ALPHABET SOUP FOR BEETS: STATUS OF ESTS, BACS, RILS AND OTHER GENOMIC SUNDRIES

J.M MCGRATH

USDA-ARS, Sugarbeet and Bean Research Unit, 494 PSSB, Michigan State University, East Lansing, MI 48824-1325 USA

ABSTRACT

Dogma holds that phenotype = genotype + environment; DNA makes RNA makes protein; and form follows function. What this means is that the beet's work is accomplished in large part by proteins; that proteins (via genes) are inherited from the parents; and expression of genes is influenced by environment (and also development). By understanding beet proteins deduced from gene sequences, whose function can be inferred from other well-characterized protein forms, we can begin to build a conceptual framework for the types of work that a beet must accomplish in order to be profitable to growers and industry. This report considers the progress in building the tools that will enable such a framework. These prerequisites are mainly the tools of genomics, which encompass everything you wanted to know about the inner workings of beets (but were afraid to ask). For instance, as of February 2003, over 19,500 Expressed Sequence Tags are available, a 5X coverage Bacterial Artificial Chromosome library has been constructed, and 5,000 Recombinant Inbred Lines are being developed. These efforts have and will continue to require close cooperation among ARS, industry, and academic scientists. These tools are freely available now and will likely remain so in the future. Already, problems previously considered intractable are beginning to yield insight upon application of these tools. Progress is likely to accelerate in the future, as these genomics investments can be leveraged with scientific expertise inside and outside of the sugar beet community.

INTRODUCTION

Genomics is a logical extension of concepts developed over the past 200 years that combine plant and animal breeding, cell biology and biochemistry, genetics, and molecular biology and physiology. It is integrative in that these disparate disciplines are united at the description and function of the myriad cell types in various tissues at the level of the gene. Genomics attempts, therefore, to describe the structure and function of every gene in the genome, in every cell and tissue type, and begin to understand the hierarchy of gene interactions that ultimately result in phenotype.

SUGAR BEET GENOMICS

Functional genomics, or global analyses of gene expression, fills a gap between traditional biochemical analyses and the genetic instructions for these gene products encoded in the DNA, as represented by expressed RNA molecules (e. g. genes). Analyses of these transcripts by nucleotide sequencing (or other methods) reveals information about the identity and abundance of a specific transcript, the diversity of transcripts present in specific cells and tissues, and the biochemical complexity of an organism. Potentially novel or unexpected solutions to specific developmental or environmental cues may also be evident from such analyses.

Gene expression is the realization of genetic potential. Gene expression results in phenotype, which itself is the interaction of genotype and environment. While the genotype is accessible through inheritance and selection over generations in different environments, phenotypic expression of specific traits is often limited spatially or temporally, in effect accumulating throughout the growing season. Phenotypic responses to abiotic (and biotic) stresses may be predictable, but complex, particularly in combinations that would be expected under the diversity of field environments where beets are grown. Global gene expression analyses can help understand this complexity by determining which gene products are regulated under each type of stress. Each regulated gene product would have some probability for involvement in response to stress, and would represent a target for breeding and selection. Global gene expression analyses have been unavailable to plant breeding prior to large scale sequencing projects.

A prerequisite to global gene expression profiling is knowledge of nucleotide sequences of expressed genes. The typical approach to gaining this information has been to sequence cDNA (complimentary or copy DNA, reverse transcribed from mRNA) clones, which by definition are derived from expressed genes. These sequences are compared for similarity to the ever-increasing number of nucleotide sequences held in databases, and sequences with high similarity to genes with known function are used to assign putative functions. A catalog of expressed genes is often generated by mass sequencing of cDNA libraries, each cDNA clone being sequenced a single time, resulting in a collection of Expressed Sequence Tags (ESTs). ESTs by definition are preliminary and unsubstantiated indications of actual nucleotide sequences.

Depending on the level of similarity between any two sequences, a putative protein functional class can be assigned for many ESTs (Burks, 1999). Most functional classes belong to biochemical pathways, so a complete set of nucleotide sequences for an organism defines the biochemical reactions that can occur. Differential expression of genes, particularly among genes of a common pathway, provides a measure of the importance of any particular biochemical process in a particular environment. This information can be used directly for selection and breeding.

The first complete nucleotide sequence of a plant genome, *Arabidopsis thaliana*, was completed in 2000 (The *Arabidopsis* Genome Initiative, 2000), providing an unparalleled opportunity to access plant genes. Gross characterization of *Arabidopsis* gene content revealed features that have relevance for plant improvement. First, *Arabidopsis* has on the order of 25,000 genes. In relation to other fully sequenced multicellular eukaryotic genomes of fruit fly (*Drosophila melanogaster*) and nematode (*Caenorhabditis elegans*), *Arabidopsis* shared most similarity in genes with basic metabolic functions and shared least similarity in genes that sense and respond to environmental and developmental signals. Sugar beet is expected to be similar at a gross level to *Arabidopsis*, although differences in gene regulation, gene copy number, and presence or absence of specific gene classes might be expected.

One of the tasks for sugar beet research will be to determine which specific genes are of interest in germplasm improvement. Having a list of genes expressed in sugar beet is one of the earliest objectives that need to be accomplished. It is perhaps cost prohibitive to sequence the entire sugar beet genome presently, but EST projects are more affordable and a great deal of progress on this objective has occurred recently. Insufficient time has elapsed to have fully explored these new resources, and rapid progress can be expected. As of February 2003, over 20,000 *Beta vulgaris* nucleotide sequences had been deposited in the National Center for Biotechnology Information (i.e. GenBank, www.ncbi.nlm.nih.gov). The majority of these are ESTs (19,617 sequences). ESTs have been submitted by three independent groups (USDA-ARS East Lansing, Max Plank - Cologne, and the GABI project) from mRNAs expressed in seedlings germinating under stress, four week old roots, mature roots, storage roots, leaves, and inflorescences. The GABI set is unique in that clones were pre-selected prior to sequencing to remove a large proportion of redundant transcripts (Herwig et al. 2002), and thus represents a 'unigene' set of over 10,000 unique expressed gene sequences covering the important developmental stages of beet growth. Sugar beet researchers now have perhaps one third of the expected expressed genes available to evaluate.

Expressed gene sequences contain information required for the translation of the genetic code into proteins, the molecules that accomplish much of the cell's work. They do not generally contain information required for the correct expression of their respective proteins; however these instructions are often located close to the gene, generally immediately adjacent. Thus, complete characterization of a gene involved in expression of an agronomic trait requires sequencing of these promotor regions. As perhaps 60% of the 750 million base pair beet genome is comprised of non-protein encoding sequences, isolating the adjacent sequences to expressed genes can be problematic. Bacterial clones are available with very large segments of sugar beet DNA inserted with them and the task of screening such large insert libraries is proportionately less intensive. This strategy has been successful in numerous genomics programs, and a number of BAC (bacterial artificial chromosome) libraries have been constructed for beet. Additionally, BAC libraries are useful starting points for complete genome sequencing, as well as estimating the number of genes similar to any particular EST, as a measure of genetic redundancy in the beet genome.

In collaboration with USDA-ARS scientists at Fargo, ND; Ft. Collins, CO; and Salinas, CA, a BAC library with five-fold genome coverage (38,400 clones) was constructed from *Hin*DIII-digested sugar beet hybrid USH20, with an average insert size of 100 – 125 kb. Filter arrays were prepared that contained all clones and were used to assess the abundance and distribution of particular types of nucleotide sequences via filter-hybridization approaches. Using a ribosomal

RNA gene probe, 1.2% (450 clones, estimated to total 9,500 copies of a presumed 10 kb repeat unit) of the library carried sequences similar to these highly repetitive, highly conserved sequences located on Chromosome 1 of the Butterfass trisomic series (Schondelmaier & Jung 1997). A simple sequence repeat element $(CA)_8$ thought to be predominantly distributed throughout centromere regions of all chromosomes was present in 1.6% of clones (Schmidt & Heslop-Harrison 1996). A probe for the telomere canonical sequence $(TTTAGGG)_7$ only hybridized with seven BAC clones; however this region at the end of chromosomes is difficult to clone and was not expected to be well represented in this library, and may represent interstitial relics from previous inversion events. Organelle DNA (plastid and mitochondria) contamination was assessed with organelle-specific DNA probes. Chloroplast DNA contamination was greater than mitochondrial DNA (1.6 % of clones vs. 0.01%, respectively).

Twenty-eight randomly chosen ESTs were screened against nylon filter arrays of the BAC library (Table 1). These sequences represent a small sampling of structural and regulatory gene sequences. Assuming 5X coverage, the number of gene copies similar to a particular EST in the beet genome was estimated from the number of hybridization signals. For over half of the ESTs used as probes, a greater than expected number of hybridization signals were observed for a single copy sequence, suggesting that many genes are duplicated in the beet genome. It is possible that some of these duplicated sequences provide strict redundancy of gene function, while others may have sufficiently diverged and may have altered gene expression patterns or functions.

ESTs, as representatives of expressed genes, and BACs, as representatives of the position and number of these genes in the beet genome, provide virtually no information on the agronomic importance of these nucleotide sequences. These traits can be correlated with genetic position through various genetic mapping approaches yet knowledge of the gene functions that underlie agronomic traits are not easily discerned. Correlating gene identity with agronomic function is an important goal, and one approach to achieve this is via integration of gene expression profiling and physical and genetic maps. Since beets are out-crossing and wind pollinated, relatively large amounts of heterozygosity are present in populations available for genetic analyses. Heterozygosity, i.e. genetic variability, adds to environmental variability in measurements of field performance, and reduces precision of genetic analyses in beets.

Recombinant Inbred Lines (RILs) help to accomplish two goals simultaneously. First is the reduction of heterozygosity through inbreeding, and the attendant advantage of potentially allowing better environmental variance estimates. Second is genetically mapping agronomic traits more precisely by allowing large seed productions of defined, identical-by-decent genotypes for multi-location, multi-year estimates of quantitative agronomic traits. Currently, a target for development is 50 RIL populations of 100 individuals each, derived from single seed descent of individual self-fertile hybrid plants for five or six generations. A large range of germplasm is being used, including disease resistance donor germplasm, high and low sucrose breeding lines, and various crop and wild relatives of sugar beet.

Table 1: Estimated gene number of ESTs deduced with filter hybridization of the BAC library.

Putative EST function	Genbank ID	Estimated # genes
ABC transporter	BI543560	1
adenine triphosphatase	BI543538	3
allergen	BI095948	2
beta amyrin synthase	BF011005	1
germin-like protein	AF310017	8
calmodulin	BI096069	1
carboxyphosphonoenol pyruvate mutase	AW697745	1
cystein protease	BE590278	1
enolase	BI543290	1
heat shock protein 81-2	AW697750	1
heat shock protein	BI543424	1
hydroxymethyltransferase	BI095900	7
malate dehydrogenase	BI073206	2
UDP glucose pyrophosphorylase	BI096068	3
alcohol dehydrogenase	AW697786	1
aquaporin	BI643109	4
sucrose synthase	BI543240	8
ribosomal RNA genes	pTA71	950
ribulose bisphosphate carboxylase	BI643066	3
MAP kinase	BQ060614	1
UDP-glucose glucosyltransferase	BI073142	2
ribulose phosphate 3-epimerase	BI073233	2
pyruvate dehydrogenase-1	BI073208	3
pyruvate dehydrogenase-2	BI096005	6
hexokinase	BI543276	1
glyceraldehyde 3P dehydrogenase	BI095991	4
isocitrate lyase	BI095941	1
14-3-3 like protein	BI543270	1
phosphofructokinase	BI096032	2

CONCLUSION

Sugar beet genomics is an extension of traditional breeding and modern genetic methods. Its fundamental utility lies in the ability to define specific elements of the genome, initially in terms of nucleotide sequence and later in terms of specific function. Many long-standing production problems related to variety will be accessible through genomic analyses, and biochemical mechanisms for breeding and selection efficiency will be evident. Sugar beet breeders need to incorporate this knowledge into breeding programs, and should be involved in interpreting genomic information

REFERENCES

1. Burks, C. 1999. Molecular biology database list. Nucleic Acids Research 27:1-9.

2. Gindullis F, Dechyeva D, Schmidt T. 2001. Construction and characterization of a BAC library for the molecular dissection of a single wild beet centromere and sugar beet (*Beta vulgaris*) genome analysis. Genome 44: 846-855.

3. Herwig R, Schulz B, Weisshaar B, Hennig S, Steinfath M, Drungowski M, Stahl D, Wruck W, Menze A, O'Brien J, Lehrach H, Radelof U. (2002) Construction of a 'unigene' cDNA clone set by oligonucleotide fingerprinting allows access to 25 000 potential sugar beet genes. Plant J. 32: 845-857.

4. Loudet O, Chaillou S, Merigout P, Talbotec J, Daniel-Vedele F. (2003) Quantitative trait loci analysis of nitrogen use efficiency in Arabidopsis. Plany Physiol. 131: 345-358.

5. Schmidt T, Heslop-Harrison JS (1996) The physical and genomic organization of microsatellites in the sugar beet. Proc. National Academy of Sciences USA 93: 8761-8765.

6. Schondelmaier J, Jung, C (1997) Chromosomal assignment of the nine linkage groups of sugar beet (*Beta vulgaris* L.) using primary trisomics. Theoretical and Applied Genetics 95:590-596.

7. The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature (London) 408:796-815.