BIOINFORMATICS DISCOVERY NOTE Vol. 28 no. 10 2012, pages 1303–1306
doi:10.1093/bioinformatics/bts133

Databases and ontologies Advance Access publication March 25, 2012

1001 Proteomes: a functional proteomics portal for the analysis of *Arabidopsis thaliana* accessions

Hiren J. Joshi[1], Katy M. Christiansen[1], Joffrey Fitz[2], Jun Cao[2], Anna Lipzen[3], Joel Martin[3], A. Michelle Smith-Moritz[1], Len A. Pennacchio[3], Wendy S. Schackwitz[3], Detlef Weigel[2] and Joshua L. Heazlewood[1,*]

[1]Joint BioEnergy Institute and Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [2]Molecular Biology Department, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany and [3]US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

Associate Editor: Trey Ideker

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 accessions. 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 whom correspondence should be addressed.

Schneeberger *et al.*, 2011) and eight unpublished accessions by the Joint Genome Institute and the Joint BioEnergy Institute. These data contained pre-computed nucleotide variances (SNPs) against the *A. thaliana* reference genome strain Col-0. Accession specific pseudo chromosomes were constructed and protein sets extracted using a pipeline employing Perl and the Bio::Perl libraries (Stajich *et al.*, 2002) as well as gene reference information available at TAIR.

2.2 Identification of protein substitutions

Protein substitutions in the accession specific protein sets were identified using a simple non-aligning difference algorithm that compared amino acids at equal positions along the protein. This process was implemented using an in-house program written in C and the resultant output optimized so that subsequent analyses could simply extract the specific amino acid differences for each protein without re-calculating differences. This program is available for download at http://code.glycocode.com/SNP-server/.

2.3 Construction of the web interface

A web interface was constructed to visualize the amino acid substitutions derived from each *A. thaliana* accession for an individual protein. 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 development 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

3.1 Expected amino acid substitution per accession

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 re-sequenced *A. thaliana* accessions (Fig. 1B). The probability that all amino acids are substituted on a protein n amino acids long, having seen t accessions, given a probability that any one amino acid is substituted of p is calculated by the following formula.

$$P(n,t) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} (1-p)^{it}$$

Given the probability of any random amino acid being substituted (p) (580 000 unique substitutions on ~13 million amino acids) and a median protein length (n) of 340 amino acids (Swarbreck *et al.*, 2008), we could be 99.5 % certain of total amino acid substitution with sequence data from over 2370 *A. thaliana* accessions (Fig. 1B).

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.

3.3 Post-translation modifications in A. thaliana accessions

Post-translational modifications are important functional components of a protein and are likely to be conserved across *A. thaliana* accessions. Consequently if the underlying modified amino acid is crucial to the function of a protein, it will likely be substituted at a lower rate. 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* derived from the reference accession Col-0 and was used to assess the conservation of phosphorylation sites among accessions. A comparison of the conservation rate among accessions of the ca. 5300 experimentally determined phosphorylation sites from RIPP-DB and random sets of amino acids comprising Ser, Thr and Tyr (total 6478) indicated that experimentally determined phosphorylation sites were significantly more conserved (p = 0.03). Statistical significance was calculated by creating a contingency table, and using Pearson's chi-squared test.

In order to test whether the number of random theoretical amino acids selected per accession affected the result or not, varying multipliers of selected amino acids were tested. Placing these data into a contingency table, and using Pearson's chi-squared test, the data was not significantly different (p = 0.97). To control for any effects from promiscuous substitutions on proteins across many accessions, the distribution of nsSNPs found in single to close to a hundred accessions were compared with respect to the distribution of nsSNPs substituting phosphorylated sites, and were identical (p = 1, using chi-squared test), suggesting that there is no relation between the post-translational modification state of an amino acid, and the number of accessions a particular nsSNP is observed in. Thus,

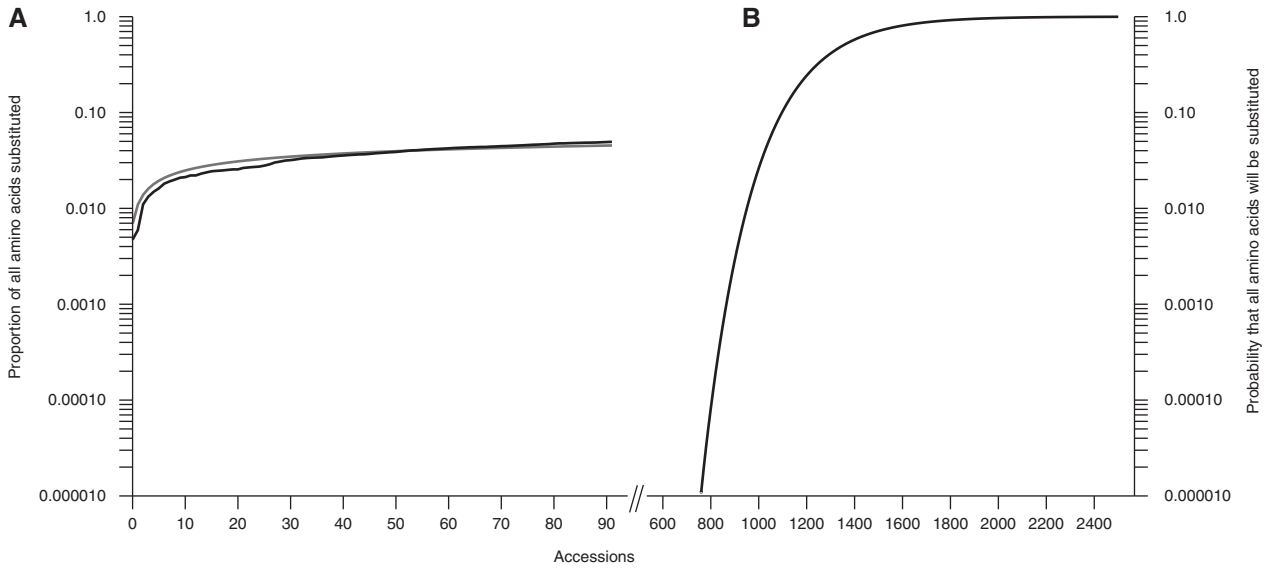

Fig. 1. Proportions/probabilities of amino acid substitution with an increasing number of *A. thaliana* accessions. Plot of an experimentally derived proportion of amino acids in a protein that have been substituted (**A**) and the theoretical probability that all amino acids will be substituted (**B**) given a static probability of substitution across all amino acids. The experimental data suggests that the regions of substitution will be constrained. The grey line (A) is a plot of ln(x)/100.

amino acid substitutions in proteins of *A. thaliana* are constrained by conserved functional motifs such as sites of phosphorylation.

3.4 1001 Proteomes web interface

The principal focus of the 1001 Genomes project is the analysis of genomic sequence variation among the *A. thaliana* natural strains collection. In order to make this information useful for functional proteomics studies, we have created a web interface for which facilitates the visualization of protein variation, 1001 Proteomes. A simple and intuitive interface was created where a user inputs an AGI and retrieves data for the corresponding protein (http://1001proteomes.masc-proteomics.org). A consolidated amino acid substitution track comprising nsSNPs is shown under the sequence of the reference *A. thaliana* protein from the accession Col-0. The size of the nsSNP on the consolidated track is scaled depending on the number of accessions that contribute to a specific substitution. The specific amino acid substitution and each accession that contributes to the consolidated track can be visualized by toggling a switch on the consolidated track. Thus all nsSNP information for each accession for a given protein can be observed and assessed thorough a simple visual medium. Finally, pre-computed accession specific proteome sets are available for download to enable these data to be used in proteomic studies.

4 DISCUSSION

The 1001 Genomes program involving the sequencing of *A. thaliana* 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

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

AGI	Ambiguous (PhosPhAt)	nsSNP assisted	Site	Acc.	Source
AT1G07620.1	NE(s)(s)SPNHGKYNHK	NEN(pS)SPNHGKYNHK	Ser^{10}	75	Unpublished (PhosPhAt)
AT1G31440.1	LH(s)E(oxM)IAEEEAIG(s)(s)PK	LHAE(oxM)IAEEEAIGA(pS)PK[a]	Ser^{264}	27/69	(Reiland *et al.*, 2009)
AT1G72150.1	(s)V(s)VKEEETVVVAEK	(pS)VPVKEEETVVVAEK	Ser^{69}	56	(Whiteman *et al.*, 2008)
AT2G37340.1	NSVV(pS)PVVGAGGD(s)(s)K	NSVV(pS)PVVGAGGD(pS)PK[a]	Ser^{265}	13	(Nakagami *et al.*, 2010)
AT3G59820.1	LGSKPEENATEEE(s)(s)	LGSKPEENATEEE(pS)N[a]	Ser^{754}	52	(Nakagami *et al.*, 2010)
AT5G20200.1	LIEMI(s)(s)R	LIEMIN(pS)R	Ser^{184}	24	Unpublished (PhosPhAt)
AT5G45060.1	DVNL(t)(s)LK	DVNLM(pS)LK	Ser^{706}	60	Unpublished (PhosPhAt)

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

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 1001 Proteomes. New data will be made available both in the browser and as downloadable proteins sets for functional proteomics studies as they are made publicly available to the community.

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

Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.

Atwell,S. *et al.* (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **465**, 627–631.

Cao,J. *et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations *Nat. Genet.*, **43**, 956–963.

Clark,R.M. *et al.* (2007) Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*, **317**, 338–342.

Durek,P., *et al.* (2010) PhosPhAt: the Arabidopsis thaliana phosphorylation site database. An update. *Nucleic Acids Res.*, **38**, D828–D834.

Joshi,H.J. *et al.* (2011) MASCP Gator: an aggregation portal for the visualization of Arabidopsis proteomics data. *Plant Physiol.*, **155**, 259–270.

Macek,B. *et al.* (2009) Global and site-specific quantitative phosphoproteomics: principles and applications. *Annu. Rev. Pharmacol. Toxicol.*, **49**, 199–221.

Nakagami,H. *et al.* (2010) Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants. *Plant Physiol.*, **153**, 1161–1174.

Nordborg,M. *et al.* (2005) The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol.*, **3**, e196.

Ossowski,S. *et al.* (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res.*, **18**, 2024–2033.

Reiland,S. *et al.* (2009) Large-scale Arabidopsis phosphoproteome profiling reveals novel chloroplast kinase substrates and phosphorylation networks. *Plant Physiol.*, **150**, 889–903.

Ren,J. *et al.* (2010) PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. *Mol. Cell. Proteomics*, **9**, 623–634.

Riano-Pachon,D.M. *et al.* (2010) Proteome-wide survey of phosphorylation patterns affected by nuclear DNA polymorphisms in Arabidopsis thaliana. *BMC Genomics*, **11**, 411.

Schneeberger,K. *et al.* (2011) Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proc. Natl Acad. Sci. USA*, **108**, 10249–10254.

Stajich,J.E. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

Swarbreck,D. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.

Weckwerth,W. *et al.* (2008) The multinational Arabidopsis steering subcommittee for proteomics assembles the largest proteome database resource for plant systems biology. *J. Proteome Res.*, **7**, 4209–4210.

Weigel,D. and Mott,R. (2009) The 1001 Genomes project for Arabidopsis thaliana. *Genome Biol.*, **10**, 107.

Whiteman,S.A. *et al.* (2008) Identification of novel proteins and phosphorylation sites in a tonoplast enriched membrane fraction of Arabidopsis thaliana. *Proteomics*, **8**, 3536–3547.