



## OPEN Meta-analysis of RNA-Seq datasets allows a better understanding of *P. tricornutum* cellular biology, a requirement to improve the production of Biologics

Isabelle Boulogne<sup>1,3</sup>✉, Charlotte Toustou<sup>1</sup> & Muriel Bardor<sup>1,2</sup>✉

The marine diatom *Phaeodactylum tricornutum* is currently used for various industrial applications, including the pharmaceutical industry as a cost-effective cell biofactory to produce Biologics. Recent studies demonstrated that *P. tricornutum* can produce functional monoclonal antibodies, such application is currently limited by the production yield that hinders industrialization. Therefore, it is necessary to understand and control the cell biology of *P. tricornutum* to improve the Biologics production yield. Transcriptomic analyses have recently been used by the pharmaceutical industry to improve the production of Biologics in mammalian cells, especially Biologics titer and cell productivity. Hence, in the present work, we performed a meta-analysis of seven publicly available RNA-Seq datasets from different strains of *P. tricornutum*, for which the culture conditions were chosen as similar as possible. We analyzed the differential expression of genes that are involved in biological processes that are well known to potentially impact the bioproduction and critical quality attributes of Biologics. Therefore, the expression of genes involved in the *N*-glycan biosynthesis, protein export and secretion, protein quality control and proteasome, as well as those encoding proteases were analyzed and compared. The results pave the way towards optimizing Biologics production in *P. tricornutum* and highlight that the Pt4, Pt3 Ov and Pt8 strains seem to be the most promising *P. tricornutum* strains.

**Keywords** Meta-analysis, *P. tricornutum*, Bioproduction, Biologics, Glycosylation

The marine diatom *Phaeodactylum tricornutum* is currently used for various industrial applications such as food, nutraceuticals, cosmetics, biotechnology and bioproduction. In this context, the capacity of *P. tricornutum* to be a cost-effective cell biofactory for the production of high value-added molecules has been highlighted<sup>1</sup>. For example, the development of genome-editing tools in *P. tricornutum* enabled the production of bioplastics<sup>2</sup> and terpenoids<sup>3,4</sup>. Furthermore, *P. tricornutum* has also been shown to be a reliable and more sustainable means of producing Biologics. Biologics are used to treat life-threatening diseases, and include gene or cell therapies as well as recombinant proteins. Due to the complexity of these recombinant proteins, their production relies on the use of living systems such as mammalian cell systems. However, alternatives such as *P. tricornutum* are currently being explored. For example, the production and secretion of monoclonal antibodies (mAbs) directed against two deadly viruses (Hepatitis B and Marburg) have been demonstrated<sup>5–7</sup>. More recently, the successful production of the RBD (receptor-binding domain) of the SARS-CoV-2 in *P. tricornutum* allows this microalgae to be considered as a viable platform for the production of sensitive diagnostic tools<sup>8</sup>. In addition, most recombinant proteins of interest to human health are glycosylated. Unlike *Chlamydomonas reinhardtii*, another microalgae used as a cell biofactory, for which the *N*-glycosylation of the recombinant erythropoietin has been shown to be very different from the human *N*-glycosylation<sup>9</sup>, *P. tricornutum* is able to synthesize *N*-glycan structures that are similar to the human ones<sup>10,11</sup>. Currently, the development of the production of these

<sup>1</sup>University of Rouen Normandie, UNIROUEN, UFR des Sciences et Techniques, GlycoMEV UR4358, Innovation Chimie Carnot, Fédération de Recherche Normandie-Végétal FED 4277, 76821 Mont-Saint-Aignan, France.

<sup>2</sup>Alga Biologics, CURIB, 25 Rue Tesnières, 76821 Mont Saint Aignan Cedex, France. <sup>3</sup>Present address: ECOTERCA - ÉCOlogie TERrestre CARibéenne, Université des Antilles, Faculté des Sciences Exactes et Naturelles, Campus de Fouillole, BP 592, 97159 Pointe-à-Pitre Cedex, France. ✉email: isabelle.boulogne@univ-rouen.fr; isabelle.boulogne@univ-antilles.fr; muriel.bardor@univ-rouen.fr

high value-added molecules at an industrial scale remains restrained by their limited production yield. However, since the approval of the first recombinant protein produced in Chinese Hamster Ovary (CHO) cells, these cells have benefited from more than four decades of active international research<sup>12</sup>. This includes CHO cell engineering, vector design and control of culture media<sup>13–16</sup>. Moreover, although numerous cell lines of CHO cells have been used as models since the 1950s, the genomes of CHO cell lines have only been available 13 years ago. The sequencing of the CHO-K1 cell line genome and its publication in 2011<sup>17</sup> has allowed an increased use of “omics” techniques and more accurate annotation of CHO omics data<sup>18,19</sup>. The first applications of omics tools to CHO cell lines were mainly focused on a fundamental understanding of the cell biology. For example, initial transcriptomic works were focused on understanding the regulation of growth-phase-dependent genes, central sugar metabolism, and the *N*-glycosylation pathway<sup>18–20</sup>. Transcriptomic analyses are now being used directly for industrial purposes to identify genes involved in cellular pathways that could be used to optimize mAb titer and quality as well as cell productivity<sup>21–23</sup>.

The first fully assembled genome of *P. tricornutum* (Pt1.8.6 strain) is available since 2008<sup>24</sup>, allowing the generation of numerous genomic<sup>25,26</sup>, transcriptomic<sup>27–31</sup> and even proteomic data<sup>30,32–35</sup>. Since 2020, genomes of several other ecotypes of *P. tricornutum* have been sequenced<sup>36,37</sup>. Most of the *P. tricornutum*'s omics data focus on the effect of environmental conditions, such as nutrient element depletion/starvation (nitrogen, phosphorus) or light/dark stress on gene expression for a single strain<sup>27,30,33,35</sup>. Most of these data are now available in the Diatomics database (<https://www.diatomicsbase.bio.ens.psl.eu/>) and the PhaeoEpiView database (<https://phaeoeview.univ-nantes.fr/>). Unlike other diatoms, *P. tricornutum* has the unique ability to be pleiomorphic, as evidenced by the three main morphotypes observed for a single strain such as Pt3 (fusiform, oval, or triradiate)<sup>29,38,39</sup>. The fusiform morphotype is the most common in the natural environment. Indeed, out of the ten most studied ecotypes, seven of them are predominantly fusiform (>95% of cells). However, some ecotypes present a natural predominance for other morphotypes, such as Pt8, which was originally described as mainly triradiate<sup>38</sup>. However, this morphotype was reported to be unstable, and was not always observed in this ecotype<sup>31,37</sup>. Additionally, depending on the culture conditions and often under stress, other predominances are reported<sup>38,39</sup>. For example, the Pt3 strain, which is derived from Pt2, was selected for its ability to grow in freshwater media under an oval morphotype. Thus, the appearance of the oval morphotype demonstrated the ability of *P. tricornutum* to adapt to stress conditions (salinity, temperature, light). Such ability varies from one ecotype to another<sup>38,39</sup>. Transcriptomic and proteomic analyses of the three morphotypes of the Pt3 strain highlighted that the oval morphotype is able to secrete proteins faster, and in larger amounts as compared to the other two morphotypes<sup>29,32</sup>. Such observation was confirmed by microscopic analyses and protein quantification<sup>40,41</sup>. In addition, a comparative proteomic analysis of the proteins secreted within the culture medium, called secretome, of the three morphotypes, revealed that more proteins secreted in the culture media of the Pt3 oval morphotype were involved in proteolysis<sup>32</sup>. Besides the phenotypic and genotypic differences, variations in biological processes have also been noticed between the ecotypes, such as a higher capacity of Pt4 to assimilate nitrate than the reference strain Pt1<sup>42</sup>, probably due to a higher copy number of the gene encoding for the nitrate reductase<sup>36</sup>.

In this context, we performed a meta-analysis of transcriptomic data retrieved from the literature. Among the dozens of articles related to RNA-Seq analyses, those concerning different strains of *P. tricornutum*, for which the culture conditions were chosen as similar as possible were selected (Table 1). We especially focused on biological processes that are well-known to significantly impact the bioproduction and the critical quality attributes of Biologics. Thus, in agreement, the expression of genes involved in the *N*-glycan biosynthesis, protein export and secretion, protein quality control and proteasome as well as those encoding proteases were analyzed and compared. The results provide first elements to consider when selecting an optimal *P. tricornutum* strain for the production of Biologics.

## Results and discussion

Based on our search criteria, which included information on culture conditions, ecotypes, and morphotypes, seven publicly available RNA-Seq datasets were selected from the literature and used in the present work (Table 1). Unfortunately, we were unable to find RNA-seq data with culture conditions as similar as possible for the 10 *P. tricornutum* ecotypes (Pt1 to Pt10). The selected datasets include the RNA-Seq dataset of the strain Pt1.8.6, for which the genome has been sequenced and annotated. As this is the most widely used and studied strain, we selected it as the reference strain<sup>24,36</sup>. This reference strain was reported to be 95–100% fusiform. Other datasets concerning the Pt3, Pt4 and Pt8 strains were also included as they were cultured under similar conditions. Pt4 was described as mainly fusiform (95–100%) as reported for Pt1 whereas Pt3 and Pt8 were predominantly oval (60–75%) and triradiate (80–85%) morphotypes, respectively. Another RNA-seq study regarding the Pt3 strain, which was specifically enriched in the different morphotypes (Pt3 Fus, Pt3 Ov, Pt3 Tr), was also included in the study for comparison. For all the strains, RNA extractions were performed on 6 to 8 days of culture under similar conditions, using seawater supplemented with either Conway or *f*/2, which are very close in terms of composition. The temperature used for the cultivation was comprised between 18 and 20°C.

### Differentially expressed genes

The differential expression of genes (DEG) analysis performed in the different ecotypes of *P. tricornutum* using Pt1.8.6 as a reference (Supplemental data S1) showed that 2,284, 2,817, 1,975, 2,415, 2,332 and 2,457 genes are significantly more expressed in Pt3 Fu, Pt3 Ov, Pt3 Tr, Pt3, Pt4 and Pt8, respectively, than in the reference strain. Regarding down-regulated genes, 2,071, 2,219, 1,756, 2,460, 2,935 and 2,645 genes are significantly less expressed in Pt3 Fu, Pt3 Ov, Pt3 Tr, Pt3, Pt4 and Pt8, respectively, compared to Pt1.8.6. Overall, the number of DEG is relatively well balanced between the ecotypes and more or less expressed genes (over 2,400 DEG), except for Pt3 Tr, which presents less than 2,000 more or less expressed DEG. It should also be noted that Pt3 Ov is the

Ecotype accession	Major morphotype	Medium	T°C	Photoperiod (L:D)	Light intensity ( $\mu\text{mol m}^{-2} \text{s}^{-1}$ )	Days of culture/harvest	SRA	Reference
Pt 1.8.6	Natural fusiform 95–100%	100% artificial seawater + f/2-Si	20	12:12	100	8 days, exponentially growing cells	SRX957920 SRX957921	27
Pt 3 Fu	Enriched fusiform 89–91.2%	100% artificial seawater + Conway + Si	19	16:8	68	8 days, exponentially growing cells	ERR3285011 ERR3285012 ERR3285013 ERR3285014	29
Pt 3 Ov	Enriched oval 97.6–98.4%	10% artificial seawater + Conway + Si	19	16:8	68	8 days, exponentially growing cells	ERR3285015 ERR3285016 ERR3285017 ERR3285018	29
Pt 3 Tr	Enriched triradiate 76.5–77.5%	100% artificial seawater + Conway + Si	19	16:8	68	8 days, exponentially growing cells	ERR3285019 ERR3285020 ERR3285021 ERR3285022	29
Pt 3	Natural oval 60–75%	No details available in the paper*	20	12:12	75	Exponentially growing cells	SRR9945641	36
Pt 4	Natural fusiform 95–100%	No details available in the paper*	18	16:8	40	6 days (4 days in light + 2 days in darkness)	SRR5298736 SRR5298737 SRR5298738	28
Pt 8	Natural triradiate 80–85%	100% natural seawater + f/2	18	12:12	175	Cultures to reach $2.10^6$ cells. $\text{mL}^{-1}$ + 2 days	SRR10120340 SRR10120339	31

**Table 1.** Based on our search criteria (culture conditions, ecotypes, and morphotypes), the seven RNA-seq datasets included in our study are described as follows. \*As not specific details were provided in the publication regarding the composition of the culture media, we assumed that the culture media were either Conway or f/2, that are medium commonly used for the culture of *P. tricornutum*. All data were downloaded from SRA-NCBI in February 2023.

ecotype with the most genes more expressed than the reference strain (2,817) and Pt4 is the ecotype with the most genes less expressed than the reference strain (2,935). The ecotypes Pt3, Pt4 and especially Pt8 also present a greater number of genes with a significantly higher difference in expression (high  $-\log_{10}$  ( $p$  value)) compared to Pt3 Fu, Pt3 Ov and Pt3 Tr, especially for less expressed genes.

In order to investigate more precisely the transcriptomic differences between the ecotypes, a pathway analysis was carried out using ShinyGO (Table 2). This analysis showed that no significant enrichment was found for Pt3 Tr, either for more or less expressed genes, despite the morphotype difference between the two ecotypes (fusiform for Pt1.8.6 and triradiate for Pt3 Tr). The lack of significant enrichment can be attributed to the low number of DEGs between the two accessions. This result is in agreement with a previous transcriptomic study carried out on the 3 morphotypes of the Pt3 strain<sup>29</sup>. It was shown that less than 1% of the transcriptome was differentially expressed between the two morphotypes of the same ecotype. No significant enrichment was also found for Pt3 Fu and Pt3 Ov in terms of genes more expressed than the reference strain. On the other hand, cultures enriched with these morphotypes showed an enrichment in ‘Ribosome biogenesis in Eukaryotes’ and ‘Ribosome’ pathways with less expressed genes than the reference. This is also the case for Pt4 and Pt8. It should be noted that such enrichment is independent of the morphotypes, as Pt3 Ov is mainly oval, Pt4 mainly fusiform, while Pt8 is mainly triradiate. The enrichment of these pathways ‘Ribosome biogenesis in Eukaryotes’ and ‘Ribosome’ has been observed previously, but as over-expressed genes in studies dealing with the effect of naphthenic acids as a stress condition on *P. tricornutum*<sup>43</sup>,  $\text{H}_2\text{O}_2$  on *Chlorella pyrenoidosa*, tetracycline on *Scenedesmus obliquus*<sup>44</sup> or cold stress on *Neoporphyra haitanensis*<sup>45</sup>. The authors of these studies correlated the overexpression of these pathways with the adaptation of the microalgae to stress conditions. Pt3 and Pt8 were significantly enriched with more expressed genes than the reference strain in ‘Biosynthesis of secondary metabolites’, ‘Fatty acid biosynthesis’, ‘Porphyrin and chlorophyll metabolism’ and ‘Protein processing in endoplasmic reticulum’. On the other hand, Pt3 exhibited significant enrichment with less expressed genes than the reference strain in ‘Valine, leucine and isoleucine degradation’. Pt4 showed significant enrichments with more expressed genes than the reference strain in two pathways related to the protein life cycle: ‘Protein export’ and ‘Protein processing in endoplasmic reticulum’. In contrast, the less expressed genes of this ecotype compared to the reference strain presented enrichments in several pathways of the carbohydrate metabolism category such as ‘Citrate cycle (TCA cycle)’, ‘Pyruvate metabolism’, ‘Glycolysis / Gluconeogenesis’ and ‘Carbon metabolism’.

This first analysis could imply that Pt3 Fu, Pt3 Ov, Pt4 and Pt8 downregulate genes involved in stress pathways under control conditions. It could also show that Pt3 and Pt8 are more effective in lipid and specialized metabolism and Pt4 in folding, sorting and degradation, strategic processes in protein and *N*-glycan export pathways. However, at this first level of study, no correlation between a specific pathway and the morphotypes of *P. tricornutum* was observed.

To confirm the trend observed at this first level of investigation, we focused specifically on biological processes known in other organisms to have a significant impact on the production of Biologics and that can influence their critical quality attributes<sup>46,47</sup>. Therefore, we focused on the differential expression of genes involved in

	Pathways significantly enriched with more expressed than the reference strain genes	Pathways significantly enriched with less expressed than the reference strain genes
Pt 3 Fu	No significant enrichment found	Ribosome biogenesis in Eukaryotes Ribosome
Pt 3 Ov	No significant enrichment found	Ribosome biogenesis in Eukaryotes Ribosome
Pt 3 Tr	No significant enrichment found	No significant enrichment found
Pt 3	Fatty acid biosynthesis Ribosome Porphyrin and chlorophyll metabolism Pyrimidine metabolism Biosynthesis of cofactors Protein processing in endoplasmic reticulum Biosynthesis of secondary metabolites Metabolic pathways	Valine, leucine and isoleucine degradation
Pt 4	Protein export Protein processing in endoplasmic reticulum	Citrate cycle (TCA cycle) Pyruvate metabolism Porphyrin and chlorophyll metabolism Glycolysis / Gluconeogenesis Carbon metabolism Biosynthesis of amino acids Ribosome biogenesis in eukaryotes Ribosome Biosynthesis of secondary metabolites Biosynthesis of cofactors Metabolic pathways
Pt 8	Proteasome Photosynthesis Fatty acid biosynthesis Porphyrin and chlorophyll metabolism Protein processing in endoplasmic reticulum Biosynthesis of secondary metabolites Metabolic pathways	Ribosome biogenesis in eukaryotes

**Table 2.** Fold enrichment pathways analysis with DEG found in Pt3 Fu, Pt3 Ov, Pt3 Tr, Pt3, Pt4 or Pt8 (compared to Pt1.8.6) with ShinyGO 0.77 (Pathway database: KEGG, FDR cutoff 0.05).

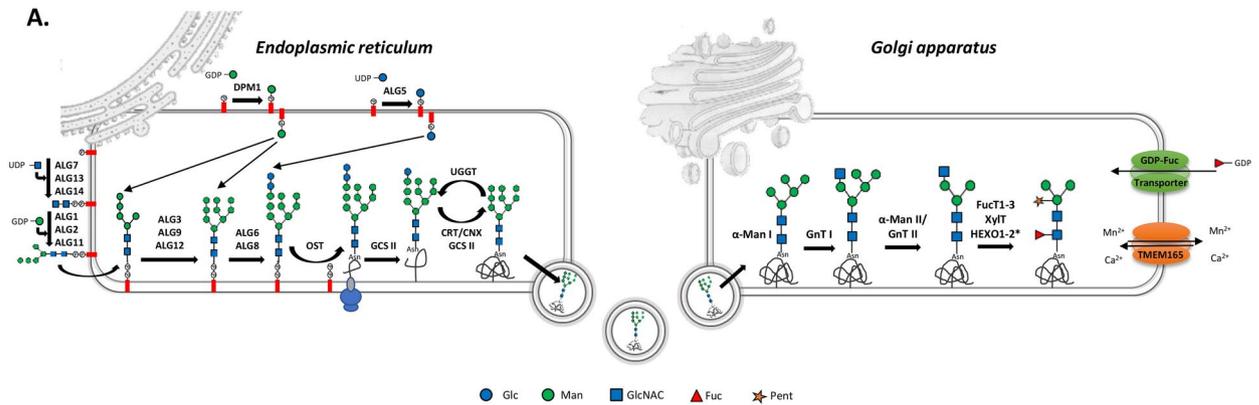
the *N*-glycan biosynthesis (Fig. 1), protein export and secretion (Fig. 2), protein quality control (Fig. 3), and proteasome (Fig. 4) in the different ecotypes of *P. tricornutum*.

### ***N*-glycan biosynthesis**

Most of the Biologics are glycoproteins for which the *N*-glycosylation is well known to influence their biological activities, potential immunogenicity and allergenicity<sup>46,48</sup>. Therefore, when selecting an organism as a host for the production of glycoproteins as Biologics, particular attention should be paid to the *N*-glycosylation pathway. The diatom *P. tricornutum* has been shown to be able to produce glycoproteins such as mAbs without immunogenic glycoepitopes on their *N*-glycans<sup>10</sup>. Moreover, these recombinant mAbs expressed in *P. tricornutum* have been shown to be glycosylated with mammalian-like oligomannosides resulting from processing steps in the endoplasmic reticulum (ER) and early Golgi apparatus (Fig. 1A). It is noteworthy that, despite the presence of glycoenzymes potentially involved in complex-type *N*-glycan biosynthesis within the Golgi apparatus of *P. tricornutum* (Fig. 1A), including *N*-acetylglucosaminyltransferase I (GnT I), fucosyltransferases or xylosyltransferases, a glycoproteomic study of the mAbs produced in *P. tricornutum* revealed that oligomannosidic *N*-glycan structures constitute the major *N*-glycan population of mAbs produced in *P. tricornutum*<sup>10</sup>. Based on this knowledge, we screened all the genes known so far to be involved in the different steps of the *N*-glycosylation pathway in *P. tricornutum* (Fig. 1B). Pt3 Ov, Pt4 and Pt8 have more differentially expressed genes (10, 13 and 12, respectively) involved in the *N*-glycosylation pathway as compared to the others, which present 7 differentially expressed genes. These are also the strains with the most under-expressed genes, with 6, 10 and 8 under-expressed genes, respectively, compared to the reference strain, Pt1.8.6 (Fig. 1B).

*N*-glycan synthesis is initiated in the ER with the synthesis of a dolichol-linked intermediate oligosaccharide precursor, commonly known as the lipid-linked oligosaccharide (LLO). The enzymes that synthesize this precursor, called ALGs (Asparagine-Linked Glycosylation), use nucleotide sugars and specific sugar acceptors to elongate the LLO (Fig. 1A). Among the 13 genes involved in the synthesis of this LLO, three are more highly expressed in the Pt3 Fu, Pt3 Ov and Pt3 Tr strains. The Pt3 Fu and Pt3 Tr strains have the same three genes that are more expressed than the reference, and the same two genes that are less expressed than the reference. The Pt4 and Pt8 strains have four and two genes, respectively, that are less expressed than in the reference strain, (Fig. 1B).

The Glc<sub>2</sub>Man<sub>9</sub>GlcNAc<sub>2</sub>-P-P-Dol precursor synthesized in *P. tricornutum* is then transferred *en bloc* to the asparagine at the consensus *N*-glycosylation sites (Asn-X-Thr/Ser/Cys), by an enzymatic complex called oligosaccharyltransferase (OST). This step takes place in the lumen of the ER. Two genes encoding putative catalytic subunits of the OST, named Phatr3\_J55197 (SST3A) and Phatr3\_J55198 (SST3B), have been identified in the *P. tricornutum*'s genome. The gene encoding SST3A is not differentially expressed between the different strains compared to the reference strain. In contrast, SST3B is less expressed in most samples, including Pt3 Ov,



**Fig. 1.** (A) Proposed *N*-glycan biosynthesis pathway in *P. tricornutum* and differential genes expression in the ten *P. tricornutum* strains. Representation of the putative mechanism of action of the different enzymes identified in the *N*-glycosylation biosynthesis pathway of *P. tricornutum*. The first steps take place in the endoplasmic reticulum leading to the synthesis of the lipid-linked oligosaccharide precursor, Glc<sub>2</sub>Man<sub>9</sub>GlcNAC<sub>2</sub>, which is then transferred en bloc onto the nascent protein thanks to the oligosaccharyltransferase. After the protein quality control steps, involving some chaperone proteins, the glycoprotein is transported to the Golgi apparatus where the *N*-glycan undergoes maturation steps leading to a paucimannosidic structure Pent1Fuc1Man3GlcNAC<sub>2</sub>. The *N*-glycan structures were drawn according to the Symbol Nomenclature for Glycans (Varki, 2017). Glc: glucose, Man: mannose, GlcNAC: N-acetylglucosamine, Fuc: fucose, Pent: Pentose like xylose or arabinose. HEXO1-2\* : the localization of the hexosaminidases is not yet determined in *P. tricornutum*. (B). Genes implicated in the *N*-glycan biosynthesis pathway of *P. tricornutum*. In green, genes more expressed than the reference strain with significant fold change > 2 (i.e.  $\log_2FC > 1$ ),  $pvalue < 0,05$ ; in red genes less expressed than the reference strain with significant fold change < 0,5 (i.e.  $-1 < \log_2FC$ ),  $pvalue < 0,05$ ; in white genes with non-significant fold change. Genes in Pt3 Fu, Pt3 Ov, Pt3 Tr, Pt3, Pt4 or Pt8 compared to Pt1.8.6.

Pt3, Pt4 and Pt8. After being transferred by the OST, the *N*-glycan attached to the glycoprotein is then matured by two enzymes located in the lumen of the ER: the Glucosidase II (GCS II) and the UDP-glucose: glycoprotein glucosyltransferase (UGGT). Two predicted genes have been described to encode such GCS II in *P. tricornutum*: Phatr3\_EG02351 gene, that is more expressed than the reference strain in Pt3, Pt4 and Pt8 and Phatr3\_J54169, that is less expressed than the reference strain in Pt8 (Fig. 1B). The UGGT gene in *P. tricornutum* itself is less expressed in all the samples except in Pt3 and Pt8. Once correctly folded, the glycoproteins leave the ER to reach the Golgi apparatus, where maturation steps involve a variety of enzymes that result in species-specific *N*-glycan structures.

Upon entry into the Golgi apparatus, mannose residues are trimmed from the *N*-glycans by the alpha-mannosidase I, resulting in the Man<sub>5</sub>GlcNAC<sub>2</sub> structure (Fig. 1A), which is the substrate for the GnT I. GnT I has been shown to be a key enzyme in the synthesis of complex *N*-glycan structures<sup>49,50</sup>. The gene encoding GnT I (Phatr3\_J54844) is more expressed in the Pt4 strain than in the reference strain whereas it is less expressed in Pt3 Fu, Pt3 Ov, Pt3 Tr and Pt8 than in the reference strain (Fig. 1B). Therefore, the Pt4 strain might be a better choice for the production of recombinant glycoproteins carrying complex-type *N*-glycans as compared to others. However, the biochemical characterization of the anti-VHB mAbs produced in the Pt4 strain of *P. tricornutum* showed that mainly Man<sub>9</sub>GlcNAC<sub>2</sub> and Man<sub>8</sub>GlcNAC<sub>2</sub> *N*-glycans were found on the surface of the mAbs without any detection of complex-type *N*-glycans<sup>10</sup>. This suggests that there is no direct relationship between the expression level of GnT I and the final *N*-glycan structure and that the maturation of the oligomannosidic *N*-glycans in the Golgi apparatus seems to be limited, at least on mAbs. Indeed, the absence of complex-type *N*-glycans on the surface of the mAbs could be explained by the nature of the protein, since the *N*-glycosylation sites of mAbs are buried within the CH<sub>2</sub> constant domain. Thus, the folding of the mAbs limits the accessibility of the *N*-glycans to the Golgi processing glycosyltransferases, leading to incomplete maturation of the *N*-glycans. Nevertheless, the *N*-glycosylation profile of a total protein extract from the Pt1 strain showed that more mature *N*-glycans can be found. However, the complex *N*-glycans were present at low level on the surface of *P. tricornutum* endogenous proteins<sup>51</sup>. Thus, it would be interesting in future studies to analyze the *N*-glycan profiles of all the ecotypes of *P. tricornutum*.

Special attention was also paid to fucosylation, as it has been previously shown that the  $\alpha(1,3)$ -fucose glycoepitopes of plant-made biopharmaceuticals can lead to an immunogenic response in humans<sup>52</sup>. Looking at the fucosylation in *P. tricornutum*, it was demonstrated that the overexpression of PtFucT1 leads to a significant increase in  $\alpha(1,3)$ -fucosylation on *P. tricornutum* endogenous glycoproteins<sup>53</sup>. In the present work, the expression of the gene encoding PtFucT1 (Phatr3\_J54599) is not modified within the different ecotypes. In contrast, the putative PtFucT2 is more expressed in Pt3 and Pt8 than in the reference strain while the expression of the gene encoding for the putative PtFucT3 is less expressed in Pt3, Pt4, and Pt8 than in the reference strain. The GDP-Fuc transporter of *P. tricornutum* has also been characterized and shown to restore fucosylation of proteins in a mutant CHO cell line lacking its endogenous GDP-Fuc transporter activity<sup>53</sup>. The gene encoding for this

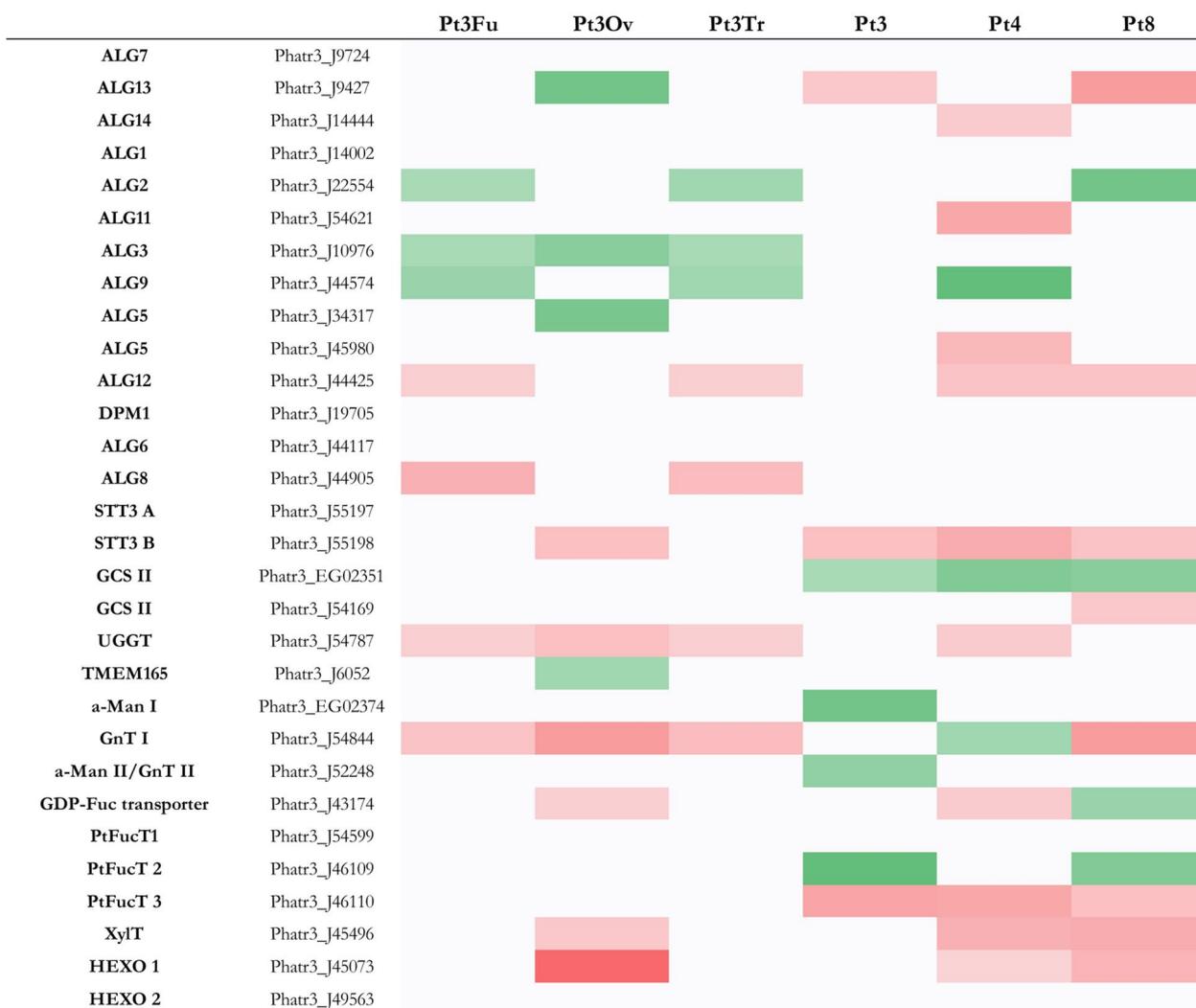


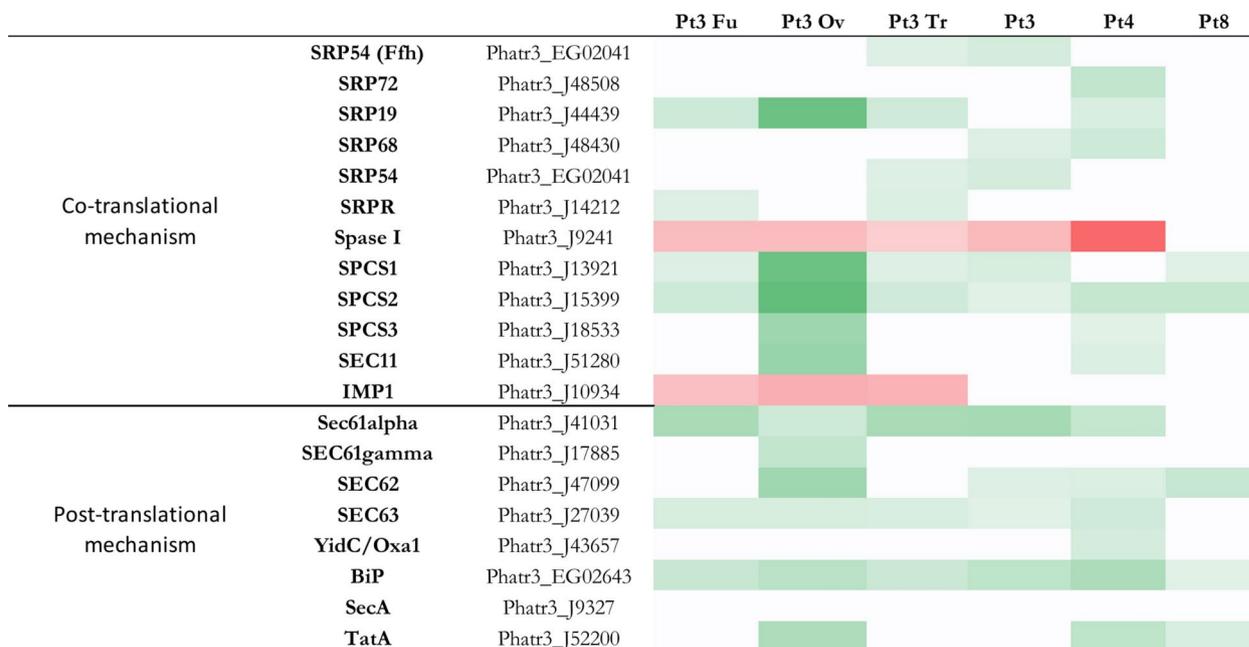
Figure 1. (continued)

transporter (Phatr3\_J43174) is less expressed in Pt3 Ov and Pt4 than in the reference strain, but more expressed in Pt8 than in the reference strain (Fig. 1B). Once again, the Pt4 strain seems to be more interesting for the production of Biologics dedicated to human health, since the genes involved in the addition of immunogenic epitopes are less expressed in this ecotype.

The regulation of the processing of oligomannosidic to complex type *N*-glycans depends on the accessibility of the *N*-glycans to glycosyltransferases (GTs) and glycosidases in the Golgi apparatus, but also on the regulation of these enzymes by the Golgi Mn<sup>2+</sup> homeostasis<sup>54</sup>. Recent studies carried out on human congenital disorders of glycosylation have highlighted the role of TMEM165 (transmembrane protein 165) in Golgi cation homeostasis and the impact of *TME165* deficiency on the *N*-glycan processing<sup>55,56</sup>. *TMEM165* homologous sequences have been identified in several eukaryotic organisms including *P. tricornutum* but their role is not yet understood<sup>54</sup>. The gene Phatr3\_J6058 from *P. tricornutum* encodes a putative homolog of TMEM165, which is only more expressed in Pt3 Ov than in the reference strain compared to other strains or morphotypes (Fig. 1B). The presence of TMEM165 in organisms that do not produce many complex glycans, such as *Phaeodactylum*, and its overexpression in stress-related morphotypes could potentially suggest a role in osmoregulation. Indeed, comparative EST-based analyses showed that the oval morphotype exhibited an up-regulation of genes encoding proteins involved in stress responses with putative roles in signaling, membrane remodeling, protein degradation but also in cell homeostasis<sup>38</sup>. More recently, Kleiner et al., have also shown an increase in cytosolic Ca<sup>2+</sup> concentration in *P. tricornutum* in response to cold stress<sup>37</sup>.

### Protein export and secretion

In mammalian cell factories, it is well known that improving the efficiency and stability of recombinant protein production relies on maximizing the protein export and secretory capacity of the cells<sup>47</sup>. Recently, it has been shown that microalgae are also capable of secreting recombinant proteins into their culture media<sup>9,58,59</sup>, including *P. tricornutum*<sup>57</sup>. The protein export is the active transport of proteins from the cytoplasm to the outside of the cell. Both prokaryotic and eukaryotic organisms have two major modes of protein export: (1) the co-translational



**Fig. 2.** Heatmap of genes implicated in the protein export pathway of *P. tricornutum*. In green, genes more expressed than the reference strain with significant fold change  $> 2$  (i.e.  $\log_2FC > 1$ ),  $p$  value  $< 0,05$ ; in red genes less expressed than the reference strain with significant fold change  $< 0,5$  (i.e.  $-1 < \log_2FC$ ),  $p$  value  $< 0,05$ ; in white genes with non-significant fold change. Genes in Pt3 Fu, Pt3 Ov, Pt3 Tr, Pt3, Pt4 or Pt8 compared to Pt1.8.6.

delivery mechanism, in which the nascent polypeptide can be delivered to the export membrane along with the ribosome during protein synthesis, or (2) the post-translational mechanism, in which protein delivery can occur after its synthesis is complete<sup>60</sup>. In *P. tricornutum*, the predicted genes related to protein export and secretion that have been described are derived from both prokaryotes and eukaryotes. In all strains, most of the genes involved in the co- and post-translational delivery mechanism are more expressed than the reference strain, especially in Pt3 Ov and Pt4, which have the highest number of genes more expressed than the reference strain compared to other strains, with 11 and 12 genes overexpressed, respectively (Fig. 2).

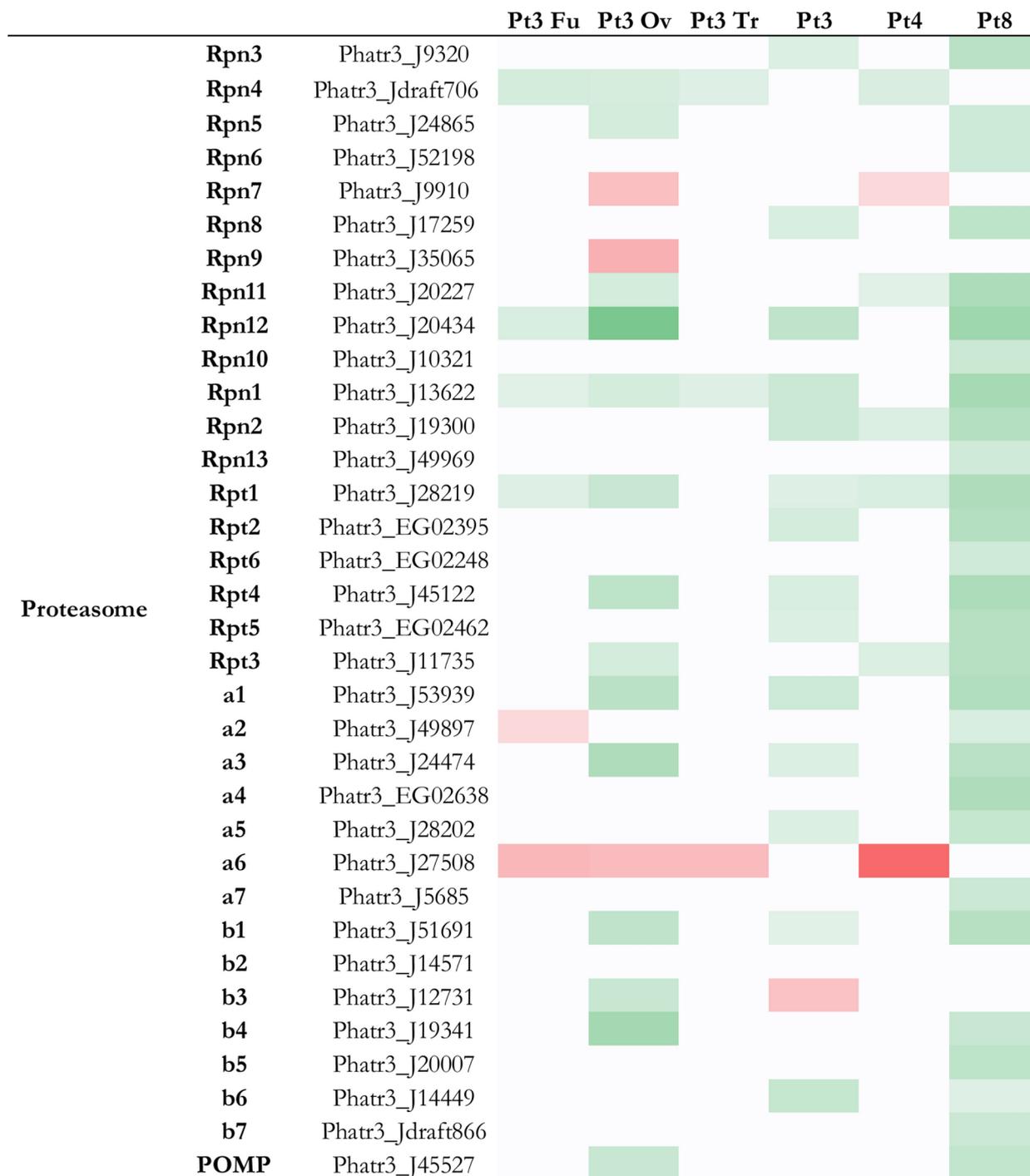
In the co-translational mechanism, the peptide chains are first synthesized by the ribosome. Then, signal recognition particles (SRPs) recognize and bind to the signal peptide. Simultaneously with the signal peptide binding, the SRP recognizes the SRP receptor (SRPR) in the ER membrane and then forms a complex with the ribosome on the membrane. Among the SRPs, SRP19 is the most differentially expressed among the six different accessions, especially in Pt3 Ov. Only Pt3 Fu and Pt3 Tr present a significant differential expression for SRPR, which it is more expressed than the reference strain in both strains (Fig. 2). Pt8 has no gene related to SRP or SRPR that is differentially expressed. Subsequently, the translocation of the peptide occurs, leading to the entry of the peptide into the lumen of the ER. During this process, signal peptides are removed by the signal peptidase in the lumen of the ER. In prokaryotes, signal peptidases (SPase I and II) are involved. In eukaryotes, this step is performed by signal peptidase complex subunits (SPCS 1, 2 and 3), signal peptidase I (SEC11) and mitochondrial inner membrane protease subunits (IMP)<sup>60</sup>. Our study reveals that *Spase I* is less expressed than the reference strain in all strains except for Pt8, where no significant differential expression was observed. *IMP1* is also less expressed than the reference strain in Pt3 Fu, Pt3 Ov and Pt3 Tr, no significant differential expression was observed in the other strains compared to the reference strain. Moreover, in Pt3 Ov, all the other genes (*SPC1*, *SPC2*, *SPC3* and *SEC11*) are more expressed than the reference strain, in the other strains only two or three of them are more expressed than the reference strain. *SPCS2* is more expressed than the reference strain in all strains.

Regarding the genes involved in the post-translational mechanism, they are all more expressed than the reference strain or with no significant difference in all strains. Moreover, Pt3 Ov and Pt4 are the most predominant with six out of eight genes (Fig. 3). In eukaryotic organisms, the post-translational mechanism involves the translocation channel consisting of SEC61, SEC62, SEC63 (transport protein Sec61 subunits alpha/gamma and translocation proteins) and BiP (an ER chaperone). All these actors are more expressed in Pt3 Ov compared to the reference strain, and BiP is more expressed in all the strains compared to Pt1.8.6. In prokaryotic organisms, this mechanism involves, among others, the membrane protein insertase of the YidC/Oxa1 family and the preprotein translocase subunit SecA<sup>60</sup>. Almost no difference in expression is observed for these two actors. Finally, another protein transport system that transports folded proteins in bacteria, archaea, and chloroplasts is the twin-arginine translocation (Tat) pathway, which involves several membrane proteins with overlapping functions, of which TatA is the most important<sup>61</sup>. In our study, Tat A was more expressed than the reference strain only in Pt3 Ov, Pt4 and Pt8, while the other three strains did not show any significant expression.



**Fig. 3.** Heatmap of genes implicated in quality control of *P. tricornutum*. In green, genes more expressed than the reference strain with significant fold change  $> 2$  (i.e.  $\log_2FC > 1$ ),  $p$  value  $< 0,05$ ; in red genes less expressed than the reference strain with significant fold change  $< 0,5$  (i.e.  $-1 < \log_2FC$ ),  $p$  value  $< 0,05$ ; in white genes with non-significant fold change. Genes in Pt3 Fu, Pt3 Ov, Pt3 Tr, Pt3, Pt4 or Pt8 compared to Pt1.8.6.

The fact that more genes involved in the protein export pathway are more expressed in Pt3 Ov seems to correlate with previous results. Indeed, it has been shown that more proteins are expressed at higher levels in oval cell cultures compared to triradiate or fusiform cell cultures<sup>32,41</sup>. A proteomic analysis of the secretome of the three morphotypes of Pt3 reveals that 385 proteins are secreted in the culture media of Pt3 Ov, compared to 275 and 231 proteins secreted for Pt3 Fu and Pt3 Tr, respectively<sup>32</sup>. A study led by Galas et al., in 2021 also confirmed that Pt3 Ov cells can secrete proteins faster and in higher amounts than the other two morphotypes<sup>40</sup>. As mentioned above, since secretory capacity is important for the Biologics accumulation in the culture medium, it seemed important to focus on protein secretion, regulation of exocytosis and vesicle docking involved in exocytosis. However, very little data are currently available on identified or predicted genes involved in these pathways in *P. tricornutum*. Since this microalgae is a metabolic intermediate between plants and mammals<sup>1</sup>



**Fig. 4.** Heatmap of genes implicated in proteasome of *P. tricornutum*. In green, genes more expressed than the reference strain with significant fold change  $> 2$  (i.e.  $\log_2FC > 1$ ),  $p$  value  $< 0,05$ ; in red genes less expressed than the reference strain with significant fold change  $< 0,5$  (i.e.  $-1 < \log_2FC$ ),  $p$  value  $< 0,05$ ; in white genes with non-significant fold change. Genes in Pt3 Fu, Pt3 Ov, Pt3 Tr, Pt3, Pt4 or Pt8 compared to Pt1.8.6.

and to extend our knowledge of the secretion and exocytosis process in *P. tricornutum*, we performed a BLAST analysis on the *Phaeodactylum* genome to identify homologous genes compared to *A. thaliana* (AT) and *Homo sapiens* (HS) genes that are involved in these pathways. The expression of the identified genes was then described in the selected RNA-Seq (Supplemental 2 and 3). It was noted that Pt4 has more genes involved in 'protein secretion and exocytosis' that are more expressed than the reference strain (8 genes). Pt3 Ov has specifically more expressed genes than the reference strain, such as Ras-related protein RAB8 (Phatr3\_J22713) and Kish-A-like protein (Phatr3\_J12535). Pt4 has specifically more expressed than the reference strain genes such as Ras-related protein RABE1e (Phatr3\_J22095), Golgi apparatus membrane protein-like ECHIDNA (Phatr3\_J6254),

PHF1 (a SEC12-like protein 1) (Phatr3\_J7966), and exocyst complex component SEC3A (Phatr3\_EG02390). Finally, three Ras-related proteins, Rab-10 (Phatr3\_J44216), Rab-8A (Phatr3\_EG02235) and Rab-3C (Phatr3\_J22095) are specifically more expressed in Pt4 than in the reference strain.

Regarding the expression of genes involved in the protein export pathway and the secretion, Pt3 Ov and Pt4 seem to have more potential to enhance the protein production and its export in the culture media.

### Quality control and proteasome

Before leaving the ER and being delivered to various intracellular compartments or the extracellular environment, proteins interact with several quality control factors whose aim is to release well-folded, functional proteins. Analysis of the DEGs encoding these factors reveals that Pt3, Pt4, and Pt8 have more more-expressed genes than the reference strain that are involved in the protein quality control than others (15, 18, and 18 genes respectively) (Fig. 3). Among these factors, nascent proteins interact with chaperones such as calreticulin, calnexin, heat shock proteins (HSP) such as BiP (Binding Immunoglobulin Protein) and GRP (Glucose related proteins). The genes encoding BiP (Phatr3\_EG02643) and GRP94 (Phatr3\_J16786) are more expressed in all strains compared to Pt1.8.6, but there is no significant differential expression of the gene encoding HSP70\_2 (Bip2-Phatr3\_J21519). Recently, some studies have demonstrated a correlation between BiP expression levels and mAb production in mammalian cell lines producing mAbs<sup>62,63</sup>. For *N*-glycosylated proteins, the oligosaccharide on their surface is recognized by two other ER luminal enzymes: glucosidases I and II (GCSI and GCSII)<sup>64</sup>. These enzymes cleave the terminal glucose residues. The GCSI cleaves the first terminal residue and the GCSII cleaves the other two. Before the last glucose residue is trimmed, the glycoprotein is recognized by the chaperone proteins and, if properly folded, the third glucose is removed by the GCSII and the protein is directed to the secretory pathway. The gene Phatr3\_EG02531, encoding a GCSII is more highly expressed in Pt3, Pt4 and Pt8 than in the reference strain. Misfolded proteins whose three glucoses are cleaved are recognized by UGGT (UDP-glucose/glycoprotein glucosyltransferase), which adds glucose residues to these misfolded proteins to send them back to the chaperones. If the protein is still misfolded after the ER quality control cycle, it is recognized by  $\alpha$ 1,2-mannosidase I via the EDEM (ER degradation enhancing 1,2-mannosidase protein). The gene Phatr3\_J52346, which encodes the EDEM protein, is expressed at higher levels than Pt1.8.6 in all strains except Pt3 Ov, for which the expression is not significantly different from Pt1.8.6. The  $\alpha$ 1,2-mannosidase I specifically cleaves mannose residues, allowing the misfolded proteins to be directed to the ER-associated degradation (ERAD)<sup>64</sup>. It has been shown that, among other stress factor, the expression of proteins known to be difficult to produce, such as mAbs, can disrupt the ER homeostasis, leading to the accumulation of misfolded proteins, and subsequently of the activation of the ERAD pathway to restore the ER homeostasis<sup>62,63</sup>. Regarding the ERAD pathway, Pt4 has more genes more expressed than the reference strain (8 genes) and Pt3 has more 1 genes less expressed than the reference strain. The genes encoding Ubc W-B (Phatr3\_J10724), Ubx1 (Phatr3\_J47158) and SCDC48 (Phatr3\_J36597) are more expressed in the seven strains than in the reference strain.

In the ERAD pathway, most substrate proteins are ubiquitin-modified prior to proteasomal degradation. The proteasome is a protein-degrading apparatus composed of two subcomplexes: a 20S core particle (CP), which carries the catalytic activity and a regulatory 19S regulatory particle (RP). The CP is composed of heptameric rings: outer identical  $\alpha$ -rings (a1 to 7) and inner identical  $\beta$ -rings (b1 to 7) and a maturation protein (POMP). The RP consists of two eight-subunit subcomplexes: the lid (Rpn3 to Rpn12), and the base (Rpn1, 2 and 13, Rpt 1 to 6). Rpn10 can interact with either the lid or the base, stabilizing the interaction between the two<sup>65</sup>. Regarding the proteasome, Pt8 has more and only more expressed genes than the reference strain compared to others (i.e. 28 genes) (Fig. 4), followed by Pt3 Ov and Pt3, which both have 14 genes that are more expressed compared to Pt1.8.6. A correlation between the overexpression of QC-related genes, especially those involved in ERAD, and the overexpression of proteasome-related genes could be expected, since the activation of the proteasome follows that of the ERAD pathway. This correlation is observed for Pt8, which has the most overexpressed genes in both cases. However, this is not the case for Pt3 Ov and Pt3.

Considering the elements provided by the analysis of DEGs involved in QC and proteasome, we can suggest that the Pt3 Fu and Pt3 Tr could be a bit more interesting for the production of Biologics. Indeed, these strains have more down-regulated genes and less up-regulated genes that are involved in QC and proteasome than the other strains. It can be assumed that the lower activation of genes related to these pathways implies a better capacity to produce correctly folded proteins.

### Proteases

As mentioned above, the mAbs production yields obtained in *P. tricornutum* are approximately 40-fold lower than the 10 g/L observed in CHO cells<sup>18</sup>. The currently observed low yields may be related to the proteolytic degradation of mAbs accumulated in the culture medium. Although host cell proteases are critical due to their essential functions in catalytic and metabolic pathways, as well as in the renewal of extracellular waste<sup>66</sup>, when present in the cell culture medium they can degrade the product of interest<sup>67</sup>. Indeed, previous studies in plants or CHO have mentioned that proteolytic degradation of mAbs that accumulate in the culture media of cell expression systems can dramatically affect the yield of some recombinant proteins<sup>67,68</sup>. Proteases can accumulate in the culture media either by secretion from the host cell or by release during cell lysis. Moreover, an *in silico* gene analysis<sup>69</sup> and a proteomic analysis<sup>32</sup> of the secretome of *P. tricornutum* suggest that this microalga secretes a large number of putative proteases. It is essential to consider the occurrence of proteases in the selected datasets (Supplemental 4).

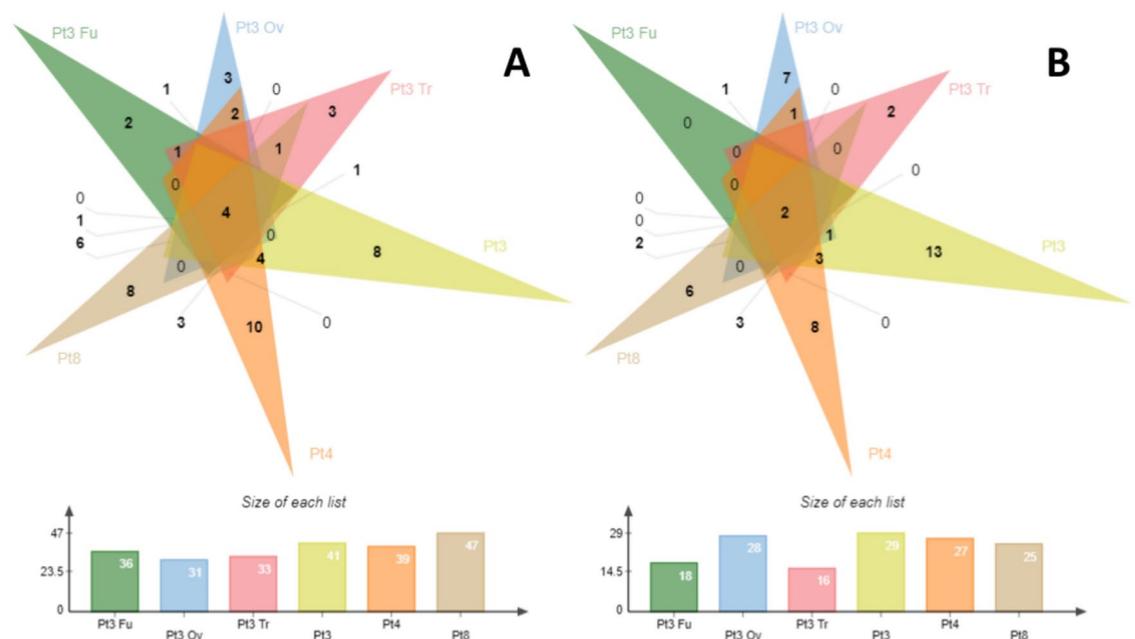
First, for all strains, most of the identified proteases have no significant expression (about 110 out of the 173 identified proteases). Then, the overall number of proteases is relatively well balanced between the ecotypes for more expressed than the reference strain genes (about 35 proteases) or less expressed than the reference strain genes (about 25) (Supplemental 4). We then wanted to know if there were any proteases that were specifically

expressed in each strain. For the proteases more expressed than the reference strain, 2, 3, 3, 8, 10, and 8 proteases were found only in Pt3 Fu, Pt3 Ov, Pt3 Tr, Pt3, Pt4, and, Pt8 respectively. For proteases less expressed than the reference strain, 7, 2, 13, 8 and 6 DEGs were found only in Pt3 Ov, Pt3 Tr, Pt3, Pt4, and, Pt8, respectively (Fig. 5). Finally, among these specific proteases, we identified (in red) those secreted according to Chuberre et al. 2021 and Dorrell et al. 2021 (Table 3). Our analysis showed that Pt3 and Pt4 have the most specifically secreted proteases, which are less expressed than the reference strain (six proteases). However, Pt3 is also the strain with the most specifically secreted proteases that are more expressed than the reference strain (five proteases), while Pt4 has only two. It is also worth mentioning that Pt3 Ov has 2 secreted proteases more expressed than the reference strain and 4 secreted proteases less expressed than the reference strain. The protease categories identified include serine-type proteases, cysteine-type proteases, metalloproteases, ATP-dependent proteases, and thiol-dependent proteases. In all strains, at least one gene encoding a secreted serine-type protease is specifically more expressed than in the reference strain. Serine-type proteases are commonly released as host cell proteins (HCPs) into the culture media, causing the degradation of recombinant proteins during production<sup>70,71</sup>. Their removal during purification is often challenging, with some, even co-eluting with mAbs during protein A chromatography<sup>72</sup>.

Considering the level of information provided by the expression of proteases, and especially of those secreted in the culture media, the Pt4 ecotype seems to be more interesting for the production of Biologics, given the low number of less expressed than the reference secreted proteases and the high number of more expressed than the reference secreted proteases. Thus, this accession is likely to lead to a higher accumulation of recombinant proteins in the culture medium. Nevertheless, it would be interesting to precisely identify the proteases directly involved in the degradation of recombinant proteins, in order to strengthen the conclusion regarding the interest of one strain over another.

## Conclusion

In conclusion, our meta-analysis of publicly available transcriptomic data from different strains of *P. tricornutum* showed that compared to the others, Pt3 Ov, Pt4 and Pt8 seemed to be more efficient in the *N*-glycosylation pathway, Pt3 Ov and Pt4 in protein export and secretion, Pt3 Tr and Pt3 Fu in quality control and proteasome and Pt3 Ov and Pt4 were less affected by secreted proteases. Therefore, based on this analysis and considered it as a whole, Pt4 appears to be the optimal strain for the production of recombinant proteins, confirming existing data that identify the Pt4 strain as the most promising for the production of Biologics<sup>5–7,41,42,73</sup>, followed by Pt3 Ov<sup>32,40</sup>. It is worth mentioning that the Pt8 strain, which also has interesting characteristics even though, has never been mentioned for the production Biologics. In addition, it is important to note that the laboratory conditions to which the strains are exposed, may favor mutations, especially due to a high growth rate, which has already been observed in strains coming from different laboratories<sup>74</sup>. For example, it has recently been shown that Pt1 strains issued from two different laboratories, present different lipid contents<sup>42</sup>. This is certainly a limitation of this meta-analysis, which uses datasets from different RNA-seq studies with similar but not strictly identical culture conditions. Due to our selection criteria for RNA-seq data regarding the closest similarity of culture conditions, this meta-analysis concerns only 4 different *P. tricornutum* ecotypes, whereas 10 ecotypes are currently commonly described in the literature. Recently, an RNA-seq analysis of the 10 most commonly used



**Fig. 5.** Venn diagram of genes implicated in protease activities in *P. tricornutum*. (A) More expressed genes than the reference strain in Pt3 Fu, Pt3 Ov, Pt3 Tr, Pt3, Pt4 or Pt8 compared to Pt1.8.6. (B) Less expressed genes than the reference strain in Pt3 Fu, Pt3 Ov, Pt3 Tr, Pt3, Pt4 or Pt8 compared to Pt1.8.6.

	Genes of proteases up-regulated		Genes of proteases down-regulated	
	Gene ID	Predicted protease	Gene ID	Predicted protease
Pt3 Fu	Phatr3_EG02027 Phatr3_J11204	Extracellular metalloprotease Serine-type carboxypeptidase	-	-
Pt3 Ov	Phatr3_J44664 Phatr3_J50206 Phatr3_J31772	Cysteine-type peptidase Serine-type endopeptidase Thiol-dependent deubiquitinase	Phatr3_J12595 Phatr3_J40690 Phatr3_J24006 Phatr3_J42588 Phatr3_J13331 Phatr3_J46562 Phatr3_J3924	Metalloaminopeptidase Metalloendopeptidase Metalloendopeptidase Metalloendopeptidase Serine-type peptidase Serine-type peptidase Thiol-dependent deubiquitinase
Pt3 Tr	Phatr3_J11217 Phatr3_J38448 Phatr3_J46548	ATP-dependent Clp protease Metalloaminopeptidase serine-type endopeptidase	Phatr3_J48319 Phatr3_J47594	Metalloprotease Serine-type endopeptidase
Pt3	Phatr3_J12595 Phatr3_EG02434 Phatr3_J12106 Phatr3_J12241 Phatr3_J5396 Phatr3_J35781 Phatr3_J3061 Phatr3_J42729	Metalloaminopeptidase Metalloaminopeptidase Metalloendopeptidase Peptidase Peptidase Protease Serine-type peptidase Ubiquitin carboxyl-terminal hydrolase	Phatr3_J9388 Phatr3_J16168 Phatr3_J11217 Phatr3_J25433 Phatr3_J4936 Phatr3_J44664 Phatr3_J42694 Phatr3_J49223 Phatr3_J14380 Phatr3_J49652 Phatr3_J44745 Phatr3_J31772 Phatr3_EG02322	Aspartic-type endopeptidase Aspartic-type endopeptidase Clp protease Cysteine-type peptidase Cysteine-type peptidase Cysteine-type peptidase Lys48-specific deubiquitinase Serine-type endopeptidase Serine-type endopeptidase Serine-type peptidase Serine-type peptidase Thiol-dependent deubiquitinase Thiol-dependent deubiquitinase Thiol-dependent deubiquitinase
Pt4	Phatr3_J9625 Phatr3_J10593 Phatr3_J45606 Phatr3_J13602 Phatr3_EG02320 Phatr3_J46087 Phatr3_J46562 Phatr3_J46565 Phatr3_J29340 Phatr3_Jdraft1763	Aspartic-type endopeptidase ATP-dependent Clp protease Metalloendopeptidase Metallopeptidase Metallopeptidase Serine-type endopeptidase Serine-type peptidase Serine-type peptidase Thiol-dependent deubiquitinase Thiol-dependent ubiquitinyl hydrolase activity	Phatr3_J1884 Phatr3_J38448 Phatr3_J45960 Phatr3_J48491 Phatr3_J4014 Phatr3_J40462 Phatr3_J42439 Phatr3_J48493	Metalloaminopeptidase Metalloaminopeptidase Serine-type carboxypeptidase Serine-type endopeptidase Serine-type endopeptidase Serine-type endopeptidase Serine-type endopeptidase Serine-type endopeptidase
Pt8	Phatr3_J16455 Phatr3_EG02217 Phatr3_J9309 Phatr3_J8670 Phatr3_J45960 Phatr3_J47107 Phatr3_J13240 Phatr3_J4014	clpC peptidase Cysteine-type peptidase Metalloendopeptidase Metalloendopeptidase Serine-type carboxypeptidase Serine-type endopeptidase Serine-type endopeptidase Serine-type endopeptidase	Phatr3_J9625 Phatr3_J45606 Phatr3_J42587 Phatr3_J50206 Phatr3_J50346 Phatr3_J37460	Aspartic-type endopeptidase Metalloendopeptidase Metalloendopeptidase Serine-type endopeptidase Serine-type endopeptidase Serine-type peptidase

**Table 3.** Genes of proteases more or less expressed only found in Pt3 Fu, Pt3 Ov, Pt3 Tr, Pt3, Pt4 or Pt8 (compared to Pt1.8.6). In red, secreted proteases identified according to Chuberre et al., 2021 and Dorrell et al., 2021).

*P. tricornutum* strains cultured under exactly the same conditions was published<sup>75</sup>. This study in agreement with our meta-analysis, demonstrated that the Pt4 strain appears to be the most promising for Biologics production<sup>75</sup>. In future work, it would be interesting to produce antibodies in the different ecotypes of *P. tricornutum* and evaluate whether these strains are still the most interesting on an industrial scale. Thus, further work needs to be performed in the coming years to understand and better control the cell biology of *P. tricornutum* in order to finally improve the Biologics production yield in this diatom.

Methods.

### Identification and selection of the RNA-seq datasets

RNA-seq studies used were chosen according to (1) procedures of culture, (2) ecotypes and (3) morphotypes. The chosen culture conditions were those available where microalgae cells were cultured in natural or artificial seawater media supplemented with Conway or f/2, at a temperature comprised between 18 and 20 °C and a photoperiod of 12:12 or 16:8 with light intensity comprised between 40 and 175  $\mu\text{mol m}^{-2} \text{s}^{-1}$  and harvested on the 6th to 8th day of culture. The chosen datasets are from the three morphotypes (oval, fusiform and triradiate as well as enriched versions of these morphotypes) (Table 1). The datasets used are those of cells grown without any stress (control). All libraries used were sequenced on an Illumina instrument. Based on those search criteria, seven RNA-seq datasets were included in our study. All information about these datasets is summarized in Table 1. FASTQ files were downloaded from the Sequence Read Archive (SRA)-NCBI (<https://www.ncbi.nlm.nih.gov/sra/docs/srdownload>). A PRISMA checklist and flow diagram are available in Supplemental data S5.

### Trimming, alignment, and gene count

FastQ files were uploaded on the Galaxy platform<sup>76</sup> (<https://plants.usegalaxy.eu>), and trimmed with Trimmomatic V0.36.6. Reads were aligned to the *P. tricornutum* genome (Phaeodactylum\_tricornutum.ASM15095v2) with Hisat2 V2.1.0 and a gene counting was done with feature-counts V1.6.4. The alignment results showed that 64.3–80.1% of the reads were assigned to unique positions on the reference genome (Supplemental data S6).

## Differential gene expression analysis

Analyses of the differential gene expression (Pt3 Fu, Pt3 Ov, Pt3 Tr, Pt3, Pt4 or Pt8) were performed against Pt1.8.6 with DESeq2 V2.11.40.6 (which also normalized the datas) and visualized by volcano plots (DESeq2), heatmaps (Microsoft Excel) and Venn diagrams (JVenn <https://jvenn.toulouse.inrae.fr/app/example.html>). The chosen Fold Change (FC) was  $0,5 < FC > 2$  (i.e.  $-1 < \log_2 FC > 1$ ) with a  $p$  value  $< 0,05$  (i.e.  $\log_{10} p$  value  $> 1,3$ ).

## Gene ontology

Gene ontology was explored using the overrepresentation test performed through ShinyGO 0.76 (<http://bioinformatics.sdstate.edu/go/>, Pathway database: KEGG, FDR cutoff 0.05). Genes involved in protease activities were obtained from UniprotKB (keyword: KW-064 Protease) and from DiatomicBase (<https://www.diatomicsbase.io.ens.psl.eu/resources>).

## Data availability

All information on these datasets is summarized in Table 3 and a PRISMA checklist and flow diagram are available in Supplemental data S5. All data analysed during this study are included in these published articles: <https://doi.org/10.1111/nph.13787> for Pt1.8.6, <https://doi.org/10.1038/s41598-018-32519-7> for enriched Pt3, <https://doi.org/10.1038/s41598-018-23106-x> for natural Pt3, <https://doi.org/10.1093/pcp/pcx127> for Pt4, and <https://doi.org/10.1111/nph.1712> for Pt8.

Received: 20 February 2024; Accepted: 21 January 2025

Published online: 29 January 2025

## References

- Butler, T., Kapoore, R. V. & Vaidyanathan, S. *Phaeodactylum tricornutum*: A diatom cell factory. *Trends Biotechnol.* **38**, 606–622 (2020).
- Hempel, F. et al. Microalgae as bioreactors for bioplastic production. *Microb. Cell Factories* **10**, 81 (2011).
- D'Adamo, S. et al. Engineering the unicellular alga *Phaeodactylum tricornutum* for high-value plant triterpenoid production. *Plant Biotechnol. J.* **17**, 75–87 (2019).
- Fabris, M. et al. Extrachromosomal genetic engineering of the marine diatom *Phaeodactylum tricornutum* enables the heterologous production of monoterpenoids. *ACS Synth. Biol.* **9**, 598–612 (2020).
- Hempel, F. et al. From hybridomas to a robust microalgal-based production platform: Molecular design of a diatom secreting monoclonal antibodies directed against the Marburg virus nucleoprotein. *Microb. Cell Factor.* **16**, 131 (2017).
- Hempel, F., Lau, J., Klingl, A. & Maier, U. G. Algae as protein factories: Expression of a human antibody and the respective antigen in the diatom *Phaeodactylum tricornutum*. *PLoS ONE* **6**, e28424 (2011).
- Hempel, F. & Maier, U. G. An engineered diatom acting like a plasma cell secreting human IgG antibodies with high efficiency. *Microb. Cell Factor.* **11**, 126 (2012).
- Slattery, S. S. et al. Phosphate-regulated expression of the SARS-CoV-2 receptor-binding domain in the diatom *Phaeodactylum tricornutum* for pandemic diagnostics. *Sci. Rep.* **12**, 7010 (2022).
- Leprovost, S. et al. Fine-tuning the N-glycosylation of recombinant human erythropoietin using *Chlamydomonas reinhardtii* mutants. *Plant Biotechnol. J.*
- Vanier, G. et al. Biochemical characterization of human anti-hepatitis B monoclonal antibody produced in the microalgae *Phaeodactylum tricornutum*. *PLoS ONE* **10**, e0139282 (2015).
- Dumontier, R. et al. Identification of N-glycan oligomannoside isomers in the diatom *Phaeodactylum tricornutum*. *Carbohydr. Polym.* **259**, 117660 (2021).
- Kunert, R. & Reinhart, D. Advances in recombinant antibody manufacturing. *Appl. Microbiol. Biotechnol.* **100**, 3451–3461 (2016).
- Ho, S. C. L. et al. Comparison of internal ribosome entry site (IRES) and Furin-2A (F2A) for monoclonal antibody expression level and quality in CHO cells. *PLoS ONE* **8**, e63247 (2013).
- Hoseinpoor, R., Kazemi, B., Rajabibazl, M. & Rahimpour, A. Improving the expression of anti-IL-2Ra monoclonal antibody in the CHO cells through optimization of the expression vector and translation efficiency. *J. Biotechnol.* **324**, 112–120 (2020).
- Hossler, P., Racicot, C., Chumsae, C., McDermott, S. & Cochran, K. Cell culture media supplementation of infrequently used sugars for the targeted shifting of protein glycosylation profiles. *Biotechnol. Prog.* **33**, 511–522 (2017).
- O'Flaherty, R. et al. Mammalian cell culture for production of recombinant proteins: A review of the critical steps in their biomanufacturing. *Biotechnol. Adv.* **43**, 107552 (2020).
- Xu, X. et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* **29**, 735–741 (2011).
- Stolfa, G. et al. CHO-omics review: The impact of current and emerging technologies on Chinese hamster ovary based bioproduction. *Biotechnol. J.* **13**, 1700227 (2018).
- Lewis, A. M., Abu-Absi, N. R., Borys, M. C. & Li, Z. J. The use of 'Omics technology to rationally improve industrial mammalian cell line performance. *Biotechnol. Bioeng.* **113**, 26–38 (2016).
- Becker, J. et al. Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. *J. Biotechnol.* **156**, 227–235 (2011).
- Könitzer, J. D. et al. A global RNA-seq-driven analysis of CHO host and production cell lines reveals distinct differential expression patterns of genes contributing to recombinant antibody glycosylation. *Biotechnol. J.* **10**, 1412–1423 (2015).
- Jamnikar, U. et al. Transcriptome study and identification of potential marker genes related to the stable expression of recombinant proteins in CHO clones. *BMC Biotechnol.* **15**, 98 (2015).
- Clarke, C. et al. Predicting cell-specific productivity from CHO gene expression. *J. Biotechnol.* **151**, 159–165 (2011).
- Bowler, C. et al. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239–244 (2008).
- Montsant, A., Jabbari, K., Maheswari, U. & Bowler, C. Comparative genomics of the pennate diatom *Phaeodactylum tricornutum*. *Plant Physiol.* **137**, 500–513 (2005).
- Hoguin, A., Rastogi, A., Bowler, C. & Tirichine, L. Genome-wide analysis of allele-specific expression of genes in the model diatom *Phaeodactylum tricornutum*. *Sci. Rep.* **11**, 2954 (2021).
- Cruz de Carvalho, M. H., Sun, H.-X., Bowler, C. & Chua, N.-H. Noncoding and coding transcriptome responses of a marine diatom to phosphate fluctuations. *New Phytol.* **210**, 497–510 (2016).
- König, S. et al. The influence of a cryptochrome on the gene expression profile in the diatom *Phaeodactylum tricornutum* under blue light and in darkness. *Plant Cell Physiol.* **58**, 1914–1923 (2017).
- Ovide, C. et al. Comparative in depth RNA sequencing of *P. tricornutum*'s morphotypes reveals specific features of the oval morphotype. *Sci. Rep.* **8**, 1 (2018).

30. Remmers, I. M. et al. Orchestration of transcriptome, proteome and metabolome in the diatom *Phaeodactylum tricornerutum* during nitrogen limitation. *Algal Res.* **35**, 33–49 (2018).
31. Zhao, X. et al. Genome wide natural variation of H3K27me3 selectively marks genes predicted to be important for cell differentiation in *Phaeodactylum tricornerutum*. *New Phytol.* **229**, 3208–3220 (2021).
32. Chuberre, C. et al. Comparative proteomic analysis of Q12 the diatom *Phaeodactylum tricornerutum* reveals new insights into intra- and extra-cellular protein contents of its oval, fusiform, and triradiate morphotypes. *Front. Plant Sci.* **13**, (2022).
33. Feng, T.-Y. et al. Examination of metabolic responses to phosphorus limitation via proteomic analyses in the marine diatom *Phaeodactylum tricornerutum*. *Sci. Rep.* **5**, 10373 (2015).
34. Bai, X. et al. Proteomic analyses bring new insights into the effect of a dark stress on lipid biosynthesis in *Phaeodactylum tricornerutum*. *Sci. Rep.* **6**, 25494 (2016).
35. Longworth, J., Wu, D., Huete-Ortega, M., Wright, P. C. & Vaidyanathan, S. Proteome response of *Phaeodactylum tricornerutum*, during lipid accumulation induced by nitrogen depletion. *Algal Res.* **18**, 213–224 (2016).
36. Rastogi, A. et al. A genomics approach reveals the global genetic polymorphism, structure, and functional diversity of ten accessions of the marine model diatom *Phaeodactylum tricornerutum*. *ISME J.* **14**, 347–363 (2020).
37. Chaumier, T. et al. Genome-wide assessment of genetic diversity and transcript variations in 17 accessions of the model diatom *Phaeodactylum tricornerutum*. *ISME Commun.* **4**, ycad008 (2024).
38. De Martino, A., Meichenin, A., Shi, J., Pan, K. & Bowler, C. Genetic and phenotypic characterization of *Phaeodactylum tricornerutum* (Bacillariophyceae) accessions. *J. Phycol.* **43**, 992–1009 (2007).
39. De Martino, A. et al. Physiological and molecular evidence that environmental changes elicit morphological interconversion in the model diatom *Phaeodactylum tricornerutum*. *Protist* **162**, 462–481 (2011).
40. Galas, L. et al. Comparative structural and functional analyses of the fusiform, oval, and triradiate morphotypes of *Phaeodactylum tricornerutum* Pt3 strain. *Front. Plant Sci.* **12**, (2021).
41. Song, Z., Lye, G. J. & Parker, B. M. Morphological and biochemical changes in *Phaeodactylum tricornerutum* triggered by culture media: Implications for industrial exploitation. *Algal Res.* **47**, 101822 (2020).
42. Scarsini, M. et al. Metabolite quantification by fourier transform infrared spectroscopy in diatoms: Proof of Concept on *Phaeodactylum tricornerutum*. *Front. Plant Sci.* **12**, (2021).
43. Zhang, H. et al. Transcriptome aberration in marine microalgae *Phaeodactylum tricornerutum* induced by commercial naphthenic acids. *Environ. Pollut.* **268**, 115735 (2021).
44. Chen, Z. et al. Effects of tetracycline on *Scenedesmus obliquus* microalgae photosynthetic processes. *Int. J. Mol. Sci.* **23**, 10544 (2022).
45. Zhu, S. et al. Cold stress tolerance of the intertidal red alga *Neoporphyra haitanensis*. *BMC Plant Biol.* **22**, 114 (2022).
46. van Beers, M. M. C. & Bardor, M. Minimizing immunogenicity of biopharmaceuticals by controlling critical quality attributes of proteins. *Biotechnol. J.* **7**, 1473–1484 (2012).
47. Barnes, L. M. & Dickson, A. J. Mammalian cell factories for efficient and stable protein expression. *Curr. Opin. Biotechnol.* **17**, 381–386 (2006).
48. Lingg, N., Zhang, P., Song, Z. & Bardor, M. The sweet tooth of biopharmaceuticals: Importance of recombinant protein glycosylation analysis. *Biotechnol. J.* **7**, 1462–1472 (2012).
49. Stanley, P., Schachter, H. & Taniguchi, N. N-Glycans. in *Essentials of Glycobiology* (eds Varki, A. et al.) (Cold Spring Harbor Laboratory Press, 2009).
50. Lerouge, P. et al. N-Glycoprotein biosynthesis in plants: Recent developments and future trends. in *Protein Trafficking in Plant Cells* (ed. Soll, J.) 31–48 (1998). [https://doi.org/10.1007/978-94-011-5298-3\\_2](https://doi.org/10.1007/978-94-011-5298-3_2).
51. Bäiet, B. et al. N-Glycans of *Phaeodactylum tricornerutum* diatom and functional characterization of its N-acetylglucosaminyltransferase I enzyme. *J. Biol. Chem.* **286**, 6152–6164 (2011).
52. Bardor, M. et al. Immunoreactivity in mammals of two typical plant glyco-epitopes, core alpha(1,3)-fucose and core xylose. *Glycobiology* **13**, 427–434 (2003).
53. Zhang, P. et al. Characterization of a GDP-fucose transporter and a Fucosyltransferase involved in the Fucosylation of glycoproteins in the diatom *Phaeodactylum tricornerutum*. *Front. Plant Sci.* **10**, (2019).
54. Toustou, C. et al. Towards understanding the extensive diversity of protein N-glycan structures in eukaryotes. *Biol. Rev.* **97**, 732–748 (2022).
55. Foulquier, F. et al. TMEM165 deficiency causes a congenital disorder of glycosylation. *Am. J. Hum. Genet.* **91**, 15–26 (2012).
56. Reinhardt, T. A., Lippolis, J. D. & Sacco, R. E. The Ca<sup>2+</sup>/H<sup>+</sup> antiporter TMEM165 expression, localization in the developing, lactating and involuting mammary gland parallels the secretory pathway Ca<sup>2+</sup> ATPase (SPCA1). *Biochem. Biophys. Res. Commun.* **445**, 417–421 (2014).
57. Kleiner, F. H. et al. Cold-induced [Ca<sup>2+</sup>]<sub>cyt</sub> elevations function to support osmoregulation in marine diatoms. *Plant Physiol.* **190**, 1384–1399 (2022).
58. Eichler-Stahlberg, A., Weisheit, W., Ruecker, O. & Heitzer, M. Strategies to facilitate transgene expression in *Chlamydomonas reinhardtii*. *Planta* **229**, 873–883 (2009).
59. Kiefer, A. M., Niemeyer, J., Probst, A., Erkel, G. & Schroda, M. Production and secretion of functional SARS-CoV-2 spike protein in *Chlamydomonas reinhardtii*. *Front. Plant Sci.* **13**, 988870 (2022).
60. Cross, B. C. S., Sinning, I., Luirink, J. & High, S. Delivering proteins for export from the cytosol. *Nat. Rev. Mol. Cell Biol.* **10**, 255–264 (2009).
61. Müller, M. & Bernd Klösgen, R. The Tat pathway in bacteria and chloroplasts (Review). *Mol. Membr. Biol.* **22**, 113–121 (2005).
62. Kyeong, M. & Lee, J. S. Endogenous BiP reporter system for simultaneous identification of ER stress and antibody production in Chinese hamster ovary cells. *Metab. Eng.* **72**, 35–45 (2022).
63. Du, Z. et al. Non-invasive UPR monitoring system and its applications in CHO production cultures. *Biotechnol. Bioeng.* **110**, 2184–2194 (2013).
64. Hebert, D. N. & Molinari, M. In and out of the ER: Protein folding, quality control, degradation, and related human diseases. *Physiol. Rev.* **87**, 1377–1408 (2007).
65. Glickman, M. H. & Ciechanover, A. The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. *Physiol. Rev.* **82**, 373–428 (2002).
66. Bond, J. S. Proteases: History, discovery, and roles in health and disease. *J. Biol. Chem.* **294**, 1643–1651 (2019).
67. Gilgunn, S. & Bones, J. Challenges to industrial mAb bioprocessing—Removal of host cell proteins in CHO cell bioprocesses. *Curr. Opin. Chem. Eng.* **22**, 98–106 (2018).
68. Jutras, P. V. et al. An accessory protease inhibitor to increase the yield and quality of a tumour-targeting mAb in *Nicotiana benthamiana* leaves. *PLoS ONE* **11**, e0167086 (2016).
69. Dorrell, R. G. et al. Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes. *Proc. Natl. Acad. Sci. USA* **118**, e2009974118 (2021).
70. Sandberg, H. et al. Mapping and partial characterization of proteases expressed by a CHO production cell line. *Biotechnol. Bioeng.* **95**, 961–971 (2006).
71. Laux, H. et al. Degradation of recombinant proteins by Chinese hamster ovary host cell proteases is prevented by matriptase-1 knockout. *Biotechnol. Bioeng.* **115**, 2530–2540 (2018).

72. Zhang, Q. et al. Characterization of the co-elution of host cell proteins with monoclonal antibodies during protein A purification. *Biotechnol. Prog.* **32**, 708–717 (2016).
73. Leyland, B., Zarka, A., Didi-Cohen, S., Boussiba, S. & Khozin-Goldberg, I. High resolution proteome of lipid droplets isolated from the pennate diatom *Phaeodactylum tricorutum* (Bacillariophyceae) Strain pt4 provides mechanistic insights into complex intracellular coordination during nitrogen deprivation. *J. Phycol.* **56**, 1642–1663 (2020).
74. Lakeman, M. B., von Dassow, P. & Cattolico, R. A. The strain concept in phytoplankton ecology. *Harmful Algae* **8**, 746–758 (2009).
75. Toustou, C., Boulogne, I., Gonzalez, A.-A. & Bardor, M. Comparative RNA-Seq of ten *Phaeodactylum tricorutum* accessions: unravelling criteria for robust strain selection from a bioproduction point of view. *Mar. Drugs* **22**, 353 (2024).
76. Afgan, E. et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).

## Acknowledgements

The Galaxy server that was used for some calculations is partly funded by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and German Federal Ministry of Education and Research (BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi, 031L0106 de.STAIR (de.NBI)).

## Author contributions

Methodology and bioinformatics analysis, I.B. and C.T.; Data Analysis: I.B., C.T. and M.B.; writing—original draft preparation, I.B., C.T. and M.B.; writing—review and editing, I.B., C.T. and M.B.; funding acquisition, M.B. All authors have read and agreed to the published version of the manuscript.

## Funding

This work was financially supported by the French government through the ANR agency under the program «Grand défi Biomédicament: améliorer les rendements et maîtriser les coûts de production: Nouveaux Systèmes d'Expression – 2020» (PHAEOMABS project—ANR-21-F2II-0005).

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-87620-5>.

**Correspondence** and requests for materials should be addressed to I.B. or M.B.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025