Vitis **55**, 129–133 (2016) DOI: 10.5073/vitis.2016.55.129-133

VitisPathways: gene pathway analysis for *V. vinifera*

M. V. Osier

Rochester Institute of Technology, Rochester, New York, USA

Summary

Pathway enrichment analysis of genetic and proteomic data is fraught with multiple testing and other interpretive issues. A web-based tool, VitisPathways, was developed to simplify the process of pathway analyses for *Vitis* researchers while maintaining statistical robustness. Because enrichment analysis tools outside of pathway analysis have shown non-regularity to multiple test corrections, simulations were used to assess the degree of regularity in *Vitis* knowledgebases and its impact on interpretation. This tool is freely available and can be an aide to hypothesis generation in transcriptomic studies of *Vitis*.

K e y w o r d s : *Vitis vinifera*; pathway; Fishers exact test; gene expression; genomics; proteomics.

Introduction

Large-scale transcriptomic and genomic studies are becoming more common for understanding aspects of *Vitis vinifera* transcriptional control such as development (Sweetman *et al*. 2012, Venturini *et al*. 2013, Chitwood *et al*. 2014) and host-pathogen interactions (Perazzolli *et al*. 2012). These studies often generate large numbers of genes determined to be of interest. On their own, these large lists gene are often daunting. Resources such as VitisNet (Grimplet *et al.* 2012) and GrapeCyc (Plant Metabolic Network, http://www.plantcyc.org/about/databases_overview.faces#grapecyc) connect genes and specific metabolic or signaling pathways as means to structure the results and assist researchers in understanding their large gene lists in the context of metabolic pathways. However, utilizing the resources often require installing and learning to use external tools such as Cytoscape (http://www.cytoscape.org/), or doing pathway analysis "by hand".

To make preliminary pathway analysis quantitative and as accessible as possible to the grape genetics community, a web-based resource, VitisPathways, was developed to assist in prioritizing gene lists through application of the Fishers Exact Test. VitisPathways will work from nearly any browser and requires little input or technical knowledge. No assumptions are made about the nature of the molecular experiment (microarray, RNA Seq, proteomics, etc.). VitisPathways is publicly available from "http://www.rit.edu/VitisPathways". In similar enrichment analysis methods for highly structured data such as the Gene Ontology (Ashburner *et al*. 2000), it has been noted that the structure of the ontology can impact the false positive rate in a non-linear and highly unpredictable fashion (Osier *et al*. 2004). Metabolomics databases also have degrees of structure because the pathways themselves have overlap in genes and metabolites. Due to the large number of functional overlaps of genes, as demonstrated by the large number of accessions at each edge in the VitisNet pathway picture files, on average any two pathways with any shared genes will have 12.1 genes overlapping between them. Some of these are unique genes and some are isoforms. However, in both cases there may be unique activities to these genes; compressing the genes into a single activity instead of counting them individually would be inappropriate. Pathways with more genes will likely have larger overlaps, which could create a bias in which larger pathways would have a larger than linear increase in hits. To test for similar impact from the structure of one of the pathway databases (VitisNet), simulations were also carried out to identify unpredictable effects of structure between pathways on expected results.

Material and Methods

VitisPathways reads in the significant gene list as either VitisNet Unique ID (typically VITxxxxxxxx) or accession numbers used by GrapeCyc (typically GSVIVT01xxxxxxxxx), a short description, the user's email address, and the number of permutations desired (100 or 1000). Genes are associated with pathways and a Fishers Exact Test (FET) is performed using the "fisher.test" subroutine in R (http://www.r-project.org/). All associated pathways are reported by email to the user in a tab-delimited format. In addition to statistical information, all genes associated with a given pathway are reported to allow follow-up hypothesis testing.

Permutations are used to correct for multiple testing. Permuted gene lists are generated from the entire gene list of VitisNet or GrapeCyc respectively. For each pathway with any gene association, FET is again calculated on the permuted gene list. The number of permuted *p*-values at least as significant for the permutated data set as for the user data set are tallied and reported as a corrected *p*-value.

Correspondence to: Dr. M. V. Osier, Rochester Institute of Technology, Gosnell Building 08 - Rm 1338, 85 Lomb Memorial Drive, Rochester, New York 14623, USA. E-mail: mvoscl@rit.edu

Output is sent to the user as a tab-delimited email body containing each pathway, the four gene counts used in the FET analyses, the FET results, the permutation results, and the list of user genes associated with that pathway.

To explore expected distributions for VitisNet analysis and the possible effect of structure on results, simulated data sets were generated as 1000 data files with a specific number of genes (increments of 50 from 50 to 1250) picked randomly from the total gene list. Analysis of simulated sets by VitisPathways was performed without permutations. In these analyses, a "hit" for a pathway means that for a given gene set of a given gene set size, that pathway was determined to be significant by FET. For pathway mean hits, the number of hits is determined across all 1000 simulated data sets for each gene set size, and the mean of those numbers is taken. For gene set mean hits, the number of hits for each pathway across all 1000 simulations is taken, and the mean across all pathways is calculated.

Results

As a test of function and demonstration of the simplicity of the interface, the Switch Genes from the Table of Palumbo *et al*. (2014) were used to compare gene function identification. The VIT accession numbers were input into the Web interface (Fig. 1). Soon thereafter the results were emailed (Fig. 2). Note that the categories are similar to those identified in the Gene Ontology by Palumbo, but more specific. For example, the four genes identified as being related to Carbohydrate Metabolic Process in the Gene ontology are in more specific pathways in VitisNet: vv10010Glycolysis (VIT_14s0068g01760, VIT_14s0060g00420), vv10620Pyruvate_metabolism (VIT_09s0002g06420, VIT_14s0060g00420), and vv10052Galactose_metabolism (VIT_07s0005g01680). However, some of these genes are also associated with other, less obvious, pathways through VitisNet: vv10290Valine_leucine_and_isoleucine_biosynthesis (VIT_14s0060g00420), vv10071Fatty_acid_metabolism (VIT_14s0068g01760), vv10252Alanine_and_aspartate_metabolism (VIT_14s0060g00420), and vv10350Tyrosine_metabolism (VIT_14s0068g01760). The increased specificity and wider range of functional annotation is a byproduct of the level of the Gene Ontology used in the Palumbo *et al*. analysis. The Gene Ontology also allows for a single gene to be associated with multiple terms at different levels of specificity. This should not be taken as a criticism of the Palumbo results, as the number of terms the Gene Ontology will indirectly associate with those four genes is somewhat staggering. Indeed, Palumbo *et al*. picked an excellent level of Gene Ontology to focus on for those gene functions. Also note that two of the non-carbohydrate related pathways were insignificant by FET and all four were insignificant by permutations. The author would deem these four VitisNet pathways to most likely be spurious pathway associations. However, it would be worth considering the relationships between all seven VitisNet pathways, as they clearly have some overlaps.

Simulated data sets of randomly sampled grape genes were used to identify any unpredictable effects such as structure between pathways impacting permutation results. As the size of each gene set increased from 50 to 1250, the mean number of hits across pathways increased linearly ($r^2 = 0.988$). It was also observed in the simulations that some pathways are more likely to appear significant than others. This may be due in part to some data pathways having more genes than others, or those same pathways including genes that are also in other pathways, making these pathways commonly associated with each other. The latter would be the complex structural effects such as was observed in the previously discussed Gene Ontology study. The logarithmic distribution in Fig. 3 is similar to what would be expected if there is no effect of pathway structure (e.g. each gene belongs to only one pathway). Therefore, it appears that the overlap in genes between pathways is not unpredictably impacting results as was observed previously in the Gene Ontology results (Osier *et al*. 2004).

The ten pathways with highest mean hits (number of times significant by FET for a given gene set size) are presented in the Table. In general, these are broad categories which obviously have many genes associated with them. In addition, the categories are well studied in *Vitis*, meaning it is more likely that we have identified the functions of a large number of genes in these categories. Given the large number of genes in these pathways, they also have more overlaps with more pathways. Therefore, it is not unexpected

Table

The top ten most common pathways by simulation hits across gene set sizes

Accession	Name	Genes in pathway	Mean hits
vv23010	Ribosome	612	671.7
vv40006	Cell wall	518	627.9
vv34627	R_proteins_from_Plant-pathogen interaction	472	591.8
vv50101	Channels and pores	450	587.4
vv44810	Regulation of actin cytoskeleton	397	546.2
vv50109	Incompletely characterized transport systems	383	537.9
vv10500	Starch and sucrose metabolism	373	527.8
vv10190	Oxidative phosphorylation	383	526.1
vv34626	Plant-pathogen interaction	360	515.7
vv44110	Cell cycle	355	508.1

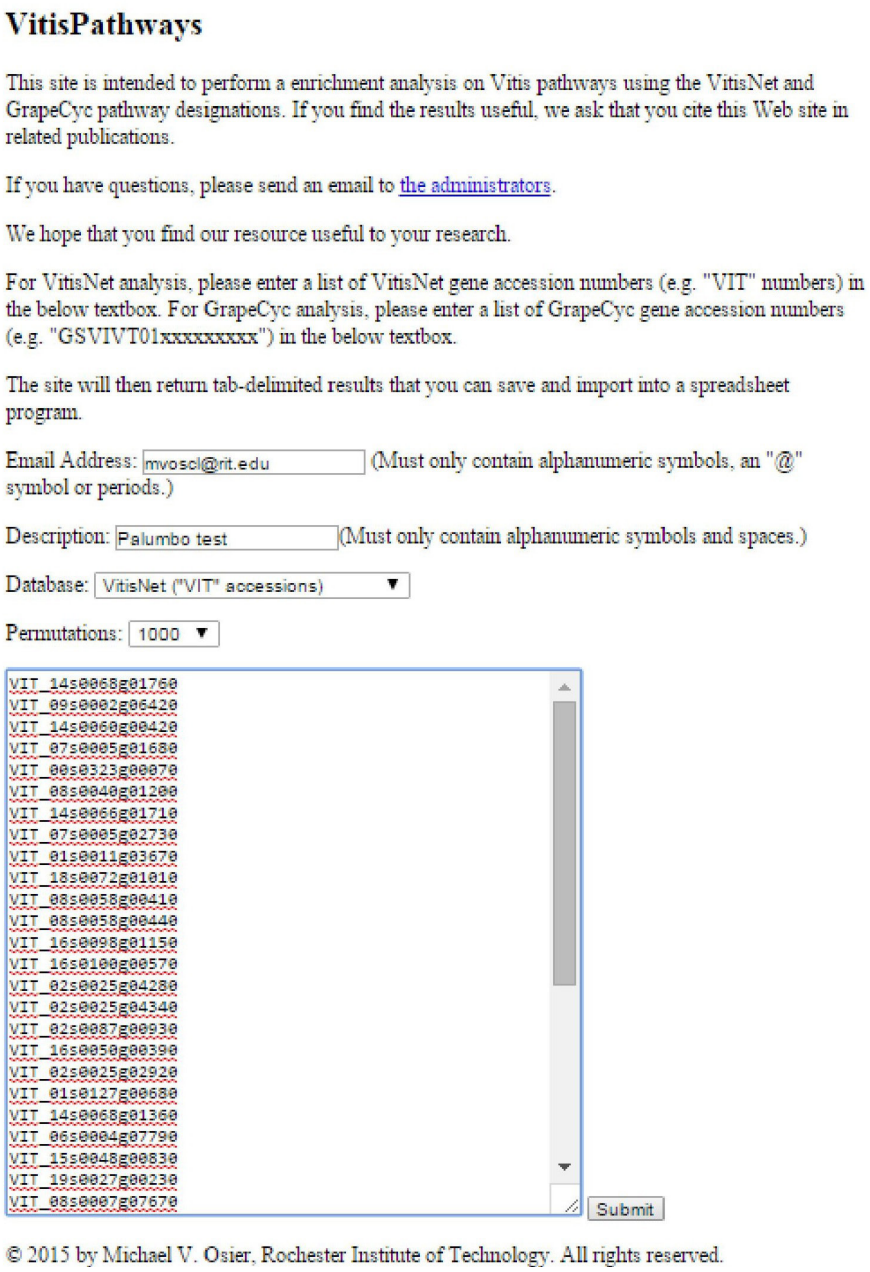

Fig. 1: Entry of Palumbo *et al.* (2014). VIT accessions into the VitisPathways interface.

that any randomly picked gene is more likely to belong to these pathways than to others, which results in more hits by chance.

Discussion

VitisPathways is already being used by scientists researching *Vitis* (Majumdar *et al*. 2015). Key summaries and rankings of pathways allow researchers to develop and prioritize new hypotheses focused on the most robust pathway sets. In the future, the relationships between significant pathways, due to shared genes, will be presented to the user to identify larger biological networks and further refine hypothesis generation.

For the genes in the Table and other terms with large numbers of associated genes, interpretation of a significant result should be viewed with a small degree of caution and in the context of the experiment. For example, in a study of plant tumorigenesis, pathway vv44110 (Cell_cycle), a term with 355 associated genes, would make sense. However, in a study of grape sugar content, the same term should be given more scrutiny before declaring the result to be meaningful. As was suggested in the reanalysis of the Palumbo *et al*. data set, if a large pathway has overlaps with other pathways, it would be worth considering whether interaction between these pathways impacts the experimental condition. A future release of VitisPathways will detail minimal distances between pathways associated with user data, making these inter-pathway relationships clear to users.

As discussed in Methods, the entire gene list is used to create permuted sets because experimental data may also have any genes from the same pool. It should be noted that this assumption may be violated if only custom gene

VitisPathways results 43, Palumbo test
Michael Osier [ ]
Sent: Thursday, July 23, 2015 7:01 AM
To: Michael Osier

Pathway	DE in pathway	not DE in pathway	DE not in pathway	not DE not in pathway	Fisher's Exact Test	permuted p-value	DEGs
vv60007AS2	2	50	21	51405	0.0002498	0.003	VIT_15s0048g00830 VIT_06s0004g07790
vv10941Flavonoid_biosynthesis	1	189	22	51266	0.08155	0.239	VIT_02s0025g02920
vv10071Fatty_acid_metabolism	1	110	22	51345	0.04845	0.14	VIT_14s0068g01760
vv10052Galactose_metabolism	1	175	22	51280	0.07576	0.199	VIT_07s0005g01680
vv10760Nicotinate_and_nicotinamide_metabolism	1	27	22	51428	0.01244	0.041	VIT_01s0127g00680
vv30001ABA_signaling	1	166	22	51289	0.07203	0.209	VIT_02s0087g00930
vv60078Other_zf-C3HC4	3	274	20	51181	0.000252	0.004	VIT_14s0219g00040 VIT_18s0001g01060 VIT_08s0040g01950
vv10940Phenylpropanoid_biosynthesis	1	242	22	51213	0.1031	0.304	VIT_16s0050g00390
vv11013ABA_biosynthesis	1	17	22	51438	0.008013	0.034	VIT_02s0087g00930
vv10860Porphyrin_and_chlorophyll_metabolism	2	73	21	51382	0.0005195	0.004	VIT_08s0058g00410 VIT_08s0058g00440
vv10906Carotenoid_biosynthesis	1	48	22	51407	0.02167	0.063	VIT_02s0087g00930
vv50104Group_translocators	1	41	22	51414	0.0186	0.054	VIT_16s0050g00390
vv10010Glycolysis	2	221	21	51234	0.004451	0.035	VIT_14s0060g00420 VIT_14s0068g01760
vv50121Porters_cat_1_to_6	1	196	22	51259	0.08443	0.242	VIT_19s0014g04790
vv10620Pyruvate_metabolism	2	228	21	51227	0.004727	0.046	VIT_09s0002g06420 VIT_14s0060g00420
vv30003Auxin_signaling	1	309	22	51146	0.1297	0.325	VIT_16s0098g01150
vv50123Porters_cat_18_to_29	1	224	22	51231	0.09586	0.251	VIT_06s0009g01140
vv30009Flower_development	1	203	22	51252	0.0873	0.252	VIT_07s0005g02730
vv10290Valine_leucine_and_isoleucine_biosynthesis	1	72	22	51383	0.03212	0.086	VIT_14s0060g00420
vv40006Cell_wall	1	517	22	50938	0.2076	0.541	VIT_00s0323g00070
vv10350Tyrosine_metabolism	1	164	22	51291	0.07119	0.189	VIT_14s0068g01760
vv60046NAC	2	84	21	51371	0.0006822	0.003	VIT_08s0007g07670 VIT_19s0027g00230
vv60045MYBrelated	1	70	22	51385	0.03125	0.101	VIT_07s0005g02730
vv10252Alanine_and_aspartate_metabolism	1	116	22	51339	0.051	0.149	VIT_14s0060g00420
vv23015mRNA_surveillance_pathway	1	134	22	51321	0.05862	0.163	VIT_18s0072g01010

Fig. 2: Emailed results of test data set in Figure Y. The tab-delimited email body can be easily exported from the email message and imported into a spreadsheet.

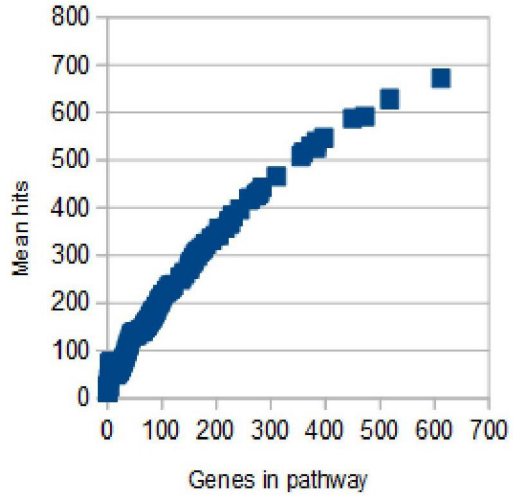

Fig. 3: Relationship between the number of genes and number of hits in simulation. As the number of genes in a pathway increases, the number of times that pathway was significant (hits) in random data sets increased asymptotically toward 100 % (1000 simulations). Mean hits are the mean number of times that pathway was significant by FET across all data set sizes (e.g. 50 to 1250 genes).

lists are considered in the experiment generating the gene list (e.g. custom microarray). Under such conditions, all statistical results should be considered suspect. However, with modern full transcriptome methods such as RNA-Seq, there will be no problem.

The statistical and interpretation considerations should be the same for any pathway analysis which uses VitisNet as its knowledgebase. The results suggest that the observed relationships are due to the internal structure of the knowledgebase, and not the method of analysis. Given that other metabolomic knowledgebases have similar structure, as differences in metabolic pathways should be minimal, results should extend to other grape and non-grape knowledgebases, such as GrapeCyc or AraCyc (https://www.arabidopsis.org/biocyc/). However, it would be worth further study to test if these concerns are general across species. This would require extensive simulation.

Conclusions

Pathway analyses can be a very useful tool to comprehend large scale patterns among genes deemed to be significantly differentially expressed. However, some pathways are more likely to be identified by these methods than others, especially as the number of genes analyzed increases. Therefore, pathway analysis should be treated as a tool for further hypothesis generation and not necessarily given as much weight as the results of formal inference. Although permutation-based correction of FET p-values can be useful in identifying false leads, secondary validation of genes in highlighted pathways and mechanisms of interaction is critical. VitisPathways provides the analytical methods in a single, freely available tool with easily utilized interface to guide understanding of experimental results and new hypothesis generation in *Vitis* genetics.

Acknowledgements

The author wishes to thank Dr. L. Cadle-Davidson for his suggestions regarding this manuscript and his assistance in understanding the genomic peculiarities of *Vitis*.

References

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G.; 2000: Gene Ontology: tool for unification of biology. Nat. Genet. **25**, 25-29.

Chitwood, D. H.; Ranjan, A.; Martinez, C. C.; Headland, L. R.; Thiem, T.; Kumar, R.; Covington, M. F.; Hatcher, T.; Naylor, D. T.; Zimmerman, S.; Downs, N.; Raymundo, N.; Buckler, E. S.; Maloof, J. N.; Aradhya, M.; Prins, B.; Li, L.; Myles, S; Sinha, N. R.; 2015: A modern ampelography: A genetic basis for leaf shape and venation patterning in grape. Plant Physiol. **164**, 259-272.

Grimplet, J.; Van Hemert, J.; Carbonell-Bejerano, P.; Diaz-Riquelme, J.; Dickerson, J.; Fennell, A.; Pezzotti, M.; Martinez-Zapater, J. M.; 2012: Comparative analysis of grapevine whole-genome gene predictions, functional annotation, categorization and integration of the predicted gene sequences. BMC Res. Notes **5**, 213.

Majumdar, R.; Brooks, S.; Lillis , J.; Osier, M.; Reisch, B.; Cadle-Davidson, L.; 2015: Silencing of Grapevine Pectate Lyase-like Genes VvPLL2 and VvPLL3 Confers Resistance Against *Erysiphe necator* and Differentially Modulates Gene Expression. Int. Plant and Animal Genome Conference, Jan 10-14, 2015, San Diego, CA, USA (Poster presentation).

Osier, M. V.; Zhao, H.; Cheung, K. H.; 2004: Handling multiple testing while interpreting microarrays with the Gene Ontology. BMC Bioinformatics **5**, 124.

Palumbo, M. C.; Zenoni., S.; Fasoli, M.; Massonnet, M.; Farina, L.; Castiglione, F.; Pezzotti., M.; Paci, P.; 2014: Integrated network analysis identifies fight-club nodes as a class of hubs encompassing key putative switch Genes that induce major transcriptome reprogramming during grapevine development. Plant Cell **26**, 4617-4635.

Perazzoli, M.; Moretto, M.; Fontana, P.; Ferrarini,A.; Velasco, R.; Moser, C.; Delledonne, M.; Pertot, I.; 2012: Downy mildew resistance induced by *Trichoderma harzianum* T39 in susceptible grapevines partially mimics transcriptional changes of resistant genotypes. BMC Genomics **13**, 660.

Sweetman, C.; Wong, D. C. J.; Ford, C. M.; Drew, D. P.; 2012: Transcriptome analysis at four developmental stages of grape berry (*Vitis vinifera* cv. Shiraz) provides insights into regulated and coordinated gene expression. BMC Genomics **13**, 691.

Venturini, L.; Ferrarini, A.; Zenoni, S.; Tornielli, G. B.; Fasoli, M.; Santo, S. D.; Andrea, M.; Buson, G.; Tononi, P.; Zago, E. D.; Zamperin, G.; Bellin, D.; Pezzotti, M.; Delledonne, M.; 2013: *De novo* transcriptome characterization of *Vitis vinifera* cv. Corvina unveils varietal diversity. BMC Genomics **14**, 41.

Received March 1, 2016