

Copyright: © 2017 Wiley
It is posted here for your personal use. No further distribution is permitted.
Plant proteogenomics: Improvements to the grapevine genome annotation

Brett Chapman1, Matthew Bellgard 1§

1 Centre for Comparative Genomics, Murdoch University, Western Australia, 6150

§Corresponding author

Email addresses:

BC: bchapman@ccg.murdoch.edu.au

MB: mbellgard@ccg.murdoch.edu.au

Keywords: Bioinformatics, Proteogenomics, Proteomics, Vitis vinifera

Total word count: 5,794

Abstract

Grapevine is an important perennial fruit to the wine industry, and has implications for the health industry with some causative agents proven to reduce heart disease. Since the sequencing and assembly of grapevine cultivar Pinot Noir, several studies have contributed to its genome

Received: 05 20, 2017; Revised: 07 28, 2017; Accepted: 08 31, 2017

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pmic.201700197.

This article is protected by copyright. All rights reserved.
annotation. This new study further contributes towards genome annotation efforts by conducting a proteogenomics analysis using the latest genome annotation from CRIBI, legacy proteomics dataset from cultivar Cabernet Sauvignon and a large RNA-seq dataset. A total of 341 novel annotation events were identified consisting of 5 frame-shifts, 37 translated UTRs, 15 exon boundaries, 1 novel splice, 9 novel exons, 159 gene boundaries, 112 reverse strands and 1 novel gene event in 213 genes and 323 proteins. From this proteogenomics evidence, the Augustus gene prediction tool predicted 52 novel and revised genes (54 protein isoforms), 11 genes of which were associated with key traits such as stress tolerance and floral and fruity wine characteristics. This study also highlighted a likely over-assembly with the genome, particularly on chromosome 7.

1 INTRODUCTION

The grapevine industry has global market access and large economic support worldwide. The genus Vitis is important to the wine industry and is a perennial fruit consumed as part of the staple Mediterranean diet where evidence indicates a reduced prevalence of heart disease [1]. The putative causative agents reducing the prevalence of heart disease may well be derived from grapes, with a number of key candidates being resveratrol, quercetin and ellagic acid [2], with resveratrol attracting media attention in recent years as a potential life-extending drug [3], and which further adds to the Vitis market value and opens up the potential for many unexplored medical benefits. The broad spectrum of commercial applications of grapevines challenges the industry to improve yields, quality, resistance to diseases and abiotic stress conditions across the globe. One particularly important Vitis species is Vitis vinifera, with the heterozygous variety Pinot Noir [4], and a 93% homozygous Pinot Noir (genotype PN40024) [5], recently being sequenced. The sequencing and assembly of the latter variety has since been improved from 8X coverage to 12X coverage resulting
in a 487.1 Mbp assembled genome. Genomic annotation of the 8X and 12X was also undertaken, with the 8X gene prediction being performed using GAZE [6], published along with its sequencing [5], while the 12X sequence and assembly has since resulted in three different iterative improvements in annotation. The first named 12Xv0 was performed using GAZE. The second named 12Xv1 resulted from the combination of 12Xv0 and gene predictions by the tool JIGSAW [7], undertaken at CRIBI in Padova, Italy [8]. The third improvement named 12Xv2 (since updated to 12Xv2.1) was undertaken recently, using assembled transcripts from RNA-seq, ab initio predictions, proteins, and ESTs [9].

Proteogenomics is a new methodology for improving and refining genome annotations, which utilizes peptide mass spectrometry, first pioneered over 15 years ago [10, 11]. The proteogenomics approach assists with improving gene predictions as well as the validation of previously annotated genes as being protein coding. There have been many different methodologies over the last several years [12-19]. One novel approach implemented within the Enosi [13] tool utilizes a holistic approach to assigning confidence to annotation events, using an annotation event probability (eventProb) [13], previously explored by the authors in a bacterial case study [20], and which has also been extensively used throughout this study.

1.1 Outline of this study
The aim of this study was to apply proteogenomics to further improve on genomic annotation of the grape genome and to highlight the complexity of performing proteogenomics annotation in larger plant genomes. In addition, the benefits and shortcomings of current proteogenomics strategies were outlined. The latest grape Pinot Noir 12X genome assembly and 12Xv2.1 genome annotation were obtained from CRIBI, while the proteomics data were in the form of 177,174 MS/MS spectra derived from Cabernet Sauvignon grape berry skins, used in an earlier proteogenomics study in Chapman et al [21], identifying 29 annotation events. The present study expands on that work by using an improved proteogenomics pipeline and an additional 2,701,718 MS/MS spectra derived.
from a previous proteomics study [22], RNA-seq data [23], and a large currently unpublished RNA-seq dataset to contribute towards the identification of splice regions.

2 MATERIALS AND METHODS

2.1 Proteomics, genomics and RNA-seq datasets
The latest assembled grape Pinot Noir 12X genome [5] with the genotype identifier PN40024, and the 12Xv2.1 genome annotation and protein predictions [9] were downloaded for use from the CRIBI web site (http://genomes.cribi.unipd.it/DATA/). The MS/MS spectra were derived from finely ground shoot tips of Cabernet Sauvignon [22]. The samples were digested with trypsin, aided by Lys-C digestion, and were run on a LTQ XL (Thermo) mass spectrometer. Further MS/MS spectra were derived from berry skins, obtained from a previous proteogenomics study [21], running on an LTQ Velos Pro (Thermo) mass spectrometer. It should be noted that the mass spectrometers used are of low MS and MS/MS precision, however, they are capable of high sensitivity and scan speeds, producing large numbers of MS/MS spectra in short time.

A total of 2,701,718 MS/MS spectra were downloaded from the ProteomeXchange Consortium (http://proteomcentral.proteomexchange.org) via the PRIDE partner repository [24], using the identifier PXD000123. An additional 177,174 MS/MS spectra from the previous proteogenomics study in Chapman et al [21] were also used, derived from Cabernet Sauvignon berry skins. A contaminant database was also used (ftp://ftp.thegpm.org/fasta/cRAP/crap.fasta).

RNA-seq datasets were obtained from 1) transcriptomes of V. vinifera cultivar Corvina [23], under different time-points and abiotic stresses, downloaded from the Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) identifier SRA055265 and 2) a currently unpublished dataset of different V. vinifera cultivars from the grape skins of Chardonnay, Cabernet Sauvignon, Merlot, Pinot
Noir, Semillon, Cabernet Franc and Sauvignon Blanc, also under different time-points and abiotic stresses, obtained from collaborator Ryan Ghan from the University of Nevada.

### 2.2 MS/MS database searching

The MS/MS database search was performed by MS-GF+ (version v9949 2/10/2014) [25] and used the following settings: 1) The protease used was trypsin; 2) precursor mass tolerance was 2.0 Da and 3.0 Da for two different MS/MS database searches (see parameter optimization in Supporting Information); 3) number of modifications per peptide was set to 2, with modifications: carbamidomethylation of cysteine (C+57), oxidation of methionine (M+16) and protein N-terminal acetylation (+42); 4) maximum peptide length was 30 amino acids (aa); 5) isotope error range was “0.2”; 6) instrument was set to low-res Linear Trap Quadrupole (LTQ) (e.g. Ion Trap); 7) number of tolerable termini was 1; 8) number of reported peptide-spectrum matches (PSMs) was set to 10, and 9) a reverse sequence decoy database was generated.

### 2.3 Proteogenomics pipeline

Proteomics, genomics and RNA-seq datasets were processed, formatted and optimized. The RNA-seq reads were mapped to the genome and a splice database was generated, along with the six-frame translation of the genome. The MS/MS database searching was performed, peptides mapped, clustered, 1% PSM FDR, 5% local FDR (IFDR) and 5% peptide-level FDR (pepFDR) filtering applied across 4 different merged result files, and annotation events inferred, followed by annotation event probability (eventProb) filtering, using the Enosi tool (version 1.0) [26].

Based on a pre-processing and optimization step (Figure 1, Supplementary File 1, Supplementary Figure 1 and 2), the MS/MS spectral dataset was clustered by a factor of 1.5, quality filtered to a PepNovo score of 0.01, and the selection of two separate precursor mass tolerances of 2.0 Da and 3.0 Da were chosen for two separate proteogenomics runs (Supporting Information and Supplementary File 1) and results later aggregated.
The proteogenomics pipeline was run with a minimum eventProb of 90%, a peptide linkage distance of 18,000 bp representing >95% of gene sizes in the 12Xv2.1 annotation, a minimum cluster size of 1 (total peptides per cluster) and to ensure a relatively high specificity of annotation event identification, peptide clusters with at least 1 unique peptide were accepted.

Validation of single unique peptide annotation events was performed with consideration of spectral counts, peptide length, homology to sequences in NCBI NR, NCBI RefSeq protein and NCBI SwissProt protein repositories using BLASTP (Supplementary File 2 and Supporting Information).

For all previously annotated protein identifications, using an in-house script, identifications were divided into two confidence levels: all proteins that contained any mapped peptides and all proteins that contained ≥2 peptides and ≥1 unique peptide (Figure 1), to address the protein inference problem.

Using evidence from the proteogenomics mapping, gene predictions were carried out using Augustus (version 3.02) [27], as outlined in Supporting Information.

3 RESULTS AND DISCUSSION

3.1 Proteogenomics analysis
The proteogenomics analysis was performed based on the refinement of the precursor mass tolerance, MS/MS clustering and quality filtering parameters (Supplementary File 1, Supplementary Figure 1 and 2) and the orthogonal validation of novel annotations. To account for the inflated numbers of gene boundary and reverse strand events for each peptide cluster, due to the use of a fixed peptide linkage distance, a non-redundant count of the 1:1 association of peptide cluster to annotation event was also determined, outlined in parenthesis and with an explanatory note in
Table 1.

After annotation events were screened the final eventProbs were 99.80% for novel genes and 98.37% for distal events and proximal events, leading to the identification of a total of 129 novel peptides and 339 novel annotation events (102 non-redundant associations) among 213 genes (67 non-redundant associations). These identified annotation events included 5 frame shifts, 37 translated UTRs, 15 exon boundaries, 1 novel splice, 9 novel exons, 159 (24 non-redundant associations) gene boundaries, 112 (10 non-redundant associations) reverse strands and 1 novel gene (Table 1 and Supplementary File 2).

This study showed an improvement over the 29 annotation events identified during the preliminary study [21]. The proteogenomics evidence was then used as hints for Augustus gene prediction. A total of 84,948 genes and 93,754 proteins (≥66 aa in length) were predicted, and of these 54 predicted proteins had 106 novel peptides incorporated (Table 1), of which 90 novel peptides were unique and identified in 51 of the 54 predicted proteins (Supplementary File 3). The number of protein-coding genes and proteins predicted by Augustus was higher than the original reference 12Xv2.1 predictions (Table 1), which would be in part due to the fragmented genome and incomplete coverage of orthogonal evidence, outlined in Supporting Information.

A BLASTP search was performed, searching all 93,754 Augustus-predicted proteins (Table 1) against the 12Xv2.1 proteins, taking the top match with E-value ≤1E-10. Any sequences that did not match were considered novel predictions, sequences that had a query coverage ≥95% with at least 1 mismatch were considered to be the same prediction as the reference protein, and the remaining matches were considered to be modified predictions, either due to Augustus predicting different gene models or modified as a direct result from the supporting evidence. From this analysis, there were 42,257 non-paralogous novel protein predictions, 32,837 modified predictions and 18,660...
predictions considered to be essentially the same as the reference.

Searching all 54 protein predictions that had the novel peptide evidence incorporated, against the 12Xv2.1 proteins, taking the top match with E-value ≤1E-10, identified 47 protein predictions likely to be modified predictions, leaving 7 protein predictions that found no match and were considered as non-paralogous novel protein predictions (Table 1).

Based on the annotation events incorporated into the Augustus gene predictions, the minimum eventProbs which led to a new Augustus gene prediction were: gene boundary, translated UTR, reverse strand and exon boundary event 98.374%, frame-shift event 99.80%, and novel exon event 99.193%.

Insert Table 1 here

3.2 Novel annotation events and gene model revisions
There were a total of 339 annotation events identified. The following sections details 5 of these annotations events (Figures 2 – 4, Supplementary Figure 3) and provides orthogonal evidence and Augustus gene predictions. Additional annotation events not discussed in this manuscript can be found in Supplementary File 2, Supplementary Figures 4 – 8, putative over-predictions from N-terminal peptides in Supplementary File 4, Supplementary Figures 9 - 13, and MS/MS spectra for all these annotations in Supplementary Figures 15 - 28.

3.2.1 Gene boundary annotations
There were 159 (24 non-redundant associations) gene boundary events identified (Table 1). An example of a gene boundary event is on chromosome 18, spanning positions 4,109,211 to 4,112,683, with gene VIT_218s0001g04980, consisting of two protein isoforms with an eventProb of 100% and with 13 PSMs identifying 4 unique peptides. In addition, an Augustus gene prediction improved the gene model (Figure 2, and with 13 supporting annotated MS/MS spectra in Supplementary Figure
Performing a BLASTP search against the grape family in NR revealed that all novel peptides matched acetyl-CoA carboxylase 1-like protein (XP_002285808.2 with E-value range: 5E-07 – 3E-16), with 100% query coverage and identity. The two reference protein-coding transcripts also matched acetyl-CoA carboxylase 1-like protein (XP_002285808.2 both with E-values = 0.0); protein-coding transcript 1 with 100% query coverage and protein-coding transcript 2 with 98% query coverage, both with 100% identity. The Augustus gene prediction also matched acetyl-CoA carboxylase 1-like protein (XP_002285808.2 with E-value = 0.0), with 100% query coverage and identity, which showed that the original prediction was under-predicted, requiring a further extension of the gene towards the 5’ region on the reverse strand.

### 3.2.2 Evidence for genome over-assembly

A potential over-assembly of chromosome 7 was indirectly found through the identification of a translated UTR and exon boundary annotation event for gene VIT_207s0031g03000 (Supplementary File 2 and samples of supporting MS/MS spectra in Supplementary Figure 15 and 16, for translated UTR and exon boundary annotations, respectively). Additionally, Augustus predicted a gene model incorporating the novel peptides (Supplementary Figure 3).

Performing a BLASTP search against the grape family in NR revealed that novel peptides from the translated UTR annotation and exon boundary annotation events significantly matched Ribulose bisphosphate carboxylase/oxygenase (AFG24212.1 and CBI21646.3, respectively) with E-values ranging from 2E-04 to 8E-20, with 100% query coverage and identity. The reference protein...
and Augustus prediction also matched a Ribulose bisphosphate carboxylase/oxygenase (CBI21646.3 with E-value = 2E-62 and 100% query coverage and identity, and CAN63541.1 with E-value = 5E-166 and 77% query coverage and 99% identity, respectively).

Ribulose bisphosphate carboxylase/oxygenase, commonly called RuBisCo, is widely known as a dominant protein in plants, found more abundantly in leaves, a major source of peptides in this study. However, the protein is found exclusively in chloroplasts. This is evidence of over-assembly of the reference genome, particularly chromosome 7. With the genome already in a fragmented state and the presence of a large unassigned chromosome [28], there is a strong likelihood that the genome is over-assembled.

### 3.2.3 Frame-shift annotation

There were 5 frame-shift annotations identified (Table 1). An example of a frame-shift annotation was with gene VIT_200s1339g00010 and its single protein-coding transcript. This annotation was identified on chromosome Un, spanning positions 38,615,528 to 38,615,653, with an eventProb of 99.998% and 3 PSMs identifying 2 novel and unique peptides. The novel peptides were incorporated into an Augustus gene model (Figure 3, and with 3 supporting annotated MS/MS spectra in Supplementary Figure 23).

Performing a BLASTP search against the grape family in NR revealed the novel peptides matched a hypothetical protein (CAN63109.1 with E-value ranges: 2E-07 – 5E-21), with 100% query coverage and identity, described as containing a Ribonuclease T2 domain. The reference protein matched a hypothetical protein (CAN63794.1 with E-value = 3.3), with 65% query coverage and 50% identity, described as containing a retrotransposon gag protein domain. The Augustus gene prediction matched the same hypothetical protein as the novel peptides (CAN63109.1 with E-value = 0.0), with 97% query coverage and 98% identity. The BLASTP evidence indicated that the original
reference protein had a poor match to a hypothetical protein, while the Augustus gene prediction found a significant match to a hypothetical protein, with a good protein alignment that included the novel peptides. In addition, the novel peptides and Augustus gene prediction both matched exactly the same protein, described as containing a Ribonuclease T2 domain implicated in plant leaf senescence, which correlates well with the source of the proteomics data, derived from plant leaf shoot tips. This evidence has led to a significant overall improvement to the annotation of this gene.

Insert Figure 3 here

3.2.4 Novel exon annotation

There were 9 novel exon annotations identified (Table 1). An example of a novel exon annotation was with gene VIT_217s0000g02480. This annotation was identified on chromosome 17, spanning positions 2,264,427 to 2,264,591, with an eventProb of 99.999% and 15 PSMs identifying 3 novel and unique peptides. The novel peptides were incorporated into an Augustus gene model, with complete removal of the intron (Figure 4, and with 15 supporting annotated MS/MS spectra in Supplementary Figure 17).

Performing a BLASTP search against the grape family in NR revealed the novel peptides matched calcium-binding allergen Ole e 8 (XP_010663678.1 with E-value ranges: 2E-07 – 2E-15), with 100% query coverage and identity. The reference protein matched unnamed protein product (CBI15562.3 with E-value = 2E-97), with 100% query coverage and identity, described as containing an EF-hand calcium-binding domain. The Augustus gene prediction matched calcium-binding allergen Ole e 8 (XP_010663678.1 with E-value = 2E-176), with 100% query coverage and identity. Overall, the new Augustus gene prediction had a better match to the calcium-binding protein when the novel peptides were included and the intron was removed.

Insert Figure 4 here
3.2.5 N-terminal acetylated peptides

Protein N-terminal acetylation contributes to many functional changes in proteins, from signalling, regulation of protein-protein interactions, and transportation of proteins to their target, such as embedding in membranes [29]. The identification of any N-terminal acetylated peptides not at the N-terminal ends of a previously annotated protein could potentially indicate an over-predicted gene requiring re-annotation, but could also indicate an alternative protein isoform with a different translation initiation start (TIS) site. No N-terminal acetylated peptides were identified from the 129 novel peptides (Table 1), however, a total of 192 N-terminal acetylated peptides were identified among the 12Xv2.1 predicted proteins and 80 were identified from 77 high confidence proteins (≥2 peptides and ≥1 unique peptides). Of the total proteins, 5 shared N-terminal acetylated peptides were identified as conflicting with the 12Xv2.1 annotation (Supplementary File 4). However, no over-predicted genes or alternative TIS sites could be identified with any confidence, as they were shared peptides (Supplementary File 4, Supplementary Figures 9–13 & 24–28).

3.3 Impact of search space

During a proteogenomics search the size of the search space can negatively impact the sensitivity of the search. Compared to previous studies [20], which were limited to small bacterial genomes, in this study with the larger grape genome, the impact was more pronounced. A total of 2,773 out of 55,373 proteins were identified, while the total number of mapped proteins was 7,536. Of these, 1,117 high confidence proteins had ≥2 peptides and ≥1 unique peptides (Table 1). When the same search was conducted against only the 12Xv2.1 proteins a total of 5,795 proteins were identified. Comparisons between proteomics- and proteogenomics-only searches revealed a loss of 3,022 proteins out of 5,795 or a loss of 52%, which would also infer a significant loss to novel identifications. This was significantly higher than the 30% loss found in previous studies [20, 30], due
to the larger genome size of *V. vinifera*. This trend in the loss of sensitivity would most likely continue to increase as the genome size increases in other studies and needs to be addressed.

### 3.4 Controlling for false positives in proteogenomics analysis

During analysis the FDR can be negatively impacted through the introduction of false positives through addition of contamination sequences into the database. In this analysis many of the novel peptides were found to match chloroplast and mitochondrial proteins, accompanied by higher numbers of PSMs (Supplementary File 2), as expected as chloroplasts and mitochondria are numerous, particularly in rapidly growing grape shoot tips used in this study. A cell component isolation step during sampling could have controlled this, had the data not been legacy data. Other sources of false positives would be the low precision MS/MS spectra, which would have larger mass error windows, and could be accounted for in future studies by using higher mass precision instruments such as a QTOF or LTQ Orbitrap. The use of an incomplete fragmented genome with large unassigned chromosome regions [28] and over-assembled genomic regions as outlined in this study (Section 3.2.2) would have also contributed to false identifications. Additionally, discrepancies in the datasets used could also introduce false positives, with the genome being the cultivar Pinot Noir, the proteomic data derived from Cabernet Sauvignon and RNA-seq data covering various different cultivars, variations between the peptide sequences and target genomic and RNA-seq sequences could occur. This could lead to potential misinterpretations of the variation as a post-translational modification (PTM) or assignment of peptides to incorrect genomic regions. In future studies, apart from acquiring appropriately suitable matching datasets, variations between proteomic and genomic datasets could be accounted for by including RNA-seq derived variant sequences in the database [31].
3.5  **Corrections to genes important to the grapevine industry**

The pursuit of improved genomic annotation through proteogenomics cannot be over-emphasised. There were 11 different biologically significant genes identified whose annotation had been putatively improved (5 exon boundaries, 4 translated UTRs, and 2 reverse strands). The eleven genes are involved in a variety of important biological processes, such as abiotic and biotic stress responses, and production of floral and fruity wine characteristics (Supplementary File 3). The identification and improvement of the annotation of such genes is important to the wine industry, where knowing the detailed characterisation of genes involved in important grape traits lends a competitive advantage to the development of new and improved grape cultivars.

4  **CONCLUDING REMARKS**

The present study was able to identify improvements to the 12Xv2.1 annotation of *V. Vinifera*, with significantly more annotation events than previously reported [21], by using an improved methodology with additional datasets. However, the loss in sensitivity due to a proteogenomics search resulted in a loss of 52% of previously annotated proteins.

This study confirmed the benefits and caveats of previous proteogenomics works, and which has provided a good example of how to share and bring different legacy –omics datasets together for the genomic annotation of *Vitis vinifera*.

Among the 339 annotations identified, 106 novel peptides directly led to 54 predicted proteins. Through the identification of these annotation events a possible over-assembly of chromosome 7 was identified. The methods employed in this study have identified improvements as well as gaps in the understanding of proteogenomics approaches. With the reduced sensitive of a
larger genome there is a more pressing need for a proteogenomics approach that has refined control on the search space and FDR filtering stage of analysis. This requirement could be achieved by segregating and reducing the search space to only necessary sequences for more sensitive identification of novel and previously annotated peptides and to provide better discrimination between true and false positives, as well as reducing the post-processing overhead.

Overall this study contributed significantly to grape genome annotation, correcting several genes important to the grapevine industry, and further understanding the field of proteogenomics. However, further studies are still warranted to utilise larger, diverse and precise MS/MS spectra (QTOF/LTQ Orbitrap) to identify more novel annotations and to provide tighter control on false positives by using varietal specific proteomics and RNA-seq data to complement the target genome. The addition of divergent proteomics and RNA-seq datasets could be included for specific studies targeting the identification of sequence variations between different varieties [31], as well as the inclusion of N-terminomics [32] to improve confident identification of TIS sites. Studies of this type would prove worthwhile once the question of genomic coverage and suspected over-assembly have been addressed, to reduce any spurious identifications. In the meantime, the assembly and annotation would benefit greatly from a review and assessment using tools such as BUSCO [33] to identify and resolve any caveats moving forward with improving the assembly and annotation.

AUTHORS’ CONTRIBUTIONS
BC conceived the project in discussion with MB, BC designed the proteogenomics experiment and BC conducted bioinformatics and statistical analysis. BC and MB wrote the manuscript.

ACKNOWLEDGEMENTS
The authors wish to acknowledge Natalie Castellana for her work on developing the proteogenomics tool Enosi, and for receiving the authors suggestions and assistance with bug fixes during its early development. The authors also wish to thank Sangtae Kim for his work on developing the MS/MS
database search tool, MS-GF+, Steven Van Sluyter for providing guidance and the proteomics dataset, and Ryan Ghan for providing the larger RNA-seq dataset across multiple cultivars. The authors would also like to thank the Centre for Comparative Genomics for their compute resources and guidance and the Pawsey Supercomputing Centre for the use of their compute resources, which were supported by funding from the Australian Government and the Government of Western Australia.

**CONFLICTS OF INTEREST STATEMENT**
The authors have declared that there are no financial conflicts of interest.

5 REFERENCES


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.
predictions, functional annotation, categorization and integration of the predicted gene sequences. 
*BMC Res Notes* 2012, 5, 213.


Figure legends

Figure 1 Customized proteogenomics pipeline

In module 1, a MS/MS spectral pre-processing and optimization step was carried out, following a MS/MS database search against the six-frame translation of the genome, previously annotated proteome and a splice graph. In module 2, 1% PSM FDR, 5% local FDR and 5% peptide-level FDR filtering was applied, and then separated into novel and previously annotated PSMs. Identified previously annotated proteins were filtered by parsimony (≥2 peptides and ≥1 unique peptides). Novel peptides were mapped to the genome. In module 3, novel peptides were clustered within a peptide linkage distance of 18,000bp, peptide clusters with at least 1 unique peptide, annotation events inferred, then filtered based on event probability, spectral counts, peptide length, and sequence homology to annotated protein databases, followed with identification of novel and/or refined gene models.
**Figure 2 Gene boundary annotation**

The gene boundary event inferred from the novel and unique peptides closely flanked reference gene VIT\_218s0001g04980 on chromosome 18. Both the VIT\_218s0001g04980 protein, novel peptides, and Augustus prediction g68936 were annotated as an acetyl-CoA carboxylase 1-like protein (XP\_002285808.2), which is involved in the biosynthesis of fatty acids. The novel peptides were incorporated into the Augustus gene prediction. A group of peptides were also found mapped to gene VIT\_218s0001g04980 indicating its expression and adding confidence to the annotation.
Figure 3 Frame-shift annotation

A frame-shift event for gene VIT_200s1339g00010 on chromosome Un, inferred from the novel and unique peptides. While the VIT_200s1339g00010 protein was annotated poorly as a hypothetical protein with a retrotransposon gag protein (CAN63794.1), the novel peptides and Augustus prediction g83988 were annotated as a hypothetical protein with a Ribonuclease T2 domain (CAN63109.1), which is involved in leaf senescence in order to scavenge phosphate from ribonucleotides. The novel peptides were incorporated into the Augustus gene prediction, changing the frame of the first exon from the original prediction. A region further downstream contained a number of repeats.
**Figure 4 Novel exon annotation**

A novel exon event for gene VIT_217s0000g02480 on chromosome 17, inferred from the novel and unique peptides.

While the VIT_217s0000g02480 protein was annotated as unnamed protein product, with an EF-hand calcium-binding domain (CB15562.3), the novel peptides and Augustus prediction had an improved annotation as a calcium-binding allergen Ole e 8 (XP_010663678.1), which is involved in the activation or inactivation of target proteins. The novel peptides were incorporated into the Augustus gene prediction, bridging the two exon/CDS regions in the original prediction into one exon/CDS region.
Table legends

Table 1 Summary of grape proteogenomics annotations

The results of the proteogenomics analysis of grape 12Xv2.1 annotation.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total 12Xv2.1 genes</td>
<td>31,845</td>
</tr>
<tr>
<td>Total 'previously annotated' protein-coding genes</td>
<td>31,654*</td>
</tr>
<tr>
<td>Total 'previously annotated' proteins</td>
<td>55,373*</td>
</tr>
<tr>
<td>Raw MS/MS search 'previously annotated' protein matches ≤1% PSM FDR</td>
<td>2,773</td>
</tr>
<tr>
<td>Proteogenomics mapping: Total 'previously annotated' proteins ≤1% PSM FDR</td>
<td>7,536</td>
</tr>
<tr>
<td>Proteogenomics mapping: Total 'previously annotated' proteins ≤1% PSM FDR  (≥2 peptides AND ≥1 unique peptides)</td>
<td>1,117</td>
</tr>
<tr>
<td>Total identified 'novel' peptides ≤1% PSM FDR</td>
<td>325</td>
</tr>
<tr>
<td>Total identified 'novel' peptides ≤1% PSM FDR, &lt;5% local FDR AND &lt;5% peptide-level FDR</td>
<td>255</td>
</tr>
<tr>
<td>Raw MS/MS search 'previously annotated' peptides ≤1% PSM FDR</td>
<td>7,886</td>
</tr>
<tr>
<td>Proteogenomics mapping: Total identified 'previously annotated' peptides ≤1% PSM FDR</td>
<td>11,779</td>
</tr>
<tr>
<td>Proteogenomics mapping: Total identified 'previously annotated' peptides ≤1% PSM FDR  (≥2 peptides AND ≥1 unique peptides)</td>
<td>5,048</td>
</tr>
<tr>
<td>Frame-shifts</td>
<td>5</td>
</tr>
<tr>
<td>Translated UTRs</td>
<td>37</td>
</tr>
<tr>
<td>Exon boundaries</td>
<td>15</td>
</tr>
<tr>
<td>Novel splices</td>
<td>1</td>
</tr>
<tr>
<td>Novel exons</td>
<td>9</td>
</tr>
<tr>
<td>Gene boundaries</td>
<td>159 (24)</td>
</tr>
<tr>
<td>Reverse strands</td>
<td>112 (10)</td>
</tr>
<tr>
<td>Novel genes</td>
<td>1</td>
</tr>
<tr>
<td>Total annotation events (eventProb ≥99.80% for novel genes &amp; ≥98.374% for distal/proximal annotation events)</td>
<td>339 (102)</td>
</tr>
</tbody>
</table>

This article is protected by copyright. All rights reserved.
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total genes affected</td>
<td>213 (67)</td>
</tr>
<tr>
<td>Total proteins affected</td>
<td>323 (101)</td>
</tr>
<tr>
<td>Total novel peptides in affected genes/proteins</td>
<td>129</td>
</tr>
<tr>
<td>Total Augustus protein-coding gene predictions</td>
<td>84,948</td>
</tr>
<tr>
<td>Total Augustus protein predictions</td>
<td>93,754</td>
</tr>
<tr>
<td>Total Augustus gene predictions with incorporated novel peptides</td>
<td>52</td>
</tr>
<tr>
<td>Total Augustus protein predictions with incorporated novel peptides</td>
<td>54</td>
</tr>
<tr>
<td>Total novel peptides incorporated into Augustus protein predictions</td>
<td>106</td>
</tr>
<tr>
<td>Improved protein predictions with incorporated novel peptides</td>
<td>47</td>
</tr>
<tr>
<td>Novel non-paralogous protein predictions with incorporated novel peptides</td>
<td>7</td>
</tr>
</tbody>
</table>

* The original consisted of 31,845 protein-coding genes coding for 55,564 proteins. A total of 191 proteins which were <20 aa in length were removed from the analysis.

Note: Numbers in parenthesis represent the non-redundant count of annotation events which have a 1:1 peptide cluster to annotation event association. The inflationary effect of a large peptide linkage distance on gene boundaries and reverse strands was removed by assigning a peptide cluster as either a proximal or distal event, not both, with preference placed on proximal events.