The sequencing of over a thousand natural strains of the model plant Arabidopsis thaliana is producing unparalleled information at the genetic level for plant researchers. To enable the rapid exploitation of these data for functional proteomics studies, we have created a resource for the visualization of protein information and proteomic datasets for sequenced natural strains of A. thaliana.

Results: The 1001 Proteomes portal can be used to visualize amino acid substitutions or non-synonymous single-nucleotide polymorphisms in individual proteins of A. thaliana based on the reference genome Col-0. We have used the available processed sequence information to analyze the conservation of known residues subject to protein phosphorylation among these natural strains. The substitution of amino acids in A. thaliana natural strains is heavily constrained and is likely a result of the conservation of functional attributes within proteins. At a practical level, we demonstrate that this information can be used to clarify ambiguously defined phosphorylation sites from phosphoproteomic studies. Protein sets of available natural variants are available for download to enable proteomic studies on these accesses. Together this information can be used to uncover the possible roles of specific amino acids in determining the structure and function of proteins in the model plant A. thaliana. An online portal to enable the community to exploit these data can be accessed at http://1001proteomes.masc-proteomics.org/.

ABSTRACT

Motivation: The sequencing of over a thousand natural strains of the model plant Arabidopsis thaliana is producing unparalleled information at the genetic level for plant researchers. To enable the rapid exploitation of these data for functional proteomics studies, we have created a resource for the visualization of protein information and proteomic datasets for sequenced natural strains of A. thaliana.

Results: The 1001 Proteomes portal can be used to visualize amino acid substitutions or non-synonymous single-nucleotide polymorphisms in individual proteins of A. thaliana based on the reference genome Col-0. We have used the available processed sequence information to analyze the conservation of known residues subject to protein phosphorylation among these natural strains. The substitution of amino acids in A. thaliana natural strains is heavily constrained and is likely a result of the conservation of functional attributes within proteins. At a practical level, we demonstrate that this information can be used to clarify ambiguously defined phosphorylation sites from phosphoproteomic studies. Protein sets of available natural variants are available for download to enable proteomic studies on these accesses. Together this information can be used to uncover the possible roles of specific amino acids in determining the structure and function of proteins in the model plant A. thaliana. An online portal to enable the community to exploit these data can be accessed at http://1001proteomes.masc-proteomics.org/.

Contact: jlheazlewood@lbl.gov

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on December 13, 2011; revised on February 14, 2012; accepted on March 12, 2012

1 INTRODUCTION

The genome sequence of the model plant Arabidopsis thaliana was completed over 10 years ago by a consortium of international research groups and facilities (Arabidopsis Genome Initiative, 2000). Recently, a project to re-sequence over one thousand genomes from A. thaliana accessions or natural variants using next-generation sequencing platforms was initiated (Weigel and Mott, 2009). The 1001 Genomes project will provide unparalleled information for plant molecular geneticists, as natural variation caused by single-nucleotide polymorphisms (SNPs) can be readily mapped using genome-wide association studies (Atwell et al., 2010). Currently, sequence information for over 450 A. thaliana natural strains are available at 1001 Genomes with a significant number of accessions in the sequencing pipeline (http://1001genomes.org/). Initial analyses of two re-sequenced A. thaliana natural strains (Bur-0 and Tsu-1) indicated that SNPs in coding regions resulted in over 80 000 changes to amino acids (Ossowski et al., 2008). More recently, with the analysis of 80 A. thaliana accessions by whole genome sequencing, over 12 000 SNPs were identified that potentially resulted in drastic effects to coding regions (Cao et al., 2011). These findings prompted us to investigate whether this emerging dataset could be exploited to analyze conservation of functional aspects of protein sequences. In recent years, the large-scale characterization of protein phosphorylation has become relatively straightforward due to advances in enrichment and analysis strategies (Macek et al., 2009). In A. thaliana, many thousands of phosphorylation sites have been experimentally characterized using emerging phosphoproteomic techniques (Durek et al., 2010; Nakagami et al., 2010). We have employed these distinct datasets to examine whether it is possible to assess the conservation and corroborate post-translational modifications in proteins of A. thaliana. Finally, in order to provide additional utility to proteomic researchers, we have developed a portal to provide easy access to the proteomic sequence data resulting from this newly developed information. This portal is part of a collection of resources developed by the Proteomics Subcommittee (Weckwerth et al., 2008) of the Multinational Arabidopsis Steering Committee (MASCP) and the 1001 Genomes consortium (Weigel and Mott, 2009) and is available at http://1001proteomes.masc-proteomics.org/.

2 METHODS

2.1 Construction of protein datasets

Pseudo chromosomes for the A. thaliana reference genome (Col-0) were obtained from The Arabidopsis Information Resource (TAIR) and corresponded to genome release TAIR10 (Swarbreck et al., 2008). SNP datasets were obtained from available data at the 1001 Genomes portal (Weigel and Mott, 2009) comprising published sets (Cao et al., 2011;
To demonstrate the utility of the 1001 Proteomes portal a median protein length (A. thaliana derived from each accession) was calculated. The utility uses the Arabidopsis gene identifier (AGI), a unique gene and protein identifier for Arabidopsis, as the principal input. The web interface was constructed using previously developed tools and requires a web browser (Joshi et al., 2011). Pre-computed protein sets in FASTA format for each accession are available from the Data Center at the 1001 Proteomes portal (http://1001proteomes.masc-proteomics.org/).

3 RESULTS

To demonstrate the utility of the 1001 Proteomes portal and the potential impact of non-synonymous single-nucleotide polymorphisms (nsSNPs) on protein function, we initially determined the number of accessions that would result in a complete amino acid substitution of an A. thaliana protein. An analysis of the average nsSNP rate per protein for the publicly available re-sequenced A. thaliana accessions was undertaken (Supplementary Table S1). On average, there are 1.9 (σ = 6.8) nsSNPs per protein per accession, when compared with protein sequences from the reference accession Col-0. Initially, we made the assumption that nsSNPs occur randomly throughout a protein and that with enough accessions, nsSNPs would occur at every amino acid. We initially calculated the probability of total coverage after taking into account the effect of an increasing number of accessions, nsSNPs occur at every amino acid. Consequently if the underlying modified residues we analyzed the effect of nsSNPs upon protein phosphorylation represents one of the most common post-translational modifications and generally occurs to a very specific subset of amino acids, namely Ser, Thr and Tyr. Recent large-scale phosphoproteomic studies in A. thaliana (Col-0) have resulted in large publicly available datasets (Nakagami et al., 2010). In order to examine the importance of modified residues we analyzed the effect of nsSNPs upon experimentally determined phosphorylation sites. The RIPP-DB database (https://database.riken.jp) contains ~5300 experimentally determined phosphorylation sites in A. thaliana (Col-0) have resulted in large publicly available datasets. The conservation rate among accessions of the ca. 5300 experimentally determined phosphorylation sites was significantly more conserved (p = 0.03). Statistical significance was calculated by creating a contingency table, and using Pearson’s chi-squared test.

3.2 Experimentally determined amino acid substitutions in accessions

The analysis of the expected nsSNPs that could occur in A. thaliana proteins indicates that data from over 2370 accessions would be necessary to confidently define important residues and regions within a protein. To assess whether these theoretical substitution rates were matched in the experimental datasets, nsSNPs occurring in each of the randomly selected accessions were collated. To compare the experimental rates of substitution, we calculated the actual average rate of substitution for all proteins (essentially the probability that any selected amino acid has been substituted) under the effect of an increasing number of accessions. To achieve this, substitutions observed in previous accessions were discarded, such that the average number of unique substitutions per protein for any accession could be established (Fig. 1A). The distribution of actual rates of substitution is closer to a log distribution, and is clearly different to the theoretical rates of substitution. These data support the notion that nsSNPs do not occur randomly on amino acids throughout proteins of A. thaliana and their distribution within a protein is constrained.
The 1001 Genomes program involving the sequencing of *A. thaliana* of genomic sequence variation among the natural strains has the capacity to significantly enhance our ability to identify subtleties in gene regulation. To allow proteomic researchers to more readily interface with this information, we have created a portal called 1001 Proteomes to distribute translated protein information and to provide interface to readily browse data at the protein level. As an illustrative example of the utility of such a resource to understand protein function, we demonstrated that these data can be employed to examine features such as conservation of protein modifications such as phosphorylation in the model plant *A. thaliana*. Experimentally determined phosphorylation sites from large-scale phosphoproteomics studies of the reference *A. thaliana* strain Col-0, were significantly less likely to be substituted by nsSNPs in accessions.

Our initial analyses indicated that several thousand accessions would be required to obtain adequate or complete amino acid substitutions based on coding region changes in natural strains of *A. thaliana*. In reality, substitutions in the coding regions of these proteins is heavily constrained and likely controlled by conservation of functional attributes within a protein. Although the actual amino acid substitution space across *A. thaliana* accessions is constrained, it was still possible to use these data to validate the conservation of phosphorylation sites in protein sequences. It is likely that this information can also be used to tease apart important and redundant residues within functional domains of proteins. The utility of this portal for functional proteomics can be further highlighted when examining ambiguities in phosphorylation site assignments from large-scale phosphoproteomics. Variable phosphorylation assignments are common when multiple S, T or Y residues are present in an identified phosphopeptide. An examination of variably assigned phosphorylation sites on a handful of phosphopeptides revealed that it was possible to readily determine the likely phosphorylated residue (Table 1). These results further demonstrate the value in visually presenting nsSNP data to the research community as they can address simple but diverse questions on protein structure and function.

Recently a similar analysis of phosphorylation site conservation across natural strains of *A. thaliana* concluded that there was no association between experimental sites and nsSNPs (Riano-Pachon et al., 2010). This contrary finding was likely due to the non-uniformity of the SNP datasets which comprised data from re-sequencing arrays of 20 accession (Clark et al., 2007), short
fragment sequencing of 96 accessions (Nordborg et al., 2005) and early versions of next generation re-sequencing data for only two accessions (Ossowski et al., 2008). Overall this dataset likely lacked the resolution to obtain adequate coverage of SNPs and specifically nsSNPs from A. thaliana accessions. Furthermore recent work examining nsSNPs associated with phosphorylation sites in humans (phosSNPs) argued that phenotypic effects may be attributable to nsSNP substitutions of phosphorylated residues (Ren et al., 2010). These observations further reinforce the likely conservation of nsSNPs that alter protein function and effect phenotype.

The development of the 1001 Proteomes portal (http://1001proteomes.masc-proteomics.org) provides a simple means to analyze the role of amino acids on protein attributes such as post-translational modifications. We have now developed an automated pipeline that efficiently converts the processed re-sequencing data from the Arabidopsis 1001 Genomes project into visual tracks at Arabidopsis proteomics data. These observations further reinforce the likely conservation of nsSNPs that alter protein function and effect phenotype.

Table 1. Utilization of nsSNPs to determine ambiguous phosphorylation sites in phosphoproteomics data

<table>
<thead>
<tr>
<th>AGI</th>
<th>Ambiguous (PhosPhAt)</th>
<th>nsSNP assisted</th>
<th>Site</th>
<th>Acc.</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G07620.1</td>
<td>NE(s)SPNHGKYNHK</td>
<td>NENg5SPNHGKYNHK</td>
<td>Ser10</td>
<td>75</td>
<td>Unpublished</td>
</tr>
<tr>
<td>AT1G31440.1</td>
<td>LH(s)ieoxMSAEEAIGoxo</td>
<td>LHAeoxMSAEEAALGsgPKx</td>
<td>Ser294</td>
<td>27/69</td>
<td>(Rieland et al., 2009)</td>
</tr>
<tr>
<td>AT1G72150.1</td>
<td>(s)V(s)VKEEETVVV</td>
<td>AEK(V)pSVPVKEEETVVV</td>
<td>AEK</td>
<td>56</td>
<td>(Whiteman et al., 2008)</td>
</tr>
<tr>
<td>AT1G31440.1</td>
<td>(s)V(s)VKEEETVVV</td>
<td>AEK(V)pSVPVKEEETVVV</td>
<td>AEK</td>
<td>56</td>
<td>(Whiteman et al., 2008)</td>
</tr>
<tr>
<td>AT1G31440.1</td>
<td>LH(s)E(oxM)IAEEEAIG(s)(s)PK</td>
<td>LHAE(oxM)IAEEEAIGA(pS)PK</td>
<td>Ser754</td>
<td>52</td>
<td>(Nakagami et al., 2010)</td>
</tr>
<tr>
<td>AT3G59820.1</td>
<td>LGSKPEENATEEE(s)(s)</td>
<td>LGSKPEENATEEE(pS)N</td>
<td>Ser184</td>
<td>24</td>
<td>Unpublished (PhosPhAt)</td>
</tr>
<tr>
<td>AT2G37340.1</td>
<td>NSVV(pS)PVVGAGGD(s)(s)K</td>
<td>NSVV(pS)PVVGAGGD(pS)PK</td>
<td>Ser184</td>
<td>24</td>
<td>Unpublished (PhosPhAt)</td>
</tr>
</tbody>
</table>

The amino acid substitution is underlined in the nsSNP-assisted column. Site indicates the likely phosphorylation site after taking into account the substitution. Acc. (accession) indicates the number of accessions with the nsSNP. *Phosphorylation site also confirmed experimentally by other studies, see PhosPhAt (http://phosphat.mpimp-golm.mpg.de/).

ACKNOWLEDGEMENTS

We wish to acknowledge the RIPP-DB resource (https://database.riken.jp) and the PhosPhAt resource (http://phosphat.mpimp-golm.mpg.de/) for collecting and providing publicly accessible phosphorylation data.

Funding: The work conducted by the Joint BioEnergy Institute was supported by the Office of Science, Office of Biological and Environmental Research of the US Department of Energy under Contract No. DE-AC02-05CH11231. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. Work on the 1001 Genomes project in the Weigel lab is supported by a Gottfried Wilhelm Leibniz Award of the Deutsche Forschungsgemeinschaft and by the Max Planck Society.

Conflict of Interest: none declared.

REFERENCES


H.J. Joshi et al.