Phylogenetic footprinting: Using the natural variation among related species to identify transcription factor binding sites upstream of developmentally regulated genes

Kenneth Morales[1] and Brian G. Ayre[2]

[1] Texas Academy of Math and Science, University of North Texas, Denton TX

[2] Department of Biological Sciences, University of North Texas, Denton, TX

Background: A Prelude to Footprinting

Phloemeomics (*flō-mē-ō-mics*): the science of understanding phloem structure and function by cataloguing and characterizing all genes expressed in the phloem.

Identifying phloem specific genes

To identify genes that are specifically expressed in leaf phloem, an enhancer trap genetic screen was carried out.

➢Tissue-specific gene expression is mediated by transcription factors binding to enhancers in DNA to promote transcription (Figure 1).

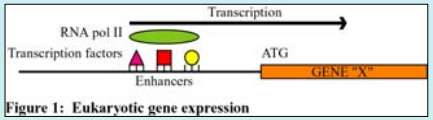


Figure 1: Eukaryotic gene expression

➢In an enhancer trap screen, a promoterless reporter gene is randomly integrated into the genomes of a large population of individuals. In the model plant *Arabidopsis thaliana*, this random insertion is achieved by *Agrobacterium tumefaciens* mediated transformation, and the reporter gene encodes the ß-glucuronidase, or GUS, enzyme (Figure 2).

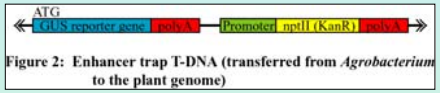


Figure 2: Enhancer trap T-DNA (transferred from *Agrobacterium* to the plant genome)

➢The integrated *GUS* gene is expressed in response to nearby enhancer elements, and may assume the expression pattern of a tissue-specific, flanking gene (Figure 3).

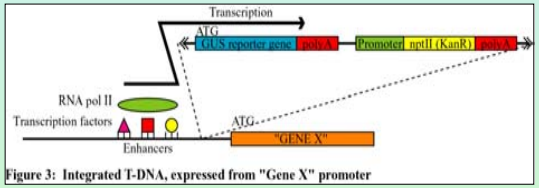


Figure 3: Integrated T-DNA, expressed from "Gene X" promoter

➢Of 2000 independent enhancer-trap lines, 1% demonstrated GUS expression in leaf veins. A sampling: 64-1F, minor vein expression following the sink-to-source transition; 64-8A, expression in large veins of mature and immature leaves; 59-B4, expression in large veins of mature leaves; 64-12C, expression in all veins of mature and immature leaves.

The enhancer-trap DNA of line 64-1F was inserted between two oppositely orientated genes on chromosome 4: At4g28630, encoding an ATP-binding cassette (ABC) transporter, and At4g28640, an auxin inducible transcription factor (IAA 11). The intergenic sequence was fused to GUS and resulted in phloem-specific gene expression in both orientations (Figure 4). See poster by McGarry, Turgeon, and Ayre for more information.

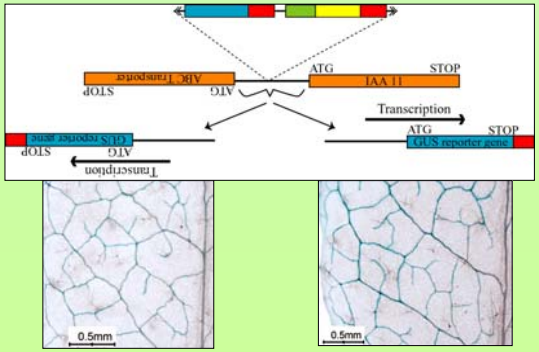


Figure 4: Top: Representation of the enhancer trap T-DNA inserted between two oppositely orientated genes. **Middle:** GUS reporter gene fused to either end of the intergenic region. **Bottom:** Representative blue coloration resulting from the GUS enzyme activity.

Phylogenetic Footprinting: The science of identifying potential transcription factor binding sites by comparing promoter sequences from related species, based on the principle that these sites are conserved over evolutionary time to maintain the proper expression of orthologous genes (Figure 5; Ayre et al. 2003).

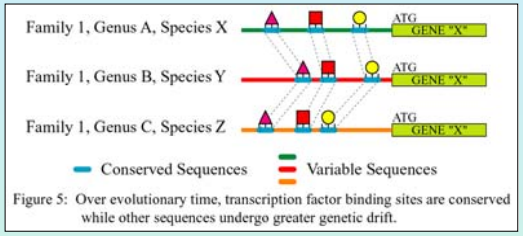


Figure 5: Over evolutionary time, transcription factor binding sites are conserved while other sequences undergo greater genetic drift.

Based on Arabidopsis and *Brassica oleracea* (cauliflower) sequences available in public databases, oligonucleotides corresponding to conserved sequences where used to obtain the intergenic regions from diverse Brassicaceae (mustard family) members using the polymerase chain reaction (PCR). *Brassica oleracea* and *Sinapis alba* (white mustard) sequences are complete and aligned with Arabidopsis sequence to identify conserved sequences (Figure 6).

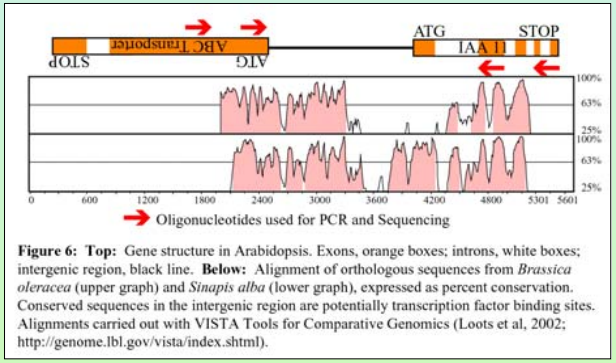
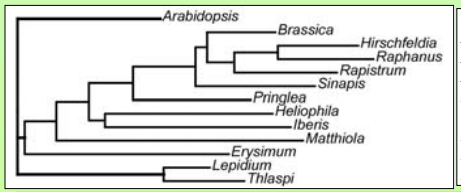


Figure 6: Top: Gene structure in Arabidopsis. Exons, orange boxes; introns, white boxes; intergenic region, black line. **Below:** Alignment of orthologous sequences from *Brassica oleracea* (upper graph) and *Sinapis alba* (lower graph), expressed as percent conservation. Conserved sequences in the intergenic region are potentially transcription factor binding sites. Alignments carried out with VISTA Tools for Comparative Genomics (Loots et al, 2002; http://genome.lbl.gov/vista/index.shtml).

Future Work: Sequencing orthologus regions from other Brassicaceae family members continues (Figure 7). These sequences will be aligned using several different algorithms. We anticipate that addition of more species to the alignment will increase the resolution of conserved sequences. If sequence conservation in the intergenic region is too high (*i.e.*, insufficient genetic drift at non-essential sequences), DNA from species outside the Brassicaceae will be obtained [other Brassicales (capers), Malvales (cotton), or Sapindales (citrus)].



Genomic DNA for future analysis was isolated from:	
Anastatica hierochuntica	*Iberis amara*
Arabidopsis thaliana	*Ionopsidium abulense*
Arabis purpurea	*Lepidum sativum*
Brassica oleracea	*Morisa monanthos*
Chorispora tenella	*Matthiola ovatifolia*
Erysidium cheiri	*Rapistrum Rugosum*
Hirschfeldia incana	*Sinapsis alba*
Heliophila longifolia	*Thlapsi arvense*

Figure 7: Left: Phylogenetic relationships in the Brassicaceae based on the internal transcribed spacer 1 sequence flanking the 5.8S ribosomal RNA gene. **Right:** DNA was isolated and a three kilobase orthologous fragment was PCR amplified from the indicated species for future analysis.

Ayre BG, Blair JE, Turgeon R (2003) Functional and phylogenetic analyses of a conserved regulatory program in the phloem of minor veins. Plant Physiol **133:** 1229-1239.

Loots G, Ovcharenko I, Pachter L, Dubchak I, Rubin E (2002) rVISTA for comparative sequence-based discovery of functional transcription factor binding sites. Genome. Res. **12:** 832-839.