Journal of Plant Biochemistry & Physiology

Alagna, J Plant Biochem Physiol 2013, 1:2
doi: 10.4172/jpbp.1000e107

Editorial Open Access

Innovative Transcriptomics Approaches for Large Scale Identification of Genes Involved In Plant Secondary Metabolism

Fiammetta Alagna*

CNR – Institute of Plant Genetics (IGV), Via Madonna Alta 130, 06128 Perugia, Italy

Importance of Plant Secondary Metabolites

Plants synthesize an impressive number of different secondary metabolites involved directly in the interaction with other organisms and with environment or indirectly in the regulation of plant responses [1,2]. Moreover many secondary metabolites hold beneficial effects on human health and are used as bioactive components of drugs [3]. Plant nutraceuticals, natural plant food and use of nutritional therapies and phytotherapies have become progressively popular to improve health and to prevent and treat diseases. Improving the traits related to plant secondary metabolism (i.e. defense responses, fitness, stress tolerance, nutraceutical value, etc.) has become a main target of plant breeding and biotechnology industry [4].

Despite the importance of plant secondary metabolites there is a big gap of knowledge on the genes involved in their biosynthesis, accumulation, and degradation. A limited number of pathways have been completely clarified and in most of the cases the study was limited to model plants. Nevertheless, this scenario is rapidly changing as new tools for the identification of genes encoding for entire metabolic pathways and those involved in their regulation emerged.

Advances in Sequencing Technologies and Data Mining for the Prediction of Genes for Secondary Metabolite Pathways

The deep-sequencing technologies (NGS - Next generation sequencing), recently developed, allow to deliver large amount of fast and inexpensive sequence information [5], at the same time the progress in bioinformatic approaches, both as hardware and software for the analysis of data, permit an ever improving management and mining of such large datasets. RNA-Seq, a recently developed approach for transcriptome profiling by mean of deep-sequencing technologies, provides a far more precise measurement of the level of gene transcripts and their isoforms than other methods [6].

The diffusion of new technologies have rapidly increased the release rate of sequence and expression data. The large amount of sequence information now available for plant genomes and transcriptomes, provides an opportunity to identify genes involved in secondary metabolic pathways in many plant species of agronomical interest.

The analysis of differential gene expression is the main approach in the identification of transcripts involved in particular plant traits. The approach is based on the individuation of two experimental conditions that possibly determine differences circumscribed to the investigated phenomenon (i.e. a mutant *vs.* wild type, presence or absence of an elicitor, different tissues or developmental stages etc.) [7].

An effective approach to predict genes involved in the same metabolic pathway is the co-expression analysis. A set of genes involved in a biological process can be co-regulated and thus co-expressed under the control of a shared regulatory system, therefore, if a gene with unknown function is co-expressed with known genes of a particular metabolic pathway, this gene has potentially a role in the same pathway [8]. Co-expression analysis can be conducted using datasets from RNA-seq or microarray obtained in expressly designed experiments or also by comparing already existing data publicly available [9]. Further functional characterization of the identified genes demonstrated the effectiveness of this approach [10].

Some mutations or stress conditions share a similar effect on particular metabolic pathways. Combining co-expression analysis and metabolic profiling in the appropriate conditions it is possible to realize a prioritization of the genes involved in such pathways [8,11].

Gene Clusters for Secondary Metabolites

Until recently, it was thought that genes for plant metabolic pathways were randomly spread along the genome, and this was certainly true in some cases; however it is now becoming increasingly clear that genes for the synthesis of many major classes of plant-derived secondary metabolites are organized into clusters, reminiscent of the operons and metabolic gene clusters found in microbes [12,13]. These clusters consist of groups of physically linked genes that are functionally related and co-regulated. Unlike operons, genes within these clusters are transcribed separately [13].

Operon-like gene clusters can be identified by integrating expression data and information about physical position of genes in the genome [14]. Recently, Field and Osbourn [15] have shown that computational approaches are effective to predict gene clusters in plants. Extensive gene expression data can be used to identify clusters of co-expressed genes with predicted functions in secondary metabolism. It can be evaluated whether physically linked genes show statistically significant co-expression compared to other randomly selected genes. A high number of candidate clusters have been recently predicted utilizing publicly available expression profile data jointly with genome assembly data [16].

Expression Quantitative Trait Locus (eQTL)

Phenotypic differences are both the result of sequence polymorphisms that produce altered (or absent) proteins and the result of qualitative and quantitative differences in gene expression that generate varying amounts of protein in a cell or tissue [17]. Up to know, the contribution of transcript abundance variation to the phenotypic diversity still remains under appreciated. Nevertheless, in the last years these studies are undergoing a strong impulse thanks to the availability of large scale mRNA profiling technologies [18,19].

***Corresponding author:** Fiammetta Alagna, CNR – Institute of Plant Genetics (IGV), Via Madonna Alta 130, 06128 Perugia, Italy, E-mail: fiammetta.alagna@igv.cnr.it

Received May 06, 2013; **Accepted** May 08, 2013; **Published** May 15, 2013

Citation: Alagna F (2013) Innovative Transcriptomics Approaches for Large Scale Identification of Genes Involved In Plant Secondary Metabolism. J Plant Biochem Physiol 1: e107. doi:10.4172/jpbp.1000e107

If transcript levels are measured across individuals of a genetic mapping population the recorded variation in mRNA transcript abundance for each gene may be treated as a heritable trait that can be subjected to statistical genetic analyses. These analyses allow identifying the chromosomal region that controls the observed variation (eQTLs) [20].

Large scale mRNA profiling technologies allow the genome-wide mapping of thousands of eQTLs in a single experiment. A single gene can have one or multiple eQTLs. When combined with classical or trait QTLs, correlation analyses can directly suggest candidates for genes underlying these traits [17]. The mRNA profiling data sets can also be used to infer the chromosomal positions of thousands of genes, an outcome that is particularly valuable for species with unsequenced genomes where the chromosomal location of the majority of genes remains unknown.

As plant genotypes strongly differ for the content of specific secondary metabolites and a transcriptional regulation of the pathways has been observed in some cases, the eQTL analyses can contribute to clarify the genetic regulatory networks which control these traits.

In the next years, improved eQTL mapping strategies will strongly increase the knowledge of the contribution of non-coding polymorphisms to gene expression regulation providing new tools to determine the genetic variants underpinnig the broad diversity of the plants.

Integration of Transcriptomics, Proteomics and Metabolomics Data

As it was discussed in the previous paragraphs, the transcriptomics is a powerful tool for the identification of the genes involved in plant secondary metabolism, however the integration of transcriptomics with proteomics or metabolomics can strongly increase the quantity of information that can be obtained and offer new possibilities to explore the extraordinary complexity of plant biochemical capacity.

One important aspect regards the identification of variably spliced transcripts and the discrete proteins they encode. The integration of next-generation gene expression data with proteomics data provides information on the functional significance of these isoforms clarifying which is the percentage expressed as proteins. Proteomic data can also be used to inform genome annotation and characterize post-translational modification as has been demonstrated in a number of recent studies [21,22]. These studies will lead to novel testable hypotheses regarding the connection between genotype and phenotype.

In recent years, technologies for the analysis of metabolites have made spectacular advances both in terms of the number of metabolites that are identified and of throughput. Recent developments allow the construction of metabolic networks and study of the role of these networks in the plant biological processes. Metabolite data sets in combination with other types of data, such as expression profiles or proteomics data can be used to generate hypotheses about functional relationships [23,24]. At the simplest level, correlation networks can be used to identify which components might be functionally related based on the "guilt by association" principle. As discussed below, the integration of transcriptomics and metabolomics have been successfully used to identify many genes regulating the metabolic pathways as those involved in glucosynolate and flavonoid biosynthesis [23,25,26]. Moreover, correlating mass peaks with transcripts could be a powerful strategy for identifying metabolites in complex extracts [27].

Much effort to create bioinformatics tools able to combine data coming from different sources are on-going. It is expected that these tools will strongly increase the level of comprehension of the plant system.

Closing Remarks and Future Perspectives

In the last years, thanks to the next generation sequencing methodologies, a huge amount of sequence data on plant genomes and transcriptomes has been released and much more will be made available in the coming years. The availability of these resources makes the use of genome mining and other computational approaches powerful tools for the study of genes involved in secondary metabolite pathways. Now more than ever this kind of studies can have a strong impulse and can be extended to a large number of plant species.

Co-expression analysis and genome mining approaches based on these data sets have become a powerful tool to elucidate gene function. They can be used to find effective candidate genes for entire secondary metabolic pathways. This is particularly useful in those species where finding the genes of a pathway is particularly difficult by traditional functional genomic techniques (e.g. gene silencing, overexpression), due to the long time required for the regeneration and growth of plants and the lack of effective mutagenesis populations.

The discovery that genes for plant secondary metabolic pathways are organized in clusters has open unprecedented opportunities to discover entirely new metabolic pathways and chemistries of agronomical and pharmaceutical importance, an approach that has been highly successful in microbes. As the volume of available genome sequence information for different plant species is continuously increasing, it will be possible to determine how common secondary metabolic clusters are organized in plant genomes and what kind of compounds these clusters may produce.

Reduced sequencing costs will also increase the number of studies that use RNA sequencing data to perform eQTL mapping, which will increase our knowledge of how gene expression variation contributes to phenotypic variation.

Finally, the combination and the integration of different –omics data will contribute, in the near future, to a comprehensive systems-level understanding of plants.

References

1. Metlen KL, Aschehoug ET, Callaway RM (2009) Plant behavioural ecology: dynamic plasticity in secondary metabolites. Plant Cell Environ 32: 641-653.
2. Forester SC, Waterhouse AL (2009) Metabolites are key to understanding health effects of wine polyphenolics. J Nutr 139: 1824S-31S.
3. Pereira DM, Valentao P, Correia-da-Silva G, Teixeira N, Andrade PB (2012) Plant secondary metabolites in cancer chemotherapy: where are we? Curr Pharm Biotechnol 13: 632-650.
4. Zhao J (2007) Nutraceuticals, nutritional therapy, phytonutrients, and phytotherapy for improvement of human health: a perspective on plant biotechnology application. Recent Pat Biotechnol 1: 75-97.
5. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31-46.
6. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.
7. Alagna F, D'Agostino N, Torchia L, Servili M, Rao R, et al. (2009) Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. BMC Genomics 10: 399.
8. Saito K, Hirai MY, Yonekura-Sakakibara K (2008) Decoding genes with

coexpression networks and metabolomics - 'majority report by precogs'. Trends Plant Sci 13: 36-43.

9. Naoumkina MA, Modolo LV, Huhman DV, Urbanczyk-Wochniak E, Tang Y, et al. (2010) Genomic and coexpression analyses predict multiple genes involved in triterpene saponin biosynthesis in Medicago truncatula. Plant Cell 22: 850-866.
10. Carelli M, Biazzi E, Panara F, Tava A, Scaramelli L, et al. (2011) Medicago truncatula CYP716A12 is a multifunctional oxidase involved in the biosynthesis of hemolytic saponins. Plant Cell 23: 3070-3081.
11. Urbanczyk-Wochniak E, Baxter C, Kolbe A, Kopka J, Sweetlove LJ, et al. (2005) Profiling of diurnal patterns of metabolite and transcript abundance in potato (Solanum tuberosum) leaves. Planta 221: 891-903.
12. Chu HY, Wegel E, Osbourn A (2011) From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. Plant J 66: 66-79.
13. Osbourn A (2010) Gene clusters for secondary metabolic pathways: an emerging theme in plant biology. Plant Physiol 154: 531-535.
14. Field B, Fiston-Lavier AS, Kemen A, Geisler K, Quesneville H, et al. (2011) Formation of plant metabolic gene clusters within dynamic chromosomal regions. Proc Natl Acad Sci U S A 108: 16116-16121.
15. Field B, Osbourn AE (2008) Metabolic diversification--independent assembly of operon-like gene clusters in different plants. Science 320: 543-547.
16. Wada M, Takahashi H, Altaf-Ul-Amin M, Nakamura K, Hirai MY, et al. (2012) Prediction of operon-like gene clusters in the Arabidopsis thaliana genome based on co-expression analysis of neighboring genes. Gene 503: 56-64.
17. Druka A, Potokina E, Luo Z, Jiang N, Chen X, et al. (2010) Expression quantitative trait loci analysis in plants. Plant Biotechnol J 8: 10-27.
18. Cubillos FA, Coustham V, Loudet O (2012) Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. Curr Opin Plant Biol 15: 192-198.
19. Majewski J, Pastinen T (2011) The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet 27: 72-79.
20. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17: 388-391.
21. Baginsky S, Gruissem W (2006) Arabidopsis thaliana proteomics: from proteome to genome. J Exp Bot 57: 1485-1491.
22. Castellana N, Bafna V (2010) Proteogenomics to discover the full coding content of genomes: a computational perspective. J Proteomics 73: 2124-2135.
23. Hirai MY, Klein M, Fujikawa Y, Yano M, Goodenowe DB, et al. (2005) Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. J Biol Chem 280: 25590-25595.
24. Stitt M, Sulpice R, Keurentjes J (2010) Metabolic networks: how to identify key components in the regulation of metabolism and growth. Plant Physiol 152: 428-444.
25. Yonekura-Sakakibara K, Tohge T, Matsuda F, Nakabayashi R, Takayama H, et al. (2008) Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in Arabidopsis. Plant Cell 20: 2160-2176.
26. Sawada Y, Kuwahara A, Nagano M, Narisawa T, Sakata A, et al. (2009) Omics-based approaches to methionine side chain elongation in Arabidopsis: characterization of the genes encoding methylthioalkylmalate isomerase and methylthioalkylmalate dehydrogenase. Plant Cell Physiol 50: 1181-1190.
27. Liberman LM, Sozzani R, Benfey PN (2012) Integrative systems biology: an attempt to describe a simple weed. Curr Opin Plant Biol 15: 162-167.