn#7	HPRC1AsynK174R_K241F	α	K12 TOP10F'
3084	pPpT4_alpha_SmiI_HRPC1ASyn#8	HPRC1AsynK174R_K232Q	α	K12 TOP10F'
3086	pPpT4_alpha_SmiI_HRPC1ASyn#9	HPRC1AsynK174R_K232N	α	K12 TOP10F'
3076	pPpT4_alpha_SmiI_HRPC1ASyn#10	HRPC1AsynK174Q/K241N	α	K12 TOP10F'
3077	pPpT4_alpha_SmiI_HRPC1ASyn#11	HRPC1AsynK174Q_K241F	α	K12 TOP10F'
3078	pPpT4_alpha_SmiI_HRPC1ASyn#12	HRPC1AsynK174Q_K232Q	α	K12 TOP10F'
3079	pPpT4_alpha_SmiI_HRPC1ASyn#13	HRPC1AsynK174Q_K232N	α	K12 TOP10F'
3087	pPpT4_alpha_SmiI_HRPC1ASyn#14	HPRC1AsynT110V	α	K12 TOP10F'
3085	pPpT4_alpha_SmiI_HRPC1ASyn#15	HPRC1AsynK241F	α	K12 TOP10F'
6245	pPpT4_alpha_SmiI_HRPC1A	Wild-type aa sequence	β γ	K12 TOP10F'
6246	pPpT4_alpha_SmiI_HRPC1B_15901	Wild-type aa sequence	β γ	K12 TOP10F'
6247	pPpT4_alpha_SmiI_C1C_25148	Wild-type aa sequence	β γ	K12 TOP10F'
6248	pPpT4_alpha_SmiI_C1D	Wild-type aa sequence	β γ	K12 TOP10F'
6249	pPpT4_alpha_SmiI_C2_04627	Wild-type aa sequence	β γ	K12 TOP10F'
6251	pPpT4_alpha_SmiI_C3	Wild-type aa sequence	β γ	K12 TOP10F'
6302	pPpT4_alpha_SmiI_A2A	Wild-type aa sequence	β γ	K12 TOP10F'
6303	pPpT4_alpha_SmiI_A2B	Wild-type aa sequence	β γ	K12 TOP10F'
6304	pPpT4_alpha_SmiI_E5	Wild-type aa sequence	β γ	K12 TOP10F'
6305	pPpT4_alpha_SmiI_01805	Wild-type aa sequence	β γ	K12 TOP10F'
6306	pPpT4_alpha_SmiI_22684.1	Wild-type aa sequence	β γ	K12 TOP10F'
6307	pPpT4_alpha_SmiI_22684.2	Wild-type aa sequence	β γ	K12 TOP10F'
6308	pPpT4_alpha_SmiI_01350	Wild-type aa sequence	β γ	K12 TOP10F'
6309	pPpT4_alpha_SmiI_02021	Wild-type aa sequence	β γ	K12 TOP10F'
6362	pPpT4_alpha_SmiI_23190.1noSS1	Wild-type aa sequence	α γ	K12 TOP10F'
6363	pPpT4_alpha_SmiI_23190.1noSS2	Wild-type aa sequence	α γ	K12 TOP10F'

6358	pPpT4_alpha_SmiI_04663.1noSS1	Wild-type aa sequence	α γ	K12 TOP10F'
6361	pPpT4_alpha_SmiI_06351noSS1	Wild-type aa sequence	α γ	K12 TOP10F'
6335	pPpT4_alpha_SmiI_03523.noSS1	Wild-type aa sequence	α γ	K12 TOP10F'
6359	pPpT4_alpha_SmiI_05508.1noSS1	Wild-type aa sequence	α γ	K12 TOP10F'
6360	pPpT4_alpha_SmiI_05508.1noSS2	Wild-type aa sequence	α γ	K12 TOP10F'
6458	pPpT4_alpha_SmiI_22489.1noSS1	Wild-type aa sequence	α γ	K12 TOP10F'
6459	pPpT4_alpha_SmiI_22489.1noSS2	Wild-type aa sequence	α γ	K12 TOP10F'
6460	pPpT4_alpha_SmiI_22489.2noSS1	Wild-type aa sequence	α γ	K12 TOP10F'
6461	pPpT4_alpha_SmiI_22489.2noSS2	Wild-type aa sequence	α γ	K12 TOP10F'
6462	pPpT4_alpha_SmiI_04791noSS1	Wild-type aa sequence	α γ	K12 TOP10F'
6463	pPpT4_alpha_SmiI_06117noSS1	Wild-type aa sequence	α γ	K12 TOP10F'
6464	pPpT4_alpha_SmiI_06117noSS2	Wild-type aa sequence	α γ	K12 TOP10F'
6340	pPpT4_alpha_SmiI_17517.1noSS1	Wild-type aa sequence	α γ	K12 TOP10F'
6341	pPpT4_alpha_SmiI_17517.1noSS2	Wild-type aa sequence	α γ	K12 TOP10F'
6342	pPpT4_alpha_SmiI_17517.2noSS1	Wild-type aa sequence	α γ	K12 TOP10F'
6344	pPpT4_alpha_SmiI_17517.2noSS2	Wild-type aa sequence	α γ	K12 TOP10F'
6334	pPpT4_alpha_SmiI_08562.1noSS1	Wild-type aa sequence	α γ	K12 TOP10F'
6336	pPpT4_alpha_SmiI_08562.1noSS2	Wild-type aa sequence	α γ	K12 TOP10F'
6345	pUC57_03523	Wild-type aa sequence	αε	K12 TOP10F'
6346	pUC57_04663	Wild-type aa sequence	αε	K12 TOP10F'
6347	pUC57_05508	Wild-type aa sequence	αε	K12 TOP10F'
6348	pUC57_06351	Wild-type aa sequence	αε	K12 TOP10F'
6349	pUC57_23190	Wild-type aa sequence	αε	K12 TOP10F'
6350	pUC57_22489.1	Wild-type aa sequence	αε	K12 TOP10F'
6351	pUC57_22489.2	Wild-type aa sequence	αε	K12 TOP10F'
6352	pUC57_04791	Wild-type aa sequence	αε	K12 TOP10F'
6353	pUC57_06117	Wild-type aa sequence	αε	K12 TOP10F'
6354	pUC57_17517.1	Wild-type aa sequence	αε	K12 TOP10F'
6355	pUC57_17517.2	Wild-type aa sequence	αε	K12 TOP10F'
6356	pUC57_08562.4	Wild-type aa sequence	αε	K12 TOP10F'
6357	pUC57_08562.1	Wild-type aa sequence	αε	K12 TOP10F'
6374	pPpT4_alpha_SmiI_HRPC1A	Wild-type aa sequence	β γ	3444/mut^{S}
6375	pPpT4_alpha_SmiI_HRPC1B	Wild-type aa sequence	β γ	3444/mut^{S}
6376	pPpT4_alpha_SmiI_HRPC1C	Wild-type aa sequence	β γ	3444/mut^{S}

6377	pPpT4_alpha_SmiI_HRPC1D	Wild-type aa sequence	β γ	3444/mut^{s}
6378	pPpT4_alpha_SmiI_HRPC2	Wild-type aa sequence	β γ	3444/mut^{s}
6379	pPpT4_alpha_SmiI_HRPC3	Wild-type aa sequence	β γ	3444/mut^{s}
6380	pPpT4_alpha_SmiI_HRPA2A	Wild-type aa sequence	β γ	3444/mut^{s}
6381	pPpT4_alpha_SmiI_HRPA2B	Wild-type aa sequence	β γ	3444/mut^{s}
6323	pPpT4_alpha_SmiI_HRPE5	Wild-type aa sequence	β γ	3444/mut^{s}
6324	pPpT4_alpha_SmiI_HRP01805	Wild-type aa sequence	β γ	3444/mut^{s}
6325	pPpT4_alpha_SmiI_HRP22684.1	Wild-type aa sequence	β γ	3444/mut^{s}
6382	pPpT4_alpha_SmiI_HRP22684.2	Wild-type aa sequence	β γ	3444/mut^{s}
6373	pPpT4_alpha_SmiI_HRP01350	Wild-type aa sequence	β γ	3444/mut^{s}
6500	pPpT4_alpha_SmiI_HRPC1A	Wild-type aa sequence	β γ λ	3444/mut^{s}
6501	pPpT4_alpha_SmiI_HRPC1B	Wild-type aa sequence	β γ λ	3444/mut^{s}
6502	pPpT4_alpha_SmiI_HRPC1C	Wild-type aa sequence	β γ λ	3444/mut^{s}
6503	pPpT4_alpha_SmiI_HRPC1D	Wild-type aa sequence	β γ λ	3444/mut^{s}
6504	pPpT4_alpha_SmiI_HRPC2	Wild-type aa sequence	β γ λ	3444/mut^{s}
6505	pPpT4_alpha_SmiI_HRPC3	Wild-type aa sequence	β γ λ	3444/mut^{s}
6506	pPpT4_alpha_SmiI_HRPA2A	Wild-type aa sequence	β γ λ	3444/mut^{s}
6507	pPpT4_alpha_SmiI_HRPA2B	Wild-type aa sequence	β γ λ	3444/mut^{s}
6508	pPpT4_alpha_SmiI_HRPE5	Wild-type aa sequence	β γ λ	3444/mut^{s}
6509	pPpT4_alpha_SmiI_HRP01805	Wild-type aa sequence	β γ λ	3444/mut^{s}
6510	pPpT4_alpha_SmiI_HRP22684.1	Wild-type aa sequence	β γ λ	3444/mut^{s}
6511	pPpT4_alpha_SmiI_HRP22684.2	Wild-type aa sequence	β γ λ	3444/mut^{s}
6512	pPpT4_alpha_SmiI_HRP01350	Wild-type aa sequence	β γ λ	3444/mut^{s}
6513	pPpT4_alpha_SmiI_HRP02021	Wild-type aa sequence	β γ λ	3444/mut^{s}
6514	pPpT4_alpha_SmiI_03523.noSS1	Wild-type aa sequence	α γ λ	3444/mut^{s}
6515	pPpT4_alpha_SmiI_04663.1noSS1	Wild-type aa sequence	α γ λ	3444/mut^{s}
6516	pPpT4_alpha_SmiI_05508.1noSS2	Wild-type aa sequence	α γ λ	3444/mut^{s}
6517	pPpT4_alpha_SmiI_06351noSS1	Wild-type aa sequence	α γ λ	3444/mut^{s}
6518	pPpT4_alpha_SmiI_23190.1noSS2	Wild-type aa sequence	α γ λ	3444/mut^{s}
6519	pPpT4_alpha_SmiI_22489.1noSS2	Wild-type aa sequence	α γ λ	3444/mut^{s}
6520	pPpT4_alpha_SmiI_22489.2noSS2	Wild-type aa sequence	α γ λ	3444/mut^{s}
6521	pPpT4_alpha_SmiI_04791noSS1	Wild-type aa sequence	α γ λ	3444/mut^{s}
6522	pPpT4_alpha_SmiI_06117noSS1	Wild-type aa sequence	α γ λ	3444/mut^{s}
6523	pPpT4_alpha_SmiI_17517.1noSS2	Wild-type aa sequence	α γ λ	3444/mut^{s}
6524	pPpT4_alpha_SmiI_17517.2noSS2	Wild-type aa sequence	α γ λ	3444/mut^{s}

6525	pPpT4_alpha_SmiI_08562.4noSS2	Wild-type aa sequence	α γ λ	3444/mut^{s}
6526	pPpT4_alpha_SmiI_08562.1noSS1	Wild-type aa sequence	α γ λ	3444/mut^{s}

Sequences of the HRP genes. Variable positions are underlined. Introns and other non-coding sequences are marked with a grey colour.

```
>C1A
TCATCTTCTCAGTAATATAGTTTTCCCCTTTAAAAATGCATTTCTCTTCTTCTTCTACTTTGTTCACTTGTATAACCTTAATCCCATTGGTATG
TCTTATTCTTCATGCTTCTTTGTCTGATGCTCAACTTACCCCTACCTTCTACGACAATTCATGTCCTAATGTCTCTAACATCGTACGGGATACT
ATTGTCAATGAGCTAAGATCAGACCCTCGTATTGCCGCGAGCATCCTTCGTCTTCACTTCCACGACTGCTTTGTTAATGTAAATTACTACTTTT
CATATTTCTATTTCGTTACGAATTATTATATGTTTCATGTAACTGATTTTTAAATGCTTATCTATTCATACTGTCTAATTACATTTATCATTTG
AATTTGATTATTAAATCAGTTTTAATGTGATTAATATAAATTTTAAAAAAATGTGATTAATATATAAACATTTTGATGATAAAATTTTTAATTT
GTTTATTTTTCTTTTTTTTTACTGAATAAAATTTTTTTTAGTGGATGACAAAAATGTTACATTGTTGCCGTGGTAAAGATTCAACTGTATATGG
ATGTAATACATTGAATAATAATTTATTAAATGATATATTGGATTTTTGAAAGGGTTGTGACGCATCGATCTTGTTAGACAACACAACATCATTT
CGAACAGAGAAAGATGCGTTTGGAAACGCAAACTCGGCAAGAGGATTTCCAGTGATTGATAGAATGAAAGCCGCGGTGGAGAGTGCATGCCCAA
GAACCGTTTCATGCGCAGATTTGCTCACCATTGCAGCTCAACAATCTGTCACTTTGGTATGCTCCATTGATCCCACTAACTTTTATTCATTACA
ATTATTGCTTTTAATTTTAATATATCAAATAGTCACTTTACATATCAACACGGCAACCTAAGTTTAGAAAAACAAAAATCGGGATATTTTTAGC
TGTTTGAGAACGACATGGAAATTATTTAGTTGTTTCAGAAGGGTCACAATCAATATATAGTATCAACTTTCCAATCATATGACTAGTATTCAAA
ATTAATCGGGAGATATTTGAAACGCGTGGTCCCCGTAGAGAAAACTTGAAATGTTTAATATCACTGAAAAAAATTAATCTCATTAACAAGAATA
TACTTACTTGGCAATACTACGTTTTTAGTTAACCAAAAAGAATGGTTTTCATATTTTTCAAGAAASAGTGGAGACTTAAAAACTTCACTTCAAG
CATATATATCATCAATGAATTTAAATTACACATATTTTTCTCTTAACACATTGAAACTTCTAAATGAGGAAAATAATAATCAAAACAAAATGGT
TATTATTATAGGCGGGAGGTCCTTCTTGGAGAGTTCCTTTGGGCAGAAGAGATAGCTTACAAGCATTTCTGGATCTTGCTAATGCAAATCTTCC
AGCTCCATTCTTCACACTTCCACAACTTAAAGACAGCTTTAGAAATGTTGGCCTCAACCGTTCTTCTGATCTCGTTGCACTGTCCGGTAATTAA
CAAAAATATATTAAACACACCATTTGATATAGTTGTATTTAGTGAGTTTATTAAAGATCTCTCTTTCTTTTGTTAGGGGGCCACACATTTGGTA
AAAATCAGTGTCGGTTTATTATGGACAGATTATACAACTTCAGCAACACCGGTTTACCCGATCCTACTCTCAACACTACTTATCTCCAAACTCT
TCGTGGACTATGTCCCCTCAATGGTAATCTAAGCGCTTTGGTGGATTTTGATCTACGTACGCCAACGATTTTTGACAACAAATACTATGTGAAT
CTCGAAGAGCAAAAAGGACTTATCCAAAGCGACCAAGAGTTGTTCTCTAGCCCCAATGCCACTGACACAATCCCTTTGGTGAGATCATTTGCTA
ATAGCACACAAACATTCTTCAATGCGTTTGTGGAGGCGATGGATAGGATGGGAAACATTACACCTCTTACAGGAACTCAAGGACAGATCAGGTT
GAATTGTAGGGTGGTGAACTCCAACTCTCTACTCCATGATATGGTGGAGGTCGTTGACTTTGTTAGCTCTATGTGAGCATAGTCGACGCCA

>C1B
TTAATTAGTTTTCTTTTTATCTTTAAAAATATGCATTCCCCTTCTTCTACTTCGTTTACTTGGATCTTAATCACATTGGGATGTCTTGCGTTTT
ATGCGTCTTTGTCCGATGCTCAGCTTACCCCTACCTTCTACGACACTTCATGTCCTAATGTCTCAAACATCGTACGAGACATCATTATTAATGA
GCTACGATCGGACCCTCGTATCACCGCGAGTATCCTTCGTCTTCACTTCCACGACTGCTTTGTTAATGTAAGATAATACTTTTTCATATTTCTA
YTGCGTTATGAATTATTGTGCGTTTTATCCTTTAGATATTGATAAATCACCTCAAGTCAAAATTTAATAAAACATTAAATAAAATATGATACAA
AGAGTCAATTTGTTTTGTAGGAATATAATAGAAATTTCAACATGTTTTTAAATATGGGTCCAAAATGTTGAAAACGACATTTCTTATGAAAAAG
AGTGAGTATGTATTAAATCATAATTTGCTATAATTATCGGGTTGTGAAAGTAGTTCTATTCATTTTGACATGTAGATGGTTGCATAGTACGTTT
TGTCTACAACATTTTTTTCTTGATTCATTTTACAAAATTACAAGTTCACTTGCCTCCGAAAAATATGTATAGCCTTACTATGACATAATTACAT
AATTTACATTCAATAATAATTTTTATTTTTTATATAATAATTTTTATTTTTATTTATATAAAAAAGAAAGATATTATTGTTTGTGGTGTCAGTT
GGGTGAAATCGTATCCTAAATAAAAGTCACTCGAGTAACGGTTCTGATCCAGATTAAAAAATCAGTACCAAATTTCACATGGTTAGATGTCGTG
TGTTAGATTTTGCTGTTGAATAATTAAATACTTAACTCTCGTYGACAATGATAATGCTAAAATATTTATGAAATCGGATTCACGCCCGTGTTAC
AGTATTAAGAGCATGGTGCCGTACCAAACACGATACGAATTTAATGGTGACAAAAAATCTCTGTTAAATTGTTACCGGTGGTAAAGAGTTAGCT
ATGGATGTAATACATGTACTAATAATTTGTTAATTAATATTTTGGATGTTTGAAAGGGTTGTGACGCATCGATATTGTTAGACAACACAACATC
ATTTCTAACAGAGAAAGATGCGCTTGGAAACGCAAACTCGGCTAGAGGATTTCCTACAGTTGACAGAATCAAGGCCGCGGTGGAGAGGGCATGC
CCAAGAACAGTTTCATGCGCAGATGTGCTTACCATTGCAGCTCAACAATCTGTTAATTTGGTATGCTCCATTCATTACAAACATTGTTTTTTAA
TTTTAACATATTTTTTAGTTGTTTGAGAGCGTCACAATCTATATTTAGTATCAACAACTGTTCGACTATATGAGGTATTCATGGATTAATCGAG
AAATATTCAAAACGCGTGGTCCCGTTAAGAAAATTTGACAAGTTTAATATCATTGGAAAAATTAGGTCTCATCACAAGGTTTTACCTTGGTAGG
CAAAGCTACTTTATAATTAACCAAAAAGGATAATTTTCATTTTTTCCAAAAATAGTGGAGATAAAACAAACTCCACTTTGAGTATATCATCAAT
CAATTTATACTACATGTTTATCTTTTTCTCTTTACATATTGAAACTTCCGAGTGACAAATTAATCTCACAAAAATAATTATTTTGAATATAATG
GTTACTATTATATTATAGGCAGGAGGTCCTTCTTGGAGGGTTCCTTTGGGAAGAAGAGACAGCTTACAAGCATTTTTAGATCTTGCTAATGCAA
ATCTTCCAGCTCCATTCTTTACACTTCCACAACTTAAGGATGCCTTTGCAAAAGTTGGCCTCGACCGTCCTTCTGATCTCGTTGCTCTCTCCGG
TAGTTAACAAAAGAAAATTAAACACCATTTGATATAAGTTCAATTAGATATTTCATTATTGATCTTATTATATGGTCTTTCTTTTGTTAGGTGG
TCACACATTTGGAAAAAATCAGTGTAGATTTATTATGGACAGATTATACAACTTCAGCAACACCGGACTACCCGACCCTACCCTCAACACTACT
TACCTTCAAACTCTTCGTCAACAATGTCCCCTAAATGGAAACCAAAGTGTATTGGTGGATTTCGATCTGCGTACGCCAACGGTTTTCGATAACA
AATACTATGTGAATCTTAAAGAGCAAAAAGGTCTCATTCAGAGTGACCAAGAGTTGTTCTCTAGCCCCAATGCCACTGACACAATCCCCTTGGT
GAGGTCATTTGCTGATGGCACACAAAAATTCTTCAATGCGTTTGTGGAGGCTATGAATAGGATGGGAAATATTACACCTCTTACAGGAACTCAA
GGAGAAATCAGGTTGAATTGTAGGGTGGTGAACTCCAACTCTCTACTCCATGATATAGTGGAGGTCGTTGACTTTGTTAGCTCTATGTGAGAAA
AGTTGAGTCAATATCTGGCTACCAGAGTACACGTTAAGATAAATAAAGCGCTCTCAAGATGTTACTTG


>C1C
TTTTTTTTTTCTCTTAAAAAATGCATTCCCCTTCTTCTACTTCGTTTACTTGGGCAACCTTAATCACATTGGGATGTCTTATGCTTCATGCATC
TTTTTCCAATGCTCAACTTACCCCTACCTTCTACGACAATTCATGTCCTAACGTCTCAAACATAGTACGGGACATCATTATCAATGAGTTACGA
TCGGACCCTCGTATCGCCGCGAGCATCCTTCGTCTTCACTTCCACGACTGCTTTGTTAATGTAAGATATTACTTTTCATATTTCTATTGCGTTG
TGAATTATTGTGTTTTATCTTTTTTTGAAGATATGTGTTTTATCTTTTAGATATTGATATATCACACCTCTAGTCAAAATTTAATAAAACATAA
AATAAAATATAATACAAAGAGGAGTATAATTAATAAAAAATTTAACATGTTTTTAATCTGGCTCAAAAATGTCGAAACGAGATTTCTTATGAAA
AAGAGTAAGTATGTATTAAATCATAATTTGCTACAATTATTGGGTGGTAAAAGTAAGGAATTCTATTCATTTTGAAATTTAGCTGGTTGCATAG
CTTTATTCTACAACATTTTTACTGGATTCATTTTATAAAGTTATAATTTCACCTTTTTTTTTTGTTAAATAGTTATAATTTCACTTGCCTTCGG
AAAATATGTATAGCCTTACTATAACATAATTACATTCAATAATAATTTTTATTTTATTTATATAAAAAGATTATATTATTTTTTGTTTTGTTCA
```

```
GTTAGAGTGAACCGTATCCTAAATAAAAGTCACTCGAGTAACGGGTCTGATCCAGACAAAAAAAAAAAATTCAATACCAAATTTCACATGGTTA
GATGTCGTATGTTTAATTTTGCTGGTGAATAATTAATACTTAACTCTCGTCGACAATGATAATGCCAAAACATTTATGAAATCGGATTCATGAA
TTGTCCGTGTTACAGTATTAAGAGCATGGTGCCGTACCCCAAACACGATATGAATTTAATGGATGACAAAAAAAAAATTCATTGTTAAATTGTT
ACCGGTGGTAAAGATTTAGCTATGGATGTAATACATTTACTAATAATTTGTTAATTAATATTTTGGATTTTTGATAGGGTTGTGACGCATCGAT
CTTGTTAGACAACACAACATCATTTCGAACAGAGAAAGATGCGTTTGGAAACGCAAACTCGGCTCGAGGATTTCCTGTGGTTGACAGAATCAAG
GCCGCGGTGGAGAGGGCATGCCCAAGAACTGTTTCATGCGCAGATGTGCTTACCATTGCAGCTCAACAATCTGTTAATTTGGTATGTTCCAATA
ACTTCTGTCATTACAAACATTGTTTTAATTTTTTTTTWTTTTTGTTAAAATCATTGTTTTAAATTTTAATATATCAATTCCTCACTTTTACAT
ATATCAAACGTACAACGATAGCCTAAGTTTGAAAAGAAAAGTATTCAAAACGTCACAATCTATATATTTAGCATCAACAACTTTTCAACTATAT
ATATGAGGTATGTATGGATTAATCTAGAAATATTCAAAACGCGTGGTCGCGTTATTTGACAAGTTTAATAATATCATTGAAAAAGTAGGTATCA
TCACAAGAATTTACCTTGTAATAGGCAAAGCCTACCTTTATAGTTAACCAAAAAGAATAATTTTCATTTTTTTCCAAAAATTAGTGGAGATAAA
ACAAAGCTCCACTACTTTGAGTATATATTATCAATCAATTTATACTACTTGTTTATCTTTTTCTCTTTACATATTGAAACTTCCGAGTGACAAA
ATTAATCTCAAAAAAAATAATTATTTTGAATATAATGGTTATTATATAGGCAGGAGGTCCTTCTTGGAGGGTTCCTTTGGGAAGAAGAGACAGC
CGACAAGCATTTTTAGATCTCGCTAATRYGAATCTTCCAGCTCCATCCTTCACACTTCCASAACTTAAGGCTGCTTTTGCAAATGTTGGCCTCA
ACCGTCCTTCTGATCTCGTTGCTCTCTCTGGTAATTAACAAAAGAAAACTAAACACCATTTGGTATAGTTTAATGAGCGATTTCATTATTAATC
TTATTATGGTCTTTCTTTTGTTAGGTGGTCACACATTTGGTAAAAATCAATGTCGATTTATTATGGACAGATTATACAACTTCAGCAACACCGG
ACTACCCGACCCTACCCTCAACACTACTTACCTTCAAACTCTTCGTCAACAATGTCCCCGTAATGGTAACCAAAGCGTCTTGGTGGATTTCGAT
CTGCGTACGCCAACAGTTTTCGATAACAAATACTATGTGAATCTTAAAGAGCAAAAAGGTCTCATCCAGAGTGACCAAGAGTTGTTCTCTAGCC
CCAATGCCACTGACACAATCCCCTTGGTGAGATCATATGCTGATGGCACACAAACATTCTTCAATGCCTTTGTGGAGGCCATGAATAGGATGGG
AAACATTACACCTCTTACAGGAACTCAAGGAGAAATCAGGTTGAACTGTAGGGTGGTGAACTCCAACTCTCTACTCCATGATATAGTGGAGGTC
GTTGACTTTGTTAGCTCTATGTGAGAAAAGTTGACTCAATATCTGGCTACCAGAGTATACGTTAAGATAAATAAAGCGCTCTCAAGA


>C2
AAACCTAACCAAAGAATTTTATCTTAGAGAGCAAAGAAAATGCATTCCTCTTCCAGTTTGATAAAATTGGGATTTCTTCTTCTTCTTCTTAATG
TATCATTGTCTCACGCTCAACTAAGCCCTTCATTTTACGATAAAACATGTCCACAAGTCTTTGACATTGCAACCAATACCATTAAAACTGCGCT
GAGATCAGACCCTCGCATCGCTGCGAGCATCCTTCGTCTTCATTTCCACGACTGCTTTGTTAATGTAAGATACCACTAGATTCAATATTTTCAC
GTATATATTAATTAAGGCCAAATGACTTTTTATTAACGCTAGTAAAAAACCTGTTTAAGTACTTCAACTGTAAACTACCATCTTAAATCTAGCT
AATCTATATGGGTTACAATATTTTTGTACACAAGATTATATTAGATATTATATAAAATTAAAATTATTCATATGGTGTCTAATACTTGTCGCCG
TTCACCAATATATCCATAAATTATTGTCCTGTCGCAGAAAAAAATACAAAAGAGTTGAAATATACTAAAAATATAATCTTACTAGAGATGAGTA
TTTATCATAGATTTGTTGTCGTTTTCGGATATGATTGATTAAGAAATTTTATTATATAGAATGTCTTTTGGTTCCCTGATATATCACAATCACA
AGCTTGATATTAGGATATCAAAAAGTGATAGACATTTTGATTTGTGGTTCGACTCAAATTGTTTTTTTTGCTTAAAATCGACTCAAATTGTTCG
TTTCTCGGTTTATTCATTTCAAATAGAATATACTCAATAAAGTCTCAAACTGTATAAAAGTTTGTGTTAGTTTTGGTTATTGAAATGTTGGGTT
TCTTTTGGTTCAGATTAAAACAAAAGAAAACACCGGAGTTAAATAAAAATAGAAATAACAAAATAAAATTATTATATCAACAAAAAATATGTCT
TTTAGTTAAATTATTTTATAAGAACTTGATTCGACTTATTTTGACTAATTTTGGTTCAATTTGGAGGGGCAGTATTTATAAATGGACCAATCGC
TTCTTAAGCAAATTAATATGATTTAAATAGTACAACGTAGATTGCTTTCTTATCTTAACTCTCAAGTTTCAGATTCTGTCGTCAAAGTATAGTC
AAAGATGATTAGTTAATAATTAGCGAACTCATGGCTGTTTCAATTAACTCGGCAAAAACAAAAACAAGTATTGGACTTTTGATGGATTTTTATT
TGGGAGATCAAATCAGTTTTTTTGTCCACCATCGAAATTTGAACTCTAAATTTTCAAAATCTCTCTCTTTTTTTTTTTTTTTTTGTTAAACCTC
TTTATAATTTTTTTTTCTCACAGCAAATCATTATTTACTGATTAGATAGCTCAGGAATAGGATGGGAAACACAAATTCTTGCACTGTTTACGGA
TGTTAAGATGTTTTGATTTAACTATGAATATGGTAGTTGACTAAATGTGAAGAACTATATTTAATTTTGAAACAGGGATGTGATGCATCGATAT
TGCTAGACAACACTACATCATTTAGGACTGAGAAAGATGCGTTTGGGAACGCAAGATCAGCTCGAGGCTTCGACGTAATCGACACAATGAAGGC
TGCCGTGGAGAAAGCATGTCCTAAAACCGTTTCATGTGCTGATTTGCTCGCCATTGCAGCTCAAAAATCTGTCGTTTTGGTATTCTTCTTTTAC
AATGCTATGCATCTATACATCTTTATTTTTCTTTCCTTTTTCATTTAAATGGTTTTCATATCATCTTATTGATTTTAATTTTTAGGAGAAACAA
GTATTTAAACCATGCAAATATAATTGTTTACTTGAAAATGAAAATAAAATAGGCGGGAGGTCCTTCATGGAAGGTTCCAAGTGGAAGAAGAGAC
AGCTTAAGAGGGTTCATGGATCTCGCTAATGATAACCTTCCAGGTCCATCCTCTACACTTCAAGTACTTAAGGACAAATTCAGAAATGTCGGAC
TCGACCGTCCTTCTGATCTCGTTGCTCTTTCTGGTACATTATAGTTAAAAACATTTTCATTTTCATATATAACCTATAGTTGTAGTCATTAATA
AACTCTAAATTATTATTTGCTTTGGTTTATTAATTCTTTTGATATTTTTCTTTGGTTAGGTGGTCACACCTTTGGCAAAAACCAATGTCAGTTC
ATAATGGATCGGCTTTACAACTTCAGTAACTCCGGTAAACCCGACCCAACCCTTGATAAATCGTACCTCAGCACGCTAAGAAAACAATGCCCAC
GTAATGGAAACCTGAGTGTATTGGTAGATTTTGATTTACGTACACCGACAATCTTTGACAACAAATACTATGTGAATCTCAAAGAGAACAAAGG
TCTTATCCAGAGTGACCAAGAGTTATTCTCTAGCCCTGATGCTTCTGACACTATCCCTTTGGTCCGAGCATACGCTGATGGTCAAGGAAAGTTT
TTTGATGCATTCGTGGAGGCAATGATAAGGATGGGAAATCTTTCACCTTCAACTGGGAAACAAGGAGAAATTAGATTGAATTGTAGAGTGGTGA
ATTCTAAACCTAAAATCATGGATGTGGTTGATACTAATGACTTTGCCAGCTCCATCTGAAGAAATGACTTTCTCCTAATAATAAATGATCAAT


>C3_NCBI
GAATTCAATTTTCAGAAGGAATCATTTAATTTATGTTTTGAAATAAATATTGAATCAAGAATATAGGCGGGATCACCATCTCTTTATAGTATTT
AATGTCGATTTGCAATAAATCTTTATGGTAATATGATCAATATCAATCTTAGTAGCTTATTATCTTAGGGTTGATTTAATTATCACAGCGACAA
ATACAGCCAAAAACTAAAAGTTATAAATTATGTGCAAATAATCTAGCATTGTCTAAGAAATCGGTGAATAGATTTGAAAATTTAATTGTCCTTC
AACCACTCTATAGTATGGCCTTAGAACAATACCTAATACCAAAAGAAAAAAATCAACCTAACAGTGTAAAATACTATAAAAGAAAATACTCCAG
CTACTCTATATAGTGTTCCTTCTACGCCTTAGCTAGATTCACACCATCAGCCACACTCTCAACTGATCAAATCATAGTTTGTCTTCTTCCTAAA
AAAGAAAAAGAAAAATGGGGTTTTCTCCTCTCATTTCCTGCAGTGCTATGGGAGCCCTAATATTGAGTTGCCTTCTGCTTCAAGCTTCAAACTC
TAATGCTCAGTTGAGGCCTGACTTCTACTTTAGGACTTGCCCATCTGTTTTCAATATTATTGGGGATATCATTGTCGATGAACTGAGGACTGAT
CCTCGTATTGCCGCTAGCCTTCTTCGCCTTCACTTTCATGACTGCTTTGTTCGTGTAAGTGTAAGGACTTAACTTTTTTTTTTTTAAACTATGAC
GTGTTCATTGGACGTAACTACTTTTCACCATTTAATTCACATATAGAATAGAGGCCAAAAGGAATATTCGAATCAATAAATACAAGCGTCATAT
AATGTCATATATATATATATATATAATTTTGTAGGGTTGTGATGCATCGATCCTGCTTGACAATTCCACGTCGTTCCGAACCGAAAAAGATGCTGC
TCCAAACGCAAATTCAGCTAGAGGATTTGGTGTCATAGATAGAATGAAAACATCCCTTGAGAGAGCTTGCCCAAGAACAGTGTCTTGTGCAGAT
GTTCTCACCATCGCCTCTCAAATATCAGTGCTTTTGGTATGTACATGATTTATAACGGATGATATTAATCCAATATGTTATGGATTTGACGTCA
ATGCTTTATAAGTTATGAAATTTGATTCAAAATGTTATGAATTTGATGTCAATTCTTTATATATATGTTATAGTCGGGAGGTCCATGGTGGCCG
GTTCCGTTGGGGAGGAGAGACAGCGTCGAAGCTTTCTTCGATTTGGCTAATACAGCTCTTCCCTCTCCATTTTTCACTCTTGCTCAACTTAAAA
AAGCTTTCGCTGACGTTGGCCTAAACCGCCCCTCAGATCTAGTCGCTCTTTCTGGTAAAATATTCATGATGTTTCTAATATAAGTGTTTTTGAT
CTAGCTAGATCTATGCAATTCATTTTATATAATGATAGCTAAATGGATGCACTCCTCCACTAAGTCTGGAATTTACATATTAATTTATAAGTTA
TAGAATTACAAATTTATAATTCAAATTTTATTATTTTGAATTCTATTAATATGATCTACTATTATAATTCATATTTATTTTGTAATTTCTTAGG
AATGATTTGCTTTTGGCTTTGAAATGCATGACCAAGTATAAAATAAATTAAAAAAGGATAATATAATTAAATAATAAATAATACCATCTCAAAC
AAAATGTCTTCAAGCACGTGATTTTGAATTAAAACATGAATTAGTAAAACGAAAAAATCATGACTTTTTTGTTGTCAAACTTTTTGTTTTTTTT
TTTTAATTTTCCAACTGACAAACCATAAAAACTTAAACATGCAATGTATCTAAACCTTAGTTATTGTCAAAATGGTAGGTGGTCACACATTTGG
AAGAGCACAATGCCAATTTGTGACACCTCGTCTCTACAACTTCAACGGTACAAACAGACCAGACCCAACTCTGGACCCAACTTACCTTGTCCAA
CTCCGTGCATTGTGCCCTCAAAACGGAAACGGCACCGTTCTGGTCAACTTCGATGTCGTGACTCCGAATACTTTTGATCGTCAATACTACACCA
ATCTTCGTAATGGGAAAGGTCTGATTCAGAGTGACCAAGAGCTCTTCTCGACTCCAGGAGCCGACACGATCCCACTAGTAAACCTATACAGCAG
CAACACGTTCGCGTTCTTCGGAGCATTCGTTGATGCAATGATTAGGATGGGAAATCTTAGACCTTTGACTGGAACTCAAGGCGAGATAAGACAG
```

```
AATTGTAGGGTTGTGAATTCGCGAATTAGGGGTATGGAGAACGATGATGGAGTTGTGAGTTCTATTTGATTATGTTGGGAATATGGTTATGTAA
CAAATCATAAAATGTGTGGGAACATGCATGTCGACTAAATAAAAGCTCTCACGAGTTATGACTTGTGAGATTACAACTGAAAAAACCAAAGGAA
AGAAAAGATCAGATTTTGGATCCAAGTAAGGTGATGAGGAAGTCCTTGTAGTCTCCAGAAACATCACCGATAATAGCGTTATCCATGCTTGAAT
TGTACATGTTGAAATACTCTCCTCTCACCTTCATCAAATCAATCTCTGCACGCGTCACAATTGCTCTCATCAACGAATCTTCATCTGTTCCAAA
ACCCTCAATCGAATCTCTTACAACCTGCCAAATAATAAACACAAACTTGAGGAACACAAACACCTTCTCTACTTAATCTATGGCAATTTCACAC
CTTGCAA

>A2
ATCATACCTCTAAAATCATTATTTTGTAAAACCTAATTAATGGCTGTAACAAATCTATCTACTACTTGTGATGGTTTGTTTATCATCAGCCTTC
TTGTTATCGTTTCTTCATTGTTTGGAACATCATCTGCGCAGCTAAATGCAACGTTTTACTCCGGGACTTGCCCTAACGCATCTGCCATCGTACG
CAGCACTATTCAGCAAGCTTTTCAATCCGATACAAGAATCGGAGCCAGCCTCATCCGCCTTCATTTTCACGACTGCTTTGTTAATGTATACTAA
TCTTCCCAATGCAGCTCTTTACATAAAGGCTTCTTGATATTTTTCGCTCTAAACCGCTACTTTGCTTCTTTATTTTTTCAAAGGGTTGTGATGC
GTCGATCTTGCTTGACGACAGTGGAAGCATCCAGAGCGAGAAGAACGCTGGTCCGAATGCAAACTCAGCTAGAGGATTCAATGTTGTCGATAAC
ATCAAGACTGCCCTTGAAAACACTTGCCCTGGTGTTGTCTCTTGCTCTGACATTTTAGCCCTTGCCTCAGAGGCTTCCGTGTCTTTGGTAATTA
GTAATTACACTTTCTTTGTGAACATATGAAACAAAACATAACTAAAAATTTCAAATTTGCTCTTAATTTCTTGTTATATATATATAATTTGTCT
TATAAATTATGTTTTAGTAATAATATAGATACGTATATGTTCATATATATGTTTGATCATCTTCAGACAGGAGGGCCATCATGGACTGTATTAT
TAGGAAGACGAGATAGTCTCACCGCAAACCTCGCCGGGGCAAATTCGGCTATTCCTTCTCCCTTCGAAGGCCTTAGCAATATCACATCTAAATT
TTCGGCTGTCGGGCTAAACACGAACGATCTAGTAGCCTTATCTGGTAAGTTCATCTACATGTTTAGTTACTTGCGGTTCAAGTTAATTCAAAAC
CCTGACGTCATCTCTTGTCTACGTAGGTGCGCATACGTTCGGGCGTGCTCGATGTGGAGTGTTCAACAACAGACTATTTAACTTCAGCGGGACA
GGAAATCCAGACCCGACTCTAAACTCAACGCTACTGAGCAGTCTTCAACAGCTATGTCCTCAAAACGGCAGCGCATCAACCATCACCAATCTCG
ATCTGAGCACACCTGATGCGTTCGATAACAATTACTTCGCCAACCTTCAGAGCAACAATGGGCTTCTTCAGTCAGACCAAGAGCTGTTCTCTAC
CACGGGTTCAGCCACAATARCGGTTGTTACCTCCTTTGCAAGTAACCAGACTCTGTTTTTTCAGGCCTTTGCTCAGTCCATGATCAACATGGGG
AATATTAGTCCCTTGACAGGGAGTAATGGAGAGATTAGATTAGACTGTAAGAAGGTTAATGGAAGTTGATTTCCATAAAGCTCTTGTTTTTCAA
AAAACACATAAT

>E5
CATAGTCTATCATCCTCCTAAAAATTAAAGAGAAATGGTGGTTTCTCCTTTCTTTTCTTGCAGTGCTATGGGAGCCCTAATATTGGGTTGCCTT
CTGCTTCAAGCATCTAATGCTCAGTTGAGGCCTGACTTCTACTCTAGGACTTGCCCATCTGTTTTCAATATTATTAAGAATGTCATCGTCGATG
AACTGCAGACTGATCCTCGCATTGCCGCTAGTATCCTTCGCCTTCACTTTCATGACTGCTTTGTTCGTGTAAGTACTTAACTTATGTTTTTATT
ATTAATAAAAAAAAACCATGACGAATTCATATTGGACGAAACTACTTTTTATACCATTTAATTTACTTATAGGTTAGAGGCCAAAAGGCATATT
TGAATCAACAAGTACAATCGTCATATAACGTATAATATCTATGGTTTTTGTAGGGTTGTGATGCATCGATCCTGCTYGAYACTTCCAAATCGTT
CCGAACCGAAAAAGATGCTGCTCCAAACGTAAATTCGGCTCGAGGGTTCAATGTCATAGATAGAATGAAAACAGCACTTGAGAGAGCTTGTCCT
AGAACAGTGTCTTGYGCAGATATTCTCACCATCGCCTCTCAAATATCAGTGCTTTTGGTATGTACATATACCTATATATGACTTATAATATCGG
GTGAAATTAATAAAAATATGTTATGAAATTTTGACGTCAATGCTTTATATGTTATAGTCGGGAGGTCCATCTTGGGCAGTTCCGTTGGGGAGGA
GAGACAGCGTAGAAGCTTTCTTTGACCTAGCTAATACAGCTCTTCCCTCTCCATTTTTCACTCTTGCTCAACTTAAAAAAGCTTTTGCTGACGT
TGGTTTAAACCGCCCCTCAGATCTAGTCGCTCTTTCTGGTAAACTATATATATTCATGTTGTTTATAATATAAAGTGTTTTATATAAAATGATA
GCTAACCACACCCACTCCGCGATCGATGAGGTCTAGAATTTACACACTAATTTATAAGTTATAGAATTACAAAATTTTATTATTTTGTATTTAT
TAATATGATTTGCTTTTGGCTTTGAAATACATGACCAAGTATAAAATGAAAAACAAAATGGATAATATAATCAAATAATAATACTATTTCAGAC
AAAACATCATCTAGCACGTGATTTTGAATTATAAGATAAGATGGTAAAACGACAAAACATGACTATTATTTTTTTACCTTTTTTTTTTTTAATT
GTCAAACTGACACATACAATGTATCTAAACCTTATTGTCAAAATGGTAGGTGGTCACACATTTGGAAGAGCACGATGCCTATTTGTGACAGCTC
GTCTCTACAACTTCAACGGTACAAACAGACCAGACCCAACTCTGAACCCATCTTACCTCGCCGACCTCCGTCGATTGTGCCCTCGAAACGGAAA
CGGCACCGTTCTGGTCAACTTCGATGTCATGACTCCGAATACTTTCGATAATCAATTCTACACTAATCTTAGAAATGGGAAAGGTCTGATTCAG
AGTGACCAAGAGCTCTTCTCGACTCCAGGAGCCGACACGATCCCACTAGTAAACCTATACAGCAGCAACACGTTATCGTTCTTCGGAGCATTCG
CTGATGCAATGATTAGGATGGGAAATCTTAGACCTTTGACTGGAACTCAAGGCGAGATAAGACAGAATTGTAGGGTTGTGAATTCGCGAATTAG
GGGTATGGAGAACGATGATGGAGTTGTGAGTTCTATGTGATTATGTTGGGAATATATATCATATATGGTTATGTATCAAATCATAAAATGTGTG
GGAACATGCATGTCGACTAAATAAAAGTTCTAACGAGTTGTGAAATGACTATGAAATTTTG


>01805
CTTAAACCAATAAAAGATAAGTTTCCTCTTACCAAAAATGCATTTCTCTACTTCTTCTTCTTCCTTGTCTACTTGGACAACCCTAATAACATTG
GGGTGTCTTATGCTTCATTCATTTAAGTCCAGTGCTCAACTAACCCCTACCTTTTACGACAGTACCTGCCCCAGCGTCTTTAGCATCGTACGGG
ACACCATCGTGAACGAGCTAAGATCAGATCCTCGAATTGCTGCAAGTATCCTTCGTCTTCACTTCCACGACTGCTTCGTTAATGTAAGATATTA
TTTTTCAGTTTAATTGATGATTTTGTGGATACGCTAGTATCTTTTATAATTGAATAAAAGAATTATTTTATCTAATTTTCACACATATAATTGA
TATATTTGGTACTAATATAAGTTCTTCAAAAACAATTTCAACGATTTGTGTTGTCAACTTAGTTGCAATTCGGATCGACTTGTAGTTAACGTAT
TTGTTTTCTGAAATAAGTTTTTTGGGTGGGGATAATGGCGCCACTCAAGAAAAACCAGTTATTTCTTTGTAAGTTTTAATTGTTGAGAAAGACA
ACAAAAAAAAACGCGTGGATACGGCACTTACATTTATTAAACTAAAGGTTTGTCAGAATTAATTAATTATAGTAATAGTTGTTTCTGTGGTATT
AGTTTTTTTGTTTGTTAAAGCGTTTCTGTGGTATTAGTTACCATGACATGTTATCATCAATACAAAAGTTCACGTAGTTAGATGTAATATTTTT
GGTTTTTACTGTCGAGTATTTTTCTTCTTGTCGTCGACTCATGACATAAAATAATGATTCCAAAACATCTATTAAACTTTGACCTCCCATTCGG
CGAACATGAATTATTCGTACAATAATTTAATAAAAGCATGGACCCGTACAAAGCACGTTATGAAAATTTAAATCAGTCCCTTTTTTTAGCATAA
AAATAATTTTTTTTGTGAAATCATTGAAAAGTCTCAAGCTTATCCTCAACCCCAAATAACCTACAGATATGAAAAAGAAACAGAAATGTACGTC
AAAACCTACTATATAATTTAATTTATTATTTTTGGGATTTTCAACTCTCGAATCTGATGTAGTTGTTAGCATACTTTATTTGAACTTTTAATGT
TTTTATTCGAAAACATAGTATATACCCAGACCCAGATATTATAGCCAAATAATGTAGACCAACAACTTTTTGAATTTTGTGTGAAAAGATTTTA
GATTTATATGTAAACTAATCTGTTTATAATTATGTATATATTATCATGCCTTGTTGAAGAGGAACAAATGGTTTCAAGGAATGAAAATGCAATC
TAACAATGTTTATAGGTAGTAATTAAAGATTTGATTATGATTTCAACTTATTTCTTTGCAAAAGGGCTGTGATGCATCGATTTTGTTAGACAAC
ACAACATCATTTCGAACAGAGAAAGATGCAGCTCCAAACGCAAACTCAGCTCGAGGATTTCCAGTGATTGATACAATGAAAGCTGCAGTGGAAA
GAGCATGTCCAAGAACCGTATCATGCGCAGATTTGCTTACCATCGCAGCTCAACAATCTGTGAATTTGGTATGCAACATTAATTTATCTCTTTT
TTTTAGTTTTAATTTTATCTATTTTTTCTCCACTTCTTGCAAACATGGTCTCAAATTCCTTGTTGAAGCCGTAACTGAACCCGTTAAACATGAC
TACATAGCATTGAAAAATGAGATGGACTGTTGCATGTTTTATAATTTAGATTTAATATTACCATTCTTAAATATTAATTACCATACTTGTCTTG
AAAATTCATTGTAAGAAATAATCAAAAACACATGGTAATAATATTGTCTTGTTTTTGTAAAAGGGTTGGGTTTTCTAATAATATAAAGTCTCAT
TACAATAATAAACTTGATTAGATAAAAACTCTTTTCTGAATTAACTAAAGTGAATATAAGCAATATTTGTAAAAGAACTGGGAAAATAACAGTA
CAATTGAGAACCTAATTTTGGGATTTATATTTTATATATCCTTTTTAATAATTATGACTTTGACGTGGTAAGAATTATTTTCATCTATAAAATC
ATTATTTTAAGATATGGGTATTTTAGGCAGGAGGTCCTTCTTGGAGGGTTCCTTTGGGGAGAAGAGACAGCGTACAAGCATTTTTTGATCTTGC
CAATACAAATCTTCCCGCTCCATTCTTCACGCTTCCACAACTTAAGGCCAGCTTTAGTAATGTTGGACTTGACCGTCCAGAAGATCTCGTTGCA
CTCTCTGGTAATTATGAGGAGTAATAGTAACAACCAAAACTTTTATTTGATCTAATTAGTTTATAAAATATTATTAATTTATCATCTTTTGATT
AGGTGGTCACACATTTGGTAAAAACCAATGCCAATTTATTATGGACAGACTATACAACTTTAGTAACACTGGTTTACCCGACCCTACTCTCAAC
ACTACTTATCTCCAGACACTTCGTGTACAATGTCCCCGTAATGGTAACCAGTCCGTCTTGGTCGATTTCGATCTACGCACACCGACAGTTTTTG
ACAACAAATACTATGTGAATCTGAAAGAGCACAAGGGACTTATCCAGACCGATCAAGAGTTGTTCTCCAGCCCTAATGCCGCTGATACAATCCC
CTTGGTAAGATCATATGCTGATGGCACTCAGAAGTTCTTCAATGCTTTTATGGAGGCCATGAACAGAATGGGAAACATTACCCCTCTCACTGGA
```

```
ACTCAAGGACAGATCAGGCAAAATTGTAGGGTGATCAACTCCAACTCGCTGCTCCATGATATTGTTGAAATCGTTGACTTTGTGAGCTCTATGT
AACAATAGTTGTCTCAATATATGTGGCAACCAAAATTATATGTTCTTATGAAAATAAAATGTTCTCGAAACATTACTTAAG

>22684
AAAATGGGGTTTTCTCCTTCATTTTCTTCCAGTTCTATAGGAGTCCTAATATTGGGTTGCCTTCTGCTTCAAGCTTCAAACTCTAATGCTAAGT
TGAGGCCTGACTTCTACTTAAAGACATGTCCATCAGTTTTCCAAATCATTGGGAATGTCATCGTCGATGAACTGCAGAGTGATCCTCGTATTGC
AGCTAGTCTCCTTCGCCTTCACTTCCATGACTGTTTTGTTCGTGTAAGGACTTAATTACTCAACTTATCTTTTTATCCGAAAAAGAAAAAAACA
TGACGTGTTCAATGGACAAAATGACTTTTCAATACGGAAGTAGAGGCTAAAAGCAATTTTTAATTAATAAATACAATCTTCATGTATTATAATA
TGGTTTTGTAGGGTTGTGATGCATCGGTCCTGCTCGACAATTCCACATCATTTCAGTCCGAGAAAGATGCTGCTCCAAACGCAAATTCGGCTCG
AGGGTTCGACGTCGTAGATAGAATGAAAGCAGCCCTTGAGAAAGCTTGTCCTGGAACAGTGTCTTGTGCAGATGTTCTTGCCATCTCCGCTCAA
ATATCAGTGCTTTTGGTATGTACATATACCTATATATGACTTATATCGGGTGAAACTAATCAAAATATATTATGAAATTTTGACGTTATGTTAT
ATATTATATAGTCGGGAGGCCCATGGTGGCCGGTTTTGTTGGGGAGGAGAGACGGCGTAGAAGCTTTCTTCGATTTGGCTAATACAGCTCTTCC
CAATCCATTTGCCCCTCTTACTGAACTTAAAGAAAAATTTGCTGACGTTGGCCTAAAGCGCGCCTCAGATCTAGTTGCTCTTTCCGGTAAAATT
TTCATATTTTTCAATCTTTCTTGTTTTGGTCAACCATATTGTTTCAATCTATATAACCTACTCTATGTATTATTGTTTTTTTTATAATCTAACT
AGATATAATTCATTTTATCTCGATAACTAGGTAGATATCGAGTTAGTCTGGAATTAACTTAGTAAAGTTATCTTCTAGGTAGATATCGAGTTAG
TCTGGAATTAAGATCAAATTTGCAATGTCAACACAAGAAAAGTGACTTGAAATTAAAGATGAGTTGGTCAAACGACATGACATGAGTCATCTTT
AATAAAGTTAAACATATATACAATGTATCTAAACTTTACTTTTTTTTGTGGGGTCAAAATGGAAGGTGCTCACACATTTGGAAGAGCACAATGT
CTACTTGTGACACCTCGTCTCTACAACTTCAGCGGCACCAATAAACCAGACCCAACTCTGAACCCATCTTACCTCGTCGAACTCCGTCGATTGT
GCCCTCAAAACGGAAACGGCACCGTTCTGCTCAACTTCGATCTCGTGACTCCAAATGCTTTCGATCGTCAATACTACACCAATCTTCGAAATGG
GAAAGGTCTGATTCAGAGTGACCAAGAGCTCTTCTCGACTCCAGGAGCCGACACGATCCCACTAGTAAACCTATACAGCAAGAACACGTTCGCG
TTCTTCGGTGCATTCGTTGACGCAATAATTAGGATGGGAAATATTCAACCTTTGACTGGAACTCAAGGCGAGATAAGACAGAATTGTAGGGTTG
TGAATTCGCGAATTARGGGTATGGAGAACGAYGRTGGAGTTGTGAGTTCTATTTGATTATGTTGGGAATATGGTTATGTAACAAATCATAAAAT
GTGTGGGAACATGCATGTCGACTAAATAAAAGCTCTCACGAGTTATGACTTGTGAGATTACAACTGAAAAAACCAAAGGAAAGAAAAGATCAGA
TTTTGGATCACCAGCCCGGGCCGTCGACCACGC

>01350
CAACTCCAACTCCAAGTCTATTCAAAGTCTTTGTTTAACCTAAACATGGCTTCAAATCAACGTATTTCCATTCTAGTTCTCGTAGTTACATTTT
TAGTGCAAGGTAATTACAATAACGTCGTTGAAGCACAACTGACGCCCAATTTCTACTCAACCTCTTGCCCTAACCTCCTCTCCACCGTCCAATC
CGCCGTTAAGTCTGCCGTTAACAGCGAGGCTCGAATGGGTGCATCTATCGTACGCCTTTTCTTCCACGATTGCTTCGTCAACGTTAGTTTTTTT
TTAACTTTTTTTTTTTGTTTTTAGTTTTCCTTGCATTCCAAGAAACTTAGCGTAATGTTTTTTTTTTTTTAATTTTCTTTTCTAATCATAGTCATA
ACTTGCAAATATATATAAAAATATAGGGATGCGATGGTTCGATTTTACTAGATGACACATCAAGCTTCACGGGAGAACAAAATGCGAACCCAAA
CCGCAATTCCGCTCGCGGGTTTAATGTGATCGACAACATCAAAGCAGCGGTCGAGAAAGCATGTCCCGGGGTCGTGTCTTGTGCTGATATCTTA
GCCATCGCAGCTAGAGACTCCGTCGTAGTCGTAAGCTCCCTATGTCCTCCTCTCTTAGTCCGGTTTTGCAAATTTAAAAGATTAATTAAACGGG
TCTAAAAAATCGTCTTTGTTCTCGTTGAAAGCTTGGAGGGCCTAACTGGACTGTGAAAGTAGGAAGAAGAGATGCGAGAACGGCGAGTCAAGCG
GCGGCGAATAGCAACATTCCGGCGCCCACTTCTAGTCTGAGCCAACTCATTAGTAGTTTCAGTGCCGTTGGACTCTCCACCAGAGATATGGTTG
CTCTCTCCGGTCCGTTTCATCTCTCTTCTTAACTCAAATTTTTTTTTTTATCATATATGCAAATATTTGTTTCATAGATTAAGGCTGATTGCAA
TTGTTAACGGTTACTAATATATACATTACTAAAATTGGTGACAATAAATTTCTTTTGTTAATGTGTAATTGCTTTCACTAATTTGATTTACAAA
AAAGTAAAAAACAAAATCACAAAAATAAATATTCAAAGGACTAAGAGTCAACAGAATCGACATAAAAATAAACTAATATGATTGTGTCGTTTGA
TTTTTGTCGACTCTCCCCCTTTAAGTTCTTGTTTCTTTGTGATTCTTTTTTCAACACATTTTTTGTAACGCACCACCGACTTTTCTGTTTCTTA
ACCAAAGTTACGTAATCATATAGATTACTGTTACGCGACACCGACAAAACAATTGACTGATTAATCAACAAATAATTAAGTCCACTATTAAACT
AATTGCTATATGTTCATTTTTTTCAGGCGCGCACACGATCGGGCAATCCCGTTGCACGAGCTTCCGAACGAGAATCTATAACGAGACAAACATC
AACGCCGCATTCGCCACAACACGTCAACGAACTTGCCCTAGAACCTCCGGCTCCGGCGACGGGAATTTAGCTCCACTTGACGTCACCACGGCGG
CTTCTTTCGACAACAACTATTTCAAGAATCTCATGACTCAAAGAGGTCTTCTCCATTCCGACCAAGAGCTCTTCAACGGCGGCTCCACTGACTC
CATAGTCCGTGGATACAGCAACAATCCGTCAAGCTTTAGCTCCGATTTTGCGGCGGCGATGATTAAAATGGGTGATATTAGCCCCTTGACCGGT
AGTAGCGGTGAGATCCGGAAGGTTTGCGGGAGGACCAACTGATATTTCTTTTTCCCTATTGGAATTTGACTTTTGTTAGTTGATTCGGTGAGAA
TAAGATTTGATC

>02021
AAGAGTATTGAGAAACAAGATCGAGGAAACTAATCAATGAGGACGATGAAGCGATTGAACGTGGCGGTGGCGGTTGCGGTTACAGCGACGGTTC
TTATGGGAATGTTAGGATCATCAGAGGCTCAGCTTCAAATGAATTTCTACGCGAAGAGCTGTCCAAACGCAGAGAAAATAATTTCAGATCATAT
TCAAAAGCATATCCCTAGTGGTCCTTCTCTTGCAGCTCCTCTCATCAGAATGCACTTCCATGATTGCTTCGTCAGGGTATTTAATCTCTAATCT
ATCTACATATATAGTTGCAAGTGTTTAGATATATTCGACTTTTATGTAACATATGTAGGAAATTAGTATTCACACATCCAGTTCAATAAATGAT
GGGATAGTCCAGAAGATGTACTAATGTATATTTTAAAAAAATGGTAATTTGATAGTGAACATGAGGTAGATCTAGAATATTTGATATTTATTGC
ATTTTTTAAATATTGACAATGTTTTTGAAAAAAAAAACATAATAATCTAGGGATGTGATGGATCGGTGTTGATAAATTCGACATCAGGGAACGC
AGAGAAAGATTCAGCACCAAATCTAACACTTAGAGGCTTCGGTTTCGTAGAGAGGATTAAGACTCTTCTTGAAGCAGAGTGTCCTAAGACTGTT
TCTTGCGCCGACATCATCGCACTGACCGCTAGAGACGCAGTTGTTGCCACCGTAAGTAAACAAATTATAACTTCAAGACTCAAAACATTATTTA
ATCTAATTAATCGAAATTATAATCTAATTTTTTTTAATAGGGAGGTCCTTCATGGAAAGTTCCGACAGGAAGAAGAGACGGTAGGATCTCAAAT
ACGACGGAGGCTTTGAACAACATTCCACCGCCGACGAGTAATTTCACGACGTTACAGCGACTTTTCGCTAATCAAGGCCTTAATCTCAAAGACC
TTGTTCTGCTTTCCGGTAAGTTTAGTAACCGGAAATAACCAGATTGAATTTAACAACCTAACGGTGTTTAACTTTTTGTTGTTGTTGTTGTTGT
TTAGGAGCTCACACGATCGGTGTCTCGCATTGTTCTTCCATGAATACTCGTCTCTACAACTTCTCGACGACAGTCAAACAAGATCCATCTCTGG
ATAGCGAGTACGCAGCAAATCTAAAGGCTAACAAATGTAAGAGTCTTAACGATAACACCACCATCCTCGAGATGGATCCTGGTAGTAGCAAAAC
CTTTGATCTCAGTTATTATAGGCTTGTCTTGAAGAGGAGAGGTTTGTTTCAGTCTGATTCTGCCTTAACGACAAACTCAGCTACGTTGAAGATG
ATCAACGACTTGGTCAACGGTCCTGAAAAGAAGTTTTTAAAGGCTTTCGCTAAGTCAATGGAGAAGATGGGGAGAGTTAAAGTGAAGACGGGCT
CAGCCGGTGTGATTAGGACACGTTGTTCTGTTGCCGGAAGTTAGTTAGTTTGGTCGGAAAGTGATGTTTTCTGTT

>04663
AACTAATTAGATTAAAGTATCATAAGTTCTGAATCAAAATCCGTCAAAGGAAAGATTAATCAAAGCTCTATCATTATTTGCAACAAACTGATTA
ATGGCTGCAACAAGCTCTTCTACTACTTGTGATGGTCTCTTCATCATTAGCCTTCTTGTTATCGCTTCTTCATTGTTTGGGACATCATCTGCGC
AGTTAAACGCTACGTTTTACTCCGGGACGTGCCCTAATGCCTCTGCCATCGTACGCAGCACTATCCAGCAAGCTCTTCAATCCGACCCGAGGAT
CGGAGCCAGCCTCATCCGCCTTCATTTTCACGACTGTTTTGTTAATGTATAAGTCAATAAGAAGCTCAGAAGATCTTTAAATAAGCTTCTTGGT
TTTATTTTATCATAAGGCCTTTTGATTTCTGTTCCACTAATTAACCGCCCTATTTGGCTTCTTTACTTTTGACAAGGGCTGCGACGGGTCGCTC
TTGCTTGACGACACTGGAAGTATCCAGAGCGAGAAGAACGCTCCTGCCAACGCAAACTCAGCTAGAGGATTTAATGTTGTCGACGATATCAAAA
CTGCCCTCGAGAATGCTTGTCCCGGCATTGTCTCTTGCTCTGACATTCTAGCTCTTGCCTCAGAGGCTTCCGTGTCTTTGGTAACAACACTTTC
TTCATAAACATATCAAACAAAACAGACACACATATATATTTAAACTCACACACATATATAGGGATATTAGTATGGATCAAAAGTCCACGGGGTT
CCAATCCCGGCCAAGAATCCCACCGGCTTAATTTTAAAACAAATTTATGTTAAAACGACGTCATTTTGGCCAGTTAAAAATATGAGAAAAAAAA
TAAATTACAGTTGACTAAAAATATTGCATGTGTTTAATCGCACTGATCCTGGCCCTGTGGGCCTTCCCACAGGCTTAATCGCCAATAATGAGTT
TACATAATTACATTCTTATACACAAATAAAAATATGCATGTTTGATCATCTTCAGGCAGGAGGTCCTTCATGGACTGTGTTAGTAGGAAGAAGA
GATGGTCTCACCGCAAACCTGTCCGGGGCCAATTCGTCGCTTCCCTCTCCCTTCGAAGGCCTTAACAACATCACATCTAAATTTTTAGCTGTCG
```

```
GGCTAAATACAACCGATGTAGTAGTCTTGTCTGGTAACTCATCGACATATTTAATTACTTGCGGCTCAATTCAAACAAAACCTTACCTAATGAC
ATATCTCTTGTGTATGTTAAATGTGCATAGGAGCTCATACGTTTGGGCGTGGCCAATGTGTAACCTTCAACAATAGACTTTTCAACTTCAACGG
AACAGGAAGTCCCGACCCGACTCTGAACTCAACACTTCTCAGCAGTCTTCAACAGATATGTCCTCAAAACGGCAGCGGATCAGCGATCACCAAT
CTCGATCTGACTACACCTGATGCATTTGATAGCAACTACTACACGAACCTTCAGAGTAACAATGGGCTTCTTCAGTCAGACCAAGAACTATTCT
CCAACACCGGTTCACCCACCATCGCGATTGTTAATTCCTTTGCAAGTAACCAAACCCTGTTTTTTGAGGCTTTTGCTCAGTCTATGATCAAGAT
GGGTAACATTAGTCCCCTGACTGGGACTAGTGGAGAGATTAGACAAGATTGTAAGGCGGTTAATGGACAGTCATCAGCCACTAAAGCAGAGGAC
ATTCAGATGCAATCTGACGGACCAGTGAGTTTAGCAGATATGTGAACAATAATGGGATCAG

>05508
TACAAGATGGGTTTGATTAGATCATTATGCGTATTCATAACTTTCCTCAGTTGTATCATCAGCTCGGCCCATGGCCAAGCCATCTCGATTTCTA
TCACAATTAGGATCGGGTTTTACTTGACCACGTGTCCCACAGCTGAAATCATTGTTCGAAACGCCGTGAGAGCTGGTTTCAATTCTGACCCGAG
AATCGCACCCGGAATATTGAGAATGCATTTCCACGACTGCTTCGTTCAAGGTTGTGACGGTTCAGTCCTTATATCAGGAAGTAACACCGAGAGA
ACCGCCGTTCCAAACCTCAGCCTCCGTGGATTTGAAGTCATAGAAAACGCCAAAACGCAGCTCGAAGCCRCGTGCCCAGGAGTTGTCTCTTGTG
CTGATATTTTAGCCTTAGCTGCTCGTGATACTGTAGTCCTTGTAAGCCCTAATCCATAAGCGCAATTGCATTAATACTACTTTTCTTGTTATAT
ATATATATATATATATATATATGTGAGAAACGTTTTACGTGCGTGGATTGATTTGCGTGCAGACGAGAGGGATAGGCTGGCAAGTACCAACGGG
ACGTAGAGATGGTCGAGTTTCTGTGGCCTCGAACGCTAATAATCTTCCAGGTCCCCGTGACTCCGTCGCCGTTCAACAACAGAAATTCTCCGCT
CTCGGACTCAATACCCGCGATCTCGTCGTCCTCGCCGGTACGTAGTAGTTTACACTTTACATACAATACTATAGATTACCACTAATTCGAAAAA
TTAACGTTGAGAAACTCTTCATGTTAATTTAAAAAAAAAAAAAAAAAAACTCTTCATGTGCATGTTCGTTAAACGATTATTTTTTTAAAAAATT
TAATTTTAACTGATTTCACCATTTTTTTTGTCCGCTTTAAAACTTCTTTGAAAATATAGAAAACTTTGTTAATATTAGGATGTGAATAATTACA
ACATATACTGAATTATTCATAAATTTTTAACAGGAGGACACACGCTCGGAACAGCTGGATGCGGTGTATTCAGGGACAGACTATTCAATAACAC
GGATCCTAACGTCGACCAGCCATTTTTGACGCAGCTTCAAACAAAATGTCCCCGAAACGGAGACGGTTCAGTGCGCGTGGATCTCGATACCGGA
AGCGGAACCACTTTTGATAATTCCTACTTCATCAACCTAAGTCGTGGCCGCGGAGTCCTCGAATCCGATCATGTACTTTGGACCGATCCAGCCA
CTAGACCCATCGTGCAACAGTTGATGAGTTCTAGTGGCAACTTCAACGCTGAATTTGCGAGGTCAATGGTCAAGATGAGTAATATCGGTGTGGT
TACGGGGACTAATGGGGAAATTCGTAAGGTTTGCTCTGCGATTAATTAATTAACCGATTAAAATCAGTGGTGAAACT


>06351
AATTAATCTGATTACAAGATTTAAGATAGAAAATAATAAGATGGTTAGGGCAAATTTAGTGAGCGTGATTCTGTTAATGCATGTTATTGTTGGG
TTTCCTTTTCATGCGAGGGGCTTAAGTATGACTTATTACATGATGAGCTGTCCTATGGCTGAACAAATTGTGAAAAACAGTGTTAACAATGCTC
TTCAAGCCGATCCCACTTTAGCCGCAGGTCTTATACGTATGTTGTTCCACGACTGTTTCATTGAGGTATAGTAATGTTTTTTTTTCTTTTCTTA
ATTAGTTTCCAATATTCTAATTGTGTCGTTTACGTAGAGAACCCCTTAAATTAGTTATTTTTGGTCTTATTAAGGGATGTGATGCGTCGATTCT
GCTAGATTCAACAAAAGACAACACTGCGGAAAAGGATTCTCCTGCGAATCTGAGTCTACGTGGCTACGAGATCATAGATGATGCAAAAGAGAAA
GTTGAGAATATGTGTCCAGGAGTTGTATCTTGCGCAGATATTGTTGCCATGGCTGCTAGAGATGCTGTCTTTTGGGTAATTATATAAGCTGATC
AAAATTGACATTGTGATTACATTAAGATAATCTTTTTAATTACTTAATATAATTACTAGTGATTGGTTTGTTGATTAGGCTGGTGGTCCATATT
ATGACATACCAAAAGGAAGATTTGATGGTAAAAGATCGAAGATAGAAGATACAAGAAACCTTCCTTCACCTTTTCTCAATGCCTCTCAACTCAT
TCAAACCTTTGGCAACCGTGGCTTCTCTCCGCAAGATGTTGTTGCTCTCTCTGGTGAGTTCTTTACTACATACGTACTTGGATTCCGTATATGT
CGATATTTTTGTATATCTACTAGGTCTTATAGCTCAATGAAGAAAAATATTGGAACATTAACAAGAAGAAAATTAATTTATTATGAATACGTTT
AAAATTAAATTTCTAAATTTTAGAACTTAATTATATGTTTTGCATGACCATCAGATGTTTTTAAAAAAAAATTGTATATATACAAATAACGTAC
CACAAATTAATTTTAGTAGCACCATTTTTAAAAAATTTTCCAAAATACCACAAATGTTAGGCTTTTTCGTAATTTTAAAATTGTGGTGTTTTTG
GAATTCTAGAATTTTGGTGGTAACCTAGAAAATAAAACATCAAATATGTATATTTTCGTAGTTTCAAAATTGATATATTGTTTTAACTTTTATT
AATTCAAAATTTGTTGATTAATTTTTTTTCTATATAATTATTTCACTGGCCAAATTAAAAAGTAGTTTTGGAGAACAAAATTTGTTTATGTTTC
TCATTTGTACCTTCTTTTTAGTAATATAAGATCTTATTAATTTCTATAATCATATAAAATTATTATTATTATTCTATATGACATTTTTCACATC
ACTTGGGATGAAGTCCATATAAAATGTTTAGCATGACAGTGAATGCCCCACAAGATGTTAATTTTGGTTGTAACTTAACTTGTGGAATATATTA
TAGGAGCACATACCCTTGGAGTTGCACGATGCTCCTCCTTCAAGGCTAGACTTACCACTCCAGATTCTTCACTGGACTCCACTTTTGCAAACAC
TCTCACTAGAACTTGCAATGCGGGGGACAATGCAGAGCAACCCTTTGATGCGACCCGCAACGATTTCGACAATGCCTACTTCAATGCGCTTCAG
AGGAAATCAGGAGTCCTCTTTTCAGACCAGACCTTATTCAACACTCCAAGGACCAGGAATCTTGTTAATGGTTATGCCCTTAATCAAGCTAAGT
TTTTCTTTGATTTCCAACAGGCCATGCGCAAAATGAGCAATCTTGATGTTAAACTTGGCTCTCAAGGTGAAATACGTCAAAATTGCCGGACTAT
TAACTAAGCCTAGGCCGATTTTTGTATTTATGCTCCCACCTTTAATTATACTTACCTACTCTGTCATTAATTCGAGTCATAATGTCCTATGCTA
CCATGTAAAATTAGTGTGCCTAATGTGATATGCAGCTGTATTGTACTTATTGTTTGTGGGTTTCAGATGTCCATCATCAAAACGTAATATATAT
ACTTGGTGATCTTG

>23190
CCCTATAGTGAGTCGTATTACGGCCGGGGGAACAACAAGAAGCAGAGAAGAGAGAGGCTTCGACTGAAACAACAAAAAATGGCAATGAGTTATT
CGATACGTGTCCTGACGTTTCTGATGTTGATCTCGTTAATGGCAGTGACACTGAACCTTCTGTCAACGGCGGAAGCAAAGAAGCCGAGGAGAGA
TGTTCCTATAGTGAAAGGTCTCTCTTGGAACTTTTACCAGAGAGCATGTCCGAAAGTGGAAAAGATTATCAAAAAAGAACTCAAAAAAGTCTTC
AAGAGAGATATTGGTTTAGCCGCAGCCATCCTTCGTATACATTTCCATGACTGCTTCGTTCAGGTTCTATCTTTTTCGTTCCCTAATTTTTTGT
ATTAAAACCTAATTAAGAACATTTAATTTTATGTTCTTTAAGCCTACTTTAAACGGTTCTACGTATGTGATTAACATAATTTTCACTACTTGAA
AACTCCTTTTCTTTTAGTATTAGAAACCATATCCACTCATTAGTCATTACTCATCATCATACCCACACAAATAAAATAATAGGATTCGCGGTGT
GTGTCCGGAATGCATGAAAATGAAATTCGTAAAATCAATAAACAAGTTTGACTTTTATATTTAAAATCCGACCAAAATAGCCTAAAGAAAAAA
CACAACTCGAAATTCTCACGAACAAGAAATTAGAGTATTGTCCGCTTCTTACTTTCACTTCTTTTCAAACAAAAGATTTTTTAGTGTAAACAGT
AGTTAAGTGACCAAGATATTATTGATCGAAGAGTATTGTACTATTGTTTAATACTACAGTAGTTGTACATGCGTTTTAGAATAGACTCGAATAT
AGGACATGTCTCATTAAAACTATTTACCCACTTTCCGTGGATTAGTTGGGCTTCACTTATGGAACGTAAATATCTTACATAAATGAATATACAT
CGAACCAATGGAAGAAATAATAAATATAAGGATATGATAGTATATACAGGTTTATATCTAGATATTTTTTTTAATCAGGAGATGATCCATAGTT
TAAAATCCGAATCATATGATAAATATATAGCAGTCGAGTAACCCATTCACCAGATATCTATGTAGTTGGATAAAATATCTATGTCTATGGTTTT
TCGTTCCAGATTATATATACTTATAAAATCAACAATATAGACTTTTCTAACTAAATACATATAAAACAAAACAATTACATACTAATCAGATATT
TATAGTCTTACTTGCAAGTTGCAACACCTTTTACTGAAGTGTGGCCTATTTTGATCAGATAAATAGTTTCGGAGAACTAAAAGCTTTTATTATG
AGAGTTCATAAAATAAATAATTCTTTTTTATTATTTTAAAAAAACAGGGGTGTGAAGCATCTGTGCTGCTAGCTGGATCAGCAAGTGGACCAGG
AGAACAATCATCKATCCCGAACCTAACACTCCGTCAACAAGCCTTTGTTGTCATCAATAACCTGCGTGCCCTCGTCCAGAAACAGTGTGGYCAA
GTCGTCTCTTGCTCCGACATCCTCGCTCTCGCCGCTCGCGATTCCATCGTCCTTGTAACCAACTATCTCTCGTCTATAAATCAATCACACAATA
ATGGGTTTAAATATGGATTATAACACGGCAGTGACACTAATATGAACCATTAGCTTATACAACATCGGTTTGTGTCGATATATGCCTTTTTCTA
ATTAATAAAAAAATGTGTCAATGTAGTCAGGAGGGCCAGACTATGCTGTGCCACTTGGCCGACGTGACTCGCTAGCGTTTGCGACCCCGGAAAC
GACGTTAGCTAACTTACCGCCACCGTTTGCCAACGCAAGCCAGCTCATCAGCGACTTCAACGACAGAAACCTCAACATCACCGACTTAGTAGCA
CTTTCCGGTGGTCACACCATCGGAATTGCGCATTGTCCGYCTTTCACMGACCGGCTCTACCCAAACCAAGATCCAACCATGAACAAGTCTTTCG
CCAACAGCCTCAAACGCACCTGTCCCACGGCGAACTCGAGCAACACGCAAGTGAATGACATAAGGAGTCCTGACGTGTTTGACAACAAGTACTA
TGTTGATCTCATGAACCGACAAGGGCTGTTCACTTCCGACCAGGATCTGTTCGTTGACAAGAGGACACGTGGCATAGTGGAAAGCTTTGCGATC
GACCAGAACTTGTTTTTTGATCATTTCACGGTGGCAATGATTAAGATGGGTCAGATGAGTGTCTTGACGGGGACACAAGGGGAGATCCGTTCCA
ACTGTTCAGCCAGAAACACCGCAAGTTTCATATCCGTTTTGGWAGAAGGCATAGTCGAGGAAGCTCTTTCCATGATCTAAAATAACCATAAATC
```

```
TCAGACTTTTCTTTTCTTTAACTTTGTTTTTATTTAGTTGTCGCACTTGTGGTTTGTGGAATGCCTAAGACTATCTCATAAATAAGAGCATTGC
TTTCATCTTAATTTCTTCTTCTTTTTTTTTTCTCTAGTTTGGTCAATGTCTGGGACTATAATGAATAAAGAATGTTGCTACATCTT


>22489(non-verified isotig)
ACTACAACTAAAAACACACTTGATCTTCTCTAAAACACTAAAATTATATATCCAATATGGAGTTTGTTAGATCATTATGCGTATTCATAACTTT
CCTCGGTTGTCTCATCAGCTCGGCCCATGGCCAAGCCGCCGCAAGGCGACCTGGTCCGATTTCTGGCACAAGGATCGGGTTTTACTTGACCACG
TGTCCCACCGCTGAAATCATTGTCCGAAACGCCGTGAGAGCTGGTTTCAATTCTGACCCAAGAATCGCACCCGGAATATTGAGAATGCATTTCC
ACGACTGCTTCGTTCTAGGTTGTGACGGTTCAGTCCTTATATCAGGAAGTAACACTGAGAGAACCGCCGTTCCGAACCTCAACCTCCGTGGATT
TGAAGTCATAGACAACGCCAAAACGCAGCTCGAAGCCACATGCCCAGGAGTTGTCTCTTGTGCTGATATTTTAGCCTTAGCTGCTCGTGATACT
GTAGTCCTTACGAGAGGGTTAGGCTGGCAAGTACCAACGGGACGTAGAGATGGTCGAGTTTCTGTGGCCTCGAACGCTAATAATCTTCCAGGTC
CCCGTGACTCCGTCGCCGTTCAACAACAGAAATTCTCCGCTGTCGGACTCAATACCCGCGATCTTGTCGTCCTCGCCGGAGGACACACRATCGG
AACAGCTGGATGCGGTGTTTTCAGGGACAGGCTATTCAATAACACGGATCCTAACGTCAACCAGCTATTTTTGACGCAGCTTCAAACACAATGT
CCCCAAAACGGAGACGGTKCAGTGCGCGTGGATCTCGACACCGGAAGTGGAACCACTTTTGACAATTCCTACTTCATCAACCTAAGCCGTGGCC
GCGGAGTCCTCGAATCCGACCATGTACTTTGGACCGATCCAGCCACTAGACCGATCGTGCAACAGTTGATGAGTCCTAGAGGCAACTTCAACGC
TGAATTTGCGAGGTCAATGGTCAGGATGAGTAATATCGGTGTGGTTACGGGGGCTAATGGGGAAATTCGTAGGGTTTGCTCTGCGGTTAATTAA
TTAACCGATTAAAATCAGTGGTGATAAACAGA


>04791_03523
TCTCACTTTCTCTCTTCCGCTACAACAATGGCGGAACTCAAATCTCTCTCCCTCATCCTCCTCTTCACACTCCTCACCACCACCATCGAATCTC
GTTTAACCACAAACTTCTACTCAAAATCATGTCCAAGATTCTTCGACATAGTCAGAGATACAATCTCAAACAAACAAATCACAACACCAACCAC
GGCAGCCGCCACAATCCGTCTCTTCTTCCACGACTGTTTCCCCAACGGCTGCGACGCCTCAATCCTAATCTCCTCAACTGCCTTCAACACCGCA
GAACGTGACTCATCAATCAATCTCTCACTTCCCGGCGACGGCTTTGACGTCATAGTCCGAGCTAAAACCGCAATCGAACTCGCTTGTCCCAACA
CTGTTTCTTGCTCCGATATAATCACCGTCGCTACTCGTGACCTTCTTGTCACCGTCGGTGGTCCTTACTACGACGTTTACCTCGGCCGTCGTGA
TTCAAGAATATCTAAATCATCTCTTTTAACCGATCTTCTTCCTCTTCCTTCATCTCCGATCTCAAAAACCATTCGTCAGTTTGAATCTAAAGGT
TTCACTATTCAAGAAATGGTTGCTCTTAGTGGGGCCCACTCAATCGGGTTTTCACATTGTAAAGAGTTTGTTAATCGGGTCGCCGGTAATAATA
CCGGGTATAACCCGAGATTTGCTCAGGCGTTGAAGCAAGCTTGTTCTAATTACCCGAAAGATCCGACGTTATCTGTGTTTAATGATATTATGAC
TCCGAATAGGTTTGATAATATGTATTATCAGAATATTCCAAAGGGTCTTGGGTTACTTGAATCGGATCATGGGTTATATTCTGACCCGAGAACC
CGACCTTTTGTTGATCTTTATGCTAGAGATCAAGATTTGTTCTTTAAAGATTTTGCTAGAGCTATGCAGAAGTTGAGTCTCTTTGGTGTTAAGA
CTGGTCGACGAGGAGAGATCCGACGAAGGTGCGATGCGATTAACTGAGTTTATGTTTTTTCTTTTATGTATTTATTTTTTGGATTATATTTGGG
GAAGAAATGTGAGATTTGATGAATTGTGAATCTTCTGAGTTTTTTTTTTTT


>06117
CTTCTTCTTTCTTCTGGATCTAGTGGAAACTGACATTATGGCAAGAATTGGAAGCTTTCTCGTTGTYATCTCTCTCGCTTGCGTTCTTACTCTC
TGCATCTGCGACGACGAGAGTAATTATGGCGGCCAAGGGAAACTCTTCCCAGGTTTCTACAGCAGCTCGTGCCCTAAAGCTGAGGAGATCGTGA
GGTCTGTTGTAGCCAAAGCTGTTGCAAGAGAGACTCGTATGGCTGCTTCTCTCATGAGGCTCCATTTTCACGACTGTTTTGTTCAGGTACTCAA
GAGCTATTTTAAGTAACTTTCTCATCAAGCTAACGACTAATTTTGAGAATTTTAATTAATTATAGCTGTTTGCATAAAAATGTGTAGGGTTGTG
ATGGATCGTTGCTTCTAGACAGCAGTGGAAGTATAGTTACTGAGAAGAACTCTAACCCTAACAGCAGATCAGCTCGTGGGTTTGAAGTTGTTGA
CGAGATCAAAGCTGCATTGGAGAATGAATGCCCTAACACTGTTTCTTGCGCTGACGCCCTAACTCTAGCCGCTAGAGACTCCTCTGTTCTTGTA
AGTCGCCTTATTCCACTTTCTCTCTTTACCGCTTCCTTAAGTTCTCAAGAGATCTAATTTTGTTCGGTTTGAATATAGACTGGTGGACCAAGCT
GGATGGTTCCTTTGGGAAGAAGAGATTCGACAAGTGCAAGCTTGAGTGGATCAAATAACAACATTCCTGCACCAAACAACACTTTCAACACAAT
TCTCTCGAGATTTAACAGCCAAGGTCTCGATCTCACCAATGTCGTTGCTCTCTCCGGTAAGCTTACTTAAAACACAAGGAAAATTTTACTTTCC
TTGCTACTCAAGTTAAAGTTAAATCCTAAAAAGGATTTTAGTATAACCTGACTCAATTGTTTCTCAGGGAGCCACACAATTGGATTCTCAAGAT
GTACTAGTTTTAGACAGAGACTTTACAACCAATCCGGAAACGGAAGTCCCGACACAACCTTAGAGCAATCCTACGCTGCTAACTTGCGCCATCG
GTGCCCTAGATCAGGTGGGGACCAGAACCTGTCGGAGCTTGACATCAACAGTGCTGGAAGGTTTGATAACAGCTACTTTAAGAATTTGATYGAG
AACATGGGACTGTTGAATTCCGACCAGGTCTTGTTCTCTAGCAACGACGAATCGAGAGAGCTAGTGAAGAAGTATGCAGAGGATCAAGAAGAGT
TCTTCGAGCAGTTCGCGGAATCGATGGTCAAGATGGGGAATATCTCTCCCTTGACTGGTTCAAGTGGTCAAATCAGGAAGAATTGCAGGAAGAT
TAACTCTTGATTTCATAATATTGAATTGGGCGAAATAAAAATGAGAGATGTTTTGGG


>17517
CAACAACAACTTTTACAAAGCTCAAAGAGTTTCTTATTTACCAAAACAAAAACAAAATGGGTCGTGGTTATAATTTACTATTAATTCTAGTTAC
GTTTTTAGTATTGGTCGCAGCTGTAACCGCTCGAAGACCACGAGTTGGGTTTTATGGGAATAGATGCCGAAAGGTAGAGTCTATCGTGAGATCG
GTGGTTCGATCTCATTTCCGGTGTAATCCGGCAAATGCACCGGGAATTTTGCGTATGYATTTTCACGATTGTTTTGTCAATGGCTGCGATGGCT
CGATCCTCCTCGCTGGTAACACTTCGGAGAGAACTGCGGGTCCTAACCGTTCATTGAGAGGGTTCGAAGCTATTGAAGAAGCTAAGACTCGGCT
TGAGAATGCTTGTCCTAATACCGTTTCTTGTGCGGATATCCTCACCCTTGCTGCTCGAGACGCCGTCGTTTGGGTAAAACATATTGAAATTAGT
TTCAGTTTTAAGAAATTTTAATTATATATTGTGTGTTTAATGGTAATATAATTTTGGGTGATGCAGACCGGTGGAAAAGGTTGGTCGGTGCCAT
TGGGACGTCTTGACGGCCGAAGATCAGAAGCCTCAGATGTAAATTTGCCCGGACCAAGCGACCCCGTCGCTAAGCAGAAGCAAGACTTTGCAGC
TAAAAATCTCAACACCTTGGACCTCGTAACTCTTGTTGGTTCGTTGTAATTATATATAATCAAACAATGTATATTAATCAAATAGTTACATATT
TGGTCGGTTGTATTAAATTTTCGATACAGTTCTGATTACGTCGGTATGTATGAGGTTATGAAATATTTTTTTTCTTGATTTAAGTTTAGAAATT
TAGTTTCAATTTGAACCGAATTCGGTTTACATGTTCAATTTTATTAGCTAATATTGCAAATTGATATTGATCAGGTGGACACACAATTGGAACT
GCTGGTTGCGGTTTGGTAAGAGGCAGATTCTTTAACTTCAATGGCACGGGACAACCTGACCCATCAATCGACCCGAGTTTCGTTCCTCTAGTTC
AGGCTCGTTGCCCTCAAAACGGTAACGCAACGACCCGAGTCGACTTAGACACTGGAAGTGCAGGTGATTTCGATACATCGTACCTAAGTAACGT
GAGGTCAAGCCGCGTGGTTCTCCAATCCGATCTAGTCCTGTGGAAGGACACCGAAACCAGAGCCATCATAGAACGTTTATTAGGYTTACGCCGG
CCCGTTTTGAGGTTCGGATCAGAATTTGGGAAGTCGATGACCAAGATGAGTCTCATAGAAGTTAARACCGATGGGGAGATTCGTAGGGTTTGCT
CTGCGATCAATTAAGTATTAAAAACACACAAATGTTTGGTTTGATTTTATCACTTATTTTATGGAATAAGCTTGCTATAG

>08562.1
GTAATGGCAAGACTCACTAGCATTCTCCTTCTTCTTTCTCTTTTATGCTTTTTCCCTCTCTGTCTCTGCGACAAGAGCTATGGAGGCAAACTCT
TCCCTGGTTTTTACGCCCACTCATGCCCACAAGCCGGGGAAATCGTGAGATCAGTCGTAGCTAAAGCTGTTGCTAGAGAGACCCGTATGGCTGC
TTCTTTGATGAGACTTCATTTCCACGACTGTTTCGTTCAGGTTTGGTTAATTTCTTCTACGCCCACTATTCTAAAGATTTTTTTATTGAGCAAG
GTAACTGTGAAATGCAGGGTTGTGATGGCTCTTTGCTTCTAGACAGCAGTGGGAAAATAGTGAGTGAGAAAGGCTCAAACCCTAACAGCAGATC
AGCTCGTGGGTTCGACGTAGTTGACCAAATCAAAGCTGAATTGGAGAAACAATGCCCTGGAACTGTTTCTTGCGCTGATGCTCTAACTCTAGCC
GCTAGAGACTCCTCTGTTCTTGTAAGTCCCCTCCATAGTTTCCAAATCAAATTTAAAACATCAGCTAACTCGGTGTGGTTTTTGTTTTAGACCG
GTGGACCGAGCTGGGTGGTTTCATTAGGAAGAAGAGATTCAAGAAGTGCAAGCTTGAGTGGTTCGAACAACAATATCCCTGCACCAAACAACAC
TTTCCAGACCATTCTATCGAAGTTTAACCGTCAAGGACTCGATGTCACCGACCTTGTTGCTCTCTCCGGTAAGCTTTCTTCACTTGCACGCAAC
```

```
ACAGTTAAAAGAAACCCCATTGCCTTACTTTTTTCTCAACCCACCACACTTCTTAACTGTTTCTCAGGGAGCCACACCATTGGATTCTCGAGAT
GCACGAGTTTCAGACAAAGATTGTACAACCAGTCCGGAAACGGACGTCCAGACATGACATTGGAACAATCCTTCGCTGCTAACTTGCGCCAAAG
GTGTCCGAGATCCGGCGGAGACCAGATTCTCTCAGTGTTGGACATCATCAGCGCCGCGAAATTCGACAACAGCTACTTCAAGAACTTGATAGAA
AACAAGGGTTTGTTGAACTCGGACCAGGTTTTGTTCAGCAGTAATGAGAAATCTAGAGAGCTTGTGAAGAAGTATGCAGAGGACCAAGGAGAGT
TTTTTGAGCAGTTTGCGGAATCGATGATCAAGATGGGAAATATATCTCCCTTGACGGGTTCGAGTGGCGAAATCAGAAAGAATTGCAGGAAGAT
AAACTCTTGAATTCTTGAAATGAGGAAAGTATTGGG

>08562.4
GTAATGGCAAGACTCACTAGCATTCTCCTTCTTCTTTCTCTTCTATGCTTTTTCCCTCTCTGTCTCTGCGACAAGAGCTATGGAGGCAAACTCT
TCCCTGGTTTTTACGCCCACTCATGCCCACAAGCCGGGGAAATCGTGAGATCAGTCGTAGCTAAAGCTGTTGCTAGAGAGACCCGTATGGCTGC
TTCTTTGATGAGACTTCATTTCCACGACTGTTTCGTTCAGGTTTGGTTAATTTCTTCTACGCCCACTATTCTACAGATTTTTTTATTGAGCAAG
GTAACTGTGAAATGCAGGGTTGTGATGGCTCTTTGCTTCTAGACAGCAGTGGGAGAATAGTGAGTGAGAAAGGCTCAAACCCTAACAGCAGATC
AGCTCGTGGGTTCGACGTAGTTGACCAAATCAAAGCTGAATTGGAGAAACAATGCCCTGGAACTGTTTCTTGCGCTGATGCTCTAACTCTAGCC
GCTAGAGACTCCTCTGTTCTTGTAAGTCCCCTCCATAGTTTCCAAATCAAATTTAAAACATCAGCTAACTCGGTGTGGTTTTTGTTTTAGACCG
GTGGACCGAGCTGGGTGGTTTCATTAGGAAGAAGAGATTCAAGAAGTGCAAGCTTGAGTGGTTCGAACAACAATATCCCTGCACCAAACAACAC
TTTCCAGACCATTCTATCGAAGTTTAACCGTCAAGGACTCGATGTCACCGACCTTGTTGCTCTCTCCGGTAAGCTTTCTTCACTTGCACGCAAC
ACAGTTAAAAGAAACCCCATTGCTTAACTTCTTTCTCAAACCGCTACACTTCTTCACTGTTTCCCAGGGAGCCACACCATTGGATTCTCCAGAT
GCACGAGTTTCAGACAAAGGTTGTACAACCAGTCCGGAAACGGACGTCCAGACATGACACTGGAACAATCCTTCGCTGCTAACTTGCGCCAAAG
GTGTCCGAGATCCGGCGGGGACCAGATTCTCTCGGTGCTGGACATCATCAGCGCCGCGAAATTCGACAACAGCTACTTCAAGAACTTGATAGAG
AACAAGGGTTTGTTGAACTCGGACCAGGTTCTGTTCAACAGTAACGAGAAATCTAGAGAGCTTGTGAAGAAGTATGCAGAGGACCAAGGAGAGT
TTTTTGAACAGTTTGCGGAATCAATGATCAAGATGGGAAATATATCTCCCTTGACGGGTTCGAGTGGCGAAATCAGAAAGAATTGCAGGAAGAT
AAACTCTTGAATTCTTGAAATGAGGAAAGTATTGGG
```

CURRICULUM VITAE

PERSONAL DATA

Name:	Laura Hannele Näätsaari
Date of Birth:	November 6th, 1980
Place of Birth:	Tornio, Finland
Nationality:	Finland

EDUCATION

1987-1996:	Elementary school in Kemi, Finland
1996-1998:	Kemin Lyseon Lukio High School in Kemi, Finland
1998-1999:	Tranquillity Union High School in San Joaquin, CA, USA
1999-2000:	Kemin Lyseon Lukio High School in Kemi, Finland
2000-2007:	University of Helsinki, Master’s degree in Biosciences • Major: genetics • Minors: general biology, chemistry, biochemistry
2005:	Subject teacher education and practical training for upper elementary school and high school in Viikki teacher training school, Helsinki, Finland
2008-2012:	Graz University of Technology, PhD studies

RESEARCH AND TEACHING EXPERIENCE

Graz University of Technology, Institute of Molecular Biotechnology 18.9.2007-

- Project work and PhD thesis in the research group of Anton Glieder
- Teaching in courses CHE.177 LU aus Biotechnologie (2009) and MOL.912 Molekulare Biotechnologie (2009 and 2010)
- Research stay in the group of Lloyd Ruddock in the Institute of Biochemistry, University of Oulu 2/2011-8/2011

University of Helsinki, Department of Medical Genetics, 1.1.2006-31.12.2006

- Research group of Lauri A. Aaltonen (Cancer genetics)
- Master’s thesis: Modifier-effect in HLRCC tumorigenesis

Viikin normaalikoulu (Viikki High School), 2005, 2006

- Practical training in teacher’s education, short temporary posts

Suomen Ympäristökeskus (Finnish Environmental Center) 1.7.2004-31.8.2004

- Practical laboratory work connected with genetic diversity

University of Helsinki, Department of Medical Genetics, 03.06.2002-31.12.2002

- Research group of Lauri A. Aaltonen (Cancer genetics, Finnish Academy Top Unit)

PUBLICATIONS

PAPERS

Alhopuro P, Katajisto P, Lehtonen R, Ylisaukko-Oja SK, Näätsaari L, Karhu A, Westerman AM, Wilson JH, de Rooij FW, Vogel T, Moeslein G, Tomlinson IP, Aaltonen LA, Mäkelä TP, Launonen V. Br J Cancer. *Mutation analysis of three genes encoding novel LKB1-interacting proteins, BRG1, STRADalpha, and MO25alpha, in Peutz-Jeghers syndrome*. 2005 Mar 28;92(6):1126-9.

Vahteristo P, Koski TA, Näätsaari L, Kiuru M, Karhu A, Herva R, Sallinen SL, Vierimaa O, Björck E, Richard S, Gardie B, Bessis D, Van Glabeke E, Blanco I, Houlston R, Senter L, Hietala M, Aittomäki K, Aaltonen LA, Launonen V, Lehtonen R. *No evidence for a genetic modifier for renal cell cancer risk in HLRCC syndrome*. Fam Cancer. 2010 Jun;9(2):245-51.

Näätsaari L, Mistlberger B, Ruth C, Hajek T, Hartner F, Glieder A (2012).
Deletion of the Pichia pastoris KU70 Homologue Facilitates Platform Strain Generation for Protein Expression and Synthetic Biology, PLoS ONE, accepted with revision

Näätsaari L, Krainer FW, Schubert M, Thallinger GG, Glieder A (2012). *Peroxidase gene discovery from the horseradish transcriptome,* manuscript prepared for submission in NAR

Näätsaari L, Kulterer M, Nothdurft P, Ribitsch V, Glieder A (2012). *Horseradish peroxidase surface variants for oriented enzyme immobilization,* manuscript in preparation for submission in “Protein Expression and Purification”

PATENTS

UG001/EP „Horseradish Peroxidase Isoenzymes“: The present invention relates to recombinant heme-containing horseradish peroxidase isoenzymes with improved properties.

POSTERS AND PRESENTATIONS:

12/2010 Biocatalysis Conference 2010, Cancun, Mexico
06/2010 OxiZymes 2010 and the 9th International Symposium on Peroxidases, Leipzig, Germany
06/2010 Research 2010, Graz, Austria
10/2009 *Pichia* 2009 protein expression conference, Tucson, USA
01/2009 PepTalk 2009, San Diego, USA
05/ 2008 BioTech 2008 and 4th Swiss-Czech Symposium, Wädenswil, Switzerland