Trait directed de novo population transcriptome dissects genetic regulation of a balanced polymorphism in phosphorus nutrition/arsenate tolerance in a wild grass Holcus lanatus

Trait directed *de novo* population transcriptome dissects genetic regulation of a balanced polymorphism in phosphorus nutrition/arsenate tolerance in a wild grass *Holcus lanatus* L.

Caroline Meharg[1]*, Bayezid Khan[2], Gareth Norton[2], Claire Deacon[2], David Johnson[2], Richard Reinhardt[3], Bruno Huettel[3], Andrew A. Meharg[1*]

1. Institute for Global Food Security, Queen's University Belfast, David Keir Building, Malone Road, Belfast BT9 5BN.
2. Institute of Biological and Environmental Sciences, University of Aberdeen, Cruickshank Building, St Machar Drive, Aberdeen AB24 3UU.
3. Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Köln, Germany.

* Corresponding authors

Summary

- Here we characterize the genetic regulation of a single gene balanced polymorphism for phosphate fertilizer responsiveness, as well as arsenate tolerance, in wild grass *Holcus lanatus* L. genotypes screened from the same habitat, treated with high and low phosphorus (P) as phosphate.
- *De novo* transcriptome sequencing, RNAseq and SNP calling was conducted on extracted RNA. Roche 454 sequencing data was assembled into ~22,000 isotigs and Illumina reads for phosphorous starved (P-) and phosphorous treated (P+) genovars of tolerant (T) and non-tolerant (N) phenotype were mapped to this reference transcriptome.
- Heatmaps of the gene expression data showed strong clustering of each P+/P- treated genovar, as well as clustering by N/T phenotype. Statistical analysis identified 87 isotigs to be significantly differentially expressed between N and T phenotypes and 258 between P+ and P- treated plants.
- This single gene for tolerance led to distinct phenotype transcriptomes, with large differences in post-translational and post-transcriptional regulation rather than in P nutrition transport and metabolism; and the identification of SNPs that systematically differ between phenotype, again in transcripts with predicted regulatory function including several proteases, kinases and ribonuclear RNA binding protein and concurrent up-regulation of expressed transposons in the T phenotype.

Introduction

Grasses are known to have complex genomes of various size, often large, with extensive repetitive elements, local rearrangements, differences in genomic structure, ploidy level and chromosome number, and are consequently challenging with respect to whole genome sequencing and assembly (Buckler *et al*., 2001; Feuillet & Keller, 2002; Jackson *et al.*, 2011, Hamilton & Buell, 2012). Transcriptome sequencing bypasses genomic complexity by focusing on protein coding genes (Hamilton & Buell, 2012) (~25,000 to 56,000 in grasses) derived from a well-conserved and relatively small proportion of the genome (~84% of gene families shared between grass sub-families) (The International Brachypodium Initiative, 2010). As transcripts are sensitive to environment, RNA sequencing can be clustered by function through use of appropriate experimental manipulation to further aid in gene pathway identification (Suarez Rodriguez *et al.,* 2010; Urzica *et al*., 2012; O'Rourke *et al.*, 2013).

While genomic sequencing has led to unparalleled advance in our understanding of the physiology and ecology of model organisms, it is not currently feasible (due to expense) to conduct trait led genome sequencing investigations in all wild species of interest (Nawy, 2012), rather there is a reliance in finding traits that exist (naturally or through mutation) in model species (Jackson *et al.*, 2011). While laboratory generated (site directed, knockdown, overexpression, chemical or radiation induced) mutants are essential for studying gene function and serve as breeding resource (Kuromori *et al.*, 2009), natural polymorphisms in non-model species are a greatly untapped resource.

The wild grass *Holcus lanatus* L., an outcrossing diploid (2n=14) and closely related to *B. distachyon* (Aliscioni *et al.*, 2012), has a remarkable balanced polymorphism in arsenate tolerance, screened from a semi-natural, non-arsenic contaminated populations (Meharg *et al.*, 1993), coded by a single gene (Macnair *et al.,* 1992). As arsenate is a phosphate analogue it has been postulated that this polymorphism is maintained due to phosphorus nutrition, not arsenate tolerance *per se*, particularly as the tolerance gene co-segregates with suppression of High affinity Phosphate Transport (HAPT) (Meharg *et al.,* 1992a; Meharg & Macnair, 1992b), though an explicit ecological link to phosphorus status of soils has yet to be proven (Naylor *et al.*, 1996). This soil P responsiveness is addressed in this current study along with the transcriptomic consequences of being of tolerant (T) or non-tolerant (N) phenotype to ascertain why and how this polymorphism is maintained.

Materials and methods

Chemicals used in experiments were Trace element grade or better, while all chemicals used for analytical purposes were Aristar grade.

Plant & soil collection

Single tillers of *H. lanatus* were collected from a semi-natural grassland, Cruickshank Botanic Gardens (CBG), University of Aberdeen, UK, where *H. lanatus* is a dominant species. Only one tiller was taken from each of 250 individual plants, and only isolated plants were selected, spaced at least 5 m from each other. Surface soil (0-10 cm) from the CBG population was collected and 2 mm sieved and stored at field moist conditions until use. Tillers were cultivated in a temperate glasshouse following potting (10 cm wide pots) into John Innes number 2 compost.

Plant growth characterization

Tiller testing was conducted according to procedures outlined in Macnair *et al.* (1992). Following tiller tolerance testing, genotypes were selected for further study where their longest root length after 2 wk growth in arsenate free solution exceeded 100 mm. The plants were either classified as non-tolerant (N) or tolerant (T) based on their tolerance index (TI), where TI = 100 * (root growth in 0.013 mM arsenate/root growth in absence of arsenic), with this segregation shown in Fig. 1. Out of the 250 genotypes screened, ~30 of each phenotype were selected for detailed study.

A P fertilization experiment was conducted on these ~60 genotypes, split into N and T phenotype, where each genotype was a single replicate, and where the CBG soil was fertilized with phosphate (100 P mg/kg soil d.wt. as disodium phosphate). Plants were grown in control (no fertilization) and with fertilization for 60 d before harvesting. At harvest roots were washed free of soil and root and shoot dry weight recorded along with shoot P. P was analysed in powderised shoot by Inductively Coupled Plasma – Mass Spectrometry (ICP-MS) on an Agilent 4000 instrument, following microwave assisted digestion (CEM-Technologies) in concentrated nitric acid.

RNA preparation

Tillers were grown hydroponically for 2 wk in 50 ml either complete Hoaglands solution, or that solution minus the phosphate, in individual centrifuge tubes with base of the tubes covered in tin foil to block out light. Plants were grown in a heated glasshouse (~18^0C), under supplemental sodium lamps. Roche 454 sequencing was conducted on one non-tolerant replicate grown plus (P+) or minus (P-) phosphate. A genotype of the non-tolerant phenotype was chosen for 454 transcriptome sequencing and assembly as it was assumed that the non-tolerant phenotype would be the most P responsive from previous physiological studies (Meharg & Macnair, 1992a). For Illumina sequencing a further 4 replicate individual genotypes (not including the genotype used for 454 sequencing), of each phenotype (T or N) were grown in P+ and P- in a factorial design, phenotype (tolerance) by phosphorous [P] treatment, *i.e.* 16 samples in total. The N genotype used for 454 sequencing, and another T genotype, where analysed by Illumina in the same experiment used for

454 sequencing, but where not included in the phenotyope*P Illumina analysis experiment analysis as there RNA was extracted at a different time point which lead to differential gene expression. However, these additional T and N genotype were used in SNP analysis so that SNP calling was conducted on 20 samples.

On harvesting blotted lightly dry roots and shoots for each replicate had their RNA extracted. Samples were ground under liquid nitrogen and total RNA was extracted using the Plant RNeasy extraction method (Qiagen), with the additional on-column DNase treatment. The resulting material was stored at -80^{0}C until shipping on dry ice for analysis.

Transcriptome sequencing

Sequencing was conducted at the Max Planck Plant Breeding Research Institute, Cologne, Germany. RNA was reverse transcribed to cDNA, fragmented, polyA enriched and sequenced on a 454 GS-FLX (Titanium Chemistry) and HiSeq 2000 (100 bp paired end Illumina technology).

For generation of a reference transcriptome assembly, two normalized Roche 454 libraries were prepared for one arsenic non-tolerant genovar (N) and sequenced on a Roche 454 GS-FLX using Titanium chemistry. Half a plate was used for transcriptome sequencing of phosphorous treated (N+) and half a plate for non-treated (N-) plants.

For gene expression and SNP analysis (RNAseq), a 100bp paired end Illumina sequencing library was generated for each of the 20 samples. These included all 4

samples from the first experiment (T-, N-, T+, N+), as well as another 16 samples (experiment 2) consisting of 4 tolerant (T) and 4 non-tolerant (N) genovars, receiving P- and P+ treatment (4 N-, 4 N+, 4 T-, 4 T+). Illumina reads were mapped to the 454 reference transcriptome assembly.

For annotation of the reference transcriptome with standalone BLAST (blast-2.2.22, ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/2.2.22/), the Osativa_193_transcript and Osativa_193_peptide databases were downloaded from ftp://ftp.plantgdb.org/download/Genomes/OsGDB/, the plant-refseq database from ftp://ftp.ncbi.nih.gov/refseq/release/plant/ and the nucleotide (nt) database from ftp://ftp.ncbi.nih.gov/blast/db/.

For generation of the reference transcriptome, the 454 reads were adapter and quality trimmed and assembled with Newbler version 2.6. After assembly, isotigs were annotated by blasting them against *Oryza sativa* transcript (BLASTn), *O. sativa* peptide, plant-refseq (BLASTx) and nt (BLASTn) using an e-value cutoff of 1.00 E-08. A BLAST report was compiled by parsing of all BLAST results with Perl scripts using BioPerl modules (Stajich *et al.,* 2002) to extract the top hit accession, description, E-value, percent identification (id) from each BLAST search. Annotation of putative function for *O. sativa*-peptides and *O. sativa*-transcripts were inserted into the Blast report after submission to http://rice.plantbiology.msu.edu/downloads_gad.shtml (annotation of all isotigs is reported in SI Annotated-DESeq-Result.xlsx).

For identification of differentially expressed transcripts/genes, 100bp paired end Illumina reads from experiment 1 (4 samples) and experiment2 (16 samples) were aligned to the assembled 454 reference transcriptome (isotigs) using bowtie (Langmead *et al*., 2009), allowing multiple matches (option -a) reporting only the best hits obtained for each read pair (option --best --strata) with an allowed maximum of 3 end-to-end mismatches (option –v 3) to an isotig. For each of the 20 samples the number of reported reads aligning to each isotig was counted with use of a Perl script.

Differentially expressed transcripts/isotigs were identified in R (http://www.r-project.org/) with package DESeq (Anders & Huber, 2010) using the FDR<0.1 as cutoff for significance. DESeq analysis was carried out on the Illumina data from the second experiment (n=4) for all 4 pairwise comparisons (N- *vs.* T-, N+ *vs.* T+, N- *vs.* N+, T- *vs.* T+) to allow identification of genes relevant to P+/ P- treatment and those involved in tolerant *vs.* non-tolerant phenotype. After DESeq analysis the log2 Fold Change (log2FC) of lowly expressed genes was recalculated as follows: DESeq calculated normalized counts of <5 were set to a baseline of 5. This allowed estimation of log2FC even if one of the treatments showed a mean normalized count of 0 and furthermore ensured that all 2-fold changes reported had normalized mean expression values of at least ≥10 in one of the treatments in question. Isotigs with False Discovery Rate (FDR) < 0.1 for any treatment comparison (344 isotigs) were submitted to BLAST2GO (default settings: BLASTx, nr, BLAST expect value 1.0E-03, number of BLAST hits 20) (Conesa & Götz, 2008) for further functional analysis

including Gene Ontology, Enzyme Code and Interpro domain search. The full annotated DESeq result is reported in SI Annotated-DESeq-Result.xlsx.

For identification of isotigs showing homology to specific proteins of interest proteins were blasted against all isotigs (tBLASTn). For visualisation of assigned protein homologies, some selected isotigs were translated into protein (expasy, http://web.expasy.org/translate/) and aligned with homologous plant protein sequences using mafft version7 (Katoh & Standley 2013). Mafft alignments were imported into Seaview (Gouy *et al*., 2010), and the alignments exported (SI Figs. 1-8).

Samtools was used for identification of SNPs with variant and mapping quality >20 (Li *et al*., 2009). SNP tables of all 20 plants were merged and homozygous and heterozygous SNPs consistent across N and T phenotype (n=10) were extracted as potentially relevant drivers of N versus T phenotype using a Perl script.

The data is in the process of being submitted to Short Read Archive (SRA) at the European Nucleotide Archive (www.ebi.ac.uk/ena/), to be released upon publication.

Results and Discussion

There was a clear segregation into tolerant and non-tolerant classes (Fig. 1). This was expected as previous studies have shown the ubiquity of this balanced polymorphism (Meharg *et al.* ,1993; Naylor *et al.*, 1996). When grown in their soil of origin the plants responded differentially to phosphate fertilizer (Fig. 2). General Linear Modeling (GLM) (using Minitab (v.16)) of ranked data (due to non-normality of untransformed data) found a significant (P=0.004) phenotype*P fertilization interaction for shoot/root ratio, all other model terms were not significant for this comparison. This phenotype*P fertilization interaction term was caused by a decrease in shoot/root ratio, in tolerants treated with P, showing that the tolerant phenotype is not reducing relative root growth in response to P nutrition. Plants normally reduce relative root production under P nutrition (Gojon *et al.*, 2009). This is the first characterization of a phosphate specific response for this polymorphism and the differential allocation to root and shoot biomass in response to P availability likely to be the reason why this polymorphism is maintained, and this will explained and outlined in more detail in subsequent publications. There were no significantly different (*i.e.* P>0.005) terms for both shoot weight and shoot P, while only the treatment term was significant for root weight. These data show that shoot P (root P was not measured as it is impossible to remove all adhering soil which greatly confounds interpretation) does not differ between phenotypes, and that this shoot P status is not P fertilizer responsive, all indicating tight homeostasis, a known characteristic with respect to plant P nutrition (Gojon *et al.*, 2009; Hill *et al.*, 2006).

Roche 454 sequencing generated ~1 million reads and a total of 474 megabases (MB) of sequence data for assembly of a reference transcriptome. *Circa.* 82% of all reads and ~85% of all bases (400MB) aligned. The inferred read error was 1.06% (Supporting Information [SI] Illumina-and-454-Stats.xlsx). Assembly with Newbler 2.6 generated a reference transcriptome (SI 454Isotigs.fna) with a total of 22,313 isotigs. The overall size of the assembled reference transcriptome was 29 megabases (MB). The average isotig size obtained was 1,302bp, the N50 Isotig size 1,489bp, the number of isotigs ≥1kb was 12,828, (SI Illumina-and-454-Stats.xlsx, Sheet1). When BLASTed against *O. sativa*, plant refseq transcriptome databases and non-redundant nucleotide database (nt) (SI Annotated-DESeq-Result.xlsx)). Of these, 18,204 returned a match against *O. sativa* transcripts, 18,954 returned a match against *O. sativa* peptides, 19,344 against plant refseq and 19,589 against nt (cutoff for significance < 1.00 E-08). BLAST against nt returned hits almost exclusively against plant cDNA/mRNA, predominantly *Hordeum vulgare* (barley), *Triticum aestivum* (wheat). Highest homologous matches against Plant-refseq were invariably identified against protein/transcript sequences of *Brachypodium*, *Sorghum bicolor* and *O. sativa* (rice), of these *Brachypodium* being the most frequent hit and most closely related to *H. lanatus* (Aliscioni *et al.*, 2012). Further statistics on assembly are reported in the SI Illumina-and-454-Stats.xlsx. The only previously reported *H. lanatus* gene sequence, AY704470, a CDC25 phosphatases homologue (Bleeker *et al.,* 2006), was identified in all genotypes with isotig19077 showing 100% identity to this published sequence (tBLASTn) (SI Fig. 1), giving further verification of the transcriptome assembly.

Paired end Illumina sequencing produced an average of ~53 million reads for each of the 20 libraries giving a total of 1052 million 100bp paired end reads or 210 gigabases (GB) of sequence data. Forty-eight percent (100GB) of the HiSeq data aligned successfully to the assembled 454 transcriptome without any prior clipping of reads, with an average of 25 million full length 100bp paired end reads (5GB) for each of the 20 libraries mapping successfully to isotigs with 3 or less end-to-end basepair mismatches (SI Illumina-and-454-Stats.xlsx). Of 22313 isotigs only 52 isotigs did not obtain any mapped paired end Illumina reads when aligning all Illumina reads from (all 20 individual samples) giving good verification of the 454 reference transcriptome assembly (SI RawCounts-per-isotig.xlsx). The isotigs, which did not obtain any mapped Illumina reads may either have been too short (less than 250bp) or misassembled. Non-aligned Illumina reads on the other hand are likely to be a mixture of poor quality reads, reads containing adapter sequence, which would require end-trimming, or alternatively originate from lowly expressed transcripts, which may not be represented in the 454 reference library, but may still be picked up amongst the Illumina sequences due to the much deeper coverage achieved with Illumina technology. Transcript expression result showed strong clustering of each P+ and P- treated genotype verifying that each plant was indeed a different genotype (Fig. 3). T genotypes clustered together, as did N, showing that the phenotypes had distinctive transcript expression signatures. Differential expression analysis was carried out with DESeq (Anders & Huber, 2010) and identified 344 isotigs, with a False Discovery Rate (FDR) of 0.1, significantly up or down-regulated in response to either P+/P- treatment or N/T phenotype (Fig. 4, SI Annotated-DESeq-Result.xlsx). Of these 87 isotigs were shown to be differentially expressed between N

and T phenotype, while the majority, 258 isotigs, were shown to be differentially expressed in response to different P± nutrition treatment. There was no overlap between the significant isotigs identified for P+/- and N/T phenotype response. There are a number of potential explanations for the independence of transcripts between N/T and P+/P- nutrition. The first is that a single regulatory gene, given that we know that arsenate tolerance is under single gene control (Macnair *et al.*, 1992), is controlling transcript production of a host of genes in the T/N comparison. The second, which is not exclusive of the first, is that differences in metabolism resulting from differential function of a gene(s) may lead to feedback regulating transcripts of interrelated functions such as the obvious impact of P starvation on P metabolism observed here (Figs. 3 and 4). Differences in P stress perception by plants are known to induce a host of differential responses, such as tillering, root biomass production, arbuscular mycorrhizal regulation, and rhizosphere excretion of dicarboxylic acids to mobilize phosphate from iron minerals, and this will lead to differential regulation of a network of genes (Chiou *et al.*, 2011; Gojon *et al.*, 2009; Hill *et al.*, 2006). Again, the arsenate tolerance gene has a range of pleiotropic consequences (shoot/root biomass, HAPT suppression and arsenate tolerance itself), fitting well such a feedback and/or upstream regulator model.

The amplitude of fold-change was greater between phenotypes than between P treatments, while annotation was better for P responsive genes (Fig. 4). Annotated genes in classes that are highly relevant to the current study show that transcripts significantly differentially regulated between T and N phenotype are dominated by kinases, pathogen resistance, plant growth regulators (PGR), proteases, transposable

elements (TEs) and RNA directed activity, but none involved in phosphate transport (Annotated-DESeq-Results.xlsx, Fig. 4). It has previously been shown that one consequence of having the arsenate tolerance gene is suppression of high affinity phosphate/arsenate transport (Meharg *et al.*, 1992a,b). The results presented here indicate, therefore, that post-transcriptional and post-translational mechanisms play a key role in this suppression, as there was no differential expression of transcripts involved in phosphate transport between phenotype.

With respect to T phenotype the only annotated gene absent compared to N, where it is highly expressed, is a kinase receptor (isotig09647). Furthermore, isotigs with significant homology to cbl-interacting kinase, MAPK kinase and serine/threonine kinases had systematic differences in SNPs between N and T phenotype as did isotigs showing homology to proteasome associated protein, transferases and a ribonucloprotein/RNA recognition protein (Fig. 5, SI Table 1).

A gene that codes for rice adaption to soil P stress, PSTOL1, has been characterized (Gamuyao *et al.,* 2012). This gene is an enhancer of early root growth and over expression leads to increased grain yields, hypothesized to be due to more efficient P capture due to larger root systems, with larger root systems characterizing the *H. lanatus* T phenotype here (Fig. 2). This gene is a kinase and some homologous sequences were identified in *H. lanatus*, such as isotig20112 which showed ~ 88% (identities 145/164) homology to the serine/threonine protein kinase LOC_Os01g04570.2 and 63% (identities 108/169) to Pstol1/OsPupK46-2, also annotated as serine/threonine protein kinase (SI Fig. 2) and isotig03216, which

showed 50% (identities 158/312) to protein serine/threonine kinase LOC_Os01g04570.2 and 49% (identities 154/313) to Pstol1/OsPupK46-2. Isotig20112 was expressed in all T and N phenotypes but was ~4 fold down-regulated in 2 out of 4 N phenotypes, with no P effect. isotig03216 was again expressed in all N and T phenotypes, but N5, N4, N2 as well as T4 showed ~>3 fold lower expression compared to T2, T3, T5, N3, but the overall observed ~2 fold up-regulation in T- versus N- was not statistically significant (SI AnnotatedDeseqResult.xlsx).

Other transcripts belonging kinases, some of these showing homology to cbl-interacting kinases 9, 14 & 23, were found to be up-regulated under low P status (Fig. 4, SI AnnotatedDeseqResult.xlsx). Cbl-interacting kinases are serine/threonine protein kinases, as is PSTOL1 (Gamuyao *et al*., 2012). The only transcript absent in T compared to N phenotype (present in 3 out of 5 N phenotypes and absent in all 5 T phenotypes), isotig09647, noting that there was also a transposon severely suppressed, was a receptor protein kinase. Another probable serine/threonine protein kinase WNK2-like, isotig18018, was ~60 fold up-regulated in T compared to N. It is apparent that kinases play a role in both T/N phenotypic and P responsive differences in the *H. lanatus* transcriptomes presented here and that there is possibly some level functional redundancy. It is pertinent in this context that some kinases have recently been identified as being central in phenotypic differences in plant root response to P status (Gamuyao *et al.,* 2012) as well as shown to be up regulated in response to arsenate stress (Huang *et al.,* 2012).

Also of note is that an auxin binding protein (isotig16840) was highly expressed in the T phenotype compared to the N phenotype. The alignment of the translated isotig16840 is shown in SI Fig. 3 and shows a high homology to auxin binding rice and *Brachypodium* proteins. One of the key differences between T and N phenotype identified here was in root biomass (Fig. 2). Auxins, and associated expansins, have a key role in root growth (Cosgrove, 1999). The higher expression of an auxin-binding protein with decreased expression of an expansin precursor transcript in T phenotype, compared to N, and the enhanced expression of an auxin responsive protein (isotig11028) under high P (SI AnnotatedDeseqResult.xlxs) suggest that auxin signaling is central to the difference in root biomass regulation between T and N. The alignment of translated isotig12721 with expansin precursor can be seen in SI Fig. 4.

While most expressed transposons, some of these annotated with RNA directed activity, were up regulated in T compared to N (isotig13038, 13187, 13075, 15546, 07417), one transposon of the en spm subclass (isotig02887/02888) and one unclassified nucleic acid binding transposon (isotig05913) were suppressed in T compared to N, (SI AnnotatedDeseqResult.xlxs). Transposons and retrotransposons, the most frequent and strongly differentially expressed class of isotigs between phenotypes, are thought to play a role in post-transcriptional regulation with silencing of TEs involving both transcriptional and post-transcriptional mechanisms (Mirouze & Paszkowski, 2011; Okamoto & Hirochika, 2001). In plants, retrotransposons are commonly known to be expressed under conditions of stress (Grandbastien, 1998). It is also thought that retrotransposon activation is sensitive to environment (Mirouuze & Paszkowski, 2011; Grandbastien, 1998), further enhancing

their candidacy for regulating stress responses, such as nutritional deficiencies. Transposable Elements (TEs) can evoke gene variation and functional changes (Gao *et al*., 2012) and are a source of small RNAs and implicated in gene-regulation in both animals and plants (McCue & Slotkin, 2012). A role for small RNAs in regulation of P starvation is emerging (Fang *et al.*, 2009; Hsieh *et al.*, 2009; Chiou & Lin, 2011; O'Rourke *et al.*, 2013). Further to this, a potential role of a small RNA targeting transcripts involved in posttranscriptional/posttranslational regulation leading to T and N phenotype in *H. lanatus* is worth further investigation.

A suite of isotigs involved in protein degradation including two with homology to aspartic proteinase nepenthesin precursor (isotig17128, isotig10719), one FtsH protease (03772) and a ubiquitination like protein (isotig20448) were suppressed in T compared to N phenotype, indicating that post-translational protein degradation may be an important factor for the N phenotype. With respect to P treatment, ubiquitin-domain protein (isotig17019), protease inhibitor (isotig19161), putative subtilisin homologue, a non-specific protease (isotig03645) were significantly up regulated in response to P starvation, while ubiquitin conjugating enzyme like (isotig08177), ICE-like protease p20 domain containing protein (isotig10848), putative Deg protease homologue (isotig09222) and LTPL113 - Protease inhibitor (isotig19161) were significantly down regulated in response to P starvation.

Furthermore, 6 consistent homozygous and 30 consistent heterozygous SNPs for N versus T phenotype (n=10) were identified (Fig. 5). These occurred predominantly in transcripts associated with regulatory function such as proteolysis (protease,

proteasome subunit, heat shock), protein modification (cysteine desulferase, glycosyl transferase, transferase, kinase) and RNA recognition (RNA-binding, ribonucleoprotein) and were a mixture of transitions and transversion SNPs (Fig. 5). Ubiquitins, which mark proteins for proteasome mediated degradation are thought to have a key role in regulation of plant SPX domains in response to P stress (Wu *et al.,* 2012), so it is particularly interesting that we find that the largest number of SNPs (3 transversion and 1 transition SNP) between T and N in an putative proteasome subunit (isotig11038) as well as SNPs in FtsH protease (isotig03772) and heat shock protein (isotig03795/03796). Also while we assume that these SNPs are of genomic origin, without the genomic sequence of *H. lanatus* it is not possible to rule out that targeted mRNA editing may be involved in some of these cases. RNA-editing, which has first been identified in the cox2 mRNA of *Trypanosoma brucei*, is thought to play an important role in organelles (plastids and mitochondria) of plants, with those identified typically involving a change of specific C to U, but other changes can as yet not be ruled out (Grennan, 2011; Jiang *et al*., 2012). The role of RNA-editing in plant plastid as well as nuclear encoded RNA/mRNA, remains to be further investigated by systematic sequencing of plant genome (DNA) and transcriptome (cDNA) as has been described for identification of RNA editing sites in human studies (Ramaswami *et al*., 2012). While RNA-binding proteins of the pentatricopeptide repeat family, multiple organellar RNA editing factor and chloroplast ribonucleoproteins are known to be involved, further proteins remain to be identified (Tillich *et al*., 2009; Grennan, 2011; Takenaka *et al*., 2012). So it is noteworthy in this context that one of the homozygous SNPs identified between T versus N phenotype in this study is in isotig05374 which shows 70% homology to a

RNA recognition motif containing protein/predicted ribonucleoprotein. Further to that, exonuclease (isotig08248), which mediates RNA degradation (Stoppel *et al.,* 2012), is strongly up regulated in the N phenotype. Thus both the N/T gene expression result as well as the SNPs obtained for N/T phenotype suggest that post-translational regulation of proteins via the ubiquitin-proteasome system plays an important role in determining the N and T phenotype and furthermore points to a potential role of post-transcriptional regulation (RNA degradation and a possible role of RNA editing). A causative upstream master regulatory gene, inducing post-translational and maybe also post-transcriptional events of consequence for arsenate resistance and P uptake efficiency in these plants remains to be identified and potential involvement of small RNAs should be investigated in this context.

The transcripts differentially regulated by P treatment were more completely annotated compared to between phenotype (SI AnnotatedDeseqResult.xlxs), with many of the genes identified as being significant having well known roles in P transport and metabolism, as well as those involved in post-transcription, post-translation, and signaling (Fig. 4, SI Annotated-DESeq-Results.xlsx). Some gene expression responses involved phosphate transport. These include 3 isotigs annotated as phosphate co-transporters and 2 with SPX domains. All showed the same general pattern, with respect to transcript counts, with up-regulation under P starvation, and down-regulation in P replete treatment, but there was little difference with respect to the expression changes in response to P-/P+ in the T and N phenotype. Alignments of phosphate transport translated isotigs to proteins are shown in SI Fig. 5 & 6. Isotigs01092, 18981, 09507 showed strong homology to

AAM49810.1 a putative rice HAPT (SI Fig. 5), and isotig16690 to Q651J5 phosphate transporter PHO1-3 known to be induced under P deficient conditions (SI Fig. 6). Of the few isotigs with significantly higher counts in P+ compared to P- in T, but not in N, was an auxin induced protein (isotigs11028). Given the strong induction in an auxin-binding transcript (isotig16840) in the T versus N phenotype and the observed differences in root growth, auxin mediated response appears to be an important differentiator of the T/N phenotype. The rice and *Brachypodium* matches to the translated isotig16840 protein sequence are shown in SI Fig. 3.

A number of recent studies have found up-stream regulators of low phosphorus adaption responses in plants with ALFIN—LIKE proteins implicated in regulating root hair growth in Arabidopsis under P stress (Chandrika *et al.,* 2013). ALFIN-LIKE proteins are a small family of plant Homeo Domain (PHD) containing putative transcription factors with a methylated histone residue binding component and ALFIN-LIKe 6 was shown to control the transcription of a range of genes involved in growth in particular root hair growth (Chandrika *et al.,* 2013). A homologue of this gene was identified in our *H. lanatus* transcriptome (isotig02053) and the translated isotig showed high similarity to *A. thaliana* and even higher similarity to *B. distachyon* ALFIN-LIKE protein (SI Fig. 8), providing further evidence of the overall quality of the 454 transcriptome assembly, though this transcript was equally highly expressed in all N and T plants and not differentially regulated between either P treatment or phenotype.

Many of the transcripts responsive to P nutrition are regulated as expected, with many obviously involved in P nutrition and generally strongly up regulated in response to P stress. For example the 4 phosphate transport related transcripts, and the 2 SPX transcripts, identified in *H. lanatus* in this study were induced by P stress, in line with what has been observed generically in literature studies, as exemplified in the phosphate responsive transcriptome of white lupin (O'Rourke *et al.*, 2013). Other transcripts coding for proteins involved in P regulation or transport identified (SI Annotated-DESeq-Result.xlsx) include glycerophosphodiester/phosphodiesterase, sucrose-phosphate synthase, glucose pyrophosphorylase, purple acid phosphatase, phosphatase, glucose-6-phosphate phosphate translocator, for example, also differentially regulated when comparing P stressed and P repleat lupin (O'Rourke *et al.*, 2013).

With respect to arsenate tolerance, the character used to screen the phenotype understudy, one gene with a putative role in arsenic transport/metabolism, besides phosphate transporters and their regulators, was an arsB-like gene. The alignment of isotig09604 with ars-B like protein can be seen in SI Fig. 9. ArsB, is widely present in arsenic resistant bacteria where its role is as an arsenite efflux channel (Yang *et al.*, 2012), though in plants these class of aqua-glycerin prions also are involved in silicic acid transport (Zhao *et al.*, 2009). This gene (isotig09604) was more strongly expressed in T compared to N phenotype as was aquaporin Pip1-5 like (Plasmamembrane Intrinsic Protein) (isotig00796/00797/00798), but this is only involved in transport of water and small neutral solutes. Logoteta *et al.* (2009) found that differential efflux of arsenite by *H. lanatus* was not found between T and N

phenotype. A protein, HLASR, has been identified in *H. lanatus* to have a constitutive, but not an adaptive role, in metabolism of arsenate, namely as an arsenate reductase (Bleeker *et al.* 2006). It was thought that this HLASR gene only had a secondary role with respect to arsenic metabolism and that its primary role was homologous to CDC25 phosphatases, which activate cycline-dependent kinases in *A. thaliana*, which are involved in cell cycle regulation. HLASR is also thought to have a role in GSH oxidation (Bleeker *et al.,* 2006). An exact protein match (isotig19077) to this Cdc25-like *H. lanatus* ASR, the only gene sequence previously published for this species, was found in all 10 N and 10 T transcriptomes (SI Fig. 3), and was shown not to be differentially regulated, confirming that it has no adaptive role in tolerance.

Returning to the fact that the trait under study here is due to a single *loci* and that it leads to a range of distinct phenotypic traits is indicative that an upstream regulatory gene is involved. It is now well established that P responsive genes a) give rise to transcriptome signal cascades and/or b) are involved in these signal cascades (*Wu et al.* 2013; Chiou & Lin, 2011). We identified 87 transcripts whose transcript expression significantly differed between T and P phenotype, and 19 transcripts (17 with functional annotation) with consistent SNPs (36 SNPs in total) in all 10 T versus N genotypes (SI Table 1). It is noteworthy that consistent SNPs and significant gene expression changes between N versus T phenotypes are aggregated in transcripts with regulatory functions (Fig.s 4 and 5). Differential post-translational and/or post-transcriptional regulation (involving ubiquitin, proteases, kinases, methylation, transponsons, retrotransposons and RNA binding proteins) of HAPT or proteins

acting on HAPT, therefore, appear to define the N/T phenotype. Whether identified SNPs are all genomic SNPs or whether in some cases mRNA editing may be involved remains to be elucidated. A master regulatory gene, potentially to be found within our list of target genes or possibly in form of an as yet to be identified small RNA, leading to the observed effect on genes involved in post-translational events including protein degradation via the ubiquitin/proteasome complex and potentially also post-transcriptional events mediated by RNA binding proteins, is as yet to be identified. This characterization of the genetic consequences the P response polymorphism in *H. lanatus* provides an unparalleled insight into the signal cascades, optimized under natural selection, involved in P nutrition and has major consequences for understanding how plants respond to phosphorus nutrition and adaptation to arsenate in their environment. We anticipate that this as yet unknown master regulatory gene and it's downstream targets, which we have already identified, will be of significant consequence for future study and breeding of P-efficient forage plants and cereal crops.

References

Aliscioni S, Bell HL, Besnard G, Christin P-A, Columbus JT, Duvall MR, Edwards EJ, Giussani L, Hasenstab-Lehman K, Hilu KW, Hodkinson TR, Ingram AL, Kellogg EA, Mashayekhi S, Morrone O, Osborne CP, Salamin N, Schaefer H, Spriggs E, Smith SA, Zuloaga F. 2012. New grass phylogeny resolves deep evolutionary relationships and discobers C4 origins. *New Phytologist*: 193, 304-312.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biology*: 11, R106.

Bleeker PM, Hakvoort HWJ, Bliek M, Souer E, Schat H. 2006. Enhanced arsenate reduction by a CDC25-like tyrosine phosphatase explains increased phytochelatin accumulation in arsenate-tolerant *Holcus lanatus*. *Plant Journal*: 45, 917-929.

Buckler ES, Thornsberry JM, Kresovich S. 2001. Molecular diversity, structure and domestication of grasses. *Genetics Research Cambridge*: 77, 213-218.

Chandrka NNP, Sundarevelpandian K, Yu S-M, Schmidt W. 2013. ALFIN-LIKE 6 is involved in root hair elongation during phosphate deficiency in Arabidopsis. *New Phytologist*: advanced online, doi:10.1111/nph.12194.

Chiou TJ, Lin SI. 2011. Signaling network in sensing phosphate availability in plants. *Annual Review Plant Biology*: 62, 185-206.

Conesa A, Gotz S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal Plant Genomics*: Article ID 619832.

Cosgrove DJ. 1999. Enzymes and other agents that enhance cell wall extensibility. *Annual Review Plant Physiology Plant Molecular Biology*: 50, 391-417.

Fang ZY, Shao C, Meng YJ, Wu P, Chen M. 2009. Phosphate signaling in *Arabidopsis* and *Oryza sativa*. *Plant Science*: 176, 170-180.

Feuille C, Keller B. 2001. Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution. *Annals Botany*: 89, 3-10.

Gamuyao R, Chin JH, Pariasca-Tanaka J, Catausan S, Dalid D, Slamet-Loedin I, Tecson-Mendoza EM, Wissuwa M, Heur S. 2012. *Nature*: 488. 535-541.

Gao C, Xiao M, Jiang L, Li J, Yin J, Ren X, Qian W, Oscar O, Fu D, Tang Z. 2012. Characterization of transcriptional activation and inserted-into-gene preference of various transposable elements in the Brassica species. *Molecular Biology Reports*: 39, 7513-7523.

Gojon A, Nacry P, Davidian JC. 2009. Root uptake regulation: a central process for NPS homeostasis in plants. *Current Opinion Plant Biology*: 12, 328-338.

Grandbastien MA. 1998. Activation of plant retrotransposons under stress conditions, *Trends Plant Science* 3: 181-187.

Grennan AK. 2011. To thy proteins be true: RNA editing in plants. *Plant Physiology*: 156, 453-454.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology Evolution:* 27, 221-224.

Hamilton JP, Buell CR. 2012. Advances in plant genome sequencing. *The Plant Journal*: 70, 177-190.

Hill JO, Simpson RJ, Moore AD, Chapman DF. 2006. Morphology and response of roots of pasture species to phosphorus and nitrogen nutrition. *Plant Soil*: 286, 7-19.

Hsieh LC, Lin SI, Shih ACC, Chen JW, Lin WY, Tseng CY, Li WH, Chiou TJ. 2009. Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. *Plant Physiology*: 151, 2120-2132.

Huang TL, Nguyen QT, Fu SF, Lin CY, Chen YC, Huang HJ. 2012. Transcriptomic changes and signaling pathways induced by arsenic stress in rice roots. *Plant Molecular Biology*: 80, 587-608.

Jackson SA, Iwata A, Lee SH, Schmutz J, Shoemaker R. 2011. Sequencing crop genomes: approaches and applications. *New Phytologist*: 191, 915-925.

Jiang Y, Fan SL, Song MZ, Yu JN, Yu SX. 2012. Identification of RNA editing sites in cotton (*Gossypium hirsutum*) chloroplasts and editing events that affect secondary and three-dimensional protein structures. *Genetics Molecular Research*: 11, 987-1001.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:improvements in performance and usability. *Molecular Biology Evolution*: 30, 772-80.

Kuromori T, Takahashi S, Kondou Y, Shinozaki K, Matsui. 2009. Phenome analysis in plant species using loss-of-function and gain-of-function metants. *Plant Cell Physiology*: 50, 1215-1231.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*: 10, R25.

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science*: 324, 1210-1213.

Logoteta V, Xu XY, Macnair MR, McGrath SP, Zhao FJ. 2009. Arsenite efflux is not enhanced in the arsenate-tolerant phenotype of *Holcus lanatus*. *New Phytologist*: 183, 340-348.

Macnair MR, Cumbes QJ, Meharg AA. 1992. The genetics of arsenate tolerance in Yorkshire Fog, *Holcus lanatus* L. *Heredity*: 69, 325-335.

McCue AD, Slotkin RK. 2012. Transposable element small RNAs as regulators of gene expression. *Trends Genetics*: 28, 616-623.

Meharg AA, Cumbes QJ, Macnair MR. 1993. Pre-adaptation of *Holcus lanatus* L. to arsenate tolerance. *Evolution*: 47, 313-316.

Meharg AA, Macnair MR. 1992a. Suppression of the phosphate uptake system: A mechanism of arsenate tolerance in *Holcus lanatus* L. *Journal Experimental Botany*: 43, 519-524.

Meharg AA, Macnair MR. 1992b. Genetic correlation between arsenate tolerance and the rate of arsenate and phosphate uptake in *Holcus lanatus* L. *Heredity*: 69, 336-341.

Mirouze M, Paszkowski J. 2011. Epigenetic contribution to stress adaptation in plants. *Current Opinion Plant Biology*: 14, 267-274.

Nawy T. 2012. Non-model organisms. *Nature Methods*: 9, 37.

Naylor J, Macnair MR, Williams END, Poulton PR. 1996. A polymorphism for phosphate/arsenate tolerance in *Holcus lanatus* L.: is their a correlation with edaphic or environmental factors? *Heridity*: 77, 509-517.

O'Rourke JA, Yang SS, Miller SS, Bucciarelli B, Liu J, Rydeen A, Bozsoki Z, Uhde-Stone C, Zheng JT, Allan D, Gronwald JW, Vance CP. 2013. An RNA-Seq transcriptome analysis of othophosphate-deficient white lupin reveals novel insights into phosphorus acclimation in plants. *Plant Physiology*: 161, 705-724.

Okamoto H, Hirochika H. 2001. Silencing of transposable elements in plants. *Trends Plant Sciences*: 6, 527-534.

Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB. 2012. Accurate identification of human Alu and non-Alu RNA editing sites. *Nature Methods*: 9, 579-81.

Takenaka M, Zehrmann A, Verbitskiy D, Kugelmann M, Härtel B, Brennicke A. 2012. Multiple organellar RNA editing factor (MORF) family proteins are required for RNA editing in mitochondria and plastids of plants. *PNAS*: 109, 5104-5109.

Stajich JE, Block D, Boulez K, Brenne SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Kork I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR,

Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. 2002. The Bioperl Toolkit: Perl modules for life sciences. *Genome Research:* 12, 1611-1618.

Stoppel R, Meurer J. 2012. The cutting crew - ribonucleases are key players in the control of plastid gene expression. *Journal Experimental Botany*: 63, 1663-1673.

Suarez Rodriguez MC, Edsgard D, Hussain SS, Alquezar D, Rasmussen M, Gilbert T, Nielsen BH, Bartels D, Mundy J. 2010. Transcriptome of the dessication-tolerant resurrection plant *Craterostigma plantagineum*. *Plant Journal*: 63, 212-228.

The International Brachypodium Initiative. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*: 463, 763-768.

Tillich M, Hardel SL, Kupsch C, Armbruster U, Delannoy E, Gualberto JM, Lehwark P, Leister D, Small ID, Schmitz-Linneweber C. 2009. Chloroplast ribonucleoprotein CP31A is required for editing and stability of specific chloroplast mRNAs. *PNAS*: 7106, 6002-6007.

Urzica EI, Yamassaki H, Hsieh SI, Adler LN, Karpowicz SJ, Blaby-Hass CE, Clarke SG, Loo JA, Pellegrini M, Merchant SS. 2012. Systems and *trans*-system level analysis identifies conserved iron deficiency resonponses in the plant lineage. *Plant Cell*: 24, 3921-3948.

Wu P, Shou H, Xu G, Lian X. 2013. Improvement of phosphorus efficiency in rice on the basis of understanding phosphate signaling and homeostasis. *Current Opinions Plant Biology*: in press: http://dx.doi.org/10.1016/j.pbi.2013.03.002

Yang H-C, Fu H-L, Lin Y-F, Rosen BP. 2012. Pathways of arsenic uptake and efflux. *In* Metal Transporters *eds.* Lutsenko S and Arguello. Elsevier Inc..

Zhao FJ, Ma JF, Meharg AA, McGrath SP. 2009. Arsenic uptake and metabolism in plants. *New Phytologist*: 181, 777-794.

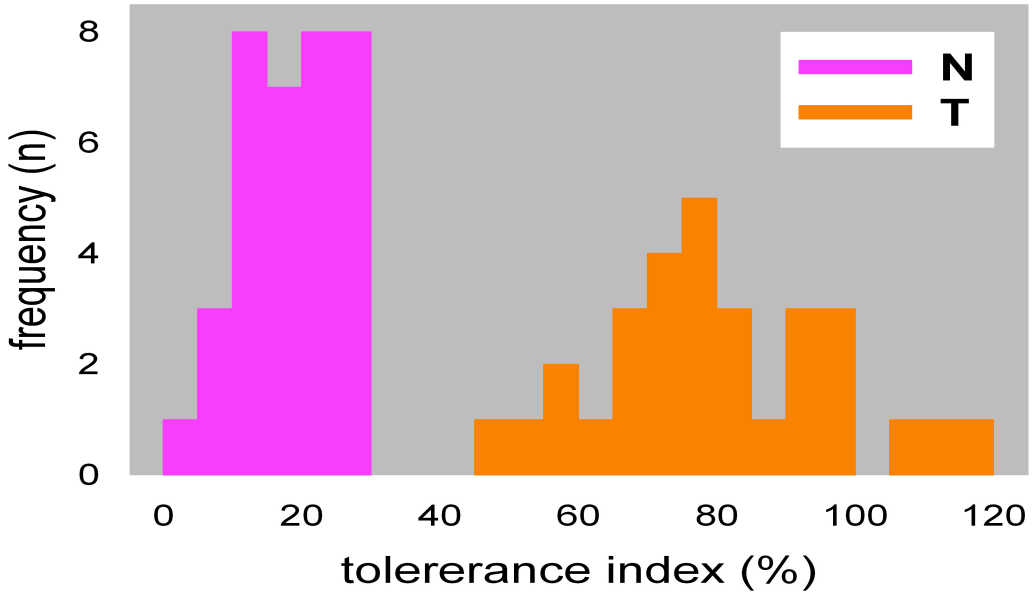

Figure 1. Tolerance index (100*root length in +As/root length in –As) of Cruickshank garden *H. lanatus* population grown hydroponically in 0.013 mM arsenic as arsenate for 2 wk..

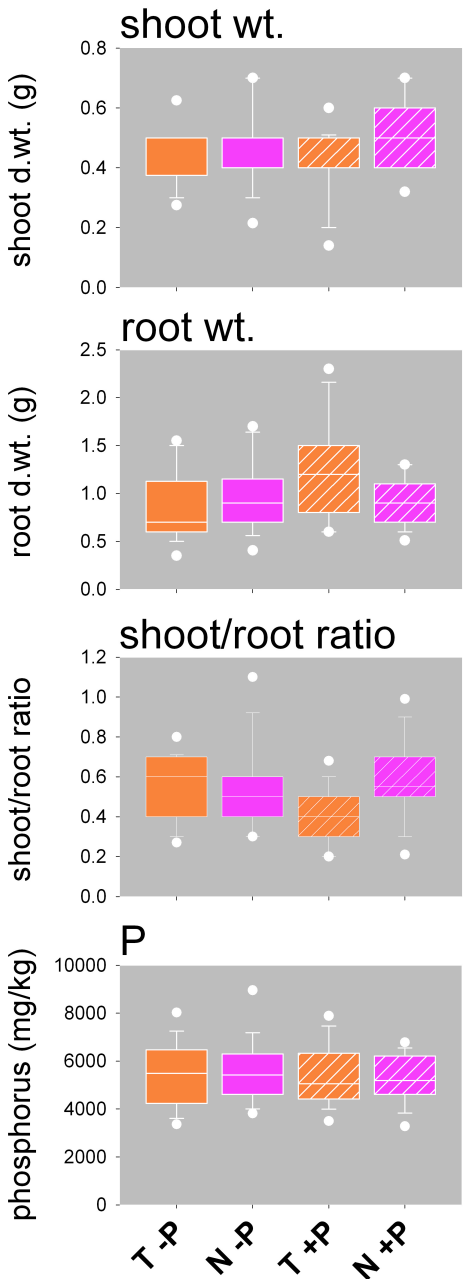

Figure 2. Box plots of Cruickshank *H. lanatus* response to phosphate fertilization (100 mg/kg P on a soil d.wt. basis) when grown from tillers for 2 mth. on their soil of origin. Arsenate tolerant plants are orange bars, non-tolerant pink bars. No P fertilization is without shading while P fertilized plants are shaded. Shown on the bars are the median, 25th & 27th (outer box), 10th and 90th (whiskers) and 5th and 95th (dots) percentiles.

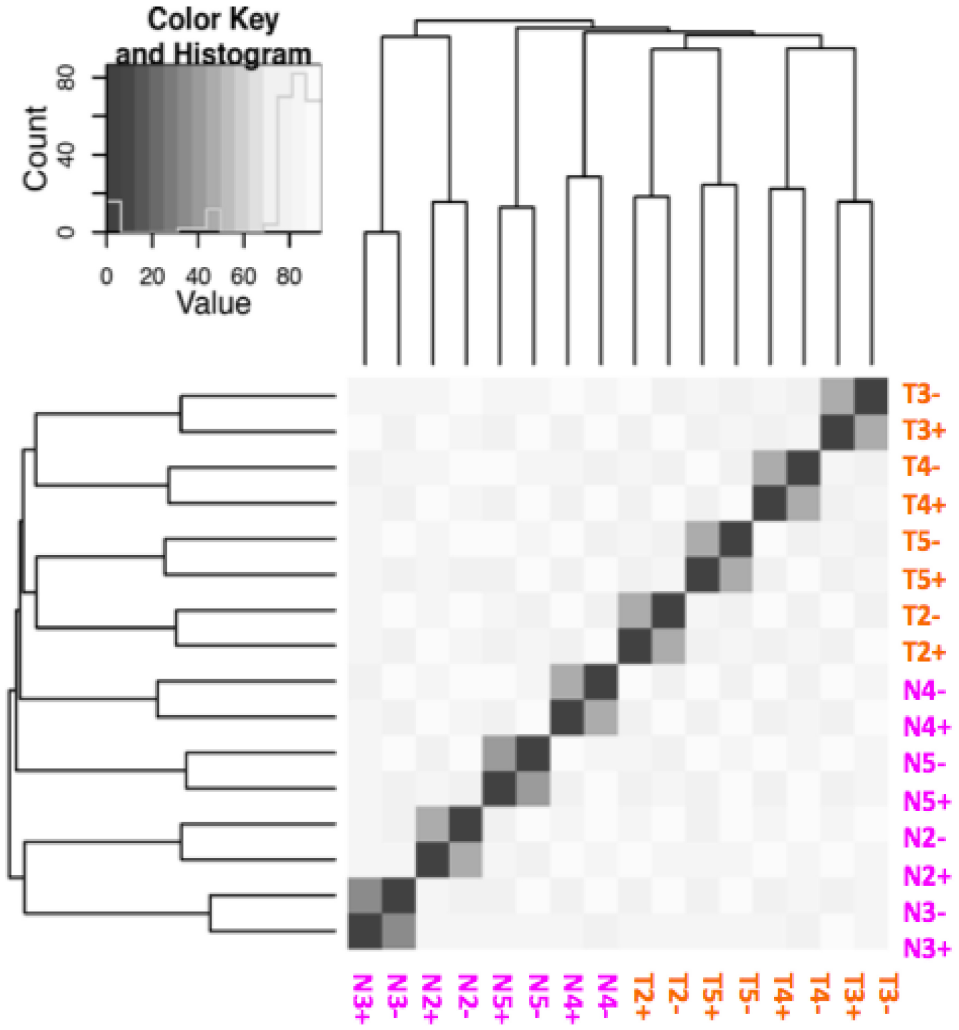

Figure 3: Heatmap of the RNAseq gene expression result showing clustering of P+/- of each genotytpe as well as T and N phenotype

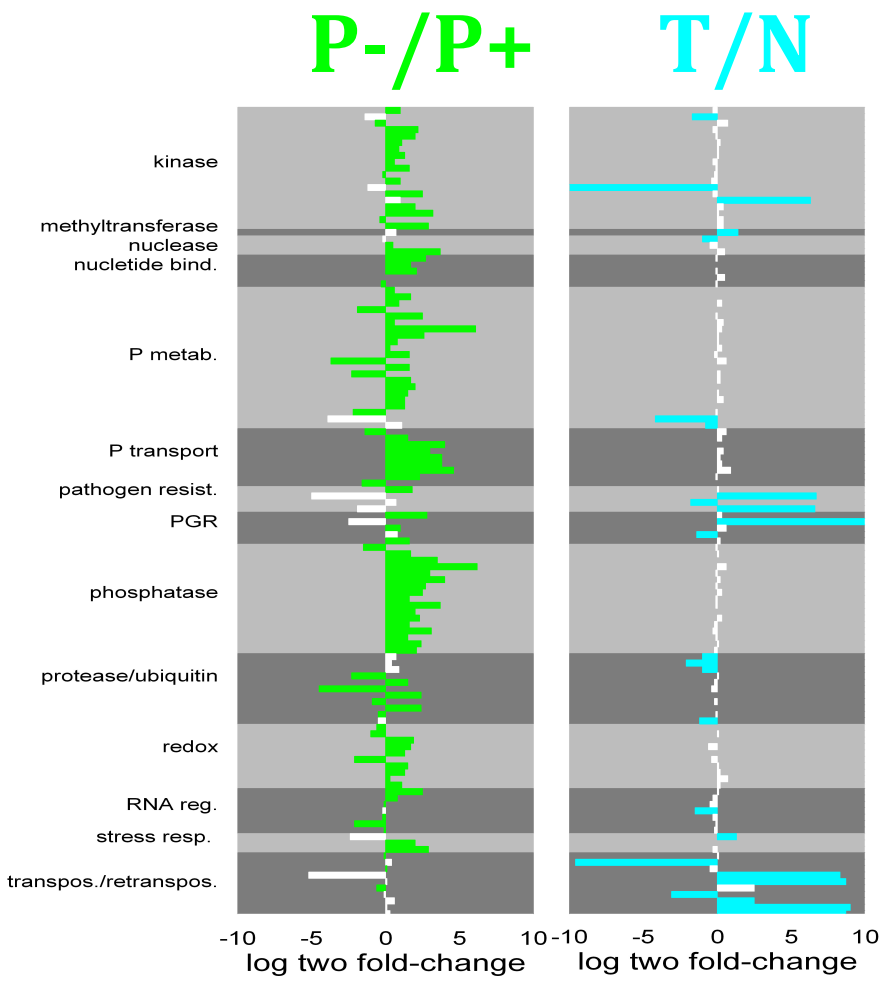

Figure 4. RNAseq gene expression result: up and down regulation identified for P-/P+ and T/N. As there was no statistical difference between phosphorus treatments and tolerance, the average of all treatments by either P treatment or by phenotype was used in this analysis. All genes significantly differentially regulated (at FDR = 0.01) in either treatment (P or phenotype) are shown. Where the ratio was significant between treatments the bars are coloured, where not significant, the bars are white.

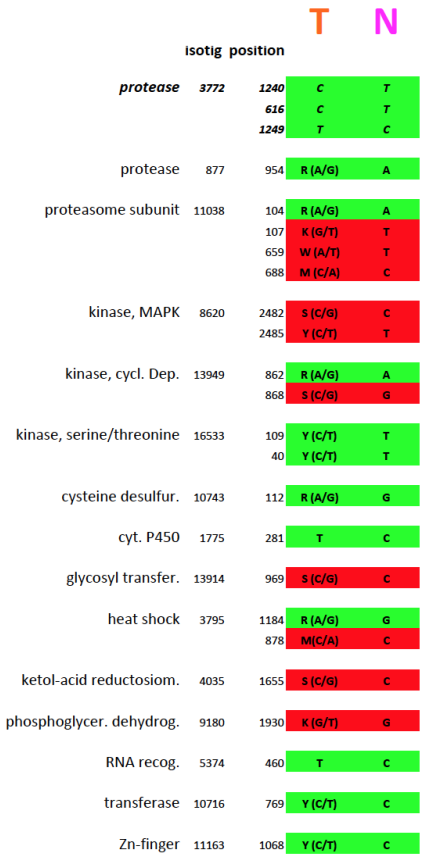

Figure 5. Consistent T versus N (n=10) homozygous (~100%) and heterozygous (~50%) SNPs identified are found predominantly in transcripts with regulatory function. In red transversions A <-> C, A <-> T, G <-> C, G <-> T (interchange of two ring purine (AG) for one ring pyrimidine (CT)). In green the more common transitions A<->G, C<->T (interchange of two ring purines (AG) or one ring pyrimidines (CT)).